Guided Association Mining through Dynamic Constraint Refinement

by Aaron John Ceglar, *BIT.(Hons)*

School of Informatics,
Faculty of Science and Engineering

March 28, 2005

Flinders University of South Australia

in total fulfillment of the requirements for the degree of
Doctor of Philosophy

Adelaide, South Australia, 2005

Abstract

Association mining, the discovery of *interesting* inferences from within a dataset, is ultimately subjective as only the user can assess the practical usefulness of an inference. To this effect, an association mining system harnesses the user's perceptual capabilities and the computer's processing power to improve the quality of a set of inferences. Although current association mining systems tightly involve the user within the pre-processing and presentation stages, the analysis stage of the association mining process remains relatively autonomous and opaque. This lack of user involvement constrains domain space exploration and subsequent inference derivation, potentially reducing inference quality, due to the lack of user-computer synergy.

The theory of guided association mining and its realisation represents a timely and logical step in the progression of association mining research. Early research focused upon algorithmic efficiency, addressing issues such as I/O reduction and scalability, however this seems to have reached a point of diminishing return. The research focus has therefore shifted to improving result quality, or improving inference interest, rather than the speed at which the results are generated, including areas of research such as measures of interestingness and semantic inclusion. However, these areas of research which attempt to incorporate domain knowledge within analysis, fall short of providing user-computer synergy as the specified constraints are statically included within an automated process. Given this static constraint inclusion, the derivation of quality inferences often requires an iterative analysis process, whereby a set of quality inferences is converged upon through iterative constraint refinement.

This thesis argues that by maintaining the user-computer synergy during analysis, the quality of discovered inferences can be improved. This is achieved by opening the opaque "black box" analysis process and providing functionality through which the user can interact, and subsequently guide, domain space exploration. Thus by enabling the user to dynamically focus exploration upon concept areas of specific interest, the quality of the derived inferences will improve.

This thesis addresses the next step in providing *analysis synergy* by enabling the user to dynamically refine constraints during analysis instead of between analysis iterations. To this end a guided mining architecture is proposed that merges the currently accepted knowledge discovery architecture with the model-view-controller architecture, enabling analysis synergy through the provision of a transparent and interactive analysis environment. Furthermore this thesis also makes novel contributions to the foundation fields of analysis and rule presentation, by way of an incremental closed-set association mining algorithm and an association visualisation technique that accommodates hierarchical semantics.

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

Aaron John Ceglar

Use of this thesis

Acknowledgements

This thesis was inspired by a casual conversation between two gifted scholars: John Roddick and Paul Calder, regarding the potential benefits of incorporating the user within association analysis. This field of research has proved rewarding and I believe the contributions we have made within it are useful. However, my research apprenticeship has been much more than this thesis represents, I believe my true reward is my growth as a person during this time, which I credit to the social environment in which I have been immersed.

I would first like to thank my supervisor, John Roddick for suggesting this area of research, providing invaluable insights, timely encouragement as well as guidance balanced by the freedom to express myself through this work. Most importantly however, I would like to thank John for his patience.

I would like to acknowledge the academic contributions of Paul Calder, David Powers, Scott Vallance, Denise de Vries, Carl Mooney, Darius Pfitzner and Darin Chan. Also to Annette Shillabeer and Vaughan Hobbs for reading my thesis and providing critical feedback.

From a philosophical view I am eternally grateful to Ron Porter, for challenging my philosophical beliefs. I have really enjoyed our heated conversations, you have had a significant impact upon my life.

I would like to thank my colleagues and friends: Darius Pfizner, Trent Lewis, Martin Leurssen, Scott Vallance, Darin Chan, Ron Porter, Denise de Vries, Carl Mooney and Graham Roberts for their support and conversation. I would also like to thank the technical staff Murray, Paul, Tristan, Michael and Rino for their support over the years and Flinders University for the provision of the scholarship that made all this possible.

Most importantly, I would like to thank my family. To Mama, Nathan and Alisha for being there. To Kurtis, Eryn and Corban for always being happy to see me, and to Jo for absolutely everything.

Aaron Ceglar
November 2004

Table of Contents

List of Figures

List of Tables

List of Algorithms

Notation

Set Nomenclature

C	Candidate element subset.
L	Closed lattice.
D	Data set.
E	Element set.
N	GAM model.
Q	GAM queue.
δ	Increment dataset.
I	Increment lattice.
O	Object set.
R	Rule set (inference set).
CL	Set of closed elementsets.
V	Set of valid elementsets.
U	Universal association mining context.

Miscellaneous Nomenclature

$\wedge$	And.
$\vee$	Or.
$\supset$	Superset.
$\supseteq$	Superset or equal.
$\subset$	Subset.
$\subseteq$	Subset or equal.
$\cap$	Intersection.
$\cup$	Union.
$\in$	Exists in.
γ	Confidence.
$minconf$	Minimum Confidence Threshold.
σ	Support.
$minsup$	Minimum Support Threshold.
κ	Elementset length.
cl	Elementset closure.
$\Rightarrow$	Infers.
$\Re$	Root Node.
$tidList$	Object identifier list. Note that *Tid* is inherited from Transaction identifier - a domain specific concept.
ω	The number of increment datasets (δ) specified in the inclusion of windowing functionality.

Acronyms

BFT	Breadth First Traversal.
DFT	Depth First Traversal.
DCR	Dynamic Constraint Refinement.
GUI	Graphical User Interface.
HCI	Human Computer Interaction.
I/O	Input / Output.
LIM	Layered Interaction Model.
MOI	Measures of Interest.
Tid	Transaction Identifier.
UCP	Upward Closure Principle.

