## 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, Spain
{magc,mcdiaz,amontejo,maite}@ujaen.es

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.
Four experiments were run:

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.

The obtained results are not successful and 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.

## The University of Glasgow at ImageCLEFPhoto 2009

Guido Zuccon, Teerapong Leelanupab, Anuj Goyal, Martin Halvey, P. Punitha and Joemon M. Jose

University of Glasgow, Glasgow, G12 8RZ, United Kingdom
{guido,kimm,anuj,halvey,punitha,jj}@dcs.gla.ac.uk

In this extended abstract we briefly describe the approaches adopted to generate the five runs submitted to ImageCLEFPhoto 2009 by the University of Glasgow. The aim of our methods is to exploit document diversity in the rankings. In four out of five submitted runs, we have tackled the problem of diversity considering topicality and semantics. Under these approaches, the content of the documents have been assumed to coincide with the text (the caption) associated with each image. All the four runs based on text aim to exploit the statistical features drawn from the provided textual captions, promoting topically and semantically different documents. In particular, we implement the well known Maximal Marginal Relevance (MMR) method in "Glasgow−run−1" to serve as benchmark for our runs named "Glasgow−run−3" and "Glasgow−run−4". These runs rely on a clustering procedure (on the textual data) to individuate topically and semantically diverse facets of the retrieved documents and re-rank the initial ranking combining those evidences and the MMR methodology. While the previous approaches start from empirical observations to derive a model, our approach (Glasgow−run−2) based on the Quantum Probability Ranking Principle (QPRP) is motivated employing quantum theory. Finally, the approach called "VisualDiversity" (Glasgow−run−5) is our only method that uses both the visual features of the images and the text features of the associated captions. Visual features are processed using factor analysis and bi-clustering to derive an image ranking which promotes diversity amongst the results.

The results obtained by evaluating our rankings using the ground truth data suggest that our methods based on text captions significantly improve the performance of the respective baselines for cluster recall, up to 19.15%. Also, we show that "Glasgow−run−3" and "Glasgow−run−4" outperform the original formulation of MMR. However, the approach that combines visual features with text statistics (Glasgow−run−5) shows lower levels of improvements. This might be due to the fact that the results diversity required in the ImageCLEFPhoto 2009 dataset is preeminently topical rather than visual.

## SINAI at ImageCLEF 2009 WikipediaMM task

M.C. Díaz-Galiano, M.T. Martín-Valdivia, L.A. Ureña-López and J.M. Perea-Ortega

University of Jaén. Departamento de Informática
Grupo Sistemas Inteligentes de Acceso a la Información
Campus Las Lagunillas, Ed. A3, E-23071, Jaén, Spain
{mcdiaz,maite,laurena,jmperea}@ujaen.es

This paper describes the first participation of the SINAI team in the CLEF 2009 wikipediaMM task. This year, we only want to establish a first contact with the task and the collections. Thus, we have generated a new collection expanding with WordNet terms in order to perform the information included in this collection. In addition, we have expanded de queries with WordNet too. We have used the LEMUR toolkit as the Information Retrieval system in our experiments.

The experiments are named with the name of the group, the name of the collection and the name of the topic set.

- **sinai_NT_T**: Baseline experiment. Collection includes narrative and title and topics include only title.
- **sinai_NTWn_T**: Collection includes narrative and title expanded with WordNet and topics include only title.
- **sinai_NT_TWn**: Collection includes narrative and title and topics include title expanded with WordNet.
- **sinai_NTWn_TWn**: Collection includes narrative and title expanded with WordNet and topics include title expanded with WordNet.

The query expansion does not improve the results. Only the experiment named *sinai_NTWn_T* improve the baseline results.

The obtained results show that it is necessary to continue investigating the expansion methodology. Thus, our next goal will be to improve the expansion by applying some more techniques.

## DCU at WikipediaMM 2009: Document Expansion from Wikipedia Abstracts

Jinming Min, Peter Wilkins, Johannes Leveling and Gareth Jones

Centre for Next Generation Localisation
School of Computing, Dublin City University
Dublin 9, Ireland
{jmin,pwilkins,jleveling,gjones}@computing.dcu.ie

In this paper, we describe our participation in the WikipediaMM task at ImageCLEF 2009. Our main efforts concern the expansion of the image metadata from the Wikipedia abstracts collection DBpedia. Since the metadata is short for retrieval by query words, we decided to expand the metadata using a typical query expansion method. In our experiments, we use the Rocchio algorithm for document expansion. And we choose the Wikipedia abstracts collection (DBpedia) as the external resource for document expansion. The reason to choose DBpedia is that it has similar characteristic with the Wikipedia image metadata collections. Our best run is the combination of the document expansion from DBpedia and typical query expansion with MAP 0.1752. It is in the 26th rank of all 57 runs which is under our expectation, and we think that the main reason is that our document expansion method uses all the words from the metadata documents which contain words which are unrelated to the content of the images. Compared with our text retrieval baseline, our best document expansion run improves MAP by 11.17%. As one of our conclusions, we think that the document expansion can play an effective factor in the image metadata retrieval task. Our content-based image retrieval uses the same approach as in our participation in ImageCLEF 2008.

