In Search of a Common Thread: Enhancing the LBD Workflow with a view to its Widespread Applicability

by

Kumarage Menasha Silva Thilakaratne

Supervised by

Professor Katrina Falkner

Dr Thushari Atapattu

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in the
Faculty of Engineering Computer & Mathematical Sciences
School of Computer Science
THE UNIVERSITY OF ADELAIDE

November 2020

Abstract

Literature-Based Discovery (LBD) research focuses on discovering implicit knowledge linkages in existing scientific literature to provide impetus to innovation and research productivity. Despite significant advancements in LBD research, previous studies contain several open problems and shortcomings that are hindering its progress. The overarching goal of this thesis is to address these issues, not only to enhance the *discovery component* of LBD, but also to shed light on new directions that can further strengthen the existing understanding of the LBD workflow. In accordance with this goal, the thesis aims to enhance the LBD workflow with a view to ensuring its *widespread applicability.*

The goal of *widespread applicability* is twofold. Firstly, it relates to the adaptability of the proposed solutions to a diverse range of *problem settings.* These problem settings are not necessarily application areas that are closely related to the LBD context, but could include a wide range of problems beyond the typical scope of LBD, which has traditionally been applied to scientific literature. Adapting the LBD workflow to problems outside the typical scope of LBD is a worthwhile goal, since the intrinsic objective of LBD research, which is discovering novel linkages in text corpora is valid across a vast range of problem settings.

Secondly, the idea of *widespread applicability* also denotes the capability of the proposed solutions to be executed in *new environments.* These 'new environments' are various academic disciplines (i.e., *cross-domain* knowledge discovery) and publication languages (i.e., *cross-lingual* knowledge discovery). The application of LBD models to new environments is timely, since the massive growth of the scientific literature has engendered huge challenges to academics, irrespective of their domain.

This thesis is divided into five main research objectives that address the following topics: *literature synthesis*, *the input component*, *the discovery component*, *reusability*, and *portability.* The objective of the *literature synthesis* is to address the gaps in existing LBD reviews by conducting the first systematic literature review. The *input component* section aims to provide generalised insights on the suitability of various *input types* in the LBD workflow, focusing on their role and potential impact on the information retrieval cycle of LBD.

The *discovery component* section aims to intermingle two research directions that have been under-investigated in the LBD literature, 'modern word embedding techniques' and 'temporal dimension' by proposing *diachronic semantic inferences.* Their potential positive influence in knowledge discovery is verified through both *direct* and *indirect* uses. The *reusability* section aims to present a new, distinct viewpoint on these LBD models by verifying their reusability in a timely application area using a methodical reuse plan. The last section, *portability*, proposes an *interdisciplinary LBD framework* that can be applied to *new environments.* While highly cost-efficient and easily pluggable, this framework also gives rise to a new perspective on knowledge discovery through its generalisable capabilities.

Succinctly, this thesis presents novel and distinct viewpoints to accomplish five main research objectives, enhancing the existing understanding of the LBD workflow. The thesis offers new insights which future LBD research could further explore and expand to create more efficient, widely applicable LBD models to enable broader community benefits.

Declaration of Authorship

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I acknowledge that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Kumarage Menasha Silva Thilakaratne

To all researchers involved in the pursuit of new knowledge

Acknowledgements

I take this opportunity to express my profound gratitude and deepest appreciation for all of those who helped and contributed in every possible way to make my PhD journey a success. First of all, I am truly grateful to my two supervisors, Professor Katrina Falkner and Dr Thushari Atapattu, for their immense guidance, constantly monitoring my progress, inspiring me towards excellence and encouraging me throughout this research. Thank you for trusting me to develop new ideas in the way I saw fit, enabling me to explore so many novel areas which I was never previously aware. This is something that I appreciate greatly, since it built my confidence in approaching a problem (despite the positive or negative results) and deep inside myself I feel like I am becoming a bigger and better LBD researcher. Without your encouragement, deeply insightful comments and incredible support, I would never have been able to explore my topic so deeply. I am very fortunate to have you both throughout my PhD journey.

I am thankful to my research group, CSER (Computer Science Education Research) for their tremendous support and encouragement. Even though my research topic deviated slightly from the remaining research in my group, they never stopped supporting me by attending to my university milestones, 3MT presentations, and so on. I thank Dr Rebbeca Vivian for her continuous interest in my research topic, which was a huge encouragement for me. I would also like to thank my internal examiner, Dr Christoph Treude, for his insightful comments and feedback, all of which helped to make this research a great success. Furthermore, I wish to express my thanks to the postgraduate coordinators, Professor Ian Reid, Professor Frank Neumann and Professor Markus Wagner, for their readiness to help me with the Graduate Center milestones.

Moreover, I would like to extend my gratitude towards the worldwide community of LBD researchers who communicated with me via emails amid their busy schedules. I am truly grateful for their cooperation and support. I still remember how I accidentally found the topic of LBD while searching for research topics at the beginning of my candidature. Looking back, I feel very fortunate to have picked this topic, as it helped me to shape my future alongside a very friendly research community. Furthermore, I express my gratitude to other research communities, including those involved in researching the semantic web, sequence analysis and signal processing, for inspiring me to explore these new directions and providing technical guidance when required. I would like to thank Mr. Joseph Miller for his professional service in proofreading my thesis. I am grateful to my fellow PhD students (including the optimisation research group) in our lab, for their friendly chats and support with completing forms and with other university procedures.

Last but not least, I thank my lovely family for being with me during the days and nights, providing the fullest support in all my pursuits and encouraging my academic interests from day one. In a way, PhD was an emotionally rough journey for me, since I often see my loved ones virtually rather than physically. I always wished that I could have had my family nearby. I also want to thank my beloved husband for being very understanding and supportive throughout the process. You never stopped encouraging me and taking care of me, which is invaluable.

Pursuing a PhD is undoubtedly the biggest dream I had in my life, so I cannot thank you all enough for your incredible support in making my dream a reality.

Publications

The following *peer-reviewed publications* arose from the work conducted in this thesis. The publications have been organised chronologically.

1. **Information Extraction in Digital Libraries: First Steps towards Portability of LBD Workflow** (*related to Chapter 7 in the thesis*)
   - *Authors:* *Menasha Thilakaratne*, Katrina Falkner, Thushari Atapattu
   - *Conference Venue:* JCDL 2020 (CORE Rank: A*)

2. **Garbage in, Garbage out? An Empirical Look at Information Richness of LBD Input Types** (*related to Chapter 4 in the thesis*)
   - *Authors:* *Menasha Thilakaratne*, Katrina Falkner, Thushari Atapattu
   - *Conference Venue:* JCDL 2020 (CORE Rank: A*)

3. **Connecting the Dots: Hypotheses Generation by Leveraging Semantic Shifts** (*related to Chapter 5 in the thesis*)
   - *Authors:* *Menasha Thilakaratne*, Katrina Falkner, Thushari Atapattu
   - *Conference Venue:* PAKDD 2020 (CORE Rank: A)
   - *Travel Award:* ECMS Travel Scholarship

4. **A Systematic Review on Literature-Based Discovery: General Overview, Methodology, & Statistical Analysis** (*related to Chapter 2 in the thesis*)
   - *Authors:* *Menasha Thilakaratne*, Katrina Falkner, Thushari Atapattu
   - *Journal Venue:* ACM Computing Surveys 2019 (CORE Rank: A*, Impact Factor: 6.131)

5. **A Systematic Review on Literature-Based Discovery Workflow** (*related to Chapter 2 in the thesis*)
   - *Authors:* *Menasha Thilakaratne*, Katrina Falkner, Thushari Atapattu
   - *Journal Venue:* PeerJ-CS 2019 (Impact Factor: 3.09)

6. **Automatic Detection of Cross-Disciplinary Knowledge Associations** (*Doctoral Consortium Proposal*)
   - *Authors:* *Menasha Thilakaratne*, Katrina Falkner, Thushari Atapattu
   - *Venue:* Student Research Workshop, ACL 2018 (CORE Rank of Main Conference: A*)
   - *Travel Award:* Student Research Workshop Travel Grant by Roam Analytics
   - *Others:* Student volunteer @ ACL 2018

Presentations

- **Let Computers Discover Your Next 'Winning' Research Idea**
  - *Presenter:* *Menasha Thilakaratne*
  - *Competition:* Three Minute Thesis (3MT) - 2018
  - *Awards:* Engineering, Computer & Mathematical Sciences Faculty Finalist

Other Publications

- **Detecting Cognitive Engagement using Word Embeddings within an Online Teacher Professional Development Community**
  - *Authors:* Thushari Atapattu, *Menasha Thilakaratne*, Rebecca Vivian, Katrina Falkner
  - *Journal Venue:* Computers & Education 2019 (Impact Factor: 5.296)

- **What Do Linguistic Expressions Tell Us about Learners? Confusion? A Domain-independent Analysis in MOOCs**
  - *Authors:* Thushari Atapattu, Katrina Falkner, *Menasha Thilakaratne*, Lavendini Sivaneasharajah, Rangana Jayashanka
  - *Journal Venue:* IEEE TLT 2020 (Impact Factor: 2.315)

Contents

List of Figures

List of Tables

Nomenclature

Acronyms/Abbreviations

LBD	Literature-Based Discovery
ML	Machine Learning
DNN	Deep Neural Network
LSTM	Long Short Term Memory
CNN	Convolutional Neural Network
DTM	Dedicated Trajectory Model
FTM	Feature-based Trajectory Model
TAM	Trajectory Alignment Model
RDF	Resource Description Framework
URI	Uniform Resource Identifier
LOD	Linked Open Data
W3C	World Wide Web Consortium
SPARQL	SPARQL Protocol And RDF Query Language
DCTERMS/DCT	Dublin Core Metadata Terms
SKOS	Simple Knowledge Organization System

Notations

$(x_1, ..., x_n)$	Ordered list of n pairwise distinct elements $(\mathrm{x_i})_{i=1}^{n}$
*dbr:*x	Denoting that x is a DBpedia resource
*dbc:*x	Denoting that x is a DBpedia category

Chapter 1

Introduction

1.1 Problem Definition

The scientific literature is growing at an unprecedented rate and it is estimated that the global scientific output doubles every nine years (Bornmann & Mutz 2015). To date, scientific digital libraries consist of millions of research publications, with thousands of these being added every day (Masic & Milinovic 2012). For instance, consider *MEDLINE*[1], a popular bibliographical database. It contains more than 26 million journal articles, mainly in the fields of *life sciences* and *biomedicine* (Guo et al. 2020, Jha et al. 2018). The MEDLINE database is updated with nearly 2000-4000 scientific papers on a daily basis (Masic & Milinovic 2012, Lu et al. 2015). This enormous growth of scientific literature and its easy accessibility via World Wide Web (WWW) has opened up massive opportunities for scientists to explore novel research directions (Jha, Xun, Wang & Zhang 2019).

However, at the same time, this overwhelming amount of information has created huge barriers for scientists to make connections with their work from other disciplines (Pratt & Yetisgen-Yildiz 2003, Cohen & Hersh 2005). It is widely accepted that solutions derived through *interdisciplinary scientific problem solving* are more impactful and innovative than solutions proposed within the same problem domain (Chen 2016, Lavrač et al. 2020, Tang et al. 2012, Rzhetsky et al. 2015, Kostoff 2002). Nevertheless, this massive influx of scientific literature has made it extremely difficult for scientists to identify suitable

[1] `https://www.nlm.nih.gov/bsd/medline.html`

cross-domain topics that complement their own areas of study (Hristovski et al. 2005, Weeber 2007). More specifically, researchers typically specialise in limited branches of knowledge. Thus, researchers from each area of academic specialisation only see a part of the big picture, which often leads to difficulty in identifying complementary cross-domain topics (Hristovski et al. 2005, Lindsay & Gordon 1999).

Consider a scientist who is interested in exploring novel research directions in *dementia*. To construct a scientifically sensible novel research hypothesis, the scientist is required to analyse the existing and emerging knowledge in the literature and combine the observations in a creative way to form a hypothesis (Weeber et al. 2005, Brown 2020). At the time of writing, a simple search in *MEDLINE alone* for the query 'dementia' results in more than 210,000 scientific articles. Even if the scientist decided only to investigate research published in the past 12 months, MEDLINE would still return more than 13,000 records.

Despite this staggering amount of information, the *reading ability* of humans has remained the same over the years. In 2012, it was reported that US scientists read 264 papers per year on average, which is similar to the figure recorded in an identical survey conducted in 2005 (Wang et al. 2019). In light of this sheer volume and the rapid growth of scientific literature, it is obvious that no one will be able to keep abreast of all the advancements across the entire body of the literature (Preiss et al. 2015, Pratt & Yetisgen-Yildiz 2003, Yetisgen-Yildiz 2006). Consequently, potentially valuable cross-silo linkages in the literature tend to remain unnoticed. This indicates the need to develop tools that efficiently search knowledge in the literature to assist researchers in forging novel research hypotheses (Swanson 2008, Smalheiser 2017). In this regard, novel advances in *text summarisation* techniques may assist researchers to some extent by providing them with a high-level overview of the literature (Jha et al. 2018). However, such tools are not tailored to capture the novel knowledge linkages made between seemingly distinct knowledge areas in the literature (Jha et al. 2018, Jha, Xun, Wang & Zhang 2019).

Motivated by this, *Literature-Based Discovery (LBD)* research focuses on developing efficient knowledge discovery models that elicit new, implicit knowledge linkages from existing cross-domain scientific facts (Gopalakrishnan et al. 2019). Given the sheer volume of scientific knowledge, LBD is becoming an increasingly important tool in the research development process. For instance, *Arrowsmith* (Torvik & Smalheiser 2007), which was initiated by the pioneers of the LBD discipline and is considered to be the most popular and well-maintained LBD tool in the discipline (Sebastian et al. 2017*a*) has approximately 1200 unique monthly users (Smalheiser et al. 2009). The escalating benefits that LBD tools offer, as well as their practicality and capacity to accelerate innovation have attracted more and more research contributions from the text mining community. Smalheiser, a pioneer of the discipline, defines LBD as follows (Smalheiser 2012):

"LBD refers to a particular type of text mining that seeks to identify nontrivial assertions that are implicit, and not explicitly stated, within (generally a large body of) documents."

1.2 Role of Literature-Based Discovery (LBD)

The ultimate goal of LBD research is to bridge undiscovered research gaps in the existing scientific knowledge to provide impetus to *research progress* and increase *research productivity* (Jha et al. 2018, Xun et al. 2017). This process will also connect isolated facts into one interconnected knowledge space by introducing new interdisciplinary research directions (Palmer & Fenlon 2010, Skeels et al. 2005).

For instance, consider the research collaborations between *biology* and *computer science*, which evoked the revolutionary *bioinformatics* discipline (Tang et al. 2012). Due to these cross-domain collaborations, biology tasks such as *DNA sequencing* and *protein structure modelling* (which were originally very time-consuming and expensive) have become more scalable and affordable (He et al. 2008, Eswar & Sali 2009, Wei & Zou 2016). Similarly, the field of *medical informatics* was created from cross-domain collaborations between *medicine* and *data mining*, which undoubtedly had a massive impact on the development of medicine as a discipline (Tang et al. 2012). In essence, *interdisciplinary scientific problem solving* has a huge influence on society (Tang et al. 2012). Thus, insights derived through LBD models for such cross-domain research directions are becoming increasingly important (Chen 2016).

LBD was developed as a research field based on the ground-breaking studies of Swanson since 1986. These studies demonstrated the possibility of detecting undiscovered cross-silo knowledge in the literature (Swanson 1986, 1988). The underlying notion of

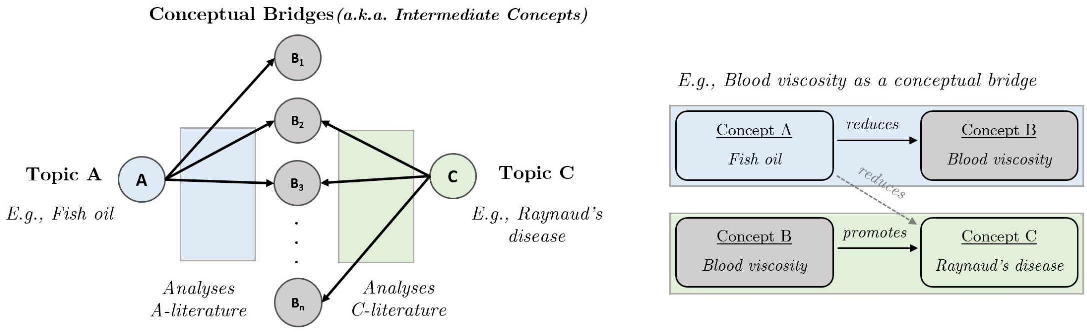


FIGURE 1.1: Schematic overview of the LBD setting

TABLE 1.1: LBD terminology

Component	**Alternative Terminology**
Topic A	Start topic/concept, Source topic/concept
A-literature	Start literature, Source domain
Topic C	Target topic/concept
C-literature	Target literature/domain
Conceptual bridges	B-concepts, Intermediate concepts, Novel knowledge bridges, Novel knowledge linkages

Swanson's work is that within the scientific literature, there exist *complementary* and *non-interactive* structures that can lead to interesting and novel discoveries (Maclean & Seltzer 2012, Hu et al. 2006). As such important linkages are not indexed or cross-cited; they may not be accessible through mere *customary methods* of keyword and citation searching, and, thus require a more detailed and systematic knowledge discovery process, as in LBD (Lindsay & Gordon 1999).

For instance, consider two disjointed topics of interest A and C (see Figure 1.1); a therapeutic substance (e.g., *fish oil*) and a disease (e.g., *Raynaud's disease*), where the objective of the LBD process is to explore novel ways to meaningfully connect these two disjointed areas of knowledge (e.g., *blood viscosity*, as illustrated in Figure 1.1) (Jha et al. 2018). The most critical characteristic of LBD models is their ability to identify novel cross-silo knowledge, even if the articles in the two domains A and C have not cited or co-cited each other. This aspect of LBD ensures that it is able to detect knowledge bridges between seemingly uncorrelated pieces of information (Swanson 1986).

Table 1.1 summarises the *terminology* used in connection with the LBD setting illustrated in Figure 1.1. The rest of the thesis utilises the terminology outlined in Table 1.1 interchangeably to denote each of the main components of the LBD setting.

1.3 Research Context and Objectives

The problem of eliciting novel knowledge from unstructured text started gaining attention following the publications of Swanson's seminal studies since 1986, as discussed in Section 1.2. Even though Swanson's initial studies laid the groundwork for the discipline, the knowledge synthesis method underlying his LBD process was labour-intensive and time-consuming (Jha et al. 2018). Since then, different computational models were proposed by the LBD community to facilitate the knowledge discovery process in a more automated manner (Smalheiser 2017, Sebastian et al. 2017*a*, Henry & McInnes 2017, Gopalakrishnan et al. 2019, Cohen & Hersh 2005).

Despite the significant progress in the field of LBD over the past few decades, there are several open research issues and technical shortcomings in the discipline that this thesis intends to address. This section discusses five main research objectives, which were proposed with respect to these identified research deficiencies in the LBD literature. The main aim of this thesis is to enhance the existing understanding of the LBD workflow in order to enable its *widespread applicability.*

1.3.1 Main Research Objective 1

To integrate a large-scale systematic literature review procedure of LBD studies, in order to address the limitations in the existing traditional narrative-based LBD reviews, while shedding light on novel focus areas in the LBD workflow.

Problem Setting

Literature reviews are an essential part of any research discipline, since they involve assessing and analysing pertinent literature as well as providing valuable insights for future research. Even though several literature reviews have been published on the subject of LBD (Gopalakrishnan et al. 2019, Henry & McInnes 2017, Sebastian et al. 2017*a*, Smalheiser 2017, Ahmed 2016, Smalheiser 2012, Kostoff et al. 2007, Bekhuis 2006, Ganiz et al. 2005, Weeber et al. 2005, Davies 1989), these follow the *traditional narrative form* of collecting, analysing and synthesising the literature. Despite the valuable contributions of these LBD reviews in shaping the field of LBD and its position today in the text mining community, these traditional narrative-based LBD reviews suffer from several limitations, including their *restrictive scope* and *limited focus points*. For instance, it is evident that none of the existing LBD reviews focuses on the *LBD workflow* (i.e., the input component, discovery component, output component, evaluation component, and the overall process - including reusability and portability) as a whole. Furthermore, most of the existing LBD reviews have restricted their scope to medical-related LBD studies. To strengthen the existing understanding of the LBD workflow and to promote its *widespread applicability*, this thesis conducts a large-scale, domain-independent literature synthesis with a broader, more comprehensive scope than that of existing reviews.

Following this notion, conducting a *systematic literature review* (a well-known research method with multiple strengths, including *transparency*, *clarity*, *equality*, *accessibility*, *impartial inclusive coverage*, *replicability*, *objectivity*, *scientific rigour*, *focus* and *unity* (Frangieh & Yaacoub 2017, Boell & Cecez-Kecmanovic 2015)) is particularly critical in the LBD field for two main reasons. Firstly, there are now almost 35 years of published LBD research; secondly, the field is continuously growing and evolving. As such, there is ample scope for a systematic literature review of the subject.

Systematic literature reviews play a pivotal role in any academic discipline since they are considered the *gold standard* among reviews (Snyder 2019). They follow a rigorous and transparent approach to ensure the future replicability of results through the use of a clear systematic *review protocol*, and to minimise any bias in results by focusing on empirical evidence rather than preconceived knowledge (Mallett et al. 2012). While addressing the limitations of traditional narrative-based reviews in the LBD discipline (such as *restrictive scope* and *limited focus points*), this systematic literature review also aims to shed light on several new areas that future LBD research could contribute towards enabling its *widespread applicability*.

1.3.2 Main Research Objective 2

To investigate the input component of the LBD workflow in order to deduce the suitability of different input types in the LBD process.

Problem Setting

The *input* is one of the most critical components in the LBD workflow as the entire knowledge representation and reasoning of the discovery process relies on it (Henry & McInnes 2017). As with other text mining tasks, low-quality input will impact the LBD results and will ultimately impact decisions that are made based on those results (Corrales et al. 2015). However, there is no consistent selection of the LBD input and different studies have picked different input types (Henry & McInnes 2017). These include *title only* (Swanson & Smalheiser 1997), *title and abstract* (Sebastian et al. 2017*b*), *full-text* (Lever et al. 2018), *keywords* (Jha et al. 2018), and highly specialised input resources such as *clinical patient records* (Symonds, Bruza & Sitbon 2014), and *case reports* (Smalheiser et al. 2015).

Among these input types, *title and abstract* is the most common selection. However, LBD pioneers have consistently adopted the *title* of the research publications as the LBD input since the inception of the field (Swanson et al. 2006, Swanson & Smalheiser 1999). Exemplifying this practice, *Arrowsmith*, the most popular and well-maintained LBD tool in the field (Sebastian et al. 2017*a*), only supports the analysis of titles when making predictions (Torvik & Smalheiser 2007). Some studies argue that using *title/abstract* may introduce noise and be computationally expensive, thereby using a special form of keywords called *MeSH (Medical Subject Headings)*[2] as their input (Jha et al. 2018).

Despite these discussions in the LBD literature, to the best of our knowledge, *no previous LBD research* has explicitly attempted to perform *any sort of assessment* to verify these conclusions. Considering the cruciality of the input component in the LBD process, it is vital to understand the role of input types in the LBD workflow, as well as their impact on the overall knowledge discovery process. Such explorations will allow for the construction of better LBD models in the future. Thus, this thesis explores potential definitions to assist in the comprehension of different LBD input types to establish the first steps towards understanding the input component in a *generalisable* manner.

[2] `https://www.nlm.nih.gov/mesh/meshhome.html`

1.3.3 Main Research Objective 3

To enhance the discovery component of the LBD workflow using fine-grained diachronic semantic inferences by conjoining global semantic relationships with the temporal dimension to enrich the typical static cues used in the LBD literature.

Problem Setting

Notwithstanding the significant progress in LBD research over the last few decades, almost all prior LBD studies have neglected the importance of scrutinising the temporal evolution of scientific topics in digital libraries (Jha et al. 2018, Jha, Xun, Wang & Zhang 2019). Consequently, these LBD studies have mainly relied on a *static snapshot* of literature (i.e., assuming that the knowledge in the domain remains static) to discover novel knowledge linkages. This may be limiting, as scientific knowledge evolves continuously with the constant addition of new information from on-going research (Jha et al. 2018, Jha, Xun, Wang & Zhang 2019).

Therefore, integrating the *dynamic nature* of knowledge into the LBD workflow may provide rich cues to further enhance the identification of novel knowledge linkages in the scientific literature. More recently, a few studies have attempted to mitigate the assumption of *static domains* made in previous LBD studies through the infusion of temporal information of scientific topics into the LBD process (Jha et al. 2018, Jha, Xun, Wang & Zhang 2019, Xun et al. 2017). Even though these few recent studies undoubtedly ameliorate the typical knowledge discovery process, the *temporal analysis component* of these studies is fairly shallow. For example, Xun et al. (2017) only considered the first and last values of the time series when measuring the temporal trend of a scientific topic neglecting the subtle patterns that could reside in the time series as a whole. Nevertheless, a fine-grained analysis of the time series may provide further promising cues towards discovering the novel knowledge linkages with high precision. With this in mind, the current thesis explores the need to perform a *circumstantial temporal analysis* in the context of LBD, in order to capture novel cross-silo connections. Such an analysis, may represent an improvement over *static cues* (employed in almost all previous LBD studies) and *shallow temporal cues* (employed in emerging LBD studies that incorporate temporal information into the LBD workflow (Xun et al. 2017)).

Despite the wide spectrum of techniques employed to enhance the *predictions of the LBD process* over the last few decades, this thesis also observes that most of these previous LBD studies rely on *one (or at most two to three) characteristic(s)* to elicit new knowledge (Sebastian et al. 2017*a*, Henry & McInnes 2017). For instance, in a recent LBD study, Jha et al. (2018) have only considered two characteristics, namely *global transformation* and *local transformation* to discover potential novel linkages. The use of one or a few characteristic(s) to define novel knowledge linkages may be limiting for two main reasons.

Firstly, due to the complexity of natural language usage (that causes intricate structures in the scientific literature), identification of novel knowledge linkages using one (or a handful) of characteristic(s) may not be sufficient. In other words, for a knowledge linkage to be labelled a *potential novel knowledge linkage*, it may need to fulfil multiple factors or characteristics. Therefore, the use of one or limited characteristics may inhibit the model's ability to discover novel knowledge linkages more precisely.

Secondly, in the theoretical LBD literature, it has been identified that novel knowledge can reside in the literature in different forms. For example, Davies (1989) identified five forms of novel knowledge in the *'Fish oil-Raynaud's disease'* and *'Migraine-Magnesium'* test cases. Therefore, reliance on one or limited characteristics in the knowledge discovery process may hinder the model's ability to identify novel knowledge linkages in different forms. It may also result in situations where the LBD model may disproportionately be picking only one or limited forms of novel knowledge based on the single or limited characteristics utilised in the LBD workflow. With this problem in mind, this thesis attempts to verify the potential benefits of *defining multiple meaningful characteristics* in the knowledge discovery process, with the goal of further enhancing the prediction performance of novel knowledge linkages.

Most prior LBD research relies on a *query-specific local corpus* to discover potential new knowledge in the LBD process (Jha et al. 2018). Otherwise stated, to capture the interactions of scientific topics, they focus on cues at the *local scale*. This may be limiting, since a local-scale analysis may not necessarily convey a detailed picture of scientific topic interactions. For example, when analysing *'COVID-19'* literature in the LBD workflow, it may be important to identify how scientific topics in the *COVID-19* literature have interacted with other related topics such as *'SARS'*. However, a query-specific local corpus could fail to convey such implicit interactions that may require for complex semantic deductions. Thus, this thesis intends to view the interactions of the scientific topics through a wider lens by incorporating the *global picture of topic interactions* in the LBD workflow.

Bearing in mind, the aforementioned research deficiencies observed in the discovery component of the LBD workflow, this thesis attempts to make the most of these neglected components in the prior LBD studies by accommodating ideas involving the *temporal dimension of the scientific literature* and *large-scale feature analysis* using *the global picture of topic interactions* to enhance its *discovery component*.

1.3.4 Main Research Objective 4

To validate the predictive power of the proposed LBD models through reuse research, with the goal of providing broader community benefits.

Problem Setting

Reuse research assists in creatively uncovering novel application areas for the proposed models (in contrast to the LBD models, which cater to one single problem), while also increasing their dependability (or reliability) (Ahmaro et al. 2014). Therefore, integrating *reusability* into the LBD workflow (which involves identifying new applications of LBD models using proper evaluations) will provide an extended platform to further verify the predictive performances of such models. Furthermore, ensuring reusability will facilitate the marking of new research directions in order to further improve existing LBD models as well as expanding the potential benefits of these models to the community. Contemplating the positive impact and numerous benefits of *reuse research* on the proposed LBD models (in contrast to LBD models specialised to a single problem), this thesis explores potential application areas in validating and comparing the predictive effects of the LBD models proposed as part of the main research objective 3.

1.3.5 Main Research Objective 5

To demonstrate the portability of the LBD workflow by proposing an interdisciplinary (or generalisable) LBD framework to assist scientific problem solving in a domain-agnostic manner.

Problem Setting

Even though LBD plays a critical role in speeding up innovation and research productivity regardless of the domain, most existing LBD research efforts suffer from a major research deficiency which is *lack of portability* of their LBD models. The main reason for this is that their LBD models depend on domain-specific knowledge resources, which hinders their applicability in other domains. More specifically, to date, LBD research has primarily been restricted to the *medical domain*, relying on semantic inferences that are made using *medicine-specific* knowledge resources (e.g., *UMLS*, *MeSH* and *SemRep*) (Henry & McInnes 2017, Sebastian et al. 2017*a*). The enormous growth of scientific literature (i.e., the *'data deluge'* (Khan et al. 2017)) has imposed challenges on researchers in almost every discipline; thus, those with stakes in LBD models can be found in almost every discipline. Therefore, the reliance on semantic inferences made using medicine-specific knowledge resources restricts the benefits that LBD could offer to the researchers outside the medical domain (Hui & Lau 2019).

Developing an *interdisciplinary (or generalisable)* LBD framework that could easily be applied to *general scientific problem solving* is important not only in order to equip a large and diverse community with the tools of LBD, but also to enhance LBD research outside the medical domain (where it is still in a nascent stage) (Hui & Lau 2019). To the best of our knowledge, no previous LBD studies have attempted to fulfil this research deficiency. Motivated by the broader opportunities that a portable LBD framework could offer to expand the existing constrained environments of LBD models, this thesis puts forward the first steps toward achieving *portability* in the LBD workflow, by proposing a highly cost-efficient and easily pluggable *interdisciplinary (or generalisable) LBD framework*. While enabling the *widespread applicability* of the LBD workflow, this proposed portable framework also alleviates one of the most often-cited challenges observed in non-medical LBD studies, which is the unavailability of a comprehensive knowledge base (Hui & Lau 2019).

1.4 New Contributions in the LBD Discipline

This section provides a high-level overview of new research contributions made through the thesis. More details on these new research contributions (along with the remaining major contributions) are outlined methodically at the end of each chapter and discussed in detail in Chapter 8.

- Integrating a systematic literature procedure into the LBD discipline to address the limitations of traditional narrative-based LBD reviews, while also shedding light on novel focus points in the field.

- Exploring the suitability of different input types in the LBD workflow by quantitatively assessing and comparing them by taking inspiration from the *subjective understanding of information* and *optimality theory*.

- Integrating a comprehensive temporal component into the LBD workflow to perform a nuanced analysis of semantically infused temporal signals.

- Introducing patterns based on relativity by integrating a *trajectory binding method*, taking inspiration from the molecular docking engine used in structured drug design.

- Proposing an interdisciplinary (or generalisable) LBD framework by circumventing existing domain-specific impediments to facilitate cross-domain and cross-lingual knowledge discovery with little or no cost.

- Integrating the vast range of knowledge encoded in *DBpedia* into the LBD workflow to build a robust platform from which to facilitate the formation of deep semantic inferences in a cross-domain and cross-lingual manner.

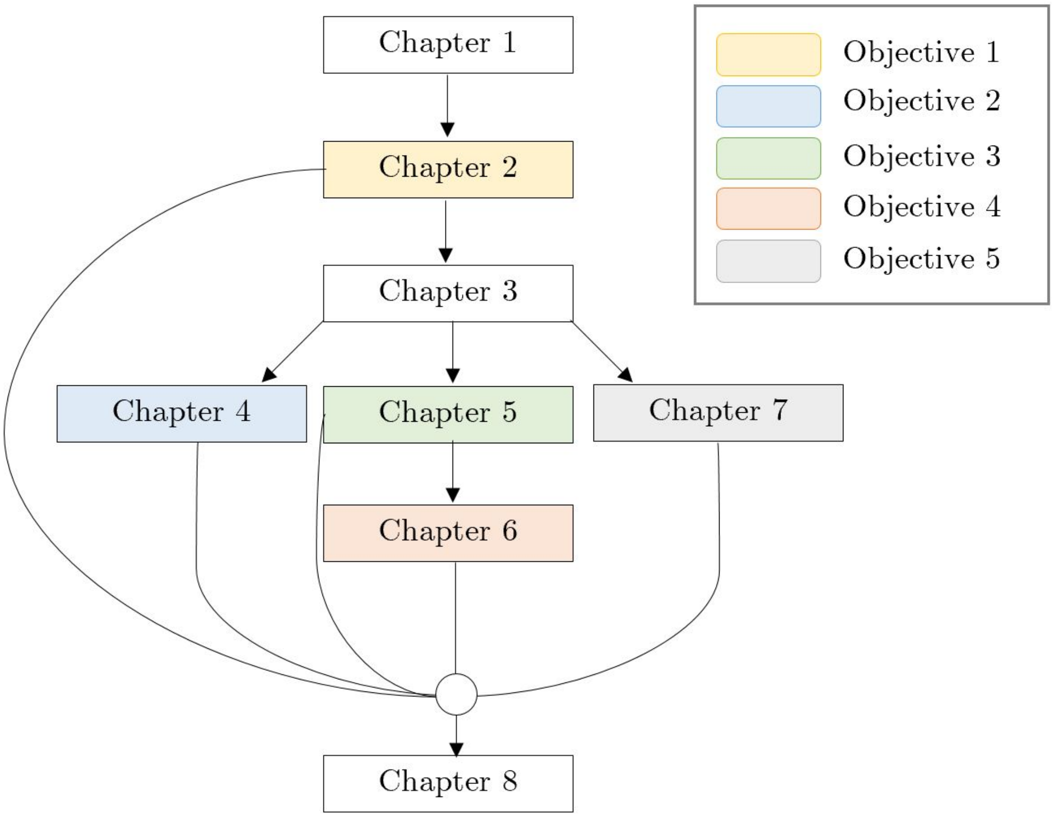


FIGURE 1.2: Dependency relations among chapters

1.5 Thesis Organisation

This section outlines the remaining *main chapters* of the thesis with a brief summary of their content. The dependency relations of these chapters are depicted in Figure 1.2 that could also be used as a guide to reading the thesis.

- *Chapter 2 (Systematic Literature Review):* Chapter 2 reviews the existing LBD literature by mainly considering on areas related to *general overview*, *methodology*, *statistical analysis*, and *components of the LBD workflow* (i.e., *input component*, *process component*, *output component* and *evaluation component*). This thesis adheres to a *systematic review protocol* to collect, appraise and synthesise the literature to answer clearly formulated research questions. The purpose of following this protocol is to establish a broad and comprehensive evidence base which can be used to form conclusions that serve as the main theoretical foundation for the remaining chapters of the thesis.

- *Chapter 3 (Research Design):* The intention of this chapter is to provide details on the underlying design considerations that will be utilised in the studies performed in the ensuing chapters. This chapter opens up by outlining the scope of this research and the components of the LBD workflow which correspond to the defined research scope. Subsequently, the experimental setups are discussed with a focus on the selected datasets and test cases. This is followed by a discussion on current challenges in LBD evaluation and how this thesis selected the most suited evaluation technique by outlining their advantages and disadvantages. The latter part of this chapter describes the theoretical foundation of the machine learning framework adopted in Chapters 5 and 6, while also discussing the evaluation metrics and baselines which were selected to facilitate performance comparisons.

- *Chapter 4 (Input Types):* This chapter is dedicated to establishing the first steps in investigating the *input component* of the LBD process, in order to understand the role of LBD input types and their contributions to the overall knowledge discovery process. More specifically, this chapter looks closely at the *information richness* of different LBD input types in the information retrieval cycle of the LBD workflow. The main aim of this analysis is to ascertain the suitability of different input types for the LBD framework, which will ideally serve as a guide towards developing better LBD models in the future. This analysis entails quantitatively measuring the information richness of different LBD input types using a *subjective understanding of information* (Tague-Sutcliffe 1992), while mapping the major ingredients of *optimal foraging theory* (Stephens & Krebs 1986) with the information retrieval cycle of the LBD workflow.

- *Chapter 5 (Semantic Evolution):* Chapter 5 concentrates on intermingling *modern word embedding techniques* with the *temporal dimension* to enhance the *discovery component* of the LBD workflow. This chapter discusses how the thesis disentangles multiple types of semantic shifts from *diachronic word embeddings*, in order to better understand the semantic evolution of scientific topics. More specifically, this chapter focuses on three broader categories of diachronic semantic inferences, namely *individual*, *pairwise* and *neighbourhood* to perform a circumstantial analysis of the *semantically infused temporal trajectories* of the scientific topics. The holistic integration of *vector semantics* with *temporally charged semantic deductions* substantiates the efficacy of the proposed LBD models as a means of discovering new knowledge linkages.

- *Chapter 6 (Reusability):* Chapter 6 provides a distinctive perspective to the proposed LBD models by validating their reusability in a timely reuse application area. Since this study follows a method similar to *opportunistic reuse* (i.e., gluing together pieces of components constructed for distinct problem setting(s) to create new capabilities), adaptations are made to the selected reuse setting using a methodical reuse plan. The experimental results of this reuse research corroborate the *vertical reuse* of the proposed LBD models, further verifying their robust predictive performances, as well as the positive influence of the complementary integration of *vector semantics* with the *temporal dimension*.

- *Chapter 7 (Portability):* The purpose of Chapter 7 is to describe the portability research performed as part of this thesis. To this end, this chapter describes how the thesis leverages the revolutionary opportunities offered through *Semantic Web* (more specifically, *Linked Open Data (LOD)*) to alleviate the domain-dependent impediments that are typical of the LBD workflow, which restrict the LBD models' applicability to limited problems or domains. Subsequently, this chapter investigates how well the proposed solutions meet the ultimate research objective of developing an interdisciplinary (or generalisable) LBD framework and the costs involved in the process of portability, in order to assess the cost-effectiveness of the proposed portable framework.

- *Chapter 8 (Conclusions and Future Work):* Chapter 8 concludes the thesis with a detailed reflection on the solutions proposed to overcome the identified research issues in the LBD discipline. More specifically, it restates the main research objectives of the thesis, provides a summary of studies performed and a detailed discussion on how these studies contribute to the field of LBD research. The latter section of this chapter describes the proposed research directions for the future considering each of the main objectives of the thesis. Lastly, the purpose of this thesis and how it contributes towards enhancing the existing understanding of the LBD workflow are discussed.

Chapter 2

Systematic Literature Review

2.1 Introduction

With the seemingly boundless growth of scientific literature, researchers struggle to deal with this amount of knowledge that ultimately has led to knowledge fragmentation (Liu & Rastegar-Mojarad 2016). Consequently, useful and interesting knowledge linkages among these fragmented knowledge isolations remain unnoticed (Pratt & Yetisgen-Yildiz 2003, Choudhury et al. 2020). Classical techniques, such as computer-aided literature searches or even recent advancements in text summarisation, may assist researchers to some extent by providing them with a high-level overview of the discipline. Nevertheless, such tools or techniques are not tailored towards capturing novel knowledge linkages between seemingly distinct knowledge fragments in the literature (Jha, Xun, Wang & Zhang 2019). *Literature-Based Discovery* (LBD) aims to elicit latent novel knowledge linkages in digital libraries by logically integrating complementary and non-interactive scientific literature. Discovering such meaningful novel knowledge linkages contributes to stimulating human creativity, which increases scientific productivity and research innovation (Jha et al. 2018, Xun et al. 2017).

The LBD research progressed through the groundbreaking studies of Swanson since 1986. These studies demonstrated the possibility of detecting undiscovered knowledge from the literature (Swanson 1986). In his first LBD study, Swanson discovered that *fish oil* might serve as a treatment for *Raynaud's disease*. This deduction was made by logically integrating the circulatory effects observed in the fish oil literature with the literature on Raynaud's disease (Swanson 1986). This implicit connection that Swanson identified through his unique bibliographic analysis was later supported by evidence from laboratory experiments (Kastrin & Hristovski 2020). Swanson labelled his initial finding *undiscovered public knowledge* - *public*, since every piece of knowledge required for his knowledge synthesis already existed publicly in the literature, and *undiscovered* because no researcher had previously brought these pieces together to form such a hypothesis (Bekhuis 2006, Garten et al. 2010). Later, Swanson further verified the importance of detecting such undiscovered knowledge through a series of other LBD discoveries (Swanson 1988, 1990*a*, Smalheiser & Swanson 1996, 1998, Swanson & Smalheiser 1996). Swanson's seminal discoveries demonstrate the potential for detecting undiscovered public knowledge that could provide valuable insights and lead to the formation of novel scientific hypotheses (Jha et al. 2018).

While Swanson's LBD discoveries form the groundwork in the discipline, the underlying knowledge synthesis processes that he followed to elicit these implicit novel knowledge linkages were both time and labour intensive (Jha, Xun, Wang & Zhang 2019). Therefore, different computational models were proposed in the LBD discipline to facilitate the knowledge discovery process in a more automated manner. While the initial computational methods in the LBD field were based purely on statistical techniques, with time, a wide spectrum of techniques was introduced to the field, facilitating the further automation of knowledge synthesis and making LBD knowledge discovery more efficient (Sebastian et al. 2017*a*, Henry & McInnes 2017). More specifically, LBD research focuses on developing novel knowledge discovery models that elicit such implicit linkages from the existing scientific knowledge in the literature (Xun et al. 2017).

While several *traditional narrative-based review papers* on LBD have been published (Gopalakrishnan et al. 2019, Henry & McInnes 2017, Sebastian et al. 2017*a*, Smalheiser 2017, Ahmed 2016, Smalheiser 2012, Kostoff et al. 2007, Bekhuis 2006, Ganiz et al. 2005, Weeber et al. 2005, Davies 1989), there are no published systematic literature reviews on LBD. Conducting such a systematic literature review is pivotal to the discipline, due to its ever-increasing growth of research contributions across 35 years of study. With this in mind, the current thesis performs a large-scale systematic literature review that circumvents the limitations of traditional narrative-based literature reviews such as *restrictive scope* and *limited focus points*. Systematic reviews employ a rigorous, transparent, well-defined as well as reproducible approach to synthesise the literature in a manner designed

Table 2.1: Procedural differences between systematic reviews and traditional reviews

Component	**Systematic Review**	**Traditional Review**
Protocol	Includes an explicit and detailed review protocol	No protocol
Focus	Clear objectives are identified; uses focused research questions	Covers several aspects of the topics, including context and current thinking, often with no specific research questions
Inclusion/ exclusion criteria	Inclusion and exclusion criteria are identified prior to conducting the review	No criteria specified
Search strategy	Comprehensive, reproducible and systematic search is conducted using several specified databases with precise search terms. There is an attempt to identify all relevant publications on the topic	Search strategy is not mentioned; papers are found using a random process. Usually involves few literature databases
Process of selecting articles	Clear and explicit selection process is performed using explicit inclusion and exclusion criteria	No details on the selection process
Results and data synthesis	Clear conclusions based on high-quality evidence. The findings of the review are unbiased, balanced and reproducible	May be influenced by the reviewer's needs, beliefs and theories

to minimise bias (Snyder 2019, Kitchenham et al. 2009). The following key principles of systematic literature reviews can be considered their main strengths: *transparency*, *clarity*, *equality*, *accessibility*, *impartial inclusive coverage*, *replicability*, *objectivity*, *scientific rigour*, *focus* and *unity*. Such attributes are lacking in the traditional reviews (Frangieh & Yaacoub 2017, Boell & Cecez-Kecmanovic 2015, Pittaway & Cope 2007). Table 2.1 outlines the procedural differences between systematic literature reviews and traditional narrative-based reviews (Keele 2007, Cook et al. 1997, Egger & Smith 1997, Khan et al. 2003, Snyder 2019).

2.2 Research Questions of the Systematic Literature Review

The research questions designed as part of the systematic literature procedure (i.e., the *focus* component in Table 2.1) are also compatible with the main research objectives of this thesis, as mentioned below.

- **Main Research Objective 2:** Since main objective 2 of this thesis (discussed in Section 1.3.2) is related to LBD input types the following two research questions were defined to better understand the *input component* of the LBD workflow.

    1. What *input types* are used in the knowledge discovery process of the LBD workflow?
    2. What *data sources* are used in LBD research to extract these identified input types?

- **Main Research Objective 3:** Since main objective 3 of this thesis (discussed in Section 1.3.3) is related to the knowledge discovery process of the LBD workflow, the following nine research questions were defined in order to identify potential directions in enhancing the current understanding of the *discovery component*.

    3. What *computational techniques* are used in LBD research?
    4. What *topics/central themes* emerged over time in the LBD discipline?
    5. What *filtering techniques* are used in the LBD process?
    6. What *ranking/thresholding mechanisms* are used in the LBD process?
    7. What is the *evidence* that LBD generates discovery?
    8. What are the *LBD evaluation types* and how suitable are they to non-medical domains?
    9. What *quantitative measurements* are used to assess the effectiveness of the results?
    10. What *visualisation techniques* are used to display results in LBD research?
    11. What are the *trends* in LBD research in terms of publications over the years, top-cited papers and top authors?

- **Main Research Objective 4:** Since main objective 4 of this thesis (discussed in Section 1.3.4) is related to reusability in the LBD context, the following research question was designed to better understand *potential application areas* for LBD models.

    12. What are the applications of LBD research?

- **Main Research Objective 5:** Since main objective 5 of this thesis (discussed in Section 1.3.5) is related to portability of the LBD workflow, the following three research questions were designed to better understand the *potential reasons* that restrict the ability of LBD models to serve non-medical domains.

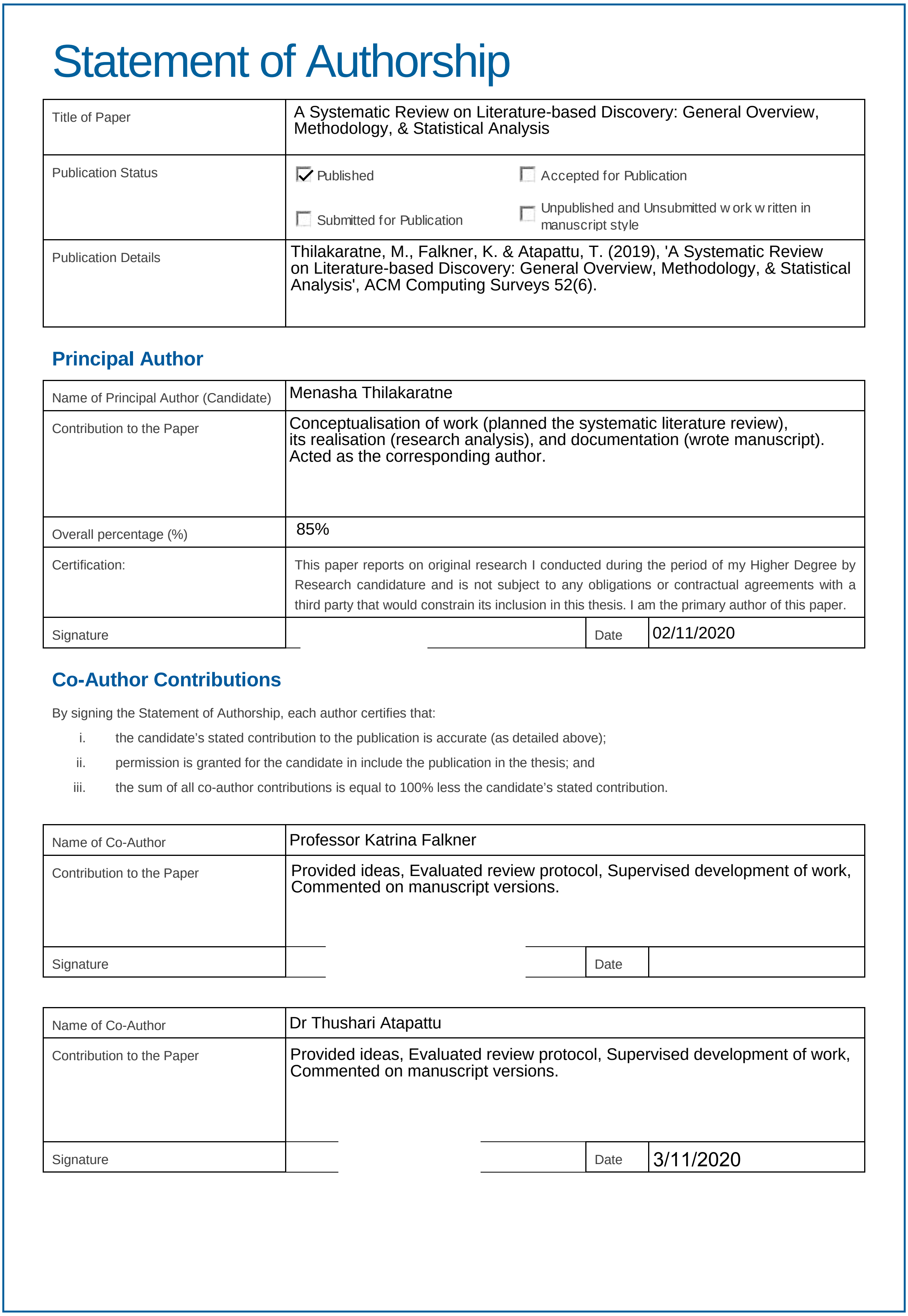
  13. What domains are considered in LBD research, and what are the levels of generalisability for these domains?
  14. What domain-independent and domain-dependent resources are utilised in LBD research?
  15. What are the main LBD tools available, and what are their supported domains?

2.3 Findings of the Systematic Literature Review

Since this thesis follows the *hybrid publications-narrative format*, the findings of the systematic literature review are presented methodically in the following two publications, which are enclosed in this chapter.

- **Publication I:**
  Title: A Systematic Review on Literature-based Discovery: General Overview, Methodology, & Statistical Analysis
  Authors: Menasha Thilakaratne, Katrina Falkner, Thushari Atapattu
  Venue: ACM Computing Surveys 2019 (CORE Rank: A*, Impact Factor: 6.131)

- **Publication II:**
  Title: A Systematic Review on Literature-based Discovery Workflow
  Authors: Menasha Thilakaratne, Katrina Falkner, Thushari Atapattu
  Venue: PeerJ-CS 2019 (Impact Factor: 3.09)



Statement of Authorship

Title of Paper	A Systematic Review on Literature-based Discovery: General Overview, Methodology, & Statistical Analysis
Publication Status	☑ Published □ Accepted for Publication □ Submitted for Publication □ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Thilakaratne, M., Falkner, K. & Atapattu, T. (2019), 'A Systematic Review on Literature-based Discovery: General Overview, Methodology, & Statistical Analysis', ACM Computing Surveys 52(6).

Principal Author

Name of Principal Author (Candidate)	Menasha Thilakaratne		
Contribution to the Paper	Conceptualisation of work (planned the systematic literature review), its realisation (research analysis), and documentation (wrote manuscript). Acted as the corresponding author.		
Overall percentage (%)	85%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	02/11/2020

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Professor Katrina Falkner		
Contribution to the Paper	Provided ideas, Evaluated review protocol, Supervised development of work, Commented on manuscript versions.		
Signature		Date	

Name of Co-Author	Dr Thushari Atapattu		
Contribution to the Paper	Provided ideas, Evaluated review protocol, Supervised development of work, Commented on manuscript versions.		
Signature		Date	3/11/2020

2.4 Publication I

A Systematic Review on Literature-Based Discovery: General Overview, Methodology, & Statistical Analysis

The vast nature of scientific publications brings out the importance of *Literature-Based Discovery (LBD)* research that is highly beneficial to accelerate knowledge acquisition and the research development process. LBD is a knowledge discovery workflow that automatically detects significant, implicit knowledge associations hidden in fragmented knowledge areas by analysing the existing scientific literature. Therefore, the LBD output not only assists in formulating scientifically sensible novel research hypotheses, but also encourages the development of cross-disciplinary research. In this systematic review, we provide an in-depth analysis of the computational techniques used in the LBD process using a novel, up-to-date and detailed classification. Moreover, we also summarise the key milestones of the discipline through a timeline of topics. To provide a general overview of the discipline, the review outlines LBD validation checks, major LBD tools, application areas, domains and generalisability of LBD methodologies. We also outline the insights gathered through our statistical analysis that capture the trends in the LBD literature. To conclude, we discuss the prevailing research deficiencies in the discipline by highlighting the challenges and opportunities for future LBD research.

i Introduction

Formulation of scientifically sensible novel research hypotheses requires a comprehensive analysis of the existing domain-specific knowledge presented in the literature. However, the massive influx of research publications (Cheadle et al. 2017) makes the hypotheses generation process extremely difficult and time-consuming even in the narrow specialisation of a scientist. Developing a tool that assists in eliciting novel knowledge linkages can significantly reduce the time and the effort the scientists must put in to manually articulating and validating research hypotheses, which will ultimately accelerate scientific productivity and research innovation. In this regard, Literature-Based Discovery (LBD) research is highly beneficial as it aims to detect non-trivial implicit associations in the literature that have the potential to generate novel research hypotheses (Swanson 2001, Ganiz et al. 2005, Su & Zhou 2009). A simple definition by Hristovski et al. (2015*b*) is; *"Literature-Based Discovery (LBD) generates discoveries, or hypotheses, by combining what is already known in the literature"*.

Swanson (the pioneer of the LBD discipline) demonstrated the importance of detecting such non-apparent associations between *disjointed* knowledge fragments by manually discovering the role of *fish oil* in preventing *Raynaud's disease* (Swanson 1986). He followed a simple procedure, namely the *ABC model* to make this discovery. The ABC model is built on the assumption that 'if concept A is associated with a concept B and that concept B is associated with another concept C, then concept A is associated with concept C, where the B-concept denotes the association/relationship between the two concepts A and C'. Thus, concept A can be treated as the starting concept/term, B concept(s) as the intermediate association(s), and concept C as the target concept/term. Later on, Swanson followed the same process in unrevealing the hidden associations between *Migraine*↔*Magnesium* literature (Swanson 1988). Subsequently, his observations were proven from laboratory experiments that demonstrate the validity of his thinking process (Ramadan et al. 1989). These two discoveries of Swanson formed the groundwork of the LBD discipline. Even though the ABC model is simple, it is still widely used as a discovery framework of the existing LBD studies (Sebastian et al. 2017*a*).

This model has two variants termed *open discovery* and *closed discovery*. In *open discovery*, the user requires to specify a topic of interest (concept A), and the LBD process identifies the B and C concepts, respectively by exploring the scientific literature. On the contrary, *closed discovery* requires the user to input a pair of topics (concepts A and C) and the LBD process detects the implicit relationships between these two concepts

(B concepts). Later, various other discovery frameworks were introduced into the field, such as the *AnC model* (Wilkowski et al. 2011*b*), *heterogeneous bibliographic information network* (Sebastian et al. 2017*b*, 2015), *network structures* (Ding et al. 2013), *outlier detection* (Petrič et al. 2012), and *analogical reasoning* (Mower et al. 2016) to elicit more complex associations that the ABC model fails to detect (Smalheiser 2017, 2012).

ii Purpose of the Review

Although there have been several literature reviews published in the LBD discipline over time (Henry & McInnes 2017, Sebastian et al. 2017*a*, Smalheiser 2017, Ahmed 2016, Smalheiser 2012, Kostoff et al. 2007, Bekhuis 2006, Ganiz et al. 2005, Weeber et al. 2005, Davies 1989), the field is still lacking a *systematic literature review*. Different from the traditional literature reviews, systematic reviews follow a rigorous, transparent, explicit, and reproducible methodology with a predefined review protocol to minimise bias in the results. This enables systematic reviews to provide more reliable findings and conclusions in the discipline (Higgins & Green 2008). With the intention of filling this gap, we present a large-scale systematic review that critique the research progress in the LBD discipline in a wide scope. In a nutshell, our major contributions are; 1) being the first systematic literature review in the LBD discipline, 2) providing novel, up-to-date and comprehensive classifications to answer our research questions, and 3) reviewing independently from the domain without only limiting to the medical LBD studies.

iii Research Questions

This review attempts to answer the below-mentioned seven research questions that are categorised into *methodology*, *general overview*, and *statistical analysis*.

1. **LBD Methodology**

   What computational techniques are used in LBD research?

   What topics/central themes emerged over time in the LBD discipline?

2. **General Overview**

   What is the evidence that LBD generates discovery?

   What are the main LBD tools available, and what are their supported domains?

   What are the applications of LBD research?

   What domains are considered in LBD research, and what are the levels of generalisability for these domains?

3. **Statistics Analysis**

   What are the trends in LBD research in terms of publications over the years, top-cited papers and top authors?

iv Methods

This review follows the typical workflow of systematic literature reviews in computer science to retrieve and select articles for analysis (Weidt & Silva 2016).

Table 2.2: Statistics of the article retrieval process

Keyword	**Web of Science**	**Scopus**	**PubMed**	**ACM Digital Library**	**IEEE Xplore**	**Springer -Link**	**Total Count**
Query 1[a]	161	68	75	15	15	8	342
Query 2[b]	14	0	4	1	2	1	22
Query 3[c]	14	0	0	0	0	1	15
References from Henry & McInnes (2017)							96
Total Article count							**475**

[a] "literature based discovery" OR "literature based discoveries"
[b] "literature based knowledge discovery" OR "literature based knowledge discoveries"
[c] "literature related discovery" OR "literature related discoveries"

iv.1 Article Retrieval Process

We used six keywords and six literature databases to retrieve articles, as summarised in Table 2.2. The search was performed using title, abstract or keywords depending on the search options given by the databases. To minimise the risk of losing important articles that are outside the keywords and the databases used, we also obtained the references list from the latest LBD review (Henry & McInnes 2017).

iv.2 Article Selection Process

The article types that we considered for the review are only journals and conference proceedings. We excluded articles that are reviews, book chapters, books, editorials, keynotes, and lesson learned reports. The language of the articles considered is English. Our article selection process is comprised of three stages (Weidt & Silva 2016); *Stage 1:* analyse only title and abstract, *Stage 2:* analyse introduction and conclusion, and *Stage 3:* read complete article and quality checklist. We did not include articles that are less than or equal to 4 pages in our analysis as they mainly reflect work-in-progress. However, we included such papers only to answer *RQ5*, as such papers tend to propose novel application areas in LBD. Our article selection process resulted in 176 papers, and for RQ5, we used additional 18 papers. The complete list of articles is available at `https://tinyurl.com/selected-list`.

v LBD Methodology

v.1 What computational techniques are used in LBD research?

Even though the early work in LBD was mostly performed manually (Swanson 1986, 1988), over time, different computational techniques were adopted to automate the knowledge discovery process. In this review, we provide a detailed classification of the existing LBD techniques, as illustrated in Figure 2.1.

v.1.1 Statistical/Probabilistic/Co-occurrence Models

This section reviews the LBD methodologies that rely on statistical measures to determine the frequencies/likelihood or co-occurrence patterns of the relationships between terms. The main disadvantage of solely depending on the techniques in this category is that they do not consider the semantic aspects of the terms in the knowledge discovery process. However, *distributional semantic models* deviate from the remaining techniques

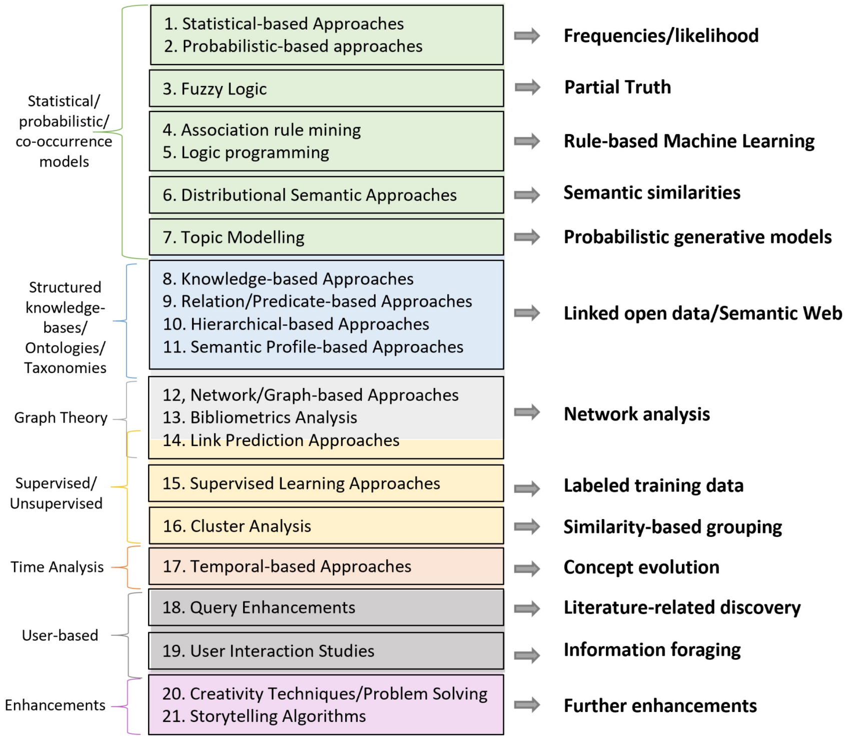

Statistical/ probabilistic/ co-occurrence models
1. Statistical-based Approaches
2. Probabilistic-based approaches
Frequencies/likelihood
3. Fuzzy Logic
Partial Truth
4. Association rule mining
5. Logic programming
Rule-based Machine Learning
6. Distributional Semantic Approaches
Semantic similarities
7. Topic Modelling
Probabilistic generative models
Structured knowledge-bases/ Ontologies/ Taxonomies
8. Knowledge-based Approaches
9. Relation/Predicate-based Approaches
10. Hierarchical-based Approaches
11. Semantic Profile-based Approaches
Linked open data/Semantic Web
Graph Theory
12, Network/Graph-based Approaches
13. Bibliometrics Analysis
14. Link Prediction Approaches
Network analysis
Supervised/ Unsupervised
15. Supervised Learning Approaches
Labeled training data
16. Cluster Analysis
Similarity-based grouping
Time Analysis
17. Temporal-based Approaches
Concept evolution
User-based
18. Query Enhancements
Literature-related discovery
19. User Interaction Studies
Information foraging
Enhancements
20. Creativity Techniques/Problem Solving
21. Storytelling Algorithms
Further enhancements

Figure 2.1: Main computational techniques in LBD

as they also capture the context of the terms (patterns of their positions in the content) to construct the vector space by adhering to the *distributional hypothesis*.

Statistical-based Approaches: Statistical approaches often rely on frequencies of concepts and their statistical distributions to discover implicit knowledge associations between disjointed literature sets (Swanson & Smalheiser 1997, Lindsay & Gordon 1999, Gordon & Lindsay 1996, Petric et al. 2014, Petrič et al. 2009, Spinak et al. 1999, Workman et al. 2016, Ittipanuvat et al. 2012, Gordon et al. 2002, Kibwami & Tutesigensi 2014, Yao et al. 2008). Early studies in the LBD discipline mostly relied on statistical measures, which can be considered as the most primitive technique used in the literature.

For instance, Swanson & Smalheiser (1997) initiated the automation of the LBD process by following a simple frequency-based metric of word occurrences to obtain the target concepts. Subsequently, Gordon & Lindsay (1999, 1996) further extended this work by using scores such as token frequency, record frequency, term frequency-inverse document frequency (TF-IDF) and relative frequency. However, these statistical approaches tend to pick terms that frequently co-occur. Thus, they fail to identify important associations that are formed using less frequent words. As a result, Petric et al. (2014, 2009, 2007) exploited the notion of rarity in their LBD process. That is, if a concept rarely appears in a given set of literature, they believe that it is less researched in the given field. Thus, they argue that exploring these concepts may lead to innovative research pathways. Similarly, *Spark* LBD system (Workman et al. 2016) also exploits the use of rarity as a signal to provide new knowledge to the users.

Even though these statistical techniques are easy to compute, they require high intervention of human experts because their success vastly depends on prior knowledge. For example, Lindsay & Gordon (1999) manually removed highly ranked B-concepts during their discovery process to reach the target C-concept, which shows the bias and the requisite of the prior knowledge. Moreover, the success of these statistical approaches is highly limited as they do not consider the semantic meaning of terms.

Probabilistic Approaches: Some LBD approaches (Vidal et al. 2014, Seki & Mostafa 2009, 2007) have utilised probabilistic techniques to detect potential knowledge associations between disjointed literature sets. For instance, Vidal et al. (2014) proposed an authority-flow based ranking mechanism by modelling a Bayesian network using two sampling techniques; *direct sampling reasoning algorithm* and *conditional probability*. Similarly, Seki & Mostafa (2009, 2007) also utilised an inference network (Turtle & Croft 1991) to predict novel *gene↔disease* associations based on probabilities.

Fuzzy Logic: Fuzzy logic (Steimann 1997) is a foam of *multivalued logic* that computes *degrees of truth* (which ranges from 0 to 1) by handling the concept of *partial truth*. Therefore, fuzzy logic is different to *Boolean logic*, which is only based on two-valued logic, *true or false (1 or 0)*. Wren et al. (2004) argue that the term co-occurrences do not necessarily indicate meaningful relationships between the terms. Therefore, they have exploited the use of fuzzy logic to weight the importance of the term co-occurrences by assigning a fuzzy score to model relationships in their LBD process.

Association Rule Mining: Association Rule Mining (ARM) helps to uncover associations between data objects by observing frequent patterns/behaviours, and correlations among objects. Although ARM and co-occurrence analysis are similar, ARM can detect tri-occurrences, quad-occurrences that can be utilised to identify the correlations between terms (Ganiz et al. 2005). An association rule can be denoted as the expression A→B, where A and B are set of objects. Every association rule must satisfy the user-defined two constraints, namely *support* and *confidence*. 'Support' measures the count of articles in which both starting and linking terms co-occur, whereas 'confidence' measures the fraction of articles that contain the linking term, given that the starting concept occurs in the document (Yetisgen-Yildiz & Pratt 2009).

The typical procedure involving ARM in the LBD process is; 1) For a given starting concept A, find all linking terms B such that A→B, 2) Find all target concepts C such that B→C, 3) Remove those C concepts for which A→C already exists, and 4) The remaining C concepts are the candidates of novel associations (between A and C) (Hu et al. 2010, Hristovski et al. 2005, Berardi et al. 2005, Huang & Nakamori 2004, Hristovski et al. 2003, Jha & Jin 2016*b*, Hristovski et al. 2001, Pratt & Yetisgen-Yildiz 2003, Yetisgen-Yildiz 2006). Generally, the produced candidate list is extensive, which requires some mechanism to handle this combinatorial problem. Hristovski et al. (2001) proposed the use of UMLS semantic types to limit the results. For example, if the starting concept belongs to a semantic type 'disease', the user can select 'pathologic function' and 'pharmacologic substance' to be the semantic types of B and C, respectively. Similarly, Hu et al. (2010) also utilised a *semantic-based* association rule system by using semantic type filters. Berardi et al. (2005) proposed the use of *generalised* association rules by exploring the hierarchy of MeSH taxonomy. Generalised association rules (Srikant & Agrawal 1995) signify association rules A→B, where no term in B is an ancestor of any term in A. With regards to the ARM algorithms, LBD literature has commonly used Apriori algorithm in the knowledge discovery process (Cherdioui & Boubekeur 2013, Hu et al. 2010, Pratt & Yetisgen-Yildiz 2003, Yetisgen-Yildiz 2006). A comparison by Yetisgen-Yildiz & Pratt (2009) has revealed that association rules outperformed statistical measures such as TF-IDF, Mutual Information Measure (MIM) and z-score.

Logic Programming: Thaicharoen et al. (2009) proposed *Inductive Logic Programming (ILP)* to detect meaningful associations in forms of relational frequent patterns. The expressive data representations of ILP and its ability to integrate background knowledge are the main benefits of this technique. There are several popular ILP algorithms, such as FOIL, WARMR and PROGOL. For instance, Thaichareon et al. (2009) used WSRMR (which is an extension of the Apriori algorithm) in their LBD process.

Distributional Semantic Approaches: Distributional semantic models involve constructing semantic representations of terms in the form of dense vectors by analysing their statistical distribution across documents, syntactic dependency relations, collocational profiles, and other contextual features. These models are based on the assumption that two words in a similar context are semantically related (a.k.a. *distributional hypothesis*). As a result, semantically related terms tend to have similar vector representations in the vector space.

Various distributional semantic techniques have been proposed in the LBD literature such as *Latent Semantic Indexing (LSI)/Latent Semantic Analysis (LSA)* (Kostoff, Solka, Rushenberg & Wyatt 2008, Gordon & Dumais 1998), *Reflective Random Indexing (RRI)* (Shang et al. 2014, Cohen et al. 2012, 2011, Cohen, Schvaneveldt & Widdows 2010, Mower et al. 2016, Malec et al. 2016, Cohen, Widdows, Stephan, Zinner, Kim, Rindflesch & Davies 2014, Cohen, Whitfield, Schvaneveldt, Mukund & Rindflesch 2010), *Predication-based Semantic Indexing (PSI)* (Shang et al. 2014, Cohen et al. 2012, 2011, Mower et al. 2016, Malec et al. 2016, Cohen, Widdows, Stephan, Zinner, Kim, Rindflesch & Davies 2014, Cohen, Whitfield, Schvaneveldt, Mukund & Rindflesch 2010, Cohen et al. 2009, Cohen, Widdows & Rindflesch 2014, Cohen, Widdows, Schvaneveldt & Rindflesch 2010), *Associative Concept Space (ACS)* (Van der Eijk et al. 2004, 2002), *Semantic Vectors package* (McClure 2012), *Tensor Encoding (TE)* (Symonds, Bruza & Sitbon 2014), *Symmetric Random Indexing (SRI)* (Cohen & Schvaneveldt 2010), *Hyperspace Analogue to Language (HAL)* (Bruza et al. 2006, Cole & Bruza 2005, Bruza et al. 2004), *Word embeddings* (Xun et al. 2017) and *Graph embeddings* (Gopalakrishnan et al. 2017).

Typically, *nearest neighbour analysis* or *vector operations in the semantic space* is performed to identify the implicit and novel associations using distributional models. For instance, the work of Gordon & Dumais (1998) followed a nearest neighbour search to detect the potential target concepts. More specifically, they have employed LSI to identify semantically similar neighbouring terms of the A-concept by calculating the cosine similarity to derive the target C-concepts. They have reported that LSI analysis provided slightly better results than the traditional frequency-based statistical metrics (Gordon & Lindsay 1996).

Cohen et al. (2011) proposed a vector operations-based distributional semantic model that uses the PSI technique based on Kanerva's Binary Spatter Code. The PSI space was built using *SemRep* predications that are encoded into a high-dimensional vector space. Afterwards, the predication space was searched using a process similar to Kanerva's XOR-based analogical mapping to facilitate analogical retrieval that in the form of *"A is to B as C is to ?" (e.g., "prozac is to depression as what is to schizophrenia?").*

Topic Modelling: How topic-level information is propagated among documents can be observed using topic modelling algorithms, instead of performing a term-level analysis. This approach can also be viewed as a topic-based profiling technique (see Section v.1.2). However, the effects of such algorithms are rarely experimented in LBD research (Sebastian et al. 2017*b*). Few studies have involved topic modelling in the knowledge discovery process using Latent Dirichlet Allocation (LDA) algorithm (Sebastian et al. 2017*b*, Qi & Ohsawa 2016, Bisgin et al. 2011).

v.1.2 Structured Knowledge bases/Ontologies/Taxonomies

Semantic augmentation (a.k.a. *semantic annotation* or *semantic tagging*) is the process of attaching semantics to terms in texts to assist automatic interpretation of their meaning. In this section, we summarise how LBD research have used structured data to facilitate *semantic augmentation* with the intention of enhancing the reasoning and inferencing ability of the knowledge discovery process. However, this also opens up questions such as *word sense disambiguation/entity resolution* (Preiss & Stevenson 2016).

Knowledge-based Approaches: The involvement of knowledge-based techniques has become an integral component of the LBD process (Lever et al. 2018, Vlietstra et al. 2017, Preiss & Stevenson 2017, Huang et al. 2016, Preiss & Stevenson 2016, Zhou et al. 2015, Song et al. 2015, Preiss et al. 2015, Cairelli et al. 2015, Cameron et al. 2015, Rastegar-Mojarad et al. 2015, Srinivasan et al. 2015, Shang et al. 2014, Hanauer et al. 2014, Dong et al. 2014, Tsafnat et al. 2014, Kastrin et al. 2014*b*, Vidal et al. 2014, Petric et al. 2014, Ding et al. 2013, Liang et al. 2013, Cameron et al. 2013, Gabetta et al. 2013, Cherdioui & Boubekeur 2013, Cohen et al. 2012, Miller et al. 2012, Bhattacharya & Srinivasan 2012, Goodwin et al. 2012, Faro et al. 2011, Guo & Kraines 2009*b*, Maclean & Seltzer 2011, Loglisci & Ceci 2011, Hur et al. 2010, Baker & Hemminger 2010, Ijaz et al. 2009, Hu et al. 2010, Hristovski et al. 2010, Petrič et al. 2009, Vidal et al. 2010, Yetisgen-Yildiz & Pratt 2009, Kostoff & Briggs 2008, Kostoff, Briggs & Lyons 2008, Yetisgen-Yildiz & Pratt 2006, Swanson et al. 2006, Hu et al. 2006, Hristovski et al. 2005, Hu et al. 2005, Berardi et al. 2005, Huang et al. 2005*b*, Srinivasan 2004, Van der Eijk et al. 2004, Huang et al. 2005*a*, Huang & Nakamori 2004, Stegmann & Grohmann 2003, Weeber et al. 2001, 2000, Park et al. 2017, Jha & Jin 2016*b*, Rastegar-Mojarad et al. 2016, Mower et al. 2016, Gulec et al. 2010, Sang, Yang, Wang, Liu, Lin & Wang 2018, Gopalakrishnan et al. 2017, Peng et al. 2017, Malec et al. 2016, Cohen, Widdows, Stephan, Zinner, Kim, Rindflesch & Davies 2014, Cairelli et al. 2013, Wilkowski et al. 2011*b*, Özgür et al. 2011, Cohen, Whitfield, Schvaneveldt, Mukund & Rindflesch 2010, Hristovski et al. 2006, 2003, Qian et al. 2012, Huang et al. 2012, Jelier et al. 2008, Srinivasan & Libbus 2004, Zhang et al. 2014, Preiss 2014, Cohen et al. 2009, Cohen, Widdows & Rindflesch 2014, Cohen, Widdows, Schvaneveldt & Rindflesch 2010, Workman et al. 2016, Pratt & Yetisgen-Yildiz 2003, Yetisgen-Yildiz 2006, Wren 2004, Frijters et al. 2010). These approaches utilise external structured knowledge-based resources to acquire domain-specific background knowledge.

To date, LBD literature has only focused on knowledge-based resources that are in the medical domain to gain additional knowledge. The most common practice of the existing literature is to utilise *Unified Medical Language System (UMLS)* (Bodenreider 2004) with the help of tools, such as *MetaMap* for concept detection. The advantage of MetaMap tool (Aronson 2001) is that it automatically identifies the medical concepts in a text, and maps them to UMLS medical entities. Using such concept-based controlled vocabularies greatly assist in detecting words that are biologically relevant. Moreover, this also enables to explore additional information of concepts, such as semantic types, hierarchical relations, synonyms etc. For instance, Weeber et al. (2001) proposed a semantic type filtering approach using MetaMap tool to detect medical concepts and to filter these identified concepts by user-specified semantic types.

As knowledge-based approaches explore the semantics of concepts, they tend to produce more meaningful knowledge associations. Other types of approaches that are evolved from knowledge-based approaches are *relation-based* and *hierarchical-based* approaches.

Relation/Predicate-based Approaches: Relation based approaches use explicit relations between concepts by analysing *subject-predicate-object* triples (*semantic predications*) to detect the meaning of knowledge associations (Vlietstra et al. 2017, Yang et al. 2017, Preiss & Stevenson 2017, Kim et al. 2016, Song et al. 2015, Preiss et al. 2015, Cairelli et al. 2015, Cameron et al. 2015, Rastegar-Mojarad et al. 2015, Shang et al. 2014, Vicente-Gomila 2014, Marsi et al. 2014, Cohen et al. 2012, Miller et al. 2012, Bhattacharya & Srinivasan 2012, Goodwin et al. 2012, Cohen et al. 2011, Guo & Kraines 2009*b*, Kraines et al. 2010, Hristovski et al. 2010, Guo & Kraines 2009*a*, Hu et al. 2005, Sang, Yang, Wang, Liu, Lin & Wang 2018, Malec et al. 2016, Cohen, Widdows, Stephan, Zinner, Kim, Rindflesch & Davies 2014, Cairelli et al. 2013, Wilkowski et al. 2011*b*, Cohen, Whitfield, Schvaneveldt, Mukund & Rindflesch 2010, Ahlers et al. 2007, Hristovski et al. 2006, Kraines et al. 2013, Zhang et al. 2014, Preiss 2014, Hristovski, Kastrin, Dinevski & Rindflesch 2015*a*, Cohen et al. 2009, Cohen, Widdows & Rindflesch 2014, Cohen, Widdows, Schvaneveldt & Rindflesch 2010, Workman et al. 2016). This will not only filter out the meaningless connections, but also enable the user to clearly understand the derived implicit associations. The most commonly used semantic interpreter to extract these semantic predications from the biomedical text is *SemRep* (Rindflesch & Fiszman 2003).

An example of this technique is the work proposed by Hristovski et al. (2006) that exploited a discovery pattern-based technique. More specifically, they introduced two forms of discovery patterns named *Maybe_Treats1* and *Maybe_Treats2* to propose novel treatments for a given disease. For a given disease (concept A), the first pattern identifies any change in body function, substance or body measurement (concept B), and proposes treatments C, which are associated with the opposite change of concept B. For a starting disease A, the second pattern analyses the diseases B with similar characteristics and suggests their treatments as potential C concepts.

Another useful resource available for relations-based approaches is *Semantic MEDLINE*, a web-based tool that visualises SemRep-generated semantic predications of MEDLINE stored in SemMedDB database. Some research studies have directly utilised Semantic MEDLINE in their LBD process (Cairelli et al. 2013, Wilkowski et al. 2011*b*). The main advantage of the relation-based approaches is the better interpretation of associations that assists in detecting more accurate results. However, this technique is restricted to problems where such explicit associations between concepts are known in advance or in the domains where such resources (e.g., SemRep) are available.

Hierarchical-based Approaches: Hierarchical-based approaches exploit the hierarchical structure of concepts in a given knowledge base/taxonomy to gain additional knowledge (Berardi et al. 2005, Huang et al. 2005*b*,*a*, Pratt & Yetisgen-Yildiz 2003, Gulec et al. 2010). This could be done by analysing details such as; 1) position/level of a concept in the hierarchy, and 2) relationships between ancestor-descendant and siblings.

The study of Pratt & Yetisgen-Yildiz (2003) can be considered as an example for the first category that analyses the position/level of concepts in the hierarchy. They observed that concepts which reside on second and third levels of UMLS hierarchy tend to be general (e.g., *drug*, *disease*). Thus, they utilised the position details of concepts as a filtering mechanism to remove the general terms.

An example that denotes the second category is the study of Huang et al. (2005*a*) which analysed sibling relationships to eliminate meaningless candidate associations. More specifically, for a starting concept A, they have identified B concepts that co-occur with A. Afterwards, the siblings of B concepts are extracted as C-concepts. Then, the already known $A \rightarrow C$ connections are removed to obtain the novel association list. The authors state that since $A \rightarrow B$ is reported as a valid association in the literature, and B and C tend to be similar due to the sibling relationship, there is a high chance of establishing a connection between A and C. However, these hierarchical-based techniques are limited to problems/domains where such hierarchical taxonomies are available.

Semantic Profile-based Approaches: In the *Manjal* project (Srinivasan & Libbus 2004), Srinivasan (2004) proposed a semantic profile-based approach for the first time in the LBD discipline. In her methodology, a profile is denoted by vectors of weighted MeSH terms, which are each assigned to one of the 124 UMLS semantic types. She has used the TF-IDF metric as the weighting mechanism. These generated MeSH-based profiles are used in both open and closed discovery setting to identify the potential associations (Srinivasan 2004). Similarly, the system *Anni* also leverages a profile-based technique using biomedical concept profiles (Jelier et al. 2008). The system constructs profiles using related biomedical concepts that are weighted using symmetric uncertainty coefficient to denote their importance within a profile. Moreover, Cheung et al. (2012*a*) utilised a MeSH-based weighted profile, namely *Medical Subject Heading Over-representation Profiles (MeSHOPs)* in the LBD process, where Fisher's Exact Test was used to determine the over-represented MeSH terms in profiles (Cheung et al. 2012*b*).

v.1.3 Graph Theory

In this section, we present how graph theory is integrated into the LBD framework. Different types of graphs have been analysed in the discipline representing both *directed* (e.g., *Entitymetrics (Ding et al. 2013)*, *heterogeneous bibliographic information network* (Sebastian et al. 2017*b*, 2015)) and *undirected* (e.g., *co-occurrence networks* (Kastrin et al. 2016)) graphs. These constructed graphs are typically analysed using one or more of the following three levels; macro-level (i.e., global graph metrics such as *degree distribution*, and *shortest distance*), meso-level (i.e., cluster characters such as *clustering coefficient*, and *modularity-based clustering*), and micro-level (i.e., node properties such as *centrality measures*).

Network/Graph-based Approaches: Graph-based approaches use graph properties and theories to identify the novel associations between concepts (Baek et al. 2017, Vlietstra et al. 2017, Pusala et al. 2017, Sebastian et al. 2017*b*, Jha & Jin 2016*a*, Kastrin et al. 2016, Kim et al. 2016, Song et al. 2015, Cairelli et al. 2015, Cameron et al. 2015, Lee et al. 2015, Kastrin et al. 2014*b*, Ding et al. 2013, Liang et al. 2013, Cameron et al. 2013, Goodwin et al. 2012, Maciel et al. 2011, Guo & Kraines 2009*b*, Özgür et al. 2010, Schroeder et al. 2007, Park et al. 2017, Gopalakrishnan et al. 2017, Wilkowski et al. 2011*b*, Özgür et al. 2011, Hu et al. 2006). They typically rely on the AnC discovery model and mostly output graph paths that include a number of bridging terms, connecting the start (A) and target (C) concepts. Hence, the output of graph-based approaches greatly assists in generating more comprehensive research hypotheses.

Wilkowski et al. (2011*a*) developed a graph-based approach with semantic predications by adhering to the AnC model, which is the first approach reported in the literature that did not follow the canonical ABC model. This work mainly utilised graph theories, such as the degree centrality of nodes and path analysis using depth first search to output graph paths that represent relationship chains. Same as Wilkowski et al. (2011*a*), Özgür et al. (2011) also made use of network centrality analysis to elicit hidden linkages. Furthermore, some approaches (Baek et al. 2017, Kim et al. 2016) have performed shortest path analysis by using algorithms such as Dijkstra to output the implicit discovery paths. More recently, Cameron et al. (2015) suggested a model that uses SemMedDB database to extract the semantic predications to build the knowledge graph. The main strength of their work is the automatic generation of sub-graphs based on the context/thematic dimension of paths.

Most of the existing graph-based approaches heavily rely on external knowledge resources in their knowledge discovery process. This limits the applicability of these approaches in situations/problems where such resources are unavailable.

Bibliometrics Analysis: Several approaches have utilised bibliographic link structures such as direct citation links, co-citation links, and bibliographic coupling in their knowledge discovery process (Kostoff 2014, Sebastian et al. 2017*b*, 2015, Lee et al. 2015, Ittipanuvat et al. 2014, Nakamura et al. 2014, Ding et al. 2013, Ittipanuvat et al. 2012). The concept of bibliographic coupling is first introduced to LBD research by Kostoff (2014) through his LBD approach that inspected shared references between two disjointed medical literature sets. When two publications cite many common references, then it is said that their bibliometric coupling is strong. Their results (Kostoff 2014) verified the importance of analysing the content in research papers along with their shared references.

Shibata et al. (2009) have shown that *direct* citations are the most effective way of detecting emerging research fronts in a field. As a result, Ittipanuvat et al. (2014) considered direct citation links in their knowledge discovery process. Ding et al. (2013) proposed a network-based approach that utilised biological entities extracted from the literature along with the citation details to construct a network, namely *entitymetrics*. In other words, they constructed an entity-entity citation network by linking biological entities extracted from paper 1 with the biological entities extracted from paper 2, given that paper 1 cites paper 2. The constructed entity-entity citation network is analysed by considering both node-level and cluster-level features to predict the novel entity interactions. The integration of both biological entities and bibliographic entities to the same network is useful as the same network can be utilised to obtain different network-based features.

Sebastian et al. (2017*b*) further extend the bibliometric-based research by using *heterogeneous bibliographic information network* to extract more complex bibliometric-based relationships (e.g., core paper shares a term with other core paper's citer). More specifically, they analysed 16 different bibliometric-based relationship features as cues to detect potential knowledge links.

Link Prediction Approaches: Several approaches have viewed the LBD process as a link prediction problem (Pusala et al. 2017, Sebastian et al. 2017*b*, Kastrin et al. 2016, Sebastian et al. 2015, Kastrin et al. 2014*b*, Crichton et al. 2018, Kastrin et al. 2014*a*). They analyse the attributes of concepts and observed links from the current literature to predict the existence of new links between concepts in the future. The existing link prediction studies can be divided into two main groups; predicting future links between *homogeneous entities*, and predicting future links between *heterogeneous entities*.

Homogeneous networks only consider the 'terms' as 'nodes' and the 'connection' of terms obtained from evidence (e.g., literature and databases) as the 'edges' of their network. More specifically, these LBD studies have considered biological entities as their nodes and the co-occurrence of biological entities extracted from the literature (Pusala et al. 2017, Kastrin et al. 2014*b*, Crichton et al. 2018) or the entity associations extracted from curated databases (Crichton et al. 2018) as their edges. Subsequently, the constructed networks were used to predict the future links between nodes, which are treated as the novel associations in the field. In the second category of link prediction, the networks are created from nodes and edges with diverse entity types. For instance, Sebastian et al. (2017*b*) introduced *Heterogeneous Bibliographic Information Network (HBIN)* to the LBD discipline using terms in the paper, venue of the journal/conference, and author details as *nodes*, and authorship links, citation links, semantic links, and publication links as *edges*. Through HBIN graphs, they have attempted to predict the future co-citation links in the disjointed literature.

Most of the link prediction LBD approaches have employed the state-of-the-art link prediction techniques, such as *Adamic-Adar*, *Common Neighbours*, and *Jaccard Index* in their methodologies or as the baselines.

v.1.4 Supervised/Unsupervised Learning

Incorporating machine learning techniques to analyse and interpret patterns and structures of the literature using *supervised*, *semi-supervised* or *unsupervised learning* is described in this section. The integration of machine learning is not only limited to *knowledge discovery* component of the LBD workflow, but also in other phases, such as *pre-processing* (Hossain et al. 2012, Özgür et al. 2010, Petrič et al. 2012) and *ranking* (Torvik & Smalheiser 2007).

Supervised Learning Approaches: Several approaches have used supervised machine learning techniques in their LBD process (Kastrin et al. 2016, Sang et al. 2015, Özgür et al. 2010, Torvik & Smalheiser 2007, Park et al. 2017, Mower et al. 2016, Sang, Yang, Wang, Liu, Lin & Wang 2018, Gopalakrishnan et al. 2017). For instance, Torvik & Smalheiser (2007) have experimented how machine learning can be adopted to rank the intermediate concepts. More specifically, they have utilised a manually created dataset for the training of a Logistic Regression model, which was employed to rank the generated intermediate concepts. Since manual annotation of data is expensive and time-consuming, some studies have directly used the data from databases such as

AIMED (Özgür et al. 2010), CB (Özgür et al. 2010), SemMedDB (Rastegar-Mojarad et al. 2016), CTD (Rastegar-Mojarad et al. 2016), OMOP (Mower et al. 2016) and Therapeutic Target Database (TTD) (Sang, Yang, Wang, Liu, Lin & Wang 2018) for model training.

As discussed above, supervised learning-based techniques require a large, high-quality dataset to train the model, which is challenging. As a result, some approaches (Sang et al. 2015, Park et al. 2017) have performed semi-supervised learning techniques that make use of a few labelled data with a large amount of unlabeled data. For instance, Sang et al. (2015) have used 5% of the data to create the gold-standard. Unsupervised learning techniques do not require any labelled data and learn from the test data itself to identify potential knowledge associations. The work of Xun et al. (2017), Kastrin et al. (2016), and Bisgin et al. (2011) can be considered as examples of unsupervised learning.

Cluster Analysis: Several approaches have used cluster analysis to detect potential associations between disconnected knowledge areas (Qi & Ohsawa 2016, Cameron et al. 2015, Ittipanuvat et al. 2014, Nakamura et al. 2014, Ittipanuvat et al. 2012, Faro et al. 2011, Kostoff 2011, Kostoff, Block, Stump & Johnson 2008, Kostoff 2008, Kostoff & Briggs 2008, Kostoff, Briggs & Lyons 2008, Kostoff, Solka, Rushenberg & Wyatt 2008, Van der Eijk et al. 2004, Stegmann & Grohmann 2003, Gopalakrishnan et al. 2017, Kostoff et al. 2004, Ye et al. 2010, Petrič et al. 2012, Kostoff & Patel 2015, Kostoff 2014). These approaches can be divided into two categories by considering the *data* used for clustering; *term/document-based clustering* and *citation-based clustering*. The work of Stegmann & Grohmann (2003) that uses co-word analysis along with clustering can be provided as an example for the first category. They analysed cluster properties (e.g., external centrality and internal density) of their keyword-based clusters to identify potential regions where the intermediate terms can be found. Their results revealed that such linking terms reside in regions of below-median centrality and density. The second category of cluster analysis, which is citation-based clustering, can be represented by the work of Ittipanuvat et al. (2014). Initially, they constructed a direct citation network that was classified into clusters using Newman's community detection algorithm. Each cluster was then represented as a term vector to measure the cluster similarity using similarity measures (e.g., cosine, Jaccard Index, and Dice Coefficient) to pair clusters with high similarity from the two domains. For each paired cluster, lexical statistics such as term frequency, document frequency and TF-IDF were calculated to identify potential linking terms.

When considering the *unit of analysis*, cluster-based LBD approaches can be categorised into two groups; *analysis of major clusters*, and *analysis of outliers*. For instance, Ittipanuvat et al. (2014) have considered the top 10 clusters after ordering by cluster size to analyse the potential linking terms. In contrast, Petric et al. (2012) only considered the detected outliers as their unit of analysis in the LBD process. Overall, using cluster analysis techniques, the authors could discover some interesting insights into LBD discipline.

v.1.5 Time Analysis

Dynamic representation of knowledge in the literature can be represented as a time series of snapshots where each snapshot represents the state of the knowledge over an interval of time (e.g., 5 years, 10 years). The knowledge evolution of these dynamic representations facilitates the analysis of different patterns of evolutionary aspects, as described in this section.

Temporal-based Approaches: Few LBD approaches have analysed the evolutionary behaviour of terms to detect the interesting associations between disjointed literature sets (Xun et al. 2017, Loglisci & Ceci 2011, Cohen & Schvaneveldt 2010). For instance, Xun et al. (2017) have observed how the semantics of terms evolve by utilising dynamic

MeSH-based embeddings to track the evolutionary trajectories of MeSH terms in the vector space. More specifically, their research is based on the assumption that if two terms have an evolutionary trend towards each other, it implies that the two terms are more likely to form a relationship in the future. Moreover, Loglisci & Ceci (2011) also considered temporal factor into consideration to analyse the dynamic behaviour of domains using a series of static representations over time. In other words, they have observed several snapshots at different time points to discover potential bisociations between concepts using ARM as the main discovery technique.

v.1.6 User-based Approaches

How the user can be involved in the LBD workflow to enhance the prediction accuracy of the knowledge discovery process is described in this section. We consider two classes of user-based approaches; *query enhancements* (i.e., expanding and enhancing queries based on observations) and *user-interaction* (i.e., incorporating theories of human information-seeking behaviours).

Query Enhancements: This technique is based on query development and enhancements in the literature search engines, which falls under *Literature-Related Discovery (LRD)* methodology proposed by Kostoff et al. (2008). The major steps of LRD are manually creating and executing various queries in the literature search engines, analysing the retrieved articles using CLUTO clustering software, and selecting important themes and phrases in the literature set (Kostoff & Patel 2015, Kostoff 2014, Kostoff & Briggs 2008, Kostoff 2011, 2008, Kostoff, Briggs & Lyons 2008, Kostoff, Solka, Rushenberg & Wyatt 2008, Kostoff & Lau 2013). Furthermore, LRD also includes a multi-step vetting process to filter out associations that are false-positives. The author has proposed several hypotheses using his LRD process, such as chronic kidney disease (Kostoff & Patel 2015), Parkinson's disease (Kostoff 2014), SARS (Kostoff 2011), cataracts (Kostoff 2008), multiple sclerosis (Kostoff, Briggs & Lyons 2008) and water purification (Kostoff, Solka, Rushenberg & Wyatt 2008).

User Interaction Studies: User interaction LBD studies use the theories related to human information-seeking behaviours with the intention of assisting human creativity in generating new knowledge (Wilkowski et al. 2011*b*, Hristovski et al. 2006, Cairelli et al. 2013, Workman et al. 2016, 2014). For example, *information foraging theory* (Pirolli 2007) assesses the information-seeking behaviour of users regarding cost and benefit. In other words, if the user can get the highest amount of benefit spending the lowest amount of energy in the information-seeking activity, it can be considered as optimal foraging. Discovery browsing (a technique based on information foraging theory), was introduced by Wilkowski et al. (2011*b*) using a graph-based approach with semantic predications. This work was an extension of the discovery pattern approach of Hristovski et al. (2006). Wilkowski et al. (2011*b*) allowed the users to iteratively navigate through graph paths of LBD output to gain novel insights about poorly understood relationships. The main advantage of this approach is that it allows the user to fine-tune the LBD process by controlling the growth of the graph. Later, the discovery browsing technique introduced by Wilkowski et al. (2011*b*) was further extended by Cairelli et al. (2013) and Goodwin et al. (2012).

v.1.7 Enhancements

This section summarises the potential enhancement techniques that can be incorporated to uplift the typical workflow of the LBD process. Up to now, LBD is treated as a support tool for researchers as it requires human assistance and creativity to interpret the predicted knowledge associations, and to formulate them into a research hypothesis. However, LBD is an *innovation* problem where the ultimate goal is to construct a model of human-level creativity in detecting novel knowledge. This is where the techniques in *computational creativity* (more specifically, *linguistic creativity*) can be useful. This section discusses the studies that have attempted to integrate techniques involving linguistic creativity into the LBD workflow.

Creativity Techniques and Problem Solving: Vicente-Gomila (2014) pointed out the importance of identifying the relationship between logical causes and effects, in order to enhance the traditional LBD process. To facilitate this, they have utilised TRIZ (Moehrle 2005), which suggests innovative solutions for problem-solving by identifying and generalising patterns across various disciplines. Vicente-Gomila (2014) suggested that incorporating such human-like logic sense will also reduce human intervention during the knowledge discovery process with less degree of implication.

Storytelling Algorithms: Storytelling algorithms provide a different perspective to LBD research and can be considered as a more improved version of the AnC model. However, the LBD community seems to have overlooked the importance of such algorithms in the LBD process. For example, Hossain et al. (2012) utilised an approach based on storytelling algorithm on PubMed abstracts to build story chains involving biological entities. The authors of this study argue that this type of work can be served as a valuable discovery aid to develop hypotheses for the users.

v.2 What topics/central themes emerged over time in the LBD discipline?

In order to analyse topics/central themes emerged in the discipline over time, we created a timeline indicating the events that occurred first in the field of LBD (Figure 2.2). For example, *lexical statistics* was marked in 1996 since the very first LBD experiment that used *lexical statistics* was published in 1996. Note that this timeline has been created by only using the articles selected for the review.

The events in *black* show the evolution of computational techniques discussed in Section v.1. As seen in the timeline, the computational techniques have roughly evolved in the sequence of: *lexical statistics* → *distributional semantics* → *ARM* → *knowledge-based* → *semantic profiling* → *hierarchical-based* → *relations-based* → *machine learning* → *LRD* → *ILP* → *network analysis* → *temporal analysis* → *topic modelling* → *user interaction studies* → *linguistic creativity* → *bibliometrics-based* → *embedding techniques*. According to the timeline, the most recently emerged techniques include latest embedding techniques, such as GloVE, DeepWalk, LINE, node2vec, and SDNE. These techniques have also been successfully applied in many other recent NLP tasks (Hashimoto et al. 2015). Embedding methods based on neural networks (Collobert & Weston 2008, Mnih & Hinton 2009, Mikolov, Sutskever, Chen, Corrado & Dean 2013) are at the forefront of this trend due to their scalability, simplicity and semantic richness (Hashimoto et al. 2015). However, non-parametric embedding techniques have also proven to exhibit similar properties as embeddings based on neural networks (Levy & Goldberg 2014*a*). Therefore, future LBD research could be further expanded by experimenting with the efficiency of these techniques in the knowledge discovery process.

In addition, we have also denoted other special events in the discipline, where *orange* denotes the changes in the data sources, *blue* represents the popular LBD tools, and *green* shows the different evaluation techniques developed. When it comes to data sources, Gordon et al. (2002) have initially attempted to utilise a non-traditional data source by extracting data from the *World Wide Web (WWW)*. Later, different other traditional and non-traditional sources, such as *drug labels* (Bisgin et al. 2011), *patents* (Vicente-Gomila 2014, Maciel et al. 2011), *Tweets* (Bhattacharya & Srinivasan 2012), *Google news* (Maclean & Seltzer 2011) and *non-English data resources* (Su & Zhou 2009, Gao, Wang, Tao, Liu, Li, Yu, Yu, Tian & Zhang 2015, Qian et al. 2012, Yao et al. 2008) have been experimented in the field. With regards to LBD tools, the Arrowsmith project

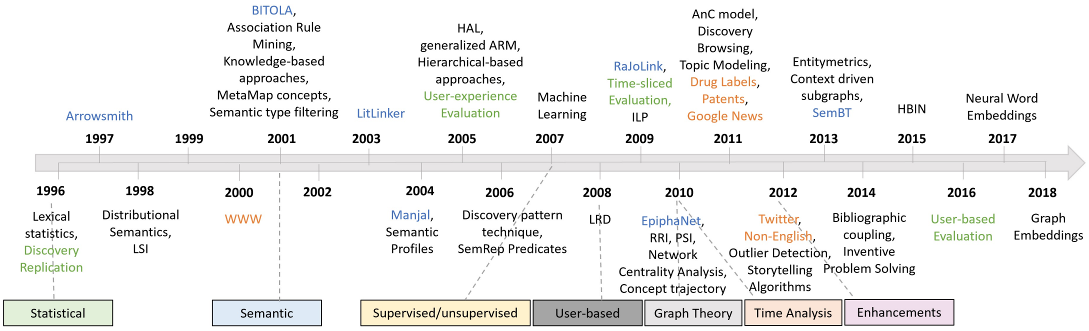


FIGURE 2.2: Evolution of the LBD discipline over time

initiated by Swanson & Smalheiser (1997, 1999) can be considered as the first tool that supported the LBD process. Subsequently, different other tools were introduced in the discipline roughly in the sequence of: BITOLA → LitLinker → Manjal → RaJoLink → EpiphaNet → SemBT.

vi General Overview

vi.1 What is the evidence that LBD generates discovery?

The main objective of LBD is to generate new knowledge by combining existing literature, as demonstrated through Swanson's discoveries since 1986 (Swanson 1986). In the first discovery of Swanson, he manually analysed the titles of *fish oil* and *Raynaud's disease* literature, where he observed that the patients with Raynaud's disease tend to have high *blood viscosity* and high *platelet aggregation*, and fish oil (eicosapentaenoic acid) helps to decrease the *blood viscosity* and *platelet aggregation*. By combining these knowledge pairs, he concluded that *'Raynaud's disease can be cured using fish oil'*. The significance of his hypothesis is due to the disjointedness of the two literature sets. That is, the articles of *fish oil* and *Raynaud's disease* have not mentioned, cited or co-cited each other.

Swanson followed the same thinking process in uncovering the implicit relationship between *Migraine and Magnesium*. Subsequently, Swanson introduced several other new medical discoveries (Swanson & Smalheiser 1996), such as *Somatomedin C and Arginine*, *Dietary Magnesium and Neurologic disease*, *Indomethacin and Alzheimer's disease*, *Estrogen and Alzheimer's disease*, and *Phospholipases and Sleep*.

In addition to Swanson's discoveries, other proposals of novel knowledge discovered through the LBD process include *Alzheimer's disease and Gut microbiota* (Gubiani et al. 2017), *Oral Lichen Planus and Depression* (Zhan et al. 2017), *Alzheimer's disease and Parkinson's disease* (Kim et al. 2016), *Neovascularization in Diabetic Retinopathy* (Maver et al. 2013), *Parkinson's disease* (Hristovski et al. 2010), *Autism and Calcineurin* (Petrič et al. 2009, Petrič et al. 2012), *Down syndrome and Cell polarity* (Thaicharoen et al. 2009), *Deafness and Macular Dystrophy* (Van der Eijk et al. 2004), *Insulin and Ferritin* (Van der Eijk et al. 2004), *Prions and Manganese* (Stegmann & Grohmann 2003), *Obese patients* (Cairelli et al. 2013), *Serotonin* (Wilkowski et al. 2011*b*), *Chronic kidney disease* (Kostoff & Patel 2015) and *Huntington disease* (Hristovski et al. 2006).

Even though LBD contributes to detect anticipated novel knowledge linkages (as discussed above), this raises the question *'how do authors claim that the detected discoveries are actual discoveries?'*. Kostoff et al. (2009) argue that to label an anticipated discovery candidate as a potential discovery, the following four checks are required as a minimum: 1) check for the co-occurrence of the discovery candidate and the core problem in the core problem literature, 2) check the discovery candidate citing papers for mention of the core problem, 3) check for the co-occurrence of the discovery candidate and the core problem in the patent literature, and 4) involve an expert(s) in the core problem area to check whether the discovery candidate is an actual discovery. Kostoff et al. (2009) emphasis the importance of passing these four checks to consider a potential discovery candidate as an actual discovery by revising the novel knowledge proposed in the past LBD studies.

Extending the discussion of Kostoff et al. (2009), we summarise the validation techniques used in the literature to determine whether the detected knowledge associations of the LBD process are true discoveries.

vi.1.1 Evidence-based Validation

In this section, we discuss the sources that are used in the LBD literature to validate whether the proposed discovery candidates are scientifically sensible valid research discoveries.

Using literature databases (such as *MEDLINE* and *Web of Science*) to check whether the detected discovery candidate has co-occurred with the core problem (starting concept) is the commonly used validation resource in the LBD literature (Lever et al. 2018, Yang et al. 2017, Xun et al. 2017, Pusala et al. 2017, Preiss & Stevenson 2017, Sebastian et al. 2017*b*, Kim et al. 2016, Preiss et al. 2015, Kastrin et al. 2014*b*, Petric et al. 2014, Cheung et al. 2012*a*, Maciel et al. 2011, Baker & Hemminger 2010, Cohen, Schvaneveldt & Widdows 2010, Yetisgen-Yildiz & Pratt 2009, 2006, Huang et al. 2005*b*,*a*, Rastegar-Mojarad et al. 2016, Gulec et al. 2010, Peng et al. 2017, Cohen, Whitfield, Schvaneveldt, Mukund & Rindflesch 2010, Crichton et al. 2018, Hristovski et al. 2001, Yetisgen-Yildiz 2006, Frijters et al. 2010). Some studies have verified the validity of their LBD output by cross-referencing their results with curated databases. Examples of curated databases include *Comparative Toxicogenomics Database (CTD)* (Rastegar-Mojarad et al. 2015, Ding et al. 2013, Cheung et al. 2012*a*), *SIDER2* (Shang et al. 2014), *GEO* (Faro et al. 2011) and *GAD* (Seki & Mostafa 2009, 2007). Few studies have used discussion forums (such as *UseNet* (Gordon et al. 2002)) and public websites (such as *Mayo Clinic* (Vidal et al. 2014)) as their validation resources. Extracting reference sets from the previous literature reviews (Vlietstra et al. 2017) and the results of other publications (Malec et al. 2016) as the validation resources is another technique used in the LBD literature. Few studies have attempted to prove the validity of their proposed discovery candidates through the results of laboratory experiments (e.g., *clinical trials*) (Baek et al. 2017, Ramadan et al. 1989). However, validating all detected discovery candidates of LBD systems using laboratory experiments is infeasible. Hence, the most likely to be successful discovery candidate is usually picked to prove its validity.

vi.1.2 Expert/User-oriented Validation

In expert-based validation, typically one (Gubiani et al. 2017, Ittipanuvat et al. 2012, Guo & Kraines 2009*b*, Petrič et al. 2009, Gordon et al. 2002, Urbančič et al. 2007) or two (Baek et al. 2017, Hanauer et al. 2014) domain experts inspect the LBD output to validate whether the detected discovery candidates are meaningful. Alternatively, the domain expert(s) may provide with a more open-ended validation (Gordon et al. 2002) by asking them to provide anticipated future associations in the domain from their experience, without actually looking at the LBD output. Afterwards, the list of potential associations provided by the expert is cross-checked against the actual LBD results as the validation source. However, expert-based evaluation is expensive, time-consuming and suffers from subjectivity. Qi & Ohsawa (2016) have involved both experts and non-experts in validating the detected knowledge candidates using a score derived using three criteria: *utility* (how useful is the generated hypothesis?), *interestingness* (how interesting is the generated hypothesis?), and *feasibility* (to what extent the generated

Table 2.3: LBD tools and their main computational techniques

Tool	**Main Computational Techniques**
Arrowsmith	Statistical-based (*relative and absolute frequencies*), Knowledge-based (*MeSH, UMLS*), Machine Learning (*aggregate local and global features to obtain a composite ranking function*)
LitLinker	Probabilistic (*z-score*), Knowledge-based (*MeSH*), Association Rule Mining (*Support and Confidence*)
RaJoLink	Statistical-based (*rarity*), Knowledge-based (*MeSH, HUGO, ToppGene, Endeavour, STRING*)
Bitola	Knowledge-based (*UMLS*), Predicate-based (*SemRep, BioMedLEE*), Discovery patterns (*two disease treatment patterns*)
Manjal	Statistical-based (*TF-IDF*), Knowledge-based (*MeSH, UMLS*), Profile-based (*weighted vectors with semantic type groupings*)
DAD	Statistical-based (*frequency*), Knowledge-based (*UMLS*)
Anni	Knowledge-based (*UMLS, Gene Ontology*), Profile-based (*weighted related concepts*))
IRIDESCENT	Statistical-based (*Mutual Information Measure*), Knowledge-based (*MeSH, OMIM, drug names from FDA, Locuslink, Gene Ontology*)

hypothesis can be realised?). The derived scores have been used to verify the validity of their LBD process.

vi.2 What are the main LBD tools available, and what are their supported domains?

In this section, we outline the popular LBD tools developed over time in the field. Even though some of these tools are no longer available for online use, we still discuss how they have been developed as the underlying algorithms of these tools will be useful for future LBD research. Table 2.3 summarises the key computational techniques utilised in each tool.

vi.2.1 Arrowsmith

After the manual discoveries of Swanson (1986, 1988), he initiated the Arrowsmith project, which is the very first semi-automatic tool in the LBD literature (Swanson & Smalheiser 1997, 1999). The system uses a simple frequency-based metric of word occurrences to obtain the C-concept. Even though Arrowsmith sets a promising start to develop tools that automate the LBD process, its scope was limited as it did not contain a strong lexical approach.

Later, Smalheiser (Smalheiser 2005, Smalheiser et al. 2009, 2006) initiated the second version of the Arrowsmith project[1]. Even though Arrowsmith initially utilised a basic statistical approach, over time, Smalheiser improved the tool by incorporating knowledge from medical resources (such as MeSH (Swanson et al. 2006) and UMLS (Torvik & Smalheiser 2007)), and machine learning techniques (Torvik & Smalheiser 2007).

[1] http://arrowsmith.psych.uic.edu/arrowsmith_uic/index.html

vi.2.2 LitLinker

Pratt & Yetisgen-Yildiz (2003) developed an LBD tool that supports the open discovery process, namely *LitLinker*. LitLinker utilises a knowledge-based approach using UMLS for concept extraction, filtering and clustering. Their methodology employs association rule mining to identify the associations between concepts. They replicated Swanson's *Migraine↔Magnesium* to evaluate their methodology.

A later version of LitLinker utilises MeSH descriptors, instead of UMLS concepts to represent the documents (Yetisgen-Yildiz & Pratt 2006). Moreover, to identify the associated concepts, they used z-score as a metric that is based on the background distribution of the term probabilities. Through this approach, they were able to provide new insights into Swanson's discoveries by identifying the following new associations; *Alzheimer disease↔Endocannabinoids*, *Migraine↔AMPA receptors* and *Schizophrenia↔Secretin*. However, Kostoff et al. (2009, 2007) have questioned whether the above-claimed discoveries are truly novel as they were not cross-checked against patent databases (where they found evidence that the suggested knowledge discoveries have actually been published prior in patent databases). Later, Yestisgen-Yildiz & Pratt (2009) proposed a way to automatically create a gold standard dataset using *time-sliced evaluation*. They used this evaluation technique to compare the performance of LitLinker with other techniques using 100 disease names.

vi.2.3 RaJoLink

Even though most existing LBD approaches heavily rely on the idea of frequent terms, RaJoLink (Petrič et al. 2009) make use of 'rarity' as the main underlying principle of the knowledge discovery process based on *associationist creativity theory* (Mednick 1962). Another distinguishing feature of RaJoLink is their unique discovery model employed to identify the hidden connections. Instead of using the ABC model as it is, they have combined both *open* and *closed* discovery models into one model to identify potential associations. To reduce the search space and speed up the whole process, a human expert focuses on the neighbouring documents in the similarity graph to detect potential linking terms. This process was done with the use of OntoGen tool using TF-IDF as a metric to measure the document similarity. As for evaluation, the authors applied their model in Autism literature to detect novel, implicit associations that were confirmed through expert evaluation. Later, an enhanced rarity based RaJoLink model (Petric et al. 2014) was proposed using open discovery. It combined text mining and gene prioritising techniques to identify gene↔disease associations using MeSH and HUGO based term identification. In this method, the authors proposed four types of re-ranking methods using the two web servers; *ToppGene* and *Endeavour*, and two propagation-based methods; *personalised PageRank* and *diffusion kernel method*.

vi.2.4 Bitola

Hristovski et al. (2003, 2005, 2006) developed BITOLA that supports both open and closed discovery of LBD. BITOLA detects novel *disease↔gene* associations using association rule mining and background medical knowledge. The system makes use of MeSH, gene symbols and chromosomal locations. Given a disease X, the system initially identifies the disease characteristics (concepts B) as linking terms using association rule A→B. Afterwards, the genes related to the previously identified disease characteristics are selected as the target C concepts using association rule B→C. The system also employs UMLS semantic type filter, chromosomal locations filter, and relationship strength filter (support and confidence) to limit the derived associations.

However, using co-occurrence to identify novel associations suffers from several drawbacks: 1) system tend to identify many semantically unrelated connections (false positives), 2) all co-occurrence pairs identified in MEDLINE are not necessarily interesting and 3) user needs to understand the nature of the specified association by manually reviewing the research articles. As a result, Hristovski et al. (2006) proposed a predication-based approach using SemRep and BioMedLEE with the use of discovery patterns based approach (discussed in Section v.1.2). BITOLA was further improved by integrating a pre-processor for the gene symbol disambiguation process using a Chi-square based scoring method (Kastrin et al. 2010, Kastrin & Hristovski 2008).

Subsequently, Hristovski et al. (2010) also developed a semantic version of the BITOLA, namely *SemBT*[2] that integrates semantic relations with microarray results. They have demonstrated the use of their system using microarray data on Parkinson's disease along with semantic relations to detect novel and implicit treatments.

vi.2.5 Manjal

Srinivasan (2004) introduced Manjal that supports both open and closed knowledge discovery process. Her proposed approach leverages semantic profiling technique based on the relationships between MeSH terms and UMLS semantic types. In her work, a profile is a set of weighted MeSH terms that are grouped to denote a specific UMLS semantic type. More specifically, all the MeSH terms in the given article set are retrieved and weighted based on TF-IDF scheme. Each UMLS semantic type constructs a vector of MeSH terms and normalises their weights within each vector. The user can restrict the intermediate and target concepts by specifying UMLS semantic types of interest. Srinivasan has replicated five of Swanson's discoveries to evaluate her Manjal system.

vi.2.6 Other LBD Tools

Weeber et al. (2000) developed *DAD (Drug-ADR-Disease)*, a knowledge-based tool that utilises UMLS and MetaMap. *Anni* is a concept profile-based LBD tool, where a profile contains weighted concepts co-occurred with the target concept (Jelier et al. 2008). Wren et al. (2004) developed *IRIDESCENT* by integrating fuzzy logic, as discussed in Section v.1.1. In a later study, Wren (2004) extended mutual information measure to detect potential associations.

Apart from the above discussed main LBD tools, other LBD tools such as *CrossBee*[3] (Juršič et al. 2012, 2013), *EpiphaNet*[4] (Cohen, Whitfield, Schvaneveldt, Mukund & Rindflesch 2010), *Spark* (Workman et al. 2016), *Transcriptional Regulatory Modules Extracted from Literature (TREMEL)*[5] (Roy et al. 2017), *Biolab Experiment Assistant (BAE)*[6] (Persidis et al. 2004) and *Dragon Exploratory System on Hepatitis C Virus (DESHCV)*[7] (Kwofie et al. 2011) have also been built to support the knowledge discovery process. These tools have been utilised by several LBD studies to identify novel knowledge associations, which supports the need for such systems to solve real-world problems. For example, Swanson et al. (2001), Gao et al. (2015) and Dong et al. (2014) have used *Arrowsmith*, Gubiani et al. (2017) have used *CrossBee*, Zhan et al. (2017) have used *BITOLA*, Vos et al. (2013) have used *Anni*, and Maver et al. (2013) have used *SemBT* to predict potential discoveries. Unfortunately, the existing LBD tools only support medical-based searches. This clearly showcases the requirement of constructing cross-domain LBD systems.

[2] http://sembt.mf.uni-lj.si

[3] http://crossbee.ijs.si/

[4] http://epiphanet.uth.tmc.edu

[5] http://binf1.memphis.edu/tremel

[6] https://www.biovista.com/research/bea/

[7] http://apps.sanbi.ac.za/DESHCV/

vi.3 What are the applications of LBD research?

As discussed in Section i, the main objectives of LBD research are; 1) alleviating the problem of the knowledge over-specialisation, and 2) assisting to formulate scientifically sensible novel research hypotheses. However, LBD techniques have also been successfully applied to other application areas as described below.

vi.3.1 Drug Development

Developing a novel drug for existing diseases is of vital importance as it could save millions of human lives. Some studies have contributed to this application area by employing their LBD process to discover novel treatments for existing diseases (Kostoff 2008, Kostoff & Briggs 2008, Kostoff, Briggs & Lyons 2008, Kostoff 2014, Zhang et al. 2014, Hu et al. 2003, Kim & Park 2016). Several examples of the diseases that have been explored in LBD are *Parkinson's disease* (Kostoff & Briggs 2008, Kostoff 2014), *multiple sclerosis* (Kostoff, Briggs & Lyons 2008) and *cataracts* (Kostoff 2008). Some of these proposed treatments have even been verified through clinical tests (Baek et al. 2017), which shows the potential usage of LBD in developing new drugs.

vi.3.2 Adverse Drug Reactions

Prevention of fatal adverse drug events is another application area where the LBD process has been successfully applied (Hristovski, Kastrin, Dinevski, Burgun, Žiberna & Rindflesch 2016, Shang et al. 2014, Rastegar-Mojarad et al. 2016, Mower et al. 2016, Malec et al. 2016, Banerjee et al. 2014). Therefore, LBD can be considered as a useful technique for early prediction of unanticipated adverse drug reactions by automatically analysing clinical notes and literature.

vi.3.3 Drug Repositioning

Drug repositioning is the process of detecting novel therapeutic uses and applications for existing drugs (Andronis et al. 2011). This is a highly useful application as it involves 500-2000 million of dollars with 10 -15 years of effort to invent a new drug (Henry & McInnes 2017). However, the success rate of a new drug is less than 10% (Rastegar-Mojarad & Prasad 2015, Henry & McInnes 2017). As a result, several LBD studies have developed drug repositioning LBD systems to cater to this issue (Yang et al. 2017, Rastegar-Mojarad et al. 2015, Park et al. 2017, Rastegar-Mojarad et al. 2016, Sang, Yang, Wang, Liu, Lin & Wang 2018, Lekka et al. 2011, Rastegar-Mojarad & Prasad 2015).

In addition to the above discussed popular application areas, the LBD process has also been employed in the following problem areas.

vi.3.4 Cross-domain Research Collaboration Recommendation

Hristovski et al. (2015, 2016) have utilised LBD paradigm to recommend novel cross-domain research collaborations. To facilitate this, they have developed a network with *author names* and *biomedical concepts* as the major node types, and *writes_about* and *co_author* as the major edge types. Using the suggestions proposed through the open discovery process, they select author profiles that write about these suggested concepts to propose potential collaborations. Kothari & Payne (2015) have also attempted to identify cross-disciplinary research teams by using a keyword-based approach.

vi.3.5 Clinical Guidelines Update Process

The procedure of creating, reviewing and updating clinical guidelines is expensive and laborious. As a result, the guidelines update usually occurs based on a fixed schedule (such as every two years) which often leads into situations where guidelines get out of date by the time they are published (Iruetaguena et al. 2013). Inspired from the LBD open discovery model, Iruetaguena et al. (2013) have attempted to support the decision-making process of clinical guidelines update by recommending new articles to the medical committee as the starting point to update clinical guidelines without manually searching and reading the articles.

vi.3.6 Co-citation Prediction

While most existing LBD approaches have attempted to elicit future links among concepts, Sebastian et al. (2017*b*, 2015) have formulated the problem of LBD as a co-citation prediction task. They have utilised heterogeneous bibliographic information network analysis to predict the potential novel co-citations that are likely to occur in the future, as discussed in Section v.1.3.

vi.4 What domains are considered in LBD research, and what are the levels of generalisability for these domains?

When analysing the literature, it is evident that the majority of the studies have only contributed to the medical domain and its applications. The reason for this might be the availability of highly specific and descriptive content in medical research papers, which is necessary for the LBD process (Ittipanuvat et al. 2014). To date, only a handful of research studies have been performed in non-medical domains, as shown in Table 2.4. For example, the only LBD study that has been performed in the *computer science* domain is to find suitable implicit applications of the Genetic algorithm (Gordon et al. 2002). This points out the importance of contributing to non-medical LBD research, since automated knowledge discovery is beneficial to research development despite the domain.

Furthermore, we also analysed the domain generalisability of LBD research. In other words, we examined the extent to which the existing methodologies can be applied in other domains. For this purpose, we used the following categories to divide the literature, as summarised in Table 2.5.

- *Category 1 (only limited to specific medical problem/subdomain):* This category represents LBD methodologies that can only be applied to a specific medical problem/-subdomain. That is, these methodologies are specialised to a certain problem and cannot be generalised even within the medical domain itself (e.g., LBD methodologies that are specialised only to find associations between 'diseases' and 'drugs', in order to fulfil purposes such as drug repositioning (Rastegar-Mojarad et al. 2015) or adverse drug reactions (Shang et al. 2014)).
- *Category 2 (can be used in the medical domain in general):* If an LBD methodology can be used in any problem/area related to the medical domain (but not in other domains), we consider it as a Category 2 methodology.
- *Category 3 (can be used in other domains in addition to the medical domain):* This category denotes LBD methodologies that are originally proposed in the medical domain but can also be used in other domains since they do not specifically use any medical domain related resources in their methodologies. However, their usage in other domains has not explicitly been verified or tested by the authors.

Table 2.4: Domains in which LBD experiments have been conducted

Domain	**Past Studies**
Medical	(Lever et al. 2018, Gubiani et al. 2017, Baek et al. 2017, Vlietstra et al. 2017, Yang et al. 2017, Zhan et al. 2017, Xun et al. 2017, Pusala et al. 2017, Preiss & Stevenson 2017, Sebastian et al. 2017*b*, Huang et al. 2016, Hristovski, Kastrin, Dinevski, Burgun, Žiberna & Rindflesch 2016, Preiss & Stevenson 2016, Kastrin et al. 2016, Qi & Ohsawa 2016, Kim et al. 2016, Zhou et al. 2015, Song et al. 2015, Preiss et al. 2015, Cairelli et al. 2015, Cameron et al. 2015, Rastegar-Mojarad et al. 2015, Srinivasan et al. 2015, Sebastian et al. 2015, Sang et al. 2015, Lee et al. 2015, Shang et al. 2014, Hanauer et al. 2014, Vicente-Gomila 2014, Dong et al. 2014, Tsafnat et al. 2014, Workman et al. 2014, Kastrin et al. 2014*b*, Vidal et al. 2014, Petric et al. 2014, Vos et al. 2013, Ding et al. 2013, Liang et al. 2013, Iruetaguena et al. 2013, Cameron et al. 2013, Gabetta et al. 2013, Cherdioui & Boubekeur 2013, Maver et al. 2013, Cohen et al. 2012, Cheung et al. 2012*a*, Miller et al. 2012, Hossain et al. 2012, Bhattacharya & Srinivasan 2012, Goodwin et al. 2012, Faro et al. 2011, Maciel et al. 2011, Bisgin et al. 2011, Kostoff 2011, Kwofie et al. 2011, Cohen et al. 2011, Guo & Kraines 2009*b*, Maclean & Seltzer 2011, Loglisci & Ceci 2011, Hur et al. 2010, Baker & Hemminger 2010, Ijaz et al. 2009, Cohen, Schvaneveldt & Widdows 2010, Hu et al. 2010, Özgür et al. 2010, Kraines et al. 2010, Hristovski et al. 2010, Cohen & Schvaneveldt 2010, Kastrin et al. 2010, Vidal et al. 2010, Yetisgen-Yildiz & Pratt 2009, Smalheiser et al. 2009), ...
Other domains	Industrial domain (electric vehicles energy storage systems) (Vicente-Gomila 2014), Water purification techniques (Kostoff, Solka, Rushenberg & Wyatt 2008), Robotics↔Gerontology (Ittipanuvat et al. 2014, 2012), Chance discovery↔Olympic games (Qi & Ohsawa 2016), Counterterrorism (Jha & Jin 2016*a*), Built environment (Kibwami & Tutesigensi 2014), Genetic algorithms (Gordon et al. 2002), Chinese agricultural economics (Huang et al. 2012), Crime investigation (Schroeder et al. 2007), Climate science (Marsi et al. 2014), Sustainability issues↔Aviation industry (Nakamura et al. 2014)

- *Category 4 (proven to be useful in medical and other domains):* This category represents LBD studies same as Category 3, except for the fact that the authors have verified or tested the suitability of their medical LBD approach in other domains as well.
- *Category 5 (other domains):* If the LBD methodology is proposed in a non-medical domain, it is categorised under this category.

