A REVIEW OF WEB MINING RESEARCH

Ms. Poonam Sawant*

Dr. R.V. Kulkarni**

Abstract:

Today, with the huge amount of information available online, the World Wide Web is a fertile area for data mining research. Nowadays Web users are facing the problems of information overload and drowning due to the significant and rapid growth in the amount of information and the number of users. As a result, how to provide Web users with more exactly needed information is becoming a critical issue in web-based information retrieval and Web applications. The Web mining research is at the cross road of research from several research communities such as database, information retrieval, AI, Machine learning and natural language processing etc. In this paper authors discuss the number of researches carried out in the area of web mining. This paper describes the importance of web mining process to improve the performance of Web information retrieval . They also aim to address how the adoption of Web Mining is essential for banking organizations to identify, preserve and disseminate best context of e-service practices which is useful to tell, how to provide customer with more exactly needed and accurate on line information to achieve business goal.

Keywords: Web Mining, Web Mining Categories, Information Retrieval and Information Extraction.

* Research Student, SIBER, Kolhapur.

** Professor and Head, SIBER, Kolhapur.

1. Introduction:

With the explosive growth of information sources available on the World Wide Web, it has become increasingly necessary for users to utilize automated tools in find the desired information resources, and to track and analyze their usage patterns. These factors give rise to the necessity of creating server side and client side intelligent systems that can effectively mine for knowledge. Web mining is the application of data mining techniques to extract knowledge from web data, including web documents, hyperlinks between documents, usage logs of web sites, etc. Web mining is the search for relevant information from the World Wide Web. As all users of Web search engines are aware, accurate answers are not always at the top of the result list. Web mining is the extraction of interesting and potentially useful patterns and implicit information from artifacts or activity related to the World.

It is also defined as the Information Retrieval(IR) and Information Extraction(IE) from the world Wide Web. Information Retrieval Is the area of study concerned with searching for documents, for information within documents, and for metadata about documents, as well as that of searching structured storage, relational databases, and the World Wide Web. Information Extraction is a type of information retrieval whose goal is to automatically extract structured information from unstructured and/or semi-structured machine-readable documents. Wide Web. There are three categories of web mining: Web Content Mining, Web Structure Mining, and Web Usage Mining. Web content mining is the process of extracting knowledge from the content of documents or their descriptions. It is an automatic process that goes beyond keyword extraction. Web structure mining is the process of inferring knowledge from the World Wide Web organization and links between references and referents in the Web. And Web usage mining is the process of extracting interesting patterns in web access logs. Web servers record and accumulate data about user interactions whenever requests for resources are received.

2. Literature Review:

Recently, many research projects are dealing with Web usage mining and Web personalization areas. Most of the efforts focus on extracting useful patterns and rules using data mining techniques in order to understand the users' navigational behavior, so that decisions concerning

site restructuring or modification can then be made by humans. In several cases, a recommendation engine helps the user navigate through a site. Some of the more advanced systems provide much more functionality, introducing the notion of adaptive Web sites and providing means of dynamically changing a site's structure. All research efforts combine more than one of the aforementioned methods in Web personalization, namely, user profiling, Web usage mining techniques, content management and publishing mechanisms.

One of the earliest attempts to take advantage of the information that can be gained through exploring a visitor's navigation through a Web site resulted in Letizia [Lieberman 1995], a client-site agent that monitors the user's browsing behavior and searches for potentially interesting pages for recommendations. The agent looks ahead at the neighboring pages using a best-first search augmented by heuristics inferring user interest, in as much as they're derived from the user's navigational behavior, and offers suggestions.

An approach for automatically classifying a Web site's visitors according to their access patterns is presented in the work of Yan et al. [1996]. The model they propose consists of two modules; an offline module that performs cluster analysis on the Web logs and an online module aiming at dynamic link generation. Every user is assigned to a single cluster based on his current traversal patterns. The authors have implemented the offline module (Analog) and have given a brief description of the way the online module should function.

