РАСШИРЕНИЕ ЗАПРОСА В ИНФОРМАЦИОННОМ ПОИСКЕ: ЧТО МЫ МОЖЕМ УЗНАТЬ ИЗ ГЛУБИННОГО АНАЛИЗА ЗАПРОСА?

Ермакова Л. М. (liana.ermakova@irit.fr)
Institut de Recherche en Informatique de Toulouse, CNRS UMR 5505, Université de Toulouse, Тулуза, Франция;
Пермский государственный национальный исследовательский университет, Пермь, Россия

Мот Ж. (josiane.mothe@irit.fr)
Institut de Recherche en Informatique de Toulouse, CNRS UMR 5505, Université de Toulouse, Тулуза, Франция

Овчинникова И. Г. (ira.ovchi@gmail.com)
Пермский государственный национальный исследовательский университет, Пермь, Россия

Одна из основных задач информационного поиска—извлечение документов, релевантных информационной потребности пользователя, выраженной запросом. Зачастую пользовательские запросы не превосходят 3 слов, что усложняет задачу. Многочисленные исследования показали, что автоматическое расширение запроса в среднем повышает точность, несмотря на то, что для некоторых запросов результаты ухудшаются. В статье предлагается новый метод автоматического расширения запроса, основанный на оценки важности слов-кандидатов, определяемой силой их связи со словами из запроса. Предлагаемый метод комбинирует локальный анализ, а именно обратную связь по релевантности, и глобальный анализ коллекции документов. Оценка метода была произведена на международных тестовых коллекциях, согласно установленным метрикам. Полученные результаты были сравнены с одной из лучших моделей, описанных в литературе. Системы показали сравнимые результаты в среднем. Однако глубинный анализ исходных и расширенных запросов позволил сделать выводы, которые могут помочь в исследовании данной области.

Ключевые слова: информационный поиск, расширение запроса, анализ слов запроса, обратная связь по релевантности, глобальный анализ, совстречаемость

QUERY EXPANSION IN INFORMATION RETRIEVAL: WHAT CAN WE LEARN FROM A DEEP ANALYSIS OF QUERIES?

Ermakova L. M. (liana.ermakova@irit.fr)
Institut de Recherche en Informatique de Toulouse, CNRS UMR 5505, Université de Toulouse, France;
State National Research University, Perm, Russia

Mothe J. (josiane.mothe@irit.fr)
Institut de Recherche en Informatique de Toulouse, CNRS UMR 5505, Université de Toulouse, France

Ovchinnikova I. G. (ira.ovchi@gmail.com)
Perm State National Research University, Perm, Russia

Information retrieval aims at retrieving relevant documents answering a user's need expressed through a query. Users' queries are generally less than 3 words which make a correct answer really difficult. Automatic query expansion (QE) improves the precision on average even if it can decrease the results for some queries. We propose a new automatic QE method that estimates the importance of expansion candidate terms by the strength of their relation to the query terms. The method combines local analysis and global analysis of texts. We evaluate the method using international benchmark collections and measures. We found comparable results on average compared to the Bo2 method. However, we show that a deep analysis of initial and expanded queries brings interesting insights that could help future research in the domain.

Key words: information retrieval, query expansion, analysis of query terms, relevance feedback, global analysis, co-occurrence

1. Introduction

Information Retrieval (IR) systems aim at retrieving information that answers a user's needs/he expresses through a query. Retrieving relevant information to a query implies a two-step process: off line, the system indexes documents, generally using a bag of words representation; online, the system computes the similarity between the user's query and the document representations (indexing terms) to retrieve the most similar documents.

IR systems have to face the problem of query term ambiguity inherent in natural language; it is even more a challenging problem since users' queries are very short (Chifu and Ionescu 2012). Indeed, more than 90% of the queries are 3 words or less long. Considering such a small number of terms, it is a challenge for the systems to "understand" the query or to disambiguate it.

