Vocabulary Reduction and Text Enrichment at WebCLEF*

[(1)]Franco Rojas, [(1)]Héctor Jiménez-Salazar & [(1,2)]David Pinto
[1]Faculty of Computer Science, BUAP, Mexico
[2]Department of Information Systems and Computation, UPV, Spain
{frlb99, hgimenezs, davideduardopinto}@gmail.com

Abstract

Nowadays, cross-lingual Information Retrieval (IR) is one of the greatest challenges to deal with. Besides, one of the most important issues in IR consists in the corpus vocabulary reduction in order to make possible to use in real situations some methods of IR such as the well-known vector space model. In this work, we have considered a vocabulary reduction process based on the selection of mid-frequency terms. Our approach enhances precision, but in order to obtain a better recall, we have conducted an enrichment process based on the addition of co-ocurrence terms. By using this approach, we have obtained an improvement of 40% in the corpus of the BiEnEs WebCLEF 2005 task. The obtained results in the current mixed monolingual task of the WebCLEF 2006 have shown that the text enrichment must be done before the vocabulary reduction process in order to get the best performance.

Categories and Subject Descriptors

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

General Terms

Measurement, Performance, Experimentation

Keywords

Text reduction, Text enrichment, Mixed-Monolingual

1 Introduction

The multi-language information published in Internet has experimented a big explosion in the last years. This fact, lead us to develop novel techniques to deal with the current cross-language feature of the World Wide Web (WWW). A common language scenario in which a user may be interested in information which is in a different language than their own native language would be when this user has some comprehension ability for a given language but s/he is not sufficiently proficient to confidently specify a search request in that language. Thus, a search system that can deal with this problem should be of a high benefit. The WWW is a natural setting for cross-lingual information retrieval and the European Union is a typical example of a multilingual scenario,

*This work was partially supported by FCC-BUAP and the BUAP-701 PROMEP/103.5/05/1536 grant.

where multiple users have to deal with information published in several languages. Therefore, evaluation environments for cross-lingual information retrieval systems are needed.

The Cross-Language Evaluation Forum (CLEF) has gathered a multi-lingual corpus and promotes the evaluation of cross-lingual information retrieval systems for different types of data [3]. WebCLEF is a particular task for the evaluation of such systems that deals with information on the Web [7]. A detailed discussion of the teams participation and the specific characteristics of the task proposed in the current WebCLEF may be found in [2]. In fact, they have proposed mainly one task for the evaluation of cross-lingual search engines: the Mixed Monolingual task. Thus, in this paper we are reporting the obtained results after the submission of one run to this competition.

We have used a text reduction with an enrichment process and, therefore, we organized this document in four sections. The next section describes the components of our search engine. In Section 3 a discussion of the corpus preprocessing as well as the obtained evaluation results are presented. Finally a conclusion of findings are given.

2 Description of the search engine

The aim of this approach was to determine the behaviour of document indexing reduction in a cross language information retrieval environment. We have used a boolean model with the Jaccard similarity formula in order to rank the obtained documents related to some query. In order to reduce the terms of every document treated, we have applied a mid-frequency terms based technique, named Transition Point, which is described as follows.

2.1 The Transition Point Technique

The Transition Point (TP) is a frequency value that splits the vocabulary of a text into two sets of terms (low and high frequency). This technique is based on the Zipf Law of Word Ocurrences [10] and also on the refined studies of Booth [1], as well as of Urbizagástegui [9]. These studies are meant to demonstrate that mid-frequency terms, of a text T, are closely related to the conceptual content of T. Therefore, it is possible to establish the hypothesis that terms closer to TP can be used as index terms of T. A typical formula used to obtain this value is: $TP = (\sqrt{8 * I_1 + 1} - 1)/2$, where I_1 represents the number of words with frequency equal to 1; see [5] [9].

Alternatively, TP can be localized by identifying the lowest frequency (from the highest frequencies) that it is not repeated in the text; this characteristic comes from the properties of the Booth's law of low frequency words [1]. In our experiments we have used this approach.

Let us consider a frequency-sorted vocabulary of a document; i.e., $V_{TP} = [(t_1, f_1), ..., (t_n, f_n)]$, with $f_i \geq f_{i+1}$, then $TP = f_{i-1}$, iif $f_i = f_{i+1}$. The most important words are those nearest to the TP, i.e.,

$$TP_{SET} = \{t_i | (t_i, f_i) \in V_{TP}, U_1 \leq f_i \leq U_2\}, \quad (1)$$

where U_1 is a lower threshold obtained by a given neighbourhood percentage of TP (NTP), thus, $U_1 = (1 - NTP) * TP$. U_2 is the upper threshold and it is calculated in a similar way ($U_2 = (1 + NTP) * TP$). Either in WebCLEF-2005 and in the current competition, we have used $NTP = 0.4$, considering that the TP technique is language independent.

