# Data Mining Course in Information System Department– Case Study of King Abdulaziz University

**[1]Farrukh Saleem, [2]Areej Malibari**

[1&2] Information Systems Department
Faculty of Computing and Information Technology
King Abdulaziz University, Jeddah, Saudi Arabia

**[1] farrukh800@yahoo.com , [2] malibari.areej@googlemail.com**

***ABSTRACT* - Data mining (DM) is an essential course to be included in the curriculum of Information systems education. This paper highlights the importance of data mining course in the Information System (IS) department. We discussed the issues related with data mining texts books and practical tools, by using our previous experience with the students of Information System (IS) Department, Faculty of Computing & Information Technology (FCIT) in King Abdulaziz University (KAU) Jeddah. The paper focuses on importance of DM course, relation of DM course with IS department, discussion about selection of the text book, and finally discussion on different DM tools can be use in the labs. This paper provides comprehensive information on DM course, text books, and tools. This paper specially guides the DM students to select better course material in the form of text book and DM software tool for practical implementation of all data mining tasks.**

***Keywords* -** Data mining, Course Material, Lab Tools.

## 1. INTRODUCTION

Data mining course plays an important role in the curriculum of IS department due to its effective implementation over management information systems, databases and data warehouses. Therefore, we presents in this paper about the importance of the course and the material we can use while studying data mining. Moreover, this paper shows clear guidance for the students for the selection of data mining text book and tool. Before moving over the discussion points, definitions, tasks and techniques of data mining has presented here using literature review, for more clear guidance on this topic.

Every day, several data has been gathered and being saved in large databases consciously or unconsciously. Most commonly data mining is there to use this large data warehouses for the extraction of hidden information from it [16]. There are several data mining techniques and algorithm available to plan proper evaluation process and extract golden information from these raw data.

Mainly, data mining tasks has been divided into descriptive and predictive methods. Classification, clustering and rule association mining are most common techniques use for predictive and descriptive analysis [1].

Furthermore, these main data mining tasks has further subcategorized into several others methods and algorithms accordingly. The need of algorithms is indeed depends on the required analysis and results, can be either predictive or descriptive.

Specifically, this paper highlights the importance of data mining course for the IS department in above mentioned university. Discussion about the importance of data mining course in general and especially in IS department presented in the immediate sections. In later sections we presented the comparison and selection of data mining text books and practical tools for teaching purpose. Specially, for the lab we try to learn our students with a tool represents comprehensive understanding to deal with the data and to provide them a professional environment which can help in their professional work after graduation.

## 2. IMPORTANCE OF DATA MINING IN GENERAL POINT OF VIEW

Data mining is a technique to find hidden rules from large amount of data like data warehouses. "Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both [2]." There are number of data mining tools available to deal with different type of data using several parameter and angles, to find out association between different parameters, convert data in several clusters according to using proper characteristics [1, 4].

Nowadays, technology advances in communication, storage and processing capabilities, which provide a facility to store more digital data. Data mining is the process to extract valuable information from such a large databases. Data mining applications have been broadly implemented to solve the problem in education, banking, marketing strategies, and

production industries, which shows that data mining has large impact in the society [3].

Lots of data is being scattered, gathered and warehoused related to e-commerce, purchase department, sales department, bank and credit card customers data, need to be evaluate properly[16]. Data mining tasks are there to analyze these large data warehouses using several techniques [1, 4] to provide advantageous knowledge for the organizations.

Moreover, for the students of IS department this paper prodive them valuable information in the selection of DM tool and text books.

## 3. SOME COMMON TASKS OF DATA MINING

### 3.1 CLUSTERING

Clustering is one of the common task of data mining, work on unsupervised data (no predefined classes). Clustering is a collection of data objects, clustered by taking similar object to one another within the same cluster, and dissimilar to the objects related in other clusters. Cluster differentiate by using similarities between data according to the characteristics found in the data and grouping similar data objects into clusters [4].

### 3.2 CLASSIFICATION

Classification is another common task of data mining. In classification we need a data has already defined a label (target) attribute. Firstly we divide the classified data into two sets; training and testing data [4]. Each data sets contains several attributes in which one of the attributes defined as class label. Jiawei Han [4] described classification task in two steps process; first is model construction and the second is model usage. The main goal of this technique is to assign previously unseen records to class as accurately as possible. We use testing data to find that check either the built model is accurate or not. While training data set is use to build the model on the other hand testing data set is use to validate the model [1].

