Faculty of Engineering and Information Technology
University of Technology, Sydney

Negative Sequential Pattern Mining

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

by

Zhigang Zheng

January 2012

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 the deepest appreciation to my supervisor, Professor Longbing Cao, for his professional guidance, persistent help and continuous support throughout my Ph.D study and research.

I would like to thank Dr. Yanchang Zhao and Professor Xiangjun Dong, for their patient guidance, and scientific advice. Without their generous support this dissertation would not have been possible.

Besides, I offer my regards and blessings to all of my co-workers at lab, and thank them for their support in my research and during the completion of this dissertation.

In addition, I would like to thank all colleagues, specially Uma Srinivasan and Sue Bird, of the CMCRC HIBIS project, for their strong support for my research by providing much domain knowledge of health insurance industry.

Last but not the least, I would like to thank my family: my wife, my parents, and my son. Without their encouragement, finishing this dissertation would be impossible; without them, nothing would have any value.

Zhigang Zheng
June 2011 @ UTS

Contents

List of Figures

List of Tables

Abstract

Sequential pattern mining provides an important way to obtain special patterns from sequence data. It produces important insights on bioinformatics data, web-logs, customer transaction data, and so on.

Different from traditional positive sequential pattern (PSP) mining, negative sequential pattern (NSP) mining takes negative itemsets into account besides positive ones. It would be more interesting in applications where non-occurring itemsets need to be considered. This thesis reports our previous and the latest research outcomes in this area. The contributions of the thesis are as following.

- A comprehensive literature review of negative frequent pattern mining is described.
- A general framework of the NSP mining is proposed. It can be used to describe the big picture of both PSP and NSP mining problems.
- Three innovative algorithms are proposed to mine NSP efficiently.
- Extensive experiments about the three algorithms on either synthetic or real-world datasets show that the proposed methods can find NSP efficiently.
- A case study describes a real-life application on customer claims analysis in health insurance industry.

Three algorithms of NSP mining are proposed in this thesis, listed as below:

(1) The first algorithm *Neg-GSP* (Zheng, Zhao, Zuo & Cao 2009) is based on a PSP mining algorithm GSP (Srikant & Agrawal 1996). Neg-GSP deals with negative problem by introducing new methods of joining and generating candidates, which borrow ideas from GSP algorithm. And also, an effective pruning method to reduce the number of candidates is proposed as well.

(2) The second one is a Genetic Algorithm based algorithm (Zheng, Zhao, Zuo & Cao 2010), which is called *GA-NSP*. It is proposed to find NSP with novel crossover and mutation operations, which are efficient at passing good genes on to next generations. An effective dynamic fitness function and a pruning method are also provided to improve performance.

(3) The third algorithm *e-NSP* (Dong, Zheng, Cao, Zhao, Zhang, Li, Wei & Ou 2011) is based on the Set Theory. It mines NSP by only involving the identified PSP, without re-scanning the database. In this way, mining NSP does not require any additional database scans. It facilitates the existing PSP mining algorithms to mine NSP. It offers a new strategy for efficient mining of NSP.

The results of extensive experiments about the three algorithms show that they can find NSP efficiently. They have good performance compared with some other existing NSP mining algorithms, such as PNSP (Hsueh, Lin & Chen 2008).

If we compare the problem statements of the above three methods, *Neg-GSP* and *GA-NSP* share the same definitions, *e-NSP* uses stronger constraints since it requires clear boundary to follow the Set Theory. When comparing their performances, *GA-NSP* algorithm slightly outperforms *Neg-GSP* in terms of execution time, but it may miss some patterns in the complete result sets due to limitations of Genetic Algorithm. Apparently, *e-NSP* is the most efficient and effective one since it does not need to scan datasets to calculate the support of NSP. Although adding stronger constraints on e-NSP makes the search space much smaller than what it is under the normal definitions, it is still very practicable while being used in some real-life applications.

Following that, NSP mining case studies coming from health insurance industry are introduced. Based on real-life customer claims datasets, we use the proposed NSP mining methods to find PSP and NSP on solving two business issues, one is in ancillary service over-service analysis, another is fraud claim detection. Both of the two case studies demonstrate the benefits gained from mining NSP.