Web Mining: A Key Enabler in E-Business

Nivedita Roy[1], Tapas Mahapaatra[2]

[1,2]ICFAI Business School, ICFAI University, IDPL Complex, Old Delhi Jaipur Road, Dundahera,
Gurgaon-122016 Haryana (India)
[1]nivedita@ibsdel.org, [2]tapas@ibsdel.org

Abstract- **The Web is becoming the apocryphal Vox Populi. There is an ever expanding amount of information "out there". Unfortunately, the morass of sources presents a formidable hurdle to effectively extract information from them. Web Mining consists of extracting knowledge from huge volumes of data available in WWW and other web sources, allowing better business decisions to be taken. Web Mining has been coined from Information Management & Retrieval (IM&R) and Artificial Intelligence (AI). In this paper, it has been shown how web mining is integrated to the knowledge discovery process, its potential applications, and techniques. Furthermore, an integrated architecture has been presented showing how web mining can contribute to e-business via the new technologies. Finally, some commercially-available architecture has been presented.**

Keywords: **Web Mining; E-business; Knowledge Discovery; Mined Technology; Mining Tools**

I. INTRODUCTION

We have truly arrived in the clichéd Information Age. There is an ever expanding amount of information "out there". Moreover, the evolution of the Internet into the Global Information Infrastructure, coupled with the immense popularity of the Web, has also enabled the ordinary citizen to become not just a consumer of information, but also its disseminator. The Web, then, is becoming the apocryphal *Vox Populi*. Given that there is this vast and ever growing amount of information, how does the average user quickly find what s/he is looking for -- a task in which the present day search engines don't seem to help much! The Web represents a key driving force for a large spectrum of applications in which users interact with or within companies, organizations, governmental agencies, and educational or collaborative environments. User preferences and expectations, together with usage, content, and structural patterns obtained from the Web, form the basis for intelligent, personalized, and business-optimal services. The development of techniques and architectures for more effective

integration and mining of structure, content, and usage data from different sources is likely to lead to the next generation of more useful and more intelligent Web applications, which can be focused to: (1) Extraction of knowledge from the Web, (2) Extraction of knowledge from the user's behavior.

Unfortunately, the morass of sources presents a formidable hurdle to effectively extract information from them. In recent years a growing number of machine learning and data mining methods have been applied to this problem. One possible approach is to personalize the web space -- create a system which responds to user queries by potentially aggregating information from several sources in a manner which is dependent on who the user is [1]. One of the most emerging and promising area in this regard is Web Mining (WM).

II. BUSINESS APPLICATION OF WEB MINING

The most dominant application area for WM is related to Internet based e-commerce (business-to-consumer) and Web-based customer relationship management (CRM) an integral part of E-business today. The e-commerce boom of the late 1990s and the plethora of Internet business start-ups gave a fillip to launch Web usage mining as the dominant Web mining application. WM provided the facility to track customer behavior for Web-based businesses more comprehensively that in any previous business model. The ability to rapidly adapt Web sites, product information and even pricing was available to those who could both collect and rapidly analyze consumer navigation patterns [2]. Web site navigation could be personalized to provide a unique experience for the consumer [3]. Consumer forums could be established to enable them to provide direct feedback on products and suppliers.

Web mining is the application of data mining or other information process techniques to WWW, to find useful patterns. Business organizations can take advantage of these patterns to access WWW more efficiently. Web mining, when looked upon in data mining terms, can be said to have three operations of interests - clustering, associations, and sequential analysis. As in most real-world problems, the clusters and associations in Web mining do not have crisp boundaries and often overlap considerably. In addition, bad exemplars and incomplete data can easily occur in the data set, due to a wide variety of reasons inherent to web browsing and logging. Thus, Web Mining and Personalization requires modeling of an unknown number of overlapping sets in the presence of significant noise and outliers. Moreover, the data sets in Web Mining are extremely large.

