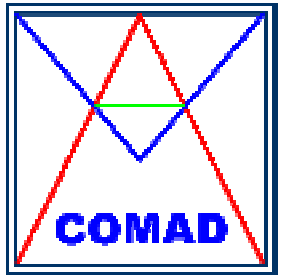


Topic-Sensitive SourceRank:

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

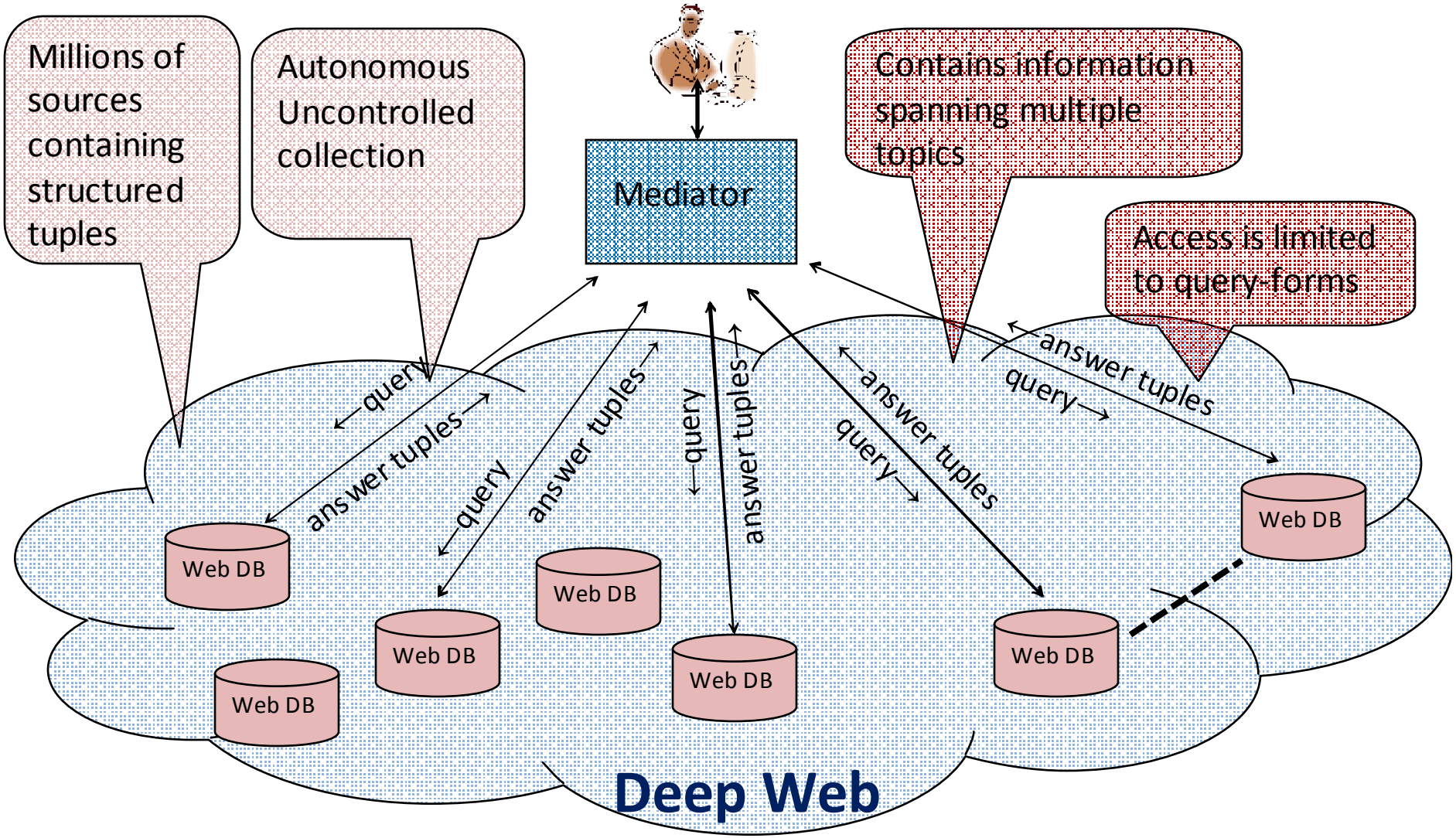
Manishkumar Jha

Raju Balakrishnan

Subbarao Kambhampati

ASU ARIZONA STATE UNIVERSITY

Deep Web Integration Scenario



Source quality and SourceRank

Source quality

- Deep-Web is
  - Uncontrolled
  - Uncurated
  - Adversarial
- Source quality is a major issue over deep-web

SourceRank

- SourceRank[1] provides a measure for assessing source quality based on source trustworthiness and result importance

[1] SourceRank:Relevance and Trust Assessment for Deep Web Sources Based on Inter-Source Agreement, WWW, 2011

Why Another Ranking?

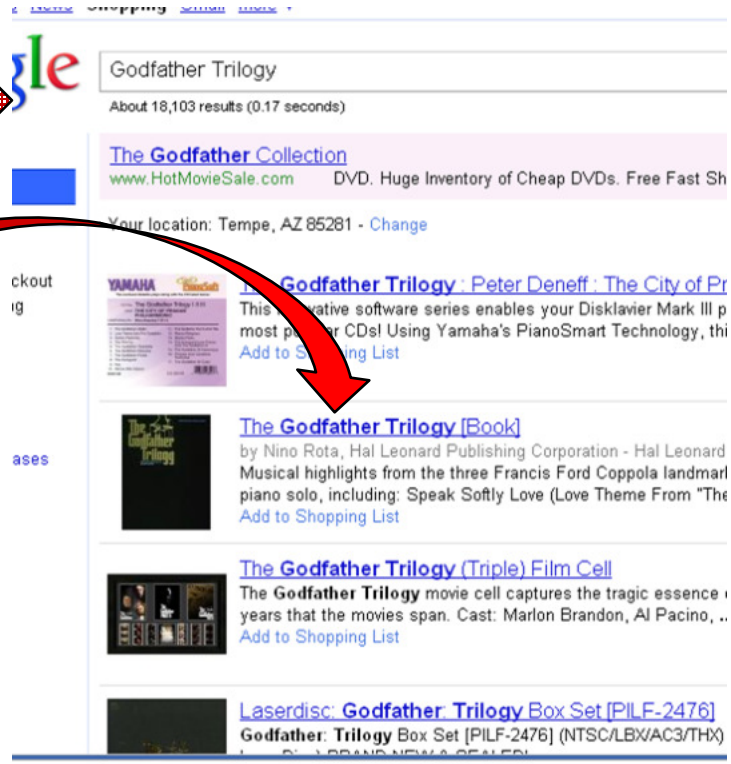
Rankings are oblivious to result Importance & Trustworthiness

Example Query: *"Godfather Trilogy" on Google Base*

Importance: Searching for titles matching with the query. None of the results are the classic Godfather

Trustworthiness (bait and switch)

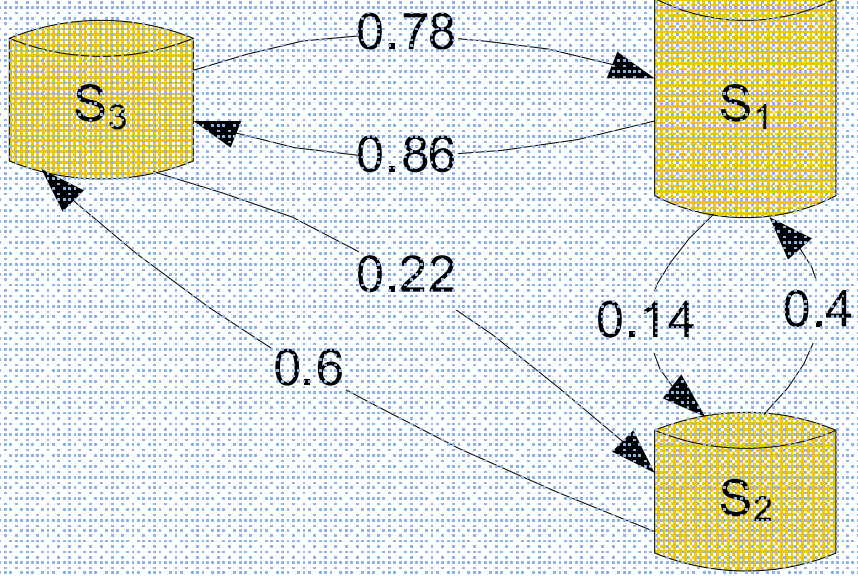
- The titles and cover image match exactly.
- Prices are low. Amazing deal!
- But when you proceed towards check out you realize that the product is a different one! (or when you open the mail package, if you are really unlucky)



SourceRank Computation

- Assesses source quality based on **trustworthiness** and **result importance**
- Introduces a *domain-agnostic agreement-based* technique for implicitly creating an endorsement structure between deep-web sources
- Agreement of answer sets returned in response to same queries manifests as a form of implicit endorsement

Method: Sampling based Agreement



Link semantics from S_i to S_j with weight w: S_j acknowledges w fraction of tuples in S_j. Since weight is the fraction, links are unsymmetrical.

$$W(S_1 \rightarrow S_2) = \beta + (1-\beta) \times \frac{A(R_1, R_2)}{|R_2|}$$

where β induces the smoothing links to account for the unseen samples. R_1, R_2 are the result sets of S_1, S_2 .

- Agreement is computed using key word queries.
- Partial titles of movies/books are used as queries.
- Mean agreement over all the queries are used as the final agreement.

