Project Proposal

Department of computer science Rhodes University

COMPUTER SCIENCE HONOURS PROJECT PROPOSAL

Title: Data Mining with Oracle using either Clustering or Classification Algorithms

By: Nhamo Mdzingwa
Supervisor: John Ebden
Date: 7 March 2005

---

1. Statement of the Problem

1.1 Aims:

Data mining also known as knowledge discovery, involves finding unexpected but interesting patterns within enormous amounts of data that are normally stored in databases and data warehouses. Data Mining has three major components Clustering or Classification, Association Rules and Sequence Analysis which come in form of many algorithms proposed. Some of the algorithms have had better success than the others.

However, the commercial world is fast reacting to the growth and potential in this area as a wide range of tools are marketed under the label of data mining. The main objective of this project is to investigate two types of algorithms available in Oracle for data mining. Apply the two algorithms to actual data. Then, analyse the results and compare outcome in terms of accuracy, efficiency and effectiveness.

2. Resources (including literature survey)

The resources listed below will basically serve as a starting point for my research project. As the project progresses, more detailed resources will be required. These may help in the understanding of Oracle data mining algorithms.

Michael J.A. Berry and Gordon S. Linoff. *Mastering Data Mining.* The Art and Science of Customer Relationship Management, Wiley Computer Publishing, 2000.This will allow for the understanding and of uses of data mining. It also discusses different methodologies used in data mining, thus it will help with evaluations of results.

Jiawei Han and Micheline Kamber, *Data mining:* concepts and techniques. San Francisco, California, Morgan Kauffmann, 2001. This book gives a wide range of information on the different algorithms used in data mining. It also discusses different uses for the algorithms. Therefore, it will be useful for analysing performance of the data mining algorithms

Robert P. Trueblood and John N. Lovett, Jnr. *Data Mining and Statistical Analysis Using SQL,* USA, Apress, 2001. This book discusses the statistical concepts to support data mining. It may be useful for understanding when to use a specific algorithm. It tackles statistics using SQL for data mining which may help if the project is extended.

David Hand, Heikki Mannila and Padhraic Smyth, *Principles of data mining*. Cambridge Massachusetts, MIT Press, 2001. This book will be important as it provides a tutorial overview of the principles underlying data mining algorithms and their application. It also devotes a chapter to data preparation which will be useful when investigating the implementation of data mining in the Oracle database. This text should provide assistance for deciding how to go about conducting experiments.

Jesus Mena, *Data mining your website.* Digital Press, 1999. This text explains how data mining is a foundation for the new field of web-based, interactive retailing, marketing and advertising. It will help in the identification of data set to test the selected algorithms for the data mining.

Richard J. Roiger and Michael W. Geatz, *Data mining: a tutorial- based primer*. Boston, Massachusetts, Addison Wesley, 2003; This book will provide the necessary background and practical knowledge required for the project research and also presents different methodologies used in data mining that may be useful.

There are many resources available on the oracle website containing investigations on Oracle9i data mining. For example, this site http://www.lc.leidenuniv.nl/awcourse/oracle/datamine.920/a95961/preface.htm provides a manual for Oracle9*i* Data Mining which is available as part of the Oracle9*i* Database Documentation Library. This manual describes how to use the Oracle9*i* Data Mining Java Application Programming Interface to perform data mining tasks, including building and testing models.

The resources at this site: http://www.oracle.com/technology/products/oracle9i/htdocs/o9idm_faq.html will provide more information on the Oracle data mining suite including benefits associated with using the suite. The site contains frequently asked questions (FAQ) associated with the Oracle9i data mining suite. This will prove very useful for getting to grips with the product as well as understanding the workings of the product.

The site below is a university / Non-profit Research Group site which also has frequently accessed sites (FAS). It provides many links relating to my research. Simon Fraser University's database group: Knowledge Discovery in Databases and Data Mining (Research papers of Prof. Jiawei Han) http://fas.sfu.ca/cs/research/groups/DB/sections/publication/kdd/kdd.html .

3. Proposed Hardware & Software (including availability)

Oracle Data Mining runs in the Oracle 10g Database. Oracle9i Data Mining is an option to Oracle9i Database Enterprise Edition (EE) that embeds data mining functionality for making classifications, predictions, and associations (all of which are already available in the CS department at Rhodes). This is all installed on the server ORA1 in the centre of excellence.

4. Timeline for implementation

This is just an initial timeline that will help me focus on major events in the course of the project. The project will however be researched and developed using an iterative approach, where analysis, design and implementation will continuously be taking places.

PROJECT MILESTONE	PROPOSED DATES:
Carry out a literature search, and an evaluation of the area to obtain background knowledge and understanding of the field.	1st term- *End 31 March*
Get to know Oracle database software, that is, oracle 10g database and Oracle9*i* Database Enterprise Edition (EE)	1st semester's work.
Get to know Oracle data mining software Oracle9i Data Mining. Study recommended papers and tutorials related to research project.	Should have reasonable understanding of oracle data mining software by end of 1st semester
Prepare project proposal presentation.	*22-31 March*
Investigate Clustering & Classification algorithms (theory) and required appropriate data types for these algorithms. (may include theoretical case studies)	2nd term- *15 to 30 April*
Find suitable computerised case studies of the use of above algorithms – with or without Oracle.	2nd term- *End of May*
Find suitable (nature and size both non-trivial) databases for testing (possibilities: AIDS data & faculty data)	2nd term- *End of May*
Apply algorithms above to data found in above step	Second semester
Critically Analyse & assess results	Second semester
Write up paper	September vacation and 3rd term
Final project write up.	Due 7/11

5. Expected Deliverable

This project aims at evaluating algorithms that will be most effective and suitable for the process of data mining on any set of data.

6. Possible Extensions

Possible extensions of this project include testing of the same algorithms with different tools offered by other vendors.