World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:9, No:10, 2015

A Methodology for Automatic Diversification of Document Categories

Dasom Kim, Chen Liu, Myungsu Lim, Soo-Hyeon Jeon, Byeoung Kug Jeon, Kee-Young Kwahk, Namgyu Kim

***Abstract*—**Recently, numerous documents including large volumes of unstructured data and text have been created because of the rapid increase in the use of social media and the Internet. Usually, these documents are categorized for the convenience of users. Because the accuracy of manual categorization is not guaranteed, and such categorization requires a large amount of time and incurs huge costs. Many studies on automatic categorization have been conducted to help mitigate the limitations of manual categorization. Unfortunately, most of these methods cannot be applied to categorize complex documents with multiple topics because they work on the assumption that individual documents can be categorized into single categories only. Therefore, to overcome this limitation, some studies have attempted to categorize each document into multiple categories. However, the learning process employed in these studies involves training using a multi-categorized document set. These methods therefore cannot be applied to the multi-categorization of most documents unless multi-categorized training sets using traditional multi-categorization algorithms are provided. To overcome this limitation, in this study, we review our novel methodology for extending the category of a single-categorized document to multiple categorizes, and then introduce a survey-based verification scenario for estimating the accuracy of our automatic categorization methodology.

***Keywords*—**Big Data Analysis, Document Classification, Text Mining, Topic Analysis.

I. Introduction

NUMEROUS documents including large volumes of unstructured data and text have been created by the increasing number of users of the Internet and social media. Usually, these documents are categorized for the convenience of users. In the past, the categorization of documents was performed manually. However, the accuracy of manual categorization is uncertain, and such categorization requires a large amount of time and incurs huge costs. Many studies on automatic categorization have been conducted to mitigate the limitations of manual categorization. Unfortunately, most of these methods are not applicable to complex documents with multiple topics because they work on the assumption that each individual document has to be categorized into individual categories. For instance, the document in Fig. 1 should be classified into "Sports", "Politics", and "Entertainment" categories because the document is about Arnold Schwarzenegger who is a politician as well as a former famous movie star and body builder.

Dasom Kim, Chen Liu, Myungsu Lim, Soo-Hyeon Jeon, Byeoung Kug Jeon, Kee-Young Kwahk and Namgyu Kim are with the Graduate School of Business IT, Kookmin University, Seoul, 136-702 Republic of Korea (e-mail: dskim1225@kookmin.ac.kr, liuchen@kookmin.ac.kr, amr2001@kookmin.ac.kr, shjeon@kookmin.ac.kr, boungkug@kookmin.ac.kr, kykwahk@kookmin.ac.kr, ngkim@kookmin.ac.kr).

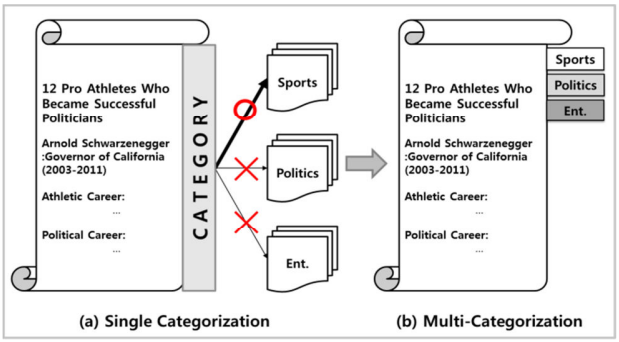

Fig. 1 Needs of multi-categorization for a complex document

Fig. 1 describes the need for multi-categorization of a complex document. In reality, most documents need to be classified into multiple categories because they usually deal with complex subjects. Therefore, many studies have attempted to devise a document classifier that can categorize each document into multiple categories. However, these studies have some common limitations that they cannot be directly applied to real cases without previously multi-categorized documents. It implies that we cannot use these classifiers if we have only single categorized documents.

Our proposed method can diversify the number of categories of a single-categorized document to multiple categories by analyzing the relationships among categories and the topics of documents [1]. The scope of our research can be graphically described by the rectangle in the upper half of Fig. 2; the dotted rectangle in the lower part of the figure represents the scope of traditional multi-classifiers using multi-categorized training sets.