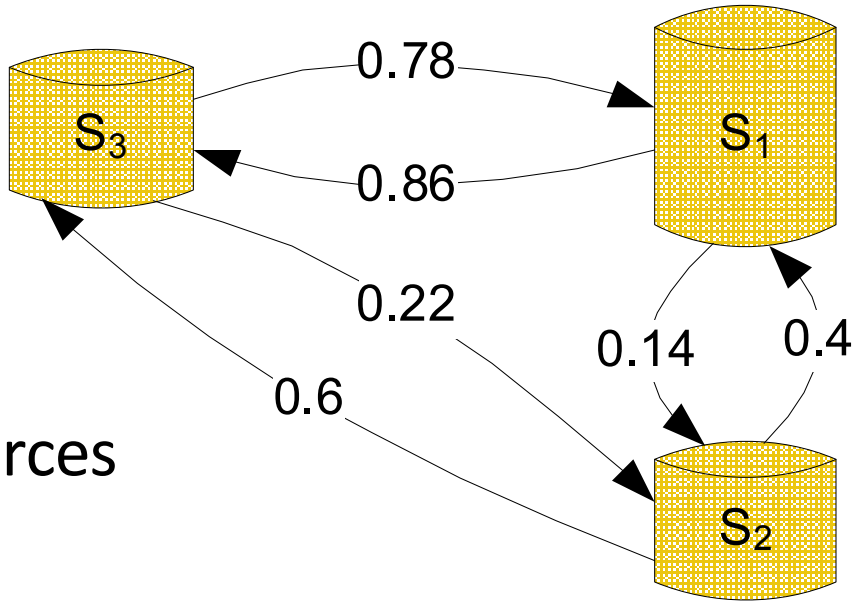
SourceRank defined as the Stationary Visit Probability on the Agreement Graph

SourceRank Computation contd.

Endorsement is modeled as directed weighted agreement graph

Nodes represent sources

Edge weights represent agreement between the sources

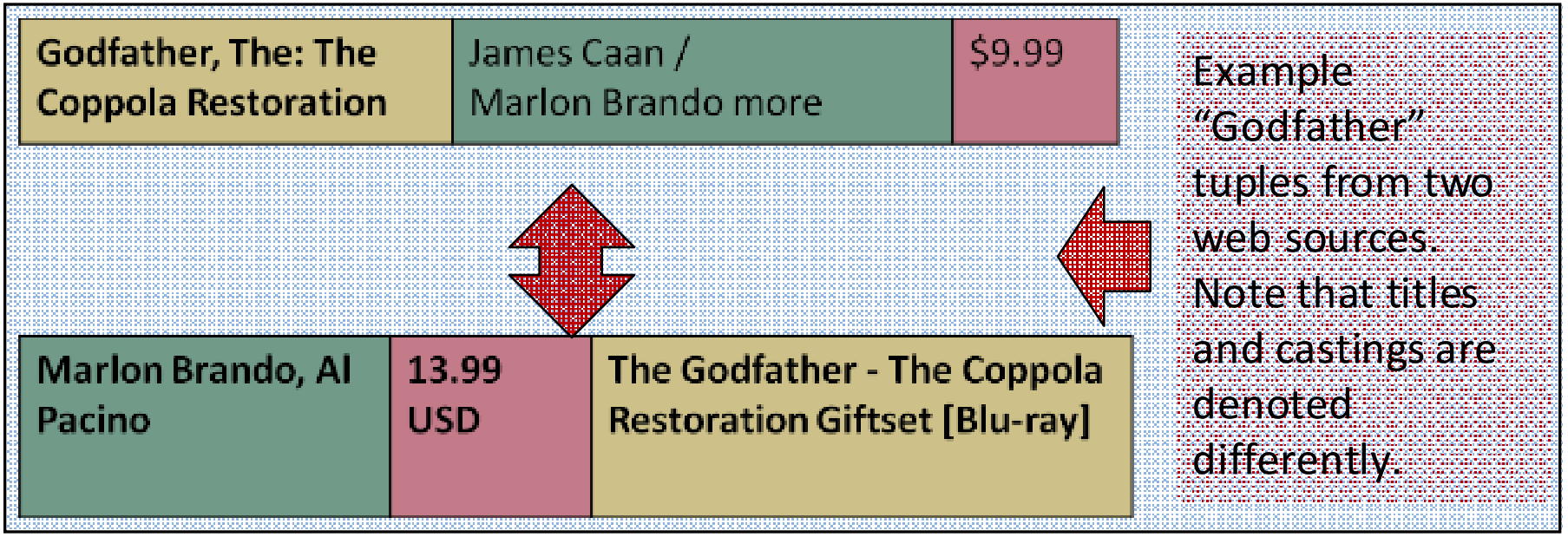


SourceRank of a source is computed as the stationary visit probability of a Markov random walk performed on this agreement graph

[WWW 2010 Best Poster; WWW 2011]

Computing Agreement is Hard

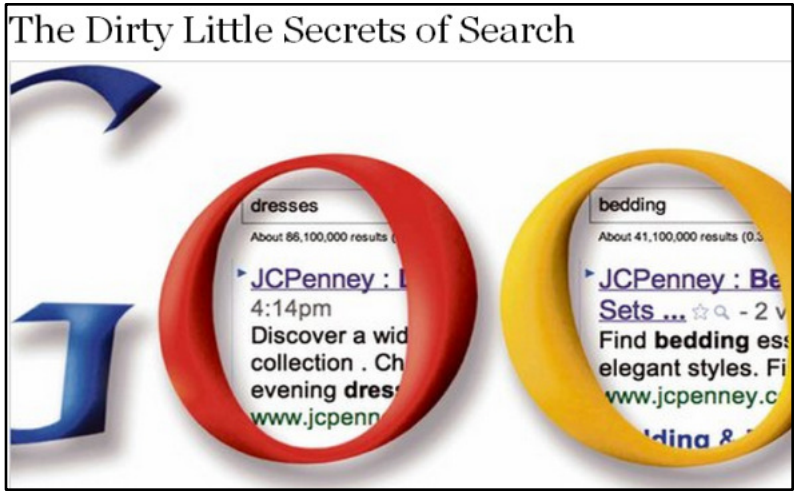
Computing semantic agreement between two records is the record linkage problem, and is known to be hard.



Semantically same entities may be represented syntactically differently by two databases (non-common domains).

Detecting Source Collusion

The sources may copy data from each other, or make mirrors, boosting SourceRank of the group.



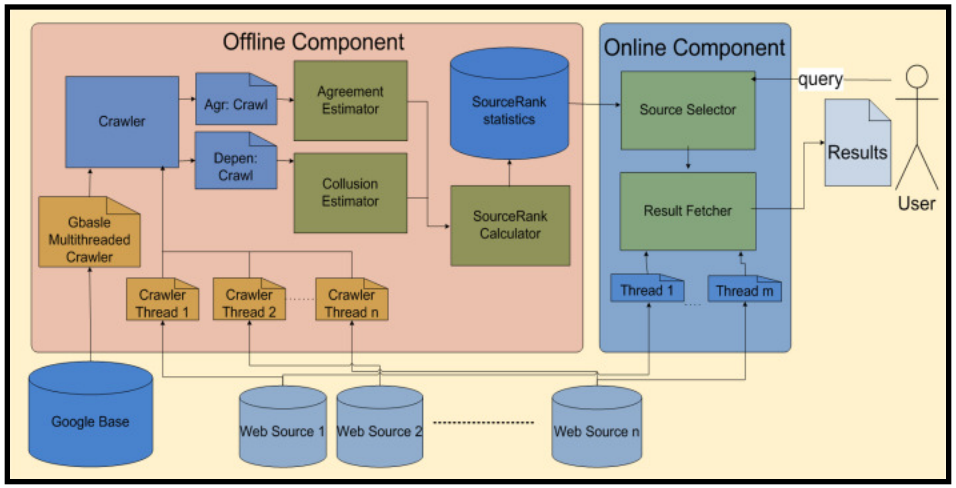
Observation 1: Even non-colluding sources in the same domain may contain same data.

e.g. Movie databases may contain all Hollywood movies.

Observation 2: Top-k answers of even non-colluding sources may be similar.

e.g. Answers to query "Godfather" may contain all the three movies in the Godfather trilogy.

Factal: Search based on SourceRank



"I personally ran a handful of test queries this way and got much better results [than Google Products] results using Factal" --- Anonymous WWW'11 Reviewer.

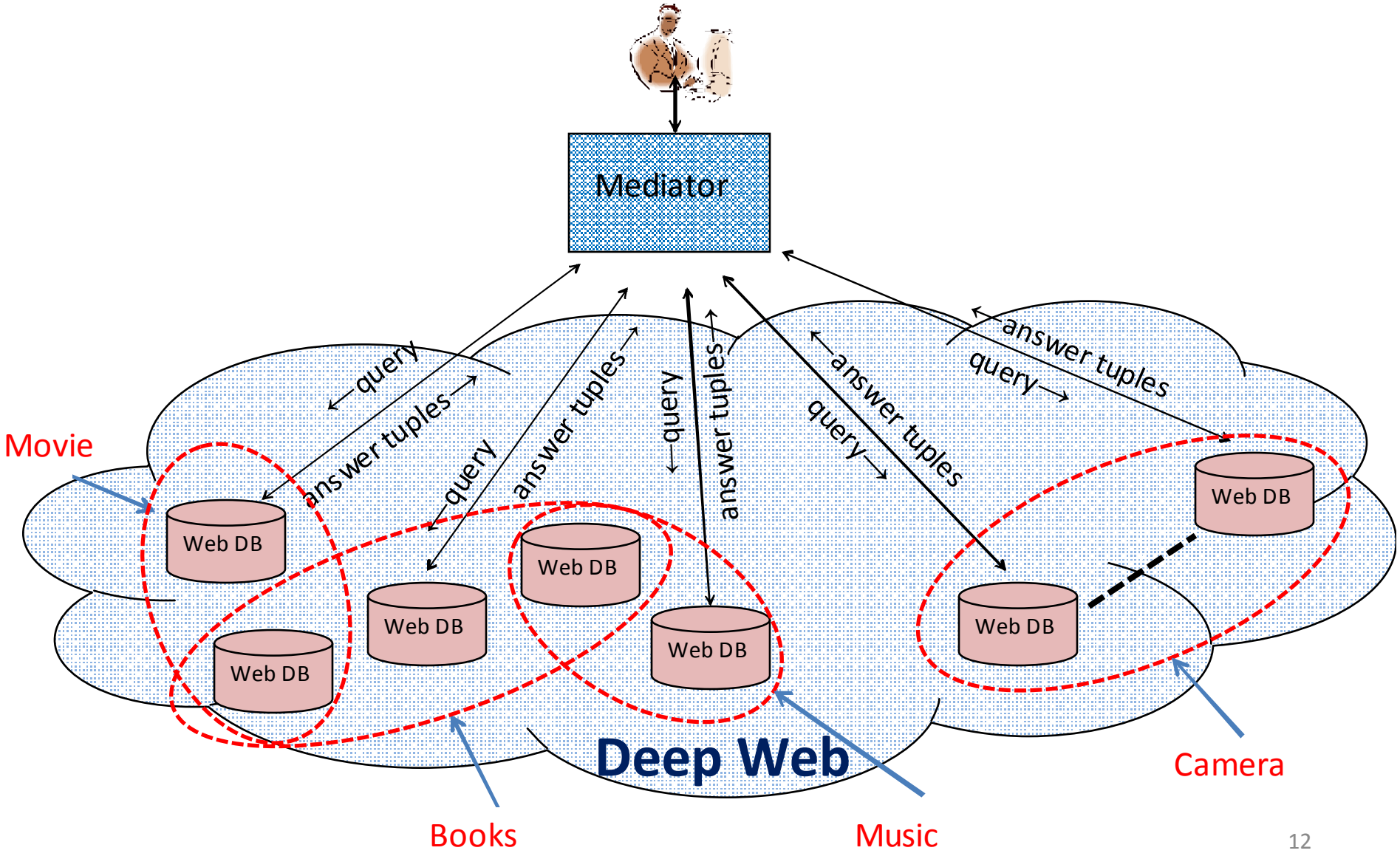
http://factal.eas.asu.edu

[WWW 2010 Best Poster; WWW 2011]

