Improving Search Engine Results Using Hits Algorithm

[1]Selva Priya.K,[2]T.Shrikkavi ,[3]B.Sailakshmi
[1]Information Technology,[2] Information Technology
[1]RMD Engineering College

Abstract - **The World Wide Web consists of millions of interconnected web pages that provide information to the user present in any part of the world. The World Wide Web is expanding and growing in size and the complexity of the web pages. That is why it is necessary to retrieve the best or the web pages that are more relevant in terms of information for the query entered by the user in the search engine. To extract the relevancy of a web page, the search engine requires applying retrieval or a ranking module that applies to a ranking algorithm on the web to fetch the web pages in order of the importance of the information entered by the user in the query. The ranking algorithm is much efficient to rank the surface web, i.e. the web pages that can be indexed by the search engine, as well as the hidden web, i.e. the web pages that cannot be indexed by the search engine. This Paper proposed an algorithm consists of Page Rank Algorithm, Term Weighting Technique and Visitor Count. In this paper introduced HITS(Hyper Induced Topic Search) algorithm for web search.**

I: INTRODUCTION

As the size of World-Wide Web, is growing continuously day by day. A large part of the web is hidden to various search engines. The term "Hidden Web" or "Deep Web" or "Invisible Web" are the web content that the search engine can't capture. To search content on the Web, search engines use web crawlers that follow hyperlinks. This technique is ideal for searching resources on the surface Web but is often ineffective at finding deep Web. For example, these crawlers are not capable of finding the dynamic pages that are the result of database queries due to the infinite number of queries that are possible. The deep Web is qualitatively different from the "surface" Web. Hidden web mainly available in searchable databases and can be retrieved only by a direct query. Without this method, the hidden web pages are not available for the user and the content is there in the databases which can't be fetched by the traditional search engines. Crawlers weave throughout the Web, indexing the url of pages they come through. When these software programs run into a page from the Invisible Web, they are not designed in a manner to decide what to do with the result pages. These crawlers can record the urls and the information in the page could not use by them.

II: RELATED WORKS

Ranking web pages using Weighted PageRank [1] algorithm gives the improved version of PageRank and HITS ranking algorithms. Weighted PageRank algorithm considers the importance of both inlinks and outlinks of the web pages and distributes PageRank of the web page on the basis of web pages popularity. This algorithm assigns higher rank values to the popular pages rather dividing the PageRank of the web pages evenly among its out links. Higher the popularity of the out link, higher will it receive the rank value. The Weighted PageRank algorithm comprises of six major activities, these activities are: Finding a website, Building a web map, finding the root set, finding the base set, Applying algorithm, evaluating the results. A. Kritikopoulos, M. Sideri and I. Varlamis proposed an algorithm to rank web pages based on content similarity. This algorithm gives a new ranking system, which evaluates the similarity between the interconnected pages [2]. It depends on the concept: "the visitor of a web page have a tendency to visit web pages with similar content rather than content irrelevant pages". Wordrank technique is more effectively used for topic based searches; it gives the high priority to the interconnected pages. Higher the interconnected links, higher the weightage of the web page.

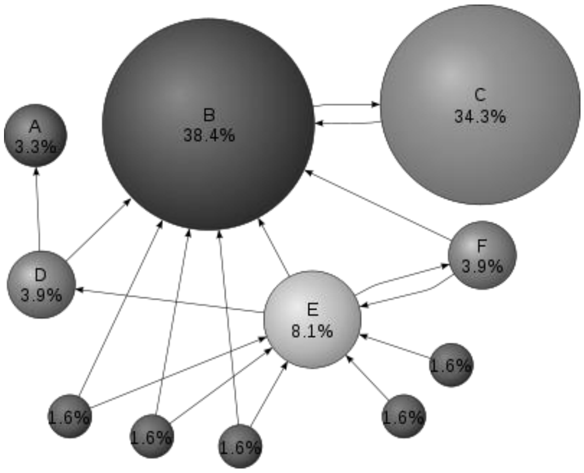
III: IMPROVING SEARCH ENGINE RESULTS USING HITS ALGORITHM:

As the World Wide Web is becoming an important data source for millions of users in the world, it has become essential to develop the new techniques to access this information efficiently from the huge amount of available data. After the pages has been retrieved using the different techniques, it's become important to rank these pages according to the relevancy. Thus many ranking algorithms have been used by the web search engines to give the results in an appropriate format to the user. The Proposed ranking algorithm consists of four different attributes. These are:

A. PageRank using HITS algorithm
B. Term Weighting Technique [TWT]
C. Active Website
D. Visitor Count

A) HITS ALGORITHM

PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites.



