Measuring Researcher Relatedness with Changes in Their Research Interests

Hiroyuki Nishizawa*, Marie Katsurai*, Ikki Ohmukai†, and Hideaki Takeda†
* Doshisha University, Kyoto, Japan
E-mail: {nishizawa, katsurai}@mm.doshisha.ac.jp Tel: +81-0774-65-7575
† National Institute of Informatics, Tokyo, Japan
E-mail: {i2k, takeda}@nii.ac.jp

***Abstract*—Relevant researcher recommendation is important for finding potential research collaborators, and several existing methods measure researcher relatedness based on their research interests. Our previous works represented a researcher with a single multidimensional topic vector calculated from the researcher's publications, ignoring the publication dates. On the other hand, recent studies on information recommendation have shown the effectiveness of modeling changes in user preferences over time. Thus, this paper proposes a new representation of researchers, which consists of *yearly topic vectors*. To measure the relatedness between researchers, we calculate the similarity between two sequences of topic vectors using Dynamic Time Warping. An experimental example visualizes topic transitions of a target researcher and demonstrates that the proposed method can effectively find researchers whose topic transitions are similar over time, when compared to the conventional method.**

I. Introduction

As a research theme becomes more complicated, the range of knowledge necessary for the research becomes greater. Collaborative research by multiple researchers has the potential to provide innovative solutions to difficult research subjects. There are numerous discussions about the relationship between collaborative research and productivity, as well as the effect of promoting collaborative research. For example, Lee and Bozeman [2] investigated the impacts of several types of collaborations on publishing productivity. Rijnsoever and Hessels [3] showed a positive relationship between innovativeness and disciplinary research collaboration. Thus, ways of encouraging collaboration have received much attention.

In computer science research, various methods for efficiently recommending relevant researchers have been proposed. These methods require the measurement of the relatedness between two researchers. Several methods calculated a social distance among researchers using existing coauthorship networks [6], while content-based similarity among researchers also plays an important role [5], [7]. In our previous study [7], we converted the textual features of a researcher's publications to topic vectors, and calculated the average of the vectors to summarize the researcher's interest. The experiments in [7] showed the importance of the content-based relatedness measure, which can find relevant researchers beyond the existing social relationships. In this previous work, the publication date of each paper was not considered in characterizing researchers, and only a single topic vector was assigned to an individual. On the other hand, in research on general information recommendation, it is known that changes in user preferences over time should be taken into account when constructing a user's profile [8], [9]. Inspired by these works, this paper presents a new representation of researchers, which describes transitions of research interests.

In the proposed method, we first extract topic vectors from each researcher's publications. Then, we calculate the average of the topic vectors by each year, producing a sequence of research interests. This is a major difference between the proposed method and our previous work [7]. Given a pair of researchers, our method uses Dynamic Time Warping (DTW) [1] to calculate the similarity between the corresponding two series of topic vectors. Our new measure can fully exploit topic transition information to find relevant researchers. Based on experiments conducted on papers published in Japan, we present an experimental example to validate whether the proposed method can search for researchers whose changes in research interests follow a pattern similar to those of a target researcher.

In summary, the main contributions of this paper are twofold. First, we introduce a new researcher representation based on publication dates to extend our previous work [7]. Second, our case study shows a visualization of topic transitions of a target researcher and demonstrates the effectiveness of the proposed method, compared to the conventional method.

II. Related Work

Characterizing researchers with their expertise and research interests is necessary for facilitating information retrieval techniques in academic databases. The research topics are generally estimated using textual features from titles and abstracts of researchers' publications, which are easily available in the databases. There are many methods that exploit topic models to represent the textual features in a low-dimensional feature space. For example, Song et al. [12] introduced variables representing authors into a topic model for name disambiguation in a bibliographic database. Similarly, Lu and Wolfram [10] used an author topic model to represent each researcher by a multinomial distribution over topics. Yan et al. [11] identified topics of research communities and analyzed the dynamics of community structures. Tang et al. [5] developed a topic model that learns a set of topics from

Fig. 1. Outline of the proposed method.

collaboration examples for collaborator recommendation. With the aim to construct a knowledge base for academic data analysis, our previous works presented a new framework that assigns a topic representation to researchers in a large-scale academic database covering all research fields in Japan [7], [13], [14]. We demonstrated the applicability of the proposed topic representation to the author disambiguation problem [13] and collaboration relationship prediction [7]. To facilitate such researcher profiling, this paper introduces a new representation that reflects the publication history.

