Faculty of Engineering and Information Technology
University of Technology, Sydney

Mining Actionable Combined Patterns Satisfied both Utility and Frequency Criteria

A thesis submitted in partial fulfillment of
the requirements for the degree of
Master of Analytics by Research

by

Jingyu Shao

June 2016

CERTIFICATE OF AUTHORSHIP/ORIGINALITY

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature of Candidate

Acknowledgments

Foremost, I would like to express my sincere gratitude to my supervisors Prof. Longbing Cao for the continuous support of my master degree study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor during my study.

I also would like to appreciate my co-supervisor Doctor Guandong Xu and Wei Liu for providing me with continuous support throughout my PhD study and research. Without their professional guidance and persistent help, this thesis would not have been possible.

I thank my fellow labmates in Advanced Analystics Institute: Junfu Yin, Xiangfu Meng, Xing Wang, Shoujin Wang, Xuhui Fan, CC Chen, Jia Xu and Liang Hu for the stimulating discussions, for the sleepless nights we were working together before deadlines, and for all the fun we have had in the last two years.

Last but not the least, I would like to thank my family: my father and my mother, for their unconditional support, both financially and emotionally throughout the whole master studying.

Jingyu
November 2015 @ UTS

Contents

List of Figures

List of Tables

List of Publications

Papers Published

- **Jingyu Shao**, Junfu Yin, Wei Liu, Longbing Cao (2015), Mining Actionable Combined Patterns of High Utility and Frequency. *in* 'Proceedings of the 2015 IEEE International Conference on Data Science and Advanced Analytics (**DSAA15**)', full paper accepted.

- **Jingyu Shao**, Junfu Yin, Wei Liu, Longbing Cao (2015), Actionable Combined High Utility Itemset Mining. *in* 'Proceedings of the 29th Association for the Advancement of Artificial Intelligence (**AAAI15**)', poster accepted.

- Xiangfu Meng, Longbing Cao, **Jingyu Shao** (2014), Semantic Approximate Keyword Query Based on Keyword and Query Coupling Relationship Analysis. *in* 'Proceedings of the 23rd ACM International Conference on Information and Knowledge Management (**CIKM14**)' , pp. 529-538.

- Xiangfu Meng, Longbing Cao, **Jingyu Shao** (2014), Finding Top-k Semantically Related Terms in Relational Keyword Search. *in* 'Proceedings of the 2014 IEEE International Conference on Data Science and Advanced Analytics (**DSAA14**)', pp. 505-511.

Papers to be Submitted/Under Review

- **Jingyu Shao**, Junfu Yin, Longbing Cao (2015), Mining Strong Associated Patterns among High Utility Itemsets. "To be submitted".

Abstract

In the last two decades, researchers have proposed numerous approaches and techniques for extracting frequent patterns. Until recent ten years, researchers have not realized the disadvantages of mining frequent patterns in several cases. One paradoxical case is that in a digital store, laptops with quite low frequency earn much higher profit than memory disks which have a high frequency. To tackle such issues, the relative importance of each item has been introduced into frequent pattern mining, and the concept "high utility itemsets mining" has been proposed. The criteria for discovering high utility patterns is a user-specified minimum utility threshold, instead of a minimum support threshold, to extract itemsets with high utilities. Even though the introduction of utility can solve some business issues better than the frequency-based measurements, the resultant patterns are still not actionable in tackling business concerns.

Accordingly, actionable knowledge discovery is proposed to identify informative and decision-making-friendly knowledge that satisfies both technical and business criteria to narrow the large gap between technically identified results and real-world user needs. Actionable pattern mining has proved to be essential for handling those impact-targeted activities and business problems, such as behavior analysis, fraud detection and government-customer debt. In addition, it has an outstanding performance especially in imbalanced datasets. For example, one of the key business concerns in the activity pattern analysis is to find out which particular activity directly triggers or is closely associated with the occurrence of a target impact.

During recent years, actionable knowledge discovery has demonstrated its value in solving business and industrial concerns, where the analysis of pattern relationship plays a foundational role, and combined pattern mining is the basic approach to generate such kind of knowledge. One approach is to develop the utility framework that is more suitable for addressing business consideration than the frequency framework, while none of existing work has been reported on discovering actionable knowledge from utility databases. Hence, it is essential to build an applicable approach for mining actionable combined patterns from utility datasets. However, there are challenges for achieving so. 1) The downward closure property does not hold in utility-based mining approaches, which means that most of the existing algorithms for frequency-based mining cannot be applied. 2) Furthermore, compared to high utility mining methods, actionable combined knowledge discovery faces the critical combinational complexity as well as the complicated structure caused by the dependence between items.

In order to address these research limitations and challenges, this thesis proposes an actionable combined knowledge discovery framework for mining actionable combined patterns that satisfy both utility and frequency requirements. The thesis is organized as follows.

Chapter 2 briefly reviews the related works on the frequent pattern mining framework, high utility itemset mining framework, and the actionable knowledge discovery approach. Chapter 3 incorporates the utility concept into combined pattern mining, and actionable patterns with high utility growth and strong associations are defined and discovered. An efficient algorithm called CUARM (Combined Utility-Association Rule Mining) is presented for actionable high utility pattern mining. A basic tree structure for mining utility growth patterns is proposed, and a measure considering both utility growth and co-occurrence rate is proposed to finalize the discovery of such combined patterns. Chapter 4 discusses how to discover those high utility patterns with highly associated relationship between one item and another. Such patterns have a significant feature, that is the utility increases with the length of such pattern increasing. That is to say, the utilities of the derivative itemsets are always higher than those of underlying itemsets. Also, a hybrid algorithm for mining both highly dependent and utility growth patterns is proposed to obtain those highly dependent actionable patterns.

All of the algorithms are examined in both synthetic and real datasets, and their performance is compared with baselines for mining frequent patterns and high utility patterns. The results show that our proposed actionable combined patterns are more informative for business decision-support.