Selecting Expansion Terms as a Set via Integer Linear Programming

Qi Zhang, Yan Wu, Xuanjing Huang
School of Computer Science, Fudan University
825 Zhangheng Road, Shanghai, P.R.China
{qi_zhang, 09110240024, xjhuang}@fudan.edu.cn

ABSTRACT

Pseudo-relevance feedback via query expansion has been widely studied from various perspectives in the past decades. Its effectiveness in improving retrieval effectiveness has been shown in many tasks. A variety of criteria were proposed to select additional terms for the original queries. However, most of the existing methods weight and select terms individually and do not consider the impact of term-to-term relationship. In this paper, we first examine the influence of combinations of terms through data analysis, which demonstrate the significant effect of term-to-term relationship on retrieval effectiveness. Then, to address this problem, we formalize the query expansion task as an integer linear programming (ILP) problem. The model combines the weights learned from a supervised method for individual terms, and integrates constraints to capture relations between terms. Finally, three standard TREC collections are used to evaluate the proposed method. Experimental results demonstrate that the proposed method can significantly improve the effectiveness of retrieval.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Relevance feedback

General Terms

Algorithms.

Keywords

Integer Linear Programming, Relevance feedback

1. INTRODUCTION

Query expansion is a topic that has been studied for a long time. It is an important technique for improving effectiveness of information retrieval. As users' queries are usually too short to describe the accurate information they need, it has received much more attention in recent years [2, 8, 12, 19, 17]. Pseudo-relevance feedback (PRF) is one of the most attractive and well-known expansion techniques. It does not require any user input and assumes that a small number of top-ranked documents in the initial retrieval result are relevant.

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

A variety of approaches and methodologies have been proposed to select expansion terms from these documents[3, 20, 16, 11]. These methods studied different criteria, thesaurus, web resources, linguistic features and many other aspects. While most of the existing approaches generally can improve the retrieval effectiveness, the closed-form term-weighting formulas cannot easily handle relations between terms. Through data analysis, we observe that combination of expansion terms can significantly impact the retrieval effectiveness.

In this paper, we propose a novel formulation, which converts the query expansion task into an integer linear programming problem (ILP) [1]. An objective function and a number of constraints which capture the relationships between terms are specified. ILP is a well-studied optimization framework and can be used to search the entire space to select terms. The formulation provides a flexible framework for integrating different criteria as objective functions or constraints.

The major contribution of this work can be summarized as follows: 1) We analyze the impact of relationships between terms and their combinations for pseudo-relevance feedback. 2) We propose a novel formulation of the query expansion task as a integer linear programming problem. 3) Expansion terms are selected as a set taking into account their relationships. 4) Experimental results using three standard TREC collections demonstrate that the proposed method performs better than state-of-the-art algorithms.

This paper is organized as follows: related work and state-of-the-art approaches are reviewed in Section 2. The proposed approach is detailed in Section 3. Experimental results using TREC collections are described and discussed in Section 4. Section 5 concludes the paper.

2. RELATED WORK

Relevance feedback (RF) and pseudo-relevance feedback (PRF) are two main types of query expansion techniques with a long history. They have received much attention due to their effectiveness in improving retrieval performance. The Rocchio formulation [15] is one classical RF approach. It reformulates the query in such a way that it gets closer to relevant documents and away from non-relevant documents in the vector space model. As no user input is required, pseudo-relevance feedback (PRF) has been widely studied in recent years. It usually assumes that the top "n" ranked documents are relevant to the query. Carpineto et al. [5] introduced several information-theoretic methods (e.g. Rocchio's weights, Robertson Selection Value (RSV), CHI-squared and so on) for query expansion. External resources, such as WordNet [13], Wikipedia [19], and user logs [10], have also been used in various pseudo-relevance feedback methods.

Recently, some approaches that take into account term-to-term

relations have been suggested. The methods proposed by Udupa et al. [17], Collins-Thompson [7] and [18] are most similar to our work. Udupa et al. claimed that the effect of including a term in an expansion set depended on the other terms in that expansion set. They proposed the use of spectral partitioning of the term-to-term interaction matrix to take into account term interactions. Collins-Thompson [7] formulated the query expansion as a convex robust optimization problem. However the risk-reward tradeoff of expansion is the main target of his work. Different with his method, we directly model the relation between terms in this work. [18] proposed the use of maximum relevance and minimum redundancy criteria to select terms as a whole. Term distributions and linguistic features are used to measure the relevance.

