Journal of the Korean Society for Geospatial Information Science
Vol.24 No.2 June 2016 pp.55-62

ISSN: 1598-2955 (Print)
ISSN: 2287-6693(Online)
http://dx.doi.org/10.7319/kogsis.2016.24.2.055

Research Paper

An Analysis on the Citizen's Health by Using the Twitter Data of Yellow Dust

Jung, Yong Han* · Seo, Min Song · Yoo, Hwan Hee*****

Abstract

Economic and social damages are expected due to yellow dust, occurring every year in Korea and risk of citizens is getting higher accordingly. This study acquired tweet data for yellow dust, which had been the greatest since 2009 for 11 days before and after February 23, 2015. After that, it conducted an analysis on the issue words and association rule. Regarding acquired tweet data, the results of analyzing issue words by using open source R, statistics language shows that 'Mask' was ranked to be the highest, followed by health-related issue words. This indicates that people put the priority in the use of mask for keeping their health, as a result of the occurrence of yellow dust, and subsequently, they showed an interest in diseases, caused by yellow dust. In addition, yellow dust-related diseases, 'cold', 'rhinitis', 'flu', 'asthma', 'bronchitis' were found as issue words, revealing that people had a high concern on the disease occurrence of the respiratory system. The analytical results are judged to reflect the citizen's thought effectively in the process of establishing measures for the prevention of yellow dust.

Keywords : Yellow Dust, Issue Word, Association Rule, R, Health, Disease

1. Research Background and Purpose

With the recent increase in population and development in the industry, air pollution has become serious as a result of increasing automobiles, industry techniques, and heating system, and diversified air pollution. In particular, the seriousness of fine dust has been emphasized according to the lapse of time, and efforts such as the reinforcement of related regulations for air pollutants and the development of control technology have been made(Kim, 2005). However, yellow dust, undergoing a change by the weather conditions, has continued to increase. Since 2012, the duration time of yellow dust has a tendency of being lengthened, affecting the increase in the concentration of fine dust in Spring. Yellow dust blocks and diffuses the sunlight, having a negative effect on people such as the deterioration of the administrative policy, hindrance to the growth of crops, and causing respiratory and eye diseases(Choi, 2014). According to the survey results targeting 12,000 people in five areas across the nation, respiratory disease patients at the time of yellow dust occurrence increased from 11% to 19%, and hospital cost increased due to eye diseases and allergy. On top of that, the death rate on the day of yellow dust occurrence was 3.7%, and the death rate of the first day and the second day after the occurrence of yellow dust increased by 2.1%(Kang, et al, 2004; Harrison et al, 2004). To understand the effect of yellow dust more substantially, what people think about the damage of yellow dust needs to be identified.

In this, relevant data are SNS(Social Network Service) data. SNS is the service in which people can communicate with others online. In particular, twitter users around the world amount to the number of 520 million people, and those in Korea reached 7.5 million people(Yang, 2014; Lim, 2014). The SNS activity of citizen's in accordance with the recent increase of smartphone distribution has increased up to 83%, which became more activated, while the

Received: 2016.05.27, revised: 2016.06.28, accepted: 2016.06.29

* Member · Master Student, BK21+, Dept. of Urban Engineering, Gyeongsang National University, jyh1315@naver.com
** Member · Master Student, BK21+, Dept. of Urban Engineering, Gyeongsang National University, ckdvh203@naver.com
*** Corresponding Author · Member · Professor, BK21+, ERI, Dept. of Urban Engineering, Gyeongsang National University, hhyoo@gnu.ac.kr

private records of communication and information exchange is stored as a form of data(Ha, 2014). After extracting social signal which is inherent in SNS data using text mining or opinion mining, researches visualizing social signal are also increasing(Achreka, et al, 2011; Bollen, et al, 2011).

