PAPER *Special Section on Data Engineering and Information Management*

Topic Representation of Researchers' Interests in a Large-Scale Academic Database and Its Application to Author Disambiguation

Marie KATSURAI[†a)], ***Member***, **Ikki OHMUKAI**[††b)], ***Nonmember***, ***and*** **Hideaki TAKEDA**[††c)], ***Member***

SUMMARY It is crucial to promote interdisciplinary research and recommend collaborators from different research fields via academic database analysis. This paper addresses a problem to characterize researchers' interests with a set of diverse research topics found in a large-scale academic database. Specifically, we first use latent Dirichlet allocation to extract topics as distributions over words from a training dataset. Then, we convert the textual features of a researcher's publications to topic vectors, and calculate the centroid of these vectors to summarize the researcher's interest as a single vector. In experiments conducted on CiNii Articles, which is the largest academic database in Japan, we show that the extracted topics reflect the diversity of the research fields in the database. The experiment results also indicate the applicability of the proposed topic representation to the author disambiguation problem.

key words: *researcher analysis, academic database, topic model, author disambiguation*

1. Introduction

Knowledge acquisition from academic papers has received much attention since huge digital libraries have become available as a result of database systems and Internet development. Various approaches have been presented to map and visualize the structures of scientific research activities at different levels of granularity including academic papers [1], [2], conferences [3], and researchers [4]–[7]. In particular, analyzing researcher relationships is useful for measuring the influence of authors [4], discovering communities [6], and recommending research collaborators [7].

Conventional methods of academic database analysis can be roughly divided into two approaches on the basis of the features they exploit: the bibliographic metadata and textual content of papers. The bibliographic metadata approach uses author lists or citations to derive co-authorship networks [8] or co-citation networks [8]–[10], to which network analysis methods are applied in order to capture the scientific structure of a target database. In contrast, the recently-proposed textual content approach, uses the words of academic papers to characterize conferences or researchers [3], [5]. One advantage of such a content-based analysis is that it can determine the potential relevance of papers, which cannot be identified by the metadata-based approach. However, these conventional methods have been applied to a limited research domain (e.g., information retrieval and natural language processing), and their usefulness in a large-scale academic database has not been adequately demonstrated. Characterizing researchers' interests with a set of diverse topics is necessary to develop applications that are useful in an academic environment, for example, promoting interdisciplinary research and recommending collaborators from different research fields.

This paper presents a topic representation of researchers' interests in a large-scale academic database covering all research fields. The proposed method focuses on the textual content of papers to characterize research fields. Specifically, we use the abstract of each paper, which is generally the longest piece of text that is available from a single paper in large-scale academic databases. In the proposed method, latent Dirichlet allocation (LDA) [11] is applied to a collection of abstracts, and this process results in word distributions for each research topic. With the use of the learned LDA, each abstract in a target researcher's publications is represented by a topic distribution. All topic distributions are finally summarized as a single vector to represent the researcher's interests. We conduct experiments using a dataset from CiNii Articles[*], the largest academic database in Japan, in order to validate the effectiveness of our approach.

The derived topic representations of researchers are potentially effective for several applications. For example, we use the topic vectors of the researchers for the *author disambiguation* task, i.e., choosing a correct author for a given paper from researchers who share the same full name [12]–[15]. An added merit of this particular application is that it can be evaluated quantitatively. The experiment results show that the presented topic representation of researchers' interests achieves better author disambiguation performance than comparative methods that directly use the textual features from abstracts or that are based on metadata.

In summary, the main contributions of this study are twofold. First, to the best of our knowledge, this work is the first to represent researchers' interests in a large-scale academic database that includes not a specific, but a diverse range of research fields in Japan. Second, we show the applicability of the topic representation of researchers

Manuscript received July 2, 2015.
Manuscript revised November 10, 2015.
Manuscript publicized January 14, 2016.

[†]The author is with Doshisha University, Kyotanabe-shi, 610–0394 Japan.
[††]The authors are with the National Institute of Informatics, Tokyo, 101–8430 Japan.
a) E-mail: katsurai@mm.doshisha.ac.jp
b) E-mail: i2k@nii.ac.jp
c) E-mail: takeda@nii.ac.jp

DOI: 10.1587/transinf.2015DAP0031

[*]http://ci.nii.ac.jp/en

to author disambiguation. The remainder of this paper is organized as follows. Section 2 provides a brief review of related studies. Section 3 presents a topic model-based approach that represents researchers' interests with the use of the textual features of academic papers. Section 4 describes an application of the proposed method to the author disambiguation problem. Section 5 presents the results of the experiments to verify the effectiveness of the proposed method. Finally, Sect. 6 gives a summary and some possible directions for future work.

2. Related Work

2.1 Knowledge Discovery from Academic Databases

Knowledge discovery from academic databases has a long history in bibliometrics research, with most conventional methods using the bibliographic metadata of papers. Small [1] assumed that more frequently co-cited papers are more related and analyzed paper co-citation relationships. This idea was extended to author analysis by White and Griffith [9]; they counted the frequency with which any paper of an author is co-cited with another author in the references of citing papers. Börner et al. [8] constructed a co-authorship network, in which each node corresponds to a researcher, and each edge represents the number of co-authored papers between the researchers. Ding et al. [4] constructed an author co-citation network among the 108 most highly cited authors in the information retrieval field, and calculated the ranks of the authors on the basis of PageRank measures. Radev et al. [10] defined network analysis-based measures to rank authors or papers, and they conducted experiments using a collection of papers published in conferences of natural language processing (NLP).

