

IPL at ImageCLEF 2011 Medical Retrieval Task

Yiannis Gkoufas, Anna Morou, Theodore Kalamboukis

Information Processing Laboratory
Department of Informatics
Athens University of Economics and Business, Greece
{gkoufas, morou, tzk}@aueb.gr

Abstract. This article describes IPL’s participation to the image CLEF ad-hoc textual and visual medical retrieval for 2011. We report on our approaches and methods from a systematic experimental investigation on fusion from visual and textual sources of images. We also explore ways to enrich our searches using external sources like Wikipedia.

1 Introduction

In this article we give an overview of the application of our methods to ad-hoc medical image retrieval and present the results of our submitted runs. The goal of the CLEF medical retrieval task is to advance the performance of multimedia objects retrieval in the medical domain combining techniques from Information Retrieval and Content Based Image Retrieval (CBIR).

Results so far, show that textual retrieval of images outperforms retrieval based only on low level features [7] [8]. This was also the case in the results reported by the CLEF campaign in the medical retrieval task this year. Thus, techniques of merging the retrieval results of multimedia objects out of multi-sources, remains an interesting and hot research topic.

Most of our efforts this year were concentrated on data fusion techniques of different low level visual and textual features. To achieve our goal we combine techniques from Information Retrieval, CBIR and Natural Language Processing. In all our submitted runs, the values of the fusion parameters were estimated based on analysis with the CLEF 2009 and 2010 collections [9], [1]. We also explore ways of enriching the textual queries applying relevance feedback on articles derived from the medical category of Wikipedia. In the following sections we overview our approaches and present the results of our runs. Finally conclusions are drawn with proposals for further work.

2 Data Preprocessing

This year’s collection contains a subset of PubMed Central of 231,000 images from various online magazines and journals [4]. Each record is identified by a unique **FigureID** and includes the following elements: the **title** of the corresponding article of the image, the **articleURL**, the **caption** which is a small

text accompanying the image, the **pmid** which assigns to the indexed article a number of MeSH terms and finally the image's **figureURL**. To enrich the collection with additional information we extracted from each article all the sentences with a **reference** to a specific image. To extract those textual references we used the sentence-splitter provided by LingPipe¹ Project and the HTML Parser of Jsoup². Figure 1 shows the final structure of the records in the database.

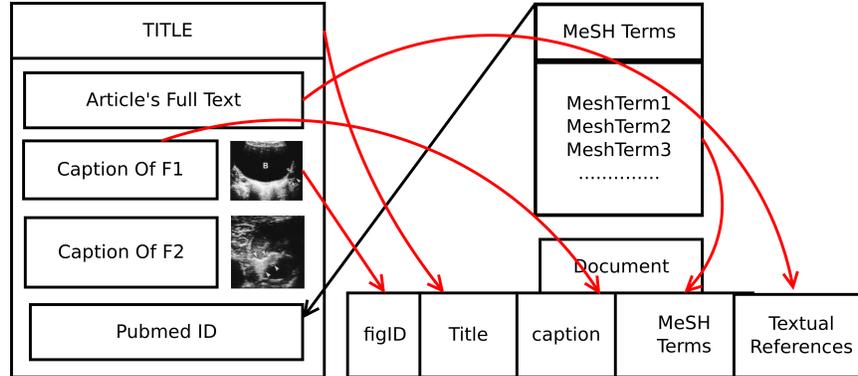


Fig. 1. The structure of the textual records

2.1 Textual Retrieval

Our retrieval system is based on the Lucene's search engine. Again this year we used Lucene³ with the Default and the BM25 scoring functions. For indexing we used our analyzer which performs tokenization, stop-words removal, short words removal (less than 2 characters) and compound word splitting. Words are transformed to lower case and suffix stripping was applied using Porter's stemmer. For the textual retrieval task we have investigated the following scoring functions [9]:

- **Default** similarity function⁴
- **BM25**, [3] and
- **BM25F**, an extension of BM25 scoring function adapted for structured documents.

¹ <http://alias-i.com/lingpipe/web/demo-sentence.html>

² <http://jsoup.org/>

³ <http://lucene.apache.org/>

⁴ <http://lucene.apache.org/java/2.4.0/api/org/apache/lucene/search/Similarity.html>

In the case of multifield retrieval weights are assigned to each field at indexing time. Since this year’s database has the same structure as the 2009 and 2010 databases we have used the same weights.

