A Data Mining Approach for Business

Hamid Rastegari, Mohd Noor Md. Sap
Faculty of Computer Science and Information Systems
University Technology Malaysia
hamid_rastegari@yahoo.com, mohdnoor@utm.my

Abstract:

Modern business is rushing toward electronic commerce. Electronic commerce has been put forward as a new method of implementing trade activities. If the transition is done properly, it enables better management, new services, lower transaction costs and better customer relations. However, successful business depends on speed and skill of knowledge discovery. Data mining has been considered to be a tool of business for knowledge discovery. This paper discusses the important role of business based on data mining knowledge development to detection the relation of data mining and electronic commerce.

Keywords: business, e-commerce, data mining.

1. Introduction

Electronic commerce has changed the face of business. It allows better customer management, new strategies for marketing, an expanded range of products and more efficient operations. A key enabler of this change is using of increasingly sophisticated data mining tools.

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 (Carbone 2000). 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.

Although much work has been done to date, more studies need to be conducted to as various subjects in a variety of e-commerce problems. The purpose of this paper is a present of data mining methods and expression application of data mining in business. It is a briefing of works that have been done in this area. This study can be useful for future work.

2. Data Mining Methods

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 personalizations in e-commerce. Data mining has been pursued as a research topic by at least three communities: the statisticians, the artificial intelligence researchers, and the database engineers (Raghavan, 2005). In this section present a brief overview of some of the features of each of these approaches.

2.1 Role of statistics in data mining

Extracting causal information from data is often one of the principal goals of data mining and more generally of statistical inference. Statisticians have done aggregate data analyses on data for decades; thus data mining has actually existed from the time large scale statistical modeling has been made possible (Carbone 2000).

Statisticians consider the causal relationship between the dependent variables and independent variables as proposed by the user (usually the domain expert), and attempt to capture the degree and nature of dependence between the variables. Modeling methods include simple linear regression, multiple regressions, and nonlinear regression. Such models are often parameter driven and are arrived at after solving attendant optimization models.

The regression methods may be considered analogous to the association rules in data mining. In the latter case, rule-mining algorithms propose the correlation of item sets in a database, across various attributes of the transactions.

Time series modeling on the other hand, is more relevant in sequential mining. This is used to unearth correlations and patterns in temporally ordered data. For a more detailed overview of time series methods, the reader may refer to Box et al (1994).

2.2 Role of AI in data mining

Artificial intelligence, on the other hand, has provided a number of useful methods for data mining. Decision tree is most useful in classification

problems. With this technique, map data into several classes and a tree is constructed to model the classification process. Machine learning is a set of methods that enable a computer to learn relations from the given data sets. With minimal or no hypothesis from the user, learning algorithms do come up with meaningful relations and also explain them well. Some of the most popular learning systems include the neural networks and support vector machines. We briefly present the relevant issues below.

Neural networks are predominantly used to learn linear and nonlinear relationships between variables of interest. The architecture, in general, consists of a perceptron with input and output nodes with weighted edges connecting the two nodes. A neural network with two layers is thus a bi-partite acyclic graph. The perceptron, which is the learning machine, is 'trained' in order to arrive at an optimal 'weight vector'. The output is then expressed as a (weighted) linear combination of the inputs. Learning consists of solving an underlying optimization model which is solved using gradient descent based methods.

Neural networks are also useful in clustering data sets. The most popular method available to cluster data sets is the K-means algorithm. Given an M-dimensional data set, the idea is to try and locate the minimal number of centroids around which the data set clusters itself. Thus the onus is to define an appropriate distance vector that helps partition the data sets into as minimally overlapping sub-sets as possible. In general, Euclidean distance metrics are proposed for 'optimally' partitioning a given data set. The optimization is again based on minimizing the sum of squares of an appropriate error term, using classical gradient based methods.

The advantages of neural networks over the conventional statistical analysis methods are as follows (Park 2000). First, neural networks are good at modeling nonlinear relationships and interaction while conventional statistical analysis in most cases assumes linear relationship between independent variables and dependent variables. Neural networks build their own models with the help of learning process whether the relationships among variables are linear or not. Secondly, neural networks perform well with missing or incomplete data. A single missing value in regression analysis leads to removal of the entire observation or removal of the associated variable from all observations in the data set being analyzed. However, neural networks update weights between input, output, and intermediate nodes, so that even incomplete data can contribute to learning and produce desired output results. Finally, neural networks do not require scale adjustment or statistical assumptions, such as normality or independent error terms.

