

## Hyperlink algorithm based on anchor texts relevance

XU Jia-shu<sup>1</sup>, XING Li-xin<sup>2</sup>, QIN Zheng<sup>1</sup>

(1. Department of Computer Science and Technology, Xian Jiaotong University, Xian 710049, China;

2. Heilongjiang Radio and TV University, Harbin 150080, China)

**Abstract:** This paper analyzes the disadvantage of the PageRank arithmetic. Calculates the relevance between the query key words and the anchor texts of hyperlink with vector space model. The new hyperlink algorithm is put forward that based on the relevance of the hyperlink anchor texts. The experiment results showed that this algorithm is effective, it can raise the precision of PageRank arithmetic.

**Key words:** hyperlink arithmetic; Web information retrieve; search engine

中图分类号: TP393

文献标识码: A

文章编号: 1672-0946(2005)03-0294-04

## 基于链接文本相关度的超链接算法

徐家树<sup>1</sup>, 邢立新<sup>2</sup>, 覃征<sup>1</sup>

(1. 西安交通大学 计算机科学与技术系, 陕西 西安 710049;

2. 黑龙江广播电视大学, 黑龙江 哈尔滨 150080)

**摘要:** 分析了超链接 PageRank 算法的不足, 采用向量空间模型计算检索关键词与超链接文本之间的相关度, 提出了基于超链接文本相关度的超链接算法。实验结果表明, 该算法可以提高 PageRank 算法的检索精确度。

**关键词:** 超链接算法; Web 信息检索; 搜索引擎

### 1 Introduction

From 20 century 90's, web technology has gotten fast development and application, web pages quantity increase very quickly. Now, web has developed to become important information resource database. What people first think of when retrieving information is that retrieve go to web, web has become the new channel to take the information, but the magnanimity of web information size has brought difficulty for people to get web information<sup>[1-4]</sup>. Therefore, people have carried out extensive and thorough research and practice, purpose is to explore the method of getting web information efficiently, thus, search engine is one of the popular retrieval tools,

as Google, Baidu etc.

Normally, the basic structure of search engine is made of 4 components, Gather (also called: WebRobot), Indexer, Searcher and User Interface (UI). Working principle of search engine can be been briefly in below. WebRobot collects web pages independent. Indexer analyses the web pages collected, and establishes index database. Searcher responds the query of users, will queue up web pages after calculated the relevance between the retrieval keywords and the web pages, and return results to user in UI. Therefore, we can find out that the search strategy of search engine is to get more and yet more web pages as far as possible, and preserve them in index database for retriever to choose web pag-

收稿日期: 2004-11-15.

作者简介: 徐家树(1962-), 男, 博士, 研究方向: Web 技术.

es. So far, the result's effect of search engine is not very ideal. There are the plenty of web pages in the results, the relevance of web pages is lack and the authority of web page is not high. Analyzing the reasons, on the one hand, search engine is to face to masses, therefore has the general or widely, lack to personalize service. On the other hand, it is deficient in the fusion arithmetic of the retrieval results. Google and Baidu are the famous search engine, PageRank arithmetic is the one of key technology.

## 2 Web Graph Model

Entire web structure can regard as a huge graph logically,  $G = (V, E)$ ,  $V$  is the set of nodes, each web page is a node in graph, web pages can regard as the carrier of web information.  $E$  is the set of directional edge, if the web pages contain a hyperlink from  $V_1$  to  $V_2$ ,  $V_1$  and  $V_2$  are a web page, therefore there is a directional edge, Hyperlink ( $V_1, V_2$ ). The anchor texts play the part of the role that navigate and recommend web pages, through the hyperlink can browse and glance over other web pages, so the web pages and hyperlink have formed a huge directed graph. Figure 1 is a simple directed graph.

