Applied Mechanics and Materials *Online: 2011-10-24*
ISSN: 1662-7482, Vols. 121-126, pp 3741-3744
doi:10.4028/www.scientific.net/AMM.121-126.3741

A novel improved web structure mining algorithm

ZHAO Li-dong[1, 2, a], LIU Sheng-Hui[1, 2, b], CHAI Qing[1, 2, c]

[1.] Lanzhou Polytechnic College, Lanzhou 730050,

[2.] Key Laboratory of Gansu Advanced Control for Industrial Processes, Lanzhou 730050

[a] zhangli3607@163.com, [b] liush@lzptc.edu.cn, [c] chaiq@lzptc.edu.cn

Keywords: web mining; HITS; time parameter

Abstract. In this paper, a compact and effective web structure mining algorithm is developed according to the hypertext induced topic search (HITS). Firstly, the environments of traditional web structure mining are explored. Secondly, by analyzing the existing techniques, the existing problem and the lack of previous methods are explored. Finally, amid to overcome the deficiency of the traditional method, the paper introduced the time parameter and proposed a novel web structure mining algorithm. It turns out that the proposed algorithm has better effective and accuracy compared to the traditional methods.

Introduction

There has been a dramatic boom of web data volume developed with the information in past few years, especially with the development of computer and internet technique. Web page or web sites ranking is an important part in the web information data mining, which related to user access recods through a group of web pages, or other similar cross referenced resources. For example, when an internet search engine is used, a ranking is used to sort the resulting pages in terms of relevance and authority [1]. So the web structure mining has become a hot focus in web data mining. HITS is a authorities and typical algorithm in the past few years, the basic think is that using the web structure link mine useful web cites in all of the web pages. Otherwise, the existing methods have the character of computing simple and objectiveness. For example, the HITS algorithm exist evident shortage, first of all, the algorithm only using the link structure to consider the importance of web cite or web page, while exception for the web page link structure, according to the practice, the content of web cite is the other very important factor to authorities cite[2]. Secondly, other type of link includes links that do not convey people's opinions such as automatically generated links, links for navigation, and banner ads. This type of link causes a problem known as topic drift, by which the most highly ranked authorities and hubs tend not to be about the topic of the given query but tend to be very popular pages such as top pages of google or yahoo[3,4]. Many researchers have tried to solve the topic drift problem such as Chakrabarti and Nishimura, and also achieve some good results, but they all ignored access time to the interesting page or sites, So we must consider the authority page include access time of page content, to solve above problem, in this study, we specifically examine the HITS algorithm and improved it by introducing time parameter, proposed a novel web structure mining algorithm and achieve a high performance in web structure data mining.

The basic think of HITS

A. The basic HITS algorithm thinks

HITS is the most important and authoritive algorithm web structure mining algorithm at present, which using the web link structure information to mining authoritive page or nodes, and has widely used in high performance search engineer such as baidu or google sites. People use a page's popularity to solve the abundance problem; technically web search engines use hyperlinks for estimation the page's popularity [5,6].

The basic idea of HITS is firstly category all web sites page into two classes; the one is authority pages which express most important content information in certain area, and linked to by many other good pages, i.e www.sohu.com, www.hao123.com or www. sina.com. the other is linked to these

authority page such as contain many interesting content for the given query web sites or page linked to these authority page, we usually called them hub page. Figure1 illustrated the principle of these two type pages.

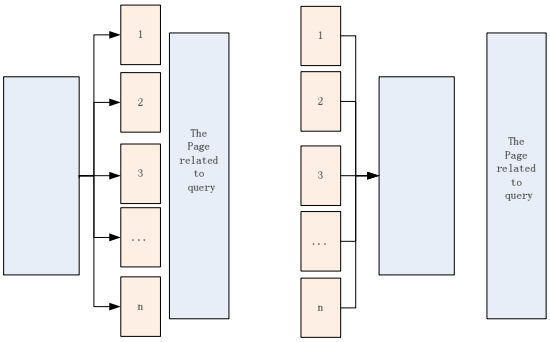

Figure 1: Hub page and authority page

Basically, the process of using hub pages to find authority page are described as following procedure:

Firstly, we can get a set of initial results sets according to query key word by search engines (we called these pages root sets or start sets). More of this pages related to search content or rely on the opinions of people who create links on the web; this page should contain the useful pointer to authority page, so root sets may extent to base sets, which include all root sets and all page that pointed to root sets.

Secondly, the algorithm does propagating weight: this process is a recursive program, aimed to make the value of hub and authority weight. At first, we set a nonnegative authority weight value ap and hub weight value hp to every page of based sets and initiate it to a constant. The weight computing method can be computing as following equation.

$$ap = \sum_{q:q\to p} hq \tag{1}$$

$$hp = \sum_{q:q\to p} aq \tag{2}$$

Equation (1) illustrate that if a page which be pointed by many hub, the weight can be increased. Equation (2) illustrate that if a page pointed to authority page, the hub weight can also be increased.

