Data Mining in E-Commerce

Prof. M. Pathak[1] Amithchand Shetty[2] Mahesh Poola[3] Pradeep Bhat[4] Dhiraj Mishra[5]

[1, 2,3,4,5] Padmabhushanvasantdadapatilpratishthan's College Of Engineering, Eastern Express Highway, Near Everard Nagar, Sion-Chunabhatti, Mumbai-400 022. Tel: 2407 0547 / 2403 8716

Abstract---The web today is different from what it used to be 10-20 years earlier. Everything nowadays is run on information. More than anything on the data matters the most. If you have accurate data you can do wonders with it. Over the years we have seen that E-Commerce applications have gained an important place in people's lives. One of the main reasons behind this can be stated as people like what they see on the web and they buy it, thus saving their time and energy. Now how do these E-Commerce websites know what the masses like and dislike is the main question, the answer to this question is Data Mining. Data mining play a major role in every E-commerce website nowadays. With the help of various data mining algorithms various patterns can be traced in the shopping behavior of an average buyer and hence e-commerce websites can have a clear indication what a buyer would like to buy and what to not.

General Terms- Data mining techniques, e-commerce applications and web mining.

Keywords: Electronic commerce, data mining, web mining.

I. Introduction

A. *What Is E-Commerce?*

E-Commerce is the ability of a company to have a dynamic presence on the Internet which allowed the company to conduct its business electronically, in essence having an electronic shop. Products can be advertised, sold and paid for all electronically without the need for it to be processed by a human being. The biggest advantage of E-Commerce is the ability to provide secure shopping transactions via the internet and coupled with almost instant verification and validation of credit card transactions.E-Commerce is not about the technology itself, it is about doing business using the technology.

B. *What is Data Mining?*

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and data base systems. The overall goal of the data mining process is to extract information from a data set and transform it into an Understandable structure for further use. Aside from the raw analysis step, it involves database and data preprocessing data management aspects, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization and online updating.The term is a buzzword, and is frequently misused to mean any form of large-scale data or information processing (collection extraction, warehousing, analysis,and Statistics) but is also generalized to any kind of computer decision support system, including artificial intelligence machine learning, and business intelligence. In the proper use of the word, the key term is *discovery*, commonly defined as "detecting something new". Even the popular book "Data mining: Practical machine learning tools and techniques with Java"(which covers mostly machine learning material) was originally to be named just "Practical machine learning", and the term "data mining" was only added for marketing reasons. Often the more general terms "(large scale) data analysis", or "analytics" – or when referring to actual methods, artificial intelligence and machine learning – are more appropriate.

II. Literature Survey

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 apposite to seek the services of data mining to make (business) sense out of these data sets.

Data mining (DM) has as its dominant goal, the generation of non-obvious yet useful information for decision makers from very large databases. The various mechanisms of this generation include abstractions, aggregations, summarizations, and characterizations of data. These forms, in turn, are the result of applying sophisticated modeling techniques from the diverse fields of statistics, artificial intelligence, database management and computer graphics.

The success of a DM exercise is driven to a very large extent by the following factors:-

Availability of data with rich descriptions: This means that unless the relations capturedin the database are of high degree, extracting hidden patterns and relationships among the various attributes will not make any practical sense.

Availability of a large volume of data: This is mostly mandated for statistical significance of the rules to hold. Absence of say, atleast a hundred thousand transactions will most likely reduce the usefulness of the rules generated from the transactional database.

Reliability of the data available: Although a given terabyte database may have hundreds of attributes per relation, the DM algorithms run on this dataset may be rendered defunct if the data itself was generated by manual and error prone means and wrong default values were set. Also, the lesser the integration with legacy applications, the better the accuracy of the dataset.

Ease of quantification of the return on investment (ROI) in DM: Although the earlier three factors may be favourable, unless a strong business case can be easily made, investments in the next level DM efforts may not be possible. In other words, the utility of the DM exercise needs to be quantified vis-a-vis the domain of application.