To face these challenges, IR systems consider several strategies. One of them is to diversify the results by providing document related to the various senses of query terms, the system optimizes the chance of providing relevant information (Vargas et al. 2013). Another strategy is to expand the query (Gauch and Smith 1991).

The principle of query expansion (QE) is to add new query terms to the initial query in order to enhance its formulation. Candidate terms for expansion are either extracted from external resources such as WordNet or from the documents themselves; based on their links with the initial query terms. In the latter types of methods, the most popular one is the pseudo-relevance feedback (Buckley 1995).

Pseudo-relevance feedback has been shown as an effective method in average; however, it can lower results for some queries. For example, it is most probable that for poor performing queries query expansion is helpless since it will be based on the first retrieved documents that are probably non-relevant documents. It is thus important to know in advance if QE will be helpful or on the contrary if it will degrade the results. Selective query expansion aims at making this decision. However, current methods use blind methodologies and uses learning methods as black boxes. On the contrary, we think that a deep analysis of queries and query expansion terms could help understanding when QE would be useful and if there are some sort of typology of QE usefulness.

This paper pursues two objectives; first of all, we suggest a new automatic query expansion method. This new method estimates the importance of expansion candidate terms by the strength of their relation to the query terms. The second objective is to deeply analyze the results: both the initial and expanded queries and the terms they are composed of, and the cases when the expansion lowers the results and when it improves them.

The remaining of the paper is organized as follows: Section 2 reports related works. In Section 3, we present the new QE method we propose. Section 4 presents the evaluation framework. Section 5 discusses the results and presents a deep analysis of query terms, on a linguistic point of view.

2. Literature Review

2.1. QE and Pseudo Relevance Feedback

QE is either based on the analysis of a document collection (Carpineto and Romano 2012) or they imply dictionary or ontology-based methods (Bhogal, Macfarlane, and Smith 2007). This study focuses on document analysis methods. This analysis may be either (1) global (corpus analysis to detect word relationships) (Carpineto and Romano 2012) or (2) local feedback (analysis of documents retrieved by the initial query) (Rocchio 1971; Xu and Croft 1996). Local analysis or local feedback methods rely on the hypothesis that relevant documents contain terms that could be useful to enhance query formulation. Rocchio defined a method in which term weights are re-computed so that the terms that occur in relevant documents contribute positively to the new query whereas the weight of the terms from non-relevant documents are lowered (Rocchio 1971). Rocchio's method implies to know document relevance. Buckley suggested pseudo-relevance feedback (PRF) in which the top retrieved document are automatically considered as relevant (Buckley 1995). PRF is now common practice and used in many expansion methods (Carpineto and Romano 2012). Global methods work alike but in that case candidate terms come from the entire document collection rather than just (pseudo-) relevant documents.

Divergence from Randomness (DFR) models were developed by the School of Computing Science, University of Glasgow (Ounis et al. 2006). These models are based on the assumption that informative words are relatively more frequent in relevant documents than in others. During QE the best-scored terms from the top-ranked documents are extracted. Terms are ranked using one of the DFR weighting model. DFR models include Kullback-Leibler divergence, Chi-square divergence, Bose-Einstein 1 (Bo1) and 2 (Bo2) models. In the DFR models QE is performed by ordering the candidate terms by their information content given the query Q. DFR models are presented in (G. Amati 2003). In this paper, we compare our results with Bo2.

2.2. Analysis of Queries and Results

A few studies have reported analysis of results. The deeper analysis has been conducted in the RIA Workshop that took place in 2004. One of the objectives of the workshop was to analyze the variability in systems: some systems answering well on some queries and badly on others; some other systems behaving oppositely. One of the conclusion of the workshop was that the comprehension of variability is complex because of various parameters: query formulation, the relation between the query and the documents as well as the characteristics of the system (Harman and Buckley 2009). Moreover, they conducted failure analysis for 45 of the TREC topics. After using various systems on "hard" topics, the workshop participants analyzed why the system failed. For 39 topics out of 45 the systems failed for the same reason. Moreover, even if they did not retrieve the same documents, they were missing the same aspect in the top documents. Predicting query difficulty remains a challenge (Mothe and Tanguy 2005).