Finally, HITS algorithm output a set of page which has high weight and high authority. If we using digraph describe web link structure, the graph include a set of nodes and a set of edges, the subset s in the node sets and the edge between node form web sub graph, HITS introduce authority weight and hub weight, and finally output the foremost important page. If we describe web page using tab $\{1,2,...,n\}$, and also define its $n*n$ adjacency matrix. If page i point to page j, the value of item (i,j) equal 1, otherwise, the value equal zero. Similarly, we also define value of authority and hub as vector, $x=(x1,x2...xn)$, y=(y1,y2…yn), the equation (1) and (2) can be represented as equation (3) and (4).

$$x \leftarrow A^T x \tag{3}$$

$$x \leftarrow A^T y \tag{4}$$

If we continue expand the equation (3) and (4), equation (5) and (6) can be obtained.

$$X \leftarrow A^T Y \leftarrow A^T Ax \leftarrow (A^T A)x \tag{5}$$

$$Y \leftarrow A^T X \leftarrow A^T Ay \leftarrow (A^T A)y \tag{6}$$

We can see obviously that vector x and vector y can be obtained by equation (3) and (4)

B．The improved algorithm based on HITS

Although the HITS algorithm introduced link structure creatively and gets a very good effective in practice. There do also some drawbacks. At first, the HITS algorithm does not consider page content, the main reason is that the algorithm set same link weight that is if two pages have link, the value in the adjacent matrix equal 1, while it is not reasonable in fact. Secondly, it is one side that the authority page weight increased if one page cited by hub page. Actually, there are many no-normal cited links such as advertisement. This must result in HITS algorithm error. The main reason is that we not classify the authority cited and virtual cited.

Amid at the first limitation, many researchers have improved the algorithm using the importance of the web page content. He assumed that the importance of the web page depend on the user request, and consider the result from the hub weight. Krishna Bharat and Monika R.Henzinger modify the authority and hub weight by introducing the relevance weight. If relevance weight value lowers than a given threshold, we can assume the influence of link omitted to the web page, the link can deleted in the sub graph, the method increased the authority page quality in some extent. The practical application instance of HITS is clever system. By using the proper control parameter and proportion of link text, the system can overcome the above limitation. To the second limitation, there is no good method to solve, so this paper attempted to introducing the time parameter to improved.

The proposed algorithm

A．the introduction of time parameter

To a certain page link such as note p pointed to note q, the link time in some extent decide the authority of cited nodes. In actua life, the user also spent long time at authority cites, otherwise, to the navigation or other condition, user only use little time. If we introducing this principle to the HITS, the performance of algorithm can get better results.

B．the time parameter control model

Defining a page point to page q, the link weight value related request key word k described by $w(p,q,k,t)$, the expression illustrate that the results decided by three factor, the the link that page p point to page q, the number of request key word in the link, and the time that page p access page q.

To control the results accuracy, we can introduce coefficient α to control the proportion of k surround the link text, and $\sqrt{t}$ control the page access time, so the expression can be described as following.

$$w(p,q,k,t) = 1 + \varphi(k) + a * \varphi(k) + \sqrt{t} \tag{7}$$

In above expression, the coefficient α can adjust according to the different page sets, and the value of w can increased in the procedure of computing the authority weight. Since we only consider the Relative size of two parameters, and also for the simple description of expression, we can standardize the value in each process of computing, $\sqrt{t}$ to control the effectiveness of time parameter to weight value. The figure of $\sqrt{t}$ is illustrated as figure2

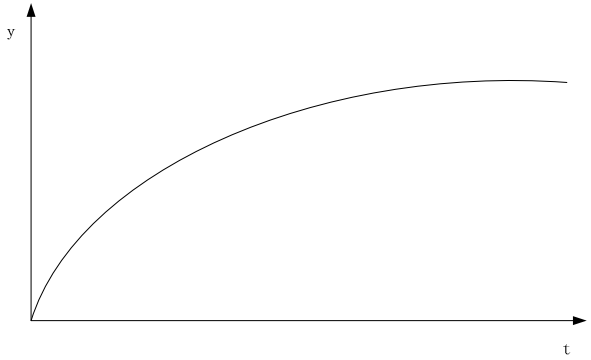

Figure 2: The figure of $\sqrt{t}$

C. The description of algorithm

Input: sub nodes, access squence, time and threshold of time (tmin).

Output: authority weight and hub page

Steps:

Step1:

If (t>tmin)

Computing nodes weight value

$$w(p,q,k,t) = 1 + \varphi(k) + a * \varphi(k) + \sqrt{t} \tag{8}$$

Else

Delete from nodes sets

Step2:

Computing authority weight (ap) and hub weight (hp)

$$ap = \sum_{q:q \to p} hq \tag{9}$$

$$hp = \sum_{q:q \to p} aq \tag{10}$$

Step 3:

Through above steps, output maximal authority weight and hub weight, we can using these output results into other application.

CONCLUSIONS

In this paper, a novel time parameter based web structure mining algorithm is developed. At the base of think to traditional HITS algorithm, the paper improved the limitation of previous method such as topic drifting by introducing the time parameter model, solve the problem of original HITS algorithm, and get a good performance in practical application. In the future work, we consider and explore better parameter model to improve existing algorithm.

References

[1] Matthew Bennett, Julie Stone, Chaoyan Zhang: a scalab le parallel HITS algorithm for page ranking, 2006 First International Multi-Symposiums on Computer and Computational Sciences

[2] Bui Quang Hung, Masanori Otsubo, Yoshinori Hijikata, Shogo Nishida: HITS algorithm improvement using semantic text portionVol. 8, no. 2, pp. 149-164(2010).

[3] Kleinberg, J: authoritative sources in a hyperlinked environment, journal of the ACM, vol.46, no.5,pp. 604-632(1999).

[4] B.Q.Hung. M. Qtsubo, Y.Hijikata and S. Nishida: extraction of semantic text portion related to anchor link, IEICE Trans. informaion and system, vol. E89, no. 6 (2006)

[5] E.J. Glover, K.Tsioutsiouliklis, S.Lawrence, D.M. Pennock, and G.W. Flake: Using web structure for classifying and describing web pages, Proc. 11th International World Wide Web Conference, pp.562-569, (2002).

[6]L. Li. Y. Shang and W. Zhang: improvement of HITS-based algorithm on web documents, proc. WWW 2002, PP. 527-535(2002).