Study on Link-Based Approaches for Web IR in TREC Experiments*

ZHANG Min[1] MA Shaoping[1] GAO Jianfeng[2]

[1](State Key Lab. of Intelligent Tech. and Sys., CST Dept. Tsinghua University, Beijing 100084, China);

[2](Microsoft Research Asia, Beijing 100080, China);

E-mail: zhangmin99@mails.tsinghua.edu.cn

Abstract: This paper studied the effects of using of link information for Web IR in TREC experiment, including link anchor text, link structure and the combination of link-based retrieval and traditional content-based retrieval. Several conclusions are drawn: Firstly, anchor text can represent precisely the topic of web page, but insufficient in describing the web page content. Secondly, comparing with traditional content-based IR technique, using link-based approach on homepage finding task can get more than 96% improvement, while it is not helpful on ad hoc task. Finally, combining link-based and content-based techniques makes consistent improvement on homepage finding task and a little progress by result re-ranking on ad hoc task.

Key words: link-based retrieval; link approach; Web IR; ad hoc; homepage finding

Introduction

One of the most important characteristics of Web Information Retrieval (IR) compared with traditional IR lies in the hyperlink structure. This motivates the so-called link-based retrieval techniques for web IR. Former researches showed that using link information can help the retrieval in web environment[1][2][3]. This conclusion, however, almost never works in Text Retrieval Conference (TREC) web track experiments up to now, except for one research on effective site finding[4]. In this paper, we investigated the use of two kinds of link information, namely link anchor text and link structure, using TREC10 collection which contains over 1.69 million web pages.

In the remainder of this paper, we first introduce the concept of anchor description document and discuss the use of link anchor text in Web IR in section 1. Then we observe the effect of using link structure in section 2. After that, three different approaches of combing link information and content information are exploited in section 3. Results of TREC experiments are described in each section to help explore the issues. Finally, we give our conclusions and present the future work.

1 Using Anchor Text

1.1 Anchor description document

In former researches, it is claimed that texts around links to a page p are descriptive of contents of page p[5]. Therefore anchor text of a link describes its target webpage while not the source page. This is the important assumption that our anchor-description-oriented approaches based on.

We firstly extracted all the anchor texts of each webpage that are used when the page linked by other pages in the whole collection. Then each page's anchor texts were merged into a new document which is called *anchor*

* Supported by the Chinese National Key Foundation Research & Development Plan (grant G1998030509), Natural Science Foundation No.69836040

description document. It represents the role of one page in the collection, and stands a good chance to be longer if there are more pages linked to this page. At the same time, the more centralized the topic of one page is, the more prominent the anchor description document is.

1.2 Experimental results of Using Anchor Text

In this method, anchor description documents are used alone to build the index. The original collection is about 10G, while the anchor description document collection is only 250M or so.

It is easy to understand that in this way, we lost too many information so that the results were bad for web track ad hoc task as shown in Table 1, compared with the content-based retrieval result 22.08% (11-point average precision). The parameter $QE(x,y)$ in the table means top y term(s) in top x document(s) got by first retrieval result is selected for pseudo relevance feedback. The reason is due to the data sparseness problem [8]. Statistics showed that about 28% web pages in the web collection have no anchor description at all. 47% anchor description documents are less than 10 words. Therefore the information that anchor text provided for ad hoc retrieval is quite limited. It is most unlikely that queries have chances to match words contained in such a short description.

While for web track homepage finding task, the result is astonishing. Using anchor text only performs much better than content based retrieval, i.e. about 96% improvements on average reciprocal rank and 48% improvements on top-10 precision, shown in Table 2. The result is reasonable. We observed that many anchor text are names of home pages (i.e. URL, or URL-like terms). Intuitively, they could be effective for page finding task, in which most queries are also a bunch of URLs or URL-like terms. Our results confirmed the intuition.

Table 1 Average precision of using anchor text only for TREC web track ad hoc task

	One-step retrieval	*QE* (5,5)	*QE* (5,10)	*QE* (5,20)	*QE* (5,30)
Using anchor text	3.12 %	2.97 %	4.94 %	4.48 %	4.85 %

Table 2 Performance of using anchor text only for TREC web track homepage finding task

	Average reciprocal rank	Top 10 precision	Not-found proportion
Content based retrieval	22.46%	44.1%	25.52%
Using anchor text	44.06%	65.5%	25.52%