The work presented in this paper combines local analysis, namely relevance feedback, and global analysis.

3. Method Description

The key idea of our new QE method is to estimate the importance of candidate terms by the strength of their relation to the query terms. In contrast to DFR models we do not compare the term frequency in PRF and the entire collection. In our approach, documents from PRF provide term candidates that are analyzed in two aspects: their frequency in PRF and their co-occurrence with query terms in the whole collection. Indeed, DFR models are based on two metrics: term frequency in PRF and the frequency of the term in the collection. Particularly, Bo2 uses the extrapolation of term frequency in PRF on the whole collection.

In our method candidate terms are selected from the PRF and are based on the underlain hypotheses: the strength of their relation to the query terms is proportional to the fraction of the number of the documents containing both candidate terms and query terms and the product of the number of documents containing at least one of these sets.

A query is first preprocessed. It is cleared from stop-words, punctuation; duplicate terms are removed. However if a query contains only stop-words, this could mean that a user is interested, for example, in grammar. For instance, the query “a and the” may imply that a user wants to find how to use English articles. Thus, if a query contains only stop-words, we keep all of them.

The importance of term combinations w_{T_j} is estimated by the formula:

$$w_{T_j} = \sum_{t_i \in T_j} (Imp(t_i) + 1)$$

$$Imp(t_i) = \frac{1}{\log doccount(t_i)}$$

where T_j is the term combination, t_i is the i-th term from T_j, $doccount(t_i)$ is the number of documents containing the i-th term. $Imp(t_i)$ is similar to IDF. For widely-spread terms with low $Imp(t_i)$ the importance of their combination is approximately equal to their number.

The importance of candidate terms w_c is computed as follows:

$$w_c = TF(c) \times \sum_{T_j \in T} MI(T_j, c)$$

where T is the set of all possible term combinations, and $MI(T_j, c)$ is the analogue of non-negative point-wise mutual information calculated by the formula:

$$MI(T_j, c) = \frac{-\log_2 \max\left(\frac{doccount(T_j, c) \times n}{doccount(T_j) \times doccount(c)}\right)}{\log_2 \frac{doccount(T_j, c)}{n}}$$

where $doccount(c)$ is the number of documents containing the candidate term c, $doccount(T_j)$ is the number of documents containing all terms from the term combination T_j, $doccount(T_j, c)$ is the length of their intersection, and n is the total number of documents in the collection.

All weights w_c are normalized. The best-scored term candidates are selected for query expansion.

4. Evaluation

4.1. Data Collections and Evaluation Measures

The evaluation was performed on two kinds of datasets: TREC (Text Retrieval Conference) Ad Hoc Track data sets for three years (1997–1999) (Voorhees and Harman 2000) containing 150 topics in total and composed of news articles, and WT10g from TREC Web track 2000–2001 (Hawking and Craswell 2002) which is a 10GB subset of the web snapshot of 1997 from Internet Archive. There are 100 topics in this second collection from track 2000 and 2001.

In our evaluation, we considered the following evaluation measures:\begin{itemize}

- Precision at 10;
- Mean Average.

Precision (P) is the fraction of retrieved documents that are relevant. P at 10 (P@10) is the fraction of the top 10 retrieved documents that are relevant. Average precision (AP) is the average of precision computed each time a relevant document is retrieved. Mean average precision (MAP) is calculated as the mean of average precision over queries.

4.2. System Details

We compared our system (Co) with the Bo2 DFR model implemented in Terrier platform.

