Federated Text Search

Luo Si

Department of Computer Science
Purdue University
Abstract

Outline

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

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

Hidden Web contained in (Hidden) information sources that provide text search engines to access the hidden information
Federated Search

U.S. EPA Publications Search

U.S. Securities and Exchange Commission

National Center for Biotechnology Information

Protein Interaction Search

Supernova Cosmology Project

Astronomy Picture of the Day Searchable Archive

PubMed

National Library of Medicine

U.S. Securities and Exchange Commission

Search the United States Code

My.ED.gov

U.S. Department of Education

United States Patent and Trademark Office

Federal Business Opportunities

Federal Business Daily

U.S. Department of Commerce

GOV

Federal Business Daily

FirstGov.gov

The U.S. Government's Official Web Portal

ARC

Archival Research Catalog

NTIS

National Technical Information Service

United States

General Accounting Office

MMWR

Morbidity and Mortality Weekly Report

SCIPP

superNova Cosmology Project

USDA

Science for a Changing World

American Memory

Historical Collections for the National Digital Library

NSF

U.S. National Science Foundation
Introduction

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

Valuable

Search by
Federated Search

Federated Search Environments:
Small companies: Probably cooperative information sources
Big companies (organizations): Probably uncooperative information sources
Web: Uncooperative information sources
Federated Search

Components of a Federated Search System and Two Important Applications

(1) Resource Representation

(2) Resource Selection

(3) Results Merging

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
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
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…

Require Thorough Simulation
Introduction

Modeling Federated Search

TREC data

- Large text corpus, thorough queries and relevance judgments
- Often be divided into $O(100)$ information sources
- Professional well-organized contents
- 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)
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 “ln.csv” 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
Abstract

Outline

• Introduction to federated search

• Main research problems
  - Resource Representation
  - Resource Selection
  - Results Merging
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
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
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 \( \text{AER} = \frac{|N - N^*|}{N^*} \)

<table>
<thead>
<tr>
<th></th>
<th>Estimated Source Size</th>
<th>Actual Source Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trec123 (Avg AER, lower is better)</td>
<td>(0.729)</td>
<td>(0.232)</td>
</tr>
<tr>
<td>Trec123-10Col (Avg AER, lower is better)</td>
<td>(0.943)</td>
<td>(0.299)</td>
</tr>
</tbody>
</table>

Cap-Recapture

Sample-Resample

Collapse every 10\(^{th}\) source of Trec123
Abstract

Outline

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

  
  DTF causes large human judgment costs

- **Lightweight probes** (Hawking & Thistlewaite, 1999)
  
  Acquire training data in an online manner, large communication costs
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
Research Problems (Resource Selection)

Language Model Resource Selection

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

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
Relevant Doc Distribution Estimation (ReDDE) Algorithm

$$\text{Rel}_Q(i) = \sum_{d \in \text{db}_i} P(\text{rel}|d) * P(d|\text{db}_i) * N_{\text{db}_i}$$

$$\approx \sum_{d \in \text{db}_i_{\text{samp}}} P(\text{rel}|d) * SF_{\text{db}_i}$$

$$P(\text{rel}|d) = \left\{ \begin{array}{ll} C_Q & \text{if Rank}_{\text{CCDB}}(Q, d) < \text{ratio} * \sum_i N_{\text{db}_i} \\ 0 & \text{otherwise} \end{array} \right.$$
Research Problems (Resource Selection)

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)

Experiments

On testbeds with uniform or moderately skewed source sizes

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

Evaluated Ranking

Desired Ranking
Research Problems (Resource Selection)

Experiments

On testbeds with skewed source sizes

![Graphs showing R Value vs Num of Selected Sources for Relevant and Nonrelevant categories with different methods (ReDDE, CORI, KL).]
Abstract

Outline

- Introduction to federated search
- Main research problems
  - Resource Representation
  - Resource Selection
  - Results Merging
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
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
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
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
Research Problems (Results Merging)

Experiments

3 Sources Selected

Trec123

Trec4-kmeans

10 Sources Selected

50 docs retrieved from each source

SSL downloads minimum docs for training

3 Sources Selected

50 docs retrieved from each source

SSL downloads minimum docs for training