International Journal of Advanced Culture Technology Vol.9 No.3 345-351 (2021)
DOI https://doi.org/10.17703/IJACT.2021.9.3.345

IJACT 21-9-45

News Article Identification Methods in Natural Language Processing on Artificial Intelligence & Bigdata

[1]Jangmook Kang, [2]Sangwon Lee*

[1]Prof., Dept. of Hacking & Security, Far East Univ., Korea
[2]Prof., Dept. of Computer & Software Engineering, Wonkwang Univ., Korea
2021035@kdu.ac.kr, sangwonlee@wku.ac.kr

Abstract

This study is designed to determine how to identify misleading news articles based on natural language processing on Artificial Intelligence & Bigdata. A misleading news discrimination system and method on natural language processing is initiated according to an embodiment of this study. The natural language processing-based misleading news identification system, which monitors the misleading vocabulary database, Internet news articles, collects misleading news articles, extracts them from the titles of the collected misleading news articles, and stores them in the misleading vocabulary database. Therefore, the use of the misleading news article identification system and methods in this study does not take much time to judge because only relatively short news titles are morphed analyzed, and the use of a misleading vocabulary database provides an effect on identifying misleading articles that attract readers with exaggerated or suggestive phrases. For the aim of our study, we propose news article identification methods in natural language processing on Artificial Intelligence & Bigdata.

Keywords: *Artificial Intelligence, Bigdata, Fake News, Identification, Natural Language Processing*

1. INTRODUCTION

As access to news quickly shifts from offline media to online media, news influence has traditionally been assessed and advertised according to the number of subscribers and news ratings of media outlets. However, recently it can measure users' responses to individual news articles posted on the Internet in real time, leading to fierce competition among journalists and journalists. What usually appears to get the reader's attention is to entice them to click on the news by putting provocative phrases in the title or false phrases that differ from the content. As such, attracting readers with misleading titles can undermine the overall confidence in Internet news, and there is a problem that readers waste their time on unwanted news. Recognizing such problems, a website has emerged, which aggregates news articles containing frequently appeared words in misleading titles. But, since the site aggregates news articles with predetermined titles such as 'surprise' and 'shock', it is difficult to reflect the new Internet terminology, fail to tell readers whether news articles are misleading in real time, and find misleading articles with different titles and contents. On the other hand, a method of determining whether or not an article is misleading was recently introduced using algorithms that learned the language patterns of portal sites and general articles based on Artificial Intelligence technology. However, since this method takes a lot of time to analyze the context of the news, there is a problem that users have to wait a certain amount of time to see if it is fake news before clicking on it. This study is about how to judge misleading news

Manuscript received: August 17, 2021 / revised: September 1, 2021 / accepted: September 4, 2021
Corresponding Author: sangwonlee@wku.ac.kr
Tel: +82-63-850-6566
Professor, Dept. of Computer & Software Engineering, Wonkwang Univ., Korea

to determine fake news.

2. RELATED WORKS

We conducted the tendency of Bigdata [1-11], which is an essential foundation for conducting this study, as well as the application of Artificial Intelligence [12-15], a key element technology in analyzing Bigdata. And then, we looked at the concept of Text Mining [16-23], a key element technology needed to determine fake news.

