A STUDY OF APPLICATION OF WEB MINING FOR E-COMMERCE: TOOLS AND METHODOLOGY

SaiMing Au

Department of Information System,
City University of Hong Kong,
Tak Chee Avenue,
Hong Kong

Abstract

Internet commerce or e-commerce brings together consumers and merchants all over the world to a virtual marketplace where customization, direct marketing, market segmentation and customer relationship management can take place. In this new marketplace, most marketers find their customer behaviors difficult to understand. Web site mining enables better understanding of the customers, for discovering meaningful business correlations and trends, and for providing better sales and marketing services over the Web. It is an active research area and the tools and methodology is still evolving. This comprehensive study reviews its application, tools and methodology to form a knowledge base for future research in the area. There are many successful commercial application cases and tools available. The web mining methodology is more involving; in its generic form, it comprises of data pre-processing, domain knowledge elicitation, methodology identification, pattern discovery, and knowledge post-processing.

Key words: e-commerce, web mining, application, tools, methodology.

Introduction

Today the Web is more than a place for information exchange. It is an important marketplace for e-commerce. With the Web, every aspect of commerce, from sales pitch to final delivery, can be automated and made available 24 hours a day all over the world. Companies can use their e-commerce platform to improve sales, increase customer satisfaction or reduce cost. E-commerce changes the B2B and B2C relationships, enabling new business models and strategies to develop. For instance, B2B developers can form vertical partnerships and co-branding as innovative business solution, and B2C marketers can find new channels to sell directly to their customers.

As only an effective web site can fulfill what it intends to achieve, marketers and web designers find it necessary to understand the effectiveness of their sites and to take

appropriate action when they fall short. They want to know who their customers are and how they react to their web sites. Although they cannot meet them face-to-face, fortunately there are many footprints left by the web surfers that enable them to study the customer behavior. The key is the computer log. With enormous data of the web surfers available, web mining can be used to learn rules relating to the behavior of customers, turning data into valuable knowledge and untapped business opportunities.

Owing to the great impact of web mining on e-commerce, both the academic and the commercial sectors are doing a lot of researches and application works on this area. With such cross-disciplinary efforts, there is the need to summarize the current research directions and results. This study performs a comprehensive survey of applications area and cases of web mining in e-commerce, and looks into examples of the tools and the details of the methodology.

Application of web mining in e-commerce

The web environment is ideal for having interactive communication and flexible transaction between the sellers and the buyers. Customers can place order anywhere at any time. More proactively, many web miners can base their offers on visitor profiles to create new products that match the results of their analysis. There are many application areas, which include the digital library, browsing enhancement, customized marketing, personalization, customer relationship management, web advertising and web site quality improvement.

- Digital libraries are essentially data management and information management systems that have to interoperate on the web. Due to the large amount of data, integration of mass storage with data management will be critical. Data mining is needed to extract information from the database. Data mining can help the users find information on the web. Commercial site like the Questia store over 35,000 books and deliver them on-line. They use intelligent agents to match the queries of the users with their stored materials.

- Browsing enhancement software can be dated back as early as in 1994, with Letizia produced as a user interface agent that assists a user browsing the World Wide Web. As the user operates a conventional Web browser such as Netscape, the agent tracks user behavior and attempts to anticipate items of interest by doing concurrent, autonomous exploration of links from the user's current position. The agent automates a browsing strategy consisting of a best-first search augmented by heuristics inferring user interest from browsing behavior. It learns user preferences and discovers Web information sources that correspond to these preferences.

- Customized marketing is one key aspect in e-commerce. The Web servers function as “pushier”, with the document to be pushed being determined by a set of association rules mined from a sample of the access log of the Web server [1]. For instance, Perkowitz and Etzioni [2] mine the data buried in Web server logs to produce adaptive Web sites that automatically improve their organization and presentation by learning from visitor access patterns. It allows the service provider to customize and adapt the site’s interface for the individual user, and to improve the site’s static structure within the underlying hypertext system.

