International Journal of Engineering & Technology, 7 (2.12) (2018) 1-4

International Journal of Engineering & Technology

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

Research Paper

A study on social big data analysis using text clustering

Jin-HeeKu [1] *, Yoon-Su Jeong [1]

[1] Division of Information Communication Convergence Engineering, Mokwon University,
88 Doanbuk-ro, Seo-gu, Daejeon, 35349, KOREA
**Corresponding author E-mail: jhku@mokwon.ac.kr*

Abstract

Background/Objectives: As the use of big data increases in various fields, the use of social big data analysis for social media is increasing rapidly.This study proposed a method to apply text clustering for analysis by related topics of texts extracted using text mining of social big data.
Methods/Statistical analysis: R was used for data collection and analysis, and social big data was collected from Twitter. The clustering model applicable to the related subject analysis of Twitter text was compared and selected and text clustering was performed. Text clustering is analyzed through a cluster dendrogram by generating a corpus, then grouping similar entities from the term-document matrix, and removing the sparse words.
Findings: In this study, text clustering improves the difficulty in analyzing by word association and subject in text mining methods such as word cloud. Especially, in the text clustering model for the related topic analysis of social big data, the hierarchical clustering model based on the cosine similarity was more suitable than the non-hierarchical model for identifying which terms in the tweet have an association with each other. In addition, cluster dendrogram has been found to be effective in analyzing text contexts by grouping several groups of similar texts repeatedly in the visualization process.
Improvements/Applications: This study can be used to confirm ideas and opinions of various participants by using Social Big Data, and to analyze more precisely the complex relationship between the prediction of social problems and the phenomenon.

Keywords: *Text Clustering; Social Big Data; Text Mining; Association Word; Cluster Analysis.*

1. Introduction

The spread of text data in business is overwhelming. 80% of business-related information is consistently generated through call center logs, e-mail, web documents, blogs, tweets, customer reviews, etc., which are mainly unstructured forms such as text[1].Text mining techniques are essential for handling unstructured text data. Text mining generally involves the process of structuring input text, such as parsing, adding derived language features, and removing unnecessary characters[2].A variety of technologies have been studied to summarize and understand the data required to obtain business insight from the rapid growth of social big data such as blogs, the web, and Twitter [3-6]. Therefore, to grasp trend or obtain insight from social big data, it is necessary to group frequent words obtained through text mining on the basis of association and to integrate them by topic.However, a web document composed of several sentences such as a blog may include two or more subjects in a document, and a tweet sentence composed of a short sentence due to the length limitation can be extracted a small amount of information from the text. Therefore, it is difficult to grasp the contextual meaning of keywords by extracting nouns or adjectives included in the text and deducing the entire contents based on the appearance frequency of the words.
This study proposes a topic oriented analysis method consisting of related word group through text clustering which improves social big data analysis method based on word frequency. In the context of text mining, clustering is divided into different groups according to similar subjects. The composition of the paper is as follows. Section 2 describes the research on text mining of social big data. Section 3 describes text clustering for social big data analysis. Section 4 describes the results and discussion, and finally Section 5 describes the conclusions.

2. Related work

The text data of social media is a major big data analysis area in that it extracts recent trends and topics and is able to know trends and topics that are currently trending.
However, the main problem with all this unstructured text data management is that there is no standard rule for writing text so that the computer understands it. Therefore, the meaning of the language differs for every document and every text. The only way to accurately retrieve and organize this unstructured data is to analyze the language and discover its meaning [7].Text mining is used as a representative method for extracting meaningful information from social media data. Text mining is a technique based on natural language processing that extracts patterns or relationships from unstructured text data and finds meaningful information. However, extracting meaningful information from text through text mining requires very specialized tools and techniques and is not simple [8].Preprocessing of text data collected from social media such as web, blog, and SNS is one of the most important and difficult tasks in the text mining process. In this process, unnecessary strings are removed and the text data is structured in a form that can be analyzed. In general, the way to remove unnecessary strings is to use regular expressions and functions that support them. The most important goal of text mining is to extract meaningful words such as

Copyright © 2018 Jin-HeeKu, Yoon-Su Jeong. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

nouns and adjectives by applying natural language processing (NLP) and analysis methods [7], [9]. Table 1 shows the output of the top 30 most frequently used words after preprocessing the tweet text collected from the US president's Twitter 'real Donald Trump'. Figure 1 shows word cloud as a text mining method that visually displays frequently used keywords in text.

