Researcher Network Visualization Using Matrix Researcher2vec

Hirata, Enna
Yamashita, Takahiro
Ozawa, Seiichi

(Citation)
Journal of Advanced Computational Intelligence and Intelligent Informatics,27(4):603-608

(Issue Date)
2023-07-20

(Resource Type)
journal article

(Version)
Version of Record

(URL)
https://hdl.handle.net/20.500.14094/0100483040

https://doi.org/10.20965/jaciii.2023.p0603

Research Paper:

Researcher Network Visualization Using Matrix Researcher2vec

Enna Hirata†, Takahiro Yamashita, and Seiichi Ozawa

Kobe University
1-1 Rokkodai-cho, Nada-ku, Kobe, Hyogo 657-8501, Japan
E-mail: enna.hirata@platinum.kobe-u.ac.jp
†Corresponding author

[Received December 8, 2022; accepted March 15, 2023]

In this study, we introduce a system called Matrix Researcher2vec (MResearcher2vec) which generates researcher embedding vectors from their papers and research projects in researchmap and KAKENHI databases. The system includes data on 276,841 researchers, 6,161,592 papers, and research projects. Utilizing natural language processing techniques, the MResearcher2vec model extracts researcher vectors from the papers and research project summaries of KAKENHI grant recipients. The similarity between reseachers is then computed to visualize inter-researcher relationships. The machine learning results have been integrated into a web service, providing a novel approach for academic relationship mining. It can be applied in the matching of research contents and researchers in evaluation of industry-government-academia collaboration and joint research. It contributes in four aspects: (1) exchanges between researchers, (2) creation of opportunities for researchers and companies to connect, (3) further promotion of interdisciplinary research, and (4) reduction of lost opportunities for research institutions to acquire talents.

Keywords: machine learning, document vectors, researcher similarity, natural language processing, Matrix Researcher2vec

1. Introduction

The focus of this study is the evaluation of the similarity between the research contents of individuals registered in the Grants-in-Aid for Scientific Research (KAKENHI), and researchmap (RM) databases. The KAKENHI database (KAKENDB) encompasses a wide range of academic fields and features up-to-date research information, consisting of research reports and self-evaluation reports. On the other hand, the RM database (RMDB) contains data about researchers, such as their name, affiliation, department, and publications, including papers, books, other publications, lectures, oral presentations, works, and research projects.

A novel approach, which is referred to as Matrix Researcher2vec (MReseacher2vec), was introduced to extract researcher embedding vectors from their research contents sourced from the KAKENDB and RMDB. Our research indicates that the similarity between research projects is accurately calculated using this method, and the outcomes obtained are in close agreement with subjective evaluations.

In addition, we created a researcher matching system that utilizes short paragraph or keyword phrases.

The paper is structured as follows. Section 2 presents a survey of previous studies that utilize natural language processing (NLP) techniques to examine academic connections. Section 3 elaborates on the attributes of the data used in the study. Section 4 details the proposed method and provides insight into the implementation outcomes. Section 5 offers a discourse on the critical findings, highlighting the importance of this research for practical applications and future investigations.

2. Related Works

Recent advances in NLP and text mining have led to a growing interest among researchers to develop applications utilizing text processing methods. In this context, academic relationship mining has gained significant attention in the era of big data. In particular, researcher similarity has been extensively investigated. For instance, Nishizawa et al. [1] have introduced temporal changes in research content and affiliations as features in researcher profiling for collaborator recommendation. They have computed the topic vector of a researcher's publications in each year and represented researcher's interests using a series of topic vectors. Okuma and Kiyoki [2] have proposed a system that extracts and presents companies that are involved in research and business similar to the researcher's research fields. They have employed a topic model to calculate the correlation between researchers' papers and documents such as technical bulletins of companies, by converting them into a vector of the degree of fitting to each topic group of the topic model.

Wang et al. [3] have introduced a novel approach for mining academic relationships, which is based on attributed collaboration network embedding. Their method learns a low-dimensional representation of scholars, while considering both scholar attributes and network topology

together.

Most recently, Scholar2vec [4] has been proposed for presenting scholar profiles through neural network embedding. In Scholar2vec, scholar vectors are generated from textual information, such as demographics, research, and influence, by mapping them to scholars' research interest vectors, and measuring similarity between scholars using vector representation.

In a similar vein, Mochihashi [5] has investigated the similarity between researchers by showing the equivalence of Word2vec to matrix factorization of the pointwise mutual information matrix. Mochihashi has proposed Researcher2vec, which computes the neural document vector of each paper by singular value decomposition to obtain the researcher vector.

