An Improved HITS Algorithm Based on Page-query Similarity and Page Popularity

Xinyue Liu[1,2]
1. School of Computer Science and Technology, Dalian University of Technology, Dalian, China
Email: xyliu@dlut.edu.cn

Hongfei Lin[1] and Cong Zhang[2]
2. School of Software, Dalian University of Technology, Dalian, China
Email: hflin@dlut.edu.cn

***Abstract*—The HITS algorithm is a very popular and effective algorithm to rank web documents based on the link information among a set of web pages. However, it assigns every link with the same weight. This assumption results in topic drift. In this paper, we firstly define the generalized similarity between a query and a page, and the popularity of a web page. Then we propose a weighted HITS algorithm which differentiates the importance of links with the query-page similarities and the popularity of web pages. Experimental results indicate that the improved HITS algorithm can find more relevant pages than HITS and improve the relevance by 30%-50%. Furthermore, it can avoid the problem of topic drift and enhance the quality of web search effectively.**

***Index Terms*—HITS Algorithm, Link Analysis, Similarity, Popularity**

I. INTRODUCTION

With the rapid growth of computer technique and the Web, 2EB (1EB≈1024*1024*1024GB) information is produced each year in the whole world and the increasing speed of information has been beyond imagination. Thus, how to effectively search relevant information from the huge information on the internet is the primary goal of the modern search engines. Link analysis plays an important role in accomplishing this task.

The most two famous link analysis algorithms are PageRank algorithm [1] by Sergey Brin and Lawrence Page and the HITS algorithm [2] by Kleinberg. They are all purely link-based algorithm which do not think about the content of the page so that resulting in the problem of topic drift. That means the results of the algorithms often contain some non-relevant pages with tightly interconnected density. In order to control the topic drift, page ranking algorithms based on hyperlinks and content have been proposed, such as the ARC algorithm [3] and the Average algorithm [4]. HITS algorithm can not treat links differently, due to its definition makes the quality of hubs is determined by it pointed to the quantity of authority page. In 2001, Allan Borodin and others proposed the Hub-Averaging (HUBAVG) algorithm which sets the hub weight of some node *i* to the average authority weight of the authorities pointed to by hub *i*. Later, Allan Borodin and others proposed the Authority-Threshold algorithm [5] which sets the hub weight of node *i* to be the sum of the *k* largest authority weights of the authorities pointed to by node *i*. This corresponds to saying that a node is a good hub if it points to at least *k* good authorities. Moreover, Lempel and Moran proposed the SALSA algorithm [6] based on the Markov chain. Cohn and Chang proposed the PHITS algorithm [7] based on the probability model, and so on.

Based on the HITS algorithm, this paper finds that the HITS algorithm treats links equally and does not take use of the content of the page when distributes rank scores. As a result, the definition of the similarity of the pages and the popularity of the pages are proposed by this paper to influence the rank result. This paper proposes the Improved HITS (I-HITS) algorithm based on similarity and popularity, which differentiates the importance of links with the similarity of pages and the query topic and the popularity of pages. And I-HITS algorithm constructs a new adjacency matrix to compute hubs and authorities. Theoretical analysis and experimental results both show that the I-HITS algorithm performances better in search precision and avoids the problems of topic drift effectively.

The rest of this paper is organized as follows. A brief background review of HITS algorithm and the problems it has been presented in Section 2. An extended HITS algorithm, the Improved HITS algorithm is described in Section 3. Experimental results of I-HITS are given in Section 4. Section 5 summarizes the conclusions.

II. HITS ALGORITHM AND PROBLEM ANALYSIS

A. HITS algorithm

Unlike PageRank algorithm, Kleinberg proposed a more refined notion for the importance of the Web pages. He believed that the importance of pages is dependent with the query topic. In his framework of Fig. 1, every page can be considered as two identities, hub and authority. The link structure can be described as a

Manuscript received December 28, 2010; revised March 1, 2011; accepted March 28, 2011.
Corresponding author: Xinyue Liu

doi:10.4304/jcp.7.1.130-134

dependent relationship: A good authority is a page pointed to by good hubs, while a good hub is a page that points to good authorities. Therefore, Kleinberg defined the following mutual reinforcing relationship between hubs and authorities: the hub weight to be the sum of the authorities of the nodes that are pointed to by the hub, and the authority weight to be the sum of the hub weights that point to this authority [2] [8] [9].