The following conclusions can be made by analysing Table 2.5; 1) most medical LBD studies rely on medical domain knowledge, making them infeasible to apply to other domains, 2) a substantial amount of medical studies are not generalised even within the medical domain itself, 3) the usage of most of the domain-independent medical approaches has not been validated or tested in other domains. 4) validating LBD methodologies in both medical and non-medical domains to demonstrate their domain-independency has not received much attention from the LBD community, and 5) the LBD approaches that have performed outside the medical domain have rarely evaluated their ability to detect medical discoveries (e.g., (Huang et al. 2012)).

Table 2.5: Level of generalisability of the existing LBD literature

Category	**Past Studies**
Category 1	Baek et al. (2017), Vlietstra et al. (2017), Yang et al. (2017), Huang et al. (2016), Hristovski, Kastrin, Dinevski, Burgun, Žiberna & Rindflesch (2016), Kim et al. (2016), Zhou et al. (2015), Cairelli et al. (2015), Rastegar-Mojarad et al. (2015), Srinivasan et al. (2015), Shang et al. (2014), Hanauer et al. (2014), Dong et al. (2014), Tsafnat et al. (2014), Vidal et al. (2014), Petric et al. (2014), Ding et al. (2013), Liang et al. (2013), Cameron et al. (2013), Gabetta et al. (2013), Maver et al. (2013), Cohen et al. (2012), Cheung et al. (2012*a*), Bhattacharya & Srinivasan (2012), Faro et al. (2011), Maciel et al. (2011), Bisgin et al. (2011), Kwofie et al. (2011), Hur et al. (2010), Baker & Hemminger (2010), Ijaz et al. (2009), Özgür et al. (2010), ...
Category 2	Lever et al. (2018), Xun et al. (2017), Pusala et al. (2017), Preiss & Stevenson (2017), Kastrin et al. (2016), Song et al. (2015), Preiss et al. (2015), Cameron et al. (2015), Sang et al. (2015), Lee et al. (2015), Kastrin et al. (2014*b*), Cherdioui & Boubekeur (2013), Miller et al. (2012), Hossain et al. (2012), Goodwin et al. (2012), Kostoff (2011), Cohen et al. (2011), Guo & Kraines (2009*b*), Maclean & Seltzer (2011), Loglisci & Ceci (2011), Cohen, Schvaneveldt & Widdows (2010), Hu et al. (2010), Kraines et al. (2010), Yetisgen-Yildiz & Pratt (2009), Smalheiser et al. (2009), Petrič et al. (2009), Guo & Kraines (2009*a*), Kostoff, Block, Stump & Johnson (2008), ...
Category 3	Sebastian et al. (2017*b*, 2015), Cohen & Schvaneveldt (2010), Thaicharoen et al. (2009), Bruza et al. (2006), Cole & Bruza (2005), Lindsay & Gordon (1999), Gordon & Dumais (1998), Gordon & Lindsay (1996), McClure (2012), Crichton et al. (2018), Petrič et al. (2012), Urbančič et al. (2007), Kostoff & Lau (2013)
Category 4	Qi & Ohsawa (2016), Vicente-Gomila (2014)
Category 5	Jha & Jin (2016*a*), Ittipanuvat et al. (2014), Nakamura et al. (2014), Marsi et al. (2014), Ittipanuvat et al. (2012), Kostoff, Solka, Rushenberg & Wyatt (2008), Schroeder et al. (2007), Gordon et al. (2002), Huang et al. (2012), Kibwami & Tutesigensi (2014)

vii Statistical Analysis

The statistical analysis of this review was performed with the intention of identifying current trends in the LBD discipline regarding publications over the years, top-cited papers and authors.

vii.1 What are the trends in LBD research in terms of publications over the years, top-cited papers and top authors?

The line chart in Figure 2.3 depicts the publication counts in the LBD discipline for each year. We only considered the publications in the dataset that we developed for this review to obtain the statistics. The publication count of the year 2018 is not mentioned, as we collected the data for the review at the beginning of May 2018. When analysing the chart, it is visible that overall there is a growing research interest in the LBD field.

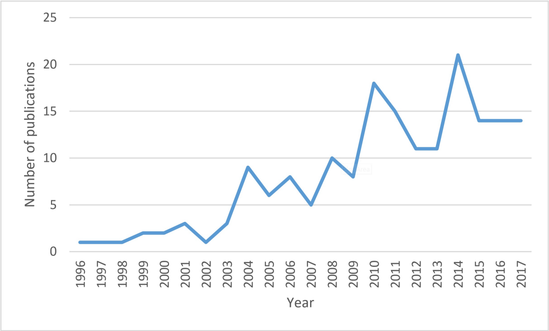


FIGURE 2.3: Publication count over time

A close inspection of the latter part of the chart reveals that LBD research has come to its peak in 2014, drops in 2015 and remains plateau over 2015-2017.

Table 2.6 mentions the top 10 cited papers in LBD that are based on the citation counts extracted from *Google Scholar*. These papers include most of the initial works in the discipline that were published in the time frame of *1996-2005*. When analysing the *purpose* of these papers, it is evident that most of the papers introduce the main LBD tools in the discipline, while others are integrating new computational techniques to the LBD framework for the first time in the discipline. Considering the categories of *techniques*, it is clear that most of the techniques belong to statistical/co-occurrence models and early attempts of incorporating domain knowledge from structured knowledge bases. One reason for the high citation counts could be that these LBD papers have set the foundation of the discipline and thus, providing the background knowledge/history of the LBD literature. However, Kostoff et al. (2009) questioned whether the predicted novel discoveries of most of these initial works are actual discoveries as they fail to fulfil the requirements of an actual discovery (see Section vi.1).

TABLE 2.6: Top cited papers in LBD

Article Title	**Count**	**Purpose**	**Techniques**
An interactive system for finding complementary literatures: a stimulus to scientific discovery (Swanson & Smalheiser 1997)	540	Introducing *Arrowsmith* (version 1) LBD tool	Relative frequency
Text Mining: Generating hypotheses from MEDLINE (Srinivasan 2004)	351	Introducing *Manjal* LBD tool	Weighted topic profiles using MeSH and UMLS
Using concepts in literature-based discovery Simulating Swanson's Raynaud-fish oil and migraine-magnesium discoveries (Weeber et al. 2001)	263	Integrating semantic information into the LBD process	Incorporating background knowledge using UMLS
Using literature-based discovery to identify disease candidate genes (Hristovski et al. 2005)	242	Introducing *BITOLA* LBD tool	Association rule mining and background domain knowledge from medical resources
Knowledge discovery by automated identification and ranking of implicit relationships (Wren et al. 2004)	219	Introducing *IRIDESCENT* LBD tool	Network structures and fuzzy logic
Toward discovery support systems: A replication, re-examination, and extension of Swanson's work on literature-based discovery of a connection between Raynaud's and fish oil (Gordon & Lindsay 1996)	205	Incorporating conventional statistical measures in LBD	Frequencies (absolute and relative) and TF-IGF
Literature-based discovery by lexical statistics (Lindsay & Gordon 1999)	199	Incorporating conventional statistical measures in LBD	Frequencies (absolute and relative) and TF-IDF

Text-based discovery in biomedicine: The architecture of the DAD-system (Weeber et al. 2000)	167	Introducing *DAD* LBD tool and its applicability in adverse drug reactions prediction	Incorporating background knowledge using UMLS
Using latent semantic indexing for literature based discovery (Gordon & Dumais 1998)	165	Integrating distributional semantics into the LBD framework	Latent semantic indexing with neighbourhood analysis
Mining MEDLINE for implicit links between dietary substances and diseases (Srinivasan & Libbus 2004)	148	Drug repositioning using *Manjal* LBD tool	Weighted topic profiles using MeSH and UMLS

Table 2.7: Top cited recent papers (2016-present)

Article Title	**Count**	**Purpose**	**Techniques**
The effect of word sense disambiguation accuracy on literature based discovery (Preiss & Stevenson 2016)	10	Emphasises the importance of Word Sense Disambiguation (WSD)	Three WSD systems: *personalized page rank*, *vector space model*, and *MetaMap*
Literature-based discovery of new candidates for drug repurposing (Yang et al. 2017)	10	Drug repurposing using the *ABC* model	Pattern-based relationship extraction and vector space-based ranking
Link Prediction on a Network of Co-occurring MeSH Terms: Towards Literature-based Discovery (Kastrin et al. 2016)	9	Devising LBD as a link prediction problem (*both supervised & unsupervised*)	Proximity measures: *common neighbor*, *Jaccard coefficient*, *Adamic/Adar index*, and *preferential attachment*
Prioritizing adverse drug reaction and drug repositioning candidates generated by literature-based discovery (Rastegar-Mojarad et al. 2016)	6	Drug repositioning, adverse drug event, & drug-disease relation detection using the *ABC* model	SemMedDB semantic predications and supervised machine learning
Learning the heterogeneous bibliographic information network for literature-based discovery (Sebastian et al. 2017*b*)	6	Devising LBD as a co-citation prediction problem (*supervised link prediction*)	Proves the importance of non-lexical information such as author, venue and citation details using Heterogeneous Bibliographic Information Network (HBIN)
Enriching plausible new hypothesis generation in PubMed Baek et al. (2017)	5	Finding implicit biological associations using the *AnC* model	Graph-based using average semantic relatedness of a path
Spark, an application based on Serendipitous Knowledge Discovery (Workman et al. 2016)	5	Integrating the potential use of information-seeking behaviour in application design	Serendipitous knowledge discovery using semantic predications
Classification-by-Analogy: Using Vector Representations of Implicit Relationships to Identify Plausibly Causal Drug/Side-effect Relationships (Mower et al. 2016)	5	Adverse drug events prediction by mimicking analogical reasoning	Semantic predications, distributional semantics (PSI) and supervised machine learning
Using Literature-Based Discovery to Explain Adverse Drug Effects (Hristovski, Kastrin, Dinevski, Burgun, Žiberna & Rindflesch 2016)	4	Adverse Drug Effects prediction using *SemBT* LBD tool	Semantic predications with discovery patterns

Table 2.8: Top authors in LBD research

Author	**Count**	**Author**	**Count**
Thomas C. Rindflesch	27	Ingrid Petric	4
Dimitar Hristovski	15	Judita Preiss	4
Trevor Cohen	13	Kishlay Jha	4
Ronald N. Kostoff	10	Michael D. Gordon	4
Neil R. Smalheiser	9	Michael J. Cairelli	4
Andrej Kastrin	8	Rein Vos	4
Borut Peterlin	8	Tanja Urbancic	4
Roger Schvaneveldt	7	Vetle I. Torvik	4
Dominic Widdows	6	Bojan Cestnik	4
Marcelo Fiszman	6	Steven B. Kraines	4
Min Song	6	Weisen Guo	4
M. Yetisgen-Yildiz	5	Erik M. van Mulligen	4
Don R. Swanson	4		

We also analysed the papers that have received high citation counts during the time-period of *2016-present* with the intention of identifying most attracted computational techniques used in the recent LBD literature. In other words, we assumed that highly cited recent LBD publications indicate an attractive technique in the LBD literature (Table 2.7).

Regarding the *purpose* of most cited recent publications, it is visible that much of them are contributing to the special-purpose applications areas of LBD (such as adverse drug events and drug repurposing). This highlights the potential of adapting the LBD framework in other problem areas to enhance the reasoning process. Another interesting pattern concerning the *purpose* of these papers is that they have deviated from the typical research setting of the LBD process, which is a ranked list of hidden associations. These redefined research settings include LBD as a *co-citation prediction task*, *link prediction task*, *supervised learning task* and/or *unsupervised learning task*. Overall, it is evident that the LBD community tends to have a special interest in involving techniques in the *link prediction* discipline to uncover hidden associations in the literature.

In terms of *techniques*, Table 2.7 reveals that *semantic predicates*, *network analysis* and *machine learning* are commonly used in most of these publications. Regarding the *network analysis*, while most of the LBD studies focus on homogeneous networks that are constructed only using concepts in the research papers, the study of Sebastian et al. (2017*b*) have incorporated multiple other metadata (such as author details, published venues, and citation details) to construct their heterogeneous network. Their results demonstrate that combining both lexical and non-lexical information tends to perform well in detecting hidden patterns. Preiss & Stevenson (2016) have attempted to measure the effect of *word sense disambiguation (WSD)* accuracy in terms of LBD performance. Their results reveal that WSD is a useful component in LBD systems, and the effectiveness of LBD is sensitive to the accuracy of WSD. Mower et al. (2016) have experimented to integrate the characteristics of *analogical reasoning* into the LBD process by incorporating distributional semantics (PSI).

We believe that the above discussed unique contributions of each study in terms of *purpose* and *techniques* are the main reason for their high citation counts. We also analysed the authors who have mostly contributed to the discipline by considering the number of times each author appeared in the author list (irrespective of the position of the author) as the metric. Table 2.8 summarises the top authors found from our statistical analysis. It is clear that most of the top-cited articles (in Table 2.6) are mostly authored by the top authors in the field.

viii Limitations

Although this review outlines the insights gleaned through our rigorous literature analysis with confidence, we could have missed articles that are outside the six databases and six keywords used. To mitigate this effect up to some extent, we have also added references from the recent review (Henry & McInnes 2017) during our paper retrieval process, as listed in Table 2.2.

ix Conclusions and Future Work

The main discussion points of this review are LBD computational techniques, key milestones of the discipline, validation checks, tools, application areas, domains and generalisability levels. The latter part of the review presents a statistical analysis that attempts to elicit patterns in the LBD literature. We also performed a comparison of the findings in this review with two most recent LBD reviews (Henry & McInnes 2017, Sebastian et al. 2017*a*), which is available at `https://tinyurl.com/review-comparisons`.

LBD was originally evolved with the intention of overcoming the *knowledge over-specialisation* and to support scientists in the *knowledge discovery* process. However, as pointed by this review, special-purpose LBD systems were successfully developed to address issues in other problem areas (such as drug discovery, drug repurposing and adverse drug reactions). Our highly cited recent publications analysis reveals that such applications have received greater attention within the LBD community. However, the application areas explored and verified so far are mainly in the medical domain. Hence, an interesting future direction would be to integrate the LBD frameworks (e.g., the *ABC* model) in other problem areas, such as e-commerce (e.g., product recommendation), entertainment (e.g., movie recommendation), and nutrition (e.g., recipe recommendation) to enhance the accuracy of the predictions in these problem settings.

As discussed in Section vi.2, the LBD discipline has few knowledge discovery tools available such as Arrowsmith, BITOLA, SemBT, LitLinker, Manjal, etc. However, these tools only support medical literature mining. This emphasises the need to develop *cross-domain* LBD tools, which can be considered as a challenging future direction. Two main reasons for their domain-dependency are; 1) underlying algorithm relies on the knowledge extracted from the domain-specific knowledge bases and databases (such as UMLS) to make the predictions, and 2) supporting the literature search only in domain-specific databases (such as PubMed). To overcome the aforementioned two limitations, the proposed algorithm should be; 1) independent of using domain-specific resources. In this regard, the usage of domain-independent knowledge bases (such as DBpedia, Freebase and YAGO) is extremely useful. Unlike hand-crafted knowledge bases, the suggested community-driven knowledge bases are up-to-date, free, multilingual and domain-independent, and 2) supporting the search in other literature databases (such as Web of Science and Scopus) to facilitate domain-independent literature search.

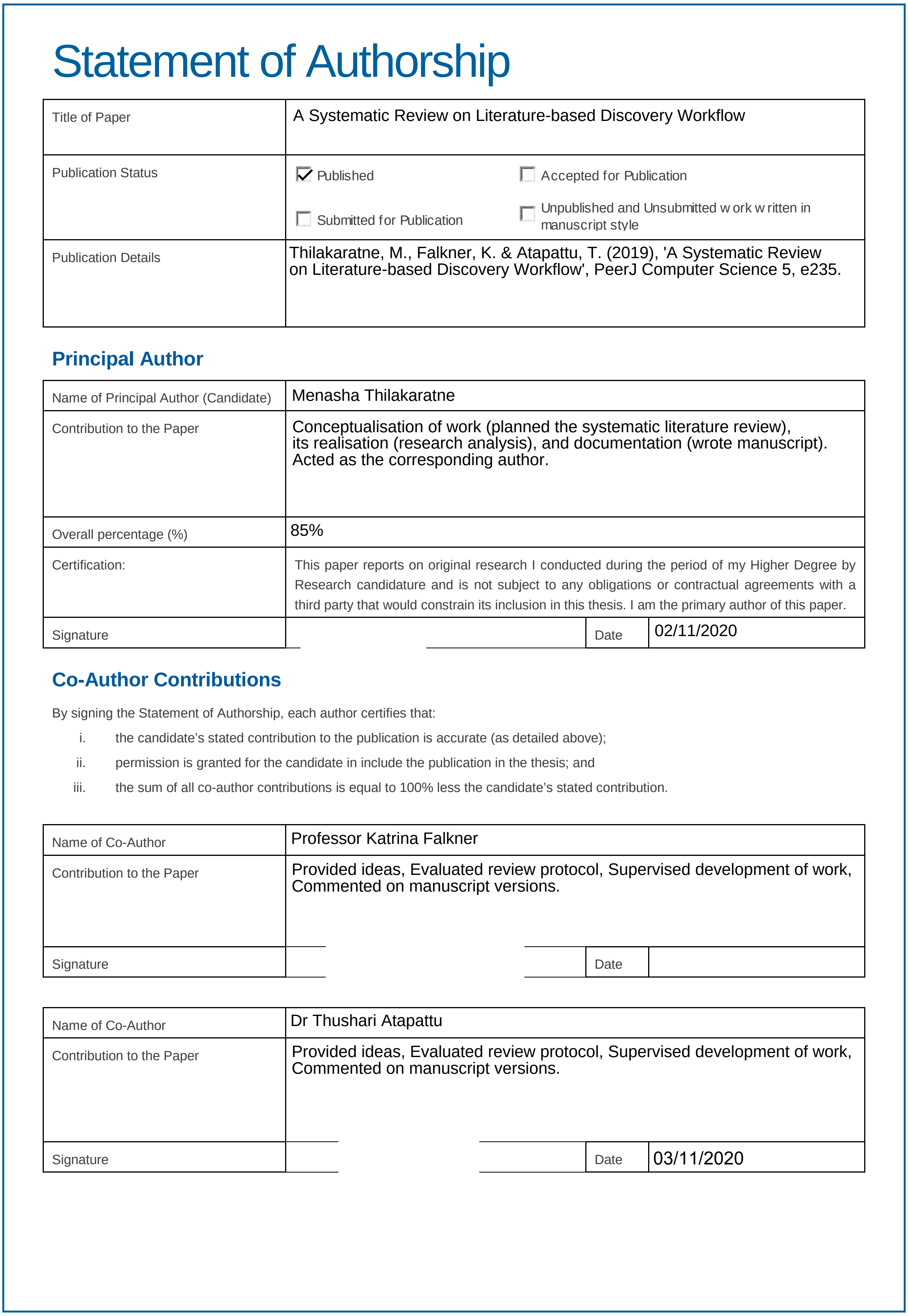
Furthermore, existing LBD tools have paid a little attention in terms of visualisation of their results, user interface and documentation (Weeber et al. 2005). Therefore, it is equally important to alleviate these issues when designing an LBD system. This brings out the importance of conducting Human-Computer Interaction (HCI) research in the field to enable LBD tools to support users with a varied range of expertise and abilities without any formal training. Currently, the usage statistics of LBD tools have been reported to be low. For example, Arrowsmith tool is only used by 1200 unique users on a monthly basis (Li et al. 2013), even though the tool is continuously maintained and available online as a free service. Hence, the involvement of HCI research will also promote awareness of the availability of such discovery tools.

As for the computational techniques, it is evident that much of the early computational approaches have utilised lexical statistics that can be considered as the most primitive technique used in the LBD literature. Later, different other techniques (such as knowledge-based, relations-based, hierarchical-based, graph-based, bibliometrics-based, link prediction-based and distributional semantics-based approaches) were introduced to the discipline. The following methodological trends were revealed from our 1) *statistical analysis of citations* (see Section vii.1), 2) *evolution analysis* (see Section v.2) and 3) *computational techniques analysis* (see Section v.1). The analysis of highly cited recent publications disclosed the trend of using predication/relation-based, network-based, machine learning and link prediction techniques. According to our timeline analysis, the recently emerged techniques include embeddings-based techniques, such as word embeddings (e.g., GloVE) and graph embeddings (e.g., DeepWalk, LINE, node2vec and SDNE). Besides, as shown in our classification of main computational techniques, creativity techniques (Vicente-Gomila 2014) and storytelling algorithms (Hossain et al. 2012) can be considered as two important enhancements in the LBD discipline.

In addition, Korhonen et al. (2014) pointed out that the existing LBD methodologies are limited as they utilise fairly shallow techniques to analyse texts. Hence, they highlight the importance of developing more accurate, dynamic and broader LBD systems through deep analysis and understanding of texts using advanced text mining methods (Korhonen et al. 2014). Moreover, as pointed through Kostoff's *LRD* studies (Kostoff, Briggs, Solka & Rushenberg 2008), it is also important to improve the information retrieval effectiveness from literature databases (such as MEDLINE) by incorporating techniques related to *query expansion* (Symonds, Bruza, Zuccon, Koopman, Sitbon & Turner 2014). The recent advancements in query expansion techniques (Azad & Deepak 2017) will be useful in this regard.

The statistical analysis of the review reveals that the LBD discipline is receiving a growing research interest from the global research fraternities. Despite the valuable contribution of the LBD studies during the last three decades, the field still requires a substantial amount of research to overcome the current limitations. In terms of methodology, the most prevailing limitation of the LBD studies is their restriction to the medical domain by developing highly specific LBD systems that lack generalisability. Most of the studies primarily focus on finding associations among a set of fixed domain concepts: *proteins*, *genes*, *diseases* and *drugs*. Surprisingly, a considerable amount of medical LBD studies are not even generalisable within the medical domain itself. To date, there are a handful of LBD research studies performed outside the medical domain. This points out the importance of developing *domain-independent* LBD solutions in future LBD research whose success do not depend on domain-specific knowledge resources.

With the increasing research trend in the field, we believe that future LBD research will attempt to alleviate these existing limitations by developing fully automated, domain-independent LBD systems with concise and informative visualisations, and robust evaluations. Such LBD systems will not only assist scientists to generate scientifically sensible novel research hypotheses in a shorter time, but also encourage cross-disciplinary research by connecting disjointed knowledge areas.



Statement of Authorship

Title of Paper	A Systematic Review on Literature-based Discovery Workflow
Publication Status	☑ Published □ Accepted for Publication □ Submitted for Publication □ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Thilakaratne, M., Falkner, K. & Atapattu, T. (2019), 'A Systematic Review on Literature-based Discovery Workflow', PeerJ Computer Science 5, e235.

Principal Author

Name of Principal Author (Candidate)	Menasha Thilakaratne		
Contribution to the Paper	Conceptualisation of work (planned the systematic literature review), its realisation (research analysis), and documentation (wrote manuscript). Acted as the corresponding author.		
Overall percentage (%)	85%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	02/11/2020

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Professor Katrina Falkner		
Contribution to the Paper	Provided ideas, Evaluated review protocol, Supervised development of work, Commented on manuscript versions.		
Signature		Date	

Name of Co-Author	Dr Thushari Atapattu		
Contribution to the Paper	Provided ideas, Evaluated review protocol, Supervised development of work, Commented on manuscript versions.		
Signature		Date	03/11/2020

2.5 Publication II

A Systematic Review on Literature-Based Discovery Workflow

As scientific publication rates increase, knowledge acquisition and the research development process have become more complex and time-consuming. *Literature-Based Discovery (LBD)*, supporting automated knowledge discovery, helps facilitate this process by eliciting novel knowledge by analysing the existing scientific literature. This systematic review provides a comprehensive overview of the LBD workflow by answering eight research questions related to the major components of the LBD workflow (i.e., *input*, *process*, *output* and *evaluation*). With regards to the '*input*' component, we discuss the input types and data sources used in the literature. The '*process*' component presents filtering techniques, ranking/thresholding techniques and LBD resources. Subsequently, the '*output*' component focuses on the visualisation techniques used in the LBD discipline. As for the '*evaluation*' component, we outline the evaluation techniques, their generalisability and the quantitative measures used to validate results. To conclude, we summarise the findings of the review for each component by highlighting the possible future research directions.

x Introduction

Due to the exponential growth of scientific publications, keeping track of all research advances in scientific literature has become almost impossible for a scientist (Cheadle et al. 2017). As a result, scientific literature has become fragmented, and individual scientists tend to deal with fragments of knowledge based on their specialisation. Consequently, valuable implicit associations that connect these knowledge fragments tend to remain unnoticed, since scientists in each specialisation have only seen part of the big picture. *Literature-Based Discovery (LBD)* supports cross-disciplinary knowledge discovery to elicit these hidden associations to recommend new scientific knowledge. The recommended novel associations can greatly assist scientists in formulating and evaluating novel research hypotheses (Ganiz et al. 2005). While reducing the time and effort, this will also promote scientists to discover new areas of research.

x.1 Brief History

LBD was developed as a research field from the medical discoveries published by Swanson since 1986. In his first discovery, he manually analysed the titles of two literature sets: *fish oil* and *Raynaud's disease* (Swanson 1986). Swanson observed that patients with Raynaud's disease tend to have high *blood viscosity* and high *platelet aggregation*. He also noted that fish oil contains EPA (eicosapentaenoic acid), which helps to decrease the *blood viscosity* and *platelet aggregation*. By combining these knowledge pairs, he generated the hypothesis; *'Raynaud's disease can be cured using fish oil'*. Furthermore, he also observed that the two literature sets he was referring are *disjointed*. That is, the articles in the two literature sets have not mentioned, cited or co-cited each other. Consequently, he published these findings, which were deduced using the *ABC model* (see Section x.2). His second discovery followed the same process, where he manually examined the titles of *Migraine* and *Magnesium* to detect implicit associations that connects the two literature sets (Swanson 1988). Later, his observations were proven through laboratory experiments that demonstrate the validity of his thinking process (Ramadan et al. 1989).

Even though the early work of Swanson was mostly performed manually by merely analysing the article titles and their word co-occurrence frequencies, they formed the foundation of the field. In accordance with Swanson's experiments, the existing disperse knowledge fragments in the literature can be accumulated in such a way to develop novel semantic relationships that have not drawn any awareness before (a.k.a. *undiscovered public knowledge*) (Swanson 1986). These connectable disperse knowledge fragments in the literature may exist as; 1) hidden refutations or qualifications, 2) undrawn conclusion from different knowledge branches, 3) cumulative weak tests, 4) analogous problems, and/or 5) hidden correlations (Davies 1989). In a later study, Swanson also pointed out the importance of studying cases where the interaction of the two literature sets is not null (i.e., the literature sets are not disjointed), but populated by few articles (a.k.a. *literature-based resurrection* (Swanson 2011), *scientific arbitrage* (Smalheiser 2012)).

x.2 Discovery Models

Most LBD literature is based on the fundamental premise introduced by Swanson, termed *ABC model* (Swanson 1986). It employs a simple syllogism to identify the potential knowledge associations (a.k.a. *transitive inference*). That is, given two concepts A and C in two disjointed scientific literature sets, if concept A is associated with concept B, and the same concept B is associated with concept C, the model deduces that the concept A is associated with the concept C. The popular ABC model has two variants termed *open discovery* and *closed discovery.*

Open discovery is generally used when there is a single problem with limited knowledge about what concepts can be involved. The process starts with an initial concept related to the selected research question/problem (A-concept). Afterwards, the LBD process seeks the relevant concepts that ultimately lead to implicit associations (C-concepts). In other words, only the concept A is known in advance and concepts B and C are identified by the LBD process. Therefore, this model can be viewed as a knowledge discovery process that assists in generating novel research hypotheses by examining the existing literature. Unlike the open discovery process, closed discovery model attempts to discover novel implicit associations between the initially mentioned A-concept and C-concept (a.k.a. *concept bridges*). Thus it represents hypotheses testing and validation process. More explicitly, the LBD process starts with user-defined A-concept and C-concept, and the output will be the intermediate B-concepts that represents the associations between the two user-defined domains.

Even though the prevalent ABC model has contributed in numerous ways to detect new knowledge, it is merely one of several different types of discovery models that facilitates the LBD process. In this regard, Smalheiser (2012) points out the importance of thinking beyond the ABC formulation and experimenting with alternative discovery models in the discipline. Despite the simplicity and power of the ABC model, it also suffers from several limitations such as the sheer number of intermediate terms that exponentially expands the search space and producing a large number of target terms that are hard to interpret manually (Smalheiser 2012). Even though LBD research has suggested various ways to overcome the aforementioned two limitations, most of these studies rely on similarity-based measures to rank the target terms. This will result in LBD systems that merely detect *incremental* discoveries. In addition, the field requires to explore various *interestingness measures* that allow customising the LBD output to facilitate different types of scientific investigations (Smalheiser 2012).

With respect to other LBD discovery models that are enhanced based on the ABC discovery structure include the *AnC* model (where n=(B_1,...,B_n)) (Wilkowski et al. 2011*b*), combined *open* and *closed* discovery model (Petrič et al. 2009), *context-based ABC* model (Kim & Song 2019) and *context-assignment-based ABC* model (Kim & Song 2019). Moreover, recent studies have attempted to further explore alternative discovery models that deviate from the typical ABC discovery setting. These new directions include storytelling methodologies (Sebastian et al. 2017*b*), analogy mining (Mower et al. 2016), outlier detection (Gubiani et al. 2017), gaps characterisation (Peng et al. 2017) and negative consensus analysis (Smalheiser & Gomes 2015). For a comprehensive discussion of contemporary discovery models and future directions, please refer (Smalheiser 2017, 2012).

x.3 Purpose of the Review

Even though there are several review papers (Gopalakrishnan et al. 2019, Henry & McInnes 2017, Sebastian et al. 2017*a*, Ahmed 2016) published on LBD, the field still lacks systematic literature reviews. Therefore, the existing reviews merely cover a subset of the LBD literature and do not provide a comprehensive classification of the LBD discipline. To address this gap, we present a large-scale systematic review by analysing 176 papers that were selected by manually analysing 475 papers. On the contrary to the existing traditional reviews, systematic reviews adhere to a rigorous and transparent method to ensure the future replicability of the findings using a clear systematic review protocol, and to minimise the bias through the focus on empirical evidence, not preconceived knowledge (Mallett et al. 2012).

Another persistence research deficiency of other literature reviews is their limited and adhoc focus points. To date, none of the existing reviews focuses on the LBD workflow as a whole. Moreover, despite the importance of LBD components such as input, output and evaluation, the existing reviews have not paid attention to critically analyse the state-of-the-art and the limitations of these components. To overcome these two limitations, in this review, we provide a sequential walk-through of the entire LBD workflow by providing new insights on the LBD components such as input, output and evaluation.

Furthermore, we have also observed that most of the existing reviews have restricted their scope only to medical-related LBD studies. Consequently, these reviews are lacking the discussions of LBD in the non-medical and domain-independent settings. To cater this issue, we examine the LBD literature in both medical and non-medical domains in this review.

More specifically, our contributions are; 1) being the first systematic literature review that covers every component of the LBD workflow, 2) shedding light on components in the LBD workflow (such as input, output and evaluation) that have not been critically analysed or categorised by the existing reviews, 3) answering each of our research questions using novel, up-to-date and comprehensive categorisations compared to the existing reviews, and 4) critiquing the LBD literature independently from domain, without restricting to only medical-related LBD studies.

xi Methods

The overall process of this systematic review adheres the steps of systematic literature reviews in computer science (Weidt & Silva 2016), as illustrated in Figure 2.4.

xi.1 Article Retrieval Process

We used six keywords and six databases to retrieve the articles for this review. Each keyword is searched in the title, abstract or keywords depending on the search options given by the databases. To ensure that we have not missed any useful articles, we also added the full reference list of a latest LBD review (Henry & McInnes 2017). The article retrieval process (with relevant statistics) is summarised in Table 2.9.

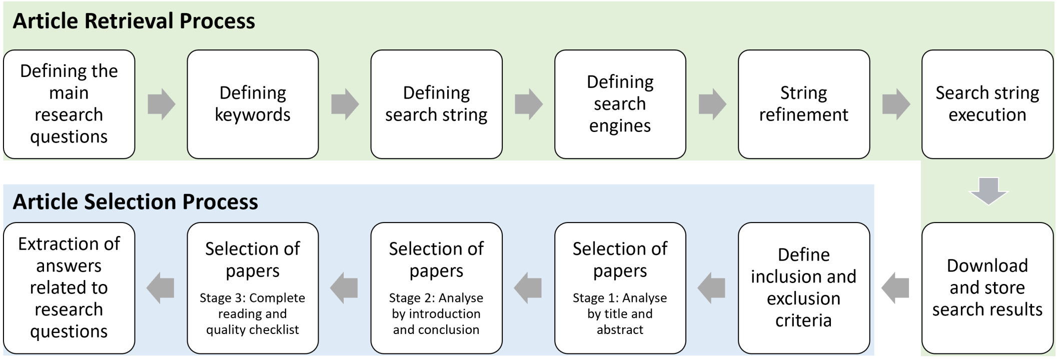

Article Retrieval Process
Defining the main research questions
Defining keywords
Defining search string
Defining search engines
String refinement
Search string execution
Article Selection Process
Extraction of answers related to research questions
Selection of papers
Stage 3: Complete reading and quality checklist
Selection of papers
Stage 2: Analyse by introduction and conclusion
Selection of papers
Stage 1: Analyse by title and abstract
Define inclusion and exclusion criteria
Download and store search results

Figure 2.4: Process of the systematic literature review

Table 2.9: Statistics of the article retrieval process

Keyword	Web of Science	Scopus	PubMed	ACM Digital Library	IEEE Xplore	Springer -Link	Total Count
Query 1[a]	161	68	75	15	15	8	342
Query 2[b]	14	0	4	1	2	1	22
Query 3[c]	14	0	0	0	0	1	15
References from Henry & McInnes (2017)							96
Total Article count							**475**

[a] "literature based discovery" OR "literature based discoveries"
[b] "literature based knowledge discovery" OR "literature based knowledge discoveries"
[c] "literature related discovery" OR "literature related discoveries"

xi.2 Article Selection Process

We only included journals and conference proceedings that are in the English language in our analysis. We excluded other types of articles such as reviews, books, book chapters, papers reporting lessons learned, keynotes and editorials. We also eliminated the papers that provide the theoretical perspective of LBD as our research questions are focused to assess the LBD discipline in terms of computational techniques. We also excluded articles of page count 4 or below as such articles mainly contain research-in-progress. The entire article selection of this review was performed in three stages (Weidt & Silva 2016); *Stage 1:* analyse only title and abstract, *Stage 2:* analyse introduction and conclusion, and *Stage 3:* read complete article and quality checklist. In total, we obtained 176 papers for this review (listed in `https://tinyurl.com/selected-LBD-articles`).

xii Review Overview

In this review we seek answers for 8 research questions that are grouped into four categories by considering the workflow of LBD process, as illustrated in Figure 2.5.

1. **Input Component**

   What input types are used in the knowledge discovery process of the LBD workflow?
   What data sources are used in LBD research to extract these identified input types?

2. **Process Component**

   What filtering techniques are used in the LBD process?
   What ranking/thresholding mechanisms are used in the LBD process?

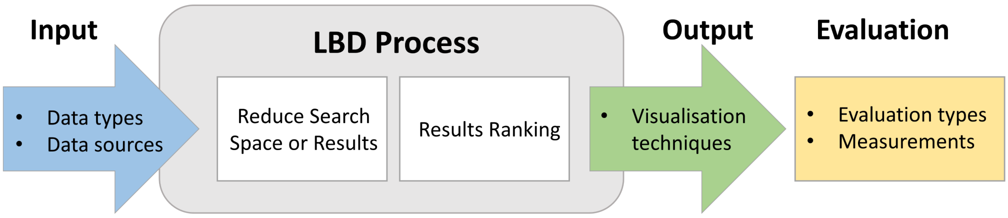


FIGURE 2.5: Main components of the LBD workflow

What domain-independent and domain-dependent resources are utilised in LBD research?

3. **Output Component**

    What visualisation techniques are used to display results in LBD research?

4. **Evaluation Component**

    What are the LBD evaluation types and how suitable are they to non-medical domains?

    What quantitative measurements are used to assess the effectiveness of the results?

To increase the readability of our review, we have cited a limited number of literature for each research question. However, a complete list of references that supports the proposed categorisations and conclusions of the research questions are listed in `https://tinyurl.com/full-references`.

xiii Input Component

This section analyses the input component of the LBD workflow to get an overview of the data structures and databases used in the literature.

xiii.1 What input types are used in the knowledge discovery process of the LBD workflow?

The LBD studies make use of different data types as their input of the knowledge discovery process. The selection of the most suitable input type is one of the key design decisions, as they should represent the most important entities and relationships of an article to perform an efficient knowledge discovery. The *input types* used in the LBD literature can be categorised as follows.

Title only: Some LBD studies (Swanson & Smalheiser 1997, Cherdioui & Boubekeur 2013) have only considered the *article title* as the input of the knowledge discovery process. This input type selection might have influenced by Swanson's initial work as he only utilised the titles to uncover the hidden associations in his discoveries such as *Raynaud's disease*↔*fish oil*. Even though the article title contains limited information, Sebastian et al. (2017*b*) have reported that using only titles for knowledge discovery tend to produce better results compared to analysing abstracts.

Title and Abstract: The most common input type selection in the literature is using both *title and abstract* (Lever et al. 2018, Sebastian et al. 2017*b*). The main reasons for this selection over full-text analysis could be; 1) *Reducing noise:* Typically, the title and abstract include the most important concepts that best describe the study than considering the full-text, 2) *Data retrieval constraints:* Most APIs of the literature databases only support metadata retrieval, and 3) *Reducing computational complexity*: As the content of title and abstract is restricted, the time and space complexities are reduced compared to full-text analysis.

Full-text: Few studies (Lever et al. 2018, Vicente-Gomila 2014) have considered the entire content of articles as their input type. It has been reported that using full-text yields better results over title and abstract analysis. (Seki & Mostafa 2009). However, it is also important to pay attention as to what sections of the full-text need to be analysed to obtain better results. For instance, does analysing only the methodological-related sections of an article produce better results than analysing the entire article? Such detailed full-text analyses have not been preformed in the LBD literature yet.

Selected articles only: While most of the studies have used data retrieved from literature database search engines (e.g., MEDLINE) for analysis, Cameron et al. (2015) have only considered the reference lists of Swanson's LBD publications. Considering only the 65 articles cited in Swanson's *Raynaud's disease↔fish oil* LBD paper (Swanson 1986) as the input of the knowledge discovery process can be taken as an example. However, since these reference lists are manually analysed and selected, whether this input type selection reflects the complexity of the real-world data is doubtful.

Entire literature database: Several research studies (Lever et al. 2018, Yang et al. 2017) have considered the entire literature database as their LBD input. That is, they have not limited to articles retrieved for a given query (e.g., subset of the articles retrieved for the query "fish oil"). Since the primary focus of LBD research is in the medical domain, the literature database that has been mainly considered for analysis is *MEDLINE*. Additionally, other sources such as SemMedDB (Cohen, Widdows, Stephan, Zinner, Kim, Rindflesch & Davies 2014) and PubMed Central Open Access Subset articles (Lever et al. 2018) have also been used as the LBD input.

Keywords: Some research approaches have employed the keywords of the articles as the input type (Pusala et al. 2017, Hu et al. 2010). The mostly utilised keyword type is *Medical Subject Headings (MeSH)*, which are associated with MEDLINE records. It is considered that MeSH descriptors are accurate and medically relevant as National Library of Medicine (NLM) employs trained indexers to assign them to the MEDLINE articles. Therefore, it is considered as a reliable source of representing the content of an article.

Other metadata: Few studies have analysed other metadata of the research articles such as author details (Sebastian et al. 2017*b*), publisher details (Sebastian et al. 2015) and reference details (Kostoff, Block, Stump & Johnson 2008) to glean additional cues for the possible links in the knowledge discovery process. The results of these studies prove that the use of such metadata enhances the predictability of implicit knowledge associations (Sebastian et al. 2017*b*).

Other traditional input types: While majority of the LBD studies have focused only on analysing the research papers, some approaches have been conducted using other traditional input types, such as patents (Vicente-Gomila 2014, Maciel et al. 2011), TREC MedTrack collection of clinical patient records (Symonds, Bruza & Sitbon 2014) and case reports (Smalheiser et al. 2015), as their input to the LBD process.

Non-traditional input types: Few research studies have attempted to perform the LBD process using non-traditional input types, such as Tweets (Bhattacharya & Srinivasan 2012), Food and Drug Administration (FDA) drug labels (Bisgin et al. 2011), Popular Medical Literature (PML) news articles (Maclean & Seltzer 2011), web content (Gordon et al. 2002), crime incident reports (Schroeder et al. 2007) and commission reports (Jha & Jin 2016*a*). Their results have proved the suitability of the LBD discovery setting in a non-traditional context to elicit hidden links.

The *data unit of analysis* denotes the types of data extracted from the above-discussed input types to represent knowledge associations. Since most LBD research is performed in medicine, the most common term representations are *UMLS* and *MeSH* (Lever et al. 2018, Preiss & Stevenson 2017). Apart from these two medical resources, other medical databases such as *Entrez Gene* (Kim et al. 2016), *HUGO* (Petric et al. 2014), *LocusLink* (Hristovski et al. 2005), *OMIM* (Hristovski et al. 2003) and *PharmGKB* (Kim & Park 2016) have also been used to extract data units. LBD studies in other domains mainly consider *word or word phrases (n-grams)* as their term representation (Qi & Ohsawa 2016) that have been extracted using techniques such as Part-Of-Speech (POS) tag patterns.

xiii.2 What data sources are used in LBD research to extract these identified input types?

MEDLINE/PubMed is extensively being used as the main data source of the LBD literature (Lever et al. 2018). Additionally, other data sources such as *PubMed Central (PMC) Open Access* (Ding et al. 2013), *Science Direct* (Vicente-Gomila 2014), *Web of Science* (Sebastian et al. 2015), *IEEE Xplore Digital Library* (Qi & Ohsawa 2016), *Engineering Village* (Kibwami & Tutesigensi 2014), *ProQuest* (Kibwami & Tutesigensi 2014), *EBSCO Host* (Kibwami & Tutesigensi 2014) and *INSPEC* (Ye et al. 2010) have also been employed by several other LBD approaches to retrieve the articles for analysis.

The patent-based LBD studies (Vicente-Gomila 2014) have considered patent databases such as *Thomson Innovation*, *United State Patent and Trade Mark Office (USPTO)* and *MAtrixware REsearch Collection (MAREC) patent document collection* to retrieve the data. Other conventional data sources include *clinical datasets* (Dong et al. 2014), *Gene Expression Omnibus (GEO) database* (Hristovski et al. 2010), *ArrayExpress (AE) database* (Maver et al. 2013), *Manually Annotated Target and Drug Online Resource (MATADOR)* (Crichton et al. 2018), *Biological General Repository for Interaction Datasets (BioGRID)* (Crichton et al. 2018), *PubTator* (Crichton et al. 2018), *Online Mendelian Inheritance in Man (OMIM)* (Cohen, Whitfield, Schvaneveldt, Mukund & Rindflesch 2010) and *TREC* (Symonds, Bruza & Sitbon 2014).

Few non-English data sources such as *Chinese Social Sciences Citation Index* (Su & Zhou 2009), *China Biology Medicine disks* (Qian et al. 2012), *Chinese Medicine Library and Information System* (Yao et al. 2008), *Traditional Chinese Medicine Database* (Gao, Wang, Tao, Liu, Li, Yu, Yu, Tian & Zhang 2015) and *Chinese Journal Full-text database* (Yao et al. 2008) have also been utilised in the LBD workflow.

The studies that have attempted to perform LBD in a non-traditional setting have extracted data from a variety of sources such as *Twitter* (Bhattacharya & Srinivasan 2012), *DailyMed: FDA drug labels* (Bisgin et al. 2011), *Google news* (Maclean & Seltzer 2011) and *World Wide Web (WWW)* (Gordon et al. 2002).

xiv Process Component

This section outlines the two major elements of the *process* component in the LBD workflow; *filtering techniques* and *ranking/thresholding techniques*. Moreover, this section also discusses the *resources* utilised in the LBD workflow.

xiv.1 What filtering techniques are used in the LBD process?

It is vital to provide a *concise output* to the user that is easily interpretable by only including the most promising knowledge associations. To achieve this, the search space of the knowledge discovery process should be reduced by eliminating spurious, general, uninteresting or invalid terms/concepts. Different types of filtering techniques used in the literature are summarised in Figure 2.6 (a).

Stop word Removal: Stop words typically denote non-topic general English terms. However, it could also include general terms used in a domain. For example, terms such as *'drug'* and *'treatment'* can be considered as general terms in the medical domain. Using stop words to remove uninformative terms is a popular filtering technique used (Lever et al. 2018, Preiss & Stevenson 2017, Sebastian et al. 2017*b*). Stop word lists could be either *manually created*, *obtained from other resources* or *automatically generated*. 1) *Manually created:* A popular example of this category is the stop word list created for the *Arrowsmith* project (Smalheiser 2005) that has nearly 9500 terms (by 2006) (Preiss & Stevenson 2017). However, manual development of stop words is costly and time-consuming. Moreover, since these stop word lists are highly domain-dependent, their applicability is also limited. 2) *Obtained from other resources:* Other resources used to obtain stop word lists include *NLTK toolkit* (Lever et al. 2018), *Corpus of Contemporary American English* (Lever et al. 2018) and *Nvivo* (Kibwami & Tutesigensi 2014). 3) *Automatically generated:* Some studies (Preiss & Stevenson 2017, Hu et al. 2010) have automatically created their stop word lists by employing different techniques. The most common way is eliminating terms that appear above a user-defined threshold (Pratt & Yetisgen-Yildiz 2003). In addition to such threshold-based removal, Xun et al. (2017) have followed *law of conformity* to remove general terms by analysing the temporal change of terms, and Jha et al. (2016*b*) have considered outliers of the box-plot as the general terms removal mechanism.

Semantic Category Filter: This technique typically utilises the *semantic type or group* information provided by *UMLS* (Lever et al. 2018, Vlietstra et al. 2017). UMLS currently provides 127 semantic types[8] and each medical concept is classified to one or more of these semantic types based on the relevance. Each semantic type is further classified into one or more of 15 UMLS semantic groups[9]. For example, *panic disorder* belongs to the semantic type *'mental or behavioural dysfunction'* and *migraine* belongs to the semantic type *'disease and syndrome'*. Both of these semantic types belong to the semantic group *'disorders'* (Yetisgen-Yildiz & Pratt 2009). This filtering technique involves imposing selected semantic type or group to restrict the linking and target concepts of the knowledge discovery process. However, selecting the most suitable semantic type or group is challenging as it varies according to the problem. If a too granular semantic category is selected, it may also remove valid associations, and if a too broader semantic category is picked, it may not filter out all meaningless associations.

Relation/predicate Type Filter: This filtering technique mostly consider the predications assigned using *SemRep* (Cameron et al. 2015, Rastegar-Mojarad et al. 2015). The typical procedure is to restrict the search space by eliminating uninteresting predicate types. For example, Cohen et al. (2010) have removed *'PROCESS_OF'* predication in their LBD process as it is less informative. Other types of predicate filtering techniques are; 1) removal of negated relations (Rastegar-Mojarad et al. 2016), 2) considering the directionality of the predicate (Baek et al. 2017) and 3) restricting the semantic type or group of the subject and object in predications (i.e., subject-relation-object triples) (Hristovski et al. 2010).

Hierarchical Filter: This technique utilises the hierarchical information such as *levels* and *relationships* of terms to filter out uninformative associations (Shang et al. 2014). The *levels* of *UMLS/MeSH hierarchy* are typically examined to remove broader terms. For example, Qian et al. (2012) have eliminated terms in the first and second level of *MeSH tree* to remove less useful, broad associations. Another approach is to analyse the *hierarchical relationships* of the concepts to eliminate terms that are too close to the starting term. For instance, Pratt & Yetisgen-Yildiz (2006) have eliminated terms in the *UMLS hierarchy* such as children, siblings, parents and grandparents as they have observed that these terms are closely related to the starting term; thus, they do not form any interesting association.

Synonym Mapping: Mapping synonyms by grouping exactly or nearly equal terms of a given term is another technique used to reduce the results (Lever et al. 2018, Baek et al. 2017). To facilitate this, resources such as *UMLS* (Preiss et al. 2015), *MeSH* (Van der

[8] https://semanticnetwork.nlm.nih.gov/SemanticNetworkArchive.html

[9] https://semanticnetwork.nlm.nih.gov/download/SemGroups.txt

Eijk et al. 2004), *Entrez gene database* (Liang et al. 2013) and *HUGO* (Özgür et al. 2011) have been utilised.

POS Tag-based Filter: Several studies have utilised POS tags to restrict the search space by limiting the terms to nouns (Qi & Ohsawa 2016), nominal phrases (Ittipanuvat et al. 2014) or verbs (Kim et al. 2016). For example, Qi & Ohsawa (2016) have only extracted *nouns* as unigrams.

Template-based Restriction: Several studies (Maver et al. 2013, Cohen et al. 2012) have reduced their search space by only extracting the associations that adhere to some specified *rules/templates*. For example, two forms of *discovery patterns* were defined by Hristovski et al. (2006) to restrict the detected associations that are in accordance with the templates of the two defined patterns.

Time-based Filter: Smalheiser (2005) have considered the time factor of the associations to reduce the search space of results. More specifically, given a user-defined year, only the associations that appear first time after the year (or before) have been considered as a filter. In addition, monitoring the temporal behaviours of words (Xun et al. 2017) have also been used to remove unnecessary terms.

Common Base Form: Deriving a common base form of terms is another technique used in the literature to reduce the vocabulary space. To facilitate this, the two popular techniques, *stemming* (Sebastian et al. 2015) and *lemmatisation* (Song et al. 2015) have been used in the LBD literature.

Article Retrieval Filter: Several studies (Cherdioui & Boubekeur 2013, Ittipanuvat et al. 2014) have attempted to limit the number of articles that need to be analysed through the LBD process with the intention of reducing their search space. For instance, Petric et al. (2012) have only considered the *outlier* documents for analysis without analysing all the documents derived from a search query.

Sentence Filter: Some studies (Hossain et al. 2012, Özgür et al. 2010) have only picked specific sentences from texts to analyse. For example, Özgür et al. (2010) have only picked sentences from abstracts that describe gene interactions for their analysis. For a sentence to qualify as a potential interaction sentence, the authors have followed a rule-based mechanism. Moreover, Hossain et al. (2012) have employed machine learning techniques to select sentences by training a Naïve Bayes classifier to differentiate *context* and *results* sentences in abstracts.

Network-based Filter: The network-based LBD approaches have utilised different techniques to reduce the size of their network. For example, Cairelli et al. (2015) have filtered their network by setting degree centrality and edge-occurrence frequency thresholds. Furthermore, Kastrin et al. (2014*b*) have performed Pearson's Chi-Square test to detect whether a particular connection occurs more often by chance. Ittipanuvat et al. (2014) have removed nodes that are not connected with any node in Largest Connected Components (LCC) of their knowledge graph.

Term Restrictions: Some studies have restricted terms in *word-level* and *character-level* to reduce the vocabulary space. Removal of unigrams from the analysis can be considered as an example for word-level restriction (Thaicharoen et al. 2009, Gordon et al. 2002). The LBD studies (Roy et al. 2017, Kibwami & Tutesigensi 2014) that have removed terms less than three characters in their LBD process can be considered as an example for character-level restrictions. However, since this filter does not consider semantic aspects of the terms into consideration, valuable short terms will be removed from the vocabulary.

Cohesion-based filter: Given two linking terms that are most similar, Smalheiser (2005) hypothesises that the term with a more narrow focus is the most useful. Hence, this filter calculates a *cohesion score* to select most granular-level terms as the results.

Expert/user-based filtering: Expert/user-based filtering (Gubiani et al. 2017, Preiss & Stevenson 2017) involves the decision of an expert/user to remove uninteresting associations. For example, most of the *semantic category filter* requires user-defined semantic

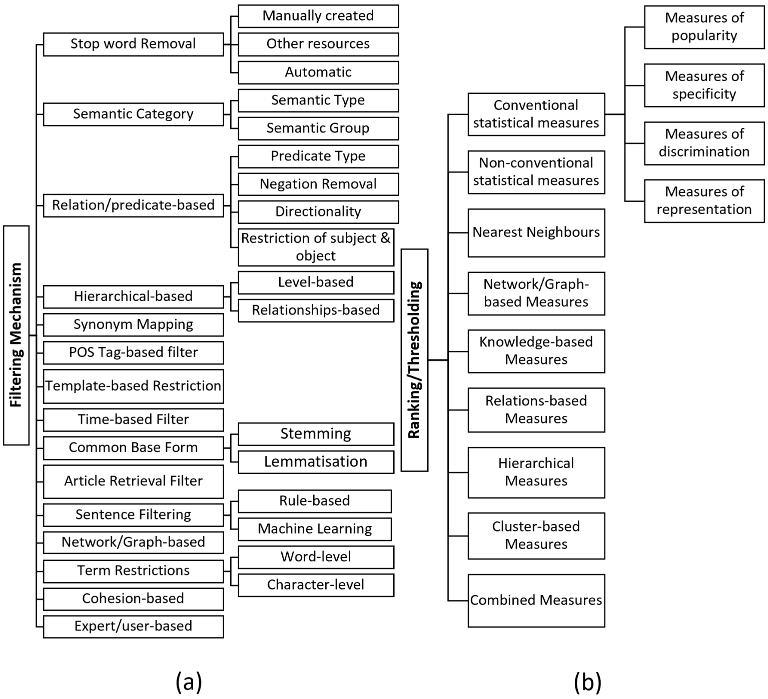


FIGURE 2.6: (a) Filtering techniques and (b) Ranking/thresholding techniques

types/groups to perform the filtering. As described in *'Semantic Category Filter'*, this selection is crucial as a more restrictive semantic category would risk at losing valid and informative associations, whereas less restrictive semantic category would result in a noisy output. As a result, the success of these approaches greatly depend on the experience and prior knowledge of the user.

xiv.2 What ranking/thresholding mechanisms are used in the LBD process?

Term ranking/thresholding is an important component of the LBD process as it should downweight or remove noisy associations, and upweight or retain the interesting and significant knowledge associations when ordering the terms. More specifically, these measures are used in two ways. 1) *Thresholding:* prune away uninteresting associations during the filtering process (e.g., setting a threshold to remove general terms), and 2) *Ranking:* rank the selected set of associations based on their significance (e.g., rank the most significant terms in the top of LBD output). Outlined below are the ranking mechanisms used in the discipline (see Figure 2.6 (b)).

Considering *conventional statistical measures* to rank/threshold terms is common in the literature. These measures can be broadly divided into four categories (Aizawa 2003) based on how they are mathematically defined; 1) *Measures of popularity:* these measures denote the frequencies of terms or probability of occurrences (e.g., concept frequency), 2) *Measures of specificity:* this category denotes the entropy or the amount of information of terms (e.g., mutual information), 3) *Measures of discrimination:* how terms are contributing to the performance of a given discrimination function is represented through these measures (e.g., information gain), and 4) *Measures of representation:* these measures denote the usefulness of terms in representing the document that they appear (e.g., TF-IDF).

Examples for conventional statistical measures used in LBD studies are; *token frequency* (Gordon & Lindsay 1996), *average token frequency* (Ittipanuvat et al. 2014), *relative token frequency* (Lindsay & Gordon 1999), *document/record frequency* (Gordon & Lindsay

1996), *average document frequency* (Ittipanuvat et al. 2014), *relative document frequency* (Thaicharoen et al. 2009), *TF-IDF* (Maciel et al. 2011), *mutual information* (Loglisci & Ceci 2011), *z-score* (Yetisgen-Yildiz & Pratt 2006), *information flow* (Bruza et al. 2006), *information gain* (Pusala et al. 2017), *odds ratio* (Bruza et al. 2006), *log likelihood* (Bruza et al. 2006), *support* (Hristovski et al. 2005), *confidence* (Hristovski et al. 2003), *F-value of support and confidence* (Hu et al. 2010), *chi-square* (Jha & Jin 2016*b*), *kulczynski* (Jha & Jin 2016*a*), *cosine* (Baek et al. 2017), *equivalence index* (Stegmann & Grohmann 2003), *coherence* (Pusala et al. 2017), *conviction* (Pusala et al. 2017), *klosgen* (Pusala et al. 2017), *least contradiction* (Pusala et al. 2017), *linear-correlation* (Pusala et al. 2017), *loevinger* (Pusala et al. 2017), *odd multiplier* (Pusala et al. 2017), *piatetsky-shapiro* (Pusala et al. 2017), *sebag-schoenauer* (Pusala et al. 2017), *zhang* (Pusala et al. 2017), *Jaccard index* (Yang et al. 2017), *dice coefficient* (Yang et al. 2017) and *conditional probability* (Seki & Mostafa 2009).

Additionally, *non-conventional statistical measures* such as *Average Minimum Weight (AMW)* (Yetisgen-Yildiz & Pratt 2009), *Linking Term Count with AMW (LTC-AMW)* (Yetisgen-Yildiz & Pratt 2009), *Averaged Mutual Information Measure (AMIM)* (Wren 2004) and *Minimum Mutual Information Measure (MMIM)* (Wren 2004) have also been proposed in the discipline to rank the potential associations. In comparison with *AMW* and *Literature Cohesiveness*, Yetisgen-Yildiz & Pratt (2009) have reported that they gained improved performance with *LTC-AMW* measure (Swanson et al. 2006). Other types of ranking and thresholding categories used in the LBD literature are summarised below.

Nearest Neighbours: In this category, the score of an association is decided by analysing its nearest neighbours. Such analyses are typically performed in distributional semantic models by employing measures such as *cosine* (Gopalakrishnan et al. 2017), *Euclidian distance* (Van der Eijk et al. 2004) and *information flow* (Bruza et al. 2006).

Network/Graph-based Measures: Network/graph-based measures analyse node-level and edge-level attributes to score an associations. Examples of measures that represent this category include *degree centrality* (Goodwin et al. 2012), *eigenvector centrality* (Özgür et al. 2010), *closeness centrality* (Özgür et al. 2011), *betweenness centrality* (Özgür et al. 2010), *common neighbours* (Kastrin et al. 2014*b*), *Jaccard index* (Kastrin et al. 2014*b*), *preferential attachment* (Kastrin et al. 2014*b*), *personalised PageRank* (Petric et al. 2014), *personalised diffusion ranking* (Petric et al. 2014) and *spreading activation* (Goodwin et al. 2012).

Knowledge-based Measures: This category denotes the scoring measures such as *MeSH-based literature cohesiveness* (Swanson et al. 2006), *semantic type co-occurrence* (Jha & Jin 2016*b*), *chemDB atomic count* (Ijaz et al. 2009) and *chemDB XLogP* (Ijaz et al. 2009) that involve the knowledge from structured resources to rank an association. The advantage of these measures is that they entangle semantic aspects into consideration to decide the potentiality of an association.

Relations-based Measures: Relations/predicates-based measures (a sub-class of *knowledge-based measures*) analyse the relations extracted from resources such as SemRep to rank/threshold associations. Scoring measures such as *semantic relations frequency* (Hristovski et al. 2010), *predicate independence* (Rastegar-Mojarad et al. 2015), *predicate interdependence* (Rastegar-Mojarad et al. 2015), *edge frequency-based weight* (Kim et al. 2016), *edge traversal probability* (Vlietstra et al. 2017), *relationship traversal probability* (Vlietstra et al. 2017), *source traversal probability* (Jha & Jin 2016*b*) and *impact factor* (Huang et al. 2016) are examples of this category.

Hierarchical Measures: This category is another sub-class of *knowledge-based measures* that utilise hierarchical information of taxonomies such as *UMLS* and *MeSH* to derive the rankings. *Child-to-parent and parent-to-child predications* (Seki & Mostafa 2009), and *MeSH tree code depth* (Gopalakrishnan et al. 2017) can be considered as examples.

Cluster-based Measures: In this category, cluster similarities are measured using techniques such as *intra-cluster similarity* (Cameron et al. 2015), *Jaccard index* (Ittipanuvat et al. 2014), *inclusion index* (Ittipanuvat et al. 2014), *dice coefficient* (Ittipanuvat et al. 2014), *cosine* (Ittipanuvat et al. 2014), *cosine similarity of tf-idf* (Ittipanuvat et al. 2014) and *cosine similarity of tf-lidf* (Ittipanuvat et al. 2014) to derive the ranking scores of associations.

Combined Measures: The idea of combined measures is to incorporate multiple characteristics of an association to decide its potential ranking. For example, Torvik & Smalheiser (2007) have utilised machine learning techniques to combine seven characteristics of an association (such as *absolute and relative term frequencies*, *cohesion*, *recency*, etc.) to obtain the final ranking score. Song et al. (2015) have also proposed a combined ranking measure by considering an average of three *semantic similarity measures*, and *SemRep score*. The characteristics that have been considered in the study of Ijaz et al. (2009) include *UMLS semantic type*, *structural similarity*, *chemDB atomic count* and *chemDB XLogP*. Similarly, Gopalakrishnan et al. (2017) have also introduced a combined ranking measure by using global (*node centrality* and *MeSH tree code depth*) and local (*semantic co-occurrence* and *betweenness centrality*) measures. Overall, combined ranking measures are more flexible as they rely on multiple characteristics to prioritise the derived associations.

xiv.3 What domain-independent and domain-dependent resources are utilised in LBD research?

xiv.3.1 Domain-Dependent Resources

Since the majority of LBD research are in medicine, we refer *medical resources* as domain-dependent resources. These resources are further categorised as; 1) *Resources that provide background domain knowledge*, and 2) *Resources that are used in content analysis.*

Resources to acquire background domain knowledge: The main purposes of extracting the domain knowledge are; 1) *input data preparation* (e.g., concept extraction), 2) *filtering the noisy, uninteresting or unrelated associations* (e.g., semantic type filtering), 3) *prepare a ranking mechanism* (e.g., hierarchical ranking), 4) *evaluate the results* (e.g., compare results with curated databases), and 5) *training data preparation.* The popular domain-dependent resources used in the discipline are;

- UMLS: Lever et al. (2018), Vlietstra et al. (2017), Preiss & Stevenson (2017)
- MeSH: Baek et al. (2017), Xun et al. (2017), Pusala et al. (2017)
- SemMedDB/Semantic MEDLINE: Vlietstra et al. (2017), Cairelli et al. (2015)
- Gene Ontology: Baek et al. (2017), Huang et al. (2016), Kim et al. (2016)
- Entrez Gene Database: Baek et al. (2017), Liang et al. (2013), Kwofie et al. (2011)
- Kyoto Encyclopedia of Genes and Genomes (KEGG): Kwofie et al. (2011)
- HGNC/HUGO: Petric et al. (2014), Ding et al. (2013), Maciel et al. (2011)
- UNIPROT: Baek et al. (2017), Vlietstra et al. (2017), Swiss-Prot Jelier et al. (2008)
- Therapeutic Target Database (TTD): Yang et al. (2017), Maciel et al. (2011)
- LocusLink: Smalheiser (2005), Hristovski et al. (2003)
- Online Mendelian Inheritance in Man (OMIM) Hristovski et al. (2003), Wren et al. (2004)
- Drug Bank: Vlietstra et al. (2017), Maciel et al. (2011), Ding et al. (2013)
- Comparative Toxicogenomics Database (CTD): Vlietstra et al. (2017), Yang et al. (2017)
- BioGRID: Huang et al. (2016), Crichton et al. (2018)

- Gene2pubmed: Cheung et al. (2012*a*), Roy et al. (2017)
- Drugs.com: Maciel et al. (2011), Banerjee et al. (2014)
- SIDER Side Effect Resource: Vlietstra et al. (2017), Shang et al. (2014)

Additionally, other medical resources such as *Medical Dictionary for Regulatory Activities (MedDRA)* (Bisgin et al. 2011), *Reactome Pathway Database* (Kwofie et al. 2011), *Orphanet* (Baek et al. 2017), *Human Metabolome Database (HMDB)* (Baek et al. 2017), *Lipid Maps* (Baek et al. 2017), *MassBank* (Baek et al. 2017), *DailyMed* (Vlietstra et al. 2017), *miRBase* (Huang et al. 2016), *miRGate* (Huang et al. 2016), *Transcriptional Regulatory Relationships Unraveled by Sentence-based Text mining (TRRUST)* (Huang et al. 2016), *PAZAR* (Huang et al. 2016), *Biomedical Knowledge Repository (BKR)* (Cameron et al. 2015), *MEDI* (Shang et al. 2014), *Tanabe-Wilbur list* (Smalheiser 2005), *ChemDB* (Ijaz et al. 2009), *BioVerb* (Kim et al. 2016), *AIMED* (Özgür et al. 2010), *CB* (Özgür et al. 2010), *STRING* (Petric et al. 2014), *ToppGene* (Petric et al. 2014), *Endeavour* (Petric et al. 2014), *MIPS* (Liang et al. 2013), *Proteomics Standards Initiative Molecular Interactions (PSI-MI)* (Song et al. 2015), *Cell Line Knowledge Base (CLKB)* (Song et al. 2015), *Observational Medical Outcomes Partnership (OMOP)* (Mower et al. 2016), *METADOR* (Crichton et al. 2018), *Animal Transcription Factor Database (AnimalTFDB)* (Roy et al. 2017), *RxNorm* (Malec et al. 2016), *Vaccine Ontology (VO)* (Özgür et al. 2011), *Gene Reference Into Function (GeneRIF)* (Cheung et al. 2012*a*), *Homologene* (Jelier et al. 2008), *Pharmacogenomics Knowledge base (PharmGKB)* (Kim & Park 2016), *Chinese Medical Terminology* (Qian et al. 2012), *Food and Drug Administration approved drug names* (Wren 2004) and *Rush University Medical Center's health encyclopedia* (Banerjee et al. 2014) have also been employed in the LBD workflow.

Our analysis reveals that *UMLS* and *MeSH* are most extensively used as the domain-dependent resources in the literature. The databases such as *SemMedDB/Semantic MEDLINE*, *Gene Ontology*, *Entrez Gene Database* and *HUGO/HGNC* are also popular among other resources.

Resources for content analysis: The following resources have been used in the LBD systems to process and analyse contents.

- MetaMap (medical concept extraction): Preiss & Stevenson (2017, 2016), Cairelli et al. (2015)
- SemRep (semantic predications extraction): Vlietstra et al. (2017), Preiss et al. (2015)
- Genia Tagger (biological NER): Lever et al. (2018), Özgür et al. (2010)
- ABNER (biological NER): Liang et al. (2013)
- Peregrine software (biological NER): Jelier et al. (2008)
- DAVID tool (gene annotation enrichment analysis): Maver et al. (2013), Özgür et al. (2010)
- RankProd Package (meta analysis): Maver et al. (2013)
- BioTeKS Text Analysis Engine (text annotation): Berardi et al. (2005)
- PubTator (PubMed citations annotation): Crichton et al. (2018)
- MedLEE (structure and encode clinical reports): Malec et al. (2016)
- BioMedLEE (semantic predications extraction): Hristovski et al. (2006)
- EpiphaNet (interactive visual representation): Malec et al. (2016)
- SciMiner (literature mining and functional enrichment analysis): Hur et al. (2010)
- Biovista (drug repurposing, systems literature analysis environment): Persidis et al. (2004)

Among the content analysis tools, we observed that *MetaMap* and *SemRep* are the most popular selections. *MetaMap* is a tool that recognises *UMLS* concepts in texts, whereas *SemRep* is used to extract semantic predications from texts. The predications in SemRep are formal representations of text content that comprises of *subject-predicate-object* triples.

xiv.3.2 Domain-Independent Resources

In this section, we summarise the resources that can be used in a cross-domain LBD setting. For Named Entity Recognition (NER) resources such as *GATE* (Loglisci & Ceci 2011), *PKDE4J* (Baek et al. 2017), *Open Calais* (Jha & Jin 2016*a*), *Sementax* (Jha & Jin 2016*a*) and *Lingpipe* (Hossain et al. 2012) have been employed in the LBD literature.

Other text analytics resources include *NLTK*: to identify noun phrases (Sebastian et al. 2017*b*) and stop words (Lever et al. 2018), *ReVerb:* to extract relations (Preiss et al. 2015), *Stanford parser:* for dependency tree parsing (Sang et al. 2015) and extract relations (Preiss et al. 2015), *Stanford CoreNLP:* for sentence boundary detection, POS tagging and lemmatisation (Song et al. 2015), *WordNet:* for word sense disambiguation (Sebastian et al. 2017*b*), *RacerPro:* for logical and rule-based reasoning (Guo & Kraines 2009*a*), *Link Grammar Parser:* for sentence parsing (Ijaz et al. 2009), *Vantage Point:* for document clustering, auto-correction mapping and factor matrix analysis (Kostoff 2011), *Nvivo:* to extract terms, stop words, coding and matrix coding queries (Kibwami & Tutesigensi 2014), *CLUTO:* for document clustering (Kostoff, Briggs, Solka & Rushenberg 2008), *Lucene:* for information retrieval (Malec et al. 2016) and *OntoGen:* for topic ontology construction (Petrič et al. 2009).

To facilitate tasks such as network construction and visualisation, the following resources have been utilised in the literature; *Neo4j* (Vlietstra et al. 2017), *JUNG* (Kim et al. 2016), *Gephi* (Song et al. 2015), *NetworkX* (Wilkowski et al. 2011*b*) and *Large Graph Layout (LGL)* (Ittipanuvat et al. 2014).

The importance of using the aforementioned resources in LBD systems is that they support the systems' functionalities not only in medical domain, but also in a wide variety of other domains. To date, such domain-independent LBD methodologies have been rarely experimented.

xv Output Component

This section discusses the existing LBD output types, their drawbacks and the important characteristics that need to be fulfilled in terms of output visualisation to meet the objectives of the LBD discipline.

xv.1 What visualisation techniques are used to display results in LBD research?

The most commonly used output of LBD systems is a ranked list of associations (Gubiani et al. 2017, Baek et al. 2017), where the top associations reflect the most probable knowledge links. However, providing only a ranked list may not be the best way of visualising the results due to the following two reasons; 1) ranked associations are isolated in nature and do not provide an overall picture of all suggested associations, and 2) ranked associations do not reflect how they are linked with the start and/or target concepts to better understand an association. As a result, the user needs to manually analyse the ranked associations individually to get an overview of the entire results and to interpret the linkage of a proposed associations with the start and/or target concepts. This points out the importance of exploring better visualisation techniques that can reduce the manual investigations the user requires to perform. Discussed below are other visualisation techniques employed in the literature.

Group based on semantic type: In Manjal LBD system (Srinivasan 2004), the outputted terms are organised by UMLS semantic types and ranked based on their estimated potential within these semantic types.

Rank based on templates: SemBT LBD system (Hristovski et al. 2010) ranks the identified novel associations using frequency of semantic relations (relation triples) by specifying the subject and object of the relation. Ijaz et al. (2009) have ranked the detected associations based on an information model that includes substance, effects, processes, disease and body part.

Graph-based visualisations: Several studies have utilised graphs to visualise their LBD results. For instance, Kim et al. (2016) have used directed gene-gene network to clearly illustrate the discovery pathways suggested by their LBD methodology. A more advanced graph-based visualisation was proposed by Cameron et al. (2015) that outputs multiple context driven sub-graphs. Since the graph is divided into subgraphs by grouping the paths with similar context, the results can be easily interpreted by the user.

Ranking the discovery pathways: From the LBD perspective, this technique can also be viewed as an output of the *AnC* model. While *graph-based visualisations* (discussed above) display graphs as output, this technique only lists down the potential paths from the graph. Examples of this category include the study of Wilkowski et al. (2011*b*) where the graph paths with high degree centrality are shown as the output, and the study of Kim et al. (2016) that considers the shortest paths in the graph as the output.

Story chain construction: Hossain et al. (2012) have attempted to build story chains by focusing on biological entities in PubMed abstracts. Their storytelling algorithm provides new insights on LBD visualisation and can be viewed as a next step of the *ranking the discovery pathways* technique (discussed above).

Word clouds: Malec et al. (2016) have used word clouds to present their results where the font size is proportionate to term frequencies.

Matrix-like visualisation: Qi & Ohsawa (2016) have proposed a matrix-like visualisation to detect mixed topics of their experiments. Moreover, they have also performed a user-based evaluation by providing their visualisation to the users to detect and interpret mixed topics.

Using Existing Tools: Some studies have utilised existing tools such as Semantic MEDLINE (Miller et al. 2012), OntoGen (Petrič et al. 2012), EpiphaNet (Cohen et al. 2009) and Biolab Experiment Assistant (BAE) (Persidis et al. 2004) for the LBD visualisation.

Improving output visualisation is an essential component of the LBD workflow as it highly influences the user acceptance of LBD systems. However, the existing literature has a little contribution towards output visualisation. This suggests the importance of involving Human Computer Interaction (HCI) techniques in the field. Some important characteristics that should be taken into consideration when developing a visualisation technique are; 1) concise output, 2) easily interpretable, 3) less complex, 4) visually attractive and 5) assist users to gain new insights. Moreover, it is also vital to evaluate the efficiency of the visualisation techniques by performing user-based evaluations (Santos 2008). For instance, one could organise sessions for the participants to use LBD tools (Cohen, Whitfield, Schvaneveldt, Mukund & Rindflesch 2010), observe how they interact with tools and obtain their feedback. Santos (2008) suggests two types of participants for such evaluations; *target users* and *graphic designers*. The author pointed out that the target users will assist to elicit new ideas, whereas graphic designers will detect problems and provide suggestions with visual aspects. Furthermore, another interesting avenue is to involve target users with different level of expertise (i.e., expert and novice) to evaluate how users with each level of expertise interact and benefit with the LBD process (Qi & Ohsawa 2016).

xvi Evaluation Component

xvi.1 What are the LBD evaluation types and how suitable are they to non-medical domains?

Evaluating the effectiveness of the LBD results is challenging and remains to be an open issue. The main reason for this is that the LBD process detects novel knowledge that has not been publicly published anywhere and thus needs to be proven that they are useful. Moreover, there are no comprehensive gold standard datasets or consistent formal evaluation approaches in LBD (Ganiz et al. 2005). This review provides an in-depth classification of the existing evaluation techniques as summarised below.

xvi.1.1 Evidence-based Evaluation

This category of evaluation asserts if a given association is accurate by using evidence from reliable sources such as existing discoveries, literature or curated databases.

Replicating existing medical discoveries: By far, this is the most commonly used evaluation technique. It measures the capability of the LBD methodology to reproduce the popular historical discoveries (see Table 2.10). The most popular selections of discovery replication are Swanson's initial two medical discoveries; *Raynaud's disease↔Fish oil* and *Migraine↔Magnesium*. The normal procedure used for discovery replication is to only use the literature before the original paper of discovery as the input data of the LBD process and to verify if the mentioned associations detected in the original paper could be replicated. For example, if we consider Swanson's *Raynaud's disease↔Fish oil* to replicate, the literature prior to 1986 (the published year of the paper) should only be considered.

However, discovery replication may not be the most effective way of evaluating an LBD methodology due to the following reasons. 1) These existing discoveries have not developed rigorously as a gold standard (Ganiz et al. 2005). For example, in Swanson's *Raynaud's disease↔Fish oil* discovery, he only suggested three novel intermediate connections. No evidence suggest that these connections identified through his trial and error approach can be seen as the only existing novel associations that connect these two domains, 2) Only focusing on one particular discovery might result in a system that performs well for that problem, but not for other problems even within the same domain (i.e., overfitting) (Yetisgen-Yildiz & Pratt 2009). For example, Swanson & Smalheiser (1997) have replicated medical discoveries to evaluate *Arrowsmith* LBD system. The overfitting of their model is evident by the failure of it in recognising the links of *Somatomedin-C↔Arginine* (Swanson 1990*b*). As a result, it is important to accompany other evaluation techniques along with *discovery replication* to measure the true efficiency of a proposed methodology.

Time-sliced evaluation: Time-sliced method evaluates the ability of an LBD methodology to predict future co-occurrences based on a time-sliced dataset (Lever et al. 2018, Yang et al. 2017). Currently this is the most objective evaluation technique in the discipline that attempts to alleviate the following key issues (Yetisgen-Yildiz & Pratt 2009).

1) Discovery replication is limited to the associations defined in that particular discovery and merely evaluates the ability of a methodology to recreate these specific associations. As a result, the remaining associations in the LBD output are not assessed. This makes it difficult to estimate the overall performance of an LBD system. Instead, time-sliced evaluation evaluates the complete list of associations outputted from an LBD system. 2) Most LBD systems consider one or two existing medical discoveries to replicate. Hence, the true generalisability of their methodologies is not reflected. To overcome this issue, time-sliced evaluation is designed in a way it is repeatable for many starting concepts without only limiting to one or two existing medical discoveries. For example, Yetisgen-Yildiz & Pratt (2009) have considered 100 starting concepts for the evaluation of their

TABLE 2.10: Replicated discoveries in the LBD literature

Replicated Discovery	**Past Studies**
Migraine↔Magnesium	Xun et al. (2017), Preiss & Stevenson (2017), Sebastian et al. (2017*b*), Qi & Ohsawa (2016), Song et al. (2015)
Raynaud's disease↔Fish Oil	Xun et al. (2017), Preiss & Stevenson (2017), Sebastian et al. (2017*b*), Song et al. (2015), Preiss et al. (2015)
Indomethacin↔Alzheimer's	Xun et al. (2017), Preiss & Stevenson (2017), Preiss et al. (2015), Cameron et al. (2015), Sang et al. (2015)
Schizophrenia↔Calcium-Independent Phospholipase A2	Xun et al. (2017), Preiss & Stevenson (2017), Preiss et al. (2015), Cameron et al. (2015), Srinivasan (2004)
Alzheimer's↔Estrogen	Preiss & Stevenson (2017), Preiss et al. (2015), Cameron et al. (2015), Preiss (2014)
Magnesium deficiency↔Neurologic	Preiss & Stevenson (2017), Preiss et al. (2015), Preiss (2014)
Thalidomide↔Chronic Hepatitis C	Kwofie et al. (2011), Jelier et al. (2008)
Testosterone↔Sleep	Cameron et al. (2015), Goodwin et al. (2012)
Somatomedin C↔Arginine	Swanson & Smalheiser (1997), Preiss (2014)
Chlorpromazine↔Cardiac Hypertrophy	Cameron et al. (2015)
Diethylhexyl (DEHP)↔Sepsis	Cameron et al. (2015)
Sleep↔Depression	Goodwin et al. (2012)

LBD system. 3) When replicating existing medical discoveries, the required intermediate and target terms are known in advance. As a result, the parameters of a system can be tuned in a way to obtain these terms. This results in a system that performs well only for that discovery, but not in other cases. However, time-sliced evaluation is independent of prior knowledge as it does not require to know the output in advance which assists to perform an unbiased evaluation. 4) When replicating medical discoveries or in expert-based evaluation, it is difficult to compare the performance of different LBD systems. For example, if two systems claim that they could successfully replicate a particular discovery, it is hard to determine the most efficient system. Similarly, when incorporating expert decisions for evaluation, it is hard to quantify the results and compare against other LBD systems. As a result, time-sliced evaluation provides a platform to quantitatively compare the LBD outcomes with other systems.

This technique requires a *cut-off-date* to divide the dataset into two segments, namely *pre-cut-off* (data before the specified cut-off date) and *post-cut-off* (data after the cut-off date). The pre-cut-off segment is treated as the training set, where the LBD system is employed to output the potential novel associations. Afterwards, the post-cut-off segment is utilised to develop the ground truth dataset to evaluate the produced associations. The ground truth dataset is created by identifying associations present in the post-cut-off set and absent in the pre-cut-off set. More specifically, time-sliced evaluation verifies whether the identified potential associations from the LBD process have taken place in the future. Therefore, the selection of the cut-off-date is crucial because it decides the time period that turns a hypothesis into a true discovery (Yetisgen-Yildiz & Pratt 2009).

Manual literature search: Some studies have verified whether the produced associations are meaningful by manually searching the research articles that provide evidence of the existence of the specified association (Yang et al. 2017, Xun et al. 2017).

Intersection evaluation: This approach checks if the identified associations have been co-occurred with the initial concept in any of the literature databases (e.g., *Web of Science*) or other sources (e.g., *UseNet*), and remove already known associations to identify the novel associations (Gordon et al. 2002, Bhattacharya & Srinivasan 2012). Afterwards, these identified novel associations are qualitatively evaluated.

Derive reference sets from the literature: In this technique, a methodology is evaluated by using reference sets created using the past literature. For example, in the study of Vlietstra et al. (2017), they have developed the reference set from the results of a systematic literature review to compare their results. In the work of Bernstam et al. (2016), they have used curated drug-ADE associations of Patrick Ryan et al. (2013) as the reference set to facilitate comparison.

Compare results with curated databases: Cross referencing the LBD output with existing curated databases to verify the validity of results is another technique used in the LBD literature. For example, some studies (Rastegar-Mojarad et al. 2015, Cheung et al. 2012*a*) have used drug-disease interactions in Comparative Toxicogenomics Database (CTD) to validate their results. Similarly, other databases such as SIDER2 (Shang et al. 2014), GEO (Faro et al. 2011), GAD (Seki & Mostafa 2009) and StringDB (Nagarajan et al. 2015) have also been used for validation.

Compare results using other resources: In contrast to curated databases, this technique uses other reliable sources (such as websites) to validate the results. For instance, Vidal et al. (2014) have used the information published in *Mayo Clinic public website* as the ground truth to evaluate the efficacy of their ranking technique.

xvi.1.2 Comparison with Baselines

The previous LBD studies have incorporated different baseline models for comparison, as discussed below.

Comparison with existing LBD tools: Several studies have considered the output of the popular LBD tools as baselines to compare their results. The LBD tools that have been considered for results comparison are; BITOLA (Lever et al. 2018), ARROWSMITH (Loglisci & Ceci 2011), Manjal (Vidal et al. 2014), ANNI (Lever et al. 2018) and FACTA+ (Lever et al. 2018).

Comparison with previous LBD techniques: In this evaluation method, popular techniques that have already been tested by several LBD studies are considered as baselines to facilitate comparison. These include techniques such as association rule mining (e.g., Apriori (Hu et al. 2010)), distributional semantic techniques (e.g., LSI and RRI (Hu et al. 2010)), lexical statistics (e.g., TF-IDF and token frequencies (Kim et al. 2016)) and bibliographic coupling (Sebastian et al. 2015).

Comparison with previous LBD work: Several studies have recreated previous LBD methodologies as baselines to compare their results. Recreating work of Gordon et al. (1996) for comparison in (Gordon & Dumais 1998), and recreating work of Hristovski et al. (2001) for comparison in (Huang et al. 2005*a*) can be considered as examples. Some studies have only recreated subsections of the previous methodologies to evaluate the corresponding sub-section of their methodology. For instance, Rastegar-Mojarad et al. (2016) have compared their ranking method with *linking term count* mentioned in (Yetisgen-Yildiz & Pratt 2006). Others have performed a direct comparison of their results with the results of previous methodologies. For example, Qi & Ohsawa (2016) have compared their results in *Migraine*↔*Magnesium* rediscovery with five other previous work in terms of precision, recall and F-measure.

Comparison with other state-of-the-art methods: Some studies have compared their work with state-of-the-art methods in the relevant disciplines that are not necessarily tested in LBD before. For example, Crichton et al. (2018) have considered *Adamic-Adar*, *Common Neighbours* and *Jaccard Index* to compare their results as these algorithms are considered to be competitive and challenging baselines in the *link prediction* discipline.

xvi.1.3 Expert-oriented Evaluation

Expert-based evaluation: In expert-based evaluation, typically one (Gubiani et al. 2017) or two (Baek et al. 2017) domain experts inspect the LBD output to verify if the produced associations are meaningful. Alternatively, the domain expert may provide with a more open-ended evaluation (Gordon et al. 2002) by asking them to provide anticipated future associations in the domain, without actually looking at the LBD results. Afterwards, the list of potential associations provided by the expert is cross-checked against the actual LBD outcome. However, expert-based evaluation is expensive, time-consuming and suffers from subjectivity.

Qualitative analysis of selected results: A commonly used technique in the LBD evaluation is to qualitatively analyse the LBD output (typically in an ad-hoc basis) by the author(s) or domain expert(s) (Jha & Jin 2016*a*, Huang et al. 2016). Since the complete LBD result is not evaluated, it is hard to determine the true accuracy of an LBD methodology using this evaluation technique. Moreover, same as in expert-based evaluation, the analysis of results suffers from subjectivity.

xvi.1.4 User-oriented Evaluation

It is crucial to perform user-oriented evaluations to verify the use of LBD systems for real-world usage. However, such evaluations are rarely performed in the existing literature.

User-based evaluation: Evaluating user's ability to identify and formulate hypotheses from the output of the LBD process is an essential evaluation approach. However, such user-oriented evaluations are mostly neglected in the LBD literature. As defined in the study of Qi & Ohsawa (2016), criteria such as *utility* (how useful is the generated hypothesis?), *interestingness* (how interesting is the generated hypothesis?) and *feasibility* (to what extent the generated hypothesis can be realised?) can be incorporated to score these user formulated hypotheses. Such scores can be analysed to verify the extent to which LBD systems assist users to create novel scientifically meaningful research hypotheses.

User-experience evaluation: Analysing how users interact with an LBD system plays a critical role as such user behaviours provide useful insights to improve the visualisation techniques of LBD results, user-interface, and the process of knowledge discovery. However, user-experience is rarely measured in LBD research. Qi & Ohsawa (2016) have compared the performance of experts and non-experts with their matrix-like visualisation LBD process and verified that the users with no prior knowledge also benefited from their LBD process. Similarly, a user performance evaluation was conducted in the study of Cohen et al. (2010) using one domain expert and one advanced undergraduate student using a total of nearly 6.5 hours of sessions to evaluate their LBD tool, *EpiphaNet* from the users' perspective.

xvi.1.5 Proven from Experiments

Some studies have performed experiments to prove the validity of their produced hypotheses. Since most LBD methodologies are in medical domain, clinical trials are typically used to verify the derived hypotheses. However, validating all derived associations of the LBD process using laboratory experiments is infeasible. Hence, the most likely to be successful association from the top of the list is picked for validation (Baek et al. 2017). As a result, this evaluation does not assess the accuracy of the remaining associations; thus, does not reflect the overall performance of an LBD methodology.

xvi.1.6 Scalability Analysis

From query to query, the number of records that need to be analysed vary (Spangler et al. 2014). Therefore, it is important to measure the requirements in terms of time and storage for each phase in the LBD process to make the methodology more user-friendly.

Processing time analysis: Less processing time is a critical characteristic of the LBD process as the users would like to quickly obtain results for their queries. However, the time complexity is rarely measured and compared against other LBD methodologies in the literature. Few LBD studies (Hossain et al. 2012, Loglisci & Ceci 2011) have performed such processing time analyses of their algorithms.

Storage analysis: Analysing memory requirements is also important when dealing with large datasets. For instance, the study of Symond et al. (2014) have analysed the storage complexity of several distributional models. Through their analysis, they have identified that *Tensor Encoding* model is well suited for open discovery as it is efficient in storing and computing (independent of the vocabulary size).

xvi.1.7 Evaluate Ranking Technique

The algorithm used to rank the detected associations plays a vital role in an LBD methodology. It should rank the most promising associations in the top of the list by filtering the weak or false-positive associations. Therefore, the success of the LBD process greatly depends on the efficacy of the ranking algorithm.

Evaluate ranking positions: Most of the studies have evaluated the ranking positions of the LBD output to verify the effectiveness of their ranking algorithm. For instance, the LBD studies that have chosen to replicate previous medical discoveries (Gordon & Dumais 1998, Lindsay & Gordon 1999) have attempted to obtain the associations of that particular medical discovery in the top of the list. Some studies have compared their ranked list with a ranking list of previously published LBD studies to determine the superiority of their algorithms (Gordon & Dumais 1998). Moreover, in techniques such as time-sliced evaluation (Yetisgen-Yildiz & Pratt 2009), the efficiency of the ranking algorithm is measured by using information retrieval metrics (such as 11-point average interpolated precision, precision at k and mean average precision). Some studies have automatically created ground-truths using evidence from the literature to evaluate their ranking algorithms (Xun et al. 2017).

Evaluate ranking scores: Mapping the ranking scores of the detected associations with scores obtained from databases (Baek et al. 2017) or other algorithms (Pusala et al. 2017) is another evaluation technique used in the literature.

xvi.1.8 Evaluate the Quality of the Output

Evaluate the interestingness of results: Cameron et al. (2015) have used association rarity to statistically evaluate the interestingness of the LBD output. To facilitate this, they have queried MEDLINE to obtain the number of articles that contain the derived associations and divided it by the number of associations. Afterwards, an interesting score was obtained which is proportionate to the rarity score.

Evaluation of quality and coherence of stories: This evaluation metric provides a novel perspective on LBD evaluation. The quality of the produced story chains can be evaluated using dispersion coefficient, which is 1 for an ideal story (Hossain et al. 2012). This type of evaluation can be adapted when the LBD methodology outputs a chain of story path (e.g., output of the AnC model).

We also analysed the generalisability of each evaluation technique across domains. To achieve this, the previously discussed evaluation techniques are categorised into the following two groups; *Category 1:* Highly domain-dependent and only applicable to

Table 2.11: Domain-dependency of the evaluation techniques

Evaluation Technique	**Category 1**	**Category 2**
Evidence-based Evaluation:		
Replicating existing medical discoveries	✓	-
Time-sliced evaluation	-	✓
Manual literature search	-	✓
Intersection evaluation	-	✓
Derive reference sets from the literature	-	✓
Compare results with curated databases	✓	-
Compare results using other resources	✓	-
Comparison with baselines:		
Comparison with existing LBD tools	✓	-
Comparison with previous LBD techniques	-	✓
Comparison with previous LBD work	-	✓
Comparison with other state-of-the-art methods	-	✓
Expert-oriented Evaluation:		
Expert-based evaluation	-	✓
Qualitative analysis of several selected results	-	✓
User-oriented Evaluation:		
User-based evaluation	-	✓
User-experience evaluation	-	✓
Proven from Experiments:		
Clinical Tests (or relevant other experiments)	-	✓
Scalability Analysis:		
Processing time analysis	-	✓
Storage analysis	-	✓
Evaluate Ranking Technique:		
Evaluate ranking positions	-	✓
Evaluation ranking scores	-	✓
Evaluate the quality of the output:		
Evaluate the interestingness of results	-	✓
Evaluation of quality and coherence of stories	-	✓

Table 2.12: Quantitative measures used in the LBD literature

Measure	**Past Studies**
Precision	Lever et al. (2018), Yang et al. (2017), Preiss & Stevenson (2017)
Recall	Sebastian et al. (2017*b*), Jha & Jin (2016*a*), Sang et al. (2015)
F-Measure	Preiss et al. (2015), Sebastian et al. (2015), Sang et al. (2015)
Precision at k	Vlietstra et al. (2017), Shang et al. (2014), Song et al. (2015)
Recall at k	Lever et al. (2018), Vlietstra et al. (2017), Shang et al. (2014)
Average Precision	Cohen et al. (2012), Roy et al. (2017)
Mean Average Precision	Yang et al. (2017), Shang et al. (2014), Crichton et al. (2018)
Precision over time	Yetisgen-Yildiz & Pratt (2006)
Recall over time	Vlietstra et al. (2017), Yetisgen-Yildiz & Pratt (2006)
11-point average interpolated precision	Yetisgen-Yildiz & Pratt (2009)
Area Under Curve	Lever et al. (2018), Kastrin et al. (2016), Sebastian et al. (2015)
Accuracy	Sebastian et al. (2017*b*), Sang et al. (2015)
Cumulative Gain	Vlietstra et al. (2017)
Mean Reciprocal Rank	Song et al. (2015)
Correlation Analysis	Baek et al. (2017), Yang et al. (2017), Xun et al. (2017)

domains where similar resources are available, and *Category 2:* Domain-independent (Table 2.11).

The most prominent and widely used evaluation technique, which is *discovery replication*, is only limited to the medical domain. Other popular evaluation techniques such as the *use of curated databases and resources* and *comparison with existing LBD tools* are also highly domain-dependent and mostly available for the medical domain. Nevertheless, the most objective evaluation technique considered so far in the discipline, which is *time-sliced evaluation*, is domain-independent. Most of the remaining evaluation techniques are typically independent of the domain and can be utilised in non-medical LBD studies.

xvi.2 What quantitative measurements are used to assess the effectiveness of the results?

Different information retrieval metrics have been used to obtain a quantitative understanding of the performance of the LBD methodologies, as summarised in Table 2.12. From our analysis we observed that *precision* (i.e., fraction of associations obtained from the LBD process that are relevant), *recall* (i.e., fraction of relevant associations that are successfully retrieved), *F-measure* (i.e., harmonic mean of precision and recall) and *Area Under Curve (AUC)* (i.e., area under the Receiver Operating Characteristic (ROC) curve, which falls in the range from 1 to 0.5) are the popular evaluation metrics used in the previous literature.

Since most of the time the users will not able to go through the entire list of suggested associations, it is also important to evaluate the proportion of associations in the top k positions that are relevant. For this purpose, the metrics such as *precision at k*, *recall* *at k*, *11-point average interpolated precision* and *mean reciprocal rank* have been used in the LBD literature.

xvii Limitations

Even though we present the insights gleaned from our rigorous literature analysis with confidence, we may have missed LBD research articles that are outside the six databases and six keywords we used. To alleviate this issue to some extent, we also included the references from a recent review (Henry & McInnes 2017) during our paper retrieval process, as discussed in Section xi.

xviii Discussions and Future Work

The key findings and future research directions of each component of the LBD workflow are summarised below.

Input Component: The primary source of data utilised in the LBD studies is *research papers*. Different studies have extracted different details from the research papers for their analysis. Among them, using *title and abstract* is the most popular method. However, some studies have proven the use of full-text and other metadata (such as keywords, references, author details and venue details) assists to glean additional cues of the anticipated knowledge links. Lee et al. (2015) pointed out that *different perspectives* are reflected by different input types used in the content of the research papers. In their analysis, they have found that *keyphrases*, *citation relationships* and *MeSH* reflect the views of *authors*, *citers* and *indexers*, respectively. Moreover, Kostoff et al. (2004) have analysed the *information content* in various fields of a paper using four metrics; total number of phrases, number of unique phrases, factor matrix filtering and multi-link hierarchical clustering. They have identified that the selection of the field depends on the objectives of the study, as described in (Kostoff et al. 2004). Hence, selecting the suitable input type in the papers is crucial as they represent different *perspectives* (Lee et al. 2015) and *information content* (Kostoff et al. 2004) and mainly depends on the objective of the research. Furthermore, Nagarajan et al. (2015) have discovered that the LBD performance mainly depends on the richness of the information being used.

Apart from research papers, several approaches have experimented the LBD process with other traditional input types (such as patents and clinical case reports). Smalheiser et al. (2015) have identified that *information nuggets* (i.e., main findings) are surprisingly prevalent and large in clinical case reports. Mostly, the title itself reveals the main findings of the case report that enables ample opportunities for *finding-based information retrieval* (Smalheiser et al. 2015).

Interestingly, the LBD methodology was successfully adopted to non-traditional input types (such as drug labels, Tweets, news articles and web content). Therefore, an interesting future direction would be to analyse how the LBD process using research papers can be enhanced by integrating knowledge from non-traditional input types (such as Tweets). Furthermore, since most of the non-traditional input types are utilised in medical domain, another interesting avenue would be to integrate the LBD process into other domains using input types such as product descriptions (for product recommendation), movie scripts (for movie recommendation) and recipe books (for recipe recommendation).

With respect to unit of analysis, making use of controlled vocabularies such as UMLS, MeSH and Entrez Gene to extract concepts is the most popular approach. However, research outside the medical domain have followed a term-based approach by extracting n-grams. As the controlled vocabularies utilised yet in LBD research are in the medical domain, an interesting future avenue is to experiment the use of general-purpose controlled vocabularies (such as DBpedia, Freebase, and YAGO) to facilitate knowledge discovery in a cross-disciplinary manner.

Process Component: Swanson's manually detected medical discoveries have set the foundation for LBD research. Later various computational techniques such as statistical, knowledge-based, relations-based, hierarchical, graph-based, bibliometrics-based and link prediction were proposed to automate and make the process of LBD more efficient. The *filtering* and *ranking* techniques used in an LBD methodology are two equally important major components of the LBD workflow.

Many of the filtering mechanisms utilised in the LBD studies have restricted the search space using *word-level* filters. Considering the *article-level* filters (e.g., analysing the contribution of outlier documents), *section-level* filters (e.g., analysing the contribution of different sections in a research article such as introduction and conclusion) or *sentence-level* filters (e.g., analysing the contribution of sentences that describes the main findings) have received little attention in the literature. Therefore, analysing the effect of various *article*, *section* and *sentence* level filtering techniques to remove noisy associations before the word-level filtering is another important area that needs to be further explored. Ultimately, such techniques will also help to further narrow down the literature search and to eliminate the hindrances of the existing word-level filters.

As for the ranking techniques, most of the studies have utilised conventional statistical measures to rank/threshold their results. Whether using such single measure alone would be sufficient to rank the most promising associations in the top of the list is doubtful. In other words, an association may require satisfying several characteristics to become a significant and promising association among others. Therefore, it would be more interesting to develop a ranking approach that reflects the identified characteristics of potential associations to prioritise the results. For instance, Torvik & Smalheiser (2007) have attempted to derive a formula using seven features that capture various characteristics of an association into a single score by employing a machine learning model. Identifying the important characteristics of a significant and promising association and deriving a score based on these characteristics to rank the LBD results would be more successful than merely relying on standard single measures. In this regard, the analysis of different types of gaps in the literature is useful (Peng et al. 2017). Moreover, Smalheiser (2017) suggests the need of several ranking measures to customise the LBD output according to the user preferences. LION LBD system (Pyysalo et al. 2018) that supports multiple scoring functions to facilitate flexible ranking mechanism can be taken as an example.

Output Component: The typical output of the LBD process is a ranked list of terms that denote the potential associations. However, it is not an effective output technique as the users need to interpret the logical connections of the associations by manually reading the research articles, which is difficult and time-consuming. As a result, other visualisation techniques such as term groupings, graphs and discovery pathways have been proposed in the LBD literature. However, the extent to which these proposed techniques assist the user has been rarely measured. Therefore, providing a better visualisation (which is concise, easily interpretable, less complex, visually attractive and assist users to gain new knowledge) and measuring the user experience of the visualisation are two critical components of the LBD workflow that need to be further explored by incorporating HCI techniques.

Nevertheless, the importance of such techniques has been overlooked by the LBD community. To date, only a few LBD research studies (Wilkowski et al. 2011*b*, Hristovski et al. 2006) have contributed in terms of user interaction studies. These studies make use of *information foraging theory*, which is a technique that analyses the user's information retrieval behaviour. The theory evaluates the user's information seeking behaviour in terms of *costs* and *benefits*. If the user can maximise his/her rate of gaining valuable information (i.e., *maximum benefit*) by spending the lowest amount of energy (i.e., *minimum effort*), it is termed an optimal foraging. The key concepts in an information-seeking context are *information*, *information patches*, *information scents* and *information diet*, which needed to be supported effectively when designing interfaces (Ruthven & Kelly 2011). Therefore, the challenge of information visualisation is to discover effective mechanisms to represent massive amounts of data and provide effective ways to navigate through them to support users with optimal foraging. The novel advances in HCI research will be useful in this regard (Stephanidis 2019). Moreover, Smalheiser & Torvik (2008) emphasises the importance of simplicity in user-interfaces of LBD tools to support widening the target audience.

Evaluation Component: Evaluating the LBD output is challenging and remains to be an open issue as the field lacks gold standard datasets or consistent formal evaluation techniques. The most widely used evaluation technique is replicating Swanson's medical discoveries. However, relying only on discovery replication can be restrictive and may fail to reflect the true performance of LBD systems. Hence, this technique should be accompanied with other evaluation techniques to overcome these limitations. Another popular technique is qualitatively evaluating the results randomly by an expert or author. Nevertheless, this does not give an overall image of LBD systems' performance as few valid associations are taken into consideration for analysis. An LBD system that produces a handful of valid associations in a sea of invalid associations tend to be inefficient (Yetisgen-Yildiz & Pratt 2009). As a result, besides this random qualitative evaluation, LBD systems should also be validated quantitatively to measure their overall performance.

To date, *time sliced evaluation* is considered as the most objective evaluation technique proposed in the LBD field. However, this evaluation technique suffers from two major limitations; 1) The association is proven to valid if the starting and linking term co-occur in the future publications (that do not co-occur in the training set). However, co-occurrence does not necessarily mean that the proposed link has been established, and 2) Rejected associations can still be valid even though they have not been published yet.

To overcome the first limitation, it is important to perform much deeper analysis of language (Korhonen et al. 2014) to verify whether the co-occurrence display a true association, which can be considered as an interesting future direction. Additionally, some studies have attempted to utilise evidence from *curated databases* (e.g., CTD and StringDB) as an alternative for co-occurrence in time-sliced evaluation. However, such curated databases are limited to certain problems and may not be available for every domain or problem. The second limitation of time-sliced evaluation can be alleviated to some extent through domain expert involvement by further evaluating the validity of the rejected associations.

Another interesting direction for future evaluation is to incorporate the actual end users of LBD research to validate the results, which is a neglected area in the literature. For instance, involving users with a diverse range of knowledge and expertise (e.g., novice to expert) will help to understand the extent to which each user will be benefited from the LBD output. In this regard, the hypotheses scoring mechanism used by Qi & Ohsawa (2016) can be considered as a successful first step.

Due to the massive influx of scientific knowledge, the volume of data that the LBD systems expect to analyse increases with time. For instance, a simple search of *'dementia'* results in more than 150,000 records in PubMed alone. This highlights the importance of performing scalability analysis of LBD systems in terms of time and storage. This will also improve the usability of LBD systems.

xix Conclusion

In this review, we present novel, up-to-date and comprehensive categorisations to answer each of our research questions to provide a detailed overview of the discipline. The review summary and a comparison with the following recent reviews (Henry & McInnes 2017, Gopalakrishnan et al. 2019) are available at `https://tinyurl.com/workflow-summary`.

With respect to the *input* component, it is evident that the LBD community is showing a growing research interest towards incorporating knowledge from non-traditional data sources to enhance the traditional setting of the LBD framework and to explore new application areas. Nevertheless, the selection of the *input* needs to be precise and cross-checked against the research objectives, as different input types reflect different *perspectives* (Lee et al. 2015) and *information content* (Kostoff et al. 2004).

Filtering and ranking are two important constituents of the *process* component. Most of the filtering techniques examined in the discipline are at word-level. However, the importance of article-level, section-level and sentence-level filters have been rarely studied in the literature. Considering the ranking component, most of the studies have employed a single conventional ranking technique to prioritise the generated discoveries. This showcases the need of developing a series of interestingness measures that customise the LBD output that suit multiple scientific investigations (Smalheiser 2012).

The *output* component of the LBD workflow is largely neglected in the prevailing literature, which emphasises the necessity of conducting user-interaction studies to assess the user experience. Concerning the *evaluation* component, time-sliced evaluation is the current most objective technique used to validate the results. However, this technique suffers from several limitations which suggests the requirement of developing new evaluation methods and metrics to evaluate the generated output.

We hope that the future LBD studies will contribute to overcome the prevailing research deficiencies in the LBD workflow with the ultimate intention of uplifting the typical research procedures which are followed by the scientists.

2.6 Summary

This chapter surveyed the state-of-the-art in LBD field with an emphasis on the main research objectives of this thesis, including the *input component*, *knowledge discovery framework*, *reuse research* and the *portability of LBD models*. The conclusions of the review form the theoretical foundation for the rest of the thesis and also serve as a roadmap for the ensuing chapters. In addition to the rich source of conclusions obtained for this thesis, this review conducted as part of this chapter also serves as a milestone as it is the *first systematic literature review* conducted in the LBD field (as discussed in detail in the two enclosed publications).

Chapter 3

Research Design

3.1 Introduction

The main motive of LBD studies is to support the discovery of hidden knowledge linkages to assist researchers in formulating novel research hypotheses (Gordon & Dumais 1998, Guo et al. 2020). While reducing the time and effort involved in doing such divergent thinking, this will also help researchers to discover new areas of investigation. Even though significant contributions have been made in tackling this problem over the last few decades, prior research suffers from several major hindrances and shortcomings, as discussed in Section 1.3. The overarching goal of this thesis is to investigate new ways to tackle these identified open and prolonged research deficiencies in the LBD discipline. In doing so, this chapter presents the underlying research design utilised in the remaining chapters of this thesis.

This chapter is organised as follows. Section 3.2 describes the thesis' scope in relation to the main components of the LBD workflow (i.e., the main research objectives 2, 3, 4 and 5 discussed in Chapter 1). Section 3.3 discusses the experimental setup in terms of the main data sources and the golden test cases selected. Section 3.4 delineates the evaluation framework used in this thesis by outlining the advantages and disadvantages of popular existing LBD evaluation techniques. The intention of section 3.5 is to describe the theoretical foundation of the machine learning framework employed in this thesis, as well as the selected evaluation metrics. Section 3.6 discusses the baseline models considered for the comparison of the proposed LBD models. Section 3.7 summarises the

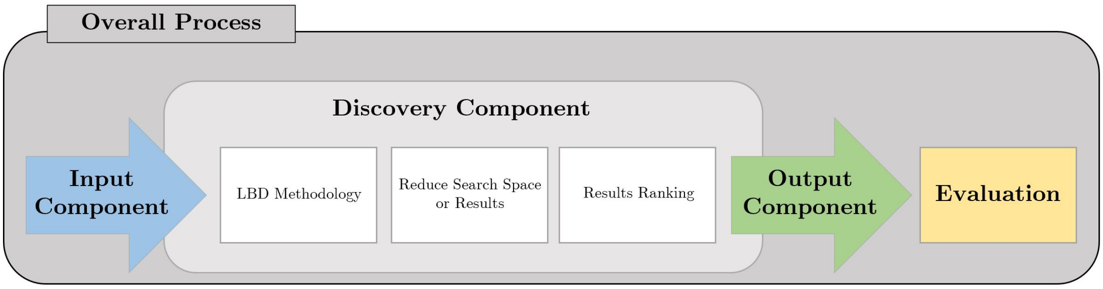


FIGURE 3.1: Research scope in terms of the main components of the LBD workflow

chapter, with a brief outline of the main design selections to be used in the subsequent chapters of this thesis.

3.2 Research Scope

This section is dedicated to discussing the LBD components that are *in scope* as part of the research conducted in this thesis. This discussion centers on the main components of the LBD process, as illustrated in Figure 3.1. It also adheres to the order of the main research objectives discussed in Chapter 1 (i.e., objective 2, 3, 4 and 5, respectively).

- *Input Component:* The input can be viewed as the fuel that drives the entire knowledge discovery process. Thus, comprehending the role of the input component in the LBD process and how it can impact the remaining components in the workflow may provide useful insights and aid in developing better LBD systems in the future. To the best of our knowledge, no studies have specifically attempted to understand the suitability of different input types to the LBD workflow (i.e., *independent input type studies* that isolate the influence of the knowledge discovery method). Contemplating the potential benefits of interpreting the role and the contribution of different input types within the LBD workflow, this thesis attempts to assess their relative suitability, taking inspiration from *information theory* (Tague-Sutcliffe 1992) and *behavioural ecology models* (Stephens & Krebs 1986) in Chapter 4. More specifically, the main research question that this chapter seeks to answer is: "*how can LBD input types be quantitatively assessed and compared so as to better understand their suitability in the LBD workflow?*".

- *Discovery Component:* The discovery component is the most researched component in the LBD literature. It involves developing *knowledge discovery methods* to *filter* unnecessary or meaningless concepts and *rank* potential novel knowledge linkages, as denoted in Figure 3.1. Despite several decades of research using a wide spectrum of computational techniques to automate and streamline the discovery process, the performance of existing LBD models is limited by several shortcomings. The key aim of this thesis is to explore novel ways to circumvent these limitations, with the ultimate goal of eliciting novel knowledge linkages with *high precision*, as discussed in Chapter 5. More precisely, this chapter is based on the research question: *"does incorporating meaningful diachronic semantic inferences in the LBD discovery process through leveraging implicit semantic relationships of word embeddings in temporally-aware vector spaces enrich the typical static cues used in the previous LBD studies?"*.

- *Reusability (i.e., involving the LBD workflow): Reusability* is the process that concentrates in adaptation and integration of the constructed components efficiently into *new applications*. Inspired by the broad benefits that reuse research can offer, this thesis explores the extent to which the proposed LBD framework can be reused in a new application area. For this purpose, this thesis adheres to a methodical reuse plan to assess the extent to which the proposed models can be reused in new settings. Chapter 6 is dedicated to discussing the reuse research performed as part of this thesis with the main research question of: *"how can the reusability of the proposed LBD models be ensured in a new application area, to further confirm their robust predictive power?"*.

- *Portability (i.e., involving the LBD workflow): Portability* refers to the adaptation of the constructed components to *new environments*, at little or no cost. Due to the importance of the problem that LBD research attempts to solve (regardless of the domain), stakeholders of LBD systems could exist in almost all academic disciplines. Therefore, fulfilling the notion of portability is crucial in LBD models to ensure its widespread applicability. Nevertheless, the existing LBD models are mostly tailored to the medical domain, relying on semantic inferences made using medicine-specific knowledge resources. This hinders the models' portability. Even though LBD has been researched for over thirty years, the lack of such portable research may explain why LBD research outside the medical domain is still in a nascent stage. Portable

LBD frameworks have the potential to expand the currently constrained environments of LBD settings, which stand to provide a wide range of benefits to the scientific community. With such a goal in mind, this thesis explores the leveraging of semantic web technologies as a way to port LBD models to a wider range of environments. Chapter 7 is dedicated to discussing this novel initiative in the LBD discipline using the main research question: *"how can an interdisciplinary (or generalisable) LBD framework be developed in a way that ensures the portability of the LBD workflow to new portable environments with little or no cost?"*.

3.3 Experimental Setup

The intention of this section is to describe the *datasets* and *test cases* utilised in the experiments performed in this thesis. This section also outlines the main reasons for selecting the datasets and test cases that were used in these experiments.

3.3.1 Datasets

The main dataset used in the subsequent chapters of this thesis is extracted from *MEDLINE* (Guo et al. 2020). The main reason for this selection is that MEDLINE has been commonly used as the *primary data source* in previous LBD studies. It is considered to be one of the largest scientific repositories that provides access to more than 25 million scientific articles (Jha et al. 2018); thus, provides the opportunity to perform a large-scale literature mining. Figure 3.2 illustrates how the scientific articles got accumulated in MEDLINE over the years that showcase the exponential growth of scientific literature over time.

The National Library of Medicine (NLM) produces baseline data for MEDLINE on an annual basis. This data contains timestamped citation records. This thesis considered the 2019 version of the *MEDLINE* data dump[1], which comprises 25,396,551 total records. The data dump consists of numerous data fields extracted from the articles, including *titles*, *abstracts*, *MeSH keywords* and other metadata such as *author names*,

[1] `https://www.nlm.nih.gov/bsd/medline.html` (downloaded as at January, 2019)

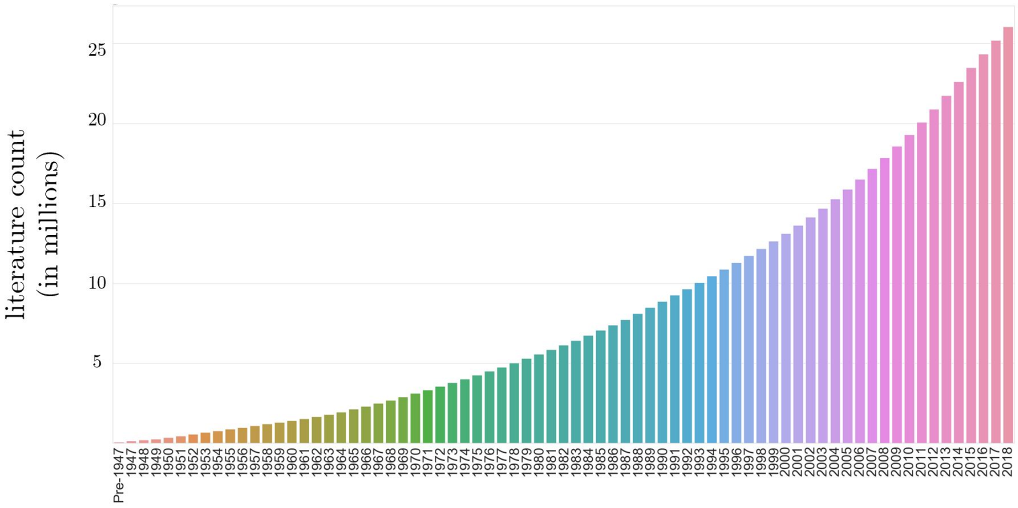


FIGURE 3.2: Yearly accumulated literature count in MEDLINE

venue details, *publication dates* etc. Since the main focus of this study is to analyse textual data using natural language processing techniques, the following three data fields are considered in this thesis: *titles*, *abstracts* and *MeSH keywords*.

Table 3.1 summarises which of these textual data fields are utilised in the remaining chapters of this thesis. In addition to these textual data fields, the publication date of the articles is also considered in order to facilitate tasks such as local topics identification (as discussed in Chapters 5 and 6), diachronic semantic inferences (as discussed in Chapters 5 and 6) and evaluation (as discussed in Section 3.4). For the experiments, this thesis considered scientific articles published from 1960 onwards in the MEDLINE data dump (further details are discussed in Section 3.4).

In addition to MEDLINE, two other datasets were employed in Chapters 6 and 7 of this thesis (Table 3.1). More specifically, Chapter 6 utilises *chemical-disease* relations as reported in the *Comparative Toxicogenomics Database (CTD)*[2] (Mattingly 2009). Further details on how these chemical-disease relations are used are discussed in Chapter 6. The other dataset utilised in Chapter 7 employs the terminology used in the LBD study by Gordon et al. (2002) (the purpose of using this dataset is discussed in Chapter 7). The main reason for this selection is that it is the only available LBD study that is directly relevant to computer science domain (Gordon et al. 2002). More specifically, in this LBD study, Gordon et al. (2002) attempted to detect novel applications of

[2] downloaded as at 5th of April, 2020

Table 3.1: Summary of the datasets used in the experiments

Chapter	**Datasets**
Chapter 4 (Input Types)	MEDLINE (titles, abstracts, MeSH keywords and publication date)
Chapter 5 (Semantic Evolution)	MEDLINE (MeSH keywords and publication date)
Chapter 6 (Reusability)	MEDLINE (MeSH keywords and publication date), CTD (chemical-disease relations)
Chapter 7 (Portability)	MEDLINE (titles, abstracts and publication date), Terminology from the only existing computer science LBD study (Gordon et al. 2002)

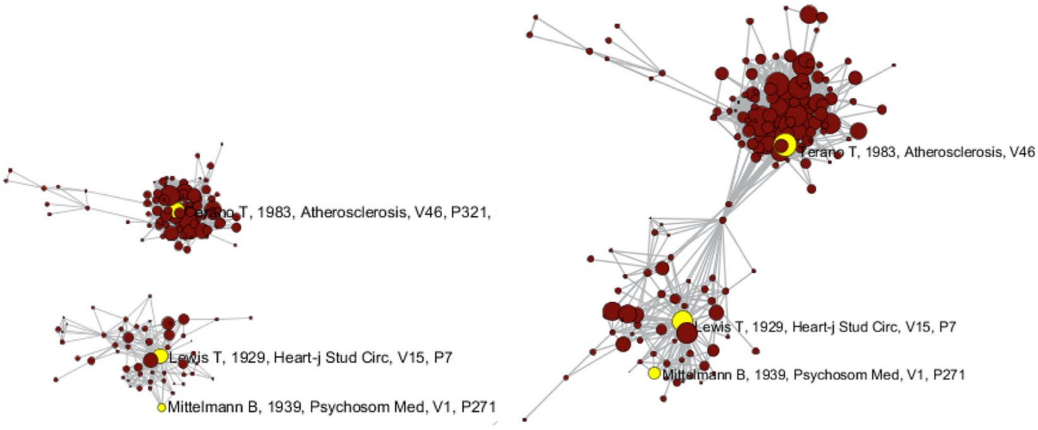
genetic algorithms (i.e., the starting concept) by viewing the A-B-C discovery path as a *technology-technology-application* problem.

3.3.2 Test Cases

To evaluate the effectiveness of the proposed solutions and to compare them with the existing LBD models and methods, test cases are required. For this purpose, this thesis considered the following five real-world test cases reported by the pioneers of the LBD discipline. The main reason for selecting these test cases is that they are commonly used for LBD evaluation and treated as *golden datasets* in the discipline (Jha et al. 2018, Jha, Xun, Wang & Zhang 2019, Xun et al. 2017).

1. *Fish-Oil (FO)* and *Raynaud's Disease (RD)* (Swanson 1986)
2. *Magnesium (MG)* and *Migraine Disorder (MIG)* (Swanson 1988)
3. *Somatomedin C (IGF1)* and *Arginine (ARG)* (Swanson 1990*a*)
4. *Alzheimer's Disease (AD)* and *Indomethacin (INN)* (Smalheiser & Swanson 1996)
5. *Schizophrenia (SZ)* and *Calcium-Independent Phospholipase A2 (PA2)* (Smalheiser & Swanson 1998)

The significance of some of these test cases in the LBD context is that they are *complementary* but *disjointed*. This means that the articles in the two topics of each test case have never been mentioned or cited together. For instance, consider Figure 3.3, which demonstrates the disjointed nature of test case 1, which is *FO-RD* before the Swanson's discovery in 1986. Therefore, the use of the aforementioned golden test cases validates


Terano T, 1983, Atherosclerosis, V46, P321,
Lewis T, 1929, Heart-j Stud Circ, V15, P7
Mittelmann B, 1939, Psychosom Med, V1, P271
Terano T, 1983, Atherosclerosis, V46
Lewis T, 1929, Heart-j Stud Circ, V15, P7
Mittelmann B, 1939, Psychosom Med, V1, P271

(a) Before Swanson's Discovery

(b) After Swanson's Discovery in 1986

FIGURE 3.3: Complementary and non-interactive nature of FO-RD test case (Sebastian et al. 2017*b*)

the LBD model's ability to accumulate existing disperse knowledge in the literature to develop novel semantic relationships that have not previously attracted any attention.

3.4 Evaluation Framework

The purpose of this section is to describe the evaluation framework employed in the subsequent chapters of this thesis. Prior to selecting an evaluation framework, the initial part of this section discusses the potential reasons why evaluation is difficult and an open issue in the LBD discipline. Subsequently, to identify the most appropriate evaluation setting, the criteria that constitute an ideal evaluation setting are discussed. The selection of an evaluation framework of this thesis was made by weighing these criteria, considering the most popular evaluation techniques in the LBD field.

3.4.1 Evaluation in the LBD Discipline

Evaluating LBD models is difficult and considered to be an open problem in the discipline. Firstly, no standard ground truth exists in the LBD field and the creation of such a ground truth remains an open issue, as it is nearly impossible to construct a comprehensive ground truth that will presumably contain all future discoveries (Jha et al. 2018). Other reasons for the difficulty of evaluation in the field of LBD include disagreements about the role of LBD models in research, difficulty in quantifying how interesting, useful or actionable a predicted discovery is, and difficulty in objectively defining what a 'discovery' is (Crichton et al. 2020). As a result of these barriers, there is no existing technique that evaluates LBD models perfectly. Therefore, it is important to weigh the advantages and disadvantages of existing LBD evaluation techniques, and to select the most suitable evaluation framework to quantify and compare the LBD outputs.

3.4.2 Ideal Evaluation Setting

Before comparing the existing LBD evaluation techniques, it is important to identify what constitutes an *ideal evaluation setting*. The following criteria need to be satisfied in order to consider an evaluation method in its ideal setting (Henry 2019).

- *Automated:* The evaluation method should be scalable and easily calculated in a reasonable amount of time, without requiring a manual process.
- *Replicable:* The evaluation method should be objective and support replication.
- *Quantifiable:* The evaluation method should provide a numeric metric indicating performance, which facilitates comparison between other LBD systems.
- *Informative:* The evaluation method should facilitate a deeper understanding of the model's behaviour.
- *Modular:* The evaluation method should not rely on the LBD workflow, since it should facilitate the evaluation of single components (or even sets of components) in isolation.

3.4.3 Selection of Evaluation Method

This section discusses the strengths and weaknesses of the popular existing LBD evaluation techniques by cross-checking the extent to which they fulfil the criteria outlined in Section 3.4.2.

The evaluation technique, *discovery replication*, focuses on reproducing historical LBD discoveries. If the terms identified in the historical discoveries are identified or ranked highly enough, according to this evaluation technique, the LBD model is considered to be successful. This evaluation method can be *automated* and *replicable (Henry 2019)*. The results are easy to understand and demonstrate the model's ability to identify at least one discovery when the target term(s) is reported at a higher rank. The consideration of ranking positions makes this evaluation method *quantitative*. Nevertheless, discovery replication is a narrow and constrained task that only reports a few hand-selected discoveries that have been reported in the LBD literature. Thus, discovery replication is prone to *overfitting* (Yetisgen-Yildiz & Pratt 2009, Henry 2019). Moreover, the reporting of a ranking position is considered to be an unstable metric (Henry 2019). Discovery replication provides little insight into how the LBD model works and the potential ways that it could be improved. Thus, this evaluation method is *not informative* (Henry 2019).

The focus of *user studies* in the LBD context is to understand how users operate the LBD system and how the system could be further improved based on users' feedback and activities (Qi & Ohsawa 2016, Cohen, Whitfield, Schvaneveldt, Mukund & Rindflesch 2010). Typically, user studies provide a good platform to understand how LBD models are actually being used. Therefore, this evaluation method is *extremely informative* (Henry 2019). Nevertheless, user studies suffer from subjectivity, making them *non-replicable* (Henry 2019, Yetisgen-Yildiz & Pratt 2009). The reliance on human users also means that this evaluation method is *non-automated*. In addition, user studies are *modular*, since they can be used to evaluate certain components in the LBD model such as the user interface and visualisation.

New discovery proposals indicate the discoveries made using an LBD model. This evaluation method provides opportunities to expose LBD to a wider community and gives LBD credibility. Nevertheless, the involvement of domain experts makes this evaluation method *non-replicable*, *non-automated* and *non-quantitative* (Henry 2019, Yetisgen-Yildiz & Pratt 2009). Furthermore, this method does not provide insights into how the LBD model works with respect to individual components or the model as a whole; thus, the method is *not informative*. Since the new discovery proposal relies on the entire LBD process, it is also *non-modular* (Henry 2019).

Time-slicing is an evaluation technique that uses a cut-off-date to divide the literature into training and testing sets. To date, time-slicing is the most objective evaluation method to have been proposed in the LBD field, circumventing most of the key issues

TABLE 3.2: Assessing the suitability of the popular evaluation methods in LBD

Evaluation Method	**Automated**	**Replicable**	**Quantifiable**	**Informative**	**Modular**
Discovery Replication	✓	✓	✓	-	-
User Studies	-	-	✓	✓	✓
New Discovery Proposal	-	-	-	-	-
Time-Slicing	✓	✓	✓	✓	✓

where '✓' denotes the support of the relevant criterion with the evaluation method and '-' denotes if the evaluation method does not fulfil the relevant criterion

in the existing evaluation methods (Yetisgen-Yildiz & Pratt 2009). Time-slicing provides the platform to evaluate in an *automated* and *quantitative* manner (Henry 2019, Yetisgen-Yildiz & Pratt 2009). This facilitates the usage of *informative* metrics, such as *precision at k* and *mean average precision* (Henry 2019, Jha et al. 2018). Moreover, this evaluation method is *replicable* due to its standardised procedure (Henry 2019, Yetisgen-Yildiz & Pratt 2009). Time-slicing is also *modular*, since it can be used to evaluate individual components in the model (Henry 2019). There are several criticisms of time-slicing in the LBD field, since the technique is mainly based on co-occurrence; thus, it can contain false discoveries or noise (Henry 2019). Nevertheless, for large scale, quantifiable evaluations, time-slicing is (so far) the only available evaluation method in the LBD discipline (Yetisgen-Yildiz & Pratt 2009, Crichton et al. 2020).

