A Survey Paper on Hyperlink-Induced Topic Search (HITS) Algorithms for Web Mining

Mr.Ramesh Prajapati

Lecturer, Information Technology, Gandhinagar Institute of Technology, Gandhinagar
ramesh.prajapati@git.org.in

Abstract--In this Paper Hyperlink-Induced Topic Search (HITS) is a link algorithm for web mining which helps in rating Web pages also known as Hubs and authorities fast and efficient HITS are query dependent and calculated at the time of the search. HITS mechanism for web crawling and retrieval remains as a challenging issue. This paper deals with Survey and comparison of web page ranking algorithms based on various parameter to find out their advantages and limitations for the ranking of the web pages using Web mining. Web mining technique is used to categorize users and pages by analyzing users behavior, the content of pages and order of URLs accessed. In this paper we discuss and compare the different used algorithms i.e.Page Rank and HITS.

Index Terms—Hub, HITS, Networking, Page Rank, Web Mining, Web Usage Mining, Web Content Mining, Web Structure Mining, Web Usage Mining,Weighted Page Rank, HITS

I. INTRODUCTION

As the volume of information on the internet is increasing Day by day so there is a challenge for website owner to Provide proper and relevant information to the internet user. Retrieving of the required web page on the web, efficiently and effectively, is becoming a challenge. Whenever a user wants to search the relevant pages, he/she prefers those relevant pages to be at hand. The bulk amount of information becomes very difficult for the users to find, extract, filter or evaluate the relevant information. This issue raises the necessity of some technique that can solve these challenges. Web mining can be easily executed with the help of other areas like Database (DB), Information retrieval (IR), Natural Language Processing (NLP), and Machine Learning etc. These can be used to discuss and analyze the useful information from WWW.
Following are some challenges:
1) Web is huge. 2) Web pages are semi structured. 3) Web information stands to be diversity in meaning. 4) Degree of quality of the information extracted. 5) Conclusion of knowledge from information extracted.

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 Scale-free network model. In Section V describes the various Links based Analysis algorithms. Page Rank algorithm and its Limitation of PageRank are presented in Section A.In Section B includes Weighted PageRank algorithms. In Section C HITS, Hub and Authorities and Motivation behind HITS.HITS Algorithm and Handling “spam” links In Section D.Handling Span links in Section E. Section VI Based on the literature analysis provides the comparison of HITS vs. PageRank algorithms. Concluding remarks are given in Section VII.

II. WEB MINING

Web mining is the Data Mining technique that automatically discovers or extracts the information from web documents. It is the extraction of interesting and potentially useful patterns and implicit information from activity related to the World.

A. *Web Mining Process*

The complete process of extracting knowledge from Web data [2] is follows in Fig.1:

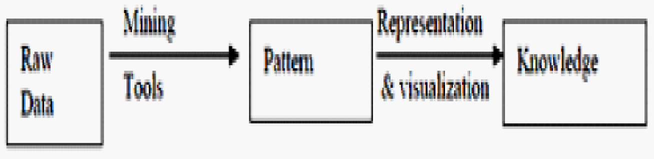


Figure 1: Web Mining Process

It consists of following tasks [4]:

1. Resource finding: It involves the task of retrieving Intended web documents.

2. Information selection and pre-processing: It Involves the automatic selection and pre processing of specific information from retrieved web Resources.

3. Generalization: It automatically discovers general Patterns at individual web sites as well as across multiple sites.

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

III.WEB MINING CATEGORIES

There are three areas of Web Mining according to the web data used as input in Web Data Mining. Web Content Mining, Web Structure Mining and Web Usage Mining.

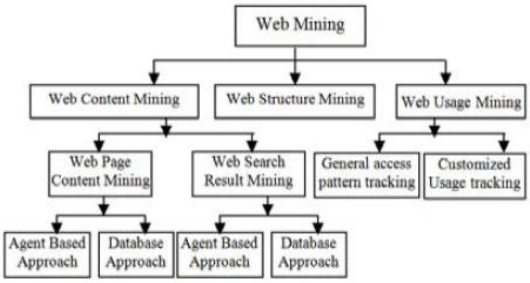


Figure 2: Classification of Web Mining

A. *Web Content Mining*

It is the process of retrieving the information from Web document 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.

B. *Web Structure Mining*

It is the process by which we discover the model of link structure of the web pages. We catalog the links; generate the information such as the similarity and relations among them by taking the advantage of hyperlink topology. The goal of Web Structure Mining is to generate structured summary about the website and web page. Page Rank and hyperlink analysis also fall in this category. It tries to discover the link structure of hyper links at inter document level. As it is very common that the web documents contain links and they use both the real or primary data on the web so it can be concluded that Web Structure Mining has a relation with Web Content Mining. It is using the tree-like structure to analyze and describe the HTML (Hyper Text Markup Language).

