Chapter 22
Query Expansion Based on Central Tendency and PRF for Monolingual Retrieval

Rekha Vaidyanathan
MANIT Bhopal, India

Sujoy Das
MANIT Bhopal, India

Namita Srivastava
MANIT Bhopal, India

ABSTRACT

Query Expansion is the process of selecting relevant words that are closest in meaning and context to that of the keyword(s) of query. In this paper, a statistical method of automatically selecting contextually related words for expansion, after identifying a pattern in their score, is proposed. Words appearing in top 10 relevant document is given a score w.r.t partitions they appear in. Proposed statistical method, identifies a pattern of central tendency in the high scores and selects the right group of words for query expansion. The objective of the method is to keep the expanded query with minimum words (light), and still give statistically significant MAP values compared to the original query. Experimental results show 17-21% improvement of MAP over the original unexpanded query as baseline but achieves a performance similar to that of the state of the art query expansion models - Bo1 and KL. FIRE 2011 Adhoc English and Hindi data for 50 topics each were used for experiments with Terrier as the Retrieval Engine.

INTRODUCTION

Query Expansion is the process of selecting relevant words that are closest in meaning and context to that of the keyword in query. It overcomes the problem of word mismatch, where different words are used in query and documents to describe the same concept (Xu & Croft, 1996). Query Expansion is a successful technique in most of the cases but largely depends on the variation in retrieval performance of

DOI: 10.4018/978-1-5225-5191-1.ch022

queries (Amati, Carpineto & Romano, 2004). One of the most popular techniques is Pseudo Relevance Feedback, where the user submits a short query and from the initial set of retrieved results, an expanded query is reformulated. The expanded query contains terms from the initially retrieved documents that closely match with the query words: synonyms, plurals, modifiers etc. (Jones, Rey, Madani, & Greiner, 2006). Originally, query expansion is performed on the PRF information extracted from top N documents- selected from an initial search, on the same collection where the target documents are in (Evans & Lefferts 1994). Pseudo Relevance Feedback method is fully automatic compared to explicit feedback (Farah, 2009) as it does not require any user input, thus making it more attractive (Wu, Zhang, Zhou & Huang, 2010; Buckley, Salton, Allan & Singhal, 1995; Yu, Cai, Wen, & Ma, 2003). As this is a fully automated process this method can hurt or improve the query. Recent experiments show that with large query sets (with 50 queries) significant improvements are shown in overall performance (Mitra, Singhal & Buckley, 1998).

Automatically identifying expansion terms from the documents is a challenging task. A popular technique in IR is to give weights to the words and defining them in a vector space. It is used for Relevance Feedback and a classical model was proposed by Rocchio to find Text similarity and identifying relevant and non-relevant documents (Rocchio & Salton, 1971). Other methods for relevance feedback cite contextual and word similarity modeled as co-occurrence (Kilgarriff, Rychly, Smrz & Tugwell, 2004; Matsuo & Ishizuka, 2004), frequency estimates (Terra & Clarke, 2003) etc. Among the term weighting methods, T*erm frequency* and *inverse document frequency* is regarded as an empirical method with several possible variations (Aizawa, 2003). Studies related included frequency of words (Luhn, 1957), using inverse document frequency as term specifity (Spark-Jones, 1972) *tf.idf* and its variations (Salton & Buckley, 1988). More recent studies included, the *tf-rf scheme* for text categorization (Lan, Tan, Low & Sung, 2005), a local relevance scheme (Wu, Luk, Wong, & Kwok, 2008), and *tf.idf* weighting based on the length of the query (Paik, 2013), to cite a few. In this paper, a variation of the ***tf.idf*** is applied to derive a score for the words. It is assigned to the words in *initially* retrieved partitioned document instead of whole document, after applying Pseudo Relevance Feedback (PRF).

The measure *tf.idf* is extensively used in text retrieval due to its robustness (Robertson, 2004). In the basic formula of *tf.idf*, the measure of term specificity called the *inverse document frequency* (IDF, proposed by Karen Spärck Jones in 1972) is based on the number of *documents* containing the word. Thus ***idf*** is calculated for words across documents in a collection.

$$\boldsymbol{idf}\left(\boldsymbol{t},\boldsymbol{D}\right) = log\left[\left|D\right| / \left|\left\{d \in D : t \in d\right\}\right|\right] \tag{1}$$

Where, **t** = term in query; D = total number of documents in a collection;

$\mathbf{d} \in \mathbf{D} : \mathbf{t} \in \mathbf{d}$ = number of documents where term **t** occurred.

In this paper, we propose a technique that selects the most relevant words for query expansion automatically, from top *k* feedback documents, retrieved using PRF, in the following manner:

1. Identify good candidate words related to query words based on sections or partitions of a document by assigning scores for the words, using a variation of the tf.idf formula called tf.ipf.