

LB/DOA/92/2019
IT 01/221

An Application of Data Mining for a Library Management System

LIBRARY
UNIVERSITY OF MORATUWA, SRI LANKA
MORATUWA

T.A.U.I. Ranaweera

169327K

004 "19"

004 (043)

University of Moratuwa



TH3897

Dissertation submitted to the Faculty of Information Technology, University of Moratuwa, Sri Lanka for the partial fulfilment of the requirements of the Master of Science in Information Technology

February 2019

TH3897

+
CD-2000

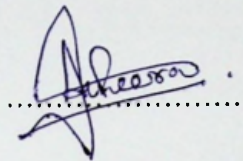
ii

TH3897

Declaration

We declare that this thesis is our own work and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education. Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.

T A U I Ranaweera

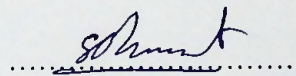
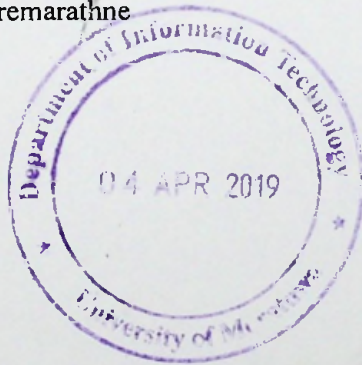


Signature

Date: 04/04/2019

Supervised by

Mr. Saminda Premarathne



Signature

Date: 04/04/2019



Dedication

Completing this research would be impossible without the support and guidance of several friends and people who were around me during this entire period. First of all, I would like to express my deepest gratitude to my Supervisor, Professor [Name], for his expertise and providing excellent guidance throughout the project. The knowledge I gained from you made me strong enough to complete this challenge. Furthermore, my gratitude goes to the broader panel of M.Sc. degree programmes and I appreciate the effort they took to improve our subject knowledge throughout these two years.

Obtaining relevant data for the analysis was vital and I would like to thank the [Name] Committee, [Name] of University of [Name] for granting approval to access the [Name] database. It is with great honor, I acknowledge the cooperation of the [Name] researchers who provided necessary information on the [Name] branch of University of [Name]. Without your support, my efforts could have been in vain.

At last, but not least, I would like to thank my family members for being supportive to always. My family is the greatest strength of my life and I have my words to express my gratitude towards them.

To my family

Acknowledgement

Completing this research would be impossible without the support and guidance of whole bunch of people who were around me during this entire period. First of all I would like to express my deepest gratitude to Mr. Saminda Premarathne for being my supervisor and providing excellent guidance throughout the project. The knowledge I gained from you made me strong enough to complete this challenge. Furthermore, my gratitude goes to the lecturer panel of M.Sc. degree programme and I appreciate the effort they took to improve our subject knowledge throughout these two years.

Obtaining relevant data for the analysis was vital and I would like to thank Mr. Ananda Karunaratne, Librarian of University of Ruhuna for granting approval to access data from the library database. It's with great honor, I acknowledge the cooperation of Mr. Nimal Hettiarachi who provided necessary information on the library system of University of Ruhuna. Without your support, my efforts could have been in vain.

At last, but not least, I would like to thank my family members for being supportive as always. My family is the greatest strength of my life and I have no words to express my gratitude towards them.

List of Abbreviations

DDC – Dewey Decimal Classification

EG - Faculty of Engineering

HS- Faculty of Humanities and Social Science

MD - Faculty of Medicine

MF- Faculty of Management

PCA –Principle Component Analysis

SC – Faculty of Science

Abstract

Data mining has become an emerging concept in today's world. With the development of the technology in every field, a large amount of data can be collected and stored very easily. The challenge is to analyse these data from traditional analysing techniques. The concept of data mining; which enables users to analyze data and draw conclusions through clearly defined procedures comes into play as a solution for this matter.

The purpose of this research is to test the adequacy of data mining techniques to improve the library usage in Sri Lankan State Universities. A data warehouse was designed and implemented using the most important variables in the raw dataset. After cleaning and preprocessing of data, association rule mining and clustering were basically used in the data analysis phase. Interesting rules were identified and the said results were used in the next stage of the study. A Book Recommendation System was implemented in Java based on the results obtained in the previous stage. The system enables users to select library materials according to their prior borrowing patterns. My SQL, R and Java were basically used in the analysis.

This research will be beneficial to enrich the library usage in state universities in Sri Lanka as well as the researchers those who are interested in Data Mining.

Table of Contents

| | |
|---|----|
| Chapter 1..... | 1 |
| Introduction..... | 1 |
| 1.1. Introduction..... | 1 |
| 1.2. Background..... | 1 |
| 1.3. Statement of Research Problem | 2 |
| 1.4. Aim | 3 |
| 1.5. Specific Objectives | 3 |
| 1.6. Solution..... | 3 |
| Chapter 2..... | 4 |
| Literature Review..... | 4 |
| 2.1. Knowledge Discovery Process | 4 |
| 2.2 Data Mining..... | 5 |
| 2.3 Applications of Data Mining | 6 |
| 2.4 Data Mining in Library Management Systems | 7 |
| 2.5 Association Rule Mining for Library Systems | 9 |
| 2.6 Clustering for Library Systems..... | 11 |
| 2.7 Book Recommendation Systems | 12 |
| Chapter 3..... | 14 |
| Technology Adapted..... | 14 |
| 3.1 Theoretical Background..... | 14 |
| 3.1.1. Association Rule Mining..... | 14 |
| 3.1.1.1. Rule Assessment Measures..... | 15 |
| 3.1.2. K-means Clustering | 17 |
| 3.1.2.1. Cluster Validation..... | 17 |
| 3.2. Technologies Used for the Application | 18 |
| 3.2.1. Java..... | 18 |
| 3.2.2. R Statistical Software | 18 |
| 3.2.3. My SQL | 18 |
| Chapter 4..... | 19 |
| Methodology..... | 19 |
| 4.1 Data Selection Procedure..... | 19 |

| | |
|--|----|
| 4.2. Data Cleaning and Preprocessing | 22 |
| 4.3. Building the Data Warehouse..... | 23 |
| 4.4. Creation of the Connection between R and Other Sources..... | 24 |
| 4.5. Establishment of the Connection between R and MySQL..... | 25 |
| 4.6. Establishment of the Connection between R and Java..... | 25 |
| Chapter 5..... | 26 |
| Analysis and Design | 26 |
| 5.1. Identifying Borrowing Patterns of Patrons | 26 |
| 5.1.1 Identify Frequent Patterns of Borrowing Books by the Users..... | 26 |
| 5.1.1.1 Data Preparation | 26 |
| 5.1.1.2 Test for Statistical Significance of Rules..... | 27 |
| 5.2.2 Identify User Behavior when Lending Books | 28 |
| 5.2. Clustering Library Users with k-means Clustering..... | 30 |
| 5.2.1 Clustering the Library Users without Dimension Reduction..... | 30 |
| 5.2.2 Clustering the Library Users by Dimension Reduction Using Principle Components (Two Step Clustering) | 31 |
| Chapter 6..... | 33 |
| Implementation of the Book Recommendations System..... | 33 |
| 6.1. Providing Recommendations for the Users without Logging-in | 33 |
| 6.1.1. Recommendations Based on the Support Value of the Dewy Numbers of Borrowed Books | 33 |
| 6.1.1.1. Recommendations Based on Association Rules between Faculty and Dewy Numbers of Borrowed Books. | 34 |
| 6.1.1.2. Recommendations Based on Association Rules between Dewy Numbers of Borrowed Books | 35 |
| 6.2 Providing Recommendations for the Users after Logging-in | 37 |
| 6.2.1. Recommendations Based on the Most Popular Books in the Cluster..... | 37 |
| 6.2.2. Recommendations Based on Association Rules between Dewy Numbers of Borrowed Books within the Cluster | 38 |
| Chapter 07..... | 40 |
| Results of the Study | 40 |
| 7.1 Descriptive Analysis..... | 40 |
| 7.1.1 Composition of Student Users in 2013 | 40 |
| 7.2 Association Rule Mining | 42 |
| 7.2.1 Mining Associations between the Dewy Numbers of Borrowed Books | 42 |

| | |
|--|----|
| 7.2.2 Association Rules between Faculty and Dewey Decimal Number | 50 |
| 7.3 Clustering the Library Users (Students)..... | 52 |
| 7.3.1 Clustering the Library Users in the Faculty of Science | 52 |
| 7.4 Book Recommendation System | 54 |
| 7.4.1 Main Interface..... | 54 |
| 7.4.2 Interface after Logging-in..... | 55 |
| 7.5 Evaluation | 56 |
| Chapter 7..... | 57 |
| Conclusion & Further Work | 57 |
| 7.1 Conclusions..... | 57 |
| 7.2 Limitations | 59 |
| 7.3 Future Work..... | 59 |
| References..... | 61 |



List of Tables

| | |
|--|----|
| Table 4.1: Table of "dewycodes" | 20 |
| Table 4.2: Tables Selected from the Database for the Study | 20 |
| Table 4.3: Summary Table of Selected Variables..... | 21 |
| Table 4.4: Summary of New Variables..... | 22 |
| Table 5.1: Arrangement of Data in the Database..... | 26 |
| Table 5.2: Layout of the Rearranged Data..... | 27 |
| Table 7.1: Summary of Association Rules..... | 45 |
| Table 7.2: Summary of Association Rules..... | 46 |
| Table 7.3: Summary of Association Rules..... | 47 |
| Table 7.4: Summary of Association Rules..... | 48 |
| Table 7.5: Summary of Association Rules..... | 50 |
| Table 7.6: Summary of Association Rules..... | 51 |

List of Figures

| | |
|---|----|
| Figure 2.1: Knowledge Discovery Process..... | 4 |
| Figure 4.1: Implementation of "transactional" Table..... | 24 |
| Figure 5.1: Flow Chart of Generating Association Rules..... | 28 |
| Figure 5.2: Flow Chart for Generating Associations Considering More Than One Variable . | 29 |
| Figure 5.3 : Flow Chart for the Process of Clustering..... | 31 |
| Figure 5.4: Flow Chart for the Process of Two Step Clustering..... | 32 |
| Figure 6.1 : Process for the Method 01..... | 34 |
| Figure 6.2 : Process of the Method 02..... | 35 |
| Figure 6.3: Process of the Method 03..... | 36 |
| Figure 6.4: Process of the Recommendations after Logging-in..... | 37 |
| Figure 6.5: Process of Recommendations when Searching for a Certain Book..... | 38 |
| Figure 7.1: Composition of Library Users (Students)..... | 40 |
| Figure 7.2: Faculty-wise Transactions according to the Month..... | 41 |
| Figure 7.3: Item Frequency Plot..... | 43 |
| Figure 7.4: Graph Based Visualization Rules..... | 44 |
| Figure 7.5: Plot of Number of Clusters vs. Within Groups Sum of Squares..... | 52 |
| Figure 7.6 : Cluster Validation..... | 52 |
| Figure 7.7: Main Interface of the Book Recommendation System..... | 54 |
| Figure 7.8: Interface after Logging-in to the User Account..... | 55 |

Chapter 1

Introduction

1.1. Introduction

Data mining is an emerging field and there is no concrete definition for it. According to [1], the term data mining refers to “the process of analyzing data from different perspectives and summarizing it into useful information by means of a number of analytical tools and techniques, which in turn may be useful to increase the performance of a system.”

Nowadays libraries make use of data mining techniques to improve their excellence. Librarians are increasingly finding the importance of obtaining a comprehensive and integrated view of the library operations and the services it provides. Bibliomining can help libraries to reveal the future behavior of users by predicting the circulation of documents and trends of popular subjects.

This research aims to identify the patterns of user behavior of a library system in a state university in Sri Lanka which can help to enhance the library experience of users, assist the library management in making decisions, and understand information needs of the users.

1.2. Background

University libraries in Sri Lanka have been operated since 1921 and they have developed with the changes taking place at various stages in the evolution of the present university system. However, library usage of undergraduates in most of the Sri Lankan universities in recent years is not satisfactory compared to other universities in the world. According to the library review report of University of Rajarata [5], unawareness on facilities available at the university is one of the fundamental reasons for the decrease of library usage and this is a common feature in almost every Sri Lankan university [6][2]. Nowadays most libraries have OPAC (Online Public Access Catalogue) facility which displays all the holdings of the Library such as books, thesis, e-resources, journals etc... However, it seems that students do not use these facilities well. Outdated editions and inadequate number of copies in certain subject areas are also considered responsible for this decreasing trend [5].

Furthermore, lack of statistical information on student requirements makes it hard for the library staff to purchase most needed books [6].

Considering these facts, there is no doubt that library services should be modified and promoted within the university community. In this case, a proper analysis of the past behavior of library usage should be done at least once a year. As large numbers of transactions are occurred in a library within a year, Data mining can be used effectively to analyze data. For an example, association rule mining can be used to find out patterns of lending books and most frequent books which are borrowed together [7] would be beneficial for the library management to arrange book shelves in the library and to get an idea about the books which are borrowed more frequently in each faculty to make correct decisions in acquisitions. Clustering is another practice in data mining that can be used to group borrowers who have similar behaviors [7]. Association rule mining and clustering are popular data mining techniques used in analyzing library transaction data [8], [9].

In order to enhance library usage among the university community, introducing a book recommendation system is a timely need. Based on the results obtained from the statistical analysis using data mining techniques, a book recommendation system can be implemented [7]. Then users are allowed to get more detailed information on library collection and they can find out what other users tend to borrow. Furthermore, new arrivals of books can be promoted through this system and these advancements will make the library an attractive place for the readers.

1.3. Statement of Research Problem

Libraries in state universities provide their service in an efficient and useful manner for decades. Although library management has gathered data about their collections for years, they have rarely used for better decision making. There is no doubt that library services should be modified and promoted within the university community. In this case, identifying and analyzing the patterns of user behavior in past years is a must to obtain more precious results to enhance the service in university libraries by improving library collections, staff training and equipment. Furthermore, to enhance library usage among the university community, introducing a book recommendation system is a timely need.

1.4. Aim

Applying and testing the suitability of data mining techniques to enhance the performance of library management systems.

1.5. Specific Objectives

1. Analyze and identify interesting patterns on library usage of a state university to assist the library management in decision making.
2. Develop a book recommendation system based on data mining techniques to enhance the library experience of users. (to enhance the effectiveness of OPAC)
3. Use statistical concepts together with computer science to perform comprehensive analysis and draw conclusions.
4. Gain knowledge of statistical concepts and techniques which are used in data mining applications.

1.6. Solution

As a solution for the above objectives, data will be analyzed with the use of data mining techniques such as association rule mining and clustering. It is expected to follow KDD (Knowledge Discovery in Database) process here. Apriori algorithm will be used to extract association rules which gives more valuable information on library transitions. K-means clustering and two-step clustering will be applied to group library users with common borrowing behaviors. R-statistical software will be used to analyze data. Based on the results, a book recommendation system will be implemented JAVA programming language to enhance the library experience of users. When library users search for books using this system, recommendations will be given according to their prior transactions.

This thesis consists of seven chapters. Second chapter explains the background information of the research based on the literature review. Theoretical background and the required technologies for the research is described in third chapter. In the fourth chapter, methodology of the study will be explained and design of the implementation will be explained in fifth chapter. Sixth chapter describes the implementation of the application including flow charts. A brief summary of the research and further work is mentioned in chapter seven under discussion.

Literature Review

The concept of data mining has made a revolution in modern world over the last two decades. Advancements in information technology have made storing large amount of data in databases an effortless task. However, in order to use these data in more productive way, it is essential to have effective tools to extract knowledge.

2.1. Knowledge Discovery Process

Retrieving knowledge (knowledge discovery) from data is an essential task for decision making in any circumstance and data mining is the key part of Knowledge Discovery in Database (KDD) process. The Knowledge Discovery in Database (KDD) process consists of several stages starting from raw data into knowledge and graphical representation [3] of KDD is shown below.

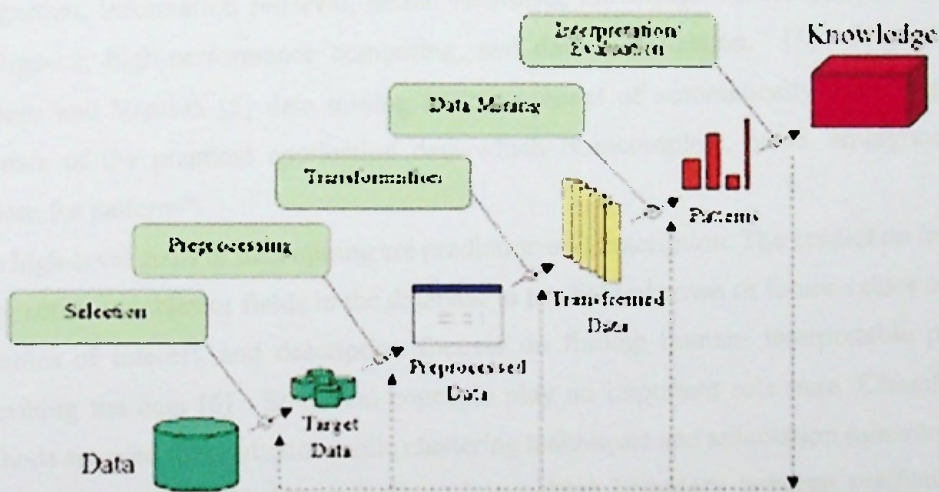


Figure 2.1: Knowledge Discovery Process

KDD process is an iterative process. First step of KDD process is the selection of relevant data according to the nature of the problem since the source of data (most of the time, transaction databases) may consists of large amount of data. Then the target data should be preprocessed to eliminate outliers, and to make the data consistent and clean. Data can be

transformed in the next step of KDD process. At this point data mining techniques can be applied to discover knowledge such as trends, patterns and characteristics. Evaluation and interpretation of identified patterns must be done in the next stage with the involvement of user. Simply the KDD process involves the evaluation and possibly interpretation of the patterns to make the decision of what qualifies as knowledge [4]. In each stage of the KDD process statistics plays an important role.

