PATTERN DISCOVERY IN UTM LIBRARY CIRCULATION DATABASE

HOSEIN JAFARKARIMI

A dissertation submitted in partial fulfillment of the
requirements for the award of the degree of
Master of Science (Information Technology – Management)

Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia

JUNE 2011

This dissertation is dedicated to my beloved mother, father and brother who have never failed to give me every support.

ACKNOWLEDGEMENT

Praises to God for giving me the patience, strength and determination to go through and complete my study. I would like to express my deep and sincere gratitude to my supervisor, Dr Alex Tze Hiang Sim. His wide knowledge and his logical way of thinking has been of great value for me. His understanding, encouragement and guidance have provided a good basis for the present thesis. I would like to express my appreciation to Dr. Mohd Shahizan Othman, Dr. Roliana Ibrahim, Dr Halina Mohamed Dahlan and Dr. Ab. Razak Che Hussin, for their support and guidance during the course of this study and Madam Lijah Rosdi for her friendly help and Madam Nor Asilkin Mohammad for her kind cooperation.

ABSTRACT

Huge databases are being used in organizations to store data. These databases contain hidden patterns which can be discovered and used in the organizations. In this project, we applied data mining techniques to uncover the patterns in the circulation database of UTM library. In order to discover worthwhile patterns we followed knowledge discovery process (KDD) to transform row data to suitable format. Weka machine learning software was applied to do the data mining task. In this project, we studied two association rules mining algorithms, Apriori and FPGrowth. The later was used to discover some patterns among borrowed books. These patterns which are presented in a list can be used to make recommendations to patrons who are searching for a certain topic based on items that previously were borrowed together. In addition, a novel rule matrix was presented to store the found rules for future use. Both the list for recommendation and rule matrix are useful to construct a recommender system for users of UTM library.

ABSTRAK

Sesebuah organisasi menggunakan pangkalan data yang besar untuk menyimpan data. Pangkalan data ini mengandungi pola tersembunyi yang boleh diterokai supaya dapat di implementasikan dalam organisasi tersebut. Dalam penyelidikan ini, teknik perlombongan data digunakan bagi mengenalpasti pola dalam pangkalan data sirkulasi Perpustakaan Sultanah Zanariah, UTM. Pola tersebut, dapat dikenalpasti melalui proses penemuan pengetahuan (KDD) bagi mengubah data mentah kepada format yang sesuai. Mesin pembelajaran perisian Weka, telah digunakan untuk melakukan tugas perlombongan data. Dalam penyelidikan ini, dua persatuan peraturan algoritma perlombongan, Apriori dan FPGrowth telah dikaji. Ia kemudiannya digunakan untuk mengenalpasti beberapa pola antara buku-buku yang dipinjam. Pola yang dibentangkan di dalam senarai, boleh digunakan sebagai cadangan kepada pengguna yang mencari topik tertentu berdasarkan bahan yang sebelum ini dipinjam bersama-sama. Di samping itu, matriks peraturan baru telah dibentangkan untuk menyimpan kaedah-kaedah yang didapati bagi kegunaan pada masa akan datang. Kedua-dua senarai cadangan dan matriks peraturan adalah berguna dalam membina satu sistem cadangan untuk pengguna Perpustakaan Sultanah Zanariah, UTM.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDIXES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Recommendation system is a system that provides information on services, movies, music, news, tools, movies, books and web pages to users. The system analyzes the existing data to work out the recommendation list. This study can be done on previous user requests or on previous related searches in an existing database. In fact the recommendation system searches the database, finds data relationships and presents it to users. As a result, users can find what they are looking for, in a shorter time and with a strikingly higher precision.

As the number of books in libraries increases, the quantity of available books in a certain topic also increases. Consequently, users might face loads of irrelevant data when searching for a book in a specific topic. They have to look at several books to find what they are seeking. Moreover, this strenuous toil would not always provide the users with the best available choice.

There are two common classifications for recommendation systems. First one is content-based approach. This technique focuses on content of items and their relation with each other. And preference-based approach, that uses collaborative filtering or other techniques, to find users with same interests, and make a recommendation list based on this (Huan-Ming, Li-Chuan *et al.* 2008). In next chapter we will talk about different kind of classifications and different approaches in detail.

Nowadays recommendation systems are used in e-commerce, web services and web searching. These systems help people to find what they are looking for, in a shorter time. Also it can show them the best choice among a variety of items. Hence, it can fulfill their satisfaction which is a goal for all organizations. In this project the effort is to propose a book recommendation system model for UTM's library as a case study. First, we will discuss about an appropriate way to find books which are related to each other. Next, we will propose a Recommendation model for users in UTM's library.

