IIAI Open Conference Publication Series
IIAI Letters on Institutional Research
Vol. 001, LIR019
DOI: https://doi.org/10.52731/lir.v001.019

Feature Selection by Thematic and Temporal Distinction in Research Grant Applications

Michiko Yasukawa [*] , Koichi Yamazaki [†]

Abstract

We propose an effective method for selecting feature documents from a research grant database. The goal is to build a useful corpus for analytical tasks. While grant applications adopted in the past contain abundant information for institutional research, older applications are not assigned newer category labels for research areas. It is often difficult to apply unlabeled data to established techniques for data science and text analysis. To deal with this issue, our method automatically categorizes unlabeled grant applications into existing research categories. Using a document-by-document search technique, our method selects the best feature documents that are effective for improving the classification accuracy. To confirm the effectiveness of our proposed method, we conducted experiments using actual grant applications. The useful findings obtained in this study are as follows. (i) Using labeled grant applications, unlabeled grant applications are assigned labels to build a well-assorted corpus that includes the same number of grant applications from each research category of each year. (ii) By selecting a certain number of best feature documents from each research category of each year, the classification accuracy can be improved compared to that obtained using the initial dataset of labeled documents.

Keywords: open data, institutional research, faculty development, text analysis

1 Introduction

Research funding is important for university management. In this study, we investigate feature selection from research grant applications in Japan. Specifically, our database of interest includes grant application documents in the KAKEN database that is provided as open data from National Institute of Informatics (NII) [1].

To explain the background of our study, we first discuss the relationship between funding and university management. Hayashi [2] reviewed the reform policies of universities in Japan over the past few decades and highlighted that the competitive funding system has hindered the stable management of universities. Kikuchi [3] used the KAKEN database to

[*] Faculty of Informatics, Gunma University, Gunma, JAPAN
[†] School of Science and Engineering, Tokyo Denki University, Tokyo, JAPAN

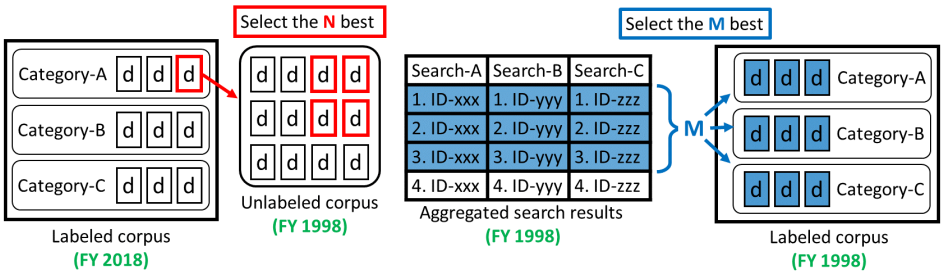


Figure 1: Feature selection for each category and year

quantitatively measure the impact of the partial privatization of Japan's national universities on research performance by research area. Ito and Watanabe [4] used the KAKEN database to analyze the role of research management specialists in universities. Nishizawa et al. surveyed the number of research funds obtained by universities in their pioneering study analyzing the KAKEN database [5]. Mizunuma and Tsuji [6] analyzed the role assignment and research outcomes of researchers who have been awarded research grants. Fujita et al. [7] proposed a method to quantitatively analyze interdisciplinary research projects according to disciplines and their research organizations.

Next, we discuss the relationship between the KAKEN database and its educational usage. The KAKEN database is provided on the NII's web server as open data and can be used to promote open science. To study the scientific concepts used in higher education, NII ' s website provides technical tips for citizens to search on the KAKEN database.[1] However, the search procedures of this database are detailed and employ complex conditions that are difficult to use. Regarding the usability of the information system, the study conducted by Ono et al. [8] suggested that a wiki service was preferred by users to enhance their understanding of science and technology. The study conducted by Ozeki et al. [9] noted that Japan's faculty members did not show remarkable concerns in acquiring funding, although faculty development was acknowledged to be essential in higher education. The study conducted by Ying et al. [10] developed a system to investigate the research trends based on the KAKEN database and article information in social sciences disciplines[2] for the purpose of active-learning and self-learning to enhance the literacy of the researchers.

