Agreement Based Source Selection for the Multi-Topic Deep Web Integration

Manishkumar Jha #[1],Raju Balakrishnan #[2], Subbarao Kambhampati #[3]

#Computer Science and Engineering, Arizona State University
Tempe AZ USA 85287
{[1]mjha1,[2]rajub,[3]rao}@asu.edu

Abstract

One immediate challenge in searching the deep web databases is *source selection*—i.e. selecting the most relevant web databases for answering a given query. For open collections like the deep web, the source selection must be sensitive to trustworthiness and importance of sources. Recent advances solve these problems for a single topic deep web search adapting an agreement based approach (*c.f.* SourceRank [10]). In this paper we introduce a source selection method sensitive to trust and importance for *multi topic* deep web search. We compute multiple quality scores of a source tailored to different topics, based on the topic specific crawl data. At the query time, we classify the query to determine its probability of membership in different topics. These fractional memberships are used as the weights to the topic specific quality scores of sources to select sources for the query. Extensive experiments on more than a thousand sources in multiple topics show 18-85% improvements in result quality over Google Product Search and other existing methods[1].

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 a welter of web-accessible relational databases. By some estimates, the data contained in this collection—popularly referred to as the deep web—is estimated to be in tens of millions spanning across numerous topics [28]. Searching the deep web has been identified as the next big challenge in information management [32]. The most promising approach that has emerged for searching and exploiting the sources on the deep web is data integration. A critical advantage of integration to surface web search is that the integration system (mediator) can leverage the semantics implied in the structure of deep web tuples. Realizing this approach however poses several fundamental challenges, the most immediate of which is that of *source selection.* Briefly, given a query, the source selection problem involves selecting the best subset of sources for answering the query.

Recent advancements in deep web source selection—specifically SourceRank [10, 8]—consider the trustworthiness and relevance of sources. A straightforward idea for extending SourceRank for multi-topic deep web search is a weighted combination with query similarity, like PageRank [14]. But in general, agreement by sources in the same topic is likely to be much more indicative of importance of a source than endorsement by out of topic sources. Moreover, sources might have data corresponding to multiple topics, with the importance of the data varying across those topics. For example, Barnes & Noble might be quite good as a book source but might not be as good as a movie source (even though it has information about both topics). These problems are noted for surface web (e.g. Haveliwala [24]), but is more critical for the deep web since sources are even

17th *International Conference on Management of Data*
COMAD 2011, Bangalore, India, December 19–21, 2011
©*Computer Society of India, 2011*

[1] This research is supported by ONR grant N000140910032 and two Google research awards

more likely to cross topics than single web pages. To account for this fact, we extend the deep web source selection by assessing a topic-specific quality metric for the sources and assessing the resulting improvement in search quality.

To give a brief background of the problem, source selection received some attention in the context of text and relational databases even prior to advent of deep web specific source selection [29, 11, 17, 30, 25]. But these classic approaches are focused on assessing relevance of a source based on local measures of similarity between the query and the answers expected from the source. These local approaches are agnostic to the trustworthiness and importance of the sources [9]. A global measure of trust and importance is particularly important for uncontrolled collections like deep web, since sources try to artificially boost their rankings. SourceRank is a method of assessing trustworthiness and importance of sources based on the agreement between the sources. Please refer to Section 3 for details of SourceRank computation. Though SoureRank was found to be effective for trust and importance assessment, the best way to adapt the method to a multi-topic deep web environment remains an open problem.

To adapt the SourceRank for multiple-topics, we assess the quality of a source predominantly based on the endorsement by sources in the same topic. For this, we use different sampling query sets for different topics. The quality score of the source for a topic solely depends on the answers to the queries in that topic. To rank the sources while the user enters a query, a classifier is used to determine the topic of the query. The classifier gives the probability with which the query may belong to different topics. These probabilities are used to weight the topic-specific quality scores of the sources to compute a single source quality score. These combined scores are used to rank the sources. In contrast to the SourceRank, these rankings are specific to the query topic. This work may be considered as a parallel to the important work of topic sensitive PageRank by Haveliwala for the surface web [24].

For empirical evaluation, we compare the precision of the topic sensitive SourceRank to: (i) Google Product Search (ii) CORI [16] used for text databases (ii) Topic-Oblivious universal SourceRank and (iv) a combination of oracular source-classification and SourceRank. Experiments are performed on a multi-topic environment comprising of hundreds of databases from four popular topics: movies, books, camera and music. Our precision values show 18-85% of improvement over all the baselines with statistical confidence of 0.95 or more. These evaluations establish the need for a topic sensitive source selection for integrating multi-topic environments like the deep web.

Rest of the paper is organized as the following. Next section reviews the related work. We give background of the SourceRank computation in Section 3. Section 4 provides a detailed description of the proposed topic sensitive SourceRank computation. Further we discuss the classification of a query to different topics in Section 5. Experimental results are discussed in Section 7. Finally we present our conclusions derived from the experimental evaluations.

2 Related Work

Balakrishnan and Kambhampati [10, 8, 9] present a method for selecting deep web sources considering trust and relevance—namely SourceRank. Like the original Pagerank [14], SourceRank work was done with the assumption that importance will be measured in a topic-independent way, and that the topic sensitivity can be handled through the similarity metric (which is linearly combined with importance). Haveliwala's work [24] showed that topic-sensitive page importance measures can be more effective. We started our work with the belief that the topic-sensitivity is even more important for sources than it is to pages, as sources are even more likely than pages to cross topics.

