International
Journal
Of Advanced Research in Engineering & Management (IJAREM)

A survey on Web Mining Techniques

YOGITA RAWAT
Assistant Professor, PGDM-IT
ITM business school
ITM campus,25/26,Institutional Area, Sec-4, Kharghar(E), Navi Mumbai, Maharashtra 410210

ABSTRACT: As web is a vast collection of completely uncontrolled heterogeneous documents. Due to these characteristic, the web poses a fertile area of data mining research with the huge amount of information available online. This paper consists of brief description about existing methods for Web mining techniques. First section of the paper, consists of the concept of web mining & their categories. Secondly, it covers web mining techniques useful in search engines (like AltaVista , google) which has become inevitable part of our day to life. Lastly, a brief introduction about a Japanese search Engine- "Mondou (RCAAU)", based on the emerging technologies of data mining have been discussed.
Keywords: WSM,WUM,Mondou,WCM

I. Introduction

Web mining aims to discover useful information or knowledge from the Web hyperlink structure, page content and usage log, based on the primary kind of data used in the mining process.

A. Web Mining & Data Mining:

1.1. Data mining :
It is the process of extracting previously unknown information from (usually large quantities of) data, which can, in the right context, lead to knowledge. When data mining techniques are applied to web data, we speak of web-data mining or web mining. The research of web mining is also related to many different research studies, such as database, information retrieval, artificial intelligence, machine learning, natural language processing and many others.

1.2. Web mining:
It refers to the whole of data mining and related techniques that are used to automatically discover and extract information from web documents and services. When used in a business context and applied to some type of personal data, it helps companies to build detailed customer profiles, and gain marketing intelligence. The World Wide Web can be seen as the largest database in the world. This huge and ever-growing amount of data is a fertile area for data mining research.[1]
"It is the application of data mining techniques to discover patterns from the Web".
or
"It is used as a general term for any Data Mining application on data originating from the Web"

1.3. Search Engines:
These are websites which store information about Web Pages and which allow you to search through this information to find the specific page that you are looking for. Some of the most popular search engines are Yahoo, Alta Vista and Google. A web search engine is designed to search for information on the World Wide Web and FTP servers.

The search results are generally presented in a list of results and are often called *hits*. The information may consist of web pages, images, information and other types of files. Some search engines also mine data available in databases or open directories. Unlike web directories, which are maintained by human editors, search engines operate algorithmically or are a mixture of algorithmic and human input.

1.4How web search engines work :A search engine operates, in the following order:

1.4.1 Web crawler: It is a computer program that browses the World Wide Web this process is called *Web crawling* or *spidering*. Many sites, in particular search engines, use spidering as a means of providing up-to-date data. Web crawlers are mainly used to create a copy of all the visited pages for later processing by a search engine that will index the downloaded pages to provide fast searches.[1] Crawlers can also be used for automating maintenance tasks on a Web site, such as checking links or validating HTML code methodical, automated manner or in an orderly fashion.

1.4.2 Search engine indexing : It collects, parses, and stores data to facilitate fast and accurate information retrieval. Index design incorporates interdisciplinary concepts from linguistics, cognitive psychology, mathematics, informatics, physics, and computer science. An alternate name for the process in the context of search engines designed to find web pages on the Internet is **Web indexing**.
1.4.3 A web search query : It is a query that a user enters into web search engine to satisfy his or her information needs. Web search queries are distinctive in that they are unstructured and often ambiguous; they vary greatly from standard query languages which are governed by strict syntax rules.

Relevance of web mining in search engines :
1.5 Web is expanding day by day and people generally rely on search engine to explore the web. In such a scenario it the duty of service provider to provide proper, relevant and quality information to the internet user against their query submitted to the search engine.
1.5.1 It deals with analysis and comparison of web page ranking algorithms based on various parameters to find out their advantages and limitations for the ranking of the web pages.
1.5.2 Based on the analysis of different web page ranking algorithms, a comparative study is done to find out their relative strengths and limitations to find out the further scope of research in web page ranking algorithm.

