New Perspectives in Science Education
International Conference
NEW PERSPECTIVES in SCIENCE EDUCATION

From Text to Insight: Leveraging Free-Text Responses to Inform University Educational Improvements

Mio Tsubakimoto[1], Takashi Nishide[2], Sachio Hirokawa[3], Tetsuya Oishi[4], Kunihiko Takamatsu[5]

Tokyo Metropolitan University, Japan[1]
Otaru University of Commerce, Japan[2]
Advanced Institute of Industrial Technology / MONOLITHIC DESIGN Co., Ltd., Japan[3]
Kyushu Institute of Technology, Japan[4]
Institute of Science Tokyo, Japan[5]

Abstract

Some universities collect free-text answers in student surveys. Despite the valuable information in the texts, they are not sufficiently analyzed and utilized to improve university education due to the difficulty of directly linking uniform visualization expressions to educational improvement measures. Our study will investigate "how the texts of student surveys can be analyzed and visualized to extract information that contributes to the improvement of university education." We will develop new visualization expressions and analyze methods that can facilitate discussions among university constituents. Moreover, in student surveys, free-text questions are often used in addition to multiple-choice questions, and text data analysis has become important. Various methods are used to analyze text data, but in recent years, analysis methods using machine learning, such as topic models and large-scale language models (LLMs), have made rapid progress, and the findings obtained from text data have improved dramatically. Conversely, it is often difficult to understand the real voices of students' feelings with only a summary or quantified and visualized information obtained as a result of these analyses, so careful reading of texts is also essential. However, it is not realistic to read and comprehend all of the large number of texts. Therefore, we consider using a randomly sampled subset of text data as an intermediate analysis method between careful reading of the entire text and summarization or quantification. Suppose the similarity of semantic content between the subset and the whole is high. Reading it in real-time may allow us to get a general overview of students' opinions with a limited amount of text, even though we may miss individual opinions with few mentions. Here, we explore this possibility by quantitatively analyzing the representativeness and accuracy of textual data through this sampling. When humans generate hypotheses, bias may occur due to stereotypes when analyzing issues. However, if hypotheses can be generated mechanically in large quantities as a result of this study, it is expected that multifaceted analysis will be possible without being influenced by stereotypes.

Keywords: *Text Analysis in Education, Data-Driven Educational Insights, Text Data Sampling Methods, University Education Improvement*

1. Background and Objectives

1.1 Problem Statement

Free-response data collected through student surveys have the potential to play a crucial role in improving and quality-assuring university education. However, utilizing such small-scale text data presents fundamental challenges [1][2]. First, text data contain diverse student opinions, emotions, experiences, and topics, making it difficult to analyze using conventional hypothesis-testing approaches (Problem 1) [3][4][5]. Student voices are multifaceted and complex, possessing a richness that cannot be captured by a single perspective or existing framework.

The second challenge lies in the difficulty of using current text mining methods and visualization techniques to provide information that promotes constructive dialogue and practical discussion for specific educational improvements and quality assurance (Problem 2). This is better understood as a technical limitation rather than a gap between the presentation of analysis results and their practical application.

1.2 Research Objectives and Significance

This study addresses the aforementioned challenges from two perspectives. The first is the development of analysis and visualization methods for text data obtained from student surveys. Here,

we specifically aim to establish methods for extracting and effectively presenting information that directly contributes to university education. This approach goes beyond mere technical method development, seeking to explore ways of providing truly useful information for education practitioners and decision-makers.

The second approach examines the possibility of using sampling methods to efficiently grasp the overall picture while preserving the semantic content of text data as much as possible. This effort aims to bridge quantitative analysis and qualitative understanding, particularly considering practical applications in settings dealing with large volumes of text data.

These research objectives are positioned within the broader context of evidence-based decision-making support for educational improvement. The results of this study are expected to provide IR (Institutional Research) practitioners and education improvement professionals with both concrete methodologies and new perspectives.

2. Theoretical Framework for Information Visualization

2.1 Purposes and Directions of Information Visualization

In information visualization practice, purposes can be understood from three perspectives. The first aspect is information exploration and monitoring, which provide a foundation for continuous data observation and understanding. By grasping data trends and identifying important changes and characteristics, we can find starting points for deeper analysis.

