Analysis of Web Page links and Web Content Similarity using Hits Algorithm

X. Leela Mary, G. Silambarasan

Abstract: HITS (Hyper Link-Induced Topic Search) are a classical link analysis algorithm for analyzing WSM (Web Structure Mining). The algorithm takes into consideration of the structural information of links but ignores the correlation between pages and topics. In some cases, the problem of "topic drift"-a deviation between search and topic-would appear. For this purpose, the current paper presents an improved algorithm, by taking into account both of the web content similarity and link analysis. Our experiment shows that the improved algorithm has enhanced the correlation of search results and limited the occurrence of topic waft to some degree.

Keywords: HITS algorithm; Web content similarity; Authority page; Hub page

I. INTRODUCTION

With the rapid development of information technologies, internet begins to pervade our daily life and has become into a new life style that enriches people's living contents. Search engine a crucial part of the internet, is an important tool for us to acquire information. In searching process, how to find and download pages that are most relevant to users' topics has now become the key for the topical search engine. At present, there are two common types of searching strategies .The first is a content evaluation-based searching strategy such as Fish-Search and Best-First The content describes topics accurately and thus the relevancy between them can be calculated well and truly. This type of strategy, however, ignores structural information of links. Hence, it has disadvantages when forecasting the accuracy of link values and finding out the direction of the focused crawler. The second type of searching strategy is based on evaluation of the structure of web links. Representatives of this type are Page Rank and HITS. This strategy determines the importance of a page by analyzing the reference relationship between pages, so as to determine the crawling sequence of the focused crawler on the page. It takes into consideration of structural information of links, though ignores the correlation between page and topic. In some cases, a HITS algorithm would result in the topic drift that a deviation exists between search and topic; while a Page Rank algorithm is more suitable for authority pages rather than topic resources. In order to solve the problem of topic drift in the HITS algorithm, this paper combines the relevancy of web content with the authority of link analysis to present an optimized strategy that can improve the accuracy of topic searching.

II. HITS ALGORITHM

A. Principle of the HITS algorithm

HITS (Hyper Link—Induced Topic Search) is an algorithm for analyzing WSM (Web Structure Mining). The algorithm is to find out the authority pages and hub pages from a set of web pages, according to user query and through analyzing the forward and backward linkages. An authority page is an authoritative page most relative to the query topic(authority is of influence and is accepted by most of people); while a hub page is a web page that points to the link set of authoritative pages . Each page requires two measurement values: authority weight and hub weight, according to which the importance of a page to a specific topic can be judged.

B. Process of the HITS algorithm

Many algorithms, including the HITS, are based on hypothesis. The HITS algorithm uses two basic hypotheses:

Hypothesis 1: a good "authority page" is linked by many good "hub pages", and;

Hypothesis 2: a good "hub pages" points to many good "authority pages".

Hypothesis 1 describes what a "good" authority page is: a page linked by many good "hub pages". Hypothesis 2 describes what a "good" hub page is: a page points to many good "authority pages". Form above basic hypotheses we can see a mutually reinforcing relationship between the hub and authority page, i.e. the higher quality of a hub page, the better the authority page will be, and vice versa, if an authority page has higher quality, then the hub page that points to the authority page has a higher quality. Based of such a mutually reinforcing relationship, those hub and authority pages that have highest quality can be figured out through iterative computations. The HITS algorithm can be divided into following steps

(1) Input a keyword to the searching engine, to retrieve the top n (most relevant) pages to the search query. This set is called the root set (R) and is required to meet following three conditions:

1. R has a relatively small number of pages;
2. Most of the pages in R is relevant to the query keyword q, and;
3. R contains many authority pages.

(2) A base set (B) can be generated by augmenting the R with all the web pages that are linked from it and some of the pages that link to it. The rule for augmenting is: add in all of the web pages in the root set and add d (the largest number) links to these web pages.

Revised Version Manuscript Received on June 15, 2017.

X. Leela Mary, M.Phil Scholar, Department of Computer Science, Annai Vailankanni of Arts & Science College, Thanjavur (Tamil Nadu), India.

