International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 6, June - 2013

Comparative Study of Link Analysis Algorithms

Aditya Upadhyay[1],
Department of CS&E,
Amity University, Haryana

Pooja Batra[2], Asha Sohal[3]
Department of CS&E,
Amity University, Haryana,

Abstract – Web is a collection of various inter-linked web pages. It is a repository of heterogeneous information which is continuously increasing day by day in size and complexity. Due to the vast nature of web, searching has been become a challenging task for a user. One can easily get lost in this rich hyper structure. The main objective of website's owner is to provide the desired information to the user to satisfy their needs. Web mining, the application of data mining, is described as the process that make the use of data mining techniques to extract knowledge from web data. In this paper, we discuss and compare PageRank, Weighted Page rank and HITS algorithm on the basis of case studies.

Keywords: Web Mining, Web Content Mining, Web Structure Mining, Web Usage Mining, PageRank, Weighted PageRank, HITS.

I. INTRODUCTION

Web is a huge collection of static and dynamic web pages. It is an infinite source of information which includes countless hyperlinks. The amount of information is increasing that creates challenges for information retrieval. Retrieval of the desired information on the web, effectively and efficiently, has been becoming a challenge [3]. As the web is unstructured data repository, which provides the large amount f information to the users. User always wants the relevant information when he/she performs searching on the web. But the bulk amount of information increases the complexity of the users to find, extract, filter or evaluate the relevant information.

Web mining [8][6] 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 can be divided into three based on the kinds of data to be mined. The following challenges [4] in Web Mining are:

1) The amount of information is huge on the web.
2) The web is noisy and dynamic
3) Much of the web information is linked
4) Web pages are semi structured.
5) Much of the web information is redundant
6) The web is a virtual society.

This paper is organized as follows- Web Mining is introduced in Section II. The areas of Web Mining i.e. Web Content Mining, Web Structure Mining and Web Usage Mining are discussed in Section III. Section IV describes the various Link analysis algorithms. Section IV (A) defines Page Rank, IV (B) defines Weighted Page Rank and IV(C) defines HITS Algorithm. Section V provides the comparison of various Link Analysis Algorithms.

II. WEB MINING

Web Mining is the process of applying data mining techniques to extract useful information from Web data [4]. Web mining helps the internet user about the web pages to be viewed in future. The kinds of data that can be collected and used in Web Mining analysis include content data, structure data, and usage data [10].

The entire process of extracting knowledge from Web data [8] is follows in Fig.1:

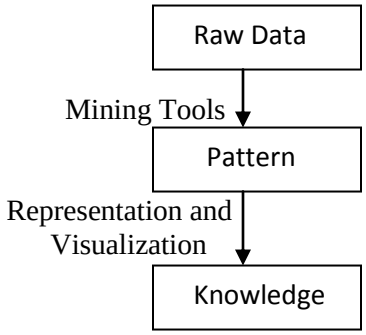


Fig. 1: Web Mining Process

Web Mining [6] consists the following tasks:

1. **Resource finding**: It is the process by which we extract the data from 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.
3. **Generalization**: Automatically discovers general patterns at individual Web site as well as 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.

III. WEB MINING CATEGORIES

Web Mining [6] is divided into three distinct categories, based on the kinds of data to be mined. Web Content Mining (WCM), Web Structure Mining (WSM) and Web Usage Mining (WUM).

Web Content Mining

Web content mining [12] is the technology that discovers web characteristics and properties from various data-types and attributes values. Web content mining focuses on the discovery of knowledge from the content of web pages.
Web pages consist of different types of data attributes, such as text, image, audio, video, meta-data, hyper-link and others.
WCM [7] is the process of retrieving the information from web into more structured forms and indexing the information to retrieve it quickly. It focuses mainly on the structure within a document i.e. inner document level.
Web Content Mining is related to Data Mining because many Data Mining techniques can be applied in Web Content Mining. It is also related with text mining because much of the web contents are text, but is also quite different from these because web data is mainly semi structured in nature and text mining focuses on unstructured text.
The technologies that are normally used in web content mining are NLP (Natural language processing) and IR (Information retrieval) [4].

Web Structure Mining

Web structure mining is an approach based on directory structures and web graph structures of hyperlinks [4]. Web structure mining is closely related to analyzing hyperlinks and link structure on the web for information retrieval and knowledge discovery.
WSM is used to find out the relation between different web pages by processing the structure of web. Web Structure Mining is useful for extracting structure information from the Web. The structure of a typical Web graph consists of Web pages as nodes, and hyperlinks as edges connecting two related pages

WSM can be performed at two levels:

1. **Document structure analysis:** deals with the structure of a document such as the Document Object Model.
2. **Link type analysis:** deals with links that may be inter-document or intra-document.

