Information Fusion 39 (2018) 56–71

Contents lists available at ScienceDirect

Information Fusion

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

A rank fusion approach based on score distributions for prioritizing relevance assessments in information retrieval evaluation

David E. Losada[a,*], Javier Parapar[b], Alvaro Barreiro[b]

[a] Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS), Universidade de Santiago de Compostela, Spain
[b] Information Retrieval Lab, Department of Computer Science, University of A Coruña, Spain

ARTICLE INFO

Article history:
Received 22 November 2016
Revised 1 March 2017
Accepted 2 April 2017
Available online 3 April 2017

Keywords:
Rank fusion
Information retrieval
Evaluation
Pooling
Score distributions
Pseudo-relevance

ABSTRACT

In this paper we study how to prioritize relevance assessments in the process of creating an Information Retrieval test collection. A test collection consists of a set of queries, a document collection, and a set of relevance assessments. For each query, only a sample of documents from the collection can be manually assessed for relevance. Multiple retrieval strategies are typically used to obtain such sample of documents. And rank fusion plays a fundamental role in creating the sample by combining multiple search results. We propose effective rank fusion models that are adapted to the characteristics of this evaluation task. Our models are based on the distribution of retrieval scores supplied by the search systems and our experiments show that this formal approach leads to natural and competitive solutions when compared to state of the art methods. We also demonstrate the benefits of including pseudo-relevance evidence into the estimation of the score distribution models.

1. Introduction

Evaluation is crucial to making progress in science. In data intensive disciplines, creating the *gold standard* is often a major bottleneck in the process of building a test collection or benchmark for evaluation. Gold standards –also known as ground truth– are typically produced by humans and, therefore, they are expensive. A case in point is the creation of Information Retrieval (IR) test collections for evaluating search algorithms. Given a set of test queries, representing different information needs, and a large collection of documents, we would like to have exhaustive judgments (relevance information on every query-document pair). But this is unfeasible with current large-scale collections. For each query, we can only afford to judge a selected sample of documents from the collection. It is standard practice to run each query against multiple search engines, fuse the rankings supplied, and manually assess for relevance the documents at the top of the list. Focusing the judgment effort on these selected documents makes the most of human assessors' time and leads to pools of judged documents that can be reliably used to compare retrieval strategies [41].

With multiple search systems contributing to the evaluation task, rank fusion becomes an essential component. An effective rank fusion strategy leads to high counts of relevant documents in the judged set of documents and, consequently, to a robust benchmark. However, most fusion methods employed to date for ranking to-be-judged documents are rather simplistic. We claim that the process of prioritization of to-be-judged documents should take into account all available evidence. In particular, most search systems participating into a given evaluation initiative supply scores that measure the degree of relevance between documents and test queries. Modeling the distributions of these scores has found to be effective for a broad range of tasks [3] but it was never used for rank fusion in evaluating IR systems. We show here that fusion models based on Score Distributions (SDs) lead to highly competitive methods for allocating documents for judgment.

We define and experiment with two main classes of prioritization strategies: i) *static methods*, which build a merged ranking of documents that remains unchanged during the whole assessment process, and ii) *dynamic methods*, where the priority of documents changes as we obtain relevance assessments. Dynamic methods fit well with SD models because we can update the estimations of the distributions of relevant and non-relevant documents after each assessment. This iterative update of distributions is a natural consequence of employing SD models for ranking to-be-judged documents. Furthermore, we propose an innovative way to estimate the initial score distributions of relevant and non-relevant documents for each search system. In the absence of relevance information, it is customary to build these two distributions using only the list of scores supplied by each system. But we argue that pseudo-

* Corresponding author.
E-mail addresses: david.losada@usc.es (D.E. Losada), javierparapar@udc.es (J. Parapar), barreiro@udc.es (A. Barreiro).

http://dx.doi.org/10.1016/j.inffus.2017.04.001
1566-2535/

relevance information can be inferred from the list of available rankings, and such pseudo-relevance evidence can be effectively used for initializing the SD models. Initializing SD models of metasearch in this way is novel, and our experiments demonstrate that incorporating pseudo-relevance information is beneficial.

This paper addresses the following research questions:

- Are rank fusion models based on SDs effective for prioritizing assessments in the context of search system evaluation? How do they compare with state-of-the-art static and dynamic prioritization strategies?
- How can we use pseudo-relevance information to estimate the initial score distributions of relevant and non-relevant documents? Does the incorporation of pseudo-relevance information into SD models lead to improved models?

The rest of the paper is organized as follows. Section 2 presents models of score distributions that have been employed in different Information Access tasks, and Section 3 explains our proposal to use SD models for rank fusion in pooling-based evaluation of IR systems. The experiments are reported in Section 4 and Section 5 offers some discussing remarks. Section 6 reviews some studies that are related to our research. The paper ends with some conclusions and future lines of work.

2. Modeling score distributions of search systems

Given a user query most retrieval systems calculate a score per document that measures the degree of relevance to the query. These scores are employed for ranking retrieved documents, and their range and distribution varies across different systems. Score distributions have been effectively modeled in multiple Information Access areas, such as Information Filtering or Distributed Information Retrieval. For instance, Manmatha and colleagues [31] exploited score distributions for combining the outputs of different search engines (meta-search problem). Arampatzis and his colleagues [1,2] formulated the threshold optimization problem and worked with score distributions models for locating a good cut-off point in a legal search task. Other researchers have applied score distributions to tasks such as query performance prediction [17], image retrieval [4] and pseudo-relevance feedback [36].