SourceRank is Query Independent

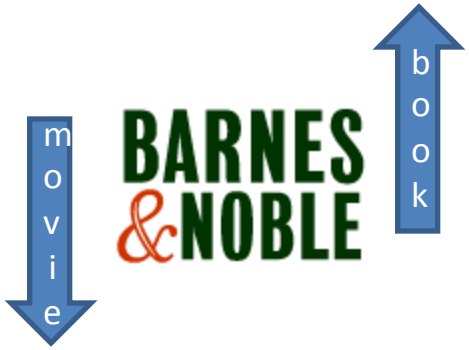
- SourceRank computes a single measure of importance for each source, independent of the query…

..But, Sources May Straddle Topics



… And Source quality is topic-sensitive

- Sources might have data corresponding to multiple topics. Importance may vary across topics



- SourceRank will *fail* to capture this fact

- Issues were noted for surface-web[2]. But are much more critical for deep-web as sources are even more likely to cross topics

[2] Topic-sensitive PageRank, WWW, 2002

SourceRank is Query Independent

- SourceRank computes a single measure of importance for each source, independent of the query…
  - Large source that has high quality for one topic will also be considered to be high quality for topic B
- Ideally, the importance of the source should depend on the query
  - But, too many queries..
  - ..and too costly to compute query specific quality at run time..

Problem Definition

Problem Definition: Performing effective multi-topic source selection sensitive to trustworthiness for deep-web

This Paper: Topic sensitive-SourceRank

- Compute multiple topic-sensitive SourceRanks
  - Source quality is a vector in the topic space
  - Query itself is a vector in the topic space
- At query-time, using query-topic, combine these rankings into composite importance ranking

- Challenges
  - Computing topic-sensitive SourceRanks
  - Identifying query-topic
  - Combining topic-sensitive SourceRanks

Agenda

- SourceRank
- Topic-sensitive SourceRank
- Experimental setup
- Results
- Conclusion

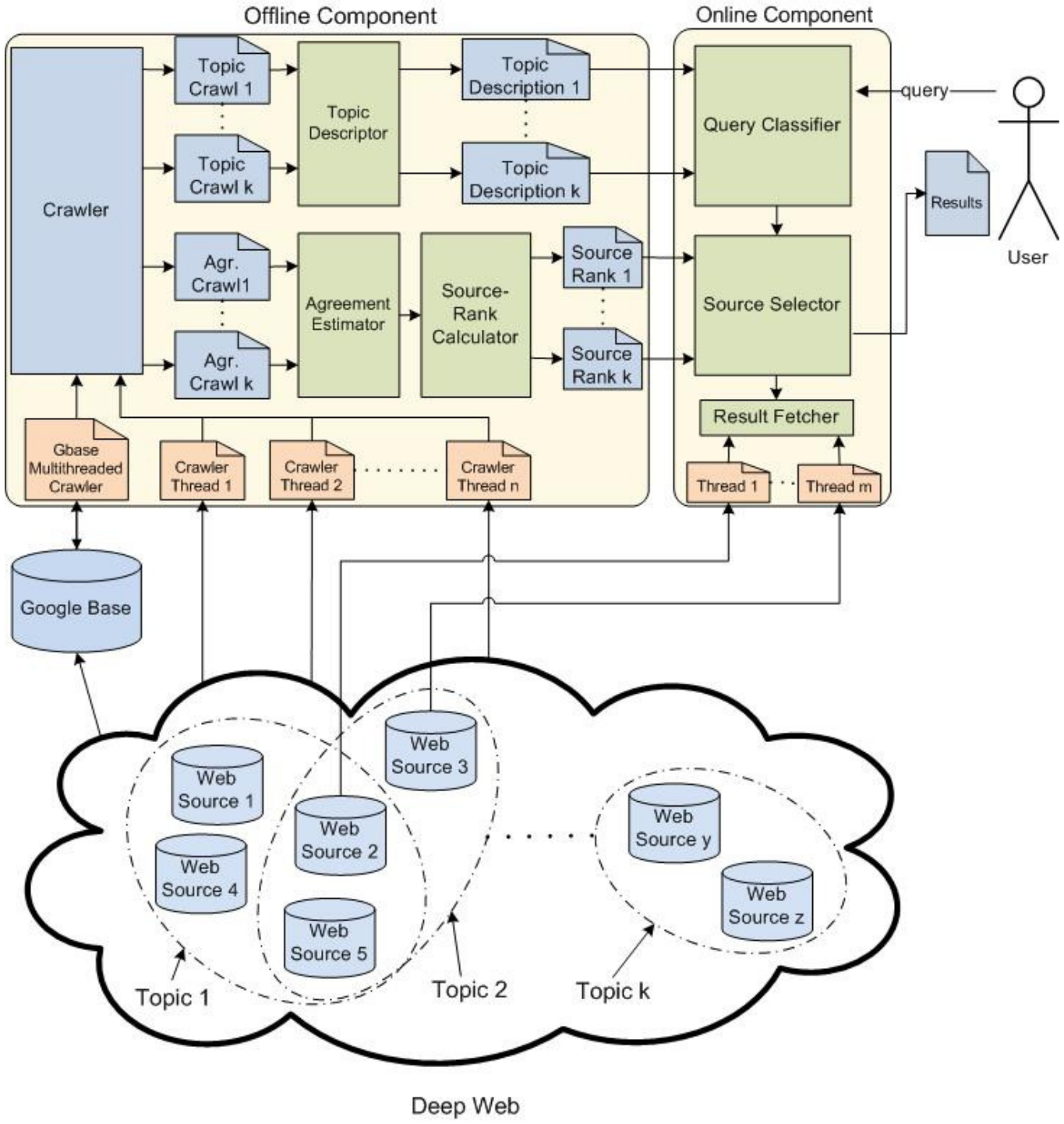
Trust-based measure for multi-topic deep-web

- Issues with SourceRank for multi-topic deep-web
  - Single importance ranking
  - Is query-agnostic
- We propose Topic-sensitive SourceRank, TSR for effectively performing multi-topic selection sensitive to trustworthiness
- TSR overcomes the drawbacks of SourceRank

Topic-sensitive SourceRank overview

- Multiple importance rankings
  - Each importance ranking biased towards a particular topic

- At query-time, using query information, composite importance ranking is computed, biased towards the query

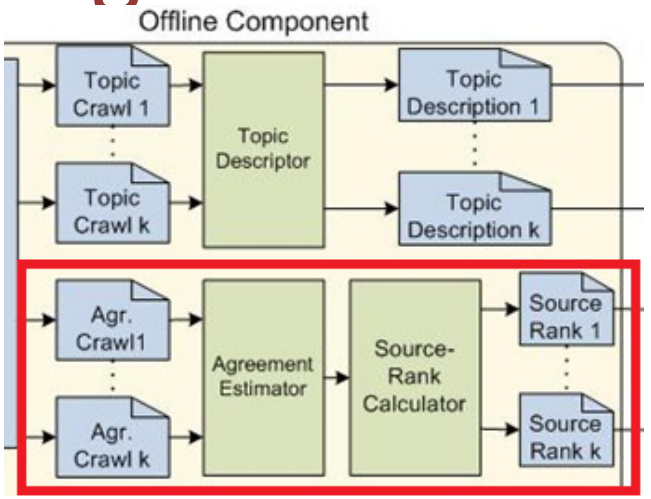

Offline Component
Online Component
Crawler
Topic Crawl 1
Topic Crawl k
Topic Descriptor
Topic Description 1
Topic Description k
Agr. Crawl1
Agr. Crawl k
Agreement Estimator
Source-Rank Calculator
Source Rank 1
Source Rank k
Query Classifier
Source Selector
Result Fetcher
query
Results
User
Gbase Multithreaded Crawler
Crawler Thread 1
Crawler Thread 2
Crawler Thread n
Thread 1
Thread m
Google Base
Web Source 1
Web Source 2
Web Source 3
Web Source 4
Web Source 5
Web Source y
Web Source z
Topic 1
Topic 2
Topic k
Deep Web

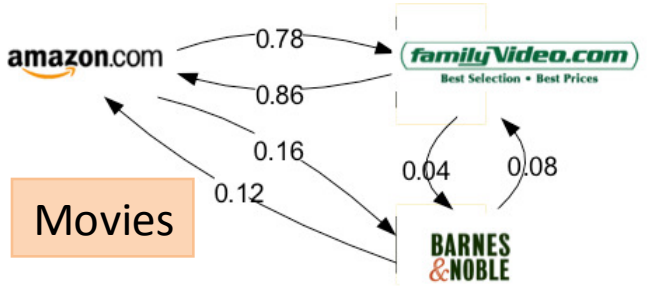
Challenges for TSR

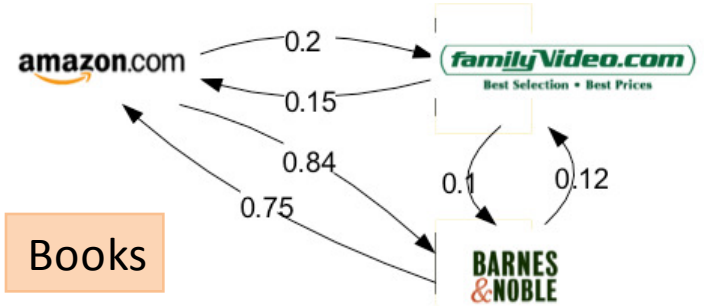
- Computing topic-specific importance rankings is not trivial
- Inferring query information
  - Identifying query-topic
  - Computing composite importance ranking