Web Usage Mining

Web usage mining focuses on techniques that could predict the behavior of users while they are interacting with the WWW [12][6]. Web usage mining collects the data from Web log records to discover user access patterns of Web pages. It can discover the browsing patterns of user and some kind of correlations between the web pages.

Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. WUM is responsible for recording the user profile and user behavior inside the log file of the web.

Typical sources of data in web usage mining are automatically generated data stored in server access logs, referrer logs, agent logs, and client-side cookies, ands user profiles.

A Web usage mining [12] system performs five major tasks:

(i) Data gathering
(ii) Data preparation
(iii) Navigation pattern discovery
(iv) Pattern analysis and visualization
(v) Pattern applications

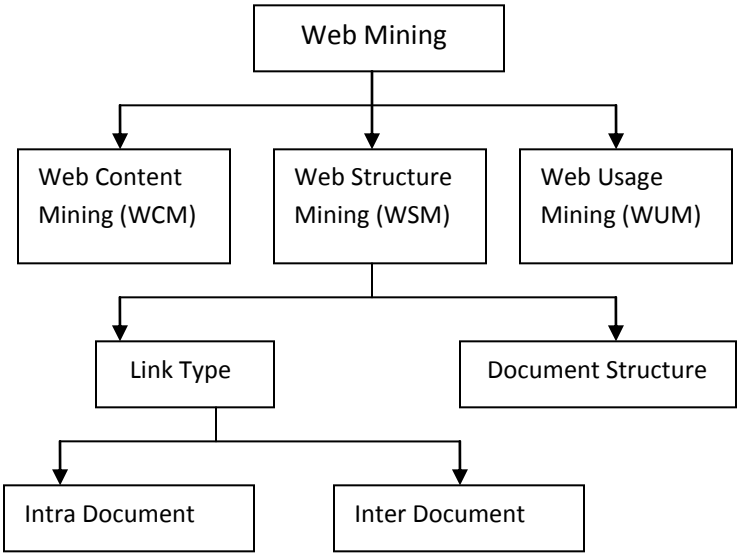


Fig 2: Descriptive Web Mining categorization

IV. LINK ANALYSIS ALGORITHMS

Link-based ranking algorithms propagate page importance through links. The web is described 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.

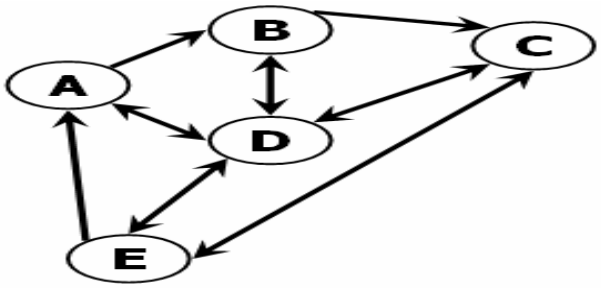


Fig 3: Web Graph

Website structure can be easily understood from figure (3). As we know that website is a collection of related web pages containing images, videos or other digital assets .
In figure (3) A, B, C, D, E is different pages of website. It is clear that if hyperlinks are available then we can easily move between pages. In web mining website structure is also important.
Three important algorithms PageRank, Weighted PageRank and HITS (Hyper-link Induced Topic Search) are proposed based on link analysis.

IV. (A) PAGERANK ALGORITHM

PageRank algorithm is developed by Brin and Page during their Ph. D at Stanford University [1]. PageRank algorithm is used by the famous search engine Google.
PageRank is the most widely used algorithm for ranking the various pages and the working of this algorithm depends upon link structure of the web pages. PageRank is a metric for ranking hypertext documents based on their quality
PageRank [5] takes the backlinks into account and propagates the ranking through links: a page has a high rank if the sum of the ranks of its backlinks is high.

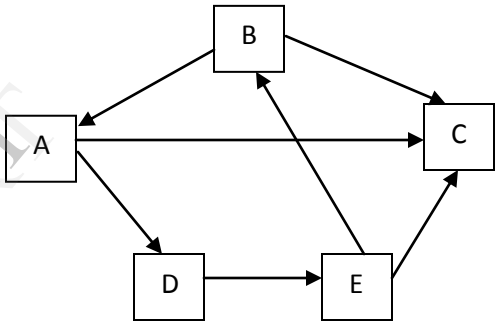


Figure 4. An example of backlinks

Figure 4 shows an example of backlinks: page A is a backlink of page D and page C, page B is a backlink of page A and page C, page D is a backlink of page E, while page E is a backlink of page B and page C.