C. *Web Usage Mining*

It is the process by which we identify the browsing patterns by analyzing the navigational behavior of user. It focuses on techniques that can be used to predict the user behavior while the user interacts with the web. It uses the secondary data on the web. This activity involves the automatic discovery of user access patterns from one or more web servers. Through this mining technique we can ascertain what users are looking for on Internet. It consists of three phases, namely preprocessing, pattern discovery, and pattern analysis. Web servers, proxies, and client applications can quite easily capture data about Web usage.

IV SCALE-FREE NETWORK MODEL

A simple model for generating “scale-free” networks in following point.

1. Evolution: networks expand continuously by the addition of new vertices, and 2. Preferential-attachment (rich get richer): new vertices attach preferentially to sites that are already well connected. Growing the network (evolution): Starting with a small number *(m*0) of vertices, at every time step we add a new vertex with m (≤*m*0) edges that link the new vertex to *m* different vertices already present in the system.
Growing the network (preferential attachment): To incorporate Preferential attachment, we assume that the probability P that a new vertex will be connected to vertex *i* depends on the Connectivity *k i* of that vertex, so that P(*k i*) = *k i* /□ *k j j* .

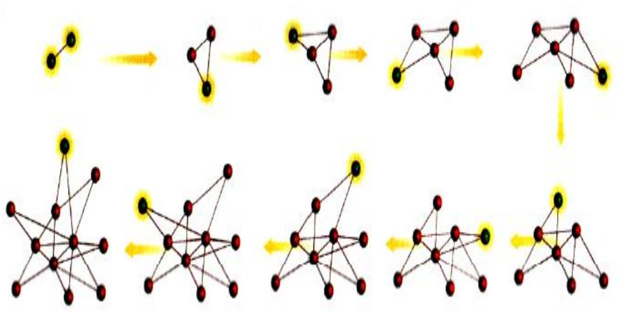


Figure 3: Scale-free network model

V. LINK BASED ANALYSIS

Web mining technique provides the additional 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 based on link analysis. Three important algorithms Page Rank,Weighted PageRank and HITS are discussed below.

A. *Page Rank Algorithm*

Page Rank is a numeric value that represents how important 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 taken into account, Google uses Page Rank to adjust 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 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.Its provides a better approach that can compute the importance of web page by simply counting the number of pages that are linking to it. These links are called as backlinks.If a backlink comes from an important page than this link is given higher weightage than those which are coming from non-important pages. The link from one page to another is considered as a vote. Not only the number of votes that a page receives is important but the importance of pages that casts The algorithm of Page Rank as follows:

Page Rank takes the back links into account and propagates the ranking through links. A page has a higher rank, if the sum of the ranks of its backlinks is high. 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 original Page Rank algorithm is given in following equation

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

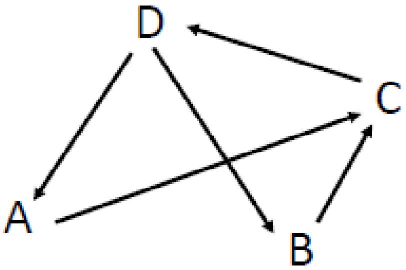


Figure 4: Backlinks Page Rank

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.

A link from A to B is a vote for B cast by A. Votes cast by pages that are *important* weigh more heavily.But there are different types of important nodes:

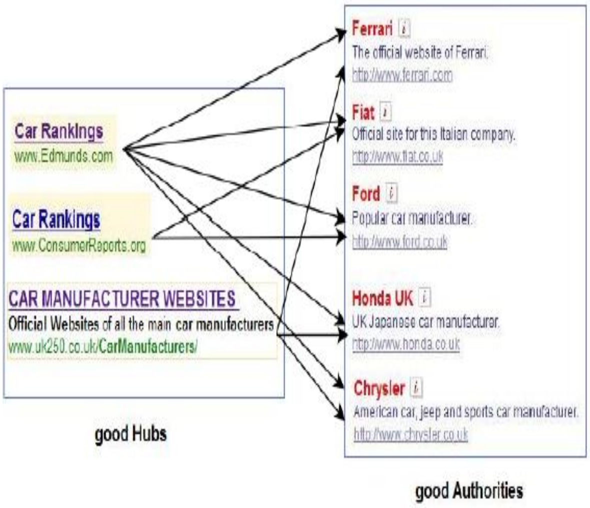

Car Rankings
www.Edmunds.com
Car Rankings
www.ConsumerReports.org
CAR MANUFACTURER WEBSITES
Official Websites of all the main car manufacturers
www.uk250.co.uk/CarManufacturers/
good Hubs
Ferrari
The official website of Ferrari.
http://www.ferrari.com
Fiat
Official site for this Italian company.
http://www.fiat.co.uk
Ford
Popular car manufacturer.
http://www.ford.co.uk
Honda UK
UK Japanese car manufacturer.
http://www.honda.co.uk
Chrysler
American car, jeep and sports car manufacturer.
http://www.chrysler.co.uk
good Authorities