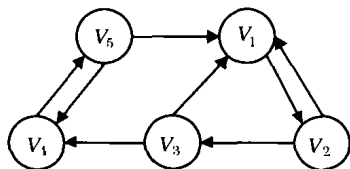


Figure 1 Directed graph

In arithmetic based on the hyperlink analysis, the anchor texts and the quantity of hyperlink have important reference value to appraisement the authority of a web page, therefore should be setup the different weight for each hyperlink. Following, we will introduce PageRank arithmetic, and analysis its existent problems.

## 3 PageRank Arithmetic

PageRank is a typical arithmetic that based on analysis of hyperlink, is supposed based on two following probabilities.

• Many web pages link go to this web page, this web page is important probably. This web page is also important probably, the web page has no repeatedly (or

though the amount of hyperlink is lower), but it is linked by the important web page. The importance of a web page has been average to transmit the web page linked by it. This kind of important web page is called as, is the authority page.

• Set up a web page random visited for user begin to browse the web, afterwards follow the extrovert hyperlink of the web page glance over forward, do not return to back, that is the PageRank value that is the probability to browse the next web page.

PageRank arithmetic can be described following:

$$R(u) = c \sum_{v \in B(u)} \frac{R(v)}{N(v)} + cE(u). \quad (1)$$

$u, v$  is a web page.  $R(u)$  is a set of the web pages that all hyperlinks assemble to  $u$ .  $N(v)$  is the amount of hyperlinks that are the outside hyperlink from  $u$ .  $C$  is the standardize factor, and taken between (0, 1).  $E(u)$  is the attenuation factor.

From Equation (1), we can discover that the advantage of PageRank arithmetic is that can avoid the simple text matching to calculate the importance of the web pages. This arithmetic makes the best of the quantity of hyperlinks and the quality of the web pages through the web graph model  $R$  to evaluate the authoritativeness of the web page, has offered the effective method to measure objectively the web resource.

At the same time, can also find out that PageRank is a static arithmetic, and showed that the quality of web pages is unconcerned with the semantic of queries.

Every web page had been preserved in index database, PageRank value is calculated periodicity for every web page.

After user submits the query keywords, all web pages that are satisfying the query condition according the order of successively decrease of PageRank value to return to UI.

However, possess the higher PageRank value of the web page, and not definitely with the semantic on have correlation between the query keywords submitted and the web pages, in the results of retrieval, can contain the plenty of the pages that are unconcerned with the query keywords.

Generally, in index database of search engine, can retain some information related to web pages, for instance, web page's URL, Title, Anchor texts and the

outside hyperlink number. To raise the accuracy of retrieval can will use these important information properly, calculating dynamic the semantic correlation of the web pages, then according to the order of related value successively decrease, returning those web pages to users. To reduce the complexity of the algorithm, this paper considers the anchor texts only.

The anchor texts of hyperlink can play the part of the role of navigation and the recommendation of the web page. Before doing not enter the web page yet, “surf-boarder” could glance over the web page through the anchor text information, and can know the related information of the Web page probably. Therefore, it can calculate the relevance between the web page and the query keywords using the anchor texts can raise the retrieve accuracy.

#### 4 Hyperlink algorithm based on anchor texts

With VSM (Vector Space Model), calculate the relevance of the keywords and the anchor texts.

In traditional text retrieval system, VSM is the model that to calculate the similarity of two documents. In this model, the vector denoted the document, and the words in document are expressed with the term of the vector, this term value is the weight of this word. Usually, it is the frequency of appearance. By this reason, the combination of query keywords can be also expressed with the weight of vector; the similar degree of the keywords and document is inner product of two vectors. A common measure of similarity is the cosine of the angle between vectors.