MCL	Maintained Closed Lattice analysis algorithm.
CARV	Concentric Association Rule Visualiser.
GAM	Guided Association Mining tool.
HND	Hierarchical Non-monotonic Dynamic analysis algorithm.
HPTid	Hierarchical Prioritised Tidlist analysis algorithm.

Preface

This thesis presents a guided knowledge discovery architecture that facilitates enhanced user-computer synergy within knowledge discovery analysis by providing an interactive analysis environment. Although this architecture has generic connotations, as it is designed to be applicable to all explorative knowledge discovery tasks, the research has been undertaken in the context of association mining, effectively enabling the guidance of association analysis through dynamic constraint refinement. To this end, the thesis builds towards the proposed guided architecture through significant research into the critical foundation areas of analysis and presentation, which has resulted in additional contributions to these areas. The thesis is presented in five logical parts: 1)introduction, 2) association mining, 3) rule presentation, 4) guided association mining and 5) conclusion. Furthermore, for example purposes the thesis uses the simple concept hierarchy presented in Figure 1.

Part I introduces the thesis by providing the problem statement and thesis hypothesis, which is supported by recent statements by prominent researchers regarding the need for further research into interactive analysis. The major areas in which this thesis aims to contribute are then introduced, namely knowledge discovery and association mining, as well as a section that introduces the possible effects of user participation based upon research in the fields of psychology and Human Computer Interaction. The introduction concludes by presenting the general approach of this thesis and addressing issues of terminology.

The next three parts present the thesis contributions, each of which contains a review of the pertinent area and a contribution. Parts II and III present research into the foundation fields of association analysis and rule presentation, while Part

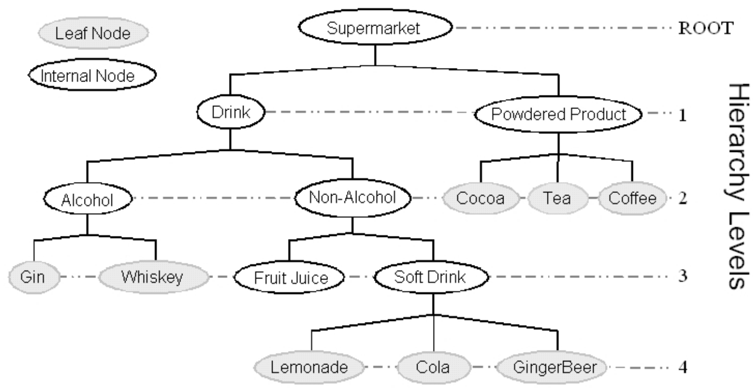

Figure 1: Simple Concept Hierarchy

IV culminates in the presentation of the guided association architecture.

Part II discusses association analysis and is divided into two chapters. The first chapter presents a comprehensive review of current techniques used in the discovery of inferences, focusing upon data structures, traversal strategies and semantic inclusion. The second chapter presents a novel closed set incremental association mining algorithm, MCL, that improves on the state-of-the-art in incremental association mining through the maintenance of a smaller concise representation of the data based upon the concept of closed-sets, defined in Section 1.3.1. Given that knowledge discovery is user centric, reducing the size of the maintained structure facilitates user interpretation. MCL also creates a closed-set representation of the increment dataset, providing the user with insight to the increment's effect upon the maintained lattice and an effective means of incorporating windowing functionality.

Part III discusses the presentation of association rules or inferences and is divided into two chapters. The first presents a review of current presentation techniques, with a focus on graphical visualisation. The second chapter presents CARV, a novel visualisation technique that enables the presentation of inferences within a hierarchical context.

Part IV presents the culmination of this thesis over four chapters, the first two chapters of which are surveys. The first chapter discusses methods by which exploration is constrained within association analysis, presenting a review of current techniques and identifying the different types of constraints that need to be implemented to realise a holistic guided association analysis environment. The second chapter reviews the current techniques used to enable constraint refinement within a knowledge discovery session, which falls into iterative and interactive refinement. Iterative refinement is discussed in relation to association analysis only, while interactive refinement (or guidance), being central to this thesis, is discussed in relation to the knowledge discovery process itself and in regard to the exploratory tasks of clustering, classification and association mining.

The third chapter of Part IV presents the proposed guided architecture, discussing the role of each architectural component in facilitating user interaction. The final chapter presents GAM, a proof-of-concept tool that, based upon the proposed architecture, provides a guided association mining system that dynamically incorporates the refinement of an example constraint for each constraint class identified (see Chapter 5). The thesis concludes in Part V with a discussion of the thesis contributions, areas of further work and a conclusion.

Publications

The following publications have resulted from material presented within this thesis. Publications 1 and 2 relate to initial research efforts into guided association mining and although much has been superseded, remnants can be found in Chapters 3 and 7. Publication 3 directly relates to material presented in Chapter 1, while publication 4 relates to Chapter 4.

1. Ceglar, A., Roddick, J.F. and Calder, P. (2003), Guiding Knowledge Discovery through Interactive Data Mining in Managing Data Mining Technologies in Organisations: Techniques and Applications, Pendharker, P., IDEA Group Publishing, 45-90.

2. Ceglar, A., Roddick, J.F., Mooney, C.H. and Calder, P. (2003). From Rule Visualisation to Guided Knowledge Discovery. In Proc. Second Australasian Data Mining Workshop (AusDM'03), Canberra. Simoff, S. J., Williams, G. J. and Hegland, M., 59-94.

3. Ceglar, A. and Roddick, J.F (2003), Association Mining, ACM Computing Surveys (submitted), 2003.

4. Ceglar, A., Roddick, J.F., Calder, P and Rainsford, C.P. (2005), Visualising Hierarchical Associations, Knowledge and Information Systems (to appear), 2005.