Data Mining for Web-Enabled Electronic Business Applications

Richi Nayak

School of Information Systems
Queensland University of Technology
Brisbane QLD 4001, Australia
r.nayak@qut.edu.au

ABSTRACT

Web-Enabled Electronic Business is generating massive amount of data on customer purchases, browsing patterns, usage times and preferences at an increasing rate. Data mining techniques can be applied on all the data being collected for obtaining useful information. This chapter attempts to present issues associated with data mining for web-enabled electronic-business.

INTRODUCTION

Web-Enabled Electronic Business (e-business) is generating massive amount of data such as customer purchases, browsing patterns, usage times and preferences at an increasing rate. What can be done to utilize this large volume of web data with rich description? One possible solution is processing of all the data being collected and obtaining some useful information. For an instance, mining of such web-enabled e-business data can provide valuable information on consumer buying behaviour, which is buried deep within the data otherwise, resulting in an improved quality of business strategies.

As corporations look toward the next phase of e-business (i.e. web-enabled), one thing is clear that it will be hard to continue to capture customers in the future without the help of data mining. Data mining techniques assist e-businesses to seek and retain the most profitable customers by analysing demographic data, customer-buying and traversing patterns collected online or offline (Kohavi & Provost, 2001). Essentially, e-business companies can improve products quality or sales by anticipating problems before they occur with the use of data mining techniques. Data mining, in general, is the task of extracting implicit, previously unknown, valid and potentially useful information from data (Fayyad et al., 1995).

Examples of data mining in e-business applications are generation of user profiles, enabling customer relationship management, and targeting web advertising based on user access patterns that can be extracted from the web data. Data mining in web-enabled e-business domain is currently a `hot' research area. The objective of this chapter is to present and discuss issues associated with data mining for web-enabled e-business applications.

This chapter starts with brief description of basic concepts and techniques of data mining. This chapter then extends these basic concepts for the web-enabled e-business domain. This chapter also discusses challenges for data mining techniques when faced with e-business data and strategies that should be implemented for better use of web-enabled electronic business.

WHAT IS DATA MINING?

A typical data mining process starts with identifying a data mining problem depending on the goals and interest of a data analyst. Next, all sources of information are identified and a subset of data is generated from the accumulated data for the data mining application. To ensure quality, the data set is preprocessed by removing noise, handling missing information and transforming to an appropriate format. A data mining technique or a combination of techniques appropriate for the type of knowledge to be discovered is then applied to the derived data set. The discovered knowledge is then evaluated and interpreted, typically involving some visualization techniques. Finally the information is presented to user to incorporate into the company's business strategies.

A data mining task can be decomposed into many sub tasks when dealing with web-enabled e-business data. Figure 1 illustrates a typical data mining process for web documents. The process starts with locating and then retrieving intended web documents or web access logs. The next and most important task is analysis of data obtained from web document(s) or logs. This includes preprocessing, actual mining process and knowledge assimilation. In the end, the discovered knowledge is presented to user in a format that is appropriate to its goal. The analysis may indicate how a web site is useful in making decision for a user or not. Information for a company to improve its web site can be concluded from this analysis. The analysis may indicate business strategies to acquire new customers and retaining the existing one.

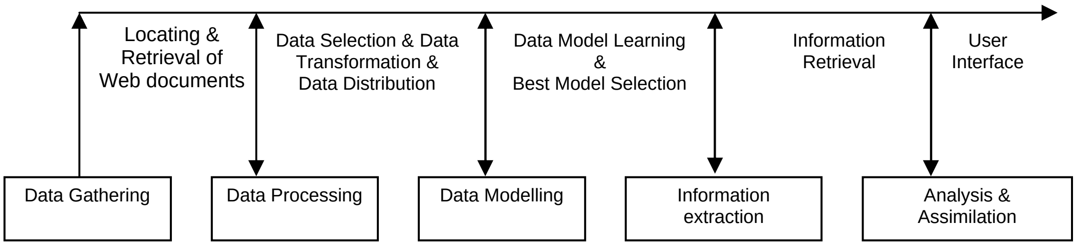

Figure 1: A mining process for web-enabled e-business data

Various data mining tasks and techniques

Depending on the goals and interests of an end user, a data mining process can have three possible tasks - predictive modelling, clustering and link analysis[1].

