

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

Certificate	i
Acknowledgment	ii
List of Figures	viii
List of Tables	x
Abstract	xii
Chapter 1 Introduction	1
1.1 Background	1
1.2 Research Issues & Significance	3
1.3 The Profile of Research Work	6
1.3.1 Research Introduction	6
1.3.2 Research Problem Statement	7
1.3.3 NSP Mining Methodologies	7
1.3.4 Experiments and Evaluation	8
1.3.5 Case Studies	8
1.4 Main Research Objectives	9
1.4.1 Improving PSP Mining Algorithm	9
1.4.2 Mining NSP Based on Genetic Algorithms	10
1.4.3 Mining NSP by Deducible Theories	11
1.5 Research Contributions	11
1.6 Thesis Organization	12
Chapter 2 Related Work	13
2.1 Positive/Negative Association Rule Mining	14

2.1.1	(Positive) Association Rule Mining	14
2.1.2	Negative Association Rule Mining	17
2.2	Sequential Pattern Mining	20
2.2.1	Sequential Pattern Mining Algorithms	21
2.2.2	Comparison & Computational Complexity Analysis	24
2.3	Negative Sequential Pattern Mining	26
2.3.1	PSP Mining Algorithms in NSP Mining Problem	26
2.3.2	State-of-the-art Algorithms	26
2.4	Summary and Conclusion	28
 Chapter 3 Problem Statement		31
3.1	Basic Definitions	32
3.1.1	Positive/Negative Sequence	32
3.1.2	Data Sequence & Candidate Sequence	35
3.2	Constraints of Positive/Negative Candidates	35
3.3	Reconstruct Data Sequence	37
3.4	Supporting	39
3.4.1	Basic Operation	39
3.4.2	Criteria of Supporting	40
3.4.3	Properties of Negative Supporting	43
3.4.4	Positive/Negative Sequential Pattern	44
3.5	A Framework of NSP Mining	45
3.5.1	NSP Mining Problem	47
3.5.2	PSP Mining Problem	48
3.6	NSP Mining Problem in the Thesis	49
3.6.1	Constraints on Interesting NSP	49
3.6.2	Criteria of Negative Supporting	50
3.7	Conclusions	51
 Chapter 4 Neg-GSP Algorithm		53
4.1	GSP Algorithm	53
4.1.1	General Description of GSP	53

4.1.2	Generating Candidates	54
4.1.3	Pruning Candidates	54
4.1.4	Counting Candidates	55
4.1.5	Procedure of GSP	55
4.1.6	Improving GSP to Find NSP	56
4.2	Process of Neg-GSP	57
4.3	Neg-GSP Algorithm	59
4.3.1	Joining to Generate Candidates	59
4.3.2	Pruning Invalid Candidates	60
4.3.3	Generating Seed Set for Next Pass	61
4.3.4	Algorithm Description	61
4.3.5	Computational Complexity Analysis	62
4.4	Experiments	65
4.4.1	Datasets	65
4.4.2	Performance Evaluation	66
4.4.3	Comparison with PNSP Algorithm	66
4.5	Conclusions	69
 Chapter 5 Genetic Algorithm Based Algorithm: GA-NSP		71
5.1	Genetic Algorithm	71
5.1.1	Procedure of Genetic Algorithm	72
5.1.2	Encoding	74
5.1.3	Fitness Function	74
5.1.4	Selection	75
5.1.5	Crossover	75
5.1.6	Mutation	76
5.2	Genetic Algorithm Based NSP Mining	77
5.3	GA-NSP Algorithm	77
5.3.1	Encoding	78
5.3.2	Population	79
5.3.3	Selection	80
5.3.4	Crossover	80

5.3.5	Mutation	81
5.3.6	Pruning	81
5.3.7	Fitness Function	81
5.3.8	Algorithm Description	83
5.3.9	An Example of GA-NSP Algorithm	85
5.4	Experiments	90
5.4.1	Analysis of Crossover Rate	94
5.4.2	Analysis of Mutation Rate	94
5.4.3	Analysis of Decay Rate	94
5.4.4	Performance Comparison with Other Methods	94
5.5	Conclusions	99
 Chapter 6 Effective NSP (e-NSP) Mining Algorithm		 105
6.1	Problem Statement	106
6.1.1	Related Definitions	106
6.1.2	Constraints of Negative Candidates	108
6.1.3	Negative Containment	109
6.1.4	Brief Introduction of Set Theory	112
6.1.5	Negative Supporting	113
6.2	e-NSP Algorithm	115
6.2.1	Negative Sequential Candidates Generation	115
6.2.2	Supports of Negative Sequences / Candidates	116
6.2.3	Negative Conversion Strategy and Proof	117
6.2.4	e-NSP Data Structure and Optimization	118
6.2.5	Pseudocode of e-NSP Algorithm	120
6.2.6	An Example	122
6.2.7	Computational Complexity Analysis	123
6.3	Experiments and Evaluation	125
6.3.1	Data Sets	125
6.3.2	Computational Cost	127
6.3.3	Dataset Characteristics Analysis	127
6.3.4	Scalability Test	129

