Factal: Integrating Deep Web Based on Trust and Relevance*

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

ABSTRACT

We demonstrate *Factal*—a system for integrating deep web sources. Factal is based on the recently introduced source selection method *SourceRank*; which is a measure of trust and relevance based on the agreement between the sources. SourceRank selects popular and trustworthy sources from autonomous and open collections like the deep web. This trust and popularity awareness distinguishes Factal from the existing systems like Google Product Search. Factal selects and searches active online databases on multiple domains. The demonstration scenarios include improved trustworthiness, relevance of results, and comparison shopping. We believe that by incorporating effective source selection based on the SourceRank, Factal demonstrates a significant step towards a deep-web-scale integration system.

Categories and Subject Descriptors

H.3.5 [**INFORMATION STORAGE AND RETRIEVAL**]: Online Information Services—*Web-based services*

General Terms

Algorithms, Design

1. INTRODUCTION

By many accounts, surface web containing HTML pages is only a fraction of the overall information available on the web. The remaining is hidden behind millions of web-accessible relational databases [7]. The most promising approach that has emerged for searching and exploiting sources on the deep web is data integration. One immediate challenge in realizing deep web integration is source selection—selecting the most relevant subset of sources for answering a query.

Source selection involving coverage and latency of the source, and the overlaps between sources has received some previous attention in data integration (c.f. [8, 3]). Existing approaches are focused on assessing the relevance of a source based on local measures; as they evaluate the quality of the source based on the similarity between the answers provided by the source and the query. When applied to the deep web, this local approach for source selection has the following two deficiencies:

1. A purely query based relevance assessment is insensitive to the importance of the source results. For example, Figure 1(a) shows the results to the query *Godfather Trilogy* by Google Product Search. Intuitively users will be searching for the classic Godfather movie trilogy. Since the search assesses the relevance by similarity of query with the product title, titles containing words "godfather" and "trilogy" are returned as the top results as observed in the Figure 1(a). Unfortunately, none of these results refers to the intended classic movie trilogy (or even the classic book by the same name).

2. Existing source selection is agnostic to the trustworthiness of the answers. Relevance of a tuple is a measure of whether the tuple is answering the query. On the other hand, trustworthiness is a measure of the correctness of the answer. Insensitivity to trustworthiness exposes the users to bait and switch behavior

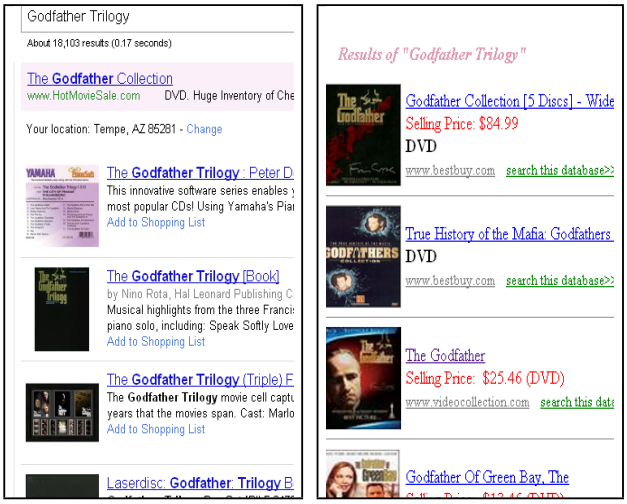

Figure 1: Comparison of results to the query *Godfather Trilogy* from (a) Google Product Search and (b) Factal. None of the top results of Google Products refer to the classic Godfather, whereas many results in Factal including top result are correct.

*This research is supported by ONR grant N000140910032 and two Google research awards.

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

of shopping sites. Databases in Google Base may return copies of the the same book with very low prices. While the user proceeds towards the checkout, the item would turn out to be out of stock, and many times a different item with the same title and cover. For example, the second result titled "The Godfather Trilogy [Book]" in Figure 1(a) happens to be a totally different book.

