Modeling Document Scores for Distributed Information Retrieval

Ilya Markov
Faculty of Informatics, University of Lugano
Lugano, Switzerland
ilya.markov@usi.ch

ABSTRACT

Distributed Information Retrieval (DIR), also known as Federated Search, integrates multiple searchable collections and provides direct access to them through a unified interface [3]. This is done by a centralized *broker*, that receives user queries, forwards them to appropriate collections and returns merged results to users.

In practice, most of federated resources do not cooperate with a broker and do not provide neither their content nor the statistics used for retrieval. This is known as *uncooperative* DIR. In this case a broker creates a *resource representation* by sending sample queries to a collection and analyzing retrieved documents. This process is called *query-based sampling.* The key issue here is the following:

1.1 *How many documents have to be retrieved from a resource in order to obtain a representative sample?*

Although there have been a number of attempts to address this issue it is still not solved appropriately.

For a given user query resources are ranked according to their similarity to the query or based on the number of relevant documents they contain. Since resource representations are usually incomplete, the similarity or the number of relevant documents cannot be calculated precisely. Resource selection algorithms proposed in the literature estimate these numbers based on incomplete samples. However these estimates are subjects to error. In practice, inaccurate estimates that have high error should be trusted less then the more accurate estimates with low error. Unfortunately none of the existing algorithms can make the calculation of the estimation errors possible. Therefore the following questions arise:

2.1 *How to estimate resource scores so that the estimation errors can be calculated?*

2.2 *How to use these errors in order to improve the resource selection performance?*

Existing results merging algorithms estimate normalized document scores based on scores of documents that appear both in a sample and in a result list. The problem similar to the resource selection one arises. The normalized document scores are only the estimates and are subjects to error. Inaccurate estimates should be trusted less then the more accurate ones. Again none of the existing algorithms provide a way for calculating these errors. Thus the two question to be address on the results merging phase are similar to the resource selection ones:

3.1 *How to estimate normalized document scores so that the estimation errors can be calculated?*

3.2 *How to use these errors in order to improve the results merging performance?*

In this work we address the above issues by applying score distribution models (SDM) to different phases of DIR [2]. In particular, we discuss the SDM-based resource selection technique that allows the calculation of resource score estimation errors and can be extended in order to calculate the number of documents to be sampled from each resource for a given query. We have performed initial experiments comparing the SDM-based resource selection technique to the state-of-the-art algorithms and we are currently experimenting with the SDM-based results merging method.

We plan to apply the existing score normalization techniques from meta-search to the DIR results merging problem [1]. However, the SDM-based results merging approaches require the relevance scores to be returned together with retrieved documents. It is not yet clear how to relax this strong assumption that does not always hold in practice.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms

Keywords

Distributed Information Retrieval, Federated Search, Score Distribution Models

1. REFERENCES

[1] A. Arampatzis and J. Kamps. A signal-to-noise approach to score normalization. In *Proceedings of the ACM CIKM*, pages 797–806. ACM, 2009.

[2] A. Arampatzis and S. Robertson. Modeling score distributions in information retrieval. *Inf. Retr.*, 14(1):26–46, 2011.

[3] J. Callan. *Advances in Information Retrieval*, chapter Chapter 5. Distributed Information Retrieval, pages 127–150. Kluwer Academic Publishers, 2000.

SIGIR'11, July 24–28, 2011, Beijing, China.
ACM 978-1-4503-0757-4/11/07.