2.2 Data Mining

As mentioned earlier, data mining is an emerging field and there is no concrete definition for it. According to [1], the term data mining refers to "the process of analyzing data from different perspectives and summarizing it into useful information by means of a number of analytical tools and techniques, which in turn may be useful to increase the performance of a system." Another definition is that "Data mining is a multidisciplinary field, drawing work from areas including database technology, machine learning, statistics, pattern recognition, information retrieval, neural networks, knowledge-based systems, artificial intelligence, high-performance computing, and data visualization." [3]. Regarding to Romero and Ventura [5] data mining is "the process of automatically searching large volumes of the practical application data which is incomplete, noise, ambiguous and random for patterns".

Two high-level goals of data mining are prediction and description. The prediction involves using some variables or fields in the database to predict unknown or future values of other variables of interest, and description focuses on finding human- interpretable patterns describing the data [6]. Statistical concepts play an important role here. Classification methods are used for prediction while clustering techniques and association rule mining are used for description [7]. However there is no sharp boundary between prediction and description [6].

2.3 Applications of Data Mining

Nowadays data mining techniques are applied in numerous areas such as Medical, Marketing, Banking, Education, Agriculture, Astronomy etc. In the Banking sector customer retention, fraud prevention and fraud detection can be accomplished by using decision trees, logistic regression and clustering [8]. Almasoud et al. [9] describe recent developments in data mining applications and techniques in several fields. According to their research, educational data mining, healthcare data mining, web data mining, data stream mining and big data mining are five most active data mining research fields.

Slater et al. [10] have discussed the appropriateness of applying data mining techniques in educational data. Several tools such as Text mining, Process and sequence mining and Bayesian knowledge tracing (BKT) in analyzing educational data have been discussed there [10].

In the Medical sector, evaluating the effectiveness of a treatment, detecting means to deliver better medicine with low cost by mining treatment record data, passing warnings regarding to a severe disease by analyzing patient records can be accomplished with the use of data mining [11]. In the field of molecular biology, data mining has become a major analysis tool in analyzing DNA sequence patterns. DNA sequence patterns can be analyzed with the help of data mining techniques such as neural networks, classification and clustering [11]. In addition, this concept is widely applied in disease diagnosis, disease related factor analysis, disease prediction and analyzing physiological parameters of patients [12].

Moreover, data mining is widely used in market researches. After analyzing the history of consumer's likes/ dislikes and the purchases, new products can be proposed to match their preferences. Most of the online purchasing websites are equipped with this technique and Amazon¹ is one of the best examples. When customer log-in to his/her account, recommendations are displayed based on the behavior of prior transactions and then the customer can select the best product available to them. Furthermore, LinkedIn and Netflix are also commercial websites where recommendation systems are used to accomplish customer needs.

¹ <http://www.amazon.com/>

Association rule mining is widely used in supermarket data to discover customers' purchasing habits. This is called as "market basket analysis". Kaur.D and Kaur.J [13] have analyzed supermarket data using apriori algorithm and eclat algorithm and proved that both the algorithms work better in providing best results [13]. It is emphasized that apriori algorithm serves better for small datasets whereas eclat algorithm works better for large datasets. More often the layouts of the supermarkets are arranged based on the results of market basket analysis.

2.4 Data Mining in Library Management Systems

Library Management Systems are another field where data mining is widely used. Large number of transactions are occurred in any library within a day and an explosive growth of library data is a huge challenge for the libraries. So the work of the present library professionals is increasing in satisfying the needs of the users and for better utilization of resources available in the library justifying the huge expenditure on online resources [14]. Data mining techniques can be applied to overcome the difficulties in handling such data by identifying patterns of user behavior and making necessary recommendations to the management for better decision making.

The term "Bibliomining" introduced by Scott Nicholson and Stanton in 2003[15] is the process of applying data mining techniques to extract patterns of behavior from library databases. It is derived or a combination from the terms "bibliometrics" and "data mining". Bibliomining can help libraries to reveal the future behavior of users by predicting the circulation of documents and trends of popular subjects. These predictions can then be used to make decisions on the acquisition and allocation of funds. Furthermore, the knowledge gained by the circulation of documents can be further used to decide on the space allocation for different resources in the library, to decide on the opening hours and the number of librarians, etc.[16],[17].

Dwivedi and Bajpai[18] pointed out possibilities of data mining in the field of library and information science such as Classification: to develop system that will replace the manual classification with the automatic classification of library contents, Sequence analysis: to identify unlinked documents that users are likely to want to read together, Link analysis: to

identify frequently linked-to-documents at the top of a list or to identify documents that are associated with each other, Clustering: to find natural grouping in data.[18]

Chang and Chen [19] had mined library data and an upward trend of reading preferences of readers in reading digital publications had been noticed. Five clusters of users had been identified and frequency of graduates and associate researchers borrowing multimedia data, such as CDs, VCDs, etc., was much higher [19] .

The research paper presented by Littman and Connaway [20] pointed out the adoption of electronic books (e-books) using data mining. Out of 7,880 titles that were available in both print and e-book format at the Duke University Libraries the usage of e-books was greater than the printed books [20].

However, automating the library or developing digital library is not the only solution unless it is not able to explore the hidden information from large database [21]. In that case, data mining comes into play. In their research, they are discussing advantages as well as disadvantages of clustering and other statistical techniques in library database.

Text mining is another approach widely used in text summarization, digital libraries, life science, social media and business intelligence. [22] have explained the appropriateness of applying text mining in digital libraries. GATE, Net Owl and Aylien are frequently used tools for text mining in digital libraries. Selection and use of right techniques and tools according to the domain help to make the text mining process easy and efficient.

In the research paper of Zhang [23], it is described technologies related with data mining, specially for digital libraries. Association analysis, clustering, concept description, time series patterns and text mining are some of the techniques discussed in the paper. Moreover, analysis regarding to low utilization readers in Linyi university library has been carried out using decision tree algorithm in this study [23].

Data mining on circulation data can be used to perform monthly circulation variation tendency analysis. It provides a clear picture on variation of book circulations over the

months of a year [24]. Eg: before examination periods, higher number of books might be issued to the students and issuing rate may be low when there are holidays. Hence by analyzing the ratio of issued books at regular intervals, librarian can plan the purchasing, maintenance and enhancement of resources.

Kensuke, Mitsuro & Toshiro [25] classified books in Kyushu University Library according to Nippon Decimal Classification (NDC), and investigated the turnover rate for each category to find useful information for book selection [25]. According to the mining results, it has been identified that the library was conducting the most appropriate book selections.

With the development of modern technologies, it becomes increasingly important to find a way to effectively utilize and management the intelligent library. For this reason, Xie.F[26] have designed an intelligent library management system and implemented based on RFID / GRPS [26]. The final experimental results suggested that the proposed approach is feasible and correct.

2.5 Association Rule Mining for Library Systems

Association rule mining is a widely used data mining technique to mine library data to find out borrowing patterns of users. A rule consists of a left hand side proposition (the antecedent or condition) and a right hand side. Both left and right hand side consists of Boolean statements. The rule states that if the left hand side is true, then the right hand side is also true. A probabilistic rule modifies this definition so that the right hand side is true with probability p , given that the left hand side is true [27]. Patterns of these rules are frequently used for decision making.

A model was created for the user's behavior in Suan Sunandha Rajabhat University by Kanyarat & Kunyanuth [29] using association rules. Fourteen rules were identified in their experiment using apriori algorithm in WEKA (the Waikato Environment for Knowledge Analysis) which is a collection of machine learning algorithms to analyze data set for data mining tasks[28]. Furthermore results were tested using a testing data set. [29].

Uppal and Chindawani [24] had found out that readers who borrow books of biography and Chinese literature at the same time will borrow books of common foreign languages using

association rules [24]. They also suggested to arrange the layout of the library to match their findings, so that readers can find required books easily. Furthermore they propose in implementing a book recommendation system to readers based on association rules.

Mehta et al. [30] explain how Apriori algorithm can be applied on the university's library transactional database in order to find out the frequent book items and generate rules on these book items so as to predict the book borrowing behavior of the students [30]. Minimum support and confidence are key factors that they have considered in identifying most interesting association rules through out the research. It further explains how incremental mining when incorporated by adding five more transactions to the original set of ten transactions changes the number of frequent item-sets and association rules generated by the algorithm.

In order to assist the library management in acquisitions and organizing the arrangement of the library, association rule mining is widely used. Anuradha.T et al.[31] have mined frequent patterns and association rules on 2000 university library records to predict frequently borrowed books based on many factors such as their subject requirements, number of books issued, and duration for each book [31]. Improving the performance of the library by avoiding the delay, maintaining sufficient number of highly required books according to the current subjects of the different streams are the achieved objectives of the study.

Yan.L et al.[32] have studied the application of association rule mining for a university management system [32]. It has been showed that association rule mining in college management system was more realistic and reasonable, which has been solved a series of problems in college caused by diversification of data resource data in the resource release difficult.

Adoption of association rule mining for library management systems has been compared with other approaches in several researches. According to the study of Bansal.M et al [33] , several important information can be extracted from real-time database of university libraries using association rule mining [33]. Attained results have been compared with the results of SQL based mining of the same data set. It has been proved that association rule

mining can be effectively applied to find the relationships between different factors compared with other tested methods.

Nowadays association rule mining is applied in developing book recommendation systems. Joshua et al. [34] have developed a book recommendation system using frequent patterns and association rules. Frequent patterns of borrowed books have been extracted using the Frequent Pattern growth algorithm and generated results have been used in recommending books for library users. Furthermore, findings of the study have been supported the library management in making decisions more effectively [34].

2.6 Clustering for Library Systems

Presently educational institutions compile and store huge volumes of data such as student enrolment and attendance records, as well as their examination results. This rapid growth in educational data points to the fact that distilling massive amounts of data requires a more sophisticated set of algorithms. Clustering is one such preprocessing algorithm in Education Data Mining [35].

Clustering is a technique of “partitioning the points into natural groups called clusters, such that points within a group are very similar, whereas points across clusters are as dissimilar as possible”[36]. In the paper of Peter & Jan in 2012, they have explained an application of clustering methods on real data from a public library to obtain behavioral patterns of representative users to achieve more efficient management of public libraries[17]. K-means clustering had been used to identify groups with similar characteristics.

In the case study which was done by Karno, Noordin, Talib and Rahman in 2009 had proposed the implementation of bibliomining to the library data as a tool to assist library management in making decisions. Library users had been clustered according to their characteristics and borrowing trends, popular subjects had been observed in each cluster [16].

Uppal and Chindawani in their research, clustered readers in a library according to the grades and major and by the nature of books they borrow. They had observed that the freshman usually borrows Literary and Language [24].

Sahoo and Mishra [14] have used K-means algorithm of the cluster analysis for a library management system on various characteristics of readers' grades and departments. The cluster results have shown that the library procurement department shall add the books of English reading materials, computational linguistics, computer operation, social-romantic novels, etc. so as to satisfy the demand of the student readers [14].

2.7 Book Recommendation Systems

A recommendation system can be implemented using numerous techniques. Random prediction algorithm, Frequent sequences, Collaborative filtering algorithms (CF) and Content based algorithms are some of the algorithms which are used in recommender systems [37]. Among them Collaborative filtering algorithms (CF) and Content based algorithms are widely used. "CF is based on the premise that users looking for information should be able to make use of what others have already found and evaluated"[38] and Content based algorithms approaches based on "finding correlations between content of items as opposed to correlation between users as is the case in CF approaches. CB approach is employed where the items can be evaluated (or rated) by keywords such as textual documents and web pages (HTML)"[38].

To enhance the library experience of library users, introducing a book recommendation system is one of the best options. Rajagopal and Kwan [39] describe a theoretical design of a library recommendation system using k-means algorithm with subject headings of borrowed items as the basis for generating pertinent recommendations. A data warehouse was designed in their study and two clusters were identified based on data in the data warehouse. Based on the results, they have evidenced that extraction of user profiles into a data warehouse and implementing a library recommendation system using k-means clustering was feasible[39].

Two phase data mining recommendation system for a digital library was proposed by Chia-Chen and An-Pin [40]. In their research, users had been clustered using "Ant Colony Optimization Algorithm" and rules had been built by mining with association rules in each cluster to provide recommendations[40]. Uppal and Chindawani [24] pointed out the appropriateness of applying association rule mining in identifying frequent patterns and rules in implementing a book recommendation system[24].

Recommendation systems has been proposed using combination of collaborative filtering and association mining in such a way that collaborative filtering is used for finding similarity between items which would help the system to recommend items and association mining is used for filling the vacant ratings where necessary [41]. According to the said research, it is obvious that the problem of data sparsity can be solved by combining the collaborative-based filtering and association rule mining to achieve better performance.

Data mining has marked a revolution in many fields within a short period of time. Various software such as WEKA has been introduced to accomplish data mining tasks. Furthermore, this is a new research area and web sites like Keggale conduct competitions to enhance the popularity of this field. Even though introducing these novel concepts is a challenging task, it is essential to make use of them when necessary, for the betterment of the whole society.

Technology Adapted

3.1 Theoretical Background

3.1.1. Association Rule Mining

In this research most of the results will be based on association rule mining. When recommending books, it is important to identify which books are more likely to borrow together, what are the most frequently borrowed books and so on. These results may be more beneficial for library management in their decision making.

Association rule mining is widely used in discovering interesting patterns in a large database. Let $I = \{x_1, x_2, x_3, \dots, x_m\}$ be set of elements called *items* (For example, the collection of all books in a library) . A set $X \subseteq I$ is expressed as an *itemset*. Furthermore, $T = \{t_1, t_2, \dots, t_n\}$ can be denoted as a set of elements known as transaction identifiers or *tids* (ex: set of all library users in the library) and $T \subseteq T$ is called as a *tidset*. A *transaction* can be in the form of (t, X) , where $t \in T$ is unique transactions identifier and X is an itemset.[36]

An association rule is in the form of $X \rightarrow Y$, where X and Y are itemsets such that $X, Y \subseteq I$ and $X \cap Y = \phi$ [36]. Left hand side of the rule is known as the antecedent (or condition) and the right hand side is called as consequent. *Support* of a rule is defined as number of transactions which both X and Y appear together as subsets. The conditional probability that a transaction contains Y given that it contains X is defined as the *confidence* of a rule. A rule is *frequent* if the *support* of itemset XY ($X \cup Y$) is greater than or equal to the *minimum support*². Furthermore, a rule is strong if the *confident* of the rule is greater than or equal to the *minimum confidence*³.

Detection of frequent item sets is the first step of generating association rules with frequent and high confidence. Several algorithms have been introduced to find out frequent item sets and "Apriori" algorithm is one of the widely used algorithms. This algorithm is more

² Minimum support is specified by the user

³ Minimum confident is specified by the user



appropriate when the number of transactions as well as the number of features in transactional data are extremely large. Starting from the empty set, it employs a level-wise exploration of all possible itemsets in I and prunes all the supersets of any infrequent candidate (no superset of an infrequent itemset can be frequent).[36].

Subsequently, using the collection of frequent itemsets (F), association rules are generated and the relevant algorithm is given below.[36]

associatonRules (F , *minimum confidence*):

for each $Z \in F$, *such that* $|Z| \geq 2$ **do**

$\mathcal{A} \leftarrow \{X | X \subset Z, X \neq \emptyset\}$

While $\mathcal{A} \neq \emptyset$ **do**

$X \leftarrow$ *maximal*⁴ *element in* \mathcal{A}

$\mathcal{A} \leftarrow \mathcal{A} \setminus X$ // *remove* X *from* \mathcal{A}

$c \leftarrow \text{sup}(Z) / \text{sup}(X)$

if $c \geq$ *minimum confidence* **then**

print $X \rightarrow Y, \text{sup}(Z), c$

else

$\mathcal{A} \leftarrow \mathcal{A} \setminus \{W | W \subset X\}$ //*remove all subsets of* X *from* \mathcal{A}

3.1.1.1. Rule Assessment Measures

Supposing X and Y be two itemsets of dewy numbers of borrowed books, following measures can be obtained.