Therefore, tweet data that mentioned yellow dust for 11 days before and after February 23, 2015, which had been recorded as the date of the greatest yellow dust since 2009 were analyzed by tweet data, R. The analysis extracted issue words that drew people's attention according to the yellow dust to analyze people's thinking about yellow dust-related diseases with a view to utilizing the study results in establishing measures, required for the management of citizen's health.

2. Research Theory and Methods

Tweet data, collected in this study, were analyzed by using R. R is the programming language and open source based statistics language. R community members offered a package which can enhance the function of R periodically. In addition, with the offering of built-in statistical function along with the function of the unique language, R programming language takes up the high recognition in the field of statistical analysis(Choi and Yoo, 2014). This study, conducted an analysis on the issue words and association rule by using the package, offered by the open source based statistics language, R.

2.1 Analysis on the Issue words

The issue words, mentioned the most in the entire twitters, were selected by using the KoNLP package of R to single out the issue words in association with the occurrence of yellow dust. KoNLP package is the package, suitable for dealing with Hangul, among the unstructured data, including a variety of functions such as Hangul extractNoun. By using the KoNLP package, it extracted the nouns of twitter data and visualized the entire words by using the wordcloud package. Therefore, key words, which are high in the occurrence frequency, can be recognized easily by being placed in the center.

2.2 Analysis on the Association Rule

This study conducted an association rule analysis to determine the relevance of each key word within a single tweet. The data, to which an association analysis is applicable, are composed of Transaction and Item. The analysis of association rule in the data mining is the method of exploring a meaningful rule based on Support, Confidence, and Lift in a large number of association rules among the items of two or more than two. Support is the ratio of the transaction, purchasing the item A and the item B during the whole transaction at the same time (Eq. (1)), Confidence is the ratio of the transaction that contains the item B during the purchase of the item A (Eq. (2)), and Lift means the ratio of the transaction, purchasing transaction that includes item B with the transaction, purchasing item B during the purchase of item A(Eq. (3)).

$$\text{Support}(A \Rightarrow B) = \Pr(A \cap B) = \frac{n(A \cap B)}{N} \quad (1)$$

$$\text{Confidence}(A \Rightarrow B) = \frac{\Pr(A \cap B)}{\Pr(A)} = \frac{n(A \cap B)}{r(A)} \quad (2)$$

$$\text{life}(A \Rightarrow B) = \frac{\Pr(A \cap B)}{\Pr(A) \cdot \Pr(B)} = \frac{n(A \cap B)}{r(A) \cdot r(B)} \quad (3)$$

An association rule analysis is the analysis technique, which is commonly utilized in the analysis case of the market basket like "men often buy beer at the same time when they buy diapers." In this study, individual tweets were comparable to the market baskets, and words used in tweets were considered products in the market basket(Yoo and Hong, 2015). In addition, it set Support to represent how often two words are included in one tweet as an important indicator and kept Confidence constant at 60% to do an analysis. Arules package, offered by R was used for the analysis, and the rules of issue words were visualized through the arrows by using arulesViz package for the identification.

3. Result Analysis

3.1 Data Collection

In this study, tweet data referring to yellow dust, occurred for 11 days between February 18, 2015 to February 28, 2015, were collected through the pulse-K to do analyze yellow dust of the day, February 23, 2015, which had been recorded as the greatest date of yellow dust in Seoul since 2009. Pulse-K is the social media analysis & monitoring service, developed by the company, Conan Technology, performing the collection, search, and analysis of big data. In this study, raw data, offered by the pulse-k, were collected according to the dates and data, collected for 11 days were a total of 16,497(Fig. 1).

The tweet data that had mentioned yellow dust were rarely seen from February 18 to February 21, which is the time before the occurrence of yellow dust. The tweet data were the highest in February 23 as 8,751 cases, and significantly decreased after that. This shows that the citizen's interest was drastically changed under the influence of yellow dust.

