Improving the Support System of Public Sports Facilities Applying Text Mining and Multiple Focused on the support facilities for the National Sports Promotion Fund

Il-Gwang Kim, Mi-Suk Kim, Su-Sun Park, Jialei Jiang, Seong-Taek Park

Abstract Background/Objectives: ***This study used text mining to analyze the user complaints about public sports facilities supported by the Korea Sports Promotion Fund and seek measures for improvement.***

Methods/Statistical analysis: ***A framework for sports texts should be designed to include diverse features for collecting and analyzing sports-related texts. Among other methods of topic modelling, this study used the most widely used probability model, LDA(Latent Dirichlet Allocation). Word2vec models are applicable for different purposes. This study used Word2vec to identify key words highly associated to relevant key words.***

Findings: ***The analysis highlighted the following. First, the LDA topic clustering analysis by type identified 4 important key words (instructors, members, swimming and failure), which were in turn explored further with Word2Vec. Second, the analysis of associated words found such salient words as swimming, members, time, center, class and fitness acceptance in relation to the general type, whereas members, swimming, time, center, exercise, class and lesson proved important in the complex type. Third, as for the frequency of words, swimming, members and center frequently appeared in the general type in the order named, whereas the complex and gymnasium types were associated with the importance of swimming, members and time, in the order named.***

Improvements/Applications: ***The present findings may serve as a guideline for public sports facilities as public goods to improve the quality of service for users based on the user complaints.***

Keywords: ***Public Sports Facilities, Text Mining, Multiple Comparative Study, National Sports Promotion Fund, bigdata***

I. INTRODUCTION

As data have been generated and accumulated by dint of ever-evolving ICT, the need for big data analysis has been rising as a means of gaining new insights[1]. A wide range of instruments and methods for data collection and analysis have been developed. Still, relevant research in the field of sports is in its infancy and lacks in cases, which warrants the need for lots of research efforts to lay the foundation. Also, the ultimate purpose of text mining in research on sports is to maximize the automation of data collection and analysis, and continuously use them to minimize the manual work needed for research.

So far, diverse findings have been derived from text mining. Still, the focus has been put on analysis without taking into account how to apply and utilize methodologies from multiple perspectives. Therefore, it is now necessary to seek some methods of analyzing diverse sports-related texts for text-specific issues and details and to derive complex information from text mining. Currently, data or source codes used for scholarly purposes are hardly reused in most cases[2,3].

Yet, the primary strength of text mining lies in the timely findings through automated continuous analysis. To that end, this study developed a sports-related text analysis framework involving diverse features of collecting and analyzing sports-related text data.

Text analysis mostly is based on the frequency of key words and assumptions, which causes quite a few errors, and sensitive to the types and contents of texts, which warrants the application of an algorithm that fits the characteristics of texts given[4,5]. Hence, it is critical to formulate an algorithm optimized after countless rounds of trial and error, so as to minimize errors and derive domain-specific findings.

This study concerns the methods of using text mining to derive salient sports-related issues. To that end, it collects sports-related texts to analyze relevant issues, and determines which text mining or big data analysis method should be used to derive certain results. Particularly, this study analyzes different sports-related texts collected from a set of documents carrying different characteristics to identify relevant issues, sub-classify those texts based on their characteristics and derive complex information.

II. THEORETICAL BACKGROUND

With ICT having taken long strides, diverse data have been generated and accumulated to the extent that the need for big data analysis emerges across

Revised Manuscript Received on May 22, 2019.

Il-Gwang Kim, Department of Leisure Sport Industy, Korea national Sport University, 1239, Yangjae-daero, Songpa-gu, Seoul, 05541, South Korea

Mi-Suk Kim, Korea Institute of Sport Science, 727, Hwarang-ro, Nowon-gu, Seoul, 01794, South Korea

Su-Sun Park, Department of Social Welfare, Seowon University, 377-3, Musimseo-ro, Seowon-gu, Cheongju-si, Chungbuk,28674, South Korea

JIALEI JIANG, [5,*]**Seong-Taek Park,** Department of MIS, Chungbuk National University, Chungdae-ro 1, Seowon-gu, Cheongju, Chungbuk, 28644, South Korea

solpherd@cbnu.ac.kr

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

research and industrial sectors. Big data are generally categorized into structured and unstructured data. Text data belong to the latter as they are not structured in pre-determined forms or formats (e.g. tables, vectors, numbers). Compared with structured data, unstructured text data need much manual work, undergo complex procedures for analysis, become structured or are analyzed with an algorithm specialized for each type of unstructured data[6].