1. **Support** - number of transactions that contain both X and Y

⁴ X is "maximal" if all of its supersets are not frequent

$$\text{sup}(X \rightarrow Y) = \text{sup}(XY) = |t(XY)|$$

$$\text{rsup}(X \rightarrow Y) = \text{rsup}(XY) = \frac{|t(XY)|}{|D|}$$

where $|D|$ is the total number of transactions

Rules consisted with higher support were interested.

2. **Confident** – the conditional probability of that a transaction contains the consequent Y given that it contains the antecedent X .

$$\text{conf}(X \rightarrow Y) = P(Y|X) = \frac{\text{sup}(XY)}{\text{sup}(X)}$$

Rules with high confident were interested.

3. **Lift** – the ratio of the joint probability of X and Y to the expected joint probability if they were statistically independent.

$$\text{lift}(X \rightarrow Y) = \frac{P(XY)}{P(X) \cdot P(Y)} = \frac{\text{conf}(X \rightarrow Y)}{r \text{sup}(Y)}$$

4. **Odds ratio** – ratio of the odds of Y occurring in the presence of X to the odds of Y occurring in the in the presence of the complement of X .

$$\text{oddsratio}(X \rightarrow Y) = \frac{\text{odds}(Y|X)}{\text{odds}(Y|\neg X)} = \frac{P(XY)/P(X)}{P(X\neg Y)/P(X)}$$

Odds greater than one imply higher odds of Y occurring in the presence of X opposed to its complement $\neg X$.

3.1.2. K-means Clustering

Clustering is another important data mining technique which helps to group objects with common behaviors. In this research this technique is applied to cluster library users with common borrowing patterns. This will be used in the recommendation system to recommend books.

Clustering is an unsupervised technique introduced to group elements which are close to each other and k-means clustering is one of the popular clustering techniques. k-means clustering employs a iterative process and the goal is to find the clustering that minimizes the *error sum of squares* within a cluster[36].

i.e.
$$C^* = \operatorname{argmin}_c \{SSE(C)\}$$

$$\text{where } SSE(C) = \sum_{i=1}^k \sum_{x \in C_i} \|x_j - \mu_i\|^2 \text{ and } C = \{C_1, C_2, \dots, C_K\}$$

Number of clusters should be pre-determined in this technique and the algorithm begins with randomly generated k-points in the data space. Assignment of elements to the clusters and updating the centroids are employed in each iteration. Objects are assigned to a cluster which has the minimum distance between the cluster mean and the object. The procedure is repeated until the centroid has a small change between two iterations[36].

3.1.2.1. Cluster Validation

For internal validation of clustering, measurements such as Connectedness, Dunn Index and Silhouette Width Index can be employed [42]. Connectedness expresses “the extent to which observations are placed in the same group as their nearest neighbors in the data space” [43] and smaller values of connectivity implies better clustering. Similarity of an observation between the assigned cluster and a different one is measured in Silhouette Width Index and values close to one indicates proper grouping. The Dunn’s index combines dissimilarity between clusters and their diameters to estimate the most reliable number of clusters and smaller values than one indicates better clustering[43].

3.2. Technologies Used for the Application

3.2.1. Java

Java is a programming language that produces software for multiple platforms. When a programmer writes a Java application, the compiled code (known as bytecode) runs on most operating systems (OS), including Windows, Linux and Mac OS. Java derives much of its syntax from the C and C++ programming languages.

3.2.2. R Statistical Software

R is a programming language and free software environment for statistical computing and graphics that is supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. Polls, surveys of data miners, and studies of scholarly literature databases show that R's popularity has increased substantially in recent years. As of June 2018, R ranks 10th in the TIOBE index, a measure of popularity of programming languages.

3.2.3. My SQL

MySQL is an open-source relational database management system (RDBMS). It is written in C and C++. MySQL works on many system platforms, including AIX, BSDi, FreeBSD, HP-UX, eComStation, i5/OS, IRIX, Linux, macOS, Microsoft Windows, NetBSD, Novell NetWare, OpenBSD, OpenSolaris, OS/2 Warp, QNX, Oracle Solaris, Symbian, SunOS, SCO OpenServer, SCO UnixWare, Sanos and Tru64.

Chapter 4

Methodology

Ruhuna library is automated with KOHA open source library management software and all the records of library activities are stored in the central database using MySQL.

Following data were identified as necessary data for the study.

1. Data on bibliographic information
2. Information about borrowers
3. Information about transactions
4. Information on dewey decimal classification of books

4.1 Data Selection Procedure

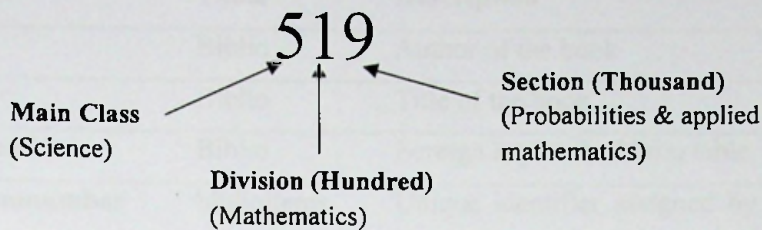
Data selection is the first stage of the KDD (Knowledge Discovery in Databases) process. In order to meet the objectives of the study, the following procedure was undertaken.

The SQL database was implemented for the library data using MySQL Workbench. This was a relational database with 165 tables.

In University of Ruhuna library, Dewey Decimal Classification (DDC) system is used to organize library materials according to the field of study. In DDC, basic classes are organized by fields of study. Fundamentally, the DDC is divided into ten main classes which cover the entire world of knowledge. Each main class is further divided into ten divisions, and each division into ten sections. For an example, the dewey number 519 can be explained as;

500- Science, 510- mathematics and 519- Probabilities & applied mathematics

TH 3897



Results of the study were based on the dewy numbers. Therefore, a new table called “dewy codes” was created and “id”, “dewynumber”, “dewyname”, “subdewyname”, “primarydewyname” were its fields.

| id | dewynumber | dewyname | subdewyname | primarydewyname |
|----|------------|---|---|---|
| 1 | 000 | Computer science- information & general works | Computer science- knowledge and systems | Computer science- information and general works |
| 2 | 001 | Knowledge | Computer science- knowledge and systems | Computer science- information and general works |
| 3 | 002 | The book | Computer science- knowledge and systems | Computer science- information and general works |
| 4 | 003 | Systems | Computer science- knowledge and systems | Computer science- information and general works |
| 5 | 004 | Computer science | Computer science- knowledge and systems | Computer science- information and general works |
| 6 | 005 | Computer programming- programs and data | Computer science- knowledge and systems | Computer science- information and general works |
| 7 | 006 | Special computer methods | Computer science- knowledge and systems | Computer science- information and general works |

Table 4.1: Table of "dewycodes"

Out of 165 tables created, following tables were used to select required attributes for the study. Certain attributes were selected from each table and their summary is given below in the Table 4.2.

| Table Name | Data |
|-------------|---|
| Bibilio | Information on library materials |
| Biblioitems | Information on library materials |
| Items | Information on library materials |
| Borrowers | Information on library users |
| Statistics | Information on issuing details |
| Dewycodes | Dewy decimal classification information |

Table 4.2: Tables Selected from the Database for the Study

| Attribute | Table | Description |
|-------------------------|--------------|---|
| Author | Biblio | Author of the book |
| Title | Biblio | Title of the book |
| Biblionenumber | Biblio | Foreign key to the biblio table |
| biblioitemnumber | biblioitems | Unique identifier assigned by "Koha" for a library material |
| Isbn | biblioitems | Unique numeric commercial book identifier |
| Itemtype | biblioitems | Item type of the book (lending, reference and etc.) |
| Publicationyear | biblioitems | Publication year of the book |
| publishercode | biblioitems | Publisher of the book |
| Dewyname | dewycodes | Dewy names of thousand sections |
| subdewyname | dewycodes | Names of hundred divisions |
| primarydewyname | dewycodes | Names of ten main classes |
| borrowernumber | statistics | Borrower number of the user |
| branchcode | borrowers | Branch of the borrower |
| Sex | borrowers | Gender of the borrower |
| Cardnumber | borrowers | library assigned ID number for borrowers |
| Surname | borrowers | Surname of the borrower |
| Firstname | borrowers | First name of the borrower |
| Itemnumber | statistics | Unique identifier assigned by Koha |
| itemcallnumber | Items | Unique description of each item in a library collection.(the "address" of materials on the shelf) |
| Datetime | statistics | Date and time for the issued book |

Table 4.3: Summary Table of Selected Variables

Certain additional data were required for the study and these were obtained from existing attributes using SQL queries.

| Attribute | Used attribute | description |
|--------------------------|-----------------------|---|
| Faculty | cardnumber | Faculty of the borrower |
| Year | datetime | Year of the issued book |
| Month | datetime | Month of the issued book |
| dewynumber | itemcallnumber | Dewy number of the book |
| subdewynumber | dewynumber | Dewy number which represent hundred divisions |
| primarydewynumber | dewynumber | Dewy number of ten main classes |

Table 4.4: Summary of New Variables

4.2. Data Cleaning and Preprocessing

Cleaning and preprocessing of data is an essential part in KDD process since unnecessary data can have an effect on conclusions. Therefore, noise and irrelevant data are removed in this stage[3].

Borrowing history of certain user types (temporary lecturers, temporary nonacademic staff, temporary professors, special medical students) were not considered for the study.

Furthermore, some missing values for the attribute "sex" were observed and they were imputed using the last three digits of the borrowers' identity card numbers.

4.3. Building the Data Warehouse

Data warehouse is “subject-oriented historical data that is organized to be accessible in a form readily acceptable for analytical processing activities (such as data mining, decision support querying, and other applications)”[6]. Accessing data within few seconds is one of the major benefits of a data warehouse. Following characteristics were considered when creating the data warehouse .[6]

- Organization – data are organized according to the objective.
- Consistency – In the warehouse data will be coded in a consistent manner.
- Time variant – data are stored for several years.
- Non-volatile – data are not updated after entering to the ware house.
- Relational – Normally the data warehouse uses a relational structure.
- Integration – Data from various sources are integrated.

Having required data in several tables, it was not easy to access them. Therefore, after cleaning and preprocessing of data, a data warehouse was created to reach data quickly and easily. Transaction details of library users were stored in the data warehouse called “transactional”, under selected attributes.

| biblio | biblioitems | items | Statistics | dewycodes | borrowers |
|--------------------------------|--|----------------|--|--|--|
| biblonumber author title | biblioitemnumber isbn itemtype publicationyear publishercode | itemcallnumber | datetime year month borrowernumber item number | dewynumber dewynome subdewynumber subdewynome primarydewynumber primarydewynome | branchcode sex cardnumber surname firstname faculty |

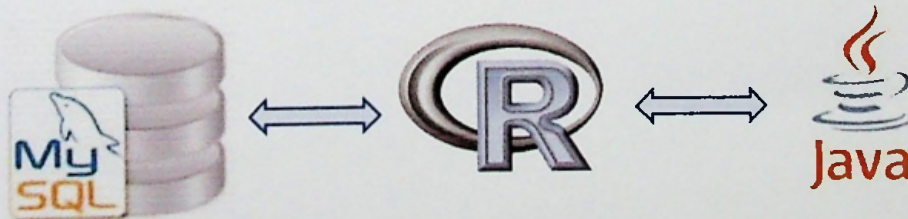
| biblonumber | biblioitemnumber | borrowernr | isbn | author | title | itemtype | publicatonyear | publishercode | datetime | year | month | categorycode |
|-------------|------------------|------------|------------|---------------------|------------------|----------|----------------|----------------|---------------------|------|-------|--------------|
| 3813 | 57417 | 3813 | 9552023092 | රජකරයා, මහජන රජ. | වැනි ව්‍යාපෘති | L | 1962 | සොවිලෝ | 2011-06-13 10:16:08 | 2011 | June | Student |
| 3813 | 54949 | 3813 | 9551387791 | සමිති, විල්ලව් | කුඩා කුරුල්ලා | L | 2006 | විලේප්පුරා | 2011-06-13 10:16:11 | 2011 | June | Student |
| 3813 | 52796 | 3813 | 9551387791 | සේනානායක, සෙනරත් | යනපත්තිය | L | 1985 | දොවැල ප්‍රකාශන | 2011-06-28 10:55:38 | 2011 | June | Student |
| 3813 | 26675 | 3813 | 9555513171 | සෙනෙවිරත්න, ජයවර්ධන | විද්‍යාගාර... | L | 2003 | දොවැල ප්‍රකාශන | 2011-06-28 10:55:42 | 2011 | June | Student |
| 3813 | 37747 | 3813 | 8121903459 | Dass, H.K. | Advanced eng... | L | 2004 | S. Chand | 2011-06-28 10:56:02 | 2011 | June | Student |
| 3813 | 10013 | 3813 | 9559175173 | Fonseka, Kubera | Ways of the w... | L | 1998 | Subha | 2011-07-13 10:03:29 | 2011 | July | Student |

Figure 4.1: Implementation of "transactionall" Table

After preprocessing and necessary transformation of data, connections between the statistical software and other sources were created.

4.4. Creation of the Connection between R and Other Sources

R⁵ was used to perform most of the statistical analysis in this research. One of the most fascinating features of this study was, setting up the connections between R and MySQL⁶ as well as R and Java⁷ in order to achieve the research objectives.



⁵ A statistical software

⁶ An open source database

⁷ A programming language



4.5. Establishment of the Connection between R and MySQL

Data required for the analysis were to be retrieved from the “transactionall” data warehouse. Therefore, the establishment of the connection between the MySQL database and R software was required. The package “RMySQL” [44] was installed and some of the computer settings were changed to meet the requirements. Subsequently, the connection was established by providing the username, password, host and database name of the existing MySQL database.

4.6. Establishment of the Connection between R and Java

Java programming language was used in developing the book recommendation system in the second stage of the study. Libraries; JRI, Rserve and RserveEngine, which allow to run R codes in Java applications were installed to build the connection between R and Java. The function “Rserve()” was called in R and a new local connection was established in Java as the next step. Finally, R commands were placed in the Java program and the necessary results were obtained.

Analysis and Design

5.1. Identifying Borrowing Patterns of Patrons

Association rule mining was performed for library transaction data as follows.

1. Identify frequent patterns of borrowing books by the users
2. Identify user characteristics in lending books

5.1.1 Identify Frequent Patterns of Borrowing Books by the Users

5.1.1.1 Data Preparation

Transaction data from 1st January 2013 to 31st December 2013 of five faculties, Science, Engineering, Medical, Management and Humanities and Social Science were used to identify the type of books which were more likely to be borrowed together by the students. Dewy number of the borrowed book was used for the analysis and lending records of the patrons were recorded in the “transactional” table in following manner.

| borrowernumber | dewynumber |
|-----------------------|-------------------|
| 3810 | 891 |
| 3810 | 581 |
| 3811 | 576 |
| 3811 | 571 |
| 3811 | 547 |
| 3812 | 574 |
| 3812 | 545 |
| 3812 | 576 |

Table 5.1: Arrangement of Data in the Database

The data set was rearranged by using MySQL queries in a way that each row of the table represents all the dewy numbers of the borrowed books per user throughout the year (2013) and each row is considered as a transaction.

| borrowernumber | dewynumber |
|-----------------------|-------------------|
| 3810 | 891,581 |
| 3811 | 576,571,547 |
| 3812 | 574,545,576 |
| : | : |
| : | : |

Table 5.2: Layout of the Rearranged Data

A sparse matrix⁸ was then created where each row represents a transaction and there is a column for each dewy number (feature) that could appear in any of the transactions. Summary of the data set was inspected and item frequency (relative) plot was obtained to identify the behavior of items in the data set.

Relevant rules for each faculty in the university and different user types were attained using Apriori algorithm[36] with pre-defined minimum confidence and minimum support. The package “arules” [45] was used to preform association rule mining in R and the package “arulesViz”[46] was used for the visualization purpose of extracted rules. Redundant rules (more general rules) were eliminated in the next stage and the rule assessment measures (explained above) were obtained to find out strong rules.

5.1.1.2 Test for Statistical Significance of Rules

Fisher Exact Test[36] (which tests whether the given rule is productive⁹ by comparing its confidence with that of each of its generalizations) was applied for the generated rules to validate them.

⁸ Most of the cells in the matrix are zero

⁹ A rule is productive if its Improvement (minimum difference between the confidence of a rule and any of its generalizations) is greater than zero

Rules with p-value less than the significance level (0.05) were excluded from the selected association rules and remaining rules were sorted according to the *support* as well as *lift*. Frequent patterns of lending books were identified by the selected association rules.