Other studies analyzing the KAKEN database and Web of Science[3], which is bibliography information, include studies by Igami et al. [11] and Kurakawa et al. [12] The study conducted by Shimada et al. [13] analyzed mission-driven research grants and curiosity-driven grants using the KAKEN and CREST databases as research funding information and the J-Global database as bibliographic information. The study conducted by Kawashima et al. [14] used the KAKEN database and Scopus[4] as bibliographic information to verify the accuracy of author IDs in the analysis of researcher information.

Based on the above discussion, we propose a method to build a useful corpus, in which the best feature documents are selected with the same number in each research area to avoid unbalanced feature allocation. Such a corpus is called a *well-assorted corpus*, hereinafter. While grant applications of the past can be used as abundant knowledge resources, they are

[1] `https://support.nii.ac.jp/en/kaken/howtouse`
[2] `https://clarivate.com/webofsciencegroup/solutions/webofscience-ssci/`
[3] Web of Science, `https://www.webofknowledge.com/`
[4] Scopus, `https://www.scopus.com/`

Copyright © by IIAI. Unauthorized reproduction of this article is prohibited.

not assigned up-to-date category labels. To improve accessibility to the database, both new and old grant documents must be assigned recent category labels. To address this issue, our method automatically assigns labels to unlabeled grant applications using manually labeled grant applications to create a well-assorted corpus for analytical tasks. In a simple word, our method is an innovative combination of conventional text mining methods.

Previous studies reported that text mining techniques can be applied effectively to institutional research. For example, Sugihara et al. [15] employed SVM-based feature analysis to investigate questionnaires collected from athlete students and other students. Ogashiwa et al. [16] employed an SVM-based method to obtain feature words in a mid-term plan of higher education. In these prior studies [15] [16], SVM was used as a numerical model to obtain word vectors for documents with positive/negative labels. Different from their approaches, our method obtains document vectors rather than word vectors. It should be emphasized that our approach has a salient characteristic whereby a thorough document-by-document search is performed for each piece of data separated by category (thematic distinction) and fiscal year (temporal distinction). Notably, the numerical model for document vectors in our approach is not limited to a particular weighting model; thus, basic methods to obtain document rankings can be used. In addition, any document classifiers, including SVM classifiers and classification measures, e.g., F1-score and classification accuracy, can be selected in the proposed method according to the practical requirements of the constructed corpus.

The remainder of this paper is organized as follows. The proposed method is described in Section 2, the experimental data in Section 3, and the experimental results in Section 4, followed by a discussion in Section 5 and a summary and future work in Section 6.

2 Method

The motivation for this study is to use grant-application documents to create a useful corpus. Hereinafter, we refer to a piece of text in a grant-application document as a *document*. Our method consists of two phases. The first phase (Phase-1) is the process of "selecting the N best documents via document search," as illustrated on the left side of Figure 1. The second phase (Phase-2) is the process of "selecting the M best documents from the aggregated search results," as illustrated on the right side of Figure 1.

In Phase-1, each unlabeled document is used to search for similar documents from the categorized documents. All of the categorized documents are used for this document-by-document search. For each search, the N best documents are selected from the search results. In Figure 1, a concrete example (shown in red) illustrates the case of selecting the four best documents from the search result using a document in Category-A. The selection of the N best documents is performed for each search across categories to create a group of search results by category. InFigure 1, each of the three categories contains three documents. In the example, there are a total of nine categorized documents. Therefore, nine search processes are performed. Because the four best results are selected in each search, a total of 36 results are obtained. However, we have only 12 unlabeled documents. This means that duplicated documents must be included among the 36 search results. The duplicated documents are aggregated into groups of search results. The number of duplicates is counted for each searched document for each search group and sorted in the descending order of the number of duplicates to create a ranking list. This ranking list is referred to as *aggregated search results* (ASR).

Copyright © by IIAI. Unauthorized reproduction of this article is prohibited.

In Phase-2, we select the M best documents from each of the search groups in the ASR to obtain a group of categorized documents that is an intended labeled corpus. The concrete example (shown in blue) in Figure 1 illustrates the case of selecting the three best documents from each search group.

