Evaluating Fusion Techniques at Different Domains at ImageCLEF Subtasks

Sergio Navarro, Rafael Muñoz and Fernando Llopis
Natural Language Processing and Information Systems Group. University of Alicante. Spain.
snavarro,rafael,llopis@dlsi.ua.es

Abstract

In our participation in the 2009 edition of the ImageCLEF task we pursued two objectives, first to expand the number of subtasks in which we evaluate our proposed multimodal fusion presented in previous works, MultiModal Local Context Analysis (MMLCA). Furthermore, we evaluated three new proposals: a subquery generation technique based on clustering, a new variant of Multimodal Re-ranking TF-IDF (MMRR TF-IDF), finally we evaluated a term filtering technique for the medical domain which is based on WordNet Domains. [4]. From the experiments conducted: On the one hand, we have confirmed that MMLCA performs better than the other local expansion techniques evaluated. On the other hand, the results show that our proposal of subquery generation based on clustering combined with PRF obtains especially good results (the 5th best textual run of the Photo Retrieval subtask in terms of F-Meassure).

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 Managment**]: Languages—*Query Languages*

General Terms

MMLCA, LCA, MMRR TF-IDF, MeSH, WordNet

Keywords

Image Retrieval, Multimodal Fusion Techniques

1 Introduction

The wide variety of digital formats on the Internet and the boom of multimedia content create the need to develop and/or adapt tools for finding information with these new characteristics such as video and image among others. The development of search engines able to manage properly these new sources is beyond the scope of the VIR research field and specifically beyond the scope of the ImageCLEF subtasks. We can say that the VIR is a specific area within the Information Retrieval (IR), which in fact initially used traditional IR systems without any specific adaptation to the VIR, performing searches only using the annotations related to the images. Thus, the collections used by VIR systems are composed of images and their related annotations describing their content.

Historically in VIR area there were two approaches used to carry out the IR of images: In the beginning of the VIR in the late 70s, VIR systems were based on the image annotations, therefore, these were Text-Based VIR (TBIR) systems. Later in the early 90's, in an attempt to overcome the dependence of TBIR systems from the existence of textual annotations to perform the indexing of an image, the image Content Based VIR (CBIR) systems appear [2].

Finally, in recent years as the technologies used by CBIR systems matured, a third approach to tackle the problem of the VIR emerged, these systems combine textual and image based technologies. In this context are organized competitions like ImageCLEF[1] which is a specific VIR task which takes place within the framework of the annual competitions of the CLEF[2] campaigns. These competitions aim the development of multimodal systems using image collections with their related short annotations.

In our participation in the 2009 edition of the ImageCLEF task we pursued two objectives, first to expand the number of subtasks in which we evaluate our proposed multimodal fusion presented in previous works, MultiModal Local Context Analysis (MMLCA) [6]. Furthermore, we evaluated three new proposals: a subquery generation technique based on clustering, a new variant of Multimodal Re-ranking TF-IDF (MMRR TF-IDF), finally we evaluated a term filtering technique for the medical domain which is based on WordNet Domains. As in previous editions we have customized our TBIR system, IR-n [3], to be able to harness the image list returned by a CBIR system.

This paper is structured as follows: Firstly, it presents the main characteristics of the IR-n system focusing on the techniques used, and then it moves on to describe the experiments and the results obtained. Finally, it presents conclusions and future work.

2 The IR-n System

To perform the experiments we used IR-n, an information retrieval system based on passages. Such systems treat each document as a set of passages; each passage defines a portion of text from the document. Unlike systems based on documents, passages based systems give greater relevance to those documents where the query terms appear in closer positions to each other [3].

2.1 Local Query Expansion

The IR-n architecture allows us to use local query expansion based on either the most relevant passages or the most relevant documents. Furthermore, it allows choosing between two different term selection strategies for the query expansion. These strategies are Probabilistic Relevance Feedback (PRF)[9] and Local Context Analysis (LCA) [11]. Moreover it supports the multimodal versions of these two strategies, MMPRF and MMLCA. These multimodal versions use the top-ranked documents returned by the CBIR system as input for the term selection strategy (PRF or LCA).

