Available online at www.sciencedirect.com
SCIENCE DIRECT®
Information Processing and Management 42 (2006) 1151–1162
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

Does pseudo-relevance feedback improve distributed information retrieval systems?

Fernando Martínez-Santiago, Miguel A. García-Cumbreras *, L. Alfonso Ureña-Lòpez

Department of Computer Science, University of Jaén, Jaén, Spain

Received 3 October 2005; received in revised form 4 January 2006; accepted 4 January 2006
Available online 23 February 2006

Abstract

This paper presents a thorough analysis of the capabilities of the pseudo-relevance feedback (PRF) technique applied to distributed information retrieval (DIR). Previous studies have researched the application of PRF to improve the selection process of the best set of collections from a ranked list. This work emphasizes the effectiveness of PRF applied to the collection fusion problem. Usually, DIR systems apply PRF in the same way as traditional Information Retrieval systems. For each collection, local results are improved through PRF. A first question which arises is whether this local improvement is preserved in the final result. In addition, DIR systems merge the documents of rankings that are returned from a set of collections. Since a new global list of documents is available, we could use that list to apply PRF again, but on global level rather than on a local level. In order to apply global PRF, we have developed a merging approach called two-step RSV. Finally, we describe a number of experiments involving the two levels, local and global, of application of the PRF techniques.

Keywords: DIR; Collection fusion; TREC; CORI; Pseudo-relevance feedback

1. Introduction

Typically, a distributed information retrieval (DIR) system must rank document collections for query relevance. It selects the best set of collections from a ranked list, and merges the document rankings, returned from a set of collections. This last issue is the problem called *collection fusion problem* (Voorhees, Gupta, & Johnson-Laird, 1995). The aim of this paper is a thorough analysis of the pseudo-relevance feedback (PRF), also named blind feedback, applied to the collection fusion problem.

* Corresponding author.
E-mail addresses: dofer@ujaen.es (F. Martínez-Santiago), magc@ujaen.es (M.A. García-Cumbreras), laurena@ujaen.es (L.A. Ureña-Lòpez).

0306-4573/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2006.01.003

Relevance feedback (Rocchio, 1971) is an appreciated process that improves the performance of the information retrieval. The goal of relevance feedback is to retrieve and rank those documents highly, which are similar to the documents that are found relevant by the user. On the other hand, (PRF) (Salton & Buckley, 1990) technique does not need user interaction but it makes the assumption that the top N retrieved documents are relevant. Relevance feedback has been applied to non-distributed scenarios as well as DIR systems, but it needs user collaboration to decide what documents are relevant.

Usually DIR systems apply PRF in the same way that traditional information retrieval system: PRF or query expansion is applied in a local environment by each individual IR System. In this paper, we use *local feedback* to refer to this way of application of PRF. In addition our proposal is to apply PRF globally over the final top ranking merged document list. In this paper it is said that this kind of PRF is *global PRF*. Previous works are focused on local PRF as a way of improving the collection selection process, obtaining poor results (Ogilve & Callan, 2001). We explore global PRF as a way of improving the document merging process, not the collections selection process.

The documents returned for each IR engine are merged by using an algorithm called two-step RSV (Martínez-Santiago, Martín, & Ureña, 2003a, 2005). This algorithm works well in CLIR systems based on query translation, but the application of two-step RSV at DIR environments requires an additional effort: learning of collection issues such as document frequency, collection size and so on. On the other hand, since two-step RSV makes up a new global index based on query terms and the whole of retrieved documents, it is possible the application blind feedback at global level, by means of the DIR monitor, better than at local one, by means of each individual IR engine. Note that two-step RSV algorithm does not implement the whole of the DIR problems. Two-step RSV deals only with the document merging problem. Thus, ranking and selection of collections are realized by using the known CORI algorithm, described in the next section (see Table 1).

1.1. The Collection Retrieval Inference Network (CORI)

The Collection Retrieval Inference Network (CORI) model (Callan, Lu, & Croft, 1995) is a well-known algorithm used in DIR, and it addresses three problems. Briefly, CORI is inspired by the TF * IDF document ranking method (TF is the term frequency and IDF is the inverse document frequency) as an analogy for collection ranking. Each collection is depicted as a virtual document. The whole of those virtual documents build up a virtual collection. Given an user query, collections are selected in the same way that traditional IR systems select documents. Thus, CORI uses the TF * IDF formula by replacing TF with DF (document frequency) and IDF with ICF (inverse collection frequency). The required resource description is built up by document frequency, size of the collection and so on. (Callan, Connell, & Du, 1999) and (Callan, French, Powell, & Connell, 2000) propose a sampling technique (query-based sampling) in order to learn the description of the collections by interacting with every database by sending queries and analyzing the outcomes.

In a recent paper (Si & Callan, 2003c) studied the limitations of CORI when collection size varies. They found that CORI rarely ranks large collections highly, even though the collections are often the best source of relevant documents. They propose to modify CORI based on estimated database size to compensate for this effect but they do not address the difficult issue of parameter choice.

Table 1
DIR systems and algorithms implemented for the experiments

	CORI	Two-step RSV
Ranking collections	CORI	CORI
Selecting collections	CORI	CORI
Merging documents	CORI	Two-step RSV
Local PRF available	Yes	Yes
Global feedback available	No	Yes