Table 3.2 summarises the extent to which popular existing LBD evaluation methods fulfil the criteria defined in Section 3.4.2 (Henry 2019). Overall, time-slicing fulfils every defined criterion, making it the most suitable evaluation setting in the LBD field. Thus, this thesis selected *time-slicing* for the evaluation of the proposed LBD models.

3.4.4 Time-Slicing Setting

In a time-slicing setting, the LBD system uses *known knowledge* to make its predictions and verifies whether the proposed novel knowledge linkages have actually taken place in the *future*. The proposed novel linkage is *legitimate* if it is *absent* in the *known knowledge* and *present* in *future knowledge* (Yetisgen-Yildiz & Pratt 2009, Jha et al. 2018, Xun et al. 2017). For instance, consider Figure 3.4 that depicts how scientific knowledge evolves over time by forming connections with different topics. The known connections between scientific topics are limited in timestamp t. However, with the ongoing research findings,

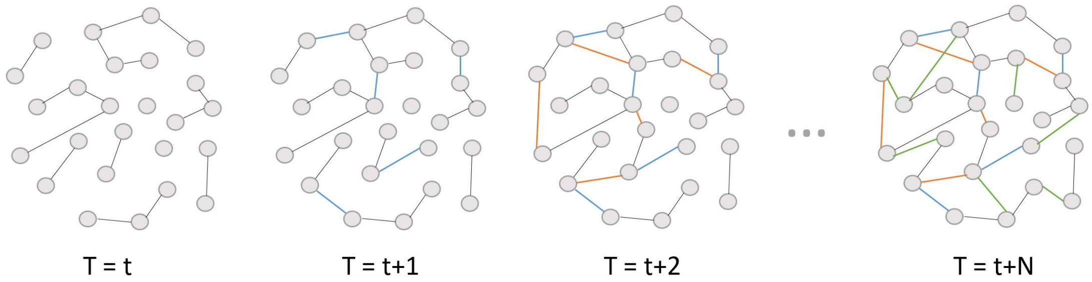


FIGURE 3.4: Temporal evolution of scientific knowledge at different timestamps (where nodes represent scientific topics and edges represent if there is a connection between two scientific topics at a given timestamp)

more connections have been established between topics that are depicted by edges in blue, orange and green colours in timestamps *t+1*, *t+2* and *t+N*, respectively. The purpose of the time-slicing setting is to predict such future connections between topics (i.e., shown in timestamps *t+1*, *t+2* and *t+N*) by only using the *known knowledge* (i.e., using the topic interactions at timestamp t).

In LBD context, time-slicing is achieved by dividing the literature repository into two segments: *pre-cut-off* and *post-cut-off*[3]. The *pre-cut-off* segment represents *known knowledge*, and the LBD system uses the knowledge in this segment to discover the potential future discoveries. The *post-cut-off* segment that represents *future knowledge* is used to evaluate the legitimacy of the predictions. The legitimacy of a discovered knowledge linkage is established if it is present in the post-cut-off segment and absent in the pre-cut-off segment (Yetisgen-Yildiz & Pratt 2009). Typically, *co-occurrence* is used to detect such newly established knowledge linkages in the literature (Xun et al. 2017, Jha et al. 2018, Jha, Xun, Wang & Zhang 2019). Figure 3.5 summarises the above-discussed *pre-cut-off* and *post-cut-off* setting used in time-slicing. Table 3.3 summarises the details of the pre-cut-off and post-cut-off segments of the selected five golden test cases discussed in Section 3.3.2.

As in previous LBD studies (Jha et al. 2018, Xun et al. 2017, Jha, Xun, Gopalakrishnan & Zhang 2019), Chapter 5 uses the following equation: $gt(k) = \frac{\#(k,A)+\#(k,C)}{\#(k)}$, where $\#(i,j)$ is number of times concepts i and j co-occur and $\#(i) = \sum_j \#(i,j)$ to rank the ground truth conceptual bridge k for the two given topics of interest (A and C). For example, consider the FO-RD test case where the ground truth intermediate concepts k are ranked using $gt(k) = \frac{\#(k,\text{“}FO\text{”})+\#(k,\text{“}RD\text{”})}{\#(k)}$ in the post-cut-off segment of 1986-2019

[3] https://github.com/Menasha/LBD/

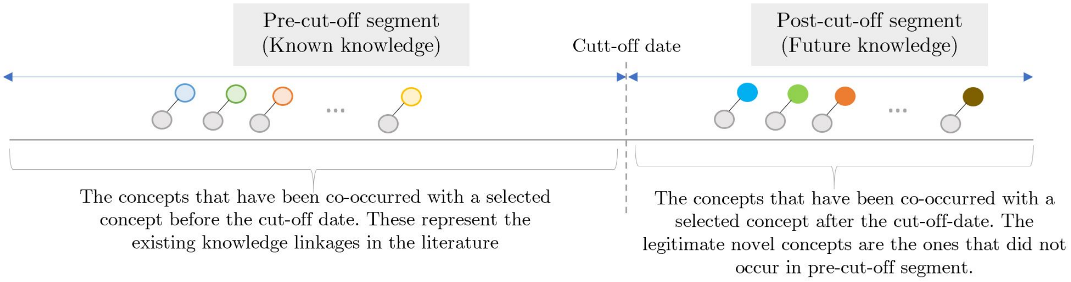

Pre-cut-off segment
(Known knowledge)
Cutt-off date
Post-cut-off segment
(Future knowledge)
...
...
The concepts that have been co-occurred with a selected concept before the cut-off date. These represent the existing knowledge linkages in the literature
The concepts that have been co-occurred with a selected concept after the cut-off-date. The legitimate novel concepts are the ones that did not occur in pre-cut-off segment.

Figure 3.5: Pre-cut-off and post-cut-off segments of the time-slicing setting

Table 3.3: Time-slicing setting of the golden test cases

Test case	**First Discovery**	**Pre-cut-off Segment**	**Post-cut-off Segment**
FO-RD	Swanson (1986)	1960-1985	1986-Jan 2019
MG-MIG	Swanson (1988)	1960-1987	1988-Jan 2019
IGF1-ARG	Swanson (1990)	1960-1989	1990-Jan 2019
AD-INN	Smalheiser & Swanson (1996)	1960-1995	1996-Jan 2019
SZ-PA2	Smalheiser & Swanson (1998)	1960-1997	1998-Jan 2019

(Table 3.3). To retain only the legitimate novel knowledge linkages, all the existing connections in the pre-cut-off segment (i.e., 1960-1985 as listed in Table 3.3) are removed from the ranked list. Some examples of the identified novel knowledge linkages from this time-slicing setup include *blood viscosity*, *platelet aggregation*, *vasoconstriction*, *vasodilation* and *prostaglandins e*. Highly frequent terms (i.e., the bottom 5% of this ranked list) were removed from the ground truth as a post-processing step. As in previous LBD studies, Chapter 6 employs one-node time-slicing, where $\#(k,A)$ is used to decide the legitimacy of the novel knowledge linkage (i.e., the co-occurrence pair is unavailable in the pre-cut-off segment and available in the post-cut-off segment) (Yetisgen-Yildiz & Pratt 2009). In the instance of the FO-RD test case, the novel knowledge linkages are identified using $\#(k,$ *"RD"*$)$ in the post-cut-off segment of 1986-2019 by removing all the existing connections that occurred in the pre-cut-off segment of 1960-1985 (Table 3.3). Some examples of the identified novel knowledge linkages in this process include *fish oils*, *eicosapentaenoic acid*, *lipoproteins ldl*, *oils* and *cardiolipins*. Table 3.4 summarises the number of local topics and the number of legitimate novel knowledge linkages identified for each golden test case in the time-slicing setups of Chapters 5 and 6.

Table 3.4: Time-slicing setup used in Chapters 5 and 6

Test case	**Setting**	**Local Topics**	**Novel Linkages**
FO-RD	Chapter 5	3014	914
	Chapter 6	2964	250
MG-MIG	Chapter 5	9487	3064
	Chapter 6	3026	536
IGF1-ARG	Chapter 5	6298	3819
	Chapter 6	–	–
AD-INN	Chapter 5	9632	4126
	Chapter 6	3182	1067
SZ-PA2	Chapter 5	9179	2567
	Chapter 6	3105	409

3.5 Machine Learning Framework

The automation of data analysis techniques became possible with the development of digital computers in the mid 20th century. Fuelled by rapid advancements in algorithms and computer power over the past half-century, machine learning methods have become powerful tools for discovering complex and subtle patterns in data (Biamonte et al. 2017, Anzai 2012). The purpose of *machine learning* (or its subfield, *deep learning*) is to elicit patterns from large volumes of data (Nguyen et al. 2019, Ongsulee 2017). This aligns with the aim of *LBD* research, which is to uncover patterns of potential novel knowledge linkages from vast quantities of literature. Therefore, the integration of machine learning methods into the *discovery component* of the LBD workflow opens up ample opportunities to perform large-scale knowledge discovery, in order to perceive complex and subtle patterns in the literature. Discovering such intricate structures in the scientific literature is essential to the automated generation of high-quality predictive decisions. With this idea in mind, this thesis incorporates machine learning techniques in order to discover potential novel knowledge linkages in the scientific literature with high precision.

The purpose of this section is to describe the foundation of the machine learning setup used in Chapters 5 and 6. In doing so, the first part of this section discusses how the machine learning setup was mapped to the *discovery component* of the LBD workflow. Subsequently, the process of *stratified cross-validation* is discussed. Stratified cross-validation was used to obtain prediction probabilities of scientific topics, indicating their likelihood of becoming a novel knowledge linkage (when they were in the test sample). Subsequently, the procedure of *cost-sensitive learning*, which was used in the training phase of the stratified cross-validation, is discussed. Next, the theoretical foundations of deep learning and machine learning settings are discussed. These settings are used in the construction of the machine learning models in Chapters 5 and 6. The latter part of this section discusses the evaluation metrics employed to quantify and compare the performance of the LBD models.

3.5.1 Setup of the Discovery Component

The *discovery component* of a typical LBD workflow comprises two main tasks: *filtering* and *ranking* (Henry & McInnes 2017). The purpose of the 'filtering' component is to discard uninteresting or meaningless scientific topics during knowledge discovery (Figure 3.6). The 'ranking' component attempts to efficiently order the remaining scientific topics (i.e., the scientific topics retained in the 'filtering' component) to assist researchers to develop novel hypotheses (Figure 3.6). In a typical machine learning setup, these two tasks from the *discovery component* can be mapped to two classes: *negative instances* and *positive instances* (Bunescu & Mooney 2007, Settles et al. 2008, Jha, Xun, Gopalakrishnan & Zhang 2019). The *positive instances* signify potential novel knowledge linkages (resembling the 'ranking' component). In the time-slicing setup (discussed in Section 3.4), positive instances denote the scientific topics that were realised in the *post-cut-off segment*, but which were absent in the *pre-cut-off segment* (a.k.a. *legitimate novel knowledge linkages*(Jha et al. 2018)). The *negative instances* are uninteresting or meaningless concepts (resembling the 'filtering' component). This category denotes the remaining scientific topics that are not identified as legitimate novel knowledge linkages.

With the mapping of negative and positive instances to the 'filtering' and 'ranking' tasks in the discovery component, the goal of the machine learning model is as follows: given a scientific topic, the machine learning model predicts the probability with which it will belong to the negative class P_{neg}, or the positive class P_{pos} (Figure 3.6). In the context of LBD, these two prediction probabilities can be interpreted as follows. P_{pos} denotes the probability of a scientific topic becoming a novel knowledge linkage. Thus, the higher P_{pos} is, the higher the chance that the relevant scientific topic to be a potential novel knowledge linkage. Similarly, P_{neg} (i.e., 1 - P_{pos}) signifies the likelihood that a scientific topic is meaningless or uninteresting. Thus, the higher P_{neg} is, greater the chance that

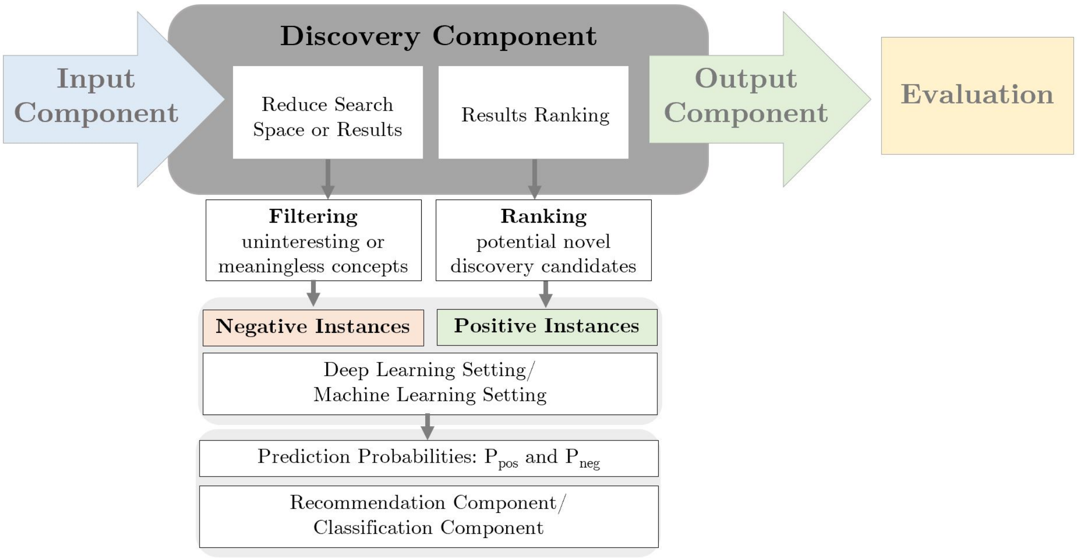

Discovery Component
Input Component
Reduce Search Space or Results
Results Ranking
Output Component
Evaluation
Filtering uninteresting or meaningless concepts
Ranking potential novel discovery candidates
Negative Instances
Positive Instances
Deep Learning Setting/ Machine Learning Setting
Prediction Probabilities: P_{pos} and P_{neg}
Recommendation Component/ Classification Component

Figure 3.6: Mapping the machine learning setup to the discovery component in the LBD workflow

the scientific topic is an unnecessary or meaningless concept in the context of knowledge discovery.

The two probabilities P_{pos} and P_{neg} can be employed in two settings: the *recommendation component* and the *classification component* (Figure 3.6). The purpose of the recommendation component is to evaluate the validity of the scientific topics that the machine learning model predicts with high P_{pos} probability (i.e., the P_{pos} in the descending order). In contrast, the classification component gauges how well the machine learning model is able to classify the positive and negative instances by mapping each scientific topic to the class in which it indicates the highest probability, from P_{pos} and P_{neg}. Further details on these two setups are discussed in Section 3.5.6 and Chapter 5.

3.5.2 Stratified Cross-Validation

Cross-validation (more specifically, *k-fold cross-validation*) partitions the available learning set (i.e., the *positive and negative instances* discussed in Section 3.5.1) into k number of disjoint subsets (or *folds*) of approximately equal size (Zhang et al. 2016). For the purpose of model training, *k-1* subsets are used, which represents the *training data* of the machine learning framework. The remaining fold, which is known as the *test set*, is used to apply the machine learning model to obtaining prediction probabilities (i.e.,

P_{pos} and P_{neg}). This process is repeated until all k subsets have served as a test set in the iterations (Niu et al. 2018, Berrar 2019).

Figure 3.7 denotes the cross-validation process in which k is set to 10 (i.e., *10-fold cross-validation*). Cross-validation often uses *stratified* random sampling (a.k.a. *stratified k-fold cross-validation*), which denotes that the sampling is performed while preserving the class proportions in the learning set in the individual folds. The underlying aim of stratification is to avoid biased evaluation, since the data folds used for evaluating the machine learning model reflect the class ratio in the population (Berrar 2019, Purushotham & Tripathy 2011). Therefore, the machine learning setups used in Chapters 5 and 6 are based on the *stratified k-fold cross-validation* variant. More specifically, every instance in the dataset becomes an ‘unknown’ (i.e., not included in the training set) in one of the iterations in stratified cross-validation, where the prediction probabilities (P_{pos} and P_{neg}) are computed by the machine learning model. Thus, these probabilities are used to evaluate the validity of the machine decision (Bannach-Brown et al. 2019, Kwon et al. 2019, Purushotham & Tripathy 2011).

It should also be noted that the same training data were used in each fold of the stratified cross-validation, in order to obtain the prediction probabilities P_{pos} and P_{neg} of the corresponding test set in all the proposed LBD models, including the baselines (Kwon et al. 2019). The main reason for ensuring the consistency of the training and testing folds among the LBD models is to facilitate a uniform and unbiased performance comparison. In other words, the prediction probabilities (P_{pos} and P_{neg}) of the test sets are obtained using the same training sets across all LBD models.

3.5.3 Cost-Sensitive Learning

In the machine learning settings, it is important to integrate some strategy to balance if there are any imbalances between the two classes in the training sample (i.e., positive and negative instances in the stratified cross-validation, as discussed in Section 3.5.2). For this purpose, this thesis incorporates *cost-sensitive learning* to ensure that if there are any imbalances among the classes, the machine learning model is aware of the fact (Zadrozny et al. 2003). Cost-sensitive learning is incorporated by mapping class weights inversely proportional to class frequencies (a.k.a. *balanced mode*) (Zadrozny et al. 2003, Liu & Zhou 2006). If the classes in the training sample are balanced, the weight will be

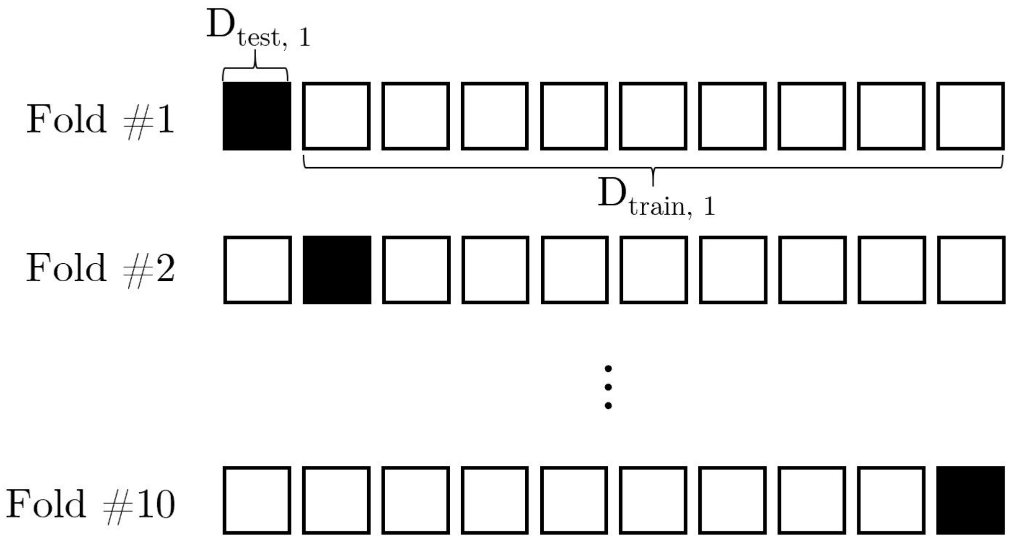


FIGURE 3.7: 10-fold cross-validation in which $D_{\text{test},1}$ denotes the first fold that is served as the testing set and $\text{D}_{\text{train},1}$ denotes the remaining nine folds used for training

1 for each class. More specifically, the class weights are defined as $\frac{n_{samples}}{(n_{classes} \times [n_1, n_2, ...])}$, where $n_{samples}$ is the number of instances in the training sample, $n_{classes}$ is the number of classes, and n_1, n_2, ... are the number of instances in each class (Elkins et al. 2019).

3.5.4 Deep Learning Setting

The purpose of this section is to discuss the theoretical foundation of the deep learning setting used in this thesis. More specifically, The deep learning models are constructed using two main building blocks, namely *Long Short-Term Memory (LSTM)* and *Convolutional Neural Networks (CNN)*. The main reason for using these two building blocks is that *LSTM* has shown superior performance in modelling *temporal dynamics* (Lee et al. 2017), while *CNN* excels at detecting *low-level to high-level features* using its series of feature extractors (Yu et al. 2018). Therefore, the use of these two variants of deep neural networks provides an extended platform from which to decide which setting is most appropriate in the LBD context (i.e., comparing the suitability of modelling temporal dependencies using LSTM or the multiple layers of feature hierarchies using CNN (Ordóñez & Roggen 2016)). The deep learning models constructed using these two primary building blocks are used to obtain the prediction probabilities P_{pos} (i.e., the probability of becoming a novel knowledge linkage) and P_{neg} (i.e., the probability of being an unnecessary or meaningless concept; 1 - P_{pos}). Further details on the

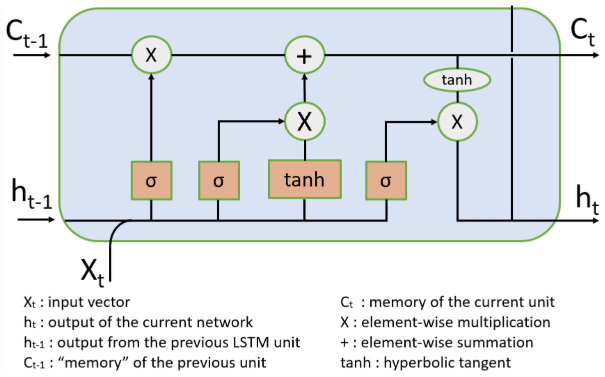

C_{t-1}
C_t
h_{t-1}
h_t
X_t
X
+
X
tanh
X
σ
σ
tanh
σ
X_t : input vector
h_t : output of the current network
h_{t-1} : output from the previous LSTM unit
C_{t-1} : "memory" of the previous unit
C_t : memory of the current unit
X : element-wise multiplication
+ : element-wise summation
tanh : hyperbolic tangent

FIGURE 3.8: Internal structure of the LSTM unit (Chen et al. 2020)

constructed deep learning models (using *LSTMs* and *CNNs*) are discussed in Chapter 5.

3.5.4.1 Long Short-Term Memory (LSTM)

LSTM (Hochreiter & Schmidhuber 1997) is a specialised version of the *Recurrent Neural Network (RNN)*, which is capable of learning long-term dependencies and detecting long-range features in sequences (Selvin et al. 2017, Chen et al. 2020). The unit structure of an LSTM network is illustrated in Figure 3.8 (Zhou et al. 2019). The main component of the LSTM unit is its *cell*, which keeps track of the dependencies between elements in the input sequence by maintaining a cell state c_t in time. More specifically, LSTM has the ability to remove and add information to the cell state through its three types of gates (Selvin et al. 2017), as outlined below.

- *input gate:* this gate has control over a new value that flows into the cell.
- *forget gate:* this gate has the control to decide the amount of value remains in the cell.
- *output gate:* this gate controls which portion of the value in the cell is used to calculate the LSTM unit's output activation.

The process in LSTM unit can be summarised as follows (Zhou et al. 2019). The *forget gate* (which is a *sigmoid* layer) uses $\mathrm{h}_{t\text{-}1}$ and x_t to decide whether to increase or decrease the data flow by imposing a threshold denoted as $f_t = \sigma(\mathrm{w}_f \mathrm{h}_{t-1} + \mathrm{u}_f \mathrm{x}_t + \mathrm{b}_f)$, where u and w are the values of weights, σ is the activation function, and b is the bias value.

In the next step, the *input gate* (which is also a *sigmoid* layer) determines which values to update (i_t), and is followed by a *tanh* layer that constructs a vector of new candidate values ($\tilde{c}_t$) to store in the cell state. These two functionalities are expressed using: $i_t = \sigma(\mathrm{w}_i\mathrm{h}_{t-1} + \mathrm{u}_i\mathrm{x}_t + \mathrm{b}_i)$ and $\tilde{c}_t = \tanh(\mathrm{w}_c\mathrm{h}_{t-1} + \mathrm{u}_c\mathrm{x}_t + \mathrm{b}_c)$, respectively.

Subsequently, the cell state is updated using the old cell state $c_{t\text{-}1}$ and the new cell state c_t, as denoted in $c_t = c_{t-1} \odot f_t + i_t \odot \tilde{c}$, where $\odot$ is the Hadamard product. In essence, the old cell state $c_{t\text{-}1}$ is scaled according to how much the forget gate has decided to forget, and the new state c_t is scaled according to how much the input gate has decided to update.

Finally, the *output gate* decides the output h_t in two steps (i.e., through *sigmoid* layer and *tanh* filter) as defined in: $o_t = \sigma(\mathrm{w}_0\mathrm{h}_{t-1} + \mathrm{u}_0\mathrm{x}_t + \mathrm{b}_0)$ and $h_t = o_t \odot \tanh(c_t)$. More specifically, the output of the previous moment $h_{t\text{-}1}$ and the input of the current moment x_t are processed first using a *sigmoid* layer, which is then passed to the next stage to filter the current version of the cell state.

3.5.4.2 Convolutional Neural Network (CNN)

CNN is a type of deep neural network that is used to process data with grid patterns (such as images) to automatically and adaptively learn spatial features from low-level to high-level patterns (Yamashita et al. 2018). Thus, CNNs are most commonly applied in research areas related to image analysis (e.g., *computer vision* (Le Guennec et al. 2016)), in which spatial convolutions are cascaded to represent the spatial content in images (Tijskens et al. 2019, Liu et al. 2018). More recently, CNNs have been successfully applied to *learning sequences* (Tijskens et al. 2019) using temporal convolutions in areas such as signal processing (Yang et al. 2020), speech recognition (Fawaz et al. 2019) and time series analysis (Le Guennec et al. 2016). The core idea of CNN was derived from the organisation of the visual cortex in animals (Hubel & Wiesel 1968, Fukushima & Miyake 1982). CNN can be considered a mathematical construct that typically contains three types of layers/building blocks: *convolution*, *pooling* and a *fully connected layer* (Zhao et al. 2017). The purpose of the first two layers is to extract meaningful features, while the latter layer maps these extracted features into a final output (Yamashita et al. 2018).

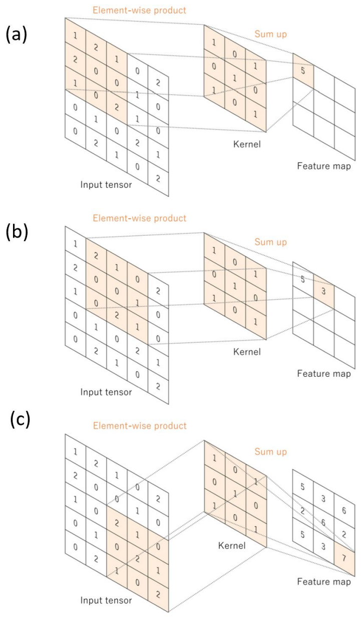


FIGURE 3.9: Simplified example of convolution operation with 3×3 kernel size *(note: a stride of 1 is used in this example, with no padding)* (Yamashita et al. 2018)

The *convolutional layer* can be considered the fundamental unit in CNN. It comprises a stack of mathematical operations, including convolution. Convolution is a specialised form of linear operation in which a small array of numbers (the *kernel*) is applied through the input, which is an array of numbers called a *tensor*. Subsequently, an element-wise product between the elements of kernel and tensor is performed at each location in the tensor to produce a *feature map* (Yamashita et al. 2018). This process is repeated through multiple kernels, in order to construct an arbitrary number of feature maps that denote different characteristics of tensors. Thus, different kernels can be considered to be different *feature extractors or filters* (Yamashita et al. 2018). In essence, the core purpose of this layer is to learn convolutional filters in a data-driven manner, with the ultimate aim of extracting features that efficiently describe the inputs (see Figure 3.9) (Le Guennec et al. 2016).

The *pooling layer* provides a platform to reduce the in-plane dimensionality of the constructed feature maps. Therefore, it involves a down-sampling operation such as *max pooling*, *average pooling*, *probabilistic max pooling* or *differentiable pooling* (Shin et al. 2016). The most popular pooling operation is *max pooling*, which is illustrated in Figure 3.10 (Yamashita et al. 2018). The main intention of this layer is to make feature maps *translation-invariant* with regard to distortions and small shifts, and to preserve

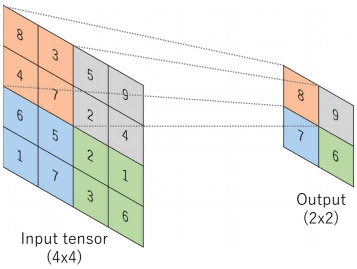


Figure 3.10: Simplified example of max pooling with a 2×2 filter size *(note: a stride of 2 is used in this example, with no padding)* (Yamashita et al. 2018)

important information (Yang et al. 2020). Once the features are extracted and downsampled using the convolution layer and pooling layer, respectively, the *fully connected layer* maps them to the final output of the network (i.e., the probabilities, P_{pos} and P_{neg}) (Yamashita et al. 2018).

3.5.5 Machine Learning Setting

This section describes the setup of the traditional machine learning setting utilised in Chapters 5 and 6. More specifically, these traditional machine learning setups, which are based on handcrafted features, incorporate conventional machine learning algorithms so as to make predictions. To this end, this thesis employed *random forest* as the base learning algorithm. The main reason for this selection was that *random forest* is based on *ensemble learning algorithms*, which are considered to be more accurate than a single machine learning model, since the notion of ensembles is based on the premise that a set of models tend to perform better in comparison to a individual models (Rodriguez-Galiano et al. 2012, Breiman 1996, Xuan et al. 2018). Additionally, random forest presents other advantages, such as its ability to estimate which features are important, its ability to generate an internal unbiased estimation of the generalisation error, its relative robustness to noise and outliers, and its relative computational lightness in comparison to other tree ensemble methods (Rodriguez-Galiano et al. 2012, Cutler et al. 2012, Khoshgoftaar et al. 2007). *Random forest* has also commonly been used as a learning algorithm in previous LBD studies and has demonstrated the highest or competitive results relative to other learning algorithms (Kastrin et al. 2016, Rastegar-Mojarad et al. 2016, Sang, Yang, Liu, Wang, Lin, Wang & Dumontier 2018). As in the deep learning setting (discussed in Section 3.5.4), the output of the machine learning setting is composed of the two prediction probabilities P_{pos} and P_{neg}.

3.5.6 Evaluation Metrics

The selection of evaluation metrics is strongly influenced by the problem objective (Lemnaru 2012). Evaluation metrics that perfectly suit a given scenario may not fit expectations in a different problem setting. For instance, the focus in a medical diagnosis setting is on maximising the true positive rate (Horn et al. 2011), while in contextual advertising problems precision is important (Ciaramita et al. 2008). Therefore, identifying an appropriate performance metric that adheres with specific problem goals is important (Lemnaru 2012).

3.5.6.1 Recommendation Component

The purpose of the recommendation component is to determine how well the identified *characteristics* (or *features*) contribute towards deciding the correct recommendations while discarding the irrelevant ones (discussed in Section 3.5.1). In *recommendation systems*, precision is typically considered to be more important than recall (Tyler & Zhang 2008). In large-scale knowledge retrieval systems, such as LBD, it is unrealistic to assume that the user will read all the predicted recommendations. Therefore, as in previous LBD studies (Jha, Xun, Wang & Zhang 2019, Jha et al. 2018), *Precision at k (P@k)* and *Mean Average Precision (MAP)* are utilised as the key evaluation metrics to assess the recommendation component. In addition to these key metrics, this thesis also uses *Geometric Mean Average Precision (GMAP)* to evaluate the consistency of the predictions.

Precision@k (P@k) denotes the proportion of the top k records that are relevant, as defined in the form $\frac{r}{k}$, where r is the number of relevant records (Craswell 2009). P@k gives every record in the ranked list an equal weight. For instance, when calculating *P@1000*, the 1000th record in the ranked list has a equal weight to the 1st record. Nevertheless, in ranked retrieval systems, a greater emphasis should be placed on early ranks than on the later records.

To alleviate this issue, information retrieval metrics such as *Mean Average Precision at k (MAP@k)* (which are sensitive to the ranking order) could be utilised. More specifically, the relevant records that are ranked more highly contribute more to this metric than the relevant records that are ranked lower in the ranking list. *Mean Average Precision*

Total number of instances			**P**	**N**
			Actual class	
			positive	negative
p	**Predicted class**	positive	**TP**	**FP**
n		negative	**FN**	**TN**

FIGURE 3.11: Confusion matrix

(MAP) denotes the arithmetic mean of average precision (AP) values over a set of n query topics as defined using $\frac{1}{n}\sum_{n} AP_n$. This measure is widely used as the *de facto gold standard* in the evaluation of information retrieval systems (Beitzel et al. 2009*b*). Average precision in MAP denotes the mean of the precision scores obtained after each relevant record is retrieved. In essence, this measure combines both *recall* and *precision* in the retrieval results, as defined using $\frac{\sum_r P@r}{R}$ where r represents the rank of each relevant record, R is the total number of relevant records, and $P@r$ denotes the precision of the top r retrieved records (Zhang & Zhang 2009*a*).

While $MAP@k$ showcases the *overall performance*, *Geometric Mean Average Precision at k (GMAP@k)* examines whether a model demonstrates *consistently good performance* across all queries (Beitzel et al. 2009*a*). $GMAP$ is defined as $\sqrt[n]{\prod_{n} AP_n}$, where AP is the average precision over n queries. This is alternatively calculated as the arithmetic mean of logs as expressed in $\exp\left(\frac{1}{n}\sum_{n}\log AP_n\right)$ (Beitzel et al. 2009*a*). To avoid logs of 0.0, AP scores lower than 0.00001 are set to 0.00001 (Voorhees 2006).

3.5.6.2 Classification Component

Classification problems attempt to determine the *characteristics* (or *features*) that correctly distinguish the class to which each of the test instances belongs (discussed in Section 3.5.1). The performance metrics used to compare classification performance are typically represented using elements in the confusion matrix, which is generated by the machine learning model on a test sample (Lemnaru 2012). Figure 3.11 denotes the template of a confusion matrix for a two-class classification problem, where the class of an instance is either positive or negative (discussed in Section 3.5.1).

In the confusion matrix, columns represent actual classes, while rows represent the predicted classes. The number of instances in the test sample is depicted on the top of the confusion matrix, where P is the total number of positive instances and N is the total number of negative instances. The number of instances predicted by the model in each class is shown in the left of the confusion matrix, where p is the total number of instances predicted to be positive and n is the total number of instances predicted to be negative. *True Positives (TP)* denotes the number of instances correctly predicted to be positive examples. *False Negatives (FN)* denotes the number of positive instances predicted to be negative. Similarly, *True Negatives (TN)* is the number of correctly predicted negative instances, and *False Positives (FP)* denotes the number of negative instances predicted to be positive. The *True Positive rate (TP_{rate})*, which is represented as $TP_{rate} = \frac{TP}{TP+FN}$, depicts the rate at which the positive class is recognised. This is also known as *recall* or *sensitivity* (Zhang & Zhang 2009*d*). The corresponding metric of the negative class is the *true negative rate (TN_{rate})*, which is measured as $TN_{rate} = \frac{TN}{TN+FP}$. This is also known as *specificity* and indicates the number of negative instances that are correctly detected. The purpose of *Positive Predictive Value (PPV)* and *Negative Predictive Value (NPV)* is to quantify how many instances which are detected as belonging to a given class actually represent that class. PPV, which is also known as *precision*, measures the number of actual instances identified as positive (i.e., $PPV = \frac{TP}{TP+FP}$) (Zhang & Zhang 2009*c*). NPV denotes the number of negative instances that are correctly detected out of all instances predicted to be negative (i.e., $NPV = \frac{TN}{TN+FP}$).

From the elementary performance metrics discussed above, several *composite measures* have been constructed, such as *F-measure* and *ROC curves*. F-measure (more specifically, F_1) is the harmonic mean of precision and recall, and is denoted as $2\times\frac{precision\times recall}{precision+recall}$ (Zhang & Zhang 2009*b*). The ROC (Receiver Operating Characteristic) curve plots *true positive rate* (or *sensitivity* denoted as $\frac{TP}{TP+FN}$) against *false positive rate* (or *1-specificity* denoted as $\frac{FP}{FP+TN}$), at different classification thresholds (Tan 2009). Typically, a good classification model should reside in the upper left region of the plot (Figure 3.12). Point (0,0) indicates a model that detects all instances as negative. Point(1,1) denotes all instances as positive, while a random classifier signifies y=x curve. The ideal classification model generates the point (0,1) indicating that its false positive rate is zero (i.e., none of the negative instances are predicted to be positive) and the true positive rate is equal to 1 (i.e., every positive instance is identified). The AUC (Area Under the ROC Curve)

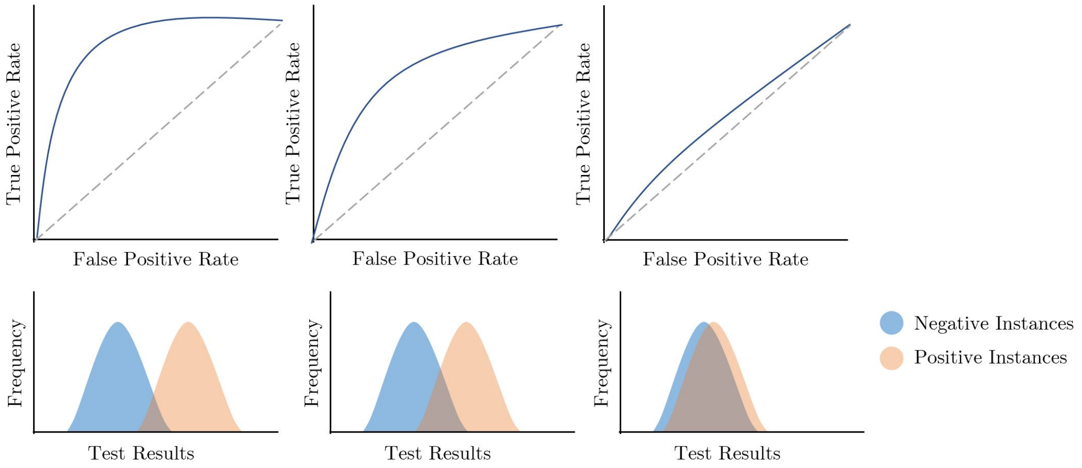


FIGURE 3.12: Classifier performance with ROC curve

is the aggregated measure of the ROC curve that indicates the performance across all possible thresholds . More specifically, the AUC denotes the entire two-dimensional area under the ROC curve from point (0,0) to (1,1). Simply put, it indicates the probability with which classifier will rank a random positive instance more highly than a random negative instance.

In the classification setup of this thesis, *negative class* denotes topics that are *not interesting or meaningless* in the knowledge discovery process. In contrast, the *positive class* denotes topics that are *potential novel knowledge linkages*, as discussed at the outset of this section. In a typical LBD workflow, these two classes are equivalent to *filtering* and *ranking*, respectively (Figure 3.6). Therefore, given an instance, it is important to understand how well the model is capable of distinguishing its class (whether it is an uninteresting, meaningless concept used in *filtering* or a potential novel knowledge linkage used in *ranking*). To facilitate this, the *weighted average* composite measures of precision, recall and F-measure are utilised (Zhou et al. 2016, Mohammed & Omar 2020, Maharjan et al. 2018). More specifically, these composite measures provide the opportunity to get an understanding of the overall performance of an LBD model in terms of how well each instance was classified in the testing sample. This thesis also considers AUC in order to quantify how well the LBD model separates negative instances from positive instances.

3.6 Baselines

As in previous LBD research (Jha et al. 2018, Jha, Xun, Wang & Zhang 2019, Lever et al. 2018, Xun et al. 2017), this thesis considered the following eight baseline algorithms in order to facilitate a comparison of the proposed LBD models' performance (discussed in Chapters 5 and 6). The discussion in this section focuses on two aspects: the characteristic(s) that are considered in each baseline and the motivation for using each selected baseline.

Arrowsmith (AR) is the oldest LBD project in the discipline. It was initiated by the pioneers of the LBD field and is considered to be the most popular and well-maintained LBD system (Sebastian et al. 2017*a*). It is reported that Arrowsmith has approximately 1200 unique monthly users. The system uses seven features to decide potential novel knowledge linkages (Torvik & Smalheiser 2007), namely *does the B concept occur in more than a paper in A and C literature? (f_1)*, *do the sub-literatures AB and BC have any common MeSH terms? (f_2)*, *does B concept has a mapping to at least one semantic category in UMLS? (f_3)*, *does the B concept demonstrate a high literature cohesion score? (f_4)*, *does the B concept extremely common or extremely rare in MEDLINE? (f_5)*, *does the first occurrence of B concept recent in MEDLINE? (f_6)*, and *does the B concept highly characteristic in A and C literature? (f_7)* (Torvik & Smalheiser 2007). The features proposed in Arrowsmith include both *global* (i.e., f_3, f_4, f_5 and f_6) and *local* (i.e., f_1, f_2 and f_7) properties of the literature. Even though the feature f_6, which is the first occurrence of the B-concept in the literature, could potentially be used to obtain some basic understanding of the temporal aspect of the concept, the remaining features in Arrowsmith are based on *static cues* taken from the literature. Therefore, the use of this baseline in this thesis provides the opportunity to assess whether *meticulous temporal cues* really matter in determining potential novel knowledge linkages. To facilitate the comparison, the features proposed in Arrowsmith are used in the same machine learning setting that is used in this thesis (as discussed in Section 3.5). Since Chapters 5 and 6 are based on *MeSH keywords* (as discussed in Section 3.3.1), feature f_3 will be meaningless in this setting, since *MeSH* terms are integrated into *UMLS*; thus, the MeSH terms are assigned to UMLS semantic categories (Bodenreider 2004). With that in mind, this feature is removed to facilitate a fair comparison of results.

Bitola (BI) (Hristovski et al. 2001) is one of the longest-established and most popular LBD tools in the discipline. It uses association rule mining (more specifically, *confidence* or *support*) to rank potential knowledge linkages. *Confidence* measures the percentage of all records in which A appears that contain B, whereas *support* indicates the number of records in which A and B co-occur (Hristovski et al. 2005). These two measurements can be denoted in the form: $\frac{|D_A \cap D_B|}{|D_A|}$ and $\mid D_A \cap D_B \mid$, respectively, where D_i is the set of records in which the term i is included (Yetisgen-Yildiz & Pratt 2009). In the default setting, Bitola LBD system uses *confidence* for ranking (Hristovski et al. 2001). The use of this baseline model provides the opportunity to understand whether a *single conventional statistical metric* would be sufficient in the knowledge discovery process or whether the knowledge discovery process favours the integration of *multiple semantically infused features* to elicit potential novel knowledge linkages with high precision.

Dynamic Embeddings (DE) (Xun et al. 2017) represent a recently developed LBD algorithm that mainly relies on diachronic word embeddings. This study is based on three global semantic measures, *local topic's cosine similarity with topic A and C at cut-off timestamp t*, *trend between local topic, topic A and topic C with reference to the timestamp of first occurrence and cut-off timestamp t*, and *generality of the local topic* (Xun et al. 2017). Even though this study undoubtedly provides a novel perspective to the LBD field through the use of diachronic semantic inferences, this study suffers from several inherent limitations. One of these is its relatively shallow temporal component. For instance, to measure the temporal trend, this study simply considers the first and last values in the diachronic vector spaces, ignoring the concept's behaviour in the remaining timestamps. Secondly, the number of semantic measures incorporated in this study is limited to three. Therefore, the use of this LBD algorithm as a baseline assists in the task of assessing whether a *meticulous temporal analysis* with *multiple temporal characteristics* is required in the LBD process.

Static Embeddings (SE) baseline algorithm uses word embeddings that are generated without integrating any temporal analysis, where the bridge terms are ranked using cosine similarity (Jha et al. 2018, Jha, Xun, Wang & Zhang 2019). Since this baseline does not incorporate any temporal cues in the vector space, it provides the opportunity to assess whether *static similarity analysis* among words alone is sufficient in the LBD knowledge discovery workflow.

Term-frequency and Inverse-document frequency (TI) is a popular metric that represents the importance of a concept to a document in the corpus (Jha et al. 2018, Liu & Rastegar-Mojarad 2016). This metric has been widely used in the LBD literature since the inception of LBD; thus, it is selected as a baseline in this thesis (Yetisgen-Yildiz & Pratt 2009, Ittipanuvat et al. 2014). The use of *TF-IDF* as a baseline helps to identify whether such a *standard statistical measure* alone is capable of capturing novel knowledge linkages, or whether the knowledge discovery process requires more problem-specific measures to detect latent novel knowledge linkages with high precision.

More recently, there has been growing research interest in incorporating *link prediction techniques* in the LBD field (Kastrin et al. 2016, Yang et al. 2017, Li 2020). More specifically, these LBD studies have attempted to predict the links between terms that are not present in the current timestamp, but that have a tendency of occurring in the future. With these studies in mind, this thesis uses three popular link prediction techniques that have been employed in prior LBD studies (*Common Neighbours (CN)*, *Jaccard's Index (JI)* and *Preferential Attachment (PA)*) as baselines. As in the case of TF-IDF, the use of these link prediction methods helps to gauge whether the direct use of such standard measures alone would be sufficient to discover potential novel knowledge linkages, or whether the knowledge discovery process favours the development of methods tailored to the focus of LBD. The three link prediction metrics can be denoted in the form: $\mid \Gamma(x) \cap \Gamma(y) \mid$, $\mid \frac{\Gamma(x) \cap \Gamma(y)}{\Gamma(x) \cup \Gamma(y)} \mid$, and $\mid \Gamma(x) \mid \times \mid \Gamma(y) \mid$, respectively, where $\Gamma(i)$ denotes a set of terms that co-occur with the term i (Gao, Musial, Cooper & Tsoka 2015, Jha, Xun, Wang & Zhang 2019, Lever et al. 2018).

A summary of the selected eight baselines is outlined in Table 3.5, along with the chapter number in which they will be utilised. Note that some baselines are not used in both Chapters 5 and 6 due to their incompatibility with the setting and focus of the chapters. Details on the incompatibility of these baselines are discussed in the *Experimental Setup* section of Chapters 5 and 6. The selected eight baselines include the only two long-established LBD models that are still available online for public use, which are *Arrowsmith*[4] and *Bitola*[5] (Kastrin & Hristovski 2020).

[4] http://arrowsmith.psych.uic.edu/arrowsmith_uic/index.html

[5] https://ibmi.mf.uni-lj.si/sl/bitola

Table 3.5: Summary of baseline models

Baseline	**Prominent Properties**	**Chapter**
Arrowsmith (AR)	• The oldest LBD project in the discipline • Considered to be the most popular and well-maintained LBD tool with approximately 1200 unique monthly users • Multi-characteristic • Global and local features • Static literature analysis	Chapter 5, Chapter 6
Bitola (BI)	• One of the long-established and popular LBD tool in the discipline • Co-occurrence frequencies • Local feature • Static literature analysis	Chapter 5
Dynamic Embeddings (DE)	• Considering recent advancements in word embeddings • Multi-characteristic • Global features • Dynamic literature analysis using temporal cues	Chapter 5, Chapter 6
Static Embeddings (SE)	• Considering recent advancements in word embeddings • Global feature • Static literature analysis	Chapter 5, Chapter 6
TF-IDF (TI)	• Widely used metric in the LBD literature • Co-occurrence frequencies • Static literature analysis	Chapter 5
Common Neighbours (CN)	• State-of-the-art link prediction technique • Neighbourhood analysis • Static literature analysis	Chapter 6
Jaccard's Index (JI)	• State-of-the-art link prediction technique • Neighbourhood analysis • Static literature analysis	Chapter 6
Preferential Attachment (PA)	• State-of-the-art link prediction technique • Neighbourhood analysis • Static literature analysis	Chapter 6

3.7 Summary

The purpose of this chapter is to outline the underlying research design of the remaining chapters of this thesis. Initially, the scope of the research and the connections between the subsequent four chapters of this thesis and the main components of the LBD workflow are discussed. The main scope of this thesis is to address the *input component*, the *discovery component* and the overall LBD workflow, with an emphasise on *reusability* and *portability*. Subsequently, the datasets and the test cases selected for the experiments (and the reasons behind these selections) are discussed. More specifically, this thesis uses *MEDLINE* as its main data source, due to the popularity of the database in the LBD literature. MEDLINE also contains more than 25 million timestamped article records, making it suitable for large-scale literature mining. With regard to test cases, this thesis considered real-world test cases reported in the LBD literature; these are commonly considered *golden datasets* for the purposes of evaluating results.

Subsequently, the evaluation framework adhered to in this thesis is discussed. The selection of the most suited evaluation technique was performed by initially identifying the criteria that constitute an ideal evaluation setting. Next, popular LBD evaluation techniques were cross-checked with these defined criteria to identify the evaluation technique that most closely resembles the ideal evaluation setting. Based on this assessment, *time-slicing* was selected as the main evaluation framework for this thesis. Time-slicing enables large-scale knowledge discovery through the incorporation of machine learning techniques. In the machine learning framework, the *legitimate novel knowledge linkages* identified through time-slicing can be considered *positive instances*, while *remaining local topics* are considered *negative instances*. In the conventional LBD workflow, *negative class* is equivalent to *filtering process*, while *positive class* indicates the potential candidates used for the *ranking process*. Subsequently, the machine learning framework that will be used in Chapters 5 and 6 is discussed. This discussion also covers the selection of metrics that can be used to evaluate the recommendation component and the classification component. These selections were made by contemplating the problem setting of LBD and the qualities of the LBD models which needed to be highlighted in the experiments. Finally, the baseline models used to compare the performance of the proposed LBD models are discussed. This discussion includes not only the characteristic(s) used in each baseline LBD model, but also the reason why this thesis selected it as a baseline.

Such a broader understanding of the baseline models is important to critically compare and discuss the results.

Throughout this chapter, design selections were made carefully based on logical reasoning. A strong evidence base was used to ensure that the best selections were made. The subsequent chapters of this thesis mainly rely on the research design selections discussed in this chapter. When design selections mentioned in this chapter are used in the subsequent chapters, this thesis makes relevant references to this chapter, indicating the relevant details and explaining why such selections were made.

Chapter 4

Input Types

4.1 Introduction

To initiate the LBD process, the user is required to input two scientific topics of interest A and C. The LBD model elicits potential new knowledge linkages between the two user-defined knowledge fragments that are most likely to occur in the future. For this purpose, the literature related to the two topics A and C is collected from a digital library that is collectively termed the *local corpus* (Figure 4.1). The local corpus represents the *input component* of the LBD workflow. This derived local corpus could consist of *different input types*. For instance, it could include data on *titles*, *keywords*, or even the scientific articles' *full content* in the literature database (Henry & McInnes 2017).

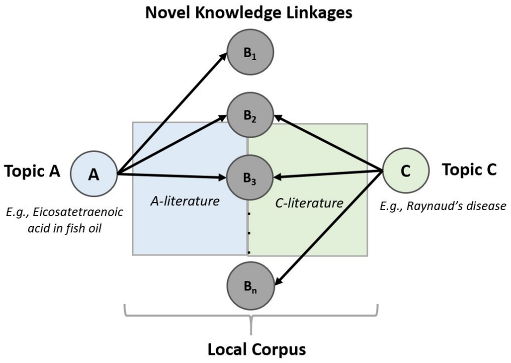


FIGURE 4.1: Input component of the LBD workflow

The input is one of the most critical components in the LBD workflow, as the entire knowledge representation and reasoning of the discovery process relies on it (Henry & McInnes 2017). As with other text mining tasks, low-quality input will negatively impact the LBD results and, ultimately, the decisions based on it (Corrales et al. 2015). Existing studies are not consistent in their choice of LBD input types, since different studies have picked different input types (e.g., *titles*, *abstracts* and *keywords*) for their LBD process. This varied selection of input types leads to the following question: *'which input types are best-suited to the LBD workflow?'*.

Despite the importance of LBD input in the overall knowledge discovery process, *no previous studies* have explicitly attempted to assess the suitability of different input types to the knowledge discovery process. Some prior LBD studies have *implicitly* compared the performance of their LBD models with various input types. For instance, Sebastian et al. (2017*b*) reported that they obtained better results using *titles* in comparison to *abstracts*. Nagarajan et al. (2015) mentioned that their LBD performances mainly depended on the *richness of the information* being used (i.e., with *more edges* in the knowledge network). However, these conclusions are potentially biased to their methodologies as they have not isolated the *input component* from their proposed discovery methodology during the evaluation. Otherwise stated, these conclusions may differ when a different discovery method is utilised in the LBD process. Thus, they may not necessarily provide insights that can be broadly applied to determine the suitability of each input type in the LBD workflow.

Selecting a suitable input type representing the LBD workflow's input component is not straightforward. This is because different *data fields* in research papers have their own *perspectives* and *information content*. For instance, Lee et al. (2015) have found that *keyphrases*, *citation relationships*, and *MeSH* reflect the views of *authors*, *citers*, and *indexers*, respectively. Kostoff et al. (2004) have identified that the *information content* in different data fields of the research papers varies; thus, the selection of the field depends on the objectives of the study. This highlights the importance of exploring the input component of the LBD workflow, as proper decisions about input types in the LBD context may ultimately contribute to developing better LBD models in the future.

With this goal in mind, this study performs a quantitative analysis of the *LBD input component* to understand its performance using *different input types*. The main research objective of this study is:

"to investigate the input component of the LBD workflow in order to deduce the suitability of different input types in the LBD process"

as defined at the outset of this thesis (i.e., *main research objective 2 (RO2)* in Chapter 1). To the best of our knowledge, this is the *first study* in the LBD discipline that explores the *input component* of the LBD workflow with the ultimate aim of deciding the suitability of each LBD input type in the knowledge discovery process. In doing so, this study attempts to answer the following main research question (*RQ2*): *'how can LBD input types be quantitatively assessed and compared so as to better understand their suitability in the LBD workflow?'*. This study is split into two stages to accomplish the main research objective and answer the primary research question, considering the following two sub research objectives.

- *RO2.1. Identifying the most influential characteristics that should be considered to understand the role of the input types in the context of LBD* (discussed in Sections 4.2 and 4.3).

- *RO2.2. Leveraging the identified characteristics to quantitatively assess and compare the input types, to validate their contribution to the overall knowledge discovery process of LBD workflow* (discussed in Section 4.4).

This chapter is organised as follows. Section 4.2 attempts to identify factors that differentiate each input type in the information retrieval cycle of the LBD process. In accordance with the knowledge gained from Section 4.2, Section 4.3 explores potential characteristics to facilitate a comprehension of LBD input types by exploring informativeness definitions in information theory. In this regard, this study explores *subjective definitions* of input types, since, in information retrieval cycles (as in the LBD workflow), subjective definitions of information are considered to be more meaningful and sensible than objective definitions (Tague-Sutcliffe 1992) (as discussed in Section 4.4). Following this notion, this section also discusses the main proposed subjective definition of information which is used in this study to quantitatively assess and compare LBD input types, and to decide their suitability to the LBD workflow. Section 4.5 outlines the existing LBD input types used in the literature in terms of their popularity and viability.

Section 4.6 extends the selected input types through the use of 'local neighbourhood', to verify whether such extensions of the input types would benefit the LBD process. Section 4.7 outlines the setup used in experiments in terms of input types, dataset and test cases. Section 4.8 presents the results alongside an extended discussion on the key observations, while also modifying the main subjective definition of information so as to unravel various other perspectives on input types. Furthermore, this section contains a compatibility check of the observations, drawing connections to the text mining literature to explain potential reasons for the key findings. Section 4.9 summarises the main findings of the study, along with its major contributions.

4.2 Information Retrieval Cycle of the LBD Workflow

LBD is an information retrieval process that is initiated when the user enters the scientific topics of interest into the LBD system. Subsequently, the user's input is transformed into an extended query to identify all the relevant literature that the user is interested in. For instance, consider a situation where a user inputted "fish oil" as an input topic. The query formulation component will identify every possible mapping to the user input such as synonyms, abbreviations and syntactic variations to ensure a high literature coverage in the subsequent phases. In the case of "fish oil", the potential mappings identified in the query formulation component would include terms such as *fish-oil*, *marine oil*, *fish oils*, and *fish liver oils*. These identified mappings are queried in a digital library to extract all the relevant literature related to the user input. This derived literature set from the digital library represents the *local corpus* (more specifically, the *input type*) in the LBD workflow. To perform a topic-level analysis, domain-related scientific topics need to be extracted from the derived textual data in the *local corpus*. Since the aforementioned example (i.e., *fish oil*) is from the medical domain, all the relevant medical-related scientific topics (such as *platelet aggregation*, *blood viscosity*, *vasodilation*, etc.) in the textual data needs to be extracted. These identified scientific topics are termed *local topics* since they are extracted from the local corpus. Subsequently, these local topics are processed through the knowledge discovery component in order to detect potential new knowledge linkages that are most likely to occur in the future. Finally, these elicited novel knowledge linkages are provided as output to the user. Figure 4.2 illustrates the information retrieval cycle of the LBD workflow as discussed above.

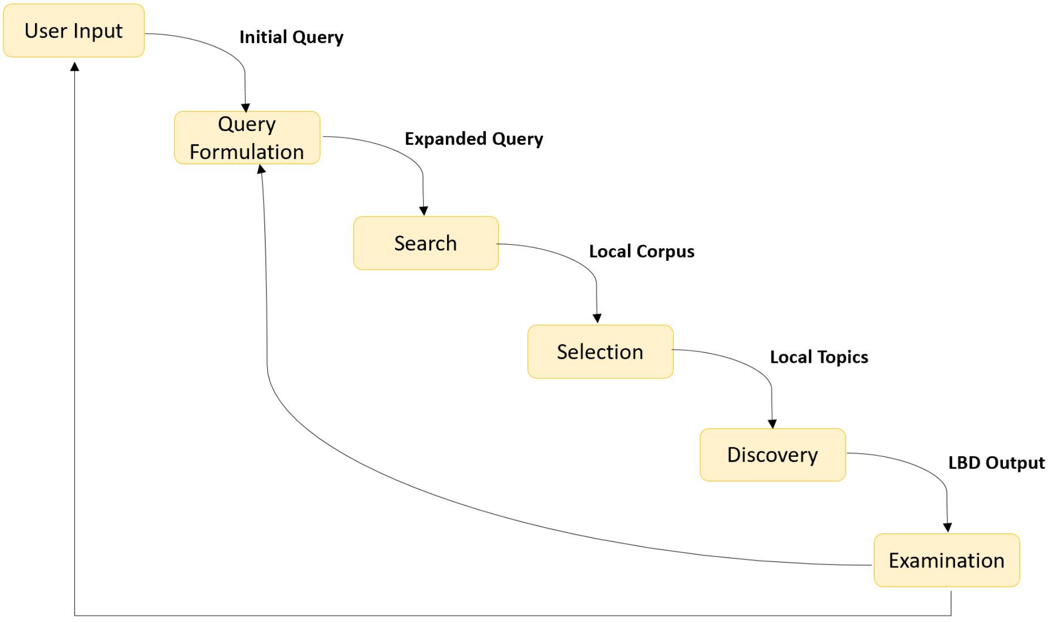


FIGURE 4.2: Information retrieval cycle in the LBD workflow

When closely inspecting this information retrieval cycle, it is evident that the only difference between each input type in this cycle is the *'information'* that it provides to the knowledge discovery process, which will ultimately determine the output provided to the user. In other words, the process that each LBD input type flows through in the information retrieval cycle is identical, in spite of the difference in content (or information) that it carries across the components in the LBD workflow.

Thus, it can be deduced that the input types that are *'most informative'* (i.e., the input types that demonstrate the 'greatest degree of *information richness*') in the LBD workflow are the most suitable input types. Nevertheless, it is difficult to quantify or make appropriate decisions about input types without defining what it means by *'informativeness'* (or *'information richness'*) that it provides to the information retrieval cycle in the context of LBD.

4.3 Different Perspectives on Information Richness

Informativeness (or *information richness*) is the main factor that differentiates input types from each other in the information retrieval cycle of LBD (as discussed in Section 4.2). With that in mind, the most important characteristics in terms of comprehending the relative suitability of different input types should be those related to *informativeness* (or *information richness*). In this regard, the first question to emerge from this study was: *'how can one define the information that resides in input types?'*. Within information theory, there are two main viewpoints on the definition or understanding of information: *objective phenomena* and *subjective phenomena* (Hjørland 2007, Capurro & Hjørland 2003).

- *Objective phenomena:* The *objective definitions* consider information to be an attribute that is mainly based on the text (or record) itself (Tague-Sutcliffe 1992). Therefore, the idea of objective perspectives of information are *observer-independent* as well as *situation-independent* (Hjørland 2007).

- *Subjective phenomena:* On the contrary, *subjective definitions* consider information to be an attribute of the transaction between the text and the user, and what the user learns from reading the output (Tague-Sutcliffe 1992, Bates 2005). Therefore, the subjective understanding of information is user-centered while involving the information retrieval cycle. This is also known as the *situational understanding* of information.

These two key understandings of information influenced this study's definition of *informativeness* (or *information richness*) in the context of LBD input types. Decisions as to which perspective to choose will vary according to the goal of the study (Tague-Sutcliffe 1992). Therefore, this study investigates how each of the two perspectives on information can be transformed into the LBD context, to aid the process of choosing the most suitable perspective.

4.3.1 Objective Perspectives in the Context of LBD

The objective understanding of information has intrinsic value and a definite meaning, since it is user-independent and situation-independent (Pervez 2009). With this definition in mind, consider a situation in which *readability formulas* (as proposed in text mining literature (Shams 2014)) are used to assess input types. Even though such readability measures provide a quantitative metric that facilitates the comparison of input types to decide their suitability, they are solely based on the attributes of a *text*, as summarised below.

- *Flesch Reading Ease Score (FRES):* This measure is based on the average number of syllables per word and the average sentence length in a text.
- *SMOG Index:* This measure is based on the number of polysyllabic words in a text.
- *Gunning Fox Index:* This measure is based on the percentage of long words and the average sentence length in a text.
- *FORCAST Index:* This measure is based on the number of monosyllabic words in a text.

When such readability measures are used, they cannot tell us anything about the impact these input types had on the user, or whether the selected input types fulfilled the user's needs. Because of the *observer-independent* and *situation-independent* nature of such readability measures, they do not capture the way in which input types interact with the information retrieval cycle of the LBD process. Since objective measures (such as *readability measures*, as discussed above) only reflect the properties of textual elements in input types, rather than how these input types interact with the whole information retrieval process, their use in the context of LBD can be limiting.

4.3.2 Subjective Perspectives in the Context of LBD

The subjective understanding of information is situation-dependent, interpretive and constructivist (Pervez 2009). In the information retrieval context, where the information outputted from the system depends on the user's needs, *subjective definitions* of information are considered to be more reasonable and sensible than *objective definitions* (Tague-Sutcliffe 1992). The purpose of the information retrieval cycle of LBD is to inform users about potential latent knowledge linkages (i.e., *information-as-process* (Buckland 1991)) to support the user to gain new knowledge (i.e., *information-as-knowledge* (Buckland 1991)). As such, this indicates a case in which *subjective definitions* should be used to define informativeness (or information richness), in order to quantitatively validate the suitability of each input type. With this aim in mind, this study explores potential *subjective definitions* that consider the interactivities between texts and the user, in order to measure the *informativeness* or the *information richness* of the LBD input types.

4.4 Defining Information Richness in the Context of LBD

Since a subjective understanding of information is the best approach in the LBD scenario, this section explores potential *subjective definitions* that can be used to quantify *informativeness* or *information richness* for each LBD input type. Otherwise stated, the proposed metric should pay close attention to the information retrieval cycle in the context of LBD (i.e., *information-as-process*) and what users gained from it (i.e., *information-as-knowledge*). Nevertheless, it remains unclear which fundamental aspect(s) should be captured in order to quantify informativeness within the subjective phenomenon.

In this regard, this study revisited the main objective of this study, which is identifying the LBD input types that demonstrate maximum information richness (or informativeness), as discussed in Section 4.2. In essence, this can be viewed as an *optimisation problem*, where the most suitable LBD input type is the most optimised solution. With this objective in mind, this study leverages the *optimality theory* as the fundamental aspect (or focus point) of the subjective understanding of information to quantify the *information richness* of LBD input types.

The main inspiration for the proposed metric came from *optimal foraging theory*, which is based on a cost-benefit analysis (Stephens & Krebs 1986). Simply put, the goal of the theory is to assess the amount of resources consumed (i.e., *cost*) and resulting gains (i.e., *benefit*) in the information retrieval cycle. Originally, the idea came from a *behavioural ecology model* that predicts how animals behave when searching for food. In the case of a predator, it adopts an *optimality model* where with the lowest effort to obtain the maximum amount of energy. The process of gaining the highest benefit by spending the least amount of energy is called *optimal foraging*. This theory has also been widely used in the context of information-related research, namely *information foraging theory* (Pirolli 2007). Inspired by the key interpretation of the theorem, this thesis uses *optimal foraging* as the primary setting to assess *information richness*. More specifically, the intention is to measure which input types provide the *maximum benefit* at the *lowest cost* in the information retrieval cycle of the LBD workflow. The two main components of the foraging theory, *cost* and *benefit*, are analogically mapped to the information retrieval cycle of the LBD workflow, as described below.

4.4.1 Cost Assessment

In the context of LBD, the *cost* is mapped to the *number of local topics* in the local corpus. The main reason for this mapping is that the local topics selected from the input component are used as the main data source of the entire knowledge discovery process (Figure 4.3). These local topics consume both time (i.e., denoting *time complexity*) and space (i.e., denoting *space complexity*) in the LBD workflow, both of which can be analogously mapped to *energy* in optimal foraging theory.

4.4.2 Benefit Assessment

LBD is designed to infer potential novel knowledge linkages which have been previously unknown but are probably going to occur in the future. Thus, the *benefit (or the gain)* in the LBD workflow is the *number of legitimate novel knowledge predictions* (Figure 4.3). This demonstrates how the information retrieval cycle of LBD (i.e., *information-as-process*) helps users to gain or perceive knowledge (i.e., *information-as-knowledge*). Thus, it is fair to say that these legitimate novel knowledge predictions signify the interactivities between the LBD model and the user, and how satisfied the user was from the information retrieval output.

It should also be noted for something to be considered informative, several individuals need to agree that it is so (Buckland 1991). This is known as *information by consensus.* Thus, if one of the LBD model's predictions is considered a legitimate novel knowledge by a mere individual, this does not necessarily indicate *informativeness.* Thus, it is necessary to account for multiple users' consensuses in order to assess *benefit* (which is the *number of legitimate predictions*). Due to the time- and cost-intensive nature of such large-scale user studies, this study considers *time-slicing* as a substitute for user studies to denote the *legitimacy* of a knowledge linkage (discussed in Chapter 3). Since the number of times the proposed novel knowledge linkages have taken place in the *future* is incorporated in time-slicing, this method also caters to the need for *information by consensus.* In addition, time-slicing also ensures the reproducibility of results, which is lacking in actual user studies.

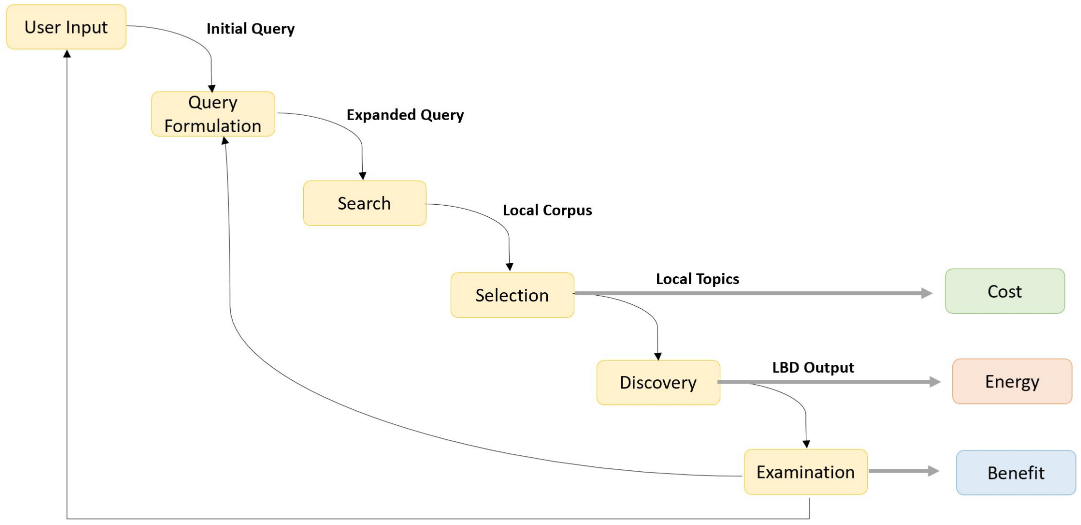


FIGURE 4.3: Subjective perspectives involving optimal foraging

TABLE 4.1: Mapping to the Optimal Foraging Theory (OFT)

OFT	**Mapping**
Cost	Number of local topics (depicting the amount of information extracted from each input type)
Energy	Computations in the knowledge discovery process (denoting the time and space complexity of the discovery process in the information retrieval cycle)
Benefit/Gain	Number of legitimate novel knowledge linkages (signifying how satisfied the user was with the information retrieval output)

4.4.3 Optimal Foraging

To summarise, this study maps foraging theory setup to the process of measuring the *Information Richness (IR)* of LBD input types, as summarised in Table 4.1. Succinctly, this study attempts to identify the input types that provide the *greatest benefit* by consuming the *least energy* (i.e., *optimal foraging behaviour*), as denoted in equation 4.1. More specifically, the notion of optimal foraging answers the following question: *'how much important information does the information retrieval cycle (i.e., information-as-process) provide to the user (i.e., information-as-knowledge)?'*. Figure 4.3 illustrates how the key ingredients of optimal foraging interact with the information retrieval cycle to preserve the subjective perspective of *information richness*.

$$IR\ (input\_type) = \frac{\#legitimate\ novel\ topics}{\#local\ topics} \times 100 \tag{4.1}$$

4.5 Input Types

The LBD literature has utilised different variants of input types in its LBD models in order to facilitate the knowledge discovery process. These variants include *title only*, *title and abstract*, *full-text*, *keywords*, and even some highly specialised input type variants, such as *clinical patient records* and *case reports* (discussed in Chapter 2). Among these variants, *title and abstract* are the most commonly selected. Nevertheless, the pioneers of the LBD disciple have continuously employed *only the title* of research publications as their LBD input since the inception of the LBD field (Swanson & Smalheiser 1997). Following this notion, *Arrowsmith* (the most popular and well-maintained LBD tool in the discipline (Sebastian et al. 2017*a*)), only supports the analysis of titles to elicit new knowledge (Torvik & Smalheiser 2007). To date, the most widely used keyword type in the LBD literature is *Medical Subject Headings (MeSH)*. MeSH is a controlled vocabulary thesaurus maintained and updated annually by the *National Library of Medicine (NLM)* (Lipscomb 2000). There are several LBD studies reported in the literature that have used the *full-text* of articles as their input (Lever et al. 2018). However, most APIs of literature databases merely support metadata retrieval; thus, the use of full-text may limit the applicability of the LBD system in real-world settings (Cohen & Hersh 2005). Input types which are rarely used in the LBD discipline include *selected articles only* (Cameron et al. 2015), *other metadata* (Kostoff 2014), and *non-traditional input types* (Bhattacharya & Srinivasan 2012). This study picked the three most *popular* and *feasible* input types for our investigations: *title only*, *title and abstract*, and *MeSH keywords*.

4.6 Influence of Local Neighbourhood

Given the novel advancements in *word embedding techniques*, recent LBD studies have paid special attention to integrating the *local semantic neighbourhood* into the analysis in the LBD workflow (Jha et al. 2018, Jha, Xun, Gopalakrishnan & Zhang 2019). In the same spirit, this study also aimed to verify whether the addition of local neighbouring research publications to the selected three input types: *title only*, *title and abstract*, and *MeSH keywords* would benefit the knowledge discovery workflow. To facilitate the inclusion of such neighbouring documents, some method is required to identify which documents are the most similar to the local corpus. In this regard, this study uses novel advancements in *document embeddings* that were emerged due to the success of modern word embedding techniques (such as *word2vec*). More specifically, this study employs the popular *doc2vec document embedding* technique to identify semantically similar neighbouring documents (Le & Mikolov 2014). *Doc2vec* is an extended version of *word2vec* that determines an adequate d-dimensional and continuous vector for each *document* (or *paragraph*), while preserving semantic relationships among the documents (or paragraphs) in the corpus (Kim et al. 2019).

To facilitate the identification of local semantic neighbouring documents, first, this study learnt the document embeddings of the *entire literature in the digital library* using the Doc2vec model (more specifically, using the *'distributed memory'* variant, since it has been found to work well in most situations (Le & Mikolov 2014)). Subsequently, the nearest k neighbours of the original articles in the local corpus were added to each selected input type: *title only*, *title and abstract* and *MeSH keywords.* These additional input types constructed using the nearest local neighbourhood are referred to as the *extended input types* for brevity.

4.7 Experimental Setup

This section is dedicated to describing the experimental setup to which this study adheres to evaluate the information richness of the LBD input types. To this end, the first part of this section describes the different input type variants incorporated in this study, while the latter part discusses the main dataset and test cases used.

4.7.1 Input Type Variants

This study uses two different k values (*5* and *10*) to construct extended datasets (as discussed in Section 4.6). In summary, the study intends to analyse nine variants of the LBD input types, as summarised in Table 4.2.

Table 4.2: Selected input type variants

Dataset Type	**k value**	**Input Type Variant**
Default datasets	$k = 0$	1. title only (T) 2. title and abstract (TA)

		3. MeSH keywords (K)
Extended datasets	$k = 5$	4. title only (*Ex5-T*) 5. title and abstract (*Ex5-TA*) 6. MeSH keywords (*Ex5-K*)
Extended datasets	$k = 10$	7. title only (*Ex10-T*) 8. title and abstract (*Ex10-TA*) 9. MeSH keywords (*Ex10-K*)

4.7.2 Dataset and Test Cases

This study uses the entire *MEDLINE* literature repository to extract local corpora, construct document vectors (discussed in Section 4.6), and determine the legitimacy of the novel knowledge linkage. The following five test cases are used to evaluate the *information richness* of each input type. Further details on these aforementioned selections are described in Chapter 3.

- *Fish-Oil (FO)* and *Raynaud's Disease (RD)* (Swanson 1986)
- *Magnesium (MG)* and *Migraine Disorder (MIG)* (Swanson 1988)
- *Somatomedin C (IGF1)* and *Arginine (ARG)* (Swanson 1990*a*)
- *Alzheimer's Disease (AD)* and *Indomethacin (INN)* (Smalheiser & Swanson 1996)
- *Schizophrenia (SZ)* and *Calcium-Independent Phospholipase A2 (PA2)* (Smalheiser & Swanson 1998)

4.8 Results and Discussion

This section assesses the information richness of the selected input type variants to analyse their foraging behaviours. Moreover, this section also redefines the proposed information richness metric to capture several other perspectives of the input types to verify whether the observed foraging behaviours are consistent with these perspectives. The latter part of this section draws connections with the findings and conclusions reported in the text mining literature. This allows for a description of the observed foraging behaviours of the main input types.

4.8.1 Information Richness (IR)

This section uses equation 4.1 to assess the informativeness (or information richness) of the selected nine variants of LBD input types. Table 4.3 outlines the *information richness (IR)* scores obtained for each of the nine input type variants in the context of the five golden test cases. When analysing Table 4.3, it is evident that the input type *title and abstract* consistently achieved the highest IR score across all the datasets. Due to the independence of the test cases, this thesis also analysed how the IR score correlates with the sizes of the local corpora. This yielded -0.483 of *Pearson's correlation coefficient* for *title and abstract* that demonstrates that the IR score is marginally sensitive to the size of the local corpus in each test case. The second highest IR score was obtained through the use of *MeSH keywords.* The mean IR score increase of *title and abstract* over *MeSH keywords* was 10.7%. Furthermore, it was evident that the use of *only titles* yielded the lowest IR out of the three main input types.

A similar IR score pattern was observed for *extended* input types: *Ex5* and *Ex10.* In other words, the IR scores of the main three input types occur in the following order (from highest to lowest): *title and abstract*, *MeSH keywords*, and *title only* for both the *Ex5* and *Ex10* datasets. Overall, the involvement of neighbouring documents reduced the IR score of the three main input types. In other words, the IR score was negatively correlated with the number of k nearest neighbours added to the original local corpus. This study observed an average *Pearson's correlation coefficient* of -0.942 between k (i.e., for k values 0, 5 and 10) and *IR score* for *title and abstract.* Overall in this experimental setup, the most *optimal foraging behaviour* was achieved using *title and abstract* as the LBD input type, and the second-best optimal foraging was obtained through the use of *MeSH keywords.*

4.8.2 Intrigue Information Richness

Despite the consistency of the IR score based patterns observed over the five golden test cases (Table 4.3), this study aimed to further confirm the observed optimal foraging behaviours of the input type variants by disentangling IR score in several other perspectives. To this end, the following questions emerged: 1) what input types contain the highest number of *intriguing novel knowledge topics* (not just the count of novel knowledge linkages, as captured in equation 4.1)?, and 2) does the inclusion of implicit neighbouring documents compensate for its low IR gain by increasing the opportunity to include more *intriguing novel knowledge topics*?

To answer these questions, the *intrigue* score of a legitimate novel topic (n) for the two input topics A and C was calculated using equation 4.2 in the *post-cut-off segment* (as in previous LBD studies (Jha et al. 2018, Xun et al. 2017)). That is, all the legitimate novel topics were ranked using the scores gained from equation 4.2, where the topmost topics reflected the most intriguing new knowledge (Jha et al. 2018, Xun et al. 2017).

$$rank\_score\ (n) = \frac{\#(n, A) + \#(n, C)}{\#n} \times 100 \tag{4.2}$$

The *total intrigue* for each test case was calculated using the *cumulative gain* of scores derived from equation 4.2. Subsequently, this study redefined equation 4.1 using the derived total intrigue score from cumulative gain (equation 4.3).

$$intrigue\_IR\ (input\_type) = \frac{total\ intrigue}{\#local\ topics} \times 100 \tag{4.3}$$

The results of the intrigue IR are reported in Table 4.4. This study observed similar patterns as those in Table 4.3 by using *intrigue IR*. As with the IR score, the input type *title and abstract* consistently engendered the highest *intrigue IR* across all the golden test cases. Therefore, based on the evaluation results, this study can confirm that using the *title and abstract* in the LBD workflow not only ensures the maximum IR in terms of legitimate novel topic *count*, but also the highest *intrigue score* for these legitimate novel topics. As in the previous evaluation setting, the use of *MeSH keywords* resulted in the second-highest *intrigue IR* across the datasets. The lowest *intrigue IR* was obtained when *titles* were used as input type.

The observations pertaining to the extended datasets are compatible with those from the previous evaluation setting. More specifically, the three main input types were in the following order (from highest to lowest): *titles and abstracts*, *MeSH keywords* and *titles only* in both the extended datasets: *Ex5* and *Ex10*. Furthermore, this evaluation setting also confirms that the inclusion of neighbouring documents to the main input types is not rewarding, since the inclusion of these documents consistently results in a loss for every test case.

4.8.3 Average Intrigue Score

This study also analysed the *average intrigue score* of legitimate novel topics for each input type, as defined in equation 4.4. The main reason for conducting this analysis was to verify whether *extended input types* dilate the opportunity of including the *most intriguing novel topics*, which could possibly indemnify the constant IR loss these extended input types incur.

$$average\_intrigue\_score = \frac{total\ intrigue}{\#legitimate\ novel\ topics} \tag{4.4}$$

Table 4.5 summarises the results obtained for the five golden test cases in this analysis. Overall, using *title and abstract* as the input type resulted in the maximum average intrigue score. As was the case under the previous evaluation settings, *extended input types* exhibited the minimum results in comparison with their main input type. This further confirms that the integration of the local neighbourhood not only incurs IR loss, but also lowers the average intrigue score.

4.8.4 Key Observations

Succinctly, the input type *title and abstract* conclusively reported the highest *IR*, *intrigue IR*, and *average intrigue score* in all three evaluation settings. The consistent optimal foraging behaviour of the *title and abstract* in all three evaluation settings confirms that it is the most suitable LBD input type. This study observes that *MeSH keywords* are the second-best LBD input type, since they often achieved the second highest optimal foraging behaviours. Overall, *titles only* demonstrated the least optimal foraging behaviours. Furthermore, the evaluation results indicate that the inclusion of local neighbouring documents to the input types was not rewarding, as they consistently demonstrated a *loss* in each evaluation metric. More specifically, the foraging behaviours can be placed in the following order (from highest- to lowest-performing): *TA*, *K*, *T*, *Ex5-TA*, *Ex5-K*, *Ex5-T*, *Ex10-TA*, *Ex10-K*, and *Ex10-T*.

4.8.5 Compatibility Check

This section explores findings/conclusions relating to the selected three main input types as reported in the literature, in order to locate potential reasons for the observed foraging behaviours. Since such input-based discussions are rare in the LBD literature, this study mainly relies on the studies reported in the text mining literature to build this discussion.

4.8.5.1 Titles Only

The following three findings relating to titles (i.e., *limited character length*, *inverse impact and influence*, and *single characteristic*) could support why we observed titles to have the lowest information richness in each experimental setup (i.e., the least optimal foraging).

- *Limited character length:* Titles have a limited character length (or word length) (Moattarian & Alibabaee 2015, Nagano 2015). For instance, Hudson (2016, 2017) has analysed the *average character length* of titles in numerous disciplines and observed that the longest character length for titles occurs in disciplines such as *public health* (117.1 characters), *clinical medicine* (113 characters), and *agriculture* (110.4 characters), whereas the shortest occurs in disciplines like *philosophy* (51.1 characters) and *economics* (66 characters) (Hudson 2016). It is interesting to see that even the longest title is about *41.8%* of the length of the longest possible Tweet, indicating the potential paucity of information or facts that can be conveyed through titles.

- *Inverse impact and influence:* It has also been identified that using *long titles* or *wider diversity of concepts in titles* can adversely affect the *impact* and the *influence* of research publications (Hudson 2016, Paiva et al. 2012, Milojević 2017, Elgendi 2019, Jamali & Nikzad 2011, Subotic & Mukherjee 2014). Such findings may discourage researchers from including a large number of details in their titles, which may further reduce the possibility of capturing rich information through knowledge discovery.

- *Single characteristic:* There are different classifications of title types (Bahadoran et al. 2019). For instance, Hartley (2008, 2007) recognises 13 types of titles, including titles with *a general subject*, *a specific theme*, *a controlling question*, *findings*, *an indication of an answer to a question*, *an indication of the direction of an argument*, *an emphasis* *on methodology*, *guidelines and/or comparisons*, *a bid for attention*, *alliteration*, *literary elements*, *puns* and *mystifying utterances*. Therefore, it is fair to conclude that the title always informs one single characteristic of the study (e.g., either *findings* or *methodology*), which may not be sufficient for the process of knowledge discovery. Moreover, title types that contain literary elements, humour, irony, or puns might not reveal details that are pertinent to knowledge discovery, since machines are not as intelligent as humans when it comes to understanding the meaning conveyed through them.

4.8.5.2 MeSH Keywords

The following factor may often have influenced MeSH keywords to manifest the second highest *optimal foraging*.

- *Manual indexing:* MeSH terms are manually assigned to research papers by trained indexers with the required qualifications (Lipscomb 2000). Since MeSH keywords are selected based on a systematic procedure by subject matter experts, it is safe to assume that they represent the important content of a research paper (Jha et al. 2018).

Typically, MeSH is limited to 10-12 terms per each article (Chapman 2009). This may be the reason why it did not surpass the foraging behaviours of title and abstract.

The other form of keywords available for research papers (in addition to the indexed keywords such as MeSH) is *author keywords*, where the authors select keywords during their manuscript submissions (Oermann & Murphy 2018). Since MeSH keywords demonstrated the second-highest optimal foraging behaviour, this thesis also investigated whether author keywords would potentially demonstrate a similar information richness behaviour by making references to the text mining literature.

In the study of Névéol et al. (2010), they have identified that 60% of author keywords can be closely linked with the MeSH keywords. Even though there is a high similarity of author keywords and MeSH keywords in terms of their content, a small subset of the biomedical research papers has author keywords recorded. For instance, the cumulative

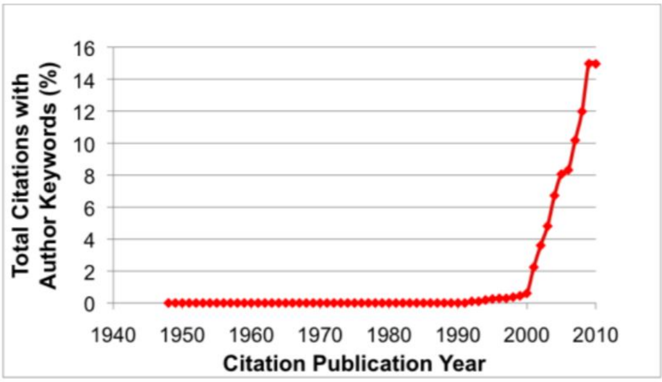


FIGURE 4.4: Cumulative percentage of papers with author keywords in PubMed Central (PMC) Open Access set (Névéol et al. 2010)

percentage of research papers with author keywords in the PubMed Central Open Access set is estimated to be nearly 15% as of 2010 (Figure 4.4). This thesis observed a similar conclusion outside the medical domain. More specifically, it has been identified that a large portion of research papers in non-medical domains (such as Ethnology, Economics, Physics, Sociology, Library and Information Science (LIS), Fluids & Plasma, and Acoustics) do not also have author keywords (Mao et al. 2018).

The above-discussed *limited availability* of author keywords (in both medical and non-medical domains) suggests that using author keywords in LBD workflow may not necessarily demonstrate a high information richness, as shown by MeSH keywords.

4.8.5.3 Title and Abstract

The following factors may have caused the abstracts to demonstrate continuous optimal foraging in every evaluation setting: *well-structured elements*, *handy synopsis of a paper's content* and *length*.

- *Well structured elements:* Unlike titles, there are various standards that authors should follow when constructing abstracts for research papers (e.g., *American National Standards Institute (ANSI)*) (Tenopir & Jasco 1993, Hartley 2008). In accordance with the ANSI standards, research papers should include *informative abstracts*. These abstracts are considered a *condensed version* of the important ideas presented in the paper, incorporating the following elements: *purpose*, *methodology*, *results* and *conclusions* (Tenopir & Jasco 1993, Hartley 2008). Thus, it is fair to say that the abstract contains the *main content of the paper*, yet in a concise manner.

- *Handy synopsis of the paper content:* It has been identified that abstracts (analysed from *1930-2013*) are becoming more *generous* (or *representative*) with time, and cannot merely be considered 'teasers' (Ermakova et al. 2018). The generous (or representative) nature of abstracts in comparison with their corresponding full-texts may have influenced abstracts to exhibit consistent optimal foraging behaviours.
- *Length:* In regard to the length of abstracts for research publications, the ANSI recommendation is 250 words (Tenopir & Jasco 1993). Thus, abstracts have a greater chance of providing rich information in the knowledge discovery process than input types such as *titles*.

In addition to the aforementioned findings and conclusions, this study observes that the text mining community has identified that the *readability* of abstracts is lower (i.e., *text difficulty* is high) across all disciplines using measures such as the *Flesch Reading Ease score* (Gazni 2011, Hartley et al. 2003). However, to quantify text difficulty, these readability scores mainly rely on metrics like the average number of words per sentence and the average number of syllables in words, rather than the semantic aspects of abstracts (Shams 2014, Farr et al. 1951). Based on the observations in our study, it can be concluded that the readability of abstracts is not an important consideration in the LBD workflow. The main reason for this could be that the readability scores used in the text mining community are mostly syntactic and do not factor in semantic aspects of abstracts. This ensures that in complex reasoning tasks like LBD, semantic details are more important than syntactic details.

4.8.6 Limitations

Due to the time- and cost-intensive nature of large-scale user studies, this study utilised *time-slicing* as a substitute for *benefit* assessments. While the use of time-slicing enables the *replicability* of results and *information by consensus*, reliance on *co-occurrence* in time-slicing may introduce noise, since co-occurrence does not necessarily imply a legitimate relationship between two topics. Therefore, time-slicing is merely an approximated substitute for such large-scale user studies.

4.9 Summary

The input can be considered one of the most critical components of the LBD process, as the entire knowledge discovery depends on the content and quality of the input selection. However, different LBD studies have made use of different input types (e.g., *titles*, *abstracts* and *keywords*) (Henry & McInnes 2017). Choosing the most suitable input types is a key design decision, as input types should be able to convey the most important entities and relationships contained in an academic article, in order to permit efficient knowledge discovery (Henry & McInnes 2017). This indicates the need to assess the informativeness (or *information richness*) of inputs in order to choose the most suitable input types in the LBD workflow.

Accordingly, this study performed a large-scale quantitative assessment of nine variants of LBD input types, taking inspiration from the *subjective understanding of information* and *optimal foraging theory*. More specifically, amalgamating these two notions enabled an assessment of different LBD input types in the form of: *'how much important information does the information retrieval cycle (i.e., information-as-process) provide to the user (i.e., information-as-knowledge)?'*. In terms of the foraging behaviours, the input types can be ordered as follows (from highest to lowest): *title and abstract*, *MeSH keywords* and *titles only*. This study also observed that the inclusion of semantic neighbouring documents in the LBD workflow is ineffective due to their consistent loss of information richness scores. Lastly, a compatibility check was performed to explain potential reasons for the foraging behaviours observed in the three main LBD input types. To summarise, this study put forward the first paving stones on the path towards assessing and comparing input types. This process is crucial to the construction of better LBD models in the future.

4.9.1 Major Contributions

Through this study, this thesis was able to shed light on a new direction for the LBD discipline. The major contributions of this chapter are summarised below, and are discussed in detail in Chapter 8.

- Being the first study in the LBD discipline that comprehensively analyses and evaluates the input component of the LBD workflow.

- Proposing a novel perspective on assessing the *information richness* of LBD input types, taking inspiration from foraging theory and subjective understandings of information that make use of the information retrieval cycle of the LBD workflow.

TABLE 4.3: Assessing IR of the input types, as defined in equation 4.1

Input Type	FO-RD	MG-MIG	IGF1-ARG	AD-INN	SZ-PA2
T	$\frac{169}{680}\times100 = 24.85\%$	$\frac{882}{3538}\times100 = 24.93\%$	$\frac{1051}{2262}\times100 = 46.46\%$	$\frac{740}{4360}\times100 = 16.97\%$	$\frac{739}{4767}\times100 = 15.50\%$
TA	$\frac{1538}{2960}\times100 = \mathbf{51.96\%}$	$\frac{5665}{11601}\times100 = \mathbf{48.83\%}$	$\frac{6233}{9709}\times100= \mathbf{64.20\%}$	$\frac{6643}{14856}\times100 = \mathbf{44.72\%}$	$\frac{4581}{13556}\times100 = \mathbf{33.79\%}$
K	$\frac{914}{3062}\times100 = 29.85\%$	$\frac{3064}{9804}\times100 = 31.25\%$	$\frac{3819}{6437}\times100 = 59.33\%$	$\frac{4126}{9754}\times100 = 42.30\%$	$\frac{2567}{9334}\times100 = 27.50\%$
Ex5-T	$\frac{284}{4185}\times100 = 6.79\%$	$\frac{1292}{14037}\times100 = 9.20\%$	$\frac{1875}{10097}\times100 = 18.57\%$	$\frac{910}{17435}\times100 = 5.22\%$	$\frac{963}{21947}\times100 = 4.39\%$
Ex5-TA	$\frac{2320}{7171}\times100 = 32.35\%$	$\frac{7624}{24172}\times100 = 31.54\%$	$\frac{9192}{21228}\times100 = 43.30\%$	$\frac{7827}{29561}\times100 = 26.48\%$	$\frac{5607}{30816}\times100 = 18.20\%$
Ex5-K	$\frac{1336}{9544}\times100 = 14.00\%$	$\frac{3669}{15955}\times100 = 23.00\%$	$\frac{5719}{14571}\times100 = 39.25\%$	$\frac{4898}{17978}\times100= 27.24\%$	$\frac{2986}{18701}\times100 = 15.97\%$
Ex10-T	$\frac{321}{6232}\times100 = 5.15\%$	$\frac{1339}{18800}\times100 = 7.12\%$	$\frac{2012}{13949}\times100 = 14.42\%$	$\frac{932}{23122}\times100 = 4.03\%$	$\frac{973}{28571}\times100 = 3.41\%$
Ex10-TA	$\frac{2598}{9800}\times100 = 26.51\%$	$\frac{8048}{29323}\times100 = 27.45\%$	$\frac{9988}{26421}\times100 = 37.80\%$	$\frac{8054}{35608}\times100= 22.62\%$	$\frac{5745}{37377}\times100 = 15.37\%$
Ex10-K	$\frac{1402}{11309}\times100 = 12.40\%$	$\frac{3770}{16997}\times100 = 22.18\%$	$\frac{5954}{16009}\times100 = 37.19\%$	$\frac{4983}{19006}\times100 = 26.22\%$	$\frac{3047}{19866}\times100 = 15.34\%$

**The highest IR scores are highlighted*

Table 4.4: Assessing intrigue IR of the input types, as defined in equation 4.3

Input Type	**FO-RD**	**MG-MIG**	**IGF1-ARG**	**AD-INN**	**SZ-PA2**
T	$\frac{14.31}{680}\times100 = 2.10\%$	$\frac{626.23}{3538}\times100 = 17.70\%$	$\frac{495.44}{2262}\times100 = 21.90\%$	$\frac{462.88}{4360}\times100 = 10.62\%$	$\frac{519.13}{4767}\times100 = 10.89\%$
TA	$\frac{279.96}{2960}\times100 =$ **9.46%**	$\frac{5389.02}{11601}\times100 =$ **46.45%**	$\frac{9200.04}{9709}\times100=$ **94.76%**	$\frac{8538.15}{14856}\times100 =$ **57.47%**	$\frac{4789.24}{13556}\times100 =$ **35.33%**
K	$\frac{179.46}{3062}\times100 = 5.86\%$	$\frac{1709.53}{9804}\times100 = 17.44\%$	$\frac{3409.15}{6437}\times100 = 52.96\%$	$\frac{3380.60}{9754}\times100 = 34.66\%$	$\frac{1667.08}{9334}\times100 = 17.86\%$
Ex5-T	$\frac{20.12}{4185}\times100 = 0.48\%$	$\frac{745.04}{14037}\times100 = 5.31\%$	$\frac{773.07}{10097}\times100 = 7.66\%$	$\frac{519.38}{17435}\times100 = 2.98\%$	$\frac{631.81}{21947}\times100 = 2.88\%$
Ex5-TA	$\frac{343.51}{7171}\times100 = 4.79\%$	$\frac{6484.57}{24172}\times100 = 26.83\%$	$\frac{11553.05}{21228}\times100 = 54.42\%$	$\frac{9455.79}{29561}\times100 = 31.99\%$	$\frac{5215.52}{30816}\times100 = 16.92\%$
Ex5-K	$\frac{207.06}{9544}\times100 = 2.17\%$	$\frac{1889.42}{15955}\times100 = 11.84\%$	$\frac{4229.73}{14571}\times100 = 29.03\%$	$\frac{3717.61}{17978}\times100= 20.68\%$	$\frac{1813.16}{18701}\times100 = 9.70\%$
Ex10-T	$\frac{23.08}{6232}\times100 = 0.37\%$	$\frac{763.98}{18800}\times100 = 4.06\%$	$\frac{829.56}{13949}\times100 = 5.95\%$	$\frac{535.68}{23122}\times100 = 2.32\%$	$\frac{638.17}{28571}\times100 = 2.23\%$
Ex10-TA	$\frac{379.37}{9800}\times100 = 3.87\%$	$\frac{6782.82}{29323}\times100 = 23.13\%$	$\frac{12399.44}{26421}\times100 = 46.93\%$	$\frac{9697.73}{35608}\times100= 27.23\%$	$\frac{5344.08}{37377}\times100 = 14.30\%$
Ex10-K	$\frac{215.45}{11309}\times100 = 1.91\%$	$\frac{1940.32}{16997}\times100 = 11.42\%$	$\frac{4378.26}{16009}\times100 = 27.35\%$	$\frac{3763.80}{19006}\times100 = 19.80\%$	$\frac{1838.00}{19866}\times100 = 9.25\%$

**The highest intrigue IR scores are highlighted*

Table 4.5: Assessing average intrigue score of the input types, as defined in equation 4.4

Input Type	**FO-RD**	**MG-MIG**	**IGF1-ARG**	**AD-INN**	**SZ-PA2**
T	$\frac{14.31}{169} = 0.08$	$\frac{626.23}{882} = 0.71$	$\frac{495.44}{1051} = 0.47$	$\frac{462.88}{740} = 0.63$	$\frac{519.13}{739} = 0.70$
TA	$\frac{279.96}{1538} = 0.18$	$\frac{5389.02}{5665} = \mathbf{0.95}$	$\frac{9200.04}{6233} = \mathbf{1.48}$	$\frac{8538.15}{6643} = \mathbf{1.29}$	$\frac{4789.24}{4581} = \mathbf{1.05}$
K	$\frac{179.46}{914} = \mathbf{0.20}$	$\frac{1709.53}{3064} = 0.56$	$\frac{3409.15}{3819} = 0.89$	$\frac{3380.60}{4126} = 0.82$	$\frac{1667.08}{2567} = 0.65$
Ex5-T	$\frac{20.12}{284} = 0.07$	$\frac{745.04}{1292} = 0.58$	$\frac{773.07}{1875} = 0.41$	$\frac{519.38}{910} = 0.57$	$\frac{631.81}{963} = 0.66$
Ex5-TA	$\frac{343.51}{2320} = 0.15$	$\frac{6484.57}{7624} = 0.85$	$\frac{11553.05}{9192} = 1.26$	$\frac{9455.80}{7827} = 1.21$	$\frac{5215.52}{5607} = 0.93$
Ex5-K	$\frac{207.06}{1336} = 0.15$	$\frac{1889.42}{3669} = 0.51$	$\frac{4229.73}{5719} = 0.74$	$\frac{3717.61}{4898} = 0.76$	$\frac{1813.16}{2986} = 0.61$
Ex10-T	$\frac{23.08}{321} = 0.07$	$\frac{763.98}{1339} = 0.57$	$\frac{829.56}{2012} = 0.41$	$\frac{535.68}{932} = 0.57$	$\frac{638.17}{973} = 0.66$
Ex10-TA	$\frac{379.37}{2598} = 0.15$	$\frac{6782.82}{8048} = 0.84$	$\frac{12399.44}{9988} = 1.24$	$\frac{9697.73}{8054} = 1.20$	$\frac{5344.08}{5745} = 0.93$
Ex10-K	$\frac{215.45}{1402} = 0.15$	$\frac{1940.32}{3770} = 0.51$	$\frac{4378.26}{5954} = 0.74$	$\frac{3763.80}{4983} = 0.76$	$\frac{1838.00}{3047} = 0.60$

**The highest average intrigue scores are highlighted*

Chapter 5

Semantic Evolution

5.1 Introduction

Even though perceiving the meaning of words in the text is at the heart of natural language processing research, understanding them deeply at a human-level remains elusive (Levy et al. 2015). Nevertheless, in recent times, vector representations of words (developed using *word embeddings*) have demonstrated huge success in recovering certain semantic properties of words (Levy et al. 2015, Hashimoto et al. 2016). Word embeddings represent words as vectors in a multi-dimensional, continuous vector space where the geometrical relationships between vectors are vital. For instance, words that have higher semantic similarity to each other tend to reside in close proximity within the vector space (i.e., *distributional hypotheses*), and analogical relationships can be discovered through distance and angle properties (i.e., *vector arithmetic*) (Mikolov, Sutskever, Chen, Corrado & Dean 2013). Word embeddings have been successfully applied in a wide variety of natural language processing applications including *sentence classification* (Kim 2014), *machine translation* (Zou et al. 2013), *part-of-speech tagging* (Al-Rfou' et al. 2013) and *recommender systems* (Musto et al. 2016). Most of these application areas entail using word embeddings to learn a detailed representation of input data, which is crucial for downstream natural language processing tasks (Palangi et al. 2016, Hashimoto et al. 2016). From the *timeline analysis* of LBD computational techniques (discussed in Chapter 2), this thesis observed that the incorporation of modern word embedding techniques in the knowledge discovery process is the most recent type of computational technique utilised in the LBD literature. Nevertheless, only a *handful* of recent LBD studies use such techniques (Jha et al. 2018, Jha, Xun, Wang & Zhang 2019, Xun et al. 2017). In complex natural language processing application areas such as *LBD* (where *rich semantic inferences* are crucial) circumstantial analysis of vector semantics through the leveraging of word embeddings could be highly beneficial.

This thesis also observed from the *categorisation* of the LBD computational techniques (discussed in Chapter 2) that almost all prior LBD studies have neglected the temporal evolution of topics in the scientific literature. That is, they have used a static snapshot of digital libraries to discover novel knowledge linkages (Jha et al. 2018, Jha, Xun, Wang & Zhang 2019, Xun et al. 2017). However, scientific knowledge evolves rapidly, with the constant addition of new knowledge from on-going research (Jha et al. 2018, Jha, Xun, Wang & Zhang 2019). Therefore, the use of a mere static snapshot of the literature restrains the opportunity of detecting dynamic cues in the knowledge discovery process (Xun et al. 2017, Jha et al. 2018). Encoding the *temporal dynamics of scientific knowledge* in the LBD process may offer the opportunity to unravel meaningful temporal signals in differentiating new knowledge that cannot be captured using *static analysis of literature*.

Contemplating the complementary strengths of *modern embedding techniques* (observed in *timeline literature analysis* as discussed above) and *temporal dynamics* (observed in *categorisations of computational techniques* as discussed above), providing a holistic solution that encodes the *global scale implicit semantics* into an *informative temporal setting* may represent an improvement on existing LBD models. The main objective of this chapter is:

"to enhance the discovery component of the LBD workflow using fine-grained diachronic semantic inferences by conjoining global semantic relationships with the temporal dimension to enrich the typical static cues used in the LBD literature"

as defined at the outset of this thesis (i.e., *main objective 3 (RO3)* in Chapter 1). Otherwise stated, this study intends to analyse the implicit semantic relationships of scientific topics in a time-sensitive environment with the ultimate goal of detecting novel knowledge linkages with high precision. With this goal in mind, this chapter attempts to answer the main research question (*RQ3*): *'does incorporating meaningful diachronic semantic inferences in the LBD discovery process through leveraging implicit semantic relationships of word embeddings in temporally-aware vector spaces enrich the typical* *static cues used in the previous LBD studies?'*. To support the main objective of this chapter and to systematically answer the aforementioned research question, this study is sub-divided into several stages by focusing on the following sub research objectives.

- *RO3.1. Incorporating a global picture of topic interactions into temporally encoded schemata to capture the semantic relationships of the topics in a wide scope* (discussed in Section 5.3).

- *RO3.2. Integrating temporal information of the scientific topics with word embeddings to construct temporally encoded schemata to model and understand the semantic behaviour of scientific topics across time* (discussed in Section 5.4).

- *RO3.3. Disentangling temporal semantics of the scientific topics from the temporally encoded schemata that reflect the potential characteristics of novel knowledge linkages* (discussed in Section 5.5).

- *RO3.4. Scrutinising the derived diachronic semantic inferences of the scientific topics using a circumstantial temporal analysis component to unravel meaningful semantically infused temporal cues* (discussed in Sections 5.6, 5.7, 5.8 and 5.9).

This chapter is organised as follows. Section 5.2 provides a high-level overview of the major phases of the *proposed LBD framework* by summarising the key functionalities and objectives of each phase. Section 5.3 discusses the way in which global topic interactions are induced to learn the latent vector representations of scientific topics in the literature through the use of the *time-specific global corpus*. Section 5.4 describes how the corpora prepared in the previous phase are used to construct diachronic word embeddings that co-model both *vector semantics* and the *temporal dimension*. Section 5.5 presents the *core discovery setting* of this study by elaborating how the semantics and temporal aspects of the scientific topics are combined to provide a holistic solution to the problem of discovering novel knowledge linkages. In this regard, this section leverages the idea of *semantic shifts* to capture the *semantically infused temporal trajectories* of scientific topics. Section 5.6 is dedicated to describing how these extracted semantically infused temporal trajectories (i.e., *diachronic semantic inferences*) of scientific topics are sifted to elicit novel knowledge linkages patterns. In essence, Section 5.6 provides a high-level overview of the *core analysis setting* of this study, which is constituted of three

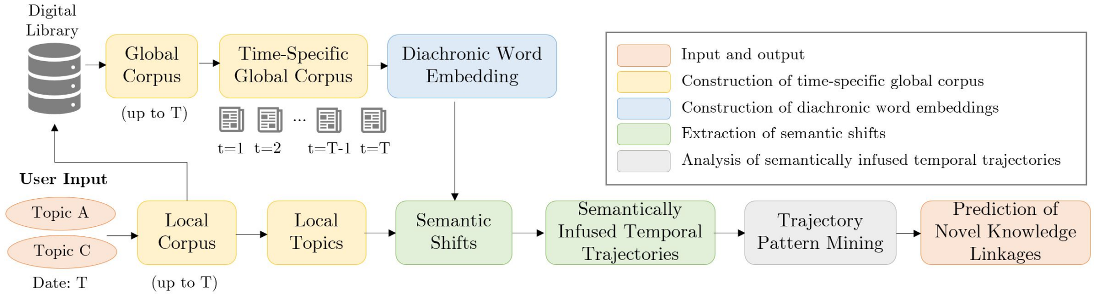

Digital Library
Global Corpus
(up to T)
Time-Specific Global Corpus
t=1 t=2 ... t=T-1 t=T
Diachronic Word Embedding
Input and output
Construction of time-specific global corpus
Construction of diachronic word embeddings
Extraction of semantic shifts
Analysis of semantically infused temporal trajectories
User Input
Topic A
Topic C
Date: T
Local Corpus
(up to T)
Local Topics
Semantic Shifts
Semantically Infused Temporal Trajectories
Trajectory Pattern Mining
Prediction of Novel Knowledge Linkages

Figure 5.1: Schematic overview of the proposed LBD framework

main models: the *dedicated trajectory model*, the *feature-based trajectory model* and the *trajectory alignment model*. These three models are discussed in detail in Sections 5.7, 5.8 and 5.9, respectively. Section 5.10 outlines the experimental setup and the experimental results of the proposed models, as well as comparing them to the baseline LBD models. Section 5.11 summarises the key findings of this study, while also highlighting its major contributions.

5.2 Overview of the Proposed LBD Framework

The purpose of this section is to briefly outline the four major components in the *proposed LBD framework* and their key objectives. These components are *construction of a time-specific global corpus*, *construction of diachronic word embeddings*, *extraction of semantic shifts*, and *analysis of semantically infused temporal trajectories*. Figure 5.1 denotes a high-level overview of how these four components are connected in the proposed LBD framework. This framework is considered the main blueprint of all the LBD models proposed in the latter part of this chapter. Further details on these four components are discussed in Sections 5.3, 5.4, 5.5 and 5.6, respectively.

The main purpose in the *construction of a time-specific global corpus* component is to prepare the scientific literature corpora for the analysis of remaining phases in the proposed LBD framework. In preparing the corpora, this study deviates from most prior LBD studies, which rely on a *query-specific local corpus* to extract potential patterns in identifying novel knowledge linkages. Otherwise stated, this study aims to detect large-scale global patterns in the local corpora by enriching concepts' semantic neighbourhoods with the idea of the *global corpus*. The key objective of this component is to incorporate semantic relationships of scientific topics in a wide scope that would ultimately benefit the semantic deductions made in the latter components of the proposed framework.

The intention of the *construction of diachronic word embeddings* component is to combine the word embeddings with the time dimension. This will allow for the construction of schemata that better represent the evolution of knowledge in the scientific literature. In this regard, this thesis focuses on an *emerging research field* that was initiated with the development of modern word embedding techniques (such as *word2vec*), namely *diachronic word embeddings* (a.k.a. *temporal word embeddings*, *dynamic word embeddings*), where the idea is to capture how words change across time in a data-driven manner (Kutuzov et al. 2018). More specifically, given the corpora of text ($\mathcal{C}_{t_1}$, $\mathcal{C}_{t_2}$, ..., $\mathcal{C}_{t_{n-1}}$ $\mathcal{C}_{t_n}$) in time slices (t_1, t_2, ..., t_{n-1}, t_n), the task of *diachronic word embeddings* is to analyse the dynamics of relationships among words across time (i.e., from t_1 to t_n). These dynamics reflect complicated processes in the natural language usage displayed in the corpora. The use of such diachronic embedding settings (which are rich in both *semantics* and *temporal details*), facilitates this study's main objective of inspecting the semantic behaviour of scientific topics in a time-sensitive environment.

The main objective of the *extraction of semantic shifts* component is to extract meaningful measures to demonstrate the benefits of amalgamating semantic aspects and temporal dynamics of scientific topics towards discovering novel knowledge linkages. To facilitate this objective, the thesis leverages the idea of *semantic shifts*, which denotes how a concept's semantics change across time. In disentangling semantic shifts, this thesis focuses on three different perspectives of the concepts, namely *individual*, *pairwise* and *neighbourhood*. The extracted semantic shifts are prepared in the form of *semantically infused temporal trajectories* (i.e., *diachronic semantic inferences*). These trajectories are used as the key source to mine *semantically infused temporal patterns* in the subsequent phase (a.k.a. *trajectory pattern mining*). By mining these patterns, it may be possible to unravel strong temporal signals to detect novel knowledge linkages in the literature with high precision.

The final component of the proposed LBD framework, the *analysis of semantically infused temporal trajectories* entails scrutinising the derived semantically infused temporal trajectories (i.e., the extracted *diachronic semantic inferences*) to detect patterns of potential novel knowledge linkages. In this regard, this study proposes three types of LBD

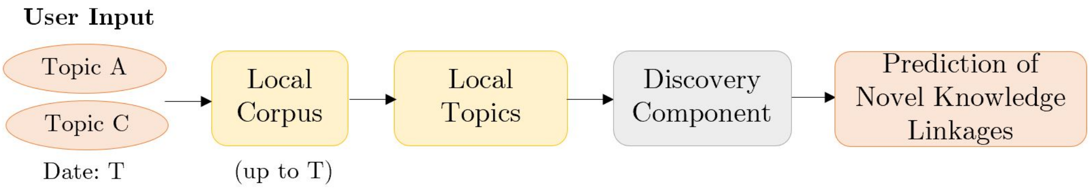


Figure 5.2: Schematic overview of the typical LBD workflow

models, namely the *dedicated trajectory model*, *feature-based trajectory model* and *trajectory alignment model*, where the first two models demonstrate the *direct uses* of the proposed diachronic semantic inferences. In contrast, the latter model manifests the *indirect uses* of the proposed diachronic semantic inferences. Unlike most previous LBD studies, the design of these three proposed LBD models does not incorporate semantic inferences from any external knowledge resources to support the idea of *reusability* (discussed in Chapters 6 and 8) and *portability* (discussed in Chapter 7) of this thesis.

5.3 Construction of Time-specific Global Corpus

To analyse the semantic properties of scientific topics in a temporal setting, a *time-specific corpus* is required. In this regard, this study leverages the *entire literature in the selected digital library/text repository.* The key objective behind incorporating the entire text repository is that it provides a rich platform to analyse the semantic relationships of *local scientific topics* in a global setting. In other words, the inclusion of the global semantic relationships in the entire text repository allows us to harness weak signals of novel knowledge that are not visible in a *query-specific local corpus.*

To further elaborate on this idea, consider the typical LBD framework used by most prior LBD studies depicted in Figure 5.2. In the typical framework, only the *query-specific local corpus* is used for the purposes of knowledge discovery. The major disadvantage of employing the query-specific local corpus is that it may be lacking crucial semantic relationships; thus, it may provide weak signals in eliciting new knowledge. For example, consider a situation where the user needs to explore *coronavirus* literature. In such situations, the query-specific local corpus merely contains the literature on *coronavirus.* However, when eliciting new knowledge on coronavirus, the semantic relationships of *'coronavirus'* with other related areas, such as *'SARS'*, may be vitally important. Due to the query-restrictive nature of *local corpora*, accommodating such vital semantic details

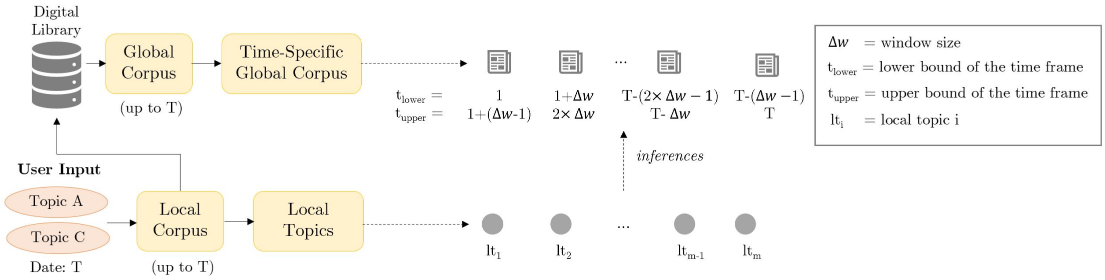


FIGURE 5.3: Schematic overview of the time-specific global corpus

into the analysis is difficult. To circumvent this issue, this study analyses the topics in the *local corpus* in a global semantic space by incorporating the *entire text repository* (namely, *global corpus*), with the ultimate aim of performing the semantic analysis in a wide perspective (Figure 5.3).

To facilitate the temporal analysis of the current study, the *global corpus* is divided into equivalent-sized time slices according to the *window size* (Figure 5.3). Supposing, the window size is set at five years, the global corpus is divided into five years slices, with each slice containing the literature published in the corresponding five years. This corpus is termed a *time-specific global corpus*, and it is used as the main data source in knowledge discovery.

5.4 Construction of Diachronic Word Embeddings

To prepare the *diachronic word embeddings* using the time-specific global corpus constructed in the previous phase, this study considers the following two main steps: *embedding construction* and *embedding alignment*.

5.4.1 Embedding Construction

This study utilises the popular neural word embedding technique *word2vec* to construct the distributional embeddings of the *global corpus*. The technique was chosen because its vector representations are efficient and expressive in comparison to those of other modern word embedding techniques, such as *GloVE* (Naili et al. 2017, Levy et al. 2015). There are two variants of word2vec, namely *CBOW (Continuous bag of words)* and *Skip-Gram* (Figure 5.4).

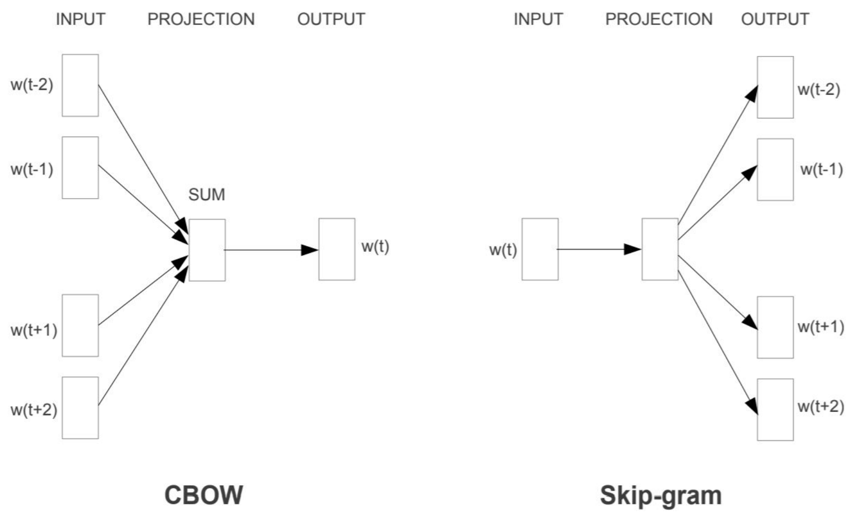


FIGURE 5.4: The architectures of CBOW and Skip-Gram models (Mikolov, Chen, Corrado & Dean 2013)

This study employs the *Skip-Gram* variant of word2vec (more specifically, *Skip-Gram with Negative Sampling (SGNS)*) to learn latent embedding spaces due to the following reasons.

- SGNS is considered to be the most popular variant to learn monolingual vector representations due to its robustness and training efficiency (Ruder et al. 2019).

- Levy et al. (2015) found that SGNS consistently outperformed the recent embedding technique GloVE on most of the tasks, such as word similarity and analogy. In the same study, they concluded that SGNS as a *robust baseline*, since even if it underperformed in some tasks, its reduction was not significant.

- SGNS has established its reputation by providing state-of-the-art results in numerous linguistic tasks (Levy & Goldberg 2014*b*)

- Levy et al. (2015) have identified SGNS as the fastest and cheapest embedding method to train in terms of *memory consumption* and *disk space*.

- SGNS is considered to be a powerful diachronic tool in the study of Hamilton et al. (2016*b*) that analyses the evolution of language.

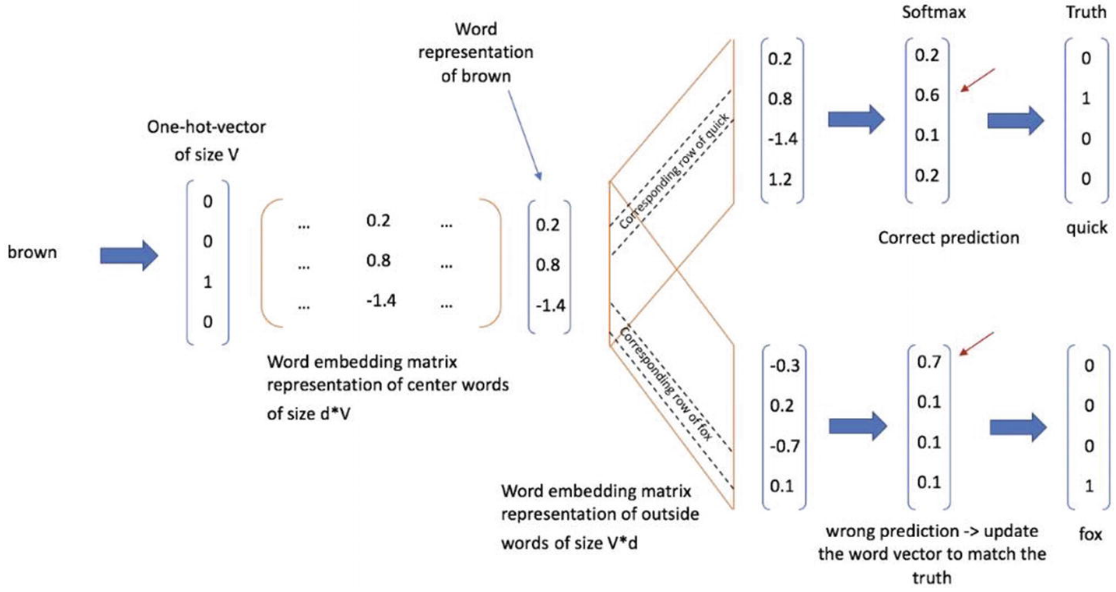


Figure 5.5: Simple example illustrating the three layers of the SGNS neural network (El-Amir & Hamdy 2020)

5.4.1.1 Skip-Gram with Negative Sampling (SGNS)

Given a target word w_k, *skip-gram* predicts the surrounding context words (see Figure 5.4) under the training objective defined in equation 5.1, where $\mathcal{C}$ is the corpus and C is the window size of each word.

$$\mathcal{L}_{SGNS} = -\frac{1}{|\mathcal{C}| - C} \sum_{k=C+1}^{|\mathcal{C}|-C} \sum_{-C \leq j \leq C, j \neq 0} \log P(w_{k+j}|w_k) \tag{5.1}$$

To calculate P($w_{k+j} \mid w_k$) in equation 5.1, a softmax function is used as denoted in equation 5.2, where $\tilde{\mathbf{x}}$ and $\mathbf{x}$ represent the input and output word embeddings of w_i.

$$P(w_{k+j}|w_k) = \frac{\exp(\tilde{\mathbf{x}}_{w_{k+j}} \top \mathbf{x}_{w_k})}{\sum_{i=1}^{|V|} \exp(\tilde{\mathbf{x}}_{w_i} \top \mathbf{x}_{w_k})} \tag{5.2}$$

Figure 5.5 demonstrates a simplified example of how the three layers in the neural network structure of SGNS (i.e., *input layer*, *projection layer* and *output layer*, illustrated in Figure 5.4) works to predict the vectors of the context words.

Since the partition function in the softmax's denominator in equation 5.2 is computationally expensive, SGNS utilises *Negative Sampling* (a simplification of *Noise Contrastive Estimation*) to approximate softmax. Negative sampling can be defined as in equation

5.3, where N is the number of negative samples, σ is the sigmoid function and P_n is the noise distribution.

$$P(w_{k+j}|w_k) = \log\sigma(\tilde{\mathbf{x}}_{w_{k+j}}\top\mathbf{x}_{w_k}) + \sum_{i=1}^{N}\mathbb{E}_{w_i\sim P_n}\log\sigma(-\tilde{\mathbf{x}}_{w_i}\top\mathbf{x}_{w_k}) \tag{5.3}$$

P_n is empirically defined as in equation 5.4, where *U(w)* represents the unigram distribution and Z is a normalisation factor (Mikolov, Sutskever, Chen, Corrado & Dean 2013).

$$P_n(w) = \frac{U(w)^{\frac{3}{4}}}{Z} \tag{5.4}$$

5.4.1.2 Limitations of Word Embedding Techniques

Despite the significant advances achieved in natural language processing applications using modern word embedding techniques (such as *word2vec*), the use of mere word embeddings may be limited. This is mainly because such models are based on static context, such that the meaning of words remains the same across time (Jha, Xun, Gopalakrishnan & Zhang 2019). Such static contexts are unable to capture complex phenomena involving language usage over time. However, analysis of language usage across time is crucial for areas such as *scientific literature mining*, where the knowledge is evolving rapidly on a daily basis (e.g., *MEDLINE* alone updates its data repository with nearly 2000-4000 scientific articles daily (Lu et al. 2015)).

To further illustrate this idea, consider the task of tracking the neighbourhood of a word over time. Figure 5.6 illustrates the evolution of the word *'cell'*, using three different timestamps (Boukhaled et al. 2019). In the 18th century, the word *cell* referred to a prison cell. However, the meaning of *cell* has changed drastically over time, and it is now mostly used to refer to the microscopic part of living beings (Figure 5.6).

