

Source Profiling for Smart City Sensing



THE UNIVERSITY
of ADELAIDE

Yihong Zhang

School of Computer Science

The University of Adelaide

This dissertation is submitted for the degree of

Doctor of Philosophy

Supervisors: Prof. Michael Sheng and Dr. Claudia Szabo

October 2016

© Copyright by

Yihong Zhang

October 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.

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.

Yihong Zhang

October 2016

To My Mother and Father
Who are in my flesh and soul

Acknowledgements

This thesis is the result of contributions from many people. I would like to take this opportunity to acknowledge their great effort.

I am deeply indebted to my supervisor, Dr. Claudia Szabo, who showed me the gate to a research career when I was still a master student and uncertain about my future. At the beginning of my PhD, she trained me and instilled in me the necessary skills for conducting scientific researches, through our numerous meetings, discussions, and arguments. I would not have acquired these skills without her enormous patience and persistence. When I started writing research papers, she also assisted me in working through numerous revisions for each of these papers, some of which fortunately get published and form a large part of this thesis.

I am also thankful to the support of my supervisor, Prof. Michael Sheng, whose wisdom and generosity have helped me through the more difficult times in my PhD. Prof. Michael Sheng always encouraged me to aim high and strive for excellence, setting himself as a good example. These encouragements in the end allowed me to avoid many problems and traps in a research career. Without his kind guidance, my progress would certainly be much more tedious and bitter.

I would also like extend my thanks to my senior colleagues Lina, Yongrui, Kewen, Ali and Scott, who have offered useful advises and shared their experiences to help me get through my PhD. And also to my colleague Susie, who helped me conduct some of the experiments presented in this thesis. And also to my colleagues Javier, Lachlan, and Nguyen, who proof-read many chapters in this thesis and provided interesting feedbacks.

I would also like to thank my friends Tony, David and Nick. The time I spent hanging out with them provide me the much-needed relief in a rigid process.

Finally I thank my mother and father for their patience, encouragement and support in this extraordinary long journey of three-year-and-a-half.

This thesis is supported by the Faculty of ECMS Divisional Scholarship, the University of Adelaide.

Abstract

Source Profiling for Smart City Sensing

by

Yihong Zhang

Doctor of Philosophy in Computer Science

The University of Adelaide, 2016

Recent years have seen the emergence of smart cities, which utilize various sensing data for applications such as pollution monitoring, infrastructure planning and traffic control. Current sensing projects tend to deploy a large number of low-cost and unreliable sensing sources, rather than a small number of high-quality sensing sources. It is therefore critical to provide data analysis in the face of unreliable sources.

This thesis focuses on two types of sensing sources that have been used in smart city sensing projects, namely, environmental sensors and human sensors. The environmental sensors are physical sensors that are made to monitor certain environmental features, such as temperature, humidity, and pollutant concentration. An environmental sensor can fail frequently and will start generating faulty data when there is chemical compound decay, battery exhaustion, or calibration problems. Human sensors, as recently proposed in a new area called social sensing, are online messaging platform users who post observations about their surrounding environments. The data generated by human sensors can be erroneous because the natural language used in their messages does not conform to a machine-readable

standard. Based on a survey of existing literature, this thesis presents source profiling-based solutions for three data analysis problems, data cleaning in environmental sensing, observation message classification in social sensing, and message location inference. Each of the solutions is validated with various real-world data and extensive experiments.

For data cleaning in environmental sensing, we propose two solutions, approaching from a frequentist perspective and a Bayesian perspective, respectively. The frequentist approach determines sensor reliability based on the frequency of reliable behavior in the past, and in each data collection iteration updates a reliability score, which can be used to weight down or remove the data from unreliable sources. The Bayesian approach models sensor reliability as a latent variable, and applies the Expectation Maximization framework to discover the latent sensor reliability and correct reading values for the environmental feature.

For observation message classification, we propose supervised and unsupervised solutions. We propose a supervised solution to distinguish messages according to three perspectives, namely, observation, affection, and speculation. We next propose a supervised solution based on user features such as trending activity, communication status, and writing styles. And finally, we propose an unsupervised solution based on lexical analysis and user profiling in four user attributes, namely, originality, interactivity, objectivity, and topic focus.

