An Improvement of Link Analysis Algorithm to Mine Pertinent Links: Weighted HITS Algorithm based on additive fusion of graphs by Query Similarity

Hemangini S. Patel
Bhagwan Mahavir College of Computer Application (BCA)
Bharthana, Vesu,
Surat, India

Apurva A. Desai
Department of Computer Science,
Veer Narmad South Gujarat University,
Surat, India

ABSTRACT

In recent days, link analysis has been found to increase the performance of web search significantly to extract pertinent links which are valuable. Generally, short term queries matches to link anchors and titles. We give a weighted Input to HITs and proposed algorithm weighted HITs (WHITs) in which the adjacency matrix is weighted double if link anchors and titles are matched with query term by additive fusion of graphs. Experimental results provided evidences that weighted input to HITs (WHITs) returns unique rankings for authoritative pages, for link anchors and link titles which are similar to query term. The proposed algorithm, namely weighted HITs (WHITS) helps to extract a pertinent and valuable links based on similarity of link anchors, titles, and query term.

General Terms

Information Retrieval, Link Analysis, HITs, weighted HITs.

Keywords

Web Mining, Web Structure Mining, Information Retrieval, Link Analysis, Anchor Text, WWW.

1. INTRODUCTION

Information retrieval becomes a challenge, as search engine database contains the huge bulk of data. Ranking is the main aspect of any information retrieval system. The main aim of ranking algorithm is to discover the highest ranked authority pages in the large collection of the pages. With web novel resources of information became accessible, one of them being the hyperlinks between pages. For web information retrieval, a hyperlink offers valuable resources of information [1]. Pertinent and excellent documents are ranked by Link analysis ranking algorithms. The present research presented improved link based ranking algorithm weighted HITs (WHITs) to enhance the quality of web search effectively. However, existing link analysis ranking algorithms assigns every link by the equal weight. These hypothesis outcomes in topic drift due to some non-relevant pages tightly interconnected densely. The primary hypothesis is that the Web is a plane graph, where the entire pages are equal and their significance is found by only the link associations. Since, the ranking of pages assigned by HITs algorithm just by using the in-links and out-links of links, it will be inconvenient in several cases. This difficulty can be improved by rising weights towards edges based on text within the pages or their anchors [2, 3]. To increase the accuracy of HITs algorithm weight adjustment of a link by considering techniques such as content investigations [2, 3], user opinion [4], or Web log records [5], has been known to be proficient techniques.

It has been verified that to get better accuracy, merging link and content investigations is very helpful. To estimate the relevance of document, the likeness is computed in vector space model [2] or by considering frequency of occurrence of the query terms within the text surround to the link anchors [3]. It might possibly be the finest technique as the majority of the queries sent to search engines are shorter, two or three terms or less than three terms.

HITs algorithm is most important for mining link structure for web search as it is query reliant. It determines the importance of pages among esteem to a specified query. In the present study, relative effectiveness of the ranking algorithm, namely, HITs [6], is evaluated with further algorithm proposed namely weighted HITs.

The link weight is computed based on the connection of its adjacent pages and based on the likeness of every page towards the query. Moreover, the additional trendy pages are further associated with other web pages which are likely to direct to them or are directed to by them. In this paper, we present how a webpage is similar to query by acquiring a broad view that the query induced similarity and suggested web pages by considering target page and query and its similarity with anchor text and titles of links to discover “in-formativeness". Then, we present enhanced weighted HITs algorithm via handing over suitable weights towards links among the likeness of link anchors, titles with user query. Also, WHITs algorithm is used to builds a new adjacency matrix to calculate authorities and hubs. According to experimental results, it is demonstrated that our enhanced weighted HITs technique outperforms considerably superior than HITs and removes the problem of topic drift efficiently to some extent.

The paper is ordered as follows in rest of sections. In 2nd Section, concise background review of assigning weights to links in various algorithms. In 3rd Section, we describe the existing link analysis web page ranking algorithms. We propose to merge hyperlink graph and anchor, title similarity graph jointly to refine HITs algorithm and experimental result and details in 4th Section. Further, we give the concluding remarks and future work in 5th Section.

2. BACKGROUND

PageRank [4, 5] and HITs [6] algorithms meant for ranking web search results. There are various attempts to improve better effectiveness of link analysis algorithms. The ARC (Automatic Resource Compilation) [7] and the average [8] ranking algorithms are proposed based on hyperlinks and content in order to manage topic drift.