The topics are also in XML format. Each topic is identified by a unique number and is characterized as visual, semantic or mixed. Visual topics may contain more than one image.

2.2 Data Fusion

In our data fusion strategy we used a linear function defined by:

$$WeightedSUM(q, d) = \sum_i w_i Score_i(q, d) \quad (1)$$

where w_i is a weight proportional to the performance of the i -th retrieval component. In the case of CBIR since different retrieval systems generate different ranges of similarity scores, it is necessary to normalize these scores by the formulae (2).

$$NormScore_i = \frac{Score_i - MinScore}{MaxScore - MinScore} \quad (2)$$

All variables in (2) are related to a given query, q , and a given result list. MaxScore and MinScore are the maximum and minimum scores in the result list, respectively; $Score_i$ is the score that a document (image), d , obtained initially; and $NormScore_i$ is the normalized score that d should obtain.

For our runs we have used two different sets of values for the weights w_i :

- A set of values estimated empirically from the CLEF ’09 and ’10 collections.
- and a set of values $w_i = MAP_i$. The weight of each system is determined as a function of its performance [10].

A third set of values was also investigated based on a machine learning technique. Although no runs were submitted with those values we present briefly the approach and results are given in section 4.2. These weights are estimated by training a linear classifier. For training set was served the 2010 CLEF database with the 2010 topics and the set of Qrels. A training example is defined by the normalized tuple of scores $(score_1, \dots, score_k)$ together with the label of the retrieved document: +1 if it is relevant and -1 if it is nonrelevant. A value $score_i$ denotes the score of the image with respect to a field for the textual case or to a low level feature in the case of the visual retrieval. We have used a Modified Perceptron linear classifier proposed in [2]. The train set contains 1600 examples of relevant and an equal number of nonrelevant documents. The estimated weights were $W_1 = 0.05$ (for title) $W_2 = 0.31$ (for caption) $W_3 = 0.12$ (for meshterms).

2.3 Splitting Compound Words

In medical articles we encounter a large number of compound words, i.e. words that contain two or more single words. In the 2010 CLEF campaign we have

noticed a significantly low performance in queries containing compound words. For example, the query "images of dermatofibroma" returns zero documents since the word "dermatofibroma" does not occur in the text collection (the words "images" and "of" are stopwords in our system). If we split the compound word in its constituent words (derma and fibroma) the system retrieves 10 relevant documents out of 29 in total.

There are two data sources which we could take advantage of in order to modify our analyzer to deal with compound words: The UMLS⁵ (Unified Medical Language System), a set of files that brings together many health and biomedical vocabularies and the Merriam Websters Medical Dictionary⁶. We have used these two sources together with the Lucene's API, DictionaryCompoundWord-TokenFilter⁷ to split up the compound words. However, in order for this class to work, one should provide a dictionary of simple words. The two language resources were used to construct such a dictionary. Initially we construct a table of medical terms from the UMLS meta-thesauruses. Then we proceed constructing for each term in the table all the n-grams of size greater or equal to 4. From these, we remove the n-grams with frequency less than 5. For the remaining n-grams we are assisted by the online dictionary Merriam Webster (MW), which has a medical subdomain. We have developed a script which for each candidate morpheme of the table, looks it up in the MW dictionary by using a simple HTTP call. It parses the response text and decides whether or not it is a medical term. In the case of a medical term it stores the primary form of its entry. Thus from all the n-grams we keep in our dictionary only those which exist in the MW dictionary in terms of medical lemmas or morphemes in combining form. For example, the word dermatofibroma is indexed as *dermat*, *fibr*, *dermatofibroma*.

2.4 Relevance Feedback from Wikipedia

Usually the initial query reflects the user's first attempt in the process of information seeking. The initial query, however, might be more generic or more specific than it should be. As a result, the user should modify his original query to get better results. In our implementation we have applied a pseudo-relevance feedback on the top k retrieved documents using Rocchio's algorithm [6] and we select the top p terms with the the higher weight. For our runs $k = 5$ and $p = 2$.

