IRIT at ImageCLEF 2011: medical retrieval task

Duy Dinh, Lynda Tamine

IRIT laboratory - University of Toulouse,
118 route de Narbonne, 31062 Toulouse, France
{Duy.Dinh,Lynda.Tamine}@irit.fr

Abstract. In this paper, we reported some experiments conducted by our members in the SIG team at the IRIT laboratory in the University of Toulouse within the context of the medical information retrieval (IR) task. As in our previous participation in ImageCLEF, in 2011, our research focuses on the *case-based retrieval* task. We compared the performance of different state-of-the-art term weighting models for retrieving *patient cases* that might best suit the clinical information need. Furthermore, we also combined term scores obtained by two state-of-the-art weighting models using a particular data fusion technique. Finally, a state-of-the-art query expansion (QE) technique is used for improving biomedical IR performance.

Key words: Biomedical information retrieval, Term weighting models, Query expansion, Model fusion

1 Introduction

This paper describes the contribution of the SIG team (Generalized Information Systems) at the IRIT[1] (Institute for Research in Informatics of Toulouse) laboratory in its second year participation at the medical retrieval track. We focused in particular on the case-based retrieval task, where patient demographics, limited symptoms and test results are provided to answer the medical professionals' information need [1].

We first investigate the effectiveness of two different state-of-the-art term weighting models that have been shown to work well in the past: LGD (a log logistic model) [2], In_expB2 (Inverse Expected Document Frequency model with the Bernoulli ratio normalisation) [3]. These models are then combined using a particular data fusion technique [4]. Next, we experiment with a state-of-the-art pseudo or blind feedback query expansion algorithm implemented in the DFR framework [3].

The rest of this paper is organized as follows: Section 2 describes our indexing and retrieval framework. Experimental results will be presented and discussed in section 3. We conclude the paper in section 4 and outline some perspectives for our future work.

[1] http://www.irit.fr

2 Indexing and retrieval framework

The indexing aims to organize, structure and store statistical and/or linguistic information about terms and documents in the collection allowing a rapid and efficient search. We use the Terrier IR platform for indexing documents [5]: stop-words are removed from documents and queries before stemming using the Porter algorithm [6].

The document retrieval aims to match the user query and document representations in order to retrieve a list of results that may satisfy the user information need. In our work, a document D containing terms used for formulating query Q is weighted by summing the score of each term figuring in document D:

$$RSV(D,Q) = \sum_{t \in Q} score(t \in D) \tag{1}$$

where $score(t \in D)$ is the query term weight calculated using a particular term weighting model. In this section, we first describe two different term weighting models used in our experiments, namely LGD [7] and In_expB2 [3]. These models are fused to obtain a combined score for each query term figuring in documents. We then applied a state-of-the-art pseudo-relevance feedback technique in order to improve the information retrieval (IR) performance.

2.1 The LGD model

In the LGD model, query terms are weighted using the log logistic distribution [7]. Formally:

$$score_{LGD}(t \in D) = qtf \times \left[log_2(\frac{N_t}{N} + tf_n) - log_2(\frac{N_t}{N}) \right] \tag{2}$$

where

- t is a query term occurring in document D,
- N_t is the document frequency (i.e., number of documents containing term t),
- N is the total number of documents in the collection,
- qtf is the query term frequency,
- tf_n is the normalised within-document term frequency, given by:

$$tf_n = tf \times log_2(1 + c \times \frac{avg\_dl}{dl}) \tag{3}$$

where avg_dl is the average document length (in tokens), dl is the document length (in tokens) and c is a multiplying factor or tuning parameter.

2.2 The In_expB2 model

For the In_expB2 model, query terms are weighted using the Inverse Expected Document Frequency model with Bernoulli after-effect and term frequency normalisation [3]. Formally:

$$score_{In\_expB2}(t \in D) = \frac{qtf \times (tf+1) \times tfn_2}{N_t \times (tfn_2+1) \times ln2} \times \log_2 \frac{N+1}{N \times (1-e^{\frac{-tf}{N}}) + 0.5} \quad (4)$$

where

- t is a query term occurring in document D,
- N_t is the document frequency,
- N is the total number of documents in the collection,
- qtf is the query term frequency,
- tf is the within-document term frequency,
- tfn_2 is the normalised within-document term frequency, given by:

$$tfn_2 = \frac{tf}{ln2} \times log_2 \left[1 + c \times \frac{avg\_dl}{dl}\right] \quad (5)$$