## Medical Image Retrieval: ISSR at CLEF 2009

Waleed Arafa and Ragia Ibrahim

Department ofex Computer & Information Sciences
Institute of Statistical Studies and Research (ISSR), Cairo University, Egypt
{waleed_arafa,ribrahim}@issr.cu.edu.eg

This paper represents the first participation of the Institute of Statistical Studies and Research at Cairo University group in CLEF 2009-Medical image retrieval track. Our system uses Lemur toolkit for text retrieval. Our main objective is to carry out retrieving medical image depending on associated image text. We experimented with different text features such as article title, image caption and the article paragraph(s) denoting to the image. We propose a simple and effective extraction method to find relevant paragraphs based on the structure of HTML files. We experimented also the automatic translation of queries in different languages other than collection language. We represent results of 9 runs in order to compare retrieval based on different text features and the effect of stop word lists and the use of relevance feed back.

We used CLEF 2008 Medical image collection to test our approach. Our approach increased the recall of 60% of the 25 English queries but the mean average precision is decreased since the added paragraphs may refer to more than one figure and may contain many irrelevant terms to the image which caused the relevant documents get low rank in the retrieved list.

We used Google statistical machine translation to translate French and German queries to English before retrieval, French translation increased the mean average precision about 15% while the German translation decreased it by about 26%. However, this result is considered acceptable for statistical automatic translation.

We intend to enhance this approach using semantic extraction methods such as shallow NLP techniques or statistical approaches to extract only relevant sentences from the paragraph denoting the image instead of adding the whole paragraph in order to reduce noise terms.

## SINAI at ImageCLEF 2009 Medical Task

M.C. Díaz-Galiano, M.T. Martín-Valdivia, L.A. Ureña-López and M.A. García-Cumbreras

University of Jaén. Departamento de Informática
Grupo Sistemas Inteligentes de Acceso a la Información
Campus Las Lagunillas, Ed. A3, E-23071, Jaén, Spain
{mcdiaz,maite,laurena,magc}@ujaen.es

This paper describes the SINAI team participation in the ImageCLEF 2009 medical task.

We explain the experiments accomplished in the medical retrieval task (ImageCLEFmed). We have experimented with query and collection expansion. For expansion, we carry out experiments using MeSH ontology.

This year we have two types of topics sets: image based retrieval topics (adhoc) and case based retrieval topics. We have expanded these topics sets and we have obtained the next topics sets:

- t: Original topics set to use in adhoc retrieval.
- tM: Topics set to adhoc retrieval expanded with MeSH ontology.
- cbt: Original topics set to use in case based retrieval.
- cbtM: Topics set to case based retrieval expanded with MeSH ontology.

With respect to text collection, we have used different collections:

- C: It contains caption of image to use in image based retrieval.
- CT: Constains caption of image and title of the article to use in image based retrieval.
- CM: It contains caption of image expanded with MeSH to use in image based retrieval.
- CTM: Constains caption of image and title of the article expanded with MeSH to use in image based retrieval.
- TA: Constains title and text of the full article to use in medical case based retrieval.

Moreover, we have experimented only with textual search, using the LEMUR toolkit as the Information Retrieval system.

The obtained results are not successful and we are investigating the reasons of these unexpected results.

## UAIC: Participation in INFILE@CLEF Task

Cristian-Alexandru Drăguşanu, Alecsandru Grigoriu, Andreea-Loredana Andreşan, Daniela Epure, Dan Anton and Adrian Iftene

UAIC: Faculty of Computer Science, "Alexandru Ioan Cuza" University, Romania

{cristian.dragusanu, alecsandru.grigoriu, andreea.andresan, daniela.epure, dananton, adiftene}@info.uaic.ro

INFILE@CLEF (information filtering evaluation) extends the TREC 2002 filtering track. In comparison, it uses a corpus of 100,000 Agence France Press comparable newswires for Arabic, English and French (Besançon et al., 2008). The participants received news collections contains 100,000 news articles for each language (English, French and Arabic) and 50 topics and competitors must compare each topic in a source language to the documents in the target languages.