2.1. *LDA(Latent Dirichlet Allocation)*

Topic modelling is a methodology for finding out abstract topics from a set of documents or corpus. Among other methods of topic modelling, this study used the most widely used probability model, LDA(Latent Dirichlet Allocation) [7,8]. The result from LDA is a collection of words corresponding to a topic. Researchers use those words to determine a topic. LDA assumes each document may contain multiple topics, each of which involves multiple words, and each word has a probability of a topic[9]. That is, a topic comprises a weighted combination of diverse topics. Likewise, a document may represent a weighted combination of diverse topics[10].

2.2.Word2vec model

Word2vec is a method of embedding words based on neural networks by converting a high-dimensional one-hot encoding vector to a low-dimensional vector without considering adjacent word[11,12]s. For example, suppose there are 10K words. Word2vec considers the distribution of those words and converts it into a 5D vector to make the words used in similar contexts have a similar cosine distance, which does not depend on sizes but is used to calculate a distance based on relative proportions[13].

Word2vec models include CBOW(Continuous Bag Of Words) and Skip-gram models. The CBOW model uses adjacent words to infer a central word, whereas the Skip-gram model uses a central word to infer its adjacent words. Word2vec models are applicable for different purposes. This study used Word2vec to identify key words highly associated to relevant key words.

III. RESEARCH MODEL AND HYPOTHESIS

3.1. Overview of the sports text framework development

A framework for sports texts should be designed to include diverse features for collecting and analyzing sports-related texts. This study separately implemented a collector and an analyzer and included them in the framework. The sports text analysis framework consists of three parts, i.e. a feature to regularly collect and save environment texts, a second part for collecting and saving texts, and a third part for analyzing sports texts[14].

3.2. Collecting and saving sports texts

In this study, JAVA and Python's request libraries were used for crawling and scraping, and Python – Beautiful Soup library was used to parse titles, contents and dates, which were in turn saved in a file. First, it is necessary to identify the structures and URL rules of the web pages where relevant documents are found, so as to collect data. The fields used in the indexing file include dates, titles, URLs(content_url) containing the contents, and the paths to save data in files (file_path) [15,16,17].

While collecting the data, it is designed to refer to the indexing file, compare the dates first with the latest indexing dates of articles, add the list of data which are not crawled to the indexing file, and perform the crawling and parsing of the contents of texts.

3.3. Analyzing sports texts

To analyze the sports texts, diverse features were implemented in this study, i.e. calling different environment texts from the server, filtering documents by date, filtering documents by content, topic modelling for thematic analysis, analyzing the frequency of key words, analyzing similar contexts for key words, and extracting sentences containing key words.

3.4. Sample

A total of 20 civic sports centers run by 17 municipalities in 2018 were selected: 6 general types, 8 complex types and 6 gymnasium types.

IV. RESULTS

4.1.LDA analysis

4.1.1. Pools general type

No. 1 and No.3 topics are similar to each other. The key words include people, sports and time, from which it can be inferred that 'time' is the topic. The verb key words include such positive key words as love, be like and become. The lower the gamma value, the longer the red bar. The greater the gamma, the longer the blue bar(Fig 1.).