The work that is most related to this paper is Kong et al. [15], which characterized researchers with topics over time. That method measured the relatedness between researchers by accumulating the topical similarity at the same year. However, in the case that two researchers worked in the same topic but in different time periods, this measure fails in finding their potential relevance. On the other had, we do not perform the similarity computation in exactly the same periods; our method detects similar patterns of changes in research interests.

III. Proposed Method

This section presents a method for measuring researcher relatedness with changes in their research interests. Figure 1 shows an overview of the proposed method. For each researcher, we extract a sequence of topic vectors from the researcher's publications (see Section III-A). Then, we calculate the relatedness between two researchers as the similarity between the corresponding sequences (see Section III-B).

A. Characterizing researchers with sequences of topic vectors

Academic papers written by a target researcher are supposed to reflect the researcher's specific interests. Therefore, we treat the title and abstract of a paper as a single document, from which we extract textual features. Given a collection of documents with author information, we first construct a vocabulary of unique words. After removing words having the highest or lowest frequency as stop words, we calculate Bag-of-Words (BoW) for each paper.

The dimensionality of the BoW feature space is usually very large. To reduce the dimensions and grasp the main subjects of documents, we use Latent Dirichlet Allocation (LDA) [16], which is one of the typical topic models. We learn the LDA model to extract T topics from all the papers in the dataset and calculate the topic distribution over words. Using LDA, we represent each document as a vector of T topic scores.

Next, we divide a set of documents written by researcher a by year; the resulting set for year y is denoted by $S_{a,y}$. For each document in $S_{a,y}$, we calculate its topic vector with the learned LDA model. Finally, topic vectors in $S_{a,y}$ is averaged as follows:

$$\boldsymbol{v}_{a,y} = \frac{1}{|S_{a,y}|} \sum_{d \in S_{a,y}} \boldsymbol{\theta}_d, \tag{1}$$

where $\boldsymbol{\theta}_d \in \mathbb{R}^T$ is the topic vector of document d. We use $\boldsymbol{v}_{a,y}$ as a topic vector of researcher a at year y.

Focusing on each element of vector $\boldsymbol{v}_{a,y}$, some topics may be assigned very small values. Such topics are considered to be unimportant in expressing the researchers' interests. Therefore, we apply thresholding on the element $v_{a,y,t}$ for the t-th topic as follows:

$$v_{a,y,k} = \begin{cases} v_{a,y,k} & (v_{a,y,k} \geq Threshold) \\ 0 & (v_{a,y,k} < Threshold) \end{cases} \tag{2}$$

As a result, weak topics are discarded, and only main topics representing the researcher's major interests are left. We perform L2 normalization on the vector $\boldsymbol{v}_{a,y}$. Finally, during time interval $[y_1, y_n]$, research interests of researcher a is represented as: $\{\boldsymbol{v}_{a,y_1}, \boldsymbol{v}_{a,y_2}, \cdots, \boldsymbol{v}_{a,y_n}\}$.

B. Measuring researcher relatedness

This subsection describes how we measure the relatedness between researchers based on their sequences of topic vectors. Our aim is to detect the relevance of two publication histories with time lags. DTW is one of popular algorithms for measuring similarity between sequences. It consists in finding the optimal alignment between the two sequences and then accumulating the individual vector-to-vector distances along the alignment.

Given a pair of researchers a and b, For each (y_i, y_j) $(i, j > 1)$, we calculate the cumulative score $\gamma_{a,b}(y_i, y_j)$ as follows:

$$\begin{aligned} \gamma_{a,b}(y_i, y_j) =& Sim(\boldsymbol{v}_{a,y_i}, \boldsymbol{v}_{b,y_j}) + \\ & \max[\gamma_{a,b}(y_{i-1}, y_j), \gamma_{a,b}(y_{i-1}, y_{j-1}), \gamma_{a,b}(y_i, y_{j-1})], \end{aligned} \tag{3}$$

where $Sim(\boldsymbol{v}_{a,y_i}, \boldsymbol{v}_{b,y_j})$ is the similarity between vectors $\boldsymbol{v}_{a,y_i}$ and $\boldsymbol{v}_{b,y_j}$. In this paper, we exploit the cosine similarity. Starting from (y_n, y_n), backtracking along the maximum score index pairs yields the optimal alignment called warping path. Finally, the value of $\gamma_{a,b}(y_n, y_n)$ obtained with the warping path is used as the relatedness between researchers a and b.

IV. Experiments

This section presents results of relevant researcher recommendations that verify the effectiveness of the proposed method. We describe the details of the dataset used and the results of the recommendation in Sections IV-A and IV-B, respectively.

