How Effective is Query Expansion for Finding Novel Information?

Min Zhang[1], Chuan Lin[2], and Shaoping Ma[1]

[1] State Key Lab of Intelligent Tech. and Sys., Tsinghua University, Beijing, 100084, China
{z-m, msp}@tsinghua.edu.cn
[2] State Key Lab of Intelligent Tech. and Sys., Tsinghua University, Beijing, 100084, China
lch@mail.lits.tsinghua.edu.cn

Abstract. The task of finding novel information in information retrieval (IR) has been pro-posed recently and paid more attention to. Compared with techniques in traditional document-level retrieval, query expansion (QE) is dominant in the new task. This paper gives an empirical study on the effectiveness of different QE techniques on finding novel information. The conclusion is drawn according to experiments on two standard test collections of TREC2002 and TREC2003 novelty tracks. Local co-occurrence-based QE approach performs best and makes more than 15% consistent improvement, which enhances both precision and recall in some cases. Proximity-based and dependency-based QE are also effective that both make about 10% progress. Pseudo relevance feedback works better than semantics-based QE and the latter one is not helpful on finding novel information. [3]

1 Introduction

Information retrieval (IR) techniques have become dominant in finding information in people's daily life. Current systems return ranked lists of documents as the answer for an information request. It is most possibly, however, that not whole documents are useful to the user, and lots of redundancy exists.

One of approaches to provide direct information to users is question-answering. Another one would be to return only relevant AND new sentences (within context) rather than whole documents containing duplicate and extraneous information. The latter one is named finding novel information, which has been paid more attention to recently. TREC (Text REtrieval Conference), which is one of the most famous conferences in IR, proposed a new track named novelty in 2002.

Being compared to document-level IR, term mismatch problem is more considerable in sentence-level novel information finding because of the short content in sentences. In all of the current IR models, information is represented as terms,

[3] Supported by the Chinese Natural Science Foundation (60223004, 60321002, 60303005), and partially sponsored by the joint project with IBM China research.

namely characters, words or phrases. Only if at least one query term appears in a document, the document may be selected. In natural language, however, one concept can always be expressed using different terms. It leads to term mismatch and the possible missing of useful information.

To solve this term mismatch problem, query expansion (QE) techniques have been proposed. Generally there're three branches of QE approaches. One is to expand query using a global thesaurus which can be constructed according to semantic knowledge [1][2], or learned by statistical relations such as co-occurrence and mutual information [3][4][5][6][7], or got by some syntax-based learning [8][9][10][11] such as dependency-based word similarity [12]. The second kind of QE is to use thesaurus learned by local collection information [11]. And the third one is to expand query by pseudo relevance feedback [13][14].

In traditional IR, the effectiveness of QE technologies is instable. Voorhees [10] tried different weights to expansion terms, even manually selecting terms, but got less than 2% improvement. In the new task of finding novel information, however, how effective are QE-based technologies? This paper makes an empirical study on three kinds of QE approaches for finding novel information, and gives a comparison of their effectiveness.

The remaining part of the paper is constructed as follow: Section 2 describes the global thesaurus-based expansion, including thesauri based on semantics, statistical proximity, and dependency. Section 3 gives a brief introduction to pseudo relevance feedback. Section 4 shows an algorithm of finding novel information with QE based on local co-occurrence. Experiments and analysis are addressed in section 5. Finally the conclusion is drawn.

2 Global Thesaurus-Based Expansion

2.1 Thesaurus Based on Semantics

In this kind of approaches, people construct a thesaurus manually and select terms that have the same or similar semantic meanings. Therefore the noise taken into the thesaurus is relatively less. But the manual classifications of words are always too sensitive or too rough, hence it is difficult to be decided to what extent terms should be added.

Since WordNet [1](http://www.cogsci.princeton.edu/~wn/) is such a semantic thesaurus for English words that is used most widely, it was selected in our study. Totally three kinds of information were observed in experiments: hyponyms (descendants), synonyms and coordinated words. Effects of different levels of hyponyms have been studied.

