Spoken document retrieval using extended query model and web documents

Kiichi Hasegawa
Gifu University
1-1 Yanagido Gifu
Gifu 501-1193 Japan
hasegawa@asr.info.gifu-u.ac.jp

Masanori Takehara
Gifu University
1-1 Yanagido Gifu
Gifu 501-1193 Japan
takehara@asr.info.gifu-u.ac.jp

Satoshi Tamura
Gifu University
1-1 Yanagido Gifu
Gifu 501-1193 Japan
tamura@info.gifu-u.ac.jp

Satoru Hayamizu
Gifu University
1-1 Yanagido Gifu
Gifu 501-1193 Japan
hayamizu@gifu-u.ac.jp

ABSTRACT

This paper proposes a novel approach for spoken document retrieval. In our method, a query model which is one of the probabilistic language models is adopted, in order to computes a probability to generate a given query from each document. We employ not only a “static” document collection consisting of targeted documents but also a “dynamic” document collection including web documents related with queries. We expand the query model so as to incorporate probabilities obtained from “static” and “dynamic” language models using the Dirichlet smoothing. Furthermore, in order to improve retrieved results, we develop a weighting method for web documents. Experiments using NTCIR-9 SpokenDoc Dry-run and NTCIR-10 SpokenDoc-2 Formal-run were conducted, and it is found our proposed scheme has enough performance compared with conventional methods.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Retrieval models.*

General Terms

Algorithms, Experimentation, Theory.

Keywords

Information retrieval, Query model, Web document, Dirichlet smoothing, Latent Dirichlet allocation (LDA).

Subtask : [SCR Subtask] [Lecture retrieval]

Team name : HYM

1. INTRODUCTION

This paper shows a novel approach for NTCIR-10 SpokenDoc-2 [1]. Since speech is one of the most common communication modalities, spoken document is increasingly attracting attention as developing information technologies. For example in a conference, it becomes usual to record speech data and transcribe them using speech recognition, instead of taking minutes manually. As spoken documents are rapidly increasing, however, one issue arises: how the spoken documents should be automatically retrieved. This task is called spoken document retrieval.

In the document retrieval literature, the TF-IDF score has been widely used. On the other hand, the TF-IDF-based method has several disadvantages for spoken document retrieval; TF-IDF is not suitable for spoken documents including recognition errors. If there are few words in the comparing document (just like a query), the TF-IDF method often chooses inappropriate documents.

In this paper, we try to apply a probabilistic model and build a logical framework for spoken document retrieval. As the probabilistic model, we focus on a query model. In the query model, a probability to generate a query is considered, and it is computed using a language model. A smoothing technique is often employed in the query model to deal with the zero probability problem. The conventional smoothing uses a “static” document collection which is obtained prior to retrieval. However, the document collection has a problem that it cannot deal with terms appeared in a query and never appeared in the collection.

This paper proposes a new modeling approach by expanding the query model mentioned above; in our model, not only the “static” document collection but also “dynamic” documents obtained from web pages are used in the Dirichlet smoothing. In addition, we also propose a weighting method to web pages; actually some web pages are important and informative, and some pages are not. Thus web page weighting is essential.

The rest of our paper is organized as follows: Section 2 describes a conventional query model and smoothing techniques. Our proposed approaches are introduced in Section 3. Experiments were conducted in Section 4. Finally Section5 concludes this paper.

2. LANGUAGE MODELING APPROACH

2.1 Query model

The document retrieval issue can be formulated by estimating a probability $\mathrm{P(d|q)}$, where q is a given query and d is a document. Applying the Bayesian theorem, this probability is calculated by:

$$\mathrm{P(d|q)} = \frac{\mathrm{P(q|d)p(d)}}{\mathrm{p(q)}} \qquad (1)$$

In Equation (1), the denominator $\mathrm{P(q)}$ can be treated as a constant because the query is not dependent on any document. And the probability $\mathrm{P(d)}$ can be ignored when no previous knowledge is available. Then Equation (1) is rewritten as:

$$P(d|q) \propto P(q|d) \quad (2)$$

In Equation (2), the query likelihood $P(q|d)$ means a probability to generate a query under the condition that a document is found. This probability can be then computed using a language model. This strategy is known as language modeling approach.

In the language modeling approach for information retrieval, a multinomial model is commonly used. In the multinomial model, each term is assumed to be independent of the other terms, and query likelihood $P(q|\theta_d)$ can be estimated by a unigram language model θ_d as:

$$P(q|\theta_d) = \prod_{w_i \in V} P(w_i|\theta_d)^{C(w_i,q)} \quad (3)$$

where $w_i \in V = \{w_1, w_2, \cdots, w_{|V|}\}$ is a term in a given query, and $C(w_i, q)$ is a count for w_i in q. As a result, the document retrieval is now equivalent to the issue to estimate $P(w_i|\theta_d)$.

