Information Processing and Management 50 (2014) 342–360

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Detecting verbose queries and improving information retrieval

Emanuele Di Buccio *, Massimo Melucci, Federica Moro

Department of Information Engineering, University of Padua, Via Gradenigo 6/B, 35131 Padua, Italy

ARTICLE INFO

Article history:
Received 21 November 2012
Received in revised form 8 September 2013
Accepted 18 September 2013
Available online 28 October 2013

Keywords:
Information retrieval
Verbose queries
Query modification
Query expansion

ABSTRACT

Although most of the queries submitted to search engines are composed of a few keywords and have a length that ranges from three to six words, more than 15% of the total volume of the queries are verbose, introduce ambiguity and cause topic drifts. We consider verbosity a different property of queries from length since a verbose query is not necessarily long, it might be succinct and a short query might be verbose. This paper proposes a methodology to automatically detect verbose queries and conditionally modify queries. The methodology proposed in this paper exploits state-of-the-art classification algorithms, combines concepts from a large linguistic database and uses a topic gisting algorithm we designed for verbose query modification purposes. Our experimental results have been obtained using the TREC Robust track collection, thirty topics classified by difficulty degree, four queries per topic classified by verbosity and length, and human assessment of query verbosity. Our results suggest that the methodology for query modification conditioned to query verbosity detection and topic gisting is significantly effective and that query modification should be refined when topic difficulty and query verbosity are considered since these two properties interact and query verbosity is not straightforwardly related to query length.

1. Introduction

Most of the queries submitted to search engines are composed of a few keywords and have a length that ranges from three to six words as reported by Kumaran and Carvalho (2009); short queries became common due to the advent of the WWW search engines. According to Bogatin (2006), which was cited by Kumaran and Carvalho (2009), more than 15% of the total volume of the queries are written in natural language using more than four terms that might contain redundant and unnecessary keywords for the purposes of IR rather than selecting a small number of well-focused keywords; these keywords may even introduce ambiguity and cause topic drifts, thus reducing the effectiveness of retrieval. As well as the (small) fraction of long and/or verbose queries typed at web search engines, this will be an issue with voice querying, queries by example, and longer interactions with search engines – all of which are likely to increase. The detection and processing of these queries, which are called verbose queries, is the topic of this paper.

Whereas "verbose query" has the same meaning as "long query" in the relevant literature, we prefer to distinguish between "long" and "verbose" – while the former is basically measured by the number of keywords, the latter means that *the query is too long, detailed, uses or is expressed in more words than are needed*. This definition of "verbose" does not imply that a query must necessarily be long to be verbose – a short query may also be verbose and a long query may also be succinct.

The main objectives of the paper are to:

* Corresponding author. Tel.: +39 0498277929; fax: +39 0498277799.
E-mail address: dibuccio@dei.unipd.it (E. Di Buccio).

0306-4573/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ipm.2013.09.003

- assess the extent to which verbose queries affect the performance of a retrieval system;
- design and evaluate a methodology for processing verbose queries in order to improve system performance;
- provide a scheme for deciding which methods have to be applied for improving retrieval effectiveness depending on query length and verbosity.

Topic difficulty can be a cause of query verbosity and, in general, may affect query length. In this paper a topic is considered difficult when the IR system is unable capture one or more aspects of the topic, or relationship between aspects. When a user perceives a topic as difficult, he might try to either add words to his queries, thus inflating queries with useless or noisy words, or remove words, thus making queries very short without necessarily removing one or two unnecessary and noisy words. When considered in the context of query reformulation, the former can be the case of reformulations which involve specialization, while the latter can be the case of query generalization – see (Huang & Efthimiadis, 2009; Jansen, Spink, Blakely, & Koshman, 2007) for a taxonomy of web queries. Checking whether a query is difficult is a crucial step for query expansion. Indeed, the TREC Robust 2004 test collection was used in the experiments reported in this paper, since it contains more difficult topics than the other TIPSTER test collections.

As regards the study of query verbosity, research in IR exemplified by Bendersky and Croft (2008), Hoenkamp, Bruza, Song, and Huang (2009), and Park and Croft (2010) has considered long queries without assessing the actual information content and verbosity of queries. Because we have separated query verbosity from query length, we engaged a group of human assessors who assessed verbosity for each query on the basis of the definition of query verbosity introduced in this paper. The query verbosity assessment provided four sets of queries: short-and-succinct, long-and-succinct, short-and-verbose, long-and-verbose.

To automate the recognition of verbose queries, we implemented a predictor based on a C4.5 decision tree. The training data used to implement the predictor include a score based on a variant of the method of Lesk (1986) and features based on the semantic relationships (e.g. generalization, aggregation, inclusion, temporal collocation relationships) among terms contained in a linguistic database. These semantic relationships have been exploited to implement some algorithms for extracting the query term synonyms and for topic gisting. In particular, this approach to verbose query detection complements those reported by Bendersky and Croft (2008) and Park and Croft (2010) which are mainly focused on key concept or key term identification, which are then combined with the original query (which is actually the description of the TREC topic). Moreover, the work reported in this paper extends those works since our algorithms for verbose and succinct queries processing exploit both query modification by concept expansion and syntactic features.

This paper reports the experimental results about the impact of the above mentioned verbose query detection and processing algorithms on the overall retrieval effectiveness of an IR system. The results suggest that the proposed query verbosity detection and processing algorithms can increase retrieval effectiveness to a statistically significant extent. Finally, the paper summarizes the methodological and experimental results and provides a scheme that is a blueprint for future use of query expansion and reduction algorithms in verbose query detection and processing. We would like to emphasize on the methodological contribution of this paper; the experiment on the robust track is only as a thorough application that shows the feasibility of the proposed methodology.

The proposed methodology that combines an ad hoc incremental query expansion based on external resources with query term filtering based on part-of-speech tagging and query classification sounds natural since most of these techniques are used in automatic summarization by sentence extraction (Dang & Owczarzak, 2008) and the experiments reported in this paper give some evidence that it could work in IR.[1]

In this paper, we argue that queries can automatically be classified as verbose or not. The classifier used in the experiments reported in the paper can correctly detect verbose queries with small yet non-zero error. Moreover, the paper provides a blueprint for exploiting the information provided by the classifier to improve retrieval.

This paper is organized as follows: Section 2 provides a background and reports on the literature devoted to verbose query processing and detection; Section 3 describes the methodology we have proposed for verbose query processing and detection; Section 4 describes the experiments; Section 5 discusses the results and provides a scheme for query expansion and reduction algorithms in verbose query detection and processing.

2. Background

The main aim of this section is to discuss the issues related to verbose queries. In particular, we want to motivate the study of this type of query and highlight possible applications. After presenting a brief overview of the historical evolution of IR systems, paying attention to the "query evolution" (Croft, 2009), we will explore the application context of query verbosity. Furthermore, some of the approaches described and the strategies used for processing (and, in particular, for reducing) queries will be presented.

Over the years the structure of the query submitted to an IR system has significantly changed. In the 1970s, most of systems were based on Boolean logic and queries were characterized by a complex structure, the formulation of which required familiarity with the syntax and some expertise in retrieving information. For this reason, in general, only qualified personnel

[1] We thank the anonymous reviewer who pointed this out.