CLUSTERING ALGORITHM FOR MARKET-BASKET ANALYSIS: THE UNDERLYING CONCEPT OF DATA MINING TECHNOLOGY

KHAIRIL ANNUAR B. ABDUL KADIR

FSKTM 2003 7

Clustering Algorithm for Market-basket Analysis: The Underlying Concept of Data Mining Technology

By

KHAIRIL ANNUAR B. ABDUL KADIR

The Submitted in Partial Fulfillment of the Requirement for the
Degree of Master of Science in Information Technology in the
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia

March 2003

PREFACE

One of the prerequisites to successfully completing the course is to be able to deliver the final project in the Master of Science in Information Technology. This is a compulsory subject with three credit hours. The topic of this final project is based on the data mining concept. The title of the project is “Clustering Algorithm for Market-basket Analysis: The Underlying Concept of Data Mining Technology.” The objective of the project is to analyze the underlying concept of data mining technology by exploring the concept of clustering algorithm and the application of market-basket analysis in PolyAnalyst 4.5 (data mining software). To achieve this, some information about what data mining is all about, what are the algorithms used in market-basket analysis, and how the software works have to be gathered.

The application of the software seeks for hidden patterns of a sample data set of different products in supermarket. The readers are encouraged to download the evaluation version of the software from the Internet (please refer to the references section). The concept of clustering algorithm is presented in a manner that helps readers to understand the concept of data mining technology in market-basket analysis.

ABSTRACT

The goal of data mining is to extract interesting correlated information from large databases. This thesis seeks to understand the underlying concept of data mining technology in market-basket analysis. The clustering algorithm based on Small Large Ratios, SLR is presented in a manner that helps to understand the concept of data mining technology in market-basket analysis. The author used a data mining software called PolyAnalyst 4.5 to perform analysis on the set of items that customers have bought in supermarket for market-basket application. In this research, the author tried to relate the algorithm presented with the experiment. Then, the author discussed the results by showing an application of market-basket analysis. The statistical results from the PolyAnalyst's reports are explained and elaborated further in the results section. The author summarized the findings and tried to relate them to the benefits of data mining towards organizations.

ABSTRAK

Fungsi utama perlombongan data adalah untuk melombong keluar maklumat menarik yang diperlukan daripada pengkalan data. Tesis ini mencari untuk memahami konsep disebalik teknologi perlombongan data di dalam analisa pasar-raga. Algoritma 'clustering' berdasarkan 'Small Large Ratios, SLR' dipersembahkan untuk membantu pemahaman konsep tersebut. Penulis menggunakan perisian perlombongan data yang dipanggil PolyAnalyst 4.5 untuk menganalisa satu kumpulan barangan yang dibeli oleh pelanggan di pasaraya untuk applikasi pasar-raga. Di dalam penyelidikan ini, penulis cuba mengaitkan algoritma yang dikaji dengan eksperimen yang dijalankan. Kemudian, penulis membincangkan hasil penemuannya daripada applikasi pasar-raga. Hasil keputusan statistik daripada repot yang dikeluarkan oleh perisian PolyAnalyst telah dibincangkan dan dihuraikan dengan lebih lanjut di bahagian penemuan tesis ini. Penulis membuat kesimpulan terhadap penemuan yang telah diperolehi dan cuba menghubungkan penemuan tersebut terhadap kepentingan perlombongan data kepada setiap organisasi.

ACKNOWLEDGEMENT

I would like to take this opportunity to thank the individuals who have fully supported my works towards realizing this project. I would like to thank my supervisor, Prof. Madya Dr. Md. Nasir b. Sulaiman, for his assistance in providing the framework of my research. He was very helpful with his knowledge and this gives me the advantage to complete my project.

I would like to thank all of the lecturers, staffs, and fellow classmates who were directly or indirectly involved throughout the whole program that we had gone through.

I would like to acknowledge the invaluable support that I have received from my friends and to all of my colleagues at work, especially to Shahril Ayob, Syahrun Nazri, Maria Azah, Fakruzrazi, and Muhaini Yusof.

Most importantly, I would like to thank my family especially both of my parents who are fully supportive towards my studies. Their moral support always reminds me to be serious in my studies. Thank you God and again I wish to thank everyone.

APPROVAL

This Thesis was partially submitted to the Senate of Universiti Putra Malaysia and was accepted as partly of the requirements for the degree of **Master of Science in Information Technology**.

