DATA MINING TECHNIQUES IN E-COMMERCE APPLICATIONS

Pabitra Kumar Tripathy*
Assistant Professor,
Department of Computer Science & Engineering
Kalam Institute of Technology
Berhampur – 761013 (Odisha).

Dr. Subhendu Kumar Rath**
Deputy Director, Examination
Biju Patnaik University of Technology (BPUT)
Chhend Colony, Rourkela, Odisha-769004

Data mining can be defined as the art of extracting non-obvious, useful information from large databases. This emerging field brings a set of powerful techniques which have relevance for companies to focus their efforts in taking advantage of their data.

In the *Business Data Definition* component the e-commerce business user defines the data and metadata associated with their business. This data includes merchandising information like products, assortments, and price lists etc., content information like web page templates, articles, images, multimedia etc. and business rules like personalized content rules, promotion rules, and rules for cross-sells and up-sells. From a data mining perspective the key to the *Business Data Definition* component is the ability to define a rich set of attributes (metadata) for any type of data. For example, products can have attributes like size, color, and targeted age group, and can be arranged in a hierarchy representing categories like men's and women's, and subcategories like sarees and shirts. Having a diverse set of available attributes is not only essential for data mining, but also for personalizing the customer experience.

Data mining tools generate new information for decision makers from very large databases. The various mechanisms of this generation include abstractions, aggregations, summarizations, and characterizations of data[i].

Having a huge amount of data, make some problems for detection of hidden relationships among various attributes of data and between several snapshots of data over a period of time. These hidden patterns have enormous potential in predictions and personalization in e-commerce.

History of E-Commerce

Until about 1994, electronic commerce was not web-based. The term referred to the use of computers and telecommunications to automatically forward and process commercial documents, such as invoices and inventory requests. Ross Perot, was the first one who founded the Electronic Data Systems Company (EDS) in 1962 with a goal to streamline municipal parking ticket billing. The core technology was an industry standard for electronic data interchange (EDI), which allowed communication between computers, originally through shipment of magnetic tapes. Till the 1990s, early versions of electronic commerce focused on business-to-business (B2B) transactions, because personal computers were relatively rare and the EDI systems were expensive.

Most e-commerce start-ups had simple business models rooted in traditional bricks-and-mortar perspectives. A typical paradigm was to have customers place orders over the Internet, or to sell advertisement space (pop-ups and banners) during web browsing sessions. However, emergent complexity soon began to dominate the evolution, and business biodiversity blossomed.

Computer security software became a necessary enabler, as well as spam filters and technology such as PayPal to facilitate small-scale commercial purchases. Gray commerce became pervasive, notably in pornography and services such as Napster that support music file-sharing. The norms and laws of society are struggling to adapt to new circumstances; this is complicated by the fact that businesses can easily be located outside of national jurisdiction where different rules apply. Much of the growth and diversification cited above depends on data mining methods of varying degrees of sophistication, but e-commerce also allows businesses to access new data streams that inform management in ways that were not previously possible.

A wide verity of multimedia data such as images, videos, signals, and text that are available in electronic form with temporal and spatial characteristics shows variety of data types and structures. These are basic resources in the present generation data applications. Many people into business and research uses internet as the basic infrastructure in data mining to revolutionize business and scientific landscape. The web has declared itself as a powerful global connecting force. Internet established world merge to one community and technology created intranets and extranets creating communities within companies[ii]. In these circumstances it is not surprising to note that more and more companies depend

on data mining techniques for their e-commerce. Some of the important data mining techniques may be summarized as follows.

Data mining Techniques

Clustering

Sometimes elements can be categorized even when the set of categories **n** are not available. This problem is known as data clustering. Compared to data classification it is a challenging task. Here, a mathematical model receives the data without the class labels and infers groups of elements just by merely examining their similarities. The output is an estimated class membership. In contrast to the classification problem where there is a set of possible classes known *a priori,* in the clustering problem different groups are created. The objective is to group similar instances in the same group, while at the same time, assign to distinct groups those elements which are different. This type of learning is sometimes referred to as unsupervised learning because the function lacks of a teacher that tells the correct class label of a particular pattern. Formally, data clustering, consist in assigning a class label $\mathbf{l_f}$ to each pattern **Xj** in the set $X_v = \{x_{1(}x_2, ...,x_{„}\}$, identifying its respective label. The set of all labels for a pattern set Xu is L = **{li,l$_2$' - .In},** where \ e {1,2..,k}, and where k is the number of clusters.