When one page links to another page, it is effectively casting a vote for the other page. More votes implies more importance. Google calculates a page's importance from the votes cast for it. If a backlink comes from an important page than this link is given higher weightage than those which are coming from Non - important pages [2]. So, the rank of a page depends upon the ranks of the pages pointing to it.

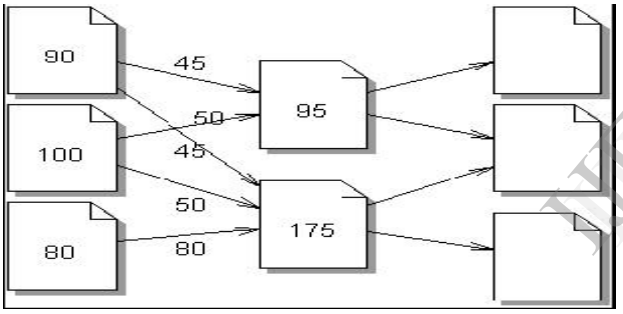
Pagerank is not the only factor that Google uses to rank pages, but it is an important one. The order of ranking in Google works like this:
(1) Find all pages matching the keywords of the search.
(2) Adjust the results by PageRank scores.

Page Rank is a numeric value that represents the importance of a page present on the web.
Brin S. and L. Page described PageRank formula as below

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

Where, **PR (A)** = PageRank of page A
T1….Tn=All pages that link to page A
PR (Ti) =Page rank of page Ti
C (Ti) =the number of pages to which Ti links to
d= damping factor which is set between 0 and 1
PR (Ti)/C (Ti) = PageRank of Ti distributing to all pages that Ti links to.
(1-d)= To make up for some pages that do not have any out-links to avoid losing some page ranks.

In PageRank, rank score of a page *p* is evenly divided among outgoing links. Values assigned to the outgoing links of page *p* are in turn used to calculate the ranks of the pages to which page *p* is pointing.



To understand the working of PageRank algorithm, consider the example hyperlinked structure where X, Y and Z are three web pages.

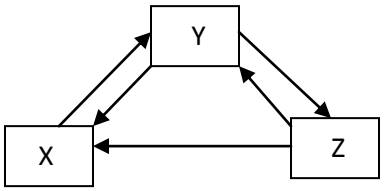


Fig 5 Web Graph

The PageRank for pages X, Y, Z can be calculated using the formula

PR (X) = (1-d) +d ((PR(Y)/2+PR(Z)/2) (a)
PR (Y) = (1-d) +d (PR(X)/1+PR(Z)/2) (b)
PR (Z) = (1-d) +d (PR(Y)/2) (c)

By calculating the above equations with d=0.5 (say), the page ranks of pages X, Y and Z become:
PR (X) =1.2, PR (Y) =1.2, PR (Z) =0.8

Strengths of PageRank algorithm

The strengths of PageRank algorithm are as follows:

- **Less Query time cost:** PageRank has a clear advantage over the HITS algorithm, as the query-time cost of incorporating the precomputed PageRank importance score for a page is low [14].
- **Less susceptibility to localized links:** Furthermore, as PageRank is generated using the entire Web graph, rather than a small subset, it is less susceptible to localized link spam.
- **More Efficient:** In contrast, PageRank computes a single measure of quality for a page at crawl time. This measure is then combined with a traditional information retrieval score at query time. Compared with HITS, this has the advantage of much greater efficiency [14].
- **Feasibility:** As compared to Hits algorithm the PageRank algorithm is more feasible in today's scenario since it performs computations at crawl time rather than query time.

Drawbacks of PageRank algorithm

The following are the problems or disadvantages [1] of PageRank:

- **Rank Sinks**: The Rank sinks problem occurs when in a network pages get in infinite link.
- **Spider Traps**: Another problem in PageRank is Spider Traps. A group of pages is a spider trap if there are no links from within the group to outside the group.
- **Dangling Links**: This occurs when a page contains a link such that the hypertext points to a page with no outgoing links. Such a link is known as Dangling Link.
- **Dead Ends**: Dead Ends are simply pages with no outgoing links.
- PageRank doesn't handle pages with no outedges very well, because they decrease the PageRank overall.
- **Circular References:** If you have circle references in your website, then it will reduce your front page's PageRank [5].
- PageRank score of a page ignores whether or not the page is relevant to the query at hand.

