DEMIR at CLEF eHealth: The Effects of Selective Query Expansion to Information Retrieval

Okan Ozturkmenoglu[1], Adil Alpkocak[1], and Deniz Kilinc[2]

[1] Dokuz Eylül University, Dept. Computer Engineering,
DEMIR Dokuz Eylül Multimedia Information Retrieval Research Group,
Tinaztepe Izmir 35390, Turkey
[2] Celal Bayar University, HFT Technology Faculty
Turgutlu, Manisa, Turkey
okan.ozturkmenoglu@deu.edu.tr, alpkocak@cs.deu.edu.tr,
deniz.kilinc@cbu.edu.tr

Abstract. This paper presents the details of participation of DEMIR (Dokuz Eylül University Multimedia Information Retrieval) research team to the Share/CLEF eHealth 2014. This year, we participated to task 3a: monolingual user-centered health information retrieval. In this task, we focused to apply query expansion techniques selectively to some queries to improve the performance of information retrieval. Thus, we first extracted some statistical features from queries such as length of query, sum and intersect of document frequencies of each query term etc. We develop a system to predict if a query is to be expanded or not. Then, we trained our system with previous year's data. Then, we applied a query expansion method only to the queries, which are selected by the system. The results show that the approach we proposed slightly improves our baseline retrieval performance in terms of P@10.

Keywords: Query classification, selective query expansion, information retrieval

1 Introduction

In this paper, we present the experiments performed by Dokuz Eylül University Multimedia Information Retrieval (DEMIR) Research Group, in the context of our participation to the ShARe/CLEF eHealth Evaluation Lab [1]. This year, we participated to task 3a: monolingual user-centered health information retrieval [2]. This task is a standard information retrieval task as retrieving the relevant documents for a given set of user topics/queries. This task uses 2012 crawl of approximately one million medical documents made available by the EU-FP7 Khresmoi project (http://www.khresmoi.eu/). The main focus of our participation is to apply query expansion methods to a set of queries selectively instead of the whole set. This is because, there is no query expansion method improves retrieval performance for all queries. A query expansion method works well on some queries or query type while it doesn't work for other queries. If we can predict the queries to be expanded before-

hand, it would be great for retrieval performance. Then we can apply the query expansion on a selected set of queries predicted to be potential to improve the information retrieval performance. We called this approach as selective query expansion [3-6]. We experiment our proposed approach at 2013 dataset in CLEF eHealth. To do this, we extracted some statistical features for type of queries from indexed documents and used them as attributes in classification. After classification, we predict if a query is going to be expanded or not. We obtained a slight improvement with 2013 data. Thus, we applied this approach to this year data.

Rest of the paper is organized as follow: Section 2 provides an explanation of statistical features of query text and describes query classification process. In the next section 3 we present our experimental results. Section 4 concludes the paper by pointing out the open issues and possible avenues of further research for applying query expansion methods selectively in information retrieval system.

2 Classification of Queries

Task 3a contains approximately one million medical documents, which are collected different web sources. They are in eight-part folder and each document is taken formatted style in own source file which extension is *dat*. We processed these *dat* files and extracted all documents as a single file, which can be indexed and retrieved in IR system.

In original document collection, data structure contains HTML tags in *content* tag so when we processed them, we stripped out HTML tags and used *title*, *heading* and *body* information in it. We created a new content data and document data structure using these tags information. So we used this new data structure in IR system.

In this work, we tested effect of selective query expansion method in IR system performance. As in standard IR system, we preprocess document collection and indexed them. We used Terrier IR Platform API, which is an open source search engine written in Java and is developed at the School of Computing Science, University of Glasgow, to generate vector space model [7]. Terrier provides efficient and effective search methods supported by many different parameters.

Before retrieval, we processed the queries and extracted statistics of query terms from data collection [8]. The extracted features for query types are as follows:

- Query length (*QE*): The number of terms in a query.
- Intersect document frequency (*IntersectDF*): The number of documents in the collection that contain all terms in a query.
- Maximum document frequency (*MaxDF*): The maximum document frequency (*df*) of query terms.
- Minimum document frequency (*MinDF*): The minimum document frequency of query terms.
- Summation document frequency (*SumDF*): The summation document frequencies of all query terms.
- Average document frequency (*AvgDF*): The average of document frequencies of all query terms.

