COMBINING INDEXING METHODS AND QUERY SIZES IN INFORMATION RETRIEVAL IN FRENCH

Désiré Kompaoré, Josiane Mothe
Institut de Recherche en Informatique de Toulouse, Université de Toulouse, France
josiane.mothe@irit.fr

Ludovic Tanguy
ERSS, Université de Toulouse, France
tanguy@univ-tlse2.fr

Keywords: Information retrieval, data fusion, indexing, information retrieval in French

Abstract: This paper analyses three type of different indexing methods applied on French test collections (CLEF from 2000 to 2005): lemmas, truncated terms and single words. The same search engine and the same characteristics are used independently to the indexing method to avoid variability in the analysis. When evaluated on French CLEF collections, indexing by lemmas is the best method compared to single words and truncated term methods. We also analyse the impact of combining indexing methods by using the CombMNZ function. As CLEF topics are composed of different parts, we also examine the influence of these topic parts by comparing the results when topic parts are considered individually, and when they are combined. Finally, we combine both indexing methods and query parts. We show that MAP can be improved up to 8% compared to the best individual methods.

1 INTRODUCTION

Information retrieval is composed of various processes which distinguish systems from each other. *Indexing* aims at building a reduced representation of the contents of the documents and queries. The majority of indexing techniques analyzes the contents of texts, and eliminates stop words to keep only representative descriptors. Depending on the method used, these descriptors can be represented in many forms: original terms in documents, stems or lemmas. When considering documents, these descriptors are generally weighted in order to depict how a term represents the document content and how they can separate relevant and non-relevant documents when a query contains this term. The *weighting function* also helps in comparing systems performances. The *searching function* (or model) is used to match query and document representations resulting from indexing, in order to decide which documents to retrieve. This function calculates the document scores and characterizes systems. These scores indicate the degree of possible relevance of the retrieved documents.

Specific studies focus on the influence of such or such parameter on the efficiency of the search, by choosing a parameter which varies while trying to keep the other parameters identical. Other works study the combination of various searches for a given query: different representation of queries (Fox and Shaw, 1994), various search models (McCabe et al., 1999), various search systems (Hubert et al., 2006) to improve the results.

Such studies are made possible by the existence of test collections, such as those of TREC, CLEF, or INEX, and of criteria to measure the efficiency of search engines. A collection of evaluation is

composed of a set of documents, a set of topics and the set of documents judged as relevant for each topic. Evaluation is usually based on various criteria that are calculated using the trec_eval tool (trec.nist.gov). Basically, evaluation is based on recall, which measures if the relevant documents are retrieved, and on precision, which measures if the retrieved documents are relevant.

Our study focuses on monolingual information retrieval in French and analyses three indexing techniques where indexes are terms (documents or queries terms), truncated terms, or lemmas. We analyse in a first step the various indexing modes considered individually, and in a second step we combine them with the CombMNZ function (Fox and Shaw, 1994). We also studied the influence of the size of queries on the results. In this study, the query size is based on the structure of topics from evaluation campaigns. The evaluation is carried out on 6 collections of French ad-hoc collection of CLEF.

This paper is organized as follows: section 2 presents some related works. Section 3 presents the various modes of indexing we analyse. Section 3 presents the collections as well as the criteria of evaluation. Section 4 reports the results obtained and discuss these results. We then conclude this paper and indicate some directions for future works.

2 SOME RELATED WORKS

Some studies consider the various parts of a topic when analysing system results. (Savoy, 2003) indicates that, on the French collection of CLEF 2000 (34 queries), the majority of the ten studied systems improve the average precision (MAP) when the title and the description are taken into account. The improvement compared to title only is on average of 5%. Improvement when considering the complete query compared to title only is 10.21%. Within the framework of the TREC TetraByte track, (Metzler et al., 2005) show that the average precision is improved on average by 5% when one adds the description to the title, and by 10,4% when one adds the narrative to the title and description. (Ahlgren and Kekäläinen, 2006) study the Swedish collection of CLEF 2003 and various strategies of indexing. 4 combinations out of 7 based on a morphological analyser; one is based on a truncation and one on a stemming-based approach. Truncation gives the best results.