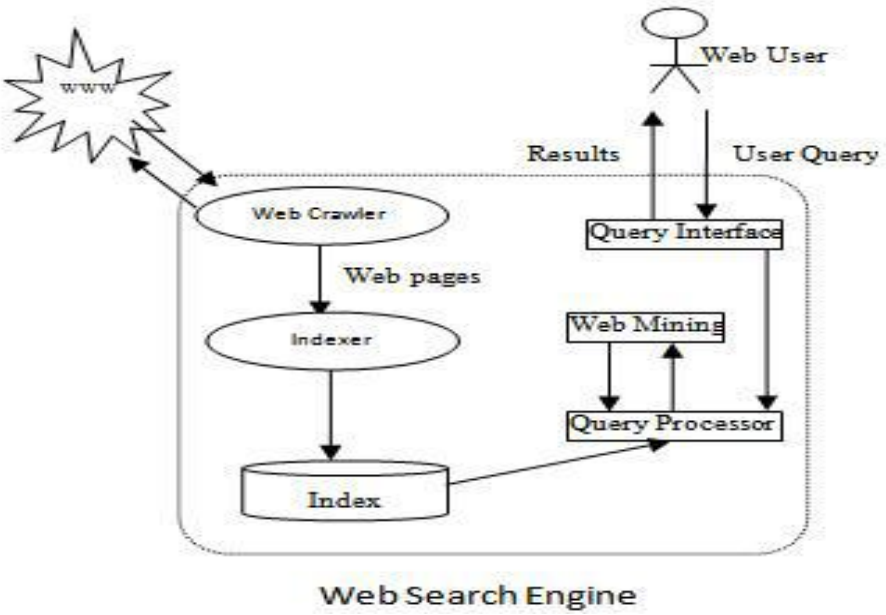

Fig 1. Web search Engine[1]

1.6 Need of Web Mining: Web Data collected at the client and server level can help in better performance and providing better features for Web Services like it helps in understanding of client-server interactions. The data from the interactions can be mined for analyzing interesting patterns. The client level data can provide information to personalize Web services for the users.

II. Web Mining Categories:

Web mining is the technique to classify the web pages and internet users by taking into consideration the contents of the page and behavior of internet user in the past. Web mining helps the internet user about the web pages to be viewed in future.

2.1 Web Mining Taxonomy:

Web mining is divided into three mining categories according to the different sources of data analyzed.

Following three categories of Web Mining**:**

(a) Web content mining (WCM)
(b) Web usage mining(WUM)
(c) Web structure mining (WSM)

2.1 Web content mining(WCM):

It focus on the discovery of knowledge from the content of web pages and therefore the target data consist of multivariate type of data contained in a web page as text, images, multimedia etc.

Web content mining is the technologies that discovers web characteristics and properties from various data-types and attribute values. Data mining and knowledge discovery in web pages and semi-structured documents are extensively studied by many researchers, and web/text mining targets are classified into three major application categories:

1. *Unstructured documents*
2. *Semi structured documents*
3. *Structured documents.*

WCM is responsible for exploring the proper and relevant information from the contents of web.

2.2 Web structure mining (WSM):

It is process of discovering structured information from web. This type of mining is performed at either document level (intrapage) or at Hyperlink level (interpage). WSM is used to find out the relation between different web pages by processing the structure of web.

2.3 Web usage mining(WUM) :

It focuses on the discovery of knowledge from user navigation data when visiting a website. The target data are requests from users recorded in special files stored in the website's servers called log files.

Web usage mining is the technique by analyzing various logs recorded on web server, proxy server, browser's cache, navigation, click-streams and other monitoring mechanisms.

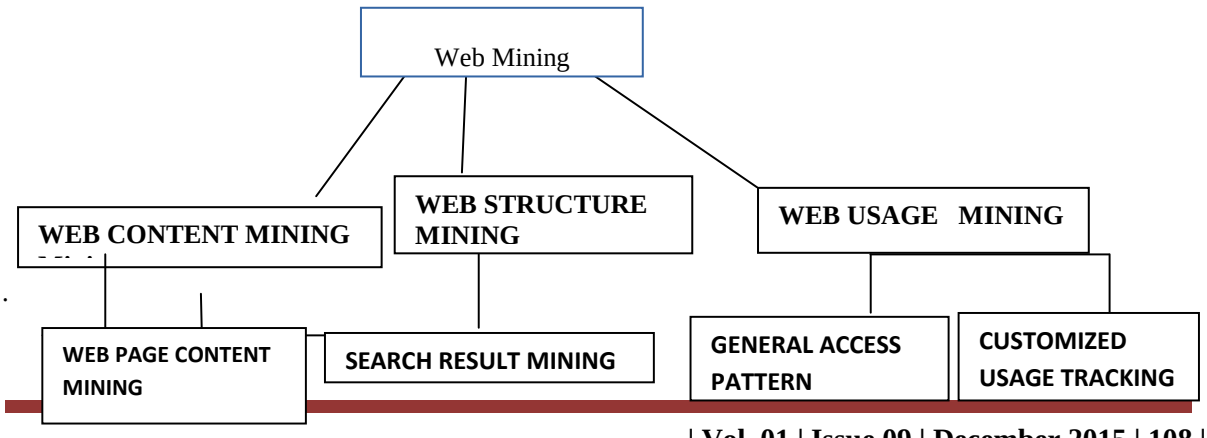

Fig.2Web mining architecture [2]

