Information Processing and Management 49 (2013) 558–575

Contents lists available at SciVerse ScienceDirect

Information Processing and Management

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

Probabilistic co-relevance for query-sensitive similarity measurement in information retrieval

Seung-Hoon Na

Natural Language Processing Research Term, Electronics and Telecommunications Research Institute, South Korea

ARTICLE INFO

Article history:
Received 23 February 2012
Received in revised form 28 September 2012
Accepted 16 October 2012
Available online 24 November 2012

Keywords:
Probabilistic co-relevance
Query-sensitive similarity
Inter-document similarity
Cluster hypothesis
Cluster-based retrieval

ABSTRACT

Interdocument similarities are the fundamental information source required in cluster-based retrieval, which is an advanced retrieval approach that significantly improves performance during information retrieval (IR). An effective similarity metric is *query-sensitive similarity*, which was introduced by Tombros and Rijsbergen as method to more directly satisfy the cluster hypothesis that forms the basis of cluster-based retrieval. Although this method is reported to be effective, existing applications of query-specific similarity are still limited to vector space models wherein there is no connection to probabilistic approaches. We suggest a probabilistic framework that defines query-sensitive similarity based on *probabilistic co-relevance*, where the similarity between two documents is proportional to the probability that they are both *co-relevant* to a specific given query. We further simplify the proposed co-relevance-based similarity by decomposing it into two separate relevance models. We then formulate all the requisite components for the proposed similarity metric in terms of scoring functions used by language modeling methods. Experimental results obtained using standard TREC test collections consistently showed that the proposed query-sensitive similarity measure performs better than term-based similarity and existing query-sensitive similarity in the context of Voorhees' nearest neighbor test (NNT).

1. Introduction

Interdocument similarity is a major factor affecting the enhancement of cluster-based retrieval because it is one of the most crucial factors according to the cluster hypothesis. Typically, *term-based similarity* has been widely used as a similarity metric for inter-document similarity where a document is represented as terms and the similarities between documents then straight-forwardly derived by applying matching functions or retrieval models. They include the Dice coefficient and Jaccard's coefficient, cosine similarity in the vector-space model (Salton, Wong, & Yang, 1975), the probabilistic retrieval model (Robertson & Jones, 1976; Robertson & Walker, 1994), and the KL-divergence between query model and document model in language modeling approaches (Hiemstra, 1998; Lafferty & Zhai, 2001; Ponte & Croft, 1998).

However, it is not known whether term-based similarity is the most suitable metric in terms of the cluster hypothesis. This is because the cluster hypothesis declares the expected properties of *similar* or *closely associated* documents, but does not specify details of the type of interdocument similarities. Therefore, we cannot assume that interdocument similarities necessarily take the form of term-based similarities. Instead of using existing term-based similarity, a more straight-forward way of better similarity metric is to apply the inverse of the cluster hypothesis to produce a similarity metric that might better satisfy the hypothesis.

E-mail address: nash@etri.re.kr

0306-4573/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ipm.2012.10.002

The inspiration for this new method was provided by Tombros and Rijsbergen (2001, 2004), who argued that a form of similarity that better fits the cluster hypothesis should take into account the query context. Based on this argument, they suggested the use of *query-sensitive similarity*, which imparts a query-specific bias on any interdocument similarities found between two documents. This similarity metric makes one pair of documents more similar than others, when both are more similar according to a given query. The similarity obtained is a dynamic quantity that needs to be computed differently for specific goals, in specific retrieval situations, or for each specific query. Evidence for the use of this dynamic similarity metric is found in studies on *query-specific clustering* (Hearst & Pedersen, 1996; Willett, 1985) and *structural re-ranking* on the basis of top-retrieved documents for each query (Kurland & Lee, 2005).

In this pioneering work of Tombros and Rijsbergen (2001, 2004), query-sensitive similarity was only explored in the setting of a vector-space model based on cosine similarity. However, most modern retrieval methods are based on probabilistic frameworks such as the probabilistic retrieval model and language modeling. Therefore, it would be an interesting challenge to develop a query-sensitive similarity method in a probabilistic framework without losing the original insight, and to connect the derived similarity with the cluster hypothesis in a formal manner.

