Pseudo-Relevance Feedback Based on mRMR Criteria

Yuanbin Wu, Qi Zhang, Yaqian Zhou, and Xuanjing Huang

School of Computer Science, Fudan University
825 Zhangheng Road
Shanghai, P.R. China, 201203
{ybwu,qz,zhouyaqian,xjhuang}@fudan.edu.cn

Abstract. Pseudo-relevance feedback has shown to be an effective method in many information retrieval tasks. Various criteria have been proposed to rank terms extracted from the top ranked document of the initial retrieval results. However, most existing methods extract terms individually and do not consider the impacts of relationships among terms and their combinations. In this study, we first re-examine this assumption and show that combinations of terms may heavily impact the final results. We then present a novel clustering based method to select expansion terms as a whole set. The main idea is to use first simultaneously cluster terms and documents using non-negative matrix factorization, and then use the Maximum Relevance and Minimum Redundancy criteria to select terms based on their clusters, term distributions, and other features. Experimental results on serval TREC collections show that our proposed method significantly improves performances.

Keywords: Pseudo-relevance Feedback, NMF, mRMR Criteria.

1 Introduction

With the ever-increasing growth of the World-Wide Web, the number of casual search engines users has grown rapidly. However user queries are usually too short to describe the accurate information they need. The analysis [1] shows that the average query length is 1.7 terms for popular queries and 2.2 terms over all queries. In order to address this problem, query expansion has been receiving much attention in a long time [2,3,4,5,6,7,8]. Among all the query expansion methods, pseudo-relevance feedback (PRF) [9,10] is nearly the most attractive one because it does not require any user input. PRF assumes that top ranked documents are relevant. Many approaches have been presented to extract useful terms from those pseudo-relevant documents. Serval criteria have also been proposed to select expansion terms based on term distributions, such as *idf*, *tf*, χ^2 statistic, web resource (e.g Wikipedia), linguistic feature, ontology and so on. More recently, supervised learning methods have also been studied[5,11].

However, previous research efforts have mainly focused on extracting terms separately without considering the impacts of relationships among terms and their combinations. Selection of terms is typically performed in a greedy manner using some type of score or rank. As a result, the performances of current models are usually unstable and the improvements are limited. Several experiments in Section 3 show that term combinations can highly impact the final results. This in turn leads to a natural question: how can we select *a set* of expansion terms from pseudo-feedback documents?

P.-J. Cheng et al. (Eds.): AIRS 2010, LNCS 6458, pp. 211–220, 2010.

To address the above question, we propose to seek a set of terms by requiring that: 1) the terms are related to the query and useful for Information Retrieval (IR) according to their distributions, linguistic features, and so on; 2) redundancy of the terms is minimum, in other words, the terms should be maximally dissimilar to each other. In this paper, we use non-negative matrix factorization method to cluster the terms and capture their pairwise correlations. Then a novel expansion term selection algorithm is used to extract a set of terms. We compare the proposed algorithm with the traditional approaches on five TREC test collections. The experimental results show that our proposed term selection method achieves good performance and is able to improve the retrieval effectiveness significantly.

The contributions of our work can be summarized as follows: 1) We thoroughly evaluate impacts of relationships among terms and their combinations for pseudo relevance feedback. 2) We propose a novel maximum relevance and minimum redundancy criteria to select expansion terms. In contrast to existing work, expansion terms are selected as a set, and the relationships among them are considered.

The remaining of the paper is organized as follows: In Section 2, we review a number of related work and the state-of-the-art approaches in query expansion. Section 3 provides experimental examinations of the hypothesis about term combinations and shows that it does hold in practice. In Section 4, we present our expansion criteria. Experimental results on five TREC test collections are shown in Section 5. Finally, Section 6 concludes and suggests some future work.

2 Related Work

The approach presented in this paper is related to previous work on pseudo-relevance feedback and Non-negative Matrix Factorization.

