University of Jaén at ImagePhoto 2008: Filtering the Results with the Cluster Term

Miguel Angel García-Cumbreras, Manuel Carlos Díaz-Galiano, María Teresa Martín-Valdivia, and L. Alfonso Ureña-López

SINAI Research Group*, Computer Science Department, University of Jaén, Spain
{magc,mcdiaz,maite,laurena}@ujaen.es
http://sinai.ujaen.es

Abstract. This paper describes the University of Jan system presented at ImagePhoto CLEF 2008. Previous systems used translation approaches and different information retrieval systems to obtain good results. The queries used are monolingual, so translation methods are not necessary. The new system uses the parameters that obtain the best results in the past. The novelty of our method consists of some filtered methods that are used to improve the results, with the cluster terms and its WordNet synonyms. The combination of different weighting functions (Okapi and Tfidf), the results obtained by the information retrieval systems (Lemur and Jirs), and the use or not of automatic feedback complete the experimentation.

1 Introduction

In this paper our system has been tested with the framework provided by ImagePhoto CLEF organization[1].

Our system only uses textual information, not visual information, to improve the retrieval methods. Two Information Retrieval (IR) systems have been run, and the experiments test the use of automatic feedback and different weighting functions (Okapi and Tfidf). A simple method has been developed to filter the results with the cluster term and its WordNet[1] synonyms. It has been applied in some experiments.

Section 2 describes the complete system. In Section 3 the experiments and results are shown. Finally, Section 4 contains the analysis of the results and the main conclusions.

2 System Description

In our system there is not user interaction, and we have used the English textual information (not visual information).

* http://sinai.ujaen.es
[1] Available at http://wordnet.princeton.edu

C. Peters et al. (Eds.): CLEF 2008, LNCS 5706, pp. 593–596, 2009.

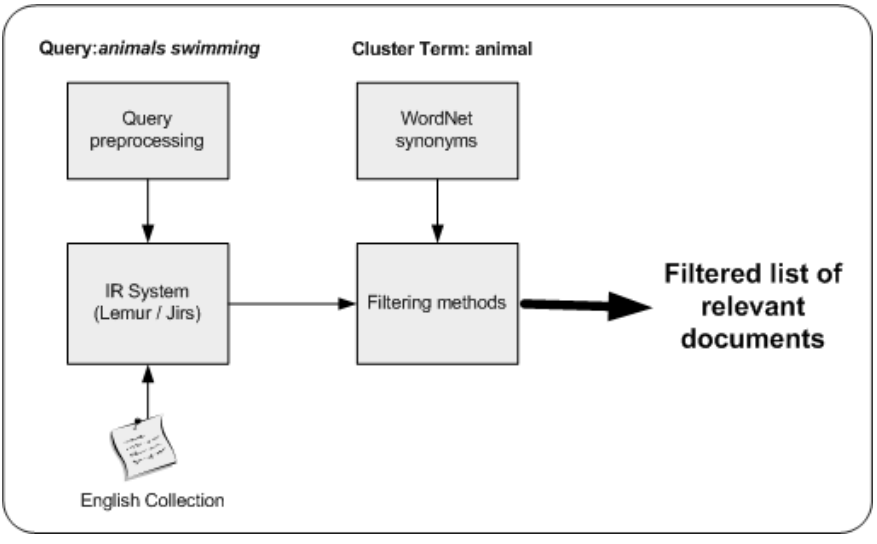

Fig. 1. General scheme of the SINAI system at ImagePhoto 2008

The English collection documents have been preprocessed as usual (English stopwords removal and the Porter's stemmer[2]). Then, the documents have been indexed using as IR systems: Lemur[2] and Jirs[3].

Previous results of our system shown that precision results obtained with both IR systems were very similar. Only the results with Italian queries were quite different[4]. A simple combination method with both IR results was developed, and the evaluation of the combined list of relevant documents fix the parameter that weight each list in 0.8 for Lemur documents and 0.2 for Jirs documents. Using the same combination parameters, the new system try to improve the results with different combinations of methods and the application of a filter with the cluster term. The weighting function of the IR systems is a parameter changed to test the results.

The use of Psedo-Relevance Feedback (PRF) to improve the retrieval process is not conclusive, but in general the precision is increased in past experiments, so it is used always with Lemur.

After the retrieval process, the documents or passages marked as relevant are filtered, using the cluster terms, as follows:

1. The cluster term is expanded with its WordNet synonyms (the first sense).
2. The list of relevant documents generated by the IR system is filtered. If the relevant document contains the cluster term or a synonym its docid (the identifier of the document) is written in another list.
3. Finally, the new list with the filtered documents is combined with the original ones (Lemur and Jirs) in order to improve them. A simple method to do this

[2] Available at `http://www.lemurproject.org/`

was to duplicate the score value of the documents in the filtered list and to add them to the original ones.