Nomura et al. [9] have proposed two various techniques which make use of link-only in turn of the Web to solve the topic drift problem of HITs: (i) the projection technique, in which most pages inside the root set will be pertinent to main topic due to it projects eigenvectors on the root set, and (ii) the base set scaling down technique, in which it filters away the pages with no links to numerous pages inside the root set. Yan et al. [10] performed page importance analysis with site organization constraints by developing three novel algorithms: optimization-based algorithm, additive flat and tree graph fusion algorithm, and multiplicative graph fusion algorithm. An experiment on TREC2003 demonstrates the topic distillation task by all of novel algorithms outperformed the benchmark Page Rank algorithm.

Lempel and Moran projected the SALSA algorithm [11] to assign weights by considering together rows and columns weights to calculate its hub scores as well as authority scores. Cohn and Chang [12] proposed a model based on the probability known as PHITS algorithm. In 2001, Borodin et al. proposed the HUB-Averaging (HUBAVG) algorithm by considering average weight of authority by all authorities pointed to by hub to assign hub weight. Later, Borodin et al. proposed the Authority-Threshold algorithm [13] by considering summation of k major authority weights of authorities directed to by hub to assign hub weight.

It was experimented by Jin, Hauptmann, and Zhai [14] that there is a close similarity among page titles and queries, and that they are formed by alike intellectual method. Thus it is innate to imagine that together titles and anchor text imprison a few concept of what a page is in relation to, even if they are linguistically unlike [15].

Yue et al. [16] introduced the information of web click and timeliness to improve HITs by improved algorithms including improved HITs with click-through rate (CTR). The influence weight of page's age considered and integrated with CTR and the influence weight of page's age were experimented and shown that the improved algorithms had a certain advantages over the original algorithm to reduce topic drift.

Jaganathan and Desikan [17] introduced weight based page rank algorithm stands on in-link and out-link, and noticed that their algorithm is additionally efficient while judged with the page rank algorithm with admire towards time. Xing and Ghorbani [18] anticipated on the computation of the weight of the page with the concern of the out links, in links and allocate rank scores on the basis of the reputation of the pages.

Ricardo and Davis [19] presented WLRank algorithm that considers various Web page attributes to provide additional weight to several links to rank i.e. relative location in page, tag anywhere link is contained and distance end to end of anchor text rather than providing uniform link weights. Hebert et al. [20] suggested an authority score calculation technique that considers the associations existing along with dissimilar skills such as LINKEDIN and RESEARCHGATE permit user endorsements for particular skills. This technique is stand on inspiring the information enclosed in the digraph of endorsements related to a particular skill, and then performing a ranking method like PAGERANK to declare weighted digraphs. Benzi et al. [21] unambiguously find out the exponential of original adjacency matrix, directed network, and provide an understanding of centrality and communicability in new perspective. It is most important technique for ranking hubs and authorities.

Based on the existing defects of PageRank algorithm in the application, Luo and Xue [22] proposed a distribution of weights based on ant colony optimization search engine link scheduling model of PageRank algorithm. The experimental simulation results showed that the proposed algorithm is superior to the traditional PageRank algorithm in terms of accuracy and recall rate. Andri and Masashi [23] expands the thought of HITs by authority and hub scores via establishing two diagonal matrices which holds fixed values that operate as weights to build authority pages more authoritative and hub pages more important hub.

Hamed [24] on the other hand introduced Randomize HITs which converges faster than the other algorithms especially it is better than PageRank Algorithm in converging and stability features. Hung et al. [25] improved the HITS algorithm using STPs (Semantic Text Portion) for weighting each link to assign bigger weight to link for identifying authority pages and contrast STPs among anchor-based texts of further types. In aspect, they contrast STPs with techniques utilizing the: (i) descriptive text anchor, (ii) text contains within the subsection straightly contains the anchor, (iii) text contains within the rigid-window of 50 terms in the anchor region, and (iv) text contains in the entire superior-level headers of the anchor. Tiana et al. [26] – proposed improved HITS algorithm based on the theory of triadic closure and VSM. According to page topic likeness and general reference degree, this technique initially calculates the relevance among random two pages. Then, by utilizing the relevance, a novel adjacency matrix is build to iteratively compute authorities and hubs. Preliminary experiments showed that new algorithm which improves the efficiency and quality of query, reduce the theme drifts. Our study is based on link analysis ranking based modified HITs version which collects anchors of out-links, titles of in-links, and double weighting links if anchors and titles matched with query term, known as weighted HITs (WHITs).

3. PROPOSED ALGORITHM:

In the HITs algorithm, for a specified query term Q, a set of t highest ranked pages is selected which is known as root set. From this, base set S is constructed by comprising any page directed to by a page and any page directs to a page. The adjacency matrix L of the directed Web graph can be defined in HITs is as.

Lij = {1, if an edge exist between node i towards node j,

0, else.