Artificial intelligence based methods using neural networks are used in clustering and classification methods of data mining. They can also be used in sequential mining (Park 2000). For instance, market basket analysis which concerns itself with identifying hidden customer segments could be solved using neural networks with unsupervised learning. An online web store may want to provide different grades of service to its users depending on the frequency of customers' visits to their websites. Identifying the basket of such customer segments could be done using clustering methods. Finally, if the web store wants to identify the factors contributing to repeat customers, they could use the nonlinear regression expressions obtained using neural networks.

2.3 Role of database in data mining

Keeping in mind that data mining approaches rely heavily on the availability of high quality data sets, the database community has invented an array of relevant methods and mechanisms that need to be used prior to any data mining exercise. Extract, transform and load (ETL) applications are worthy of mention in this context. Given an enterprise system like an enterprise resource planning system (ERP), it is likely that the number of transactions that happen by the minute could run into hundreds, if not thousands. Data mining can certainly not be run on the transaction databases in their native state. It requires to be extracted at periodic intervals, transformed into a form usable for analysis, and loaded on to the servers and applications that work on the transformed data. Today, software systems exist in the form of data warehousing solutions that are often bundled with the ERP system, to perform this complex and important task.

It is to be observed that data warehouses are essentially snapshots of transactional data aggregated along various dimensions (including time, geographies, demographies, products etc.) In order to run data mining algorithms, it is common practice to use the data available in the data warehouse rather than by running real time scripts to fetch transactional data. This is for the simple reason that for practical purposes, it is sufficient to include snapshots of data taken at say, weekly or monthly basis, for analysis. Real-time data is not relevant for tactical decision making, which is where data mining is used. Data warehousing is nevertheless fraught with technological challenges.

When one has terabytes of data available, the goal of database engineers in data mining is to create

structures and mechanisms to efficiently read in data into memory and run algorithms like A priori (Agrawal & Srikant 1994). Such algorithms assume the so-called item sets. Consider a database where the transactions pertain to a retail store. Customers buy various products and each transaction records the products bought by the customer. Observe that such databases can grow enormously in size, especially for large retailers who have web storefronts, like amazon.com. Each item set is a record in the database, with attributes mentioning if a particular product was purchased or not. The algorithms compute, given a certain support and confidence, the rules that apply on the given item sets.

It is to be noted that A priori-like algorithms work on a given item set only. If the underlying transactional database is dynamic, then there are methods known as incremental mining proposed by the database community. Such methods resort to minimizing the number of passes over a given database for computing the support and confidence values in rule-mining (Woon et al 2002).

3. E-Commerce and Data Mining

In this section, we survey articles that are very specific to data mining implementations in e-commerce. The salient applications of data mining techniques are presented first. Later in this section, architecture and data collection issues are discussed.

3.1 Data mining 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 behavior will improve the availability of products and services and hence the profits. Thus the end goal of any data mining 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 detect what prospective customers are seeking from a site.

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 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.

3.2 Data mining 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 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 (Zhang et al 2007). Techniques including fuzzy computing and interval computing are used to generate if-then-else rules.

Niu et al (2002) present a method to build customer profiles in e-commerce settings, based on product hierarchy for more effective personalization. They divide each customer profile into three parts: basic profile learned from customer demographic data; preference profile learned from behavioral data, and rule profile mainly referring to association rules. Based on customer profiles, the authors generate two kinds of recommendations, which are interest recommendation and association recommendation. They also propose a special data structure called profile tree for effective searching and matching.

3.3 Data mining 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 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.

3.4 Data mining and 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 (2003) describe an approach to predict user behavior 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 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 (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 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 (2002) propose an algorithm based on sequence alignment to measure similarities between web sessions where sessions are chronologically ordered sequences of page accesses.

4. Data Mining and Business Intelligence

Data mining is about finding useful patterns in data. This word useful can be unpacked to expose many of the key properties of successful data mining. The patterns discovered by data mining are useful because they extend existing business knowledge in useful ways. But new business knowledge is not created "in a vacuum"; it builds on existing business knowledge, and this existing knowledge is in the mind of the business expert. The business expert therefore plays a critical role in data mining, both as an essential source of input (business knowledge) and as the consumer of the results of data mining. The business expert not only uses the results of data mining but also evaluates them, and this evaluation should be a continual source of guidance for the data

mining process. Data mining can reveal patterns in data, but only the business expert can judge their usefulness. It is important to remember that the data is not the business, but only a dim reflection of it. This gap, between the data and the business reality it represents, is called the chasm of representation to emphasize the effort needed to cross it.

