

The Discovery and Retrieval of Temporal Rules in Interval Sequence Data

by

Edi Winarko, *B.Sc., M.Sc.*
School of Informatics and Engineering,
Faculty of Science and Engineering

March 19, 2007

A thesis presented to the
Flinders University
in total fulfillment of the requirements for the degree of
Doctor of Philosophy

Adelaide, South Australia, 2007
© (Edi Winarko, 2007)

Abstract

Data mining is increasingly becoming important tool in extracting interesting knowledge from large databases. Many industries are now using data mining tools for analysing their large collections of databases and making business decisions. Many data mining problems involve temporal aspects, with examples ranging from engineering to scientific research, finance and medicine. Temporal data mining is an extension of data mining which deals with temporal data. Mining temporal data poses more challenges than mining static data. While the analysis of static data sets often comes down to the question of data items, with temporal data there are many additional possible relations.

One of the tasks in temporal data mining is the pattern discovery task, whose objective is to discover time-dependent correlations, patterns or rules between events in large volumes of data. To date, most temporal pattern discovery research has focused on events existing at a point in time rather than over a temporal interval. In comparison to static rules, mining with respect to time points provides semantically richer rules. However, accommodating temporal intervals offers rules that are richer still.

This thesis addresses several issues related to the pattern discovery from interval sequence data. Despite its importance, this area of research has received relatively little attention and there are still many issues that need to be addressed. Three main issues that this thesis considers include the definition of what constitutes an interesting pattern in interval sequence data, the efficient mining for patterns in the data, and the identification of interesting patterns from a large

number of discovered patterns.

In order to deal with these issues, this thesis formulates the problem of discovering rules, which we term *richer temporal association rules*, from interval sequence databases. Furthermore, this thesis develops an efficient algorithm, **ARMADA**, for discovering richer temporal association rules. The algorithm does not require candidate generation. It utilizes a simple index, and only requires at most two database scans. In this thesis, a retrieval system is proposed to facilitate the selection of interesting rules from a set of discovered richer temporal association rules. To this end, a high-level query language specification, **TAR-QL**, is proposed to specify the criteria of the rules to be retrieved from the rule sets. Three low-level methods are developed to evaluate queries involving rule format conditions. In order to improve the performance of the methods, signature file based indexes are proposed. In addition, this thesis proposes the discovery of *inter-transaction relative temporal association rules* from event sequence databases.

Certification

I certify that this thesis does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any university; and that to the best of my knowledge and belief it does not contain any material previously published or written by another person except where due reference is made in the text.

As requested under Clause 14 of Appendix D of the *Flinders University Research Higher Degree Student Information Manual* I hereby agree to waive the conditions referred to in Clause 13(b) and (c), and thus

- Flinders University may lend this thesis to other institutions or individuals for the purpose of scholarly research;
- Flinders University may reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Signed

Dated

Edi Winarko

Use of this thesis

I hereby acknowledge that I have been given access to the thesis for consultation only and that no part will be published or paraphrased without the prior consent of the author and that the author's intellectual property rights will be respected.

Acknowledgements

My great gratitude goes to my supervisor Professor John F. Roddick who has helped me a lot, not only giving me valuable resources, corrections and room for expressing my idea but also encouragement.

I am particularly grateful to the Australian Government for awarding me an AusAid Scholarship, which gave me an opportunity to undertake my study at the Flinders University. I also gratefully acknowledge Prof. Subanar at Gadjah Mada University, for the full support and opportunity to pursue my study program.

My heartfelt thanks go to the special persons in my family, Tutik and Triska, who have contributed a great deal of time with sharing, understanding and support during my candidature and especially during the difficult time. Thanks to my parents and parents-in-law for their love and support.

Then, there are many friends who have made my study so pleasant, fruitful, and so rich in experience. In particular, the member of the Knowledge Discovery and Intelligent System Group: Aaron, Denise, Carl, Sally, Anna, Da, and Ping. You all contributed in one or the other way, through technical discussions, sharing experience, or ‘simply’ by your company and your friendship.

Edi Winarko

March 2007

Adelaide.

