Text Mining of University Midterm Reports for Comparison of Educational Strength

Kahori Ogashiwa[1], Masao Mori[2], Sachio Hirokawa[3]

Abstract

We are studying to find distinction among Japanese national universities, and strength of each university by analyzing their midterm goals and plans. In this paper, we propose a system that extracts keywords from documents, analyze and visualize features of the documents based on the extracted keywords. In our study, we adopt the midterm goals and plans of Japanese national universities as the object documents. We regard the features from the keyword analysis from those documents as the features of universities. In this paper, we report the experimental analysis focused on "regional" contribution and "international" contribution by university, which are important factors of social relevance.

We collected the documents of midterm goals, extracted the segments on education and constructed a "mind map" search engine and "cross tabulation" search engine. The mind map displays the co-occurrence relation of words that appear in the search result of a query. The cross tabulation shows a table whose axis is either names of universities, keywords in the search results or the focused keywords such as "regional", "local", "global", "foreign", "English" etc. We used the two search engine to compare 7 universities that contain the focused keyword "region".

1. Introduction

As a result of the establishment of the National University Corporation Law on April 1, 2004, national universities became incorporated and obliged to undergo national university corporate evaluations (hereinafter referred to as "corporate evaluation"). The purpose of corporate evaluation is to continuously improve the quality of the universities, their accountability to society, and to calculate their subsidies for operating expenses. The calculation of subsidies for operating expenses, the acquisition of subsidies for competitive funds, and the strengthening of functions are closely related; thus, corporate evaluation tends to be taken very seriously by universities. However, the original purpose of evaluation is to contribute to the extension of the individuality of each university. It is important for universities to know their strengths and to develop and extend them.

In order to identify the strengths of each university (including the brand or image of the university and the accomplishments of its management), correspondence to the ex-post evaluation (preparation of the report and the presentation of the basis data) is conducted to understand its own characteristics. In addition, it is essential to conduct heuristic institutional research (IR) (a new idea or discovery based on data with support from information technology). Various studies are being conducted to address these aspects. Takada researched the progress management system of the midterm goals of universities and their midterm plans [1]. Fujii studied the types and characteristics of indicators included in the midterm plans [2]. Sun conducted research on university personality and indicators using fundamental numerical data from university portraits [3]. Funamori conducted a comparative analysis of universities using data from the School Basic Survey [4]. Mori conducted research on a system capable of benchmarking comparative analysis between universities in real time using data from the School Basic Survey analyzed by Funamori [5]. Recently, research on the analysis of universities using Twitter text data has become a trend [6] [7].

However, these studies are analyses from a limited viewpoint, and do not analyze the strengths of the universities as organizations. As far as the authors know, at present, no method of analyzing the strengths of universities as organizations has been established. Therefore, in the present paper, we propose a system to visualize the comparative analysis results from different universities to identify their strengths using the text mining method.

[1] Utsunomiya University, Japan
[2] Tokyo Institute of Technology, Japan
[3] Kyushu University, Japan

2. Development of a system to analyze and visualize the characteristics of universities and identify their strengths

This study aims to identify the strengths of universities concerning educational activities. We selected keywords from the midterm goals that described the universities' educational goals and developed a system to analyze and visualize the universities' strengths from the number of keywords and their relevance. The system procedures are outlined below.

(1) "Cross tabulation" search engine
First, enter keywords (search words) and select the items to be displayed on each axis of the cross table from among the university name (univ), the feature (related word), and the focused word (keyword). Set the focused word in advance. For example, if you enter "region" as a keyword and specify [university name] × [focused word] for each axis, university documents including the keyword "region" are extracted. The vertical axis shows the list of universities and the focused word is displayed on the axis. It is possible to display not only the focused word but also the search results of the feature word. This system visualizes the distribution of [university] × [keyword].

(2) "Mind map" search engine
Enter a keyword to display a mind map. Extract the feature words, the co-occurrence information, generating the mind map. Co-occurrence refers to a phenomenon in which, when a certain keyword appears in an arbitrary sentence, words that are likely to appear at the same time around the word are found. Further details of this system are explained in Refs. [8] and [9].