The problem of how to automatically extract useful terms from the top rank initial retrieval set is a long-studied task. Since no user input is required, pseudo-relevance feedback (PRF) has received considerable attentions. Statistical based query expansion is one of the classical methods. Okapi [12] addes the 20 top ranking terms, which is scored by $BM25$ weight, for query expansion. Carpineto et al.[8] introduced an information-theoretic method, including Rocchio's weights, Robertson Selection Value (RSV), CHI-squared, and Kullback-Leibler distance, for query expansion.

Many approaches have been proposed to improve the effectiveness of PRF with external resources, such as Wordnet[13], dependency relations[14], and so on. Xu et al.[2] explored the utilization of wikipedia in PRF. They categorized the TREC topics into three types based on wikipedia and proposed different methods for term selection with wikipedia entity pages. Collins-Thompson and Callan [15] described a Markov chain model that combines multiple sources of knowledge on term associations, and allows chaining of multiple inference steps with different link types to perform "semantic smoothing" on language models, and applied this model to query expansion.

Recently, there has been work focused on selecting better documents for pseudo-relevance feedback. Sakai et al. [16] proposed an approach to skip documents in the initial ranked documents to look for more "novel" pseudo-relevant documents.

They used cluster method to collect novel documents rather than separating relevant documents from non-relevant ones. However, their experiments on NTCIR collections did no show significant improvements. Lee et al.[3] proposed an approach to resample the top-ranked documents using clusters. They assumed that document that appears in multiple highly-ranked clusters would contribute more to the query terms than other documents.

Huang and Croft [17] proposed a framework to expand queries with a small number of opinion words. A number of sentiment expansion approaches were used to find the most appropriate query-independent or query-dependent opinion words. However, the query expansion method they used is query independent. It only forces on the opinion retreval domain.

Supervised learning methods have also been proposed to classify expansion terms. Zhang et al. [11] proposed a method to automatically evaluate the retrieval effectiveness of terms. Then SVM is used to select terms directly based on statistical features. Cao et al. [5] re-examined the assumption of pseudo-relevance feedback and used the similar idea to select expansion terms.

The most similar work to our proposal is the method proposed by Udupa et al. [18]. They claimed that the effect of including a term into an expansion set depended on the rest of the terms in the expansion set. They proposed to use spectral partitioning of term-term interaction matrix to takes into account term interactions. Different from their approach, we propose to use maximum relevance and minimum redundancy criteria to select terms as a whole. The redundancy is captured by term clusters, which is obtained by constrained non-negative matrix factorization method. Term distributions and linguistic features are used to measure the relevance.

3 Motivation

As we mentioned in Section 1, our approach selects expansion terms as a whole set. The general assumption behind it is that the relationships among terms and their combinations can impact the final retrieval result. To evaluate this assumption, we consider all the "good" expansion terms and their combinations.

Suppose $MAP(q)$ represents the mean-average-precision of the original query q and $MAP(q \bigcup t_i)$ represents the MAP of the expanded query(original query with term t_i). Following the formula proposed in [5], the performance change due to t_i is measured by $chg(t_i) = \frac{MAP(q \bigcup t_i) - MAP(q)}{MAP(q)}$. *Good expansion terms* are those whose $chg(t_i)$ is bigger than 0.05[1].

Since the size of vocabulary is too big and the evaluation is a time consuming task, we use the following score function to select top 200 words as candidate terms:

$$Score(t_i, q) = \log \frac{(r_i + 0.5) * (N - n_i - R + r_i + 0.5)}{(R - r_i + 0.5) * (n_i - r_i + 0.5)} \tag{1}$$

where N is the total number of documents, R is the number of relevant documents, the number of documents and relevant documents containing term t_i are respectively

[1] Cao et al. [5] set up this threshold to 0.005. Because of the computational limitations, we use a higher threshold to reduce the number of expansion sets.

Table 1. MAP at all test collections with different expansion queries

Collection	BM25	Best	Worst	Top 5
TREC 7	0.1815	0.3792	0.1565	0.3320
TREC 8	0.2385	0.3518	0.2032	0.3194
TREC 10 Web	0.1923	0.4212	0.1636	0.2767
Blog2006	0.3040	0.4560	0.2135	0.3840
Blog2007	0.3744	0.5296	0.2992	0.4687

represented by n_i and r_i[19]. The top 20 documents[2] are regarded as relevant in our experiments. Then, good expansion terms are measured and selected by $chg(t)$ from the candidate terms.