Current relational database selection methods minimize cost by retrieving maximum number of distinct records from minimum number of sources [29]. The parameter widely considered for this minimum cost access is the coverage of sources. Coverage of a database is a measure of number of relevant tuples to the query. Hence the cost based web database selection is formulated as selecting the least number of databases maximizing sum of coverages. Related problem of collecting source statistics [29, 25] has been researched. The problem of ranking database tuples for key word search is addressed [12]. Note that database selection reduces number of sources accessed before ranking tuples and is indispensable for the huge size of the deep web [28].

Considering research in the text databases selection, Callan *et al.* [17] formulated CORI method for query specific selection based on relevance. Cooperative and non-cooperative text database sampling [15, 25] and selection considering coverage and overlap to minimize cost [31, 30] are addressed by a number of papers.

Combining multiple retrieval methods for text documents has been used for improved accuracy [18]. In his early work of combining multiple retrieval methods to improve the retrieval accuracy for text documents, Lee [27] observes that the different methods are likely to agree on the same relevant documents than on irrelevant documents. This observation rhymes with our argument in Section 3 in giving a basis for agreement-based relevance assessment. For the surface web, Gyöngyi *et al.* [23] proposed trust rank, and extension of page rank considering trustworthiness of sources of hyperlinks. Agrawal *et al.* [6] explored ranking database search records by comparing to corresponding web search results.

A probabilistic framework for trust assessment based on agreement of web pages for question answering has been presented by Yin *et al.* [33]. 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.* [20, 19] extended this model considering source dependence using the same basic model as Yin *et al.* The collusion detection specific to the deep web is discussed by Balakrishnan and Kambhampati [10]. Gupta and Han [22] give an overview of network based trust analysis methods.

3 Deep Web Source Selection Background

In this section we illustrate why considering trustworthiness and importance of sources is critical for the deep web integration. Subsequently, we formalize the argument that the relevance and trustworthiness of a database manifests as the agreement of its results with those from other databases. We also explain the 2-step SourceRank calculation process: (i) create a source graph based on agreement between the sources (ii) assessing the source reputation based on this source graph.

3.1 Considering Trust and Importance

As we discussed in the Introduction, classical approaches to source selection in relational and text databases are purely local. In the context of deep web, such a source-local approach has important deficiencies:

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

2. The source selection is agnostic to the trustworthiness of the answers. Trustworthiness is a measure of correctness of the answer (keep in mind that relevance assess whether tuples is answering the query, not the correctness of the information). For example, for the query *The Godfather* many databases in Google Base return copies of the book with unrealistically low prices to attract the user attention. When the user proceeds towards the checkout, these low priced items would turn out to be out of stock, and many times a different item with the same title and cover (e.g. solution manual of the text book).

A global measure of trust and importance is particularly important for uncontrolled collections like deep web, since sources try to artificially boost their rankings. A global relevance measure should consider popularity of a result, since the popular results tends to be relevant. Moreover, it is impossible to measure trustworthiness of sources based on local measures; since the measure of trustworthiness of a source should not depend on any information the source provides about itself. In general, the trustworthiness of a particular source has to be evaluated in terms of the endorsement of the source by other sources.

Given that the source selection challenges are similar in a way to “page” selection challenges on the web, an initial idea is to adapt a hyper-link based method like PageRank [14] or authorities and hubs [26] from the surface web. However, the hyper-link based endorsement is not directly applicable to the web databases since there are no explicit links across source records. To overcome this problem, we create an implicit endorsement structure between the sources based on the *agreement* between the results returned by the sources. Two sources agree with each other if they return the same records in answer to the same query. It is easy to see that this agreement based analysis will solve the result importance and source trust problems mentioned above. Result importance is handled by the fact that the important results are likely to be returned by a large number of sources. For example, the classic *Godfather* movie is returned by hundreds of sources while the *Little Godfather* is returned by less than ten sources on a Google Products search [1]. A global relevance assessment based on the agreement of the results would thus have ranked the classic Godfather high. Similarly regrading trust, the corruption of results can be captured by an agreement based method, since other legitimate sources answering the same query are likely to disagree with the incorrect result attribute (e.g. disagree with unrealistically low price of the book result).

3.2 Agreement as Endorsement

In this section we show that the result set agreement is an implicit form of endorsement. Let R_T be the set of relevant and trustworthy tuples for a query, and U be the search space (the universal set of tuples searched). Let r_1 and r_2 be two tuples independently picked by two sources from R_T (i.e. they are relevant and trustworthy), and $P_A(r_1, r_2)$ be the probability of agreement of the tuples (we may think of "agreement" of tuples in terms of high degree of similarity; please refer to Balakrishnan and Kambhampati [10] for details of agreement computation)

$$P_A(r_1, r_2) = \frac{1}{|R_T|} \tag{1}$$

Similarly let f_1 and f_2 be two irrelevant (or untrustworthy) tuples picked by two sources and $P_A(f_1, f_2)$ be the agreement probability of these two tuples. Since f_1 and f_2 are from $U - R_T$

$$P_A(f_1, f_2) = \frac{1}{|U - R_T|} \tag{2}$$

For any web database search, the search space is much larger than the set of relevant tuples, i.e. $|U| \gg |R_T|$. Applying this in Equation 1 and 2 implies

$$P_A(r_1, r_2) \gg P_A(f_1, f_2) \tag{3}$$

For example, assume that the user issues the query *Godfather* for the Godfather movie trilogy. Three movies in the trilogy *The Godfather I*, *II* and *III* are the results relevant to the user. Let us assume that the total number of movies searched by all the databases (search space U) is 10^4. In this case $P_A(r_1, r_2) = \frac{1}{3}$ and $P_A(f_1, f_2) = \frac{1}{10^4}$ (strictly speaking $\frac{1}{10^4-3}$). Similarly the probability of three sources agreeing are $\frac{1}{9}$ and $\frac{1}{10^8}$ for relevant and irrelevant results respectively. It is easy to extend this argument to set of results, rather than a single result.