3. ILP-BASED QUERY EXPANSION

The objective of query expansion is to select words related to the query and use all the words rather than expanding each word separately [9]. The selected terms should be relevant to the query and subject to constraints which include the number of terms, and whether terms can be simultaneously selected. These constraints are global, and can not be adequately satisfied by selecting terms individually. Therefore, in this work, we formalize the problem as an integer linear programming (ILP) problem. This is a well-studied optimization framework and can be efficiently solved using standard optimization tools.

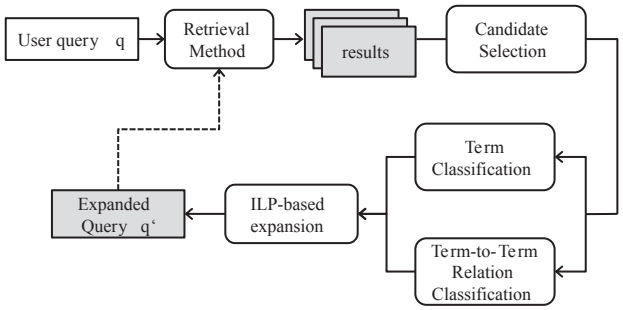

Figure 1: Processing flow of the ILP-based query expansion method.

Fig.1 shows the process flow of the proposed approach. Generally there are four phases in expanding queries: (1) select candidate terms, (2) classify individual terms according to helpfulness, (3) determine the term-to-term relations, (4) select expansion terms. These phases are detailed in the following sections.

3.1 Candidate selection

As there are too many terms that can be extracted from the pseudo-relevant documents, we use the following score function to select top words as candidates:

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

where N is the total number of documents and R is the number of relevant documents. The number of documents and relevant documents containing term t_i are respectively represented by n_i and r_i [14].

3.2 Term classification

In order to determine whether the expansion terms extracted are useful for improving retrieval performance we use SVM, a supervised learning method, for individual term selection. The training corpus is constructed according to the MAP change rate $chg(t_i)$ of terms, which is described in the previous section. Terms whose MAP change rate is higher than 0.01 are selected as good expansion terms. Bad expansion terms are those that diminish performance.

Each expansion term is represented by a feature vector. Previous studies have introduced a number of useful features. In this work, we use the following features:

• **Term distribution**. Term distribution in the pseudo-relevant documents has been used in many related works. In this paper, we define it as follows:

$$f_1(t) = \log \frac{\sum_{D \in F} tf(t, D)}{\sum_w \sum_{D \in F} tf(w, D)}$$

where F is the set of feedback documents.

• **Document frequency**. Document frequency in the pseudo-relevant documents is defined as follows:

$$f_2(t) = \log \frac{dn(t, F)}{|F|}$$

where $dn(t, F)$ represents the number of documents where the term t appears in the set of documents F.

• **Co-occurrence with the single query term**. It has been shown that the terms that co-occur with the query terms are usually related to the query.

$$f_3(t) = \log \frac{1}{n} \sum_{i=1}^{n} \frac{\sum_{D \in F} C(q_i, t|D)}{\sum_w \sum_{D \in F} tf(w, D)}$$

Where $C(q_i, t|D)$ is the frequency of co-occurrence of query term q_i and the expansion term t within text windows of the document D.

• **Dice's coefficient**. This is another popular association measure which has been widely studied.

$$f_4(t) = \frac{1}{n} \sum_{i=1}^{n} \log \sum_{D \in F} \frac{C(q_i, t|D)}{C(q_i|D) + C(t|D)}$$

Where $C(q_i, t|D)$ is same as in f_3, $C(q_i|D)$ is the number of occurrences of query term q_i in the document D.

3.3 Term-to-term relation classification

As mentioned in the previous section, although each individual term can improve retrieval performance, their combination may have a negative impact on the final result. In order to model this, we convert the problem into a classification task. We try to identify whether two terms together have a harmful or helpful effect on retrieval effectiveness.