The second aspect, data analysis, is the process of gaining deeper insights. The goal is to discover patterns and relationships inherent in the data and extract meaningful findings. Visualization plays a crucial role in this process, helping to understand data characteristics intuitively and generate new hypotheses.

The third aspect, presentation, is critical to sharing findings with others and promoting understanding. Effective visualization acts as a catalyst that communicates complex data and analysis results clearly and promotes constructive discussion.

For these purposes, two fundamental approaches exist in data analysis: exploratory and explanatory thinking. The exploratory thinking approach emphasizes discovering unknown patterns and relationships inherent in the data. In contrast, the explanatory thinking approach aims to clearly explain specific aspects of the data based on particular viewpoints or hypotheses.

2.2 Organization of "Information Visualization" in IR for Education

Human behavior and social phenomena collected through digital data include, for example, social networking services, written content of digital documents, computer system logs, open data from companies and local governments, and records of human behavior by various sensors. These data are also considered to be subject to analysis and visualization in educational IR, and the method is collectively called "information visualization (InfoVis)" [6].

"Information visualization" is a term that combines "information" and "visualization" but here we would like to clarify the definitions of "information" and "visualization." First, "information" is sometimes used synonymously with the term "data," which also appears in this paper, but some consider them to be separate terms. In this study, "data" is defined as a set of qualitative and quantitative facts about a phenomenon, based on previous studies [7]. Furthermore, it is assumed that "information" that is meaningful to data users can be obtained by processing, analyzing, and visualizing data. In other words, "information" can only be obtained through appropriate analysis and visualization of "data." Next, "visualization" is defined in this context as a method for obtaining information from data. Generally, aesthetic aspects of graphs, including the effective and visually appealing representation of data, are considered and organized under visualization. However, in this study, these elements are positioned as corresponding to the scope of C2 "Explanation is the purpose" in Table 1. While it is important to prepare visualization results that can be used in C2, the "visualization" targeted in this study is a method to support the process in which the analyst interacts with the data, explores the data, determines the axis of analysis, and elicits the interests and arguments of those with whom one interacts (stakeholders within the university) through the teaching and learning data. It is a method to support the process of eliciting the interests and discussion point of the people with whom one interacts (stakeholders within the university) through educational data.

Table 1. Purpose and Direction of Information Visualization and Utilization of Student Survey Text Data in Academic IR

		Objectives	
	(A) Monitoring Conveying Information from Machines to Humans (e.g., clearly presenting machine status or displaying database search results to users) / Monitoring dynamic data to detect anomalies or changes	**(B) Data Analysis** **Verification-Oriented:** Hypothesis-driven approach where a pre-existing hypothesis is tested for validity through analysis. Statistical methods are effective. **Exploration-Oriented:** Hypothesis-free approach aimed at discovering hypotheses or knowledge (e.g., patterns or relationships in data)	**(C) Presentation** Communication that allows non-expert audiences to understand data and information without additional explanation. The sender selects efficient and effective media to convey the data or information and employs simple visual representations wherever possible.
Directions **(1) Exploratory-Oriented** * Visualization aimed at discovering data characteristics, patterns, and mechanisms * Visualization to help analysts deepen their understanding during the data analysis process * Capable of displaying a large amount of information and patterns	Not Applicable (N/A)	**Extensive Exploration and Experimentation:** * Investigating features within text * Combining text with other variables (e.g., GPA, course satisfaction) * Generating numerous hypotheses	**Explanation as a Means:** * Sharing progress and engaging in discussions with IR professionals or related internal stakeholders * Facilitating analysts' dialogue with themselves and the data to generate ideas and deepen understanding
(2) Explanation-Oriented * Visualization used to report and present analysis results (e.g., in academic papers, reports, general media coverage, press releases, or briefings for decision-makers) * Unnecessary information is removed to clearly emphasize the key features of the data	**Automatable Analysis and Visualization:** * Conducting regular analyses and reports * Confirming the absence of changes or the presence of consistent patterns * Ensuring reliable detection of anomalies in critical situations	**Focus on Specific Hypotheses, Variables, and Visualizations:** * Analysis based on predetermined features, variables, and hypotheses * Regularly verify and report on standardized content * Often results in "analysis for the sake of analysis"	**Explanation as the Primary Goal:** * Explaining to executives or board members * Publishing in reports or on websites

Note: The descriptions of the three objectives are based on [8], and the descriptions of the two orientations are based on [9].