On the other hand, recently presented methods do not rely on bibliographic metadata but use the textual content of papers. In these methods, topic models have played an important role to describe the content of papers in a low-dimensional feature space. Hall et al. [3] used LDA to calculate the topic distributions of three international conferences related to NLP, and they calculated the similarity between the topic distributions of the conferences. In addition, by calculating the entropy of the topic distribution, the authors quantified the range of research themes that each conference covers. The Hall's method has been effectively applied to analyze other research fields such as software systems [16]. Lu and Wolfram [5] presented a topic model-based approach to measure the relatedness between researchers. Specifically, they applied the author topic model (ATM) [17], an extension of LDA with author information, to represent each author by a multinomial distribution over topics. However, the analysis in [5] is limited to only 50 researchers, and their research field involved is information science only. Compared with the work in [5], the present study aims to characterize researchers' interests in a diverse range of research topics, and it shows the usefulness of such an approach in facilitating the management of large-scale academic databases.

2.2 Author Disambiguation in Academic Databases

Author name ambiguity is a well-known and important issue to overcome in an academic database that offers a paper search service [18]. Clustering [12], [13] and identifier assignment [14], [15] are two approaches for author disambiguation. Some digital libraries have introduced the clustering approach in practice to improve the efficiency of author search [19]–[21]. However, the authors in [22] reported that the existing systems have yet to achieve a highly accurate identification.

Author identifier assignment is another name disambiguation approach, which provides direct links between academic papers and a list of researchers. Thomson Reuters provides this type of systems, which is called ResearcherID[†], and the Open Researcher and Contributor ID (ORCID) initiative [23] attempts to assign global identifiers to authors with the aim of linking several publishers. CiNii Articles, whose overview is described in the Appendix, is also equipped with an author identifier assignment system. These identifiers can offer precise information about an arbitrary author's research activity in a database, so the evaluation of a researcher's performance is possible. However, current systems require manual operation by the users themselves or by the database managers. The systems should present to users the results of classifying an author of a given paper into one of the candidate researchers who share the same full name to reduce the time for manual operation. Thus, in this paper, we quantitatively show that our topic representation of researchers' interests is applicable to the distinguishment of a correct author.

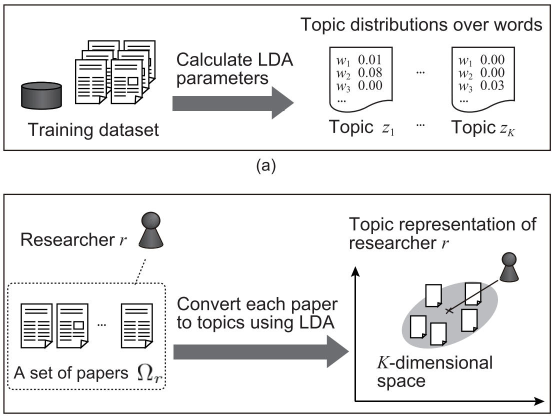
3. Topic Representation of Researchers' Interests in a Large-Scale Academic Database

This section presents a topic model-based approach for representing researcher's interests in a large-scale academic database. An overview of the proposed method is shown in Fig. 1. Given a training dataset, we first extract the textual features of each paper. Then, we extract the topics by calculating the distributions over words with the use of LDA (see Sect. 3.1). This step enables the characterization of each paper with a set of topics. All publications of the researcher are converted to topic vectors, and the centroid of these vectors is calculated to model a target researchers' interests (see Sect. 3.2).

3.1 Topic Extraction from an Academic Database

This subsection describes how to extract topics from a target academic database. Suppose that we are given a training dataset consisting of N abstracts from the database, which is denoted by D. For each abstract $d \in D$, we extract a vector

[†] http://www.researcherid.com/



(b)

Fig. 1 The proposed steps to represent researchers' interest: (a) topic extraction through LDA training and (b) topic summarization from the researchers' publications.

of words, $\boldsymbol{x}_d = [x_{d1}, x_{d2}, \ldots, x_{dW_d}]$, where W_d is the number of words in abstract d. Let W be the number of unique words in the training dataset, and $V = [w_1, w_2, \ldots, w_W]$ represents a vocabulary of words.

The words that appear in an academic paper usually reflect a particular set of topics the paper deals with. We use LDA to determine the set of topics that are used in the training dataset. With the assumption that K topics are found in the database, LDA regards an abstract as a probabilistic mixture of these topics. The generative process of LDA can be formalized as follows [11]:

- For each topic $z \in \{1, 2, \ldots, K\}$, draw a W-dimensional multinomial distribution $\boldsymbol{\phi}_z$ from a Dirichlet distribution with prior β;
- For each document $d \in D$, draw a K-dimensional multinomial distribution $\boldsymbol{\theta}_d$ from a Dirichlet distribution with prior α;
- For each word x_{di} in document d, draw a topic $z_{di} \in \{1, 2, \cdots, K\}$ from the multinomial distribution $\boldsymbol{\theta}_d$;
- Draw a word x_{di} from the multinomial distribution $\boldsymbol{\phi}_{z_{di}}$.

After the model training, the resulting distribution $\boldsymbol{\phi}_z$ shows which words are important in topic z, whereas the distribution $\boldsymbol{\theta}_d$ represents which topics are included in paper d. To calculate these distributions, we exploit the collapsed Gibbs sampling [24]. Given a new abstract, we can estimate its topic distribution by using the fixed distribution $\{\boldsymbol{\phi}_z\}_{z=1}^{K}$ in the sampling framework. The calculation of the topic distribution corresponds to reducing the dimensionality of the textual feature space from W to K, which can deal with the variability in word choice across research fields in a large-scale academic database.

Unlike ATM used in [5], LDA itself does not provide information about the interests of authors during training. However, we solve that using a two-stage approach that first estimates a topic proportion of a given paper; and then uses it for calculating the researchers' interests. The detail of this approach is described in the following subsection.