Computing topic-specific agreement

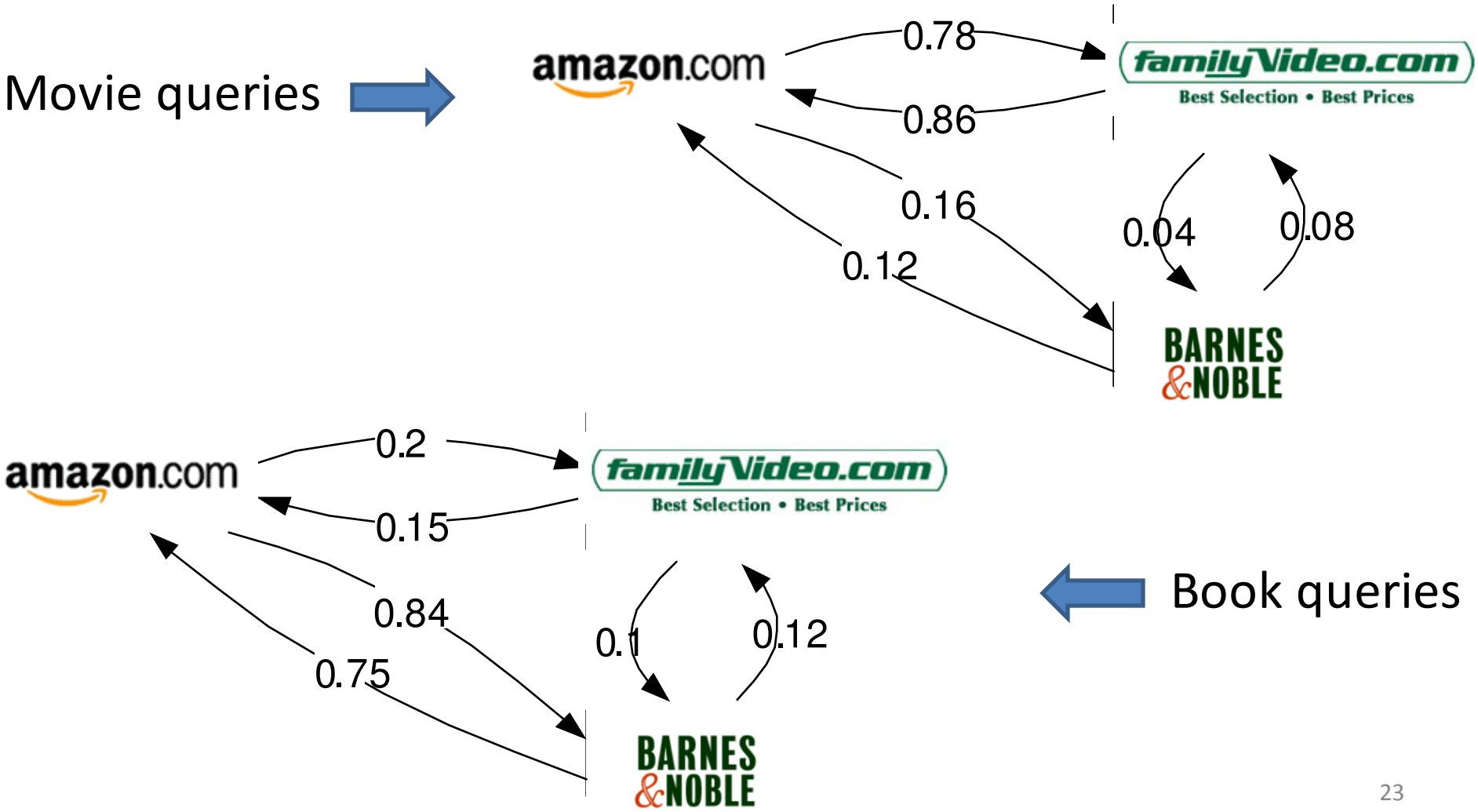
- For a deep-web source, its SourceRank score for topic will depend on the answers to queries of same topic
- Topic-specific sampling queries will result in an endorsement structure biased towards the same topic
- Topic Specific Source Ranks are stationary visit prob on the topic-specific agreement graphs



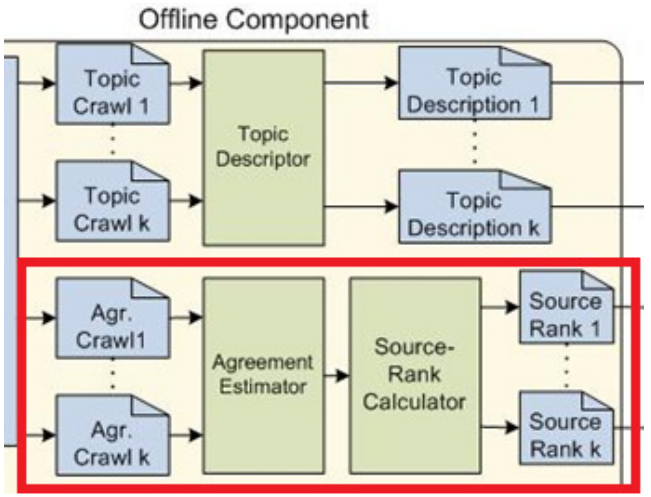




Sampling queries influencing agreement computation



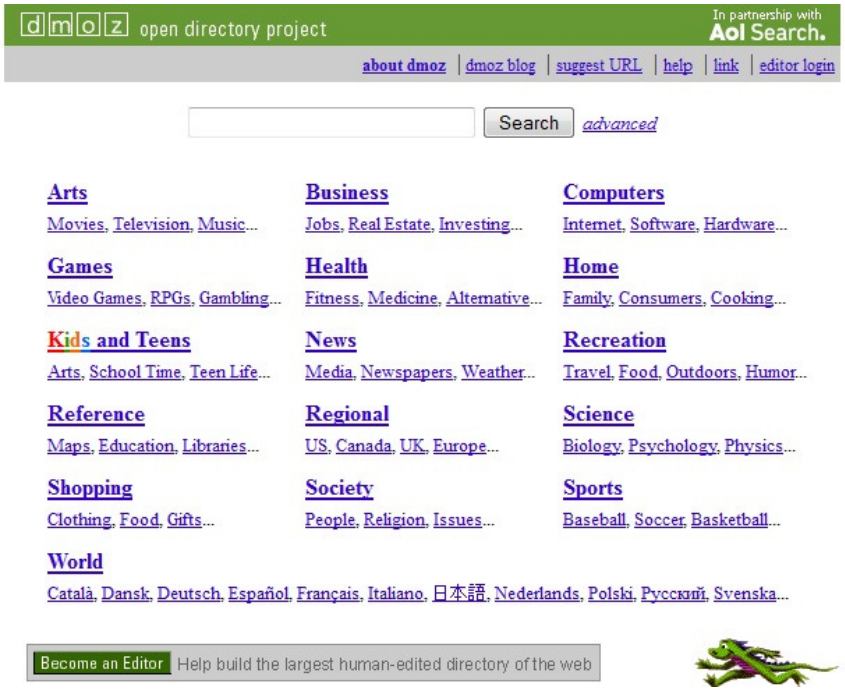
Computing Topic-specific SourceRanks



- Partial topic-specific sampling queries are used for obtaining source crawls
- Biased agreement graphs are computed using topic-specifc source crawls
- Performing a weighted random walk on the biased agreement graphs would result in topic-specific SourceRanks, TSR's

Topic-specific sampling queries

- Publicly available online directories such as **ODP**, **Yahoo Directory** provide hand-constructed topic hierarchies



- Directories are a good source for obtaining topic-specific sampling queries

Query Processing

- Computing query-topic
- Computing query-topic sensitive importance scores

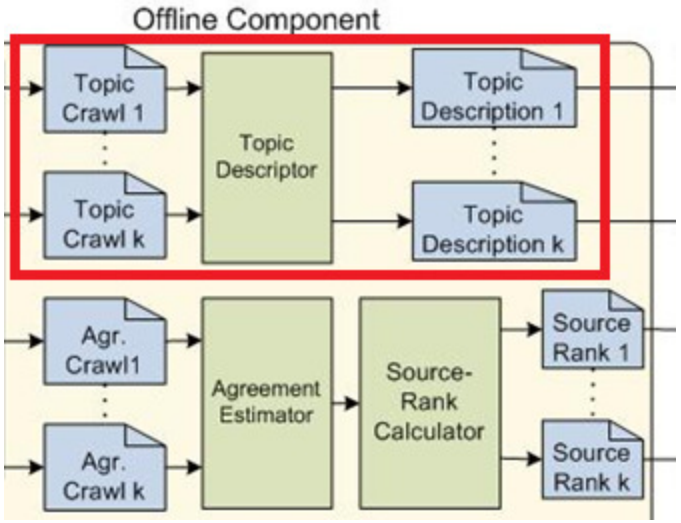
Computing query-topic

- Query-topic
  - Likelihood of the query belonging to topics
  - Soft classification problem

topic →	Camera	Book	Movie	Music
query-topic → For Query="godfather"	0	0.3	0.6	0.1

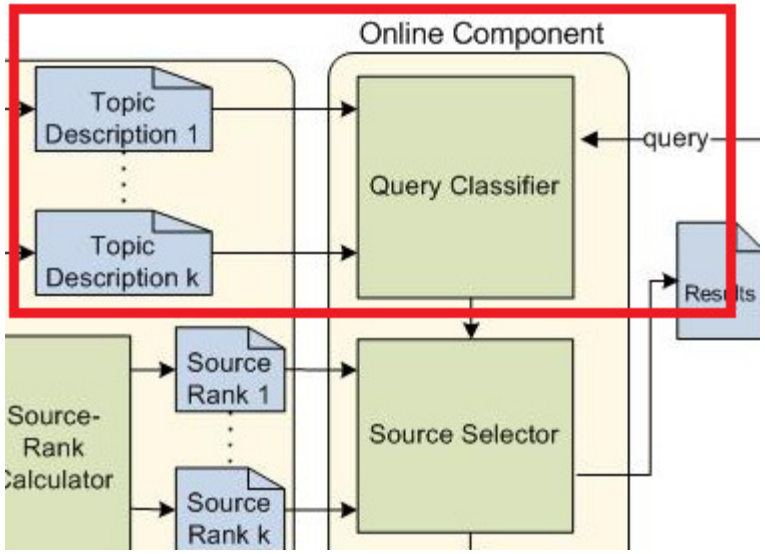
Computing query-topic – Training Data

- Training data
  - Description of topics
  - Topic-specific source crawls act as topic descriptions
  - Bag of words model



Computing query-topic – Classifier

- Classifier
  - Naïve Bayes Classifier (NBC) with parameters set to maximum likelihood estimates
  - NBC uses topic-description to estimate topic probability conditioned on query *q*



$$P(c_i \mid q) = \frac{P(q \mid c_i) \times P(c_i)}{P(q)} \propto P(c_i) \prod_j P(q_j \mid c_i)$$

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

Computing query-topic – Classifier contd.