Interestingly enough, the human brain is particularly good at this task. For generations we have used this type of reasoning to distinguish between ripe fruit before collecting it or to build an entire taxonomy of the animal kingdom based on the observed characteristics: Nobody told us to categorize animals according to whether they produce milk or not! Applications of data clustering in e-commerce include recommendation systems[iii], search engines[iv], etc.

Semi-supervised Classification

Classification is an example of supervised learning, assuming the knowledge of well-defined training sets with a clear specification of the identity of *all* the training samples. A distinct and intriguing learning paradigm that has emerged in the recent years is semi-supervised learning. This paradigm combines labeled and unlabeled instances simultaneously to perform classification [v]. This specific type of classifiers does not demand the specification of the class labels of every sample. Usually this type of learning appears in situations where many instances are available, but only few of them possess labels because the cost of acquiring them is high. One common way to learn in this context is to perform a

clustering-like mechanism, assigning the training samples into different groups, and subsequently, a class label is assigned to each group using a small subset of the training instances whose class identities are known. Given a clustering algorithm, **c#$_c$,** a set of labeled instances, X_L, a set of unlabeled instances, Xy, and a supervised learning algorithm, $<A_S$, the Cluster-then-Label method works as follows [4]: First, we identify the clusters of the input manifold using the clustering algorithm $<A_C$. Secondly, we determine which of the labeled samples fall in each cluster. For each cluster we determine a decision boundary based on the supervised algorithm **c#$_s$,** and the labeled samples assigned to that cluster, which, in turn, allows the prediction of the label of every cluster. Finally, each uncategorized item is labeled according to the predicted class of the cluster in which it is contained. Recently, semi-supervised classification has been successfully applied in the estimation of the quality of online reviews [vi].