III. Web mining algorithms used in Search Engines

3.1 Efficient Algorithms used for Web Mining:

An efficient ranking of query words has a major role in efficient searching for query words. There are various challenges associated with the ranking of web pages such that some web pages are made only for navigation purpose and some pages of the web do not possess the quality of self descriptiveness. For ranking of web pages, several algorithms are proposed in the literatures.
The motive behind this paper to analyse the currently important algorithms for ranking of web pages to find out their relative strengths, limitations and provide a future direction for the research in the field of efficient algorithm.

3.1.1 The Page Rank Algorithm:

Page Rank algorithm is the most commonly used algorithm for ranking the various pages. Working of the Page Rank algorithm depends upon link structure of the web pages. The Page Rank algorithm is based on the concepts that if a page contains important links towards it then the links of this page towards the other page are also to be considered as important pages.
The Page Rank considers the back link in deciding the rank score. If the addition of the all the ranks of the back links is large then the page then it is provided a large rank.

A simplified version of PageRank is given by:

PR(u) =□ PR(v)/L(v)

Where Page Rank value for a web page u is dependent on the Page Rank values for each web page v out of the set Bu(this set contains all pages linking to web page u), divided by the number L(v) of links from page v.

This algorithm uses a random surfing model to describe the probability that a page is visited and taking the probability as the importance measurement of the page. They approximated this probability with the famous Page Rank algorithm, which computes the probability scores in an iterative manner.

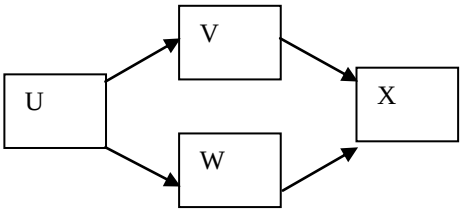

Fig. 3An example of back link [1]

Where U is the back link of V & W and V & W are the back links of X.

3.1.2 HITS(Hyperlink-Induced Topic Search):

HITS algorithm is an iterative algorithm developed to quantify each page's value as an authority and as a hub.

The premise of the algorithm is that a web page serves two purposes: to provide information on a topic, and to provide links to other pages giving information on a topic. So it categorizes a web page in two ways:

- **Authority**: pages that provide important and trustworthy information on a given topic. So an authority is a page that is pointed by many hubs.
- **Hub**: pages that contain links to authorities i.e pointing to many pages.[8]

HITS algorithm ranks the web page by processing in links and out links of the web pages. In this algorithm a web page is named as authority if the web page is pointed by many hyper links and a web page is named as authority if the web page is pointed by many hyper links and a web page is named as HUB if the page point to various hyper links HITS is technically, a link based algorithm.

In HITS algorithm, ranking of the web page is decided by analysing their textual contents against a given query. After collection of the web pages, the HITS algorithm concentrates on the structure of the web only, neglecting their textual contents.

Hyperlink-Induced Topic Search (HITS) (also known as Hubs and authorities) is a link analysis algorithm that rates Web pages, developed by Jon Kleinberg. It was a precursor to PageRank.

The main Idea behind Hubs and Authorities stemmed from a particular insight into the creation of web pages when the Internet was originally forming; that is, certain web pages, known as hubs, served as large directories that were not actually authoritative in the information that it held, but were used as compilations of a broad catalog of information that led users directly to other authoritative pages where

Good hub=a page that pointed to many other pages
& Good authority = a page that was linked by many different hubs.

Some Problems Related to original HITS algorithm below:

Problem 1:

(a) High rank value is given to some popular website that is not highly relevant to the given query.
(b) Drift of the topic occurs when the hub has multiple topics as equivalent weights are given to all Of the out links of a hub page.

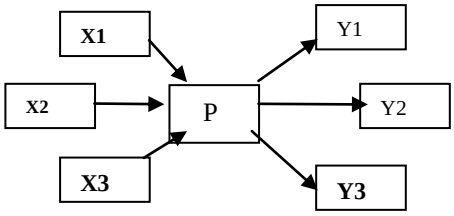

Fig. 4shows an Illustration of HITS process.