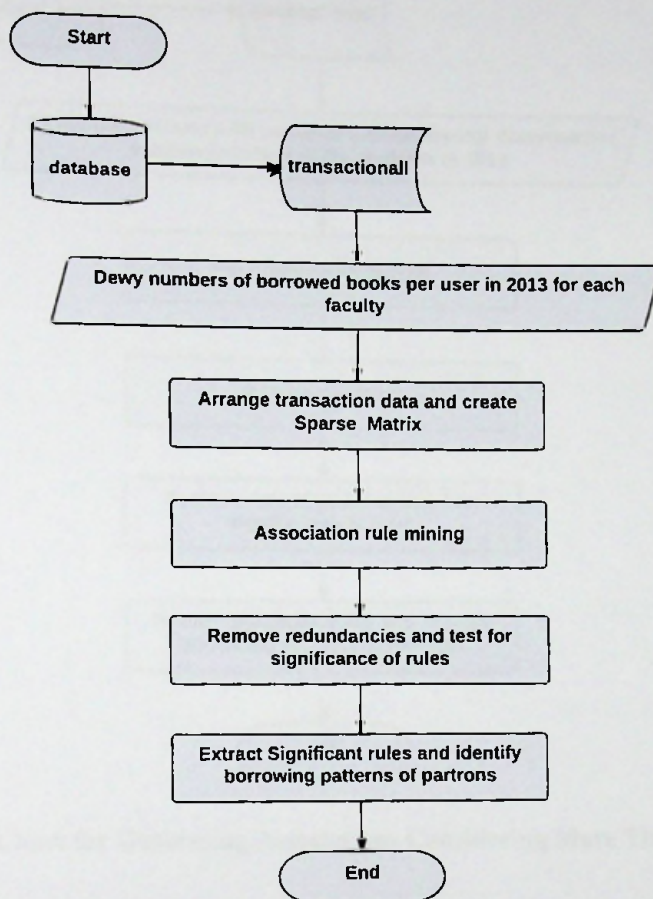


Figure 5.1: Flow Chart of Generating Association Rules

5.2.2 Identify User Behavior when Lending Books

More variables related to a transaction such as gender of the user, patron's faculty, month of the year, user type, dewy number, and sub dewy number were analyzed using association rules extracted by the procedure given below.

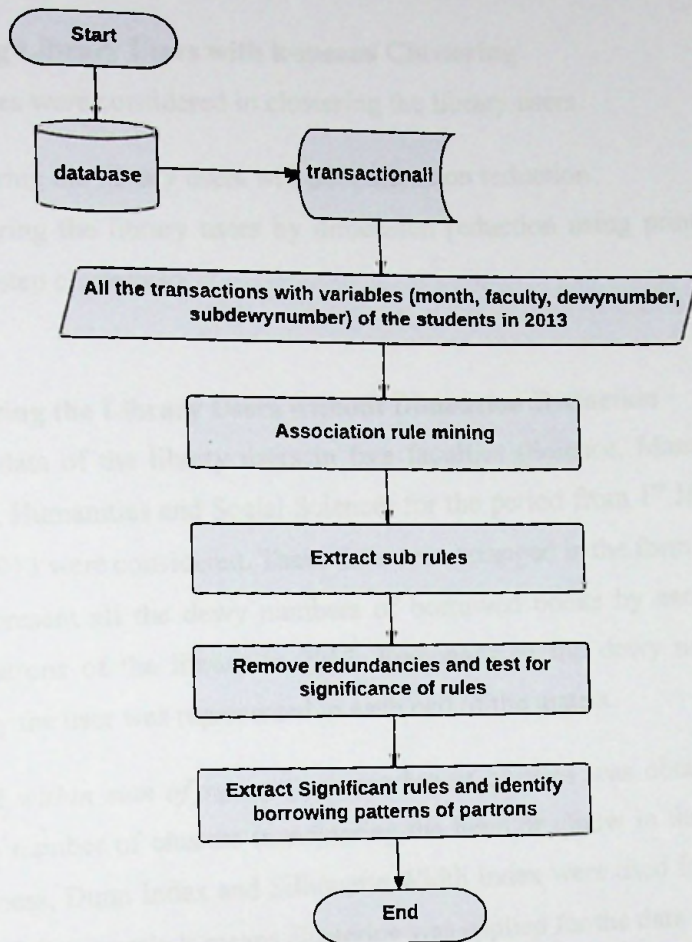


Figure 5. 2: Flow Chart for Generating Associations Considering More Than One Variable

5.2. Clustering Library Users with k-means Clustering

Two approaches were considered in clustering the library users.

1. Clustering the library users without dimension reduction
2. Clustering the library users by dimension reduction using principle components (Two step clustering)

5.2.1 Clustering the Library Users without Dimension Reduction

Transaction data of the library users in five faculties (Science, Management, Medical, Engineering, Humanities and Social Science) for the period from 1st January 2013 to 31st December 2013 were considered. These data were arranged in the form of a matrix, so that columns represent all the dewy numbers of borrowed books by each patron and rows represent patrons of the library in 2013. Frequency of the dewy number of the book borrowed by the user was represented in each cell of the matrix.

The plot of *within sum of squares vs. number of clusters* was obtained to identify the appropriate number of clusters (considering the bend or elbow in the plot). Furthermore Connectedness, Dunn Index and Silhouette Width Index were used for internal validation of clusters. Subsequently k-means clustering was applied for the data set using “kmeans()” function in R and properties of each cluster were identified.

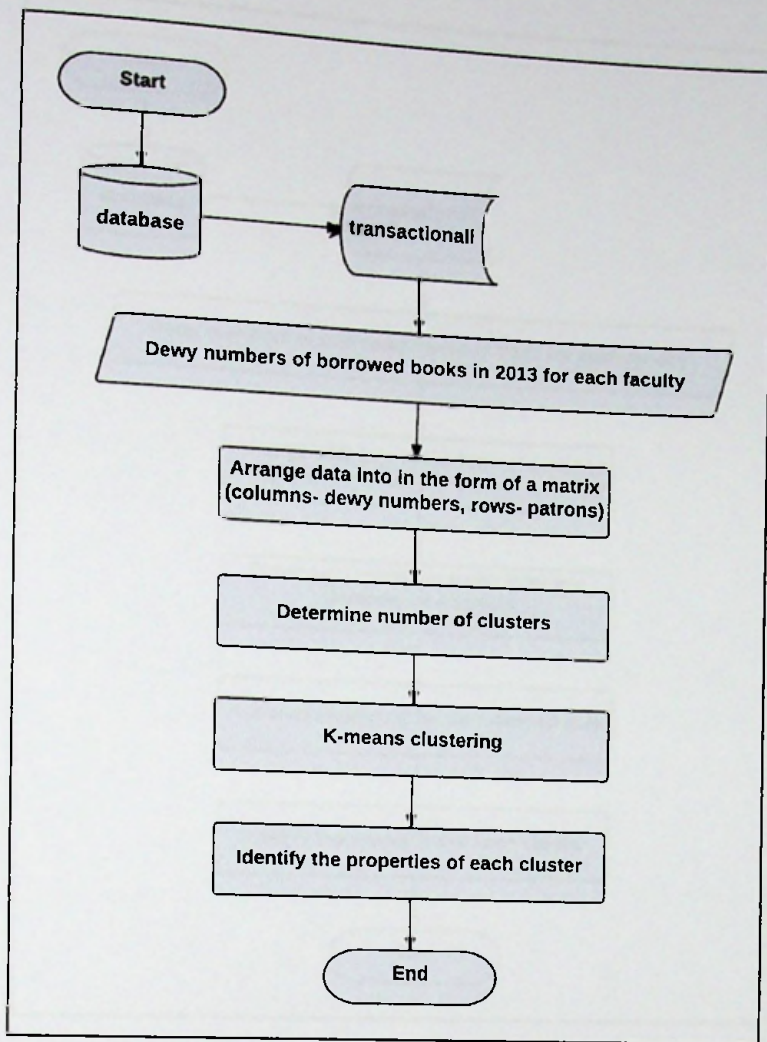


Figure 5.3 : Flow Chart for the Process of Clustering

5.2.2 Clustering the Library Users by Dimension Reduction Using Principle

Components (Two Step Clustering)

Principle component analysis (PCA) was applied for the arranged dataset and selected components covered 95% of the total variation. Subsequently the scores of selected components were used as variables for k-means clustering and cluster validation was done as mentioned in the previous method.

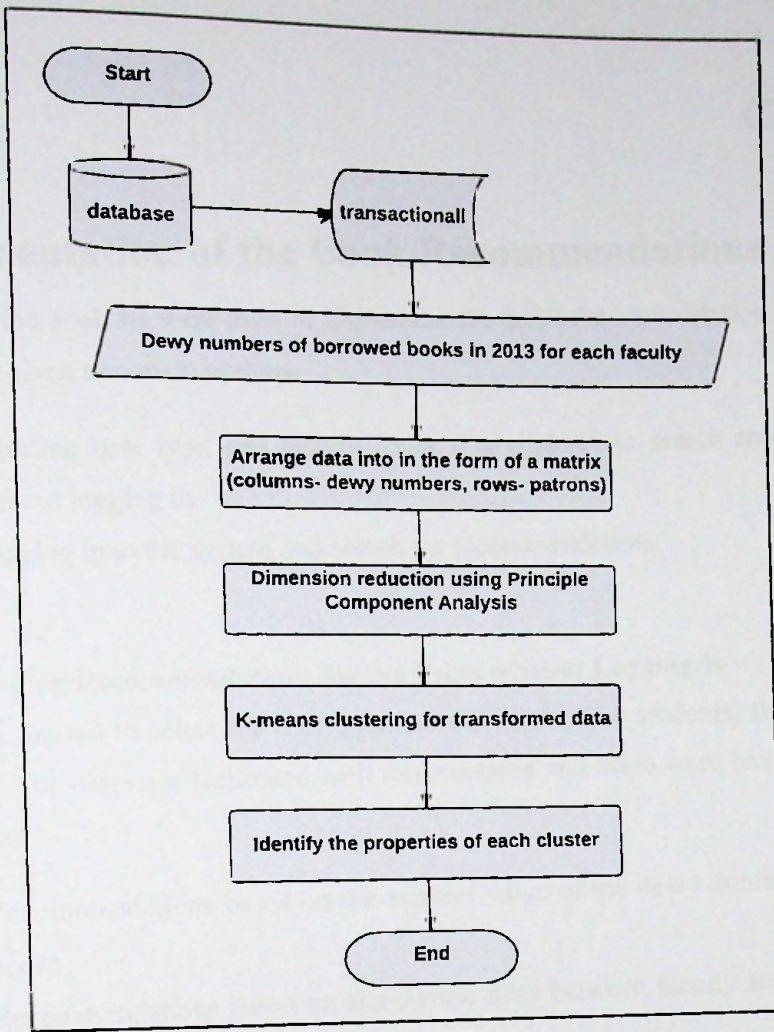


Figure 5.4: Flow Chart for the Process of Two Step Clustering

Results of the two approaches were compared to identify the most suitable clustering approach.

Chapter 6

Implementation of the Book Recommendations System

Results of the analysis were used to implement the book recommendation system where users were given two main options.

1. Selecting user type and faculty (only if a student) to search recommendations without logging in.
2. Logging in to the system and search for recommendation.

6.1. Providing Recommendations for the Users without Logging-in

User is requested to select the user type and the faculty (for students) through the main interface. The users are facilitated with three options and these were based on following techniques.

1. Recommendations based on the *support* value of the dewy numbers of borrowed books.
2. Recommendations based on association rules between faculty and dewy numbers of borrowed books.
3. Recommendations based on association rules between dewy numbers of borrowed books.

6.1.1. Recommendations Based on the Support Value of the Dewy Numbers of Borrowed Books

In this method, dewy numbers of the borrowed books per user throughout the year 2013 for the given faculty and user type were considered as transactions. Support for each dewy number were calculated and three dewy numbers with highest support were selected using R program. A MySQL query is sent to the database through R, to select two books which have highest usage from each dewy number. "Title of the book", "author", "item call number", "dewy name" and frequency of borrowing each book are displayed through main interface.

6.1.1.1. Recommendations Based on Association Rules between Faculty and Dewy Numbers of Borrowed Books.

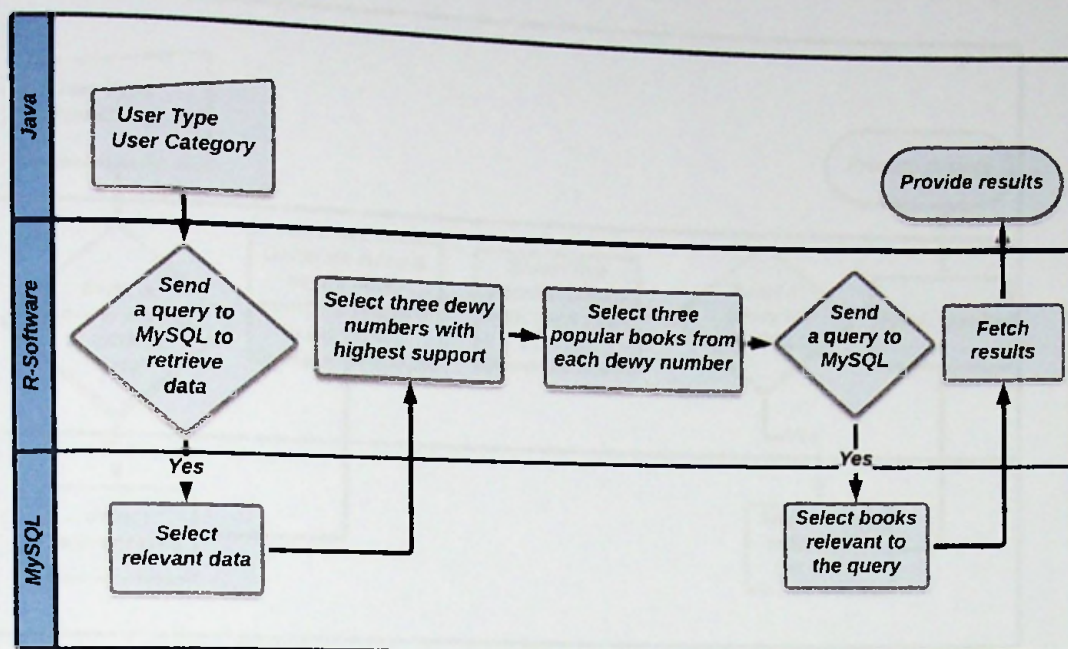


Figure 6.1 : Process for the Method 01

For the given faculty, association rules between faculty and dewy numbers were generated and dewy numbers relevant to the particular faculty were selected. Most popular five books were selected for each dewy number and "Title of the book", "author", "item call number", "dewy name" and frequency of borrowing each book will be displayed through the main interface.

6.1.1.2. Recommendations Based on Association Rules between Dewy Numbers of Borrowed Books

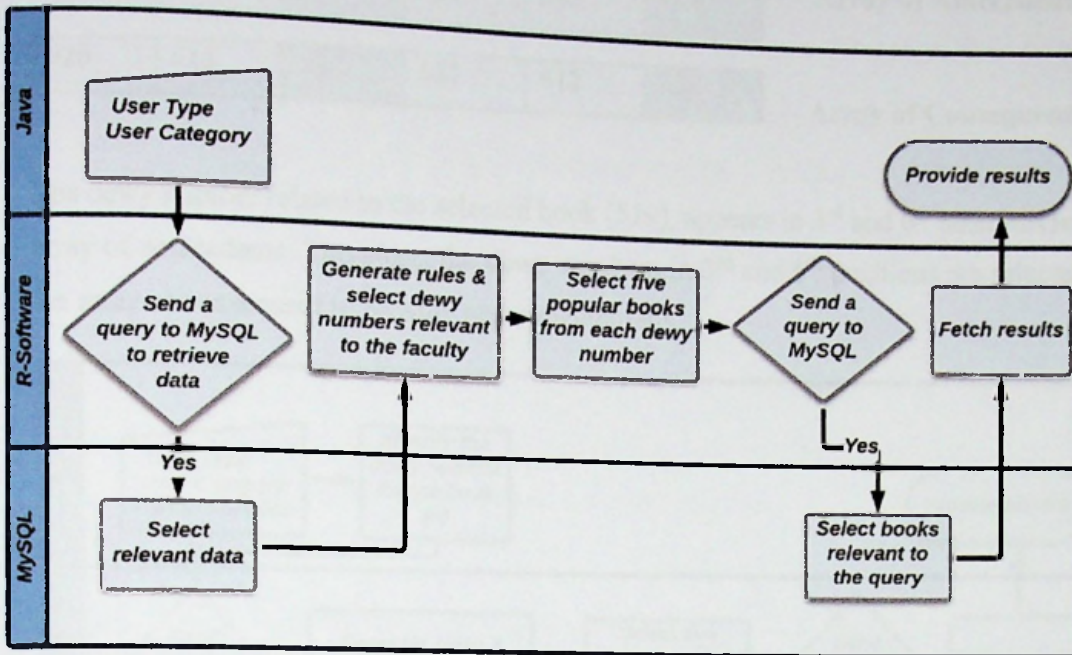


Figure 6.2 : Process of the Method 02

When the user searches for a book, extra recommendation materials will be suggested using association rules.