Through the above steps, we can obtain a set of good expansion terms for each topic. In order to evaluate the impact of their combinations, $C^5_{|G_k|}$ expanded queries are generated where $|G_k|$ represents the count of good expansion terms in topic k and $C^5_{|G_k|}$, is the number of 5-combinations from the good expansion terms. In our examination, the maximum size of $|G_k|$ is set to 20. Each expanded query contains both the original query and five additional terms, which are selected by $chg(t_i)$. We use Lemur toolkit[3] to conduct the experiments in this paper.

Five TREC collections are used to examine the assumption and their description can be found in Section 6.1. in TREC 7 (from topic 351 to $C^5_{|G_k|}$ Table 1 shows the summary results in different collections. The fourth column "Top 5" represents the result of combining initial query with five additional terms whose $chg(t_i)$ are the highest among all the candidate terms.

As we can see, pseudo-relevance feedback is effective when good expansion set is selected. In TREC 7, the relative improvement of the best expansion set over the initial result is 108.9%. However, although every single term can improve the retrieval result, their combinations may deteriorate the performance. In all collections, the worst results with query expansions of some topics are even much lower than the result of the initial query without any expansion terms. We also note from the Table 1 that the best expansion set achieves better result than the *Top 5*. In other words, although each single term can only improve the result slightly, their combination can be more effective.

4 Query Expansion Methods

Figure 1 shows the overview of our proposed query expansion methods. First of all the system returns an initial set of retrieval results with the given query. Then the unigram language model is applied to generate term-document matrix. After that non-negative matrix factorization methods are used on term-document matrix. Upon convergence of the matrix factorization algorithm, term-cluster and document-cluster matrices are

[2] Both Huang & Croft's results [17] and our experimental results show that the number of pseudo relevance documents from five to twenty are reasonable estimations.

[3] http://www.lemurproject.org/

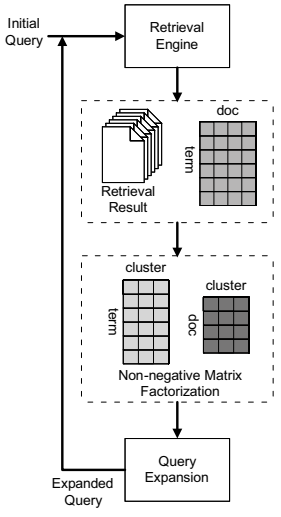

Fig. 1. Overview of our methods

obtained. Given the two matrices, query expansion set can be generated by different query expansion methods, which will be detailedly described in the following parts.

4.1 Coverage Criterion for Query Expansion

From analyzing clustering result of the top-ranked documents, we observe that those documents usually correspond to several different topics. Different clusters usually contain different concepts. Buckley et al.[20] also mentioned the observation. Since the intention of the query cannot be easily detected, including all major topics should be useful for the expansion queries.

In order to improve the coverage of expansion queries, one of the straightforward method is to select terms from each clusters. Since clusters have different size, a balanced way to select terms from the multiple clusters should be considered. We assume that clusters with more documents are more important. Based on this assumption, the number of terms extracted from cluster C_i is set to $\lceil \frac{|C_i|}{N} \rceil$, where N is the total number of documents and $|C_i|$ represents the number of documents in the cluster. In each cluster, terms are ranked by Eq.(1).

4.2 Maximum Relevance and Minimum Redundancy Criterion for Query Expansion (mRMR-QE)

Ding and Peng [21] proposed *minimal-redundancy-maximal-relevance*(mRMR) framework to select promising features. Inspired by their work on feature selection, we redefine the *maximal relevance* constraint and *minimal redundancy* constraint for text retrieval, and combine them as mRMR-QE criterion for query expansion.

