

The Discovery of Interacting Episodes and Temporal Rule Determination in Sequential Pattern Mining

by

Carl Howard Mooney, *B.Sc.(Comp.Sc)(Hons)*
School of Informatics and Engineering,
Faculty of Science and Engineering

December 1, 2006

A thesis presented to the
Flinders University of South Australia
in total fulfilment of the requirements for the degree of
Doctor of Philosophy

Abstract

The reason for data mining is to generate rules that can be used as the basis for making decisions. One such area is sequence mining which, in terms of transactional datasets, can be stated as the discovery of *inter-transaction* associations or associations between different transactions. The data used for sequence mining is not limited to data stored in overtly temporal or longitudinally maintained datasets and in such domains data can be viewed as a series of events, or episodes, occurring at specific times. The problem thus becomes a search for collections of events that occur frequently together.

While the mining of frequent episodes is an important capability, the manner in which such episodes interact can provide further useful knowledge in the search for a description of the behaviour of a phenomenon but as yet has received little investigation. Moreover, while many sequences are associated with absolute time values, most sequence mining routines treat time in a relative sense, returning only patterns that can be described in terms of Allen-style relationships (or simpler), ie. nothing about the relative pace of occurrence. They thus produce rules with a more limited expressive power. Up to this point in time temporal interval patterns have been based on the endpoints of the intervals, however in many cases the ‘natural’ point of reference is the midpoint of an interval and it is therefore appropriate to develop a mechanism for reasoning between intervals when midpoint information is known.

This thesis presents a method for discovering interacting episodes from temporal sequences and the analysis of them using temporal patterns. The mining can be conducted both with and without the mechanism for handling the pace of events and the analysis is conducted using both the traditional interval algebras and a midpoint algebra presented in this thesis.

The visualisation of rules in data mining is a large and dynamic field in its own right and although there has been a great deal of research in the visualisation of associations, there has been little in the area of sequence or episodic mining. Add to this the emerging field of mining stream data and there is a need to pursue methods and structures for such visualisations, and as such this thesis also contributes toward research in this important area of visualisation.

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

Carl Howard Mooney

Acknowledgements

I would like to thank my supervisor John Roddick for his support and guidance throughout this journey that has ultimately become this dissertation. John your patience, enthusiasm and work ethic are inspirational. To those colleagues and friends who have shared their insights into this work; Denise de Vries, Aaron Ceglar, Anna Shillabeer, Paul Calder, Darius Pfitzner, and Eddie Winarko, thank you all. I would like to add a special thank you to both Anna and Denise for reading the drafts and making critical comment.

To the School of Informatics and Engineering and the computer support staff who have provided me with materials and support for my studies I thank you. Rino and Michael you are great in a crisis and you always find time to help. Thanks to Janet for the casual chats over lunch and coffee and to Graham who didn't mind interruptions at any time. This work was funded by an APA scholarship and I would like to thank both the Commonwealth Government for its financial assistance as well as the Flinders University for the casual work that was offered to me.

To be able to complete this thesis has been a journey of nearly a decade and I owe a great deal to Tony Mykolajenko and his staff for both putting me on a new path in life and also for the belief that I could make it.

To my family who has supported me throughout my life in all my endeavours and especially through this time, my heartfelt thanks go to Mum, Dad, Ian, Heather and Grace and to Grandma Jane who will live to see me finally complete my studies. Finally I would like to say thank you to Louise for the understanding, support and encouragement she has given me for the last year, it has been quite a strain on the relationship, but we have prevailed.

Carl Howard Mooney
December 2006
Adelaide.

