# 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**.

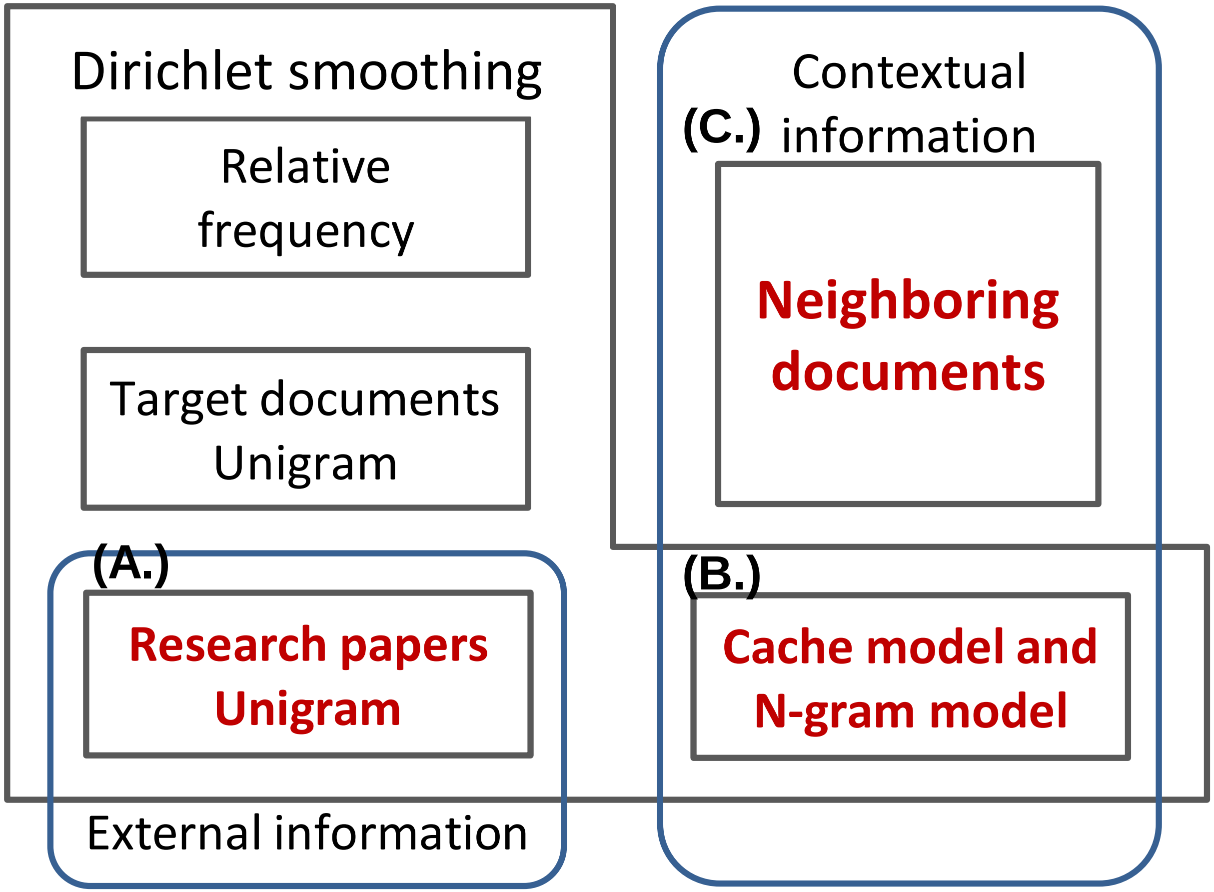
$$P(w_i|\theta_d;\mu,\nu) = \frac{|d|}{|d|+\mu+\nu}P(w_i|\theta_d) + \frac{\mu}{|d|+\mu+\nu}P(w_i|\theta_C) + \frac{\nu}{|d|+\mu+\nu}P(w_i|\theta_W)$$

# 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.) 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.

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

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

- $P_{CH}(w_i|M,d)$: 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)

- Experimental condition(Formal-run)

| Task | SpokenQuery&Doc-2 |
|---|---|
| Sub task | SQ-SCR SGS retrieval |
| Query | K-REF-WORD-MATCH |
| Target | K-REF-WORD-MATCH |
| The number of queries | Dry-run:35 ｜ Formal-run:80 |
| The number of target documents | 2807 |
| Static document collection | Target documents |
| Dynamic document collection | Research papers in ASJ<br>(published 2005 to 2014) |