Figure 5: Top automobile markers

Problem to be solved relevant terms may not appear on the pages of authoritative websites. Many prominent pages are not self descriptive. Car manufacturers may not use the term “automobile manufacturers” on their home page. The term “search engine” is not used by any of natural authorities like Yahoo, Google, and AltaVista etc.

B. *Weighted Page Rank*

Extended Page Rank algorithm- Weighted Page Rank assigns large rank value to more important pages instead of dividing the rank value of a page evenly among its outlink pages. The importance is assigned in terms of weight values to incoming and outgoing links denoted as and respectively. This algorithm was proposed by Wenpu Xing and Ali Ghorbani which is an extension of PageRank algorithm.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 . 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.

$$W^{in}_{(m,n)} = \frac{I_n}{\sum_{p \in R(m)} I_p} \qquad \text{….(2)}$$

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.$W^{out}(m,n)$ is the weight of link(m,n)as given in (3). It is calculated on the basis of the number of outgoing links of page n and the number of utgoing links of all the reference pages of page m.

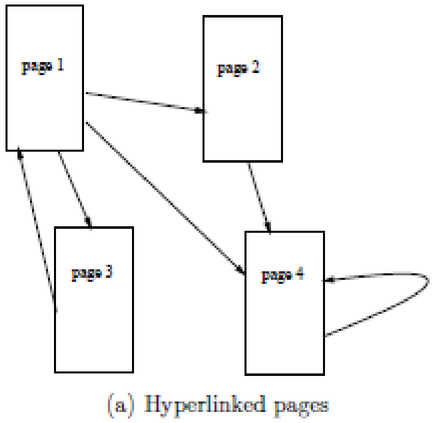
$$W^{out}_{(m,n)} = \frac{O_n}{\sum_{p \in R(m)} O_p} \qquad \text{….(3)}$$

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 (4)

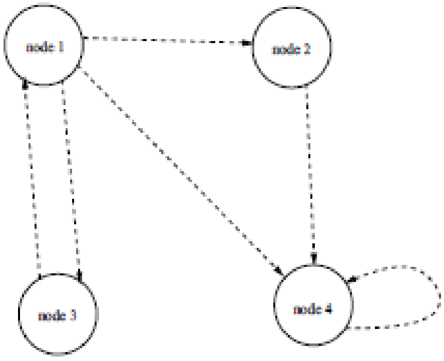
$$WPR(n) = (1-d) + d \sum_{m \in B(n)} WPR(m) W^{in}_{(m,n)} W^{out}_{(m,n)} \qquad \text{…(4)}$$

C. *HITS (Hyper-link Induced Topic Search)*

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 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. In other words, a good hub represented a page that pointed to many other pages, and a good authority represented a page that was linked by many different hubs. The scheme therefore assigns two scores for each page: its authority, which estimates the value of the content of the page, and its hub value, which estimates the value of its links to other pages. 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. Attempts to computationally determine hubs and authorities on a particular topic through analysis of a relevant sub graph of the web.



(a) Hyperlinked pages



(b) Directed graph

Figure 6: Hyperlinked Pages Modeled as Directed Graph

Based on mutually recursive facts: Hubs point to lots of authorities. Authorities are pointed to by lots of hubs. Authority: A valuable and informative webpage usually pointed to by a large number of hyperlinks • Hub: A webpage that points to many authority pages is itself a resource and is called a hub • Authorities and hubs reinforce one another. A good authority is pointed to by many good hubs. A good hub points to many good authorities

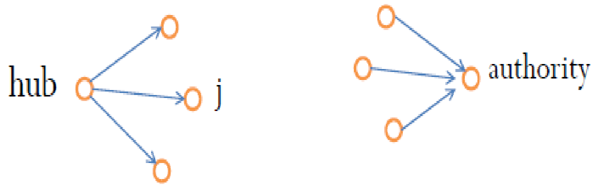


Figure 7: Hub and Authority

Motivation behind HITS:
The creator of page p, by including a link to page q, has in some measure conferred authority on q Links afford us the opportunity to find potential authorities purely through the pages that point to them

What is the problem here? Some links are just navigational "Click here to return to the main menu" Some links are advertisements Difficulty in finding balance between relevance and popularity Solution: Based on relationship between the authorities for a topic and those pages that link to many related authorities-hubs.