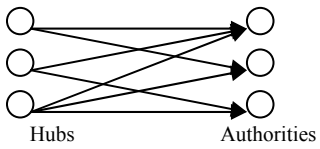

Figure 1. Hubs and authorities

$$a_i = \sum_{j \in B(i)} h_j \quad \text{and} \quad h_i = \sum_{j \in F(i)} a_j \tag{1}$$

Let a_i and h_i represent the authority and hub weight of page i, respectively. $B(i)$ and $F(i)$ denote the set of referrer and reference pages of page i, respectively.

B. The problems of the HITS algorithm

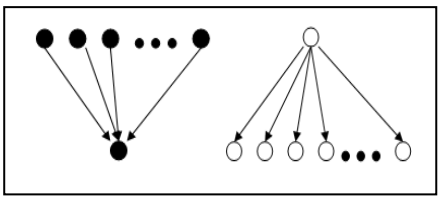

Figure 2. A bad example for HITS

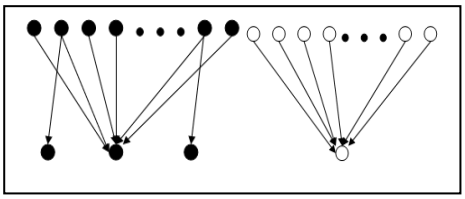

Figure 3. Another bad example for HITS

The definition of HITS algorithm has the following two implicit properties. It is symmetric, in the sense that both hub and authority weights are defined in the same way. If we reverse the orientation of the edges in the graph, authority and hub weights are swapped. The HITS algorithm is also egalitarian, in the sense that when computing the authority weight of some page p, the hub weights of the pages that point to page p are all treated equally (same with computing the hubs weights). However, these two properties may sometimes lead to non-intuitive results. Consider the example graph in Fig. 2 demonstrates that if the number of white authorities is larger than the number of black hubs, the HITS algorithm will allocate all authority weight to the white authorities, while giving little weight to the black authority and easily cause topic drift. However, intuition suggests that the black authority is better than the white authorities and should be ranked higher. Similarly, in Fig. 3, after computing, the middle black authority will have higher authority weight than the white authority, but actually they should be equally good. Therefore, we seek to change the symmetric and the egalitarian of the HITS algorithm, and aim at treating links differently.

III. I-HITS ALGORITHM

In view of the problem of the HITS algorithm, the paper proposes the I-HITS algorithm, which is according to the product of the relevance of Web page and query topic, the relevance of Web page linked to and query topic and the popularity of the page to weight the link instead of the average transfer strategy.

A. Definition 1: The similarity of the page and the query

If page i points to page j, then i is called the source page and j is called the target page. In most cases, the more information a source page contains with the query Q, this source page is more relevant with the query topic, and the similarity is S_i. Similarly, the more information a target page contained with the query Q, this target page is more relevant with the query topic as well, and the similarity is S_j. Furthermore, we find that anchor text is used to describe the target document, not to describe the current document [1] [10], and it summarizes the topic of the target document with a high degree of accuracy [11]. Therefore, in order to reduce the computational complexity, computing the similarity of the target page and the query Q is simplified by computing the similarity of the anchor text and the query Q. Thus, in the adjacency matrix, if page i points page j, the item in it is $(1+ S_i)*(1+ S_j)$; if not, it is 0.

B. Definition 2: The popularity of the page

The more popular a page is, the more other pages tend to point to it or it will be linked to by other pages. The proposed extended I-HITS algorithm contributes larger rank values to more important (popular) pages instead of dividing the rank value of a page evenly among its outlink pages.

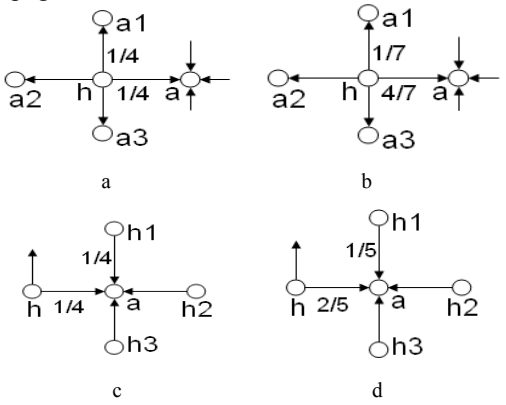

Figure 4. How to distribute weight with popularity

Fig. 4.a represents the HITS algorithm. According to the out-degree of h, h allocates its weight as the probability of a quarter. That is, a, a_1, a_2 and a_3 will get the same value of hub. However, in Fig. 4.b, with the definition of the popularity, the hub value of each webpage is determined by the proportion of its in-degree