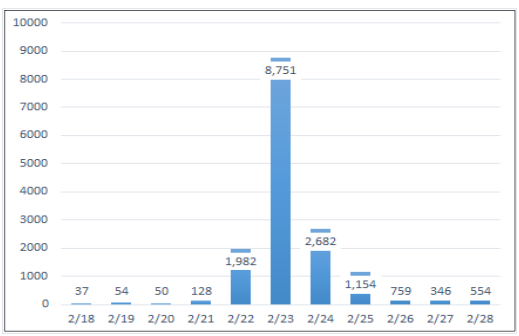

Figure 1. Number of Tweet data

Figure 2. Wordcloud of the entire data

3.2 Analysis of Issue Words

It indicated the entire data, collected for 11 days through R in the wordcloud(Fig. 2). Yellow dust is located in the center, and issue words such as 'mask', 'fine', 'dust', 'health' and 'weather' was identified around the center. To confirm the order of the lists, it organized the number of issue words(Table 1).

Issue words, mentioned the most was 'mask', followed by 'today', and 'dust', shown in Table 1. In this study, 'fine' and 'dust' were combined with two words(fine dust) by the spacing. The percentage of 'dust' was higher than 'fine', regardless of the word, 'fine'. In addition, issue words; 'going out' related to the life of citizens; 'alarm' related to the weather, 'cold' and 'health' related to the weather and health, were found. This demonstrates that citizens have a great interest in the occurrence of yellow dust in many aspects.

3.3 Analysis of Health-Related Issue Words

For a detailed analysis of the health, it classified and rearranged tweet data that include health-related issue words as a group, and conducted an analysis of the issue words by using this data group(Table 2).

Table 1. The number of issue words for the entire data

Issue words(Count)		
Mask(2551)	Today(2418)	Dust(1536)
Fine(1374)	caution(997)	Nation(608)
Going out(554)	Alarm(479)	Weather(430)
Seoul(408)	Brother(395)	Cold(376)
Event(365)	Health(358)	···

Table 2. Health-related issue words

Issue words(Count)		
Cold(376)	Heath(358)	Respiratory(138)
Bronchial(114)	Disease(108)	Health Care(73)
precautionary(63)	Pharmacy(56)	Face(56)
Head(46)	Pollen(38)	Rhinitis(36)
Hospital(35)	Flu(27)	Sore throat(22)
Asthma(16)	Bronchitis(12)	···

Table 3. Issue words related to respiratory diseases

Issue words(Count)		
Cold(376)	Rhinitis(36)	Flu(27)
Sore throat(22)	Asthma(16)	Bronchitis(12)

The study results reveal that ‘cold’ was mentioned the most, and the words of the body part for yellow dust-related diseases, ‘respiratory’ and ‘bronchial’ appeared, and issue words related to health care and treatment were prevention method, 'pharmacy' and ‘hospital’. Also, six diseases such as ‘cold’, ‘rhinitis’, ‘flu’, ‘sore throat’, ‘asthma’ and ‘bronchitis’, which are highly associated with yellow dust appeared. Among the entire issue words, related to the health, six respiratory disease-related issue words were selected(Table 3).

When classifying tweet data based on the six respiratory diseases, ‘sore throat’ was included in the category of ‘cold’ so that it was integrated into the word, ‘cold’. Thus, the association rule analysis was conducted on total five issue words.

3.4 Analysis of the Association Rule in Issue Words of ‘Yellow Dust’ related diseases

An association rule analysis was conducted by using R on the five disease-related tweet data and highly relevant issue words were identified by visualizing them graphically. The support in the analysis of the association rule for each disease-related tweet data was adjusted at the interval of 5% respectively. The meaning that support is 5% is that issue words appeared over five times in 100 sentences. When the number of data was deficient, support was raised to draw the connection of issue words. In the association rule graph, circle, aside from the text, shows the relationship with the issue word, and one line, connecting this relationship represents one rule, and the size of the circle indicates the size of support in the association rule.