Figure 2 shows a general architecture of the system delevoped.

3 Experiments Description and Results

We have made the following experiments:

1. **SINAI exp1 Baseline**. It is the baseline. Lemur is used as IR system with automatic feedback. The weighting function applied was Okapi. There was no combination of results, nor filtering method with the cluster term.
2. **SINAI exp2 LemurJirs**. This experiment combines the IR lists of relevant documents. Lemur also uses Okapi as weighting function and PRF. Before the combination of results Lemur and Jirs lists are filtered, only with the cluster term.
3. **SINAI exp3 Lemur fb okapi**. The Lemur list of relevant documents is filtered with the cluster term and its WordNet synonyms. Okapi is used as weighting funcion, and PRF is applied automatically.
4. **SINAI exp4 Lemur fb tfidf**. It is the same experiment as before, but in this case the weighting function used was Tfidf.
5. **SINAI exp5 Lemur simple okapi**. Lemur IR system has been run with Okapi as weighting function and without feedback. The list of relevant documents has been filtered with the cluster term and its WordNet synonyms.
6. **SINAI exp6 Lemur simple tfidf**. Lemur IR system has been used with Tfidf as weighting function and without feedback. The list of relevant documents has not been filtered.

Table 1 presents the results, with and without filtering. All the results are based on textual information of the English queries. The last column shows the mean F1-Measure obtained in ImagePhoto 2008, with an automatic system and only text.

Table 1. Results

Id	Filtering	Modality	FB	Expansion	MAP	P@5	P@10	Best F1
(1)	No	Text	Yes	No	**0.2125**	**0.3744**	**0.3308**	0.2957
(6)	No	Text	No	No	0.2016	0.3077	0.2872	0.2957
(2)	Yes	Text	Yes	No	0.2063	0.3385	0.2949	0.2957
(3)	Yes	Text	Yes	No	**0.2089**	**0.3538**	0.3128	0.2957
(4)	Yes	Text	Yes	No	0.2043	0.2872	0.2949	0.2957
(5)	Yes	Text	No	No	0.1972	0.3385	**0.3179**	0.2957

4 Discussion and Conclusions

In this paper we have presented the results of our architecture to retrieve information from a multimedia corpus, presented in the ImageCLEF 2008 Photo task. We have experimented with two major variables, a filtering process that used the cluster term, and its synonyms in one case, and some changes in the retrieval parameters, such as the weighting function or the use or automatic feedback.

The results show that a filtering method is not useful if the cluster term or related words are used to filter the IR retrieved documents, because some good documents are deleted and none of non retrieved relevant documents are included in the second step. In general, the results in term of MAP or other precision values are not so different. Between the best MAP and the worse one the difference is less than 8%. Filtering methods have not improved the baseline cases. After an analysis of the performance we can write some reasons:

- Some relevant documents that appear in the first retrieval phase have been deleted because they not contain the cluster term, so the cluster term is not useful in a filtering process.
- Other documents retrieved by the IR, that are not relevant, contains synonyms of the cluster term, so they are not deleted and the precision decrease.

The final conclusion is that this filtering process is not good with the cluster term to improved the results. As future work a clustering or classifying method will be developed, working with textual information, to classify and improved the baseline results.

Acknowledgements

This project has been partially supported by a grant from the Spanish Government, project TIMOM (TIN2006-15265-C06-03).

References

1. Arni, T., Clough, P., Sanderson, M., Grubinger, M.: Overview of the ImageCLEFphoto 2008 Photographic Retrieval Task. In: Working Notes of the 2008 CLEF Workshop, Aarhus, Denmark (September 2008)
2. Porter, M.F.: An algorithm for suffix stripping. In Readings in information retrieval, pp. 313–316. Morgan Kaufmann Publishers Inc., San Francisco (1997)
3. Gómez-Soriano, J.M., Montes-y-Gómez, M., Sanchis-Arnal, E., Rosso, P.: A Passage Retrieval System for Multilingual Question Answering. In: Matoušek, V., Mautner, P., Pavelka, T. (eds.) TSD 2005. LNCS (LNAI), vol. 3658, pp. 443–450. Springer, Heidelberg (2005)
4. Díaz-Galiano, M.C., García-Cumbreras, M.A., Martín-Valdivia, M.T., Montejo-Raez, A., Ureña-López, L.A.: Integrating MeSH Ontology to Improve Medical Information Retrieval. In: Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D.W., Peñas, A., Petras, V., Santos, D. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 601–606. Springer, Heidelberg (2008)