In a previous work we evaluated these techniques on different generic domain image collections. We concluded that MMLCA is the best of these techniques in terms of precision and the only one of the four which does not hurt the diversity of the results returned by the VIR system. [6]

2.2 Multimodal Re-ranking Strategy

This strategy involves the merging of the list returned by a TBIR system and the list returned by a CBIR system. Following a standard re-ranking strategy this is done by giving a different weight to the normalized relevance value or ranking position for a document in each list, the Textual List (TL) and the Visual List (VL).

IR-n allows to use the standard re-ranking strategy and also a variation of our MMRR TF-IDF proposal used in previous works [6].

[1] *http://www.imageclef.org*

[2] *http://www.clef-campaign.org*

Our previous TF-IDF Re-ranking proposal was based on the following assumptions: on the one hand, the list based on image annotations is more confident than the list based on images and on the other hand, we assumed that TF-IDF is a suitable way to measure the quantity and the quality of a text. Thus, we used a TF-IDF threshold to decide whether for the final relevace value of an image we only use the relevance value returned by the TBIR or we take the risk of adding also the relevance value returned by the CBIR.

These assumptions proved to be effective but were not enough. There were two problems with this approach: The first one is that we consider that any relevance value obtained from a CBIR system always have the same level of risk, independently of the position of its related image in the VL. Nervertheless, this is not true since that although CBIR systems obtain lower MAP results than TBIR systems, in general the first ones obtain not bad P20 results (see Photo Retrieval 2008 results).

The second problem in our previous approach was that the TF-IDF threshold was very dependent of the images returned for each query. The reason is that the system used a percentage value as parammeter for the threshold. Later for each query it obtains the value of the threshold applying the percentage parammeter to the maximum TF-IDF value obtained from the documents in the two lists. This dependence from the image subset retrieved could cause erratic results.

In order to solve these problems our proposal should use a new threshold formula. This formula should allow us to manage different levels of risk in order to achieve an optimal performance and should depend from the characteristics of the entire colletion instead of only depend on the image subset retrieved for each query.

The new threshold (1) used is based on two points: firstly, on to work out the average of the TF-IDF value per sentence ($TFIDFperSentence$) for the whole collection. We use this value for avoiding the dependence of the characteristics of the image set returned for a query. Secondly, on to use a simple lineal function for decreasing the threshold value as the position of the image increases in the VL.

$$threshold(pos) = numSentences * TFIDFperSentence * (\frac{-pos}{maxPosCBIR} + 1) \tag{1}$$

- $TFIDFperSentence$ is the average TF-IDF value per sentence for the whole collection, this value is worked out automatically by the system.
- $numSentences$, this parameter is a multiplier which represents the maximum number of sentences that the annotations related to an image can contain in order to take into account the relevance value returned by the CBIR for this image.
- pos is the position of the image in the VL.
- $maxPosCBIR$, this parameter indicates the limit position in the visual ranking to use the relevance values returned by the CBIR.

Reviewing the TF-IDF MMRR formula (2) we can observe that the *threshold* is used to avoid the risk of use the CBIR relevance values for those images which are low ranked in the VL and for those ones which have image annotations with enough quantity and quality to perform a suitable TBIR.

$$FL(d) = \begin{cases} TL(d) + VL(d), & \text{if } 0 <= pos(d) <= maxPosCBIR \\ & \text{and threshold(pos(}d\text{))} > TFIDF(d) \\ TR(d), & \text{else} \end{cases} \tag{2}$$

- $TFIDF(d)$ is the TF-IDF value of the text document d related to an image.
- $pos(d)$ returns the position of the document d if it exists in the visual list else it returns -1.

It is important to point out that when the TF-IDF MMRR strategy is enabled it is executed for each retrieval iteration. It implies that if the system is using a local expansion technique the TF-IDF MMRR will be executed for the two retrieval iterations involved. Firstly, in the first iteration of the retrieval it will modify the document list used to feed the local expansion strategy. Finally, in the second iteration it will modify the document list returned after retrieve the expanded query.

