DEU at ImageCLEFmed 2009: Evaluating Re-Ranking and Integrated Retrieval Model

Tolga BERBER, Adil ALPKOÇAK
Dokuz Eylul University, Department of Computer Engineering
{tberber,alpkocak}@cs.deu.edu.tr

Abstract

This paper presents DEU team participation in imageCLEF2009med. Main goal of our participation is to evaluate two different approaches: First, a new re-ranking method which aims to model information system behaviour depending on several aspects of both documents and query. Secondly, we compare a new retrieval approach which integrates textual and visual contents into one single model. Initially we extract textual features of all documents using standard vector space model and assume as a baseline. Then this baseline is combined with re-ranking and integrated retrieval model. Re-ranking approach is trained using ImageCLEFmed 2008 ground truth data. However, re-ranking approach did not produced satisfactory results. On the other hand, our integrated retrieval model resulted top rank among all submissions in automatic mixed runs.

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.3.7 Digital Libraries

General Terms

Information Retrieval, Re-Ranking, Content-Based Image Retrieval

Keywords

Re-Ranking, Information Retrieval, Machine Learning, Content-Based Image Retrieval, ImageCLEF Medical Retrieval Task.

1 Introduction

This paper describes our work on medical retrieval task of ImageCLEF 2009 track. The main goal of our participation is to evaluate two new approaches; one of them is a new re-ranking approach and other one is a new integrated retrieval model which integrates both textual and visual contents into one single model.

Proposed re-ranking method considers several aspects of both document and query (e.g. degree of query/document generality, document/query length etc.) in re-ranking. System learns to re-rank of initially retrieved documents, using all these features. System is trained with ground truth data of imageCLEFmed 2008. Second approach we present is an integrated retrieval system which extends textual document vectors with visual terms. The method aims to close well-known semantic gap problem in content-based image retrieval by mapping low-level image features to high level semantics. Moreover, this combination of textual and visual modality into one also helps to query a textual database with visual content or, visual database with textual content. Consequently, images could be defined semantic concepts instead of low-level features.

Rest of this paper is organized as follows. In Section 2, our retrieval framework is described. Section 3 gives results of both methods on imageCLEFmed 2009 dataset. Finally, Section 4 concludes the paper and gives a look at future work.

2 Retrieval Framework

Our retrieval framework contains two new methods based on classical vector space model (VSM).In VSM a document is represented as a vector of terms. Hence, a document repository, D, becomes a sparse matrix whose rows are document vectors, columns are term vectors.

$$D = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m1} & w_{m2} & \cdots & w_{mn} \end{bmatrix} \tag{1}$$

where w_{ij} is the weight of term j in document i, n is term count, m is document count. Literature proposes a plenty number of weighting schemes. We used pivoted unique term weighting scheme [4]. In general, definition of term weighting scheme is shown below.

$$w_{ij} = L_{ij} G_i N_j \tag{2}$$

where w_{ij} is term weight, L_{ij} is local weight for term i in document j, G_i is the global term weight for term i in whole document collection and N_j is the normalization factor for document j. In classical *tf*idf* weighting scheme, N_j is always 1. Actually, this assumption is not the best case for all situations. So, in our work we prefer to use pivoted unique term weighting scheme.

$$w_{ij} = \frac{\log(dtf) + 1}{sumdtf} \times \log\left(\frac{N - nf}{nf}\right) \times \frac{U}{1 + 0.0115U} \tag{3}$$

where dtf is the number of the times term appear in document, $sumdtf$ is sum of $\log(dtf) + 1$ for document, N is the total number of documents, nf is number of terms containing the term and U is the number of unique terms in the document [1].

2.1 Re-Ranking Approach

After generating initial results, re-ranking phase re-orders initially generated results to obtain high precision at top ranks. Re-ranking phase includes training. So, it first requires to extract features from both documents and query [2]. Table 1 has a list of used features in two groups.