By developing query-specific similarity in a probabilistic context, this study proposes the use of *probabilistic co-relevance* to define similarity more directly and to satisfy the cluster hypothesis. Our main hypothesis is postulated in the co-relevance principle for similarity metric which is stated as follows: *the similarity between two documents should be proportional to the probability in a query context that they are co-relevant to a 'given query' which we call the 'co-relevance probability'*. We consider two different cases; (i) the relevance of a document is independent of the relevance of other documents and (ii) the relevance of a document is dependent on others. We then integrate the resulting estimations from both cases and decompose the co-relevance probability into two *relevance probabilities*. We adopt the same assumptions made in previous works (Lafferty & Zhai, 2003; Roelleke & Wang, 2006), where each relevance probability is further simplified into ranking formulae of retrieval model. Finally, the co-relevance probability is easily and tractably estimated on the basis of simple formulae that rely on top-retrieved documents without resorting to the actual relevant documents. Specifically, we applied language modeling methods for estimating co-relevance probability and obtained query-sensitive similarities which are similar to the interpolation style of RM3, a variant of relevance model (Abdul-jaleel et al., 2004; Lavrenko & Croft, 2001), which is a widely used pseudo-relevance feedback method employed by language modeling approaches.

The results of experiments carried out with standard TREC collections consistently show that the proposed query-specific similarity significantly outperforms the state-of-the-art method developed by Tombros and Rijsbergen (2001) in the setting of Voorhees' nearest neighbor test (NNT) (Voorhees, 1985), thus supporting our claim that co-relevance-based similarity is an improvement over existing metrics.

This paper is organized as follows. Section 2 briefly reviews related works. Section 3 presents our proposed query-sensitive similarity method based on probabilistic co-relevance. Section 4 presents experimental results, and Section 5 presents our conclusions and future work.

2. Related work

Cluster-based retrieval is an approach that generates clusters from collections in order to enhance retrieval performance. This method was motivated by the *cluster hypothesis*, which states that "*closely associated documents tend to be relevant to the same requests*" (Jardine & Rijsbergen, 1971; Rijsbergen, 1979). Thus far, numerous studies have been conducted, for instance, initial trials based on hierarchical clustering that employed different types of merging criteria, i.e., single linkage, complete linkage, group average, and Ward's method (Croft, 1980; El-Hamdouchi & Willett, 1986; Griffiths, Robinson, & Willett, 1984; Jardine & Rijsbergen, 1971; Rijsbergen & Croft, 1975; Voorhees, 1985). There are also more recent language modeling approaches based on partitional clustering (Liu & Croft, 2004; Na, Kang, Roh, & Lee, 2007) and document expansion using nearest neighbors as a cluster (Kurland & Lee, 2004; Tao, Wang, Mei, & Zhai, 2006). The early studies on cluster-based retrieval have delivered inconclusive results; however, recent works based on language modeling methods indicate a significant improvement over the baseline retrieval method (Liu & Croft, 2004; Kurland & Lee, 2004).

Initially, cluster-based retrieval studies were focused on using static clusters from an entire collection as the cluster type. In contrast to these studies, Willett (1985) developed *query-specific clusters*, which were clusters formed from top-retrieved documents in a collection rather than whole documents. However, Willett (1985)'s experiments reported only a limited performance for query-specific clusters when compared with static clusters, possibly due to his limitation in search method as pointed out in Tombros, Villa, and Rijsbergen (2002). Hearst and Pedersen (1996) further motivated the use of query-specific clusters by revising the underlying assumption of the cluster hypothesis and stating that the co-relevance event of two documents should not be fixed statically; it should be dependent on a specific query instead. Hearst and Pedersen (1996) examined the use of query-specific clusters in the setting of an enhanced user interface where a user could choose the best relevant cluster from those recommended by a system. They showed that the retrieval effectiveness can be improved substantially with the interface.

Tombros et al. (2002) further investigated query-specific clusters, but with hierarchical clusters based on four widely used merging methods: single linkage, complete linkage, group average, and Ward's method. They concluded that query-specific clusters showed much better potential effectiveness compared to static clusters.