International Journal of Institutional Research and Management
International Institute of Applied Informatics
2023, Vol. 7, No. 1, IJIRM766

Detecting Transition of Research Themes using Time-oriented Attributes in Governmental Funding

Michiko Yasukawa [*] , Koichi Yamazaki [†]

Abstract

We investigate a method for detecting yearly difference between new and old scientific research themes in grant applications. While open data for such analysis is available, there has not yet been sufficient study to fill in the gap between theory and practice of quantitative analysis of actual data. In our approach, binary document classification and regression analysis are combined to examine a large corpus of grant applications. From a theoretical viewpoint, we analyzed artificial corpora that emulates heterogeneity in the target text data. Then, we experimented on the real data of research themes in governmental funding in Japan to confirm the effectiveness of our approach. Our contribution in this study is represented by the notable findings as follows. (1) As research themes in competitive grants somewhat changed each year, newer themes gradually became dissimilar to old themes. (2) While the differences in a shorter span is generally smaller and different research areas have different tendencies in a longer span, the time-oriented tendency in research themes for 20 years were detectable and the differences between the baseline and our methods were statistically significant.[1]

Keywords: text mining, regression analysis, database, grant-in-aid for scientific research

1 Introduction

Scientific studies in higher education have been strenuously promoted around the world. There is an old Japanese saying, "*Jū nen hito mukashi.*" meaning *10 years of time can bring a lot of changes*. Then, what changes in scientific themes can be expected for 20 years? In institutional research, decision making based on data science, rather than subjectivity of individuals, is important. Our motivation in this study is to explore how to detect transition in scientific research themes in governmental funding in Japan.

To clarify research background of the current study, we discuss some prior studies as follows. In the study conducted by Yamashita et al. [2], researchers proposed a method to predict the trendiness of research fields using funding data in JSPS Grants-in-Aid for Scientific Research [3]. While their study analyzed temporal differences for several years, our method in the current study intends to detect temporal differences during 20 years in

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

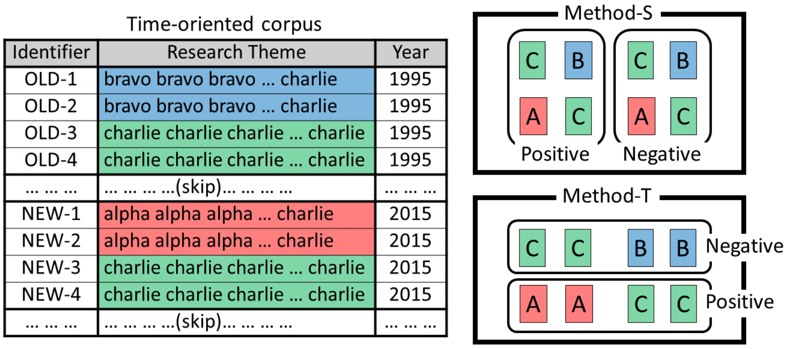
[1] An earlier version of this paper was presented at [1].

the KAKEN database [4]. A pioneering method for dealing with funding documents was introduced by Cohen et al. [5] and their method makes semantic reasoning in searching research topics. For example, search keywords, such as “mitral valve prolapse” issued by funding applicants need to link to relevant topics, such as “heart disease” in funding advertisements. While their study aimed at efficient document search by eliminating keyword mismatches between queries and documents, our approach takes an advantage of the differences between new and old keywords to realize time-oriented analysis of funding documents. Recent studies analyzing research funding include the analysis of research funding in two research areas in four European countries [6], the study of the relationship between valorization of science and research funding [7], the study for measuring the relationship between research quality and research budget [8].

The challenges in analyzing tendency of research areas in terms of bibliographical information have a long history of more than 50 years. The study conducted by Kuhn [9] introduced an analogy between scientific revolutions and political revolutions, explaining that scientific revolutions were generally difficult to observe. The study conducted by Small [10], on the other hand, reported that co-citation frequencies in the Science Citation Index (SCI) from 1973 and 1974 were measurable and observation of such a quantity could provide insight into the very rapidly changing frontiers of scientific research. In the study conducted by Griffiths et al. [11], they proposed a method for identifying hot topics in peer reviewed journals in biology, physics, and social sciences while Krenn et al. [12] used the content of 750,000 scientific papers published since 1919 to detect influential and prize-winning research topics in the discipline of Quantum Physics. In our study, we use the KAKEN database that includes over 900,000 adopted grant applications in all research areas in Japan. Specifically, our study uses grant documents from all research areas as well as grant documents from Medical Science, Social Science, Literature Science, and Information Science.

Our goal in this study is to detect what differences exist between two different corpora (i.e., new and old research themes). It is humanly infeasible to manually examine features in high-dimensional data, such as large corpora. For this reason, *visual feature analysis*, by which the frequency of occurrence of tokens is automatically counted and visualized, is used for assisting humans’ analytical tasks. Bird et al. [13] demonstrates a visualization of the frequency of male and female names with each letter of the English alphabet at the end while Bengfort et al. [14] introduces a method to visualize the frequency of words in the speeches of three different US presidents using a line graph, in which the x-axis indicates the publication date of each text data. By visual feature analysis, qualitative analysis of our interested text data, i.e., funding data, is possible. However, no quantitative analysis method for our target issue in the current study has been proposed so far. While our prior study [1] indicated the feasibility of the new/old binary classification in academic disciplines for bibliographical data, its applicability to funding data had not been examined. The main contribution of the current study is to address methodology for quantitative analysis that detects differences between old and new research themes in governmental funding in Japan. As our methodology exploits well-developed machine learning algorithms in addition to the simple linear regression analysis to examine a large database of grant applications, the obtained results are expected to provide innovative insights in the research community of institutional research.

In the following sections, we describe baseline and proposed methods (Section 2), experimental data (Section 3), and experimental results (Section 4), followed by discussion (Section 5) and conclusion (Section 6).



Time-oriented corpus

Identifier	Research Theme	Year
OLD-1	bravo bravo bravo ... charlie	1995
OLD-2	bravo bravo bravo ... charlie	1995
OLD-3	charlie charlie charlie ... charlie	1995
OLD-4	charlie charlie charlie ... charlie	1995
... ...	... ...(skip)...	... ...
NEW-1	alpha alpha alpha ... charlie	2015
NEW-2	alpha alpha alpha ... charlie	2015
NEW-3	charlie charlie charlie ... charlie	2015
NEW-4	charlie charlie charlie ... charlie	2015
... ...	... ...(skip)...	... ...

