Information Processing and Management 52 (2016) 174–192

Contents lists available at ScienceDirect

Information Processing and Management

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

A linguistically driven framework for query expansion via grammatical constituent highlighting and role-based concept weighting

Bhawani Selvaretnam[a], Mohammed Belkhatir[b,*]

[a] Faculty of Computing and Informatics, Multimedia University, Malaysia
[b] Faculty of Computer Science, University of Lyon, France

ARTICLE INFO

Article history:
Received 4 May 2011
Revised 25 February 2015
Accepted 10 April 2015
Available online 28 November 2015

Keyword:
Query expansion
Information retrieval

ABSTRACT

In this paper, we propose a linguistically-motivated query expansion framework that recognizes and encodes significant query constituents characterizing query intent in order to improve retrieval performance. *Concepts-of-Interest* are recognized as the core concepts that represent the gist of the search goal whilst the remaining query constituents which serve to specify the search goal and complete the query structure are classified as descriptive, relational or structural. Acknowledging the need to form semantically-associated base pairs for the purpose of extracting related potential expansion concepts, an algorithm which capitalizes on syntactical dependencies to capture relationships between adjacent and non-adjacent query concepts is proposed. Lastly, a robust weighting scheme that duly emphasizes the importance of query constituents based on their linguistic role within the expanded query is presented. We demonstrate improvements in retrieval effectiveness in terms of increased mean average precision garnered by the proposed linguistic-based query expansion framework through experimentation on the TREC ad hoc test collections.

1. Introduction

Natural language is not always well-structured and may be semantically ambiguous, thus making it difficult to formulate the most appropriate and restrictive search query that is in line with the vocabulary of the documents being searched for. In recognition of this problem, several query expansion efforts have emerged over the years in an attempt to minimize query-document vocabulary mismatch. However, the query expansion process requires the comprehension of the intended search goal through the identification of key concepts prior to spawning and integrating additional terms to an initial query.

It is, however, inherent that query constituents possess two distinct functionalities which are disregarded in the aforementioned research efforts. Concepts either represent the content in accordance to the search goal or serve to connect query constituents. This is shown for example in the query "*coping with overcrowded prisons*" where the noun "*prisons*" provides the content whilst the verb "*coping*" and adjective "*overcrowded*" give specificity to the search goal. It is therefore necessary to acknowledge that query constituents take on multiple roles which if recognized and encoded appropriately would render a

* Corresponding author.
E-mail addresses: bhawani@mmu.edu.my (B. Selvaretnam), mohammed.belkhatir@univ-lyon1.fr (M. Belkhatir).

http://dx.doi.org/10.1016/j.ipm.2015.04.002
0306-4573/

more accurate understanding of the intended search goal. In particular, the recognition of formal non-compositional phrases (i.e. phrasal verbs, modals, fixed phrases, idioms, collocations, proper names and acronyms) which exist within queries is of particular importance.

Furthermore, the very nature of natural language dictates that there are intrinsic relationships between adjacent and non-adjacent concepts that reveal semantic notions pertaining to a search goal. However, earlier works fail to fully capitalize on these relationships between query terms which if considered appropriately would improve retrieval performance (Song, Taylor, Wen, Hon, & Yu, 2008). Term dependencies are applied in query expansion in Metzler and Croft (2007) where it is reported that retrieval performance through sequential modeling improves significantly in comparison to the unigram language model (full independence) and some improved effectiveness is seen in about 65 to 80% of the queries examined. We postulate that retrieval performance is possibly hindered by the fact that only relationships between adjacent concepts are considered. One might argue that both adjacent and non-adjacent dependencies can be captured through full dependence modeling. Contrarily, this will prove costly, in especially long queries, as multiple concept pairs will be derived, from which possibly a large number would not be very meaningful. Query expansion based on these pairs would generate unrelated concepts which in turn cause digression from the original search goal. Another imperative fact is also that the proposed dependence models are purely concerned with position and proximity of concepts. We hypothesize that both adjacent and non-adjacent association among query concepts can be effectively capitalized from syntactical dependencies within queries. This then translates into meaningful query concept pairing for expansion. We exemplify this conviction through the following discussion of the queries which consist of multiple concepts representing a single information need (e.g. "US President Barack Obama's inaugural address") and multiple concepts representing multiple information needs (e.g. "Barack Obama's policies and inaugural address"). Apart from recognizing that the terms "*Barack*" and "*Obama*" form a concept representing a specific person; it is intuitively evident that "*US*" and "*President*" refer to "*Barack Obama*" in the former, whilst "*policies*" and "*inaugural address*" are linked to "*Barack Obama*" in the latter.

Finally, a crucial element in query expansion is the process of weighting the original and expansion concepts to adequately reflect the search goal of a query. The choice of concept weighting schemes for query expansion models in existing approaches is very much dependent on the retrieval model utilized. The drawback of such models is that key concepts are established based on the statistical occurrence of a concept which is not necessarily reflective of the search goal of a query. Instead, we believe concepts should be given due emphasis based on the role they play within a query in representing the information need.

Our goal in this paper is to improve retrieval performance through query expansion by capitalizing on linguistic characteristics of queries. Queries are composed of multiple concepts of varying parts-of-speech and grammatical relations. These syntactical characteristics play distinct roles, whether it be representing the query content or providing links between the concepts, thus emphasizing semantic attributes of an expressed information need. To this end, we present a framework that consists of three major elements. We first propose a linguistically-motivated scheme for recognizing and encoding significant query constituents that characterize the intent of a query. We then extract potential expansion concepts through a proximity-based statistical query expansion method that capitalizes on grammatically linked base pairs of query concepts. Finally, we reconcile original and expansion concepts through a robust weighting scheme that is reflective of the role types of query constituents in representing an information need. One important factor is that, contrarily to state-of-the-art solutions, we are not dependent on the processed datasets to generate expansion concepts but instead consider a non-domain specific knowledge base (i.e. an n-gram corpus).

In Section 2, we cover the related works then discuss in Section 3 the details of the proposed framework, whilst in Section 4, we explain the experimental setup of the retrieval experiments, as well as the results obtained from the evaluation of our proposed scheme.

2. Related work

Depending on their familiarity with the search process, users construct queries which are both short and straight to the point or long-winded (Aula, 2003). These queries may take the form of grammatically correct sentences or merely a group of keywords associated to their search goals. This basic variance of query formats motivates the need to analyze the structure and varying lengths of queries in order to decipher the intended search goal. However, as far as query structure is concerned, even though linguistic characteristics such as parts-of-speech are utilized for the purpose of key concept identification (Cao, Nie, & Jing, 2005), they are not considered when extracting expansion concepts.

Voorhees (1994) highlights that long queries, if not well detailed, would also benefit from query expansion as much as short queries. This is similar to Lau and Goh (2006) who theorize that as query length increases, there is a higher probability that users would encounter unsuccessful searches. Di Buccio, Melucci, and Moro (2014) perform query expansion on all query terms upon distinguishing short and verbose queries with promising results. However, long queries may not always have done well with direct query expansion as such queries may consist of multiple important concepts. A verbose query may not always contain explicit information in the description itself to indicate which of these concepts is more important.