Journal of Engineering and Applied Sciences 14 (11): 3492-3497, 2019
ISSN: 1816-949X

A Study on Clustering News Articles in Korean

[1]Jeong-Soo Kim, [2]Yuchul Jung, [3]Heung-Seon Oh, [1]Kwangyoung Kim and [1]Jungsun Yoon
[1]Division of Advanced Information Convergence,
Korea Institute of Science and Technology Information, Daejeon, Korea
[2]Department of Computer Engineering, Kumoh National Institute of Technology (KIT),
Gumi, Korea
[3]School of Computer Science and Engineering,
Korea University of Technology and Education (KORATECH), Cheonan, Korea

Abstract: The research on news clustering for Korean news articles based on specific issues or topics is very few. This study proposed a method of constructing a correct answer data set for monthly issues for 1 year's worth of Korean news (approximately 4.58 million articles) and the appropriate preprocessing method and clustering algorithm for effective clustering of this news were researched. The clustering results were visualized and the differences among the different methods were comparatively analyzed.

Key words: Korean news, clustering, preprocessing, issue, comparatively analyzed, visualized

INTRODUCTION

Many news articles are created every day and it is not easy for users to quickly find specific articles that they are interested in. News services on portal sites such as naver and daum cluster similar news articles and place them under the same heading. This allows users to quickly identify news topics and easily access detailed news articles after selecting a specific news topic. In addition to being convenient, it offers the advantage of reducing time wasted in repetitive searches. However, many issues need to be considered before applying the clustering method to efficiently process vast amounts of news.

The news section of the widely used web portal naver serves approximately 250,000 news articles in 1 month. Applying the general clustering method directly to such a vast number of news articles is not easy because it requires excessively large amounts of memory and computing powers. Many data handling methods exist for solving this problem. Dimensionality reduction relieves the problem by expressing large quantities of data using a small number of features by reducing the number of features. Another problem is that naver web portal news articles contain many unnecessary sentences (introduction of newspaper or reporter, advertisements and related news) and it often happens that specific words or sentences appear in almost all news articles. This characteristic of Naver news inhibits the clustering effectiveness. Therefore, this study proposed a method of removing the news characteristics that inhibit clustering effectiveness and verified its performance through comparative tests among different clustering algorithms. In this study, the performance of the clustering algorithm was verified using 4.58 million Korean news articles collected between January and December 2016. Because it is inefficient to use more than 250,000 news articles per month for performance verification, a data set for verifying the performance of the proposed clustering algorithm was created and used for the experiments. The proposed data preprocessing method was applied to improve the performance of the clustering algorithm and a performance improvement of up to 70% was verified using the silhouette coefficient score.

Literature review: To cluster text documents, the documents are expressed as features in a feature space and entered into a clustering algorithm before the clustering is carried out (Aggarwal and Zhai, 2012). Similarity is compared between documents or between document and cluster; documents having a high similarity are allocated to the same cluster and documents that have low similarity are allocated to different clusters. Documents are expressed as vectors consisting of a pair of feature and feature value. Features are expressed as words based on the bag-of-words model and feature values are calculated as weight values that can properly express the characteristics of the feature such as TF*IDF.

Corresponding Author: Yuchul Jung, Department of Computer Engineering,
Kumoh National Institute of Technology (KIT), Gumi, Korea

Clustering algorithms can be classified into many categories (Aggarwal and Zhai, 2012; Fahad *et al*., 2014). They can be divided into hierarchical clustering and flat clustering algorithms according to whether the data are clustered hierarchically. Hierarchical clustering in turn can be classified into top-down divisive clustering which places all documents in the data into one cluster and then subdivides them and bottom-up agglomerative clustering which creates one cluster for each document and combines them. A hierarchy of upper and lower levels is generated for the clusters created.