Figure 1: Schematic illustration of the baseline and proposed methods

2 Methods

Figure 1 shows a schematic diagram of our approach. In the example, the corpus is stored in a spreadsheet format, wherein each row contains a research theme, identifier and time-oriented attribute. The time-oriented attribute is represented in *fiscal year* (hereinafter, FY), in which the research project initiated. Each research theme is text data that consists of a sufficient number of words. Henceforth, we will refer to this piece of text (research theme) as a *document*. On the right side of Figure 1, Method-S indicates our baseline method, which ignores time-oriented attributes. Method-T is our proposed method, which exploits time-oriented attributes. By applying the convention of machine learning, we regard one document class is positive and the other is negative in both of the methods. Our approach for transition detection is that we perform many times of the binary classification for a given period of years to examine how the new/old classification succeeds thanks to the corresponding time-attributes. Then, a linear regression is used to analyze the temporal heterogeneity in the given period. The details are explained as follows.

For our discussion, let us focus on the colored documents in the figure. As can be seen in the figure, *alpha* and *bravo* are exclusively included in the new and old documents, respectively while at the same time *charlie* appears in both old and new documents. In the example, A, B and C indicates topics in documents. When time-oriented attributes are meaningful, Method-T is expected to be more successful than Method-S does. Conversely, there must be no significant differences between the two methods when time-oriented attributes do not contribute to the binary classification. Formally speaking, our approach in this study is based on the hypothesis as follows. The binary classification (new/old classification) should be easy when the given document classes are heterogeneous. If the classification is easier, the classification accuracy must be higher. Hence, we could deduce the heterogeneity of the two document classes by measuring the classification accuracy in the binary document classification. In information retrieval, some probabilistic models are designed by deducing unobservable values utilizing observable values [15]. We apply this common known technique to our approach.

To verify the aforementioned hypothesis on an empirical manner, we synthesize heterogeneous corpora, as follows. First, we define the heterogeneity parameter R by the equation, $R = K/L$. Here, K, L indicates the number of heterogeneous documents and the number of documents in each document class, respectively. Some examples of different heterogeneity values are shown in Figure 2.

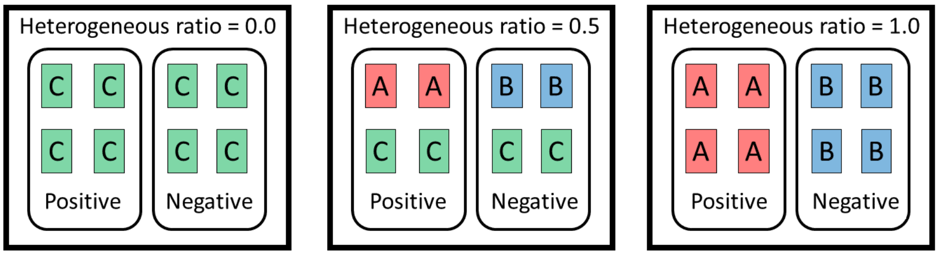


Figure 2: Example of heterogeneous corpora

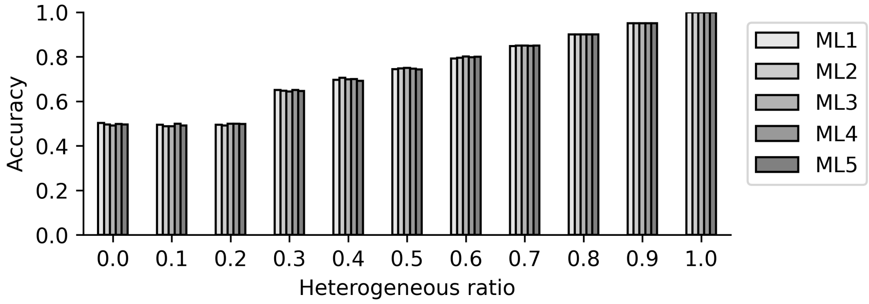


Figure 3: Heterogeneous ratio and accuracy classification score

Based on the aforementioned definition of L and R, artificial corpora can be synthesized as follows.

- Both of the two classes contain $L \times R$ heterogeneous documents.
- Both of the two classes contain $(L \times (1 - R))$ homogeneous documents.
- Each heterogeneous document contains p content words and q non-content words.
- Each homogeneous document contains (p+q) non-content words.

While the values for p and q may be chosen from arbitrary natural numbers, they should be realistic to emulate our target text data, which contains funding information. For example, a document that contains only a few content words (e.g., mitral, valve, prolapse) and millions of non-content words (e.g., a, an, the, of, and, so, on) should be unrealistic. Incidentally, a meaningful phrase may consist of only non-content words (e.g., “to be, or not to be”) in a realistic setting. Hence, non-content words should be retained in the artificial corpora. Once the corpora are synthesized, we can perform the binary document classification. There is, however, no single classifier that can be applied for any data. Instead, the machine learning library called scikit-learn [16] provides several choices of implementations for text classification [17]. In our approach, the following well-established implementations are applied:

ML1 K Neighbors Classifier [18]
ML2 Naive Bayes Classifier [19]
ML3 Ensemble Classifiers (commonly known as *random forest classifier*) [20]
ML4 SGD Classifier [21]
ML5 Linear SVC Classifier (commonly known as *support vector machine classifier*) [22]

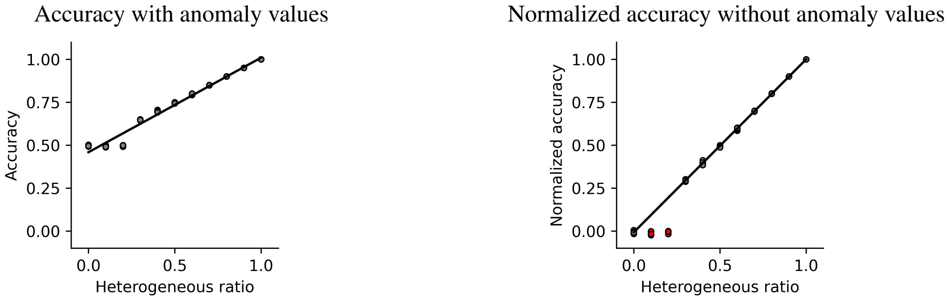


Figure 4: Comparison of accuracy and normalized accuracy

By using each of the aforementioned classifiers (ML1, ML2, ML3, ML4, and ML5), we can obtain the classification accuracy for an arbitrary corpus, as follows.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

