Applied Mathematics & Information Sciences
An International Journal

Improvement on HITS Algorithm

Yue He, Minghong Qiu, Maozhu Jin[*]*,Tao Xiong*

School of Business and Administration, Sichuan University, Chengdu 610064, China

Received: Jul 8, 2011; Revised Oct. 4, 2011; Accepted Oct. 6, 2011

Abstract: In this paper, we did a theoretical analysis on the typical link analysis algorithm HITS and proposed a new improved algorithms to the problems of the algorithm. To the problem of topic drift, we proposed an improved HITS algorithm considering user's click-through rate. To the problem that the authority value leans to old web pages, we proposed an improved HITS algorithm considering web age impact weight. In order to solve two problems at the same time, we combined the first two improved algorithms and proposed a new synthesis algorithm. Finally, these three improved algorithms are proved through the experiments that we designed.

Keywords: HITS algorithm, User's Click-Through Rate, Web Age Impact Weight

1. Introduction

With the vast development and popularization, the Internet has become a huge and widely distributed center of global information service since the 1990s. How to help the users find useful knowledge from the information ocean is an urgent problem to be solved. Web data mining is just the technique developed to meet the needs. It can extract implied information and useful patterns that users are interested in from relevant internet resources and user browsing behavior [1]. Many users find information on the internet by search engine. The search engine collects and finds information in a certain strategy and then understands, extracts, processes and organizes information to provide searching and information navigation services [2]. Web data mining and search engine can be mutual complement and promotion. The search engine can improve its efficiency and meet users' needs through learning from the Web mining. The link analysis algorithm that is widely used by search engine at present belongs to the structure mining category of Web data mining. Through analyzing the hyperlink and mining the latent semantic information, link analysis algorithm can help users to find information they need from the sea of information on the internet, and thus improve the retrieval efficiency of search engine.

In 1999, Dr. Jon Kleinberg of Cornell University put forward HITS (Hyperlink-Induced Topic Search) algorithm [3]. Based on the result set getting from traditional search engine, the algorithm acquires page set related to query topic, calculates quality value of every page, and selects a few high-quality information sources according to it. This algorithm is also known as the topic distillation algorithm. HITS algorithm proposes the definition of hub pages and authority pages, and they reinforce each other. As mentioned earlier, HITS algorithm and query topics are closely related. After users inputting the query key words, HITS algorithm firstly acquires a root set related to query topic by a traditional search engine. Next, according to the expanded rules, the algorithm expands root set into base set. Finally, it calculates the authority weight and hub weight of each page in base set.

HITS algorithm creatively puts the hyperlink into practice and makes good effect. However, further studies [13, 14, **?**] find that the algorithm has some inadequacies. First of all, HITS algorithm does not consider the content of the pages at all, so it easily generates topic drift phenomena. Secondly, when it comes to calculating authority weights, HITS algorithm easily leans to old web pages and ignores the new ones. Meanwhile, HITS is often influenced by the unreasonable relationships between different websites, as well as other human disturbances.

To the inadequacies of HITS algorithm, domestic and foreign researchers have improved it a lot. The Clever Engineering Group or Almaden Research Center in IBM proposes improved HITS algorithm–ARC(Automatic Resource Compilation) algorithm [5]. When it assigns initial value to Web adjacency graph's corresponding adjacency matrix, this algorithm combines anchor text information of hyperlink. Different hyperlinks have different weights; as a result, it improves topic drift. In literature [16], Lem-

[*] Corresponding author: e-mail: jinmaozhu@scu.edu.cn

pel and Moran improved HITS algorithm with Markov Chain. They desalted the relationship between authority pages and hub pages, and proposed SALSE (Stochastic Approach for Lin-Structure Analysis) algorithm to analyze the structure of hyperlink. Based on the detailed analysis of link relations among Web pages, Allan Borodin etc. scholars improved the iterative computations of HITS algorithm, and proposed Hub-Averaging-Kleinberg algorithm and three kinds of threshold control algorithms [6]. Aiming at unreasonable reinforcing relationship among web pages in HITS algorithm, Krishna Bharat etc. scholars proposed distribute-influence IMP algorithm [7]. Liu Fangfang in Dalian University of Technology and Liu Jun in Central South University etc. have studied the expand course from root set into base set, and proposed root set extend mode MCJITS algorithm [8]. Christopher S. Withers et. al considered N-transmit M-receive antenna systems with multiple frequencies and delay spread [15].

As mentioned earlier, although many scholars make some improvements on the HITS, these improvements often focus on a particular aspect of the shortages. Thus, it's difficult to meet the various demands of users and there still exists difference compared with users' expectation. Furthermore, most of current researches are in the scope of Web structure mining and have certain limitations. to the problems of topic drift and emphasis on the authority value of old webpages, the paper proposed a new improved algorithm in order to help users to find the authoritative source of information.

2. HITS algorithm

Based on the link analysis algorithm of Web structure, search engine proposes a measurement process of the authority of webpages: topic distillation [4]. The topic distillation, whose essence is attempted to find a commonly accepted objective evaluation conclusion from a vast amount of quality evaluation opinions, is a process that the search engine finds the high-quality authority source of information relevant to the query subject based on the user's query request [2]. HITS algorithm is a typical topic distillation algorithm.

HITS algorithm divides the webpages into two types, called hub pages and authority pages. The authority pages are generally recognized as the important pages on a particular topic. The hub pages, which can be regarded as the pages of evaluation pages, are the pages that link to a collection of authority pages on a particular topic. There is a mutually reinforcing relationship between authority pages and hub pages: a good authority page should be pointed to by many good hub pages while a good hub page should point to many authority pages. HITS algorithm makes the use of the mutually reinforcing relationship between them and gets the page ranks by iterative computation.