HITs algorithm finds authority pages by handover two statistics to page authority weight and hub weight. It assumes that all links are equal and assigns equivalent weight to every page which results in topic drift.

3.1 ADDITIVE GRAPH FUSION FOR IMPROVEMENT TO HITS ALGORITHM

Analogous to the hyperlink connectivity graph, we can build a similarity of anchors, titles and query graph. In this way, we can obtain two graphs. Let L and L1 signify the adjacency matrix of the connectivity of hyperlink graph and similarity of anchors, titles and query graph separately. To integrate the similarity of anchors, titles, and query graph in turn with hyperlink connectivity graph for Web page ranking, it is required to fuse these two flat graphs for the improvement of HITs algorithm. One of the simplest fusion techniques to adjoin the two flat graphs directly to fuse the adjacency matrix L and L1 to get a novel adjacency matrix W;

$$W = L + L1 \qquad (1)$$

Modification to HITs is weighted HITs in which it needs to create an adjacency matrix W by, double-weighting of links i.e. 2, if anchors (links and outgoing links) and titles (incoming links) in base set contain the query term and rest to 1, if connection exist but not match with query term, otherwise 0, then calculate hubs and authorities.

$$Wij = Lij + L1ij \qquad (2)$$

Where,

Lij = 1, if an edge exist between node i towards node j,

L1ij = 1, if anchor text and title matched with query term,

0, otherwise.

For the recent graph with adjacency matrix W which is weighted input, we can go after the standard HITs algorithm to calculate the authority and hub scores for every page within the Web graph. We describe it by additive flat graph fusion for HITs algorithm namely, weighted HITs (WHITs).

Hence, in order to decrease the computation complication, calculating the likeliness of the destination page and the query term Q is based on calculating the likeliness of the anchor text of links along with out-links as well as titles of in-links with the query Q and then it calculates authority scores and hub scores using the weighted matrix, and compared to HITs algorithm. So it will ultimately increase weights of links which are mostly not self-descriptive. For a given page i in S, a weighted authority score Wa(i) and weighted hub score Wh(i) are assigned using weighted matrix;

$$Wa(i) = \sum_{(j,i)\in E} Wh(j) \qquad (3)$$

$$Wh(i) = \sum_{(i,j)\in E} Wa(j) \qquad (4)$$

3.2 Weighted Hits (Whits) Algorithm

WHITs algorithm is also work for the eigenvalues calculation. A HITs derives the hub and authority matrices form adjacency matrix L and the transpose of matrix L (i.e. LT) where as in weighted HITs, matrix W is considered which increases the weight of links from single to double (i.e. 2) if anchor text, title, and query Q are matched, otherwise 1, and if no linkage than 0. WTW is an authority matrix used to discover an authority vector. Matrix WWT is a hub matrix used to discover a hub vector. An authority vector and a hub vector are eigenvectors corresponding to highest eigenvalue of the authority matrix and the hub matrix respectively. The pseudo code of weighted HITs (WHITs) algorithm is shown below;

```
1. Initiate the entire weights to double if anchors and titles contain the query term and rest towards 1.
2. Reiterate until the double weights converge:
3. For each hub i ∈ H
       $Wh(i) = \sum_{(i,j)\in E} Wa(j)$
4. For each authority i ∈ A
       $Wa(i) = \sum_{(j,i)\in E} Wh(j)$
5. Normalize
```

Start with to find out main authority by maximum hubs pointing to it by highly weighted links.

4. EXPERIMENTS

4.1 DATA SET

In our experiments, 5 short term queries and two existing search engines Google and Bing are used. For every query, to construct a base set, we underway two threads concurrently to gather around 100 highest ranked nodes (t) known as root set and their neighbourhood are used to constructs the base set. By using nodes of Root set collection of out-links, anchors of out-links and in-links, and titles of in-links are performed to build base set (S). Duplicate or intra-domain links are then removed. On this web graph, HITs, weighted HITs algorithms are applied to check the relative effectiveness of ranking orders. The base sets used for query (Q) are build, as shown by Kleinberg and numerical statistics are shown in Table 1.

Table 1: Experimental data for various queries
Experimental data for various queries

Query (Q)	Root Set	Out-links	In-links	Links	Base Set (S)
Java	102	11546	191	13560	10806
Jaguar	102	16527	744	17373	12711
Harvard	95	27243	427	31609	13192
Search Engine	100	8264	227	10637	9152
Kyoto University	94	6393	700	7187	6070
Toyota	107	9116	497	9720	7802

4.2 Experimental Results

The calculation of HITs is performed on our data set as described and compared with weighted HITs (WHITs).