Both systems used InL2c1.0 model for PRF, 3 documents from which 10 best scored terms were extracted. InL2c1.0 is a DFR (divergence from randomness) model based on TF-IDF measure with L2 term frequency normalization (Gianni Amati and Van Rijsbergen 2002; He and Ounis 2005)Heidelberg”,”page”:”200–214”, ”event-place”: ”Berlin, Heidelberg”,”URL”:”http://dx.doi.org/10.1007/978-3-540-31865-1_15”,”DOI”: ”10.1007/978-3-540-31865-1_15”,”ISBN”:”3-540-25295-9, 978-3-540-25295-5”,”author”: [{“family”:”He”,”given”:”Ben”},{“family”:”Ounis”,”given”:”Iadh”}],”issued”:{“date-parts”: [[“2005”]]}}}],”schema”:”https://github.com/citation-style-language/schema/raw/master/csl-citation.json”}.

4.3. Performance Results

Table 1 provides information about the number of relevant retrieved documents (RRD), MAP, and P@10. On adhoc data set by all metrics our systems showed the best results, which are much higher than the baseline. The Student’s test confirmed that the differences between MAP values of Co and Bo2 is not significant at the level $p < 0.05$. Significant difference is marked by *, the best results are marked-up in bold. Our system showed lower results than Bo2 according to P@10. In case of web data, Bo2 obtained the same MAP score as Co. Co remains the best according to P@10.

Table 1. General results

		Co	Bo2	Baseline
TREC 6-8 data	RRD	8230	8184	7225
	MAP	0.2507	0.2491	0.2105*
	P@10	0.4400	0.4413	0.4180
Web data	RRD	3897	3935	3810
	MAP	0.2190	0.2190	0.1894*
	P@10	0.3327	0.3296	0.2816

5. Discussion and Further Analysis

5.1. Detailed Results

In the previous section, we reported the results when averaged over the set of topics. In this section, we aim at analyzing the results deeper.

Table 2. Detailed statistics

	# topics	# Baseline best	# Bo2 best	# Co best
Adhoc	150	33 (22%)	52 (34.6%)	63 (42%)
Adhoc 25 hardest	25	8 (32%)	7 (28%)	10 (40%)
Adhoc 25 easiest	25	7 (28%)	10 (40%)	7 (28%)
Web	98	34 (34.7%)	30 (30.6%)	31 (31.6%)
Web 25 hardest	25	11 (44%)	5 (20%)	7 (28%)
Web 25 easiest	25	10 (40%)	10 (40%)	4 (16%)

Table 2 reports the number of topics in each collection for which the method (column) got the best results according to AP. We also report these numbers when the 25 hardest and 25 easiest topics are considered. The hardest and easiest topics are defined as the ones that got the highest and lowest AP using the initial query. For example, in the Adhoc collection, from 150 topics, 33 are best treated without QE, 52 best when using Bo2 and 63 when using our method. The percentage of best treated topic per method is slightly different when considering the easiest topics: 7 are best without QE, 10 are best treated using Bo2 and 7 using our method.

Two examples are provided below. Each topic is composed of the title part which was used as a submitted query to the system, as well as descriptive and narrative parts that helps in understanding the user's need.

Example 1
<num>530
<title>do pheromone scents work?
<desc>What is the scientific evidence that suggests pheromones stimulate the opposite sex?
<narr>A relevant document will discuss how pheromones act as an attractor or repellent among humans, other animals, or plants.

Example 2.
<num>494
<title>nirvana
<desc>Find information on members of the rock group Nirvana.
<narr>Descriptions of members' behavior at various concerts and their performing style is relevant. Information on who wrote certain songs or a band member's role in producing a song is relevant. Biographical information on members is also relevant.

Table 3 provides the results in terms of AP and the various query formulations the system really used as well as the term weights. For example, with regard to topic 530, "do" was removed as a stop word and the other words have been stemmed.

