An Analysis on Differences of Word Usage between Full and Short Conference Papers

Toshiro Minami and Yoko Ohura

***Abstract*—We have two aims in this study. The first one is to investigate differences of full and short papers in an academic conference. Full papers are supposed to contribute the field by providing with original outcome. Short papers, on the other hand, are considered to report ongoing specific research topics. We would like to know how these difference in characters appear in the papers, especially in their word usage. The results of this study should provide tips for writing valuable papers. The second aim of this study is to develop a methodology of analysis for investigating small data such as those we can collect individually. Our another research topic is to analyze educational data from our everyday university classes. The approach presented in this paper can be generalized and is applicable to other kinds of data analysis. As the result of the study in this paper, we recognize that short papers use technical words more frequently than full papers, which should reflect the characteristic difference of short and full papers.**

***Index Terms*—Academic material, bibliometrics, data mining, feature finding, small data analysis.**

I. Introduction

In this paper, we pursue research on differences between full, or regular, papers and short papers in a conference. The aims of this study are twofold. Firstly, we would like to find some kind of tips for writing better papers. It is quite important for academic members to write high quality papers. Academic papers submitted to a conference are assessed by reviewers whether they deserve to present at the conference. The papers are often classified into full, or regular, papers, short papers, and other types of papers. The classification is done based on not only on their types of research, but also on their quality. Therefore, it is crucial to know the characteristic differences between full and short papers, so that we can get tips to write high quality papers.

Secondly, we would like to develop a methodology appropriate in analytics for specific domains. One topic for this aim is to find features of the target data that differentiate them from other data. In this paper, they are the features that discriminate full and short papers. Dataset with relatively small in size is a disadvantage in general as we compare datasets with big sizes because it is quite hard to generalize what we find in the analysis of the dataset.

However, smallness may become an advantage. The datasets we come up with in our daily life are often very small, and easy to analyze in deep with our personal computers. Also, even though the findings in our small data analysis [1] may not be applicable to wider domains, they may be useful in our daily life domains because the findings are domain-specific and it would be quite difficult to extract useful finding from a very big dataset that is effectively applicable to such specific small domains.

As an example, we take another research topic of us. We have been doing research on relationship between university students' behaviour and their academic achievement [2], [3]. In our analysis of texts obtained as the answers to the question about their retrospective evaluation of the lectures, we found that students with wider view to learning have tend to have better achievement than those with narrower view.

In this study, we extracted words that may differentiate the students who use them and estimate the difference in their attitudes to learning. This approach is common with our study on finding differences of full and short papers.

The rest of this paper is organized as follows: In Section II, we show overview of the original data and how to create the data for analysis in the following steps.

In Section III, we explain our approach to the study and define basic notations and concepts as a preparation to the following sections.

In Section IV, we investigate the words that appear in the papers. An index is introduced, which indicates how much weight the word is used in full and short papers. Then, we introduce an index for a paper which is the counter concept of the index for a word. Using the index for a paper, we calculate its maximum accuracy in order to evaluate how much it can discriminate full and short papers. The result shows it is more accurate than to discriminate by number of pages of papers.

Finally, in Section V, we review what we have discussed in this paper, and show our possible future directions.

II. Data

A. Overview of Original Data

The data we use in this study are the full and short papers presented in the 9th International Conference on Computer Supported Education (CSEDU 2017) [4]. CSEDU 2017 conference consists of 10 sessions, namely, "Artificial Intelligence in Education (AIE, for short)," "Domain Applications and Case Studies (DACS)," "Information Technologies Supporting Learning (ITSL)," "Social Context and Learning Environments (SCLE)," "Special Session on Analytics in Educational Environments (SSAEE)," "Special Session on Fostering Open Leadership in School Culture (SSFOLSC)," "Special Session on Lifelong Learning (SSLL)," "Special Session on Serious Games on Computer Science Learning (SSSGCSL)," "Teaching Methodologies

Manuscript received May 4, 2020; revised November 12, 2020. This work was supported in part by JSPS KAKENHI Grant Number JP17K00502.

