Web Mining Overview, Techniques, Tools, algorithms and Applications: A Survey

Mrs.A.Vaishnavi[1], Dr.N.Balakumar[2]

Research Scholar (P.hD) & Assistant Professor Department of Computer Applications, Pioneer College of Arts and Science, Coimbatore, India

Associate Professor, Department of PG and Research, Pioneer College of Arts and Science, Coimbatore, India

Abstract: *Web mining is application of data mining is very wide. Although web mining may have some difficult structure, long training time ,and difficult to understandable representation of results, web mining have acceptance ability for noisy data and high accuracy and are preferable in data mining. In this paper the data mining based on web mining is researched in detail, and the key technology and ways to achieve the data mining based on web mining networks are also researched.*

Keywords: *data mining, web mining, web personalization, data mining process, implementation.*

I. INTRODUCTION

Data Mining is defined as the procedure of extracting information from massive sets of data. The term data mining is mining knowledge from data. There is a vast amount of data available in the Information Industry. It is necessary to analyze this huge set of data and extract useful data from it.

Extraction of data is not the only process we need to perform; data mining also involves other processes such as Data Cleaning, Data Integration, Data Transformation, Data Mining, Pattern Evaluation and Data Presentation. All these processes are over, we will able to use this information in many applications such as Fraud Detection, Market Analysis, Production Control, Science Exploration, etc. The overview and the terminologies involved in data mining such as knowledge discovery, query language, classification and prediction, decision tree induction, cluster analysis, and how to mine the Web.

Data mining is not to perform an easy task, for the algorithms used can get very complex and data is not always available at one place. It needs to be included from various data sources.

II. WEB MINING

Web mining is the application of data mining techniques and it is used to extract knowledge from web data including web documents, hyperlinks between documents, and usage logs of web sites [1].

Web mining is based on discovery of knowledge form the web. It is used to extract the information , text,audio,video and multimedia document. The main aim of the web mining is include the improvement of web design and structure and generation of dynamic recommendation. Web is huge and broadly distributed data, all the information are interconnected in the storage area. It provides some information service such as news, advertisements, customer information, financial management, education, government and ecommerce etc., there are four sub task of web mining they are

1. Resource finding: It involve the mission of retrieve intended web documents. It is the procedure by which we mine the data and also from online or offline text resources available on web.
2. Information selection and pre-processing: It involves the routine selection and preprocessing of exact 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 repeatedly 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 extract patterns. It plays an important role in pattern mining. A human plays an important role in information on knowledge discovery process(KDD) on web.

III. WEB MINING CATEGORIES

Web mining has three important categories. They are Web Content Mining, Web Structure Mining, and Web Usage Mining is as well as described.

A.WEB CONTENT MINING

The Web content mining refers to the finding of useful information from web contents which include text, image, audio, video, etc. The mining of link structure aims at developing techniques to take advantage of the collective result of web page value which is available in the form of hyperlinks that is web structure mining [7]. It includes extraction of structured data/information from web pages, identification, similarity and integration of data's with similar meaning, view extraction from online sources, and concept hierarchy, knowledge incorporation [8].

It includes extraction of structured data from web pages, similarity and integration of data's with similar meaning, view extraction from online sources, and concept hierarchy, knowledge incorporation [10].
Some of the prominent web content mining techniques
are:

- Unstructured text mining,
- Structured mining,
- Semi structured text mining, and
- Multimedia mining.

1. Unstructured Text Data Mining:

Most of the web pages are in the form of textual data. Content mining requires different application of data mining and text mining techniques [4]. The data mining techniques to unstructured textual data is known as Knowledge Discovery in Texts (KDT), or text data mining, or text mining. Some of the techniques used in text mining known are

- information Extraction,
- Topic Tracking
- Summarization
- Categorization
- Clustering
- Information Visualization

2. Structured Data Mining

The Structured information on the Web represents their host pages. Structured data is without difficulty to extracted compared to unstructured texts. The techniques used for mining structured data are

- Web Crawler
- Wrapper Generation,
- Page content Mining.

3. Semi-Structured Data Mining

Semi-structured data developing from strictly structured relational tables with numbers and strings to enable the natural representation of complex real world objects without sending the application writer into contortions. HTML is a special case of such intra-document structure [11]. The techniques used for semi structured data mining,

- Object Exchange Model (OEM),
- Top down Extraction
- Web Data Extraction language

4. Multimedia Data Mining

Multimedia data mining can be defined as the process of discover exciting patterns from media data such as audio, video, image and text that are not ordinarily accessible by basic queries and associated results. The aim of doing Multimedia data mining is to use the exposed patterns to improve decision making. Comparison of Multimedia data mining techniques with state of the art video processing, audio processing and image processing techniques is also provided [12].The techniques of Multimedia data mining
are:

- SKICAT
- Color Histogram Matching
- Multimedia Miner
- Shot Boundary Detection.