The objectives of information visualization (or "data visualization" from the standpoint of this study) can be organized into three categories: (A) information search/monitoring, (B) data analysis, and (C) presentation [8]. There are two types of analysis: (1) exploratory data analysis and (2) explanatory data analysis [9]. Table 1 shows these objectives and directions as well as their use in IR for teaching and learning. The column directions (A-C) in Table 1 indicate the purposes of visualization, and the rows and columns (1-2) indicate the directions of visualization. The six cells where the rows and columns intersect (A1 to C2) show specific examples of the analysis, visualization, and presentation of text data from student surveys in IR for teaching and learning.

Cells B1, "Massive search and trial (thinking)," and C1, "Explanation as a means," represent actions we aim to promote among IR staff and other members of the university through the findings of this study. First, it is difficult to apply a hypothesis-testing approach to the analysis of text data in B1, "Massive search and trial (thinking)." Therefore, both qualitative and quantitative analyses involve coding from textual content and classification using relevant variables. However, since hypotheses are essential to this process, it is important to closely read and analyze the data through trial and error to develop the "hypotheses for generating hypotheses." Although it is possible to use qualitative text analysis methodologies (e.g.,[10]) for this task, there are problems with time-consuming analyses and subjective bias. For example, if there is a system that can perform an exploratory analysis of texts by quickly switching between various angles (variables), as described in [11], it would be possible to find a coding or classification axis by dividing and comparing texts using both extrinsic and intrinsic axes. This solves Problem 1. To address Problem 2, C1, "Explanation as a means" can promote cooperative discussion in which explanation is a means rather than an end, through the realization of B1, "mass search and trial (thinking).

2.3 What is Information Visualization of Text Data that Contributes to the Improvement of University Education?

In future, we will develop data analysis and visualization methods to solve Problems 1 and 2 and develop a system that implements the proposed methods. For example, there is a system that displays the appearance of words when cross-tabulating variables associated with text data to be analyzed (e.g., grade, year of admission, GPA, and responses to other items in the questionnaire). It

allows users to quickly switch the variables to be crossed while checking the results (Figure 2) [11]. Utilizing such a function for exploratory analysis of text data may support the flexible modification of hypotheses and “deeper digging.”

