УДК 004.78:025.4.036

DO FREQUENT MEDIA WORDS WORSEN QUERY EXPANSION?

Irina Ovchinnikova

Perm State University, Bukireva, 15, Perm 614990, Russia, ira.ovchi@gmail.com

Liana Ermakova,

LISIS – Université de Lorraine, 5, boulevard Descartes, 77454 Champs-sur-Marne, France, liana87@mail.ru

Josiane Mothe,

ESPE, Université de Toulouse, Institut de Recherche en Informatique de Toulouse, UMR5505 CNRS, 118 Route de Narbonne, F-31062 Toulouse Cedex 9, France, josiane.mothe@irit.fr

This paper offers a linguistic approach to the study of the potency of query expansion while retrieving information from the web. The expansion allows enhancing the results; however, some queries show lower effectiveness after expansion. The objective of the study is to analyze linguistic features of initial query (IQ) as predictors for the expansion potency by different systems. The IQ is considered as a 'bag of words' with their linguistic descriptions, frequency first of all. The interdependence of different linguistic features of a query term determines the term value and its validity for the expansion. Analyzing two sets of terms from IQ (from queries that failed and from queries that were improved after expansion), we found out the negative impact of frequent terms from media on query expansion. This effect reflects the semantic variety of the frequent term connections in texts of different genres.

Initial query, information retrieval, query expansion, word frequency, pragmatics, query analysis

1. IQ processing by a search engine results in extraction of documents with the query terms as frequent words. The small number of query terms (more than 90% of queries are 3 words or less long) generates a problem of ambiguity of users' queries. To solve this problem, some systems diversify their response by expanding the query with terms either being extracted from external resources such as WordNet or co-occurring in the documents containing initial query terms. In the latter case, the query expansion (QE) can use pseudo-relevance feedback which considers the first retrieved documents using the IQ to extract the terms to expand it. Sometimes pseudo-relevance feedback lowers results, demonstrating poorer relevance to users' interests

than the IQ. The poor results can be caused by non-relevant expansion of IQ with terms from non-relevant documents.

The objective of this paper is to reconsider the basis for prognosis of the necessity and productivity of the IQ expansion. We argue that the relevance of the expansion depends on the linguistic parameters of the IQ terms. Since IR systems deal with words and word chunks, their frequencies in texts of different genres and styles should have an impact. In this paper, we examine a possibility to recognize by virtue of frequency, whether the expansion will improve or impair the results.

2. We consider an IQ as a "bag of words" with their linguistic characteristics. We examine impact of the frequency in interaction with pragmatics on the effectiveness of IQ processing. Word frequency correlates with semantics and pragmatics on the basis of different rates of word occurrence in the various discourses.

In our previous publication [1] we analyze the effectiveness of QE using two different QE systems. Based on these results, we chose two sets of query terms. The first set consists of the terms from queries which failed after expansion by the systems (*Low* set). IQs for the control set (*Imp* set) were randomly chosen from the queries which were improved after expanding. Our research question is "Are there cues to decide whether a frequent term in an IQ need to be expanded or not?" To answer the question, we test the distributions of the word frequencies[1] for two sets of query terms. Our analysis is based on WT10G TREC collection[2]. TREC WT10G is a subset of the Internet archive making 10GB, containing more than 1.6 million documents and 98 requests with judgments of relevance.

3. The data are represented in the Table 1. Word frequencies in the *Low* set vary in wider limits than ones in the *Imp*.

Table 1

Frequency of terms in WordAndPhrase for *Low* and *Imp*

Set	**Average Total F**	**Me Total F**	**Average F newsp.**	**Me F newsp.**	**Average F magaz.**	**Me F magaz.**
Low	51796	27922	10309	6896	12151	7313

[1] http://www.wordandphrase.info/frequencyList.asp WordAndPhrase resource, based on COCA, represents 1,500,000 frequent English words in 60,000 lists 25 words each.

[2] trec.nist.gov.

Imp	39922	11103	5589	2022	6926	3081

Frequency of words in *Low* is essentially higher in media (F newspaper plus F magazine), as well as in all communicative spheres (Total F). Terms from *Low* occur in the corpora almost twice often than terms from *Imp*. We found out that the most frequent words from media tend to fail as a query terms after being expanded by their collocations.

Reference

1. Ermakova, L., Mothe, J., & Ovchinnikova, I. (2014). Query expansion in information retrieval: What can we learn from a deep analysis of queries? In: International Conference on Computational Linguistics-Dialogue 2014 (Vol. 20, No. 13, pp. 152-162).

ЧАСТОТНЫЕ В МЕДИА СЛОВА КАК ТЕРМИНЫ ПОИСКОВОГО ЗАПРОСА

Ирина Овчинникова

Пермский государственный национальный исследовательский университет, 614990, Россия, г. Пермь, ул. Букирева, 15, ira.ovchi@gmail.com

Лиана Ермакова,

Университет Лотарингии, 5, boulevard Descartes, 77454 Champs-sur-Marne, France, liana87@mail.ru

Жозианна Мот,

Институт Исследований Информационных Технологий, Тулуза, 118 Route de Narbonne, F-31062 Toulouse Cedex 9, France, josiane.mothe@irit.fr

В статье исследуется продуктивность автоматического расширения первичного запроса на основе анализа частотности и сферы использования его терминов. Расширение осуществляют различные системы извлечения информации. Нередко расширение запроса оказывается неудачным, ухудшает результативность поиска. Цель нашего исследования – выявить прогностические свойства терминов первичного запроса, которые позволили бы предсказать необходимость его расширения любой из систем и, если таковое необходимо, эффективность расширенного запроса. Первичный запрос представлен как набор слов с некоторыми лингвистическими параметрами. Основным параметром слова для ИП является его частотность. Частотность слова в определенном дискурсе предопределяет его ценность для запроса и

пригодность для расширения за счет коллокаций, синонимов и ассоциаций. В результате анализа двух наборов из терминов первичных запросов (ухудшенного (Low) и улучшенного (Imp) при расширении различными системами) мы обнаружили негативное влияние частотных в медиа слов на результаты поиска посредством расширенного запроса. Это негативное воздействие обусловлено разнообразием связей таких слов в текстах различных жанров и тематики.

Первичный запрос, извлечение информации, расширение запроса, частотность, прагматика, дискурс, анализ запроса

Acknowledgements

We would like to thank Université Paul Sabatier in Toulouse for the grant to welcome Irina Ovchinnikova for 1 month in 2015. It helped us to initiate this work.