IV.(B)WEIGHTED PAGERANK ALGORITHM

Weighted Page Rank Algorithm is proposed by Wenpu Xing and Ali Ghorbani [3].WPR is a modification of the original page rank algorithm. In WPR, the rank scores of web pages are decided based on the popularity of the pages by taking into consideration the importance of both the in-links and out-links of the pages.
WPR Algorithm assigns larger rank values to more popular pages instead of dividing the rank value of a page evenly among its outlinks pages.
Each out-link page gets a value proportional to its popularity which is decided by observing its number of in-links and out-links. The popularity from the number of inlinks and outlinks is recorded as $W^{in}(V, U)$ and $W^{out}(V, U)$ respectively.

Weighted PageRank formula is

$$PR(u) = (1-d) + d \sum PR(v) W^{in}(v,u) W^{out}(v,u)$$

$W^{in}(V, U)$ is the weight of link (v, u) which is calculated based on the number of in-links of page u and the number of in-links of all reference pages of page v.

$$W^{in}(v, u) = (Iu) / \sum_{p \epsilon R(v)} Ip$$

Where I_u and I_p represent the number of inlinks of page u and page p, respectively. R(v) denotes the reference page list of page v.

$W^{out}(V, U)$ is the weight of link (v, u) which is calculated based on the number of outlinks of page u and the number of outlinks of all reference pages of page v.

$$W^{out}(v, u) = Ou / \sum_{p \epsilon R(v)} Ip$$

Where O_u and O_p represent the number of outlinks of page u and page p, respectively. R(v) denotes the reference page list of page v.

To compare WPR with PageRank, the resultant pages of a query are categorized into four categories based on their relevancy to the given query. [3] They are

1. Very Relevant Pages (VR): These are the pages that contain very important information related to a given query.

2. Relevant Pages (R): These Pages are relevant but not having important information about a given query.

3. Weakly Relevant Pages (WR): These Pages may have the query keywords but they do not have the relevant information.

4. Irrelevant Pages (IR): These Pages are not having any relevant information and query keywords.

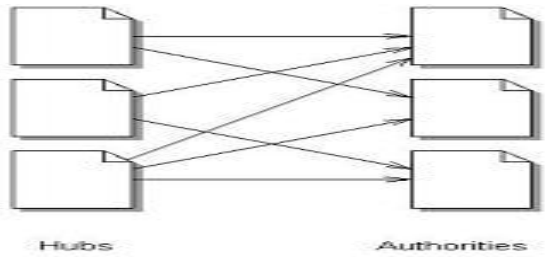
The PageRank and WPR algorithms both provide ranked pages in the sorting order to users based on the given query. So, in the resultant list, the number of relevant pages and their order are very important for users. Relevance Rule is used to calculate the relevancy value of each page in the list of pages. That makes WPR different from PageRank.

IV. (C) HITS

Hypertext Induced Topic Search (HITS) or hubs and authorities is a link analysis algorithm developed by Jon Kleinberg to rate Web pages [7]. HITS algorithm is a search query dependent algorithm that ranks the web page by processing its entire in links and out links. Thus, ranking of the web page is decided by analyzing its textual contents against a given query. 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 HUB if the page point to various hyperlinks [13]. The algorithm produces two types of pages:

- **Authority***:* pages that provide important, trustworthy information on a given topic
- **Hub***:* pages that contain links to authorities.

Authorities and hubs exhibit a mutually reinforcing relationship: a better hub points to many good authorities, and a better authority is pointed to by many good hubs [9].
An Illustration of HUB and authority are shown in figure



Authority Update Rule

$\forall p$, we update $\text{auth}(p)$ to be:

$$\sum_{i=1}^{n} hub(i)$$

Where n is the total number of pages connected to p and i is a page connected to p. That is, the Authority score of a page is the sum of all the Hub scores of pages that point to it.

Hub Update Rule

$\forall p$, we update $\text{hub}(p)$ to be:

$$\sum_{i=1}^{n} auth(i)$$

Where n is the total number of pages p connects to and i is a page which p connects to. Thus a page's Hub score is the sum of the Authority scores of all its linking pages.