Table 1: Term Frequency of Twitter Text

Term	Freq	Term	Freq
great	79	hard	15
tax	43	first	14
people	39	biggest	14
news	33	united	14
fake	30	dems	14
thank	28	china	14
us	28	puerto	14
cuts	27	rico	14
today	22	clinton	13
years	20	fbi	13
time	20	job	13
country	19	day	13
working	19	new	13
big	19	stock	12
crooked	16	hillary	12

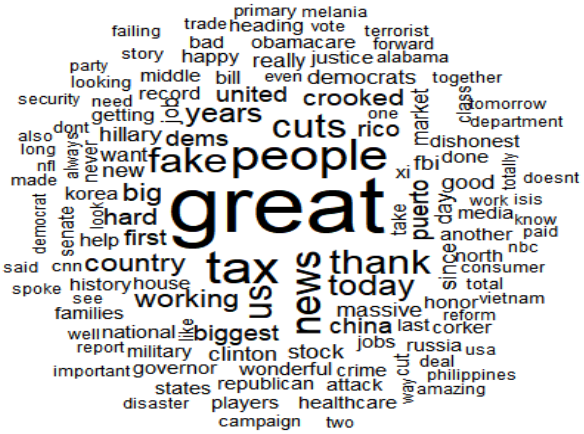


Fig. 1: Word Cloud from Word Frequency.

Figure 1 shows that the words 'great', 'fake', 'people', 'cuts', 'news', 'tax', 'china'. However, the word cloud process often extracts nouns that are not nouns. In this case, it is necessary to draw more meaningful word clouds by repeatedly removing words that are not nouns. In addition, the text mining method based on the frequency of words may not be able to know the word used in a certain context in the task of removing words or it may be difficult to extract insight.

3. Social big data analysis using text clustering

3.1. Clustering model

In this study, we applied the traditional clustering model to the text data to improve the accuracy of the social big data analysis.

The hierarchical clustering model is one of the clustering models that is used variously in various fields of unsupervised learning. Hierarchical clustering model is a method to combine the most similar entities according to the distance between entities and to make them into a binary tree form [10]. Hierarchical clustering models include agglomerative clustering and divisive clustering. Agglomerative model is more common. Agglomerative methods can do a better job of grouping together documents that form small and reasonably cohesive clusters than divisive methods [11], [12]. The agglomerative hierarchical clustering process starts with all the points of the data set at the origin of the clusters at the beginning and continues until the number of clusters reaches the desired number of clusters. Initially, all learning data is a cluster itself. That is, assume that n data pieces are n different clusters. In the next step, each cluster searches all the other learning data to find the closest data to itself, and the two data are grouped into one cluster. Repeat this operation in the same way. If the number of clusters becomes one, the cluster process is completed. The algorithm of the agglomerative hierarchical clustering model is as follows [4], [10], [11].

1) Define all data as a single cluster.
2) Calculate the similarity between clusters.
3) Combine the two clusters with the highest similarity.
4) Repeat steps 2) and 3) until the total number of clusters = 1 is satisfied.

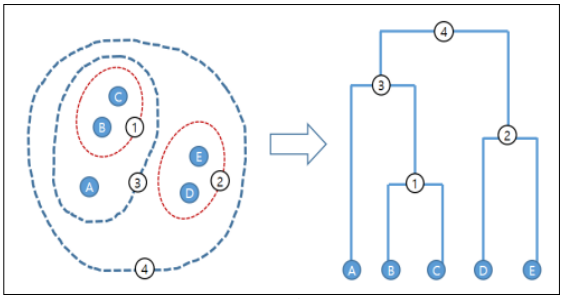


Fig. 2: Conceptual Diagram of Hierarchical Cluster [10].

Figure 2 is a conceptual diagram of the hierarchical clustering model. Text clustering is useful for obtaining high-quality text mining results by classifying Twitter text into several clusters according to association such as interests or hobbies.

3.2. Text clustering process for social big data

Text clustering is applied to group similar texts into several groups by calculating the similarity between the texts, and then to grasp the context to each text group.Text clustering can provide a higher level of text mining analysis results.

