

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.

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

$$w_i \in V = \{w_1, w_2, \dots, w_{|V|}\}$$

- Relative frequency of each term;

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

Dirichlet smoothing

- 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 μ 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:

$$\begin{aligned} P(w_i | \theta_d, \mu, \nu) = & \frac{|d|}{|d| + \mu + \nu} \cdot P(w_i | \theta_d) \\ & + \frac{\mu}{|d| + \mu + \nu} \cdot P(w_i | \theta_c) \\ & + \frac{\nu}{|d| + \mu + \nu} \cdot P(w_i | \theta_W) \end{aligned}$$

- where $P(w_i | \theta_W)$ is for the dynamic documents (web pages) and μ and ν are the smoothing parameters.

LDA (latent Dirichlet allocation)

Topics

gene	0.04
dna	0.02
genetic	0.01
...	

life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

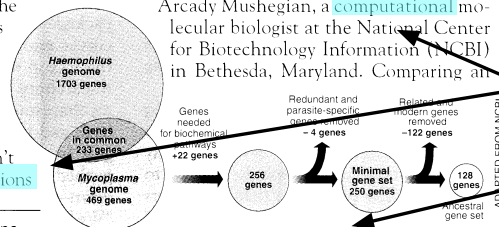
Documents

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many **genes** does an **organism** need to **survive**? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for **life**. One research team, using **computer** analyses to compare known **genomes**, concluded that today's **organisms** can be sustained with just 250 genes, and that the earliest life forms required a mere 128 **genes**. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those **predictions**

"are not all that far apart," especially in comparison to the 75,000 **genes** in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a **genetic numbers game**, particularly as more and more **genomes** are completely mapped and sequenced. "It may be a way of organizing any newly **sequenced genome**," explains Arcady Mushegian, a **computational** molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