The business benefits that Web mining affords to digital service providers include personalization,

0-7803-8971-9/05/$20.00 ©2005 IEEE

collaborative filtering, enhanced customer support, product and service strategy definition, particle marketing and fraud detection. Today businesses talk more and more about e-business as they incorporate Internet technology into their core business processes. This new and modern business requires the key web mining process to be merged with the new technologies. This coupling consists of integrating on-line data into the databases to be mined, and allowing businesses to access the extracted knowledge through the web and internet/intranet technologies. These technologies can be also used for enhancing the performance of the web mining process. Indeed, the powerful computing available on the Internet can be exploited through the use of meta-computing technology, i.e. large scale (internet-based) parallel and distributed computing. The result is an integrated architecture that supports the data mining process and at the same time the exploitation of the discovered business knowledge.

III. WEB MINING CATEGORIES

Web Mining has been coined from Information Management & Retrieval (IM&R) and Artificial Intelligence (AI). Much has been taken from the lead application areas from data mining, such as fraud detection, customer behavior analysis and Customer Relationship Management (CRM). Web mining can be divided into two categories, keeping the structure mining at the base, we can have content mining and usage mining [4]. *Web structure mining* is a research field focused on using the analysis of the link structure of the web, and one of its purposes is to identify more preferable documents. *Web content mining* is an automatic process that extracts patterns from on-line information, such as the HTML files, images, or E-mails, and it already goes beyond only keyword extraction or some simple statistics of words and phrases in documents. *Web Usage Mining* facilitates to understand the user behavior and the web structure, thereby improving the design of this colossal collection of resources. All these ultimately are used for either the retrieval of existing information or discovering insights from the existing information. Figure 1 gives a schematic representation of different categories.

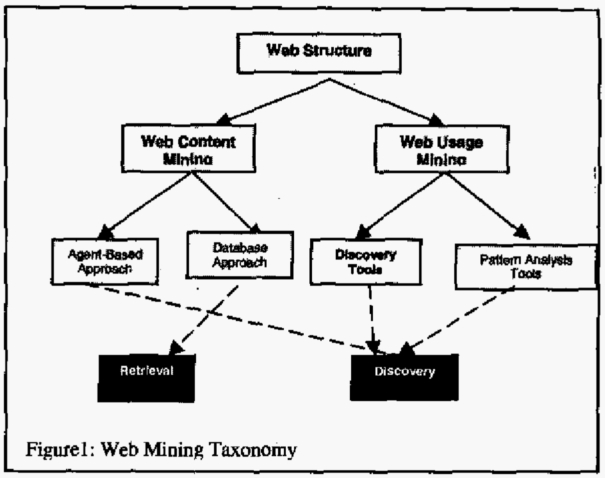

Figure1: Web Mining Taxonomy

IV. APPLICATIONS AND TECHNOLOGIES

A. Applications

Web Mining has a wide range of applications. The more common of them are: market analysis and a management (market basket analysis, cross-selling, market segmentation, etc.), risk analysis and management (forecasting, customer retention, etc.) and fraud detection and management (e-commerce, etc.). Business Intelligence, Customer Relationship Management, Bio-Informatics, and Knowledge Management are also candidate applications. In customer relationship management, Web mining is the integration of information gathered by traditional data mining methodologies and techniques with information gathered over the World Wide Web. It is used to understand customer behavior, evaluate the effectiveness of a particular Web site, and help quantify the success of a marketing campaign.

Web mining allows you to look for patterns in data through content mining, structure mining, and usage mining in different sectors. The application of Web usage mining techniques to data gathered from customers' online activity helps them to acquire business intelligence by providing high-level knowledge in the form of rules and patterns that describe consumer navigational and purchasing behavior [5]. Thus consumer profiles and market segmentation can be achieved giving these companies a competitive advantage. Even in the case of smaller organizations or individuals, the outcome of log analysis and Web usage mining can help them improve the performance of their systems, identify their Web site's visitors, and even customize their Web site making it more efficient and user-friendly. Content mining is used to examine data collected by search engines and Web spiders. Structure mining is used to examine data related to the structure of a particular Web site and usage mining is used to examine data related to a particular user's browser as well as data gathered by forms the user may have submitted during Web transactions.

