AUDR at ImageCLEF2011: medical retrieval task

Junwu Luo[1] , Xianglong Liu[1], Han Wang[1], Bo Lang[1]

[1] Department of Computer Science and Technology, Beihang University,
100191 Beijing, China
luojunwu@nlsde.buaa.edu.cn

Abstract. This paper describes the participation of AUDR group in the ImageCLEF 2011 medical retrieval task. We are particularly interested in the adhoc retrieval task. Two existing retrieval engines are used: LIRE for visual retrieval and Apache Lucene for textual retrieval. Based on the two tools, we consider two information models for text: vector space model and topic model, and the query expasion mechansim based on KL distance is used for improving the performance. Concerning the visual retrieval, a purely visual approach named CEDD is used with LIRE. Fusion strategies are also tried out to combine results from two engines. The experiments carried out on the ImageCLEFmed datasets show baselines provided by LIRE and Lucene are ranked close to the average among visual and textual runs respectively, and the fusion of vector spacing model and topic model will perform better than mono-model.

Keywords: ImageCLEF, medical image retrieval, topic model, query expasion

1 Introduction

This paper describes the contribution of the AUDR team in its first year participation at the medical retrieval track.

ImageCLEF is the cross-language image retrieval track of the Cross Language Evaluation Forum. ImageCLEFmed is part of ImageCLEF focusing on medical images. The ImageCLEFmed 2011[1] task consists of three subtasks: modality classification, ad-hoc image-based retrieval and case-based retrieval. We are particularly interested in the ad-hoc image-based retrieval task, which is the classic medical retrieval task, similar to those in organized in 2005-2010. Participants will be given a set of 30 textual queries with 2-3 sample images for each query. The queries will be classified into visual, textual and mixed, based on the methods that are expected to yield the best results.

Concerning the textual retrieval, two information models are considered: vector space model[2] and topic model[3]. The fusion of this two model has shown good performances. We also aim at improving it combined with query expasion based on KL distance[4], but the results are lower than expected yet reveal new insights which we will improve in the future. The visual baseline is based on LIRE[5], whereas Apache Lucene[6] is used for textual baseline.

The rest of this paper is organized as follows: In Section 2 we describe our visual retrieval methods. In Section 3 we briefly describe textual retrieval techniques. Fusion approaches are presented in section 4. In section 5 Submitted results will be compared and discussed and we conclude in section 6.

2 Visual Retrieval

LIRE is a light-weight visual library for content-based image retrieval(CBIR), three of the available image features are taken from the MPEG-7 standard: ScalableColor, ColorLayout and EdgeHistogram. Futhermore LIRE provides other global features such as CEDD and FCTH, local feature such as SIFT.

Research related shows composite feature model used to describe the image content is more effective than a mono-model. CEDD feature implemented by LIRE[7], which incorporates color and texture information in a histogram, is used for indexing and retrieval in the task. It outputs a vector of 144 bins and the value of every bin is between 0 and 7.

Based on LIRE Java library, we develop a CBIR engine and use the kNN search techniques. For the measurement of the distance of CEDD feature between the images, Tanimoto coefficient is adapted.The system allows query with multiple input images and uses combSUM fusion strategy[8] for scoring.

3 Textual Retrieval

The text retrieval approaches are all based on the bag-of-words model, a text (such as image caption or full-article) is represented as an unordered collection of words after tokenization and standard stopword removal. After the pre-processing step, two information model are considered: vector space model and topic model. We also use a query expansion mechanism, pesudo relevance feedback based on information-theoretic. We will briefly describe the details specific to the techniques mentioned above.

3.1 Vector Space Model

Vector Space Model is a standard language model and is widely used in text retrieval. Document D_j can be represented as an N-dimensional feature vector using VSM.

$$V(D_i) =< (t_1, w_{1i}), (t_2, w_{2i}), \cdots, (t_m, w_{mi}) > \tag{1}$$

where w_{ji} denotes the weight (tf-idf value) of keyword t_k in document D_i.

The keywords are obtained from document after tokenizaiton and stopword removal. Image caption and full-article are indexed and scored separately. All the approaches are based on Lucene using standard settings. The scoring formula used in Lucene is as follows:

$$score(D,Q) = coord(Q,D) \times queryNorm(Q) \times \sum_{t\ in\ Q} (tf(t \text{ in D}) \times idf(t)^2 \times t.getBoost() \times norm(t,D)) \quad (2)$$

This formula can be derived from cosine distance between two vectors, whereas lucene has added some boosts and coords into it.

