SourceRank: Relevance and Trust Assessment for Deep Web Sources Based on Inter-Source Agreement

Raju Balakrishnan, and Subbarao Kambhampati *
Computer Science and Engineering, Arizona State University
Tempe AZ USA 85287
rajub@asu.edu, rao@asu.edu

ABSTRACT

We consider the problem of deep web source selection and argue that existing source selection methods are inadequate as they are based on local similarity assessment. Specifically, they fail to account for the fact that sources can vary in trustworthiness and individual results can vary in importance. In response, we formulate a global measure to calculate relevance and trustworthiness of a source based on agreement between the answers provided by different sources. Agreement is modeled as a graph with sources at the vertices. On this agreement graph, source quality scores—namely *SourceRank*—are calculated as the stationary visit probability of a weighted random walk. Our experiments on online databases and 675 book sources from Google Base show that SourceRank improves relevance of the results by 25-40% over existing methods and Google Base ranking. SourceRank also reduces linearly with the corruption levels of the sources.

Categories and Subject Descriptors

H.3.3 Information Search and Retrieval [**Selection Process, Search Process.**]:

General Terms

Algorithms, Experimentation.

Keywords

web databases, source selection, deep web, source trust.

1. INTRODUCTION

Selecting the most relevant subset of web databases for answering a given query is an important problem in deep web integration. Source selection for text and relational databases involving relevance, coverage, and the overlaps between sources has received some previous attention ([5, 3]). These existing approaches are focused on assessing relevance of a source based on local measures, as they evaluate quality of a source based on the similarity between the answers provided by the source and the query. For application to the deep web, this pure query based local approach for source selection has the following two deficiencies:

(i) Query based relevance is insensitive to the importance of source results. For example, the query *godfather* matches the classic movie *The Godfather* and the little known movie *Little Godfather*. Intuitively, most users would be looking for the classic movie.

(ii) The source selection is insensitive to the trustworthiness of the answers. For example, many queries in Google Products return answers with unrealistically low prices. Only when the user proceeds towards the checkout, many of these low priced results turn out to be non-existing, a different product with same title (e.g. solution manual of the text book) etc.

A global measure of trust and relevance is particularly important for uncontrolled collections like deep web, since sources generally try to artificially boost their rankings. Our broad plan of attack is to adapt the link-analysis techniques used for ranking pages on the surface web [2]. The main stumbling block is that there are no explicit hyper-link based endorsements among deep web sources. We surmount this by defining implicit endorsement structure among sources in terms of the *agreement* between the results returned by sources for sample queries. Two sources agree with each other if both return the same tuples in answer to a query.

Agreement based analysis would be able to solve the problems (i) and (ii) mentioned above. Considering problem (i) above, the important results are likely to be returned by large number of sources. For example, the classic Godfather is returned by hundreds of sources while the *Little Godfather* is returned by less than ten on a Google Product Search. Similarly regarding trust, source corruption can be captured since other fair sources are unlikely to agree with corrupted databases (*c.f.* [6]). Please refer to Balakrishnan *et al.* [1] for a formal argument.

2. AGREEMENT ANALYSIS AND SOURCERANK COMPUTATION

We represent the agreement between the source result sets as an agreement graph. In the agreement graph, vertices are sources, and edge weight $w(S_1 \rightarrow S_2)$ of the link from S_1 to S_2 is computed as,

$$A_Q(S_1, S_2) = \sum_{q \in Q} \frac{A(R_{1q}, R_{2q})}{|R_{2q}|} \quad (1)$$

$$w(S_1 \rightarrow S_2) = \beta + (1 - \beta) \times \frac{A_Q(S_1, S_2)}{|Q|} \quad (2)$$

*This research is supported by ONR grant N000140910032 and a 2008 Google research award.

WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

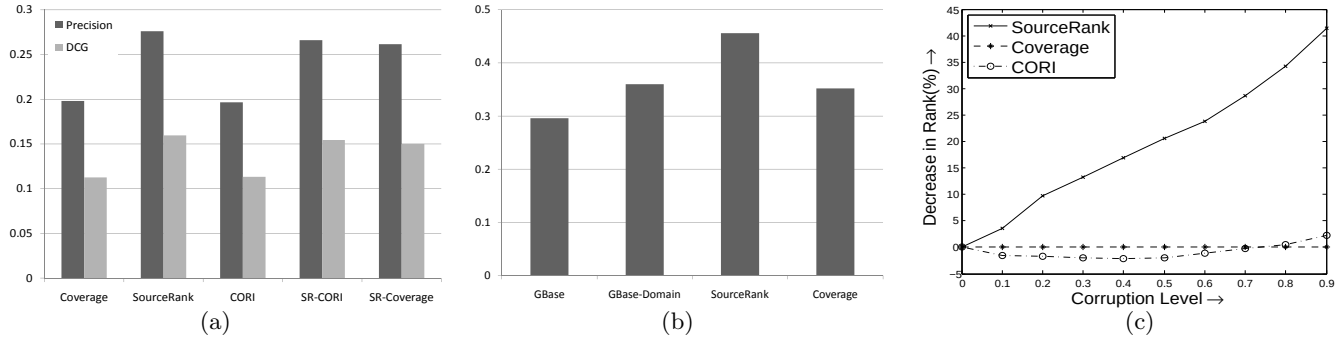

Figure 1: (a) Comparison of precision and DCG of top-4 sources selected by Coverage, SourceRank, CORI, two combinations for online databases. (b) Comparison of top-5 precision of SourceRank, Google Base and Coverage. (c) Decrease in ranks of the sources with increasing source corruption levels.

