Segmented spoken document retrieval using word co-occurrence information

Kensuke Hara, Hiroaki Taguchi, Koudai Nakajima,
Masanori Takehara, Satoshi Tamura, Satoru Hayamizu

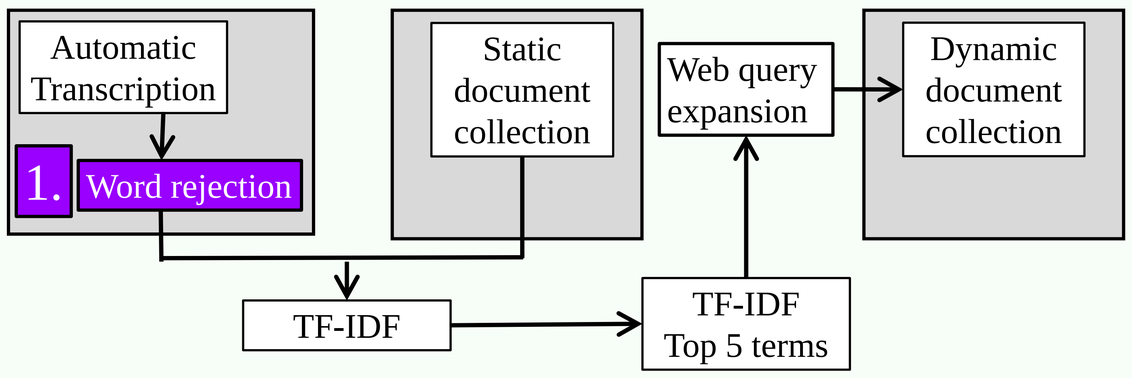
GIFU UNIVERSITY

Introduction

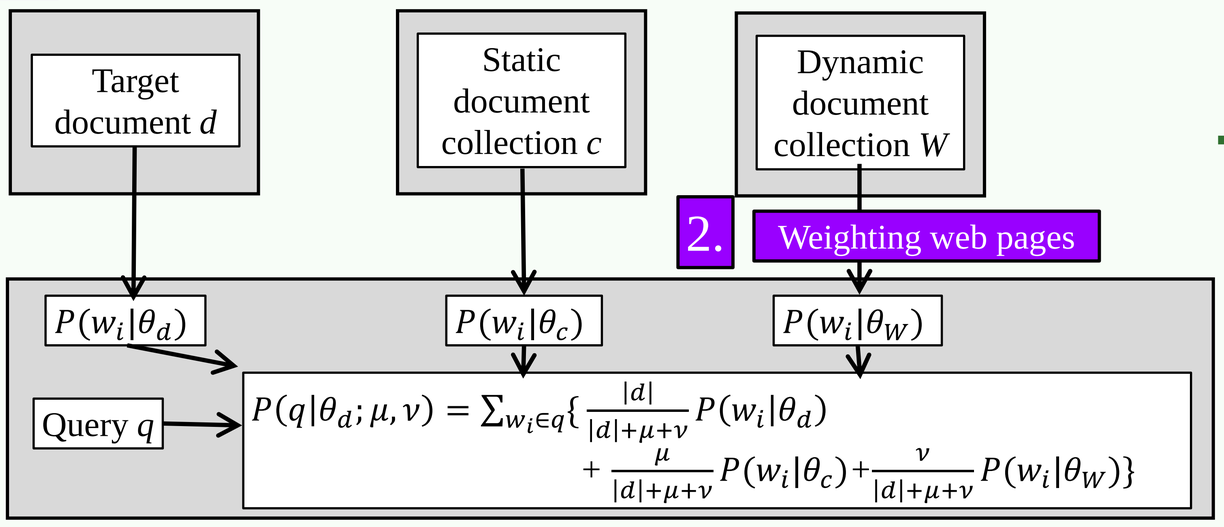
- Spoken document retrieval(SDR) is attracting attention in searching news shows and movies.
- In SDR, mis-recognized words have bad effects.
  ⇨ TF-IDF values of mis-recognized words sometimes become inappropriate.
- The cosine similarity is widely used for comparison.
  ⇨ But the cosine similarity treats words having the same meaning (e.g. ASR and speech recognition) as different ones.
- To overcome these issues, pointwise mutual information (PMI) is employed.
  - PMI represents a relationship between two words.
  - To reject mis-recognized words, PMI is used to compute a contextual coherency of a word.
  - For query-document comparison, PMI is used to consider the similarity of different words.