2 Using of Link Structure

2.1 Spreading activation

On using link structure, the important assumption is links between documents indicate useful semantic relationships between web pages. Depending on that, we tried spreading activation (SA) approach [6] [7] and observed the effects of using this method on TREC10 web track ad hoc task. In SA method, the degree of match between a web page D_i and a query Q, as initially computed by the IR system (denoted $SIM(D_i,Q)$), is propagated to the linked documents through a certain number of cycles using a propagation factor [7]. Our experiments showed that considering outgoing links will give negative effect on retrieval results. Therefore only the most similar incoming link is considered in our method. In this case, the final retrieval value of a document D_i with m incoming linked documents is computed as:

$$\text{SAscore}(D_i) = SIM(D_i,Q) + \lambda \cdot \max\{SIM(D_j,Q) \mid j = 1,...m\} \quad (1)$$

2.2 Experimental results of SA

We made the full study on SA approach applying on TREC10 web track ad hoc task whose results were shown in Figure 1. When SA weight λ is zero, The *SAscore* is the same as that of content-based retrieval. The performances are worse with increase of λ. Therefore the conclusion is this approach could not help to improve the retrieval effectiveness in TREC web track ad hoc task.

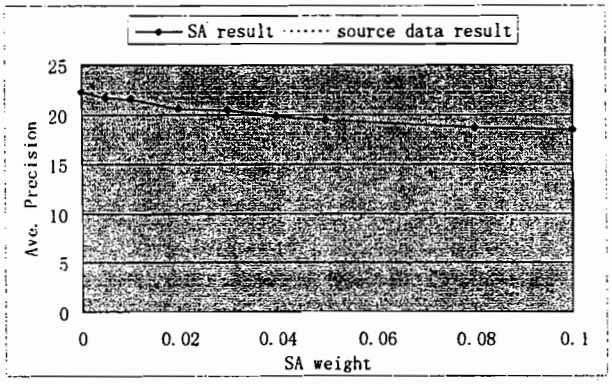

Figure 1 The effect of Spreading Activation approach on TREC10 web track ad hoc task

3 Combine link-based retrieval and content-based retrieval

In what follows, we propose three combination approaches. First, combine anchor description document and initial source data into a full corpus to build index. Second, combine anchor-description-only retrieval result and full corpus retrieval result. Third, re-rank the results of full corpus retrieval according to anchor-description-only result.

3.1 Combing Corpus

As mentioned before, anchor text of a link describes its target page. This makes it reasonable to index anchor text as though it occurred in the target document, rather than the source. Based on this viewpoint, we merged the two kinds of corpus, i.e. anchor description document and source document to build index and retrieve queries. The influence on web track ad hoc task in TREC data and queries is shown in table 3, which achieved a little improvement. Results of TREC10's web track homepage finding task are shown in table 4. It is clear that in homepage finding tasks combined corpus was also much better than only using source data.

Table 3 Results of combing anchor text and source data on TREC9 web track ad-hoc task

	Source data only	Anchor description + source data
One-step retrieval	22.08 %	22.23 %
With query expansion	22.21 %	22.84 %

Table 4 Results of combing anchor text and source data on TREC10 page finding task

	Average reciprocal rank	Top 10 precision	Not-found proportion
Source data only	22.46 %	44.14 %	25.52 %
Corpus combing	42.4 %	65.52%	13.1 %

3.2 Combing Results

With method 1.1 (anchor text only indexing) and method 3.1 (combine corpus indexing), we can get two ranking lists to each query. This method is to linear combine these two ranking lists to form the final results. Suppose scores of documents in the two ranking lists are *sa* and *sc* respectively, then the combined score is:

$$s = \lambda * sc + (1-\lambda) * sa \qquad (2)$$

where λ ($1 \geq \lambda \geq 0$) is the combination weight.

The average precision of this method on TREC web track ad hoc task is shown in table 5. Experimental results of combing ranking list on TREC-10 homepage finding task is shown in table 6.

The conclusion is for ad hoc task, no positive effect was got, while for home page finding task, rich effectiveness improvement was made.

Table 5 Performance of combing results on TREC web track ad hoc task

λ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
A.P.(%)	22.08	21.55	20.39	19.69	18.13	15.08	10.5	6.47	5.29	5.15	4.98

Table 6 Performance of combing results on TREC 10 homepage finding task

	Average reciprocal rank	Top 10 precision	Not-found
Corpus combing	42.4 %	65.5 %	13.1 %
Corpus combine+ result combine	50.5 %	69.0 %	15.2 %

3.3 Re-ranking

This method is based on our observation that although the average precision of the anchor-text-only based retrieval is not good, the precision of its top-*N* documents is much higher, especially when *N* is small enough. Therefore for each document in the content-based retrieval ranking list which is also contained in the top-*N* list of the anchor retrieval result, we set a higher score with proportion λ ($\lambda \geq 1$). Our results showed trivial while consistent improvement in a belief interval. Results are shown in figure 2.