3. Data and methods

We targeted third-term midterm documents (text data) [10]. The midterm goals were roughly divided into five major items, as follows: I. Goals concerning quality improvement of university educational research. II. Goals concerning improvement and the efficiency of business operations. III. Targets for improvement of financial contents. IV. Self-inspection / evaluation of targets relating to the provision of information pertaining to the situation. V. Other important objectives relating to the operation of the business. Since this study aims to identify the strengths related to educational activities, we targeted a part of the description of "1. objectives relating to education" in "I. Goals concerning quality improvement of university educational research." Regarding educational goals, nine universities with descriptions of universities and long sentences with different descriptions of item names were excluded. For that reason, we used data from only 77 of 86 national universities.
We investigate the strengths of university's educational activities in terms of "regional contribution" and "international contribution." We analyzed the data based on the following research items: (a) A list of focus words and (b) university names. We used the following procedures:
(1) We set the focus words. In this study, words such as "region," "international," and "global" were set as focus words in order to analyze the educational activities on regional contribution and international contribution, which are important elements with social relevance. (2) We used the cross tabulation search engine and extracted the universities that contained the focused word "region" in their documents. (3) Using the mind map search engine, we extract the universities that contain "region." For the words appearing in the search result document, the top N words of the focused words were selected and if co-occurring, the branch between the two words is displayed.

4. Results and discussion

Using the cross tabulation search engine, we displayed a cross table of "university name" and "top 5 focused words" and studied universities containing only the focused word "region" (Fig. 1). First, "international," "global," "region," "world," and "innovation" were listed as the five most frequently occurring words in the data from all 77 universities. In this context, we extracted seven universities with patterns flagged only with the focused word "region" (international: none, global: none, region: yes, world: none, innovation: none). The results of the cross tabulation search engine showed that "international" and "global" were the two most frequently occurring target words. From this, it was found that the educational activities at the university had an overall emphasis on international contribution. As a result of analyzing the third-most frequently occurring keyword "region," we were

able to extract a group of universities that are considered to have strengths in regional contribution regarding educational activities.

	k:国際 (38/38)	k:グローバル (35/35)	k:地域 (30/30)	k:世界 (14/14)	k:イノベーション (9/9)
北見工業大学 (1/1)					
名古屋大学 (1/1)	1			1	
大分大学 (1/1)		1	1		
熊本大学 (1/1)		1			1
総合研究大学院大学 (1/1)					
大阪大学 (1/1)					1
北海道大学 (1/1)	1				
東京工業大学 (1/1)					
琉球大学 (1/1)	1		1		
名古屋工業大学 (1/1)		1	1		
宮崎大学 (1/1)	1	1			
弘前大学 (1/1)			1	1	
高知大学 (1/1)	1		1		
岡山大学 (1/1)	1				
九州工業大学 (1/1)		1			
鹿児島大学 (1/1)			1		
北海道教育大学 (1/1)			1		
埼玉大学 (1/1)					

Figure 1: Keyword appearance using the cross tabulation search engine. A list of university names is displayed on the left side and a list of the five most frequently occurring focused words is displayed on the right side. If a focused word was found for the university, a flag was set and the cell was highlighted in pink. This makes it possible to understand the distribution of "university × keyword" at a glance.
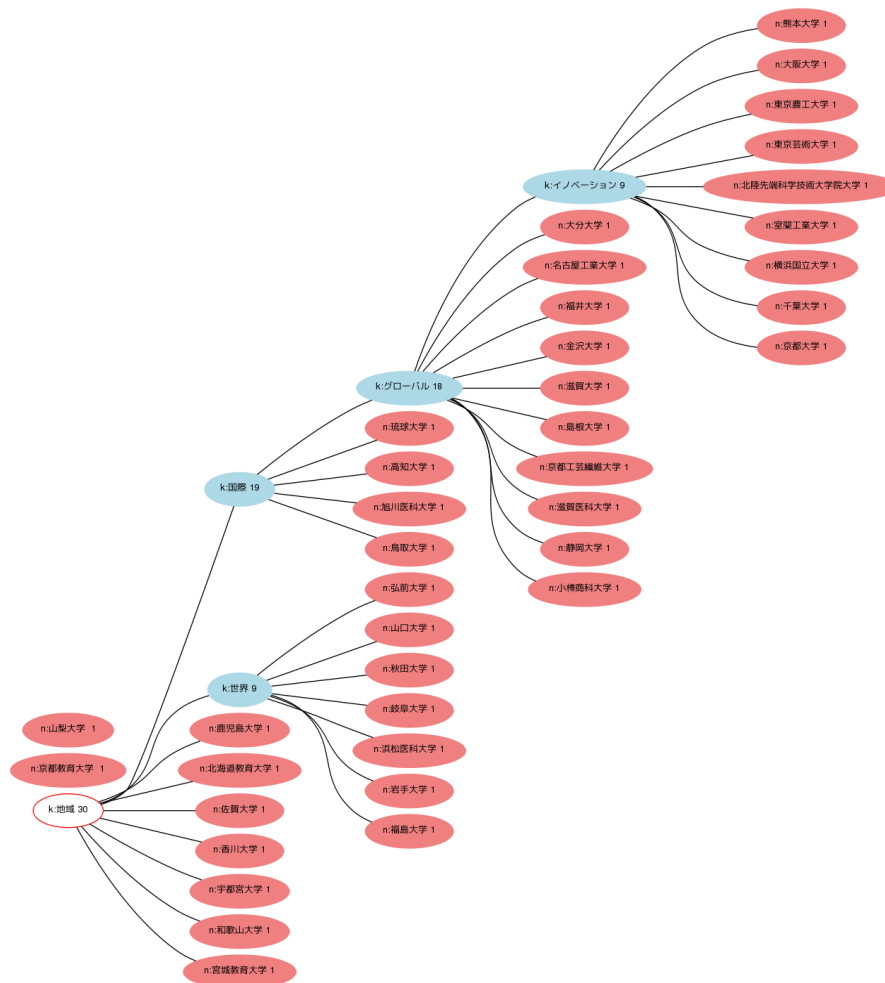