Wikipedia is a free, web-based, collaborative, multilingual encyclopedia project and provides for public use the latest dump of its entire database. This huge amount of data can be very useful for several tasks in Information Retrieval and Natural Language Processing. In our runs we applied pseudo-relevance feedback using the Wikipedia subcollection on the medical domain to expand our initial query. Thus, the initial query is submitted to wikipedia database and we keep the top 5 documents retrieved for further analysis using Rocchio formula. Finally we select the two top terms, those with the highest weights (other than the terms in the initial query) to form the refined query.

⁵ <http://www.nlm.nih.gov/research/umls/quickstart.html>

⁶ <http://www.merriam-webster.com/>

⁷ <http://lucene.apache.org/java/3.1.0/api/all/org/apache/lucene/analysis/compound/>

3 Visual retrieval

For CBIR we have used LIRE (Lucene Image Retrieval)⁸, a light weight open source Java library [5]. Lire provides a simple way to retrieve images based on several low level features of MPEG-7. We have used several combinations of features and fusion techniques described in previous section.

4 Experimental results

4.1 Results from Textual Retrieval

In table 1 we give the definitions of our textual runs. The last part of the run's identifier denotes the similarity function used. As we have already mentioned in section 2.2 we have used two types of weights all derived from the analysis in the CLEF '09 and '10 databases.

Table 1. Definitions of IPL's runs on textual retrieval.

Run	Description
IPL2011AdHocTCM-DEFAULT IPL2011AdHocTCM-BM25	title, caption, mesh-terms all in one field.
IPL2011AdHocTC0.9-M0.1-DEFAULT IPL2011AdHocTC0.9-M0.1-BM25F	title and caption in one field with weight 0.9, and mesh-terms in a second field with weight 0.1 (same as IPL-2010 run)
IPL2011AdHocT1-C6-M0.2-DEFAULT IPL2011AdHocT1-C6-M0.2-BM25F	title, caption, mesh-terms in three fields with weights 1, 6, 0.2 respectively.
IPL2011AdHocT1-C6-M0.2-R0.01-DEFAULT IPL2011AdHocT1-C6-M0.2-R0.01-BM25F	title, caption, mesh-terms, references in 4 fields with weights 1, 6, 0.2, 0.01. These empirical values were estimated on the 2009, and 2010 databases.
IPL2011AdHocT0.113-C0.335-M0.1-DEFAULT IPL2011AdHocT0.113-C0.335-M0.1-BM25F	title, caption, mesh-terms in 3 fields with weights 0.113, 0.335, 0.1. wi= MAPi s separately on the '10 database.

From our textual results, in Table 2 it is evident that the weighted multifield retrieval outperforms retrieval in a single field. Another interesting observation was that BM25 and BM25F were worst than the default similarity function in all the cases.

4.2 Results from Visual Retrieval

Visual retrieval compiles several low level features in a linear weighted function. Our submitted visual runs with the weights of the constituent systems were the following:

⁸ <http://www.semanticmetadata.net/lire/>

Table 2. IPL’s Performance Results from Textual Retrieval

runid	MAP	P10	P20	Rprec	bpref	rel_ret
IPL2011AdHocT1-C6-M0.2-R0.01-DEFAULT	0.2145	0.4033	0.3333	0.2536	0.2434	1451
IPL2011AdHocT1-C6-M0.2-DEFAULT	0.2130	0.3567	0.3167	0.2392	0.2370	1433
IPL2011AdHocT0.113-C0.335-M0.1-DEFAULT	0.2016	0.3733	0.3183	0.2307	0.2269	1385
IPL2011AdHocTC0.9-M0.1-DEFAULT	0.1945	0.3700	0.3100	0.2227	0.2255	1380
IPL2011AdHocTCM-DEFAULT	0.1599	0.3367	0.2850	0.1799	0.1874	1336
IPL2011AdHocTC0.9-M0.1-BM25F	0.1510	0.3033	0.2633	0.1971	0.1909	1245
IPL2011AdHocT1-C6-M0.2-R0.01-BM25F	0.1492	0.3067	0.2550	0.1883	0.1848	1240
IPL2011AdHocT1-C6-M0.2-BM25F	0.1485	0.3067	0.2583	0.1882	0.1839	1233
IPL2011AdHocT0.113-C0.335-M0.1-BM25F	0.1312	0.2767	0.2433	0.1665	0.1670	1210
IPL2011AdHocTCM-BM25	0.1289	0.2867	0.2500	0.1710	0.1744	1168