Here, TP, TN, FP, FN indicate true positive, true negative, false positive, false negative in the classification, respectively. Figure 3 presents the classification accuracies with varied values for heterogeneous ratio R. To obtain this results, values for p and q were assigned, 25 and 5, respectively so that (p+q) became 30, based on the average values in our interested data. As can be seen in Figure 3, the classification accuracy is approximately 0.5 for the minimum heterogeneous ratio (i.e., $R = 0.0$) as the chances of a successful classification are one in two. When the heterogeneity was small (i.e., $R = \{0.1, 0.2\}$), the classification was poorly performed and anomaly values less than 0.5 were yielded as shown in Figure 3. By excluding the anomaly values, the observed values of the classification accuracies ranged 0.5 (minimum) to 1.0 (maximum). Hence, we can apply the min-max normalization to obtain the normalized accuracy, as follows.

$$y_i = \frac{\text{Accuracy}_i - \text{minimum}}{\text{maximum} - \text{minimum}} \quad (i = 1, ..., n)$$

In the equation, y and n indicate the normalized classification accuracy and the number of obtained accuracies, respectively. Finally, the relationship between the heterogeneity ratio x_i and the normalized accuracy y_i is analyzed by the linear regression, which is represented in the equation $y_i = ax_i + b$. As a result of the empirical analysis, we confirmed that the regression coefficient a and intercept b were approximately 1 and 0, respectively. Figure 4 visualizes the regression analysis for the classification accuracy (with anomaly values) and the normalized classification accuracy (without anomaly values). As can be seen, the normalized classification accuracy (observable from the data) corresponds to the heterogeneity of the corpus (unobservable from the data). Thus, the value for the normalized classification accuracy is used for the theoretical heterogeneity in the two classes. To observe the temporal differences for a long span (e.g., 20 years), our approach attempts the binary classification for many possible intervals in the span (e.g., 1, 2, ..., 20 years). In the same way as the accuracy, the min-max normalization is applied to the interval and called the normalized interval. Finally, transition detection in our approach is achieved by regression analysis using the normalized interval (explanatory variable) and the normalized accuracy (objective variable).

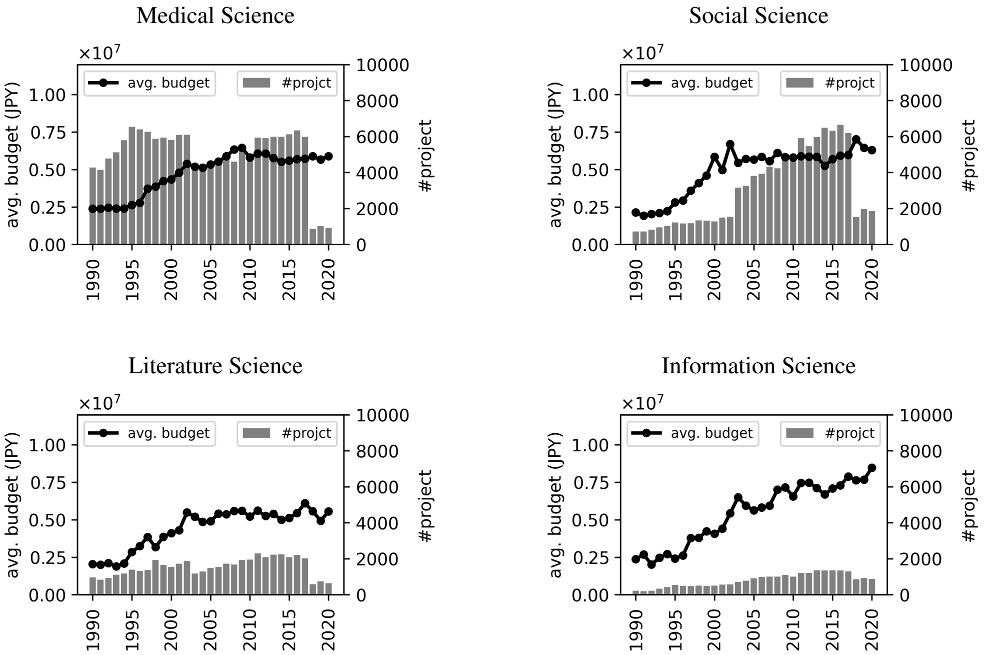


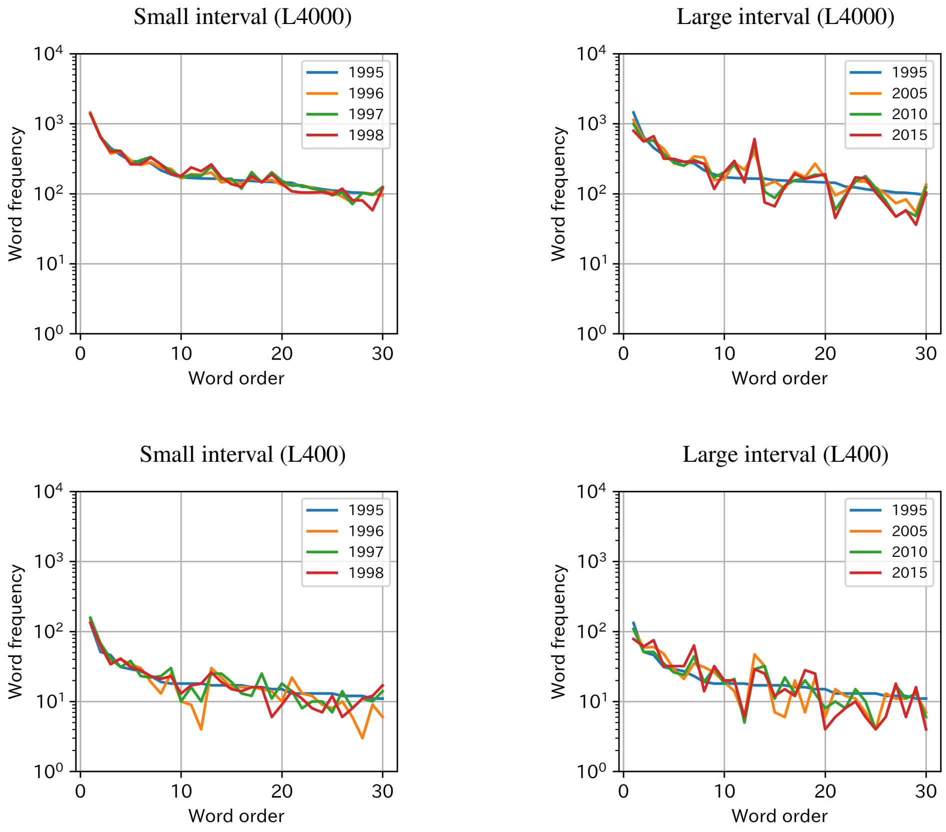
Figure 5: Average research budget and the number of projects (for the four research areas)

