Toward improvement of SDR accuracy using LDA and query expansion for SpokenDoc

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

Hideki Sekiya
Gifu University
1-1 Yanagito Gifu
Gifu 501-1193 Japan
sekiya@asr.info.gifu-u.ac.jp

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

Taro Niinomi
Gifu University
1-1 Yanagito Gifu
Gifu 501-1193 Japan
niinomi@asr.info.gifu-u.ac.jp

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

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

ABSTRACT

This paper investigates several techniques for spoken document retrieval, toward improvement of retrieval performance based on the conventional method i.e. TF-IDF. The first approach employs rescaled unigrams of LDA to compute a similarity score. The second technique employs query expansion by web retrieval using Yahoo!API. And the third technique is Prioritized And-operator Retrieval based on TF-IDF techniques. We tested these methods using a dry-run data, then it turned out that the third technique is most promising.

Categories and Subject Descriptors

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

General Terms

Theory, Experimentation.

Keywords

Information Retrieval, Latent Dirichlet Allocation(LDA), Query Expansion, TFIDF value,

Team Name

ASR

Subtasks

Spoken Document Retrieval

Eternal Resource Used

Yahoo!API, Mainichi Newspaper Account.

1. INTRODUCTION

This paper shows three approaches for NTCIR-9 SpokenDoc [1].

As developing information technology, we can treat and deal with enormous data including video contents; online video sharing and broadcasting websites such as YouTube [2], become so popular, and video retrieval systems have been investigated. In order to use such the state-of-the-art retrieval system efficiently, it is usual to utilize tag information consisting of keywords and related terms. Most conventional retrieval systems use the tag information taken by a presenter or viewers, however, providing the tag information automatically is nowadays essential as the number of video contents increases explosively. Therefore, video mining that generates index terms or keywords as well as other information, e.g. about speakers in a movie, is paid more attention. We have developed a video content viewer and an automatic captioning method for speech recognition by using Latent Dirichlet Allocation (LDA) [3]. In this study, we attempt to automatically conduct video-content mining applying our previous techniques.

If a query topic is given like SpokeDoc, it is common to compute a similarity score between spoken documents and the query, so as to obtain the best retrieval result. The basic feature of similarity is TF-IDF, however, TF-IDF can estimate the similarity only if the words in the query appear in the document. In most cases, other terms that have the same meaning as the words in the query appear in retrieving documents, or inappropriate terms are sometimes contained due to speech recognition errors. Furthermore, TF-IDF is not appropriate if the number of words in the query is too small.

In this study, we attempt to enhance the precision of spoken document retrieval by three ways: Section 2 describes LDA [4] and a retrieval scheme using LDA. Section 3 describes a query expansion technique by web retrieval. Section 4 introduces our proposed method based on the TF-IDF technique. And we conclude this paper in section 5.

2. DOCUMENT RETRIEVAL USING LDA

2.1 LDA : Latent Dirichlet Allocation

LDA is a probabilistic model where a probability to generate a particular word can be calculated assuming a topic distribution. In general, LDA achieves better performance than pLSI (probabilistic Latent Semantic Indexing) [5] in terms of representing the topic distribution and the relationship between topics, because LDA employs Dirichlet distribution for estimate its prior distribution. And LDA data are assumed to be observed from a generative probabilistic process that includes hidden variables. Furthermore,

it is an advantage that LDA is robust against the over-adaptation problem since LDA is based on Bayesian estimation.

Let us denote C latent topics by $\mathbf{Z} = (z_1, z_2, \ldots, z_C)$, and a probability of a k-th topic z_k by θ_k. In LDA, it is assumed that a set of topic probabilities $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_C)$ is given by the Dirichlet distribution $\mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha})$ for each document. A probability of a document $d = (w_1, w_2, \ldots, w_N)$ is then expressed by:

$$P(d|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \int \mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha}) \left\{ \prod_{j=1}^{N} \sum_{\mathbf{Z}} P(w_j|z_k, \boldsymbol{\beta}) P(z_k|\boldsymbol{\theta}) \right\} d\boldsymbol{\theta} \quad (1)$$

where

$$\sum_{\mathbf{Z}} = \sum_{z_1=1}^{C} \sum_{z_2=1}^{C} \cdots \sum_{z_n=1}^{C} \quad (2)$$