G. Silambarasan, Assistant Professor, Department of Computer Science, Annai Vailankanni of Arts & Science College, Thanjavur (Tamil Nadu), India.

(3) In the base set (B), hub pages are defined into a vertex set (V1), and authority pages are defined into another vertex set (V2). Hyperlinks from pages in V1 to those in V2 are defined into an edge set (E), forming a binary directed graph G = (V1, V2, E). For any vertex (v) in V1, h(v) is used to represent the hub value of page V, and for any vertex (u) in V2, a(u) is used to represent the authority value. Assume that the structural subgraph (G) of web links contains n nodes (pages), which are numbered by: 1, 2, n, then A denotes the adjacency matrix (n×n) of the graph G. The th entry of A is equal to 1 if page i points to page j, and is equal to 0 otherwise. Similarly, use vectors to define the authority and hub values at all nodes: a=(a1, a2, …, an) and h=(h1, h2, …, hn).

According to the theory of linear algebra, the vectors a and h will converge, after calculation, to the principal eigenvectors of the symmetric matrixes ATA and AAT . The principal eigenvector of ATA represents the authority page. A larger value of the principal eigenvector means a higher authority weight of the page. In the same way, the principal eigenvector of AAT represents the hub page. A larger value of the principal eigenvector means a higher hub weight of the page. From above process we can see that, the authority and hub of each page can be obtained after several times of iterative computations. Fundamentally, the authority and hub weight of a page in the base set B is decided by links between pages in B. More specifically, they are decided by the symmetric matrixes ATA and AAT.

C. Problems of the HITS algorithm

(1) Low computational efficiency

HITS is an algorithm related to query, thus the real-time computation will be performed after receiving the user query. However, the HITS algorithm itself requires times of iterative computations before getting the final results, leading to a low computational efficiency.

(2) The problem of topic waft

For some query topics, the HITS algorithm can accurately extract out the authority pages, but serious “topic waft” (a phenomenon that authorities are linked to many unrelated pages) problem would happen in some cases. It excludes totally the content or text of a page, with considering only thelink structure to analyze the authority. Compared with an authority page in real network, it is obviously unscientific.

(3) Results can be easily manipulated by cheaters

The HITS algorithm can be easily manipulated by cheaters at the mechanism level. For example, a cheater can build a page, in which the content is added with many links that point to high-quality pages or famous websites to get a good hub page; then the cheater lets this hub page point to his cheating page, in this way he can increase the authority score of the cheating page.

III. IMPROVEMENT OF THE HITS ALGORITHM

A. Ideas for improvement

The HITS algorithm has problems is mostly because that it is an algorithm purely based on the link analysis, ignoring the relevancy between the page and topic, without considering the content and text. In some situations, the HITS algorithm would cause the problem of “topic waft”. To solve this problem, we add the similarity weight of web page content into the iterative computations of authority and hub values. In this way we are able to get a balance between the link and the importance of the content, and thus can return better searching results [9]-[13].

B. Detailed steps for improving the algorithm

For the HITS algorithm, we here add the similarity value into the authority and hub values and then we get the improved algorithm as follows:

$$a(i) = \Sigma j \quad B(i)\{ \quad 1-\lambda \quad *Sim \ \ content \ \ +\lambda*h\} \qquad 3$$

$$h(i) = \Sigma j \quad F(j)\{ \quad 1-\lambda \quad *Sim \ \ content \ \ +\lambda*a\}$$

where, a(i) and h(i) is the authority and hub value, respectively; Sim(content) means the similarity of the web content, and; λ is an impact factor, varying from 0 to

IV. RELATED EXPIREMENT

A. Process of the Experiment

Here in the experiment, five pages A, B, C, D and E make

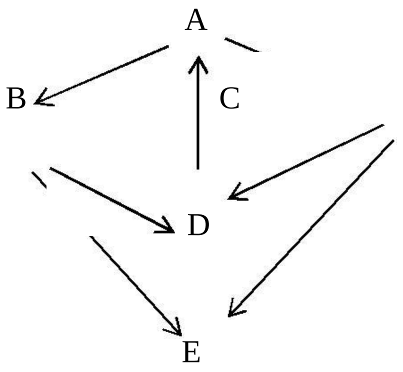
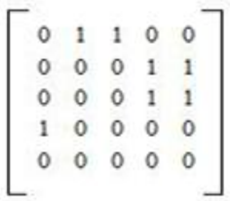


