

Managing Data Dynamics, Streams and Sharing in the Internet of Things



THE UNIVERSITY
of ADELAIDE

Yongrui (Louie) Qin

School of Computer Science

The University of Adelaide

This dissertation is submitted for the degree of

Doctor of Philosophy

Supervisors: A/Prof. Michael Sheng, Dr. Nickolas J.G. Falkner

and Prof. Hua Wang

August 2015

© Copyright by
Yongrui (Louie) Qin
August 2015

All rights reserved.

No part of the publication may be reproduced in any form by print, photoprint,
microfilm or any other means without written permission from the author.

*To my mother and father,
my wife and my little princess,
my brother,
who made all of this possible,
for their endless encouragement and patience.*

Declaration

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 give consent to this copy of my thesis, when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

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.

Yongrui (Louie) Qin

August 2015

Acknowledgements

This thesis would not have been possible without the support and help from some important people in my life. I would like to take the opportunity to thank all those who have helped me during my PhD journey.

First of all, my sincere thanks will go to my principal supervisor Prof. Michael Sheng, who has taught me how to do good research as well as how to be a person with better personality. He has always been patient, passionate, encouraging throughout the whole journey of my PhD study. His many insightful suggestions and comments on my research have significantly improved the work in this thesis.

Second, I am very thankful to my co-supervisors, Dr. Nickolas J.G. Falkner and Prof. Hua Wang. I want to thank Dr. Falkner for his insightful suggestions on improving my research work and drafts of my papers. I want to thank Prof. Wang for his guidance on my research and for his encouragement to do high-quality research.

I would also like to express my gratitude to Dr. Edward Curry at the Insight Centre for Data Analytics at the National University of Ireland, Galway for providing me the opportunity to work with him and his team as a research intern. My internship experience there has significantly contributed to the work in this thesis.

It has been a great pleasure working with the faculty, staff, students at the University of Adelaide, during my PhD study, and I would like to thank them all for such a great graduate school experience.

I express my sincere appreciation to the University of Adelaide, who provided the Adelaide Scholarship International (ASI), to financially support my work in this dissertation.

No need to mention the great love and support from my family. I am thankful to my parents, who have given up so much and worked so hard to earn money for my education and better future. I am forever indebted to them. I would like to thank my lovely younger brother for his company with me during my childhood and primary and secondary education. I sincerely wish that he will be able to escape from schizophrenia soon and live a better life in future. I am also deeply grateful to my wife and little princess for their constant love and support.

Abstract

Recently, the Internet of Things (IoT) has gained momentum in connecting everyday objects to the Internet and facilitating machine-to-human and machine-to-machine communication with the physical world. IoT offers the capability to connect and integrate both digital and physical entities, enabling a whole new class of applications and services.

This thesis firstly reviews the state-of-the-art research efforts in IoT from data-centric perspectives, including data stream processing, data storage models, complex event processing, and searching in IoT by identifying an IoT data taxonomy, which includes ten key data elements of IoT data under three categorizations. In this thesis, we focus ourselves on three aspects of data management in IoT: data dynamics, data velocity, and data incompleteness. More specifically, we study data dynamics in dynamic graphs, handle data velocity in streams, and tackle data incompleteness via sharing.

In IoT, connections and relations between things are universal and highly dynamic. It is natural to model these connections and relations using dynamic graphs. Meanwhile, shortest path computation is one of the most fundamental operations for managing and analyzing graphs. In this thesis, we focus on the problem of computing the shortest path distance in graphs subject to edge failures. We propose SIEF, a Supplemental Index for Edge Failures in a dynamic graph, which is based on distance labeling, to support distance queries in dynamic graphs with edge failures efficiently.

In IoT, one challenging issue is how to disseminate streaming data to relevant consumers efficiently. Semantic technologies aim to facilitate machine-to-machine (M2M) communication and are attracting more and more interest from both academia and industry, especially in the emerging IoT. This thesis leverages semantic technologies, such as Linked Data, which can facilitate M2M communications to build an efficient information dissemination system for semantic IoT. The system integrates Linked Data streams generated from various data collectors and disseminates matched data to relevant data consumers based on triple pattern queries registered in the system by the consumers. We also design new data structures, *TP-automata* and *CTP-automata*, to meet the high performance needs of Linked Data dissemination.