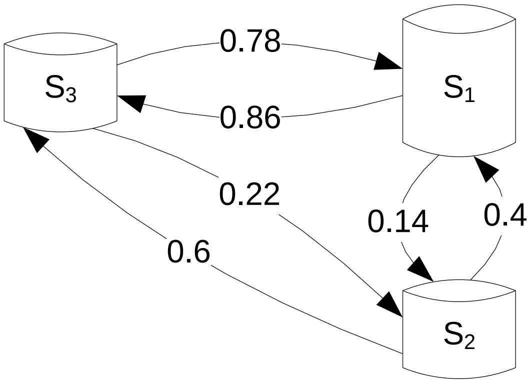


Figure 1: A sample agreement graph structure of three sources. The weight of the edge from S_i to S_j is computed by Equation 5

Though the explanation above assumes independent sources, it holds even for partially dependent sources. However, the ratio of two probabilities (i.e. the ratio of probability in Equation 1 to Equation 2) will be smaller than that for the independent sources.

3.3 Creating The Agreement Graph

To facilitate the computation of SourceRank, we represent the agreement between the source result sets as an agreement graph. Agreement graph is a directed weighted graph as shown in example Figure 1. In this graph, the vertices represent the sources, and weighted edges represent the agreement between the sources. The edge weights correspond to the normalized agreement values between the sources. For example, let R_1 and R_2 be the result sets of the source S_1 and S_2 respectively. Let $a = A(R_1, R_2)$ be the agreement between the results sets. In the agreement graph we create two edges: one from S_1 to S_2 with weight equal to $\frac{a}{|R_2|}$; and one from S_2 to S_1 with weight equal to $\frac{a}{|R_1|}$. The semantics of the weighted link from S_1 to S_2 is that S_1 endorses S_2, where the fraction of tuples endorsed in S_2 is equal to the weight. Since the inter-source weights are equal to the fraction of tuples, rather than the absolute number, they are asymmetric.

As described in Balakrishnan and Kambhampati [10], the agreement weights are estimated based on the results to a set of sample queries To account for the "sampling bias" in addition to the agreement links described above, we also add "*smoothing links*" with small weights between every pair of vertices.

Adding this smoothing probability, the overall weight $w(S_1 \rightarrow S_2)$ of the link from S_1 to S_2 is:

$$A_Q(S_1, S_2) = \sum_{q \in Q} \frac{A(R_{1q}, R_{2q})}{|R_{2q}|} \tag{4}$$

$$w(S_1 \rightarrow S_2) = \beta + (1 - \beta) \times \frac{A_Q(S_1, S_2)}{|Q|} \tag{5}$$

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. We set β at empirically determined value of 0.1 for our experiments (before normalization). Empirical studies like Gleich *et al.* [21] may help more accurate estimation. These smoothing links strongly connect agreement graph (we shall see that strong connectivity is important for the convergence of SourceRank calculation). Finally we normalize the weights of out links from every vertex by dividing the edge weights by sum of the out edge weights from the vertex. This normalization would make the edge weights equal to the transition probabilities for the random walk computations.

3.4 Calculating Scores

Let us start by considering certain reasonable desiderata that measures of reputation defined with respect to the agreement graph must satisfy:

1. Nodes with high in-degree should get higher rank—since high in-degree sources are agreed upon by large number of sources, they are likely to be more trustworthy and relevant.

2. Endorsed (agreed) by a source with a high in-degree should be more respected than endorsed by a source having smaller in-degree. Since a highly endorsed source is likely to be more relevant and trustworthy, the source endorsed by a highly endorsed source is also likely to be of high quality.

The agreement graph described above provides important guidance in selecting relevant and trustworthy sources. Any source that has a high degree of agreement by other relevant sources is itself 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 [14]. In particular, if we view the agreement graph as 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. This stationary visit probability gives the quality score of the source.

The markov random walk based ranking does satisfy the two desiderata described above. The graph is strongly connected and irreducible, hence the random walk is guaranteed to converge to the unique stationary visit probabilities for every node. This stationary visit probability of a a node is used as the SourceRank of that source.

4 Topic Sensitive Source Selection

Having described the details of computing agreement based source quality assessment, we describe the specifics of extending the source selection for multiple topics. Next section describes the multi-topic query based sampling. Subsequently we enumerate details of computing topic-sensitive source scores—namely TSR.

4.1 Sampling Sources

The deep-web sources are non-cooperative—i.e. they may not share the data statistics—and may allow only limited form based key word query access [15]. Hence we use a basic key word query based approach for sampling the sources. For generating the sampling queries are generated from publicly available online listings. We used two hundred titles or names in each topic as our sampling queries. Specifically, we randomly selected cameras from pbase.com [4], books from New York Times best sellers [2], movies from dmoz.org [3] and music albums from Wikipedia's top-100, 1986-2010 [5].