In order to provide extra recommendations, dewy number of the certain book is identified and association rules are generated relevant to the faculty and user type using "Apriori" algorithm. Antecedent and consequent of each generated rule are stored in separate arrays. If the dewy number of the selected book is included in the array of antecedents, its positions are saved and the dewy numbers at the same positions in the array of consequent are selected for recommendations.

Then five books with highest usage from each dewy number are recommended for the user and "title of the book", "author", "item call number", "dewy name" and frequency borrowing each book will be displayed through the main interface.

For an example, assume that the dewy number related to the selected book is "519" and following associations can be identified.

512 => 620, 616 => 612, 519 =>330, 520 =>521, 543 =>512, 519 => 548.

Two arrays are created as follows.

| | | | | | |
|-----|-----|-----|-----|-----|-----|
| 512 | 616 | 519 | 520 | 543 | 519 |
|-----|-----|-----|-----|-----|-----|

Array of Antecedent

| | | | | | |
|-----|-----|-----|-----|-----|-----|
| 620 | 612 | 330 | 521 | 512 | 548 |
|-----|-----|-----|-----|-----|-----|

Array of Consequent

The dewy number related to the selected book (519), appears in 3rd and 6th positions in the array of antecedents. Therefore, the dewy numbers in 3rd and 6th positions are selected in the array of consequent to provide recommendations.

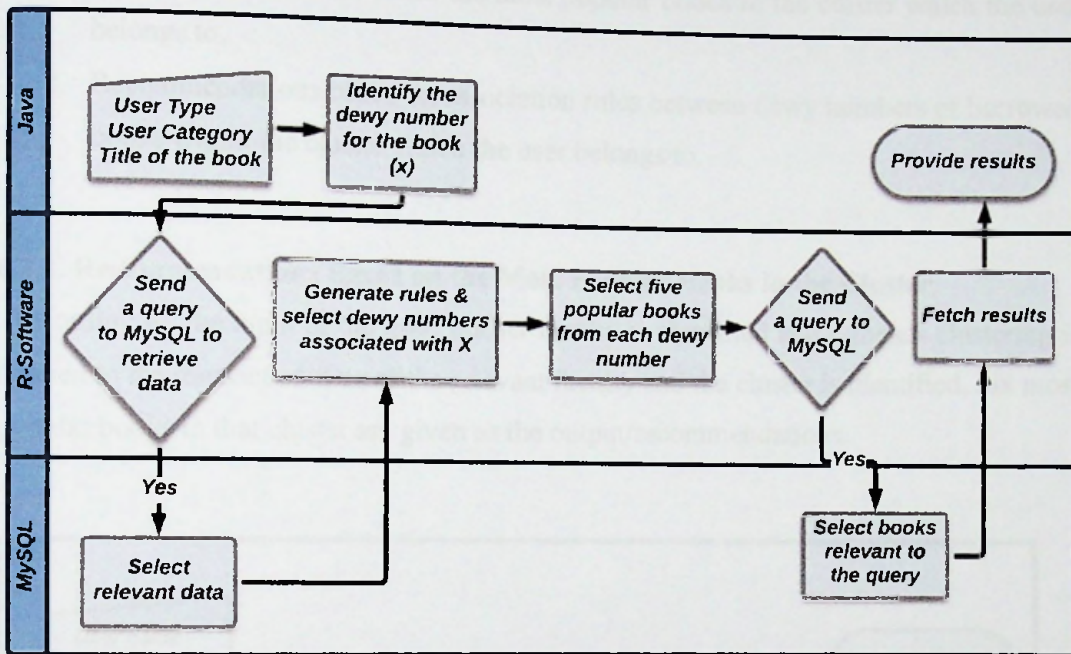


Figure 6.3: Process of the Method 03

6.2 Providing Recommendations for the Users after Logging-in

Results of the cluster analysis were used to provide recommendations after logging-in to the user account.

User is requested to enter his/her username and password into the system. Summary of the user; name, gender faculty, total count of his/her transactions and information on most read book is displayed in the interface. The users are facilitated with two options and these were based on following techniques.

1. Recommendations based on the most popular books in the cluster which the user belongs to.
2. Recommendations based on association rules between dewey numbers of borrowed books within the cluster which the user belongs to.

6.2.1. Recommendations Based on the Most Popular Books in the Cluster

According to the input of the user, his/her faculty is identified and k-means clustering is applied to the transaction data of the relevant faculty and the cluster is identified. Six most popular books in that cluster are given as the output/recommendations.

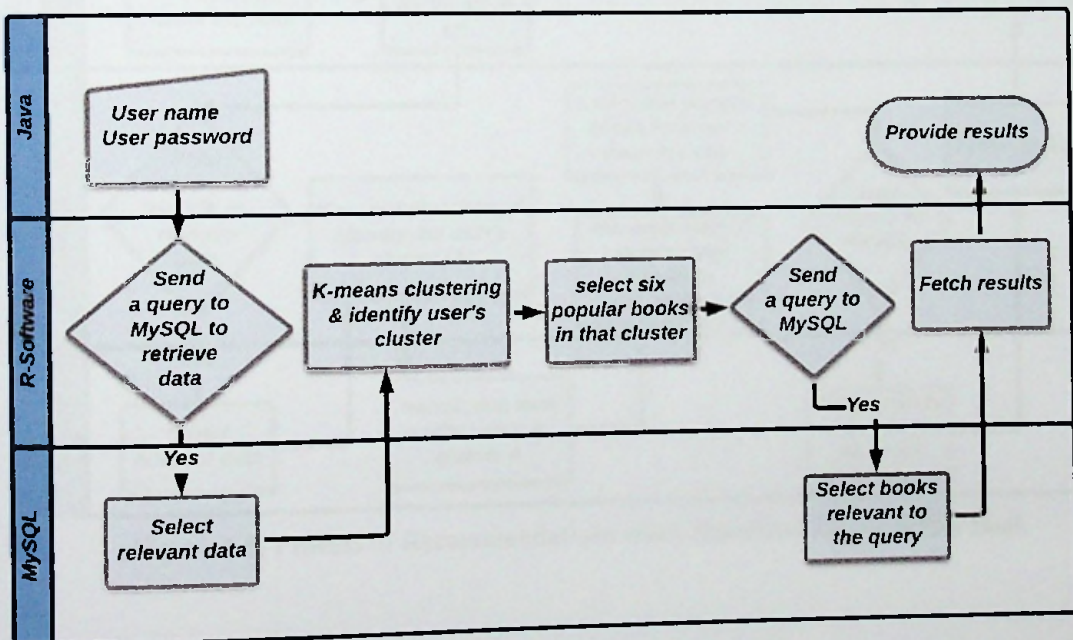


Figure 6.4: Process of the Recommendations after Logging-in

6.2.2. Recommendations Based on Association Rules between Dewey Numbers of Borrowed Books within the Cluster

When the user searches for a book, extra recommendation material will be suggested using association rules in the cluster which the user belongs to.

According to the input of the user, his/her faculty is identified and k-means clustering is applied to the transaction data of the relevant faculty and the cluster is identified. Transactions of all the users in that cluster for the year 2013 are fetched and associations between dewey numbers of borrowed books are generated using “Apriori” algorithm.

In order to provide extra recommendations, dewey number of the certain book is identified and antecedent and consequent of each generated rule are stored in separate arrays. If the dewey number of the selected book is included in the array of antecedents, its positions is saved and the dewey numbers at the same positions in the array of consequent are selected for recommendations. Then five books with highest usage from each dewey number are recommended for the user and “title of the book”, “author”, “item call number”, “dewey name” and frequency borrowing each book will be displayed through main interface.

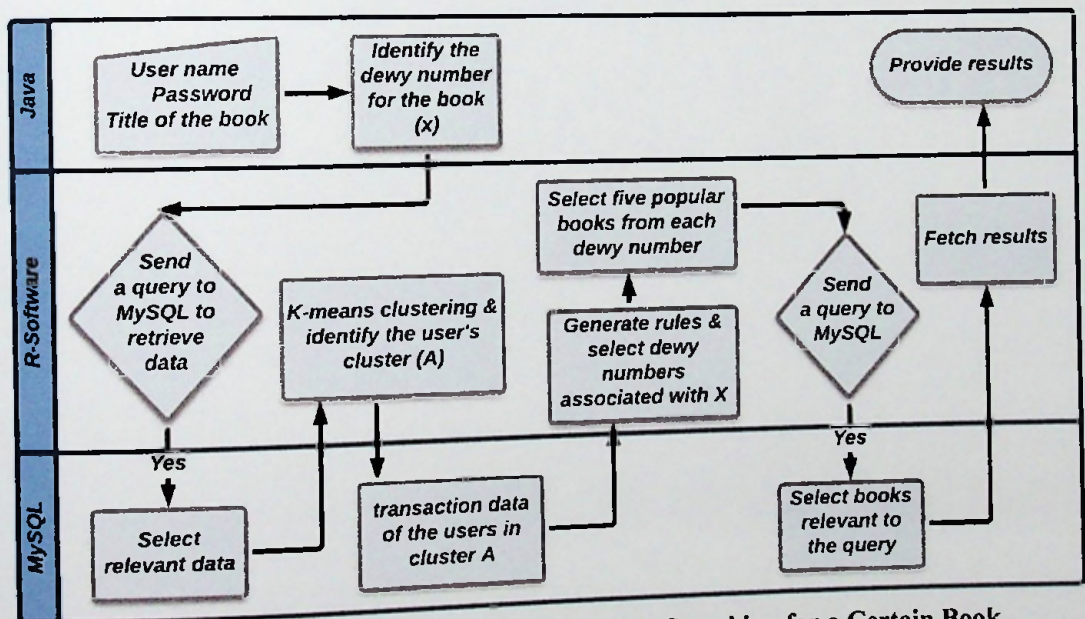


Figure 6.5: Process of Recommendations when Searching for a Certain Book

To evaluate the user satisfaction level on the book recommendation system, the user is asked to provide his/her feedback at the bottom of the main interface. According to their responses, performance of the recommendation system is evaluated.

7.2 Descriptive Statistics

7.2.1 Comparison of Current Users of the System



Figure 7.2 Comparison of Current Users of the System

- 1. Number of the library users who worked in the faculty of Education and Social Sciences.
- 2. Number of the library users who worked in the faculty of Engineering.

Results of the Study

7.1 Descriptive Analysis

7.1.1 Composition of Student Users in 2013

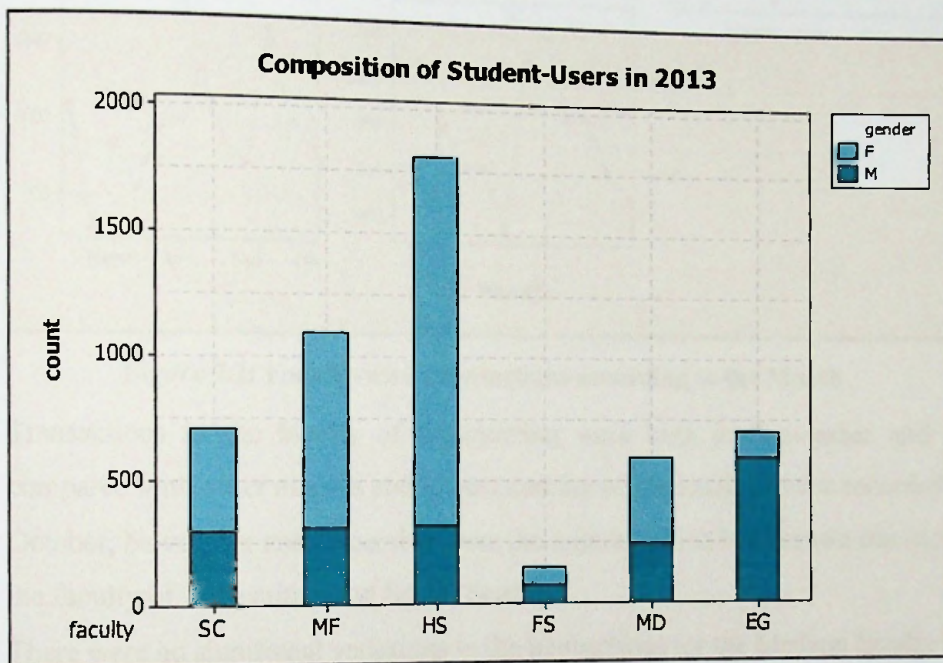


Figure 7.1: Composition of Library Users (Students)

- Majority of the library users were students in the faculty of Humanities and Social Science.
- Female students who use the library were greater than male students except for the Faculty of Engineering.

7.1.2 Faculty-wise Transactions according to the Month

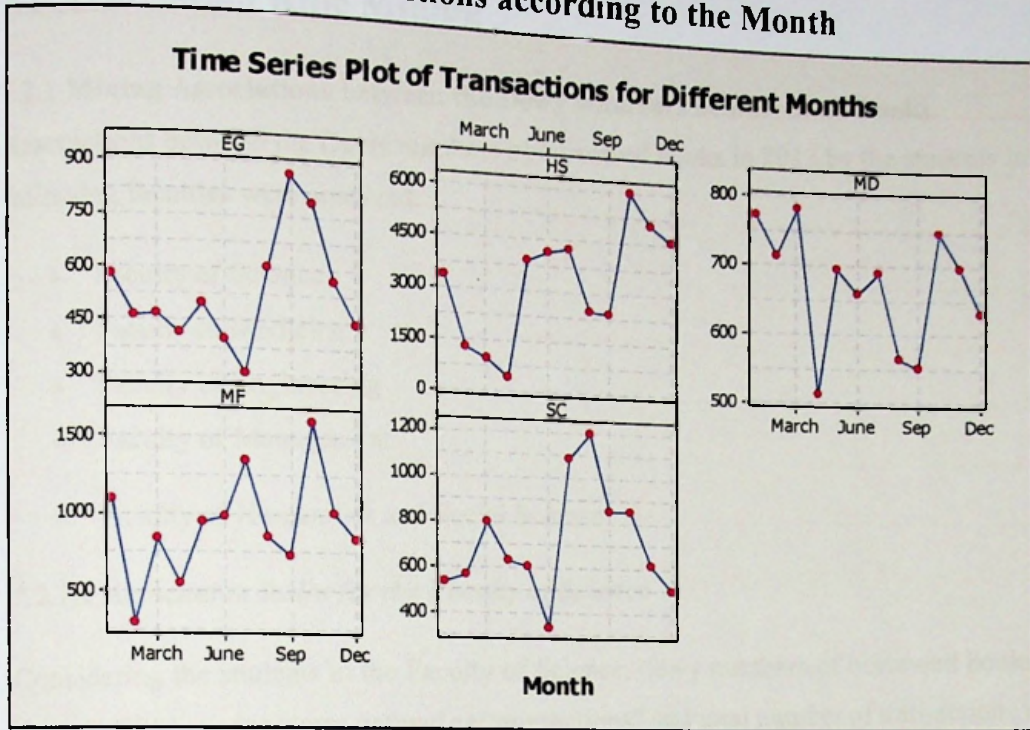


Figure 7.2: Faculty-wise Transactions according to the Month

- Transactions in the faculty of Engineering were high in September and October compared with other months and lowest number of transactions were recorded in July.
- October, November and December were the months which had highest transactions for the faculty of Humanities and Social Science.
- There were no significant variations in the transactions for the Medical faculty between the months. Transactions for the faculty of Medicine had lowest records in April.
- January, July and August are the months which had highest number of transactions in the faculty of Management.
- Transactions in the faculty of Science were lowest in June where July and August indicated highest number of transactions.

7.2 Association Rule Mining

7.2.1 Mining Associations between the Dewy Numbers of Borrowed Books

Associations between the Dewy numbers of borrowed books in 2013 by the students in following faculties were analyzed.

- Faculty of Science
- Faculty of Medicine
- Faculty of Engineering
- Faculty of Management

- Faculty of Humanities and Social Science

7.2.1.1 Association Rules for the Faculty of Science

Considering the students in the Faculty of Science, dewy numbers of borrowed books for the year 2013 per user were defined as “transactions” and total number of transactions were 707. Summary of the transactions are given below.

Summary of Transactions

Average items contained in a transaction were 4.605.

Considering the proportion of transactions for an item (*support*), the item frequency (relative) plot was obtained.