The feature expressions based on the bag-of-words model have a very large vocabulary. Not all words in the vocabulary are effective for use in clustering. Furthermore, using a large vocabulary requires a large memory and many calculations. These can be reduced by selecting useful words through feature selection. The widely used Latent Semantic Analysis (LSA) method is a feature selection method based on principal component analysis (Deerwester *et al*., 1990). In addition, there are many effective feature selection methods for text document clustering (Yang and Pedersen, 1997; Liu *et al*., 2003). In most cases, a feature selection method based on unsupervised learning such as term strength and entropy-based ranking is used. However, such feature selection methods require many calculations to select a feature because the similarity between all features or documents must be compared and there is no correct answer label.

By Park *et al*. (2014), stop words were collected manually from Korean news articles and used for news clustering. Manual collection of stop words takes a long time and requires much effort and stop words must be collected again for news that have content different from the existing news which is cumbersome. Therefore, this study proposed a stop word selection method that considers the number and length of news items.

This study has laid a foundation for increasing the clustering effectiveness of large amounts of Korean news by proposing a simple, effective feature selection method based on an understanding of the characteristics of Korean news items.

MATERIALS AND METHODS

This study proposed two methods. The first method was one to create a data set for comparing clustering performance and the second method was a data preprocessing method to improve the clustering effectiveness when using the conventional clustering algorithm.

First, the method for creating a data set for comparing clustering performance is as follows. Naver's news section which was the object of clustering in this study, served more than 4.58 million news articles for 1 year in 2016.

As shown in Table 1 in order to create a data set for comparing clustering performance, news articles from four of the categories (politics, society, culture and sports) having the largest monthly average number of news articles are selected first. The news articles that have a similarity above the standard value (in this study, threshold = 0.75) based on the cosine similarity in each category are selected for each group. Among these groups, 10 news articles from each of 10 groups are selected in descending order of the number of news articles that belong to the group. By performing this task for the four categories, a data set for comparing clustering performance that has around 400 articles per month is created. The appropriate number of clusters to be created during clustering each month was derived using the elbow method (Raschka, 2014).

Next, there are two preprocessing methods for improving the clustering effectiveness. As shown in Fig. 1, the first method is the Purging of Dispensable Sentences (PDS) which removes unnecessary sentences from the news data and the second method is the Stop Word Selection (SWS) which selects stop words to be excluded from clustering.

The PDS method removes unnecessary sentences from the news data. As shown in Fig. 2, news articles contain unnecessary sentences that interfere with clustering such as the newspaper name, reporter's name, advertisements, e-mail addresses and related news. If such unnecessary elements exist in many news articles, the clustering effectiveness is reduced. To solve this problem, this study proposed the PDS method which removes specific content in the top and bottom parts of the news items.

First, the top part removal method removes the top part up to the point where a specific symbol or word first appears. This removes the unnecessary parts at the top. Next, the bottom part removal method removes all data starting from the point where a specific symbol or e-mail address first appears. This removes all unnecessary parts at the bottom.

Table 1: Statistics of news articles in 2016

Subject area	Number of news articles	Share (%)
Politics	753,275	16.4
Economy	760,103	16.6
Society	981,394	21.4
Life/culture	334,771	7.3
World	862,843	18.8
IT/science	174,992	3.8
Entertainment	2,492	0.1
Sports	715,367	15.6
Total	4,585,237	

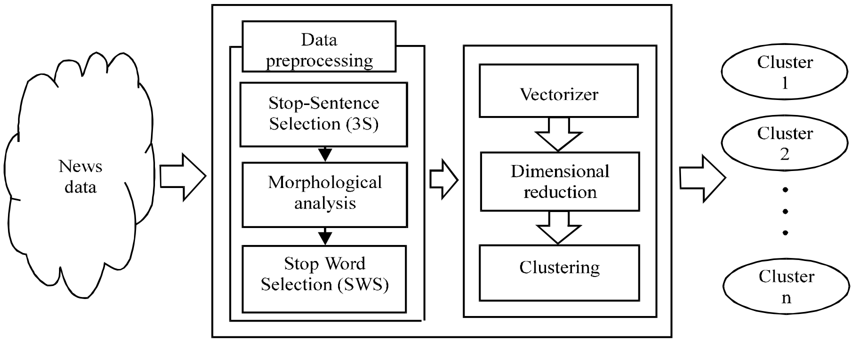
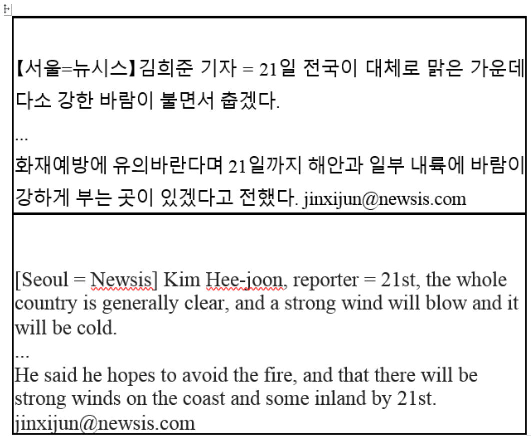


