The Effectiveness of Query Expansion for Distributed Information Retrieval

Paul Ogilvie, Jamie Callan
Carnegie Mellon University
{pto,callan}@cs.cmu.edu

Outline
- Previous Work
- Distributed Information Retrieval and Query Expansion
- Sampled Information in Single Database IR
- Sampled Information in Distributed IR
- Conclusions

Previous Work
- Xu introduced local context analysis for automatic query expansion [Xu & Croft, 1996]
- In single database setting, boost of 20% to average precision
- In multiple database setting, with complete information, boosts of 25% to 40% for precision at 20 documents.

Unrealistic Assumptions in Prior Work
- Query expansion used complete information
  - Complete information = All documents in all databases
- Cooperative environment assumed for Distributed IR
  - Cooperative environment = Complete information is available
    - Document scores from different databases are directly comparable

Questions About Previous Work
- What happens when using partial information for query expansion?
  - Single database IR
  - Distributed IR
- Partial information = Incomplete information, a sample of documents from each database

Questions About Previous Work
- Does query expansion work in an uncooperative environment for Distributed IR?
  - Uncooperative environment =
    - Complete information is not available
    - Partial information is available
    - Scores given by different databases are not directly comparable
Outline

- Previous Work
- Distributed Information Retrieval and Query Expansion
- Sampled Information in Single Database IR
- Sampled Information in Distributed IR
- Conclusions

Distributed Information Retrieval

- Decide where to search
- Search (possibly search several places)
- Merge results returned by different searches

Resource Description

- All documents from database
  - Works in cooperative environments
  - Gives the best description of the databases contents
- Collect a portion of the database
  - Useful when in an uncooperative environment
  - Approximates what is in a database
- Use query-based sampling
  - Random 1-word queries
  - Gather a documents from each database
  - Generates fairly representative descriptions

Query-Based Sampling

- Resource description
- Resource/collection selection
  - Search a database about databases to choose which ones are best
  - Each resource description is a "document"
- Results merging

Resource Merging

- Use global idf statistics
  - Works in cooperative distributed IR
  - Simulates single database scores
- Use normalized scores
  - Used in uncooperative environments
  - Traditional InQuery merging
    \[
    R_i' = \frac{R_i - R_{min}}{(R_{max} - R_{min})} \left( 1 + \frac{D_i' - D_{min}}{D_{max} - D_{min}} \right)
    \]
  - Simulate single database scores
    \[
    D' = D' + 0.4 \cdot D' \cdot R'
    \]
  - Use normalized scores
    \[
    D' = \frac{D' + 0.4 \cdot D' \cdot R'}{1.4}
    \]
Query Expansion in Distributed Information Retrieval

Where does the passage database come from?

Our hypothesis:
- We already gather documents during resource description
- We can use these same documents to build the passage database

Questions Raised
- How many documents do we need to expand queries effectively?
- Does query-based sampling choose good documents for query expansion?
- How well does query expansion work in distributed information retrieval?
  - cooperative environment
  - uncooperative environment

Test Environment
- TREC CDs 1,2,3
  - About 3 GB of text
- Split into 100 databases
  - by source
  - database contents differ from db to db
    - Associated Press
    - Department of Energy Abstracts
    - Federal Registrar Rules and Regulations
    - Wall Street Journal
    - ...
- TREC topics 51-150

Query Expansion
Local Context Analysis [Xu & Croft, 1996]:
- Automatic query expansion
- Uses a database of n-word passages
- To expand a query:
  1. Retrieve top x passages
  2. Select best terms
  3. Add terms to query

Questions Raised
- How many documents do we need to expand queries effectively?
- Does query-based sampling choose good documents for query expansion?
- How well does query expansion work in distributed information retrieval?
  - cooperative environment
  - uncooperative environment

Resource Merging

Database Ranking

DOCNO D
ZF207-965-790
ZF207-965-790
ZF207-965-790
...

Search Results
DOCNO D
AP880212-0001 0.5432
AP880212-0016 0.5103
AP880212-0097 0.4069
...

Merged Results
DOCNO D'
AP880212-0001 0.4145
ZF207-965-790 0.3241
AP880212-0016 0.3123
...

Test Environment
- TREC CDs 1,2,3
  - About 3 GB of text
- Split into 100 databases
  - by source
  - database contents differ from db to db
    - Associated Press
    - Department of Energy Abstracts
    - Federal Registrar Rules and Regulations
    - Wall Street Journal
    - ...
- TREC topics 51-150

Query Expansion in
Distributed Information Retrieval

Expand Queries
Decide where to search
Search and merge results
Sampled Information in Single Database IR

How many documents do we need to expand queries effectively?

- Single database setting
- Partial information for query expansion
- Representative sampling
- Take every \( n \)th document for the passage database – documents in NIST order

Sampled Information in Single Database IR

Every \( n \)th Sampling for Query Expansion

Average Precision

<table>
<thead>
<tr>
<th># docs</th>
<th>No QE</th>
<th>( n=4 )</th>
<th>( n=8 )</th>
<th>( n=16 )</th>
<th>( n=32 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>TITLE</td>
<td>0.1759</td>
<td>0.2144</td>
<td>0.2133</td>
<td>0.2058</td>
<td>0.1958</td>
</tr>
<tr>
<td></td>
<td>+21.9%</td>
<td>+22.2%</td>
<td>+16.9%</td>
<td>+11.3%</td>
<td></td>
</tr>
<tr>
<td>DESCRIPTION</td>
<td>0.1659</td>
<td>0.2112</td>
<td>0.1885</td>
<td>0.1898</td>
<td>0.1784</td>
</tr>
<tr>
<td></td>
<td>+24.4%</td>
<td>+11.1%</td>
<td>+11.8%</td>
<td>+5.0%</td>
<td></td>
</tr>
</tbody>
</table>

Sampled Information in Single Database IR

Does query-based sampling select good documents for query expansion?

