MIRACLE-GSI at ImageCLEFphoto 2009: Comparing Clustering vs. Classification for Result Reranking

Julio Villena-Román[1,3], Sara Lana-Serrano[2,3], José C. González-Cristóbal[2,3]

[1] Universidad Carlos III de Madrid
[2] Universidad Politécnica de Madrid
[3] DAEDALUS - Data, Decisions and Language, S.A.

jvillena@it.uc3m.es, slana@diatel.upm.es, josecarlos.gonzalez@upm.es

Abstract

This paper describes the participation of MIRACLE-GSI research consortium at the ImageCLEF 2009 Photo Retrieval Task. For this campaign, the main purpose of our experiments was to compare the performance of a “standard” clustering algorithm, based on the k-Medoids algorithm, against a more simple classification technique that makes use of the cluster assignment that was provided for a subset of topics by the task organizers. First a common baseline algorithm was used in all experiments to process the document collection: text extraction, tokenization, conversion to lowercase, filtering, stemming and finally, indexing and retrieval. Then this baseline algorithm is combined with these two different result reranking techniques. As expected, results show that any reranking method outperforms a standard non-clustering image search baseline algorithm in terms of cluster recall. In addition, using the information of cluster assignments leads to the best results.

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 Management]**: H.2.5 Heterogeneous Databases; **E.2 [Data Storage Representations]**.

Keywords

Image retrieval, domain-specific vocabulary, thesaurus, linguistic engineering, information retrieval, indexing, relevance feedback, topic expansion, ImageCLEF Photo Retrieval task, ImageCLEF, CLEF, 2009.

1. Introduction

MIRACLE is a research consortium formed by research groups of three different universities in Madrid (Universidad Politécnica de Madrid, Universidad Autónoma de Madrid and Universidad Carlos III de Madrid) along with DAEDALUS, a small/medium size enterprise (SME) founded in 1998 as a spin-off of two of these groups and a leading company in the field of linguistic technologies in Spain. MIRACLE has taken part in CLEF since 2003 in many different tracks and tasks.

The basic goal of the ImageCLEF 2009 Photo Retrieval task [1] was, similar to previous campaigns, given a multilingual statement describing a user specific information need, find as many relevant images as possible from a given multilingual document collections containing images and text. However, the task introduces a different approach to evaluation by studying image clustering. The idea is that the top results for the given topics must contain diverse items representing different subtopics within the results. This is because a search engine that retrieves a diverse, yet relevant set of images at the top of a ranked list is supposed to be more likely to satisfy its users.

Participants are provided with a set of topics, which are run on their image search system to produce a ranking that in the top 20, holds as many relevant images that are representative of the different subtopics within the results. Evaluation is based on two measures: precision at 20 and instance recall at rank 20 (also called S-recall), which calculates the percentage of different clusters represented in the top 20. This campaign a new data set containing half a million images was used.

MIRACLE team decided to split into two subgroups, MIRACLE-GSI (Grupo de Sistemas Inteligentes – Intelligent System Group) in charge of purely textual runs and MIRACLE-FI (Facultad de Informática, Computer Science Faculty) in charge of visual and mixed runs. This paper reviews the participation of MIRACLE-GSI at ImageCLEFphoto 2009. The participation of the other subgroup is described in an accompanying paper.

Our idea for this campaign was to continue the open line of research [2] [3] in clustering techniques applied to result reranking. The main purpose of our experiments was to compare the performance of a "standard" clustering algorithm, based on the k-Medoids algorithm [4], against a more simple classification technique that makes use of the cluster assignment that was provided for a subset of topics by the task organizers. All experiments were fully automatic, with no manual intervention, and are described in the following sections.

2. Experiments

Based on our experience in previous campaigns, we designed a flexible system in order to be able to execute a large number of runs that exhaustively many combinations of different techniques. Our system is composed of a set of small components that are easily combined in different configurations and executed sequentially to build the final result set. Specifically, our system is composed of five modules:

- **Linguistic processing module**, which extract, parses and prepares the input text for subsequent modules,
- **Expander module**, which expands documents and/or topics with additional related terms using textual and/or statistical methods,
- **Textual (text-based) retrieval module**, which indexes image annotations in order to search and find the list of images that are most relevant to the text of the topic,
- **Result combination module**, which uses OR/AND operators to combine, if necessary, two different result lists,
- **Clustering module**, which reranks the result list to allow cluster diversity**.**

Figure 1 shows an overview of the system architecture.