First of all, the visualization of the tweet data for the entire disease by applying the support of 5% is shown in Fig. 3. This indicates that ‘cold’, among the diseases is at the center of yellow dust and the rule, revealing the closest correlation with yellow dust.

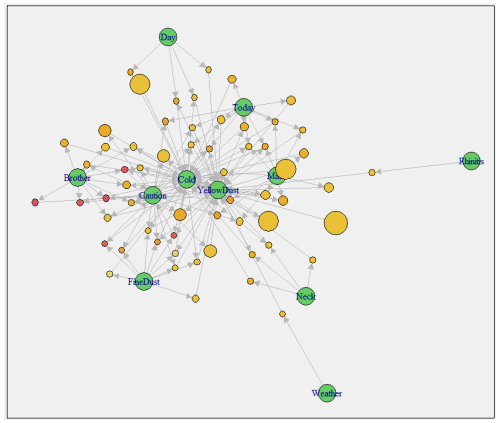

Figure 3. Association rule graph for the entire diseases (support:0.05)

The distribution of issue words; ‘mask’, ‘fine dust’, ‘today’ and ‘caution’ appeared to be connected. ‘Rhinitis’ was shown to have a distant connection. The rest three diseases were not seen when applying the support of 5%, which was judged to have a relatively farther connection. For the detailed analysis, an association rule analysis was conducted on each disease.

3.4.1 Association Rule Analysis on ‘Cold’

The visualization of the tweet data for the ’cold’ by applying the support of 5% in an association rule analysis is shown in Fig. 4. ‘cold’ was located at the

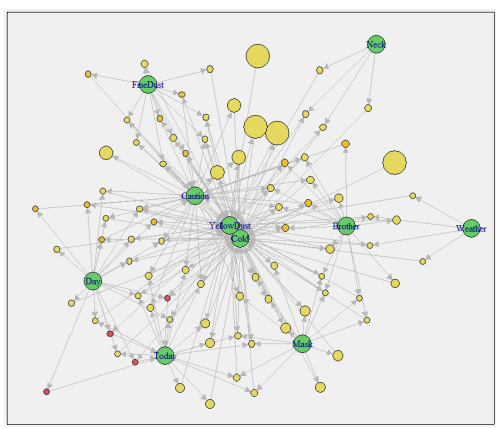

Figure 4. Association rule graph of ‘Cold’ (support:0.05)

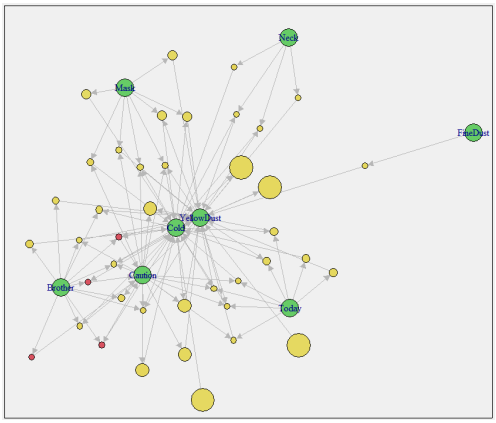

Figure 5. Association rule graph of 'Cold' (support:0.10)

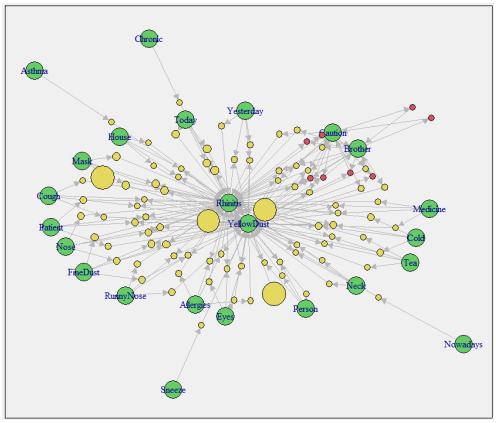

Figure 6. Association rule graph of 'Rhinitis' (support:0.05)

center of the rule along with 'yellow dust', showing a close relationship between cold and yellow dust. Issue words; 'mask', 'fine dust', 'today' and 'caution' appeared to be distributed condense by showing a correlation around the 'cold'. To identify a higher connection, this research analyzed these issue words by applying the support of 10%(Fig. 5). The study results show that 'cold' became closer to 'yellow dust', showing a close connection, followed by 'caution', 'neck', 'mask' and 'fine dust'