Toshiro Minami and Yoko Ohura were with Kyushu Institute of Information Sciences, 6-3-1 Saifu, Dazaifu, Fukuoka 818-0117 Japan (e-mail: minamitoshiro@gmail.com, ohura@kiis.ac.jp).

doi: 10.18178/ijke.2020.6.2.132

and Assessment (TMA)," and "Ubiquitous Learning (UL)." The total number of papers is 132.

TABLE I: NUMBER OF FULL/SHORT PAPERS OF SESSIONS

Session (Abr.)	Full	Short	Total
AIE	1	4	5
DACS	0	6	6
ITSL	19	49	68
SCLE	5	5	10
SSAEE	3	2	5
SSFOLSC	2	3	5
SSLL	3	2	5
SSSGCSL	0	5	5
TMA	6	13	19
UL	0	4	4
Total	39	93	132

Table I shows the session names and the numbers of full and short papers. Among 132 papers, 39 (30%) papers are classified as full and 93 (70%) are as short papers. We can see that ITSL is the biggest session of this conference, which has 68 papers. Among 68, 19 (28%) are full papers and 49 (72%) are short papers.

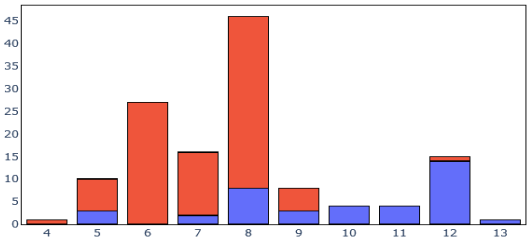

Fig. 1. Histogram of number of pages.

Fig. 1 shows the histogram of the number of pages of the papers. Lower, or blue, part of a bar shows the number of full papers and upper, or orange, part shows the number of short papers. Full papers are assigned a 12-page limit, whereas 8-page limit for short papers. Extra 4 pages are allowed with additional fee.

According to Fig. 1, the numbers of pages range from 4 to 13. More precisely, full papers' pages range from 5 to 13, and short papers, 4 to 12. The range difference is small because there are some papers exceptional in the number of pages; there is 1 short paper with 12 pages, and 3 full papers with only 5 pages.

It is interesting to see that there are three peaks; 6, 8, and 12. The numbers 8 and 12 are the page limit for short and full papers as was mentioned above, and can be easily understood that most authors might try hard to include as much outcomes they had in their studies. Papers with the number of pages of 6 and 7 might inspire us that some authors could not fill up the pages to the limit, possibly because of the time limit, or because of the amount of contents to be presented.

B. *Data for Analysis*

Fig. 2 shows the outline of how the papers are processed to the data that are available for analysis. At the first step of the process, we extract text data from the original papers in pdf format by using Python library of Apache Tika [5]. At the same time, we also get the number of pages of papers with Tika.

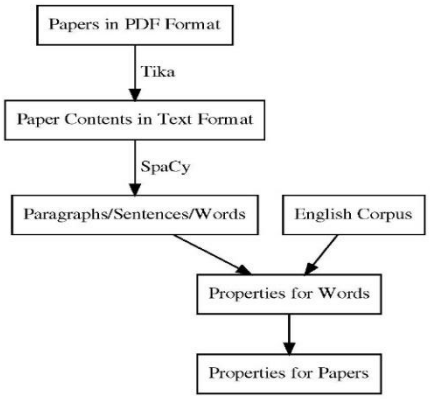

Fig. 2. Outline of process for analysis.

What we need to do next is to obtain the list of paragraphs from the original text data obtained by Tika. Since the papers are written in two-columns format, we need to connect the last line of the left column to the first line of the right column in a page. Also, we need to connect the last line of the right column of a page to the first line of the left column in the next page. There are footer part, page number, and header part in-between these two lines. We detect these parts and eliminate them.

Then, we use spaCy [6] for getting information about part of speech (POS). SpaCy provides with various functions for English such as tokenizer, tagger, parser, named entity recognition (NER) and word vectors. With the help of spaCy, we can sentencize paragraphs and obtain word tokens used in the papers.