query="k:地域 k:イノベーション", 39

Figure 2: Generating maps of related keywords using the mind map search engine. "K: region" is displayed in a red circle in the center of the map on the left side. In this figure, the relationships of the top five focused words and university names are shown.

Using the mind map search engine, we mapped the relationships of the top 5 focused words to 30 universities that included "region" (Fig. 2). Branches were displayed between "Region – Region," "Region – World," "Region – International," "International – Global," and "Global – Innovation." From the results of the mind map search engine, "world" and "international" were cited as keywords related (co-occurring) to "region".

In the cross tabulation search engine, there is an advantage that it can understand the relationship between the university name and the keyword at a glance, however it does not know the relation of the keyword. In the mind map search engine, the relation of keywords is also displayed on the map, so detailed analysis can be performed.

5. Conclusion

This paper proposes a system to visualize the results of a comparative analysis of universities to identify their strengths using the text mining method. Utilizing two search engines, we were able to find a group of universities with strength in the area of regional contribution. In fact, as a result of investigating the composition of graduates and the graduates by employment area, it was confirmed by some universities that the employment rate in their prefecture is the highest [11], and thus, we were able to verify the usefulness of this system.

Since the data used in this study was obtained using only one of the educational goals of the midterm goals documents for analysis, in future research we will incorporate more sources of data, such as the midterm plan documents.

Acknowledgements

References

[1] Takata, E. "The subject about the data collection in the progress management of a mid-term plan", Proc. Of IIAI International Conference on Advanced Applied Informatics (IIAIAAI), 2014, pp.527-528.
[2] Fujii, T. "University evaluation and institutional research (in japanese)", vol.7, 2016, pp.3-10, http://iir.ibaraki.ac.jp/jcache/lib/docu/007_h2811/007-h2811-01_fujii.pdf, (accessed 2016-12-31).
[3] Sun, Y., Todo, N., Inoue, S. "Exploration of higher education indicators and universities' characteristics in Japan", Proc. Of IIAI International Conference on Advanced Applied Informatics (IIAIAAI), 2014, pp.411-416.
[4] Funamori, M. "Thorough analysis of School Basic Survey -enrollment and retention management", http://researchmap.jp/?action=cv_download_main&upload_id=94797, (accessed 2016-12-31).
[5] Mori, M. "Benchmark system of School Basic Survey", http://www.masaomori.net/ir-analysis/BSJE_bench/, (accessed 2016-12-31).
[6] Pu, X., Chatti, MA., Schroeder, U. "Wiki-LDA: A mixed-method approach for effective interest mining on twitter data", Proc. Of 8th International Conference on Computer Supported Education (CSEDU), 2016, pp.426-433.
[7] Funamori, M., Mori, M., "Conceptual design toward a visualization system of university's web presence -simple analysis and system development using twitter-", Proc. Of IIAI International Conference on Advanced Applied Informatics (IIAIAAI), 2016, pp.456-461.
[8] Hirokawa, S., Ito, E., Baba, K. "Visualization of keywords for exploratory search of research", Journal of Information Processing and Management, vol.58, no.6, 2015, pp.447-454.
[9] Yin, C. Tabata, Y. Wu, X. Nakatoh, T. Hirokawa, S. "Building a search engine for scientific projects survey", Proc. Of IEEE International Conferences on Internet of Things and Cyber, Physical and Social Computing, 2011, pp.558-563.

[10] Midterm goals and midterm plans of national universities, Ministry of Education, Culture, Sports, Science and Technology, http://www.mext.go.jp/a_menu/koutou/houjin/1368750.htm, (accessed 2016-12-31).

[11] Japanese College and University Portraits, http://portraits.niad.ac.jp, (accessed 2016-12-31).