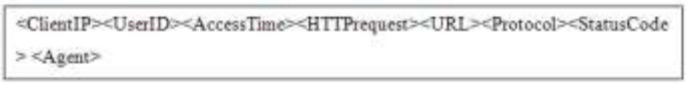
B. WEB USAGE MINING

It is also known as Web log mining, is used to inspect the performance of website users. It tries to find out useful information secondary data resultant from the interaction of users while surfing web [8]. Web usage mining collects the data from Web log report to conclude user access patterns of Web pages. This information is often collected automatically into access logs via the Web server. Typically there are four types of data sources present in which usage data is recorded at dissimilar levels they are: client level collection, browser level collection, server level collection and proxy level collection. It contains four processing stages including

- Data collection
- Preprocessing
- Pattern discovery and Analysis

Data Preprocessing

The input to the preprocessing phase is web logs. Thisfile is stored in a common log format. This format [9] isgiven by W3C (World Wide Web Community). But whichis in unstructured format for Web Usage Mining. The belowfigure shows the common log file data format.



<ClientIP><UserID><AccessTime><HTTPrequest><URL><Protocol><StatusCode
> <Agent>

Pattern Discovery

It focuses on to predict uncover patterns from the abstractions produced as an effect of pre-processing phase. Pattern discovery drawn upon various methods and techniques developed from various fields like, data mining, machine learning, statistics and pattern recognition. Discovery of desired patterns and to extract easily understandable knowledge from them is a tough task. This phase explains some of algorithms.

Pattern Analysis

Pattern analysis is final step in the all web usage mining process. The motivation at the rear this phase is to separates the interesting and uninteresting patterns from the overall patterns discovered during pattern discovery phase.

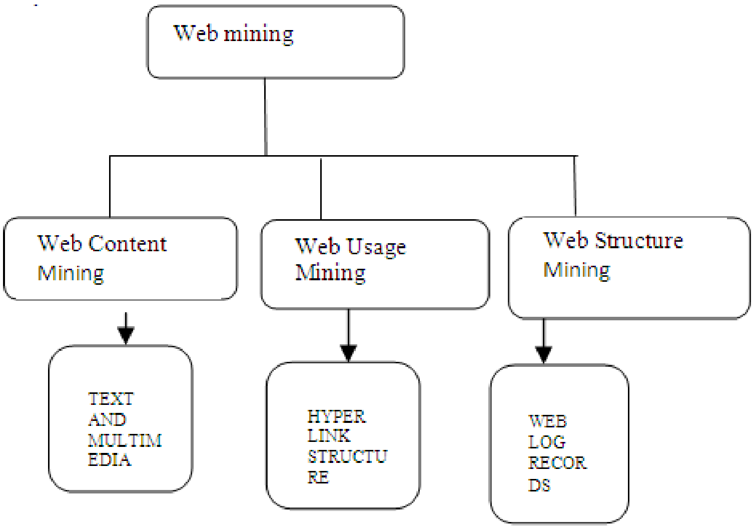


Fig 1.WEB MINING

C.WEB STRUCTURE MINING

In confront for Web structure mining is to deal with the structure of the hyperlinks within the Web only. The growing interest in Web mining, the research of structure analysis had enlarged and these efforts had resulted in a newly emerging research area called Link Mining [], which is located at the juncture of the work in link analysis, hypertext and web mining, relational learning and inductive logic programming, and graph mining. There is a potentially broad range of application areas for this new area of research, including Internet. The Web contains a variety of objects with approximately no unifying structure, with differences in the authoring style and content much greater than in usual collections of text documents. The substance in the

WWW are web pages, and links are in-, out- and co-citation. Attributes include HTML tags, text appearance and anchor texts [9]. This variety of objects creates new problems and challenges, since is not possible to directly made use of existing techniques such as from database management or information retrieval. Link mining had produced some agitation on some of the conventional data mining tasks. As follows, we summarize some of these likely tasks of link mining which are applicable in Web Structure Mining.