1.2 Problem Background

There are more than 350,000 books titles in UTM's library (UTM-library 2010), and even more as the time pass. The variety of items might confuse users when they are searching for a subject. In this case they may take several books out of shelves and after spending their time revising them, before choosing one and leaving the other books. Consequently, librarians need to spend some time to put those books back in shelves. This is time consuming for both users and library staffs. It will be great for library management and also users to have a system to help users find what they need in a shorter time with less effort.

One of problems that library management faces, is returning books back to shelves(JianWei and PingHua 2008). Commonly, librarians do the shelving which is very time consuming task. For example, in UTM's library 9 people do this duty 3 times a day and it takes about 2 hours each time. On the other hand, figures show that number of library users increased from 33,164 members in 2006(UTM-library 2007) to 42,534 members in 2009(UTM-library 2010). It shows that year by year the process of returning books to shelves will take more time, and requires more manpower. Also, books may be misplaced in shelves that can happen due to simple mistakes. In this case, finding such books is practically impossible for another user.

Another important matter for library managers is customer satisfaction. To gain that, managers can provide library with more effective services. But to increase customer satisfaction, commonly there is a strong need to additional staffs, which inevitably depends on more budgets. Even without adding services it is natural to add library staffs, because the number of users and visitors[1] is increasing (from 870,000 visitors in 2006 (UTM-library 2007) to 1,080,000 in 2009 (UTM-library 2010)). In this case library management needs to think about automatic services which do not need additional manpower.

One of the services that libraries can offer to their users could be a system that utilizes collective experience of others to suggest best item to other users. In this case, users would be grateful if they could use other users' experience for selecting books. Book recommendation systems in the library management play an important role to provide new services to patrons (Yongcheng, Jiajin *et al.* 2009). In addition, it can help to gain users satisfaction, without additional manpower. As a result, implementation of this system is satisfactory for both users and library management.

[1] Here visitor means the total entering and exiting people during a year. For example a student can visit library 300 times during a year which will increase the total amount of visitors by 300 in a year.

1.3 Problem Statement

There is a huge database consisting of approximately one million transactions since 2007, which needs an appropriate strategy to extract hidden knowledge with minimum loss. The main challenge in designing a recommendation system is the data mining part, including preprocess and association rules mining. Moreover, the design of this system is important itself. It should have a user friendly interface, to provoke patrons to use it. In addition, it should be fast enough to respond to users' query in a short time.

1.4 Project Objectives

1. To collect the information about borrowed book list and transactions from UTM's library.
2. To pre-process the data and prepare it for pattern discovery.
3. To study and choose an appropriate technique for pattern discovery.
4. To uncover transferable knowledge from the data in order to present it to library's members.

1.5 Project Scope

1. The study will focus on postgraduates' transactions in UTM's library.
2. The dataset will cover postgraduates' transaction during 2007- 2009.
3. The study will use Weka software.

1.6 Project Importance

As it was mentioned before, in recent years, the number of library visitors increased. To cover this demand, the number of staffs increased too (from 150 in 2006 (UTM-library 2007) to 170 in 2009 (UTM-library 2010)). More manpower means overload in costs, which managers try to avoid it. With implementation of Book Recommendation System, the library can provide users new service, without the need for additional staffs. And as well it can decrease the irrelevant books that a user takes out from shelves during a search for a certain topic, which will directly decrease the work hours of library staffs. As a result, Book recommendation system is a cheap, useful and important for UTM's developing library in IT era.

1.7 Summary

In this chapter, we discussed about the importance of recommendation systems importance in business. We explained how these systems work a total overview about content-based and preference-based approach for recommendation systems; and role of association rule mining in recommendation systems. Also, we explained some difficulties that library management faces to keep current services working and provide new ones. Then we tried to highlight the importance of Book recommendation system for UTM's library. We will talk about all these techniques and several important approaches in chapter 2.

REFERENCES

Agrawal, R., T. Imieli\, et al. (1993). *Mining association rules between sets of items in large databases.* Proceedings of the 1993 ACM SIGMOD international conference on Management of data. Washington, D.C., United States, ACM: 207-216.

Agrawal, R. and R. Srikant (1994). *Fast Algorithms for Mining Association Rules in Large Databases*. Proceedings of the 20th International Conference on Very Large Data Bases, Morgan Kaufmann Publishers Inc.: 487-499.

Balabanović and Y. Shoham (1997). "Fab: content-based, collaborative recommendation." Commun. ACM 40(3): 66-72.

Breese, J. S., D. Heckerman, et al. (1998). *Empirical Analysis of Predictive Algorithms for Collaborative Filtering*. 14th Conf. Uncertainty in Artificial Intelligence,: 21.