PROF. MADYA DR. MD. NASIR B. SULAIMAN

Faculty of Computer Science and Information Technology,

Universiti Putra Malaysia

Date

TABLE OF CONTENTS

Page

PREFACE 2

ABSTRACT 3

ABSTRAK 4

ACKNOWLEDGEMENTS 5

APPROVAL 6

DECLARATION 7

LIST OF TABLES 10

LIST OF FIGURES 11

LIST OF ABBREVIATIONS/GLOSSARY OF TERMS 12

CHAPTERS

1. INTRODUCTION 14

2. LITERATURE REVIEW

    2.1 Data Mining Techniques and Applications 18

    2.2 Market-Basket Analysis 20

    2.3 Clustering Technique 21

    2.4 PolyAnalyst 4.5 – Data Mining Software 22

3. METHODOLOGY

    3.1 Theory: Understanding the Clustering Algorithm 24

        3.1.1 The Concept of Large Items and Small Items 26

        3.1.2 SLR Algorithm 28

        3.1.3 Discussion 30

    3.2 Experiment: Market-basket using PolyAnalyst 4.5 31

3.2.1 System Requirements 31

4. RESULTS AND DISCUSSIONS

4.1 Experiment: Market-basket using PolyAnalyst 4.5 43

4.1.1 First Analysis 43

4.1.2 Second Analysis 48

4.1.2.1 Discussion 50

4.1.3 Third Analysis 51

4.1.3.1 Discussion 52

5. CONCLUSION

5.1.1 Benefits to Organizations 53

5.1.2 Meeting the Objectives 54

5.2 Future Research and Demands 55

REFERENCES/BIBLIOGRAPHY 56

APPENDICES

Appendix A: The cost function of SLR algorithm 58

Appendix B: Localization of Anomalies (LA) 59

Appendix C: Options in the Basket Analysis Dialog Box 61

Appendix D: What is New in PolyAnalyst v4.5 63

BIODATA OF AUTHOR 66

LIST OF TABLES

Page

Table 3.1: The sample dataset for market-basket analysis 25

Table 3.2: Predetermined clustering U_0 = {C1, C2, C3} 26

Table 3.3: The large, middle, and small items in each cluster 27

Table 3.4: The calculation of SLR in cluster C1 28

Table 3.5: Summary of the experiment 42

LIST OF FIGURES

	Page
Fig. 3.1: The implementation of algorithms in PolyAnalyst Server	31
Fig. 3.2: The dialog box for project editing	34
Fig. 3.3: The tree view of PolyAnalyst	35
Fig. 3.4: The dataset	35
Fig. 3.5: Basket Analysis dialog box	36
Fig. 3.6: Different values of 15 percent and 70 percent	40
Fig. 3.7: Different values of 25 percent and 75 percent	41
Fig. 3.8: Summary statistic of market-basket analysis	42
Fig. 4.1: The clustering result of the first analysis	44
Fig. 4.2: The product association rules of the first analysis	46
Fig. 4.3: The clustering result of the second analysis	49
Fig. 4.4: The association rules of the second analysis	50
Fig. 4.5: The result of the third analysis	51

LIST OF ABBREVIATIONS/ GLOSSARY OF TERMS

Abbreviations

LA	Localization of Anomalies
SLR	Small Large Ratios
SKAT	Symbolic Knowledge Acquisition Technology

Glossary

Clustering. Machine learning task aimed at identifying groups of records in a database that are similar between themselves but very different from the rest of data. Clustering is an unsupervised data mining algorithm – it does not require a target attribute. PolyAnalyst offers Cluster algorithm based on Localization of Anomalies technique.

Data Mining. “Data Mining is the process of identifying valid, novel, potentially useful, and ultimately comprehensible knowledge from databases that is used to make crucial business decisions.” -- G. Piatecki-Shapiro, kdnuggets.com. Synonym: Knowledge Discovery.

Empirical Model. Predictive model developed through mining data in a database with the help of a machine learning algorithm.

Exploration Engine. Refers to machine learning algorithms implemented in PolyAnalyst.

Knowledge Discovery. Synonym to Data Mining in our interpretation.

Market-basket Analysis. Data mining technique aimed at finding groups of features that frequently occur together in transactional data and identifying directed association rules inside these groups of features. Derives its name from its original use in retailing, but can be applied successfully in many other fields.

OLE DB for Data Mining. An OLE DB extension that supports data mining operations over OLE DB data providers. The goal of this specification is to provide an industry standard for data mining so that different data mining algorithms from various data mining vendors can be easily plugged into user applications.

SKAT. Symbolic Knowledge Acquisition Technology - develops an evolving model from a set of elementary blocks, sufficient to describe an arbitrarily complex algorithm hidden in data, instead of routine searching for the best coefficients for a solution that belongs to some predetermined group of functions. SKAT is implemented in PolyAnalyst as the Find Laws exploration engine.