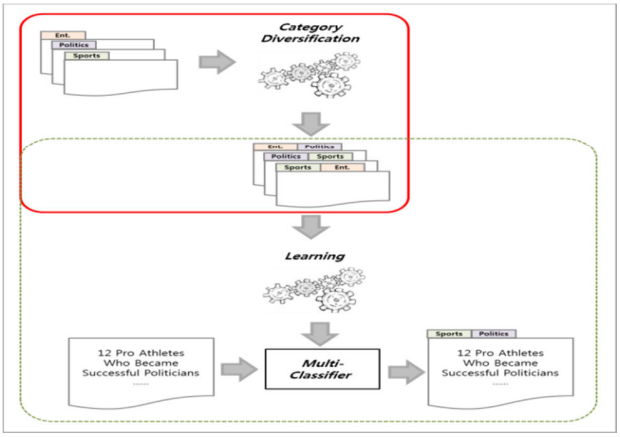

Fig. 2 Research scope and related application area

Although the main idea of category diversification was originally introduced in our previous paper [1], we did not provide any concrete method for evaluating the performance of the proposed approach. In this paper, therefore, we propose a survey-based verification scenario for estimating the accuracy of our automatic category diversification method.

The remainder of this paper is organized as follows: The next section introduces related works on text mining and document classification. In Section III, our methodology for automatic category diversification [1] is revisited, and the experimental results of the methodology are summarized. Section IV proposes a new survey-based verification scenario for estimating the accuracy of our methodology. Finally, Section V concludes this study.

II. Related Works

A. Text Mining

Text is the most representative method to express and communicate information in the real world [2]. Text mining is a sequence of analyzing processes employed to extract useful information from voluminous text [3]–[6]. Currently, there are attempts being made to discover valuable knowledge by using text mining techniques. Identifying the original document of some documents [7]; discovering new crimes by analyzing patterns in previous crimes, structuring unstructured storage by text categorization are some examples of text mining applications.

Text mining not only utilizes the association, classification, and clustering techniques that have been used in traditional data mining applications, but also additional techniques of natural language processing, information retrieval, issue tracking, and text categorization areas [4], [8], [9]. Among the techniques mentioned above, natural language processing is regarded as the core technique used by text mining applications. While traditional structured data is presented in the form of two-dimensional tables, text data is presented in the form of unstructured documents. Therefore, many useful techniques for structuring text data into matrices, hierarchies, and vectors have been developed. The most fundamental and widely used technique is the vector space model [10], [11] that summarizes the frequencies of terms in each document.

Among the various contemporary text mining related applications, topic analysis draws the most attention from researchers and practitioners. With the theoretical foundations of the vector space model and a TF-IDF (term frequency–inverse document frequency) measure [12], the main process of topic analysis is usually performed on structured documents after parsing and filtering. In the parsing stage, sentences of the documents are separated into tokens, and some tokens are eliminated based on certain predefined conditions in the filtering stage. Topic analysis is similar to traditional clustering techniques, in that their goals are to group similar objects and separate the dissimilar ones. However, topic analysis can map each document to multiple topics, whereas in traditional clustering algorithms, each element can belong to only one specific cluster.

B. Document Classification

Conventionally, document classification was processed manually, and therefore, it was a time consuming and error-prone process. However, in recent times, the rapidly increasing volume of online documents has accelerated the need for automatic document classification services. As a result, some automatic document classifiers [13]–[15] have been devised to classify each document into a certain category based on certain predefined rules.

Most studies on automatic document classifiers have been conducted using machine learning methods. For example, many classifiers are devised based on the KNN (K-nearest neighbors) [16], naïve Bayesian model [17], ANN (artificial neural network) [18], and SVM (support vector machine) methods [19]. Attempts to optimize the performance using existing classifiers have also been made [15], [20]. However, traditional document classifiers, unfortunately, adopt the strict and unrealistic assumption that each document can be classified into only one category. Therefore, such single-category classifiers cannot be used to identify multiple categories of a complex document. Consequently, recent studies for classifying multiple categories have drawn considerable attention from the research fraternity. For example, in [21], a multi-category (i.e., multi-label) classifier was devised by utilizing three types of correlations: between categories, between features, and between categories and features. In [22], a method to enhance the performance of multi-label classifiers utilizing subsets of features of each document was proposed. However, most multi-category classifiers have limitations, in that they involve training using a multi-categorized document set in their learning process. It implies that traditional multi-category classifiers cannot be used for multi-categorization of documents unless multi-categorized training sets are provided.

