How Different Are Full and Short Papers in Word-Usage?

Toshiro Minami and Yoko Ohura

***Abstract*—Academic papers submitted to a conference are assessed by reviewers and judged if they deserve to be presented at the conference. The accepted papers are often classified into full papers, short papers, and other types, according mainly to the reviewers' assessment. The major aim of the study presented in this paper is to find tips which are effective for a paper to be improved so that a paper supposed to be classified as a short paper becomes a full paper. In this study, we investigate a scenario for finding the differences between full and short papers on the usage of words/terms. Then, we extract words which are characteristic for either full or short papers through an experimental study. In order to find these words, we introduce a couple of indexes of a word. The results inspire that we can obtain practical tips in this approach by refining this method.**

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

I. Introduction

It is a big issue for researchers and graduate students in academic organizations how to write highly assessed papers so that they can obtain higher academic degrees and/or higher reputations. Academic papers submitted to a conference are reviewed and judged according to their assessment if they deserve to be presented at the conference. The accepted papers are often classified to full papers, short papers, and other types, according mainly to the reviewer's assessment result.

Our major aim of the study presented in this paper is to find any kind of effective tips for improving a paper which should be evaluated as a short paper to improve so that it becomes evaluated as a full paper. Among various considerable candidates, we take an approach of finding tips by analyzing sample papers in this study as the very first step toward our goal. More specifically, we investigate a scenario for finding the differences between full and short papers in their usage of words, or terms, and show some of the most discriminating words in the process of analysis.

One of the contributions of this paper is to propose a couple of indexes of words which show how much they are used in full and short papers. Precisely, we define a function called FS-index (full-short-index) for a word. FS-index shows a ratio which represents how much amount is the word used in full papers in comparison with the amount for short papers. Furthermore, we define a function called μ-index (mixed index), which considers both FS-index value and the popularity of usage of the word.

Even with an analysis on word/term-usage in this paper, the results inspire us that we should obtain some kind of practical tips if we investigate not only the differences in word-usage, but also the differences of the organization, and other features of the papers by refining the analysis method used in this paper.

In our different study, we proposed a method for discriminating full and short papers for the same data used in this paper. The proposed method gives better performance than discriminating by the number of pages [1].

We have been experiencing studies of analyzing other types of data in a similar approach [2], [3]. From the free texts for retrospective evaluations of students, we have found that the students with wide perspective for learning have better outcome, or examination score, than those who have narrow perspective. We take the approach of introducing appropriate measuring index(es) considering the specific needs of the problem, and finding some properties which are interesting and informative.

Use of citation data is a different approach from that in this study for assessing importance of papers. Nakatoh *et al*. proposed the concept of focused citation index in this approach [4].

The rest of this paper is organized as follows: In Section II, we describe about the target data for analysis, including the tools we use.

In Section III, we start with analyzing the total frequencies of words, so that we recognize how the words are used in the target set of papers. Then, we go forward to analyze the usage of words how the words appear in the full and short papers.

Finally, in Section IV, we summarize our discussions and results in this paper and show our future possible directions.

II. The Target Data

The target data used in this paper are the academic papers presented in the session of "Area 2 - Information Technologies Supporting Learning" of the 9th International Conference on Computer Supported Education (CSEDU 2017) [5]. Among 68 papers, 19 of them (28%) are full papers and the remaining 49 (72%) are short papers.

Full papers are assigned a 12-page limit, whereas 8-page limit for short papers. Extra 4 pages are allowed if necessary with additional fee.

The numbers of pages of the actual papers vary from 5 to 13. Table I shows the numbers of full and short papers for the given number of pages. We can see that 12 out of 19 (63%) full papers have smaller papers than the limit of 12 and 1 paper has more than the limit number of pages. For the short papers, 24 papers out of 49 (49%) short papers have smaller number of pages than the limit number of 8, and 2 papers

Manuscript received December 9, 2019; revised April 18, 2020.

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

Y. Ohura is with Kyushu Institute of Information Sciences (KIIS), 6-3-1 Saifu, Dazaifu, Fukuoka 818-0117 Japan (e-mail: ohura@kiis.ac.jp).

doi: 10.18178/ijmlc.2020.10.4.976

have more than the limit number of pages.