The simplest model of $P(w_i|\theta_d)$ is accomplished by using relative frequency of each term, given by:

$$P(w_i|\theta_d) = \frac{C(w_i, d)}{|d|} \quad (4)$$

where $|d|$ means the total number of terms in d.

2.2 Smoothing

According to Equation (4), if the model θ_d has zero probability for a term w_i, the query likelihood $P(q|\theta_d)$ estimated by Equation (3) becomes zero. It means that we cannot estimate $P(q|\theta_d)$ appropriately. To avoid this, a smoothing technique is often employed for a language model. In a smoothing scheme, a non-zero probability is assigned to terms in the query, which do not appear in the document. The smoothing technique finally enables us to approximate the query likelihood, and compare the query with any targeted document.

2.2.1 Linear interpolation

One of the simplest ways for smoothing is linear interpolation. The linear interpolation is formulated by:

$$P(w_i|\theta_d; \lambda) = \lambda \cdot P(w_i|\theta_d) + (1 - \lambda) \cdot P(w_i|\theta_C) \quad (5)$$

where λ is a smoothing parameter $(0 \leq \lambda \leq 1)$. $P(w_i|\theta_C)$ is a collection model, equivalent to a unigram language model for a documents collection C. The unigram probability is obtained as:

$$P(w_i|\theta_C) = \frac{\sum_{d \in C} C(w_i, d)}{\sum_{d \in C} |d|} \quad (6)$$

2.2.2 Dirichlet smoothing

Dirichlet smoothing is another common smoothing technique. With a parameter μ, the Dirichlet smoothing is given by:

$$P(w_i|\theta_d; \mu) = \frac{|d|}{|d| + \mu} \cdot P(w_i|\theta_d) + \frac{\mu}{|d| + \mu} \cdot P(w_i|\theta_C) \quad (7)$$

In Equation (7), the parameter is determined according to the length of the document; for a long document the smoothing effect becomes smaller, and the effect becomes stronger for a short document. Since the document length is taken into account, the Dirichlet smoothing is a better interpolation scheme than the linear interpolation. Therefore in this paper, we adopt the Dirichlet smoothing technique.

3. PROPOSED METOD

3.1 Smoothing using dynamic documents

Regarding Equation (7), in general, a targeted document set is used as the document collection C. If there is any term in the given query that does not appear in the document set, however, the model θ_C cannot deal with the term. Therefore, we use another document set obtained from web pages as a "dynamic" collection W, in addition to the "static" document collection C. In order to combine the "static" and "dynamic" document collections, we extend the Dirichlet smoothing method. In our method, the probability $P(w_i|\theta_d)$ is obtained as:

$$\begin{aligned} P(w_i|\theta_d; \mu, \nu) &= \frac{|d|}{|d| + \mu + \nu} \cdot P(w_i|\theta_d) \\ &+ \frac{\mu}{|d| + \mu + \nu} \cdot P(w_i|\theta_C) \\ &+ \frac{\nu}{|d| + \mu + \nu} \cdot P(w_i|\theta_W) \end{aligned} \quad (8)$$

where $P(w_i|\theta_W)$ is a collection model built using the dynamic document collection W, and ν is another smoothing parameter for W. Figure 1 shows a graphical model of our proposed method.

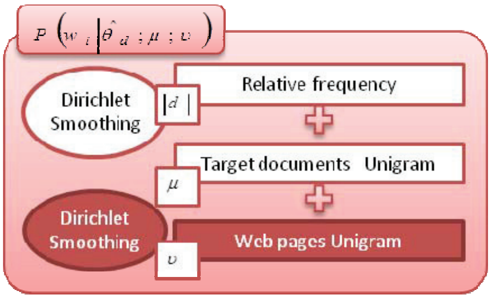

Figure 1. Smoothing using static and dynamic collections.

The web pages are obtained according to given queries. In a simple crawling scheme, the web pages are obtained for each query. In recent web retrieval technologies, on the other hand, it is often handled to utilize past queries in order to improve the results. Consequently, another crawling method is considered that uses a query history, or in the case of NTCIR-10 SpokenDoc-2, all the queries are utilized.

3.2 Weighting web pages

In order to employ the dynamic document collection, it might be necessary to evaluate each web page; some pages contain important information, while some pages have less information. Hence, in this paper, we try to weight web pages using a topic model.