2.2 Thesaurus Based on Statistical Proximity

Research of using statistical approaches on a large corpus is based on a distribution hypothesis which states that two words are semantically similar to the extent that they share contexts [15].

Dr Dekang Lin has made an in-depth study and provided an online dictionary (http://www.cs.ualberta.ca/~lindek/demos/proxysim.htm) based on statistical proximity on a generally corpus [7], which is one of the best thesauri of this kind of approaches. Hence this thesaurus is used as the representative in our study.

2.3 Thesaurus Based on Dependency

This kind of researches is to combine the statistical and the syntax information. If two terms are frequently have same or similar dependencies accord-ing to a general corpus, they are taken as similar or having a tight relationship.

In this paper, we use Dr Dekang Lin's online thesaurus based on statistical dependency [12] (http://www.cs.ualberta.ca/~lindek/demos/depsim.htm).

3 Pseudo Relevance Feedback

Pseudo relevance feedback strategies are to expand the query with top n terms extracted from the top m initial retrieved documents after initial search. In the task of finding novel information, each sentence is taken as an individual document. The n terms are chosen based on their similarities to the query [14].

4 Local Thesaurus-Based Expansion

Compared with global thesauri, thesaurus learned from local collection has the advantage that relations between words reflect the characteristics of retrieving collections directly. Therefore it may be more helpful. Following gives an algorithm of finding novel information with QE based on local co-occurrence (called *LCE*). It expands terms highly co-occurred with any of query terms in a fixed window size within a sentence in the retrieving collection.

The algorithm is described as following:

Suppose given a user query Q, the set of sentences in the collection is S;

1. Filter all stop-words in sentences in S;

2. To each $q_i \in Q$:

1) Construct a co-occurrence vector T_i:
$T_i = ((t_{i1}, f_{i1}), (t_{i2}, f_{i2}) \dots (t_{in}, f_{in}))$,
where t_{ij} is the j^{th} term co-occurred with q_i in the window (size N),
f_{ij} is the co-occurrence frequency of t_{ij} and q_i;

2) Normalize co-occurrence frequencies in T_i, and get a new vector T_i':
$T_i' = ((t_{i1}, f_{i1}'), (t_{i2}, f_{i2}') \dots (t_{in}, f_{in}'))$, *$f_{ij}'$ is the normalized score of f_{ij};*

3) Select terms that $f_{i1}' \geq \theta_i (\theta_i \geq 0)$ to expand q_i and forms a new query Q';

3. Find novel information in S using expanded new query Q'.

5 Experiments and Analysis

For finding novel information task, two standard test collections are available: TREC (Text REtrieval Conference)'2002 & TREC'2003 novelty tracks test sets. In each set, there are 50 queries, a collection of supposed relevant documents and a set of identified sentences with relevant and new information (called *qrels*) for each query. The *qrels* are generated by assessors.

The two test sets, referred as novelty 2002 and novelty 2003, respectively, are extremely different from each other. Collection of novelty 2002 is poorly relevant to the given user queries. And that of novelty 2003 can be taken as highly reliable relevant sets of documents for queries. Empirical studies have been performed on both test sets.

The results are evaluated in terms of *precision*, *recall* and *F-measure*. Sometimes $P \times R$ is also used as one metric.

$$precision = \frac{\#\ relevant\ (or\ new)\ sentences\ matched}{\#\ sentences\ retrieved}$$

$$recall = \frac{\#\ relevant\ (or\ new)\ sentences\ matched}{\#\ relevant\ (or\ new)\ sentences}$$

Since the task of finding novel information is quite difficult than traditional document-level IR task, the precision and recall are much lower.

5.1 Using Global Thesaurus

Semantics-Based QE. Table 1 and Table 2 show the QE effects of using different semantic relations, namely hyponyms, synonyms and coordinates, using WordNet on novelty 2002 and 2003 respectively.

