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

Adelaide, South Australia, 2006

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

List of Figures

List of Tables

List of Algorithms