Predictive Modelling The goal of predictive modelling is to make predictions based on essential characteristics about the data (Berry & Linoff, 2000). These goals are achieved by classification and regression tasks of data mining. The classification task of data mining builds a model to map (or classify) a data item into one of several predefined classes. The regression task of data mining builds a model to map a data item to a real-valued prediction variable. Both the tasks have same basic objective - to make prediction about variable(s) of interest. The difference lies in the nature of the variable(s) being predicted - categorical variable(s) for the classification data mining task and continuous variable(s) for the regression data mining task.

Any supervised machine learning algorithm, that learns a model on previous or existing data, can be used to perform this type of data mining task. The model is given some already known facts with correct answers, from which the model learns to make accurate predictions. Mainly three techniques namely neural induction, tree induction and bayesian classifiers are used for classification data mining tasks (Lim & Loh, 2000). Some other classification methods are K-nearest neighbour classifiers, case based reasoning, genetic algorithms, rough set and fuzzy set approaches (Berry & Linoff, 2000; Han & Kamber, 2001). Mainly three techniques namely linear regression, nonlinear regression and radial basis function are used for regression data mining tasks (Cabena et al., 1997).

Clustering The goal of clustering data mining task is to identify items with similar characteristics, and thus creating a hierarchy of classes from the existing set of events. A data set is partitioned into segments of elements (homogeneous) that share a number of properties. Elements in a cluster are in close proximity to each other, and elements in different clusters are far apart to each other. Usually the proximity is measured by some distance between elements or clusters.

Any unsupervised machine learning algorithm, for which a predetermined set of data categories is not known for the input data set, can be used to perform this type of data mining task. The model is given some already known facts, from which the model derives categories of data with similar characteristics. When a new fact or event comes across, the learned model is capable of categorizing that fact to an appropriate cluster. Some major clustering methods are partitioning, hierarchical, density based and model based algorithms (Han & Kamber, 2001).

Link analysis The goal of link analysis is to establish internal relationship among items in a given data set. This goal is achieved by association discovery, sequential pattern discovery and similar time sequence discovery tasks (Cabena et al., 1997). These data mining tasks expose samples and trends by predicting correlation of items that are otherwise not obvious. Association discovery builds a model to find items implying the presence of other items (with a certain degree of confidence and support) in the given data set. This process reveals hidden affinity among the items i.e. which items are frequently purchased together or which web sites are accessed together. Sequential discovery builds a model to detect an interesting trend between actions or events such that the presence of one set of item is followed by other set of items in a sequence of actions or events over a period of time. The resulting model detects association among events with certain temporal relationship. Similar time sequence discovery builds a model to find similar occurrences in a time series data set. This process reveals hidden information (similar or dissimilar) about patterns of sales (or browsing) of two different products (or web sites) over time.

The link analysis techniques are based on counting occurrences of all possible combination of items. The basic association discovery algorithms are considered very simple. Some of the most widely used algorithms are Apriori and its variation (Agrawal & Srikant, 1994).

DATA MINING IN WEB ENABLED E-BUSINESS DOMAIN

A small shop owner builds relationship with its customers by noticing their needs, remembering their preferences and buying behaviour. A web-enabled e-business will like to accomplish something similar. It is a relatively easy job for the small shop owner to learn from past interactions to serve his customers better in the future. But, this may not be easy for web-enabled e-businesses when most customers may never interact personally with its employee, and there may be a lot more customers than a small shop owner has. Data mining techniques can be applied to understand and analyse such data, and turned into actionable information, that can support a web enabled e-business to improve its marketing, sales and customer support operations. This seems to be more appealing, especially when (1) data is being produced and stored with advance electronic data interchange methods, (2) the computing power is affordable, (3) the competitive pressure among businesses is strong, and lastly (4) efficient and commercial data mining tools have become available for data analysis.

The general statistical approaches of data analysis fail due to large amount of data available for analysis (Cabena et al., 1997). These traditional approaches to data analysis generally start by reducing the size of data. The reduced data facilitates data analysis on the available hardware and software systems. Data mining, on the other hand, is the process to searching for trends and valuable anomalies in the entire data. The process gets benefited with the availability of large amount of data with rich description. The rich descriptions of data such as wide customer records with many potentially useful fields allow data mining algorithms to search beyond obvious correlations.

Data mining opportunities

One of the challenges in web-enabled e-businesses is to develop ways of gaining deep understanding into the behaviour of customers based on the data collected from a web site. Observing customer behaviour is important information for predicting customer behaviour in future. Data mining provides a new capability to company managers by analysing data derived from the interaction of users with the web.