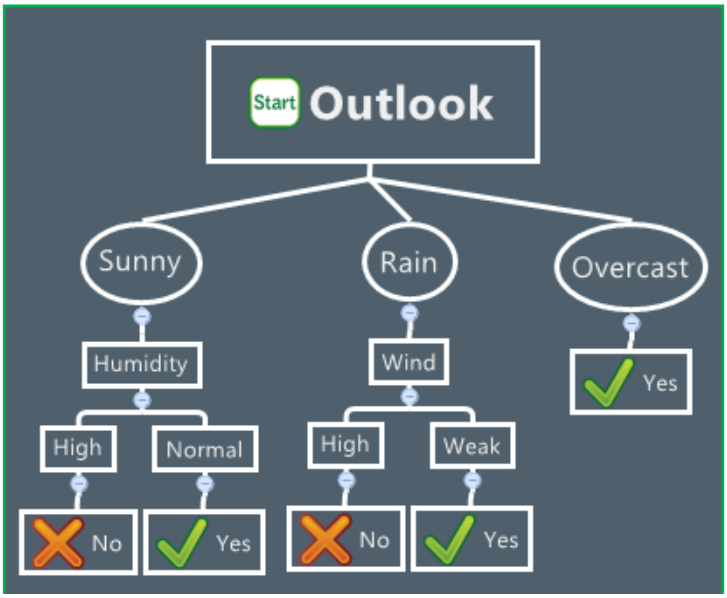
Association Analysis

In this context, the data is conformed of transactions, e.g., the bill that includes a list of products that we bought in a grocery store. The nature of the data is unique: items do not necessarily repeat in two bills, but usually people tend to behave similarly in their buying trends. Association analysis attempts to discover those trends. In this context, one may be interested in knowing which patterns are more frequent. A famous example is the relationship between diapers and beer in grocery store bills. Information like this provides useful information at the time a grocery store is designed: if you know that people will buy beer and diapers you can put them together, or place them in opposite corners, increasing the probability that the customers will see other products they might be interested in Association rule mining can be formally defined as follows [2]: Let **I =** $\{a_{lt}a_2$**,...,a$_n$}** be a collection of **n** elements called items. Also, let **D =** ***{T-L,T$_2$,*** **-,T$_m$}** be a collection of transactions called the database. Each transaction **T e D** contains a subset of the items in **I.** Additionally, an itemset is a set of items. Given an itemset **X ^ I** and a given transaction **T,** it is said that **T** contains **X** if and only if XET. The support count of a given itemset **X,** denoted by cr_x, is defined as the number of transactions in **D** that contain **X.** Let s be the support threshold and **|D|** be the total number of transactions in **D.** An itemset is said to be frequent if $o_x >$ **D** xs% . An association rule corresponds to an implication **X=»Y** where **X . Y E I ,** and **X n Y = 0 .** One of the main goals of association analysis is to discover association rules or sets of frequent items.

Applications of association analysis include customer relationship management (CRM)[vii], building recommendation systems [viii], personalization applications and collaborative filtering [ix].

Decision trees

The decision tree is one of the most common used data mining techniques because its model is easy to understand for users. In decision tree technique, the root of the decision tree is a simple question or condition that has multiple answers. Each answer then leads to a set of questions or conditions that help us determine the data so that we can make the final decision based on it. For example, we use the following decision tree to determine whether or not to play tennis:



Starting at the root node, if the outlook is overcast then we should definitely play tennis. If it is rainy, we should only play tennis if the wind is the weak. And if it is sunny then we should play tennis in case the humidity is normal.

We often combine two or more of the data mining techniques together to form an appropriate process that meets the business needs.

Prediction

The prediction is one of a data mining techniques that discover the relationship between independent variables and relationship between dependent and independent variables. For instance, the prediction analysis technique can be used in the sale to predict profit for the future if we

consider the sale is an independent variable, profit could be a dependent variable. Then based on the historical sale and profit data, we can draw a fitted regression curve that is used for profit prediction.

Sequential Patterns

Sequential patterns analysis is one of data mining technique that seeks to discover or identify similar patterns, regular events or trends in transaction data over a business period. In sales, with historical transaction data, businesses can identify a set of items that customers buy together different times in a year. Then businesses can use this information to recommend customers buy it with better deals based on their purchasing frequency in the past.

APPLICATIONS DATA MINING IN E-COMMERCE

Customer Profiling

Many companies 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 data mining tools that take care of knowledge representation of customer profiles and predictive modeling 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.

Recommendation Systems

The article by Jeng & Drissi [x] 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 activated 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[xi].

Web Personalization

A comprehensive overview of the personalization process based on web usage mining [xii] is presented by Mobasher. Here, 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 the Web and application servers represent the fine-grained navigational behavior of visitors.

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 page views. 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. The operational database(s) for the site may include additional user profile information. Such data may include demographic or other identifying information on registered users, user ratings on various objects such as pages, products, or movies, past purchase or visit histories of users, as well as other explicit or implicit representations of users' interests.

Customer 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 behavior. Vallamkondu & Gruenwald describe an approach to predict user behavior in e-commerce sites[xiii]. The core of their approach involves extracting knowledge from integrated data of purchase and path traversal patterns of past users (obtainable from web server logs) to predict the purchase and traversal behavior 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 propose a methodology to improve the success of web sites, based on the exploitation of navigation-pattern discovery[xiv]. In particular, the authors present a theory, in which success is modeled on the basis of the navigation behavior 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' behavior. With WUM the authors measure the success of a site's components and obtain concrete indications of how the site should be improved.

In the context of web mining, clustering could be used to cluster similar click-streams to determine learning behaviors in the case of e-learning or general site access behaviors 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. Wang & Zaiane propose an algorithm based on sequence alignment to measure similarities between web sessions where sessions are chronologically ordered sequences of page accesses[xv].

Conclusion

Data mining tools aid the discovery of patterns in data. Until recently, companies that have concentrated on building horizontal data mining modeling tools, have had little commercial success. Many companies were bought, including the acquisition of Compression Sciences by Gentia for $3 million, Hyper Parallel by Yahoo for about $2.3 million, Clementine by SPSS for $7 million, and Thinking

Machines's Darwin by Oracle for less than $25 million. Recently, a phase shift has occurred in the valuation of such companies, and recent acquisitions have given rise to valuations 10 to 100 times higher. KD1 was acquired by Net Perceptions for $116M, RightPoint (previously DataMind) was acquired by E.piphany for $400M, DataSage was acquired by Vignette for $577M, and NeoVista was acquired by Accrue for $140M. The shift in valuations indicates wider recognition of the value of data mining modeling techniques for e-commerce.