TABLE I: THE NUMBERS OF FULL AND SHORT PAPERS FOR THE GIVEN NUMBER OF PAGES

Number of Pages	Number of Full Papers	Number of Short Papers
5	1	1
6		17
7		6
8	5	23
9	1	2
10	2	
11	3	
12	6	
13	1	
Total	19	49

We can see also that 6 full papers have the number of pages that are in the range for short papers. Therefore, we cannot determine if a given paper is full or short only from the number of pages.

The pre-processing of data consists of three steps:

1) To convert the papers in pdf format into text data by applying a pdf to text converter software.
2) To formulate the text file for analysis suitable to the analysis using KH Coder [6], i.e. by adding HTML tag like information.
3) To apply KH Coder and extract frequency data of words, or terms.

In the first step (1) for pre-processing the data, the information about the arrangement, non-text part such as figures and tables, and some other organizational information is eliminated. The remaining data contain only the text parts of the original papers. We use the obtained text data without editing.

KH Coder is a free software which is convenient to analyze text data. It can deal with hierarchical data by using HTML-like tag <h1> to <h5>. The <h1> tags show the topmost group of text data, and <h2> tags show the lower level data, and so on. We can change the target range by specifying the part for analysis to KH Coder.

The file obtained in step (2) in the study of this paper has the following hierarchical organization: The topmost <h1> tag is used for grouping full and short papers. For example, "<h1>FullPapers</h1>" specifies that the full papers will follow to this tag.

The second <h2> tag is used for specifying the header of each paper of full and short papers. For example, "<h2>file:FCSEDU_2017_101_CR.pdf.txt </h2>" specifies the paper with the number 101. The "F" character after the string "file:" indicates that this paper is classified as a full paper. The content of the paper follows the "<h2>" tag. For example, the content of the paper number 101 starts with the string as follows: "Personalized, Affect and Performance-driven Computer-based Learning Christos Athanasiadis, Enrique Hortal, Dimitrios Koutsoukos, ".

After the step (iii), i.e., by applying KH Coder, we obtain the statistical data of all the papers: The file contains 58,008 sentences, 47,168 paragraphs, 68 <h2> tags (the number of papers), and 2 <h1> tags (full and short).

By choosing the Tools>Words>Frequency List menu and choosing the options "By POS tags" and "Term frequency", we obtain the word frequency data by POS tags, which are used in the following analysis steps.

III. WORD USAGE ANALYSIS

In this section, we analyze the word-usage data and investigate the characteristic features for discriminating the full and short papers.

This section is organized as follows: We start the analysis with investigating how the words used in the papers are distributed in terms of the frequencies in the full and short papers in Section A.

Then in Section B, we define an index for characterizing whether and how much a word is used in full and short papers. We also investigate how they are distributed by their histogram.

In Section C, we introduce another index called Pop(w) for measuring popularity of words and see how the words are used from these two indexes.

Finally, in Section D, we define an index for measuring importance of words and we show some of the most characteristic words for full and short papers.

A. Preliminary Analysis of Word-Usage as a Whole

We start with capturing what words/terms are used in the papers; especially, what are the differences in the full and short papers. Fig. 1 and Table II show how words are used in full papers and short papers in terms of frequency (the number of occurrences).

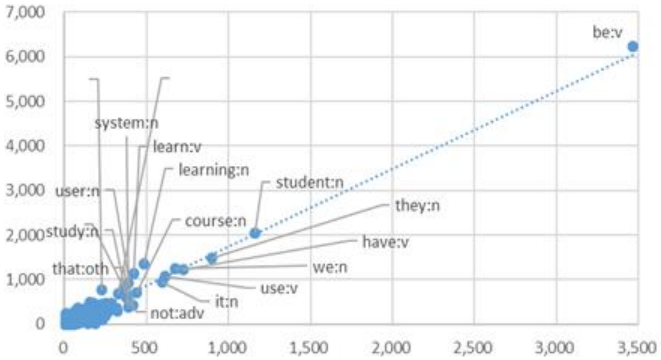

Fig. 1. Frequencies of words used in full papers (x-axis) and in short papers (y-axis).