In general, data obtained from web-enabled e-business transactions is (1) primary data that includes actual web contents, and (2) secondary data that includes web server access logs, proxy server logs, browser logs, registration data if any, user sessions, user queries, cookies, etc (Cooley et al., 1997; Kosala & Blockeel, 2000).

Mining of primary web data Given the primary web data, the goal is to effectively interpret the searched web documents. Web search engines discover resources on the web but have many problems such as (1) the abundance problem, where hundreds of irrelevant data are returned in response to a search query, (2) limited coverage problem, where only a few sites are searched for the query instead of searching the entire web, (3) limited query interface, where user can only interact by providing few keywords, (4) limited customization to individual users, etc (Garofalakis et al., 1999). Mining of primary data i.e. actual web contents can help e-business customers to improve the organization of retrieved result and to increase the precision of information retrieval (Jicang et al., 1997). The basic categorization, clustering, association analysis and trend prediction techniques can be utilized within the retrieved information for better organization. Some of the data mining applications appropriate for such type of data are:

- applying trend prediction within the retrieved information to indicate future values. For example, an e-auction company provides information about items to auction, previous auction details, etc. Predictive modelling can be utilized to analyse the existing information, and to estimate the values for auctioneer items or number of people participating in future auctions.

- applying text clustering within the retrieved information to understand efficiently. For example structured relations can be extracted from unstructured text collections by finding the structure of web documents and present a hierarchical structure to represent the relation among text data in web documents (Wong & Fu, 2000).

- applying association analysis to monitor a competitor's web site. Data mining techniques can help e-businesses to find unexpected information from its competitor's web sites e.g. offering unexpected services and products (Liu et al., 2001). Because of the large number of competitor's web sites and huge information in them, automatic discovery is required. For instance, association rule mining can be used to discover frequent word combination in a page that will lead a company to learn about competitors (Liu et al., 2001).

- discovering similarity and relationships between different web sites so to categorize web pages. This categorization will lead in efficiently searching the web for the requested web documents within the categories rather than the entire web. The categorization can be obtained by using either clustering or classification techniques. Cluster hierarchies of hypertext documents can be created by analysing semantic information embedded in link structures and document contents (Kosala & Blockeel, 2000). Documents can also be given classification codes according to keywords present in them.

- using web query languages to providing a higher level of organization for semi-structured or unstructured data available on the web. Users do not have to scan the entire web site to find the required information, whereas they can use web query languages to search within the document or to obtain structural information about web documents. A web query language restructures extracted information from web information sources that are heterogenous and semi-structured (Abiteboul et al., 1997; Fernandez & Suciu, 1999). An agent based approach involving artificial intelligent systems can also be used to organize web based information (Dignum & Cortes, 2001).

Mining of secondary web data Secondary web data includes web transaction data extracted from web logs. Given the secondary web data, the goal is to capture the buying and traversing habits of customers in an e-business environment. Any existing pattern recognition method such as a traditional classification and clustering method can be utilized for this task after applying some preprocessing steps on the data. Some of the data mining applications appropriate for such type of data are:

- promoting campaign by cross-marketing strategies across products. Data mining techniques can analyse logs of different sales indicating customer's buying patterns (Cooley et al., 1997). Classification and clustering of web access log can help a company to target their marketing (advertising) strategies to a certain group of customers. For example, classification rule mining is able to discover that a certain age group of people from a certain locality are likely to buy a certain group of products. Web enabled e-business can also take benefit of link analysis for repeat buying recommendations. Schulz et al (1999) applied link analysis in traditional retail chains and have found that 70% cross-selling potential exists. Associative rule mining can be applied to find frequent products bought together. For example, association rule mining can discover rules such as “75% customers who place an order for product1 from the /company/product1/ page place the order for product2 from the /company/product2/ page as well”.

- maintaining or restructuring web sites in order to better serve the needs of customers. Data mining techniques can assist in web navigation by discovering authority sites of a user's interest, and overview sites for those authority sites. For instance, association rule mining can be applied to discover correlation between documents in a web site and thus estimate the probability of documents being requested together (Lan et al., 1999). An example association rule resulting from analysis of a travelling e-business company web data is: “79% of visitors who browsed pages about *Hotel* also browsed pages on *visitor information: places to visit*”. This rule can be used in redesigning the web site by directly linking the authority and overview web sites.