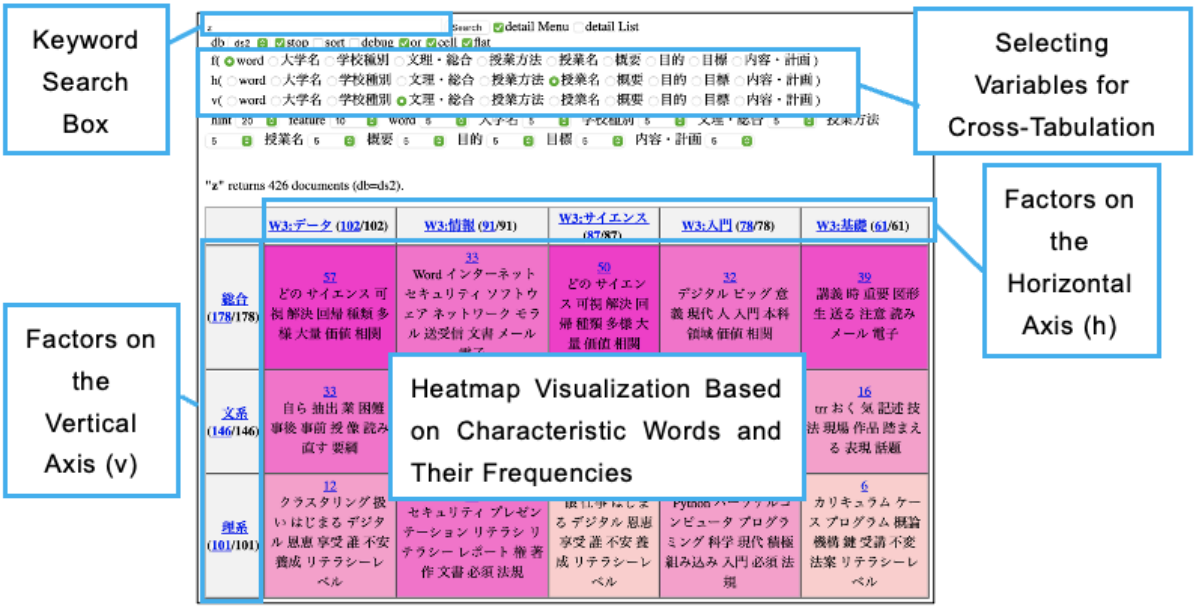

Fig. 1. Cross tabulation system

3. Verification of Representativeness through Sampling

Universities conduct student surveys at various time points, including new students and graduating seniors. These questionnaires use a response format in which the respondents select from a list of pre-prepared options. However, open-ended items are also often used to obtain students' candid opinions, which are difficult to capture in choice-type surveys. In contrast to multiple-choice items, which are generally more familiar with quantitative analysis and can ensure reproducibility of analysis to some extent, the analysis of open-ended items, in which response data are obtained in writing, requires judgment and interpretation of the written responses. Thus, objectivity, reproducibility, and the effort required for analysis may be problematic.

In recent years, with the availability of digital data in the form of response texts and the development of text mining, it has become possible to conduct analyses while ensuring objectivity and reproducibility to some extent, such as by detecting word frequencies and co-occurrence relationships, detecting topics through machine learning, and classifying documents [12]. Furthermore, the rapid development of large-scale language models (LLMs) has enabled the highly accurate summarization and classification of text data, greatly expanding the potential use of texts, such as open-ended responses to questionnaires and descriptive answers [13].

3.1 Research Methodology

This study attempts to verify the representativeness of the sample using free-response data from graduation surveys at University A. The analysis covered 2,007 descriptive responses, for which we conducted staged sampling and evaluation.

The research procedure began by sampling at different extraction rates. Specifically, we set the extraction rates starting from 1%, 5%, and 7.5%, and then from 10% to 90% in 10% increments, executing 1,000 simulations of sampling without replacement for each condition. This allowed us to observe in detail how representativeness changes with different sample sizes.

For text data analysis, we adopted the Bag of Words (BoW) model, converting each text into a vector representation. This process enabled the quantitative capture of word occurrence patterns in the texts. Furthermore, to evaluate the similarity between the entire text dataset and the subsets obtained through sampling, we conducted comparisons using cosine similarity.

3.2 Analysis Results

The analysis provided several important insights into the representativeness of the text data sampling (Fig.2). First, the analysis based on word frequency (TF/BM25) showed a high similarity to the whole, even at relatively low extraction rates. Notably, even at a 10% extraction rate, the cosine similarity exhibited high values of approximately 0.8. This result suggests that the major content characteristics of the text data can be captured using relatively small samples.