A. Dataset

To construct a list of researchers for experiments, we first collected an initial set of papers published from 2005 to 2009 in a certain domestic conference regarding information science and communication from CiNii Articles[1]. Because CiNii

[1] https://ci.nii.ac.jp/en

TABLE I
THE MOST FREQUENT WORDS IN THE DATASET; WORDS IN ITALICS ARE ENGLISH TRANSLATIONS.

確認 (*confirmation*), 環境 (*environment*), 情報 (*information*), 有効 (*valid*), 通信 (*communication*), 方法 (*approach*), 評価 (*valuation*), データ (*data*), 本稿 (*this paper*), 実現 (*realization*), 実験 (*experiment*), 方式 (*technique*), 特性 (*special quality*), 可能 (*possibility*), 技術 (*technology*), 論文 (*paper*), ネットワーク (*network*), 処理 (*processing*), 画像 (*image*), 手法 (*method*), シミュレーション (*simulation*), 開発 (*development*), 検討 (*examination*), 一般 (*generally*), 計算 (*calculat*), 利用 (*use*), システム (*system*), 提案 (*suggest*), 報告 (*report*), 制御 (*control*), 問題 (*matter*), 設計 (*design*), 必要 (*necessary*), 解析 (*analysis*), 研究 (*research*), モデル (*model*)

Articles have an author identifier assignment system, we were able to obtain each individual researcher's publication history with high precision. Each paper in CiNii Articles generally has author IDs, the title of the paper, the abstract, publication venue, and the publication date. We removed papers that lack at least one of these attributes. Most of the papers in CiNii Articles are written in Japanese, while some papers only have English titles and abstracts. We also removed these English papers to avoid the necessity of word translation. The resulting dataset contains 13,643 Japanese papers written by 583 authors. We applied the Japanese morphological analysis engine MeCab[2] to the texts and then extracted noun words only. To reduce the noise, we further applied the following preprocessing:

- Remove symbols.
- Remove numbers.
- Remove double-byte spaces.
- Remove parentheses.
- Remove URI strings.
- Convert half-width katakana to full-width katakana.
- Convert full-width alphabet to half-width alphabet.
- Convert uppercase alphabet to lowercase.

Finally, we obtained a vocabulary of 32,014 unique words. Table I shows the most frequent words used as stop words. In the LDA model, we set the number of topics as $T = 150$. The threshold value in Eq. (2) was set to 0.1.

B. Case study of relevant researcher recommendation

Because there is no ground truth about relevant researcher recommendation, this paper presents an experimental example for a target researcher (ID: 1000000127143). Figure 2 visualizes the sequences of topic vectors for a query researcher and the top three relevant researchers found by the proposed method. In these heatmaps, the darker cells indicate a higher assignment of the corresponding topic. For reference, the word distributions of eight topics found for the query researcher are shown in Table II. In Fig. 2, we can find that the query researcher specialized in topics 22, 42, and 50 in 2005, and the main interests changed to topic 104 in 2007, and to topics 131 and 141 in 2009. As shown in fig. 2(b)-(d), The top three relevant researchers found by the proposed method show the topic transitions similar to that of the query researcher. For

[2]http://taku910.github.io/mecab/

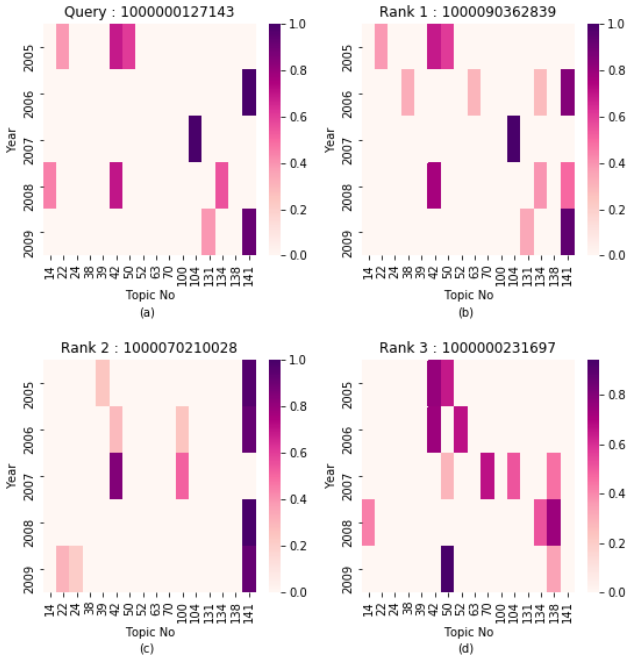

Fig. 2. Heatmaps of the yearly topic representation for (a) a query researcher and (b)(c)(d) the top three relevant researchers found by the proposed method. Darker cells indicate a higher assignment of the corresponding topic.