To clarify the two phases in the proposed method, let us consider a realistic example. If Category-A, Category-B, and Category-C are the categories of pedagogy, medical science, and informatics, respectively, then Search-A, Search-B, and Search-C are expected to be the search result rankings for pedagogy, medical science, and informatics, respectively. Selecting the top ranking documents, one would obtain the best feature documents for Category-A, Category-B, and Category-C. The example in Figure 1 shows a small-sized group of documents for simplicity. As the proposed method assumes Zipf's law [17] in the natural language, a sufficiently large group of documents is required to collect the best feature documents at the top ranking in the ASR. In addition, to assign relevant labels to unlabeled documents (i.e., old documents), the labels for search-key documents (i.e., new documents) must be associated with the relevant research areas. When the input (a search key) is given irrelevant labels, the output (its search results) would be automatically associated with the same labels as those of the input, resulting in irrelevant labeling. To validate the initial dataset (Category-A, Category-B, Category-C on the left side of Figure 1) and the final dataset (Category-A, Category-B, Category-C on the left side of Figure 1), we use *classification accuracy* (Acc) to evaluate the classification performance. Acc is defined as follows.

$$\mathrm{Acc} = \frac{\mathrm{Tp} + \mathrm{Tn}}{\mathrm{Tp} + \mathrm{Tn} + \mathrm{Fp} + \mathrm{Fn}}$$

In the equation, Tp, Tn, Fp, and Fn indicate true positive, true negative, false positive, and false negative, respectively, in the classification. Most importantly, the proposed method establishes data boundaries not only by category distinction but also by year distinction for unlabeled documents. The reason for this is that feature documents in a research area in one particular year may involve different trends from feature documents in other years. When the unlabeled document set is divided into segments of temporal attributes (e.g., a fiscal year), Phase-1 and Phase-2 are executed for each category of each year. Subsequently, the feature documents obtained from all segments are contained in a single large corpus.

Notably, the proposed method selects category feature documents (or feature documents) rather than document feature words (or feature words). Documents labeled with the same research area can have common features; however, each document should contain unique feature words. Thus, the document-by-document search is performed using each of the documents in a category as a single query rather than combining documents in a category as a single query.

In the proposed method, the similarity measure in the document search and document classifier for multiclass classification can be anything reasonable for the target documents. Realistically, however, the number of documents should be large in actual databases; thus, it is necessary to consider not only the effectiveness but also the efficiency of Phase-1 and Phase-2. For this reason, the experiments in this paper used an efficient variant of TFIDF for similarity measure and SVM classifier for the document classifier. The details of the experimental conditions are described in the following sections.

Copyright © by IIAI. Unauthorized reproduction of this article is prohibited.

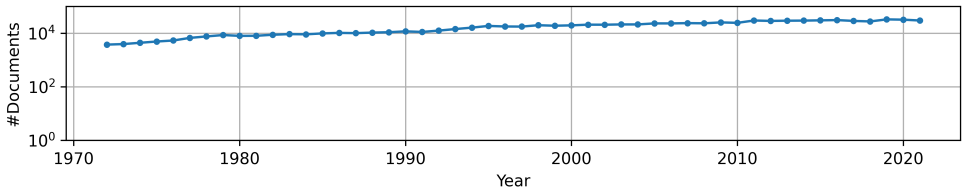
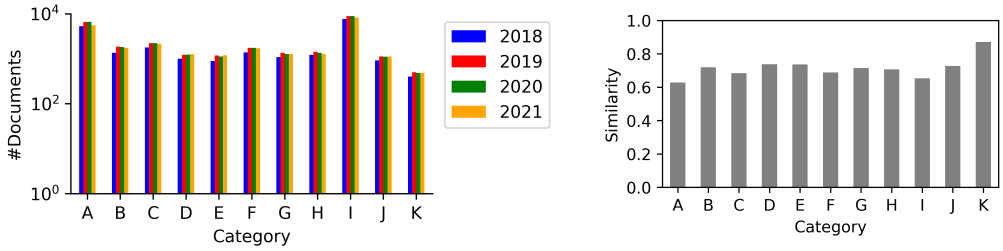