3.2 Characterization of Researchers with the Use of Topics

This section presents an approach to characterize researchers' interests in a topic space calculated by LDA. Let Ω_r be a set of abstracts authored by a target researcher r. From each abstract $d \in \Omega_r$, its word vector $\boldsymbol{x}_d$ can be converted to a K-dimensional topic distribution $\boldsymbol{\theta}_d$ by an inference in LDA. Our aim is to discover the set of topics expressed by the papers in Ω_r and use these to characterize the researcher r. Several possible ways can be used to summarize a set of vectors $\{\boldsymbol{\theta}_d | d \in \Omega_r\}$. In this paper, we calculate a centroid of the vectors by using a simple average approach described as follows:

$$\boldsymbol{m}_r = \frac{1}{|\Omega_r|} \sum_{d \in \Omega_r} \boldsymbol{\theta}_d, \tag{1}$$

where $|\Omega_r|$ denotes the number of abstracts in Ω_r. The resulting vector encodes the information about research topics that the papers focus on. Such a vector representation of researchers' interest has several potential applications. For example, to calculate the relatedness between researchers r_1 and r_2, we can exploit the similarity between their topic vectors $\boldsymbol{m}_{r_1}$ and $\boldsymbol{m}_{r_2}$.

The proposed method can be applied to an arbitrary academic database if a set of papers is accurately linked to researchers in the database. As mentioned in Sect. 2.2, current academic databases usually require the manual author identifier assignment because some researchers who share the same full name: this requirement entails time-consuming work. Thus, we try to use the proposed method to solve an author disambiguation problem in the following section.

4. Author Disambiguation Using Topic Representation of Researchers' Interests

This section describes how to use the proposed topic representation of researchers' interests as one of its possible applications. We assume that a target database has researcher identifiers. We consider the following situation: when a new abstract d_{new} is given to the database, more than one researcher has the same name as that of an author of d_{new}. The task is to assign the correct researcher identifier to the author. Let us denote a set of researchers whose full names are a by R_a. To disambiguate the author name a in d_{new}, we use a framework illustrated in Fig. 2. First, words from the paper d_{new} are converted to a topic distribution $\boldsymbol{\theta}_{d_{new}}$. Then, using the topic vectors of the candidate researchers $\{\boldsymbol{m}_r | r \in R_a\}$, we select the correct author as follows:

$$r^* = \arg\min_{r \in R_a} D(\boldsymbol{m}_r, \boldsymbol{\theta}_{d_{new}}), \tag{2}$$

where $D(\boldsymbol{m}_r, \boldsymbol{\theta}_d)$ is a distance measure between two vectors $\boldsymbol{m}_r$ and $\boldsymbol{\theta}_{d_{new}}$. By comparing the topic vector of the paper with the topic vector of each researcher, we can disambiguate its author name.

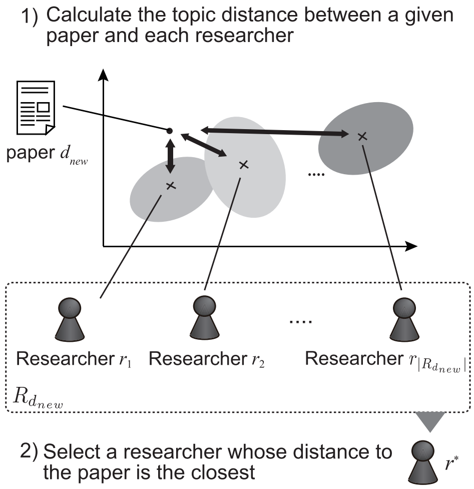

1) Calculate the topic distance between a given paper and each researcher
paper d_{new}
....
Researcher r_1
Researcher r_2
....
Researcher $r_{|R_{d_{new}}|}$
$R_{d_{new}}$
2) Select a researcher whose distance to the paper is the closest
r^*

Fig. 2 An illustration of the application of the researcher's topic representation to author disambiguation.

Presenting the result of author disambiguation to users can reduce the manual effort in maintaining the database. After the identifier of paper d_{new} is fixed, we update the topic representation of the researcher r^* as follows:

$$\boldsymbol{m}_{r^*} \leftarrow \frac{1}{1+|\Omega_{r^*}|}\left(\boldsymbol{\theta}_{d_{new}} + |\Omega_{r^*}|\boldsymbol{m}_{r^*}\right), \tag{3}$$

$$\Omega_{r^*} \leftarrow \{d_{new} \cup \Omega_{r^*}\}. \tag{4}$$

With the use of these equations, the topics of the given abstract can be easily reflected in the latest topic vector of the researcher. This update method is also a contribution of our method to manage the large-scale academic database compared with the most related work [5]: the conventional method [5] assumes that the target academic database is static over time, and it does not show any scheme to update the researcher's information based on his/her new publications. Thus, if a new paper is added to the database, the method in [5] requires to be trained again using a whole dataset, which is very time-consuming and intractable in a large-scale academic database. On the other hand, our method assumes that the target database is populated with new papers and can efficiently update a researcher's topic vector using Eqs. (3) and (4) when given his/her new publications.

5. Experiments

In this section, we present experimental results that verify the effectiveness of the proposed method. The details of the datasets used in our experiments are described in Sect. 5.1. Then, the results of topic extraction from the dataset, which represent some researchers' interests, are presented in Sect. 5.2. Finally, an application of the topic representation of researchers' interests to author disambiguation is evaluated in Sect. 5.3.

Table 1 Details of the training dataset used in the experiments.