Table 1: Experimental corpora (for the period from FY 1995 and FY 2015)

Corpus name	Research area	Aggregated by year			Total	L
		Min.	Max.	Avg.		
Corpus-L4000	(unspecified)	17732	30329	23317	489652	4000
Corpus-L400	*do.*	*do.*	*do.*	*do.*	*do.*	400
Corpus-Med	Medical Science	4148	6535	5519	115900	400
Corpus-Soc	Social Science	1171	6491	3483	73147	400
Corpus-Lit	Literature Science	1174	2295	1740	36547	400
Corpus-Inf	Information Science	478	1365	866	18181	400

3 Data

In this study, we used the KAKEN database [4]. This database contained adopted grant applications for the Grant-in-Aid for Scientific Research (or, KAKENHI) [3], which was the largest governmental funding in Japan. The database included detailed information such as the research identification number, the research theme, the project principal investigator, research collaborators (if any), the fiscal year in which the research project initiated, the total budget amount allocated, etc. The database is updated on a daily basis. In the current study, the data downloaded from February 10th to 11th, 2022 was used. The downloaded data contained 974,826 grant applications from FY 1964 to FY 2021. Considering the KAKEN database was intended to provide funding information, we confirmed the number of research projects and the average budget amount. For the database as a whole, there was a monotonic increase in budget amount and the number of projects. However, older data was insufficient and the most recent data was still being updated and unstable.


Small interval (L4000)

Large interval (L4000)

Small interval (L400)

Large interval (L400)

Figure 6: Small interval vs. large interval (for unspecified research areas)

In addition, the numbers of projects differed depending on the research areas. To compare the holistic and partial features of the database, we experimented on the following four research areas: Medical Science, Social Science, Literature Science and Information Science, respectively. While the research areas were reformed every several years, these four research areas included adequate amount of data in the past. Figure 5 visualizes the budget and the project number for the four research areas[2]. To make the experimental results stable and reproducible, we focused on the data from FY 1995 to FY 2015, for approximately 20 years of time. For the experiment, six corpora shown in Table 1 were constructed. In the table, "Aggregated by year" indicates the minimum, maximum, and average number of grant applications per year. In the right most column, L indicates the number of documents in new/old document classes (see the explanation in Section 2).

As a preliminary experiment, we confirmed that the experimental corpora contained small heterogeneity in the short interval and large heterogeneity in the long interval by using visual feature analysis [13] [14]. Figure 6 and Figure 7 show the word order and word frequency of top 30 high frequency words from each of the corpora. In the figure, the blue line indicates the word frequency in research themes in FY 1995. The left column visualizes cases with small intervals from 1995, i.e., one, two, three years. The right column visualizes the cases with large intervals since FY 1995, i.e., 10, 15, and 20 years of time.

[2] Only a few research projects were explicitly linked to the four conventional research areas since FY 2018. The latest research area code (alpha/numeric characters) were defined more interdisciplinary.

