Text-mess in the ImageCLEFphoto08 Task

S. Navarro[1], M.A. García[4], F. Llopis[3], M.C. Díaz[2], R. Muñoz[5],
M.T. Martín[6], L.A. Ureña[7], A. Montejo[8]
[1,3,5]Natural Language Processing and Information Systems Group. University of Alicante, Spain
[2,4,6,7,8]Sistemas Inteligentes de Acceso a la Información, SINAI Group. University of Jaén, Spain
{snavarro[1],llopis[3],rafael[5]}@dlsi.ua.es
{mcdiaz[2],magc[4],maite[6],laurena[7],amontejo[8]}@ujaen.es

Abstract

This paper describes our participation in the ImagePhoto task at CLEF 2008. We present the joint work of two teams belonging to the TEXT-MESS project using a new system that combines the individual systems of these teams, one based on filtering and the other one based on clustering. We have submited experiments using SINAI filtering method with the IR-n output, and the IR-n clustering module with the SINAI output. Our objective was to study the behaviour of these methods with a large number of configurations in order to increase our chances of success. The results show that a filtering method is not useful when we use the cluster terms or related words to filter retrieved documents, and that a clustering method can improve the results of cluster detection although at the expense of a decrease in precision of the results that is greater than the gain obtained for the CR20 measure with this method.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.2 Information Storage H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries; H.2 [**Database Managment**]: H.2.5 Heterogenous Databases

General Terms

Measurement, Performance, Experimentation

Keywords

Information Retrieval, Image Retrieval, Multimodal Reranking, Late Fusion, Intermedia Pseudo Relevance Feedback, Multimodal Relevance Feedback, PRF, LCA, Visual Concepts

1 Introduction

Given a monolingual English query and related topic images the goal of the ImagePhoto task is to find as many relevant images as possible from an image collection.

In 2008 this task takes a different approach to evaluate the image clustering. Given a query the goal is to retrieve a diverse, yet 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* (PRF) [6] and *Local context Analisy* (LCA) [7] in a textual and multimodal way, IR systems with different weighting functions and clustering or filtering methods applied over the cluster terms.

We have focused our participation, on analysing the effect of using SINAI filtering method with the IR-n output, and IR-n clustering module with the SINAI output. In order, to analyse the behaviour of these methods with a broader number of configurations, in order to improve the performance of the systems and to increase our chances of success.

This paper is structured as follows: Firstly, it presents the main characteristics of the SINAI and IR-n system focusing on their filtering and clustering strategies respectively, then it moves on to explain the experiments we have made to evaluate the system, and finally it describes the results and conclusions.

2 The System

The complete system is composed by two main modules that work in a serial mode. The output of the first module is the input of the second one.

2.1 The SINAI System

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*[1] and *Jirs*[1].

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[4]. 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[6].

2.1.1 Filtering Method

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.

2.2 The IR-n System

IR-n is an information retrieval system based on passages. Those type of IR systems, unlike document-based systems, can consider the proximity of words with each other, that appear in a document in order to evaluate their relevance [2].

This system has added for its current participation in this task three common approaches to the multimodal issue.

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

On the one hand, it allows to use automatic relevance feedback in a multimodal way - feeding it with the top documents retrieved from a CBIR system - and has added another relevance feedback technique - LCA strategy - as alternative to PRF in an attempt of skipping the great number of non relevant documents top ranked in a CBIR list which are used for the relevance feedback [3].

On the other hand, the system has added a multimodal reranking module. It allows two working modes. The first one is the standard one for merging two lists, based on set values to the weighting factor of each list in order to create a joined list. The second one is the TF-IDF multimodal reranking, it is a variation of the standard one. It bases the calculus of the relevance of an image on the quantity and the quality of its annotations in order to decide if the relevance value returned by the textual IR system is enough to rank a document or if it is needed to add the relevance returned by the CBIR system[3].

Finally, it has a module for enrich the documents with visual concepts, and a clustering module in order to improve the recall of the different image cluster detected within the 20 top ranked documents. This clustering module is based on the image annotations and the visual concepts related to each image [3].