3.2.1 LDA : Latent Dirichlet Allocation

LDA is a probabilistic model where a probability to generate a particular term can be calculated assuming a topic distribution [2]. In general, LDA achieves better performance than pLSI (probabilistic Latent Semantic Indexing) [3] in terms of representing the topic distribution and the relationship between topics. In LDA, each sample is assumed to be observed from a generative probabilistic process that includes hidden variables. LDA is robust against the over-adaptation problem since LDA is based on Bayesian estimation.

Let us denote a latent topic by $z_k \in Z = \{z_1, z_2, \cdots, z_{|Z|}\}$ and a probability of a topic z_k by ϑ_k. In LDA, it is assumed that a set of topic probabilities $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_{|Z|})$ is given by the Dirichlet distribution $\mathrm{Dir}(\boldsymbol{\vartheta}|\boldsymbol{\alpha})$ for each document. A probability of a document $d = (w_1, w_2, \ldots, w_N)$ is then expressed by:

$$P(d|\boldsymbol{\alpha},\boldsymbol{\beta}) = \int \mathrm{Dir}(\boldsymbol{\vartheta}|\boldsymbol{\alpha}) \left\{ \prod_{i=1}^{N} \sum_{Z} P(w_i|z_k,\boldsymbol{\beta}) P(z_k|\boldsymbol{\vartheta}) \right\} d\boldsymbol{\vartheta} \qquad (9)$$

where

$$\sum_{Z} = \sum_{z_1=1}^{|Z|} \sum_{z_2=1}^{|Z|} \cdots \sum_{z_N=1}^{|Z|} \qquad (10)$$

In Equation (9), $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are LDA model parameters, where $\beta_{k,i}$ denotes $P(w_i|z_k)$: a unigram probability of a term w_i in a topic z_k. These parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ can be trained using the variational Bayesian method.

3.2.2 Weighting method

Let us consider to weight a web page $p \in W = \{p_1, p_2, \cdots, p_{|W|}\}$ based on the hypothesis that web pages which are close to target documents are important. At first, a topic mixture ratio vector $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \cdots, \gamma_{|Z|})$ for each web page as well as target document is respectively computed. Secondly, the distance $\delta(p, d)$, between a web page p and a target document d, is estimated. Figure 2 illustrates a concept of the distance. In this work, we estimate the distance using the cosine distance. Thirdly, the distance of the web page and the document collection is computed as:

$$\delta(p, C) = \frac{1}{|C|} \sum_{m=1}^{|C|} \delta(p, d_m) \qquad (11)$$

where $C = \{d_1, d_2, \cdots, d_{|C|}\}$. The value $\delta(p, C)$ in Equation (11) is then used to weight the corresponding web page.

Finally, a probability to observe a term w_i in the dynamic document collection W is formulated as Equation (12), where N_j is the total number of terms appeared in a web page p_j:

$$P(w_i|\theta_W) = \frac{\sum_{j=1}^{|W|} \delta(p_j, C) \cdot C(w_i, p_j)}{\sum_{j=1}^{|W|} \sum_{l=1}^{N_j} \delta(p_j, C) \cdot C(w_l, p_j)} \qquad (12)$$

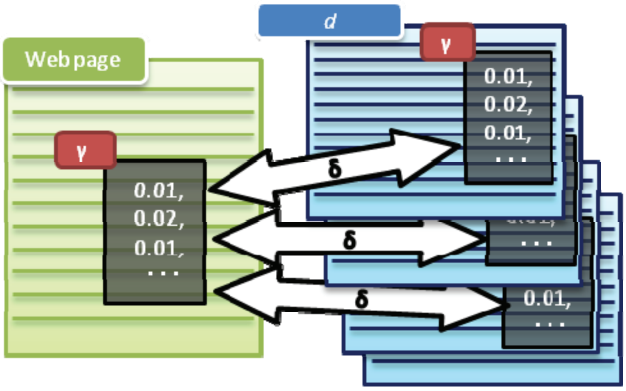

Figure 2. Distance between a web page and each document.

4. EXPERIMENTS

4.1 Experimental setup

We conducted experiments to evaluate our proposed method for SpokenDoc-2 SCR subtask in NTCIR-10. Table 1 shows experimental conditions. For the target documents and the static document collection, we used the provided recognition results (we did not improve speech recognition results, and simply employed the provided results as they were). For the web page collection, we obtained 30 web pages per query. These pages were collected using Yahoo!API [4]. The other newspaper corpus was prepared to build the LDA model for weighting. Retrieved results were evaluated by the Mean Average Precision (MAP) score. Note that smoothing parameters are manually optimized and fixed through the experiments.

Table 1. Experimental conditions.