A word is represented in the form "word:POS" so that it shows in what part-of-speech the word is used in the paper. For example, "be:v" represents the verb "be" and "student:n" represents the noun "student".

As we can see easily, the verb "be" is used much more than other words, and it is followed by the popular nouns such as "student", "they" in both full and short papers.

For full papers, "have" (verb), "we" (noun), "use" (verb), "it" (noun), "learning" (noun), "course" (noun), "learn" (verb), and "user" (noun) are following, whereas for short papers, "learning" (noun), "we" (noun), "have" (verb), "learn" (verb), "use" (verb), "it" (noun), and "system" (noun) are following.

The underline in the list for full papers in Table II shows that the rank of the word is higher (smaller ranking number) than that in the list of short papers; and vice versa for those

for short papers. These words could be considered to somehow represent the differences of the full and short papers. Especially if a word has big difference in their ranking numbers in full and short papers, it may represent a kind of characteristic difference between full and short papers.

TABLE II: THE TOP 15 MOST FREQUENTLY-USED WORDS IN FULL AND SHORT PAPERS

Full Papers			Short Papers		
Rank	Word:POS	Freq.	Rank	Word:POS	Freq.
1	be:v	3,466	1	be:v	6,234
2	student:n	1,159	2	student:n	2,047
3	they:n	896	3	they:n	1,487
4	have:v	727	4	learning:n	1,345
5	we:n	674	5	we:n	1,245
6	*use:v*	614	6	have:v	1,218
7	*it:n*	596	7	learn:v	1,138
8	learning:n	483	8	use:v	1,070
9	course:n	438	9	it:n	948
10	learn:v	423	10	system:n	921
11	user:n	417	11	education:n	823
12	system:n	389	12	datum:n	761
13	that:oth	389	13	course:n	707
14	study:n	387	14	which:oth	679
15	not:adv	378	15	that:oth	614

By comparing these words, we are inclined to think that the words in short papers such as "system" and "datum" inspire that the short papers deal with more about specific topics such as educational systems than full papers.

On the other hand, the words in full papers such as "course" and "user" inspire that full papers intend to deal with more about the topics from wider points of view such as the curriculums, educational framework, and students who are the main participants in education.

B. Analysis of Word-Usage of Nouns in Full and Short Papers

In this section as well as the rest of this paper, we mainly deal with nouns as the target words for analysis. The nouns are the most important part of speech (POS) because they carry the main idea, or the subject, of the sentences.

Now we would like to adjust the weight of occurrence of words. As the numbers of word occurrences in the full papers and in the short papers are different, one occurrence of a word in a full paper and in a short paper has different weight among all the occurrences of words. Thus, we would use the ratios of occurrences in full papers and in short papers.

For a word w, we define the ratio of w for full papers by

$$Fr(w) = \frac{\#\text{occurrences of } w \text{ in full papers}}{\#\text{all occurrences of words in full papers}},$$

where # symbols stand for the number of the following set.

From this definition, $Fr(w)$ is the ratio of the word w among all words in their occurrences in full papers.

We define $Sr(w)$ in the similar way for short papers.

$$Sr(w) = \frac{\#\text{occurrences of } w \text{ in short papers}}{\#\text{all occurrences of words in short papers}}$$

Note that $0 \le Fr(w) < 1$ and $0 \le Sr(w) < 1$.

Now, we define FS index of w by

$$FS(w) = \frac{Fr(w) - Sr(w)}{Fr(w) + Sr(w)}.$$

We also define $FS(w)=0$ if $Fr(w)=Sr(W)=0$.

Note that $-1 \le Fs(w) < 1$ and $FS(w) = 0$ if and only if (iff) $Fr(w) = Sr(w)$, According to the definition $FS(w)$ shows the weight of usage of the word w how much it is used in full papers and how much in short papers. $FS(w)>0$ shows that w is used more in full papers and $FS(w)=1$ means w is used only in full papers. Similarly, $FS(w)<0$ means that w is more used in short papers than in full papers.

