Provenance in Keyword Search

Wensheng Wu
University of North Carolina, Charlotte
w.wu@uncc.edu

1. INTRODUCTION

While database and scientific communities have extensively researched data and workflow provenance in the past several years [1], the provenance in keyword search has received very little attention. However, the provenance of search results is important to end users since they often want to know why document A is ranked higher than document B and why document C is not included in the results. It is also important for developers of search systems, e.g., for debugging and tuning purposes.

Existing keyword search systems provide very primitive facilities, if any, that show how the scores of search results are derived. Such explanations are typically targeted at developers. For example, consider searching DBLP for publications about "big data". Figure 1 shows a snippet of the explanation that Lucene [2] generates. It shows that the score of "from big data to big knowledge" is 5.9 and is obtained by summing up the individual scores for "big" and "data", and so on.

```
8. 3770855      From big data to big knowledge  5.924873
5.924873 = (MATCH) sum of:
  5.017788 = (MATCH) weight(title:big in 15160) [DefaultS
    5.017788 = score(doc=15160,freq=2.0 = termFreq=2.0
), product of:
      0.8924116 = queryWeight, product of:
        9.087702 = idf(docFreq=1177, maxDocs=3833449)
        0.09819992 = queryNorm
      5.6227283 = fieldWeight in 15160, product of:
```

Figure 1: Explanation facility in Lucene

However, such an explanation is inadequate in several aspects. (1) *Incomplete*: It might not include all steps (e.g., stemming and prefix matching) or details (e.g., formulas for computing term weights) necessary to determine exactly how the final score is derived. (2) *Hard to interpret*: It uses tool-specific names and notations to specify the concepts and steps in the derivation process. For example, it is not clear what queryNorm means without consulting with the Lucene documentation. (3) *Lacking support for why- and where-provenances*: Users need to manually compare the derivations of different documents to understand their relative ranking (why-provenance), and search the documents for the evidences that support their rankings (where-provenance).

7th Biennial Conference on Innovative Data Systems Research (CIDR'15) January 4–7, 2015, Asilomar, California, USA.

2. TOWARD EXPLAINABLE SEARCH

To address the drawbacks of existing explanation facilities, we propose to systematically study the provenance of keyword search, with a focus on three areas.

(1) Scoring ontology: Develop an ontology that describes common concepts in scoring (e.g., document, query, and term) and their relationships such as subclass (e.g., a query term is a term), compositional (e.g., a document is formed by a set of terms), and derivational (e.g., IDF is derived from document frequency) relationships. The scoring ontology is essentially a conceptual model that captures the commonality (and differences) among varied scoring functions.

(2) Scoring schema: Translate the ontology into an XML schema for querying provenance, where both concepts and relationships in the ontology may be represented using elements in the schema. The process is analogous to that of converting ER model to relational schema in database design. Schema elements (e.g., for "document term frequency") may have parametric natural language annotations (e.g., "the number of times a term t appears in a document d") that can be instantiated into user-friendly explanations.

(3) Querying provenance: Design a declarative language for expressing queries on search provenance recorded using the scoring schema. Develop an intuitive query interface that enables users to ask provenance questions without the knowledge of the query language. User questions will be translated into provenance queries over the scoring schema. The provenance query language may support three types of queries: (a) *How:* For example, explain how the final (partial) score of a document is derived; and how different scoring functions, e.g., Cosine and Okapi BM25, affect the ranking of documents. (b) *Why:* For example, explain why two given documents have different scores; and why a particular document has a very low/high score. (c) *Where:* For example, "drill-through" to the documents to answer questions such as "show me the places in the document D where term t occurs or terms t and s are in close proximity".

3. REFERENCES

[1] P. Buneman, S. Khanna, and T. Wang-Chiew, "Why and where: A characterization of data provenance," in *ICDT*, 2001, pp. 316–330.

[2] "Lucene," http://lucene.apache.org/.