Contents

Abstract	ii
Certification	iii
Acknowledgements	iv
Table of Contents	v
List of Figures	xi
List of Tables	xiii
List of Algorithms	xv
1 Introduction	1
2 Sequential Pattern Mining	4
2.1 The Sequential Pattern Mining Problem	5
2.2 Types of Constraints	6
2.3 Counting Techniques	9
2.4 Apriori-based Algorithms	10
2.4.1 Problem Statement and Notation	11
2.4.2 Horizontal Database Format	12
2.4.3 Horizontal Database Format Algorithms	12
2.4.4 Vertical Database Format	20
2.4.5 Vertical Database Format Algorithms	21
2.4.6 Summary of Apriori-based Algorithms	26

2.5	Projection-based Algorithms	27
2.5.1	Pattern Growth	27
2.5.2	Summary of Pattern Growth Algorithms	31
2.6	Temporal Sequences	31
2.6.1	Problem Statement and Notation for Episode and Event-based Algorithms	31
2.6.2	WINEPI	33
2.6.3	PROWL	35
2.6.4	Event-Oriented Patterns	36
2.6.5	Pattern Directed Mining	37
2.6.6	Summary of Temporal Sequence Algorithms	38
2.7	Extensions	38
2.7.1	Closed Frequent Patterns	38
2.7.2	Hybrid Methods	40
2.7.3	Approximate Methods	40
2.7.4	Parallel Algorithms	41
2.7.5	Other Methods	42
2.7.6	Time Series Mining	42
2.8	Incremental Mining Algorithms	43
2.8.1	Incremental Discovery of Sequential Patterns	43
2.8.2	ISM: Interactive Sequence Mining	43
2.8.3	ISE: Incremental Sequence Extraction	43
2.8.4	IUS/DUS: Incrementally/Decreasingly Updating Sequences	44
2.8.5	GSP+ and MFS+	44
2.8.6	IncSpan: Incremental Sequential Pattern mining	45
2.8.7	Improvements of IncSpan	46
2.9	Areas of related research	47
2.9.1	Streaming Data	47
2.9.2	String Matching and Searching	47
2.10	Rule Inference	48
2.11	Discussion	48

3	Temporal Logic	50
3.1	Temporal Logic Models	50
3.2	Allen's Interval Algebra	52
3.3	Freksa's Semi-Intervals	53
3.4	Extensions	56
3.4.1	Fuzzy Time Intervals	56
3.4.2	Fuzzy Interval Algebra	57
3.5	Discussion	58
4	Temporal Intervals with Midpoints	59
4.1	Midpoints in Relation to Existing Models	60
4.2	Linear Temporal Sequences	60
4.2.1	Implied Order and Implied Simultaneity	62
4.3	Midpoint Preliminaries	65
4.4	Equal Length Intervals	67
4.5	Variable Length Intervals	68
4.5.1	Naming Conventions	68
4.5.2	The Set of Variable-Length Midpoint Interval Relationships	71
4.6	Transformations	73
4.6.1	Conceptual Hierarchies	73
4.7	Iconic Representation	77
4.7.1	Extensions to Freksa's Iconic Representations	77
4.8	Discussion	80
5	Mining Interacting Episodes	81
5.1	The Framework	82
5.1.1	Data considerations	82
5.2	Frequent Episode Discovery	83
5.2.1	Problem Definition	83
5.2.2	Algorithmic Considerations	84
5.3	Interacting Episode Discovery	86
5.3.1	Problem Definition	86

<i>CONTENTS</i>	viii
5.3.2 An Algorithm for Interaction Discovery	89
5.3.3 Interaction Classes	91
5.3.4 Common Tokens	96
5.3.5 Interruptions at different locations	96
5.4 Discussion	97
6 Timing Considerations	99
6.1 Timing Marks	100
6.1.1 Timing Marks as Tokens	101
6.1.2 Timing Marks Added as Delimiters	101
6.1.3 Timing Marks as Absolute Time	101
6.1.4 The Value of Timing Marks	101
6.2 Rule semantics	102
6.3 Algorithmic Considerations	103
6.3.1 Timing Mark Pruning	104
6.4 Discussion	104
7 Transitive Relationships	106
7.1 The Structure of Transitive Relationships	106
7.1.1 Terminology	107
7.2 Transitivity and Known Lengths	108
7.2.1 Transitivity for Variable-Length Intervals	108
7.2.2 Transitivity for Equal-Length Intervals	113
7.3 Rule Inference using Transitive Relationships	115
7.3.1 Rule Inference: An Overview	115
7.3.2 Limiting the Number of Itemsets, Sequences and Rules	115
7.3.3 Outcome Discovery	116
7.3.4 Presentation of Outcomes	116
7.3.5 Implications Arising from Interacting Episodes	117
7.4 Visualising the Outcomes from Transitive Relationships	118
7.5 Discussion	119

