REINA at WebCLEF 2008

Carlos G. Figuerola, José L. Alonso Berrocal, Ángel F. Zazo Rodríguez, Montserrat Mateos
REINA Research Group, University of Salamanca
reina@usal.es

Abstract

The task for this year is very similar to last year. However, this time we incorporate last year's experience, in particular, we explored the possibility of improving the selection of snippets, eliminating those that do not make sense, as well as those containing duplicate information. Also, it is intended to explore the real impact of the use of several languages in obtaining relevant fragments.

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; H.3.7 Digital Libraries; H.2.3 [**Database Managment**]: Languages—*Query Languages*

General Terms

Measurement, Performance, Experimentation

Keywords

Web retrieval, Text segmentation

1 Introduction

This year, the WebCLEF track is similar to the 2007 edition, that is: retrieving text snippets or fragments of web pages which bring up information about a topic; additionally, snippets must be in a language from a set of accepted ones. As in 2007, we have a set of topics, each with a title and a short description, as well as several documents or *known sources* about the topic. Additionally, for each topic, we have one or several searches in Google, with the first 1000 documents retrieved.

The system used is basically the same as last year, for each topic we considered all documents retrieved after queries to Google as the collection of documents with which to work. These documents are to be fragmented into pieces, each of whom will be treated as a separate document.

For the queries, we use the description that we have for each topic. This query can be enriched with more terms from the *known sources*. So, the task can be approached like a classic problem of retrieval, and apply, consequently, conventional techniques.

2 Segmentation od web pages

Organizers of task provide us the translation to plain text of documents retrieved by Google. We have valued equal all the searches in Google for same topic. UTF-8 worked fine in almost all cases, something important as there was documents in several languages and with different alphabets (including Cyrillic, for example). That freed us of the many problems experienced in

previous editions with the detection of the coding system of each document. So, for each document translated to plain text, we have to segment it in fragments, to obtain the terms of each fragment and to calculate its weights.

To segment documents and to obtain fragments or short text passages can be applied diverse techniques. Basically, ones are based on the size in bytes, or words; and others are oriented in the separation in phrases or paragraphs [6]. The former techniques produce, of course, pieces more homogenous in size, but often devoid of sense, as the partition point is blind. The other techniques tend to produce fragments of very different size. In addition, its application not always is simple; in many cases the conversion to plain text of a web document loses the separations between paragraphs, nondifference between soft and hard line feeds, or blurs structural elements, like the tables [5].

A simplist approach, like the election of a orthographic character, as the period (.) like reference to fragment the text [2], tends to produce passages too short and, therefore, little useful for the objectives of this task. In our case, we adopted a mixed approach. After several tests, we decided that the suitable size for each fragment was around the 1500 bytes, but as we wanted fragments that had informative sense, our fragmenter looks for the period closest the 1500 bytes, and part by that point.

Some other transformations were carried out: conversion to small letters, removing accents, removing stopwords, (with a long list of stop words for all the accepted languages), application of a simple s-stemmer.

Each fragment thus obtained and transformed was considered an independent document. Terms were extracted and they were weighed according to scheme ATU (slope=0.2) [3], applying to the good well-known vectorial model.

3 Formation of queries

From the document collection formed with snippets, we must select those that are more usefull for each topic. The key is in composing suitable queries that can produce this selection. As sources of information to compose those queries, we have topics with a short title and a brief description. Additionally, we also have, for each topic, a few documents denominated *known sources*, in full text.

So we can use topics (title and description) like nucleus of each query, and enrich this one with terms coming from the *known sources*. The *known sources* are complete documents, which can contain many terms.

4 What's new for this year

Last year we worked on the formation of the queries. The basic dilemma was whether the core of each topic (title and description) could be enriched by the terms of the *known sources*, and to what extent the use of these terms adds useful information to the query.

Results showed little difference in using terms of *known sources* or not. It's better to use these terms, *ma non tropo*. However, runs of last year showed more things. Web pages are not conventional documents; in addition to hyperlinks and hypermedia elements, they have a structure that is not always sequentially. Many web pages are viewed by the user as a set of visual blocks that have different functions and containing different types of information [4]. From the standpoint of obtaining this information, some blocks are more useful than others. The conventional tools of conversion to plain text are not able to reproduce this visual structure, the result is that many of the fragments that we get are meaningless. Others contain information not relevant to our purposes: navigational aids, copyright notices, advertising, etc..

Unfortunately, this visual structure can not be obtained from the elements of HTML and this a difficult area to address, although there are some jobs that are trying to solve these problems [1].

We tried a very naive approach, filtering and dropping snippets based on a simple heuristics: fragments with too many blank lines, with very short lines, with a few words in relation to the size of the fragment, and so on. So, from 639 215 snippets obtained from documents , our filter deleted 165 442 (=25.88 %). This would suggest that we work with a database with a lot of noise, a deepening in the way of extracting fragments could possibly significantly improve results.

In similar way, last year we observed a lot of duplicated snippets. Information is replicated across the web, and so we have fragments of different pages that have the same information. However, as visual presentation is not always the same, the results of the conversion to plain text produces different strings. We used the Dice Coefficient as measure to compare snippets and discover duplicates and almost duplicates. In this case, we applied detection of duplicates on the retrieved snippets for each topic. So, if we consider snippets with Dice similarity greater than 0.7, we found that 11.08 % are duplicates ones.

On the other hand, one may wonder whether the retrieval of fragments in different languages provide more relevant information, and to what extent. All topics allow at least fragments into English as a useful response and, additionally, in other languages. It is expected that these fragments in other languages are derived from queries which include terms in those other languages. We made a run using queries in English only, which should allow us to compare results and assess whether the extent to which the use of other languages aid in retrieval.

5 Conclusions

We have described our approach to WebCLEF task, similar to last year, but incorporating the experience of the last edition. On this occasion, although on the same basis, we have focused on the effect of filtering meaningless fragments, detect duplicates, as well as in estimating the importance of documents in languages other than English. At the time of writing these notes does not yet have the results of experiments.

References

[1] J. Kang and J. Choi. A preliminary report for an information extraction system based on visual block segmentation. Technical Report TR-IS-2007-1, Hanyang University, Intelligent Systems Laboratory, 2007.

[2] Andrei Mikheev. Tagging sentence boundaries. In *Proceedings of the First Meeting of the North American Chapter of the Computational Linguistics (NAACL2000)*, page 264271. Morgan Kaufmann, 2000.

[3] Amit Singhal, Chris Buckley, and Mandar Mitra. Pivoted document length normalization. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 18–22, 1996, Zurich, Switzerland (Special Issue of the SIGIR Forum)*, pages 21–29. ACM, 1996.

[4] Yudong Yang and HongJiang Zhang. Html page analysis based on visual cues. In *ICDAR*, pages 859–864. IEEE Computer Society, 2001.

[5] Shipeng Yu, Deng Cai, Ji-Rong Wen, and Wei-Ying Ma. Improving pseudo-relevance feedback in web information retrieval using web page segmentation. In *Proceedings of the Twelfth International World Wide Web Conference, WWW2003, Budapest, Hungary, 20-24 May 2003*, pages 11–18. ACM, 2003.

[6] Ángel F. Zazo, Carlos G. Figuerola, José Luis Alonso Berrocal, and Emilio Rodríguez. Reformulation of queries using similarity thesauri. *Information Processing & Management*, 41(5):1163–1173, 2005.