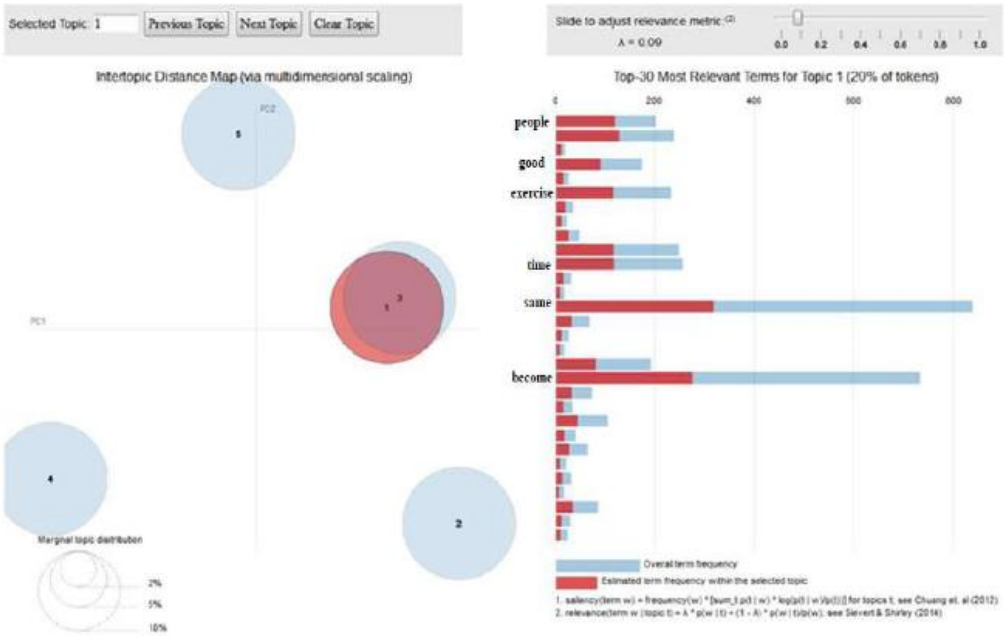

Figure 1. Pools general type LDA

You can see that the first topic and the third topic are similar. Keywords: people, exercise, time, etc. (You can deduce that the topic is time), Verb keywords: good, same, become, and so on.

4.1.2. Multi-purpose gymnasiums

No. 1 and No.3 topics are similar to each other. The key words include people, sports and time, from which it can be inferred that 'time' is the topic. The verb key words include such positive key words as love, be like and become. The lower the gamma value, the longer the red bar. The greater the gamma, the longer the blue bar(Fig. 2).