- Web personalization tailors the Web experience according to the user's preferences. A good example of e-commerce site using personalization is the Amazon.com, in which customer profiles are stored in the database and appropriate recommendations are pushed to different customers. Most customers welcome this service as very often they find the recommended products really meeting their needs. The technology applies collaborative-filtering to recommend items liked by similar users. Thus users are grouped by sharing similar interest. Then a user is recommended items, which his similar users have rated highly and he has not seen before. Recent development like the Inductive Logic Programming based INDWEB helps Internet users browse the Web by learning a model of their preference [3].

- Customer relationship management adopts a total quality management approach to serve the customers. Marketing experts divide the customer relationship life cycle into three distinct steps, which cover attraction, retention, and cross sales. Buchner and Mulvenna [4] suggest using adaptive web sites to attract customers, using sequential patterns to display special offers dynamically to keep a customer interest in the site, and using customer segments for cross-selling.

- Advertising accounts for the highest sales revenue in the e-commerce. At present, there are several commercial services and software tools for evaluating effectiveness of Web advertising in terms of traffic and sales driven by them. They use metrics such as click-through rates and ad banner ROI. Commercial agents, like the NetZero, track subscribers' traffic patterns throughout their online session and uses the information it collects to display advertisements and content that may be of interest to subscribers. Advancement in the area like the Latent Semantic Analysis (LSA) information retrieval technique by Murray and Durrell [5] is used for targeting advertisement. They construct a vector space to represent the usage data associated with each Internet user of interest. This enables the marketer to infer the demographic attributes of the Web users.
- Web mining can be helpful in the development of strategy to improve the web sites. Spiliopoulou et al [6] propose a new methodology based on the discovery and comparison of navigation patterns of customers and non-customers. The comparison leads to rules on how the site's topology should be improved. Web caching, prefetching and swapping can be applied to improve access efficiency. The problem of classifying customers can also be solved by using a clustering method based on the access pattern [7]. Using attribute-oriented induction, the sessions are then generalized according to a page hierarchy, thereby organizing pages based on their contents. These generalized sessions are finally clustered using a hierarchical clustering method.

Overview of web mining tools

Web mining is more than a simple application of ordinary data mining to the web data. The lack of structure and dynamic nature of the Web content adds difficulty to data extraction and mining. Moreover, instead

of using conventional market research data or customer database showing demographics, researchers need to rebuild the profiles of their customers using computer logs, web content and new transaction variables. Luckily the log data are relatively easy and cheap to collect. The recent rapid development and growing interest in Web mining for the e-commerce is aided by the technical advancement in the use of scripting and CGI that replaces the static web pages by dynamic contents using web page generation on request and applets-like applications. This allows for better logging of the truly personalized and interactive web behaviors of the customers.

There are many computer programs (Table 1) that log the visitors and provide some statistical information. They are not web mining software as they provide little analysis and no data mining facilities. They provide basic statistics pertaining to the visitor categories (by visit frequency), referral, browsing pattern, traffic pattern, entry and leaving pattern, etc.

Table 1: Some Web log / Web site traffic analysis programs

Product	Author / Company	Feature	Function
Analog	University of Cambridge Statistical Laboratory	Measures the usage on web server. It tells which pages are most popular, which countries people are visiting from, which sites they tried to follow broken links from.	Log file analyzer
Webalizer	GNU project	Supports standard Common Log file Format server logs. In addition, several variations of the Combined Log file Format are supported, allowing statistics to be generated for referring sites and browser types as well.	Web server log analyzer
NetTracker	Sane Solution	Analyze multiple web sites, as well as proxy server and firewall log files to monitor the organizations' web surfing patterns, plus FTP log files	Web server log analyzer
Weblog	Webscripts	Relies on Datalog-like rules to represent web documents	Web content mining

For genuine web mining software, in the taxonomy of web mining, three types of web mining are identified according to their main purpose, viz. web content mining, web link structure mining and web usage mining.

