SINAI at ImagePhoto 2009

M.A. García-Cumbreras, M.C. Díaz-Galiano, A. Montejo-Raez, M.T. Martín-Valdivia
University of Jaén. Computers Department. SINAI Group.
{magc,mcdiaz,amontejo,maite}@ujaen.es

Abstract

This paper presents the fourth participation of the SINAI group, University of Jaén, in the Photo Retrieval task at Image CLEF 2009. Our system uses only the text of the queries, and a clustering system (based on kmeans) that combines different approaches based on a different use of the cluster data of the queries. The official results shown that the combination between the title of each query and the other titles of the clusters obtain our best performance and that our clustering system did not work well.

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

Text retrieval, Indexing, Clustering, PRF

1 Introduction

In this paper we describe our approach to the ImagePhoto 2009 evaluation campaign at CLEF, over the new collection, which contains 498,920 images from Belga News Agency. Given a monolingual English query the goal of the ImagePhoto task is to find as many relevant images as possible from an image collection[1].

In 2008 this task took a different approach to evaluate the image clustering. This year the organizers give special value to the diversity of results. 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-relevant feedback (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, not visual information, to improve the retrieval methods. This year, a new method has been implemented to cluster images.

The following section describes the queries built and the new algorithm of clustering. In Section 3 we explain the experiment carried out. Finally, the obtained results and conclusions are presented in Section 4.

2 System description

In our system we have used an automatic modality, without user interaction, with English text information (not visual information). The English collection documents have been preprocessed as usual (English stopwords removal and the Porter's stemmer[2]). Then, it has been indexed using as Information Retrieval (IR) system Lemur[1].

Past campaigns our adhoc system worked with different IR systems, and test different approaches, such as combination of retrieved lists or fusion of a filtering method that used some clusters terms. The precision results obtained were very similar to different languages[3, 4]. In 2007, 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[5]. In 2008 the use of the cluster term was oriented in a filtering way, and after the retrieval process the documents or passages marked as relevant are filtered[6].

The weighting function of the IR system is a parameter changed to test previous results and based on them we have used Okapi. 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 always used with Lemur. The blind feedback algorithm is based on the probabilistic term relevance weighting formula developed by Robertson and Sparck Jones[7].

The evolution of our system introduces a clustering module, based on the algorithm kmeans, and the creation of the final topics using the information of the title and the clusters terms.

2.1 Building the queries

The first module of our system uses the data of the title and the cluster title, in some cases, and combines them to obtain the final topics. These topics, in a second step, are run over the IR system, and a final list of relevant documents is obtained.

The following four sets of topics have been built for these experimentation:

1. In the baseline case only the title of the query is used against the index of the IR system.
2. Each final topic is a combination between the query title and the title of the last cluster.
3. Each final topic is a combination between the query title and the other words of each cluster.

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

2.2 Clustering subsystem

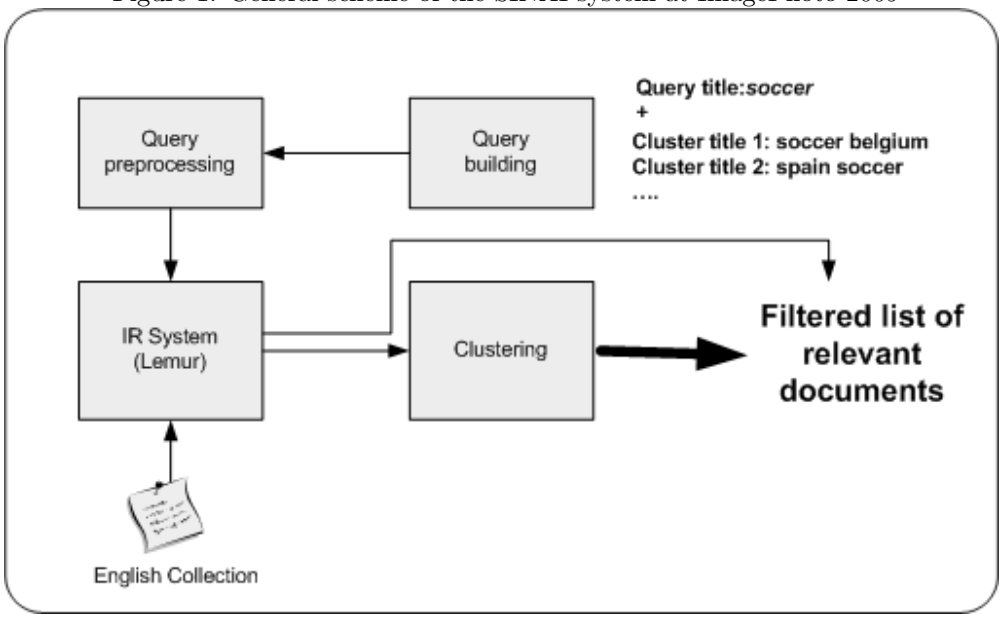
It has been found that the variability of top results in a list of documents retrieved as answer to a query, the performance of the retrieval systems increases too, being in some cases more desirable to have less but more varied items in this list [8]. In order to increase variability, a clustering system has been applied. This has also been used in other systems with the same aim [9]. The idea behind is rather simple: re-arrange most relevant documents so that documents belonging to different clusters are promoted to the top of the list.

We have applied kmeans on each list returned by the Lemur IR system. For this, Rapid Miner tool was used[2]. The clustering algorithm has tried to group these results, without any concern on ranking, into 4 different groups. The number of groups has been stablished at this value as documents in the training set have this average number of clusters specified in their metadata.