Item frequency (relative) plot

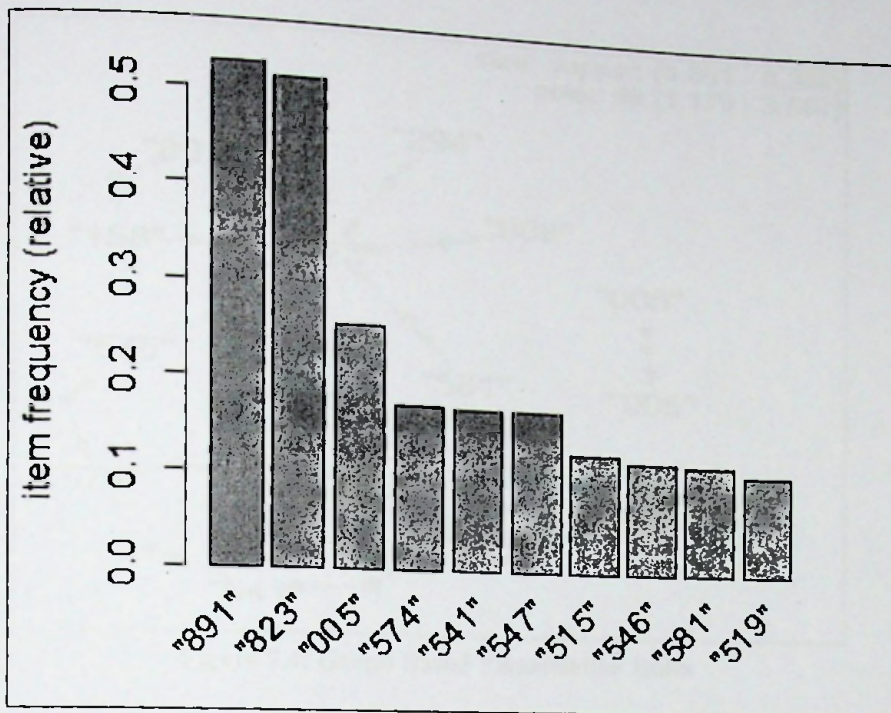


Figure 7.3: Item Frequency Plot

According to the Figure 7.3, books under the DDC categories of; East Indo-European & Celtic literatures (891), English fiction (823) had highest usage. Other than the literature categories, Computer programming, programs & data (005), Biology (574), Physical chemistry (541) had been appeared in most of the transactions.

Generating Association Rules

Association rule mining using "Apriori" algorithm with minimum confidence = 0.05 and minimum support = 0.05 was performed and 84 results were obtained. Redundancies were removed from the existing rules and 16 productive rules were obtained using Fisher Exact Test.

Graph-based Visualization of Rules

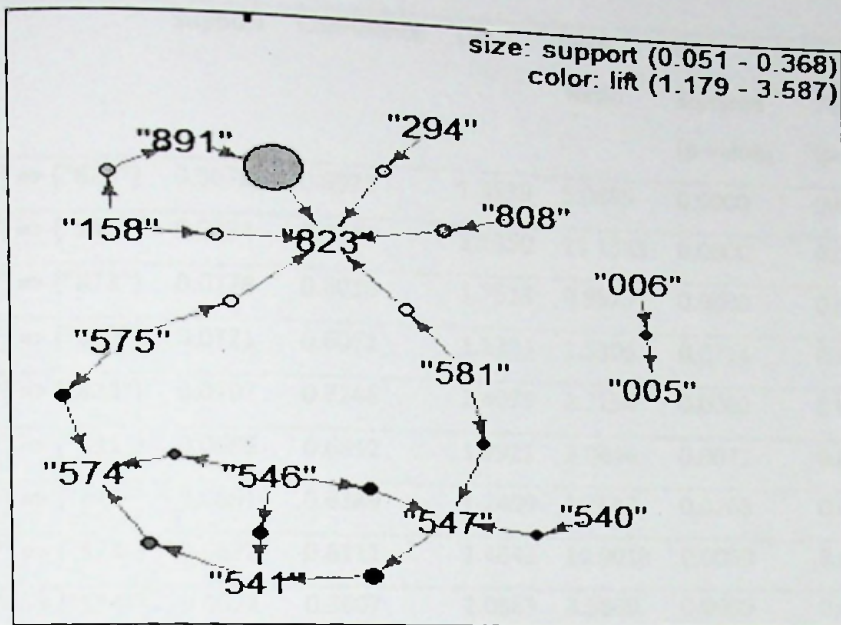


Figure 7.4: Graph Based Visualization Rules

- Books in the categories of “East Indo-European & Celtic literatures” (891) and “English fiction” (823) seemed to appear together in most of the transactions.
- The *lift* values for the association rules which contained following categories seemed to be large;
 - Physical Chemistry (541), Organic Chemistry (547), Specific parts of & systems in plants (575), Inorganic Chemistry (546)

Association Rules Extracted from the Records of Faculty of Science Students in 2013

| Rules | Support | Confidence | Lift | Odds Ratio | Chi Squared (p-value) | Fisher Test (p-value) |
|--------------------|---------|------------|--------|------------|-----------------------|-----------------------|
| {"891"} => {"823"} | 0.3678 | 0.6971 | 1.3539 | 5.0885 | 0.0000 | 0.0000 |
| {"547"} => {"541"} | 0.0934 | 0.5410 | 3.1350 | 11.1333 | 0.0000 | 0.0000 |
| {"808"} => {"823"} | 0.0778 | 0.9016 | 1.7513 | 9.9973 | 0.0000 | 0.0000 |
| {"581"} => {"823"} | 0.0721 | 0.6071 | 1.1793 | 1.5306 | 0.0714 | 0.0454 |
| {"158"} => {"823"} | 0.0707 | 0.7246 | 1.4075 | 2.7154 | 0.0002 | 0.0002 |
| {"158"} => {"891"} | 0.0665 | 0.6812 | 1.2911 | 2.0446 | 0.0071 | 0.0048 |
| {"575"} => {"823"} | 0.0651 | 0.6389 | 1.2409 | 1.7637 | 0.0263 | 0.0175 |
| {"575"} => {"574"} | 0.0622 | 0.6111 | 3.4843 | 10.9018 | 0.0000 | 0.0000 |
| {"541"} => {"574"} | 0.0622 | 0.3607 | 2.0563 | 3.5609 | 0.0000 | 0.0000 |
| {"294"} => {"823"} | 0.0608 | 0.8113 | 1.5758 | 4.4607 | 0.0000 | 0.0000 |
| {"546"} => {"541"} | 0.0608 | 0.5000 | 2.8975 | 6.8608 | 0.0000 | 0.0000 |
| {"546"} => {"547"} | 0.0594 | 0.4884 | 2.8302 | 6.4551 | 0.0000 | 0.0000 |
| {"540"} => {"547"} | 0.0552 | 0.6190 | 3.5874 | 10.9834 | 0.0000 | 0.0000 |
| {"581"} => {"547"} | 0.0537 | 0.4524 | 2.6216 | 5.3007 | 0.0000 | 0.0000 |
| {"006"} => {"005"} | 0.0523 | 0.6167 | 2.3955 | 5.5694 | 0.0000 | 0.0000 |
| {"546"} => {"574"} | 0.0509 | 0.4186 | 2.3867 | 4.3609 | 0.0000 | 0.0000 |

Table 7.1: Summary of Association Rules

- Most number of rules were associated with the categories of Chemistry division.
- Considering the rule {"547"} => {"541"},
 - 54% of students who borrowed books in the category of "Organic chemistry" (547) may borrow the books in "Physical chemistry" (541) category.
 - Out of 707 transactions, these two categories were appeared together in 66 transactions.
 - According to the *lift* value, it was three times more likely that a student borrow a book in "Physical chemistry" category relative to the average borrower, given that a book in "Organic chemistry" category has been borrowed by the same student.

- Odds ratio for the rule was greater than one. This implies, odds of borrowing a book in "Physical chemistry" category given that a book in "Organic chemistry" category has been borrowed is eleven times odds of borrowing a book in "Physical chemistry" category given that a book in "Organic chemistry" category has not been borrowed.

(All the other rules can be interpreted in the same manner)

- 45% of students who borrowed books in the category of "Specific topics in natural history" (581) are likely to borrow the books in "Organic chemistry" (547) category.
- Students who borrowed books in the category of "Special computer methods" (006) are likely to borrow the books in "Computer programming, programs & data" (005) category with 61% of confidence.

As explained above, association rules were generated to identify the borrowing patterns of the other faculties using "Apriori" algorithm with 0.05 of minimum support and minimum confidence. Some of the results are given below.

7.2.1.2 Association Rule Mining for the Faculty of Management

| Rules | Support | Confidence | Lift | Odds Ratio | Chi Squared (p-value) | Fisher Test (p-value) |
|--------------------|---------|------------|--------|------------|-----------------------|-----------------------|
| {"823"} => {"891"} | 0.3854 | 0.7339 | 1.2364 | 3.5324 | 0.0000 | 0.0000 |
| {"158"} => {"891"} | 0.1142 | 0.7062 | 1.1897 | 1.7995 | 0.0009 | 0.0005 |
| {"158"} => {"823"} | 0.1050 | 0.6497 | 1.2373 | 1.8468 | 0.0003 | 0.0002 |
| {"330"} => {"658"} | 0.0959 | 0.7394 | 1.3450 | 2.6037 | 0.0000 | 0.0000 |
| {"330"} => {"891"} | 0.0913 | 0.7042 | 1.1863 | 1.7446 | 0.0040 | 0.0024 |
| {"657"} => {"658"} | 0.0658 | 0.6545 | 1.1906 | 1.6266 | 0.0199 | 0.0124 |
| {"338"} => {"658"} | 0.0648 | 0.6961 | 1.2661 | 1.9927 | 0.0018 | 0.0011 |
| {"294"} => {"823"} | 0.0575 | 0.6495 | 1.2368 | 1.7588 | 0.0102 | 0.0066 |
| {"808"} => {"823"} | 0.0566 | 0.8378 | 1.5955 | 5.1163 | 0.0000 | 0.0000 |
| {"808"} => {"891"} | 0.0530 | 0.7838 | 1.3204 | 2.6269 | 0.0006 | 0.0003 |

Table 7.2: Summary of Association Rules

7.2.1.3 Association Rules Mining for the Faculty of Engineering

| Rules | Support | Confidence | Lift | Odds Ratio | Chi Squared (p-value) | Fisher Test (p-value) |
|--------------------|---------|------------|--------|------------|-----------------------|-----------------------|
| {"004"} => {"621"} | 0.2051 | 0.8521 | 1.1752 | 2.6512 | 0.0000 | 0.0000 |
| {"510"} => {"621"} | 0.1496 | 0.8678 | 1.1968 | 2.8752 | 0.0001 | 0.0000 |
| {"517"} => {"621"} | 0.1225 | 0.9149 | 1.2618 | 4.7015 | 0.0000 | 0.0000 |
| {"821"} => {"624"} | 0.0912 | 0.4267 | 1.5439 | 2.4157 | 0.0000 | 0.0000 |
| {"620"} => {"621"} | 0.0826 | 0.8657 | 1.1939 | 2.6292 | 0.0067 | 0.0034 |
| {"628"} => {"624"} | 0.0812 | 0.7215 | 2.6109 | 9.1911 | 0.0000 | 0.0000 |
| {"681"} => {"621"} | 0.0798 | 0.9333 | 1.2872 | 5.8411 | 0.0002 | 0.0000 |
| {"626"} => {"624"} | 0.0755 | 0.7260 | 2.6272 | 9.1716 | 0.0000 | 0.0000 |
| {"519"} => {"624"} | 0.0726 | 0.4286 | 1.5508 | 2.3077 | 0.0000 | 0.0001 |
| {"528"} => {"624"} | 0.0655 | 0.6866 | 2.4844 | 7.2079 | 0.0000 | 0.0000 |
| {"519"} => {"004"} | 0.0570 | 0.3361 | 1.3963 | 1.7820 | 0.0076 | 0.0063 |
| {"519"} => {"821"} | 0.0513 | 0.3025 | 1.4158 | 1.7844 | 0.0095 | 0.0080 |
| {"626"} => {"628"} | 0.0470 | 0.4521 | 4.0170 | 10.4560 | 0.0000 | 0.0000 |
| {"625"} => {"624"} | 0.0442 | 0.8857 | 3.2050 | 23.9632 | 0.0000 | 0.0000 |

Table 7.3: Summary of Association Rules

7.2.1.4 Association Rules Mining for the Faculty of Medicine

| Rules | Support | Confidence | Lift | Odds Ratio | Chi Squared (p-value) | Fisher Test (p-value) |
|--------------------|---------|------------|--------|------------|-----------------------|-----------------------|
| {"617"} => {"616"} | 0.3653 | 0.8543 | 1.2025 | 3.8622 | 0.0000 | 0.0000 |
| {"618"} => {"616"} | 0.3114 | 0.8079 | 1.1371 | 2.2708 | 0.0000 | 0.0000 |
| {"615"} => {"616"} | 0.2542 | 0.8882 | 1.2503 | 4.4869 | 0.0000 | 0.0000 |
| {"618"} => {"617"} | 0.2155 | 0.5590 | 1.3072 | 2.4039 | 0.0000 | 0.0000 |
| {"891"} => {"616"} | 0.1936 | 0.8394 | 1.1815 | 2.5540 | 0.0001 | 0.0001 |
| {"615"} => {"617"} | 0.1835 | 0.6412 | 1.4994 | 3.4382 | 0.0000 | 0.0000 |
| {"612"} => {"611"} | 0.1566 | 0.6159 | 1.8955 | 5.4998 | 0.0000 | 0.0000 |



| | | | | | | |
|--------------------|--------|--------|--------|--------|--------|--------|
| {"614"} => {"616"} | 0.1566 | 0.9208 | 1.2961 | 5.7948 | 0.0000 | 0.0000 |
| {"823"} => {"618"} | 0.1532 | 0.4892 | 1.2691 | 1.8741 | 0.0005 | 0.0003 |
| {"823"} => {"617"} | 0.1515 | 0.4839 | 1.1316 | 1.3948 | 0.0613 | 0.0376 |
| {"891"} => {"823"} | 0.1498 | 0.6496 | 2.0746 | 6.8814 | 0.0000 | 0.0000 |
| {"614"} => {"617"} | 0.1313 | 0.7723 | 1.8060 | 6.1082 | 0.0000 | 0.0000 |
| {"813"} => {"823"} | 0.1145 | 0.6415 | 2.0487 | 5.6111 | 0.0000 | 0.0000 |
| {"891"} => {"618"} | 0.1145 | 0.4964 | 1.2875 | 1.8119 | 0.0024 | 0.0018 |
| {"614"} => {"615"} | 0.1094 | 0.6436 | 2.2487 | 6.6720 | 0.0000 | 0.0000 |
| {"813"} => {"617"} | 0.0976 | 0.5472 | 1.2796 | 1.8002 | 0.0060 | 0.0043 |
| {"813"} => {"618"} | 0.0875 | 0.4906 | 1.2725 | 1.6920 | 0.0142 | 0.0101 |
| {"813"} => {"891"} | 0.0774 | 0.4340 | 1.8816 | 3.3447 | 0.0000 | 0.0000 |
| {"614"} => {"823"} | 0.0690 | 0.4059 | 1.2964 | 1.6400 | 0.0273 | 0.0196 |
| {"614"} => {"891"} | 0.0572 | 0.3366 | 1.4596 | 1.9215 | 0.0055 | 0.0050 |