Computing $P(c_i|q)$

$$P(c_i \mid q) = \frac{P(q \mid c_i) \times P(c_i)}{P(q)} \propto P(c_i) \prod_j P(q_j \mid c_i)$$

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

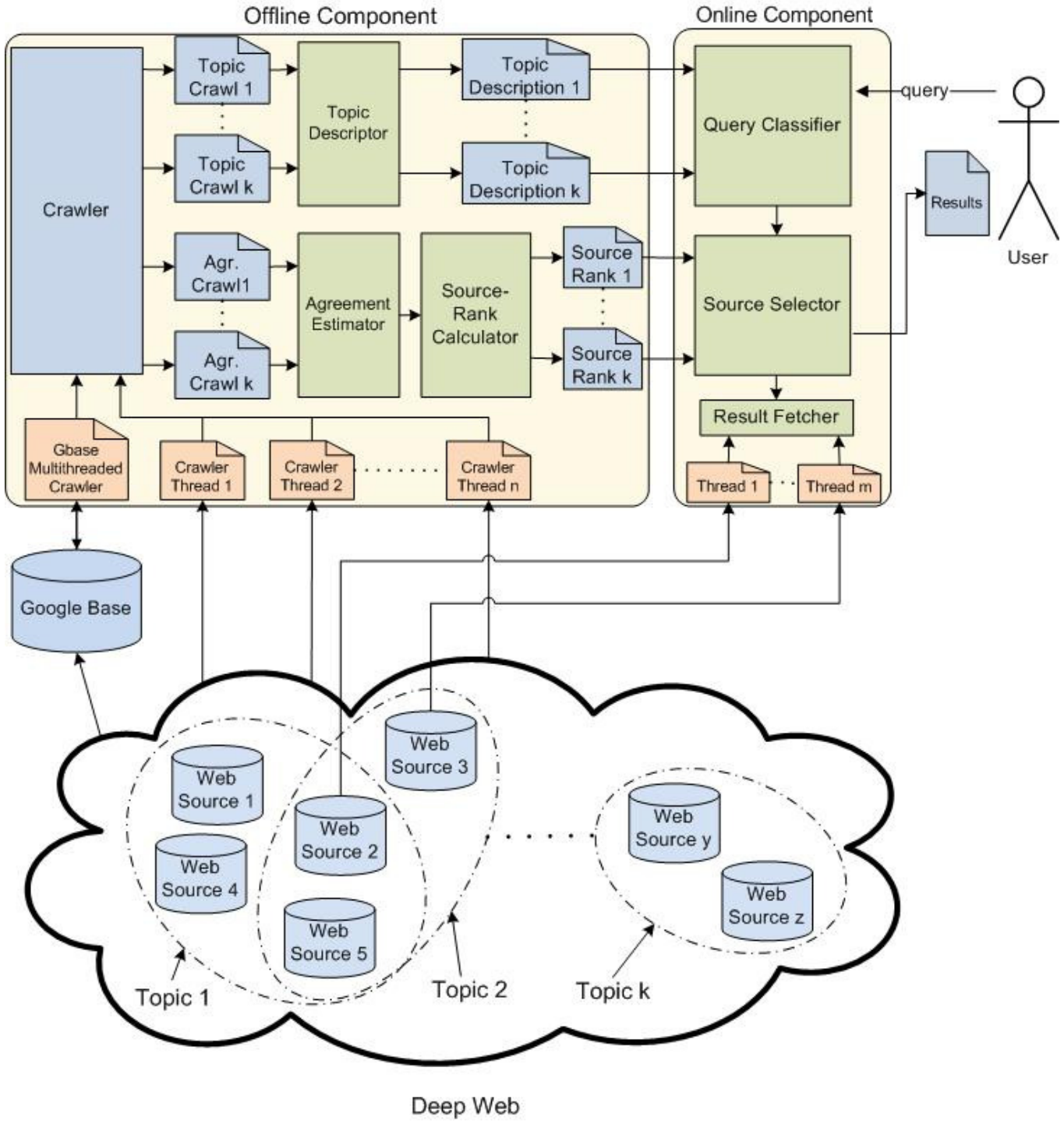
$$P(c_i \mid q) = \prod_j P(q_j \mid c_i)$$

Computing query-topic sensitive importance scores

- Topic-specific SourceRanks are linearly combined, weighted based on query-topic, to form a single composite importance ranking

$$CSR_k = \sum_i P(c_i \mid q) \times TSR_{ki}$$

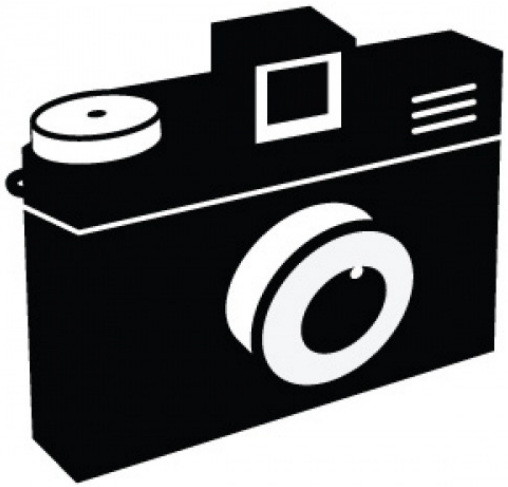
where TSR_{ki} is the topic-specific SourceRank score of source s_k for topic c_i

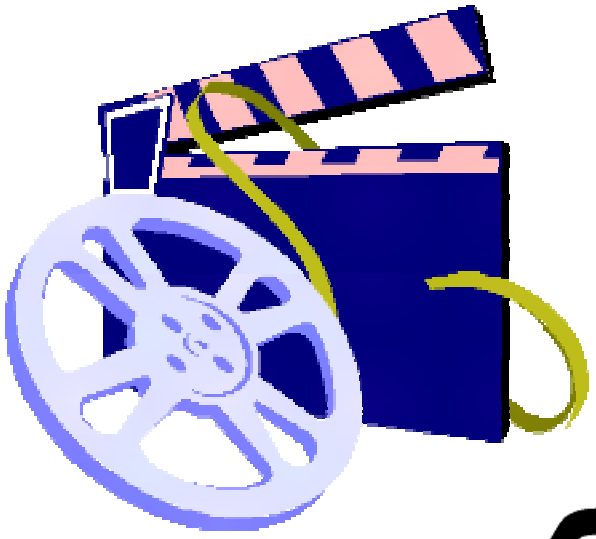

Offline Component
Online Component
Crawler
Topic Crawl 1
Topic Crawl k
Topic Descriptor
Topic Description 1
Topic Description k
Agr. Crawl1
Agr. Crawl k
Agreement Estimator
Source-Rank Calculator
Source Rank 1
Source Rank k
Query Classifier
Source Selector
Result Fetcher
query
Results
User
Gbase Multithreaded Crawler
Crawler Thread 1
Crawler Thread 2
Crawler Thread n
Thread 1
Thread m
Google Base
Web Source 1
Web Source 2
Web Source 3
Web Source 4
Web Source 5
Web Source y
Web Source z
Topic 1
Topic 2
Topic k
Deep Web

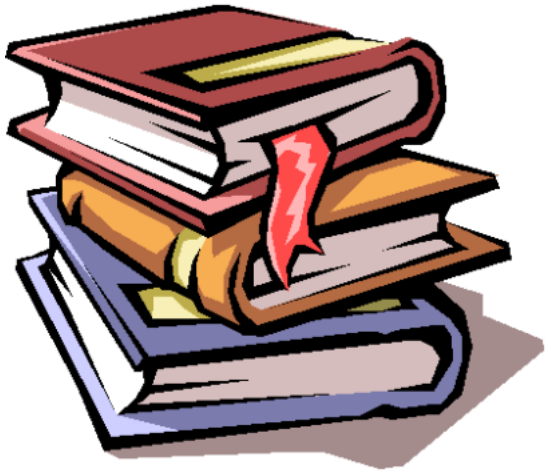
Agenda

- SourceRank
- Topic-sensitive SourceRank
- Experimental setup
- Results
- Conclusion

Four-topic deep-web environment



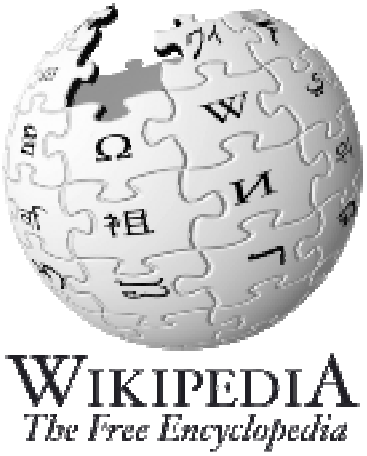




Deep-web sources

- Collected via Google Base
- 1440 sources: 276 camera, 556 book, 572 movie and 281 music sources

Sampling queries



The New York Times

- Used 200 random titles or names for each topic

Test queries

- Mix of queries from all four topics
- Generated by randomly removing words from titles or names with 0.5 probability
- Number of test queries varied for different topics to obtain the required (*0.95*) statistical significance

Baseline Methods

- Used four baseline methods for comparison
  1. CORI (A standard Collection Selection approach)
  2. GoogleBase
     1. All Sources
     2. Only on the crawled sources
  3. USR: Undifferentiated Source Rank
     - One rank per source, independent of query topic
  4. DSR: Domain Specific Source Rank
     - Assumes oracular information on the topic of the source as well as query

Baseline 1- CORI

- Source statistics collected using highest document frequency terms
- Source selection performed using the same parameters as found optimal in CORI paper