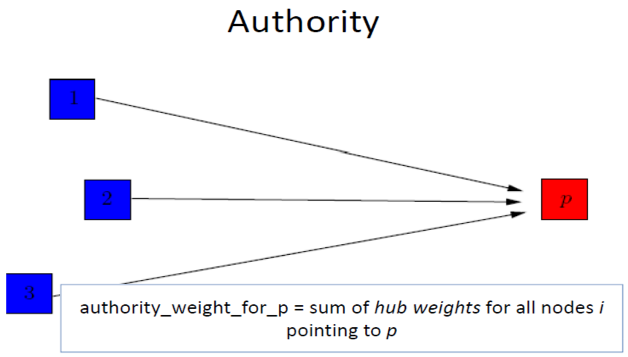

Authority
1
2
3
p
authority_weight_for_p = sum of hub weights for all nodes i pointing to p

Figure 8: Relationship between Authority and Hub

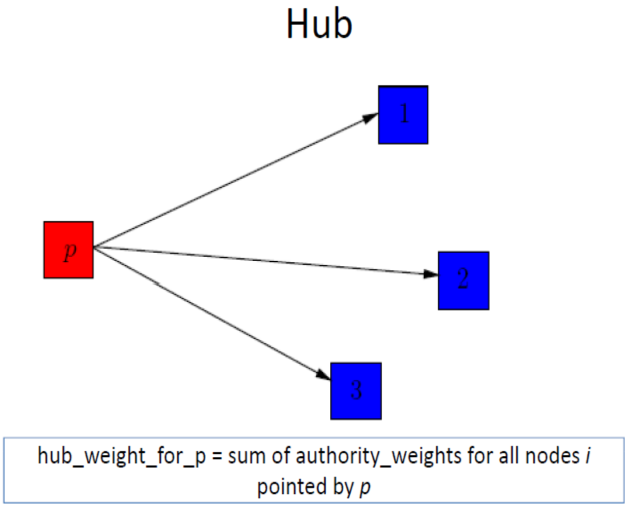

Hub
1
p
2
3
hub_weight_for_p = sum of authority_weights for all nodes i pointed by p

Figure 9: Relationship between Hub and Authority

D. HITS Algorithm

Computes hubs and authorities for a particular topic specified by a normal query.• First determines a set of relevant pages for the query called the *base* set *S*.Analyze the link structure of the web subgraph defined by *S* to find authority and hub pages in this set. Following point's construction of focused sub graph.

- We have a set created by text-based search engine.
- Why do we need subset?
- The set may contain too many pages and entail a Considerable computational cost
- Most of the best authorities may not belong to this set
- Subset properties:
- Relatively small
- Rich in relevant pages
- Contains most (or many) of the strongest authorities

You need relevance – Start filtering

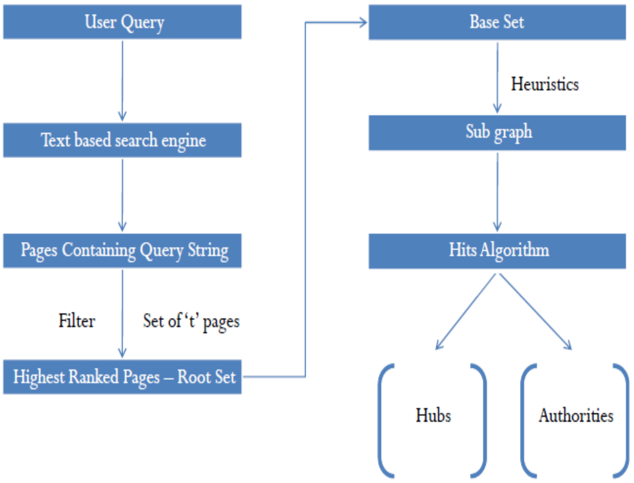


Figure 10: Computes hubs and authorities for a particular topic specified by a normal query

First find a set of relevant pages

- For a specific query *Q*, let the set of documents returned by a standard search engine be called the *root* set *R*.
- Initialize *S* to *R*.
- Add to *S* all pages pointed to by any page in *R*.
- Add to *S* all pages that point to any page in *R*.

