BOOSTING WEB RETRIEVAL THROUGH QUERY OPERATIONS

Gilad Mishne Maarten de Rijke

Informatics Institute, University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam

Abstract

We explore the use of phrase and proximity terms in the context of web retrieval, which is different from traditional ad-hoc retrieval both in document structure and in query characteristics. We show that for this type of task, the usage of both phrase and proximity terms is highly beneficial for early precision as well as for overall retrieval effectiveness.

1 Introduction

An important aspect in which web retrieval differs from ad-hoc retrieval concerns the users needs. User studies and anecdotal evidence suggest that web users wish to spend as little time as possible going through the results, and are mostly interested in a small number of relevant documents in the topmost ranks. Most users look only at the first page of results (usually, containing 10 results) [11], and this trend is strengthening over time [10]. Moreover, web search users usually have short search sessions, indicating that once a user has followed a link to a document which she finds relevant, she will in most cases not return to the result list and examine further hits [1].

Accordingly, recent large-scale web search evaluations such as the web track at TREC [3] have broadened the traditional focus on evaluation measures such as Mean Average Precision (MAP) and Precision/Recall graphs to also include early precision based measures such as Precision@10, Precision@20 and Success@10; in some cases, even higher precision is evaluated, e.g., Mean Reciprocal Rank (MRR, mostly for tasks with a single relevant document).

The web continues to be an inspiring domain for retrieval research. For instance, the layout information embedded within HTML documents gave rise to retrieval models that take non-content features of documents into account [2]. Our focus in this paper is not on web retrieval models but on web *queries*. How can we boost web retrieval effectiveness, measured using any of the measures just mentioned, by means of automatic operations on queries?

2 Operations on Web Queries

An important difference between web retrieval and other retrieval tasks is the average query length. Web search user studies such as those mentioned earlier report on average lengths of 1.5 to 2.6 terms; web search engines report similar numbers [8]. In contrast, closed-domain searches have significantly higher average lengths, e.g., 4.9 terms for the TREC 2004 Genomics track [4].

For short queries it is especially important to make the most out of what little information they give us. We examine the effect of automatic query rewrites, specifically phrasal and proximity-based retrieval, on the performance of web retrieval. A phrase match between a document and a query is usually an accurate indication that the document deals with the aspect of the query described by the phrase. Intuitively, the ability to detect overlap between a document and a query aspect is particularly important if queries are short and may have very few aspects.

In the paper we are especially interested in the effectiveness of "light-weight" query operations for web retrieval. Thus, we do not consider phrases as indexing units, but submit queries that exploit phrases or proximity terms against an index consisting of single terms only. Also, our phrases are not syntactic or even statistical in nature; we simply treat every word n-gram from the query as a phrase. For us, proximity based retrieval is a natural extension of phrasal retrieval where the restriction on the nearness of the terms is somewhat more relaxed.

3 Our Main Findings

The usage of proximity and phrases has been studied extensively for ad-hoc retrieval. Reports on their contribution are mixed, and it is generally accepted now that with a good basic ranking formula, the effectiveness of phrases is negligible or even negative [7], while recent evaluations of the use of automatically generated proximity terms suggest that term proximity may improve retrieval effectiveness especially at the top documents retrieved [9].

To determine the effectiveness of query operations for web retrieval we used the experimental setup of the web tracks at TREC 2003 and 2004 [3]. The corpus used for the experiments is the .GOV corpus, a crawl of a subset of the .gov domain performed in 2002, which contains 18.1Gb of data in 1.25M documents, the vast majority of which are HTML documents, and it preserves the link information between the documents. Our test set consists of the two topic distillation topic (and assessment) sets released with TREC 2003 and 2004, adding up to 125 queries.

The main research results obtained in the paper are the following:

- Even on top of a good basic ranking scheme for web retrieval, phrases and proximity terms may bring improvements in retrieval effectiveness.
- While we observed improvements both when documents are represented as a single field, and as aggregates of multiple fields, the latter setting gave more substantial improvements.
- Somewhat suprisingly, we found that phrases and proximity terms improve scores for traditional mean average precision as well as for high precision measures, although the former tended to be more substantial.
- Another important finding was that phrases and proximity terms have a strong positive impact on web retrieval effectiveness for extremely short queries (2 or 3 terms), while they have less, or even negative, effects on longer queries.

The full paper appeared as [6]; [5] repeats our findings, with similar scores, over additional corpora.

References

[1] F. Cacheda and A. Vina. Understanding how people use search engines: a statistical analysis for e-business. In *Proc. e-Business and e-Work Conf. and Exhibition*, pages 319–325, 2001.

[2] S. Chakrabarti. *Mining the Web: Analysis of Hypertext and Semi Structured Data.* Morgan Kaufmann, 2002.

[3] N. Craswell, D. Hawking, R. Wilkinson, and M. Wu. Overview of the TREC-2003 web track. In *Proc. of TREC 2003*, Gaithersburg, Maryland USA, November 2003.

[4] W. Hersh and R.T. Bhupatiraju. TREC GENOMICS Track Overview. In *Proc. TREC 2003*, pages 14–23, 2004.

[5] D. Metzler and W.B. Croft. A Markov random field model for term dependencies. In *Proc. SIGIR '05*, pages 472–479, 2005.

[6] G. Mishne and M. de Rijke. Boosting Web Retrieval through Query Operations. In *Advances in Information Retrieval. Proc. 27th European Conf. on IR Research*, pages 502–516, 2005.

[7] M. Mitra, C. Buckley, A. Singhal, and C. Cardie. An analysis of statistical and syntactic phrases. In *Proc. of RIAO-97*, 1997.

[8] V. Mittal, S. Baluja, and M. Sahami. Google tutorial on web information retrieval. In *RIAO-2004*, 2004.

[9] Y. Rasolofo and J. Savoy. Term proximity scoring for keyword-based retrieval systems. In *Proc. 25th European Conference on IR Research (ECIR 2003)*, pages 207–218, 2003.

[10] A. Spink, B.J. Jansen, D. Wolfram, and T. Saracevic. From e-sex to e-commerce: Web search changes. *Computer*, 35(3):107–109, 2002.

[11] A. Spink, D. Wolfram, B.J. Jansen, and T. Saracevic. Searching the web: the public and their queries. *J. American Society for Information Science and Technology*, 52(3):226–234, 2001.