We evaluated some machine learning algorithms to use in training phase. Table 2 shows results of 10-fold cross-validation of each method. Because C4.5 Decision tree generation algorithm shown best performance among the other methods, we chosen to use C4.5 in training phase [3]. After all, new ranks, R_{new}, of all documents in the initial results is recalculated using following formula:

$$R_{new} = \begin{cases} \delta + \alpha R_{initial} & \text{; if document is relevant} \\ R_{initial} & \text{; if document is not relevant} \end{cases} \tag{4}$$

where δ and α are shift and bias factors; $R_{initial}$ and R_{new} are initial and newly calculated ranks of the documents, respectively.

	Textual Features	
Feature Name	Scope	Feature Definition
Number of Matched Terms	Document	$\sum_{t \in d \cap q} 1$
Number of Bi-Word Matches	Document	$\sum_{t_i, t_j \in d \cap q} 1, i \neq j, j = i - 1$
Term Count	Document	$\sum_{t \in d} 1$
Term Count	Query	$\sum_{t \in q} 1$
Generality	Document	$\sum_{t \in d} idf(t)$
Generality	Query	$\sum_{t \in q} idf(t)$
Depth of Indexing	Document	$\sum_{t' \in d'} 1, d'$ unique terms of d
Relevance Score	Document	$\vec{d} \cdot \vec{q}$
Initial Rank	Document	-
	Visual Features	
Feature Name	Scope	Feature Definition
Grayscaleness	Document	$P(G)$
Avg. Grayscaleness	Query	$\frac{1}{N_i} \sum_{i=1}^{N_i} P(G)$
Color Amount	Document	$P(C)$
Avg. Color Amount	Query	$\frac{1}{N_i} \sum_{i=1}^{N_i} P(C)$

Table 1: Features used to re-rank documents, where d is term set of document, q is term set of the query, $P(G)$ is probability of being grayscale image, $P(C)$ is probability of being color image.

Method	Class	TP (%)	FP (%)	Precision	Recall	F	ROC
C4.5	Not Rel.	0,969	0,575	0,916	0,969	0,942	0,836
	Relevants	0,425	0,031	0,678	0,425	0,523	0,836
	Total	0,896	0,502	0,885	0,896	0,886	0,836
Neural Network	Not Relevants	0,982	0,807	0,888	0,982	0,933	0,794
	Relevants	0,193	0,018	0,624	0,193	0,295	0,794
	Total	0,877	0,702	0,853	0,877	0,847	0,794
Naive Bayes	Not Relevants	0,960	0,810	0,885	0,960	0,921	0,674
	Relevants	0,190	0,040	0,421	0,190	0,262	0,674
	Total	0,857	0,708	0,823	0,857	0,833	0,674

Table 2: Results of 10-fold Cross-Validation for Machine Learning Algorithms.

2.2 Integrated Retrieval Model

Our second approach focuses on closing semantic gap problem of content-based image retrieval. Our approach aims to integrate both textual and visual contents in same space. System proposes a modification on D matrix (Eq. 1) by adding visual terms representing visual contents. Formally, document-term matrix becomes as follows:

$$D' = \left[\begin{array}{cccc|cccc} w_{1,1} & w_{1,2} & \cdots & w_{1,n} & i_{1,n+1} & i_{1,n+2} & \cdots & i_{1,n+k} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,n} & i_{1,n+1} & i_{2,n+2} & \cdots & i_{2,n+k} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ w_{m,1} & w_{m,2} & \cdots & w_{m,n} & i_{m,n+1} & i_{m,n+2} & \cdots & i_{m,n+k} \end{array} \right] \quad (5)$$

where i_{ij} is the weight of image term j in document i, k is the number of visual terms. Visual and textual features are normalized independently. In sum, Integrated Retrieval Model (IRM) combines visual features to traditional text-based VSM. Initially, we chosen to use two simple visual terms which aim to model color information of the whole image. In Algorithm 1 extraction of one term is given. Algorithm counts pixels that has same pixel value in each color channel. This visual term determines what amount of image is grayscale. Second visual term is the complement of grayscaleness term. In other words, second visual term determines what amount of image is coloured.

