Indian Journal of Science and Technology, *Vol 9(24), DOI: 10.17485/ijst/2016/v9i24/96148, June 2016*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Comparison of Performance in Text Mining using Categorization of Unstructured Data

Lee Junyeon[1]*, Shin Seungsoo[2] and Kim Jungju[3]

[1]Department of Media Engineering, Tongmyung University, Korea; jylee@tu.ac.kr
[2]Department of Information Security, Tongmyung University, Korea; shinss@tu.ac.kr
[3]Choonhae College of Health, Korea; jj790105@hanmail.net

Abstract

Background/Objectives: The text mining would help finding information to the users in the enormous documents. The text mining has been actively developed and utilized in various fields, mainly English-based document, but Study on the Korean text mining has been relatively limited. The importance of the Korean text mining has emerged with increasing big data including Korean text data, the needs for the intensive study and application of Big Data are increasing. **Methods/ Statistical Analysis:** In this study, we compared the performance of these classifications by applying the method of Bayesian methods, k-NN, decision trees, SVM, and as a neural network in classification of unstructured newspaper article into given categories. **Findings:** In the experiment result, the SVM model has a high F-measure value relative to other models, and has shown stable results in the classification information and recall rate. Also, this model showed a high F-measure value in the classification of a more granular list. **Application/Improvements**: The methods of k-nn and decision tree show slightly lower performance than SVM, they are turned out to be appropriate models using classification problem cause of having advantages to easy interpretation and short learning time.

Keywords: Categorization, Decision Tree, k-NN, Naive Bayes, Text Mining

1. Introduction

The goal of information access is to help users find documents that satisfy their information needs. The standard procedure is akin to looking for needles in a needlestack - the problem isn't so much that the desired information is not known, but rather that the desired information coexists with many other valid pieces of information. The goal of data mining is to discover or derive new information from data, finding patterns across datasets, and/or separating signal from noise. The fact that an information retrieval system can return a document that contains the information a user requested does not imply that a new discovery has been made: the information had to have already been known to the author of the text; otherwise the author could not have written it down[1].

A large database treated in data mining can be classified into an unstructured db and structured db according to the structure. Data mining has been developed mostly focused based on the structured database techniques up to now. Therefore, data mining investigators were aware of the need to research on the unstructured databases, there have recently been increasing studies with them[2].

Text mining can be divided into data processing and data analysis. Data processing is a step of machining to facilitate the unstructured data to the data analysis, and data analysis is extracting meaningful information from the text using data mining, machine learning, statistics.

<Table 1> shows the classification of data mining and text data mining applications. The researches in text mining of textual material are in progress due to the increasing of amount of data. Extracting a pattern or relationship from the text data, such as web pages, electronic documents, electronic mail, a process of managing the new information, the classification techniques to mining techniques to the text data based on a given keyword assigning article and without advance information, there are similar binding document clustering techniques to group. Document

*Author for correspondence

Table 1. A classification of data mining and text data mining applications

	Finding Patterns	Finding Nuggets	
		Novel	Non-Novel
Non-textual data	standard data mining	?	database queries
Textual data	computational linguistics	real Textual Data Mining	information retrieval

classification scheme is classified according to a given keyword, and according to a given set of keywords, the document is determined whether classified into the category or not. Document classification allows convenient access to the user a number of documents by structuring categories automatically. So, it has emerged as a very important factor for efficient information management and search.

Utilizing a document classification scheme of text mining in previous studies using a neural network and k-NN classification into four categories, it was compared to the performance[3].

In this paper, we classify the big data consisting of Hangul and compared the performance between the various classification techniques. In chapter 2, we describe how to formalize unstructured text data. And chapter 3 says theoretical background of various classification techniques, and chapter 4 extracts the result by applying real data. Finally chapter 5 presents the conclusions.

2. Text Mining

The purpose of text mining is to find useful patterns from the large documents and it means to apply an algorithm to the method of the text data and the statistical machine learning. In the aspect of finding a pattern and extracting the information from a big data, text mining is similar to data mining, except using unformulated data.

