ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org

http://dx.doi.org/10.13088/jiis.2018.24.1.141

Mining Intellectual History Using Unstructured Data Analytics to Classify Thoughts for Digital Humanities*

Hansol Seo
Researcher at School of Management
Kyung Hee University
(power2sky@khu.ac.kr)

Ohbyung Kwon
School of Management,
Kyung Hee University
(*obkwon@khu.ac.kr*)

Information technology improves the efficiency of humanities research. In humanities research, information technology can be used to analyze a given topic or document automatically, facilitate connections to other ideas, and increase our understanding of intellectual history. We suggest a method to identify and automatically analyze the relationships between arguments contained in unstructured data collected from humanities writings such as books, papers, and articles. Our method, which is called history mining, reveals influential relationships between arguments and the philosophers who present them. We utilize several classification algorithms, including a deep learning method.

To verify the performance of the methodology proposed in this paper, empiricists and rationalism - related philosophers were collected from among the philosophical specimens and collected related writings or articles accessible on the internet. The performance of the classification algorithm was measured by Recall, Precision, F-Score and Elapsed Time. DNN, Random Forest, and Ensemble showed better performance than other algorithms. Using the selected classification algorithm, we classified rationalism or empiricism into the writings of specific philosophers, and generated the history map considering the philosopher’s year of activity.

Key Words : Digital Humanities, History Mining, Text Analysis, Philosophy, Classification Algorithms

Received : November 10, 2017 Revised : February 28, 2018 Accepted : March 9, 2018
Publication Type : Regular Paper Corresponding Author : Ohbyung Kwon

1. Introduction

“Digital humanities” refers to “the use and application of computational tools and methods to humanist domains of study, but also the opposite, the application of humanistic questions to computer science” (Martin, 2013). Digital humanities researchers “use information technology to illuminate the human record, and [bring] an understanding of the human record to bear on the development and use of information technology” (Schreibman et al., 2004, xviii). For these reasons,

* This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIP) (No.2017-0-01122, Development of personal profiling and personalized chatbot technology based on language usage pattern analysis for intelligent It's My Story service)

digital humanities research has been conducted in various disciplines such as literature (Berry, 2011; Berry et al., 2015), cultural anthropology (Wilkens, 2015), and geography (Jessop, 2008). Furthermore, universities and businesses such as Oxford University and Google are currently involved in digital humanities projects.

Big data analysis, one of the latest techniques in academic research, has been applied to areas in which digital humanities has received attention. Big data visualization tools are useful for grasping the relations among constructs in humanities literature. For example, Data Sprint utilizes Amazon API, a visualizing social network that includes brief bibliographic information such as authors' names and book titles (Berry et al., 2015). Data retrieval services can be supported by big data analysis as well. Currently, a study on retrieving the requested data within a fixed time frame from massive amounts of material is being conducted.

In digital humanities research, intellectual history is of great value in analyzing, and organizing the thoughts and ideas of a specific field. Thus, studies on intellectual history have been implemented in almost all fields of humanities and sociology, including religious studies (Cross, 2015), urban design studies (Hall, 2014), sociology (Kerber, 2014), and bibliography (Sattelmeyer, 2014). However, studying intellectual history is difficult since thoughts and ideas are very subjective, empirical, and subject to variation depending on the source. Furthermore, it is indirectly expressed in speeches or writings left by fallible human beings (Higham, 1954). In the study of intellectual history, it is important to analyze the connections between ideas and the similarities between a single person's thought and ideas in the broader culture. Identifying the ideological tendencies and traits of some intellectuals of the time and extracting, selecting, and cataloging the related literature are also important. However, such work is too extensive in scope and requires a lot of time; therefore, researchers tend to rely on existing studies or narrow the scope of their studies rather than trying to manage it on their own. Moreover, depending on the amount of literature to analyze, classifying them manually may be impossible. To overcome these difficulties, very little supportive information technology has been developed. For this reason, in this study we propose a method of automatically classifying unstructured data to search for certain ideas through data mining and analyzing relationships among them.

The remainder of this paper is organized as follows: after reviewing existing research involving digital humanities and document classification in Chapter 2, we provide a description in Chapter 3 of our history mining method that identifies and verifies the accuracy of relationships among intellectuals through their ideological writings. Chapter 4 outlines the experiment using the text data containing actual ideas and analyzes the results. Finally, Chapter 5 concludes this study and describes the necessity for subsequent research.

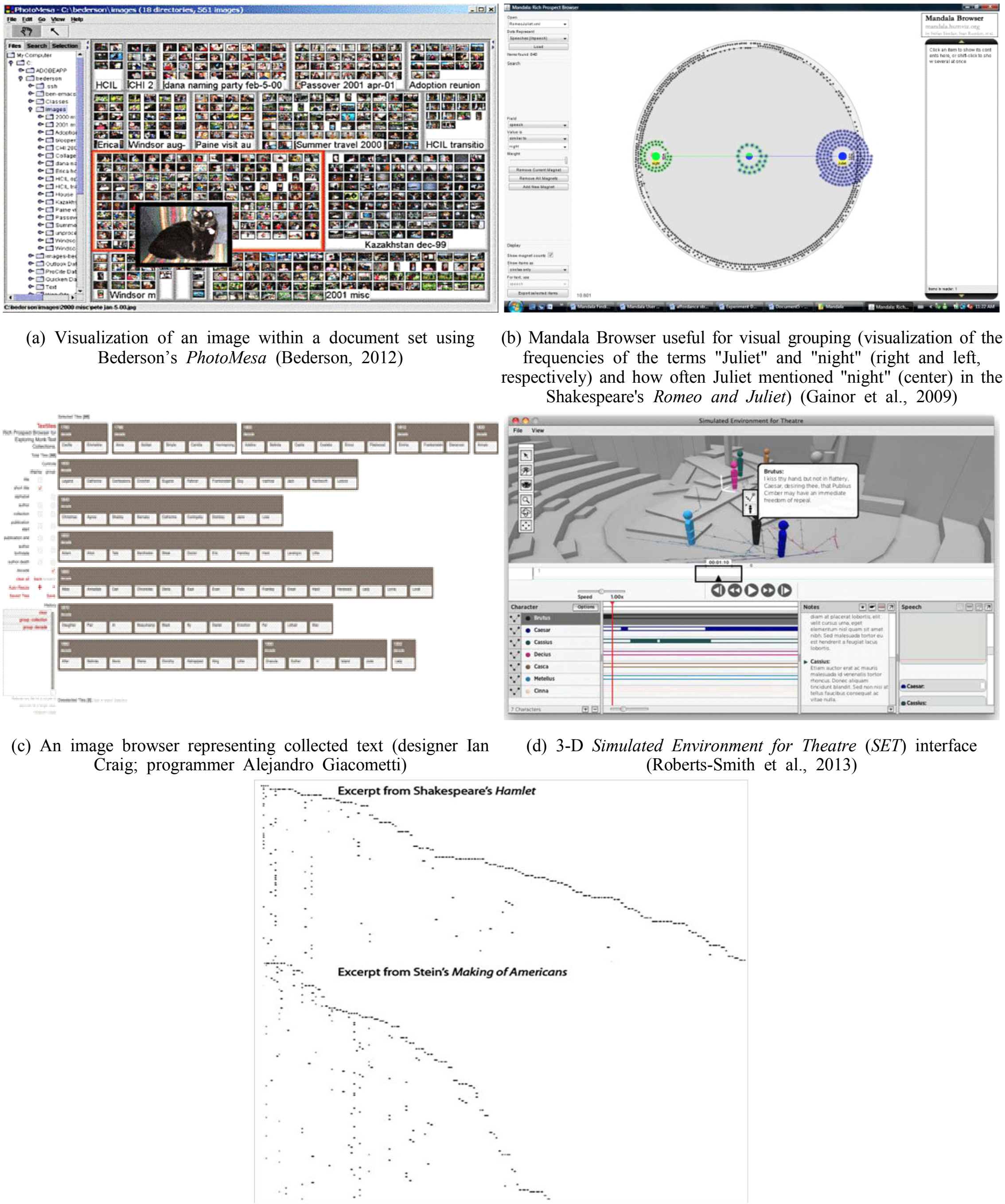
2. Digital Humanities

Digital humanities research is defined as the articulation of technical competences in computing with a critical approach to a given humanities topic, both within and beyond the academy (Ross and Sayers, 2014). Digital humanities, which is also called humanities computing, involves creative use of digital technologies to support humanities education and research (Gole, 2012). It may be considered as a vast ecosystem in which related activities are fostered and facilitated. However, some regard it as a secondary branch of existing disciplines (Nelson, 2016).