Many business practices can benefit from mining e-commerce information, even if they are not directly using it to promote new services or better handle their customers. Information collected by e-commerce transactions can inform businesses that practice no e-commerce at all. Companies that do international business need constant guidance on ever changing regulations, tariffs, price differentiation across geography, weather conditions, leading fees and spot prices for commodities, etc. This field is complex and highly localized; one needs smart people on the ground in every country with which business is done. It is expensive to build this kind of expertise in-house, and probably impossible for it to be truly expert, since complex optimization problems must be solved. As an e-business opportunity, a start-up could hire locals to input this kind of information from all over the world into a common data base.

Suppose customers had access to the kind of high-end data mining tools that e-commerce uses. People could use intelligent personalized shopping bots to churn endlessly, always demanding the best value. This would drive down profit margins by forcing tight competition. This will enable businesses to become presbyopic: they can plan better, control better and adapt faster. There is reason to hope that the well-documented inefficiencies of the market will be reduced, which economists suggest will make the world better, of course, in the long run.

References:

[i] P.L. Carbone, "Expanding the meaning of and applications for data mining," *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics,* vol. 3, 2000, pp. 1872-1873.

[ii] Mehmed M.Kantardzic, Jozef Zurada, Next Generation of Data Mining Applications, Wiley interscience, IEEE,ISBN 0-471-65605-4

[iii] J. Wu, Q. Liu and S. Luo, Clustering technology application in e-commerce recommendation system, in Proceedings ICMECG International Conference Management of e-Commerce and e-Government, Jiangxi, 2008, pp. 200-203.

[iv] R. Mikut and M. Reischl, Data mining tools, WIREs: Data Min Knowl Discov, Vol. 1, no. 5, pp. 431-443, 2011.

[v] X. Zheng, S. Zhu and Z. Lin, Capturing the essence of word-of-mouth for social commerce: Assessing the quality of online e-commerce reviews by a semi-supervised approach, Decision Support Systems, Vol. 56, 2013, pp. 211-222.

[vi] Ibid.

[vii] R. Vedala and B. Kumar, An application of naive bayes classification for credit scoring in e-lending platform, in Proceedings International Conference on Data Science Engineering (ICDSE), Kochi, 2012, pp. 81-84

[viii] X.Z. Zhang, Building personalized recommendation system in e-commerce using association rule-based mining and classification, in Proceedings International Conference on Machine Learning and Cybernetics, Hong Kong, 2007, pp. 4113-4118.

[ix] R. Natarajan and B. Shekar, Interestingness of association rules in data mining: Issues relevant to e-commerce, Sadhana, vol. 30, no. 2-3, 2005, pp. 291-309.

[x] J.J. Jeng and Y. Drissi, "PENS: A predictive event notification system for e-Commerce environment," *Proceedings - IEEE Computer Society's International Computer Software and Applications Conference,* 2000, pp. 93-98,

[xi] X. Z. Zhang, "Building personalized recommendation system in E-Commerce using association rule-based mining and classification," in *Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, ICMLC 2007,* 2007, pp. 4113-4118.

[xii] B. Mobasher, "Web usage mining and personalization," *Practical Handbook of Internet Computing,* 2004.

[xiii] S. Vallamkondu and L. Gruenwald, "Integrating purchase patterns and traversal patterns to predict http requests in e-commerce sites," *IEEE Int. Conf. on e- commerce,* 2003, pp. 256-263.

[xiv] M. Spiliopoulou and C. Pohle, "Data mining to measure and improve the success of web sites," *J. Data Mining and Knowledge Discovery,* 2000.

[xv] W. Zhu, J. Chen, and J. Yin, "Application of data mining in E-business," *Jisuanji Gongcheng/Computer Engineering,* Vol. 28, 2002, p. 73.