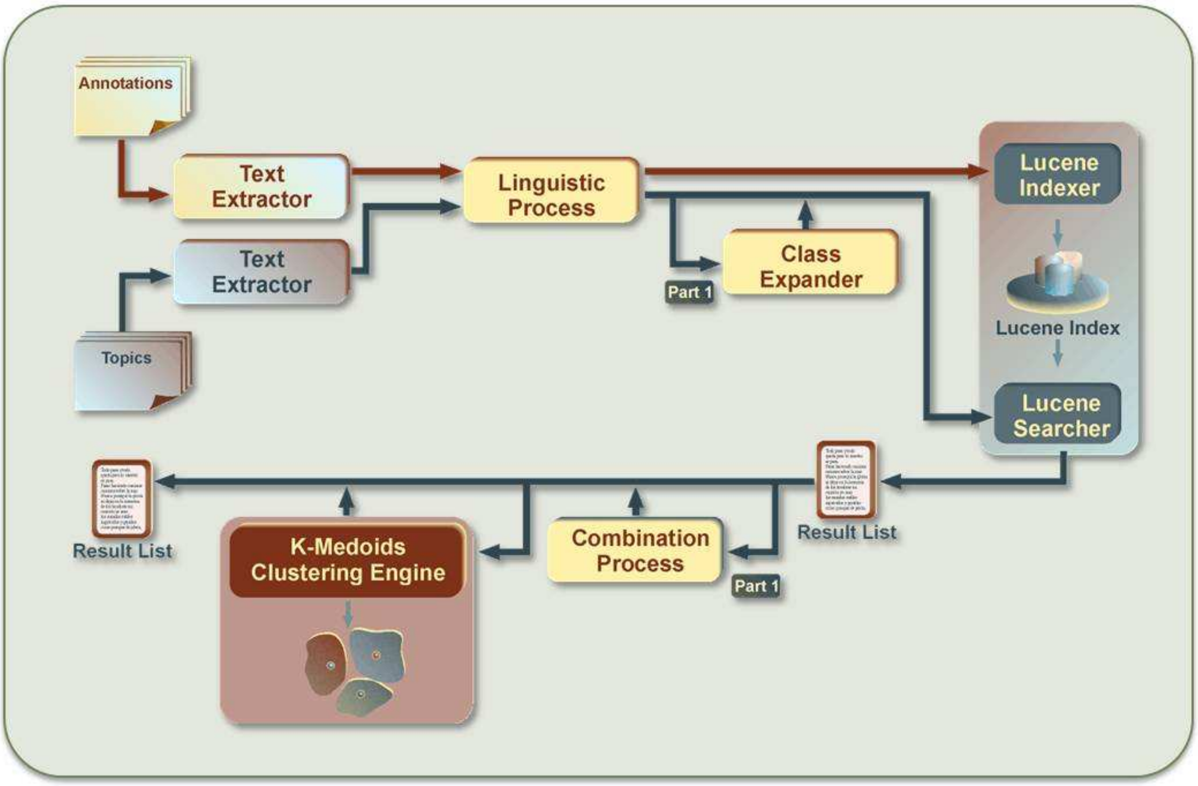

Figure 1. Overview of the system.

A common baseline algorithm was used in all experiments to process the collection, following these steps:

1. **Text Extraction:** Ad-hoc scripts are run on the files that contain image annotations in XML format.
2. **Tokenization:** This process extracts basic textual components. Some basic entities are also detected, such as numbers, initials, abbreviations, and years. So far, compounds, proper nouns, acronyms or other types of entity are not specifically considered. The outcomes of this process are only single words, years in numbers and tagged entities.
3. **Conversion to lowercase:** All document terms are normalized by changing all letters to lowercase.
4. **Filtering:** All words recognized as stopwords are filtered out. Stopwords in the target languages were initially obtained from the University of Neuchatel's resources page [5] and afterwards extended using our own developed resources [2].
5. **Stemming:** This process is applied to each one of the words to be indexed or used for retrieval. Standard Porter stemmers [6] for each considered language have been used.
6. **Indexing and retrieval:** Lucene [7] was used as the information retrieval engine for the whole textual indexing and retrieval task.

The topic set was divided into two subgroups. The first 25 topics include a cluster assignment provided by the task organizers, i.e., some clues were given to guide the clustering process. For those topics, classification techniques can be used to produce the final result list. The rest of the topics did not include any cluster assignment, so "standard" clustering techniques had to be used to produce the final result list.

