CS47300: Web Information Search and Management

Deep Web & Federated Search

Prof. Chris Clifton

30 October 2020

Hidden Web

Visible Web vs. Hidden Web

- Visible Web: Information can be copied (crawled) and accessed by conventional search engines like Google or Yahoo!
- Hidden Web: Information hidden from conventional engines. Provide source-specific search engine but no arbitrary crawling of the data
  - No arbitrary crawl of the data
  - Updated too frequently to be crawled

  → **Can NOT Index (promptly)**
- Hidden Web contained in (Hidden) information sources that provide text search engines to access the hidden information

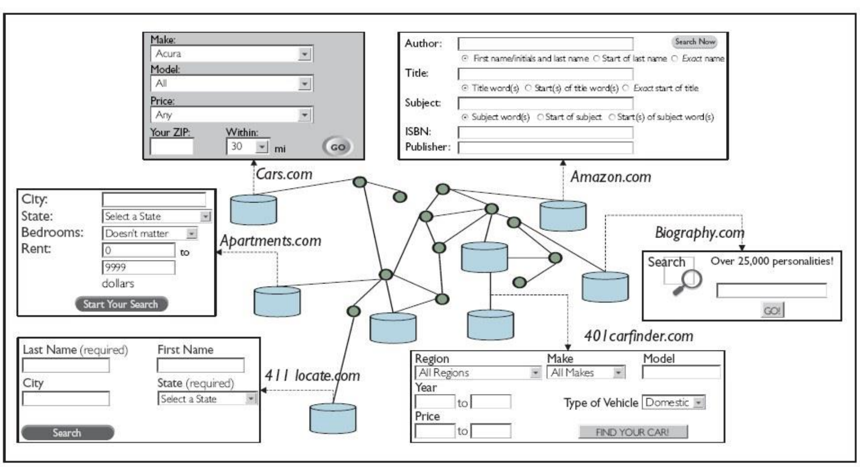
Department of Computer Science

Deep Web vs. Dark Web

- Dark Web: Hidden intentionally
  - Largely to support illegal or socially unacceptable activity
  - *But legality and acceptability vary, web is trans-national and trans-cultural*
  - We won't go here…
- Deep Web: Data hidden behind interfaces
  - Can we crawl this data?

Department of Computer Science

Conceptual View
(He, Patel, Zhang, Chang '07)



Why can't we crawl the entire web?

A. Pages with no incoming links
B. Dynamically created content
C. Web servers forbid crawling
D. All of the above
E. We CAN crawl the entire web!

Does the Deep Web Matter?

- Where are the entry points?
- What is the scale?
- How "structured" is the data?
- What topics are covered?
- How well do search engines already cover this?
- Wat about existing specialized portals?

PURDUE UNIVERSITY
Department of Computer Science

Size Estimate of the Deep Web

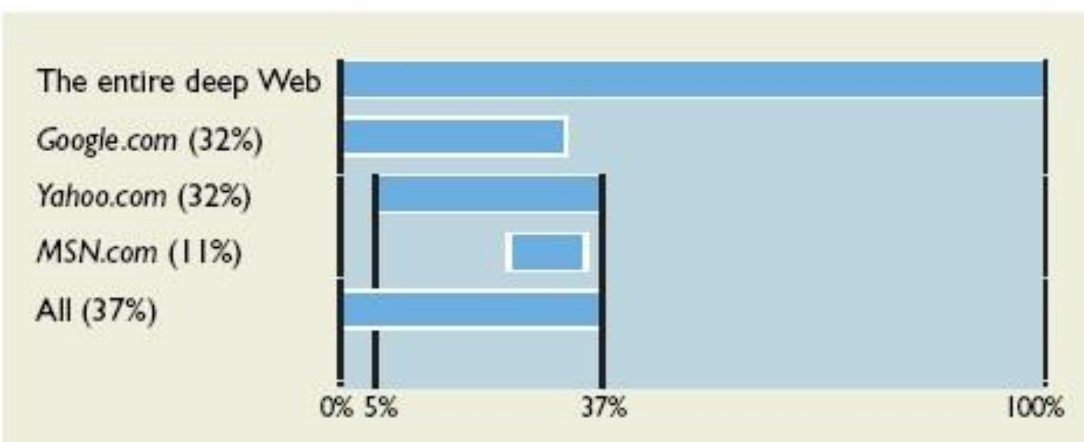
	Sampling Results	Total Estimate	99% Confidence Interval
Deep Web sites	126	307,000	236,000 - 377,000
Web databases	190	450,000	366,000 - 535,000
–unstructured	*43*	*102,000*	62,000 - 142,000
–structured	*147*	*348,000*	275,000 - 423,000
Query interfaces	406	1,258,000	1,097,000 - 1,419,000