Choonho, K. and K. Juntae (2003). *A recommendation algorithm using multi-level association rules*. Web Intelligence, 2003. WI 2003. Proceedings. IEEE/WIC International Conference on.

H.Witten, I. and E. Frank (2005). *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann Publishers.

Han, J., J. Pei, et al. (2000). *Mining frequent patterns without candidate generation.* SIGMOD Rec. 29(2): 1-12.

Han, J., J. Pei, et al. (2004). *Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach.* Data Mining and Knowledge Discovery 8(1): 53-87.

Huan-Ming, C., W. Li-Chuan, et al. (2008). *A Study on the Comparison between Content-Based and Preference-Based Recommendation Systems*. Semantics, Knowledge and Grid, 2008. SKG '08. Fourth International Conference on.

JianWei, L. and C. PingHua (2008). *The application of Association rule in Library system*. Knowledge Acquisition and Modeling Workshop, 2008. KAM Workshop 2008. IEEE International Symposium on.

JingHui, W., L. Qiang, et al. (2008). *Clustering Technology Application in e-Commerce Recommendation System*. Management of e-Commerce and e-Government, 2008. ICMECG '08. International Conference on.

Krulwich, B. and C. Burkey (1997). *The InfoFinder agent: learning user interests through heuristic phrase extraction.* IEEE Expert 12(5): 22-27.

Lang, K. (1995). *Newsweeder: Learning to filter netnews*. Proceedings of the 12th ICML. San Francisco, Morgan Kaufmann: 331-339.

Linden, G., B. Smith, et al. (2003). *Amazon.com recommendations: item-to-item collaborative filtering.* Internet Computing, IEEE 7(1): 76-80.

Longjun, H., D. Liping, et al. (2008). *A Personalized Recommendation System Based on Multi-agent.* Genetic and Evolutionary Computing, 2008. WGEC '08. Second International Conference on.

Lovins, J. and B. R. Date (1968). *Development of a stemming* Mechanical Translation and Computational Linguistics 11(1).

Pazzani, M. J. and D. Billsus (2007). *Content-based recommendation systems.* The adaptive web: methods and strategies of web personalization, Springer-Verlag: 325-341.

Porter, M. F. (1980). *An algorithm for suffix stripping*. Program: electronic library and information systems 40.

Resnick, P., N. Iacovou, et al. (1994). *GroupLens: an open architecture for collaborative filtering of netnews*. Proceedings of the 1994 ACM conference on Computer supported cooperative work. Chapel Hill, North Carolina, United States, ACM: 175-186.

Sarwar, B., G. Karypis, et al. (2000). *Analysis of recommendation algorithms for e-commerce*. Proceedings of the 2nd ACM conference on Electronic commerce. Minneapolis, Minnesota, United States, ACM: 158-167.

Sobecki, J. and K. Piwowar (2009). *Comparison of Different Recommendation Methods for an e-Commerce Application*. Intelligent Information and Database Systems, 2009. ACIIDS 2009. First Asian Conference on.

UTM-library. (2007, 19/8/2010). *Basic Information of UTM's Library in year 2006.* Basic inforamion of library Retrieved 21/8/2010, 2010, from

http://portal.psz.utm.my/psz/images/stories/2009/about/factsandfigures/06basicinfo.pdf.

UTM-library. (2010, 19/8/2010). *Basic Information of UTM's Library in year 2009.* Basic inforamion of library Retrieved 21/8/2010, 2010, from http://portal.psz.utm.my/psz/images/stories/2009/about/factsandfigures/06basicinfo.pdf.

Wang, Y., M.-Y. Kan, et al. (2004). *LyricAlly: automatic synchronization of acoustic musical signals and textual lyrics*. Proceedings of the 12th annual ACM international conference on Multimedia. New York, NY, USA, ACM: 212-219.

Wild, M. (2003). *From Usenet to CoWebs - Interacting with social information spaces*. Educational Technology & Society 6(4): 164-165.

Xuejun, Y., Z. Hongchun, et al. (2009). *ARTMAP-Based Data Mining Approach and Its Application to Library Book Recommendation*. Intelligent Ubiquitous Computing and Education, 2009 International Symposium on.

Yongcheng, L., L. Jiajin, et al. (2009). *A Privacy-Preserving Book Recommendation Model Based on Multi-agent*. Computer Science and Engineering, 2009. WCSE '09. Second International Workshop on.

Zhen, Z. and W. Jing-Yan (2007). *Book Recommendation Service by Improved Association Rule Mining Algorithm.* Machine Learning and Cybernetics, 2007 International Conference on.