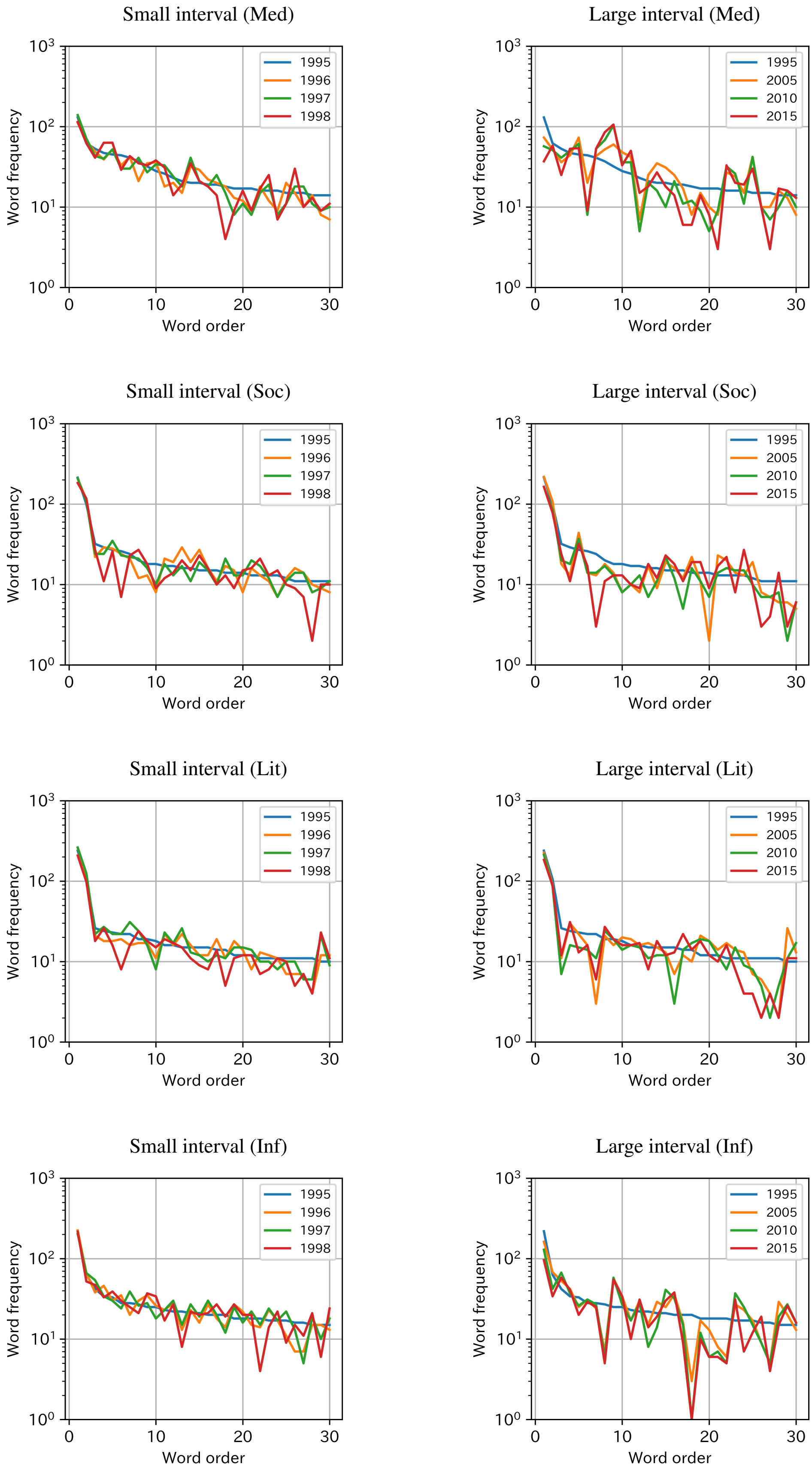


Figure 7: Small interval vs. large interval (for specified research areas)

It can be seen that the deviation from the blue line is larger for large intervals than for small intervals. It indicates that the corpora included small heterogeneity in a short span and large heterogeneity in a long span because some old-fashioned technical words that were frequently used in research themes in FY 1995 have become obsolete gradually in few years, and have rarely or never used in recent years. The results of a series of qualitative analysis indicated that there were observable time-oriented differences between the old (FY1995) and new (FY1996 to FY 2015) data for the experimental corpora. In the next section, we attempt quantitative analysis on the experimental data.

4 Experiments

To confirm the effectiveness of the proposed method, we conducted experiments on the aforementioned data, which included (i) research themes as documents and (ii) time-oriented attributes for the documents. For the experiment, documents in Japanese were used and the standard morphological analyzer MeCab[3] with IPADIC were used in the text pre-processing.

Table 2 compares the results by Method-S and Method-T for Corpus-L4000 and Corpus-L400. In the table, "Norm. Interval" and "Norm. Accuracy" indicate the normalized interval and the normalized accuracy, respectively. The values, 0.05, 0.50, 1.00 for the normalized interval correspond to one year, 10 years, and 20 years of time. While the results with the two methods for a short span of time (one year) were low, Method-T marked higher values than Method-S did for a long span of time (10 years, 20 years). This result suggests that non-negligible numbers of research themes in the KAKEN database should be highly heterogeneous, or obsolete in 10 to 20 years while there was no significant heterogeneity within a year or two.

Figure 8 visualizes the correlation of the normalized interval and the normalized accuracy for Corpus-L4000 and Corpus-L400 for a 20-year period from FY 1995 in the KAKEN database. In the figure, ML1, ML2, ML3, ML4 and ML5 indicate the five machine learning algorithms in our approach (see the explanation in Section 2). While the obtained results were higher and more uniformed when L is larger (i.e., L=4000), the visualization demonstrates a positive correlation, even for the relatively smaller value of L (i.e., L=400).

Table 3 presents detailed results of the single regression analysis. In the table, "Coef." indicates the coefficient obtained by the regression analysis. "Avg. N. Acc" indicates the average normalized accuracy. As can be seen, Method-T marked higher values for the coefficient, R2 score, and the average normalize accuracy than Method-S did. For further investigation, we performed paired *t*-test to compare the two methods. It was confirmed that the differences between the proposed method (i.e., Method-T) and the baseline method (i.e., Method-S) were statistically significant. In the table, ‡ indicates $p < 0.0001$. The results indicate that our proposed method is applicable to not only a larger corpus (i.e., L=4000) but also a smaller corpus (i.e., L=400).

Next, we describe the experimental results for Corpus-Med, Corpus-Soc, Corpus-Lit and Corpus-Inf. Table 4 compares the results by Method-S and Method-T for the four research areas. In all cases, the normalized accuracies were nearly zero in a short span. It indicates that no significant differences exist in research themes in the proximate years. In a long span, the results obtained by Method-T showed higher results than those by Method-S.

[3] `https://taku910.github.io/mecab/`

Table 2: Normalized interval and accuracy (for unspecified research areas)

Corpus name	Year1	Year2	Norm. Interval	Method	Norm. Accuracy
Corpus-L4000	1995	1996	0.05	Method-S	-0.0024
				Method-T	-0.0272
		2005	0.50	Method-S	0.0218
				Method-T	0.2244
		2015	1.00	Method-S	0.0007
				Method-T	0.3988
Corpus-L400	1995	1996	0.05	Method-S	-0.0320
				Method-T	0.0200
		2005	0.50	Method-S	-0.0525
				Method-T	0.1005
		2015	1.00	Method-S	-0.0265
				Method-T	0.2210