2.3 Subqueries Generation based on Clustering

Usually, when the user uses a VIR system, they find that there are several similar images between the first results returned by the system. Thus, if they want to find different relevant images they have to navigate through the next pages of the returned results.

Our approach is an attempt to solve this problem. It is based on Lingo [7], the default clustering algorithm of Carrot2[3] an open source clustering engine for text. Our clustering module passes to Carrot2 a preprocessed version of the top relevant documents in the ranking list. The preprocessing step removes the stopwords and extracts the stem from the original documents before those are passed to Carrot2.

Opposite to the clustering approach we used in our previous participation, instead of using the documents returned by the clustering tool, our system uses the set of terms which represent each one of the clusters returned by Carrot2, the label of the clusters. The system builds a subquery per each term set representing a cluster. The new subqueries are compounded by the terms of the original query jointly with those terms representative of the cluster, discarding those terms which already exist in the original query and that are not repeated in other cluster labels returned by Carrot2.

Finally, the system performs the retrieval for each one of the new generated subqueries obtaining a ranking list per subquery. These ranking lists are fused using an standard re-ranking strategy in order to produce the final ranking result.

It is important to point out that when the clustering module is used to create subqueries, the system uses the same local expansion and re-ranking configuration for obtaining the ranking used to feed this clustering module and for the retrieval of each one of the subqueries generated.

2.4 Medical Stopwords based on WordNet Domains

In our previous experiments using generic domain image collections we observed that especially when the precision of the initial ranking used for the expansion has a poor precision, in general LCA shows better precision results than PRF. [6]

However, this behavior was not observed in the experiments performed at the 2008 Medical Retrieval subtask. Reviewing the behavior of the system, we concluded that a reason for this different behavior could be that in the medical domain the terms used in the collections follow a different distribution of significance in comparison with their distribution in a generic domain. This different distribution could hurt the performance of LCA, an strategy based on term coocurrence. In fact, we believe that while some terms of the query are especially relevant under a medical viewpoint other of its terms are not so important as their frequency figures in the collection indicate. Thus, a possible solution could be to filter the terms returned by the local expansion strategy for the expansion, in order to only allow terms with enough significance under the medical domain.

For this purpose, we generated a stopword list specifically for the medical domain. In order to generate it we used all the WordNet terms which not pertain to the following medical WordNet Domains: *medicine*, *dentistry*, *pharmacy*, *psychiatry*, *radiology*, *surgery*, *purescience*, *chemistry*, *biology*, *biochemistry*, *zoology*, *anatomy*, *physiology* and *genetics*.

[3] *http://www.carrot2.org*

2.5 MeSH Query expansion

The system uses a query expansion module based on MeSH. It is the same used in our past participation in the Medical Retrieval task at ImageCLEF 2008.

3 ImageCLEF Participation

Table 1: Results in Photo Retrieval Subtask. Part 1 Topic Set.

run	relFB	rrtfidf	clust	MAP	P10	CR10	F-Mea	rkCLEF
Southampton TXT	-	-	-	0.3709	0.868	0.7730	0.8178	1/84
Southampton TXTIMG	-	-	-	0.3329	0.804	0.8063	0.8052	2/84
Alicante4 TXTIMG	MMLCA (1/0/20/5)	yes	yes	0.4232	0.8000	0.7500	0.7742	11/84
Alicante2 TXT	PRF (1/5/0/10)	no	yes	0.3902	0.8280	0.7056	0.7619	15/84
Alicante3 TXTIMG	MMLCA (1/0/20/5)	yes	no	0.4038	0.736	0.4899	0.5882	55/84
Alicante1 TXT	PRF (1/5/0/10)	no	no	0.4230	0.7960	0.4301	0.5584	62/84

This section shows the results of the system and describes the configuration used for each one of the tasks in which we have been involved. Below there is the description of each one of the IR-n configuration parameters used in our different participations:

- **Relevance Feedback** (*relFB*): Indicates which relevance feedback is used PRF, LCA, MMPRF or MMLCA.