Figure 2: Number of grant applications per year



(a) #Documents per year

(b) Average similarity

Figure 3: Numbers of categorized documents and document similarity

3 Data

In the experiments, we used documents downloaded from the KAKEN database. This database is available online and modified from time to time; thus, it is possible that a version of database at a time in the past may differ from another version in the future. We used documents downloaded in the period of February 10–11, 2022. The downloaded documents contained information for 974,826 unique research projects. To understand yearly trends, the documents were aggregated by the year of research initiation, and the number of applications were counted. Most documents were between FY 1972 and FY 2021. The database included a limited number of documents in FY 1971 or earlier. Documents for FY 2022 were not added to the database. The number of documents per year for the 50-years period (FY 1972–2021) is presented in Figure 2. The X-axis indicates the year, and the Y-axis indicates the number of applications. The number of annual applications varied only slightly between the two years in proximity. However, the number of applications per year increased approximately 7.9 times over the past 50 years, i.e., 3,766 applications in FY 1972 to 29,908 applications in FY 2021. Research areas for the research projects to be reviewed were selected manually by the corresponding applicant; thus, the labels must have been assigned accurately to appropriate category labels by human experts. In addition, the definitions of research areas changed drastically in FY 2018; thus, the documents for FY 2017 or older were not assigned the most recent category labels. For the experiments, we used documents from FY 2018 and FY 2021 as categorized documents, and we used documents from FY 2017 and prior fiscal years as unlabeled documents. While the category labels are inconsistent between the old and new documents, each document unexceptionally includes a research project identification number (including numerals, English alphabet characters, and symbols) and a research project title that indicates the research theme of the project.

Copyright © by IIAI. Unauthorized reproduction of this article is prohibited.

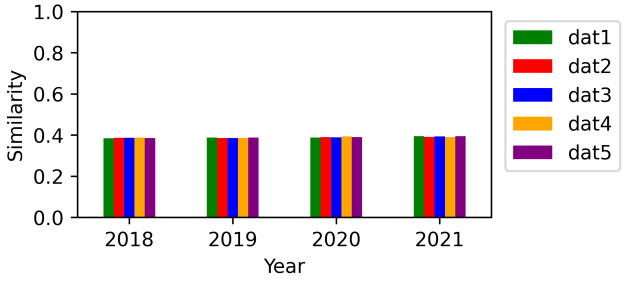


Figure 4: Document similarity for random samples

For example, an actual document included the research title, "A Study on Systematization of Culinary Vocabulary by Approximate String Matching and Related Terms Clustering" in English and the identification number, KAKENHI-PROJECT-26330363[5]. Both English and Japanese are used in the documents in the KAKEN database. The project titles in Japanese were used as text data to be analyzed in the experiments. For text preprocessing, we applied standard morphological analysis by MeCab[6] to the Japanese text data. After the morphological process, the Japanese text was separated with spaces to resemble the English text.

In the latest research categories, 11 categories have been defined and assigned identification labels using English alphabet characters from A to K. These 11 categories were used as the categories in our experiments. The number of documents in the 11 categories, aggregated by each year between FY 2018 and FY 2021, is shown on the left side of Figure 3. The X-axis indicates the category label, and the Y-axis indicates the document number. As shown, the number of documents varied significantly according to the category label information; however, the number of documents did not differ from year to year in the same classification. To verify the inter-categorical similarity among documents, the cosine similarity was obtained in a pairwise manner. Here, the documents in each category were combined to create one pseudo-document, and subsequently, the average cosine similarity between two pseudo-documents was calculated. The average cosine similarity for each category is shown on the right side of Figure 3. The X-axis indicates the classification, and the Y-axis indicates the average similarity. By comparing the left and right sides of Figure 3, we see that categories with more documents have lower similarity than those with fewer documents. As the number of documents in a given category increases, the number of unique features that contribute to this difference from other categories also increases.

However, when a single category has more unique features that differ from other categories, the Acc is essentially higher than that of the other categories. Thus, it is necessary to avoid bias in the number of documents when evaluating the categories. For this reason, random sampling from each of the thematic/temporal segments was performed with a fixed number of documents.

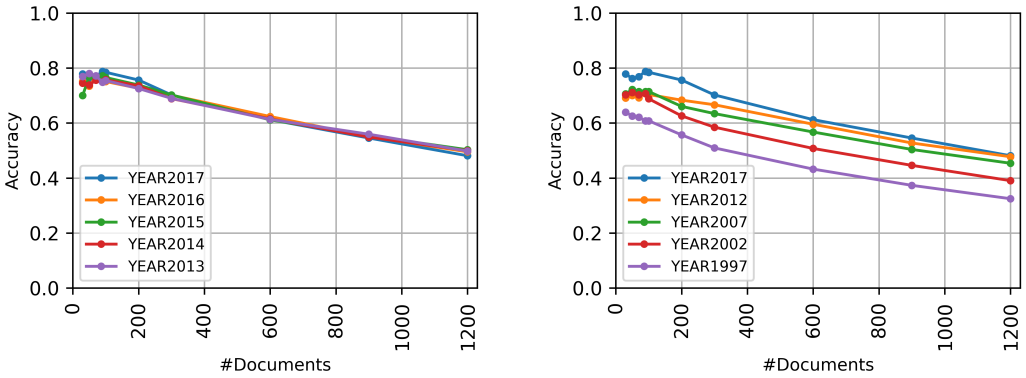
Specifically, Category-K had 402 documents in FY 2018. Considering the minimum number of documents in all categories, 300 was selected as a value sufficiently smaller than the minimum for Category-K (i.e., 402) and sufficiently large to obtain feature documents for random sampling. We performed five random samplings of 300 documents to obtain the