2.3 Model fusion

In the context of information retrieval, data fusion refers to the task of combining the output of multiple ranking strategies into a single list of objects (documents, concepts, etc.) [4,8]. Objects in different lists can be merged together by using a variety of aggregate functions such as MIN, MAX, SUM, AVERAGE, MEDIAN, MNZ, etc. The combination technique based on each of those functions can be referred to as $CombXXX$, where 'XXX' stands for the name of the aggregate function. For example, using the SUM operator for two ranking algorithms A1 and A2, scores are summed to obtain a final score, which is the sum of the term score obtained by ranking algorithm A1 and the one obtained by ranking algorithm A2. CombSUM and CombMNZ have been widely studied and have demonstrated state-of-the-art performance [9]. Such techniques are important in distributed IR where results obtained from several corpora or IR ranking strategies must be coordinated. Authors in [10] distinguished two classes of data fusion techniques : one has access to query-document score (term weighting model), and one does not, with access only to system rankings (document relevance scoring).

With the development of information technology and communication, especially in the IR field, a large number of IR models have been developed and integrated into the Terrier IR platform [5]. We study here the impact of using data fusion technique on the performance of term weighting model. More specifically, we combine term scores obtained by summing the scores obtained by the two state-of-the-art term weighting models described earlier, i.e., LGD and In_expB2 model. Given a query term t, its combined score is computed as follows:

$$
\begin{aligned}
score(t \in D) &= score_{LGD}(t \in D) + score_{In\_expB2}(t \in D) \\
&= qtf \times \left[log_2(\frac{N_t}{N} + tf_n) - log_2(\frac{N_t}{N}) \right] + \\
&\quad \frac{qtf}{ln2} \times \frac{(tf+1) \times tfn_2}{N_t \times (tfn_2 + 1)} \times \log_2 \frac{N+1}{N \times (1 - e^{-tf/N}) + 0.5}
\end{aligned} \tag{6}
$$

2.4 Query expansion

The DFR framework employs a query expansion (QE) mechanism that is a generalisation of Rocchio's method [11]: terms in the top-ranked documents retrieved in the first stage are weighted using a particular DFR term weighting model. In general, the weight of a term of the expanded query q^* derived from the original query q is obtained as follows:

$$
weight(t \in q^*) = qtfn + \beta * \frac{Info_{DFR}}{MaxInfo} \tag{7}
$$

where

- $qtfn$ is the normalised within-query term frequency,
- $MaxInfo = \arg_{t \in q^*} \max Info_{DFR}$,
- $Info_{DFR}$ is the term frequency in the expanded query induced by using a DFR model, that is:

$$
Info_{DFR} = -\log_2 Prob(Freq(w|K)|Freq(w|C)) \tag{8}
$$

where Prob is the probability of obtaining a given within-query term frequency from the top-ranked documents retrieved in the first stage. In the DFR framework, several measures are used to compute this probability such as: Bose-Einstein (Bo) statistics and Kullback-Leibler (KL) measure [3]. The former gives the following term frequency normalisation:

$$
\begin{aligned}
\text{Info}_{\text{Bo}} &= -\log_2 \text{Prob}(\text{Freq}(\text{w}|\text{K})|\text{Freq}(\text{w}|\text{C})) \\
&= -\log_2(\tfrac{1}{1+\lambda}) - \text{Freq}(\text{w}|\text{K}) * \log_2(\tfrac{1}{1+\lambda})
\end{aligned} \tag{9}
$$

where

- $Freq(w|K)$ (resp. $Freq(w|C)$) is the the term frequency within the top ranked documents (resp. the collection)
- $\lambda_{\text{Bo1}} = \frac{\text{Freq(w|C)}}{\text{N}}$ and $\lambda_{\text{Bo2}} = \frac{\text{TotalFreq(K)} * \text{Freq(w|C)}}{\text{TotalFreq(C)}}$
- $\beta = 0.4$

while the latter gives the following term frequency normalisation:

$$
\text{Info}_{\text{KL}} = \frac{Freq(w|K)}{TotalFreq(K)} * \log_2 \frac{Freq(w|K) * TotalFreq(C)}{Freq(w|C) * TotalFreq(K)} \tag{10}
$$

3 Experimental evaluation

3.1 Collection statistics

Some statistical characteristics of the Case-based 2011 collection is depicted in table 1. The histogram in figure 1 shows the variation of document length in the collection.

