Federated Search

Outline

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

  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
Language Model Resource Selection

\[ P(db_i | Q) = \frac{P(Q | db_i) \cdot P(db_i)}{P(Q)} \]

\[ P(Q | db_i) = \prod_{q \in Q} \left( \lambda P(q | db_i) + (1 - \lambda) P(q | G) \right) \]

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

\[ P(C_i) = \frac{N_{C_i}}{\sum_j N_{C_j}} \]

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)

Relevant Doc Distribution Estimation (ReDDE) Algorithm

\[
\text{Rel}_Q(i) = \sum_{d \in \text{db}} P(\text{rel}|d) \cdot P(d|\text{db}_i) \cdot N_{\text{db}_i} \\
\approx \sum_{d \in \text{db}_{\text{samp}}} P(\text{rel}|d) \cdot SF_{\text{db}_i}
\]

Rank on Centralized Complete DB

\[
P(\text{rel}|d) = \begin{cases} 
C_Q & \text{if } \text{Rank}_{\text{CCDB}}(Q,d) < \text{ratio} \cdot \sum_{i} N_{\text{db}_i} \\
0 & \text{otherwise}
\end{cases}
\]

Problem: To estimate doc ranking on Centralized Complete DB

“Everything at the top is (equally) relevant”

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

Centralized Sample DB

Resource Representation

Engine 1

Engine 2

Engine N

CSDB Ranking

CCDB Ranking

Threshold

© 2016 Christopher W. Clifton

4
Research Problems (Resource Selection)

Experiments

On testbeds with uniform or moderately skewed source sizes

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

Evaluated Ranking

Desired Ranking

On testbeds with skewed source sizes
Federated Search

Outline

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

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

Experiments
Trec123
3 Sources Selected
50 docs retrieved from each source
SSL downloads minimum docs for training

Trec4-kmeans
10 Sources Selected
SSL
CORI, k=0.4
Final Project

• Self-directed final project
  – You must decide what to do

• First step: Proposal
  – What is the problem?
  – How is it solved today?
  – What is your approach? Why should it work?
  – How long will it take? (Milestones)
  – What is your measure for success?
  – Deliverables

Final Project: Ideas

• Identify an unsolved (or poorly solved) problem
  – Try a new solution

• Take an existing approach and try to improve it

• Compare existing approaches

• “Reproducibility”: Validate existing work
  – Does it hold in different conditions / data?
Final Project: Deliverables

Dependent on the project
- Written report describing outcomes / experiments
- (Taped) oral presentation describing outcomes
  – Include system demonstration?
- System that can be tried out
  – Runs on SSLab machines
  – Web accessible
- Other ideas?

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
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

Diversity

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