- Single database setting
- Partial information for query expansion
- Query-based sampling for passage database

Sampled Information in Single Database IR

Query-Based Sampling for Query Expansion

Average Precision

<table>
<thead>
<tr>
<th># docs/db</th>
<th>No QE</th>
<th>2900</th>
<th>1450</th>
<th>725</th>
<th>362</th>
</tr>
</thead>
<tbody>
<tr>
<td>TITLE</td>
<td>0.1759</td>
<td>0.2124</td>
<td>0.2096</td>
<td>0.2025</td>
<td>0.1910</td>
</tr>
<tr>
<td></td>
<td>+20.7%</td>
<td>+19.1%</td>
<td>+15.1%</td>
<td>+8.5%</td>
<td></td>
</tr>
<tr>
<td>DESCRIPTION</td>
<td>0.1659</td>
<td>0.2018</td>
<td>0.1909</td>
<td>0.1828</td>
<td>0.1700</td>
</tr>
<tr>
<td></td>
<td>+18.9%</td>
<td>+12.4%</td>
<td>+7.7%</td>
<td>+0.1%</td>
<td></td>
</tr>
</tbody>
</table>

Sampled Information in Single Database IR

- Partial information can be used for query expansion in single database retrieval
- Every \( n \)th sampling works at \( n=16 \) or about 68k documents for Title and Description queries.
- Query-based sampling works almost as well as every \( n \)th sampling.
- Query-based sampling works at 725 documents per database or 73k documents.
Outline

- Previous Work
- Distributed Information Retrieval and Query Expansion
- Sampled Information in Single Database IR
- Sampled Information in Distributed IR
- Conclusions

Sampled Information in Distributed IR

Cooperative Distributed IR with Partial Information for Query Expansion
Precision at 20 docs

<table>
<thead>
<tr>
<th># docs/db</th>
<th>No QE</th>
<th>2900</th>
<th>1450</th>
<th>725</th>
</tr>
</thead>
<tbody>
<tr>
<td>TITLE</td>
<td>0.3810</td>
<td>0.4263</td>
<td>0.4190</td>
<td>0.4052</td>
</tr>
<tr>
<td></td>
<td>+11.8%</td>
<td>+9.3%</td>
<td>+6.3%</td>
<td></td>
</tr>
<tr>
<td>DESCRIPTION</td>
<td>0.4085</td>
<td>0.4031</td>
<td>0.3745</td>
<td>0.3522</td>
</tr>
<tr>
<td></td>
<td>-1.3%</td>
<td>-8.3%</td>
<td>-12.7%</td>
<td></td>
</tr>
</tbody>
</table>

Sampled Information in Distributed IR

Uncooperative Distributed IR with Partial Information for Query Expansion
Precision at 20 docs

<table>
<thead>
<tr>
<th># Docs per db</th>
<th>2900</th>
<th>1450</th>
<th>725</th>
</tr>
</thead>
<tbody>
<tr>
<td>TITLE</td>
<td>0.3959</td>
<td>0.4040</td>
<td>0.3955</td>
</tr>
<tr>
<td></td>
<td>+2.0%</td>
<td>+1.5%</td>
<td>+0.8%</td>
</tr>
<tr>
<td>DESCRIPTION</td>
<td>0.4167</td>
<td>0.3585</td>
<td>0.4063</td>
</tr>
<tr>
<td></td>
<td>-13.9%</td>
<td>-17.6%</td>
<td>-20.3%</td>
</tr>
</tbody>
</table>

Sampled Information in Distributed IR

- Boosts to Title queries in cooperative environments
- In uncooperative environments, query expansion does poorly
- No improvements in collection selection

Conclusions

- Explains previous work of Xu
  - Query expansion can give boosts to Title queries in cooperative distributed environments
  - Global ids for merging
  - Query expansion does not work well in uncooperative distributed environments
  - Global ids not available
- Sampled information can be used for query expansion
- Query-based sampling does well for building a passage database
# Sampled Information in Distributed

Merging with Global *idf* and Sampled Info for Collection Selection

Precision at 20 docs

<table>
<thead>
<tr>
<th># Docs</th>
<th>2900</th>
<th>1450</th>
<th>725</th>
</tr>
</thead>
<tbody>
<tr>
<td>TITLE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No QE</td>
<td>0.3926</td>
<td>0.4238</td>
<td>0.3865</td>
</tr>
<tr>
<td>W/ QE</td>
<td>0.3926</td>
<td>0.4174</td>
<td>0.3819</td>
</tr>
<tr>
<td>DESC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No QE</td>
<td>0.3918</td>
<td>0.3979</td>
<td>0.3838</td>
</tr>
<tr>
<td>W/ QE</td>
<td>0.3918</td>
<td>0.3740</td>
<td>0.3839</td>
</tr>
</tbody>
</table>

*Note*: The table above shows the precision at 20 docs for different numbers of documents (2900, 1450, 725) for both title and description sections. The columns labelled 'No QE' and 'W/ QE' represent the precision without and with quality estimation, respectively. The percentage changes indicate the improvement or deterioration in precision with quality estimation.