

School of Computer Science and Software Engineering

Monash University

Bachelor of Software Engineering Honours (2770),

Clayton Campus

Research Proposal - Semester 1, 2003

Data Mining Environment Development

Name: Mark Hollands

Id: 13079042

Supervisor: Trevor Dix

Contents

1	Introduction	3
2	Research Context	3
3	Research Plan and Methods	5
3.1	Research Methods	5
3.2	Proposed Thesis Chapter Headings	5
3.3	Timetable	6
3.4	Special Facilities Required	6
4	Relevance of the Research	6
5	References	7

1 Introduction

Data mining, otherwise known as Knowledge Discovery in Databases, is defined as *"the exploration and analysis by automatic or semi-automatic means of large quantities of data in order to discover meaningful patterns or rules"* [1]. It first received recognition as a significant field of study during the early 1990s. This was stimulated by research from a number of different fields as the potential in extracting information from large real world datasets was discovered. Over the past decade, data mining has made the transition from a largely academic venture to become a major development area for information technology.

There are now more than one hundred commercial data mining software packages, sometimes called sftware, on the market. Despite the considerable commercial investment in data mining in both development and usage, generally software products have struggled in the marketplace. It has been speculated that this is because the initial attempts at creating data mining systems were constructed with little forethought for system extensibility or consideration for ease of use [2]. Due to the inherent diversity in how individual users think and interact with a system means that the process of assisting users in making accurate inferences about a model can be extremely complex.

The process of data mining is conducted over a number of distinct stages. Throughout these stages a model of the data is built, analysed and revised. By adding a graphical interface to the existing data mining system, Snob, this study will examine the effects upon system usability over the various stages in the knowledge discovery (KD) process. In the process of constructing a more usable interface around the existing system, interfaces will be built to offer better flexibility for further development upon the system.

2 Research Context

Essentially the purpose of a data mining system is to assist the user in discovering *"meaningful patterns or rules"* [1]. This is achieved through the knowledge discovery process, which can be divided into the following phases [3]:

1. Determining what problem is to be solved, i.e. what domains are to be explored. These constitute hypotheses that the data mining process is being conducted to test.
2. Extracting the relevant data set for mining from the source database.
3. Pre-processing of the data set, dealing with missing values and/or invalid data.
4. Using the system to build the required model from the data.
5. Analysing the model against the hypotheses.
6. Modifying the domain of model being processed based upon the information provided by the DM. The last two phases will be conducted iteratively towards the goal of providing better understanding from a more comprehensive model.

Historically data mining systems have placed considerable importance upon the mathematical analysis involved in building the model, in phase four. The task of successfully conveying the information discovered to the user is too often forgotten [4]. This simplification of the knowledge discovery process down to a fixation upon data and outcomes leads to the user becoming disassociated with the model that has been created. Subsequently the user can struggle to come to terms with knowledge discovery [5].

Some data mining software is still yet to reach an adequate level of maturity associated with commercial software products [3]. The academic beginnings are still evident in the major flaws these software products exhibit:

- They are based upon non-extensible and inflexible system frameworks.
- They provide a non-uniform mining environment, differing interface(s) across implementations of different data mining techniques.

From examining the shortcomings of previous data mining systems, it is important that the system be designed with the user as the focus in relation to the knowledge discovery process [7]. The data mining system must be a consistent environment, seamlessly integrating all of the KD phases. This will provide the minimum level of usability exhibited in successful commercial software. To enable data mining to be commercially successful it is essential that the system be usable by users who are not expert analysts. End users of data mining systems are interested in using data mining as technique for solving business problems.

Considerable gains in usability can be established at a number of different phases within the KD process. The first two phases of the KD process are outside the scope of this project. By pre-processing the data before it is input into the system, the system will be able to recognize and either automatically deal with invalid data, or prompt the user for an action. This will minimize time that would be otherwise wasted in attempting to isolate the cause of spurious results at the end of the process. Allowing the user to define the domain of the model from a larger data set will force the user to think more carefully about the model that is being created. Through the use of visualization and interpretation of the results, the system will assist the user in judging the accuracy of hypotheses that are being tested. After testing the hypotheses the domain of the model can be modified. This will in effect streamline the iterative process through phase four and five. Instead of having to manually having to reformat the data sets, this process will be automated.