III. A Methodology for Automatic Category Diversification

A. Research Overview

In this section, we introduce a method to identify multiple categories from single-categorized documents. The overall process overview of our research is shown in Fig. 3. First, we attempt to determine the relationship between documents and topics by using the results of topic analysis for single-categorized documents. Second, we construct a correspondence table between topics and categories by investigating the relationship between them. Finally, we identify the possible additional categories for each document by calculating the matching scores of individual documents to each category. Details and examples of the above three modules are presented in the successive subsections.

The overall process shown in Fig. 3 is an extended version of the model presented in our previous work [1]. The extended part is included in this research for performance evaluation; this part is represented by a dotted rectangle in Fig. 3. The accuracy analysis plan is presented in the next section by using a simple example.

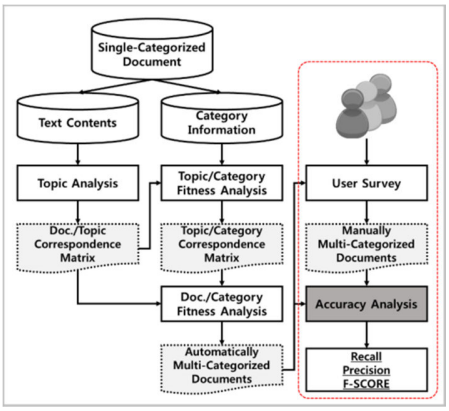

Fig. 3 Research overview

B. Topic Analysis

In the topic analysis stage, each document is featured by the frequencies of its containing terms. Frequency can be represented by binary values, absolute frequencies, or TF–IDF values. Subsequently, the similarity between documents is calculated based on the term frequencies. Finally, similar documents are grouped together and a set of representative keywords for each group is selected. Topic analysis differs from traditional clustering methods, in that each document in topic analysis can belong to multiple topics concurrently. We will not repeat the details of topic analysis in this paper as it has already been introduced in many papers and can be performed easily by using many commercial mining tools.

A simple example of topic analysis for single-categorized documents is presented in Fig. 4. In the figure, "Original Category" refers to the original category of the document. Each numeric cell contains a D/T Score that represents the degree of correspondence between each topic and each document.

Doc. No	Original Category	Topic 1	Topic 2	Topic 3	Topic 4
DOC_1	Sports	1.2	0.0	1.7	1.3
DOC_2	Sports	1.9	1.7	0.0	0.0
DOC_3	World	1.7	0.0	1.2	0.0
DOC_4	World	0.0	1.8	0.0	1.9
DOC_5	Entertainment	2.1	0.0	0.0	2.1
DOC_6	Entertainment	0.0	1.2	0.0	0.9

Fig. 4 Document/Topic correspondence score (D/T Score)

C. Analyzing Topic/Category Correspondence

In this stage, the degree of correspondence between each topic and each category (i.e., T/C Score) is calculated based on the D/T Score. In Fig. 4, for example, "Topic1" is related to the category "Sports" through documents "DOC_1" and "DOC_2". Additionally, "Topic1" is related to the category "World" through the document "DOC_3" and related to the category "Entertainment" through the document "DOC_5". It should be noted here that "Topic1" is related to three categories. A process to calculate T/C Scores from the table in Fig. 4 is described in Fig. 5. In Fig. 4, for example, "Topic 1" is related to the category "Sports" through documents "DOC_1" and "DOC_2". The D/T Scores of the two mappings are "1.2" and "1.9", respectively, in Fig. 4. By summing up the two D/T Scores, we can acquire the T/C Score of "Topic 1" to "Sports" in Fig. 5.

