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

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

List of Figures

List of Tables

List of Algorithms