Interpreting words based on their neighbourhood (as in Figure 5.6) is simply one of the many tasks that such *time-sensitive word embeddings* can offer. For instance, one could analyse how the word's neighbourhood *density* changes in time (Naili et al. 2017). In Figure 5.7 (Li et al. 2019), it is clear that the word *cell* does not have a *dense neighbourhood* in the 1900s. However, the neighbourhood of the word *cell* becomes

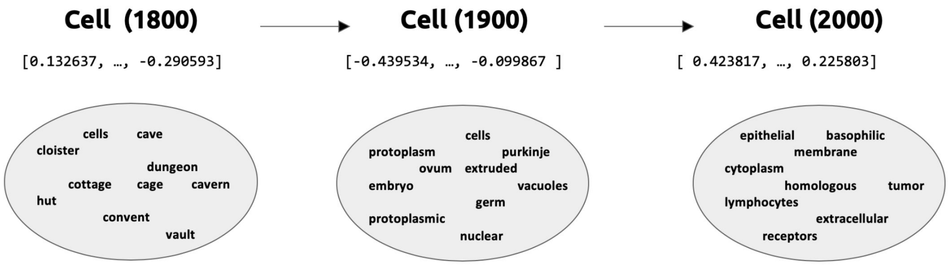


FIGURE 5.6: Neighbouring words of the word 'cell' across time (Boukhaled et al. 2019)

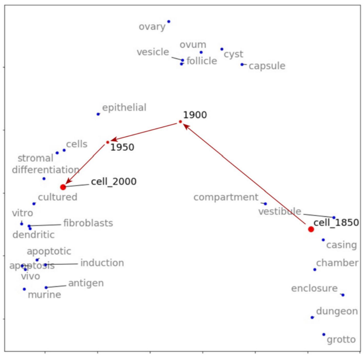


FIGURE 5.7: Semantic change of the word 'cell' across time (Li et al. 2019)

denser across time. In contrast to such density analysis, one could measure how much a word has *moved* in the semantic space across time. For example, Figure 5.7 clearly illustrates that the word *cell* has moved drastically (i.e., it displays a higher semantic distance) from 1850 to 1900. However, from 1950-2000, the semantic movement of the word *cell* is less prominent.

To further elaborate the potentiality of time-sensitive semantic inferences in the context of LBD, consider the classic example of *fish oil-blood viscosity-Raynaud's disease*. Figure 5.8 illustrates how the semantic meaning of the two topics, *fish oil* and *Raynaud's disease* evolved over time with respect to the intermediate concept of *blood viscosity*. More specifically, the concept of *blood viscosity* was semantically distinct from the two main topics in 1953 (Xun et al. 2017). Nevertheless, with more research findings getting published on these topics over time, the concept of *blood viscosity* has come closer to the main topics in the semantic spaces indicating their implicit semantic relatedness, which was eventually identified by Swanson in 1986 (Swanson 1986).

Correspondingly, *temporal word embeddings* can be used to make in-depth semantic inferences about words in a way that static word embeddings cannot facilitate. This emphasises the need to develop *dynamic language models* wherein the semantic change

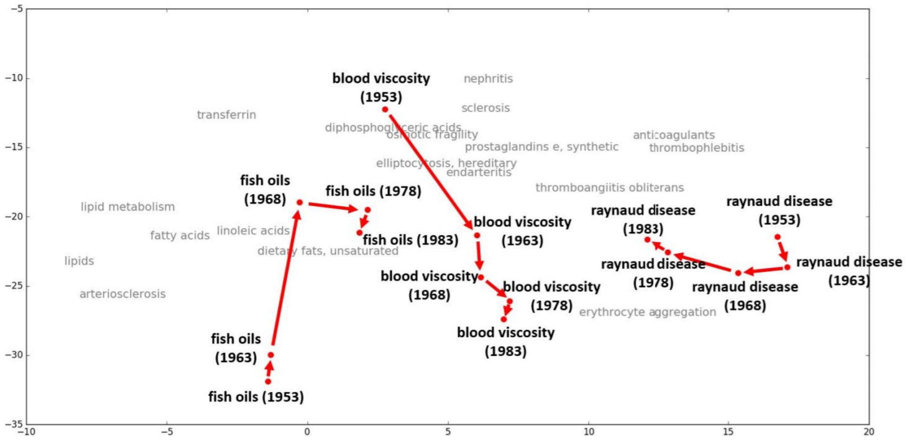


FIGURE 5.8: Semantic change of words across time using the classic example of *fish oil-blood viscosity-Raynaud's disease* in the LBD field (Xun et al. 2017)

of words across time is encapsulated. With such a goal in mind, this study leverages the revolutionary opportunities afforded by *diachronic word embeddings* in order to better understand the way that the semantics of scientific topics change over time. This allows for the detection of new temporal signals that could be beneficial in capturing novel knowledge linkages more precisely.

5.4.1.3 Construction of Time-specific Embedding Spaces

To construct time-specific embedding spaces, this study learnt the distributed representation of scientific topics for each time-slice in the *time-specific global corpus*, employing *SGNS*. That is, this phase entailed constructing n latent embedding spaces, assuming the existence of n time-slices in the *time-specific global corpus*. In the constructed embedding spaces, each scientific topic w_{i} has a vector representation $\mathbf{w}^{(t)}$ in each time slice of the global corpus.

5.4.2 Embedding Alignment

It is not possible to directly compare the constructed word vectors in each time slice of vector spaces. This is because most modern word embedding methods (including *SGNS*) are inherently stochastic; thus, the produced word embedding sets could occur in arbitrary orthogonal transformations (Hamilton et al. 2016*b*,*a*). Consequently, even if word embeddings are trained on the same data, the produced numerical vectors will be different in separate learning runs (however, the pairwise similarities between vectors will

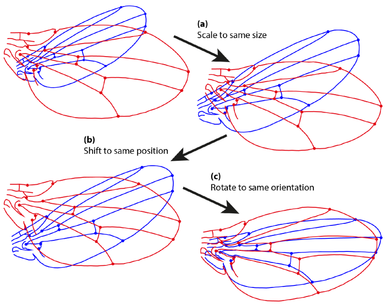


Figure 5.9: Simplified example of orthogonal Procrustes alignment
Source: https://en.wikipedia.org/wiki/Procrustes_analysis

be roughly equivalent) (Levy et al. 2015). Therefore, it is crucial to perform an alignment of the word vectors in each time slice to the same co-ordinate axes before extracting the semantic shifts of local topics (e.g., for measures such as *individual semantic shifts*). To facilitate this alignment process, *orthogonal Procrustes alignment* is utilised in this study, which finds the optimal rotational alignment of embedding spaces. Figure 5.9 illustrates a simplified example of the *orthogonal Procrustes alignment* of two different shapes.

In embedding alignment, the orthogonal Procrustes problem can be considered as a *matrix approximation* in linear algebra. Simply put, given two matrices M_1 and M_2, the orthogonal Procrustes problem attempts to find the *orthogonal matrix* which most closely maps M_1 to M_2. Considering a matrix of word embeddings trained at time slice t ($\mathbf{W}^{(t)} \in \mathbb{R}^{d \times |\mathcal{V}|}$), orthogonal Procrustes alignment is conducted across time, as defined in equation 5.5 where $\mathbf{R}^{(t)} \in \mathbb{R}^{d \times d}$. The solution corresponds to the best rotational alignment while preserving cosine similarity (Hamilton et al. 2016*b*). In essence, given two matrices $\mathbf{W}^{(t)}$ and $\mathbf{W}^{(t+1)}$ in arbitrary coordinate systems, equation 5.5 minimises over all the possible orthogonal matrices $\mathbf{Q}$ to find the most *optimal solution.*

$$\mathbf{R}^{(t)} = \underset{\mathbf{Q}^{\top}\mathbf{Q}=\mathbf{I}}{\arg\min} \|\mathbf{Q}\mathbf{W}^{(t)} - \mathbf{W}^{(t+1)}\|_F \tag{5.5}$$

In equation 5.5, $\|\cdot\|_F$ denotes the Frobenius norm. For a matrix M, the Frobenius norm can be calculated as per equation 5.6.

$$\| M \|_F = (\sum_i \sum_j M_{ij}^2)^{\frac{1}{2}} \tag{5.6}$$

5.5 Extraction of Semantic Shifts

The purpose of this section is to explain how meaningful measures are extracted from the constructed diachronic embedding spaces in the previous phase, in order to quantify the semantic evolution of scientific topics in the literature (more specifically, *local topics*, as denoted in Figure 5.1). In this regard, the *semantic shifts* of the scientific topics are the crux of this phase. Accordingly, this study unravels the way in which scientific topics' semantics evolve over time from three broad perspectives, *individual semantic shifts*, *pairwise semantic shifts* and *neighbourhood semantic shifts*. The ultimate rationale behind extracting such measures in the form of semantic shifts is to unravel new temporal patterns to distinguish potential novel knowledge linkages from the remaining scientific topics in the literature.

5.5.1 Individual Semantic Shifts

This category captures how the semantics of each scientific topic changes across time by focusing on the scientific topic itself. In this regard, two types of individual semantic shift were employed, namely *individual global shifts* and *individual local shifts*.

Individual Global Shift (IGS) quantifies the linguistic drift of a concept by analysing how far a scientific topic has shifted in the embedding spaces in two consecutive time slices *t* and *t+1*, as defined in equation 5.7. More specifically, equation 5.7 extracts the cosine distance of the concept's word vector $\mathbf{w}_i$ in the vector spaces modelled at time *t* and *t+1*. This process is illustrated in Figure 5.10. The subtle usage changes and other global effects encountered as a result of the shifting of the entire semantic space are reflected in this measure (Hamilton et al. 2016*a*).

$$d^{\,\mathrm{IGS}}(w_i^{(t)}, w_i^{(t+1)}) = \text{cos-dist}(\mathbf{w}_i^{(t)}, \mathbf{w}_i^{(t+1)}) \tag{5.7}$$

Individual Local Shift (ILS) focuses on semantic change at the local scale by observing the concept's ($\mathbf{w}_i$) nearest semantic neighbours in two consecutive time slices, *t* and *t+1* (Figure 5.11). As such, ILS is sensitive to the concept's paradigmatic relations and less concerned with global shifts in syntagmatic contexts. Since this measure is based on the local semantic neighbours, initially, the concept w_i's $\mathcal{K}$ nearest neighbours at

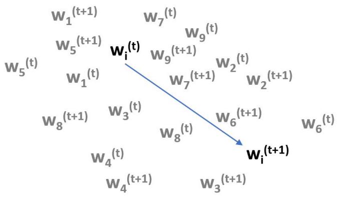


Figure 5.10: Individual global shifts

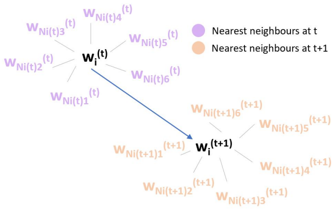


Figure 5.11: Individual local shifts

time t are obtained $(\mathcal{N}_{\mathcal{K}}(w_i^{(t)}))$. Subsequently, to quantify the change between the two time-periods t and $t{+}1$, a second-order similarity vector is computed for $w_i^{(t)}$ based on these nearest neighbour sets, as defined in equation 5.8. The computed vectors for $w_i^{(t)}$ and $w_i^{(t+1)}$ are used to quantify the local neighbourhood change, as denoted in equation 5.9 (Hamilton et al. 2016*a*,*b*).

$$\mathbf{s}^{(t)}(j) = \text{cos-sim}(\mathbf{w}_i^{(t)}, \mathbf{w}_j^{(t)}) \text{ where } \forall w_{\mathrm{j}} \in \mathcal{N}_{\mathcal{K}}(w_i^{(t)}) \cup \mathcal{N}_{\mathcal{K}}(w_i^{(t+1)}) \tag{5.8}$$

$$d^{\,\mathrm{ILS}}(w_i^{(t)}, w_i^{(t+1)}) = \text{cos-dist}(\mathbf{s}_i^{(t)}, \mathbf{s}_i^{(t+1)}) \tag{5.9}$$

5.5.2 Pairwise Semantic Shifts

This category assesses how the semantics of each scientific topic change across time with respect to the two user-defined input topics A and C. In terms of pairwise semantic shifts, this study leverages two types of measure, namely *pairwise semantic displacement* and *pairwise distance proximity.*

Pairwise Semantic Displacement (PSD) is intended to capture how a concept's $(\mathbf{w}_i)$ semantic similarity changes across time relative to topics A $(\mathbf{w}_A)$ and C $(\mathbf{w}_C)$, as shown in Figure 5.12. Thus, this measure provides a platform from which to assess whether

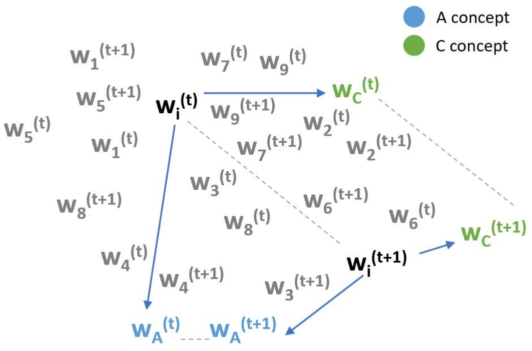


FIGURE 5.12: Pairwise semantic displacement

a concept displays a growing semantic similarity with topics A and C over time. To facilitate this process, the cosine similarity of the vectors in each time-slice is used, as defined in equation 5.10.

$$s^{\,\mathrm{PSD}}(w_i^{(t)}, w_A^{(t)}, w_C^{(t)}) = \mathrm{avg}(\text{cos-sim}(\mathbf{w}_i^{(t)}, \mathbf{w}_A^{(t)}), \text{cos-sim}(\mathbf{w}_i^{(t)}, \mathbf{w}_C^{(t)})) \tag{5.10}$$

The purpose of the *Pairwise Distance Proximity (PDP)* measure is to identify whether a concept's ($\mathbf{w}_i$) temporal trajectory is leaning towards (i.e., in close proximity to) both topics A ($\mathbf{w}_A$) and C ($\mathbf{w}_C$). The reason for adopting this measure is that LBD seeks latent conceptual bridges that connect topics A and C; thus, the concept's trajectory should incline towards both the input topics. Note that in Figure 5.13, $\mathbf{w}_j$ only favours $\mathbf{w}_C$ at time *t+1*, while $\mathbf{w}_i$ favours both $\mathbf{w}_A$ and $\mathbf{w}_C$. The purpose of this measure is to capture such details in the knowledge discovery process. PDP is calculated as defined in equation 5.11, where β denotes a penalising factor, equal to or greater than zero.

$$\begin{aligned} d^{\,\mathrm{PDP}}(w_i^{(t)}, w_A^{(t)}, w_C^{(t)}) &= \max(\text{cos-dist}(\mathbf{w}_i^{(t)}, \mathbf{w}_A^{(t)}), \text{cos-dist}(\mathbf{w}_i^{(t)}, \mathbf{w}_C^{(t)})) \\ &\quad + \beta \mid \text{cos-dist}(\mathbf{w}_i^{(t)}, \mathbf{w}_A^{(t)}) - \text{cos-dist}(\mathbf{w}_i^{(t)}, \mathbf{w}_C^{(t)}) \mid \;\; \text{where } \beta \geq 0 \end{aligned} \tag{5.11}$$

5.5.3 Neighbourhood Semantic Shifts

This category of semantic shifts detects how the semantics of each scientific topic change over time, focusing not only on the user-defined A and C topics alone, but also on their *recent core meaning*. With reference to neighbourhood semantic shifts, this study uses

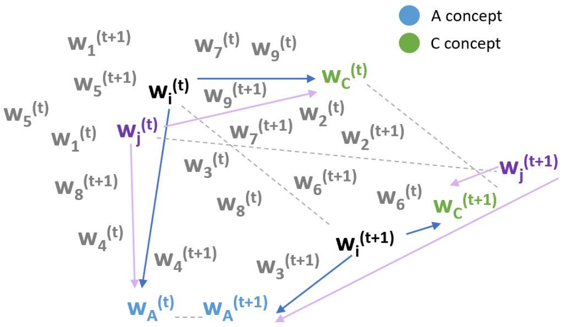


FIGURE 5.13: Pairwise distance proximity

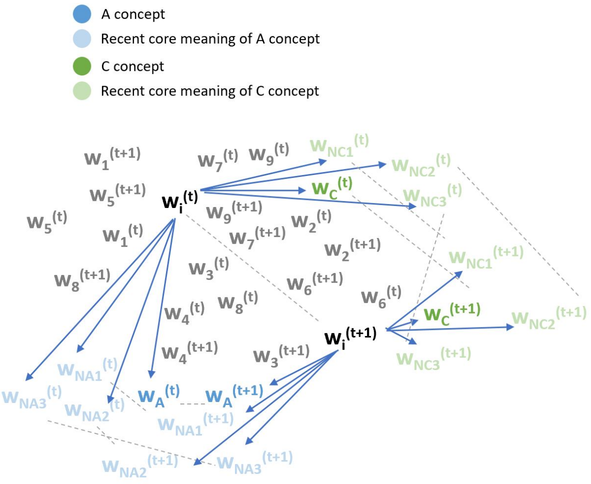


FIGURE 5.14: Neighbourhood semantic displacement

the same two measures introduced in *pairwise semantic shifts*, except that instead of using A and C themselves, their *recent core meanings* are utilised. The recent neighbours of topic A (N_A) and C (N_C) in a time window $\mathcal{W}$ are calculated as in equation 5.12.

$$N_A = \bigcap_{t=T-\mathcal{W}}^{T} \mathcal{N}_{\mathcal{K}}(w_A^{(t)}), \ N_C = \bigcap_{t=T-\mathcal{W}}^{T} \mathcal{N}_{\mathcal{K}}(w_C^{(t)}) \tag{5.12}$$

The only difference between *Neighbourhood Semantic Displacement (NSD)* and *Pairwise Semantic Displacement* is that NSD includes the *recent core meaning* of input topics A ($\mathbf{w}_A$) and C ($\mathbf{w}_C$) in the calculation, as denoted in Figure 5.14. Thus, NSD captures the extent to which a concept ($\mathbf{w}_i$) forms semantic relationships not only with the two input topics, but also with their recent core meaning. Therefore, this measure provides the opportunity to evaluate the need to assess the semantic neighbourhood of the input topics in the LBD context.

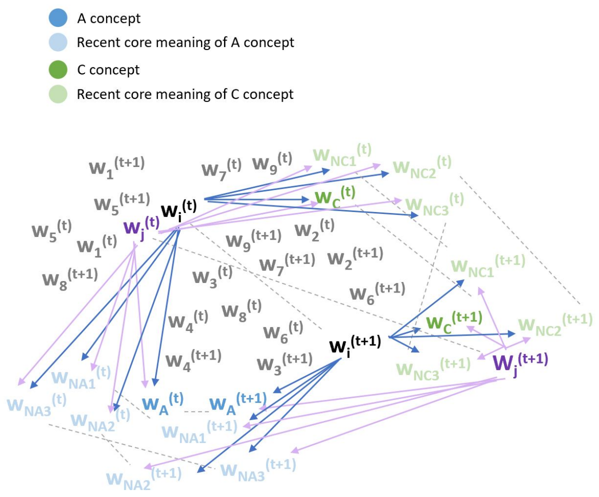


FIGURE 5.15: Neighbourhood distance proximity

Neighbourhood Distance Proximity (NDP) extends the idea of *Pairwise Distance Proximity* by incorporating the recent neighbourhoods of topics A ($\mathbf{w}_A$) and C ($\mathbf{w}_C$). Thus, this measure assesses whether a concept's ($\mathbf{w}_i$) temporal trajectory inclines not just to topics A and C, but also to their semantic neighbours, as illustrated in Figure 5.15.

5.5.4 Semantically Infused Temporal Trajectories

The six types of semantic shift (as described in Sections 5.5.1, 5.5.2 and 5.5.3; two from each semantic shift category, *individual*, *pairwise* and *neighbourhood*) are devised in the form of trajectories. Since these trajectories reflect both semantics and temporal aspects of concepts, this study refers to them as *semantically infused temporal trajectories* (i.e., *diachronic semantic inferences*). More specifically, the six semantically infused temporal trajectories that are extracted for a scientific topic w_i can be denoted in the form of:

$$\mathrm{TJ}^{\mathrm{IGS}}(w_i) = (d^{\mathrm{IGS}}(w_i^y),\ d^{\mathrm{IGS}}(w_i^{y+1})),\ ...,\ d^{\mathrm{IGS}}(w_i^{T-1}),\ d^{\mathrm{IGS}}(w_i^T))$$

$$\mathrm{TJ}^{\mathrm{ILS}}(w_i) = (d^{\mathrm{ILS}}(w_i^y),\ d^{\mathrm{ILS}}(w_i^{y+1})),\ ...,\ d^{\mathrm{ILS}}(w_i^{T-1}),\ d^{\mathrm{ILS}}(w_i^T))$$

$$\mathrm{TJ}^{\mathrm{PSD}}(w_i) = (s^{\mathrm{PSD}}(w_i^y),\ s^{\mathrm{PSD}}(w_i^{y+1})),\ ...,\ s^{\mathrm{PSD}}(w_i^{T-1}),\ s^{\mathrm{PSD}}(w_i^T))$$

$$\mathrm{TJ}^{\mathrm{PDP}}(w_i) = (d^{\mathrm{PDP}}(w_i^y),\ d^{\mathrm{PDP}}(w_i^{y+1})),\ ...,\ d^{\mathrm{PDP}}(w_i^{T-1}),\ d^{\mathrm{PDP}}(w_i^T))$$

$$\mathrm{TJ}^{\mathrm{NSD}}(w_i) = (s^{\mathrm{NSD}}(w_i^y),\ s^{\mathrm{NSD}}(w_i^{y+1})),\ ...,\ s^{\mathrm{NSD}}(w_i^{T-1}),\ s^{\mathrm{NSD}}(w_i^T))$$

$$\mathrm{TJ}^{\mathrm{NDP}}(w_i) = (d^{\mathrm{NDP}}(w_i^y),\ d^{\mathrm{NDP}}(w_i^{y+1})),\ ...,\ d^{\mathrm{NDP}}(w_i^{T-1}),\ d^{\mathrm{NDP}}(w_i^T))$$

where y is the first occurrence of w_i in the dataset, s is a similarity measure and d is a distance measure.

Since local topics are the *potential discovery candidates* of the two user-defined input topics A and C, each local topic is represented by the six semantically infused temporal trajectories, as defined above (Figure 5.1). To summarise, the main ingredients of these semantically infused temporal trajectories are *global-scale semantics* and *time-specific behaviours* of concepts in the scientific literature. The ultimate objective of this analysis is to deduce whether these six temporal trajectories demonstrate potential *semantically infused temporal cues*, which can be used to distinguish novel knowledge linkages from the remaining scientific topics with high precision.

5.5.5 Frequency Heuristics

In addition to the proposed semantically infused temporal trajectories discussed in Section 5.5.4, this study also uses *two frequency heuristics* that have been developed in past LBD research, namely the *Local Frequency Heuristic (LFH)* and *Global Frequency Heuristic (GFH).*

The intention of *LFH* is to capture the frequency with which a *local topic (lp_i)* appears in the local corpus, since prior LBD research has identified that local topics which occur only once in A or C literature are less prominent in the LBD workflow (Torvik & Smalheiser 2007). More specifically, this feature is set to *1*, if ($\mathrm{n}_{(A,lp_i)}$ >1 OR $\mathrm{n}_{(A)}$ <1000) AND ($\mathrm{n}_{(lp_i,C)}$ >1 OR $\mathrm{n}_{(C)}$ <1000), and *0* otherwise (Torvik & Smalheiser 2007). The intention of *GFH* is to capture the global frequency of a *local topic (lp_i)*, since it has also been identified in the LBD literature that very frequent or rare local topics in the global corpus are less prominent. Thus, this feature is set to $|3 - \log_{10}(\mathrm{n}_{(lp_i)})\ |$ (Torvik & Smalheiser 2007).

5.6 Analysis of Semantically Infused Temporal Trajectories

This section briefly introduces the three types of LBD models proposed in this study, each of which leverages the derived semantically infused temporal trajectories as their *core discovery source*. These LBD models are the *dedicated trajectory model*, *feature-based trajectory model* and *trajectory alignment model.*

In recent times, *deep learning models* have shown promise in many application areas, including *time series* and *sequential data analysis* (Fawaz et al. 2019, Längkvist et al. 2014). Inspired by such research *outside LBD*, the *Dedicated Trajectory Model (DTM)* leverages deep learning techniques to perform feature learning using the derived semantically infused temporal trajectories. More specifically, this study proposes multiple *Deep Neural Network (DNN) architectures* to unravel meaningful semantically infused temporal signals to discover potential novel knowledge linkages. For this purpose, the study leverages *LSTM* to detect long term temporal dependencies and *CNN* to capture the spatial sparsity and heterogeneity in data (Du et al. 2018). Further details on this proposed LBD model are discussed in Section 5.7.

In the *Feature-based Trajectory Model (FTM)*, semantically infused temporal trajectories are represented using *hand-crafted features.* To extract hand-crafted features, this study incorporates both the *trajectory values* and *trajectory shape*, since these are the major components that constitute a semantically infused temporal trajectory. Therefore, intermingling these two types of hand-crafted features facilitates the derivation of meaningful temporal patterns that are otherwise hidden (i.e., when using these two feature types in isolation) in the knowledge discovery process. Details on this proposed LBD model are presented in Section 5.8.

Unlike *DTM* and *FTM*, the *Trajectory Alignment Model (TAM)* does not incorporate the proposed semantically infused temporal trajectories directly into the analysis. Instead, this model demonstrates the potential *indirect uses* of the semantically infused temporal trajectories. More specifically, this LBD model leverages the idea of incorporating multiple forms of new knowledge types by maintaining a *template repository*, which includes the *trajectory samples of actual new knowledge.* Subsequently, these trajectory samples are aligned with the *trajectories of local topics* to identify the extent of their correspondence. In essence, this LBD model focuses on large-scale integration of patterns from multiple forms of new knowledge to provide a different perspective on enhancing the knowledge discovery process. The main inspiration for this proposed model comes from the *docking mechanism*, which is popular in molecular modelling (Jacob et al. 2012, Ferreira et al. 2015). This model is discussed in detail in Section 5.9.

5.7 Dedicated Trajectory Model (DTM)

This section describes the first LBD model proposed in this study, the *Dedicated Trajectory Model (DTM)*. This model leverages modern deep learning techniques (more specifically, *LSTM* and *CNN*, as discussed in Section 3.5.4) to sift important characteristics of the semantically infused temporal trajectories, with the ultimate motive of detecting novel knowledge linkages with high precision. This section commences by summarising the key motivation for this study and providing an overview of the proposed model. The succeeding sub-sections describe the main phases of this LBD model (including the setup of the DNN framework, design considerations relating to DNNs, and the construction of DNNs) in detail.

5.7.1 Rationale

More recently, a few studies (Jha et al. 2018, Jha, Xun, Wang & Zhang 2019, Xun et al. 2017) have attempted to mitigate the limitation of *static domain* in previous LBD research by integrating temporal information on scientific topics into the LBD workflow. Even though these studies undoubtedly enhance the traditional LBD setting, they still contain several inherent limitations.

One of these limitations is their *fairly shallow temporal analysis component*. For example, when measuring the temporal trend of implicit connections, Xun et al. (2017) only consider the first and last values of the temporal sequence, ignoring the patterns in the overall sequence. The focus of this study is to overcome this limitation by scrutinising semantically infused temporal trajectories using a higher level of granularity (Shoemark et al. 2019), which may aid in identifying novel knowledge linkages more precisely. More specifically, this study attempts to answer the following question: *'does analysing the proposed semantically infused temporal trajectories in greater detail assist in the unravelling of novel knowledge linkages with high precision?'*. To the best of our knowledge, this is the *first study* in the LBD field to integrate such circumstantial temporal analysis in order to deduce semantically infused temporal cues. In this regard, this study explores the massive opportunities afforded by modern deep learning techniques to unwind new signals of potential novel knowledge linkages. Unlike handcrafted features, using DNN models may offer the opportunity to discover unforeseen structures of novel knowledge.

Secondly, as in most existing LBD literature, these recent temporal studies rely on one or two temporal characteristics to discover potential new knowledge linkages. Such a reliance on few temporal characteristics may be limited due to two reasons (as discussed in detail in Chapter 1). Firstly, due to the complexity of natural language usage, incorporating limited characteristics may hinder the LBD model's ability to discover novel knowledge linkages with high precision. Secondly, these LBD models may be biased in favour of picking only one or limited types of novel knowledge, since it has been observed in the theoretical LBD literature that novel knowledge may reside in multiple forms in the literature (Davies 1989). Therefore, the integration of multiple factors/characteristics in the knowledge discovery process may assist in overcoming these two limitations. In essence, this study attempts to answer the following question: *'does providing a comprehensive solution that incorporates multiple factors/characteristics (e.g., multiple semantically infused temporal cues) yield better predictive effects in comparison to single or limited characteristics?'*. In this regard, this study focuses on different DNN architecture setups by contemplating the strengths of *LSTM* and *CNN*, with the main objective of broadly identifying features/characteristics (e.g., from *low-level features* to *high-level features*) that will ultimately be beneficial in increasing prediction precision as well as recovering multiple forms of novel knowledge linkages in the literature.

5.7.2 Overview of Proposed LBD Model (DTM)

This section provides a high-level overview of the proposed LBD model by outlining the key functionalities of its main phases. Recall that the input to an LBD model is two topics of interest (A and C) and a date T, where the goal is to analyse the literature up to time T, and to detect latent conceptual bridges that are most likely to connect the two topics in the future. To facilitate this process, the same initial phases in the blueprint of the proposed LBD framework (discussed in Section 5.2) are utilised. To sum up, first, the local corpus is preprocessed in order to identify scientific topics that are relevant to the user-defined input topics A and C (i.e., local topics in Figure 5.16). Subsequently, semantic inferences relating to these extracted local topics are performed using the global corpus. The main reason for adopting the global corpus is that it is rich in semantic details compared to the query-specific local corpus. To perform this analysis in a temporal setting, the global corpus is divided into equivalent-sized time-slices named *time-specific global corpus* (Figure 5.16). For the scientific topics in each

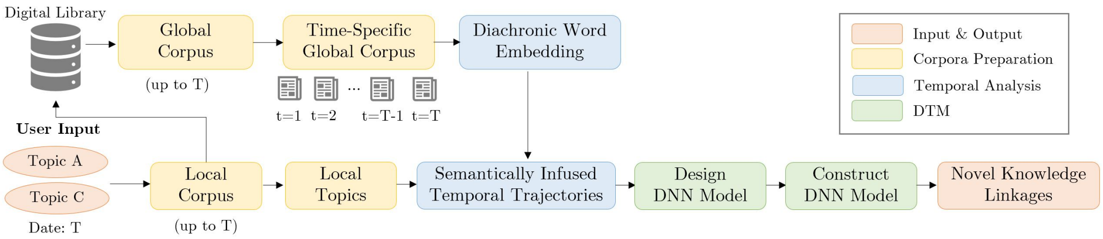


FIGURE 5.16: Schematic overview of the Dedicated Trajectory Model (DTM)

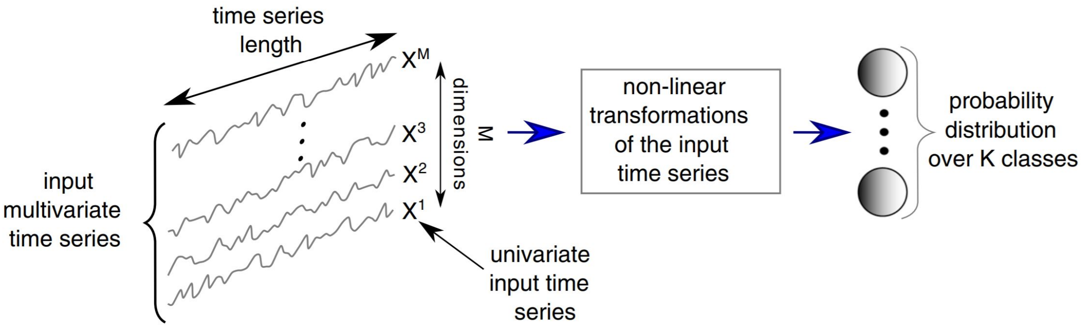

time series length
input multivariate time series
X^M
X^3
X^2
X^1
M dimensions
univariate input time series
non-linear transformations of the input time series
probability distribution over K classes

FIGURE 5.17: Unified deep learning framework of multivariate time series (Fawaz et al. 2019)

time-slice of the time-specific global corpus, latent embedding spaces are constructed to reason upon them to detect interesting global semantic relationship patterns of the local topics. More specifically, this study investigates the evolution of global semantic relationships of local topics in the embedding spaces across time, in order to extract six types of semantically infused temporal trajectories (i.e., *diachronic semantic inferences*), as discussed in Section 5.5.

Prior to the design of deep neural network models, as denoted in Figure 5.16, this study redefines the six extracted semantically infused temporal trajectories of each local topic as a *multivariate time series problem*. For this purpose, this study introduces the notions of univariate time series and multivariate time series, and transforms the six temporal trajectories in the setting of a multivariate time series. The next stage of this model incorporates deep neural network (DNN) models that excel at interpreting sequence/time series data to detect patterns in a data-driven manner. To this end, two variants of deep neural networks (*LSTM* and *CNN*) are used as the *main building blocks* of this study for the purpose of designing DNN architectures. Subsequently, the temporal trajectories that are in the setting of multivariate time series are used with the designed DNNs, as denoted in Figure 5.17 (Fawaz et al. 2019).

5.7.3 Multivariate Time Series Setting

Typically, a *univariate time series* is an ordered set of data points measured at successive time-spaced points with uniform time intervals (Fawaz et al. 2019, Zheng et al. 2014). The univariate time series can be denoted in the form $\mathrm{U_i} = (\mathrm{x_1}, \mathrm{x_2}, ..., \mathrm{x_n})$, where n is the length of $\mathrm{U_i}$. In *multivariate time series* M, each component $\mathrm{m_i}$ is a univariate time series $\mathrm{U_i}$. In essence, multivariate time series is a collection of time series that have the same timestamps. In any timestamp t, m_t can be defined as $\mathrm{m_t} = (\mathrm{m}_{\mathrm{U_1}t}, \mathrm{m}_{\mathrm{U_2}t}, ...,$ $\mathrm{m}_{\mathrm{U_j}t})$, where j is the number of univariate time series collected in M.

Following these definitions, it is safe to assume that the derived six semantically infused temporal trajectories as six univariate time series. Next, this study articulated these six temporal trajectories in the form of M (defined above), wherein for time slice t, m_t is defined as in equation 5.13. In the equation, x in $\mathrm{TJ^x}$ corresponds to the six temporal trajectories defined in Section 5.5.4.

$$\mathrm{m}_t = (\mathrm{m}_{\mathrm{TJ^{IGS}}t}, \mathrm{m}_{\mathrm{TJ^{ILS}}t}, \mathrm{m}_{\mathrm{TJ^{PSD}}t}, \mathrm{m}_{\mathrm{TJ^{PDP}}t}, \mathrm{m}_{\mathrm{TJ^{NSD}}t}, \mathrm{m}_{\mathrm{TJ^{NDP}}t}) \tag{5.13}$$

5.7.4 Main Building Blocks of DNN Models

This section is dedicated to discussing the design considerations of DNN architectures that are used to analyse the derived temporal trajectories. The two main building blocks used to construct the proposed DNN architectures are *Long Short-Term Memory (LSTM)* and *Convolutional Neural Network (CNN)*. The theoretical foundation of these two key building blocks is discussed in Chapter 3.

While *LSTMs* are inherently designed to analyse *time series or sequence data* (much like the proposed *semantically infused temporal trajectories*), *CNNs* were originally used to analyse data with grid patterns, such as *images* (discussed in Chapter 3). Nevertheless, the proposed *semantically infused temporal trajectories* display different characteristics in contrast to images, since these trajectories are *1D sequences*, not 2D pixels (Zheng et al. 2014). Therefore, in *multivariate temporal trajectory* problems, as in this study (discussed in Section 5.7.3), *1D convolutions* can be employed to circumvent this issue (Figure 5.18).

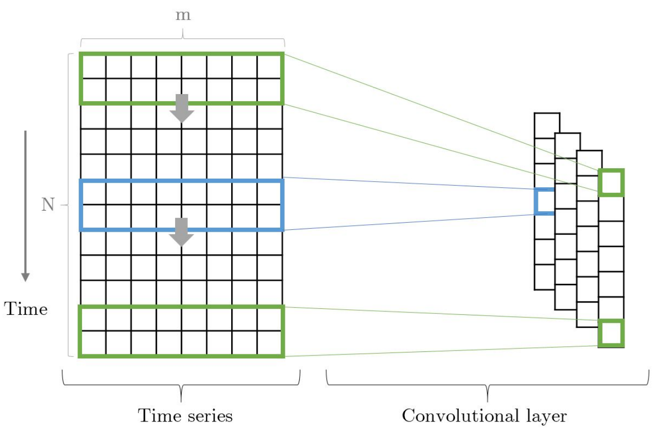


FIGURE 5.18: Example of 1D convolutions for temporal trajectories where N is the number of time steps and m is number of data points in each time step (note that in 1D convolutions kernel width is similar to m)

5.7.4.1 Complementary Integration of LSTMs and CNNs

Even though the objectives of LSTM and CNN are different (i.e., extracting *long-term temporal dependencies* vs. extracting *spatial features*), successful attempts have been reported in the *time series analysis*, *sequence mining* and *signal processing* research areas when the two models are combined (Kim & Cho 2019, Liu et al. 2017, Kim & Cho 2018). Such combinations are feasible, since CNN can typically be utilised as *feature extractors* in any kind of network (Le Guennec et al. 2016). Inspired by such research from *outside the field of LBD*, this study proposes several DNN architectures that employ both LSTM and CNN, as described in Section 5.7.4. The main reason for adopting such architectures is to verify the suitability of both temporal and spatial features in the LBD context.

5.7.5 Design of DNN Models

Considering the strengths of each main building block (i.e., *LSTM* and *CNN*), this section provides details on the proposed DNN architectures used in this study to sift the proposed semantically infused temporal trajectories.

5.7.5.1 Proposed LSTM Architectures

The design of *sequence problems* (similar to this study) can be broadly defined into four categories: *one-to-one*, *one-to-many*, *many-to-one* and *many-to-many* (Gulli & Kapoor

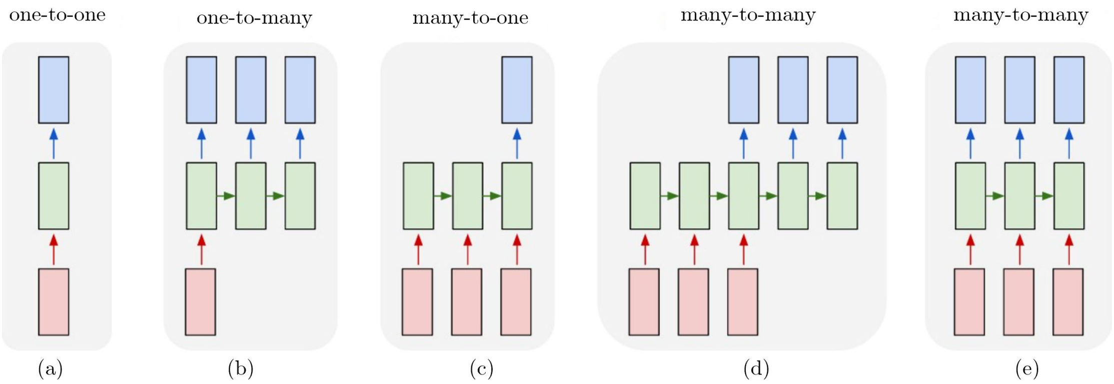


FIGURE 5.19: Types of sequence problems in which rectangles represent vectors and arrows denote functions such as matrix multiplication. (input, output and state of LSTM are presented in red, blue and green, respectively) (Gulli & Kapoor 2017)

2017, Ayyadevara 2019).

- *One-to-one* sequence problems contain a single fixed-sized input and output (e.g., *image classification*). Such problems can directly utilise vanilla processing mode (without *LSTMs*), as denoted in Figure 5.19 (a).

- *One-to-many* sequence problems contain a single fixed-sized input and a sequence output. *Image captioning* can be considered an example of this category. During image captioning, an image is used as the input, and the output made up of multiple words (Figure 5.19 (b)).

- *Many-to-one* sequence problems contain a sequence input and one fixed-sized output, as illustrated in Figure 5.19 (c). One example of this category is *sentiment analysis*, in which a model can determine the sentiment of a sentence as positive or negative.

- *Many-to-one* sequence problems can be occur in two possible forms; one with a desynced input sequence and output sequence (Figure 5.19 (d)), and one with a synced input sequence and output sequence (Figure 5.19 (e)). For instance, consider a *machine translation* problem in which LSTMs read a sentence in English and output a French sentence. This denotes the first type of many-to-one sequence architectures. The latter type can be illustrated through the example of *video classification*, wherein each frame of the video is labelled by the constructed model.

This study falls under *many-to-one* type architectures in which the six semantically infused temporal trajectories are considered the *input* (i.e., *many inputs*) and the prediction probability that denotes the potential of a local topic to become a new knowledge linkage is the *output* (i.e., *one output*). With this in mind, the study proposes the following three LSTM architectures in order to analyse the temporal trajectories. In the LSTM designs, dropout layers are used to prevent model overfitting, and the Adam algorithm is used to optimise the loss function.

LSTM model architecture 1 (*LSTM_1*) uses the six semantically infused temporal trajectories prepared in the form of multivariate time series (discussed in Section 5.7.3) to construct its input layer. This model incorporates vertically stacked LSTM layers through the use of two LSTM layers to sift the temporal trajectories, as depicted in Figure 5.20. The colour differences between the two LSTM layers in Figure 5.20 indicate that the first LSTM layer will output the full sequence of hidden states, (h_1, h_2, ..., h_n), where n is the final time step, while the second LSTM layer will only output the hidden state at the final time step. Subsequently, the LSTM output is concatenated with the two frequency heuristics (i.e., *LFH* and *GFH*, discussed in Section 5.5.5) followed by a fully connected layer (Figure 5.20). The final output of the model is the predicted probability of a local topic becoming a new knowledge linkage, which is denoted through *sigmoid* in Figure 5.20.

Much like *LSTM_1*, the remaining two LSTM model architectures (*LSTM_2* and *LSTM_3*) follow the idea of stacked LSTM. The only difference between the structure of *LSTM_2* and *LSTM_3* and that of *LSTM_1* is in the number of LSTM layers included in the architectures. Specifically, *LSTM_2* uses three LSTM layers, whereas *LSTM_3* incorporates four. As with *LSTM_1*, the output of these two models is a prediction probability that denotes the potential of a local topic becoming a new knowledge linkage.

5.7.5.2 Proposed CNN Architectures

As in the case of LSTM architectures, the *many-to-one* setting is utilised in CNN, as illustrated in Figure 5.21. More specifically, this study employs two 1D convolution layers followed by a max-pooling layer as *feature extractors* (see Chapter 3 for details). Subsequently, the *feature maps* constructed through convolution layers and filtered through a pooling layer are passed to the flatten layer. The output of CNN is concatenated with

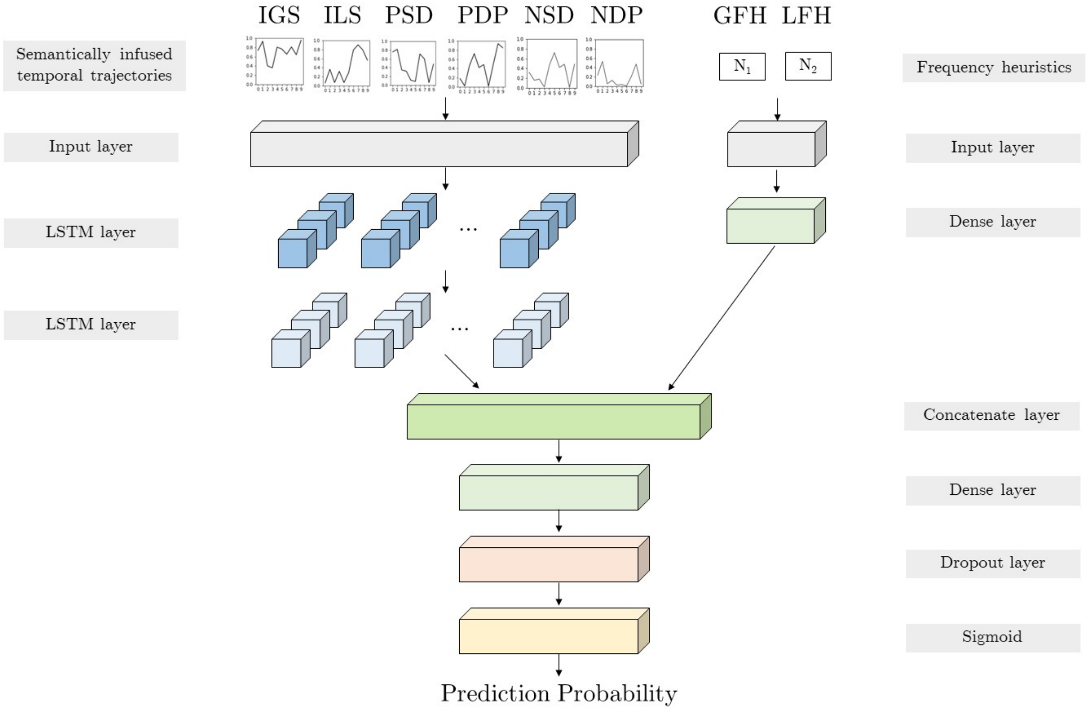


FIGURE 5.20: LSTM_1 model architecture

the two frequency heuristics (*LFH* and *GFH*) and then connected via a fully connected layer. As with LSTM models, the final output of this model is a prediction probability indicating the extent to which a local topic is likely to become a novel knowledge linkage.

5.7.5.3 Proposed LSTM and CNN Architectures

In addition to the use of *LSTMs* and *CNNs* separately (discussed above), this study also leverages the idea of hybrid architectures, which use both *LSTMs* and *CNNs* as their main building blocks. The ultimate motive of these proposed architectures is to verify the suitability of both *temporal* and *spatial* features in the context of LBD. In this regard, this study proposes two DNN architectures, namely *CNN_LSTM* and *LSTM_CNN*.

In *CNN_LSTM model architecture* (Figure 5.22), CNN layers are employed first to extract features from the temporal trajectories. Next, these extracted features (which are represented as feature maps) are passed to the LSTM layers. The LSTM layers are vertically stacked, as discussed in *LSTM_2*. The output of the LSTM is connected to the two frequency heuristics via a concatenation layer, followed by a fully connected

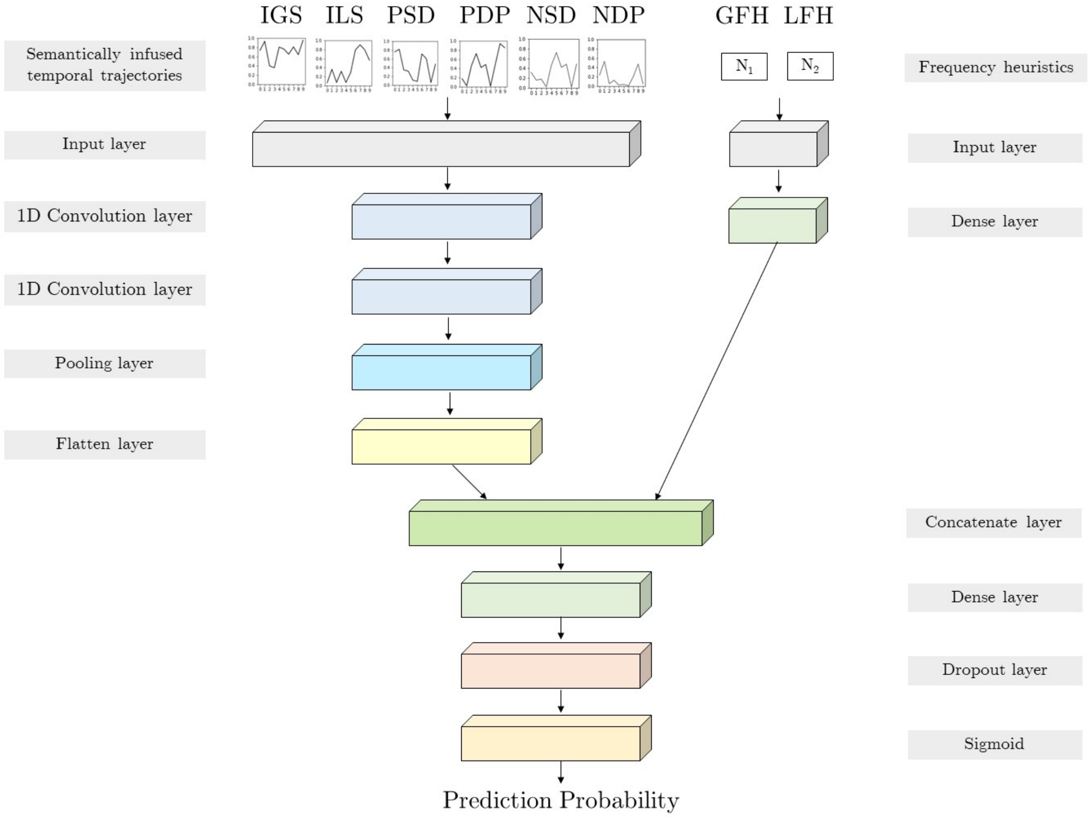


FIGURE 5.21: CNN model architecture

layer. As in the case of other models, the output of this model is a probability that denotes the potential of a local topic becoming a novel knowledge linkage. Note that a pooling layer is not employed in this model just after the convolution layers (as in the proposed CNN model, depicted in Figure 5.21). The main reason for this is that using a pooling layer reduces the amount of inputs passed on to the LSTM layer. Since LSTMs typically excel at processing with sequences of any length, inserting a pooling layer is not necessarily important in this instance.

The functionalities of the proposed *LSTM_CNN model architecture* can be considered the inverse of those in the previous model (see Figure 5.23). In essence, this model first extracts temporal features from the temporal trajectories using vertically stacked three LSTM layers. Subsequently, these temporal features are passed to the convolution layers for the extraction of spatial features. In contrast to the previous model, a pooling layer is employed prior to the fully connected layer to make feature maps translation invariance using max-pooling as the down-sampling operation (discussed in Chapter 3).

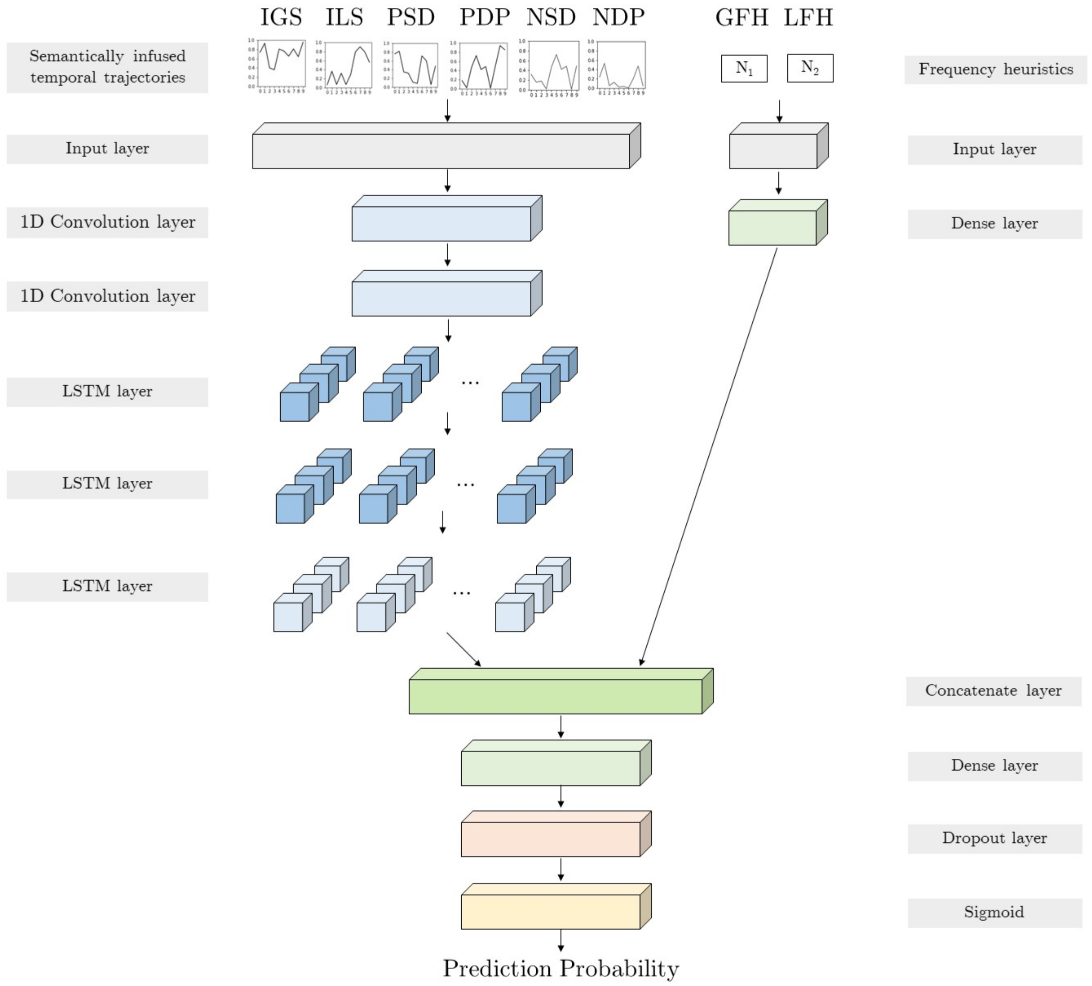


Figure 5.22: CNN_LSTM model architecture

As with other models, the output of this architecture is a probability that expresses the extent to which a local topic will become a new knowledge linkage.

5.8 Feature-based Trajectory Model (FTM)

This section describes the second proposed LBD model of this study, the *Feature-based Trajectory Model (FTM)*. The only difference between *FTM* and *DTM* is in the underlying process used by the model to analyse the derived semantically infused temporal trajectories. In essence, FTM employs the *traditional ML setting*, using hand-crafted features derived from temporal trajectories to perform knowledge discovery. The first part of this section presents an overview of the model and the reasons for adopting it, while the latter part of this section presents details relating to this model's setting.

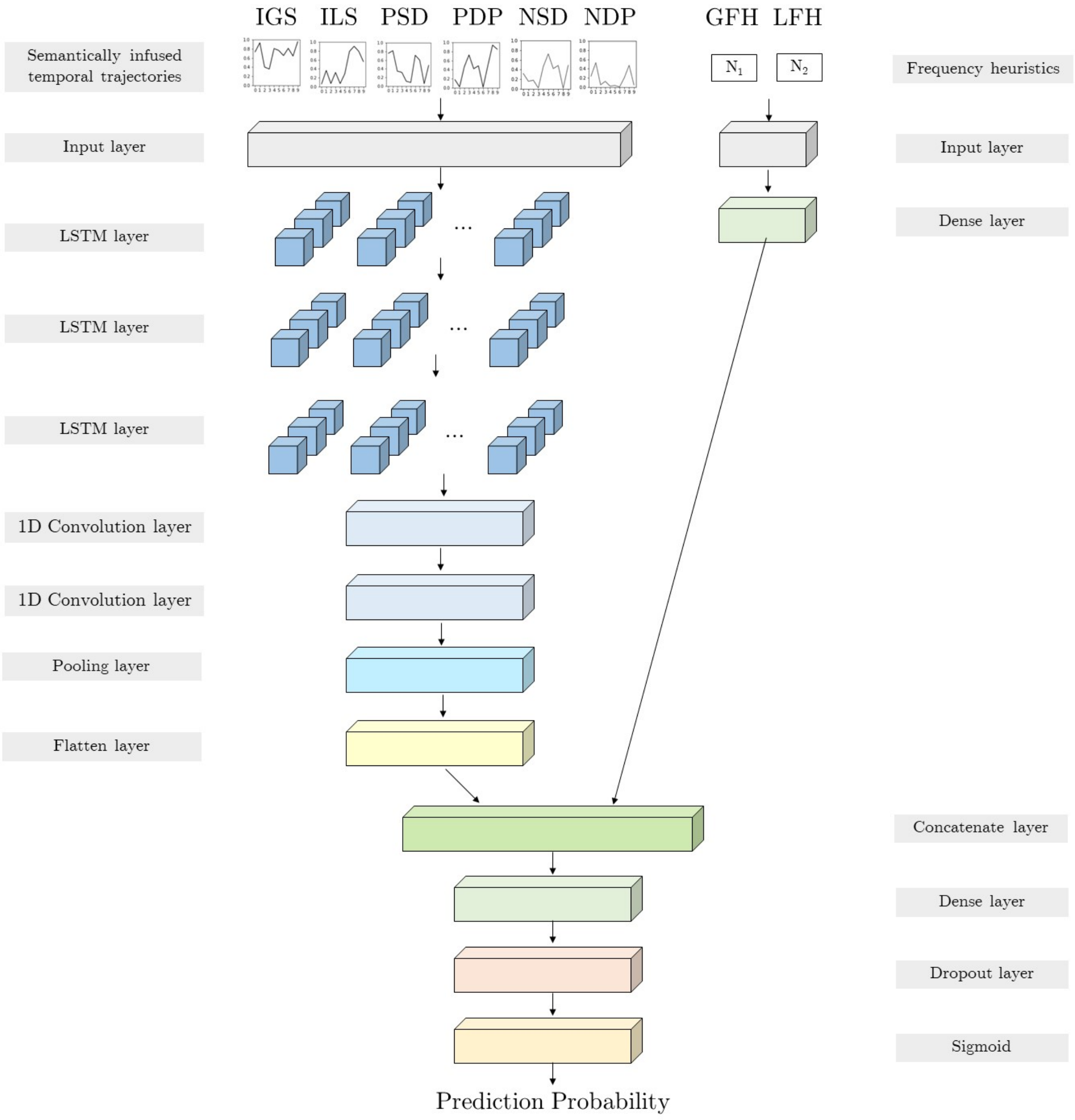


FIGURE 5.23: LSTM_CNN model architecture

5.8.1 Rationale

The key aim of this model is similar to that of *DTM*, as discussed in Section 5.7.1. Succinctly, FTM explores the need for *detailed temporal analysis* by incorporating *multiple characteristics* to elicit novel knowledge linkages. In doing so, this model exploits the traditional feature-based ML setting, in contrast to the deep learning setting that is used in *DTM* (discussed in Section 5.7). More specifically, this study exploits salient/noteworthy features from the proposed semantically infused temporal trajectories by focusing on both the *trajectory values* and *trajectory shape*. Like DTM, this model can

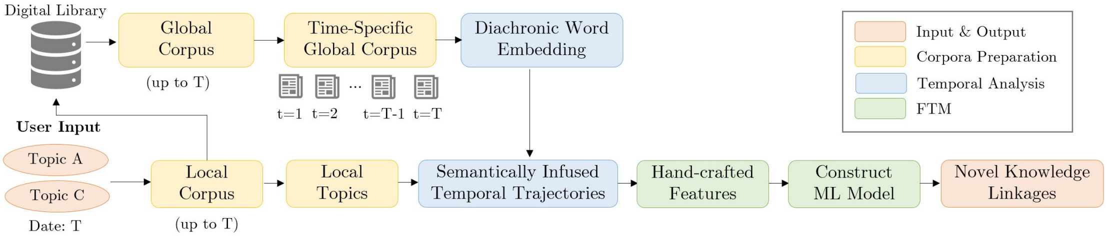


Figure 5.24: Schematic overview of the Feature-based Trajectory Model (FTM)

also be considered a model which provides the opportunity to understand the *direct uses* of the proposed semantically infused temporal trajectories, with the ultimate goal of detecting novel knowledge linkages.

5.8.2 Overview of Proposed LBD Model (FTM)

This section provides a high-level overview of the main phases involved in this model. The objective of this model is similar to that of *DTM*. That is, given two user-defined input topics A and C, and a date T, the model seeks novel knowledge bridges that are likely to occur in the future by analysing literature up to time T. Like *DTM*, this LBD model follows the initial phases of the proposed LBD framework, which are constituted of phases such as *corpora preparation*, *construction of diachronic word embeddings* and *semantic shifts extraction* (discussed in Section 5.2). This model follows traditional ML techniques by manually extracting features from the six extracted semantically infused temporal trajectories (Figure 5.24). *Features* are an important consideration in pattern recognition tasks and are also related to prediction performance (Fu 1968). Thus, this study focuses on both the key components of the proposed temporal trajectories, which are their *values* and *shapes*.

5.8.3 Hand-crafted Features

When analysing a trajectory, both its *values* and *shape* play a crucial role. For instance, consider the example trajectories denoted as t_1, t_2, ..., t_{10} in Figure 5.25, where the *trajectory values* represent the *cosine similarity*. This example can be analysed using three different scenarios: 1) *analysing only trajectory values*, 2) *analysing only trajectory shapes* and 3) *analysing both trajectory values and trajectory shapes.*

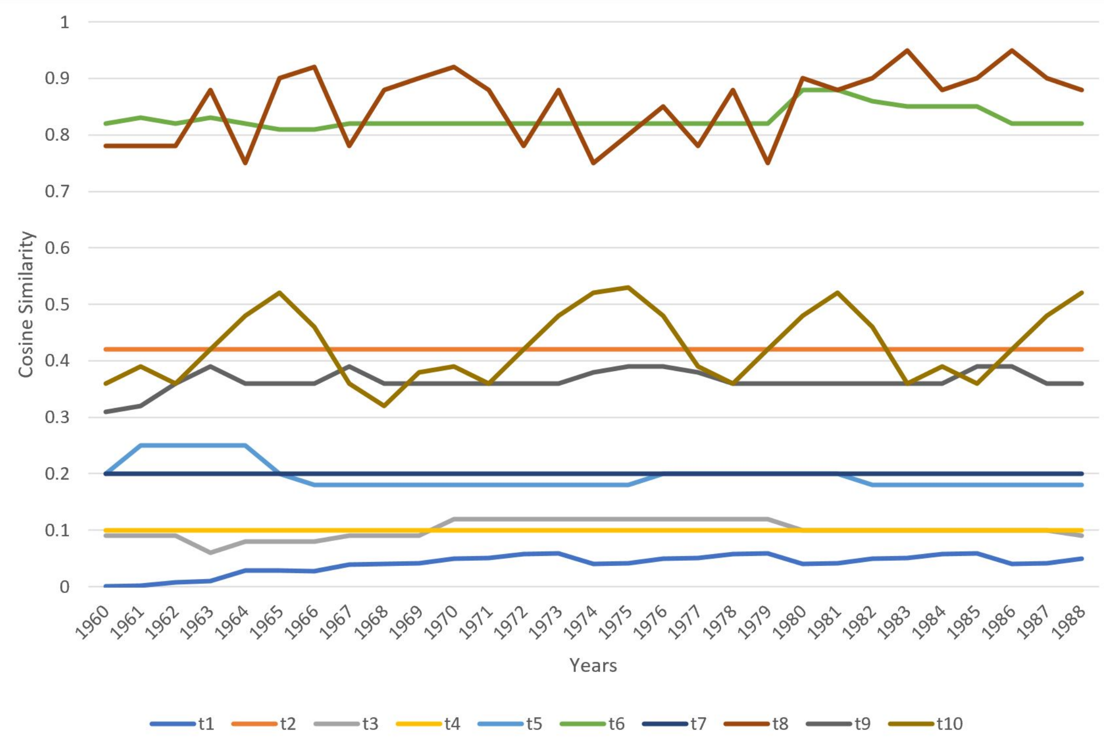


Figure 5.25: Nature of trajectory values and shapes

When analysing only the trajectory values from Figure 5.25 (i.e., *scenario 1*), the most obvious observation is that t_8 and t_6 have higher semantic similarity in comparison to remaining trajectories. When looking at Figure 5.25, by focusing on *scenario 2*, it is visible that t_8 and t_{10} demonstrate distinguishing patterns in their trajectory shapes when compared to remaining trajectories. However, when focusing on *scenario 3*, it can be observed that the scenario reflects not only the two observations in *scenario 1* and *2*, but also further interpretations of the trajectories. For example, in *scenario 3*, it is possible to say that even though both t_8 and t_{10} have distinguishing trajectory shapes, t_8 and t_{10} are entirely different in terms of trajectory values (i.e., the trajectory values of t_8 have high semantic similarities, whereas the trajectory values of t_{10} have low semantic similarities). Therefore, it is important to accommodate both the trajectory's *values* and *shape* in order to perform a rich pattern mining of the trajectories. Following this reasoning, this study attempts to sift the derived semantically infused temporal trajectories using signals from both trajectory values and shape to differentiate potential novel knowledge linkages from the remaining scientific topics in the literature. In essence, the values and shape of the semantically infused temporal trajectories are the main focus points of this analysis.

With this in mind, given a temporal trajectory $t = (t_1, t_2, ..., t_n)$ and assuming t_1 is the first occurrence of a concept in the literature, this study considers the following descriptive statistics to represent the feature category *trajectory values*.

- *Minimum:* indicating the lowest value of the trajectory *t*.
- *Index of the minimum:* indicating the point at which the lowest value is encountered in the trajectory *t*.
- *Maximum:* indicating the highest value of the trajectory *t*.
- *Index of the maximum:* indicating the point at which the highest value is encountered in the trajectory *t*.
- *Mean*: indicating the average of the values in the trajectory *t*.
- *Median:* indicating the median (Q_2) of the values in the trajectory *t*.
- *Standard deviation:* indicating the standard deviation of the values in the trajectory *t*.
- *Variance:* indicating the variance of the values in the trajectory *t*.
- *Sum:* indicating the total of the values in the trajectory *t*.
- *Count above mean:* indicating how many values are above the mean of the trajectory *t*.
- *Count below mean:* indicating how many values are below the mean of the trajectory *t*.
- *Length ratio:* indicating the proportion of unique values in the trajectory *t*.
- *Subsequence above mean:* indicating the length of the longest consecutive sub-sequence in the trajectory *t* in which the values are higher than its mean.
- *Subsequence below mean:* indicating the length of the longest consecutive sub-sequence in the trajectory *t* in which the values are lower than its mean.
- *Mean change:* indicating the mean of the differences between subsequent values in the trajectory *t*, which can be denoted as in equation 5.14.

$$\frac{1}{n-1}\sum_{i=1}^{n-1} t_{i+1} - t_i = \frac{1}{n-1}(t_n - t_1) \tag{5.14}$$

- *Absolute mean change:* indicating the mean of the absolute differences between subsequent values in the trajectory *t*, which can be denoted as in equation 5.15.

$$\frac{1}{n}\sum_{i=1}^{n-1} \mid t_{i+1} - t_i \mid \tag{5.15}$$

This study considers the following features to denote the *trajectory shape*-based features.

- *Smoothness:* indicating the roughness of the trajectory line using the equation 5.16 in which d_i is an element of the first-order differences vector of t (t^{FOD} = (d_1, d_2, ..., $d_{n\text{-}1}$)), where $d_i = t_{i+1} - t_i$.

$$\frac{\sqrt{\frac{1}{n-2}\sum_{i=1}^{n-1}(d_i - \bar{d})^2}}{\mid \bar{d} \mid} \tag{5.16}$$

- *Skewness:* indicating the skewness of the trajectory *t*, using the adjusted Fisher-Pearson standardised moment coefficient G1 (Joanes & Gill 1998, Doane & Seward 2011) (equation 5.17).

$$\frac{\sqrt{n(n-1)}}{n-2} = \left[\frac{\frac{1}{n}\sum_{i=1}^{n}(t_i - \bar{t})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(t_i - \bar{t})^2\right)^{\frac{3}{2}}}\right] \tag{5.17}$$

- *Number of peaks:* indicating peaks inside the trajectory *t* based on peak properties. More specifically, a peak is considered to be a local maximum based on comparison with its neighbours in terms of their height, prominence, width, threshold and mutual distance (Bills et al. 2020).

- *Length:* indicating the length of the trajectory *t* starting from the first occurrence of the concept in the literature.

- *Trend:* indicating the slope of the values in the trajectory *t* using *ordinary least-squares approximation* (i.e., *m* in *f(x)* = *mx* + *b*)).

For each of the six semantically infused temporal trajectories of a local topic lt_i, the discussed hand-crafted features are extracted in the form of feature profiles FP^{x}, where x denotes the semantic shift variant discussed in Section 5.5.4.

$\mathrm{FP}^{\mathrm{IGS}}(lt_i) = (\mathrm{f_1}^{\mathrm{IGS}}(lt_i), \mathrm{f_2}^{\mathrm{IGS}}(lt_i), ..., \mathrm{f_{n\text{-}1}}^{\mathrm{IGS}}(lt_i), \mathrm{f_n}^{\mathrm{IGS}}(lt_i))$

$\mathrm{FP}^{\mathrm{ILS}}(lt_i) = (\mathrm{f_1}^{\mathrm{ILS}}(lt_i), \mathrm{f_2}^{\mathrm{ILS}}(lt_i), ..., \mathrm{f_{n\text{-}1}}^{\mathrm{ILS}}(lt_i), \mathrm{f_n}^{\mathrm{ILS}}(lt_i))$

$$\mathrm{FP}^{\mathrm{PSD}}(lt_i) = (\mathrm{f_1}^{\mathrm{PSD}}(lt_i),\ \mathrm{f_2}^{\mathrm{PSD}}(lt_i),\ ...,\ \mathrm{f_{n\text{-}1}}^{\mathrm{PSD}}(lt_i),\ \mathrm{f_n}^{\mathrm{PSD}}(lt_i))$$

$$\mathrm{FP}^{\mathrm{PDP}}(lt_i) = (\mathrm{f_1}^{\mathrm{PDP}}(lt_i),\ \mathrm{f_2}^{\mathrm{PDP}}(lt_i),\ ...,\ \mathrm{f_{n\text{-}1}}^{\mathrm{PDP}}(lt_i),\ \mathrm{f_n}^{\mathrm{PDP}}(lt_i))$$

$$\mathrm{FP}^{\mathrm{NSD}}(lt_i) = (\mathrm{f_1}^{\mathrm{NSD}}(lt_i),\ \mathrm{f_2}^{\mathrm{NSD}}(lt_i),\ ...,\ \mathrm{f_{n\text{-}1}}^{\mathrm{NSD}}(lt_i),\ \mathrm{f_n}^{\mathrm{NSD}}(lt_i))$$

$$\mathrm{FP}^{\mathrm{NDP}}(lt_i) = (\mathrm{f_1}^{\mathrm{NDP}}(lt_i),\ \mathrm{f_2}^{\mathrm{NDP}}(lt_i),\ ...,\ \mathrm{f_{n\text{-}1}}^{\mathrm{NDP}}(lt_i),\ \mathrm{f_n}^{\mathrm{NDP}}(lt_i))$$

where $\mathrm{f_i}$ denotes a hand-crafted feature and n is the number of hand-crafted features used in this model. Finally, the six feature profiles of local topic lt_i are aggregated to construct the final feature profile of lt_i, as defined in equation 5.18.

$$\mathrm{FP}(lt_i) = \mathrm{FP}^{\mathrm{IGS}}(lt_i) \cup \mathrm{FP}^{\mathrm{ILS}}(lt_i) \cup \mathrm{FP}^{\mathrm{PSD}}(lt_i) \cup \mathrm{FP}^{\mathrm{PDP}}(lt_i) \cup \mathrm{FP}^{\mathrm{NSD}}(lt_i) \cup \mathrm{FP}^{\mathrm{NDP}}(lt_i) \tag{5.18}$$

These constructed *feature profiles* along with the *two frequency heuristics* are utilised in a traditional ML framework (discussed in Chapter 3) to predict the probability with which each local topic lt_i will become a novel knowledge linkage (as in the case of the *DNN models* proposed in *DTM*).

5.9 Trajectory Alignment Model (TAM)

The purpose of this section is to describe the third proposed LBD model, which is the *Trajectory Alignment Model (TAM)*. This model is different from *DTM* (discussed in Section 5.7) and *FTM* (discussed in Section 5.8) in terms of its *rationale*, *objectives* and *questions*. These differences are described in Section 5.9.1. Subsequently, a high-level overview of the proposed model is presented, in which the key functionalities of its major phases are outlined. The remaining part of this section describes each of these major phases of *TAM* in detail. The major phases are as follows: the construction of the *template repository* using large-scale actual novel knowledge linkages, the alignment of the temporal trajectories in the template repository in the form of *docking engine*, and the use of ML techniques to sift the extracted patterns in the trajectory alignment process so as to distinguish novel knowledge linkages from the remaining scientific topics.

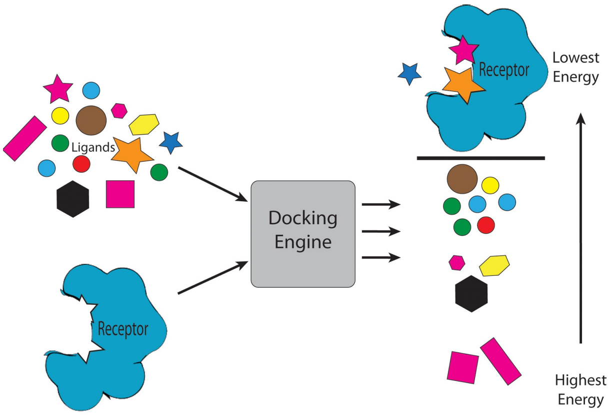


FIGURE 5.26: Schematic overview of molecular docking used in structure-based drug design (Jacob et al. 2012)

5.9.1 Rationale

The ultimate purpose underlying the LBD research is the elicitation of meaningful patterns that could be employed to identify novel knowledge linkages in the scientific literature. To this end, previous LBD studies have employed a wide spectrum of techniques, from basic statistical methods to complex graph-theoretic methods. Nevertheless, to the best of our knowledge, *none* of the previous LBD studies has attempted to learn patterns *relative* to *actual novel knowledge linkages* that could serve as a meaningful metric in identifying whether the *potential candidates* (or *local topics*) exhibit the same patterns as those demonstrated in actual novel knowledge linkages.

The main impetus for the development of the proposed LBD model came from the approach of *docking* in molecular modelling, which is the most frequently used method in *structure-based drug design* (Ferreira et al. 2015). The easiest way to understand molecular docking is to think of it as a 'lock and key' problem in which the *lock* is the *receiving molecule* (or *receptor*) - most commonly a protein or biopolymer, and the *key* is the *complementary partner molecule* that binds to the receptor (a.k.a. the *ligand*). The purpose of the docking engine is to measure the free energy of binding ΔE between the receptor and a ligand, as illustrated in Figure 5.26. Subsequently, ligands are ranked by ΔE, where a lower ΔE denotes more favourable ligand bindings, while a higher ΔE denotes less favourable bindings (see Figure 5.26) (Jacob et al. 2012). Similarly, the idea of this model is to bind the trajectories of local topics (analogously, *ligands*) with the trajectories of actual novel knowledge linkages (analogously, *receptors*) to deduce some cost metric that denotes whether the binding is less or more favourable.

The idea of *docking* as described above could be highly beneficial in the context of LBD, for the following two reasons.

- Firstly, it is fair to assume that the concealed patterns of potential novel linkages sought within LBD studies across several decades are encapsulated in these *actual novel knowledge linkages* (a.k.a. *templates*) since they have been realised in the real-world with time, thereby providing a platform which is rich in cues for knowledge discovery. Nevertheless, these patterns (encapsulated in templates) may not be *salient* when they are considered as *separate entities*. This is where the idea of *relativity* may assist. In other words, instead of directly mining these templates, one could verify the extent to which the *patterns of local topics* match with the *patterns of these templates* (as in the case of *docking*).

- There are several discussions in the *theoretical LBD literature* that novel knowledge can exist in multiple forms. Despite decades of research in LBD, only a limited number of such forms have been identified (Davies 1989). The main reason for this could be the complexity of natural language usage that causes intricate structures in the literature. Thus, there could be several hundred or even thousands of other forms that are not salient, and yet to be discovered in such theoretical LBD studies. Nevertheless, the idea of maintaining a large collection of templates in a *template repository* may assist to overcome this constraint to some extent. This is because such a repository could accommodate a large number of novel knowledge linkage forms in a single place.

Considering all these facts, this LBD model utilises a *large-scale and data-driven 'template docking' approach* to discover novel knowledge linkages, namely the *trajectory alignment model*. More precisely, this study attempts to answer the following question: *'can actual novel knowledge linkages serve as templates to deduce the potentiality of a local topic becoming a new knowledge linkage?'*. To the best of our knowledge, this is the first study in the LBD field that employs the idea of deriving patterns from actual novel knowledge linkages in the form of *docking* to deduce the potentiality of local topics representing new knowledge.

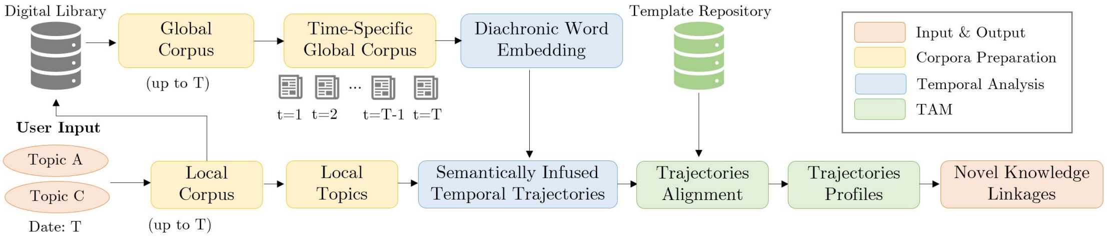


FIGURE 5.27: Schematic overview of the Trajectory Alignment Model (TAM)

5.9.2 Overview of Proposed LBD Model (TAM)

This section outlines the key functionalities of the main phases in this proposed LBD model. Like *DTM* and *FTM*, the objective of this model (given two user-defined input topics A and C, and a date T) is to analyse the literature up to time T and detect the temporally charged novel knowledge bridges that are most likely to occur in the future. The initial phases of this LBD model (i.e., *corpora preparation*, *construction of diachronic word embeddings* and *semantic shifts extraction*) follow the blueprint of the proposed LBD framework discussed in Section 5.2 (Figure 5.27). Thus, this section provides an overview of the steps specific to *TAM* and aligns with the study's rationale.

This study revolves around the *template repository* (Figure 5.27), which comprises novel knowledge linkages that have been realised in the real-world. The *template repository* maintains a collection of temporal trajectories that denote how the semantics of actual novel knowledge linkages have evolved across time. These serve as *templates* that can be used to analyse the trajectories of local topics (Figure 5.27). Otherwise stated, this study scrutinises how closely the semantically infused temporal trajectories of the local topics resemble the trajectories in the template repository. This phase is termed *trajectory alignment*, as shown in Figure 5.27. Next, the output of the trajectory alignment is leveraged to construct a profile for each local topic denoting the similarity or difference between each local topic with the templates in the template repository. Finally, these profiles of local topics are analysed using ML techniques, in order to discover potential novel knowledge relating to the user-defined input topics A and C.

5.9.3 Constructing the Template Repository

One of the main components of this model is the construction of the *template repository* which will be used as the *core analysis source* in the remaining phases (see Figure 5.27).

To this end, this study leverages *historical test cases* to create a rich and comprehensive platform from which to extract multiple forms of novel knowledge (i.e., *templates*).

For instance, consider the historical test case of *fish oil and Raynaud's disease*, where Swanson initially identified three novel conceptual bridges in order to meaningfully connect these two knowledge isolations in 1986 (Swanson 1986). The prominence of this test case is due to the fact that the two subjects (fish oil and Raynaud's disease) were *complementary* but *non-interactive* before Swanson's LBD-facilitated discovery (Figure 3.3). That is, articles relating to these two knowledge fragments had never mentioned, cited or co-cited each other. With time, researchers identify more and more creative ways to combine such knowledge fragments. Thus, such historical test cases provide a rich and comprehensive platform from which to extract multiple forms of novel knowledge in a data-driven manner.

Simply put, this study first identifies the actual novel knowledge linkages in such historical test cases that got realised over time. Next, this study extracts the same six types of semantically infused temporal trajectories discussed in Section 5.5.4, before the selected historical test case gets bridged (i.e., when it was *non-interactive*). This thesis calls these derived semantically infused temporal trajectories of such identified new knowledge linkages in the selected historical test case *historical trajectories*.

5.9.4 Trajectory Alignment

The derived historical trajectories represent their temporal behaviour in semantic space across time, before the bridging of the two knowledge isolations A and C in the historical test case. Therefore, these historical trajectories serve as *templates* of potential novel knowledge. Simply put, this study assumes that potential novel knowledge relating to the local topics is correlated with the temporal behaviour of historical trajectories. Thus, for each local topic extracted using a user-defined query, this study measures how closely their temporal trajectories resemble the historical trajectories. For this purpose, this study utilises Dynamic Time Wrapping (DTW), for the following two reasons.

- DTW measures the similarity between two trajectories that might differ in time scale, but which are similar in shape (Figure 5.28) (Keogh & Ratanamahatana 2005). This

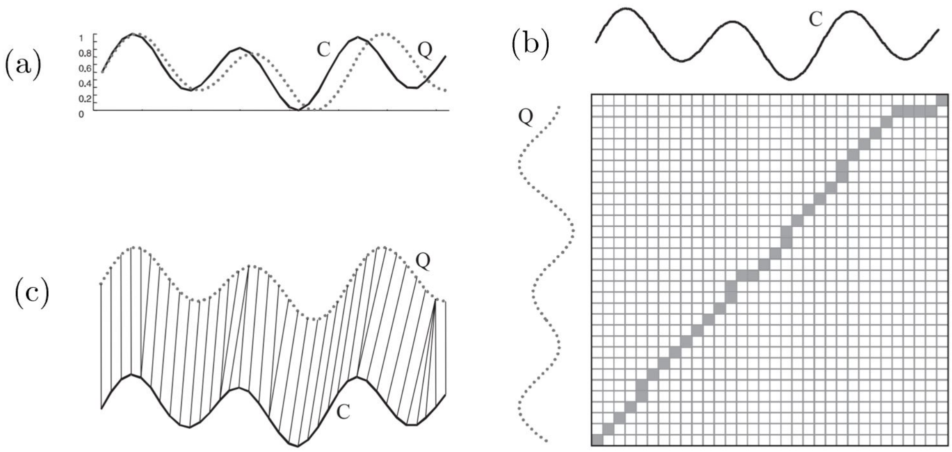


Figure 5.28: Dynamic time warping in the context of (a) two sequences C and Q that are similar in shape but different in time scale, (b) construction of a cost matrix (warping matrix) to search optimal warping path, illustrated using solid squares, and (c) optimal alignment found, using a cost matrix (Keogh & Ratanamahatana 2005)

is in line with the objective of this study, i.e., identifying the extent to which the trajectories of local topics resemble historical trajectories.

- Unlike other measurements (such as *Euclidean distance* and *edit distance*), DTW has proven to be an exceptionally strong distance measure for time series (Kate 2016).

5.9.4.1 Dynamic Time Warping (DTW)

Consider two temporal trajectories t_1 and t_2, where the DTW algorithm first defines a local cost matrix $C \in \mathbb{R}^{|\mathrm{t}_1| \times |\mathrm{t}_2|}$ as in equation 5.19, where $\| t_1[i] - t_2[j] \|$ denotes a distance between two points in the trajectories (Radinsky et al. 2011).

$$C_{i,j} = \| t_1[i] - t_2[j] \|, i \in \langle 1... \mid t_1 \mid \rangle, j \in \langle 1... \mid t_2 \mid \rangle \tag{5.19}$$

After defining this cost matrix, DTW creates an alignment path p that minimises the cost over the constructed cost matrix. This alignment p is known as the *warping path* and can be expressed as a sequence of point pairs from the two trajectories p = ($pair_1$, $pair_2$, ..., $pair_k$), in which $pair_l$ = $(i{,}j) \in \langle 1...|t_1|\rangle \times \langle 1...|t_2|\rangle$ is the index of points in t_1 and t_2, respectively. Each subsequent pair in the warping path p preserves the point ordering in t_1 and t_2, while enforcing the initial points and endpoints of the warping path to become the initial points and endpoints of t_1 and t_2. For each warping path p, its cost is calculated as $c(p) = \sum_{l=1}^{k} C(pair_l)$. The DTW is the minimum optimal warping path of all possible warping paths P, as denoted in equation 5.20 (Radinsky et al. 2011).

$$DTW(t_1, t_2) = min\{c(p) \mid p \in P^{|t_1 \times t_2|}\} \tag{5.20}$$

Since DTW is computationally expensive, a dynamic programming algorithm is typically used to calculate the optimal warping path of two trajectories (Radinsky et al. 2011). According to this study, t_1 and t_2 (discussed above) will be the *trajectory of a local topic* and the *historical trajectory*, as denoted in Figure 5.29. In summary, for each of the six semantically infused temporal trajectories of a local topic lp_i, the aforementioned trajectory alignment process was performed with the corresponding variant of the historical trajectories. In essence, the local topic lp_i can be denoted as a *cost profile* as summarised below.

$\mathrm{CP}(lp_i)^{\mathrm{IGS}} = (C(\mathrm{TJ}^{\mathrm{IGS}}(lp_i), \mathrm{HT}_1^{\mathrm{IGS}}), ..., C(\mathrm{TJ}^{\mathrm{IGS}}(lp_i), \mathrm{HT}_{\text{N-1}}^{\mathrm{IGS}}), C(\mathrm{TJ}^{\mathrm{IGS}}(lp_i), \mathrm{HT}_{\mathrm{N}}^{\mathrm{IGS}}))$

$\mathrm{CP}(lp_i)^{\mathrm{ILS}} = (C(\mathrm{TJ}^{\mathrm{ILS}}(lp_i), \mathrm{HT}_1^{\mathrm{ILS}}), ..., C(\mathrm{TJ}^{\mathrm{ILS}}(lp_i), \mathrm{HT}_{\text{N-1}}^{\mathrm{ILS}}), C(\mathrm{TJ}^{\mathrm{ILS}}(lp_i), \mathrm{HT}_{\mathrm{N}}^{\mathrm{ILS}}))$

$\mathrm{CP}(lp_i)^{\mathrm{PSD}} = (C(\mathrm{TJ}^{\mathrm{PSD}}(lp_i), \mathrm{HT}_1^{\mathrm{PSD}}), ..., C(\mathrm{TJ}^{\mathrm{PSD}}(lp_i), \mathrm{HT}_{\text{N-1}}^{\mathrm{PSD}}), C(\mathrm{TJ}^{\mathrm{PSD}}(lp_i), \mathrm{HT}_{\mathrm{N}}^{\mathrm{PSD}}))$

$\mathrm{CP}(lp_i)^{\mathrm{PDP}} = (C(\mathrm{TJ}^{\mathrm{PDP}}(lp_i), \mathrm{HT}_1^{\mathrm{PDP}}), ..., C(\mathrm{TJ}^{\mathrm{PDP}}(lp_i), \mathrm{HT}_{\text{N-1}}^{\mathrm{PDP}}), C(\mathrm{TJ}^{\mathrm{PDP}}(lp_i), \mathrm{HT}_{\mathrm{N}}^{\mathrm{PDP}}))$

$\mathrm{CP}(lp_i)^{\mathrm{NSD}} = (C(\mathrm{TJ}^{\mathrm{NSD}}(lp_i), \mathrm{HT}_1^{\mathrm{NSD}}), ..., C(\mathrm{TJ}^{\mathrm{NSD}}(lp_i), \mathrm{HT}_{\text{N-1}}^{\mathrm{NSD}}), C(\mathrm{TJ}^{\mathrm{NSD}}(lp_i), \mathrm{HT}_{\mathrm{N}}^{\mathrm{NSD}}))$

$\mathrm{CP}(lp_i)^{\mathrm{NDP}} = (C(\mathrm{TJ}^{\mathrm{NDP}}(lp_i), \mathrm{HT}_1^{\mathrm{NDP}}), ..., C(\mathrm{TJ}^{\mathrm{NDP}}(lp_i), \mathrm{HT}_{\text{N-1}}^{\mathrm{NDP}}), C(\mathrm{TJ}^{\mathrm{NDP}}(lp_i), \mathrm{HT}_{\mathrm{N}}^{\mathrm{NDP}}))$

in which $C(,)$ is the cost of the optimal warping path, $\mathrm{HT}_{\mathrm{i}}^{\mathrm{x}}$ is a historical trajectory and N is the number of historical trajectories in the template repository. Figure 5.30 illustrates an example of a *cost profile*.

5.9.5 Extracting Patterns from the Cost Profiles of Local Topics

The derived cost profile of each local topic epitomises its similarity or dissimilarity with the historical trajectories in the template repository. For instance, consider the *cost profiles* of three local topics (lt_1, lt_2 and lt_3) denoted in Figure 5.31 as CP_1, CP_2 and CP_3, assuming that there are only six *historical trajectories* in the template repository,

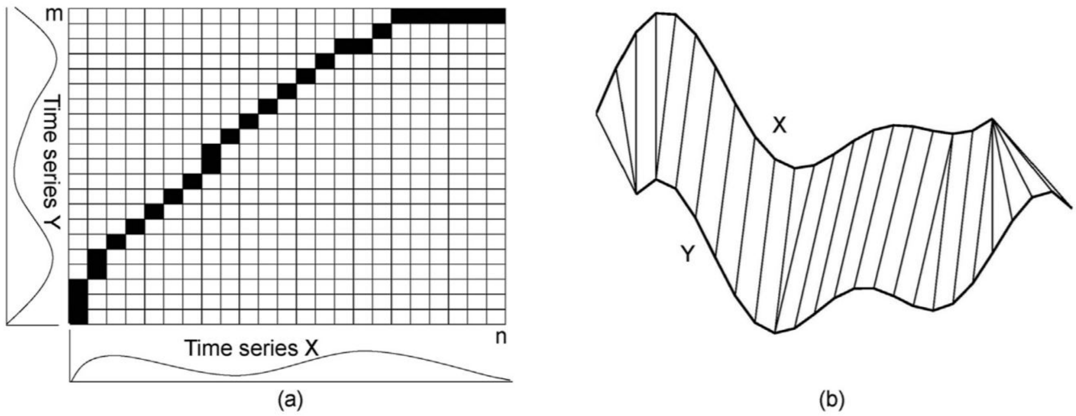


FIGURE 5.29: Trajectory alignment using dynamic time warping where time series X represents a trajectory of local topic and time series Y represents a historical trajectory (a) Cost matrix (b) Optimal warping path (Yang, Scholz, Shao, Wang & Liu 2019)

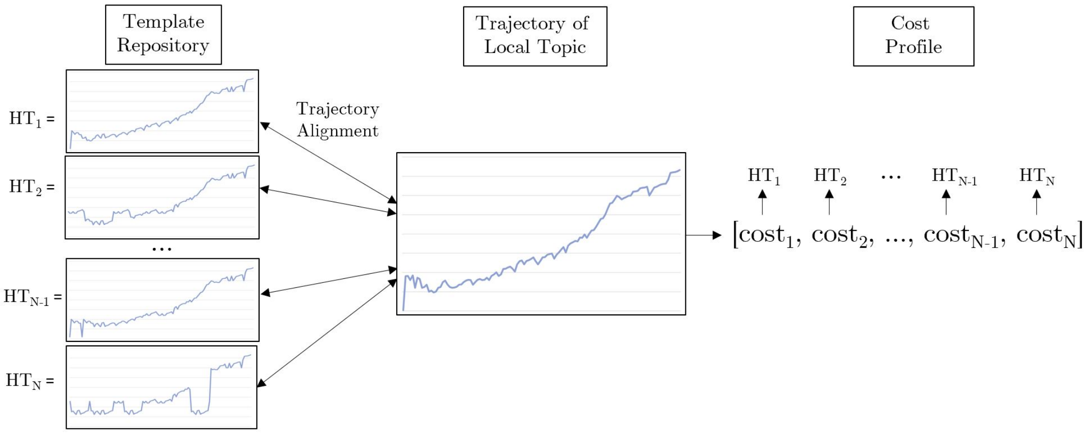

Template Repository
Trajectory of Local Topic
Cost Profile
HT_1 =
HT_2 =
...
HT_{N-1} =
HT_N =
Trajectory Alignment
HT_1 HT_2 ... HT_{N-1} HT_N
$[cost_1, cost_2, ..., cost_{N-1}, cost_N]$

FIGURE 5.30: Construction of cost profile for a local topic using historical trajectories from the template repository

namely HT_1, HT_2, HT_3, HT_4, HT_5 and HT_6. When closely inspecting each of the cost profiles, the following observations can be made.

- *Cost profile CP_1:* The trajectory of the local topic lt_1 is almost identical with that of HT_3, and nearly identical to that of HT_6. The trajectory of lt_1 is quite dissimilar to those of HT_2, HT_4 and HT_5 and mostly dissimilar with that of HT_1.

- *Cost profile CP_2:* This cost profile denotes that the local topic lt_2 has a nearly similar trajectory to that of HT_4. However, unlike lt_1, lt_2 does not have any identical trajectories in the template repository. Moreover, the trajectory of lt_2 demonstrates low dissimilarity with HT_2. The remaining historical trajectories (i.e., HT_1, HT_3, HT_5 and HT_6) are almost dissimilar with the trajectory of lt_2.

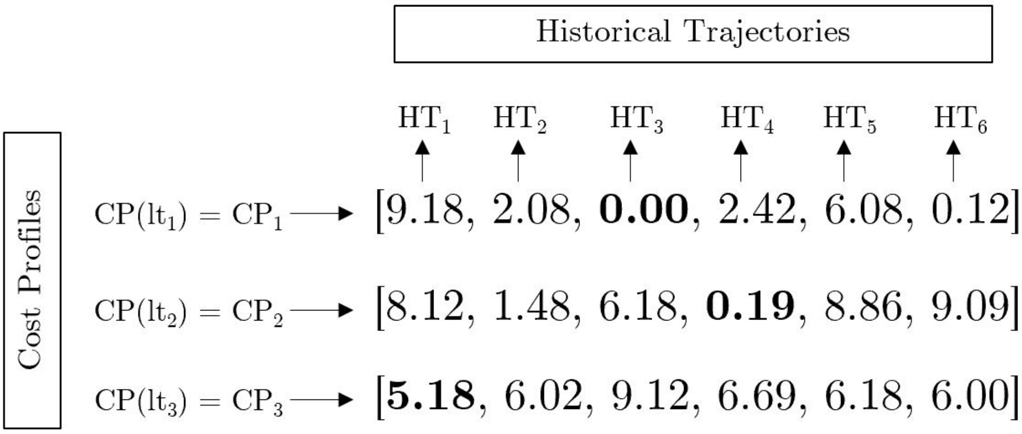

Historical Trajectories
HT$_1$ HT$_2$ HT$_3$ HT$_4$ HT$_5$ HT$_6$
Cost Profiles
CP(lt$_1$) = CP$_1$ → [9.18, 2.08, **0.00**, 2.42, 6.08, 0.12]
CP(lt$_2$) = CP$_2$ → [8.12, 1.48, 6.18, **0.19**, 8.86, 9.09]
CP(lt$_3$) = CP$_3$ → [**5.18**, 6.02, 9.12, 6.69, 6.18, 6.00]

FIGURE 5.31: Simplified example of cost profiles, where the lowest DTW value in each profile is highlighted

- *Cost profile CP$_3$:* When observing the cost profile *CP$_3$*, it is evident that every historical trajectory is almost dissimilar with the trajectory of *lt$_3$*. The historical trajectory displaying the greatest dissimilarity to *lt$_3$* is *HT$_3$*.

Overall, the interpretations of the cost profiles summarise the degree of *similarity* (or *dissimilarity*) between each trajectory of a local topic and the historical trajectories in the template repository. Following this notion, this study extracts the following descriptive features from the cost profiles to capture the extent to which each semantically infused temporal trajectory of the local topics resembles the historical trajectories.

- *Minimum:* denoting the highest cost in the CP$_\mathrm{i}$.
- *Maximum:* denoting the minimum cost in the CP$_\mathrm{i}$
- *Mean:* denoting the average cost in the CP$_\mathrm{i}$
- *Standard deviation:* denoting the dispersion of costs in the CP$_\mathrm{i}$ relative to its mean.
- *Variance:* denoting the variability of costs in the CP$_\mathrm{i}$ from the mean.
- *Q1:* denoting the middle value in the first half of the rank-ordered costs in the CP$_\mathrm{i}$.
- *Q2:* denoting the median of the costs in the CP$_\mathrm{i}$.

- *Q3:* denoting the middle value in the second half of the rank-ordered costs in the $\mathrm{CP_i}$.

Note that the features introduced in *FTM* (discussed in Section 5.8.3) have no relationship with this study, since this model uses *cost profiles*, not *time series*. In essence, the aforementioned descriptive features provide a rough estimation of the similarity (or dissimilarity) between a local topic and the historical trajectories in the template repository. Similarly, for each of the six cost profiles constructed for a local topic lt_i, the aforementioned descriptive features are extracted in the form of:

$CFP^{IGS}(lt_i) = (d^{IGS}(lt_i, ht_1), d^{IGS}(lt_i, ht_2), ..., d^{IGS}(lt_i, ht_{n-1}), d^{IGS}(lt_i, ht_n))$

$CFP^{ILS}(lt_i) = (d^{ILS}(lt_i, ht_1), d^{ILS}(lt_i, ht_2), ..., d^{ILS}(lt_i, ht_{n-1}), d^{ILS}(lt_i, ht_n))$

$CFP^{PSD}(lt_i) = (d^{PSD}(lt_i, ht_1), d^{PSD}(lt_i, ht_2), ..., d^{PSD}(lt_i, ht_{n-1}), d^{PSD}(lt_i, ht_n))$

$CFP^{PDP}(lt_i) = (d^{PDP}(lt_i, ht_1), d^{PDP}(lt_i, ht_2), ..., d^{PDP}(lt_i, ht_{n-1}), d^{PDP}(lt_i, ht_n))$

$CFP^{NSD}(lt_i) = (d^{NSD}(lt_i, ht_1), d^{NSD}(lt_i, ht_2), ..., d^{NSD}(lt_i, ht_{n-1}), d^{NSD}(lt_i, ht_n))$

$CFP^{NDP}(lt_i) = (d^{NDP}(lt_i, ht_1), d^{NDP}(lt_i, ht_2), ..., d^{NDP}(lt_i, ht_{n-1}), d^{NDP}(lt_i, ht_n))$

where ht_i represents historical trajectories, $d(,)$ is the cost of the optimal warping path and n is number of historical trajectories in the template repository.

Subsequently, the six cost feature profiles $CFP^{IGS}(lt_i)$, $CFP^{ILS}(lt_i)$, $CFP^{PSD}(lt_i)$, $CFP^{PDP}(lt_i)$, $CFP^{NSD}(lt_i)$, and $CFP^{NDP}(lt_i)$ are concatenated to represent the final feature set of local topic lt_i, as in equation 5.21.

$$\begin{split}\mathrm{CFP}(lt_i) = CFP^{IGS}(lt_i) \cup CFP^{ILS}(lt_i) \cup CFP^{PSD}(lt_i) \cup CFP^{PDP}(lt_i) \cup CFP^{NSD}(lt_i) \\ \cup\, CFP^{NDP}(lt_i)\end{split} \tag{5.21}$$

Finally, the traditional ML setting (discussed in Chapter 3) is used to analyse the patterns in *cost feature profiles* with the two *frequency heuristics* to distinguish new knowledge. In this regard, this study considers the *prediction probability* in a similar way to the proposed LBD models, *DTM* and *FTM*.

5.10 Experiments

The purpose of this section is to validate the predictive performance of the proposed LBD models using a variety of experiments conducted under different settings. The first part of this section outlines the experimental setup used, including data sources, test cases and other design selections. Subsequently, the results are presented with a discussion of the observations, along with a comparison with the baselines. The latter part of this section describes the strengths of the proposed LBD models while also revisiting the research objectives of this study.

5.10.1 Experimental Setup

This section briefly outlines the setup used for the experiments that was discussed in detail in Chapter 3. In doing so, the first part summarises the dataset and test cases used in the experiments. Subsequently, the setups used in extracting *semantic shifts* (discussed in Section 5.5) and constructing *template repositories* for the proposed LBD model: *TAM* (discussed in Section 5.9) in each of the five selected golden test cases are discussed.

5.10.1.1 Dataset and Test Cases

The main data source used for the experiments was obtained using *MeSH keywords* in *MEDLINE*, as discussed in Chapter 3. To evaluate the effectiveness of the proposed LBD models and to compare it with the baseline models, the following five *golden test cases* (which have commonly been used as evaluation datasets in the previous LBD studies) were utilised.

- *Fish-Oil (FO)* and *Raynaud's Disease (RD)* (Swanson 1986)
- *Magnesium (MG)* and *Migraine Disorder (MIG)* (Swanson 1988)
- *Somatomedin C (IGF1)* and *Arginine (ARG)* (Swanson 1990*a*)
- *Alzheimer's Disease (AD)* and *Indomethacin (INN)* (Smalheiser & Swanson 1996)
- *Schizophrenia (SZ)* and *Calcium-Independent Phospholipase A2 (PA2)* (Smalheiser & Swanson 1998)

Table 5.1: Parameters of semantic shift measures

Measure	**Parameters**
Individual global shifts	–
Individual local shifts	The nearest neighbour count was set to 100 for each local topic (i.e., $\mathcal{K} = 100$)
Pairwise semantic displacement	–
Pairwise distance proximity	The penalising factor was set to 5 to quantify the distance differences of topic A and C (i.e., $\beta = 5$)
Neighbourhood semantic displacement	To compute the constant neighbours of topic A and C, this study considered five years of time span, while sieving neighbouring concepts within a proximity of 500 nearest neighbours (i.e., $\mathcal{W} = 5$ and $\mathcal{K} = 500$)
Neighbourhood distance proximity	Similar to *Neighbourhood semantic displacement*

5.10.1.2 Extraction of Semantic Shifts

To construct the time-specific global corpus in the experiments, this study divided the *MEDLINE* dataset into *one-year* time unit slices (e.g., 1960, 1961, etc.). In each time-slice, an SGNS model was trained, with the dimensionality of the word embeddings set to *300* and the window size set to *5*. The parameters used for each of the proposed semantic shift measures are summarised in Table 5.1.

5.10.1.3 Construction of the Template Repository

This study considered the actual novel knowledge linkages from *historical test cases* (which were realised with time) as *templates* to construct the template repository, as discussed in Section 5.9.3. More specifically, this study utilised the actual novel knowledge linkages in the oldest historical test case (*FO-RD*) as templates to construct the template repository in the test cases: *MG-MIG*, *IGF1-ARG*, *AD-INN* and *SZ-PA2*. Since employing the actual novel knowledge linkages from *FO-RD* as templates for the *FO-RD* test case itself is biased, the template repository for *FO-RD* was constructed using the actual novel knowledge linkages from the *MIG-MG* test case.

5.10.2 Results and Discussion

Tables 5.2, 5.3, 5.4, 5.5 and 5.6 report the *Precision at k (P@k)* results of the five golden test cases, where the value of *k* is gradually increased by an interval of 10. Overall, *P@k* results indicate the robust predictive performance of the proposed LBD models across all five golden test cases. More specifically, the overall highest predictive performances of *P@k* were often exhibited across every *k* value of the golden test cases from the two proposed LBD models, *TAM* (discussed in Section 5.9) and *FTM* (discussed in Section 5.8). While *TAM* got 1.0 of *P@10* for the test cases *MG-MIG* and *IGF1-ARG*, it got 0.9, 0.9 and 0.8 of *P@10* for the test cases, *FO-RD*, *AD-INN* and *SZ-PA2*, respectively. The *P@10* errors of *TAM* are atropine (in *FO-RD*), epilepsies partial (in *AD-INN*), and growth hormone and adrenalectomy (in *SZ-PA2*).

Since *P@k* is not sensitive to the ranking order of the correct predictions (Craswell 2009), this study also focused on *Mean Average Precision at k (MAP@k)*, which quantifies the *Average Precision (AP)* in each test case. *MAP@k* is not only sensitive to the number of correct predictions, but also evaluates how well the ranking of these predicted correct instances are ordered. As such this metric is considered the *de-facto gold standard* for quantifying information retrieval systems (Beitzel et al. 2009*b*). Figure 5.32 presents the *MAP@k* results obtained across all five golden test cases (also mentioned in Table A.1). As in the case of *P@k*, the *k* value in *MAP@k* was incremented from 10 to 100 with an interval of 10.

Table 5.2: Precision@k results for FO-RD test case

Method	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**
AR (baseline)	0.8	0.7	0.733	0.675	0.68	0.667	0.657	0.638	0.611	0.6
BI (baseline)	0.0	0.0	0.1	0.225	0.38	0.45	0.486	0.525	0.556	0.58
DE (baseline)	0.3	0.25	0.4	0.425	0.4	0.383	0.4	0.438	0.422	0.4
SE (baseline)	0.2	0.2	0.2	0.25	0.24	0.283	0.271	0.3	0.322	0.32
TI (baseline)	0.2	0.25	0.333	0.425	0.46	0.483	0.5	0.5	0.522	0.53
DTM: LSTM_1	0.5	0.7	0.667	0.725	0.66	0.683	0.671	0.688	0.689	0.71
DTM: LSTM_2	0.7	0.8	0.833	0.775	0.78	0.783	0.743	0.763	0.756	0.76
DTM: LSTM_3	0.5	0.5	0.567	0.55	0.6	0.65	0.686	0.7	0.733	0.71
DTM: CNN	0.5	0.45	0.433	0.5	0.48	0.533	0.571	0.6	0.611	0.63
DTM: CNN_LSTM	0.0	0.15	0.233	0.275	0.32	0.417	0.443	0.475	0.511	0.55
DTM: LSTM_CNN	0.0	0.1	0.2	0.3	0.36	0.417	0.443	0.488	0.544	0.57
FTM	**0.9**	0.8	**0.867**	**0.85**	**0.82**	**0.817**	**0.8**	**0.775**	**0.789**	**0.77**
TAM	**0.9**	**0.85**	**0.867**	**0.85**	0.78	0.783	0.757	0.75	0.756	**0.77**

Table 5.3: Precision@k results for MG-MIG test case

Method	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**
AR (baseline)	0.6	0.55	0.567	0.575	0.6	0.567	0.529	0.575	0.567	0.57
BI (baseline)	0.0	0.0	0.0	0.0	0.06	0.117	0.143	0.15	0.2	0.26
DE (baseline)	0.6	0.5	0.467	0.425	0.44	0.467	0.5	0.488	0.456	0.44
SE (baseline)	0.5	0.5	0.633	0.6	0.56	0.55	0.529	0.525	0.544	0.55
TI (baseline)	0.1	0.15	0.2	0.3	0.34	0.4	0.386	0.425	0.456	0.47
DTM: LSTM_1	0.4	0.65	0.633	0.7	0.74	0.767	0.8	0.825	0.822	0.81
DTM: LSTM_2	0.5	0.7	0.667	0.675	0.68	0.7	0.714	0.725	0.733	0.73
DTM: LSTM_3	0.4	0.65	0.7	0.7	0.74	0.75	0.757	0.75	0.756	0.75
DTM: CNN	0.7	0.6	0.733	0.775	0.8	0.8	0.829	0.825	0.822	**0.82**
DTM: CNN_LSTM	**1.0**	**0.9**	0.8	0.725	0.7	0.733	0.757	0.763	0.789	0.79
DTM: LSTM_CNN	0.9	0.8	0.833	0.8	0.82	0.85	0.843	0.85	0.811	0.81
FTM	0.9	**0.9**	**0.867**	**0.9**	**0.9**	0.85	0.857	0.813	0.8	0.8
TAM	**1.0**	**0.9**	0.833	0.85	0.84	**0.867**	**0.871**	**0.863**	**0.833**	**0.82**

Table 5.4: Precision@k results for IGF1-ARG test case

Method	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**
AR (baseline)	**1.0**	0.85	0.8	0.775	0.78	0.783	0.786	**0.813**	**0.811**	**0.8**
BI (baseline)	0.0	0.0	0.033	0.05	0.04	0.05	0.057	0.075	0.089	0.12
DE (baseline)	0.5	0.5	0.6	0.575	0.62	0.583	0.6	0.613	0.611	0.61
SE (baseline)	0.4	0.55	0.6	0.6	0.64	0.633	0.6	0.6	0.622	0.61
TI (baseline)	0.1	0.2	0.267	0.3	0.4	0.383	0.4	0.388	0.411	0.4
DTM: LSTM_1	0.4	0.6	0.6	0.65	0.68	0.683	0.7	0.713	0.722	0.72
DTM: LSTM_2	0.6	0.55	0.567	0.65	0.68	0.7	0.714	0.713	0.688	0.7
DTM: LSTM_3	0.5	0.65	0.7	0.625	0.68	0.7	0.7	0.688	0.7	0.7
DTM: CNN	0.8	0.65	0.767	0.825	0.8	0.767	0.743	0.75	0.767	0.75
DTM: CNN_LSTM	0.6	0.6	0.633	0.65	0.68	0.717	0.743	0.763	0.778	0.79
DTM: LSTM_CNN	0.5	0.65	0.667	0.725	0.76	0.783	0.771	0.763	0.778	0.77
FTM	0.7	0.8	0.8	0.75	0.74	0.683	0.671	0.663	0.644	0.66
TAM	**1.0**	**0.9**	**0.9**	**0.925**	**0.92**	**0.867**	**0.829**	**0.813**	0.778	0.75

Table 5.5: Precision@k results for AD-INN test case

Method	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**
AR (baseline)	0.8	0.85	0.8	0.775	0.76	0.767	0.7	0.7	0.689	0.67
BI (baseline)	0.0	0.0	0.0	0.025	0.02	0.033	0.1	0.1	0.167	0.17
DE (baseline)	0.6	0.45	0.5	0.525	0.48	0.467	0.486	0.488	0.5	0.48
SE (baseline)	0.2	0.6	0.6	0.575	0.62	0.617	0.629	0.65	0.656	0.67
TI (baseline)	0.0	0.1	0.1	0.125	0.22	0.283	0.257	0.275	0.3	0.31
DTM: LSTM_1	0.8	0.85	0.9	0.825	0.82	0.833	0.814	0.8	0.822	0.84
DTM: LSTM_2	0.7	0.75	0.8	0.825	0.86	0.85	0.871	0.85	0.856	0.86
DTM: LSTM_3	0.8	0.85	0.833	0.8	0.8	0.767	0.771	0.788	0.778	0.78
DTM: CNN	0.7	0.85	0.833	0.85	0.86	0.867	0.871	0.875	0.889	0.89
DTM: CNN_LSTM	**1.0**	**0.9**	**0.933**	**0.925**	**0.92**	0.867	0.871	0.863	0.867	0.88
DTM: LSTM_CNN	0.8	0.8	0.867	0.85	0.82	0.833	0.814	0.825	0.822	0.82
FTM	0.9	**0.9**	0.9	0.9	**0.92**	**0.933**	**0.943**	**0.95**	**0.944**	**0.93**
TAM	0.9	**0.9**	0.867	0.875	0.9	0.883	0.871	0.875	0.867	0.88

Table 5.6: Precision@k results for SZ-PA2 test case

Method	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**
AR (baseline)	0.7	0.6	0.633	0.6	0.58	0.6	0.629	0.625	0.622	0.61
BI (baseline)	0.0	0.0	0.0	0.0	0.02	0.067	0.086	0.1	0.111	0.15
DE (baseline)	0.2	0.4	0.4	0.45	0.48	0.467	0.429	0.425	0.433	0.43
SE (baseline)	0.0	0.15	0.267	0.325	0.34	0.367	0.386	0.4	0.4	0.42
TI (baseline)	0.1	0.15	0.133	0.125	0.14	0.167	0.171	0.213	0.233	0.25
DTM: LSTM_1	0.5	0.55	0.633	0.65	0.66	0.667	0.671	0.663	0.678	0.68
DTM: LSTM_2	0.6	0.45	0.367	0.4	0.44	0.45	0.5	0.538	0.567	0.59
DTM: LSTM_3	0.7	0.75	0.7	0.725	0.76	0.767	0.757	0.75	0.733	0.72
DTM: CNN	0.7	0.6	0.6	0.6	0.66	0.617	0.614	0.575	0.589	0.6
DTM: CNN_LSTM	0.7	0.75	0.733	0.75	0.72	0.683	0.714	0.7	0.7	0.71
DTM: LSTM_CNN	0.1	0.25	0.233	0.3	0.34	0.317	0.329	0.4	0.411	0.44
FTM	**0.9**	**0.95**	**0.9**	**0.85**	**0.86**	0.833	**0.857**	0.825	0.811	0.82
TAM	0.8	0.8	0.8	0.8	0.84	**0.85**	0.843	**0.838**	**0.833**	**0.83**

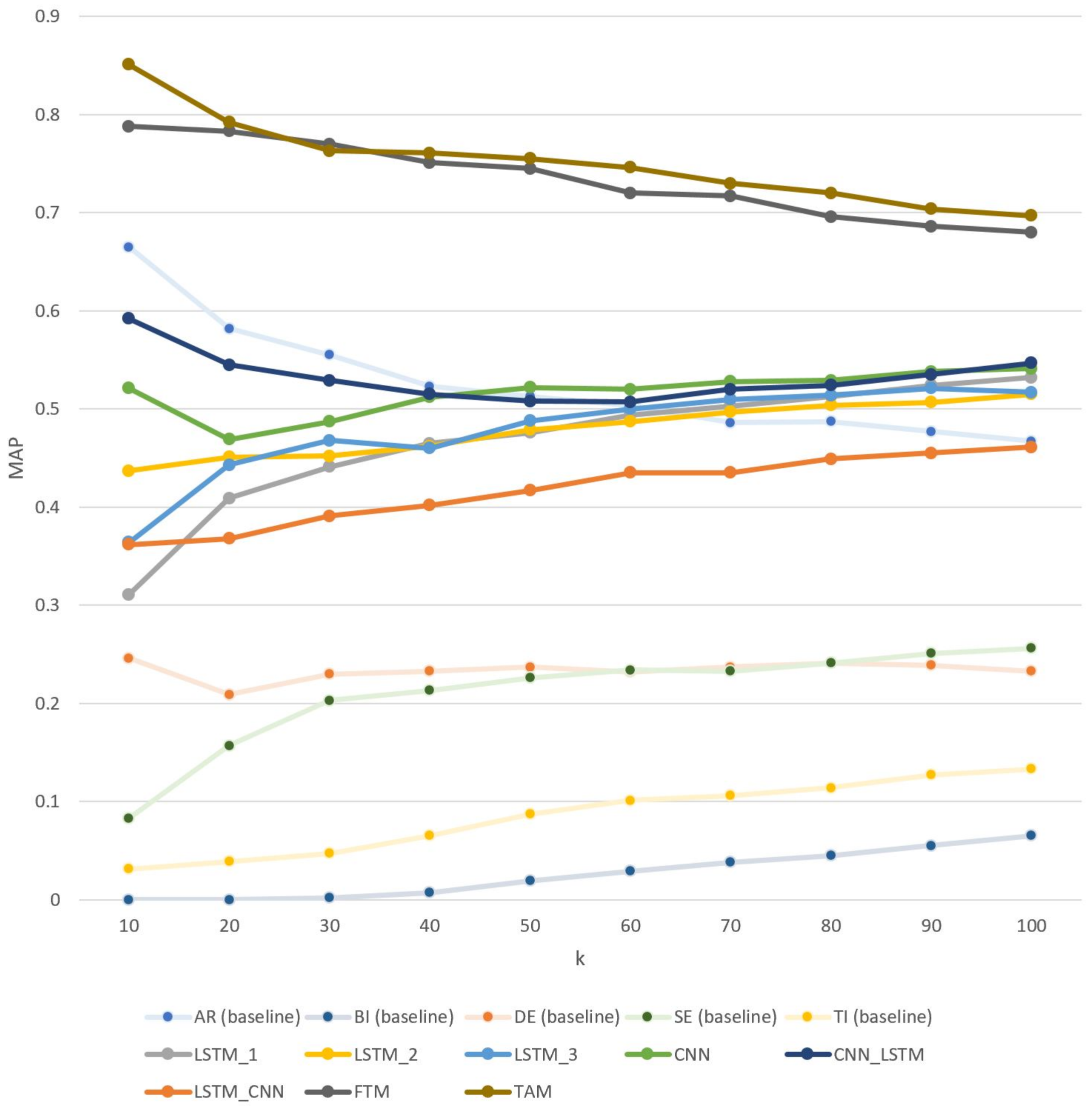


FIGURE 5.32: MAP@k results for the five golden test cases: FO-RD, MG-MIG, IGF1-ARG, AD-INN and SZ-PA2

When observing *MAP@k* results for the five golden test cases (Figure 5.32), it is evident that the proposed LBD model *TAM* displayed the highest predictive performance. The proposed LBD model *FTM* displayed the second highest predictive performance. Even though *TAM* exihibited a 6.3% performance increase over *FTM* at *MAP@10*, the remaining performance increases of *TAM* over *FTM* were in the range of 1% to 3%.

Ordered from highest to lowest, the predictive performances of the baseline models were as follows: *AR*, *DE*, *SE*, *TI* and *BI*. The performance increases of the two highest-performing predictive models: *TAM* and *FTM* over the baseline models are illustrated in Figures 5.33 and 5.34, respectively. It is evident from Figure 5.33 that *TAM* exhibited significant performance increases over the baselines at every k value. The most competitive baseline model was AR, yet TAM demonstrated consistent performance increases of nearly 20% across every k value. More specifically, the average performance increases of TAM over the baselines were *0.226* (with AR), *0.518* (with DE), *0.542* (with SE), *0.667* (with TI) and *0.726* (with BI). FTM also demonstrated slightly similar performance increases over the baselines across the k values as depicted in Figure 5.34. More precisely, the average MAP performance increases of FTM over the baselines were *0.208* (with AR), *0.5* (with DE), *0.524* (with SE), *0.649* (with TI) and *0.708* (with BI).

With respect to DTM model variants, it is evident from Figure 5.32 that they performed better than the baseline models: DE, SE, TI and BI. Nevertheless, the baseline model AR displayed a higher performance than the DTM model variants until *MAP@60*. The ensuing MAP performances of DTM model variants (except *DTM: LSTM_CNN* model) after the k value reached 60 demonstrated better performance compared to the AR baseline. From *MAP@k* results, *DTM: CNN_LSTM* displayed the highest predictive performance. The second-highest performance of the DTM model variants was observed through the use of *DTM: CNN*. One of the key differences between the DTM model variants and the remaining two proposed LBD models (TAM and FTM) is that their MAP performances increased as values of k increased. Nevertheless, these DTM model

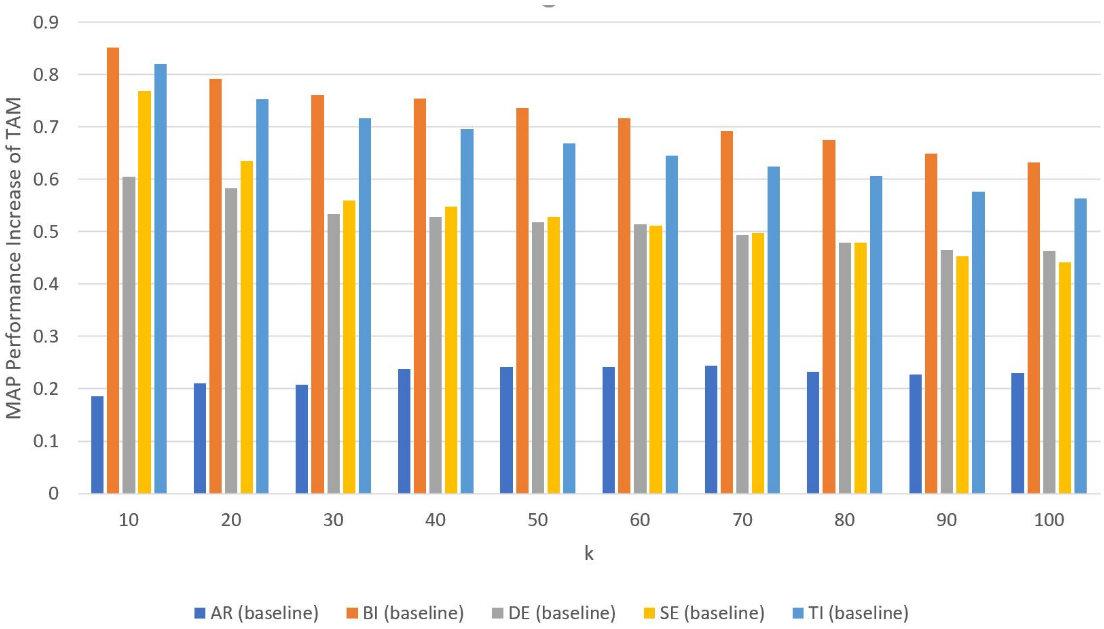


Figure 5.33: The performance increase of TAM over the baseline models

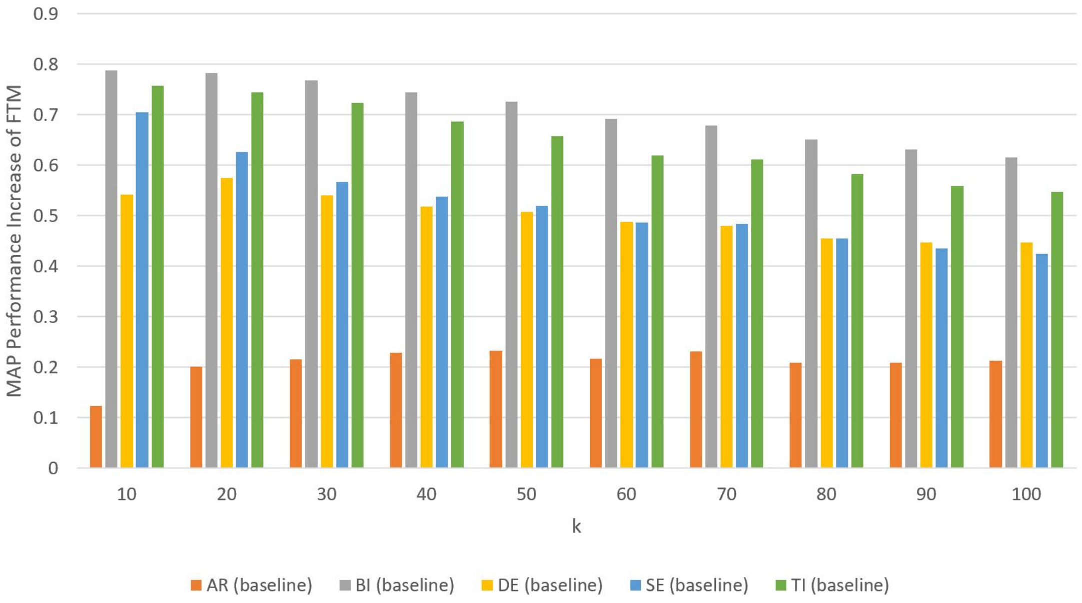


Figure 5.34: The performance increase of FTM over the baseline models

variants did not surpass the predictive performance of *TAM* and *FTM*, which demonstrates that knowledge discovery often favours measures that are tailored to the LBD problem (i.e., *handcrafted features*) over features extracted using deep learning models. Otherwise stated, the results demonstrate that LBD performance is more sensitive to temporal patterns extracted using empirical observations while also focusing on LBD's problem setting and objective.

The following conclusions were obtained through an analysis of the predictive performances of the baseline models. The *AR* baseline consistently outperformed the other baselines in terms of prediction. This could be due to the *AR* baseline's *usage of multiple characteristics*, *focus on both global and local features* and *LBD-tailored heuristics*. Despite these strengths, this baseline relies on semantic inferences performed using domain-specific knowledge resources (i.e., *MeSH* and *UMLS*), which may inhibit the *reusability* and *portability* of this LBD model in *other problem settings* and *other portable scientific domains*.

The second-best performance was exhibited by the *DE* (*Dynamic Embedding*) baseline, potentially due to its focus on integrating *temporal characteristics* in *semantic spaces* into the knowledge discovery. Nevertheless, the baseline's use of *limited characteristics* to define new knowledge, as well as its *shallow temporal analysis component* may have reduced its performance over the proposed LBD models in this study. The next highest performance was obtained through the use of *SE* (*Static Embedding*) baseline. The prediction performances of *DE* and *SE* indicate the need for semantic inferences in the knowledge discovery process to detect novel knowledge linkages more precisely than purely statistical-based baseline models, such as *TI* and *BI*. The fact that *DE* outperformed *SE* in terms of MAP until k is equal to 60 showcases the limited nature of *static semantic cues* in comparison to the *shallow temporal semantic cues* used in *DE*. With the integration of *large-scale temporal semantic cues* through the circumstantial temporal analysis component (as in this study's proposed LBD models), the predictive performance was significantly improved in comparison to the *SE* and *DE* baseline models, which use *static* and *shallow temporal semantic cues*, respectively.

