Spoken document retrieval using extended query model and web documents

Kiichi Hasegawa, Masanori Takehara, Satoshi Tamura, Satoru Hayamizu
Gifu University

Our approach

- Our basic framework is a query model.
- Two types of extension:
  - 1) One is to use web documents to expand the corpus as dynamic documents.
  - 2) The other is to use a topic model (LDA) to estimate similarities between web documents and the corpus in the test collection.
- These two extensions are incorporated in a smoothing formula with Dirichlet smoothing.

Query model

- The probabilities where q is a given query and d is a document.
  \[ P(d|q) = \frac{P(q|d)P(d)}{P(q)} \propto P(q|d) \]
- In language modeling of multinomial model, each term is assumed to be independent.
  \[ P(q|\theta_q) = \prod_{w_{i} \in V} P(w_{i} | \theta_{d})^{C(w_{i}, q)} \]
- Relative frequency of each term:
  \[ P(w_{i} | \theta_{d}) = \frac{C(w_{i}, q)}{|d|} \]

Smoothing using dynamic documents

- Dynamic documents are web pages obtained according to given queries.
- Dirichlet smoothing is extended as follows:
  \[ P(w_{i} | \theta_{d}, \mu, \nu) = \frac{|d| + \mu + \nu}{|d| + \mu + \nu} \cdot P(w_{i} | \theta_{d}) + \frac{\mu}{|d| + \mu + \nu} \cdot P(w_{i} | \theta_{w}) + \frac{\nu}{|d| + \mu + \nu} \cdot P(w_{i} | \theta_{w}) \]
- where \( P(w_{i} | \theta_{w}) \) is for the dynamic documents (web pages) and \( \mu \) and \( \nu \) are the smoothing parameters.

Weighting method

- Weighted score is used for probability of the web page which is extracted by the query.
- Its weight is average similarity between the web page and documents in the collection.

Similarity by topic mixture

- Similarity between a document and a web page is defined as cosine distance between topic mixture ratio vectors.
  \[ \gamma = (\gamma_{1}, \gamma_{2}, \ldots, \gamma_{T}) \]
- For each document, a topic mixture ratio vector (topic proportion) is estimated using LDA with the parameters \( \alpha_{i}, \beta_{i} \) from the whole document collection.
- For a web page, a topic mixture ratio vector is estimated using the same parameters \( \alpha_{i}, \beta_{i} \).
- Finally, cosine distance between two vectors are calculated as the similarity measure.

Graphical model of LDA

- \( K \) is the number of topics
- \( N \) is the number of web pages
- for each \( d \), \( \theta_{d} \) is the topic distribution
- \( \phi_{w} \) is the word distribution
- Dirichlet parameters \( \alpha, \beta \)

Experiments

- Experimental setup
- SpokenDoc-2 SCR subtask in NTCIR-10
- Sub-task: Lecture retrieval
- Spoken document: Ref-Word-Matched
- LDA training data: Mainichi newspaper corpus (2007–2008)
- Web search engine: Yahoo! API

NTCIR-9 Dry-run results

<table>
<thead>
<tr>
<th>TF-IDF</th>
<th>Query model</th>
<th>+ web page</th>
<th>+ LDA weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.366</td>
<td>0.369</td>
<td>0.371</td>
<td>0.372</td>
</tr>
</tbody>
</table>

NTCIR-10 Formal-run results

<table>
<thead>
<tr>
<th>Table 2. Results for NTCIR-10 SpokenDoc-2 Formal-run.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query model + LDA (RunID L36)</td>
</tr>
<tr>
<td>Query model + Web (RunID L37)</td>
</tr>
<tr>
<td>Query Expansion (RunID L38)</td>
</tr>
</tbody>
</table>

Note: since queries in NTCIR-10 Formal-run were longer than those in NTCIR-9 Dry-run, it seemed that more related and informative web pages were obtained.

Conclusion

- Our spoken document retrieval method uses the language model approach.
- We extend query model in two ways.
- One is to use web page retrieval for dynamic document collection.
- The other is to employ a topic model (Latent Dirichlet allocation) for the measure between documents and retrieved web pages.
- We expand the Dirichlet smoothing for dynamic documents and the topic model.
- We showed improvements at NTCIR-9 Dry-run and NTCIR-10 Formal-run.