Report on the TREC 2003 Experiments using Web Topic-Centric Link Analysis

P. Ingongngam and A. Rungsawang
Massive **I**nformation & **K**nowledge **E**ngineering
Department of Computer Engineering
Faculty of Engineering
Kasetsart University, Bangkok Thailand.
pain@csl.cpe.ku.ac.th, arnon@mikelab.net

Abstract

In TREC 2003, our experiments have been concentrated only on the topic distillation task. We first simply apply the term-based technique to the .GOV web collection, and then re-rank the retrieval results using a link analysis algorithm in order to boost the retrieval precision. Our link analysis has been inspired from the original PageRank, but focused on the web topic during the iterative score propagation. We hybridize the term-based retrieval scores with our link analysis approach. From the experiments, the results show that the combination of those scores still provides inadequate precision improvement.

1 Introduction

The information retrieval focuses on the quality of result the users obtain. However, the traditional information retrieval methods give the dissatisfying results for the web collections as the queries the users enter are highly ambiguous. In order to increase the precision of the retrieval results, many techniques have been developed. Link Analysis is one of the techniques extensively used in web-IR community. HITS [10] and PageRank [5] are the two well known algorithms that are developed based on link analysis technique, and widely studied by several TREC participants [9, 11, 6, 7, 12].

Hyperlinked-Induced Topic Search (HITS) has been developed by Kleinberg and implemented in the CLEVER Project [2], while PageRank [5] is the core mechanism of the most successful search engine, Google [1]. Both of these algorithms have different advantages and disadvantages. HITS algorithm calculates a page score based on the the user's topic. However, HITS must be computed on-line at query time. On the other hand, PageRank algorithm calculates a page score based on link relations. This calculation is done off-line only once, but the calculated scores are not related either to the topics of the web pages or to that the users are interested.

For the TREC 2003, we introduce another off-line link analysis that allows the web topic to influence the propagation of link scores, called "Topic-Centric" (TC) [8]. Analyzing the connectivity of the web graph in the same way as PageRank, TC algorithm iteratively propagates the portion of rank score of a source web page to the rank score of the destination one in accordance with the topics of both web pages. Following the hyperlink, the destination web page will then appropriately receive a high rank score when the topic of the source web page is really referred to that of the destination one.

The report is structured as follows. Section 2 reviews briefly the PageRank algorithm. Section 3 explains the new TC algorithm. Section 4 shortly provides the experimental setup and results. Finally, section 5 concludes the report.

2 Basic PageRank Algorithm

Brin and Page suggest a link-based search model called "PageRank" that evaluates the importance of each web page based on its citation pattern. The basic idea of PageRank is as follows. When a page u has a hyperlink to a page v, it is assumed that the author of the page u suggests the page v with some reasons, e.g., related context, individual favor, or popular reference. PageRank employs this hint to compute the page scores. Since it considers all web pages equally important, if we let N_u be the number of pages which page u points out, and $Rank(u)$ represent the rank score of a page u, then a hyperlink $u \rightarrow v$ confers $1/N_u$ units of rank to page v.

To compute the rank vector for all web pages, we then simply iteratively perform the following fixed-point computation. If we let B_v represent the set of pages pointing to page v, for each iteration, the successive rank scores of pages are recursively propagated from the previously computed rank scores of all other pages pointing to them:

$$\forall_v Rank_{i+1}(v) = \sum_{u \epsilon B_v} \frac{Rank_i(u)}{N_u} \tag{1}$$

In general, the web graph is not strongly connected, and this may lead the PageRank computation of some pages to be trapped in a small isolated cluster of the web graph. This problem is usually resolved by pruning nodes with zero out-degree, and by adding random jumps to the random surfer process [5]. This leads to the following modification of Equation (1) to:

$$\forall_v Rank_{i+1}(v) = (1 - \alpha) + \alpha \sum_{u \epsilon B_v} \frac{Rank_i(u)}{N_u} \tag{2}$$

where α, called "damping factor", is the value that we use to modify the transitional probability of the random surfer model of an underlying web graph.

3 Web Topic-Centric Algorithm

There is a big difference between PageRank and TC algorithms. PageRank treats every web page equally important. It iteratively propagates the link score from the source page to the destination one with the fraction of $\frac{1}{N}$ where N is the amount of outbound links of the former. Link score propagation with no regard to the web content, or the "topic" of that web page, may be inappropriate and independent to the user query. Besides, our TC algorithm propagates link scores by considering the similarity between the topics of the source and destination pages.