pointed to the overall in-degree. That is, allocating hub weights according to the migration probability of links. The in-degree of *a*, a_1, a_2, a_3 are 4, 1, 1, 1, then *a* gets 4/7 of the hub weight and a_1, a_2, a_3 gets 1/7. Similarly, in Fig. 4.c, when HITS computing authority values, the chances of the authority pages accumulating the authority values are equal. All the *h*, h_1, h_2, h_3 pointing to "*a*" will have the same chance of gaining authority values. In the I-HITS algorithm, the authority value of each webpage pointing to "*a*" is determined by the proportion of its out-degree to the overall out-degree, as shown in Fig. 4.d. E.g. with the respective out-degree values of *h*, h_1, h_2, h_3 at 2、1、1、1, the authority values of *h*, h_1, h_2, h_3 are 2/5, 1/5, 1/5, 1/5 respectively. Advantage of this way is that allocating more rank values to the most popular pages from an objective point of view. Like in a certain area, the more fans he has, more famous he should be.

The popularity from the number of outlinks is recorded as $W_{(j,i)}^{out}$ [12]. $W_{(j,i)}^{out}$ is calculated based on the number of outlinks of page *i* and the number of outlinks of all reference pages of page *j*. Where O_i and O_p represent the number of outlinks of page *i* and page *j*, respectively. R(*j*) denotes the reference page list of page *j*.

$$W_{(j,i)}^{out} = \frac{O_i}{\sum_{p \in R(j)} O_p} \tag{2}$$

The popularity from the number of inlinks is recorded as $W_{(j,i)}^{in}$ [12]. $W_{(j,i)}^{in}$ is calculated based on the number of inlinks of page *i* and the number of inlinks of all reference pages of page *j*. Where I_i and I_p represent the number of inlinks of page *i* and page *j*, respectively. R(*j*) denotes the reference page list of page *j*.

$$W_{(j,i)}^{in} = \frac{I_i}{\sum_{p \in R(j)} I_p} \tag{3}$$

In this example of Fig. 5, page A points page C and page D. The inlinks and outlinks of these two pages are I_C=2, I_D=1, O_C=2, O_D=3. Therefore, $W_{(A,C)}^{out} = O_C / (O_C + O_D)$ =2/5 and $W_{(A,C)}^{in} = I_C/(I_C + I_C)$ =2/3.

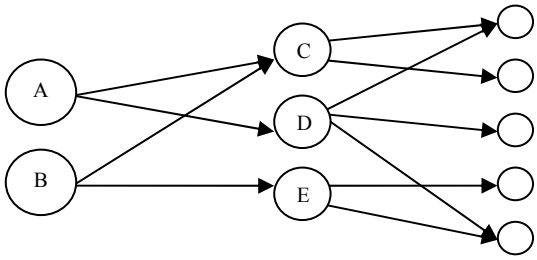

Fig 5. Links of Web

C. The I-HITS algorithm

(1) Assign each page with two value: a_i and h_i, represent authority weight and hub weight.

(2) Initialize: a_i=1, h_i=1(i=1, 2, 3, … , n).

(3) Compute:

$$a_i = \sum_{j \in B(i)} h_j * (1 + s_i) * (1 + s_{ji}) * \frac{I(i)}{\sum_{p \in F(j)} I(p)}$$

$$h_i = \sum_{j \in F(i)} a_j * (1 + s_i) * (1 + s_{ij}) * \frac{O(i)}{\sum_{p \in B(j)} O(p)} \tag{4}$$

Normalization.

(5) If a_i and h_i do not converge, turn to step (3).

IV. EXPERIMENTS

A. Evaluation method

TREC (text retrieval conference) is the greatest impact of evaluation conference in the text information retrieval areas. This experiment uses TREC's P@10 evaluation criteria. That is, P@10 is the number of relevant documents in the top 10 documents in the ranked list returned for a topic.

TREC usually use binary and ternary evaluation. This experiment uses a more specific and detailed ternary evaluation and classifies a document as:

(1) Highly relevant (HR): Contain very important and very authoritative information about the given query.

(2) Relevant(R): Have relevant but not important information about the given query.

(3) Non-relevant (NR): Include neither the keywords of the given query nor relevant information about it.

An objective categorization of the results is achieved by integrating the responses from several people who are able to offer impartial opinions: for each page, we compared the count of each category and chose the category with the largest count as the type of that page.

B. Experiment data

In order to evaluate the I-HITS algorithm, this experiment compared with HITS and I-HITS on 5 different queries, four in English and one in Chinese: abortion, alcohol, basketball, movies, 搜索引擎. All of these queries have already appeared in previous works [2] [13] [14] and are representative. The base sets for these queries are constructed in the fashion described by Kleinberg. Statistical data is in Table 1.