MAP change rates, like the term classification task, can also be calculated from existing TREC collections. This is used to generate the training corpus. Any classifier can be used for the term-to-term relation classification. In the current work, we use SVM. In addition to the features used in the term classification, we also take into account the following additional feature:

• **Co-occurrence of two terms**. It has been shown that co-occurrence of two terms can be described as follows:

$$f_5(t_i, t_j) = \frac{\sum_{D \in F} C(t_i, t_j|D)}{\left(\sum_{D \in F} C(t_i|D) \cdot \sum_{D \in F} C(t_j|D)\right)^{\frac{1}{2}}}$$

where $C(t_i, t_j|D)$ is the frequency of co-occurrence of the two terms t_i and t_j within text windows in the document D.

3.4 ILP-based query expansion

Integer Linear Programming (ILP) denotes a set of constraint optimization problems which have a linear objective function,

are subject to linear equality and linear inequality constraints, and require the objective variables to be integers. With ILP formalization, the query expansion task is treated as a two class labeling problem. Given a candidate set of terms S, for each term $t \in S$, a term is selected as an expansion term (assign label "1" to the term), or not (assign label "0"). A vector of binary variables $X = (x_1, x_2, ..., x_n)$ is used over term $t \in D$, to indicate whether the candidate term should be selected or not. When the objective function is put together with all the constraints, the ILP algorithm to select expansion terms is determined as follows:

$$\text{maximize} \quad C^T X \tag{2}$$

$$\text{subject to} \quad \sum_{i=1}^{n} x_i \leq K \tag{3}$$

$$x_i + x_j \leq 1\,, \ \text{if SC}(t_i, t_j) \leq \xi \tag{4}$$

$$0 \leq x_i \leq 1 \tag{5}$$

$$X \in \mathbf{Z}^n \tag{6}$$

The objective function Eq.(2) denotes the expected scores over all the words of a solution X. $C = (c_1, c_2, ..., c_n)$ is defined as the assignment value. The variable c_i gives the value of labeling t_i as a expansion term. We use result of the term classification to model the importance c_i of the candidate term t_i.

In order to restrict the number of selected terms, we introduce Eq.(3) as one of the constraints. Eq.(4) ensures that the terms which have a negative influence will not be selected together. The SC(t_i, t_j) in Eq.(4) is given through term-to-term relation classification, which is described in the section 4.3.

4. RESULTS AND DISCUSSION

4.1 Data collection and implementation

Methods were evaluated with three TREC corpora: Disk4&5 and WT10g. Three test collections, TREC 7, TREC 8, and TREC 10 Web were used in the experiments. Table 4.1 shows statistics related to the collections. Disk4&5 are part of NIST TREC Document Databases which are distributed for testing of IR systems. A number of TREC ad hoc tracks have used this corpus to evaluate systems. Topics 351 to 450 are used as queries. In order to evaluate our methods in a more realistic environment, we also assessed them in WT10g, used by the TREC 10 Web track. This contains more than 1.6 million documents collected from about 11,000 servers.

Test Collection	Topics	#Docs	Size
TREC 7 TREC 8	351-400 401-450	556,077	1.86GB
TREC 10 Web	501-550	1,692,096	11GB

Table 1: Statistics of the evaluation corpus

The SVM implementation SVM^{light}[1] is applied to perform the classifications. For the ILP solver, we use YALMIP [2] to estimate the optimal solution from Eq.(2) to Eq.(6). The implementation of our expansion method is based on Indri 5.0 [3], where the retrieval model is based on a combination of language modeling and inference network retrieval frameworks. Only the title of each TREC topic is used for the initial query. The main evaluation metric are Mean Average Precision (MAP) and the precision at top 10 (P@10) for top 1000 documents.

[1] http://svmlight.joachims.org/
[2] http://users.isy.liu.se/johanl/yalmip/
[3] http://www.lemurproject.org/indri/

4.2 Experimental results

The main purpose of the experiment was to demonstrate the performance of the proposed ILP-based query expansion method. Several experiments were designed for this purpose. As the term classification has already been evaluated by [4], we focused in this work on the term-to-term relation classification and the ILP-based expansion method.