To tackle data incompleteness, we consider large-scale information sharing scenarios among mobile objects in IoT. By leveraging semantic techniques, we propose broadcasting Linked Data on-air to allow simultaneous access to the information and to achieve better scalability. We introduce a novel air indexing method to reduce the information access latency and energy consumption. We also study the data placement problem of periodic XML data broadcast in IoT environments to facilitate data sharing in IoT. Taking advantage of the structured characteristics of XML data, we present a theoretical analysis on the XML data placement on a wireless channel, which forms the basis of our novel data placement algorithm.

This thesis also discusses on-going and emerging IoT applications, and open research issues for processing and managing IoT data. Several representative domains where IoT can make profound changes are explored, and some key directions for future research and development from a data-centric perspective are identified.

Table of contents

List of figures	xvii
List of tables	xxi
1 Introduction	1
1.1 Background	1
1.2 Research Scope and Contributions	4
1.3 List of Publications	6
1.4 Thesis Organization	7
2 Literature Review	9
2.1 IoT Data Taxonomy	9
2.1.1 Data Generation	9
2.1.2 Data Quality	11
2.1.3 Data Interoperability	12
2.2 Data Streams	13
2.2.1 General Data Stream Processing	13
2.2.2 RFID Data Stream Processing	18
2.2.3 RDF Triple Stream Processing	22
2.3 Data Storage Models	25
2.3.1 New Architecture	25
2.3.2 Large-Scale Storage in Distributed Environments	27

2.3.3	Storage on Resource-Constrained Devices	29
2.4	Search Techniques	30
2.4.1	Deep Web and Semantic Web	30
2.4.2	Web Search	32
2.4.3	Search of Things in IoT	34
2.5	Complex Event Processing	36
2.5.1	Complex Event Processing	37
2.5.2	Semantic Complex Event Processing	39
2.6	Summary	43
3	Shortest Path Computation in Dynamic Graphs	45
3.1	Overview	46
3.2	Related Work	51
3.3	Preliminaries	53
3.3.1	2-Hop Distance Labeling	53
3.3.2	Well-Ordering 2-Hop Distance Labeling	54
3.3.3	Properties of Well-Ordering 2-Hop Distance Labeling	56
3.4	The SIEF Approach	59
3.4.1	SIEF Overview	59
3.4.2	Identification of Affected Vertices	60
3.4.3	Relabeling: Supplemental Index Construction	64
3.4.4	Distance Query Evaluation on SIEF	67
3.4.5	Some Remarks	68
3.4.6	Applying SIEF in Multi-Edge Failure Graphs	69
3.5	Experiments	71
3.5.1	Datasets	71
3.5.2	Performance on Single-Edge Failure	73
3.5.3	Performance on Multi-Edge Failure	79

3.6	Summary	82
4	Matching over Linked Data Streams	85
4.1	Overview	85
4.2	Linked Data Dissemination System	88
4.2.1	System Overview	89
4.2.2	TP-automata for Single Triple Pattern Query Matching	91
4.2.3	CTP-automata for Conjunctive Triple Pattern Query Matching	92
4.3	Experimental Evaluation	98
4.3.1	Experimental Setup	98
4.3.2	Evaluation of TP-automata	99
4.3.3	Evaluation of CTP-automata	101
4.3.4	Limitations	106
4.4	Related Work	106
4.5	Summary	108
5	Data Sharing in IoT Environments Part I	111
5.1	Overview	112
5.2	Related Work	114
5.3	Air Indexing for Linked Data	115
5.3.1	Wireless Broadcast Model	115
5.3.2	Mapping between Triples and 3D Points	116
5.3.3	3D Hilbert Curve Index	116
5.4	Experiments	122
5.4.1	Experimental Setup	122
5.4.2	Performance Analysis	123
5.5	Summary	126

6	Data Sharing in IoT Environments Part II	129
6.1	Overview	130
6.2	Background	133
6.2.1	Application Scenario	133
6.2.2	A Wireless XML Data Broadcast System	135
6.2.3	XML Structural Sharing	136
6.3	Analysis of the Data Placement Problem	138
6.4	The Data Placement Algorithm	147
6.4.1	Structural Sharing in XML Data	147
6.4.2	The Greedy Data Placement Algorithm	151
6.4.3	Index Distribution Strategy	153
6.5	Experiments	155
6.5.1	Experimental Setup	156
6.5.2	Performance of GDPA	158
6.6	Related Work	167
6.7	Summary	169
7	Emerging IoT Applications and Open Issues	171
7.1	Emerging IoT Applications	171
7.1.1	Smart Cities and Homes	171
7.1.2	Environment Monitoring	173
7.1.3	Health	174
7.1.4	Energy	175
7.1.5	Business	176
7.2	Open Issues	177
8	Conclusions and Future Work	183
8.1	Conclusions	183
8.2	Directions for Future Work	185