Mathematical PageRanks for a simple network, expressed as percentages. (Google uses a logarithmic scale.) Page C has a higher PageRank than Page E, even though there are fewer links to C; the one link to C comes from an important page and hence is of high value. If web surfers who start on a random page have an 85% likelihood of choosing a random link from the page they are currently visiting, and a 15% likelihood of jumping to a page chosen at random from the entire web, they will reach Page E 8.1% of the time. (The 15% likelihood of jumping to an arbitrary page corresponds to a damping factor of 85%.) Without damping, all web surfers would eventually end up on Pages A, B, or C, and all other pages would have PageRank zero. In the presence of damping, Page A effectively links to all pages in the web, even though it has no outgoing links of its own.

Problems of PageRank Algorithm : PageRank algorithm is dependent only on the link structure rather than the query. The PR value of pages only depends on the number of in-links and out-links of a page. PageRank algorithm is a static algorithm and has nothing to do with the query. In order to overcome this concept in page rank we use HITS algorithm which is based on query based search.

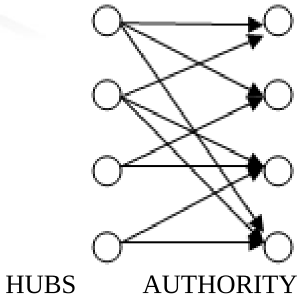
The HITS algorithm was developed by J. Kleinberg and is now a part of the CLEVER Searching project of the IBM Almaden Research Center

The premise of the algorithm is that a web page serves two purposes:

- To provide information on a top
- To provide links to other pages giving information on a topic.

This gives rise to two ways of categorizing a web page.

- First, we consider a web page to be an authority on a subject if it provides good information about the subject.
- Second, we consider the web page to be a hub if it provides links to good
- Authorities on the subject.
- Authorities– pages that are relevant and are linked to by many other pages
- Hubs– pages that link to many related authorities
- The HITS algorithm is an iterative algorithm developed to quantify each page’s value as an authority and as a hub. The HITS algorithm is not applied to the graph representing the whole web, but rather to a sub graph, typically of 1000–5000 nodes, derived from traditional text matching of the query terms in the search topic.



Intuitive Idea to find authoritative results using link analysis:

- Not all hyperlinks related to the conferral of authority.
- Find the pattern authoritative pages have:

Authoritative Pages share considerable overlap in the sets of pages that point to them.

- First Step:– Constructing a focused subgraph of the WWW based on query
- Second Step– Iteratively calculate authority weight and hub weight for each page in the subgraph

B. TERM WEIGHTING TECHNIQUE

The term weighting technique is based on probabilistic and vector space model. There are three main parameters used in calculating TWT. The parameters are document length, document frequency and term frequency.

C: ACTIVE WEBSITE:

Active websites are those which is being currently available for user in the server. For example if an organization had been shutdown , they remove all their official website. Thus this website become inactive now.

D. VISITOR COUNT:

Hits on the web page are considered as the visitor count. It is assumed that more the number of hits on the web page and higher the popularity of the web page

IV) IMPLEMENTATION AND EXPERIMENTS

1. LOGIN

In this we have login in both user side and admin side. User side login for user accessing. Admin side login for authenticating purpose. Admin has the activities like Add new website, View all web site, Top search website, Active websites.

- Add new website - Adding a new website for particular keyword
- View all website - Viewing all the websites which is stored.
- Active website - Checking whether that website is currently working or not.
- Top search website – Frequently used website will be in top while searching.

2 SIGN UP OR REGISTER

When a new user enter they need to fill the necessary details in the form to create an account for them, Again whenever they need they can login with username and password.

Fig New user side registration

3 KEYWORD MAP MODULES

When the user searches for something it will map with keywords user searched and display the most related websites for the searching keyword.

E.g.: When the user search for "top colleges in Chennai" the websites like www.srmuniversity.in , www.rmd.ac.in will be relevant solution for user search.

4 ACTIVE WEBSITE

Active website is checking whether that website is currently in use. Before adding a website admin must check whether that website is active or not.

E.g.: Now a day's orkut is not in use (not active) if admin had already added it now they should remove it.

5 TERM WEIGHTING URL WEBSITES

Term weighting is how frequently a particular term is used. If single term of a website had more weight age then that website URL get priority.

6 RANKING WEBSIE

The rank for a website is given based on how many users visited that particular website and how long they stayed up on that particular website.

Eg: Consider the both website www.srmuniversity.in ,www.ssn.in have same number of users visited so now we check for time if user stayed up for 50minutes in srmuniversity page and 40minutes in ssn page then priority (rank) will be given to srmuniversity page.

7 KEY WORD SEARCH

This module is to display the website according to the content user search. This will give most appropriate and relevant website as result. In this module we defined previously as for this keyword these website has to be shown as result.

E.g.: When user search as "best online shopping website" for this keyword admin would set a website as result like www.flipkart.com

8 ADDING NEW WEBSITE

The keyword for the website must be given in the add keyword. Title of the website should be given in the title about website column URL must be given in the target website and check for its availability. Content description should be same as the add keyword column.

