Faculty of Information and Communication Technology

IMPLEMENTATION OF ASSOCIATION TECHNIQUE IN IDENTIFY THE FREQUENT PAIRING SUBJECT FROM PRIVATE TUTOR DATA

NUR HAMIZAH BINTI ASNOR

Master of Computer Science (Database Technology)

2017

IMPLEMENTATION OF ASSOCIATION IN IDENTIFY THE FREQUENT PAIRING SUBJECT FROM PRIVATE TUTOR DATA

NUR HAMIZAH BINTI ASNOR

A thesis submitted
in fulfillment of the requirements for the Master of Computer Science (Database Technology)

Faculty of Information and Communication Technology

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2016

DECLARATION

I declare that this thesis entitled “Implementation of association in identify the frequent pairing subject from private tutor data” is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature : ……..

Name : ……..

Date : ……..……

APPROVAL

I hereby declare that I have read this thesis and in my opinion this thesis is sufficient in term of scope and quality for the award of Master of Computer Science (Database Technology).

Signature	:	……..
Supervisor Name	:	……..
Date	:	……..……

DEDICATION

I dedicate the work and effort poured into this project to lecturer and friends who encouraged me in this project. A special thanks to my mother Aumikalsom binti Abdullah and father Asnor bin Hamid also my beloved brother, sister and best friends. Your blessing and support make this project complete. This success belongs to ours.

ABSTRACT

In this project we are doing research on implementation data mining in big data. We have explored about Market Basket Analysis is that consist of the algorithm, technique and implementation method. In market basket analysis, normally we can identify the business study and research were focusing on the marketing and business domain. There are no research's about education domain. So in this research, we would like to implement the market basket analysis in education domain. We would like to identify the frequent pairing request subject by customer. We are using real data that has been scrapped from online tuition website. The implementation process will be doing using rapid miner tools. An Apriori and Association algorithm was selected to implemented this project.

ABSTRAK

Dalam projek ini kami melakukan penyelidikan mengenai pelaksanaan data di dalam data yang besar. Kami telah menerokai mengenai analisis pembelian produk ia yang termasuklah mengkaji kaedah algoritma, teknik dan pelaksanaan. Dalam analisis bakul pasaran, biasanya kita dapat mengenal pasti kajian perniagaan dan penyelidikan telah memberi tumpuan kepada domain pemasaran dan perniagaan. Tiada pengkajian di dalam domain pendidikan. Jadi, dalam kajian ini, kami ingin melaksanakan analisis bakul pasaran dalam domain pendidikan. Kami ingin mengenal pasti kekerapan permintaan berbelian tertakluk oleh pelanggan. Kami menggunakan data sebenar yang telah diambil dari laman sesawang tuisyen. Proses pelaksanaan akan melakukan menggunakan "Rapid Miner". Algoritma Apriori dan Association telah dipilih untuk melaksanakan projek ini.

ACKNOWLEDGEMENTS

Apart from the efforts contribution by myself, the success of my project relies highly n on the encouragement and guidelines of many others. I would like to take this opportunity to express my gratitude to the people who have been instrumental in the successful completion of this project. Special Thanks to my parent for providing me with the opportunity to be where I am. Without them, none of this would be possible. I would also like to thank you my friends for their encouragement, input and constructive criticism which are priceless and also for their moral support.

I would like to express sincere gratitude to my supervisor, Dr Sharifah Sakinah binti Syed Ahmad for continuous support and guidance in completing this project and research. A special thanks to Profesor Madya Dr Azah Kamilah binti Draman @ Muda for this guidance in helping me in my research. I could have imagined having a better advisor and mentor for completion of my final year project.

Not forgetting friends and all my educators in Computer Science (Database Technology), for their perceptiveness, understanding and their skilful teaching. And above all many thank you to Allah s.w.t who helps me get all required resources for completing this research.

Table of Contents

CHAPTER I

INTRODUCTION

1. Background

Market basket analysis is to determine the customer frequently bought products and it is one of the data mining methods that focus on a purchasing pattern by doing an extracting or co-occurrence from stores transaction data. Restructure supermarket layout and design or create promotion campaign is one of the purpose of Market Basket Analysis. This process can be improved and increase supermarket sales and profit. The market consumer behaviour need to be analysed to identify different segments of customer and improve customer satisfaction. In this project, we are analysed in educational sector which is focused on the subject that recently requested by customer for private tutors. This analysis will have identified the recent pairing subject from history of private tutor data. We have studied several implementations in data mining technique in literature review chapter and also explored about market basket analysis and association included association rules, apriori and FP-Growth. The implementation was used rapid miner tools because the result and interface are visualise, it is easy to understand. This paper establishes the association of tutor data subject by identify the frequent requested subject by using Rapid Miner.