### 3.3 ASSOCIATION

Data can be use to find association between several attributes, generate rules from data sets, this task is known as association rule mining. Given a set of transactions, find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction. The goal of association rule mining is to find all rules having support ≥ minsup (minimum support) threshold and confidence ≥ minconf (minimum confidence) threshold [1].

Moreover, association rule mining can be viewed as a two-step process, first, find all frequent itemsets: items satisfying minimum support. Second, generate strong association rules from the frequent itemsets: these rules must satisfy minimum support and minimum confidence [4].

## 4. DATA MINING COURSE IN INFORMATION SYSTEM DEPARTMENT

Courses include in IS department are focus on the business environment including, people, processes, data, information, businesses, and information technology. The main concern of IS department is to learn the students, database management systems, as well as information and computer technology. Database management systems, Management information system, data mining and decision support systems are some common courses include in IS department.

An information systems discipline therefore is supported by the theoretical foundations
of information and computations such that learned scholars have unique opportunities to explore the academics of various business models as well as related algorithmic processes within a computer science discipline [5]. Therefore, Information Systems are being implemented in the organizations to support and improve their processes. Specially, high performance and efficiency is always being a motto of every organization, IS plays a significant role for increasing the performance and making the process better. Concerning the main objective of IS in the organization, data mining is a course which provide a facility to learn such methods which can be beneficial for the development of any organization. Classification, clustering and regression are some fundamental techniques use for analyzing the huge amount of data for extracting hidden gold from it.

Fulfilling the requirement of IS department, Data Mining has became essential course to be included in the IS curriculum, as data mining has tasks to analyze the stored data in any format and extract new information from it, for making departmental activities more stronger and beneficial for the organizations. In fact, this course is available for both type of backgrounds; either technical or business.

## 5. DATA MINING BOOK - COMPARISON AND SELECTION

We reviewed many data mining books [1, 4, 6, 15] to be prescribed as the textbook for the course. The criteria include topic coverage, information presentation, instructor's resources, target audience and the book's usage as a textbook by other universities, in the given order of importance. While a very good data mining read, [15] is not really designed as a textbook for an undergraduate level data mining course. No instructor's resources are present and data warehousing is not absent in the book. The topics covered in [6] matches our syllabus but its target audience are Management Information System students rather than Computer Information Systems students, hence the book lacks the technical depth of the algorithms covered. Both [1] and [4] are among the most famous data mining textbooks [8], but the topic of data warehousing is not covered in [1]. For the above mentioned criteria, we decided to use [6] as the textbook for the course. The book is a comprehensive overview to data mining and warehousing and it can be used for an introductory as well as advanced data mining course. It discusses the key concepts in details and provides a thorough understanding of the different techniques used in data mining. It discusses data, exploratory data analysis, and the techniques used to store and mine this data in detail. Proper attention is given to the major concepts of classification, clustering and association analysis with an excellent coverage of the different algorithms used. Data warehousing, OLAP and cubes are also covered in detail. The book comes with a companion website [9], that includes lecture slides, instructor's manual and course plan among other resources.

## 6. DATA MINING TOOL - COMPARISON AND SELECTION

Practical implementation of theory classes covers the major part of studies in computer science education. For data mining course we reviewed many tools [12, 13, 17] for the practical implementation in the computer labs. Selection of the tool is depends mainly on the topics covering in theory classes to facilitate students by practical execution in the lab's session. Moreover, easy to use and installation steps can also consider important criteria in the selection of practical tool. Initially in the first term, we selected the DM tool ODMiner version 10.1.0. [12], developed by ORACLE [18]. ODMiner has the graphical user interface that help data analysts to use the application in an easy way. While using [12] we experienced some issues raised by students regarding installation of ODMiner as we found that ODMiner cannot work without Oracle database installation as mentioned in the installation guide, issued by Oracle Technology Network [12]. The second issue as Oracle database/developer packages requires network identifications/authentications at the time of installation steps [12]. Moreover, compatibility between Oracle database and ODMiner versions was also the issue, as both should have the same version [12]. According to the syllabus, ODMiner version 10.1.0, not support range of classification, association and clustering tasks [11]. These point was majorly highlighted from the students that having tough time in the installation process for those students who have not enough knowledge about Oracle databases.