[5] `https://kaken.nii.ac.jp/en/grant/KAKENHI-PROJECT-26330363/`
[6] `https://taku910.github.io/mecab/`

Copyright © by IIAI. Unauthorized reproduction of this article is prohibited.

Table 1: Classification accuracy for the baseline method

	Year 2018	Year 2019	Year 2020	Year 2021	Years 2018-2021
#Doc	300	300	300	300	1200
Accuracy	0.4991	0.4958	0.4894	0.4882	0.5820



(a) Short span

(b) Long span

Figure 5: Classification accuracy for documents extracted from a single year

average cosine similarity for each dataset. The obtained average cosine similarity is shown in Figure 4. The X-axis shows the year, and the Y-axis shows the average cosine similarity. The average cosine similarity obtained from random sampling was nearly the same, which confirms that these datasets are unbiased in terms of features. In the following experiments, we employed this random sampling document dataset as labeled documents.

Notably, the constructed initial dataset (i.e., the labeled document dataset) contained randomly shuffled documents within each categorization and fiscal year. The Acc of the multiclass classification for this dataset was approximately 0.5 (details are given in Section IV). On the other hand, the Acc of the 11-class classification for a completely random dataset, in which fiscal year and categorization are ignored, was approximately 0.1. This result is considered legitimate because the 11-class classification with the completely random dataset theoretically has a 1/11 (approximately 9.091%) probability of being correct. Under the assumption that the random sampling in each segment (by category and fiscal year) resulted in the inheritance of 11-class features that are effective for multiclass classification, we consider that the prepared dataset contained more effective features for classification than the completely random dataset.

4 Experiments

To confirm the effectiveness of the proposed method, we conducted evaluation experiments using the documents in the KAKEN database. For the efficiency and effectiveness of the document search and classification of the documents, a variant of TFIDF with pivoted document length normalization (PDLN normalization) and SVM were used. Specifically, we used a machine learning library called scikit-learn to implement the experimental programs

Copyright © by IIAI. Unauthorized reproduction of this article is prohibited.

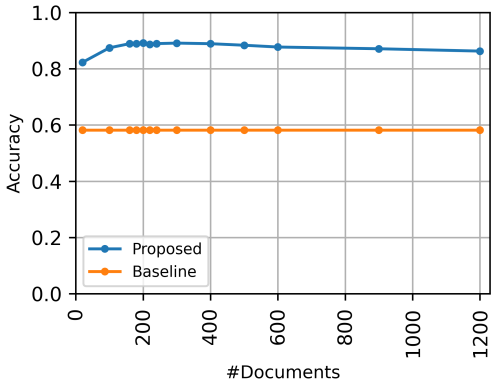


Figure 6: Classification accuracy for the baseline and proposed methods

in Python. The program for document-by-document search was implemented in the C programming language. To avoid dependency on a particular choice for the train/test pair sets, a procedure called cross-validation (CV) was adopted. Specifically, we used the default five-fold CV in scikit-learn's CV module. As explained in the previous section, our proposed method uses a group of labeled documents as initial data to categorize a group of unlabeled documents. Thus, the baseline method simply utilized the initial data and applied no subsequent procedures. The documents in the initial data were the grant applications that were categorized in the research categories (A to K) from FY 2018 to FY 2021 in the KAKEN database. As the baseline method also used the same data (i.e., from FY 2018 to 2021), we first confirmed the validity of the document classification for this data. To avoid a bias in the number of documents in the dataset, 300 documents were obtained by random sampling for each segment, as explained in the previous section. The obtained multiclass (A to K) classification accuracy for each year is shown on the left side of Table 1. Here, the multiclass classification of the combined four-year span represents the baseline method. The combined data included 1,200 documents in total. The classification accuracy obtained by the baseline method is shown on the rightmost column of Table 1. The higher accuracy in the four-year period than that in the single year period is attributed to the increased number of feature documents for effective classification.