The two statistical baselines (*TI* and *BI*) displayed the lowest predictive performance. There are three possible reasons for this. Firstly the complexity of the problem that LBD attempts to solve may require *detailed semantic inferences* that these conventional statistical-based techniques may not capture sufficiently. Secondly, overall, the results indicate that knowledge discovery in the LBD process favours measures tailored to the problem, rather than to the direct use of conventional statistical measures. Third, the use of a single characteristic may not be sufficient to capture novel knowledge linkages precisely.

Despite the promising *MAP@k* results observed in the previous setting that showcase *overall predictive performances*, this study requires to further verify how *consistent the predictive performances* were across all five golden test cases. To quantify the *consistency* of the prediction, *Geometric Mean Average Precision at k (GMAP@k)* was utilised, as discussed in Chapter 3. The predictive performances of the LBD models in terms of *GMAP* are shown in Figure 5.35. It is evident that the two proposed LBD models, *TAM* and *FTM* also demonstrated significant performance increases in terms of *GMAP*. This verifies that these two LBD models not only demonstrated the *overall* highest predictive performances, but also *consistently* highest predictive performances across the test cases. Considering the *DTM* variants of the proposed LBD models, it is plain that *DTM: LSTM_1*, *DTM: LSTM_2*, *DTM: LSTM_3*, and *DTM: CNN* displayed similar *GMAP* performances in comparison to their corresponding *MAP* performances. It is interesting to observe that even *DTM: CNN_LSTM* exhibited overall high performance (i.e., in *MAP* results) due to its superior performances in test cases such as *MG-MIG* and *AD-INN*, its *GMAP* performance was significantly lower, since it did not perform

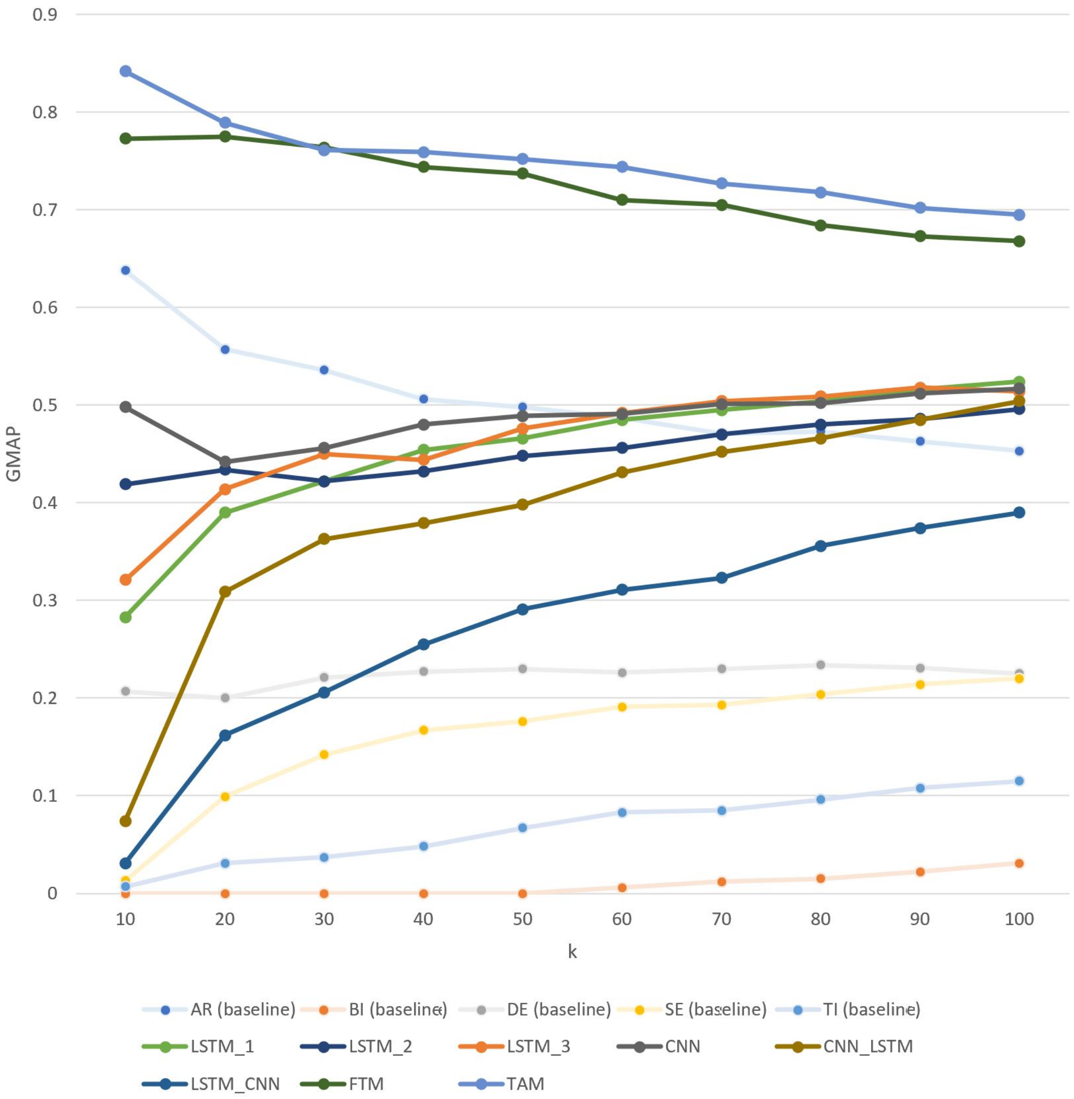


Figure 5.35: GMAP@k results for the five golden test cases: FO-RD, MG-MIG, IGF1-ARG, AD-INN and SZ-PA2

well for test cases such as *FO-RD*. Otherwise stated, these findings indicate that the use of the *DTM: CNN_LSTM* model may not necessarily guarantee good performance for every user query, since its predictive performances are unstable, especially for the initial k values. Among the proposed LBD models, *DTM: LSTM_CNN* displayed the lowest *GMAP* predictive performance, as was the case for the previous *MAP* setup.

This study also aimed to detect potential contributions of semantic shift types in isolation to predictive performance, with the aim of identifying the most effective semantic shift types in the LBD knowledge discovery process. Figures 5.36 and 5.37 represent the performance differences exhibited through each of the semantic shift types (i.e., *Individual Semantic Shifts (ISS)*, *Pairwise Semantic Shifts (PSS)* and *Neighbourhood Semantic*

Shifts (NSS)). Overall, it is evident that the complementary integration of these three types of semantic shifts (i.e., *ISS+PSS+NSS*) is more promising in terms of discovering novel knowledge linkages than using each semantic shift type alone. Otherwise stated, semantic shift types in isolation underperformed relative to the integrated version. Nevertheless, each of the semantic shift types alone performed better than the baseline models. This indicates the significant influence of diachronic semantic inferences on the high-precision discovery of novel knowledge linkages in the LBD process.

Comparing the predictive performance of the semantic shift types alone, the most consistent predictive performance was displayed by *ISS* in both of the two proposed LBD models, *TAM* and *FTM* (Figures 5.36 and 5.37, respectively). The predictive performances of the *PSS* and *NSS* semantic shift types contrasted with those of *TAM* and *FTM*. More specifically, while *NSS* exhibited high predictive performance in *TAM* in the initial phase (i.e., until MAP@50), its predictive performance dropped in the latter phase, while *PSS* demonstrated consistent performance throughout the k values. In the model *FTM*, *PSS* showed high predictive performance at MAP@10. Nonetheless, there was a performance drop after the k value 10, while *NSS* indicated consistent predictive performance across every k value. Furthermore, this study observed that the predictive performance variance of each semantic shift type alone (i.e., ISS, PSS and NSS individually) was slightly low in *TAM* in comparison to *FTM*, further verifying the consistent predictive performance of *TAM* relative to *FTM*.

This study aimed to further assess the robust predictive performance of the proposed LBD models in an extended experimental setup, with the goal of indicating how the models perform in the *long run* (in contrast to the previous setup). More specifically, this experimental setup resembles a situation where a user is keen to explore novel knowledge linkages of more than 100, as described in the previous setup. For this purpose, this study evaluated the predictive performances of the proposed LBD models and baselines up to the k value 500. Similarly to the previous experimental setup, the k value was incrementally increased from 10 to 500, with an interval of 10. The MAP results are depicted in Figure 5.38 that indicate the *overall performances* of the LBD models in the long run. As in the previous experimental setup, the two proposed LBD models, *TAM* and *FTM* consistently outperformed in this experimental setup, demonstrating their robust predictive performance. It is also evident that in this extended experimental setup, the *DTM* model variants are becoming to be better than the baseline *AR*, indicating

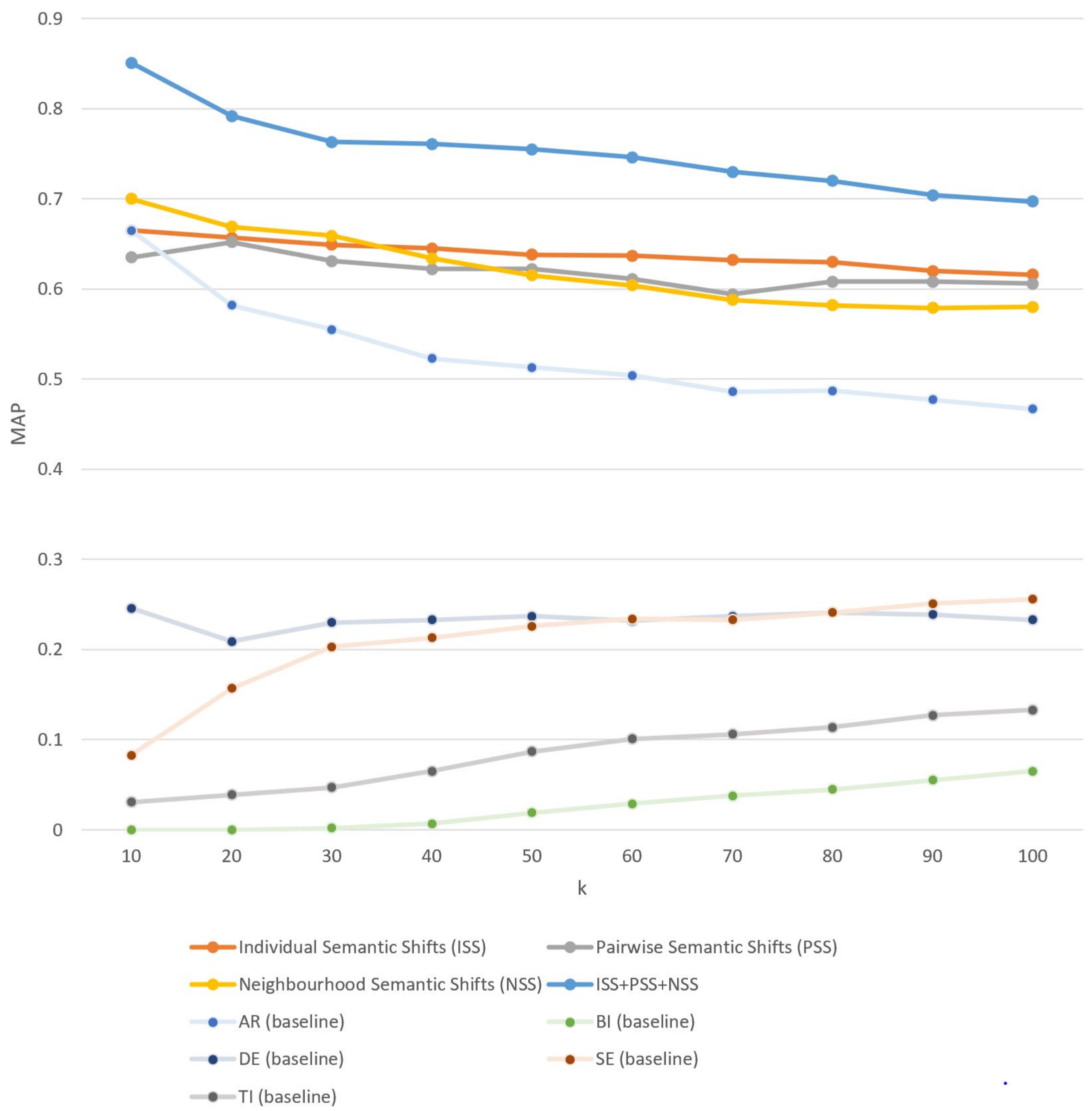


FIGURE 5.36: Contribution of each semantic shift type towards the predictive performance of TAM

the potential contributions that deep neural network techniques may have in the LBD context. Furthermore, this study observed that *TAM* performed better than *FTM* until k was equal to 200. For the ensuing k values, FTM demonstrated a slight performance increase over *TAM*. More specifically, the average performance increase of *FTM* over *TAM* after $k = 200$ was 0.5%. The most visible performance improvements of *TAM* over *FTM* were within the range of $k = 240$ to $k = 300$, with an increase of nearly 1%. It is evident that the performance of *BI* was increasing over the two semantic baselines *DE* and *SE* after the $k = 360$, whereas *TI* displayed nearly similar performances after $k =$ 470. Overall, the proposed LBD models demonstrated significant performance increases over the baseline models.

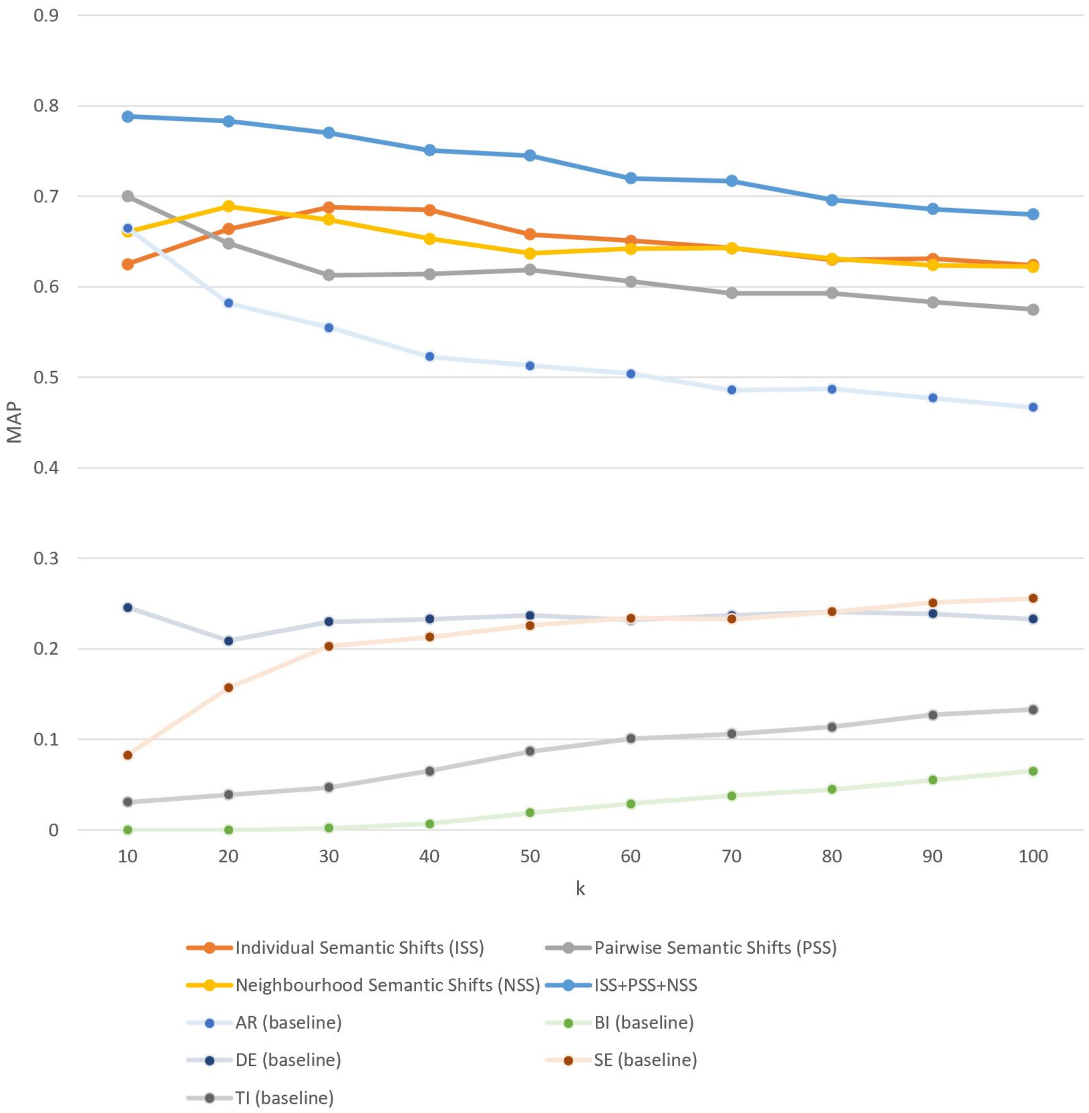


FIGURE 5.37: Contribution of each semantic shift type towards the predictive performance of FTM

As with the previous *MAP* setup, this extended study also analysed the *consistency* of the predictive performances in the *long run* through the use of *GMAP*, as denoted in Figure 5.39. As in the case of all the previous experimental setups, *TAM* and *FTM* demonstrated the highest predictive performances, indicating that they had the *highest* as well as the most *consistent* predictive performances in the *long run*. Moreover, it is also evident that the prediction consistency of the two LBD models, *DTM: CNN_LSTM* and *DTM: LSTM_CNN* that were penalised in the *short run* (due to their low performances in test cases such as *FO-RD*) increased in the *long run*. More specifically, these two DTM variants surpassed the performance of the most competitive baseline, which is *AR* at *k* values 90 and 180, respectively. Moreover, these two DTM variants also

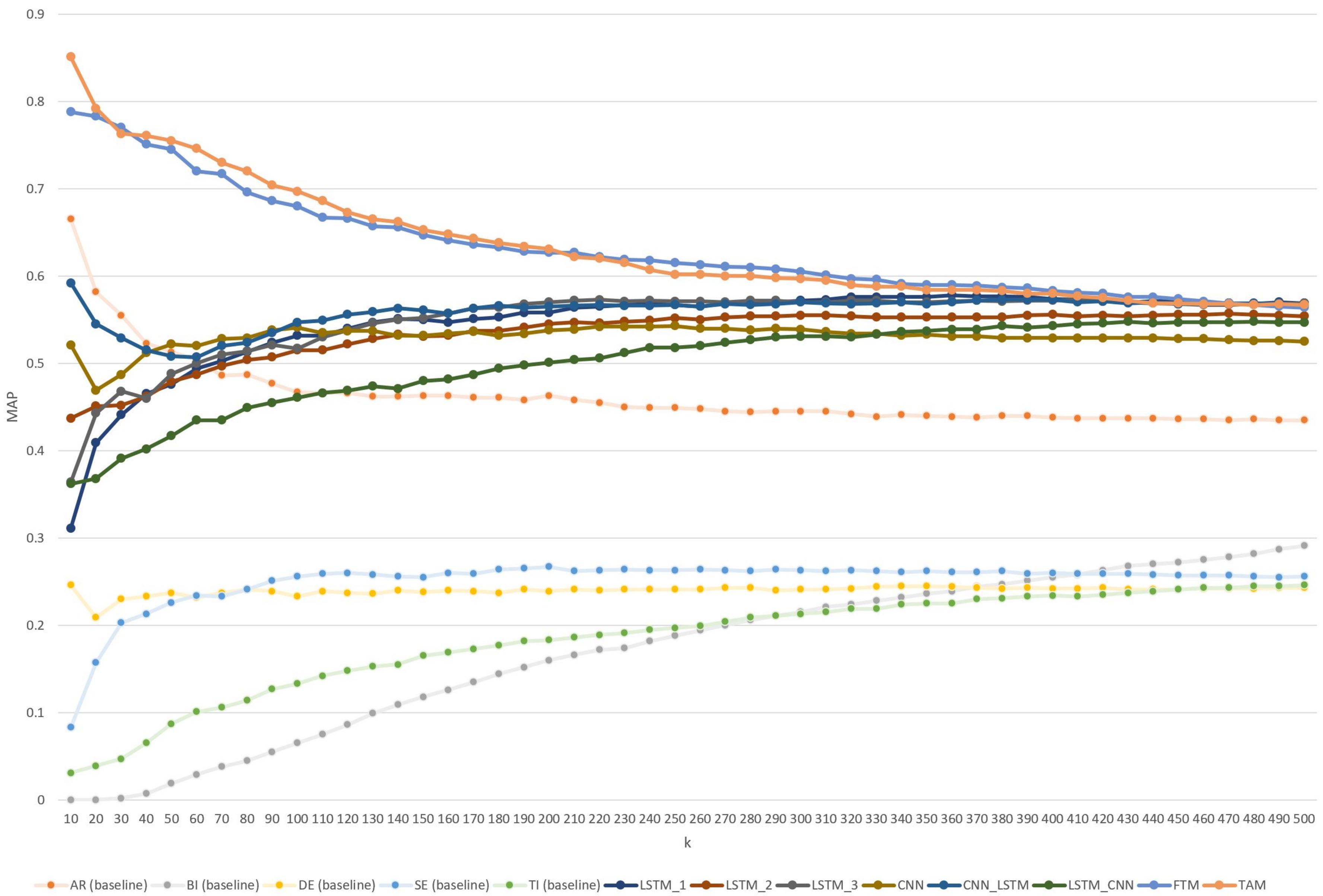


FIGURE 5.38: MAP@k results for the five golden test cases: FO-RD, MG-MIG, IGF1-ARG, AD-INN and SZ-PA2 in the long run

aligned with the remaining DTM model variants at k values 100 and 330. Overall, the proposed LBD models outperformed the baseline models considering their *consistency* of the predictions in the long run.

Next, the same extended experimental setup was conducted with the use of diachronic semantic inferences alone (i.e., *ISS*, *PSS* and *NSS* individually). The purpose of this experiment was to further analyse whether these semantic shift types alone (i.e., most simplified versions of the proposed LBD models) outperformed the baseline models in the *long run*. Figures 5.40 and 5.41 denote the predictive performances of each semantic shift type in the two proposed LBD models, *TAM* and *FTM* that demonstrated the highest predictive performances in all experimental setups. It is evident from the results of this experiment that even the most simplified versions of the proposed LBD models (i.e., each semantic shift type alone) outperformed the baseline models not only in the *short run* (as demonstrated in the previous experimental setup), but also in the *long run*. These results demonstrate the robust positive influence of the complementary integration of *vector semantics* with *temporal dimension* in the LBD knowledge discovery workflow.

Despite the interesting results observed in terms of *P@k*, *MAP@k* and *GMAP@k* in the *short run* and *long run*, to further verify the robust predictive performance of the two best-performing models (*TAM* and *FTM*) in the experimental setups described above, this study evaluated how well these proposed LBD models separated *negative instances* from *positive instances* using standard classification metrics (as discussed in Chapter 3). This setup reflects the way in which the LBD models perform in the *longer run*. To summarise, negative instances in the LBD workflow denote uninteresting or meaningless concepts, indicating the *filtering phase* of knowledge discovery. In contrast, positive instances in the LBD workflow denote potential novel knowledge linkages, which can be viewed as the *ranking component* in knowledge discovery (discussed in Chapter 3).

Table 5.7: Classification results for FO-RD test case

Method	**ROC Area**	**Precision**	**Recall**	**F-Measure**
AR (baseline)	0.69	0.69	0.697	0.692
DE (baseline)	0.554	0.614	0.672	0.621
FTM	0.734	0.723	0.74	0.703
TAM	**0.749**	**0.737**	**0.751**	**0.727**

Tables 5.7, 5.8, 5.9, 5.10 and 5.11 summarise the classification performance of the two best-performing models in comparison to the two multi-characteristic baselines, *AR* and

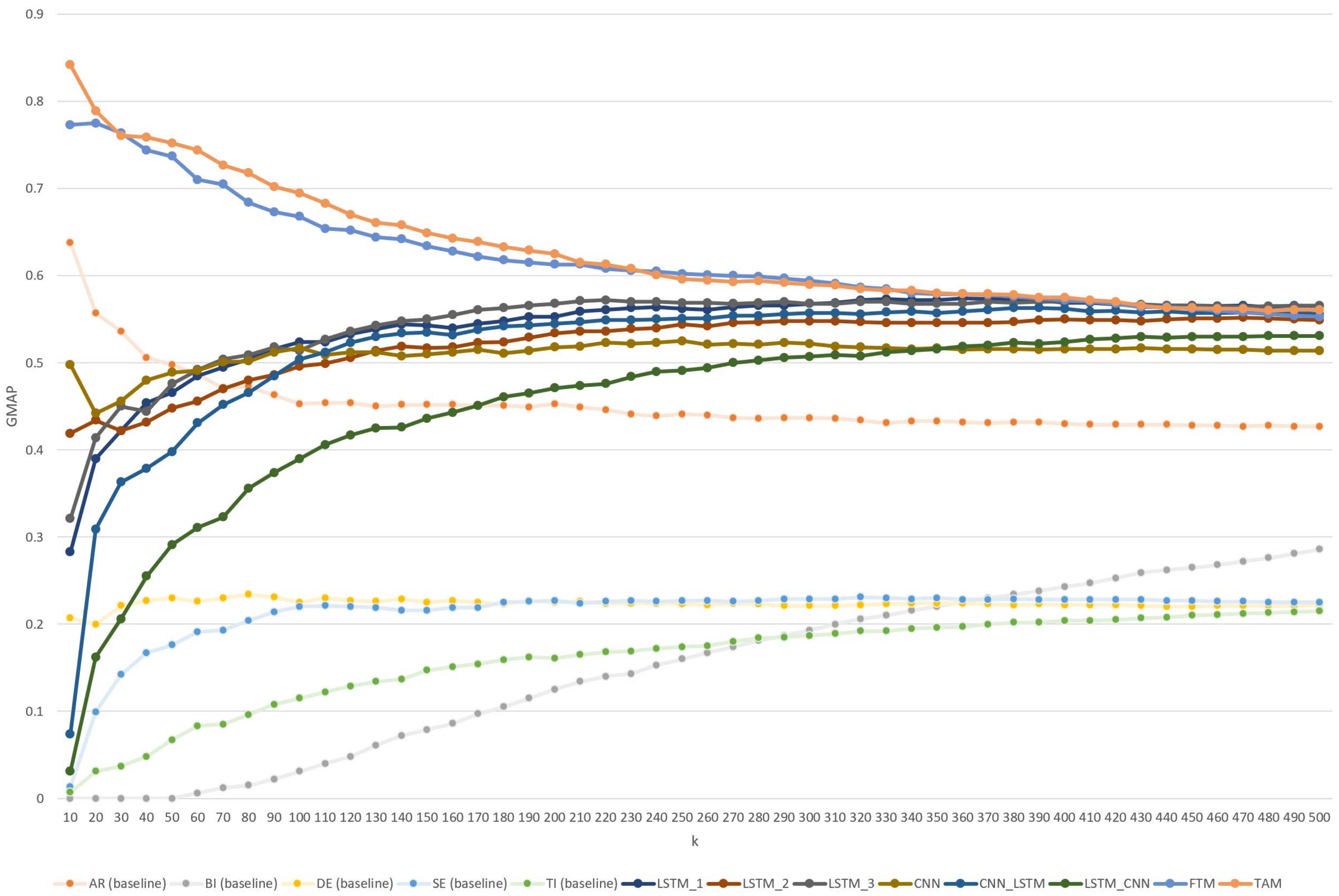


Figure 5.39: GMAP@k results for the five golden test cases: FO-RD, MG-MIG, IGF1-ARG, AD-INN and SZ-PA2 in the long run

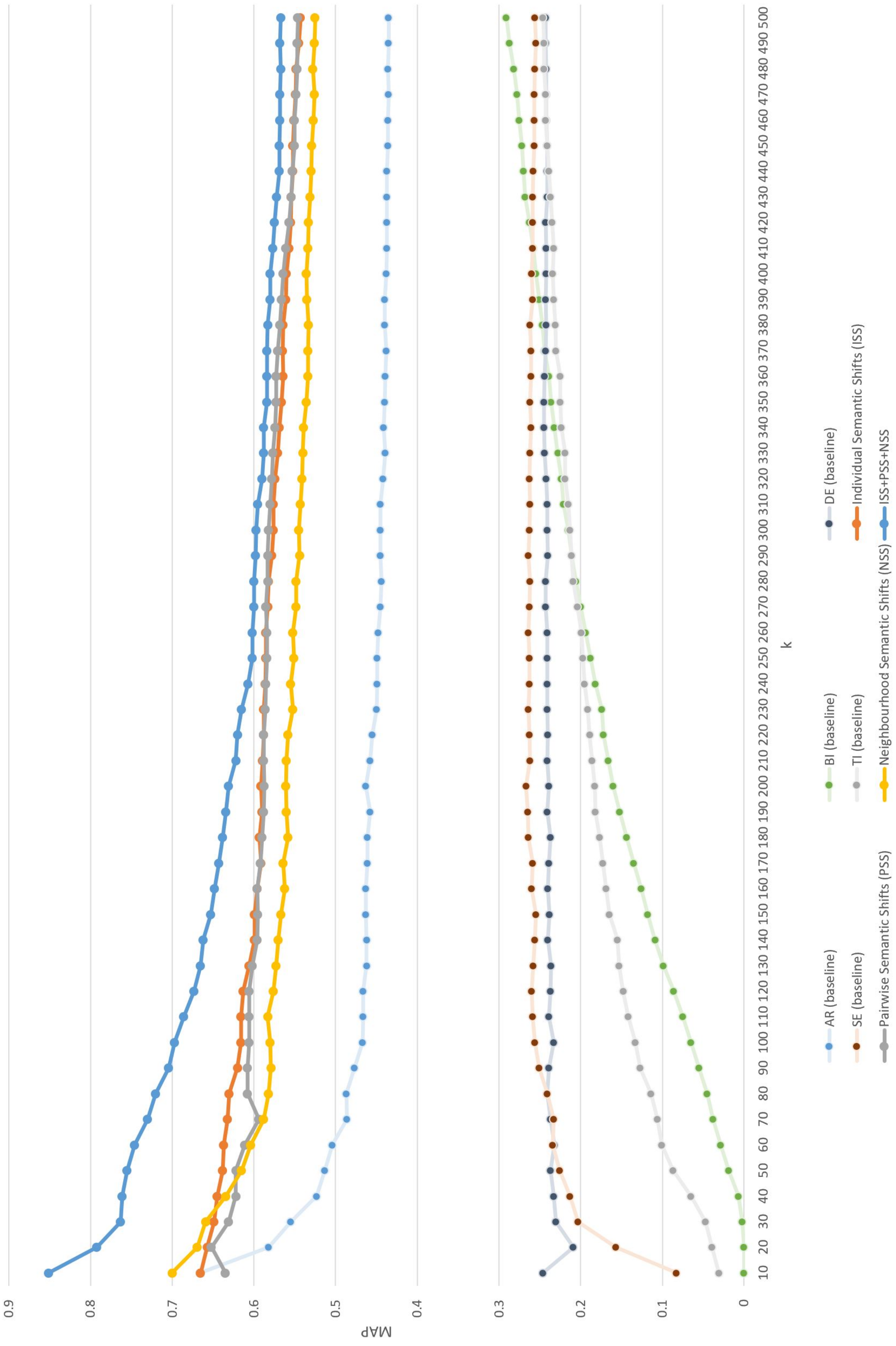


FIGURE 5.40: Contribution of each semantic shift type towards the predictive performance of TAM in the long run

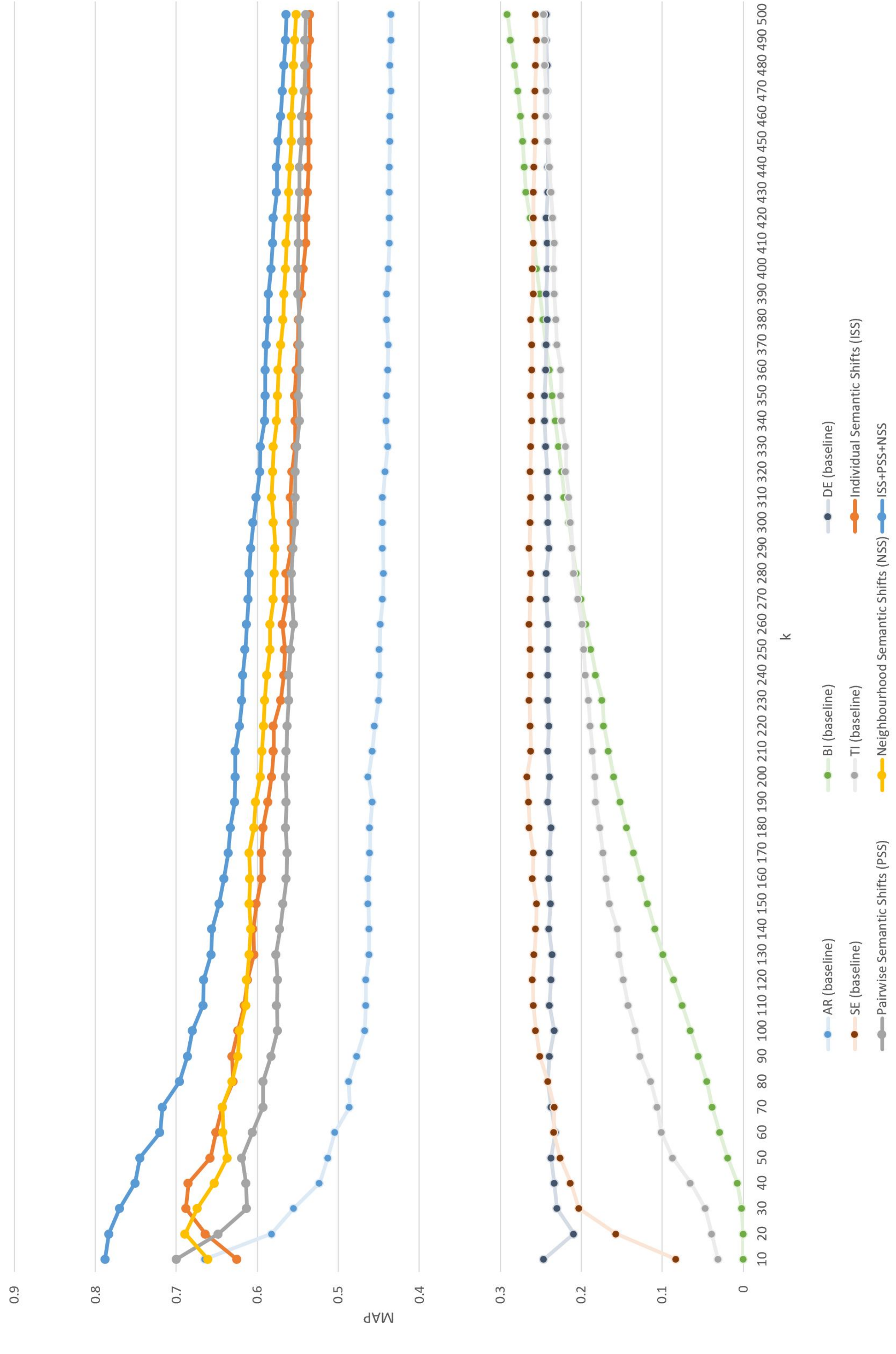


Figure 5.41: Contribution of each semantic shift type towards the predictive performance of FTM in the long run

Table 5.8: Classification results for MG-MIG test case

Method	**ROC Area**	**Precision**	**Recall**	**F-Measure**
AR (baseline)	0.658	0.676	0.684	0.679
DE (baseline)	0.603	0.618	0.661	0.623
FTM	0.75	0.724	0.74	0.708
TAM	**0.755**	**0.736**	**0.747**	**0.724**

Table 5.9: Classification results for IGF1-ARG test case

Method	**ROC Area**	**Precision**	**Recall**	**F-Measure**
AR (baseline)	0.623	0.607	0.604	0.605
DE (baseline)	0.564	0.564	0.567	0.565
FTM	0.7	0.663	0.664	0.663
TAM	**0.715**	**0.681**	**0.682**	**0.681**

Table 5.10: Classification results for AD-INN test case

Method	**ROC Area**	**Precision**	**Recall**	**F-Measure**
AR (baseline)	0.651	0.625	0.626	0.625
DE (baseline)	0.59	0.565	0.577	0.563
FTM	0.743	0.686	0.687	0.677
TAM	**0.744**	**0.692**	**0.693**	**0.684**

Table 5.11: Classification results for SZ-PA2 test case

Method	**ROC Area**	**Precision**	**Recall**	**F-Measure**
AR (baseline)	0.686	0.706	0.713	0.709
DE (baseline)	0.571	0.63	0.693	0.641
FTM	**0.765**	0.744	0.762	0.73
TAM	0.755	**0.747**	**0.764**	**0.737**

DE. Overall, it is evident that the two models, *TAM* and *FTM* consistently outperformed the two competitive baselines in all five golden test cases in terms of *area under ROC curve* (i.e., the *AUC*), and weighted average versions of *precision*, *recall* and *F-measure*. As in the case of the previous setup, the predictive performance of these two proposed LBD models are in the order of *TAM* and *FTM*. Moreover, the performance of the two baseline models indicated that *AR* performed better than *DE* in every test case. The increased performance behaviour of *AR* in comparison to the *DE* baseline was consistent with the previous setup.

5.10.3 Strengths of the Proposed LBD Models

This study observes the following eight strengths of the proposed LBD models in comparison to the previous LBD research: *integration of global semantics*, *intermingling* *vector semantics with the temporal dimension*, *disentangling semantic shifts from different perspectives*, *nuanced temporal analysis*, *use of multiple characteristics*, *integration of machine learning and deep learning techniques*, *domain independency*, and *potential direct and indirect uses of diachronic semantic inferences.* These strengths may explain why the proposed LBD models consistently outperformed the baseline models throughout every experimental setup.

Most of the prior LBD research mainly uses the *query-specific local corpora* to facilitate knowledge discovery; thus, the potential cues identified in this process denote signals extracted from local scale topic interactions. Since query-specific local corpora lack the global-level topic interactions due to its restrictive nature, harnessing these signals through the *integration of global semantics* may be particularly important for tasks that require complex semantic deductions, as in the case of LBD. Therefore, the proposed LBD models incorporated a wide scope of topic interactions through the integration of time-specific global corpus. This ensured that the implicit semantic relationships of scientific topics that are invisible to the local corpus were also captured when forming diachronic semantic inferences.

The following two observations (discussed in Chapter 2), which are limited research contributions that incorporate modern word embedding techniques (observed in *timeline analysis*) and overlooking the importance of temporal dimension (observed in *categorisations* of computational techniques) prompted this thesis to amalgamate *vector semantics* with the *temporal dimension.* Unlike previous LBD studies, which are based on a static snapshot of the literature, the complementary integration of these two notions enables the opportunity to capture the dynamic nature of scientific knowledge that is invisible to mere static literature analysis. Therefore, this study has been influenced by an emerging research field, *diachronic word embeddings*, which has emerged due to recent advancements in word embedding techniques. The experimental results indicate the importance of co-modelling the complementary strengths of vector semantics and temporal dynamicity, due to its robust predictive performance in terms of LBD knowledge discovery, compared to models that use static cues. More specifically, scrutinising the temporal behaviour of scientific topics in the semantic spaces to discover latent novel knowledge linkages demonstrated consistently highest predictive performance in every experimental setup across all golden test cases.

The constructed diachronic vector spaces enable a rich platform to scrutinise deep semantic inferences. Yet it is important to define meaningful diachronic semantic inferences tailored to focus and objective of the LBD research. To this end, this thesis disentangles diachronic semantic inferences from three main perspectives (i.e., *individual semantic shifts*, *pairwise semantic shifts* and *neighbourhood semantic shifts*). Individual semantic shifts were quantified at both the global and local context, namely *individual global shifts* and *individual local shifts*. Pairwise semantic shifts included two measures relative to user-defined input topics, namely *pairwise semantic displacement* and *pairwise distance proximity*. Neighbourhood semantic shifts extended pairwise semantic shifts by incorporating the recent core meaning into the semantic inferences through the use of *neighbourhood semantic displacement* and *neighbourhood distance proximity*. Unravelling such meaningful semantic shifts at multiple levels enriches the semantic deductions and ultimately strengthens the prediction capabilities of the LBD models.

Even though there are a few recent LBD studies that have attempted to mitigate the issue of static literature analysis of the previous LBD research through the inclusion of temporal information, the underlying temporal analysis component of these studies is fairly shallow (Jha et al. 2018, Jha, Xun, Wang & Zhang 2019, Xun et al. 2017). For instance, to identify the temporal trend of the scientific topics, the study of Xun et al. (2017) has only considered the first and last value of the time series while neglecting potential temporal signals that could reside in the time series as a whole. Nevertheless, circumstantial analysis of patterns in time series (Shoemark et al. 2019) could be beneficial to further comprehend the temporal behaviours of the scientific topics that would contribute to enhancing predictions in the knowledge discovery component. To circumvent this hindrance, this study performs a *nuanced temporal analysis* by using *semantically infused temporal trajectories* as the main analysis unit in the proposed LBD models. The robust predictive performances of the proposed LBD models showcase the necessity of integrating such fine-grained temporal analysis components into the LBD workflow. More precisely, such subtle temporal analysis provides the opportunity to identify strong temporal cues which are otherwise concealed.

Most of the previous LBD studies mainly rely on one (or at most two to three) characteristic(s) to discover potential new knowledge linkages in the literature. This may be limited due to two main reasons. Firstly, it may inhibit a model's capability to detect novel knowledge linkages precisely, due to the complexities involved in natural language usage that result in intricate structures of scientific literature. Secondly, it is observed in the theoretical LBD literature that new knowledge can be in multiple forms; thus, the use of limited characteristics may hinder the model's capability of identifying novel knowledge forms in a wide perspective. Therefore, this study attempts to consolidate multiple characteristics in the knowledge discovery process to circumvent the above two limitations. More specifically, this study used two different levels to integrate multiple characteristics that could have a potentially positive impact on the knowledge discovery process. The first level was at the formation of diachronic semantic inferences, where this study disentangled semantic shifts at three main perspectives, resulting in six semantically infused temporal trajectories. The second level involved the temporal analysis component, in which numerous temporal signals were extracted to elicit potential novel knowledge linkages. These temporal signals were either handcrafted as used in traditional machine learning setting or extracted using deep learning models. The experimental results demonstrate the importance of incorporating such multiple meaningful characteristics in the knowledge discovery process to enhance its predictive performance.

Machine Learning (ML) is concerned with eliciting patterns from large volumes of data (Nguyen et al. 2019, Ongsulee 2017), whereas *LBD* aims to uncover novel knowledge patterns from vast quantities of literature. This indicates that there is a case for using *machine learning* (or even its subfield *deep learning*) in LBD. More specifically, this study attempts to accommodate ML and deep learning techniques, both of which provide ample opportunities for identifying intricate structures in data (LeCun et al. 2015). This process is crucial for complex reasoning tasks such as LBD. Otherwise stated, perceiving complex patterns in data through the integration of machine learning techniques provides an extended platform to make better predictive decisions in an automated manner. The experimental results highlighted the need to perform such large-scale pattern mining (with the use of meaningful data or characteristics) to distinguish novel knowledge linkages more precisely.

Most of the previous LBD models rely on semantic inferences performed using domain-specific knowledge resources to discover novel knowledge linkages from the literature. Nevertheless, the use of such external knowledge inferences based on domain-specific knowledge resources inhibits the LBD model's *reusability* and *portability*. This is because of their restrictive prediction settings, which may not be supportive and may even be unavailable in other problem settings (i.e., when focusing on *reusable applications*) or other domains (i.e., when focusing on *portable domains*). The benefits that LBD models offer are domain-agnostic and could be broadly applicable to almost every discipline due to the escalating scientific knowledge growth, which is commonly visible in all disciplines. Moreover, given the objective of discovering potentially novel linkages, the use of the discovery component in LBD models could also be broadly applicable to numerous other problem settings. Therefore, reusability and portability are two crucial design properties that should be considered when developing LBD models to ensure their widespread applicability. To support the idea of reusability and portability, the LBD models proposed in this study are completely free from external knowledge inferences that depend on domain-dependent knowledge resources. More specifically, the proposed LBD models' predictive performances do not rely on domain-specialised knowledge resources. As a result, they are portable and can be broadly reused, ensuring their widespread applicability towards providing broader community benefits.

The LBD models proposed in this study incorporated *semantically infused temporal trajectories* (i.e., *diachronic semantic inferences*) as the core analysis setting. Therefore, these models can be broadly divided into two categories based on how these proposed temporal trajectories are being analysed: *manifesting the direct usage* and *manifesting the indirect usage.* The first category refers to LBD models that directly extract potential semantically infused temporal signals from the proposed diachronic semantic inferences to make predictions. Therefore, the two proposed LBD models; *dedicated trajectory model* and *feature-based trajectory model* represent this category. The latter category refer to LBD models that do not directly extract potential signals from the diachronic semantic inferences, and instead use them as a *medium* to discover potential new knowledge linkages, as in the *trajectory alignment model.* The experimental results indicate that the proposed diachronic semantic inferences are efficient in both direct and indirect usages. Succinctly, the robust predictive performance evident in both the direct and indirect settings of the proposed diachronic inferences further supports their contribution towards the discovery of new knowledge linkages.

5.10.4 Limitations

The *semantically infused temporal trajectories* of scientific topics are the *core analysis unit* of the proposed LBD models. Therefore, new scientific topics that only appear in the last time-slice of the literature (i.e., time-slice T in Figure 5.2) do not have temporal trajectories, as they only exist in that final time-slice. This means that implicit knowledge linkages involving such recently added scientific topics are not captured by the proposed LBD models. In other words, reliance on the temporal trajectories of scientific topics limits the proposed LBD models' ability to discover novel knowledge linkages involving scientific topics that are *emergent* in the literature.

This study employs *time-sliced evaluation* to analyse and compare results since it resembles the characteristics of an ideal evaluation setting such as *automated*, *replicable*, *quantifiable*, *informative* and *modular*. These characteristics are lacking in other LBD evaluation techniques as discussed in detail in Chapter 3. Nevertheless, the reliance on *co-occurrence* in the time-sliced evaluation may introduce noise, since co-occurrence does not necessarily imply a legitimate relationship between two topics. Therefore, the use of time-sliced evaluation only provides an approximated platform from which to analyse and compare results.

5.11 Summary

The main focus of this study was to understand the potential contribution of co-modelling *vector semantics* with the *temporal dimension* to discovering novel knowledge linkages. The main incentive for intermingling these two notions came from two key observations in systematic literature review discussed in Chapter 2. Firstly, the sparsity of LBD studies that attempt to incorporate modern word embedding techniques, which was observed in the *timeline analysis*. Secondly, previous LBD research contains scant usage of temporal details due to a focus on static snapshots of the literature, as observed in the *categorisations* of computational techniques. More specifically, the main objective of this study was to verify whether the complementary integration of modern word embedding techniques with temporally charged environments and scrutinising the *semantic evolution* of scientific topics enriches the typical static cues used in the LBD literature. To construct temporally encoded semantic spaces, this study incorporated diachronic word embeddings, a research field which is emerging as a result of modern developments of word embedding techniques. In the constructed diachronic word embeddings, this study disentangled the semantic evolution of scientific topics at multiple levels. This facilitates a comprehension of the dynamic behaviour of scientific topics, and allows for the capture of meaningful semantically infused temporal signals, in order to discover novel knowledge linkages with high precision. In this regard, this study considered the derived semantically infused temporal trajectories as the main analysis unit in the experiments. Overall, the experimental results showcase the strength of the holistic integration of these two notions in the LBD context to enhance the predictive performance.

5.11.1 Major Contributions

Through this study, the thesis was able to provide several insights which, to the best of our knowledge, are new in the LBD discipline. The major contributions of this study are summarised below and are discussed in detail in Chapter 8.

- Being the first study in the LBD discipline to incorporate a circumstantial temporal component by utilising a wide range of techniques from areas such as *sequence mining*, *time series analysis* and *signal processing*, in order to perform a fine-grained analysis of semantically infused temporal trajectories.

- Being the first study to introduce patterns based on *relativity* by taking inspiration from molecular docking mechanism.

- Demonstrating not only the *direct uses* of the proposed diachronic semantic inferences, but also their *indirect uses* through the trajectory alignment model.

- The experimental results verified the efficacy of the proposed LBD models (i.e., both *direct* and *indirect* usage of diachronic semantic inferences) in all experiments, performed under different settings.

- The proposed semantic shift types in isolation (i.e., *ISS*, *PSS* and *NSS*) also demonstrated high prediction performances (in both *direct* and *indirect* uses of diachronic semantic inferences) compared to the baseline models, indicating the predictive power of the proposed semantically infused temporal trajectories, even individually.

- The prediction performance of the proposed LBD models does not depend on semantic inferences performed using external domain-dependent knowledge resources, which ensures their *reusability* (in various problem settings) and *portability* (in various academic domains), offering the opportunity to provide broader community benefits.

Chapter 6

Reusability

6.1 Introduction

Reuse research focuses on efficiently reusing components (or similar artifacts) in new applications (Mooney 1995). Creatively uncovering new application areas of reusability increases the *dependability* (or *reliability*) of the reused components (Ahmaro et al. 2014, Singh et al. 2010). With this aim in mind, the thesis required to further assess the robust predictive performances of the proposed LBD models (discussed in Chapter 5) by conducting reuse research. To this end, the following question was evoked: *'how can the reusability of the proposed LBD models be ensured in a new application area, to further confirm their robust predictive power?'*. To seek potential extensions of such reuse research in the context of LBD, this chapter considered the existing application areas of the LBD discipline, following a method similar to *opportunistic reuse* (i.e., making new capabilities by welding together pieces of components originally developed for distinct problem setting(s)) (Katz et al. 1994, Ncube et al. 2008).

Even though the primary objective of LBD models is to mitigate the effects of knowledge over-specialisation by helping researchers to formulate novel research hypotheses (Swanson & Smalheiser 1996), there are several special-purpose LBD models that have been developed to cater to specific problem areas (Henry & McInnes 2017). Among these application areas, *drug development*, *drug repositioning* and *adverse drug reactions* can be considered the most popular selections (Henry & McInnes 2017, Thilakaratne et al.

2019*b*). The COVID-19 pandemic underscores the need to contribute to such special-purpose application areas of LBD more urgently than ever. This critical situation provides an impetus to explore this timely direction, and to demonstrate the reusability of the proposed LBD models described in Chapter 5.

The development of new drugs is a costly and time-intensive procedure that involves costs between 500 million to 2 billion dollars and takes 10 to 15 years to bring a new drug from the laboratory to market (Wei et al. 2015, Henry & McInnes 2017). Nevertheless, the success rate of such newly developed drugs is less than 10%, and FDA approval of new drugs is declining (Henry & McInnes 2017). Identifying potential *chemical-disease* interactions play a crucial role in drug discovery, biocuration and pharmacovigilance (Chen et al. 2015, Li et al. 2015). It has been reported that chemicals, diseases and their relationships are some of the most searched topics in PubMed (Wei et al. 2015). Despite the importance of such chemical-disease relations in numerous biomedical research and healthcare, including *drug discovery* and *safety surveillance*, many undiscovered interactions could be buried in the literature due to its exponential growth (Wei et al. 2015). This suggests the need to elicit such latent interactions from the unstructured text using natural language processing techniques (Wei et al. 2015, Li et al. 2015).

The intrinsic objective of LBD studies is to discover implicit novel knowledge linkages hidden in the vast academic literature, this indicates the potential benefits that LBD models could offer to the discovery of hidden chemical-disease relations. The recent COVID-19 pandemic illustrates the urgent need for research on this selected reuse setting. The main research objective of this study is:

"to validate the predictive power of the proposed LBD models through reuse research, with the goal of providing broader community benefits"

as defined at the outset of this thesis (i.e., *main research objective 4 (RO4)* in Chapter 1). With the overarching goal of contributing to this timely application area, this chapter revolves around the main research question (*RQ4*): *'are the proposed LBD models reusable for the purpose of discovering latent chemicals that may have potential interactions for a given disease?'*. More specifically, since a closely related problem area to the LBD context is selected to demonstrate the reusability of the proposed LBD models, this can be considered *vertical reuse* (Jalender et al. 2010). Bearing in mind this chapter's focus on reusability, RQ4 is further divided into the following two sub research objectives.

- *RO4.1. Defining a methodical reuse plan in consideration of the opportunistic reuse nature of the problem setting, to ensure that the meaning of reusability is preserved during the adaptations* (discussed in Section 6.2).

- *RO4.2. Adapting the proposed LBD models to this new reuse setting in accordance with the defined reuse plan to make predictions about potential chemical-disease relations* (discussed in Section 6.3).

This chapter is organised as follows. Section 6.2 outlines the underlying reuse framework that is employed, as well as discussing reuse considerations and reuse objectives. Section 6.3 discusses the adaptation of the proposed LBD framework (discussed in Chapter 5) to this new reuse setting. Section 6.4 describes the experimental setup of this chapter, including the datasets, test cases, baselines and other design considerations of this study. Section 6.5 presents the results of the experiments, along with an extended discussion of the key observations. The latter part of this discussion highlights the strengths of the proposed LBD models that were evident in this new reuse setting. Furthermore, this study looks closely at the predictive performances of the adapted LBD models with the intention of understanding potential future improvements that could be considered in the next iterations of the selected reuse framework (discussed in Section 6.2). Section 6.6 concludes the chapter by outlining the key findings and major contributions.

6.2 Structure of the Reuse Research

The purpose of this section is to describe the structure underpinning this study, which is performed as part of the thesis' reuse research. In this regard, the first section discusses the underlying reusability framework, which is employed to ensure that the study falls within the boundaries of reuse research. Subsequently, the major differences between this problem setting and the setting utilised in Chapter 5 are discussed. The main reason for this discussion is that these differences indicate the instances where adaptations are required during the process of assembling the reuse components, in accordance with the reusability framework employed. Lastly, the main focus of this reuse research is outlined by revisiting the definition of reusability.

6.2.1 Reusability Framework

This study uses *grab-and-glue* as its underlying reusability framework. This framework is based on the assembling of components rather than building the components to demonstrate their reusability (Robinson et al. 2004). More specifically, the components that are intended to be reused are grabbed and glued, so as to quickly assemble a model in a new reuse setting. Subsequently, this quickly assembled model is validated to verify its fitness for the intended purpose. If its fitness is judged to be satisfactory, it can be concluded that an understanding of the problem has been attained. If its fitness is judged unsatisfactory, the assembled model is rejected, and grab-and-glue is performed differently (Robinson et al. 2004). In essence, this process can be performed iteratively until fitness for purpose is established. Nevertheless, this study only considers one iteration in the grab-and-glue framework to verify the reusability of the LBD models.

6.2.2 Reuse Considerations

Prior to conducting reuse research, it is important to identify the key differences between the two problem settings (i.e., the problem setting discussed in Chapter 5 and the setting used in this chapter). These key differences can be considered as areas of adaptation when assembling the reuse components.

The major difference between the setting used in this chapter and the previous setting is the *input* that the user provides to the LBD model to initiate the knowledge discovery process. In the current setting, the user merely enters a disease name as the input topic (i.e., only one user topic, namely *topic* A). In essence, this setting does not have a topic C (whereas the previous setting does). Therefore, in this new reuse setting, the LBD model is required to elicit novel knowledge in a more open-ended manner, (based only on *topic* A) to discover meaningful novel knowledge linkages (or chemicals that are currently unknown, but potentially related to the user-defined disease).

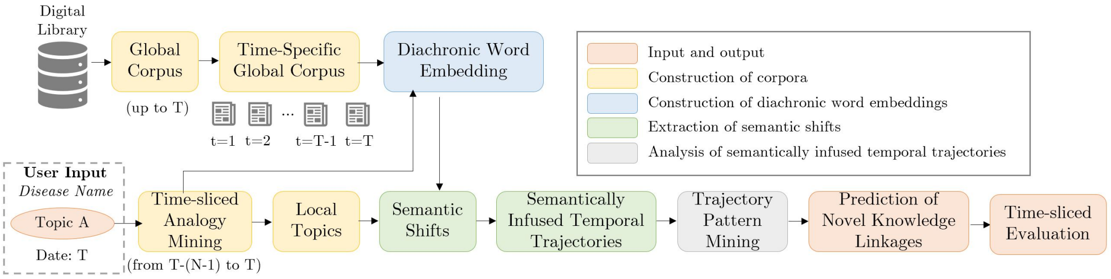

Digital Library
Global Corpus
Time-Specific Global Corpus
Diachronic Word Embedding
(up to T)
t=1
t=2
...
t=T-1
t=T
Input and output
Construction of corpora
Construction of diachronic word embeddings
Extraction of semantic shifts
Analysis of semantically infused temporal trajectories
User Input
Disease Name
Topic A
Date: T
Time-sliced Analogy Mining
(from T-(N-1) to T)
Local Topics
Semantic Shifts
Semantically Infused Temporal Trajectories
Trajectory Pattern Mining
Prediction of Novel Knowledge Linkages
Time-sliced Evaluation

FIGURE 6.1: Schematic overview of the adaptation of the proposed LBD framework

6.2.3 Reuse Focus

Recall that reuse denotes the process of efficiently using components designed for one application in new applications (Mooney 1995, Katz et al. 1994). Reuse can be demonstrated in application areas that are closely related to the original area (i.e., *vertical reuse*, which is similar to the kind of reuse demonstrated in this study) or even in broadly different application areas (i.e., *horizontal reuse*) (Jalender et al. 2010, Katz et al. 1994). The cost of adapting components to facilitate a new function in a new reuse setting (relative to the original purpose of those components) should be *little or none* (Katz et al. 1994). With this in mind, this chapter is not about developing new LBD models or features from scratch. Instead, the focus of this study is on quickly adapting and assembling the proposed LBD models (discussed in Chapter 5) to this new problem setting (i.e., *grab-and-glue*) to assess their fitness for the intended purpose.

6.3 Adaptation of the Proposed LBD Framework

This section discusses how the proposed LBD framework discussed in Chapter 5 was adapted to this new problem setting (Figure 6.1). As discussed in Section 6.2, the adaptation is performed with minimal effort (to adhere to the earlier definition of reusability), and only one iteration of the *grab-and-glue* process is performed to identify whether fitness for purpose was established. The new objective of this adapted version of the LBD framework is to discover *chemicals* with potential novel relationships to a user-defined input disease (i.e., *topic A*).

6.3.1 Local Topic Extraction

Since the new reuse setting only employs one topic of interest to elicit new knowledge, the local topics relevant to the input topic (i.e., *topic A*) need to be identified in a more open-ended manner. In this regard, this study exploits the analogical reasoning power of word embeddings to construct an initial list of local topics to initiate the knowledge discovery process.

6.3.1.1 Analogy Mining

The vector representations of words generated through neural network methods such as *word2vec* (discussed in Chapter 5) have shown surprising capacity to detect verbal *analogies* (Allen & Hospedales 2019, Chen et al. 2017). These verbal analogies between vectors can be represented through the *parallelogram model*. Parallelogram model states that the four elements involved in an analogy adhere to a regularity rule, much like a parallelogram in vector space (Murena et al. 2018). The parallelogram model was reincarnated in recent machine learning research through popular embedding methods such as word2vec, which have been successfully applied to a wide variety of natural language processing tasks (Allen & Hospedales 2019, Chen et al. 2017). These studies suggest that the verbal analogies enabled through these vector representations may accommodate sufficient information to allow for relationships to be directly inferred from them (Chen et al. 2017). The application of the parallelogram model of an analogy using vector representations is considered to be domain-agnostic and broadly usable in both semantic and perceptual domains (Chen et al. 2017). For instance, consider the verbal analogy 'w_a is to w_{a^*} as w_b is to w_{b^*}', which often satisfies $\mathbf{w_{a^*}}$ - $\mathbf{w_a}$ + $\mathbf{w_b} \approx \mathbf{w_{b^*}}$ where $\mathbf{w_i}$ defines the vector representation of the word w_i (Figure 6.2) (Allen & Hospedales 2019). Following this notion, this study explores the topics that may be potentially relevant to *topic A* using analogy mining to initiate the knowledge discovery process.

6.3.1.2 Time-sliced Analogy Mining

This study explores the notion of *time-sliced analogy mining*, as illustrated in Figure 6.1, where recent N vector spaces are considered for analogy mining (i.e., from *T-(N-1)* to

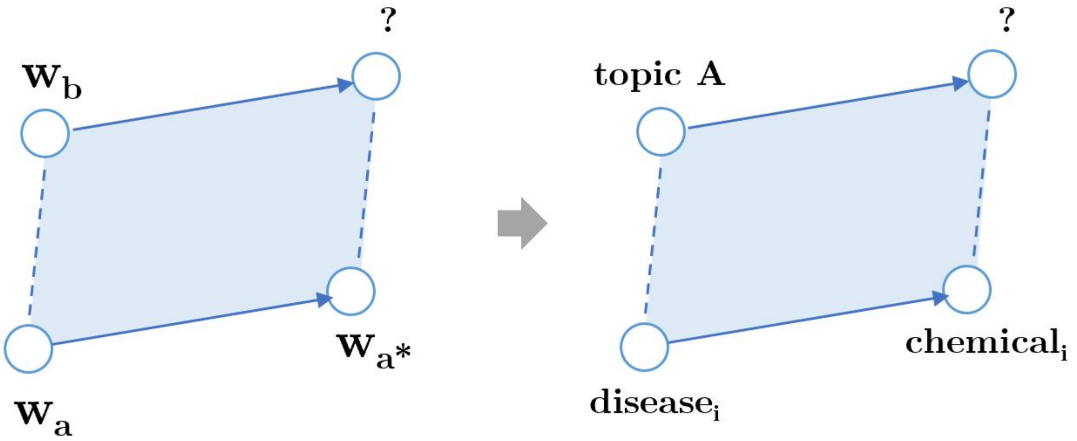


FIGURE 6.2: Completing the analogy $\mathbf{w_a} : \mathbf{w_{a^*}} :: \mathbf{w_b} : ?$ by adding the difference vector between $\mathbf{w_a}$ and $\mathbf{w_{a^*}}$ to $\mathbf{w_b}$, forming a parallelogram in vector space

T). The reason for not considering only the last time-slice (i.e., T) is that this study aims to dilate the search scope in order to perform large-scale knowledge discovery. Such a broader search may be beneficial during the knowledge discovery process, allowing for the capture of *surprising* or *radical* knowledge linkages (Jha et al. 2018). The aforementioned process of analogy mining in each vector space is performed in the form of $disease_i$: $chemical_i$ :: *topic A* : *?*, as illustrated in Figure 6.2. The chemicals derived through this phase for *topic A* are considered *local topics* in this setting (Figure 6.1). It is important to note that these extracted local topics merely indicate chemicals that *may be* interested with reference to the user-defined disease, not the potential novel knowledge. Thus, it is important to sieve these local topics using the knowledge discovery process in order to retain chemicals that are potentially relevant and novel with respect to the user-specified disease. For this purpose, (and in a manner similar to that in the previous setting discussed in Chapter 5), *semantic shifts* are employed to perform knowledge discovery.

6.3.2 Semantic Shifts

In this new reuse setting, the user merely inserts the *name of a disease* to initiate the discovery process (i.e., *topic A*). Nevertheless, recall that in Chapter 5, semantic shifts were defined using two topics (i.e., *topic A* and *topic C*). Thus, measures that use both topics A and C are required to be adapted to this new reuse setting by only focusing on *topic A*. Table 6.1 summarises the adaptation of the semantic shifts proposed in Chapter 5 to this new reuse setting.

6.3.2.1 Individual Semantic Shifts

Since this category focuses on the semantic change of topics based on the topic itself (i.e., not involving topic A and/or C), the two semantic shift measures defined under this category (*Individual Global Shifts (IGS)* and *Individual Local Shifts (ILS)*) remain *unchanged* in this new setting (Table 6.1).

6.3.2.2 Pairwise Semantic Shifts

This category analyses the semantic change of topics relative to the two user-defined input topics A and C. Thus, this category is required to be adapted to this new setting by only considering *topic* A. The two types of measures proposed in this category were previously named *pairwise semantic displacement* and *pairwise distance proximity*. Note that the word *pairwise* in these two measures was originally used to indicate that they are based on the two user-defined input topics A and C.

- *Pairwise Semantic Displacement (PSD):* Originally, this measure was intended to capture the concept's ($\mathbf{w}_i$) semantic change over time relative to topics A ($\mathbf{w}_A$) and C ($\mathbf{w}_C$). Due to the unavailability of *topic* C in this new setting, PSD was redefined with reference to the concept's semantic change relative to *topic* A only (i.e., cos-sim($\mathbf{w}_i^{(t)}, \mathbf{w}_A^{(t)}$)). More specifically, in this setting, PSD verifies whether the concept displays any growing semantic similarity with respect to topic A over time.

- *Pairwise Distance Proximity (PDP):* Originally, the idea of this measure was to verify whether the temporal trajectory of a concept was leaning towards (i.e., in close proximity to) both user-defined topics A and C. This was because Chapter 5 was seeking topics that bridge the two topics A and C. Thus, the concept's trajectory ought to have inclined towards both the input topics. Due to the unavailability of two topics in the new setting, this measure is not compatible with this setting. Since the idea of this chapter is not to develop new features (as discussed in Section 6.2) but to adapt the existing features (if compatible) with minimum effort, this semantic shift has been removed from the knowledge discovery process employed in this setting (Table 6.1).

6.3.2.3 Neighbourhood Semantic Shifts

This category originally denoted the extended measures of *pairwise semantic displacement* and *pairwise distance proximity* by incorporating not only A and C, but also their *recent core meanings*. Thus, the measures proposed under this category of semantic shifts need to be adapted to a setting featuring one topic only.

- *Neighbourhood Semantic Displacement (NSD):* The main difference between this measure and the *pairwise semantic displacement* measure is that it also involves the *recent core meaning* of the input topics A ($\mathbf{w}_A$) and C ($\mathbf{w}_C$). As there is no topic C in the new reuse setting, NSD is adapted in such a way that the concept's semantic shift is measured relative to topic A and its recent core meaning only (i.e., topic C and its recent core meaning are excluded). Thus, the adapted measure captures how well the concept ($\mathbf{w}_i$) semantically connects with topic A and its recent core meaning.

- *Neighbourhood Distance Proximity (NDP):* This measure represents the neighbourhood variant of the *pairwise distance proximity* measure. Since *pairwise distance proximity* was removed from this study due to its incompatibility with the reuse setting, this measure is also removed in this new setting.

Table 6.1: Adaptation of the proposed semantic shift measures in the new reuse setting

Semantic Shift	**Compatibility**	**Cost of Adaptation**	**Description**
Individual Global Shifts (IGS)	✓	None	Does not involve A and/or C topics; thus, no adaptation is required.
Individual Local Shifts (ILS)	✓	None	Does not involve A and/or C topics; thus, no adaptation is required.
Pairwise Semantic Displacement (PSD)	✓	Negligible	Involves both topics A and C; thus, adaptation is performed by retaining the semantic inference corresponding to topic A only.

Pairwise Distance Proximity (PDP)	×	–	Requires both topics A and C; thus, not compatible.
Neighbourhood Semantic Displacement (NSD)	✓	Negligible	Involves both A and C topics; thus, adaptation is performed by retaining the semantic inference corresponding to topic A only.
Neighbourhood Distance Proximity (NDP)	×	–	Requires both topics A and C; thus, not compatible.

6.3.2.4 Frequency Heuristics

The two frequency heuristics considered with the aforementioned semantic shifts in Chapter 5 are also adapted to the new reuse setting if topic A and/or C are utilised. These two frequency heuristics obtained from previous LBD research (Torvik & Smalheiser 2007) are *Global Frequency Heuristic (GFH)* and *Local Frequency Heuristic (LFH)*. *GFH* penalises concepts that are extremely common or extremely rare in the literature. Since GFH does not incorporate topic A or C, it remains unchanged in this new reuse setting. Originally, *LFH* was employed to penalise concepts that only occurred once in A or C literature, where the corresponding literature set had over 1000 records. Since this measure involves both topics A and C, this required to be altered by only incorporating topic A. Thus, in the new setting, this feature is set by only focusing on the concept's frequency with reference to topic A (i.e., $n(A,\ w_i)$>0). Therefore, in this new setting, LFH indicates whether the extracted local topics already have an established relationship with topic A on or before time T (i.e., indicating that it is not a novel knowledge linkage).

To summarise, in this reuse setting, this study only considered four types of semantic shifts (Table 6.1) along with the two frequency heuristics. As in the previous setting (discussed in Chapter 5), the semantic shifts are denoted in the form of *semantically infused temporal trajectories*, to facilitate temporal analysis. More specifically, the four semantically infused temporal trajectories of concept w_i are constructed in the form of: $\mathrm{TJ}^{\mathrm{IGS}}(w_i) = (d^{\mathrm{IGS}}(w_i^{y}),\ d^{\mathrm{IGS}}(w_i^{y+1})),\ ...,\ d^{\mathrm{IGS}}(w_i^{T-1}),\ d^{\mathrm{IGS}}(w_i^{T}))$

$$\mathrm{TJ}^{\mathrm{ILS}}(w_i) = (d^{\mathrm{ILS}}(w_i^{y}),\ d^{\mathrm{ILS}}(w_i^{y+1})),\ ...,\ d^{\mathrm{ILS}}(w_i^{T-1}),\ d^{\mathrm{ILS}}(w_i^{T}))$$

$$\mathrm{TJ}^{\mathrm{PSD}}(w_i) = (s^{\mathrm{PSD}}(w_i^{y}),\ s^{\mathrm{PSD}}(w_i^{y+1})),\ ...,\ s^{\mathrm{PSD}}(w_i^{T-1}),\ s^{\mathrm{PSD}}(w_i^{T}))$$

$$\mathrm{TJ}^{\mathrm{NSD}}(w_i) = (s^{\mathrm{NSD}}(w_i^{y}),\ s^{\mathrm{NSD}}(w_i^{y+1})),\ ...,\ s^{\mathrm{NSD}}(w_i^{T-1}),\ s^{\mathrm{NSD}}(w_i^{T}))$$

where y is the first occurrence of w_i in the dataset, s is a similarity measure and d is a distance measure.

6.3.3 Reuse of Proposed LBD Models

The main three LBD models proposed in Chapter 5, the *Dedicated Trajectory Model (DTM)*, *Feature-based Trajectory Model (FTM)* and *Trajectory Alignment Model (TAM)* are used in this new experimental setting to sieve novel knowledge linkages from the remaining local topics. To summarise, *DTM* uses novel advancements in deep learning techniques by employing *LSTM* and *CNN* as the two main building blocks of the proposed neural network architectures. The main purpose of these developed neural network architectures is to scrutinise extracted semantically infused temporal trajectories in order to discover patterns of potential novel knowledge linkages. Further details on this model can be found in Section 5.7 of Chapter 5. *FTM* follows the traditional machine learning process using hand-crafted features from the semantically infused temporal trajectories to discover potential novel knowledge linkages. These features mainly comprise two feature categories, *trajectory values-based* and *trajectory shape-based* features. Further details on this model can be found in Section 5.8 of Chapter 5. *TAM* uses the semantically infused temporal trajectories of actual novel knowledge linkages as templates (in a *trajectory repository*) to analyse the extent to which the trajectories of local topics demonstrate the patterns exhibited in these templates. To measure how similar or dissimilar the trajectories of local topics are to the templates, a *trajectory alignment* procedure is proposed in this LBD model. Further details on this model can be found in Section 5.9 of Chapter 5.

6.4 Experimental Setup

The purpose of this section is to discuss the experimental setup employed in this reuse research. In this regard, the initial part of this section discusses the datasets, test cases and baselines utilised. The latter section provides details on how *time-sliced analogy mining* was performed and how the *template repository* of the proposed LBD model, *TAM* was constructed using the additional data sources used in this study.

6.4.1 Datasets and Test Cases

As with previous experiments in this thesis, *MEDLINE* was used as the main data source for this study. The *MEDLINE* data field used in the study's experiments was *MeSH* (discussed in Chapter 3). In addition to MEDLINE, the study also made use of the *chemical-disease* interactions reported in the *Comparative Toxicogenomics Database (CTD)* (Davis et al. 2019). The chemical-disease associations stored in CTD are either *curated* (i.e., extracted from the published literature) or *inferred* (i.e., extracted using transitive inferences from the literature) (Yang, Zhao, Waxman & Zhao 2019, Zhang et al. 2018, Wang et al. 2017). CTD is considered a primary data source that facilitates an understanding of the way environmental exposures impact human health (Yang, Zhao, Waxman & Zhao 2019).

With regard to test cases (and for the sake of consistency with the remaining chapters of the thesis), this study used the disease names from the *golden test cases* (where available) as *topic A* to initiate knowledge discovery in this reuse setting. The selected disease names from the golden test cases included *Raynaud's Disease (RD)* (Swanson 1986), *Migraine Disorder (MIG)* (Swanson 1988), *Alzheimer's Disease (AD)* (Smalheiser & Swanson 1996) and *Schizophrenia (SZ)* (Smalheiser & Swanson 1998).

6.4.2 Baselines

Bearing in mind the previous experimental setup discussed in Chapter 5, this study incorporated the following baseline models: *Arrowsmith (AR)*, *Dynamic Embeddings (DE)* and *Static Embeddings (SE)*. Moreover, within the LBD field, there is a growing research interest in integrating *link prediction techniques* to discover future links between concepts (Yang et al. 2017, Kastrin et al. 2014*b*). Since link prediction techniques are suited to this reuse setting and have been used as baselines in previous LBD studies for the purpose of comparing results (Jha, Xun, Wang & Zhang 2019, Lever et al. 2018), this study also incorporated three popular and classical link prediction techniques, namely

Common Neighbours, *Jaccard's Index* and *Preferential Attachment* as baselines in this reuse setting (discussed in Chapter 3).

Note that this reuse setting did not employ the two LBD models: *Bitola (BI)* and *TF-IDF (TI)* as baselines, as in Chapter 5. The main reason for excluding these two LBD models was that they follow the traditional ABC model in facilitating *one-node searches*, similar to this reuse setting. Thus, when these two LBD models get adapted to this reuse setting, the meanings conveyed through these measures become irrelevant, as summarised below.

- *Bitola (BI):* The default metric used by *BI* is *confidence*. This is is expressed as $\frac{|D_A \cap D_{lp_i}|}{|D_A|}$, where D_x is the set of records in which the term x is included (Hristovski et al. 2001, Yetisgen-Yildiz & Pratt 2009). In this reuse setting, lp_i denotes chemicals that may have potential relationships with the user-defined disease. Thus, if a local topic lp_i has $\mid D_A \cap D_{lp_i} \mid > 0$, this indicates that the relevant chemical already has a connection with the user-defined disease. Thus, the knowledge linkage between the chemical and the disease is not a novel one. Therefore, this measure becomes meaningless when adapted to this reuse setting. More specifically, this metric is only make sense if knowledge discovery is performed using the typical ABC setting.

- *TF-IDF (TI):* Much like *BI*, the initial component of TF-IDF (which is term frequency; TF) calculates the number of times a local topic lp_i and the user-defined disease have occurred together. Thus, when TF >0, this indicates that the local topic already has a connection with the disease. Thus, this metric becomes meaningless in the process of adaptation to this reuse setting. In other words, this metric is only valid in the typical ABC discovery setting.

When adapting the three baseline models used in Chapter 5 *Arrowsmith (AR)*, *Dynamic Embeddings (DE)* and *Static Embeddings (SE)* to this reuse setting, the inference related to topic A only is retained by excluding topic C. This is similar to the cases involving our proposed LBD models (discussed in Section 6.3.2). In this process of adapting the three baseline models to the current setting, the following two features in *Arrowsmith (AR)* baseline are removed, due to their incompatibility with this reuse setting (similar to ours, as discussed in Section 6.3.2 to adhere with the reuse plan).

- Feature f_2 (discussed in Chapter 3) which is characterised by the question: *do sub-literatures AB and BC have any common MeSH terms?* involves both topics A and C; thus, adaptation needs to be performed by retaining the semantic inference corresponding to topic A only. Nevertheless, if only topic A is considered, this measure becomes meaningless in our present context, as it is infeasible to calculate shared MeSH terms in the context of AB literature only.

- Feature f_7 (discussed in Chapter 3) which is characterised by the question: *does the B concept highly characteristic in A and C literature?* also includes both topics A and C; thus, needs to be adapted to this reuse setting by retaining the semantic inference corresponding to topic A in isolation. In situations similar to this, the expected term occurrence in the literature used in f_7 could be calculated using a hypergeometric distribution expressed as follows: $Pr(X = x) = \binom{f_1}{x}\binom{N-f_1}{f_2-x}/\binom{N}{x}$ for $x = 0, 1, 2, ..., min(f_1, f_2)$, where N is the total paper count, f_1 represents the papers that have a local topic lp_i and f_2 represents the papers specific to the user-defined disease. Since N is relatively large, the aforementioned hypergeometric distribution can be approximated using a Poisson distribution defined using: $Pr(X = x) \approx {}^{e^{-\lambda}\lambda^x}/_{x!}$ for $x = 0, 1, 2, ...$ where $\lambda = {}^{f_1 f_2}/_N$ (Smalheiser et al. 2011, 2008). This could alternatively be considered as a problem with a number of balls N in an urn, where f_1 is denoted using black balls. When randomly selecting f_2 distinct balls, the number of black balls selected (Smalheiser et al. 2011) resembles f_7 in this setting. Nevertheless, as in the case of *BI* and *TI* (discussed above), this is a situation where a local topic lp_i already has a connection with the user-defined disease; thus, this measure becomes meaningless when adapted to this reuse setting.

In summary, this study incorporated the following six baseline models in this reuse setting: *Arrowsmith (AR)*, *Dynamic Embeddings (DE)*, *Static Embeddings (SE)*, *Common Neighbours (CN)*, *Jaccard's Index (JI)* and *Preferential Attachment (PA)*.

6.4.3 Construction of Local Topics via Time-sliced Analogy Mining

This study extracted the *chemical-disease* pairs from the CTD as *seed pairs* in order to perform analogical mining, as discussed in Section 6.3.1. The only purpose of the CTD seed pairs at this stage was to develop an initial local topic list, denoting potential chemicals that *may be* worth exploring with regard to the disease mentioned as topic A to initiate the knowledge discovery process. More specifically, this study used the CTD chemicals available at time T to extract a maximum of ten *chemical-disease* associations (i.e., the first ten entries in CTD[1]) for each chemical, in order to perform time-sliced analogy mining to construct the initial list of local topics. Note that these local topics derived through analogy mining are not potential novel knowledge linkages. They merely serve as an independent initial vocabulary to initiate the knowledge discovery process. The novel knowledge linkages within these constructed local topics are discovered by employing the proposed LBD models, as discussed in Section 6.3.

6.4.4 Construction of the Template Repository of Trajectory Alignment Model

To construct a *template repository* of the trajectory alignment model (discussed in Section 6.3.3), this study incorporated the trajectories of chemicals in CTD, which are available at time T. Note that when constructing the *pairwise semantic displacement trajectory* and *neighbourhood semantic displacement trajectory* of these chemicals, the chemical-disease relationships that have not been realised by time T (i.e., no direct co-occurrences) are employed. The main reason for this is that the key purpose of the *template repository* is to collect potential trajectory shapes that showcase their semantic evolution before they actually get realised in the future. Moreover, chemical-disease instances for which the disease name is equivalent to topic A are not included when constructing the template repository. This helps to avoid biasing the trajectory alignment procedure. For instance, consider a chemical-disease pair in CTD where the chemical name is $chemical_x$, and the disease name is topic A. If the same chemical name (i.e., $chemical_x$) were present in local topics, the trajectory alignment would incur zero costs, which could ease the decision of the ML component. The main purpose of excluding chemical-disease relations where the disease name is topic A in the process of constructing the template repository to avoid such bias decisions.

[1] downloaded as at 5th of April, 2020

6.5 Results and Discussion

This section validates the predictive effects of the proposed LBD models in this new reuse setting. Tables 6.2, 6.3, 6.4, and 6.5 report the results of *precision at k (P@k)* for the test cases *RD*, *MIG*, *AD* and *SZ*, respectively. As in the case of Chapter 5, the k value was gradually increased from 10 to 100, at an interval of 10. When observing $P@k$ results, it is evident that the following three proposed LBD models often exhibited the highest predictive performances in every golden test cases across all the k values: *FTM* (discussed in Section 5.8), *DTM: LSTM_1* (discussed in Section 5.7) and *TAM* (discussed in Section 5.9). This verifies the potential positive influence of the proposed diachronic semantic inferences, not only in terms of their *direct uses*, but also their *indirect uses* in the LBD knowledge discovery process. This thesis observed the same conclusion across all the experimental setups in Chapter 5. The robust predictive performances evident even in the very first iteration of the grab-and-glue framework (discussed in Section 6.2) are indicative of the efficient reuse capabilities of the proposed LBD models.

Since $P@k$ is not sensitive to the ranking order of the correct predictions, this study used *Mean Average Precision at k (MAP@k)*, which favours models that often front-load the correct predictions (i.e., the relevant novel knowledge linkages that are ranked at high positions make a higher contribution to the average than the relevant novel knowledge linkages that are ranked at low positions). More specifically, the *MAP* is considered to be the *de facto gold standard* for evaluating information retrieval systems (Beitzel et al. 2009*b*), and it captures the *overall performance* of the LBD models across the golden test cases. Figure 6.3 presents the $MAP@k$ results across the selected golden test cases (i.e., *RD*, *MIG*, *AD* and *SZ*), where the value of k was gradually increased from 10 to 100, at an interval of 10 (also outlined in Table B.1).

From Figure 6.3, it is evident that all the variants of the proposed LBD models outperformed the baseline models (*AR*, *DE*, *SE*, *CN*, *JI* and *PA*). This indicates the robust predictive performance of the proposed LBD models, while also highlighting their potential reuse capabilities. Overall, *FTM* demonstrated the highest performance across the golden test cases. The second-highest performance was displayed by *DTM: LSTM_1*. *TAM* demonstrated the third-highest overall performance, especially in terms of the initial k values.

Table 6.2: P@k results for FO-RD test case where topic A is RD

Method	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**
AR (baseline)	0.2	0.1	0.067	0.075	0.1	0.117	0.114	0.113	0.122	0.12
DE (baseline)	0.1	0.15	0.1	0.1	0.1	0.117	0.143	0.163	0.156	0.15
SE (baseline)	0.2	0.25	0.167	0.15	**0.2**	**0.217**	**0.186**	**0.163**	**0.178**	**0.18**
CN (baseline)	0.0	0.0	0.0	0.0	0.02	0.033	0.057	0.05	0.056	0.06
JI (baseline)	0.1	0.1	0.067	0.075	0.08	0.083	0.071	0.075	0.067	0.06
PA (baseline)	0.0	0.0	0.067	0.05	0.08	0.083	0.086	0.088	0.089	0.08
DTM: LSTM_1	0.2	0.25	0.2	**0.2**	0.18	0.183	0.157	**0.163**	0.144	0.16
DTM: LSTM_2	0.0	0.05	0.067	0.075	0.1	0.083	0.071	0.075	0.078	0.1
DTM: LSTM_3	0.0	0.1	0.133	0.1	0.08	0.067	0.071	0.075	0.089	0.08
DTM: CNN	0.0	0.05	0.033	0.05	0.08	0.1	0.1	0.125	0.122	0.11
DTM: CNN_LSTM	0.0	0.05	0.033	0.05	0.06	0.067	0.071	0.088	0.089	0.09
DTM: LSTM_CNN	0.0	0.05	0.133	0.125	0.12	0.117	0.1	0.088	0.078	0.09
FTM	**0.4**	0.25	**0.233**	**0.2**	**0.2**	0.167	**0.186**	**0.163**	0.144	0.16
TAM	0.3	**0.3**	**0.233**	**0.2**	0.16	0.15	0.143	0.15	0.144	0.14

Table 6.3: P@k results for MG-MIG test case where topic A is MIG

Method	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**
AR (baseline)	0.1	0.1	0.1	0.15	0.18	0.2	0.186	0.163	0.178	0.16
DE (baseline)	0.5	0.35	0.233	0.25	0.22	0.183	0.171	0.175	0.178	0.19
SE (baseline)	0.1	0.1	0.133	0.175	0.18	0.183	0.186	0.175	0.167	0.17
CN (baseline)	0.0	0.0	0.0	0.0	0.0	0.017	0.029	0.038	0.033	0.04
JI (baseline)	0.0	0.0	0.0	0.0	0.0	0.017	0.029	0.05	0.044	0.04
PA (baseline)	0.0	0.0	0.033	0.075	0.08	0.1	0.114	0.125	0.122	0.13
DTM: LSTM_1	0.3	0.4	0.4	0.4	0.42	**0.4**	**0.371**	0.338	0.333	0.33
DTM: LSTM_2	0.2	0.25	0.233	0.2	0.26	0.3	0.314	0.338	0.344	0.33
DTM: LSTM_3	0.4	0.35	0.267	0.25	0.26	0.267	0.257	0.263	0.244	0.25
DTM: CNN	0.4	0.35	0.367	0.375	0.36	0.317	0.329	0.325	0.3	0.29
DTM: CNN_LSTM	0.1	0.25	0.333	0.275	0.32	0.35	0.329	0.338	0.344	0.33
DTM: LSTM_CNN	0.4	**0.5**	0.333	0.325	0.32	0.283	0.286	0.288	0.3	0.29
FTM	**0.6**	**0.5**	**0.5**	**0.5**	**0.46**	**0.4**	**0.371**	**0.4**	**0.389**	**0.39**
TAM	0.5	0.45	0.333	0.325	0.3	0.3	0.3	0.325	0.311	0.31

Table 6.4: P@k results for AD-INN test case where topic A is AD

Method	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**
AR (baseline)	0.6	0.7	0.7	0.65	0.62	0.6	0.6	0.588	0.589	0.58
DE (baseline)	0.2	0.2	0.233	0.3	0.3	0.267	0.286	0.288	0.3	0.29
SE (baseline)	0.0	0.05	0.033	0.05	0.06	0.05	0.086	0.088	0.089	0.13
CN (baseline)	0.0	0.0	0.0	0.0	0.0	0.0	0.014	0.013	0.011	0.03
JI (baseline)	0.0	0.05	0.033	0.05	0.04	0.033	0.043	0.038	0.044	0.04
PA (baseline)	0.0	0.0	0.0	0.0	0.0	0.017	0.029	0.038	0.056	0.05
DTM: LSTM_1	**0.9**	0.8	0.833	0.8	0.8	0.833	**0.829**	0.813	**0.822**	**0.83**
DTM: LSTM_2	0.8	0.7	0.767	**0.825**	**0.82**	0.833	**0.829**	0.813	**0.822**	**0.83**
DTM: LSTM_3	0.8	**0.9**	**0.867**	**0.825**	**0.82**	**0.85**	**0.829**	**0.838**	0.811	0.8
DTM: CNN	0.7	0.55	0.567	0.55	0.52	0.517	0.557	0.567	0.578	0.58
DTM: CNN_LSTM	0.8	0.65	0.667	0.675	0.7	0.717	0.757	0.763	0.756	0.75
DTM: LSTM_CNN	0.8	0.75	0.733	0.8	**0.82**	0.833	0.8	0.813	0.811	0.8
FTM	0.6	0.65	0.733	0.725	0.72	0.717	0.714	0.713	0.711	0.72
TAM	0.6	0.65	0.6	0.55	0.54	0.55	0.529	0.55	0.556	0.56

Table 6.5: P@k results for SZ-PA2 test case where topic A is SZ

Method	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**
AR (baseline)	0.1	0.1	0.167	0.175	0.18	0.2	0.214	0.25	0.267	0.26
DE (baseline)	0.0	0.05	0.033	0.1	0.1	0.1	0.114	0.113	0.1	0.11
SE (baseline)	0.0	0.0	0.0	0.0	0.0	0.033	0.043	0.05	0.067	0.08
CN (baseline)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
JI (baseline)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
PA (baseline)	0.0	0.0	0.0	0.0	0.0	0.0	0.029	0.025	0.044	0.04
DTM: LSTM_1	0.8	0.5	0.5	0.45	0.46	0.467	0.443	0.463	0.467	0.45
DTM: LSTM_2	0.5	0.35	0.3	0.25	0.28	0.317	0.329	0.363	0.367	0.35
DTM: LSTM_3	0.3	0.3	0.3	0.35	0.36	0.367	0.371	0.375	0.378	0.36
DTM: CNN	0.5	0.35	0.367	0.375	0.36	0.317	0.271	0.263	0.233	0.24
DTM: CNN_LSTM	0.7	0.6	**0.667**	0.6	0.56	0.55	0.514	0.5	0.467	0.43
DTM: LSTM_CNN	0.3	0.45	0.467	0.475	0.42	0.4	0.357	0.35	0.344	0.34
FTM	0.8	**0.7**	**0.667**	**0.65**	**0.62**	**0.633**	**0.571**	**0.563**	**0.522**	**0.51**
TAM	**0.9**	0.65	0.6	0.575	0.56	0.533	0.486	0.438	0.444	0.41

The predictive performances of the baseline models can be ranked in the following order, from highest to lowest: *AR*, *DE*, *SE*, *JI*, *PA* and *CN*. The superiority of *AR* to *DE*, and *DE* to *SE* is consistent with results from the previous setting (Chapter 5). The inheritance of the following strong points: *multi-characteristic nature*, *use of both local and global features* and *LBD-tailored heuristics* may have caused *AR* to outperform the other baseline models. The inclusion of dynamic semantic inferences through *shallow temporal cues* may have caused *DE* to perform second-highest relative to the baseline models, not only in the previous setting (Chapter 5), but also in this setting. The other semantic baseline (*SE*) exhibited the third-highest predictive performance over the other baseline models. The inclusion of vector semantics may have caused *SE* to showcase performance to this level. The performance differences of the two semantic baselines *DE* and *SE* indicate the need for *dynamic vector semantics* over *static vector semantics*. Overall, the three link prediction baselines, *CN*, *JI* and *PA*, performed poorly across all k values indicating that the LBD process favours techniques tailored to LBD type problems rather than to the direct usage of *conventional measures*. A similar conclusion is reached in Chapter 5.

Next, this study evaluated the predictive performances of the three highest-performing proposed LBD models (*FTM*, *DTM: LSTM_1* and *TAM*) in comparison to the three competitive baselines: *AR*, *DE* and *SE*. It is evident that *FTM* initially demonstrated a 30.2% performance increase over *AR*. The average performance increase of *FTM* over *AR* was 19.88%. Relative to the two semantic baselines (*DE* and *SE*), *FTM* demonstrated average performance increase of 27.33% and 30.89%, respectively. The proposed DTM variant, *DTM: LSTM_1* displayed the following average performance increases over the baselines: 17.43% (compared to *AR*), 24.64% (compared to *DE*) and 28.44% (compared to *SE*). The third highest-performing proposed LBD model, TAM displayed the following average performance increases 12.04%, 19.24%, 23.05% over the baselines, *AR*, *DE* and *SE*, respectively. Overall, the prediction increases indicate that the proposed LBD models demonstrated significant performance increases over the baselines.

Overall, both the *P@k* and *MAP@k* results indicate that the proposed LBD models not only detected novel knowledge linkages with high precision (i.e., the *P@k* results), but also demonstrated a better ordering of new knowledge (i.e., the *MAP@k* results). However, despite these promising results, this study aimed to verify the *consistency* of the predictive performances through the use of the *Geometric Mean Average Precision*

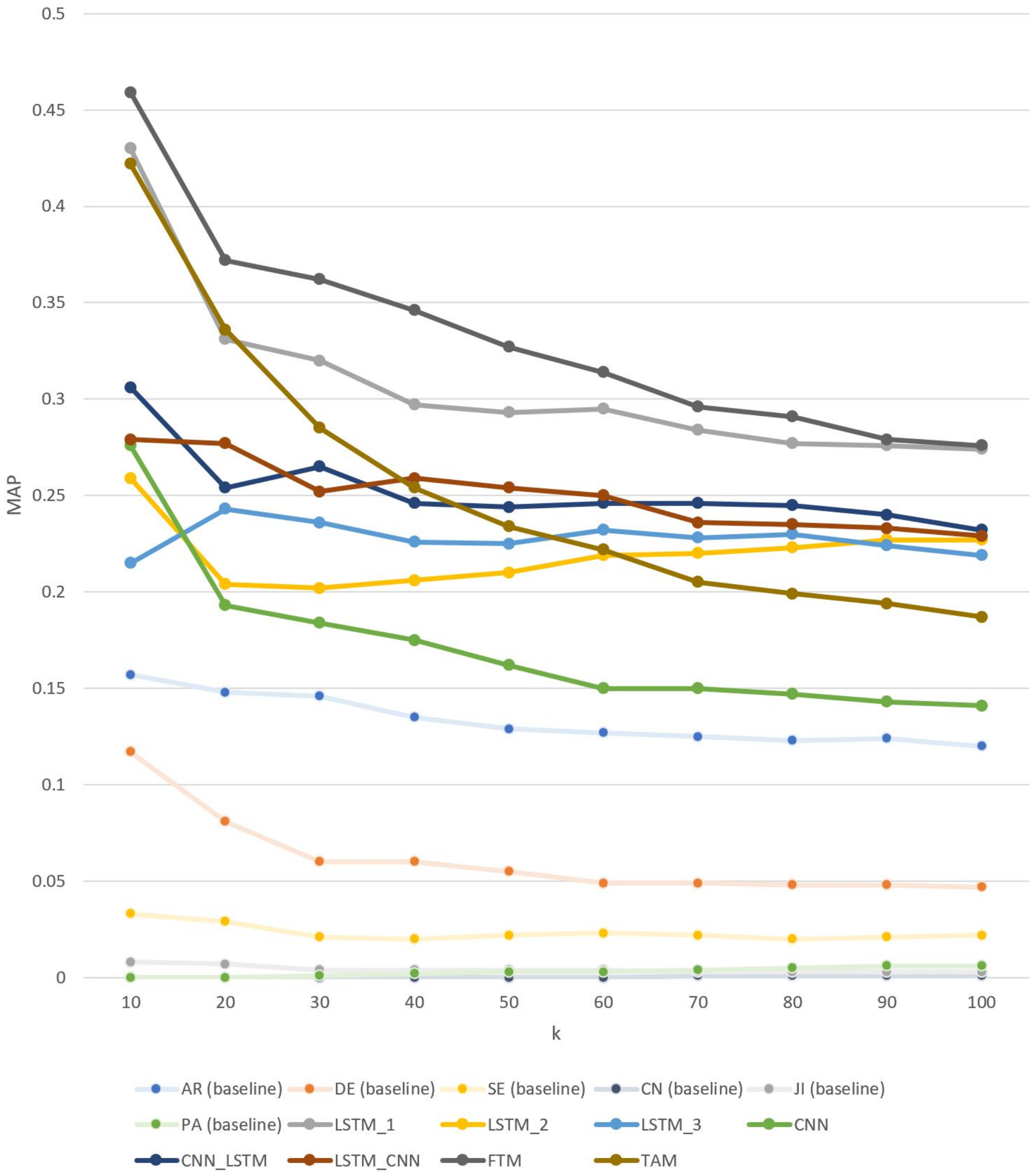


FIGURE 6.3: MAP@k results for the four golden test cases

at k (GMAP@k) evaluation metric. More specifically, *GMAP* penalises LBD models with unstable predictive performances in the test cases, as discussed in Chapter 3. As in the previous setups, the k value was increased by increments of 10, beginning at 10 and going up to 100. Figure 6.4 denotes the predictive performances of the LBD models in terms of *GMAP*. The three highest-performing proposed LBD models: *FTM*, *DTM: LSTM_1* and *TAM* also displayed the highest *GMAP* performances. This shows not only that they had the highest *overall* predictive performances, but also the highest *consistent* predictive performances. It is interesting to observe that the other DTM variants (i.e., all except *DTM: LSTM_1*) were penalised, especially when $k = 10$, due to

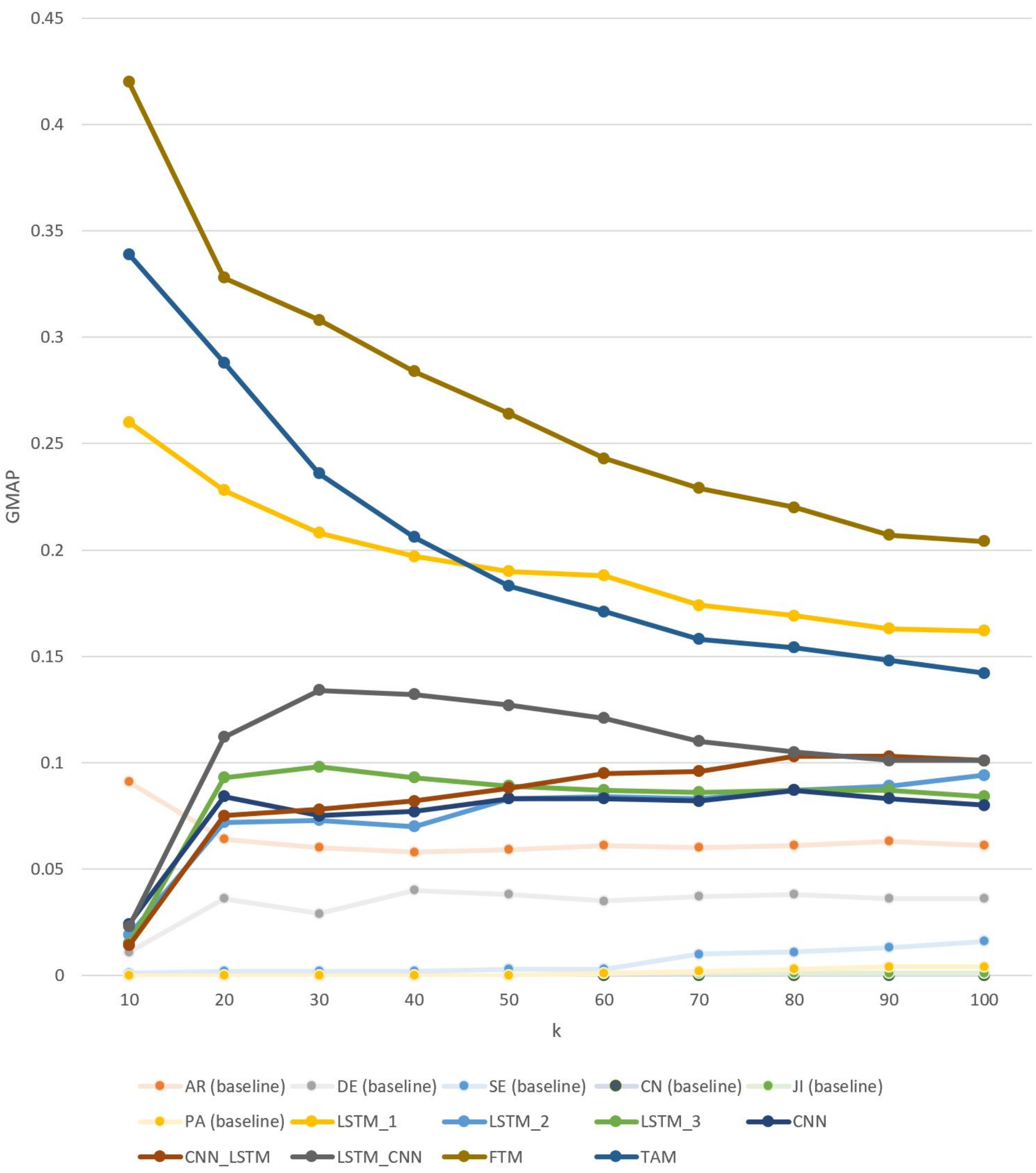


FIGURE 6.4: GMAP@k results for the four golden test cases

their unstable predictive performances in test cases such as *FO-RD*. Nevertheless, these DTM variants displayed better *GMAP* performances for the ensuing *k* values relative to the most competitive baseline *AR*, indicating the potential contributions of feature learning using deep learning models in the LBD workflow (that are worth exploring and expanding in the future LBD models).

In the next experimental setting, the search for novel knowledge was limited to *drugs* by retaining the CTD chemicals that had a corresponding mapping to the drugs in *Drug-Bank* (Yang, Zhao, Waxman & Zhao 2019). Thus, this setting is relatively similar to *drug repurposing*, where the idea is to propose existing drugs that may have potential relationships to a user-specified disease. Tables 6.6, 6.7, 6.8, and 6.9 present the *P@k* results for golden test cases *RD*, *MIG*, *AD* and *SZ*, respectively. As in the previous setups, *FTM*, *DTM: LSTM_1* and *TAM* often demonstrated the highest predictive performances in every test case across all *k* values. Figure 6.5 illustrates the *MAP@k* results of the golden test cases (i.e., *RD*, *MIG*, *AD* and *SZ*) (also reported in Table B.2).

Table 6.6: P@k results for FO-RD test case using only drugs, where topic A is RD

Method	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**
AR (baseline)	0.0	0.0	0.033	0.075	0.12	0.117	0.1	0.088	0.089	0.08
DE (baseline)	0.1	0.1	0.067	0.15	0.2	0.167	0.143	0.125	0.122	0.13
SE (baseline)	0.1	0.15	0.1	0.15	0.2	0.183	0.157	0.175	0.178	0.17
CN (baseline)	0.0	0.0	0.0	0.05	0.08	0.1	0.086	0.088	0.1	0.11
JI (baseline)	0.2	0.1	0.067	0.05	0.04	0.033	0.043	0.05	0.056	0.05
PA (baseline)	0.0	0.05	0.1	0.075	0.1	0.083	0.1	0.1	0.1	0.11
DTM: LSTM_1	**0.4**	**0.4**	**0.3**	**0.25**	**0.22**	**0.2**	**0.214**	**0.188**	**0.2**	**0.2**
DTM: LSTM_2	0.1	0.1	0.1	0.075	0.08	0.083	0.086	0.1	0.1	0.1
DTM: LSTM_3	0.1	0.15	0.1	0.1	0.1	0.083	0.1	0.1	0.1	0.09
DTM: CNN	0.1	0.05	0.067	0.075	0.14	0.133	0.129	0.125	0.122	0.14
DTM: CNN_LSTM	0.1	0.1	0.067	0.05	0.08	0.067	0.071	0.088	0.1	0.11
DTM: LSTM_CNN	0.0	0.05	0.133	0.1	0.08	0.083	0.086	0.1	0.089	0.08
FTM	**0.4**	0.25	0.233	0.175	0.18	0.15	0.129	0.125	0.122	0.14
TAM	0.2	0.2	0.167	0.125	0.14	0.15	0.143	0.138	0.133	0.12

Table 6.7: P@k results for MG-MIG test case using only drugs, where topic A is MIG

Method	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**
AR (baseline)	0.1	0.2	0.233	0.2	0.16	0.167	0.157	0.188	0.189	0.21
DE (baseline)	0.4	0.2	0.167	0.125	0.16	0.183	0.214	0.213	0.211	0.21
SE (baseline)	0.1	0.15	0.167	0.2	0.18	0.183	0.186	0.188	0.167	0.17
CN (baseline)	0.0	0.0	0.033	0.05	0.06	0.067	0.071	0.088	0.111	0.11
JI (baseline)	0.0	0.0	0.0	0.05	0.06	0.05	0.057	0.088	0.1	0.1
PA (baseline)	0.0	0.1	0.067	0.1	0.1	0.117	0.129	0.125	0.156	0.17
DTM: LSTM_1	0.4	0.5	0.433	0.425	0.44	0.4	0.4	0.375	0.356	0.35
DTM: LSTM_2	0.2	0.3	0.3	0.325	0.36	0.333	0.371	0.363	0.333	0.35
DTM: LSTM_3	0.2	0.2	0.267	0.25	0.24	0.25	0.243	0.225	0.233	0.25
DTM: CNN	0.5	0.4	0.4	0.375	0.36	0.333	0.314	0.363	0.356	0.34
DTM: CNN_LSTM	0.3	0.3	0.4	0.375	0.38	0.367	0.386	0.35	0.344	0.37
DTM: LSTM_CNN	0.4	0.35	0.367	0.35	0.34	0.333	0.343	0.35	0.333	0.34
FTM	**0.7**	**0.65**	**0.5**	**0.475**	**0.46**	**0.45**	**0.429**	**0.438**	**0.411**	**0.39**
TAM	0.6	0.45	0.367	0.35	0.34	0.367	0.371	0.35	0.333	0.33

Table 6.8: P@k results for AD-INN test case using only drugs, where topic A is AD

Method	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**
AR (baseline)	0.5	0.45	0.4	0.4	0.4	0.417	0.429	0.413	0.411	0.41
DE (baseline)	0.2	0.25	0.233	0.225	0.24	0.25	0.257	0.238	0.233	0.24
SE (baseline)	0.0	0.0	0.033	0.075	0.08	0.083	0.1	0.113	0.144	0.15
CN (baseline)	0.0	0.0	0.0	0.0	0.02	0.017	0.014	0.025	0.044	0.05
JI (baseline)	0.0	0.0	0.033	0.05	0.04	0.033	0.043	0.038	0.067	0.07
PA (baseline)	0.0	0.0	0.0	0.0	0.0	0.017	0.014	0.038	0.044	0.05
DTM: LSTM_1	0.7	0.65	0.667	0.7	0.68	0.717	**0.743**	0.713	0.711	0.7
DTM: LSTM_2	0.7	0.65	**0.733**	**0.75**	**0.74**	**0.733**	0.729	**0.75**	**0.744**	**0.73**
DTM: LSTM_3	**0.8**	**0.75**	0.667	0.7	0.7	0.667	0.657	0.65	0.622	0.64
DTM: CNN	0.7	0.6	0.533	0.5	0.5	0.467	0.471	0.513	0.5	0.5
DTM: CNN_LSTM	0.6	0.55	0.633	0.65	0.62	0.6	0.614	0.625	0.622	0.62
DTM: LSTM_CNN	0.5	0.65	**0.733**	0.7	0.68	0.7	0.7	0.675	0.644	0.65
FTM	**0.8**	0.65	0.7	0.65	0.62	0.6	0.586	0.588	0.556	0.58
TAM	0.5	0.45	0.433	0.45	0.46	0.45	0.457	0.488	0.5	0.5

Table 6.9: P@k results for SZ-PA2 test case using only drugs, where topic A is SZ

Method	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**
AR (baseline)	0.1	0.15	0.167	0.225	0.26	0.283	0.257	0.238	0.222	0.2
DE (baseline)	0.1	0.05	0.133	0.15	0.12	0.117	0.114	0.125	0.122	0.13
SE (baseline)	0.0	0.0	0.0	0.0	0.04	0.05	0.071	0.1	0.089	0.09
CN (baseline)	0.0	0.0	0.0	0.0	0.0	0.017	0.014	0.013	0.022	0.03
JI (baseline)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.011	0.03
PA (baseline)	0.0	0.0	0.0	0.05	0.04	0.033	0.043	0.038	0.033	0.04
DTM: LSTM_1	0.5	0.4	0.367	0.35	0.38	0.4	0.4	0.375	0.367	0.35
DTM: LSTM_2	0.3	0.25	0.233	0.2	0.24	0.267	0.271	0.313	0.322	0.32
DTM: LSTM_3	0.1	0.2	0.267	0.25	0.28	0.3	0.271	0.275	0.267	0.27
DTM: CNN	0.2	0.25	0.267	0.225	0.18	0.167	0.171	0.213	0.211	0.23
DTM: CNN_LSTM	**0.7**	0.6	0.567	0.5	0.44	0.4	0.357	0.375	0.378	0.4
DTM: LSTM_CNN	0.3	0.35	0.4	0.375	0.34	0.367	0.343	0.35	0.344	0.33
FTM	0.6	**0.65**	**0.633**	**0.625**	**0.6**	**0.55**	**0.529**	**0.5**	**0.467**	**0.43**
TAM	0.6	0.5	0.567	0.525	0.44	0.4	0.4	0.4	0.4	0.39

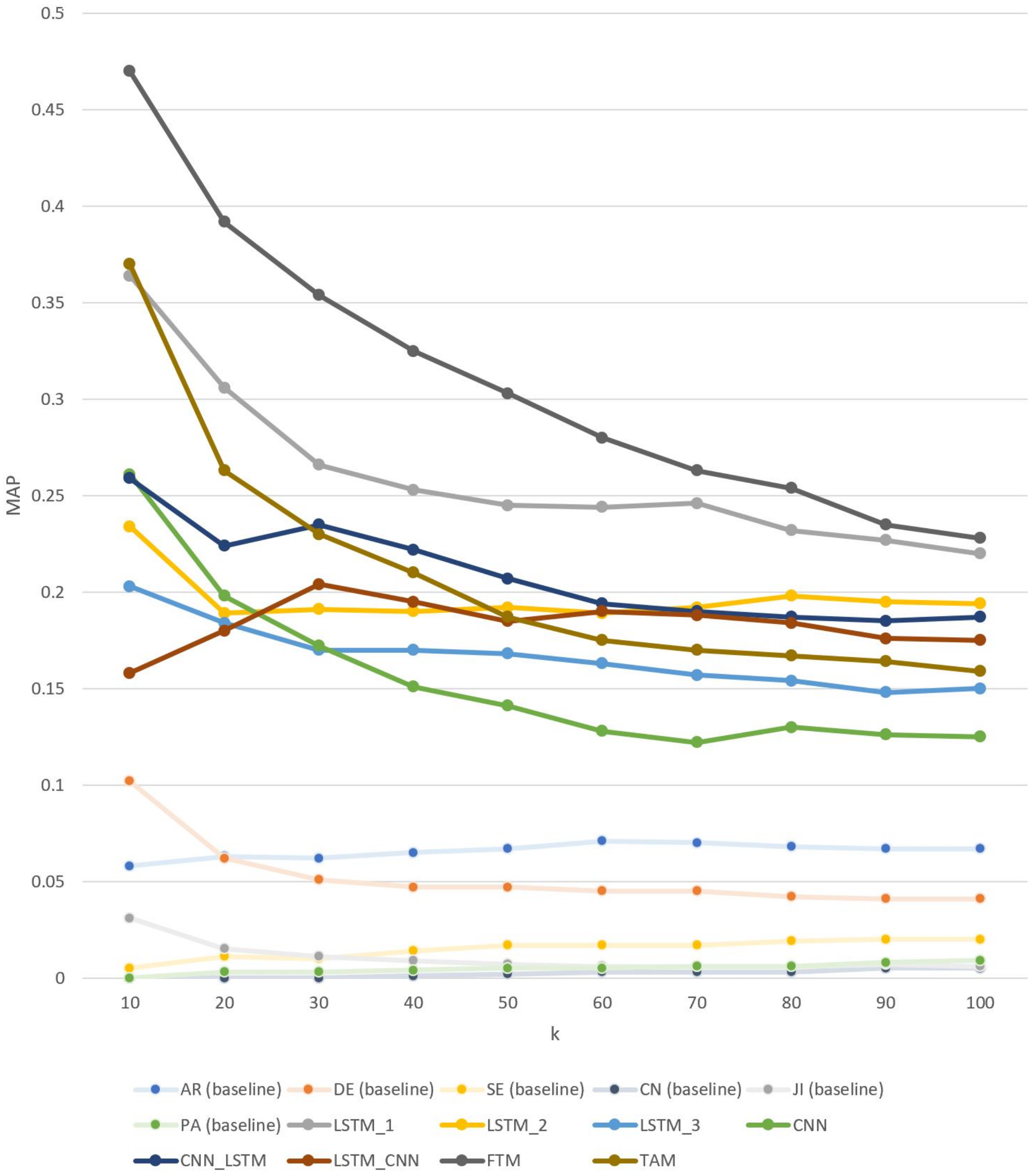


FIGURE 6.5: MAP@k results for the four golden test cases using only drugs

The results obtained through the use of *MAP* (Figure 6.5) are consistent with observations from the previous settings, in which *FTM*, *DTM: LSTM_1* and *TAM* demonstrated the highest predictive performances. More specifically, *FTM* demonstrated a 41.2% performance increase over the most competitive baseline *AR* at the outset of the *k* values. The average performance increases of FTM, compared to the three most competitive baselines: *AR*, *DE*, *SE* were 24.46%, 25.81% and 29.54%, respectively. The second-highest performing model, *DTM: LSTM_1* exhibited average performance increases of 19.45% (compared to *AR*), 20.8% (compared to *DE*) and 24.53% (compared to *SE*). The proposed LBD model *TAM* displayed average performance increases of 14.37%, 15.72% and 19.45% compared to the three baseline models *AR*, *DE* and *SE*, respectively. It is also noteworthy that all the proposed LBD models showcased significant performance increases (Figure 6.5), providing further support for their potential reuse capabilities, even in the very first iteration of the grab-and-glue framework. As in the previous setup, link prediction baseline models showcased poor predictive performances, further substantiating the need for LBD-tailored measures rather than the direct use of *conventional measures*. The main reason for the poor performances of such conventional measures could be the complexity of the problem that LBD attempts to address, which requires more comprehensive and detailed semantic deductions.

Subsequently, *GMAP@k* was used to verify the *consistency* of the predictive performances in this experimental setup. Figure 6.6 depicts the results obtained using *GMAP*. The *GMAP* results also verify the robust predictive performances of the three LBD models, *FTM*, *DTM: LSTM_1* and *TAM*. Except for *DTM: LSTM_CNN* at the *k* value of 10, all the other proposed LBD models demonstrated consistent predictive performances in comparison to the baselines. This further supports the potential reusability of the proposed LBD models.

In spite of the promising results observed, it was necessary for this study to further verify the potential reusability of the proposed LBD models in an extended setup. Within an extended setup, a user is interested in exploring potential novel knowledge linkages greater than 100 (i.e., in the *long run*). To model this situation, the predictive performances of the LBD models are observed until *k* is equal to 250. The *MAP* and *GMAP* results of this extended experimental setup are illustrated in Figures 6.7 and 6.8, respectively. When observing the results in Figures 6.7 and 6.8, it is clear that all the proposed LBD models outperformed the baseline models, not only in terms of overall predictive performance, but also in the consistency of their predictive performance across the test cases even in the *long run*. This provides evidence for two important things: firstly, the efficient reusability of the proposed LBD models, and secondly, the power of diachronic semantic inferences to aid LBD models in discovering implicit linkages in the knowledge discovery process of the LBD workflow.

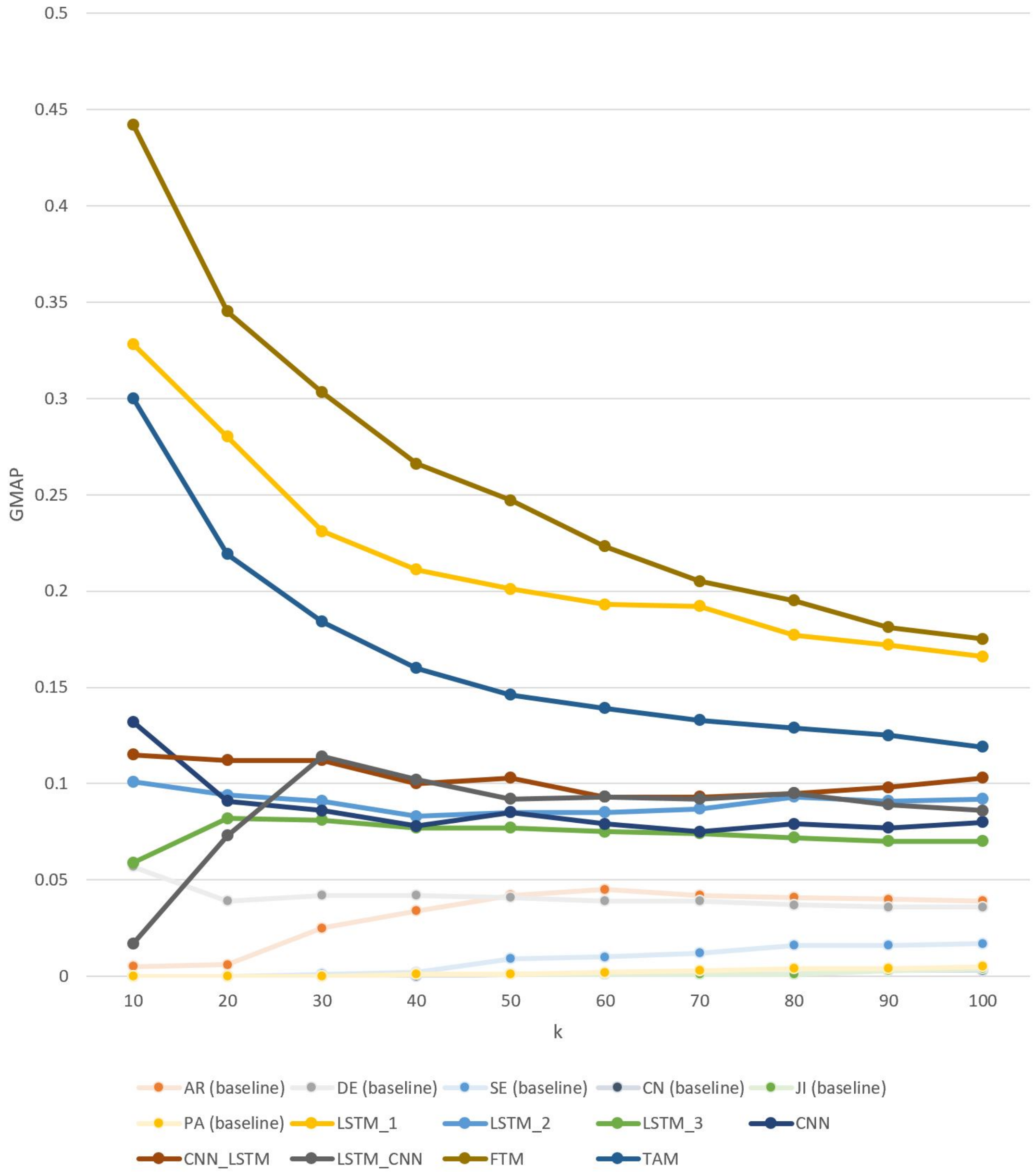


FIGURE 6.6: GMAP@k results for the four golden test cases using only drugs

6.5.1 Strengths of the Proposed LBD models

As in the previous setting (discussed in Section 5.10.3 of Chapter 5), the following eight strengths may explain why the proposed LBD models demonstrated robust predictive performances in all experimental setups in this reuse setting (even in the very first iteration of the grab-and-glue framework). These strengths are as follows: *integration of global semantics*, *intermingling of vector semantics with temporal dimension*, *disentangling of semantic shifts from different perspectives*, *nuanced temporal analysis*, *use of multiple characteristics*, *integration of machine learning and deep learning techniques*,

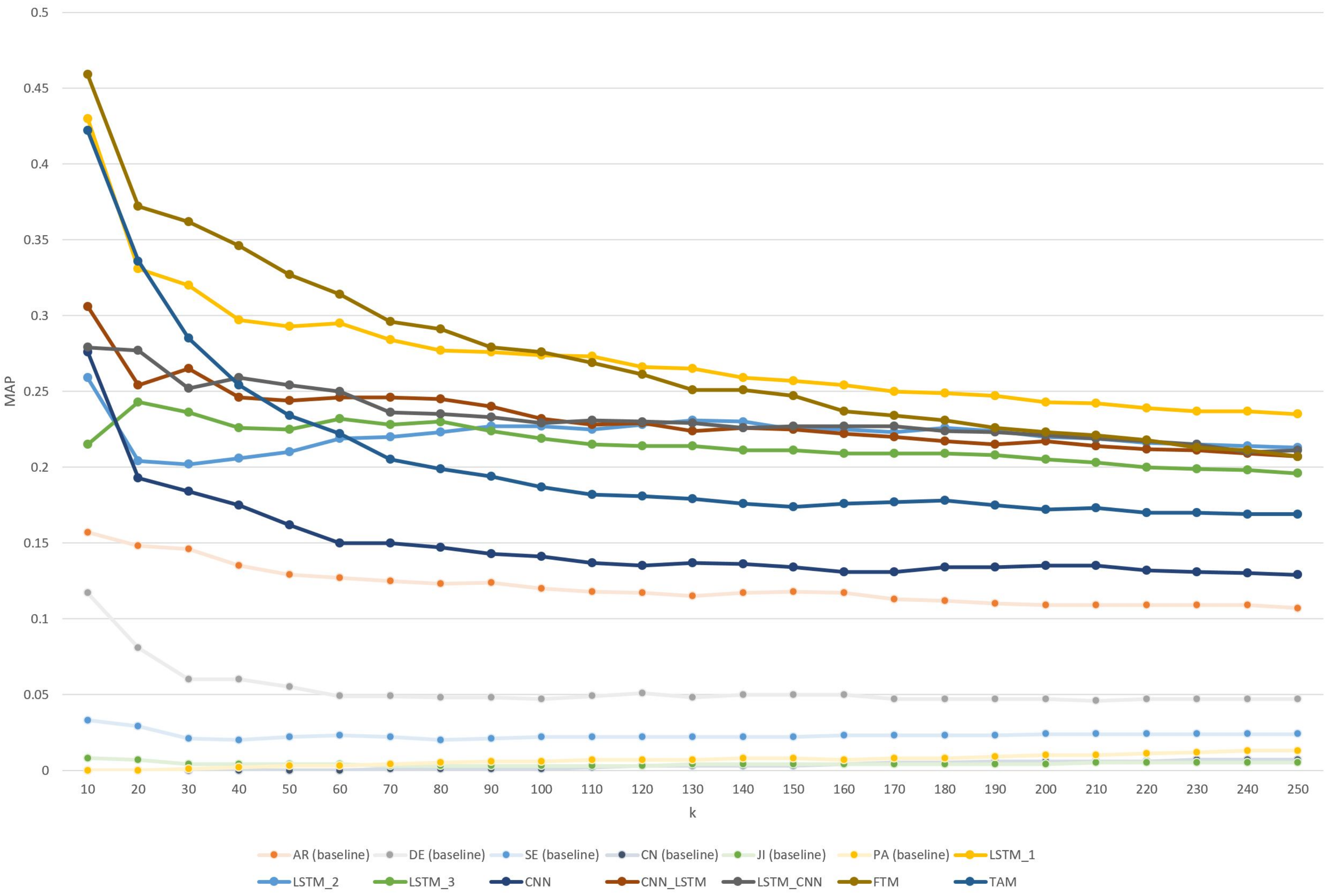


Figure 6.7: MAP@k results for the four golden test cases in the long run

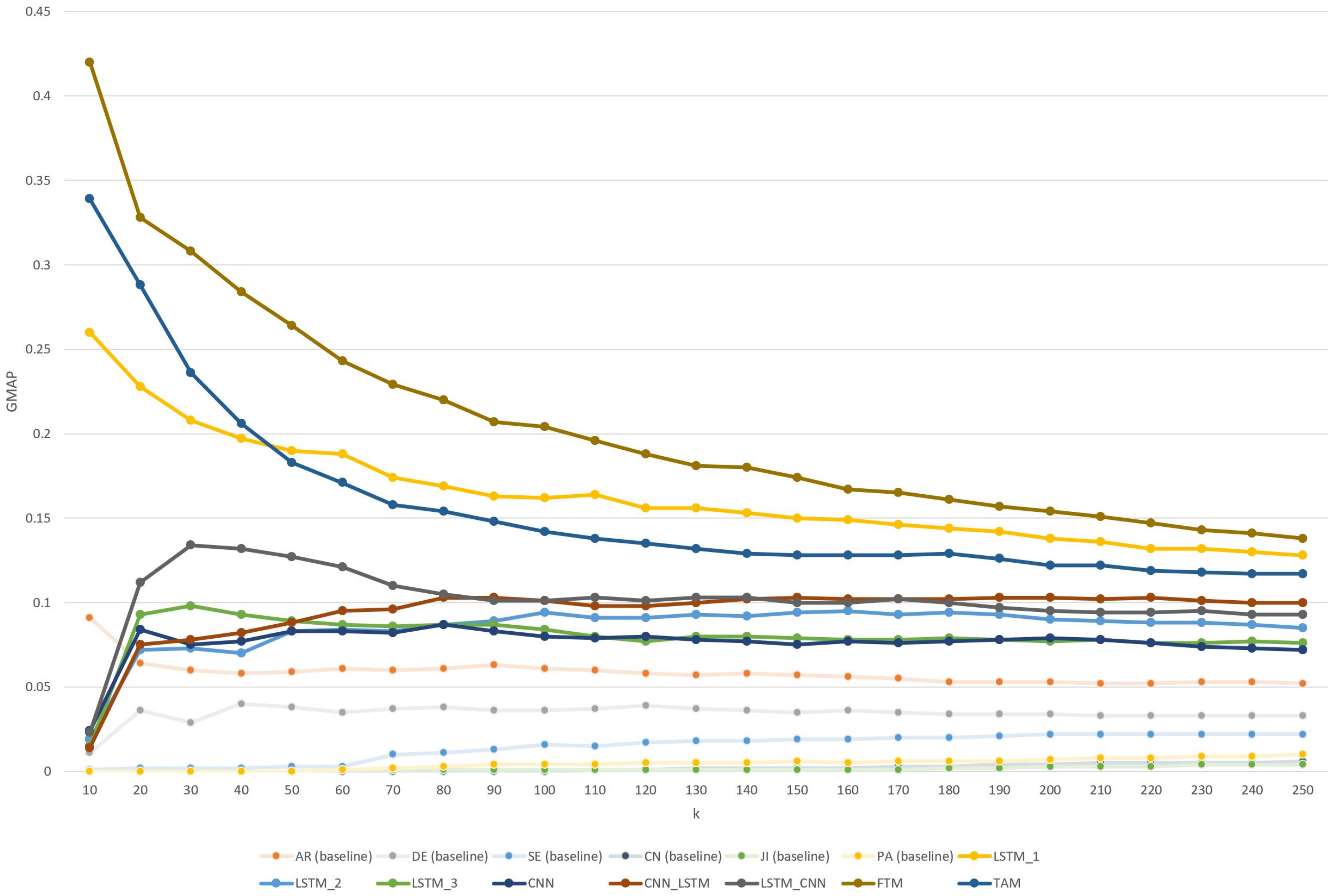


Figure 6.8: GMAP@k results for the four golden test cases in the long run

domain independency, and *potential direct and indirect uses of diachronic semantic inferences*.