As far as we are aware, there is a lack of research that calculates the similarity between researchers in a large database while taking into account their research keywords and content. Our proposed method aims to fill this research gap by generating a vector for each researcher based on a matrix of papers and research projects.

3. Description of Data

The data used in this study are sourced from the following two databases.

3.1. RMDB

The RMDB[1] offers a Web API that facilitated the development of a directory of researchers and a performance management system intended for universities and research institutes. It is worth noting, however, that the database has a constraint in which it can only retrieve information on researchers who affiliate with the same research institution as users, in this instance, Kobe University. To overcome this limitation, we have integrated data from the KAKENDB to obtain information on researchers who are associated with institutions other than Kobe University.

3.2. KAKENDB

The KAKENDB[2] includes information on research projects that were conducted under the purview of the Ministry of Education, Culture, Sports, Science and Technology (MEXT) and the Japan Society for the Promotion of Science (JSPS) through the KAKENHI program. This information covers various aspects, such as the adopted proposals at the time of initial selection, research implementation status reports, research performance reports, summaries of research result reports, research result reports, and self-evaluation reports. It is important to note that the database encompasses data for the past decade.

1. https://researchmap.jp/public/about/operations?lang=en [Accessed November 25, 2022]
2. https://support.nii.ac.jp/en/kaken [Accessed November 25, 2022]

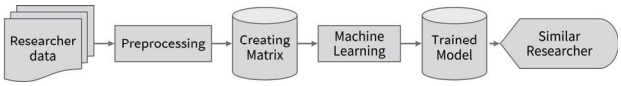


Fig. 1. Analysis process.

3.3. Data Pre-Processing

As part of data preprocessing stage, our first step involves eliminating non-disclosure researchers from the data that has been obtained. We selectively utilize certain pieces of information from both the RMDB and the KAKENDB. Specifically, we extract the researcher's name, affiliation, article title, and article abstract from the RMDB. Similarly, we collect the researcher's name, researcher's ID, researcher's affiliation, project subject, and summary of project results from the KAKENDB.

3.3.1. Translation from English to Japanese

To translate the English contents into Japanese, we use the DeepL[3] translation tool. Initially, we explored the possibility of adopting a method for generating word embeddings across multiple languages. However, we discovered a limitation in this approach, that is, words with similar meanings between languages may not always appear in similar contexts. In previous study [6], it was observed that while an accuracy of 81.4% was achieved in translating words between English and Spanish–two languages with similar grammar and etymology, only 1.7% accuracy was achieved in translating words between English and Japanese, which have very distinct grammar and etymology. For this reason, we decide to create a Japanese corpus by translating English content into Japanese.

The analysis process is summarized in **Fig. 1**.

In the analysis, we excluded researchers who contributed less than five papers to achieve better accuracy.

3.3.2. Part of Speech Processing

The data pre-processing phase consists of three steps. Firstly, we generate the corpus. Secondly, we employ the technique of tokenization, which involves breaking down a sentence or document into smaller units known as tokens. Lastly, we undertake the task of eliminating stop words, which are the most frequently occurring words in a given natural language, such as "a," "the," "is," "in," etc., in the case of English. These stop words are generally removed as they do not convey significant meaning in a sentence, and their exclusion can lead to improved accuracy without compromising the sentence's overall meaning.

The part of speech (POS)-tagging process categorizes words based on their grammatical function, and this can be accomplished using the natural language toolkit (NLTK), as shown in in **Table 1**. To enhance the performance of our model, we selectively excluded words that did not belong to noun, verb, or adjective category. This selection was performed by utilizing POS-tagging.

3. https://www.deepl.com/ja/docs-api/ [Accessed November 25, 2022]

Table 1. Collection of POS tags in NLTK.