CHAPTER 1

INTRODUCTION

1.1 Overview

In most organizations, there are thousands and millions of accumulated data, which are meaningless if not converted to relevant information. Data mining technology is a tool that can help the organizations to extract information from large databases. “Data mining is part of a larger process called knowledge discovery” (Wipro, 2001).

To be more specific, data mining is a decision support process in which we search for patterns of information in data i.e. by performing queries, graphs, and reports. It uses sophisticated statistical analysis and modeling techniques to uncover patterns and relationships hidden in databases. It has attracted a growing amount of attention by many industrial companies due to its wide ability to improving marketing strategies with an opportunity of major revenues.

1.2 Problem Statement

Most organizations nowadays do not realize the importance of data mining technology towards the benefits of the organizations. This is probably because of the emergence of hundreds of organizational-related tools and software in the market that causes confusion to many corporate people. Therefore, this research would directly and indirectly focus on addressing the importance of data mining to the organization.

1.3 Objectives

The objectives of the project are:

1) To study the underlying concept of data mining technology by exploring the concept of clustering algorithm.
2) To run the application of market-basket analysis to extract hidden patterns among different products in supermarket by using a data mining software called PolyAnalyst 4.5.

The goal of market-basket experiment is to find out which of these products sell well together, so that when a customer goes shopping in a supermarket, the related products can be arranged together to increase the chance of making a sale.

Before performing the experiment, the author analyzed the modeling technique used in market-basket analysis, which is the analysis of clustering algorithm based on Small Large Ratios, SLR. The analysis is based on the example in (Yun *et al.*, 2001) that utilizes the concept of SLR to divide the transactions into clusters such that similar transactions are in the same cluster and dissimilar transactions are in different clusters.

From the results of the experiment, the author discussed the benefits of the extracted information to the sales department of the supermarket, and concluded with the benefits of data mining towards organizations as a whole.

1.4 Scope of works

This work concerns only with the analysis of the application of market-basket analysis and the algorithm behind it. Therefore, the work will include the study of only one type of several clustering algorithms, some examples of the market-basket analysis for a supermarket transaction history and the explanation of the findings.

1.5 Organization of the Report

This thesis is divided into five chapters:

Chapter 1 provides some background knowledge of the project and highlights the scope of research and its objectives.

Chapter 2 discusses the literature review of the research. This chapter basically explains the techniques and applications of data mining in the organizations, the application of market-basket analysis and the algorithm behind it, which is the clustering algorithm, and the software of data mining used in this research, which is PolyAnalyst 4.5.

Chapter 3 deals with the methodology of this research. It describes the study of a clustering algorithm called Small Large Ratios, SLR. Then, it analyzes an application of market-basket analysis using PolyAnalyst software. The results of the experiment are further discussed in the chapter 4.

And finally, Chapter 5 concludes this research by summarizing the findings and relates them to the benefit of data mining towards organizations.

CHAPTER 2

LITERATURE REVIEW

2.1 Data Mining Techniques and Applications

“Data mining can use programming methods to identify patterns among data objects” (Dennis *et al.*, 2001). Data mining comprises of three techniques that are commonly used by programmers. The techniques are clustering or classification, association rules and sequence analysis.

The clustering techniques generate a set of grouping rules to classify data. Once the clusters are generated, data mining tools or programs can be used to identify and study the patterns of each cluster. The association rules are rules “that implies certain association relationships among a set of objects in a database” (Wipro, 2001).

In sequential analysis, programmers seek to discover patterns in data that occur in sequence. For example, if a customer buys a computer, he is also expected to purchase software games or printer cartridges.

Many processes from marketing to customer service can be analyzed using data mining. The three major applications of data mining are in the marketing area, including customer profiling, targeted marketing, and market-basket analysis.

Customer profiling identifies the characteristics of good customers that can predict who will become good customers. This helps the marketing department to target for new prospect customers. Targeted marketing seeks for patterns in a customer database so that customer acquisition can be appropriately targeted. Specific promotions and direct marketing can be customized according to the customers' needs. This helps to reduce expenses and increase sales.

Market-basket analysis helps retailers to identify which products are mostly purchased together by customers. Retailers could use this information for promotion by displaying the products that are frequently sold together in the same area.

Some other applications of data mining are fraud detection, stock prediction, and in medical field. Fraud detection is very popular among telecommunications firms, banking institutions, insurance companies, and government agencies. Data mining technology enables companies to identify and manage potentially fraudulent transactions, thus helps to reduce losses by reducing the number of fraud activities.