1.1. Problem Statement

The problem happens when one in unclear information because of the information are too large and also the difficulties to imagine the relationship between the information. Investigated the data is one of the challenges for one the company. How to scrape an information's from the vast amount customer in the database and extract the product feature to gain the competitive advantage is to change. The problem happens when users often get lost in the vast amount of information and customer hard to making a decision. The power of data mining is not yet to be fully exploited by an education domain. For example, in online education or private tutor sector, especially in Malaysia there is no analysis and comparative study about a frequent pairing subject that can be offered to the student. The market basket analysis is normally focusing on business and purchasing like supermarket and stores. The researcher also not found relative study in an Association, market basket analysis, Apriori that focusing on education domain and about the analysis of frequent pairing subject that recently request together by parents or students for private tutors or online education and so on.

1.2. Research Motivation

An address the problem of facing a complicated choice by customer, data mining was used to analyse the data. Data mining recently implemented in the marketing sector, health, e-commerce, etc. But there are no comparative study and analyse about learning education especially private tutor. What recently subject that user request? Which subject was recently request together? What packages of subjects can be added to improve the sales of private tutor subject?

In this project we will analyse about online education related by using real data of history of the tutor requested data on the website. This study may improve the online education sector to create the package of the subject and improve increase sales of private tutor. This study will explore about Association technique which is included Apriori, FP-Growth and Association rules. We have chosen the visualize and easy to understand implementation tools, the tools used is Rapid Miner. In this research, we will explore about market basket analysis, study about frequent pairing itemset and get used of Rapid Miner tools.

1.3. Research Question

To answer the purpose of this study, the researcher's question in the following:

1. The data selected was the real data?

   In this study, we scrape the real data from websites. The data contained private tutor data which are historical requested subject by the customer.

2. What do you analyse from the private tutor or online education data?

   In this research, we have analysed the pairing subject that can be done to improve the private tutor sales, to provide the package for students and this analysis can be used to generate recommendation system for private tutor.

3. How to analyse the subject requested private tutor data?

   There are many types of algorithm can be used in this study. We will study some types of algorithm and technique like Decision Tree, Apriori, FP-Growth and Association before make a decision to choose specific algorithm.

1.4. Research Objective

This research aims to achieve these main objectives:

1. To study about implementation of data mining

2. To study about market basket analysis.

3. To an identify pairing subject based on the historical requested subject by the customer.

1.5. Research Contribution

1. This study will describe the implementation of data mining in the big data. Where the description covered the algorithms and technique of data mining.

2. This research analyses the real data of the education sector, which focusing on the private tutor requested in Malaysia.

3. This study will also describe the implementation of data mining in the selected sector.

1.6. Thesis Organization

This thesis is organized in six chapters. The following chapters are organized and briefly described as follows:

Chapter 1: Introduction

In this chapter, we have introduced this project research by deliberating on the background, research motivations, research questions, research objectives and contribution about implementation data mining in big data.

Chapter 2: Preliminary Study and Literature Review

In the chapter 2, the preliminary study on the implementation data mining in the big data. Which is this study may help to analyse the data. The algorithm is also described in this section.

Chapter 3: Research Methodology

This chapter describes the research methodology that was used in this project. This research methodology involves in five parts.

Chapter 4: Descriptive Implementation

This chapter describes the implementation of data mining that had been used in selected algorithm, topics and tools which is an online education sector.

Chapter 5: Result and Discussions

This chapter presents the result of the analysis, steps of analysis technique and algorithm used. This sector also include discussion, result and testing of the research project.

Chapter 6: Conclusion

The conclusion, is the last chapter in this project report. The conclusion concludes our research limitation and future works for improvement of this project. It is also included the improvement, ideas and difficulties during this research done.

1.7. Summary

Chapter one discusses the background of this project, which is contained research study, problem discussion, objective, research motivation, research question, research contribution and thesis organization of the project. In the next chapter, we will further elaborate about project flow.