Table 7.4: Summary of Association Rules

7.2.1.5 Association Rules Mining for the Faculty of Humanities and Social Science

| Rules | Support | Confidence | Lift | Odds Ratio | Chi Squared (p-value) | Fisher Test (p-value) |
|--------------------|---------|------------|--------|------------|-----------------------|-----------------------|
| {"823"} => {"891"} | 0.4437 | 0.8963 | 1.1803 | 5.1841 | 0.0000 | 0.0000 |
| {"954"} => {"891"} | 0.3168 | 0.8085 | 1.0646 | 1.5788 | 0.0001 | 0.0000 |
| {"301"} => {"891"} | 0.2483 | 0.8287 | 1.0913 | 1.7923 | 0.0000 | 0.0000 |
| {"294"} => {"891"} | 0.2384 | 0.8045 | 1.0594 | 1.4426 | 0.0036 | 0.0019 |
| {"158"} => {"891"} | 0.2351 | 0.8469 | 1.1153 | 2.0907 | 0.0000 | 0.0000 |
| {"301"} => {"823"} | 0.1700 | 0.5672 | 1.1458 | 1.5131 | 0.0001 | 0.0000 |
| {"930"} => {"891"} | 0.1656 | 0.8772 | 1.1551 | 2.6155 | 0.0000 | 0.0000 |
| {"158"} => {"823"} | 0.1639 | 0.5905 | 1.1928 | 1.7037 | 0.0000 | 0.0000 |
| {"294"} => {"954"} | 0.1606 | 0.5419 | 1.3830 | 2.4167 | 0.0000 | 0.0000 |
| {"301"} => {"954"} | 0.1302 | 0.4346 | 1.1092 | 1.2893 | 0.0147 | 0.0086 |
| {"158"} => {"954"} | 0.1280 | 0.4612 | 1.1771 | 1.4883 | 0.0002 | 0.0001 |
| {"158"} => {"301"} | 0.1192 | 0.4294 | 1.4330 | 2.2601 | 0.0000 | 0.0000 |

| | | | | | | |
|--------------------|--------|--------|--------|---------|--------|--------|
| {"930"} => {"954"} | 0.1159 | 0.6140 | 1.5671 | 3.0864 | 0.0000 | 0.0000 |
| {"930"} => {"823"} | 0.1098 | 0.5819 | 1.1754 | 1.5391 | 0.0004 | 0.0002 |
| {"158"} => {"294"} | 0.1065 | 0.3837 | 1.2947 | 1.7465 | 0.0000 | 0.0000 |
| {"398"} => {"891"} | 0.1015 | 0.8440 | 1.1115 | 1.8251 | 0.0018 | 0.0008 |
| {"491"} => {"891"} | 0.0999 | 0.9141 | 1.2038 | 3.7332 | 0.0000 | 0.0000 |
| {"294"} => {"301"} | 0.0999 | 0.3371 | 1.1248 | 1.2823 | 0.0242 | 0.0143 |
| {"320"} => {"301"} | 0.0949 | 0.4687 | 1.5639 | 2.5534 | 0.0000 | 0.0000 |
| {"306"} => {"891"} | 0.0938 | 0.8333 | 1.0974 | 1.6667 | 0.0087 | 0.0045 |
| {"320"} => {"954"} | 0.0911 | 0.4496 | 1.1474 | 1.3489 | 0.0111 | 0.0068 |
| {"910"} => {"823"} | 0.0861 | 0.5417 | 1.0942 | 1.2488 | 0.0844 | 0.0483 |
| {"808"} => {"891"} | 0.0817 | 0.9427 | 1.2414 | 5.7181 | 0.0000 | 0.0000 |
| {"910"} => {"954"} | 0.0795 | 0.5000 | 1.2761 | 1.6926 | 0.0000 | 0.0000 |
| {"398"} => {"954"} | 0.0745 | 0.6193 | 1.5804 | 2.8825 | 0.0000 | 0.0000 |
| {"398"} => {"823"} | 0.0723 | 0.6009 | 1.2139 | 1.6276 | 0.0009 | 0.0005 |
| {"320"} => {"158"} | 0.0712 | 0.3515 | 1.2662 | 1.5521 | 0.0004 | 0.0003 |
| {"305"} => {"320"} | 0.0690 | 0.5531 | 2.7308 | 6.8734 | 0.0000 | 0.0000 |
| {"930"} => {"301"} | 0.0690 | 0.3655 | 1.2197 | 1.4497 | 0.0032 | 0.0022 |
| {"155"} => {"891"} | 0.0684 | 0.8857 | 1.1664 | 2.5998 | 0.0003 | 0.0001 |
| {"330"} => {"823"} | 0.0684 | 0.5511 | 1.1133 | 1.2928 | 0.0722 | 0.0421 |
| {"150"} => {"891"} | 0.0668 | 0.8897 | 1.1716 | 2.7060 | 0.0002 | 0.0001 |
| {"930"} => {"294"} | 0.0668 | 0.3538 | 1.1938 | 1.3872 | 0.0098 | 0.0064 |
| {"300"} => {"301"} | 0.0657 | 0.5920 | 1.9756 | 4.0627 | 0.0000 | 0.0000 |
| {"306"} => {"823"} | 0.0657 | 0.5833 | 1.1784 | 1.4936 | 0.0074 | 0.0046 |
| {"930"} => {"158"} | 0.0657 | 0.3480 | 1.2535 | 1.5092 | 0.0013 | 0.0009 |
| {"306"} => {"301"} | 0.0651 | 0.5784 | 1.9302 | 3.8193 | 0.0000 | 0.0000 |
| {"491"} => {"823"} | 0.0607 | 0.5556 | 1.1223 | 1.3135 | 0.0711 | 0.0418 |
| {"306"} => {"954"} | 0.0602 | 0.5343 | 1.3636 | 1.9225 | 0.0000 | 0.0000 |
| {"303"} => {"320"} | 0.0602 | 0.5023 | 2.4800 | 5.2302 | 0.0000 | 0.0000 |
| {"303"} => {"301"} | 0.0596 | 0.4977 | 1.6608 | 2.6422 | 0.0000 | 0.0000 |
| {"551"} => {"823"} | 0.0591 | 0.5691 | 1.1497 | 1.3946 | 0.0318 | 0.0191 |
| {"305"} => {"954"} | 0.0591 | 0.4735 | 1.2083 | 1.4658 | 0.0072 | 0.0047 |
| {"330"} => {"338"} | 0.0574 | 0.4622 | 3.7224 | 10.4135 | 0.0000 | 0.0000 |
| {"303"} => {"954"} | 0.0568 | 0.4747 | 1.2114 | 1.4706 | 0.0077 | 0.0051 |

| | | | | | | |
|--------------------|--------|--------|--------|--------|--------|--------|
| {"398"} => {"294"} | 0.0524 | 0.4358 | 1.4705 | 2.0130 | 0.0000 | 0.0000 |
| {"658"} => {"954"} | 0.0519 | 0.4947 | 1.2626 | 1.5991 | 0.0021 | 0.0015 |
| {"398"} => {"301"} | 0.0519 | 0.4312 | 1.4389 | 1.9331 | 0.0000 | 0.0000 |
| {"306"} => {"320"} | 0.0513 | 0.4559 | 2.2508 | 4.0791 | 0.0000 | 0.0000 |
| {"910"} => {"158"} | 0.0513 | 0.3229 | 1.1633 | 1.2958 | 0.0611 | 0.0370 |
| {"155"} => {"823"} | 0.0502 | 0.6500 | 1.3130 | 1.9954 | 0.0001 | 0.0001 |

Table 7.5: Summary of Association Rules

7.2.2 Association Rules between Faculty and Dewey Decimal Number

Association rules between the Faculty and Dewey Number were important in the recommendation system. Hence, following rules were identified with the use of Association Rule Mining.

| Rules | Support | Confidence | Lift | Odds Ratio | Chi Squared (p-value) | Fisher Test (p-value) |
|----------------------------------|---------|------------|---------|------------|-----------------------|-----------------------|
| {faculty=HS} => {dewynumber=954} | 0.0679 | 0.1335 | 1.9217 | 47.7877 | 0.0000 | 0.0000 |
| {faculty=MD} => {dewynumber=616} | 0.0609 | 0.4240 | 6.7641 | 343.9604 | 0.0000 | 0.0000 |
| {faculty=MF} => {dewynumber=658} | 0.0421 | 0.4221 | 7.5164 | 46.0695 | 0.0000 | 0.0000 |
| {faculty=RU} => {dewynumber=621} | 0.0348 | 0.4216 | 7.1154 | 26.6319 | 0.0000 | 0.0000 |
| {faculty=EG} => {dewynumber=621} | 0.0240 | 0.6531 | 11.0226 | 49.5976 | 0.0000 | 0.0000 |
| {faculty=MD} => {dewynumber=615} | 0.0189 | 0.1315 | 6.4886 | 92.9632 | 0.0000 | 0.0000 |
| {faculty=MD} => {dewynumber=611} | 0.0159 | 0.1108 | 6.8953 | 641.5953 | 0.0000 | 0.0000 |
| {faculty=MD} => {dewynumber=617} | 0.0145 | 0.1009 | 6.9376 | 1541.9539 | 0.0000 | 0.0000 |

| | | | | | | |
|-------------------------------------|--------|--------|---------|----------|--------|--------|
| {faculty=RU} => {dewynumber=624} | 0.0139 | 0.1686 | 11.0100 | 133.4454 | 0.0000 | 0.0000 |
| {faculty=SC} => {dewynumber=005} | 0.0119 | 0.1198 | 7.5821 | 31.5536 | 0.0000 | 0.0000 |

Table 7.6: Summary of Association Rules

According to the Table 7.6,

- Students in the Faculty of Medicine were more likely to borrow books in “Human anatomy, cytology & histology” (611), “Pharmacology & therapeutics” (615), “Diseases” (616) and “Surgery & related medical specialties” (617) categories.
- Students in the Faculty of Management were likely to borrow books in the category of “General management” (658) with 42% confidence.

Remaining rules can be interpreted in the same manner.

7.3 Clustering the Library Users (Students)

7.3.1 Clustering the Library Users in the Faculty of Science

7.3.1.1 Results of K-means Clustering

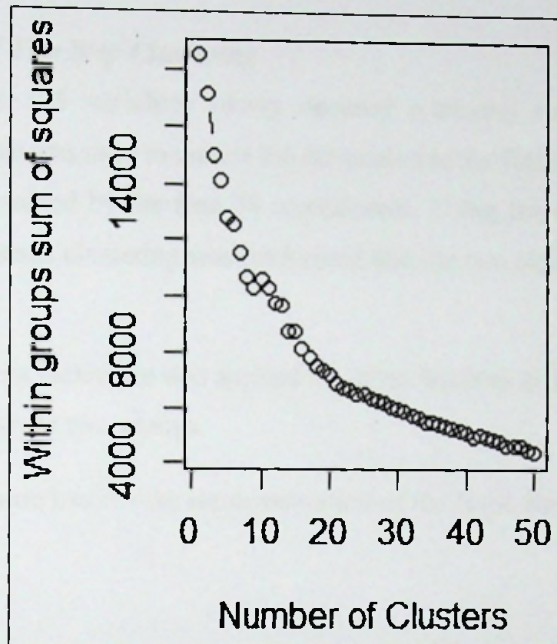


Figure 7.5: Plot of Number of Clusters vs. Within Groups Sum of Squares

Cluster Validation

| | | | | | | | | | | |
|----------------------|--------------|---------|----------|---------|---------|----------|----------|----------|----------|----------|
| Clustering Methods: | | | | | | | | | | |
| kmeans | | | | | | | | | | |
| Cluster sizes: | | | | | | | | | | |
| 2 3 4 5 6 7 8 9 10 | | | | | | | | | | |
| Validation Measures: | | | | | | | | | | |
| | | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| kmeans | Connectivity | 11.1972 | 31.0734 | 31.1734 | 38.7718 | 126.3877 | 129.8290 | 145.9595 | 138.7988 | 185.5512 |
| | Dunn | 0.1537 | 0.1660 | 0.1748 | 0.1236 | 0.0485 | 0.0546 | 0.0591 | 0.0650 | 0.0695 |
| | Silhouette | 0.6554 | 0.6535 | 0.6327 | 0.6368 | 0.3539 | 0.3586 | 0.3521 | 0.3522 | 0.3155 |
| Optimal scores: | | | | | | | | | | |
| | Score | Method | Clusters | | | | | | | |
| Connectivity | 11.1972 | kmeans | 2 | | | | | | | |
| Dunn | 0.1748 | kmeans | 4 | | | | | | | |
| Silhouette | 0.6554 | kmeans | 2 | | | | | | | |

Figure 7.6 : Cluster Validation

Considering the validation measures in Figure 7.6, "Connectivity" was small and "Silhouette index" was close to 1 when there were two clusters. Therefore, two clusters were appropriate for the users in the Faculty of Science.

7.3.1.2 Results of Two Step Clustering

Since there were 175 variables (dewy decimal numbers) in the data set, principle component analysis was used to reduce the dimension in the feature space. 95% of the total variation was explained by the first 35 components. Using the factor scores of those 35 components, K-means clustering was performed and the two clusters were appropriate for the data set.

Similarly, clustering technique was applied for other faculties and the users in each faculty could be clustered into two groups.

All these results were used in the implementation of the Book Recommendation System.

7.4 Book Recommendation System

7.4.1 Main Interface

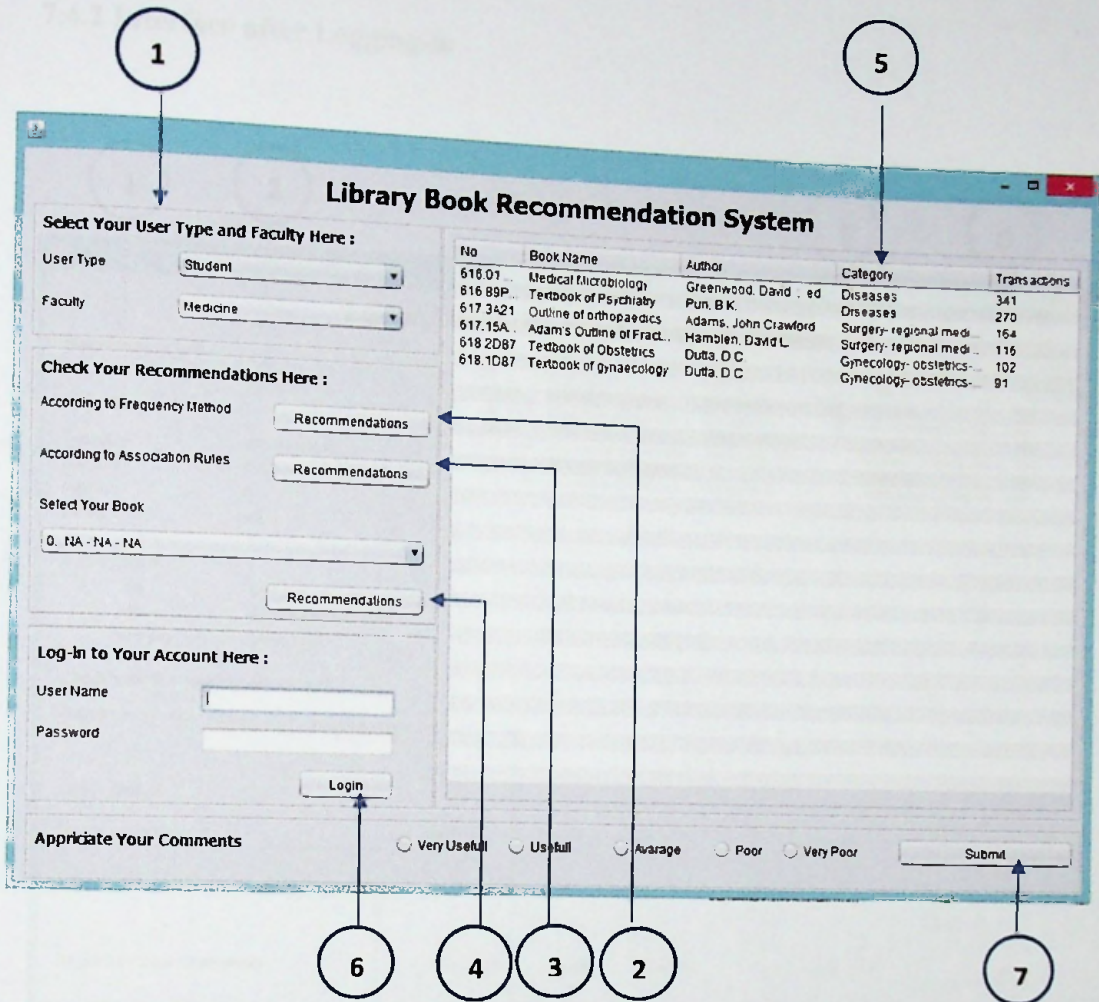


Figure 7.7: Main Interface of the Book Recommendation System

- 1- User can select his/her user type and faculty.
- 2- User is facilitated to search for recommendations using the procedure explained in the section 6.1.1. (considering the support value)
- 3- User is facilitated to search for recommendations using the procedure explained in the section 6.1.1.1. (using association rules related to the faculty)
- 4- User is facilitated to search for recommendations when searching for a certain book as explained in the 6.1.1.2.
- 5- Recommendations are displayed in this table.
- 6- User is facilitated to log-in to his/her account.

- 7- User is given the opportunity to comment on the book recommendation system.

