Data Mining Research Challenges in E-Commerce

Md. Zahid Hasan
Green University of Bangladesh

Mohiuddin Ahmed
Green University of Bangladesh

Md. Elias Mollah
Green University of Bangladesh

Abstract

The concepts and techniques of data mining, a promising and flourishing frontier in database systems and new database applications. The major reason that data mining has attracted a great deal of attention in information industry in recent years is due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. In this paper we are focusing the important role of business, based on data mining knowledge development for detecting the e-commerce and express some research directions in these areas in order to facilitate the reader's comprehension of their respective roles in data mining.

Keywords: Data Mining, Data Engineering, Knowledge Discovery, e-commerce, web mining, business intelligence

1. Introduction

Data mining, also popularly referred to as knowledge discovery in databases (KDD), is the automated or convenient extraction of patterns representing knowledge implicitly stored in large databases, data warehouses, and other massive information repositories and applying multidisciplinary field including database technology, artificial intelligence, machine learning, neural networks, statistics, pattern recognition, knowledge based systems, knowledge acquisition, information retrieval, high performance computing, and data visualization[1]. The revolution of data mining tools and the concept of Electronic Commerce (E-Commerce) is growing fast which grow the interests of many companies to improve the online experience [2]. At the same time, faster and cheaper storage technology ahead us to store greater amounts of data online, and better database-management-system software provides easy access to those databases [3][4]. Many industrial and researchers are interested and invested billions of dollars in the area of business intelligence which encompasses data mining. To ensure that the advances of data mining research and technology will effectively benefit the progress of science and engineering, it is important to examine the challenges on data mining posed in data-intensive science and engineering and explore how to further develop the technology to facilitate new discoveries and advances in science and engineering [5]. Many application domains of data mining are following the standard data mining process. The process incorporates the accumulates of information or data cleaning of mapping data to a single naming convention, uniformly representing and handling missing data, and handling noise and errors when possible, data pre-processing and transforming a subset of data to a flat file, building one or more models that can be predictive models, clusters or data visualizations that lead to the formulation of rules as shown in figure 1 [6].

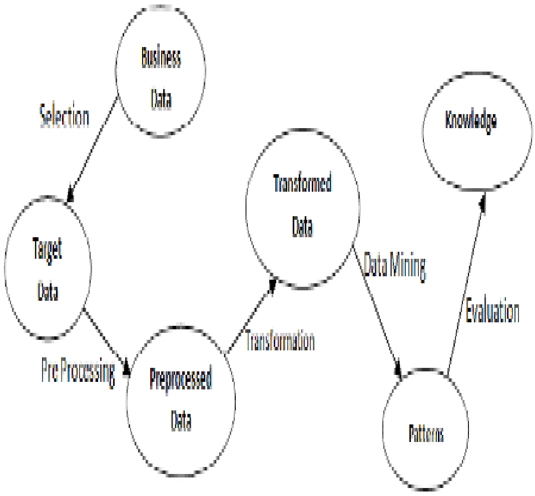

Figure 1: An overview of the data mining process of Business perspective.

Finally, the article enumerates challenges for future research and development and in some insights into future directions.

2. Objectives of Data Mining for E-commerce

The traditional method of turning data into knowledge relies on manual analysis and interpretation. For example, in the Web server to manipulate billions of information, it is common for server analysts of the system to periodically analyze current trends and changes in the information. The analyst then provides a report detailing the analysis to this organization or institution; this report becomes the basis for future decision making and planning for systems. So, manual probing of a data set is slow, expensive and highly subjective. The volumes of data are growing dramatically. The sizes of Databases are increasing in two ways: (1) the number N of records or objects in the database and (2) the number d of fields or attributes to an object. Databases containing on the order of $N = 10^9$ objects are becoming increasingly common, for example, in the astronomical sciences. Similarly, the number of fields d can easily be on the order of 10^2 or even 10^3 [7]. With the development of Google and other effective web search engines become an important role of E-Commerce. In a statistics, it is seen that the reported size and page depth of Google search engine is 8.1 million and 101k [9] which is gradually increases day by day. Data mining is a term coined to describe the process of sifting through large databases in search of interesting patterns and relationships. Practically, Data Mining provides tools by which large quantities of data can be automatically analyzed. Data Mining can be considered as a central step of the overall process of the Knowledge Discovery in Databases (KDD) process. Several researchers have proposed different ways to divide the KDD process into phases. But the hybridization of these proposals and breaking the KDD process integrate with business mind as follows [6] [8]:

1. Business Understanding: The goal of this phase is to gain a deeper understanding of the project objectives and further circumstances strictly from the business perspective and finally the initial phases are turning to be a data mining problem definition.
2. Selecting Useful Data: To select data set on which discovery is to be performed.
3. Data Preprocessing: This stage includes operations for Dimension Reduction (such as Feature Selection and Sampling), Data Cleaning (such as Handling Missing Values, Removal of Noise or Outliers), and Data Reduction and projection.
4. Choosing appropriate method: Based on the identified business goals and the assessment of the available data choosing an appropriate data mining task with a particular data mining method such as: classification, regression, clustering and summarization.
5. Choosing the Data Mining algorithm: This stage includes selecting the specific method to be used for searching patterns and decides which models and parameters of the methods used may be appropriate.
6. Evaluating and interpreting the mined patterns and visualization of the data based on the technical point of view and able to interpret the prepared dataset and designated information actually given to the algorithm.
7. Deployment: After mining the data and assessing the data mining results one need to transfer the results back into the business environment.

3. Research Challenges

In this section there will be an extensive discussion on the research challenges of Data mining. To simplify the concept of data mining, we point out the current and present research challenges.

3.1 Mapping Business question to make data transformation easily

Today it is very complex mechanism to map business questions to data transformation. The researchers make it easier while built a user interface that supports many useful transformations; fundamental operations like aggregations remain a complex concept to grasp. The programmers in early 70s were designing language called SQL for non-programmer which is very beneficial for the non programmers which is very beneficial for the non-programmer to interact with databases [10]. Now the programmers have to be built a transformation language and user interface that is significantly easier.

3.2 To construct automate feature

A mix of domain knowledge and a data miner's expertise is needed to construct feature. The goal of feature construction is therefore to add features that have large information gains. That is the added features will be selected by the learning algorithm and resulted classification models will be more accurate. While we are able to provide many features for our domain, we build hundreds of unique attributes with every client as a customer signature to run. Now these should be easier to construct the features automatically [11].

3.3 Build Intelligible Model

The objective of data mining is to provide an interesting insight with business users. Our building models which are restricted the ability to perceive clearly, such as decision trees, decision rules and Naïve Bayes. Now the other approachable model have to be build which are easy to understand by business user.

3.4 Data Transformation

There are two significant rules to transfer the data that need to place the first data must be brought from operational system to data warehouse, and second data may need to transform to answer a specific business question and process some operations such as new columns define, data binning, and aggregating it [13]. While the first set of transformations need not to be frequently modified. But the second set of transformations provides a significant challenges faced by many data mining tools today.

3.5 Scalability of Data Mining Algorithm

Data mining tools handles with large amount of data where two scalability issues are needed. (a) Most data mining algorithms cannot process the amount of data gathered at websites in reasonable time because they scale nonlinearly and (b) generated models are so much complicated for humans to understand [14].

3.6 Experiment because correlation does not imply causation

When interpreting the results it is often the case that the variables do not automatically imply that correlation is confused with causality. Business users need to be made aware that correlation does not necessarily imply causality. It should conduct control experiments to establish a casual relationship.

3.7 Explain counter intuitive insights

In a few occasions, it becomes to present insights that are counter intuitive. For an example of Simpson's Paradox [12] to analyze the client's data. It occurs when the correlation between two variables is reserved when a third variable is controlled.

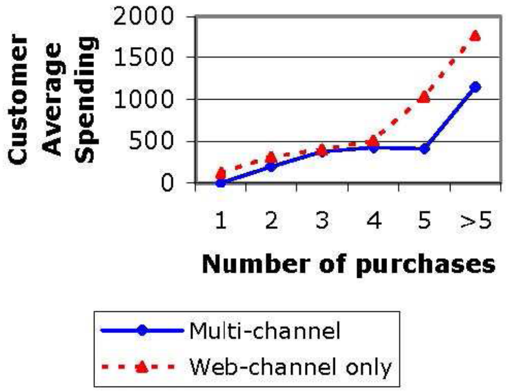

Figure 2: Customer Average spending vs. Number of purchases on Multi channel and Web channel [12]

In the above figure 2, the reversal trends is happening because a weighted average is being computed and it is seen that the number of customers who shopped on the web more than five times is much smaller than who shopped more than five times across multiple channels. Such insights must be explained to business users.