By using the relationship between words and papers, we can define properties about words as well as about papers. In our previous studies [7], [8], we have observed that the words used in short papers are more specific than those used in full papers in general. In order to investigate further about this topic, we use the British National Corpus (BNC) [9].

III. CONCEPT AND NOTATION

Our aim in this study is to find properties of papers that discriminate full papers and short papers. We use accuracy for evaluating effectiveness of distinction. Even though our main concern in this study is the properties on word usage in the papers, we experiment with how much the number of pages discriminates full and short papers.

We can classify a prediction and its result into 4 cases:

TP (True Positive): The case when predicting some property is satisfied, and it is satisfied in reality.

TN (True Negative): The case when predicting some property is not satisfied, and it is not satisfied in reality.

FP (False Positive): The case when predicting some property is satisfied, but it is not satisfied in reality.

FN (False Negative): The case when predicting some property is not satisfied, but it is satisfied in reality.

Then, the accuracy measure for the prediction is generally defined as follows:

$$Accuracy(p) = \frac{\#TP + \#TN}{\#TP + \#TN + \#FP + \#FN} \quad (1)$$

where, "#" is the number of elements of the set.

As an example, we show how we set properties and how

to calculate accuracy of the property by predicting a paper to be full from its number of pages. Let n be the integer in-between 4 and 13, the minimum and maximum numbers of pages, respectively. We predict a paper full if its number of pages is greater than or equals to n, and short if not. According to the general definition, accuracy in this case is calculated as follows:

$$\text{Accuracy for } n = \frac{\text{TruePositive} + \text{TrueNegative}}{\text{Number of Papers}} \quad (2)$$

where, TruePositive is the number of full papers that have number of pages ≥ *n*, and TrueNegative is the number of short papers that have number of pages < *n*.

In our case, the number of papers is 132. As we calculate the accuracies by varying the number *n*, we have the best accuracy 0.87 when *n* = 10. This result is a kind of reasonable one, because the mean of pages for full papers is 9.87 and mean for short papers is 7.10. From this result, we say that the property of number of pages distinguishes full and short papers with accuracy 0.87. We consider this accuracy value as the baseline of our study. We would like to find more accurate property for discriminating full and short papers.

IV. Analysis of Words

One of our aims in the study of this paper is to find properties about word usage of papers and clarify the difference of full and short papers. In this section, we focus on the frequencies of words used in papers. By chunking the paragraphs of papers we have the word list and its frequency data.

A. Word Frequency

By making the union of the set of words used in a paper, we have 30,105 words that appear in at least one paper. Each word is used 26.0 times, 748.7 words appear in a page, 5927.08 words appear in a paper in average. The most frequently used words are ",", "the", ".", "of", "and", "be", "to", "in", "-pron-", and ")." Note that punctuation symbols such as period and comma are considered to be words.

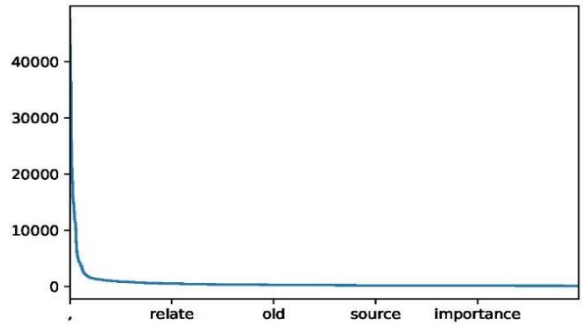

Fig. 3. Frequencies of words.

Fig. 3 shows the frequencies of words in descending order. We can see that frequencies drop very rapidly.

One of our ideas for finding differences of word-usage between full and short papers is to define appropriate indexes about words, with which we can see which words are used more popularly in full papers than short papers, or, vice versa, and how much they are different.

B. Index of Word for Measuring Full/Short-Orientedness

In order to describe our ideas, we give some notational definitions here. Let *P* be the set of papers, *F* be the set of full papers, and *S* be the set of short papers. From the definitions, $P = F \cup S$ and $F \cap S = \varnothing$ hold. Let *W* be the set of all words that appear in *P*. For a paper p ($\in P$) and a word w ($\in W$), we use the notation $w \in p$ if w is included, or is used, in p.

