International Journal of Engineering & Technology, 7 (2.33) (2018) 161-164

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

A study on improving the quality inspection on national information by using levenshtein distance algorithm

Sanggi Lee [1], Inje Kang [2], Eungyeong Kim [3], Kangryul Shon [3], Chulsu Lim [3] *

[1] First Author, National Science & Technology Information Service Center, Korea Institute of Science and Technology, 245, Daehak-ro, Yuseong-gu, Daejeon, 34141, Republic of Korea
[2] University of Science and Technology, 217, Gajeong-ro, Yuseong-Gu, Daejeon, 34113, Republic of Korea
[3] National Science & Technology Information Service Center, Korea Institute of Science and Technology Information, 245, Daehak-ro, Yuseong-gu, Daejeon, 34141, Republic of Korea
**Corresponding author E-mail: cslim@kisti.re.kr*

Abstract

Background/Objectives: In Korea, much effort and budget were spent to improve national R&D information management. However yet, project summaries of national R&D are not accurate enough to be utilized.
Methods/Statistical analysis: To examine the accuracy of project summaries, Levenshtein Distance Algorithm (LDA) was applied. LDA is expected to extract improper project summaries of which some parts of sentences are repeatedly used. To evaluate how the algorithm performs with national R&D information in Korea, project summaries of 53,492 national R&D projects that were conducted in 2014 were used.
Findings: Unlike other algorithms, LDA was able to detect project summaries consisted of repeatedly used phrases. According to the test with LDA, from 53,492 cases, 3,445 projects had inaccurate contents in project summaries. In details, 2,707 projects had improper research objective, while 712 projects and 26 projects had improper contents in research summary and expected impact, respectively. Although the algorithm allowed extracting repeatedly used phrases, it had problems of time; thus, it was only applied offline. Also, a research had to confirm once more to verify the accuracy of the result.
Improvements/Applications: This paper applied LDA to detect inappropriate project summaries. The result implies that by applying LDA, the quality of the information can be improved to facilitate the utilization.

Keywords: *Use about five key words or phrases in alphabetical order, Separated by Semicolon.*

1. Introduction

Republic of Korea (ROK) was once one of the world's poorest nation in 1950s; however, is now one of the advanced country. The quick elevating transition to an advanced country is due to continuous investment on R&D [1], [2]. As demonstrated in Table 1, according to Gross Expenditure on R&D (GERD), the statehas spent substantial proportion of GDP on R&D. For the last 20 years, from 1996 to 2015, the average GERD in ROKwas 3.0%, while the average GERD in Organization for Economic Co-operation and Development (OECD) countries were 2.2 than many other advanced countries [3]. The average growth rate of GERD in ROK was 3.4%, which is one of the highest average growth rates.

Table 1: Average GERD and Growth Rate [3] (Unit: %)

Country	Average GERD	Average Growth Rate
Republic of Korea	3.0	3.4
OECD	2.2	1.0
The U.S.	2.6	0.7
Japan	3.1	1.1
China	1.3	7.2
Germany	2.5	1.7
France	2.1	0.0
Israel	3.9	2.8

Due to the increasing expenditure on R&D, the need of efficient management and investment on R&D has increased in ROK. Thus, the stated developed "the world's first R&D information portal" in 2008, which is called the National Science and Technology Information Service (NTIS) [4], [5]. According to 「Regulations on the Management, etc. of National Research and Development Projects」, national standard for R&D information is determined for mutual interaction with NTIS and 17 representative research management institutes. Currently, there are 422 metadata items, including information on projects, human resources, or outcomes, gathered and provided.
To secure the accuracy of the managed data, NTIS has established guidelines for researchers or the research management institutes to demonstrate how to input the data. One of the guidelines is information on project summary. As researchers need to input the summary themselves, some of the project summaries weren't filled out properly. For instance, unnecessary or meaningless words or sentences were repeatedly written [6].Thus, to improve the accuracy of the project summaries, this paper developed an algorithm to examine whether the summaries were dutifully written.

