A Review On Web Mining

Mr. Dushyant Rathod

Lecturer, Information Technology, Gandhinagar Institute of Technology,Gandhinagar,
dushyant.rathod@git.org.in

***Abstract*— Data mining is one of the most applicable area of research in computer applications among the various types of data mining . This paper is going to focus on web mining. This is the review paper which shows deep and intense study of various techniques available for web minings. Web mining - i.e. the application of data mining techniques to extract knowledge from Web content, structure, and usage - is the collection of technologies to fulfill this potential. Above definition of web mining is explored in this paper.**

***Index Terms*—Web Mining , Web Structure Mining, Web Content Mining, Web Usage Mining.**

I. INTRODUCTION

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. Two different approaches were taken in initially defining Web mining. First was a 'process-centric view', which defined Web mining as a sequence of tasks. Second was a 'data-centric view', which defined Web mining in terms of the types of Web data that was being used in the mining process . The second definition has become more acceptable, as is evident from the approach adopted in most recent papers that have addressed the issue. In this paper we follow the data-centric view, and refine the definition of Web mining as, **Web mining** is the application of data mining techniques to extract knowledge from Web data, where **at least one of structure (hyperlink) or usage (Web log) data is used in the mining process** (with or without other types of Web data)[1].

The attention paid to Web mining, in research, software industry, and Web-based organizations, has led to the accumulation of a lot of experiences. It is our attempt in this paper to capture them in a systematic manner, and identify directions for future research.[2]

II. WEB MINING

Web mining is the Data Mining technique that automatically discovers or extracts the information from web documents. It consists of following tasks[4]:

1. *Resource finding:* It involves the task of retrieving intended web documents. It is the process by which we extract the data either from online or offline text resources available on web.

2. *Information selection and pre-processing:* It involves the automatic selection and pre processing of specific information from retrieved web resources. This process transforms the original retrieved data into information. The transformation could be renewal of stop words, stemming or it may be aimed for obtaining the desired representation such as finding phrases in training corpus.

3. *Generalization:* It automatically discovers general patterns at individual web sites as well as across multiple sites. Data Mining techniques and machine learning are used in generalization

4. *Analysis:* It involves the validation and interpretation of the mined patterns. It plays an important role in pattern mining. A human plays an important role in information on knowledge discovery process on web[3].

III. WEB MINING TAXONOMY

Web Mining can be broadly divided into three distinct categories, according to the kinds of data to be mined:

A. *Web Content Mining*

Web content mining is the process of extracting useful information from the contents of web documents. Content data is the collection of facts a web page is designed to contain. It may consist of text, images, audio, video, or structured records such as lists and tables. Application of text mining to web content has been the most widely researched. Issues addressed in text mining include topic discovery and tracking, extracting association patterns, clustering of web documents and classification of web pages. Research activities on this topic have drawn heavily on techniques developed in other disciplines such as Information Retrieval (IR) and Natural Language Processing (NLP). While there exists a significant body of work in extracting knowledge from images in the fields of image processing and computer vision, the application of these techniques to web content mining has been limited.

B. *Web Structure Mining*

The structure of a typical web graph consists of web pages as nodes, and hyperlinks as edges connecting related pages. Web structure mining is the process of discovering structure information from the web. This can be further divided into two kinds based on the kind of structure information used.

Hyperlinks

A hyperlink is a structural unit that connects a location in a web page to a different location, either within the same web page or on a different web page. A hyperlink that connects to a different part of the same page is called an *intra-document hyperlink*, and a hyperlink that connects two different pages is called an *inter-document hyperlink*. There has been a significant body of work on hyperlink analysis, of which Desikan, Srivastava, Kumar, and Tan (2002) provide an up-to-date survey.

Document Structure

In addition, the content within a Web page can also be organized in a treestructured format, based on the various HTML and XML tags within the page. Mining efforts here have focused on automatically extracting document object model (DOM) structures out of documents (Wang and Liu 1998; Moh, Lim, and Ng 2000).

