Information Processing and Management 51 (2015) 179–193

Contents lists available at ScienceDirect

Information Processing and Management

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

Weighted Word Pairs for query expansion ☆

Francesco Colace [a], Massimo De Santo [a], Luca Greco [a,*], Paolo Napoletano [b]

[a] DIEM, University of Salerno, Fisciano, Italy
[b] DISCo, University of Milano-Bicocca, Italy

ARTICLE INFO

Article history:
Received 25 June 2013
Received in revised form 4 July 2014
Accepted 11 July 2014
Available online 7 August 2014

Keywords:
Text retrieval
Query expansion
Explicit relevance feedback
Pseudo-relevance feedback
Probabilistic Topic Model

ABSTRACT

This paper proposes a novel query expansion method to improve accuracy of text retrieval systems. Our method makes use of a minimal relevance feedback to expand the initial query with a structured representation composed of weighted pairs of words. Such a structure is obtained from the relevance feedback through a method for pairs of words selection based on the Probabilistic Topic Model. We compared our method with other baseline query expansion schemes and methods. Evaluations performed on TREC-8 demonstrated the effectiveness of the proposed method with respect to the baseline.

1. Introduction

Most retrieval systems show relative weaknesses in retrieving relevant documents, especially when few keywords are used to model user information needs. Information retrieval models, that have been proposed through the years, often rely on the *bag of words* model for document and query representation and can be grouped into three main categories: set-theoretic (including boolean), algebraic and probabilistic models (Christopher, Manning, & Schtze, 2008; Baeza-Yates & Ribeiro-Neto, 1999). It is well known that the "bag of words" model assumes both documents and queries representable as feature vectors. The elements of such vectors can indicate the presence (or absence) of a word or take into account its occurrence frequency, but the information about the position of that word within the document is completely lost (Christopher et al., 2008); then, the elements of the vector are simply weights computed in different ways. In this context, the relevance of a document to a query can be measured as the distance between the corresponding vector representations in the space of features.

It has been found that common users are used to perform short queries, 2 or 3 words on average (Jansen, Spink, & Saracevic, 2000; Jansen, Booth, & Spink, 2008). Unfortunately, the shortness of a query can cause common information retrieval systems failures due to the inherent ambiguity of language (polysemy, etc.). Since most text retrieval systems relying on a term-frequency based index generally suffer from low precision (or low quality document retrieval), a typical solution adopted to reduce this query/document mismatch is expanding the initial query using words or phrases with a similar meaning or some other statistical relation to the set of relevant documents (Carpineto, de Mori, Romano, & Bigi, 2001): this strategy is often referred as *query expansion*.

☆ The authors contributed equally to this work.

* Corresponding author.
E-mail address: lgreco@unisa.it (L. Greco).

http://dx.doi.org/10.1016/j.ipm.2014.07.004
0306-4573/

In this work we propose a query expansion method that automatically extracts a set of Weighted Word Pairs from a set of topic-related documents provided by the relevance feedback. Such a structured set of terms is obtained by using a method of *term extraction* previously investigated in Colace, De Santo, Greco, and Napoletano (2013, 2014), Clarizia, Greco, and Napoletano (2011) and based on the *Latent Dirichlet Allocation* model (Blei, Ng, & Jordan, 2003) implemented as the *Probabilistic Topic Model* (Griffiths, Steyvers, & Tenenbaum, 2007).

Evaluation has been conducted on TREC-8 repository. We compared the proposed Weighted Word Pairs (WWP) with a method for term extraction based on the Kullback Leibler divergency (Carpineto et al., 2001). Our approach achieves overall better performances and demonstrates that a structured feature representation has a greater discriminating power than a feature vector made of weighted words.

2. Problem formulation

According to the Information Retrieval (IR) theory, the representation of queries and documents is based on the *Vector Space Model* (Christopher et al., 2008): a document or query is a vector of weighted words belonging to a vocabulary $\mathcal{T}$:

$$\mathbf{d} = \{w_1, \ldots, w_{|\mathcal{T}|}\}.$$

Each weight w_n is such that $0 \leqslant w_n \leqslant 1$ and represents how much the term t_n contributes to the semantics of the document $\mathbf{d}$ (in the same way for $\mathbf{q}$). In the *term frequency-inverse document frequency* (*tf-idf*) model, the weight is typically proportional to the term frequency and inversely proportional to the frequency and length of the documents containing the term.

Given a query, the IR system assigns the relevance to each document of the collection with respect to the query, by using a similarity function as defined in the following:

$$sim(\mathbf{q}, \mathbf{d}) = \sum_{t \in \mathbf{q} \cap \mathbf{d}} w_{t,\mathbf{q}} \cdot w_{t,\mathbf{d}}, \tag{1}$$

where $w_{t,\mathbf{q}}$ and $w_{t,\mathbf{d}}$ are the weights of the term t in the query $\mathbf{q}$ and document $\mathbf{d}$ respectively.

2.1. Query expansion by relevance feedback

Performance of IR systems can be improved by expanding the initial query with other topics-related terms. These query expansion terms can be manually typed or extracted from feedback documents selected by the user himself (*explicit relevance feedback*) or automatically chosen by the system (*pseudo-relevance feedback*) (Baeza-Yates & Ribeiro-Neto, 1999).

A general query expansion framework is a modular system including one or several instances, properly chained, of the following modules: Information Retrieval (IR), Feedback (F), Feature Extraction (FE), Query Reformulation (QR).

A general scheme is represented in Fig. 1 and can be explained as follows. Let us consider a generic IR system and a collection of indexed documents $\mathcal{D}$. The user performs a search in the IR system by typing a query $\mathbf{q}$. The IR system computes the *relevance* of each document of the corpus with respect to the query through the Eq. (1). As a result of the search, a set of ranked documents $\Omega_{res} = \{\mathbf{d}_1, \ldots, \mathbf{d}_N\} \subseteq \mathcal{D}$ is returned to the user.

Once the result is available, the module F assigns a judgement of relevance, also known as *relevance feedback*, to each document of Ω_{res}. The relevance can be manually or automatically (pseudo-relevance) assigned. In case of manual, the user provides the *explicit feedback* by assigning a positive judgment of relevance to a subset of documents $\Omega_{fback} = \{\mathbf{d}_1, \ldots, \mathbf{d}_M\} \subseteq \Omega_{res}$. In case of automatic feedback, the module F arbitrarily assigns a positive judgment of relevance to a subset of documents, usually the top M documents retrieved from Ω_{res}.

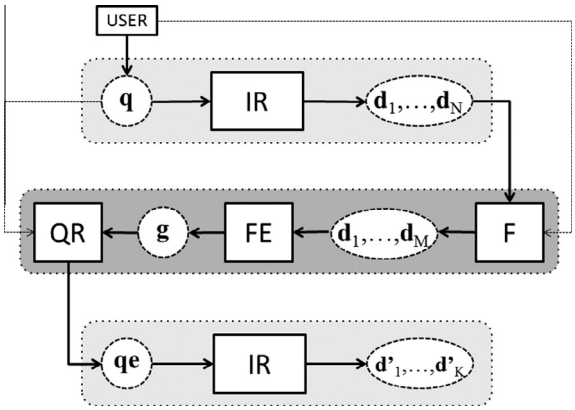

Fig. 1. General framework for query expansion.