Fig.1. Link relation

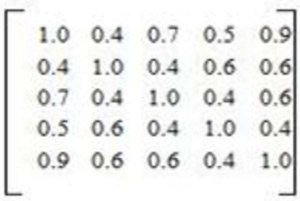
up a small network. Figure 1 shows its link relation. Link relation matrix: in the graph, A points to B, then AB=1; while B doesn’t point to A, so BA=0; the same for others.

Similarity matrix: this is a matrix assumed to satisfy the requirements of the experiment. It is a symmetric matrix. Assume that the similarity of A to itself is 1; the similarity of A to B is 0.4; likewise, the similarity of B to A is also 0.4; the same for others.

According to above analysis we get the link relation matrix and similarity matrix, as shown in Fig. 2 and 3.



$$\begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$



$$\begin{bmatrix} 1.0 & 0.4 & 0.7 & 0.5 & 0.9 \\ 0.4 & 1.0 & 0.4 & 0.6 & 0.6 \\ 0.7 & 0.4 & 1.0 & 0.4 & 0.6 \\ 0.5 & 0.6 & 0.4 & 1.0 & 0.4 \\ 0.9 & 0.6 & 0.6 & 0.4 & 1.0 \end{bmatrix}$$

Fig. 2.Link relation matrix **Fig. 3.Similarity**

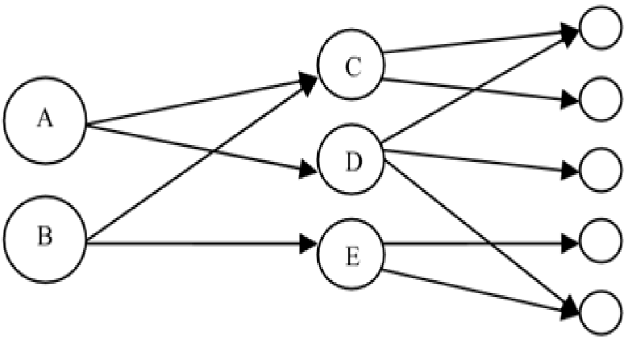
Different similarities are endowed to B and C that have the same links. A relatively higher similarity is endowed to the page E, which has no link pointing out. In this way we can compare and analyze the results later in the experiment. Program the HITS algorithm and compute the authorities of contents and hubs of links of each page. In the experiment, the algorithm is achieved by using java programming, in which the input matrix is read in form of TXT.

V. EXISTING SYSTEM

The research of page ranking algorithm is significant because it will increase the accuracy of search engines.At the beginning this article briefly introduced mainstream page ranking algorithms, especially analyzed topic drift its existing problem.For some query topics, the HITS algorithm can accurately extract out the authority pages, but serious “topic drift” (a phenomenon that authorities are linked to many unrelated pages) problem would happen in some cases. It excludes totally the content or text of a page, with considering only theWhen performing the link analysis, λ is used to reasonably control the impact of the content similarity on the authority and hub values, so as to eventually control the problem of topic drift. The results can fully represent the similarity as well as the authority, thus meet the user query in a better way.

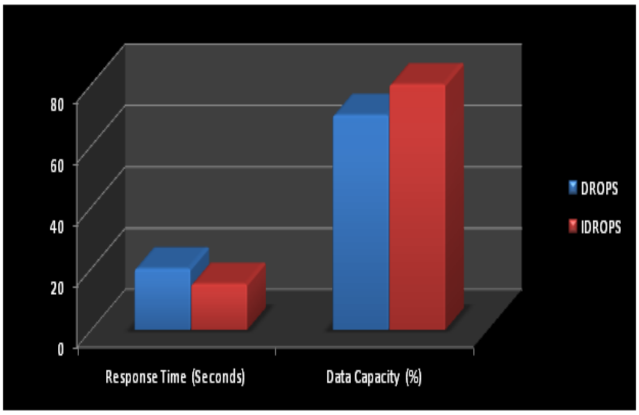
VI. PROPOSED SYSTEM

HITS (Hyper Link—Induced Topic Search) is an algorithm for analyzing WSM (Web Structure Mining). The algorithm is to find out the authority pages and hub pages from a set of web pages, according to user query and through analyzing the forward and backward linkages. An authority page is an authoritative page most relative to the query topic (authority is of influence and is accepted by most of people); while a hub page is a web page that points to the link set of authoritative pages . Each page requires two measurement values: authority weigh and hub weight, according to which the importance of a page to a specific topic can be judged.