For location inference, we propose a solution based on name entity extraction and user message histories. The proposed solution extracts location names from text messages using a gazetteer, and after retrieving a number of past locations from the message history of a user, it applies outlier removal before inferring the current location. Incorporating observation classification and location inference, we propose an event detection system called Sense and Focus (SNAF), which detects real world events based on discussions exchanged on Twitter. A prototype implementation of the system has shown a number of detection results, 54% of which corresponding to real-world events, and in many case detected earlier than news reports, and with less than 1.5km location error.

Table of contents

List of figures	xv
List of tables	xvii
Nomenclature	xix
1 Introduction	1
1.1 Research Objectives	3
1.2 Contributions	4
1.3 Publications Related to This Thesis	6
1.4 Thesis Structure	8
2 Smart City Sensing: Sensors and Twitter Users	11
2.1 Introduction	12
2.2 Internet-of-Things Environmental Sensors	14
2.3 Participatory Sensing on Twitter	18
2.4 Opportunistic Sensing on Twitter	20
2.5 Summary	23
3 Environmental Sensor Profiling for Predicting Correct Reading Values	25
3.1 Overview	26
3.2 Related Work	29

3.3	Frequentist Approach: Incremental Reliability Update	32
3.3.1	Faulty Data and Reliability	33
3.3.2	Influence Mean	35
3.3.3	Incremental Reliability Update	36
3.3.4	IM-Reduce and IM-Remove	38
3.4	Bayesian Approach: Expectation Maximization	39
3.4.1	Background	39
3.4.2	Likelihood Model	42
3.4.3	Expectation Maximization for Finding Reliable Sensors	43
3.4.4	Final Algorithm	46
3.5	Experimental Analysis	47
3.5.1	Simulated Air Pollution Sensing Data	48
3.5.2	Real Data with Synthetic Faults	50
3.5.3	Melbourne Weather Data	56
3.5.4	OpenSense Ozone Data	59
3.5.5	Discussion	62
3.6	Summary	62
4	User Profiling for Classifying Observation Tweets	65
4.1	Overview	66
4.2	Related Work	69
4.3	Message Perspective Classification	73
4.3.1	Message Perspectives	74
4.3.2	Identifying Lexical Features	76
4.3.3	Identifying Textual Features	78
4.3.4	Machine Learning Classifiers	78
4.4	Supervised Observation Classification with User Features	80

4.4.1	Generating User Features	80
4.4.2	Trending Activity Features	81
4.4.3	Communication Status Features	82
4.4.4	Writing Style Features	82
4.5	Unsupervised Observation Classification	84
4.5.1	Observation Filtering	84
4.5.2	User Profiling for Personal Account Classification	88
4.5.3	Personal Account Classification with Profiles	91
4.5.4	Overall Algorithm	95
4.6	Experimental Analysis	95
4.6.1	Message Perspective Classification	96
4.6.2	Supervised Classification with User Features	100
4.6.3	Unsupervised Observation Classification	104
4.7	Summary	109
5	User Location Profiling for Localized Event Detection	111
5.1	Overview	112
5.2	Related Work	114
5.3	SNAF: Location Inference and Event Detection	116
5.3.1	Location Extraction for A Single Tweet	117
5.3.2	Location Resolution Given Past Locations	120
5.3.3	Realtime Event Monitoring	125
5.4	Experimental Analysis	127
5.4.1	Datasets	128
5.4.2	Measurements and Baseline Methods	129
5.4.3	Location Accuracy Results	129
5.4.4	Event Detection Results	131