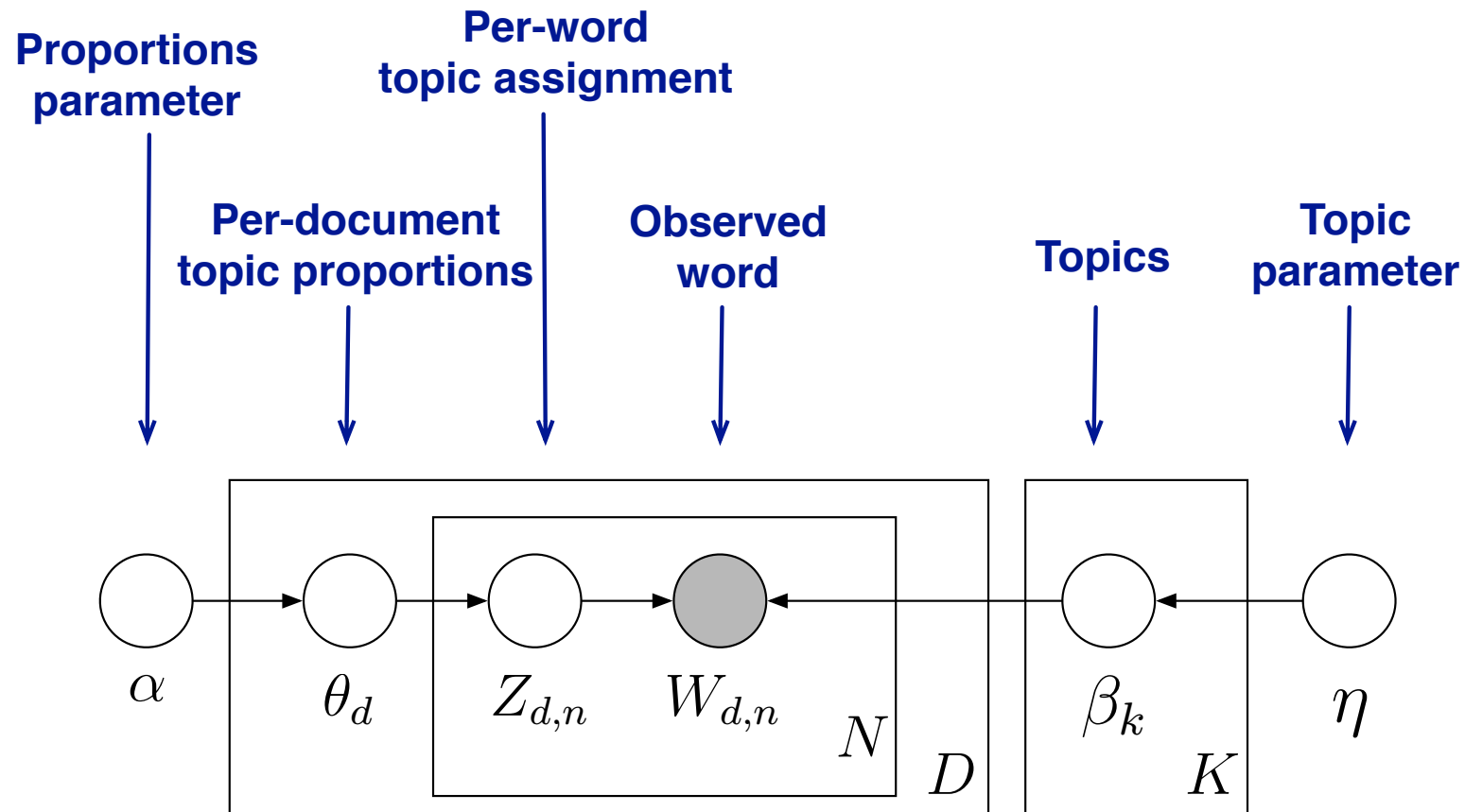
Topic proportions and assignments

- 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

graphical model of LDA

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



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

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

p: extracted web page
 d_m : m-th document

$$C = \{d_1, d_2, \dots, d_{|C|}\}$$

Similarity by topic mixture

- Similarity between a document and a web page is defined as cosine distance between topic mixture ratio vectors.

$$\gamma = (\gamma_1, \gamma_2, \dots, \gamma_{|Z|})$$

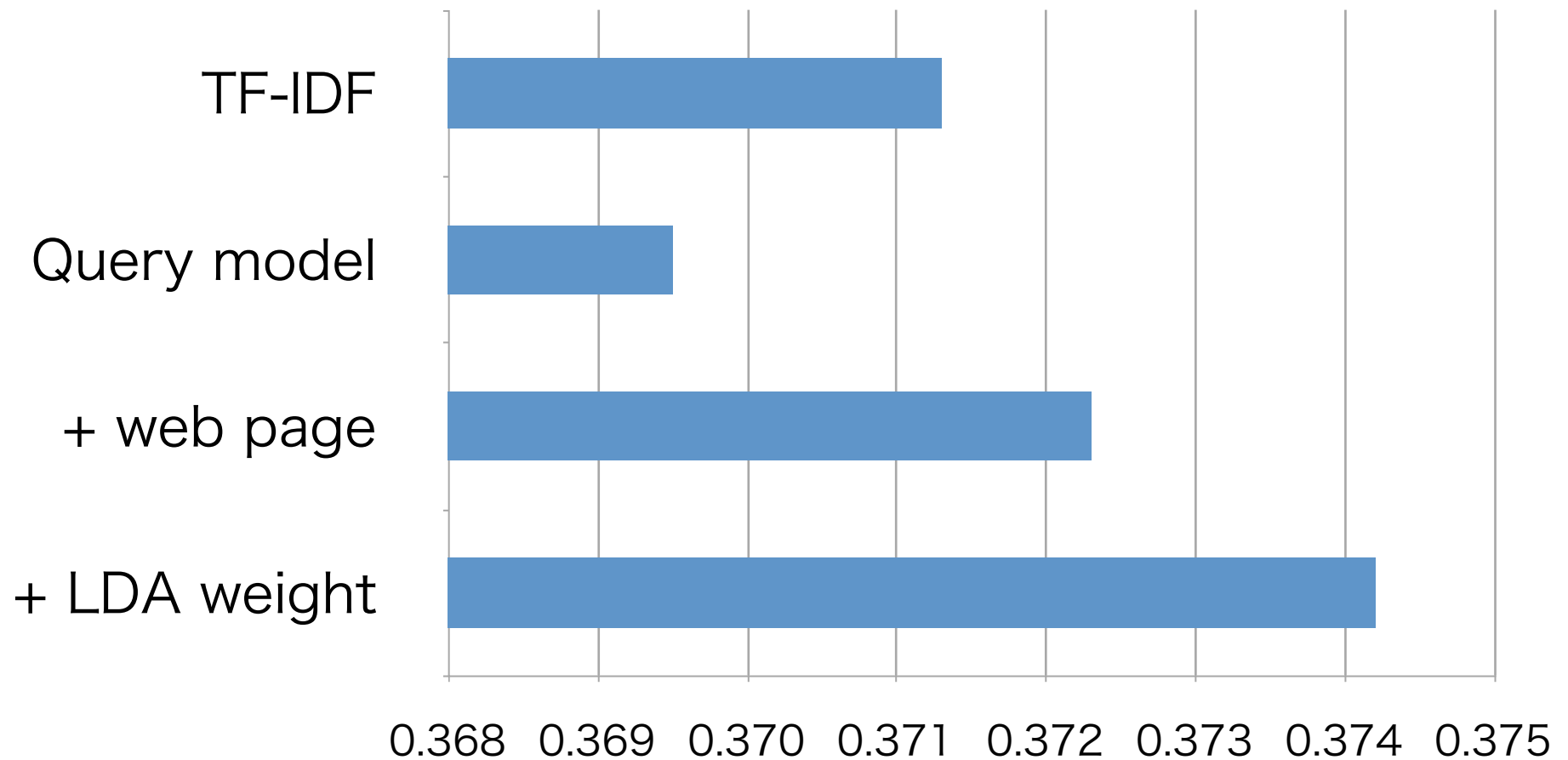
- For each document, a topic mixture ratio vector (topic proportion) is estimated using LDA with the parameters α, β_k from the whole document collection.
- For a web page, a topic mixture ratio vector is estimated using the same parameters α, β_k .
- Finally, cosine distance between two vectors are 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).



NTCIR-10 Formal-run results

■ 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.