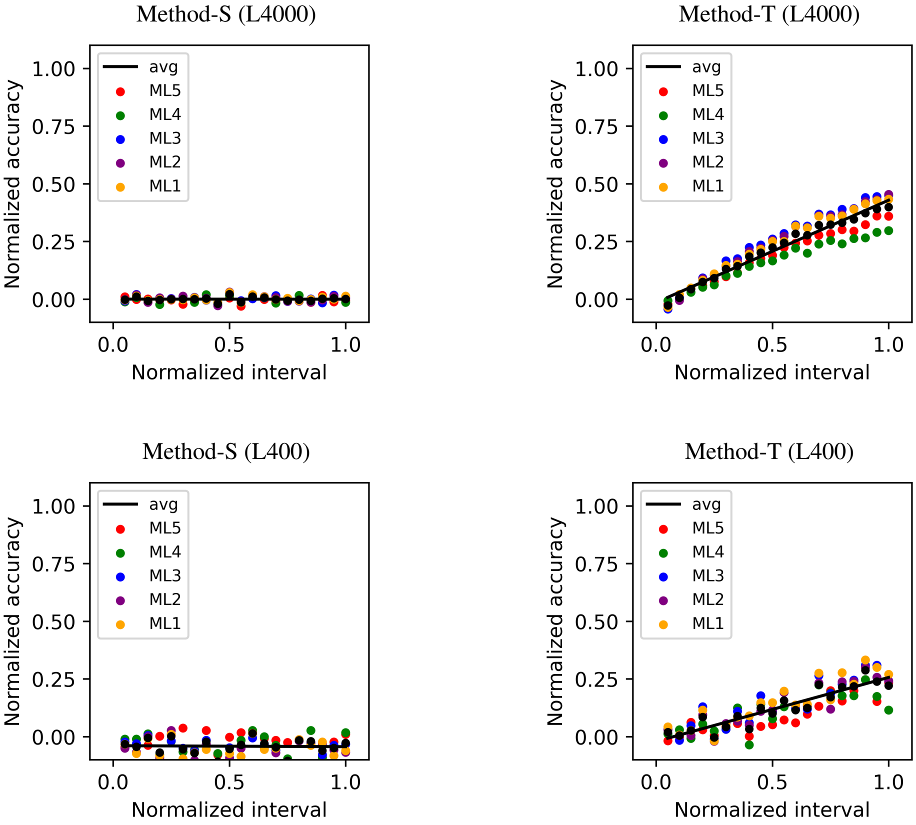


Figure 8: Transition detection (unspecified research areas)

Table 3: Regression analysis (for unspecified research areas)

Corpus name	Method	Coef.	Intercept	R2 score	Avg. N. Acc
Corpus-L4000	Method-S	-0.0003	0.0003	0.0001	0.0001
	Method-T	0.4427	-0.0141	0.9798	0.2183 ‡
Corpus-L400	Method-S	-0.0043	-0.0390	0.0016	-0.0413
	Method-T	0.2773	-0.0206	0.8578	0.1250 ‡

Table 4: Normalized interval and accuracy (for specified research areas)

Corpus name	Year1	Year2	Norm. Interval	Method	Norm. Accuracy
Corpus-Med	1995	1996	0.05	Method-S	-0.0375
				Method-T	-0.0445
		2005	0.50	Method-S	-0.0575
				Method-T	0.2340
		2015	1.00	Method-S	-0.0435
				Method-T	0.4485
Corpus-Soc	1995	1996	0.05	Method-S	0.0485
				Method-T	0.0315
		2005	0.50	Method-S	0.0275
				Method-T	0.3250
		2015	1.00	Method-S	0.0025
				Method-T	0.3555
Corpus-Lit	1995	1996	0.05	Method-S	0.0340
				Method-T	-0.0060
		2005	0.50	Method-S	0.0055
				Method-T	0.1690
		2015	1.00	Method-S	0.0045
				Method-T	0.2255
Corpus-Inf	1995	1996	0.05	Method-S	0.0175
				Method-T	-0.0495
		2005	0.50	Method-S	0.0285
				Method-T	0.3075
		2015	1.00	Method-S	0.0120
				Method-T	0.4655

Table 5: Regression analysis (for specified research areas)

Corpus name	Method	Coef.	Intercept	R2 score	Avg. N. Acc
Corpus-Med	Method-S	0.0134	-0.0428	0.0175	-0.0358
	Method-T	0.4930	-0.0385	0.9469	0.2203 ‡
Corpus-Soc	Method-S	-0.0213	0.0410	0.0550	0.0298
	Method-T	0.3869	0.0531	0.7891	0.2562 ‡
Corpus-Lit	Method-S	-0.0109	0.0073	0.0245	0.0016
	Method-T	0.3290	-0.0595	0.7956	0.1133 †
Corpus-Inf	Method-S	-0.0019	0.0209	0.0004	0.0199
	Method-T	0.5518	-0.0386	0.9345	0.2511 ‡