3 DOCUMENTS AND QUERIES INDEXING METHODS

Text indexing consists in two principal steps: extracting the terms that characterize the document contents and assigning a weight to each of these indexing terms. This weight reflects the capacity of characterization of the document by the term.

The three indexing approaches we study vary according to the unit of indexing. The weighting function remains the same and is based on the BM25 function (Roberston et al., 1995). We use in these experiments the Mercure system (Boughanem et al, 1998). Search is based on matching queries and documents indexed by the same method.

3.1. Single words indexing

According to this indexing method, stop words are eliminated but accents are preserved. The remaining single terms correspond to indexes. Indexes thus correspond exactly to terms in documents and queries. This mode of indexing, called SW indexing in the paper, is supposed to increase precision as the documents that will be retrieved contains the terms exactly as they appear in the query.

3.2. Indexing by truncated terms

Truncated terms (TT) are used to conflate the different forms of a term into a single one. This should improve recall due to the use of various forms of terms in the documents and queries. Using TT is a simple method to reduce morphological variation of words. Several methods for generating truncated terms exist and are based on statistical considerations: deletion of the most frequent word endings, application of rules such as in Porter-like algorithms. In our approach, we use a truncation at 7 characters as (Denjean, 1989) suggested. Stop words and accents are also removed.

3.3. Indexing by lemmas

Indexing by lemmas (L) has the same objective than stem indexing: to limit the risks of not retrieving documents containing various forms of query terms. However, lemmas extracting is more efficient than word truncation. For example, “computers” and “compute” could be associated to the same radical “compute” by truncation at 7 characters. Lemmatization will not make this type of association and will lead to two indexing terms (“computer” as a noun in singular form and “compute” as a verb).

In the experimentation that we report here, lemmas are extracted using TreeTagger (*TreeTagger*, H. Schmidt; www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger), stop words are removed, but accents are preserved since it is consistent with the fact to retrieve lemmas.

4 EXPERIMENTS

4.1 Collections of evaluation

The evaluation is based on 5 years of the French monolingual ad-hoc CLEF collection. We chose this collection because this is the main used when considering French.

The collection contains documents from ATS (SDA) 1994 and 1995, and articles from Le Monde newspaper 1994. It also includes about 50 topics per year. Table 1 indicates various characteristics of the collections and the number of indexing terms resulting from the various indexing methods. CLEF 2001 and CLEF 2002 use the same document collection, for that reason, we consider a set of (49+50) topics on this collection.

Topics are composed of three parts (see figure 1): a *title* (notes T) which is limited to a few words, a *descriptive* (noted D) which explains in one or two sentences the title, and a *narrative* (noted N) which indicates what a relevant document is and eventually what will not be relevant. In our study, 9 queries can be built from one topic that result from the three types of indexes (SW, TT, L) and the use of *T* only, *T+D,* or *T+D+N*.

<num> C001
<title> Architecture à Berlin
<desc> Trouver des documents au sujet de l'architecture à Berlin.
<narr> Les documents pertinents parlent, en général, des caractéristiques architecturales de Berlin ou, en particulier, de la reconstruction de certaines parties de cette ville après la chute du mur.

Figure 1: Example of CLEF topic

4.2 Evaluation and methodology

Trec_eval (trec.nist.gov) is a program which is used to evaluate systems performances according to a certain number of measures. In this study, we choose the following measures:

- MAP Mean Average Precision. Average precision for a topic is the average of the precision obtained after a new relevant document is retrieved. MAP is the average of the average precision over a topic set.
- Average precision at 5 documents. Precision at 5 documents corresponds to the proportion of relevant documents in the first 5 retrieved documents. It is averaged over the topics.

MAP is used for global comparisons (Voorhees, 2007). On the other hand, precision at 5 is a good indicator of users' satisfaction since users generally have a look to the top retrieved documents.

In a first step, we want to analyze the impact of the indexing methods on systems performances. These indexing methods are applied successively on the *T*, *T+D*, and *T+D+N* topic fields ; results are then compared. In a second stage, we combine the different type of indexing (SW, TT, L). We use CombMNZ function to combine the results.