Furthermore, $FS(w) > 0$ iff $Fr(w) > Sr(w)$, $FS(w) < 0$ iff $Sr(w) > Fr(w)$, $FS(w) = 1$ iff $Fs(w) \neq 0$ and $\mathrm{Sr}(w) = 0$ (the word w appears only in one or more full papers), and $FS(w) = -1$ iff $Sr(w) \neq 0$ and $Fr(w) = 0$ (the word w appears only in one or more short papers). From these properties, we say a word w by F-word if $FS(w) > 0$ and S-word if $FS(w) < 0$.

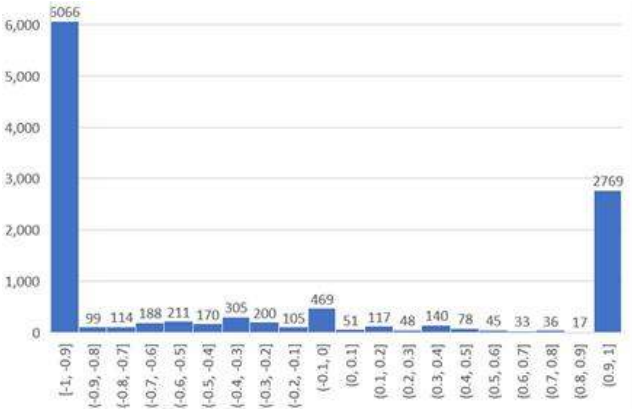

Fig. 2. Histogram for FS index of Nouns.

Fig. 2 shows the histogram of the FS index values of words. We can see that quite a few words have FS indexes close to either 1 or -1. Actually, 2,766 words have value 1, i.e., they appear only in full papers. Among them, 1,804 (65%) words appear only once. Thus, they appear only in a paper. For short papers, 6,028 words only appear in short papers, and 3,605 (60%) of them appear only once.

Among these words, the number of S-words is greater than that of F-words. However, by dividing the number of papers, the number of S-words per one paper is 124, whereas that of F-words is 146. Thus, we could say that these values are almost the same with full and short papers.

C. Word-Usage Analysis with Two Indexes

As we have seen in the previous subsection, the word having FS index of 1 do not mean that they characterize full papers. For example, a word with occurrence 1 has FS index value of 1 if it happens to be used in a full paper. Also, if it happens to be used in a short paper, FS index value becomes -1. Therefore, we cannot say that the word characterize either full papers or short papers from the fact that FS value is 1 or -1.

Thus, the characteristic words we are looking for should be chosen by considering the popularity, or the number of

occurrences, so that total frequency ratio should not be very low. Let us define the popularity of word w by the mean value of Fr(w) and Sr(w); i.e.,

$$\text{Pop(w)} = \frac{Fr(w) + Sr(w)}{2}$$

Fig. 3 shows a scatter diagram between FS index and popularity of words. We can see that the words "student", "they", "we", "it", "course", etc. are located close to the line for FS index=0, i.e., they are used both in full and short papers in a similar ratio even if their popularity is higher than other words.

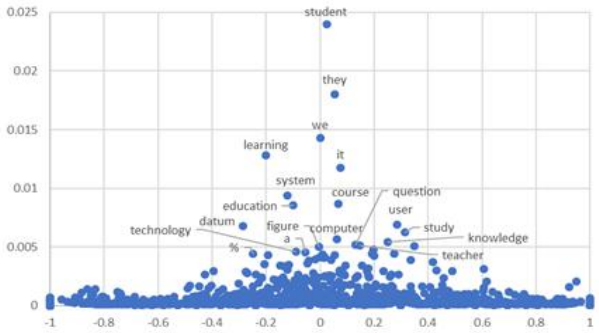

Fig. 3. Position of words by FS index (x-axis) and popularity (y-axis).

In Fig. 3, we see many noun words that appear in Table II locate in the area in the middle popularity area and in the range from 0.1 to 0.4 in their absolute values FS index. For example, for example, the words "course", "user", and "study" appear in the right area, i.e., where $FS(w)>0$, in Fig. 3. Their ranking orders in full papers are 9, 11, 14, respectively in Table II.