3.2.1. Numerical representation of text data

In order to know the similarity between text documents, text documents should be represented by numerical data. TF (Term Frequency) is used to represent the document as numerical data.In equation (1), TF is the number of term appearances in a document. The more often a word appears, the more important it is. However, there is a problem in using only TF as a weight. Some words in the sentence may not have significant meaning. Therefore, if a lot of certain words are found in the whole document, the weight is adjusted by lowering the degree of importance. Table 1 shows TF. If a word appears frequently in the document, it can be considered that the word is related to the document.

$$tf(t,d) = \frac{f(t,d)}{max\{f(w,d):w \in d\}} \quad (1)$$

$$idf(t,D) = log\frac{|D|}{|\{d \in D: t \in d\}|} \quad (2)$$

$$tfidf(t,d,D) = tf(t,d) \cdot idf(t,D) \quad (3)$$

In equation (2), IDF (Inverse Document Frequency) means the number of documents in which a particular word appears in the entire document. When documents are digitized using TF, words appear frequently in the document (for example, the, is, a, are, was, etc.) do not contribute much to the meaning of the document, but they have a large weight because of the frequency of appearance.The TF-IDF in equation(3) is a method of assigning the weighted IDF to the original TF value in consideration of the importance of words in order to solve this disadvantage. Where D is the total number of documents and DF is the number of documents in which the word appears in the document frequency. The smaller the DF value, the greater the importance of the word. The TF-IDF is calculated as the product of the TF value and the IDF value as shown in equation (3).

3.2.2. Similarity matrix

Text clustering computes the similarity between text entities and forms clusters of entities with many similarities. Euclidian similarity and cosine similarity are often used to measure similarity. Equation (4) defines the Euclidean distance based Euclidean similarity. The cosine similarity of equation (5) is a radian-based distance calculation method between two vectors.

$$dist(x \cdot y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (4)$$

$$cos\theta = \frac{x \cdot y}{\|x\|\|y\|} = \frac{\sum_{i=1}^{n}(x \times y)}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{j=1}^{n} y_j^2}} \quad (5)$$

Each of the agglomerative hierarchical clusters becomes one cluster, and a pair of combinations is formed between the respective data, and the similarity in each combination is calculated. Since the clusters have more than two data after the clustering in the first step, the reference points must be set to calculate the distance between the clusters. The agglomerativehierarchical clustering includes single link clustering, full link clustering, average link clustering, and the word link method. Non-hierarchical clustering is a way to include the closest individuals in the center of a set with a predetermined number of clusters. K-means clustering is a typical method. For K-means clustering, the cosine measure is used to calculate the document center closest to a given document.Single link clustering selects the distance between the closest entities belonging to the cluster, the distance between two clusters, the distance between the furthest entities is selected by the complete link clustering, and the average link clustering is the distance between all entities belonging to the cluster Select the average distance.

$$SSE_i = \sum_{j=1}^{n_i}\sum_{k=1}^{m}(X_{ijk} - \bar{X}_{ik})^2 \quad (6)$$

The ward link method is a method that is used most practically and is a method of merging clusters based on the sum of squares of deviations within a cluster, that is, Sum of Square Error (SSE) rather than linking individuals according to the distance between clusters. Equation (6) defines the word-link method.While singlelink clustering and complete link clustering are sensitive to noise and outliers, ward's method minimizes loss of information between clusters and is less susceptible to noise or outliers [8], [10], [11].

4. Results and discussion

4.1. Structuring tweets with text clustering

In this paper, text clustering is performed by applying anagglomerative clustering model for semantic analysis efficiency of social big data. The text is typically mapped into a vector space. That is, a document is represented asthebag-of-words, and each document is a vector using a weighting scheme. Then clustering was performed by measuring the distance between specific vectors.Table 2 shows the Term Document Matrix for TF in Table 1. In Table 2, rows are words and columns are documents (tweets). A total of 394 documents and 1936 words were extracted, indicating the mapping between 10 documents and 15 terms.

There are many ways to calculate similarity between documents, but the most common method is to define it as a cosine measure. The cosine similarity is characterized by the fact that it is not influenced by the size of the vector. The range of values is -1 to 1, and the closer to 1, the more similar. Table 3 shows the distance matrix obtained by the dist () function of R to measure the dissimilarity between observations. Similarity between entities was calculated by applying 'cosine' in the method argument of this function.

