Distributed Query-Sampling: A Quality-Conscious Approach

James Caverlee  Georgia Tech
Ling Liu  Georgia Tech
Joonsoo Bae  Chonbuk Natl U.
Why is Sampling Useful?

Surface Web ➔ Crawl and Index Everything

Web Index
Why is Sampling Useful?

Deep Web ➔
Crawl and Index
Everything?? NO!

92,000 TB vs. 167 TB on surface

Web Index

Caverlee, Liu, Bae: Distributed Query Sampling

SIGIR '06
How Do We Sample?

- **Query-Based Sampling**
  - [Callan et al. TOIS '01, SIGMOD '99]

1. Select Query
2. Issue Query
3. Collect Sample Docs
4. Update Query Dictionary
Key Challenges for...
Popular Solution - Uniform Sampling

Total # of sample docs = 10,000
Total # of databases = 5

Sample Size

2,000 2,000 2,000 2,000 2,000
Uniform Sampling Problems

- **Size differences**
  - A has 1,000 docs; B has 1 million

- **Diversity of information**
  - B has basketball info; C has basketball, football, baseball, ...

- **Duplicate (or near duplicate) databases**
  - D and E are mirrors

- ...

**Key Ideas:**
- Databases vary in quality
- Adaptively allocate resources to reflect this quality
Our Solution: Quality-Conscious Adaptive Sampling

1. Use fraction of total resources to collect seed samples from each database
2. Use collected samples to estimate quality of each database
3. Generate a new sampling allocation based on relative quality
4. Collect additional quality-conscious samples
5. Goto 2 until resources exhausted

Iterate
Adaptive Sampling: Simple Example

10,000 sample docs total; 5 databases

Caverlee, Liu, Bae: Distributed Query Sampling
Challenges for Adaptive Sampling

1. How do we estimate database quality?
   - We describe three schemes in the paper.

2. How do we tell if one scheme is better than another?

3. How should we divide the total resources between the seed and adaptive phases?

4. For how many iterations should we adaptively sample?
1. How Do We Estimate Database Quality?

- Proportional Document Ratio [PD]
- Proportional Vocabulary Ratio [PV]
- Vocabulary Growth [VG]
Sampling Scheme 1: Proportional Document Ratio

- Idea: Collect the same document proportion from each database

How do we estimate $|D|$??

$\Rightarrow$ sample-resample

[Si & Callan, SIGIR 2003]

This estimates ideal sample size for each database – need to scale to account for seed sampling [see paper]

<table>
<thead>
<tr>
<th>Uniform:</th>
<th>2,000</th>
<th>2,000</th>
<th>2,000</th>
<th>2,000</th>
<th>2,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>PD:</td>
<td>200</td>
<td>1,000</td>
<td>2,000</td>
<td>5,000</td>
<td>1,800</td>
</tr>
</tbody>
</table>
Sampling Scheme 2: Proportional Vocabulary Ratio

- Idea: Collect the same vocabulary proportion from each database
  - ratio_{PV}
  - E.g., 10% of the vocabulary terms at D_1, D_2, ..., D_n
- Favors databases with large vocabularies

- Challenge:
  - Each database will have different size vocabularies and different growth rates of vocabulary
  - Need to estimate vocabulary size and growth rate
Sampling Scheme 2: Proportional Vocabulary Ratio

How big should our sample be to collect ratio_{PV} = 50% of vocabulary terms?

Caverlee, Liu, Bae: Distributed Query Sampling

SIGIR '06
Challenge: Estimate Vocabulary Size

\[ |V| = K \times (\text{text\_size})^\beta \]

\[ = K \times (\text{avg\_doc\_size} \times |D|)^\beta \]

But these parameters are unknown!
Rely on seed sample to estimate:

- \( \text{avg\_doc\_size} \) \( \rightarrow \) straightforward
- \( |D| \) \( \rightarrow \) use sample-resample [Si & Callan, SIGIR 2003]

K and Beta? curve fitting to estimate

Caverlee, Liu, Bae: Distributed Query Sampling
Challenge: Estimate Sample Size

- **n equations like:**
  - \( \text{ratioPV} \times |V| = K \times (\text{avg_doc_size} \times \text{sample_size})^B \)

- **Solve for `sample_size`**
  - \( \text{sample_size} = e^{\text{... ratioPV ...}} \)
  - But `ratioPV` is still unknown

- **Vary `ratioPV` until we have:**
  - \( N = \text{sample_size}(A) + \text{sample_size}(B) + \ldots \)
Sampling Scheme 3: Vocabulary Growth

- Favors databases with a fast growing vocabulary, regardless of total size

Database A

Vocab

| Docs | |V_A| |
|------|---|---|

Database B

Vocab

| Docs | |V_B| |
|------|---|---|

Expected Marginal Vocabulary Terms

<table>
<thead>
<tr>
<th>Doc</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>180</td>
</tr>
<tr>
<td>2</td>
<td>80</td>
<td>140</td>
</tr>
<tr>
<td>3</td>
<td>60</td>
<td>110</td>
</tr>
<tr>
<td>4</td>
<td>35</td>
<td>85</td>
</tr>
<tr>
<td>5</td>
<td>25</td>
<td>65</td>
</tr>
</tbody>
</table>

Caverlee, Liu, Bae: Distributed Query Sampling

SIGIR '06
2. How Do We Evaluate Sampling Schemes?

- **Compare unigram language models for Sample Docs vs. Complete Database**
  - Quality of term weights
    - Weighted Common Terms
    - Jensen-Shannon Divergence
  - Quality of term rankings
    - **Spearman** rank correlation coefficient

- **Application scenario**
  - Distributed IR - Database Selection
  - Measure Recall @ 1, 2, ..., 20
Experimental Setup

- **Data**
  - TREC 123
    - 100 databases, by source and publication date
    - Avg docs ~11,000
  - TREC 4
    - 100 databases, organized by topic
    - Avg docs ~5,700
  - TREC 123 Large Databases
    - Aggregated several smaller dbs
    - Range from 45,000 docs to 242,000 docs

- **Sampling**
  - Stopwords removed, Porter’s Stemmer
  - Indexed & Searched by Lucene
  - Max 4 docs per query
  - Initial queries selected from UNIX dictionary; subsequent queries selected randomly weighted by term frequency
Vocabulary Size Estimation Error

![Graph showing Relative Error vs Docs Sampled for TREC123, TREC4, and TREC123 (Large).](image)
Sample Quality - Spearman

Total Sample Docs = 300 * # of DBs

½ Total Sample Docs Used For Seed Sampling

See Paper For More
Total Collection Vocabulary Size

- **TREC4**
- **TREC123**
- **TREC123-A**
- **TREC123-B**
- **TREC123-C**

Legend:
- VG
- U
- PV
- PD
Higher-quality samples should support smarter database selection.
Best case: biggest DBs are most relevant to query mix

Recall vs. Baseline Ranking [by # of relevant docs]

Databases Selected

Coverlee, Liu, Bae
Database Selection Recall TREC123B

Worst case: biggest DBs are least relevant to query mix
Contributions

- An adaptive distributed query-sampling framework that is quality-conscious for extracting high-quality text database samples

- Relies self-configuring ability based on the overall quality of all text databases under consideration

- Three quality-conscious sampling schemes for estimating database quality

- Higher-quality document sampling over multiple metrics compared to existing approaches
Future Work

- Alternative types of quality
  - Novelty, Freshness, Topic-sensitivity

- Augment with other info
  - Web: link data, access patterns
  - P2P: peer longevity, reputation

- Non-uniform sampling costs
Thank You!

caverlee@cc.gatech.edu
www.caverlee.com