CHAPTER II

LITERATURE REVIEW

2. Introduction

In this literature review chapter, we have been doing a comprehensive literature review on implementing Data mining in Big Data. The implementation of big data in data mining have many types of implementation, algorithm, technique and tools. This paper described some of the algorithm and technique that can be used in the data mining included decision tree, classification tree, regression tree, C4.5 algorithms, Apriori Algorithm and k-means algorithm. Other than data mining terms such as web mining, text mining also described in this paper. At the beginning of this literature review section. We have studied several types of implementation of data mining of big data due to less understanding of implementation method and to determine the consumer behavior can be done through different data mining technique. After that, we are focused and done deep study on the market basket analysis, Apriori, FP-Growth, association and all related to the implementation of this project, which in implementation of association technique and process implemented in this research.

2.1. Big Data

Big Data is not the quantity of data that is revolutionary, but the big data revolution is that now we can do something with the data (Jonathan Shaw 2014). From the data, we need to analyze to make a decision making and make an information. An information is being and stored at an unprecedented rate because it comes from variety of sources. An implementation of big data mining technology affected by many factors. Processing big data by using data mining technologies promises to create several benefits which can contribute to the success of information system by providing analytical tools which helped in decision making. Big data play an important role in drawing up effective new promotional policies appropriate pattern like buying and shopping pattern, an improving predictability, Provide creative opportunities and effectiveness of the decision making where it is the key resources for business intelligence (Almoqren 2016).

In Information Technology (IT) word, an Enormous amount of data rapidly increases in the daily life. Data comes from varied sources of data such as phone, computer and sensors. The World Wide Web is an extremely large collection of information. According to (Sharma 2016) web rising dreadfully as approximately 70 million pages added daily. Big data alone is insufficient to make valid causal inference. However, having more data certainly can improve causal inference in large-scale datasets. As an example using matching methods and characteristics of observation to make treatment and control units comparable (Grimmer 2015). The researcher tried to describe about big data using data lifecycle as shown below (Academy n.d.). The complexity increase as volume of data increases.

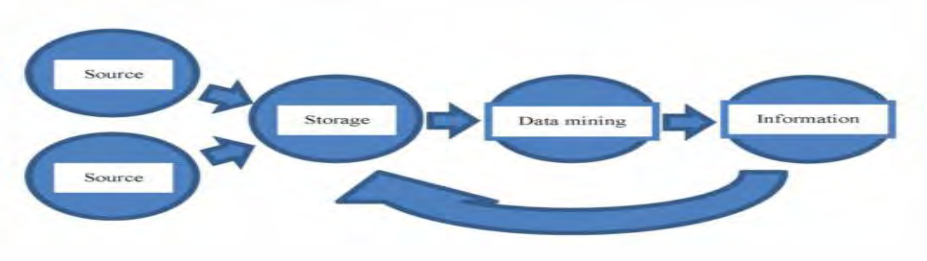

Figure 1: Big Data Lifecycle

Big Data term defining data that have three main characteristics. First, it involves a great volume of data. Second, the data cannot be structured in regular database tables. Third, data is produced with great velocity and must be captured and processed rapidly. An oracle adds a fourth characteristic for kinds of data and that is low value density, means something very big volume of data to process before finding valuable needed information (Garlasu et al. 2013). Big data is relatively new terms that came from the need of big companies like Google, Yahoo, Facebook to analyze big amount of unstructured data, but this need could be identified in a number of other big enterprises as a week in the research development field.

2.2. Big Data Mining

Based on the author(Fan and Bifet 2013) review, "Big Data" terms first time appeared in 1998 in a Silicon Graphics (SGI) slide deck by John Mashey with the title of "Big Data and the Next Wave of InfraStress". The first book mentioning is a data mining book also published in the year 1998 by Weiss and Indrukya the big data mining is very relevant from the beginning. From the large amount of data, an organization, an individual or government using data mining to extract the information. Data mining (DM) and Big Data (BD) are among the technologies that become increasing popular.

Big data mining defined as “the collection and interpretation of massive datasets, made possible by the vast computing power that monitor a variety of digital streams like social information exchanges and analysis them using “smart algorithm”(Almoqren 2016). Enormous advantages and benefits can be resulted from the adoption and implementation of big data technologies, including improving decision-making, understanding customer’s needs, deal with huge amount of data and complexity of data, expectation of future and knowledge discovery for economic community. Process of analyzing data from different perspective and summarizing it into information is a data mining process. Concept of big data and data mining are completely different. However, both of them handle the collection of large data sets or reporting of data that may help business, company or client make better decisions.

