MIRACLE-GSI at ImageCLEFphoto 2008: Experiments on Semantic and Statistical Topic Expansion

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 ImageCLEFphoto task of ImageCLEF 2008. For this campaign, the main purpose of our experiments was to evaluate different strategies for topic expansion in a pure textual retrieval context. Two approaches were used: methods based on linguistic information such as thesauri, and statistical methods that use term frequency. 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 different expansion techniques. For the semantic expansion, we used WordNet to expand topic terms with related terms. The statistical method consisted of expanding the topics using Agrawal's apriori algorithm. Relevance-feedback techniques were also used. Last, the result list is reranked using an implementation of k-Medoids clustering algorithm with the target number of clusters set to 20. 14 fully-automatic runs were finally submitted. In general, results are on the average, comparing to other groups.

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 Photographical Retrieval task, ImageCLEF, CLEF, 2008.

1. Introduction

MIRACLE team 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, including the main bilingual, monolingual and cross lingual tasks as well as in ImageCLEF [13] [9], Question Answering, WebCLEF, GeoCLEF and VideoCLEF (VID2RSS) tracks.

To simplify our internal coordination, MIRACLE team decided to split for this task into two subgroups, MIRACLE-GSI (Grupo de Sistemas Inteligentes – Intelligent System Group) in charge of purely textual experiments, and MIRACLE-FI (Facultad de Informática, Computer Science Faculty) in charge of visual and mixed runs. This paper describes the participation of MIRACLE-GSI at the ImageCLEF Photographic Retrieval task of ImageCLEF 2008. The participation of the other subgroup is described in an accompanying paper.

The basic goal of the task [7] is, 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. This campaign the task introduced 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, reused from the previous campaigns, 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 be 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. The reference database for this campaign is the same as last year, IAPR TC-12 Benchmark [6]. This collection contains 20,000 photos (mainly colour photographs) taken from locations around the world and comprises a varying cross-section of still natural images, annotated with captions in English and German.

For this campaign, the main purpose of our experiments was to compare among different strategies for topic expansion in a pure textual context. Two approaches were used: methods based on linguistic information such as thesauri, and statistical methods that use term frequency. We also participated in the ImageCLEF Medical Retrieval task with the same approach, which allows for comparison between two different domains. All experiments were fully automatic, with no manual intervention. Finally 14 runs were submitted, as described next.

2. Description of the System

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 cover all the combinations of the 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 different modules: the textual (text-based) retrieval module, which indexes image annotations in order to search and find the most relevant ones to the text of the topic; the expander module, which expands documents and/or topics with additional related terms using textual and/or statistical methods; the relevance-feedback module, which allows to execute reformulated queries that include the results of previous queries; the result combination module, which uses OR operator to combine, if necessary, the results of the previous subsystems; and, finally, a clustering module that reranks the result list to allow cluster diversity. Figure 1 shows an overview of the system architecture.

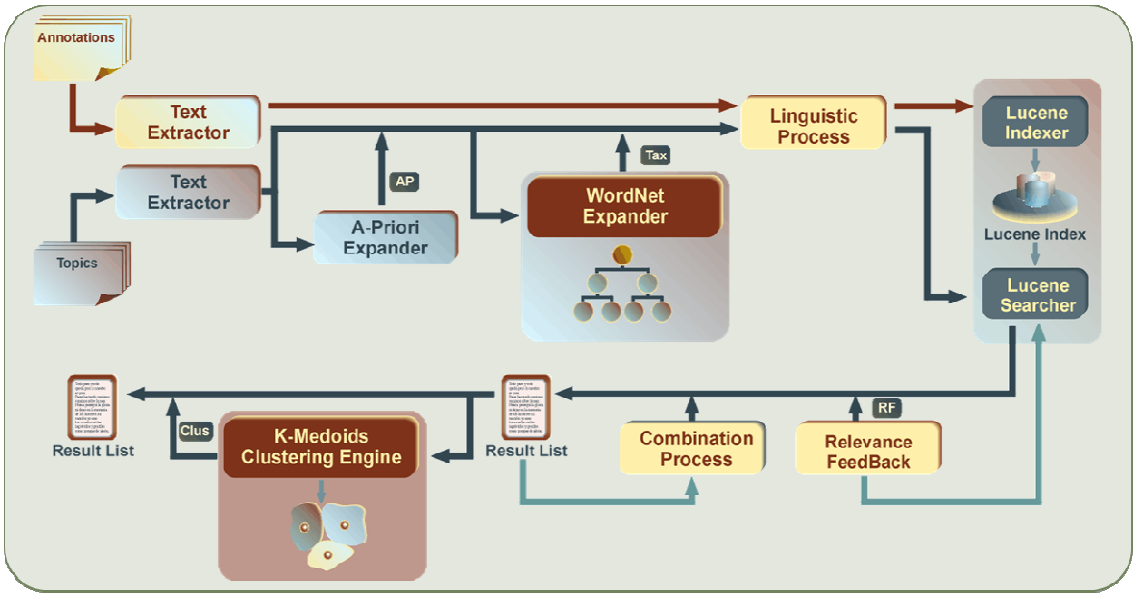

Figure 1. Overview of the system.