TABLE I.
EXPERIMENT DATA

query	nodes	hubs	authorities	links
abortion	1652	949	933	2849
alcohol	1964	1441	1213	11083
basketball	1153	930	641	3588
movies	2934	2051	1885	18210
搜索引擎	2884	2142	1744	37941

C. Evaluation

The following table shows three queries results of alcohol, movies and 搜索引擎. In these tables, the results that are labeled highly relevant appear in boldface, while the relevant ones appear in italics, and the rest are non-relevant.

Table 2 shows the results of alcohol. The top 10 results returned by HITS are all relevant, 7 of them are highly

relevant. The top 10 results returned by I-HITS are all relevant, 8 of them are highly relevant, the proportion of highly relevant improves by 10%.

TABLE II.
RESULTS OF ALCOHOL

rank	HITS	I-HITS
1	**http://www.niaaa.nih.gov/**	*http://www.ndp.govt.nz/publications/review-*
2	**http://www.health.org/**	**http://www.wrap.org/**
3	**http://faculty.washington.edu/chudler/alco.html**	**http://www.alcoholmedicalscholars.org/**
4	*http://www.alcoholfreechildren.org*	**http://www.niaaa.nih.gov**
5	**http://www.nasadad.org/**	**http://www.health.org/**
6	*http://www.nofas.org/*	**http://www.alcoholconcern.org.uk/**
7	**http://ncadi.samhsa.gov/govpubs/ph323**	**http://ncadi.samhsa.gov/govpubs/ph323/**
8	**http://www.alcoholconcern.org.uk/**	**http://faculty.washington.edu/chudler/alc**
9	*http://www.atf.treas.gov/*	**http://www.cdc.gov/alcohol/**
10	**http://www.hsph.harvard.edu/nutritionso**	*http://www.atf.treas.gov/*

TABLE III.
RESULTS OF MOVIES

rank	HITS	I-HITS
1	**http://movies.yahoo.com/**	http://www.gazettenet.com/dining
2	**http://www.nytimes.com/pages/movies/in**	**http://www.onwisconsin.com/movies/**
3	**http://movies.msn.com/**	*http://us.lrd.yahoo.com/_ylt=A9FJq6Dw5x*
4	http://www.wmm.com/	**http://www.hotfreelayouts.com/layouts.p**
5	http://dmoz.org/Arts/Movies/	*http://www.bio.unc.edu/faculty/goldstein/la*
6	*http://www.moviesunlimited.com/*	**http://www.frazy.com/**
7	http://service.real.com/filmcom/	*http://www.bio.davidson.edu/Courses/movie*
8	**http://movies.about.com/**	*http://www.gnovies.com/*
9	*http://www.empiremovies.com/links.php*	**http://www.yourmovies.com.au/**
10	http://www.teachwithmovies.org/	**http://www.reelclassics.com/**

TABLE IV.
RESULTS OF 搜索引擎

rank	HITS	I-HITS
1	**http://www.google.com/intl/zh-CN/**	*http://www.xpue.net*
2	*http://www.gseeker.com*	*http://www.tooooold.com*
3	http://www.wangtam.com/50226711/c_wav	*http://www.bbssearch.cn*
4	http://www.yuleguan.com	**http://www.google.com/intl/zh-CN/**
5	http://www.chinaventurenews.com	http://www.1hd.cn
6	http://www.tjacobi.com	**http://www.baidu.com/**
7	http://www.money-courier.com	*http://bizsite.sina.com.cn/*
8	http://www.geekervision.com	http://it.sohu.com/7/0903/35/column213613
9	http://www.in-women.com	**http://so.163.com/**
10	http://www.tracingadget.com	http://news.qq.com/a/20070405/001271.ht

Table 3 shows the results of movies. 6 of the top 10 results returned by HITS are relevant, 4 of them are highly relevant. 9 of the top 10 returned by I-HITS are relevant, 5 of them are highly relevant. The proportion of highly relevant improves by 10% and the proportion of relevant improves by 30%.

Table 4 shows the results of 搜索引擎. 2 of the top 10 results returned by HITS are relevant, 1 of them are highly relevant. 7 of the top 10 returned by I-HITS are relevant, 3 of them are highly relevant. The proportion of highly relevant improves by 20% and the proportion of relevant improves by 50%.