2.1 Bigdata & Artificial Intelligence

Bigdata refers to a sea of information consisting of vast zettabytes of data flowing from computers, mobile devices, and mechanical sensors we use every day-to-day. Bigdata is used by organizations to make decisions, improve processes and policies, and build customer-centric products, services and experiences. Bigdata is defined as big not just because of the amount of data, but because of the diversity and complexity of characteristics. Bigdata typically exceeds the capacity of existing databases that collect, manage, and process data. Bigdata can also be generated from any object and location around the world that can be monitored digitally. Meteorological satellites, IoT devices, transportation cameras, and social media trends are some of the data sources that are mined and analyzed to strengthen business resilience and competitiveness. The true value of Bigdata is measured by the degree to which it can be analyzed and understood. Artificial Intelligence, Machine Learning, and the latest database technologies can provide actionable real-time insights through visualization and analysis of Bigdata. Bigdata analytics can help companies realize new opportunities and build business models. Bigdata management leverages systems with the ability to process and significantly analyze vast amounts of different and complex information. In this respect, Bigdata and Artificial Intelligence have a somewhat collaborative relationship. Without Artificial Intelligence to organize and analyze Bigdata, Bigdata will be less practical. And Artificial Intelligence provides powerful analytics that can be implemented according to the range of datasets included in Bigdata. Machine learning algorithms define incoming data and identify patterns within the data. These insights provide information about business decision making and help automate processes. Machine learning is based on Bigdata because the stronger the analysis of datasets, the greater the system's process learning, continuous evolution, and adaptation opportunities.

2.2 Text Mining

McKinsey Global Institute, an economic research institute under the global consulting group McKinsey & Company, divided Bigdata technology into analysis techniques for Bigdata analysis and processing technologies for collecting, manipulating, managing, and analyzing data. A typical method in the analysis technique is mining. Mining technology, which means mining minerals from mines, extracts information that can predict the future by identifying patterns and relationships hidden in large amounts of data. Currently, it is not only applied to corporate decision-making, marketing, and customer management, but it is also being used in various areas such as finance and education. Recently, various analytical techniques are drawing attention due to the increase in amorphous data, which is data that is not stereotyped due to complex shapes and structures such as pictures, videos, and documents. Text mining refers to the extraction of meaningful information from large-scale documents. There is a difference from data mining in that the target of analysis is unstructured document information. Text mining is sometimes referred to as text analysis, knowledge discovery from text databases, and document mining. Opinion mining is a Bigdata processing technology that collects and analyzes public opinion or information on a particular topic on websites and social media to produce results. It refers to a technique that analyzes public opinions, evaluations, and feelings about people and issues. Through this, it is possible to predict the size of the new product market or to understand consumer responses in advance. This is because structured and unstructured text posted on social media can determine whether the intention to deliver is positive or negative. It extracts opinions by separating facts and opinions, divides them into positive and negative, and measures their strength. It uses automated analysis methods

mainly because they are large web documents such as blogs and shopping malls. Since the analysis target is text, it utilizes natural language processing methods and computer linguistics that are utilized in text mining. It is becoming an important technology in the social media era. Web mining is data mining for web targets that extracts useful information from web log information or search terms generated by the Internet. Web mining requires separate analytical techniques because the properties of web data are semi-structured or amorphous and form a link structure. Web mining is divided into web structure mining, web usage mining, and web content mining, depending on the analysis target. Among them, web content mining is often used in search engines as a technique to quickly find information that web users want from content stored on web pages.

3. THE MODEL OF MISLEADING NEWS DISCRIMINATION SYSTEM

The purpose of this study is to provide a misleading news discrimination system that allows users to immediately tell whether they are misleading articles without having to wait. Another purpose is to provide a method of identifying misleading news that allows users to immediately tell if they are misleading articles without having to wait.

3.1 Architecture of the Proposed System

Figure 1 shows the architecture of our proposed system. In general, users of Internet news tend to collect news from various media outlets from Naver or the following news portals and use news that is placed according to portal standards or rules, or search keywords, rather than visiting certain media sites. In this case, users get a lot of news titles and summaries on a single screen, and click on one of the news to read the content, which is reflected in the number of views, affecting the placement of news articles on the portal and the media's advertising sales. Therefore, news titles are a significant factor in Internet news selection because they are an important factor in attracting users to increase the number of views, and for users, they are a means to access news in areas of interest with minimal search effort. Thus, the study initiates a system and a method that can inform users that the news title is wrong as soon as they pass their cursor over the news title or touch their fingers.