- **Relevance Feedback parameters** (*exp*/*num*/*ncbir*/*term*): If *exp* has value 1, this denotes we use relevance feedback based on passages. But, if *exp* has value 2, the relevance feedback is based on documents. Moreover, *num* denotes the number of passages or documents that the local expansion strategy will use from the textual ranking, *ncbir* denotes the number of documents that the multimodal local expansion strategy will use from an image based list and finally, *term* indicates the number of terms that the local expansion strategy will use for the query expansion.

- **Subqueries Generation based on Clustering (clust)**: Indicate if this module is used or not.

- **Multimodal Re-ranking Strategy (rrtfidf)**: Indicate if the system uses the MMRR TFIDF or not.

- **Automatic query expansion based on MeSH (mesh)**: Indicates if it is used or not.

- **Medical Stopwords based on WordNet Domains(stopWN)**: Indicates if the system uses medical stopwords filtering for the terms selected by the local expansion technique.

For all the experiments we have use DFR as the weighting schema. We based this decision on our training results at ImageCLEF 2007 edition.

3.1 Photo Retrieval Subtask

For our participation in the Photo Retrieval subtask the CBIR system used was FIRE [1]. We found problems to process with this CBIR the big collection used this year in this subtask. The time constraints and our lack of knowledge to modify the initial configuration of FIRE, in order to make it able to process the entire collection, forced us to split the collection in three parts and launch each query for each one of the FIRE indexes created. After this step we fused the three visual lists putting their results in descending order and selecting only the 1000 highest results. Despite we understand this could hurt the precision of the CBIR we carry on the experiments in order to see how it affects to the multimodal fusion techniques used by IR-n.

Further information regarding the collections and the topic sets used in this task can be found at [8].

We divide the results achieved in the competition in two parts. The Table 1 shows the Part 1 query set results ordered by F-Meassure. It shows the results of our submitted runs jointly with the best TXT run and the best TXTIMG run of the subtask for this topic set. The characteristic of the queries within this query set is that they provide a number of subqueries and images, per query, which define the different clusters that should be retrieved in order to return a diverse list of images to the user.

Using the extra information provided for this query set IR-n replaces the cluster labels returned by Carrot2 with the subqueries provided for each query. Furthermore, the system obtains the VL for each subquery using the image provided with each subquery of the topic.

For this topic set our group was the 5th best group in terms of F-Meassure for a total of 19 participants.

Table 2: Results in Photo Retrieval Subtask. Part 2 Topic Set.

run	**relFB**	**rrtfidf**	**clust**	**MAP**	**P10**	**CR10**	**F-Mea10**	**rkCLEF**
XRCEXKNND TXTIMG	-	-	-	0.3729	0.8200	0.8189	0.8194	1/80
INFOCOMM TXT	-	-	-	0.4256	0.8280	0.6901	0.7528	3/80
Alicante2 TXT	PRF (1/5/0/10)	no	yes	0.2879	0.7520	0.6031	0.6694	20/80
Alicante3 TXTIMG	MMLCA (1/0/20/5)	yes	no	0.4428	0.7360	0.5758	0.6461	37/80
Alicante5 TXTIMG	MMLCA (1/0/20/5)	yes	yes	0.2513	0.7400	0.5321	0.6191	49/80
Alicante1 TXT	PRF (1/5/0/10)	no	no	0.3976	0.6600	0.5478	0.5987	54/80
Alicante4 TXTIMG	MMLCA (1/0/20/5)	yes	yes	0.2540	0.6880	0.5208	0.5928	55/80

Table 2 shows the results achieved for the Part 2 query set ordered by F-Meassure. This query set did not provide the subqueries per query as the previous one. Therefore, the runs which used the clustering module used Carrot2 in order to produce the different subqueries per cluster. Furthermore, for all these runs the VL used for the retrieval is shared between all the subqueries created from a query. This VL is obtained querying the CBIR system using the three images provided with the query. There was an exception, the *Alicante*5 run. This one used a different VL for each subquery generated by the clustering module. Each of these visual lists is obtained using the first image in the TL returned by IR-n in response to each textual subquery created.

