Faculty of Engineering and Information Technology
University of Technology, Sydney

Mining High Utility Sequential Patterns

A thesis submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

by

Junfu Yin

July 2015

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

To My Father and Mother
For Your Love and Support

Acknowledgments

Foremost, I would like to express my sincere appreciation to my supervisor Prof. Longbing Cao for his continuous support of my Ph.D study and research, for his patience, motivation, enthusiasm, and immense knowledge. Unlike other PhD students, I was recruited by Prof. Cao once I had finished my undergraduate studies. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having had a better advisor and mentor for my Ph.D study.

I also would like to extend gratitude to my co-worker Zhigang Zheng for his hard work on our collaborated papers. Thanks to David Wei and Yin Song for the sleepless nights when we worked together before deadlines, and our co-authored papers were finally accepted. Thanks to all other members in the Advanced Analytics Institute for their selfless support of my research, my life, and all the good times we have had.

I place on record my gratitude to Dr. Haixun Wang and other team members at Microsoft Research Asia for their valuable suggestions on my research. I also thank the workmates in the Shanghai Stock Exchange. They have always been patient in teaching me about the financial markets.

Last but not least, I would like to thank my parents for their unconditional support. Without their endless love, it would never have been possible for me to finish this dissertation.

Junfu Yin
December 2014 @ UTS

Contents

List of Figures

List of Tables

List of Publications

Papers Published

- Jingyu Shao, **Junfu Yin**, Wei Liu, Longbing Cao (2012), Actionable Combined High Utility Itemset Mining. *in* 'Twenty-Ninth AAAI Conference on Artificial Intelligence, AAAI '15, Austin, Texas, USA, January 25-29, 2015 (**AAAI 2015**)' (Poster Accepted).

- Wei Wei, **Junfu Yin**, Jinyan Li, Longbing Cao (2014), Modelling Asymmetry and Tail Dependence among Multiple Variables by Using Partial Regular Vine. *in* 'Proceedings of the 2014 SIAM International Conference on Data Mining, Philadelphia, Pennsylvania, USA, April 24-26, 2014 (**SDM 2014**)', pp. 776-784.

- **Junfu Yin**, Zhigang Zheng, Longbing Cao, Yin Song, Wei Weig (2013), Efficiently Mining Top-K High Utility Sequential Patterns. *in* '2013 IEEE 13th International Conference on Data Mining, Dallas, TX, USA, December 7-10, 2013 (**ICDM 2013**)', pp. 1259-1264.

- Yin Song, Longbing Cao, **Junfu Yin**, Cheng Wang (2013), Extracting discriminative features for identifying abnormal sequences in one-class mode. *in* 'The 2013 International Joint Conference on Neural Network-s, IJCNN 2013, Dallas, TX, USA, August 4-9, 2013 (**IJCNN 2013**)', pp. 1-8.

- **Junfu Yin**, Zhigang Zheng, Longbing Cao (2012), USpan: an efficient algorithm for mining high utility sequential patterns. *in* 'The 18th

ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, Beijing, China, August 12-16, 2012 (**KDD 2012**)', pp. 660-668.

Papers to be Submitted/Under Review

- Chunyang Liu, Ling Chen, **Junfu Yin**, Chengqi Zhang (2014), P^3-Mining: A Profile-based Approach to Summarize Probabilistic Frequent Patterns. *to be submitted*.

- **Junfu Yin**, Zhigang Zheng, Longbing Cao (2014), Efficient Algorithms for Mining High Utility Sequential Patterns. *to be submitted*.

- **Junfu Yin**, Longbing Cao, Chunyang Liu, Zhigang Zheng (2014), CloUSpan: Mining Concise and Lossless High Utility Sequential Patterns. *to be submitted*.

- Jingyu Shao, **Junfu Yin**, Wei Liu, Longbing Cao (2014), Mining Combined High Utility Patterns. *submitted to DSAA 2015*.

- **Junfu Yin**, Longbing Cao, UIP-Miner: An Efficient Algorithm for High Utility Inter-transaction Pattern Mining. *to be submitted*.