3.4.2 Association Rule Analysis on 'Rhinitis'

As a result of an association rule analysis on 'Rhinitis' tweet data by applying the Support of 5% to analyze the correlation between rhinitis-related issue words and yellow dust(Fig. 6), 'rhinitis' and 'yellow dust' were located at the center of the rule, showing the highest connection, followed by 'eyes', 'nose' and 'neck', which are body-related issue words, and issue words which are related to the cause of 'rhinitis', 'allergies' and 'fine dust' appeared. Furthermore, issue words; 'asthma' and 'cold', which belong to the respiratory disease, appeared, and issue words which can identify the symptoms of the cold; 'cough', 'runny nose' appeared as well. To confirm a higher correlation, it conducted an analysis by applying the support of 10%(Fig. 7). The analysis results reveal that 'rhinitis' was closer to the 'yellow dust', followed by 'mask', 'cold' and 'nose'.

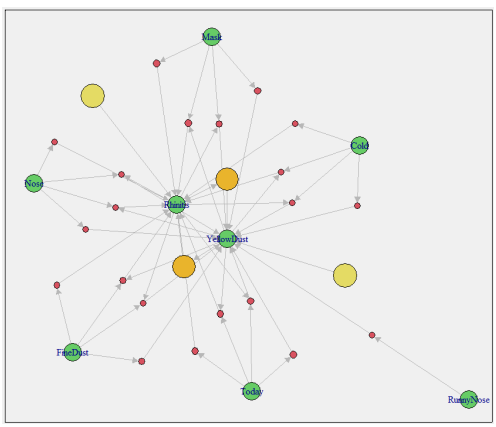

Figure 7. Association rule graph of 'Rhinitis' (support:0.10)

3.4.3 Association Rule Analysis on 'Flu'

Since the number of data for 'flu' is relatively less, an association analysis on flu with the support of 10% was conducted (Fig. 8), in which relatively many rules with a total of 280 occurred, but issue words between rules were relatively less, showing a close correlation with other issue words. 'Flu' and 'yellow dust', had the highest correlation by being located at the center of the rule. These words were highly correlated to the issue words, 'newsroom', 'JTBC', and 'winter' at the top of the left in Fig. 8. It is judged that news or articles have reported the risk of 'winter flue' very significantly. Issue words of 'fine dust' and 'mask' related to yellow dust and

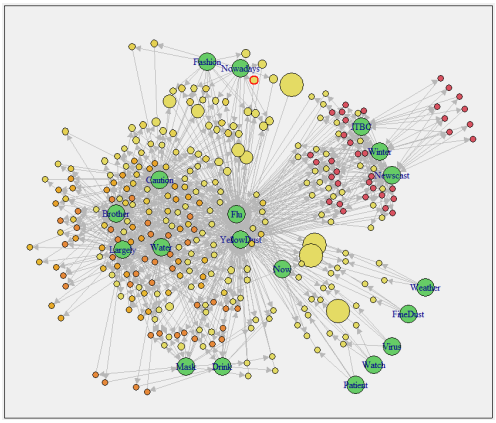

Figure 8. Association rule graph of 'Flu' (support:0.10)