1. Link-based Classification: The most current advance of a classic data mining task to linked Domains. The task is to meeting point on the prediction of the category of a web page, based on terms that occur on the page, links between pages, anchor text, html tags and additional feasible attribute found on the web page.
2. Link-based Cluster Analysis The data is segmented into groups, where analogous objects are grouped together, and dissimilar objects are grouped into different groups. Different than the previous task, link-based cluster analysis is unsupervised and can be used to determine hidden patterns from data.
3. Link Type. There are a broad range of tasks relating to the prediction of the existence of links, such as predicting the type of link between two entities, or predicting the purpose of a link.
4. Link Strength. Links could be connected with weights.
5. Link Cardinality. The main mission here is to predict the number of links between objects, Page categorization, finding related pages, predicting duplicated web sites and to find out similarity between them.

IV. WEBMINING TOOLS

In web mining we have sub categories such as, Web Content Mining, Web Structure Mining, and Web Usage Mining. Various types of tools used in all these mining categories.

Web Content Mining Tool [4]

(i) Web Info Extractor

This tool is useful in mining extract structure or unstructured data from web page, extract the web content, and monitoring content update.

(ii) Mozenda

To extract simplify the web data and to manage it affordably Mozenda is valuable. Mozenda support the logins, paging throughout lists of results, AJAX, frames, with other difficult web sites. Mined data can be accessed on online, the exported, as well as used throughout an Application Programming Interface.

(iii)Screen-Scraper [13]

Screen-scraper allowed to mining the content from the web page, like searching a database, SQL server or SQL database, which interfaces with the software, to achieve the content mining requirements. Recently screen scrapers provide the information in HTML, thus it be able to access with a browser.

Commonalities and Differences between the Above Tools

Commonalities

All the tools automate the business mission and to recover the web data in an efficient way.

Differences

- Screen-scrapper needs previous knowledge of proxy server and some knowledge of HTML and HTTP where as other equipment do not require any such knowledge and it need Internet connection to run.
- Automation-Anywhere 7 allow recording of actions this facility is not provided in the other tools.
- Though we have executable file, Mozenda will not allow us to set up without Internet connection, other tools can be set up offline.

V. WEB MINING APPLICATIONS [14]

In past years web applications are being developed at a much faster rate in the industry and also research in web related technologies. Most of these are based on the use of web mining techniques, even though the organizations that developed these applications. We explain some of the most successful applications in this segment.

(i)Web Search—Google

Google is one of the most well-liked and broadly used searches engines. It suggests users access to information from over 2 billion web pages that it has indexed on its server. The value and quickness of the search facility makes it the most prominent search engine. Past search engines are concentrated on web content alone to back the relevant pages to a query. Google was the first to initiate the importance of the link structure in mining information from the web. PageRank, which measures the importance of a page, is the underlying technology in all Google search produce, and uses structural information of the web graph to return high quality grades.

The Google toolbar is another service provided by Google that seeks to make search easier and informative by providing additional features such as highlighting the query words on the returned web pages. Google's web directory provides a fast and easy way to search within a certain topic or related topics. The advertising program introduced by Google targets users by providing advertisements that are relevant to a search question. One of the latest services offered by Google is Google News. It integrates news from the online versions of all newspapers and organizes them categorically to make it easier for users to read "the most applicable news."

It seeks to provide most recent information by constantly retrieve web pages from news site worldwide that are being updated on a regular basis.

(ii)Web-Wide Tracking

"Web-wide tracking," is an entity across all sites he visits, is an intriguing and controversial technology. It can provide an understanding of an individual's lifestyle and habits to a level that is unprecedented, which is clearly of incredible interest to marketers.
Example-DoubleClick Inc.

(iii)Understanding Web Communities-AOL

It is One of the major successes of America Online (AOL) has been its sizeable and loyal customer base. A large portion of this customer base participates in various AOL communities, which are collections of users with similar interests. AOL provides them with useful information and services. Over time these communities have grown-up to be well-visited waterholes for AOL users with shared interests. Applying web mining to the information collected from community interactions provides AOL with a very good understanding of its communities, which it has used for targeted marketing through advertisements and e-mail solicitation.

(iv) EBay

The genius of eBay's founders was to create an infrastructure that gave this urge a global reach, with the convenience of doing it from one's home PC. E-bay has detailed data on bid history, participant rating, bid data, usage data. In addition, it famous auctions as a product selling and buying mechanism and provides the thrill of gambling without the trouble of having to go to Las Vegas. All of this has made eBay as one of the most successful businesses of the internet era. eBay is now using web mining techniques to improve bidding behaviour to determine if a bid is fraudulent .Recent efforts are geared towards understanding participants' bidding behaviours/patterns to create a more efficient auction market.