The CombMNZ function (Fox and Shaw, 1994) is widely used in data fusion studies (Beitzel et al., 2004). Formula (1) indicates how the CombMNZ function calculates the score of a document J after fusion. Function (1) takes into account two parameters:

- the score of the document in each fused result,
- the number of systems which retrieved a document

$$ScoreCombMNZ_j = \sum_{i=1}^{nbre\_syst} Score_{ij} \cdot Count_j \quad (1)$$

where $Score_{ij}$ is the score calculated by system *I* for document *J*, and $Count_j$ is the number of fused systems which retrieve document *J*.

5 RESULTS

5.1 Results without combinations

Table 2 indicates the mean average precision (MAP) and precision at 5 documents (P5) on the sets of topics. In this table, *T+D* sections of the topics are considered and the three methods of indexing are used independently on these sections. We chose to present first these results since using *T+D* is the most used method in CLEF by participants. Values in bold font indicate the best results for a measure for a given year. The line entitled "average" indicates the MAP and P5 value for each method, averaged over years. The line entitled "Var. in %" indicates the variation in percentage of performance.

The baseline is the SW indexing method.

For four collections, lemma-based indexing outperforms the others methods. For example, when one considers the CLEF 2000 collection, the MAP is improved by approximately 7% compared to the SW indexing. Lemmas are the most effective indexes for all CLEF, except 2001/02. None of the collections should use SW indexing. TT and L lead to improvement of both MAP and P5. To be more precise, on average MAP is improved by 6.06% (resp. 7.39%) compared to SW and P5 by 3.24% (resp. 5.6%). Considering statistical significance, we use the wilcoxon test with a p-value < 0.05 to consider the results significant. TT is better than SW (statistically significant); however, the difference between L and TT indexing is not statistically significant.

Table 3 shows the results obtained by the indexing method on the *title* section of the topic. The additional line (Var. in % TD) indicates the average variations observed compared to the *T+D* section of Table 2 (same measure and same indexing units). When titles only are considered, the performances are overall lower than those obtained when the descriptive section is also considered. This result is not surprising since the title is enriched by the descriptive section and thus potentially makes it possible to better meet the user's needs. On average, the best method is the one that uses L indexing. Using SW remains the worth method. L or TT indexing is on average better than using SW. In average, when considering all collections, L indexing improves by 11.36% the MAP compared to the technique of the SW when *titles* are considered whereas this improvement is of 7.39% when *T+D* are considered. This means that the indexing method has a greater impact when considering short queries. Again, considering statistical significance, TT is better than SW; however, L and TT are not different.

When complete topics are considered, there is no notable improvement on the best results compared to search when queries are built considering title and descriptive sections. Lemma-based indexing is still the best method and gets about the same results than using *T+D* whatever the measure (MAP or P5) when averaged over the years. However, SW indexing benefits from the enrichment of the query by the *N* section (+3.40% for MAP and +5.14% for P5) compared to *T+D*. Detailed results using *T+D+N* are not presented here.

The comparative results obtained when the various sections of the queries are used tend to indicate that taking into account a more semantic indexing (by lemmas) is especially effective on short queries (title only). Indeed, it is within this framework that we obtain the greatest variations between the techniques of indexing, with a superiority of the L indexing.

5.3. COMBINATION OF INDEXING METHODS

In this section, we study the combination of the indexing methods and the influence of these combinations on the performances. The variability of MAP according to the indexing method used suggested it was relevant to combine them. Depending on the topic, this variation can be up to 600%.

For this reason, we apply CombMNZ on pairs of methods and also to combine the three indexing methods. Surprisingly, the results we obtained slowed that one unique method (without fusion) obtains comparable or higher MAP than any of the combinations. The only exception occurs for the collection of CLEF 2003 for which a combination (SW + TT + L) improves the results compared to the best of the simple techniques. When one considers the average over the years, the combination of the three methods of indexing does not improve MAP compared to the best indexing, by lemmas.

If we analyse the results of the precision at 5 documents, we can also select a unique indexing method that obtains performance equal or better than any of the combinations. Even if this best simple indexing method is not the same one on each collection, on average, the non-fused indexing method based on L remains the best. The details of these results are not displayed in this paper.