Despite some skepticism about computational analysis and digital communication in the humanities, scholarly techniques and interpretation of modernist literature have been expanded through digitization and computation (Ross and Sayers, 2014). More and more people leave their thoughts and ideas about topics in the humanities on mobile apps, websites, or social media. Samuel-beckett.net, Joycesociety.org, and Virginiawoolfsociety.co.uk (for Samuel Beckett, James Joyce, and Virginia Woolf, respectively) are representative sites. Also, many informal documents, slides, photographs, and visualizations that could not be included in formally published journals have become more widely available (Gold, 2012). However, most of these data are not officially scholarly; they do not come in standardized or structured forms. Furthermore, many sites include the ideas of low-profile thinkers, humanities scholars, and ordinary people who would not usually be recognized in academic circles. Many of these ideas and opinions could be worthy of analysis and interpretation, although they may not have necessarily been written for academic purposes.

Using our data mining technique, it may be possible to determine the philosophical category to which a person’s idea belongs based on his or her writings shared on the internet. As a result, communities may arise made up of people with similar ideas, in which related books and cultural products may be shared and recommended. Humanities scholars, in their quest to understand the human condition, may find it meaningful to elucidate the unstructured writings of various authors from a wide range of sources, even those without copyright. Due to the effort required to search, analyze, and classify these writings, we conducted this research and develop the data mining technique described herein to perform these kinds of analyses accurately and cost-effectively.

In the field of digital humanities or humanities computing, research on information visualization is currently also under way (Bae and Watson, 2014). This research shows a combination of facts and arguments from a given text in visual format (Sinclair et al., 2013), ii) it describes tools like Many Eyes, which provides eye-catching summaries of text-related statistics to users, making them more intuitively accessible, iii) it presents distributions of user-selected terms that appeared in various documents decoratively, rather than functionally, iv) it offers an overall view of multiple documents using tools like Mandala


PhotoMesa - C:\bederson\images (18 directories, 561 images)
HCIL
CHI 2
dana naming party feb-5-00
Passover 2001 apr-01
Adoption reunion
Erica
Windsor aug-
Paine visit au
Summer travel 2000
HCIL transitio
Kazakhstan dec-99
Windsor m
2001 misc
Mandala Browser
Simulated Environment for Theatre
Brutus: I kiss thy hand, but not in flattery, Caesar, desiring thee, that Publius Cimber may have an immediate freedom of repeal.
Character
Notes
Speech
Brutus
Caesar
Cassius
Decius
Casca
Metellus
Cinna
Excerpt from Shakespeare's *Hamlet*
Excerpt from Stein's *Making of Americans*

(a) Visualization of an image within a document set using Bederson's *PhotoMesa* (Bederson, 2012)

(b) Mandala Browser useful for visual grouping (visualization of the frequencies of the terms "Juliet" and "night" (right and left, respectively) and how often Juliet mentioned "night" (center) in the Shakespeare's *Romeo and Juliet*) (Gainor et al., 2009)

(c) An image browser representing collected text (designer Ian Craig; programmer Alejandro Giacometti)

(d) 3-D *Simulated Environment for Theatre* (*SET*) interface (Roberts-Smith et al., 2013)

(e) The repetition grid, designed by Piotr Michura (Michura et al., 2007)

〈Figure 1〉 Information Visualizations of Unstructured Text (Sinclair, 2013)

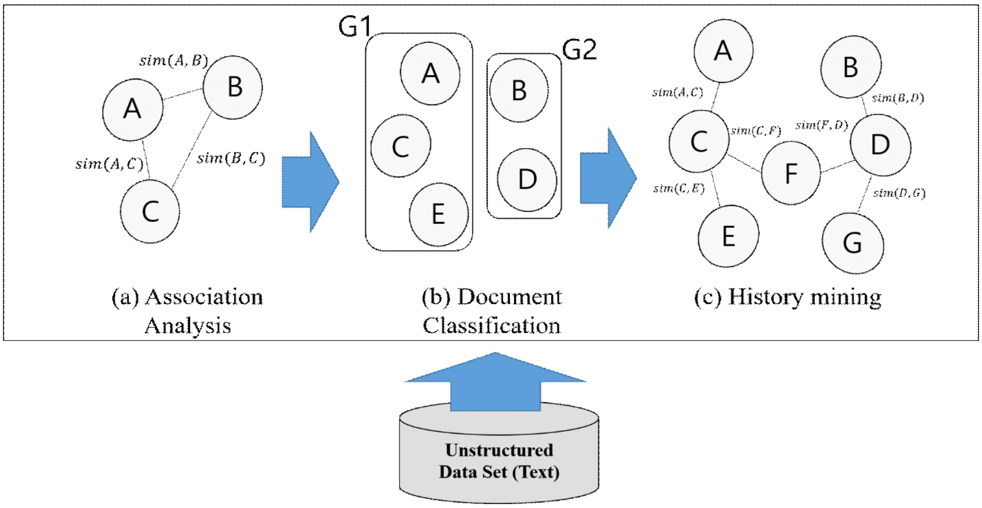
Browser (e.g., showing the novel *Romeo and Juliet* as XML code). In the field of information visualization, various methods of explaining unstructured text or documents are elucidated in research on humanities computing.

Figure 1 depicts various information visualization methods. Among them, the visual grouping approach is known to be useful from a usability perspective (i.e., in terms of efficiency, effectiveness, and satisfaction with the user's ability to access information) (Gainor et al., 2009). Recently, several studies have been published on how to improve the effectiveness for visual grouping by including sound (Yano et al., 2016), utilize the hierarchy derived from visual grouping for better comprehension of the information flow within a data set (Hunnicutt, 2016), and explore the effects and implications of visual grouping (Powell, 2016).

Some of the information visualization tools used for visual grouping, like Mandala, depend on the frequency of words appearing in the document or the frequency of certain terms appearing in the speech. Therefore, considerable effort is still required to interpret the meaning of the frequency. In our opinion, more in-depth studies about grouping based on ideas and consisting of a richer set of search terms or statistics are needed. Little progress has been made in this area.

3. Mining Intellectual History

To analyze and classify ideas in humanities texts and visualize temporal associations between similar thoughts in unstructured text datasets, we suggest a three-step analytical method: association analysis, document classification, and history mining (see Figure 2). In association analysis, Euclidean distance and cosine similarity are used



〈Figure 2〉 Conceptual Framework

to determine the similarities between ideas contained in the documents and evaluate them for the purposes of classification. In the document classification stage, collected documents are classified using existing classification algorithms and the performances of trained classifiers are evaluated. Finally, in the history mining stage, the unstructured text data in the document classification step is regarded as the data at time point t, after which we collect additional data at the t+n time point data, which is affected by time point t, and data at the t-n time point, which affects time point t. Using unstructured data from all time points, we then evaluate whether the data at the t-n and t+n time points are classified correctly.

3.1 Pre-processing

Before analyzing the collected unstructured data, it must be structuralized. For that reason, a vector space model is used. The vector space model, also called bag-of-words, is a method used to analyze data from a large number of documents by expressing them as *n*-dimensional vectors (Hotho et al., 2005). Documents can be regarded as thoughts or ideas expressed in letters (Small, 1978) contained in keywords, mostly nouns (Bouras & Tsogkas, 2008). In this study, we extract only nouns from the collected text to create a vector consisting of weights of nouns. Then, we create a Document-Term Matrix (DTM) composed of these vectors to structuralize the unstructured text data. To determine the weights of nouns, the Term Frequency-Inverse Document Frequency (TF-IDF) weights that are widely utilized in text mining are used (Kim and Kwon, 2015). Generally, use of a vector space model in text classification results in high dimensionality problems. Therefore, in this study, the Chi-squared test was adopted as a feature selection method. In order to avoid the overfitting problem, the keywords used to collect the text data were removed in advance.