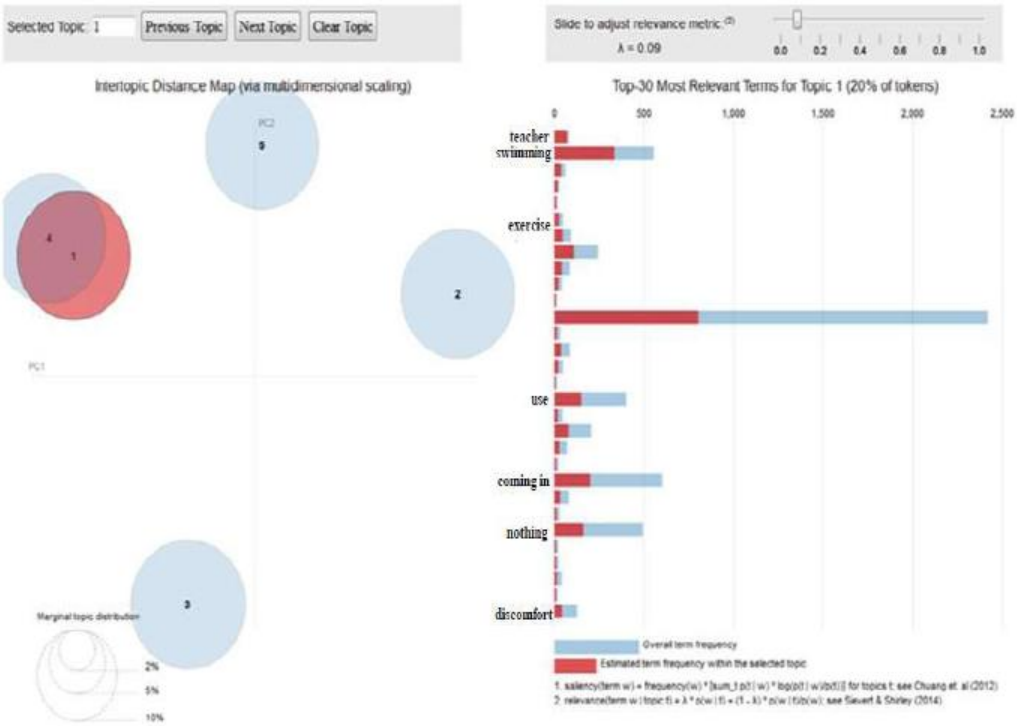

Figure 2. Multi-purpose gymnasiums LDA

You can see that the first topic and the fourth topic are similar. Keywords: swimming, membership, use, teacher, discomfort (can be inferred that the topic is a member), Verb keywords: negative keywords such as coming in, nothing, and so on

4.1.3. Complex gymnasiums

No. 1 and No.3 topics are similar to each other. The key words include people, sports and time, from which it can be inferred that 'time' is the topic. The verb key words include such positive key words as love, be like and become. The lower the gamma value, the longer the red bar. The greater the gamma, the longer the blue bar(Fig. 3).

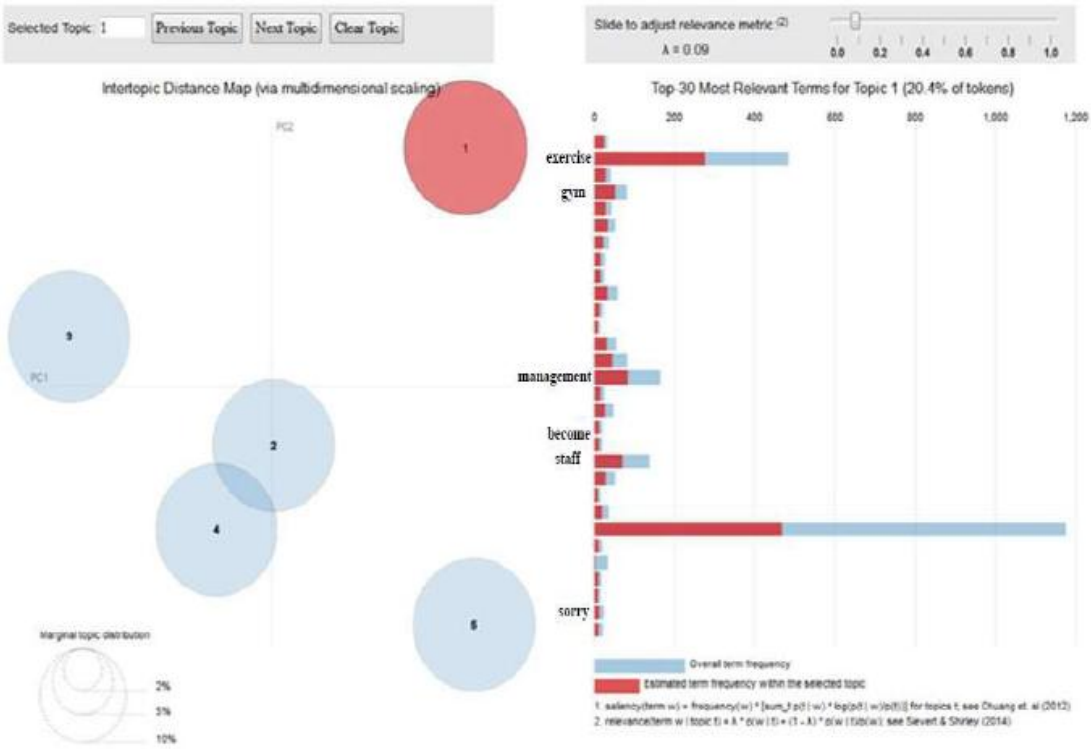

Figure 3. Complex gymnasiums LDA

You can see that the first topic and the fourth topic are similar. Keywords: Exercise, management, staff, gym, etc. (You can deduce that the topic is exercise.), Verb keywords: negative keywords such as falling, sorry, etc.

4.2.Word2Vec analysis

30 words close to "instructors" were found. Quite a few negative key words (e.g. atmosphere, leave, too much, appointment, combined classes) appeared.30 words close to "members" were found. Quite a few negative key words (e.g. existing, damages, complaints, application, enrollment) appeared.Word2Vec analysis is written in Korean. The paper did not use it.