Fully automated data mining tools using a web interface have displayed increased usability for users new to the concept of data mining [4]. Experiments in supporting the iterative nature of the KD process through the use of visual data mining frameworks have produced encouraging results [6]. Further evaluation is still needed of the effectiveness in applications of a more general nature. Highly graphical metaphor based visualisation has provided new insights into data mining interface design and methods with which to support the KD process [3].

3 Research Plan and Methods

3.1 Research Methods

This project will develop a web accessible graphical user environment around the existing data mining system, Snob. The version written in C, known as Snob-Vanilla, will be used. The surrounding environment will be coded in an internet scripting language. Quantitative analysis techniques will compare the time spent by users as they interact with the new and original versions of the system. For users who are unfamiliar with the system, testing focused upon their learning curve will show which system they can become proficient in the fastest. For users that have had prior experience with using Snob, qualitative analyses will be performed to gauge the users feel for the system. This will assess the level of control that the user experiences when operating a known system through a new interface.

3.2 Proposed Thesis Chapter Headings

1. Introduction
 - 1.1 Purpose of Research
 - 1.2 Objectives of Research
2. Visual Data Mining Systems
 - 2.1 Interface design
 - 2.2 User centred design
 - 2.3 Data Mining Process
3. System Design and Implementation
 - 3.1 Interface Layering
 - 3.2 Flexibility
4. Conclusion and Future Work
5. Bibliography
6. Appendix A Sample Data
7. Appendix B Programs

3.3 Timetable

Date	Week No.	Activity
21st March	3/13	Familiarize with project
24th March	4/13	Study existing system
7th April	6/13	Literature Study
28th April	8/13	Finalize Research Proposal
5th May	9/13	Begin framework implementation
12th May	10/13	First draft of Literature Review
19th May	11/13	Interface Design
2nd June	13/13	Finalize Literature Review
9th June	1/6	Finalize system framework
16th June	2/6	Interface Implementation
1st July	3/6	Visualisation Implemenation
21st July	1/13	Draft progress report
4th August	3/13	Design usability study
18th August	5/13	Finalize implementation and testing
25th August	6/13	Thesis Chapter 1 - Introduction
22nd September	10/13	Prepare for presentation
20th October	13/13	Finalize Thesis

3.4 Special Facilities Required

The facilities that are offered to Honours students at Monash University will be sufficient to do the research.

4 Relevance of the Research

Existing data mining systems often neglect the needs of the user in favour of design which is focused strongly upon mathematical analysis. This project will provide a web based interactive environment for an existing data mining system to give improved usability. This environment will be designed with flexibility as a key feature, as to enable further development of the system in the future. The design decisions made towards the goal of improving the usability and flexibility of the system will be useful in the future development of data mining systems.

5 References

- [1] M. Berry, G. LinHoff, Data Mining Techniques for Marketing, Slaes and Customer Support, Chapter 1, Wiley Computer, Publishing, 1997.
- [2] J. Hipp et al. Data Quality Mining. DMKD2001 Workshop on Research Issues in Data Mining and Knowledge Discovery DMKD2001, 2001.
- [3] N. Robinson, M. Shapcott. Data Mining information visualisation - beyond charts and graphs. In: Sixth International Conference on Information Visualisation. IEEE. 2002. Page(s): 577-583
- [4] P. Myllymaki, T. Silander, H. Tirri, P. Uronen. Bayesian Data Mining on the Web with B-Course. In: Proceedings IEEE International Conference on Data Mining. IEEE. 2001. Page(s): 626-629
- [5] M. Ankerst, Human Involvement and Interactivity of the the Next Generation's Data Mining Tools, Workshop on Research Issues in Data Mining and Knowledge Discovery. 2001.
- [6] M. Kreuseler, H. Schumann. A flexible approach for visual data mining. Visualization and Computer Graphics, IEEE Transactions on. 2002. 8(1): 39-51
- [7] T. Ho, T. Nguyen. Visualization support for user-centered model selection in knowledge discovery in databases. In: Tools with Artificial Intelligence, Proceedings of the 13th International Conference on. IEEE. 2001. Page(s): 228-235