3.2 Association Analysis

When documents are expressed as vectors of words, measuring similarities between those vectors reveals the associations among vectors (Huang, 2008). To analyze associations between documents, a DTM of the collected data is averaged for each class to create a single vector per class. After that, a distance matrix is constructed by measuring similarities between vectors of classes using Euclidean distance and cosine similarity. Euclidean distance is a method of measuring the distance between two points in a multidimensional space and it is widely used for clustering using techniques such as K-means in the field of text mining (Huang, 2008). When there are two documents, the Euclidean distance (D_E) between the term vectors of the documents, V_1 and V_2, is obtained using the formula below. The total number of words included in the two documents is presented as *n*. In addition, $w_{1,t}$ and $w_{2,t}$, which represent the weights of words, are expressed as TF-IDF values.

$$D_E(V_1, V_2) = \sqrt{\sum_{t=1}^{n} \left(w_{1,t} - w_{2,t}\right)^2}$$

Cosine similarity is a method of measuring similarities between two documents based on the inner product between those documents expressed as vectors (Dhillon & Modha, 2001). Cosine similarity is widely used for text mining as well as information retrieval because it is easy to interpret and is simple to use for calculation of sparse vectors.

$$D_C(V_1, V_2) = \frac{\sum_{t=1}^{n} w_{1,t} * w_{2,t}}{\sqrt{\sum_{t=1}^{n}\left(w_{1,t}\right)^2} * \sqrt{\sum_{t=1}^{n}\left(w_{2,t}\right)^2}}$$

3.3 Document Classification

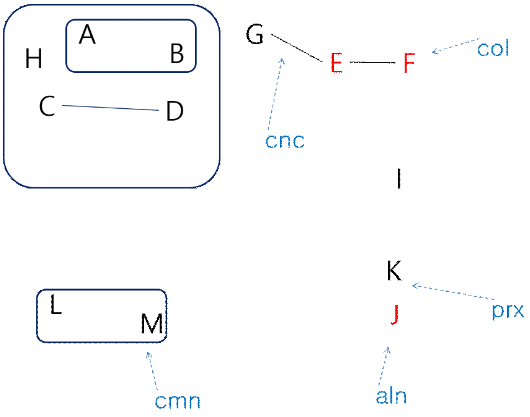
After conducting the association analysis, we classify the documents using existing classification algorithms, and performances of the trained classifiers are evaluated (Kwon and Lee, 2014). As classification algorithms, decision trees, such as CART and C5.0, deep neural network, k-NN classifier, multinomial logistic regression, naïve Bayes, random forest, and SVM are used. The parameters of each algorithm are optimized with tuning functions of libraries or a heuristic way. For validation of the trained classifiers, the holdout method is used to construct independent sets consisting of training and test sets at a random 7:3 ratio of unstructured to structured text data. Learning of the classification algorithms is performed by a training set, and evaluation of trained classifiers is performed by a test set. In order to evaluate the classification algorithms accurately, random sampling, which iterates the holdout method k times, is used. In this study, k is set to 100 and the holdout method is utilized 100 times. To measure the performances of classifiers, a confusion matrix is formed and then the F-score metric of the macro-average, which is a harmonic average of recall and precision, is used to check the average performance of the classifiers from all classes.

History mining, the construct developed in this study, is a method of analyzing unstructured documents and classifying all the people, such as scholars and researchers, who appear in the documents into a finite number of certain assertions and theories. It is a kind of information visualization that uses a visual grouping approach as a technique for showing relationships between ideas.

The history mining technique is used as follows. The data acquired in the document classification step are associated with time *t*, and the data of time point *t-n* (which affect time point *t*) and time point *t+n* (which is affected by time point *t*) are also collected. Like the data of time *t*, the additionally collected data is also pre-processed by removing search keywords and extracting nouns. Then, labelling is performed on the data at time points *t+n* and *t-n* based on the protocol outlined in previous literature.

After labelling is complete, document classification is performed using classification algorithms as in the previous step. The parameters of each algorithm are optimized with tuning functions of libraries or in a heuristic way. In addition, to ensure optimal performance of the history mining method, ensemble techniques are used: the simple voting (i.e., majority voting) method and the weighted ensemble method (i.e., F-scores of individual classification algorithms are used as weights for each algorithm). With these techniques, higher scores are assigned to classification algorithms with superior performance. For evaluation of the history mining method, the macro-average F-score metric is used.

Next, a history network is constructed to visualize the results of the history mining evaluation. History network modeling is depicted in Figure 3 based on the approach of Bae and Watson (2014), which is a visual grouping that includes five grouping cues: common region of the terms or the conceptual sets of terms (cmn), connectedness (cnc), color similarity (col), proximity (prx), and alignment (aln).



〈Figure 3〉 Bae and Watson (2014): Approach for Visual Grouping

(cmn = common region, cnc = connectedness, col = color similarity, prx = proximity, aln = alignment)

As in the association analysis step, all data collected from texts at time points *t-n*, *t*, *t+n* are converted into a DTM, and the average value for the DTMs is calculated for each thinker. Then, similarities between thinkers are determined using cosine similarity, the distance matrix is constructed based on these similarities, and the matrix is used to express the relations among thinkers as a network. The thickness of the colored edges between nodes in Figure 3 represents the intensity of the relationships between thinkers. The location of the nodes is determined by the similarities in ideas among thinkers and the year in which the activity occurs. The range of the network is represented by the x and y coordinates, respectively, from -1 to +1. The y-axis is determined by the year of activity of each thinker, which is determined by extracting from the collected text data and eliminating the outlier years. Using Euclidean distance, the x-axis is determined by measuring the similarities between ideas and thinkers. Then, if there are many similarities, the nodes of the thinkers are placed closer to -1 and +1 on the x-axis, and if there are few similarities, the nodes are placed closer to 0. The results of the history network are then interpreted in comparison with existing literature on the thinkers.

4. Experiment

4.1 Data

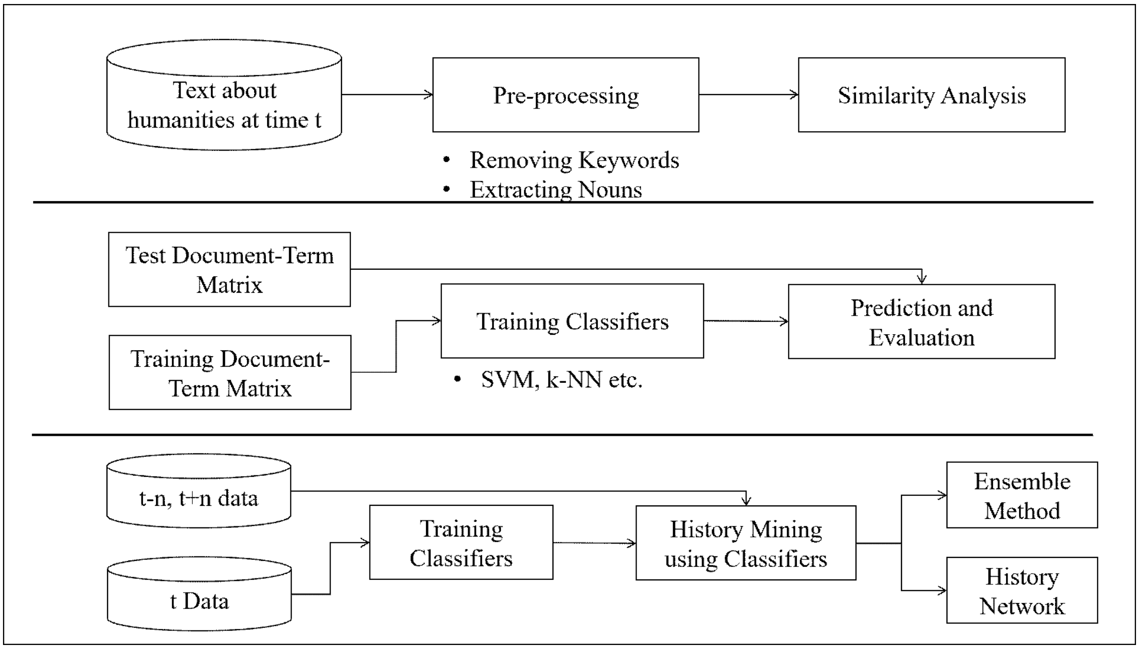
To verify the performance of the methodology proposed in this work, experiments were conducted on texts containing modern philosophical thought. For this purpose, the authors collected unstructured text data related to modern empiricist and rationalist philosophical thought. Among the modern philosophers, Descartes, Spinoza, and Leibniz represent rationalism and Locke, Berkeley, and Hume represent empiricism. Based on these philosophical categories, the names of these thinkers were searched on internet portal sites such as SNS and web blogs. In total, 308 philosophical thought texts, including 261 cases acquired through Internet searches and 47 cases from philosophical books, were collected and summarized (Table 1). For reference, all of these data are free from privacy-related problems and can be shared.

4.2 Procedure

The process for the experiment can be divided into three steps, as shown in Figure 4. In the first step, the authors collect the data of time *t*, which is the era in which modern empiricist/rationalist philosophical thought was prevalent, in the form of unstructured text data. The collected data is preprocessed by removing the name of the philosopher who is the subject of the search and extracting only nouns. The association analysis, Euclidean distance, and cosine similarity are then performed to see how philosophical thoughts in the texts are related to the ideas of the philosopher. In the second stage, modern philosophical thought data is divided into a training data set and a test data set, and then a DTM composed of TF-IDF weights is conducted. The training DTM is used for learning with classifiers such as SVM and k-NN, and the test DTM is used for evaluating the learned classifiers. The evaluation metric uses the F-score. In the third stage, the data at time *t-1*, which influenced modern philosophical thought,

〈Table 1〉 Sample Data Set

School of Philosophical Thought	Philosopher	Text
Rationalism	Descartes	Descartes tries to establish absolute certainty in his famous reasoning: “Cogito, ergo sum,” or “I think, therefore I am.”
Empiricism	Locke	In their natural state, all people are equal and independent, and everyone has a natural right to defend his "Life, health, liberty, or possessions"
Empiricism	Hume	A person’s imagination, regardless of how boundless it may seem, is confined to the mind’s ability to recombine the information it has already acquired from the body’s sensory experience (the ideas that have been derived from impressions).
⋮	⋮	⋮



〈Figure 4〉 Experimental Procedure

and the data at time *t+1*, which is influenced by modern philosophical thought, are collected. Then, history mining is performed using data from times *t-1*, *t*, and *t+1*.

First, as part of preprocessing, the Internet search engine shows the search results to users using TF-IDF (Blei et al., 2003). The names of the six philosophers, Descartes, Spinoza, Leibniz, Locke, Berkeley, and Hume, are removed to avoid overfitting in the classifier learning process. Noun extraction is performed using the morphological analyzer RHINO 2.5.4 (Choi et al., 2015). The extracted nouns form a corpus created using the text mining package of the open source statistical program R, and labelled “tm” using the Corpus and VectorSource functions. The data, which are collected using the holdout method, are divided into a training set for classifier learning and a test set for predictive performance evaluation, and the data set is divided into a 7:3 ratio randomly. Since the holdout method is divided by class (philosopher’s name) by using “createDataPartition” function of the “caret” package, the class ratio of the training and test sets is even. Nouns are extracted from the data set of philosophical thought collected from the articles using RHINO 2.5.4, and the data, which are stored in a text file, are organized into a .csv file using R. Before using the holdout method, the philosopher's name is removed from the form to avoid over-sum problems with the classifier. The DTM is constructed as shown in Table 2 through

〈Table 2〉 Document-Term Matrix of a Sample Training Set

Document ID	Sense	Experience	Cognition	Existence	Cause	Philosopher
1	0.017359	0.003962	0.003626	0.001526	0.004109	Descartes
2	0.046697	0.003849	0.003131	0.000988	0.002661	Descartes
⋮	⋮	⋮	⋮	⋮	⋮	⋮
41	0.021479	0.027086	0.029158	0.004908	0.003305	Locke
42	0.050914	0.012589	0.019199	0	0	Locke
⋮	⋮	⋮	⋮	⋮	⋮	⋮
105	0.010902	0.048524	0.019734	0.01661	0.005592	Hume
106	0	0.031681	0.025769	0.004067	0	Hume
⋮	⋮	⋮	⋮	⋮	⋮	⋮

the vector space model drawing input values from the data. Table 2 represents a part of the resulting training set matrix; each component represents a TF-IDF value of a specific word in the corresponding document. In total, 5,739 nouns were extracted from 308 documents. Since the entire data set was divided using the holdout method, the data of the training set did not affect the test set. The training set and test set DTM words are used in only 5,739 words since the DTM is generated. On average, about 4,948 words are used in DTM generation of training sets and test sets.

The authors then confirm the similarities among the philosophical thought texts of the six philosophers collected through association analysis. The Euclidean distance and cosine similarity methods were used for this purpose. Average TF-IDF values of the DTM were calculated for each philosopher, and the distances between the feature vectors were compared between them.

After the association analysis, dimension reduction is performed using the chi-squared test, which is the most widely used feature selection method, to remove nonsignificant features prior to document classification. Then, the authors classify the classifiers using methods such as decision tree (C5.0, CART), deep neural networks, k-NN, logistic regression, naïve Bayes, random forest, SVM1, and SVM2 using the training DTM. For the parameters of the algorithms, a tuning function provided by the library is used. If there is no tuning function, the researchers perform a heuristic operation.

This paper focuses on the discriminant algorithm, which shows excellent performance in the document classification stage. The authors collected unstructured textual data related to modern philosophical thought and ancient philosophical thought, which influenced modern rationalism and empirical philosophical thought. History mining is performed by using texts containing modern philosophical thought as learning data and other related data as test data. The philosophers who influenced modern empiricism were Gorgias and Protagoras; these influential philosophers collected the philosophical texts of Bentham and Mill. Plato and Aquinas are philosophers who influenced modern rationalist philosophy, and philosophical texts of Schopenhauer and Hegel, who were influenced by them, were collected. In addition, the authors also collected data from Kant's philosophical ideological texts, which were influenced by both modern rationalist and empiricist philosophical thought, to confirm that the classification was done correctly even for philosophers who were influenced by both schools of thought. As an outcome of the history mining process, ensemble techniques were utilized to improve the performance of the method using individual classifiers that yielded better performance. The authors compared the weighted ensemble method with the weighted F-score of individual classifiers to determine which method was superior.

For the experiment, we used a hardware which contains Processor Intel® Core™ i7-5500U CPU @ 2.40GHz 2.39 GHz, 8.00GB RAM, 64-bit Operating System, x64-based processor and runs on Windows 8.1.

4.3 Results

4.3.1 Association analysis

As a result of the association analysis, we observe that Euclidean distance (Table 3) was the most similar between the rationalist philosopher Descartes and the Spinoza documents (0.097). In the cosine similarity analysis (Table 4), the documents of the empiricist philosophers Berkeley and Locke were found to be most similar (0.586). However, the document of Hume, an empiricist philosopher, was very similar to that of the rationalist philosopher Descartes (Euclidean distance 0.134, cosine similarity 0.353) rather than to those of the other empirical philosophers Berkeley and Locke. It seems that Hume’s skepticism was unlike the thinking of empiricist philosophers such as Berkeley and Locke.

<Table 3> Similarities between Philosophers Computed by Euclidean Distance

	Berkeley	Descartes	Hume	Leibniz	Locke	Spinoza
Berkeley	0					
Descartes	0.104013	0				
Hume	0.139714	0.134218	0			
Leibniz	0.140758	0.130063	0.166499	0		
Locke	0.098182	0.113638	0.140942	0.148367	0	
Spinoza	0.116909	0.097906	0.144216	0.131543	0.125264	0

〈Table 4〉 Similarities between Philosophers Computed by Cosine Similarity

	Berkeley	Descartes	Hume	Leibniz	Locke	Spinoza
Berkeley	1					
Descartes	0.447761	1				
Hume	0.336501	0.352719	1			
Leibniz	0.291569	0.359281	0.209868	1		
Locke	0.586265	0.396486	0.357334	0.253208	1	
Spinoza	0.35182	0.498597	0.280601	0.37575	0.311997	1

4.3.2 Document classification

A confusion matrix was obtained as shown in Table 5. Based on this matrix, the average recall factor, precision, F-score, and elapsed time for learning are shown in Table 6. As a result, the F-score of DNN showed the best performance with a value of about 0.95, the value using the random forest method was about 0.94, that with the SVM with the RBF (radial basis) kernel function was about 0.93, and that with the logistic regression

〈Table 5〉 Confusion Matrix of an Example Classifier

Random Forest	Berkeley	Descartes	Hume	Leibniz	Locke	Spinoza	Predicted
Berkeley	14	0	0	0	0	0	14
Descartes	1	14	0	0	0	0	15
Hume	0	0	15	0	0	0	15
Leibniz	0	0	0	14	0	0	14
Locke	0	0	0	0	15	0	15
Spinoza	0	1	0	1	0	15	17
Actual	15	15	15	15	15	15	90
Logistic Regression	Berkeley	Descartes	Hume	Leibniz	Locke	Spinoza	Predicted
Berkeley	14	1	1	1	1	0	18
Descartes	0	14	1	0	0	0	15
Hume	0	0	13	0	0	0	13
Leibniz	0	0	0	14	0	0	14
Locke	1	0	0	0	14	0	15
Spinoza	0	0	0	0	0	15	15
Actual	15	15	15	15	15	15	90

〈Table 6〉 Recall, Precision, F-score and Elapsed Time of the Classifiers Considered in the Experiment

	Recall	Precision	F-score	Mean elapsed Time (sec)
C5.0	0.7337	0.7541	0.7437	1.259
CART	0.7322	0.7648	0.7480	0.514
DNN	0.9437	0.9467	0.9452	27.047
k-NN	0.4474	0.6019	0.5114	0.971
Logistic	0.9107	0.9186	0.9146	11.439
Naïve Bayes	0.7677	0.7802	0.7739	0.519
Random Forest	0.9377	0.9423	0.9400	4.144
SVM (linear kernel)	0.8892	0.8979	0.8936	253.136
SVM (RBF kernel)	0.928519	0.933253	0.930875	8.287

method was about 0.92. For the k-NN classifier, the F-score was about 0.51, which was not good for high-dimension K-NN classifiers. It seems that the features of the training and test sets are still high-dimension even after the feature selection is performed, because the features are larger than the number of documents used for learning. SVM has the best performance of the RBF kernel function and linear kernel function among several possible kernel functions. In the case of SVM using the linear function, parameters were optimized using the tuning function provided by the library, and the average execution time was the longest. However, the F-score is about 0.89, which is not superior to that with the SVM using the RBF kernel function without parameter optimization. Lastly, when it comes to compare DNN with Random Forest, even though DNN seems a little bit better than Random Forest, Random Forest is more competitive than DNN in terms of elapsed time for learning. Hence, choosing DNN or Random Forest depends on the size of the learning data set.

4.3.3 History mining

The results of the history mining evaluation, including philosophers who have influenced modern empiricist and rationalist thought and other philosophers, are shown in Table 7. As in the document classification of modern philosophical thought, the deep neural network classifier has the highest F-score of about 0.9685.

However, the random forest method, which showed excellent performance in the classification of documents related to modern philosophical thought, performed relatively poorly with an F-score of about 0.7672 in the history mining analysis. Because it checks the confusion matrix, the random forest algorithm tends not to classify Kant's data, which are all influenced by both schools of philosophical thought. Therefore, the

<Table 7> Recall, Precision, F-scores, and Elapsed Time of History Mining

	Recall	Precision	F-score	Mean elapsed Time (sec)
C5.0	0.8128	0.8037	0.8078	2.276
CART	0.7492	0.7749	0.7613	0.994
DNN	0.9684	0.9686	0.9685	60.889
k-NN	0.5050	0.7356	0.6130	1.865
Logistic	0.9540	0.9545	0.9542	3.439
Naïve Bayes	0.8186	0.7890	0.8033	0.763
Random Forest	0.6955	0.8413	0.7671	8.399
SVM (linear kernel)	0.8648	0.9108	0.8867	498.726
SVM (RBF kernel)	0.7864	0.8810	0.8323	8.603
Simple Ensemble	0.9448	0.9499	0.9472	587.479
Weighted Ensemble	0.9484	0.9539	0.9511	587.525

overall F-score decreased. Logistic regression resulted in an F-score of about 0.9685, which is higher than that of the classification of texts of modern philosophical thought. SVM using the linear kernel function performed better than SVM using the RBF kernel function (F-scores and 0.8323 and 0.8868, respectively), unlike the document classification related to modern philosophical thought. The performance of SVM using logistic regression and the linear kernel function is excellent because the history mining technique is able to separate the data set more linearly with a smaller number of classes to classify i than that of the modern philosophical math class.

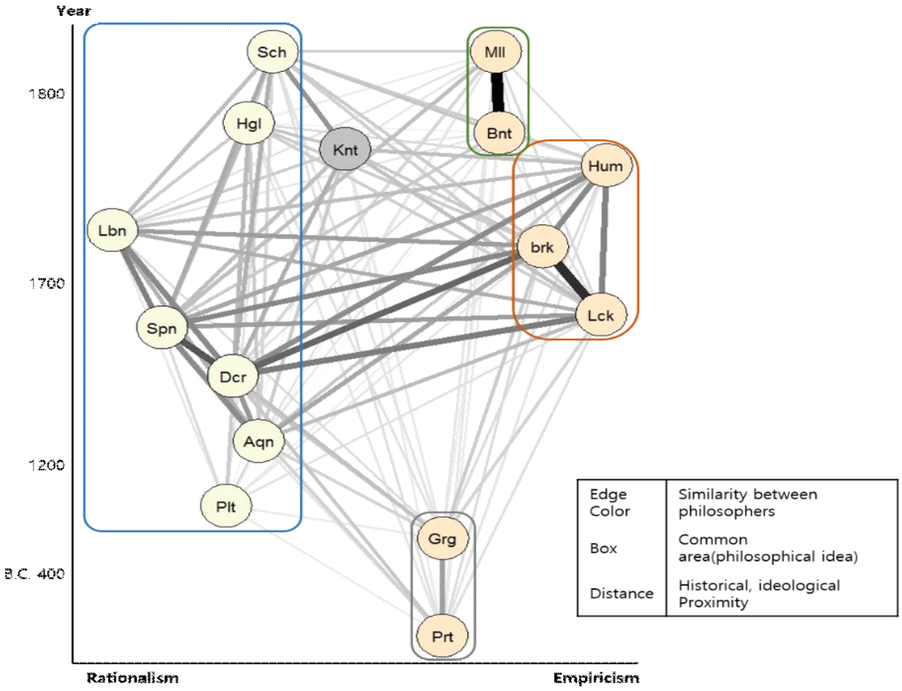
In order to improve the performance of the history mining technique, the authors performed a simple ensemble analysis using simple voting of the whole algorithm and a weighted ensemble analysis using F-scores of individual algorithms as weights. The performances of the two ensembles were relatively good, with F-scores of 0.9473 and 0.9511, respectively, but they did not perform better than the deep neural network and logistic regression individual classifiers. These results show that the algorithms such as CART, k-NN, and random forest obtained values below the F-score of 0.8, although the DNN and logistic regression algorithms performed well. Other algorithms yielded F-scores of around 0.8 except for the linear kernel function. Thus, no significant improvement in performance was observed despite the use of all nine algorithms in an ensemble.

In the field of text mining, machine learning techniques are used rather than statistical techniques for document classification. In general, the SVM technique is considered to be the best among the individual determination algorithms

(Korde and Mahender, 2012). Deep neural network, a machine learning technique, has performed well in modern philosophy classification and history mining studies. However, in the case of the random forest method as a machine learning technique and logistic regression as a statistical technique, there were significant differences in performance between the two experiments. In classification evaluations of philosophical literature, the performance of the random forest machine learning method was excellent in cases with no imbalance between classes, but its performance degraded markedly when there was an imbalance between classes. The logistic regression statistical technique was found to be superior to all the others except DNN.

According to research using ensembles of individual algorithms, ensemble with weights (weighted voting) yielded better performance than that with simple majority voting (Fung et al., 2006; Xia et al., 2011). The results of this study also show that the ensemble method with the F-score as a weighted value is superior to that of the simple majority ensemble.