Patterns found in the data may fail to be useful for many different reasons. They may reflect properties of the data, which do not represent reality at all, for example when an artifact of data collection, such as the time a snapshot is taken, distorts its reflection of the business. Alternatively, the patterns found may be true reflections of the business, but they merely describe the problem that data mining was intended to solve – for example arriving at the conclusion that "purchasers of this product have high incomes" in a project to market the product to a broader range of income groups. Finally, patterns may be a true and pertinent reflection of the business, but nevertheless merely repeat "truisms" about the business, already well known to those within it. It is all too easy for data mining, which is insufficiently informed by business knowledge to produce useless results for reasons like the above. To prevent this, the business expert must be at the very heart of the data mining process, spotting "false starts" before they consume significant effort. The expert must either literally "sit with" the data miner, or actually perform the data mining. In either case, the close involvement of the business expert has far-reaching consequences for the field of data mining.

5. Data Collection

It may be observed that there are various ways of procuring data relevant to e-commerce data mining. Web server log files, web server plug-ins(instrument-ation), TCP/IP packet sniffing, application server instrumentation are the primary means of collecting data. Other sources include transactions that the user performs, marketing programs (banner adver-tisements, emails etc), demographic (obtainable from site registrations and subscriptions), call centres and ERP systems. It is quite common to expend about 80% of any data mining 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 behavior-related information along with the product details. Ansari et al (2001) describe an architecture for supporting the integration of data mining and e-commerce. The architecture is found to dramatically reduce the preprocessing, cleaning, and data understanding effort. They 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. They also describe the data transformation bridges required from the transaction processing systems and customer event streams (e.g. click streams) to the data warehouse.

5.1 Analyzing 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).

The view of web transactions as sequences of page views allows one to employ a number of useful and well-studied models which can be used to discover or analyze user navigation patterns. One such approach (Sarukkai 2000) is to model the navigational activity in the website as a Markov chain. In the context of web transactions, Markov chains can be used to model transition probabilities between page views. In web-usage analysis, they have been proposed as the underlying modeling machinery for web prefetching applications or to minimize system latencies.

Hu&Cercone (2002) present a new approach called on-line analytical mining for web data. Their approach consists of data capture, web house construction, and 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 web house is akin to the data warehouse.

5.2 An architecture for data mining

In a B2B e-commerce setting, it is very likely that vendors, customers and application service providers (ASP) (usually the middlemen) have varying data mining requirements. Vendors would be interested in data mining 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 data mining architecture that enables a data mining 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). The costing infrastructure refers to the system having a framework for estimating the costs of different tasks. This implies that 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 optimize 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). The authors have indeed designed and implemented such a framework.

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 authors 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 on the site itself. The system is extensible to provide for a wide range of mining algorithms (Krishnaswamy et al 2000). The authors provide a facility wherein the user can register their algorithms with the ASP for use in their specific distributed data mining jobs.

6. Cases in e-commerce data mining

In this section, we first present an interesting application of data mining in e-commerce. We then present some important lessons learnt by some authors while implementing data mining in e-commerce.

6.1 Distributed spatial 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 discriminate 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 modeling 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 (e.g. nutrient concentrations).

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. A 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.

6.2 data mining 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 technology related. 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-full web logs.

(2) Designing user interface forms needs to consider the data mining 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 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 data mining 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 data mining 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 data mining exercise may not be correct.

7. Conclusions and future work

In this paper, we have presented how web mining (in a broad sense, data mining applied to ecommerce) is applicable to improving the services provided by e-commerce based enterprises. Specifically, we first discussed some popular tools and techniques used in data mining. Statistics, AI and database methods were surveyed and their relevance to data mining in general was discussed. We then presented a host of applications of these tools to data mining in e-commerce. Later, we also highlighted architectural and implementation issues. We now present some ways in which web mining can be extended for further research. With the growing interest in the notion of semantic web, an increasing number of sites use structured semantics and domain ontologies as part of the site design, creation, and content delivery. The notion of Semantic Web Mining was introduced by Berendt et al (2002). 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, including the data preparation, pattern discovery, and recommendation phases. Such a process must involve some or all of the following tasks and activities (Mobasher 2004).

(1) Ontology learning, extraction, and preprocessing: Given a page in the web site, we must be able extract domain-level structured objects as semantic entities contained within this page.

(2) Semantic data mining: In the pattern discovery phase, data mining algorithms must be able to deal with complex semantic objects.

(3) Domain-level aggregation and representation: Given a set of structured objects representing a discovered pattern, we must then be able to create an aggregated representation as a set of pseudo objects, each characterizing objects of different types occurring commonly across the user sessions.