Similarly, the words "learning", "system", "education", and "datum" appear in the left area. Their ranking orders in Table II are 4, 10, 11, and 12, respectively.

A characteristic word needs to satisfy the following two conditions:

1) It has high absolute values in FS index because it shows how much the word is used specifically in full or short papers.
2) It has a high popularity value as well, because if the popularity is very small, it means that the word appears only in a small number of papers, and thus, it is highly possible that the word happens to appear in one or some small number of papers.

D. Use of Mixed Index for Measuring the Amount of Characteristic Feature

In order to find the characteristic words that discriminate the full and short papers more appropriately, we propose an index by mixing up the two indexes of FS index and popularity in this section. Firstly, we consider the necessary properties of the mixing up function, and then we take an example function and investigate what words we can obtain from the target data.

1) Defining an index of word for measuring the amount of discrimination between full and short papers

Before defining the specific mixing-up function, we would like to investigate the conditions required to a mixing-up function. Such an indexing function $\mu(i, p)$ for mixing-up a FS index value i and a popularity p should satisfy the following properties.

(i) $\text{sgn}(\mu(i,p)) = sgn(i)$, i.e., $\mu(i,p) > 0$ *iff* $i > 0$, $\mu(i,p) < 0$ *iff* $i < 0$, *and* $\mu(i,p) = 0$ *iff* $i = 0$. Where, $sgn(x) = 1$ if $x > 0$, $sgn(x) = 0$ if $x = 0$, and $sgn(x) = -1$ if $x < 0$.

(ii) Strictly increasing regarding i, i.e., $\mu(i_1,p) > \mu(i_2,p)$ if $i_1 > i_2$.

(iii) Absolute value is strictly increasing regarding p, i.e., $|\mu(i,p_1)| > |\mu(i,p_2)|$ if $p_1 > p_2$.

In this paper, we define the mixing-up function by using multiplication as follows, which we call m-index: For FS index i and popularity p, $\mu(i,p) = ip$

Note that some functions in the form $\mu(i,p) = i^r p^s$ also satisfy the three conditions that are shown above; e.g., when $r = 3$ and $s = 2$.

2) Histogram of words with m-Index

Fig. 4 shows the histogram of m-index values for nouns.

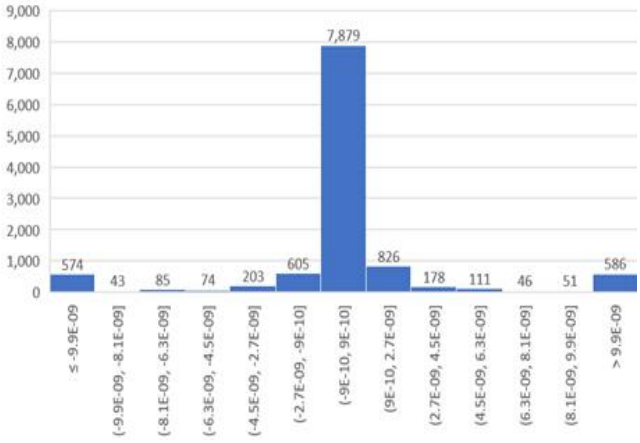

Fig. 4. Histogram for m-Index (Nouns).

Most words are located at the central part where the absolute values are small. At the same time, some 500 words are located near 1 as well as near -1. These words are considered to be the characteristic words for full and short papers, respectively.

3) Highly characteristic words

The words having big absolute values in their m-index discriminate the full papers and short papers more than other words. Among them, the words having positive m-index values are those that characteristically appear in full papers, which we call F-words, and we call S-words for those having negative values.

Table III shows the list of top 30 F-words and S-words. According to the words in the list, here again, we find that F-words contain generic words concerning study and lectures of university students, such as, together with ranking, "student/2", "user/3", "study/4", "knowledge/7", "course/11", "classroom/14", "teacher/16", "participant/17", "question/20", "response/22", etc.

Different from these words, the words "programming/8", "design/9", "application/10", "computer/24" seem to be related to programming. The word "problem/12" and "response/22" may be related to programming also.