Table 2: Term Document Matrixfor TF

	doc1	doc2	doc3	doc4	doc5	doc6	doc7	doc8	doc9	doc10
clinton	1	0	1	2	0	0	0	0	1	0
fbi	1	0	1	1	0	0	0	0	2	0
people	0	1	0	0	0	0	0	1	0	0
stock	0	1	0	0	0	0	0	0	0	0
years	0	0	1	0	0	0	0	0	0	0
fake	0	0	0	0	0	1	0	0	0	0
news	0	0	0	0	0	1	1	0	0	0
country	0	0	0	0	0	0	0	1	0	0
crooked	0	0	0	0	0	0	0	0	1	0
hillary	0	0	0	0	0	0	0	0	1	0
first	0	0	0	0	0	0	0	0	0	1
great	0	0	0	0	0	0	0	0	0	1
job	0	0	0	0	0	0	0	0	0	1

Table 3: Cosine Similarity between Entities

	clinton	fbi	people	stock	years	fake	news	country	crooked	hillary
fbi	0.53333									
people	1.00000	0.96063								
stock	1.00000	1.00000	0.95598							
years	0.94227	0.94227	1.00000	0.87090						
fake	0.81144	0.95286	0.91647	0.89459	0.95918					
news	0.86907	1.00000	0.89689	0.90241	0.96220	0.16676				
country	1.00000	1.00000	0.93003	1.00000	0.94870	0.95811	0.96122			
crooked	0.61270	0.74180	1.00000	1.00000	0.94410	0.90871	0.91548	1.00000		
hillary	0.55279	0.70186	1.00000	1.00000	0.93545	0.89459	0.90241	1.00000	0.13397	
first	1.00000	1.00000	1.00000	1.00000	1.00000	0.90241	0.86447	1.00000	1.00000	1.00000

4.2. Comparing clustering results

In this study, clustering results were compared and analyzed by visualizing the process of merging or separating clusters as a dendrogram. The agglomerative hierarchical clustering algorithms build a cluster hierarchy that is commonly displayed as a tree diagram called a dendrogram[13].When visualizing the results of text clustering as a dendrogram, it is necessary to remove some sparse words to get a simple figure.Sparse means zero or how many terms do not appear. In the case of a matrix of many zeros, the algorithm wastes unnecessary time with unnecessary performance. In this case, efficient data analysis can be done by removing the sparse term. But what is important is how many lines will be deleted. In Table 4, when sparse = 0.97, sparsity = 95%, and the actual data values (‘non-sparse entities’) are 626.This means that the term in the corpus that is not commonly used in at least 97% of the terms in other documents has been removed. However, if sparse is set to 89%, the number of meaningful data can be made 75, which makes it difficult to analyze meaningfully. Although there is no rule for the number

of words suitable for dendrogram the results of text clustering, about 10-30 words are appropriate to simplify the dendrogram. For meaningful analysis, the sparse term should be removed at an appropriate level.

Table 4: Sparse Term Matrix

Sparse	documents	terms	Non-sparse	sparse	total	Sparsity (%)
0.97	394	30	626	11194	11820	95
0.96	394	15	429	5481	5910	93
0.95	394	10	337	3603	3940	91
0.89	394	1	75	319	394	81

Figure 3 shows the dendrogram visualized using the ward link method at sparse 97%. Text clustering visualization using dendrograms can identify which words are related to each other in tweet texts. In R, rect.hclust () function is useful for dividing an appropriate number of clusters in a dendrogram. In this study, we used a scree chart as a way to select the most appropriate number of clusters.

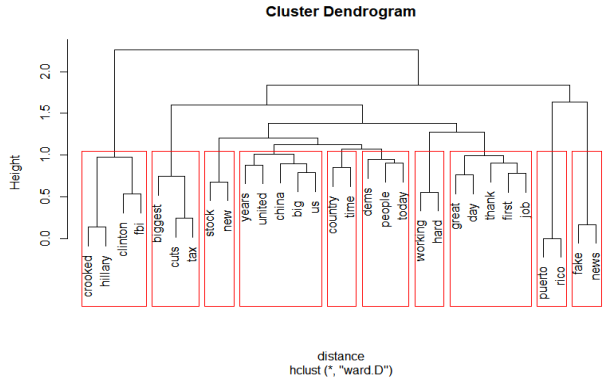


Fig. 3: Dendrogram Using Ward Link Method at Sparse 97%.