The maximal relevance condition is to search a set of terms T satisfying Eq.(2). $Score_{c_i}(t_i, q)$ is similar to that in Eq.(1), except that R is the number of relevant

documents in cluster c_i. $\max_{c_i \in C} Score_{c_i}(t_i, q)$ represents the maximum relevance score of the term to a cluster. The relevance between T and q is measured by the sum of individual term t_i and query q.

$$\max Rel(T, q), \quad Rel(T, q) = \sum_{t_i \in T} \max_{c_i \in C} Score_{c_i}(t_i, q) \tag{2}$$

However, it is likely that the redundancy among the terms selected according to the maximal relevance criterion could be rich. More over, from the analysis of Section 3, we observe that the including terms which are highly dependent on each other may degrade the result. Therefore, the following minimal redundancy criterion can be used to select mutually exclusive terms.

$$\min Red(T, q), \quad Red(T, q) = \frac{1}{|T|^2} \sum_{t_i, t_j \in T} I(t_i, t_j), \tag{3}$$

where $I(x_i, x_j)$ represents the mutual information between term t_i and t_j.

Combing the maximal relevance and minimal redundancy creation, the maximum relevance and minimum redundancy Criterion for query expansion (mRMR-QE) is defined by the Eq.(4).

$$\arg \max_{T \in \mathcal{J}} (Rel(T, q) - Red(T, q)) \tag{4}$$

Following the method proposed in [22], an incremental search method is used in practice. Suppose we already have T_{m-1}, which contains $m-1$ terms. Then Eq.(5) is used to select the mth term from the remaining set $\{W - T_{m-1}\}$.

$$\arg \max_{t_i \in \{W - T_{m-1}\}} \{Rel(T_{m-1} \cup t_i, q) - Red(T_{m-1} \cup t_i, q)\} \tag{5}$$

5 Experiments

5.1 Collections

We evaluate our methods with three TREC corpus Disk4&5, WT10g and BLOGS06. Five test collections, TREC 7, TREC 8, TREC 10 Web, TREC Blog 2006, and TREC Blog 2007, are used in the experiments. We implement our expansion methods based on Lemur 4.10 [4]. Okapi BM25 ranking function is used as the retrieval model. All test collections and corpus were stemmed using the Porter stemmer provided as part of Lemur. As for performance measures, the mean average precision (MAP) for top 1000 documents is the primary evaluation metric in all the test collections. Other metrics include precision at five documents (P@5), precision at ten documents (P@10), R-precision (R-prec), and binary Preference (bPref). We also conduct τ-test to determine where the improvement on performance statistically significant.

[4] `http://www.lemurproject.org`

Table 2. Ad-hoc retrieval results on all the test collections with Okapi BM25 function and pseudo relevance feedback using coverage criterion. * indicates that the improvement over "BM25+PRF" is statistically significant($\rho < 0.05$).

Collection	P@10	bPref	R-prec	MAP-P	MAP
TREC 7	0.3920	0.2423	0.2636	0.2303	0.2355
TREC 8	0.4640	0.2574	0.2812	0.2489	0.2555 *
Blog2006	0.6812	0.3615	0.3961	0.3150	0.3095
Blog2007	0.6700	0.4001	0.4111	0.3912	0.3849
TREC10Web	0.3086	0.1809	0.2319	0.2070	0.2086

Table 3. Ad-hoc retrieval results on all the test collections with Okapi BM25 function and pseudo relevance feedback using mRMR-QE criterion. * indicates that the improvement over coverage criterion is statistically significant at level $\rho < 0.05$.

Collection	P@10	bPref	R-prec	MAP-P	MAP
TREC 7	0.4440	0.2557	0.2775	0.2303	0.2494 *
TREC 8	0.4860	0.2759	0.3046	0.2489	0.2776 *
Blog2006	0.6937	0.3676	0.4106	0.3150	0.3389 *
Blog2007	0.7060	0.4072	0.4227	0.3912	0.4017
TREC10Web	0.3216	0.1862	0.2343	0.2070	0.2109

5.2 Coverage Criterion Evaluation

Table 2 contains the results of experiments in all test collections using Okapi BM25 ranking function and pseudo relevance feedback with coverage criterion. The left column in the table shows the test collections. Each row in the table represents the results of different performance metrics. The column "MAP-P" represents results of the Okapi BM25 ranking function with traditional pseudo relevance feedback. For all four collections, 20 terms are sorted by Eq. 1 and extracted with coverage criterion from 5 top-ranked documents.