The Hub score and Authority score for a node is calculated with the following algorithm [11]:

- Start with each node having a hub score and authority score of 1.
- Run the Authority Update Rule
- Run the Hub Update Rule
- Normalize the values by dividing each Hub score by the sum of the squares of all Hub scores, and dividing each Authority score by the sum of the squares of all Authority scores.
- Repeat from the second step as necessary

Advantages of HITS

There are following advantages of HITS:

- HITS scores due to its ability to rank pages according to the query string, resulting in relevant authority and hub pages.
- The ranking may also be combined with other information retrieval based rankings.
- HITS is sensitive to user query (as compared to PageRank).
- Important pages are obtained on basis of calculated authority and hubs value.
- HITS is a general algorithm for calculating authority and hubs in order to rank the retrieved data.
- HITS induces Web graph by finding set of pages with a search on a given query string.
- Results demonstrate that HITS calculates authority nodes and hubness correctly.

Limitation of HITS algorithm

There are some limitations of HITS algorithm [10]

- **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.

Conclusion

Web Mining is powerful technique which is used to extract the data from past behavior of web users. Web Structure Mining plays an important role in this approach. PageRank, Weighted PageRank and HITS are used in Web Structure Mining to rank the relevant web pages.
PageRank and Weighted PageRank are used in Web Structure Mining. HITS algorithm is used in both structure Mining and Web Content Mining. PageRank, Weighted PageRank calculates the score at indexing time and sort them according to importance of page where as HITS calculates the hub and authority score of n highly relevant pages.
In this paper, the working of different page ranking algorithms is explained which is used to retrieve the relevant pages through search engines.
The main objective is the comparative analysis based on web graphs which shows importance of different ranking algorithms for the web pages..
These comparisons can be considered to generate a new algorithm in future to eliminate the problems and enhance the working of ranking algorithm to get the better relevant results.

References

[1] S. Brin, and L. Page, The Anatomy of a Large Scale Hypertextual Web Search Engine,, Computer Network and ISDN Systems, Vol. 30, Issue 1-7, pp. 107-117,
1998.

[2] Pooja Sharma, Deepak Tyagi, Pawan Bhadana "Weighted Page Content Rank for Ordering Web Search Result" International Journal of Engineering Science and Technology Vol. 2 (12), 2010, 7301-7310.

[3] Wenpu Xing and Ali Ghorbani, "Weighted Page Rank Algorithm", Proceedings of the Second Annual Conference on Communication Networks and Services Research (CNSR'04), 2004 IEEE

[4] M.G. da Gomes Jr. and Z. Gong, Web Structure Mining: An Introduction, Proceedings of the IEEE International Conference on Information Acquisition, 2005.

[5] Tamanna Bhatia," Link Analysis Algorithms For Web Mining ", IJCST Vol. 2, Issue 2, June 2011.

[6] Raymond Kosala, Hendrik Blockee, "Web Mining Research: A Survey", ACM Sigkdd Explorations Newsletter, June 2000, Volume 2.

[7] Rekha Jain, Dr G.N.Purohit, "Page Ranking Algorithms for Web Mining," International Journal of Computer application,Vol 13, Jan 2011.

[8] Cooley, R., Mobasher, B., and Srivastava, J. *"Web mining: Information and pattern discovery on the World Wide Web"*. In proceedings of the 9th IEEE International conference on Tools with Artificial Intelligence (ICTAI' 97), Newposrt Beach, CA, 1997.

[9] J. Kleinberg, Authoritative Sources in a Hyper-Linked Environment, Journal of the ACM 46(5), pp. 604-632, 1999.

[10] S. Chakrabarti, B.Dom, D.Gibson, J. Kleinberg, R. Kumar, P. Raghavan,S. Rajagopalan, and A. Tomkins, Mining the Link Structure of the World Wide Web, IEEE Computer, Vol. 32, pp. 60-67, 1999.

[11] Association Rule Mining based on Ontological Relational Weights, N. Radhika, K.Vidya, Department of Computer Science and Engineering, Aurora's Technological and Research Institute, India.

[12] J. Srivastava, R. Cooley, M. Deshpande, and P. –N. Tan. "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data" (2000), SIGKDD Explorations, Vol. 1, Issue 2, 2000.