2.2 Term Enrichment

Certainly TP reduction may increase precision, but furthermore it decreases recall. Due to this fact, we enriched the selected terms by obtaining new terms, those with similar characteristics to the initial ones. Specifically, given a text T, with selected terms TP_{SET}, y is a new term if it co-occurs in T with some $x \in TP_{SET}$, i.e.,

$$TP'_{SET} = TP_{SET} \cup \{y | x \in TP_{SET} \wedge (fr(xy) > 1 \vee fr(yx) > 1)\}. \quad (2)$$

Considering the text length, we only selected a window of size 1 around each term of TP_{SET}, and a minimum frequency of two for each bigram was required as condition to include new terms.

2.3 Information Retrieval Model

Our information retrieval is based on the Boolean Model and, in order to rank the documents retrieved, we have used the Jaccard similarity function applied to both, the query and every document of the corpus used. Previously, each document was preprocessed and its index terms were selected (the preprocessing phase is described in section 3.1). As we will see in Section 3.3, we have represented each text by using the selection formula given in the Equation (1). Additionally, after the reduction step, we have carried out an enrichment process (see Equation (2)) based on the identification of related terms to those resulted from the selection.

3 Evaluation

3.1 Corpus

For the experiments carried out, we have used the EuroGOV corpus provided by the WebCLEF forum which is well described in [6]. We have indexed only 20 from 27 domains, namely: DE, AT, BE, DK, SI, ES, EE, IE, IT, SK, LU, MT, NL, LV, PT, FR, CY, GR, HU, and UK (we did not indexed the following domains: EU, RU, FI, PL, SE, CZ, LT). Due to this fact, only 1,470 from 1,939 topics were evaluated, which is approximately a 75.81% of the whole topics. Although we presented in Section 3.3 the MRR over 1,939 topics, there were 469 topics not indexed.

The preprocessing phase of the EuroGOV corpus was carried out by writing two scripts to obtain index terms for each document. The first script uses regular expressions for excluding all the information which is enclosed by the characters < and >. Although this script obtains very good results, it is very slow and therefore we decided to used it only with three domains of the EuroGOV collection, namely Spanish (ES), French (FR), and German (DE).

On the other hand, we wrote a script based in the *html* syntax for obtaining all the terms considered interesting for indexing, i.e., those different than script codes (javascript, vbscript, style cascade sheet, etc), *html* codes, etc. This script speeded up our indexing process but it did not take into account that some web pages were incorrectly written and, therefore, we missed important information from those documents.

For every page compiled in the EuroGOV corpus, we also determine its language by using TexCat [8], a language identification program widely used. We constructed our evaluation corpus with those documents identified as a language of the above list.

Another preprocessing problem consisted in the charset codification, which led to a even more difficult analysis. Although the EuroGOV corpus is given in UTF-8, the documents that made up this corpus do not neccesarily keep this charset. We have seen that for some domains, the charset codification is given in the *html* metadata tag, but also we found that very often this codification is wrong. We consider the charset codification detection the most difficult problem in the preprocessing step. Finally, we eliminated stopwords for each language (except for Greek language) and punctuation symbols. For the evaluation of this corpus, a set of queries was provided by WebCLEF-2006, which were applied with the same preprocessing process described above.

3.2 Indexing reduction

After our first participation in WebCLEF [4], we carried out more experiments using only those documents in Spanish language from the EuroGOV corpus. We observed that a value of $NTP = 0.4$ using the reduction process shown in the Equation 1 was adequated. Therefore, in this test we carried out one run with that value. Moreover, this run took the evaluation corpus composed by the reduction of every text, using TP technique with a neighbourhood of 40% around TP, an enriched this set of terms using related terms as described by Equation (2).

Table 1 shows the size of every evaluation corpus used; the vocabulary composed by representation of all texts, $|TP'_{SET}|$, as well as the percentage of reduction obtained by each one with respect to the original vocabulary. As we can see, the TP technique obtained a vocabulary reduction percentage of more than 95%, which implies a time reduction for any search engine indexing process.

Table 1: Vocabulary size and percentage of reduction.