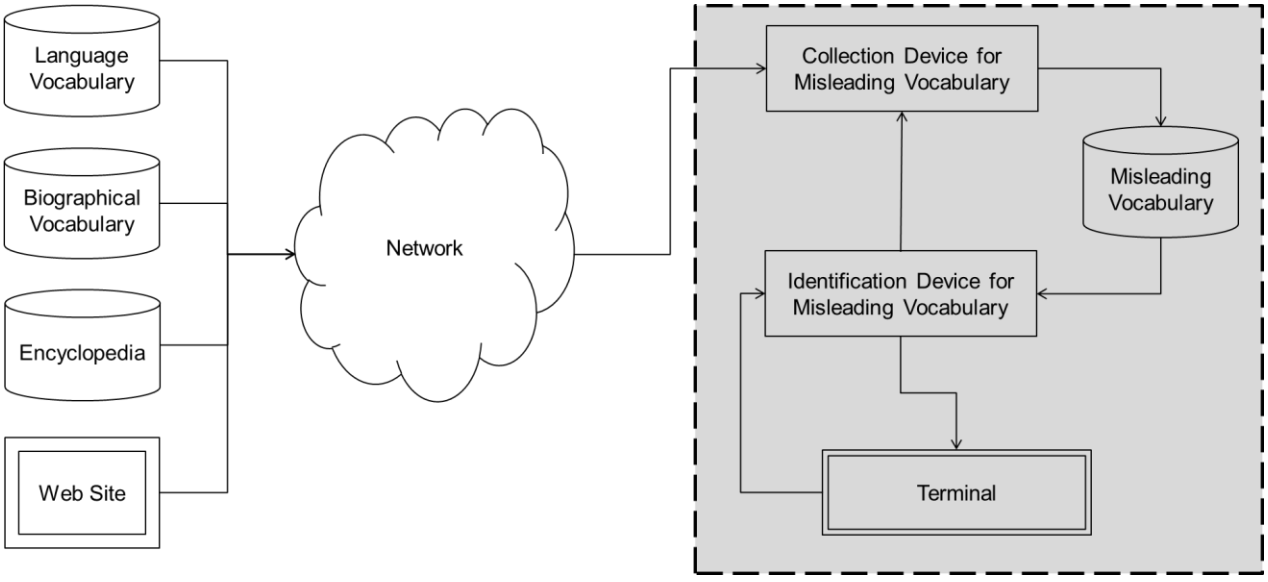

Figure 1. Architecture of the system

3.2 Notion of the Proposed System

Desirable embodiments in accordance with this study are described in detail with reference to the accompanying drawings. Figure 1 is a conceptual diagram of the misleading news discrimination system

according to an embodiment of this study. Our proposing model, the misleading news identification system, according to an embodiment of this study may consist of a misleading vocabulary database, misleading vocabulary gathering device (that is, collection device for misleading vocabulary), and misleading news identification device (that is, identification device for misleading vocabulary). The misleading news identification system according to an embodiment of this study can obtain language dictionary database, biographical dictionary database, encyclopedic (that is, encyclopedia) database, and various web content (that is web site) based on the Internet.

Hereinafter, each composition of the misleading news identification system according to an embodiment of this study is described in more detail. The misleading vocabulary database is a database containing misleading vocabulary extracted from Internet news articles. The misleading vocabulary may have been obtained from a public database that collected the misleading vocabulary, or extracted by the misleading vocabulary collection device while monitoring the Internet news article. Alternatively, the misleading news discriminator may be stored in the database by extracting the misleading vocabulary after determining whether it is misleading for articles requested by the user. Misleading vocabulary collection devices can collect misleading news articles while monitoring Internet news articles, extracting misleading vocabulary from the titles of the collected misleading news articles and storing them in the misleading vocabulary database. The misleading news identifier can refer to the misleading vocabulary database to determine the relevance of news articles associated with news titles selected by users on web pages where multiple news titles are exposed. At this time, language dictionaries, biographical dictionaries, and encyclopedias, which are released on the Internet, can be used to analyze vocabulary such as searching consent/similar words or extracting parts of index words. In addition, a web page with multiple news titles may be a web page on an Internet portal site that collects or edits news from multiple sources, or a web page on a search site that lists news collected associated with keywords you enter in the search window, but is not limited to this embodiment.

