Spoken document retrieval using neighboring documents and extended language models for query likelihood model

Gifu University

Kazuaki Ogawa, Tatsuaki Murahashi, Hiroaki Taguchi, Kodai Nakajima, Masanori Takehara, Satoshi Tamura, Satoru Hayamizu
Contents

1. Overview
2. SDR based on query likelihood model
3. Proposed smoothing methods
4. Experiments
5. Conclusions
6. Future works
Contents

1. Overview
2. SDR based on query likelihood model
3. Proposed smoothing methods
4. Experiments
5. Conclusions
6. Future works
Overview

• Our methods are based on the **query likelihood model** and **smoothing**.

• We use extended language models and neighboring documents.

• We propose three smoothing methods.
  A) Using a language model obtained from **research papers**.
  B) Using a **cache model** and an **N-gram model**.
  C) Using **neighboring documents**.
Background

• A vector space model is used widely in document retrieval.

• However, the vector scape model is not robust to retrieve spoken document.
  ➢ Including speech recognition errors.

• **Probabilistic models** are employed.
  ➢ We focus on a query likelihood model.

• The query likelihood model needs smoothing methods.
  ➢ We use Dirichlet smoothing.
Contents

1. Overview
2. SDR based on query likelihood model
3. Proposed smoothing methods
4. Experiments
5. Conclusions
6. Future works
Query likelihood model

• A query likelihood model $P(d|q)$:

$$P(d|q) = \frac{P(q|d)P(d)}{P(q)} \propto P(q|d)$$

$P(q|d)$ is a probability to generate a query $q$ under the condition that a document $d$ is found.

• Employing a unigram language model $\theta_d$:

$$P(q|\theta_d) = \prod_{w_i \in V} P(w_i|\theta_d)^{C(w_i,q)}$$

$$P(w_i|\theta_d) = \frac{C(w_i,d)}{|d|}$$
Dirichlet smoothing

- In the query likelihood model, the Dirichlet smoothing is used to avoid the zero probability problem.

\[
P(w_i | \theta_d; \mu) = \frac{|d|}{|d| + \mu} P(w_i | \theta_d) + \frac{\mu}{|d| + \mu} P(w_i | \theta_C)
\]

- The model \( \theta_C \) is a language model for a static document collection \( C \).
Extended Dirichlet smoothing

• The model $\theta_C$ cannot deal with terms that do not appear in a static document collection $C$.

• Some of us proposed to use a dynamic document collection $W$ obtained from web pages.

$$P(w_i | \theta_d; \mu, \nu) = \frac{|d|}{|d| + \mu + \nu} P(w_i | \theta_d) + \frac{\mu}{|d| + \mu + \nu} P(w_i | \theta_C) + \frac{\nu}{|d| + \mu + \nu} P(w_i | \theta_W)$$
Contents

1. Overview
2. SDR based on query likelihood model
3. Proposed smoothing methods
4. Experiments
5. Conclusions
6. Future works
Proposed smoothing methods

- Overview of proposed smoothing methods.

(A.) Dirichlet smoothing
- Relative frequency
- Target documents
- Unigram

(B.) Contextual information
- Neighboring documents
- Cache model and N-gram model

(C.) External information
- Research papers
- Unigram

Gifu University
(A.) Research paper corpus

- Web documents have a lot of noises.

- We propose to use a research paper corpus in ASJ instead of web documents.

\[
P(w_i | \theta_d; \mu, \nu) = \frac{|d|}{|d| + \mu + \nu} P(w_i | \theta_d) + \frac{\mu}{|d| + \mu + \nu} P(w_i | \theta_c) + \frac{\nu}{|d| + \mu + \nu} P(w_i | \theta_R)
\]

- We select papers based on cosine similarities of TF-IDF vectors between the papers and each query.
(B.) Cache model and $N$-gram model

• A cache model is based on local nature of terms:
  - Preceding terms are likely to be used again.

$$P_{CH}(w_i|M) = \frac{1}{|M|} \sum_{w_j \in M} \delta(w_i, w_j)$$

- $M$: $|M|$ terms appeared just before a target document $d$.