Table 3. Query reformulation examples

# Topic	AP / Initial query	AP / Bo2 reformulation	AP / Co reformulation
530	0.3838 / pheromon^ 1.0 scent^ 1.0 work^ 1.0	0.1551/ pheromon^1.60 scent^ 1.30 work^1.00 fragranc^ 1.00 perfum^0.96 men^ 0.45 attract^ 0.41 design^ 0.41 sex^ 0.38 natur^ 0.34 sexual^ 0.34	0.1765 / pheromon^ 2.0 scent^ 1.68 work^ 1.42perfum^ 1.0fragranc^ 0.97 aphrodisiac^ 0.36 sex^ 0.22 men^ 0.22 attract^ 0.21 sexual^ 0.17 world^ 0.16 cologn^ 0.15 natur^ 0.15 nerd^ 0.13
494	0.1706 / nirvana^ 1.0	0.3508 / nirvana^ 2.00 kurt^0.29 cobain^ 0.25 bootleg^0.17 world^ 0.17 list^0.15 song^ 0.13 new^0.11 contain^ 0.11 dedic^0.11	0.5703 / nirvana^2.0 cobain^ 1.0 kurt^0.82 bootleg^ 0.42 song^0.30 nerd^ 0.26 music^ 0.18 unplug^0.18 band^ 0.17 world^0.16 sound^ 0.15 stuff^0.14

In topic 530 both reformulation went to the "sex" concept which leaded to some noise in the answers.

In topic 494, it is clear that adding “kurt” and “cobain”, the lead singer, guitarist, and primary songwriter of the band Nirvana help in retrieving relevant documents. Either the weight of the terms (which are stronger in our reformulation than in Bo2) or some additional terms such as “band” made our reformulation better than Bo2.

5.2. Types of Initial Queries

Types of initial queries play an essential role in the prediction of successful information retrieval. As usual initial queries include, besides articles and other grammar words, nouns and entities, sometimes attributes and verbs. Grammatical structure of a title does not influence on the information retrieval process, because every title while processing the query is ruined into words and even word chunks. So types of queries are limited by a number of words and topic. Types of the query terms are restricted to words grammatical classes, such as parts of speech, and words semantic classes, such as terminology, entities, peculiarities, etc.

Potentially a document matches the initial query thanks to one term, or one term with its attribute, or two (or more) different terms. The last possibility is the best one, since a number of documents with two (or more) unconnected terms from the initial query is less, than a number of documents with the term and its attribute (noun phrase). In other words, co-occurrence of two (or more) semantically unconnected query terms in a document guarantees more accurate matching the initial query, while occurrence of one term just presupposes matching in a topic. Thus, one term query is less informative, than two and more terms queries. Hence for a one-word initial query QE is a productive way to increase the relevance of results, however, it depends on semantics of the one-word query. Our results for QE for one-word queries are slightly better regardless of the QE methods.

As a consequence of the diffusive character of the category, there are a lot of different factors which influence on the document frequency of the words associated with the topic. Thus, the more texts we use for the QE in the global analysis, the more unpredictable candidates we get for the QE. That is why the query generated on the basis of the title *What is a Bengals Cat?* (Animals) provides slightly better results with Bo2 QE system, while, with our system, we get a lot of useless extensions.

The structure of scientific categories is more compact and hierarchical. We assume that the initial query in the field of scientific categories evokes texts with less associative and more logical connections. The QE allows directing the IR process in a narrow relevant field. The title *Unsolicited Faxes* (Ad Hoc) refers to a multi-topic document, which simultaneously belongs to at least two topics in our set (“crimes” and “technology”). The results of QE performed by both systems are very good, but for our system it is significantly better: 0.6638 against 0.6015.

Therefore, the topic of the initial query is a strong factor, which influences on the necessity of the QE. Within homogenous text collection, every QE system works good, producing better results, than an initial query. Within naive topics categories the simple QE system is appropriate, while our QE generates complicated associative queries. So for the IR on the topic from naive category within heterogeneous text collection our QE system is overcomplicated, and that is why it works worse.

6. Conclusion

In this paper, we first suggest a new method for QE we call Co. The key idea of the proposed method is to estimate the importance of candidate terms by the strength of their relation to the query terms.