Results obtained are shown in Table 2 for top 10 authority ranks for a queries ”Java”, ”Jaguar”, ”Harvard” ,“Search Engine”, and ”Kyoto University”. The following Table 2 shows the results that are labelled highly authoritative and their weight are increased appear in boldface. It is observed that, the ranking of nodes is increased whose weights are doubled by considering the anchors of out-links, titles of in-links and query similarity.

According to the idea of P@10 method [27], the pertinent pages are acquired first, followed by the satisfaction evaluation of the pertinent pages. The users’ satisfaction is a deeper with the relevance. It demands not only the relevance to topic, but also the containing of the latest authority information. We have considered the authority weights used for the web pages in the web graph by utilizing the adjacency matrix related general HITs algorithm and proposed the weighted HITs (WHITs) algorithm. In experiment top 10 authority weights in Table 2, i.e. top authoritative pages which describe the major search engines for query “search engine” are listed in top 10 authorities using weighted HITs (WHITs) algorithm.

As shown in table 2, Weighted HITs (WHITs) increased the weight of the pages which are authoritative and provided almost all available search engines. In similar way, it increases a weight of pertinent links for queries as shown in Table 2. About 5 out of the top most 10 results for query “Java” are related, 2 of the top most 10 results for query “Jaguar” are related and 9 of the top most 10 results for query “Harvard” are related to top authoritative pages which describe the major search engines, using weighted HITs (WHITs) Algorithm. All 6 of the top most 10 results are related for query “Kyoto University”.

Table 2: Top ten Authorities and weighted Authorities for queries “Java”, ”Jaguar” ,“Harvard”, “Search Engine” and “Kyoto University”

Table 2 Top ten authorities and weighted authorities for queries “java”.

HITs	
0.3490	https://plus.google.com
0.2510	http://www.oracle.com/technetwork/java/index.h
0.2071	http://www.youtube.com
0.1885	http://www.oracle.com
0.1698	http://java.com
0.1661	http://www.facebook.com
0.1605	http://www.oracle.com/technetwork/java/javase/ downloads/index.html
0.1592	https://twitter.com
0.1573	http://twitter.com
0.1530	https://www.oracle.com
WHITs	
0.3221	http://www.oracle.com/technetwork/java/index.h
0.2997	http://www.oracle.com
0.2663	http://java.com
0.2590	http://www.oracle.com/technetwork/java/javase/ downloads/index.html
0.2262	https://www.oracle.com
0.2147	https://cloud.oracle.com
0.2147	http://www.java.net
0.1871	https://community.oracle.com
0.1841	http://education.oracle.com
0.1757	https://blogs.oracle.com

Table 3 Top ten authorities and weighted authorities for query ”jaguar”.

HITs	
0.2188	http://www.jaguarusa.com/index.html
0.1582	http://www.jaguar.co.uk/index.html
0.1507	http://www.jaguar.com/index.html
0.1434	http://www.jaguar.com.au/index.html
0.1390	http://www.jaguar.ie/index.html
0.1384	http://www.jaguar.in/index.html
0.1361	http://www.jaguar.co.za/index.html
0.1177	http://www.jaguar.com
0.1086	http://jaguar.pl
0.1071	http://www.jaguar.com.my
WHITs	
0.6969	http://www.jaguarusa.com/index.html
0.6147	http://www.jaguarusa.com/
0.1449	http://www.jaguar.com/index.html
0.1360	http://www.jaguar.co.uk/index.html
0.1144	http://www.jaguar.co.za/index.html
0.1141	http://www.jaguar.com.au/index.html
0.1113	http://www.jaguar.in/index.html
0.1023	http://www.jaguar.ie/index.html
0.0572	https://twitter.com
0.0366	http://instagram.com

Table 4 Top ten authorities and weighted authorities for queries “harvard”.

HITs	
0.3300	http://twitter.com
0.2914	https://twitter.com
0.2682	http://www.harvard.edu
0.2634	https://www.facebook.com
0.2239	http://www.harvard.edu/
0.2118	https://plus.google.com
0.2087	http://www.facebook.com
0.1850	http://www.youtube.com
0.1816	http://www.linkedin.com
0.1639	http://news.harvard.edu
WHITs	
0.3306	http://www.hbs.edu/
0.3048	http://hms.harvard.edu/
0.2864	https://www.hsph.harvard.edu/
0.2672	https://www.hms.harvard.edu/
0.2545	http://www.gsd.harvard.edu/
0.2500	http://alumni.harvard.edu/
0.2409	https://college.harvard.edu/
0.2284	https://www.gocrimson.com/
0.1856	https://library.harvard.edu/
0.1611	http://www.harvard.edu/

Table 5 Top ten authorities and weighted authorities for queries “search engine”.