B. *Mined Technology*

The information gathered through Web mining is evaluated by using traditional data mining parameters such as clustering and classification, association, and examination of sequential patterns. The technologies which can be used can well be summed up as under.

- Web Content and Structure mining
    - Integration of Web content, usage, and structure data for Web mining
    - Text mining techniques for generating meta-data
    - Classification and clustering of text and multimedia content
    - Detecting emerging trends or topics in text
    - Adaptive content management
    - Discovery and analysis of online communities and referral networks
- Web Usage Mining and Web Analytics
    - Web usage preprocessing
    - Novel techniques for discovery and analysis of Web usage patterns
    - Integrating semantics and domain knowledge in Web usage mining and analysis
    - Reliability and consistency of Web metrics
    - Integration of click stream data with back-end data and related metrics
    - Intelligent summarization/explanation of changes in Web usage metrics
- User Modeling and Profiling
    - Generating and updating profiles from implicit or explicit user preferences
    - Discovering misuse and fraud through outlier analysis
    - Personalized taxonomies or ontologies for navigation assistance
    - Cognitive models for Web navigation and e-commerce interactions
    - Incremental user modeling in dynamic environments
    - Permission marketing
- Applications
    - Recommendation and personalization systems
    - Intelligent Web services
    - Contextual information access and retrieval
    - Alert and Information filtering systems
    - Adaptive hypertext systems
    - Fraud and misuse detection, such as credit-card fraud and network intrusion detection
    - Web mining applications for business and competitive intelligence
    - Log analysis for security applications

C. *Mining Approaches*

Web mining is an inter-disciplinary and –functional emerging field with the potential in all the areas. Lee (2003/4) identified two major categories of areas: retrieval and discovery [6]. Most commonly web mining approaches are as: Intelligent Search Agents, Information Filtering/Categorization, Personalized Web Agents, Multilevel Databases, Web Query Systems, Pattern Analysis Tools, Pattern Discovery Web Transactions, Data Cleaning, Transaction Identification, Path Analysis, Association Rules, Sequential Patterns, Clustering and Classification, Analysis of discovered patterns, Visualization techniques, OLAP techniques, Usability analysis, etc.

Some of the above-mentioned approaches can prove to be a promising tool to address ineffective search engines, those produce incomplete indexing, retrieval of irrelevant information or unverified reliability of retrieved information. It is essential to have a system that helps the user find relevant and reliable information easily and quickly on the Web. This knowledge discovery is achieved through the following stages: (1) Web Sources Stage (XML files, databases), (2) Web Warehousing Stage (using OLAP, etc.) This is done by Web Analyst, (3) Web Mining Stage. This is done for information discovery and done by the Web Analyst, (4) Data Presentation Stage. This is done by visualization techniques and done by the business analyst, (5) Final Decision-making Stage, done by the decision-makers, usually the top management (Figure 2).

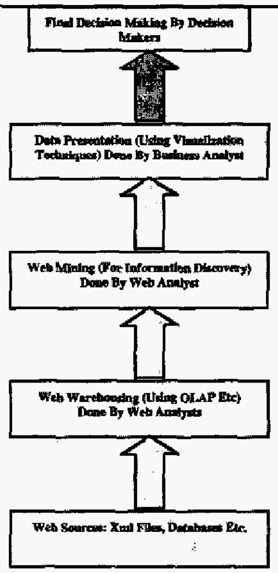
Final Decision Making By Decision Makers
Data Presentation (Using Visualization Techniques) Done By Business Analyst
Web Mining (For Information Discovery) Done By Web Analyst
Web Warehousing (Using OLAP Etc) Done By Web Analysts
Web Sources: Xml Files, Databases Etc.

Figure 2: Knowledge Discovery using Web Mining

D. *A Case Study*

A number of tools are coming up for web mining in various areas. The emerging tools for pattern discovery techniques from AI, data mining, psychology, and information theory, to mine for knowledge from collected data. For example, the WEBMINER system introduces a general architecture for Web usage mining [7].

