Distributed Information Retrieval

1. Resource Representation
2. Resource Selection
3. Results Merging
Resource Representation

Query → Resource

Documents ← Resource

Query-Based Sampling
Summary

Full Resource Content vs. **Approximate**
Resource Representation

**Kullback-Leibler Divergence**

\[
KL(S || R) = \sum_{w \in S \cup R} p_S(w) \log \frac{p_S(w)}{p_R(w)}
\]

Smoothed Language Models
Outline

Introduction - Distributed Information Retrieval

Motivation - Evolving Resources

Updating Resource Representation - Existing Approaches

Ideas for Future Work
Content Changes Over Time

- $R_t$ - resource at time $t$
- $KL(R_{t-1} || R_t)$ - Rate of change
- $R_0$ - initial resource at time 0
- $KL(R_0 || R_t)$ - Overall change

Resource content changes over time
DIR Resource Representation

- Resource representation of $R_0$
- Resource content changes over time
- $KL(S_0 \parallel R_t)$ increases with $t$
- Resource representation becomes outdated
Evolving Resources in DIR

**DIR Retrieval Accuracy**

- If **old** representation is used for retrieval...
- accuracy decreases with time
- Old representation may work for recurring events
- If **up-to-date** representation is used...
- accuracy does not change

---

**Graph**

- P@5 - old
- P@10 - old
- P@5 - fresh
- P@10 - fresh

**Axes**

- X-axis: Crawl
- Y-axis: Precision (CO=3)
Summary

- Resource content changes over time
- Resource representation becomes outdated
- **Retrieval accuracy degrades**
- Need to keep resource representation up-to-date
Outline

Introduction - Distributed Information Retrieval

Motivation - Evolving Resources

Updating Resource Representation - Existing Approaches
  Basic Process
  Updating Policies
  Modeling Content Changes

Ideas for Future Work
Updating Resource Representation

- **Old resource representation**
- Use query-based sampling
- Download $n$ documents from a resource
- Add documents to the resource representation
- **Current resource representation**
Many resources - $M$

Limited bandwidth - $N$ documents can be downloaded at a time jointly from all resources

$n_i$ - number of documents downloaded form a resource $i$

$\sum_{i=1}^{M} n_i = N$
Milad Shokouhi, Mark Baillie, Leif Azzopardi
Updating Collection Representations for Federated Search
Policies

- $\sum_{i=1}^{M} n_i = N$
- Uniform: $n_i = N \cdot \frac{1}{M}$
- Popularity: $n_i = N \cdot \frac{\rho_i}{\sum_{i=1}^{M} \rho_i}$
- Size: $n_i = N \cdot \frac{S_i}{\sum_{i=1}^{M} S_i}$
Results

- Precision is stable
- Size-based is the best
- Uniform is the worst

**Figure 5**: The P@5 and P@10 values produced by running 448 queries on different collection representations in each crawl. The baseline corresponds to the 100-doc baseline. This would indicate that by using an updating policy, the representation sets reflect the content of the collection, with those documents still contained in the representation set but not the collection increasing the KL divergence. As expected, the updating policies display a consistent estimation close to the old-baseline. In comparison, the three updating policies display a consistent estimation close to the old-baseline. In comparison, the three updating policies display a consistent estimation close to the old-baseline.
Discussion

+ Keep retrieval accuracy stable over time
+ Size-based policy is the most effective

- Size should be re-estimated periodically $\implies$ additional bandwidth is needed
- One-fits-all updating period $\implies$ some representations are updated even if underlying resources do not change
- Content of different resources may change with different speed $\implies$ rate of change should be considered
Panagiotis G. Ipeirotis, Alexandros Ntoulas, Junghoo Cho and Luis Gravano
Modeling and Managing Content Changes in Text Databases
Modeling Content Changes

1. Model
   - Content changes when $KL(R_O||R_C) > \tau$
   - Survival function $S(t) = Pr[T > t]$
   - $S(t)$ depends on linear combination of $\tau$, log Size and
     $$\Delta KL = \frac{\sum_{t=1}^{t_{train}} KL(S_{t-1}||S_t)}{t_{train}}$$