We also integrate the BM25 scoring approach[9] into Lucene for comparing with the standard scoring method of Lucene. The scoring formula of BM25 is as follows:

$$score(D,Q) = \sum_{t \in Q \cap D} IDF(t) \frac{tf(t,D) \cdot (k_1 + 1)}{tf(t,D) + k_1 \cdot (1 - b + b \cdot \frac{dl}{avgdl})} \quad (3)$$

where $f(t,D)$ is the number of occurrences of term t in document , k_1 = 3.8, b = 0.67 are the constants used in the experiments, *dl* is the length of document D, *avgdl* is the average length of all documents, the inverse document frequency (*IDF*) is computed as follows:

$$W_i = IDF(t) = \log \frac{N - n(t) + 0.5}{n(t) + 0.5} \quad (4)$$

where *N* is the number of documents in the collection and *n(t)* is the number of documents where the term *t* appears.

3.2 Topic Model

Topic model is a type of statistical model for discovering the abstract topics that occur in a collection of documents. Latent Dirichlet allocation (LDA), perhaps the most common topic model currently in use, allows documents to have a mixture of topics and and each word's creation is attributable to one of the document's topics.

The formual of monadic language model is as follows:

$$p(w \mid d) = \sum_{i=1}^{N} p(w \mid t_i) * p(t_i \mid d) \quad (5)$$

It's a Bayes chain, involved two distribution, topic~word and document~topic. LDA uses multinomial model to describe the topic~word and Dirichlet distribution to describe the document~topic.

3.3 Query Expasion

We use the concept the Kullback-Liebler Divergence to compute the divergence between the probability distributions of terms in the whole collection and in the top ranked documents obtained for a first pass retrieval using the original user query. The most likely terms to expand the query are those with a high probability in the top ranked set and low probability in the whole collection. For the term *t* the divergence is:

$$w(t) = P_R(t) \log \frac{P_R(t)}{P_C(t)} \tag{6}$$

where $P_R(t)$ is the probability of the term t in the top ranked documents, and $P_C(t)$ is the probability of the term t in the whole collection.

4 Fusion Techniques

Study[10] shows that combSUM and combMNZ proposed in 1994 are robust fusion strategies. The difference between the two techniques is small and not statistically significant.

$$S_{combSUM}(i) = \sum_{k=1}^{N_k} \overline{S_k(i)} \tag{7}$$

$$S_{combMNZ}(i) = F(i) \times S_{combSUM}(i) \tag{8}$$

where $F(i)$ is the freqence of image i being returned by one input system with a non–zero score, and $S(i)$ is the score assigned to image i.

In ImageCLEFmed2011, the fusion approach was used in three cases:

—fusing results of various query images in a visual run (combSUM was used)

—fusing results of various text model in a textual run (combSUM was used)

—fusing textual and visual runs to produce mixed runs (combMNZ was used)

5 Results and Discussions

This section describes our results for adhoc retrieval task. For this task, there are total 130 results from all participants submitted: 64 textual retrieval runs, 26 visual runs and 40 mixed runs combining textual and visual information. In total 8 runs were submitted by the AUDR group, they are 5 baselines (1 visual baseline and 4 textual baselines), 3 mixed runs.

For visual run, CEDD method is adapted, and for textual run, the methods consist of four scenarios as follows, and mixed run are produced using the combMNZ approach combining the visual results and textural results.

1. lucene based on image caption information, denoted lucene_caption.
2. lucene and Topic Model based on image caption and full articles information, denoted lucene_and_tm.
3. lucene integrated the BM25 scoring approach and Topic Model based on image caption and full articles information, denoted BM25_and_tm.
4. lucene and Topic Model based on image caption and full articles information, query expasion based on KL distance, denoted kld_ lucene_and_tm.

Results are shown in Table1. Mean average precision (MAP), and early precision (P10, P20), binary preference (Bpref) are used as evaluation indicators.