For the second topic set our group was the 7th best group in terms of F-Meassure for a total of 19 participant groups.

Finally, point out that our *Alicante*2 run was the 5th best textual run of the competition based on its results for the two subsets.

3.2 WikipediaMM Subtask

Table 3: Results in WikipediaMM Retrieval Subtask.

run	**relFB**	**rrtfidf**	**MAP**	**P10**	**P20**	**rkCLEF**
deuceng TXT	-	-	0.2397	0.4000	0.3133	1/57
lach TXTIMG	-	-	0.2178	0.3378	0.2811	5/57
Alicante7 TXTIMG	MMLCA (1/0/20/5)	no	0.1878	0.2733	0.2478	17/57
Alicante5 TXTIMG	MMLCA (2/5/20/10)	no	0.1806	0.2533	0.2356	22/57
Alicante4 TXTIMG	MMPRF (2/5/5/10)	no	0.1801	0.2644	0.2267	23/57
Alicante1 TXT	no	no	0.1784	0.2556	0.2289	25/57
Alicante2 TXT	PRF (2/5/0/10)	no	0.1745	0.2511	0.2133	27/57
Alicante6 TXTIMG	MMPRF (1/0/5/5)	no	0.1697	0.2422	0.2178	30/57
Alicante8 TXTIMG	no	yes	0.1592	0.2400	0.2200	37/57
Alicante9 TXTIMG	MMLCA (1/0/20/5)	yes	0.1222	0.1533	0.1422	46/57

For our participation in the WikipediaMM subtask we used the same Camel Case decompounding image filenames technique which we used in our participation at ImageCLEF 2008 edition. Furthermore, for this subtask we have generated the VL using the similarity matrix provided by the organizers, this similarity matrix was generated by the IMEDIA group at INRIA. Due to problems found in the similarity matrix and time constraints we were force to use VL only for those images in the queries which are not contained in the collection. It could affect negatively to the results achieved by our multimodal techniques because our system did not harness all the images provided in the queries.

The multimodal techniques used for this subtask use a VL per each image in the query instead of use a unique VL for all the images of the query. It changes slightly the behavior of the two multimodal techniques used. On the one hand, the multimodal local expansion techniques manage this issue using the annotations related with the *ncbir*-top ranked images from each VL to feed the local expansion technique. On the other hand, the MMRR TF-IDF strategy fuses these visual lists in one VL. In order to do so for those images which appear in more than one VL it sets as relevance for the final VL the maximum relevance found in the original visual lists.

Table 3 shows the results of our submitted runs jointly with the best TXT run and the best TXTIMG run of the subtask.

Further details about the dataset and the test topic set used in the task can be found at [10].

Our group was ranked at the 4th position in terms of MAP for a total of 8 participant groups.

3.3 Medical Retrieval Subtask

For our participation in the Medical Retrieval subtask we used a CBIR baseline based on 8 grey levels wchich was provided by the organization. This baseline was generated using GIFT [4] a CBIR system.

Further information regarding the collections and the topic sets used in this task can be found at [5].

Table 4 shows the results of our submitted runs jointly with the best TXT run and the best TXTIMG run of the subtask.

Table 4: Results in Medical Retrieval Ad-hoc Subtask.

run	**relFB**	**mesh**	**wnFilter**	**MAP**	**P10**	**P20**	**rkCLEF**
LIRIS TXT	-	-	-	0.4293	0.6640	0.6060	1/106
ITI TXTIMG	-	-	-	0.3775	0.7160	0.6060	10/106
Alicante2 TXT	PRF (1/5/0/10)	yes	no	0.1466	0.3280	0.2800	68/106
Alicante4 TXTIMG	MMLCA (2/5/25/10)	yes	no	0.1337	0.3440	0.3400	70/106
Alicante3 TXT	PRF (2/5/0/10)	yes	yes	0.1335	0.360	0.3440	72/106
Alicante1 TXT	no	no	no	0.1314	0.4000	0.3860	73/106
Alicante5 TXTIMG	MMLCA (2/5/25/10)	yes	yes	0.1280	0.3520	0.3380	77/106

Our group was ranked at the 13th position in terms of MAP for a total of 16 participant groups.

