International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Data Mining In E-Commerce: A Survey

Ojaswi D. Jawarkar
Student, B.E. Final Year Computer Science and Engineering,
H.V.P.M, C.O.E.T, Amravati, India

Abstract*—Data mining has matured as a field of basic and applied research in computer science in general and e-commerce in particular. In this paper, I studied some of the recent approaches and architectures where data mining has been applied in the fields of e-comerce Data mining has matured as a field of basic and e-business. My intent is not to studied the plethora of algorithms in data mining; instead, my current focus being e-commerce, I limit my discussion to data mining in the context of e-commerce. I also mention a few directions for further work in this domain, based on the survey
Keywords: Data mining; e-commerce.

I. INTRODUCTION

E-commerce has changed the face of most business functions in competitive enterprises. Internet technologies have seamlessly automated interface processes between customers and retailers, retailers and distributors, distributors and factories, and factories and their myriad suppliers. In general, e-commerce and e-business (henceforth referred to as e-commerce) have enabled on-line transactions. Also, generating large-scale real-time data has never been easier. With data pertaining to various views of business transactions being readily available, it is only opposite to seek the services of data mining to make (business)sense out of these data sets. . I studied that once the back-end databases are properly designed to capture customer buying behaviour, and provided that default data take care of missing and non-existent data, the first issue of availability of data with rich descriptions is taken care of. Similarly, the reliability of data collected is also ensured because it is possible to increase the so called no-touch-throughput in e-commerce transactions. The ROI in DM exercises related to e-commerce can be easily quantified. Improved webserver availability results in faster transactions, thus increasing the revenue. Studied that increasing the number of transactions directly results in improved profits. Data mining in e-commerce mostly relies on the controller for generating the data to m ine on. Thus integration issues also do not surface in this case. In summary, it is little surprise that e-commerce is the killer application for data mining (Kohavi 2001). This paper is organized as follows. I study various applications of these techniques for conducting DM in ecommerce, in II. The data collection and software architecture issues consititute III. Then I study case in IV, followed by conclusions and suggestions for further work.

II. E-COMMERCE AND DATA MINING

In this section , I studied articles that are very specific to DM in e-commerce. The salient applications of DM techniques are presented first. Later in this section data collection issues are discussed.

A. DM In Customer Profiling

It may be observed that customers drive the revenues of any organization. Acquiring new customers, delighting and retaining existing customers, and predicting buyer behaviour will improve the availability of products and services and hence the profits. Thus the end goal of any DM exercise in e-commerce is to improve processes that contribute to delivering value to the end customer. Consider an on-line store like http:www.dell.com where the customer can configure a PC of his/her choice, place an order for the same, track its movement, as well as pay for the product and services. With the technology behind such a web site, Dell has the opportunity to make the retail experience exceptional. At the most basic level, the information available in web log files can illuminate what prospective customers are seeking from a site. Are they purposefully shopping or just browsing? Buying something they're familiar with or something they know little about? Are they shopping from home or from work? what performance can be expected from the servers and network to support customer service and make e-business interaction productive. Companies like Dell provide their customers access to details about all of the systems and configurations they have purchased so they can incorporate the information into their capacity planning and infrastructure integration. Back-end technology systems for the website include sophisticated DM tools that take care of knowledge representation of customer profiles and predictive modelling of scenarios of customer interactions. For example, once a customer has purchased a certain number of servers, they are likely to need additional

routers, switches, load balancers, backup devices etc. Rule-mining based systems could be used to propose such alternatives to the customers.of a web personalization system. The author observe that the log data collected automatically by the Web and application servers represent the fine-grained navigational behaviour of visitors.

B. DM And Buyer Behavior In E-Commerce

For a successful e-commerce site, reducing user-perceived latency is the second most important quality after good site-navigation quality. The most successful approach towards reducing user-perceived latency has been the extraction of path traversal patterns from past users access history to predict future user traversal behavior and to prefetch the required resources. However, this approach is suited for only non-e-commerce sites where there is no purchase behaviour. Vallamkondu & Gruenwald (2003) describe an approach to predict user behaviour in e-commerce sites. The core of their approach involves extracting knowledge from integrated data of purchase and path traversal patterns of past users (obtainable from webserver logs) to predict the purchase and traversal behaviour of future users. Web sites are often used to establish a company's image, to promote and sell goods and to provide customer support. The success of a web site affects and reflects directly the success of the company in the electronic market. Spiliopoulou & Pohle (2000) propose a methodology to improve the success of web sites, based on the exploitation of navigation-pattern discovery. In particular, the author present a theory, in which success is modelled on the basis of the navigation behaviour of the site's users. In the context of web mining, clustering could be used to cluster similar click-streams to determine learning behaviours in the case of e-learning, or general site access behaviours in e-commerce. Most of the algorithms presented in the literature to deal with clustering web sessions treat sessions as sets of visited pages within a time period and do not consider the sequence of the click-stream visitation. This has a significant consequence when comparing similarities between web sessions.