8	Conclusions and Future Research	121
8.1	Mining Heuristics and Datasets	121
8.2	Transitive Relationships	121
8.3	Development of Visualisation Tools	122
8.4	Application Areas Applicable to this Approach	122
8.5	Conclusion	122
A	Transitivity Tables	123
A.1	Allen's Transitivity Table	124
A.2	Equal-Length Interval Midpoint Transitivity Table	126
A.3	Variable-Length Interval Midpoint Transitivity Table	127
A.3.1	<i>Before</i> ($<$) to <i>LargeLargeOverlap</i> (llo)	127
A.3.2	<i>is-FinishedSmall-by</i> (fsi) to <i>LastContainsFirst-of</i> (fdli)	141
A.3.3	<i>is-StartedSmall-by</i> (ssi) to <i>StartsSmall</i> (ss)	152
A.3.4	<i>FirstDuringLast</i> (fdl) to <i>FinishesSmall</i> (fs)	158
A.3.5	<i>is-LargeLargeOverlapped-by</i> (lloi) to <i>After</i> ($>$)	169
A.3.6	Symmetric Verification Table	184
B	Software Application	185
B.1	Sequence Mining	186
B.1.1	Sequence Mining Interface	186
B.1.2	Viewing the Output	187
B.1.3	Controls for Mining	189
B.1.4	Execution Information	191
B.2	Transitive Relationship Discovery	192
B.2.1	Transitive Relationships Interface	192
B.3	Experimental Results	194
B.3.1	Mining without Timing Marks	194
B.3.2	Mining with Timing Marks	196

C Algorithms	198
C.1 Interacting Episodes	198
C.1.1 Frequent Episodes	198
C.1.2 Frequent Interactions	200
C.2 Timing Marks	202
Bibliography	204

List of Figures

1.1	Structural Domain of this Thesis.	2
2.1	A comparison of different counting methods.	10
2.2	The prefix-tree of <i>PSP</i> and the hash-tree of <i>GSP</i>	16
2.3	Hackle-tree for an RE-constraint.	18
2.4	A Prefix Tree of MSPS.	20
2.5	A length-decreasing support constraint.	30
2.6	An example event sequence.	32
2.7	Depiction of a serial and parallel episode.	33
2.8	The PROWL process.	36
2.9	Sequence fragments in an event sequence.	37
2.10	Two examples of SP Trees.	38
3.1	Freksa’s iconic representation of Allen’s relations.	55
3.2	Fine-grained (Allen) reasoning using Freksa’s coarse reasoning methods.	55
3.3	A depiction of a crisp and fuzzy interval.	57
3.4	A depiction of an IA^{fuz} relation.	58
4.1	Models of temporal interval relations.	61
4.2	Data stream generated from n independent sensors.	63
4.3	Moving window of potentially simultaneous tokens.	63
4.4	Moving window (w) over tokens with larger than required granularity.	64
4.5	Allen’s constraint propagation algorithm.	66
4.6	The sections of an interval for the <i>overlap</i> relationships.	69
4.7	Depiction of a <i>SmallLargeOverlap</i> (slo) relationship.	69