Number of papers	104,705
Number of researchers	31,399
Number of unique words	85,589
Average number of words per a paper	49

5.1 Dataset

This subsection explains how we constructed a dataset based on CiNii Articles. First, to cover active researchers in all research fields in Japan, we extracted a set of researcher identifiers from the research grant awards database KAKEN[†]. The relationship between CiNii Articles and KAKEN is described in the Appendix. Then, all abstracts written in Japanese by each researcher were retrieved from CiNii Articles: a total of more than 220K papers were identified. By randomly sampling the papers so that the total number of each researcher's abstracts is less than 15, we constructed a training set of 104,705 papers. To extract a set of words from each abstract in the training set, we performed morphological analysis using MeCab[††]. Specifically, by introducing a list of page names in Wikipedia[†††] to MeCab, we regarded named entities as nouns (e.g., "support vector machine" was used as a proper noun). Finally, unigrams and bigrams of consecutive nouns were used as the textual features of the abstract. To reduce the effects of noisy words for training LDA, we discarded words that appeared in more than 5% of the total number of papers or words that appeared in fewer than five papers: a vocabulary that consists of 85,589 words was derived. The details of the training dataset constructed are shown in Table 1.

5.2 Results of Topic Extraction and Representing of Researchers' Interests

In this subsection, we first present the results of topic extraction, using the training dataset from CiNii Articles. Based on the training dataset, we calculated the topic distributions over words using LDA, as described in Sect. 3.1, in which the number of topics was empirically set as $K = 500$. The LDA hyperparameters α and β were set to $50/K$ and 0.01, respectively. Note that these model parameters can be automatically selected based on the training dataset using sophisticated methods [25], [26], which will be investigated in future work. Table 2 shows a few examples of the extracted topics[††††]. The table shows that the extracted topics span a diverse range of research fields, including *sociology*, *programming*, *architecture*, and *clinical medicine*. We can also see that the extracted topics were at different granularities: for example, Topic 358 corresponds to a specific research theme (*optical communication*), while Topic 432

[†] https://kaken.nii.ac.jp/en/

[††] https://code.google.com/p/mecab/

[†††] https://www.wikipedia.org/

[††††] Originally, all the topics obtained by the method were explained by Japanese words.

Table 2 Examples of topics extracted from the training dataset. Our interpretations of the topics are in **bold** text.

Topic 22		Topic 25		Topic 47		Topic 172	
rice cropping		**sociology**		**programming**		**histochemistry**	
Word/Phrase	Prob.	Word/Phrase	Prob.	Word/Phrase	Prob.	Word/Phrase	Prob.
flooding	0.0377	social	0.1016	program	0.0457	immune tissue	0.0323
weedkiller	0.0274	society	0.0355	object	0.0420	histochemistry	0.0315
plowing	0.0228	cultural	0.0240	description	0.0348	chemical	0.0269
cc	0.0206	sense of values	0.0216	object-oriented	0.0328	squamous epithelium	0.0268
paddy field	0.0201	community	0.0120	processing system	0.0287	cancer cell	0.0225
yield	0.0178	opinion	0.0118	implementation	0.0277	expression	0.0214
Topic 260		Topic 273		Topic 277		Topic 307	
circuits		**cropping**		**botany**		**DNA analysis**	
Word/Phrase	Prob.	Word/Phrase	Prob.	Word/Phrase	Prob.	Word/Phrase	Prob.
power consumption	0.0698	yield	0.0240	root	0.0379	base sequence	0.0433
circuit	0.0332	ripening	0.0234	above ground	0.0354	specific	0.0272
design	0.0170	leaf area	0.0211	growth	0.0311	gene	0.0259
fpga	0.0161	dry weights	0.0168	root system	0.0287	dna	0.0234
logic circuit	0.0156	heading time	0.0156	nitrogen	0.0199	pcr	0.0216
reduction	0.0130	number of ears	0.0133	underground	0.0176	amplification	0.0177
Topic 358		Topic 404		Topic 405		Topic 415	
optical communication		**architecture**		**biomass**		**clinical medicine**	
Word/Phrase	Prob.	Word/Phrase	Prob.	Word/Phrase	Prob.	Word/Phrase	Prob.
waveguide	0.0535	indoor	0.0206	soil	0.0757	tumor	0.0580
refractive index	0.0294	thermal environment	0.0169	biomass	0.0180	diagnosis	0.0334
optical fiber	0.0195	comfortability	0.0166	andosol	0.0128	lump	0.0248
resonator	0.0178	skin temperature	0.0166	pore	0.1230	case	0.0236
light wave	0.0128	heat load	0.0115	mineralization	0.0108	enforcement	0.0187
light	0.0099	indoor environment	0.0114	layer	0.0992	histopathology	0.0142
Topic 432		Topic 458		Topic 463		Topic 479	
image processing		**meteorology**		**fast algorithm**		**histology**	
Word/Phrase	Prob.	Word/Phrase	Prob.	Word/Phrase	Prob.	Word/Phrase	Prob.
image	0.1614	precipitation	0.0422	fast	0.0980	histology	0.0696
original image	0.0225	wind	0.0153	parallelization	0.0329	histopathology	0.0318
input image	0.0201	observation	0.0126	execution time	0.0235	histological	0.0203
contour	0.0183	atmosphere	0.0099	processor	0.0217	macroscopic	0.0153
photography	0.0171	develop	0.0093	execution	0.0216	tissue image	0.0151
pixel	0.0143	cloud	0.0078	performance improvement	0.0205	fibrillation	0.0130

(image processing) appears in several research fields, including *medical image analysis* and *video coding*.

Moreover, we show how researchers were represented with combinations of multiple topics by the proposed method. Based on a list of research fields designed in KAKEN, we randomly picked three research fields: *computer system and networks*, *plant nutrition and soil science*, and *fundamentals of veterinary medicine*, Then, we randomly chose five researchers registered with KAKEN from each research field, respectively, and we calculated their topic vectors using Eq. (1). Figure 3 shows a heat map of the topic representations of the fifteen researchers: the darker cells indicate a higher assignment of the corresponding topic[†]. From this figure, we can see that researchers from the same research field tend to share the same topics. This research field-based analysis shows the effectiveness of the topic vectors in representing researchers' interests.

