Spoken document retrieval using extended query model and web documents

Kiichi Hasegawa, Masanori Takehara, Satoshi Tamura, Satoru Hayamizu, Gifu University
Overview of this talk

1. Language modeling approach
   - Query model, and Smoothing
2. Smoothing using dynamic documents
   - Weighting web pages for query model
3. Topic modeling for spoken documents
   - Latent Dirichlet Allocation
4. Experiments
   - Dry-run and formal-run
5. Conclusion
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.

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

- Relative frequency of each term;

$$P(w_i|\theta_d) = \frac{C(w_i, q)}{|d|}$$
The Dirichlet smoothing is given by:

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

with a parameter \( \mu \) and,

the probability for all the collection \( P(w_i | \theta_c) \)

For a long document, the smoothing effect becomes smaller.
Smoothing using dynamic documents

- Dynamic documents are web pages obtained according to given queries.
- Dirichlet smoothing is extended as follows:

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

- where \( P(w_i|\theta_W) \) is for the dynamic documents (web pages) and \( \mu \) and \( \nu \) are the smoothing parameters.
LDA (latent Dirichlet allocation)

LDA posits that each document is a mixture of topics, and that each word's creation is attributable to one of the document's topics.

from David M. Blei, KDD2011, tutorial
LDA as a graphical model

$$\prod_{i=1}^{K} p(\beta_i | \eta) \prod_{d=1}^{D} p(\theta_d | \alpha) \left( \prod_{n=1}^{N} p(z_{d,n} | \theta_d) p(w_{d,n} | \beta_{1:K}, z_{d,n}) \right)$$

- Proportions parameter
- Per-document topic proportions
- Per-word topic assignment
- Observed word
- Topics
- Topic parameter

from David M. Blei, KDD2011, tutorial
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.

\[ P(w_i|\theta_W) = \frac{\sum_{j=1}^{|W|} \delta(p_j, C) \cdot C(w_i, p_j)}{\sum_{j=1}^{|W|} \sum_{k=1}^{N_j} \delta(p_j, C) \cdot C(w_k, p_j)} \]

where

\[ \delta(p, C) = \frac{1}{|C|} \sum_{m=1}^{|C|} \delta(p, d_m) \]

\[ C = \{d_1, d_2, ..., d_{|C|}\} \]

p: extracted web page
\(d_m\): m-th document
Similarity by topic mixture

- Similarity between a document and a web page is defined as cosine distance between topic mixture ratio vectors.
  \[ \gamma = (\gamma_1, \gamma_2, \ldots, \gamma_{|Z|}) \]

- For each document, a topic mixture ratio vector (topic proportion) is estimated using LDA with the parameters \( \alpha, \beta_k \) from the whole document collection.

- For a web page, a topic mixture ratio vector is estimated using the same parameters \( \alpha, \beta_k \).

- Finally, cosine distance between two vectors are calculated as the similarity measure.
Experiments

- Experimental setup
- SpokenDoc-2 SCR subtask in NTCIR-10
- Sub-subtask: Lecture retrieval
- Spoken document: Ref-Word-Matched
- LDA training data: Mainichi newspaper corpus (2007–2008)
- Web search engine: Yahoo! API
- Dynamic documents: 30 web pages per query
- Smoothing parameters: $\mu = 4000$, $\nu = 50$
NTCIR–9 Dry–run results

- Preliminary experiment by NTCIR–9 Dry–run.
- The score is the Mean Average Precision (MAP).

![Graph showing performance improvements with different models.](image_url)

- TF-IDF
- Query model
- + web page
- + LDA weight

Scores:

- 0.368
- 0.369
- 0.370
- 0.371
- 0.372
- 0.373
- 0.374
- 0.375
Table 2. Results for NTCIR–10 SpokenDoc–2 Formal–run.

- Query model + LDA (RunID L36) 0.408
- Query model + Web (RunID L37) 0.399
- Query Expansion (RunID L38) 0.372

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.
From teaching staff’s view

- NTCIR provides a very good opportunity for students to learn the knowledge and the evaluation skill on language technology.
- It is also good for them to have valuable experiences to be in a research community.
- In general, research purpose defines “what to measure”, and then “how to measure” provides research issues to be solved.
- It seems that future design of SpokeDoc tasks needs to be more balanced on the number of participants and diversity of research issues.