In the equation (2), $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the LDA model parameters, where $\beta_{k,j}$ denotes $P(w_j|z_k)$: a unigram probability of a word w_j in a topic z_k ($1 \le j \le J$; J: *vocaburary size*). These parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ can be trained using the variational Beysian method. Considering a topic mixture ratio vector $\boldsymbol{\gamma} = (\gamma_k)$ that is computed in each document f, a probability of a word w_j in a document f is expressed by:

$$P(w_j|f) = \frac{\sum_{k=1}^{C} \gamma_k \beta_{k,j}}{\sum_{k=1}^{C} \gamma_k} \quad (3)$$

2.2 LDA-based Document Similarity

In this paper, we estimate a similarity score $Q(f)$ for spoken document retrieval using LDA. For each document, $Q(f)$ can be measured using a probability $P(v_j|f)$ as:

$$Q(f) = \sum_{j=1}^{V} \omega_j P(v_j|f) \quad (4)$$

where ω_j is the weight of a query's word v_j, and V indicates the number of words in a query. In this paper, an IDF value in a training corpus is used as ω_j. If $Q(f)$ has a large value, it means that the document f is much similar to the given query.

2.3 Preliminary Experiment

A preliminary experiment was conducted to estimate the usefulness of the proposed scheme explained above. Table 1 shows experimental conditions. We tested our method using a dry-run data: 39 queries and their manual retrieved results as well as 2,702 spoken lectures in an academic meeting. An LDA model is built using a two-year newspaper corpus. Retrieved results were evaluated by the Mean Average Precision (MAP) score **[1]**, and each retrieval method was also evaluated by an Average Precision (AP) score.

In this experiment, we compared our scheme with a TF-IDF-based method. Table 2 shows the experimental results. From the result, it is obvious that the LDA-based scheme did not work well. This degradation might be due to the mismatch between the task and the corpus for the LDA model. The newspaper corpus covers wide topics such as society, international, entertainment, etc. On the other hand, the experimental task includes only domain-specific data. Using a domain-specific corpus, e.g. Corpus of Spontaneous Japanese (CSJ), may solve this problem. However in this case, an LDA model cannot be built from CSJ due to the lack of database size. If we prepare a proper corpus to build an LDA model, we then believe that its performance must be improved.

Table 1: Experimental conditions.

Task	Lecture retrieval
Spoken Document	Ref-Word
Query	39 queries (dry-run)
LDA Training Data	Mainichi newspaper corpus 2007 - 2008

Table 2: Experimental results for an LDA-based method.

TF-IDF	LDA
0.252	0.024

3. QUERY EXPANTION TECHNIQUE

3.1 Query Expansion

As described, sometimes TF-IDF is not suitable for document retrieval if a given query is so small. We try to propose a query expansion approach to overcome this disadvantage of TF-IDF. In this section, we describe a retrieval scheme expanding a query topic by web retrieval.

In our query-expansion scheme, Yahoo!API [6] is used as a web retrieval. Yahoo!API can find web pages by setting several web-search parameters, as if we generally do in web retrieval. Figure 1 shows a flowchart of query expansion. At first, TF-IDF values of all words appeared in each spoken document are calculated. Secondly, web search is conducted using keywords which correspond to the five-best TF-IDF words. By the web search, 20 to 30 web pages per document are obtained and these pages are sorted by the number of the 30-best TF-IDF keywords chosen from the original spoken document. Then the five-best web pages are extracted. Thirdly, a query is also expanded in the same way; using keywords extracted from the query, web search is conducted to obtain the 30-best web pages. At last, similarity scores are computed by comparing chosen five web texts from spoken documents with 30 web texts from the query. The cosine-based similarity score is employed in our scheme. Finally five-best texts are extracted as a expanded query.

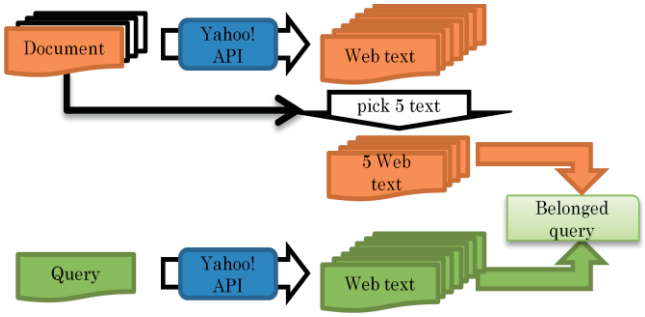

Figure 1. A query expansion scheme.

3.2 Preliminary Experiment

We conducted a preliminary experiment in the dry-run data explained in the previous section. Table 3 shows experimental results comparing TF-IDF and the proposed query expansion in

the MAP criterion. It is observed in the result that, the proposed scheme achieved better performance in some cases, on the other hand, the scheme was not superior to TF-IDF otherwise. This means our proposed query expansion has effectiveness and utility to some extent.

