A Weighted Curve Fitting Method for Result Merging in Federated Search

Chuan He
Computer Department
Beijing University of Posts and Telecommunications
Haidian District, Beijing, China
hechuanbupt@gmail.com

Dzung Hong
Department of Computer Science
Purdue University
West Lafayette, IN 47907, USA
dthong@cs.purdue.edu

Luo Si
Department of Computer Science
Purdue University
West Lafayette, IN 47907, USA
lsi@cs.purdue.edu

ABSTRACT

Result merging is an important step in federated search to merge the documents returned from multiple source-specific ranked lists for a user query. Previous result merging methods such as Semi-Supervised Learning (SSL) and Sample-Agglomerate Fitting Estimate (SAFE) use regression methods to estimate global document scores from document ranks in individual ranked lists. SSL relies on overlapping documents that exist in both individual ranked lists and a centralized sample database. SAFE goes a step further by using both overlapping documents with accurate rank information and documents with estimated rank information for regression. However, existing methods do not distinguish the accurate rank information from the estimated information. Furthermore, all documents are assigned equal weights in regression while intuitively, documents in the top should carry higher weights. This paper proposes a weighted curve fitting method for result merging in federated search. The new method explicitly models the importance of information from overlapping documents over non-overlapping ones. It also weights documents at different positions differently. Empirically results on two datasets clearly demonstrate the advantage of the proposed algorithm.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H3.3 Information Search and Retrieval; H.3.4 Systems and Software

General Terms

Algorithms, Design, Performance

Keywords

Federated Search, Result Merging, Curve Fitting

1. INTRODUCTION

Text information behind individual search engines of distributed information sources may not be easily crawled for building a centralized index because of data protection, copyrights or security. Federated text search has been proposed to search the distributed information in these environments. There are three basic problems of federated search: resource representation, resource selection and result merging. This paper focuses on result merging for generating a single ranked list of documents in different sources for a user query.

Existing result merging algorithms can be categorized into two groups. The first group assumes some level of collaboration of distributed information sources and use some statistics provided by those sources for merging [1, 3]. In non-cooperative environment, a semi-supervised learning (SSL) [5] method is a more practical approach. Basically, after the first step of query-based sampling [1], each information source is represented by a set of sample documents. The set of all sample documents is called *centralized sample database*. SSL utilizes overlapping documents in both individual ranked lists and centralized sample database to build a regression. Once regression is done, SSL can convert the rank of any document returned from an individual source to that document's centralized score. Those centralized scores are used as global scores to merge all documents.

The more recent Sample-Agglomerate Fitting Estimate (SAFE) method [4] relaxes the requirement of SSL for overlapping documents. It estimates document ranks based on the uniform sampling assumption, and uses those estimated ranks for regression. SAFE algorithm however, does not distinguish the contribution of overlapping documents with accurate ranks (i.e., existing in the source's returned list) and sample documents with estimated ranks for regression. Moreover, top ranked documents (in source-specific list) are probably more important for curve fitting because of the goal of high-precision, which is not considered in SAFE.

Based on the observation, we propose a novel result merging method called Weighted Curve Fitting (WCF) that combines the features of SSL and SAFE for result merging. The new method distinguishes accurate and estimated rank information. It also considers the importance of documents in different positions for regression. Section 2 proposes the algorithm and Section 3 presents the empirical results.

2. WEIGHTED CURVE FITTING RESULT MERGING

We describe the method of estimating document ranks in SAFE, which is now reused in WCF. Suppose that the sampling process is uniform, a centralized document of rank r in the individual sample source is estimated to have rank P in the corresponding remote source, where we define $P = r \cdot |c_k|/|\theta_k|$ ($|c_k|$ and $|\theta_k|$ are the number of documents in the

This research was partially supported by the NSF research grants IIS-0746830, CNS-1012208, IIS-1017837, and a research grant from Google Inc.

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

source c_k and its sample θ_k). P is then used for regression if the document is unseen in the source's ranked list.

As mentioned before, SAFE does not differentiate the contribution of overlapping documents and non-overlapping ones in making regression. It is better to put more weights on overlapping documents, since their ranks are more accurate than those estimated. Moreover, by assuming that top documents are more representative than the others, documents on top of the returned ranked list should contribute to the regression more than documents at the bottom. Formally, given a query and a ranked list of documents that match the query in the centralized database, let S_j be the centralized relevance score of document d_j, and let r_j be the rank of d_j in the source's individual ranked list. Note that r_j could either be real or estimated, depending on whether d_j appears in the individual ranked list. Then both SSL and SAFE attempt to fit the curve $S_j = a \cdot f(r_j) + b$, where f is any function, and a, b are the learning parameters. We fit the curve by minimizing the quadratic loss function:

$$G_{SSL} = min \sum_j ||S_j - (a \cdot f(r_j) + b)||^2$$

Now we add two more parameters for the new model:

$$G_{WCF} = min \sum_j \frac{1}{\varphi_e^2(d_j)} \cdot \frac{1}{\psi_p^2(d_j)} \Big|\Big| S_j - (a \cdot f(r_j) + b) \Big|\Big|^2$$

where φ_e is a function that gives larger value for the document with an estimated rank than a document with a true rank. ψ_p is a function depending on the position of document d_j in the centralized ranked list. In this paper, we choose a linear function for ψ_p, i.e. $\psi_p(d_j) = r_j$. We also choose a two-level function for φ_e:

$$\varphi_e(d_j) = \begin{cases} 1 & \text{if } d_j \text{ has true rank} \\ c (c > 1) & \text{if } d_j \text{ has estimated rank} \end{cases}$$

c is set to 16 in all the experiments, to indicate the penalized term for estimated ranks. We call this method Weighted Curve Fitting (WCF). Parameter estimation can be done by rewriting the objective function and solving the normal linear regression. Let $w_j = \varphi_e(d_j) * \psi_p(d_j)$, then

$$G_{WCF} = min \sum_j \Big|\Big| \frac{S_j}{w_j} - (\frac{a}{w_j} \cdot f(r_j) + \frac{b}{w_j}) \Big|\Big|^2$$

3. EXPERIMENTAL RESULTS

We conducted our experiments with two standard TREC datasets: trec123, which contains 100 collections of TREC CDs 1, 2 and 3 organized by publication sources; and trec4-kmeans, which contains 100 collections created from TREC4 by k-means clustering [1]. Each information source uses a different retrieval model among INQUERY [2], language modeling and vector space tf-idf. We sample 300 documents for each source, use CORI [1] to select the top 5 sources for each query and INQUERY for querying the centralized sample database. As for the function f, SAFE uses linear, log, square, power and hybrid functions. We follow the same approach, but only report and compare results of the hybrid choice, which was shown to obtain best results in [4].

Table 1 shows the precision results when top 20 documents returned from each source are merged. We also conduct experiments with top 50 and 100 documents selected from each source. The full result is shown in Figure 1.

Table 1: Document Precision on TREC123, TREC4 with Top 20 Documents of each Source

	TREC123				TREC4_kmeans			
P@n	@5	@10	@15	@20	@5	@10	@15	@20
SAFE	.32	.29	.27	.25	.27	.22	.19	.17
WCF	.34	.31	.29	.27	.31	.26	.24	.24

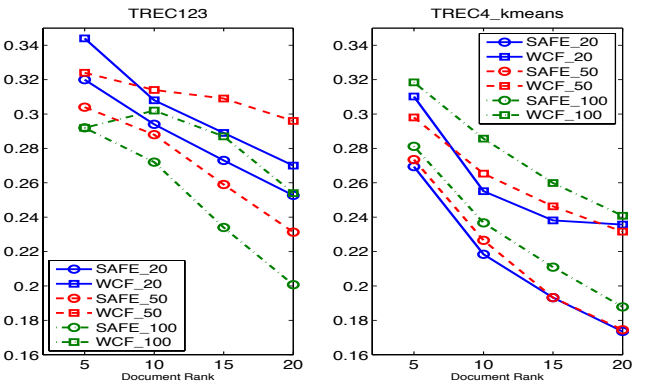

Figure 1: Document Precision with Top 20, 50, 100 Documents of each Source

In both datasets, the curves of the same color denote the performance of SAFE and WCF of the same setting. In general, WCF consistently outperforms SAFE. For trec4-kmeans, the precision is increasing when more documents from each source are merged. However, this is not the case with trec123. It may be argued that the algorithm has to deal with more noisy data when merging more documents, so a possible solution is to assign even smaller weight to documents along the tail of the returned list. The exploration of using other functions for ψ_p is reserved for the future work. In general, WCF works well without too much overhead.

4. CONCLUSION

This paper proposes a result merging algorithm for federated search. Similar to SAFE, the new method does not rely on overlapping documents. More importantly, the new method can accurately estimate global document scores for result merging by distinguishing accurate rank information and estimated rank information in regression. It also meets the high-precision criterion by putting more weights for documents in the top part of the ranked lists for curve-fitting. Empirical results on two datasets have shown the effectiveness of the proposed results merging algorithm.

5. REFERENCES

[1] J. Callan. Distributed information retrieval. *Advances in Information Retrieval*, pages 127–150, 2000.

[2] J. Callan, W. B. Croft, and S. M. Harding. The inquery retrieval system. In *Proceedings of the Third International Conference on Database and Expert Systems Applications*, 1992.

[3] S. Kirsch. Document retrieval over networks wherein ranking and relevance scores are computed at the client for multiple database documents, Aug. 19 2003.

[4] M. Shokouhi and J. Zobel. Robust result merging using sample-based score estimates. *ACM Transactions on Information Systems*, 27(3):1–29, 2009.

[5] L. Si and J. Callan. A semi-supervised learning method to merge search engine results. *ACM Transactions on Information Systems*, 21(4):457–491, 2003.