The system consists of a set of different basic components organized in three categories:

- Linguistic tools for textual analysis and retrieval.

- Sparse matrix based tools for statistical topic expansion, clustering and relevance-feedback.
- Result lists combination tools.

A common baseline algorithm was used in all experiments to process the document collection. This algorithm is based on the following sequence of 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 [12] and afterwards extended using several other sources [3] as well as our own developed resources.
5. **Stemming:** This process is applied to each one of the words to be indexed or used for retrieval. Standard Porter stemmers [11] for each considered language have been used.
6. **Indexing and retrieval:** Lucene [2] was used as the information retrieval engine for the whole textual indexing and retrieval task.

This common baseline algorithm is complemented and combined with different expansion techniques in order to compare the improvement given by semantic- versus statistical-based techniques. For the semantic expansion, we used WordNet [4] to expand topic terms with related terms corresponding to a variety of semantic relationships (mainly synonyms and hyponyms).

The statistical method consisted of expanding the topics using the Agrawal's apriori algorithm [1]. First, a term-document matrix is built using the terms in the document corpus. Then apriori algorithm is used to discover out rules having the topic terms as antecedent and a confidence value greater than 0.5. Last, the topic is expanded with the (one-term) consequent of those rules, i.e., terms related to the topic according to the document corpus.

Additionally, relevance-feedback techniques were also used. The top M indexing terms (keywords) of each of the top N result documents were extracted and weighted by a factor that is proportional to their document frequency to reformulate a new query that is executed once again to get the final result list.

To allow cluster diversity, the last step of the process is to rerank the result list, moving the discovered cluster prototypes to the top positions. An implementation of k-Medoids clustering algorithm [8] is used, with k (the target number of clusters) equal to 20 and the maximum number of epochs set to 40. For each resulting cluster, the element with higher relevance in the baseline result list is selected as the class prototype, and reranked to the top of the final result list.

3. Results

Experiments are defined by the choice of different combinations of the previous modules with the different topic expansion techniques and including relevance-feedback or not.

Table 1. Description of experiments

Run Identifier	Language	Method
TitleBaseline	EN/RND	stem + stopwords
TitleBaselineClus	EN/RND	baseline + k-Medoids clustering
TitleAPClus	EN/RND	baseline + Apriori topic expansion + k-Medoids clustering
TitleTagClus	EN/RND	baseline + WordNet topic expansion + k-Medoids clustering
TitleRF1005Clus	EN/RND	baseline + Relevance-Feedback (N=10, M=5) + k-Medoids clustering
TitleAPRF1005Clus	EN/RND	baseline + Apriori topic expansion + Relevance-Feedback (N=10, M=5) + k-Medoids clustering
TitleTagRF1005Clus	EN/RND	baseline + WordNet topic expansion + Relevance-Feedback (N=10, M=5) + k-Medoids clustering

Results are presented in the following tables. Each of them shows 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. Results for English language

	RelRet	MAP	P10	P20	P30	CR10	CR20	CR30
EN_TitleBaseline	1406	**0.1802**	0.2513	**0.2090**	**0.1957**	0.2216	0.2697	0.3034
EN_TitleBaselineClus	1406	0.1662	0.2333	0.1782	0.1769	0.2150	0.2787	0.3339
EN_TitleAPClus	1550	0.1551	0.2385	0.1590	0.1709	0.2323	0.2670	0.3209
EN_TitleTagClus	**1812**	0.1748	**0.2564**	0.1756	0.1855	**0.2366**	**0.3029**	**0.3689**
EN_TitleRF1005Clus	1333	0.0873	0.1051	0.0859	0.0889	0.1087	0.1546	0.2021
EN_TitleAPRF1005Clus	1414	0.0722	0.1359	0.1077	0.0778	0.1393	0.2037	0.2232
EN_TitleTagRF1005Clus	1047	0.0795	0.1333	0.0846	0.0966	0.1263	0.1625	0.2524

For English, the best result in terms of MAP is achieved by the baseline experiment. However, the best cluster precision (CR), which was the variable to maximize in this task, is achieved when k-Medoids algorithm is applied, thus proving to be valuable. The significant improvement in cluster precision is over 6% at CR10 and 21% at CR30.