The proposed method performs a search of the N best unlabeled documents by labeled documents before selecting the M best documents from the ASR. To validate the retrieved documents, multiclass classification was conducted for each year, and the Acc was obtained. The results are shown in Figure 5. The X-axis indicates the number N for the best documents retrieved from the search results, and the Y-axis indicates the Acc for the multiclass classification (classes A to K). Here, each line graph corresponds to each year. The left side of Figure 5 compares the Acc for each year in a short span (i.e., the five years from FY 2017), and the right side of Figure 5 compares the Acc for a long span (i.e., the 20-year period from FY 2017) at five-year intervals. For both the short and long spans, the Acc was higher for the top-ranked document set in the search results. However, the Acc was worse when lower-ranked document sets were included in the document classification process. In addition, a little difference in the classification accuracy was observed within the five-year span. Meanwhile, the Acc was lower between FY 1997 and FY 2012. In the experiments, the documents between FY 2018 and FY 2021 were used as query documents

Copyright © by IIAI. Unauthorized reproduction of this article is prohibited.

Table 2: Baseline vs. proposed

Method	Year	#Doc per Segment	#Doc per Category	#Doc per Corpus	Acc.
Baseline	2018-2021	300	1200	13200	0.5820
Proposed	1998-2017	10	200	2200	0.8918

in the document-by-document search process. These results suggest that as the grant applications became older, the retrieved document sets had fewer common features with the newer documents, and it was difficult to search similar documents.

Next, we evaluated the procedure for selecting the M best documents of the proposed method. Specifically, documents were obtained from the top search results among 11 categories (A to K) for the 20-year span from FY 1998 to FY 2017 (including the beginning and end years) and added to the document set to perform document classification and obtain the Acc. The results are shown in Figure 6. The X-axis shows the number M of the best documents to be selected, and the Y-axis shows the Acc. Here, the orange line graph shows the baseline method, and the blue line graph shows the proposed method. As Figure 6 visualizes, the proposed method outperformed the baseline method. Because the baseline method employs the initial set of documents without modification, the obtained Acc is constant. Meanwhile, in the proposed method, the Acc varies according to the M value. When M is small, the set of documents to be classified is too small and the Acc is low. With a very small number of documents, multiclass classification is considered unsuccessful as necessary features are not contained in the documents. Therefore, the highest classification accuracy was obtained when M (the number of documents in the document set) was assigned appropriately (i.e., M = 10, and the number of documents per category was 200.). When the document set was even larger, the Acc decreased due to the diversified features. Thus, to construct an optimally assorted corpus, it is recommended that the M value be chosen so that Acc is the maximum.

Finally, we confirmed the relationship between the Acc and the number of documents per segment, category, and corpus. Table 2 compares the number of documents and the Acc of the baseline and proposed methods. The baseline method combined 300 randomly selected documents per year for each category from the four-year span (FY 2018 to FY 2021), allocated 1,200 documents to each category, and collected the documents from 12 classes to build a corpus of 13,200 documents. Here, the labels for each document were assigned manually by the grant applicant, who was presumably an expert in the corresponding research area, and the corpus contained a large number of documents (i.e., greater than 10,000). The Acc obtained for this baseline corpus was 0.5820. Conversely, the proposed method combined the 10-best retrieved documents per year from each category from a 20-year span (FY 1998 to FY 2017) to include 200 documents per category and collected documents from 12 classes to construct a well-assorted corpus of 2,200 documents. The documents between FY 1998 and FY 2017 were accepted before the newly-defined research categories; thus, there were no corresponding labels given by human experts.

In our method, document labels were assigned automatically using a computer. In addition, fewer documents than the baseline method were included in the constructed corpus by the proposed method. In spite of these conditions, the Acc for the constructed corpus marked 0.8918, which was higher than that for the baseline corpus. The constructed cor-

Copyright © by IIAI. Unauthorized reproduction of this article is prohibited.

pus was concise and yet effective. Depending on the scope of application of the obtained corpus, a larger corpus may be required. The proposed method can construct a corpus of any size by varying the M value. To make the corpus size the same as that of the baseline method, the M value should be set to 60. In the experiments, when M = 60, the obtained number of documents per category was 1,200 and the Acc was 0.8628, which was still higher than that of the baseline method, i.e., 0.5820.

