SINAI at ImageCLEFPhoto 2008

M.A. García-Cumbreras, M.C. Díaz-Galiano, M.T. Martín-Valdivia, L.A. Ureña-López
SINAI Research Group. Computer Science Department. University of Jaén
{magc,mcdiaz,maite,laurena}@ujaen.es

Abstract

This paper describes the SINAI team participation in the ImagePhoto CLEF campaign. Last years translation approaches and different Information Retrieval systems were tested. In 2008 the imagePhoto task does not include multilingual queries, so translation methods are not necessary.

This year the baseline experiment uses the parameters that obtain the best results in past campaigns. The novelty of our method consists of some filtered methods that are used to improve the results, using the cluster term 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. The filtering process does not work well, because when the cluster term does not appear in a retrieved document, the document erased decrease the final precision.

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 Management**]: Languages—*English*

General Terms

Measurement, Performance, Experimentation

Keywords

1 Introduction

In this paper we describe our approach to the ImagePhoto 2008 evaluation campaign[1] at CLEF, over the IAPR TC-12 Benchmark[1] using the annotation text.

Given a monolingual English query the goal of the ImagePhoto task is to find as many relevant images as possible from an image collection[2].

In 2008 this task takes a different approach to evaluate the image clustering. Given a query the goal is to retrieve a relevant set of images at the top of a ranked list. Text and visual information can be used to improve the retrieval methods, and the main evaluation points are the use of *Pseudo-Relevance Feedback*[3] (PRF), query expansion, IR systems with different weighting functions and clustering or filtering methods applied over the cluster terms.

Our system makes use of text information, nor 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

[1] http://www.iapr.org

has been developed to filter the results with the cluster term and its *WordNet*[2][4] synonyms, in some cases.

2 System description

The SINAI system is automatic (without user interaction), and works with English text information (not visual information). The English collection documents have been preprocessed as usual (English *stopwords* removal and the Porter's *stemmer*[5]). Then, it has been indexed using as IR systems: *Lemur*[3] and *Jirs*[6].

Past campaigns our adhoc system worked with these IR systems, and the precision results obtained were very similar. Only the results with Italian queries were quite different[7, 8]. 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 main objective in 2008 has been to improve the basic case with different combinations of methods and the application of a filter with the cluster term. A similar filtering method is applied in our system that works with geographical information[9]. The weighting function of the IR systems is a parameter that changes to test the results. The use of 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. The blind feedback algorithm is based on the probabilistic term relevance weighting formula developed by Robertson and Sparck Jones[10].

The use of the cluster term has been oriented in a filtering way. After the retrieval process the documents or passages marked as relevant are filtered 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 was to duplicate the score value of the documents in the filtered list and to add them to the original ones.

The figure 1 shows a general scheme of the system developed.

3 Experiments description

The dataset is the collection IAPR TC-12 image collection, that consists of 20,000 images taken from different locations around the world and comprises a varying cross-section of still natural images. It includes pictures of a range of sports and actions, photographs of people, animals, cities, landscapes and many others of contemporary life.

Each image is associated with alphanumeric captions stored in a semi-structured format (title, creation date, location, name of the photographer, description and additional notes). The topics statements are the same of past ImagePhoto campaigns, but only the topic languages in English. A new cluster tag has been added, as appear in the Figure 2.

In our system we have proved different configurations:

1. **(1) SINAIexp1Baseline**. It is the baseline experiment. It uses Lemur 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]available at http://wordnet.princeton.edu
[3]Available at `http://www.lemurproject.org/`

Figure 1: General scheme of the SINAI system at ImagePhoto 2008

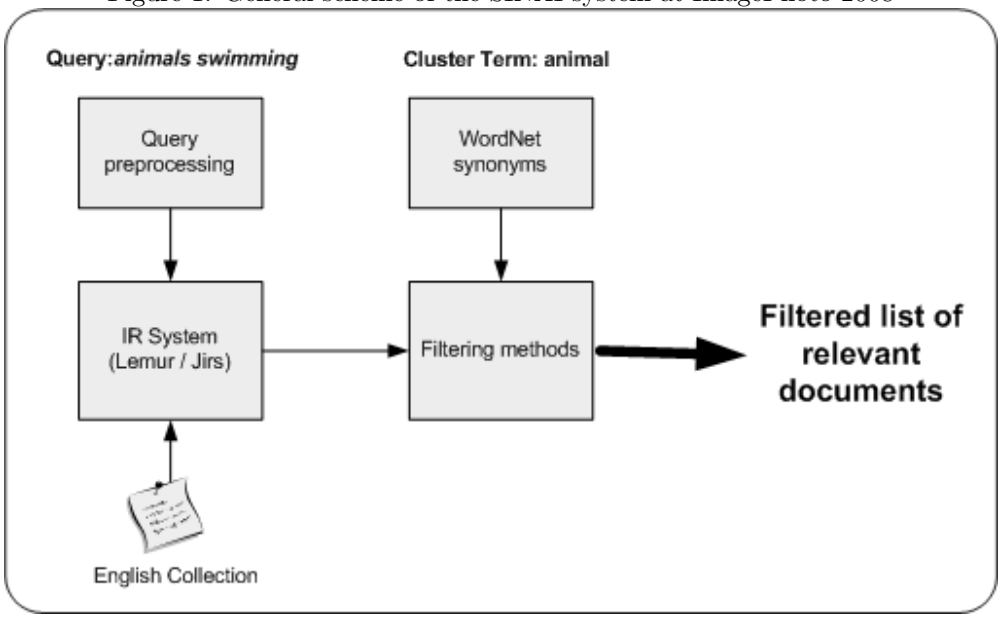

Figure 2: ImagePhoto 2008: query sample

<top>
<num> Number: 5 </num>
<title> animals swimming </title>
<cluster>animal</cluster>
<narr> </narr>
<image> 3739.jpg </image>
<image> 4968.jpg </image>
<image> 30823.jpg </image>
</top>