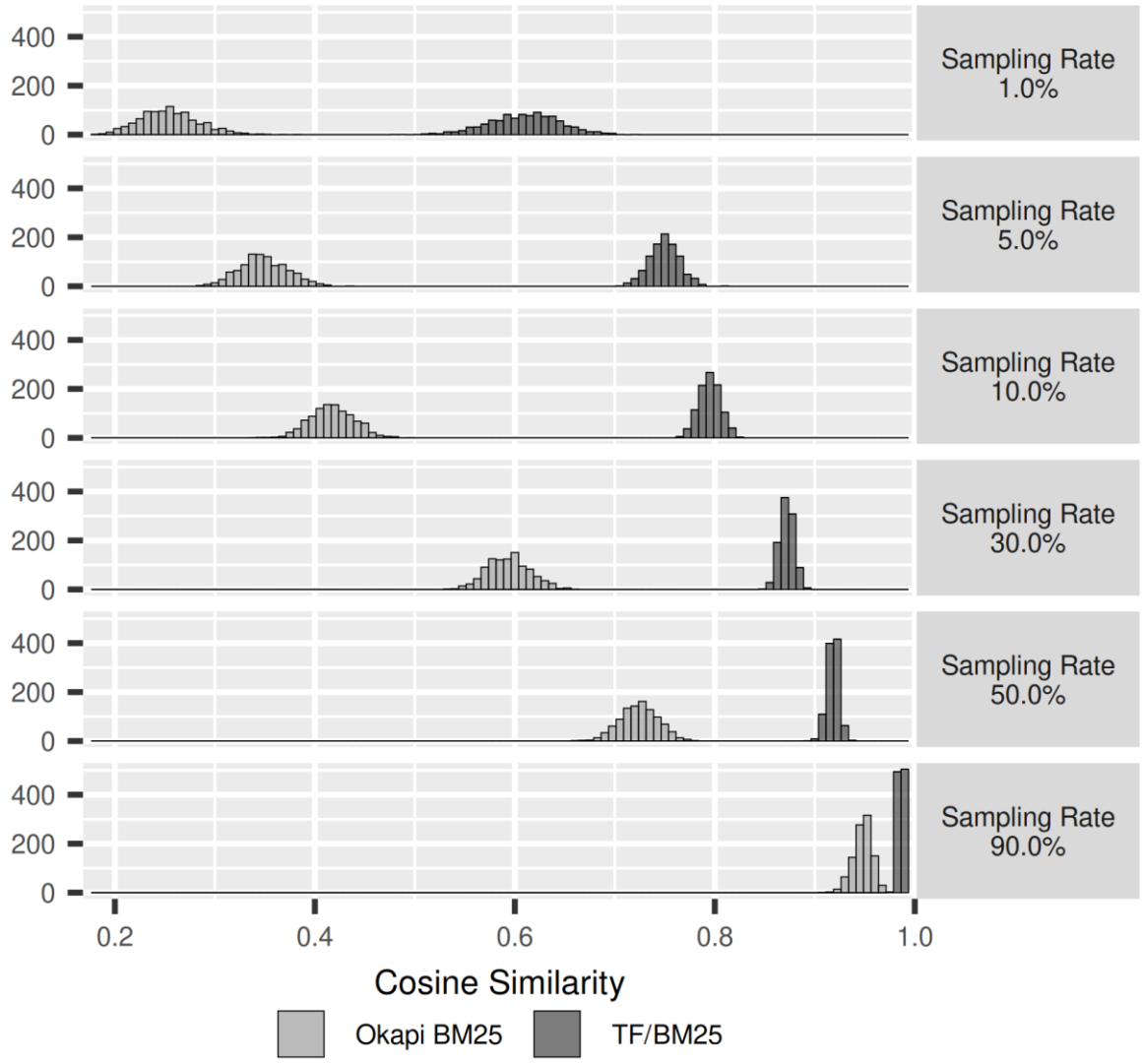

Fig.2. Similarity to the total by extraction rate (partial)

Meanwhile, the analysis based on word importance (BM25) revealed different aspects. This indicator showed relatively low overall similarity, with particularly notable differences at low extraction rates. This suggests that the distribution of characteristic expressions and words with important meanings in the text data is more sensitive to sampling.