- *Integration of global semantics:* The diachronic semantic inferences used as the core analysis unit in the proposed LBD models enclose global scale semantics. The use of a global picture of topic interactions enables models to perform comprehensive, detailed semantic deductions by analysing semantic relationships between scientific topics, with a wide scope.

- *Intermingling of vector semantics with temporal dimension:* The proposed LBD models are sensitive to the temporal semantics of scientific topics, which opens up a different dimension and allows them to detect signals of potential novel knowledge linkages. In other words, the proposed LBD models take advantage of cues that are invisible to traditional LBD models (which merely focus on *static literature analysis*) through co-modelling *vector semantics* with the *temporal dimension* of the scientific topics.

- *Disentangling of semantic shifts from different perspectives:* This thesis disentangled the diachronic semantic inferences in three broader perspectives, namely *individual*, *pairwise* and *neighbourhood*. For this reason, the proposed LBD models were able to capture the semantically infused temporal trajectories of scientific topics in different viewpoints, ultimately enriching the temporal signals on which they were based.

- *Nuanced temporal analysis:* This study integrated a circumstantial temporal analysis component, using a wide range of techniques from *sequence mining*, *time series analysis* and *signal processing* to scrutinise the derived *semantically infused temporal trajectories*. These trajectories reflect the way in which scientific topics evolved in latent embedding spaces across time. The integration of a circumstantial temporal component enabled the proposed LBD models to elicit semantically infused temporal cues in greater detail.

- *Use of multiple characteristics:* The proposed LBD models utilise multiple characteristics to elicit novel knowledge linkages. The idea of integrating multiple characteristics in the knowledge discovery is supported for the following two reasons. Firstly, it enables the precise identification of local topics that could form potential novel knowledge linkages. It does so by verifying the extent to which the local topics fulfil the characteristics of novel knowledge linkages through large-scale feature analysis. Secondly, it has been reported that novel knowledge could exist in different forms; thus, the use of multiple characteristics may provide a platform to discover novel knowledge linkages in different forms with increased coverage.

- *Integration of machine learning and deep learning techniques:* This thesis integrated machine learning as well as recent advancements in deep learning techniques to detect intricate structures in data. This is particularly important for complex reasoning tasks like LBD, since the use of such techniques unravels complex patterns in data, facilitating better predictive decisions.

- *Domain independency:* The proposed LBD models are completely free from knowledge inferences made using external knowledge resources to ensure their widespread applicability. More specifically, the robust predictive performances of the proposed LBD models do not rely on domain-dependent semantic inferences. As such, they could be broadly applicable, independent of domain or problem setting to provide broader community benefits.

- *Potential direct and indirect uses of diachronic semantic inferences:* The proposed LBD models showcase two different perspectives of the diachronic semantic inferences: firstly, *direct usage*, in which semantically infused temporal trajectories are used directly to elicit potential semantically infused temporal signals, and secondly, *indirect usage*, where the idea is to use the semantically infused temporal trajectories as a medium to facilitate knowledge discovery. As with the previous setup discussed in Chapter 5, this reuse setting provides evidence for the potential positive influence of both the direct and indirect uses of the diachronic semantic inferences on the discovery of novel knowledge linkages.

6.5.2 Potential Bottlenecks

One of the key benefits of verifying the reusability of LBD models in new application areas is that it provides better insights into potential bottlenecks. These insights can be used as a guide to further enhance LBD models. Due to the complexity of the problem that LBD attempts to solve, it is difficult (or perhaps even impossible) to produce universal LBD models that perfectly predict potential novel knowledge linkages in every possible setting. Therefore, identifying potential bottlenecks through reuse research helps to establish an extended platform from which to elicit precise future enhancements to each of the new reuse settings.

The main bottleneck that this study observed from the results obtained in the first iteration of the grab-and-glue framework was the loss of performance of the proposed LBD models in this new reuse setting, relative to the results observed in Chapter 5. One main reason for this decline in performance could be the unavailability of the two semantic shifts (*pairwise distance proximity* and *neighbourhood distance proximity*) in this new reuse setting, due to their incompatibility. Otherwise stated, this setting only employed *four semantically infused temporal trajectories* to discover novel knowledge linkages, while the setting in Chapter 5 employed *six temporal trajectories*. This leads to the following question: *'does the number of meaningful diachronic semantic inferences (i.e., the number of semantically infused temporal trajectories derived through semantic shifts for each local topic) integrated into the knowledge discovery process positively correlate with the predictive performance?'*.

To seek out answers to this question, this thesis analyses the performance impact of different combinations of semantically infused temporal trajectories in the knowledge discovery process. More specifically, the four main *trajectory combination types* summarised in Table 6.10 are considered for this analysis.

TABLE 6.10: Trajectory combination types used to analyse their performance impact

Combination Type	**Trajectory Combinations**	**Total Combinations**
1 trajectory	IGS, ILS, PSD, NSD	4
2 trajectories	IGS+ILS, IGS+PSD, IGS+NSD, ILS+PSD, ILS+NSD, PSD+NSD	6
3 trajectories	IGS+ILS+PSD, IGS+ILS+NSD, IGS+PSD+NSD, ILS+PSD+NSD	4
4 trajectories	IGS+ILS+PSD+NSD	1

Figure 6.9 presents the *MAP@k* results obtained for the 15 trajectory combinations summarised in Table 6.10. It is evident that trajectory combination types: *1 trajectory*, *2 trajectories* and *3 trajectories* often underperformed the trajectory combination type: *4*

trajectories (i.e., *IGS+ILS+PSD+NSD*). Even though *PSD+NSD* demonstrated a slight performance increase over *4 trajectories* at *MAP@10* and *MAP@20*, its performance decreased swiftly from *MAP30* onwards. Overall, *IGS+ILS+PSD+NSD* displayed the *highest* and most *consistent* predictive performance across the k values. It is also interesting to observe that even the most simplified versions of the proposed LBD model (i.e., the performances of *1 trajectory*, *2 trajectories* and *3 trajectories*) displayed better predictive performances than the baseline models (see Figure 6.9), indicating the strong positive influence of the proposed diachronic semantic inferences in the LBD workflow.

Subsequently, this study analysed the *average predictive performance* of the trajectory combination types outlined in Table 6.10. The results obtained through this analysis are presented in Figure 6.10. Overall, it is evident that at each k value, LBD predictive performance was highest for *4 trajectories*, followed by *3 trajectories*, *2 trajectories* and *1 trajectory*.

In essence, this study observed a strong positive correlation between the *number of temporal trajectories* and the *average predictive performance*. The average Pearson's correlation coefficient was *0.971* across the k values (i.e., *10* to *100*). Therefore, it can be concluded that LBD performance strongly favours the number of meaningful diachronic semantic inferences utilised in the knowledge discovery process. This also verifies a potential reason for the performance loss in this reuse setting compared to that in Chapter 5, which is the unavailability of the two diachronic inferences, *pairwise distance proximity* and *neighbourhood distance proximity*. As discussed at the outset of this section, reuse research provides a valuable opportunity to identify precise future enhancements in each new setting. This is because it is extremely difficult (or sometimes impossible) to develop LBD models that perfectly predict novel knowledge linkages in every possible setting. The bottleneck and its potential causes (observed through each iteration of the grab-and-glue framework) can be fixed in the subsequent iteration, in order to enhance the predictive performance of the LBD workflow.

6.5.3 Considerations for the Second Iteration of the Grab-and-glue Framework

There is a strong positive correlation between the number of meaningful diachronic semantic inferences and LBD predictive performance. Thus, the integration of novel

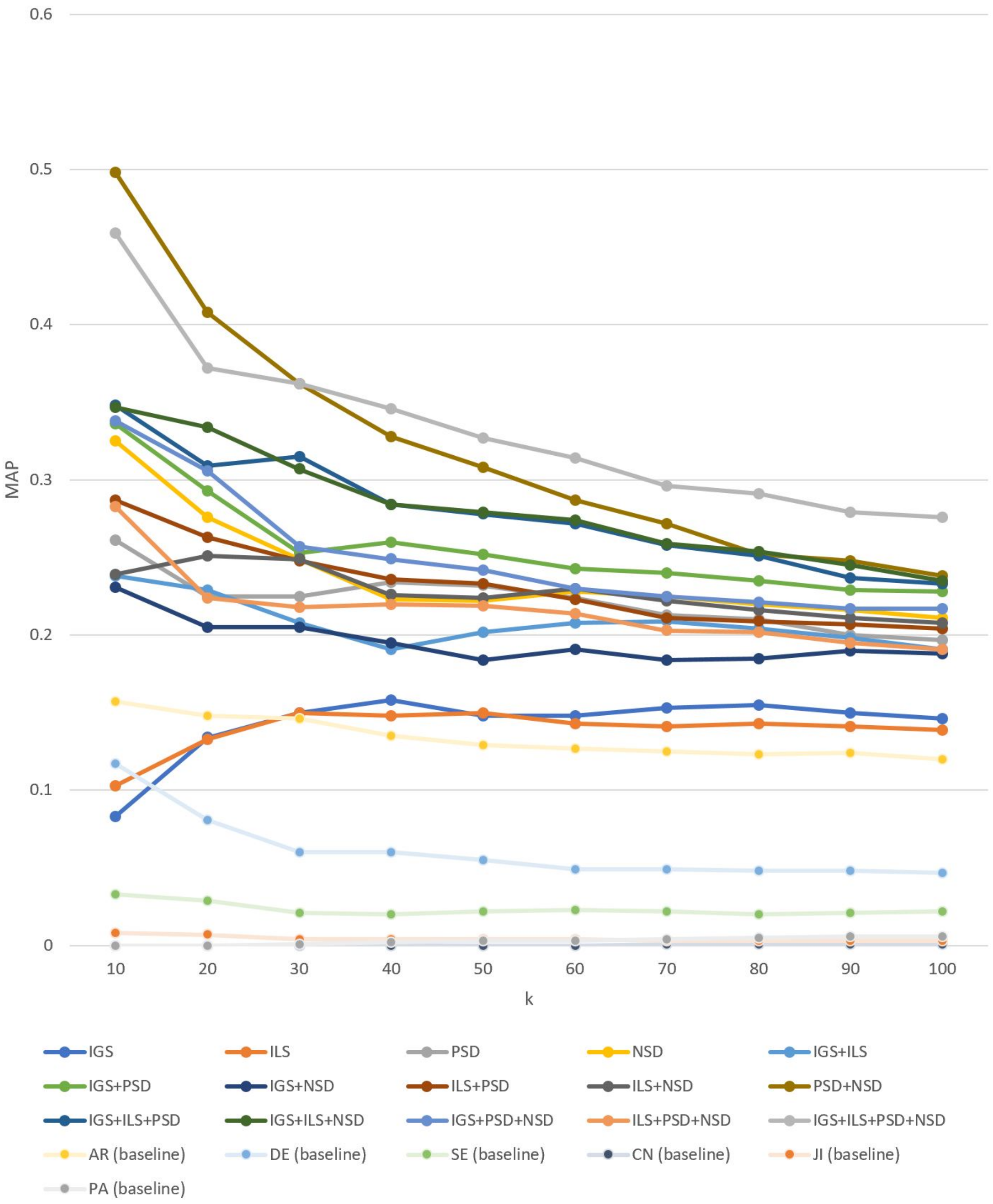


Figure 6.9: LBD predictive performance with every possible combination of the four adapted semantically infused temporal trajectories in this reuse setting

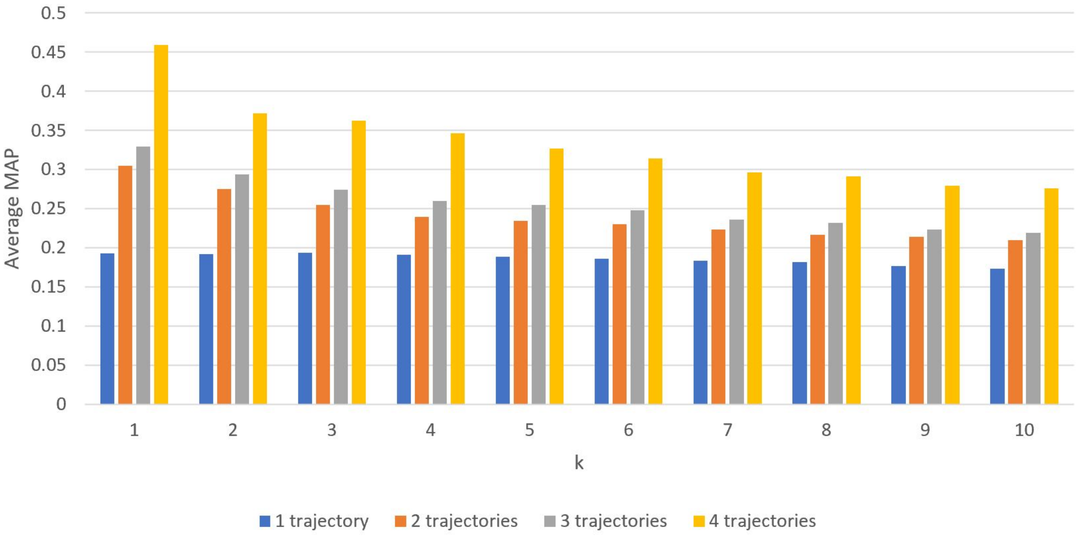


Figure 6.10: Average predictive performance with the number of semantically infused temporal trajectories

meaningful diachronic semantic inferences (i.e., more meaningful *semantically infused temporal trajectories* through semantic shifts) may have a positive impact on LBD predictive performance. Therefore, one goal for the second iteration of the grab-and-glue framework is the integration of more meaningful diachronic semantic inferences in order to make up for the absence of the two semantic shifts used in the previous setting (*pairwise distance proximity* and *neighbourhood distance proximity*). With this in mind, this study proposes semantic shifts (such as those below) in the next iteration of the grab-and-glue framework.

In this new reuse setting, *local topics* are detected through time-sliced analogy mining. Therefore, local topics indicate concepts that *may be* worth exploring with regard to *topic A*. Thus, if these local topics get condensed (i.e., becoming closer to each other) over time, the topics in such condensed clusters may demonstrate potential signals of novel knowledge linkages. For instance, consider a local topic of interest: lt_i, depicted as a blue dot in Figure 6.11. The green dots signify the *local topics* extracted via time-sliced analogy mining, and the grey dots depict the *remaining topics* in the vector spaces. When closely inspecting Figure 6.11, it is evident that at timestamp *t=1*, lt_i has only one other local topic within the r_1 sized radius of its neighbourhood. However, with time, the number of local topics in its neighbourhood increases, resulting in the formation of a slightly condensed cluster. In essence, the semantically infused temporal trajectory in this instance signifies how condensed the neighbourhood of lt_i becomes

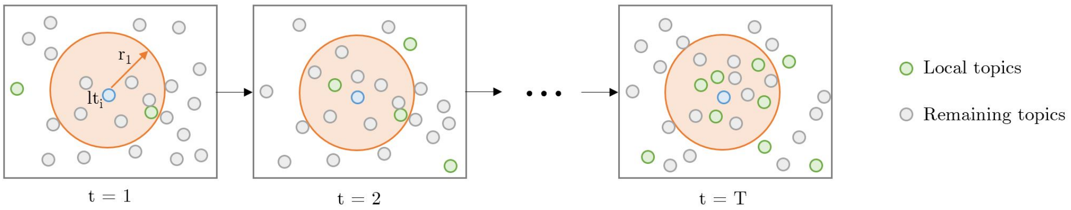


FIGURE 6.11: Formation of condensed clusters of local topics (i.e., green dots) for a local topic lt_i at a radius of r_1 over time

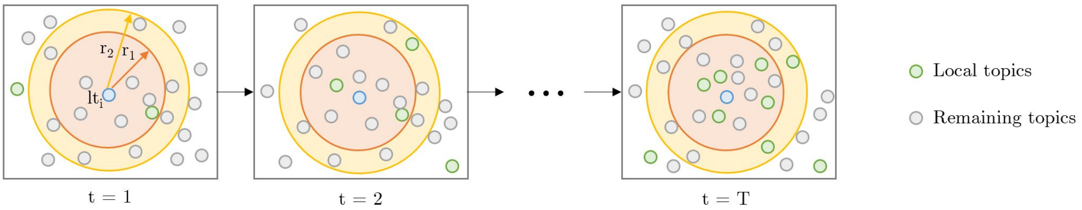


FIGURE 6.12: Formation of condensed clusters of Local topics (i.e., green dots) for a local topic lt_i at different radius values (e.g., r_1 and r_2) over time

with the local topics extracted via analogy mining over time. This idea could be further extended by integrating different neighbourhood radius sizes, as shown in Figure 6.12. Exploiting such meaningful diachronic inferences in the next iteration of the grab-and-glue framework may further increase the prediction results of the proposed LBD models reported in this chapter.

6.5.4 Limitations

This study utilises *time-sliced evaluation* to analyse and compare results due to its unique characteristics such as *automated*, *replicable*, *quantifiable*, *informative* and *modular* that are lacking in other LBD evaluation techniques (Henry 2019, Yetisgen-Yildiz & Pratt 2009). To date, time-slice is the only evaluation technique in the LBD field that is capable of performing large-scale knowledge discovery (Henry 2019). Further details on this selection are discussed in Chapter 3. The reliance on *co-occurrence* in the time-sliced evaluation may introduce noise, since co-occurrence does not necessarily imply a legitimate relationship between two topics. Thus, the use of the time-sliced evaluation only provides an approximate platform for comparing results.

Due to this chapter's focus on reusability, this study employed the *Arrowsmith* features in the *two-node search*, adapted to this new reuse setting (similar to our models and other baselines). Nevertheless, the online version of the *Arrowsmith one-node search* follows a different approach to locate potential *C concepts* (in situations similar to this reuse setting). Therefore, the Arrowsmith results reported as part of this reusability study may not necessarily indicate the performance of the online version of the *Arrowsmith's one-node search.*

6.6 Summary

Through this research into reusability, this thesis attempted to creatively uncover areas in which the proposed LBD models could be applied. The proposed three LBD models demonstrated significant performance increases even in the very first iteration of *grab-and-glue* framework compared to the baseline models not only in terms of the *direct uses* of semantically infused temporal trajectories, but also in terms of their *indirect uses.* Overall, the experimental results demonstrate the potential reusability of the proposed LBD models, which also verifies the power of diachronic semantic inferences with fine-tuned temporal analysis in the LBD workflow. The unavailability of the two semantic shifts (*pairwise distance proximity* and *neighbourhood distance proximity*) in this new reuse setting may have decreased the performance of the proposed LBD models relative to their performance in the previous setting discussed in Chapter 5. This emphasises the need to accommodate further meaningful diachronic semantic inferences in the knowledge discovery workflow, which could be considered as an improvement in the second iteration of the *grab-and-glue* framework.

6.6.1 Major Contributions

Through this reuse research, the thesis could showcase a distinct perspective of the proposed LBD models (i.e., their vertical reusability). To summarise, the major contributions of this chapter are outlined below, and are discussed in detail in Chapter 8.

- Performing large-scale reuse research by integrating considerations of reusability through a methodical reuse plan.

- Demonstrating the vertical reuse of the proposed LBD models considering an opportune application area in the LBD field.

- The proposed LBD models exhibit a greater flexibility in adapting to new reuse settings, due to their domain-agnostic nature and to the power of vector semantics on which they are based.

- Establishing the models' fitness for the intended purpose through the first iteration in the *grab-and-glue* framework, compared to the competitive baselines in the two-node search, as well as state-of-the-art link prediction techniques.

- The trajectory combination types alone also demonstrated high predictive performances compared to baseline models, which verifies the predictive power of the proposed semantically infused temporal trajectories, even when they are used individually.

Chapter 7

Portability

7.1 Introduction

Portability is characterised by the extent to which a model can be applied in *new environments*, at a cost that is lower than the model's redevelopment costs (Mooney 1995, 1997, Ghandorh et al. 2020). Despite several decades of LBD research, most proposed LBD models suffer from a major research deficiency which is lack of *portability* due to their excessive dependency on semantic inferences performed using domain-specific knowledge resources (Kastrin & Hristovski 2020, Hui & Lau 2019, Thilakaratne et al. 2019*b*). Consequently, these LBD models tend only to support knowledge discovery in a single problem setting or domain. To date, LBD research is mostly limited to the *medical domain*, thereby relying on resources that merely support medical data analysis such as *MeSH*, *UMLS*, *SemRep* and *SemMedDB* to perform semantic inferences (Kastrin & Hristovski 2020, Henry & McInnes 2017, Thilakaratne et al. 2019*b*,*a*). It is noteworthy that some of these LBD models are not even generalisable within the medical domain itself, due to their usage of highly specialised knowledge resources that are mostly available for a single or limited problem setting(s) (e.g., *Gene Ontology*, *SIDER* and *PharmGKB*).

The potentiality of LBD framework *outside the medical domain* has been experimented by few LBD studies (Hui & Lau 2019, Sebastian et al. 2017*a*). Despite the promise, these studies have tended to overlook the importance of portability since their models are mostly specific to the selected problem (Hui & Lau 2019). Most of these non-medical models require human intervention or statistical methods due to the unavailability of resources such as *MeSH*, *UMLS*, *SemRep* and *SemMedDB* outside the medical domain (Hui & Lau 2019). For instance, Kostoff et al. (2008) have stated the complexity of processing the text in non-medical domains due to the unavailability of *MeSH*. The recent LBD review by Hui & Lau (2019) have identified that lack of comprehensive ontologies outside the medical domain as an often-cited major challenge that inhibits the adaptation of the LBD workflow to other disciplines.

One of the remedies proposed by Hui & Lau (2019) is the use of domain-specific controlled vocabularies such as the ACM Computing Classification System (ACM CCS)[1] for domains such as computer science. Even though this classification was developed by computer science domain experts, it is comparatively *small-scale* (contains nearly 2,000 subject headings) and getting *updated more slowly* (the latest version was released in 2012) (Han et al. 2020). Moreover, such schemata do not capture concrete, fine-grained concepts and may only be useful for identifying relatively large areas in the computer science domain (Han et al. 2020, Salatino et al. 2020). Other prominent controlled vocabularies (such as the *Physics and Astronomy Classification Scheme (PACS)*[2] (Smith 2019), *Physics Subject Headings (PsySH)*[3] (Smith 2020), *Mathematics Subject Classification (MCS)*[4] (Lange et al. 2012, Dunne & Hulek 2020) and *Journal of Economic Literature (JEL) classification*[5] (Cherrier 2017, Heikkilä 2020)) are also limited due to their *small-scale nature*, as well as *slow and infrequent updates*. Even though classifications such as the *Library of Congress Classification (LCC)*[6] encompass several disciplines and are actively maintained (Chan et al. 2016), such schemata display a lack of breadth and depth which inhibits their capacity to carry out knowledge discovery at an adequate level of granularity. For instance, LCC only uses the three topics: *electronic computers*, *computer science* and *computer software* to characterise the computer science discipline (Salatino et al. 2020). Such examples illustrate the challenges of ensuring the *portability* of the LBD workflow.

[1] https://dl.acm.org/ccs

[2] https://journals.aps.org/PACS - small-scale (9.1K) (Han et al. 2020), latest version is from 2010 and no longer been maintained (Smith 2020, 2019)

[3] https://physh.aps.org/ - small-scale (3.5K) (Han et al. 2020) and latest version is 1.1.1 (Smith 2020)

[4] https://mathscinet.ams.org/mathscinet/msc/msc2020.html - small-scale (6.1k) (Han et al. 2020) and usually gets updated in 10 years (e.g., new MSC 2020 version is the update of its 2010 version) (Salatino et al. 2020)

[5] https://www.aeaweb.org/econlit/jelCodes.php - small-scale (1K) (Heikkilä 2020) and the last major revision was performed in 1990 (Salatino et al. 2020, Kosnik 2018)

[6] https://www.loc.gov/catdir/cpso/lcco/

More recently, Sebastian et al. (2017*b*) have attempted to deviate from the remaining LBD models by integrating *WordNet* (Fellbaum 2012) for the first time in the LBD discipline to propose an LBD model that can be easily applied to various research domains. Even though their study undoubtedly ameliorates the typical LBD setting and provides a different perspective on the LBD discipline, they merely consider WordNet to identify the *synsets* (i.e., *synonyms*). Identification of synonyms may not necessarily be the only domain-dependent impediment in the overall LBD workflow. Thus, it is unclear whether their proposal is capable of helping LBD models in general to achieve greater portability.

Contemplating the benefits of the LBD approach in terms of providing speedier innovation and enhanced research productivity, it is obvious that developing an *interdisciplinary* (or *generalisable*) LBD framework that could be *easily portable* across domains to support *general scientific problem solving* is crucial. Even though the unavailability of such an *interdisciplinary/generalisable LBD framework* remains to be a prolonged open issue in the LBD discipline (Hui & Lau 2019), to the best of our knowledge, no previous LBD studies have explicitly attempted to alleviate this issue. Motivated by the enormous potential unravels through circumventing this prolonged research deficiency, this thesis aims to propose a cost-efficient as well as easily pluggable *portable LBD framework*. To accomplish this goal, this study considers the revolutionary opportunities offered through the *Semantic Web* (more specifically, using *Linked Open Data (LOD)*). To the best of our knowledge, this is the *first study* to propose a *portable LBD framework* to assist researchers in general scientific problem solving. Our proposal also alleviates one of the top-cited major challenge faced by LBD studies outside the medical domain, which is the unavailability of comprehensive ontologies in other disciplines (Hui & Lau 2019).

The main research objective of this portability research is:

"to demonstrate the portability of the LBD workflow by proposing an interdisciplinary (or generalisable) LBD framework to assist scientific problem solving in a domain-agnostic manner"

as defined at the outset of this thesis (i.e., *main objective 5 (RO5)* in Chapter 1). With this objective in mind, this chapter attempts to answer the following main research question (*RQ5*): *'how can an interdisciplinary (or generalisable) LBD framework be* *developed in a way that ensures the portability of the LBD workflow to new portable environments with little or no cost?'*. To this end, this study is divided into several sub-components with the ultimate aim of putting forward the first steps towards this research direction in the LBD discipline by considering the following four sub research objectives.

- *RO5.1. Identifying the impediments in existing LBD models that restrict their applicability to certain problems/domains* (discussed in Section 7.3).

- *RO5.2. Identifying characteristics that need to be fulfilled in developing a portable LBD system* (discussed in Section 7.4).

- *RO5.3. Identifying potential knowledge sources in Semantic Web that support the identified portable characteristics defined, in relation to the LBD context* (discussed in Section 7.5).

- *RO5.4. Circumventing the identified domain-dependent impediments by performing semantic inferences using the selected knowledge resource, which supports portability in the LBD context* (discussed in Section 7.6).

This chapter is organised as follows. Section 7.2 is dedicated to describing the idea of the *Semantic Web* and how this thesis was inspired to explore this direction to remedy the *portability* problem that exists in the LBD workflow. Section 7.3 describes *domain-dependent impediments* that are common in existing LBD models, which restrain their applicability to other problem areas or domains. Section 7.4 defines *portability in the LBD context* and establishes six characteristics that required to be fulfilled to attain a portable LBD framework. Section 7.5 discusses how the selected knowledge resource in the *LOD cloud* (i.e., *Semantic Web*) fulfil the idea of portability by cross-checking the six characteristics defined in Section 7.4. The intention of Section 7.6 is to describe the proposed remedies to the impediments identified in the existing LBD models. With this regard, this study leverages Semantic Web technologies to perform semantic inferences in the selected knowledge resource (introduced in Section 7.2 and verified in Section 7.5). Section 7.7 evaluates the suitability of the proposals described in Section 7.6 to overcome the prevailing limitations by comparing the proposals with the commonly used domain-specific resource in the LBD literature: *MeSH* (*Medical Subject Headings*). This section also revisits the strengths and weaknesses of the proposed remedies in the form of an extended discussion. Finally, Section 7.8 summarises the main conclusions of this chapter while also outlining its major contributions.

7.2 Semantic Web

Semantic Web enables machines to browse the knowledge distributed across the Web (Bizer et al. 2011). Consider a Wikipedia page that provides *knowledge* to human readers. The knowledge contained in the Wikipedia page is *opaque* from the perspective of the machines, as they 'see' nothing but a presentation markup of the Wikipedia page. The idea of '*Semantic Web*' was developed to allow computers to explore the knowledge on the Web (i.e., to make the Web data *machine-readable*) (Coyle 2012). To realise this goal, it was crucial to have a large amount of Web data in a standard format that could be reached and managed by Semantic Web tools. In essence, to construct such *Web of Data*, Semantic Web not only requires access to the data, but also to the *relationships* among data points (as opposed to a large collection of datasets). Such *interrelated datasets* available on the Web are also known as *Linked Data* that connect and share data through *dereferenceable Uniform Resource Identifiers* (*URIs*) across a wide range of applications (Mirizzi et al. 2010).

Figure 7.1 presents historical landmarks in the evolution process of the *Semantic Web* into *Linked Data* (Méndez & Greenberg 2012). The main aim of Linked Data is to expand the Web by publishing datasets in the form of *RDF* (*Resource Description Framework*) and by setting up links between data from various other data sources. In accordance with this aim, *URIs* are the fundamental units used to identify everything and *RDF* is the fundamental linking structure that utilises URIs to name the relationships between datapoints as well as the two ends of each relationship (Mirizzi et al. 2010). In essence, *RDF* is the *W3C* (*World Wide Web Consortium*) standard for encoding the knowledge contained in the resources in World Wide Web (WWW) (Decker, Melnik, Van Harmelen, Fensel, Klein, Broekstra, Erdmann & Horrocks 2000, Decker, Mitra & Melnik 2000).

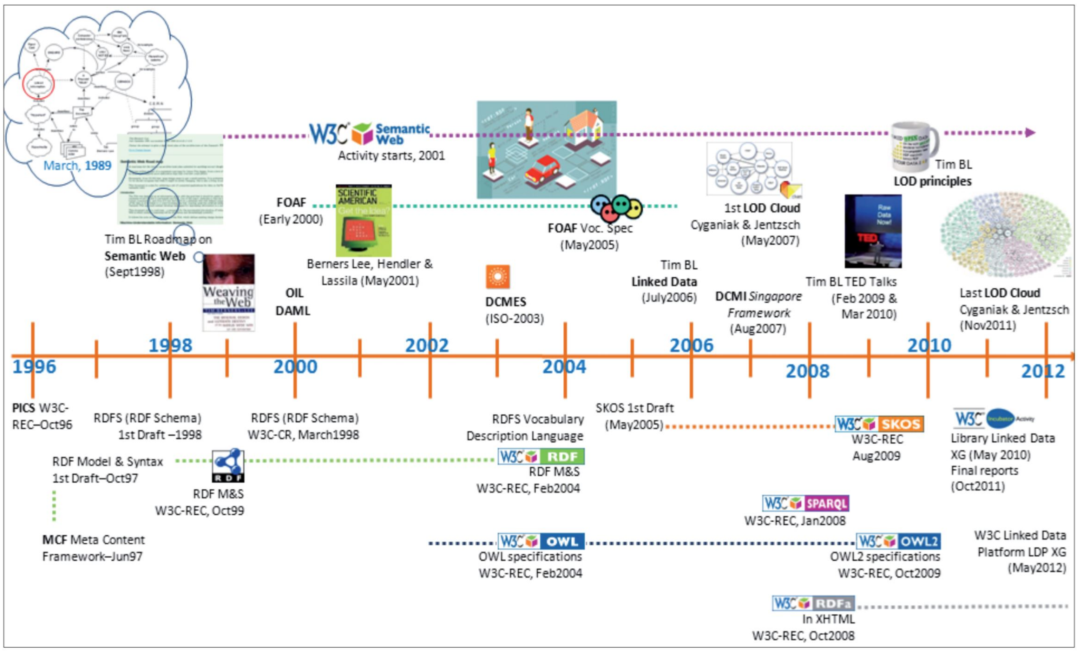


FIGURE 7.1: Historical landmarks in the evolution of Semantic Web into Linked Data (Méndez & Greenberg 2012)

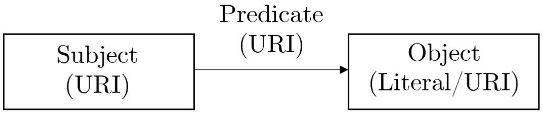


FIGURE 7.2: Structure of an RDF triple

7.2.1 RDF (Resource Description Framework) Triple

The unit structure of an RDF is a *triple*, composed of three items (Decker, Mitra & Melnik 2000). The first item is the *subject*, which represents the *resource*. It is a *URI* reference that uniquely identifies the described resource. The *subject* is followed by the second item, the *predicate*. It represents the relationship described through the RDF triple. Like the *subject*, the *predicate* is also a *URI*. The third item of the RDF triple is the *object*, which can either be a *literal* or a *URI*. It relates to the *subject* via the relationship specified by the *predicate*. Figure 7.2 depicts these three components in an RDF triple[7].

[7] Note that sometimes the *subject* and *object* could be blank nodes.

7.2.2 RDF Graph

When there is a collection of RDF triples, this becomes an *RDF graph* (Carroll et al. 2004). For instance, consider Figure 7.3, which represents a simplified example of an RDF graph. In this graph, the *subject*, *predicates* and *objects* can be defined as summarised below.

- The *subject* of this RDF graph is the resource specified by `<http://www.w3.org/People/EM/contact#me>`, which is a *URI*.
- The *predicates* of this RDF graph are:
  - `<http://www.w3.org/2000/10/swap/pim/contact#fullName>`, which is a *URI* that describes the *'whose name is'* relationship.
  - `<http://www.w3.org/2000/10/swap/pim/contact#mailbox>`, which is a *URI* that describes the *'whose email is'* relationship.
  - `<http://www.w3.org/2000/10/swap/pim/contact#personalTitle>`, which is a *URI* that describes the *'whose title is'* relationship.
  - `<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>`, which is a *URI* that describes the *'subject is a type of'* relationship.
- The *objects* of this RDF graph are:
  - *Eric Miller*, which is a *literal* that describes the name of the person identified by the *subject*.
  - `<mailto:em@w3.org>`, which is a *URI* that describes the email address of the person identified by the *subject*.
  - *Dr.*, which is a *literal* that describes the title of the person identified by the *subject*.
  - `<http://www.w3.org/2000/10/swap/pim/contact#Person>`, which is a *URI* that describes the type of the *subject* as a *Person*.

According to the standards, the RDF triple can be written as *(<subject>, <predicate>, <object>)* (Carroll et al. 2004). Therefore, the example RDF graph illustrated in Figure 7.3 can be denoted in N-triples format, as depicted in Figure 7.4.

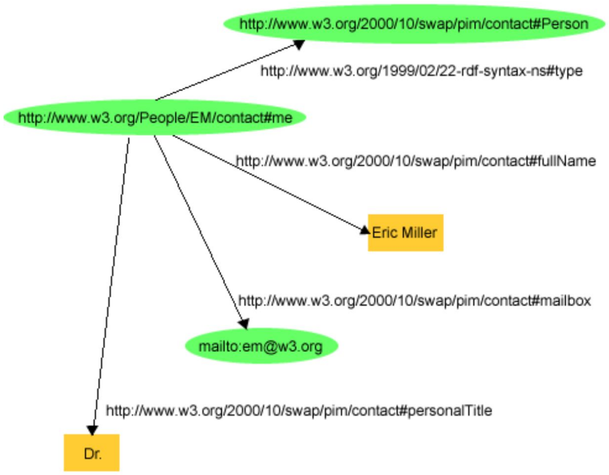

http://www.w3.org/2000/10/swap/pim/contact#Person
http://www.w3.org/1999/02/22-rdf-syntax-ns#type
http://www.w3.org/People/EM/contact#me
http://www.w3.org/2000/10/swap/pim/contact#fullName
Eric Miller
http://www.w3.org/2000/10/swap/pim/contact#mailbox
mailto:em@w3.org
http://www.w3.org/2000/10/swap/pim/contact#personalTitle
Dr.

FIGURE 7.3: Simplified example of an RDF graph
Source: https://www.w3.org/TR/rdf-primer/fig1dec16.png

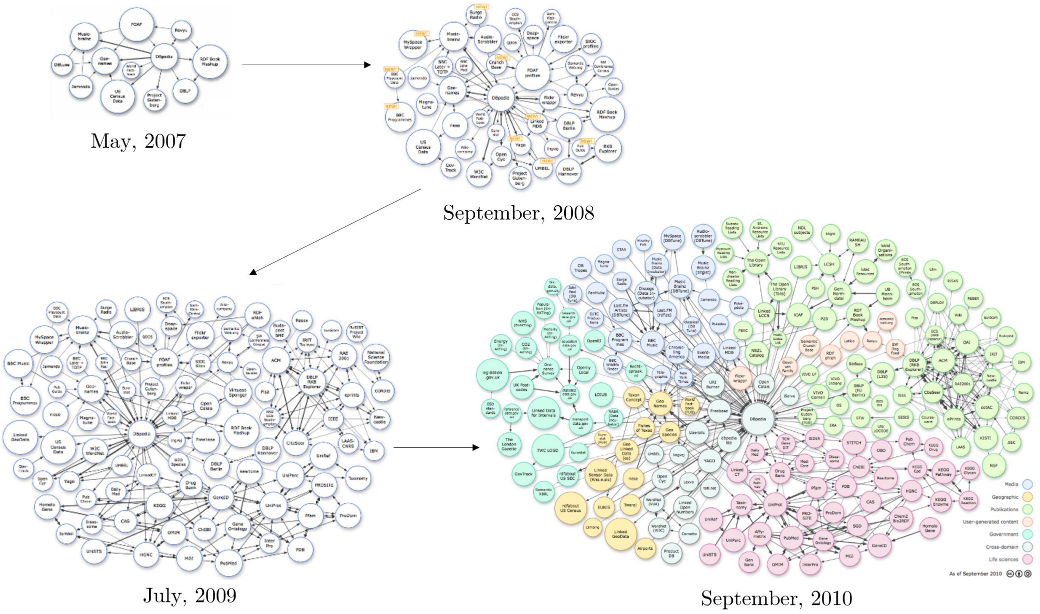
```
<http://www.w3.org/People/EM/contact#me> <http://www.w3.org/2000/10/swap/pim/contact#fullName> "Eric Miller" .
<http://www.w3.org/People/EM/contact#me> <http://www.w3.org/2000/10/swap/pim/contact#mailbox> <mailto:e.miller123(at)example> .
<http://www.w3.org/People/EM/contact#me> <http://www.w3.org/2000/10/swap/pim/contact#personalTitle> "Dr." .
<http://www.w3.org/People/EM/contact#me> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://www.w3.org/2000/10/swap/pim/contact#Person> .
```

FIGURE 7.4: N-triples format of RDF graph

7.2.3 Linked Open Data

Linked Open Data is a powerful mixture of *open data* and *linked data*. The idea of blending both open and linked data was formed in 2007 through a project entitled *linking open data* (Heath & Bizer 2011*a*). Since then, the W3C community and the semantic web research community have put tremendous effort into expanding the Linked Open Data (LOD) cloud. Figure 7.5 denotes the growth in the number of datasets published on the Web since the inception of the *linking open data project*. Currently, the LOD cloud contains over 1200 datasets, as depicted in Figure 7.6. Each *node* in Figures 7.5 and 7.6 represents *distinct open datasets* published as linked data. The *edges* indicate whether there are *links* between the items in two datasets.

From the datasets published under LOD cloud, *DBpedia* has been at the *heart of the LOD cloud* since the establishment of the *linking open data* project (Figures 7.5 and 7.6) and is considered to be the *core cross-domain knowledge base* (Heath & Bizer 2011*a*). It is the Linked Data version of *Wikipedia*, and has also been interlinked with numerous other knowledge resources since the initiation of the LOD cloud (see Figures 7.5 and 7.6).



May, 2007

September, 2008

July, 2009

September, 2010

FIGURE 7.5: Evolution of LOD cloud over the years

7.2.4 DBpedia Knowledge Base

DBpedia is a giant among current cross-domain knowledge bases, and serves as a hub in the *Web of Linked Data* (Lehmann et al. 2015, Heath & Bizer 2011*a*). It is also considered to be one of the main factors behind the success of the *linking open data project* (discussed in Section 7.2.3) (Lehmann et al. 2015). DBpedia was initiated in 2007 through the collaboration of the *Free University of Berlin* and *University of Leipzig* (Abián et al. 2017). The main aim of DBpedia was to build a large-scale, cross-domain and cross-lingual knowledge base by extracting the structured content in *Wikipedia*, which is the most widely used encyclopedia, a globally popular and heavily visited website, a central knowledge source of humankind, and a finest example of collaboratively created content (Lehmann et al. 2015, Auer et al. 2007, Kobilarov et al. 2009).

While most existing knowledge bases cover only a specific domain (Kobilarov et al. 2009), DBpedia spans multiple domains and languages by connecting isolated topical islands into one interconnected knowledge space (Heath & Bizer 2011*a*). Moreover, most of these existing knowledge bases are created by a small number of knowledge engineers; thus, it is highly cost-intensive to keep their information up-to-date, as domains change with time (Kobilarov et al. 2009). DBpedia addresses this issue through its open community vision.

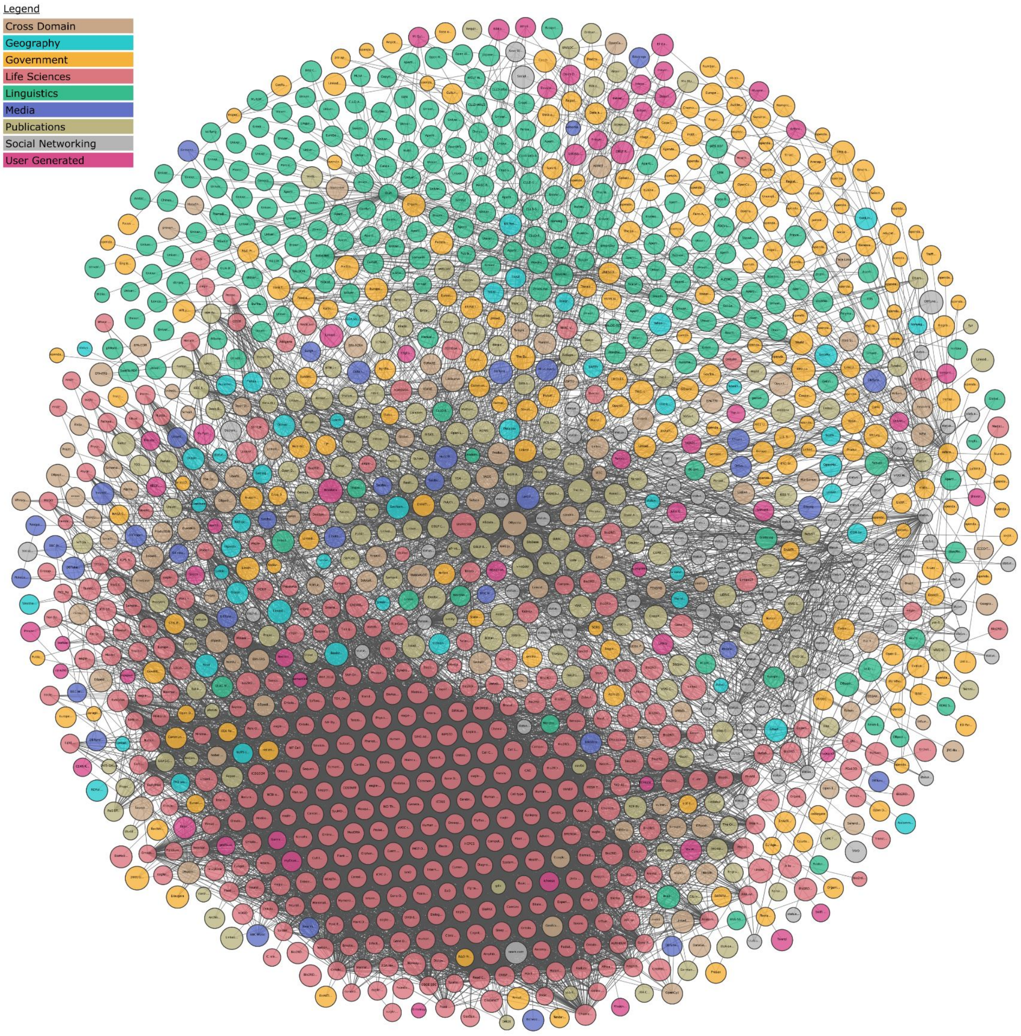


FIGURE 7.6: LOD cloud as at May, 2020 (note that the colours represent the topical domain that each dataset represents) Source: `https://lod-cloud.net/`

Due to the cross-domain and cross-lingual nature of DBpedia, it has been widely used in numerous applications, algorithms and tools including *data integration*, *document ranking*, *topic detection* and *named entity recognition* (Lehmann et al. 2015, Exner & Nugues 2012). In essence, DBpedia provides a rich platform to explore the gigantic knowledge source in Wikipedia and other datasets linked to it through sophisticated queries (Leal et al. 2012), using RDF query languages such as SPARQL (Pérez et al. 2009). Figure 7.7 demonstrates the knowledge extraction framework of DBpedia (Bizer et al. 2009).

Succinctly, DBpedia prevails over existing knowledge bases due to its inheritance of many strong points that are lacking in the existing knowledge bases as summarised below (Mirizzi et al. 2010, Kobilarov et al. 2009).

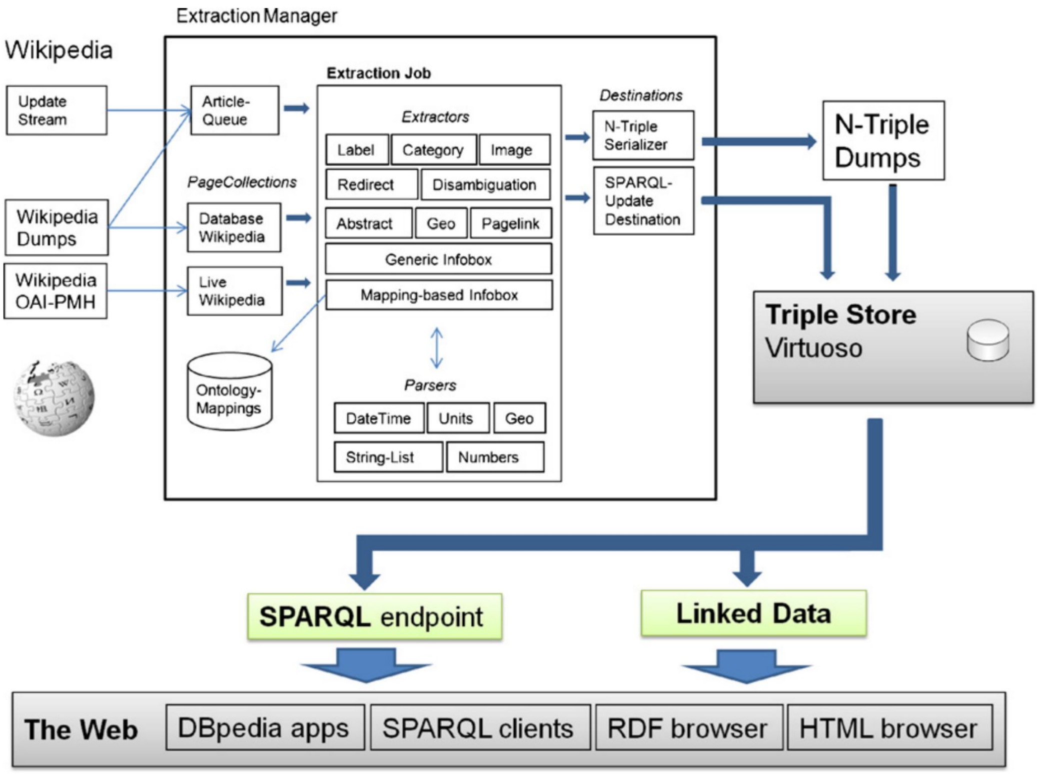


FIGURE 7.7: High-level overview of DBpedia knowledge extraction (Bizer et al. 2009)

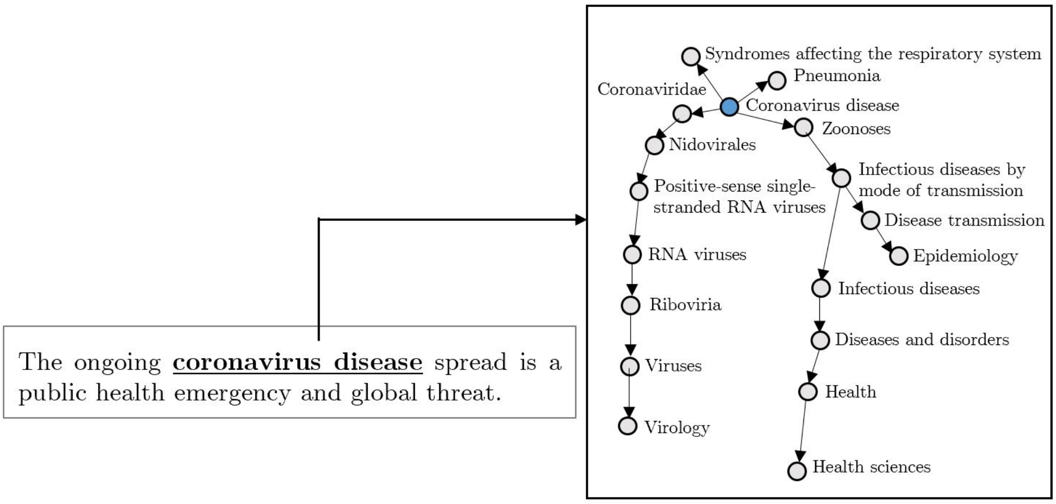
- DBpedia spans multiple domains, including more than 23 billion pieces of information (i.e., *RDF triples*).

- DBpedia plays a central role in the LOD community effort, as well as being one of the central interlinking hubs of Web of Data and the core cross-domain knowledge base in the LOD cloud.

- The real community agreement of DBpedia ensures that it is continuously updated, which reflects its dynamic, fast-growing and up-to-date nature.

- DBpedia is multilingual and contains more than 130 localised versions.

Considering the aforementioned strengths of DBpedia (all of which are rare in comparison to the knowledge bases, both generally and in the LBD domain), this study selected the *DBpedia knowledge base* for further analysis.



The ongoing **coronavirus disease** spread is a public health emergency and global threat.

FIGURE 7.8: Semantic augmentation generates additional knowledge

7.2.5 Semantic Augmentation

The key benefit of Linked Data and their accessibility through RDF query languages such as SPARQL (Pérez et al. 2009) is the ability to attach semantics to a given text as an aid to automatically interpret the meaning conveyed through a text. This procedure is termed *semantic augmentation* (a.k.a. *semantic tagging* or *semantic annotation*) (Dill et al. 2003). The main aim of semantic augmentation is to generate additional knowledge from the text, as depicted in Figure 7.8. Note how semantic augmentation has made the term *coronavirus disease* machine-interpretable. Therefore, the selection of the DBpedia knowledge base (discussed in Section 7.2.4) ensures the ability to perform such automated semantic inferences by exploring its massive machine-readable data using a semantic augmentation procedure.

7.3 Existing Impediments

The purpose of this section is to identify the existing *domain-dependent impediments* in the LBD workflow. In this regard, this study followed the framework that is typically used in most LBD systems, as suggested in a recent LBD review by Henry & McInnes (2017). This framework contains five main phases: *preprocessing*, *term linking*, *uninformative term removal*, *term ranking and thresholding*, and *evaluation/display results*, as denoted in Figure 7.9.

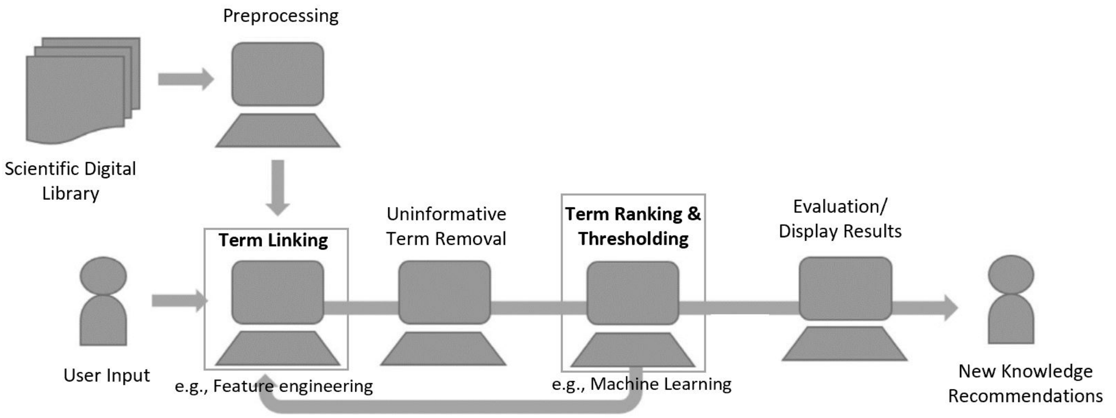

Preprocessing
Scientific Digital Library
Term Linking
Uninformative Term Removal
Term Ranking & Thresholding
Evaluation/ Display Results
User Input
e.g., Feature engineering
e.g., Machine Learning
New Knowledge Recommendations

FIGURE 7.9: Typical LBD framework followed by most LBD models (Henry & McInnes 2017)

When closely inspecting the framework, it is clear that *term linking* denotes the knowledge discovery algorithm employed in the LBD model. For instance, this could be the *feature engineering* phase of the LBD model. The *term ranking and thresholding* phase denotes the recommendation component of the model. This could be the machine learning component or even a simple statistical-based ranking mechanism employed in the LBD model. Hence, most of the *domain-dependent decisions* are commonly performed at the *preprocessing* and *uninformative term removal* stages. This is assuming that the model is capable of identifying novel knowledge linkages with *domain-independent features* at the feature engineering phase (i.e., features without the involvement of domain-specific knowledge resources) and identifying *domain-independent ranking mechanisms* during the recommendation component (i.e., employing machine learning or statistical-based ranking methods without incorporating domain-dependent knowledge resources), as highlighted in Figure 7.9.

Further disentangling tasks typically used in these two stages: *preprocessing*, and *uninformative term removal*, the following four major impediments were identified as the most common domain-dependent impediments in the existing LBD models: 1) *concept extraction (i.e., discipline-related terminology)*, 2) *semantic type filtering*, 3) *synonym identification* and 4) *granularity detection*. This study intends to address the problem of how to circumvent these impediments with high precision.

7.4 Defining Portability in the LBD Setting

This section is dedicated to explaining the notion of *portability* and the *criteria (or characteristics)* that should be fulfilled when proposing a portable framework in the context of LBD.

Portability refers to a model's ability to be executed in *new environments* (Mooney 1995). In the context of LBD, the new environments are typically different scientific domains (i.e., *cross-domain support*) and publication languages (i.e., *cross-lingual support*). To facilitate a complete and accurate knowledge search in these new environments, the LBD model should have high *coverage* of scientific facts from *reliable and up-to-date* sources. *Degree of portability* denotes the *costs* involved in portability. A model is portable if its *degree of portability* is lower than its redevelopment costs (Mooney 1995). In other words, the LBD model is portable if it is transferable across domains and languages without any heavy configurations (i.e., *easy transition*) or involvement of domain/language experts (i.e., *automation*).

In essence, the following six criteria should be fulfilled to ensure *portability* in the LBD context, where the first four criteria signify the characteristics related to the *new environments of portability* and the final two criteria indicate the characteristics related to *degree of portability.*

- New environments of portability
  1. Cross-domain support
  2. Cross-lingual support
  3. Reliability
  4. Coverage
- Degree of portability
  5. Easy transition
  6. Automation

7.5 Portability Check with DBpedia

Bearing in mind the potential benefits of employing the DBpedia knowledge base within the LBD workflow (discussed in Section 7.2), this study evaluated the extent to which DBpedia meets the defined criteria in Section 7.4.

The first criterion, *cross-domain support* is fulfilled through the use of DBpedia, since it is not restricted to a single main domain like the knowledge resources utilised so far in the LBD field. More specifically, DBpedia is a *cross-domain resource* that spans a wide range of academic domains, including (but not limited to) *medicine*, *computer science*, *sociology*, *psychology*, *geography*, *economics*, *anthropology*, *philosophy*, *law*, *languages and literature*, *history*, *arts*, *social work*, *biology*, *chemistry*, *earth science*, *space science*, *physics*, *mathematics*, *business*, *engineering* (including *chemical engineering*, *civil engineering*, *educational technology*, *electrical engineering*, *material science and engineering*, *mechanical engineering*) etc. Therefore, using DBpedia ensures that the data in these vast domains are reflected in the form of an interconnected knowledge space rather than fragmented, isolated topical islands (Mendes et al. 2012, Titze et al. 2014, Lehmann et al. 2015). This interconnected knowledge space allows us to transcend the restrictive environments of existing LBD models, which only cater to a single main domain or problem.

The second criterion, *cross-lingual support*, is also compatible with DBpedia due to its *multilingual nature*. This feature is rare among the knowledge resources utilised in the LBD field. To date, DBpedia supports more than 130 language editions, including (but not limited to) *German*, *French*, *Italian*, *Spanish*, *Polish*, *Russian*, *Portuguese*, *Catalan*, *Czech*, *Hungarian*, *Korean*, *Turkish*, *Arabic*, *Basque*, *Slovene*, *Bulgarian*, *Croatian*, *Greek* etc. Therefore, the use of DBpedia not only facilitates knowledge discovery in the *English* language, but also in a vast range of other publication languages (Aprosio et al. 2013, Lehmann et al. 2015, Chiarcos et al. 2012). This is particularly important in the LBD field given the *emerging non-English research* that exists in the LBD literature (Gao, Wang, Tao, Liu, Li, Yu, Yu, Tian & Zhang 2015, Qian et al. 2012, Yao et al. 2008).

The third criterion, *coverage* of information, is also preserved through the use of DBpedia. The main reason for this is that DBpedia is considered to be the *core cross-domain knowledge base* in the LOD cloud, lies at the *heart* of the LOD cloud, one of the *central* *interlinking hubs* of Web of Data, and plays a *pivotal role* in the LOD community's work. The main data source used in DBpedia is taken from *Wikipedia*, which is the *most widely used encyclopedia* and the *central knowledge source of humankind* (Lehmann et al. 2015, Auer et al. 2007, Kobilarov et al. 2009). In addition to knowledge from Wikipedia, DBpedia also interconnects with multitudinous knowledge resources that exist in the LOD cloud. To date, DBpedia constitutes more than 23 billion pieces of information (i.e., *RDF triples*). This wealth of information means that DBpedia can facilitate rich and informative knowledge discovery.

The fourth criterion, *reliability*, is ensured through the use of DBpedia, since it is continuously expanded and updated in line with changes to Wikipedia (a knowledge resource that is constantly improved and extended by a large global community). These information updates and additions adhere to the use of predefined collaborative procedures that ensure their reliability. One of the main negative consequences that could occur due to the notion of *open community vision* is *vandalism*, which is handled by employing a variety of vandalism removal methods including *bots*, *recent change patrols* and *watchlists* (Abián et al. 2017, Mola-Velasco 2011). The efficacy of these vandalism removal methods is evident, since the number of reported *incidental discoveries* (in which a reader identifies that vandalism has occurred) is considered to be rare (Broughton 2008). The methodical collaborative procedures involved in Wikipedia, as well as its relative freedom from vandalism, ensure that the information that DBpedia encloses is reliable and suited to knowledge discovery.

The fifth criterion, *easy transition*, is also fulfilled by the integration of DBpedia, since it interconnects isolated topical islands into one common data space (Heath & Bizer 2011*b*). Thus, DBpedia provides a single uniform view across domains and publications languages. This enables a querying of information that is more efficient than connecting numerous single domain knowledge resources into one single space that demands profuse design considerations (due to the differences in data types, data formats, programming languages, etc.). Such a process of connecting single resources into one space would be both time and labour intensive when changing domains. The use of DBpedia reduced such complexities, since it does not require any heavy configurations within the transitions among domains and publication languages.

The sixth criterion, *automation*, is ensured through the use of DBpedia, since it is based on the vision of the *Semantic Web*, i.e., to make Web data machine-readable (Taye 2010). In essence, to enable the encoding of semantics into data, W3C has defined technologies such as *RDF* and *OWL*, which make it possible for machines to access, process and understand data without human intervention. Consequently, the knowledge inferences required in the LBD process can be performed automatically without any human intervention (i.e., either without domain expert involvements (in *cross-domain* knowledge discovery) or without language expert involvements (in *cross-lingual* knowledge discovery)). Succinctly, this helps LBD discovery to be performed automatically in any domain and publication language that DBpedia supports.

Table 7.1 outlines the portability criteria check with DBpedia, along with a brief overview of the above-discussed justifications indicating how each criterion is met. Overall, it is evident that DBpedia adheres with all the characteristics that are defined considering the *new environments of portability* and the *degree of portability* (Mooney 1995). Therefore, this study mainly relies on DBpedia as the key knowledge base, in order to circumvent existing domain-dependent impediments in the LBD workflow (discussed in Section 7.3).

Table 7.1: Assessing the suitability of DBpedia in a portable LBD framework

Criteria	**Criteria Check**	**Justification**
Cross-domain Support	✓	DBpedia is not specific to a single domain, but spans *multiple domains*, and so avoids the fragmentation of data into isolated topical islands.
Cross-lingual Support	✓	DBpedia provides its *localised versions* in more than 130 languages, thereby allowing information extraction not only in English but also in other publication languages.
Coverage	✓	DBpedia is the *core* cross-domain knowledge-base in the *LOD cloud*, and it is also interlinked with numerous other data sources. Currently, DBpedia covers more than 23 billion pieces of information (*RDF triples*).

Reliability	✓	DBpedia is continuously extended and improved by a large global community with predefined collaboration procedures (e.g., the use of *templates*), which makes it an *up-to-date and reliable* resource.
Easy Transition	✓	DBpedia is a single interconnected data space connecting information from multiple domains into a *uniform view*. Therefore, its information can be queried from multiple domains and publication languages simultaneously, without any heavy configurations.
Automation	✓	The *semantic web* (including *DBpedia*) enables the machines to understand the data by encoding semantics using well-known technologies such as *RDF* and *OWL*, thereby allowing knowledge inferences to be performed automatically without any human/expert intervention.

7.6 Methodology

This section discusses how this study alleviates the detected domain-dependent impediments (discussed in Section 7.3) by performing semantic inferences that use the enormous body of machine-understandable knowledge encoded in the *DBpedia* knowledge base (in the form of RDF representation).

7.6.1 Concept Extraction

Identification of concepts from the unstructured text is one of the critical phases in the LBD process, as all of the reasoning and inference making of the discovery process relies on it. For this purpose, this study utilises *DBpedia entity names* (which represent *Wikipedia article titles*) as the main source for concept extraction. The main reason for the selection of DBpedia entity names is that they are considered to be *well-formed* and *succinct*, resembling terms in a conventional thesaurus (Milne et al. 2006, Wang et al.

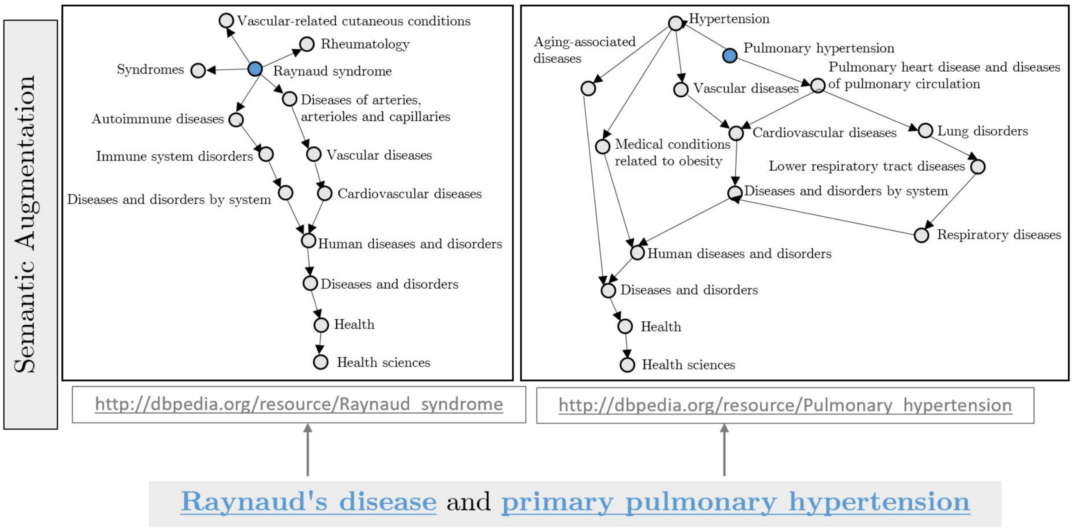

Semantic Augmentation
Vascular-related cutaneous conditions
Rheumatology
Syndromes
Raynaud syndrome
Diseases of arteries, arterioles and capillaries
Autoimmune diseases
Immune system disorders
Vascular diseases
Diseases and disorders by system
Cardiovascular diseases
Human diseases and disorders
Diseases and disorders
Health
Health sciences
Hypertension
Aging-associated diseases
Pulmonary hypertension
Pulmonary heart disease and diseases of pulmonary circulation
Vascular diseases
Cardiovascular diseases
Lung disorders
Medical conditions related to obesity
Lower respiratory tract diseases
Diseases and disorders by system
Respiratory diseases
Human diseases and disorders
Diseases and disorders
Health
Health sciences
http://dbpedia.org/resource/Raynaud_syndrome
http://dbpedia.org/resource/Pulmonary_hypertension
Raynaud's disease and primary pulmonary hypertension

FIGURE 7.10: Simplified example illustrating the semantic augmentation process

2009). Moreover, the dynamic, up-to-date and fast-growing nature of DBpedia (Wang et al. 2009) ensures a high topic coverage during the concept extraction process.

For instance, consider the article title *'Raynaud's disease and primary pulmonary hypertension'* (Celoria et al. 1960). The DBpedia entity extraction would be performed as depicted in Figure 7.10. In essence, '`http://dbpedia.org/resource/Raynaud_syndrome`' and '`http://dbpedia.org/resource/Pulmonary_hypertension`' denote the corresponding *DBpedia entries* (or *URIs*) of the two concepts: *Raynaud's disease* and *primary pulmonary hypertension*. Note that for the term *'Raynaud's disease'*, the latter part of the DBpedia URI is *'Raynaud_syndrome'*, and for the term *'primary pulmonary hypertension'*, the latter part of the DBpedia URI is *'Pulmonary_hypertension'*. The main reason for such syntactic variations in the DBpedia entries is that *Raynaud's disease* is a redirect resource (as discussed in Section 7.6.4); thus, this term is mapped to its main entity resource, which is titled 'Raynaud_syndrome'. Similarly, *primary pulmonary hypertension* is also a redirect resource in which 'Pulmonary_hypertension' is the main entity resource. Mapping the unstructured text to the corresponding DBpedia entries (a.k.a. *semantic augmentation*, as discussed in Section 7.2.5) provides the opportunity to make automated semantic deductions, in order to identify the meaning of concepts in the text (Figure 7.10).

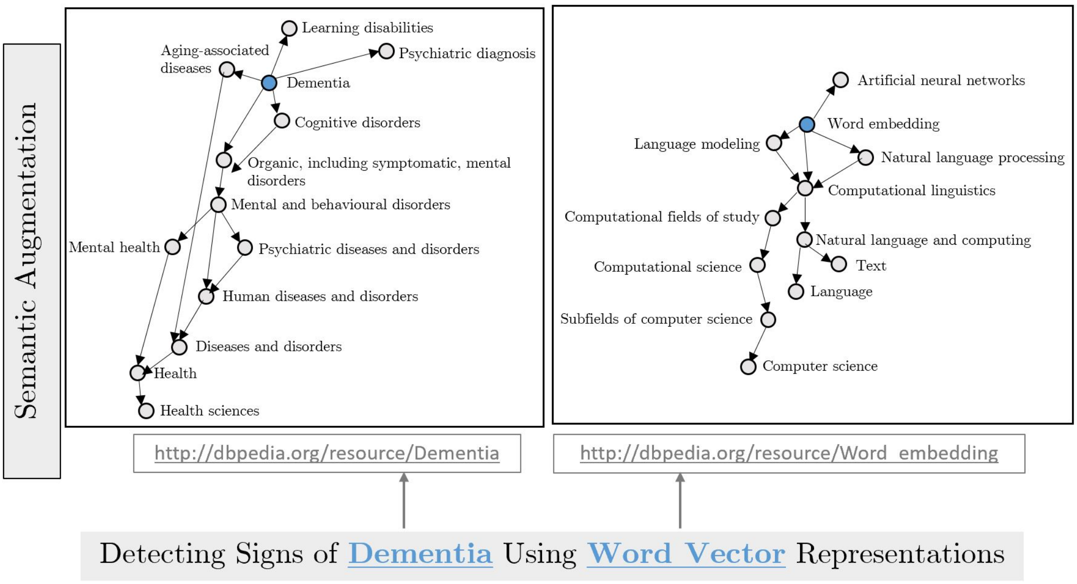

Semantic Augmentation
Learning disabilities
Aging-associated diseases
Psychiatric diagnosis
Dementia
Cognitive disorders
Organic, including symptomatic, mental disorders
Mental and behavioural disorders
Mental health
Psychiatric diseases and disorders
Human diseases and disorders
Diseases and disorders
Health
Health sciences
Artificial neural networks
Word embedding
Language modeling
Natural language processing
Computational linguistics
Computational fields of study
Natural language and computing
Computational science
Text
Language
Subfields of computer science
Computer science
http://dbpedia.org/resource/Dementia
http://dbpedia.org/resource/Word_embedding
Detecting Signs of Dementia Using Word Vector Representations

FIGURE 7.11: Exemplifying the need for discipline-related terminology extraction

7.6.2 Discipline-related Terminology Extraction

Unlike the existing resources used in the LBD domain (which merely cover concepts in a *single domain*), DBpedia spans a wide variety of domains. Hence, the concepts identified from the text are in *multiple domains*. However, the LBD user may only be interested in retrieving new knowledge from a single main domain (e.g., analysing only *medical* topics). In such situations, the extracted concepts that are not relevant to the *selected main domain* need to be filtered out. For example, consider a paper entitled, *'Detecting Signs of Dementia Using Word Vector Representations'* (Mirheidari et al. 2018). The semantic augmentation (as discussed in Section 7.6.1) can be performed, as illustrated in Figure 7.11. If the user is only interested in *medical topics* in the knowledge discovery, non-medical mappings such as *word vector* need to be removed (Figure 7.11). The purpose of the *discipline-related terminology extraction component* (performed as part of *concept extraction*) is to cope with situations like those illustrated in Figure 7.11.

To facilitate the identification of such discipline-related terminology, it is necessary to explore the machine-readable knowledge encoded in the relevant DBpedia entry. For instance, consider the paper title *'Raynaud's disease and primary pulmonary hypertension'* (Celoria et al. 1960), discussed in Section 7.6.1. The fact that the two concepts *Raynaud's disease* and *primary pulmonary hypertension* belong to the same domain (i.e., *medicine*) remains opaque to machines. This is where the mappings of the two DBpedia

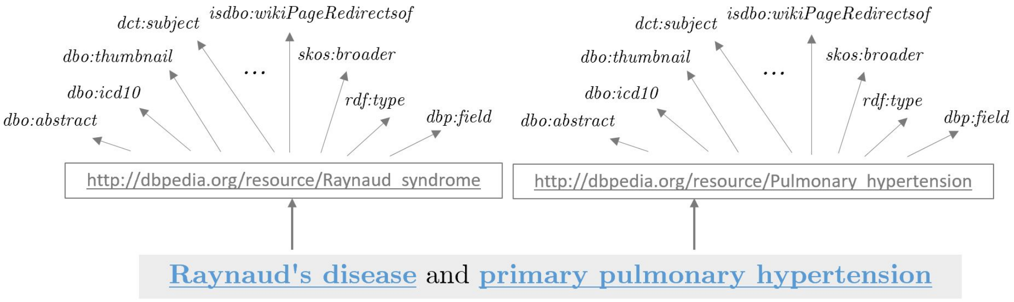

dct:subject
isdbo:wikiPageRedirectsof
dbo:thumbnail
...
skos:broader
dbo:icd10
rdf:type
dbo:abstract
dbp:field
http://dbpedia.org/resource/Raynaud_syndrome
dct:subject
isdbo:wikiPageRedirectsof
dbo:thumbnail
...
skos:broader
dbo:icd10
rdf:type
dbo:abstract
dbp:field
http://dbpedia.org/resource/Pulmonary_hypertension
Raynaud's disease and primary pulmonary hypertension

Figure 7.12: Predicates in DBpedia

URIs are crucial, since these URIs can be considered as the entry points from which to explore the knowledge encoded in DBpedia (i.e., in essence, its *RDF graph*). To access this encoded knowledge, the *predicates* (or *properties*) in the RDF graph can be utilised. Predicates of the two DBpedia URIs in the example title are denoted in Figure 7.12.

From Figure 7.12, it is evident that there are numerous predicates available in DBpedia; thus, it is necessary to carefully pick the predicate that is best suited to the problem at hand. Table 7.2 presents examples of few selected predicates that are available in DBpedia for the concept *'Raynaud_syndrome'*. Similarly, the DBpedia predicates of the remaining concept from the example title (*'Pulmonary_hypertension'*) are outlined in Table C.1.

Table 7.2: Several predicates from the DBpedia RDF graph on the subject *'Raynaud syndrome'*

No.	**Property *(Predicate)***	**Value *(Object)***	**Comments**
1	*dbo:abstract*	Raynaud syndrome, also known as Raynaud's is a medical condition in which there are episodes of reduced blood flow due to spasm of arteries. Typically the fingers and less ...	denoting the summary of the resource from Wikipedia page
2	*dbo:icd10*	173.0	denoting the corresponding mapping from ICD-10, where *173.0* maps to *Raynaud syndrome*
3	*dbo:icd9*	443.0	denoting the corresponding mapping from ICD-9, where *443.0* maps to *Raynaud's syndrome*
4	*dbo:meshid*	D011928	denoting the corresponding mapping from MeSH, where *443.0* maps to *Raynaud disease* `https://id.nlm.nih.gov/mesh/D011928.html`

5	*dbo:omim*	179600	denoting the corresponding mapping from OMIM, where *179600* maps to *Raynaud disease*
6	*dbo:wikiPage ExternalLink*	*http://niams.nih.gov/Health_Info/Raynauds_Phenomenon/default.asp*, *http://www.nhlbi.nih.gov/health/dci/Diseases/raynaud/ray_what.html*, *http://healthlink.mcw.edu/article/926055412.html*	denoting the external pages linked in Wikipedia page
7	*dbo:wikiPageID*	599203	denoting the page ID of Wikipedia
8	*dbp:diseasesdb*	25933	denoting the corresponding mapping from diseases database, where *25933* maps to *Raynaud phenomenon*
9	*dbp:field*	*dbr:Rheumatology*	denoting the corresponding field(s) mapped
10	*dct:subject*	*dbc:Vascular-related_cutaneous_conditions* *dbc:Autoimmune_diseases* *dbc:Rheumatology* *dbc:Syndromes* *dbc:Diseases_of_arteries,_arterioles_and_capillaries*	denoting the immediate categories of the subject *Raynaud syndrome*, where *'dbc:'* represents the DBpedia category
11	*rdf:type*	*owl:Thing*, *wikidata:Q12136*, *dbo:Disease*	denoting the class that the subject *Raynaud syndrome* is an instance of
12	*owl:sameAs*	*wikidata:Raynaud syndrome*, *dbpedia-cs:Raynaud syndrome*, *dbpedia-de:Raynaud syndrome*, *dbpedia-es:Raynaud syndrome*, *dbpedia-fr:Raynaud syndrome*, *dbpedia-it:Raynaud syndrome*, *dbpedia-ja:Raynaud syndrome*, *dbpedia-ko:Raynaud syndrome*, *dbpedia-nl:Raynaud syndrome*, *dbpedia-pl:Raynaud syndrome*, *dbpedia-pt:Raynaud syndrome*, *dbpedia-wikidata:Raynaud syndrome*	denoting the mappings of the connected datasets for the subject *Raynaud syndrome* (note that URIs with *'dbpedia-xx'* denotes the corresponding localised entry)

13	*is dbo:wikiPage Redirects of*	*dbr:Raynaud’s_disease_and_Raynaud’s_phenomenon*, *dbr:Reynaud’s*, *dbr:Reynaud’s_disease*, *dbr:Raynaud’s_disease*, *dbr:Raynaud_phenomenon*, *dbr:Reynaud’s_phenomenon*, *dbr:Raynauds_disease*, *dbr:Reynaud’s_Disease*, *dbr:Raynaud’s_disorder*, *dbr:Intermittent_arterial_vasospasm*, *dbr:Raynaud’s_Disease*, *dbr:Raynaud’s_syndrome*, *dbr:Raynaud’s_phenomenon*, *dbr:Raynaud’s_disease/phenomenon*, *dbr:Raynaud_disease*, *dbr:Raynauds*, *dbr:Raynauds_Syndrome*, *dbr:Raynauld’s_syndrome*, *dbr:Raynauld_syndrome*, *dbr:Reynaud’s_phenomenon*, *dbr:Primary_Raynaud’s_phenomenon*, *dbr:Raynaud’s_Phenomenon*, *dbr:Raynaud’s_Syndrome*, *dbr:Reynaud’s_Syndrome*, *dbr:Reynaud’s_syndrome*, *dbr:Primary_raynaud’s_phenomenon*, *dbr:Secondary_raynaud’s_phenomenon*, *dbr:Raynaud’s_Syndrome*, *dbr:Raynaud’s*	denoting the redirects of the subject *Raynaud syndrome*

When closely inspecting Table 7.2 and Table C.1, one could simply assume that the DBpedia entities related to the main domain *Medicine* could be located by verifying whether a DBpedia entity has properties (or ‘predicates’ in RDF terminology) that denote links to *medical classifications* or *external medical resources*. Examples of such medical classifications could include *ICD-10* (e.g., No. 2 in Table 7.2), *ICD-9* (e.g., No. 3 in Table 7.2), *OMIM* (e.g., No. 5 in Table 7.2) and *DiseasesDB* (e.g., No. 8 in Table 7.2); examples of such external medical resources could include *MeSH* (e.g., No. 4 in 7.2), *eMedicine*, *GeneReviews*, *Orphanet* and *MedlinePlus*. However, recall that the main objective of this study is to cater to *domain generalisability*. Thus, the proposed solution should fulfil the same requirements for the DBpedia resources in *other domains* too. For instance, consider the following three example concepts that are used in three separate domains outside the domain of Medicine.

- Word embedding (from the *natural language processing* domain): Table 7.3
- Big Five personality traits (from the *psychology* domain): Table C.2
- Bloom’s taxonomy (from the *education* domain): Table C.3

TABLE 7.3: Several predicates from the DBpedia RDF graph of the subject *'Word embedding'* (note that the property values indicate similar meanings to those in Table 7.2's *'comments'* column)

No.	**Property** ***(Predicate)***	**Value** ***(Object)***
1	*dbo:abstract*	Word embedding is the collective name for a set of language modeling and feature learning techniques in natural language processing (NLP) where words or phrases from the ...
2	*dbo:wikiPageID*	43561218
3	*dct:subject*	*dbc:Artificial_neural_networks*, *dbc:Language_modeling*
4	*owl:sameAs*	*freebase:Word embedding*, *wikidata:Word embedding*, *dbpedia-cs:Word embedding*, *dbpedia-eu:Word embedding*, *dbpedia-fr:Word embedding*, *dbpedia-wikidata:Word embedding*
5	*is dbo:wikiPage Redirects of*	*dbr:Thought_vectors*, *dbr:Word_vector*, *dbr:Word_vector_space*, *dbr:Word_vectors*

It is evident that such properties/predicates related to *classifications* and *external resources* (as in Tables 7.2 and C.1) are rare or almost non-existent in the DBpedia resources relating to *other domains* (Table 7.3, Table C.2 and Table C.3). Thus, locating *domain-specific terminology* in the text by merely considering such *expedients* does not meet our objective of *domain generalisability.* To provide a more comprehensive solution that is rewarding in every domain, this study explored the property *'dct:subject'*.

7.6.2.1 'dct:subject'

The property (or predicate) *dct:subject*[8] denotes the belonging of a concept/topic to its immediate *categories* (Stankovic et al. 2011). In essence, this property enables a model to bridge the *topic layer* with the *category layer* in DBpedia, as illustrated in Figure 7.13. Unlike predicates that are relevant to *classifications (e.g., ICD-10)* and *external resources (e.g., MeSH)*, the *'dct:subject'* is consistently available, in DBpedia entities in both *medical* (e.g., No. 10 in Tables 7.2 and C.1) and *non-medical* domains (e.g., No. 3 in Table 7.3, Table C.2 and Table C.3).

For instance, consider 'dct:subject' in Table 7.2, which represents the immediate categories of *Raynaud syndrome.* These include dbc:Vascular-related_cutaneous_conditions,

[8] this is one of DCTERMS (Dublin Core Metadata Terms) metadata: https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

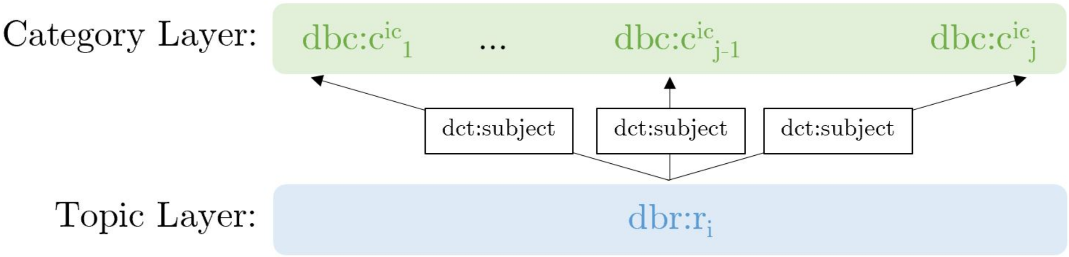


FIGURE 7.13: Deriving the immediate categories of *dbr:r_i* using 'dct:subject', where *dbc:$c^{ic}{}_j$* denotes the j^{th} immediate category in DBpedia

dbc:Autoimmune_diseases, dbc:Rheumatology, dbc:Syndromes and dbc:Diseases_of_arteries,_arterioles_and_capillaries. Using knowledge which humans have previously acquired, they can easily determine that the immediate categories of *Raynaud syndrome* belong to *Medicine*. Similarly, humans could effortlessly identify that the immediate categories of *Pulmonary Hypertension* (i.e., dbc:Hypertension and dbc:Pulmonary_heart_disease_and_diseases_of_pulmonary_circulation as denoted in Table C.1) belong to *Medicine*.

When considering non-medical domains, the immediate categories of *word embeddings* (which are dbc:Artificial_neural_networks and dbc:Language_modeling shown in Table 7.3) give an indication to humans that *word embeddings* belong in the *Natural Language Processing* domain, or, more broadly, to the *Computer Science* domain. Similarly, when inspecting the immediate categories of the two other example concepts, *Big Five personality traits* (dbc:Personality_traits, denoted in Table C.2), and *Bloom's taxonomy* (dbc:Stage_theories, dbc:Classification_systems, dbc:Educational_psychology, and dbc:Educational_technology, denoted in Table C.3) humans can easily determine the domains to which they belong through derived immediate categories (i.e., *psychology* and *education*, respectively) based on their prior knowledge.

However, such human-like deductions, concluded at a glance by inspecting the immediate categories of a concept, are not straightforward for machines. This indicates that machines are incapable of determining the main domain to which a concept belongs by considering its immediate categories derived using the *'dct:subject'* property alone. Thus, there is a need to further explore DBpedia's *category structure*. For this purpose, this study utilises the *property* (or *predicate* in RDF terminology) *'skos:broader'*.

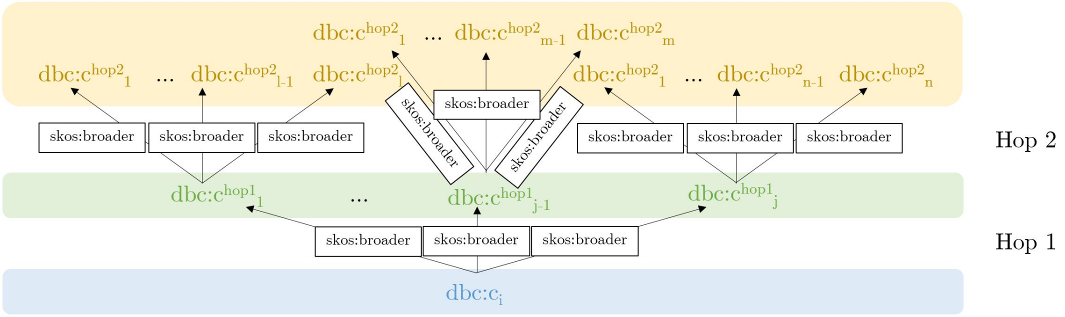

dbc:c^hop2_1 … dbc:c^hop2_m-1 dbc:c^hop2_m
dbc:c^hop2_1 … dbc:c^hop2_l-1 dbc:c^hop2_l
dbc:c^hop2_1 … dbc:c^hop2_n-1 dbc:c^hop2_n
skos:broader
Hop 2
dbc:c^hop1_1 … dbc:c^hop1_j-1 dbc:c^hop1_j
Hop 1
dbc:c_i

FIGURE 7.14: Deriving the categories of *dbr:*c_i using 'skos:broader'

7.6.2.2 'skos:broader'

The property/predicate *'skos:broader'*[9] denotes the broader category/categories of a given category (Stankovic et al. 2011). Succinctly, this provides a platform from which to analyse *category-category relationships* wherein each DBpedia category is assigned to one or more categories (Figure 7.14). Note that unlike 'dct:subject', in which there is only one hop between the topic layer and the category layer (Figure 7.13), 'skos:broader' can be performed using many hops. Figure 7.14 exemplifies how 'skos:broader' was performed in two hops for the DBpedia category *dbc:*c_i.

Therefore, the use of 'skos:broader' property facilitates a rich understanding of each of the DBpedia categories in an automated manner. For example, consider the immediate DBpedia category (derived using 'dct:subject') of *Raynaud syndrome*, which is `dbc:Vascular-related_cutaneous_conditions`. The left column of Table 7.4 denotes how to move through (or navigate) DBpedia's category structure from this immediate category (using 'skos:broader'). In this example, 'skos:broader' is performed only until six hops. If one wishes, a more in-depth navigation could be performed. Table 7.4 shows that this search (using 'skos:broader') has begun to elicit the main domain concept of `dbc:Vascular-related_cutaneous_conditions`, which is `dbc:Medicine` in the 4th, 5th and 6th hops (highlighted in Table 7.4). Also, note that the DBpedia category structure (inherited from *Wikipedia*) is a directed acyclic graph in which numerous categorisation schemes coexist simultaneously by forming a thematically organised thesaurus (Stankovic et al. 2011). This is the reason why the main domain concept `dbc:Medicine` is elicited at different hops (i.e., the 4th, 5th and 6th) rather than in a single fixed hop.

[9] this is one of SKOS (Simple Knowledge Organization System) semantic relations: https://www.w3.org/TR/skos-reference/

Overall, from the example demonstration in the left column of Table 7.4, it can be concluded that `dbc:Vascular-related_cutaneous_conditions` reaches its main domain concept `dbc:Medicine` with a shortest hop path count of four.

Much like the property 'dct:subject', 'skos:broader' is available not only in the *medical domain*, but also in *other domains*. To illustrate this fact, consider the non-medical DBpedia category `dbc:Personality_traits`, which belongs to the main domain *Psychology*. Through navigating in the DBpedia's category structure (using 'skos:broader' in six hops, as discussed above), `dbc:Personality_traits` begins to elicit its main domain concept `dbc:Psychology` (as highlighted in the right column of Table 7.4). From the example demonstration in Table 7.4, it can be concluded that the shortest hop path count from `dbc:Personality_traits` to `dbc:Psychology` is three.

TABLE 7.4: Simplified example demonstrating a sample of category relationships through the use of 'skos:broader' property, up to six hops, where the colours indicate selected DBpedia category, 1st hop, 2nd hop, 3rd hop, 4th hop, 5th hop and 6th hop

Example DBpedia category in *Medicine*	**Example DBpedia category in *Psychology***
dbc:Vascular-related_cutaneous_conditions	***dbc:Personality_traits***
dbc:Cutaneous_conditions	*dbc:Personality_theories*
—*dbc:Dermatology*	—*dbc:Personality*
——*dbc:Integumentary_system*	——*dbc:Conceptions_of_self*
———*dbc:Organ_systems*	———*dbc:Philosophical_concepts*
————*dbc:Anatomy*	————*dbc:Philosophy*
——————*dbc:Branches_of_biology*	——————*dbc:Abstraction*
——————*dbc:Morphology*	——————*dbc:Academic_disciplines*
——————*dbc:Structure*	——————*dbc:Belief*
——————*dbc:Zoology*	——————*dbc:Humanities*
————*dbc:Animal_anatomy*	———*dbc:Self*
——————*dbc:Animals*	————*dbc:Concepts_in_metaphysics*
——————*dbc:Anatomy*	——————*dbc:Metaphysics*
——————*dbc:Veterinary_medicine*	——————*dbc:Philosophical_concepts*
——————*dbc:Zoology*	————*dbc:Consciousness*
————*dbc:Biological_systems*	——————*dbc:Humans*
——————*dbc:Physical_systems*	——————*dbc:Mental_content*
——————*dbc:Systems_biology*	——————*dbc:Psychological_concepts*
————*dbc:Organs_(anatomy)*	——— etc.
——————*dbc:Anatomy*	———*dbc:Social_psychology*
——*dbc:Medical_specialties*	————*dbc:Branches_of_psychology*
———*dbc:Healthcare_occupations*	——————*dbc:Psychology*
————*dbc:Health_care*	——————*dbc:Subfields_by_academic_discipline*
——————*dbc:Health*	————*dbc:Interdisciplinary_subfields_of_sociology*
——————*dbc:Service_industries*	——————*dbc:Academic_discipline_interactions*
————*dbc:Occupations_by_type*	——————*dbc:Subfields_of_sociology*
——————*dbc:Categories_by_type*	——————*dbc:Subfields_by_academic_discipline*
——————*dbc:Occupations*	————*dbc:Psychology*
———*dbc:Medicine*	——————*dbc:Applied_sciences*
————*dbc:Branches_of_biology*	——————*dbc:Behavioural_sciences*
——————*dbc:Biology*	———*dbc:Psychological_concepts*
——————*dbc:Subfields_by_academic_discipline*	————*dbc:Psychology*
————*dbc:Health_care*	——————*dbc:Applied_sciences*
——————*dbc:Health*	——————*dbc:Behavioural_sciences*

```
        dbc:Service_industries
      dbc:Health_sciences
        dbc:Health
        dbc:Applied_sciences
        dbc:Life_sciences
    dbc:Clinical_medicine
      hldbc:Medicine
        dbc:Branches_of_biology
        dbc:Health_care
        dbc:Health_sciences
dbc:Diseases_and_disorders_by_system
  dbc:Organ_systems
    dbc:Anatomy
      dbc:Branches_of_biology
        dbc:Biology
        dbc:Subfields_by_academic_discipline
      dbc:Morphology
        dbc:Scientific_classification
        dbc:Taxonomy
      dbc:Structure
        dbc:Form
        dbc:Systems
      dbc:Zoology
        dbc:Branches_of_biology
        dbc:Animals
    dbc:Animal_anatomy
      dbc:Animals
        dbc:Eukaryotes
        dbc:Biota
      dbc:Anatomy
        dbc:Branches_of_biology
        dbc:Morphology
        dbc:Structure
        dbc:Zoology
      dbc:Veterinary_medicine
        dbc:Animals
        dbc:Health_sciences
        dbc:Medicine
      dbc:Zoology
        dbc:Branches_of_biology
        dbc:Animals
    dbc:Biological_systems
      dbc:Physical_systems
        dbc:Physics
        dbc:Systems
      dbc:Systems_biology
        dbc:Branches_of_biology
        dbc:Systems_science
    dbc:Organs_(anatomy)
      dbc:Anatomy
        dbc:Branches_of_biology
        dbc:Morphology
        dbc:Structure
        dbc:Zoology
  dbc:Diseases_and_disorders
    dbc:Health
      dbc:Main_topic_classifications
        dbc:Articles
        dbc:Container_categories
      dbc:Personal_life
        dbc:Euthenics
        dbc:Anthropology
        dbc:Philosophy_of_life
```

```
  dbc:Human_behavior
    dbc:Behavior
      dbc:Action_(philosophy)
        dbc:Free_will
        dbc:Metaphysical_theories
        dbc:Ontology
        dbc:Philosophy_of_mind
      dbc:Psychological_concepts
        dbc:Psychology
        dbc:Concepts_by_field
    dbc:Humans
      dbc:Apes
        dbc:Catarrhini
        dbc:Primate_taxonomy
      dbc:Invasive_mammal_species
        dbc:Invasive_animal_species
        dbc:Mammal_ecology
     etc.
    dbc:Behavior_by_type_of_animal
      dbc:Behavior
        dbc:Action_(philosophy)
        dbc:Psychological_concepts
  dbc:Psychological_concepts
    dbc:Psychology
      dbc:Applied_sciences
        dbc:Scientific_disciplines
        dbc:Applied_disciplines
      dbc:Behavioural_sciences
        dbc:Behavior
        dbc:Social_sciences
dbc:Psychological_theories
  dbc:Psychology
    dbc:Applied_sciences
      dbc:Scientific_disciplines
        dbc:Science
        dbc:Disciplines_by_type
        dbc:Subfields_by_academic_discipline
      dbc:Applied_disciplines
        dbc:Academic_discipline_interactions
        dbc:Academic_disciplines
        dbc:Disciplines_by_type
        dbc:Subfields_by_academic_discipline
    dbc:Behavioural_sciences
      dbc:Behavior
        dbc:Action_(philosophy)
        dbc:Psychological_concepts
      dbc:Social_sciences
        dbc:Society
        dbc:Academic_disciplines
        dbc:Scientific_disciplines
  dbc:Scientific_theories
    dbc:Theories
      dbc:Conceptual_systems
        dbc:Abstraction
        dbc:Cognitive_science
        dbc:Concepts
        dbc:Systems
        dbc:Systems_science
      dbc:Abstraction
        dbc:Innovation
        dbc:Creativity
        dbc:Philosophy_of_logic
        dbc:Structure
```

——————*dbc:Self* ———*dbc:Clinical_medicine* ————*dbc:Medicine* ——————*dbc:Branches_of_biology* ——————*dbc:Health_care* ——————*dbc:Health_sciences*	——————*dbc:Thought* ——— etc. —— etc.

7.6.2.3 dct:subject+skos:broader

This section describes how this study leverages the powerful combination of the two properties *'dct:subject'* and *'skos:broader'* in order to define *discipline-related terminology*, which is our ultimate goal. To summarise, Figure 7.15 denotes the topic-category structure of DBpedia using *'dct:subject'* and *'skos:broader'*. Succinctly, the immediate categories of each topic can be identified through *dct:subject* (i.e., *topic-category relationships*). Next, each of these categories is made up of broader categories, which can be accessed via the property *skos:broader* (i.e., *category-category relationships*; Figure 7.15).

This study entails navigating through the *topic-category* graph structure of DBpedia to elicit *domain-related terminology*. For example, consider the DBpedia graph snippet of the concept *Raynaud syndrome* (Figure 7.16). It is clear that most of the immediate categories of *Raynaud syndrome* (i.e., via *dct:subject*) reach the main domain concept *Medicine* quickly (i.e., using a lesser number of hops) via *skos:broader*.

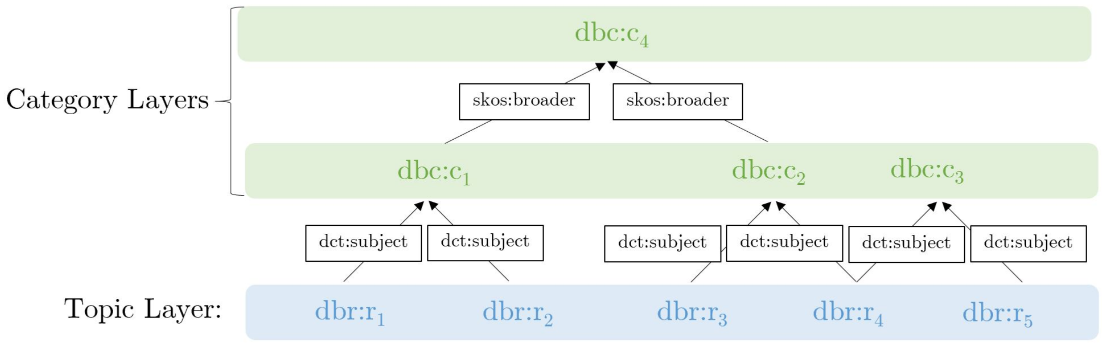


FIGURE 7.15: Exemplifying the topic-category link structure in DBpedia through the use of 'dct:subject' and 'skos:broader'

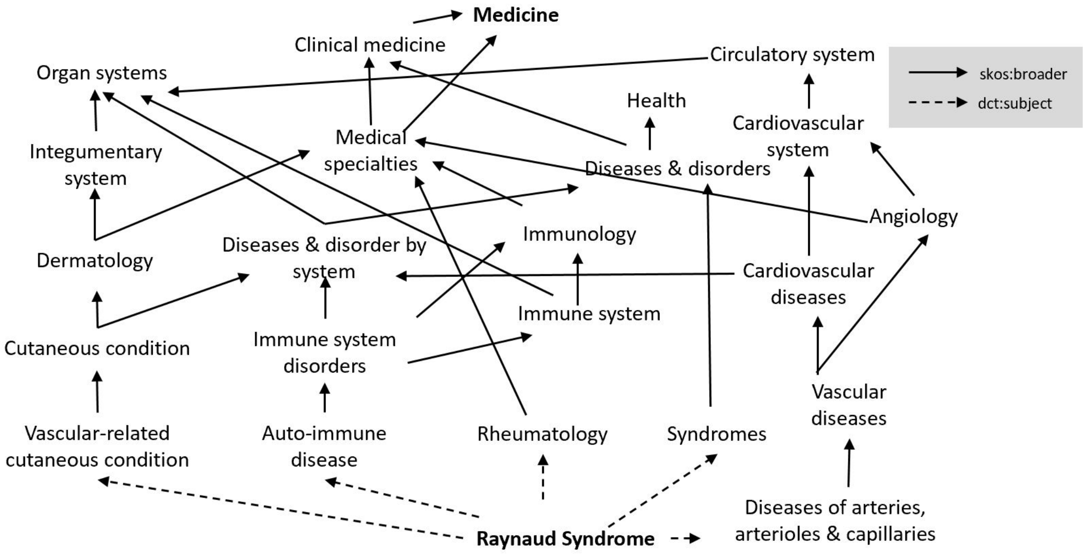


FIGURE 7.16: DBpedia graph snippet of *Raynaud Syndrome* denoting topic-category structure via 'dct:subject' and 'skos:broader'

7.6.2.4 Proposed Empirical Rules

Therefore, to decide whether a concept belongs to the main domain as selected by the user (e.g., *Medicine*), the following two empirical rules can be proposed. These rules are based on observations of DBpedia's *topic category* structure.

- *Empirical Rule 1:* This empirical rule concerns the shortest hop path count threshold n. Given a concept *dbr:*r_i (from the semantic augmentation procedure), this rule checks whether a majority of its immediate categories (i.e., *dbc:*${c^{ic}}_x$ derived via 'dct:subject') reach the main domain concept *dbc:*c_{main_domain} with fewer than n hops using 'skos:broader' (e.g., Figure 7.17). Defining the percentage of the immediate categories that reach the main domain concept using $<n$ as C(n), this empirical rule requires C(n) to be $\geqslant$ 50%. For example, suppose that n is equal to 6 in the example shown in Figure 7.17. Then, it can be concluded that three of the categories immediately next to the concept *dbr:*r_i reach the concept's main domain. Thus, the C(n) of this example is 75%.

- *Empirical Rule 2:* However, when the parameter n is increases (e.g., n>8), concepts that are not *directly relevant* to the selected main domain may also be included in the concept extraction. For instance, consider *Operant conditioning.* The user may not

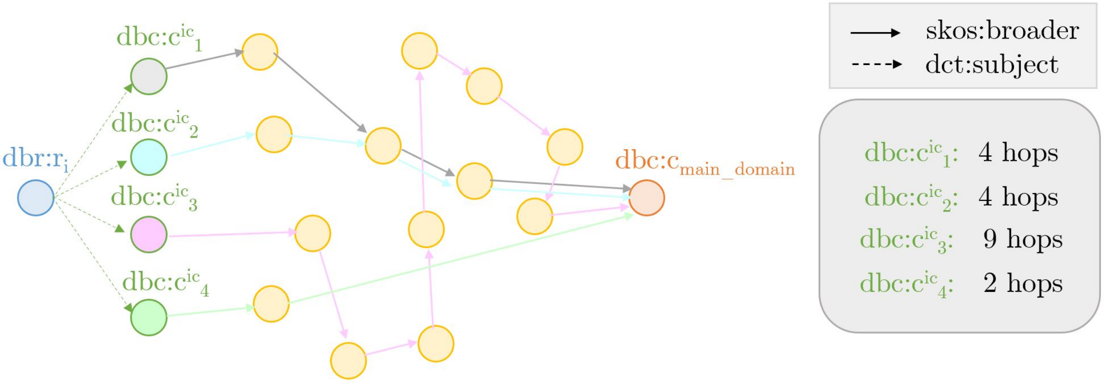


FIGURE 7.17: Simplified example of the proposed rules

necessarily expect to see this categorised as *medical* concept. However, this concept has a majority of *distant categories* which connect it to the main domain concept *Medicine* when n is increasing. To avoid the inclusion of such implicitly related concepts (to the main domain) in the discipline-related terminology, a new and more restricted rule that considers the direct semantic interactions of *immediate categories* with the main domain concept is required. For this purpose, this study defines a new parameter m ($m < n$) to identify whether a concept has at least one *immediate category* that reaches the selected main domain with fewer than m hops. That is, it is required that C(m) >0.

7.6.2.5 Word Sense Disambiguation

Due to the cross-domain nature of DBpedia, there is a possibility for a term to have multiple senses from different domains. If a term has multiple senses in DBpedia, they can be identified using the *'dbo:wikiPageDisambiguates'* predicate. For instance, consider the term "kuru", which has over 20 senses recorded in DBpedia. From the senses recorded in DBpedia, the term *kuru* could be a disease, a person, a place, a sport, a kingdom, mythology, etc., as illustrated in Figure 7.18. In situations where DBpedia has multiple senses for a given term, the most relevant sense to a given domain needs to be identified. In the example of *kuru* (in Figure 7.18), the most relevant sense to the domain of Medicine is *kuru (disease)*. To facilitate the identification of the most relevant sense in a given domain, this thesis uses the same process described in Section 7.6.2.4.

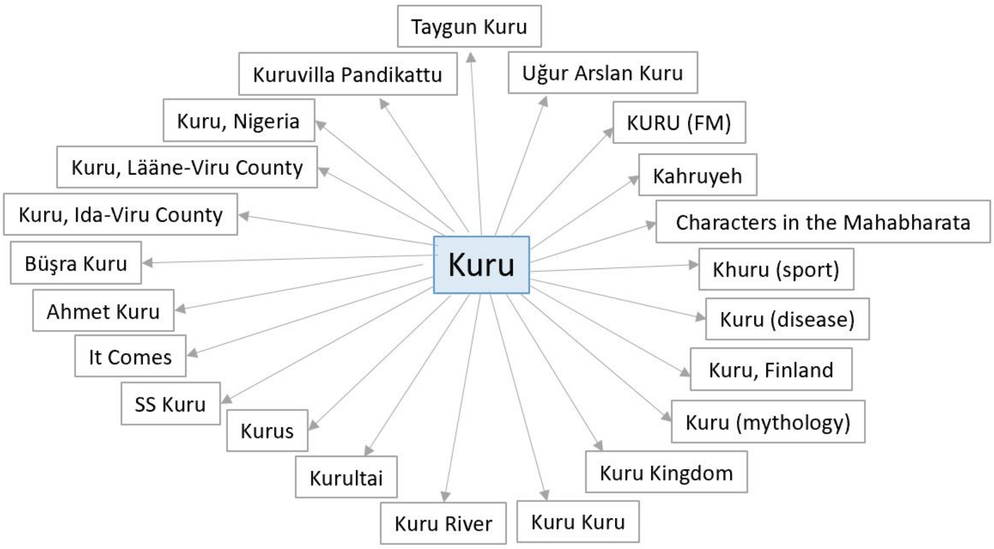


FIGURE 7.18: Senses of the term *kuru* extracted using 'dbo:wikiPageDisambiguates' predicate

To further elaborate the idea, consider the following four senses of the term kuru; *Kuru (disease)*, *Taygun Kuru*, *Kuru (Nigeria)* and *Khuru (sport)* (shown in Figure 7.18). The topic-category graph structures of these selected senses are depicted in Figures 7.19, 7.20, 7.21 and 7.22, respectively. By looking at these senses, a human can easily interpret the potential main domains that they belong, which are `dbc:Medicine` (in the sense of *Kuru (disease)*), `dbc:People` (in the sense of *Taygun Kuru*), `dbc:Places` (in the sense of *Kuru, Nigeria*) and `dbc:Sports` (in the sense of *Khuru (sport)*).

To automatically perform a similar interpretation of the senses' main domains, the two empirical rules defined in Section 7.6.2.4 can be used. More specifically, in the domain of Medicine, the sense *Kuru (disease)* obtains 83.33% of C(n) (i.e., the empirical rule 1) and 66.67% of C(m) (i.e., the empirical rule 2), given that n and m are 9 and 5, respectively. In other words, the sense *Kuru (disease)* fulfils both empirical rules defined in Section 7.6.2.4, indicating that it is the most relevant sense within the medical domain. The immediate DBpedia categories that reached the main domain concept of `dbc:Medicine` using $<n$ hops (i.e., the empirical rule 1) include `dbc:Transmissible_spongiform_encephalopathies`, `dbc:Rare_infectious_diseases`, `dbc:Foodborne_illnesses`, `dbc:Prions` and `dbc:Cannibalism_in_Oceania`. The immediate categories that reached the main domain concept of `dbc:Medicine` $<m$ hops (i.e., the empirical rule 2) are `dbc:Transmissible_spongiform_encephalopathies`, `dbc:Rare_infectious_diseases`, `dbc:Foodborne_illnesses` and `dbc:Prions`.

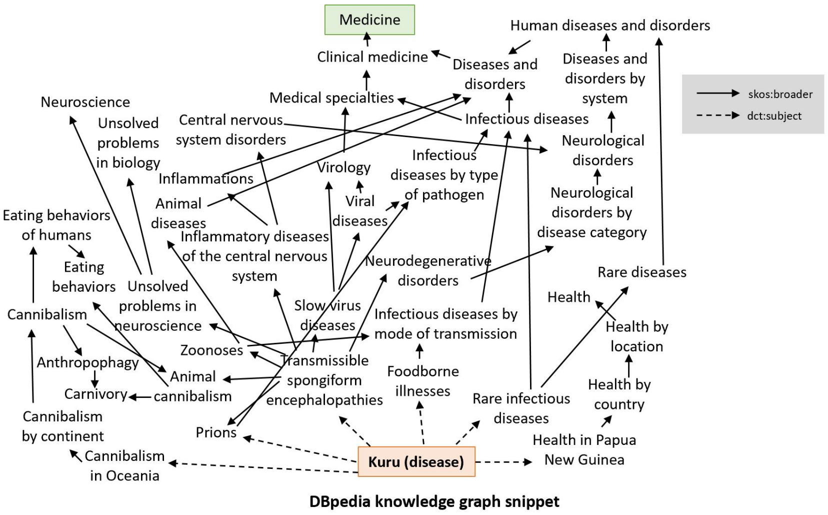

Medicine
Human diseases and disorders
Clinical medicine
Diseases and disorders
Diseases and disorders by system
skos:broader
dct:subject
Medical specialties
Neuroscience
Infectious diseases
Unsolved problems in biology
Central nervous system disorders
Neurological disorders
Infectious diseases by type of pathogen
Virology
Inflammations
Animal diseases
Viral diseases
Neurological disorders by disease category
Eating behaviors of humans
Inflammatory diseases of the central nervous system
Eating behaviors
Unsolved problems in neuroscience
Neurodegenerative disorders
Rare diseases
Cannibalism
Health
Slow virus diseases
Infectious diseases by mode of transmission
Health by location
Zoonoses
Anthropophagy
Transmissible spongiform encephalopathies
Animal cannibalism
Foodborne illnesses
Health by country
Carnivory
Rare infectious diseases
Cannibalism by continent
Prions
Health in Papua New Guinea
Cannibalism in Oceania
Kuru (disease)
DBpedia knowledge graph snippet

FIGURE 7.19: DBpedia graph snippet of the sense *kuru (disease)* using 'dct:subject' and 'skos:broader' predicates

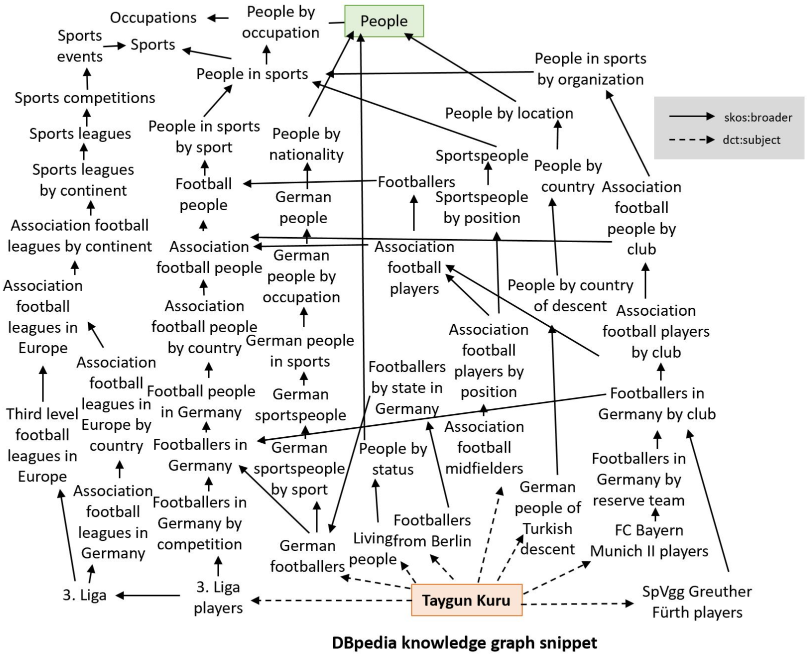

Occupations
People by occupation
People
Sports events
Sports
People in sports
People in sports by organization
Sports competitions
People by location
skos:broader
dct:subject
Sports leagues
People in sports by sport
People by nationality
Sportspeople
People by country
Sports leagues by continent
Football people
Footballers
Association football people by club
German people
Sportspeople by position
Association football leagues by continent
Association football people
Association football players
German people by occupation
People by country of descent
Association football leagues in Europe
Association football people by country
Association football players by club
German people in sports
Association football players by position
Association football leagues in Europe by country
Footballers by state in Germany
Third level football leagues in Europe
Football people in Germany
German sportspeople
Footballers in Germany by club
Association football midfielders
Footballers in Germany
German sportspeople by sport
People by status
Footballers in Germany by reserve team
Association football leagues in Germany
Footballers in Germany by competition
German people of Turkish descent
Footballers from Berlin
FC Bayern Munich II players
German footballers
Living people
3. Liga
3. Liga players
Taygun Kuru
SpVgg Greuther Fürth players
DBpedia knowledge graph snippet

FIGURE 7.20: DBpedia graph snippet of the sense *Taygun Kuru* using 'dct:subject' and 'skos:broader' predicates

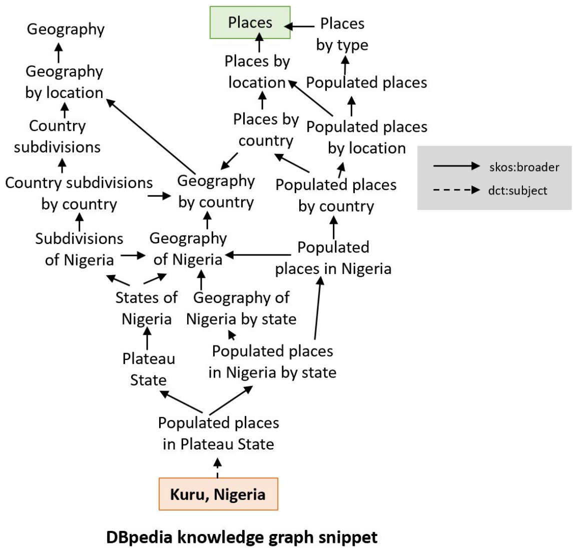


FIGURE 7.21: DBpedia graph snippet of the sense *Kuru, Nigeria* using 'dct:subject' and 'skos:broader' predicates

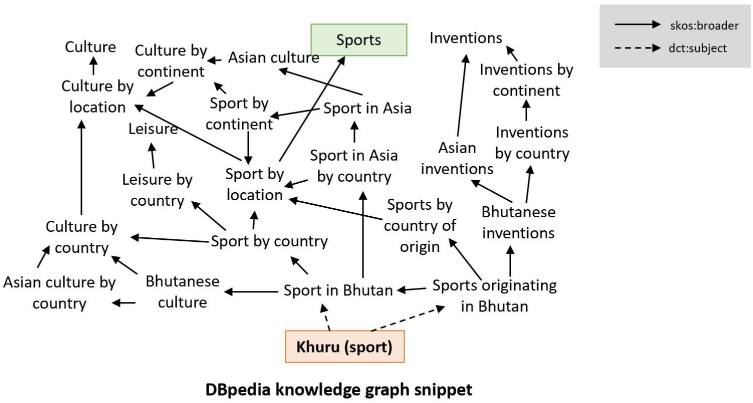


FIGURE 7.22: DBpedia graph snippet of the sense *Khuru (sport)* using 'dct:subject' and 'skos:broader' predicates

Using the same process in the domain of Medicine, *Taygun Kuru* obtains 33.33% of C(n) and 0% C(m). In other words, this sense does not fulfil the two empirical rules, indicating that it is not relevant to the medical domain. However, this sense obtains 88.89% of C(n) and 33.33% of C(m) in the context of *people*, which implies its relevancy to the selected context. Similarly, *Kuru (Nigeria)* fulfil none of the empirical rules from the domain of Medicine since it obtains 0% of C(n) and C(m) values. However, from the context of *places*, this sense obtains 100% of C(n) and C(m) values, indicating that it is the relevant sense from the perspective of places. As in *Kuru (Nigeria)*, *Khuru (sport)* obtains 0% of C(n) and C(m) in the domain of Medicine. However, it obtains 100% of C(n) and C(m) values for the context of *sports*.

As shown in the above-mentioned examples, the only sense that satisfies the two defined empirical rules in the domain of Medicine is *Kuru (disease).* Thus, in situations where a domain-specific term has multiple senses, the two empirical rules can be employed to detect the most relevant sense in the given domain.

7.6.3 Semantic Type Filtering

The LBD user may need to further narrow down the discovered novel knowledge based on semantic types. For instance, consider a situation where the user only wishes to analyse *diseases* in the literature (Jha, Xun, Wang & Zhang 2019). In such situations, the concepts need to be further filtered in such a way that those concepts only related to the selected semantic type (i.e., *diseases*) are retained in the knowledge discovery process (Jha, Xun, Wang & Zhang 2019).

As in Section 7.6.2, this study leverages properties *dct:subject* and *skos:broader* to carry out the filtering process described above. In this instance, the shortest hop path counts are calculated with respect to the specified *semantic type* (e.g., *diseases*). Moreover, the empirical rules imposed in this instance need to be *lighter* (or *less restrictive*) than those in Section 7.6.2 for the following two reasons.

- Semantic types are usually fewer hops away from the resource *dbr:*r_i than the main domain concept (e.g., *medicine*), as illustrated in Figure 7.23.
- Unlike the main domain concept, semantic types are typically linked with few immediate categories, as depicted in Figure 7.23.

For example, consider the DBpedia graph snippet of *'Raynaud syndrome'* (Figure 7.16) in which the semantic type *'diseases and disorders'* is considered to exemplify the two reasons listed above.

- It is evident that the shortest hop path count between the semantic type *diseases and disorders* and *Raynaud syndrome* is lower than the hop path count between the main domain concept *Medicine* and *Raynaud syndrome.* This is because semantic types are typically more granular than the main domain concept. Thus, when defining the

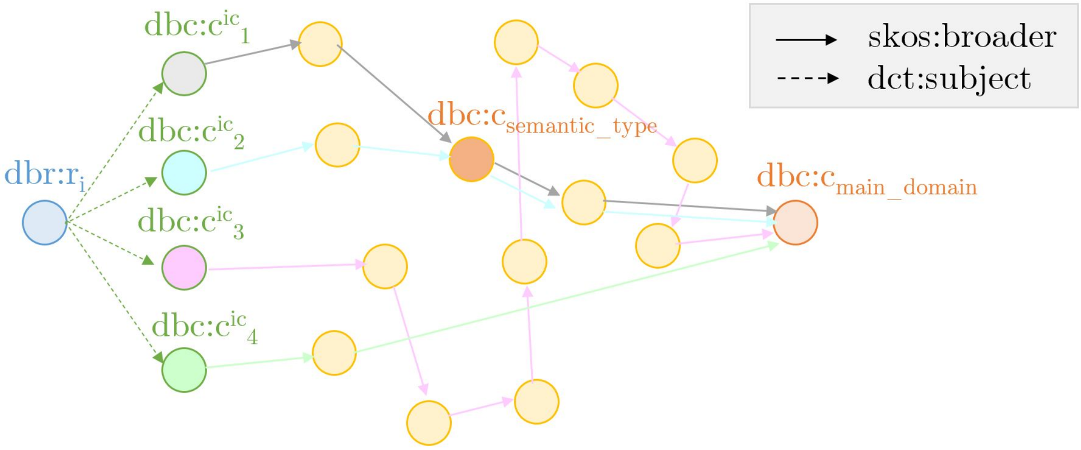


FIGURE 7.23: Schematic overview of semantic type filtering, where $\text{dbc:c}^{\text{ic}}{}_{\text{i}}$ is the i^{th} immediate category of dbr:r_{i} and $\text{dbc:c}_{\text{semantic\_type}}$ is the relevant semantic type

empirical rules for semantic type filtering, the shortest hop path count threshold n should be lower than that in Section 7.6.2.

- Most of the immediate categories of *Raynaud syndrome* link to its main domain concept, Medicine (via 'skos:broader'). But, irrespective of the length differences between the shortest hop path counts, it is unrealistic to assume that majority of the immediate categories will link with the selected semantic type (e.g., *diseases and disorders*). The main reason for this is that a semantic type typically denotes a single characteristic that describes a concept. Note that in the 'Raynaud syndrome' example, the immediate category `dbc:Rheumatology` is not connected with `dbc:Diseases_and_disorders`, even though it is connected with its main domain concept `dbc:Medicine`. Therefore, when defining rules, the immediate category count threshold $\text{C}(n)$ should be lower than in Section 7.6.2.

7.6.4 Synonym Identification

Identification of synonyms is one of the crucial steps in the LBD workflow, and it provides numerous benefits. The synonyms can be utilised from a *query expansion phase* to other tasks in remaining LBD components, for the purpose of making relevant semantic deductions. For instance, consider the *query expansion* phase of the LBD workflow as a use case. When the user inputs topics to the LBD system, the system should first extract the literature relevant to these defined input topics (i.e., the *local corpus*) (Kostoff, Briggs, Solka & Rushenberg 2008). For a complete search of potential novel

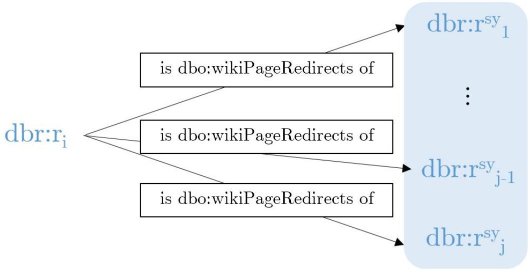

dbr:r_i
is dbo:wikiPageRedirects of
is dbo:wikiPageRedirects of
is dbo:wikiPageRedirects of
dbr:r^sy_1
⋮
dbr:r^sy_j-1
dbr:r^sy_j

FIGURE 7.24: Synonym identification of *dbr:r*$_i$ using *'is dbo:wikiPageRedirects of'*, where *dbr:r*$^{sy}{}_i$ denotes redirect DBpedia resources, and j denotes the number of redirect resources defined for *dbr:r*$_i$

knowledge, this extracted local corpus should contain all the literature relevant to the user's interests. For this purpose, the LBD system needs to construct an *expanded query* using synonyms. Thus, even if the user simply enters 'Raynaud's disease', the system will not only obtain literature that contains 'Raynaud's disease', but also other related literature containing synonymous terms such as 'Raynaud syndrome', 'Raynaud disease', etc. In essence, when the LBD system uses a *MeSH keyword-based search* in *PubMed*, the search is not performed in free-text. As such, the user does not need to think of word variations, synonyms, plural or singular forms or word endings (Chapman 2009). However, MeSH is restricted to *PubMed* (more specifically, to the *medical domain*); thus, employing such a query expansion phase using synonyms in other domains is vital.

