REAL-TIME C2C MATCHING OF SOCIAL MEDIA MESSAGES

M.R.M. RILFI

148053N

Degree of Master of Science

Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

May 2019

REAL-TIME C2C MATCHING OF SOCIAL MEDIA MESSAGES

M.R.M. RILFI

148053N

Dissertation submitted in partial fulfillment of the requirements for the degree Master of Science in Computer Science

Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

May 2019

DECLARATION

I declare that this is my own work and this thesis/dissertation does not incorporate any material previously submitted for a Degree or Diploma in any other University of institute of higher learning and to the best of my knowledge and belief, it does not contain any material previously published or written by another person except where the acknowledgment is made in the text.

Also, I hereby grant to the University of Moratuwa the non-exclusive right to reproduce and distribute my thesis/dissertation, in whole or in part in print, electronic or another medium. I retain the right to use this content in whole or part in future works (such as articles or books).

……………………………………… ………………………………………..

M.R.M. Rilfi Date

The above candidate has carried out research for the Masters' Dissertation under my supervision.

……………………………… …………..

Dr. H.M.N. Dilum Bandara Date

……………………………… …………..

Dr. Surangika Ranathunga Date

Abstract

Social media enables personalization of the Consumer to Consumer (C2C) business model where people could directly do business with each other without an intermediary by sharing their products, services, and consumer requirements. However, messages shared by both the sellers and potential buyers do not reach each other as they are embedded among other social media messages. Moreover, C2C buy/sell interest matching in real time is nontrivial due to the complexities of interpreting social media messages, number of messages, and diversity of products and services. We present a platform for real-time matching of microblogging messages related to product selling or buying in C2C. We adopt a combination of techniques from natural language processing, complex event processing, and distributed systems. First, we extract the semantics of messages such as product attributes and commercial intention of the message either buying or selling using information extraction. Then the extracted buy/sell messages are matched using a complex event processor. Moreover, NoSQL and in-memory computing are used to enhance scalability and performance. The proposed solution shows a high accuracy where commercial intent classification and Conditional random fields based named entity recognition recorded an accuracy of 98.5% and 82.07%, respectively when applied to a real-world dataset. Information extraction, in-memory data manipulation, and complex event processing steps introduced low latency were latencies were 0.5 ms, 5 ms, and 0.2 ms, respectively. For the given setup with modest hardware, we were able to process 3,400 messages per second and overall latency was 0.76 ms.

Keywords: C2C; complex event processing; information extraction; named entity recognition; stream processing;

ACKNOWLEDGEMENTS

First and foremost, I express my sincere gratitude to my advisor Dr. Dilum Bandara. His continuous support, patience, guidance, and advice made the successful completion of this research possible. I am thankful to him for regular meetings and productive discussions despite his busy schedule. I highly appreciate his tolerance during times of my slow progress due to health problems or other issues. It has been a wonderful experience working with him, during which I have acquired and improved new knowledge and skills that will be useful to my career in the future. Also, I like to thank my co-supervisor Dr. Ms. Surangika Ranathunga for her valuable guidance and support.

I would like to thank members of my review committee Dr. Shehan Perera and Dr. Lochandaka Ranathunga for their helpful feedback during my progress reviews.

My sincere thanks go to the Computer Science Department of the University of Moratuwa for facilitating me with the necessary resources throughout the course of my research. I am thankful to Dr. Shehan Perera and Prof. Sanath Jayasena for granted me an Instructor position to me which helped me to carry out my research without any hindrance. Thanks, are to the System Engineers of the Department, especially Mr. Sujith Fernando, for helping me acquire, set up and manage hardware resources. I also thank the department staff in general for their friendly interaction, making my time at the department a fruitful and pleasant one.

TABLE OF CONTENTS