Our system has three main modules: module one responsible with XML parsing, module two that indexes the XML files, and the third module that does the filtering.

**Module for XML Parsing:** First of all, we parse the XML files with aim to parse and to extract relevant content from documents (which are in News-ML format). At this step, both the XML files with news and XML files with topics are parsed.

**Indexing Module:** For indexing we use Lucene. The Indexing module receives the relevant tags from each XML file and analyzes and stores that information in the main index database.

**Filtering Module:** In the Filtering part, the file containing the 50 topics (in XML format) is parsed by XML Parsing module. Then, for each one of the 50 topics, a number of fields are stored and are sent to the Filtering module. The Filtering module receives the topic details, sorts and filters individual words from all fields and generates a search query based on the most frequent relevant words from the topic.

**Submitted Runs:** For our runs the search was made in 2 languages, English and French, using topics in English. For Run 1 and Run 2, we used different priorities for the search terms. For Run 4 we also used priorities for the terms but we looked only in *Headline* and *DataContent* fields. All terms had the standard priority only in Run 3, which returned the best result. For the last three runs we used a translation algorithm to return Eng-Fre results. The best result was obtained for Run 3, where English was considered as source and as target language.

## SINAI at INFILE 2009: Experiments with Google News

Arturo Montejo-Ráez, José M. Perea-Ortega, Manuel Carlos Díaz-Galiano and L. Alfonso Ureña-López

SINAI research group
Computer Science Department. University of Jaén
Campus Las Lagunillas, Ed. A3, E-23071, Jaén, Spain
{amontejo,jmperea,mcdiaz,laurena}@ujaen.es

This paper describes the SINAI team participation in the INFILE routing and filtering track of the CLEF campaign. This is the first participation of the SINAI research group in the INFILE task. We have participated in the batch filtering subtask, and only the English language has been considered. A supervised learning scheme was followed, by creating learning corpora and classifying documents into one of the 50 predefined topics. Two experiments have been submitted depending on the learning corpus used: one using the topics' text as learning data to train a multi-class classifier, and another one where training data has been constructed from Google News pages. The Google News corpus consists of pages downloaded from queries on this specific service. The queries were taken topics keywords and up to 50 pages per keywords were downloaded. Our results show that our use of Google News did not improved the performance of that of the classification obtained using only topics description. Google News pages showed a high level of noise in its content.

## SINAI at VideoCLEF 2009

José M. Perea-Ortega, Arturo Montejo-Ráez, M. Teresa Martín-Valdivia and L. Alfonso Ureña-López
SINAI research group. Computer Science Department. University of Jaén
Campus Las Lagunillas, Ed. A3, E-23071, Jaén, Spain
{jmperea,amontejo,maite,laurena}@ujaen.es

This paper describes the second participation of the SINAI research group in the VideoCLEF track. This year we only participated in the subject classification task. A training collection was generated using the data provided by the VideoCLEF organization. Over this data, a supervised learning approach to classify the test videos was conducted. We have used Support Vector Machines (SVM) as classification algorithm and two experiments have been submitted, using the metadata files and without using them, during the generation of the training corpus. For the experiments and analysis carried out, the Rapid Miner framework was selected. This toolkit provides several machine learning algorithms such as SVM and techniques along with other interesting features.

The results obtained show the expected increase in precision due to the use of metadata in the classification of the test videos. The use of metadata improves about 21.7% the average precision of the classification of the test videos.

## DCU at VideoCLEF 2009

Ágnes Gyarmati and Gareth J. F. Jones
Centre for Digital Video Processing
Dublin City University, Dublin 9, Ireland
{agyarmati|gjones}@computing.dcu.ie

DCU participated in the VideoCLEF 2009 Linking Task. Our approach was based on identifying relevant related content using the Lemur information retrieval toolkit. We implemented two distinctive variants of our approach. One version performs the search in the Dutch Wikipedia with the exact words (either stemmed or not) of the search query extracted from the ASR transcription, and returns the corresponding links pointing to the English Wikipedia. The other variant first performs an automatic machine translation of the Dutch query into English, and then the translated query is used to search the English Wikipedia directly. Among our four runs, we achieved the best results with the first approach, when the base of retrieval was the stemmed and stopped Dutch Wikipedia. Unfortunately for us, there is no one-to-one relation between the pages of the Dutch and the English Wikipedias, hence some hits from the Dutch Wikipedia have been lost as results due to lack of equivalent English article. In extreme cases, our system might return no output at all if none of the hits for a given anchor are linked to a page in the English Wikipedia. Although we included a preprocessing phase before indexing the article collections, some unuseful, but frequently occurring types of page escaped and had a significant negative impact of our second basic approach implemented in Run 3.