Solution to minimize the problem 1 of the original HITS algorithm:

Clever algorithm is the modification of standard original HITS algorithm. This algorithm provides a weight value to every link depending on the terms of queries and endpoints of the link.

An anchor tag is combined to decide the weights to the link and a large hub is broken down into smaller parts so that every hub page is concentrated only on one topic.

Problem 2:

Another limitation of standard HITS algorithm is that it assumes equal weights to all the links pointing to a webpage and it fails to identify the facts that some links may be more important than the other.

Solution to minimize the problem 1 of the original HITS algorithm

To resolve this problem, a probabilistic analogue of the HITS (PHITS) algorithm is proposed by reference [11]. A probabilistic explanation of relationship of term document is provided by PHITS. It is able to identify authoritative document as claimed by the author.

(a)PHITS gives better results as compared to original HITS algorithm.
(b) PHITS can estimate the probabilities of authorities compared to standard HITS algorithm, which Can provide only the scalar magnitude of authority [1].

Advantage of the Page Rank algorithm over the HITS algorithm :

The importance values of all pages are computed off-line and can be directly incorporated into ranking functions of search engines.

3.1.3 Weighted Page Rank Algorithm

Weighted Page Rank [1] Algorithm is proposed by Wenpu Xing and Ali Ghorbani. Weighted page rank algorithm (WPR) is the modification of the original page rank algorithm. WPR decides the rank score based on the popularity of the pages by taking into consideration the importance of both the in-links and out-links of the pages. This algorithm provides high value of rank to the more popular pages and does not equally divide the rank of a page among its out-link pages. Every out-link page is given a rank value based on its popularity. Popularity of a page is decided by observing its number of in links and out links.
It is shown that WPR algorithm finds larger number of relevant pages compared to standard page rank algorithm. As suggested earlier ,the performance of WPR is to be tested by using different websites and future work include to calculate the rank score by utilizing more than one level of reference page list andincreasing the number of human user to classify the web pages.

3.1.4 Weighted Links Rank Algorithm

A modification of the standard page rank algorithm is given by Ricardo Baeza-Yates and Emilio Davis [4] named as weighted links rank (WLRank). This algorithm provides weight value to the link based on three parameters i.e.

(a) Length of the anchor text
(b) tag in which the link is contained
(c) relative position in the page.

Simulation results show that the results of the search engine are improved using weighted links. The length of anchor text seems to be the best attributes in this algorithm. Relative position, which reveal that physical position does not always in synchronism with logical position is not so result oriented.

3.1.5 EigenRumor Algorithm

As the number of blogging sites is increasing day by day, there is a challenge for service provider to provide good blogs to the users.
Limitations of Page rank and HITS :
If these two algorithms are applied directly to the blogs ,the rank scores of blog entries as decided by the page rank algorithm is often very low so it cannot allow blog entries to be provided by rank score according to their importance.

Solution To resolve these limitations:
A EigenRumor algorithm [14] is proposed for ranking the blogs. This algorithm provides a rank score to every blog by weighting the scores of the hub and authority of the bloggers depending on the calculation of eigen vector.

3.1.6 Distance Rank Algorithm

An intelligent ranking algorithm named as distance rank is proposed by Ali Mohammad Zareh Bidoki and Nasser
Yazdani [3]. It is based on reinforcement learning algorithm.
In this algorithm, the distance between pages is considered as a punishment factor. In this algorithm the ranking is done on the basis of the shortest logarithmic distance between two pages and ranked according to them**.**

Advantage of this algorithm :

(a)It can find pages with high quality and more quickly with the use of distance based solution.

Limitation of this algorithm :

(a)It is that the crawler should perform a large calculation to calculate the distance vector, if new page is inserted between the two pages.

3.1.7 Time Rank Algorithm :

An algorithm named as Time Rank, for improving the rank score by using the visit time of the web page is proposed by H Jiang et al.[16] Authors have measured the visit time of the page after applying original and improved methods of web page rank algorithm to know about the degree of importance to the users. This algorithm utilizes the time factor to increase the accuracy of the web page ranking. [2]
Due to the methodology used in this algorithm, it can be assumed to be a combination of content and link structure. The results of this algorithm are very satisfactory and in agreement with the applied theory for developing the algorithm.