There are many ways to compute the page similarity, but we here only focus on the vector space approach [4]. The vector space is widely used in many information retrieval researches. The basic idea of the vector space is to imagine a web page as a vector. Each distinct word in that page is considered to be an axis of a vector in a space. The direction of a vector characterizes the content of a web page. Here, we define W_{uv} to be the page similarity score computed between the source page u and the destination page v, and calculate it using the following formula:

$$W_{uv} = \frac{\sum_{k\epsilon u\cap v} f_{ku} f_{kv}}{\sqrt{\sum_{k\epsilon u} {f_{ku}}^2 \sum_{k\epsilon v} {f_{kv}}^2}} \tag{3}$$

where f_{ku} and f_{kv} are the number of term k found in page u and page v, respectively. The rank of W_{uv} value is between 0 and 1, and the similarity increases as this value increases.

Since TC does not consider web pages as being equally important, the portion of a rank score that propagates from a page u to a page v should be dependent on their page contents or topics. We then appropriately modify the fraction of link score propagation in Equation (2) to:

$$\forall_v Rank_{i+1}(v) = (1-\alpha) + \alpha \sum_{u\epsilon B_v} \left(\frac{W_{uv}}{\sum_{x\epsilon O_u} W_{ux}} Rank_i(u) \right) \tag{4}$$

Here, B_v represents a set of pages pointing to v, O_u represents a set of pages pointed by u, and W_{uv} represents the similarity score computed between the content of the page v and the content of the in-link page u.

4 Experimental Setup and Results

We use the LEMUR toolkit [3] as our vector space based retrieval system. We first process the .GOV documents using BM25 weighting scheme, with parameters $B = 0.9$, $K_1 = 2$ and $K_3 = 7$, respectively. We only use the title section of the TREC topic without any expansion, and examine the average retrieval precision at 5, 10, 15, 100 and 1000 retrieved documents. We hereafter call the result from this step, the "base" case. We then apply both TC and PageRank algorithms to compute the rank scores of web documents in the .GOV collection, and employ those scores to re-rank the search results obtained from the base case. Table 1 as follows provides the comparison between the average precision scores of the base case, the re-ranking results obtained from the TC (denoted by "TC"), and those obtained from the PageRank (denoted by "PR"), respectively.

Table 1: The average precision scores.

retreived docs	base	TC	PR
At 5 docs	0.0920	0.0880	0.0160
At 10 docs	0.0800	0.0760	0.0280
At 15 docs	0.0667	0.0680	0.0280
At 100 docs	0.0304	0.0294	0.0272
At 500 docs	0.0125	0.0122	0.0126
At 1000 docs	0.0069	0.0069	0.0069
Avg Precision	0.0855	0.0789	0.0164

5 Conclusion

Like PageRank, TC algorithm analyzes the link connectivity, and pre-computes the rank scores of web pages. During the computation, TC propagates the portion of rank scores of the source web pages to the destination web pages in accordance with the topics found in both ends. Therefore, we expect that the final computed rank scores will be more reasonable, and be more efficient to use to re-rank the search results obtained from the traditional vector space model.

The study concluded from the TREC experiments this year shows that TC algorithm does still not provide any significant improvement when it is used to re-rank the search results obtained from the standard vector space retrieval model. Comparing with PageRank, TC algorithm however gives better re-ranking results in our experiments. More study and experiments will be conducted, e.g., we will try several other vector space based weighting scheme in similarity computation, as well as the use of weighted inter-host and intra-host link score propagation.

References

[1] Google, http://www.google.com/.

[2] IBM Almaden Researcn Center. Clever Searching. http://www.almadenibm.com/cs/k53/clever.html.

[3] LEMUR, http://www-2.cs.cmu.edu/∼lemur/.

[4] Ricardo A. Baeza-Yates and Berthier A. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, Addison-Wesley, 1999.

[5] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[6] F. Crivellari and M. Melucci. Web document retireval using passage retrieval, connectivity information, and automatic link weighting – trec9 report. In *Proceeding of the TREC-9 Conference*, 2000.

[7] J. Gevrey and S.M. Rüger. Link-based approaches for text retrieval. In *Proceeding of the TREC-10 Conference*, 2001.

[8] P. Ingongngam and A. Rungsawang. Topic-centric algorithm: A novel approach to web link analysis. In *Proceeding of the* 18^{th} *AINA Conference*, 2004. (to be appeared).

[9] T. Kanungo and J.Y. Zien. Integrating link structure and content information for ranking web documents. In *Proceeding of the TREC-10 Conference*, 2001.

[10] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[11] J. Savoy and Y. Rasolofo. Report on the trec-9 experiment: Link-based retrieval and distributed collections. In *Proceedings of the TREC-9 Conference*, 2000.

[12] K. Yang. Combining text- and link-based retrieval methods for web ir. In *Proceeding of the TREC-10 Conference*, 2001.