III. EXISTING SYSTEM

The e-commerce websites have come a long way since its inception. Gone are the days when these websites used to be static and only used to sell a single kind of product and that too home delivery as online payment was not considered to be secure back then. But the times have changed and these websites are currently the forefront runners in adapting any new technology that comes in the market which does them some good. Be it PHP or AJAX the e-commerce websites have always been keen to implement technologies which will reduce the front end load still keeping it dynamic and giving ample of execution time to the backend.

Currently all the existing e-commerce websites are up to date regarding load balancing on the front end to quick responses from the back end. The e-commerce web sites nowadays gives suggestion to customers as to which product to buy, which is the most selling product, etc. These websites also show what your recent buy at what time and which date was it bought on. A lot of websites have included the auto-completion facility on their search fields as it makes it easier for the customers. Hence even if the customer has a little idea about the product which he/she is looking for these websites can guide them perfectly.

Almost every e-commerce website today offers free home delivery service. The important thing to note here is that user don’t have to fill every time the form in which where he/she has to insert its address. You have to fill your address once and the rest will be taken care of. Many of the websites like Flipkart and SnapDeal guarantee a free home delivery within 48hrs.

The payment system is also secured in many ways. The user data is not lost and the payment gateways are absolutely secure nowadays. Visually also these websites offer the customer the best show possible. These websites are highly interactive and constantly demand users attention which keeps the user glued to their screen.

IV. DIFFERENT ALGORITHMS FOR E-COMMERCE APPLICATIONS

A. *Association*

Association (or relation) is probably the better known and most familiar and straightforward data mining technique. Here, you make a simple correlation between two or more items, often of the same type to identify patterns. For example, when tracking people's buying habits, you might identify that a customer always buys cream when they buy strawberries, and therefore suggest that the next time that they buy strawberries they might also want to buy cream.Building association or relation-based data mining tools can be achieved simply with different tools.

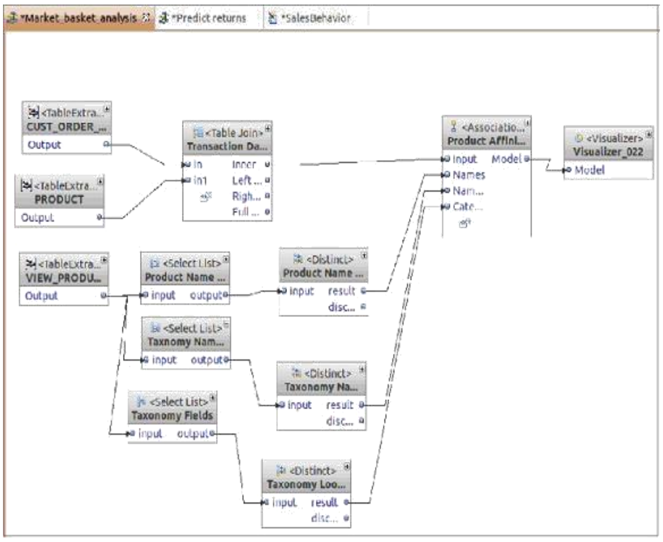
*Market_basket_analysis
*Predict returns
*SalesBehavior
<TableExtra... CUST_ORDER_...
Output
<TableExtra... PRODUCT
Output
<Table Join> Transaction Da...
In
In1
Inner
Left ...
Righ...
Full ...
<Associatio... Product Affini...
Input
Names
Nam...
Cate...
Model
<Visualizer> Visualizer_022
Model
<TableExtra... VIEW_PRODU...
Output
<Select List> Product Name ...
input
output
<Distinct> Product Name ...
input
result
disc...
<Select List> Taxonomy Nam...
input
output
<Distinct> Taxonomy Na...
input
result
disc...
<Select List> Taxonomy Fields
input
output
<Distinct> Taxonomy Loo...
input
result
disc...