Contents

Abstract	ii
Certification	iv
Use of this Thesis	v
Acknowledgements	vi
Table of Contents	vii
List of Figures	xi
List of Tables	xiii
List of Algorithms	xiv
1 Introduction	1
1.1 Temporal Data Mining	2
1.2 Research Objectives	8
1.2.1 Discovery of Point Temporal Rules	8
1.2.2 Discovery of Interval Temporal Rules	9
1.2.3 Retrieval of Discovered Temporal Rules	10
1.3 Organization of the Thesis	10
2 Review of Mining Time Point Patterns	12
2.1 Mining Temporal Association Rules	13
2.1.1 Taxonomy of Temporal Association Rules	15

2.1.2	Interval Association Rules	20
2.1.3	Cyclic Association Rules	23
2.1.4	Calendric Association Rules	25
2.1.5	Temporal Predicate Association Rules	28
2.2	Mining Sequential Patterns	29
2.2.1	Problem Definition	31
2.2.2	Sequential Pattern Mining Algorithms and Extensions	33
2.2.3	Apriori-Based using Horizontal Data Format	37
2.2.4	Apriori-Based using Vertical Data Format	41
2.2.5	Pattern-Growth using Database Projection	45
2.2.6	Pattern-Growth using Database Indexing	49
2.3	Mining Episodes and Periodic Patterns	52
2.3.1	Preliminary Definitions	53
2.3.2	Two Basic Approaches for Mining Episodes	54
2.3.3	Extensions of Episode Model	57
2.3.4	Mining Periodic Patterns	59
2.4	Summary	61
3	Mining Relative Temporal Association Rules	62
3.1	Model Description	63
3.2	Mining Relative Temporal Association Rules	66
3.2.1	Finding Frequent Relative Itemset	67
3.2.2	Generating Relative Temporal Association Rules	70
3.3	Evaluation	71
3.4	Summary	75
4	Review of Mining Time Interval Patterns	76
4.1	Sources of Interval Data	77
4.2	Time Interval Operators	82
4.3	Mining Patterns from a Long Sequence of Intervals	86
4.4	Mining Patterns from Interval Sequence Databases	89
4.5	Summary	93

5	ARMADA: Mining Richer TAR	95
5.1	Problem Statement	96
5.2	ARMADA - Mining Richer Temporal Association Rules	102
5.2.1	Discovering Frequent Temporal Patterns	102
5.2.2	Handling Large Databases	108
5.2.3	Generating Temporal Association Rules	108
5.3	Maximum Gap Time Constraint	109
5.4	Evaluation	111
5.4.1	Experiments on Synthetic Data	111
5.4.2	Experiments on Real Data	115
5.5	Summary	118
6	Review of Set and Sequence Retrieval	119
6.1	Types of Queries in Set Retrieval	120
6.2	Set Retrieval using Inverted Files	122
6.3	Set Retrieval using Signature Files	123
6.3.1	Methods for Generating Set Signatures	124
6.3.2	Processing Set Queries	126
6.4	False Drop Probability	127
6.5	Signature File Organisation	128
6.5.1	Sequential Organisation	129
6.5.2	Bit-Sliced Organisation	129
6.5.3	Hierarchical Organisation	132
6.5.4	Partitioned Organisation	133
6.6	Sequential Pattern Retrieval	134
6.6.1	Representing Sequential Patterns as Sets	134
6.6.2	Partitioning Equivalent Sets	136
6.7	Summary	137

7	Retrieval of Discovered Temporal Rules	138
7.1	Importance of Post-Processing Discovered Rules	138
7.2	Definitions	141
7.3	Framework of the Post-processing	141
7.4	Types of Queries on Temporal Rules	144
7.5	Constructing Signature Files	149
7.5.1	Converting Temporal Patterns to Equivalent Sets	149
7.5.2	Converting Equivalent Sets to Signatures	151
7.6	Processing of Queries using Signature Files	153
7.6.1	Subpattern Queries	153
7.6.2	Superpattern Queries	154
7.6.3	Equality Queries	156
7.6.4	Temporal Pattern Similarity	157
7.7	Experiments	159
7.7.1	Effect of Signature Size on the Number of False Drops	160
7.7.2	Effect of Signature Size on Query Processing Time	161
7.7.3	Effect of Database Size on the Query Processing Time	163
7.8	Summary	163
8	Conclusions	165
8.1	Contributions	166
8.2	Future Research Directions	168
	Appendices	170
A	Publications Resulting from This Thesis	170
B	Rule Discovery System	172
B.1	ARMADA	172
B.2	Interval Data Generator	173
C	TAR-QL Query Language	176
	Bibliography	182