(v)Personalized Portal for the Web—MyYahoo

Yahoo is an one of the search engine. Yahoo was the first to bring in the concept of a "personalized portal," i.e. a web site designed to have the look-and-feel and content personalized to the requirements of an individual end-user. Mining MyYahoo usage logs provides Yahoo valuable insight into an individual's web usage habits, enabling Yahoo to provide personalized content, which in turn has led to the tremendous popularity of the Yahoo web site.

vi) V-TAG Web Mining Server-Cannotate Technologies

The web mining server supports knowledge information agents that monitor, extract and summarize information from web sources. It is easily to set up GUI. Computerization of tracking and summarizing helps businesses and enterprises to analyse the various processes easily.

VI. ANALYSIS OF LINK ALGORITHMS FOR WEB MINING

Web mining technique provides the added information through hyperlinks where different documents are connected. We can view the web as a directed labeled graph whose nodes are the documents or pages and edges are the hyperlinks between them. This directed graph structure is known as web graph. There are number of algorithms proposed base on link analysis. There are three important algorithms Page Rank, Weighted Page Rank and Weighted Page Content Rank are discussed below:

A. Page Rank

Page Rank is a numeric value that represents how significant a page is on the web. Page Rank is the Google's method of measuring a page's "importance." When all other factors such as Title tag and keywords are in use into account, Google uses Page Rank to regulate results so that more "important" pages move up in the results page of a user's search result display. Google Fig.s that when a page have a hyper links to another page, it is effectively casting a vote for the other page. Google calculates a page's importance from the votes cast for it. How important each number of votes is taken into account when a page's Page Rank is calculated. It matters because it is one of the factors that determine a page's ranking in the search results. It isn't the only factor that Google uses to rank pages, but it is a key one. The order of ranking in Google works like this: Find all pages relevant the keywords of the search. Adjust the results by Page Rank scores.

The algorithm of Page Rank [3] as follows: Page Rank takes the back links into account and propagate the ranking through links. A page has a higher rank, if the sum of the ranks of its backlinks is get more high priority. Fig. 3 shows an example of back links wherein page A is a backlink of page B and page C while page B and page C are backlinks of page D. The unique Page Rank algorithm is given in following equation

PR(P)=(1-d)+d(PR(T1)/C(T1)+…..PR(Tn)/C(Tn)) .... (1)

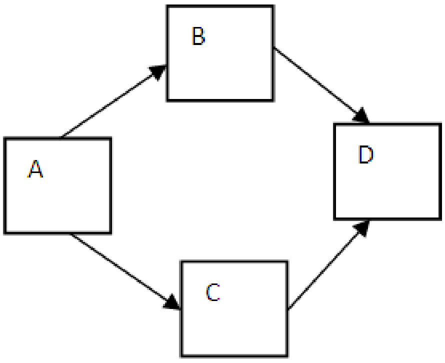


Fig.3: Example of Backlinks

Where, PR (P)= PageRank of page P PR (Ti) = PageRank of page Ti which link to page C (Ti) =Number of outbound links on page T D = Damping factor which can be set between 0 and 1.

B. Weighted Page Rank

Extended Page Rank algorithm- Weighted Page Rank assigns large rank value to more significant pages instead of dividing the rank value of a page evenly among its outlink pages. The significance is assigned in terms of weight values to incoming and outgoing links denoted as respectively and 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. (2)

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. 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. (3)

On is number of outgoing links of page n, Op is number of outgoing links of page p, Then the weighted Page Rank is given by following formula WPR(n)=(1-d)+d (4)

C. Weighted Page Content Rank

Weighted Page Content Rank Algorithm (WPCR) is a proposed page ranking algorithm which is used to give a sort ordered to the web pages returned by a search engine in response to a user question. WPCR is a numerical value based on which the web pages are given an order. This algorithm employs web structure mining as well as web content mining techniques. Importance here means the popularity of the page i.e. how many pages are pointing to or are referred by this particular page.