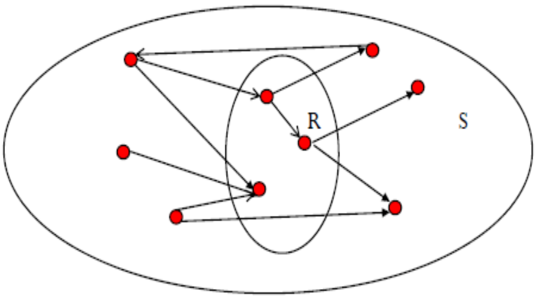


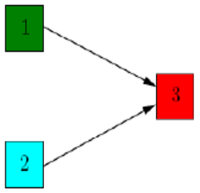
Figure 11: Determines a set of relevant pages for the query

The Subgraph reduction Offset the effect of links that serve purely a navigational function Remove all intrinsic edges from the graph,keeping only the edges corresponding to transverse link. Remove links that are mentioned in more than m pages (m=4-8).

Calculating the Hub and Authority Weights

- A is the adjacency matrix of graph G= (V,E)

- Authority weight: $\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$
- Hub weight: $\begin{bmatrix} h_1 \\ h_2 \\ h_3 \end{bmatrix}$



Iterative Algorithm

Each page p is assigned two non-negative weights, an authority weight **a** and a hub weight **h**.Update the weights of **a** and **h**

– **Authority weight:** $a(j) = \sum_{i:(i,j)\in E} h(i)$

– **Hub weight:** $h(j) = \sum_{i:(j,i)\in E} a(i)$

These operations add the weights of hubs into the authority weight and add the authority weights into the hub weight respectively Alternating these two operations will eventually result in an equilibrium value, or weight, for each page.

G: a collection of n linked pages

- Set $a_0 = [1/n, \ldots, 1/n]^T$
- Set $h_0 = [1/n, \ldots, 1/n]^T$
- For t=1,2, …,k

```
– For j= 1,2, …,n
    • Obtain new authority weights $a_t'(j) = \sum_{i:(i,j)\in E} h_{t-1}(i)$
    • Normalize weights $a_t(j) = \frac{a_t'(j)}{\sum_j a_t'(j)}$
    • Obtain new hub weights $h_t'(j) = \sum_{i:(j,i)\in E} a_t(i)$
    • Normalize weights $h_t(j) = \frac{h_t'(j)}{\sum_j h_t'(j)}$
– end
```

- End

Based on the Survey of this HITS algorithm the overall graph of with Authority and Hubness represented as following. HITS algorithm discovered, they share similar roles in terms of their email communication pattern in the data set. Our algorithm discovers this structure as well. The estimated rankings are so close to the actual ones that it is difficult to distinguish them.

Properties of HITS algorithm.

Several interesting results follow directly

(1) Webpage ordering. The authority ranking is, on average, identical to the ranking according to webpage in degrees. To see this, we have the following:

Elements of the principal eigenvector u1 are none increasing, Assuming webpages are indexed such that there in degrees are in no increasing order. We have, for any i < j,

$$\mathbf{u}_1(i) - \mathbf{u}_1(j) = \frac{d_i}{\lambda_1 - h_i} - \frac{d_j}{\lambda_1 - h_j} = \frac{(d_i - d_j)[\lambda_1 - d_i d_j/(n-1)]}{(\lambda_1 - h_i)(\lambda_1 - h_j)} \geq 0,$$

because $\lambda_1 - d_i d_j/(n-1) > h_i - d_i d_j/(n-1) = d_i(1 - (d_i + d_j)/(n-1)) > 0$, using Eq.(5.9), and $(\lambda_1 - h_i)(\lambda_1 - h_j)$ is positive. □

From this, we conclude that to the extent that the fixed degree sequence random graph approximate the web, ranking web pages by their authority scores is the same as ranking by their in degrees. Analogous results hold for hub ranking. These indicate that the duality relationship embedded in mutual reinforcement between hubs and authorities are manifested by their in degree and out degrees.

(2) Uniqueness. If d1 is larger than d2, then the principal eigenvector of LTL is unique, and is quite different from the second principal eigenvector.

(3) Convergence. The convergence for HITS can be rather fast: (1) the starting vector $x^{(0)}$ = (1,---, 1)^T has large overlap with principal eigenvector u1, but little overlap with other principal eigenvectors u_k; k = 2; ---,m, because u_k contains negative nodal values (2) In the iterations to compute u1, the convergence rate depends on Y2/Y1 ~ h1/h2~ d1/d2 ' $(1/2)^2$ = 1/4; using and the fact that in degrees follow power-law distribution [10]: d_i * 1=i^2. Thus the iteration converges rapidly. Typically 5-10 iterations are sufficient.