- personalization of web sites according to each individual's taste. Data mining techniques can assist in facilitating the development and execution of marketing strategies such as dynamically changing a particular web site for a visitor (Mobasher et al., 1999). This is achieved by building a model representing correlation of web pages and users. The goal is to build groups of users performing similar activities. The built model is capable of categorizing web pages and users, and matching between and across web pages and/or users (Mobasher et al., 1999). According to the clusters of user profiles, recommendations can be made to a visitor on return visit or to new visitors (Spiliopoulou et al., 1999). For example, people accessing educational products in a company web site between 6-8 pm on Friday can be considered as academics and can be focused accordingly.

Difficulties in applying data mining

The general idea of discovering knowledge in large amounts of data with rich description is both appealing and intuitive, but technically it is significantly challenging and difficult. There must be some data mining strategies that should be implemented for better use of data collected from web-enabled e-business sources. Some of the difficulties faced by data mining techniques in web-enabled e-business domain and their possible solutions are suggested in this section.

Data Format Data collected from web-enabled e-business sources is semi-structured and hierarchical, i.e. the data has no absolute schema fixed in advance, and the extracted structure may be irregular or incomplete (Abiteboul et al., 2000).

This type of data requires additional steps before applying to traditional data mining models and algorithms, whose source is mostly confined to structured data. This additional step includes transforming unstructured data to a format suitable for traditional data mining methods. Web query languages can be used to obtain structural information from semi-structured data. Based on this structural information, data appropriate to traditional data mining techniques are generated. Web query languages that combine path expressions with an SQL-style syntax such as Lorel (Abiteboul et al., 2000) or UnQL (Fernandez & Suciu, 1999) seem to be a good choice for extracting structural information.

Data Volume Collected e-business data sets are large in volume. The traditional data mining techniques should be able to handle such large data sets.

Enumeration of all patterns may be expensive and not necessary. In spite, selection of representative patterns that capture the essence of the entire data set and their use for mining the data set may prove a more effective approach. But then selection of such data set becomes a problem. A more efficient approach would be to use an iterative and interactive technique that takes account into real time responses and feedback into calculation. An interactive process involves human analyst in the process, so an instant feedback can be included in the process. An iterative process first considers a selected number of attributes chosen by the user for analysis, and then keeps adding other attributes for analysis until the user is satisfied. The novelty of this iterative method will be that it reduces the search space significantly (due to the less number of attributes involved). Most of the existing techniques suffer from the (very large) dimensionality of the search space (Mitchell, 1997).

Data Quality One major source of difficulties for data mining is data quality. Web server logs may not contain all the data needed for data mining. Also, noisy and corrupt data can hide patterns and make predictions harder. (Kohavi & Provost, 2001).

Nevertheless, quality of data is increased with the use of electronic interchange as there is less space for noise due to electronic storage rather than manual processing of them.

Data warehouses provide a capability for the (good quality) data storage. A warehouse integrates data from operational systems, e-business applications, and demographic data providers, and handles issues such as data inconsistency, missing values, etc. A web warehouse may be used as data source for mining data if available. There has been some initiative to warehouse the web data generated from e-business applications, but still long way to go in terms of data mining (Madria et al., 1998).

Another solution of collecting the good quality web data is the use of (1) a dedicated server recording all activities of each user individually, or (2) cookies or scripts in the absence of such server. Activities of the users include access, inspection and selection of products, retrieval of text, duration of an active session, traversing patterns of web pages (such as number, types, sequence, etc) and collection of users' demographic information such as gender, sex, and location for the user that anonymously accessing the web site, etc. The combination of tags from web pages, product correlation and feedback from the customer companies can also be used. (Chan, 1999; Kohavi, 2001)

Also, when searching for documents, methods of evaluating the usefulness of this document are important. The agent based approaches that involve artificial intelligence systems can be used to discover such web based information.

Data Adaptability Data on the web is ever changing. Data mining models and algorithms should be adapted to deal with real-time data in which new transaction data is incorporated for analysis and the constructed data model are updated as the new data approaches.

User-interface agents can be used to try to maximize the productivity of current users’ interactions with the system by adapting behaviours. Another solution can be to dynamically modifying mined information as the database changes (Cheung et al., 1996) or to incorporate user feedback to modify the actions performed by the system (Chundi & Dayal, 1997).

XML Data It is assumed that in few years XML will be the most highly used language of Internet in representing documents including business. At some stage, XML documents may not completely be in the same format thus resulting in missing values.

Assuming the metadata stored in XML, the integration of the two disparate data sources becomes much more transparent, field names can be matched more easily and semantic conflicts may be described explicitly (Abiteboul et al., 2000). As a result, the types of data input to and output from the learned models and the detailed form of the models can be determined. Various techniques, such as tag recognition, can be used to fill missing information if there is mismatch in attributes, tags or DTDs (Abiteboul et al., 2000). Moreover, many query languages such as XML-QL, XSL (Deutsch et al., 1999) and XML-GL (Ceri et al., 1999) are designed specifically for querying XML and getting structured information from these documents.