On the other hand, S-words contain the words referring educational systems and experimental results, such as

"datum/2", "system/3", "education/4", "result/6", "game/7", "video/8", "environment/9", "technology/12", "experiment/18", "evaluation/22", etc. Roughly speaking, most of these words relate to educational systems and experiments in educational situations.

TABLE III: TOP 30 F-WORDS AND S-WORDS WITH REGARD TO M-INDEX

No.	F-Word (noun)	m-Index	F-S	Pop
1	they	3.57E-05	0.00198	0.01801
2	student	2.76E-05	0.00114	0.02401
3	user	2.70E-05	0.00393	0.00687
4	study	2.45E-05	0.00392	0.00624
5	it	2.10E-05	0.00178	0.01174
6	group	1.78E-05	0.00353	0.00506
7	knowledge	1.45E-05	0.00270	0.00539
8	programming	1.15E-05	0.00375	0.00308
9	design	1.14E-05	0.00308	0.00369
10	application	1.07E-05	0.00242	0.00443
11	course	1.04E-05	0.00119	0.00869
12	problem	1.02E-05	0.00261	0.00390
13	information	8.96E-06	0.00188	0.00476
14	classroom	8.42E-06	0.00287	0.00293
15	verb	8.03E-06	0.00390	0.00205
16	teacher	7.82E-06	0.00152	0.00513
17	participant	7.81E-06	0.00260	0.00300
18	activity	7.42E-06	0.00169	0.00439
19	level	7.24E-06	0.00170	0.00426
20	question	7.06E-06	0.00136	0.00517
21	objective	5.10E-06	0.00250	0.00203
22	response	4.96E-06	0.00213	0.00232
23	dashboard	4.72E-06	0.00295	0.00160
24	computer	4.10E-06	0.00072	0.00568
25	class	4.01E-06	0.00152	0.00263
26	science	4.01E-06	0.00137	0.00291
27	experience	3.69E-06	0.00115	0.00319
28	order	3.67E-06	0.00139	0.00263
29	model	3.67E-06	0.00107	0.00343
30	feedback	3.09E-06	0.00126	0.00244

No.	S-Word (noun)	m-Index	F-S	Pop.
1	learning	-6.60E-05	-0.00515	0.01282
2	datum	-2.62E-05	-0.00387	0.00677
3	system	-2.15E-05	-0.00229	0.00939
4	education	-1.45E-05	-0.00170	0.00857
5	%	-9.68E-06	-0.00219	0.00440
6	result	-7.16E-06	-0.00166	0.00430
7	game	-6.85E-06	-0.00232	0.00294
8	video	-5.91E-06	-0.00224	0.00263
9	environment	-5.10E-06	-0.00144	0.00352
10	c	-4.63E-06	-0.00159	0.00289
11	number	-4.10E-06	-0.00148	0.00275
12	technology	-3.70E-06	-0.00081	0.00456
13	research	-3.47E-06	-0.00101	0.00343
14	ict	-3.29E-06	-0.00238	0.00138
15	university	-3.29E-06	-0.00124	0.00263
16	style	-3.17E-06	-0.00229	0.00138
17	teaching	-3.04E-06	-0.00090	0.00335
18	experiment	-2.83E-06	-0.00165	0.00171
19	school	-2.77E-06	-0.00080	0.00345
20	interaction	-2.51E-06	-0.00104	0.00240
21	management	-2.41E-06	-0.00146	0.00164
22	evaluation	-2.39E-06	-0.00101	0.00235
23	score	-2.34E-06	-0.00137	0.00170
24	a	-2.28E-06	-0.00050	0.00455
25	s	-2.26E-06	-0.00074	0.00306
26	mining	-2.10E-06	-0.00177	0.00118
27	engineering	-2.09E-06	-0.00143	0.00145
28	indicator	-2.08E-06	-0.00112	0.00185
29	discipline	-2.04E-06	-0.00189	0.00107
30	type	-1.72E-06	-0.00088	0.00194

These words might reflect that short papers pay more attentions to specific systems and experiments than full papers as we have recognized in the previous sections.