WEBMINER automatically discovers association rules and sequential patterns from server access logs. Algorithms have been introduced for finding maximal forward references and large reference sequences. These can, in turn be used to perform various types of user traversal path analysis such as identifying the most traversed paths through a Web locality. Once access patterns have been discovered, analysts need the appropriate tools and techniques to understand, visualize, and interpret these patterns. For this purpose, OLAP techniques such as data cubes can be used for the purpose of simplifying the analysis of usage statistics from server access logs. The WEBMINER system proposes an SQL-like query mechanism for querying the discovered knowledge (in the form of association rules and sequential patterns). Features, generally, supported by the WEBMINER are:

- Scan multiple sites for desired content within one session,
- Dynamically add further sites for processing while downloading,
- Add multiple sites without leaving your favorite browser (copies URL from browser),
- Unlimited authentication entries with global memory for storing passwords,
- Download by fully customizable categories such as video and audio,
- Intelligent scan to interact with servers using java for web site navigation,
- Filters (exclusive and inclusive) to control download content
- Options for dealing with duplicate files found

Once user transactions or sessions have been identified, there are several kinds of access pattern that can be performed depending on the needs of the analyst. There are many different types of graphs that can be formed for performing path analysis, since a graph represents some relation defined on Web pages.

Association rule discovery techniques are generally applied to databases of transactions where each transaction consists of a set of items [8]. In such a framework the problem is to discover all associations and correlations among data where the presence of one set of items in a transaction implies (with a certain degree of confidence) the presence of other items. In the context of Web Mining, this problem amounts to discovering the correlations among references to various files available on the server by a given client. Each transaction is comprised of a set of URLs accessed by a client in one visit to the server.

Discovery of such rules for organizations engaged in electronic commerce can help in the development of effective marketing strategies. But, in addition, association rules discovered from WWW access logs can give an indication of how to best organize the organization's Web space.

V. METHODOLOGY FOR WEB MINING

The generic architecture of the Web Mining can be divided into four steps, like Preprocessing, Transaction Identification, Data Integration, and Data Mining. First of all, clustering algorithms are used for grouping log entries into transactions. Then integration of data from various sources such as user registration databases with access log data needs to be done. Use of association rule, temporal sequence, and classification rule discovery algorithms can help in simplifying the pre-processing phase of Web mining process. Some of the information made available are: domain name, IP address of the request, user ID, timestamp, method, server status code, parameters of the script, size of the data sent, browser type, referring page, etc. These information help in summarizing the information like frequency of individual actions by user/domain/session, group actions into activities, frequency of errors and others. Pattern analysis between different users, trend analysis like users behavior change over time, network traffic change, etc. and various other types of analysis can be carried out for providing an insight on the web usage. Data mining is used to find answers to the questions like:

- In what context are the components or features being used?
- What are the sequences of activities performed?
- Are there any behavior patterns across different users?
- Does the behavior change over a period of time and how?

Figure 3 shows a diagrammatic representation of all the steps involved in Web Mining.

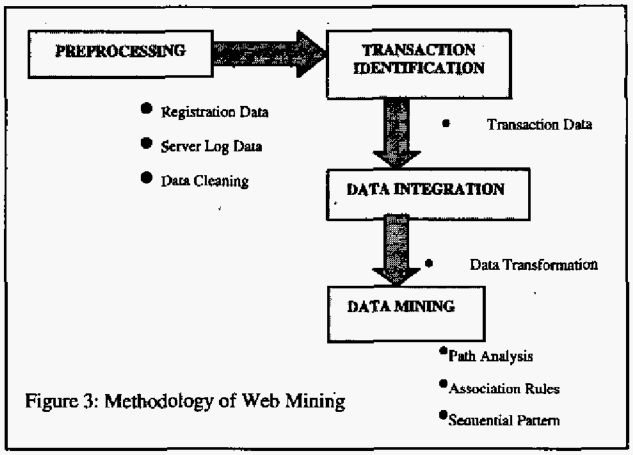

Figure 3: Methodology of Web Mining

VI. MINING TOOLS

Although web mining is emerging, many tools have already been developed. A review of the main commercially-available integrated WM architectures has been presented in the following paragraphs.