From the table we can observe that the coverage criterion achieve better results than the original pseudo relevance feedback method in all four collections. In TREC 8 collection, the criterion achieves significant improvement. The results we obtained in TREC 7 and TREC 8 are the state-of-the-art performance. Compared to "BM25+PRF ", not only the MAP but also most of the other evaluation metrics achieve better results among the collections. Those results can also demonstrate the observation we mentioned in the previous sections.

5.3 mRMR-QE Evaluation

In this experiment, we also use the TREC corpus Disk4&5, WT10g, and BLOGS06 to test the performances. In the same way as previous experiments, five test collections, TREC 7, TREC 8, TREC 10 Web, TREC Blog 2006, and TREC Blog 2007, are evaluated. Table 3 summaries the results of experiments in all test collections using Okapi

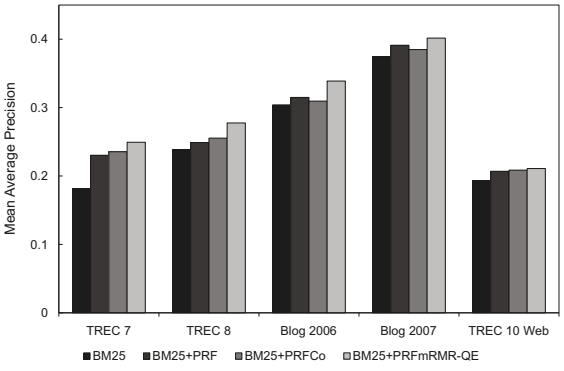

Fig. 2. Performance comparisons of different expansion methods using MAP

BM25 ranking function and pseudo relevance feedback with mRMR-QE criterion. The same parameters are used in this experiment.

From the Table 3, we observe that expansion terms extracted by pseudo relevance feedback with maximum relevance and minimum redundancy criterion can significantly improve the retrieval effectiveness. In all the collections, mRMR-QE criterion achieves better results than coverage criterion in most performance metrics. Figure 2 shows the performance comparisons of different expansion methods. Those results show that mRMR-QE criterion can capture the good expansion terms more effectively than pervious approaches. This is consistent with the observations we studied in the Section 3. We also note from the retrieval result that more than 69.2% of expansion queries give positive impact over the original queries, which is also more robust than coverage criterion.

6 Conclusions

In this paper, we studied the impacts of the relationships among terms and their combinations. Through several empirical experiments, we show that retrieval performance would significantly impacted by the combinations of expansion terms. In all five test collections, the best expansion can significantly improve the retrieval result.

In order to address this problem, we presented a novel clustering based method to select expansion terms as a set. The main idea is to first simultaneously cluster terms and documents, and then use Maximum Relevance and Minimum Redundancy criteria to select terms based on their clusters, term distributions, and other features. We evaluated the results with five different TREC collections. We also discussed the factors in our proposed method, including the number of expansion terms and the number of pseudo relevance documents.

Acknowledgements

The author wishes to thank the anonymous reviewers for their helpful comments. This work was partially funded by 973 Program (2010CB327906), The National High

Technology Research and Development Program of China (2009AA01A346), Shanghai Leading Academic Discipline Project (B114), Doctoral Fund of Ministry of Education of China (200802460066), National Natural Science Funds for Distinguished Young Scholar of China (61003092), and Shanghai Science and Technology Development Funds (08511500302).