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

[2] Available at http://rapid-i.com

Figure 1: General scheme of the SINAI system at ImagePhoto 2009



Once each of the documents in the list has been labeled to its resulting cluster index, the list has been reordered according to the described principle: we fill the list by alternating documents from different clusters.

3 Experiments description

The data set contains 498,920 images from Belga News Agency, an image search engine for news photographs. Each photograph will be up to a maximum of 512 pixels in either width or height, accompanied by a caption composed of English text up to a few sentences in length. Different to the data last year, captions are provided without a specific format to increase the challenges to participants. Caption might contain date and place where image was captured.

The topics statements include a query title with few words, and some cluster titles and descriptions, to promote the diversity of the final results, as appear in the Figure 2. In our system descriptions are not used.

Figure 2: ImagePhoto 2009: query sample

<top>
<num> Number: 0 </num>
<title> **soccer** </title>
<clusterTitle> **soccer belgium** </clusterTitle>
<clusterDesc> Relevant images contain photographs of the Belgium team in a soccer match. </clusterDesc>
<image> belga38/00704995.jpg </image>
<clusterTitle> **spain soccer** </clusterTitle>
<clusterDesc> Relevant images contain photographs of the Spain team in a soccer match. </clusterDesc>
<image> belga6/00110574.jpg </image>
<clusterTitle> **beach soccer** </clusterTitle>
<clusterDesc> Relevant images contain photographs of a soccer beach match. </clusterDesc>
<image> belga33/06278068.jpg </image>
<clusterTitle> **italy soccer** </clusterTitle>
<clusterDesc> Relevant images contain photographs of the Italy team in a soccer match. </clusterDesc>
<image> belga20/1027435.jpg </image>
<clusterTitle> **soccer netherlands** </clusterTitle>
<clusterDesc> Relevant images contain photographs of the Netherlands team in a soccer match or the teams in Netherlands' league. </clusterDesc>
<image> belga10/01214810.jpg </image>
<clusterTitle> **soccer -belgium -spain -beach -italy -netherlands** </clusterTitle>
<clusterDesc> Relevant images contain photographs of any aspects or subtopics of soccer which are not related to the above clusters.
</clusterDesc>
<image> belga20/01404831.jpg </image>

In our system we have proved the following configurations:

1. **(1) SINAI1 - Baseline**. It is the baseline experiment. It uses Lemur as IR system with automatic feedback. The weighting function applied was Okapi. The topic used is only the query title.

2. **(2) SINAI2 - title and final cluster**. This experiment combines the query title with the title of the final cluster that appear in the topics file. Lemur also uses Okapi as weighting function and PRF.

3. **(3) SINAI3 - title and all clusters**. This experiment combines the query title with all the words that appear in the titles of all the clusters. Lemur also uses Okapi as weighting function and PRF.

4. **(4) SINAI4 - clustering**. The query title and each cluster title (except the last one that combines all) are run against the index generated by the IR system. Several lists of relevant documents are retrieved, and the clustering module combines them to obtain the final list of relevant documents. The aim of this experiment is to increment the diversity of the retrieved results using a clustering algorithm.

4 Results and Discussion

The data set of the collection has been indexed using Lemur[3] IR system, by applying Okapi weighing function and using Pseudo-Relevance Feedback (PRF). We have used only textual information in English. Table 1 shows the results obtained in our four experiments submitted this year.

Experiment	Rank	CR10	P10	MAP	F-measure
sinai1_T_TXT	58	0.4580	0.796	0.4454	0.5814
sinai2_TCT_TXT	70	0.3798	58	0.3286	0.4590
sinai3_TCT_TXT	**44**	**0.5210**	**0.778**	**0.4567**	**0.6241**
sinai4_TCT_TXT	72	0.4356	0.474	0.2233	0.4540

Table 1: SINAI experiment results

We have experimented with different kinds of cluster combination. However, as we can see in Table 1, the application of clustering does not improve the results greatly. In fact, only in the run used SINAI3 the query the original title and the titles of all the clusters overcomes the baseline case SINAI1 that only uses the original title. Unfortunately, the experiment SINAI4 that applies our clustering and fusion approach has achieved the worst results. Thus, the obtained results show that it is necessary to continue investigating the clustering methodology. In addition, the use of visual information could improve the final system.

5 Acknowledgements

This work has been supported by the Regional Government of Andaluca (Spain) under excellence project GeOasis (P08-41999), the Spanish Government under project Text-Mess TIMOM (TIN2006-15265-C06-03) and the local project RFC/PP2008/UJA-08- 16-14.

References

[1] Paramita, M., Sanderson, M. and Clough, P. Diversity in photo retrieval: overview of the ImageCLEFPhoto task 2009. CLEF working notes 2009, Corfu, Greece, 2009.

[3] http://www.lemurproject.org/

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] Díaz-Galiano, M.C., García-Cumbreras, M.A., Martín-Valdivia, M.T. and Ureña-López, L.A.: SINAI at ImageCLEF 2008. In Proceedings of the Cross Language Evaluation Forum (CLEF 2008), 2008.

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

[8] Chen, H., Karger, D.R.: Less is more: probabilistic models for retrieving fewer relevant documents. SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. Seattle, Washington, 2006.

[9] Ah-Pine, J., Bressan, M., Clinchant, S., Csurka, G., Hoppenot, Y., Renders, J.M.: Crossing textual and visual content in different application scenarios. Multimedia Tools Appl., pages 31–56. Volume 41, number 1. Kluwer Academic Publishers, ISSN 1380-7501.

[10] Chevallet, J.P., Lim, J.H., and Radhouani, S.: Using Ontology Dimensions and Negative Expansion to solve Precise Queries in CLEF Medical Task. Working Notes of the 2005 CLEF Workshop. Sep, 2005. Vienna, Austria.