One of the most popular systems from the early days of Web usage mining is WebWatcher [Joachims et al. 1997]. The idea is to create a tour guide agent that provides navigation hints to the user through a given Web collection, based on its knowledge of the user's interests, the location and relevance of various items in the location, as well as the way in which other users have interacted with the collection in the past. The system starts by profiling the user, acquiring information about her interests. Each time the user requests a page, this information is routed through a proxy server in order to easily track the user session across the Web site and any links believed to be of interest for the user are highlighted. Its strategy for giving advice is learned from feedback from earlier tours. A similar system is the Personal WebWatcher [Mladenic 1999], which is structured to specialize for a particular user, modeling his interests. It solely records the addresses of pages requested by the user and highlights interesting hyperlinks

without involving the user in its learning process, asking for keywords or opinions about pages as WebWatcher does.

Chen et al. [1996] introduce the "maximal forward reference" concept in order to characterize user episodes for the mining of traversal patterns. Their work is based on statistically dominant paths and association rules discovery, and a maximal forward reference is defined as the sequence of pages requested by a user up to the last page before backtracking. The SpeedTracer project [Wu et al. 1998] is built on the work proposed by Chen et al. [1996]. SpeedTracer uses the referrer page and the URL of the requested page as a traversal step and reconstructs the user traversal paths for session identification. Each identified user session is mapped into a transaction and then data mining techniques are applied in order to discover the most frequent user traversal paths and the most frequently visited groups of pages.

A different approach is adopted by Zaiane et al. [1998]. The authors combine the OLAP and data mining techniques and a multidimensional data cube, to extract interactively implicit knowledge. Their WebLogMiner system after filtering the data contained in the Web log, transforms them into a relational database. In the next phase a data cube is built, each dimension representing a field with all possible values described by attributes. OLAP technology is then used in combination with data mining techniques for prediction, classification, and time-series analysis of Web log data. Huang et al. [2001] also propose the use of a cube model that explicitly identifies Web access sessions, maintains the order of the session's components and uses multiple attributes to describe the Web pages visited.

Buchner and Mulvenna [1998] present a knowledge discovery process in order to discover marketing intelligence from Web data. They propose an environment that combines existing online analytical mining, as well as Web usage mining approaches and incorporates marketing expertise. For this purpose, a generic Web log data hypercube is defined. In a more recent work Buchner et al. [1999] introduce the data mining algorithm MiDAS for discovering sequential patterns from Web log files, in order to perceive behavioral marketing intelligence. In this work, domain knowledge is described as flexible navigation templates that specify navigational behavior, as network structures for the capture of Web site topologies, as well as concept hierarchies and syntactic constraints.

Spiliopoulou et al. [Spiliopoulou and Faulstich 1998; Spiliopoulou et al. 1999; Spiliopoulou 2000] have designed MINT, another mining language for the implementation of WUM, a sequence mining system for the specification, discovery, and visualization of interesting navigation patterns. The Web log is preprocessed and an "aggregate materialized view" of the Web log is stored. In the data preparation phase, except for log data filtering and completion, user sessions are identified using timeout mechanisms. The path each user follows is called a "trail". Because many users access the same pages in the same order (creating similar trails), an "aggregate tree" is constructed by merging trails with the same prefix. This tree is called an "aggregated log" and navigation patterns of interest can be extracted using MINT. This language supports the specification of criteria of statistical, structural, and textual features.

Borges and Levene [1999] model the set of user navigation sessions as a hypertext probabilistic grammar whose higher probability generated strings correspond to the user's preferred trails. Shahabi et al. [1997] propose the use of a client-side agent that captures the client's behavior creating a profile. Their system then creates clusters of users with similar interests.

Masseglia et al. [1999] apply data mining techniques such as association rules and sequential pattern discovery on Web log files and then use them to customize the server hypertext organization dynamically. They regard Web usage mining as a two-phase process, consisting of the preprocessing phase where all irrelevant data are removed and log file entries are clustered based on time considerations, and the Web mining phase where data mining techniques are applied. The prototype system, WebTool, also provides a visual query language in order to improve the mining process. A generator of dynamic links uses the rules generated from sequential patterns or association rules, and each time the navigation pattern of a visitor matches a rule, the hypertext organization is dynamically modified.