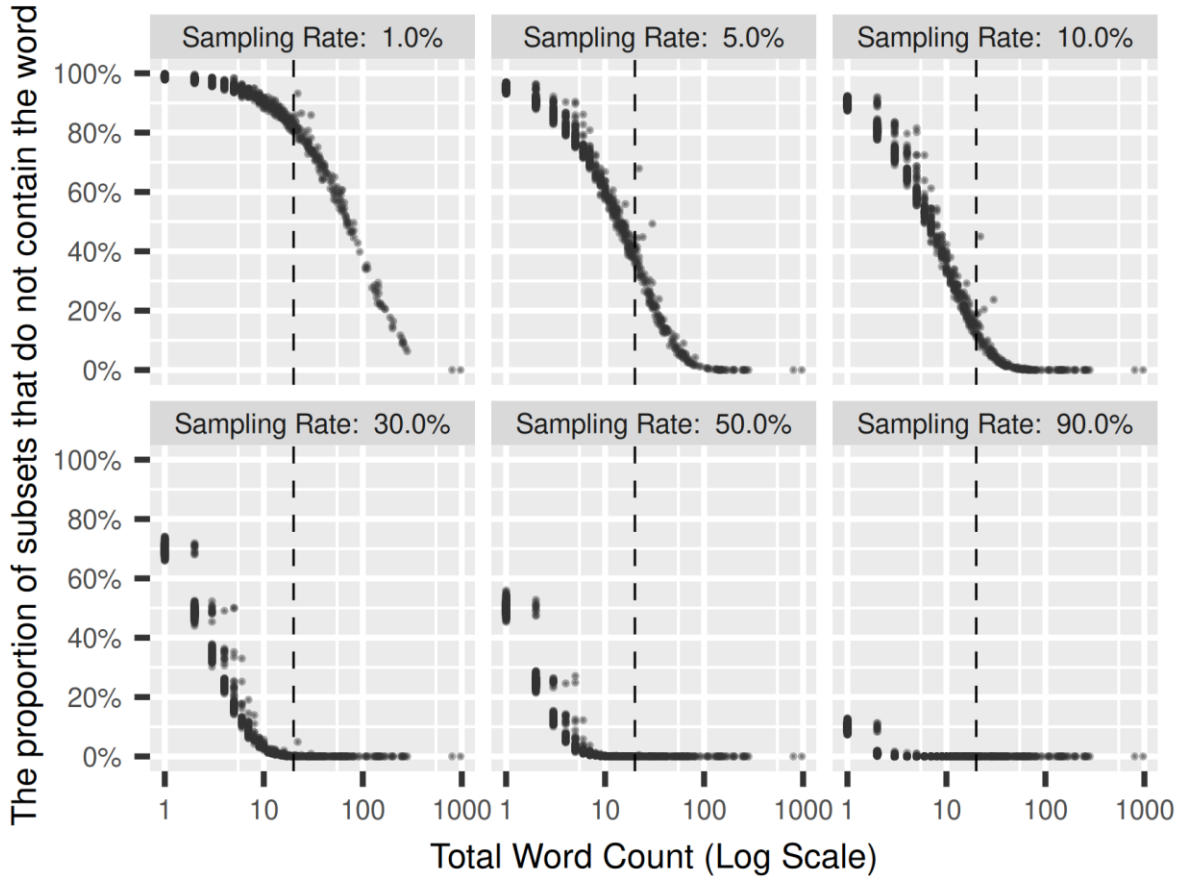

Fig.3. The proportion of subsets that do not contain the word

Furthermore, frequency-focused analysis revealed an interesting threshold phenomenon at an extraction rate of 30 %. Specifically, we confirmed that words appearing 20 or more times were included in almost all trials at this extraction rate. This finding indicates that topics and opinions mentioned at a certain frequency can be stably understood even with relatively small samples.

4. Discussion and Future Prospects

4.1 Research Significance

The findings of this study have several important implications for text data analysis in institutional research. First, the effectiveness of the proposed sampling approach as an efficient text-data analysis method was empirically demonstrated. This method can serve as a useful option for balancing analysis efficiency and accuracy, particularly in settings involving large volumes of text data.

Furthermore, the quantitative verification of representativeness through sampling represents a methodological advancement in text data analysis. In particular, the elucidation of similarity-change patterns at different extraction rates provides important guidelines for designing practical analytical procedures.

The findings of this study also suggest practical applicability in institutional research. Particularly in settings where analyses must be conducted with limited resources, the methods demonstrated in this study can serve as realistic solutions.

4.2 Future Research Challenges

Based on the findings of this study, several research directions warrant further development. First, the development of integrated approaches with rapidly evolving LLMs and other text mining methods. The

combination of these advanced technologies with the proposed sampling method is expected to enable the construction of more effective analysis frameworks.

Additionally, further refinement is needed in document similarity evaluation methods. Beyond currently used evaluation indicators, new indicators capable of capturing semantic similarity of texts more accurately is required. Particularly, establishing evaluation methods that consider subtle differences in context and meaning represents an important research challenge.

Furthermore, verification of consistency with intuitive human understanding is necessary. Clarifying the relationship between results indicated by quantitative similarity measures and similarity perceived by analysts and education practitioners is expected to lead to the establishment of more practical methods.

4.3 Visualization of IR Data and the Enhancement of Science Education

Another critical aspect of the research challenges addressed in this study is the advancement of science education. Science education often involves practical learning activities such as laboratory experiments, exercises, and fieldwork, which can lead to the fragmentation of student learning behavior data. Moreover, assessing student performance in science education requires consideration of diverse learning outcomes, including laboratory reports, project results, and collaborative learning achievements, rather than relying solely on examination scores. By employing visualization techniques for Institutional Research (IR) data, it becomes possible to facilitate sharing these diverse learning outcomes among faculty members, thereby supporting data-driven decision-making for educational improvement.

Furthermore, science education encompasses multiple interrelated disciplines, including mathematics, physics, chemistry, and biology. Using IR data visualization to analyze and represent the relationships between these subjects can contribute to the design of more effective curricula. Additionally, by actively sharing visualization results with learning support centers and other relevant institutional bodies, universities can establish a more comprehensive and systematic approach to supporting student learning in science education.