For a word *w*, let *Fcnt*(*w*) be the number of occurrences of *w* in full papers. Similarly, *Scnt*(*w*) be the number of occurrences of w in short papers. Note that *Fcnt*(*w*), *Scnt*(*w*) ≥ 0 for any word w ($\in W$).

Before defining the final version of full-short index of a word *w*, which shows how much more it occurs in full papers, or in the opposite way, how much it occurs in short papers, we define the first version of index *FSidx*1(*w*) for a word $w \in P$ as follows:

$$FSidx1(w) = \frac{Fcnt(w) - Scnt(w)}{Fcnt(w) + Scnt(w)} \quad (3)$$

if *Fcnt*(*w*) + *Scnt*(*w*) > 0, i.e., if *w* appears either in full papers or in short papers.

From this definition, we can say that $-1 \le FSidx1(w) \le 1$ for any $w \in P$, *FSidx*1(*w*) = 1 if and only if (iff) the word *w* appears only in full papers, i.e., $w \in F$ and $w \notin S$, and *FSidx*1(*w*) = −1 iff *w* appears only in short papers, i.e., $w \in S$ and $w \notin F$. Further, *FSidx*1(*w*) = 0 iff *Fcnt*(*w*) = *Scnt*(*w*), i.e., *w* appears same times in full and short papers, *FSidx*1(*w*) > 0 iff *Fcnt*(*w*) > *Scnt*(*w*), i.e., w appears more in full papers than in short papers, and *FSidx*1(*w*) < 0, vice versa.

Unfortunately, this index has a problem when the numbers of full papers and short papers are different. In our case, there are 39 full papers (#*F* = 39), whereas there are 93 short papers (#*S* = 93), and thus, the numbers are very different ($\#S \gg \#F$). Suppose a word w appears once in all papers. Then, *Fcnt*(*w*) = 39 and *Scnt*(*w*) = 93, and thus, *FSidx*1(*w*) = (39−93) / (39+93) = −0.41. From the index value, the word *w* looks like to appear much more in short papers even though it appears uniformly in all papers. In order to correct this problem, we use the ratio of occurrences instead of raw occurrence numbers of *Fcnt* and *Scnt*. Let *Fr*(*w*) be the ratio of word *w* in full papers, which is formally defined by:

$$Fr(w) = \frac{Fcnt(w)}{\sum_{w' \in F} Fcnt(w')} \quad (4)$$

where, the denominator is the total number of word occurrences in full papers.

We define the ratio of *w* in short papers *Sr*(*w*) similarly.

Then we have the revised, and final, definition as follows:

For a word w ($\in P$),

$$FSidx(w) = \frac{Fr(w) - Sr(w)}{Fr(w) + Sr(w)} \quad (5)$$

C. Distribution of FSidx Values

Now, we have *FSidx*(*w*) = 0 if the word *w* appears in the same ratio in full papers and in short papers. We call a word *w* *F*-oriented, or a *F*-word in short, iff *FSidx*(*w*) > 0, and a *S*-oriented, or *S*-word, iff *FSidx*(w) < 0. Note that the properties

discussed for *FSidx*1(*w*) holds also for *FSidx*(*w*) by definition.

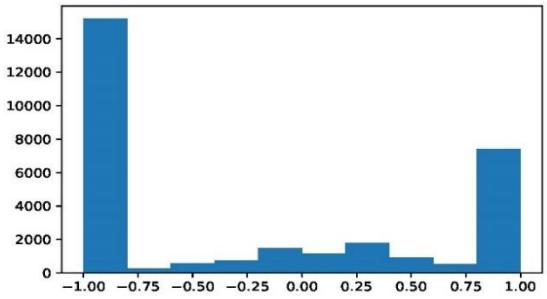

Fig. 4. Histogram of *FSidx* of words.