Cooley et al. [1999; Srivastava et al. 2000] define Web usage mining as a three-phase process, consisting of preprocessing, pattern discovery, and pattern analysis. Their prototype system, WebSIFT, first performs intelligent cleansing and preprocessing for identifying users, server sessions, and inferring cached page references through the use of the referrer field, and also performs content and structure preprocessing [Cooley et al. 1999]. Pattern discovery is accomplished through the use of general statistic algorithms and data mining techniques such as association rules, sequential pattern analysis, clustering, and classification. The results are then

analyzed through a simple knowledge query mechanism, a visualization tool, or the information filter, that makes use of the preprocessed content and structure information to automatically filter the results of the knowledge discovery algorithms.

The most advanced system is the WebPersonalizer, proposed by Mobasher et al. [1999, 2000]. WebPersonalizer provides a framework for mining Web log files to discover knowledge for the provision of recommendations to current users based on their browsing similarities to previous users. It relies solely on anonymous usage data provided by logs and the hypertext structure of a site. After data gathering and preprocessing (converting the usage, content, and structure information contained in the various data sources into various data abstractions), data mining techniques such as association rules, sequential pattern discovery, clustering, and classification are applied, in order to discover interesting usage patterns. The results are then used for the creation of aggregated usage profiles, in order to create decision rules. The recommendation engine matches each user's activity against these profiles and provides him with a list of recommended hypertext links.

This framework has been recently extended [Mobasher et al. 2000] to incorporate content profiles into the recommendation process as a way to enhance the effectiveness of personalization actions. Usage and content profiles are represented as weighted collections of page view records. The content profiles represent different ways in which pages with partly similar content can be grouped. The overall goal is to create a uniform representation for both content and usage profiles in order to integrate them more easily. The system is divided into two modules; the offline, which is comprised of data preparation and specific Web mining tasks, and the online component, which is a real-time recommendation engine.

According to Raymond Kosala and Hendrik Blockeel (2000) Web mining research is currently emerging in many aspects of e-services, aiming at improving online transactions and making them more transparent and effective. Web mining techniques could be used to solve the problems occurs in finding the relevant information, use of information available on the banks web sites in proper way, personalization of the information etc. So we can utilize the web mining techniques for decision making to fulfill the customers' expectations and expand the business.

In a recent work [Masseglia et al. 2000], the problem of incremental Web usage mining is addressed. Using the ISEWUM method, they handle the problem of mining user patterns when

new transactions are added to the Web log file by only considering user patterns obtained by an earlier mining.

Joshi et al. [2000; Krishnapuram et al. 2001; Nasraoui et al. 2000] introduce the notion of uncertainty in Web usage mining, discovering clusters of user session profiles using robust fuzzy algorithms. In their approach, a user or a page can be assigned to more than one cluster. After preprocessing the log data, they create a dissimilarity matrix that is used by the fuzzy algorithms presented in order to cluster typical user sessions. To achieve this, they introduce a similarity measure that takes into account both the individual URLs in a Web session, as well as the structure of the site.

Cooley et al. [1999; Srivastava et al. 2000] define Web usage mining as a three-phase process, consisting of preprocessing, pattern discovery, and pattern analysis. Their prototype system, WebSIFT, first performs intelligent cleansing and preprocessing for identifying users, server sessions, and inferring cached page references through the use of the referrer field, and also performs content and structure preprocessing [Cooley et al. 1999]. Pattern discovery is accomplished through the use of general statistic algorithms and data mining techniques such as association rules, sequential pattern analysis, clustering, and classification. The results are then analyzed through a simple knowledge query mechanism, a visualization tool, or the information filter, that makes use of the preprocessed content and structure information to automatically filter the results of the knowledge discovery algorithms.