In our experiments, we show that the Co method has similar results as the Bo2 method from the literature. However, our finer analysis shows that the type of initial query can have an influence on the success of QE.

In our future work, we will work on the relationship between the types of queries and the field associated to the query in order to detect correlation with these features and the best method to treat the query. We think that using more linguistic features can help in selective approaches in IR.

References

1. *Amati, G.* (2003), Probability Models for Information Retrieval Based on Divergence from Randomness: PhD Thesis. University of Glasgow.
2. *Amati, Gianni, and Cornelis Joost Van Rijsbergen* (2002), Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness. ACM Trans. Inf. Syst. Vol. 20 no. 4 pp. 357–389 (October).
3. *Bhogal, J., A. Macfarlane, and P. Smith* (2007), A Review of Ontology Based Query Expansion. Inf. Process. Manage. Vol. 43 no. 4 pp. 866–886 (July).
4. *Buckley, Chris* (1995), Automatic Query Expansion Using SMART : TREC 3. In Proceedings of The Third Text REtrieval Conference (TREC-3). NIST Special Publication 500–226. pp. 69–80. Gaithersburg, MD: National Institute of Standards and Technology (NIST).
5. *Carpineto, Claudio, and Giovanni Romano* (2012), A Survey of Automatic Query Expansion in Information Retrieval. ACM Computing Surveys. Vol. 44 no. 1 pp. 1–50 (January).
6. *Chifu, Adrian-Gabriel, and Radu-Tudor Ionescu* (2012), Word Sense Disambiguation to Improve Precision for Ambiguous Queries. Cent. Eur. J. Comp. Sci. no. in printno. in print.
7. *Gauch, Susan, and John B. Smith* (1991), Search Improvement via Automatic Query Reformulation. ACM Trans. Inf. Syst. Vol. 9no. 3pp. 249–280.
8. *Harman, Donna, and Chris Buckley* (2009), Overview of the Reliable Information Access Workshop. Information Retrieval. Vol. 12 no. 6 pp. 615–641 (July).
9. *Hawking, David, and Nick Craswell* (2002), Overview of the TREC-2001 Web Track. NIST Special Publication. pp. 61–67.
10. *He, Ben, and Iadh Ounis* (2005), Term Frequency Normalisation Tuning for BM25 and DFR Models. In Proceedings of the 27th European Conference on Advances in Information Retrieval Research. pp. 200–214. ECIR’05. Berlin, Heidelberg: Springer-Verlag. http://dx.doi.org/10.1007/978-3-540-31865-1_15.

11. *Mothe, J., and L. Tanguy* (2005), Linguistic Features to Predict Query Difficulty—A Case Study on Previous TREC Campaign. SIGIR Workshop on Predicting Query Difficulty—Methods and Applications.
12. *Ounis, I., G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma* (2006), Terrier: A High Performance and Scalable Information Retrieval Platform. In Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006). Seattle, Washington, USA.
13. *Rocchio, J.* (1971), Relevance Feedback in Information Retrieval. In The SMART Retrieval System, 313–323. http://scholar.google.com/scholar?hl=en&#38;lr=&#38;client=firefox-a&#38;q=relevance+feedback+in+information+retrieval&#38;btnG=Search.
14. *Vargas, S., R. L. T. Santos, C. Macdonald, and I. Ounis* (2013), Selecting Effective Expansion Terms for Diversity. In 10th International Conference in the RIAO Series (OAIR 2013). Lisbon, Portugal. http://ir.ii.uam.es/predict/pubs/oair2013-vargas-gla.pdf.
15. *Voorhees, Ellen M., and Donna Harman* (2000), Overview of the Ninth Text REtrieval Conference (TREC-9). In Proceedings of the Ninth Text REtrieval Conference (TREC-9). pp. 1–14.
16. *Xu, Jinxi, and W. Bruce Croft* (1996), Query Expansion Using Local and Global Document Analysis. In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 4–11. SIGIR'96. New York, NY, USA: ACM. http://doi.acm.org/10.1145/243199.243202.