Under binary relevance, the score distributions on a per query basis may be fitted as a mixture of two distributions: one for relevant documents and another one for non-relevant documents. This mixture model is used to map the scores to probabilities. This formal modeling process is essential in Information Fusion, Metasearch, Filtering or Thresholding. SD models have been shown to work for a large number of retrieval systems, particularly for those contributing to well-known IR evaluation campaigns like TREC [49].

A number of modeling alternatives have been explored in the literature. Various combinations of distributions have been employed, but we will focus on a combination of two Log-Normal distributions,[1] which is a general and consistent approach for preserving relevance information across a variety of search systems [16]. This mixture follows the recall-fallout hypothesis [39] and offers better goodness of fit than other alternatives [16,18]. A full comparison of different models was performed by Cummins [16]. He studied different combinations of distributions and concluded that a mixture of two Log-Normal distributions is the best performing model. His study considered the Normal distribution, the Log-Normal distribution and the Gamma distribution for modeling the scores of relevant documents; and the Exponential distribution, the Normal distribution, the Log-Normal distribution and the Gamma distribution for modeling the scores of non-relevant documents.

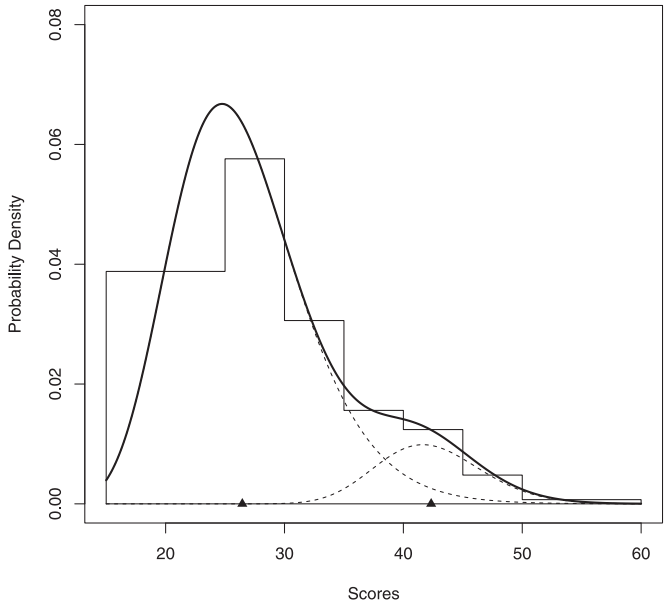

Fig. 1. A mixture of two Log-Normals fitting the scores of a retrieval system. The histogram represents the distribution of scores of the system, the dashed line on the left is the Log-Normal associated to non-relevant documents (peak centered at a low score), the dashed line on the right is the Log-Normal associated to relevant documents (peak centered at a high score), and the solid line is the mixture of both Log-Normals.

The document score distributions are modeled as a mixture of relevant and non-relevant distributions as follows:

$$p(score) = \lambda \cdot p(score|rel) + (1-\lambda) \cdot p(score|nonrel) \quad (1)$$

where $\lambda \in [0, 1]$ is the mixing weight, and $p(.|rel)$ ($p(.|nonrel)$) is the probability density function of the relevant (non-relevant) distribution. This two-component mixture model makes explicit that i) each score in a ranked list is associated to a document that is either relevant or non-relevant, and ii) the distribution of scores in relevant and non-relevant documents are not necessarily the same (the separation into two components, $p(.|rel)$ and $p(.|nonrel)$, permits to make a distinction between the pattern of scores of relevant and non-relevant documents).

Modeling the scores with two Log-Normal distribution leads to:

$$p(score|rel) = \frac{1}{score \cdot \sigma_{rel} \cdot \sqrt{2\pi}} \cdot e^{-\frac{(ln\ score\ -\ \mu_{rel})^2}{2\sigma_{rel}^2}}, score > 0 \quad (2)$$

$$p(score|nonrel) = \frac{1}{score \cdot \sigma_{nonrel} \cdot \sqrt{2\pi}} \cdot e^{-\frac{(ln\ score\ -\ \mu_{nonrel})^2}{2\sigma_{nonrel}^2}}, score > 0 \quad (3)$$

where μ_{rel} and σ_{rel} (resp. μ_{nonrel}, σ_{nonrel}) are the parameters of the Log-Normal distribution. Note that scores need to be positive.[2] Log-Normal distributions have shown to be a good fit for modeling the scores of multiple retrieval systems [16]. Fig. 1 shows an example of two Log-Normal distributions (dashed lines) that have been fitted from the scores of relevant and non-relevant documents computed by a retrieval system in response to a query. The dashed

[1] The Log-Normal distribution is a continuous distribution of a random variable whose logarithm has a Normal Distribution.

[2] The occurrence of negative scores can be overcome by shifting all scores by some constant factor. It is standard practice to make first this normalization and, next, apply the SD models. For systems that supply negative scores, we adjust each score s as follows: $s = s - min(scores) + 1$, where $min(scores)$ is the minimum score computed by the system.