5. Conclusion

This study empirically examines the possibility of using a sampling-based approach as an analysis and visualization method for free-response text data from student surveys. The results indicate that sufficient representativeness can be secured, even at relatively low extraction rates, for purposes such as understanding major opinions and overall trends. These findings contribute to improving the efficiency of practical text data analysis in institutional research.

However, it also became clear that careful consideration is necessary when dealing with characteristic minority opinions. This indicates the limitations of the sampling approach, while suggesting the necessity of comprehensive approaches combining multiple analysis methods.

Considering new technological trends, we aim to develop more integrated approaches and establish practical utilization methods. Further research is expected regarding ways to provide information that can lead to concrete improvements in educational settings.

Acknowledgments

This research was supported by JSPS KAKENHI (Grant Number 24K00452). We thank all those who participated in this study.

REFERENCES

[1] Masayuki Murakami, Yu Urata, Toshiaki Nagaoka, “A Questionnaire Survey on Educational Use of Generative AI in Osaka University,” Proceedings of the 49th National Conference of Japanese Society for Information and Systems in Education, pp. 81-82, 2024. [in Japanese]
[2] Naoki Otawa, “How Can a Student Survey be Utilized to Promote University Reform?,” Japanese journal of higher education research, 19, p. 87-106, 2016. [in Japanese]

[3] Hideya Matsukawa, Makiko Oyama, Chiharu Negishi, Yoshiko Arai, Chiaki Iwasaki, and Hiroshi Horita, “Analysis of Free Descriptions in a Class Evaluation Questionnaire Using Topic Models,” Japan journal of educational technology, 41(3), pp. 233-244, 2017. [in Japanese]
[4] Keita Nishiyama, “Student Evaluation for Online Classes Based on Free-Text Questionnaires,” Proceedings of the 10th Meeting on Japanese Institutional Research, 10, pp. 14-19, 2021. [in Japanese]
[5] Mio Tsubakimoto, “Text Mining of Reasons for “Text Mining of Reasons for Overall Satisfaction with the University in a Graduation Survey,” Proceedings of the 49th National Conference of Japanese Society for Information and Systems in Education, pp. 77-78, 2024. [in Japanese]
[6] Yuriko Takeshima, Takayuki Ito, Hideo Miyachi, and Satoru Tanaka, “Visualization for Science, Digital Humanities, and Sociology,” Media Technologies 3, CORONA PUBLISHING Co., Ltd., 2023. [in Japanese]
[7] Takashi Nishide and Tetsuya Oishi, “How to Present “Data” and the Value of “Information” in IR: Differences in the Amount of Information Depending on the Purpose of Data Use and the Way of Analysis, Visualization, and Reporting,” Proceedings of the 40th Annual Meeting of the Japan Society of Educational Information, pp. 54-57, 2024. [in Japanese]
[8] Kazuo Misue, “Introduction to Information Visualization - Human Vision and Data Representation Method,” Morikita Publishing, 2021. [in Japanese]
[9] Takahiro Ezaki, “An Introduction to Data Visualization Science, Beginning with the Design of Indicators and Features — Techniques for Turning Data into Insights,” Socym Co,.Ltd., 2023. [in Japanese]
[10] Takashi Ohtani, “SCAT A Qualitative Data Analysis Method by Four-Step Coding: Easy Startable and Small Scale Data-Applicable Process of Theorization,” Bulletin of the Graduate School of Education and Human Development. Educational Sciences, 54(2), pp. 27-44, 2008. [in Japanese]
[11] Makoto Okada, Sachio Hirokawa, and Kiyota Hashimoto, “An Investigation of Efficiency of a Method of Data Mining and Visualization: Using Questionnaire Survey Results for New Farming Applicants,” IEICE technical report, 113(65), pp. 27-31, 2013. [in Japanese]
[12] Atsushi Kawatsuma, Tomoki Koga, Yu Mizoguchi, and Hideo Narita, “On Students' Learning and Growth in Inter-University Cross-Border Learning,” Proceedings of the 46th Annual Meeting of Japan Association for College and University Education, 46, pp. 129-130, 2024. [in Japanese]
[13] Jun Saito, “The Problem and Possibility of Educational Improvement by Using Generative AI: On the Subject of Automatic Grading of Writing,” Proceedings of the 46th Annual Meeting of Japan Association for College and University Education, 46, pp.263-264, 2024. [in Japanese]