A global measure of trust and relevance is particularly important for uncontrolled collections like the deep web, since sources generally try to artificially boost their rankings. The problems described above may be significant factors in preventing the adaptation of current systems (c.f. after eight years of its introduction Google Product Search is still in beta testing stage). These problems bear resemblance to the problems in the pre-link analysis era or the surface web search engines [2]. Hence our broad plan of attack is to adapt the link-analysis techniques used for page ranking on the surface web. The main stumbling block is that there are no explicit hyper-link based endorsements among deep web sources. We overcome this hurdle 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 of importance and trust mentioned above. The importance is considered since the important results are likely to be returned by a large number of sources. For example, the classic Godfather is returned by hundreds of sources while the other results are returned only by very few sources on a Google Product Search. Similarly regarding trust, source corruption can be captured since it is unlikely that two independent sources agree on corrupt results. For example, in Figure 1(a) an agreement based analysis will reveal that the authors of the result "The Godfather Trilogy [Book]" are different from the authors of the classic Godfather returned by large number of sources. Notice that as shown in Figure 1(b) our system—Factal—is indeed able to overcome these problems and rank the Godfather Movie Trilogy at the top.

Agreement computation between the web databases poses multiple challenges that necessitate combination and extension of methods from relational and text databases. The primary problem is that different web databases represent the same entity syntactically differently, making the agreement computation hard [4]. To solve this problem, we combine record linkage models with entity matching techniques for accurate and speedy agreement computation. Further, attribute matchings are weighted against the computed attribute importance.

Like PageRank, the databases may enhance SourceRank by colluding with each other. Differentiating genuine agreement between the sources from collusion increases the robustness of SourceRank. We measure and compensate for the source collusion while computing SourceRank.

2. RELATED WORK

Current relational database selection methods minimize cost by retrieving the maximum number of distinct records from minimum number of sources [8], but do not consider the problem of trust and importance of results addressed here. Callan *et al.* [3] formulated the CORI method for query specific selection of text databases based purely on query relevance.

A probabilistic framework for trust assessment based on agreement for question answering has been presented by Yin *et al.* [9]. Their framework however does not consider the influence of relevance on agreement, multiple correct answers to a query, record linkage and non-cooperative sources; thus limiting its usability for deep web. Dong *et al.* [6] extended this model considering source dependence using the same basic model as Yin *et al.* The collusion detection in deep web needs to address different constraints like multiple true values, non-cooperative sources, and ranked answer sets.

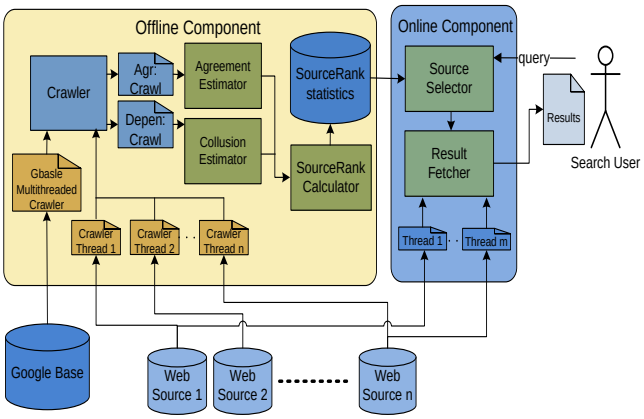

Figure 2: System Architectural Diagram. The online component contains processing steps at query time. Both the crawling and search are parallelized. (URL of the system is http://factal.eas.asu.edu).

3. THE FACTAL SYSTEM

As shown in Figure 2 the Factal system has an offline component and an online component. The offline component crawls the sources and calculates the source statistics. The online component selects the sources to search based on the statistics and fetches the results at query time. The approach is domain independent, and we use movie and book sources for this demonstration. We search 22 stand alone online sources in each domain, along with 610 book sources and 209 movie sources from Google Base.

3.1 Crawling Sources

For online sources one thread per data base is used for crawling, and for Google Base we used forty threads (acceptable for Google Base). Since users hardly go below the top few results, top-5 results from each database are used for calculating source statistics below. To generate crawling queries, we randomly selected 200 books and movies from New York Times yearly best sellers and New York Times movie guide respectively. We use partial titles as queries for crawling by removing words randomly from titles with 0.5 probability.