	NoExp	Base-FB	SVM[4]	REXP-FB[6]	ILP-FB
Trec7	0.1810	0.2081	0.2208	0.2106	**0.2465**
Trec8	0.1941	0.2226	——	0.2199	**0.3317**
WT10g	0.1724	0.1819	——	0.1990	**0.2099**

Table 2: Comparison of performance using MAP for all test collections. Values in boldface indicate statistically significant improvement over the REXP-FB method. The paired τ-test (ρ <0.05) is used to measure significance.

Table 2 summarizes results for all test collections using different query expansion methods. The mean average precision (MAP) for the top 1000 documents is used as the evaluation metric. The left-hand column in the table shows the collection names. The *NoExp* column represents the results without query expansion using Indri. For the pseudo-relevance feedback run *Base-FB*, which is an adaptation of Lavrenko's relevance models, a built-in model in Indri, 20 terms are extracted from the 50 top-ranked documents. Results of method proposed by Cao et al. [4] are shown in the *SVM*. The *REXP-FB* column represents the result obtained by the state-of-the-art method REXP feedback [6]. The values in the last column *ILP-FB* are the results of our method.

In all three collections, the results of the ILP-based expansion method are statistically better than both the Indri baseline expansion method and REXP method. In TREC 8, our method shows a better than 70.8% relative improvement over the original retrieval result. The relative improvements over the state-of-the-art method REXP is also greater than 50.8%. In TREC 7 and WT10g, the improvements of the method are also significant. This demonstrates the necessity of considering the relations between terms.

		No-Cons	ILP-FB
Trec7	Map	0.2253	**0.2465**
	P@10	0.5080	**0.5900**
Trec8	MAP	0.2500	**0.3317**
	P@10	0.4800	**0.6300**
WT10g	MAP	0.2000	**0.2099**
	P@10	0.3333	0.3384

Table 3: Comparison of performance using MAP and P@10 for the ILP-based method with and without term-to-term constraints. Values in boldface indicate statistically significant improvement over the No-Cons (no constraints) method. The paired τ-test (ρ <0.05) is used to measure significance.

Compared to previous methods, adding the term-to-term relations into constraints is one of the main contributions of this work. It is then important to see whether these constraints contribute significantly. Table 3 summaries the results of ILP-based method with

and without these constraints. These results show that ILP-based expansion methods with term-to-term constraints are statistically better than methods without these constraints. It also demonstrates the importance of considering term-to-term relations.

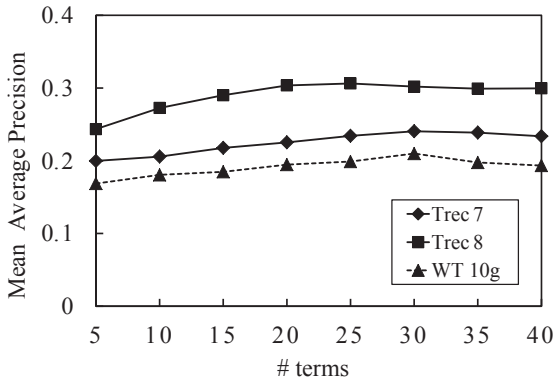

Figure 2: Results of varying the number of expansion terms for all collections.

Figure 2 shows the change of MAP as the number of expansion terms varies for all three collections. We observe that MAP is improved when expansion terms are combined with the initial query. However, MAP does not continue to improve when more than 25 terms are used for expansion. This is mainly due to a reduction in the quality of terms. Although the best parameters are different in different collections, significant improvement can be achieved when 15$\sim$20 terms are selected.

5. CONCLUSION

In this paper, we studied the impact of the relationships between terms and their combinations. In order to address the problem, we considered the query expansion task as an integer linear programming problem and used two classification models to obtain the objective function and constraints. In all three test collections, the proposed expansion method can significantly improve the retrieval result. Experimental results also demonstrate that the proposed method can significantly improve performance.

6. ACKNOWLEDGEMENT

The author wishes to thank the anonymous reviewers for their helpful comments. This work was partially funded by 973 Program (2010CB327900), National Natural Science Foundation of China (61003092, 61073069), Shanghai Leading Academic Discipline Project (B114), and “Chen Guang” project supported by Shanghai Municipal Education Commission and Shanghai Education Development Foundation (11CG05).