Privacy Issues There are always some (privacy) concerns of proper balancing between company's desire to use personal information versus individual's desire to protect it (Piastesky-Shapiro, 2000).

The possible solution is to (1) ensure users of secure and reliable data transfer by using high speed high-valued data encryption procedures, and/or (2) give a choice to user to reveal the information that he/she wants to and give some benefit in exchange of revealing their information such as discount on certain shopping product etc.

CONCLUSION

This chapter attempts to present data mining concepts and issues that are associated with web-enabled e-business applications.

It is easy to collect data from web-enabled e-business sources as all visitors to a web site leave the trail which automatically is stored in log files by web server. The data mining tools can process and analyse such web server log files or actual web contents to discover meaningful information. The data mining techniques provide companies with previously unknown buying patterns and behaviour of their online customers. More importantly, the fast feedback the companies obtained using data mining is very helpful in increasing the company's benefit.

Earlier data mining tools such as C5 (http://www.rulequest.com) and several neural network softwares (QuickLearn, Spmpack, etc) were limited to some individual researchers. These individual algorithms are capable of solving a single data mining task. But now the second generation data mining system produced by commercial companies such as clementine (http://www.spss.com/clementine/), AnswerTree (http://www.spss.com/answertree/), SAS (http://www.sas.com/), IBM Intelligent Miner (http://www.ibm.com/software/data/iminer/) and DBMiner (http://db.cs.sfu.ca/DBMiner) incorporate multiple discoveries (classification, clustering, etc), preprocessing (data cleaning, transformation, etc) and postprocessing (visualization) tasks, and becoming known to public and successful[2]. Moreover, tools that combine ad hoc query or OLAP (Online analytical processing) with data mining is also developed (Wu, 2000). Faster CPU, bigger disks and wireless net connectivity make these tools liable to analyse large volume of data.

Utilization of data mining techniques in assisting the web-enabled e-business content providers and consumers is overall a beneficial transaction (Eckerson, 1999). There are several important aspects of web-enabled e-business where data mining can be beneficial. Some of them are (1) analysis of pattern of user behaviour that reflects the acceptability and satisfaction with a web site, (2) correlation analysis between web contents be it products or documents, (3) analysis of web usage data to assist e-businesses in real-time personalization and making cross-marketing strategies.

A web-enabled e-business company that incorporates data mining results with its strategy is sure to be succeeded.

REFERENCES

Abiteboul, S., Buneman, P., & Suciu, D. (2000). *Data on the Web: From Relations to Substructured Data and XML*, California: Morgan Kaumann.

Abiteboul, S., Quass, D., McHugh, J., Widom, J., & Weiner, J. (1997) The Lorel Query Language for Semistructured Data. *Journal of Digital Libraries, 1*(1), 68-88.

Agrawal, R., & Srikant, R. (1994). *Fast Algorithms for Mining Association Rules*. IBM Research Report RJ9839, IBM Almaden Research Center.

Berry, M. & Linoff, G. S. (2000). *Data Mining: Concepts and Techniques*. John Wiley and Sons, INC.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J. & Zanasi, A. (1997). *Discovering Data Mining from Concept to Implementation*. Prentice Hall PTR.

Ceri, S., Comai, S., Damiani, E., Fraternali, P., Paraboschi, S., & Tanca, L. (1999) XML-GL: A Graphical Language for Querying and Restructuring XML

Documents. In *Proceedings of the eight International WWW Conference*, Toronto.

Chan, P. K. (1999). A non-invasive learning approach to building web user profile. In (Masand & Spiliopoulou, 1999).

Cheung, D. W., Han, J., Ng, V. T. & Wong, C. Y. (1996). Maintenance of Discovered Association Rules in Large Databases: an incremental technique. In *Proceedings of the Twelfth International Conference on Data Engineering*, New Orleans, USA, 106-114.

Chundi, P. & Dayal, U. (1997). An application of adaptive data mining: Facilitating web information access. In *Proceedings of the ACM SIGMOD Workshop on Research issues in Data Mining and Knowledge discovery*, 31-38.