Setup is the set, contains anchor texts that link to, is a web page:

$$D = \{d_1, d_2, \dots, d_i, \dots, d_k\} \quad i = 1, 2, \dots, k$$

Any an anchor text contains  $n$  vocabulary totally in, is a set, these vocabularies have formed the  $n$  - dimensional vector:

$$d_i = \{d_{i1}, d_{i2}, \dots, d_{ij}, \dots, d_{in}\}, \quad i = 1, 2, \dots, k, j = 1, 2, \dots, n$$

$d_{ij}$  is the weight of  $j$  - th word in  $d_{ij}$  document,  $d_{ij}$  is the frequency that appears in  $d_{ij}$ ,  $d_i$  is the anchor texts.

By these reasons, the combination of query keywords is expressed for a vector  $n$  - dimensional space  $Q$

$$Q = \{q_1, q_2, \dots, q_i, \dots, q_n\} \quad i = 1, 2, \dots, n$$

$q_i$  is the frequency that  $i$  - th keyword arises in query combination. According to VSM, the relevance of the keywords and  $D$  is the inner product of two vectors,  $D$  is the anchor text. Similarity value calculation is as follows

$$\text{Similarity}(D, Q) = \left[ \frac{\sum_{i=1}^k \frac{\sum_{j=1}^n d_{ij} \cdot q_j}{\sqrt{\sum_{j=1}^n d_{ij}^2 \cdot \sum_{j=1}^n q_j^2}}}{k} \right] / k. \quad (2)$$

Combining the relevance that is calculated by the keywords, the anchor texts and PageRank that is the hyperlink arithmetic based on analysis can create a new algorithm that raises the accuracy, this algorithm can be expressed with Equation (3).

$$\text{Relevance}(d) = R(d) \cdot \{1 + \delta \cdot \text{Similarity}(D, Q)\} \quad (3)$$

Where  $R(d)$  and  $\text{Similarity}(D, Q)$  are calculated respectively by Equation (1) and (2).

$\delta$  is the weight factor of the anchor texts, the scope is (0, 1), its value shows the important level of the anchor text. Usually,  $\delta$  value scope is 0.35 ~ 0.65.

#### 5 Experimental results

The reference [4] had offered the effectual experiment arithmetic, while calculate iteration arrive 19th, PageRank value has tended to steady, then PageRank of 5 web pages are:

$$R(1) = 0.28564, R(2) = 0.28554, R(3) = 0.14265, R(4) = 0.14299, R(5) = 0.14317.$$

The order of the web pages is  $V_1, V_2, V_5, V_4, V_3$ . Suppose, the anchor texts as table 1 shows, the query keyword is “Computer, Teacher”, after calculate the relevance between the anchor texts and the query keywords, the results to see table 2 ( $\delta$  takes 0.5).

The order of web pages is  $V_2, V_1, V_3, V_5$  and  $V_4$ , investigate from the semantic of the anchor texts, that order is reasonable; therefore this algorithm can raise the accuracy of PageRank arithmetic.

Table 1 Anchor texts

Edge	The Anchor text	Edge	The Anchor text
E(1,2)	Xian Jiaotong University	E(3,4)	Books
E(2,1)	Computer Science Dept.	E(4,5)	Education
E(2,3)	Computer Technology	E(5,1)	Information System
E(3,1)	Person	E(5,4)	Teacher

**Table 2 Results of the relevance calculation based on the anchor texts**

Page	Similarity	Relevance
1	0.814 20	0.401 92
2	0.816 50	0.402 11
3	1.000 00	0.213 84
4	0.707 10	0.193 54
5	0.707 10	0.193 79

## 6 Conclusion

In PageRank arithmetic, every hyperlink's weight that to be regarded as identical, and the importance of web pages can be calculated using the amount of hyperlinks.

Therefore, the retrieval effect does not be to so much satisfactory, there are the plant of web pages in the results that are not unconcerned between the query keywords and the web pages. To raise the accuracy of

PageRank, the algorithm was brought forward. First, consider the anchor texts of web pages. Second, calculate the relevance between the query keywords and the anchor texts with the Vector Space Model, and finally integrate PageRank arithmetic to raise the accuracy of retrieve. Emulation experiment results showed that the algorithm can raise the precision of PageRank.