• $M$ can be replaced to terms in a target document $d$.

$$P_{CH}(w_i|d) = \frac{1}{|d|} \sum_{w_k \in d} \delta(w_i, w_k)$$

• The model used both $P_{CH}(w_i|M)$ and $P_{CH}(w_i|d)$.

$$P_{CH}(w_i|M, d) = \frac{1}{|M| + |d|} \left\{ \sum_{w_j \in M} \delta(w_i, w_j) + \sum_{w_k \in d} \delta(w_i, w_k) \right\}$$

• Typically, the cache model is used for linear interpolation in an $N$-gram model.
(B.) Cache model and $N$-gram model

- We use linear interpolation of a cache model and an $N$-gram model in the Dirichlet smoothing.

$$P(w_i | \theta_d; \mu, \nu) = \frac{|d|}{|d| + \mu + \nu} P(w_i | \theta_d) + \frac{\mu}{|d| + \mu + \nu} P(w_i | \theta_C) + \frac{\nu}{|d| + \mu + \nu} P(w_i | \theta_{KC})$$

$$P(w_i | \theta_{KC}) = (\gamma P_{KN}(w_i | \theta_{KN}) + (1 - \gamma) P_{CH}(w_i | M, d))$$

- $P_{CH}(w_i | M, d)$: A probability based on a cache model.
- $P_{KN}(w_i | \theta_{KN})$: A probability based on an $N$-gram model.
- $\theta_{KN}$: An $N$-gram model for a target document $d$.
- We employ the $N$-gram model using Kneser-Ney smoothing.
- We train the $N$-gram model using $d$ and $M$. 
(C.) Neighboring documents

- The length of target documents is often short.

- We propose to use neighboring documents of a target document.

\[
S'(i) = \sum_{n=-L}^{L} w_n S(i + n)
\]

\[
w_n = \frac{1}{|n| + 1}
\]

- \(S(i)\): A similarity score of a partial document corresponding to \(i\)-th slide and a query.
- \(w_i\): A weighting coefficient of the inverse proportion.
Contents

1. Overview
2. SDR based on query likelihood model
3. Proposed smoothing methods
4. Experiments
5. Conclusions
6. Future works
Experimental condition (Formal-run)

<table>
<thead>
<tr>
<th>Task</th>
<th>SpokenQuery&amp;Doc-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sub task</td>
<td>SQ-SCR SGS retrieval</td>
</tr>
<tr>
<td>Query</td>
<td>K-REF-WORD-MATCH</td>
</tr>
<tr>
<td>Target</td>
<td>K-REF-WORD-MATCH</td>
</tr>
<tr>
<td>The number of queries</td>
<td>Dry-run:35</td>
</tr>
<tr>
<td>Static document collection</td>
<td>Target documents</td>
</tr>
<tr>
<td>Dynamic document collection</td>
<td>Research papers in ASJ (published 2005 to 2014)</td>
</tr>
</tbody>
</table>
Experimental condition (Formal-run)

- Experimental condition (Formal-run)

| The parameters $\mu$ and $\nu$ in the Dirichlet smoothing | Method 1 and 4: $\mu = 320$  
Method 2 and 5: $\mu = 320, \nu = 80$  
Method 3 and 6: $\mu = 320, \nu = 10$ |
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>The parameter $</td>
</tr>
<tr>
<td>The parameter $N$ in the $N$-gram model</td>
</tr>
<tr>
<td>The parameter $\gamma$ for linear interpolation</td>
</tr>
</tbody>
</table>

- We tested our following retrieval methods using NTCIR-12 Dry-run data.
NTCIR-12 Formal-run evaluation

• We tested our following six retrieval methods using NTCIR-12 Formal-run data.

2. Our Method 1 + using the unigram model from research papers.
3. Our Method 1 + using the linear interpolation with the cache model and the $N$-gram model.
4. Our Method 1 + using neighboring documents.
5. Our Method 2 + using neighboring documents.
NTCIR-12 Formal-run evaluation

- **MAP results:**

<table>
<thead>
<tr>
<th>Method</th>
<th>(A.) Research papers</th>
<th>(B.) Cache model and $N$-gram model</th>
<th>(C.) Neighboring documents</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 1</td>
<td></td>
<td></td>
<td></td>
<td>0.197</td>
</tr>
<tr>
<td>Method 2</td>
<td>○</td>
<td></td>
<td></td>
<td>0.193</td>
</tr>
<tr>
<td>Method 3</td>
<td></td>
<td>○</td>
<td></td>
<td>0.215</td>
</tr>
<tr>
<td>Method 4</td>
<td></td>
<td></td>
<td>○</td>
<td>0.242</td>
</tr>
<tr>
<td>Method 5</td>
<td>○</td>
<td></td>
<td>○</td>
<td>0.239</td>
</tr>
<tr>
<td>Method 6</td>
<td></td>
<td>○</td>
<td>○</td>
<td>0.252</td>
</tr>
</tbody>
</table>