References

1. Beitzel, S.M., Jensen, E.C., Chowdhury, A., Grossman, D., Frieder, O.: Hourly analysis of a very large topically categorized web query log. In: SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 321–328. ACM, New York (2004)
2. Xu, Y., Jones, G.J., Wang, B.: Query dependent pseudo-relevance feedback based on wikipedia. In: SIGIR 2009: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 59–66. ACM, New York (2009)
3. Lee, K.S., Croft, W.B., Allan, J.: A cluster-based resampling method for pseudo-relevance feedback. In: SIGIR 2008: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 235–242. ACM, New York (2008)
4. Collins-Thompson, K., Callan, J.: Estimation and use of uncertainty in pseudo-relevance feedback. In: SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and evelopment in Information Retrieval, pp. 303–310. ACM, New York (2007)
5. Cao, G., Nie, J.Y., Gao, J., Robertson, S.: Selecting good expansion terms for pseudo-relevance feedback. In: SIGIR 2008: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 243–250. ACM, New York (2008)
6. Tao, T., Zhai, C.: Regularized estimation of mixture models for robust pseudo-relevance feedback. In: Proceedings of SIGIR 2006, pp. 162–169. ACM, New York (2006)
7. Bhogal, J., Macfarlane, A., Smith, P.: A review of ontology based query expansion. Information Processing & Management 43(4), 866–886 (2007)
8. Carpineto, C., de Mori, R., Romano, G., Bigi, B.: An information-theoretic approach to automatic query expansion. ACM Transactions on Information Systems 19(1), 1–27 (2001)
9. Buckley, C.: Automatic query expansion using SMART: TREC 3. In: Proceedings of The Third Text REtrieval Conference (TREC-3), pp. 69–80 (1994)
10. Yu, S., Cai, D., Wen, J.R., Ma, W.Y.: Improving pseudo-relevance feedback in web information retrieval using web page segmentation. In: Proceedings of WWW 2003, pp. 11–18. ACM, New York (2003)
11. Zhang, Q., Wang, B., Huang, X.H., Wu, L.: FDU at TREC 2007: opinion retrieval of blog track. In: Proceedings of The Sixteen Text REtrieval Conference, TREC-2007 (2007)
12. Robertson, S.E., Walker, S., Hancock-Beaulieu, M.M., Gatford, M., Payne, A.: Okapi at TREC-4. In: Proceedings of The Fourth Text REtrieval Conference, TREC-4 (1996)
13. Moldovan, D.I., Mihalcea, R.: Using wordnet and lexical operators to improve internet searches. IEEE Internet Computing 4(1), 34–43 (2000)
14. Sun, R., Ong, C.H., Chua, T.S.: Mining dependency relations for query expansion in passage retrieval. In: Proceedings of SIGIR 2006, pp. 382–389. ACM, New York (2006)
15. Collins-Thompson, K., Callan, J.: Query expansion using random walk models. In: CIKM 2005: Proceedings of the 14th ACM international conference on Information and knowledge management, pp. 704–711. ACM, New York (2005)

16. Sakai, T., Manabe, T., Koyama, M.: Flexible pseudo-relevance feedback via selective sampling. ACM Transactions on Asian Language Information Processing (TALIP) 4(2), 111–135 (2005)
17. Huang, X., Croft, W.B.: A unified relevance model for opinion retrieval. In: Proceedings of 16th Conference on Information and Knowledge Management (CIKM 2009), Hong Kong, China (2009)
18. Udupa, R., Bhole, A., Bhattacharyya, P.: A term is known by the company it keeps: On selecting a good expansion set in pseudo relevance feedback. In: Azzopardi, L., Kazai, G., Robertson, S., Rüger, S., Shokouhi, M., Song, D., Yilmaz, E. (eds.) ICTIR 2009. LNCS, vol. 5766, pp. 104–115. Springer, Heidelberg (2009)
19. Robertson, S.E.: On term selection for query expansion. Journal of Documentation 46(4), 359–364 (1990)
20. Buckley, C., Mitra, M., Walz, J.A., Cardie, C.: Using clustering and superconcepts within SMART: TREC 6. Inf. Process. Manage. 36(1), 109–131 (2000)
21. Ding, C., Peng, H.: Minimum redundancy feature selection from microarray gene expression data. In: CSB 2003: Proceedings of the IEEE Computer Society Conference on Bioinformatics, Washington, DC, USA, p. 523. IEEE Computer Society Press, Los Alamitos (2003)
22. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(8), 1226–1238 (2005)