4.3.Analyzing associated words

4.3.1. Pools general type

Association analysis refers to the association between words.

That is, associative rule analysis is an algorithm that creates a set of rules that often tell which two sets of items occur. Text analysis with Market Basket Analysis shows the association between text and text.

<Fig. 4> shows the results of analyzing the words associated with the general-type 'pools'. Here, members, center, put in and the link are important.

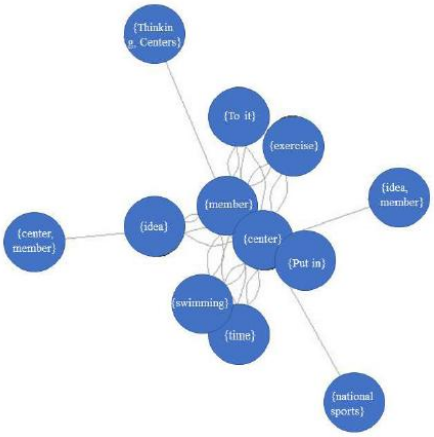

Figure 4. Pools general type of association rule

4.3.2. Multi-purpose gymnasiums

Association analysis refers to the association between words.

That is, associative rule analysis is an algorithm that creates a set of rules that often tell which two sets of items occur. Text analysis with Market Basket Analysis shows the association between text and text.

<Fig. 5> shows the results of analyzing the words associated with the multi-purpose gymnasiums. Here, members, center, builder, time, activation, and the link are important.

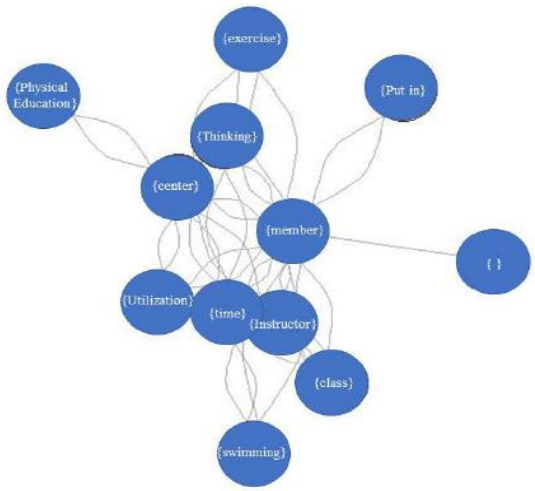

Figure 5. Multi-purpose gymnasiums of association rule

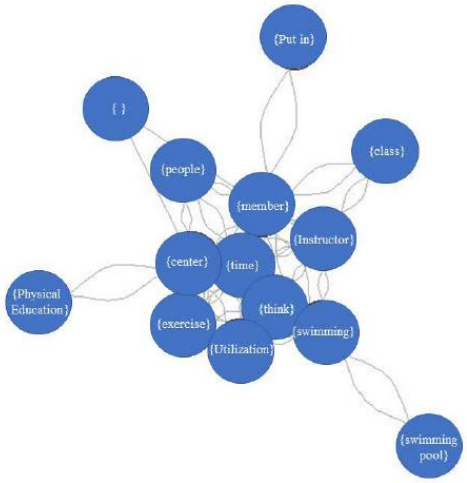

Figure 6. Complex gymnasiums of association rule

4.3.3. Complex gymnasiums

Association analysis refers to the association between words.

That is, associative rule analysis is an algorithm that creates a set of rules that often tell which two sets of items occur. Text analysis with Market Basket Analysis shows the association between text and text.

<Fig. 6> shows the results of analyzing the words associated with the complex gymnasiums. Here, time, think, center, member, exercise and the link are important.

4.4.Analyzing priorities in complaints (dissatisfaction, requests and inquires)

The analysis of complaints (dissatisfaction, requests and inquiries) relevant to the general-type 'pools' returned the following results. Out of 5,000 words in total, those words irrelevant to complaints were excluded to select 30 words for the purpose of this study.