As a result of visualizing history mining using a distance matrix with cosine similarity in the association analysis, we confirmed that the intensity of the relationship between the modern philosophers representing each idea was relatively strong, as shown in Figure 5. Among the



〈Figure 5〉 History Mining using Distance Matrix empiricist philosophers in particular, Berkeley (Luk) and Locke (Lck) had the strongest relationship. Among the rationalist philosophers, the relationship between Descartes (Dcr) and Spinoza (Spn) was strongest. The relationship between Bentham (Bnt) and Mill (Mil), both of whom were influenced by empiricist philosophy, was also strong because both philosophers were utilitarian philosophers and adhered to a similar school of philosophical thought; therefore, strong similarity was evidence in their texts. Kant (Knt), who was affected by both empiricism and rationalism and who established a new philosophical system of critical philosophy by creating a fusion of modern empiricist and rationalist philosophy, is not the most similar to the two philosophical thought of the nine philosophers except for the modern philosopher. He showed a neutral shape.

4.3.4 Comparison with survey results

To test the performance of the history mining method, the authors conducted a questionnaire survey, asking members of the public to read the texts of philosophical thought and evaluate the performance of the philosophical classification by matching each text with a philosophical ideology. The respondents to the questionnaire were 8 MBA students with majors unrelated to philosophy. The results of the survey revealed that the value for the accuracy of the respondents was 0.35 and the F-score was about 0.283, which was a significantly lower value than that of history mining. We also investigated why the questionnaire was difficult to answer through additional questionnaires. In analyzing the questions, we determined why the results of the initial survey were low. First, the survey respondents were studying in fields unrelated to philosophy and lacked background knowledge about philosophy. Second, the survey method lacks the ability to decode large amounts of information quickly. Finally, certain keywords play an important role in classifying philosophical thought; however, the ability of most classifiers to extract these keywords accurately is relatively weak.

5. Discussion

5.1 Academic Implications

The results of this study have several implications on the academic side. First, the authors propose the history mining technique that considers both time and objects (such as characters or patterns), applying it to the domain of philosophy in the field of digital humanities. There are many studies in the digital humanities field that analyze poems, novels, and other literary works using text mining techniques (Lord et al., 2006; Sculley and Pasanek, 2008; Yu, 2008), but few studies have examined the relationships among literary authors and philosophical thinkers using data mining techniques. Therefore, this study classifies philosophers' ideas, as expressed in various texts, through text analysis techniques and history mining that allow us to see the relationship between thinkers automatically.

Second, by applying the history mining algorithm proposed in this paper, the authors elucidate the influence of relationships and similarities between ideas with less effort and time than what would be required without this technique. History is, in fact, highly selective: fully reflecting the ideas of a given age is an unmanageable task (Higham, 1954). In addition, common statistical techniques have the added complexity and difficulty of analyzing historical data over contemporary data (Choi et al., 2015; Lee et al, 2016). However, the history mining method proposed in this study is not sophisticated and can overcome limitations of existing quantitative and formal data-based statistical analysis methods.

Third, the results of the classification algorithm proposed in this study can be regarded as valid because they are in line with the results of existing research. For example, the relation between British empiricism and later utilitarianism, as evidenced in the writings of philosophers such as Bentham and Mill, was relatively high and influential. This supports the claim that John Stuart Mill borrowed the inductive technique of Francis Bacon, a modern empiricist philosopher, to replace Aristotle's problem-solving methodology (Christians, 2007). Also, in a study of Jeremy Bentham of Olivecrona (1975), Bentham agrees with the assertion that all knowledge derives from a sense of confirming the reality of reality. There is no direct connection between the ancient Greek sophists Protagoras and Gorgias and modern empiricist philosophers, but the ancient Greek sophists were the first philosophers to assert empirical epistemology, and both philosophical positions emphasize sensation as the source of knowledge. According to the history mining method, sophists can be classified as empiricists rather than rationalists, which seems reasonable.

5.2 Practical Implications

This study has some significance in terms of practice. First, the authors presented a methodology for classifying thoughts or opinions of philosophers expressed in texts into groups. This analytical method can be applied to determine similarities, ideas in various fields such as investment philosophy and business philosophy. In particular, the management philosophy of opinion leaders of companies, such as CEOs, differs from one company to another, affecting the corporate culture as a whole (Gonzalez and McMillan, 1961). In future, an empirical analysis will be conducted to determine the extent to which opinion leaders' management philosophies are similar and how they relate to their business philosophies and corporate performance or culture; the history mining technique presented in this study may be applied most successfully in such cases.

Second, online communities include factual opinions such as satisfaction/dissatisfaction of customers; this construct is difficult to elucidate through surveys (Moniz and de Jong, 2014). These comments can positively impact the word-of-mouth (WOM) effect (Yoo and Gretzel, 2008) and negatively affect the reputation and image of the firm (Golob et al., 2008). Since opinions of such customers provide important information for companies, it is necessary to make decisions on how to respond to customer opinions, especially negative opinions. We suggest that the history mining method proposed in this study can be applied to opinion mining, which grasps the opinions of customers about companies.

Third, in order to develop a specific research field, it is necessary to grasp trends in the field and examine the literature systematically. Many researchers have used data mining techniques to analyze research trends, and recently, data mining techniques have been used to analyze these trends (Ananiadou et al., 2009; Tomas et al., 2011; Moro et al., 2015). Applying history mining techniques to various study topics will enable us to identify the relationships among researchers in a specific field, changes in research trends over time, and major issues.

6. Concluding Remarks

This is a digital humanities study; digital humanities is a fusion of philosophy and information technology. In this study, we suggest a history mining technique to identify and analyze connections between philosophers through their texts. For this purpose, the authors use unstructured text data, which contains data regarding empirical philosophy and rationalist philosophy that were mainstream in Western modern philosophy at various time points. We also introduce several algorithms used in previous classification techniques. Also, the authors evaluated the ability of the method to categorize philosophical atypical text data correctly. In addition, modern empiricist, philosophical thought that influenced rationalist philosophy and influenced data found in the texts on philosophical thought were used. The authors propose an ensemble method consisting of individual classifiers and a history mining method using a weighted ensemble based on the weighted F-scores of individual classifiers. Using the TF-IDF weight values among the texts of philosophers, similarities between philosophers were calculated as Euclidean distance and cosine similarity. The distance matrix is used to visualize the similarity between philosophers. It does not mean that the proposed methodology replaces the conventional analogue methods of intelligent history research, such as content analysis. This is in agreement with Edelstein's (2016) view, which suggests that research using analogue methods is aided by showing that more attention should be paid to reading and interpretation of certain documents in the study of intellectual history. However, Edelstein (2016) proposed a more advanced method than a simple text analysis method using frequency analysis as a digital tool.

References

Akbani, R., S. Kwek, and N. Japkowicz, "Applying Support Vector Machines to Imbalanced Datasets," *Machine Learning: ECML*, (2004), 39-50.

Alghoson, A. M., "Medical Document Classification Based on MeSH," *System Sciences (HICSS), 2014 47th Hawaii International Conference*, IEEE (2014), 2571-2575.

Ananiadou, S., B. Rea, N. Okazaki, R. Procter, and J. Thomas, "Supporting Systematic Reviews using Text Mining," *Social Science Computer Review*, Vol.27, No.1 (2009), 509-523.

Antonie, M. L. and O. R. Zaiane, "Text Document Categorization by Term Association," *Data Mining, 2002. ICDM 2003. Proceedings.* 2002 IEEE *International Conference,* (2002), 19-26.

Bae, J. and B. Watson, "Reinforcing Visual Grouping Cues to Communicate Complex Informational Structure," *IEEE Transactions on Visualization and Computer Graphics*, Vol.20, No.12 (2014), 1973-1982.

Bederson, B. B, "PhotoMesa: A Zoomable Image Browser Using Quantum Treemaps and Bubblemaps." *Proceedings of the Fourteenth Annual ACM Symposium on User Interface Software and Technology,* (2001), 71-80.

Berry, D., "The Computational Turn: Thinking about the Digital Humanities," *Culture Machine*, Vol.12 (2011).

Berry, D. M., E. Borra, A. Helmond, J. C. Plantin, and J. W. Rettberg, "The Data Sprint Approach: Exploring the Field of Digital Humanities through Amazon's Application Programming Interface," *Digital Humanities Quarterly*, Vol.*9*, No.4, (2015).