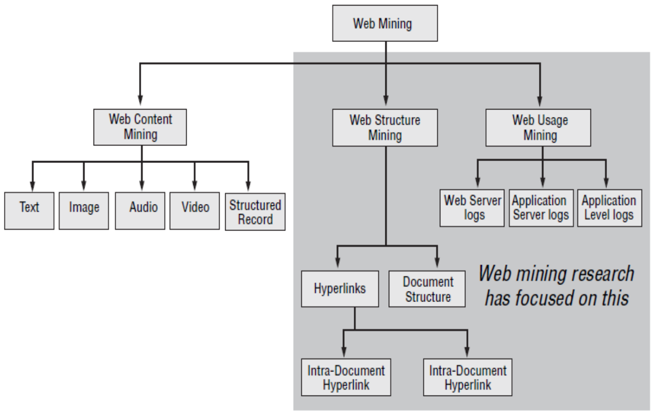


Fig. 1 . Web Mining Taxonomy

C. *Web Usage Mining*

Web usage mining is the application of data mining techniques to discover interesting usage patterns from web usage data, in order to understand and better serve the needs of web-based applications (Srivastava, Cooley, Deshpande, and Tan 2000). Usage data captures the identity or origin of web users along with their browsing behavior at a web site. web usage mining itself can be classified further depending on the kind of usage data considered:

Web Server Data

User logs are collected by the web server and typically include IP address, page reference and access time.

Application Server Data

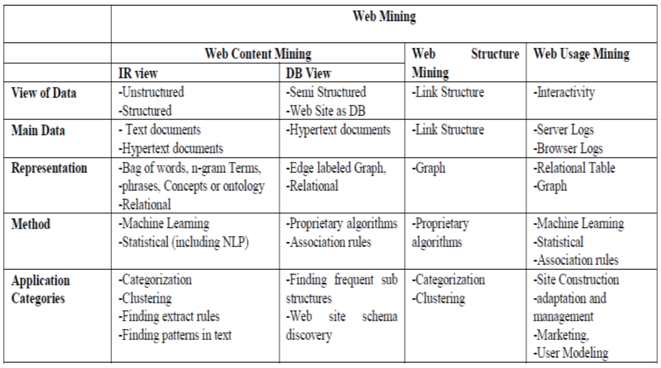
Commercial application servers such as Weblogic,[1,2] StoryServer,[3] have significant features to enable E-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and log them in application server logs.

Application Level Data

New kinds of events can be defined in an application, and logging can be turned on for them — generating histories of these events. It must be noted, however, that many end applications require a combination of one or more of the techniques applied in the above the categories.

D. *Text Mining*

Due to the continuous growth of the volumes of text data, automatic extraction of implicit previously unknown and potentially useful information becomes more necessary to properly utilize this vast source of knowledge. Text mining, therefore, corresponds to extension of the data mining approach to textual data and its concerned with various tasks, such as extraction of information implicitly contained in collection of documents or similarity- based structuring. Text collection in general, lacks the imposed structure of a traditional database. The text expresses the vast range of information, but encodes the information in a form that is difficult to decipher automatically[2].



	Web Mining			
	Web Content Mining		Web Structure Mining	Web Usage Mining
	IR view	DB View		
View of Data	-Unstructured -Structured	-Semi Structured -Web Site as DB	-Link Structure	-Interactivity
Main Data	- Text documents -Hypertext documents	-Hypertext documents	-Link Structure	-Server Logs -Browser Logs
Representation	-Bag of words, n-gram Terms, -phrases, Concepts or ontology -Relational	-Edge labeled Graph, -Relational	-Graph	-Relational Table -Graph
Method	-Machine Learning -Statistical (including NLP)	-Proprietary algorithms -Association rules	-Proprietary algorithms	-Machine Learning -Statistical -Association rules
Application Categories	-Categorization -Clustering -Finding extract rules -Finding patterns in text	-Finding frequent sub structures -Web site schema discovery	-Categorization -Clustering	-Site Construction -adaptation and management -Marketing, -User Modeling

TABLE: 1 Web Mining Categories

IV. KEY CONCEPTS WITH ALGORITHMS

In this section we briefly describe the new concepts introduced by the web mining research community .

