

Correlation Management and Search for the Internet of Things



THE UNIVERSITY
of ADELAIDE

Ali Shemshadi

School of Computer Science

The University of Adelaide

This dissertation is submitted for the degree of

Doctor of Philosophy

Supervisors: Prof. Michael Sheng

September 2016

© Copyright by

Ali Shemshadi

September 2016

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 two little princes,
my brother and sister,
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.

Ali Shemshadi

September 2016

Acknowledgements

I could not have arrived at this point without the help and support of my peers, supervisors, instructors, friends and of course, my family. I found the experience of working with the school, staff and students at the University of Adelaide to be both joyful and fruitful and thus, I would like to thank all of these people, who made my PhD journey such a great experience. Firstly, I owe a sincere gratitude towards my supervisor, Prof. Michael Sheng, a compassionate teacher, a hard-working researcher and an outstanding supervisor. His motto from the early days of my graduate study was “aim high and never compromise” keeps me inspired and motivated. This success was due to his devotion to his students, which is exemplary and unique. His kind personality made me rethink about the offers from other institutions and thoroughly changed my life and career path towards a highly positive direction. Secondly, I would like to thank my co-supervisors, Prof. Zbigniew Michalewicz and Prof. Hong Shen for their supports before and throughout my post-graduate study.

I am blessed to find the opportunity to befriend with many fellow colleagues in our research group. I appreciate every minute that I had the opportunity to be with them. In particular, I would like to acknowledge Dr. Yongrui Qin, Dr. Lina Yao and Ms. Wei Emma Zhang. It was such a pleasure for me to collaborate with and learn from them. Furthermore, I would like to thank the staff at Xively, the IoT platform for sharing their data.

Lastly, but not the least, I owe a huge debt of gratitude to my mother, my father, my wife, my both princes, my brother and my sister for their patience, encouragement and support without which, I could not be successful at any point of time in the past, present and future.

Abstract

The Internet of Things (IoT) is a compelling paradigm, which aims to enable everyday physical things embedded with electronics, software, sensors, and network connectivity to collect and exchange data on the Internet. It is anticipated that by 2020, billions of things get connected to the Internet. Creating future IoT search engines is a key step towards unlocking answering the above question. Future search engines can potentially revolutionise various applications in different domains. Existing approaches for searching the IoT use simple techniques to obtain a list of things for a query. The state of the art needs to be improved in different aspects. For instance, it is often disregarded that in the context of IoT, we have two types of users including machines and human users. In addition, many have complained about the absence of the real-world IoT data. Unsurprisingly, a common question that arises regularly nowadays is “Does the IoT already exist?”. So far, little has been known about the real-world situation on IoT, its attributes, the presentation of data and user interests. Moreover, existing approaches also disregard the attribute based correlations between things in the real-world. In this dissertation, we review the state of the art in IoT search domain and propose a novel framework to collect and analyse IoT data. Our system is also able to resolve IoT queries based on the knowledge that is acquired from the IoT data sources. Furthermore, we introduce a novel technique to extract the correlations between things. Our framework is capable of using the correlations to improve the quality of search results for both types of users. We investigate the scalability and the effectiveness of our approach using large scale and real-world datasets. Moreover, we investigate two case studies in transport systems in

our research. The first case study, challenges the complex problem of taxi ridesharing in the context of smart cities. The second case study, involves a real-time prediction method for flight delays based on the IoT sourced data.

Table of contents

List of figures	xvii
List of tables	xxi
1 Introduction	1
1.1 Motivating Scenario	3
1.2 Research Issues	7
1.3 Contributions Overview	9
1.3.1 Data Collection	10
1.3.2 Correlation Discovery	10
1.3.3 Diversified Query Resolution	11
1.3.4 Pattern Matching for Correlation Graphs	11
1.3.5 Intent-Oriented Search:Taxi Ridesharing	12
1.3.6 A Crawling and Search Engine	12
1.4 Dissertation Publications	12
1.5 Dissertation Organization	15
2 Crawling the IoT Data	19
2.1 Where is the IoT?	22
2.1.1 Cloud Based IoT Platforms	22
2.1.2 WoT Enabled Platforms	24

2.1.3	Web Mapping Enabled Data Sources	24
2.2	IoT Data Acquisition	25
2.2.1	Identification of Data Sources	27
2.2.2	IoT Data Collection	28
2.3	IoT Data Analysis	30
2.3.1	User Interests	30
2.3.2	IoT Data Characteristics	32
2.3.3	IoT vs. User Interests	44
2.4	Discussions	45
2.4.1	Challenges in IoT Data Discovery	45
2.4.2	Information Retrieval in IoT	46
2.4.3	Other Challenges	47
2.5	Related Work	49
2.6	Summary	51
3	Interlinking IoT Resources	53
3.1	The CEIoT Approach	56
3.1.1	Correlation Discovery Process	57
3.1.2	Framework Architecture and System Entities	57
3.1.3	Correlation Extraction	61
3.1.4	Correlation Representation	64
3.2	Experimental Results	66
3.2.1	System Performance	68
3.2.2	Things Correlation Graph	72
3.2.3	Message Volume	72
3.3	Related Work	73
3.4	Summary	74

