Multilingual Web Retrieval Experiments with Field Specific Indexing Strategies for CLEF 2006 at the University of Hildesheim

Ben Heuwing, Thomas Mandl, Robert Strötgen

Information Science, University of Hildesheim,
Marienburger Platz 22
D-31141 Hildesheim, Germany

mandl@uni-hildesheim.de

Abstract

For WebCLEF 2006 we experimented with the analysis and extraction of the HTML structure of the web documents. In addition, blind relevance feedback was applied in the search process. As in 2005, the experiments were carried out with a language independent indexing strategy. We experimented with HTML title, H1 element and other elements emphasizing text. Our index contained title and H1, emphasized elements, full and partial content. Blind relevance feedback was implemented for all index fields except for the full content. The best results with the WebCLEF 2005 topics were achieved with a strong weight on the title-element accomplishing a marginal improvement over the best post submission runs for the mixed-monolingual task at WebCLEF 2005. For the WebCLEF 2006 topics, improved results were achieved with the manually generated topics, while those automatically generated led to results far below average. The best performance for manual topics for CLEF 2006 was achieved with a strong weight on both HTML title as well as H1 elements, and a decreased weight for the other elements. Blind relevance feedback could not yet improve the results.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software

General Terms

Measurement, Performance, Experimentation

Keywords

Web Retrieval, Multilingual Information Retrieval, Evaluation

1 Introduction

Our participation was based on the experience gained during WebCLEF 2005 [Jensen et al. 2005]. The 80 GB multilingual EuroGOV corpus [Sigurbjörnsson et al. 2005] caused several problems during parsing and pre-processing. Nevertheless, this year all files were being processed and integrated into the index. For the multilingual task, our approach had shown competitive results in WebCLEF 2005. This year our efforts were centered on the mixed-monolingual task as the multilingual task was not offered.

2 System Description

The system developed for WebCLEF 2005 was built on Apache Lucene. We decided to take a language independent approach. All files were indexed in one multilingual index. As a consequence, no language dependent stemming algorithms could be applied. Only the character *S* at the end of words was removed by the Lucene StandardAnalyzer.

Indexing strategies were refined this year based on the structure of the HTML files. At WebCLEF 2005, retrieval based on the HTML title element proved to be extremely effective for multilingual web retrieval. We assumed that the titles might be partially of low quality and that they could be eliminated in many cases. This hypothesis was based on general observation that many web pages have the title "no title" or similar phrases in other languages. In order to identify these phrases and assemble a stop title list which would not be indexed, the frequency of title phrases was assessed. Surprisingly, the observation did not hold for the EuroGOV corpus. Some titles do occur often, but they contain valuable text and should not be eliminated. Consequently, the stopword list was merely extended with the most frequent title words.

The first order headline (H1) element was identified and added to the title in order to be indexed conjointly. A set of other elements which emphasize text (H{1-6}, strong, b, em, bold, i) was also identified and joined to form one indexing field. As in WebCLEF 2005, we indexed both the full content and partial content. Instead of choosing the first characters for the content cutoff, we adopted a more refined strategy. The most discriminating content for a webpage is often not at the beginning of the HTML code. In many cases, the beginning of the code contains navigation elements and menus which are stable for a whole site [Chen et al. 2006]. Consequently, we selected 50 tokens from the middle of the HTML code to be indexed as the partial content. A more elaborated strategy was not adopted in order not to compromise the efficiency of our indexing approach.

Blind relevance feedback is a very efficient strategy for retrieval optimization. For multilingual ad-hoc retrieval with newspaper corpora, it has been successfully applied by the University of Hildesheim [Hackl et al. 2005] and many others. We adopted the blind relevance implementation for ad-hoc retrieval and integrated it into the web retrieval scenario. Blind relevance feedback was realized for all indexed fields but the full content. Due to hardware limitations, Lucene could not store all term vectors for the full content which were used for the blind relevance feedback implementation.

In addition, a domain filter was implemented which takes advantage of the meta data provided with the topics.

3 Submitted Retrieval Experiments with EuroGOV

The parameters were optimized based on the WebCLEF 2005 topics. Results of the runs are shown in table 1.

Table 1. Results of experiments (WebCLEF 2005 topics)

	UHiBase	UHiTitle	UHi1-5-10	UHiBrf1	UHiBrf2	UHiMu
Mean reciprocal rank	0.2819	0.2807	0.2814	0.2731	0.2771	0.2443
Average success at 10	0.4168	0.4132	0.4186	0.3949	0.4040	0.3656

Compared to the best post submission experiments of WebCLEF 2005, it can be seen that the result of the multilingual experiments could be improved from 0.2117 to 0.2443. The top performing result submitted by the University of Hildesheim had been 0.1370.

For the mixed monolingual task, the performance could be improved (from an MRR of 0.2377 to 0.2819) without reaching the performance levels of the best participants in 2005. The average success rate at position ten was improved from 0.235 to 0.4168. Our new indexing strategies did improve the results overall, however, they did not lead to competitive results for the mixed monolingual task. Based on the results of the prior experiments, we decided to submit runs with the parameters shown in table 2.