Table 1: Results of analyzing priorities

	Pools general type		**Multi-purpose gymnasiums**		**Complex gymnasiums**	
	rev	Freq	rev	Freq	time	534
1	time	223	time	436	Utilization	508
2	Registration	208	Utilization	387	class	430
3	Application	205	class	368	Instructor	411
4	Instructor	201	Instructor	352	Lecture	259
5	Register	164	Lecture	189	shower	190
6	Utilization	155	shower	142	Health	182
7	Health	144	problem	140	Registration	175
8	teacher	128	Registration	135	problem	167
9	shower	114	Health	134	Register	167
10	Lecture	113	operation	123	Application	150
11	Internet	102	facilities	121	facilities	148
12	registration for courses	96	Application	116	answer	132
13	Dance	92	Register	108	lane	128
14	Position	85	answer	101	employee	124
15	Locker	81	lane	98	curious	119
16	operation	81	registration for courses	92	payment	115
17	fitting room	74	payment	87	registration for courses	115
18	Course App	70	Program	82	Position	114
19	Program	70	water quality	77	fitting room	103
20	curious	69	fitting room	73	program	102
21	employee	69	change	70	cleaning	92
22	discount	68	cleaning	69	a shower stall	90
23	refund	62	registration for courses Application	67	proposal	82

24	problem	60	a shower stall	66	water quality	82
25	weekend	59	teacher	66	Internet	80
26	payment	58	telephone	65	change	79
27	payment	58	Shuttle Bus	62	registration for courses Application	77
28	a new regulation	50	proposal	60	teacher	73
29	telephone	49	improvement	59	request	69
30	dawn	47	need	59	improvement	68

First, time, registration, application, instructors, acceptance and use were primarily considered in the order named in the complaints relevant to the general-type 'pools'. The complaints included the diversification of program sessions, and the requests for and inquiries about enrollment and application. Notably, new requests from many citizens arrived. Some argued the system prioritizing the existing users should be rectified.

Also, instructors' quality was an important aspect. Users complained about low-quality instructors, and abrupt replacement of instructors before the end of a semester.

Figure 7. general-type pools wordcloud

In addition, other complaints included shower booths. Users attending the last class daily complained about having to take a shower while workers were cleaning up the facilities and demanded more time for their shower. Also, inquiries about unfriendly staff, refund, payment, inconvenience, changing time slots and cancellation proved to be important factors.

The analyzed priorities in the complaints about the general-type 'pools' are plotted in a word cloud as follows. Larger and thicker words are more important(Fig. 7).

The analysis of complaints (dissatisfaction, requests and inquiries) relevant to the multi-purpose gymnasiums returned the following results. Out of 5,000 words in total, those words irrelevant to complaints were excluded to select 30 words for the purpose of this study.

First, time, Utilization, class, instructors, Lecture and use were primarily considered in the order named in the complaints relevant to the multi-purpose gymnasiums.

The analyzed priorities in the complaints about the multi-purpose gymnasiums are plotted in a word cloud as follows. Larger and thicker words are more important(Fig. 8).

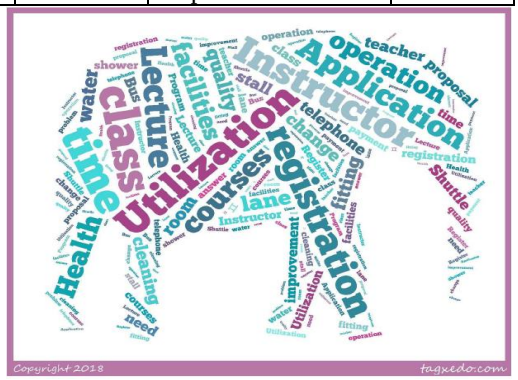

Figure 8. Multi-purpose gymnasiums wordcloud

The analysis of complaints (dissatisfaction, requests and inquiries) relevant to the complex gymnasiums returned the following results. Out of 5,000 words in total, those words irrelevant to complaints were excluded to select 30 words for the purpose of this study.

First, Utilization, class, Instructor, Lecture, shower and use were primarily considered in the order named in the complaints relevant to the complex gymnasiums.