Flows of our proposed methods

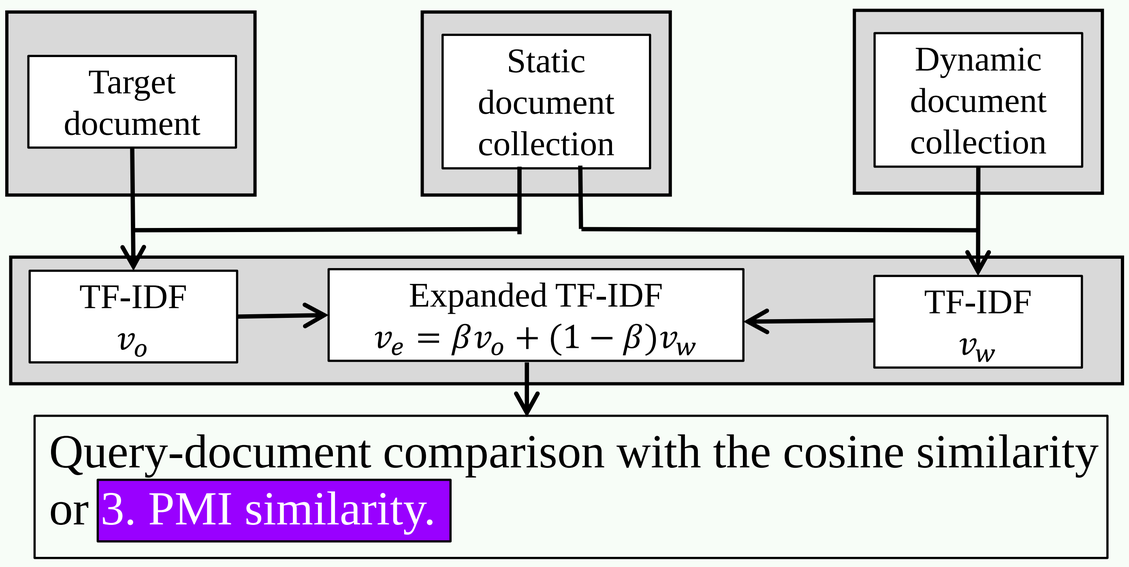
A. Building dynamic document collection



B. SDR using query model and LDA(Hasegawa,2012)



C. SDR using vector space model



Details of four techniques

1. Mis-recognized word rejection

- Mis-recognized words have less relationship to a document.
- The relationship between a word *w* and a document *d* is computed by *sumPMI(w)* used for rejection with a threshold.

$$sumPMI(w) = \sum_{w_i \in d} PMI(w, w_i)$$

2. Web page weighting using LDA

- A web-page weight is determined by the cosine distance between a query topic vector and a web page topic vector.
- A topic mixture ratio vector is computed using LDA.

3. Document comparison using PMI

- A word similarity $R(w_1, w_2)$ is computed using PMI.

$$R(w_1, w_2) = \begin{cases} 1 & (w_1 = w_2) \\ PMI(w_1, w_2) & (w_1 \neq w_2) \end{cases}$$

- It can consider a similarity of different words.

4. Segmented document retrieval method

- Linear combination of a similarity for a segment Sim_c and a similarity for a whole document Sim_d.

$$Sim = \alpha \cdot Sim_c + (1-\alpha) \cdot Sim_d$$

Experimental condition (Formal-run)

Subtask	Slide-Group-Segment retrieval
Automatic transcription(Query)	REF-WORD-MATCH
Automatic transcription(Target)	REF-WORD-MATCH
LDA training data	Mainichi news paper corpus 2007-2008
Static document collection	Manual transcription
Linear combination parameters α,β	α = 0.6, β = 0.9 (tuned in the dry-run evaluation)

Results and discussion

	model	1.	2.	3.	4.	MAP
Method 1	B. query model	○	○		○	0.161
Method 2			○			0.133
Method 3		○	○			0.114
Method 4	C. vector space model	○			○	0.143
Method 5		○		○	○	0.047

1. Rejecting mis-recognized words
3. Document comparison using PMI
   ····· not succeeded
   ⇨ PMI scores between general terms tend to be high.

4. Segmented SDR using whole documents ····· useful
   ⇨ Considering whole contents is important for segmented SDR.

2. SDR using query model and LDA ····· useful
   ⇨ The query model is useful for not only short text queries but also long spoken queries.