To facilitate domain-independent synonym identification for tasks such as *query expansion*, this study used the DBpedia *property* (or *predicate*) *'is dbo:wikiPageRedirects of'* (Figure 7.24). This property enables the identification of synonyms using the *redirect pages* of the defined entity name (or concept). *Redirects* are a special type of article that originated in *Wikipedia*. They group equivalent concepts to ensure that only one article exists for a particular concept (Wang et al. 2009). In addition to *alternative terms*, redirects also handle abbreviations (e.g., *Insulin-like growth factor 1* vs. *IGF-1*), spelling variations (e.g., *Raynaud disease* vs. *Raynaud's disease*) and even singular/plural forms where necessary (e.g., *fish oil* vs. *fish oils*). Thus, the use of *'is dbo:wikiPageRedirects of'* can be considered a good approximation of synonyms. Moreover, like *'dct:subject'* and *'skos:broader'*, *'is dbo:wikiPageRedirects of'* is not domain-specific, since it is available in DBpedia entities in both *medical* (e.g., No. 13 in Tables 7.2 and C.1) and *non-medical* domains (e.g., No. 5 in Tables 7.3, C.2 and C.3).

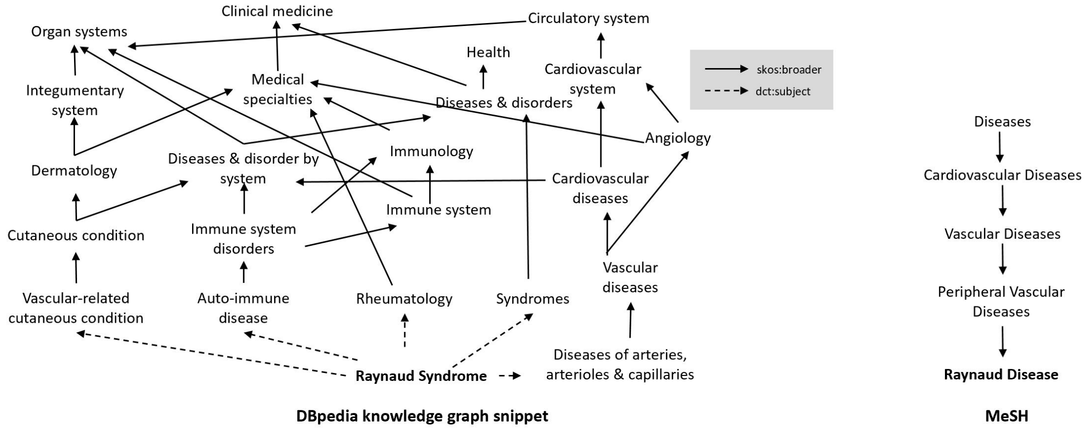


FIGURE 7.25: Structural difference between the two knowledge resources DBpedia and MeSH

7.6.5 Granularity Detection

The identification of a concept's granularity is mainly used in order to discard discipline-related stop words (a.k.a. *check-tags*). In other words, more granular concepts are typically used in the knowledge discovery process while removing more broad concepts (Swanson et al. 2006). However, performing such hierarchical-level semantic inferences are difficult using DBpedia's structure. The main reason for this is that DBpedia is not a tree, but a directed acyclic graph (Atzori & Dessi 2014). Therefore, it is not straightforward to perform hierarchical filtering similar to the tree structures which are used in the LBD domain, such as MeSH. Figure 7.25 illustrates the structural difference between DBpedia and MeSH using 'Raynaud syndrome' as an example.

Even though DBpedia is not a tree, the graph structure of DBpedia means that it can facilitate the integration of *graph theory* into the analysis. Thus, graph-related semantic inferences can be made using DBpedia. Consequently, this study attempts to verify whether using graph/network properties such as centrality can assist in approximating the granularity of concepts. In graph analysis, centrality measures are often used to capture topologically important nodes (a.k.a. hub nodes) based on the nodes' positions. These measures play a critical role in diverse types of networks (Oldham et al. 2019). There are many centrality measures that have been developed to gauge the importance of a node based on its characteristics (Srinivas & Velusamy 2015).

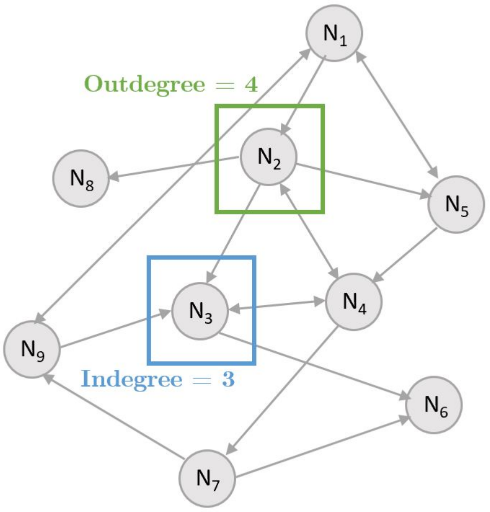


FIGURE 7.26: Illustrating the difference between in-degree centrality and out-degree centrality

This study employed *degree centrality*, which is one of the most frequently used centrality measures in network analysis studies (Valente et al. 2008). Degree centrality indicates the number of links attached to a node. In the context of directed networks, two types of degree centrality measures are used: *in-degree* and *out-degree* (Sharma & Surolia 2013). The first measure counts the number of links directed to a given node, while the latter measure counts the number of links that a given node directs to others. Figure 7.26 exemplifies the difference between these two measures of degree centrality.

This study followed the *in-degree centrality measure* to facilitate granularity detection of concepts. In the context of DBpedia, the *in-degree centrality measure* is the *in-degree resource (i.e., URIs/pages) link count* for a particular resource. The reason for this selection is that *in-degree centrality* measures the connectedness of a node in a network. Simply put, this measure facilitates the comparison of nodes in the network by considering the magnitude of their local neighbourhood. In the context of DBpedia, if a particular concept has a massive in-degree local neighbourhood, this means that the relevant node has served as a root to a large number of concepts (i.e., *hub nodes* (Oldham et al. 2019)). Thus, it is fair to assume that if a concept has a higher in-degree centrality, it is a less granular concept. Bearing this in mind, the current study assumes that concepts with excessively high in-degree resource links should represent discipline-related *check-tags*.

7.7 Evaluation

This section evaluates the suitability of the proposed solutions to circumventing the existing domain-dependent impediments in the LBD workflow. In this regard, one of the main objectives of our experiments in the validation of medical setting is to observe how well DBpedia resembles the most widely used LBD resource: *MeSH*. The main reason for pursuing this objective is that most existing LBD models are based on MeSH. Thus, if DBpedia can resemble MeSH with high precision, the LBD community does not need to perform substantial modifications to enable the portability of their models. In addition, the proposed solutions are validated in a non-medical setting by considering *computer science* as the test domain. Furthermore, this study also evaluates the suitability of *WordNet* for *synonym identification*, as proposed by Sebastian et al. (2017*b*).

The first part of this section outlines the experimental setup used in this chapter. The subsequent sections contain details of the results obtained for the proposed solutions, along with an extended discussion of their strengths and weaknesses in terms of circumventing the corresponding domain-dependent impediments. In addition to demonstrating cross-domain support of the proposed framework, the latter part of this section also demonstrates cross-lingual support for DBpedia, which will facilitate knowledge discovery in different publication languages.

7.7.1 Experimental Setup

The following five real-world test cases are used for the evaluation, as they are considered to be the *golden datasets* of the discipline. Further details on these selections are discussed in Chapter 3 of this thesis.

- *Fish-Oil (FO)* and *Raynaud's Disease (RD)* (Swanson 1986)
- *Magnesium (MG)* and *Migraine Disorder (MIG)* (Swanson 1988)
- *Somatomedin C (IGF1)* and *Arginine (ARG)* (Swanson 1990*a*)
- *Alzheimer's Disease (AD)* and *Indomethacin (INN)* (Smalheiser & Swanson 1996)
- *Schizophrenia (SZ)* and *Calcium-Independent Phospholipase A2 (PA2)* (Smalheiser & Swanson 1998)

Since the aforementioned test cases are all in the field of medicine, this study also utilises the study by Gordon et al. (2002) (the *only* available LBD study, which is directly relevant to the field of *computer science*) as a test case to demonstrate the suitability of DBpedia outside the medical domain. In their study, Gordon et al. (2002) attempted to explore novel areas using *genetic algorithms* as the start concept.

7.7.2 Concept Extraction (Discipline-related Terminology Extraction)

The purpose of this section is to verify the suitability of the proposed *discipline-related terminology extraction component* discussed in Section 7.6.2. An evaluation was performed, with reference to the *topic coverage* of local corpora from the five golden test cases, using MEDLINE's *title and abstract* fields. For the current experiments, parameters n and m of the two empirical rules (discussed in Section 7.6.2) were set to 9 and 5, respectively. To extract the MeSH concepts of the local corpora, the *MetaMap* tool[10] (Aronson & Lang 2010) was employed. The main domain concept was set to `dbc:Medicine`. The key intention of *topic coverage* was to verify how many topics extracted using MeSH were the same as those in DBpedia (denoted in equation 7.1).

$$topic\_coverage = \frac{MeSH\_topics \cap DBpedia\_topics}{MeSH\_topics} \times 100 \tag{7.1}$$

Nevertheless, it is not possible to perform direct string matching when comparing the topics extracted from two knowledge resources (*DBpedia* and *MeSH*) in order to calculate the numerator in equation 7.1. The main reason for this is that this evaluation setting compares topics from entirely different knowledge resources; thus, concepts can be in different lexical forms even if they denote similar meanings. For instance, consider the concept *Non-steroidal anti-inflammatory agent* in DBpedia, the relevant MeSH mapping of which is *Anti-Inflammatory Agents, Non Steroidal.* Note that in this scenario, in spite of the syntactic differences of these two concepts, the tokens within the concepts are also shuffled. To address this issue, this study utilises *fuzzy string matching* (more specifically, the *token set ratio* variant (Geel et al. 2012)) to measure topic similarity. The use of the *token set ratio* variant not only caters to syntactic variations, but also handles the issue of shuffled tokens. A topic is considered to be a match if it obtains more than 75% similarity. The topic coverage results for the five golden test cases are

[10] `https://metamap.nlm.nih.gov/MainDownload.shtml` - 2018 version

Table 7.5: Topic coverage of local corpora in the golden datasets

Test case	**MeSH topics**	**DBpedia topics**	**Similar topics**	**Topic coverage**
(1) FO-RD	2879	2960	2186	75.93%
(2) MG-MIG	16329	11601	9647	59.08%
(3) IGF1-ARG	14191	9709	7983	56.25%
(4) AD-INN	17672	14856	12406	70.20%
(5) SZ-PA2	15168	13556	11023	72.67%

outlined in Table 7.5. Even though *medicine* is only one of the many domains that DBpedia covers, the results in Table 7.5 indicate that DBpedia has a fair coverage of topics in MeSH, which is a specialised single-domain resource. The average topic coverage of DBpedia was 66.83%.

Similarly, to perform topic coverage in a non-medical setting, this study used all the terms mentioned in the only computer science LBD study conducted by Gordon et al. (2002) (i.e., Table 1, 2, 3, 4 and 5 in (Gordon et al. 2002)) as the main vocabulary. Subsequently, this study verified the extent to which the proposed *discipline-related terminology extraction component* identified these topics with respect to its main domain concept `dbc:Computer_science`. For this experiment, the same n and m values were utilised (i.e., 9 and 5, respectively). Through this process, 66.32% of the terms were identified as terminology related to computer science. Nevertheless, a few obvious terms in this dataset (such as *civil engineering* and *financial engineering*) may not be relevant to the computer science domain. Thus, it is impossible to get 100% topic coverage for this dataset. Even though computer science is also one of the many domains that DBpedia covers, overall, it demonstrates a fair topic coverage in terms of identifying topics related to computer science.

Note that to switch across domains, our proposal only needs to change its main domain concept name (e.g., '`dbc:Medicine` in the medical domain, and `dbc:Computer_science` in the computer science domain). The remaining computations performed as part of the proposed *discipline-related terminology extraction component* will automatically adhere to the selected main domain. This fulfils our objective of achieving portability with little or no cost (i.e., *degree of portability*). Moreover, the proposed solution facilitates knowledge discovery in a wide spectrum of domains. In addition to *medicine* and *computer science*, which were evaluated in this section, our solution could also be integrated into numerous other domains such as *sociology*, *psychology*, *geography*, *economics*, *anthropology*, *philosophy*, *law*, *languages and literature*, *history*, *arts*, *social work*, *biology*, *chemistry*, *earth science*, *space science*, *physics*, *mathematics*, *business*, *engineering* (including *chemical engineering*, *civil engineering*, *educational technology*, *electrical engineering*, *material science and engineering*, *mechanical engineering*) etc. due to the prominence of DBpedia as a cross-domain resource.

In addition to facilitating knowledge discovery in a selected single main domain, the proposed component also adheres to the use of multiple domains in single knowledge discovery (something which *none* of the existing LBD models are capable of). For instance, consider a researcher who wishes to explore knowledge in both *medicine* and *computer science* at the same time. In such a situation, the user can select both the main domain concepts (`dbc:Medicine` and `dbc:Computer_science`) in order to retain concepts from both disciplines. In this way, the user gets the opportunity to discover latent novel knowledge not only from a single main domain, but also across multiple domains. Enabling broader knowledge discovery in this way is crucial for the development of interdisciplinary research (such as bioinformatics and medical informatics). The aforementioned proof of concept enables *interdisciplinarity* (or *generalisability*) during the LBD knowledge discovery process.

Prior to this study, performing cross-disciplinary concept extraction was a long-term open issue in the LBD field, where the non-medical LBD studies were *time-intensive*, since they were mostly performed using manual concept searches. Nevertheless, with the integration of the proposed *discipline-related terminology extraction component*, these non-medical LBD studies can be improved not only in terms of *time*, but also *replicability*, *reliability*, *automation* and *easy transition.* The topic coverage results reported as part of this section may be further improved by fine-tuning the n and m parameters defined by the two empirical rules. The next stage of this component will be to integrate Machine Learning (ML) techniques. In this regard, one could use several n and m values (e.g., n = $\{n_1, ..., n_x\}$, m = $\{m_1, ..., m_y\}$) to determine C(n) and C(m) (e.g., C(n) = {C(n_1), ..., C(n_x)}, C(m) = {C(m_1), ..., C(m_y)}) as the features of a ML model designed to identify the most prominent n and m values (i.e., the most important features). Subsequently, the selected features could be utilised in the ML setting to determine discipline-related terminology with further enhanced precision.

7.7.3 Semantic Type Filtering

In this setting, this study first closely inspected semantic relationships of the medical topics from the golden datasets, in order to compare the consistency of *semantic types* between DBpedia and MeSH. Consistency of semantic types between the two knowledge resources provide evidence that similar concepts can be retrieved through semantic type filtering using the two properties *dct:subject* and *skos:broader* (as discussed in Section 7.6.3) with a closer precision to MeSH.

For instance, consider Figure 7.27, which illustrates how each semantic type in a MeSH tree (i.e., *oils*, *lipids*, *chemicals and drugs*) resembles a DBpedia knowledge graph for the topic *fish oil*. Overall, DBpedia covers almost all *semantic types* in MeSH for the topic *fish oil* (nevertheless, it does so using different wordings for some semantic types, which is inevitable given that DBpedia and MeSH are completely different knowledge resources). In addition, this study observed that DBpedia has a set of fine-grained semantic type groupings compared to MeSH. The main reason for this could be that DBpedia is not limited to a single domain, enabling it to encode the semantic relationships of a concept from a wider perspective compared to single-domain knowledge resources like MeSH. In the example in Figure 7.27, the topic *fish oil* not only interacts with semantic types such as *medical treatments* and *chemical compounds* (as is the case with MeSH), but also with a wide variety of other semantic types, from *nutrition* to *cooking*, and even to *fish industry*. In a nutshell, while demonstrating similarities with MeSH in terms of semantic types, DBpedia contains much additional knowledge encoded in a wider spectrum, due to its cross-domain support (a strength that is lacking in other knowledge sources). Such circumstantial semantic groupings in DBpedia enable it to perform subtle semantic reasoning beyond the existing LBD tasks such as *semantic type filtering*. This study observes similar conclusions (discussed above) for most of the remaining topics from the golden test cases, as illustrated from Figures 7.28, 7.29, 7.30, 7.31, 7.32, 7.33, 7.34, 7.35 and 7.36.

Overall, DBpedia displayed good coverage of the *semantic types* that exist in MeSH. Therefore, as in Section 7.6.2, the two properties *dct:subject* and *skos:broader* can be used

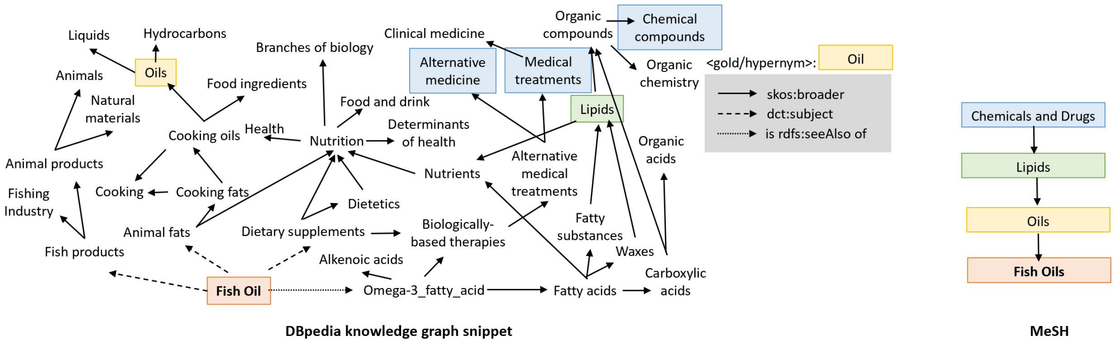


FIGURE 7.27: Comparison of semantic types for the topic *fish oil*

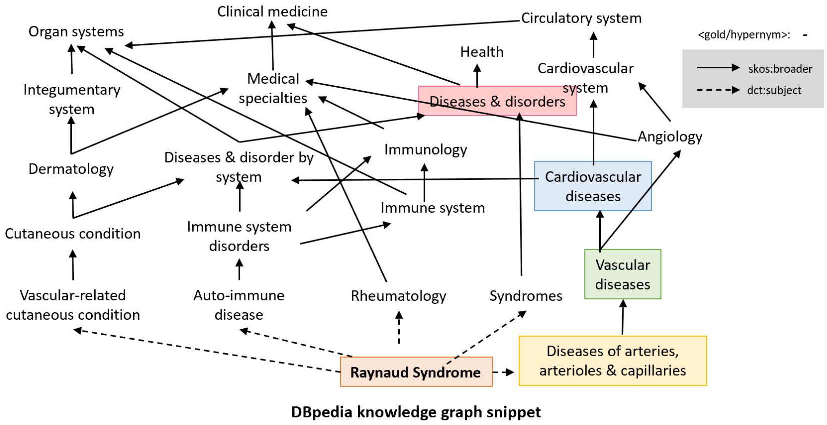


FIGURE 7.28: Comparison of semantic types for the topic *Raynaud disease*

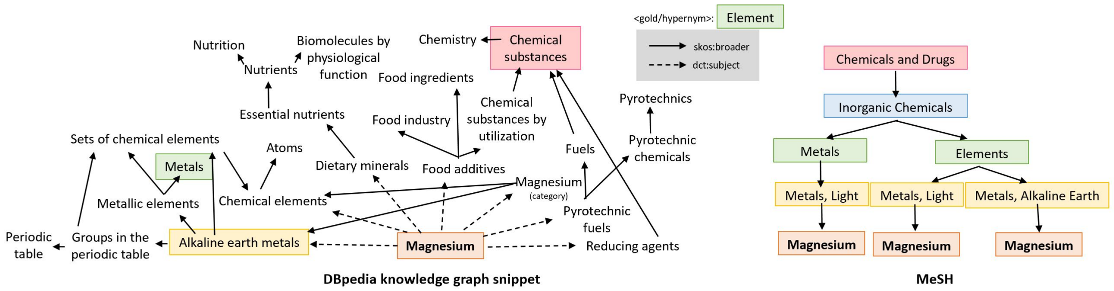


FIGURE 7.29: Comparison of semantic types for the topic *Magnesium*

to retrieve concepts under a semantic type of interest, with closer precision to MeSH. Moreover, this study observed that the two empirical observations mentioned in Section 7.6.2 were also valid in most of the topics depicted from Figure 7.27 to 7.36. The two observations were, firstly, that semantic types are usually more granular than the main domain concept, meaning that n should be lower than the *discipline-related terminology* component, and secondly, semantic types usually demonstrate a single characteristic of the concept, since limited immediate categories are connected with each semantic type;

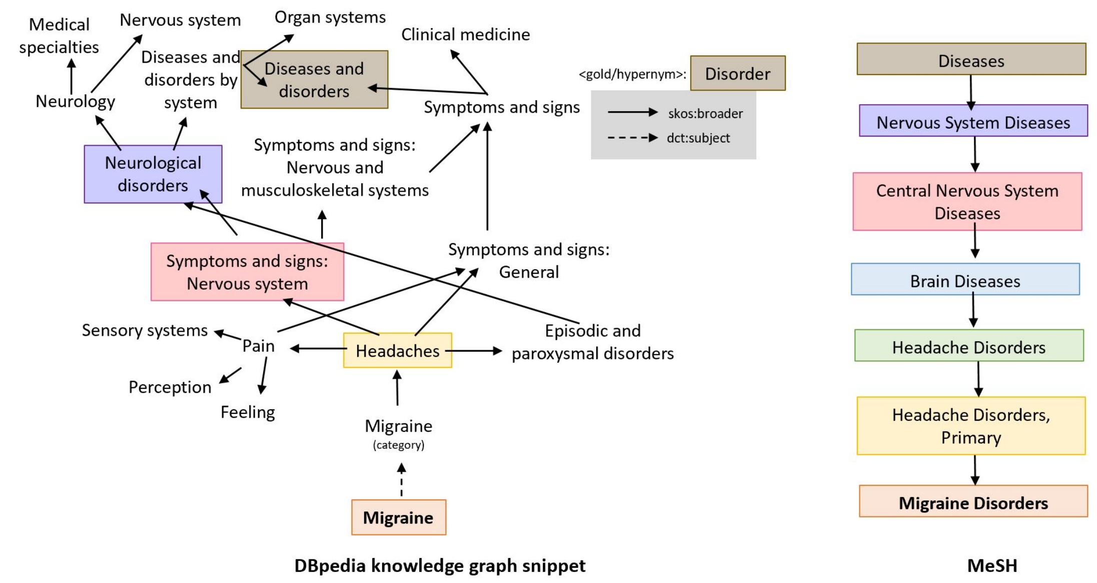


FIGURE 7.30: Comparison of semantic types for the topic *Migraine*

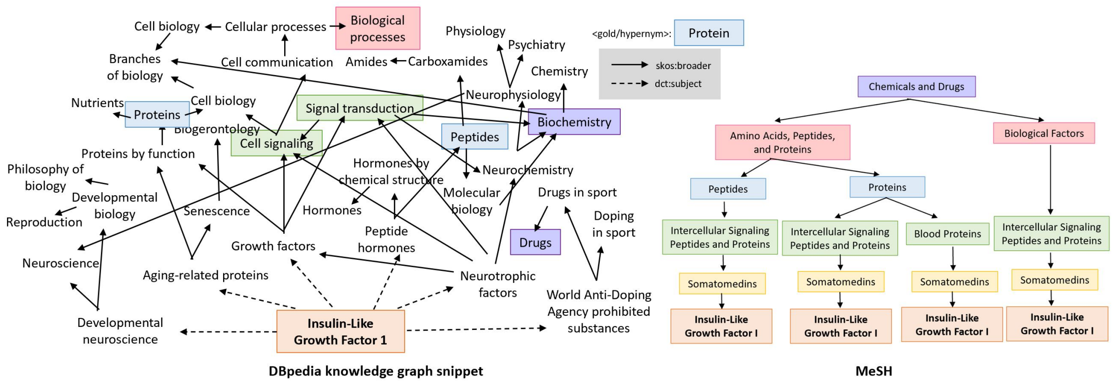


FIGURE 7.31: Comparison of semantic types for the topic *Insulin-like growth factor 1*

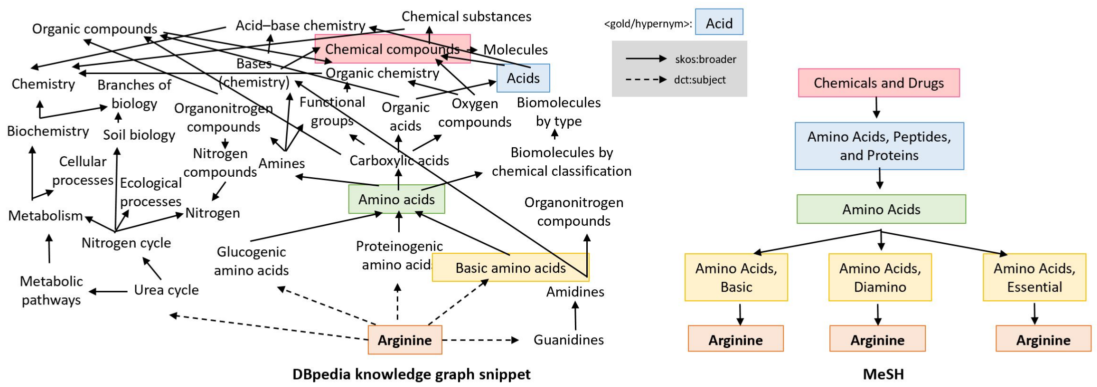


FIGURE 7.32: Comparison of semantic types for the topic *Arginine*

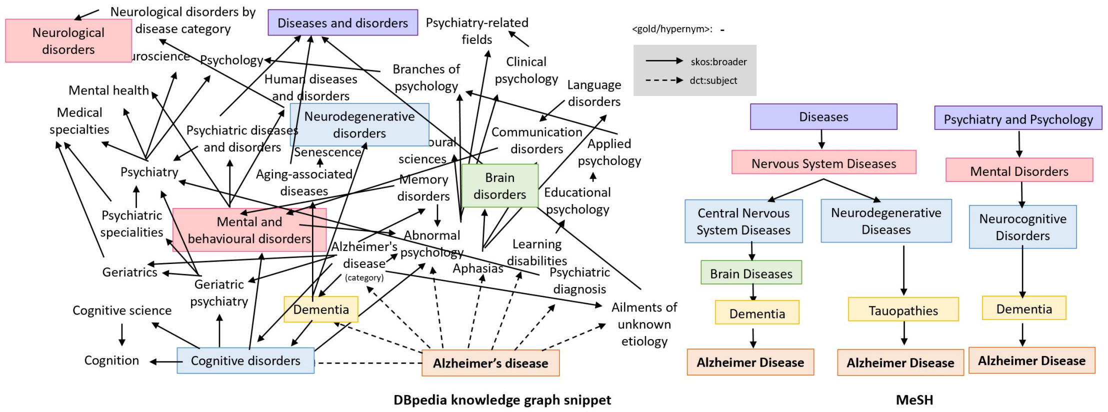


FIGURE 7.33: Comparison of semantic types for the topic *Alzheimer's disease*

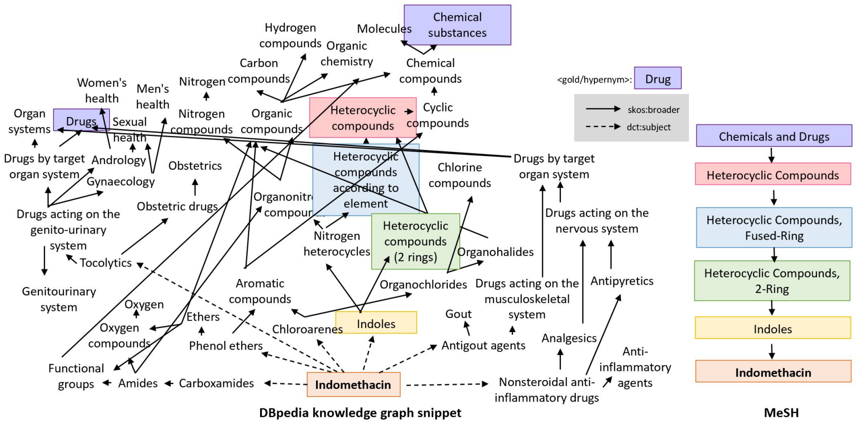


FIGURE 7.34: Comparison of semantic types for the topic *Indomethacin*

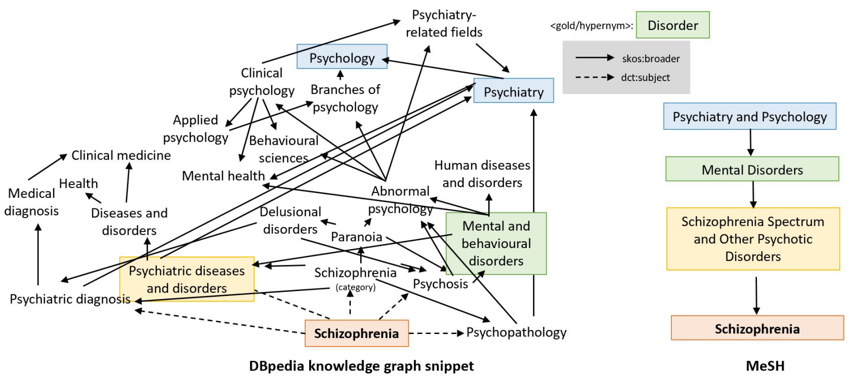


FIGURE 7.35: Comparison of semantic types for the topic *Schizophrenia*

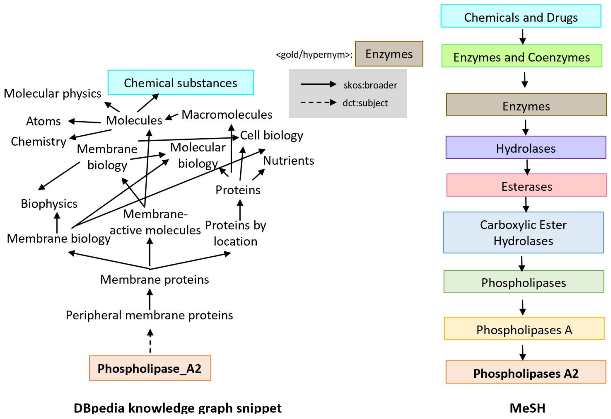


FIGURE 7.36: Comparison of semantic types for the topic *Phospholipase A2*

thus, C(n) should be lower than the *discipline-related terminology* component.

To demonstrate the semantic types in a non-medical setting, this study first considered the start concept of the study by Gordon et al. (2002): *genetic algorithm* (representing the computer science domain). Figure 7.37 denotes the semantic relationships of *genetic algorithm* in DBpedia's topic-category structure in comparison with *ACM CCS*. As in the medical domain, DBpedia has a fine-grained (a.k.a. *finer granularity*) set of semantic type groupings in non-medical settings too. Note how the genetic algorithm not only connects with semantic types such as *algorithms* and *bioinformatics*, but also with other semantic types involving *artificial intelligence*, *mathematical optimisation* and *search engines*. This study also compared the semantic type groupings of DBpedia using prominent controlled vocabularies from physics (e.g., Figure 7.38), mathematics (e.g., Figure 7.39) and economics (e.g., Figure 7.40). It is evident that DBpedia captures semantic type grouping in greater detail compared to the prominent controlled vocabularies in the corresponding domains. This provides further evidence for the depth and breadth of the semantic types that DBpedia encompasses.

As with our previous component (i.e., *discipline-related terminology extraction*), this component involved negligible costs in the process of establishing portability (i.e., denoting the *degree of portability*). This is because the only requirement when transitioning across domains in the knowledge discovery process was to set DBpedia's semantic types relevant to each domain. Moreover, in addition to the single domain knowledge discovery (through the selection of one of the numerous domains that DBpedia supports), this component also meets our objective of *interdisciplinarity* (or *generalisability*) by supporting *several domains* at once in the knowledge discovery process. For example,

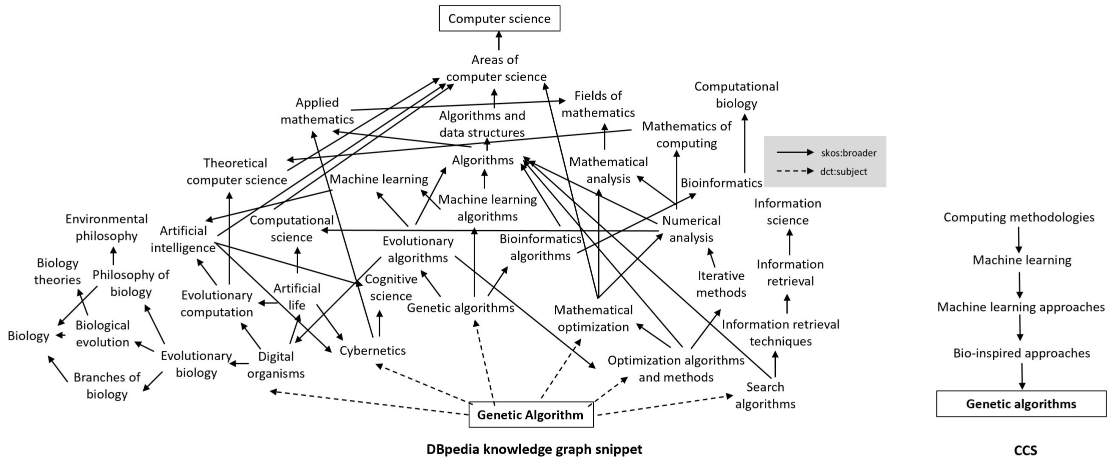


FIGURE 7.37: Semantic types of the topic *genetic algorithm* from the computer science domain (also compared with ACM CSS)

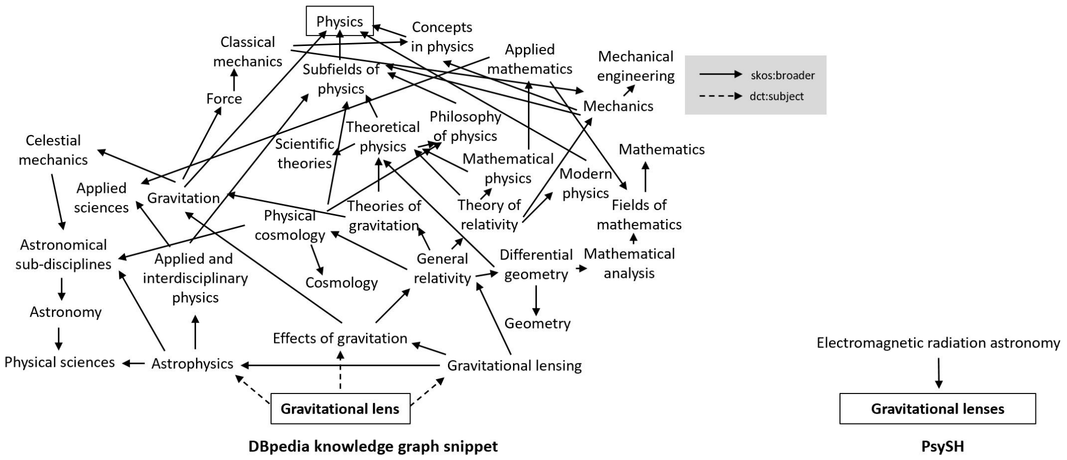


FIGURE 7.38: Semantic types of the topic *gravitational lens* from the physics domain (also compared with PsySH)

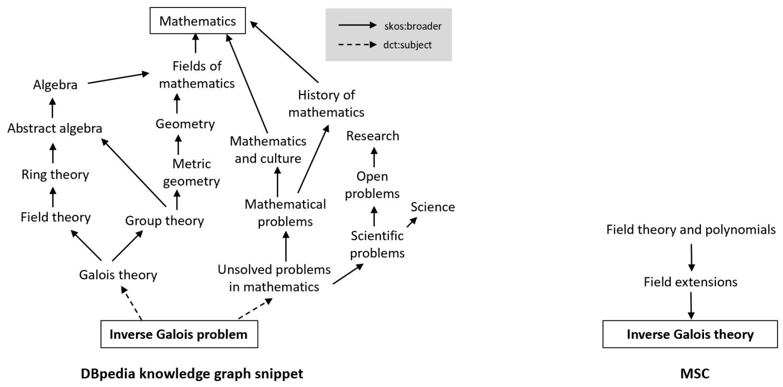
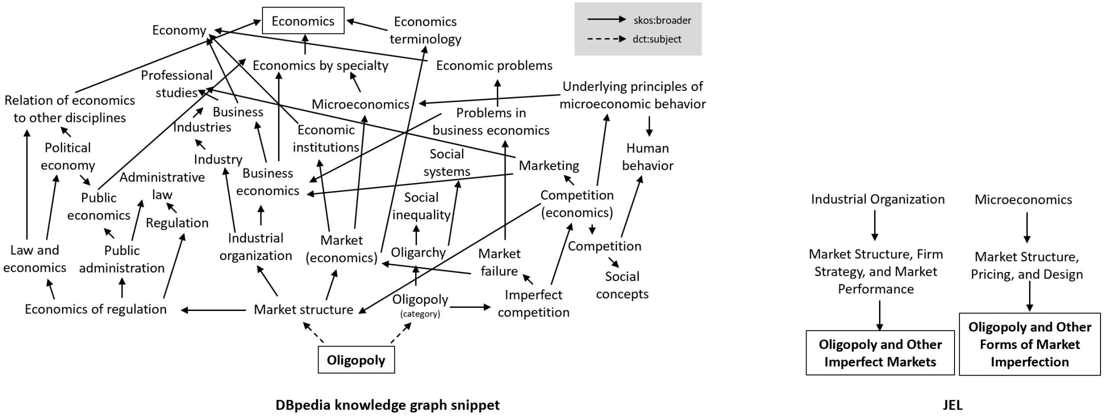


FIGURE 7.39: Semantic types of the topic *inverse galois problem* from the mathematics domain (also compared with MSC)



DBpedia knowledge graph snippet

JEL

FIGURE 7.40: Semantic types of the topic *oligopoly* from the economics domain (also compared with JEL)

consider a situation where a researcher wishes to identify new ways to combine optimisation techniques for wave energy converter placement. In this instance, the user can select the two semantic types `dbc:Optimization_algorithms_and_methods` and `dbc:Sustainable_technologies` to perform the knowledge discovery in an interdisciplinary and generalisable manner.

7.7.4 Synonym Identification

To evaluate possibility of performing synonym identification through the DBpedia predicate, *'is dbo:wikiPageRedirects of'*, this study used two settings: *synonym coverage* and *literature coverage*. The purpose of the first setting was to quantitatively evaluate the synonym coverage of *DBpedia* and *MeSH*. Note that this evaluation setting also used *WordNet*, as proposed by Sebastian et al. (2017*b*).

Table C.4 outlines the results of the *synonym coverage* setting using the main topics from the golden test cases (while Table 7.6 presents selected results from Table C.4). The synonym coverage results showed that DBpedia had higher coverage of synonyms than MeSH and WordNet. Moreover, DBpedia synonyms also include spelling variations (which are not available in MeSH), making DBpedia a suitable resource for the recent LBD research that incorporates non-traditional data sources, such as *Twitter* (Bhattacharya & Srinivasan 2012). Overall, WordNet displayed the least coverage of synonyms. Moreover, some of the topics in the golden datasets were not found in WordNet. This could be due to the fact that WordNet typically rich in general English terminology and lacking in scientific topics. Even though the use of WordNet is a good starting point for LBD, our evaluation reveals that it is unsuitable in some respects with regard to the development of a high-precision *interdisciplinary LBD framework*.

TABLE 7.6: Qualitative evaluation of synonym coverage (includes the redirects that are directly linked to the main Wikipedia page, i.e., redirects with 'no anchor')

Test case No.	**Topic**	**Resource**	**Synonyms**
(1)	FO	MeSH	Fish Oils, Fish Liver Oils, Fish Oil
		DBpedia	Fish oil, Fish oils, Fish-oil, Lovanza, Marine oil, Fish liver oils
		WordNet	Fish oil, Fish-liver oil
	RD	MeSH	Raynaud Disease, Hereditary Cold Fingers, Raynaud Phenomenon, Raynaud's Disease
		DBpedia	Raynaud syndrome, Raynaud's disease and Raynaud's phenomenon, Reynaud's, Reynaud's disease, Raynaud's disease, Raynaud phenomenon, Reynaud's phenomenon, Raynauds disease, Reynaud's Disease, Raynaud's disorder, Intermittent arterial vasospasm, Raynaud's Disease, Raynaud's syndrome, Raynaud's phenomenon, Raynaud's disease/phenomenon, Raynaud disease, Raynauds, Raynauds Syndrome, Raynauld's syndrome, Raynauld syndrome, Reynaud's phenomenon, Primary Raynaud's phenomenon, Raynaud's Phenomenon, Raynaud's Syndrome, Reynaud's Syndrome, Reynaud's syndrome, Primary raynaud's phenomenon, Secondary raynaud's phenomenon, Raynaud's Syndrome, Raynaud's
		WordNet	Raynaud's sign, Acrocyanosis

**Results pertaining to the remaining test cases can be found in Table C.4*

As in other sections, this study also validated several randomly chosen computer science terminology that was used in the study of Gordon et al. (2002) to assess *synonym coverage* in a non-medical setting. Table C.5 outlines the results obtained through this analysis[11] (while Table 7.7 presents selected results from Table C.5). As in the medical setting, DBpedia displayed the highest coverage of synonyms in this setting

[11] *CCS*, *MSC* and *JEL* do not have synonymous terms; thus, they were not included in the table. Even though *PsySH* supports synonymous terms (i.e., *alternate labels*: `https://physh.aps.org/about`), the selected physics concept does not have any synonyms recorded in *PsySH*. Thus, it was not included in the table.

too. Furthermore, this study observed that MeSH contains synonyms for some of these non-medical terms, though these are not comprehensive. As in the previous setting, WordNet displayed the lowest coverage of synonyms. Overall, the results in the non-medical setting consistently indicated the suitability of DBpedia for accomplishing our ultimate goal of portability. Since DBpedia encapsulates all the domains into a single uniform view, the costs involved in the cross-domain transitions are almost zero (i.e., *degree of portability*).

Table 7.7: Qualitative evaluation of synonym coverage in non-medical settings

Topic	**Resource**	**Synonyms**
Genetic algorithms	MeSH	–
	DBpedia	Genetic algorithm, Genetic algorithms, Darwinian algorithm, GATTO, Building block hypothesis, Theory of genetic algorithms, Genetic Algorithm, Genetic Algorithms, GEGA, Genethc algorithm
	WordNet	–
Pattern recognition	MeSH	Automated Pattern Recognition, Pattern Recognition System
	DBpedia	Pattern recognition, Pattern analysis, Visual pattern recognition, Pattern Recognition, Machine pattern recognition, Pattern recognition and learning, Pattern-recognition, Pattern Recognition and Learning, Pattern recognition (machine learning)
	WordNet	–

**Results pertaining to the remaining concepts can be found in Table C.5*

The following evaluation setting estimates *local corpus coverage* using *expanded search queries* as a use case (discussed in Section 7.6.4). For this purpose, the study used the *Web of Science* literature database. The main reason for not using MEDLINE is that MeSH terms are indexed in MEDLINE; thus, the database does not necessarily showcase the query expansion ability of MeSH in non-medical settings. The *literature coverage* results obtained through expanded queries are outlined in Table 7.8. Overall, it is evident that DBpedia also has high coverage of local corpora compared to MeSH, due to its richness of synonyms. In most of the situations, DBpedia contains all the records from MeSH as its subset (i.e., $\frac{M \cap D}{M}$ %). Since the synonym coverage of MeSH outside the medical domain is poor or non-existent in most situations, this study did not compare the coverage of local corpora in non-medical settings.

Table 7.8: Quantitative evaluation of literature coverage

Test case No.	**Topic**	**MeSH (M)**	**DBpedia (D)**	**M∩D**	$\frac{M \cap D}{M}$ %
(1)	FO	328	366	328	100%
	RD	262	1100	262	100%
(2)	MG	22780	24192	22780	100%
	MIG	3409	3520	3406	99.91%
(3)	IGF1	1838	1947	1838	100%
	ARG	6863	6966	6863	100%
(4)	AD	8972	21034	8259	92.05%
	INN	12681	12680	12671	99.92%
(5)	SZ	29240	41365	28827	98.59%
	PA2	6185	6086	5946	96.14%

7.7.5 Granularity Detection

This section evaluates whether the use of DBpedia's *in-degree resource links* approximates the identification of check-tags, as discussed in Section 7.6.5. For this purpose, the study incorporated the same concepts used by Xun et al. (2017) to empirically observe the in-degree resource link counts of each concept outlined in Table 7.9. It is clear that *check-tags* typically have a higher number of in-degree page links (i.e., *hub nodes* in knowledge networks) compared to informative, granular terms.

Subsequently, this study attempted to regenerate the *MeSH level 1 and 2 topics* (typically considered as check-tags; discussed in Chapter 2) using in-degree resource link counts. The in-degree link count threshold was set to 200. That is, terms with a >200 in-degree link count were considered as check-tags. Through this experiment, this study was able to identify 29.63% of the terms in MeSH level 1 and 2 topics as *check-tags*. When the in-degree link count threshold was set to 200, this also meant that the informative terms were the terms that had an in-degree link count of ⩽200. In this way, this study recovered 100% of the main topics from the five golden datasets as informative terms.

To validate this component in a non-medical domain, a similar analysis was performed using all the terms in the only available computer science LBD study (Gordon et al. 2002). In this non-medical setting, the study was able to identify 97.80% of terms as informative terms with the same threshold (i.e., an in-degree resource link count of ⩽200). As in other proposed components of this study, this component not only

Table 7.9: Check-tags identification

Concept	**Check-tag label**	**In-degree link count**
humans	check-tag	327
animals	check-tag	190301
female	check-tag	392
male	check-tag	506
fish oils	informative	8
Raynaud disease	informative	30
blood viscosity	informative	8
epoprostenol	informative	12

supports our idea of portability across numerous domains, but also does not involve any costs when switching across domains (denoting the *degree of portability*).

Overall, the in-degree link count is a good starting proxy from which to detect the granularity of a concept. However, our results also suggest the importance of further enhancing this component to detect check-tags with broad coverage and high precision. One of the major differences between DBpedia and MeSH that was observed in this study is that DBpedia is a *directed acyclic graph*, whereas MeSH is a tree. Thus, performing *hierarchical-level semantic inferences* is difficult in DBpedia. Nevertheless, it is possible to obtain some rough approximation using network properties in DBpedia's topic structure, as this study demonstrated. This component could be further enhanced by integrating multiple other network measures, such as *PageRank* and *structural holes*, which capture different other perspectives of nodes in the graph. For instance, consider Figure 7.41, which exemplifies how this structural transition of DBpedia could be performed. The next stage of this component is to integrate ML techniques. In essence, one could use the most prominent network properties extracted from DBpedia's graph structure in a multi-class classification problem in which each class denoted the level of a term in the MeSH tree. This would further enhance this component in terms of approximating a tree structure from DBpedia (equivalent to the structure of MeSH) to support hierarchical semantic inferences.

7.7.6 Cross-lingual Support

The purpose of this section is to demonstrate the *cross-lingual support* of DBpedia, which facilitates knowledge discovery not only across domains but also across publication

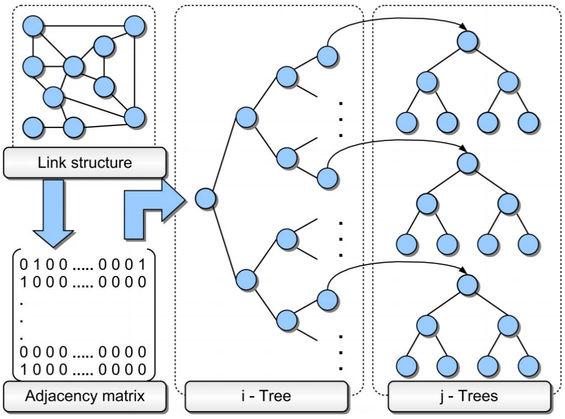


FIGURE 7.41: Converting DBpedia link structure to a tree (Nakayama et al. 2007) (e.g., like that of MeSH)

TABLE 7.10: Basic statistics on localised DBpedia editions (Lehmann et al. 2015)

Language	Inst. LD all	Inst. CD all	Inst. with MD CD	Raw Prop. CD	Map. Prop. CD	Raw Statem. CD	Map. Statem. CD
English (en)	3,769,926	3,769,926	2,359,521	48,293	1,313	65,143,840	33,742,015
German (de)	1,243,771	650,037	204,335	9,593	261	7,603,562	2,880,381
French (fr)	1,197,334	740,044	214,953	13,551	228	8,854,322	2,901,809
Italian (it)	882,127	580,620	383,643	9,716	181	12,227,870	4,804,731
Spanish (es)	879,091	542,524	310,348	14,643	476	7,740,458	4,383,206
Polish (pl)	848,298	538,641	344,875	7,306	266	7,696,193	4,511,794
Russian (ru)	822,681	439,605	123,011	13,522	76	6,973,305	1,389,473
Portuguese (pt)	699,446	460,258	272,660	12,851	602	6,255,151	4,005,527
Catalan (ca)	367,362	241,534	112,934	8,696	183	3,689,870	1,301,868
Czech (cs)	225,133	148,819	34,893	5,564	334	1,857,230	474,459
Hungarian (hu)	209,180	138,998	63,441	6,821	295	2,506,399	601,037
Korean (ko)	196,132	124,591	30,962	7,095	419	1,035,606	417,605
Turkish (tr)	187,850	106,644	40,438	7,512	440	1,350,679	556,943
Arabic (ar)	165,722	103,059	16,236	7,898	268	635,058	168,686
Basque (eu)	132,877	108,713	41,401	2,245	19	2,255,897	532,709
Slovene (sl)	129,834	73,099	22,036	4,235	470	1,213,801	222,447
Bulgarian (bg)	125,762	87,679	38,825	3,984	274	774,443	488,678
Croatian (hr)	109,890	71,469	10,343	3,334	158	701,182	151,196
Greek (el)	71,936	48,260	10,813	2,866	288	206,460	113,838

LD = Localised data sets; all = Overall number of instances in the data set, including instances without infobox data; CD = Canonicalized data sets; MD = Number of instances for which mapping-based infobox data exists; Raw Properties = Number of different properties that are generated by the raw infobox extractor; Mapping Properties = Number of different properties that are generated by the mapping-based infobox extractor; Raw Statements = Number of statements (facts) that are generated by the raw infobox extractor; Mapping Statements = Number of statements (facts) that are generated by the mapping-based infobox extractor.

languages. To date, DBpedia consists of more than 130 localised versions (Chiarcos & Pareja-Lora 2019) that have been extracted from corresponding language editions in Wikipedia (Lehmann et al. 2015). Table 7.10 summarises the basic statistics on a few localised DBpedia editions in release 3.8 (Lehmann et al. 2015). Overall, the *English* version of DBpedia includes more instances than other language editions. The second and third largest localised editions are *German* and *French*, respectively.

Table 7.11 illustrates how DBpedia entities can be mapped to its localised versions for

Table 7.11: Mapping of the main topics from golden datasets to localised DBpedia resources

Main Topic	English (en) Edition	French (fr) Edition
FO	http://dbpedia.org/resource/Fish_oil	http://fr.dbpedia.org/resource/Huile_de_poisson
RD	http://dbpedia.org/resource/Raynaud_syndrome	http://fr.dbpedia.org/resource/Syndrome_de_Raynaud
MG	http://dbpedia.org/resource/Migraine	http://fr.dbpedia.org/resource/Migraine
MIG	http://dbpedia.org/resource/Magnesium	http://fr.dbpedia.org/resource/Magn%C3%A9sium (Magnésium)
IGF1	http://dbpedia.org/resource/Arginine	http://fr.dbpedia.org/resource/Arginine
ARG	http://dbpedia.org/resource/Insulin-like_growth_factor_1	http://fr.dbpedia.org/resource/IGF-1
AD	http://dbpedia.org/resource/Alzheimer's_disease	http://fr.dbpedia.org/resource/Maladie_d'Alzheimer
INN	http://dbpedia.org/resource/Indometacin	http://fr.dbpedia.org/resource/Indom%C3%A9tacine (Indométacine)
SZ	http://dbpedia.org/resource/Schizophrenia	http://fr.dbpedia.org/resource/Schizophr%C3%A9nie (Schizophrénie)
PA2	http://dbpedia.org/resource/Phospholipase_A2	http://fr.dbpedia.org/resource/Phospholipase_A2

the main topics from the golden test cases using DBpedia's *French* language edition. The corresponding language editions of DBpedia can be located using the *‘owl:sameAs’*[12] predicate. Overall, the domain-independent solutions proposed in this study are compatible with other language editions of DBpedia, since they employ the same predicates and structures, as in the English language edition. This enables the portability of the proposed solutions across numerous publication languages with negligible costs.

To demonstrate DBpedia's cross-lingual support outside the medical domain (as in other evaluation settings), the terms used in the study of Gordon et al. (2002) were employed. Table 7.12 outlines the corresponding mapping of DBpedia entities from the English edition with entities from the French edition. This study observed that the term *text retrieval* does not have a corresponding mapping in the French edition. This may be due to the reduced content in the French edition of DBpedia (relative to the English edition),

[12] this is one of OWL (Web Ontology Language) properties: https://www.w3.org/TR/owl-ref/

TABLE 7.12: Localised DBpedia resource mapping in the computer science domain

Main Topic	English (en) Edition	French (fr) Edition
Genetic algorithm	http://dbpedia.org/resource/Genetic_algorithm	http://fr.dbpedia.org/resource/Algorithme_g%C3%A9n%C3%A9tique (Algorithme génétique)
Pattern recognition	http://dbpedia.org/resource/Pattern_recognition	http://fr.dbpedia.org/resource/Reconnaissance_de_formes
Virtual reality	http://dbpedia.org/resource/Virtual_reality	http://fr.dbpedia.org/resource/R%C3%A9alit%C3%A9_virtuelle (Réalité virtuelle)
Reinforcement learning	http://dbpedia.org/resource/Reinforcement_learning	http://fr.dbpedia.org/resource/Apprentissage_par_renforcement
Text retrieval	http://dbpedia.org/resource/Document_retrieval	-
Cluster analysis	http://dbpedia.org/resource/Cluster_analysis	http://fr.dbpedia.org/resource/Partitionnement_de_donn%C3%A9es (Partitionnement de données)
Image segmentation	http://dbpedia.org/resource/Image_segmentation	http://fr.dbpedia.org/resource/Segmentation_d'image
Speech recognition	http://dbpedia.org/resource/Speech_recognition	http://fr.dbpedia.org/resource/Reconnaissance_automatique_de_la_parole
Signal processing	http://dbpedia.org/resource/Signal_processing	http://fr.dbpedia.org/resource/Traitement_du_signal
Machine vision	http://dbpedia.org/resource/Machine_vision	http://fr.dbpedia.org/resource/Vision_industrielle

as outlined in Table 7.10. Overall, the switch across publication languages in the non-medical setting is also feasible at little or no cost, due the proposed solutions' support for predicates and structures in non-English language editions of DBpedia. Therefore, the use of proposed solutions in this study not only supports the transition *across domains*, but also *across publication languages*.

7.8 Summary

DBpedia is one of the dominant Semantic Web data sources, comprising data from Wikipedia as well as a broad range of additional knowledge gained by interlinking with other knowledge bases. The main influence for selecting DBpedia in this study is that it adhered to all the defined criteria of portability in the LBD setting. This chapter systematically compared the proposed solutions for existing domain-dependent impediments with the ultimate objective of developing a portable LBD framework to offer the benefits of LBD models beyond medicine. Overall, the proposals of this study resemble the knowledge inferences performed using *MeSH* with high precision. In some instances (e.g., synonym identification), the proposed solutions were superior to MeSH even in the medical setting itself. This ensures that existing LBD models that are mostly based on MeSH do not require to perform substantial modifications to enable the portability of their models through the integration of our solutions. Moreover, evaluations performed in non-medical settings also demonstrated the validity and reliability of the proposed solutions.

The overarching goal of this study was to develop an interdisciplinary (or generalisable) LBD framework that enables the portability of the LBD workflow in new environments at little or no cost. To assess the extent to which this goal was accomplished, the proposed solutions were validated in the context of *new environments of portability* (i.e., cross-domain and cross-lingual support; summarised in Table 7.13) and *degree of portability* (i.e., little or no cost in portability; summarised in Table 7.14). Further details on these conclusions in Tables 7.13 and 7.14 were discussed in Section 7.7. Solving the domain-dependent impediments of the LBD workflow through the proposed solutions enables a whole new level of knowledge discovery to extend LBD research beyond the medical domain, where it is still in a nascent stage.

7.8.1 Major Contributions

As a result of the portability research conducted in this chapter, this thesis presents several new insights on LBD. The major contributions of this study are outlined below. These contributions are discussed further in Chapter 8.

Table 7.13: Extent to which proposed solutions support portability to new environments

Proposed Solution	**Cross-domain support in medical setting**	**Cross-domain support in non-medical settings**	**Cross-lingual support in medical setting**	**Cross-lingual support in non-medical settings**
Disciple-related terminology extraction	✓	✓	✓	✓
Semantic type filtering	✓	✓	✓	✓
Synonym identification	✓	✓	✓	✓
Granularity detection	✓	✓	✓	✓

where ✓ denotes if the requirement is met

Table 7.14: Extent to which proposed solutions support portability in terms of associated costs

Proposed Solution	**Costs in cross-domain support**	**Costs in cross-lingual support**
Disciple-related terminology extraction	*Cost category:* Negligible *Details:* Select the domain(s) of interest (e.g., `dbc:Medicine`, `dbc:Computer_science`) in order to facilitate knowledge discovery	*Cost category:* Negligible *Details:* Set the corresponding language edition(s) required. The default setting would be English.
Semantic type filtering	*Cost category:* Negligible *Details:* Select the semantic type(s) of interest (e.g., `dbc:Optimization_algorithms_and_methods`, `dbc:Sustainable_technologies`) in order to facilitate knowledge discovery	*Cost category:* Negligible *Details:* Set the corresponding language edition(s) required. The default setting would be English.
Synonym identification	*Cost category:* None *Details:* –	*Cost category:* Negligible *Details:* Set the corresponding language edition(s) required. The default setting would be English.
Granularity detection	*Cost category:* None *Details:* –	*Cost category:* Negligible *Details:* Set the corresponding language edition(s) required. The default setting would be English.

- Being the first LBD study that proposes a comprehensive portable LBD framework to support knowledge discovery in a cross-domain and cross-lingual manner.

- Being the first LBD study to demonstrate interdisciplinarity (or generalisability) through the combination of multiple domains in a single knowledge discovery, with negligible costs in the transitions between (or among) domains.

- Being the first study to introduce DBpedia to the LBD discipline, providing opportunity to perform multifarious semantic inferences using unstructured text in a domain-agnostic and language-agnostic manner.

- Observing that the proposed solutions displayed similarities to (and sometimes outperformed) the commonly used LBD resource *MeSH*, meaning that the LBD community will be able to integrate the proposed solutions into their LBD models without substantial modifications, which will facilitate LBD research beyond medicine.

Chapter 8

Conclusions and Future Work

8.1 Introduction

To bring about advancement in a scientific field, researchers need to explore new knowledge, creatively combining observations and existing published knowledge (Foster et al. 2015, Rzhetsky et al. 2015). This requires them to keep abreast of existing and emerging scientific knowledge (Jha, Xun, Wang & Zhang 2019). However, the tremendous influx of research publications, and their easy accessibility via digital libraries, have resulted in information overload. This has made it harder for scientists to form connections between their own work and the research output from other disciplines (Xun et al. 2017, Guo et al. 2020, Su & Zhou 2009). One fundamental property of influential research is that it is richly interconnected with ideas from a broad range of domains (i.e., *divergent thinking scientific discovery*) (Chen 2016, Lavrač et al. 2020). Given the sheer volume of scientific knowledge, there is a need for models that are capable of discovering novel knowledge areas that complement scientists' niche specialisations. With this in mind, *Literature-Based Discovery (LBD)* research aims to detect hitherto undiscovered, but critical cross-silo connections in the literature (Sebastian et al. 2017*a*). Discovering such novel and potentially productive knowledge linkages serves to stimulate research development processes and increase research productivity (Jha et al. 2018, Rzhetsky et al. 2015, Swanson & Smalheiser 1997).

Notwithstanding the significant progress of LBD researchers in tackling this problem over the last few decades, there are several open issues and shortcomings in the LBD literature. The overarching goal of this thesis was to fill these identified research gaps with high precision. To this end, five primary research objectives were defined at the outset of this thesis, after which numerous studies were conducted to explore novel ways to accomplish these defined research objectives. In this process, several major contributions were made to the field of LBD research. The purpose of Section 8.2 is to provide an extended discussion of these major contributions, which are also outlined at the end of each of the chapters that are dedicated to the five main objectives. The studies conducted as part of this thesis also open up novel directions for future LBD research. These future directions are discussed in detail in Section 8.3 in light of the five main research objectives. Section 8.4 concludes by summing up the entire purpose of the thesis and how it has contributed to enhancing existing understandings of the LBD workflow to promote its *widespread applicability*.

8.2 Major Contributions

This section contains an extended discussion of the major contributions of the studies conducted in this thesis, as outlined in the latter part of the previous chapters. These major contributions represent the key outcomes of the five main research objectives that were initially defined in Chapter 1.

8.2.1 Main Research Objective 1 (RO1)

To integrate a large-scale systematic literature review procedure of LBD studies, in order to address the limitations in the existing traditional narrative-based LBD reviews, while shedding light on novel focus areas in the LBD workflow.

Due to the 35 years of LBD research and its increased knowledge production evident each year, there was a critical need for a *systematic literature review* in the LBD discipline. Systematic literature reviews are considered the *gold standard* among reviews since they provide a comprehensive overview of the evidence in a discipline (Snyder 2019). With this in mind, this thesis performed a large-scale systematic literature review, following a research method and process that adheres with standards and guidelines to collect, appraise and synthesise the literature, as discussed in Chapter 2.

To construct the review protocol for this objective, this study adhered to the standard systematic literature review procedure used in computer science. In following this protocol and considering the entire LBD workflow, this study methodically designed *fifteen research questions* that also align with the remaining four main research objectives of this thesis. The main reasons for adopting such a dilated search scope are the *restrictive scope* and *limited focus points* of existing LBD reviews. Thus, the expanded search scope of this systematic literature review offered the opportunity to appraise the LBD literature from multiple perspectives, considering each component of the LBD workflow, while also shedding light on new areas. Ultimately, this assisted in performing a comprehensive systematic literature review to gain rich, deep insights and conclusions on the historical progress and contemporary focus in the LBD field.

One of the main findings of this review was that very few existing studies have enriched the LBD workflow to support its widespread applicability. Following this notion, through the ensuing research objectives, this thesis pursued several new research directions, with the aim of promoting widespread applicability of the LBD workflow that is crucial to provide broader community benefits. In addition to providing a strong theoretical framework for the remaining research objectives of the thesis, this review is, to the best of our knowledge, the *first* systematic literature review reported in the field.

8.2.2 Main Research Objective 2 (RO2)

To investigate the input component of the LBD workflow in order to deduce the suitability of different input types in the LBD process.

The selection of an LBD input type (that denotes its input component) is not straightforward. This is because different *data fields* in the research papers have their own *perspectives* (Lee et al. 2015) and *information content* (Kostoff et al. 2004). Therefore, understanding the LBD input component in terms of how each input type contributes to the LBD workflow and what impact this has on the overall knowledge discovery is important, as the selected input type plays a central role in the information retrieval cycle of the LBD workflow. Even though there are some LBD studies that *implicitly* attempt to comprehend the performance differences of LBD input types, these studies drew conclusions without isolating the input component from their proposed discovery methodologies. Thus, these conclusions may not be generalisable (i.e., they could differ when a different discovery method is utilised). Unavailability of LBD studies that *explicitly* analyse the LBD input component explains why there is no consistent selection of LBD input types in the field, and why different studies have focused on different input types to facilitate knowledge discovery. Given this gap in the literature, this thesis sought to investigate the input component so as to scrutinise the role of input types and how they contribute to the information retrieval cycle by assessing their *informativeness*.

Informativeness (or *information richness*) of the different input types can be captured as an *objective textual feature* or as a *subjective measure* that captures the interactivity between texts and users. In information retrieval tasks such as LBD, using *subjective definitions* is considered to be more appropriate and meaningful (Tague-Sutcliffe 1992). Following this notion, this study explored a suitable setting that could define the *information richness* of each input type involving the information retrieval cycle in the LBD context (i.e., focusing on the subjective understanding of information). To this end, this study took inspiration from *optimal foraging theory*, since the main objective of this study (i.e., identifying LBD input types that demonstrate maximised information richness) can be viewed as an *optimisation problem*. More specifically, intermingling subjective understanding of information with optimal foraging theory facilitated the quantification of input types in terms of: *'how much important information does the information retrieval cycle (i.e., information-as-process) provide to the user (i.e., information-as-knowledge)?'*.

The evaluation of this study focused on a large-scale assessment of the information richness of nine different variants of the most common and viable LBD input types. Overall, information richness was showcased in the following order (from highest to lowest): *title and abstract*, *MeSH keywords* and *titles only* using the three metrics *IR*, *intrigue IR* and *average intrigue score*. This was the case not only in the *default dataset*, but also in the *extended datasets*, where the nearest neighbour count was set to 5 and 10, respectively. Furthermore, this study observed that the inclusion of neighbouring documents into the local corpus does not improve information richness in the information retrieval cycle. This is because the *nearest neighbour count* of such extended datasets is negatively correlated with the information richness score, indicating that such extended searches in the LBD knowledge discovery process are not efficient due to their negative impact on observed optimal foraging behaviours. Succinctly, this study presents the first steps towards a better understanding of input component and what impact this might have on the remaining components in the LBD workflow, with a view to developing better LBD systems in the future.

8.2.3 Main Research Objective 3 (RO3)

To enhance the discovery component of the LBD workflow using fine-grained diachronic semantic inferences by conjoining global semantic relationships with the temporal dimension to enrich the typical static cues used in the LBD literature.

The focus of RO3 was to enhance the *discovery component* of the LBD workflow. To this end, this study attempted to identify the potential contributions of *diachronic semantic inferences* in discovering potential novel knowledge linkages in the literature. This study focused on *diachronic semantic inferences* for two main reasons. Firstly, based on the *timeline analysis* of LBD computational techniques (discussed in Chapter 2), this thesis observed that the use of modern word embedding techniques is emerging, yet only a handful of LBD studies use such techniques. Secondly, based on the *categorisation* of LBD computational techniques (discussed in Chapter 2), this thesis observed that almost all previous LBD studies have overlooked the importance of integrating the temporal dimension into the knowledge discovery process, as they rely on a static snapshot of literature. Cogitating the complementary strengths of these two observations, this thesis co-modelled *vector semantics* captured through modern word embedding techniques with the *temporal dynamics*.

This study incorporated a circumstantial temporal analysis component to perform rigorous and precise analysis of *diachronic semantic inferences* through the integration of a wide range of techniques from research areas such as *sequence mining*, *time series analysis* and *signal processing* for the first time in the LBD discipline. The decision to perform such fine-grained temporal analysis was based on two observations in the few recent LBD studies that attempted to integrate temporal information of scientific topics into the LBD workflow (Jha et al. 2018, Jha, Xun, Wang & Zhang 2019, Xun et al. 2017). Firstly, even though these few LBD studies undoubtedly improved on the typical LBD setting through the integration of temporal details, the temporal analysis component used in their LBD workflow was *fairly shallow*. Secondly, these initial LBD studies (Jha et al. 2018, Jha, Xun, Wang & Zhang 2019, Xun et al. 2017) merely considered limited temporal characteristics when defining new knowledge linkages. Nevertheless, due to the complexities involved in natural language usage, as well as the availability of novel knowledge linkages in multiple forms, the use of limited temporal characteristics in the knowledge discovery process may inhibit the predictive performance of LBD models. The complementary integration of these two observations formed the groundwork for the *proposed temporal component* in this study.

The *semantically infused temporal trajectories* (i.e., *diachronic semantic inferences*) are considered the *core analysis unit* of the proposed LBD framework. These derived diachronic semantic inferences could be analysed in two broad ways to manifest their potential for detecting novel knowledge linkages in the literature. The first category would be the *direct usage* of these proposed diachronic semantic inferences to facilitate discovering such latent knowledge linkages. The two proposed LBD models, *DTM* and *FTM* represent this category, where the derived semantically infused temporal trajectories are directly utilised to mine meaningful patterns and make predictions. In contrast, the second category of analysing diachronic semantic inferences involves *indirect usage*. In accordance with this category, potential temporal signals are not directly extracted from semantically infused temporal trajectories. Instead, they serve as a *medium* to facilitate knowledge discovery. The proposed LBD model, *TAM* belongs to this category. It demonstrates how the proposed diachronic semantic inferences can be used as a medium to perform a process similar to that employed by a docking engine (Jacob et al. 2012).

More specifically, these three LBD models represent the *core trajectory analysis* component in the proposed LBD framework. The purpose of these LBD models is to scrutinise the proposed diachronic semantic inferences to identify hitherto undiscovered semantically infused temporal signals that may potentially help to discover new knowledge linkages within the remaining scientific topics in the literature. The experimental results substantiate the efficacy of both *direct* and *indirect* uses of the proposed diachronic semantic inferences through their robust predictive performance evident in every experimental setup across all test cases. Succinctly, the demonstration of both direct and indirect uses of the proposed diachronic semantic inferences indicates that the proposed semantically infused temporal trajectories are capable of enhancing prediction performance in the LBD knowledge discovery process.

This thesis also attempted to identify the potential independent contributions of each semantic shift type to prediction performance. In this analysis, this thesis observed that the semantic shift types alone also outperformed the baseline models. More precisely, the three semantic shift types alone (i.e., *ISS*, *PSS* and *NSS* individually) also tended to perform better than the baseline models. These findings indicate the potentially positive influence of semantically infused temporal signals towards discovering novel knowledge linkages more precisely. This emphasises that performing a rigorous and precise temporal analysis in the LBD workflow is rewarding, since even simplified versions of such analyses (i.e., individual performances of each semantic shift type) also display improved prediction performances in every experimental setup.

The third proposed LBD model (*Trajectory Alignment Model; TAM*) demonstrates a distinct perspective to the best of our knowledge for the *first time* in LBD discipline by incorporating semantically infused temporal patterns based on *relativity*. The main inspiration for this proposed LBD model is the *docking* method used in molecular modelling to facilitate *structure-based drug design* (Ferreira et al. 2015). The purpose of the docking engine is to quantify the free energy of binding ΔE between the receptor and a ligand to rank the ligands based on ΔE, which denotes whether the ligand bindings are more or less favourable (Jacob et al. 2012). Following the same notion, this proposed model uses the bindings of the semantically infused temporal trajectories of local topics (analogously *ligands*) with the temporal trajectories of actual novel knowledge linkages (analogously *receptors*) to derive some cost metric, which depicts whether the trajectory binding is less or more favourable.

Incorporating the idea of the aforementioned *docking* process into the LBD process could be particularly beneficial for two main reasons. Firstly, it is natural to assume that the concealed patterns of potential novel knowledge linkages in which LBD researchers have explored for more than three decades are encapsulated in these *actual novel knowledge linkages* (i.e., *templates*) due to the fact that they have been realised in real-world with time. Thus, the patterns that these templates enclose provide a rich platform for the formation of deductions that may be crucial to the knowledge discovery process. Nevertheless, these encapsulated patterns of these templates may not be *noteworthy* when they are considered as *separate entities*. For this reason, the idea of *relativity* (as similar to docking) may be appropriate in this situation. Secondly, the *theoretical LBD literature* has identified that novel knowledge linkages may reside in several forms in the scientific literature (Davies 1989). Nevertheless, to date, only a handful of such novel knowledge linkages have been identified, despite 35 years of LBD research. This may be due to the complexity of natural language usage, which may hinder the identification of such noticeable new knowledge linkage forms. More specifically, the complexity involved in natural language usage may have inhibited the identification of several hundred or even thousands of other forms of novel knowledge linkages that are not salient and may potentially be discovered in future theoretical LBD studies. However, the idea of maintaining a large collection of templates in a *template repository* may circumvent this hindrance to some extent, as such a repository could accommodate a large number of novel knowledge linkage forms in one place.

To carry out *trajectory binding*, TAM maintained a *template repository* in which the trajectories of actual novel linkages were stored. Subsequently, these trajectories were aligned with the trajectories of potential candidates in a time-invariant manner, so as to assess the extent to which the trajectories of local topics resembled the patterns of templates in the template repository. Overall, the experimental results substantiate the efficacy of incorporating such deductions made using *trajectory binding* in almost every experimental setup. This indicates the potential positive influence that patterns based on relativity could have on the knowledge discovery process, and which future LBD research could further expand and explore.

While most previous LBD models depend on semantic inferences performed using domain-specific knowledge resources to facilitate knowledge discovery, the LBD models proposed in this thesis are entirely free from such domain-dependent semantic inferences using external knowledge resources. This is because the reliance on such semantic inferences using domain-dependent external knowledge resources inhibits the LBD model's support for *reusability* and *portability*, which are vital to the provision of broader community benefits. More specifically, *reusability* and *portability* are two crucial design properties that should be considered when developing LBD models, since they inject new meaning into the LBD workflow that are otherwise obscured.

The notion of *reusability* denotes the process of creatively exploring new application areas of the proposed LBD models, in order to propose expeditious solutions. This is particularly beneficial in the context of LBD, since the fundamental purpose of LBD research is to discover novel linkages (through the integration of signals from text corpora), which could be broadly applicable across numerous other problem settings. These problem settings may not necessarily imply closely related reusability settings to the problem of LBD (indicating *vertical reuses*), but also could entail completely different reusability settings from those in the LBD context (indicating *horizontal uses*, as discussed in Section 8.3). The view of *portability* denotes the LBD model's ability to facilitate knowledge discovery in new environments. The LBD model's support of portability is vital due to the fact that the benefits of LBD research are domain-agnostic and could be broadly applied in almost any discipline. The escalating growth of scientific literature is evident in almost every discipline; thus, potential stakeholders in LBD models may be expected to exist in any discipline. Otherwise stated, the development of LBD models to assist researchers in discovering latent novel knowledge linkages in the scientific literature is crucial despite the domain. Contemplating the vast opportunities afforded by preserving these two vital properties: *reusability* and *portability*, the proposed LBD models were designed without the incorporation of semantic inferences from external knowledge resources. In essence, the prediction effects of the proposed LBD models do not rely on domain-specific semantic inferences; thus, they could be broadly applicable to *reusable applications* and *portable environments* to ensure their *widespread applicability.*

8.2.4 Main Research Objective 4 (RO4)

To validate the predictive power of the proposed LBD models through reuse research, with the goal of providing broader community benefits.

In addition to the *direct* and *indirect* uses of the proposed LBD models demonstrated in *RO3*, the reuse research performed as part of this research objective provides a distinct perspective on the proposed LBD models, which is their *vertical reusability.* The focus of reuse research is to efficiently reuse components (or similar artifacts) in new application areas. Performing such reuse research enables the identification of new application areas of proposed LBD models, ultimately providing the opportunity for broader community benefits. In addition to this, reuse research also increases the *dependability* (or *reliability*) of the proposed LBD models.

Motivated by the enormous potential of such reuse research, this study conducted large-scale reuse research following a method similar to *opportunistic reuse.* The idea of opportunistic reuse is to make new capabilities in new problem areas by gluing together components designed for distinct problem setting(s). In search of new problem areas, this study focused on several special-purpose LBD models that have been reported in the LBD literature, which cater to specific problem areas, such as *drug development*, *drug repositioning* and *adverse drug reactions* (Henry & McInnes 2017). The urgency of contributing to these special-purpose application areas of LBD has been underscored by the COVID-19 pandemic, with over 30.6 million cases across the globe resulting in more than 0.9 million deaths (as of September 2020), while antiviral medications are still under investigation (Wang et al. 2020). With this in mind, this thesis sought to explore this timely direction to demonstrate the reusability of the proposed LBD models. More specifically, this use case indicates a *vertical reuse*, wherein a closely related problem setting to the LBD discipline is used to substantiate the prediction performances of the proposed LBD models.

The experimental results corroborate the efficacy of the proposed LBD models in this new reuse setting. This verifies the potential contributions of *direct* and *indirect* uses of the proposed diachronic semantic inferences to the LBD workflow, even in reuse settings. Even though the proposed LBD models consistently outperformed the baseline models, this study observed a performance decrease in the proposed LBD models in this reuse setting, relative to the two-node search (discussed in RO3). The main difference between this reuse setting and the previous two-node setting in terms of the diachronic semantic inferences is the *number of semantically infused temporal trajectories* used during the LBD knowledge discovery process. More specifically, in the previous two-node setup, this thesis inspected six semantically infused temporal trajectories, whereas in this reuse setting, the thesis incorporated only four out of these six semantically infused temporal trajectories. This was because of the incompatibility of *pairwise distance proximity* and *neighbourhood distance proximity* in this new reuse setting. This observation arose the following question: *'does the number of meaningful diachronic semantic inferences integrated in the knowledge discovery process positively correlate with the predictive performance?'*.

To answer the question, this study scrutinised the performance differences of the best-performing proposed LBD model (FTM) for all the possible combinations of the semantically infused temporal trajectories. Through this study, it was verified that there is a strong positive correlation between the number of semantically infused temporal trajectories and the predictive performance of the LBD workflow. This study exemplifies one of the key benefits offered by reuse research, which is the identification of potential bottlenecks in reused models. Identifying such bottlenecks provides an extended platform to elicit precise enhancements in future iterations. More specifically, due to the complexity of the problem that LBD research aims to solve, it is difficult (or perhaps even impossible) to construct universal LBD models that make perfect predictions in every possible reuse setting. Therefore, verifying the reusability of the proposed LBD models in new reuse application areas provides insights into the potential bottlenecks of proposed LBD models. Such insights could be used as a guide to further improve the results in the next iterations of the reuse setting. Following this notion, this study proposed several directions (i.e., focusing on the neighbourhood density changes of the local topics) for integrating meaningful diachronic semantic inferences to make up for the absence of *pairwise distance proximity* and *neighbourhood distance proximity* in the second iteration of the grab-and-glue framework.

In the process of analysing the potential contributions of different trajectory combination types, this study observed that even the most simplified versions of the highest performing proposed LBD model in this reuse setting (*FTM*) showcased higher predictive performances than the baseline models. These trajectory combination types included the prediction performances of *one trajectory*, *two trajectories* and *three trajectories* with a total of 14 types of simplified versions of the proposed LBD model, *FTM*. Overall, 12 types of these simplified versions outperformed all the baseline models across all the k values, while the remaining 2 types of these simplified versions outperformed all the baseline models after the k value of 30. The increased predictive performances of even the most simplified versions of the trajectory combination types indicate the positive influence that the proposed diachronic semantic inferences on the knowledge discovery process of the LBD workflow.

The proposed LBD models provide greater flexibility in adapting to a wide range of application areas, as demonstrated in this main objective as well as discussed broadly in the future directions section (i.e., Section 8.3). This is evident mainly because of two reasons. Firstly, the domain-agnostic nature of the proposed LBD models, since they are completely free from knowledge inferences performed using domain-specific knowledge resources, can be used in a diverse range of application areas. Secondly, the power of embedding spaces that the proposed LBD models are based on provides a greater flexibility in performing tasks such as *vector arithmetic operations* and *analogy mining* to quickly adapt the proposed model in the new reuse setting. Such flexibility is rare when using hard-wired knowledge discovery structures such as graphs. Moreover, the LBD systems that are based on inferences using external domain-specific knowledge resources also inhibit their flexibility in terms of reusability. This is due to the unavailability or unsuitability of such knowledge inferences that are made using domain-specific knowledge resources in other application areas. Hence, the domain-agnostic nature of the proposed LBD models, and the power of vector semantics that they are based on, not only support their adaptation in *vertical reuse* (e.g., similar to this main objective) and *horizontal reuse* (discussed in Section 8.3), but could also be adapted to develop *novel components* in the LBD workflow such as *personalised knowledge discovery* (discussed in Section 8.3).

8.2.5 Main Research Objective 5 (RO5)

To demonstrate the portability of the LBD workflow by proposing an interdisciplinary (or generalisable) LBD framework to assist scientific problem solving in a domain-agnostic manner.

The key aim of this portability research was to reach a large and diverse community by proposing a *highly cost-efficient* and *easily integrable* portable LBD framework, which supports both *cross-domain* and *cross-lingual* knowledge discovery. Enabling portability is particularly important in LBD field, since the potential stakeholders in LBD systems could exist in almost any academic discipline. Thus, extricating domain-dependent hindrances (which constrain the applicability of LBD models outside the medical domain) establishes the portability of LBD models. This is crucial to *widespread applicability* of the LBD models. To the best of our knowledge, this is the *first study* on LBD that demonstrates portability with the aim of unlocking the benefits of typical LBD models to research communities beyond medicine.

To facilitate portability, this study leverages *semantic web technologies* (more specifically *Linked Open Data (LOD)*), which provide revolutionary opportunities to gain rich understandings from unstructured texts in a machine-readable manner. More specifically, this study selected *DBpedia*, which is a dominant semantic web resource since the inception of the *linking open data* project to circumvent the existing domain-dependent impediments in a typical LBD framework. The main influence for this selection is that it adhered to all the portability criteria defined in the context of LBD, which represent two main aspects of portability, namely *new environments of portability* and *degree of portability*.

The demonstrated portable framework in this thesis supports knowledge discovery across a wide range of academic domains, including (but not limited to) *medicine*, *computer science*, *sociology*, *psychology*, *geography*, *economics*, *anthropology*, *philosophy*, *law*, *languages and literature*, *history*, *arts*, *social work*, *biology*, *chemistry*, *earth science*, *space science*, *physics*, *mathematics*, *business*, *engineering* (including *chemical engineering*, *civil engineering*, *educational technology*, *electrical engineering*, *material science and engineering*, *mechanical engineering*) etc. due to two main reasons. Firstly, the prominence of DBpedia as a *cross-domain resource* enables knowledge discovery to be performed across a diverse range of domains. Secondly, the semantic inferences made in the proposed solutions (using DBpedia) did not follow expedients (or shortcuts) that are limited to only certain domains (such as *medicine*, as discussed in Chapter 7); thus, the proposed solutions are compatible in almost every domain that DBpedia supports. The proposed portable LBD framework also supports large scale *cross-lingual* knowledge discovery due to the *multilingual nature* of DBpedia.

This study demonstrated a proof of concept of one of the key specialities of the proposed portable framework, which is its *interdisciplinary (or generalisable)* nature. Such generalisable capabilities, to the best of our knowledge, have never been possible within the existing LBD models. Therefore, while avoiding the *lack of portability* which plagued previous LBD models, this proposed LBD framework also opens up a whole new level of knowledge discovery in LBD discipline by enabling *multifaceted knowledge discovery*. This is an unprecedented direction to the LBD discipline. Otherwise stated, in addition to carrying out knowledge discovery within a single main domain, the proposed portable framework also supports the use of multiple domains in a single knowledge discovery. This enables users a tremendous opportunity in discovering latent novel knowledge not only in a single main domain, but also across multiple domains. Such multifaceted knowledge discovery is crucial to advancing *interdisciplinary research* (such as bioinformatics and medical informatics). For example, consider the recent project *Neuralink*[1], which strives to develop implantable brain-machine interfaces. Such interdisciplinary research requires the discovery of knowledge from a wide range of domains including (but

[1] https://neuralink.com/

not limited to) *medicine*, *robotics*, *neural engineering*, *human-computer interaction*, *artificial intelligence* and *chemistry*. The discovery of knowledge across such a range of disciplines is beyond the ability of existing LBD models. However, the portable LBD framework proposed in this research objective is capable of facilitating such *interdisciplinary knowledge discovery*. Future LBD research should be able to take advantage of this capability. Another key advantage of the proposed portable framework is that the costs involved in the transitions of domains during knowledge discovery (i.e., the *degree of portability*) are none or negligible that reflects the *cost-efficient* nature of the proposed solution. Similar to the domains, the proposed portable LBD framework also supports the integration of multiple publication languages in a single knowledge discovery process with zero or negligible costs in the transitions between (or among) languages.

One of the key focuses of this portability research was to substantiate how well the proposed solutions resemble the knowledge inferences performed using *MeSH*. The main reason for this focus was that *MeSH* had become an integral part of most of the prior LBD studies to the formation of semantic deductions. Thus, if the proposed solutions using DBpedia closely resemble the knowledge inferences made using MeSH, the LBD community does not have to perform substantial modifications to their LBD models when integrating the proposed solutions to enable the portability. The experimental results indicate that the proposed solutions using DBpedia tend to have similarities with the corresponding knowledge inferences made using MeSH. In some instances, such as in the case of synonym identification, the proposed solutions showcased superior performances than MeSH. This ensures that existing LBD models could easily enable the portability of their models through the integration of the proposed solutions, spreading the benefits of LBD models to research communities beyond medicine. This will also assist in enhancing LBD research outside of the medical domain, where it is still in a nascent stage.

To the best of our knowledge, this is the *first study* reported in LBD literature to introduce *DBpedia* and verify its suitability in the LBD workflow. DBpedia is superior to most existing knowledge bases (not only in LBD discipline, but also in general) as a result of its many strengths that are lacking in existing knowledge bases. These distinctive features of DBpedia provide a unique and significant potential to perform multifarious deep semantic inferences. Such inferences are crucial to gain a rich understanding of the unstructured text in a *domain-agnostic* and *language-agnostic* manner. This also alleviates one of the top-cited major challenges faced by non-medical LBD studies, which is the unavailability of a comprehensive knowledge base that allows for the formation of semantic inferences during the knowledge discovery process.

TABLE 8.1: Summary of the major contributions

Objective	**Major Contributions**
RO1	• being the first *systematic literature review* in the LBD discipline. • following a rigorous review protocol with methodically designed research questions that cover the entire LBD workflow, while also shedding light on several new focal points.
RO2	• being the first study in the LBD discipline that comprehensively analyses and evaluates the input component of the LBD workflow. • proposing a novel perspective on assessing the *information richness* of LBD input types, taking inspiration from foraging theory and subjective understandings of information that make use of the information retrieval cycle of the LBD workflow.
RO3	• being the first study in the LBD discipline to incorporate a circumstantial temporal component by utilising a wide range of techniques from areas such as *sequence mining*, *time series analysis* and *signal processing*, in order to perform a fine-grained analysis of semantically infused temporal trajectories. • being the first study to introduce patterns based on *relativity* by taking inspiration from molecular docking mechanism. • demonstrating not only the *direct uses* of the proposed diachronic semantic inferences, but also their *indirect uses* through the trajectory alignment model. • the experimental results verified the efficacy of the proposed LBD models (i.e., both *direct* and *indirect* usage of diachronic semantic inferences) in all experiments, performed under different settings.

	• the proposed semantic shift types in isolation (i.e., *ISS*, *PSS* and *NSS*) also demonstrated high prediction performances (in both *direct* and *indirect* uses of diachronic semantic inferences) compared to the baseline models, indicating the predictive power of the proposed semantically infused temporal trajectories, even individually. • the prediction performance of the proposed LBD models does not depend on semantic inferences performed using external domain-dependent knowledge resources, which ensures their *reusability* (in various problem settings) and *portability* (in various academic domains), offering the opportunity to provide broader community benefits.
RO4	• performing large-scale reuse research by integrating considerations of reusability through a methodical reuse plan. • demonstrating the vertical reuse of the proposed LBD models considering an opportune application area in the LBD field. • the proposed LBD models exhibit a greater flexibility in adapting to new reuse settings, due to their domain-agnostic nature and to the power of vector semantics on which they are based. • establishing the models' fitness for the intended purpose through the first iteration in the *grab-and-glue* framework, compared to the competitive baselines in the two-node search, as well as state-of-the-art link prediction techniques. • the trajectory combination types alone also demonstrated high predictive performances compared to baseline models, which verifies the predictive power of the proposed semantically infused temporal trajectories, even when they are used individually.
RO5	• being the first LBD study that proposes a comprehensive portable LBD framework to support knowledge discovery in a cross-domain and cross-lingual manner. • being the first LBD study to demonstrate interdisciplinarity (or generalisability) through the combination of multiple domains in a single knowledge discovery, with negligible costs in the transitions between (or among) domains.

	• being the first study to introduce DBpedia to the LBD discipline, providing the opportunity to perform multifarious semantic inferences using unstructured text in a domain-agnostic and language-agnostic manner. • observing that the proposed solutions displayed similarities to (and sometimes outperformed) the commonly used LBD resource *MeSH*, meaning that the LBD community will be able to integrate the proposed solutions into their LBD models without substantial modifications, which will facilitate LBD research beyond medicine.

8.3 Future Work

There are a number of future directions opened as a result of the studies conducted in this thesis. The purpose of this section is to discuss these opportunities for future LBD research. The discussion is framed in relation to the thesis' five main objectives.

8.3.1 Main Research Objective 1 (RO1)

To accomplish RO1, this thesis performed a large-scale systematic literature review as discussed in Section 8.2. More recently, Kastrin & Hristovski (2020) have performed a large scale *scientometric analysis* of the LBD literature. They focus on evidence synthesis in a manner that is similar to methods employed in a systematic literature review. One interesting future direction of this research objective is to conduct an *umbrella review* (sometimes called a '*reviews of reviews*') (Newman & Gough 2020).

The purpose of umbrella reviews is to systematically collect and evaluate the information on previously published literature reviews. Thus, umbrella reviews offer the opportunity to obtain a more comprehensive picture of the discipline, while ascertaining whether the evidence base of the topics and questions in discipline is *consistent*, *contradictory* or *discrepant* and exploring the potential reasons to describe them. Conducting an umbrella review may be particularly important in the field of LBD, which has accrued over 30 years of research. Thus, synthesising *consistent*, *contradictory* or *discrepant* evidence base and their potential reasons may provide better future directions in the LBD discipline. Since the data in umbrella reviews are extracted from previous reviews (i.e., secondary levels of analysis) rather than primary research studies, these reviews are considered a tertiary level of research analysis (Newman & Gough 2020). Even though umbrella reviews provide an efficient way to examine previous research, they are comparatively novel. Thus, the emerging methodologies used to undertake umbrella reviews open up many challenges and questions (Newman & Gough 2020, Wiechula et al. 2016). For instance, care is required in the assessment of source reviews in terms of data inclusion, study quality and overlap among reviews (Newman & Gough 2020).

8.3.2 Main Research Objective 2 (RO2)

The RO2 of this thesis was designed with reference to the notion of the subjective understanding of information, while also incorporating optimal foraging theory. With the aim of measuring the *information richness* of the LBD input types, this thesis approximated the *benefit assessment* through the use of *time-slicing* since it ensured *information by consensus* and *replicability*, as discussed in Chapter 4. Consequentially, an interesting future research direction would be to validate whether the observed patterns are consistent with actual user studies in terms of the proposed three *informativeness* (or *information richness*) metrics. Performing such user studies will not only provide an extended platform to further validate the observations reported in Chapter 4, but also enable to gain additional understanding of user engagement with respect to each of the LBD input types, which could pave the way to deeper insights into the LBD input types.

This thesis mainly influenced from the *subjective understanding* of information to assess and compare LBD input types. However, exploring the *objective definitions* of the input types (i.e., considering *textual features themselves*, as discussed in Chapter 4) may facilitate the comparison of the observed results in this thesis to verify if they are consistent with the subjective understanding of information. In this regard, one could consider the *clustering ability* of the input types as a metric which could be used to compare various LBD input types. One possible way to analyse the clustering ability of input types is to measure the similarity of data points in each input type when they are compared to their own cluster (i.e., to capture *cohesion*), as well as their differences to other clusters (i.e., to capture *separation*), as illustrated in Figure 8.1. In this way, one can verify whether each input type indicates poor cohesion and separation (i.e., too many or too few clusters) or an appropriate clustering configuration (i.e., the data points

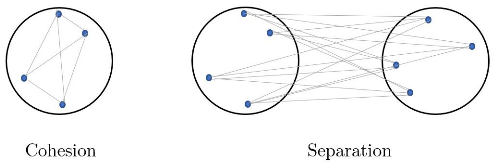


Figure 8.1: Cohesion and separation of data points in the input types

are well-matched with their own clusters and poorly matched with their neighbouring clusters). In essence, exploring objective definitions such as clustering ability enables the comparison of LBD input types according to their textual differences.

This thesis mainly focused on the *standard data types* used in the LBD discipline to assess the information richness. Following this direction, an interesting future work would be to perform information richness analysis on *non-standard data types* that involve two or three combinations of data types such as *title and MeSH*, *abstract and MeSH*, and *title, abstract and MeSH*. This will pave the way to discover efficient combinations of data types that have not been used in the LBD workflow yet. Moreover, it would also be interesting to observe how the information richness score of data types changes using test-cases outside the biomedical field such as physics, chemistry and humanities to analyse whether the information richness score is sensitive to factors such as differences in scientific expressions and linguistic styles in each discipline.

8.3.3 Main Research Objective 3 (RO3)

The RO3 of this thesis uses semantically infused temporal trajectories as the core analysis unit. The performance of the proposed LBD models could be further enhanced by identifying *important regions* in the proposed temporal trajectories. Proceeding with this idea, one could employ recent advancements in deep learning by using *attention mechanisms* to detect *important regions* in the temporal trajectories (or *segments* that are critical for the prediction). One way to perform attention would be to employ a sliding window to identify the subsequences in the temporal trajectories that are considered to be candidate segments. Then, they could be fed to the pre-trained model to obtain some metric such as entropy. The top K segments (based on the metric utilised) could be sent through some weighted ensemble mechanism, helping to identify the most discriminative parts of the trajectories (Hsu et al. 2019) (Figure 8.2). Identification of such critical segments in the temporal trajectories provides the opportunity to emphasise

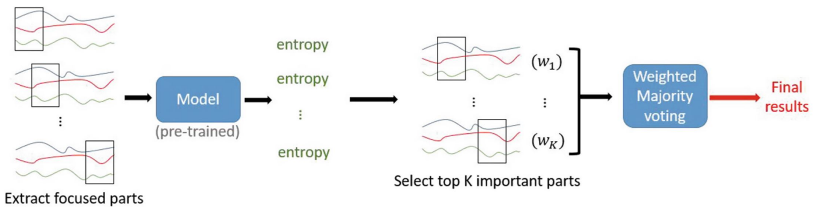


FIGURE 8.2: Identifying important segments in the trajectories (Hsu et al. 2019)

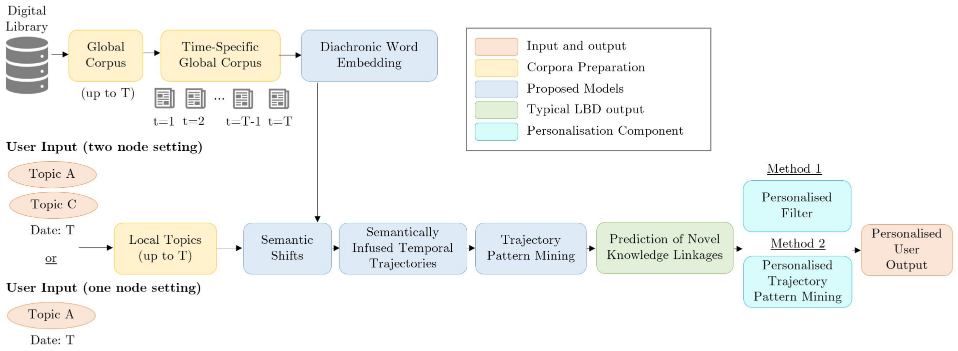

Digital Library
Global Corpus
(up to T)
Time-Specific Global Corpus
t=1 t=2 ... t=T-1 t=T
Diachronic Word Embedding
Input and output
Corpora Preparation
Proposed Models
Typical LBD output
Personalisation Component
User Input (two node setting)
Topic A
Topic C
Date: T
or
User Input (one node setting)
Topic A
Date: T
Local Topics (up to T)
Semantic Shifts
Semantically Infused Temporal Trajectories
Trajectory Pattern Mining
Prediction of Novel Knowledge Linkages
Method 1
Personalised Filter
Method 2
Personalised Trajectory Pattern Mining
Personalised User Output

FIGURE 8.3: Integrating personalisation component into the proposed LBD framework

these critical regions when making predictions, and to better understand the LBD model that may be vital for future decision making.

As with the previous LBD research, the proposed LBD models in this thesis only support static outputs. That is, for a given user input, the same output will be returned, *irrespective of the users' context*. More specifically, consider two researchers, each of whom have completely different interests and expertise. When these two users input the same user query q_1 to initiate knowledge discovery, existing LBD models (including ours) will return the same output irrespective of the users' context differences. Nevertheless, the user's context plays a critical role in information retrieval tasks (such as LBD) that ultimately decides whether the user is satisfied by the produced output. To facilitate such personalised knowledge discovery, it is important to integrate a personalisation component into the typical LBD workflow.

With reference to the LBD framework proposed in this thesis, Figure 8.3 outlines two potential mechanisms through which to fuse the idea of personalisation with the LBD workflow. The user's context (which is required for the personalisation component) could be automatically inferred by analysing papers that the user has authored, including his/her reading list (using tools such as *EndNote* or *Google Scholar*). To model the user's context in the constructed vector spaces of the proposed LBD models, a *personalised*

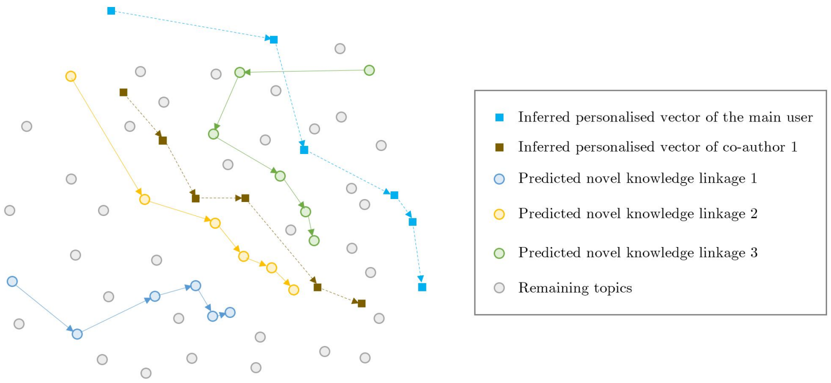

■ Inferred personalised vector of the main user
■ Inferred personalised vector of co-author 1
○ Predicted novel knowledge linkage 1
○ Predicted novel knowledge linkage 2
○ Predicted novel knowledge linkage 3
○ Remaining topics

Figure 8.4: Proposal for personalised trajectory pattern mining by adapting the proposed LBD models

vector needs to be created. For this purpose, one could identify important concepts that describe the papers that the user has authored and his/her reading list, and use the arithmetic operations of the word vectors of these identified concepts to infer a *personalised vector* (see Figure 8.4). One direction that could be used to infer such a personalised vector is to average the word vectors of the user's important concepts.

The first proposed method leverages the idea of considering personalisation as a *filter* (Figure 8.3). More specifically, the prominent novel knowledge linkages identified in the LBD process are reordered with respect to the personalised vector in the most recent embedding space. The other proposed method, which is *personalised trajectory pattern mining* involves a meticulous analysis of the user's context (Figure 8.3). Specifically, it entails analysing the way in which the trajectories of the most notable novel knowledge linkages change with the inferred personalised vector across time (Figure 8.4). Unlike method 1 (which is static, since only the last timestamp is used), this method offers greater flexibility and more diachronic cues to better model personalisation. The next stage of this personalised component will be the integration of the *inter-community context* of the user using his/her existing collaborations (Figure 8.3).

8.3.4 Main Research Objective 4 (RO4)

The RO4 of this thesis was designed by adapting the proposed LBD models in a new reuse setting to demonstrate their vertical reusability. Nevertheless, there are other closely related application areas in LBD (i.e., other *vertical reuses*) that could be potentially tackled to further verify the robust predictive effects of the proposed semantically

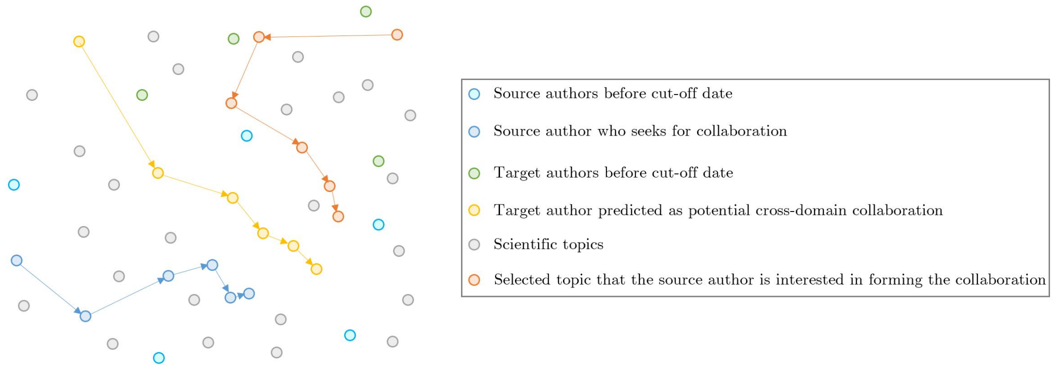

Source authors before cut-off date
Source author who seeks for collaboration
Target authors before cut-off date
Target author predicted as potential cross-domain collaboration
Scientific topics
Selected topic that the source author is interested in forming the collaboration

FIGURE 8.5: Proposal for cross-domain collaboration recommendation by adapting the proposed LBD models

infused temporal trajectories. One such application area is *cross-domain collaboration recommendation*. For example, consider a situation where a researcher in the *source domain* is seeking novel collaborations from a *target domain*. In this instance, the evolution of scientific topics of the authors in the *source domain* and *target domain* could be modelled using diachronic vector spaces (Figure 8.5). More specifically in this instance, the proposed *individual semantic shifts* of *target authors* may reveal whether the author is willing to form cross-domain collaborations. The *pairwise semantic shifts* and *neighbourhood semantic shifts* identify how well the target author matches with the interests of the source author who is seeking collaboration and the topic on which he/she is willing to collaborate (Figure 8.5). In essence, the adaptation of the proposed diachronic semantic inferences of this thesis may provide valuable insights that can further the identification of optimal collaboration candidates.

Due to the domain-agnostic nature of the proposed semantically infused temporal trajectories, it is worth verifying their potential predictive effects in *horizontal reuse* settings. Horizontal reuse denotes the process of reusing generic components in new applications. For instance, consider the development of *novel product ideas* as an application of horizontal reuse. The underpinning two core components that intricately related to such product innovation can be considered as the product's *purpose* (what it does) and its *mechanism* (how it works). One recent popular example of leveraging the similarities between purpose and mechanism in order to kindle new innovations is a device invented by a car mechanic. This device eases childbirth by drawing similarities from extracting a cork from a bottle. Therefore, the separation of purpose and mechanism, and the identification of potential repurposings for each of these core components, is demonstrably effective in terms of idea generation (Hope et al. 2017). More specifically, given a purpose and mechanism by the user that indicates what the user is interested in solving, the model/system should be able to identify products with the same purpose performed using different mechanisms (i.e., *same purpose, different mechanisms*), as well as products with the same mechanism but different purposes (i.e., *same mechanism, different purposes*). It is possible to perform such horizontal reuse research using data from crowd-sourced product innovation websites like *Quirky.com*. These datasets are large-scale, the product ideas are explained in natural language, the invention categories span a variety of domains, and the ideas posted covers several years. This make such websites an ideal setting to adapt the proposed semantically infused temporal trajectories to elicit potential purpose and mechanism suggestions to the users' queries, with the intention of providing an impetus to accelerate product innovation.

8.3.5 Main Research Objective 5 (RO5)

To achieve RO5, this thesis attempted to integrate semantic web technologies to circumvent existing domain-dependent impediments, while introducing additional benefits such as interdisciplinary usage and cross-lingual knowledge discovery. The next stage of the proposed portable LBD framework of this thesis would be the integration of machine learning techniques to further enhance the precision of each proposed component in the portability framework. More precisely, the empirical rules and evidence reported as part of this portability research can be utilised to extract features (i.e., a *feature engineering* phase) using the DBpedia knowledge base to construct machine learning models to further enhance prediction results.

This thesis also observes that the proposed portability functionalities using DBpedia are also compatible with *Wikidata* (Vrandečić & Krötzsch 2014, Piscopo & Simperl 2019). *Wikidata* is also a multidomain and multilingual knowledge graph which is collaboratively edited by a large global community and maintained by the Wikimedia foundation. It was founded more recently than DBpedia (i.e., in 2012) (Abián et al. 2017). For instance, consider the scenario that the thesis used 'dct:subject' and 'skos:broader' properties to support domain-specific terminology extraction (discussed in Chapter 7). The same functionality can be approximated using the two Wikidata properties: *'subclass* *of'* (i.e., *P279*) and *'instance of'* (i.e., *P31*) (Erxleben et al. 2014). One could locate the corresponding Wikidata entry in DBpedia using the property *'owl:sameAs'*. Following this notion, one could analyse and compare the performance of Wikidata in comparison to DBpedia in the context of LBD, or even perform a more comprehensive analysis by integrating the knowledge in both DBpedia and Wikidata, as potential future directions of this thesis.

8.4 Concluding Remarks

One of the main findings of the systematic literature review was the need to alleviate the existing constrained environments of the LBD workflow in order to reach a large and more diverse community. This is important for two main reasons. Firstly, the intrinsic aim of LBD research (i.e., discovering novel, implicit linkages by exploring signals from text corpora) could be broadly applicable to diverse problem settings. Secondly, the potential benefits of LBD research are domain-agnostic and could be broadly applicable to any discipline.

The explorations that were performed focusing on the *input component* of the LBD workflow represents the first step towards the assessment and comparison of different input types in a generalisable manner. Such explorations are crucial to the future development of better LBD models. The sparsity of the research conducted on *modern word embedding techniques* and *temporal analysis* provided a rationale for amalgamating these two methods using diachronic semantic inferences in this thesis. The results indicated that the proposed LBD models displayed robust predictive performances, not only in terms of their *direct uses*, but also their *indirect uses*.

The *reuse research* attempted to present a distinct perspective on the LBD models by demonstrating their vertical reusability in a timely application area. The results further substantiated the robust predictive performances of both the *direct* and *indirect* uses of the proposed diachronic semantic inferences. Moreover, this study demonstrated the high levels of flexibility that the proposed LBD models exhibit, due to their domain-agnostic nature and the power of the semantic spaces on which they are based. The *portability research* proposes a highly cost-efficient, easily pluggable portable LBD framework, with the ultimate goal of extending LBD research beyond the medical domain, in which it is still in a nascent stage. While ensuring broader usage of knowledge discovery through its support of multiple domains and publication languages, this study also engenders a novel perspective on knowledge discovery through its *generalisable* capabilities.

Overall, the five main objectives of this thesis involved seeking a common thread with the goal of broadening the applicability of the LBD workflow. This thesis' development of widely applicable LBD model means that a reasonably broad array of scientific problems can be tackled by a single system. This is in contrast with LBD models, which are constructed to solve only a specific problem within a particular domain. Widely applicable LBD models should offer the possibility to customise solutions in order to solve scientific problems, which are not prefigured during their construction. Moreover, these LBD models should also facilitate the execution of knowledge discovery in a domain-agnostic and language-agnostic manner. This will allow them not only to offer the benefits of LBD research to other research communities, but also to assist in solving more complex interdisciplinary problems (such as Neuralink, as discussed in this chapter). Due to their potential to adapt to a reasonably diverse range of environments and problem setups, widely applicable LBD models are highly flexible, and their potential benefits to the community are manifold. Therefore, future LBD research could further explore and expand the novel contributions established through the studies performed in this thesis, in order to further enhance the current understanding of the widespread applicability of the LBD workflow.