where R_{1q} and R_{2q} are the answer sets of S_1 and S_2 for the query q, and Q is the set of sampling queries over which the agreement is computed. β is the smoothing factor accounting for the unseen samples (set to 0.1). Semantically, the edge weight is equal to the fraction of tuples in S_2 agreed by S_1. We calculate agreement between the sources in three steps: computing (i) attribute value similarity (ii) tuple similarity (iii) answer set similarity [1] [4].

It can be shown that—if provided with the agreement graph—a rational search strategy for a searcher would be a weighted markov random walk on the graph [1]. This implies that the visit probability of the searcher on a node is equal to the stationary visit probability of the random walk. Hence we calculate the SourceRank of a database node as the static visit probability of a random walk on the node.

3. EVALUATION AND CONCLUSION

We experimented with two sets of book seller databases—(i) a set of twenty seven online book sources accessed by their own web forms from UIUC TEL-8 Repository (ii) 675 book sources on Google Base. These 675 book sources are selected by sending ten book queries to Google Base, and collecting all source ids in first 400 ranked results. For test queries, we used 25 books for youth from American Library Association. For both test sets, we compared the top-k precision and trustworthiness of results returned by SourceRank with those of (i) Coverage [5] based selection used in relational databases (coverage is calculated as sum of relevances of top−5 results), (ii) CORI [3] in text databases, and (iii) the default ranking used in Google Base.

Relevance Results: For the online databases we compared mean top-5 precision and discounted cumulative gain (DCG) of top-4 sources (normalization in NDCG since rank lists are of equal length). Five methods, namely Coverage, SourceRank, CORI, and two linear combinations of SourceRank with CORI and Coverage—$(0.1 \times SourceRank + 0.9 \times CORI)$ and $(0.5 \times Coverage + 0.5 \times SourceRank)$—are compared (the higher weight for CORI in combinations is to compensate for the higher dispersion of SourceRank scores). For the results in Figure 1(a), SourceRank improves precision over both CORI and Coverage by approximately 40% ($\frac{0.27-0.19}{0.19}$); and DCG by approximately 41%.

For the Google Base sources, we tested if the precision of Google Base search results can be improved by combining SourceRank with Google Base ranking. In Figure 1(b) the *GBase* corresponds to the stand-alone Google Base relevance ranking. *GBase-Domain* is the Google Base ranking searching only in the domain sources selected using our query probing (i.e. 675 book sources). SourceRank and Coverage are Google Base tuple rank applied to the tuples from top-10% sources selected by the SourceRank and Coverage based source selections respectively. Note that for the books domain, GBase-Domain and Coverage are performing almost equally, while SourceRank improves precision by 26%.

Trust Results: For trust evaluation, we corrupted a randomly selected subset of sources by replacing attributes not constrained in the query (i.e. attributes other than titles, since we used partial titles as queries) with random strings. (Note that the corruption in attributes not constrained in the query generates untrustworthy results; whereas difference in constrained attributes generates irrelevant results) SourceRank, Coverage and CORI ranks are recomputed using these corrupted crawls. Mean reduction in ranks of the corrupted sources are calculated over 50 runs. Since CORI is query specific, the decrease in CORI rank is calculated as the average decrease in rank over ten queries. In Figure 1(c), the Coverage and CORI are insenseitive to the corruption, whereas the SourceRank of corrupted sources reduces almost linearly with the corruption level.

These empirical results allow us to conclude that the agreement-based analysis provides an effective framework for selecting sources on the deep web.

4. REFERENCES

[1] Sourcerank: Relevance and trust assessment for deep web sources. ASU CSE TR 2009. http://www.public.asu.edu/~rbalakr2/papers/SourceRank.pdf.
[2] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, 1998.
[3] J. Callan, Z. Lu, and W. Croft. Searching distributed collections with inference networks. In *Proceedings of ACM SIGIR*, pages 21–28. ACM, NY, USA, 1995.
[4] W. Cohen. Integration of heterogeneous databases without common domains using queries based on textual similarity. *ACM SIGMOD Record*, 27(2):201–212, 1998.
[5] Z. Nie and S. Kambhampati. A Frequency-based Approach for Mining Coverage Statistics in Data Integration. *Proceedings of ICDE*, page 387, 2004.
[6] X. Yin, J. Han, and P. Yu. Truth discovery with multiple conflicting information providers on the web. *IEEE TKDE*, 20(6), 2008.