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

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

List of tables

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