2.3. Data Mining Review

The process of analysing data from different perspective and summarizing it into information is a data mining process. In this current IT technology, electronic commerce (e-commerce) site users are rapidly increasing. E-commerce technology makes a user convenient, cheap, and without being limited by space and time where users can serve e-commerce web as long as user have an internet connection. But, the problem comes when the user gets vast of information and having problem to facing the complicated choice. User often get lost in the vast amount of information. In these sections, part of the data mining like text mining, web mining will also have described. Data mining is the process of semi-automatic and analysing large databases to find useful patterns

Data mining is extracting or mining knowledge from large amounts of data. Data mining recently become one of the most progressive and promising field for the data extraction and manipulation to produce useful information. There are thousands of

businesses are using data mining application every day in order to manipulate, identify and extract useful information from the records stored in their databases, data repository and data warehouses.

Data mining tools are the best investment based on the customer's profile. Researcher results indicate the adoption and implementation of big data mining technology helps to employ the principle of running experience in decision making to deliver the right decisions to the right people in a timely manner(Almoqren 2016). From Data Mining blog Phlippe Fournier-Viger author, discuss the steps to implement a data mining algorithm. The steps start with understanding algorithm. second, implementing first draft of the algorithm step by step. 3rd, testing with other input file. 4th, cleaning the code. 5th, optimizing the code. Then make a comparison of the performance with other implementations of same algorithm or peer review.

Data mining examines large pre-existing databases a good way to generate new information. There are various duties protected below data mining and association rule mining is considered as one of the critical responsibilities amongst its. They are in form of if-then types of statements which help to find relationships amongst huge fact which do not preserve relationship with every different inside a relational database or every other information repository (Patil, Vasappanavara, and Ghorpade 2016).

2.4. Web Mining

Web mining is a kind of data mining which is becoming popular. It is from the vast number of network information. Web mining can analyse the source of the user, website advertising click-through rates, the combination of these data can be predict the user access information included analyse user behaviour trends to make classification management. The

analysis can improve the relationship between customer, seller and management. Web mining also helps user to quickly realize their shopping plan. Where web mining technology can provide user interest in information services to users and web page clustering. Web mining technology has high practicability and good practical significance(Lin and Wenzheng 2015). By using web mining technology in e-commerce, personal recommendation service system was created.

Web mining can be applied to the analysis of online user behaviour pattern and make classification management to strengthen relationship between customer and management. It is also good to improve on the quality of site, improving caching web pages and its helps in improving the performance of web pages. At the end, user more quickly realize their shopping plan. The researcher using web mining to analyse and developed personal e-commerce recommendation system. It also can improve the user complicated choice and often get lost in the vast amount of information. Knowledge discovers on web data also referred as Web Mining (Lin and Wenzheng 2015).

2.4.1. Category of Web Mining

Web Mining has four stages i.e. Data Collection, Pre-processing, Knowledge Discovery and Knowledge Analysis. Data Mining use of data mining techniques to automatically discover and extract the knowledge from World Wide Web. In general, Web Mining categories into three are: Web Content Mining, Web Structure Mining and Web Usage Mining. Web Data from variants of data source, data presentation and types. The fraction of population. According to (Sharma 2016), web data were classified into Content, Structure, Usage and User Profile.

1. **Web Content Mining:** Scraping or an extracting useful information from content, data or service available in a web document. Real context that exist on webpages are image, audio, video, text or structured record like list and tables.

2. **Web Structure Mining**: Extraction of pattern from hyperlink within the web itself. Web structure Mining also referred as a web link's structure analysis. There are four steps in web structure mining.

    a. **Data Collection**: first step in any data mining technique collects data required for analysis. In web structure mining data collection means collect hyperlinks from web pages associated with the seed URL from various servers.

    b. **Data Pre-processing:** Implements sequence of the process of web links file performing data cleaning, link validation, link identification, links uniqueness and link completion. Data Pre-processing is a process to represent data in a format as per mining techniques. There are different ways to represent data like chart, graph, etc. Data Pre-processing is classified into four categories:

        i. **Data Cleaning:** fetch cleaned data before processing. Cleaning involved identification of missing, inconsistence or mistaken values. To provide a picture of distribution and statistics like maximum, minimum or minimum value, we can use graphical tools. It in include in the process data cleaning step(Understanding and Understanding n.d.). Because of human error or evolution of reporting problem