Domain	DE	AT	BE	DK	SI
Size (KB)	2,588	2,317	6,796	1,189	6,729
Reduction (%)	95.3	97.2	98.0	97.9	97.1
Domain	ES	EE	IE	IT	SK
Size (KB)	16,271	4,838	2,632	11,913	14,668
Reduction (%)	98.5	97.2	96.0	98.4	97.5
Domain	LU	MT	NL	LV	PT
Size (KB)	3,212	4,817	20,324	21,213	9,134
Reduction (%)	99.2	95.7	97.7	97.8	97.6
Domain	FR	CY	GR	HU	UK
Size (KB)	22,083	18,814	340	10,440	14,239
Reduction (%)	95.8	96.5	97.4	98.8	96.1

3.3 Results

Table 2 shows the results for the run submitted. The first and second column indicates the number of topics evaluated and the test type. The last column shows the Mean Reciprocal Rank (MRR) obtained for each test. Additionally, the average success at different number of documents retrieved is shown; for instance, the third column indicates the average success of our search engine at the first answer.

Table 2: Evaluation results

		Average Success at					
#Topics	**Test**	**1**	**5**	**10**	**20**	**50**	**Mean Reciprocal Rank**
1939	All	0.0093	0.0217	0.0294	0.0371	0.0464	0.0157
1620	Auto	0.0025	0.0049	0.0086	0.0117	0.0160	0.0040
319	Man	0.0439	0.1066	0.1348	0.1661	0.2006	0.0750
810	A. bi.	0.0037	0.0062	0.0099	0.0123	0.0148	0.0049
124	M. new	0.0323	0.0968	0.1129	0.1613	0.2339	0.0657
810	A. uni.	0.0012	0.0037	0.0074	0.0111	0.0173	0.0031
195	M. old	0.0513	0.1128	0.1487	0.1692	0.1795	0.0810

The best overall results by using the *new topic set* is presented in Table 3. The results are reported on all topics (all), the automatic (auto) and manual (manual) subsets of topics. The average of the submitted run is shown in the fifth column.

Table 3: Results by using the new topic set

Run	**all**	**auto**	**manual**	**average**
allpt40bi	0.0272	0.0080	0.0790	0.0435

4 Conclusions

We have proposed an index reduction method for cross language search engines, which includes an enrichment step. Our proposal is based on the transition point technique which allows to index only the mid-frequency terms from every document. Our method is linear in computational time and, therefore, it can be used in a wide spectrum of practical tasks.

After submitting our run we observed enhancement if we compare the results obtained with those of the BiEnEs task in WebCLEF 2005. By using the enrichment, more than 40% on MRR was achieved. However, by using the Vector Space Model similar results to boolean model were obtained.

The TP technique has shown an effective use on diverse areas of NLP, and its best features for NLP, are mainly two: a high content of semantic information and the sparseness that can be obtained on vectors for document representation on models based on the vector space model. On the other hand, its language independence allows to use this technique in multilingual environments.

We consider that our approach may be improved by taking into account all the terms of the vocabulary in the enrichement process. Once the term expansion would be done, the mid-frequency selection technique could be applied. Further analysis will investigate this issue.

References

[1] A. Booth: *A Law of Ocurrences for Words of Low Frequency*, Information and control, 1967.

[2] K. Balog, L. Azzopardi, J. Kamps, M. d. Rijke: *Overview of WebCLEF 2006*, In proceedings of the CLEF'06: Cross-Language System Evaluation Campaign, A. Nardi, C. Peters and, J. L. Vicedo Eds., page 51, 2006.

[3] CLEF 2005: *Cross-Language Evaluation Forum*, http://www.clef-campaign.org/, 2005.

[4] D. Pinto, H. Jiménez-Salazar, P. Rosso, E. Sanchis: *TPIRS: A System for Document Indexing Reduction on WebCLEF*, Extended abstract in Working notes of CLEF'05, Viena, 2005.

[5] B. Reyes-Aguirre, E. Moyotl-Hernández & H. Jiménez-Salazar: *Reducción de Términos Indice Usando el Punto de Transición*, In proceedings of Facultad de Ciencias de Computación XX Anniversary Conferences, BUAP,2003.

[6] B. Sigurbjörnsson, J. Kamps, and M. de Rijke: *EuroGOV: Engineering a Multilingual Web Corpus*, In Proceedings of CLEF 2005, 2005.

[7] B. Sigurbjörnsson, J. Kamps, and M. de Rijke: *WebCLEF 2005: Cross-Lingual Web Retrieval*, In Proceedings of CLEF 2005, 2005.

[8] TextCat: *Language identification tool*, http://odur.let.rug.nl/ vannord/TextCat/, 2005.

[9] R. Urbizagástegui: *Las posibilidades de la Ley de Zipf en la indización automática*, Research report of the California Riverside University, 1999.

[10] G. K. Zipf: *Human Behavior and the Principle of Least-Effort*, Addison-Wesley, Cambridge MA, 1949.