Perkowitz and Etzioni [1998, 1999, 2000] were the first to define the notion of adaptive Web sites as sites that semiautomatically improve their organization and presentation by learning from visitor access patterns [Perkowitz and Etzioni 1997]. The system they propose semi automatically modifies a Web site, allowing only nondestructive transformations. Therefore, nothing is deleted or altered; instead, new index pages containing collections of links to related but currently unlinked pages are added to the Web site. The authors propose PageGather, an algorithm that uses a clustering methodology to discover Web pages visited together and to place them in the same group.

In a recent work [Masseglia et al. 2000], the problem of incremental Web usage mining is addressed. Using the ISEWUM method, they handle the problem of mining user patterns when

new transactions are added to the Web log file by only considering user patterns obtained by an earlier mining.

Berendt [2000, 2001] has implemented STRATDYN, an add-on module that extends WUM's capabilities by identifying the differences between navigation patterns and exploiting the site's semantics in the visualization of the results. In this approach, concept hierarchies are used as the basic method of grouping Web pages together. The accessed pages or paths are abstracted, because Web pages are treated as instances of a higher-level concept, based on page content, or by the kind of service requested. An "interval-based coarsening" technique is used in order to mine Web usage at different levels of abstraction using basic and coarsened stratograms for the visualization of the results.

Coenen et al. [2000] propose a framework for self-adaptive Web sites, taking into account the site structure except for the site usage. The authors underline the distinction between strategic changes, referring to the adaptations that have important influence on the original site structure, and tactical changes, referring to the adaptations that leave the site structure unaffected. The proposed approach is based on the fact that the methods used in Web usage mining produce recommendations including links that don't exist in the original site structure, resulting in the violation of the beliefs of the site designer and the possibility of making the visitor get lost following conceptual but not real links. Therefore, they suggest that any strategic adaptations based on the discovery of frequent item sets, sequences, and clusters should be made offline and the site structure should be revised. On the other hand, as far as the tactical adaptations are concerned, an algorithm for making online recommendations leaving the site structure unaffected is proposed.

In a more recent work [Perkowitz and Etzioni 2000], they move from the statistical cluster-mining algorithm PageGather to IndexFinder, which fuses statistical and logical information to synthesize index pages. In this latter work, they formalize the problem of index page synthesis as a conceptual clustering problem and try to discover coherent and cohesive link sets that can be represented to a human Webmaster as candidate index pages. The difference is that information is also derived from the site's structure and content. Therefore, IndexFinder combines the statistical patterns gleaned from the log file with logical descriptions of the contents of each Web page in order to create index pages.

Cingil et al. [2000] describe an architecture that provides a broader view of personalization, through the use of various W3C standards. They describe how standards such as XML, RDF, and P3P can be used to create personalization applications. In this architecture, a log of the user's navigation history is created as a "user agent" at the client site gathers clickstream information about the user. This information is kept in an XML file, creating a user profile that reflects the user's interests and preferences. Privacy of the user is preserved through P3P. On the server side statistical modeling is run on user profiles to match up visitors that seem to have similar interests and preferences so that the most likely content or products can be recommended to a user based on these similarities. The user profile is exploited by the user agent to discover resources on the Internet that may be of interest to the user as well as obtaining personalized information from the resources. When the metadata of the resources are expressed in RDF, it will be a lot easier for agents to discover the resources on the Web that match the user profiles. Until then, metadata tags of HTML are used in the proposed system.

According to Werner and Böttcher (2005 International Conference on Data Mining (DMIN '05) and its session on Web mining), text document, such as a Web page, contains information about itself that goes beyond the text itself. Content is emphasized in various ways to bring out the author's tone and intentions. When evaluating the meaning of a document, whether it is a Web page or other text document, the search engine must consider what text was emphasized and how it was emphasized. They wrote an article, "Supporting Text Retrieval by Typographical Term Weighting," which provides approaches for improving the classification of a document by considering the typographical information in the document's format. Web searches are improved when Web pages that are similar in content can be identified. This allows the presentation of search results, which more accurately reflect the needs of the searcher. This similarity can be accomplished by classifying the Web pages into known classes, or clustering them based solely on the content of the page.