According to this idea, HITS defines two metrics for each page: authority weight and hub weight. And then, these two weights are computed iteratively to determine the importance of a certain page. Thus, the basic idea of HITS is: firstly, use a traditional text search engine to get a root set of pages related to the query topic. Then extend the root set in order to acquire a larger base set, namely that add the pages that point to the pages from root set as well as the pages that pointed by the pages from the root set. Construct a Web adjacency graph, and acquire the authority weight and hub weight through iterative computation according to the mutually reinforcing relationship between authority pages and hub pages. Then, based on the authority weights, sort pages and acquire the authority source of information on search topic.

According to the basic idea, HITS consists of two main processes: constructing the Web adjacency graph and computing the authority weights and hub weights iteratively. The adjacency graph means using a directed graph $G = (V, E)$, to show the link structure of page set. The nodes of the graph, which define the pages in page set, are represented by set V, and the directed edges $(p, q) \in E$ means there is a link in page p pointing page q. The out-degree of a node represents the number of links in the page it defines, while the in-degree of a node represents the number of links coming into the page it defines. Generally, an adjacency matrix is used to represent the Web adjacency graph. HITS set two metrics to each node $p \in V$ in Web adjacency graph $= (V, E)$: authority weight $a(p)$ and hub weight $h(p)$. The former is used to measure the authority of a page while the latter is used to measure the hub of a page. The authority weight and hub weight of pages represented by p can be calculated through the following two steps:

I processing (calculate authority weight):

$$a(p) = \sum_{q:(q,p)\in E} h(q) \tag{1}$$

O processing (calculate hub weight):

$$h(p) = \sum_{q:(q,p)\in E} a(q) \tag{2}$$

The $a(p)$ above represents the authority weight of page p; $h(p)$ represents the hub weight of page p; directed graph $G = (V, E)$ represents the Web adjacency graph constructed; $p \in V, q \in V$ present the nodes, which correspond to the pages in the page set; directed edge $(p, q) \in E$ represents that there is a link connecting to page q in page p.

3. Improvement of HITS Algorithm

3.1. Improved HITS algorithm with user's click-through rate considered

In order to get the rank of authority weights that meet the users' needs truly and solve the problem of topic drift, it can be considered from the users' browsing behavior. When observing the users' search behavior, it can be found

that after getting the particular search results, the user read the abstract information first, and then choose a page that he think is relevant to the search. It may be one-off behavior, as well as a behavior of returning and choosing many times. It is because that the user may obtain the information they need for the first choosing, or after many times of choosing. No matter which situation it is, we believe that a rational user will choose a page relevant to the search topic according to the page abstract. This is a judgment done by user. And the fundamental of topic distillation is to find a commonly objective judgment from a large number of various subjective judgments. It not only considers the opinions of the page's author (individual or organization), but also the need of the users. Therefore, in order to reflect the user's judgment, a weight factor of user's click-behavior is introduced in the HITS algorithm.

Considering the search behavior during a period of time, the number of clicks can be defined: for a query Q, n pages are gotten. Suppose query Q is committed for M times during a period of time, and the click number of each page is w. The research on PKU Tianwang System, which is conducted by the Computer Network and Distributed Systems Laboratory of Peking University, finds that most users only browser the first few pages of the provided search list of pages [10]. Therefore, the lower-ranking search results are accessed by the users more difficultly. This leads the small number of users' click naturally. In order to eliminate the influence that a page's rank in the search page list makes on users' clicks, the pages that displayed in the same search page list are regarded as a set. The clicks proportion of each page accounts for in the whole clicks of the pages in a certain search page is regarded as a weight factor to add to the algorithm.

Definition 1 For the query Q, there are n lists of results pages. Each list can display L Web pages. And during a period of time, the click numbers of each page displayed in the same list is presented as $\{w_1, w_2, \ldots, w_L\}$. Thus user's click-through rate (CTR) can be represented as:

$$H = \frac{w_i}{\sum_{i=1}^{L} w_i} \tag{3}$$

As the authority weight and hub weight of pages increase with the increasing of the users'CTR, the iterative calculation processes of HITS algorithm with users'CTR added can be:

I processing:

$$a(p) = \sum_{q:(q,p)\in E} h(q) + H_p \tag{4}$$

O processing:

$$h(p) = \sum_{q:(q,p)\in E} a(q) + H_p \tag{5}$$

The $a(p)$ above represents the authority weight of page p; $h(p)$ represents the hub weight of page p; directed graph $G = (V, E)$ represents the Web adjacency graph constructed; $p \in V, q \in V$ present the nodes, which correspond to the Web pages; directed edge $(p, q) \in E$ represents that there is a link connecting to page q in page p. H_p represents the CTR of page p.

It should be noted that if there isn't any users click a certain page, the CTR is zero. The formula of the page's authority weight and hub weight is the same as the original algorithm. And if there are no clicks on any of the pages in the same search result list during a period of time, the CTR will not be taken into account. These are two extreme situations. Meanwhile, in order to avoid the influence that the small CTR makes on results of authority weight and hub weight, if the sum of clicks of all pages in the same search result is very small (≤ 10), the CTR of the pages in the list will not be taken into account as well. At the macro level, these processes can avoid the excessive influence that the individual user's behavior makes on the objective evaluation sort of pages.

The algorithm flow chart of improved HITS algorithm with user's click-through rate considered is shown in Fig. 1. The input of the flow chat is the root set of web pages acquired by traditional essay search engine while the output is the final page ranking.