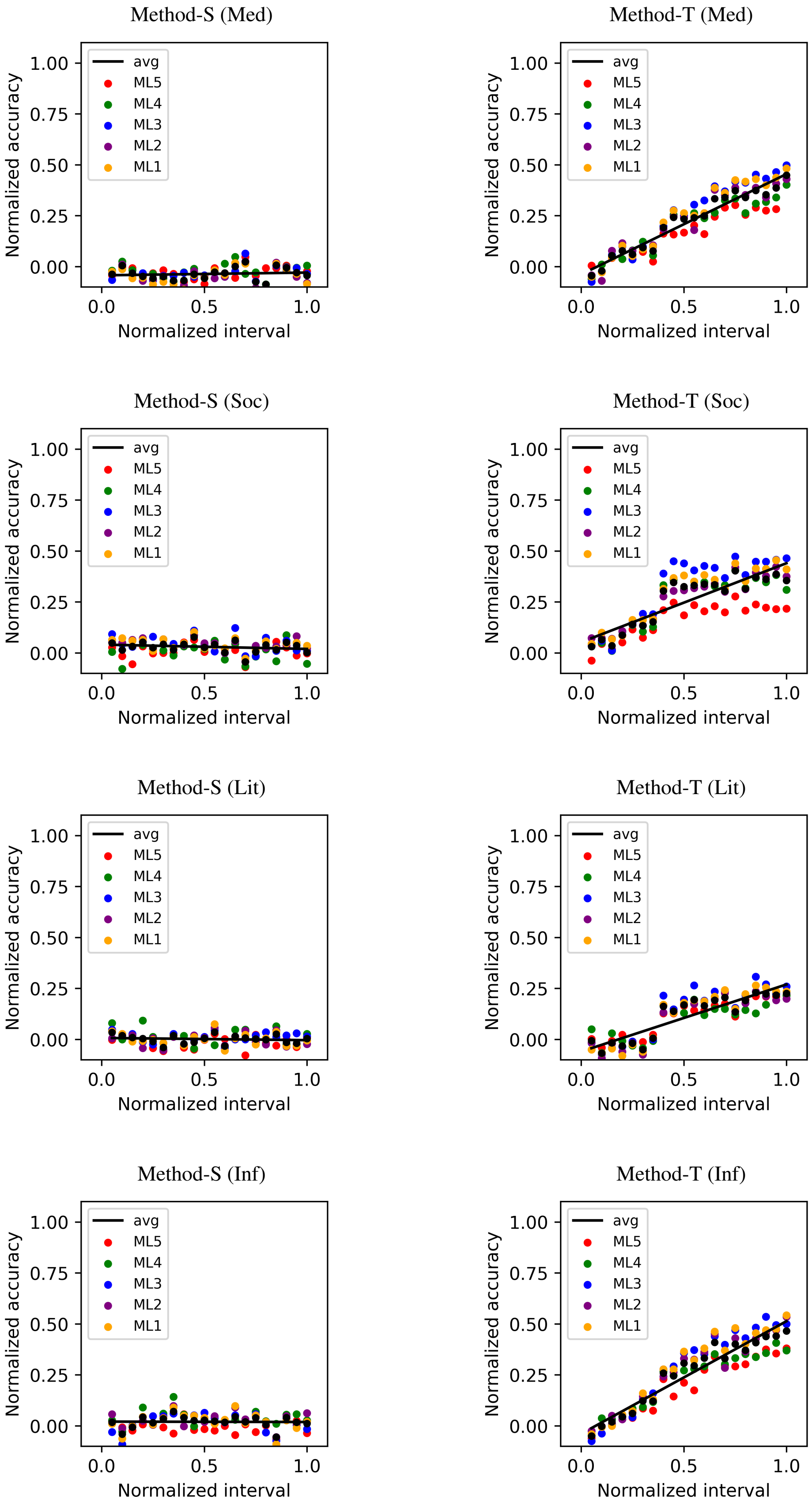


Figure 9: Transition detection (specified research areas)

Figure 9 presents the correlation between normalized interval and normalized accuracy for each of the corpora and each of the methods (Method-S and Method-T) using the five different machine learning algorithms. The transition detection applied Method-S, the normalized accuracies were all nearly zero regardless of the duration of the normalized interval. On the other hand, the transition detection applied Method-T showed a tendency for the positive relation between the normalized accuracy and the normalized interval. This figure suggests that in all four research areas, the heterogeneity between old and new research themes increased over the years. It is noteworthy that the rise in the normalized accuracy was larger in Corpus-Med and Corpus-Inf, while the rise in the normalized accuracy was not so large in Corpus-Soc and Corpus-Lit. In addition, Medical Science and Information Science have a high density of scattered points, suggesting high uniformity of document features in these corpora.

Table 5 shows the obtained values for the single regression analysis. For Corpus-Med and Corpus-Inf, the coefficient values were higher. These results suggest that the evolution of research themes in scientific research may be fast. On the other hand, the results for Corpus-Soc and Corpus-Lit showed relatively low values of the coefficient and R2 score. In these research areas, the heterogeneity of old and new data may be low while at the same time the homogeneity may be low as well. For the four corpora, we performed paired *t*-test to compare the proposed method (Method-T) and the baseline method (Method-S) and confirmed that the differences were statistically significant. In the table, † and ‡ indicate $p < 0.001$ and $p < 0.0001$, respectively. Based on the series of the experimental results, we have obtained convincing evidence that the proposed transition detection was applicable to corpora with different research areas while the trend of transitions and the homogeneity may differ according to the research areas.

5 Discussion

While human visual inspection of high-dimensional data (such as, document corpora) is limited in the capability, our proposed approach is expected to be useful in IR tasks to qualitatively analyze the transition detection in funding data. In the experiment, sampling data from a large textual database of grant applications showed a long-term trend over a 20-year period. This is an observation of a natural phenomenon: the evolution of the vocabulary usage in scientific research themes in a natural language. Once the research themes in the grant applications evolve rapidly within a 10-year period, the speed of the discrepancy between old and new research themes becomes less conspicuous. Then, the new research themes in the subsequent 10-year period gradually become diversified from the old research themes.

As can be seen from the results for the four specified research areas, the characteristics of the whole data are carried over to the partial data. In the same way as the holistic data, the research themes in the four different research areas changed rapidly in a short period of time. Then, the changing speed in the second decade became somewhat slower than the first decade. In the results for the specified research areas, different characteristics were observed in different research areas. Figure 5 shows that Medical Science included a larger number of projects for Medical Science while Information Science included a smaller number of projects. Regardless of the larger or smaller number of projects, both of these research areas showed high R2 scores in the regression analysis. This result indicates that the transition speed in each of these research areas was homogeneous and high. In con-

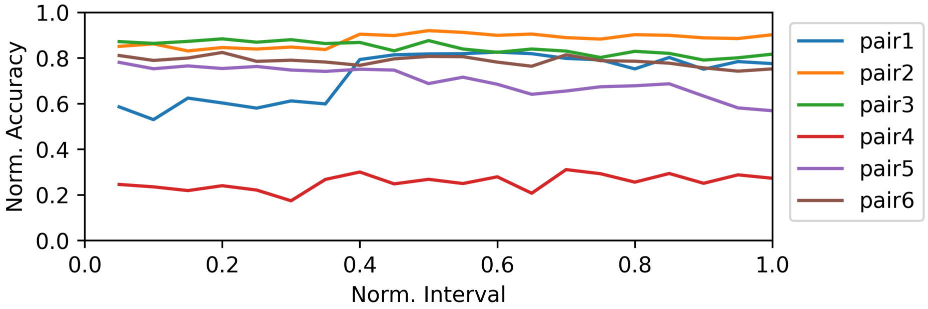


Figure 10: Pairwise comparison between different research areas for 20 years

trast, Social Science and Literature Science showed low R2 scores for the linear models, suggesting that there is a large heterogeneity of transition speed in these research areas.