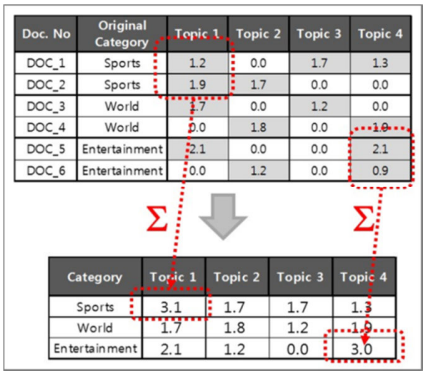

Fig. 5 Topic/Category correspondence score (T/C Score)

D. Analyzing Document/Category Correspondence

In the last stage of our methodology, we calculate the degree of correspondence between each document and each category (i.e., D/C Score). Diversified categories of each document can be easily selected based on D/C Scores. D/C Score is defined as the weighted sum of D/T Score and T/C Score. D/T Score can be regarded as a raw value, and T/C Score works as a weight. The process for calculating D/C Scores is shown in Fig. 6.

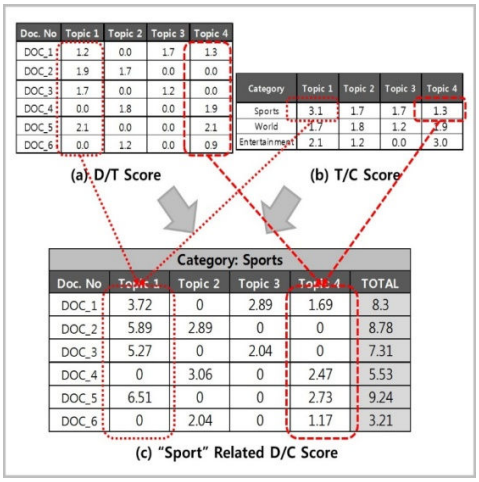

Fig. 6 Document/Category correspondence score for the "Sports" category

Fig. 6 shows the process for calculating the D/C Scores of each document for the category "Sports". For example, we can acquire the partial D/C Scores of "Topic 1" in Fig. 6 (c) by weighting the D/T Scores of "Topic 1" in Fig. 6 (a) with the T/C Scores of "Topic 1" and "Sports" in Fig. 6 (b), and then summing up the weighted D/T Scores. The partial D/C Scores of the other three topics can be calculated in a similar manner. The last column of the table in Fig. 6 (c) contains the final D/C Scores, which are calculated by summing up all values of the corresponding rows. For example, the D/C Score of "DOC_1" for the category "Sports" is "8.3", as seen in the last column of the first row in Fig. 6 (c). Although Fig. 6 only shows the D/C Scores of each document for the category "Sports", other D/C

Scores for the categories "World" and "Entertainment" can be obtained in a similar manner. The D/C Scores for the three categories are summarized in Fig. 7.

Doc. No	Original Category	Sports	World	Entertainment	1st	2rd
DOC_1	Sports	8.3	6.55	6.42	Sports	World
DOC_2	Sports	8.78	6.29	6.03	Sports	World
DOC_3	World	7.31	4.33	3.57	Sports	World
DOC_4	World	5.53	6.85	7.86	Entertainment	World
DOC_5	Entertainment	9.24	7.56	10.71	Entertainment	Sports
DOC_6	Entertainment	3.21	3.87	4.14	Entertainment	World

Fig. 7 Summarized Document/Category correspondence

Fig. 7 summarizes the D/C Scores of six documents for the categories "Sports", "World", and "Entertainment". In the case of "DOC_1", the highest D/C Score is obtained for the category "Sports" and the second highest D/C Score is obtained for the category "World". It should be noted that the category "Sports" is the original category of "DOC_1". This correspondence between the original category and a category with the highest D/C Score shows the robustness of the proposed method. The category with the second highest D/C Score helps us infer that "DOC_1" contains world-related issues/content.

E. Experiments

In this subsection, we show the experimental results obtained when our methodology is applied to diversified categories of real news articles. The news articles for the experiment were obtained from one of the largest website portals for news in Korea. The classified categories are "IT Science", "Economy", "Society", "Life and Culture", "World", "Sports", "Entertainment", and "Politics". We sampled 3,000 articles from each category; therefore, the total number of articles used in this experiment was 24,000.