HITs	
0.1199	http://www.google.com
0.1176	http://www.bing.com
0.1129	http://www.ask.com
0.1083	http://www.yahoo.com
0.1008	http://www.lycos.com
0.0975	http://www.facebook.com
0.0960	http://www.ixquick.com
0.0960	http://www.webcrawler.com
0.0929	http://www.galaxy.com
0.0929	http://www.excite.com
WHITs	
0.1209	http://www.google.com
0.1191	http://www.bing.com
0.1136	http://www.ask.com
0.1088	http://www.yahoo.com
0.1011	http://www.yahoo.com
0.0987	http://www.facebook.com
0.0971	http://www.ixquick.com
0.0962	http://www.webcrawler.com
0.0931	http://www.excite.com
0.0931	http://www.galaxy.com

Table 6 Top ten authorities and weighted authorities for queries “kyoto university”.

HITs	
0.7126	http://www.kyoto-u.ac.jp/en
0.6204	http://www.kyoto-u.ac.jp/en/
0.1158	http://www.kyoto-u.ac.jp
0.0740	http://www.opir.kyoto-u.ac.jp
0.0523	http://www.kyoto-u.ac.jp/en/faculties-and-
0.0523	http://www.oc.kyoto-u.ac.jp/en/

0.0513	http://twitter.com
0.0499	http://www.asafas.kyoto-u.ac.jp/en/
0.0494	https://www.facebook.com
0.0440	http://www.med.kyoto-u.ac.jp
WHITs	
0.7216	http://www.kyoto-u.ac.jp/en
0.5809	http://www.kyoto-u.ac.jp/en/
0.2267	http://www.kyoto-u.ac.jp
0.1037	http://www.opir.kyoto-u.ac.jp
0.0866	http://www.opir.kyoto-u.ac.jp/kuprofile/
0.0473	http://www.t.kyoto-u.ac.jp/en
0.0452	http://www.kyoto-u.ac.jp/en/faculties-and-
0.0452	http://www.oc.kyoto-u.ac.jp/en/
0.0444	http://sph.med.kyoto-u.ac.jp
0.0437	http://www.t.kyoto-u.ac.jp

Similarly, Table 3 describes the authority weights of HITs and weighted HITs (WHITs) algorithms for links returned by WHITs. It shows that the ranking of nodes is enriched whose weights are doubled by considering anchors of out-links, titles of in-links similar to query term. Relative weights after weighting of links by WHITs as compared to general HITs is outperformed for pertinent links and drastically changed the top 10 authorities itself and increased the weight of pertinent links.

Table 3 : comparison of weights of Top ten Authorities of General HITs algorithm and weights of Top 10 Authorities of Weighted HITs (WHITs) , for queries “Java”, ”Jaguar” , “Harvard”, “Search Engine” and “Kyoto University”

Table 7 Comparison of weights of Top 10 Authorities of WHITs and corresponding Authorities of HITs, for query “Java”

HITs	WHIT	Top 10 Authorities
0.2510	**0.3221**	http://www.oracle.com/technetwork/java/index.html
0.1885	**0.2997**	http://www.oracle.com
0.1698	**0.2663**	http://java.com
0.1605	**0.2590**	http://www.oracle.com/technetwork/java/javase/downloads/index.html
0.1530	**0.2262**	https://www.oracle.com
0.1398	**0.2147**	https://cloud.oracle.com
0.1398	**0.2147**	http://www.java.net
0.1507	**0.1871**	https://community.oracle.com
0.1461	**0.1841**	http://education.oracle.com
0.1101	**0.1757**	https://blogs.oracle.com

Table 8 Comparison of weights of Top 10 Authorities of WHITs and corresponding Authorities of HITs, for query ”Jaguar”.

HITs	WHIT	Top 10 Authorities
0.2188	**0.6969**	http://www.jaguarusa.com/index.html
0.0970	**0.6147**	http://www.jaguarusa.com/
0.1507	0.1449	http://www.jaguar.com/index.html
0.1582	0.1360	http://www.jaguar.co.uk/index.html
0.1361	0.1144	http://www.jaguar.co.za/index.html
0.1434	0.1141	http://www.jaguar.com.au/index.html
0.1384	0.1113	http://www.jaguar.in/index.html
0.1390	0.1023	http://www.jaguar.ie/index.html
0.0932	0.0572	https://twitter.com
0.0545	0.0366	http://instagram.com

Table 9 Comparison of weights of Top 10 Authorities of WHITs and corresponding Authorities of HITs, for query “Harvard”.