Zhongmei Yao and Ben Choi(2005 International Conference on Data Mining (DMIN '05) and its session on Web mining) have developed a bi-directional hierarchical clustering algorithm that clusters similar Web pages in an unsupervised manner. This algorithm, presented in "Clustering Web Pages into Hierarchical Categories," proves to be better than established methods and provides visualization at different granularities of similarity.

The user's ability to express their information need is a critical element in any search. It is also the most difficult to control. Search engines are needed that can compensate for the user not knowing exactly for what they are searching. This can be accomplished by relevance feedback. Picariello and Rinaldi (2005 International Conference on Data Mining (DMIN '05) and its session on Web mining)present an ontology-based relevance feedback technique for ranking search results in "User Relevance Feedback in Semantic Information Retrieval." They consider the domain of the search, develop a dynamic semantic network, and calculate a syntactic-semantic grade for each term and a semantic grade for each document. These are then used to rank the documents, improving recall and precision. As the Web is accessed more and more by mobile devices it becomes more important to efficiently navigate from Web page to Web page. "Improving Mobile Web Navigation Using N-grams Prediction Models" presents this ability.

Using Web usage mining, Fu, Paul and Shetty (2005 International Conference on Data Mining (DMIN '05) and its session on Web mining)are able to discover shortcuts to Web pages based on previous pages viewed. These shortcuts can be dynamically provided to users when navigating from a Web page and looking for a desired destination Web page. Their MINCOST algorithm is especially effective in reducing the time spent by mobile users as they search the Web.

While search engines rate Web pages to return an ordered set to a searcher, Meng, Xing and Clark (2005 International Conference on Data Mining (DMIN '05) and its session on Web mining),look at rating the search engine performance. The traditional ranking metrics for information retrieval are dependent on knowledge of the content of the information repository. This knowledge is never available with the World Wide Web. In "An Empirical Performance Measurement of Microsoft's Search Engine and its Comparison with other Major Search Engines," the *Rank- Power* metric is used, along with response time, to compare the performance of various search engines.

The transactions offered by different banks(Mr. V. S. Solankhi and Mr. Abhishek Singh(10 Nov 2009)) are continuously changing and are being improved because of some banks wants to attain competitive advantage with other banks. The banking industry should always adapt to the new technology today and basically make the necessary adjustments to gain competitive advantage with other competing banks. With these developments, customers are able to enjoy the many conveniences and lower costs that are offered by the said innovations. But no

doubt, the technological growth has considerably affected the profile of Bank risks and financial institution formation more generally. In order for users/ customers to use their banks online services like, paying bills, attaining information about accounts and loans, they need to have a proper knowledge and relevant information. We can manage the relevant information through web mining.

According to Raymond Kosala and Hendrik Blockeel Web mining research is currently emerging in many aspects of e-services, aiming at improving online transactions and making them more transparent and effective. Web mining techniques could be used to solve the problems occurs in finding the relevant information, use of information available on the banks web sites in proper way, personalization of the information etc. So we can utilize the web mining techniques for decision making to fulfill the customers' expectations and expand the business.

According to Usha P.M(2010) customer relationship management is one of the major task in banking sector and it is an important applications of Web mining. As the various banks are using wide range of e-services a website should be designed to entice the customers. Web Mining analyses visitor's behavior and makes predictions on their future interaction. This can be exploited to improve website performance and to recommend policies or links based on user's behavior. Visitors entering the site exhibits different behavior. They might just surf through or the process might end up in a purchase. For understanding customer behavior and thus improve the performance of your web site, certain standards should be used. Web metrics provide a method to evaluate the performance.

According to Usha P.M(2010) now almost all the banks have ventured into this area. Enormous amount of data gets stored through banking transactions. Success factor is the amount of valuable knowledge that is extracted from this data store. Customer profile can be generated. This helps the bank executives in identifying the appropriate customer for certain category of products and the risk in allotting loan facilities. Credit card usage patterns can be identified and special offers can be provided. Defaulters of payment can be identified easily. Banks like ICICI bank and HSBC bank identify the customers for certain offers like home equity loan using web mining.