Algorithm 1: Grayscaleness Extraction Algorithm

```
Input  : Image Pixels
Output: Probability of being grayscale
begin
    count ← 0
    channelcount ← Channel count of Image
    if channelcount=1 then
        return 1.0
    end
    if channelcount=3 then
        for $i = 1$ to image height do
            for $j = 1$ to image width do
                if $(Image(i, j, 0) = Image(i, j, 1)) \wedge (Image(i, j, 1) = Image(i, j, 2))$ then
                    count ← count + 1
                end
            end
        end
    end
    return count/totalpixelcount
end
```

3 Experimentation

In this section, we describe experimentations performed in ImageCLEFmed 2009. Our experimentations can be classified into two groups. The first one is about re-ranking and second one is about IRM. Before going further, we performed a preprocessing in all 74902 documents including combination of title and captions. First, all documents were converted to lower-case to achieve uniformity. Numbers in the documents are not removed. However, some punctuation characters like dash(-) and apostrophe(') is removed, others like comma(,), slash(/) is replaced with a space. For example, dash(-) character is removed because of the importance of terms like x-ray, T3-MR etc. We chosen words surrounded by spaces as index terms. Finally we had 33613 index terms. These index terms are normalized as described in Eq. 3. We evaluated performance of this baseline method and Figure 2 shows results of this run.

Experimentations of Re-Ranking method includes training phase where we use imageCLEFmed 2008 data. Training data contains titles and figure captions of approximately 67000 images in medical articles. Number of index terms is 30343 and same term generation technique with baseline method is used. ImageCLEFmed 2008 dataset also contains 30 queries and their relevance data. These 30 queries categorized under 3 groups; visual queries which results will be better by using visual content, mixed queries which are prepared to test performance of mixed retrieval systems and textual queries which targets to textual retrieval systems. Experimentation results of our re-ranking approach on both base-line and IRM on imgeCLEFmed 2008 data is given in Table 3 and Table 4 respectively. According to imageCLEFmed 2008 results our re-ranking approach boosts performance of both methods. However, results of imageCLEFmed 2009 did not show same impact. According to Table 5, proposed re-ranking technique reduces system performance about 6 %. Reason of this loss lies in the difference between training and evaluation datasets. Since we used ground truth data of imageCLEFmed 2008 data, system learns relevance information of imageCLEFmed 2008 dataset only.

	Query Type	MAP	P@5	P@10	P@20	P@100	P@500	P@1000	bpref
Base Line	Visual	0.197	0.320	0.290	0.235	0.142	0.089	0.058	0.278
	Mixed	0.113	0.280	0.240	0.220	0.188	0.082	0.051	0.200
	Textual	0.291	0.340	0.300	0.290	0.208	0.092	0.055	0.385
	Avg.	0.200	0.313	0.277	0.248	0.179	0.088	0.055	0.288
Re-Rank	Visual	0.248	0.420	0.420	0.370	0.211	0.107	0.061	0.368
	Mixed	0.139	0.440	0.410	0.360	0.219	0.089	0.055	0.263
	Textual	0.324	0.560	0.460	0.430	0.228	0.093	0.057	0.425
	Avg.	0.237	0.473	0.430	0.387	0.219	0.096	0.058	0.352

Table 3: Performance of Our Re-Ranking Approach on Base-Line approach using ImageCLEFmed 2008 data.

Integrated retrieval model experimentation needs image features to be extracted first. As mentioned before, we used simple grayscaleness feature. Table 4 presents results of our integrated model on ImageCLEFmed 2008 dataset. Whereas our model shows similar performance with baseline method on visual queries, it shows better performance on other two query types. Results of our integrated model can be improved by using re-ranking algorithm and combination of two methods shows the best performance in all measures.

	Query Type	MAP	P@5	P@10	P@20	P@100	P@500	P@1000	bpref
I-VSM	Visual	0.190	0.280	0.290	0.240	0.145	0.076	0.051	0.268
	Mixed	0.130	0.360	0.320	0.290	0.200	0.088	0.054	0.216
	Textual	0.312	0.380	0.320	0.365	0.240	0.096	0.059	0.423
	Avg.	0.211	0.340	0.310	0.298	0.195	0.087	0.054	0.303
Re-Rank	Visual	0.259	0.400	0.390	0.365	0.221	0.117	0.068	0.397
	Mixed	0.160	0.480	0.420	0.390	0.248	0.095	0.059	0.351
	Textual	0.384	0.620	0.550	0.540	0.265	0.098	0.060	0.528
	Avg.	0.268	0.500	0.453	0.432	0.245	0.103	0.062	0.425

Table 4: Performance of Our Mixed Retrieval System and Re-Ranking on Our Mixed Retrieval System using ImageCLEFmed 2008 data.