A . Ranking Metrics—for Page Quality

Searching the web involves two main steps: *Extracting the pages relevant to a query* and *ranking them according to their quality*. Ranking is important as it helps the user look for "quality" pages that are relevant to the query. Different metrics have been proposed to rank web pages according to their quality. We briefly discuss two of the prominent ones.

1. PageRank

PageRank is a metric for ranking hypertext documents based on their quality. Page, Brin, Motwani, and Winograd (1998) developed this metric for the popular search engine Google[4] (Brin and Page 1998). The key idea is that a page has a high rank if it is pointed to by many highly ranked pages. So, the rank of a page depends upon the ranks of the pages pointing to it. This process is done iteratively until the rank of all pages are determined. The rank of a page *p* can be written as:

$$PR(p) = d/n + (1-d) \sum_{(q,p)\in G} \left(\frac{PR(q)}{Outdegree(q)}\right)$$

Here, *n* is the number of nodes in the graph and OutDegree(q) is the number of hyperlinks on page q. Intuitively, the approach can be viewed as a stochastic analysis of a random walk on the web graph. The first term in the right hand side of the equation is the probability that a random web surfer arrives at a page p by typing the URL or from a bookmark; or may have a particular page as his/her homepage. Here d is the probability that the surfer chooses a URL directly, rather than traversing a links and 1−d is the probability that a person arrives at a page by traversing a link. The second term in the right hand side of the equation is the probability of arriving at a page by traversing a link.

2. Weighted Page Rank

This algorithm was proposed by Wenpu Xing and Ali Ghorbani which is an extension of PageRank algorithm[7]. This Algorithm assigns rank values to pages according to their importance rather than dividing it evenly. The importance is assigned in terms of weight values to incoming and outgoing links. This is denoted as $W^{in}_{(m,n)}$ and $W^{out}_{(m,n)}$ respectively.

$W^{in}_{(m,n)}$ is the weight of link(m,n) as given in (1).. It is calculated on the basis of number of incoming links to page n and the number of incoming links to all reference pages of page m.

$$W^{in}_{(m,n)} = \frac{I_n}{\sum_{p\in R(m)} I_p} \quad \dots\dots\dots\dots\dots\dots\dots (1)$$

In is number of incoming links of page n, Ip is number of incoming links of page p, R(m) is the reference page list of page m.

$W^{out}_{(m,n)}$ is the weight of link(m,n)as given in (2). It is calculated on the basis of the number of outgoing links of page n and the number of outgoing links of all the reference pages of page m.

$$W^{out}_{(m,n)} = \frac{O_n}{\sum_{p\in R(m)} O_p} \quad \dots\dots\dots\dots\dots\dots\dots (2)$$

On is number of outgoing links of page n, Op is number of outgoing links of page p, Then the weighted PageRank is given by formula in (3)

$$WPR(n) = (1-d) + d \sum_{m\in B(n)} WPR(m) W^{in}_{(m,n)} W^{out}_{(m,n)} \quad \dots\dots\dots\dots\dots\dots\dots\dots\dots (3)$$

2.1 PageRank VS Weighted PageRank

In order to compare the WPR with the PageRank, the resultant pages of a query are categorized into four categories based on their relevancy to the given query. They are

- Very Relevant Pages (VR): These are the pages that contain very important information related to a given query.
- Relevant Pages (R): These Pages are relevant but not having important information about a given query.
- Weakly Relevant Pages (WR): These Pages may have the query keywords but they do not have the relevant information