On the one hand, the classification technique finds, for each topic, the list of images that are relevant to each given cluster, and, in addition, the list of images that are relevant to the topic but do not match any of the given clusters ("Others" cluster). For that purpose, the algorithm first builds as many subtopics as different clusters have been provided for a given topic. These subtopics contain the original topic terms combined with the terms of the cluster titles. For instance, if topic A has 2 clusters associated (A1, A2), the set of subtopics would be:

{terms_A AND terms_{A1}} and {terms_A AND terms_{AB}}

Second, the algorithm builds another subtopic that includes the topic terms but excludes the terms of all clusters. For the previous example, the subtopic for "Others" cluster would be:

{terms_A AND NOT terms_{A1} AND NOT terms_{A2}}

Then those subtopics are given to the Lucene information retrieval engine to get the relevant list of images. Last, each image is assigned to the cluster that corresponds to the subtopic with which the image has the highest similarity.

On the other hand, the clustering technique is based on an implementation of k-Medoids clustering algorithm [4], with k (the target number of clusters) equal to 20 and the maximum number of epochs set to 40. This algorithm is run over a sparse term-document matrix built with the image annotations that are given as results of a textual search over the image index using each topic. For each resulting cluster, the element with higher relevance in the baseline image result list is selected as the class prototype, and reranked to the top of the final result list.

3. Results and Conclusions

Table 1 shows the complete list of submitted runs along with a brief description.

Table 1. Description of experiments

Run Identifier	Method
MIRGSI1_T_TXT	no clustering (baseline)
MIRGSI2_T_TXT	k-Medoids clustering
MIRGSI_TCT_TXT	classification with topic+cluster titles [only for topics 1-25]

Results are presented in the following tables, showing the run identifier, the number of relevant documents retrieved, the mean average precision (MAP), precision at 10, 20 and 30 first results, and cluster precision at 10, 20 and 30 first results. Table 2 shows the results for the first 25 topics and Table 3 shows the results for the remaining topics.

Table 2. Results for queries 1-25

	RelRet	MAP	P10	P20	P30	CR10	CR20	CR30
MIRGSI1_T_TXT	8777	**0.484**	**0.796**	**0.790**	**0.799**	0.417	0.531	0.600
MIRGSI2_T_TXT	8777	0.477	0.784	0.750	0.760	0.455	0.617	0.634
MIRGSI_TCT_TXT	8795	0.481	0.776	0.770	0.773	**0.643**	**0.755**	**0.782**

Table 3. Results for queries 26-50

	RelRet	MAP	P10	P20	P30	CR10	CR20	CR30
MIRGSI1_T_TXT	9596	**0.514**	**0.760**	**0.742**	0.740	0.510	0.609	0.667
MIRGSI2_T_TXT	9596	0.502	0.740	0.732	**0.745**	**0.565**	**0.668**	**0.682**

The following figures show the precision and cluster recall values for each run and allow comparing results achieved by the classification technique (MIRGSI_TCT_TXT) with respect to the clustering technique (MIRGSI2_T_TXT) in each subset of topics. Data series identified with "part 1" refer to topics 1-25 whereas data series identified with "part 2" refer to topics 26-50.