It may seem obvious that old and new research themes in a particular research area become somewhat different over the years. However, such a conjecture can be baseless and unreliable. Let us consider (a) the difference between old and new research themes in the same research area and (b) the difference of two different research areas in the same year. Which is larger, (a) or (b)? As a basis for this argument, the pairwise classification for every combination in the corpora was conducted. The results are shown in Figure 10. In the figure, pair1, pair2, pair3, pair4, pair5, pair6 indicate the pair of Corpus-Med vs. Corpus-Soc, Corpus-Med vs. Corpus-Lit, Corpus-Med vs. Corpus-Inf, Corpus-Soc vs. Corpus-Lit, Corpus-Soc vs. Corpus-Inf, and Corpus-Lit vs. Corpus-Inf, respectively. As can be seen in the figure, there was a wide range of pairwise differences and the differences changed over the years. The maximum normalized accuracy was 0.9598 (Corpus-Med and Corpus-Lit in FY 2005), and the minimum normalized accuracy was 0.5870 (Corpus-Soc and Corpus-Lit in FY 2001). By comparing Figure 9 and Figure 10, it is suggested that the heterogeneity between new/old themes (e.g., Information Science in FY 2015 and FY 1995) was larger than the heterogeneity among different research areas (e.g., Literature Science and Social Science). This analysis on the actual data revealed that the intra categorical difference could be larger than the inter categorical difference in research themes in a funding database. This surprising fact should be a take-home message in this study.

Limitations of this study are as follows. The proposed method provides a practical tool, especially from the viewpoint of time-oriented attributes in grant applications. It would be more useful to be combined with other analytical methods from other perspectives. In Figure 9, salient clusters of scatter points can be observed in Social Science and Literature Science. In addition to natural phenomena, there is a possibility that governmental policy and intrinsic characters of research areas might be manifested in the outcome. Future studies should conduct further analysis of the objective facts presented by the data and the causal relationship between observed phenomena and human interventions.

6 Conclusion

We investigated a method for detecting transition of research themes in the KAKEN database. The proposed method first performs a brute-force document classification for the old/new binary document classification. Specifically, the beginning of the span is used for one class (e.g., the oldest year of the given period as the old class) and each year in the rest is used for the other class (e.g., the succeeding year as the new classes). Then, transition detection is achieved by regression analysis of the time span and the classification results. In order to obtain reliable results, five common machine learning algorithms were applied in the binary classification, and the average result was used in the regression analysis. As the proposed method uses time-oriented attributes, we conducted an experiment using research themes for about 20 years of adopted grant applications, by comparing with the baseline method that ignores time-oriented attributes. The experimental results showed the findings as follows. (1) The proposed method was able to detect a large discrepancy between the old and the new research themes over a long period of time. (2) Statistically significant differences between the baseline method and the proposed method were confirmed. For the review of the Grants-in-Aid for Scientific Research, the concept of research areas (e.g., Medical Science, Information Science) has been defined. While data-driven, scientific decision-making is desirable in institutional research, homogeneity within the same research area and heterogeneity among different research areas may be semiconsciously over estimated. As our proposed method enables numerical analysis of the time-oriented differences in documents, it is possible for IR experts to analyze new/old research themes on a quantitative manner.

Our proposed data can be applied to other data than the KAKEN database, such as bibliography database, educational-related data, etc. We anticipate that the current study will contribute to provide the methodological foundation for advanced analytical tasks in future institutional research.

Acknowledgments

This study was supported by the ISM Cooperative Research Program (2023-ISMCRP-0006) and JSPS KAKENHI Grant Number JP23K11764.

References

[1] M. Yasukawa and K. Yamazaki, “Categorizing bibliographic data for detection of transition in academic subjects,” in *9th International Congress on Advanced Applied Informatics, IIAI-AAI 2020, Online Congress, September 1-15, 2020*. IEEE, 2020, pp. 846–848.

[2] N. Yamashita, M. Numao, and R. Ichise, “Predicting research trends identified by research histories via breakthrough researches,” *IEICE TRANSACTIONS on Information and Systems*, vol. 98, no. 2, pp. 355–362, 2015.

[3] “Grants-in-Aid for Scientific Research – KAKENHI –,” https://www.jsps.go.jp/english/e-grants/index.html.

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

[5] P. R. Cohen and R. Kjeldsen, “Information retrieval by constrained spreading activation in semantic networks,” *Information processing & management*, vol. 23, no. 4, pp. 255–268, 1987.

[6] K. Aagaard, P. Mongeon, I. Ramos-Vielba, and D. A. Thomas, "Getting to the bottom of research funding: Acknowledging the complexity of funding dynamics," *Plos one*, vol. 16, no. 5, p. e0251488, 2021.

[7] F. Munari and L. Toschi, "The impact of public funding on science valorisation: an analysis of the erc proof-of-concept programme," *Research Policy*, vol. 50, no. 6, p. 104211, 2021.

[8] M. Dzieżyc and P. Kazienko, "Effectiveness of research grants funded by european research council and polish national science centre," *Journal of Informetrics*, vol. 16, no. 1, p. 101243, 2022.

[9] T. S. Kuhn, *The structure of scientific revolutions*. Chicago University of Chicago Press, 1970, vol. 111.

[10] H. Small, "Structural dynamics of scientific literature," *KO KNOWLEDGE ORGANIZATION*, vol. 3, no. 2, pp. 67–74, 1976.

[11] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National academy of Sciences*, vol. 101, no. suppl 1, pp. 5228–5235, 2004.

[12] M. Krenn and A. Zeilinger, "Predicting research trends with semantic and neural networks with an application in quantum physics," *Proceedings of the National Academy of Sciences*, vol. 117, no. 4, pp. 1910–1916, 2020.

[13] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[14] B. Bengfort, R. Bilbro, and T. Ojeda, *Applied text analysis with python: Enabling language-aware data products with machine learning*. " O'Reilly Media, Inc.", 2018.

[15] C. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*. Cambridge university press, 2008.

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[17] F. Sebastiani, "Machine learning in automated text categorization," *ACM computing surveys (CSUR)*, vol. 34, no. 1, pp. 1–47, 2002.

[18] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992.

[19] P. Langley, W. Iba, K. Thompson *et al.*, "An analysis of bayesian classifiers," in *Proceedings of the tenth national conference on Artificial intelligence*, 1992, pp. 223–228.

[20] L. Breiman, "Arcing classifier (with discussion and a rejoinder by the author)," *The Annals of Statistics*, vol. 26, no. 3, pp. 801–849, 1998.

[21] L. Bottou, "Stochastic gradient learning in neural networks," in *Proceedings of Neuro-Nîmes 91*, vol. 91, no. 8, 1991, p. 12.

[22] B. E. Boser, I. M. Guyon, and V. N. Vapnik, “A training algorithm for optimal margin classifiers,” in *Proceedings of the fifth annual workshop on Computational learning theory*, 1992, pp. 144–152.