From the titles in these listings, words are deleted with 0.5 probability to get the partial key word queries (since partial queries gives better results [10]). All these queries are sent to every source and *top-k* (we used $k = 5$) answers returned are collected. Note that the sources are not explicitly classified into topics. The idea is that if a source gives high quality answers for a topic, the other in topic sources are likely to agree with that source. For online web databases, we need to have wrappers to extract structured tuples from these returned answers. Automatic wrapper generation has been considered separately by related research [7], and is not discussed in this paper. After these structured

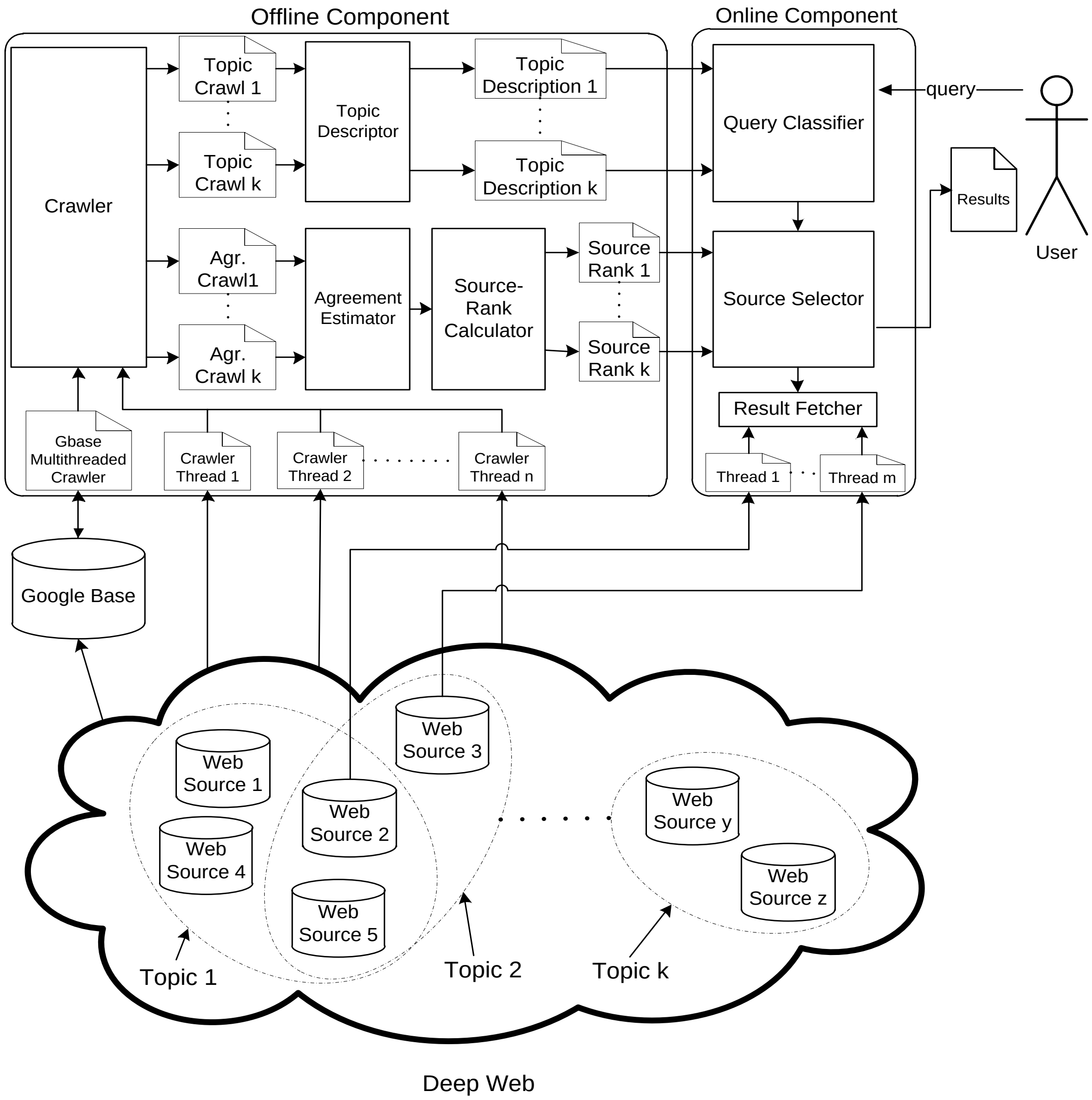


Figure 2: Multi-topics deep web integration system combining online query classification and TSR based source selection.

tuples are extracted, we computed the agreement between the sources as described above in Section 3.3.

4.2 Computing Topic Sensitive Ranking

For the Topic-sensitive SourceRank (TSR), we compute different quality scores of each source for different topics. We compute the source quality score for a topic based solely on the source crawls corresponding to the sampling queries of the topic. For example, for computing movie TSRs, we compute the agreement graph (described in Section 3.3) based on the crawl obtained by using the movie sampling queries described above in Section 4.1. After generating the agreement graph, source quality score for this topic are computed based on the static visit probability of a weighted Markov random walk on the graph as described in Section 3.4.

The acceptability of computation timings of TSR is directly inferable from the related work. The first step of computing TSR—computing the agreement graph—is shown to be scalable by Balakrishnan and Kambhampati [10]. The random walk computation is widely used including for PageRank [13] and known to be scalable. Besides, note that the TSR computation is offline, and does not add to the valuable query time. Due to the above reasons, we do not perform separate timing experiments in this paper.

Depending on the target topic of the query, we need to use the right topic TSRs to select the best sources. For example, we need to select sources ranking higher in the movie TSR for a movie query. Realistically, the membership of a query in a topic will be probabilistic. The section below describes combining topic TSRs depending on the probability of membership of the query in different topics.