3.2 Agreement Estimation

Text similarity measures work best for computing similarity between the web database tuples, since there are no

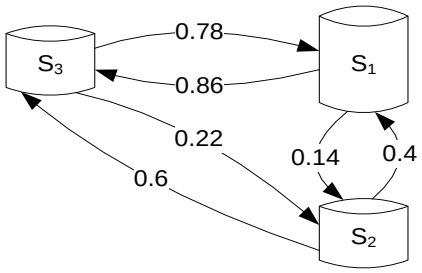

Figure 3: A sample agreement graph of three sources.

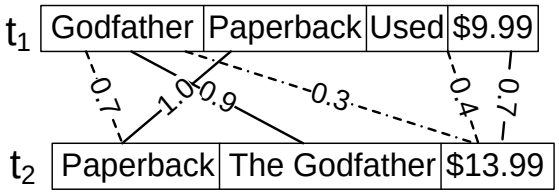

Figure 4: Example tuple similarity calculation. The dotted line edges denote the similarities computed, and the solid edges represent the matches selected by the greedy matching algorithm.

common domains for attribute values [4]. For matching between the answer sets of databases, we need three levels of similarity computations: (i) similarity between the attribute values in tuples, (ii) similarity between tuples, and (iii) similarity between answer sets.

For computing (i), we used the hybrid similarity function SoftTFIDF [5]. SoftTFIDF considers similar tokens between compared documents (attribute values), not just the exact tokens as in TF-IDF. Comparison studies have shown that SoftTFIDF performs best for named entity matching [5].

In the second level of similarity computations, we matched the values of one tuple against values of the other tuple using a greedy matching. An illustrative example of this greedy matching is shown in Figure 4. In this greedy matching, we start matching from the first attribute value in the first tuple and match it with the most similar attribute value in the second tuple. Subsequently, the second attribute value of the first tuple is matched with the most similar unmatched attribute value of the second tuple and so on. Intuitively, attribute values occurring less frequently are more indicative of semantic similarity between the tuples. To account for this, we weighted attribute value similarities against the mean inverse document frequencies of terms.

After computing the similarity values of tuples, the agreement between two result sets for every query is computed using a greedy matching. Please refer to Balakrishnan and Kambhampati [1] for further details of the agreement computations.

3.3 Calculating SourceRank

To facilitate the computation of SourceRank, we model the agreement between the sources as an agreement graph. An agreement graph is a directed weighted graph as shown in Figure 3. In the graph, the vertices are sources, and edge weights are adjusted agreements (agreement compensated for source collusion described below) between the sources. In addition to these agreement links, we add links of small weights between every pair of vertices for smoothing. These smoothing links can be seen as accounting for the unseen samples. Thus the overall weight of the link from S_1 to S_2 is,

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

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

R_{1q} and R_2q are the answer sets of S_1 and S_2 for the query q, $A(R_{1q}, R_{2q})$ is the agreement between these two answer sets, and Q is the set of sampling queries over which the agreement is computed. β is the smoothing factor and is set at 0.1.

In this agreement graph, any source that has a high degree of agreement with other relevant sources itself is likely to be a relevant and trustworthy source. This transitive propagation of source relevance (trustworthiness) through agreement links can be captured in terms of a fixed point computation [2]. In particular, if we view the agreement graph as a markov chain, with sources as the states, and the weights on agreement edges specifying the probabilities of transition from one state to another, then the asymptotic stationary visit probabilities of the markov random walk will correspond to a measure of the global relevance of that source. We call this measure the SourceRank. The graph is strongly connected and irreducible, assuring convergence of the random walk.

3.4 Collusion Estimation

We measure and compensate for source collusion while computing agreements. The collusion is computed on top-k answers sets, similar to the agreements. The basic intuition behind collusion detection is that if two sources return the same top-k answers to the queries with large number of possible answers (e.g. queries containing only stop words), they are likely to be colluding. More formally, for two ranked sets of answers, the expected agreement between top-k answers $E(A_k)$ is

$$E(A_k) = \begin{cases} \frac{k}{n}(1-e) & \text{if } k < n \\ (1-e) & \text{otherwise} \end{cases} \quad (3)$$

where k is the number of answers used to calculate agreement, n is the size of the answer set, and e is the error rate due to approximate matching. This means that for large answers sets (i.e. $n \gg k$) the expected agreement between two independent sources is very low.