Web content mining is about extracting the important knowledge from non-structured or less structured text files. It is useful to information retrieval for indexing documents and assisting users to locate information. Many applications are developed to serve this or related purposes (Table 2). Content mining techniques draw heavily from the work on information retrieval, databases, intelligent agents, etc. It is used for web

page summarization and search engine result summarization, discovering information and extracts knowledge from text documents. For e-commerce application, it is used for classifying the type of web pages that the surfer often visits before web site personalization can be done.

Table 2: Content mining programs

Product	Author / Company	Feature	Function
Intelligent Miner for Text, TextAnalyst	IBM	Implements a variety of analysis functions based on utilizing an automatically created semantic network of the investigated text.	Content mining
MetaCrawler	Selberg & Etzioni, 1995	Provides an interface for specifying a query to several engines in parallel	Search agent
WebWatcher	Amstrong et al, 1995	An agent helping the users to locate the desired information, users input required	Personal agent
Letizia	Lieberman	Uses the idle processing time available when the user is reading a document to explore links from the current position	Behavior-based interface agents Personal agent
SiteHelper	Ngu and Wu, 1997	Use log data to identify the pages viewed by a given user in previous visits to the sites	Page recommendation
LexiBot	Bright Planet	Search agent capable of identifying, retrieving, classifying and organizing "surface" and "deep" Web content.	Search agent
Webdoggie	MIT	Collaborative approach to suggests new WWW documents to the user based on WWW documents in which the user has expressed an interest in the past	Information filtering agent

The second form of web mining, the web structure mining establishes structures amongst many web pages. It identifies authoritative web pages or hubs to improve the overall structure of a series of web pages. Essentially, the Web is a body of hypertext of approximately 300 million pages that continues to grow at roughly a million pages per day. The set of web pages lacks a unifying structure and shows far more authoring style and content variation than that seen in traditional text-document collections. This level of complexity makes an "off-the-shelf" database-management and information-retrieval solution impossible and calls for the need to mine the link structure from the web pages. A way of structure mining takes the advantage of the collective judgment of web page quality in the form of hyperlinks. The most frequently visited paths

in a Web site are used as the objective assessment of the quality of web sites as perceived by the customers, as path analysis can be used to determine. This principle is used by popular programs like the PageRank and CLEVER (see Table 3). Another popular program for structure mining is the WebViz by Pitkow et al [8]. It is a system for visualizing WWW access patterns. It allows the analyst to selectively analyze the portion of the Web that is of interest by filtering out the irrelevant portions.

Table 3: Programs for web structure mining

Product	Author / Company	Feature	Function
PageRank	Larry Page	Using the Web link structure as an indicator of an individual page's value. In essence, it interprets a link from page A to page B as a vote	Search engine
CLEVER	IBM Almaden Research Center.	Incorporates several algorithms that make use of hyperlink structure for discovering high-quality information on the Web	Hypertext Classification, Mining Communities
WebViz	Tamara Munzner, Paul Burchard	*Information hierarchy visualization* Web as a graph: nodes are documents, edges are links	3D graphical representation of the structure of the Web

The third form, mining for usage pattern is the key to discover marketing intelligence in e-commerce. It helps tracking of general access pattern, personalization of web link or web content and customizing adaptive sites. It can disclose the properties and inter-relationship between potential customers, users and markets, so as to improve Web performance, on-line promotion and personalization activities. There are many popular programs for usage pattern mining (see Table 4). Web Log Mining [9] uses KDD techniques to understand general access patterns and trends to shed light on better structure and grouping of resource providers. The WEBMINER [10] discovers association rules and sequential patterns automatically from server access logs. Commercial software WebAnalyst by Megaputer learns the interests of the visitors, based on their interaction with the website. User profiles are modified in real time as more information is learned. Clementine and DB2 Intelligent Miner for Data are two general-purpose data mining tools, which can be used for web usage mining with suitable data preprocessing.