Blei, D. M., A. Y. Ng and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of machine Learning research*, Vol.3 (2003), 993-1022.

Bouras, C., and V. Tsogkas, "Improving Text Summarization using Noun Retrieval Techniques," *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems* (2008), 593-600.

Carr, O. and D. Estival, "Document Classification in Structured Military Messages," *Proceedings of the Australasian Language Technology Workshop 2003*, (2003), 134-142.

Chen, D., H. M. Müller, and P. W. Sternberg, "Automatic Document Classification of Biological Literature," *BMC bioinformatics*, Vol.7, No.1 (2006), 370.

Chen, Y., Y. Sun, and B. Q. Han, "Improving Classification of Protein Interaction Articles using Context Similarity-Based Feature Selection," *BioMed research international*, Vol.2015 (2015).

Choi, S., J. Jeon, B. Subrata, and O. Kwon, "An Efficient Estimation of Place Brand Image Power based on Text Mining Technology," *Journal of Intelligence and Information Systems*, Vol.21, No.2 (2015), 113~129.

(최석재, 전종식, 권오병, "텍스트마이닝 기반의 효율적인 장소 브랜드 이미지 강도 측정 방법," 지능정보연구, Vol.21, No.2 (2015), 113~129.)

Christians, C. G., "Utilitarianism in Media Ethics and Its Discontents," *Journal of Mass Media Ethics*, Vol.22, No.2-3 (2007), 113-131.

Cohen, M. R., “Hegel's Rationalism,” *The Philosophical Review*, Vol.41, No.3 (1932), 283-301.

Cross, W. R., *The Burned-over District: The Social and Intellectual History of Enthusiastic Religion in Western New York, 1800–1850*, Cornell University Press, New York, 2015.

Dasgupta, A., P. Drineas, B. Harb, V. Josifovski, and M. W. Mahoney, “Feature Selection Methods for Text Classification,” *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, (2007), 230-239.

Dhillon, I. S., and D. S. Modha, “Concept Decompositions for Large Sparse Text Data using Clustering,” *Machine learning*, Vol.42, No.1 (2001), 143-175.

Dodds, E. R., “Plato and the Irrational,” *The Journal of Hellenic Studies*, Vol.65, (1945), 16-25.

Edelstein, D., “Intellectual History and Digital Humanities,” *Modern Intellectual History*, Vol.13, No.1 (2016), 237-246.

Fung, G. P. C., J. X. Yu, H. Wang, D. W. Cheung, and H. Liu, “A Balanced Ensemble Approach to Weighting Classifiers for Text Classification,” *Data Mining, 2006. ICDM'06. Sixth International Conference*, (2006), 869-873.

Gainor, R., S. Sinclair, S. Ruecker, M. Patey, and S. Gabriele, “A Mandala Browser User Study: Visualizing XML Versions of Shakespeare's Plays,” *Visible Language*, Vol.43, No.1 (2009), 60.

Gold, M. K., *Debates in the Digital Humanities*, U of Minnesota Press, London, 2012.

Golob, U., M. Lah, and Z. Jančič, “Value Orientations and Consumer Expectations of Corporate Social Responsibility,” *Journal of Marketing Communications*, Vol.14, No.2 (2008), 83-96.

Gonzalez, R. F., and C. McMillian, “The Universality of American Management Philosophy,” *Academy of Management Journal*, Vol.4, No.1 (1961), 33-41.

Hall, P., *Cities of Tomorrow: An Intellectual History of Urban Planning and Design Since 1880*, John Wiley & Sons, Hoboken, 2014.

Han, B., Z. Obradovic, Z. Z. Hu, C. H. Wu, and S. Vucetic, “Substring Selection for Biomedical Document Classification,” *Bioinformatics*, Vol.22, No.17 (2006), 2136-2142.

Higham, J., “Intellectual History and its Neighbors,” *Journal of the History of Ideas*, Vol.15, No.3 (1954), 339-347.

Hossain, F. A., “A Critical Analysis of Empiricism,” *Open Journal of Philosophy*, Vol.4, No.3 (2014), 225-230.

Hotho, A., A. Nürnberger, and G. Paaß., “A Brief Survey of Text Mining,” *In Ldv Forum*, Vol.20, No.1, (2005), 19-62.

Huang, A. “Similarity Measures for Text Document Clustering,” *Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008), Christchurch, New Zealand,* (2008), 49-56.

Hunnicutt, B. J., and M. Krzywinski, “Points of View: Pathways,” *Nature methods*, Vol.13, No.1 (2016), 5-5.

Jessop, M., “Digital Visualization as a Scholarly

Activity," *Literary and Linguistic Computing*, Vol.23, No.3 (2008), 281-293.

Jessop, M., "The Inhibition of Geographical Information in Digital Humanities Scholarship," *Literary and Linguistic Computing*, Vol.23, No.1 (2007), 39-50.

Jindal, R., R. Malhotra, and A. Jain, "Techniques for Text Classification: Literature Review and Current Trends," *Webology*, Vol.12, No.2, (2015), 1-28.

Kerber, L. K., *Toward an Intellectual History of Women: Essays by Linda K. Kerber*, UNC Press Books, North Carolina, 2014.

Kim, J. and O. Kwon, "A Method of Predicting Service Time based on Voice of Customer Data," *Journal of the Korea society of IT services*, Vol. 15 (2016), 197~210.

(김정훈, 권오병, "고객의 소리 (VOC) 데이터를 활용한 서비스 처리 시간 예측방법," *한국 IT 서비스학회지*, Vol.15 (2016), 197~210.)

Korde, V. and C. N. Mahender, "Text Classification and Classifiers: A Survey," *International Journal of Artificial Intelligence & Applications*, Vol.3, No.2 (2012), 85.

Lauxtermann, P. F. H., "Hegel and Schopenhauer as Partisans of Goethe's Theory of Color," *Journal of the History of Ideas*, Vol.51, No.4 (1990), 599-624.

Kwon, O. and J. S. Lee, "Smarter Classification for Imbalanced Data Set and Its Application to Patent Evaluation," *Journal of Intelligence and Information Systems*, Vol.20, No.1 (2014), 15~34.

(권오병, 이상연, "불균형 데이터 집합에 대한 스마트 분류방법과 특허 평가에의 응용," *지능정보연구*, Vol.20, No.1 (2014), 15~34.)

Lee, H., Jin, Y., & Kwon, O. "Investigating the Impact of Corporate Social Responsibility on Firm's Short-and Long-Term Performance with Online Text Analytics," *Journal of Intelligence and Information Systems*, Vol. 22, No.2 (2016), 13-31.

Lin, Y. W., "Transdisciplinarity and Digital Humanities: Lessons Learned from Developing Text-Mining Tools for Textual Analysis," *Understanding Digital Humanities,* (2012), 295-314.

Lord, G., M. N. Smith, M. G. Kirschenbaum, T. Clement, Auvil, L. Auvil, J. Rose, B. Yu, and C. Plaisant., "Exploring Erotics in Emily Dickinson's Correspondence with Text Mining and Visual Interfaces," *Digital Libraries, 2006. JCDL'06. Proceedings of the 6th ACM/IEEE-CS Joint Conference*, (2006), 141-150.

Martin, M. *Proposal for a Digital Humanities,* Center at Princeton University, 2013. Available at https://digitalhumanities.princeton.edu/files/2013/08/Proposal-for-a-Digital-Humanities-Center-at-Princeton-University3.11.pdf. (Downloaded 21 January, 2017).

Michura, Piotr, S. Ruecker, M. Radzikowska, and C. Fiorentino, "The Novel as a List of Words." *The Potential and Limitations of a List: An International Transdisciplinary Workshop. Center for Theoretical Study, Charles U and Philosophical Inst. of the Acad. of the Sciences of the Czech Republic,* 2007.

Moniz, A., and F Jong, "Sentiment Analysis and the Impact of Employee Satisfaction on Firm Earnings," *In European Conference on Information Retrieval* (2014), 519-527.

Moro, S., P. Cortez, and P. Rita, "Business

Intelligence in Banking: A Literature Analysis from 2002 to 2013 using Text Mining and Latent Dirichlet Allocation," *Expert Systems with Applications*, Vol.42, No.3 (2015), 1314-1324.

