CZECH UNIVERSITY OF LIFE SCIENCES PRAGUE

FACULTY OF ECONOMICS AND MANAGEMENT

DEPARTMENT OF INFORMATION TECHNOLOGIES

BACHELOR THESIS APPENDICES

DATA MINING

Author: Arezoo Noushiravan

The Bachelor Thesis Supervisor: Ing. Martin Havránek, Ph.D.

Table of Contents

Summary

This thesis is concerned with Data Mining: 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. Data mining is primarily used today by companies with a strong consumer focus - retail, financial, communication, and marketing organizations. It enables these companies to determine relationships among "internal" factors such as price, product positioning, or staff skills, and "external" factors such as economic indicators, competition, and customer demographics. And, it enables them to determine the impact on sales, customer satisfaction, and corporate profits. Finally, it enables them to "drill down" into summary information to view detail transactional data. The initial phase of a data mining project focuses on understanding the project objectives and requirements. Once you have specified the project from a business perspective, you can formulate it as a data mining problem and develop a preliminary implementation plan. For example, your business problem might be: "How can I sell more of my product to customers?" You might translate this into a data mining problem such as: "Which customers are most likely to purchase the product?" A model that predicts who is most likely to purchase the product must be built on data that describes the customers who have purchased the product in the past. Before building the model, you must assemble the data that is likely to contain relationships between customers who have purchased the product and customers who have not purchased the product. Customer attributes might include age, number of children, years of residence, owners/renters, and so on.

Thesis Objective

The bachelor thesis is thematically focused on an in-depth coverage of the principles of data mining. The main goal of this bachelor thesis is to understanding the concepts of data mining and data mining techniques, the process of managing and extracting data, analyzing and establishing hidden connections and patterns in a group of data sets from different perspectives and summarizing it into useful information.

Partial goals of the thesis are:

- to provide an introduction to the multidisciplinary field of data mining
- to describe data mining tasks
- to describe data mining techniques
- to analyze practical applications of data mining

Thesis Methodology

Methodology of the thesis is based on study and specialized information resources. The practical part aim is to analyze practical applications of data mining which will be used in business intelligence, industries, computer security, etc. Based on synthesis of theoretical knowledge and results of author's own work, the conclusions of the thesis will be formulated.

Own Solution

In the solution we have incorporated the technologies for the sales system platform. Using data mining, Companies are able to maximize the effectiveness of their marketing, where data mining provide profit to the companies. We have analyzed the technologies involved from the perspective of the front-end and back-end as a law the layers of access could be implemented. Data Mining based on marketing is an ideal service which could be useful in providing profit and benefits to the owner of companies who already run their business.

Furthermore data mining for the system is a powerful tool in order to manipulate the information of customers to extract meaningful and profitable information. For example, suppose a Sales Manager wants to analyze data sets which are located in the data base.

Following are the steps that the sales manager will need to achieve his/her goal. The system we plan to implements enables to possibility for the sales manager to use their own applications to mining the useful information. In the system customers are able to log-in, select item, search item, buy item, and pay bills and finally all of data will be allocated in the company database where the knowledge engineer as a data miner with an understating of data mining technologies tries to extract meaningful information by a specific applications in order to get more profit for the company and the data expert is supervising the procedure of allocating data in the database.