Initially, we discovered 50 topics from 24,000 documents (i.e., articles) using the text miner module in SAS Enterprise Miner 12.1. The results of topic analysis were directly converted into D/T Scores. Next, we calculated the T/C Scores for 50 topics and 8 categories. The partial result of the T/C Score calculation is shown in Fig. 8. Therefore, we can infer that "Topic_1" is mainly related to the category "Sports". Additionally, "Topic_1" appears to be related to the categories "Life and Culture" and "Entertainment".

Category_ID	Category	Topic_1	Topic_2	Topic_3	Topic_4	Topic_5
1	IT Science	1311.72	1974.989	3296.735	83.58	7225.412
2	Economy	1448.035	4478.807	6226.245	212.892	615.042
3	Society	2289.974	2452.705	748.254	368.459	42.696
4	Life and Culture	3616.108	1022.092	1233.156	87.688	638.274
5	World	973.752	1709.813	435.561	492.742	46.422
6	Sports	5212.342	263.014	251.562	20.8	9.688
7	Entertainment	2739.781	161.521	36.913	54.663	6.71
8	Politics	1247.382	6563.044	358.414	7302.076	13.795

Fig. 8 Topic/Category correspondence score (T/C Score) (partial)

Fig. 9 shows a section of the D/C Score obtained from the weighted sum of D/T Score and T/C Score. In each row, the highest value can be identified by the shaded cell.

In Fig. 9, the top three categories in each document are selected and summarized in Fig. 10. Note that in each document, only one category is shaded. The shaded category implies that the category coincides with the original category of the document.

Doc. No	IT Science	Economy	Society	Life and Culture	World	Sports	Ent.	Politics
17	**46790**	5860	2298	5614	3804	744	2071	1511
29	57772	**68276**	26427	23884	13916	8392	3433	38045
170	**115525**	80428	30669	30564	22425	12126	3657	41818
171	**77590**	49748	15862	22882	7317	4079	1056	9096
185	56204	29384	6370	12223	5274	**58091**	11600	2400
5214	15734	**31693**	6428	7802	5875	2765	1392	4813
5228	**64649**	50096	17406	22025	9932	6279	1847	10414
5229	25090	**55054**	21662	13954	12014	5677	2334	37638
5255	39518	**72836**	26093	24018	14127	7761	3338	36747
5256	46674	**77628**	45523	43299	20164	35806	18807	44891
9392	2260	3979	10298	**16105**	4836	1844	3478	4311
9403	1362	1681	3422	**12296**	1645	603	1146	1290
9404	1005	1086	1113	**2541**	2009	609	722	994
9405	**46289**	39045	16836	32619	7866	26923	14407	8398
9507	4319	11630	25717	27471	9989	3296	2232	**31166**

Fig. 9 Summarized Document/Category correspondence score (partial)

Doc_Info		Ranking			Corresponded Category by Rank		
Doc. No	Original Category	1st	2nd	3rd	1st	2nd	3rd
17	IT Science	46790	5860	5614	IT Science	Economy	Life and Culture
29	IT Science	68276	57772	38045	Economy	IT Science	Politics
170	IT Science	115525	80428	41818	IT Science	Economy	Politics
171	IT Science	77590	49748	22882	IT Science	Economy	Life and Culture
185	IT Science	58091	56204	29384	Sports	IT Science	Economy
5214	Economy	31693	15734	7802	Economy	IT Science	Life and Culture
5228	Economy	64649	50096	22025	IT Science	Economy	Life and Culture
5229	Economy	55054	37638	25090	Economy	Politics	IT Science
5255	Economy	72836	39518	36747	Economy	IT Science	Politics
5256	Economy	77628	46674	45523	Economy	IT Science	Society
9392	Life and Culture	16105	10298	4836	Life and Culture	Society	World
9403	Life and Culture	12296	3422	1681	Life and Culture	Society	Economy
9404	Life and Culture	2541	2009	1113	Life and Culture	World	Society
9405	Life and Culture	46289	39045	32619	IT Science	Economy	Life and Culture
9507	Life and Culture	31166	27471	25717	Politics	Life and Culture	Society

Fig. 10 Top three categories of each document (part)

IV. Accuracy Evaluation Scenario

In this section, we introduce a survey-based verification scenario for estimating the accuracy of our automatic categorization methodology. Although the principle idea of category diversification was originally introduced in our previous paper, the paper did not provide any concrete methods for evaluating the performance of the proposed approach. Therefore, we are currently performing intensive experiments for estimating the accuracy of our methodology, as shown in the dotted rectangle in Fig. 3.