(Internationally Indexed Journal ■ www.scholarshub.net ■ Vol–II , Issue -3 March 2011) As the Web and its usage continues to grow, so grows the opportunity to analyze Web data and extract

all manner of useful knowledge from it. Various business organizations are practicing it to achieve the business goal.

3. Future Directions:

The past five years have seen the emergence of Web mining as a rapidly growing area, due to the efforts of the research community. Web mining is becoming the tool for success for those who adopt electronic means of operation (e.g. e-learning, e-banking, e-governance etc.) for conducting their business. Internet banking was offered as a "value addition"(e-services) for most customers. Though almost all the banks in private sector and public sector have ventured into this area and enormous amount of data gets stored through banking transactions, few banks like ICICI bank and HSBC bank identify the customers for certain offers like home equity loan using web mining. The authors suggested that success factor is the amount of valuable knowledge that is extracted from this data store. In e- banking the web mining research is not carried out deeply. In future, the research will expand to examine the behavior among the customer's of private and public sector banks and the relationship between the e-service quality dimensions and customer satisfaction. The research is also aims to find out the factors affecting on the Information Retrieval. The research will expand to build customer behavioral model to provide the accurate and relevant information to the customer every time. The research will also expand to generate customer profile to help the bank executives in identifying the appropriate customer for certain category of e- services like Automated Teller Machine, Credit Card, Debit Card, Smart Card, Electronic Fund Transfer System, Cheques Transaction Payment System, Mobile Banking, Internet Banking etc. in both public and private bank sector. The overall objective of this research is to targeting the right customer and ensuring excellent service to these customers to achieve the business goal.

4. Conclusion:

As we go through an inevitable phase of 'irrational despair' following a phase of 'irrational exuberance' about the commercial potential of the Web, the adoption and usage of the Web continues to grow unabated. This trend is likely to continue as Web services continue to flourish.

As the Web and its usage grows, it will continue to generate evermore content, structure, and usage data, and the value of Web mining will keep increasing, so grows the opportunity to analyze Web data and extract all manner of useful knowledge from it. Web mining is becoming the tool for success for those who adopt electronic means of operation for conducting their business. Web mining can contribute to a large extent in gaining a competitive advantage in your business. Your business goals should be well understood.

In this paper authors have briefly described the various researches carried out in the area of web mining. This paper also states that how the adoption of Web Mining is essential for banking organizations to identify, preserve and disseminate best context of e-service practices which is useful to tell, how to provide customer with more exactly needed and accurate on line information to achieve business goal.

5. REFERENCES:

- Web mining – Concepts, Applications and Research Directions by Jaydeep Shrivatava, Prasanna Deshikan, Vilpin Kumar.
- Friedman, T. J. (2005). *The world is flat:A brief history of the twenty-first century.* New York: Farrar, Straus and Giroux.
- Web Mining - An Important Tool for Molding Business by Usha P.M. Faculty Member(IT) ICFAI National College Kozhikode
- http://www.ebusinessnews.info/?action=read&article=677
- http://www.webanalyticsassociation.org/en/articles/printview.asp?467
- http://www.wikipedia.org
- Web Mining for Web Personalization by MAGDALINI EIRINAKI and MICHALIS VAZIRGIANNIS
- Cohen, A., & Nachmias, R. (2006). A quantitative cost effectiveness model for Web-supported academic instruction. *The Internet and Higher Education, 9*(2), 81-90.
- Etzioni, O. (1996). The World Wide Web: quagmire or gold mine? *Communications of ACM, 39*(11), 65-68.

- Kosala, R., & Blockeel, H. (2000). Web mining research: A survey. *SIGKDD Explorations*, *2*(1).
- www.shef.ac.uk/faculty/arts-and-humanities/news/2009/old-bailey -
- nlp.shef.ac.uk/GOTag/papers_pdf/ismb_demo_v2.ppt - 2006-03-22
- tripod.shef.ac.uk/publications/tarwochetal08.pdf - 2009-02-24