# Experimental condition(Formal-run)

- Experimental condition(Formal-run)

| | |
|---|---|
| The parameters $\mu$ and $\nu$ in the Dirichlet smoothing | Method 1 and 4: $\mu = 320$<br>Method 2 and 5: $\mu = 320,\ \nu = 80$<br>Method 3 and 6: $\mu = 320,\ \nu = 10$ |
| The parameter $\|M\|$ in the cache model | $\|M\| = 100$ |
| The parameter $N$ in the $N$-gram model | $N = 5$ |
| The parameter $\gamma$ for linear interpolation | $\gamma = 0.25$ |

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

# NTCIR-12 Formal-run evaluation

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

1. Query-likelihood-model-based method using the Dirichlet smoothing.
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**.
6. Our Method 3 + using **neighboring documents**.

# NTCIR-12 Formal-run evaluation

- MAP results:

| | (A.) Research papers | (B.) Cache model and $N$-gram model | (C.) Neighboring documents | MAP |
|---|---|---|---|---|
| Method 1 | | | | 0.197 |
| Method 2 | ○ | | | 0.193 |
| Method 3 | | ○ | | 0.215 |
| Method 4 | | | ○ | 0.242 |
| Method 5 | ○ | | ○ | 0.239 |
| Method 6 | | ○ | ○ | 0.252 |

➢ All retrieval methods are based on the query likelihood model using the Dirichlet smoothing.

# Discussion (1)

| | (A.) Research papers | (B.) Cache model and $N$-gram model | (C.) Neighboring documents | MAP |
|---|---|---|---|---|
| Method 1 | | | | 0.197 |
| Method 2 | ○ | | | 0.193 |
| Method 3 | | ○ | | 0.215 |
| Method 4 | | | ○ | 0.242 |
| Method 5 | ○ | | ○ | 0.239 |
| Method 6 | | ○ | ○ | 0.252 |

- Comparing Method 1 - 3 to 4 - 6:
  Methods using **neighboring documents** were **better** than the other methods.

# Discussion (2)

| | (A.) Research papers | (B.) Cache model and *N*-gram model | (C.) Neighboring documents | MAP |
|---|---|---|---|---|
| Method 1 | | | | 0.197 |
| Method 2 | ○ | | | 0.193 |
| Method 3 | | ○ | | 0.215 |
| Method 4 | | | ○ | 0.242 |
| Method 5 | ○ | | ○ | 0.239 |
| Method 6 | | ○ | ○ | 0.252 |

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

# Discussion (3)

| | (A.) Research papers | (B.) Cache model and $N$-gram model | (C.) Neighboring documents | MAP |
|---|---|---|---|---|
| Method 1 | | | | 0.197 |
| Method 2 | ○ | | | 0.193 |
| Method 3 | | ○ | | 0.215 |
| Method 4 | | | ○ | 0.242 |
| Method 5 | ○ | | ○ | 0.239 |
| Method 6 | | ○ | ○ | 0.252 |

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

# Discussion (3)

| | (A.) Research papers | (B.) Cache model and *N*-gram model | (C.) Neighboring documents | MAP |
|---|---|---|---|---|
| Method 1 | | | | 0.197 |
| Method 2 | ○ | | | 0.193 |

- **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.**

- We obtained the best result in Method 6.
  - 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.

# References

[1] K. Ichikawa, N. Kitaoka, S, Tsuge, K. Takeda, K. Kita.
"Improvement of Spoken Document Retrieval based on Various Text Retrieval Models".
IPSJ Journal, Vol.56, No.3, March 2015.

[2] H. Nanjo, T. Yoshimi, S. Maeda, and T. Nishio.
"Spoken Document Retrieval Experiments for SpokenQuery&Doc at Ryukoku University(RYSDT)".
In Proc. NTCIR-11, December 2014.

[3] K. Hasegawa, M. Takehara, S. Tamura, and S. Hayamizu.
"Spoken document retrieval using extended query model and web documents".
In Proc. NTCIR-10, pages 608-611, June 2013.

Thank you for your attention.