2. Literature review

Based on DQM3, Lee [7] suggested criteria to maintain preeminent information management. To examine the quality of information management, six criteria, accuracy, consistency, utility, accessibility, security, and timeliness, need to be evaluated.
According to the criteria abovementioned, national R&D data quality management system in ROK needed to be improved as shown in Table 2. To improve the accuracy, offline business rules need to be managed and shared. Also, analyzing and preventing the error is required. To consistently manage data quality, error data needs to be traced and managed. In addition, the quality of raw data of national R&D needs to be improved. Third, to improve utility, requirements such as improving the project summaries or the management of national R&D data quality, should be reflected. Lastly, to improve security and timeliness, personal information should not be leaked. Also, information life cycle and long-term preservation system needs to be constructed.

Table 2: Assessing National R&D Quality Management System Based on DQM3 [7]

Criteria	Drawbacks
Accuracy	- Offline BR management - Analyzing and preventing error data
Consistency	- Quality of raw data of national R&D
Utility	- Quality of project summaries - Quality of national R&D science data
Accessibility	- Accessibility and compatibility
Security	- Security filter in national R&D data - Personal information leakage
Timeliness	- Information life cycle and long-term preservation system

This study focused on improving the quality of project summaries. A project summary of national R&D is composed of four items, research summary, research objective, expected impact, and keyword. The summary can be used to trace duplicated projects or to examine whether the newly planned projects are similar to the previous projects. Thus, it is very important to confirm whether the summaries are precise enough to show distinct features of the projects. Common problems that cause erroneous information are summarized in Table 3. Some summaries included meaningless contents such as 'will be written later', 'R&D', or 'no contents'. Some were filled with characters such as '*'. In addition, some were filled with same words or sentences, like 'secret,secret,secret' or 'this study is to examine…', 'this study is to examine…'. Thus, as shown in Table 2, input guidelines for each input item were established [7].

Table 3: Guidelines when Inputting Project Summary [7]

Input Item	Input Rule	Input Guideline
Research Summary	- 100 ~ 2,000 letters in Korean - 200 ~ 4,000 letters in English	- Summarize the key points of the research - Do not use meaningless words - Do not repeatedly input same words or sentences
Research Objective	- 100 ~ 2,000 letters in Korean - 200 ~ 4,000 letters in English	- Designate specific and numeric goals - Do not use meaningless words - Do not repeatedly input same words or sentences
Expected Impact	- 1 ~ 2,000 letters in Korean - 1 ~ 4,000 letters in English	- Expected impact on society, economy, or culture - Do not use meaningless words - Do not repeatedly input same words or sentences
Keyword	- 1 ~ 128 letters in Korean - 1 ~ 256 letters in English	- At least one keyword that represent the project - Use comma(,) if two or more keywords are input - Do not use meaningless words

In order to examine whether project summaries are dutifully written, algorithms had been developed. First algorithm is the one to examine whether the data is meaningless or not. To develop the algorithm, the stopword dictionary was constructed in prior. The dictionary included words such as 'research objective', 'security', 'test', 'evaluation', or 'etc' as shown in Table 4. The algorithm can be applied both offline and online.

Table 4: Examples of Stop words [7]

Insincere Word Choice	General Terms	Errors
- test - security - refer to the attached file - no contents - no output - still working on the research - not applicable - null - others - planning to input later - no data	- R&D - evaluation	- #NAME? - ERROR 23

The second algorithm is to trace the words that are repeated. The project summary is divided into character string (Table 5), and the character string is compared with the original data. If more than 80% are duplicated, the summary is classified as the improper summary. The algorithm can be applied online [7].

Table 5: Length of Divided Character String [7]

Length of project summary	Length of divided character string
4 ~ 7 byte	2byte
8 ~ 11byte	2, 4byte
12 ~ 15byte	2, 4, 6byte
16byte ~	2, 4, 6, 8byte

The aspects of the algorithms abovementioned are summarized in Table 6.