Table 1. Effects of QE using WordNet semantic relations (novelty 2002)

methods	precision	recall	F-measure	$P \times R$
unexpanded	0.20	0.28	0.197	0.064
Hyponyms	0.18	0.32	0.197	0.066
Synset	0.17	0.32	0.195	0.068
Coordinate	0.18	0.29	0.189	0.061

In novelty 2002 experiments, expansion based on synonyms achieves trivial improvement in terms of average $P \times R$ while it does not help in terms of *F-measure*. On novelty 2003, results are better. A little improvement (+3.1%) of system performance has been obtained using F-measure.

The advantage of using global semantic thesaurus is that the noise taken into the thesaurus is relatively less. But there're two main disadvantages: First,

Table 2. Effects of QE using WordNet synonyms and hyponyms (novelty 2003)

methods	precision	recall	F-measure
unexpanded	0.633	0.637	0.552
Hyponyms	0.618	0.680	0.569
Synset	0.625	0.665	0.564

manual classifications of words are always too sensitive or too rough, hence it is difficult to be decided to what extent terms should be expanded. Second, because of the ambiguity of natural language, similarities of terms can not be confirmed without context, and therefore expansion based on a global semantic thesaurus is no longer reliable.

Proximity-Based QE. Table 3 and Table 4 describe the effect of proximity-based QE on both test sets. Encouraging results have been achieved, especially on novelty 2003 collection, that by using proximity-based expansion, about 10.1% improvement is obtained in terms of F-measure. Also it is shown that after expansion, precision is decreased while the recall is increased, and the overall improvement is made.

Such statistical proximity-based QE approach changes the semantic similarity to the proximity relation, and hence it avoids some difficulties such as word ambiguity, although the relationships between terms got by this approach have no clear explanation in natural language understanding.

Table 3. Effects of QE by proximity-based expansion (novelty 2002)

methods	precision	recall	F-measure	$P \times R$
unexpanded	0.20	0.28	0.197	0.064
proximity-based QE	0.19	0.30	0.200	0.066
improvement	-5.0%	+7.1%	+1.5%	+3.1%

Table 4. Effects of QE by proximity-based expansion (novelty 2003)

methods	precision	recall	F-measure
unexpanded	0.633	0.637	0.552
proximity-based QE	0.580	0.831	0.608
improvement	-8.4%	+30.4%	+10.1%

Dependency-Based QE. Experimental results of using dependency-based QE to find novel information is shown in the following two tables. The results are consistent on both test sets. This approach works a little better than proximity-based QE and gets 11.6% improvement on novelty 2003 in terms of F-measure, although the improvement in novelty 2002 is not so obvious.

Table 5. Effects of QE by dependency-based expansion (novelty 2002)

methods	precision	recall	F-measure	$P \times R$
unexpanded	0.20	0.28	0.197	0.064
dependency-based QE	0.19	0.31	0.200	0.067
improvement	-5.0%	+10.7%	+1.5%	+4.7%

Table 6. Effects of QE by proximity-based expansion (novelty 2003)

methods	precision	recall	F-measure
unexpanded	0.633	0.637	0.552
dependency-based QE	0.590	0.827	0.616
improvement	-6.8%	+29.8%	+11.6%

As mentioned in section 2.3, dependency-based expansion combines the statistical information and the syntax information. This is the most important advantage of such approaches. And it explains why this kind of QE performs better than statistical proximity-based QE. But the approach also has shortcomings. Since construct such a thesaurus should use a syntax parser, which is not precise, parsing errors will affect the quality of the thesaurus and hence hurt the performance of finding novel information.

5.2 Using Local Co-occurrence-Based QE

Effects of using LCE-based QE finding novel information are given in Table 7 and Table 8. The result is extremely good. It enhanced system performance greatly in both test sets. Two points are interesting when it is compared with other expansion techniques:

Firstly, LCE lead to consistent great improvement on both test sets which reach to 15.2% and 14.5% respectively, while other expansion approaches could hardly get much improvement in poorly relevant documents of novelty 2002.