Tag	Meaning	English examples
ADJ	adjective	*new, good, innovative, big*
ADP	adposition	*on, at, in, of, with, by, under*
ADV	adverb	*often, really, early, now*
CONJ	conjunction	*and, or, but, if, although*
DET	determiner, article	*a, the, some, most, which*
NOUN	noun	*year, home, time, Africa,*
NUM	numeral	*fourth, 1991, 14:24*
PRT	particle	*at, on, out, that, up, with*
PRON	pronoun	*he, their, her, its, my, I, us*
VERB	verb	*is, say, told, playing, would*
.	punctuation marks	*. , ; !*
X	other	*ersatz, dunno, gr8, univeristy*

Source: https://www.nltk.org/book/ch05.html

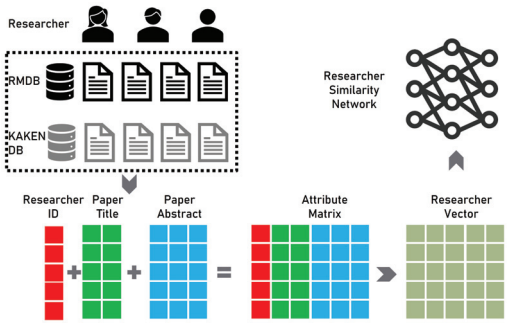
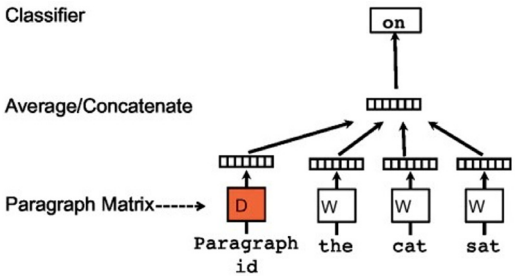


Fig. 2. Illustration of the MResearcher2vec model. The MResearcher2vec model first extracts researchers' attributes from RM and KAKENHI databases, then a matrix of researcher vectors is gained applying PV-DM algorithm. Finally, the model calculates the cosine similarity of researcher vectors.

4. Design of MResearcher2vec

Our study proposes a model called MResearcher2vec for computing researcher vectors through the distance between documents, which are comprised of a researcher's papers and research project summaries. We then calculate the similarity between researchers through the use of cosine similarity. **Fig. 2** provides an illustration of the MResearcher2vec model.

The first step of this model involves creating a researcher attribute matrix that includes a researcher ID tag, the title and abstract of each researcher's papers obtained from RMDB, and projects obtained from KAKENDB. A vector is then generated on a per-researcher basis using the Doc2vec algorithm, which is described in detail in the next section.



Source: Le and Mikolov [7]

Fig. 3. PV-DM algorithm.

4.1. Obtaining Researcher Vectors

Typically, text analysis systems consist of four distinct levels of applicability, including document level, paragraph level, sentence level, and sub-sentence level. At each level, the algorithm extracts the relevant categories of the document, paragraph, sentence, or sub-sentence, respectively. For this study, we combined the titles and abstracts of the articles to produce a paragraph-level vector.

Popular distributed representation techniques, such as Word2vec, FastText, and BERT, are used to embed feature vectors of words or documents and represent them as high-dimensional real vectors. In this study, we utilized Doc2vec, which is derived from Word2vec, to retrieve paragraph-level vectors. The model was trained in a Python 3.7 environment with NLTK and Gensim packages. The two-layer neural network model, Doc2vec, can compute vectors of sentences using two distinct algorithms [7], namely, the distributed memory model of paragraph vector (PV-DM, **Fig. 3**) and the distributed bag of words version of paragraph vectors (PV-DBOW). PV-DWOB has the advantage of faster learning speed, however, it does not consider the order of words in a sentence. Therefore, we have opted to process the data with the PV-DM algorithm in this study.

Doc2vec model uses a neural network approach to represent variable-length of text, such as sentences and paragraphs, as vectors. Vector representations have the advantage of capturing the semantics of the input text. In a vector representation, texts (either sentences or paragraphs) with similar meanings are located closer to each other than texts that are not necessarily related. Doc2vec features not only the ability to acquire a distributed representation of a document or group of documents, but also the ability to convert a sentence of any length into a vector of fixed length.

Le and MikoLov [7] offered two ways of calculating document vector from words vector–concatenation or mean. We adopt the later for computational efficiency. In the framework, each word is mapped to a unique vector represented by a column in a matrix W. The column is indexed by the position of the word in the vocabulary. The concatenation or sum of the vectors is then used as feature to predict the next word in a sentence.

More specifically, given a sequence of training words

Table 2. Parameters of Doc2vec model.