- Maximum inverse document frequency (*MaxIDF*): Inverse of maximum document frequency of query terms.
- Minimum inverse document frequency (*MinIDF*): Inverse of minimum document frequency of query terms.
- Average Term Frequency (*AvgTF*): The average term frequency, which is total number of occurrence of the term in the collection.

We used all these statistical information as attributes for classification process of queries. To classify the queries as to be expanded and not to be expanded, we used Naïve Bayes method in WEKA machine learning software [9].

We performed the training and test sets in the following way. We retrieved two groups of result for 2013 data. We applied baseline method in one of them is described in section 3.2 and is shown as $r1$ in equation 1. In addition to the first group, only we used KL method as query expansion model in another one and it is shown as $r2$ in equation 1. This expansion model was better model than others, we experiment it for 2013 data and we explained it in section 3.1. During creation of training set, we calculated difference of *map* score in retrieval result between these two groups for each query.

$$query\_diff\_map\_score = r2_{map} - r1_{map} \quad (1)$$

We used $query\_diff\_map\_score$ value in equation 1 when we decided to query's class. If this value was positive, we labelled query as positive and so we applied query expansion model in this query when we retrieved. If difference is zero or negative, we did not apply any model. We formulated this function in equation 2.

$$f(x) = \begin{cases} P, & difference\_map\_score > 0 \\ N, & difference\_map\_score \leq 0 \end{cases} \quad (2)$$

In 2013 data, we labelled 21 queries as positive, 29 queries as negative. It means that using query expansion improved retrieval performance in 21 queries. For evaluation of our labeling function is shown in Eq. 2, we used Naïve Bayes method as classifier in WEKA as test dataset, and applied 10-fold cross validation. Our classification accuracy is around 70% for 2013 queries. Likewise for 2014 data, we used 2013 queries as training set and 2014 queries as test set and performed a prediction with 2014 queries if a query is to be expanded or not. After classification, we predicted query's class as positive or negative for query expansion. If the result is positive, we expanded it using KL expansion model in Terrier and did not expand the query if otherwise.

3 Experiments and Results

3.1 Experiments

In order to assess our proposal, we set up a set of experiments on the 2013 data collection of CLEF eHealth. In experiments, our aim was which weighting-model, query field and expansion method was used in runs.

Between run 1_1 and 1_3, we used *title* tag as query field and different weighting models such as TF×IDF, BM_25 and DFR_BM25, while retrieving. By the way, we did not use any expansion method. We obtained that TF×IDF is best among them. In run 2_1 and 2_2, we tried to find out which query fields should be used in retrieval. In data collection, each query has *title*, *desc*, *profile* and *narr* information. We used only *title* tag, because its performance is the best. In run 3_1 and 3_3, we aimed to choose which query expansion method to be used. In this test, we used only *title* tag, TF×IDF weighting-model and three different query expansion models such as *Bo1*, *Bo2* and *KL* which are available in Terrier. We obtained the best result with *KL* method.

Table 1 shows our experimental results for last year data collection and queries. According to these results, we decided to use *title* tag as a query field, TF×IDF as a weighting-model, *KL* as a query expansion method in our submitted runs.

Table 1. Experimental results for CLEF eHealth 2013.

Id	Query Field	Weighting	Exp	map	gm_ map	bpref	P_10	P_30
1_1	Title	Tf×Idf	-	0.2641	0.0861	0.3557	0.4540	0.3093
1_2	Title	Bm_25	-	0.2562	0.0841	0.3498	0.4480	0.3080
1_3	Title	Dfr_Bm25	-	0.2591	0.0850	0.3512	0.4540	0.3107
2_1	Title+ Desc	Tf×Idf	-	0.2335	0.0753	0.3477	0.3980	0.2720
2_2	Title+ Narr	Tf×Idf	-	0.2483	0.0844	0.3550	0.4240	0.2887
3_1	Title	Tf×Idf	Bo1	0.2611	0.0797	0.3629	0.4200	0.2940
3_2	Title	Tf×Idf	Bo2	0.2379	0.0538	0.3710	0.4140	0.2580
3_3	Title	Tf×Idf	KL	0.2648	0.0759	0.3658	0.4320	0.3007
1_1	Title	Tf×Idf	-	0.2641	0.0861	0.3557	0.4540	0.3093

3.2 Runs

Table 2 shows the four runs we submitted to ShARe/CLEF eHealth Evaluation Lab task 3a. Below, we provide a short description of each run.

Table 2. Runs of DEMIR group for task 3a in ShARe/CLEF eHealth Evaluation Lab 2014.