5.4. Combination OF THE SIZES OF QUERIES

The combination of the query sizes consists in fusing results obtained using a single method to extract indexes but considering the three versions of the topic: T, *T+D*, *T+D+N*. That means that for example, when considering SW, we fuse the three results obtained by three runs: the one that uses *T* only, the one that uses *T+D* and the one that uses the three parts of the topic. Table 4 presents the results of this combination. On average, the best results are obtained using L indexing.

In addition, whatever the collection and whatever the indexing method, system performance is improved. This performance is compared to unique systems or fused systems that consider *T+D* only (cf table 2 and last line of table 4). Combining queries size also improves the results compared to those obtained with *titles* section only (cf table 3,

average line) and compared to those obtained with the *T+D+N*, except for P5 when indexing by lemmas. Apart from a few cases, the results are statistically significant.

5.5. COMBINATION OF THE SIZE OF QUERIES AND THE METHODS OF INDEXING

We have investigated in this section the effect of combining both indexing methods and queries size (3 sizes of query and 3 methods of indexing), see table 5. This method improves a bit more the high precision (P5). The baseline for the analysis is L indexing which is the best non-combined method on average. MAP is improved of more than 8%. Results are improved compared to each of the three preceding combinations. Apart from a few cases, the results are statistically significant.

7. CONCLUSIONS

In this study, we consider French mono-lingual information retrieval and analyze the influence of different indexing methods considering different types of indexing methods and query size. We analyzed three different indexing methods respectively single words, truncated terms and lemmatization. Our experiments are done on the French collections of CLEF from 2000 to 2005 and we have shown in an experimental way that the use of the lemmas-based indexing was the most effective unique method when one is interested in mean average precision and high precision. The difference is of more than 7% for MAP and about 6% for P5 (statistically significant). We have also shown that it was relevant to combine the results obtained using different query size (use of various sections of the topics). Therefore, combining the various methods of indexing does not make a significant improvement in the results.

Future work will focus on the contextual combination of the indexing methods and variation in the sizes of the queries. Indeed, one can think that a query for which the terms used have many alternatives but come from various concepts should be rather indexed by lemmas. On the other hand, a query for which few documents are retrieved would gain in being indexed by truncated terms in order to *expand* it. Recent works in the literature have studied how to predict query difficulty in terms of recall and precision. Others were interested in predicting the possibility of finding relevant documents in the collection. Our future work will consider these aspects: system fusion will be based on some knowledge on the difficulty of the query and on other elements of the context of the query in order to decide which sections of the query to consider (query size) and what indexes should be used (either a single type or a combination).

ACKNOWLEDGEMENTS

This work has partially been supported by the French ANR through the TCAN program (ARIEL Project) and by the European Commission through the project WS-Talk under the 6th FP (COOP-006026). However views expressed herein are ours.

REFERENCES

Ahlgren P. and Kekäläinen J., 2006. Swedish full text retrieval: Effectiveness of different combinations of indexing strategies with query terms, Information Retrieval journal, 9(6): 681-697.

Beitzel S.M. et al., 2004, Fusion of Effective Retrieval Strategies in the Same Information Retrieval System. JASIST, 55(10): 859-868.

Boughanem M., Dkaki T., Mothe J., Soulé-Dupuy C., 1998, Mercure at trec7, NIST 500-242, 413-418.

Denjean P., 1989, Interrogation d'un systeme videotex : l'indexation automatique des textes, PhD dissertation, Université de Toulouse, France.

Fox E.A. and Shaw J.A., 1994, Combination of Multiple Searches, TREC-), NIST 500-215, 243-252.

Hubert G. and Mothe J., 2007, Relevance feedback as an indicator to select the best search engine, ICEIS 2007, 184-189.

Kompaoré N. D. and Mothe J., 2007, Probabilistic fusion and categorization of queries based on linguistic features, ACM PIKM, 63-68.

Lee J., 1997, Analysis of multiple evidence combination, ACM SIGIR, 267-276.

Lu X. A. and Keefer R. B., 1994, Query Expansion/Reduction and its Impact on Retrieval Effectiveness, NIST 500-225, TREC-3, 231-240.

McCabe M. C., Chowdhury A., Grossman D.A., Frieder O., 1999, A unified Environment for Fusion of Information Retrieval, ACM CIKM, 330-334.

Metzler D., Strohman T., Zhou Y., Croft W. B., Indri at TREC 2005: Terabyte Track.