- All retrieval methods are based on the query likelihood model using the Dirichlet smoothing.
Discussion (1)

<table>
<thead>
<tr>
<th>Method</th>
<th>(A.) Research papers</th>
<th>(B.) Cache model and $N$-gram model</th>
<th>(C.) Neighboring documents</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 1</td>
<td></td>
<td></td>
<td></td>
<td>0.197</td>
</tr>
<tr>
<td>Method 2</td>
<td>○</td>
<td></td>
<td></td>
<td>0.193</td>
</tr>
<tr>
<td>Method 3</td>
<td></td>
<td>○</td>
<td></td>
<td>0.215</td>
</tr>
<tr>
<td>Method 4</td>
<td></td>
<td></td>
<td>○</td>
<td>0.242</td>
</tr>
<tr>
<td>Method 5</td>
<td>○</td>
<td></td>
<td>○</td>
<td>0.239</td>
</tr>
<tr>
<td>Method 6</td>
<td></td>
<td>○</td>
<td>○</td>
<td>0.252</td>
</tr>
</tbody>
</table>

• Comparing Method 1 - 3 to 4 - 6: Methods using **neighboring documents** were **better** than the other methods.
### Discussion (2)

<table>
<thead>
<tr>
<th></th>
<th>(A.) Research papers</th>
<th>(B.) Cache model and $N$-gram model</th>
<th>(C.) Neighboring documents</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 1</td>
<td></td>
<td></td>
<td></td>
<td>0.197</td>
</tr>
<tr>
<td>Method 2</td>
<td>○</td>
<td></td>
<td></td>
<td>0.193</td>
</tr>
<tr>
<td>Method 3</td>
<td></td>
<td>○</td>
<td></td>
<td>0.215</td>
</tr>
<tr>
<td>Method 4</td>
<td></td>
<td></td>
<td>○</td>
<td>0.242</td>
</tr>
<tr>
<td>Method 5</td>
<td>○</td>
<td></td>
<td></td>
<td>0.239</td>
</tr>
<tr>
<td>Method 6</td>
<td></td>
<td>○</td>
<td>○</td>
<td>0.252</td>
</tr>
</tbody>
</table>

- Comparing Method 1, 4 to 2, 5: The MAP score using research papers slightly **decreased**.
## Discussion (3)

<table>
<thead>
<tr>
<th>Method</th>
<th>(A.) Research papers</th>
<th>(B.) Cache model and N-gram model</th>
<th>(C.) Neighboring documents</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 1</td>
<td></td>
<td></td>
<td></td>
<td>0.197</td>
</tr>
<tr>
<td>Method 2</td>
<td>○</td>
<td></td>
<td></td>
<td>0.193</td>
</tr>
<tr>
<td>Method 3</td>
<td></td>
<td>○</td>
<td></td>
<td>0.215</td>
</tr>
<tr>
<td>Method 4</td>
<td></td>
<td></td>
<td>○</td>
<td>0.242</td>
</tr>
<tr>
<td>Method 5</td>
<td>○</td>
<td></td>
<td>○</td>
<td>0.239</td>
</tr>
<tr>
<td>Method 6</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>0.252</td>
</tr>
</tbody>
</table>

- We obtained the best result in Method 6.
  - Using **contextual information** has effectiveness for SDR.
We obtained the best result in Method 6.

However, using the cache model and the $N$-gram model has **similar effects** as using neighboring documents.

- We need to improve integration scheme in our approach.

- Using **contextual information** has effectiveness for SDR.
Contents

1. Overview
2. SDR based on query likelihood model
3. Proposed smoothing methods
4. Experiments
5. Conclusions
6. Future works
Conclusions

• We proposed three techniques for SDR.
  A) Using a research paper corpus.
  B) Using a cache model and an $N$-gram model.
  C) Using neighboring documents.

• Experiments were conducted using NTCIR-12 Formal-run data sets.
  ➢ It turns out that using contextual information is important for SDR.
Future works

• We should properly deal with **unknown words**.
  ➢ For example, when we use external information.

• We try to reconsider **weighting methods** for using neighboring documents.

• It is necessary to investigate smoothing parameters.

• We should investigate more effective **integration schemes** for our proposed smoothing methods.


Thank you for your attention.