In meaning of precision our methods outperforms baseline. In Figure 1 performance of all evaluated methods based on recall-precision scale in imageCLEF2008med dataset is given. According to the figure, our re-ranking method increases recall levels when it is applied to both of the approaches, base-line and integrated retrieval. Since results of integrated retrieval model is better than baseline method, Re-Ranking performance of mixed retrieval shows the best recall levels in all cases.

We participated ImageCLEFmed 2009 with 5 official and we evaluated our re-ranking approach on IRM run unofficially. In Table 5, performance of all our methods is given on ImageCLEFmed 2009 dataset. Our Integrated VSM approach shows best performance in all measures among others. Only Re-Ranked results of IRM run shows same performance at P@5 scale. Since simple feature is used as a visual term, performance of the approach will be expected to improve with new features. As mentioned before our re-ranking approach reduces system performance according to the ImageCLEFmed 2009 results. In Figure 2 Precision and recall graph of all our methods is given. According to the figure our IRM outperforms all of our runs in means of recall at all precision levels.

Our Integrated VSM approach has best scores in automatic mixed retrieval area of the ImageCLEFmed 2009 results. In Table 6 official results of automatic mixed retrieval area is given. Our Integreated VSM technique has the highest MAP score of Mixed Automatic runs.

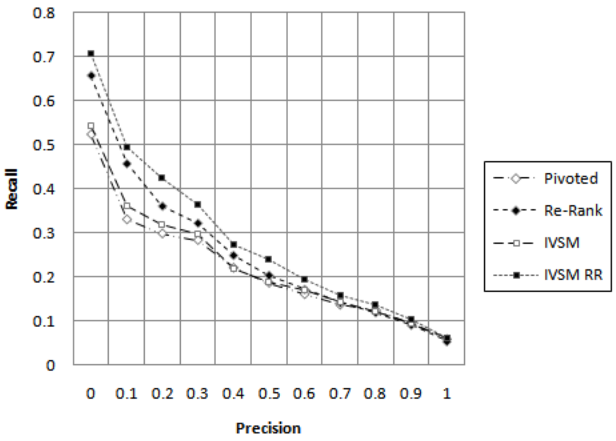

Figure 1: Precision-Recall Graph of All Methods on ImageCLEFmed 2008 Dataset.

Run Identifier	NumRel	RelRet	MAP	P@5	P@10	P@30	P@100
deu_traditionalVSM	2362	1620	0.310	0.608	0.528	0.451	0.296
deu_traditionalVSM_rerank	2362	1615	0.286	0.592	0.508	0.457	0.294
deu_baseline	2362	1742	0.339	0.584	0.520	0.448	0.303
deu_baseline_rerank	2362	1570	0.282	0.592	0.516	0.417	0.271
deu_IRM	2362	**1754**	**0.368**	**0.632**	**0.544**	**0.483**	**0.324**
deu_IRM_rerank	2362	1629	0.307	**0.632**	0.528	0.448	0.272

Table 5: Results of Our Methods on ImageCLEFmed 2009 Dataset.

Run Identifier	RelRet	MAP	P@5	P@10	P@100
DEU IRM	1754	0.3682	0.632	0.544	0.3244
York University_79_8_1244834388965	1724	0.3586	0.584	0.584	0.3312
York University_79_8_1244834655420	1722	0.3544	0.624	0.572	0.3244
York University_79_8_1244834554642	1763	0.3516	0.6	0.592	0.3356
York University_79_8_1244834740798	1719	0.3375	0.592	0.568	0.308
York University_79_8_1244834823312	1757	0.3272	0.592	0.568	0.308
medGIFT_77_8_1244842980151	1176	0.29	0.632	0.604	0.2924
ITI_26_8_1244811028909	1553	0.2732	0.488	0.52	0.2648
University of North Texas_55_8_1244879759190	1659	0.2447	0.456	0.404	0.258
medGIFT_77_8_1244752959441	848	0.2097	0.704	0.592	0.2128

Table 6: Performance of Top 10 Official Runs for Mixed Automatic Retrieval.