Sub-subtask	Lecture retrieval
Spoken document	Ref-Word-Matched
LDA training data	Mainichi newspaper corpus 2007 – 2008
Web search engine	Yahoo! API [4]
Smoothing parameters	$\mu = 4000$, $\nu = 50$

4.2 NTCIR-9 Dry-run results

As a preliminary experiment, we tested our method using NTCIR-9 SpokenDoc Dry-run data [5] including 39 queries. In this experiment, four schemes were evaluated: "Original" for the conventional Dirichlet smoothing without using web pages (Equation (7)), "Web" for our proposed method except weighting (Equation (8)), "LDA" for our method with LDA-based weighting (Equation (12)), and "TF-IDF" for the method using TF-IDF-based distance instead of Equation (11); the distance $\delta(p, C)$ is computed by TF-IDF.

Figure 3 shows experimental results for NTCIR-9 SpokenDoc Dry-run data. It is obvious that the performance is improved by using web data. And our proposed method "LDA" achieved better performance than the other schemes "Web" and "TF-IDF". As a result, it is found that our proposed method using static and dynamic collections with LDA-based weighting can successfully improve retrieved results.

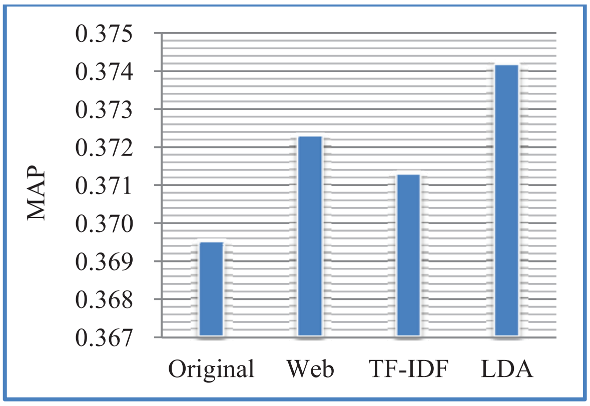

Figure 3. Results for NTCIR-9 SpokenDoc Dry-run.

4.3 NTCIR-10 Formal-run results

According to the preliminary experiments, we evaluated our methods "LDA" and "Web" using NTCIR-10 SpokenDoc-2 Formal-run data. For comparison, the query expansion method in the previous study [6] was also tested. Table 2 shows the result given by the task organizer.

Table 2. Results for NTCIR-10 SpokenDoc-2 Formal-run.

LDA (RunID L36)	**Web** (RunID L37)	**Query Expansion** (RunID L38)
0.408	0.399	0.372

This result shows that our proposed method "LDA" achieved better performance not only than "Web" and the query expansion but also than the preliminary experiments shown in section 4.2. The former indicates the effectiveness of our proposed approach. And the latter might be due to the characteristics of the queries in SpokenDoc-2. Since the queries in NTCIR-10 Formal-run were longer than those in NTCIR-9 Dry-run, more relative and informative web pages could be obtained, causing the better dynamic document collection.

5. CONCLUSION

In this paper, we described our spoken document retrieval method with employing the language model approach and without using the vector space model; we proposed to apply the query model to spoken document retrieval, employ static and dynamic document collections, and weight web pages in the dynamic collection. In general, the query model must be smoothed in order to use information retrieval. We expanded the Dirichlet smoothing to adapt web pages. Furthermore, we suggested to weight each web page using the topic model - LDA. The preliminary experiment (NTCIR-9 SpokenDoc Dry-run) shows that our proposed method can improve the MAP score. In addition, our method achieved higher performance in the formal experiment (NTCIR-10 SpokenDoc-2 Formal-run).

As our future work, we have to develop an automatic estimation method of the smoothing parameters. We also try to deal with recognition errors for further improvement.

6. REFERENCES

[1] T. Akiba et al. "Overview of the NTCIR-10 SpokenDoc-2 Task," Proc. 10th NTCIR Workshop Meeting, 2013.

[2] D. M. Blei et al., "Latent Dirichlet Allocation," Journal of Machine Learning Research, vol.3, pp.993-1022, 2003.

[3] T. Hofmann, "Probabilistic Latent Semantic Indexing," Proc. SIGIR'99, pp.50-57, 1999..

[4] "Yahoo! developer network," [Online]. Available: http://developer.yahoo.co.jp.

[5] T. Akiba et al., "Overview of the IR for Spoken Documents Task in NTCIR-9 Workshop," Proc. 9th NTCIR Workshop Meeting, 2011.

[6] S. Tsuge et al., "Spoken Document Retrieval Method Combining Query Expansion with Continuous Syllable Recognition for NTCIR-SpokenDoc.," Proc. 9th NTCIR Workshop Meeting, 2011.