Parameter name	Value
epoch	10
vector_size	200
min_count	5
window_size	5
workers	4
dm	PV-DM

w_1,w_2, w_3, …, w_t, the goal of the word vector model is to maximize the average log probability,

$$\sum_{i=1}^{N}\sum_{t=k}^{T_i-k}\log p\left(w_{i,t}|d_i,w_{i,t-k},\ldots,w_{i,t+k}\right), \quad \ldots \quad (1)$$

where N is the total number of paragraphs, T_i is the number of words in paragraph d_i, $w_{i,t}$ is the t-th word in paragraph d_i. The prediction task is typically done by a multiclass classifier such as softmax. In this case, we have

$$p\left(w_{i,t}|w_{i,t-k},\ldots,w_{i,t+k}\right)=\frac{e^{w_{i,t}}}{\sum_i e^{y_i}}. \quad \ldots \quad (2)$$

Each of y_i is an un-normalized log-probability for each output word i, computed as,

$$y=b+Uh\left(d_i,W_{i,t-k},\ldots,W_{i,t+k};D,W\right), \quad \ldots \quad (3)$$

where U and b are the parameters of activation function softmax, h is constructed by a mean of word vectors extracted from document W.

The PV-DM algorithm can be divided into two primary stages. The first stage is the training stage, which involves acquiring word vectors W, softmax weights U and b, and paragraph vectors D (a matrix built from paragraph IDs that is employed to map paragraph tokens) from previously viewed paragraphs. The second stage is the inference stage, which entails generating paragraph vectors D for new paragraphs that have not been seen before by adding additional columns in D and performing gradient descent on D while keeping W, U, and b constant.

The parameters applied in the training model are described in **Table 2**.

The epoch determines number of iterations (epochs) over the corpus during training. The vector_size describes the dimensionality of the vector. The min_count determines the minimum number of times of a word occurring in the corpus. The window-size indicates the maximum distance between the current and predicted words within a sentence. The workers determines how many worker threads to train the model (i.e., faster training with multicore machines). The dm defines the training algorithm.

As anticipated, an inadequate vector_size results in a decline in the model's accuracy since it cannot represent the requisite information. In the current instance, when the vector_size exceeds 200, it appears to surpass the optimal threshold for the amount of information present, leading to a substantial number of sparse components. This, in turn, renders the supervised learning curve somewhat unstable. Therefore, a vector_size of 200 is selected in this study.

An observable enhancement was witnessed upon increasing the epoch. Given the substantial volume of data in this instance, it became imperative to augment the epoch beyond the default setting. Consequently, epoch = 10 was employed. Elevating the value of min_count is accompanied by an improvement in accuracy, albeit slight. The accuracy exhibits an upsurge when min_count is elevated from 1 to 5, while decreasing when min_count is set to 10. In the present experiment, a min_count of 5 is ascertained to produce the utmost accuracy. Contrary to expectations, altering the window_size has a negligible impact on the accuracy, possibly due to the nature of the PV-DM model. The limited influence of window_size could be attributed to the fact that the knowledge of the word types within the document is sufficient, and the contextual details are not of much consequence. In this study, a window_size of 5 is employed. The number of workers was configured to be 4, aligning with the number of cores present in the personal computer.

4.2. Calculation and Visualization of Similarity

To evaluate the similarity in vector space between researchers, eigenvectors are utilized to determine cosine similarity. If the cosine similarity value is 0, it indicates no similarity and the angle is 90°. Conversely, if the cosine similarity value is 1, the angle is 0°, indicating a complete overlap and maximum similarity.

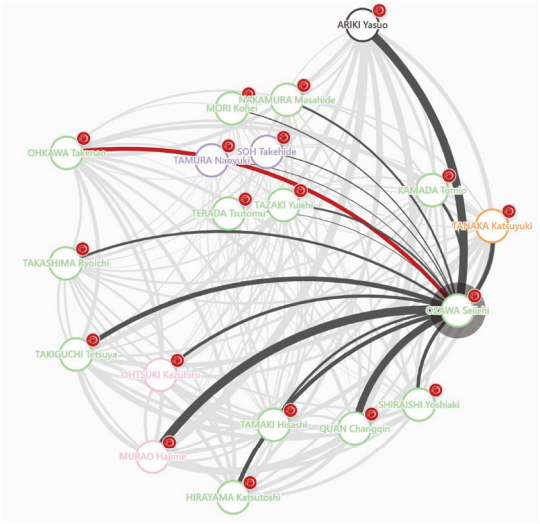
A similarity matrix is employed to generate a network graph that illustrates the connections between researchers. The network is composed of two sets of items: (1) nodes (representing the researchers) and (2) edges (indicating connections between two nodes basis on the degree of similarity). A thicker edge indicates a stronger degree of similarity between the connected researchers.

