

# Different Indexing Strategies for Multilingual Web Retrieval: Experiments with the EuroGOV Corpus



## Abstract

Experiments with a multi-lingual web collection are presented. The EuroGOV corpus is the first multi-lingual web corpus for retrieval evaluation. We show how indexes based on words and n-grams are developed for different document parts. Different indexes were based on the full document content, partial content and the title. The best results were achieved for a title only index based on words.

## Cross Language Evaluation Forum

The evaluation of information retrieval systems has gained considerable momentum in the last few years. Several evaluation initiatives are concerned with diverse applications, usage scenarios and different aspects of system performance.

The Cross Language Evaluation Forum (CLEF) is an evaluation initiative following the model of TREC. CLEF is focused on the evaluation of multilingual information retrieval systems. CLEF provides an infrastructure for research and development including test suites and multi-lingual document collections in more than ten languages. The initiatives have greatly diversified in the last years. CLEF has included a Web track.

## WebCLEF 2006

Within the Cross Language Evaluation Forum (CLEF), a web track has been established in 2005. For the first time, a large multilingual web corpus has been collected and distributed.

The EuroGOV collection represents a crawl over official pages of European governments and ministries. The 3.6 million pages are collected from 27 domains and the collection contains some 25 languages. The size of the collection is 80 GB. For a comparative analysis of the retrieval quality, 547 topics were developed in 2005. In 2006 some topics were developed automatically which led to a set of more than 1800 topics.

For all topics, the original language as well as an English translation were distributed. The participants at the Web track could either process the original version and try to retrieve pages in the same language (mixed mono-lingual task) or use both versions in order to retrieve pages in any language (multi-lingual task).

## Indexing Strategies

The experiments with the EuroGOV corpus carried out at the University of Hildesheim were based on a system developed for multi-lingual retrieval experiments in previous CLEF ad-hoc tasks. The search engine behind the system is Apache Lucene

Indexing method	Document parts indexed	Topic field usage and weighting
word index (no stemming)	full content	original (mixed-mono)
3-gram	content cut-off (first 200 chars.)	original + English (1:1)
4-gram		original + English (10:1)
5-gram	title only	original + English (1:10)

```
<topic>
<num>WC0001</num>
<title>road safety in europe</title>
<metadata>
<topicprofile>
<language language="EN" />
<translation language="EN">road
safety in europe </translation>
</topicprofile>
<targetprofile>
<language language="EN" />
<domain domain="eu.int"/>
</targetprofile>
<userprofile>
<native language="EN" />
<native_other>Tok
Pisin</native_other>
<active language="FR" />
<passive language="DE" />
</metadata>
</topic>
```

## A WebCLEF Topic

## Results and Discussion

The word based run was the best multi-lingual experiment submitted for the CLEF 2005 campaign. Surprisingly, the n-gram runs performed worse than the word based runs. The 4- and 5-grams performed even worse than the tri-gram indexes

Probably, n-gram indexes are sensitive to mixed language indexes and should not be applied in a multi-lingual environment without language identification. Although n-gram indexing performs poorly, simply boosting the original topic language 10:1 compared to the English translation version improves performance

It is especially remarkable that language independent indexing leads to such positive results.

	3-gram, title, mono	3-gram, title, multi	word part, mono	word part, multi	word, full mono	word, full multi
MRR	0.037	0.027	0.130	0.115	0.160	0.137
avg. suc. at 10	0.064	0.049	0.188	0.161	0.219	0.192

## Results 2005

## WebCLEF 2006

For the experiments with the topics for WebCLEF 2006, the indexing strategies were further developed:

- Titles were analyzed to improve the stopword list
- Headings (e.g. H1) were extracted and added to the title
- Emphasized elements in the page were identified and indexed
- Content cut-off was applied more intelligently
- Blind Relevance Feedback was integrated