Appendix A

Semantic Evolution

Algorithm A.1: Pseudocode of Individual Global Shift (IGS)

```
Input:  $model_{(t)}$ and $model_{(t+1)}$ are word2vec embedding spaces in adjacent timestamps, and $w_i$ is a string representation of a given local topic
Output: $d^{\text{IGS}}({w_i}^{(t)}, {w_i}^{(t+1)})$

Start
    1. Check if the concept $w_i$ is present in both $model_{(t)}$ and $model_{(t+1)}$
    2. Get the similarity vector of the focus concept $w_i$ from both the models
    3. Compute the cosine distance between these two similarity vectors
End
```

Algorithm A.2: Pseudocode of Individual Local Shift (ILS)

```
Input:  $model_{(t)}$ and $model_{(t+1)}$ are word2vec embedding spaces in adjacent timestamps, $w_i$ is a string representation of a given local topic, and $k$ is the size of the local neighbourhood
Output: $d^{\text{ILS}}({w_i}^{(t)}, {w_i}^{(t+1)})$

Start
    1. Check if the concept $w_i$ is present in both $model_{(t)}$ and $model_{(t+1)}$
    2. Get the two $k$-nearest neighbourhoods of $w_i$ from both models
    3. Get the 'meta' neighbourhood (both models combined)
    4. Filter the meta neighbourhood so that it contains only concepts present in both models
```

```
        5. For both models, get a similarity vector between the focus concept $w_i$ and all of the concepts in the meta neighbourhood
        6. Compute the cosine distance between those similarity vectors
End
```

Algorithm A.3: Pseudocode of Pairwise Semantic Displacement (PSD)

Input: *model*$_{(t)}$ is a word2vec embedding space in timestamp t, w_i is a string representation of a given local topic, w_A is a string representation of the user-defined concept A, and w_C is a string representation of the user-defined concept C

Output: $s^{\text{PSD}}(w_i{}^{(t)}, w_A{}^{(t)}, w_C{}^{(t)})$

```
Start
        1. Check if the concepts $w_i$, $w_A$ and $w_C$ are present in the $model_{(t)}$
        2. Get the similarity vectors of the focus concept $w_i$ and the user-defined concepts ($w_A$ and $w_C$) from the $model_{(t)}$
        3. Compute the cosine similarities between the similarity vectors of the focus concept $w_i$ and $w_A$, and the focus concept $w_i$ and $w_C$
        4. Compute the average of the two cosine similarities
End
```

Algorithm A.4: Pseudocode of Pairwise Distance Proximity (PDP)

Input: *model*$_{(t)}$ is a word2vec embedding space in timestamp t, w_i is a string representation of a given local topic, w_A is a string representation of the user-defined concept A, and w_C is a string representation of the user-defined concept C

Output: $d^{\text{PDP}}(w_i{}^{(t)}, w_A{}^{(t)}, w_C{}^{(t)})$

```
Start
        1. Check if the concepts $w_i$, $w_A$ and $w_C$ are present in the $model_{(t)}$
        2. Get the similarity vectors of the focus concept $w_i$ and the user-defined concepts ($w_A$ and $w_C$) from the $model_{(t)}$
        3. Compute the cosine distances between the similarity vectors of the focus concept $w_i$ and $w_A$, and the focus concept $w_i$ and $w_C$
        4. Compute the relative distance using the two cosine distances
End
```

Algorithm A.5: Pseudocode of Neighbourhood Semantic Displacement (NSD)

```
Input:  $model_{(t)}$ is a word2vec embedding space in timestamp $t$, $w_i$ is a string representation of a given local topic, $w_A$ is a string representation of the user-defined concept A, $w_C$ is a string representation of the user-defined concept C, $N_A$ is the most recent neighbourhood of concept A, and $N_C$ is the most recent neighbourhood of concept C
Output: $s^{\mathrm{NSD}}({w_i}^{(t)}, {w_A}^{(t)}, {w_C}^{(t)})$
Start
        1. Check if the concepts $w_i$, $w_A$ and $w_C$ are present in the $model_{(t)}$
        2. Get the similarity vectors of the focus concept $w_i$, user-defined concepts ($w_A$ and $w_C$), and recent neighbourhood ($N_A$ and $N_C$) from the $model_{(t)}$
        3. Compute the cosine similarities between the similarity vectors of the focus concept $w_i$ and $w_A$, focus concept $w_i$ and $N_A$, focus concept $w_i$ and $w_C$, and focus concept $w_i$ and $N_C$
        4. Compute the average of the derived cosine similarities
End
```

Algorithm A.6: Pseudocode of Neighbourhood Distance Proximity (NDP)

```
Input:  $model_{(t)}$ is a word2vec embedding space in timestamp $t$, $w_i$ is a string representation of a given local topic, $w_A$ is a string representation of the user-defined concept A, $w_C$ is a string representation of the user-defined concept C, $N_A$ is the most recent neighbourhood of concept A, and $N_C$ is the most recent neighbourhood of concept C
Output: $d^{\mathrm{NDP}}({w_i}^{(t)}, {w_A}^{(t)}, {w_C}^{(t)})$
Start
        1. Check if the concepts $w_i$, $w_A$ and $w_C$ are present in the $model_{(t)}$
        2. Get the similarity vectors of the focus concept $w_i$, the user-defined concepts ($w_A$ and $w_C$), and recent neighbourhood ($N_A$ and $N_C$) from the $model_{(t)}$
        3. Compute the cosine distances between the similarity vectors of the focus concept $w_i$ and $w_A$, focus concept $w_i$ and $N_A$, focus concept $w_i$ and $w_C$, and focus concept $w_i$ and $N_C$
        4. Compute the relative distance using the derived cosine distances
End
```

Table A.1: MAP@k results for the five golden test cases: FO-RD, MG-MIG, IGF1-ARG, AD-INN and SZ-PA2

Method	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**
AR (baseline)	0.665	0.582	0.555	0.523	0.513	0.504	0.486	0.487	0.477	0.467
BI (baseline)	0.0	0.0	0.002	0.007	0.019	0.029	0.038	0.045	0.055	0.065
DE (baseline)	0.246	0.209	0.23	0.233	0.237	0.232	0.237	0.241	0.239	0.233
SE (baseline)	0.083	0.157	0.203	0.213	0.226	0.234	0.233	0.241	0.251	0.256
TI (baseline)	0.031	0.039	0.047	0.065	0.087	0.101	0.106	0.114	0.127	0.133
DTM: LSTM_1	0.311	0.409	0.441	0.465	0.476	0.494	0.503	0.513	0.524	0.532
DTM: LSTM_2	0.437	0.451	0.452	0.462	0.479	0.487	0.497	0.504	0.507	0.515
DTM: LSTM_3	0.364	0.443	0.468	0.46	0.488	0.5	0.51	0.514	0.521	0.517
DTM: CNN	0.521	0.469	0.487	0.512	0.522	0.52	0.528	0.529	0.538	0.541
DTM: CNN_LSTM	0.592	0.545	0.529	0.515	0.508	0.507	0.52	0.524	0.535	0.547
DTM: LSTM_CNN	0.362	0.368	0.391	0.402	0.417	0.435	0.435	0.449	0.455	0.461
FTM	0.788	0.783	**0.77**	0.751	0.745	0.72	0.717	0.696	0.686	0.68
TAM	**0.851**	**0.792**	0.763	**0.761**	**0.755**	**0.746**	**0.73**	**0.72**	**0.704**	**0.697**

Appendix B

Reusability

Table B.1: MAP@k results for the four golden test cases

Method	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**
AR (baseline)	0.157	0.148	0.146	0.135	0.129	0.127	0.125	0.123	0.124	0.12
DE (baseline)	0.117	0.081	0.06	0.06	0.055	0.049	0.049	0.048	0.048	0.047
SE (baseline)	0.033	0.029	0.021	0.02	0.022	0.023	0.022	0.02	0.021	0.022
CN (baseline)	0.0	0.0	0.0	0.0	0.0	0.0	0.001	0.001	0.001	0.001
JI (baseline)	0.008	0.007	0.004	0.004	0.004	0.004	0.003	0.003	0.003	0.003
PA (baseline)	0.0	0.0	0.001	0.002	0.003	0.003	0.004	0.005	0.006	0.006
DTM: LSTM_1	0.43	0.331	0.32	0.297	0.293	0.295	0.284	0.277	0.276	0.274
DTM: LSTM_2	0.259	0.204	0.202	0.206	0.21	0.219	0.22	0.223	0.227	0.227
DTM: LSTM_3	0.215	0.243	0.236	0.226	0.225	0.232	0.228	0.23	0.224	0.219
DTM: CNN	0.276	0.193	0.184	0.175	0.162	0.15	0.15	0.147	0.143	0.141
DTM: CNN_LSTM	0.306	0.254	0.265	0.246	0.244	0.246	0.246	0.245	0.24	0.232
DTM: LSTM_CNN	0.279	0.277	0.252	0.259	0.254	0.25	0.236	0.235	0.233	0.229
FTM	**0.459**	**0.372**	**0.362**	**0.346**	**0.327**	**0.314**	**0.296**	**0.291**	**0.279**	**0.276**
TAM	0.422	0.336	0.285	0.254	0.234	0.222	0.205	0.199	0.194	0.187

Table B.2: MAP@k results for the four golden test cases using only drugs

Method	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**
AR (baseline)	0.058	0.063	0.062	0.065	0.067	0.071	0.07	0.068	0.067	0.067
DE (baseline)	0.102	0.062	0.051	0.047	0.047	0.045	0.045	0.042	0.041	0.041
SE (baseline)	0.005	0.011	0.01	0.014	0.017	0.017	0.017	0.019	0.02	0.02
CN (baseline)	0.0	0.0	0.0	0.001	0.002	0.003	0.003	0.003	0.005	0.005
JI (baseline)	0.031	0.015	0.011	0.009	0.007	0.006	0.006	0.006	0.007	0.006
PA (baseline)	0.0	0.003	0.003	0.004	0.005	0.005	0.006	0.006	0.008	0.009
DTM: LSTM_1	0.364	0.306	0.266	0.253	0.245	0.244	0.246	0.232	0.227	0.22
DTM: LSTM_2	0.234	0.189	0.191	0.19	0.192	0.189	0.192	0.198	0.195	0.194
DTM: LSTM_3	0.203	0.184	0.17	0.17	0.168	0.163	0.157	0.154	0.148	0.15
DTM: CNN	0.261	0.198	0.172	0.151	0.141	0.128	0.122	0.13	0.126	0.125
DTM: CNN_LSTM	0.259	0.224	0.235	0.222	0.207	0.194	0.19	0.187	0.185	0.187
DTM: LSTM_CNN	0.158	0.18	0.204	0.195	0.185	0.19	0.188	0.184	0.176	0.175
FTM	**0.47**	**0.392**	**0.354**	**0.325**	**0.303**	**0.28**	**0.263**	**0.254**	**0.235**	**0.228**
TAM	0.37	0.263	0.23	0.21	0.187	0.175	0.17	0.167	0.164	0.159

Appendix C

Portability

Table C.1: Several predicates from the DBpedia RDF graph on the subject *"Pulmonary hypertension"* (note that the property values indicate similar meanings to those in Table 7.2's *'comments'* column)

No.	**Property *(Predicate)***	**Value *(Object)***
1	*dbo:abstract*	Pulmonary hypertension (PH or PHTN) is an increase of blood pressure in the pulmonary artery, pulmonary vein, or pulmonary capillaries, together known as the lung vasculature, leading to shortness of breath, dizziness, fainting, ...
2	*dbo:icd10*	I27.0, I27.2
3	*dbo:icd9*	416.0, 416.8
4	*dbo:meshId*	D006976
5	*dbo:wikiPageExternalLink*	*http://www.cirquemeded.com/ACCP/CHEST2005/CoTherix/player.html* *http://www.merckmanuals.com/home/lung_and_airway_disorders/pulmonary_hypertension/pulmonary_hypertension.html#v727742*, *http://www.phaeurope.org/*, *http://www.phassociation.org/Page.aspx?pid=197*, *http://www.phcentral.org/*, *http://www.phaaustralia.com.au*, *http://www.annals.org/cgi/reprint/143/4/282*, *http://emedicine.medscape.com/article/1004828-overview*, *http://emedicine.medscape.com/article/303098-overview*, *http://emedicine.medscape.com/article/898437-overview*, *http://bioinfo.mc.vanderbilt.edu/PAHKB/* *http://www.ncbi.nlm.nih.gov/omim/178600,600799,178600,600799*
6	*dbo:wikiPageID*	674529
7	*dbp:diseasesdb*	10998
8	*dbp:field*	*dbr:Cardiology* *dbr:Pulmonology*

9	*dbp:wordnet_type*	*http://www.w3.org/2006/03/wn/wn20/instances/synset-disease-noun-1*
10	*dct:subject*	*dbc:Hypertension* *dbc:Pulmonary_heart_disease_and_diseases_of_pulmonary_circulation*
11	*rdf:type*	*owl:Thing*, *wikidata:Q12136*, *dbo:Disease*, *yago:Abstraction100002137*, *yago:Attribute100024264*, *yago:Condition113920835*, *yago:Disease114070360*, *yago:Disorder114052403*, *yago:IllHealth114052046*, *yago:Illness114061805*, *yago:PathologicalState114051917*, *yago:PhysicalCondition114034177*, *yago:State100024720*, *yago:WikicatLungDisorders*
12	*owl:sameAs*	*wikidata:Pulmonary hypertension*, *dbpedia-de:Pulmonary hypertension*, *dbpedia-es:Pulmonary hypertension*, *dbpedia-fr:Pulmonary hypertension*, *dbpedia-it:Pulmonary hypertension*, *dbpedia-ja:Pulmonary hypertension*, *dbpedia-pl:Pulmonary hypertension*, *dbpedia-wikidata:Pulmonary hypertension*, *dbpedia-nl:Pulmonary hypertension*, *dbpedia-pt:Pulmonary hypertension*, *http://www4.wiwiss.fu-berlin.de/sider/resource/side_*, *effects/C0020542*, *freebase:Pulmonary hypertension*, *http://purl.org/net/tcm/tcm.lifescience.ntu.edu.tw/id*, */disease/Hypertension_Pulmonary*, *yago-res:Pulmonary hypertension*
13	*is dbo:wikiPage Redirects of*	*dbr:CTEPH*, *dbr:Cteph*, *dbr:Primary_pulmonary_hypertension*, *dbr:Chronic_thromboembolic_pulmonary_hypertension*, *dbr:Pulmonary_Hypertension*, *dbr:Pulmonary_artery_hypertension*, *dbr:Secondary_pulmonary_hypertension*, *dbr:PHTN*, *dbr:Pulmonary_arterial_hypertension*, *dbr:Ayerza_syndrome*, *dbr:Hypertension,_pulmonary*, *dbr:Persistent_pulmonary_hypertension*, *dbr:Pulmonary_htn*, *dbr:Pulmonary_hypertension,_secondary*, *dbr:Pulmonary_Arterial_Hypertension*, *dbr:Pulmonary_Hypertension,_Secondary*

Table C.2: Several predicates from the DBpedia RDF graph on the subject *“Big Five personality traits”* (note that the property values indicate similar meanings to those in Table 7.2’s *‘comments’* column)

No.	**Property *(Predicate)***	**Value *(Object)***
1	*dbo:abstract*	The Big Five personality traits, also known as the five factor model (FFM), is a model based on common language descriptors of personality (lexical hypothesis). These descriptors are grouped together using a statistical technique ...
2	*dbo:wikiPageID*	1284664
3	*dct:subject*	*dbc:Personality_traits*

4	*owl:sameAs*	*wikidata:Big Five personality traits*, *dbpedia-cs:Big Five personality traits*, *dbpedia-de:Big Five personality traits*, *dbpedia-es:Big Five personality traits*, *dbpedia-fr:Big Five personality traits*, *dbpedia-it:Big Five personality traits*, *dbpedia-pl:Big Five personality traits*, *dbpedia-pt:Big Five personality traits*, *dbpedia-wikidata:Big Five personality traits*, *dbpedia-ko:Big Five personality traits*, *dbpedia-nl:Big Five personality traits*, *freebase:Big Five personality traits*, *yago-res:Big Five personality traits*
5	*is dbo:wikiPage Redirects of*	*dbr:Big_Five_Inventory*, *dbr:Big_Five_model_of_personality*, *dbr:Big_Five_personality_factors*, *dbr:OCEAN_model*, *dbr:OCEAN_model_of_personality*, *dbr:Five_factor_model*, *dbr:Five_Factor_Model*, *dbr:Big_five_personality_traits*, *dbr:OCEAN*, *dbr:"Big_Five"_factors*, *dbr:Big_Five_factors*, *dbr:Big_Five_test*, *dbr:Five_Factor_Personality_Test*, *dbr:'Five_Factor'_personality_test*, *dbr:Five_factor_inventory*, *dbr:The_Big_Five_personality_traits*, *dbr:Five_factor_model_of_personality*, *dbr:Big_Five_personality*, *dbr:Five-factor_model*, *dbr:Big_Five_Personality_Traits*, *dbr:Big_Five_model*
6	*is rdfs:seeAlso of*	*dbr:Personality_disorder*
7	*dbo:wikiPage ExternalLink*	*http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0122245* *http://www.ocf.berkeley.edu/ johnlab/bigfive.htm*, *http://ipip.ori.org/*
8	*rdf:type*	*yago:Abstraction100002137*, *yago:Attribute100024264*, *yago:Cognition100023271*, *yago:Explanation105793000*, *yago:HigherCognitiveProcess105770664*, *yago:Process105701363*, *yago:PsychologicalFeature100023100*, *yago:Theory105989479*, *yago:Thinking105770926*, *yago:Trait104616059*, *yago:WikicatPersonalityTheories*, *yago:WikicatPersonalityTraits*

Table C.3: Several predicates from the DBpedia RDF graph on the subject *"Bloom's taxonomy"* (note that the property values indicate similar meanings to those in Table 7.2's *'comments'* column)

No.	**Property *(Predicate)***	**Value *(Object)***
1	*dbo:abstract*	Bloom's taxonomy is a set of three hierarchical models used to classify educational learning objectives into levels of complexity and specificity. The three lists cover the learning objectives in cognitive, affective and sensory ...
2	*dbo:wikiPageID*	261128
3	*dct:subject*	*dbc:Stage_theories* *dbc:Classification_systems* *dbc:Educational_psychology* *dbc:Educational_technology*

4	*owl:sameAs*	*wikidata:Bloom's taxonomy*, *dbpedia-cs:Bloom's taxonomy*, *dbpedia-el:Bloom's taxonomy*, *dbpedia-es:Bloom's taxonomy*, *dbpedia-fr:Bloom's taxonomy*, *dbpedia-pl:Bloom's taxonomy*, *dbpedia-pt:Bloom's taxonomy*, *dbpedia-id:Bloom's taxonomy*, *dbpedia-it:Bloom's taxonomy*, *dbpedia-wikidata:Bloom's taxonomy*, *freebase:Bloom's taxonomy*, *yago-res:Bloom's taxonomy*, *yago-res:Bloom's taxonomy*
5	*is dbo:wikiPage Redirects of*	*dbr:Blooms_taxonomy*, *dbr:Taxonomy_of_Education_Objectives*, *dbr:Bloom's_Toxonomy*, *dbr:Bloom's_Taxonomy*, *dbr:Taxonomy_of_educational_objectives*, *dbr:Taxonomy_of_education_objectives*, *dbr:Blooms_Taxonomy_In_Education*, *dbr:Taxonomy_of_Educational_Objectives*
6	*dbo:wikiPage ExternalLink*	*http://www.nwlink.com/ donclark/hrd/bloom.html*
7	*is rdfs:seeAlso of*	*dbr:Scientific_literacy*
8	*rdf:type*	*yago:Abstraction100002137*, *yago:Arrangement105726596*, *yago:ClassificationSystem105727220*, *yago:Cognition100023271*, *yago:PsychologicalFeature100023100*, *yago:Structure105726345*, *yago:WikicatClassificationSystems*

Table C.4: Qualitative evaluation of synonym coverage (includes the redirects that are directly linked to the main Wikipedia page, i.e., redirects with 'no anchor')

Test case No.	**Topic**	**Resource**	**Synonyms**
(1)	FO	MeSH	Fish Oils, Fish Liver Oils, Fish Oil
		DBpedia	Fish oil, Fish oils, Fish-oil, Lovanza, Marine oil, Fish liver oils
		WordNet	Fish oil, Fish-liver oil
	RD	MeSH	Raynaud Disease, Hereditary Cold Fingers, Raynaud Phenomenon, Raynaud's Disease
		DBpedia	Raynaud syndrome, Raynaud's disease and Raynaud's phenomenon, Reynaud's, Reynaud's disease, Raynaud's disease, Raynaud phenomenon, Reynaud's phenomenon, Raynauds disease, Reynaud's Disease, Raynaud's disorder, Intermittent arterial vasospasm, Raynaud's Disease, Raynaud's syndrome, Raynaud's phenomenon, Raynaud's disease/phenomenon, Raynaud disease, Raynauds, Raynauds Syndrome, Raynauld's syndrome, Raynauld syndrome,

			Reynaud's phenomenon, Primary Raynaud's phenomenon, Raynaud's Phenomenon, Raynaud's Syndrome, Reynaud's Syndrome, Reynaud's syndrome, Primary raynaud's phenomenon, Secondary raynaud's phenomenon, Raynaud's Syndrome, Raynaud's
		WordNet	Raynaud's sign, Acrocyanosis
(2)	MG	MeSH	Magnesium
		DBpedia	Magnesium, Magnessium, Magnesium compounds, Magnesium ribbon, Element 12, Mg2+, C8H14MgO10, Mg^{2+}
		WordNet	Magnesium, Atomic number 12
	MIG	MeSH	Migraine Disorders, Abdominal Migraine, Acute Confusional Migraine, Cervical Migraine Syndrome, Migraine Headache, Hemicrania Migraine, Migraine, Migraine Headache, Migraine Variant, Sick Headache, Status Migrainosus
		DBpedia	Migraine, Migraines, Basilar migraine, Basilar type migraine, Migraine headaches, Facial migraine, Migrane, Migraine treatment drug, Migraine headache, Mígren, Mígreni, Bickerstaff's migraine, Classical migraine, Common migraine, Optical migraine, Migraine disorders, Anti-migraine, Migraine medication, Migreni, Migren, Migraine journal, Acute migraine, Megrims, Chronic migraine, Status migraine
		WordNet	Migraine, Megrim, Hemicrania, Sick headache
(3)	IGF1	MeSH	Insulin-Like Growth Factor I, IGF-1, IGF-I, IGF-I-SmC, Insulin Like Growth Factor I, Insulin-Like Somatomedin Peptide I, Somatomedin C
		DBpedia	Insulin-like growth factor 1, Mechano-growth factor, Insulin-like growth factor-1, IGF-I, Somatomedin C, IGF type 1 receptor, Insulin-like Growth Factor 1, Insulin-like growth factor I, IGF-1, Insulinlike growth factor I, IGF1, Insulin-like growth factor i, Sulfation factor, IGF1 (gene)
		WordNet	–
	ARG	MeSH	Arginine, Arginine Hydrochloride, Arginine, L-Isomer, DL-Arginine Acetate, Monohydrate, L-Arginine

		DBpedia	Arginine, Arginin, L-Arginine L-malate, L-arginine hydrochloride, Arginine hydrochloride, L-arginine, Arginate, L-Arginine, D-arginine, 1-Arginine, Arganine, Arginine malate, L-Arginine Malate, Argamine, Argivene, Detoxargin, Levargin, Minophagen A, L-Arg, Argenine
		WordNet	Arginine
(4)	AD	MeSH	Alzheimer Disease, Acute Confusional Senile Dementia, Alzheimer Dementia, Early Onset Alzheimer Disease, Late Onset Alzheimer Disease, Alzheimer Sclerosis, Alzheimer Syndrome, Alzheimer Type Senile Dementia, Alzheimer's Disease, Focal Onset Alzheimer's Disease, Alzheimer-Type Dementia (ATD), Alzheimer Type Dementia, Presenile Dementia, Primary Senile Degenerative Dementia, Senile Dementia, Early Onset Alzheimer Disease, Familial Alzheimer Disease (FAD), Focal Onset Alzheimer's Disease, Late Onset Alzheimer Disease, Presenile Alzheimer Dementia, Primary Senile Degenerative Dementia, Acute Confusional Senile Dementia, Alzheimer Type Senile Dementia
		DBpedia	Alzheimer's disease, Alzheimers, Alzhiemer's disease, Alzheimer's, Alzheimer's diseases, Alstimers, Altzimers, Alzheimer disease, Alzeihmers, Alzheimer's Disease, Altimers, Alzhimer, Alzhiemers, Alzheimers disease, Old timer's disease, Old timer disease, Oldtimer disease, Alzheimer's disease, Late-onset Alzheimer's Disease, Alzhemiers' disease, Presenile dementia, Old timers disease, Oldtimer's disease, Alzheimer's, Alzheimer, DAT - Dementia Alzheimer's type, Cognitive disease, Sdat, Alzheimer's dementia, Altzheimer, Alzheimer's diseases, Alzheimer's Research, Alzheimer dementia, Alzeheimer's, Alzeheimers, Alzheimer's Disease, Alzheimers Disease, Oldtimers disease, Retrogenesis, Old-timer's disease, Old-timers' disease, Alzheimer's Syndrome, Alzheimer's research directions, Alzheimer's Disease and Diet, Alzheimer's disease and diet, Alzheimer's syndrome, Primary degenerative dementia of the Alzheimer's type, Senile dementia of the Alzheimer type, Retrogenesis theory, Alzeimer's, Alzeimer's disease, Alzeimers, Alzeimers disease
		WordNet	Alzheimers, Alzheimer's disease, Alzheimer's

	INN	MeSH	Indomethacin, Amuno, Indocid, Indocin, Indomet 140, Indometacin, Indomethacin Hydrochloride, Metindol, Osmosin
		DBpedia	Indometacin, Indocin sr, ATC code C01EB03, ATCvet code QC01EB03, ATC code M01AB01, ATC code M02AA23, ATC code S01BC01, ATCvet code QM01AB01, ATCvet code QM02AA23, ATCvet code QS01BC01, Indomethacin sodium, Indocid, Indocin, Indomethacin, C19H16ClNO4, Indomethacin antenatal infection, Indophtal, Indomee, Amuno, Apo-Indomethacin, Arthrexin, Artracin, Artrinovo, Artrivia, Bonidin, Bonidon, Bonidon Gel, Catlep, Chibro-Amuno, Chrono-Indicid, Chrono-Indocid, Confortid, Dolcidium, Dolcidium Pl, Dolovin, Durametacin, Elmetacin, Flexin Continus, Hicin, Idomethine, Imbrilon, Inacid, Indacin, Indameth, Indmethacine, Indo-Lemmon, Indo-Phlogont, Indo-Rectolmin, Indo-Spray, Indo-Tablinen, Indocid Pda, Indocid Sr, Indocin I.V, Indocin I.V., Indocin Sr, Indolar Sr, Indomecol, Indomed, Indomethegan, Indomo, Indomod, Indoptic, Indoptol, Indorektal, Indoxen, Inflazon, Infrocin, Inteban Sp, Lausit, Liometacen, Metacen, Metartril, Methazine, Metindol, Miametan, Mikametan, Mobilan, Novo-Methacin, Novomethacin, Nu-Indo, Reumacide, Rhemacin La, Rheumacin La, Sadoreum, Tannex
		WordNet	Indomethacin, Indocin
(5)	SZ	MeSH	Schizophrenia, Dementia Praecox, Schizophrenic Disorders
		DBpedia	Schizophrenia, Schyzophrenia, Skitzafrenic, Schizophrene, Schizofrenia, Schizophrenic disorders, Schizophrenia, genetic types, Pathology of Schizophrenia, Schizophernia, Schizophrenic, Schitzo, Schitzophrenia, Scizophrenia, Schizo, Schizophrenic narcissism, Schizophrenics, Simple schizophrenia, Skitzophrenia, Skitsafrantic, Schizophrenia: Symptoms, Paranoid schizophrenics, SCZ, Integration disorder syndrome, Integration disorder, Failure to recognize what is real
		WordNet	Schizophrenic psychosis, Schizophrenic disorder, Schizophrenia, Dementia praecox
	PA2	MeSH	Phospholipases A2, Lecithinase A2, Phospholipase A2

		DBpedia	Phospholipase A2, PLA2, Phospholipase a2, EC 3.1.1.4
		WordNet	–

Table C.5: Qualitative evaluation of synonym coverage in non-medical settings

Topic	**Resource**	**Synonyms**
Genetic algorithms	MeSH	–
	DBpedia	Genetic algorithm, Genetic algorithms, Darwinian algorithm, GATTO, Building block hypothesis, Theory of genetic algorithms, Genetic Algorithm, Genetic Algorithms, GEGA, Genethc algorithm
	WordNet	–
Pattern recognition	MeSH	Automated Pattern Recognition, Pattern Recognition System
	DBpedia	Pattern recognition, Pattern analysis, Visual pattern recognition, Pattern Recognition, Machine pattern recognition, Pattern recognition and learning, Pattern-recognition, Pattern Recognition and Learning, Pattern recognition (machine learning)
	WordNet	–
Virtual reality	MeSH	Virtual Reality, Educational Virtual Reality, Instructional Virtual Reality
	DBpedia	Virtual reality, Virtual environment, 3d simulation, Computer-simulated environment, Computer simulated environment, Simulated environment, Virtuality, Virtual Reality, Neuron Interactive Virtual Reality, Virtual space, Virtual ⋆ terms, Virtual environments, Virtual gaming, Computer-generated environment, Virtual reality (VR), Virtual-reality, Virtual realities
	WordNet	Virtual reality
Reinforcement learning	MeSH	–
	DBpedia	Reinforcement learning, Reward function, Reinforcement Learning, Actor critic architecture, Actor critic model, Reinforcement Learning a form of Artificial Intelligence, Inverse reinforcement learning, Learning from demonstration
	WordNet	–
Text retrieval	MeSH	–
	DBpedia	Text retrieval, Document retrieval system, Document retrieval
	WordNet	–
Cluster	MeSH	Cluster Analysis, Clustering, Disease Clustering

analysis	DBpedia	Cluster analysis, Cluster analyses, Cluster Analysis, Clustered data, Clustering algorithm, Data clustering, Clustering metric, Cluster validation, Cluster (statistics), Data Clustering, Agglomerative clustering
	WordNet	Clustering, Cluster, Bunch
Image segmentation	MeSH	–
	DBpedia	Image segmentation, Segmentation (image processing), Image segment
	WordNet	–
Speech recognition	MeSH	Speech Recognition Software, Voice Recognition Software
	DBpedia	Speech recognition, Automatic speech recognizer, Speech recognizer, Voice-to-text, Voice to text, Speech to Text, Computer speech recognition, Voice recognition software, Automatic Speech Recognition, Speech-recognition, Speech-to-text, Automatic speech recognition, Speech Recognition, Spoken word recognition, Voice command, Voice typing, Voice dialing, Voice Command, Voice Recognition Command System, Speech-to-Text, Speech recognition technology, Speech recognition software, Automated speech recognition
	WordNet	–
Signal processing	MeSH	Computer-Assisted Signal Processing, Digital Signal Processing, Computer-Assisted Signal Interpretation
	DBpedia	Signal processing, Signals processing, Signal analysis, Signal Processing, Signal processor, Signal theory, Signal processsing, Multiscale signal analysis, Signal conditioner, Signal Processor
	WordNet	Signal detection
Machine vision	MeSH	–
	DBpedia	Machine vision, Machine Sight, Machine Vision, Visual navigation
	WordNet	–
Gravitational lens	MeSH	–
	DBpedia	Gravitational lens, Gravitational Lenses, Bend light, Gravitationally lensed galaxy, Einstein arc, Gravitational Lensing, Gravitational arc, Gravity lens, Gravitational lensing, Gravitational Lens, Gravitational lense, Gravitational lenses, Multiple images (gravitational lensing), Gravitatinal lensing, Macrolensing, Gravitational deflection
	WordNet	–
Inverse Galo-	MeSH	–

is problem	DBpedia	Inverse Galois problem, Inverse problem of Galois theory, Inverse Galois theory, Rigid group
	WordNet	Galois theory
Oligopoly	MeSH	–
	DBpedia	Oligopoly, Oligopolies, Desoligopolization, Desologopolization, Oligopolistic, Oligolopolistic, Oligopolist, Desoligolipolization, Oligopology, Oligopoly theory, Oligopolists
	WordNet	Oligopoly

Bibliography

Abián, D., Guerra, F., Martínez-Romanos, J. & Trillo-Lado, R. (2017), Wikidata and DBpedia: A Comparative Study, *in* 'Semanitic Keyword-based Search on Structured Data Sources', Springer, pp. 142–154.

Ahlers, C. B., Hristovski, D., Kilicoglu, H. & Rindflesch, T. C. (2007), Using the Literature-Based Discovery Paradigm to Investigate Drug Mechanisms, *in* 'AMIA Annual Symposium', Vol. 2007, American Medical Informatics Association, p. 6.

Ahmaro, I. Y., Abualkishik, A. M. & Yusoff, M. Z. M. (2014), 'Taxonomy, Definition, Approaches, Benefits, Reusability Levels, Factors and Adaption of Software Reusability: A Review of the Research Literature', *Journal of Applied Sciences* **14**(20), 2396.

Ahmed, A. (2016), 'Literature-Based Discovery: Critical Analysis and Future Directions', *International Journal of Computer Science and Network Security (IJCSNS)* **16**(7), 11.

Aizawa, A. (2003), 'An Information-theoretic Perspective of TF–IDF Measures', *Information Processing & Management* **39**(1), 45–65.

Al-Rfou', R., Perozzi, B. & Skiena, S. (2013), Polyglot: Distributed Word Representations for Multilingual NLP, *in* '17$^{\text{th}}$ Conference on Computational Natural Language Learning (CoNLL)', Association for Computational Linguistics, pp. 183–192.

Allen, C. & Hospedales, T. (2019), 'Analogies Explained: Towards Understanding Word Embeddings', *arXiv preprint arXiv:1901.09813* .

Andronis, C., Sharma, A., Virvilis, V., Deftereos, S. & Persidis, A. (2011), 'Literature Mining, Ontologies and Information Visualization for Drug Repurposing', *Briefings in Bioinformatics* **12**(4), 357–368.

Anzai, Y. (2012), *Pattern Recognition and Machine Learning*, Elsevier.

Aprosio, A. P., Giuliano, C. & Lavelli, A. (2013), Automatic Mapping of Wikipedia Templates for Fast Deployment of Localised DBpedia Datasets, *in* '13th International Conference on Knowledge Management and Knowledge Technologies (i-KNOW)', ACM, pp. 1–8.

Aronson, A. R. (2001), Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program, *in* 'AMIA Annual Symposium', American Medical Informatics Association, p. 17.

Aronson, A. R. & Lang, F.-M. (2010), 'An Overview of MetaMap: Historical Perspective and Recent Advances', *Journal of the American Medical Informatics Association* **17**(3), 229–236.

Atzori, M. & Dessi, A. (2014), Ranking DBpedia Properties, *in* '23rd International Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises', IEEE, pp. 441–446.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R. & Ives, Z. (2007), DBpedia: A Nucleus for a Web of Open Data, *in* 'The Semantic Web', Springer, pp. 722–735.

Ayyadevara, V. K. (2019), *Neural Networks with Keras Cookbook: Over 70 recipes leveraging deep learning techniques across image, text, audio, and game bots*, Packt Publishing Ltd.

Azad, H. K. & Deepak, A. (2017), 'Query Expansion Techniques for Information Retrieval: A Survey', *arXiv preprint arXiv:1708.00247* .

Baek, S. H., Lee, D., Kim, M., Lee, J. H. & Song, M. (2017), 'Enriching Plausible New Hypothesis Generation in PubMed', *PLOS ONE* **12**(7), e0180539.

Bahadoran, Z., Mirmiran, P., Kashfi, K. & Ghasemi, A. (2019), 'The Principles of Biomedical Scientific Writing', *International Journal of Endocrinology and Metabolism* **17**(4).

Baker, N. C. & Hemminger, B. M. (2010), 'Mining Connections between Chemicals, Proteins, and Diseases Extracted from Medline Annotations', *Journal of Biomedical Informatics* **43**(4), 510–519.

Banerjee, R., Choi, Y., Piyush, G., Naik, A. & Ramakrishnan, I. (2014), Automated Suggestion of Tests for Identifying Likelihood of Adverse Drug Events, *in* 'International Conference on Healthcare Informatics (ICHI)', IEEE, pp. 170–176.

Bannach-Brown, A., Przybyła, P., Thomas, J., Rice, A. S., Ananiadou, S., Liao, J. &

Macleod, M. R. (2019), ‘Machine Learning Algorithms for Systematic Review: Reducing Workload in a Preclinical Review of Animal Studies and Reducing Human Screening Error’, *Systematic Reviews* **8**(1), 1–12.

Bates, M. J. (2005), ‘Information and Knowledge: An Evolutionary Framework for Information Science’, *Information Research* **10**(4), n4.

Beitzel, S. M., Jensen, E. C. & Frieder, O. (2009*a*), *GMAP*, Springer, pp. 1256–1256.

Beitzel, S. M., Jensen, E. C. & Frieder, O. (2009*b*), *MAP*, Springer US, pp. 1691–1692.

Bekhuis, T. (2006), ‘Conceptual Biology, Hypothesis Discovery, and Text Mining: Swanson’s Legacy’, *Biomedical Digital Libraries* **3**(1), 2.

Berardi, M., Lapi, M., Leo, P. & Loglisci, C. (2005), Mining Generalized Association Rules on Biomedical Literature, *in* ‘International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (IEA/AIE)’, Springer, pp. 500–509.

Berrar, D. (2019), ‘Cross-validation’, *Encyclopedia of Bioinformatics and Computational Biology* **1**, 542–545.

Bhattacharya, S. & Srinivasan, P. (2012), A Semantic Approach to Involve Twitter in LBD Efforts, *in* ‘International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)’, IEEE, pp. 248–253.

Biamonte, J., Wittek, P., Pancotti, N., Rebentrost, P., Wiebe, N. & Lloyd, S. (2017), ‘Quantum Machine Learning’, *Nature* **549**(7671), 195–202.

Bills, M. V., Loh, A., Sosnowski, K., Nguyen, B. T., Ha, S. Y., Yim, U. H. & Yoon, J.-Y. (2020), ‘Handheld UV Fluorescence Spectrophotometer Device for the Classification and Analysis of Petroleum Oil Samples’, *Biosensors and Bioelectronics* p. 112193.

Bisgin, H., Liu, Z., Fang, H., Xu, X. & Tong, W. (2011), ‘Mining FDA Drug Labels using an Unsupervised Learning Technique – Topic Modeling’, *BMC Bioinformatics* **12**, S11.

Bizer, C., Heath, T. & Berners-Lee, T. (2011), Linked Data: The Story so far, *in* ‘Semantic services, interoperability and web applications: emerging concepts’, IGI Global, pp. 205–227.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R. & Hellmann, S. (2009), ‘DBpedia – A Crystallization Point for the Web of Data’, *Journal of Web Semantics* **7**(3), 154–165.

Bodenreider, O. (2004), 'The Unified Medical Language System (UMLS): Integrating Biomedical Terminology', *Nucleic Acids Research* **32**, D267–D270.

Boell, S. K. & Cecez-Kecmanovic, D. (2015), On being 'Systematic' in Literature Reviews, *in* 'Formulating Research Methods for Information Systems', Springer, pp. 48–78.

Bornmann, L. & Mutz, R. (2015), 'Growth Rates of Modern Science: A Bibliometric Analysis based on the Number of Publications and Cited References', *Journal of the Association for Information Science and Technology* **66**(11), 2215–2222.

Boukhaled, M., Fagard, B. & Poibeau, T. (2019), A Predictive Approach to Semantic Change Modelling, *in* 'Digital Humanities'.

Breiman, L. (1996), 'Bagging Predictors', *Machine learning* **24**(2), 123–140.

Broughton, J. (2008), *Wikipedia: The Missing Manual*, " O'Reilly Media, Inc.".

Brown, R. B. (2020), 'Breakthrough Knowledge Synthesis in the Age of Google', *Philosophies* **5**(1), 4.

Bruza, P., Cole, R., Song, D. & Bari, Z. (2006), 'Towards Operational Abduction from a Cognitive Perspective', *Logic Journal of IGPL* **14**(2), 161–177.

Bruza, P., Song, D. & McArthur, R. (2004), 'Abduction in Semantic Space: Towards a Logic of Discovery', *Logic Journal of IGPL* **12**(2), 97–109.

Buckland, M. K. (1991), 'Information as Thing', *Journal of the American Society for Information Science* **42**(5), 351–360.

Bunescu, R. C. & Mooney, R. J. (2007), Multiple Instance Learning for Sparse Positive Bags, *in* '24$^{\text{th}}$ International Conference on Machine Learning (ICML)', ACM, pp. 105–112.

Cairelli, M. J., Fiszman, M., Zhang, H. & Rindflesch, T. C. (2015), 'Networks of Neuroinjury Semantic Predications to Identify Biomarkers for Mild Traumatic Brain Injury', *Journal of Biomedical Semantics* **6**(1), 25.

Cairelli, M. J., Miller, C. M., Fiszman, M., Workman, T. E. & Rindflesch, T. C. (2013), Semantic MEDLINE for Discovery Browsing: Using Semantic Predications and the Literature-Based Discovery Paradigm to Elucidate a Mechanism for the Obesity Paradox, *in* 'AMIA Annual Symposium', Vol. 2013, American Medical Informatics Association, p. 164.

Cameron, D., Bodenreider, O., Yalamanchili, H., Danh, T., Vallabhaneni, S., Thirunarayan, K., Sheth, A. P. & Rindflesch, T. C. (2013), 'A Graph-based Recovery and Decomposition of Swanson's Hypothesis using Semantic Predications', *Journal of Biomedical Informatics* **46**(2), 238–251.

Cameron, D., Kavuluru, R., Rindflesch, T. C., Sheth, A. P., Thirunarayan, K. & Bodenreider, O. (2015), 'Context-driven Automatic Subgraph Creation for Literature-Based Discovery', *Journal of Biomedical Informatics* **54**, 141–157.

Capurro, R. & Hjørland, B. (2003), 'The Concept of Information', *Annual Review of Information Science and Technology* **37**(1), 343–411.

Carroll, J. J., Dickinson, I., Dollin, C., Reynolds, D., Seaborne, A. & Wilkinson, K. (2004), Jena: Implementing the Semantic Web Recommendations, *in* '13th International World Wide Web Conference on Alternate Track Papers & Posters (WWW Alt.)', ACM, pp. 74–83.

Celoria, G. C., Friedell, G. H. & Sommers, S. C. (1960), 'Raynaud's Disease and Primary Pulmonary Hypertension', *Circulation* **22**(6), 1055–1059.

Chan, L. M., Intner, S. S. & Weihs, J. (2016), *Guide to the Library of Congress classification*, ABC-CLIO.

Chapman, D. (2009), 'Advanced Search Features of PubMed', *Journal of the Canadian Academy of Child and Adolescent Psychiatry* **18**(1), 58.

Cheadle, C., Cao, H., Kalinin, A. & Hodgkinson, J. (2017), 'Advanced Literature Analysis in a Big Data World', *Annals of the New York Academy of Sciences* **1387**(1), 25–33.

Chen, C. (2016), 'Grand Challenges in Measuring and Characterizing Scholarly Impact', *Frontiers in Research Metrics and Analytics* **1**, 4.

Chen, C.-W., Tseng, S.-P., Kuan, T.-W. & Wang, J.-F. (2020), 'Outpatient Text Classification Using Attention-Based Bidirectional LSTM for Robot-Assisted Servicing in Hospital', *Information* **11**(2), 106.

Chen, D., Peterson, J. C. & Griffiths, T. L. (2017), 'Evaluating Vector-space Models of Analogy', *arXiv preprint arXiv:1705.04416* .

Chen, Y., Ju, J., Chien, M., Sheng, Y., Lee, T. & Chiang, J. (2015), CoTri: Extracting Chemical-disease Relations with Co-reference Resolution and Common Trigger Words, *in* '5th BioCreative Challenge Evaluation Workshop', pp. 286–291.

Cherdioui, S. & Boubekeur, F. (2013), Information Retrieval Techniques for Knowledge Discovery in Biomedical Literature, *in* '11th International Symposium on Programming and Systems (ISPS)', IEEE, pp. 137–142.

Cherrier, B. (2017), 'Classifying Economics: A History of the JEL Codes', *Journal of Economic Literature* **55**(2), 545–79.

Cheung, W. A., Ouellette, B. F. & Wasserman, W. W. (2012*a*), 'Inferring Novel Gene-Disease Associations using Medical Subject Heading Over-representation Profiles', *Genome Medicine* **4**(9), 75.

Cheung, W. A., Ouellette, B. F. & Wasserman, W. W. (2012*b*), 'Quantitative Biomedical Annotation using Medical Subject Heading Over-representation Profiles (MeSHOPs)', *BMC Bioinformatics* **13**(1), 249.

Chiarcos, C., Nordhoff, S. & Hellmann, S. (2012), *Linked Data in Linguistics*, Springer.

Chiarcos, C. & Pareja-Lora, A. (2019), 'Open Data–Linked Data–Linked Open Data–Linguistic Linked Open Data (LLOD): A General Introduction', *Development of Linguistic Linked Open Data Resources for Collaborative Data–Intensive Research in the Language Sciences* .

Choudhury, N., Faisal, F. & Khushi, M. (2020), 'Mining Temporal Evolution of Knowledge Graphs and Genealogical Features for Literature-Based Discovery Prediction', *Journal of Informetrics* **14**(3), 101057.

Ciaramita, M., Murdock, V. & Plachouras, V. (2008), 'Semantic Associations for Contextual Advertising', *Journal of Electronic Commerce Research* **9**(1).

Cohen, A. M. & Hersh, W. R. (2005), 'A Survey of Current Work in Biomedical Text Mining', *Briefings in Bioinformatics* **6**(1), 57–71.

Cohen, T. & Schvaneveldt, R. W. (2010), 'The Trajectory of Scientific Discovery: Concept Co-occurrence and Converging Semantic Distance', *Studies in Health Technology and Informatics* **160**, 661–665.

Cohen, T., Schvaneveldt, R. W. & Rindflesch, T. C. (2009), Predication-based Semantic Indexing: Permutations as a means to Encode Predications in Semantic Space, *in* 'AMIA Annual Symposium', Vol. 2009, American Medical Informatics Association, p. 114.

Cohen, T., Schvaneveldt, R. & Widdows, D. (2010), 'Reflective Random Indexing and Indirect Inference: A Scalable Method for Discovery of Implicit Connections', *Journal of Biomedical Informatics* **43**(2), 240–256.

Cohen, T., Whitfield, G. K., Schvaneveldt, R. W., Mukund, K. & Rindflesch, T. (2010), 'EpiphaNet: An Interactive Tool to Support Biomedical Discoveries', *Journal of Biomedical Discovery and Collaboration* **5**, 21.

Cohen, T., Widdows, D. & Rindflesch, T. (2014), Expansion-by-analogy: A Vector Symbolic Approach to Semantic Search, *in* 'International Symposium on Quantum Interaction (QI)', Springer, pp. 54–66.

Cohen, T., Widdows, D., Schvaneveldt, R. & Rindflesch, T. C. (2011), Finding Schizophrenia's Prozac Emergent Relational Similarity in Predication Space, *in* 'International Symposium on Quantum Interaction (QI)', Springer, pp. 48–59.

Cohen, T., Widdows, D., Schvaneveldt, R. W., Davies, P. & Rindflesch, T. C. (2012), 'Discovering Discovery Patterns with Predication-based Semantic Indexing', *Journal of Biomedical Informatics* **45**(6), 1049–1065.

Cohen, T., Widdows, D., Schvaneveldt, R. W. & Rindflesch, T. C. (2010), Logical Leaps and Quantum Connectives: Forging Paths through Predication Space, *in* 'Symposium on Quantum Informatics for Cognitive, Social, and Semantic Processes', AAAI Press.

Cohen, T., Widdows, D., Stephan, C., Zinner, R., Kim, J., Rindflesch, T. & Davies, P. (2014), 'Predicting High-Throughput Screening Results with Scalable Literature-Based Discovery Methods', *CPT: Pharmacometrics & Systems Pharmacology* **3**(10), 1–9.

Cole, R. J. & Bruza, P. D. (2005), A Bare Bones Approach to Literature-Based Discovery: An Analysis of the Raynaud's/Fish-oil and Migraine-Magnesium Discoveries in Semantic Space, *in* 'International Conference on Discovery Science (DS)', Springer, pp. 84–98.

Collobert, R. & Weston, J. (2008), A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning, *in* '25th International Conference on Machine Learning (ICML)', ACM, pp. 160–167.

Cook, D. J., Mulrow, C. D. & Haynes, R. B. (1997), 'Systematic Reviews: Synthesis of Best Evidence for Clinical Decisions', *Annals of internal medicine* **126**(5), 376–380.

Corrales, D., Ledezma, A. & Corrales, J. (2015), 'A conceptual Framework for Data Quality in Knowledge Discovery Tasks (FDQ-KDT): A Proposal', *JCP* **10**(6), 396–405.

Coyle, K. (2012), 'Semantic Web and Linked Data', *Library Technology Reports* **48**(4), 10–14.

Craswell, N. (2009), *Precision at n*, Springer, pp. 2127–2128.

Crichton, G., Baker, S., Guo, Y. & Korhonen, A. (2020), 'Neural Networks for Open and Closed Literature-based Discovery', *PLOS ONE* **15**(5), e0232891.

Crichton, G., Guo, Y., Pyysalo, S. & Korhonen, A. (2018), 'Neural Networks for Link Prediction in Realistic Biomedical Graphs: A Multi-dimensional Evaluation of Graph Embedding-based Approaches', *BMC Bioinformatics* **19**(1), 176.

Cutler, A., Cutler, D. R. & Stevens, J. R. (2012), Random Forests, *in* 'Ensemble Machine Learning', Springer, pp. 157–175.

Davies, R. (1989), 'The Creation of New Knowledge by Information Retrieval and Classification', *Journal of documentation* **45**(4), 273–301.

Davis, A. P., Grondin, C. J., Johnson, R. J., Sciaky, D., McMorran, R., Wiegers, J., Wiegers, T. C. & Mattingly, C. J. (2019), 'The Comparative Toxicogenomics Database: Update 2019', *Nucleic Acids Research* **47**(D1), D948–D954.

Decker, S., Melnik, S., Van Harmelen, F., Fensel, D., Klein, M., Broekstra, J., Erdmann, M. & Horrocks, I. (2000), 'The Semantic Web: The Roles of XML and RDF', *IEEE Internet Computing* **4**(5), 63–73.

Decker, S., Mitra, P. & Melnik, S. (2000), 'Framework for the Semantic Web: An RDF Tutorial', *IEEE Internet Computing* **4**(6), 68–73.

Dill, S., Eiron, N., Gibson, D., Gruhl, D., Guha, R., Jhingran, A., Kanungo, T., Rajagopalan, S., Tomkins, A., Tomlin, J. A. et al. (2003), SemTag and Seeker: Bootstrapping the Semantic Web via Automated Semantic Annotation, *in* '12th International World Wide Web Conference (WWW)', pp. 178–186.

Ding, Y., Song, M., Han, J., Yu, Q., Yan, E., Lin, L. & Chambers, T. (2013), 'Entitymetrics: Measuring the Impact of Entities', *PLOS ONE* **8**(8), e71416.

Doane, D. P. & Seward, L. E. (2011), 'Measuring Skewness: A Forgotten Statistic?', *Journal of Statistics Education* **19**(2).

Dong, W., Liu, Y., Zhu, W., Mou, Q., Wang, J. & Hu, Y. (2014), 'Simulation of Swanson's Literature-Based Discovery: Anandamide Treatment Inhibits Growth of Gastric Cancer Cells In Vitro and In Silico', *PLOS ONE* **9**(6), e100436.

Du, Q., Gu, W., Zhang, L. & Huang, S.-L. (2018), Attention-based LSTM-CNNs for Time-series Classification, *in* '16th Conference on Embedded Networked Sensor Systems (SenSys)', ACM, pp. 410–411.

Dunne, E. & Hulek, K. (2020), 'Mathematics Subject Classification 2020', *EMS Newsletter* **3**(115), 5–6.

Egger, M. & Smith, G. D. (1997), 'Meta-analysis: Potentials and Promise', *Bmj* **315**(7119), 1371–1374.

El-Amir, H. & Hamdy, M. (2020), Selected Topics in Natural Language Processing, *in* 'Deep Learning Pipeline', Springer, pp. 471–494.

Elgendi, M. (2019), 'Characteristics of a Highly Cited Article: A Machine Learning Perspective', *IEEE Access* **7**, 87977–87986.

Elkins, A., Freitas, F. F. & Sanz, V. (2019), 'Developing an App to interpret Chest X-rays to support the diagnosis of respiratory pathology with Artificial Intelligence', *arXiv preprint arXiv:1906.11282* .

Ermakova, L., Bordignon, F., Turenne, N. & Noel, M. (2018), 'Is the Abstract a mere Teaser? Evaluating Generosity of Article Abstracts in the Environmental Sciences', *Frontiers in Research Metrics and Analytics* **3**, 16.

Erxleben, F., Günther, M., Krötzsch, M., Mendez, J. & Vrandečić, D. (2014), Introducing Wikidata to the Linked Data Web, *in* '13th International Semantic Web Conference (ISWC)', Springer, pp. 50–65.

Eswar, N. & Sali, A. (2009), Protein Structure Modeling, *in* 'From Molecules to Medicines', Springer, pp. 139–151.

Exner, P. & Nugues, P. (2012), Entity Extraction: From Unstructured Text to DBpedia RDF Triples, *in* 'Web of Linked Entities Workshop in conjuction with the 11th International Semantic Web Conference (ISWC)', CEUR, pp. 58–69.

Faro, A., Giordano, D. & Spampinato, C. (2011), 'Combining Literature Text Mining with Microarray Data: Advances for System Biology Modeling', *Briefings in Bioinformatics* **13**(1), 61–82.

Farr, J. N., Jenkins, J. J. & Paterson, D. G. (1951), 'Simplification of Flesch Reading Ease Formula', *Journal of applied psychology* **35**(5), 333.

Fawaz, H. I., Forestier, G., Weber, J., Idoumghar, L. & Muller, P.-A. (2019), 'Deep Learning for Time Series Classification: A Review', *Data Mining and Knowledge Discovery* **33**(4), 917–963.

Fellbaum, C. (2012), 'WordNet', *The Encyclopedia of Applied Linguistics* .

Ferreira, L. G., Dos Santos, R. N., Oliva, G. & Andricopulo, A. D. (2015), 'Molecular Docking and Structure-based Drug Design Strategies', *Molecules* **20**(7), 13384–13421.

Foster, J. G., Rzhetsky, A. & Evans, J. A. (2015), 'Tradition and Innovation in Scientists' Research Strategies', *American Sociological Review* **80**(5), 875–908.

Frangieh, C. G. & Yaacoub, H. K. (2017), 'A Systematic Literature Review of Responsible Leadership', *Journal of Global Responsibility* .

Frijters, R., Van Vugt, M., Smeets, R., Van Schaik, R., De Vlieg, J. & Alkema, W. (2010), 'Literature Mining for the Discovery of Hidden Connections between Drugs, Genes and Diseases', *PLoS Computational Biology* **6**(9), e1000943.

Fu, K. (1968), *Sequential Methods in Pattern Recognition and Machine Learning*, Academic Press.

Fukushima, K. & Miyake, S. (1982), Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Visual Pattern Recognition, *in* 'Competition and Cooperation in Neural Nets', Springer, pp. 267–285.

Gabetta, M., Larizza, C. & Bellazzi, R. (2013), 'A Unified Medical Language System (UMLS) based System for Literature-Based Discovery in Medicine', *Studies in Health Technology and Informatics* **192**, 412–416.

Ganiz, M. C., Pottenger, W. M. & Janneck, C. D. (2005), 'Recent Advances in Literature-Based Discovery', *Journal of the American Society for Information Science and Technology, JASIST* .

Gao, F., Musial, K., Cooper, C. & Tsoka, S. (2015), 'Link Prediction Methods and their Accuracy for Different Social Networks and Network Metrics', *Scientific Programming* **2015**.

Gao, H., Wang, Y., Tao, J., Liu, Z., Li, J., Yu, T., Yu, Q., Tian, Y. & Zhang, H. (2015), Cordycepssinensis may have a Dual Effect on Diabetic Retinopathy, *in* '7th International Conference on Information Technology in Medicine and Education (ITME)', IEEE, pp. 63–67.

Garten, Y., Coulet, A. & Altman, R. B. (2010), 'Recent Progress in Automatically Extracting Information from the Pharmacogenomic Literature', *Pharmacogenomics* **11**(10), 1467–1489.

Gazni, A. (2011), 'Are the Abstracts of High Impact Articles more Readable? Investigating the Evidence from Top Research Institutions in the World', *Journal of Information Science* **37**(3), 273–281.

Geel, M., Church, T. & Norrie, M. (2012), Mix-n-match: Building Personal Libraries from Web Content, *in* 'Theory and Practice of Digital Libraries (TPDL)', Springer, pp. 345–356.

Ghandorh, H., Noorwali, A., Nassif, A. B., Capretz, L. F. & Eagleson, R. (2020), A Systematic Literature Review for Software Portability Measurement: Preliminary Results, *in* '9th International Conference on Software and Computer Applications

(ICSCA)', ACM, pp. 152–157.

Goodwin, J. C., Cohen, T. & Rindflesch, T. (2012), Discovery by Scent: Discovery Browsing System based on the Information Foraging Theory, *in* 'International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)', IEEE, pp. 232–239.

Gopalakrishnan, V., Jha, K., Jin, W. & Zhang, A. (2019), 'A Survey on Literature-Based Discovery Approaches in Biomedical Domain', *Journal of Biomedical Informatics* **93**, 103141.

Gopalakrishnan, V., Jha, K., Xun, G., Ngo, H. Q. & Zhang, A. (2017), 'Towards Self-learning based Hypotheses Generation in Biomedical Text Domain', *Bioinformatics* **34**(12), 2103–2115.

Gordon, M. D. & Dumais, S. (1998), 'Using Latent Semantic Indexing for Literature-Based Discovery', *Journal of the American Society for Information Science* **49**(8), 674–685.

Gordon, M. D. & Lindsay, R. K. (1996), 'Toward Discovery Support Systems: A Replication, Re-examination, and Extension of Swanson's work on Literature-Based Discovery of a Connection between Raynaud's and Fish Oil', *Journal of the American Society for Information Science* **47**(2), 116–128.

Gordon, M., Lindsay, R. K. & Fan, W. (2002), 'Literature-Based Discovery on the World Wide Web', *ACM Transactions on Internet Technology (TOIT)* **2**(4), 261–275.

Gubiani, D., Fabbretti, E., Cestnik, B., Lavrač, N. & Urbančič, T. (2017), 'Outlier based Literature Exploration for Cross-domain Linking of Alzheimer's Disease and Gut Microbiota', *Expert Systems with Applications* **85**, 386–396.

Gulec, F. M., Bicakci, T., Sezer, E. A., Sever, H. & Raghavan, V. V. (2010), Analyzing the Effectiveness of Pruning and Grouping Methods Used in Literature-Based Discovery Tools, *in* 'IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)', Vol. 3, IEEE, pp. 304–308.

Gulli, A. & Kapoor, A. (2017), *TensorFlow 1. x Deep Learning Cookbook: Over 90 unique recipes to solve artificial-intelligence driven problems with Python*, Packt Publishing Ltd.

Guo, W. & Kraines, S. B. (2009*a*), Discovering Relationship Associations in Life Sciences using Ontology and Inference, *in* 'International Conference on Knowledge Discovery and Information Retrieval (KDIR)', pp. 10–17.

Guo, W. & Kraines, S. B. (2009*b*), Extracting Relationship Associations from Semantic

Graphs in Life Sciences, *in* 'International Joint Conference on Knowledge Discovery, Knowledge Engineering, and Knowledge Management (IC3K)', Springer, pp. 53–67.

Guo, Z.-H., You, Z.-H., Huang, D.-S., Yi, H.-C., Zheng, K., Chen, Z.-H. & Wang, Y.-B. (2020), 'MeSHHeading2vec: A New Method for Representing MeSH Headings as Vectors based on Graph Embedding Algorithm', *Briefings in Bioinformatics* .

Hamilton, W. L., Leskovec, J. & Jurafsky, D. (2016*a*), Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change, *in* 'Conference on Empirical Methods in Natural Language Processing (EMNLP)', Association for Computational Linguistics, pp. 2116–2121.

Hamilton, W. L., Leskovec, J. & Jurafsky, D. (2016*b*), Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change, *in* '54$^{\text{th}}$ Annual Meeting of the Association for Computational Linguistics (ACL)', pp. 1489–1501.

Han, K., Yang, P., Mishra, S. & Diesner, J. (2020), Wikicssh: Extracting Computer Science Subject Headings from Wikipedia, *in* 'ADBIS, TPDL and EDA Common Workshops and Doctoral Consortium', Springer, pp. 207–218.

Hanauer, D. A., Saeed, M., Zheng, K., Mei, Q., Shedden, K., Aronson, A. R. & Ramakrishnan, N. (2014), 'Applying MetaMap to Medline for Identifying Novel Associations in a Large Clinical Dataset: A Feasibility Analysis', *Journal of the American Medical Informatics Association* **21**(5), 925–937.

Hartley, J. (2007), 'There's more to the title than meets the eye: Exploring the possibilities', *Journal of Technical Writing and Communication* **37**(1), 95–101.

Hartley, J. (2008), *Academic Writing and Publishing: A Practical Handbook*, Routledge.

Hartley, J., Pennebaker, J. & Fox, C. (2003), 'Abstracts, Introductions and Discussions: How far do they differ in style?', *Scientometrics* **57**(3), 389–398.

Hashimoto, T. B., Alvarez-Melis, D. & Jaakkola, T. S. (2015), 'Word, Graph and Manifold Embedding from Markov Processes', *arXiv preprint arXiv:1509.05808* .

Hashimoto, T. B., Alvarez-Melis, D. & Jaakkola, T. S. (2016), 'Word Embeddings as Metric Recovery in Semantic Spaces', *Transactions of the Association for Computational Linguistics* **4**, 273–286.

He, H., Scheicher, R. H., Pandey, R., Rocha, A. R., Sanvito, S., Grigoriev, A., Ahuja, R. & Karna, S. P. (2008), 'Functionalized Nanopore-embedded Electrodes for Rapid DNA Sequencing', *The Journal of Physical Chemistry C* **112**(10), 3456–3459.

Heath, T. & Bizer, C. (2011*a*), 'Linked Data: Evolving the Web into a Global Data Space', *Synthesis Lectures on the Semantic Web: Theory and Technology* **1**(1), 1–136.

Heath, T. & Bizer, C. (2011*b*), 'Semantic Annotation and Retrieval: Web of Data', *Handbook of Semantic Web Technologies* pp. 191–229.

Heikkilä, J. (2020), 'Classifying Economics for the Common Good: Connecting Sustainable Development Goals to JEL Codes', *SSRN: 3570112* .

Henry, S. (2019), Indirect Relatedness, Evaluation, and Visualization for Literature-Based Discovery, PhD thesis, Virginia Commonwealth University.

Henry, S. & McInnes, B. T. (2017), 'Literature-Based Discovery: Models, Methods, and Trends', *Journal of Biomedical Informatics* **74**, 20–32.

Higgins, J. P. & Green, S. (2008), 'Cochrane Handbook for Systematic Reviews of Interventions'.

Hjørland, B. (2007), 'Information: Objective or Subjective/situational?', *Journal of the American Society for Information Science and Technology* **58**(10), 1448–1456.

Hochreiter, S. & Schmidhuber, J. (1997), 'Long Short-Term Memory', *Neural computation* **9**(8), 1735–1780.

Hope, T., Chan, J., Kittur, A. & Shahaf, D. (2017), Accelerating Innovation through Analogy Mining, *in* '23rd International Conference on Knowledge Discovery and Data Mining (SIGKDD)', pp. 235–243.

Horn, A. L., Cismondi, F., Fialho, A. S., Vieira, S. M., Sousa, J. M., Reti, S., Howell, M. & Finkelstein, S. (2011), 'Multi-objective Performance Evaluation using Fuzzy Criteria: Increasing Sensitivity Prediction for Outcome of Septic Shock Patients', *IFAC Proceedings Volumes* **44**(1), 14042–14047.

Hossain, M. S., Gresock, J., Edmonds, Y., Helm, R., Potts, M. & Ramakrishnan, N. (2012), 'Connecting the Dots between PubMed Abstracts', *PLOS ONE* **7**(1), e29509.

Hristovski, D., Friedman, C., Rindflesch, T. C. & Peterlin, B. (2006), Exploiting Semantic Relations for Literature-Based Discovery, *in* 'AMIA Annual Symposium', Vol. 2006, American Medical Informatics Association, p. 349.

Hristovski, D., Kastrin, A., Dinevski, D., Burgun, A., Žiberna, L. & Rindflesch, T. C. (2016), 'Using Literature-Based Discovery to Explain Adverse Drug Effects', *Journal of Medical Systems* **40**(8), 185.

Hristovski, D., Kastrin, A., Dinevski, D. & Rindflesch, T. (2015*a*), 'Towards Implementing Semantic Literature-Based Discovery with a Graph Database', *7th International*

Conference on Advances in Databases, Knowledge, and Data Applications (DBKDA) pp. 180–184.

Hristovski, D., Kastrin, A., Dinevski, D. & Rindflesch, T. C. (2015*b*), 'Constructing a Graph Database for Semantic Literature-Based Discovery', *Studies in Health Technology and Informatics* **216**, 1094–1094.

Hristovski, D., Kastrin, A., Peterlin, B. & Rindflesch, T. C. (2010), Combining Semantic Relations and DNA Microarray Data for Novel Hypotheses Generation, *in* 'Linking Literature, Information, and Knowledge for Biology', Springer, pp. 53–61.

Hristovski, D., Kastrin, A. & Rindflesch, T. C. (2015), Semantics-based Cross-domain Collaboration Recommendation in the Life Sciences: Preliminary Results, *in* 'International Conference on Advances in Social Networks Analysis and Mining (ASONAM)', IEEE, pp. 805–806.

Hristovski, D., Kastrin, A. & Rindflesch, T. C. (2016), 'Implementing Semantics-Based Cross-domain Collaboration Recommendation in Biomedicine with a Graph Database', pp. 94–96.

Hristovski, D., Peterlin, B., Mitchell, J. A. & Humphrey (2003), 'Improving Literature-Based Discovery Support by Genetic Knowledge Integration', *Studies in Health Technology and Informatics* **95**, 68–73.

Hristovski, D., Peterlin, B., Mitchell, J. A. & Humphrey, S. M. (2005), 'Using Literature-Based Discovery to Identify Disease Candidate Genes', *International journal of medical informatics* **74**(2-4), 289–298.

Hristovski, D., Stare, J., Peterlin, B. & Dzeroski, S. (2001), 'Supporting Discovery in Medicine by Association Rule Mining in Medline and UMLS', *Studies in Health Technology and Informatics* **84**, 1344–1348.

Hsu, E.-Y., Liu, C.-L. & Tseng, V. S. (2019), Multivariate Time Series Early Classification with Interpretability using Deep Learning and Attention Mechanism, *in* 'Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)', Springer, pp. 541–553.

Hu, X., Li, G., Yoo, I., Zhang, X. & Xu, X. (2005), A Semantic-based Approach for Mining Undiscovered Public Knowledge from Biomedical Literature, *in* 'International Conference on Granular Computing (GrC)', Vol. 1, IEEE, pp. 22–27.

Hu, X., Zhang, X., Yoo, I., Wang, X. & Feng, J. (2010), 'Mining Hidden Connections among Biomedical Concepts from Disjoint Biomedical Literature Sets through Semantic-based Association Rule', *International Journal of Intelligent Systems* **25**(2), 207–223.

Hu, X., Zhang, X., Yoo, I. & Zhang, Y. (2006), A Semantic Approach for Mining Hidden Links from Complementary and Non-interactive Biomedical Literature, *in* 'SIAM International Conference on Data Mining (SDM)', SIAM, pp. 200–209.

Hu, Y., Hines, L. M., Weng, H., Zuo, D., Rivera, M., Richardson, A. & LaBaer, J. (2003), 'Analysis of Genomic and Proteomic Data using Advanced Literature Mining', *Journal of Proteome Research* **2**(4), 405–412.

Huang, S., He, L., Yang, B. & Zhang, M. (2012), A Compound Correlation Model for Disjoint Literature-Based Knowledge Discovery, *in* 'ASLIB Proceedings: New Information Perspectives', Vol. 64, Emerald Group Publishing Limited, pp. 423–436.

Huang, W. & Nakamori, Y. (2004), Fuzzy Predicting New Association Rules from Current Scientific Literature, *in* 'Annual Meeting of the North American Fuzzy Information Processing Society (NAFIPS)', Vol. 1, IEEE, pp. 450–455.

Huang, W., Nakamori, Y., Wang, S. & Ma, T. (2005*a*), Mining Medline for New Possible Relations of Concepts, *in* 'International Conference on Computational and Information Science (ICCS)', Springer, pp. 794–799.

Huang, W., Nakamori, Y., Wang, S. & Ma, T. (2005*b*), 'Mining Scientific Literature to Predict New Relationships', *Intelligent Data Analysis* **9**(2), 219–234.

Huang, Y., Wang, L. & Zan, L.-s. (2016), 'ARN: Analysis and Prediction by Adipogenic Professional Database', *BMC Systems Biology* **10**(1), 57.

Hubel, D. H. & Wiesel, T. N. (1968), 'Receptive Fields and Functional Architecture of Monkey Striate Cortex', *The Journal of Physiology* **195**(1), 215–243.

Hudson, J. (2016), 'An Analysis of the Titles of Papers Submitted to the UK REF in 2014: Authors, Disciplines, and Stylistic Details', *Scientometrics* **109**(2), 871–889.

Hudson, J. (2017), 'Identifying Economics' Place amongst Academic Disciplines: A Science or a Social Science?', *Scientometrics* **113**(2), 735–750.

Hui, W. & Lau, W. K. (2019), Application of Literature-Based Discovery in Nonmedical Disciplines: A Survey, *in* '2$^{\text{nd}}$ International Conference on Computing and Big Data (ICCBD)', ACM, pp. 7–11.

Hur, J., Sullivan, K., Schuyler, A., Hong, Y., Pande, M., States, D., Jagadish, H. & Feldman, E. (2010), 'Literature-Based Discovery of Diabetes-and ROS-related Targets', *BMC Medical Genomics* **3**(1), 49.

Ijaz, A. Z., Song, M. & Lee, D. (2009), MKEM: A Multi-Level Knowledge Emergence Model for Mining Undiscovered Public Knowledge, *in* '3rd International Workshop on Data and Text Mining in Bioinformatics (DTMBIO)', ACM, pp. 51–58.

Iruetaguena, A., Adeva, J. G., Pikatza, J. M., Segundo, U., Buenestado, D. & Barrena, R. (2013), 'Automatic Retrieval of Current Evidence to Support Update of Bibliography in Clinical Guidelines', *Expert Systems with Applications* **40**(6), 2081–2091.

Ittipanuvat, V., Fujita, K., Kajikawa, Y., Mori, J. & Sakata, I. (2012), Finding Linkage between Technology and Social Issues: A Literature-Based Discovery Approach, *in* 'Technology Management for Emerging Technologies (PICMET)', IEEE, pp. 2310–2321.

Ittipanuvat, V., Fujita, K., Sakata, I. & Kajikawa, Y. (2014), 'Finding Linkage between Technology and Social Issue: A Literature-Based Discovery Approach', *Journal of Engineering and Technology Management* **32**, 160–184.

Jacob, R. B., Andersen, T. & McDougal, O. M. (2012), 'Accessible High-throughput Virtual Screening Molecular Docking Software for Students and Educators', *PLOS Computational Biology* **8**(5), e1002499.

Jalender, B., Govardhan, A. & Premchand, P. (2010), 'A Pragmatic Approach to Software Reuse', *Journal of Theoretical & Applied Information Technology* **14**.

Jamali, H. R. & Nikzad, M. (2011), 'Article Title Type and its Relation with the Number of Downloads and Citations', *Scientometrics* **88**(2), 653–661.

Jelier, R., Schuemie, M. J., Veldhoven, A., Dorssers, L. C., Jenster, G. & Kors, J. A. (2008), 'Anni 2.0: A Multipurpose Text-Mining Tool for the Life Sciences', *Genome Biology* **9**(6), R96.

Jha, K. & Jin, W. (2016*a*), Mining Hidden Knowledge from the Counterterrorism Dataset Using Graph-Based Approach, *in* 'International Conference on Applications of Natural Language to Information Systems (NLDB)', Springer, pp. 310–317.

Jha, K. & Jin, W. (2016*b*), Mining Novel Knowledge from Biomedical Literature using Statistical Measures and Domain Knowledge, *in* '7th International Conference on Bioinformatics, Computational Biology, and Health Informatics (BCB)', ACM, pp. 317–326.

Jha, K., Xun, G., Gopalakrishnan, V. & Zhang, A. (2019), 'DWE-Med: Dynamic Word Embeddings for Medical Domain', *ACM Transactions on Knowledge Discovery from Data (TKDD)* **13**(2), 1–21.

Jha, K., Xun, G., Wang, Y., Gopalakrishnan, V. & Zhang, A. (2018), Concepts-bridges: Uncovering Conceptual Bridges based on Biomedical Concept Evolution, *in* '24th International Conference on Knowledge Discovery and Data Mining (SIGKDD)', ACM, pp. 1599–1607.

Jha, K., Xun, G., Wang, Y. & Zhang, A. (2019), Hypothesis Generation from Text Based on Co-Evolution of Biomedical Concepts, *in* '25th International Conference on Knowledge Discovery and Data Mining (SIGKDD)', ACM, pp. 843–851.

Joanes, D. & Gill, C. (1998), 'Comparing Measures of Sample Skewness and Kurtosis', *Journal of the Royal Statistical Society* **47**(1), 183–189.

Juršič, M., Cestnik, B., Urbančič, T. & Lavrač, N. (2012), Cross-domain Literature Mining: Finding Bridging Concepts with CrossBee, *in* '3rd International Conference on Computational Creativity (ICCC)', pp. 33–40.

Juršič, M., Cestnik, B., Urbančič, T. & Lavrač, N. (2013), HCI Empowered Literature Mining for Cross-domain Knowledge Discovery, *in* 'Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data', Springer, pp. 124–135.

Kastrin, A. & Hristovski, D. (2008), A Fast Document Classification Algorithm for Gene Symbol Disambiguation in the BITOLA Literature-Based Discovery Support System, *in* 'AMIA Annual Symposium', Vol. 2008, American Medical Informatics Association, p. 358.

Kastrin, A. & Hristovski, D. (2020), 'Scientometric Analysis and Knowledge Mapping of Literature-Based Discovery (1986-2020)', *arXiv preprint arXiv:2006.08486* .

Kastrin, A., Peterlin, B. & Hristovski, D. (2010), 'Chi-square-based Scoring Function for Categorization of MEDLINE Citations', *Methods of Information in Medicine* **49**(04), 371–378.

Kastrin, A., Rindflesch, T. C. & Hristovski, D. (2014*a*), 'Link Prediction in a MeSH Co-occurrence Network: Preliminary Results', *Studies in Health Technology and Informatics* **205**, 579–583.

Kastrin, A., Rindflesch, T. C. & Hristovski, D. (2014*b*), Link Prediction on the Semantic MEDLINE Network, *in* 'International Conference on Discovery Science (DS)', Springer, pp. 135–143.

Kastrin, A., Rindflesch, T. C. & Hristovski, D. (2016), 'Link Prediction on a Network of Co-occurring MeSH Terms: Towards Literature-Based Discovery', *Methods of Information in Medicine* **55**(04), 340–346.

Kate, R. J. (2016), 'Using Dynamic Time Warping Distances as Features for Improved Time Series Classification', *Data Mining and Knowledge Discovery* **30**(2), 283–312.

Katz, S., Dabrowski, C., Miles, K. & Law, M. (1994), *Glossary of Software Reuse Terms*, National Institute of Standards and Technology Gaithersburg.

Keele, S. (2007), Guidelines for Performing Systematic Literature Reviews in Software Engineering, Technical report, Technical report, Ver. 2.3. EBSE.

Keogh, E. & Ratanamahatana, C. A. (2005), 'Exact Indexing of Dynamic Time Warping', *Knowledge and information systems* **7**(3), 358–386.

Khan, K. S., Kunz, R., Kleijnen, J. & Antes, G. (2003), 'Five Steps to Conducting a Systematic Review', *Journal of the royal society of medicine* **96**(3), 118–121.

Khan, S., Liu, X., Shakil, K. A. & Alam, M. (2017), 'A Survey on Scholarly Data: From Big Data Perspective', *Information Processing & Management* **53**(4), 923–944.

Khoshgoftaar, T. M., Golawala, M. & Van Hulse, J. (2007), An Empirical Study of Learning from Imbalanced Data using Random Forest, *in* '19th International Conference on Tools with Artificial Intelligence (ICTAI)', Vol. 2, IEEE, pp. 310–317.

Kibwami, N. & Tutesigensi, A. (2014), Using the Literature-Based Discovery Research Method in a Context of Built Environment Research, *in* '30th Annual Association of Researchers in Construction Management Conference (ARCOM)', Vol. 1, ARCOM, pp. 227–236.

Kim, D., Seo, D., Cho, S. & Kang, P. (2019), 'Multi-co-training for Document Classification using various Document Representations: TF–IDF, LDA, and Doc2Vec', *Information Sciences* **477**, 15–29.

Kim, H. & Park, S. (2016), Discovering Disease-associated Drugs using Web Crawl Data, *in* '31st Annual ACM Symposium on Applied Computing (SIGAPP)', ACM, pp. 9–14.

Kim, T.-Y. & Cho, S.-B. (2018), Predicting the Household Power Consumption using CNN-LSTM Hybrid Networks, *in* 'International Conference on Intelligent Data Engineering and Automated Learning (IDEAL)', Springer, pp. 481–490.

Kim, T.-Y. & Cho, S.-B. (2019), 'Predicting Residential Energy Consumption using CNN-LSTM Neural Networks', *Energy* **182**, 72–81.

Kim, Y. (2014), Convolutional Neural Networks for Sentence Classification, *in* 'Conference on Empirical Methods in Natural Language Processing (EMNLP)', Association for Computational Linguistics, pp. 1746–1751.

Kim, Y. H., Beak, S. H., Charidimou, A. & Song, M. (2016), 'Discovering New Genes in the Pathways of Common Sporadic Neurodegenerative Diseases: A Bioinformatics Approach', *Journal of Alzheimer's Disease* **51**(1), 293–312.

Kim, Y. H. & Song, M. (2019), 'A Context-based ABC Model for Literature-Based Discovery', *PLOS ONE* **14**(4), e0215313.

Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J. & Linkman, S. (2009), 'Systematic Literature Reviews in Software Engineering – A Systematic Literature Review', *Information and Software Technology* **51**(1), 7–15.

Kobilarov, G., Bizer, C., Auer, S. & Lehmann, J. (2009), DBpedia–A Linked Data Hub and Data Source for Web and Enterprise Applications, *in* '18$^{\text{th}}$ International World Wide Web Conference (WWW)', pp. 1–3.

Korhonen, A., Guo, Y., Baker, S., Yetisgen-Yildiz, M., Stenius, U., Narita, M. & Liò, P. (2014), Improving Literature-Based Discovery with Advanced Text Mining, *in* 'International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics', Springer, pp. 89–98.

Kosnik, L.-R. (2018), 'A Survey of JEL Codes: What do they mean and are they used consistently?', *Journal of Economic Surveys* **32**(1), 249–272.

Kostoff, R. N. (2002), 'Overcoming Specialization', *BioScience* **52**(10), 937–941.

Kostoff, R. N. (2007), 'Validating Discovery in Literature-Based Discovery', *Journal of Biomedical Informatics* **40**(4), 448–450.

Kostoff, R. N. (2008), 'Literature-Related Discovery (LRD): Potential Treatments for Cataracts', *Technological Forecasting and Social Change* **75**(2), 215–225.

Kostoff, R. N. (2011), 'Literature-Related Discovery: Potential Treatments and Preventatives for SARS', *Technological Forecasting and Social Change* **78**(7), 1164–1173.

Kostoff, R. N. (2014), 'Literature-related discovery: common factors for Parkinson's Disease and Crohn's Disease', *Scientometrics* **100**(3), 623–657.

Kostoff, R. N., Block, J. A., Solka, J. L., Briggs, M. B., Rushenberg, R. L., Stump, J. A., Johnson, D., Lyons, T. J. & Wyatt, J. R. (2007), Literature-Related Discovery: A Review, Technical report, Office of Naval Research Arlington VA.

Kostoff, R. N., Block, J. A., Solka, J. L., Briggs, M. B., Rushenberg, R. L., Stump, J. A., Johnson, D., Lyons, T. J. & Wyatt, J. R. (2009), 'Literature-Related Discovery', *Annual Review of Information Science and Technology* **43**(1), 1–71.

Kostoff, R. N., Block, J. A., Stump, J. A. & Johnson, D. (2008), 'Literature-Related Discovery (LRD): Potential Treatments for Raynaud's Phenomenon', *Technological*

Forecasting and Social Change **75**(2), 203–214.

Kostoff, R. N., Block, J. A., Stump, J. A. & Pfeil, K. M. (2004), 'Information Content in Medline Record Fields', *International Journal of Medical Informatics* **73**(6), 515–527.

Kostoff, R. N. & Briggs, M. B. (2008), 'Literature-Related Discovery (LRD): Potential Treatments for Parkinson's Disease', *Technological Forecasting and Social Change* **75**(2), 226–238.

Kostoff, R. N., Briggs, M. B. & Lyons, T. J. (2008), 'Literature-Related Discovery (LRD): Potential Treatments for Multiple Sclerosis', *Technological Forecasting and Social Change* **75**(2), 239–255.

Kostoff, R. N., Briggs, M. B., Solka, J. L. & Rushenberg, R. L. (2008), 'Literature-Related Discovery (LRD): Methodology', *Technological Forecasting and Social Change* **75**(2), 186–202.

Kostoff, R. N. & Lau, C. G. (2013), 'Combined Biological and Health Effects of Electromagnetic Fields and Other Agents in the Published Literature', *Technological Forecasting and Social Change* **80**(7), 1331–1349.

Kostoff, R. N. & Patel, U. (2015), 'Literature-Related Discovery and Innovation: Chronic Kidney Disease', *Technological Forecasting and Social Change* **91**, 341–351.

Kostoff, R. N., Solka, J. L., Rushenberg, R. L. & Wyatt, J. A. (2008), 'Literature-Related Discovery (LRD): Water Purification', *Technological Forecasting and Social Change* **75**(2), 256–275.

Kothari, C. R. & Payne, P. (2015), 'A Metadata based Knowledge Discovery Methodology for Seeding Translational Research', *Studies in Health Technology and Informatics* **216**, 1071–1071.

Kraines, S. B., Guo, W., Hoshiyama, D., Makino, T., Mizutani, H., Okuda, Y., Shidahara, Y. & Takagi, T. (2013), 'Literature-Based Knowledge Discovery from Relationship Associations Based on a DL Ontology Created from MeSH', *Knowledge Discovery, Knowledge Engineering and Knowledge Management* p. 87.

Kraines, S. B., Guo, W., Hoshiyama, D., Mizutani, H. & Takagi, T. (2010), Generating Literature-based Knowledge Discoveries in Life Sciences using Relationship Associations, *in* 'International Conference on Knowledge Discovery and Information Retrieval (KDIR)', pp. 35–44.

Kutuzov, A., Øvrelid, L., Szymanski, T. & Velldal, E. (2018), Diachronic Word Embeddings and Semantic Shifts: A Survey, *in* '27th International Conference on Computational Linguistics (COLING)', Association for Computational Linguistics, pp. 1384–1397.

Kwofie, S. K., Radovanovic, A., Sundararajan, V. S., Maqungo, M., Christoffels, A. & Bajic, V. B. (2011), 'Dragon Exploratory System on Hepatitis C Virus (DESHCV)', *Infection, Genetics and Evolution* **11**(4), 734–739.

Kwon, S., Bae, H., Jo, J. & Yoon, S. (2019), 'Comprehensive Ensemble in QSAR Prediction for Drug Discovery', *BMC Bioinformatics* **20**(1), 521.

Lange, C., Ion, P., Dimou, A., Bratsas, C., Sperber, W., Kohlhase, M. & Antoniou, I. (2012), Bringing Mathematics to the Web of Data: The Case of the Mathematics subject classification, *in* 'Extended Semantic Web Conference (ESWC)', Springer, pp. 763–777.

Längkvist, M., Karlsson, L. & Loutfi, A. (2014), 'A Review of Unsupervised Feature Learning and Deep Learning for Time-series Modeling', *Pattern Recognition Letters* **42**, 11–24.

Lavrač, N., Martinc, M., Pollak, S., Novak, M. P. & Cestnik, B. (2020), 'Bisociative Literature-Based Discovery: Lessons Learned and New Word Embedding Approach', *New Generation Computing* pp. 1–28.

Le Guennec, A., Malinowski, S. & Tavenard, R. (2016), Data Augmentation for Time Series Classification using Convolutional Neural Networks, *in* 'ECML/PKDD Workshop on Advanced Analytics and Learning on Temporal Data'.

Le, Q. & Mikolov, T. (2014), Distributed Representations of Sentences and Documents, *in* '31st International Conference on Machine Learning (ICML)', JMLR, pp. 1188–1196.

Leal, J. P., Rodrigues, V. & Queirós, R. (2012), Computing Semantic Relatedness using DBpedia, *in* '1st Symposium on Languages, Applications and Technologies', Schloss DagstuhlLeibniz-Zentrum fuer Informatik, pp. 133–147.

LeCun, Y., Bengio, Y. & Hinton, G. (2015), 'Deep Learning', *nature* **521**(7553), 436–444.

Lee, D., Kim, W. C., Charidimou, A. & Song, M. (2015), 'A Bird's-eye view of Alzheimer's Disease Research: Reflecting Different Perspectives of Indexers, Authors, or Citers in Mapping the Field', *Journal of Alzheimer's Disease* **45**(4), 1207–1222.

Lee, I., Kim, D., Kang, S. & Lee, S. (2017), Ensemble Deep Learning for Skeleton-based Action Recognition using Temporal Sliding LSTM Networks, *in* 'International Conference on Computer Vision (ICCV)', IEEE, pp. 1012–1020.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., Van Kleef, P., Auer, S. et al. (2015), 'DBpedia–A Large-scale, Multilingual Knowledge Base extracted from Wikipedia', *Semantic web* **6**(2), 167–195.

Lekka, E., Deftereos, S. N., Persidis, A., Persidis, A. & Andronis, C. (2011), 'Literature Analysis for Systematic Drug Repurposing: A Case Study from Biovista', *Drug Discovery Today: Therapeutic Strategies* **8**(3-4), 103–108.

Lemnaru, C. (2012), 'Strategies for Dealing with Real World Classification Problems', *Faculty of Computer Science and Automation, Universitatea Technica, Din Cluj-Napoca* .

Lever, J., Gakkhar, S., Gottlieb, M., Rashnavadi, T., Lin, S., Siu, C., Smith, M., Jones, M. R., Krzywinski, M. & Jones, S. J. (2018), 'A Collaborative Filtering-based Approach to Biomedical Knowledge Discovery', *Bioinformatics* **34**(4), 652–659.

Levy, O. & Goldberg, Y. (2014*a*), Linguistic Regularities in Sparse and Explicit Word Representations, *in* '18th Conference on Computational Natural Language Learning (CoNLL)', pp. 171–180.

Levy, O. & Goldberg, Y. (2014*b*), Neural Word Embedding as Implicit Matrix Factorization, *in* 'Advances in Neural Information Processing Systems (NIPS)', Curran Associates, Inc., pp. 2177–2185.

Levy, O., Goldberg, Y. & Dagan, I. (2015), 'Improving Distributional Similarity with Lessons Learned from Word Embeddings', *Transactions of the Association for Computational Linguistics* **3**, 211–225.

Li, C., Liakata, M. & Rebholz-Schuhmann, D. (2013), 'Biological Network Extraction from Scientific Literature: State of the Art and Challenges', *Briefings in Bioinformatics* **15**(5), 856–877.

Li, J., Sun, Y., Johnson, R., Sciaky, D., Wei, C.-H., Leaman, R., Davis, A. P., Mattingly, C. J., Wiegers, T. C. & Lu, Z. (2015), Annotating Chemicals, Diseases, and Their Interactions in Biomedical Literature, *in* '5th BioCreative Challenge Evaluation Workshop', pp. 173–182.

Li, M.-H. (2020), Using Link Prediction Methods to Examine Networks of Co-occurring

MeSH Terms in Zika and CRISPR Research, *in* 'International Conference on Information (iConference)', Springer, pp. 782–789.

Li, Y., Engelthaler, T., Siew, C. S. & Hills, T. T. (2019), 'The Macroscope: A Tool for Examining the Historical Structure of Language', *Behavior Research Methods* **51**(4), 1864–1877.

Liang, R., Lei, W. & Gang, W. (2013), 'New Insight into Genes in Association with Asthma: Literature-Based Mining and Network Centrality Analysis', *Chinese Medical Journal* **126**(13), 2472–2479.

Lindsay, R. K. & Gordon, M. D. (1999), 'Literature-Based Discovery by Lexical Statistics', *Journal of the American Society for Information Science* **50**(7), 574–587.

Lipscomb, C. E. (2000), 'Medical Subject Headings (MeSH)', *Bulletin of the Medical Library Association* **88**(3), 265.

Liu, C.-L., Hsaio, W.-H. & Tu, Y.-C. (2018), 'Time Series Classification with Multivariate Convolutional Neural Network', *Transactions on Industrial Electronics* **66**(6), 4788–4797.

Liu, H. & Rastegar-Mojarad, M. (2016), 'Literature-Based Knowledge Discovery', *Big Data Analysis for Bioinformatics and Biomedical Discoveries* pp. 233–248.

Liu, S., Zhang, C. & Ma, J. (2017), CNN-LSTM Neural Network Model for Quantitative Strategy Analysis in Stock Markets, *in* 'International Conference on Neural Information Processing (ICONIP)', Springer, pp. 198–206.

Liu, X.-Y. & Zhou, Z.-H. (2006), The Influence of Class Imbalance on Cost-sensitive Learning: An Empirical Study, *in* '6th International Conference on Data Mining (ICDM)', IEEE, pp. 970–974.

Loglisci, C. & Ceci, M. (2011), Discovering Temporal Bisociations for Linking Concepts over Time, *in* 'Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)', Springer, pp. 358–373.

Lu, Y., Shen, D., Pietsch, M., Nagar, C., Fadli, Z., Huang, H., Tu, Y.-C. & Cheng, F. (2015), 'A Novel Algorithm for Analyzing Drug-drug Interactions from MEDLINE Literature', *Scientific Reports* **5**, 17357.

Maciel, W. D., Faria-Campos, A. C., Gonçalves, M. A. & Campos, S. V. (2011), 'Can the Vector Space Model be used to Identify Biological Entity Activities?', *BMC Genomics* **12**, S1.

Maclean, D. & Seltzer, M. (2011), Mining the Web for Medical Hypotheses: A Proof-of-Concept System, *in* 'International Conference on Health Informatics (HEALTHINF)', pp. 303–308.

Maclean, D. & Seltzer, M. I. (2012), Mining the Web for Medical Hypothesis: A Proof-of-Concept System, *in* 'International Conference on Health Informatics (HEALTHINF)'.

Maharjan, S., Montes, M., González, F. A. & Solorio, T. (2018), A Genre-aware Attention Model to Improve the Likability Prediction of Books, *in* 'Conference on Empirical Methods in Natural Language Processing (EMNLP)', pp. 3381–3391.

Malec, S. A., Wei, P., Xu, H., Bernstam, E. V., Myneni, S. & Cohen, T. (2016), Literature-Based Discovery of Confounding in Observational Clinical Data, *in* 'AMIA Annual Symposium', American Medical Informatics Association, pp. 1920–1929.

Mallett, R., Hagen-Zanker, J., Slater, R. & Duvendack, M. (2012), 'The Benefits and Challenges of using Systematic Reviews in International Development Research', *Journal of development effectiveness* **4**(3), 445–455.

Mao, J., Lu, K., Zhao, W. & Cao, Y. (2018), 'How many Keywords do Authors assign to Research Articles – A Multi-disciplinary Analysis?', *iConference 2018 Proceedings* .

Marsi, E., Øzturk, P., Aamot, E., Sizov, G. V. & Ardelan, M. V. (2014), 'Towards Text mining in Climate Science: Extraction of Quantitative Variables and Their Relations'.

Masic, I. & Milinovic, K. (2012), 'On-line Biomedical Databases–the Best Source for Quick Search of the Scientific InformatiOn in the Biomedicine', *Acta Informatica Medica* **20**(2), 72.

Mattingly, C. J. (2009), 'Chemical Databases for Environmental Health and Clinical Research', *Toxicology Letters* **186**(1), 62–65.

Maver, A., Hristovski, D., Rindflesch, T. C. & Peterlin, B. (2013), 'Integration of Data from Omic Studies with the Literature-Based Discovery towards Identification of Novel Treatments for Neovascularization in Diabetic Retinopathy', *BioMed Research International* **2013**.

McClure, M. H. (2012), Preliminary Experiments on Literature-Based Discovery using the Semantic Vectors Package, *in* 'Proceedings on the International Conference on Artificial Intelligence (ICAI)', The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), pp. 1–7.

Mednick, S. (1962), 'The Associative Basis of the Creative Process', *Psychological Review* **69**(3), 220.

Mendes, P. N., Jakob, M. & Bizer, C. (2012), DBpedia: A Multilingual Cross-domain Knowledge Base, *in* 'International Conference on Language Resources and Evaluation (LREC)', pp. 1813–1817.

Méndez, E. & Greenberg, J. (2012), 'Linked Data for Open Vocabularies and HIVE's Global Framework', *El profesional de la información* **21**(3), 236–244.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013), 'Efficient Estimation of Word Representations in Vector Space', *arXiv preprint arXiv:1301.3781* .

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. (2013), Distributed Representations of Words and Phrases and their Compositionality, *in* 'Advances in Neural Information Processing Systems (NIPS)', Curran Associates, Inc., pp. 3111–3119.

Miller, C. M., Rindflesch, T. C., Fiszman, M., Hristovski, D., Shin, D., Rosemblat, G., Zhang, H. & Strohl, K. P. (2012), 'A Closed Literature-Based Discovery Technique finds a Mechanistic Link between Hypogonadism and Diminished Sleep Quality in Aging Men', *Sleep* **35**(2), 279–285.

Milne, D., Medelyan, O. & Witten, I. H. (2006), Mining Domain-Specific Thesauri from Wikipedia: A Case Study, *in* 'IEEE/WIC/ACM International Conference on Web Intelligence (WI)', IEEE, pp. 442–448.

Milojević, S. (2017), 'The Length and Semantic Structure of Article Titles—evolving Disciplinary Practices and correlations with Impact', *Frontiers in Research Metrics and Analytics* **2**, 2.

Mirheidari, B., Blackburn, D., Walker, T., Venneri, A., Reuber, M. & Christensen, H. (2018), Detecting Signs of Dementia using Word Vector Representations, *in* 'Interspeech', pp. 1893–1897.

Mirizzi, R., Ragone, A., Di Noia, T. & Di Sciascio, E. (2010), Semantic Wonder Cloud: Exploratory Search in DBpedia, *in* 'International Conference on Web Engineering (ICWE)', Springer, pp. 138–149.

Mnih, A. & Hinton, G. E. (2009), A Scalable Hierarchical Distributed Language Model, *in* 'Advances in Neural Information Processing Systems (NIPS)', Curran Associates, Inc., pp. 1081–1088.

Moattarian, A. & Alibabaee, A. (2015), 'Syntactic Structures in Research Article Titles from Three Different Disciplines: Applied Linguistics, Civil Engineering, and Dentistry', *Journal of Teaching Language Skills* **34**(1), 27–50.

Moehrle, M. G. (2005), 'What is TRIZ? From Conceptual Basics to a Framework for Research', *Creativity and Innovation Management* **14**(1), 3–13.

Mohammed, M. & Omar, N. (2020), 'Question Classification based on Bloom's Taxonomy Cognitive Domain using Modified TF-IDF and Word2vec', *PLOS ONE* **15**(3), e0230442.

Mola-Velasco, S. M. (2011), Wikipedia Vandalism Detection, *in* '20th International Conference Companion on World Wide Web (WWW)', ACM, pp. 391–396.

Mooney, J. (1995), Portability and Reusability: Common Issues and Differences, *in* '23rd Annual Conference on Computer Science (CSC)', ACM, pp. 150–156.

Mooney, J. D. (1997), Bringing Portability to the Software Process, Technical report, Technical report, Department of Statistics and Computer Science, West Virginia University, Morgantown WV.

Mower, J., Subramanian, D., Shang, N. & Cohen, T. (2016), Classification-by-analogy: Using Vector Representations of Implicit Relationships to Identify Plausibly Causal Drug/side-effect Relationships, *in* 'AMIA Annual Symposium', Vol. 2016, American Medical Informatics Association, p. 1940.

Murena, P.-A., Cornuéjols, A. & Dessalles, J.-L. (2018), Opening the Parallelogram: Considerations on non-Euclidean Analogies, *in* 'International Conference on Case-Based Reasoning (ICCBR)', Springer, pp. 597–611.

Musto, C., Semeraro, G., de Gemmis, M. & Lops, P. (2016), Learning Word Embeddings from Wikipedia for Content-based Recommender Systems, *in* 'European Conference on Information Retrieval (ECIR)', Springer, pp. 729–734.

Nagano, R. L. (2015), 'Research Article Titles and Disciplinary Conventions: A Corpus Study of Eight Disciplines', *Journal of Academic Writing* **5**(1), 133–144.

Nagarajan, M., Wilkins, A. D., Bachman, B. J., Novikov, I. B., Bao, S., Haas, P. J., Terrón-Díaz, M. E., Bhatia, S., Adikesavan, A. K., Labrie, J. J., Regenbogen, S., Buchovecky, C. M., Pickering, C. R., Kato, L., Lisewski, A. M., Lelescu, A., Zhang, H., Boyer, S., Weber, G., Chen, Y., Donehower, L., Spangler, S. & Lichtarge, O. (2015), Predicting Future Scientific Discoveries Based on a Networked Analysis of the Past Literature, *in* '21th International Conference on Knowledge Discovery and Data Mining (SIGKDD)', ACM, pp. 2019–2028.

Naili, M., Chaibi, A. H. & Ghezala, H. H. B. (2017), 'Comparative Study of Word Embedding Methods in Topic Segmentation', *Procedia Computer Science* **112**, 340–349.

Nakamura, H., Ii, S., Chida, H., Friedl, K., Suzuki, S., Mori, J. & Kajikawa, Y. (2014), 'Shedding Light on a Neglected Area: A New Approach to Knowledge Creation', *Sustainability Science* **9**(2), 193–204.

Nakayama, K., Hara, T. & Nishio, S. (2007), Wikipedia Mining for An Association Web Thesaurus Construction, *in* 'International Conference on Web Information Systems Engineering (WISE)', Springer, pp. 322–334.

Ncube, C., Oberndorf, P. & Kark, A. W. (2008), 'Opportunistic Software Systems Development: Making Systems from What's Available', *IEEE Software* **25**(6), 38–41.

Névéol, A., Doğan, R. I. & Lu, Z. (2010), Author Keywords in Biomedical Journal Articles, *in* 'AMIA annual symposium proceedings', Vol. 2010, American Medical Informatics Association, p. 537.

Newman, M. & Gough, D. (2020), Systematic Reviews in Educational Research: Methodology, Perspectives and Application, *in* 'Systematic Reviews in Educational Research', Springer, pp. 3–22.

Nguyen, G., Dlugolinsky, S., Bobák, M., Tran, V., García, Á. L., Heredia, I., Malík, P. & Hluchỳ, L. (2019), 'Machine Learning and Deep Learning Frameworks and Libraries for Large-scale Data Mining: A Survey', *Artificial Intelligence Review* **52**(1), 77–124.

Niu, M., Li, Y., Wang, C. & Han, K. (2018), 'RFAmyloid: A Web Server for Predicting Amyloid Proteins', *International Journal of Molecular Sciences* **19**(7), 2071.

Oermann, M. H. & Murphy, B. (2018), 'Selecting Keywords for your Manuscript', *Nurse Author & Editor* **28**(4), 1–6.

Oldham, S., Fulcher, B., Parkes, L., Arnatkeviciute, A., Suo, C. & Fornito, A. (2019), 'Consistency and Differences between Centrality Measures across Distinct Classes of Networks', *PLOS ONE* **14**(7), e0220061.

Ongsulee, P. (2017), Artificial Intelligence, Machine Learning and Deep Learning, *in* '15th International Conference on ICT and Knowledge Engineering (ICT&KE)', IEEE, pp. 1–6.

Ordóñez, F. J. & Roggen, D. (2016), 'Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition', *Sensors* **16**(1), 115.

Özgür, A., Xiang, Z., Radev, D. R. & He, Y. (2010), 'Literature-Based Discovery of IFN-γ and Vaccine-Mediated Gene Interaction Networks', *BioMed Research International* **2010**.

Özgür, A., Xiang, Z., Radev, D. R. & He, Y. (2011), 'Mining of Vaccine-associated IFN-γ Gene Interaction Networks using the Vaccine Ontology', *Journal of Biomedical Semantics* **2**, S8.

Paiva, C. E., Lima, J. P. d. S. N. & Paiva, B. S. R. (2012), 'Articles with Short Titles describing the Results are Cited more often', *Clinics* **67**(5), 509–513.

Palangi, H., Deng, L., Shen, Y., Gao, J., He, X., Chen, J., Song, X. & Ward, R. (2016), 'Deep Sentence Embedding Using Long Short-Term Memory Networks: Analysis and Application to Information Retrieval', *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **24**(4), 694–707.

Palmer, C. L. & Fenlon, K. (2010), 'Information Research on Interdisciplinarity', *The Oxford handbook of Interdisciplinarity* pp. 174–188.

Park, S., Lee, D.-g. & Shin, H. (2017), 'Network Mirroring for Drug Repositioning', *BMC Medical Informatics and Decision Making* **17**(1), 55.

Peng, Y., Bonifield, G. & Smalheiser, N. R. (2017), 'Gaps within the Biomedical Literature: Initial Characterization and Assessment of Strategies for Discovery', *Frontiers in Research Metrics and Analytics* **2**, 3.

Pérez, J., Arenas, M. & Gutierrez, C. (2009), 'Semantics and Complexity of SPARQL', *ACM Transactions on Database Systems (TODS)* **34**(3), 1–45.

Persidis, A., Deftereos, S. & Persidis, A. (2004), 'Systems Literature Analysis', *Pharmacogenomics* **5**(7), 943–947.

Pervez, A. (2009), 'Information as Form', *tripleC: Communication, Capitalism & Critique. Open Access Journal for a Global Sustainable Information Society* **7**(1), 1–11.

Petrič, I., Cestnik, B., Lavrač, N. & Urbančič, T. (2012), 'Outlier Detection in Cross-context Link Discovery for Creative Literature Mining', *The Computer Journal* **55**(1), 47–61.

Petric, I., Ligeti, B., Gyorffy, B. & Pongor, S. (2014), 'Biomedical Hypothesis Generation by Text Mining and Gene Prioritization', *Protein and Peptide Letters* **21**(8), 847–857.

Petriě, I., Urbaněiě, T., Cestnik, B. & Macedoni-Lukšiě, M. (2009), 'Literature Mining Method RaJoLink for Uncovering Relations between Biomedical Concepts', *Journal of Biomedical Informatics* **42**(2), 219–227.

Pirolli, P. (2007), *Information Foraging Theory: Adaptive Interaction with Information*, Oxford University Press.

Piscopo, A. & Simperl, E. (2019), What we talk about when we talk about Wikidata Quality: A Literature Survey, *in* '15$^{\text{th}}$ International Symposium on Open Collaboration (OpenSym)', pp. 1–11.

Pittaway, L. & Cope, J. (2007), 'Entrepreneurship Education: A Systematic Review of the Evidence', *International small business journal* **25**(5), 479–510.

Pratt, W. & Yetisgen-Yildiz, M. (2003), LitLinker: Capturing Connections across the Biomedical Literature, *in* '2$^{\text{nd}}$ International Conference on Knowledge Capture (K-CAP)', pp. 105–112.

Preiss, J. (2014), Seeking Informativeness in Literature-Based Discovery, *in* 'Workshop on Biomedical Natural Language Processing (BioNLP)', Association for Computational Linguistics, pp. 112–117.

Preiss, J. & Stevenson, M. (2016), 'The Effect of Word Sense Disambiguation Accuracy on Literature-Based Discovery', *BMC Medical Informatics and Decision Making* **16**(1), 57.

Preiss, J. & Stevenson, M. (2017), 'Quantifying and Filtering Knowledge Generated by Literature-Based Discovery', *BMC Bioinformatics* **18**(7), 249.

Preiss, J., Stevenson, M. & Gaizauskas, R. (2015), 'Exploring Relation Types for Literature-Based Discovery', *Journal of the American Medical Informatics Association* **22**(5), 987–992.

Purushotham, S. & Tripathy, B. (2011), Evaluation of Classifier Models using Stratified Tenfold Cross Validation Techniques, *in* 'International Conference on Computing and Communication Systems', Springer, pp. 680–690.

Pusala, M. K., Benton, R. G., Raghavan, V. V. & Gottumukkala, R. N. (2017), Supervised Approach to Rank Predicted Links using Interestingness Measures, *in* 'International Conference on Bioinformatics and Biomedicine (BIBM)', IEEE, pp. 1085–1092.

Pyysalo, S., Baker, S., Ali, I., Haselwimmer, S., Shah, T., Young, A., Guo, Y., Högberg, J., Stenius, U., Narita, M. & Korhonen, A. (2018), 'LION LBD: A Literature-Based Discovery System for Cancer Biology', *Bioinformatics* **35**(9), 1553–1561.

Qi, J. & Ohsawa, Y. (2016), 'Matrix-like Visualization based on Topic Modeling for Discovering Connections between Disjoint Disciplines', *Intelligent Decision Technologies* **10**(3), 273–283.

Qian, Q., Hong, N. & An, X. (2012), 'Structuring the Chinese Disjointed Literature-Based Knowledge Discovery System: The Key Technologies to Success', *Journal of Information Science* **38**(6), 532–539.

Radinsky, K., Agichtein, E., Gabrilovich, E. & Markovitch, S. (2011), A Word at a Time: Computing Word Relatedness Using Temporal Semantic Analysis, *in* '20$^{\text{th}}$ International World Wide Web Conference (WWW)', ACM, pp. 337–346.

Ramadan, N., Halvorson, H., Vande-Linde, A., Levine, S. R., Helpern, J. & Welch, K. (1989), 'Low Brain Magnesium in Migraine', *Headache: The Journal of Head and Face Pain* **29**(7), 416–419.

Rastegar-Mojarad, M., Elayavilli, R. K., Li, D., Prasad, R. & Liu, H. (2015), A New Method for Prioritizing Drug Repositioning Candidates Extracted by Literature-Based Discovery, *in* 'International Conference on Bioinformatics and Biomedicine (BIBM)', IEEE, pp. 669–674.

Rastegar-Mojarad, M., Elayavilli, R. K., Wang, L., Prasad, R. & Liu, H. (2016), Prioritizing Adverse Drug Reaction and Drug Repositioning Candidates Generated by Literature-Based Discovery, *in* '7$^{\text{th}}$ ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (BCB)', ACM, pp. 289–296.

Rastegar-Mojarad, M. & Prasad, R. (2015), Toward a Complete Database of Drug Repurposing Candidates Extracted from Social Media, Biomedical Literature, and Genetic Data, *in* 'International Conference on Healthcare Informatics (ICHI)', IEEE, pp. 494–494.

Rindflesch, T. C. & Fiszman, M. (2003), 'The Interaction of Domain Knowledge and Linguistic Structure in Natural Language Processing: Interpreting Hypernymic Propositions in Biomedical Text', *Journal of Biomedical Informatics* **36**(6), 462–477.

Robinson, S., Nance, R. E., Paul, R. J., Pidd, M. & Taylor, S. J. (2004), 'Simulation Model Reuse: Definitions, Benefits and Obstacles', *Simulation Modelling Practice and Theory* **12**(7-8), 479–494.

Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M. & Rigol-Sanchez, J. P. (2012), 'An Assessment of the Effectiveness of a Random Forest Classifier for Land-cover Classification', *ISPRS Journal of Photogrammetry and Remote Sensing* **67**, 93–104.

Roy, S., Yun, D., Madahian, B., Berry, M. W., Deng, L.-Y., Goldowitz, D. & Homayouni, R. (2017), 'Navigating the Functional Landscape of Transcription Factors via

Non-negative Tensor Factorization Analysis of Medline Abstracts', *Frontiers in Bioengineering and Biotechnology* **5**, 48.

Ruder, S., Vulić, I. & Søgaard, A. (2019), 'A Survey of Cross-lingual Word Embedding Models', *Journal of Artificial Intelligence Research* **65**, 569–631.

Ruthven, I. & Kelly, D. (2011), *Interactive Information Seeking, Behaviour and Retrieval*, Facet publishing.

Ryan, P. B., Schuemie, M. J., Welebob, E., Duke, J., Valentine, S. & Hartzema, A. G. (2013), 'Defining a Reference Set to Support Methodological Research in Drug Safety', *Drug Safety* **36**(1), 33–47.

Rzhetsky, A., Foster, J. G., Foster, I. T. & Evans, J. A. (2015), 'Choosing Experiments to Accelerate Collective Discovery', *National Academy of Sciences* **112**(47), 14569–14574.

Salatino, A. A., Thanapalasingam, T., Mannocci, A., Birukou, A., Osborne, F. & Motta, E. (2020), 'The Computer Science Ontology: A Comprehensive Automatically-generated Taxonomy of Research Areas', *Data Intelligence* **2**(3), 379–416.

Sang, S., Yang, Z., Li, Z. & Lin, H. (2015), 'Supervised Learning based Hypothesis Generation from Biomedical Literature', *BioMed Research International* **2015**.

Sang, S., Yang, Z., Liu, X., Wang, L., Lin, H., Wang, J. & Dumontier, M. (2018), 'GrEDeL: A Knowledge Graph Embedding based Method for Drug Discovery from Biomedical Literatures', *IEEE Access* **7**, 8404–8415.

Sang, S., Yang, Z., Wang, L., Liu, X., Lin, H. & Wang, J. (2018), 'SemaTyP: A Knowledge Graph based Literature Mining Method for Drug Discovery', *BMC Bioinformatics* **19**(1), 193.

Santos, B. S. (2008), Evaluating Visualization Techniques and Tools: What are the Main Issues, *in* 'AVI Workshop on Beyond Time and Errors: Novel Evaluation Methods For information Visualization (BELIV)'.

Schroeder, J., Xu, J., Chen, H. & Chau, M. (2007), 'Automated Criminal Link Analysis based on Domain Knowledge', *Journal of the American Society for Information Science and Technology* **58**(6), 842–855.

Sebastian, Y., Siew, E.-G. & Orimaye, S. O. (2015), Predicting Future Links between Disjoint Research Areas using Heterogeneous Bibliographic Information Network, *in* 'Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)', Springer, pp. 610–621.

Sebastian, Y., Siew, E.-G. & Orimaye, S. O. (2017*a*), 'Emerging Approaches in Literature-Based Discovery: Techniques and Performance Review', *The Knowledge Engineering Review* **32**.

Sebastian, Y., Siew, E.-G. & Orimaye, S. O. (2017*b*), 'Learning the Heterogeneous Bibliographic Information Network for Literature-Based Discovery', *Knowledge-Based Systems* **115**, 66–79.

Seki, K. & Mostafa, J. (2007), Literature-Based Discovery by an Enhanced Information Retrieval Model, *in* 'International Conference on Discovery Science (DS)', Springer, pp. 185–196.

Seki, K. & Mostafa, J. (2009), 'Discovering Implicit Associations among Critical Biological Entities', *International Journal of Data Mining and Bioinformatics* **3**(2), 105–123.

Selvin, S., Vinayakumar, R., Gopalakrishnan, E., Menon, V. K. & Soman, K. (2017), Stock Price Prediction using LSTM, RNN and CNN-sliding Window Model, *in* 'International Conference on Advances in Computing, Communications and Informatics (ICACCI)', IEEE, pp. 1643–1647.

Settles, B., Craven, M. & Ray, S. (2008), Multiple-Instance Active Learning, *in* 'Advances in Neural Information Processing Systems (NIPS)', pp. 1289–1296.

Shams, R. (2014), Identification of Informativeness in Text using Natural Language Stylometry, PhD thesis, The University of Western Ontario.

Shang, N., Xu, H., Rindflesch, T. C. & Cohen, T. (2014), 'Identifying Plausible Adverse Drug Reactions using Knowledge Extracted from the Literature', *Journal of Biomedical Informatics* **52**, 293–310.

Sharma, D. & Surolia, A. (2013), *Degree Centrality*, Springer, pp. 558–558.

Shibata, N., Kajikawa, Y., Takeda, Y. & Matsushima, K. (2009), 'Comparative Study on Methods of Detecting Research Fronts using Different Types of Citation', *Journal of the American Society for information Science and Technology* **60**(3), 571–580.

Shin, H.-C., Orton, M., Collins, D. J., Doran, S. & Leach, M. (2016), Organ Detection using Deep Learning, *in* 'Medical Image Recognition, Segmentation and Parsing', Elsevier, pp. 123–153.

Shoemark, P., Liza, F. F., Nguyen, D., Hale, S. & McGillivray, B. (2019), Room to Glo: A Systematic Comparison of Semantic Change Detection Approaches with Word Embeddings, *in* 'Conference on Empirical Methods in Natural Language Processing and the 9^{th} International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)', pp. 66–76.

Singh, S., Singh, S. & Singh, G. (2010), 'Reusability of the Software', *International Journal of Computer Applications* **7**(14), 38–41.

Skeels, M. M., Henning, K., Yildiz, M. Y. & Pratt, W. (2005), Interaction Design for Literature-Based Discovery, *in* 'Conference on Human Factors in Computing Systems (CHI)', pp. 1785–1788.

Smalheiser, N. R. (2005), The Arrowsmith Project: 2005 Status Report, *in* 'International Conference on Discovery Science (DS)', Springer, pp. 26–43.

Smalheiser, N. R. (2012), 'Literature-Based Discovery: Beyond the ABCs', *Journal of the American Society for Information Science and Technology* **63**(2), 218–224.

Smalheiser, N. R. (2017), 'Rediscovering Don Swanson: The Past, Present and Future of Literature-Based Discovery', *Journal of Data and Information Science* **2**(4), 43–64.

Smalheiser, N. R. & Gomes, O. L. (2015), 'Mammalian Argonaute-DNA Binding?', *Biology Direct* **10**(1), 27.

Smalheiser, N. R., Shao, W. & Philip, S. Y. (2015), 'Nuggets: Findings Shared in Multiple Clinical Case Reports', *Journal of the Medical Library Association (JMLA)* **103**(4), 171.

Smalheiser, N. R. & Swanson, D. R. (1996), 'Indomethacin and Alzheimer's Disease', *Neurology* **46**(2), 583–583.

Smalheiser, N. R. & Swanson, D. R. (1998), 'Calcium-independent Phospholipase A2 and Schizophrenia', *Archives of General Psychiatry* **55**(8), 752–753.

Smalheiser, N. R. & Torvik, V. I. (2008), The Place of Literature-Based Discovery in Contemporary Scientific Practice, *in* 'Literature-Based Discovery', Springer, pp. 13–22.

Smalheiser, N. R., Torvik, V. I., Bischoff-Grethe, A., Burhans, L. B., Gabriel, M., Homayouni, R., Kashef, A., Martone, M. E., Perkins, G. A., Price, D. L., Talk, A. C. & West, R. (2006), 'Collaborative Development of the Arrowsmith Two Node Search Interface Designed for Laboratory Investigators', *Journal of Biomedical Discovery and Collaboration* **1**(1), 8.

Smalheiser, N. R., Torvik, V. I. & Zhou, W. (2009), 'Arrowsmith Two-node Search Interface: A Tutorial on Finding Meaningful Links between Two Disparate Sets of

Articles in MEDLINE', *Computer methods and programs in biomedicine* **94**(2), 190–197.

Smalheiser, N. R., Zhou, W. & Torvik, V. I. (2008), 'Anne O'Tate: A Tool to Support User-driven Summarization, Drill-down and Browsing of PubMed Search Results', *Journal of Biomedical Discovery and Collaboration* **3**(1), 2.

Smalheiser, N. R., Zhou, W. & Torvik, V. I. (2011), 'Distribution of "Characteristic" Terms in MEDLINE Literatures', *Information* **2**(2), 266–276.

Smith, A. (2019), 'From PACS to PhySH', *Nature Reviews Physics* **1**(1), 8–11.

Smith, A. (2020), 'Physics Subject Headings (PhySH)', *Knowledge Organization* **47**(3), 257–266.

Snyder, H. (2019), 'Literature Review as a Research Methodology: An Overview and Guidelines', *Journal of Business Research* **104**, 333–339.

Song, M., Heo, G. E. & Ding, Y. (2015), 'SemPathFinder: Semantic Path Analysis for Discovering Publicly Unknown Knowledge', *Journal of Informetrics* **9**(4), 686–703.

Spangler, S., Wilkins, A. D., Bachman, B. J., Nagarajan, M., Dayaram, T., Haas, P., Regenbogen, S., Pickering, C. R., Comer, A., Myers, J. N., Stanoi, I., Kato, L., Lelescu, A., Labrie, J. J., Parikh, N., Lisewski, A. M., Donehower, L., Chen, Y. & Lichtarge, O. (2014), Automated Hypothesis Generation Based on Mining Scientific Literature, *in* '20th International Conference on Knowledge Discovery and Data Mining (SIGKDD)', ACM, pp. 1877–1886.

Spinak, E. L., Launy, S. & Grillo, B. (1999), Detection of Unknown Public Knowledge through the Semantic Mapping by Disciplines in Large Databases, *in* 'International Society for Scientometrics and Informetrics (ISSI)', pp. 457–467.

Srikant, R. & Agrawal, R. (1995), Mining Generalized Association Rules, *in* '21st International Conference on Very Large Data Bases (VLDB)'.

Srinivas, A. & Velusamy, R. L. (2015), Identification of Influential Nodes from Social Networks based on Enhanced Degree Centrality Measure, *in* 'International Advance Computing Conference (IACC)', IEEE, pp. 1179–1184.

Srinivasan, M., Blackburn, C., Mohamed, M., Sivagami, A. & Blum, J. (2015), 'Literature-Based Discovery of Salivary Biomarkers for Type 2 Diabetes Mellitus', *Biomarker Insights* .

Srinivasan, P. (2004), 'Text Mining: Generating Hypotheses from MEDLINE', *Journal of the American Society for Information Science and Technology* **55**(5), 396–413.

Srinivasan, P. & Libbus, B. (2004), 'Mining MEDLINE for Implicit Links between Dietary Substances and Diseases', *Bioinformatics* **20**(suppl_1), i290–i296.

Stankovic, M., Breitfuss, W. & Laublet, P. (2011), Discovering Relevant Topics using DBpedia: Providing Non-obvious Recommendations, *in* 'IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)', IEEE, pp. 219–222.

Stegmann, J. & Grohmann, G. (2003), 'Hypothesis Generation Guided by Co-word Clustering', *Scientometrics* **56**(1), 111–135.

Steimann, F. (1997), 'Fuzzy Set Theory in Medicine', *Artificial Intelligence in Medicine* **11**(1), 1–7.

Stephanidis, C. (2019), 'New Perspectives into Human–Computer Interaction', *User Interfaces for All-Concepts, Methods and Tools* pp. 3–20.

Stephens, D. & Krebs, J. (1986), *Foraging Theory*, Princeton University Press.

Su, J. & Zhou, C. (2009), Literature-based Multidiscipline Knowledge Discovery: A New Application of Bibliometrics, *in* '12$^{\text{th}}$ International Society for Scientometrics and Informetrics (ISSI)', pp. 165–172.

Subotic, S. & Mukherjee, B. (2014), 'Short and Amusing: The Relationship between Title Characteristics, Downloads, and Citations in Psychology Articles', *Journal of Information Science* **40**(1), 115–124.

Swanson, D. (2008), Literature-Based Discovery? The Very Idea, *in* 'Literature-based discovery', Springer, pp. 3–11.

Swanson, D. R. (1986), 'Fish oil, Raynaud's syndrome, and undiscovered public knowledge', *Perspectives in Biology and Medicine* **30**(1), 7–18.

Swanson, D. R. (1988), 'Migraine and Magnesium: Eleven Neglected Connections', *Perspectives in Biology and Medicine* **31**(4), 526–557.

Swanson, D. R. (1990*a*), 'Medical Literature as a Potential Source of New Knowledge', *Bulletin of the Medical Library Association* **78**(1), 29.

Swanson, D. R. (1990*b*), 'Somatomedin C and Arginine: Implicit Connections between Mutually Isolated Literatures', *Perspectives in Biology and Medicine* **33**(2), 157–186.

Swanson, D. R. (2001), 'ASIST Award of Merit Acceptance Speech: On the Fragmentation of Knowledge, the Connection Explosion, and Assembling Other People's Ideas', *Bulletin of the American Society for Information Science and Technology* **27**(3), 12–14.

Swanson, D. R. (2011), 'Literature-based Resurrection of Neglected Medical Discoveries', *Journal of Biomedical Discovery and Collaboration* **6**, 34.

Swanson, D. R. & Smalheiser, N. R. (1996), Undiscovered Public Knowledge: A Ten-Year Update, *in* '2nd International Conference on Knowledge Discovery and Data Mining (KDD)', AAAI Press, pp. 295–298.

Swanson, D. R. & Smalheiser, N. R. (1997), 'An Interactive System for Finding Complementary Literatures: A Stimulus to Scientific Discovery', *Artificial Intelligence* **91**(2), 183–203.

Swanson, D. R. & Smalheiser, N. R. (1999), 'Implicit Text Linkages between Medline Records: Using Arrowsmith as an aid to Scientific Discovery', *Library Trends* **48**(1), 48–59.

Swanson, D. R., Smalheiser, N. R. & Bookstein, A. (2001), 'Information Discovery from Complementary Literatures: Categorizing Viruses as Potential Weapons', *Journal of the American Society for Information Science and Technology* **52**(10), 797–812.

Swanson, D. R., Smalheiser, N. R. & Torvik, V. I. (2006), 'Ranking Indirect Connections in Literature-Based Discovery: The Role of Medical Subject Headings', *Journal of the American Society for Information Science and Technology* **57**(11), 1427–1439.

Symonds, M., Bruza, P. & Sitbon, L. (2014), The Efficiency of Corpus-based Distributional Models for Literature-Based Discovery on Large Data Sets, *in* '2nd Australasian Web Conference', Australian Computer Society, Inc., pp. 49–57.

Symonds, M., Bruza, P., Zuccon, G., Koopman, B., Sitbon, L. & Turner, I. (2014), 'Automatic Query Expansion: A Structural Linguistic Perspective', *Journal of the Association for Information Science and Technology* **65**(8), 1577–1596.

Tague-Sutcliffe, J. (1992), Measuring the Informativeness of a Retrieval Process, *in* '15th International Conference on Research Development in Information Retrieval (SIGIR)', pp. 23–36.

Tan, P.-N. (2009), *Receiver Operating Characteristic*, Springer, pp. 2349–2352.

Tang, J., Wu, S., Sun, J. & Su, H. (2012), Cross-domain Collaboration Recommendation, *in* '18th International Conference on Knowledge Discovery and Data Mining (SIGKDD)', pp. 1285–1293.

Taye, M. M. (2010), 'Understanding Semantic Web and Ontologies: Theory and Applications', *arXiv preprint arXiv:1006.4567* .

Tenopir, C. & Jasco, P. (1993), 'Quality of Abstracts', *Online Vol. 17* .

Thaicharoen, S., Altman, T., Gardiner, K. & Cios, K. J. (2009), Discovering Relational Knowledge from Two Disjoint Sets of Literatures using Inductive Logic Programming, *in* 'Symposium on Computational Intelligence and Data Mining (CIDM)', IEEE, pp. 283–290.

Thilakaratne, M., Falkner, K. & Atapattu, T. (2019*a*), 'A Systematic Review on Literature-Based Discovery Workflow', *PeerJ Computer Science* **5**, e235.

Thilakaratne, M., Falkner, K. & Atapattu, T. (2019*b*), 'A systematic review on literature-based discovery: General overview, methodology, & statistical analysis', *ACM Computing Surveys* **52**(6).

Tijskens, A., Janssen, H. & Roels, S. (2019), 'Optimising Convolutional Neural Networks to Predict the Hygrothermal Performance of Building Components', *Energies* **12**(20), 3966.

Titze, G., Bryl, V., Zirn, C. & Ponzetto, S. P. (2014), DBpedia Domains: Augmenting DBpedia with Domain Information, *in* '(9th International Conference on Language Resources and Evaluation (LREC)', ELRA, pp. 1438–1442.

Torvik, V. I. & Smalheiser, N. R. (2007), 'A Quantitative Model for Linking Two Disparate Sets of Articles in MEDLINE', *Bioinformatics* **23**(13), 1658–1665.

Tsafnat, G., Jasch, D., Misra, A., Choong, M. K., Lin, F. P.-Y. & Coiera, E. (2014), 'Gene–disease Association with Literature-Based Enrichment', *Journal of Biomedical Informatics* **49**, 221–226.

Turtle, H. & Croft, W. B. (1991), 'Evaluation of An Inference Network-based Retrieval Model', *ACM Transactions on Information Systems (TOIS)* **9**(3), 187–222.

Tyler, S. K. & Zhang, Y. (2008), Open Domain Recommendation: Social Networks and Collaborative Filtering, *in* 'International Conference on Advanced Data Mining and Applications (ADMA)', Springer, pp. 330–341.

Urbančič, T., Petrič, I., Cestnik, B. & Macedoni-Lukšič, M. (2007), Literature Mining: Towards Better Understanding of Autism, *in* 'Artificial Intelligence in Medicine in Europe (AIME)', Springer, pp. 217–226.

Valente, T. W., Coronges, K., Lakon, C. & Costenbader, E. (2008), 'How Correlated are Network Centrality Measures?', *Connect* **28**(1), 16.

Van der Eijk, C. C., van Mulligen, E. M., Kors, J. A., Mons, B. & van den Berg, J. (2004), 'Constructing an Associative Concept Space for Literature-Based Discovery', *Journal of the American society for Information Science and Technology* **55**(5), 436–444.

Van der Eijk, C., Van Mulligen, E. & Van den Berg, J. (2002), Finding Complementary Scientific Concepts using a Conceptual Associative Spatial Graph, *in* ‘World Multi-Conference On Systemics, Cybernetics and Informatics (ISAS-SCI)’, pp. 46–50.

Vicente-Gomila, J. M. (2014), ‘The Contribution of Syntactic–semantic Approach to the Search for Complementary Literatures for Scientific or Technical Discovery’, *Scientometrics* **100**(3), 659–673.

Vidal, M.-E., Raschid, L., Márquez, N., Rivera, J. C. & Ruckhaus, E. (2010), BioNav: An Ontology-Based Framework to Discover Semantic Links in the Cloud of Linked Data, *in* ‘Extended Semantic Web Conference (ESWC)’, Springer, pp. 441–445.

Vidal, M.-E., Rivera, J.-C., Ibáñez, L.-D., Raschid, L., Palma, G., Rodriguez, H. & Ruckhaus, E. (2014), ‘An Authority-flow based Ranking Approach to Discover Potential Novel Associations between Linked Data’, *Semantic Web* **5**(1), 23–46.

Vlietstra, W. J., Zielman, R., van Dongen, R. M., Schultes, E. A., Wiesman, F., Vos, R., van Mulligen, E. M. & Kors, J. A. (2017), ‘Automated Extraction of Potential Migraine Biomarkers using a Semantic Graph’, *Journal of Biomedical Informatics* **71**, 178–189.

Voorhees, E. M. (2006), The TREC 2005 Robust Track, *in* ‘ACM SIGIR Forum’, Vol. 40, ACM, pp. 41–48.

Vos, R., Aarts, S., van Mulligen, E., Metsemakers, J., van Boxtel, M. P., Verhey, F. & van den Akker, M. (2013), ‘Finding Potentially New Multimorbidity Patterns of Psychiatric and Somatic Diseases: Exploring the use of Literature-Based Discovery in Primary Care Research’, *Journal of the American Medical Informatics Association* **21**(1), 139–145.

Vrandečić, D. & Krötzsch, M. (2014), ‘Wikidata: A Free Collaborative Knowledgebase’, *Communications of the ACM* **57**(10), 78–85.

Wang, P., Hao, T., Yan, J. & Jin, L. (2017), ‘Large-scale Extraction of Drug-disease Pairs from the Medical Literature’, *Journal of the Association for Information Science and Technology* **68**(11), 2649–2661.

Wang, P., Hu, J., Zeng, H. & Chen, Z. (2009), ‘Using Wikipedia Knowledge to Improve Text Classification’, *Knowledge and Information Systems* **19**(3), 265–281.

Wang, Q., Huang, L., Jiang, Z., Knight, K., Ji, H., Bansal, M. & Luan, Y.

(2019), 'Paperrobot: Incremental Draft Generation of Scientific Ideas', *arXiv preprint arXiv:1905.07870* .

Wang, Z., Hong, X., Wang, H., Liu, J. & Liu, J.-P. (2020), 'COVID-19: From Structure to Therapeutic Targeting in Studying Approved Drugs and Local DNA Vaccination'.

Weeber, M. (2007), Drug Discovery as an Example of Literature-Based Discovery, *in* 'Computational Discovery of Scientific Knowledge', Springer, pp. 290–306.

Weeber, M., Klein, H., Aronson, A. R., Mork, J. G., De Jong-van Den Berg, L. & Vos, R. (2000), Text-based Discovery in Biomedicine: The Architecture of the DAD-system, *in* 'AMIA Annual Symposium', American Medical Informatics Association, p. 903.

Weeber, M., Klein, H., De Jong-Van Den Berg, L. T. & Vos, R. (2001), 'Using concepts in literature-based discovery: Simulating Swanson's Raynaud-fish oil and migraine-magnesium discoveries', *Journal of the American Society for Information Science and Technology* **52**(7), 548–557.

Weeber, M., Kors, J. A. & Mons, B. (2005), 'Online Tools to Support Literature-Based Discovery in the Life Sciences', *Briefings in Bioinformatics* **6**(3), 277–286.

Wei, C.-H., Peng, Y., Leaman, R., Davis, A. P., Mattingly, C. J., Li, J., Wiegers, T. C. & Lu, Z. (2015), Overview of the BioCreative V Chemical Disease Relation (CDR) Task, *in* '5th BioCreative Challenge Evaluation Workshop', pp. 154–166.

Wei, L. & Zou, Q. (2016), 'Recent Progress in Machine Learning-based Methods for Protein Fold Recognition', *International journal of molecular sciences* **17**(12), 2118.

Weidt, F. & Silva, R. (2016), Systematic Literature Review in Computer Science-A Practical Guide, Technical report, Technical report, Universidade Federal de Juiz de Fora.

Wiechula, R., Conroy, T., Kitson, A. L., Marshall, R. J., Whitaker, N. & Rasmussen, P. (2016), 'Umbrella Review of the Evidence: What Factors Influence the Caring Relationship Between a Nurse and Patient?', *Journal of Advanced Nursing* **72**(4), 723–734.

Wilkowski, B., Fiszman, M., Miller, C., Hristovski, D., Arabandi, S., Rosemblat, G. & Rindflesch, T. (2011*a*), Discovery Browsing with Semantic Predications and Graph Theory, *in* 'AMIA Annual Symposium'.

Wilkowski, B., Fiszman, M., Miller, C. M., Hristovski, D., Arabandi, S., Rosemblat, G. & Rindflesch, T. C. (2011*b*), Graph-based Methods for Discovery Browsing with Semantic Predications, *in* 'AMIA Annual Symposium', Vol. 2011, American Medical

Informatics Association, p. 1514.

Workman, T. E., Fiszman, M., Cairelli, M. J., Nahl, D. & Rindflesch, T. C. (2016), 'Spark, An Application based on Serendipitous Knowledge Discovery', *Journal of Biomedical Informatics* **60**, 23–37.

Workman, T. E., Fiszman, M., Rindflesch, T. C. & Nahl, D. (2014), 'Framing Serendipitous Information-seeking Behavior for Facilitating Literature-Based Discovery: A Proposed Model', *Journal of the Association for Information Science and Technology* **65**(3), 501–512.

Wren, J. D. (2004), 'Extending the Mutual Information Measure to Rank Inferred Literature Relationships', *BMC Bioinformatics* **5**(1), 145.

Wren, J. D., Bekeredjian, R., Stewart, J. A., Shohet, R. V. & Garner, H. R. (2004), 'Knowledge Discovery by Automated Identification and Ranking of Implicit Relationships', *Bioinformatics* **20**(3), 389–398.

Xuan, S., Liu, G., Li, Z., Zheng, L., Wang, S. & Jiang, C. (2018), Random Forest for Credit Card Fraud Detection, *in* '15$^{\text{th}}$ International Conference on Networking, Sensing and Control (ICNSC)', IEEE, pp. 1–6.

Xun, G., Jha, K., Gopalakrishnan, V., Li, Y. & Zhang, A. (2017), Generating Medical Hypotheses based on Evolutionary Medical Concepts, *in* 'International Conference on Data Mining (ICDM)', IEEE, pp. 535–544.

Yamashita, R., Nishio, M., Do, R. K. G. & Togashi, K. (2018), 'Convolutional Neural Networks: An Overview and Application in Radiology', *Insights into imaging* **9**(4), 611–629.

Yang, C.-L., Chen, Z.-X. & Yang, C.-Y. (2020), 'Sensor Classification Using Convolutional Neural Network by Encoding Multivariate Time Series as Two-Dimensional Colored Images', *Sensors* **20**(1), 168.

Yang, H.-T., Ju, J.-H., Wong, Y.-T., Shmulevich, I. & Chiang, J.-H. (2017), 'Literature-Based Discovery of New Candidates for Drug Repurposing', *Briefings in Bioinformatics* **18**(3), 488–497.

Yang, K., Zhao, X., Waxman, D. & Zhao, X.-M. (2019), 'Predicting Drug-disease Associations with Heterogeneous Network Embedding', *Chaos: An Interdisciplinary Journal of Nonlinear Science* **29**(12), 123109.

Yang, Q., Scholz, M., Shao, J., Wang, G. & Liu, X. (2019), 'A generic framework to analyse the spatiotemporal variations of water quality data on a catchment scale', *Environmental Modelling & Software* **122**, 104071.

Yao, L., Yang, Z., Zhifang, S. & Zhenguo, W. (2008), Research on Non-interactive Literature-Based Knowledge Discovery, *in* 'International Conference on Computer Science and Software Engineering (CSSE)', IEEE, pp. 747–752.

Ye, C., Leng, F. & Guo, X. (2010), Clustering Algorithm in Literature-Based Discovery, *in* '7th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)', Vol. 4, IEEE, pp. 1625–1629.

Yetisgen-Yildiz, M. (2006), Litlinker: A System for Searching Potential Discoveries in Biomedical Literature, *in* '29th International Conference on Research Development in Information Retrieval (SIGIR)', ACM, pp. 6–11.

Yetisgen-Yildiz, M. & Pratt, W. (2006), 'Using Statistical and Knowledge-based Approaches for Literature-Based Discovery', *Journal of Biomedical informatics* **39**(6), 600–611.

Yetisgen-Yildiz, M. & Pratt, W. (2009), 'A New Evaluation Methodology for Literature-Based Discovery Systems', *Journal of Biomedical Informatics* **42**(4), 633–643.

Yu, W., Sun, X., Yang, K., Rui, Y. & Yao, H. (2018), 'Hierarchical Semantic Image Matching using CNN Feature Pyramid', *Computer Vision and Image Understanding* **169**, 40–51.

Zadrozny, B., Langford, J. & Abe, N. (2003), Cost-sensitive Learning by Cost-proportionate Example Weighting, *in* '3rd International Conference on Data Mining (ICDM)', IEEE, pp. 435–442.

Zhan, Y., Zhou, S., Li, Y., Mu, S., Zhang, R., Song, X., Lin, F., Zhang, R. & Zhang, B. (2017), 'Using the BITOLA System to Identify Candidate Molecules in the Interaction between Oral Lichen Planus and Depression', *Behavioural Brain Research* **320**, 136–142.

Zhang, E. & Zhang, Y. (2009*a*), *Average Precision*, Springer, pp. 192–193.

Zhang, E. & Zhang, Y. (2009*b*), *F-Measure*, Springer, pp. 1147–1147.

Zhang, E. & Zhang, Y. (2009*c*), *Precision*, Springer, pp. 2126–2126.

Zhang, E. & Zhang, Y. (2009*d*), *Recall*, Springer, pp. 2348–2348.

Zhang, R., Cairelli, M. J., Fiszman, M., Kilicoglu, H., Rindflesch, T. C., Pakhomov, S. V. & Melton, G. B. (2014), 'Exploiting Literature-derived Knowledge and Semantics to Identify Potential Prostate Cancer Drugs', *Cancer Informatics* **13**, CIN–S13889.

Zhang, W., Yue, X., Lin, W., Wu, W., Liu, R., Huang, F. & Liu, F. (2018), ‘Predicting Drug-disease Associations by using Similarity Constrained Matrix Factorization’, *BMC Bioinformatics* **19**(1), 1–12.

Zhang, Y.-D., Yang, Z.-J., Lu, H.-M., Zhou, X.-X., Phillips, P., Liu, Q.-M. & Wang, S.-H. (2016), ‘Facial Emotion Recognition based on Biorthogonal Wavelet Entropy, Fuzzy Support Vector Machine, and Stratified Cross Validation’, *IEEE Access* **4**, 8375–8385.

Zhao, B., Lu, H., Chen, S., Liu, J. & Wu, D. (2017), ‘Convolutional Neural Networks for Time Series Classification’, *Journal of Systems Engineering and Electronics* **28**(1), 162–169.

Zheng, Y., Liu, Q., Chen, E., Ge, Y. & Zhao, J. L. (2014), Time Series Classification using Multi-channels Deep Convolutional Neural Networks, *in* ‘International Conference on Web-Age Information Management (WAIM)’, Springer, pp. 298–310.

Zhou, B., Ma, X., Luo, Y. & Yang, D. (2019), ‘Wind Power Prediction based on LSTM Networks and Nonparametric Kernel Density Estimation’, *IEEE Access* **7**, 165279–165292.

Zhou, E., Hui, N., Shu, M., Wu, B. & Zhou, J. (2015), ‘Systematic Analysis of the p53-related MicroRNAs in Breast Cancer Revealing their Essential Roles in the Cell Cycle’, *Oncology Letters* **10**(6), 3488–3494.

Zhou, Y., Tong, Y., Gu, R. & Gall, H. (2016), ‘Combining Text Mining and Data Mining for Bug Report Classification’, *Journal of Software: Evolution and Process* **28**(3), 150–176.

Zou, W. Y., Socher, R., Cer, D. & Manning, C. D. (2013), Bilingual Word Embeddings for Phrase-based Machine Translation, *in* ‘Conference on Empirical Methods in Natural Language Processing (EMNLP)’, Association for Computational Linguistics, pp. 1393–1398.