We also recognize the words such as "research/13", "ICT/14", "university/15", "teaching/17", "school/19", "interaction/20", "evaluation/22", "score/23", "discipline/29" appear in Table III. These words might indicate that short papers are not just focusing specific topics about systems and something, they also concern the backgrounds about education and educational environments just like the full papers.

According to these findings, we may summarize that short papers deal with more about specific systems and experiments and their results than full papers, and full papers discuss more about philosophical aspects, models, and other topics in a more theoretical basis. We have to investigate further by deeper analysis for more precise findings and conclusions.

IV. CONCLUDING REMARKS

In this paper, we investigated the different features of full papers and short papers based on an analysis of usage of words, or terms. Our tentative conclusion is that short papers rather pay more attention to specific experiments and their results than full papers, whereas full papers pay more concerns educational discussions from general points of view than short papers.

Our objectives include not only finding differences and extracting valuable tips for improving the quality of papers, but also finding effective analysis methods. In this paper, we proposed indexes for measuring the differences of the features between full and short papers in terms of word-usage of papers.

Our target data for analysis is rather small data [7] than big data because we aim to find domain-specific tips rather than to find generic knowledge that are applicable to a wide variety of domains. Thus, we intend to pursue investigation more deeply on tools for analysis and useful findings in education according to our study approach.

The analysis methods we carried out in this paper is just the very first stage toward our goal in order to find more effective and practical tips for writing more sophisticated papers. We understand that it is quite a difficult problem to find tips for writing better papers only from the word usage.

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

1) In this paper, we used Fr and Sr for weighting importance of words. We may use other measures such as TF-IDF [8] instead of Fr/Sr, which is a candidate for our future studies. We used FS(w) for classifying words w in this paper. For this purpose we may use SVM [9], [10], which is also our candidate for future studies.
2) 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.
3) In this paper, we took all the words without thinking about parts of speech (POS) firstly, and then we took nouns only for specific analysis. Usage of words with

other parts of speech should be different in full and short papers. We need to analyze further by considering other types of POS.

4) Organizational analysis is another important topic for investigation. Organizations should be different between full and short papers. Analysis on organizational differences such as structure of contents, layout of figures and tables, use of mathematical formulas, are other important topics toward our eventual goal.

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] T. Minami and Y. Ohura, "An analysis on differences of word usage between full and short conference papers," in *Proc. the 9th International Conference on Intelligent Computing and Applications (ICICA 2020)*, 2020.

[2] T. Minami, Y. Ohura, and K. Baba, "To estimate student's viewpoint to learning from lecture/self-evaluation texts," in *Proc. the Ninth International Conference on Advanced Cognitive Technologies and Applications*, Athens, Greece, February 19-23, 2017.

[3] T. Minami, Y. Ohura, and K. Baba, "A characterization of student's viewpoint to learning and its application to learning assistance framework," in *Proc. the 9th International Conference on Computer Supported Education*, Porto, Portugal, April 21-23, 2017, vol. 1.

[4] T. Nakatoh, H. Nakanishi, T. Minami *et al*., "Bibliometric search with focused citation ratios," in *Proc. 5th International Congress on Advanced Applied Informatics*, Kumamoto, Japan, July 10-14, 2016.

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

[6] K. Higuchi. KH Coder Index Page. Available: http://khc.sourceforge.net/en/

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

[8] J. Beel, B. Gipp, S. Langer, and C. Breitinger, "Research-paper recommender systems: a literature survey," *International Journal on Digital Libraries*, vol. 17, no. 4, pp. 305-338, 2016.

[9] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273-297, 1995.

[10] Scikit Learn: "1.4. Support Vector Machines". [Online]. Available: https://scikit-learn.org/stable/modules/svm.html

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

He was a researcher of Fujitsu Limited and Fujitsu Laboratories Limited, Japan from 1984 to 1999, a research fellow of Australian National University from 1992 to 1993, 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 emeritus professor since then.

His research interests include analytics of the data obtained from educational, library, and bibliometric fields, as well as library informatics, library marketing, and multi-agent systems

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

She was a research assistant of Fukuoka University from 1978 to 1993, an associate professor of Aso Fukuoka Junior College from 1993 to 1998, and 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.