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

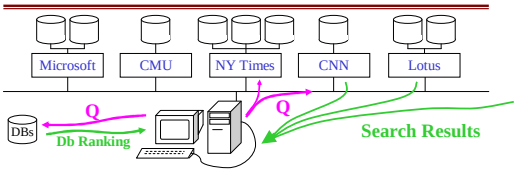
Questions About Previous Work

- 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

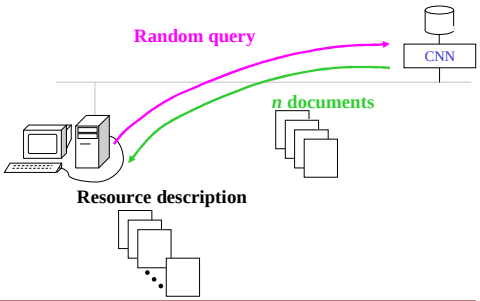
Distributed Information Retrieval

- 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 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 *d* documents from each database
    - Generates fairly representative descriptions

Query-Based Sampling



Resource Merging

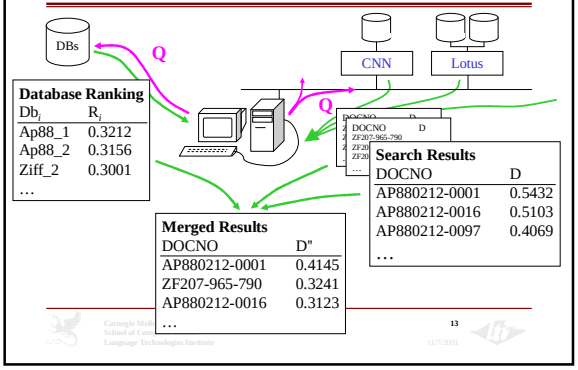
- Use global *idf* statistics
  - Works in cooperative distributed IR
  - Simulates single database scores
- Use normalized scores
  - Used in uncooperative environments
  - Traditional InQuery merging

$$R_i = (R_i - R_{\min}) / (R_{\max} - R_{\min})$$

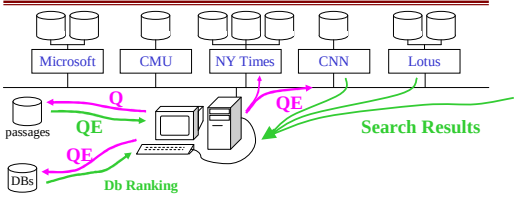
$$D' = (D - D_{\min_i}) / (D_{\max_i} - D_{\min_i})$$

$$D'' = \frac{D' + 0.4 \cdot D' \cdot R'}{1.4}$$

Resource Merging



Query Expansion in Distributed Information Retrieval



- Expand Queries
- Decide where to search
- Search and merge results

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

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

Outline

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

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

	No QE	n=4	n=8	n=16	n=32
# docs		270k	135k	68k	34k
TITLE	0.1759	0.2144 +21.9%	0.2133 +21.2%	0.2058 +16.9%	0.1958 +11.3%
DESCRIPTION	0.1659	0.2112 +24.4%	0.1885 +11.1%	0.1898 +11.8%	0.1784 + 5.0%

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

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

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

# docs/db	No QE	2900	1450	725
TITLE	0.3810	0.4263 +11.8%	0.4190 + 9.9%	0.4052 + 6.3%
DESCRIPTION	0.4085	0.4031 - 1.3%	0.3745 - 8.3%	0.3522 -13.7%

Sampled Information in Distributed IR

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

# Docs	2900		1450		725	
per db	No QE	W/ QE	No QE	W/ QE	No QE	W/ QE
TITLE	0.3959	0.4040 + 2.0%	0.3915	0.3856 - 1.5%	0.3950	0.3599 - 8.8%
DESC	0.4167	0.3585 -13.9%	0.4063	0.3345 -17.6%	0.4037	0.3222 -20.1%

Sampled Information in Distributed IR

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

Outline

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

Conclusions

- Explains previous work of Xu
  - Query expansion can give boosts to Title queries in cooperative distributed environments
    - Global idfs for merging
  - Query expansion does not work well in uncooperative distributed environments
    - Global idfs 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

# Docs	2900		1450		725	
per db	No QE	W/ QE	No QE	W/ QE	No QE	W/ QE
TITLE	0.3926	0.4238 + 7.9%	0.3865	0.4174 + 7.9%	0.3819	0.3992 + 4.5%
DESC	0.3918	0.3979 + 1.5%	0.3838	0.3740 - 2.5%	0.3839	0.3675 - 4.2%