- **IPL2011Visual-DECFc.** A combination of the Default, Extensive, CEDD and FCTH methods. (Default: 0.0097, Extensive: 0.0086, CEDD :0.0054, FCTH :0.0030).
- **IPL2011Visual-DEFC.** A combination of Default, Extensive, color-layout and CEDD. (Default: 0.0097, Extensive: 0.0086, FAST :0.0035, CEDD :0.0054)
- **IPL2011Visual-DEC.** A combination of Default, Extensive, CEDD. (Default: 0.0097, Extensive: 0.0086, CEDD: 0.0054)
- **ILP2011Visual-DEF.** A combination Default, Extensive, Fast (color-layout). (Default: 0.0097, Extensive: 0.0086, FAST: 0.0035)
- **ILP2011Visual-DTG.** A combination of Default, Tamura, Gabor (Default: 0.0097 Tamura: 0.0012 Gabor: 0.0002)

The weights were estimated by the MAP values from each individual feature in the 2009 database. The methods Default, Extensive and Fast define different combinations of color, texture and edge histograms.

For multi-image queries $Q = \{im_1, \dots, im_k\}$ the similarity score of an image, *Image*, was estimated by:

$$SCORE(Q, Image) = \sum_{j=1}^k score(im_j, Image) \quad (3)$$

4.3 Results from Mixed Retrieval

For mixed retrieval we have used two different approaches. One was with a linear combination of the textual and visual results, defined by:

$$SCORE(q, d) = 0.39 * score_{textual}(q, d) + 0.01 * score_{visual}(q, d) \quad (4)$$

where 0.39 was the MAP value of the textual retrieval and 0.01 the MAP of the visual retrieval on the CLEF 2009 database. The results from this approach

Table 3. IPL visual results

runid	MAP	P10	P20	Rprec	bpref	num_rel_ret
IPL2011Visual-DECFc	0.0338	0.1500	0.1317	0.0625	0.0717	717
IPL2011Visual-DEFC	0.0322	0.1467	0.1250	0.0640	0.0715	689
IPL2011Visual-DEC	0.0312	0.1433	0.1233	0.0616	0.0716	673
ILP2011Visual-DEF	0.0283	0.1367	0.1217	0.0583	0.0703	632
ILP2011Visual-DTG	0.0253	0.1333	0.1250	0.0538	0.0715	590

are listed in the first four lines in table 4. The second fusion approach was a filtering task of CBIR on a set of images retrieved from a textual query. The top 1000 documents retrieved from the textual queries are used in the CBIR. These documents are re-ranked according to their content based score. The results from this approach are listed in the last four lines in table 4. Finally in Table

Table 4. IPL mixed Results

runid	MAP	P10	P20	Rprec	bpref	rel_ret
IPL2011Mixed-DEF-T1-C6-M0.2-BM25F-0.39-0.01	0.1494	0.3067	0.2583	0.1882	0.1849	1241
IPL2011Mixed-DEF-T1-C6-M0.2-R0.01-BM25F-0.39-0.01	0.1493	0.3067	0.2550	0.1883	0.1849	1241
IPL2011Mixed-DTG-T1-C6-M0.2-R0.01-BM25F-0.39-0.01	0.1492	0.3067	0.2550	0.1883	0.1849	1240
IPL2011Mixed-DTG-T1-C6-M0.2-BM25F-0.39-0.01	0.1489	0.3067	0.2583	0.1882	0.1840	1239
ILP2011Mixed-DEF-T1-C6-M0.2-BM25F	0.0952	0.2967	0.2667	0.1332	0.1610	1209
ILP2011Mixed-DTG-T1-C6-M0.2-BM25F	0.0945	0.2700	0.2633	0.1314	0.1613	1232
ILP2011Mixed-DTG-T0.113-C0.335-M0.1-BM25F	0.0924	0.2733	0.2650	0.1299	0.1600	1209
ILP2011Mixed-DEF-T0.113-C0.335-M0.1-BM25F	0.0911	0.2733	0.2617	0.1338	0.1583	1186

5 we present our results based on the machine learning approach presented in section 2.2 from textual and visual retrieval on the databases over the past three years. Although no official runs were submitted from this approach results are promising and therefore we have included them in this working paper.