Table 2. Parameters for the submitted Runs

Name of Run	Weights
UHiBase	content^1 emphasised^0.1 title^20
UHiTitle	content^1 emphasised^1 title^20
UHi1-5-10	content^1 emphasised^5 title^10
UHiBrf1	content^1 emphasised^1 title^20 blind relevance feedback (weight of expanded query: 1)
UHiBrf2	content^1 emphasised^1 title^20 blind relevance feedback (weight of expanded query: 0.5)
UHiMu	(multilingual) content^1 emphasised^1 title^20 - translation^10

As a base-run, the run which showed the best results in the experiments with the WebCLEF 2005 topics was chosen. The queries were weighted, the weight of the emphasised-field actually being decreased as preliminary

results had shown that this leads to slightly better results. This finding reduces the probability that the additional elements taken into account have a significant discriminating effect. Additionally a similar run with a heavily weighted title-field (*UHiTitle*) and one with moderate weights (emphasised^5 and title^10) was submitted. In all runs the full-content field was used for search, as it leads to significantly better results than the partial-content field, which was used only to generate term-vectors for blind relevance feedback. To test the effect of blind relevance feedback, two runs with different weights on the expanded query were generated. The performance of these runs was slightly below that of the - in all other aspect equivalent - *UHiTitle*-Run. According to these findings, blind relevance feedback has not shown to have a positive effect on retrieval quality so far, even though there is still room to experiment with different methods and parameters. To test the improvements applied to the system in the multilingual context a non-official run for the multilingual task was submitted.

The improvements achieved are partly a result of the use of meta-data, restricting search to the target-domain. This of course improved the position of relevant documents in the result list. Not using the filter, the *UHiTitle*-Run has a lower MRR of 0.2552, but the 'average success at ten'-rate of 0.3784 still shows a definite improvement, probably due to the effect of an exhaustively indexed corpus (all documents as full-text) as well as the optimzed weighting of the different fields.

Considering the results of the submitted runs (shown in table 3), the difference between the results of the different topic-types is striking. While on manually generated topics (319 of 1939) the runs performed as it was to be expected from the experimental results, the performance on automatically generated topics (1620 of 1939) was poor. With the manual topics the *UHi1-5-10-Run* produced the best results with an MRR score of 0.3134 and an 'average success at 10'-rate of 0.4577. Over all topics the same run had an MRR of only 0.0718 and an 'average success at 10' of 0.1233. In other words, a relevant document was five times more likely to appear within the first ten hits as a result of a manual topic than as a result of an automatically generated one. Differences were also found between the results of the new manually created topics (124) and those of the old topics (195) of 2005. The *UHiBase*-Run resulted in an MRR of 0.2556 for the old topics while the new topics let to significantly better results (MRR 0.3893).

Table 3. Results WebCLEF 2006

	all topics		*manually generated topics*	
	MRR	Average success at 10	MRR	Average success at 10
UHiBase	0.0795	0.1377	0.3076	0.4451
UHiTitle	0.0724	0.1253	0.3061	0.4420
UHi1-5-10	0.0718	0.1233	0.3134	0.4577
UHiBrf1	0.0677	0.1104	0.3000	0.4295
UHiBrf2	0.0676	0.1124	0.2989	0.4295
UHiMu	0.0489	0.0758	0.2553	0.3824

Taking into account only the manually created topics, the results of WebCLEF 2006 show the improvements that were to be expected from the previous experiments. The results even were slightly better than those of the experiments with the WebCLEF 2005 topics. The best performance was achieved with a strong weight on HTML title and H1 elements, a moderate weight for the other elements extracted and without blind relevance feedback. Consequently, it can be said for sure whether the elements extracted additionally have a higher discriminating effect than the content of the document.

4 Conclusion and Outlook

For the second web track participation at CLEF we intended to tune our system and to index several fields. Blind relevance feedback was successfully integrated. The possibility to give different weights to the fields offered room for experiments and the discriminating effect of HTML elements was confirmed. Improving the preprocessing routines over the WebCLEF participation in 2005 also had a positive effect on the retrieval quality. The use of blind relevance feedback in this context will have to be explored further. In future experiments, we intend to test different weighting strategies for blind relevance feedback which are independent of the weights of the initial query.

In further future experiments, we intend to include advanced quality measures. Advanced quality measures which regard layout information will be applied [Mandl 2006]. To accomplish this, the preprocessing methods will have to be worked on further.

References

Chen, L.; Ye, S.; Li, X. (2006): Template Detection for Large Scale Search Engines. In: Proceedings ACM Symposium on Applied Computing ACM Press. pp. 1094-1098.

Hackl, René; Mandl, Thomas; Womser-Hacker, Christa (2005): Mono- and Cross-Lingual Retrieval Experiments at the University of Hildesheim. In: Peters, Carol; Clough, Paul; Gonzalo, Julio; Kluck, Michael; Jones, Gareth; Magnini, Bernard (eds): Multilingual Information Access for Text, Speech and Images: Results of the Fifth CLEF Evaluation Campaign. Berlin et al.: Springer [LNCS 3491] pp. 165-169.

Jensen, Niels; Hackl, René; Mandl, Thomas; Strötgen, Robert (2006): Web Retrieval Experiments with the EuroGOV Corpus at the University of Hildesheim. In: Peters, Carol; Gey, Fredric C.; Gonzalo, Julio; Jones, Gareth J.F.; Kluck, Michael; Magnini, Bernardo; Müller, Henning; de Rijke, Maarten (Eds.). Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Vienna, Austria, Revised Selected Papers. Berlin et al.: Springer [LNCS 4022] pp. 837-845.

Kamps, Jaap; de Rijke, Maarten (2006): Overview of WebCLEF 2006. In this volume.

Mandl, Thomas (2006): Implementation and Evaluation of a Quality Based Search Engine. In: Proceedings of the 17th ACM Conference on Hypertext and Hypermedia (HT '06) Odense, Denmark, August 22nd–25th. ACM Press.

Sigurbjörnsson, Börkur; Kamps, Jaap; de Rijke, Maarten (2005): Blueprint of a Cross-Lingual Web Retrieval Collection. In: Journal of Digital Information Management, vol. 3 (1) pp. 9-13.