In fact, the improved HITS with CTR introduced has evolved into the combination algorithm of Web usage mining and Web structure mining from simple Web structure mining. As one of the main algorithms in the second generation of search engine, the improvement of HITS must be close to the intelligent search engine, which is the third generation of search engine. The CTR introduced into the HITS is one of the users' feedback behaviors. As an intelligent improvement, it can make the search engine algorithm closer to people's thinking mode, and overcome the topic drift better. Thus, the rank of pages generated by it can meet the searching needs better.

3.2. Improved HITS algorithm with the influence weight of page's age considered

Generally speaking, it can be divided into new pages and old pages according to the publishing time of Web pages. Usually, users are interested in authority pages with high quality and the latest information, as well as some information in a particular time interval. Especially in the information age, people become accustomed to understand what has happened. Therefore, in the daily search behavior, the interests within a particular time interval account for an increasing proportion. When calculating the weights, original HITS tend to focus on the old pages with high quality. Thus the new pages with high quality possibly cannot get a good authority, which results their rearward positions in the final pages sort. The old pages with high quality include pages updated in time and pages out of date. It can be judged that the good authority pages which are able to meet users' need are new pages with high quality

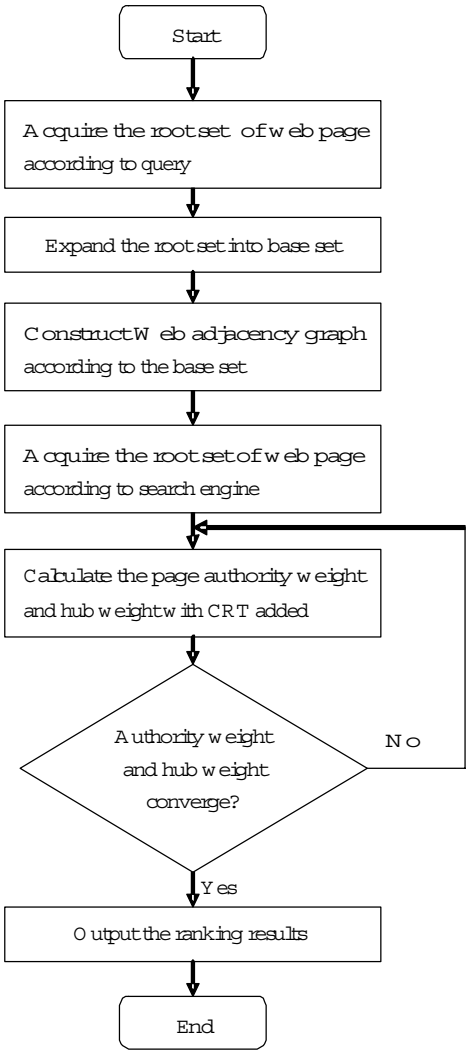

Figure 1 The processing of improved HITS with CTR considered

and old high-quality pages updated in time, for these two types of pages can both be regarded as new pages from the timeliness. The authority weight is generally selected as the weight that determines the pages sort in HITS algorithm. In order to make these pages get good positions in the search results list, the time factor is taken into account in the weight calculation.

The time information of Web pages is the time when the pages are updated. According to a research, eighty percent of Web pages will return the last update time when crawled [11]. If they won't, the latest time that the content is released in the page can be gotten by observing instead. The current time is the time when query Q happens. The page age is the difference between the current time and the pages' latest updated time.

Definition 2 The current time is the time T when query Q happens. The page's latest updating time is T_c. The time when the latest content is represented is T_p. The page's age Y is the difference between the current time and the latest updated time, which can be represented as the formula: $Y = T - T_c$. If there is no updating time, the page's age is the difference between the current time and the latest content' released time, which can be represented as the formula: $Y = T - T_p$.

Here, the age is calculated by month and the difference less than a month is ignored. For instance, if the T is $2010-1-1$ and T_c is $2008-9-1$, then $Y = (2010 - 2008)12 + (1-9) = 16$. Considering the efficiency of calculation formula, the older a page is, the more old-fashioned the content is, thus the weight is smaller. The influence weight of page's age can be represented as follows:

$$SJ = \sqrt{\frac{12}{Y+1}} \tag{6}$$

The influence weight of page's age reflects that with the increasing of the page's age, the authority weight of the page decrease nonlinearly. Therefore, the iterative calculation processes of HITS algorithm with the influence weight of page's age added can be:

I processing:

$$a(p) = \sum_{q:(q,p)\in E} h(q) SJ_p \tag{7}$$

O processing:

$$h(p) = \sum_{q:(q,p)\in E} a(q) \tag{8}$$

The $a(p)$ above represents the authority weight of page p; $h(p)$ represents the hub weight of page p; directed graph $G = (V, E)$ represents the Web adjacency graph constructed; $p \in V, q \in V$ present the nodes in the graph, which correspond to the Web pages; directed edge $(p, q) \in E$ represents that there is a link connecting to page q in page p. represents the influence weight of age of page p.