2. **(2) SINAIexp2LemurJirs**. 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. **(3) SINAIexp3Lemurfb_okapi**. The Lemur list of relevant documents is filtered with the cluster term and its WordNet synonyms. Okapi is used as weighting function, and PRF is applied automatically.

4. **(4) SINAIexp4Lemurfb_tfidf**. It is the same experiment as before, but in this case the weighting function used was Tfidf.

5. **(5) SINAIexp5Lemursimple_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. **(6) SINAIexp6Lemursimple_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: Baseline results without filtering

Id	Language	Modality	FB	Expansion	MAP	P@5	P@10
(1)	EN	Text	Yes	No	**0.2125**	**0.3744**	**0.3308**
(6)	EN	Text	No	No	0.2016	0.3077	0.2872

Table 2: Results with filtering

Id	Language	Modality	FB	Expansion	MAP	P@5	P@10
(2)	EN	Text	Yes	No	0.2063	0.3385	0.2949
(3)	EN	Text	Yes	No	**0.2089**	**0.3538**	0.3128
(4)	EN	Text	Yes	No	0.2043	0.2872	0.2949
(5)	EN	Text	No	No	0.1972	0.3385	**0.3179**

4 Results and Discussion

Our tables of results are organized as follows. Table 1 presents our baseline results, without filtering. Table 2 presents the results with filtering techniques. As we detail in the previous section all the results are based on textual information.

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 one reason is that some relevant documents that appear in the first retrieval phase have been deleted because they not contain the cluster term. For these documents the cluster term is not useful in a filtering process.

On the other hand some documents retrieved by the IR that are not relevant contain synonyms of the cluster term, so they are not deleted and the precision decrease.

5 Conclusions

In this paper we have presented results for the SINAI participation in the ImageCLEF 2008 Photo task. In our work we 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.

As future work we will develop a clustering or classifying method only with textual information.

6 Acknowledgements

This project has been partially supported by a grant from the Spanish Government, project TIMOM (TIN2006-15265-C06-03), and the RFC/PP2006/Id_514 granted by the University of Jaén.

References

[1] Grubinger, M., Clough, P., Hanbury, A., and Müller, H.: Overview of the ImageCLEF 2008 Photographic Retrieval Task. Working Notes of the 2008 CLEF Workshop. Aarhus, Denmark. 2008.

[2] Grubinger, M., Clough, P., Hanbury, A., and Müller, H.: Overview of the ImageCLEF 2007 Photographic Retrieval Task. Working Notes of the 2007 CLEF Workshop. Sep, 2007. Budapest, Hungary.

[3] Salton, G. and G. Buckley:Improving retrieval performance by relevance feedback. Journal of American Society for Information Sciences. 1990.

[4] , Rennie, Jason:WordNet::QueryData: a Perl module for accessing the WordNet database. 2000.

[5] M. F. Porter: An algorithm for suffix stripping. In Readings in information retrieval. ISBN 1-55860-454-5; pages 313-316. Morgan Kaufmann Publishers Inc., 1997.

[6] Gómez-Soriano, J.M., Montes-y-Gómez, M., Sanchis-Arnal, E., and Rosso, P.: A Passage Retrieval System for Multilingual Question Answering. 8th International Conference of Text, Speech and Dialogue 2005 (TSD'05). Lecture Notes in Artificial Intelligence (LNCS/LNAI 3658). pp. 443-450. Karlovy Vary, Czech Republic. 2005.

[7] Díaz-Galiano, M.C., García-Cumbreras, M.A., Martín-Valdivia, M.T., Montejo-Raez,A., and Ureña-López, L.A.: SINAI at ImageCLEF 2006. In Proceedings of the Cross Language Evaluation Forum (CLEF 2006), 2006.

[8] Díaz-Galiano, M.C., García-Cumbreras, M.A., Martín-Valdivia, M.T., Montejo-Raez,A., and Ureña-López, L.A.: SINAI at ImageCLEF 2007. In Proceedings of the Cross Language Evaluation Forum (CLEF 2007), 2007.

[9] Perea-Ortega, J.M, García-Cumbreras, M.A., García-Vega, M. and Montejo-Raez, A.: GEOUJA System. University of Jaén at GEOCLEF 2007. In Proceedings of the Cross Language Evaluation Forum (CLEF 2007), 2007.

[10] S. E. Robertson and K. Sparck Jones: Relevance weighting of search terms. Journal of the American Society for Information Science. 1976.