Secondly, on novelty 2002 collection, local co-occurrence-based expansion (LCE) made consistent great progress in terms of both recall and precision, while other QE-based techniques improved the recall but hurt the precision.

When using LCE, characteristics of the retrieval collection have been taken into account, and hence the information provided by local thesaurus is more helpful.

Table 7. QE by local co-occurrence expansion (novelty 2002)

methods	precision	recall	F-measure	$P \times R$
unexpanded	0.20	0.28	0.197	0.064
LCE-based QE	0.21	0.34	0.227	0.081
improvement	+5.0%	+21.4%	+15.2%	+26.6%

Table 8. QE by local co-occurrence expansion (novelty 2003)

methods	precision	recall	F-measure
unexpanded	0.633	0.637	0.552
LCE-based QE, window = 1	0.557	0.866	0.613
improvement, window = 1	-12.0%	+36.0%	+11.1%
LCE-based QE, window = 10	0.536	0.955	0.632
improvement, window = 10	-15.3%	+49.9%	+14.5%

5.3 Pseudo Relevance Feedback

Table 9 gives the effect of pseudo relevance feedback on novelty 2003. More than 5% improvement has been achieved in all cases. And it shows that when more terms are added, the better performance is achieved.

Table 9. Effects of pseudo relevance feedback (to expand top M terms in top 3 documents) (novelty 2003)

methods	precision	recall	F-measure	improvement
unexpanded	0.633	0.637	0.552	–
M=10	0.593	0.716	0.584	+5.8%
M=15	0.590	0.732	0.587	+6.3%
M=100	0.589	0.744	0.594	+7.6%

5.4 Comparisons of Approaches

The overview of effects of QE approaches in finding novel information on novelty 2002 and novelty 2003 is shown in Figure 1 and Figure 2, respectively.

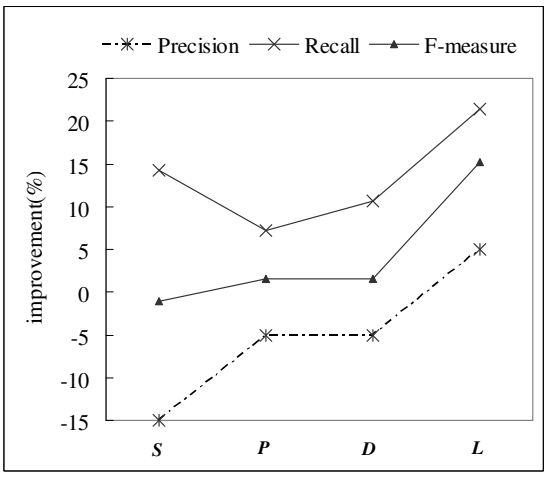

Fig. 1. Comparison of QE approaches (novelty 2002), S: Semantic-based QE, P: Proximity-based QE, D: Dependency-based QE, L: Local co-occurrence-based QE

Effects on novelty 2003 are better than that on novelty 2002. But the comparison results of effects of different QE approaches are consistent on both test sets. Local co-occurrence-based QE always performs best and makes more than 15% improvement. QE approaches based on statistical proximity and based on dependency are also helpful, improving system performance for about 10%. Pseudo relevance feedback works better than semantic-based QE and the latter one is not helpful to the task of finding novel information.

6 Conclusion

This paper gives an empirical study on the effect of different QE techniques in finding novel information. Three branches of approaches have been studied, namely QE based on global thesaurus, QE based on local co-occurrence information and pseudo relevance feedback. In global thesaurus-based approaches, following methods have been observed: QE based on semantics, statistical proximity, and dependency.

According to experiments on the two standard test collections of TREC'2002 and TREC'2003 novelty tracks, following conclusions can be drawn.

Firstly, the effect of QE-based approaches in finding novel information depends on the relevance of original documents collection, which comes from the initial results in traditional document-level retrieval.