3. Hubs and Authorities

Hubs and authorities can be viewed as “fans’ and “centers” in a bipartite core of a web graph, where the “fans” represent the hubs and the “centers” represent the authorities. The hub and authority scores computed for each web page indicate the extent to which the web page serves as a hub pointing to good authority pages or as an authority on a topic pointed to by good hubs. The scores are computed for a set of pages related to a topic using an iterative procedure called HITS (Kleinberg 1999). First a query is submitted to a search engine and a set of relevant documents is retrieved. This set, called the “root set,” is then expanded by including web pages that point to those in the “root set” and are pointed by those in the “root set.” This new set is called the “base set.” An adjacency matrix, *A* is formed such that if there exists at least one hyperlink from page *i* to page *j*, then $A_{i,j} = 1$, otherwise $A_{i,j} = 0$. HITS algorithm is then used to compute the hub and authority scores for these set of pages.

There have been modifications and improvements to the basic page rank and hubs and authorities approaches such as SALSA (Lempel and Moran 2000), topic sensitive page rank, (Haveliwala 2002) and web page reputations (Mendelzon and Rafiei 2000). These different hyperlink based metrics have been discussed by Desikan, Srivastava, Kumar, and Tan (2002).

Klienberg gives two forms of web pages called as hubs and authorities. Hubs are the pages that act as resource lists. Authorities are pages having important contents. A good hub page is a page which is pointing to many authoritative pages on that content and a good authority page is a page which is pointed by many good hub pages on the same content. A page may be a good hub and a good authority at the same time.

The HITS algorithm treats WWW as directed graph G(V,E), where V is a set of vertices representing pages and E is set of edges corresponds to link. Figure 1 shows the hubs and authorities in web [3].

It has two steps:

1. *Sampling Step*:- In this step a set of relevant pages for the given query are collected.
2. *Iterative Step*:- In this step Hubs and Authorities are found using the output of sampling step.

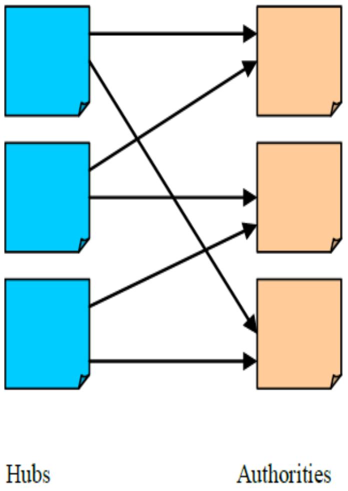


Fig. 3 Hubs And Authorities

Following expressions (1,2)are used to calculate the weight of Hub (Hp) and the weight of Authority (Ap).

$$H_p = \sum_{q \in I(p)} A_q$$

$$A_p = \sum_{q \in B(p)} H_q$$

here Hq is Hub Score of a page, Aq is authority score of a page, I(p) is set of reference pages of page p and B(p) is set of referrer pages of page p, the authority weight of a page is proportional to the sum of hub weights of pages that link to it. Similarly a hub of a page is proportional to the sum of authority weights of pages that it links to.

Constraints with HITS algorithm
Following are some constraints of HITS algorithm[3]

- *Hubs and authorities:* It is not easy to distinguish between hubs and authorities because many sites are hubs as well as authorities.
- *Topic drift:* Sometime HITS may not produce the most relevant documents to the user queries because of equivalent weights.
- *Automatically generated links:* HITS gives equal importance for automatically generated links which may not have relevant topics for the user query
- *Efficiency:* HITS algorithm is not efficient in real time.[5]

HITS was used in a prototype search engine called Clever for an IBM research project. Because of the above constraints HITS could not be implemented in a real time search engine.