When introducing the influence weight of page's age, paper only considers the formula of authority weight and the formula of hub weight doesn't change. As the center of the information sources, a page with a good hub weight contains the links of famous authority pages in a particular field, and it can stay the same for a long time. Therefore, the information of hub pages and the information of authority pages are asymmetric in time course. The information a good hub pages contains is the consensus which takes a long time before it is formed in a particular field. While a good authority page should update its information in time in order to keep it authority as a high-quality page. An old-fashioned page cannot be a good authority page. The algorithm flow chart of improved HITS algorithm with the influence weight of page's age considered is shown in Fig. 2. The input of the flow chat is the root set

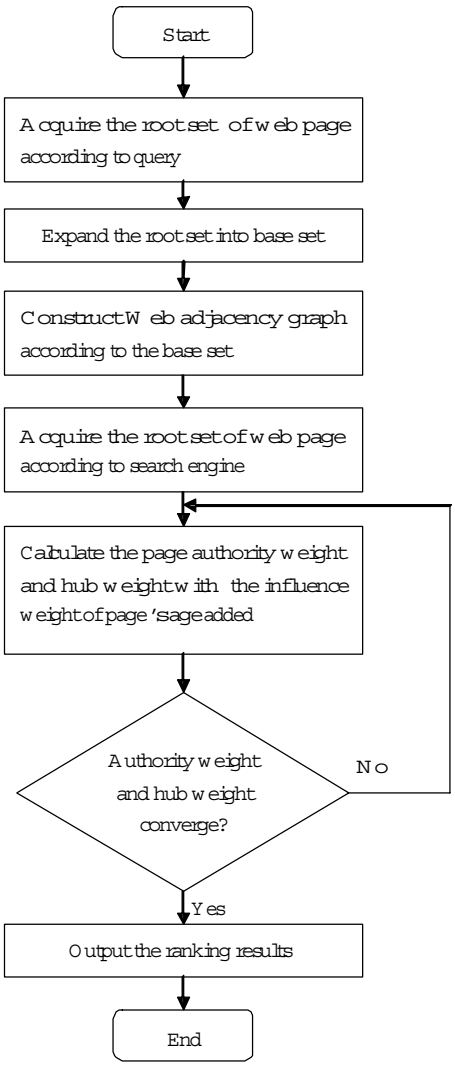

Figure 2 The processing of improved HITS with the influence weight of page's age considered

of web pages acquired by traditional essay search engine while the output is the final page ranking.

The page's age is the information of the page itself. It's one of the main directions to add the page information into the HITS, which is a pure link analysis. The ARC algorithm and the serial of its improved algorithm mentioned above introduce the page information to distinguish the different links. The page information considered in this paper is the time information. The Improved HITS algorithm with the influence weight of page's age considered can not only highlight the new high-quality pages, but also distinguish the old high-quality pages updated continuously and the old high-quality pages out of date. It can overcome the problem of emphasis on old pages when calculate the authority value, and thus, the final page rank can meet the users' timeliness demand of searching better.

3.3. Improved HITS algorithm integrated with CTR and influence weight of page's age

In HITS algorithm, generally, the authority page order list is acquired by sort the authority pages according to their authority weights. In order to meet the users' needs better, the authority pages must be the latest pages that are relevant to the users' search topics closely. The improvement on considering the CTR and the improvement on considering the influence of page's age are respectively against the problem of topic drift and the problem of emphasis on old pages. The HITS algorithm is a search engine algorithm, whose target is providing high-quality ranking information of pages according to users' needs. The results of information sources are not only relevant to the search topic, but also with a high degree of timeliness. Only meeting these two conditions can it meets user's expectation. Therefore, these two aspects are taken into account synthetically.

Based on the two improvements proposed earlier, paper proposes the improved HITS algorithm integrated with CTR and influence weight of page's to combine these two aspects of improvements effectively. A page ranking acquired by the synthetic algorithm, which considers the user's feedback behavior and the timeliness of pages at the same time, will meet users' expectation on search results better.

The processes of iterative calculation of the integrated algorithm are the follows:

I processing:

$$a(p) = \sum_{q:(q,p)\in E} h(q)SJ_p + H_p \tag{9}$$

O processing:

$$h(p) = \sum_{q:(q,p)\in E} a(q) + H_p \tag{10}$$

The $a(p)$ above represents the authority weight of page p; $h(p)$ represents the hub weight of page p; directed graph $G = (V, E)$ represents the Web adjacency graph constructed; $p \in V, q \in V$ present the nodes in the graph, which correspond to the Web pages; directed edge $(p, q) \in E$ represents that there is a link connecting to page q in page p. H_p represents the CTR of page p. represents the influence weight of age of page p. the calculation formula of refers to Eq. 3, and the calculation formula of refers to Eq. 6. The algorithm flow chart of improved HITS algorithm is shown in Fig. 3. The input of the flow chat is the root set of web pages acquired by traditional essay search engine while the output is the final page ranking.

The improved HITS algorithm integrated with CTR and influence weight of page's age, on the one hand, adds the CTR into the iterative algorithm of authority value and hub value. It enhancing the interaction between the algorithm and users and avoids the topic drift. On the other

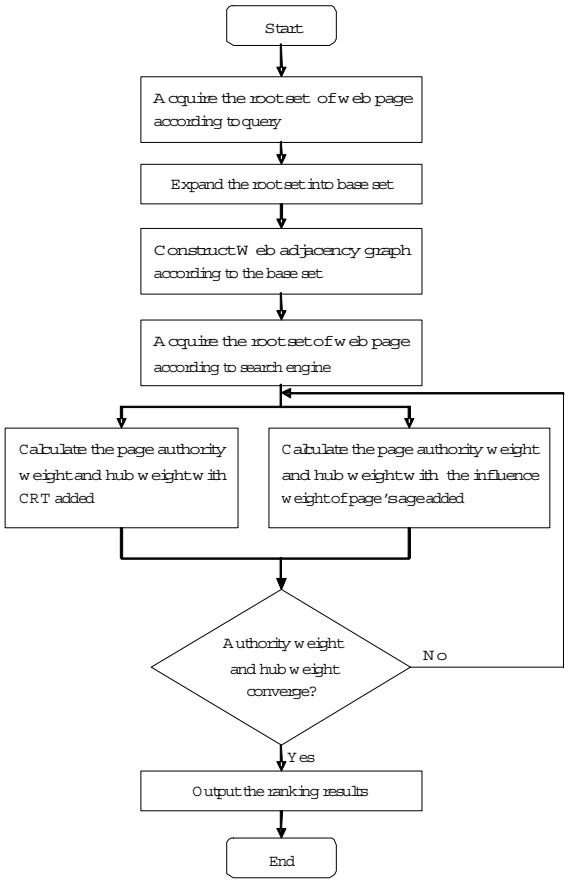

Figure 3 The processing of improved HITS integrated with CTR and influence of page's age