5.3 Results of an Application to Author Disambiguation

This subsection evaluates the applicability of the proposed method to the author disambiguation problem. For the experimental setup, we first extracted a set of researchers who share the same names from the active researchers in KAKEN. Then, we collected their publications from CiNii Articles to construct a testing dataset for author disambiguation. The details of the testing dataset constructed are shown in Table 3. To conduct the experiment, we consider the practical case, in which new papers are added to the target database day by day. The experiment procedure is as follows: for a researcher name a, all papers of researchers who share the name a, i.e., $\{d \in \Omega_r | r \in R_a\}$, were sorted in the ascending order of the publication dates. Each researcher in R_a was assigned the first paper, and from this, the researchers topic vector was calculated. Then, taking any other paper from the set, we selected the suitable author of the paper from R_a on the basis of the topic vectors calculated by this time point. To calculate Eq. (2), we used

Table 3 Details of the testing dataset used in an author disambiguation experiment.

Number of papers	6,632
Number of researchers	345
Number of unique full names	166

[†] Due to space limitations, only a set of the top three topic indices from each researcher is shown in this figure.

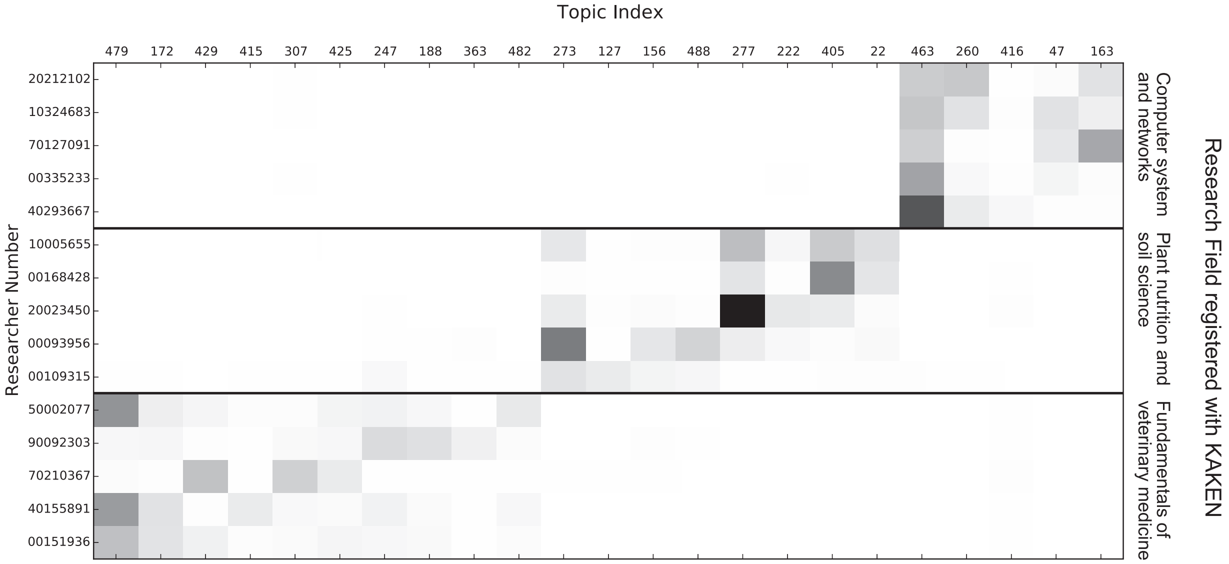


Fig. 3 A Heat map of the topic representation of each researcher, in which only a set of top three topic indexes from each researcher is shown due to the space limitation. Darker cells indicate a higher assignment of the corresponding topic.

the cosine distance between two vectors. After author identification, the topic vector of the correct author was updated according to Eqs. (3) and (4), and this process was iterated.

For performance comparison, we also applied the traditional vector space model (VSM) [27]. In VSM, each researcher is represented as the average vector of word vectors from all publications. Specifically, the following three versions of VSM were applied:

- **VSM_Stop**. VSM_Stop discards highly frequent words and very rare words in the same way as in the proposed method.
- **VSM_All**. VSM_All uses all words of papers without removing stopwords.
- **VSM_TFIDF**. VSM_TFIDF uses all words of papers, which are weighted with TF-IDF [28].

For reference, we compared the performance of our topic-based features with other metadata-based features, which are presented in conventional studies [13], [29]. Specifically, we could investigate the effectiveness of each of the following fields of metadata only which is available in CiNii Articles:

- **Affiliation.** This feature is used in [13]. It is a token-based Jaccard similarity between the name of the affiliation written in a given paper and the name of a researcher's latest affiliation.
- **JCName.** This feature is presented in [29]. It is a token-based Jaccard similarity between the name of the journal/conference that published a given paper and the most similar journal/conference name of past publications of a researcher.
- **Title.** This feature is presented in [29]. It is a token-based Jaccard similarity between the title of a given paper and the most similar title of past publications of a researcher.
- **Keywords.** This feature is presented in [29]. It is also a token-based Jaccard similarity between the keywords of a given paper and the keywords from all publications of a researcher.

Table 4 Mean accuracy of author disambiguation for the proposed method and comparative methods.

Method	Mean accuracy (standard deviation)
Affiliation	84.91% (10.92%)
JCName	91.99% (9.98%)
Title	83.12% (11.95%)
Keyword	59.58% (15.34%)
VSM_Stop	87.16% (13.16%)
VSM_All	89.27% (12.35%)
VSM_TFIDF	89.60% (12.35%)
Proposed method	**92.60% (9.67%)**

These features were used in the same experimental scheme as the VSM-based methods and the proposed method for fair comparison. Note that when each field of metadata was empty or null for a given paper, we randomly selected its author from the researchers, and averaged the accuracy of five runs for the method.