5 Query Processing

The next set of computations is performed at query time. The first task is to identify the query-topic i.e. the likelihood of the query belonging to representative topic-classes. We treat this as a soft-classification problem. For a user query q and a set of representative topic-classes $c_i \in C$, the goal is to find the probability of topic membership of q in each of these topics c_i. A Naïve Bayes Classifier (NBC) is used for this topical query classification. We describe our training data, classification approach, and the final source selection in the sections below.

5.1 Training data

In order to accurately identify query-topic, we need training data representative of the topic-classes. We use query-based sampling techniques for obtaining topic-descriptions, similar to the sampling described in the Section 4.1. We used the same set of sampling methods and list of queries described in Section 4.1. But instead of generating partial queries by deleting words randomly, full titles or names are used as queries.

5.2 Classifier

The classifier tries to identify the query-topic using query-terms and training data consisting of topic-descriptions discussed above. In our implementation we use a multinomial NBC, with maximum likelihood estimates to determine the topic probabilities of the query. For a query q, we compute the probability of membership of the query for different topic-classes as,

$$P(c_i|q) = \frac{P(q|c_i) \times P(c_i)}{P(q)} \propto P(c_i) \prod_j P(q_j|c_i) \qquad (6)$$

where q_j is the j^{th} term of user query q.

$P(c_i)$ can be set based on topic knowledge, but we assume uniform probabilities for topic-classes. Hence the above equation reduces to,

$$P(c_i|q) = \prod_j P(q_j|c_i) \qquad (7)$$

$P(q_j|c_i)$ is computed as the ratio of number of occurrences of q_j in the topic file corresponding to c_j to the total number of words in the file. If q_j is absent in the topic file we use a smoothing value of one as the count of q_j.

After computing the topic probabilities of the query, we compute the composite source rank (CSR) scores of sources based on the topical probabilities of the query. For a source s_k, CSR_k is given by

$$CSR_k = \sum_i P(c_i|q) \times TSR_{ki} \qquad (8)$$

where TSR_{ki} is the topic-sensitive SourceRank score of source s_k for topic-class c_i. CSRs give the query-topic sensitive SourceRank for all deep-web sources.

Since CSR is computed during query-time, it is important that its processing time is kept to a minimal. CSR will be used in conjunction with a relevance measure as described below. Hence CSR computation can be limited to selected *top-k* most relevant topics.

5.3 Combining Query Similarity and CSR

The CSR computed above is combined with the query similarity based relevance (we describe the details of the measure in Section 7) to get the final ranking score of the source. For a source s_k the final score combining query-similarity and the CSR is given by,

$$OverallScore_k = \alpha \times R_k + (1 - \alpha) \times CSR_k \qquad (9)$$

where R_k is the probability of relevance of the source based on the query similarity and α is the relative weight given to the query similarity. We try different values of α for our empirical evaluations.

6 System Architecture

Figure 2 provides an overview of our system. It consists of two main parts. An offline component which uses the crawled data for computing topic-sensitive SourceRanks and topic-descriptions. The online component consists of a classifier which performs user query-classification using the topic-descriptions. The source selector uses the query-classification information to combine TSRs in order to generate query specific ranking of sources.

7 Experimental Evaluation

We evaluated the effectiveness of our approach and compared it with other source selection methods. The experiments are performed on a multi-topic deep-web environment with more than thousand sources in four representative topic classes - camera, book, movie and music.

7.1 Source Data Set

For our experiments, deep-web source data was collected via Google Base. Google Base is a central repository where merchants can upload their databases thereby publishing the databases over the web. Google Product Search works over the Google Base. Google Base provides API-based access to data, returning ranked results. Google Base's Search API for shopping allows querying of data uploaded to Google Base. Each deep-web source in Google Base is associated with a source-identifier (SID). For selecting sources for our multi-topic deep-web environment, we probed Google Base with a set of 40 queries. These 40 queries contained a mix of camera model names, book, movie and music album titles. From the first 200 results of each query, we collected the SIDs and considered them as a source belonging to our multi-topic deep web environment. We collected a total of 1440 deep web sources for our multi-topic environment: 276 camera, 556 book, 572 movie, and 281 music sources.

7.2 Test Queries

Test query set contained a mix of queries from all four topic-classes and represents the possible user queries. Test queries were selected such that there is no overlap with the sampling queries. The test queries were generated by randomly removing words from camera names, book, movie and music album titles with probability 0.5, similar to the sample queries described in Section 4.1. Number of test queries are varied for different topics to obtain the required (0.95) statistical significance.

7.3 Baseline Source Selection Methods

TSR is compared with the following agreement based and query similarity based source selection methods. The agreement based methods consider the source agreement, and hence the trustworthiness and relevance of the sources are taken into account. On the other hand, pure query similarity measures like CORI [17] assesses the source quality based on similarity of content with the user query; hence agnostic to the trust and importance. The CORI and the Undifferentiated SourceRank described below may be considered as the alternative approaches to multi-topic search derived from the existing methods.