5 Conclusions and Further Work

Retrieving documents from different sources surely improves performance of information retrieval. However, efficient balancing the effect of each source to the final result needs a deeper investigation. From our experiment studies, not submitted officially this year seems that the values estimated by the machine learning technique are very promising. Further investigation, however is needed, to

Table 5. Results with weights estimated from optimization on the 2009 database queries

Textual retrieval (weights: $w_1 = 0.05, w_2 = 0.31, w_3 = 0.12$)	MAP	P10	P20	Rprec	bpref	rel_ret
2009	0.4061	0.6400	0.6120	0.4399	0.4564	1905
2010	0.3712	0.4937	0.4062	0.3836	0.8084	843
2011	0.2039	0.3933	0.3150	0.2410	0.2300	1397
Visual results (weights: $w_1 = 0.5175, w_2 = 0.3625, w_3 = 0.1198$) Default/Extensive/CEDD						
2009	0.0099	0.0328	0.0400	0.0280	0.0347	297
2010	0.0095	0.0188	0.0187	0.0162	0.0224	63
2011	0.0298	0.1533	0.1217	0.0592	0.0719	641

find out whether the training data remain linearly separable when the number of queries for training the classifier is getting larger. Another positive feature of this approach is that users may save their own queries (with positive answers) obtaining in this way a personalized classifier that will better fit to their needs.

In the medical domain, dealing with the compound words is very crucial. The medical dictionary of word lemmas and morphemes we have constructed is a positive contribution to that direction. Wikipedia is a very rich resource of data and can be used for query expansion in the medical domain.

The performance of the visual retrieval still is very poor compared to the textual retrieval. It seems that global features of the images do not have a good discrimination value. Thus techniques for image segmentation using local features may improve CBIR while keeping the complexity to acceptable levels. Finally the restriction of the CBIR retrieval to the top documents (say 1000) returned by a textual query makes CBIR scalable to large image collections.

References

1. Boutsis, I., Kalamboukis, T.: Combined content-based and semantic image retrieval. In: Working Notes of the 2009 CLEF Workshop (2009)
2. Gkanogiannis, A., Kalamboukis, T.: A perceptron-like linear supervised algorithm for text classification. In: Proceedings of the 6th international conference on Advanced data mining and applications: Part I. pp. 86–97. ADMA'10, Springer-Verlag, Berlin, Heidelberg (2010)
3. Jones, K.S., Walker, S., Robertson, S.E.: A probabilistic model of information retrieval: development and comparative experiments. *Inf. Process. Manage.* 36, 779–808 (November 2000)
4. Kalpathy-Cramer, J., Müller, H., Bedrick, S., Eggel, I., de Herrera, A.G.S., Tsirikas, T.: The clef 2011 medical image retrieval and classification tasks. In: CLEF 2011 working notes (2011)

5. Lux, M., Chatzichristofis, S.A.: Lire: lucene image retrieval: an extensible java cbir library. In: Proceeding of the 16th ACM international conference on Multimedia. pp. 1085–1088. MM '08, ACM, New York, NY, USA (2008)
6. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to information retrieval. Cambridge University Press (2008)
7. Müller, H., Kalpathy-Cramer, J., Eggel, I., Bedrick, S., Radhouani, S., Bakke, B., Kahn, Jr., C.E., Hersh, W.: Overview of the clef 2009 medical image retrieval track. In: Proceedings of the 10th international conference on Cross-language evaluation forum: multimedia experiments. pp. 72–84. CLEF'09, Springer-Verlag, Berlin, Heidelberg (2010)
8. Müller, H., Kalpathy-Cramer, J., Eggel, I., Bedrick, S., Said, R., Bakke, B., Kahn, Jr., C.E., Hersh, W.: Overview of the clef 2010 medical image retrieval track. In: Working Notes of CLEF 2010 (Cross Language Evaluation Forum). Padua Italy (September 2010)
9. Stougiannis, A., Gkanogiannis, A., Kalamboukis, T.: Ipl at imageclef 2010. In: CLEF (Notebook Papers/LABs/Workshops) (2010)
10. Wu, S., Bi, Y., Zeng, X., Han, L.: Assigning appropriate weights for the linear combination data fusion method in information retrieval. *Inf. Process. Manage.* 45, 413–426 (July 2009)