4. THE DESIGN OF MISLEADING NEWS DISCRIMINATION SYSTEM

The conceptual diagram of the misleading news discrimination system in Figure 1 is designed up of several components; misleading vocabulary gathering device (that is, collection device for misleading vocabulary), and misleading news identification device (that is, identification device for misleading vocabulary).

The misleading vocabulary collection device according to an embodiment of this study may consist of a news monitoring unit, an independent response analysis unit, a misleading vocabulary extraction unit, and a database storage unit. In addition, the misleading vocabulary collection device according to an embodiment of this study works in conjunction with the misleading news discriminator, and can be configured to store misleading vocabulary in the misleading vocabulary database. The detailed composition of the misleading vocabulary collection device according to an embodiment of the present study can be described with a following example. When the news monitoring department monitors news articles on the Internet and extracts certain news articles whose negative responses exceed certain thresholds, the independent response analysis department can analyze the above news articles. For example, in a comment on a particular news article, a reader's response analysis department may ask for a judgment on misleading by passing the news article to a misleading news discriminator if it finds a vocabulary or response that means misleading, expresses negative views, or criticizes the reporter.

The misleading vocabulary extraction unit can extract misleading vocabulary based on the results and add it to the misleading vocabulary database when it receives the results of the determination of the title consistency of a particular news article. For example, the misleading vocabulary extractor may add that vocabulary to the misleading vocabulary database if the part contains adjectives, adverbs, exclamations, and expressive nouns (e.g., astonishment, shock, etc.).

The composition of misleading vocabulary discriminator according to an embodiment of this study can show an example of a user's choice of news titles according to an embodiment of this study, in which the user determines whether or not a misleading article is displayed on the screen. The news title analysis department is a composition of misleading vocabulary identification devices according to an embodiment of this study. The misleading vocabulary identifier according to an embodiment of this study may consist of news title

identification, body text extraction, news title analysis, title consistency determination, user terminal display and misleading vocabulary transmission. In addition, misleading vocabulary discriminator according to an embodiment of this study can be operated in conjunction with user terminals and misleading vocabulary collection devices. In the news title identification unit can identify news titles selected by users among multiple news titles above on web pages where multiple news titles are exposed. For example, a news title identifier can identify the title of a user-specified news by selecting a specific news from a portal site's news list or keyword search through a news title identifier can identify the title of a user-specified by the user. At this point, the user can identify the title of the article in the location by touching the news title displayed on the screen of the terminal or by positioning the mouse cursor. On the other hand, motion sensors mounted on the screen part of the device without the user touching the screen or mouse manipulation can identify the article titles that are directed in real time by following the user's finger. Body text extractor can extract body text from news articles within web pages linked to news titles selected by the user.

The news title analysis department can extract multiple indexes and similar words through morpheme analysis of news titles, including Index Extraction Module, Ontology Generation Module, First Weighting Module, and Second Weighting Module. Indexing modules extract text from user-selected news titles, extract multiple index words through morpheme analysis, and ontology generating modules can extract paraphrases and similar words from indexers extracted using Internet dictionaries (e.g., the following dictionary, Naver dictionary, Google dictionary, encyclopedia, Wikipedia, etc.). At this time, index words are not limited to words, but may include words, compound words, phrases, etc. The method of processing text in natural language is a known method and therefore details are omitted. The first weighting module may assign a predetermined first weight to each index, depending on the part of the index word extracted. For example, if a part is an adjective, exclamation, adverb, or noun that represents emotion, or a proper noun that directs a person or animal, it can be weighted so that the index or synonym is a vocabulary that must exist in the text of the news. The second weighting module searches the misleading vocabulary database and, when the above index is retrieved, gives the index a predetermined second weight. In other words, second-weighted index words can influence the determination of whether or not the corresponding vocabulary is included in the text. For example, statistically exaggerated provocative vocabulary and advertising phrases such as huck, shock, astonishing, amazing, secret, explosive, hot etc. can be weighted to judge as misleading articles, or only in non-identical form.