Fig. 4 shows the histogram of *FSidx*(*w*). Quite a lot of words are located at the values −1 and 1. Precisely, 7,289 words are used only in full papers, or *FSidx*(*w*) = 1, and 15,126 words are used only in short papers, or *FSidx*(*w*) = −1.

Actually, most of these words appear only in one paper; 6,604 (90.6%) words out of 7,289 words appear only in one paper and 12,930 (85.5%) words are only in one short papers. The words that appear only in one full paper include "ambre," "racket," "ls," "tabletop," "opera," "ux," "gb," "pibook," and "wordnet," where those words in short papers include "ipr,""pronoun," "vw," "antecedent," "puppet," "balloon," iot," "hygiene," "irish," "cg," "vrets," "clipit," and "orchestration."

To summarize, 19,534 words out of 30,105 words (65%) appear only in one paper and have *FSidx* value either 1 or −1.

D. FSidx Index of a Paper

Now, we can define an index of a paper that indicates if a paper is full paper oriented or short paper oriented, and the amount of *F*- or *S*-orientedness by using *FSidx*(*w*) of word *w*. Our idea for this index is to define *FSidx* value of a paper as the mean of the *FSidx* values of the words that are used in the paper. Roughly speaking, a paper is *F*-oriented if it uses more full paper oriented words (*F*-words) than short paper oriented words (*S*-words), and is *S*-oriented, vice versa.

We formally define *FSidx*(*p*) for a paper *p* (∈*P*) as follows:

$$FSidx(p) = \frac{\sum_{w \in p} FSidx(w) occ_p(w)}{\sum_{w \in p} occ_p(w)} \quad (6)$$

where, $occ_p(w)$ is the number of occurrences of *w* (∈*W*) in *p* (∈*P*).

Fig. 5 shows the histogram of *FSidx*(*p*) of papers. The values range from −0.00033 to 0.00027. The range size is very small because most words in a paper are commonly used in full and short papers, and relatively small number of words are used characteristically used in either full or short papers. To be more precise, among some 78 thousand occurrences of words, 94.5% are of the words commonly used in both full and short papers, whereas only 3.9% are those that appear only in short papers, and 1.6% are sued only in full papers.

For the words themselves, among about 30 thousand words, 50.2% words appear only in short papers, 24.2% appear only in full papers, and the rest 25.6% appear in both full and short papers. In other point of view, the words commonly appear in full and short papers are used about 96.1 times in average, the words that appear only in full papers are used about 1.7 times, for the words in short papers, about 2.0 times, and for all words 26.0 times.

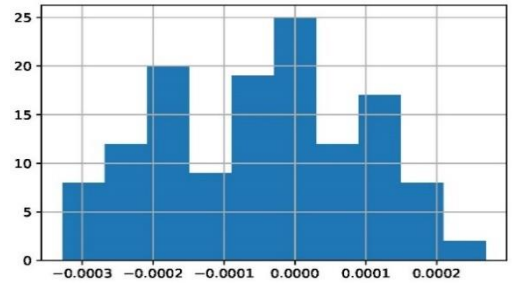

Fig. 5. Histogram of FSidx of papers.

Since the mean of *FSidx*(*p*) for full papers is 0.00010 and −0.00011 for short papers, we can roughly estimate a paper *p* if it is a full paper or a short one by its *FSidx* value.

E. Accuracy of FSidx(p) of Papers

We have the new property *FSidx*(*p*) for papers that is supposed to discriminate full and short papers. We can calculate its accuracy by using this property.

We have *FSidx* value *FSidx*(*p*) of *p* for each paper *p* (∈ *P*), Let $P = p_1, p_2, \ldots, p_n$ so that the elements are in ascending order in their *FSidx* values; i.e., $FSidex(p_i) \leq FSidx(p_{i+1})$ for $1 \leq i \leq n-1$.

For a given threshold value *t* (−1≤ *t* ≤1), we predict a paper *p* as full (positive) if *FSidx*(p) > *t*, and predict short (negative) otherwise, i.e., if *FSidx*(*p*) ≤ *t*. Then, *p* is true positive (TPt) if *p*∈ *F* and *FSidx*(*p*) > *t*, and is true negative (TNt) if *p*∈*S* and *Fsidx*(*p*) ≤ *t*. Similarly, *FPt* if *p*∈*S* and *FSidx*(*p*) > *t*, and *FNt* if *p*∈*F* and *FSidx*(*p*) ≤ *t*.