### 参考文献:

- [1] BORDER A, KUMAR R, MAGHOUL F, et al. Graph Structure in the Web [EB/OL]. Proc. WWW9 Conference, May 2000. See also: <http://www9.org/w9cdrom/160/160.html>. 309 - 320.
- [2] PAGE L, BRIN S, MOTWANI R, et al. The PageRank Citation Ranking: Bringing order to the Web [EB/OL]. <http://www-db.stanford.edu/backrub/pageranksub.ps>, January, 1998.
- [3] TAHER H. Efficient Computation of Pagerank. Technical Report 1999 - 31 [EB/OL]. Database Group, Computer Science Department, Stanford University, February 1999. <http://dbpubs.stanford.edu/pub/1999-31>.
- [4] Chao Jun. The Anatomy of Page Rank Technology of Google[J]. Journal of Information, 2002, 21(10): 15 - 18 (In Chinese).

(上接 293 页)

从图 3 可看出,出水中磷的质量浓度较高,其主要原因是由于进行水解酸化,污水中低级脂肪酸的含量增加.它是影响聚磷菌厌氧释磷的一个关键因素.在厌氧条件下,聚磷菌利用细胞内积聚的聚磷(Poly-p)和糖原(Glycogen),在分解代谢时产生的能量和还原动力,吸收细胞外的低级脂肪酸,合成聚 $\beta$ -羟基丁酸(PHB),储存在细胞内,而后在缺氧或好氧的条件下,用于微生物的生长、糖原的合成和维持细胞代谢,还为细胞外磷的吸收及其在体内 Poly-p 的合成提供能量<sup>[5,6]</sup>.所以此时磷的释放是有效的,它有利于生物除磷效果提高.从现场运行看,这几天二沉池出水 TP 都在 1.0 mg/L 以下.

## 3 结 语

根据磷的有效释放和无效释放原理<sup>[7]</sup>,在厌氧条件下,如果 pH 值降低是由于进水中含量 VFA 较高造成的,导致了磷大量释放,则属于磷的有效释放,对生物除磷有利;如果 pH 值降低是由于进水中含有强酸较多造成的,微生物由于 pH 值降低而中毒造成磷释放,则是磷的无效释放,磷的无效释放对生物除磷是有害的,释放出磷在好氧池中不能

被聚磷菌再吸收.在城市污水处理厂中,由于汇入大量工业废水,导致 pH 值变化较大,这种现象在生物除磷中要尽量避免.

### 参考文献:

- [1] 任南琪,周大石,马放.水污染控制微生物学[M].哈尔滨:黑龙江科学技术出版社,1997.
- [2] KUBA T, VAN LOOSDRECHT M C, HEIJNEN J J. Biological dephosphatation by activated sludge under denitrifying conditions: pH influence and occurrence of denitrifying dephosphatation in a full-scale waste water treatment plant[J]. Wat. Sci. Tech, 1997, 36(12): 75 - 82.
- [3] 国家城市给排水工程技术研究中心译.污水生物与化学处理技术[M].北京:中国建筑工业出版社,2001.
- [4] SMOLDERS G J, VAN DER MEIJ F. Model of the anaerobic metabolism of the biological phosphorus removal process: Stoichiometry and pH influence[J]. Biotechnology Bioeng, 1999, 44: 837 - 848.
- [5] HU Zhirong, WENTZEL M C. Anoxic growth of phosphate-accumulating organisms (PAOs) in biological nutrient removal activated sludge systems[J]. Wat. Res, 2002, 36: 4927 - 4937.
- [6] CHANG W C, CHIOU R J. Effect of anaerobic conditions on activated sludge filamentous bulking in laboratory systems[J]. Wat. Res. 2001(12): 1541 - 1546.
- [7] 郑兴灿,李亚新.污水除磷脱氮技术[M].北京:中国建筑工业出版社,1998.