7.3.1 Agreement Based Measures

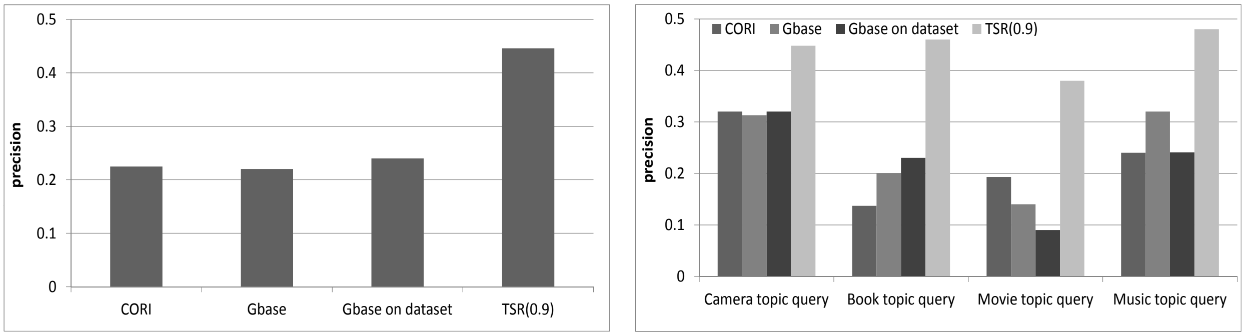
We describe the two baselines derived based on agreement based methods, which are directly derived from SourceRank [10]. We compare with these measures as standalone and in combination with query-relevance based measures.

Undifferentiated SourceRank, USR: USR does not differentiate between topics. We created a single agreement graph across the topics based on entire set of sampling queries (Section 4.1). The USR of sources are computed based on a random walk on this agreement graph. Combining USR with query similarity based source quality measures like CORI is an intuitive way of extending USR to multiple topics. Note that TSR for a topic was computed based only on the sampling queries in that topic. In other words, TSR predominantly considers endorsement of sources from the topic, where as USR considers endorsement from every sources equally. Comparing USR and TSR performances is an excellent validation of whether the topic specific endorsement improves result quality.

Oracular Source Selection, DSR: DSR assumes that a perfect classification of sources and queries are available. DSR is provided with the manually determined topic information of the sources and the test queries. DSR creates source graphs for a topic including only sources in that topic. Similarly for each test query, the sources ranking high in the topic corresponding to the test query is used. Note that TSR had to implicitly determine topic information of both the sources and the queries based on the answers to the sample queries and the NBC respectively. DSR is expected to do better than all other measures including TSR, since the topic information is available apriori. Comparison of TSR with DSR gives an idea of how well the automated topic classification of queries and sources of TSR is performing with respect to an oracular scenario.

7.3.2 Query Similarity Based Measures

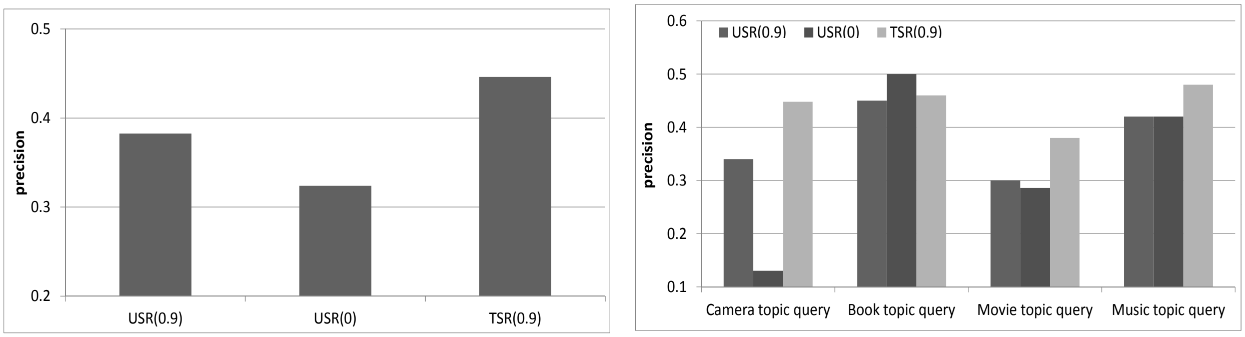
CORI: CORI is a query-based relevance measure. Source statistics for CORI were collected using highest document frequency terms from the sample crawl data. We used 800 high-tuple frequency terms as queries and



(a)

(b)

Figure 3: Comparison of *top-5* precision of TSR(0.9) with the query similarity based methods: CORI and Google Base (a) aggregation across the topics(b) topic-wise precision



(a)

(b)

Figure 4: Comparison of *top-5* precision of TSR(0.9) with the agreement based methods: USR(0), USR(0.9) (a) aggregation across the topics(b) topic-wise precision

used *top-10* results for each query to create resource descriptions for CORI. For selecting sources based on CORI, we used the same parameters as found optimal by Callan *et al.*[16].

Google Base: We compared TSR with Google Product Search results. We use two-versions of Google Base.[2] Gbase on dataset restricted to search only on our crawled sources, and stand alone Gbase in which Google Base search with no restriction i.e. considers all sources in Google Base.