Table 4: Some web usage mining programs

Product	Author / Company	Feature	Function
WebMate	Chen & Sycara, 1998	The user profile is inferred from training examples	Proxy agent
WebLogMiner	Zaiane et al	Use data mining and OLAP on treated and transformed web access files.	Mining web server log files
SpeedTracer	IBM	Use the referrer page and the URL of the requested page as a traversal step and reconstructs the user transversal paths for session identification	Mining web server log files
Web usage miner (WUM)	Myra Spiliopoulou	To analyze the navigational behavior of users, appropriate for sequential pattern discovery in any type of log. It discovers patterns comprised of not necessarily adjacent events	Discovers navigation patterns in the form of graphs
WEBMINER	R. Cooley and J. Srivastava	A general and flexible framework for Web usage mining, the application of data mining techniques, such as the discovery of association rules and sequential patterns, to extract relationships from data collected in large Web data repositories	Restructure a Web site, and in analyzing user access patterns to dynamically present information tailored to specific groups of users
Clementine	SPSS	To browse data using interactive graphics to find important features and relationships	CRM
WebAnalyst	Megaputer	Integrates the data and text mining capabilities of analytical software directly	Profiles the website resources and dynamically identifies the most appropriate resources to serve each visitor
DB2 Intelligent Miner for Data	(IBM)	Provides a single framework for database mining using proven, parallel mining techniques	User database miner

Methodology of web mining

The web mining process can be divided into 4 stages: data preprocessing, domain knowledge elicitation, methodology identification and knowledge post processing.

The details of these steps are different with different purposes for web mining and they are cross-tabulated in Table 5.

Table 5: Methodology of web mining

Stages	Content mining	Web link mining	Web usage mining
Data preprocessing	• Linguistic processing: tagger, expression, terminology, semantics, • Feature extraction, • Document analysis (Content-only), • Feature selection, • Feature weighting,	• Definition of session; • To reorganize log entries supported by meta data; • Manipulation of date and time related fields • Removal of futile entries	• Selection of data source • Definition of session • Process of web log content • Transaction identification (content / navigation- content): E.g. Page representation
Domain knowledge elicitation (feature selection)	• Incorporation of linguistic, lexical, and contextual techniques	• Traffic analysis; mining dynamic log file	• Syntactic constraint, • Navigation template, • Network topologies, • Concept hierarchy
Methodology identification	• To build a n-dimensional web log cube • Application of OLAP	• Sequence association • Build data cubes from Web server logs for data mining	• Build data cubes from Web server logs for OLAP and data mining • Research on query language
Knowledge post processing	• Directory hierarchy • Search result	• Path and node representation	• Graph and visualization • Rule extraction • Model validation

As shown by the complexity of the methodology concerned, instead of covering all the aspects, the following discussion uses web usage mining to highlight the procedure involved. Web usage mining is chosen, as it is a more popular research area relating to e-commerce and it is useful to the customer behavior. As a matter of fact, web usage mining incorporates the input from both context mining and link structure mining.

There are two approaches for web usage mining, viz. an OLAP data warehousing approach and a holistic integrated approach.

Zaiane et al's WebLogMiner [9] adopt an OLAP data warehousing approach to build up a data warehouse or data mart from Internet log files and then use OLAP or data mining technique. The most important step is to obtain a good data mart that can be summarized in Fig.1. With so many data fields and records, it calls for sophisticated techniques for data warehousing and data mining.