2.1 Extracting Keywords

Words that reflect the content and features of a document called the content word. First, we determines each part of speech by analyzing the morpheme of the sentence in order to extract the content word. Statements include various word of part of speech, and the noun is used to introduce and describe the concept of a statement. The noun is the most appeared to cent word, so we can review it the most weight.

Figure 1 is a figure showing a process to conduct a morpheme analysis to extract a keyword corresponding to the noun. In the morphological analysis, the scheme excluded the postpositional particle, verb, etc., and extracted only the classification word as a noun. And then, the extracted nouns are structured to apply a statistical analysis technique.

Because stop words do not give any information with appearing in several or every documents, the extracted nouns except with the stop words will be used.

2.2 Extracting General Format

Document should convert into structured data to be suitable for analysis because the unstructured data. A document can be expressed as a matrix of columns indicating the line that represents a keyword extracting a set of documents by using the concept of a set of words and documents[4]. This matrix is called word-document matrix, it can be expressed in the form of a matrix of m n words and documents, such as Table 2.

In the Table 2, f_{ij} means the frequency of the i-th word is included in the j-th document.

2.3 Weighted Function

To indicate a better correlation of words, word-document matrix it can be converted by using the weight function.

Text mining roughly equivalent to text analytics, refers to the process of deriving information from text. High–quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' in text mining usually refers to some combination of relevance, novelty, and interestingness.

text	Frequency
Text mining	3
pattern	3
high-quality	3
process	2
information	2
analytics	1

Figure 1. .Example of a keyword extraction.

Table 2. The World-Document Matrix

	Doc 1	…	Doc j		… Doc n
Word 1	f_{11}	…	f_{1j}	…	f_{1n}
…	…	…	…	…	…
Word i	f_{i1}	…	f_{ij}	…	f_{in}
…	…	…	…	…	…
Word m	f_{m1}	…	f_{mj}	…	f_{mn}

Weight function is composed of local weight function and global weight function[5].

Elements of the transformation matrix is equation (2.1) using a word i with the document j, as can be calculated as the product of a function of both weight L(i,j) represents a local weight function and G(i) represents global weight function.

$$a_{ij} = L(i, j) \times G(i) \tag{2.1}$$

The local weight function uses a simple frequency, or a simple log conversion or binomial conversion form generally.

Simple frequency function :

$$L(i, j) = f_{i,j}$$

Log conversion function :

$$L(i, j) = \log_2(f_{i,j} + 1)$$

Binomial conversion function :

$$L(i, j) = \begin{cases} 1 & f_{i,j} \geq 1 \\ 0 & f_{i,j} = 0 \end{cases}$$

Regional weight function gives weighted value to a word *i* included in document *j* is represented as *L(i,j)*.

The weights of the words *i* for the entire document referred to globally weight function, and represented as *G(i)*. This is also known as the term weights, and are used to supplement the disadvantages of regional weight function does not reflect the characteristics of the words for the entire collection of documents[6].

Global weighting function has a function using the function, the function representing the words the information that appears in quantity, function and standardized concept considering the frequency and the word is the ratio of the indicated number of documents in the word in the entire document using the entropy concept as follows, then and it is expressed as follows.

The function using entropy:

$$G(i) = 1 + \sum \frac{\left(\frac{f_{ij}}{g_i}\right) \log_2 \left(\frac{f_{ij}}{g_i}\right)}{\log_2(N)}$$

The function of the word by quantitative:

$$\mathrm{G(i)} = \log_2 \left(\frac{N}{d_i}\right) + 1$$

The function considering the ratio (frequency of words vs. number of documents):

$$\mathrm{G(i)} = \frac{g_i}{d_i}$$

In the global weight function g_i is the frequency with which the word appears throughout the document collection and, d_i is the number of document, and N means total number of document. Globally weighted using the concept of entropy has a large value when the emergence of a rare word in the document. If a certain word once all appearance in all documents, if the word is not a characteristic that differentiates the respective document locally weighted using the entropy concept has a value of 0, only appear a certain word one document of the entire document, It has a value of 1.

This global weighting function using the standardization concept of regional weights have the same characteristics as used in the collection of documents represents the percentage of words that appear in the document collection[7].