[13] A Comparative Analysis of Web Page Ranking Algorithms, Dilip Kumar Sharma et al. / (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 08, 2010, 2670-2676

[14] Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search Taher H. Haveliwala Stanford University taherh@cs.stanford.edu

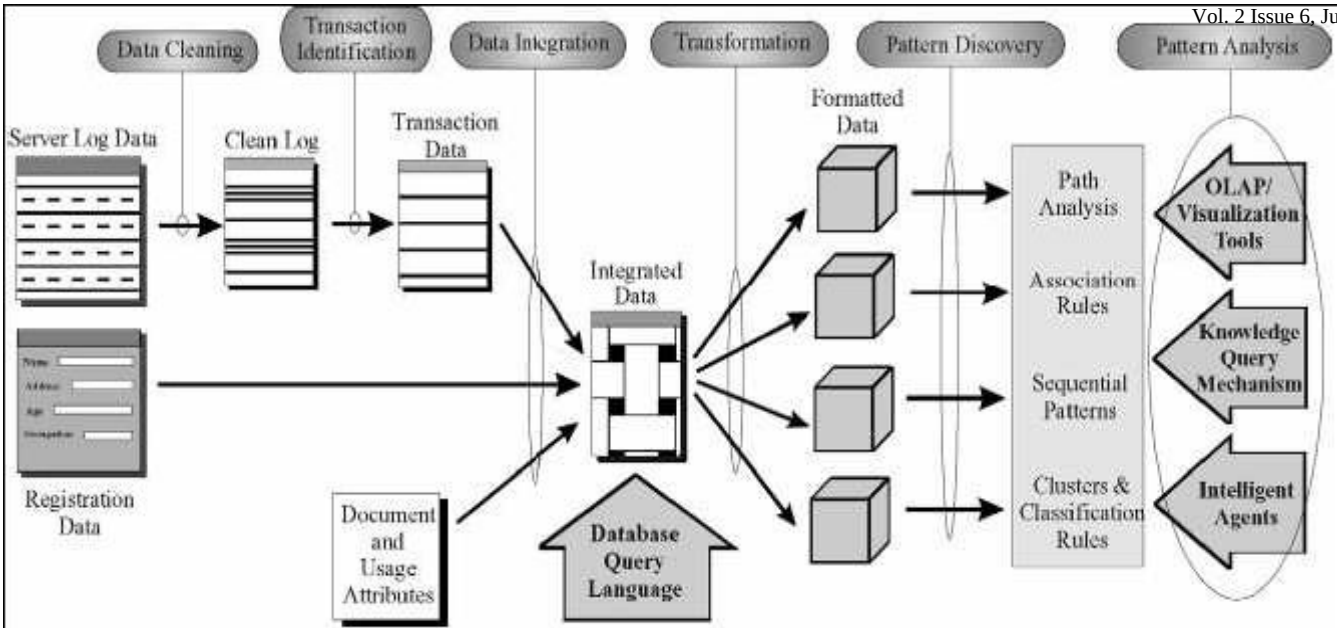


Fig.4 Web Usage Mining Process [8]

TABLE 1: Web Mining Categories

	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
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
Main data	- Text documents -Hypertext documents	-Hypertext documents	-Link Structure	-Server Logs -Browser Logs
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

ALGORITHM / Criteria	PAGERANK	WPR	HITS
Mining Technique used	WSM	WSM	WSM & WCM
Input Parameters	Backlinks	Back and forward-links	Content, Back and forward-links
Methodology	This algorithm computes the score for pages at the time of indexing of the pages.	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.	It computes the hubs and authority of the relevant pages. It relevant as well as important page as the result.
Query dependency	PageRank is Query independent	WPR is Query independent	HITS is Query dependent
Relevancy	Less (this algorithm rank the pages on the indexing time)	Less as ranking is based on the calculation of weight of the web page at the time of indexing.	More (this algorithm Uses the hyperlinks so according to Henzinger, 2001 it will give good results and also consider the content of the page)
Complexity	O(log N)	<O(log N)	<O(log N)
Quality of Result	Medium	Higher than PR	Less than PR
Importance	High. Back links are Considered.	High. The pages are sorted according to the Importance.	Moderate. Hub & authorities scores are Utilized.
Limitation	(1) Page Rank is equally Distributed to outgoing links. (2) It is purely based on the number of in-links and Out-links.	(1)While some pages may be irrelevant to a given query, it still receives the highest rank. (2) Relevancy of the pages is ignored.	(1)Topic drift. (2) Efficiency Problem. (3) Irrelevant authorities Problem. (4) Irrelevant hubs Problem.
Search Engine	It is used by Google	WPR is used in Research model	HITS algorithm is used in IBM search engine Clever