(4) Web communities. HITS algorithm has been used to identify multiple web communities using different eigenvectors [22, 16]. The principal eigenvector defines a Dominant web community. Each of other principal eigenvector u_k defines two communities, one with non-negative values {i|u_k(i) > 0}and the other with negative values {i|u_k(i) < 0}.From the pattern of eigenvectors in our solutions, the positive region of different eigenvectors overlap substantially. Thus the communities of positives regions nest with each other; so do communities of negative regions. Therefore, we believe this method to identify multiple communities is less effective. This difficulty is also noticed in practical applications .A number of web community discovery algorithms are being developed, e.g., trawling to find bipartite cores network maximum flow and graph-clustering. One advantage of these methods is that weak communities (topics) can be separated from dominant communities and thus identified. Without explicit community discovery, web pages of weak topics are typically ranked low by HITS (and by in degree ranking) and are often missed.

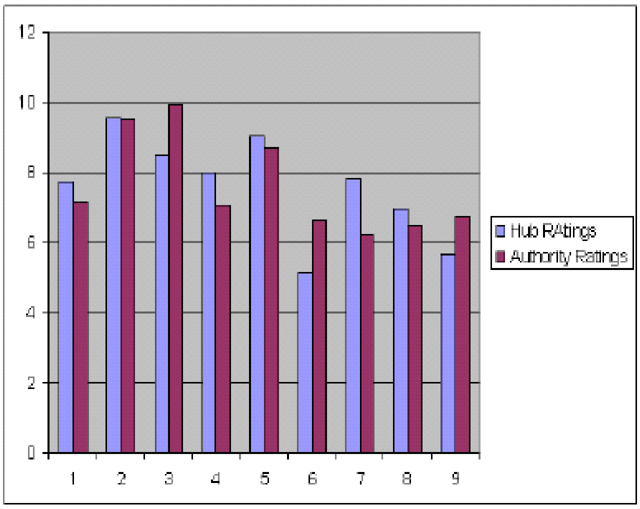


Figure12: Comparison of Hubs and Authority Scores

Evaluation of Web Search Results

Quality of Web search results is definitely a subjective matter. Importance and usefulness of pages will vary from person to person. But, can one objectively infer about the quality of a page? Based on link structure of WWW how can we define a Good page? Will it be query dependent? Or can it be query independent? Current Web search systems respond to user queries within a fraction of a second. Users will not mind having a Web search system that responds within a few seconds, provided it returns considerably better results. But as stated by Kleinberg in

"We are lacking objective functions that are both concretely defined and correspond to human notions of quality."

We describe below several parameters to objectively evaluate a page.

1 Popularity

Popularity of a node can be equated with number in links it has. Here we assume that, if many nodes point to a node then it should be a popular node.

2 Centrality

Distance from node *u* to *v* can be defined as minimum number of links via which we can reach *v* from *u*. Radius of a node is it's maximum distance from any node in the graph. Center of the graph is the node with the smallest radius. The more central the node, the more easily we can reach other parts of the graph from it.

3 Prestiges

Prestige of a node can be recursively defined as the sum of the prestiges of nodes pointing to it. Here we consider not just number of in links but also quality of those in links. This is the motivation behind Page Rank.

4 Informativeness

A node is informative if it points to several nodes that contain useful information. Here we consider not just number of out links, but also quality of nodes pointed.

5 Authority

Authority of a node is similar to the prestige of the node with the difference that authority is measured with respect to some focused tiny sub graph on a particular topic.

Selective Expansion of Root Set

Consider the step of expanding the root set. Generally root set is of the order of a few hundred pages. Although existing search systems return thousands of results for broad queries, only top few are directly relevant and important for the topic of the query. After adding all pages in one link neighborhood, the size of the base set becomes of the order of a few thousand pages. Most of the pages added are either useless or including them in the base set causes topic drift.

We attack these problems by selectively expanding the root set. So instead of expanding all the pages, we expand selective pages only. Further we are also selective in adding in links or out links of selected pages.

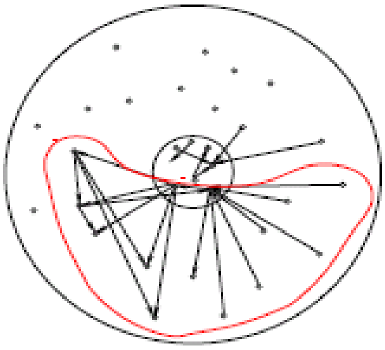


Figure13: Pages causing topic drift and topic Contamination

We first build page to host connectivity matrix of the root set. Then we carry out power iteration computation using that matrix and calculate hub and authority values as mentioned in Upper Section. Then we select the top few hubs and authorities from root set. Thus we have picked up candidate pages for expansion from the root set. Now simply adding all pages in one link neighborhood of these selected pages

Can again cause the same problems as that of simple expansion used in HITS. So we consider the following factors while adding pages to the root set.