hand, the influence weight of page's age is added in the authority value calculation to overcome the emphasis on old pages. Thus, it can not only highlight the new high-quality pages, but also distinguish the old high-quality pages updated continuously and the old high-quality pages out of date. It can meet the users' timeliness demand of searching better. Therefore, the page rank generated by the integrated improved HITS algorithm is better than single improved algorithm in meeting the users query demands and providing the latest information sources.

4. Experiment and result evaluation

4.1. Experimental design

The aim of the experiment is to examine the efficiency of the improvements of HITS. The experimental design follows these principles: first, experimental process is conducted in strict accordance with algorithm design; second, some empirical data in experimental program refer to the experiments in related research; third, when it comes to classical operations in the algorithm of search engine, it adopts general algorithm or realized works; fourth, due to the limitation of the experimental conditions, the algorithm does not adopt integrated program, and conducts as several parts.

The experiment processes include:

1. Algorithm implementation. According to the process of HITS algorithm, algorithm implementation mainly includes Web adjacency graph and iterative computations. In the algorithm implementation, a search topic is submitted to a traditional text search engine (the Alta Vista search engine is adopted here). The result list returned is regarded as the root set of Web pages. Then in order to expand the root set into a base set, use the network crawler software jspider to extract the External links in the root set of pages and remove internal links, broken links and irrelevant links. Finally, construct the Web adjacency graph based on the base set.

By means of iterative computations, calculate the page rank results of the original HITS, improved HITS with CTR considered, improved HITS with the influence weight of page's age considered, improved HITS integrated with CTR and the influence weight of page's age respectively.

2. Topic selection. In the research of HITS algorithm , 30 query topics [3, 5, 16] as follows are widely used: vintage car, recycling cans, jaguar, alcohol, Thailand tourism, parallel architecture, stamp collecting, telecommuting, sushi, abortion, classical guitar, Lyme disease, bicycling, field hockey, amusement park, table tennis, rock climbing, Olympic, computer vision, Shakespeare, cruise, gulf war, gardening, cheese, HIV, affirmative action, mutual funds, graphic design, architecture, basketball, Artificial Intelligence. The paper select six topics from them to do the experiment: alcohol, abortion, Olympic, gulf war, HIV, Artificial Intelligence.

3. The result evaluation. It's a subjective concept that evaluates the quality of page relativity in page rank results. Manual evaluation is still widely used at present [13]. The paper also adopts similar evaluation method. When analyze improved HITS with CTR considered which improve the topic drift of the original HITS, adopt the main check index of search engine relativity–recall ratio and precision ratio; When analyze improved HITS with the influence weight of page's age considered which improve the original HITS leaning to the old web pages, list the top five of page rank results, and evaluate the content of the pages by contrast. At last, aimed at improved HITS integrated with CTR and the influence weight of page's age, according to the idea of P@10 (Precision at 10) method proposed by Text Retrieval Conference (TREC) [12], a two-layer binary judgment that is similar to it is proposed. Additionally, the concept of satisfaction pages is introduced. Finally, the experiment results of the improved algorithm are evaluated from the aspect of comprehensive relevance as well as the aspect of timeliness.

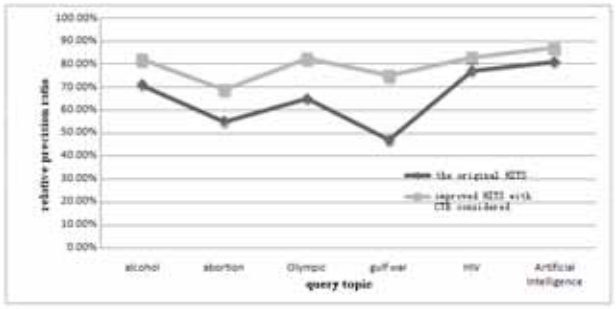

Figure 4 The precision of the original HITS and the improved HITS with CTR considered

4.2. The experiment and evaluation of the Improved HITS with CTR considered

For each query topic: alcohol, abortion, Olympus, gulf war, HIV, Artificial Intelligence, calculate the page rank results of the original HITS and the improved HITS with CTR considered respectively. Then use the recall ratio and precision ratio to conduct the comparative analysis on the results.

There are mainly two indexes in search engine evaluation: recall ratio and precision ratio. However, as the traditional calculations of recall ratio and precision ratio are not operable in the Web, the relative recall ratio and relative precision ratio [17], are used in paper.

The main data used in the calculation of relative recall ratio is the number of the Web pages, which is acquired in the process of constructing the adjacency graph. However, as the HITS algorithm with the CTR added mainly improved the iterative calculation process, the improved algorithm and the original algorithm share the same relative recall ratio. Recall and precision are the indexes that restricts mutually. For a certain search system, the precision decreases with the increasing of recall. Therefore, under the condition of the same recall, the results can be evaluated by comparing the precision directly.

For the six query topics, calculate the precision of the original HITS and the improved HITS with CTR considered respectively. Then compare the results.