3.1.8 Query Dependent Ranking Algorithm

In this approach a simple similarity measure algorithm is used to measure the similarities between the queries. A single model for ranking is made for every training query with corresponding document.
Whenever a query arises, then documents are extracted and ranked depending on the rank scores calculated by the ranking model. The ranking model in this algorithm is the combination of various models of the similar training queries. Experimental results show that query dependent ranking algorithm is better than other algorithms.[3]

3.2 Comparison of different algorithm:

Comparison of some of various web page ranking algorithms is shown in table 1 and in table 2. Comparison is done on the basis of some parameters such as main technique use, methodology, input parameter, relevancy, quality of results.

Algorithm	Page	HITS	Weighted	Eigen Rumor	Web Page

	Rank		**Page Rank**		**Ranking using Link Attributes**
Main Technique	Web Structure Mining ,	Web Content Mining & Web Structure Mining	Web Structure mining	Web Content Mining	Web Content Mining & Web Structure Mining,
Methodology .	This algorithm computes the score for pages at the time of indexing of the pages.	It computes the hubs and authority of the relevant pages. It relevant as well as important page as the result.	Weight of web page is calculated on the basis of input and outgoing links and on the basis of weight the importance of page is decided.	Eigen rumor use the adjacency matrix, which is constructed from agent to object link not page to page link.it gives different weight to web links based on 3 attributes:	Relative position in page, tag where link is contained, length of anchor text
Relevancy	Less (this algo. rank the pages on the indexing time)	More (this algo. Uses the hyperlinks so according to Henzinger, 2001 it will give good results and also consider the content of the page)	Less as ranking is based on the calculation of weight of the web page at the time of indexing.	High for Blog so it is mainly used for blog ranking.	more (it consider the relative position of the pages)
Quality of results	Medium	Less than PR	Higher than PR	Higher than PR and HITS	Medium

Table 1.Comparison between various web page algorithms

IV.Introduction of new Japanese web search engine

4.1 Introduction of Mondou search engine:

Japanese web search engine "Mondou (RCAAU)", which was based on the emerging technologies of data mining. Our search engine provides associative keywords which are tightly related to focusing web pages.

It is important to improve techniques of web mining methods which perform effective organization of search results from the huge volume of surface and hidden web pages. In order to retrieve appropriate web pages effectively, many search engines have been developed based on various web mining algorithms with score functions of usefulness, importance, freshness, popularities and others.

We focus on a relationship between search accuracy and operation cost. However, the search cost to evaluate the quality of a huge number of web pages is a disadvantageous factor in analyzing web pages. Shortly speaking, the processing speed to analyze a lot of web contents and link relations is especially significant problems.
Therefore, in order to execute web structure mining algorithm efficiently, we developed search engine "Mondou" with association rules, and we also implemented advanced information retrieval interfaces. Our proposed web mining algorithms efficiently reduces the computing cost of web search in a large web community.

4.1.1 Web search engine

In order to improve the performance of web search engines, we applied data mining techniques to our search engine "Mondou" in 1995. In this section, we have short introduction to web mining techniques and characteristics of Mondou systems.

Fig. 5 Query window of Mondou

4.2. Mondou systems :

Mondou consists of the following three main modules, *web robots*, *database systems*, *search programs*.
(a) The first module, *web robots*, is the program which collects web pages and stores them into the first module, database. In addition to the common functions of web robots, our robot parses web documents and derives important keywords by using natural language processing's and heuristic operations.
(b)The second module, *database systems*, stores web data not only of keywords, but also of the structure of hyperlinks and other attributes.
(c) The third module, *search programs*, is the most important module, and it provides search results, mining association keywords for "*bibliographic"* and the information visualization for "*distributed*" as shown in Fig. 4.

4.3 Web archiving system

As the number of pages published on the web servers is increasing, it is becoming hard to keep the correctness of web contents and preserve valuable web pages during long time. Therefore, for preserving huge volume of born-digital information in the internet, various public organizations are trying to develop digital archiving systems for web publishing contents, such as MINERVA, Kulturarw3, netarchive.dk, PANDORA, AOLA and others[2].

For example:

(a) One of well-known archiving projects is "www.archive.org" which has more than 150TB archive files, and the U.S. Congress provided a special budget to the Library for a National Digital Information Infrastructure and Preservation Program (NDIPP), as a part of the Consolidated Appropriations Act, 2001.

(b)In Japan, National Diet Library also has been developing a experimental web archiving system, WARP (http://warp.ndl.go.jp/)[6].