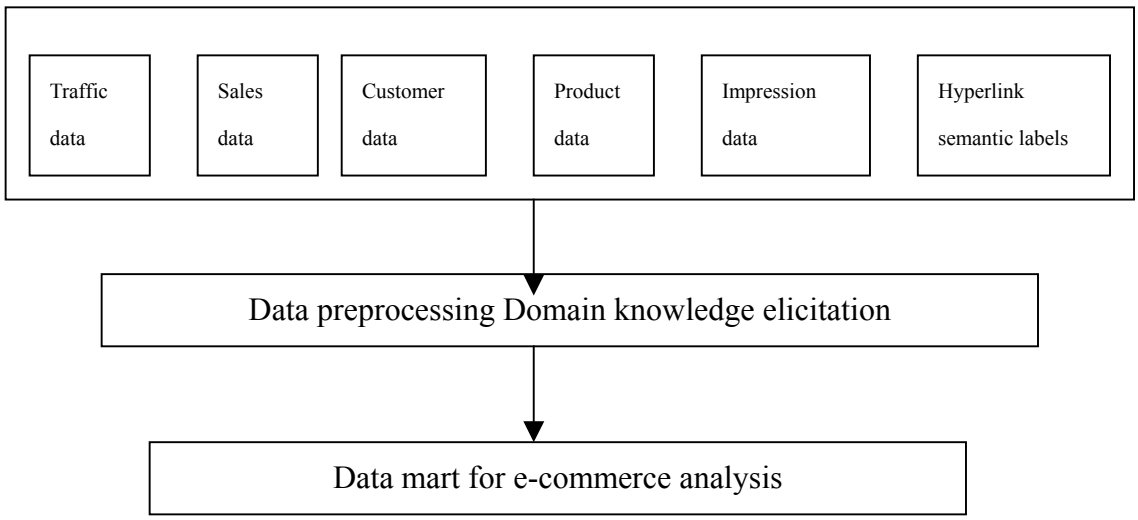

Fig. 1: Data mart preparation

On the other hand, the holistic approach proposed by Buchner et al [11] uses an integrated approach with emphasis on the interaction of different sub-systems and agents to extract information from various sources. They use the M*i*das (Mining Internet Data for Associative Sequences) to discover marketing intelligence from Internet data. The data sources include server data (server log, error log and cookie log), marketing data and knowledge (customers, products, transaction, domain expertise), and web meta data. They emphasize on the involvement of different experts required for different sub-system: a web administrator, a marketing expert, and a data-mining specialist. Web Site Information Filter (WEBSIFT) by Cooley et al [10] contains an architecture elaborated on the same theme. Spiliopoulou's [12] sequence discoverer for web data, the Web Utilization Miner WUM is very similar. It pre-processes the data and organizes the log into sessions according to user-specified criteria. Subsequently, an 'aggregation service' transforms the log of sequences into a tree structure, where sequences with the same prefix are merged. WUM processes this reduced-size structure and applies further heuristics to improve performance. Aggregated trees are generated from log files in order to discover user-driven navigation patterns.

The 4 stages of web usage mining viz, data preprocessing, domain knowledge elicitation, methodology identification and knowledge post processing are more distinctive in the holistic approach and they are discussed in the subsequent sections.

First step: data preprocessing

Data preprocessing involves data source selection, session identification and transaction identification. Firstly, the researchers need to select data sources from web server logs, referral logs, registration files, cookies log, transaction log, index server logs and the query data to a web server, etc. The server log contains information about the click-streams that shows how a web site is navigated and used by its visitors. The transaction log relates the click-stream data of the users to actual purchases and is useful for understanding the effectiveness of marketing and merchandising efforts. In addition, the transaction log can be merged with some external database containing data from

customer surveys, point-of-sale terminals, inventory databases and product-mix information.

Secondly, researchers need to define session to overcome the difficulty of identification of individual users in a web, as many web servers only record the IP address of clients shared by more than one user. A session is defined as an episode of interaction between a Web users and the Web server consisting of the pages the user visited and the time spent on each page. It is grouped by consecutive pages requested by the same user together. The use of a properly defined session is still limited by many other factors and the related results must be interpreted carefully. For instance, the session cannot tell the browsing time of the users as the actual time spent on each page is always different from the estimation because of network traffic, server load, user reading speed, etc. Moreover, the existence of caching functions and proxy servers render the server log not reflecting the actual history of browsing.

