Information Sciences 273 (2014) 171–181

Contents lists available at ScienceDirect

Information Sciences

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

Score distributions for Pseudo Relevance Feedback

Javier Parapar *, Manuel A. Presedo-Quindimil, Álvaro Barreiro

Information Retrieval Lab, Department of Computer Science, University of A Coruña, Campus de Elviña, 15071 A Coruña, Spain

ARTICLE INFO

Article history:
Received 24 July 2013
Received in revised form 18 February 2014
Accepted 10 March 2014
Available online 19 March 2014

Keywords:
Information Retrieval
Pseudo Relevance Feedback
Score distribution
Pseudo Relevance Feedback set
Relevance Model

ABSTRACT

Relevance-Based Language Models, commonly known as Relevance Models, are successful approaches to explicitly introduce the concept of relevance in the statistical language modelling framework of Information Retrieval. These models achieve state-of-the-art retrieval performance in the Pseudo Relevance Feedback task. It is known that one of the factors that more affect to the Pseudo Relevance Feedback robustness is the selection for some queries of harmful expansion terms. In order to minimise this effect in these methods a crucial point is to reduce the number of non-relevant documents in the pseudo relevant set. In this paper, we propose an original approach to tackle this problem. We try to automatically determine for each query how many documents we should select as pseudo-relevant set. For achieving this objective we will study the score distributions of the initial retrieval and trying to discern in base of their distribution between relevant and non-relevant documents. Evaluation of our proposal showed important improvements in terms of robustness.

1. Introduction and motivation

In the history of the Information Retrieval research, efforts to improve retrieval effectiveness have been centred in both developing better retrieval models by including new features or using different theoretical frameworks; and in designing new techniques to be incorporated on top of existing models to improve their performance. Particularly on the later, Query Expansion (QE) has proven to be effective from very early research stages. QE approaches can be classified between global techniques which produce a query rewriting without considering the original rank produced by the query, and local techniques in which the expanded query is generated using the information of the initial retrieval list.

In [33] Salton presented the initial efforts on exploiting the local information to improve the query formulation introducing, among others, Rocchio approach [29] working on the Vector Space Model framework. This family of local techniques is called Relevance Feedback (RF) [30] and it is based on using the relevant documents in the initial retrieval set in order to reformulate the query based on their content. Nevertheless, in a real retrieval scenario it is not realistic to assume that relevance judgements are available. Because of this, Pseudo Relevance Feedback (PRF) algorithms have been investigated [9,39]. PRF methods are based on assuming relevance of a set of documents retrieved by the original query. The set of documents which are assumed to be relevant and the way in which their information is exploited to improve the original query varies from one PRF method to another.

One crucial aspect of the Pseudo-Relevance Feedback methods is robustness. In this context, robustness is defined as the quality of not hurting the effectiveness values achieved by the retrieval model in the initial rank for every query. Most of

* Corresponding author. Tel.: +34 981 167 000x1276; fax: +34 981 167 160.
E-mail addresses: javierparapar@udc.es (J. Parapar), mpresedo@udc.es (M.A. Presedo-Quindimil), barreiro@udc.es (Á. Barreiro).

http://dx.doi.org/10.1016/j.ins.2014.03.034
0020-0255/

existing Pseudo-Relevance Feedback methods outperform the effectiveness of the initial retrieval in average but they tend to harm some of the queries. This is an important point for solving in order to popularise the use of these methods in the commercial search engines. The most common phenomenon causing the decrease of effectiveness for a query is the *topic drift*. Topic drift refers to the situation where the expansion of the query produced that the topic of the original user need has moved (drifted) away to a different one. For instance, for the TREC topic 101: *Design of the "Star Wars" Anti-missile Defense System*, a very clear example of topic drift would be the returning of documents about the film. The topic drift can be naturally produced by the addition of terms, but this problem can be greatly intensified when the pseudo-relevant set (RS) has plenty of irrelevant documents.

