International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 3 Issue: 4 2374 – 2376

Effective Utilization of Search Engine Algorithm in Relevant Information Retrieval

Jayashri D.Nagpure , Aishwarya R.Gayake, Jyoti S.Danave , Ravindra S.Tambe

Abstract -World Wide Web contains large database and to find the relevant data it is very difficult. To get more relevant data user uses search engine .Search engine use by several different ranking algorithm. Different Ranking algorithms are used to retrieve the more relevant and popular web pages and information. HITS (Hyperlink-Induced Topic Search) and page ranking algorithms are very popular that's why we use this algorithm in our proposed system.This paper proposed a system in which we merged both these page rank and HITS algorithm to find out more required information and newly created most popular web pages. The result of proposed system is better than existing individual algorithms.

Keywords- , Page Rank, Web structure, Hub, Authority, Link, Search Engine.

*****

I. INTRODUCTION

In Our Today's Generation World Wide Web is most popular and faster way of communication. Every day lots of website are newly created and many newly created web pages added in World Wide Web network.
This faster growing technology creates new challenges for information or data searching. Search engine is used to retrieve or search information about any topic on WWW.
Web search engine is responsible for to give relevant web pages of users query.

Search engine shows the webpages related to the user query in a ranked order using different ranking algorithms.

Page ranking is a Link analysis algorithm used by search engines for ranking thousands of web pages in a relative order of importance [1].
Hypertext Induced Topic Search() or hubs or authority is a link analysis algorithm.[2]

In Search Engine ,primarily focuses on link structure of the Web to find the importance of the web pages, this paper focuses on two link-base approaches for calculating rank:

- PageRanking algorithm
- HITS (Hyperlink Induced Topic Search)

These are two most popular link-based ranking algorithms.
These algorithms use the hyperlink structure of webpages, analyses it and then find the rank for web page which shows relevancy of webpage to a particular topic. The goal of this paper is to present a new algorithm which can resolve drawbacks of existing ranking algorithm. This paper is contains following sections , In Section 2 & Section 3 brief introduction of Page Rank & Page HITS is given. In Section 4 Proposed Approach is given.
& in Section 5 Comparative study is given.

II. PAGE RANK ALGORITHM

PageRank algorithm is one of the link analysis algorithm.
This algorithm is used by one of the best search engine Google[4].Firstly this algorithm calculates page rank of web pages. Web pages are maintain in hyperlinked graph structure in database called as World Wide Web. This algorithm gives rank to pages and according to ranking of pages importance to web page is given.As pages are linked with each other then it causes transfer of importance This algorithm uses linked page,hence when low ranked page is linked with another page at that time rank of that page also becomes low. Backlinks and Inlinks are used to calculate page rank.
To calculate PageRank formula is [4]

$$PR(A) = (1-d) + d(PR(T1)/C(T1) + ... + PR(Tn)/C(Tn)) \quad (1)$$

In eq.(1) the PageRank of page A is calculated where,
PR(A) = PageRank of Web Page A.
T1.....Tn =Web Pages that points to Page A (Inlinks of A)
C(A) = No of outlink of webpage A
d = Damping factor its value is set between 0 and 1
usually set to 0.85 for the web graph.

When user click on webpage during serfing then pagerank shows probability of the user arriving at that page. To reduce probability here we use Damping factor *d* having value in between 0 to 1 refer eq.(1)
To calculate PageRank simple iterative algorithm is used.
.PageRank by iterative method is very simple and easy method for small set of Web Pages but when we use large number of pages at that time it is not feasible.

Drawback of PageRank Algorithm:

- PageRank is calculated with the help of authority only and this algorithm does not distinguish between hubs and authority.
- It only give more importance to old WebPages.
- This algorithm uses backlink only.

III. HITS ALGORITHM

Hypertext Induced Topic Search (HITS) or hub and authorities is a link analysis algorithm developed by Jon Kleinberg [2] in 1998 to rate web pages. HITS algorithm is classified into two sets called Hub and Authorities. Hub and Authorities are both the web pages and that have basically Outlink and Inlink. Hubs have the out links that's pointing to another important web pages and Authorities that have Inlinks coming from WebPages.
This algorithm iteratively finds hub and authority score for a webpage, a web page can be serve as hub or authority at the same time hence they can have hub or authority score simultaneously [1]. Hubs and Authority are shown in figure 1

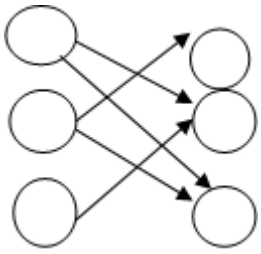

Figure 1 HUB Authorities

Fig. Hub and Authority Link Structure for a webpage p, hub and authority score is calculated by eq. (2) & eq. (3)
[1]
Authority updates Rule:
∀ p, we update auth (p) to be:

$$\sum_{i=1}^{n} \text{hub}(i) \qquad (2)$$

Where n is the total number of pages link to p.
i is a page linked to p.
Authority score of a page is the sum of all the Hub score of pages that point to it [1].
Hub updates Rule:

∀ p,we update hub (p) to be:

$$\sum_{i=1}^{n} \text{auth}(i) \qquad (3)$$

Where n is the total number of pages p links to.
i is a page which p links to.
Thus a pages Hub score is the sum of the Authority scores of all its linking pages[1].
The value of hub and authority score is initially set to 1.

Drawback of HITS Algorithm :

- HITS algorithm is query dependant algorithm.
- This algorithm is very difficult to distinguish hub and authorities.
- It is time consuming algorithm.

IV. PROPOSED APPROCH

In proposed system tries to overcome disadvantages of PageRank and HITS algorithm by applying merged algorithm. This approach gives more relevant web search than indivisual algorithm.So it gives better result than other algorithm and it reduces time required to get required data.
This system is nothing but comparative study of PageRank and HITS.
Above table tells about the actual system.

PageRank Score	Hub Score	Category
High	High	Most relevant pages
High	Low	Relevant pages
Low	High	Weakly relevant pages
Low	Low	Not relevant pages

Advantages of proposed system:

- It Avoids query dependancies in HITS algorithm .
- In PageRank algorithm older pages get more importance but in this system this disadvantage is removed.
- Topic Drift in HITS algorithm is avoided in this approach.

V. COMPARATIVE STUDY OF EXISTING AND PROPOSED SYSTEM [2].

Parameter	HITS	Page Ranking	Proposed System
Relevancy	More	Less	More
Quality of Result	Less	More	More
Technology used	Web structure, web Content	Web structure	Web Structure, Web Content
Input Parameter	Back link, Forward link, Content	Back link	Back link, As well as forward link
Query Dependency	Query dependant	Query independent	Query independent

VI. CONCLUSION

This paper consist of PageRank algorithm and HITS algorithm for search engine.In this paper these algorithms are studied in more detailed manner.Also drawbacks of these algorithm are studied and tried to overcome.Finally we come to know that after merging of these algorithm we get better result instead of using individual algorithm.

REFERENCES

[1] Shailendra G. Pawar, Pratiksha Natani Effective Utilization of Page Ranking and HITS in significant Information Retrieval International Conference on Convergence of Technology 2014

[2] Pooja Devi,Ashlesha Gupta and Ashutosh Dixit Comparative study of HITS and PageRank Link Based Ranking Algorithms International Journal of Advance Research in Computer and Communication Engg vol 3,issue 2,Feb 2014

[3] Nidhi Grover and Ritika Wason, Comparative Analysis of PageRank and HITS Algorithms International Journal of Engg Research & Tech. Vol. 1 Issue 8, October - 2012 ISSN: 2278-0181.

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