Fig.1: Information flow that is used in association

B. *Classification*

You can use classification to build up an idea of the type of customer, item, or object by describing multiple attributes to identify a particular class. For example, you can easily classify cars into different types (sedan, 4x4, convertible) by identifying different attributes (number of seats, car shape, driven wheels). Given a new car, you might apply it into a particular class by comparing the attributes with our known definition. You can apply the same principles to customers, for example by classifying them by age and social group.

C. *Clustering*

By examining one or more attributes or classes, you can group individual pieces of data together to form a structure opinion. At a simple level, clustering is using one or more attributes as your basis for identifying a cluster of correlating results. Clustering is useful to identify different information because it correlates with other examples so you can see where the similarities and ranges agree.

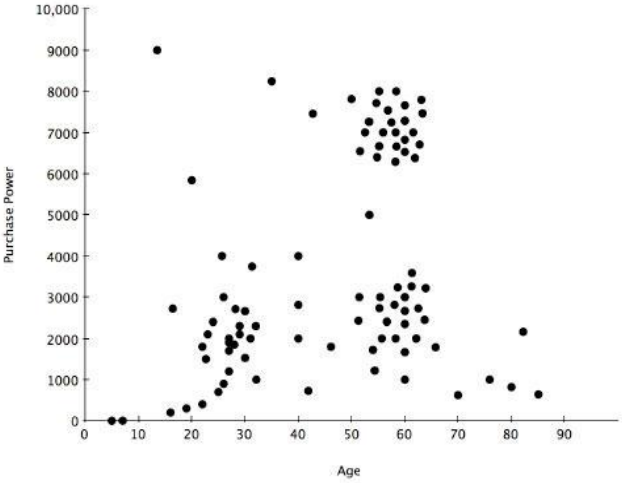

Fig. 2: Clustering

In the example, we can identify two clusters, one around the US$2,000/20-30 age group, and another at the US$7,000-8,000/50-65 age group. In this case, we've both hypothesized and proved our hypothesis with a simple graph that we can create using any suitable graphing software for a quick manual view. More complex determinations require a full analytical package, especially if you want to automatically base decisions on nearest neighbor *information.*

D. Prediction

Prediction is a wide topic and runs from predicting the failure of components or machinery, to identifying fraud and even the prediction of company profits. Used in combination with the other data mining techniques, prediction involves analyzing trends, classification, pattern matching, and relation. By analyzing past events or instances, you can make a prediction about an event.

E. Sequential patterns

Often used over longer-term data, sequential patterns are a useful method for identifying trends, or regular occurrences of similar events. For example, with customer data you can identify that customers buy a particular collection of products together at different times of the year. In a shopping basket application, you can use this information to automatically suggest that certain items be added to a basket based on their frequency and past purchasing history.

F. Decision trees

Related to most of the other techniques (primarily classification and prediction), the decision tree can be used either as a part of the selection criteria, or to support the use and selection of specific data within the overall structure. Within the decision tree, you start with a simple question that has two (or sometimes more) answers. Each answer leads to a further question to help classify or identify the data so that it can be categorized, or so that a prediction can be made based on each answer.

shows an example where you can classify an incoming error condition.

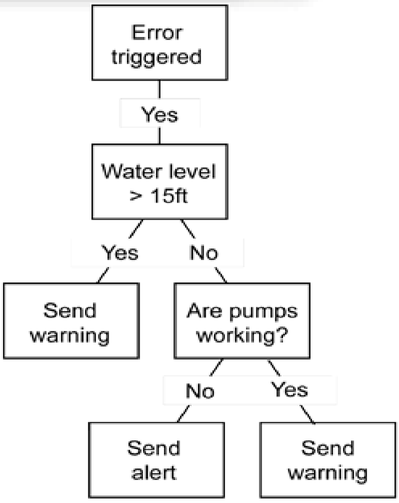

Fig.3: Decision tree

V. Advantages Of E-Commerce

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, from work, or from a hotel dial-up? The information available in log files is often used (Auguste 2001) to determine what profiling can be dynamically processed in the background and indexed into the dynamic generation of HTML, and what performance can be expected from the servers and network to support customer service and make e-business interaction productive.