4 Conclusion and Future Work

Our conclusions related with the generic domain subtasks are the following ones: On the one hand, taking into account that for the Photo Retrieval subtask we have used an irregular procedure for obtaining the visual lists and that for the WikipediaMM we do not have used all the images provided with the queries, we have confirmed that even with this problems MMLCA performs better than the other local expansion techniques evaluated (see Photo Retrieval Part 1 and WikipediaMM results).

On the other hand, regarding our subqueries generation based on clustering technique, we observe that it works better when it is combined with PRF than when it is combined with MMLCA. It explains the good results achieved by our *Alicante*2 run, which using PRF and out clustering based technique achieved the 5th best textual run of the Photo Retrieval subtask. In future works we will study deeply the reasons for this profitable relationship.

Point out that after reviewing the results we found an important bug in the MMRR TF-IDF strategy used that has hurted all the runs which used it. Thus, we are forced to delay for future works the evaluation of this technique.

Finally, for the Medical Retrieval results we conclude that our proposal of Medical stopwords based on WordNet Domains did not improve the MMLCA results as we expected. Further analysis should be performed in the future to understand why MMLCA does not work as well in medical domain as it does for generic domain.

[4] *http://www.gnu.org/software/gift/*

5 Acknowledgment

This research has been partially funded by the Spanish Government within the framework of the TEXT-MESS (TIN-2006-15265-C06-01) project and by European Union (EU) within the framework of the QALL-ME project (FP6-IST-033860).

References

[1] Tobias Gass, Tobias Weyand, Thomas Deselaers, and Hermann Ney. Fire in imageclef 2007: Support vector machines and logistic regression to fuse image descriptors in for photo retrieval. In *Advances in Multilingual and Multimodal Information Retrieval 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007*, volume 5152 of *LNCS*, Budapest, Hungary, 19/09/2007 2008. Springer, Springer.

[2] Michael Grubinger. *Analysis and Evaluation of Visual Information Systems Performance.* PhD thesis, Engineering and Science Victoria University, 2007.

[3] Fernando Llopis, José L. Vicedo, and Antonio Ferrández. IR-n System at CLEF-2002. In *3th Workshop of the Cross-Language Evaluation Forum, CLEF 2002*, Lecture notes in Computer Science, pages 291–300, 2002.

[4] Bernardo Magnini Luisa Bentivogli, Pamela Forner and Emanuele Pianta. Revising wordnet domains hierarchy: Semantics, coverage and balancing. In *Proceedings of COLING 2004 Workshop on 'Multilingual LinguisticResources'*, pages 101–108, Geneva, Switzerland, August 2004.

[5] Henning Mller, Jayashree Kalpathy-Cramer, Ivan Eggel, Steven Bedrick, Sad Radhouani, Brian Bakke, Charles Kahn Jr., and William Hersh. Overview of the clef 2009 medical image retrieval track. In *CLEF working notes 2009*, Corfu, Greece, 2009.

[6] Sergio Navarro, Fernando Llopis, and Rafael Muñoz. Using evidences based on natural language to drive the process of fusing multimodal sources. *13th International Conference on Applications of Natural Language to Information Systems, NLDB 2009. (Not yet published).*

[7] Stanislaw Osinski, Jerzy Stefanowski, and Dawid Weiss. Lingo: Search results clustering algorithm based on singular value decomposition. In *Intelligent Information Systems*, pages 359–368, 2004.

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

[9] S. E. Robertson and K. Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3):129–146, 1977.

[10] Theodora Tsikrika and Jana Kludas. Overview of the clef 2009 medical image retrieval track. In *CLEF working notes 2009*, Corfu, Greece, 2009.

[11] 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.