Non-text web pages, e.g. PDF data, may be essential for query expansion; such the pages often contain important keywords and useful information. Therefore, picking up non-text web pages and extracting meaningful information from these pages should be discussed as a coming issue. And as a future work, combination of TF-IDF and the query expansion should be investigated in order to improve the accuracy.

Table 3 : Experimental results for query expansion.

	TF-IDF	Query Expansion
MAP	0.252	0.152
AP : Dryrun-15	0.162	0.003
AP : Dryrun-16	0.015	0.610
AP : Dryrun-17	0.147	0.012
AP : Dryrun-18	0.421	0.400
AP : Dryrun-19	0.086	0.009

4. INPROVED TF-IDF-BASED RETRIEVAL

4.1 Prioritized And-operator Retrieval

This section explains our proposed method “AND-operator search first retrieval” for the SDR task.

When we compare a TF-IDF value for a query with TF-IDF values of spoken documents, we found two problems. One of them is due to speech recognition errors. Analyzing the spoken documents, some terms with high TF value may be misrecognized. If a spoken document includes a lot of speech recognition errors, query terms do not exist in the spoken document, and such the document cannot be chosen. Similarly, if a query has out-of-vocabulary terms for spoken documents, document retrieval might fail. In another problem, there are some query terms involved in most documents. These terms make topic estimation much difficult, and make similarity scores to non-related documents incorrectly small.

In our proposed method, at first we calculate a cosine distance between a TF-IDF weight of the query and each spoken document. Second, we classify spoken documents into two document sets; one has all terms on the query, and another does not have all. Third, we sort spoken documents in the ascending order in each document set according to cosine distance. Finally, the second set is lined up under the first set and ranked like Figure 2.

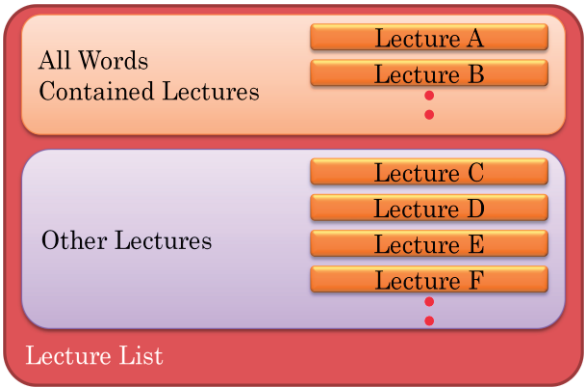

Figure 2. Document ranking in the proposed method.

4.2 Experiments

To show the effectiveness of the proposed method, evaluation experiments were conducted. Table 4 shows the MAP of the proposed method and the method with only TF-IDF technique. We retrieved the target documents from both of transcribed texts and speech recognition results. We used newspaper articles for two years to estimate IDF weights as Section 2. From the results shown in Table 4, our proposed method achieved better performance than TF-IDF, and its effectiveness of the proposed method is verified.

Table 4 : Experimental results for our proposed method.

Proposed method : MANUAL	0.330
TF-IDF : MANUAL	0.303
Proposed method : REF-WORD	0.299
TF-IDF : REF-WORD	0.252

5. CONCLUSION

We have proposed three methods to improve accuracy of spoken document retrieval. First, we described a retrieval approach using LDA. Rescaled unigrams are used to compute a similarity score. Next, we described query expansion using web search. In the scheme, the query is expanded using web texts retrieved by Yahoo!API. According to preliminary experiments, effectively of query expansion was shown in some query topics. Finally we proposed the method using a TF-IDF technique, where documents are classified according to query terms, and evaluated by TF-IDF values. Experimental results show the effectiveness of the proposed method compared to the conventional TF-IDF method.

As a future work, we have to see the method using LDA from a new view, using both TF-IDF value and query expansion. A method toward recognition errors and out-of-vocabulary words should be investigated. We would like to utilize non-text web data in the query expansion, and integration of our three methods.

6. REFERENCE

[1] Tomoyoshi Akiba et al., "*Overview of the IR for Spoken Documents Task in NTCIR-9 Workshop.*" 2011.

[2] YouTube. http://www.youtube.com/.

[3] Masanao Okamoto et al., "*Topic Based Generation of Caption and Keywords for Video Content.*" Proc. APSIPA ASC 2010, Biopolis, Singapore, 2010.

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

[5] Hofmann Thomas, "*Probabilistic Latent Semantic Indexing.*", Proc. SIGIR'99, ACM Press, pp.50-57, 1999.

[6] Yahoo! Developper Network. http://developer.yahoo.co.jp.