Fig. 1: Proposed clustering method



【서울=뉴시스】김희준 기자 = 21일 전국이 대체로 맑은 가운데 다소 강한 바람이 불면서 춥겠다.

...

화재예방에 유의바란다며 21일까지 해안과 일부 내륙에 바람이 강하게 부는 곳이 있겠다고 전했다. jinxijun@newsis.com

[Seoul = Newsis] Kim Hee-joon, reporter = 21st, the whole country is generally clear, and a strong wind will blow and it will be cold.

...

He said he hopes to avoid the fire, and that there will be strong winds on the coast and some inland by 21st. jinxijun@newsis.com

Fig. 2: Example of news text

In order to apply the SWS method, a Korean morphological analyzer is used to extract only the nouns from news articles that have been refined by applying the PDS method. Then, the SWS method which selects the stop words is applied by using the words extracted. For the SWS method, the frequency of each word in the entire content is calculated and the words are sorted in descending order of frequency. As shown in Eq. 1, the words whose appearance frequency is above a certain value are selected as stop words, taking into account the total number and length of the news articles. The value 500 which showed the best performance in multiple experiments was selected as the value for n in Eq. 1.

$$\text{Threshold} = \frac{\text{Number of news} \times \text{Avg. number of word}}{n} \quad (1)$$

RESULTS AND DISCUSSION

In this study, the preprocessing method used for news clustering, the implementation environment, the evaluation criterion and the experiment results, including the results for specific months are described in detail.

Preprocessing methods and clustering algorithms: Data preprocessing methods included PDS and SWS. In this experiment, the results of applying neither, the results of applying PDS only and the results of applying both PDS and SWS were compared.

For the implementation environment, Python Version 2.7 was used as the development language. For the vectorizer, dimensionality reduction and clustering algorithms, the modules provided by scikit-learn (Nelli, 2015) which is a Python library were used.

For the vectorizer, hashing vectorizer from the Scikit-learn library was used and features were extracted by performing Inverse Document Frequency (IDF) normalization. For dimensionality reduction, SVD (Singular Value Decomposition, a type of latent semantic analysis) from the Scikit-learn library was used. The algorithms used to compare the performance of clustering algorithms were KM (K-Means) (Hartigan and Wong, 1979; Lee, 2012), SC (Spectral Clustering) (Ng *et al*., 2001) and AC (Agglomerative Clustering) (Zhao *et al*., 2005).

Evaluation criterion: As the evaluation criterion, the silhouette coefficient score of each algorithm was used to compare the performance of the algorithms for news documents to which the proposed preprocessing method was applied. The Silhouette coefficient score is a metric from the Scikit-learn library it is calculated for each sample by Eq. 2 using the average distance inside the cluster (A) and the average distance to the nearest cluster (B).

$$\text{Silhouette coefficient score} = \frac{\text{B-A}}{\max(\text{A,B})} \quad (2)$$

The closer the silhouette coefficient score is to 1, the better and the closer to -1, the worse. If the silhouette coefficient score is close to 0, it means there are many overlaps among the clusters.