4.8	Depiction of a <i>LastDuringFirst</i> (ldf) relationship.	70
4.9	The sections of an interval for the <i>starts</i> relationships.	70
4.10	Depiction of a <i>StartsMedium</i> (sm) relationship.	71
4.11	Hierarchical structure of the <i>overlap</i> relation.	74
4.12	Hierarchical structure of the <i>during</i> relation.	75
4.13	Hierarchical structure of the <i>starts</i> and <i>finishes</i> relations.	76
4.14	Iconic representations of the Allen and VLMI relationships.	77
5.1	Possible positions for a sub-episode, e_2 , with a sub-episode, e_1	87
5.2	Section of an input string showing varying window widths.	88
6.1	Possible structure of data when <i>timing marks</i> are included.	101
7.1	Depiction of an additive and union transitive relationship.	107
7.2	Expressive power example for two meeting intervals.	109
7.3	Expressive power example for two overlapping intervals.	111
7.4	Allen <i>outcomes</i> for $A \rightarrow C$ when $A \xrightarrow{o} B \xrightarrow{o} C$	113
7.5	Equal-Length <i>outcomes</i> for $A \rightarrow C$ when $A \xrightarrow{so} B \xrightarrow{lo} C$	114
7.6	Equal-Length <i>outcomes</i> for $A \rightarrow C$ when $A \xrightarrow{lo} B \xrightarrow{lo} C$	114
7.7	Best possible Equal-Length <i>outcomes</i> for $A \rightarrow C$	115
7.8	Iconic representations of the <i>outcomes</i> for Example 7.1.	119
A.1	VLMI transitivity table highlighting the symmetry.	184
B.1	Screenshot: The $INTEM_{TM}$ application.	187
B.2	Screenshot: Viewing and Tree Panes – $INTEM_{TM}$	188
B.3	Screenshot: Control Pane for sequence mining – $INTEM_{TM}$	189
B.4	Screenshot: Mining run and Interaction discovery – $INTEM_{TM}$	191
B.5	Screenshot: Transitive Relationships interface.	192
B.6	Screenshot: All possible outputs – Transitive Relationships interface.	193
B.7	Processing time and frequent episode production.	195
B.8	Execution time and frequent interaction production.	196
B.9	Timing Marks – Processing time and frequent episode production.	197

List of Tables

2.1	Horizontal Database Format.	12
2.2	Large Itemsets.	13
2.3	A transformed database and mappings.	13
2.4	Vertical Database Format.	21
2.5	Computing Support using temporal id-list joins.	22
2.6	SPAM data representation.	24
2.7	A summary of apriori-based algorithms.	26
2.8	A sequence database.	27
2.9	A summary of pattern growth algorithms.	31
2.10	PROWL event sequence layout.	35
2.11	A summary of temporal sequence algorithms.	38
3.1	Allen’s thirteen temporal relationships.	52
3.2	Freksa’s eleven semi-interval relationships.	54
4.1	Vilains five point-interval temporal relationships.	64
4.2	Equal-length interval-interval relationships with midpoints.	67
4.3	The 49 VLMI relationships.	72
4.4	The major divisions of the VLMI relationships.	78
5.1	Constraint pattern propagation rules.	90
5.2	Possible sub-episode configurations.	96
7.1	Extract from the VLMI transitivity table.	109
7.2	Extract from the Allen transitivity table.	109
7.3	Comparison of <i>outcomes</i> for Allen and midpoint transitive relationships.	111

7.4	Output comparisons for Allen and Midpoint transitive relationships. . .	112
A.1	Allen's transitivity table.	124
A.2	ELMI transitivity table.	126
A.3	VLMI transitivity table – <i>Before</i> (<) to <i>LargeLargeOverlap</i> (llo).	127
A.4	VLMI transitivity table – <i>is-FinishedSmall-by</i> (fsi) to <i>LastContainsFirst-of</i> (fdli).	141
A.5	VLMI transitivity table – <i>is-StartedSmall-by</i> (ssi) to <i>StartsSmall</i> (ss). . .	152
A.6	VLMI transitivity table – <i>FirstDuringLast</i> (fdl) to <i>FinishesSmall</i> (fs). . .	158
A.7	VLMI transitivity table – <i>is-LargeLargeOverlapped-by</i> (lloi) to <i>After</i> (>). . .	169
B.1	Non-timing mark experimental file specifications.	195
B.2	Timing mark experimental file specifications.	196

List of Algorithms

C.1 Find Frequent Closed Episodes.	199
C.2 Find Relationships.	200
C.3 Find Any Relationships.	201
C.4 Prune Candidate Interactions.	201
C.5 Constrain using timing marks.	202
C.6 Remove timing marks.	203