VII. EXPERIMENTAL RESULTS

We can evaluate the performance of the system using the parameters such as (i) increasing the number of nodes in the system, (ii) increasing the number of objects keeping numberof nodes constant, (iii) changing the nodes storagecapacity, and (iv) varying the read/write ratio. These measurements are consolidated as capacity of replication node and time of updation. And it can be plotted as graph.



VIII. CONCLUSIONS

HITS is an important algorithm for web structure mining. In response to problems of this algorithm, many scholars have proposed various improved algorithms which are still in evolution. In order to solve the problem of topic drift in the HITS algorithm, the current study combines the relevancy of web content with the authority of link analysis to present an optimized strategy. The experiment confirms that the improved algorithm has improved the relevancy of query results and limited the occurrence of topic drift to some degree. Of course, the current algorithm Also has a problem that the factor of λ is randomly valued from 0 to 1, without proper rules for value assignment. How to assign and optimize the value of λ is a direction for further studies.

REFERENCES

1. WU An-qing,ZHANG Ying-jiang,TU Jun. Research on Integrative Crawling Strategy of Subject[J]. Journal of Wuhan UniversityofTechnology,Vol.8,No.2,pp.74-76,2006.
2. DAI Kuan,ZHAO HUI,HAN Dong,SONGTian-yong. Theme Feature Extraction of Chinese Webpage Based on Vector Spa ceModel[J].Journal of Jilin University:InformationScience,Vol.32,No.1,pp.88-94,Jan.2014.
3. SHU Ben,YINKe. Research and Design on Topical Crawler Based on Analysis of Conte nt and Link[J].Computer and Modernization,No.4,pp.77-80,2014.
4. [WEI Jing-Jing,YANG Ding-Da,LIAO Xiang-Wen. Focused Crawler Based on Improved Algorithm of Web Content Similarity[J]. Computer and Modernization,No.9,pp.1-4,2009.
5. GUO Hong.An Improved HITS Based on Texts[J]. Application of computer system,No.9,pp.38-40,2009.
6. LIU Tie-Nan,LIUBin,LIANG Fu-Gui. A genetic algorithm with local searching strategy and its application[J]. Journal of Daqing PetroleumInstitute,Vol.29,No.2,pp.76-78,Apr.2005.
7. QIAN Gong-Wei,NI Lin. Extended PageRank algorithm based on Web link and content analysis[J]. Computer engineering and applications,Vol.43,No.21,pp.160-164,2007.
8. NI Xian-Jun. Research on the Improving Algorithm of Web Structure mining[J]. Micro computer information,Vol.12,No.3,pp.163165,2007.
9. HUANG Li-Wen,QIAN Wei. An Improved HITS Approach for multi-document Text Summarization[J]. Computer application,Vol.26,No.11,pp.2625-2627,2006.
10. WANG Xiao-yu,ZHOUAo-ying. Linkage Analysis for the World Wide Web and Its Application A Survey[J]. Journal of Software,Vol.14,No.10,pp.1768-1780,2003.
11. YU Jinping, ZHU Guixiang, MEIH ongbiao. Research and improvement of HITS algorithm based on Web link analysis[J]. Computer Engineering and Applications,Vol.49,No.21,2013.
12. Cho J, Garcia-Molina H, Page L. Efficient Craw ling through URL Ordering[J].Computer Networks, Vol.30(1- 7),pp.161-172,1998.
13. HE Xiao-Yang, WUQiang, WUZhi-Rong. Comparative analysis of HITS algorithm and Page Rank algorithm. Journal of Informaiton,No.2,pp.85-86,2004.