Searching Distributed Collections
With Inference Networks

James P. Callan Zhihong Lu W. Bruce Croft

Computer Science Department, University of Massachusetts
Amherst, MA 01003-4610, USA
{callan,zlu,croft}@cs.umass.edu

Baseline 2- Google Base

- Two-versions of Google Base
  - *Gbase on dataset*: Google Base search restricted to our crawled sources
  - *Gbase*: Google Base search with no restrictions i.e. considers all sources in Google Base

Baseline 3- USR

- Undifferentiated SourceRank, USR
  - SourceRank extended to multi-topic deep-web
  - Single agreement graph is computed using sampling queries

Baseline 4 - DSR

- Oracular source selection, DSR
  - Assumes a perfect classification of sources and user queries are available
  - Creates agreement graphs and SourceRanks for a domain using just the in-domain sources
  - For each test query, sources ranking high in the domain corresponding to the test query are used

Source selection

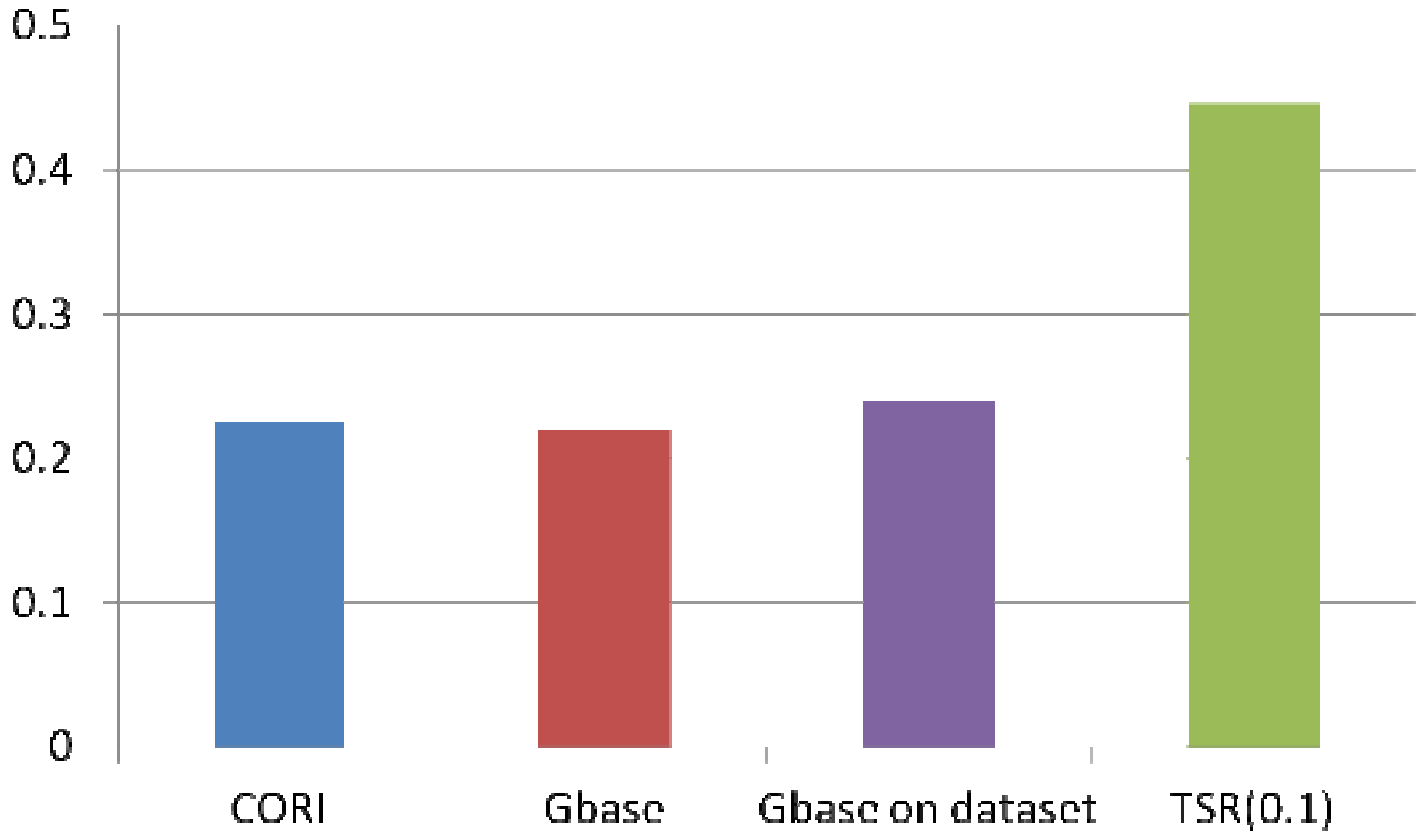
- Agreement based selection models (TSR, USR and DSR) use a weighted combination of importance and relevance scores
  - Example: TSR(0.1) represents 0.9xCORI + 0.1xTSR
- For each query q, *top-k* sources are selected
- Google Base is made to query only on these to *top-k* sources

Tuple ranking and relevance evaluation

- Google Base's tuple ranking is used for ranking resulting tuples
- *Top-5* results returned were manually classified as relevant or irrelevant
- Result classification was rule based