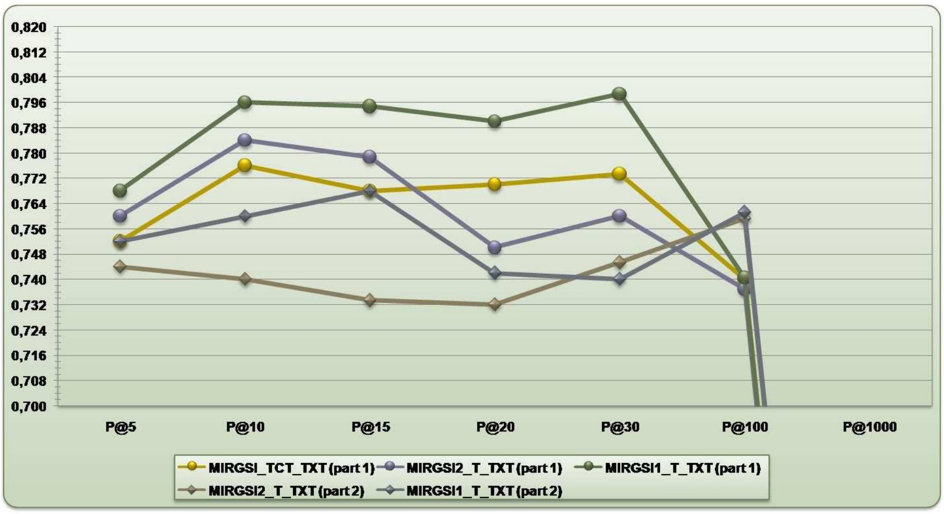

Figure 2. Precision at N, for all runs

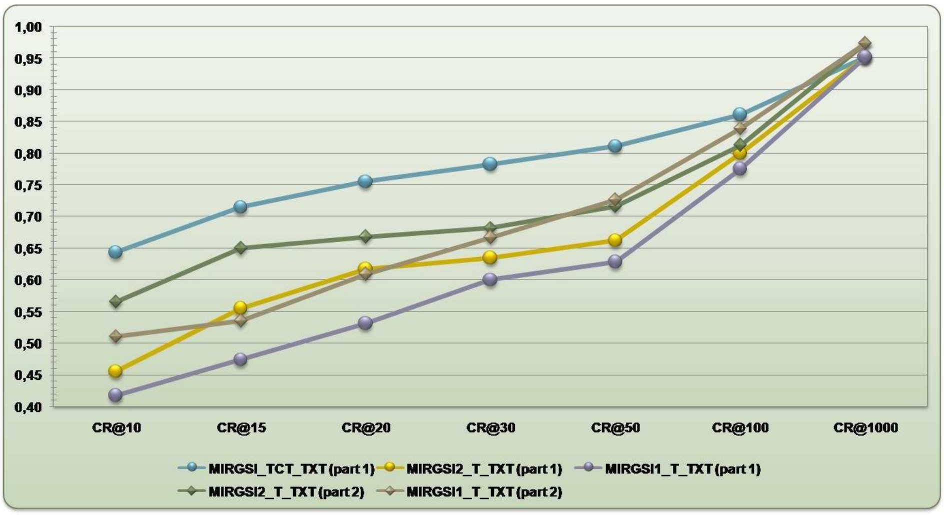

Figure 3. Cluster recall at N, for all runs

The baseline experiment achieves the best result in terms of MAP. However, the best cluster recall (CR), which was the variable to maximize in this task, is achieved when other techniques are used, thus proving to be valuable. As it could be expected, the run that makes use of the manually assigned clusters (MIRGSI_TCT_TXT) achieves the best results in terms of cluster recall, and clearly outperforms the baseline experiment (0.782 vs 0.600 at CR30, 130%). Morever, the k-Medoid clustering is slightly better than the baseline experiment in cluster recall at any value.

After this preliminary analysis, the conclusion that can be drawn is that the application of clustering techniques improves the information retrieval process and shows quite promising results.

Acknowledgements

This work has been partially supported by the Spanish R+D National Plan, by means of the project BRAVO (Multilingual and Multimodal Answers Advanced Search – Information Retrieval), TIN2007-67407-C03-03 and by Madrid R+D Regional Plan, by means of the project MAVIR (Enhancing the Access and the Visibility of Networked Multilingual Information for the Community of Madrid), S-0505/TIC/000267.

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.

2. Lana-Serrano, Sara; Villena-Román, Julio; González-Cristóbal, José Carlos. MIRACLE-GSI at ImageCLEFphoto 2008: Experiments on Semantic and Statistical Topic Expansion. Evaluating Systems for Multilingual and Multimodal Information Access 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers. Peters, Carol et al (Eds.). Lecture Notes in Computer Science, 2008 (printed in 2009).

3. Villena-Román, Julio; Lana-Serrano, Sara; Martínez-Fernández, José Luis; González-Cristóbal, José Carlos. MIRACLE at ImageCLEFphoto 2007: Evaluation of Merging Strategies for Multilingual and Multimedia Information Retrieval. Advances in Multilingual and Multimodal Information Retrieval. 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, Revised Selected Papers. Carol Peters et al (Eds.). Lecture Notes in Computer Science, Vol. 5152, 2008. ISSN: 0302-9743/1611-3349.

4. Park, Hae-sang; Lee, Jong-seok; Jun, Chi-hyuck. A K-means-like Algorithm for K-medoids Clustering and Its Performance. Proceedings of the 36th CIE Conference on Computers & In-dustrial Engineering, pp.1222-1231, Taipei, Taiwan, Jun. 20-23 (2006).

5. University of Neuchatel. Page of resources for CLEF (Stopwords, transliteration, stemmers …). On line http://www.unine.ch/info/clef [Visited 10/08/2008].

6. Porter, Martin. Snowball stemmers and resources page. On line http://www.snowball.tartarus.org [Visited 10/08/2008].

7. Apache Lucene project. http://lucene.apache.org [Visited 09/11/2008].