Table 1. Results of the runs for adhoc retrieval task

Method	run_type	MAP	P10	P20	bpref	num_rel_ret
best textual run	***Textual***	***0.2172***	***0.3467***	***0.3017***	***0.2402***	***1471***
lucene_and_tm	Textual	0.1917	0.34	0.305	0.2237	1447
lucene_caption	Textual	0.1758	0.3133	0.27	0.2187	1206
BM25_and_tm	Textual	0.0878	0.19	0.165	0.125	690
kld_ lucene_and_tm	Textual	0.0811	0.1733	0.17	0.1191	694
best visual run	***Visual***	***0.0338***	***0.15***	***0.1317***	***0.0717***	***717***
cedd	visual	0.0082	0.04	0.05	0.0427	466
best mixed run	***Mixed***	***0.2372***	***0.3933***	***0.355***	***0.2738***	***1597***
cedd_ lucene_and_tm	mixed	0.0556	0.1767	0.155	0.1018	553
cedd_kld_ lucene_and_tm	mixed	0.0341	0.1633	0.1217	0.072	496
cedd_BM25_and_tm	mixed	0.0328	0.15	0.1217	0.0719	496

In terms of MAP, the best textual run(0.2172) outperforms the best visual run (0.0338) by a factor of 6, which shows a big gap between the two approaches. The performance of the baseline produced by CEDD is slightly below the averages (rank 17/26), and the performance of the baseline produced by lucene based on image caption information is slightly above the averages (rank 28/64).

The best result in our textual runs is produced by lucene_and_tm (rank 17/64), we combined the standard methods of lucene with topic model, in contrast to lucene_caption, we take the full articles information into consideration, and use the topic model to discover the abstract topics, so the results produced by lucene_and_tm are more semantic related. Other results in our textual runs are not ideal for the value of some parameters in our methods must be set after a larger number of experiments, but here we just set the default value of these paremeters empirically.

Whereas the performance of mixed runs merging textual information with visual information has reduced the performance of textual run. The results depend largely on the fusion strategies, here we just use linear weighted methods to rescore the results, and the factor of textual run and visual run is 0.8 and 0.2 respectively, the default values are also empirical. Maybe we could enlarge the textual weights to improve the performance.

6 Conclusions

This article describes the participation of AUDR group in the ImageCLEF 2011 medical retrieval task. The results obtained by our submitted runs prove that baselines provided by LIRE and Apache Lucene are ranked close to the average among visual and textual runs respectively, and the fusion of vector spacing model and topic model will perform better than mono-model. But the pure visual based retrieval leads to poor result for semantic gap, so the performance of mixed runs merging textual with visual has reduced the performance of textual run.

However, the retrieval performance can be better improved by state-of-the-art query expasion techniques such as specific terminologies, and reducing the semantic gap is another direction we could do some research.

References

1. Kalpathy-Cramer, J., M¨uller, H., Bedrick, S., Eggel, I., de Herrera, A.G.S.,Tsikrika, T.: The CLEF 2011 medical image retrieval and classification tasks. In: CLEF 2011 working notes, Amsterdam, The Netherlands, Springer (September 2011).
2. Gerard Salton. A. Wong. And C.S. Yang. A vector space model for information retrieval. Journal of the American Society for Information Science.18(11): 613-620.November 1975.
3. D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. Journal of Machine Learning Research, 3:993–1022, 2003.
4. Thomas M. Cover and Joy A. Thomas. Elements of information theory. Wiley-Interscience, New York, NY, USA, 1991.
5. M. Lux.LIRe: Lucene image retrival an extensible java CBIR library[J]. Proceedings of the 16th ACM International Conference on Multimedia, 2008, p1085-1088.
6. http://lucene.apache.org/
7. S.A.Chatzichristofis. Cedd: Color and edge directivity descriptor. a compact descriptor for image indexing and retrieval[J].Proceedings of the 6th International Conference on Computer Vision Systems, 2008, volume 5008 of LNCS,p 312–322.
8. Fox, E.A., Shaw, J.A.: Combination of multiple searches. In: Text REtrieval Conference. (1993) 243–252
9. Perez-Iglesias J, Perez-Aguera JR, Fresno V, Feinstein YZ. Integrating the Probabilistic Models BM25/BM25F into Lucene. Computer Research Repository (CoRR), abs/0911.5046 (2009)
10. Zhou, X., Depeursinge, A., M¨uller, H.: Information fusion for combining visual and textual image retrieval. In: Pattern Recognition, International Conference on, Los Alamitos, CA, USA, IEEE Computer Society (2010)