Switching to another tool was not an easy decision; although for the next term we already had experienced from [12], therefore this time we need to be careful and keeping in mind the problems we faced. For the current term, we reviewed RapidMiner[13] and Weka[17] to be prescribed in the replacement of ODMiner. We found that both [13] and [17] has no major issue regarding installation and hardware requirements as well both provide better GUI interface covering range of operators supporting data mining tasks. Finally we selected [13] considering more advantageous than [17]. As RapidMiner provides a facility to implement Weka's methods/operators to be connected with RapidMiner using its own operators [13].

RapidMiner is an open source package provides good range of major tasks using in data mining; includes, regression, clustering, classification, and association with a good range of sub algorithms too. In addition, we can import several types of data in RapidMiner for the implementation of data mining tasks as well as after analysis we can export data, result and images. Data integration, analytical Extract, Transform and Load (ETL), data analysis and reporting options are also included in rapid miner.

RapidMiner provides an integrated development environment (IDE) (also known as integrated design environment or integrated debugging environment). RapidMiner as a software application that provides comprehensive facilities to computer programmers for software development showed in figure 1 [14].

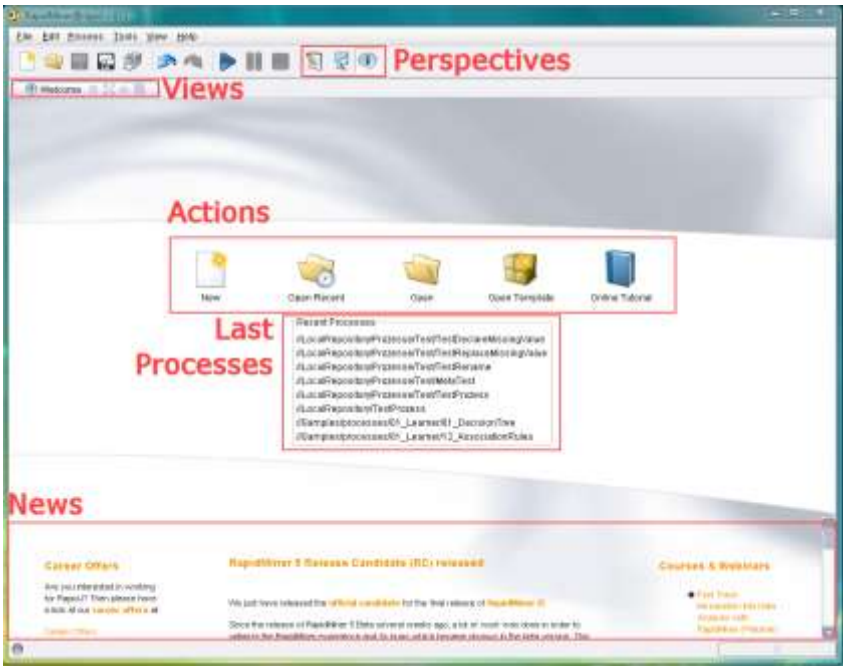


Figure-1 Rapid Miner's Interface [14]

With a huge range of operators, sample data and processes provided by RapidMiner in figure 2 , 3[14], made our decision strong to select this tool for practical implementation and give comprehensive understanding about data mining tasks.

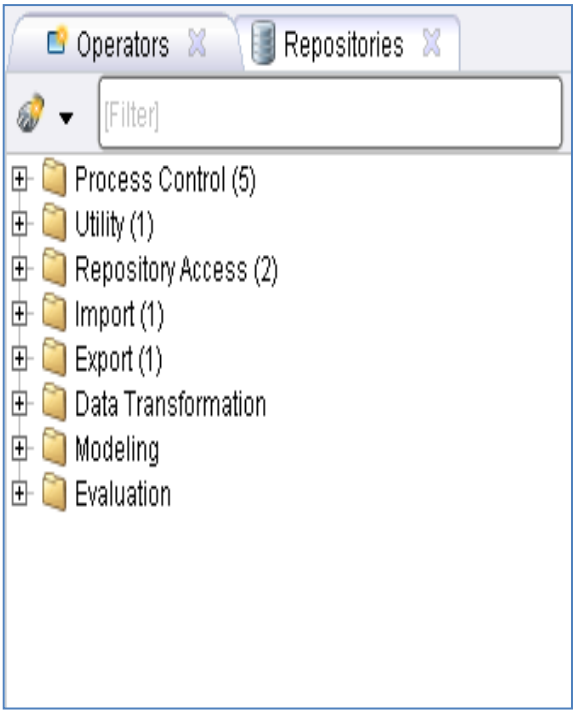


Figure-2 Sample Data and Process [14]