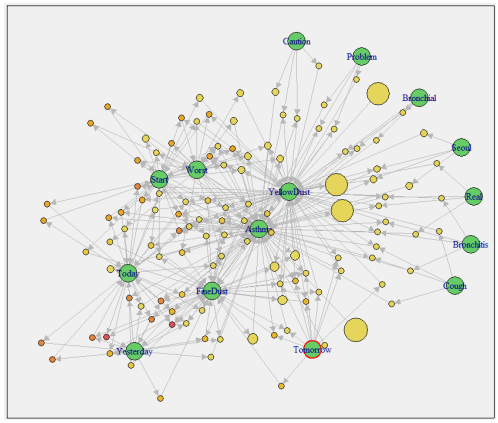

Figure 10. Association rule graph of 'Asthma' (support:0.10)

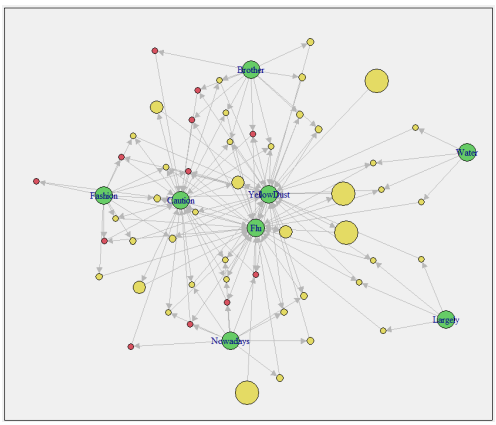

Figure 9. Association rule graph of 'Flu' (support:0.15)

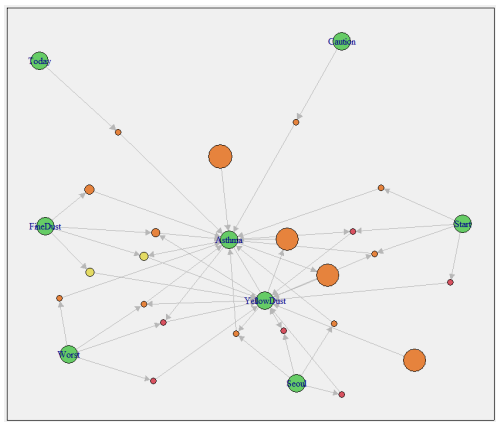

Figure 11. Association rule graph of 'Asthma' (support:0.15)

issue word, 'water' related to the disease prevention, were distributed nearby. For this, we analyzed issue words which are strongly correlated to the 'yellow dust' by increasing support up to 15%(Fig. 9).

The study results were that the correlation of 'flu' with yellow dust was the highest because it was the closest to yellow dust, followed by 'caution', 'trendy' and 'recently'.

3.4.4 Association Rule Analysis on 'Asthma'

Due to the less data in asthma, an association rule analysis of tweet data was conducted in Fig. 10 with the support of 5%. The study results were that 'rhinitis' and 'yellow dust' were located at the center of the entire rules, showing the closest correlation. Time-related issue words of 'yesterday', 'today' and 'tomorrow' were correlated with the issue word 'fine dust' on the left of Fig. 10, while issue words of 'bronchial', 'bronchitis', 'cough' were correlated and distributed nearby. In addition, issue words mentioning 'seoul' were confirmed. To identify a higher correlation, the support increased up to 15% in Fig. 11.

The results were that 'asthma' was the closest to the 'yellow dust,' followed by 'fine dust', 'the worst', 'seoul' and 'start'.

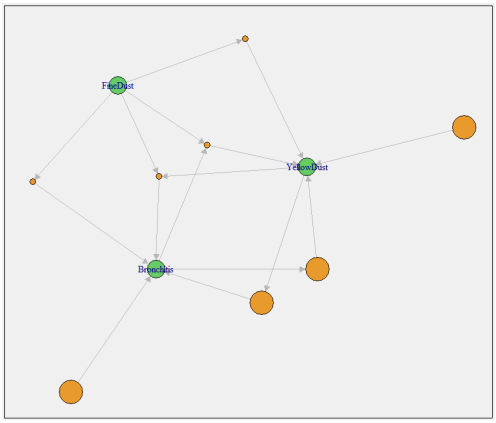

Figure 12. Association rule graph of 'Bronchitis' (support:0.30)

3.4.5 Association Rule Analysis on 'Bronchitis'

Since the relevant data to bronchitis were the least, it conducted an association analysis by increasing support up to 30%. The results were that only three issues; 'bronchitis', 'yellow dust' and 'fine dust' were correlated.