4.3.1. Problems of Web archiving system:

In order to handle monotonously increasing digital information, we have to resolve many difficult problems of long life data preservation from various technical aspects. For instance, there are optimizing problems of network bandwidth and CPU costs for execution of distributed web crawlers from surface webs. More difficult problem is how to gather digital contents from deep/hidden webs.
Furthermore, we have to improve the technologies of information retrieval techniques for multimedia web contents, and have to consider emulation and migration technologies of various application programs.

In our researches, we try to apply web mining techniques to the architecture of web archiving systems. Our proposed ideas are based on the experiences of our developed Mondou web search engine and web robots, which are based on text/web mining technologies.

4.3.2. Architecture of Web Archiving Systems:

It is possible to apply many technologies of web search engine to construction of web archiving systems. Therefore, we introduce several technical experiences of our "Mondou (RCAAU)" search engine[8] using techniques of data mining .Mondou provides related keywords to search users by using association rules derived from a set of web pages.

4.3.2.1 Web Robots:

web robots, is the program which crawls and gathers web contents from the surface web and stores them into database systems. In addition to the fundamental functions of the robot, our robots parses documents and select important keywords by several natural language processing including Japanese morphological analysis and other heuristic functions.
Our web robots have the fast gathering mechanism for more popular pages like authority pages, by analyzing the characteristics of hyper links, outside/inside connections and depth of directories.
In web archiving systems, fast crawling is most important issue in order to keep the time consistency of web pages on web servers. Therefore, our developed web robots seem to be applicable in web archiving systems.

4.3.2.2 Database systems:

The second *database systems* store the huge volume of web contents not only of keywords, but also of the number of links from other URLs and many other attributes. We store several tables with various attributes, such as keywords, URLs, hyper links, http servers, IP addresses, and control/management tables for operating Mondou systems.
In addition to these typical attributes, we have to handle time attributes effectively, in order to preserve the history of web publishing in the web archiving system.

4.3.2.3 Search programs :

The third module, *search programs*, is the most important module, which is executed by CGI, and it provides search results, patterns, trends, associative keywords derived by technologies of data mining and information visualization.

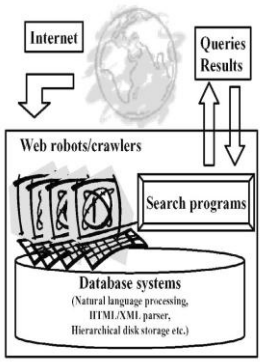

Fig.6 The architecture of web search engine and web archive

V.Conclusion

This paper consists of how web mining is useful in extracting the valuable contents from web using search engines, known as search engine mining, then it focused on different types of web mining: web content mining, web structured mining & web usage mining. Ion covered various techniques used like web crawling, indexing, filtering etc used in search engine mining along with various Web structure mining algorithms like page ranking & hits algorithm as well as different types of page rank algorithms applies in different situations. In the last section, presents introduction of new developed, Japanese search engine -"Mondou **(RCAAU)"** & architecture.

VI. References

[1]. Hiroyuki Kawano: "Applications of web mining – from web search engine to P2P filtering "Proceedings of the 12th International Conference on Informatics Research for Development of Knowledge Society Infrastructure (ICKS'04).

[2]. T. Abe and H. Kawano: "Web Structure Mining based-on Issuing Information Retrieval Model," Proc. of ICSE 2003, Coventry, 9-11 September 2003.

[3]. S. Abiteboul, G. Cob´ena, J. Masanes and G. Sedrati, "A First Experience in Archiving the French Web," Proc. of ECDL 2002, Lecture Notes in Computer Science, No. 2458, pp.1- 15, 2002.

[4]. Ackland, R., and Gibson, R. (2004). Mapping Political Party Networks on the WWW. *Australian Electronic Governance Conference*, Melbourne.

[5]. Ackland, R. (2005). Estimating the Size of Political Web Graphs, revised paper presented to ISA Research Committee on Logic and Methodology Conference.

[6]. J. Srivastava, P. Desikan, and V. Kumar, "Web Mining: Accomplishments and Future Directions," *Proc. US Nat'l Science Foundation Workshop on Next Generation Data Mining* (NGDM), Nat'l Science Foundation, 2002.

[7]. R. Kosala and H. Blockeel, "Web Mining Research: A Survey," *ACM SIGKDD Explorations*, vol. 2, no. 1, 2000, pp. 1–15

[8]. Pooja Devi, Ashlesha Gupta , Ashutosh Dixit,"Comparative Study of HITS and PageRank Link based Ranking Algorithms",International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 2 February 2014