Subject consistency judgement can search the text of the text based on the extracted index words and similar words to determine the subject consistency of the news article based on the degree to which the extracted index words and similar words are contained and the first and second weights given by news subject analysis. In this case, multiple index words and similar words extracted above are distributed across multiple sentences to complete the context of the above news title. Based on the distance between the first and final sentences of the above multiple sentences, the title consistency of the news article can be determined. For example, in the news title ‘Continuing North Korea nuclear development’ and submitting a report to the Security Council, the index terms ‘North Korea’, ‘Continuing’, ‘Report submission’ were extracted from a single sentence of the news body. In addition, if some indexers or synonyms are not found in some of the text, the suitability of the title may be underestimated. In particular, subject conformity can be assessed fairly negatively if a proper noun or a vocabulary representing figures or nominations is missing.

In other words, the vocabulary contained in the title is found in the text, but not in a single sentence, but the more distributed it is, the less consistent it is with the title, and the less consistent it is with the degree and importance of the missing index. Subject consistency can be determined by matching/mismatching, or by quantifying each discrepancy factor (e.g., whether misleading vocabulary is frequently used, the degree of variance in the body of the vocabulary, the presence of missing index words, and the frequency of use of similar words relative to the body language).

The user terminal display can be digitized and delivered to the screen of the user terminal, if the subject consistency of the above news article is below a certain threshold as a result of the subject consistency judgment. The misleading vocabulary transfer unit can communicate the results of the above news article to the misleading vocabulary collection unit if the subject consistency of the above news article is below a certain threshold. As previously described, the misleading vocabulary gathering device can add index words to the misleading vocabulary database that negatively affected subject consistency among index words extracted

from the above news articles. At this time, it may be added to the misleading vocabulary database if the index words that negatively affected the subject consistency above are adjectives, adverbs, exclamations, and emotion phenotypes nouns.

5. CONCLUSIONS

In our study, we proposed news article identification methods in natural language processing on Artificial Intelligence & Bigdata. The misleading news article identification system and method in accordance with the morpheme analysis of news titles as described above can extract index and synonyms, determine if they are fish news articles based on multiple sentences in body text to complete the context of news titles. Also, since only relatively short news titles are morphed, it does not take much time to judge misleading news, and the use of a misleading vocabulary database makes it possible to identify misleading articles that attract readers with exaggerated or sensational phrases. Meanwhile, the methods and devices of this study described so far can actually be implemented by computer programs and stored in computer-readable recording media when run on a computer. Computer-readable recording media includes all kinds of recording media where programs and data are stored to be read by computer systems, and may also be implemented as transmitted over the Internet. In other words, such media can be distributed across networked computer systems, storing and executing computer-readable code in a distributed manner. When using the misleading news article identification system and method in this study, we could use morpheme analysis of news titles to extract index and synonyms, determine whether index and synonyms are fish news articles based on multiple sentences to complete the context of news titles. Also, since only relatively short news titles are morphed, it does not take much time to judge misleading news, and the use of a misleading vocabulary database makes it possible to identify misleading articles that attract readers with exaggerated or sensational phrases. Although the above-mentioned preferred embodiment of this study has been referred to, there would be some limitations. For example, even experienced personnel in the relevant field of technology can understand that this study can be modified and altered in a variety of ways within the scope of the research.

ACKNOWLEDGEMENT

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2018S1A5A2A03038738 / Algorithm Design & Software Architecture Modeling to Judge Fake News based on Artificial Intelligence).

REFERENCES

[1] S. Park, J.S. Hwang, and S. Lee, “A Study on the Link Server Development Using B-Tree Structure in the Bigdata Environment”, Journal of Internet Computing and Services, Vol. 16. No. 1. pp. 75-82, 2015. DOI: https://doi.org/10.7472/jksii.2015.16.1.75.

[2] S.B. Park, S. Lee, S.W. Chae, and H. Zo, “An Empirical Study of the Factors Influencing the Task Performances of SaaS Users”, Asia Pacific Journal of Information Systems, Vol. 25. No. 2. pp. 265-288, 2015. DOI: https://doi.org/10.14329/apjis.2015.25.2.265.

[3] S. Park, and S. Lee, “Bigdata-oriented Analysis on Issues of the Hyper-connected Society”, The E-Business Studies, Vol. 16. No. 5. pp. 3-18, 2015. DOI: https://doi.org/10.15719/geba.16.5.201510.3.

[4] Jumin Lee, S.B. Park, and S. Lee, “Are Negative Online Consumer Reviews Always Bad? A Two-Sided Message Perspective”, Asia Pacific Journal of Information Systems, Vol. 25. No. 4. pp. 784-804, 2015. DOI: https://doi.org/10.14329/apjis.2015.25.4.784.

[5] J.K. Kim, S.W. Lee, and D.O. Choi, “Relevance Analysis Online Advertisement and e-Commerce Sales”, Journal of the Korea Entertainment Industry Association, Vol. 10. No. 2. pp. 27-35, 2016. DOI: https://doi.org/10.21184/jkeia.2016.04.10.2.27.

[6] S.W. Lee, and S.H. Kim, “Finding Industries for Bigdata Usage on the Basis of AHP”, Journal of Digital Convergence, Vol. 14. No. 7. pp. 21-27, 2016. DOI: https://doi.org/10.14400/JDC.2016.14.7.21.

[7] S. Lee, and S.Y. Shin, “Design of Health Warning Model on the Basis of CRM by use of Health Bigdata”,

Journal of the Korea Institute of Information and Communication Engineering, Vol. 20. No. 4. pp. 1460-1465, 2016. DOI: https://doi.org/10.6109/jkiice.2016.20.8.1460.

[8] M. Nam, and S. Lee, "Bigdata as a Solution to Shrinking the Shadow Economy", The E-Business Studies, Vol. 17. No. 5. pp. 107-116, 2016. DOI: https://doi.org/10.20462/TeBS.2016.10.17.5.107.

[9] S.H. Kim, S. Chang, and S.W. Lee, "Consumer Trend Platform Development for Combination Analysis of Structured and Unstructured Bigdata", Journal of Digital Convergence, Vol. 15. No. 6. pp. 133-143, 2017. DOI: https://doi.org/10.14400/JDC.2017.15.6.133.

[10] Y. Kang, S. Kim, J. Kim, and S. Lee, "Examining the Impact of Weather Factors on Yield Industry Vitalization on Bigdata Foundation Technique", Journal of the Korea Entertainment Industry Association, Vol. 11. No. 4. pp. 329-340, 2017. DOI: https://doi.org/10.21184/jkeia.2017.06.11.4.329.

[11] S. Kim, H. Hwang, J. Lee, J. Choi, J. Kang, and S. Lee, "Design of Prevention Method Against Infectious Diseases based on Mobile Bigdata and Rule to Select Subjects Using Artificial Intelligence Concept", International Journal of Engineering and Technology, Vol. 7. No. 3. pp. 174-178, 2018. DOI: https://doi.org/10.14419/ijet.v7i3.33.18603.

[12] I. Jung, H. Sun, J. Kang, C.H. Lee, and S. Lee, "Bigdata Analysis Model for MRO Business Using Artificial Intelligence System Concept", International Journal of Engineering and Technology, Vol. 7. No. 3. pp. 134-138, 2018. DOI: https://doi.org/10.14419/ijet.v7i3.33.18593.

[13] S. Kim, S. Park, J. Kang, and S. Lee, "The Model of Bigdata Analysis for MICE Using IoT (Beacon) and Artificial Intelligence Service (Recommendation, Interest, and Movement)", International Journal of Engineering and Technology, Vol. 7. No. 3. pp. 314-318, 2018. DOI: https://doi.org/10.14419/ijet.v7i3.33.21192.

[14] S.H. Kim, J.K. Choi, J.S. Kim, A.R. Jang, J.H. Lee, K.J. Cha, and S.W. Lee, "Animal Infectious Diseases Prevention through Bigdata and Deep Learning", Journal of Intelligence and Information Systems, Vol. 24. No. 4. pp. 137-154, 2018. DOI: https://doi.org/10.13088/jiis.2018.24.4.137.

[15] S. Lee, and I. Jung, "Development of a Platform Using Bigdata-Based Artificial Intelligence to Predict New Demand of Shipbuilding", The Journal of The Institute of Internet, Broadcasting and Communication, Vol. 19. No. 1. pp. 171-178, 2019. DOI: https://doi.org/10.7236/JIIBC.2019.19.1.171.

[16] H. Hwang, S. Lee, S. Kim, and S. Lee, "Building an Analytical Platform of Bigdata for Quality Inspection in the Dairy Industry: A Machine Learning Approach", Journal of Intelligence and Information Systems, Vol. 24. No. 1. pp. 125-140, 2018. DOI: https://doi.org/10.13088/jiis.2018.24.1.125.

[17] Y. Shon, J. Park, J. Kang, and S. Lee, "Design of Link Evaluation Method to Improve Reliability based on Linked Open Bigdata and Natural Language Processing", International Journal of Engineering and Technology, Vol. 7. No. 3. pp. 168-173, 2018. DOI: https://doi.org/10.14419/ijet.v7i3.33.18601.

[18] T. Minami and K. Baba, "A Study on Finding Potential Group of Patrons from Library's Loan Records", International Journal of Advanced Smart Convergence, Vol. 2, No. 2, pp. 23-26, 2013. DOI: https://doi.org/10.7236/IJASC2013.2.2.6

[19] S.H. Kim, M.S. Kang, and Y.G. Jung, "Bigdata Analysis using Python in Agriculture Forestry and Fisheries", International Journal of Advanced Smart Convergence, Vol. 5. No. 1, pp. 47-50, 2016. DOI: https://doi.org/10.7236/IJASC.2016.5.1.47

[20] W.Y. Kim, "A Practical Study on Data Analysis Framework for Teaching 3D Printing in Elementary School", International Journal of Internet, Broadcasting and Communication, Vol. 8, No. 1, pp. 73-82, 2016. DOI: https://www.earticle.net/Article/A263475

[21] H.C. Kang, K.B. Kang, H.K. Ahn, S.H. Lee, T.H. Ahn, and J.W. Jwa, "The Smart EV Charging System based on the Bigdata analysis of the Power Consumption Patterns", Vol. 9, No. 2, pp. 1-10, 2017. DOI: https://www.earticle.net/Journal/Issues/821/22509

[22] Y.I. Kim, S.S. Yang, S.S. Lee, S.C. Park, "Design and Implementation of Mobile CRM Utilizing Bigdata Analysis Techniques", The Journal of The Institute of Internet, Broadcasting and Communication, Vol. 14, No. 6, pp. 289-294, 2014. DOI: https://doi.org/10.7236/JIIBC.2014.14.6.289

[23] S.J. Oh, "Design of a Smart Application using Bigdata", The Journal of The Institute of Internet, Broadcasting and Communication, Vol. 15, No. 6, pp. 17-24, 2015. DOI: https://www.earticle.net/Article/A259710