List of figures

1.1	Internet of Computers v.s. Internet of Things	2
2.1	IoT Data Taxonomy	11
2.2	Nested CEP Query Example	39
2.3	Semantic Complex Event Processing System Overview	40
3.1	A graph example	55
3.2	Affected vertices identification	63
3.3	Supplemental index construction: BFS ALL on failed edge (0,8) . . .	66
3.4	Dual-Edge Failure Distance Estimation	70
3.5	Comparisons between supplemental label entry numbers (SLENs) and original label entry numbers (OLENs)	74
3.6	Index Size	75
3.7	Labeling Time	79
3.8	Correctness and Disconnectedness on Dual-Edge Failure	81
3.9	Correctness and Disconnectedness on Triple-Edge Failure	82
4.1	System Overview	89
4.2	Structure of TP-automata	91
4.3	Index Structure and Conjunctive Constraints of CTP-automata	94
4.4	Maintenance of Candidate Triple List	97
4.5	An Event Example	99

4.6	Average Construction Time of TP-automata	100
4.7	Average Throughput Evaluation of TP-automata	100
4.8	Average Construction Time of CTP-automata	102
4.9	Average Throughput Evaluation of CTP-automata (Varying Query Number)	103
4.10	Average Throughput Evaluation of CTP-automata (Varying Pattern Number)	104
4.11	Average Throughput Evaluation of CTP-automata (Varying Window Size)	105
5.1	2D Hilbert Curves of order 1 and 2. (a) H_1 , (b) H_1 to H_2 , (c) H_2	117
5.2	3D Hilbert Curve of order 1	118
5.3	B^+ -tree for some Points on Hilbert Curve	121
5.4	Minimal sub-region	122
5.5	Access Latency and Tuning Time	124
5.6	Index Tuning Time and Index Size	125
6.1	A wireless XML data broadcast system	136
6.2	An XML structure tree	137
6.3	A broadcast program showing positions of documents required by query q	140
6.4	$(1, m)$ index distribution	155
6.5	Evaluating <i>AAT</i> Performance on <i>DS1</i> : well-clustered data set with 250 documents	159
6.6	Evaluating <i>AAT</i> Performance on <i>DS2</i> : miscellaneous data set with 250 documents	160
6.7	Evaluating <i>AAT</i> Performance on <i>DS3</i> : a mixed set of well-clustered data and miscellaneous data with 250 documents	161

6.8	Evaluating <i>AAT</i> Performance on <i>DS4</i> : miscellaneous data set with 500 documents	162
6.9	Evaluating <i>AAT</i> Performance on <i>DS5</i> : miscellaneous data set with 1000 documents	163

List of tables

2.1	Comparisons of RFID Streaming Techniques.	22
2.2	Comparisons of Linked Stream Processing and Reasoning.	24
2.3	Comparisons of three types of distributed storage systems.	28
2.4	Comparisons of Search Techniques.	31
2.5	Comparisons of CEP Techniques.	41
3.1	2-Hop Distance Labeling L for Figure 3.1	55
3.2	Real-world Datasets and Their Statistics	72
3.3	Affected Vertices	76
3.4	Average Query Time	77
3.5	Average Identification Time	78
3.6	Average Distance Error Ratio	81
3.7	Average Query Time	82
4.1	Conjunctive Constraints	96
4.2	Matching Quality of HashMat (When the Recall is 100%) in TP-automata	101
4.3	Workload Parameters for the Experiments of CTP-automata	102
6.1	Symbols Overview	139
6.2	Matching Cases for Document d_1 and d_2 in a Document Set \mathcal{D}	149
6.3	Data Sets in Our Experiments	156

6.4	Similarity between Clusters in <i>DS1</i>	157
6.5	Workload Parameters for the Experiments	157
6.6	Query Selectivity and Document Coverage Rate	166
6.7	Data Placement Update Time (in milliseconds)	167