3. Classification Method

3.1 Baysian Method

The Bayesian method is one of the most widely used algorithms in the field of document classification, calculate the posterior probability that it will be assigned to each category receives a single document using the Bayesian theory, such as the equation (3.1)[8].

$$P(C_j \mid d) = \frac{P(d \mid C_j) P(C_j)}{P(d)} \tag{3.1}$$

In this Formula, d means random document, Cj means j-th category. P(d) has the same value to all categories, doesn't need to calculate probability. Bayesian method assumes all the words are independent from each other, assignment of the category that occurs in the document is that mutually exclusive.

So, the $P(C_j|d)$ can be calculated as follow equation (3.2)

$$P(C_j \mid d) = P(C_j) \prod_{i=1}^{n} P(w_i \mid C_j) \tag{3.2}$$

In (3.2), $P(w_i|C_j)$ means (the number of w_i's occurrence in C_j)/(the number of all words' occurrence in C_j), and $P(C_j)$ means (the number of documents allocated in C_j)/(the number of all documents).

The Baysian method calculates the probability of being classified in each category by assigning a document to a category having the maximum value, because a number of categories are categories to calculate the conditional probability for each of the categories having the highest probability belong to the document.

3.2 k-NN

The k-NN classification method is to select one pattern from among the stored pattern with a distance of at least the learning data with respect to any particular pattern and classifies the category number which belongs to the category of the given pattern[9].

The k-NN algorithm is primarily used to the degree of similarity between the new document (d_x) and the learning document (d_j) to find the degree of similarity is high top k neighboring documents in document study groups. Representative degree of similarity is the cosine similarity degree is calculated as in equation (3.3).

$$sim(d_x, d_j) = \frac{\sum_k t_{xk} \times t_{jk}}{\sqrt{\sum_k (t_{xk})^2} \times \sqrt{\sum_k (t_{jk})^2}} \quad (3.3)$$

In the equation (3.3), t_{xk}, t_{jk} means the weight value of word k appeared in d_x, d_j.

The degree of similarity between the new document and document neighbors is used as the document category, the weight of the neighborhood, if neighboring pages when the category weights, share category is high.

When extracting k learning documents among the new document and the order of similarity using the cosine similarity, the category assigned to each k learning documents is a candidate list of categories to be assigned to the new document. The order to find the best category to be assigned to the new document category from the candidate list, calculate the compliance with total frequency of each category or the similarity score per category such as expression (3.4) in the learning document classification.

$$\text{rel}(C_k \mid d_x) \approx \sum_j sim(d_x, d_j) \times \{(C_k \mid d_j) \quad (3.4)$$

In the k-NN algorithm selection of k it will have a large impact on the classification result. Likely k is not too small, there is a possibility of overcharging as the sum of the noise data for training, as opposed to the data groups classified as near to that k is too large classification exist.

3.3 Decision Tree

Decision tree is made a decision rule to classify the function. In supervised learning problem it is sometimes important to the prediction of the final model and analysis even more emphasis on predictive analysis or according to circumstances. By dividing the regions of each variable repeatedly with supervised learning techniques, decision tree creates a rule with incorrect prediction and correct analysis relatively to the other supervised learning techniques for the whole area.

Rules created by the decision tree has the advantage of being easy to understand and easy to implement, because they have if-then structure and similar to SQL database language format.

The analysis of the decision tree is composed of growth of the decision tree, pruning, feasibility study, analysis and forecasting as shown in Table 3.

The most widely used in decision tree algorithm is CHAID. This can be applied to all types of the target and classification variables, such as nominal, the order-type, continuous type.

3.4 SVM

SVM is receiving attention not only in the classification problem, and also applicable to a regression problem, accurate and applicable to various types of data of the predicted prediction easier problem. Unlike the probability estimation of the object to minimize the empirical risk, such as the logistic regression and discriminant analysis conventional methods classification SVM is got with the purpose to minimize structural risk look place only the classification efficiency itself over existing probability estimation method overall, higher predictability.