HITs	WHIT	Top 10 Authorities
0.1419	**0.3306**	http://www.hbs.edu/
0.1423	**0.3048**	http://hms.harvard.edu/
0.0443	**0.2864**	https://www.hsph.harvard.edu/
0.0751	**0.2672**	https://www.hms.harvard.edu/
0.0785	**0.2545**	http://www.gsd.harvard.edu/
0.0326	**0.2500**	http://alumni.harvard.edu/
0.0657	**0.2409**	https://college.harvard.edu/
0.0697	0.2284	https://www.gocrimson.com/
0.0196	**0.1856**	https://library.harvard.edu/
0.2239	0.1611	http://www.harvard.edu/

Table 10 Comparison of weights of Top 10 Authorities of WHITs and corresponding Authorities of HITs, for query “Search Engine”.

HITs	WHIT	Top 10 Authorities
0.1199	**0.1209**	http://www.google.com
0.1176	**0.1191**	http://www.bing.com
0.1129	**0.1136**	http://www.ask.com
0.1083	**0.1088**	http://www.yahoo.com
0.1008	**0.1011**	http://www.yahoo.com
0.0975	**0.0987**	http://www.facebook.com
0.0960	**0.0971**	http://www.ixquick.com
0.0960	**0.0962**	http://www.webcrawler.com
0.0929	**0.0931**	http://www.excite.com
0.0929	**0.0931**	http://www.galaxy.com

Table 11 Comparison of weights of Top 10 Authorities of WHITs and corresponding Authorities of HITs, for query “Kyoto University”.

HITs	WHIT	Top 10 Authorities
0.7126	**0.7216**	http://www.kyoto-u.ac.jp/en
0.6204	0.5809	http://www.kyoto-u.ac.jp/en/
0.1158	**0.2267**	http://www.kyoto-u.ac.jp
0.0740	**0.1037**	http://www.opir.kyoto-u.ac.jp
0.0440	**0.0866**	http://www.opir.kyoto-
0.0384	**0.0473**	http://www.t.kyoto-u.ac.jp/en
0.0523	0.0452	http://www.kyoto-
0.0523	0.0452	http://www.oc.kyoto-u.ac.jp/en/
0.0330	**0.0444**	http://sph.med.kyoto-u.ac.jp
0.0276	**0.0437**	http://www.t.kyoto-u.ac.jp

Fig1 to 5 shows graphical representation of weights of HITs and WHITs algorithms for our dataset queries. It can be seen that weights of WHITs are increased as compared to HITs algorithm. By applying dual weight to web pages performs well as compared to HITs algorithm.

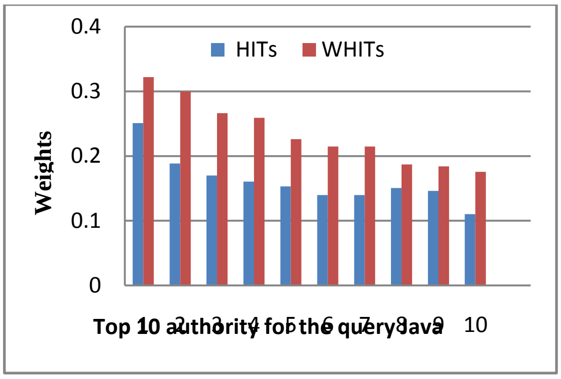

Fig1. Top 10 authority weights for query "Java"

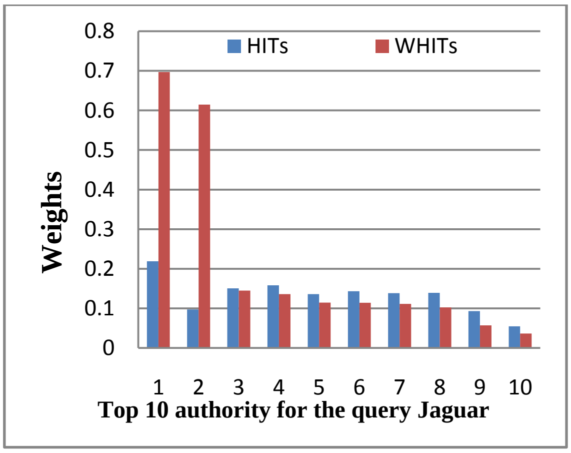

Fig 2. Top 10 authority weights for query "Jaguar"

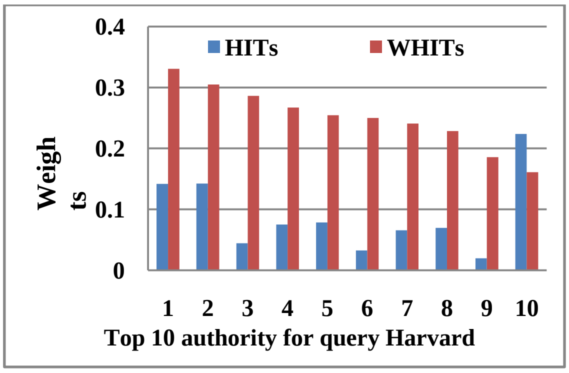

Fig 3. Top 10 authority weights for query "Harvard"