Table 3. Results for Random language

	RelRet	MAP	P10	P20	P30	CR10	CR20	CR30
RND_TitleBaseline	900	0.0995	0.1692	0.1692	0.1487	0.1858	0.2398	0.2781
RND_TitleBaselineClus	900	0.0954	0.1872	0.1295	0.1333	0.1797	0.2393	0.2943
RND_TitleAPClus	984	0.0892	0.1897	0.1192	0.1325	0.1786	0.2110	0.2846
RND_TitleTagClus	**1270**	**0.1048**	**0.2154**	0.1449	**0.1658**	**0.2133**	**0.2758**	**0.3477**
RND_TitleRF1005Clus	801	0.0536	0.0949	0.0654	0.0615	0.1114	0.1456	0.1942
RND_TitleAPRF1005Clus	732	0.0357	0.0795	0.0487	0.0547	0.0930	0.1108	0.1689
RND_TitleTagRF1005Clus	724	0.0537	0.1000	0.0667	0.0846	0.1234	0.1406	0.2066

Again, as in the case of English, the best results in terms of cluster relevance are obtained in runs that include k-Medoids clustering. MAP value for English is significantly better than for the Random (mixed) language, probably due to the noisy nature of the multi-language annotation.

In general, with respect to MAP, the highest value is obtained with the baseline experiment; MAP values are similar in practice for experiments using topic expansion (Tag and AP) and significantly worse (0.08 against 0.18) in the case of relevance-feedback (RF). This shows that no strategy for topic expansion nor specially relevance-feedback has proved to be useful.

Results are on the average, comparing to other groups.

4. Conclusions and Future Work

A preliminary analysis of the results, given the low precision values obtained in the experiments that make use of the relevance-feedback methods, shows that the reranking algorithm used for combining the different result lists is likely to be the main reason for the disappointing results. However, this impression has to be confirmed with a more in-depth analysis. However, even though all expansion processes produce a decrease in the appropriateness of the results, their recall, as shown in the number of relevant document retrieved) improves in a significant manner.

Another probable cause is the choice of the OR operator to combine the terms in the topic to build up the query. Due to time constraints to prepare this report, we were unable to repeat our experiments with the AND operator, but we think that MAP values should be significantly higher using this operator.

The last conclusion that can be drawn is that the application of clustering techniques smoothes the negative effect of the expansion processes, showing 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] Agrawal, Rakesh; Srikan, Ramakrishnan. Fast algorithms for mining association rules. In Proceedings of the International Conference on Very Large Data Bases, pp. 407-419, 1994.

[2] Apache Lucene project. On line http://lucene.apache.org [Visited 10/08/2008].

[3] CLEF 2005 Multilingual Information Retrieval resources page. On line http://www.computing.dcu.ie/~gjones/CLEF2005/Multi-8/ [Visited 10/08/2008].

[4] Eurowordnet: Building a Multilingual Database with Wordnets for several European Languages. March (1996). On line http://www.illc.uva.nl/EuroWordNet/ [Visited 10/08/2008].

[5] Grubinger, Michael; Clough, Paul; Müller, Henning; Deselaers, Thomas. The IAPR-TC12 benchmark: A new evaluation resource for visual information systems. In International Workshop OntoImage'2006 Language Resources for Content-Based Image Retrieval, held in conjunction with LREC'06, pages 13–23, Genoa, Italy, May 2006.

[6] IAPR: On line http://www.iapr.org/ [Visited 10/08/2008]

[7] ImageCLEF Photo Task: On line http://www.imageclef.org/2008/photo [Visited 14/08/2008].

[8] Krishnapuram, R.; Joshi, A.; Yi, Liyu. A fuzzy relative of the k-medoids algorithm with application to web document and snippet clustering, in: Proceedings of the 1999 IEEE International Conference on Fuzzy Systems, vol. 3, 1999, pp. 1281-1286.

[9] Martínez-Fernández, J.L.; Villena-Román, Julio; García-Serrano, Ana M.; González-Cristóbal, José Carlos. Combining Textual and Visual Features for Image Retrieval. Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Vienna, Austria, Revised Selected Papers. Carol Peters et al (Eds.). Lecture Notes in Computer Science, Vol. 4022, 2006. ISSN: 0302-9743.

[10] 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 & Industrial Engineering, pp.1222-1231, Taipei, Taiwan, Jun. 20-23, 2006.

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

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

[13] Villena-Román, Julio; Lana-Serrano, Sara; González-Cristóbal, José Carlos. MIRACLE at ImageCLEFmed 2007: Merging Textual and Visual Strategies to Improve Medical Image Retrieval. Working Notes of the 2007 CLEF Workshop, Budapest, Hungary, September 2007.

[14] 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. Working Notes of the 2007 CLEF Workshop, Budapest, Hungary, September 2007.