Word Cloud Techniques for Data Analysis

Se In Jung, Shin Dong Ho
Graduates and Professor, My Paul School
12-11, Dowontongmi-gil, Cheongcheon-myeon, Goesan-gun
Chungcheongbuk-do, Republic of Korea
eavatar@hanmail.net

Jeongwon Kim
Graduates, Department of Economics, College of Economics, Nihon University
3-2 Kanda-Misakicho, 1-chome, Chiyoda-ku, Tokyo, Japan
eavatar@hanmail.net

Abstract

In text data big data analysis, the majority of raw data is large and atypical, and analysis techniques cannot be applied. Therefore, the collected raw data removes unnecessary data through heuristic purification and old terms through purification. After that, the vocabulary frequency is calculated, visualized through word cloud technology, the core task is extracted, and the result is analyzed after the information is made. In this study, we propose a new method for improving unused terms using the external unused terms Set (DB) of Word Cloud, and derive the problems and usefulness of this method through real-world case analysis. Through this verification result, we present the utility for practical application of word cloud analysis using the proposed subdivision method.

Keywords
Artificial intelligence, AI, deep learning, big data and Word Cloud

1. Introduction
I wanted to propose a new method of improving unused terms using external unused terms in Word Cloud, and derive the problems and usefulness of this method through actual case analysis.
Therefore, the collected raw data removes unnecessary data through heuristic purification and old terms through purification. After that, the vocabulary frequency is calculated, visualized through word cloud technology, the core task is extracted, and the result is analyzed after the information is made.

2. Body
2.1 Word Cloud Analysis Techniques
Word Cloud is a representative technique for analyzing unstructured text data, and manually preprocessed text data is extracted through the R program, frequency is calculated, and visualized and analyzed using Word Cloud techniques. Tags obtained from metadata are analyzed, visually placed in consideration of importance or popularity, and displayed on the website. Usually, tags are placed in the same form as a two-dimensional table, and the order is arranged in alphabetical/alpha order. To emphasize the visual importance, each tag changes its shape, such as the color or thickness of the letter, depending on its importance. The user finds a favorite keyword among the displayed tags and selects it to move to a web page originally connected to the metadata. Words with high frequency of appearance are displayed large, and each word is displayed in a different color. Here, the word with a high frequency of appearance means high importance or high interest. The following is a pre-processing regulation for the number of words according to the frequency type. Frequency calculation result processing is divided into the following five cases.

(1) Frequent and unimportant words
(2) Frequent and important words
(3) Low frequency and low importance words
(4) a word of low frequency and high importance
(5) Frequent but ineligible values

From the five types extracted from the above R program processing, (1) Remove paragraphs (3) and (5) and converge paragraphs (2) and (4) and use a common Korean dictionary words to be corrected are manually added to the dictionary and preprocessed.

2.2. Text Mining

Text mining is the process of mining unstructured data. Mining is the process of extracting statistically meaningful characteristics and concepts from data and extracting high-quality information such as patterns in between. Data composed of structural forms fixed by form are divided into fixed form data, and if there is no fixed structure, it is divided into general data. Mining using fixed data is called data mining, and mining using unstructured data is called text mining.

2.2.1 Characteristic Extraction

Characteristic extraction is the recognition and extraction of important terms from text, which are typically converted into word prototypes to form characteristic vectors. These feature vectors are used as basic information for classifying or demonizing documents, and the calculation of weights and support functions that indicate the importance of that feature is based on the location and number of occurrences of the word. Therefore, the inverse function value of the number of documents in which the word is generated is used to calculate the weight function. The techniques based on this are used to discover and extract information and knowledge of text and are largely classified into three categories: document clustering, classification, and magic. Before classification is performed, clustering is performed first to obtain an overview of the entire document set and obtain classification criteria**.**

1) Topic tracking techniques
Topic tracking is a system that predicts what kind of document a user is interested in based on a user profile. The user's profile may be written according to the user's direct keyword or category or classification of documents read so far. For example, the system allows companies to easily observe trends in competitors, and the medical industry can provide the latest research results and new drug information without missing it.

2) Question answering
It is also known as a question-and-answer system, but when a user asks a question in natural language, the system provides an answer to the question.

3) Duo mining
For example, if you want to apply data mining and text mining together, carriers that want to perform CRM can use data mining technology to analyze customer's monthly call volume to extract customer groups for CRM.

4) Opinion mining
Analyze why people like or dislike certain products or services. It also checks in real time on what issues and how the public's interest changes.

2.3 Web Mining

We extract and analyze interesting and potentially useful patterns, profiles, trends, and explicit information from information collected on the World Wide Web. Information obtained from the web can be used in real time, such as traffic, registration information, and transaction information. It is a very useful technology that can be applied to ERP, CRM, and SCM so that web data can be collected and analyzed in real time to enable true personalization services.

2.4 Analyzing Unstructured Data

1) Text data analysis model
In big data analysis, unstructured text data usually undergo a purification process to convert it into analytical type of data. It is very important to organize it with data suitable for big data analysis, but if it is not properly organized, the analysis is impossible or inconsistent, resulting in poor reliability. Therefore, Figure 1 below shows the purification process of existing unstructured data, and Figure 2 shows the purification process of the proposed unstructured data. In the proposed model, the analyst repeats R program and heuristic post-processing refinement to remove the ineligible values and add data that are not in the Korean dictionary. In addition, results are interpreted and

informatized through word cloud visualization analysis along with post-processed data.

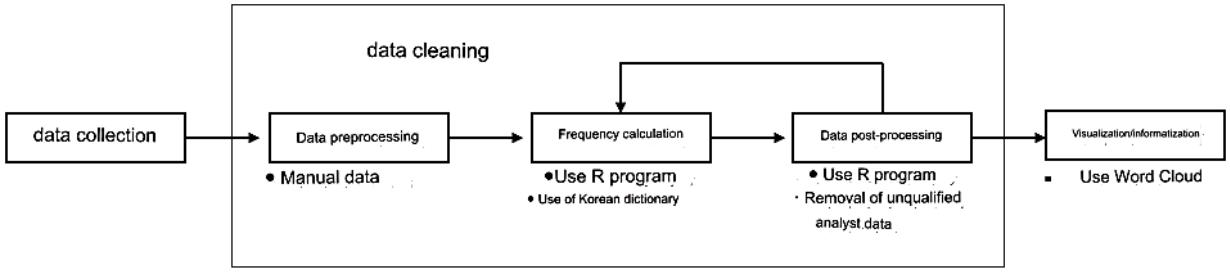

Figure 1. Existing atypical data purification process

The existing process of refining structured data is to process the collected data so that it can be efficiently analyzed. Various types of data collected from various media are converted into processable types, standardized, and integrated and stored. It also checks its usefulness through data quality inspection and continuously manages it.

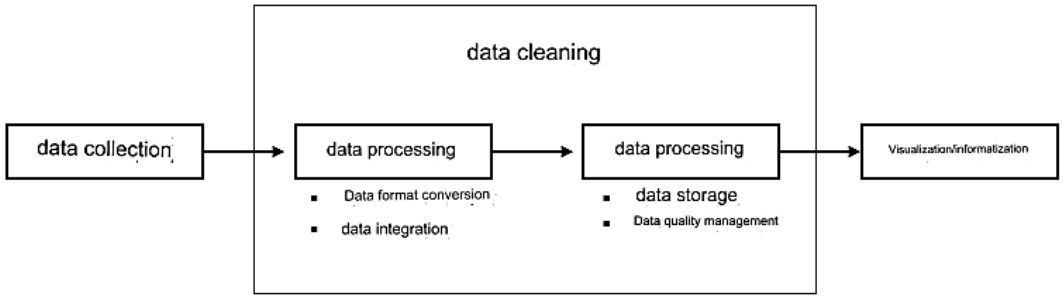

Figure 2. Proposed unstructured data purification model

2.5 Text Mining Methodology

Text mining methodology is a methodology that uses information processing technology and infrastructure to acquire information from text documents such as news and the Internet, analyze keyword patterns, and predict based on them. Text mining is a similar concept to data mining, but traditional data mining can only handle structured data such as relational databases and XML, while text mining separately classifies methodologies for processing unstructured or semi-formatted data such as text documents, e-mails, and HTML files.
Text mining methodologies have been defined by several leading researchers. Most deviant.text mining is defined as the process of discovering new knowledge in large unstructured text populations.

Bae et al. divided text mining into four stages: document collection, document preprocessing, text analysis, and result analysis and refinement. The preprocessing process was again divided into refining unnecessary words or symbols and normalizing the stems of each word to grasp the exact meaning of the sentence. The normalization process was also divided into a morpheme analysis step for extracting the minimum semantic unit from a sentence for Korean processing, a syntactic structure analysis step for understanding the common historical structure, a semantic analysis step, and a context analysis step for sentence analysis. The text analysis process was divided into text clustering, text classification, and text summary. Text clustering has the advantage of being able to analyze results early in the analysis without knowledge of data in the process of dividing text groups into smaller groups according to content similarity.

2.6 Big Data - Data Visualization

Data visualization refers to the effective delivery of information through a visual means called a diagram so that data analysis results can be easily understood. Infographic, which summarizes numerous data into a single picture, and word clouds that visually express the frequency and importance of words used in documents are drawing attention.

(1) Visualize data
Data visualization refers to the effective transmission of information through a visual means called a diagram so that data analysis results can be easily understood. Infographic, which summarizes numerous data into one picture, and WordCloud, which visually represents the frequency and importance of words used in documents, are eye-catching.

Professor Hans Rosling's graphic animation is typical. A professor and statistician at the University of Sweden's medical school, he persuasively expresses changes in each country through a GDP chart that moves with the average life expectancy of the country and over 200 years. The photo shows Korea entering the developed world from a backward country.

Data visualization concepts include information visualization, scientific visualization, visual design, and information graphics. Information visualization generally refers to the visual representation of large amounts of unconventional information, such as file and program code, library bibliographic databases, and Internet relational networks in software systems. Scientific visualization refers to a three-dimensional representation of changes over time in the fields of architecture, meteorology, medicine, and biology. In science, a person who expresses research results in pictures so that the general public can easily understand them is called a visualization scientist. Beautifully represented pictures of planets and galaxies are their works.

Information graphics, also known as infographic, are visual representations of information, data, and knowledge. The key is to express complex information occurring in the fields of signs, maps, media, technical reporting, and education quickly and clearly. Typical infographic is the traffic signs and complex subway lines that we often encounter.

Infographic, which expresses a large amount of data in one sheet, is especially attracting attention in media such as newspapers and broadcasting. It has long been used to graph weather charts that show local weather and statistics in the content of articles. The task of selecting important information from large amounts of data and maximizing visual effects is called data journalism.

(2) Statistical graphic
Friendly defines data visualization as the science of visually representing data and divides statistical graphics and topics into maps. The two fields are similar in that they represent data visually but have different directions. Map visualization focuses mainly on visualization of areas related to space, but statistical graphics are characterized by comprehensive visualization of various fields related to statistical analysis.

Statistical graphics are closely related to data visualization due to the academic nature of statistics that extract meaningful information from data and express it effectively. Various representative attempts in descriptive statistics to effectively represent numerical information include pole plots, cumulative frequency distribution tables, stem leaf plots, and box beard plots.

Tukey, also called Picasso in statistics, emphasized the importance of graphics in statistics. Tuckey, who has also contributed greatly to the field of theoretical statistics, has achieved a lot in the visual representation of statistical data through his book 'Exploration Data Analysis'. Boxes and mustache emoticons were also designed by him. Boxes and whiskers concisely express data characteristics such as minimum, maximum, and quarter in simple box-type pictures.

Multivariate statistics dealing with complex data containing multiple variables are also areas that focus on developing various visual methods to effectively express analysis results. Cluster analysis, one of the multivariate analysis methodologies, is a way to express the process of grouping objects with similar characteristics and to easily grasp the data structure. Decision tree analysis, an analysis method that creates pictures in the form of a tree structure and uses them for classification and prediction, is one of the easy-to-understand visualization methods.

The Chernobyl plane is one of the representative ways to visualize multidimensional statistical data. Each part of the face, such as the forehead, chin, eyes, nose, mouth, and ears, is replaced with variables so that the characteristics of the data can be grasped at a glance.

(3) Word Cloud
Word Cloud refers to calculating the frequency of words used in documents and visually expressing them. A large number of frequently appearing words are displayed, allowing you to grasp the core contents of the document at a glance.

Figure 3. Word cloud

Word Cloud is also called Tag Cloud. A tag is a label attached to clothes and objects to describe materials and handling methods. Keywords added to describe the content of web pages and social network services are called tags. Tag clouds indicate the importance of tags in text size and color on websites.

Depending on the characteristics of the content you want to express, it can be divided into a text cloud and a data cloud. When words contained in a document are visually represented, the data cloud refers to words that represent numerical information in size and color instead of words. For example, the size and color of the country's name vary depending on the size of the population, and the size and color of the company's name are determined by reflecting stock price fluctuations and trading volume in the stock market.

Code analysis is an analysis method that pays attention to the correlation between words. Word analysis is a content analysis technique that examines the rules of words used together in sentences and identifies key concepts and relationships related to the subject of the document. The technology was developed in France in the 1980s. The relationship between words is expressed by converting the frequency of simultaneous occurrence and intimacy between words into indices, and based on this index, it represents a collaborative relationship. For example, if all papers published in academic journals are searched and visualized, the relationship of keywords can be divided into detailed fields such as big data technology, big data use cases, and data scientists. Accordingly, it is possible to easily grasp the trends in the research field.

(4) Data visualization tools
Tools that help you visualize your data include tools you create for data management and graphing, such as Microsoft's Excel and Google's spreadsheets. Google's spreadsheet stores data on Google servers, allowing users to work on computers such as Internet access. You can collaborate with others in real time and create charts that move with time, like Hans Rosling's graph.

Programming languages for specialized analysis include Python and P.A.P., open source processing, and R. R is statistical analysis software, but it is also a free program with excellent statistical graphics.

3. Conclusion
The problems of visualizing and analyzing key issues using R-program word cloud techniques are first caused by the omission of technical terms and new words in the Korean dictionary (KoNLP), secondly, the analyst's poor use of R-program and poor interpretation of third visualization. Heuristic preprocessing and post-processing processes are

very important, which must be extracted from low-frequency but high-importance words, and which are not included in the commercial Korean dictionary are manually added. Therefore, the proposed problem solving and refining model are evaluated to be very useful for increasing the reliability of the verification result data and analysis results, and these measures are meaningful as practical application guidelines for word cloud techniques.
Future research should be conducted on commercial word cloud techniques that can improve the reliability of unstructured text analysis, not interest-oriented or text-design word cloud, by complementing the problems of the development and research results for big data analysis.

References

KO, I. S., "Revisiting Asimov's 3 Laws of Robotics", *Philosophy Research*,vol. 93, pp. 97-120, 2011.
W. Lee, A Study on Word Cloud Techniques for Analysis of Unstructured Text Data, *JCCT,* vol. 6, no. 3, pp. 337-341, 2021.
J. Lee, D. Yun, S. O, C. Lee, *A Big Data Analysis of Civel Complaint Texts Using R Language*, KIICE, 2020.
I. Chun, D. Park, Y. Kang, Python and data science, *Saengneun Publishing*, pp. 222-233, 2019.
M. Chi, S. Lin, S. Chen, C. Lin, T. Lee, *Morphab1e word Clouds for Time-Varying Text Data Visualization*, IEEE, 2015.
Kumar, P. Thakur, K. Gupta, and A. Pal, *Text mining approach to analyse the relation between obesity and breast cancer data*, ILNS, 2015.
M. Han, Y. Kim, C. Lee, Analysis of News Regarding New southeastem Airport Using Text Mining Techniques, *Smart Media Journal*, vol. 6, no. 1, 2017.
Jong Suk Lee and 3 others, *Big data analysis of civil complaint texts using R language*, 2020.
Insun Lee and 1 others, Unstructured data analysis and visualization, *Korean Psychology Association*, 2018.
Dongnyeok Sim, *Research on ICT issue detection and analysis methodology using text data*, 2020.
Software Engineering Center Webzine Materials, *Big data purification process*, 2019.
Giseop Noh, An Analysis on Internet Information using Real Time Search Words, *JCCT*, vol. 4, no. 4, pp. 337-341, 2018.
Jongyong LEE, A Study on Tourism Analysis in Uijeongbu Region Using Big Data, *JCCT,* vol. 6, no. 1, pp. 413-419, 2020.
Sunghuk Moon, Big data environment analysis and research on ways to secure global competitiveness, *JCCT,* vol. 5 no. 2, pp. 361-367
Web Mining, IT Glossary, Korea Information and Communication Technology Association Text mining, *Biochemistry Encyclopedia*
Sejong Oh, R data analysis for everyone, R data analysis for everyone*, Hanbit Media*, 2019.

Biography

Se In Jung is graduates in My Paul School. She is interested in artificial intelligence, deep learning, cryptography, robots, mechanical engineering, automotive engineering, architectural engineering, block chains, drones, autonomous vehicles, etc., and is conducting related research.

Jeongwon Kim is graduates in College of Economics, Nihon University. She is interested in artificial intelligence, deep learning, cryptography, robots, block chains, drones, autonomous vehicles, etc., and is conducting related research.

Shin Dong Ho is Professor and Teacher in MY PAUL SCHOOL. He obtained his Ph.D. in semiconductor physics in 2000. He is interested in artificial intelligence, deep learning, cryptography, robots, block chains, drones, autonomous vehicles, mechanical engineering, the Internet of Things, metaverse, virtual reality, and space science, and is conducting related research.