Based on the experimental results, we consider that the proposed method is more effective than the baseline method in terms of classification accuracy. In addition, the proposed method is more useful than the baseline method because it can vary the corpus size according to the requirements of the target application.

5 Discussion

The proposed method sorts unlabeled documents grouped by category in such a manner that highly similar documents to the labeled documents are ranked higher in each category. In data analytics, it is a common practice to sort large amounts of data by the same group to extract the head of each group, and the proposed method provides a useful function to sort documents by similarity in each category. One notable aspect of the proposed method is that each of the labeled documents is used as a search query to obtain the N best documents from each temporal segment. As the overall size of the dataset increases, the best documents with greater support from others will be more emphasized according to Zipf's law. The proposed method creates a ranked list of searched documents by year to select the best feature documents on a yearly basis. The reason for this is that research themes in research grants may differ from one research area to another as well as from one year to another. While proximate years may contain similar feature documents, distinguishing between years prevents small differences from being outweighed by large differences of the overall dataset. This distinction allows feature selection with high Acc without missing important feature documents.

The corpus constructed by the proposed method could be used for a variety of analytical tasks. For example, it could be used to investigate the differences in the institutional affiliations and budgets of the researchers by research area and by year, based on the feature documents in the category. In addition, the well-assorted corpus, which is arranged in equal numbers for each research area, enables feature analyses using basic methods in textbooks. For example, the corpus constructed by the proposed method could be used for analytical tasks, such as feature word analysis. Examples of feature words extracted using "mutual information"(MI)[7] from categories A to K are listed in Table 3. Categories A to K are well represented by the extracted feature words. However, some specific collocations and compound words in specialized areas were excessively separated or redundant. For example, the phrase "large-scaled data" was separated into "large-scale" and "data," as shown in the lower part of Table 3. For feature word extraction, we used the Japanese morphological analyzer MeCab accompanied with Neologism dictionary for MeCab[8]. The dictionary contained newly emergent words including words in the Japanese Wikipedia titles. While some specialized words in scientific themes were preprocessed accurately, others caused unstable results in the morphological analysis. This issue is outside the scope of the current study and will be discussed in our future research. Another limitation of the current study is the

[7] https://nlp.stanford.edu/IR-book/html/htmledition/mutual-information-1.html
[8] https://github.com/neologd/mecab-ipadic-neologd

Copyright © by IIAI. Unauthorized reproduction of this article is prohibited.

Table 3: Examples of feature words

Category	Feature words
A	japanese, history, culture, comparison
B	nonlinear, geometry, x-ray, quantum
C	structure, steel, flow, measurement
D	oxide, nanoparticles, heat, foundry
E	reaction, catalyst, complex, metal
F	plant, symbiosis, fungi, bacteria
G	nerve, species, dependent, plasticity
H	medicine, treatment, immunity, t-cell
I	nursing, cancer, mesenchymal stem cell, stem cell
J	machine learning, large-scale, data, deep learning
K	climate change, ocean, forest, isotope

time efficiency of the proposed method. As the proposed method intensively searches unlabeled documents, it requires more computation time than the baseline method. Devising a more computationally efficient method will be investigated in future.

6 Conclusion

With the goal of constructing a useful corpus for analytical tasks, we investigated highly effective feature selection methods for grant-application documents. Owing to the reform of research categories every few years, older documents are not labeled with the latest research categories. Our method utilizes labeled documents to assign labels to unlabeled documents. First, a document-by-document search is conducted according to thematic and temporal segments to select a group of candidates for the best feature documents. Subsequently, the best feature documents are obtained from the aggregated search results. Finally, the selected best feature documents on a yearly basis are concatenated to construct a well-assorted corpus. We found that the corpus constructed using the proposed method exhibited a smaller file size than the initial dataset and demonstrated higher classification accuracy than the manually categorized dataset. The constructed corpus can be applied to various text analytic tasks, e.g., feature word extraction. We confirmed that the feature words extracted from the obtained corpus were plausible for each research category. For example, *machine learning*, *large-scaled data*, and *deep learning* were obtained for the informatics discipline (Category-J). We believe that the proposed method can function as a fundamental text processing method for grant applications. In the future, we plan to ameliorate the efficacy of the morphological analysis and the computation time efficiency of the proposed method.