Table 3. The Analysis Process of Decision Tree

Division	Description
Growing decision tree	Grow the trees to find an appropriate optimum separation criteria, and stop if they meet the appropriate stopping rules.
pruning	Remove a branch which will increase the risk of an error significantly or has an improper inference rules.
feasibility study	Evaluate the decision tree using gain chart, risk chart or verification data
analysis and forecasting	Interpret and apply build a wooden model after setting a prediction model predictions.

The reason of effectiveness in SVM is that it performs linear classification quickly and easily when the linearity see the non-linear data set at a low level at a higher level and expand its dimensions. Most of the pattern and can not be linearly separated, to separate the non-linear patterns and converts the input space of the non-linear pattern in the feature space of the linear pattern. At this time, the kernel functions are used to classify non-linear patterns.

$$K(x, x_i) = ((x \cdot x_i) + 1)^d \quad (3.5)$$

Polynomial kernel function is dependent on the direction between the two vectors, a vector having the same direction, they eventually result there is given a high value is for a polynomial kernel function, a polynomial kernel function is the same as equation (3.5).

4. Experiments

In this study, we used a Korea Herald article for one month (in January 2016). Total 340 articles related to "North Korea" and "nuclear" are used in the analysis. It was repeated 10 times performed to ensure the validity of the generalized model building, and analyzed through the same procedure as in <Figure 2>.

4.1 Structuring Data

Since the newspaper article is unstructured data, we perform the structuring process at first to apply formal analysis techniques. The first step of structuring is to extract a keyword from the data, the newspaper article, comprises a variety of combinations, such as letters, special characters, numbers. Table 4 shows a part of newspaper article and the keywords extracted.

Even though some nouns appeared in all documents without meaning, we exclude them and put to the stop words. The vector of the rough with the keywords extracted noun is created for every article. Using this vector, word-document matrix such as Table 5 was created.

Table 5 is a part of word-document matrix generated by configuring each of the newspaper articles. In the vector, column is words extracted from article, and row is each newspaper article number. In Table 5, the number 1 means Doc 8, Doc 9 and Doc 12 include the word 'coal' in

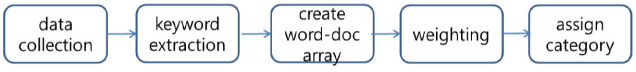

Figure 2. Analysis Process.

Table 4. Keywords Extraction

Article	Extracted nouns
China could stop buying coal from North Korea to punish the ally for its recent nuclear test, a South Korean expert suggested Thursday amid calls for Beijing to take a firmer stance against Pyongyang. China has come under growing pressure from South Korea and the United States to help draw a strong sanctions resolution from the U.N. Security Council to punish the North for its fourth nuclear test last week. Choi Kyung-soo, president of the North Korea Resources Institute in Seoul, noted the North's high reliance on trade with China. "Coal exports account for nearly half of all North Korean exports to China," he said in a phone interview with Yonhap News Agency. North Korea earned $2.84 billion from exports to China in 2014, nearly 90 percent of the $3.16 billion earned in total, according to data from the institute and the Korea Trade-Investment Promotion Agency.	coal, North Korea, ally, nuclear, test, South Korea, Thursday, calls, Beijing, stance, Pyongyang, China, pressure, United States, sanction, resolution, U.N. Security Council, week, Choi Kyung-soo, president, Resources Institute, Seoul, reliance, trade, half, he, phone, interview, Yonhap News Agency, billion, percent, data, Korea Trade Investment Pormotion Agency

Table 5. Word-Document Matrix

terms	documents														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Coach	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Coagulation	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Coal	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0
Coalition	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Coast	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
Coat	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Coater	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Coating	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Coatrack	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Coatroom	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

the article. Word-document matrix is the basic form of a newspaper article which converts unstructured data into structured data.

This study is due to the newspaper's view to applying classification techniques to formalize the newspaper into one object, word-document matrix was applied to

the pre-analysis. The matrix transposition is located on the line newspapers, and a considerable number of words having the meaning of the word matrix rather than a variable number of newspaper articles that correspond to the nature of the data object to the location column.