To get queries with large answer sets, we fetched a set of two hundred keywords with the highest document frequencies from the crawl described in the Section 3.1. Sources are crawled with these queries. The agreement between the answer sets are computed based on this crawl. These agreements give a measure of the the collusion between the sources. The agreement computed between the same two sources is multiplied by $(1 - collusion)$ to get the adjusted agreement used for computing SourceRank above.

3.5 Online Query Processing

To process the queries, the top-k sources with the highest SourceRank are selected. We set the value of k at five for the online sources and 10% of the total number of sources

(a) (b)

Figure 5: (a) Comparing trustworthiness of result of SourceRank and baseline methods. The corrupted results are marked as red against the ground truth. (b) A sample book query results in Factal comparison shopping integration system

for the google base. Queries are dispatched to these sources in parallel spawning a separate thread for each source. Top-5 results are fetched from each source, and the results are combined and presented to the user.

4. RESULTS AND DEMONSTRATION

Our relevance evaluations (reported in Balakrishnan and Kambhampati [1]) show that SourceRank improves precision and DCG (discounted cumulative gain) by 22-60% over the the Google Base and the existing methods. Further, existing source selection measures are agnostic to the corruption, whereas the SourceRank of sources reduces almost linear to the corruption levels. For details of experiments and results please refer to Balakrishnan and Kamhampati [1]. In this demonstration, we focus on three scenarios (i) trustworthiness of results (ii) relevance of results and (iii) SourceRank based comparison shopping.

4.1 Trust Comparison

This scenario is a visually compelling illustration of trustworthiness of results. We set up our databases using tuples crawled from Google Base, and corrupt them to varying degrees. Subsequently we compute SourceRank, Coverage and CORI ranks for each of the databases, and compare the search results from each method. The screenshot in Figure 5(a) shows the layout of the results presented. The corrupted tuples are marked with red background, for an easy interpretation. The left pane shows the results from SourceRank and right pane shows the results from CORI or Coverage—as selected by the user.

4.2 Relevance Comparison

Based on direct searches to Google Base, we illustrate the improvement over Google Base ranking and Coverage. The screen looks similar to that of trust comparison shown in Figure 5(a). Though users may try any query, we provide a predefined set of queries with relevant results computed in advance for ground truth. The relevance of each result set is displayed, and irrelevant results are marked red. To emphasize the contrast in relevance between the results of methods compared, we eliminate the common tuples between two methods on display.

4.3 Comparison Shopping

As a practical application, we present a comparison shopping engine prototype powered by SourceRank. The prototype searches on forty online databases in movie and book domains. The results are returned to the user, and upon clicking, the user is redirected to corresponding online database to continue shopping. Unlike most existing warehousing approaches, Factal is based on a web integration approach with direct searchers in selected sources at query time. We believe that this scenario illustrates fast response time, ranking quality, and feasibility of the approach. The system—*Factal*—may be accessed at `http://factal.eas.asu.edu`.

5. REFERENCES

[1] R. Balakrishnan and S. Kambhampati. Sourcerank: Relevance and trust assessment for deep web sources based on inter-source agreement. In *Proceedings of WWW*, 2011.

[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] W. Cohen, P. Ravikumar, and S. Fienberg. A comparison of string distance metrics for name-matching tasks. In *IIWeb Workshop*, 2003.

[6] X. Dong, L. Berti-Equille, and D. Srivastava. Integrating conflicting data: the role of source dependence. In *PVLDB*, 2009.

[7] J. Madhavan, A. Halevy, S. Cohen, X. Dong, S. Jeffery, D. Ko, and C. Yu. Structured Data Meets the Web: A Few Observations. *Data Engineering*, 31(4), 2006.

[8] Z. Nie and S. Kambhampati. A Frequency-based Approach for Mining Coverage Statistics in Data Integration. *Proceedings of ICDE*, page 387, 2004.

[9] X. Yin, J. Han, and P. Yu. Truth discovery with multiple conflicting information providers on the web. *TKDE*, 2008.