7. REFERENCES

[1] D. Alevras and M. W. Padberg. *Linear Optimization and Extensions: Problems and Solutions*. Springer, 2001.

[2] J. Bhogal, A. Macfarlane, and P. Smith. A review of ontology based query expansion. *Information Processing & Management*, 43(4):866–886, 2007.

[3] C. Buckley. Automatic query expansion using smart : Trec 3. In *Proceedings of The third Text REtrieval Conference (TREC-3)*, pages 69–80, 1994.

[4] G. Cao, J.-Y. Nie, J. Gao, and S. Robertson. Selecting good expansion terms for pseudo-relevance feedback. In *SIGIR ’08*, pages 243–250, 2008.

[5] C. Carpineto, R. de Mori, G. Romano, and B. Bigi. An information-theoretic approach to automatic query expansion. *ACM Transactions on Information Systems*, 19(1):1–27, 2001.

[6] K. Collins-Thompson. Estimating robust query models with convex optimization. In *Advances in Neural Information Processing Systems 21 (NIPS)*, pages 329–336, 2008.

[7] K. Collins-Thompson. Reducing the risk of query expansion via robust constrained optimization. In *Proceedings of the Eighteenth International Conference on Information and Knowledge Management (CIKM 2009)*, pages 329–336, Hong Kong, China, 2009. ACM.

[8] K. Collins-Thompson and J. Callan. Estimation and use of uncertainty in pseudo-relevance feedback. In *SIGIR ’07*, pages 303–310, 2007.

[9] B. Croft, D. Metzler, and T. Strohman. *Search Engines: Information Retrieval in Practice*. Addison-Wesley Publishing Company, USA, 1st edition, 2009.

[10] H. Cui, J.-R. Wen, J.-Y. Nie, and W.-Y. Ma. Query expansion by mining user logs. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):829–839, 2003.

[11] X. Huang and W. B. Croft. A unified relevance model for opinion retrieval. In *Proceedings of 16th Conference on Information and Knowledge Management(CIKM 2009)*, Hong Kong, China, 2009.

[12] K. S. Lee, W. B. Croft, and J. Allan. A cluster-based resampling method for pseudo-relevance feedback. In *SIGIR ’08*, pages 235–242, 2008.

[13] D. I. Moldovan and R. Mihalcea. Using wordnet and lexical operators to improve internet searches. *IEEE Internet Computing*, 4(1):34–43, 2000.

[14] S. E. Robertson. On term selection for query expansion. *Journal of Documentation*, 46(4):359–364, 1990.

[15] J. Rocchio. *Relevance Feedback in Information Retrieval*, pages 313–323. 1971.

[16] R. Sun, C.-H. Ong, and T.-S. Chua. Mining dependency relations for query expansion in passage retrieval. In *Proceedings of SIGIR 2006*, pages 382–389, 2006.

[17] R. Udupa, A. Bhole, and P. Bhattacharyya. "a term is known by the company it keeps": On selecting a good expansion set in pseudo-relevance feedback. In *ICTIR ’09: Proceedings of the 2nd International Conference on Theory of Information Retrieval*, pages 104–115, Berlin, Heidelberg, 2009. Springer-Verlag.

[18] Y. Wu, Q. Zhang, Y. Zhou, and X. Huang. Pseudo-relevance feedback based on mrmr criteria. In P.-J. Cheng, M.-Y. Kan, W. Lam, and P. Nakov, editors, *Information Retrieval Technology*, volume 6458 of *Lecture Notes in Computer Science*, pages 211–220. Springer Berlin / Heidelberg, 2010.

[19] Y. Xu, G. J. Jones, and B. Wang. Query dependent pseudo-relevance feedback based on wikipedia. In *SIGIR ’09*, pages 59–66, 2009.

[20] S. Yu, D. Cai, J.-R. Wen, and W.-Y. Ma. Improving pseudo-relevance feedback in web information retrieval using web page segmentation. In *Proceedings of WWW 2003*, pages 11–18, 2003.