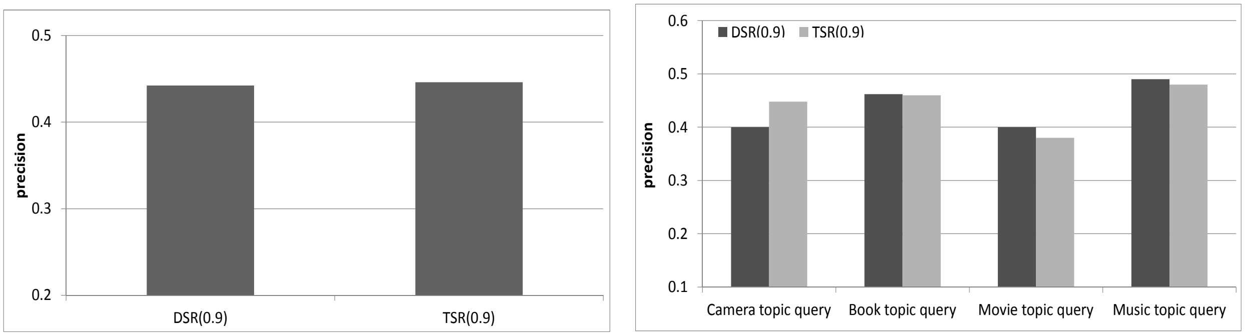
7.4 Result Merging and Ranking

Using our source selection strategies, we selected *top-k* sources for every test query and made Google Base query only on these *top-k* sources. We experimented with three different values of *k*—*top-10* sources, *top-5%* and *top-10%* sources—and found that best precision was obtained for *k=10.* We used Google Base's tuple ranking for ranking the resulting tuples and return *top-5* tuples in response to test queries. After ranking the tuples, the methods can be directly compared with each other.

7.5 Relevance Evaluation

For assessing the relevance, we used the test queries defined above. The queries were issued to *top-k* sources selected by different source selection methods. The *top-5* results returned were manually classified as relevant or irrelevant. The classification was rule based. For example, if the test query is “Pirates Caribbean Chest” and the original movie name is “Pirates of Caribbean and Dead Man's Chest” then if the result entity refers to the movie “Pirates of Caribbean and Dead Man's Chest” (DVD, Blue-Ray etc.) then the

[2] Google Product Search implements a search on Google Base, and provides API based access as well. Though the exact searching method of Google Base in unknown, we assume that Google Base predominantly fetch results based on query similarity based on our examination of Google Base results.



(a)

(b)

Figure 5: Comparison of *top-5* precision of TSR(0.9) with oracular DSR(0.9) (a) aggregation across the topics(b) topic-wise precision

result is classified as relevant and otherwise irrelevant. First author did the classification of the results. To avoid author bias, we merged results from different source selection methods in a single file so that the evaluator does not know which method each result came from while he does the classification.

7.6 Results

We compared TSR with the baselines described in Section 7.3. Instead of using stand-alone TSR, we combined TSR with query similarity based CORI measure. We experimented with different values of weighted combination of CORI and TSR, and found that $TSR \times 0.1 + CORI \times 0.9$ gives best precision. For rest of this section we denote this combination as $TSR(0.9)$. Note that the the higher weightage of CORI compared to TSR is to compensate for the fact that TSR scores have much higher dispersion compared to CORI scores, and not an indication of relative importance of these measures.

7.6.1 Comparison with Query Similarity Based Source Selection

Our first set of experiments compare precision of TSR(0.9) with query similarity based measures i.e. CORI and Google Base discussed above. The results are illustrated in Figure 3(a). Note that the improvement in precision for TSR is significant as the precision improves approximately 85% over all competitors, including Google Base. This considerable improvement in precision is not surprising in the light of prior research on agreement based source selection with query based measures [10].

A per topic-class analysis of test queries, Figure 3(b), reveals that TSR(0.9) significantly outperforms the relevance-based source selection models for all topic-classes. As a note on the seemingly low precision values, these are mean relevance of the top-5 results. Many of the queries used have less than five possible relevant answers (e.g. a book title query may have only paperback and hard cover for the book as relevant answers). But since we count the *top-5* results always, the mean precision is bound to be low. For example, if a method returns one relevant answer on in *top-5* for all queries, the *top-5* precision value will be only 20%. We get better values since some queries have more than one relevant results in *top-5* (e.g. Blu-Ray and DVD of a movie).

7.6.2 Comparison with Agreement Based Source Selection

We compare TSR(0.9) with the linear combination of USR and CORI. We used $USR \times 0.1 + CORI \times 0.9$ for these comparisons. Linear combination of USR with a query specific relevance is a highly intuitive way of extending a static SourceRank multi-topic deep web search. Note that the comparison of TSR and USR is isomorphic to the comparison of topic-sensitive PageRank [24], and PageRank [14] for the surface web.

The aggregated results across the topics are illustrated in Figure 4(a). TSR(0.9) precision exceeds USR(0.9) by 18% and USR(0) by 40%. Since the difference are small we evaluated the statistical significance of these results. We used sufficient number of queries to guarantee that TSR(0.9) out-performs both USR(0.9) and USR(0) (i.e. stand alone USR, not combining with CORI) with confidence levels of 0.95 or more.

Figure 4(b) provides per topic results. For three out of four topic-classes (Camera, Movies, and Music), TSR(0.9) out-performs USR(0.9) and USR(0) with confidence levels 0.95 or more. For books we found no statistical significant difference between USR(0.9) and TSR(0.9). This may be attributed to the fact that the source set was dominated by large number of good quality book sources, biasing the ranking towards book topic. Further, our analysis revealed that there are many multi-topic sources providing good quality results for books, movies and music topics (e.g. Amazon, eBay). These versatile sources occupy top positions in USR as well as USR(0.9) for these three topics. Consequently the topic independent USR performs comparable to topic specific USR(0.9) for these three topics: music, movies and books.