Acknowledgments

This study was supported by the ISM Cooperative Research Program (2022-ISMCRP-0006) and JSPS KAKENHI Grant Number JP18K11986.

Copyright © by IIAI. Unauthorized reproduction of this article is prohibited.

References

[1] "KAKEN: Grants-in-Aid for Scientific Research Database (The National Institute of Informatics)," https://kaken.nii.ac.jp/.

[2] T. HAYASHI, "University Reform Policy and the New Image of University," *The Journal of Science Policy and Research Management*, vol. 36, no. 3, pp. 257–270, 2021.

[3] Y. Kikuchi, "Impact of university reform on research performance aggregated and disaggregated across research fields: a case study of the partial privatization of Japanese national universities," *The Japanese Economic Review*, pp. 1–27, 2021.

[4] S. Ito and T. Watanabe, "Multilevel Analysis of Research Management Professionals and External Funding at Universities: Empirical Evidence from Japan," *Science and Public Policy*, vol. 47, no. 6, pp. 747–757, 2020.

[5] M. Nishizawa, M. Negishi, M. Shibayama, Y. Sun, H. Nomura, M. Maeda, and Y. Mitsuda, "Evaluation of Japanese universities' research activity based on the number of awards of Grants-in-Aid for Scientific Research from 1998 to 2002 and in 2003."

[6] Y. Mizunuma and K. Tsuji, "An Investigation on the Researchers Who Received Japanese Grant-in-Aid for Scientific Research (KAKENHI) with a Focus on Their Roles and Research Achievements," in *2019 8th International Congress on Advanced Applied Informatics (IIAI-AAI)*. IEEE, 2019, pp. 103–108.

[7] M. Fujita, T. Okudo, T. Terano, and H. Nagane, "Analyzing Two Approaches in Interdisciplinary Research: Individual and Collaborative," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 25, no. 3, pp. 301–309, 2021.

[8] E. Ono and Y. Ikkatai, "Internet-based Services to Obtain Information on Science and Technology according to the Degree of Interest," in *2020 9th International Congress on Advanced Applied Informatics (IIAI-AAI)*. IEEE, 2020, pp. 328–331.

[9] S. Ozeki, T. Hayashi, M. Fukano, S. Yamazaki, A. L. Beach, and M. D. Sorcinelli, "Exploring the Future Trends of Faculty Development in Japanese Higher Education," in *2021 10th International Congress on Advanced Applied Informatics (IIAI-AAI)*. IEEE, 2021, pp. 291–294.

[10] C. Yin, Y. Tabata, and S. Hirokawa, "A " Milky Way Research Trend " system for Survey of Scientific Literature," in *International Conference on Web-Based Learning*. Springer, 2012, pp. 90–99.

[11] M. Igami and A. Saka, "Decreasing diversity in Japanese science, evidence from in-depth analyses of science maps," *Scientometrics*, vol. 106, no. 1, pp. 383–403, 2016.

[12] K. Kurakawa, Y. Sun, and S. Ando, "Application of a Novel Subject Classification Scheme for a Bibliographic Database Using a Data-Driven Correspondence," *Frontiers in big Data*, p. 48, 2020.

[13] Y. Shimada, N. Tsukada, and J. Suzuki, "Promoting diversity in science in Japan through mission-oriented research grants," *Scientometrics*, vol. 110, no. 3, pp. 1415–1435, 2017.

Copyright © by IIAI. Unauthorized reproduction of this article is prohibited.

[14] H. Kawashima and H. Tomizawa, "Accuracy evaluation of Scopus Author ID based on the largest funding database in Japan," *Scientometrics*, vol. 103, no. 3, pp. 1061–1071, 2015.

[15] T. Sugihara, S. Aihara, S. Hirokawa, and T. Nara, "An Analysis of Characteristics of Student-Athletes from Questionnaire by SVM," in *2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*. IEEE, 2017, pp. 163–166.

[16] K. Ogashiwa, E. Takata, T. Oishi, M. Mori, and S. Hirokawa, "Automatic Estimation and Feature Word Analysis of Universities Using University Medium-term Plans," in *2019 International Congress on Applied Information Technology (AIT)*. IEEE, 2019, pp. 1–6.

[17] G. K. Zipf, *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley, 1949.

Copyright © by IIAI. Unauthorized reproduction of this article is prohibited.