Table 2: Average silhouette coefficient scores

Variables	None	PDS	PDS+SWS
KM	0.109	0.158	0.201
SC	0.132	0.173	0.225
AC	0.14	0.147	0.204

Experiment results: Table 2 shows the average silhouette coefficient score for each method in each algorithm. For each algorithm, the case in which both the PDS and SWS methods were applied shows the highest silhouette coefficient score, followed by the case in which only PDS was applied. The lowest silhouette coefficient score was obtained when neither the PDS nor the SWS method was applied. The difference between the case in which neither PDS nor SWS was applied and the case in which both PDS and SWS were applied was approximately 84%. The difference between the case in which neither PDS nor SWS was applied and the case in which only PDS was applied was approximately 45%.

In a comparison of the silhouette coefficient scores for each algorithm, SC generally showed a high silhouette coefficient score. In particular, the silhouette coefficient score of SC for the case in which both PDS and SWS were applied was 0.225 which is the highest among all the values. On the other hand, the silhouette coefficient score of KM for the case in which neither PDS nor SWS was applied was 0.109 which is the lowest among all the values.

However, when only the silhouette coefficient score is considered, the change in the silhouette coefficient score depends much more on the preprocessing method applied than on the algorithm applied.

As shown in Fig. 3, all values <0.38. This means that there are many overlaps among the clusters when news data are clustered. Owing to the nature of Korean Naver news, there are many unnecessary or overlapping parts (related news, advertisements, etc.,) resulting in high overlap between clusters. When cases in which both PDS and SWS were used and those in which neither were used are compared, the difference in silhouette coefficient score is generally large.

Although, there are differences between months, the cases in which both PDS and SWS were applied show a higher silhouette coefficient score than the other cases. In particular, the difference in silhouette coefficient score between the case in which both PDS and SWS were applied and the other cases was greatest in February.

Detailed analysis: In a comparison of the monthly silhouette coefficient scores between algorithms, for SC, the cases in which only PDS was applied had a very large monthly variation in the silhouette coefficient score. However, in the cases in which both PDS and SWS were applied, the monthly variation in the silhouette coefficient score of SC was less. The SWS method which removes nouns that appear too frequently or too rarely, improved the performance of news clustering and showed stable clustering results. As for AC, the difference in silhouette coefficient scores between the cases in which only PDS was applied and the cases in which neither PDS nor SWS was applied was not large. In other words, in the case of AC, it is difficult to expect improved performance of news clustering using only PDS which removes the unnecessary or overlapping parts in the top and bottom parts of the news articles.

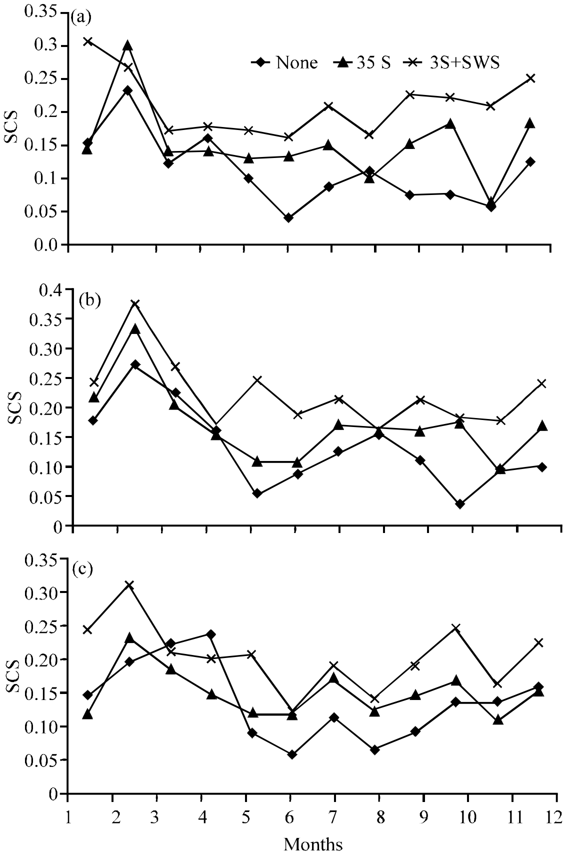


Fig. 3: Silhouette coefficient scores for each algorithm: a) k-means; b) Spectral clustering and c) Agglomerative clustring

In a comparison of the silhouette coefficient scores by month, February 2016 shows large differences in silhouette coefficient score between the methods and August 2016 shows relatively small differences in silhouette coefficient score. In this section, therefore, the news clustering results for February and August are examined in more detail.