List of Figures

1.1	Knowledge discovery process	3
2.1	Example database	20
2.2	A database shown as a set of transactions and a set of sequences .	32
2.3	A set of sequential patterns	33
2.4	Classification of sequential pattern mining algorithms	35
2.5	Vertical data format used in SPADE	42
2.6	Vertical bitmap representation of the example database	45
2.7	Database projected on frequent 1-sequences	47
2.8	Example of index sets	50
3.1	Example database	64
3.2	Output of each phase of the algorithm	68
3.3	Effect of decreasing minimum support on the processing time . . .	73
3.4	Effect of decreasing minimum support on the number of patterns	73
3.5	Effect of increasing database size on the processing time	74
3.6	Processing time required by each phase of the algorithm	74
4.1	Transform time series into a sequence of intervals	78
4.2	Seven basic shapes	79
4.3	Example of partitioning time series using supervised approach . .	80
4.4	Example of partitioning time series using unsupervised approach .	81
4.5	Allen's interval relationships	83
4.6	Three relationships resulting from two almost equal intervals . . .	85
4.7	Example of temporal patterns	87

4.8	Example of an interval sequence	90
4.9	An interval sequence database and fragment of mining process	91
4.10	Five relations used in arrangements	92
4.11	Two arrangements of size 3	93
5.1	Seven relations used in normalized temporal patterns	97
5.2	Three normalized temporal patterns	98
5.3	Example database consisting of clinical records	100
5.4	Example of index sets	105
5.5	Determining gap and maximum gap in the state sequence	110
5.6	Effect of decreasing minimum support	113
5.7	Effect of increasing maximum gap	113
5.8	Effect of increasing number of states	114
5.9	Effect of increasing database size	115
5.10	Effect of decreasing minimum support (ASL database)	117
6.1	Example of market basket database	121
6.2	Inverted file of the market basket database	122
6.3	Generating signature using superimposed coding	125
6.4	SSF file organisation	129
6.5	BSSF file organisation	130
6.6	Structure of S-tree (K=4 and k=2)	132
7.1	Example of temporal patterns and temporal rules	142
7.2	Classical vs proposed approaches	143
7.3	The post-processing framework	143
7.4	Processing temporal pattern query using signature files	154
7.5	Effect of signature size on the number of false drops	161
7.6	Effect of signature size on query processing time	162
7.7	Effect of database size on query processing time	162
B.1	Screen shot of the ARMADA interface	173
B.2	Screen shot of the data generator interface	174

List of Tables

2.1	Temporal association rule classification	16
2.2	Temporal association rule algorithms and models	18
3.1	Parameters	72
3.2	Eight datasets for the experiments	72
5.1	A set of frequent temporal patterns	101
5.2	List of field names and labels	116
6.1	Symbols	128
7.1	Equivalent sets and signatures of temporal patterns	151
7.2	Parameters	159
B.1	Parameters	173

List of Algorithms

2.1	Pseudo code for generating frequent itemsets (in their lifespan) . . .	23
2.2	Pseudo code for generating precise-match frequent itemsets	27
2.3	Pseudo code of the GSP Algorithm	38
3.1	Pseudo code for Generating frequent relative itemsets	70
5.1	Pseudo code of ARMADA	103
5.2	Pseudo code for constructing an index set	104
5.3	Pseudo code for mining an index set	105
5.4	Pseudo code of ARMADA for processing large databases	109
5.5	Pseudo code for generating richer temporal association rules . . .	109
7.1	Constructing a signature file of temporal patterns	153
7.2	Pseudo code of evaluateSubPattern using SSF	155
7.3	Pseudo code of evaluateSubPattern using BSSF	155
7.4	Pseudo code of evaluateSuperPattern using BSSF	156