It can be seen from the Fig. 4 that the line of the original HITS has a large fluctuation due to the topic drift. However, despite the lower precision of topic abortion, the precision of improved HITS with CTR considered is smooth generally and is higher than that of the original HITS. The addition of feedback information can improve the topic drift and acquire a higher correlation of page rank result than the original HITS obviously.

Table 1 Rank list on alcohol acquired by the original HITS

Rank	Pages
1	http://www.niaaa.nih.gov/
2	http://ncadi.samhsa.gov/
3	http://faculty.washington.edu/ chudler/neurok.html
4	http://www.alcoholfreechildren.org/
5	http://www.nasadad.org/

Table 2 Rank list on alcohol acquired by the improved HITS with page's age considered

Rank	Pages
1	http://www.alcoholfreechildren.org/
2	http://faculty.washington.edu/chudler /neurok.html
3	http://www.nasadad.org/
4	http://www.alcoholconcern.org.uk/
5	http://www.wrap.org/

Table 3 Rank list on abortion acquired by the original HITS

Rank	Pages
1	http://www.prochoice.org/
2	http://www.policyalmanac.org/
3	http://www.adoption.org/
4	http://www.adoption.com/
5	http://www.gynpages.com/

4.3. The experiment and evaluation of the Improved HITS with the influence weight of page's age considered

For each query topic: alcohol, abortion, Olympus, gulf war, HIV, Artificial Intelligence, calculate the page rank results of the original HITS and the improved HITS with the influence weight of page's age considered respectively. Then analysis the top five pages acquired by these two algorithms for each query topic comparatively.

The Tab. 1 shows that the query results on alcohol are mainly organizations pages that publicize the harm of alcohol. They are relevant to alcohol. There are some difference between the Tab. 1 and Tab. 2. The first and the second pages on the list are old and are not update in time. This leads their absence in the Tab. 2. The pages in Tab. 2 are authoritative and updated in time. So the improved HITS with the influence weight of page's age considered can meet the timeliness demands better.

The Tab. 3 shows that the top drift takes palace on abortion, for the third and fourth pages deviate from the topic. However, as the improved HITS with the influence weight of page's age considered is aimed at solving the timeliness problem, it keeps some deviation. Comparing the Tab. 3 and Tab. 4, it can be found that most pages are

Table 4 Rank list on abortion acquired by the improved HITS with page's age considered

Rank	Pages
1	http://www.gynpages.com/
2	http://www.now.org/issues/abortion/
3	http://www.prochoice.org/
4	http://www.adoption.com/
5	http://www.policyalmanac.org/

Table 5 Rank list on Olympic acquired by original HITS

Rank	Pages
1	http://www.olympic.cn/
2	http://www.olympic.org/
3	http://www.olympic.edu/index.htm/
4	http://www.vancouver2010.com/ paralympic-games/
5	http://www.cbssports.com/u/olympics/

Table 6 Rank list on Olympic acquired by the improved HITS with page's age considered

Rank	Pages
1	http://www.olympic.cn/
2	http://www.olympic.org/
3	http://www.cbssports.com/olympics/ photos/OTHER
4	http://www.vancouver2010.com/ paralympic-games/
5	http://www.olympic.edu/index.htm

about abortion policy and protecting women's rights. The latest authoritative pages are placed at the top position by the improved HITS to meet the users' demand.

The third page in Tab. 5 is the Olympic university, which is deviate from Olympic. The rest are the home pages of the Olympic committee or news and have strong relevance. Compare the Tab. 5 and Tab. 6, the rank list in Tab. 6 does not change much. The first and the second pages stay the same, while the third and the fourth pages are replaced by the Winter Games in Vancouver. Compared with the original HITS, the improved HITS can acquire the latest authoritative information sources better.

There is some business information in Tab. 7. The topic drift takes place and the pages are old. In Tab. 8, there are some pages that don't exist in Tab. 7, such as the pages about the films of gulf war and the information about gulf war. The improved HITS not only highlights the authoritative pages updated in time, but also reduces topic drift to a certain extent, for it reduces the weights of the old disturbing pages that link each other tightly.

The rank list in Tab. 9 contains governments, schools and business websites, and they all provide the authoritative information about HIV. Compare Tab. 9and Tab. 10,

Table 7 Rank list on gulf war acquired by the original HITS

Rank	Pages
1	http://admin-amos.shop.com/ss_sign_in +2260.xhtml
2	http://www.law.cornell.edu/uscode/17/107.shtml
3	http://www.pbs.org/wgbh/pages/frontline/ gulf/index.html
4	http://www.ngwrc.org/
5	http://www.vetshome.com/army_national _guard_patches_his_2.htm

Table 8 Rank list on gulf war acquired by the improved HITS with page's age considered

Rank	Pages
1	http://www.gulfweb.org/
2	http://www.pbs.org/wgbh/pages/frontline /gulf/index.html
3	http://admin-amos.shop.com/ss_sign_in +2260.xhtml
4	http://www.ngwrc.org/
5	http://www.gulfwarvets.com/

Table 9 Rank list on HIV acquired by the original HITS

Rank	Pages
1	http://www.cdc.gov/hiv/default.htm
2	http://www.thebody.com/
3	http://www.hopkins-aids.edu/
4	http://hivinsite.ucsf.edu/
5	http://www.gmhc.org/

Table 10 Rank list on HIV acquired by the improved HITS with page's age considered

Rank	Pages
1	http://www.gmhc.org/
2	http://hivinsite.ucsf.edu/
3	http://www.cdc.gov/hiv/default.htm
4	http://www.thebody.com/
5	http://www.aidsmap.com/cms1038153.aspx

most pages keep the same. The third pages in Table 9 don't exist in Tab. 10 for without updating for a period of time. However, the fifth page in Tab. 9 gets the first place for it contains the latest information.