4.2 Comparision of Models

To evaluate the performance of the classification model, it was used for Correct Classified Rate (CCR), Recall Rate (RR), and F-measure to be used in document classification criteria. The Correct Classified Rate is the rate of correct prediction, and Recall Rate is the ratio actually hit accurate predictions. And F-measure means the combinational mean of CCR and RR, and this is convenient expression method to compare models.

The calculation of CCR and RR and F-measure can be written as equation (4.1) using the actual situation and prediction for the classification in Table 6.

$$\mathrm{CCR}:\mathrm{p}=\frac{A}{A+B}$$

$$\mathrm{RR}:\mathrm{r}=\frac{A}{A+C}$$

$$\mathrm{F}(\mathrm{p},\mathrm{r})=\frac{2pr}{p+r} \tag{4.1}$$

In order to compare the performance of the article were classified using the F-M value using a CCR and RR, exhibited the model F-Measure specific value corresponding to 10 times the Table 7.

Table 6. The Classification Table of Correctness in Prediction Model

Category		Real	
		Congruity	*Incongruity*
Prediction	Congruity	A	B
	Incongruity	C	D

Table 7. The Mean and Stand Deviation in each Model

Model	F-measure	
	Mean	*Standard Deviation*
Baysian	0.5404	0.002
k-NN	0.5760	0.002
Decision Tree	0.5757	0.001
SVM	0.5909	0.001

SVM model has the best F-measure value (59%) model showed the value compared to the other models in this category.

5. Conclusions

In this paper, we converted the unstructured data to structured using text mining techniques. And we compared those data by applying the material to the conventional techniques of Bayesian statistical classification, k-NN, decision tree, SVM to form a prediction model. After morphological analysis, we extracted only for the classification word as a noun. By applying a stopwords dictionary with respect to the extracted to remove words that can cause noise in the analysis and the word-document matrix was produced using the remaining words.

By weighting the matrix to indicate a good correlation between the word and document, it was converted into a simple frequency matrix. Model Evaluation criteria include the performance was compared using a common information classification and rate of recall and F-measure that evaluation criteria in the classification document. It was also run for 10 iterations performed to ensure the validity and generalization of the model building.

The F-measure value of SVM model shows better performance than other models. After using the standardized Unstructured newspaper article in this paper, applying to each analysis were compared to the results. Overall, the performance was shown in the model is relatively low, the reason is estimated to be limitations of the automatic document classification algorithm.

5. Acknowledgement

This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2014S1A5B6035600).

6. References

1. Hearst Marti A. Untangling Text Data Mining. Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. 1999; 3(10). DOI: 10.3115/1034678.1034679.
2. Kho Kwangsoo, Chung Wonkyo, Shin Youngkeun, Park Sangsung, Jang Dongsik. A Study on Development of Patent Information Retrieval Using Textmining. Journal of the Korea Academia-Industrial cooperation Society. 2011; 12(8):3677-88.

3. Cho Taeho. Comparison of Neural Network and k-NN Algorithm for News Article Classification. 1998 Conference on Korea Information Science Society. 1998; 25(211):363-65.
4. Bartere MM and Deshmukh PR. Cluster Oriented Image Retrieval Systems. IJCA Proceedings on Emerging Trends in Computer Science and Information Technology. 2012; ETCSIT(3):25-27.
5. Mittermayer M and Knolmayer G. Text Mining Systems for Market Response to News: A survey, Working paper in Institut fur Wirtschaftsinformatik der Universitat Bern. 2006; 184:1-17.
6. Chin KK. The Graduate School of the University of Darwin College: Support Vector Machines applied to speech pattern classification. Dissertation of PhD. 1998.
7. Cart, NC, USA: SAS Institute Inc.: SAS Publishing. SAS® Text Miner™ 4.2 Reference. 2009.
8. Van Driel MA, Bruggeman J, Vriend G, Brunner HG and Leunissen JA. A Text-Mining Analysis of the Human Phenome. European Journal of Human Genetics. 2006; 14(50):535-42.
9. Sebastiani F. Machine learning in automated text categorization. ACM Computing Surveys. 2002; 34(1):1-47.