2.2.1 Clustering Module

Usually, when the users performs a query in a IR system that works with images, they find that there are several similar images between top 20 results. Thus, if they want to find different relevant images they have to navigate through the rest of the returned list.

Our approach is a naive attempt to solve this problem using Carrot2 [2] an open source clustering engine for text. Our clustering module use as input for Carrot2 the query and the documents of the ranking list - which optionally can be enriched with its related visual concepts -. It uses the texts to perform the clustering. For each cluster returned, the image with the best relevance in the ranking list is selected. If there are a number of clusters lower than twenty, the other images are selected between those images without cluster assigned that has been better ranked by the system until complete the selection of twenty images. Afterwards, the module adds to the relevance value of the selected images the maximum relevance value in the whole list, in order to take up this images to the top 20 positions in the ranking.

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 also have a semi-structured format, this year are the same of past ImagePhoto campaigns, but only the topic languages in English. Two new tags have been added this year. The cluster tag and the narrative tag.

We have used for our experiments the following configurations of the SINAI system:

- **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.
- **Lemur fb 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.

[2] Available at `http://www.carrot2.org/`

- **Lemur fb tfidf**: It is the same experiment as before, but in this case the weighting function used was Tfidf.
- **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.
- **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.

Next we can see the configurations used for the IRn system experiments -All the experiment uses DFR as weighting schema and a passage size of 4 sentences, moreover none of the runs uses neither the narrative of the topic nor the concept -:

- **IRnExp**: This experiment uses PRF as relevance feedback strategy.
- **IRnExpClust**: This experiment uses PRF as relevance feedback strategy and clustering based on the image annotations
- **IRnFBFIRE**: It uses a baseline experiment of the FIRE system and LCA as multimodal relevance feedback strategy.
- **IRnFBFIREClustC**: It first uses a a baseline experiment of the FIRE system and LCA as multimodal relevance feedback strategy. Afterwards it uses visual concepts extracted from the images to enrich the returned image annotations in order to use it as input for the clustering module.
- **IRnConcepFBFIRE**: The image annotations indexed by IR-n are previously enriched with visual concepts extracted from the image. For the retrieval phase, the system uses a baseline run of the FIRE system and LCA as multimodal relevance feedback strategy.
- **IRnConcepFBFIREClustC**: The image annotations indexed by IR-n are previously enriched with visual concepts extracted from the image. For the retrieval phase, the system uses a baseline run of the FIRE system and LCA as multimodal relevance feedback strategy. Afterwards it uses the returned image annotations enriched with visual concepts in order to use it as input for the clustering module.

We have focused our participation, on analysing the effect of using SINAI filtering method and IR-n clustering module in order to improve the performance of the systems. Thus, we have added the filtering method to the work flow of the IR-n system - using filtering before or after the clustering phase -, and vice versa, - using the IR-n clustering module to process the output of the SINAI system -.

4 Results in ImageCLEFphoto08

In the Table 1 and Table 2 we can see the textual runs and the mixed runs respectively. They show the official results obtained by each run - Official MAP, P20, CR20 and F-Mean - using filtering - with IR-n - and clustering - with SINAI system -. Furthermore, we can see the results previously obtained by the standalone runs - without adding the external filtering or clustering module -, in order to observe the improvement or worsening obtained with the added module. For each run name we show a term in bold letters which identifies the external module which has been added to that base configuration - **Filt** or **Clust** -.

We can observe in the Table 1 that the CR20 value has increased its value for almost all the experiments which have used the clustering module. Indeed the best CR20 value for the textual runs has been obtained using the clustering module with SINAI system.