- As per definition hubs should point to good authorities. So pages pointed to by top hubs in the root set can be good authority pages. So pages pointed to by top hubs are added to root set.
- As per definition authorities are pointed to by good hubs. So pages pointing to top authorities in the root set can be good hubs. So these pages are added to the root set.

The Web can be viewed as a directed graph whose nodes are the documents and the edges are the hyperlinks between them, as shown in below figure. The graph structure of the World Wide Web can be used for analysis to improve the retrieval performance and classification accuracy the rank value indicates an importance of a particular page. A hyperlink to a page counts as a vote of support. The HITS of a page is defined and depends on the number and Page Rank metric of all pages that link to it (i.e.). A page that is linked to by many pages with high Page Rank receives a high rank itself. If there are no links to a web page there is no support for that page.

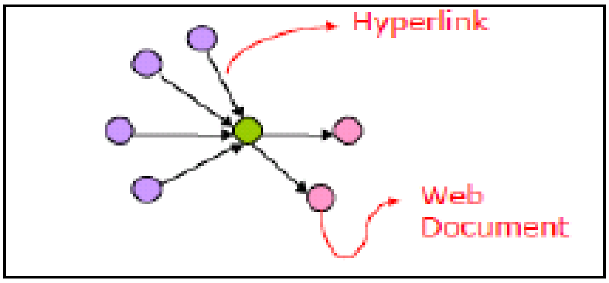
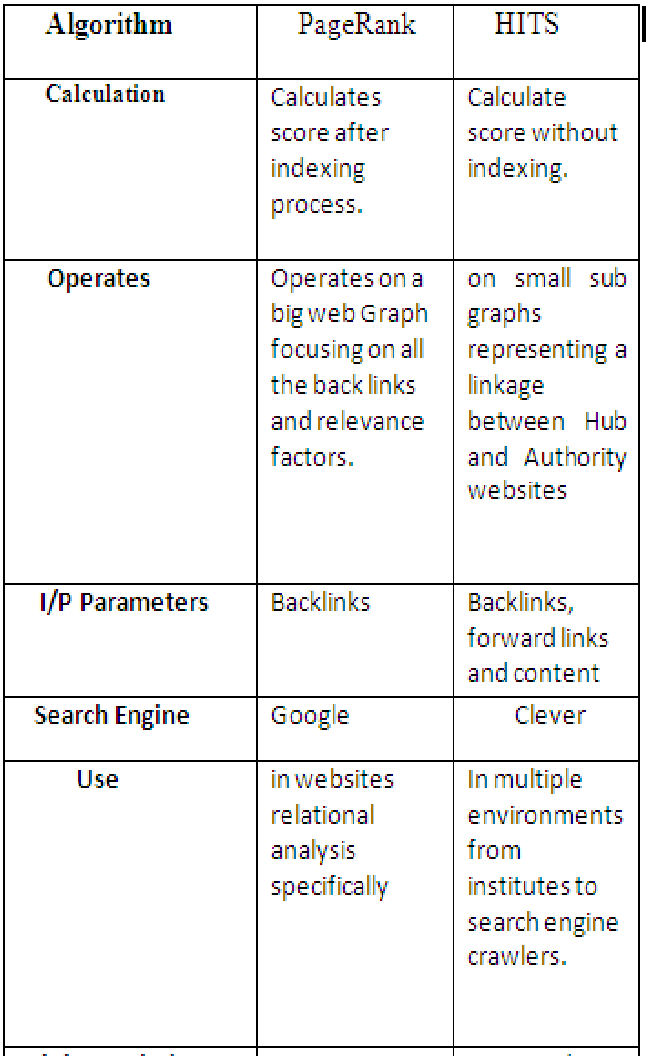


FIGURE14: WEB GRAPH STRUCTURE

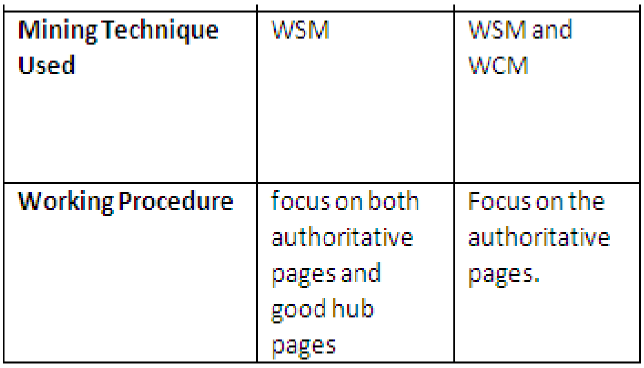
VI. COMPARISON

Table1 shows the difference between above two algorithms:

Table 1: Comparison of Page Rank and HITS



Algorithm	PageRank	HITS
Calculation	Calculates score after indexing process.	Calculate score without indexing.
Operates	Operates on a big web Graph focusing on all the back links and relevance factors.	on small sub graphs representing a linkage between Hub and Authority websites
I/P Parameters	Backlinks	Backlinks, forward links and content
Search Engine	Google	Clever
Use	in websites relational analysis specifically	In multiple environments from institutes to search engine crawlers.