There is nearly no difference between the Tab. 11 and Tab. 12 on topic Artificial Intelligence. Both of these two algorithms acquire the strong relevant pages, such as institutions, laboratories in universities and comprehensive websites. The third page in Tab. 11 takes the firth place in Tab. 12 for its publishing the latest information.

Through the comparative analysis on the rank lists of these topics acquired by the original HITS and the im-

Table 11 Rank list on Artificial Intelligence acquired by the original HITS

Rank	Pages
1	http://www.aaai.org/home.html
2	http://library.thinkquest.org/ 2705/?tqskip1=1
3	http://www.csail.mit.edu/
4	http://ai-depot.com/
5	http://sigart.acm.org/

Table 12 Rank list on Artificial Intelligence acquired by the improved HITS with page's age

Rank	Pages
1	http://www.csail.mit.edu/
2	http://www.aaai.org/home.html
3	http://library.thinkquest.org/ 2705/?$tqskip1 = 1$
4	http://ai-depot.com/
5	http://sigart.acm.org/

proved HITS with the influence weight of page's age considered, we can conclude that: the addition of time factor not only highlights new authoritative pages, but also differentiates the old authoritative pages updated continuously from the old pages out of date. It overcomes the emphasis on old authoritative pages and meet the users' timeliness needs better.

4.4. Experiment and evaluation on improved HITS integrated with CTR and the influence weight of page's age

For the six search topics: alcohol, abortion, Olympic, gulf war, HIV, Artificial Intelligence, calculate page rank results of the original HITS, the influence weight of page's age respectively. Then, combined with the page rank results of improved HITS with CTR considered mentioned in 4.2 chapter, improved HITS with the influence weight of page's age considered mentioned in 4.3 chapter, compare and analyze the four algorithms.

Text Retrieval Conference proposed the evaluation criteria P@10 $Precisionat10$ of information retrieval technology: P@10 refers the precision of top 10 results in the search results list. When the users are viewing results of search engine, in ideal circumstances, they can find the information they need at the first page $usually10results$. So that it set such a personified index. P@10 can reflect the search engine's performance in actual application effectively, and it's widely used. According to the idea of P@10 method, the relevant pages are acquired first, followed by the satisfaction evaluation of the relevant pages. The users' satisfaction is a deeper and more comprehensive concept compared with the relevance. It demands not

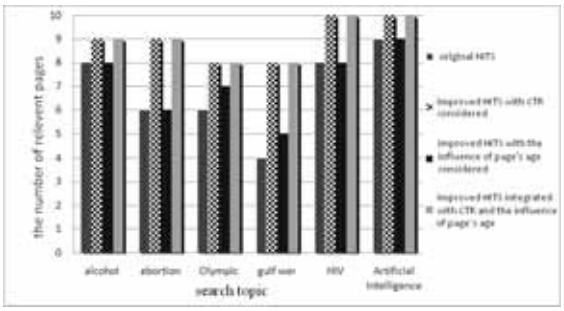

Figure 5 The relevance comparison histogram of the four algorithms

only the relevance to topic, but also the containing of the latest authority information.

The paper adopts the two-layer binary judgment. Firstly, divide the pages into two types: relevant pages and irrelevant pages. Then use the binary judgment again to divide the relevant pages into two types: satisfying pages, which contain new information; unsatisfying pages, where the content is old-fashioned.

In the evaluation, the Pooling method, which is generally used in TREC [12], is adopted. Its idea is: for each search topic, merge the search results of all algorithms that are going to be evaluated into a page pool. As the candidate set that maybe relevant to the search topic, after removed the repeated pages, the page pool are sent to the evaluation group to judge. An evaluation group of nineteenth volunteers was set up to evaluate the top 10 pages in the page ranking list. Then sum the evaluation of each page. If the sum of evaluation of a page is greater than 9, this page can be judged as relevant page. Otherwise, it is judged as irrelevant page. Similarly, if the sum of satisfaction of a relevant page is greater than 9, it can be judged as satisfying page, otherwise it is unsatisfying page.

For the six search topics, calculate the sums of evaluation of the top 10 pages of the original HITS, improved HITS with CTR considered, improved HITS with the influence weight of page's age considered, improved HITS integrated with CTR and the influence weight of page's age respectively, and then compare and analyze the results.

It can be seen from the Fig. 5 that there are different degrees of topic drift in the topic of abortion, Olympic and gulf war. For these topics, improved HITS with CTR considered gets a good evaluation, as well as the improved HITS integrated with CTR and the influence weight of page's age. For the topics of alcohol, HIV, and artificial intelligence, all of the four algorithms get good evaluations. This experiment shows that the improved HITS with CTR considered has a certain advantage, together with the improved HITS integrated with CTR and the influence weight of page's age.

It can be seen from the Fig. 6 that for the topics of alcohol, HIV and Artificial Intelligence, improved HITS with

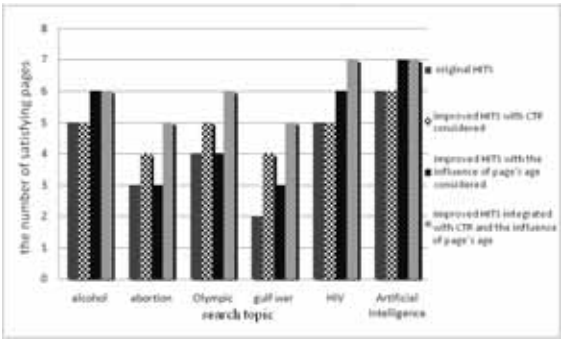

Figure 6 Satisfaction comparison histogram of the four algorithms

the influence weight of page's age considered gets a good evaluation. For these topics, in Fig. 5, the relevant analyses of all these four algorithms are good. However, the satisfaction evaluation not only considers the relevance, but also the timeliness. Therefore, the two algorithms that consider the influence weight of page's age are better naturally. For the topics of abortion, Olympus and gulf war, where the topic drift takes place, the algorithms that consider the CTR are still better. To sum up, for the indexes of relevance and timeliness of page, the improved HITS integrated with CTR and the influence weight of page's age gets a better evaluation. Its result can meet the users' demands of relevance and timeliness at the same time.