7.6.3 Comparison with Oracular Source selection

We compared TSR with oracular source selection, DSR described above in Section 7.3.1. We compared TSR(0.9) with DSR(0.9) (i.e. linear combination $0.1 \times DSR + 0.9 \times CORI$). As shown in Figures 5(a) and 5(b), TSR(0.9) is able to match DSR(0.9) performance for the test queries. The aggregate results across the topics is shown in Figure 5(a) and topic-wise result is shown in Figure 5(b). Result shows that the TSR precisions are quite comparable with that of DSR. This implies that TSR is highly effective in categorizing sources and queries, almost matching with oracular DSR. A note on the DSR's performance for camera-topic. After investigating our deep-web environment for camera-topic, we found that the source-rank for camera-topic was dominated by sources which answered less than 25% of sampling queries. This could be attributed to the fact that our source selection technique led to selection of relatively more number of cross-topic sources than pure sources for camera topic. As a result, selecting top-ranked camera-topic sources infact led to a drop in performance.

8 Conclusion

We investigated multi-topic source selection sensitive to trustworthiness and importance for the deep web. Although SourceRank is shown to be effective in solving this problem in single topic environments, there is a need for extending SourceRank to multiple-topics. We introduced topic-sensitive SourceRank (TSR) as an efficient and effective technique for evaluating source importance in a multi-topic deep web environment. We combined TSR source selection with a Naïve Bayes Classifier for queries to build our final multi-topic deep web search system. Our experiments on more than s thousand sources spanning across multiple topics shows that a TSR-based source selection is highly effective in extending SourceRank for multi-topic deep web search. TSR is able to significantly out-perform query similarity based retrieval selection models including Google Product Search by around 85% in precision. Comparison with other baseline agreement-based source selection models showed that using TSR results in statistically significant precision improvements over baseline methods; including a topic oblivious SourceRank combined with query similarity. Comparison with oracular DSR approach reveals effectiveness of TSR for topic-wise query and source classification and subsequent source selection.

References

[1] Goolge products. http://www.google.com/products.

[2] Ny times. http://www.nytimes.com/.

[3] Open directory project. http://www.dmoz.org.

[4] Pbase. http://www.pbase.com/.

[5] Wikipedia. http://www.wikipedia.org/.

[6] S. Agrawal, K. Chakrabarti, S. Chaudhuri, V. Ganti, A. Konig, and D. Xin. Exploiting web search engines to search structured databases. In *Proceedings of WWW*, pages 501–510. ACM, 2009.

[7] A. Arasu and H. Garcia-Molina. Extracting structured data from Web pages. In *Proceedings of SIGMOD.*

[8] R. Balakrishnan and S. Kambhampati. SourceRank: relevance and trust assessment for deep web sources based on inter-source agreement. In *Proceedings of WWW*, pages 1055–1056. ACM, 2010.

[9] R. Balakrishnan and S. Kambhampati. Factal: integrating deep web based on trust and relevance. In *Proceedings of the 20th international conference companion on World wide web*, pages 181–184. ACM, 2011.

[10] R. Balakrishnan and S. Kambhampati. Sourcerank: Relevance and trust assessment for deep web sources based on inter-source agreement. In *Proceedings of the 20th international conference on World wide web*, pages 227–236. ACM, 2011.

[11] M. Bender, S. Michel, P. Triantafillou, G. Weikum, and C. Zimmer. Improving collection selection with overlap awareness in P2P search engines. *SIGIR*, pages 67–74, 2005.

[12] G. Bhalotia, A. Hulgeri, C. Nakhe, S. Chakrabarti, and S. Sudarshan. Keyword searching and browsing in databases using BANKS. In *ICDE*, page 0431, 2002.

[13] S. Brin and L. Page. The anatomy of a large scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30, April 1998.

[14] 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.

[15] J. Callan and M. Connell. Query-based sampling of text databases. *ACM TOIS*, 19(2):97–130, 2001.

[16] J. Callan, Z. Lu, and B. Croft. Searching distributed collections with inference networks. *SIGIR '95 Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, 1995.

[17] 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.

[18] W. Croft. Combining approaches to information retrieval. *Advances in information retrieval*, 7:1–36, 2000.

[19] X. Dong, L. Berti-Equille, Y. Hu, and D. Srivastava. Global detection of complex copying relationships between sources. *Proceedings of the VLDB Endowment*, 3(1), 2010.

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

[21] D. Gleich, P. Constantine, A. Flaxman, and A. Gunawardana. Tracking the random surfer: empirically measured teleportation parameters in PageRank. In *Proceedings of WWW*, 2010.

[22] M. Gupta and J. Han. Heterogeneous network-based trust analysis: a survey. *ACM SIGKDD Explorations Newsletter*, 13(1):54–71, 2011.

[23] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with trustrank. In *Proceedings of VLDB-Volume 30*, page 587, 2004.

[24] T. Haveliwala. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE Transactions on Knowledge and Data Engineering*, 15, 2003.

[25] P. Ipeirotis and L. Gravano. When one sample is not enough: improving text database selection using shrinkage. *SIGMOD*, pages 767–778, 2004.

[26] J. KLEINBERG. Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, 46(5):604–632, 1999.

[27] J. Lee. Analyses of multiple evidence combination. In *ACM SIGIR Forum*, volume 31, page 276. ACM, 1997.

[28] 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.

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

[30] M. Shokouhi and J. Zobel. Federated text retrieval from uncooperative overlapped collections. In *Proceedings of the ACM SIGIR*. ACM, 2007.

[31] L. Si and J. Callan. Relevant document distribution estimation method for resource selection. In *Proceedings of ACM SIGIR*, pages 298–305, 2003.

[32] A. Wright. Searching the deep web. *Commmunications of ACM*, 2008.

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