RunID	map	gm_map	bpref	P_10	P_30
DEMIR_EN_Run.1	0.3644	0.3065	0.5154	0.6300	0.5280
DEMIR_EN_Run.5	0.3714	0.3079	0.5490	0.6700	0.5200
DEMIR_EN_Run.6	0.3049	0.2470	0.5199	0.6740	0.4687
DEMIR_EN_Run.7	0.3261	0.2518	0.5281	0.6120	0.4720

- DEMIR_EN_Run.1: This run is our baseline retrieval result. In this run, *title* and *content* tag is used to index documents. Term-weighting model is TF×IDF. UTF tokenizer and stopword list were used and we applied porter stemmer. Query expansion model was not used. We used *title* field in query

file for each topic when retrieved. We obtained the best result in CLEF eHealth 2013, using this method so we have it as a baseline for this year.

- DEMIR_EN_Run.5: This run is exactly the same with baseline run except for all queries we used KL query expansion method available in Terrier.
- DEMIR_EN_Run.6: In this run, we applied similar pre-process and indexing operations on documents like run 1 and 5. For query expansion model, we extracted term statistics such as *QE, IntersectDF, MaxDF, MinDF, SumDF, AvgDF, MaxIDF, MinIDF* and *AvgTF* from queries. We used them as attributes and applied Naïve Bayes classification method. As a result of classification, we expect to predict which query to expand. 27 queries are positive and other 23 queries are negative. We expanded positive queries using KL query expansion model, which were selected by classification process. We did not expanded negative queries.
- DEMIR_EN_Run.7: In this run, we selected queries manually which were expanded. We called it as blind query expansion. We labelled 16 queries as positive and others were negative. We expanded positive queries.

Table 3 shows the results obtained with the graded relevance assessment.

Table 3. Runs of DEMIR group for task 3a in ShARe/CLEF eHealth Evaluation Lab 2014 with graded relevance assessment (nDCG).

Run ID	cut_ 5	cut_ 10	cut_ 15	cut_ 20	cut_ 30	cut_ 100	cut_ 200	cut_ 500	cut_ 1000
1	0.654	0.632	0.610	0.593	0.579	0.537	0.591	0.642	0.667
5	0.696	0.672	0.647	0.621	0.584	0.542	0.601	0.649	0.677
6	0.656	0.652	0.619	0.576	0.533	0.469	0.515	0.579	0.614
7	0.667	0.621	0.588	0.558	0.529	0.492	0.548	0.602	0.635

4 Conclusion

In this year, we tried to classify queries as to be expanded or not. This is because there is no query expansion methods works for all type of queries. In other say, a query expansion method improves only a set of queries and worsens the rest. Thus, it would be great if we can predict which queries to be expanded. This is the basic idea of our study. Hence, we tried to classify or predict the queries to which an expansion method will work effectively and we apply it on them and expect to improve retrieval performance.

So in this work, we performed query expansion on a selected set of queries instead of the whole queries and expect a performance improvement. The results we obtained showed that the approach we proposed slightly improves our baseline retrieval performance in terms of P@10. It shows that it is promising and needs further studies on this topic.

References

1. Kelly, L., Goeuriot, L., Suominen, H., Schrek, T., Leroy, G., Mowery, D.L., Velupillai, S., Chapman, W.W., Martinez, D., Zuccon, G., Palotti, J.: Overview of the ShARe/CLEF eHealth Evaluation Lab 2014. Springer, (2014)
2. Goeuriot, L., Kelly, L., Li, W., Palotti, J., Pecina, P., Zuccon, G., Hanbury, A., Jones, G., Mueller, H.: ShARe/CLEF eHealth Evaluation Lab 2014, Task 3: User-centred health information retrieval. (2014)
3. Cronen-Townsend, S., Zhou, Y., Croft, W.B.: A language modeling framework for selective query expansion. DTIC Document (2004)
4. Bashir, S., Rauber, A.: On the relationship between query characteristics and IR functions retrieval bias. Journal of the American Society for Information Science and Technology 62, 1515-1532 (2011)
5. He, B., Ounis, I.: Query performance prediction. Information Systems 31, 585-594 (2006)
6. Kumaran, G., Carvalho, V.R.: Reducing long queries using query quality predictors. In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, pp. 564-571. ACM, (2009)
7. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Lioma, C.: Terrier: A High Performance and Scalable Information Retrieval Platform. In: Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006). (2006)
8. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press (2008)
9. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explor. Newsl. 11, 10-18 (2009)