Nelson, R. K., "Digital Humanities as Appendix," *American Quarterly*, Vol.68, No.1 (2016), 131-136.

Olivecrona, K., "The Will of the Sovereign: Some Reflections on Bentham's Concept of a Law," *The American Journal of Jurisprudence*, Vol.20, No.1 (1975), 95-110.

Powell, R. J., *An Experimental Examination of Visual Grouping Techniques in Skip Patterns on Respondent Navigation Errors*, University of Nebraska – Lincoln, 2016, Available at http://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1008&context=sramdiss (Downloaded 21 January, 2017).

Roberts-Smith, J., S. DeSAouza-Coelho, T. M. Dobson, S. Gabriele, O. Rodriguez-Arenas, S. Ruecker, and D. Jakacki, "Visualizing Theatrical Text: From Watching the Script to the Simulated Environment for Theatre (SET)," *Digital Humanities Quarterly*, Vol.7, No.3, (2013).

Rosa, K. D., J. Ellen, "Text Classification Methodologies Applied to Micro-text in Military Chat," *Machine Learning and Applications, 2009. ICMLA'09. International Conference*, (2009), 710-714.

Ross, S., amd J. Sayers, "Modernism Meets Digital Humanities," *Literature Compass*, Vol.11, No.9 (2014), 625-633.

Sattelmeyer, R. *Thoreau's Reading: A Study in Intellectual History with Bibliographical Catalogue*, Princeton University Press, New Jersey, 2014.

Schreibman, S., R. Siemens, and J. Unsworth. *Introduction*, in Schreibman et al. (eds.) A Companion to Digital Humanities. Oxford: Blackwell, 2004.

Sculley, D. and B. M. Pasanek, "Meaning and Mining: the Impact of Implicit Assumptions in Data Mining for the Humanities," *Literary and Linguistic Computing*, Vol.23, No.4 (2008), 409-424.

Sebastiani, F., "Machine Learning in Automated Text Categorization," *ACM Computing Surveys*, Vol.34, No.1 (2002), 1-47.

Sinclair, S., S. Ruecker, and M. Radzikowska, "Information Visualization for Humanities Scholars," *Literary Studies in the Digital Age-An Evolving Anthology*, (2013)

Sinclair, S., D. Sondheim, C. Warwick, and J. Windsor, "Introduction to Designing Interactive Reading Environments for the Online Scholarly Edition," *Digital Humanities 2012*, (2012), 36.

Skorupski, J., *The Place of Utilitarianism in Mill's Philosophy*. Utilitarianism, Wiley-Blackwell, New Jersey, 2008.

Small, H. G., "Cited Documents as Concept Symbols," *Social Studies of Science*, Vol.8, No.3 (1978), 327-340.

Stiltner, B., "Who can Understand Abraham? The Relation of God and Morality in Kierkegaard and Aquinas," *The Journal of Religious Ethics*, Vol.12, No.2 (1993), 221-245.

Thomas, J., J. McNaught, and S. Ananiadou, "Applications of Text Mining within Systematic Reviews," *Research Synthesis Methods*, Vol.2, No.1 (2011), 1-14.

Vanzo, A., “Kant on Empiricism and Rationalism,” *History of Philosophy Quarterly*, Vol.30, No.1 (2013), 53-74.

Wang, T. Y. and H. M. Chiang, “Solving Multi-Label Text Categorization Problem using Support Vector Machine Approach with Membership Function,” *Neurocomputing*, Vol.74, No.17 (2011), 3682-3689.

Wilkens, M., “Digital Humanities and Its Application in the Study of Literature and Culture,” *Comparative Literature*, Vol.67, No.1 (2015), 11-20.

Xia, R., C. Zong, and S. Li, “Ensemble of Feature Sets and Classification Algorithms for Sentiment Classification. *Information Sciences*, Vol.181, No.6 (2011), 1138-1152.

Yadav, K., E. Sarioglu, M. Smith, H. A. Choi, and C. D. Newgard, “Automated Outcome Classification of Emergency Department Computed Tomography Imaging Reports,” *Academic Emergency Medicine*, Vol.20, No.8 (2013), 848-854.

Yano, H., Y. Nakajima, K. Ueda, and G. B. Remijn, “The Effect of Sound on Visual Grouping in a Multi-Stable Stimulus,” *International Journal of Psychology*, Vol.51, (2016), 1027.

Yoo, K. H. and U. Gretzel, “What Motivates Consumers to Write Online Travel Reviews?,” *Information Technology & Tourism*, Vol.10, No.4 (2008), 283-295.

Yu, B., “An Evaluation of Text Classification Methods for Literary Study,” *Literary and Linguistic Computing*, Vol.23, No.3 (2008), 327-343.

국문요약

디지털 인문학에서 비정형 데이터 분석을 이용한 사조 분류 방법

서한솔* · 권오병**

최근 디지털 인문학 (Digital humanities) 연구분야의 등장으로 정보기술을 활용하여 인문학 연구의 효율성 제고에 기여하고 있다. 특히 인문학 연구에서 특정한 인물 혹은 문서가 어떠한 사상 (idea)을 담고 있는지와 다른 사상과의 어떤 연결성을 가지는지를 자동적인 방법으로 분석하는 것은 지성사 (intellectual history)를 파악하는 데 중요한 도전이 될 것이다. 본 연구의 목적은 책이나 논문, 기사와 같은 비정형 데이터 (unstructured data)에 포함된 주장을 파악하고 이를 다른 주장이나 사상과 어떠한 관련이 있는지를 자동으로 분석하는 방법을 제안하는 것이다. 특히 본 연구에서는 주장과 주장 사이의 영향관계를 밝히는 히스토리 마이닝 (History Mining)이라는 방법도 제안하였다. 이를 위해 딥러닝 기법 (deep learning method)을 포함한 분류알고리즘 기법 (classification algorithm)을 활용하였다.

본 연구가 제안하는 방법론의 성능을 검증하기 위하여 철학 사조 중에서 대표적으로 대비되는 경험주의와 합리주의 관련 철학자들을 선정하고 관련된 저서 혹은 인터넷 상의 글을 수집하였다. 분류 알고리즘의 성능은 Recall, Precision, F-Score 및 Elapsed Time으로 측정하였으며 DNN, Random Forest, 그리고 앙상블 등이 우수한 성능을 보였다. 선정된 분류 알고리즘으로 특정 철학자의 글에 대해 합리주의 혹은 경험주의로 분류하였으며, 그 철학자의 활동 연도를 고려하여 히스토리 맵을 생성할 수 있었다.

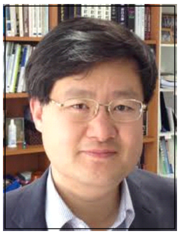
주제어 : 디지털 인문학, 히스토리 마이닝, 텍스트 분석, 철학, 분류 알고리즘

논문접수일 : 2017년 11월 10일 논문수정일 : 2018년 2월 28일 게재확정일 : 2018년 3월 9일
원고유형 : 일반논문 교신저자 : 권오병

* Researcher at School of Management Kyung Hee University
** 교신저자 : 권오병
Professor, School of Management, Kyung Hee University
26, Kyungheedae-ro, Dongdaemun-gu, Seoul 130-701, Republic of Korea
Tel: +82 2 961 2148, Fax: +82 2 961 0515, obkwon@khu.ac.kr

저 자 소 개

서 한 솔
현재 마크로젠에서 연구원으로 재직 중이다. 경희대학교 경영대학에서 학사 및 빅데이터 전공으로 석사학위를 취득하였다. 로보틱스 및 행복지수 기반의 큐레이션 시스템 관련 정부과제를 수행한 바 있으며, 관심분야로는 빅데이터 분석, 휴먼로봇인터페이스, 경영정보시스템 등이다.



권 오 병
현재 경희대학교 경영대학 교수로 재직 중이다. 서울대학교 경영학과에서 학사학위를 한국과학기술원에서 석사 및 박사학위를 취득하였고, 카네기멜론대학 ISRI연구소에서 유비쿼터스 컴퓨팅 프로젝트를 수행한 바 있다. 관심분야는 텍스트 분석, 휴먼로봇인터페이스, 상황인식 서비스, IT비즈니스, 의사결정지원시스템 등이다.