B. DM in recommendation systems

Systems have also been developed to keep the customers automatically informed of important events of interest to them. The article by Jeng&Drissi (2000) discusses an intelligent framework called PENS that has the ability to not only notify customers of events, but also to predict events and event classes that are likely to be triggered by customers. The event notification system in PENS has the following components: Event manager, event channel manager, registries, and proxy manager. The event-prediction system is based on association rule-mining and clustering algorithms. The PENS system is used to actively help an e-commerce service provider to forecast the demand of product categories better. Data mining. has also been applied in detecting how customers may respond to promotional offers made by a credit card e-commerce company (Zhang *et al* 2003). Techniques including fuzzy computing and interval computing are used to generate if-then-else rules.

C. DM in web personalization

Mobasher (2004) presents a comprehensive overview of the personalization process based on web usage mining. In this context, the author discusses a host of web usage mining activities required for this process, including the preprocessing and integration of data from multiple sources, and common pattern discovery techniques that are applied to the integrated usage data. The goal of this paper is to show how pattern discovery techniques such as clustering, association rule-mining, and sequential pattern discovery, performed on web usage data, can be leveraged effectively as an integrated part of a web personalization system. The author observes that the log data collected automatically by theWeb and application servers represent the fine-grained navigational behaviour of visitors.

D. Data to be captured by weblogs

Depending on the goals of the analysis, e-commerce data need to be transformed and aggregated at different levels of abstraction. e-Commerce data are also further classified as usage data, content data, structure data, and user data. Usage data contain details of user sessions and pageviews. The content data in a site are the collection of objects and relationships that are conveyed to the user. For the most part, the data comprise combinations of textual material and images. The data sources used to deliver or generate data include static HTML/XML pages, images, video clips, sound files, dynamically generated page segments from scripts or other applications, and collections of records from the operational database(s). Site content data also include semantic or structural metadata embedded within the site or individual pages, such as descriptive keywords, document attributes, semantic tags, or HTTP variables. Structure data represent the designer's view of

the content organization within the site. This organization is captured via the inter-page linkage structure among pages, as reflected through hyperlinks. Structure data also include the intra-page structure of the content represented in the arrangement of HTML or XML tags within a page. Structure data for a site are normally captured by an automatically generated *site map* which represents the hyperlink structure of the site.

E. *DM and multimedia e-commerce*

Applications in virtual multimedia catalogs are highly interactive, as in e-malls selling multimedia content based products. It is difficult in such situations to estimate resource demands required for presentation of catalog contents. Hollfelder*et al* (2000) propose a method to predict presentation resource demands in interactive multimedia catalogs. The prediction is based on the results of mining the virtual mall action log file that contains information about previous user interests and browsing and buying behaviour.

F. *DM and buyer behaviour 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 behaviour 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 authors present a theory, in which success is modelled on the basis of the navigation behaviour of the site's users. They then exploit web usage miner (WUM), a navigation pattern discovery miner, to study how the success of a site is reflected in the users' behaviour. With WUM the authors measure the success of a site's components and obtain concrete indications of how the site should be improved.

G. *Analysing web transactions*

Once the data are collected via any of the above mentioned 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).