Table 6: Algorithms to Detect the Accuracy of Project Summaries [7]

Algorithm	Details	Aspect
Algorithm examining meaningless data (A-M)	- Extracting data that does specify distinctive features of projects	- Stopword dictionary - Applied offline and online
Algorithm examining repeated words (A-R)	- Extracting data with words that are repeatedly used	- Considered as 'repeated words' if the character strings matches with the original data more than 80% - Applied online

3. Levenshtein distance algorithm

Although the abovementioned algorithms are expected to detect the projects summaries with improper contents, it is difficult for them to identify inadequate contents if only some parts of a sentence are repeatedly used. Therefore, to determine whether project summaries were filled out properly or not, this paper applied Levenshtein Distance Algorithm (LDA). The algorithm is developed to compare two character strings. For example, if a character string [s] is 'boy' and the other string [t] is also 'boy', then Levenshtein Distance(s, t) is 0 as there is no additional work to do to make [s] equal to [t]. On the other hand, if [s] is 'boy' but [t] is 'bay', Levenshtein Distance(s, t) becomes 1. It is because an additional work, changing 'o' to 'a' is required to make [s] equal to [t].
In details, the process of measuring is as follows.
First, if a character string, 'ABCDBCDA' is input; the given string is divided into two variables, var_i and var_j. Thus, var_i is 'ABCD', and $var_j$is 'BCDA'. Then, it is organized in a matrix. Thus, i[1], i[2], i[3], and i[4] is A, B, C, D, respectively, while [j]1, j[2], j[3], and j[4] are B, C, D, A, respectively.
Second, as i[1] and j[1] are different, the distance between i[1] and j[2] is 1. The distance is the minimum value of each adjacent cell plus 1. The adjacent cells are the cells highlighted in blue color. As the adjacent cells are 0, 1, and 1, the distance is 1, as highlighted in yellow.

Third, as both i[2] and j[1] are B, the distance between the two is 1, which is the value of the adjacent cell that is in the second row.
The same method of measuring the distance is applied until the end. In sum, if i and j are the same, the distance between the two is the value of the adjacent that is in lower row. On the other hand, if i and j are not the same, the distance is the minimum value of each adjacent cell plus 1.

The process above is demonstrated in Figure 2. The cells highlighted in blue color are the adjacent cells, and the ones highlighted in yellow color are the measured distance. The characters in the circles are the ones that are compared in each phase.
According to the Figure 1, the distance of the divided string is 2, while the length is 4. The result implies that 50% ($\frac{2}{4} \times 100$) are identical.