B. Robot Detection and Filtering—Separating Human and NonhumanWeb Behavior

Web robots are software programs that automatically traverse the hyperlink structure of the web to locate and retrieve information. The importance of separating robot behavior from human behavior prior to building user behavior models has been illustrated by Kohavi (2001). First, e-commerce retailers are particularly concerned about the unauthorized deployment of robots for gathering business intelligence at their web sites. Second, web robots tend to consume considerable network bandwidth at the expense of other users. Sessions due to web robots also make it difficult to perform click-stream analysis effectively on the web data. Conventional techniques for detecting web robots are based on identifying the IP address and user agent of the web clients. While these techniques are applicable to many well-known robots, they are not sufficient to detect camouflaged and previously unknown robots. Tan and Kumar (2002) proposed a classification based approach that uses the navigational patterns in click-stream data to determine if it is due to a robot. Experimental results have shown that highly accurate classification models can be built using this approach. Furthermore, these models are able to discover many camouflaged and previously unidentified robots.[6]

C. User Profiles— Understanding How Users Behave

The web has taken user profiling to new levels. For example, in a "brick-andmortar" store, data collection happens only at the checkout counter, usually called the "point-of-sale." This provides information only about the final outcome of a complex human decision making process, with no direct information about the process itself. In an on-line store, the complete click-stream is recorded, which provides a detailed record of every action taken by the user, providing a much more detailed insight into the decision making process.
Adding such behavioral information to other kinds of information about users, for example demographic, psychographic, and so on, allows a comprehensive user profile to be built, which can be used for many different purposes (Masand, Spiliopoulou, Srivastava, and Zaiane 2002). While most organizations build profiles of user behavior limited to visits to their own sites, there are successful examples of building web-wide behavioral profiles such as Alexa Research6 and DoubleClick7. These approaches require browser cookies of some sort, and can provide a fairly detailed view of a user's browsing behavior across the web.[8]

D. Preprocessing—MakingWeb Data Suitable for Mining

In the panel discussion referred to earlier (Srivastava and Mobasher 1997), preprocessing of web data to make it suitable for mining was identified as one of the key issues for web mining. A significant amount of work has been done in this area for web usage data, including user identification and session creation (Cooley, Mobasher, and Srivastava 1999), robot detection and filtering (Tan and Kumar 2002), and extracting usage path patterns (Spiliopoulou 1999). Cooley's Ph.D. dissertation (Cooley 2000) provides a comprehensive overview of the work in web usage data preprocessing.

Preprocessing of web structure data, especially link information, has been carried out for some applications, the most notable being Google style web search (Brin and Page 1998). An up-to-date survey of structure preprocessing is provided by Desikan, Srivastava, Kumar, and Tan (2002).

E. Online Bibiliometrics

With the web having become the fastest growing and most up to date source of information, the research community has found it extremely useful to have online repositories of publications. Lawrence observed (Lawrence 2001) that having articles online makes them more easily accessible and hence more often cited than articles that are offline. Such online repositories not only keep the researchers updated on work carried out at different centers, but also makes the interaction and exchange of information much easier.

With such information stored in the web, it becomes easier to point to the most frequent papers that are cited for a topic and also related papers that have been published earlier or later than a given paper. This helps in understanding the state of the art in a particular field, helping researchers to explore new areas. Fundamental web mining techniques are applied to improve the search and categorization of research papers, and citing related articles. Some of the prominent digital libraries are Science Citation Index (SCI),8 the Association for Computing Machinery's ACM portal,9, the Scientific Literature Digital Library (CiteSeer),10 and the DBLP Bibliography.

F. Visualization of the World WideWeb

Mining web data provides a lot of information, which can be better understood with visualization tools. This makes oncepts clearer than is possible with pure textual representation.Hence, there is a need to develop tools that provide a graphical interface that aids in visualizing results of web mining.