5.5	A Prototype Realtime Event Monitoring System	133
5.6	Discussion	136
5.7	Summary	137
6	Conclusion	139
6.1	Thesis Summary	139
6.1.1	Insights of Current Situation	139
6.1.2	Developing Data Cleaning Techniques	140
6.1.3	Applying Cleaned Data	142
6.2	Limitations	143
6.2.1	Unreliable Sensor Behavior Assumption	143
6.2.2	Consistent User Feature Assumption	143
6.2.3	Multiple Sensing Source Requirement	144
6.2.4	Access to Message History Requirement	145
6.3	Extending Results	145
6.3.1	Data Integration with Unreliable Sources	145
6.3.2	Online Rumor Detection	146
6.4	Future Works	147
6.4.1	Sensor Fault Detection with Multiple Features	147
6.4.2	Improving Social Sensing by Exploiting Friend Network	148
6.4.3	Mash-up of Environmental Sensing and Social Sensing	148
	References	151
	Appendix A Curriculum Vitae	167

List of figures

3.1	Incremental Reliability Update	32
3.2	Faulty sensor data patterns	33
3.3	Environmental readings within one hour	41
3.4	Simulated Air Pollution Data	49
3.5	Simulated Air Pollution Noisy readings	50
3.6	IM methods on Simulated Air Pollution Data	51
3.7	Intel Lab Data with Synthetic Faults	52
3.8	IM methods on Intel Lab Data	54
3.9	EM method on Intel Lab Data	55
3.10	The Melbourne Weather Data Sensor Location	56
3.11	The Melbourne Weather Data	57
3.12	IM methods on Melbourne Weather Dataset	58
3.13	EM method on Melbourne Weather Dataset	59
3.14	The OpenSense Ozone Data	60
3.15	IM methods on OpenSense Dataset	61
3.16	EM method on OpenSense Dataset	61
4.1	Supervised Message Perspective Classification	73
4.2	Unsupervised Observation Classification	85
4.3	Perspective classification accuracy with lexical features	98

4.4	Perspective classification accuracy results with lexical and textual features	99
5.1	The overall architecture of SNAF: Sense and Focus	117
5.2	Mean location differences and percentages for tweets in different past months	121
5.3	Event detection based on connected components	126
5.4	Identified report locations in North America	130
5.5	The map view of the event detection system	134
5.6	The tweet view of the event detection system	135

List of tables

2.1	Traditional Sensor Network vs. IoT Sensing Project	16
2.2	Sensor-generated Twitter accounts and typical refresh rate	20
3.1	Reading Prediction Mean Square Error	62
4.1	Examples of tweets in a given perspective	74
4.2	Effective lexical categories for different perspectives	77
4.3	Selected lexical categories and example words	77
4.4	Textual features for perspective classification	78
4.5	Example feature vector	79
4.6	User Feature Sets	81
4.7	Business Account Identifier	83
4.8	Originality Test Rules	83
4.9	Non-observation tweets filtered by POS tagging, for monitoring flight delay, shooting incidents, and rainbows	86
4.10	Examples of specific-purpose accounts	89
4.11	Perspective classification accuracy with lexical features	97
4.12	Perspective classification accuracy results with lexical and textual features .	97
4.13	Classification Precision	102
4.14	Classification F-value	103

4.15	Filtering accuracy for hailstorm and car accident datasets	106
4.16	Filtering accuracy for the crisis dataset	108
5.1	Examples of non-placenames in the geo-coordinate dataset	118
5.2	Precision of location extraction on single tweets using refined gazetteer . .	119
5.3	Examples of placename in tweets not indicating user location	120
5.4	Precision and recall for Random Forest	128
5.5	Accuracy of location inference methods	131
5.6	Examples of detected events	132

Nomenclature

Roman Symbols

CDE	Crime and Disaster Events
EM	Expectation Maximization
GPS	Global Positioning System
IM	Influence Mean
KNN	k-Nearest Neighborhood
LDA	Linear Discriminant Analysis
LIWC	Linguistic Inquiry and Word Count
LLSE	Linear Least-Squares Estimation
MSE	Mean Square Error
NGO	Non-governmental organization
POI	Point-of-interest
POS	Part-of-speech
PSO	Particle swarm optimization

RF Random Forrest

RFID Radio-frequency identification

SNAF Sense and Focus Event Monitor System

SVM Support Vector Machine