Chag, He, Li, Patel, Zhang SIGMoD Record 2004

8

PURDUE UNIVERSITY
Department of Computer Science

Search Engine Coverage

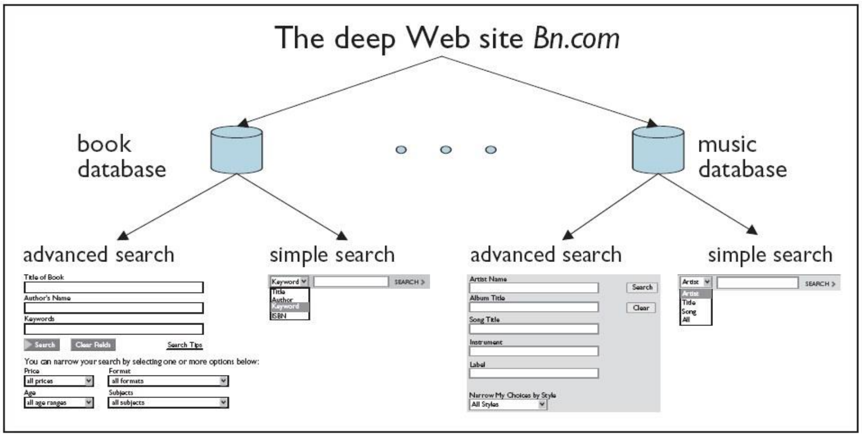


9

PURDUE UNIVERSITY

Department of Computer Science

Deep Web Components
(He, Patel, Zhang, Chang '07)



10

PURDUE UNIVERSITY

Department of Computer Science

Challenges

- How do we know what is in a database?
  - Sample queries?
  - Search page
    - Descriptive information
    - Form fields
- How do we query it?
- How do we process results?

11

Can this be real?

- "General" search
  - See Google, etc.
- "Specialized" search
  - [Metaquerier]
  - [Cazoodle]
- Federated Search

12

Federated Search

Outline

- Introduction to federated search
- Main research problems
  - Resource Representation
  - Resource Selection
  - Results Merging

PURDUE UNIVERSITY
Department of Computer Science

Federated Search

PURDUE UNIVERSITY
Department of Computer Science

Introduction

Hidden Web is:

- Larger than Visible Web (2-50 times, Sherman 2001)
- Created by professionals

} **Valuable** → **Searched by** ***Federated Search***

Federated Search Environments:

Small companies: Probably cooperative information sources

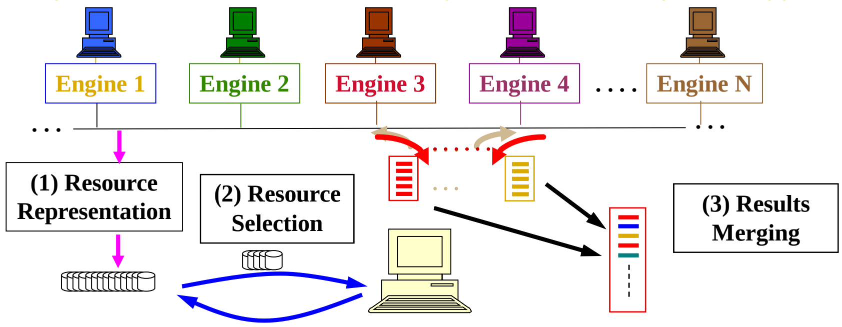
Big companies (organizations): Probably uncooperative information sources

Web: Uncooperative information sources

PURDUE UNIVERSITY
Department of Computer Science

Federated Search

Components of a Federated Search System and Two Important Applications



Information source recommendation: **Recommend information sources for users' text queries (e.g., completeplanet.com)**: Steps 1 and 2

