University of Indonesia Participation at WEBIR-CLEF 2005

Mirna Adriani and Rama Pandugita
Faculty of Computer Science
University of Indonesia
Depok 16424, Indonesia
(mirna@cs.ui.ac.id, ramap101@mhs.cs.ui.ac.id)

Abstract. We present a report on our participation in the mixed monolingual web task of the 2005 Cross-Language Evaluation Forum (CLEF). We compared the result of web page retrieval based on the content of the page, the target domain and the page content, and a combination of the page title and the target domain. The result shows that combining the page title and the target domain resulted in better retrieval performance than using only the page content or the target domain and the page content.

1 Introduction

The fast growing amount of information on the web motivated many researchers to study how to deal with such information efficiently. Information retrieval forums such as the Cross Language Evaluation Forum (CLEF) have included research in the web area. In fact, this year 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 2005 task.

2 The Query Process

The mixed monolingual task searches for web pages in a number of languages. The queries and the documents were processed using the *Lucene* information retrieval system (see http://lucene.apache.org). Stopword removal was applied only to the English queries and documents. We used three different techniques for indexing the documents in the collection, i.e., based only on the content of the page, based on the target domain and the page content, and based on the combination of the page title and the target domain.

The first technique considers the content of the page in order to find the most relevant web pages to the query. We used the *vector space model* [1, 2] to find the similarity value between the query and the pages.

Since the collection contains documents not only in English, we also used the information in the metadata of the target document. The information about the target domain of the query is matched with the domain metadata of the pages. We then applied the vector space model to obtain the most similar pages to the query based on the page content. For example:

<topic_id>	<domain>	<LANGUAGE>	<query>
WC0001	eu	EN	road safety in europe
WC0003	nl	NL	list of dutch eco-labels

Query number 1 (WC0001) was used for searching only in the *eu* (European Union) domain and query number 3 (WC0003) was used for searching only in the nl (Netherland) domain.

The third technique uses a combination of the page title and the target domain. We used only the terms in the page title as the query terms once the domain of the target pages is found. We then match the title of the pages with the query using the vector space model.

3 Experiment

The web collection contains over two million documents from the EUROGOV collection. The collection is divided into 27 European language domains. In this mixed monolingual task, the queries are in various languages and used to find documents in the same language as the queries. There are 547 queries to be used for searching in two categories, namely, the name page search and the homepage search. The average number of words in the queries is 6.29 words.

In these experiments, we used the *Lucene* information retrieval system to index and retrieve the documents. *Lucene* retrieval system is based on the vector space model. The documents were indexed in separate indexing files according to the domain. For example, documents from internet domain 'uk' are indexed separately from documents from internet domain 'de'. Lucene has the capability to build separate indexes and to search according to the specified index.

Lucene is also capable of indexing documents using two separate fields such as the title page and the content, and then searching can be done using either the title page or the content page.

4 Results

The results that we have submitted were produced using the three techniques, namely, based on the content only, based on the target domain and page content, and based on a combination of the title page and the target domain.

Table 1 shows the result of our experiments. The mean reciprocal rank (MRR) over 547 queries is 0.2165 for using the content, 0.2714 for using the target domain and the content, and 0.2860 for using the combination of the page title and the target domain. The combination of the page title and the target domain resulted in retrieval performance 24.82% better than that of using only the content, and 5.10% better than that of using the target domain and page content.

Task : Mixed Monolingual	**Mean Reciprocal Rank (MRR)**
Page content	0.2165
Target domain + Page content	0.2714
Page title + Target domain	0.2860

Table 1. Mean reciprocal Mean (MRR) of the mixed monolingual queries.

The average of success values of each technique in several ranks (Table 2) show some differences. The best result was achieved by using the combination of the target domain and the page title, with average success at 1: 0.2249, which is consistent with the MRR result. The retrieval performance of the combination of the target domain and the page title was 33.34% better than that of using the content only and 15.47% better than that of using the target domain and the content. However, the average of success of using the target domain and the page content shows better results at higher ranks than the other two techniques.

Our result is similar to the work by Westerveld et al. [2] who obtained better results by using other information in addition to the content.

Task : Mixed Monolingual	Content	Target domain	Target domain + title page
Average success at 1:	0.1499	0.1901	0.2249
Average success at 5:	0.2834	0.3638	0.3583
Average success at 10:	0.3583	0.4223	0.4186
Average success at 20:	0.3931	0.4936	0.4662
Average success at 50:	0.4826	0.5978	0.5320

Table 2. Average of success of the mixed monolingual runs using content only, target domain and content, and a combination of the target domain and the page title.

Since this was our first participation in the WEB task, it took us quite a lot of effort to cope with such large collections. There were several document sets that were damaged, possibly in the process of downloading the files. As a result, we could not index those corrupt files. It is possible that those files were relevant to some of the queries.

The other problem that we had was that we did not prepare Lucene to handle non-Latin characters, and so, the retrieval of documents using queries containing such characters was erroneous.

4 Summary

Our results demonstrate that combining the target domain metadata and the page title resulted in better mean reciprocal rank (MRR) compared to searching using the content only and using the target domain metadata and the page content. However the combination of the target domain metadata and the page title achieved best performance only at 1 rank. For the other ranks, using the target domain metadata and page content showed better results compared to the other two techniques. We hope to improve our results in the future by exploring still other methods.

5 References

1. Baeza-Yates, Richardo, and Berthier Ribeiro-Neto. *Modern Information Retrieval*, New York: Addison-Wesley, 1999.
2. Salton, Gerard, and McGill, Michael J. *Introduction to Modern Information Retrieval*, New York: McGraw-Hill, 1983.
3. Westerveld, Thisj, Wessel Kraaij and Djoerd Hiemstra. Retrieving Web Pages using Content, Links, URLs, and Anchors. In *NIST Special Publication: The 10th Text Retrieval Conference (TREC-10)*. 2001.