III. DATA COLLECTION

A. Enabling Data Collection In E-Commerce.

It may be observed that there are various ways of procuring data relevant to e-commerce DM. Web server log files, web server plug-ins (instrumentation) application server instrumentation are the primary means of collecting data. Other sources include transactions that the user performs, marketing programs (banner advertisements, emails etc), demographic (obtainable from site registrations and subscriptions), call centres and ERP systems. It is quite common to expend about 80% of any DM effort in e-commerce in data filtering. This is largely in part to the heavy reliance on the web logs that are generated by the HTTP protocol. This protocol being stateless, it becomes very difficult to cull out customer buying behaviour-related information along with the product details. The proposed DM is found to dramatically reduce the preprocessing, cleaning, and data understanding effort. It emphasize the need for data collection at the application server layer and not the web server, in order to support tagging of data and metadata that is essential to the discovery process. It also describe the data transformation bridges required from the transaction processing systems and customer event streams (e.g. click streams) to the data warehouse.

.

B. Analysing Web Transaction.

Once the data are collected via any mechanisms, data analysis could follow suit. This could be done along session level attributes, customer attributes, product attributes and abstract attributes. Session level analysis could highlight the number of page views per session, unique pages per session , time spent per session, average time per page, fast vs. slow connection etc. Additionally, this could throw light on whether users went through registration, if so, when, did the users look at the privacy statement; did they use search facilities, etc. The user level analysis could reveal whether the user is an initial or repeat or recent visitor/purchaser; whether the users are readers, browsers, heavy spenders, original referrers etc. (Kohavi 2001). Hu & Cercone(2002) present a new approach called on-line analytical mining for web data. Their approach consists of data capture, webhouse construction, pattern discovery and pattern evaluation. The authors describe the challenges in each of these phases and present their approach for web usage mining. Their approach is useful in determining the most profitable customers, the difference between buyers and non-buyers, identification of website parts that attract most visits, parts of website that are session killers, parts of the site that lead to the most purchases, identifying the typical path of customers that leads to a purchase or otherwise etc. The webhouse is akin to the data warehouse.

C. Proposed DM.

In a B2B e-commerce setting, it is very likely that vendors, customers and application service providers (ASP) (usually the

middlemen) have varying DM requirements. Vendors would be interested in DM tailored for market basket analysis to know customer segments. On the other hand, end customers are keen to know updates on seasonal offerings and discounts all the while. The role of the ASP is then to be the common meeting ground for vendors and customers. Krishnaswamy et al (2000) propose a distributed DM that enables a DM to be conducted in such a naturally distributed environment. The proposed distributed data mining system is intended for the ASP to provide generic data mining services to its subscribers. In order to support the robust functioning of the system it possesses certain characteristics such as heterogeneity, costing infrastructure availability, presence of a generic optimization engine, security and extensibility. Heterogeneity implies that the system can mine data from heterogeneous and distributed locations. The proposed system is designed to support user requirements with respect to different distributed computing paradigms (including the client-server and mobile agent based models). A task that requires higher computational resources and/or faster response time should cost the users more on a relative scale of costs. Further, the system should be able to optimise the distributed data mining process to provide the users with the best response time possible (given the constraints of the mining environment and the expenses the user is willing to incur). Maintaining security implies that in some instances, the user might be mining highly sensitive data that should not leave the owner's site. In such cases, the author provide the option to use the mobile-agent model where the mining algorithm and the relevant parameters are shipped to the data site and at the end of the process the mobile agent is destroyed onthe site itself. The system is extensible to provide for a wide range of mining algorithms (Krishnaswamy et al 2000). The author provide a facility wherein the user can register their algorithms with the ASP for use in their specific distributed DM jobs.

IV. CASES IN E-COMMERCE DATA MINING.

In this section, I study some important lessons learnt by some author while studying DM in e-commerce.