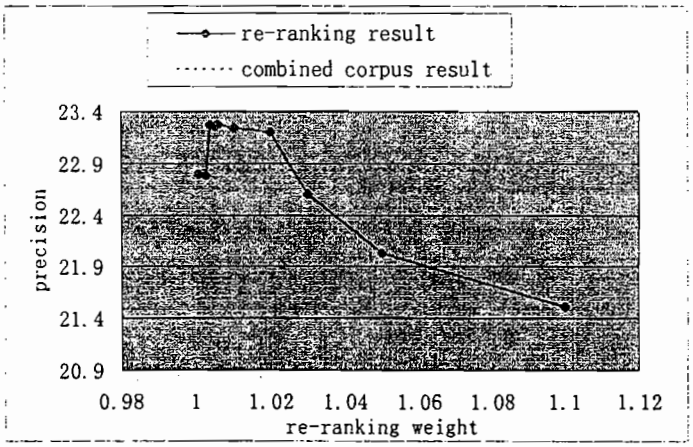

Figure 2 Effect of re-ranking on TREC9 web track ad hoc task

4 Conclusion

In this paper, we studied the effect of link-based approaches in web IR on TREC experiments. We firstly observed the performance of using link anchor text information, then investigate the effect of using link structure, finally studied three different combinations of link-based retrieval and tradition content-based retrieval. These link-based experiments carried out that: 1. Anchor text can represent precisely the topic of web page, but insufficient in describing the web page content; 2. Comparing with traditional content based IR technique, using anchor text on homepage finding task can get more than 96% improvement in terms of average reciprocal rank, while it is not helpful on ad hoc task; 3. combining link-based and content-based techniques can make consistent improvement on homepage finding task and a little progress by result re-ranking on ad hoc task.

In the future, we'll focus our study to enrich the anchor description by using context information of the anchor.

Reference

[1] S. Brin and L. Page, The Anatomy of a Large-Scale Hypertextual Web Search Engine, www7, 1998

[2] K. Bharat and M. R. Henzinger, Improved Algorithms for Topic Distillation in a Hyperlinked Environment, SIGIR 1998

[3] Jon M. Kleinberg, Authoritative Sources in a Hyperlinked Environment, Proceeding of the 9th annual ACM-SIAM symposium on Discrete Algorithms, pp 668-677, 1997

[4] N. Craswell, D. Hawking, S. E. Robertson, effective site finding using link anchor information, SIGIR 2001

[5] S. Chakrabarti, B. Dom, D. Gibson, H. Kleinberg, P. Raghavan, S. Rajagopalan, Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text, www7, 1998

[6] F. Crestani and P. L. Lee, Searching the web by constrained spreading activation, Information Processing & Management, 36(4), 2000, page 585-605

[7] J. Savoy and Y. Rasolofo, Report on the TREC-9 Experiment: Link-Based Retrieval and Distributed Collections, The Ninth Text Retrieval Conference (TREC-9), 2000

[8] Jianfeng Gao, etc, TREC-10 Web Track Experiments at MSRA, Proceedings of Text Retrieval conference, 2001.

Authors: ZHANG min (1977 -), female, Ph.D. candidate in the Department of Computer Science and Technology, Tsinghua University, research interests: Information Retrieval, Language Modeling; MA Shaoping (1961 -), male, professor and doctoral supervisor in the Department of Computer Science and Technology, Tsinghua University, research interests: Chinese Character Recognition, Information Retrieval; GAO Jianfeng, male, researcher in Microsoft Research Asia, research interests: Language Modeling, Information Retrieval, Statistical Machine Learning.

基于链接的方法进行 Web 信息检索的 TREC 实验研究*

张敏[1] 马少平[1] 高剑锋[2]

[1](清华大学计算机系, 智能技术与系统国家重点实验室, 北京 100084);
[2](微软亚洲研究院, 北京 100080)
E-mail: zhangmin99@mails.tsinghua.edu.cn

摘要: 本文通过 TREC 实验研究链接信息对 Web 信息检索的影响, 包括使用链接描述文本, 链接结构以及将基于链接的方法和传统基于内容检索的方法合并。我们得到结论如下: 首先, 链接描述文档对网页主题的概括有高度的精确性, 但是对网页内容的描述有极大的不完全性; 其次, 与传统检索方法相比, 使用链接文本在网页定位的任务上能够使系统性能提高 96%, 但是在信息查询任务上没有帮助; 将两种技术合并, 在网页定位任务上总能有很好的效果, 而对信息查询任务则只有重新调权的方法能够得到一定的改善。

关键词: 基于链接检索; 基于链接的方法; Web 信息检索; 信息查询; 网页定位

*本项目受到国家重点基础研究 (973) (G1998030509) 和自然科学基金项目 (No.69836040) 资助