Accuracy in this case is defined similarly as follows:

$$Accuracy_t(p) = \frac{\#TP_t + \#TN_t}{\#TP_t + \#TN_t + \#FP_t + \#FN_t} \quad (7)$$

Note that "#" is the number of elements of the set. In order to calculate the maximum value of accuracy in this case, it is sufficient to take the threshold values *t* such that $t = FSidx(p_i)$ for some *i* = 1, 2, . . ., *n*. It is unnecessary to calculate other values of *t* in-between these values because accuracy values are the same for any *t* if $FSidx(p_i) \leq t < FSidx(p_{i+1})$.

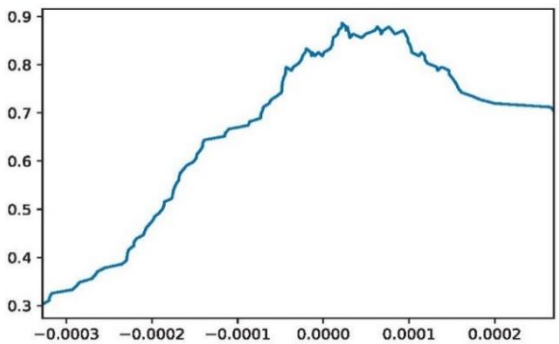

Fig. 6. Relation between accuracy (y-axis) with Threshold value of FSidx (x-axis).

Fig. 6 shows accuracy values for different thresholds. The best accuracy is 0.886 at the threshold value *t*=0.000022 (at the paper "S003"), where #*TP* = 34, #*TN* = 83, #*FP* = 5, and

#*FN* = 10. Thus, accuracy measure for *FSidx*(*p*) becomes 0.886 (88.6%).

We can evaluate that *FSidx*(*p*) is a good measure to discriminate full and short papers because its accuracy is better than that of using number of pages (0.87), which we have shown at the end of Section III.

V. Concluding Remarks

In this paper, we have investigated differences between full and short conference papers in terms of word-usage. One of our aims is to find properties about papers that characterize/discriminate full and short papers.

We defined an index FSidx(p) of a paper p which shows how much amount the paper uses the words that appear in full/short papers. Experimental results show that the maximum accuracy of this index to discriminate full and short papers is 0.886, which outperforms the property of number of pages. This is the first outcome to be noted in this paper.

Another outcome of this paper is that we can find such property by following our approach in small data analysis. If we extract properties from big data, we can apply the properties to a wide variety of applications. On the other hand, it is difficult to extract results that are effectively applicable to specific application fields because the extracted properties are supposedly satisfied in a wide area uniformly.

The study in this paper can be classified as that of bibliometrics, the major approach in this research field is citation analysis [10], [11], which evaluate a paper based on how and how much it is cited by other papers. We did not take this approach because we are interested more in evaluating papers before publishing rather than after they are published.

In small data analysis, we would need to have many trial and error experiments by considering characteristic properties of the dataset. Our approach is to start with small data and extract domain specific results at the first step. Then, we expand the target data and see if the properties we have found are also applicable to different and/or new data. Most of the cases, we would need some modifications to analysis methods in the first experiment. We tune our analysis methods so that the new methods are applicable to wider datasets. In this way, we would develop effective analysis methods for our target field.

The analysis methods we carried out in this paper is still in a very early stage toward our goal to find effective and practical methods to get tips in order to write more sophisticated papers.

In order to obtain satisfactory results, we need to investigate the papers more deeply including the following topics.

(1) Analysis not only with usage of specific words but also with usage of types of words; what features of word usage are more full-paper oriented and what are more short-paper oriented. Even though the features might be highly domain-specific, it must be very effective.

(2) In this paper, we partly consider the parts of speech (POS) of words. By considering POS of words, we would be able to find more informative results that characterize full papers from short papers, and vice versa.