(4) Ontology-based recommendations: Finally, the recommendation process must also incorporate semantic knowledge from the domain ontologies.

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

- Crawler/bot/spider/robot 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 data mining algorithms to yield sensible and pragmatic proposals.

- Data transformations: There are two sets of transformations that need to take place first data must be brought in from the operational system to build a data warehouse, and second 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. While the first set of transformations needs to be modified infrequently (only when the site changes), the second set of transformations provides a significant challenge faced by many data mining tools today.

- Scalability of data mining algorithms: With a large amount of data, two scalability issues arise: (i) most data mining algorithms cannot process the amount of data gathered at web sites in reasonable time, especially because they scale nonlinearly; and (ii) generated models are too complicated for humans to comprehend.

The above challenges need to be better addressed in real world tools.

Episode mining involves mining not one-time events, but mining for a historical pattern of events. Episode-mining methods rely on extensions of rule-mining methods. Alternate approaches could be explored here. Support vector machines (Haykin 1998) have taken the centre stage of late, in learning linear and nonlinear relationships. Their applications in episode mining could be an exciting area for further work.

REFERENCES

Agrawal R, Srikant R 1994 Fast algorithms for mining association rules. In 20th Int. Conf. on VeryLarge Databases (New York: Morgan Kaufmann) p 487–499

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

Banks, D. L., and Said, Y. H. (2006). Data mining in electronic commerce. Statistical Science, 21(2), 234-246

Box G, Jenkins G, Reinsel G 1994 Time series analysis: Forecasting and control 3rd edn (Englewood Cliffs, NJ: Prentice Hall)

Carbone P L 2000 Expanding the meaning of and applications for data mining. In IEEE Int. Conf. on Systems, Man, and Cybernetics (New York: IEEE) pp 1872–1873

Gujarati D 2002 Basic econometrics (New York: McGraw-Hill/Irwin)

Hu X, Cercone N 2002 An olam framework for web usage mining and business intelligence reporting.

In Proc. IEEE Int. Conf. on Fuzzy Systems, FUZZ-IEEE'02 (New York: IEEE Comput. Soc.)pp 950–955

Kohavi, R. (2001). Mining e-commerce data: The good, the bad, and the ugly. Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 8-13.

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)

Koperski K, Adhikary J, Han J 1996 Spatial data mining: Progress and challenges. J. Data Mining Knowledge Discovery Washington

Krishnaswamy, S., Zaslavsky, A., and Loke, S. W. (2000). An architecture to support distributed data mining services in e-commerce environments. Advanced Issues of E-Commerce and Web-Based Information Systems, Second International Workshop on, 239-246.

Lazarevic A, Xu X, Fiez T, Obradovic Z (1999) Clustering-regression-ordering steps for knowledge discovery in spatial databases. In Proc. of IEEE/INNS Int. Conf. on Neural Neural Networks

MobasherB2004Web usage mining and personalization. In Practical handbook of internet computing (ed.) M P Singh (CRC Press)

Neter J, Kutner M, Nachtsheim C J,WassermanW1996 Applied linear statistical models (New York: McGraw-Hill/Irwin)

Niu L, Yan XW, Zhang C Q, Zhang S C 2002 Product hierarchy-based customer profiles for electronic commerce recommendation. In Int. Conf. on Machine Learning and Cybernetics pp 1075–1080

Olson, D. L. (2007). Data mining in business services. Service Business, 1(3), 181-193

Park S 2000 Neural networks and customer grouping in e-commerce: a framework using fuzzy art. In Academia/Industry Working Conference on Research Challenges, pp 331–336

Raghavan, N. R. S. (2005). Data mining in e-commerce: A survey. Sadhana - Academy Proceedings in Engineering Sciences, 30(2-3), 275-289

Sarukkai R R 2000 Link prediction and path analysis using markov chains. In Proceedings of the 9th International World Wide Web Conference, Amsterdam

Smith, M., Wenerstrom, B., Giraud-Carrier, C., Lawyer, S., and Liu, W. (2007). Personalizing e-commerce with data mining, Studies in Computational Intelligence (Vol. 37, pp. 273-286).

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

Vallamkondu S, Gruenwald L (2003) Integrating purchase patterns and traversal patterns to predict http requests in e-commerce sites. In IEEE Int. Conf. on e-commerce, pp 256–263

WangW, Zaiane O R (2007) Clustering web sessions by sequence alignment. In 13th Int. Workshop on Database and Expert Systems Applications pp 394–398

Wang, H., and Wang, S. (2008). A knowledge management approach to data mining process for business intelligence. Industrial Management and Data Systems, 108(5), 622-634

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