- QL2 Software is the inventor of WebQL, the Web equivalent of SQL. The value of WebQL is its ability to interrogate or look inside particular Web pages for answers to queries in the same way that SQL is used to look inside databases and respond to queries [9].

- Touchgraph appears to be an experimental metasearch engine product that can map results from Google or Amazon [10]. Google Touchgraph accepts a starting URL and then builds a map of related sites as determined by Google. The Amazon Touchgraph is perhaps more useful as it can provide a full map of potential reading around a nominated topic.
- Megaputer has recently released Web Mining products (Web Analyst and X-Sell Analyst) to support typical online retail sites in maximizing the returns from their customer base [11]. While these products are relatively new, they are built on the base of a fairly mature data mining package called Poly Analyst. Poly Analyst integrates a collection of data mining and text mining techniques into a single tool. It is built on the DCOM-based architecture, which enables the data mining solution to be easily integrated into larger applications or used as a stand-alone tool.
- Visible Path is an example of a growing number of products aimed at mining relationships and social networks [12]. The product's focus is on large corporate environments, where understanding relationship networks can enhance the sales function in particular but could also enhance support functions. The software works by mining a company's messaging sources like e-mail, instant messaging, calendars, meeting rooms and directories to develop a database of weighted relationship network information.
- The IBM Almaden research center has developed WebFoundation which is a large text analytic solution that collects massive amounts of unstructured and semi-structured text and converts it to XML tagged information, prior to mining for patterns and trends [13]. Text sources include Internet data, Weblogs, bulletin boards, enterprise data, licensed content, newspapers, magazines and trade journals.

VII. CONCLUSION

Through this paper, we have tried to highlight how to gain business advantage in e-Business, by using the WWW, which is considered to be the largest pool of information resources. Before starting the web mining process the knowledge to be extracted must be clearly identified. This technology helps in gaining meaningful insights related to the day-to-day business activities by using the useful information made accessible through the web. It not only enables discovery of relevant information from mounds of data on the WWW, but it also monitors and predicts user visit habits. Using various web mining tools and techniques discussed in the paper, helps in digging out layers of information about the markets, and the data collected from Web sites present enormous potential for direct marketing. Marketers can fine-tune their selling strategies by building customer or prospect profiles and using these to identify the segments upon which marketing activities are focused. Various mining approaches can well be used catering to the specific business requirements. In a nutshell, Web Mining can be summed up as a very viable technology, application and a product suite meant for discovering knowledge pertaining to the routine business activities and can prove to be a very important approach for gaining competitive advantage.

REFERENCES

[1] Eirinaki, M, and Vazirgiannis, 2003, *Web Mining for Web Personalization*, ACM Trans. On Internet Technology, vo. 3, No. 1, Feb., pp. 1-27

[2] Iyer, G., A. Miyazaki, et al, 2002, *Linking Web-based segmentation to pricing tactics*, Journal of Product & Brand Management, 11(5): pp. 288 -302

[3] Mulvenna, M.D., Anand, S.S., and Buchner, A.G., 2000, *Personalization on the net using web mining*, Commun., ACM 43.8 (August), pp. 123-125.

[4] Cooley,R., B. Mobasher, et al, (1997), *Web Mining: Information and Pattern Discovery on the World Wide Web*, Proc. Of the 9th IEEE Int. Conf. on Tools with Artificial Intelligence (ICTAI'97), Dept. of Computer Science, Univ. of Minnesota.

[5] Buchner, A. and Mulvenna, M.D., 1998, *Discovering Internet Marketing Intelligence through online analytical web usage mining*, SIGMOD Rec. 27.4.54-61

[6] Lee, L.L., 2003/4, *Web Mining*, Leading Edge Forum, Australia Group

[7] http://tribolic.com/webminer

[8] Bamshad Mobasher, 1997. "Association Rules," at http:// maya.cs.depaul.edu /~mobasher /webminer

[9] www.ql2.com

[10] www.touchgraph.com

[11] www.megaputer.com

[12] www.visiblepath.com

[13] www.almaden.ibm.com/webfountain/