To quantitatively measure the performance of each method, we calculated the mean accuracy of author disambiguation over unique full names. The results are shown in Table 4, from which we find that the proposed method achieved the best mean accuracy among all methods. This result validates the effectiveness of the textual features of abstracts and its dimensionality reduction by LDA. We also find that JCName outperformed other metadata-based features, but its performance is not higher than that of our method. This means that our topic representation can characterize each researcher's domain better than the journal/conference names. The collaborative use of the pro-

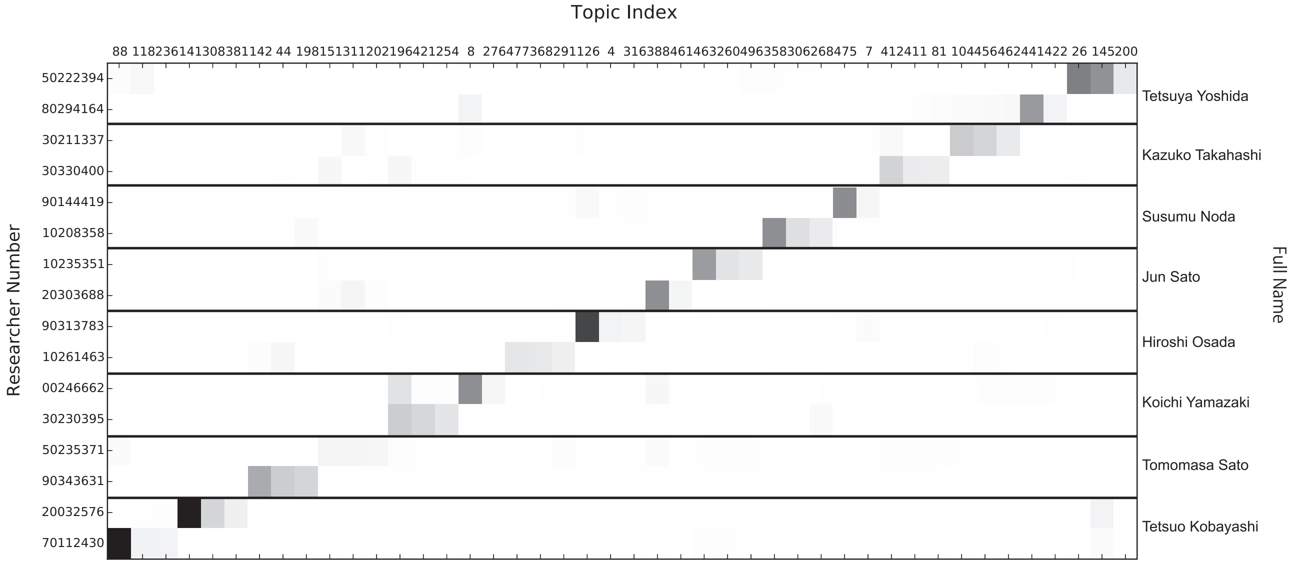


Fig. 4 A heat map of the topic representation of researcher pairs who share the same full name. Darker cells indicate a higher assignment of the corresponding topic.

posed method and these metadata-based features will better identify the correct researchers, which should be investigated in future work.

For further examination, we randomly extract pairs of researchers who share the same name and show their topic vectors as a heat map in Fig. 4. Comparing Fig. 4 and Fig. 3 (i.e., the heat map of researchers extracted from the same research field), we cannot find many topic overlaps in each pair of researchers having the same full names. This means that representing researchers' interests in a topic space can contribute to name disambiguation in a large-scale academic database including all research fields.

In this experiment, we also found some difficult cases that cannot be distinguished by the proposed method. For example, when two researchers who share the same full name are experts in the almost same research theme, our topic representation cannot accurately disambiguate the author for a given paper. For example, distinguishing between two researchers who specialize in agricultural research, *crop science*, was difficult for the proposed method. We observed that in such a case, VSM_TFIDF and VSM_All worked slightly better than the proposed method. From this observation, the distributions of stop words in a researcher's papers can be effective to characterize his/her writing style. Although our experiments removed stopwords that are generally irrelevant to the topics of papers, it could be useful to combine topic information with distinctive stylistic features for an author disambiguation problem. Thus, we will investigate the effectiveness of the use of stopwords for further improvement. Furthermore, to facilitate the management of the academic database, we should develop not only author identification among candidate researchers but also a framework that classifies whether the author is a new researcher to register in the database. In our future work, we will introduce a novel scheme that classifies the authors of a given paper into existing or new researchers to increase the usefulness of the proposed method.

6. Conclusions and Future Work

In this paper, we present a topic representation of researchers' interests in a large-scale academic database. In the proposed method, we first calculate topic distributions over words by using LDA based on the training dataset. Then, we convert all papers of a target researcher to topic vectors by using LDA, and then calculate a centroid of the topic vectors to characterize the researcher. The results of experiments conducted on the CiNii Articles show that the topics extracted by our method actually span a diverse range of research fields. In the experiments, the applicability of the presented topic representation to author disambiguation is also demonstrated: our approach achieves a mean accuracy of 92.60%, which indicates that it outperforms comparative methods that directly use textual features without dimensionality reduction or that are based on metadata.

Our work in this paper is the first step towards representing the researchers' interests in a large-scale academic database covering all research fields in Japan. Further room for investigation exists. For example, although our quantitative evaluation showed the effectiveness of simply averaging researcher's topic vectors, their weighted average or a more complicated distribution such as Gaussian mixture models might be suitable. How to summarize the researcher's topic vectors should be studied for further performance improvement.