Federated document retrieval**: Also search selected sources and merge individual ranked lists into a single list:** Steps 1, 2 and 3

PURDUE UNIVERSITY
Department of Computer Science

Introduction

Solutions of Federated Search

Information source recommendation: Recommend information sources for users' text queries

- Useful when users want to browse the selected sources
- Contain resource representation and resource selection components

Federated document retrieval: Search selected sources and merge individual ranked lists

- Most complete solution
- Contain all of resource representation, resource selection and results merging

PURDUE UNIVERSITY
Department of Computer Science

Introduction

Modeling Federated Search

Application in real world

- FedStats project: Web site to connect dozens of government agencies with uncooperative search engines
  - Previously use centralized solution (ad-hoc retrieval), but suffer a lot from missing new information and broken links
  - Require federated search solution: A prototype of federated search solution for FedStats is on-going in Carnegie Mellon University
- Good candidate for evaluation of federated search algorithms
- But, not enough relevance judgments, not enough control… → **Requires Thorough Simulation**

PURDUE UNIVERSITY
Department of Computer Science

Introduction

Modeling Federated Search

TREC data

- Large text corpus, thorough queries and relevance judgments

Simulation with TREC news/government data

- Professional well-organized contents
- Can be divided into O(100) information sources
- Simulate environments of large companies or domain specific hidden Web
- Most commonly used, many baselines (Lu et al., 1996) (Callan, 2000) ...
- Normal or moderately skewed size testbeds: Trec123 or Trec4_Kmeans
- Skewed: Representative (large source with the same relevant doc density),
  Relevant (large source with higher relevant doc density),
  Nonrelevant (large source with lower relevant doc density)

PURDUE UNIVERSITY
Department of Computer Science

Introduction

Modeling Federated Search

Simulation multiple types of search engines

- **INQUERY**: Bayesian inference network with Okapi term formula, doc score range [0.4, 1]
- **Language Model**: Generation probabilities of query given docs doc score range [-60, -30] (log of the probabilities)
- **Vector Space Model**: SMART "lnc.ltc" weighting doc score range [0.0, 1.0]

Federated search metric

- Information source size estimation: Error rate in source size estimation
- Information source recommendation: **High-Recall**, select information sources with most relevant docs
- Federated doc retrieval: **High-Precision** at top ranked docs

PURDUE UNIVERSITY
Department of Computer Science

Federated Search

Outline

- Introduction to federated search
- Main research problems
  - ➢Resource Representation
  - – Resource Selection
  - – Results Merging

Research Problems (Resource Representation)

- Previous Research on Resource Representation

Resource descriptions of words and the occurrences

- **STARTS protocol** (Gravano et al., 1997): Cooperative protocol
- **Query-Based Sampling** (Callan et al., 1999):
  - Send random queries and analyze returned docs
  - Good for uncooperative environments

Centralized sample database: Collect docs from Query-Based Sampling (QBS)

- For query-expansion (Ogilvie & Callan, 2001), not very successful
- Successful utilization for other problems, throughout this proposal

Research Problems (Resource Representation)

- Research on Resource Representation

Information source size estimation

Important for resource selection and provide users useful information

- Capture-Recapture Model (Liu and Yu, 1999)

  Use two sets of independent queries, analyze overlap of returned doc ids

  But require large number of interactions with information sources

Sample-Resample Model (Si and Callan, 2003)

Assume: Search engine indicates num of docs matching a one-term query

Strategy: Estimate df of a term in sampled docs

Get total df from by resample query from source

Scale the number of sampled docs to estimate source size

PURDUE UNIVERSITY
Department of Computer Science

Research Problems (Resource Representation)

Experiments

To conduct component-level study

- Capture-Recapture: about 385 queries (transactions)

- Sample-Resample: 80 queries and 300 docs for sampled docs (sample) + 5 queries (resample) = 385 transactions

Measure:

Absolute error ratio $AER=\frac{|N-N^*|}{N^*}$

Estimated Source Size

Actual Source Size

Collapse every 10th source of Trec123

	Trec123 (Avg AER, lower is better)	Trec123-10Col (Avg AER, lower is better)
Cap-Recapture	**0.729**	**0.943**
Sample-Resample	0.232	0.299