Cooley, R., Mobasher, B., & Srivastave, J., (1997). Web mining: Information and Pattern Recovery on the World Wide Web. In *Proceedings of the ninth International Conference on Tools with Artificial Intelligence.*

Deutsch, A., Florescu, D., Fernandez, M., Levy, A., & Suciu, D. (1999). A query langauge for XML. In *Proceedings of the eight International WWW Conference*, Toronto.

Dignum, F. & Cortes, U. (Eds.). (2001). *Agent-Mediated Electronic Commerce III: Current Issues in Agent-Based Electronic Commerce Systems*. Lecture Notes in Artificial Intelligence. Springer Verlag.

Eckerson, W. E. (1999). Marrying E-commerce and Customer Intelligence. In *Patricia Seybold Group's Information Assets Service*. June. Retrieved October 1, 2001, from http://www.psgroup.com/doc/products/1999/6/PSGP6-18-99IA/PSGP6-18-99IA.asp

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1995). From Data Mining to Knowledge Discovery: An Overview. In Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P. & Uthurusamy, R. (Eds.). *Advances in Knowledge Discovery and Data Mining*. 1-34. AAAI Press.

Fernandez, M. & Suciu, D. (1999). UNQL: A Query Language for Web Site. Retrieved September 21, 2000, from http://www.cs.huij.ac.il/~yarivi/unql-htom.html.

Garofalakis, M. N., Rastogi, R., Seshadri, S., & Hyuseok, S. (1999). Data Mining and the Web: past, Present and Future. In *Proceedings of the second International Workshop on Web Information and Data Management*. 43-47.

Han, J. & Kamber, M. (2001). *Mastering Data Mining*. San Francisco: Morgan Kaufmann.

Jichang, W., & Huan, H., Gangshan, W., & Fugan, Z. (1997). Web mining: Knowledge Discovery on the web. In *Proceedings of the ninth International Conference on Tools with Artificial Intelligence*. Nov.

Kohavi, R. (2001). Mining E-commerce Data: The Good, the Bad and the Ugly. In *Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD 2001).

Kohavi, R. & Provost, F. (2001). Applications of data mining to electronic commerce. *Data Mining and Knowledge Discovery, 5* (1/2).

Kosala, R. & Blockeel, H. (2000). Web Mining Research: A survey. *SIGKDD Explorations. 2*(1), 1-15, July

Lan, B., Bressan, S., & Ooi, B. C. (1999). Making web servers pushier. In (Masand & Spiliopoulou, 1999).

Lim, T. S. & Loh, W. Y. (2000). A comparison of prediction accuracy, complexity and training time of thirty three old and new classification algorithms. *Machine Learning, 40*(3), sep. 203-228.

Liu, B., Ma, Y., & Yu, P. H. (2001). Discovering unexpected information from your competitor's web sites. In *Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2001)*. Aug. SanFrancisco, USA.

Madria, S. K., & Bhowmick, S. S., Ng, W. K., & Lim, E. P. (1998). Research issues in Web data mining. *Applied Artificial Intelligence*, *12*, 303-312.

Masand, B. & Spiliopoulou, M. (1999). *KDD'99 workshop on web usage analysis and user profiling (WEBKDD'99)*. Aug. San Diego, CA. ACM.

Mitchell, T. M. (1997). *Machine Learning*. The McGraw-Hill Companies, Inc.

Mobasher, B., Cooley, R., & Srivastave, J. (1999). Automatic Personalization based on Web Usage Mining. In (Masand & Spiliopoulou, 1999).

Piastesky-Shapiro, G. (2000). Knowledge Discovery in Databases: 10 years after. *SIGKDD Explorations, 1*(2), Jan, 59-61. ACM SIGKDD.

Schulz, A. G., Hahsler, M. & Jahn, M. (1999). A customer purchase incidence model applied to recommendation service. In (Masand & Spiliopoulou, 1999).

Spiliopoulou, M., Pohle, C., & Faulstich, L. C. (1999). Improving the effectiveness of a web site with web usage mining. In (Masand & Spiliopoulou, 1999).

Wong, W. C. & Fu, A. W. (2000). Finding Structure and Characteristic of Web Documents for Classification. In *Proceedings of the ACM SIGMOD Workshop on Research issues in Data Mining and Knowledge discovery*. July. ACM.

Wu, J. (2000). Business Intelligence: What is data mining. In *Data Mining Review Online*. August.

[1] This chapter will not go into the depth of data mining techniques. Interested readers can refer data mining textbooks for the detailed description of these techniques.
[2] An interesting review of data mining softwares compiled by Peter Spirtes can be found in http://crl.research.compaq.com/vision/multimedia/dm/DataMiningSurvey.html