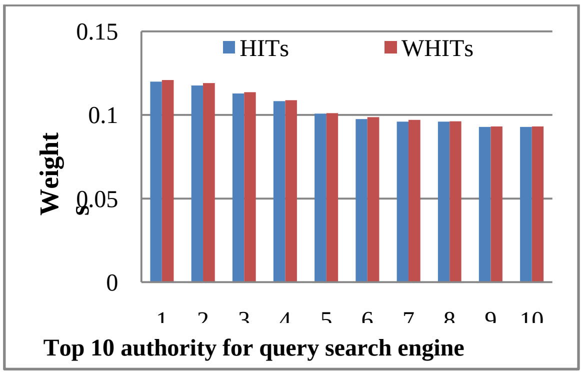

Fig4. Top 10 authority weights for query "Search Engine"

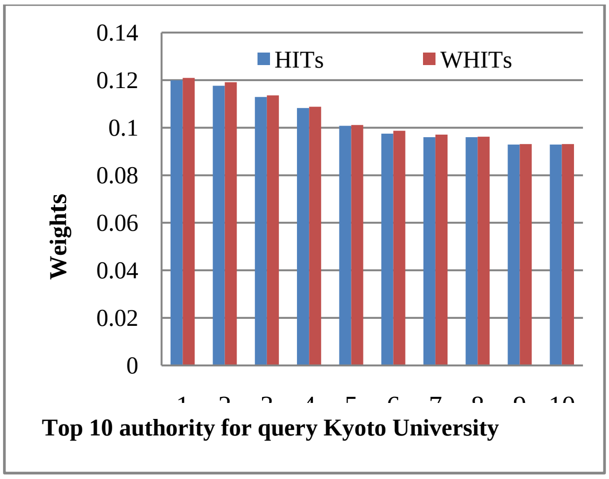

Fig 5. Top 10 authority weights for query "Kyoto University"

5. CONCLUSION AND FUTURE WORKS

In this paper, we implemented improved versions of HITs algorithm, which is called as WHITs and showed enhanced ranks than HITs. It has been verified that to get better accuracy, merging link and content investigations is very helpful. The aim and contributions of this present study were to discover the link analysis algorithms for ranking and utilization of anchor text and titles in IR. Experimentation with the double weighting of links of nodes which matches with query and anchors of out-links and titles of in-links in Weighted HITs (WHITs), outperforms for authority pages which are not probably self-descriptive as compared to general HITs algorithm. Hence, how anchor text and titles can be utilized to get better search superiority in authority finding is shown by assigning different weights to links. By considering anchor texts and titles, one can improve Web search rankings, especially for the pages which are not self-descriptive and queries which are short terms and generally matched with anchor texts. Also, it is helpful for homepage discovery, named page discovery, navigational queries and ad hoc search tasks.

6. REFERENCES

[1] M. Henzinger, "Link analysis in web information retrieval," IEEE Data Engineering Bulleitin, 1-6, 2000.

[2] S. Chakrabati, B. Dom, D. Gibson, J. Kleinberg, S. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, "Mining the link structure of the World Wide Web," IEEE Computer, 32, no. 8, 60-67, 1999.

[3] K. Bharat and M. R. Henzinger,"Improved algorithms for topic distillation in a hyperlinked environment," In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 104-111, 1998.

[4] S. Brin and L. Page, "The anatomy of a large-scale hyper textual Web search engine," Computer Networks and ISDN Systems, 30(1–7): 107–117, 1998.

[5] L. Page, S. Brin, R. Motwani, and T. Winograd, "The page rank citation ranking: Bringing order to the web," 1999.

[6] J. M. Kleinberg, "Authoritative sources in a hyperlinked envi-ronment," Journal of the ACM (JACM). 46, no. 5,

604–632, 1999.

[7] S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, and J. Kleinberg, "Automatic resource compilation by analyzing hyperlink structure and associated text," Computer Networks and ISDN Systems, 30 no. 1, 65-74, 1998.

[8] J. Gevrey, and S. Ruger, "Link-based Approaches for Text Retrieval," Proceedings of TREC-10, NIST Special Publication, 2002, 279-285, 2001.

[9] S. Nomura, S. Oyama, T. Hayamizu and T. Ishida, "Analysis and improvement of HITS algorithm for detecting Web communities," Systems and Computers in Japan, IEEE, 32-42, 2004.