The analyzed priorities in the complaints about the complex gymnasiums are plotted in a word cloud as follows. Larger and thicker words are more important(Fig. 9).

Figure 9. Complex gymnasiums wordcloud

V. CONCLUSION

This study used text mining to analyze the user complaints about public sports facilities supported by the Korea Sports Promotion Fund and seek measures for improvement. The findings revealed the complaints raised by citizens using such public sports facilities founded by the Fund and helped derive some measures for improving the support system.

The analysis highlighted

the following.

First, the LDA topic clustering analysis by type identified 4 important key words (instructors, members, swimming and failure), which were in turn explored further with Word2Vec.

Second, the analysis of associated words found such salient words as swimming, members, time, center, class and fitness acceptance in relation to the general type, whereas members, swimming, time, center, exercise, class and lesson proved important in the complex type. Also, members, swimming, instructors, time, center and class were important in relation to the gymnasium type.

Third, as for the frequency of words, swimming, members and center frequently appeared in the general type in the order named, whereas the complex and gymnasium types were associated with the importance of swimming, members and time, in the order named.

As a scholarly implication, this study used such big data analysis methods as crawling and text mining techniques to determine the user complaints in public sports centers. As a practical implication, the present findings may serve as a guideline for public sports facilities as public goods to improve the quality of service for users based on the user complaints.

This study has limitations as well including the fact that the analysis was limited to the comments posted on the home pages of the public sports centers, while excluding those facilities that were less used or did not have home pages. Future studies need to expand the sample group, and investigate some systematic measures for rectifying the matters manifested in complaints.

ACKNOWLEDGMENT

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2017S1A5A2A03068219)

REFERENCES

1. Park EM, SeoJH, Ko MH. The effects of leadership by types of soccer instruction on big data analysis. Cluster computing. 2016; 19(3):1647-1658.
2. Park ST, Lee SW, Kang TG. A study on the trend of cloud service and security through textmining technique. International Journal of Engineering & Technology. 2018; 7 (2.33): 127-132.
3. Kim DY, Park ST, Ko MH. A Study on the Analysis of IT-related Occupational Cluster using Big Data. IAENG International Journal of Computer Science.2018: 45(1): 7-11.
4. Jeon SH, Jo YY, Lim CJ, Park ST, Ko MH. An Exploratory Study on Measures for Aging Society based on Big Data Convergence. Indian Journal of Science and Technology. 2016 Dec 27;9(S1).
5. Li G, DaiJS, Park EM, Park ST. A study on the service and trend of Fintech security based on text-mining: focused on the data of Korean online news. Journal of Computer Virology and Hacking Techniques. 2017; 13(4): 249-255.
6. BenitoB, Solana J, Moreno MR. Assessing the efficiency of local entities in the provision of public sports facilities. International Journal of Sport Finance.2012; 7(1).
7. Da-chao Z, Min L. Studies on evaluation index system of public sports facilities development level in China. China Sport Science. 2013; 33(4): 3-23.
8. Kung SP, Taylor P. The use of public sports facilities by the disabled in England. Sport management review. 2014; 17(1): 8-22.
9. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of machine Learning research, 2003; 3(Jan): 993-1022.
10. Hong L, Davison BD. Empirical study of topic modeling in twitter. InProceedings of the first workshop on social media analytics 2010 Jul 25 (pp. 80-88). ACM.
11. Goldberg Y, Levy O. word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. arXiv preprint arXiv:1402.3722. 2014 Feb 15.
12. Rong X. word2vec parameter learning explained. arXiv preprint arXiv:1411.2738. 2014 Nov 11.
13. Pennington J, Socher R, Manning C. Glove: Global vectors for word representation. InProceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) 2014 (pp. 1532-1543).
14. Lazaridou A, Pham NT, Baroni M. Combining language and vision with a multimodal skip-gram model. arXiv preprint arXiv:1501.02598. 2015 Jan 12.
15. Hedley J. jsoup: Java html parser. 2009-11-29)[2015-06-12] http://jsoup. org. 2009.
16. Richardson L. Beautiful soup documentation.
17. Nair VG. Getting Started with Beautiful Soup. Packt Publishing Ltd; 2014 Jan 24.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication