Bilkent University

Distributed Information Retrieval

By: Jamie Callan

7th of March, 2013

Group members:

Amir R. Ilkhechi (Emir R. Yılkıcı) – Yağız Salor – M.İlker Saraç – Hakan Sözer

1. Motivation

- Web search engines and large networks are usually based on a single database model of text retrieval, in which documents from around the network are copied to a centralized database, where it is indexed and made searchable.

2. Problem Definition

- However information that cannot be copied is not accessible under single database model.

- Information that is proprietary or that publisher wishes to control carefully is essentially invisible to the single database model.

3. Solution

- Multi-database model is the alternative to the single database model.

- A central site stores brief descriptions of each database, and a database selection service uses these resource descriptions to identify the databases that are most likely to satisfy each information need.

- Multi-database model can be applied to this problem due to the central site does not require copies of the documents in each database. However, it is also more complex than the single database model of information retrieval so multi-database model requires several additional problems to be addressed:

a. **Resource description:** The contents of each text database must be described. Simple and robust solution is to represent each database by a description consisting of the words that occur in the database, and their frequencies of occurrence. This type of representation is called *unigram language model.*

b. **Resource Selection:** Given information need and a set of resource descriptions, a decision must be made about which database(s) to search. The main approach is to

apply the techniques of document ranking to the problem of resource ranking, using variants of *tf.idf* approaches.; and

c. **Resource Merging:** Integrating the ranked lists returned by each database into a single, coherent ranked list. For that purpose the most accurate solution is to normalize the scores of documents from different databases, either by using global corpus statistics or by re-computing document scores at the search client.

4. Results

Acquiring Resource Description

Acquiring resource descriptions can be a difficult problem, especially in a wide-area network containing resources controlled by many parties. Solutions that require cooperation are appropriate in controlled environments. If a resource provider can't cooperate or refuses to cooperate, or is deceptive, the cooperative approach fails.

An alternative solution is for the resource selection service to learn what each resource contains by submitting queries and observing the documents that are returned. (query-based sampling)

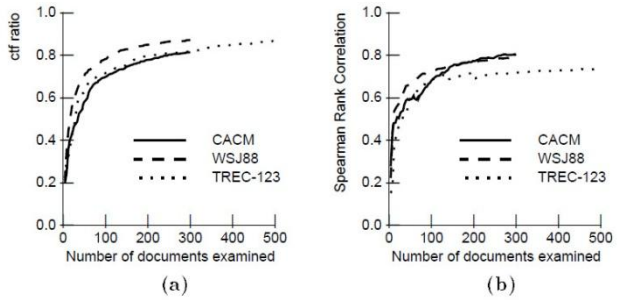
4.1 Accuracy of Unigram Language Models

The first tests of query based sampling studied how well the learned language models matched the actual or complete language model of a database. Ctf ratio is the proportion of term occurrences in the database that are covered by terms in the learned resource description. For a learned vocabulary V and an actual vocabulary V, ctf ratio is:

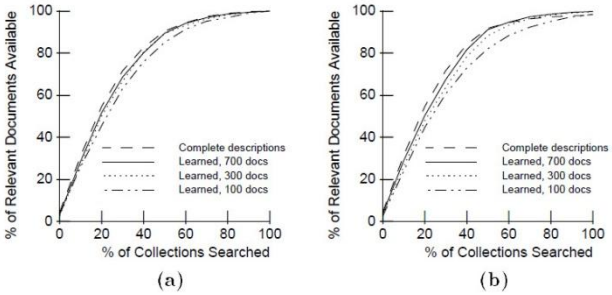
$$\frac{\sum_{i \in V'} ctf_i}{\sum_{i \in V} ctf_i}$$

The Spearman Rank Correlation Coefficient is defined as:

$$R = \frac{1 - \frac{6}{n^3-n}(\Sigma d_i^2 + \frac{1}{12}\Sigma(f_k^3 - f_k) + \frac{1}{12}\Sigma(g_m^3 - g_m))}{\sqrt{(1 - \frac{\Sigma(f_k^3-f_k)}{n^3-n})}\sqrt{(1 - \frac{\Sigma(g_m^3-g_m)}{n^3-n})}}$$



4.2 ACCURACY OF RESOURCE RANKINGS



4.3 ACCURACY OF DOCUMENT RANKINGS

	Topics 51-100 (INQ026 queries)		*Topics 101-150 (INQ001 queries)*	
Document Rank	*Complete Resource Descriptions*	*Learned Resource Descriptions*	*Complete Resource Descriptions*	*Learned Resource Descriptions*
5	0.5800	0.6280 (+8.3%)	0.5960	0.5600 (−6.0%)
10	0.5640	0.6040 (+7.1%)	0.5540	0.5520 (−0.3%)
15	0.5493	0.5853 (+6.6%)	0.5453	0.5307 (−2.7%)
20	0.5470	0.5830 (+6.6%)	0.5360	0.5270 (−1.7%)
30	0.5227	0.5593 (+7.0%)	0.5013	0.4993 (−0.4%)

Summary statistics for the query sets used with the testbed.

Query Set Name	*TREC Topic Set*	*TREC Topic Field*	*Average Length (Words)*
Title queries, 51-100	51-100	Title	3
Title queries, 101-150	101-150	Title	4
Description queries, 51-100	51-100	Description	14
Description queries, 101-150	101-150	Description	16

5 Conclusion

The research reported in this paper addresses many of the problems that arise when full-text information retrieval is applied in environments containing many text databases controlled by many independent parties. The solutions include techniques for acquiring descriptions of resources controlled by uncooperative parties using resource descriptions to rank text databases by their likelihood of satisfying a query and merging the document rankings returned by different text databases. Collectively, these techniques represent an end-to-end solution to the problems that arise in distributed information retrieval.