PURDUE UNIVERSITY
Department of Computer Science

Federated Search

Outline

- Introduction to federated search
- Main research problems
  - Resource Representation
  - ➢Resource Selection
  - Results Merging

PURDUE UNIVERSITY
Department of Computer Science

Research Problems (Resource Selection)

Goal of Resource Selection of Information Source Recommendation

High-Recall**: Select the (few) information sources that have the most relevant documents**

Research on Resource Selection

Resource selection algorithms that need training data

- **Decision-Theoretic Framework** (DTF) (Nottelmann & Fuhr, 1999, 2003)

  DTF causes large human judgment costs

- **Lightweight probes** (Hawking & Thistlewaite, 1999)

  Acquire training data in an online manner, large communication costs

PURDUE UNIVERSITY
Department of Computer Science

Research Problems (Resource Selection)

Research on Resource Representation

"Big document" resource selection approach: Treat information sources as big documents, rank them by similarity of user query

- **Cue Validity Variance (CVV)** (Yuwono & Lee, 1997)
- **CORI** (Bayesian Inference Network) (Callan,1995)
- **KL-divergence** (Xu & Croft, 1999)(Si & Callan, 2002), Calculate KL divergence between distribution of information sources and user query

CORI and KL were the state-of-the-art (French et al., 1999)(Craswell et al,, 2000)

But "Big document" approach loses doc boundaries and does not optimize the goal of High-Recall

Language Model Resource Selection

$$P\left(db_i \mid Q\right) = \frac{P(Q \mid db_i) * P(db_i)}{P(Q)}$$

DB independent constant

$$P\left(Q \mid db_i\right) = \prod_{q \in Q} \left(\lambda P\left(q \mid db_i\right) + (1-\lambda) P\left(q \mid G\right)\right)$$

Calculate on Sample Docs

In Language Model Framework, P(C_i) is set according to DB Size

$$P(C_i) = \frac{\hat{N}_{C_i}}{\sum_j \hat{N}_{C_j}}$$

Research Problems (Resource Selection)

Research on Resource Representation

But "Big document" approach loses doc boundaries and does not optimize the goal of High-Recall

Relevant document distribution estimation (ReDDE) (Si & Callan, 2003)

Estimate the percentage of relevant docs among sources and rank sources with no need for relevance data, much more efficient

Research Problems (Resource Selection)

Purdue University, Department of Computer Science

Relevant Doc Distribution Estimation (ReDDE) Algorithm

Source Scale Factor

Estimated Source Size

Number of Sampled Docs

$$Rel\_Q(i) = \sum_{d \in db_i} P(rel|d) * P(d|db_i) * N_{db_i} \approx \sum_{d \in db_i\_samp} P(rel|d) * SF_{db_i}$$

$$SF_{db_i} = \frac{\hat{N}_{db_i}}{N_{db_i\_samp}}$$

Rank on Centralized Complete DB

"Everything at the top is (equally) relevant"

$$P(rel|d) = \begin{cases} C_Q & \text{if } Rank_{CCDB}(Q,d) < ratio * \sum_i N_{db_i} \\ 0 & \text{otherwise} \end{cases}$$

Problem: To estimate doc ranking on Centralized Complete DB

Research Problems (Resource Selection)

Purdue University, Department of Computer Science

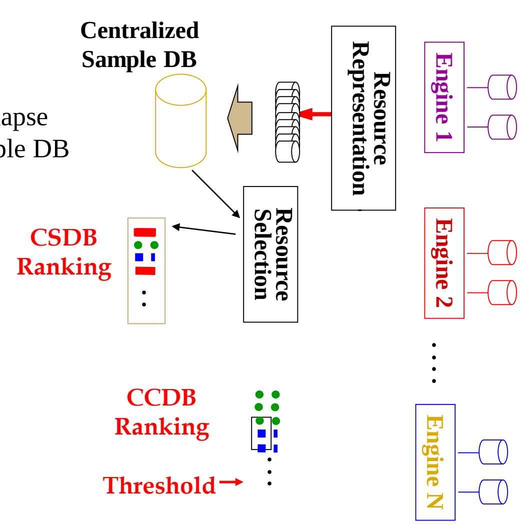
ReDDE Algorithm (Cont)