A. DM Applied To Retail E-Commerce.

Kohavi et al (2004) have attempted a practical implementation of data mining in retail ecommerce data. They share their experience in terms of lessons that they learnt. There are the important issues in practical studies, into two categories: business-related and technology related. First I study then summarize their findings on the technical issues here.

Collecting data at the right level of abstraction is very important. Web server logs were originally meant for debugging the server software. Hence they convey very little useful information on customer-related transactions. Approaches including sessionising the web logs may yield better results. A preferred alternative would be have the application server itself log the user related activities. This is certainly going to be richer in semantics compared to the state-less web logs, and is easier to maintain compared to state-ful web logs.

Designing user interface forms needs to consider the DM issues in mind. For instance, disabling default values on various important attributes like Gender, Marital status, Employment status, etc., will result in richer data collected for demographical analysis. The users should be made to enter these values, since it was found by Kohaviet al (2004) that several users left the default values untouched.

Certain important implementation parameters in retail e-commerce sites like the automatic time outs of user sessions due to perceived inactivity at the user end, need to be based not purely on DM algorithms, but on the relative importance of the users to the organization. It should not turn out that large clients are made to lose their shopping carts due to the time outs that were fixed based on a DM of the application logs.

Auditing of data procured for mining, from data warehouses, is mandatory. This is due to the fact that the data warehouse might have collated data from several disparate systems with a high chance of data being duplicated or lost during the ETL operations.

Mining data at the right level of granularity is essential. Otherwise, the results from the DM exercise may not be correct.

V. CONCLUSION AND FUTURE WORK.

This paper presented how web mining (in a broad sense, DM applied to ecommerce) is applicable to improving the services provided by e-commerce based enterprises.Now I studying some ways in which web mining can be extended for further research. as part of the site design, creation, and content delivery.With the growing interest in the notion of semantic web, an increasing number of sites use structured semantics and domain ontologies. The primary challenge for the next-generation of personalization systems is to effectively integrate semantic knowledge from domain ontologies into the various parts of the process.

Some of the challenges in e-commerce DM include the following (Kohavi 2001).

Crawler/bot identification: Bots and crawlers can dramatically change clickstream patterns at a web site. For example, some websites like (www.keynote.com) provide site performance measurements. The Keynote bot can generate a request multiple times a minute, 24 hours a day, 7 days a week, skewing the statistics about the number of sessions, page hits, and exit pages (last page at each session). Search engines conduct breadth-first scans of the site, generating many requests in short duration. Tools need to have mechanisms to automatically sieve such noisy data in order for DM algorithms to yield sensible and pragmatic proposals.

Data transformations: There are two sets of transformations that need to take place: (i) data must be brought in from the operational system to build a data warehouse, and (ii) data may need to undergo transformations to answer a specific business question, a process that involves operations such as defining new columns, binning data, and aggregating it.

REFERENCES

[1] N R Srinivasa raghavan Department of Management studies, Indian Institute of Science, Bangalore 560 012, India.
[2] Agrawal R, Srikant R 1994 Fast algorithms for mining association rules. In 20th Int. Conf. on Very Large Databases (New York: Morgan Kaufmann) p 487–499
[3] Ansari S, Kohavi R, Mason L, Zheng Z 2001 Integrating e-commerce and data mining: architecture and challenges. In Proc. 2001 IEEE Int. Conf. on Data Mining (New York: IEEE Comput. Soc.)pp 27–34
[4] AugusteDM2001Customerserviceine-business.IEEE Internet Comput. 5(5): 90–91
[5] Berendt B, Hotho A, Stumme G 2002 Towards semantic web mining. In Proc. First Int. Semantic Web Conference, Sardinia, Italy Box G, Jenkins G, Reinsel G 1994 Time series analysis: Forecasting and control 3rd edn (Englewood Cliffs, NJ: Prentice Hall)
[6] CarbonePL2000Expandingthemeaningofandapplicationsfordatamining.In IEEE Int. Conf. on Systems, Man, and Cybernetics (New York: IEEE) pp 1872–1873
[7] Glymour C, Madigan D, Pregibon D, Smyth P 1996 Statistical inference and data mining. Commun. ACM 39(11):
[8] Gujarati D 2002 Basic econometrics (New York: McGraw-Hill/Irwin) [9] Haykin S 1998 Neural networks: A comprehensive foundation 2nd edn (Englewood Cliffs, NJ: Prentice-Hall)
[9] Hertz J, Krogh A, Palmer R G 1994 Introduction to the theory of neural computation (Reading, MA: Addison-Wesley).