LIT at TREC-2002: Web Track

Nie Yu, Ji Donghong, Yang Lingpeng
Laboratories for Information Technology, Singapore
{ynie, dhji, lpyang}@lit.a-star.edu.sg
http://www.lit.org.sg

Abstract

In Trec-2002, we participated in the Web Trec (named page finding task). There are two kinds of information that can be used while finding the expected page, content information and link information. We exploited both of them. That is to say, our system is content-based and link-based. As to link information, we only used anchor text and connections, and topology between pages is ignored. We submitted two runs. One is based on traditional contented-based retrieval, the other try to combine content-based retrieval and link-based retrieval to get better result.

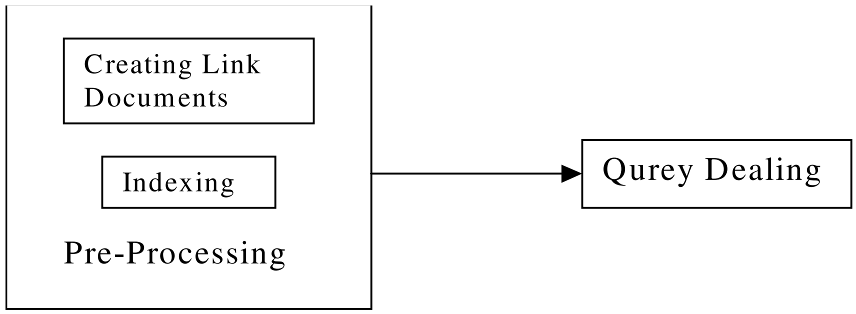
1 Introduction

This is the first time we participate in Web Trec. We focused our work on named page finding task. To exploit link information more explicitly, we extracted link information of each page that is used to create the link document (Craswell, et al., 2001; Dumais, 2001; Savoy, 2001). So each page has its respective link document.

For each query, we extracts source terms from the query, creates vectors for each web page and its link document according to source terms, and then calculates evaluating value of each web page (Robertson et al., 2001). Finally we get the ranked page list via sorting evaluating values of web pages.

2 System Description

We did the pre-processing work before dealing with queries. The process is as the diagram below.



We did the pre-processing work with three desktop computers (750MHz and 128M memory, Windows 98). Queries dealing runs on one desktop computer (750MHz, 512M memory, Windows 98).

3 Pre-processing

Before dealing with queries, we created link document for each web page, and indexed all web pages. The glossary and stop word list are from Wordnet.

After pre-processing, we got two tables and indexed result. The title&anchor table contains titles and anchor descriptions of web pages. The link table denotes which pages are pointing to a given page.

3.1 Creating Link Documents

For each web page d, a link document ld is created. The link document ld contains all anchor text of hyperlinks which pointing to web page d in other pages. The sum of those hyperlinks is recorded as sum_d. We defined the title of a web page as the anchor text of one additional hyperlink, because we thought title should be similarly accurate with pointing-in hyperlinks on indicating a page's meaning.

3.2 Indexing

For each web page, all words were stemmed and stop words were removed before indexing.

4 Named Page Finding

First, we extract source terms from the query, and then we built context vectors and link document vectors for pages based on source terms. By comparing the evaluating value of all pages, we get the final result sequence.

4.1 Extracting Source Terms

All stop list words are ignored before source term extracting. For one query q such as w_1 w_2 $w_{3...}$ w_n, w_i is a single word, $1 \le i \le n$, there are n(n-1)/2 possible source term t.

$$t \; \varepsilon T$$

$$T = \{ w_j \, w_{j+1...} \, w_{j+l} \mid 1 \le j \le n, 1 \le l \le n-j \} \cup \{ w_j \mid 1 \le j \le n \}$$

Each source term t is given a weight value wt_t:

$$wt_t = F(l_t, l_q)$$

where l_t is the length of source term t, l_q is the length of query q. So for source terms T, a weight vector wt is created.

$$wt = (wt_{t1}, wt_{t2}, \ldots, wt_{tm}), \; m = n(n-1)/2$$

4.2 Creating Context Vector and Link Document Vector

We create a context vector v and a link document vector vl for each web page d regarding source terms T.

$$v = (f_1, f_2, \ldots, f_{n(n-1)/2})$$

$$vl = (fl_1, fl_2, \ldots, fl_{n(n-1)/2})$$

where f_i ($1 \le i \le n(n-1)/2$) is the frequency of t_i ($1 \le i \le n(n-1)/2$) in web page d, fl_i ($1 \le i \le n(n-1)/2$) is the frequency of t_i ($1 \le i \le n(n-1)/2$) in link document ld. Frequencies will not be repeatedly counted for different terms. Longer terms have priority over shorter terms.

4.3 Evaluating Value

For each web page d, an evaluating value e is calculated following the equation below:

$$e = f(v/f1(l_d), wt) + \mu f(vl/f2(sum_d), wt)$$

where f,f1,f2 are functions, μ is a parameter to weight effect of link information, which is valued after a lot of tests.

In fact, we delivered two runs, one is content-based, and the other is content & link-based. In the content-based run, we used equation below:

$$e = f(v/f1(l_d), wt)$$

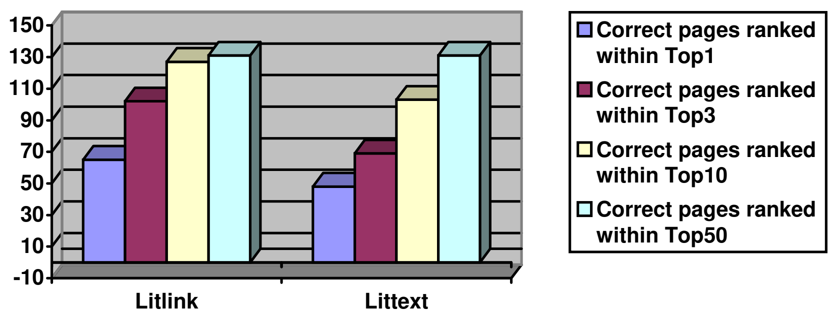
4.4 Get Ranked Page List

After comparing and sorting evaluating value e of each web page, we can get the ranked page list.

5 Result

We submitted 2 runs for Web Trec task. The results are as Chart 1 below:

Chart 1



In the chart, the value of "top n" column means how many named pages were found in the top n ranked results within 150 given queries. The run 'litlink ' used both content information and link information, run 'littext' only used content information. The accuracy of two runs is compared in the table below. We can easily find out that with link information, much better results are obtained.

Run Id	Top 1	Top 3	Top 10	Top 50
Litlink	43.3%	68%	84.7%	87.3%
Littext	32%	46%	68.7%	87.3%

6 Conclusion and Future Work

Unlike the set of plain text documents, web pages are document set with topology structure and relationships. The titles and anchor descriptions actually provide the summary indicating kernel content of pages. They can remarkably improve the efficiency and performance of retrieval, should and must be considered in retrieval process.

We will continue focusing on retrieval based on link information. Different process should be taken to consider different kind of link information e.g., title, anchor.

In addition, we will try to combine content-based retrieval and link-based retrieval better to achieve higher performance.

References

Craswell, N., Hawking, D., and Robertson, S.E.(2001). Effective site finding using link anchor information. In SIGIR-01.

Dumais, S., and Jin, R.(2001). Probalistic combination of content and links. In SIGIR-01.

Savoy, J., and Rasolofo, Y. (2000). Report on the TREC-9 Experiment: Link-Based Retrieval and Distributed Collections. In TREC-9.

Robertson, S.E, Spark-Jones K., Relevance weighting of search terms. In Journal of ASIS, 27, pp.129-146, 1976