In resource representation:

- Build representations by QBS, collapse sampled docs into centralized sample DB

In resource selection:

- Construct ranking on CCDB with ranking on CSDB



Research Problems (Resource Selection)

Purdue University — Department of Computer Science

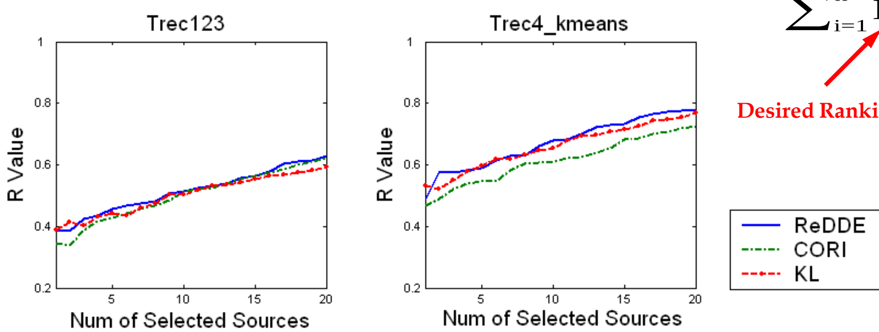
Experiments

On testbeds with uniform or moderately skewed source sizes

Evaluated Ranking

$$R_k = \frac{\sum_{i=1}^{k} E_i}{\sum_{i=1}^{k} B_i}$$

Desired Ranking

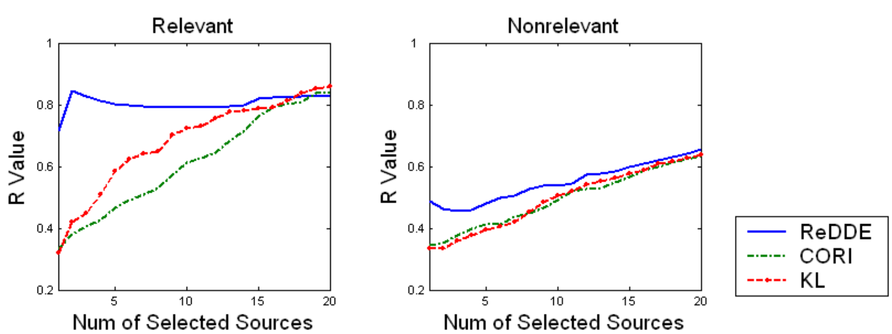


Research Problems (Resource Selection)

Purdue University — Department of Computer Science

Experiments

On testbeds with skewed source sizes



Federated Search

Outline

- Introduction to federated search
- Main research problems
  - – Resource Representation
  - – Resource Selection
  - ➢ Result Merging

Why can't we just rank based on scores?

A. Scores are relative, and only are comparable within a single corpus

B. Different scoring methodologies

C. Search engines provide ranking, not scores

39

PURDUE UNIVERSITY
Department of Computer Science

Research Problems (Results Merging)

Goal of Results Merging

Make different result lists comparable and merge them into a single list

Difficulties:

- Information sources may use **different retrieval algorithms**
- Information sources have **different corpus statistics**

Previous Research on Results Merging

Most accurate methods directly calculate comparable scores

- **Use same retrieval algorithm and same corpus statistics** (Viles & French, 1997)(Xu and Callan, 1998), need source cooperation
- **Download retrieved docs and recalculate scores** (Kirsch, 1997), large communication and computation costs

PURDUE UNIVERSITY
Department of Computer Science

Research Problems (Results Merging)

Research on Results Merging

Methods approximate comparable scores

- **Round Robin** (Voorhees et al., 1997), only use source rank information and doc rank information, fast but less effective
- **CORI merging formula** (Callan et al., 1995), linear combination of doc scores and source scores
    - Use linear transformation, a hint for other method
    - Work in uncooperative environment, effective but need improvement

PURDUE UNIVERSITY

Department of Computer Science

Research Problems (Results Merging)

Thought

Previous algorithms either try to **calculate** or to **mimic** the effect of the centralized scores

Can we estimate the centralized scores effectively and efficiently?