In the experiments, we successfully extracted a set of sensible topics from short documents like abstracts, as well as the conventional methods on topic modeling [17], [24]. LDA has also been used to analyze other short documents, e.g., IMDB movie plots or reviews [30], [31]. However, some papers argued that extracting meaningful topics from short documents sometimes fails due to the lack of word co-occurrences [32]. To avoid the data sparsity problem caused by very short abstracts, using external information resources such as Wikipedia can reduce the sparsity [33]. In addition, introducing additional variables such as journal/conference names to topic modeling can also provide good features to characterize researchers. We will investigate the effectiveness of additional features and information sources for topic modeling in future work.

Adaptively combining the presented topic representation with bibliographic metadata will be effective for improving the performance of author disambiguation by the proposed method. Related to this, it will also be useful to develop a novel topic model that can consider the personalities of the researcher's writings. Furthermore, the availability of the presented topic representation is not limited to author disambiguation. Other applications, including researcher network construction and collaborator recommendation, can be realized with the use of the researchers' topic vectors. In addition, visualizing the research theme transitions of a particular researcher or a research group can also provide us new insights. Thus, our future work also includes developing these applications to promote interdisciplinary research and facilitate collaboration and research activities.

References

[1] H. Small, "Co-citation in the scientific literature: A new measure of the relationship between two documents," J. Am. Soc. Inf. Sci., vol.24, no.4, pp.265–269, 1973.

[2] K.W. Boyack and R. Klavans, "Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately?," J. Am. Soc. Inf. Sci. Technol., vol.61, no.12, pp.2389–2404, 2010.

[3] D. Hall, D. Jurafsky, and C.D. Manning, "Studying the history of ideas using topic models," Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), pp.363–371, 2008.

[4] Y. Ding, E. Yan, A. Frazho, and J. Caverlee, "PageRank for ranking authors in co-citation networks," J. Am. Soc. Inf. Sci. Technol., vol.60, no.11, pp.2229–2243, 2009.

[5] K. Lu and D. Wolfram, "Measuring author research relatedness: A comparison of word-based, topic-based, and author cocitation approaches," J. Am. Soc. Inf. Sci. Technol., vol.63, no.10, pp.1973–1986, 2012.

[6] R. Ichise, H. Takeda, and K. Ueyama, "Community mining tool using bibliography data," Proc. Int. Conf. Information Visualisation (IV), pp.953–958, 2005.

[7] J. Li, F. Xia, W. Wang, Z. Chen, N.Y. Asabere, and H. Jiang, "ACRec: A co-authorship based random walk model for academic collaboration recommendation," Proc. Companion Publication of the 23rd Int. Conf. World Wide Web Companion, pp.1209–1214, 2014.

[8] K. Börner, L. Dall'Asta, W. Ke, and A. Vespignani, "Studying the emerging global brain: Analyzing and visualizing the impact of co-authorship teams," Complexity, vol.10, no.4, pp.57–67, 2005.

[9] H.D. White and B.C. Griffith, "Author cocitation: A literature measure of intellectual structure," J. Am. Soc. Inf. Sci., vol.32, no.3, pp.163–171, 1981.

[10] D.R. Radev, P. Muthukrishnan, V. Qazvinian, and A. Abu-Jbara, "The ACL anthology network corpus," Language Resources and Evaluation, vol.47, no.4, pp.919–944, 2013.

[11] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet allocation," J. Machine Learning Research, vol.3, pp.993–1022, March 2003.

[12] H. Han, H. Zha, and C.L. Giles, "Name disambiguation in author citations using a K-way spectral clustering method," Proc. 5th ACM/IEEE-CS Joint Conf. Digital Libraries (JCDL), pp.334–343, June 2005.

[13] J. Huang, S. Ertekin, and C.L. Giles, "Efficient name disambiguation for large-scale databases," Knowledge Discovery in Databases: PKDD 2006, Lecture Notes in Computer Science, vol.4213, pp.536–544, Springer Berlin Heidelberg, 2006.

[14] D.A. Dervos, N. Samaras, G. Evangelidis, J. Hyvärinen, and Y. Asmanidis, "The universal author identifier system (UAI_Sys)," Proc. Int. Scientif. Conf., eRa: Contribution of Information Technology in Science, Economy, Society and Education, 2006.

[15] K. Kurakawa, H. Takeda, M. Takaku, A. Aizawa, R. Shiozaki, S. Morimoto, and H. Uchijima, "Researcher name resolver: Identifier management system for Japanese researchers," International Journal on Digital Libraries, vol.14, no.1-2, pp.39–58, 2014.

[16] S.W. Thomas, B. Adams, A.E. Hassan, and D. Blostein, "Studying software evolution using topic models," Science of Computer Programming, vol.80, pp.457–479, Feb. 2014.

[17] M. Rosen-Zvi, C. Chemudugunta, T. Griffiths, P. Smyth, and M. Steyvers, "Learning author-topic models from text corpora," ACM Trans. Information Systems, vol.28, no.1, pp.4:1–4:38, Jan. 2010.

[18] N.R. Smalheiser and V.I. Torvik, "Author name disambiguation," Annual Review of Information Science and Technology, vol.43, no.1, pp.1–43, Jan. 2009.

[19] Thomson Reuters Science, "Distinct author identification system." http://images.webofknowledge.com/WOKRS512B4/help/WOS/hp_das1.html, Last accessed: 21/06/2015.

[20] Scopus, "Scopus author identifier." http://help.scopus.com/Content/h_autsrch_intro.htm, Last accessed: 21/06/2015.

[21] National Institute of Informatics, "CiNii Articles — About CiNii Articles." http://support.nii.ac.jp/en/cia/cinii_articles, Last accessed: 21/06/2015.

[22] L. Tang and J.P. Walsh, "Bibliometric fingerprints: Name disambiguation based on approximate structure equivalence of cognitive maps," Scientometrics, vol.84, no.3, pp.763–784, 2010.