Figuer 4 shows graphs indicating the ratios of documents classified in correspondence with the group of each document classified when the data set was created

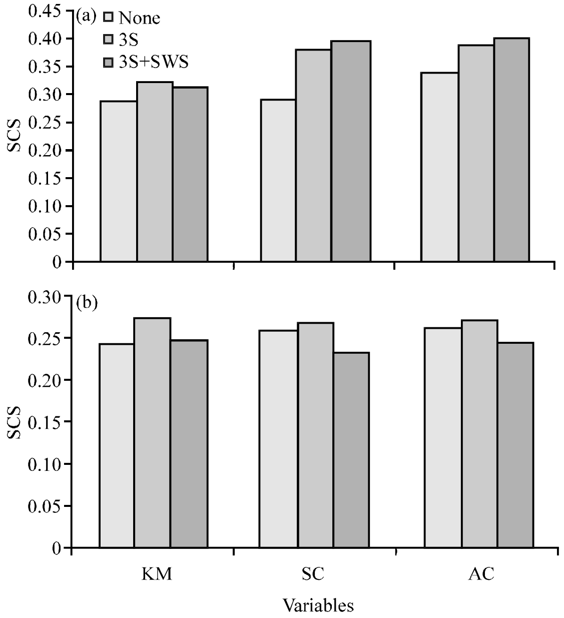


Fig. 4: Silhouette coefficient scores for February and August: a) February 2016 and b) August 2016

after clustering the news articles of February and of August 2016. In the February results, KM shows generally low success rates and the overall success rate is around 30%. On the other hand, SC and AC show higher success rates than KM. In particular, the cases in which the PDS and SWS methods proposed in this study were used and the other cases show large differences. When the cases in which only PDS was applied and the cases in which both PDS and SWS were applied are examined, SC and AC show almost identical success rates. However, in the cases in which neither method was applied, AC shows a higher success rate than SC.

When the August results are examined, it can be seen that all cases show low success rates. The differences in success rate between algorithms also are not large. In other words, the news data of August 2016 had very low clustering success rates regardless of which method or algorithm was used. In August 2016, the Olympics were held, it was a very hot Summer and the Sewol Ferry Special Law was being processed. Because almost all news articles in each category were related to these issues, the differences between news items were not large and as a result, all clustering methods failed to show good performance.

Figuer 5 shows graphs comparing the correct answer set when both PDS and SWS were applied to Naver news articles and the clustering results for each algorithm in February 2016. The positions of the clusters are different in the correct answer set and the clustering algorithm performance results because some features of the news data were changed by applying the PDS and SWS methods.

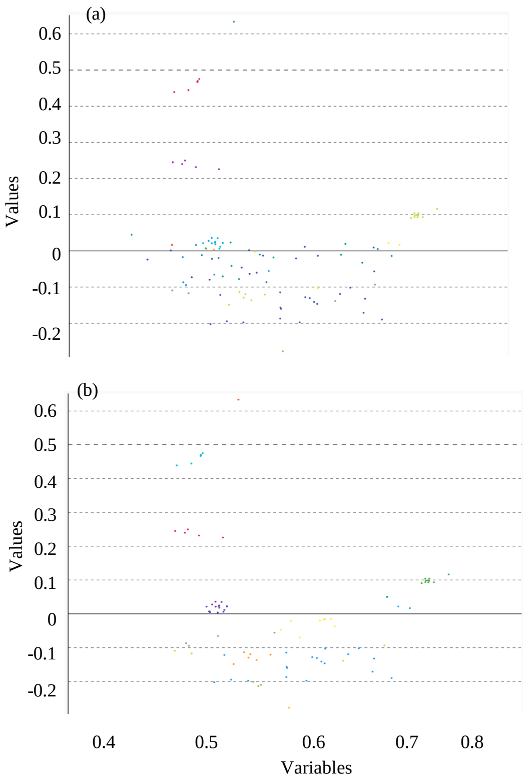


Fig. 5: Details of clustering results: a) Clusters found by k-means and b) Clusters found by spectral clustering

As shown graphically in Fig. 4, application of the SC algorithm resulted in a success rate of around 40% but KM shows large differences from SC. The success rate of KM was also lower by around 30% than that of SC as also shown in Fig. 4.