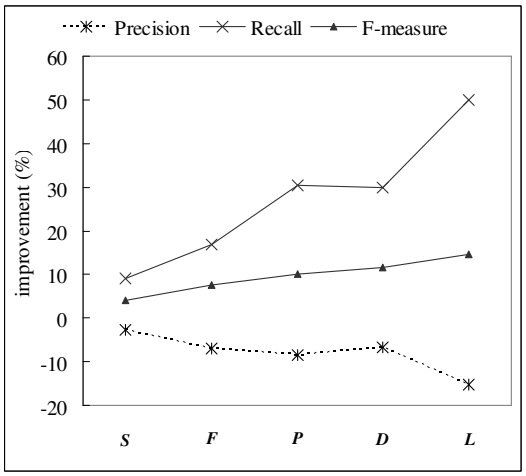

Fig. 2. Comparison of QE approaches (novelty 2003), S: Semantic-based QE, F: relevance feedback, P: Proximity-based QE, D: Dependency-based QE, L: Local co-occurrence-based QE

Secondly, comparison results of effects of different QE approaches are consistent. (1) LCE-based QE performs best and made more than 15% great and consistent improvement in both tests. (2) QE approaches based on statistical proximity and based on dependency are both effective to the task, making about 10% enhancement of performances in novelty 2003. And dependency-based QE is a little better. (3) Pseudo relevance feedback works better than semantics-based QE, and improves system performance at about 5%. (4) Semantic-based QE is not helpful on finding novel information.

Thirdly, LCE is the only QE approach that improves system performance in terms of both precision and recall in novelty 2002. Except for that, all QE approaches improve recall but hurt precision as the same time.

In the future, the effects of more QE techniques, especially those most recently proposed ones, will be studied on more test collections. And further analysis will be made.

References

1. Miller G. A., et al.: Introduction to WordNet: An on-line lexical database. International Journal of Lexicography (special issue), (1990) 3(4):235–312
2. Smeaton A. F. and Berrut. C.: Thresholding postings lists, query expansion by word-word distances and POS tagging of Spanish text. In Proceedings of the 4th Text Retrieval Conference (1996)
3. Van Rijbergen.: A theoretical basis for the use of co-occurrence data in information retrieval. Journal of Documentation, (1977) 106–119

4. Crouch, C. J., Yong, B.: Experiments in automatic statistical thesaurus construction. In Proceedings of 15th Int. ACM/SIGIR Conf on R&D in Information Retrieval, Copenhagen, Denmark (1992) 77–87
5. Schutze, H. and Pedersen, J.O.: A cooccurrence-based thesaurus and two applications to information retrieval. In Proceedings of RIAO'94. (1994) 266–274
6. Chen H., et al: Automatic thesaurus generation for an electronic community system. Journal of American Society for Information Science. 46(3): 175-193. (1995)
7. Lin D., et al.: Identifying Synonyms among Distributionally Similar Words. In Proceedings of IJCAI-03 (2003)
8. Ruge G.: Experiments on linguistically-based term associations. Information Processing and Management. (1992) 28(3):317–332
9. Grefenstette G.: Explorations in automatic thesaurus discovery. Kluwer Academic Publisher (1994)
10. Voorhees. E. M.: Query Expansion Using Lexical-Semantic Relations. In 17th Annual International ACM SIGIR conference (1994)
11. Xu J. and Croft. W.B.: Query Expansion Using Local and Global Document Analysis. In Proceedings of the 19th Annual International ACM SIGIR Conference. (1996) 4–11
12. Lin D. and Pantel. P.: Concept Discovery from Text. In Proceedings of Conference on Computational Linguistics 2002. Taipei, Taiwan. (2002) 577–583
13. Rocchio. J.: Relevance feedback in information retrieval. The Smart retrieval system— experiments in automatic document processing, Prentice-Hall, Englewood Cliffs, NJ, (1971) 313–323
14. Attar R. and Fraenkel A. S.: Local feedback in full-text retrieval systems. Journal of the ACM, 24(3): 397-417. (1977)
15. Harries Z.S.: Mathematical Structures of Language. New York, Wiley Publisher, (1968)