Thirdly, researchers need to identify transaction by extracting appropriate fields and merge them together into meaningful clusters of references for each user. This process can be extended into multiple steps of merge or divide in order to create transactions for a given data mining task. One way is to classify a given page to either content-only or navigation-content, based on the time spent on it [9].

Content-only transactions consist of all of the content references for a given user session. These transactions can be used to discover associations between the content pages of a site. This kind of "page typing" is delineated by Pirolli et al [13], using various page types such as index pages, personal home pages, etc. in the discovery of user patterns. Their page representation represents the content of the visited pages using a vector space model in which documents are represented as real-valued vectors. Each element corresponds to the frequency of occurrence of a particular term in the document.

For the navigation-content transactions, they consist of a single content reference and all of the navigation references in the traversal path leading to the content reference. These transactions can be used to mine for path traversal patterns. They are represented in many different ways:

- Maximal forward reference is defined by Chen et al [14] as the set of pages in the path from the first page in the log for a user up the page before a backward reference is made. A new transaction is started when the next forward reference is made. A forward reference is defined to be a page not already in the set of pages for the current transaction. A backward reference is defined to be a page that is already contained in the set of pages for the current transaction. The WEBMINER system [11] currently has reference length, maximal forward reference, and time window divide modules, and a time window merge module to achieve the transaction definition. This can identify the most traversed paths through a Web locality.

- Page Interest Estimator (PIE) argues that the History, Bookmarks and Links of the users indicate their interest [15]. In the History, a higher frequency and more recent visits of an URL indicate stronger user interest of that URL. For the Bookmarks, pages that are bookmarked are of strong interest to the user. Similarly, a higher percentage of links visited from a page indicates a stronger user interest in that page. Considering all these factors thus can form an index reflecting the interest of the users.

- Web Access Graph (WAG) looks at the path transversal patterns of the customers

instead of looking at the content of individual page. A WAG is a weighted directed graph to represent a user's access behavior. Each vertex in the graph represents a web page and stores the access frequency of that page. The intensity of a vertex indicates the interest level of the corresponding page and the thickness of an edge depicts the degree of association between two different pages. Chan estimates the user's interest of a web page by locating multi-word phrases to enrich the common bag-of-words representation for text documents [15]. PIE's then learns to predict the user's interest on any web page, and a WAG is used to summarize the web page access patterns of a user. The user profile can be utilized to analyze search results and recommend new and interesting pages.

- Sequential patterns are special patterns discovered from the log, which can be constrained by a number of factors, such as support, minimum and maximum time gaps (between user sessions). These patterns are then recorded and mined. This allows any number of log files to be combined in any order determined by the analyst. Sequential pattern is first proposed by Agrawal & Srikant, using the GSP algorithm [16]. The MiDAS proposed by Baumgarten et al is an improvement over the GSP algorithm [11], using depth-dependent pattern trees.

- Hypertext probabilistic grammar is used by Borges and Levene to capture the user navigation behavior patterns [17]. Higher probability strings are used to reflect the user's preferred trials. They propose using entropy as an estimator of the grammar's statistical properties.

Second step: Domain knowledge specification

Effective data mining relies on further process of the transaction log and the web page content supported by external knowledge. In order to discover web-specific sequential patterns, domain knowledge may be incorporated with the objective to constrain the search space of the algorithm, reduce the quantity of pattern discovered and increase the quality of the discovered patterns. From the raw data, the modeler can generate some relationships. For instance it can report the most frequent visitors to a set of web pages by demographic classification. Sometimes some business models may be useful. For example, visitors may be classified as short time visitors, active investigators and customers as appropriate and the data will be drawn and reported accordingly. A popular bunch of tools use some flexible navigation templates for domain knowledge representation, which is summarized by Baumgarten et al [11]:

- Syntactic constraint uses threshold sextuples representing the minimum support and minimum confidence. It is possible to eliminate shallow navigational patterns and enables the researchers to focus on more important navigation patterns.