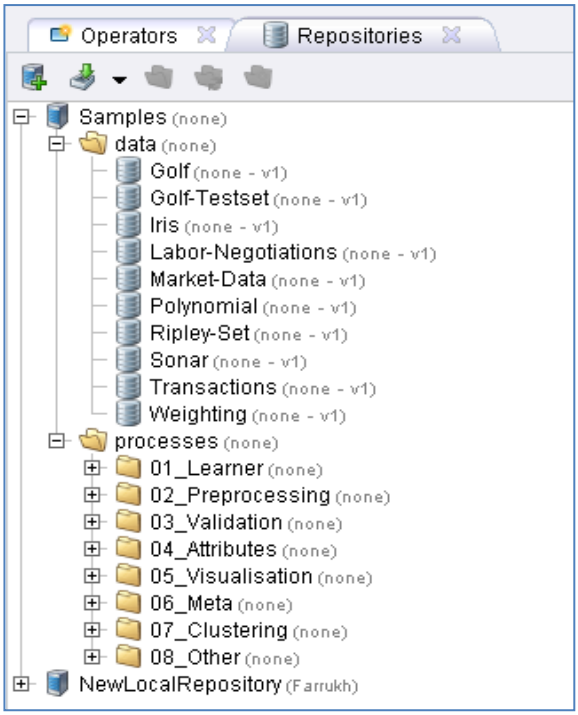


Figure-3 List of Operators [14]

## 7. CONCLUSION

Data mining is an obligatory course for every student studying in IS department in the FCIT college. To provide quality and creative education we as a course coordinators need to follow course objective in the selction of the books and practical tools. At the moment, we applied our best efforts to make the things easier to understand and applicable. Although, current book and lab tool covering all the topics included in the course syllabus but there can be possibility to improve the course material and tool. For this, we will continue our efforts according to student's feeback and education requirements in the next semesters. However, in the current semester we asked from the students to submit their experience for prescribed text book and DM lab tool. In this regards, complete statistical analysis using student's feedback and suggestions are our main concern in the near future.

## REFERENCES


[1] Pang-Ning Tan, Michael Steinbach & Vipin Kumar, "Introduction to Data Mining", Addison Wesley, 2005, ISBN 0321321367

[2] http://www.anderson.ucla.edu/faculty/ jason. frand/teacher/technologies/palace/datamining.htm, accessed date, 25th March, 2011.

[3] http://www.sigkdd.org/curriculum.php, accessed date, 25th March, 2011.

[4] Jiawei Han, Micheline Kamber, Jian Pei, "Data Mining: Concepts and Techniques", 2nd edition, 2005, Morgan Kaufmann, ISBN 1558609016.

[5] Denning, Peter (June 2007). Ubiquity a new interview with Peter Denning on the great principles of computing. 2007. pp. 1–1

[6] George M. Marakas, "Modern Data Warehousing, Mining, and Visualization: Core Concepts", 1st edition, 2002, Prentice Hall, ISBN 0131014595, Page ix.

[7] http://www.amazon.com/Data-Mining-Concepts-TechniquesManagement/product reviews/1558609016/ref=cm_cr_dp_all_summary?ie=UTF8&showViewpoints=1&sortBy=bySubmissionDateDescending, accessed date, 25th March, 2011.

[8] http://www.kdnuggets.com/polls/2005/data _mining_textbooks.htm, accessed date, 30th March, 2011.

[9] http://www.cs.uiuc.edu/~hanj/bk2/, accessed date 4th April, 2011.

[10] http://fcit.kau.edu.sa/moodle/course/view .php?id=43, accessed date, 1st May, 2011.

[11] http://www.oracle.com/technetwork /database/options/odm/downloads/index.html, accessed date, 1st May, 2011.

[12] http://www.oracle.com/technetwork/ database/options/odm/index.html, accessed date, 30th April, 2011.

[13] http://www.rapidminer.com/, accessed date, 5th May, 2011.

[14] http://rapidi.com/content/view/ 181/196/, accessed date, 10th May, 2011.

[15] Kantardzic, M., "Data Mining: Concepts, Models, Methods, and Algorithms" 1st edition, 2002, Wiley IEEE press.

[16] http://www.laits.utexas.edu/~norman/BUS.FOR /course.mat/Alex/.

[17] http://www.cs.waikato.ac.nz/ml/weka/ , accessed date, 10th May, 2011.

[18] http://www.oracle.com/index.html, accessed date, 5th May, 2011.