(3) Analysis of different types of data, such as organization of papers, citation data is another challenging topic. By combining the properties obtained from different approaches, we are able to find better properties for discriminating full and short papers.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

All authors had approved the final version.

Acknowledgment

This work was supported in part by JSPS KAKENHI Grant Number JP17K00502.

References

[1] R. Kitchin and T. P. Lauriault, "Small data in the era of big data," *GeoJournal*, vol. 80, pp. 463-475, 2015.

[2] T. Minami, S. Hirokawa, Y. Ohura, and K. Hashimoto, "A part-of-speech-based exploratory text mining of students' looking-back evaluation," *Advances in Natural Language Processing, Intelligent Informatics and Smart Technology* ed. T. Theeramunkong *et al.*, Advances in Intelligent Systems and Computing, vol. 684, pp. 61-72, 2018. Springer-Verlag. https://doi.org/10.1007/978-3-319--70016-8_6

[3] T. Minami, Y. Ohura, and T. Baba, "A characterization of student's viewpoint to learning and its application to learning assistance framework," in *Proc. the 9th International Conf. on Computer Supported Education (CSEDU 2017)*, 2017, vol. 1, pp. 619-630.

[4] P. Escudeiro, G. Costagliola, S. Zvacek, *et al.* (Eds.). *Proceedings of the 9th International Conference on Computer Supported Education (CSEDU 2017)*. [Online]. Available: https.//www.scitepress.org/ProceedingsDetails.aspx?ID=dor1zNQw36c=&t=1

[5] Apache. *Apache Tika - A Content Analysis Toolkit.* [Online]. Available: https://tika.apache.org/

[6] SpaCy. *Industrial-Strength Natural Language Processing in Python.* [Online]. Available: https://spacy.io/

[7] T. Minami and Y. Ohura, "Difference analysis of word-usage between full and short papers," in *Proc. the 2018 International Conf. on Big Data and Education (ICBDE 2018)*, 2018.

[8] T. Minami and Y. Ohura, "How different are full and short papers in word-usage?" *International Journal of Machine Learning and Computing (IJMLC)*, vol. 10, no. 1, 2020.

[9] British National Corpus (BNC). [Online]. Available: https://www.english-corpora.org/bnc/

[10] T. Nakatoh, S. Hirokawa, T. Minami, T. Nanri, and M. Funamori, "Attribute-based quality classification of academic papers," *Journal of Artificial Life and Robotics*, vol. 23, no. 2, pp. 235-240, 2018. Springer-Verlag New York, Inc. [Online]. Available: https://doi.org/10.1007/s10015-017-0412-z

[11] M. Tang, H. Liao, Z. Wan, E. Herrera-Viedma, and M. A. Rose. (2018). Ten years of sustainability (2009 to 2018): A bibliometric overview. *Sustainability*. [Online]. 10(5). P. 1655. Available: https://doi.org/10.3390/su10051655

Toshiro Minami was born in Japan. He received his BS degree in engineering from Kyushu Institute of Technology, Japan in 1973, and received the MS and D.Sc. degrees in mathematics and computer science from Kyushu University, Japan in 1975 and 1999.

He was a researcher of Fujitsu Limited and Fujitsu Laboratories Limited, Japan from 1984 to 1999. He is a research fellow of Australian National University from 1992 to 1993. He is an associate professor of Kyushu

University Library, Japan from 1999 to 2001. He has been a professor of Kyushu Institute of Information Sciences from 2001 to 2016, and is an emeritus professor since then.

Professor Minami's research interests include educational, bibliometric, and library data analytics, as well as library informatics, library marketing, artificial intelligence, and multi-agent systems.

Yoko Ohura was born in Japan. She received the BS and D.Ph. degree from Saga University in 1978 and 1998, respectively.

She was a research assistant of Fukuoka University from 1978 to 1993. She is an associate professor of Aso Fukuoka Junior College from 1993 to 1998, and she reorganized to Kyushu Institute of Information Sciences since 1998. She has been a professor of there since 2005.

Her research interests include numerical simulation, statistics and data analysis.