- Navigation template can specify the pattern in the start, middle and end with constants, wildcards, and predicates restricting the permissible values of the pattern. This can prepare tuples that are specified by the researchers.

- Network topologies can derive the topology from log files, based on all site internal http referrers - URL document name links.

- Concept hierarchy redefines page relationships other than the URL document name links. A typical application is the topological organizations of Internet domain levels. In addition, marketing-related hierarchies, such as product categorizations or customer locations can also be used.

Third step: Methodology identification and pattern discovery

After the data are preprocessed, the data will be transformed, cleansed, normalized, integrated with some well-established procedures. The processed data then can be used for On-Line Analytical Processing (OLAP) or data mining.

OLAP is a special category of query and reporting tools that can be used to pulled data out of a database. They are designed to support complex, multi-dimensional and multi-level on-line analysis of large volumes of data stored in data warehouses. OLAP can also be used in path analysis to determine frequent traversal patterns or large reference sequences from the physical layout type of graph. It can be used to determine the most frequently visited paths in a Web site. The WEBMINER system [18] proposes an SQL-like query mechanism for querying the discovered knowledge (in the form of association rules and sequential patterns). Besides, new algorithm like M*i*Das supports sequence discovery from multidimensional data to detect sequence across monitored attributes, such as URLs and http referrers [12]. The mechanism has been incorporated in an SQL-like query language (called MINT), which together form the key components of the Web Utilization Analysis platform.

Data mining tools work in a way very similar to that of statistical tools, but the user is much less active in the analysis process than when using the statistical tools. Due to its own objectives and data representation, web mining employs a special subset of the general data mining tools. Web mining tools can be grouped according to the desired outcomes: classification, sequence detection, data dependency analysis and deviation analysis. Some of the more popular tools for web mining include the Association rule, Classification rules (CHAID), cluster analysis and neural network.

Fourth step: knowledge post processing

Knowledge post processing is important to convey the finding to the decision makers. The managers like decision rules, as they are easier to understand and apply. Very often, some other summarization and visualization techniques are required to make the results more meaningful to the market managers. In addition, the result of web mining can be processed by the XML that is a standard language for formatting the responses from database systems so that the web clients can understand the results.

Concluding remarks

The potential of using a website as a data collection tool for e-commerce is enormous, because of its interactiveness, simplicity and unobtrusiveness. The results of the data mining would ideally be integrated into the dynamic website to provide an automated, end-to-end functional system for target marketing and customer relationship management. Most of the web mining tools are evolving and the present web mining techniques still have rooms for improvement to make them prevail in the e-commerce. Some problems like the need for greater integration, scalability problem, and the need for better mining tools are just some problems mentioned by many researchers [19].

The sharpening on the mining tools in many different aspects are important for the future development in this area:

- Web usage mining must handle the integration of offline data with e-business

analytic tools, RDBMS, catalogs of products and services and other applications.

- Some new variables or logs should be sought that can be used for finding more natural, meaningful and useful patterns.

- New tools are sought which will not use up too much resources or process time during the web mining process.

- There will always be a need to have benchmark tests to improve the performance of mining algorithms, as the efficiency and effectiveness of a mining algorithm can be measured and a better tool for web data mining can be used.

- It is important to improve visualization, as much of the data is unorganized and difficult for the user to understand.

Web mining is a new and rapidly developing research and application area. With more collaborative research across different disciplines like database, artificial intelligence, statistics and marketing, we will be able to development web mining applications that are very useful to the e-commerce community.

References

1. Lan, B. Stephae Bressan, BengChin Ooi and Y.C. Tay (1999): "Making Web Servers Pushier", *International WEBKDD'99 Workshop* San Diego, Ca, USA, August 1999. Revised Papers.