4.3. Implementation

The Institute of Promoting Academic Research Programs at Kobe University has commissioned a contractor to implement the machine learning model and results of the research as a service, which is called "Kobe University Research Hub."[4] Registered users can search for researcher name or a short paragraph of less than 500 words, and the network of researchers with similar vectors will be displayed. The search can be limited to researchers affiliated with Kobe University or to all researchers registered in KAKENDB, excluding those with less than 5 papers or those who choose not to be disclosed.

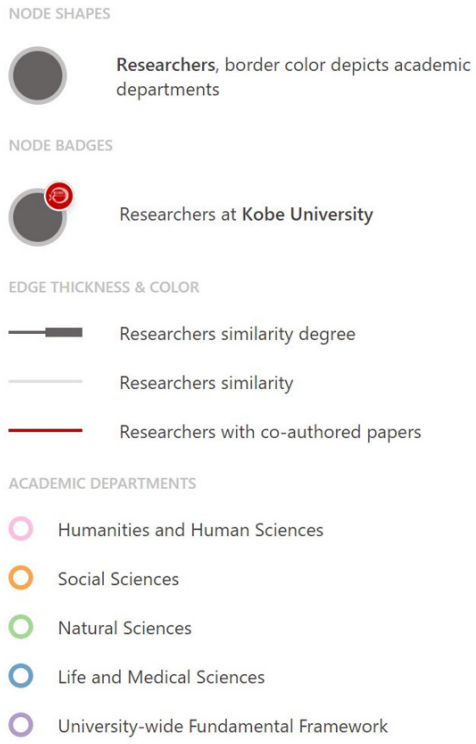
Figure 4 provides an example network graph of a researcher search, where the search parameters are defined by the name of the researcher "OZAWA Seiichi"[5] and the

4. https://www.researchhub.innov.kobe-u.ac.jp/ja [Accessed November 25, 2022]
5. https://researchmap.jp/ozawasei?lang=en [Accessed November 25, 2022]



Source: Kobe University Research Hub

Fig. 4. A demo of researcher search.



Source: Kobe University Research Hub

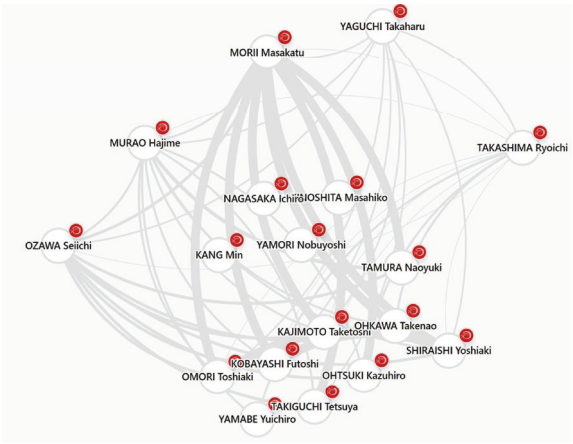
Fig. 5. Legend of colors.

display settings are configured to show top 20 researchers at Kobe University with the highest similarity scores.

The legend is shown in **Fig. 5**.

Figure 6 demonstrates the search capacity for short paragraphs, where the search query is a phrase containing keywords such as "information and communication," "information security," "cyber security," "intelligent informatics," "machine learning," "soft computing," and "neural networks."

The availability of correct answer dataset is absent in this study unless similarity between researchers is assessed subjectively. As a result, a qualitative evaluation method was employed, which entails the visual inspection and judgment of a human observer regarding the relationship between the "learned vector cosine similarity" and the "sentence similarity."



Source: Kobe University Research Hub

Fig. 6. A demo of short paragraph search.

Based on the analysis of a group of highly-published researchers across various fields, the search outcomes for both researcher name and short paragraph searches are deemed to be highly reliable.

Our study employs a matrix-based researcher-to-vector approach to compute similarities between researchers, which addresses the research gap outlined in Section 2. Furthermore, our research findings have the potential to facilitate the matching of research contents and researchers in industry-government-academia collaborations and joint research. This could foster exchanges between researchers, create opportunities for researchers and companies to establish connections, promote interdisciplinary research, and reduce lost opportunities for research institutions to attract expert personnel.

5. Summary

Our study introduces a system for visualizing the connections among researchers. To achieve this, we propose the MResearcher2vec method, which employs vectors derived from the contents of researchers' papers and research projects to compute their similarities. Our experimental results demonstrate the effectiveness of the proposed approach. For future studies, we aim to expand the application of the method to even larger datasets and compare its performance to other machine learning methods.