Table 1. Test collection statistics

Number of documents	55,634
Average document length	3,078
Total number of tokens	171,251,809
Size of the vocabulary	815,708
Number of queries	10
Average query length	30

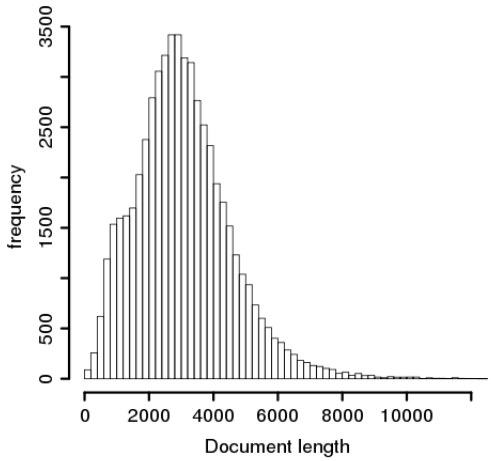

Fig. 1. Variation of document length in the collection

3.2 Evaluation metrics

Retrieval performance was evaluated using standard measures: *P@10, P@20*, and *MAP*. P@10, P@20 represent respectively the mean precisions at the top 10, 20 returned documents. MAP (Mean Average Precision) is the average precision of

a query which is computed by averaging the precision values computed for each relevant retrieved document of rank $x \in (1..K)$, where $K = 1000$ is the number of retrieved documents. Our results are generated by the *trec_eval* standard tool[2] used by the TREC community for evaluating ad hoc retrieval runs.

3.3 Run description

We submitted ten official runs to the case-based medical retrieval track [1]. Our submitted runs are divided into two groups: the first one (6 runs) includes terms with low inverse document frequency (IDF) while the second one (4 runs) excludes them from the document index. In table 2, runs without query expansion are distinguished by an asterisk (*). For the first group of runs, we used the two state-of-the-art weighting models namely LGD [7] (run 1) and In_expB2 [3] (run 2). The $CombSUM$ technique [4] is used in run 3 at the level of term scoring instead of document re-ranking, i.e. the fusion technique modifies directly the scores obtained by retrieval models and not the final scores of output documents. Runs 4, 5 and 6 are combined with a blind feedback query expansion based on the Kullback-Leibler (KL) statistics. Runs in the second group (7, 8, 9, 10) are submitted with the exclusion of low IDF terms using each of the two state-of-the-art weighting models, the $CombSUM$ fusion technique [4] and the KL QE technique. For QE, a maximum number of twenty terms are extracted from the top twenty returned documents. All runs are submitted with the default configuration in Terrier: c=1.0, stopword removal, Porter stemmer.

3.4 Results and discussion

According to the results presented in table 2, we see that ignoring low IDF terms does not help and even harms the IR performance. Normally, low IDF terms are not useful for describing the semantics of the document and can be ignored from the document index [12, 13]. However, in the biomedical domain, especially in medical records retrieval, low IDF terms may be used to mention or distinguish medical concepts such as 'low', 'high', 'right', 'left', ... (e.g., low back pain *vs.* high back pain, right lung *vs.* left lung), etc.

Here, we compare the performance of the two mentioned state-of-the-art models LGD and In_expB2. We notice that the LGD model is better than the In_expB2 model in terms of MAP with an improvement of +17.36%. In terms of P@10, the two models yield the same performance, but in terms of P@20, the former is better than the latter with an improvement of +22.2%. For this reason, we chose the $IRIT\_LGDc1.0$ run as our **strong baseline** to compare to other runs.

The $CombSUM$ method combining term scores obtained by those models outperforms the In_expB2 model and similar to the LGD model (baseline) in terms of MAP. In terms of P@10, the $CombSUM$ fusion technique outperforms the baseline with an improvement of +29.97%. However, in terms of P@20, the

[2] http://trec.nist.gov/trec_eval/

Table 2. Official results of submitted runs in the case-based retrieval task. Runs with an asterisk are without using pseudo relevance feedback.

ID	Run	MAP	P@10	P@20
Includes terms with low IDF				
1	IRIT_In_expB2c1.0_1*	0.0743	0.1111	0.1000
2	IRIT_LGDc1.0* (*baseline*)	0.0872	0.1111	0.1222
3	IRIT_CombSUMc1.0_3*	0.0859	0.1444	0.1000
4	IRIT_CombSUMc1.0_KLbfree_d_20_t_20_2	0.0874	0.1111	0.1000
5	IRIT_LGDc1.0_KLbfree_d_20_t_20_1	**0.1030**	**0.1556**	**0.1278**
6	IRIT_In_expB2c1.0_KLbfree_d_20_t_20_0	0.0772	0.1000	0.1000
Ignores terms with low IDF from index				
7	~~IRIT_CombSUMc1.0_KLbfree_d_20_t_20_2~~	~~0.0874~~	~~0.1111~~	~~0.1000~~
8	IRIT_In_expB2c1.0_KLbfree_d_20_t_20_0_ignore_low_idf	0.0793	0.1444	0.0889
9	IRIT_LGDc1.0_KLbfree_d_20_t_20_1_ignore_low_idf	0.0937	0.1111	0.0889
10	IRIT_CombSUMc1.0_2_ignore_low_idf*	0.0721	0.1333	0.0778