Research Reports of Industry Projects

- **Junfu Yin**, Cheng Zheng (Fudan University), Lei Chen (Shanghai Stock Exchange). IPO Stock Manipulation Analysis, Shanghai Stock Analysis ,Oct 2013 - Jan 2014.

Abstract

Sequential pattern mining refers to the identification of frequent subsequences in sequence databases as patterns. It provides an effective way to analyze the sequential data. The selection of interesting sequences is generally based on the frequency/support framework: sequences of high frequency are treated as significant. In the last two decades, researchers have proposed many techniques and algorithms for extracting the frequent sequential patterns, in which the *downward closure property* (also known as *Apriori property*) plays a fundamental role. At the same time, the relative importance of each item has been introduced in frequent pattern mining, and "high utility itemset mining" has been proposed. Instead of selecting high frequency patterns, the utility-based methods extract itemsets with high utilities, and many algorithms and strategies have been proposed. These methods can only process the itemsets in the utility framework.

However, all the above methods suffer from the following common issues and problems to varying extents: 1) Sometimes, most of frequent patterns may not be informative to business decision-making, since they do not show the business value and impact. 2) Even if there is an algorithm that considers the business impact (namely utility), it can only obtain high utility sequences based on a given minimum utility threshold, thus it is very difficult for users to specify an appropriate minimum utility and to directly obtain the most valuable patterns. 3) The algorithm in the utility framework may generate a large number of patterns, many of which maybe redundant.

Although high utility sequential pattern mining is essential, discovering the patterns is challenging for the following reasons: 1) The downward closure property does not hold in utility-based sequence mining. This means that most of the existing algorithms cannot be directly transferred, e.g. from frequent sequential pattern mining to high utility sequential pattern mining. Furthermore, compared to high utility itemset mining, utility-based sequence analysis faces the critical combinational explosion and computational complexity caused by sequencing between sequential elements (itemsets). 2) Since the minimum utility is not given in advance, the algorithm essentially starts searching from 0 minimum support. This not only incurs very high computational costs, but also the challenge of how to raise the minimum threshold without missing any top-k high utility sequences. 3) Due to the fundamental difference, incorporating the traditional closure concept into high utility sequential pattern mining makes the outcome patterns irreversibly lossy and no longer recoverable, which will be reasoned in the following chapters. Therefore, it is exceedingly challenging to address the above issues by designing a novel representation for high utility sequential patterns.

To address these research limitations and challenges, this thesis proposes a high utility sequential pattern mining framework, and proposes both a threshold-based and top-k-based mining algorithm. Furthermore, a compact and lossless representation of utility-based sequence is presented, and an efficient algorithm is provided to mine such kind of patterns.

Chapter 2 thoroughly reviews the related works in the frequent sequential pattern mining and high utility itemset/sequence mining.

Chapter 3 incorporates utility into sequential pattern mining, and a generic framework for high utility sequence mining is defined. Two efficient algorithms, namely USpan and USpan+, are presented to mine for high utility sequential patterns. In USpan and USpan+, we introduce the lexicographic quantitative sequence tree to extract the complete set of high utility sequences and design concatenation mechanisms for calculating the utility of a node and its children with three effective pruning strategies.

Chapter 4 proposes a novel framework called *top-k high utility sequential pattern mining* to tackle this critical problem. Accordingly, an efficient algorithm, Top-k high Utility Sequence (TUS for short) mining, is designed to identify top-k high utility sequential patterns without minimum utility. In addition, three effective features are introduced to handle the efficiency problem, including two strategies for raising the threshold and one pruning for filtering unpromising items.

Chapter 5 proposes a novel concise framework to discover US-closed (Utility Sequence closed) high utility sequential patterns, with theoretical proof that it expresses the lossless representation of high-utility patterns. An efficient algorithm named CloUSpan is introduced to extract the US-closed patterns. Two effective strategies are used to enhance the performance of CloUSpan.

All of the algorithms are examined in both synthetic and real datasets. The performances, including the running time and memory consumption, are compared. Furthermore, the utility-based sequential patterns are compared with the patterns in the frequency/support framework. The results show that high utility sequential patterns provide insightful knowledge for users.