[10] H. M. Yan, T. Qin, T. Y. Liu, X. D. Zhang, G. Feng, and W. Y. Ma, "Calculating webpage importance with site structure constraints," In Information Retrieval Technology. Springer Berlin Heidelberg, 546-551, 2005.

[11] R. Lempel, and S. Moran, "The stochastic approach for link-structure analysis (SALSA) and the TKC effect," Computer Networks. 387–401, 2000.

[12] D. Cohn, and H. Chang, "Learning to probabilistically identify authoritative documents," In Proceedings of the 17th International Conference on Machine Learning (ICML), Stanford University, United States, 167-174,2000.

[13] A. Borodin, G. O. Roberts, J. S. Rosenthal, and P. Tsaparas, "Finding authorities and hubs from link structures on the world wide web," Proceedings of the 10th International World Wide Web Conference, ACM, 415–429, 2001.

[14] J. Rong, A. G. Hauptmann, and C. X. Zhai, "Title language model for information retrieval," In Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval, Tampere, Finland, Association for Computing Machinery, ACM, 42-48,2002.

[15] N. Eiron and K. S. McCurley, "Analysis of anchor text for web search," In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, 459–460, 2003.

[16] Y. He, M. Qiu, M. Jin, and T. Xiong, "Improvement on HITS Algorithm," Applied mathematics and information sciences, 6, no. 3, 1075-1085, 2012.

[17] B. Jaganathan, and K. Desikan, "Weighted Page Rank Algo-rithm based on In-Out Weight of Webpages," Indian Journal of Science and Technology 8, no. 34, 1-6, 2015.

[18] W. Xing and A. Ghorbani, "Weighted PageRank Algorithm," in proceedings of the 2rd Annual Conference on Communication Networks & Services Research,IEEE, 305-314, 2004.

[19] R. Baeza-Yates and E. Davis, "Web page ranking using link attributes," In proceedings of the 13th international World Wide Web conference on Alternate track papers & posters, ACM, 328-329, 2004.

[20] H. Pérez-Rosés, F. Sebé, and J. M. Ribó, "Endorsement deduction and ranking in social networks," Computer Communications, 73, 200-210, 2016.

[21] M. Benzi, E. Estrada, and C. Klymko, "Ranking hubs and authorities using matrix functions," Linear Algebra and its Applications, 438, no. 5, 2447-2474, 2013.

[22] X. Luo, and H. Xue, "Weights Allocation Optimization of Search Engine Links Sorted Pagerank Algorithm," 355-360, 2015.

[23] A. Mirzal and M. Furukawa, "A Method for Accelerating the HITS Algorithm," Journal of Advanced Computational Intelligence, 1-10, 2009.

[24] H. Hamed, "Link Analysis and web page ranking algorithms,". 1-8, 2015.

[25] B. Q. Hung, M. Otsubo, Y. Hijikata, and S. Nishida, "HITS algorithm improvement using semantic text portion," Web Intelligence and Agent Systems: An International Journal, 8, no. 2, 149-164, 2010.

[26] X. Tiana, Y. Dua, W. Songa, W. Liua, and Y. Xieb, "Improve-ments of HITS Algorithm Based on Triadic Closure," Journal of Information & Computational Science, 1861–1868, 2014.

[27] J. Thom and F. Scholer, "A comparison of evaluation measures given how users perform on search tasks," In ADCS2007 Australasian Document Computing Symposium, RMIT University, School of Computer Science and Information Technology, 100-103, 2007.

Hemangini S. Patel completed her graduation and post graduation from Veer Narmad South Gujarat University in the Information Technology. She has completed her Ph.D. from Veer Narmad South Gujarat University in the field of computer science. She has been with the Bhagwan Mahavir College of Computer Application, Surat, Gujarat, India since 2008, as Assistant Professor. She has 12 years of teaching experience since 2006 at under graduate level. Her fields of interest are Data mining, Link Analysis, Web structure mining and web mining.

Apurva A. Desai completed his graduation and post graduation from Veer Narmad South Gujarat University, Surat, Gujarat, India. He earned his Ph.D. in the year 1997 in the joint fields of Operation Research and Computer Science. He has got a long teaching and research experience since 1990. He has published many research papers at national and international level and four books to his credit. Dr. Desai is a Dean of faculty of Computer Science and Information Technology and Chairman Board of Studies (Computer Science). He has attended many International and National conferences. He is and Editor in Chief of VNSGU Journal of Science and Technology and also serving as a member of Editorial board for some of the national and international journals. His fields of interest are Optical Character Recognition, Image Processing, Computer Graphics, Data Mining and Soft Computing.