2. Optimality problem
   - $\max \sum_{i=1}^{M} S_i(t)$, with the constraint $\sum_{i=1}^{M} n_i = N$

3. Optimal solution with Lagrange-multiplier method
Results

- $KL(S_O || R_C)$ increases
- $KL(S_C || R_C)$ is stable
Discussion

+ **Models** content changes
+ Rate of change of the resource’s content ($\Delta KL$) is considered as well as the size
+ Resource representation is kept up-to-date

− Rate of change $\Delta KL$ is considered to be linear, but this is not always the case
− Survival function $S(t)$ can only be used for updating a resource representation, while a general content changes model may be used also for Resource Selection and Results Merging
Outline

Introduction - Distributed Information Retrieval

Motivation - Evolving Resources

Updating Resource Representation - Existing Approaches

Ideas for Future Work
  - More Updating Policies
  - More Content Changes Modeling
Preface

TREC collections with timestamps (FR, AP, WSJ)
Documents are added in chronological order

Cumulative - $KL(R_0 \| R_t)$
Marginal - $KL(R_{t-1} \| R_t)$
<table>
<thead>
<tr>
<th>Introduction - Distributed Information Retrieval</th>
<th>More Updating Policies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motivation - Evolving Resources</td>
<td>More Content Changes Modeling</td>
</tr>
<tr>
<td>Updating Resource Representation - Existing Approaches</td>
<td></td>
</tr>
<tr>
<td>Ideas for Future Work</td>
<td></td>
</tr>
</tbody>
</table>

More Updating Policies
Best Existing Method

Size-based updating policy: \( n_i = N \cdot \frac{S_i}{\sum_{i=1}^{M} S_i} \)

- Size should be re-estimated periodically
- One-fits-all updating period
- Content of different resources may change with different speed \( \Rightarrow \) rate of change should be considered
"Rate of Change"-based Method

\[ n_i(t) = N \cdot \frac{\Delta KL_i(t)}{\sum_{i=1}^{M} \Delta KL_i(t)} \]

\[ \Delta KL_i(t) = KL(R_{t-1} \parallel R_t) \]
Discussion

- Size should be re-estimated periodically
- \( \Delta KL_i(t) = KL(R_{t-1} || R_t) \) should be estimated based on resource representation
- One-fits-all updating period
- Content of different resources may change with different speed \( \Rightarrow \) rate of change should be considered
More Content Changes Modeling
Existing Modeling Approach

1. Model changes with survival function $S(t) \sim \text{Size}, \Delta KL$
2. State optimality problem: $\max \sum_{i=1}^{M} S_i(t)$.
3. Solve it with Lagrange-multiplier method

- **Rate of change** $\Delta KL$ is considered to be linear
- Survival function $S(t)$ can only be used for updating a resource representation
Rate of Change

Fit rate of change into a curve: $\Delta KL \simeq f(t)$

$KL(R_{t-1} || R_t)$

Documents are added in chronological order
Modeling Content Changes

1. Model changes with $f(t)$
2. State optimality problem: $\min \sum_{i=1}^{M} F_i(t)$, where $F_i(t) = \int_{t} f_i(t)$
3. Solve it with Lagrange-multiplier method: $f(t) = F'(t)$
Problems

- Shape of \( f(t) \)
- \( f(t) \) depends on:
  - Initial size of a resource - can be estimated based on resource representation
  - Number of documents changed in each time unit - can be estimated based on resource representation
  - Training data \( \Delta KL(t) \) for \( t \in [0..t_{\text{train}}] \) - needs to be estimated based on resource representation
Discussion

+ Model content changes explicitly
+ Find optimal resource representation updating periods
+ Predict $\Delta KL$ at any point in time $t$
+ Additional information for Resource Selection and Results Merging

- Rate of change $\Delta KL$ is considered to be linear
- Survival function $S(t)$ can only be used for updating a resource representation

- Problems stated above
Summary

- Content of resources changes over time
  - Retrieval accuracy of DIR systems degrades
- Resource representation updating techniques
  - Size-based updating policy
  - Modeling with survival analysis
- Future work - utilize changes in content
  - Estimate based on resource representation
  - Fit into a curve
  - Apply to Resource Selection and Results Merging
Answers