Chapter 70
Lexical Co-Occurrence and Contextual Window-Based Approach With Semantic Similarity for Query Expansion

Jagendra Singh
Jawaharlal Nehru University, India

Rakesh Kumar
Jawaharlal Nehru University, India

ABSTRACT

Query expansion (QE) is an efficient method for enhancing the efficiency of information retrieval system. In this work, we try to capture the limitations of pseudo-feedback based QE approach and propose a hybrid approach for enhancing the efficiency of feedback based QE by combining corpus-based, contextual based information of query terms, and semantic based knowledge of query terms. First of all, this paper explores the use of different corpus-based lexical co-occurrence approaches to select an optimal combination of query terms from a pool of terms obtained using pseudo-feedback based QE. Next, we explore semantic similarity approach based on word2vec for ranking the QE terms obtained from top pseudo-feedback documents. Further, we combine co-occurrence statistics, contextual window statistics, and semantic similarity based approaches together to select the best expansion terms for query reformulation. The experiments were performed on FIRE ad-hoc and TREC-3 benchmark datasets. The statistics of our proposed experimental results show significant improvement over baseline method.

INTRODUCTION

In this section, we present an overview of information retrieval, information retrieval system, and the need for query expansion. Further, it discusses appropriateness and drawbacks of term co-occurrence approaches for query expansion and the need for incorporating query terms context window and semantics in the field of automatic query expansion.

DOI: 10.4018/978-1-5225-5191-1.ch070

Information Retrieval

The discipline of information retrieval is almost as old as the computer itself. An old definition of information retrieval is the following by Mooers (1950):

Information retrieval is the name of the process or method whereby a prospective user of information is able to convert his need for information into an actual list of citations to documents in storage containing information useful to him.

An information retrieval system is a software program that is used to retrieve, store and manages needed information in a large collection. The system assists users to find the information need like the question answering system that returns the existence and location of documents instead of returning needed information or answer the question explicitly. Some system suggested documents may satisfy the user's information need. These kinds of documents are called *relevant* documents. A perfect retrieval system would retrieve only the relevant documents, not the irrelevant documents. However, there are no perfect retrieval systems because the searching statements are necessarily incomplete, and relevance of documents is the user's subjective opinion.

There are a large number of applications in which information retrieval is useful such as digital libraries, information filtering, recommender system, media search, search engines and many other and there is a constant need for improving such systems. In this context, information retrieval is an active field of research in computer science.

Query Expansion and Term Co-Occurrence Approach

Query Expansion and It's Need

The main objective of an information retrieval system is returning the maximum number of relevant documents for corresponding user query. However, there are many problems in developing an efficient IR system. The most critical problem for retrieval effectiveness is the term mismatch problem (Grossman & Frieder, 2004; Singh et al., 2015; Xu, 2000; Singh & Sharan, 2015a): the indexers and the users do often not use the same words for the same concept or idea. The term mismatch problem compounded by synonymy (the same word that has different meanings, for example - "java") and polysemy (the different words that has same or similar meanings, for example - "tv" and "television"). Synonymy words, together with its inflections form (such as plural forms, "television" versus "televisions"), may failure to retrieve relevant documents, that may decrease in recall (the ability of the system to retrieve all the relevant documents). Polysemy words may retrieve irrelevant documents that may decrease in precision (the ability of the system to retrieve only the relevant documents).

One of the most feasible and successful technique to handle the problem of term mismatch is to expand the original query with other words that describes the user intention or a query that is more likely to retrieve only the relevant documents. To consider the above problem, there is a need for automatic query expansion techniques that can assist the user in formulating the query. In last some years, it has been observed that the volume of data available online has dramatically increased while the number of query terms searched remained very less (Xu & Croft, 2000; Carpineto & Romano, 2012, Singh &