3.8 Estimate the ROI (Return on Investment) of insights

Data mining is evolving quickly into a mainstream that puts predictive insight in the hands of decision maker. The evolution is driven by the tangible return on investment (ROI) that organizations of every industry are realizing. But it is difficult to compute a quantities value of ROI. While the potential business value of predictive anilities is clear, process for realizing that value- strategically deploying data mining results to create predictive analytic solutions are often less clear. Many organizations are successfully applying data mining to discover new insights, As for example in the scenario of large automotive manufacturer, they changed their websites and after changing resulted in a 30% improvement in revenue. However, in other cases, the successful improvement achieved by customer's browsing experience and satisfaction, so, the results of which are hard to measure quantitatively. And sometimes failing to execute decision based on those insights to achieve their desired return.

4. Conclusion

In this paper, we have presented how the data mining is applicable for improving the services provided by e-commerce based enterprises. We also highlighted the KDD process of the E-commerce perspective. Existing algorithms and features of data mining tools is not reliable for modeling and extracting the knowledge from the data warehouses and so the specified challenges in this paper need to be better addressed in real world.

5. References

[1] J. Han, J. Ago, "Research Challenges for Data Mining in Science and Engineering", in Next Generation of Data Mining, Taylor and Francis Group LLC 2008.

[2] S. Ansari, R. Kohavi, L. Mason, and Z. Zheng, "Integrating E-Commerce and Data Mining: Architecture and Challenges", IEEE International Conference on Data Mining, pp. 27-34, 2001, ISBN: 0-7695-1119-8.

[3] Piatetsky-Shapiro, Gregory, "The Data-Mining Industry Coming of Age", *in IEEE Intelligent Systems*, vol. 14, issue 6, Nov 1999. Doi. 10.1109/5254.809566.

[4] J. Hsu,"Rise of Data Mining: Current and Future Application Areas", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 1, September 2011 ISSN (Online): 1694-0814.

[5] Venkatadri.M and Dr. Lokanatha C. Reddy, "A Review on Data mining from Past to the Future", International Journal of Computer Applications, pp. 19-22, Volume 15, No.7, February 2011.

[6] Fayadd, U., Piatesky -Shapiro, G., and Smyth, P. "From Data Mining To Knowledge Discovery in Databases", AAAI Press / The MIT Press, Massachusetts Institute Of Technology. ISBN 0–26256097–6. MIT 1996.

[7] Vladimir Brusic and John Zeleznikow, "Knowledge discovery and data mining in biological databases", journal of The Knowledge Engineering Review, Volume 14 Issue 3, September 1999, pp. 257 - 277.

[8] Reinartz, T. 2002. "A unifying view on instance selection. Data Mining and Knowledge Discovery", An International Journal, 6(2), pp.191–210.

[9] David L. Banks and Yasmin H. Said, "Data Mining in Electronic Commerce", Journal of The Institute of Mathematical Statistics, Vol. 21, No. 2, May 2006, pp. 234-246.

[10] Ron Kohavi Lle,w Mason, Rajesh Parekh, Zijian Zheng , **"**Lessons and Challenges from Mining Retail E-Commerce Data**",** Journal Machine Learning , Vol. 57 Issue 1-2, October-November 2004, pp. 83-113.

[11] W. Lee, " A data mining framework for constructing features and models for intrusion detection system", PhD Thesis, Columbia University, 1999.

[12] R. Sanchati, P.C. Patidar, G. Kulkarni, "Path breaking case studies in E-commerce using data mining", IJCTEE, vol.1, issue 1, pp. 20-25.

[13] R. Kohavi, R. M. Henne, and D. Sommerfield, "Practical guide to controlled experiments on the web: Listen to your customers not to the hippo", in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2007, pp. 959-967.

[14] N. R. S. Raghavan, "Data mining in e-commerce: A survey", Journal Article ,Vol. 30, Parts 2 & 3, April/June 2005, pp. 275–289.