This problem has been exposed very early in the literature [24] and caused lots of works on areas such as query performance prediction [11,8] which investigates how to predict the performance of a query anticipating those queries that will be negatively affected by the expansion, selective Pseudo-Relevance Feedback [32,2] which tries to decide for which queries PRF should or not be applied, and adaptive Pseudo-Relevance Feedback [21] that is centred on adjust the weight of the expansion terms over the original query automatically depending on the nature of the given query.

The different approaches to decide when of how much apply PRF have considered pre-query processing indicators and initial ranking examination. Several evidences have been considered such as the number of query terms in the pseudo-relevant documents, the similarity between query and the relevant set, and term proximity measures. But it was only recently when some works started to consider the scores of the initial retrieval [34]. Shtok et al. argue that query-drift can potentially be estimated by measuring the diversity (e.g., standard deviation) of the retrieval scores of the documents in the ranking.

In this paper we also exploit the scoring information but in a different way, we use the scores of the initial retrieval for determining the pseudo-relevant set itself, trying to minimise the amount of non-relevant documents in it. For achieving this objective we used a framework for modelling the score distributions of a retrieval model [23] and adapt the threshold optimisation solution for recall-oriented retrieval [4] for our particular problem, where we want to stop selecting documents from the top of the initial retrieval when non-relevant documents appear. Score distributions research investigates the idea of using the documents' scores for separating relevant and non-relevant documents. For doing this, different statistical modelling choices over both groups of documents are taken and the parameters of the statistical distributions are inferred from the observed scores. Although it has been already used for other task such as meta-search and high recall oriented task such as legal retrieval, this is a novel and especially adequate use of the score distributions analysis. We are really pursuing a high precision for our task in such a way that ideally if no relevant documents are present on the top of the initial retrieval we want to return an empty RS producing a way of selective PRF. Furthermore, and not less important, our approach reduces the number of parameters to tune in the training phase of PRF methods by suppressing the necessity of tune r, the number of documents on the RS.

For assessing our proposal we will use one of the most successful PRF methods in the state-of-the art: Relevance-Based Language Models (RM) [18]. In particular, we will use the best performing estimation for RM the so called RM3 estimation [1]. Although, when averaged over a query set the differences in performance in terms of Average Precision when selecting different top sizes for RS in a particular collection may not differ too much for RM3, it varies a lot at query level (see Fig. 1). Meanwhile some queries present a stable behaviour (as query 81), most of them have either an increasing behaviour (as que-

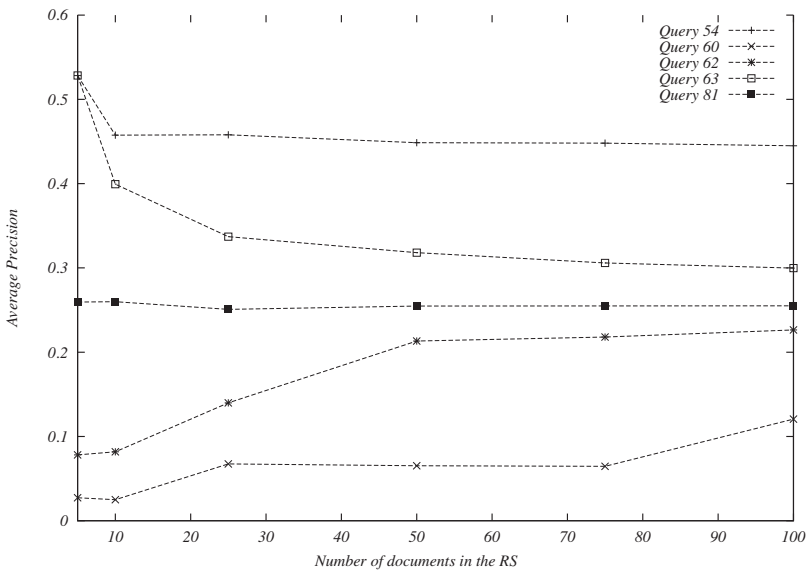

Fig. 1. RM3 behaviour in terms of Average Precision for different queries from the training query set of the AP88–89 collection with $t = 100$ and $\lambda = 0.8$ and $\mu = 1000$.