Financial companies use data mining to predict the performance of the stock market. Regression methods are commonly used to perform statistical analysis of investment portfolio and financial data. In the medical field, data mining helps to “predict the effectiveness of surgical procedures, diagnostic tests, medications, service management, and process control” (Wipro, 2001).

2.2 Market-basket Analysis

Market-basket analysis is one of the most common methods of data mining for marketing and retailing. The goal is to determine what products customers are likely to purchase together. It takes its name from the idea of customers throwing all their purchases into a shopping cart (a "market-basket") during grocery shopping.

Market-basket analysis concerns the rules of subset of items taken from a larger set of items. For example, a rule is denoted as A → B, where A is referred to as the rule’s antecedent condition and B is the consequent. The rule is interpreted as “If A occurs in the market basket, then B also occurs in the same basket” (Gayle, 2000).

A well-known example is that WalMart in the USA discovered from their basket analysis that diapers and beer sell well together on Friday evenings (Dennis *et al.*, 2001). Though the result does make sense – fathers were taking home extra beers when they went to buy diapers before the weekend starts – it's not the sort of thing that someone would normally think of right away.

The strength of market-basket analysis is that it is a data-driven marketing since the customers' sales data does the entire job in the analysis. A sales person or a marketer does not have to think of what products customers would logically buy together.

2.3 Clustering Techniques

Clustering is a process of grouping a set of data into subsets, "such that the degree of similarity between cases in one group is significantly higher than between members of different groups" (Moore *et al.*, 2002). A cluster has a property of a concrete dataset which has been explored. Moore *et al.* (2002) considers a cluster as a region where the concentration of data is significantly higher than in other regions.

There are several clustering algorithms that have been derived by researchers such as Localization of Anomalies (LA), Apriori, Fast Distributed Mining (FDM), and Small Large Ratios (SLR). Correspondingly, the clustering algorithm analyzed in this thesis is Small Large Ratios (SLR).

2.4 PolyAnalyst 4.5 – Data Mining Software

PolyAnalyst 4.5 is a new knowledge discovery system, which is a trademark of Megaputer Intelligence, Inc. (copyright © 2002). It provides wide range of data mining solutions for various industries like Finance, Sales, Marketing, Healthcare, R&D, and Manufacturing.

PolyAnalyst contains sixteen advance knowledge discovery algorithms that perform a thorough analysis of data, automatically extracting the previous knowledge from an investigated database and presenting it in an easily understood form. One of the most advance exploration engines or algorithms is Find Laws, which utilizes a unique “Symbolic Knowledge Acquisition Technology” (SKAT) – a next generation data mining technique.

PolyAnalyst automatically finds dependencies and laws hidden in data, presenting them explicitly in the form of rules and algorithms. Relationships in the data can be discovered, predictions made, and the data classified and organized using analytical algorithms. It can perform tasks beyond the scope of statistical analysis software such as Microsoft Excel.

PolyAnalyst is a complete suite of data mining algorithms. It includes sixteen different approaches of data mining in one package. The following data mining algorithms are: Find Laws, Nearest Neighbor, Neural Network, Find Dependencies, Stepwise Linear Regression, Market-basket Analysis, Transaction Basket Analysis, Cluster, Classify, Discriminate, Summary Statistics, Decision Tree, Decision Forest, Text Analysis, Text Categorization, and Link Analysis.

CHAPTER 3

METHODOLOGY

3.1 Theory: Understanding the Clustering Algorithm

The goal of this sub-chapter is to study on how clustering is performed using SLR. The clustering algorithm is meant to divide a set of data items into proper groups. In market-basket analysis, data is represented by a set of transactions. Table 3.1 is an example of a list of transactions of items that customers bought at supermarket.

In this dataset, rows represent transactions and columns represent product codes. Each cell contains a yes/no value representing if that product was purchased in that transaction.

The transactions can be denoted by D = $\{t_1, t_2, \ldots, t_h\}$, where each transaction t_i, has a set of items $\{i_1, i_2, \ldots, i_h\}$. To illustrate the clustering algorithm, the dataset is assumed to have only fifteen transactions and ten items. The dataset is then simplified into a predetermined clustering U_0 = {C1, C2, C3}, where C1 = $\{t_1, t_2, t_3, t_4, t_5\}$, C2 = $\{t_6, t_7, t_8, t_9, t_{10}\}$, and C3 = $\{t_{11}, t_{12}, t_{13}, t_{14}, t_{15}\}$ as shown in Table 3.2.