VI. Cases In E-Commerce Regarding To Data Mining

In various e-commerce domains involving spatial data (real estate, environmental planning, precision agriculture), participating businesses may increase their economic returns using knowledge extracted from spatial databases. However, in practice, spatial data is often inherently distributed at multiple sites. Due to security, competition and a lack of appropriate knowledge discovery algorithms, spatial information from such physically dispersed sites is often not properly exploited. Lazarevic*et al* (1999) develop a distributed spatial knowledge discovery system for precision agriculture. In the proposed system, a centralized server collects proprietary site-specific spatial data from subscribed businesses as well as relevant data from public and commercial sources and integrates knowledge in order to provide valuable management information to subscribed customers. Spatial data mining software (Koperski *et al* 1996) interfaces this database to extract interesting and novel knowledge from data. Specific objectives include a better understanding of spatial data, discovering relationships between spatial and nonspatial data, construction of spatial knowledge-bases, query optimization and data reorganization in spatial databases. Knowledge extracted from spatial data can consist of characteristic and discriminant rules, prominent structures or clusters, spatial associations and other forms. Challenges involved in spatial data mining include multiple layers of data, missing attributes and high noise due to a low sensibility of instruments and to spatial interpolation on sparsely collected attributes. To address some of these problems, data are cleaned by removing duplicates, removing outliers and by filtering through a median filter with a specified window size (Lazarevic*et al* 1999). The goal of precision agriculture management is to estimate and perform site-specific crop treatment in order to maximize profit and minimize environmental damage. Through a knowledge discovery (KDD) process, Lazarevic*et al* (1999) propose learning algorithms that perform data modelling using data sets from different fields in possibly different regions and years. Each dataset may contain attributes whose values are not manageable (e.g. topographic data), as well as those attributes that are manageable. In order to improve prediction ability when dealing with heterogeneous spatial data, an approach employed in the proposed system by Lazarevic*et al* (1999) is based on identifying spatial regions having similar characteristics using a clustering algorithm. Clustering algorithm is used for partitioning multivariate data into meaningful subgroups (clusters), so that patterns within a cluster are more similar to each other than are patterns belonging to different clusters. Local regression models are built on each of these spatial regions describing the relationship between the spatial data characteristics and the target attribute.

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. They classify the important issues in practical studies, into two categories: business-related and technologyrelated.

We now summarize their findings on the technical issues here.

1) 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 having 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.
2) Designing user interface forms needs to consider theDMissues 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 Kohavi*et al* (2004) that several users left the default values untouched.
3) 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.
4) Generating logs for several million transactions is a costly exercise. It may be wise to generate appropriate logs by conducting random sampling, as is done in statistical quality control. But such a sampling may not capture rare events, and in some cases like in advertisement referral based compensations, the data capture may be mandatory. Techniques thus need to be in place that can do this sampling in an *intelligent* fashion.
5) 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.
6) Mining data at the right level of granularity is essential. Otherwise, the results from the DM exercise may not be correct.

VII. CONCLUSION

E-commerce has taken the world by storm and one of the main reasons behind it is Data Mining. If it wouldn’t have been for data mining then ecommerce industry wouldn’t have reached the heights in which it is currently present. Data mining allows the ecommerce websites to create a target audience and cater to their needs successfully. Data mining in ecommerce also tries to attract new customer to the website along with maintaining the current ones by keeping the shopping logs updated and highlighting those products which are red hot in the market.

REFERENCES

[1] http://www.ibm.com/developerworks/opensource/library/ba-data-mining-techniques/index.html
[2] http://en.wikipedia.org/wiki/Data_mining#Pre-processing
[3] http://en.wikipedia.org/wiki/Ecommerce
[4] A Study of Ethical and Social Issues in E-Commerce by HimaniGrewal and Shivani
[5] Data mining in e-commerce: Asurvey by N R Srinivasa Raghavan
[6] 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
[7] Kohavi R 2001 Mining e-commerce data: The good, the bad, and the ugly. In Proceedings of the Seventh ACMSIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2001) (New York: ACM Press) pp 8–13
[8] Kohavi R, Mason L, Parekh R, Zheng Z 2004 Lessons and challenges from mining retail e-commerce data. Machine Learning J. (Special Issue on Data Mining Lessons Learned)applications for data mining. In IEEE Int. Conf. on Systems, Man, and Cybernetics (New York: IEEE) pp 1872–1873
[9] Glymour C, Madigan D, Pregibon D, Smyth P 1996 Statistical inference and data mining. Commun. ACM 39(11): Gujarati D 2002 Basic econometrics (New York: McGraw-Hill/Irwin)