Table 1: Textual Results in ImageCLEFphoto08

	Standalone Run				Official Run			
run name	**MAP**	**P20**	**CR20**	**FMea**	**MAP**	**P20**	**CR20**	**FMea**
IRnExp**Filt**	0.2699	**0.3244**	0.2816	0.3015	**0.2671**	0.3154	0.2875	0.3008
IRnExpClust**Filt**	0.2699	**0.3244**	0.2816	0.3015	0.2287	0.2090	0.3011	0.2467
LemurSimpleOkapiFilt**Clust**	0.1972	0.2795	0.2930	0.2861	0.1750	0.1987	**0.3241**	0.2464
LemurFbOkapiFilt**Clust**	0.2089	0.2808	0.2682	0.2744	0.1804	0.1897	0.2764	0.2250
LemurJirs**Clust**	0.2063	0.2769	0.2900	0.2833	0.1840	0.2051	0.2815	0.2373
LemurFbTfidfFilt**Clust**	0.2043	0.2679	0.2704	0.2691	0.1786	0.1974	0.3185	0.2437

Table 2: Mixed Results in ImageCLEFphoto08

	Without Filtering or Clustering				With Filtering or Clustering			
run name	**Of. MAP**	**P20**	**CR20**	**FMea**	**Of. MAP**	**Of. P20**	**Of. CR20**	**FMea**
IRnFBFIRE**Filt**	**0.3436**	**0.4564**	0.3119	0.3706	0.3354	0.4333	0.3041	0.3574
IRnFBFIREClustC**Filt**	0.3032	0.3782	0.3483	0.3626	0.3183	0.3808	0.3178	0.3465
IRnFBFIRE**Filt**ClustC	0.3032	0.3782	0.3483	0.3626	0.3097	0.3564	0.3223	0.3385
IRnConcepFBFIRE**Filt**	0.3333	0.4333	0.3316	**0.3757**	0.3272	0.4115	0.3311	0.3669
IRnConcepFBFIRE**Filt**ClustC	0.3032	0.3782	0.3483	0.3626	0.2917	0.3410	**0.3483**	0.3446
IRnConcepFBFIREClustC**Filt**	0.3032	0.3782	0.3483	0.3626	0.2973	0.3603	0.3446	0.3523

5 Conclusion and Future Work

In this paper we have presented results for the Text-mess participation in the ImageCLEF 2008 Photo task. In our work we have focused our efforts on analysing the effect of using SINAI filtering method based on the used the cluster term and IR-n clustering module - not based on the clustering term -.

On one hand the results shows that a filtering method is not useful when we use the cluster term or related words to filter retrieved documents, because some relevant documents are deleted and none of non retrieved relevant documents are included in the second step. On the other hand the clustering method without using the cluster term, has showed that it can improve the results of cluster detection, although at the expense of a decrease in precision of the results that is greater than the gain obtained for the CR20.

As future work we will develop clustering or classifying method with textual information. This method should take in account the cluster term of the topic in order to select which annotation tags of the topic are more useful for the clustering phase.

6 Acknowledgement

This research has been partially funded by the Spanish Government within the framework of the TEXT-MESS (TIN-2006-15265-C06-01) project.

References

[1] J.M. Gómez-Soriano, M. Montes-y Gómez, E. Sanchis-Arnal, and P. Rosso. A passage retrieval system for multilingual question answering. 2005.

[2] Fernando Llopis. *IR-n: Un Sistema de Recuperacin de Informacin Basado en Pasajes*. PhD thesis, University of Alicante, 2003.

[3] Sergio Navarro, Fernando Llopis, and Rafael Muñoz. Different Multimodal Approaches using IR-n in ImageCLEFphoto 2008. In *In on-line Working Notes, CLEF 2008*, 2008.

[4] García-Cumbreras Perea-Ortega, J.M, García-Vega M.A., M., and A. Montejo-Raez. Geouja system. university of jaén at geoclef 2007. 2007.

[5] M. F. Porter. An algorithm for suffix stripping. 1997.

[6] S. E. Robertson and K. Sparck Jones. Relevance weighting of search terms. 1976.

[7] Jinxi Xu and W. Bruce Croft. Improving the effectiveness of information retrieval with local context analysis. *ACM Trans. Inf. Syst.*, 18(1):79–112, 2000.