Semi-Supervised Learning (SSL) Merging **(Si & Callan, 2002, 2003)**

- Some docs exist in both centralized sample DB and retrieved docs

From Centralized sampled DB and individual ranked lists when long ranked lists are available

Download minimum number of docs with only short ranked lists

- Linear transformation maps source specific doc scores to source independent scores on centralized sample DB

PURDUE UNIVERSITY

Department of Computer Science

Research Problems (Results Merging)

SSL Results Merging (cont)

In resource representation:

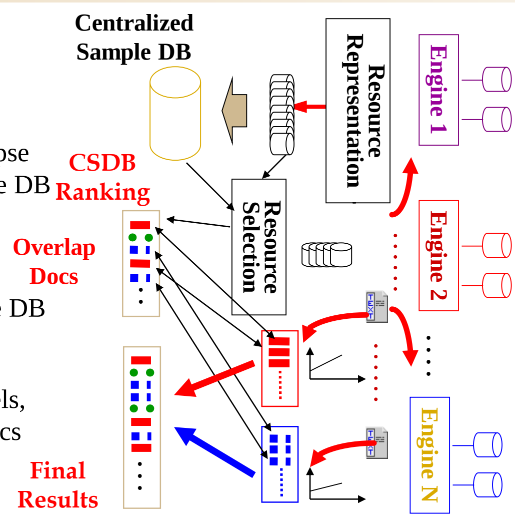
- Build representations by QBS, collapse sampled docs into centralized sample DB

In resource selection:

- Rank sources, calculate centralized scores for docs in centralized sample DB

In results merging:

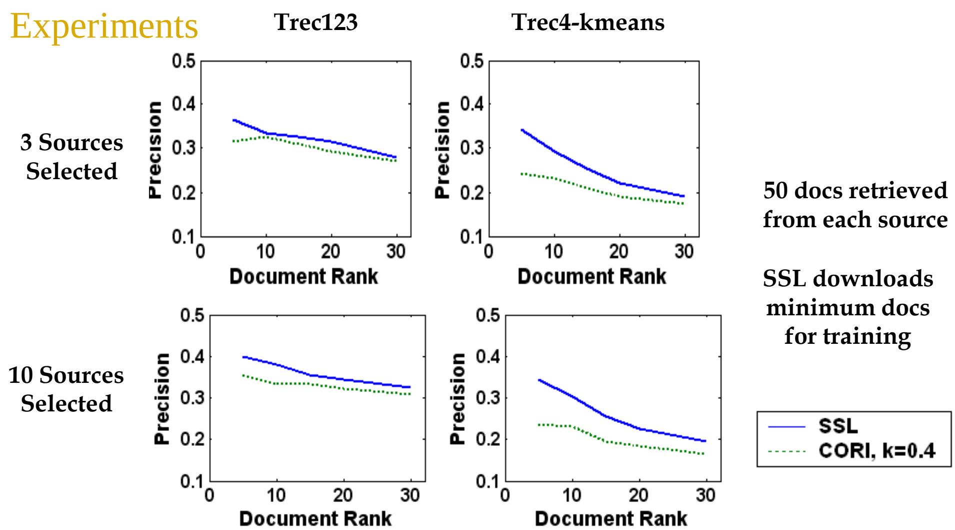
- Find overlap docs, build linear models, estimate centralized scores for all docs



PURDUE UNIVERSITY
Department of Computer Science

Research Problems (Results Merging)

Experiments



PURDUE UNIVERSITY
Department of Computer Science

More on Federated Search

- Search Result Diversification (Hong&Si SIGIR’13)
- Problem: Lack of diversity in results
  - E.g., several copies of the same document
- Key contribution: Metric
  - Need to be able to measure diversity
- Builds on ReDDE and others

46

Base: R-Metric

- Ranking algorithm independent metric
  - Based on top, or ranked list, of documents
- $R_k = \frac{\sum_{i=1}^{k} E_i}{\sum_{i=1}^{k} B_i}$
  - E_i is relevant documents in source *i* according to algorithm *E*
  - B_i is true relevant documents in source *i*
- Basic idea: Replace “Relevant” with a diversity metric

47

Diversity

- Query has multiple *aspects*
  - Evaluate each aspect separately
  - Remember something like this?
  - *Macro vs. Micro F1*
- What is an aspect?
  - *Topic*

48