4	Pattern Matching for Things Correlation Graphs	77
4.1	Problem Formulation	82
4.2	The Naive Approach	83
4.3	Background	85
4.3.1	Markov Chains	85
4.3.2	Markov Chain Monte-Carlo	87
4.4	Pattern Matching	88
4.4.1	Identifying Matches	89
4.4.2	Top-k Matches	93
4.5	Experimental Evaluation	96
4.5.1	Experimental Setting	96
4.5.2	Efficiency	100
4.5.3	Discussion	105
4.6	Related Work	107
4.7	Summary	109
5	Diversifying Top-k Query Matches	111
5.1	Problem Statement	114
5.1.1	Problem Formulation	114
5.1.2	Methodology	116
5.2	ECS Approach	117
5.2.1	TCG Construction	117
5.2.2	Clustering	118
5.2.3	Selection	120
5.3	Experimental Results	122
5.3.1	Datasets	123
5.3.2	Results	130

5.4	Related Work	133
5.4.1	Search Based on Social Relationships	133
5.4.2	IoT Search Engines	134
5.5	Summary	136
6	Intent Based Search: A Case Study in Taxi Ridesharing	137
6.1	Preliminaries	143
6.1.1	Problem Definition	143
6.1.2	Design Basics	145
6.2	Background: IS vs. DS Approach	151
6.3	The TRIPS Framework	153
6.3.1	Traffic Modeling Layer	156
6.3.2	TRIPS Application Layer	157
6.3.3	Distribution Management Layer	162
6.4	Experiment	163
6.4.1	The Dataset	163
6.4.2	Performance	164
6.4.3	Cost Savings	167
6.4.4	Comparison with Other Solutions	169
6.5	Related Work	170
6.5.1	Dynamic Ridesharing	170
6.5.2	Travel Time Estimation	172
6.5.3	Uncertainty in Spatio-Temporal Data	173
6.6	Summary	174
7	ThingSeek: An Enriched Interface for IoT Search Engine	177
7.1	ThingSeek: An Overview	179

7.1.1	ThingSeek Crawler Engine	179
7.1.2	Query Results Preparation	182
7.2	Demonstration	184
7.2.1	Crawling an IoT Data Source	185
7.2.2	IoT Search by Human Users	185
7.2.3	IoT Search by Smart Machines	186
7.3	ThingSeek in Application: Flight Delay Analysis	187
7.3.1	Model Features	187
7.3.2	Feature Analysis Results	187
7.4	Related Work	192
7.5	Summary	196
8	Conclusion	199
8.1	Summary	199
8.2	Future Research	202
	References	205
	Appendix A Curriculum Vitae	223

List of figures

1.1	Motivating scenario for IoT search	4
2.1	Illustration of the sequential-spatial access to things data	26
2.2	Ranking of the popular IoT services	31
2.3	Query frequency per day in Thingful	31
2.4	The distribution of things trajectories on a map	33
2.5	The distribution of IoT queries on a map	34
2.6	Technologies used by IoT platforms	36
2.7	Comparison of the densities of the query logs and IoT data	37
2.8	EMD through time for IoT data sources	38
2.9	Sensor update percentage	41
3.1	ThingSpeak mashup example	54
3.2	Correlation discovery process for IoT	58
3.3	CEIoT Framework	59
3.4	Search area breakdown	62
3.5	Multiple correlations example	66
3.6	CEIoT system view	69
3.7	Characteristics of the data set and the final results	70
3.8	CLOR graphs for connections with different thresholds	71

4.1	Example query and data graph	79
4.2	Graphs after removing non-matching edges	92
4.3	Visualization of pattern graphs	98
4.4	Results for the total runtime	100
4.5	Matching results for the first experiment	101
4.6	Matching results for the second experiment	102
4.7	Matching results for the third experiment	103
4.8	Matching results for the synthetic datasets	104
4.9	Comparison with the baseline	106
5.1	Correlations in IoT	112
5.2	ECS Components	116
5.3	Distribution of query keywords from <i>Thoughtful</i> dataset	122
5.4	Visualization of two TCGs	126
5.5	Visualization of different selection methods on the trajectory dataset	127
5.6	Results for weather stations dataset	128
5.7	Distribution of 3D location data for weather stations	129
5.8	Experimental results for IoT search results diversification	131
6.1	Intent-based search in ridesharing scenario	140
6.2	Extended search areas for a query	145
6.3	Possible schedule sequences for a query	147
6.4	TRIPS Framework	155
6.5	Symmetric vs. asymmetric extensions of search	158
6.6	Availability of the index for all origins and destinations	164
6.7	The results of the experimental evaluation	165
6.8	Effectiveness results	168

7.1	Architecture of the ThingSeek crawler engine	180
7.2	Query resolution in ThingSeek	183
7.3	Query resolution for smart things in ThingSeek framework	185
7.4	ThingSeek Web based visualization	186
7.5	Features affecting flight delays from each source	188
7.6	Example of the flight records	190
7.7	Example of the weather records	191
7.8	Example of the air quality records	191
7.9	Delay at departure performance for airlines	192
7.10	Delay at departure performance for airports	193
7.11	Results of the correlation analysis	193

List of tables

2.1	Examples of IoT cloud services	24
2.2	Most popular keywords and their categories	35
2.3	WoT vs. IoT cloud services	36
2.4	Sensor readings from Xively platform	39
2.5	Transportation data sources with overlapping set of objects	43
3.1	Requirements of traditional WWW hyperlinks vs. novel IoT-links	54
4.1	The set of label assignments	79
4.2	Nodes with label similarity above threshold	80
4.3	Mapping of the edges of G to edges in the query graph Q	93
4.4	Index construction summary for different datasets	99
5.1	Object data structure for taxi trajectories dataset	124
5.2	Object data structure for weather stations dataset	124
6.1	List of important notations	146
6.2	Taxi speed estimation based on GPS reading analysis	170
7.1	Feature description for flight delay search	189