Mothe J. and Tanguy L., 2005, Linguistic features to predict query difficulty, SIGIR wkshop on Predicting Query Difficulty - Methods and Applications.

Robertson S E, et al., 1995, Okapi at TREC-3, Overview of the Third Text REtrieval Conference, 109-128.

Savoy J., 2003, Cross-language information retrieval: experiments based on CLEF 2000 corpa, IPM, V. 39, 75-115.

Voorhees, E.M., 2007. Overview of TREC 2006, NIST, MD 20899, 1-16.

	CLEF 2000	CLEF 2001/2002	CLEF 2003	CLEF 2004	CLEF 2005
Nber of topics	34	49+50	52	49	50
Nber of documents	44 13	87 191	129 806	90 261	177 452
Nber of SW	295 156	339 879	390 742	387 386	563 199
Nber of TT	195 47	228 354	270 281	290 646	410 155
Nber of Lemmas	246 400	290 037	340 887	340 811	509 475

Table 1: Characteristics of the test collections used

	SW		TT		L	
Year	MAP	P5	MAP	P5	MAP	P5
2000	0.3946	0.4118	0.4067	0.4000	**0.4333**	**0.4588**
2001/2002	0.3916	0.4524	**0.4401**	**0.5168**	0.4326	0.4786
2003	0.4888	0.4654	**0.5043**	0.4615	0.4831	**0.4731**
2004	0.4174	0.4612	0.4311	0.4408	**0.4479**	**0.4612**
2005	0.2827	0.4400	0.3125	0.4840	**0.3241**	**0.4840**
Average	0.3950	0.4462	0.4190	0.4606	**0.4242**	**0.4711**
Var. in %	-	-	+6.06	+3.24	+7.39	+5.60

Table 2: Results of each indexing method– Title and descriptive

	SW		TT		L	
Year	MAP	P5	MAP	P5	MAP	P5
2000	0.4002	0.4059	0.3963	0.3941	**0.4107**	**0.4235**
2001/2002	0.335	0.3859	0.3933	0.4324	**0.3939**	**0.4346**
2003	0.4212	0.4000	**0.4567**	**0.4192**	0.4382	**0.4192**
2004	0.3644	0.3918	**0.3995**	**0.4286**	0.3962	0.4245
2005	0.2158	0.4160	0.2867	0.4760	**0.2948**	**0.4920**
Average	0.3473	0.3999	0.3865	0.4301	0.3868	0.4388
Var. in %	-	-	+11.28	+7.54	+11.36	+9.71
Var. in % TD	-12.08	-10.36	-7.74	-6.63	-8.83	-6.87

Table 3: Results of each indexing method – Title only

	SW		TT		L	
Year	MAP	P5	MAP	P5	MAP	P5
2000	0.4320	0.4529	0.4354	0.4294	**0.4568**	**0.4765**
2001/2002	0.4352	**0.5229**	0.4516	0.4972	**0.4740**	0.4352
2003	0.5041	**0.4962**	0.5309	**0.4962**	**0.5392**	0.4846
2004	0.4331	0.4612	**0.4472**	**0.4735**	0.4442	0.4531
2005	0.2914	0.4800	0.3386	**0.5360**	**0.3454**	0.5000
Average	0.4192	0.4826	0.4407	**0.4865**	**0.4519**	0.4699
Var % vs TD	+6.13	+8.16	+5.18	+5.62	+6.53	-0.25

Table 4: Results of each method of indexing – Combining the sizes of the queries

MAP	L	Comb.	%		**P5**	L	Comb.	%
2000	0.4333	0.4647	+7.25%		2000	0.4588	0.4464	-2.70%
2001/2002	0.4326	0.5336	+23.35%		2001/2002	0.4786	0.4635	-3.16%
2003	0.4831	0.5115	+5.88%		2003	0.4731	0.5355	+13.19%
2004	0.4479	0.4497	+0.40%		2004	0.4612	0.4571	-0.89%
2005	0.3241	0.3477	+7.28%		2005	0.4840	0.5280	+9.09%
Average.	0.4242	0.4614	+8.78%		Average.	0.4711	0.4861	+3.18%

Table 5: MAP and P5 - Results obtained from indexing methods and query size