[23] L.L. Haak, M. Fenner, L. Paglione, E. Pentz, and H. Ratner, "ORCID: A system to uniquely identify researchers," Learned Publishing, vol.25, no.4, pp.259–264, Oct. 2012.

[24] T.L. Griffiths and M. Steyvers, "Finding scientific topics," Proc. National Academy of Sciences, vol.101, no.suppl 1, pp.5228–5235, 2004.

[25] D. Newman, A. Asuncion, P. Smyth, and M. Welling, "Distributed algorithms for topic models," J. Machine Learning Research, vol.10, pp.1801–1828, Dec. 2009.

[26] G. Heinrich, ""infinite LDA" — Implementing the HDP with minimum code complexity," Tech. Rep. 170, Technical University Darmstadt, 2011.

[27] G. Salton, A. Wong, and C.S. Yang, "A vector space model for automatic indexing," Commun. ACM, vol.18, no.11, pp.613–620, Nov. 1975.

[28] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," Information Processing & Management, vol.24, no.5, pp.513–523, 1988.

[29] T. Gurney, E. Horlings, and P.V.D. Besselaar, "Author disambiguation using multi-aspect similarity indicators," Scientometrics, vol.91, no.2, pp.435–449, May 2012.

[30] S. Maneeroj and A. Takasu, "Hybrid recommender system using latent features," Proc. Int. Conf. Advanced Information Networking and Applications Workshops (WAINA), pp.661–666, May 2009.

[31] J. Liu, S. Cyphers, P. Pasupat, I. McGraw, and J.R. Glass, "A conversational movie search system based on conditional random fields," Proc. INTERSPEECH, pp.2454–2457, 2012.

[32] L. Hong and B.D. Davison, "Empirical study of topic modeling in Twitter," Proc. First Workshop on Social Media Analytics (SOMA),

pp.80–88, 2010.

[33] S. Banerjee, K. Ramanathan, and A. Gupta, "Clustering short texts using wikipedia," Proc. Annual Int. ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), pp.787–788, 2007.

Appendix

This appendix describes an overview of CiNii Articles, the subject of this paper's experiments. CiNii Articles is a database service maintained by the National Institute of Informatics that can search for information in academic articles published in academic society journals, university research bulletins or articles included in the National Diet Library's Japanese Periodicals Index Database [21]. As of June 23, 2015, it contains 19,042,439 articles, covering all research fields including the humanities, law, economics, pure sciences, engineering, agriculture, and medicine[†]. Duplicate information from different databases can be treated as a single article due to article identifiers.

In CiNii Articles, for each author name *a*, papers written by persons whose names are *a* are first clustered to discover groups of papers authored by identical researchers; this is done in such a way that precision is kept very high. Then, to accurately merge paper clusters for the same identical researcher, CiNii Articles uses a publication attached to grant reports that have been submitted by researchers to KAKEN. KAKEN is the awards database of the Grants-in-Aid for Scientific Research administered by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) and the Japan Society for the Promotion of Science (JSPS). After each researcher who obtained a grant submits a report with a publication list to KAKEN, the researcher is linked to clusters that include papers contained in the submitted publication list. Thus, the current system requires a manual procedure to associate researchers with papers. Note that we can easily access the publication information and grant information of a target researcher using Researcher Name Resolver[††], which is a researcher identifier management system [15]. Figure A· 1 provides an overview of the researcher identifiers in CiNii Articles, Researcher Name Resolver, and KAKEN. Because these databases cover all research fields in Japan, there are many authors who share the same names. Thus, author disambiguation is a major issue that needs to be solved in order to provide information about researcher activities. By developing our method (i.e., by representing researchers' interests in a topic space), we can facilitate the usability of searching the publications or activities of a target researcher.

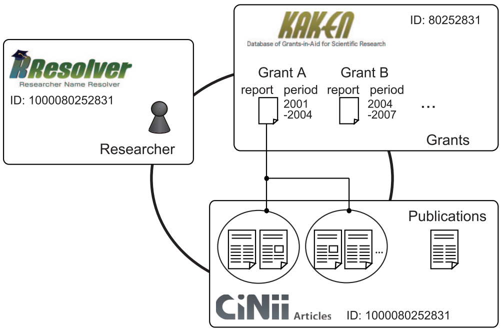


Fig. A· 1 An overview of author identifiers in CiNii articles, researcher name resolver, and KAKEN.

[†] http://ci.nii.ac.jp/cinii/servlet/DirTop?lang=en

[††] http://rns.nii.ac.jp/

Marie Katsurai received the B.S. degree in Engineering, the M.S. degree and the Ph.D. degree in Information Science and Technology from Hokkaido University in Sapporo, Japan in 2010, 2012, and 2014, respectively. She was a Research Fellow of the Japan Society for the Promotion of Science from 2013 to 2015, and she was with the National Institute of Informatics from 2014 to 2015. She is currently an Assistant Professor in the Department of Information Systems Design, Doshisha University. Her research interests include multimedia information retrieval and data mining. She is a member of the IEICE, IEEE, and ACM.

Ikki Ohmukai received his Ph.D. degree in informatics from the Graduate University for Advanced Studies in 2005. He joined National Institute of Informatics in 2005 and has been an Associate Professor since 2009. His research interests are the semantic web and social media. He is a member of IPSJ and JSAI.

Hideaki Takeda is a professor at National Institute of Informatics (NII) Japan, and a professor at the Graduate University for Advanced Studies (Sokendai). He received B.Eng., M.Eng. and Dr.Eng. degrees in Precision Machinery Engineering from the University of Tokyo, Japan, in 1986, 1988 and 1991, respectively. He worked at the Norwegian Institute of Technology and the Nara Institute of Technology prior to joining the current institution. He has been the Sumitomo Endowed Professor in the University of Tokyo between 2005 and 2010. His research interests include the semantic Web, knowledge sharing systems and design theory.