As confirmed in the above discussion, the performance of news clustering decreased when neither the PDS nor the SWS preprocessing method was applied. This is because of the nature of Naver news: content unrelated to the news such as advertising, the newspaper name and related news occurs repeatedly in various news articles. The PDS method proposed in this study was applied to solve this problem and the news clustering performance improved by about 9%. In addition, the SWS method which removes stop words that have a negative effect on clustering was applied, achieving a final performance improvement of around 10%.

CONCLUSION

This study experimentally examined which methods are required to automatically cluster major Korean news articles on issues in each month and which clustering algorithm is most effective. The performance of clustering Korean news data is considerably lower than that of the known results of clustering English news data.

RECOMMENDATIONS

To improve the clustering performance, methods that consider vocabulary features at a more semantic level are required in addition to methods to filter out words causing noise. In addition, further research should be conducted on methods that can automatically extract features that are effective for distinguishing different issues and make effective use of them during clustering.

ACKNOWLEDGEMENT

"This research was supported by Korea Institute of Science and Technology Information (KISTI)".

REFERENCES

Aggarwal, C.C. and C.X. Zhai, 2012. A Survey of Text Clustering Algorithms. In: Mining Text Data, Charu, C.A., and X.Z. Cheng (Eds.). Springer, New York, USA., ISBN:978-1-4614-3222-7, pp: 77-128.

Deerwester, S., S.T. Umais, G.W. Furnas, T.K. Landauer and R. Harshman, 1990. Indexing by latent semantic analysis. J. Soc. Inform. Sci., 41: 391-407.

Fahad, A., N. Alshatri, Z. Tari, A. Alamri and I. Khalil *et al.*, 2014. A survey of clustering algorithms for big data: Taxonomy and empirical analysis. IEEE. Trans. Emerging Top. Comput., 2: 267-267.

Hartigan, J.A. and M.A. Wong, 1979. Algorithm AS 136: A K-means clustering algorithm. J. R. Stat. Soc. Ser. C. Applied Stat., 28: 100-108.

Lee, S., 2012. Comparison of initial seeds methods for K-Means clustering. J. Internet Comput. Serv., 13: 1-8.

Liu, T., S. Liu, Z. Chen and W.Y. Ma, 2003. An evaluation on feature selection for text clustering. Proceedings of the 20th International Conference on Machine Learning (ICML-03), August 21-24, 2003, ICML AAAI Press, Washington DC., California, pp: 488-495.

Nelli, F., 2015. Machine Learning with Scikit-Learn. In: Python Data Analytics, Fabio, N. (Ed.). Apress, Berkeley, California, USA., ISBN:978-1-4842-0959-2, pp: 237-264.

Ng, A., M. Jordan and Y. Weiss, 2001. On Spectral clustering: Analysis and an algorithm. Adv. Neural Inform. Proc. Sys., 14: 849-856.

Park, Y., B. Kim, S. Kwak and J.S. Lee, 2014. [Two-level clustering for sub-topic labeling of social media data (In Korean)]. J. KISS. Software Appl., 41: 225-232.

Raschka, S., 2014. Python Machine Learning. Packt, Birmingham, England, UK.,.

Yang, Y. and J.O. Pedersen, 1997. A comparative study on feature selection in text categorization. Proceedings of the 14th International Conference on Machine Learning (ICML'97) Vol. 97, July 08-12, 1997, Morgan Kaufmann Publishers, San Francisco, California, USA., ISBN:1-55860-486-3, pp: 412-420.

Zhao, Y., G. Karypis and U. Fayyad, 2005. Hierarchical clustering algorithms for document datasets. Data Min. Knowl. Discov., 10: 141-168.