7.4.2 Interface after Logging-in

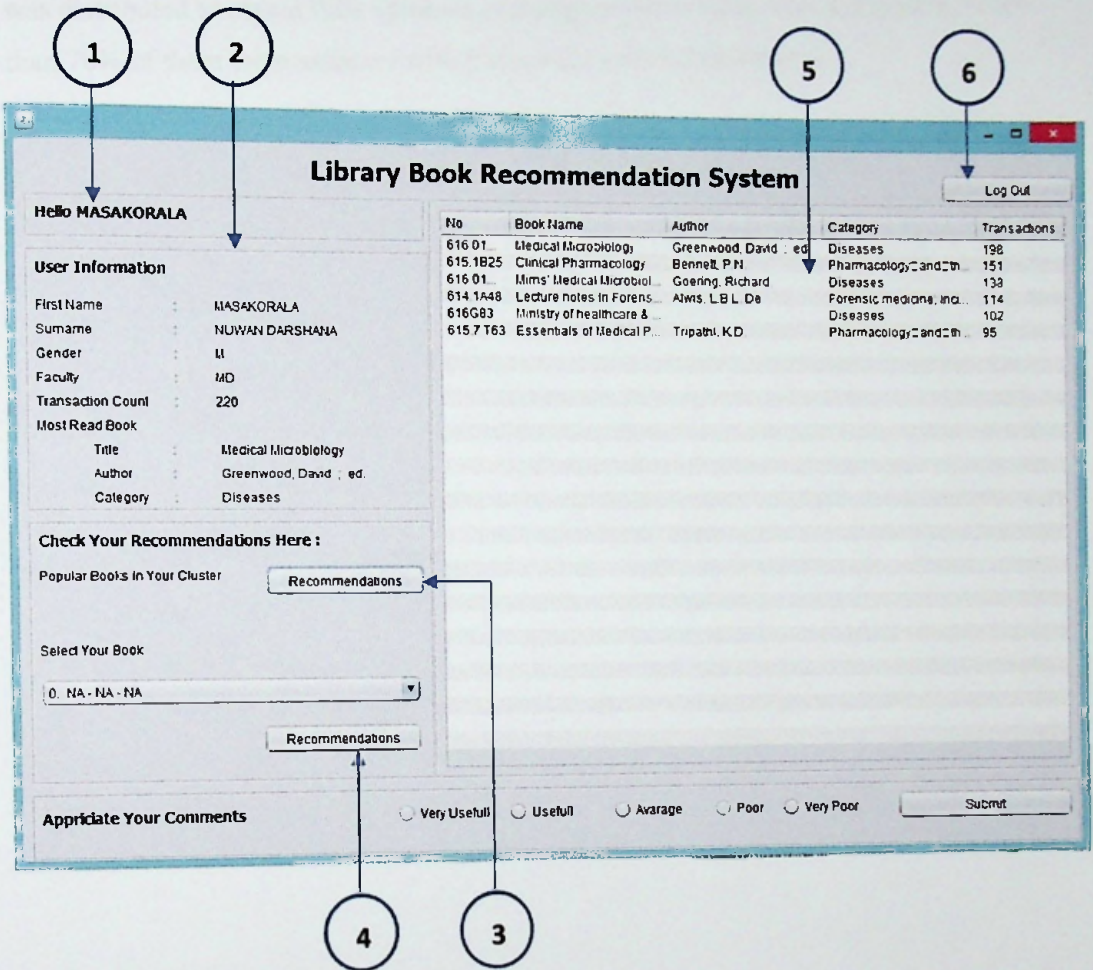


Figure 7.15: Interface after Logging-in to the User Account

- 2- Displays the user information.
- 3- User is facilitated to search for recommendations as explained in the section 6.2.1.
- 4- Search for recommendations when searching for a certain book explained in the section 6.2.2 (using association rules within the cluster)
- 5- Recommendations are displayed in this table.
- 6- User can log-out from his/her account.



7.5 Evaluation

Evaluation of the Book Recommendation system was carried out using a questionnaire. Thirty five students were selected from four state universities in Sri Lanka and allowed them to use the implemented Book Recommendation System. Then, a questionnaire was distributed to obtain their opinions and suggestions towards the said system. More than 70% of them were satisfied with the system's recommendations.

Conclusion & Further Work

7.1 Conclusions

As per the pre-defined objectives of the study, data mining techniques could be used effectively to enhance library experience of the library users of University of Ruhuna. Students' transactions in five faculties (Science, Medical, Management, Engineering, Humanities and Social Science) were basically analyzed in this study to retrieve interesting patterns and behaviors of borrowing books and the findings were useful in assisting library management to make decisions and extra recommendations could be provided for the library users.

Considering the students in five faculties; Science, Medicine, Management, Engineering and Humanities and Social Science, the highest number of transactions were recorded among the students in Humanities and Social Science. The reason could be that the total number of students in the faculty of Humanities and Social Science is high when compared to other faculties. Borrowing books in the categories of "East Indo-European & Celtic literatures" (891) and "English fiction" (823) of Dewy Decimal classifications was a common feature in the faculties of Science, Management and Humanities and Social Science, where highest number of transactions were recorded for said categories. It was observed that borrowing books related to the subject areas is not satisfactory. On the other hand students in the Faculties of Medicine and Engineering seem to borrow more books related to their subject areas. Without considering the categories of "East Indo-European & Celtic literatures" (891) and "English fiction" (823), the most popular categories of books within the students in these five faculties are as follows.

Faculty of Science - Computer programming, programs & data (005)

Faculty of Management - General management (658)

Faculty of Medicine – Diseases (616),

Faculty of Engineering - Applied physics (621)

Faculty of Humanities and Social Science - South Asia; India (954)

Library management can give priority to these categories in acquisitions. Moreover they can facilitate the users by providing e-books related to these categories, so that the problem of lack of copies in books can be solved.

According to the results of the study, management can decide on the most popular categories within these divisions and it is helpful in acquisitions. For an example, when they are going to purchase books related to the division of "Earth sciences & geology" (550-560), they can give priority to the books in the category of "Geology, hydrology & meteorology" (551).

When considering the borrowing patterns of each faculty, some interesting patterns were identified. Students tend to borrow books related to Chemistry and Computer science together in the Faculty of Science. On the other hand according, students are likely to borrow books in the divisions of "Biology" (570-580) or "Plant" (580-590) when they are borrowing books in the division of "Chemistry" (540-550). Therefore layout of the Library can be re-arranged (arrange these divisions closer to each other) to minimize searching time. When considering the students in the faculty of Management, they are more likely to borrow books in "General management" (658) when they are borrowing books in "Economics" (330) category and these categories can be arranged together.

Students in the Faculty of Engineering are more likely to borrow books in the categories of "Data processing & computer science" (004) and "Applied physics" (621) together. Therefore, as per the results, books in the divisions of 000-010, 510-520, 520-530 and 620-630 can be arranged together.

To group the students in each faculty, two clustering methods (K-means clustering and two step clustering) were tested. Results from these two methods were not significantly differ from each other. Students in each faculty could be grouped into two clusters. In the Faculty of Science, considering the users in two clusters, it was identified that most of the students had borrowed books related to the "chemistry"(540-550) in one of the clusters and the other cluster suggested that books in "Computer programming, programs & data" (005), "Analysis" (515), "Classical mechanics; solid mechanics"(531), were also popular among them.

In the management faculty books in the category of "General management" (658) were prominent in one cluster and books in the division of "Economics" (330-340) appeared mostly in the other cluster.

When considering the faculty of Humanities and Social Science, books in the main class of DDC "Social sciences" (300-400) were popular in one cluster and "South Asia; India" (954), "Religions of Indic origin" (294) and "History of ancient world to ca. 499" (930) were popular (had borrowed) among the students in the other cluster.

Other than providing conclusions for the library management to make their decisions, a book recommendation system was implemented to enhance the library experience of users based on the results of association rule mining and clustering. K-means clustering was used for the recommendation system to group the students in each faculty. The book recommendation system was a combination of R software, Java and MySQL and depicts how different subject areas can be integrated to satisfy the users in every aspect.

This system can be used effectively in selecting relevant books easily. Since this was implemented based on the prior behavior of the users, current students can search for more relevant books. These recommendations will increase the library usage by making students aware of the library collection. Even a student who has no idea on which book to borrow, will be facilitated to select relevant books without wasting time.

7.2 Limitations

In this research, association rule mining and clustering were performed to analysis the transaction data of only five faculties in University of Ruhuna. Hence, more important results could be obtained by analyzing the transactions of all the other user categories. Furthermore in future researches, another clustering method (density based clustering) can be used to analyze library data and a comparison can be made with results obtained for K-means clustering method.

Furthermore, there were variations in borrowing books according to the month and the reasons for these variations can be further analyzed.

7.3 Future Work

The recommendation system can be further developed to search books by providing key words using "text mining". The content of the book is required to adopt this procedure and TF-IDF (term frequency-inverse document frequency) approach can be

used to accomplish this task. In this method, TF (Term frequency) measures how frequently a term occurs in a document. TF-IDF measures, “values for each word in a document through an inverse proportion of the frequency of the word in a particular document to the percentage of documents the word appears in”. If a user searches for a book using key words, TF-IDF weights of each key word for each book can be calculated and books with highest TF-IDF weights can be recommended for the users.

Nowadays new concepts and techniques are tried out and discovered by the scientists to overcome the problems in managing large datasets. R has introduced various packages like “dat.table”, “ff”, “ffbase” to perform operations much faster. Furthermore, scientists have introduced the concept of “parallel computing”, which uses a set of computers (networks of multiple computers called clusters) to solve problems in handling large amount of data. The “MapReduce” programming model which was developed by Google is used to process data on a large cluster of networked computers. In MapReduce, a problem is divided into smaller tasks that are distributed across the computers in the cluster in the “Map” step and the results of small work pieces are collected and summarized into a final solution in the “Reduce”. A popular commercial open source alternative to the MapReduce framework is Apache Hadoop. These new concepts and techniques can be used when developing applications on data mining to enhance the performance. Furthermore, this is a new research area and web sites like Keggel conduct competitions to enhance the popularity of this field. Even though introducing these novel concepts is a challenging task, it is essential to make use of them when necessary, for the betterment of the whole society.



References

- [1] J. Pal, "Usefulness and applications of data mining in extracting information from different perspectives," vol. 58, no. March, pp. 7–16, 2011.
- [2] H. Perera, P. Wijetungs, and R. Mahaliyanaarachchi, "Library Review – Wayamba University of SL," 2009.
- [3] H. Sahu, S. Shurma, and S. Gondhalakar, "A Brief Overview on Data Mining Survey," *Int. J. Comput. Technol. ...*, vol. 1, no. 3, pp. 114–121, 2008.
- [4] A. Selvadoss Thanamani, "An Overview of Knowledge Discovery Database and Data mining Techniques," *Int. J. Innov. Res. Comput. Commun. Eng. (An ISO Certif. Organ.)*, vol. 32972, no. 1, pp. 1571–1578, 2014.
- [5] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," *Expert Syst. Appl.*, vol. 33, no. 1, pp. 135–146, Jul. 2007.
- [6] Y. Alsultanny, "Database Preprocessing and Comparison between Data Mining Methods," *Int. J. New Comput. Archit. ...*, vol. 1, no. 1, pp. 61–73, 2011.
- [7] Y. Koren and R. Bell, "Advances in Collaborative Filtering," *Recomm. Syst. Handb.*, pp. 145–186, 2011.
- [8] K. Chitra and B. Subashini, "Data Mining Techniques and its Applications in Banking Sector," *Int. J. Emerg. Technol. ...*, vol. 3, no. 8, pp. 219–226, 2013.
- [9] A. M. . b Almasoud, H. S. . Al-Khalifa, and A. . Al-Salman, "Recent developments in data mining applications and techniques," *10th Int. Conf. Digit. Inf. Manag. ICDIM 2015*, no. Icdim, pp. 36–42, 2015.
- [10] S. Slater, S. Joksimovic, V. Kovanovic, R. S. Baker, and D. Gasevic, "Tools for Educational Data Mining: A Review," *J. Educ. Behav. Stat.*, vol. 42, no. 1, pp. 85–106, 2016.
- [11] H. C. Koh and G. Tan, "Data mining applications in healthcare.," *J. Healthc. Inf. Manag.*, vol. 19, pp. 64–72, 2005.
- [12] Z.-G. Hul, J.-P. Li !', L. Hul, and Y. Yang !, "Research and Application of

- Data Warehouse and Data Mining Technology in Medical Field," pp. 457–460, 2015.
- [13] D. Kaur and J. Kaur, "Available Online at www.ijarcs.info International Journal of Advanced Research in Computer Science A Countermeasure Technique for Email Spoofing," vol. 8, pp. 1110–1112, 2017.
- [14] B. Sahoo and B. S. P. Mishra, "DATA MINING IS A PERPETUAL CONCEPT FOR LIBRARY AND INFORMATION SCIENCE : AN ESTIMATED OVERVIEW," vol. 5, no. September, pp. 14–21, 2015.
- [15] S. Nicholson, "The Bibliomining Process: Data Warehousing and Data Mining for Library Decision-Making," *Inf. Technol. Libr.*, vol. 22, pp. 1–10, 2003.
- [16] B. R. Karno, S. A. Noordin, A. Talib, and S. A. Rahman, "Bibliomining in UTM digital library : Using data mining to discover user pattern behavior," pp. 1–11, 2009.
- [17] P. Hajek and J. Stejskal, "Analysis of user behavior in a public library using bibliomining," *Adv. Environ. Comput. Chem.*, pp. 339–344, 2012.
- [18] R. K. Dwivedi and R. P. Bajpai, "Use of Data Mining in the field of Library and Information Science : An Overview," pp. 11–13, 2004.
- [19] C.-C. Chang and R.-S. Chen, "Using data mining technology to solve classification problems: A case study of campus digital library," *Electron. Libr.*, vol. 24, no. 3, pp. 307–321, 2006.
- [20] J. Littman and L. Connaway, "A Circulation Analysis of Print Books and E-Books in on Academic Research Library," *Libr. Resour. Tech. Serv.*, vol. 48, no. 4, pp. 256–263, 2004.
- [21] V. A. Moses, C. S. Jeevan, and S. Kala, "Library Management System Using Association Rule Mining," pp. 10–16, 2011.
- [22] R. Talib, M. K. Hanif, S. Ayesha, and F. Fatima, "Text Mining: Techniques, Applications and Issues," *IJACSA) Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 11, pp. 414–418, 2016.
- [23] M. Zhang, "Application of Data Mining Technology in Digital Library," vol. 6,

- no. 4, pp. 761–768, 2011.
- [24] V. Uppal and G. Chindawani, “An Empirical Study of Application of Data Mining Techniques in Library System,” *Int. J. Comput. Appl.*, vol. 74, no. 11, pp. 42–46, 2013.
- [25] K. Baba, M. Kitajima, and T. Minami, “An Evaluation of Book Selection in a University Library by Loan Record Analysis,” *Int. Conf. Educ. Technol. Comput.*, vol. 5, no. 10, 2015.
- [26] F. Xie, “The design and implementation of intelligent library management system based on RFID / GRPS,” vol. 6, no. 4, pp. 839–844, 2014.
- [27] D. Hand, H. Mannila, and P. Smyth, *Principles of data mining*, vol. 30. 2001.
- [28] “WEKA.”
- [29] K. Bussaban and K. Kularbphetpong, “Analysis of Users ’ Behavior on Book Loan Log Based On Association Rule Mining,” *Wold Acad. Sci. Eng. Technol.*, vol. 8, no. 1, pp. 18–20, 2014.
- [30] G. Mehta, D. Sharma, and E. Chauhan, “Article: Application of Incremental Mining and Apriori Algorithm on Library Transactional Database,” *Int. J. Comput. Appl.*, vol. 73, no. 8, pp. 12–18, 2013.
- [31] T. Anuradha, K. Sathyatej, and S. S. A. Krishana, “Frequent Pattern Mining for Efficient Library Management,” *Int. J. Comput. Sci. Eng. Freq.*, vol. 3, no. 11, pp. 3582–3587, 2011.
- [32] L. Yan, W. Cunrui, and W. Nan-nan, “Application of Apriori Association Rule Mining Algorithm in University Management System,” vol. 13, no. 5, pp. 1–4, 2016.
- [33] M. Bansal and M. Kaur, “Analysis and Comparison of Data Mining Tools Using Case Study of Library Management System,” *Int. J. Inf. Electron. Eng.*, vol. 3, no. 5, pp. 466–469, 2013.
- [34] J. V Joshua, O. D. Alao, A. O. Adebayo, G. A. Onanuga, E. O. Ehinlafa, and O. E. Ajayi, “Data Mining : A Book Recommender System Using Frequent Pattern Algorithm,” vol. 3, no. 3, pp. 1–13, 2016.

- [35] A. Dutt, M. A. Ismail, and T. Herawan, "A Systematic Review on Educational Data Mining," *IEEE Access*, vol. 5, no. c, pp. 15991–16005, 2017.
- [36] M. J. Zaki and W. Meira, *Fundamental concepts and Algorithms*. Cambridge University Press (May 12, 2014), 2014.
- [37] D. Almazro, G. Shahatah, L. Abdulkarim, M. Kherees, R. Martinez, and W. Nzoukou, "A Survey Paper on Recommender Systems," 2010.
- [38] E. Neuhold, C. Niederée, A. Stewart, and I. Frommholz, "The role of context for information mediation in digital libraries," *Digit. Libr.*, vol. 5, no. 11, 2005.
- [39] S. Rajagopal and A. Kwan, "Book Recommendation System using Data Mining for the University of Hong Kong Libraries," no. i, pp. 1–8, 2011.
- [40] C.-C. Chen and A.-P. Chen, "Using data mining technology to provide a recommendation service in the digital library," *Electron. Libr.*, vol. 25, pp. 711–724, 2007.
- [41] S. Parvatikar and B. Joshi, "Online book recommendation system by using collaborative filtering and association mining," *2015 IEEE Int. Conf. Comput. Intell. Comput. Res. ICCIC 2015*, pp. 6–9, 2016.
- [42] G. Brock, V. Pihur, S. Datta, and S. Datta, "clValid , an R package for cluster validation," no. March 2008, pp. 1–32, 2011.
- [43] I. J. Pelczar, "Identification of rainfall patterns over the Valley of Mexico," no. 1997, pp. 1–9, 2008.
- [44] D. James, "Package ' RMySQL ,'" 2015.
- [45] M. Hahsler, K. Hornik, and C. Buchta, "Introduction to arules – A computational environment for mining association rules and frequent item sets," *J. Stat. Softw.*, vol. 14, pp. 1–25, 2005.
- [46] M. Hahsler and S. Chelluboina, "Visualizing Association Rules: Introduction to the R-extension Package arulesViz," *R Proj. Modul.*, 2011.