4. Conclusions

This study conducted an issue word analysis and an association rule analysis by using R for tweet data for 11 days before and after February 23, 2015, which had been a date of the greatest yellow dust since 2009 and came to a conclusion as follows:

First, as a result of analyzing tweet data for 11 days around the date of the greatest yellow dust, 'Mask' was ranked the highest in issue words, followed by 'today', 'fine', 'dust' and 'caution'. This shows that citizens think the mask as the most necessary thing to keep their health from the yellow dust. Subsequently, they had a high interest in the complications, caused by yellow dust.

Second, it extracted five issue words; 'cold', 'rhinitis', 'flu', 'asthma' and 'bronchitis' through the search of issue words for the types of disease people are concerned about in relation to yellow dust. This reveals that citizens showed a high interest in the outbreak of respiratory diseases. In addition, the significant concern of people for the relevant diseases to yellow dust was revealed by conducting an association rule analysis on the five respiratory diseases.

In view of the situation that the frequency of yellow dust occurrence has been increasing year by year, mentioned above, it is expected that more effective measures to cope with the yellow dust can be taken when citizen's thoughts are analyzed in the process of establishing prevention methods against yellow dust through the analysis of SNS data like tweet data. Also, this research analyzed the tweet data of the attribute data of the spatial big data, and will be utilized to conduct the spatial analysis in connection with the locational data contained in the tweet data.

References

1. Achrekar, H., Gandhe, A., Lazarus, L., Yu, S. H. and Liu, B., 2011, Predicting Flu trends using Twitter data, Proc. of 2011 IEEE Conference on Computer Communications Workshops, IEEE, Shanghai, China, pp. 702-707.
2. Bollen, J., Mao, H. and Pepe, A., 2011, Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena, Proc. of the 5th International AAAI Conference on Weblogs and Social Media, AAAI, Barcelona, Spain, pp. 450-453.
3. Choi, W., 2014, A study improving observation method of Asian Dust using COMS data, Master's thesis, Cheongju University, pp. 13-14.
4. Choi, K. H. and Yoo, J. A., 2014, A reviews on the social network analysis using R, Journal of Korea Convergence Society, Vol. 6, No. 1, pp.77-83.
5. Ha, B. K., 2015, A study on geotagged SNS data analysis methodology to select the Tweets hotspots, Doctoral thesis, Kwangwoon University, pp. 1-4.
6. Harrison, R. M., Smith, D. J. T. and Kibble, A. J., 2004, What is responsible for the carcinogenicity of PM2.5?, International Journal of Occupational and Environmental Medicine, Vol. 61, No. 10, pp. 799-805.
7. Kang, K. G., Chu, J. M., Jeong, H. S., Han, H. J. and Yoo, N. M., 2004, A study on the analysis of damages from the northeast Asian Dust and Sand Storm and the regional cooperation strategies, Korea Environment Institute, Vol. 2004/re-01, pp. 55-78.

8. Kim, G. H., 2005, Assessment of PM10 and heavy metals compare with Yellow-Sand period and non Yellow-Sand period in the Daegu area, Master's thesis, Kyungpook National University, p. 1.
9. Lim, S. Y., Lim, Y. M. and Lee, J. Y., 2014, Study on the trends of U-City and smart city researches using Text Mining technology, Journal of the Korean Society for Geospatial Information System, Vol. 22, No. 3, pp. 87-88.
10. Yang, M. G., 2014, An awareness identification and preference analysis for domestic university using SNS data, Master's thesis, Chungbuk National University, pp. 1-2.
11. Yoo, C. H. and Hong, S. H., 2015, R visualization, Insightbook, pp. 676-679.