5. Conclusion

To the problems in the original HITS algorithm, the information of web click and timeliness was introduced to improve it. Three improved algorithms, including improved HITS with CTR considered, improved HITS with the influence weight of page's age considered and improved HITS integrated with CTR and the influence weight of page's age were proposed in this paper. The experiment showed that the improved algorithms had a certain advantages over the original algorithm. It is left to future work to combine the Web usage mining and link and analysis algorithm to make more fully use of the usage information and provide personalized search service. This is one of important directions of search engine development. Meanwhile, in the synthesis algorithm, it can try more combinations, such as linear combination, to further improve the synthesis algorithm.

Acknowledgement

The authors acknowledge the financial support Humanities and social science planning fund project of the Ministry of education of China, project No. 11YJA630029. This paper was also supported by the National Science Foundation of China under Grant Number 71001075, 71131006 and 71020107027. The author is grateful to the anonymous referee for a careful checking of the details and for helpful comments that improved this paper.

References

[1] H. Qin, Web-Based Data Mining. Journal of University of Electronic Science and Technology of China, **31**, 7 (2002).
[2] L. Liu, The Meta Search Engine Research on Topic Distillation Algorithms, (Liaoning Technical University, Shenyang, (2010).
[3] J. Kleinberg, Authoritative Sources in a Hyperlinked Environment, Journal of the ACM, **45**, (1999).
[4] Y. Jiang, Algorithm Research for WEB Structure Mining Based on Hyperlink, (Xidian University, Xian, 1999).
[5] S. Chakrabarti, B. Dom, P. Raghavan, etal., Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text, Computer Networks and ISDN Systems, **30**, (1998).
[6] A. Borodin, O. Gareth, etal., Finding Authorities and Hubs From Link Structures on the World Wide Web, In Proceedings of the 10th international conference on World Wide Web, (2001).
[7] K. Bharat, etal., A Comparison of Techniques to Find Mirrored Hosts on the WWW, American Society for Information Science, **51**, (2001).
[8] F. Liu, Study of HITS Algorithm in Web Hyperlink Analysis, (Dalian University of Technology, Dalian, 2006).
[9] J. Liu, Research on HITS Algorithm of Web Structure Mining, (Central South University, Changsha, 2008).
[10] N. Wu,J. Zhang, Research on a Fuzzy Clustering for Web Log Mining, Science Journal of Harbin Normal University, **19**, (2003).
[11] X. Wang, X. Zhang, X. Li, The Application of Time Parameter in HITS Algrithom, Modern Computer, 6 (2006).
[12] C. Zhang, Research and Improvement on Link Analysis Based on HITS Algorithm, (Dalian University of Technology, Dalian, 2007).
[13] K. Bharat, M. 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,(1998).
[14] J. Xu, X. Yang, A Study on the Ranking Algorithms of Specific Search Engine, New Technology of Library and Information Service, 7 (2006).
[15] Ayaz Isazadeh, Habib Izadkhah, A. H. Mokarram. A Learning based Evolutionary Approach for Minimization of Matrix Bandwidth Problem, 6, 51-57 (2012).
[16] R. Lempel, S. Moran, The Stochastic Approach for Link-Structure Analysis(SALSA) and the TKC Effect, Computer Networks, 1 (2000).
[17] Y. Feng, Z. Liu, J. Wang, Research on Information Retrieval Evaluation Measure System for Search Engine Functions, Journal of the China Society for Scientific and Technical Information, 1 (2004).

Yue He is a leading figure in data mining and management information system and is presently employed as Professor at SCU, China. He obtained his PhD from SCU, China. He is the general secretary and executive member of the Enterprise branch of China Operations Research Society, the executive member of Sichuan Province Association of quantitative economics, leader of SCU Information management and decision making Research Institute. In recent 5 years, Prof. He took part in 5 projects supported by the National Natural Science Foundation of China, Social Science Fund, leaded 1 project supported by Humanities and social science planning fund of Chinese Ministry of Education, also leaded more than 20 research projects of Sichuan government and some other enterprises and institutions. He published more than 100 research papers in domestic and foreign academic journals, over 10 textbooks. Prof. He obtained first prize of Sichuan province outstanding software, second prize of Sichuan province science and technology progress prize for the year 2005, third prize of Sichuan university excellent teaching award.

Minghong Qiu was born in 1985. She received his Bachelor's degree in management from Sichuan University of China. She is currently a Master degree candidate at Sichuan University. Her research interests include data mining and management information system.

Dr. Maozhu Jin is an instructor of Business School,the tutor of MBA operations management and innovation and entrepreneurship management in Sichuan University. His current research interests include the areas of operations management, service operations management, platform-based mass customization and risk management. He has presided over three ministerial and provincial projects, one project supported by National Natural Science Foundation of China, named Research on service modularity based on service platform under the circumstance of mass customization. As a main researcher, he has participated in and completed three projects supported by National Natural Science Foundation of China and two surface projects. He has published two books and over ten research papers in authoritative journals of high quality both at home and abroad, and ten of them are retrieved by SCI and EI.