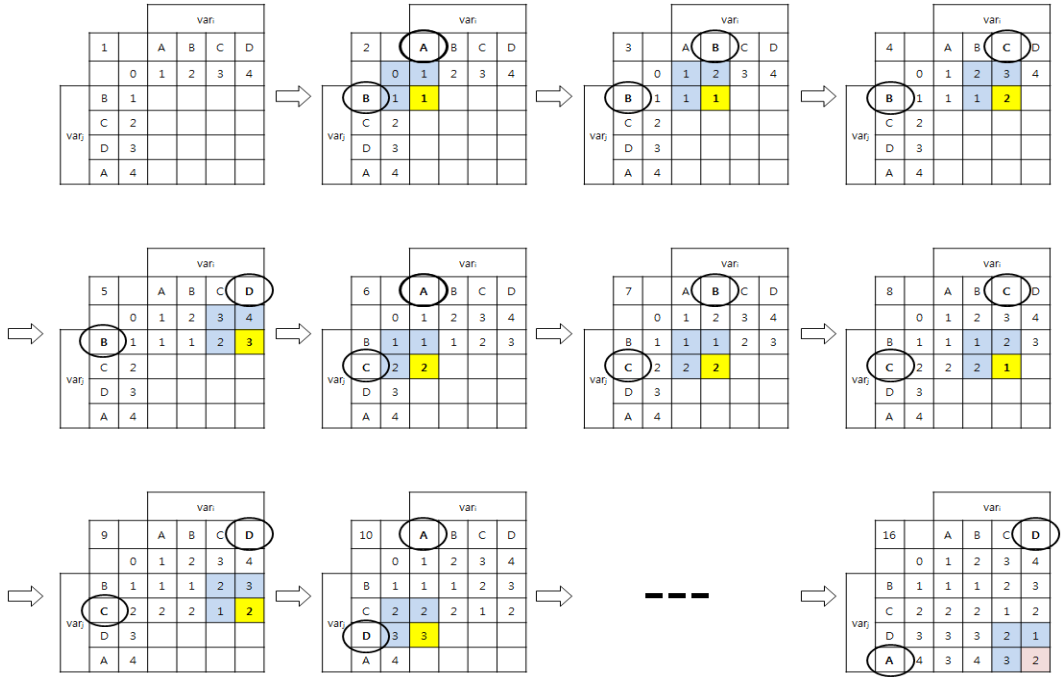

Fig. 1: Process of Measuring Levenshtein Distance [7].

4. Result

To examine whether LDA accurately performs detecting the accuracy of project summaries, this paper applied the algorithm and examined 53,492 cases of national R&D projects that were conducted in 2014.As shown in Table 5, 3,445 projects had inaccurate project summaries. 2,707 projects did not have proper research objective. 712 projects had improper research summary, while 26 projects had improper expected impact [6], [7].
Although LDA improved the accuracy of national R&D projects by detecting improperly filled project summaries, it takes too much time. Thus, LDA could only be applied offline. Also, as there is possibility of the algorithm making false classification, the person in charge (POC) needed to confirm the result.

Table 7: Algorithms to Detect the Accuracy of Project Summaries [6], [7]

Input Items	Criterion of an 'inaccurate project summary'	Number of cases
Research Objective	- if some parts of sentence is repeatedly used (over 95% of similarity)	2,707
Research Summary	- if some parts of sentence is repeatedly used (over 95% of similarity)	712
Expected Impact	- if some parts of sentence is repeatedly used (over 95% of similarity)	26

5. Conclusion

Although the project summaries are crucial as they allow detecting duplicated or similar projects, some researchers or research management institutes do not sincerely input the information. Thus, this paper applied LDA in order to extract inaccurate project summaries. To examine how LDA performs with data from NTIS, this paper conducted a test on the algorithm with information on national R&D projects conducted in 2014.
According to the test, about 6.4% of the projects had inaccurate project summaries. 2,707 projects had inaccurate research objective, while 712 and 26 projects had improper research summary and expected impact, respectively. The result implies that applying three algorithms together, A-M, A-R, and LDA, the accuracy of project summaries would be improved, and eventually contribute to promote utilizing the information.
However, because of precision and required time due to the unstructured text forms of project summaries, the POC needs to additionally confirm the result.

Acknowledgement

This research was supported by Maximize the Value of National Science and Technology by Strengthen Sharing/Collaboration of National R&D Information funded by Korea Institute of Science and Technology Information (KISTI).

References

[1] Fernando H S G. Relevance of development assistance to the economy and its impact after Sri Lanka's elevation to upper middle income status, *Sri Lanak Forum of University Economists (SLFUE), Department of Economics, Faculty of Social Sciences, University of Kelaniya*, 2016, pp.202-210.
[2] Herrera M E F. Contrasting the strategic role of firms in the economic development of Ecuador with that of South Korea using Ghemawat CAGE distance framework, 2017.
[3] Gross domestic spending on R&D. https://data.oecd.org/rd/gross-domestic-spending-on-r-d.htm. Date accessed: 11/29/2017.
[4] Kang N, Park M, Choi K, Kim T, Joo W, Kown O. A development of service model for mapping the ecology of scientific research us-

ing national science & technology information service. *Indian Journal of Science and Technology*. 2015, 8 (S1), pp. 121-130.

[5] Kang I, Lee B, Kim S, Lim C, Choi K. Design and pilot test of software prototype for linking national R&D information with result materials. *Indian Journal of Science and Technology*. 2016, 9 (46), pp. 1-8.

[6] Lee S, Kim E, Shon K, Lim C. A study of algorithm for quality inspection on the summary of national R&D program information, *The 7th International Conference on Convergence Technology 2017*, 2017, pp.58-59.

[7] Lee S. Improving the quality management system of national R&D data. *Doctoral dissertation,* University of Seoul, Seoul, Republic of Korea, 2015.