Figure 4 is a scree chart for selecting the number of clusters for text clustering in this study. The number of clusters can be selected at the point where the steep slope becomes gentle. Figure 4 shows the number of clusters at the point where there is a sharp increase in the sum of the square of the distance between the clusters (between_SS) / the sum of squares of the total distances (total_SS). As a result, it was found that there was no significant improvement at 84.5% for k = 11 and 80.1% for k = 10.

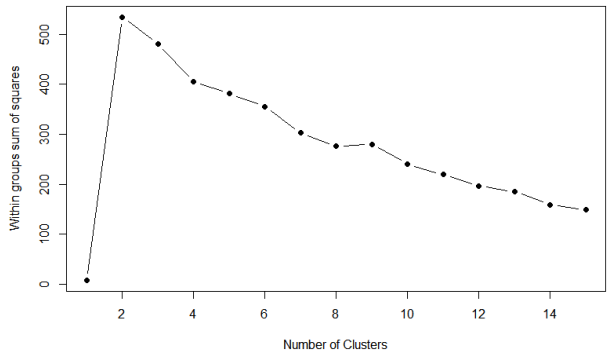


Fig. 4: Scree Plot to Select the Number of Clusters.

5. Conclusion

High-quality text mining technology is required to extract meaningful information from social big data such as web, blog, and Twitter and to improve business operation and performance. This study proposed a topic oriented analysis method consisting of related word group through text clustering which improves social big data analysis method based on word frequency.In this study, text clustering improves the difficulty in analyzing by word association and subject in text mining methods such as word cloud. In the text clustering model for the related topic analysis of social big data, the hierarchical clustering model based on the cosine similarity was more suitable than the non - hierarchical model for identifying which terms in the tweet have an association with each other. But in the case of Twitter text, which has a small amount of information that can be extracted from the text and has a large number of documents, the K-means clustering method shows a large number of clusters but no difference in subject classification. It is necessary to apply the number of clusters based on the scree plot and to derive optimal clustering results because the clustering results may differ depending on the number of clusters.This study can be used to confirm ideas and opinions of various participants by using Social Big Data, and to analyze more precisely the complex relationship between the prediction of social problems and the phenomenon.

References

[1] Chakraborty G, Pagolu M, Garla S, Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS, SAS Institute Inc.:North Carolina, USA, 2013.
[2] Wikipedia, https://en.wikipedia.org/wiki/Text_mining, 2017.
[3] GrossO, Doucet A, Toivonen H, Document Summarization Based on Word Associations, Proceedings of the 37th international ACM SIGIR conference, 2014, pp. 1023-1026.
[4] Park Y M, Kim B G, Kwak S J, Lee J S, Two-Level Clustering for Sub-Topic Labeling of Social Media Data, Journal of KISS : Software and Applications, 2014, 41(3), pp. 225-232.
[5] Gao D, Li W, Zhang R, Sequential Summarization: A New Application for Timely Updated Twitter Trending Topics, Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, 2013, pp. 567-571.
[6] Park W J, Yu K Y, Spatial Clustering Analysis based on Text Mining of Location-Based Social Media Data, Journal of the Korean Society for Geospatial Information Science, 2015, 23(2), pp. 89-96.
[7] IBM, IBM SPSS Modeler Text Analytics 17 User's Guide, IBM Corporation 2003: USA, 2015.
[8] Yu C H, Hong S H, R Visualization, Insight: Seoul, KOREA, 2015.
[9] Vijayarani S, Ilamathi J, Nithya, Preprocessing Techniques for Text Mining - An Overview, International Journal of Computer Science & Communication Networks, 2015, 5(1), pp. 7-16.
[10] Kim U J, Introduction to Artificial Intelligence Machine Learning and Deep Learning, Wiki Books: Seoul, KOREA, 2016.
[11] Steinbach M, Karypis G, Kumar V, A Comparison of Document Clustering Techniques, the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2000, pp. 1-20.
[12] Zhao Y,Karypis G, Comparison of Agglomerative and Partitional Document Clustering Algorithms, 2002, University of Minnesota, Technical Report#02-014, pp. 1-13.
[13] NCSS, Hierarchical Clustering / Dendrograms, http://ncss.wpengine.netdna-cdn.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Hierarchical_Clustering-Dendrograms.pdf.