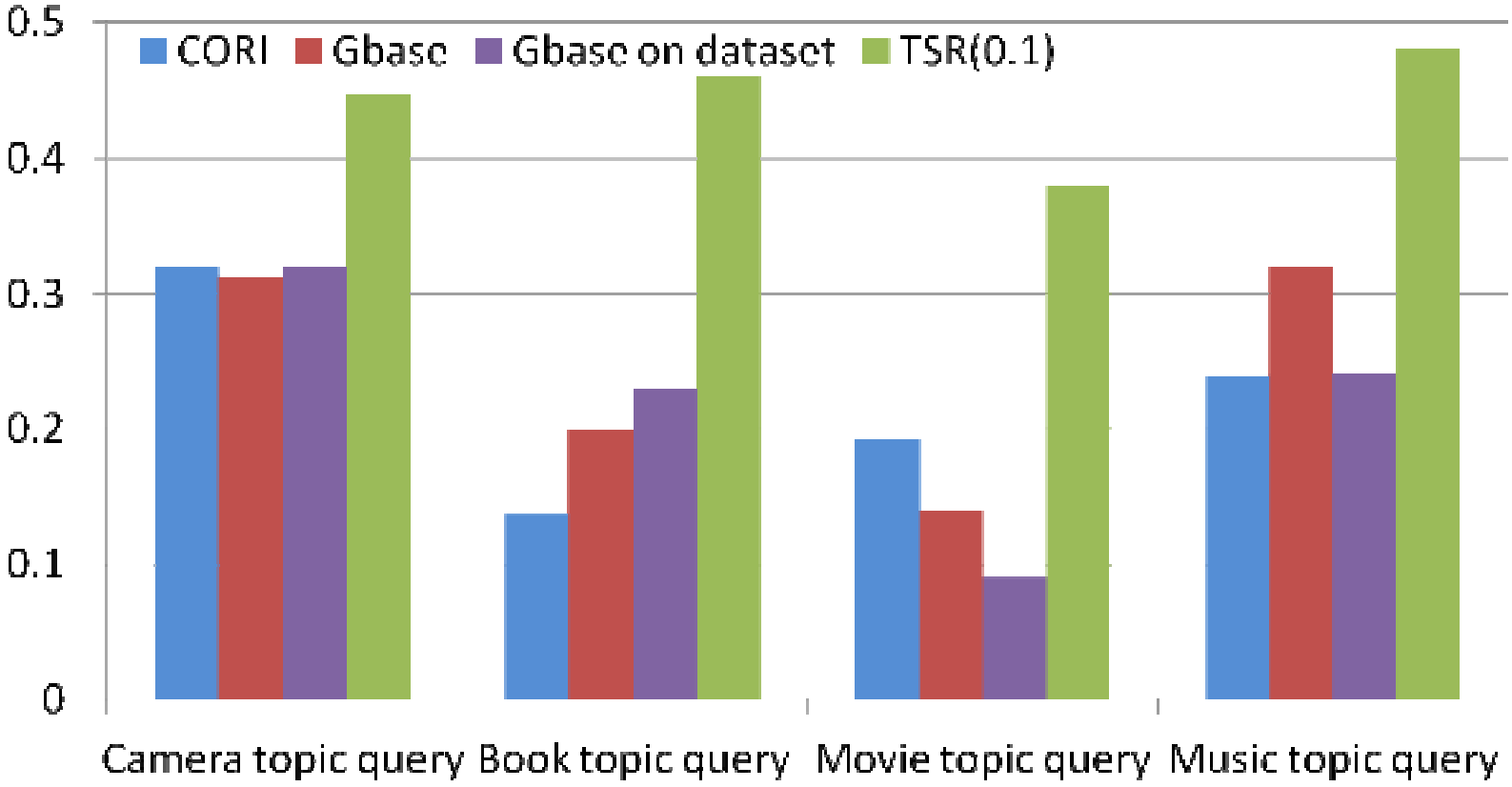
Agenda

- SourceRank
- Topic-sensitive SourceRank
- Experiments
- Results
- Conclusion



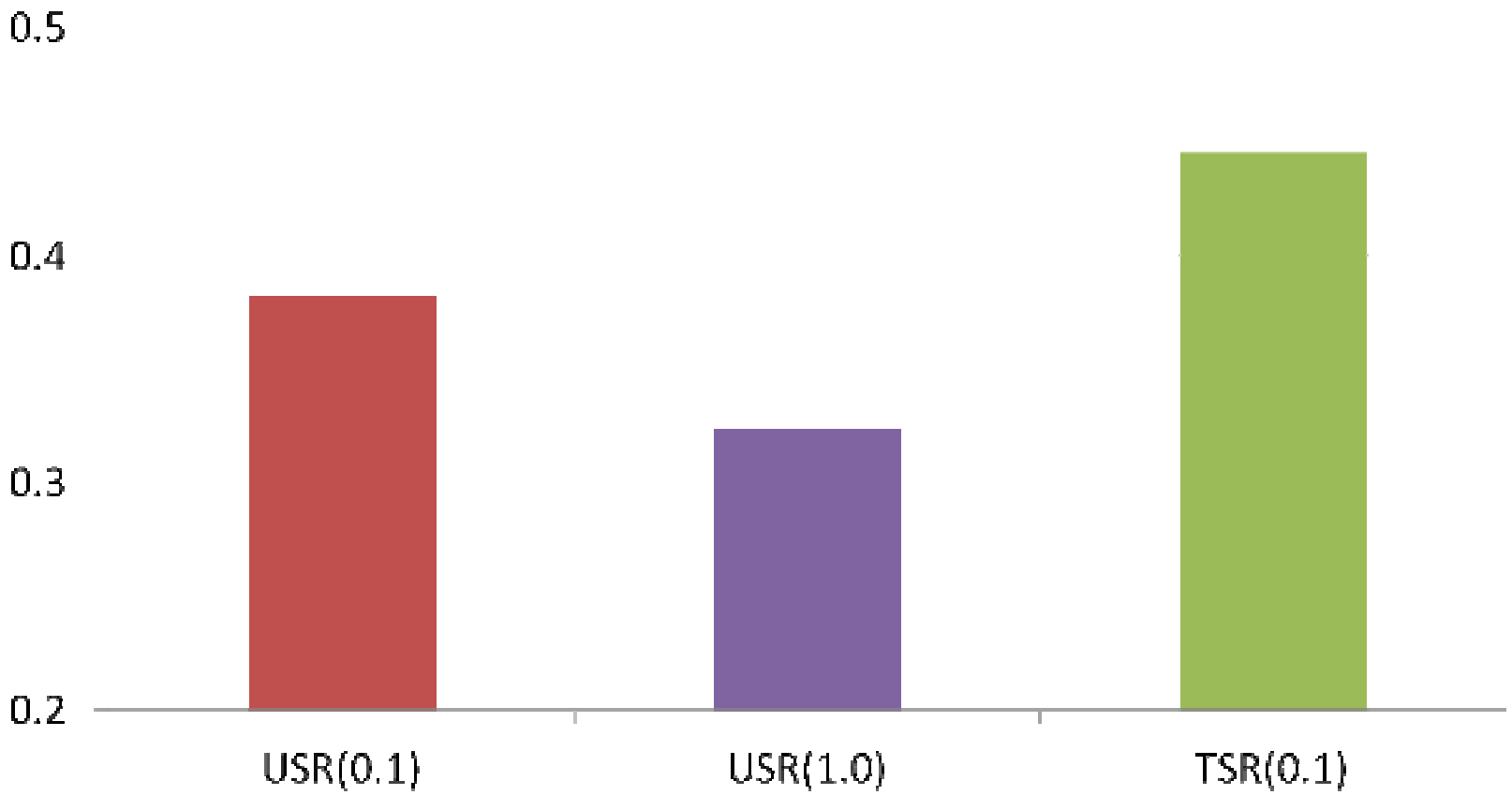
Comparison of *top-5* precision of TSR(0.1) and query similarity based methods: CORI and Google Base

- TSR precision exceeds that of similarity-based measures by 85%



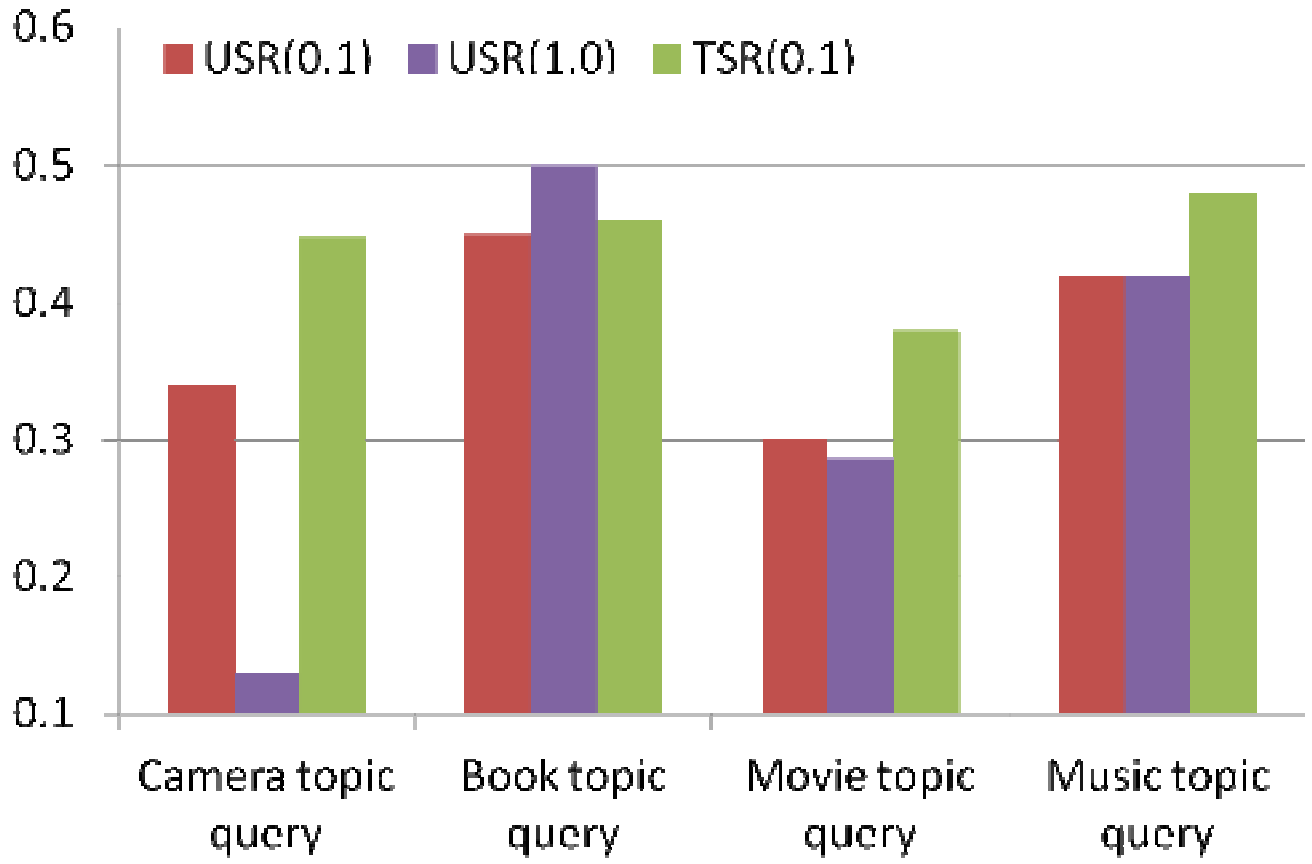
Comparison of topic-wise *top-5* precision of TSR(0.1) and query similarity based methods: CORI and Google Base

- TSR significantly out-performs all query-similarity based measures for all topics



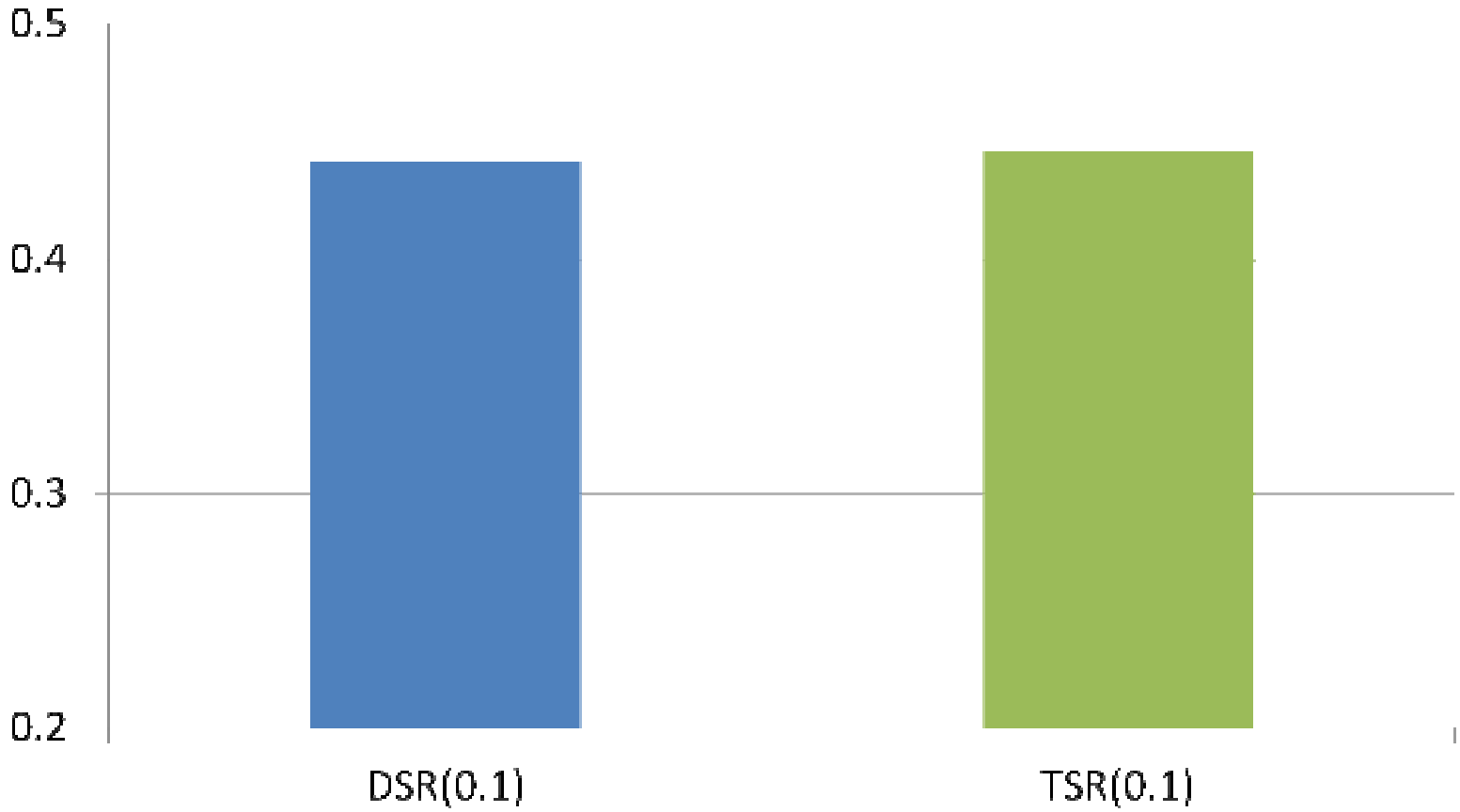
Comparison of *top-5* precision of TSR(0.1) and agreement based methods: USR(0.1) and USR(1.0)