performance comparison, we also show a result obtained using the conventional method [7]. Specifically, the conventional method assigned a single topic vector to a researcher without considering publication dates. For given a pair of researchers, it calculated the researcher relatedness measure using the cosine similarity between the corresponding two vectors. Figure 3 visualizes the sequences of topic vectors for the query researcher and the top three relevant researchers found by the conventional method [7]. In the figure, the topic vectors estimated using the conventional method have actually weak topics, but these were removed in our visualization (especially in Fig. 3(d)). Compared with the result of the proposed method, the topic transitions of the top three researchers found by the conventional method are incoherent to that of the query researcher. This demonstrates that aggregating all papers to a single topic vector makes the transition information obscure, and the resulting similarity is affected by the sum of similarities over weak topics. The experiment in this paper corresponds to a preliminary investigation of the effectiveness of the proposed researcher representation with a small dataset. We will perform more experiments such as representing researchers with yearly topics over a long span of time.

V. CONCLUSIONS AND FUTURE WORK

This paper presented a new representation of researchers with yearly topic vectors. As a researcher relatedness measure, we used DTW to calculate the similarity between two sequences of topic vectors. In the experiments conducted using

TABLE II
EXAMPLES OF TOPICS ESTIMATED USING LDA. OUR INTERPRETATION OF THE TOPICS ARE INDICATED IN **BOLD** TEXT. ONLY ENGLISH WORDS TRANSLATED BY OURSELVES ARE SHOWN.

Topic 14 **Software**		Topic 22 **Medical**		Topic 42 **Interface**		Topic 50 **Motion perception**	
Word	Prob.	Word	Prob.	Word	Prob.	Word	Prob.
action	0.141	example	0.029	presentation	0.075	color	0.058
reliability	0.135	diagnosis	0.024	vision	0.056	posture	0.052
software	0.043	right	0.023	haptics	0.039	walking	0.036
development	0.031	case	0.022	stimulation	0.037	change	0.025
simulator	0.025	inspection	0.020	human	0.034	leg	0.023
test	0.023	left	0.020	sense	0.029	marker	0.022

Topic 104 **Volume imaging**		Topic 131 **Coding**		Topic 134 **Medical effect**		Topic 141 **Bioimaging**	
Word	Prob.	Word	Prob.	Word	Prob.	Word	Prob.
velocity	0.062	coding	0.081	medical	0.073	dimension	0.054
quantity	0.042	block	0.040	effect	0.028	image	0.042
image	0.032	image	0.040	machine	0.024	surgery	0.030
change	0.031	predict	0.037	tissue	0.020	type	0.025
display	0.031	frame	0.035	additive	0.020	nerve	0.021
visualize	0.030	compression	0.033	influence	0.019	mri	0.016

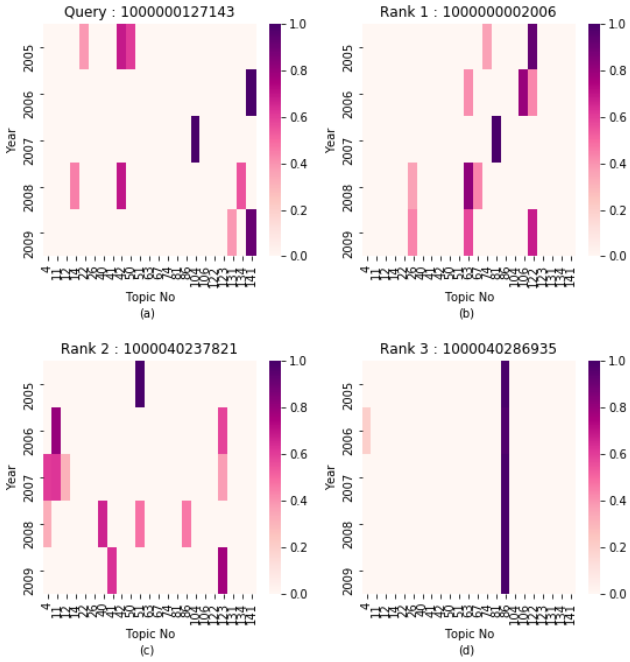

Fig. 3. Heatmaps of the yearly topic representation for (a) a query researcher and (b)(c)(d) the top three relevant researchers found by the conventional method [7]. Darker cells indicate a higher assignment of the corresponding topic.