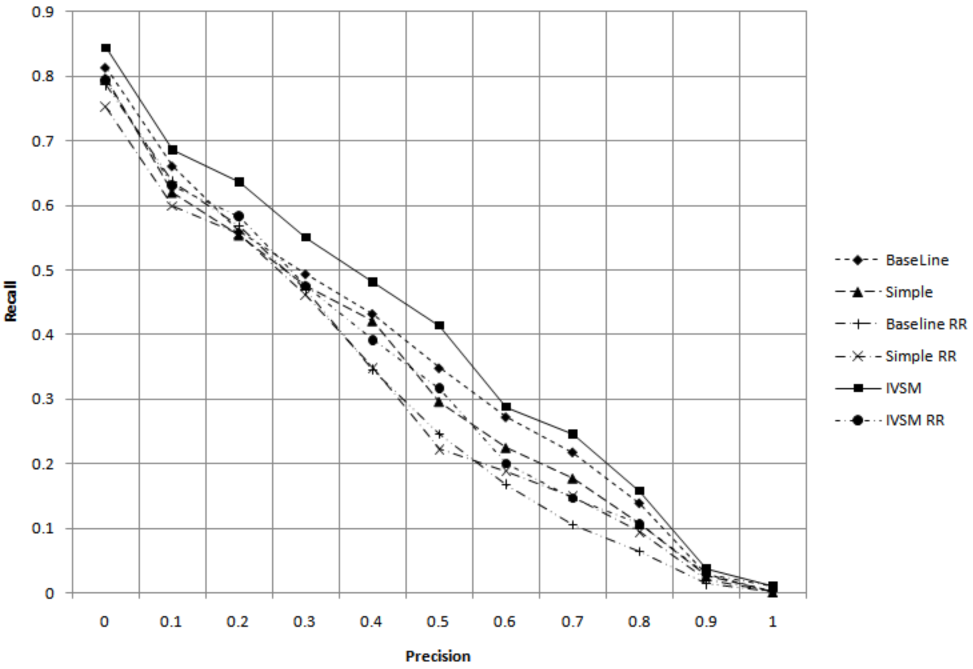

Figure 2: Precision-Recall Graph of Top-5 Automatic Mixed Runs of ImageCLEFmed 2009.

4 Discussion and Future Work

In this paper, we summarize our participation to imageCLEFmed task. In this study, we propose and evaluate two new methods: first method is a re-ranking method which considers several aspects of both document and query such as degree of query/document generality, document/query length etc. System learns to re-rank of initially retrieved documents, using all such features. Ground truth data of imageCLEF 2008 used to train re-ranking system. Second method we present is an integrated retrieval system which extends textual document vectors with visual terms.

Results of ImageCLEFmed 2009 shows that proposed re-ranking approach boosts performance of information retrieval system if modeled dataset is not subject to change. But proposed technique could modified to adopt itself to dataset changes. As a result, we obtain an adaptive re-ranking system. On the other hand, experiments on our second proposal that integrates both textual and visual content ranked first among all submissions for mixed automatic runs.

In this study we proposed an integrated model that ultimately aims to close semantic gap in visual information retrieval. We used a single visual term, however results are satisfactory. In sum, it is promising that usage of visual term gives better results than most of the textual only models.

Acknowledgements

This work is supported by Turkish National Science Foundation (TÜBİTAK) under project number 107E217.

References

[1] Erica Chisholm and Tamara G. Kolda. New term weighting formulas for the vector space method in information retrieval. Technical report, 1999.

[2] Tao Qin, Tie-Yan Liu, Jun Xu, and Hang Li. How to make letor more useful and reliable. In *Proceedings of Learning to Rank for Information Retrieval Workshop*, pages 52–58. ACM, ACM, 2008.

[3] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, 1993.

[4] Amit Singhal, Chris Buckley, and Mandar Mitra. Pivoted document length normalization. In *SIGIR '96: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 21–29, New York, NY, USA, 1996. ACM.