2. Perkowitz, M and O. Etzioni (2000), "Towards Adaptive Web Sites: Conceptual Framework and Case Study", *Artificial Intelligence*, 118 (2000) 245-275.

3. Jacquent, F. and P. Brenot (1998), "Learning User Preferences on the WEB", *Research and Development in Knowledge Discovery and Data Mining*, Second Pacific-Asia Conference, PAKDD-98, Melbourne, Australia, *1998 Proceedings*.

4. Buchner, A., Maurice Mulvennan, Sarabjot Anand and John Huges (2000): "An Internet-enabled Knowledge Discovery Process", (Internet resources, to be identified later)

5. Murray, D and Kevan Durrell (1999): "Inferring Demographic Attributes of Anonymous Internet Users", *International WEBKDD'99 Workshop* San Diego, Ca, USA, August 1999. Revised Papers.

6. Berendt, B. and Myra Spiliopoulou (2000): "Analysis of Navigation Behavior in Web Sites Integrating Multiple Information Systems", *The VLDB Journal* 9:56-75.

7. Fu, Y., Kanwalpreet Sandhu and Ming-Yi Shih (1999): "A Generalization-Based Approach to Clustering of Web Usage Sessions", *International WEBKDD'99 Workshop* San Diego, Ca, USA, August 1999. Revised Papers.

8. Pitkow, J. and Krishna K. Bharat. Webviz (1994), "A tool for world-wide web access log analysis", *First International WWW Conference*.

9. Zanine, O.R., M.Xin, J. Han (1998), "Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs", *Prod. Advances in Digital Libraries Conf.* (1998) 19-29.

10. Cooley, R., Bamshad Mobasher,

Jaideep Srivastava (2000), "Web Mining: Information and Pattern Discovery on the World Wide Web", http://maya.cs.depaul.edu/~mobasher/webminer/survey/survey.html

11. Baumgarten, M., Alex Buchner, Sarabjot Anand, Maurice Mulvennan and John Huges (1999): "User-Driven Navigation Pattern Discovery from Internet Data", *International WEBKDD'99 Workshop* San Diego, Ca, USA, August 1999. Revised Papers.

12. Spiliopoulou, M., Carsten Pohle and Lukas Faulstich (1999): "Improving the Effectiveness of a Web Site with Web Usage Mining", *International WEBKDD'99 Workshop* San Diego, Ca, USA, August 1999. Revised Papers.

13. Pirolli, P., J. Pitkow, and R. Rao (1996): "Silk from a sow's ear: Extracting usable structures from the web", *Proc. of 1996 Conference on Human Factors in Computing Systems (CHI-96)*, Vancouver, British Columbia, Canada, 1996.

14. Chen, M.S., J.S. Park, and P.S. Yu (1996): "Data mining for path traversal patterns in a web environment", *In Proceedings of the 16th International Conference on Distributed Computing Systems*, pages 385--392, 1996.

15. Chan, P.K. (1999): "Constructing Web User Profiles: A Non-invasive Learning Approach", *International WEBKDD'99 Workshop* San Diego, Ca, USA, August 1999. Revised Papers.

16. Agrawal, R. and R. Srikant (1995), "Mining Sequential Pattern", Proc. *Int'l Conf. On Data Engineering*, 3-15.

17. Borge, J. and M. Levene (1999), "Data Mining of User Navigation Patterns", *International WEBKDD'99 Workshop* San Diego, Ca, USA, August 1999. Revised Papers.

18. Mobasher, B., H. Dai, T. Luo, Y. Sun and J. Zhu (1999), "Integrating Web Usage and Content Mining for More Effective Personalization", *International WEBKDD'99 Workshop* San Diego, Ca, USA, August 1999. Revised Papers.

19. Torrent System Inc. (2000), "Driving e-Commerce Profitability from Online and Offline Data", Torrent Systems White Paper.

———