CombSUM technique gives the same performance as the *In_expB2*, which is lower than the baseline. We conclude that combining term scores at the level of weighting models can be useful for improving the search precision (P@10), without losing the MAP performance.

At this level, we present the results of submitted runs obtained using query expansion. The In_expB2 model in combination with the KL QE method shows a small improvement in terms of MAP (+03.90%), a decrease in terms of precision P@10 (-10.00%) and no effect in terms P@20 compared to the In_expB2 model without QE. This is probably due to the fact that the clinical query length is long (about 30 in average) and the number of extracted terms for QE is smaller than or equal to 20; therefore extracted terms may be observed as in the original terms or also they can be different from the latter but the top-ranked documents do not or slightly change after expansion. The LGD model in combination with the KL QE method (run *IRIT_LGDc1.0_KLbfree_d_20_t_20_1*) outperforms the baseline with an improvement rate of +17.85% in terms of MAP, +40.05% in terms of P@10 and +4.58% in terms of P@20. This proves that query expansion is only effective if it is based on an underlying effective ranking model. Indeed, the *In_expB2* model in combination with the KL QE method performs worse than the baseline. This also explains why the *CombSUM* method in combination with the KL QE gives no improvement compared to the baseline.

4 Conclusion

In this work, we have compared and evaluated the IR performance of two state-of-the-art term weighting models, a state-of-the-art query expansion approach. In our empirical studies, we proposed to combine term scores obtained by different term weighting models to improve the retrieval performance, especially the search precision.

Within the case-based retrieval task, we noticed that low IDF terms are also useful for indexing and retrieval because they can be used to mention or distinguish medical concepts. The LGD model proposed by [2] shows the best performance on the case-based retrieval task and consistently outperforms the In_expB2 model with or without query expansion. The combination of the LGD model, which is based on the log logistic distribution, and the KL query expansion method gives the best results. We conclude that an effective ranking model in conjunction with a appropriate query expansion strategy could be combined together to improve the IR performance.

Since documents in the case-based collection contains a lot of medical concepts, in our future work, we aim to extract concepts from documents for better representing the document's semantics. In addition, we'll also focus on adjusting query by expanding the query with related terms denoting concepts in ontologies or removing non informative terms from the query.

References

1. Kalpathy-Cramer, J., Müller, 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. Clinchant, S., Gaussier, É.: Information-based models for ad hoc IR. In: SIGIR. (2010) 234–241
3. Amati, G.: Probabilistic models for Information Retrieval based on Divergence from Randomness. PhD thesis, University of Glasgow (2003)
4. Fox, E.A., Shaw, J.A.: Combination of Multiple Searches. In: TREC 1994. (1994) 243–252
5. Ounis, I.;Lioma, C.C.V.: Research directions in terrier. Novatica/UPGRADE Special Issue on Web Information Access, Ricardo Baeza-Yates et al. (Eds), Invited Paper (2007)
6. Porter, M.F. In: An algorithm for suffix stripping. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1997) 313–316
7. Clinchant, S., Gaussier, É.: Retrieval constraints and word frequency distributions a log-logistic model for ir. Information Retrieval **14**(1) (2011) 5–25
8. Dinh, D., Tamine, L.: Voting techniques for a multi-terminology based biomedical information retrieval (regular paper). In: Conference on Artificial Intelligence in Medicine (AIME), Bled, Slovenia, 02/07/2011-06/07/2011. Volume 6747 of LNAI., Springer (2011) 184–193
9. Lee, J.H.: Analyses of multiple evidence combination. In: Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval. SIGIR '97, New York, NY, USA, ACM (1997) 267–276
10. Efron, M.: Generative model-based metasearch for data fusion in information retrieval. In: Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries. JCDL '09, New York, NY, USA, ACM (2009) 153–162
11. Rocchio, J. In: Relevance Feedback in Information Retrieval. (1971) 313–323
12. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. 1st edn. Addison Wesley (May 1999)
13. Croft, W.B., Metzler, D., Strohman, T.: Search Engines - Information Retrieval in Practice. Pearson Education (2009)