From Table 2 to Table 4, to some extent, HITS algorithm and I-HITS algorithm all will product the problem of topic drift. However, Table 5 and Table 6 show that I-HITS algorithm generates more relevant pages than HITS, which illustrates I-HITS is better than HITS in the precision. Furthermore, the proportion of highly relevant improves by 10%-20% and the proportion of relevant improves by 30%-50%.

V. CONCLUSION

TABLE V.
PROPORTION OF HR

query	HITS	I-HITS
alcohol	70%	**80%**
movies	40%	**50%**
搜索引擎	10%	**30%**

TABLE VI.
PROPORTION OF R

query	HITS	I-HITS
alcohol	**100%**	**100%**
movies	60%	**90%**
搜索引擎	20%	**70%**

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 HITS algorithm and its existing problem. Based on HITS algorithm, this article proposed I-HITS algorithm, which

is an algorithm according to parameters of similarity and popularity. I-HITS algorithm measured the importance of page link based on two page attributes, which includes source page, target page, relevancy of query topic and the popularity of page in Network subgraph. As a result, I-HITS increased the ability of distinguish link importance, and avoid top drift. Furthermore, this article compared HITS algorithm with I-HITS algorithm by experiment. The result shows that I-HITS algorithm is better than HITS algorithm when searching related page. The quality of query is increased obviously.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China under Grant No. 60873180 and the Fundamental Research Funds for the Central Universities.

REFERENCES

[1] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," *Computer Networks and ISDN Systems*, 1998, vol. 30, pp. 107-117.

[2] J. Kleinberg, "Authoritative sources in a hyperlinked environment," *In Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, 1998, pp. 668-677.

[3] Chakrabarti S, Dom B, Raghavan P, etal, "Automatic resource compilation by analyzing hyperlink structure and associated text," *Computer Networks and ISDN Systems*, 1998, vol. 30, pp. 65-74.

[4] Gevrey J, Ruger S, "Link-based approaches for text retrieval," *Proceedings of TREC-10, NIST (Gaithersburg, MD, 13-16, Nov 2001), NIST Special Publication*, 2002, pp. 279-285.

[5] Borodin A, Roberts G O, Rosenthal J S, "Finding Authorities and Hubs from Link Structures on the World Wide Web," *In Proceedings of the 10th International Conference on World Wide Web, Hong Kong*, *China,* 2001, pp. 415-429.

[6] Lempel R, Moran S, "The stochastic approach for link-structure analysis (SALSA) and the TKC effect," *Computer Networks*, 2000, vol. 33, pp. 387-401.

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

[8] S. Chakrabarti, B. E. Dom, S. R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, and J. Kleinberg, "Mining the Web's link structure," *Computer*, 1999, vol. 32, pp. 60-67.

[9] C. Ding, X. He, P. Husbands, H. Zha, and H. Simon, "Link analysis: Hubs and authorities on the world," *LBNL Tech Report 47847*, 2001.

[10] N Craswell, D Hawking, S E Robertson, "Effective site finding using link anchor information," *In Research and Development in Information Retrieval*, 2001, pp. 250-257.

[11] Zhang Min, Gao Jianfeng, Ma Shaoping, "Anchor Text and Its Context Based Web Information Retrieval," *Journal of Computer Research and Development*, 2004, vol. 41, pp. 221~226.

[12] Wenpu Xing and Ali Ghorbani, "Weighted PageRank Algorithm," *In Proceedings of the Second Annual Conference on Communication Networks and Services Research*, 2004.

[13] C. Dwork, R.Kumar, M. Naor, and D. Sivakumar, "Rank aggregation methods for the Web," *In Proceedings of the 10th International World Wide Web Conference*, *Hong Kong*, 2001, pp. 613-622.

[14] R. Lempel and S. Moran, "The Stochastic Approach for Link-Structure Analysis (SALSA) and the TKC Effect," *In Proceedings of the 9th International World Wide Web Conference*, 2000, pp. 387-401.

Xinyue Liu received the M.S degree in Computer Science and technology from Northeast Normal University, China, in 2006. She is currently working toward the Ph.D. degree in the School of Computer Science and Technology, Dalian University of Technology, Dalian, China. Her research interests include multimedia information retrieval, web mining and machine learning.

Hongfei Lin received the Ph.D degree from Northeastern University, China. He is a professor in the School of Computer Science and Technology, Dalian University of Technology, Dalian, China. His professional interests lie in the broad area of information retrieval, web mining and machine learning, affective computing.

Cong Zhang received her M.S. degree in Software Engineering from Dalian University of Technology in 2008. Her major interests lie in web link analysis.