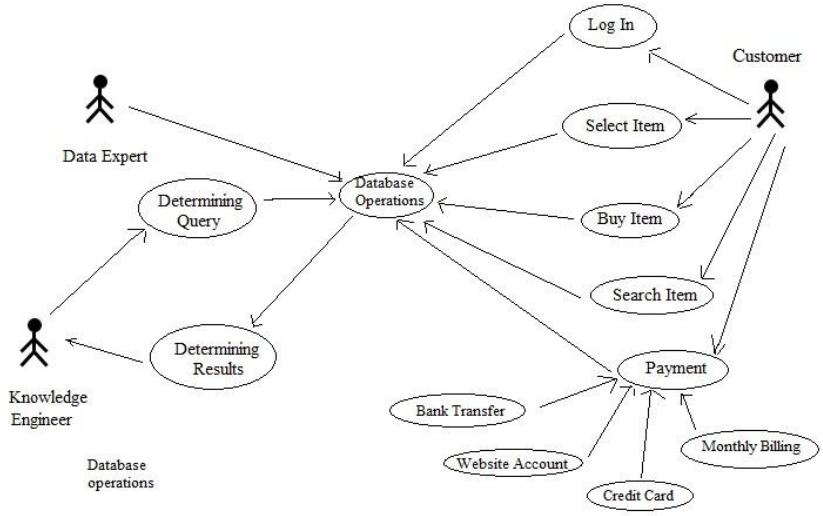

Figure 1- Use Case Diagram

Conclusion

Databases today may contains more than 1000^4 Terabyte (TB) of data and human analysts is unable to discover importance information or potentially useful information that can be hidden within these huge of data, therefore scientists with a many kind of data analysis tools tries to extract valid model and relationships in data in order to predict future trend or valid predictions.
Particularly innovative organization worldwide already tend to use data mining methods to find higher value customers in order to reconfigure their product offerings, to increase level of sales which can be lead to more profit, to minimize the cost of products, losses due to a specific error and also to increase revenues in company lifetime.

Scientists in order to utilize Data mining to solve their problem at first they should prepare and clean data from any error or missing values according their needs and also they need to know the exact problem in order to choose appropriate datamining method since Data mining has two different methods Prediction and Description Methods that both is used in a specific area, and according to the problem and their interesting goal apply appropriate data mining method.

Scientists or analyzers in prediction Methods such as Classification, Deviation Detection, and Regression use an interesting variables to predict unknown class of objects or value of a specific variable and in description method such as Association Rule Discovery, Sequential Pattern Discovery, Clustering analyzers extract interpretable patterns in order to describe the data.

Based on my practical example, RapidMinder software is used as a data mining tool that provides various technique and algorithms to be applied on data sets. In practical part decision tree had been chosen to classify customers according to type of class due to many reasons. Decision tree excellent predictive models in data mining, Decision tree is also able to handle both continuous and discrete variables, it is easy to interpret for small-sized trees and extremely fast at classifying unknown objects, to help predict the future trends and is easy to understand, also work more efficiently with discrete attributes. Briefly Decision tree generates predictions for the scoring observations.

Also decision tree is a popular technique for supervised classification, especially when the results are interpreted by human. And as I noticed, in Decision tree method Irrelevant attributes may affect badly the construction of a decision tree (E.g. ID numbers).

Most of retailers can use the decision tree approach to deal in marketing situations and identify important components of the decision process.

References

1. ***wikipedia.*** **[Online] https://en.wikipedia.org/wiki/Data.**

2. What is the difference between categorical, ordinal and interval variables? ***UCLA (Institute for Digital Research and Education).*** **[Online] http://www.unesco.org/webworld/idams/advguide/Chapt1_3.htm.**

3. Jiawei Han and Micheline Kamber. ***Western Michigan University.*** **[Online] 2000. https://cs.wmich.edu/~yang/teach/cs595/han/ch01.pdf.**

4. S. Sumathi, S.N. Sivanandam. . ***Introduction to Data Mining and Its Applications.*** **Oct 12, 2006.**

5. Data Mining Concepts. ***microsoft.com.*** **[Online] https://msdn.microsoft.com/en-us/library/ms174949.aspx.**

6. ZAKI, MOHAMMED J. SPADE: An Efficient Algorithm for Minings. ***philippe Fournier-Viger.*** **[Online] 2001. http://www.philippe-fournier-viger.com/spmf/SPADE.pdf.**

7. Aggarwal, Charu C. ***Data Mining.*** **New York, 2015.**

8. Building Classification Models: ID3 and C4.5. [Online] http://cis-linux1.temple.edu/~giorgio/cis587/readings/id3-c45.html.

9. k-means clustering algorithm. ***Data Clustering Algorithms.*** **[Online] https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm.**

10. Rakesh Agrawal - Ramakrishnan Srikant. Rakesh Agrawal. [Online] http://rakesh.agrawal-family.com/papers/vldb94apriori.pdf.

11. wikipedia. [Online] https://en.wikipedia.org/wiki/Regression_analysis.

12. Dr. Saed Sayad. KNN - Classification. ***Data Mining.*** **[Online] http://www.saedsayad.com/k_nearest_neighbors.htm.**

13. ***Wikipedia.*** **[Online] http://en.wikipedia.org/wiki/Anomaly_detection.**

14. Zaki, Mohammed J. Efficient Enumeration of Frequent Sequences. ***Ca' Foscari University of Venice.*** **[Online] Ca' Foscari University. http://www.dsi.unive.it/~dm/CIKM98_ps.pdf.**

15. Microsoft. Data Mining Algorithms . ***Microsoft.*** **[Online] https://msdn.microsoft.com/en-us/library/ms175595.aspx.**