IV) CONCLUSION

The difference between Page Rank and HITS is not essential because a PageRank can be equally well applied to a subgraph and HITS can be equally applied to entire web graph an important characteristics of web page ranking is that we strongly emphasize the top ranked web pages. Therefore in assessing and ranking, only the top ranked web pages are considered. The results are displayed in two types: one is Highly Ranked authority web pages, but relatively with small indegrees and another is WebPages with large indegrees, but ranked low by HITS or PageRank. These WebPages would have been incorrectly ranked if we simply count indegrees. But the result should be displayed as one of the following:

- Corresponds to highly related papers though they are not popularly visited.
- Corresponds to frequently visited but not highly related paper.

In web ranking schemes for most queries, the number of retrieved webpages is usually very large, easily be larger than 10,000 webpages. Ranking the webpages such that the most informative webpages are placed within top 20 is therefore a truly challenging task. HITS or Page Rank effectively gives a slightly different but useful perturbation (deviation) from the indegree ranking. For this reason, we believe a search engine should provide multiple ranking schemes for user to choose from.

V) FUTURE ENHANCEMENTS

The ranking of a web page for the search engine is one of the significant problems at present. The ranking algorithm (CHWRA) as discussed above gives us the desired result. To make the result more appropriate the new concept of "Access Time Length" can also be usedwith the given algorithm. The "Access Time Length" uses a biasing feature to provide better rank prediction. The feature used by ATL is "**Length of the time spent on a web page**". When the ATL spent on a web page by a user goes beyond the average time-length spent on this web page by a larger ratio, then this means access time length can be used as the user's access time-length on this webpage.We have discussed a personalization extension of search us at the group level. However, there are straightforward extension that could produce personalized results at the user level. For example, we could devise a distance metric to compare the URLs of a user with the URLs of all other users. Then, when a specific user poses a query we could limit the users that participate on the results based on that metric.

VI) REFERENCES

[1] Wenpu Xing and Ali Ghorbani Faculty of Computer Science University of New Brunswick Fredericton, NB, E3B 5A3, Canada "Weighted Page Rank Algorithm".

[2] Apostolos Kiriakopoulos, Martha Sideri, Iraklis Varlamis, "Wordrank: A Method for Ranking Web Pages Based on Content Similarity," bncod,pp.92-100, 24th British National Conference on Databases(BNCOD'07), 2007.

[3] Amit Singhal, Google, Inc. "Modern Information Retrieval: A Brief Overview", http://singhal.info/ieee2001.pdf

[4] Crawling the HiddenWeb,Sriram Raghavan Hector Garcia-Molina,Computer Science Department, Stanford University,Stanford,CA94305,SA,frsram,hectorg@cs.stanford.edu."http://ilpubs.stanford.edu:8090/456/1/2000-6.pdf".

[5] Shiguang Ju, Zheng Wang, Xia Lv School of Computer and Telecommunication Engineering, Jiangsu University, Zhenjiang, P.R.China "Improvement of Page Ranking Algorithm Based on Timestamp and Link", 2008 International Symposiums on Information Processing.

[6] Chia-Chen Yen Department of Information Management National Yunlin University of Science and Technology Yunlin, Taiwan, Jih-Shih Hsu Department of Information Management National Yunlin University of Science and Technology Yunlin, Taiwan. "Page rank Algorithm Improvement by Page Relevance Measurement".

[7] Zhou Cailan,Computer & Science Technology Department,Wuhan University of Technology Hubei, China. Chen Kai, Li Shasha Computer & Science Technology Department Wuhan University of Technology Hubei, China. "Improved Page Rank Algorithm Based on Feedback of User Clicks".

[8] Gyanendra Kumar, Neelam Duhan, A. K. Sharma,Department of Computer Engineering, YMCA University of Science & Technology, Faridabad, India. "Page Ranking Based on Number of Visits of Links of Web Page", eInternational Conference on Computer & Communication Technology (ICCCT)-2011.

[9] Dilip Kumar Sharma, GLA University, Mathura UP, India. A.K Sharma, YMCA University of Science and Technology, Faridabad, Haryana, India. "A Novel Ranking Algorithm of Query Words Stored in QIIIEP Server", Dilip Kumar Sharma et.al. / International Journal of Engineering Science and Technology, Vol. 2(11), 2010, 6097-6107.

[10] Yi, J., Nasukawa, T., Niblack,W., & Bunescu, R. (2003), 'Sentiment Analyzer: extracting sentiments about a given topic using natural.

[11] Natural Language Processing, MIT Press, Cambridge, MA, USA,1999, vol.3, pp 12-34.

[12] Empirical Methods in Natural Language Processing (HLT/EMNLP), Vancouver, pp. 347–354