Output: Rank score

Step 1: Relevance calculation:

a) Find all meaningful word strings of Q (say N)

b) Find whether the N strings are occurring in P or not? Z= Sum of frequencies of all N strings.

c) S= Set of the maximum possible strings occurring in P.

d) X= Sum of frequencies of strings in S.

e) Content Weight (CW)= X/Z
f) C= No. of query terms in P
g) D= No. of all query terms of Q while ignoring stop words.
h) Probability Weight (PW)= C/D
Step 2: Rank calculation:
a) Find all backlinks of P (say set B).
b)PR(P)=(1-d)+d[
c) Output PR(P) i.e. the Rank score

VII. COMPARISON OF ALGORITHMS

Table1 shows the difference between above three algorithms:

Table 1: Comparison of Page Rank, Weighted Page Rank and Weighted Page Content Rank

Contents	Comparison		
	Page Rank	Weighted Page Rank	Weighted Page Content
Mining technique used	WSM	WSM	WSM and WCM
complexity	O(logn)	< O(logn)	< O(logn)
Working procedure	Computers scores at index time. The results are sorted on the importance of pages.	assign large value to more important web pages instead of diving the rank value of a page evenly among its outlink web pages.	Gives sort ordered to the web pages returned by a search engine as numerical value in response to a user question
Inout/Output parameters	backlinks	Backlinks and forward links	Backlinks and forward links content
advantages	It provide important information about given query by diving rank value equally among its outline pages	It provide important information about given query and assigning importance in terms of weight values to incoming and outgoing links	It provide important information and relevancy about a given query by using web structure and web content mining
Search engine	google	google	Research Model
limitations	1)Page rank is equally distributed to outgoing links(connections). 2)It is purely based on the number of the inlinks and outlinks.	1) while some of the web pages may be irrelevant to a given query, it still receives the highest rank. 2) there is a less determination of the relevancy of the pages to a given query.	No limitation best as contrast to page rank and weighted page rank

VIII. CONCLUSION

Web mining is the Data Mining technique that automatically discover or extracts the data or information from web documents. Page Rank and Weighted Page Rank algorithms are used in Web Structure Mining to rank the significant pages. In this paper we focused that by using Page Rank and Weighted Page Rank algorithms users may not get the necessary relevant documents easily, but in new algorithm Weighted Page Content Rank user can get relevant and important pages easily as it employs web structure mining and web content mining. The key parameters used in Page Rank are Backlinks, Weighted Page Rank uses Backlinks and Forward Links as Input Parameter and Weighted Page Content Rank uses Backlinks, Forward Link and Content as Input Parameters. As part of our future work, we are planning to carry out performance analysis of Weighted Page Content and working on finding required relevant and important pages more easily and fastly.

REFERENCES

[1]. J. Srivastva, P. Desikan, and V. Kumar, Web mining – Concepts, Application and Research direction, pp. 51, 2009.

[2]. Preeti Chopra and Md. Ataullah. 2013. A Survey on improving thee Efficiency of different Web Structure Mining Algorithms.

[3]. Kleinberg, J.M., Authoritative sources in a hyperlinked environment. In Proceedings of ACM-SIAM Symposium on Discrete Algorithms, 1998, pages 668-677 – 1998.

[4]. Han, J., Kamber, M. Kamber. "Data mining: concepts and techniques". Morgan Kaufmann Publishers, 2000

[5]. G. Srivastava, K. Sharma, V. Kumar," Web Mining: Today and Tomorrow", in the Proceedings of 2011 3rd International Conference on Electronics Computer Technology (ICECT), pp.399-403, April 2011.

[6]. L. Getoor, Link Mining: A New Data Mining Challenge. SIGKDD Explorations, vol. 4, issue 2, 2003.

[7]. Horowitz, E., S. Sahni, and S. Rajasekaran, eds. Fundamentals of Computer Algorithms. 2008, Galgotia Publications Pvt. Ltd

[8]. Brin, S. and L. Page, The anatomy of a large-scale hypertextual Web search engine. Comput. Netw.

[9]. www.w3c.org/clf.html

[10]. D. Sridevi, Dr. A. Pandurangan, Dr. S. Gunasekaran,"Survey on Latest Trends in Web Mining", International Journal of Research in Advent Technology, Vol. 2, No.3, March 2014.

[11]. Web Mininghttps://www.techopedia.com/definition/ 15634/web-mining

[12]. Chidansh Amitkumar Bhatt, Mohan S. Kankanhalli Multimedia data mining: state of the art and challenge. Journal Multimedia Tools and Applications archive Volume 51 Issue 1, January 2011

[13]. Screen-scraper, http://www.screen -scraper.com Viewed 19 February 2013.

[14]. V. Bharanipriya & V. Kamakshi Prasad, Web Content Mining tools: A Comparative Study in International Journal of Information Technology and Knowledge Management January-June 2011, Volume 4, No. 1, pp. 211-215.

[15]. Jaideep Srivastava, Prasanna Desikan, Vipin Kumar, Web Mining —Concepts, Applications, and Research Directions", AHPCRC technical report 2003 -110,July 2003