a paper collection, we demonstrated that the proposed method can effectively find relevant researchers whose interests change over time in similar way to a query researcher. The scope of our project including this paper and the conventional methods [7], [13] is to profile interests of researchers in all disciplines in Japan. It will be necessary to use not only Japanese papers but also English papers to characterize researchers. There is still room for future study in this direction. Our future work includes increasing the size of the experiment dataset and constructing yearly topic vectors of researchers with a long span of time. In addition, we will quantitatively evaluate the performance of the proposed method via investigating user satisfaction and in several applications such as collaborator prediction. It would be valuable to apply the proposed method to retrieval and data mining, e.g., constructing a researcher retrieval interface and mining frequent patterns of research topic transitions.

VI. ACKNOWLEDGMENT

This work was supported in part by a Grant-in-Aid for Young Scientists (B) 17K12794 and the open collaborative research program at National Institute of Informatics (NII) Japan (FY2018).

REFERENCES

[1] D. J. Berndt and J. Clifford, "Using Dynamic Time Warping to Find Patterns in Time Series," *In Proc. of Int. Conf. Knowledge Discovery and Data Mining (AAAIWS'94)*, vol. 10, pp. 359–370, 1994.
[2] S. Lee and B. Bozeman, "The Impact of Research Collaboration on Scientific Productivity," *Social Studies of Science*, vol. 35, pp. 673–702, 2005.
[3] F. J. van Rijnsoever and L. K. Hessels, "Factors associated with disciplinary and interdisciplinary research collaboration," *Research Policy*, vol. 40, no. 4, pp. 463–472, 2011.
[4] W. Wang, Z. Cui, T. Gao, S. Yu, X. Kong and F. Xia, "Is Scientific Collaboration Sustainability Predictable?," *In Proc. of Int. Conf. World Wide Web Companion*, pp. 853–854, 2017.
[5] J. Tang, S. Wu, J. Sun and H. Su, "Cross-domain Collaboration Recommendation," *In Proc. of ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD)*, pp. 1285–1293, 2012.
[6] 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," *In Proc. of Int. Conf. on World Wide Web*, pp. 1209–1214, 2014.
[7] M. Araki, M. Katsurai, I. Ohmukai and H. Takeda, "Interdisciplinary Collaborator Recommendation Based on Research Content Similarity," *IEICE Trans. on Information and Systems*, pp. 785–792, 2017.
[8] H. Liang, Y. Xu, D. Tjondronegoro and P. Christen, "Time-aware Topic Recommendation Based on Micro-blogs," *In Proc. the 21st ACM Int. Conf. Information and Knowledge Management*, pp. 1657 – -1661, 2012.
[9] M. Canut, S. On-At, A. Péninou and F.Sèdes, "Time-aware Egocentric network-based User Profiling," *In Proc. of IEEE/ACM Int. Conf. Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 569–572, 2015.
[10] K. Lu and D. Wolfram, "Measuring Author Research Relatedness: A Comparison of Word-based, Topic-based, and Author Cocitation Approaches," *Journal of the American Society for Information Science and Technology*, pp. 1973–1986, 2012.
[11] E. Yan, Y. Ding, S. Milojević, and C. R. Sugimoto, "Topics in dynamic research communities: An exploratory study for the field of information retrieval," *Journal of Informetrics*, pp. 140–153, 2012.
[12] Y. Song, J. Huang, I. G. Councill, J. Li, and C. L. Giles, "Efficient Topic-based Unsupervised Name Disambiguation," *In Proc. of ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*, pp. 342–351, 2007.
[13] M. Katsurai, I. Ohmukai and H. Takeda, "Topic Representation of Researchers' Interests in a Large-Scale Academic Database and Its Application to Author Disambiguation," *IEICE Trans. on Information and Systems*, pp. 1010-1018, 2016.
[14] M. Katsurai, I. Ohmukai and H. Takeda, "Topic Representation of Researchers' Interests in a Large-Scale Academic Database," *In Proc. Int. Tech. Conf. Circuits Systems Computers and Communications (ITC-CSCC)*, pp. 35-37, 2015.
[15] X. Kong, H. Jiang, W. Wang and T. M. Bekele, Z. Xu and M. Wang, "Exploring dynamic research interest andacademic influence for scientific collaborator recommendation," *Scientometrics*, pp. 369–385, 2017.
[16] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *J. Mach. Learn. Res*, vol. 3, pp. 993-1022, March 2003.