6.3.5 Experiments Summary	131
6.4 Conclusions	131
Chapter 7 Case Studies: NSP Applications	135
7.1 Case 1: Ancillary Services Over-service Analysis	135
7.1.1 Data Preparation	136
7.1.2 PSP/NSP Mining by Neg-GSP Algorithm	137
7.1.3 Risk Scoring	138
7.2 Case 2: Fraud Claim Detection	138
7.2.1 Data Preparation	139
7.2.2 A Fraud Scenario	140
7.2.3 Fraud Claim Detection by e-NSP Algorithm	141
Chapter 8 Conclusions and Future Work	143
8.1 Conclusions	143
8.2 Future Work	145
8.2.1 Future Work of Current Topics	145
8.2.2 Order-First and Negative-First Problems	145
8.2.3 Negative Sequential Pattern Classification	145
8.2.4 Post Mining of Negative Sequential Pattern	146
Chapter A Appendix: List of Publications	147
Chapter B Appendix: List of Symbols	149
Bibliography	152

List of Figures

1.1	The Profile of Research Work	6
2.1	Systematization of Association Rule Mining	17
2.2	Association Rule And Sequential Pattern Mining Algorithms .	21
3.1	A Framework of Negative Sequential Pattern Mining	46
4.1	The Process Flow of Neg-GSP	58
4.2	Neg-GSP: An Example	63
4.3	Neg-GSP: Execution Time	67
4.4	Neg-GSP: Patterns Counts	68
4.5	Neg-GSP: Comparison with PNSP Algorithm	70
5.1	GA-NSP Algorithm: Process Flow	78
5.2	Experiments: Different Crossover Rates On DS1	92
5.3	Experiments: Different Crossover Rates On DS2	93
5.4	Experiments: Different Mutation Rates On DS1	95
5.5	Experiments: Different Mutation Rates On DS2	96
5.6	Experiments: Different Decay Rates On DS1	97
5.7	Experiments: Different Decay Rates On DS2	98
5.8	GA-NSP Algorithm: Execution Time Comparison On DS1 . .	100
5.9	GA-NSP Algorithm: Execution Time Comparison On DS2 . .	101
5.10	GA-NSP Algorithm: Execution Time Comparison On DS3 . .	102
5.11	GA-NSP Algorithm: Execution Time Comparison On DS4 . .	103

6.1	e-NSP Algorithm: the Intersection of $\{< a >\}$ and $\{< b >\}$. .	114
6.2	e-NSP Algorithm: the Meaning of $sup(< a \neg b c \neg d e \neg f >)$.	114
6.3	e-NSP Algorithm: Framework	115
6.4	e-NSP Algorithm: Execution Time Comparison	128
6.5	e-NSP Algorithm: Maximum Length and Patterns Counts . .	133
6.6	Dataset Characteristics Analysis	134
6.7	e-NSP Algorithm: Scalability Test	134
7.1	Case Study: Example of Customers Risk Scoring	139
7.2	Samples of Customer Claims Dataset	139

List of Tables

1.1	A Transactional Data Table	2
1.2	Synthetic Dataset Factors	9
3.1	Example of Maximum Equivalent Sequence	38
3.2	Examples of Sequence Containing	39
3.3	Examples of Sequence Absolutely Containing	39
3.4	Examples of the Three Criteria of Supporting	42
3.5	Apriori-property in a Positive-first Problem	44
3.6	Neg-GSP: Example of Negative Supporting	51
4.1	Examples of base-support and support	57
4.2	An Example of Joining	59
4.3	Neg-GSP: Features of Synthetic Datasets	66
5.1	Genetic Algorithm: Examples of Encoding	74
5.2	Genetic Algorithm: Single Point Crossover	76
5.3	Genetic Algorithm: Two Point Crossover	76
5.4	Genetic Algorithm: Cut and Splice Crossover	76
5.5	Genetic Algorithm: Mutation	77
5.6	GA-NSP Algorithm: Encoding	78
5.7	GA-NSP Algorithm: Crossover	80
5.8	GA-NSP Algorithm: Crossover at Head/End	81
5.9	GA-NSP Algorithm: Features of Synthetic Datasets	91
6.1	e-NSP: Data Structure and An Example	119

6.2	e-NSP: Data Set for Example	122
6.3	e-NSP: Example Results - Positive Patterns	123
6.4	e-NSP: Example Results - NSC and Supports	124
6.5	e-NSP: Experiments of Dataset Factors Analysis	130

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.