Using Document Structure on Retrieving Webpages at the Web-CLEF 2006

Syntia Wijaya, Bimo Widhi, Tommy Khoerniawan, and Mirna Adriani

Faculty of Computer Science
University of Indonesia
Depok 16424, Indonesia
{swd20, bimo20, tokh20}@mhs.cs.ui.ac.id, mirna@cs.ui.ac.id

Abstract. We present a report on our participation in the mixed monolingual web task of the 2006 Cross-Language Evaluation Forum (CLEF). We compared the result of web page retrieval based on the page content, page title, and anchor page. The retrieval effectiveness for the combination of page content, page title, and anchor texts was better than that of the combination of page title and page title only. Applying the pseudo-relevance feedback improved the retrieval performance of the queries.

Keywords : web retrieval

1 Introduction

The fast growing amount of information on the web motivated many researchers to come up with a way to deal with such information efficiently [2, 5]. Information retrieval forums such as the Cross Language Evaluation Forum (CLEF) have included research in the web area. In fact, since 2005, CLEF includes a WEBIR topic as one of the research tracks. This year we, the University of Indonesia IR-Group, participated in the mixed monolingual WEBIR - CLEF 2006 task.

2 The Retrieval Process

The mixed monolingual task searches for web pages in a number of languages. The queries and the documents were processed using the *Lemur*[1] information retrieval system. Stop-word removal, as is done by many IR systems, was applied only to the English queries and documents.

2.1 Web-page Scoring Techniques

We employed five different techniques for scoring the relevance of documents (web pages) in the collection, i.e., based on the combination of the content of the page, the title of the page, and the anchor texts that appear on the pages.

The first technique takes into account only the content of a web page to find the most relevant web pages to the query. We used the *language model* [3] to find the probability value between the query and the pages. The second technique considers the title of the web page as the only source in finding the relevant pages. The third technique uses the content and the title of the page to find the relevant pages.

[1] See “http://www.lemurproject.org”.

3 Experiment

The web collection contains over two million documents from the EUROGOV collection. In these experiments, we used *Lemur* information retrieval system to index and retrieve the documents. *Lemur* is built based on the *language model* [3]. We index the webpages according to their content pages, title pages, and anchors. Stopwords were removed from the collection, but word stemming was not applied to the collection.

4 Results

We were very surprised to see the results of our participation this year. All of the results that we submitted are very low compared to our last year's result. In 2005, we indexed the collection using a different information retrieval system, i.e., *Lucene*[2] which is built based on the *vector similarity* model [1, 4]. The first result is shown in Table 1. In the retrieval, we compute the total relevance score by summing up the relevance scores based on page content, page title, and anchor texts found on the webpages.

Table 1. Mean Reciprocal Mean (MRR) of the combined relevance score for page content, page title, and anchor texts on a webpage.

Task : Mixed Monolingual	UI1DTA
MRR	0.0404
Average success at 1:	0.0258
Average success at 5:	0.0531
Average success at 10:	0.0707

Table 2 shows the result of combining the relevance scores based on page content and page title. As can be seen, the MRR dropped from 0.0404 (see Table 1) to 0.0116.

Table 2. Mean Reciprocal Mean (MRR) of the combined relevance score for page content and page title.

Task : Mixed Monolingual	UI4DTW
MRR	0.0116
Average success at 1:	0.0067
Average success at 5:	0.0150
Average success at 10:	0.0201

The third technique applies the pseudo-relevance feedback to the retrieval that uses the combined score of page content, page title, and anchor texts. As shown in Table 3, the feedback reduced the performance of the queries where the MRR dropped to 0.0253. The pseudo-relevance feedback was done using the top-5 relevant documents retrieved.

Table 3. Mean Reciprocal Mean (MRR) of the combined score of page content, page title, and anchor texts with top-5 documents pseudo-relevance feedback.

Task : Mixed Monolingual	UI3DTAF
MRR	0.0253
Average success at 1:	0.0160
Average success at 5:	0.0309
Average success at 10:	0.0423

[2] See "http://lucene.apache.org/".

Finally, the last result was obtained by applying the pseudo-relevance feedback to the combined relevance score of page content and page title only. As shown in Table 4, we obtained the highest retrieval performance with MRR of 0.0918.

Table 4. Mean Reciprocal Mean (MRR) of the combined score of page content and page with top-5 documents pseudo-relevance feedback.

Task : Mixed Monolingual	UI1DTF
MRR	0.0918
Average success at 1:	0.0634
Average success at 5:	0.1202
Average success at 10:	0.1516

To investigate the cause of our poor retrieval performance, we conducted some further experiments. We used the queries from last year's task and ran them on the same index that was built using *Lemur*. The result is as shown in Table 5, which is much better than for this year's queries. However, we found a sign of indexing error, i.e., there were some domains that *Lemur* was unable to index. This resulted in Lemur's not being able to retrieve any documents for a number of queries. We also suspected that the index for documents in languages containing non-latin characters was corrupt, as indicated by the fact that documents in some domains such as Russian and Greek were never retrieved.

Table 5. Mean Reciprocal Mean (MRR) of the combined relevance score of page content, page title, and anchor texts using the 2005 query-topics.

Task : Mixed Monolingual	DTA-2005
MRR	0.2069
Average success at 1:	0.1444
Average success at 5:	0.2742
Average success at 10:	0.3254

5 Summary

Our results demonstrate that combining the page content, the page title, and anchor texts resulted in a better mean reciprocal rank (MRR) compared to searching using the page content and page title only. The pseudo-relevance feedback that we employed increased the retrieval performance of the queries. However, we had some problems with indexing the collection, which resulted in our poor retrieval performance in our participation this year. We hope to improve our results in the future by exploring still other methods.

References

1. Baeza-Yates, Richardo, and Berthier Ribeiro-Neto. *Modern Information Retrieval*, New York: Addison-Wesley, 1999.
2. Hawking, David. Overview of the TREC-9 Web Track. In *NIST Special Publication: The 10th Text Retrieval Conference (TREC-10)*. 2001
3. Ponte, J. and Croft, W.B. A Language Modeling Approach to Information Retrieval. In Proceedings of the 21st ACM SIGIR Conference on Research and development in Information Retrieval, p.275-281. ACM: 1998.
4. Salton, Gerard, and McGill, Michael J. *Introduction to Modern Information Retrieval*, New York: McGraw-Hill, 1983.
5. Zobel, J. How reliable are the results of large-scale information retrieval experiments? In Proceedings of ACM SIGIR'98. Melbourne, Australia: August 1998.