Analyzing the web log data with visualization tools has evoked a lot of interest in the research community. Chi, Pitkow, Mackinlay, Pirolli, Gossweiler, and Card (1998) developed a web ecology and evolution visualization (WEEV) tool to understand the relationship between web content, web structure and web usage over a period of time. The site hierarchy is represented in a circular form called the "Disk Tree" and the evolution of the web is viewed as a "Time Tube." Cadez, Heckerman, Meek, Smyth, and White (2000) present a tool called WebCANVAS that displays clusters of users with similar navigation behavior. Prasetyo, Pramudiono, Takahashi, Toyoda, and Kitsuregawa developed Naviz, an interactive web log visualization tool that is designed to display the user browsing pattern on the web site at a global level, and then display each browsing path on the pattern displayed earlier in an incremental manner. The support of each traversal is represented by the thickness of the edge between the pages. Such a tool is very useful in analyzing user behavior and improving web sites.[7]

V. CONCLUSION

In this article, we have outlined three different modes of web mining, namely web content mining, web structure mining and web usage mining. Needless to say, these three approaches can not be independent, and any efficient mining of the web would require a judicious combination of information from all the three sources. We have presented in this paper the significance of introducing the web mining techniques. The development and application of Web mining techniques in the context of Web content, usage, and structure data will lead to tangible improvements in many Web applications, from search engines and Web agents to Web analytics and personalization. Future efforts, investigating architectures and algorithms that can exploit and enable a more effective integration and mining of content, usage, and structure data from different sources promise to lead to the next generation of intelligent Web applications.

REFERENCES

[1] Srivastava J, Desikan P and V Kumar , *"Web Mining-Concepts,Applications & Research Direction*" in 2002 Conference

[2] Srivastava J, Desikan P and V Kumar , "*Web Mining- Accomplishment & Future Directuins*" in 2004 Conference

[3] Rekha Jain and Dr G. N Purohit,"*Page Ranking Algorithms for Web Mining*" International Journal of Computer Applications (0975 – 8887 Volume 13– No.5, January 2011

[4] Srivastava, J., Cooley, R., Deshpande, M., And Tan, P-N. (2000)." *Web usage mining: Discovery and applications of usage patterns from web data"*, SIGKDD Explorations, 1(2), 12-23.H. Poor, An Introduction to Signal Detection and Estimation. New York: Springer-Verlag, 1985, ch.4.

[5] Maier T. (2004). A Formal Model of the ETL Process for OLAP-Based Web Usage Analysis. In Proc. of "WebKDD- 2004 workshop on "*Web Mining and WebUsage Analysis"*, part of the ACM KDD: Knowledge Discovery and Data Mining

[6] Meo R., Lanzi P., Matera M., Esposito R. (2004). Integrating Web Conceptual Modeling and Web Usage Mining. In Proc. of "*Web KDD-2004 workshop on Web Mining and Web Usage Analysis*", part of the ACM KDD: Knowledge Discovery and Data Mining Conference, Seattle, WA.

[7] Desikan P. and Srivastava J. (2004), Mining Temporally Evolving Graphs. In Proceedings of "*Web KDD- 2004 workshop on Web Mining and Web Usage Analysis*", B. Mobasher, B. Liu, B. Masand, O. Nasraoui, Eds. part of the ACM KDD: Knowledge Discovery and Data Mining Conference, Seattle, WA.

[8] Berendt B., Bamshad M, Spiliopoulou M., and Wiltshire J. (2001). Measuring the accuracy of sessionizers for web usage analysis, In Workshop on Web Mining, at the First SIAM International Conference on Data Mining, 7-14.

[9] Srivastava, J., Cooley, R., Deshpande, M., And Tan, P-N. (2000). "*Web usage mining: Discovery and applications of usage patterns from web data*", SIGKDD Explorations, 1(2), 12-23.

[10] J. Hou and Y. Zhang, Effectively Finding Relevant Web Pages from Linkage Information, IEEE Transactions on Knowledge and Data Engineering, Vol. 15, No. 4, 2003.

[11] R. Kosala, and H. Blockeel, "*Web Mining Research: A Survey"*, SIGKDD Explorations, Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining Vol. 2, No. 1 pp 1-15, 2000.