Mining Technique Used	WSM	WSM and WCM
Working Procedure	focus on both authoritative pages and good hub pages	Focus on the authoritative pages.

VII. CONCLUSION

Web Mining is powerful technique used to extract the Information from past behavior of users. Selective expansion Of root set and a different way of calculating hub and authority values. As a result we had a very small base set and we were able to distill results only for one topic even if a query was ambiguous. Various Algorithms are used in Web Mining to rank the relevant pages. The main focus of web structure mining is on link information. Web usage mining focuses on understanding user behavior as depicted in the web access logs while interacting with a website. PageRank, Weighted PageRank and HITS treat all links equally when distributing the rank score. In the Problem of page rank and weight page algorithm relevant terms may not appear on the pages of authoritative websites. Many prominent pages are not self descriptive. In HITS algorithm all links should be equally treated so we considerations two problem. Some links may be more meaningful than other links.Further.we also observed that selective expansion of the root set is also rich in quality, as many pages from the expanded root set topped the hub and authority list. For the future works, there are still many issues that need to be explored With the HITS algorithm, Being HITS algorithms are not good enough to be applied in mining the informative structures, the phenomenon that authorities converge into densely linked irrelevant pages is called *topic drift problem*. This problem is notorious in the area of Information Retrieval. To address this problem, we propose some other types of link analysis- based modification.

REFERENCES

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

[2] Cooley, R, Mobasher, B., 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.

[3] Pooja Sharma, Pawan Bhadana, “Weighted Page Content Rank For Ordering Web Search Result”, International Journal of Engineering Science and Technology, Vol 2, 2010.

[4] R. Kosala, H. Blockeel “Web mining research” A survey. ACM Sigkdd Explorations,2(1):1-15, 2000.

[5] Wang jicheng, Huang Yuan,Wu Gangshan, Zhang Fuyan, "Web mining: Knowledge discovery on the Web Systems", Man and Cybernetics 1999 IEEE SMC 99 conference Proceedings. 1999 IEEE International conference

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

[7] Taher H. Haveliwala, "Topic-Sensitive Page Rank: A Context-Sensitive Ranking Algorithms for Web Search", IEEE transactions on Knowledge and Data Engineering Vol.15, No 4 July/August 2003.

[8] 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.

[9] P Ravi Kumar, and Singh Ashutosh kumar, Web Structure Mining Exploring Hyperlinks and Algorithms for Information Retrieval, American Journal of applied sciences, 7 (6) 840-845 2010.

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

[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.

[12] HITS Algorithm - Hubs and Authorities on the Internet", Available:http://www.math.cornell.edu/~mec/Winter2009/RalucaRemus/Lecture4/lecture4.html

[13] HITS ", Available: http://en.wikipedia.org/wiki/PageRank.

[14] R. Weiss, B. Velez, M. Sheldon, C. Nemprempre, P. Szilagyi, D.K. Gifford,, HyPursuit: A Hierarchical Network Search Engine that Exploits Content-Link Hypertext Clustering," *Proceedings of the Seventh ACM Conference on Hypertext*, 1996.

[15] M.R. Hen zinger. Hyperlink analysis for the web. IEEE Internet Computing, 5:45{50, 2001.

[16] M. Kessler. Bibliographic coupling between scientific papers. American documentation, 14:10-25, 1963.

[17] J. M. Kleinberg. Authoritative sources in a hyperlinked Environment. J. ACM, 48:604-632, 1999

[18] R. Lempel and S. Moran. SALSA: stochastic approach for link-Structure analysis and the TKC effect. ACM Trans. Information Systems, 19:131-160, 2001.

[19] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, P. Raghavan, S. Rajagopalan,Au-tomatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text," *Proc. 7th International World Wide Web Conference*, 1998.

[20] D. Gibson, J. Kleinberg, and P. Raghavan. Inferring Web Communities from link topology. In *Proc. 9th ACM Conference On Hypertext and Hypermedia (HyperText 98)*, pages 225–234, Pittsburgh PA, June 1998.