- TSR precision exceeds USR(0.1) by 18% and USR(1.0) by 40%



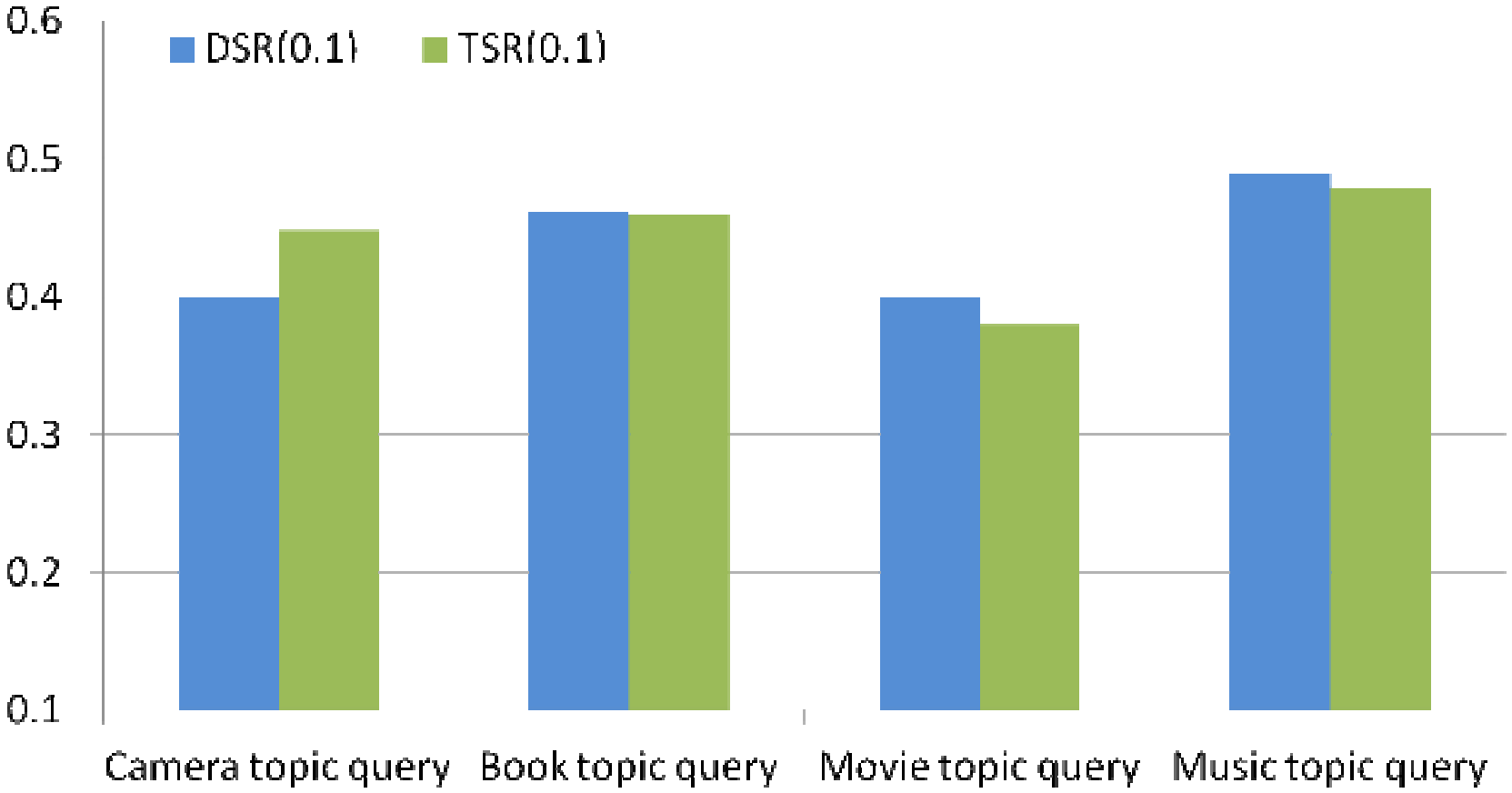
Comparison of topic-wise *top-5* precision of TSR(0.1) and agreement based methods: USR(0.1) and USR(1.0)

- For three out of the four topics, TSR(0.1) out-performs USR(0.1) and USR(1.0) with confidence levels 0.95 or more



Comparison of *top-5* precision of TSR(0.1) and oracular DSR(0.1)

- TSR(0.1) is able to match DSR(0.1)'s performance



Comparison of topic-wise *top-5* precision of TSR(0.1) and oracular DSR(0.1)

- TSR(0.1) matches DSR(0.1) performance across all topics indicating its effectiveness in identifying important sources across all topics

Agenda

- SourceRank
- Topic-sensitive SourceRank
- Experimental setup
- Results
- Conclusion

Conclusion

- Attempted multi-topic source selection sensitive to trustworthiness and importance for the deep-web
- Introduced topic-sensitive SourceRank (TSR)
- Our experiments on more than a thousand deep-web sources show that a TSR-based approach is highly effective in extending SourceRank to multi-topic deep-web

Conclusion contd.

- TSR out-performs query-similarity based measures by around 85% in precision
- TSR results in statistically significant precision improvements over other baseline agreement-based methods
- Comparison with oracular DSR approach reveals effectiveness of TSR for topic-specific query and source classification and subsequent source selection

Questions?

Source selection

- Linearly combines relevance-scores with importance scores

- Overall score of a source s_k is computed as

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

where R_k: relevancy score of s_k

CSR_k: query-topic sensitive score of s_k

Paper submitted to Comad'11

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

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

#Computer Science and Engineering, Arizona State University
Tempe AZ USA 85287
{1mjha1,2rajub,3rao}@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 domain deep web search adapting an agreement based approach (*e.f.* SourceRank [10]). In this paper we introduce a source selection method sensitive to trust and importance for *multi domain* deep web search. We compute multiple quality scores of a source specific to different domains, based on the domain specific crawl data. At the query time, we classify the query to determine its probability of membership in different domains. These fractional memberships are used as the weights to the domain specific scores of sources to select sources for the query. Extensive experiments on more than a thousand sources in multiple domains show 18-85% improvements in result quality over Google Product Search and other existing methods.

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 domains [26]. Searching the deep web has been identified as the next big challenge in information management [30]. 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 straight forward idea for extending SourceRank for multi-domain deep web search is a weighted combination with query similarity, like PageRank [13]. But in general, agreement by sources in the same domain is likely to be much more indicative of importance of a source than endorsement by out of domain sources. Moreover, sources might have data corresponding to multiple topics. The importance of the source might vary 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. Havelwala [22]), but is more critical for the deep web since sources are even more likely to cross topics/domains than single web pages. To account for this fact, we extend the deep web source selection by assessing a domain-specific

17th International Conference on Management of Data
COMAD 2011, Bangalore, India, December 19-21, 2011

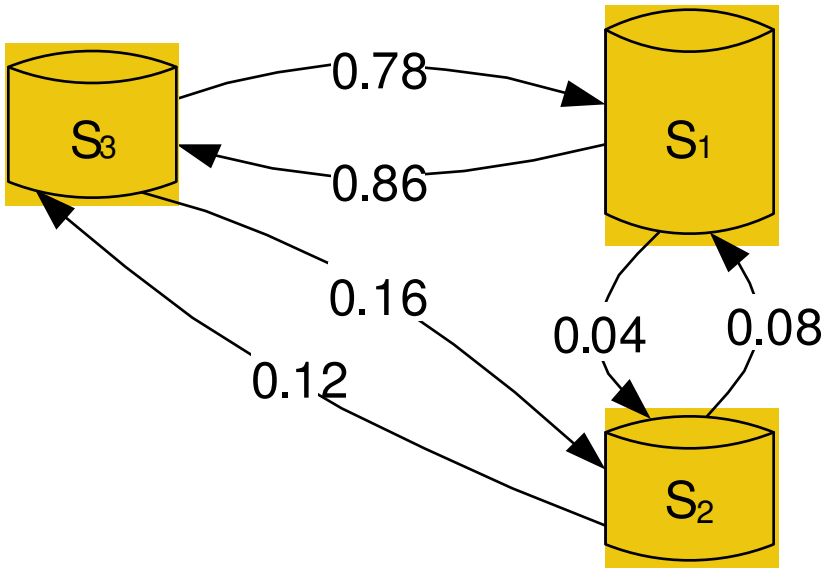
Ex: a two-topic deep web consisting of three sources: S_1, S_2 and S_3

Movie topic: S_1 and S_3

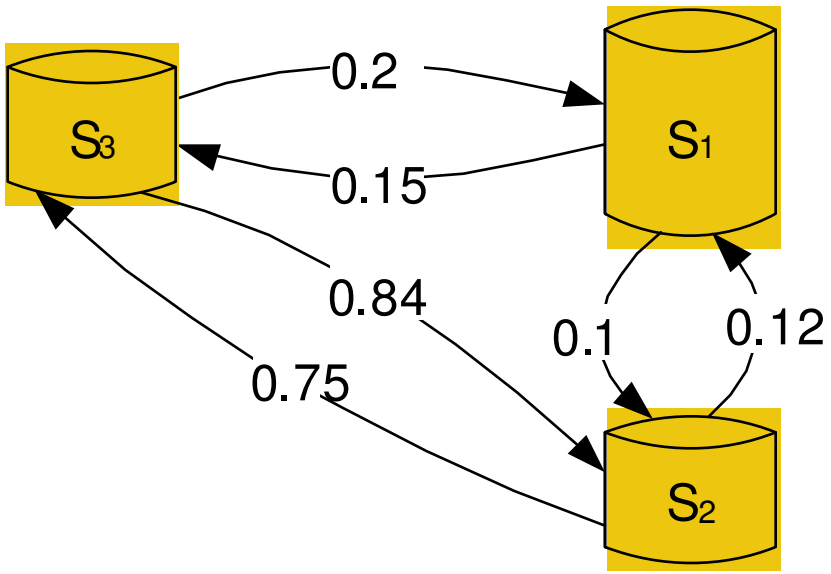
Book topic: S_2 and S_3

Agreement graphs

Movie queries



Book queries



Computing topic-specific SourceRank contd.

- SourceRank computed on biased agreement graph for a topic will capture topic-specific source importance ranking for the same topic