Acknowledgments

The authors thank Ms. Eri Anno, Mr. Makoto Yoshida, and Dr. Takashi Kita at Kobe University for their useful inputs and valuable comments.

References:
[1] H. Nishizawa, M. Katsurai, I. Omukai, and H. Takeda, "A Note on Similar Researcher Retrieval Considering Temporal Changes of Research Content and Affiliations," Proc. of the 32nd Annual Conf. of the Japanese Society for Artificial Intelligence (JSAI), Article No.4Pin1, 2018 (in Japanese).
[2] A. Okuma and Y. Kiyoki, "Seiki: Proposal of a Matching System for Companies and Researchers by Analyzing Papers Using Topic Models," 11th Forum on Data Engineering and Information Management (DEIM2019), Poster Session, 2019 (in Japanese).
[3] W. Wang, J. Liu, T. Tang, S. Tuarob, F. Xia, Z. Gong, and I. King, "Attributed Collaboration Network Embedding for Academic Relationship Mining," ACM Trans. Web, Vol.15, No.1, 2020. https://doi.org/10.1145/3409736
[4] W. Wang, F. Xia, J. Wu, Z. Gong, H. Tong, and B. Davison, "Scholar2vec: Vector representation of scholars for lifetime collaborator prediction," ACM Trans. on Knowledge Discovery from Data (TKDD), Vol.15, No.3, Article No.40, 2021. https://doi.org/10.1145/3442199
[5] D. Mochihashi, "Researcher2Vec: Visualization and recommendation of natural language processing researchers using neural linear models," Proc. of the Annual Conf. of the Association for Natural Language Processing, 2021. https://www.anlp.jp/proceedings/annual_meeting/2021/pdf_dir/B2-2.pdf [Accessed November 25, 2022]
[6] M. Zhang, K. Xu, K. Kawarabayashi, S. Jegelka, and J. Boyd-Graber, "Are Girls Neko or Shojo? Cross-Lingual Alignment of Non-Isomorphic Embeddings with Iterative Normalization," Proc. of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 3180-3189, 2019. https://doi.org/10.18653/v1/P19-1307
[7] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," Proc. of the 31st Int. Conf. on Machine Learning, Vol.32, No.2, pp. 1188-1196, 2014.

Name:
Enna Hirata

ORCID:
0000-0002-3127-3170

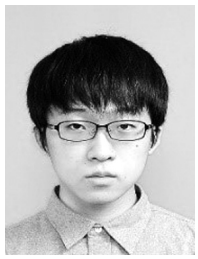
Affiliation:
Associate Professor, Center for Mathematical and Data Sciences, Graduate School of Maritime Science, Kobe University

Address:
1-1 Rokkodai-cho, Nada-ku, Kobe, Hyogo 657-8501, Japan

Brief Biographical History:
1998- A.P. Moller-Maersk A.S.
2019- Kobe University

Main Works:
• "Blockchain technology in supply chain management: insights from machine learning algorithms," Marit. Bus. Rev., Vol.6, No.2, pp. 114-128, 2020 (Emerald Literati Awards 2022 Outstanding Paper).

Membership in Academic Societies:
• The Japanese Society for Artificial Intelligence (JSAI)
• International Association of Maritime Economists (IAME)
• Maritime Policy & Management (MPM), Taylor & Francis. Associate Editor



Name:
Takahiro Yamashita

Affiliation:
Faculty of Engineering, Kobe University

Address:
1-1 Rokkodai-cho, Nada-ku, Kobe, Hyogo 657-8501, Japan

Brief Biographical History:
2020- Kobe University

Main Works:
• Conducting natural language process related studies.

Name:
Seiichi Ozawa

ORCID:
0000-0002-0965-0064

Affiliation:
Professor, Center for Mathematical and Data Sciences, Center for Advanced Medical Engineering Research & Development, Graduate School of Engineering, Kobe University

Address:
1-1 Rokkodai-cho, Nada-ku, Kobe, Hyogo 657-8501, Japan

Brief Biographical History:
2000- Associate Professor, Kobe University
2011- Professor, Kobe University

Main Works:
• "eFL-Boost: Efficient Federated Learning for Gradient Boosting Decision Trees," IEEE Access, Vol.10, pp. 43954-43963, 2022.

Membership in Academic Societies:
• International Neural Network Society (INNS), Vice-President for Membership
• Institute of Electrical and Electronics Engineers (IEEE) Trans on Cybernetics, Associate Editor
• Asia Pacific Neural Network Society (APNNS), Immediate-Past President