An overview of our evaluation scenario is as follows. First, we sample 10 articles from each of the 8 categories, which imply that the total number of articles sampled is 80. Next, we compose 10 document groups, each of which comprises 8 articles from each category. Subsequently, we compose 10 user groups, each of which comprises 5 users of Internet news. After that, we assign all the document groups to individual user groups such that one user reviews 8 articles from each of the 8 categories, and every document is reviewed by 5 users. Fig. 11 illustrates this scenario. Finally, users select the best and the second best category for each document they review. The obtained 400 responses are used for evaluating the accuracy of our methodology. Now, we have two sets of documents with diversified categories. The first set contains automatically diversified categories using the proposed method, whereas the

other contains manually selected categories from the survey in Fig. 11. We can measure the accuracy of our methodology by investigating the consistency between the categories of the two sets.

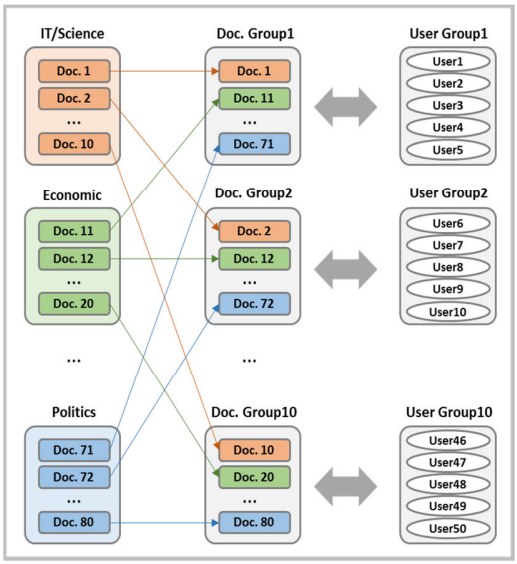

Fig. 11 Mapping plan between user groups and document groups

In this evaluation scenario, the accuracy of the classifier can be measured using an F-Score, which is defined by the harmonic mean of Precision and Recall. To exclude unreliable responses, we consider a response invalid if the "best" category remarked in the response does not coincide with the original category. This implies that the selected two categories of a valid response will always contain the original category of the document. Let us assume that a notation Doc_d^{man} represents a set of categories that are manually selected at least once as the best or second best category of a document d in valid responses. Additionally, assume that a notation Doc_d^{auto} a set of categories that are automatically recommended at least once as the most or second most appropriate category of a document d by using our methodology.

If a number of elements of a set A can be represented by *count(A)*, the *Precision, Recall,* and *F-Score* can be calculated using:

$$(Presion) = \frac{\sum_{d=1}^{80} count(Doc_d^{auto} \cap Doc_d^{man})}{\sum_{d=1}^{80} count(Doc_d^{auto})}$$

$$(Recall) = \frac{\sum_{d=1}^{80} count(Doc_d^{auto} \cap Doc_d^{man})}{\sum_{d=1}^{80} count(Doc_d^{man})}$$

$$(F-Score) = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

This study is still in progress. In accordance with the evaluation criteria and performance measures, we are currently undertaking user surveys. After the completion of the surveys, we will not only estimate the overall accuracy of our methodology, but also compare the differences in accuracies among various categories.

V. Conclusion

Many kinds of multi-category classifiers have been invented so far. However, they cannot work properly unless multi-categorized training sets are provided. To overcome such limitations, we previously proposed a new methodology that could extend the category of a single-categorized document to multiple categorizes by analyzing the relationships among categories, topics, and documents. In this paper, we proposed a survey-based verification scenario for estimating the accuracy of our methodology. As mentioned earlier, this study is still in progress, and we are performing user surveys currently in accordance with the evaluation criteria and performance measures proposed in this paper. In future studies, we aim to present the difference in accuracies among various categories as well as the overall accuracy of our methodology.

References

[1] J. Hong, N. Kim, and S. Lee, "A Methodology for Automatic Multi-Categorization of Single-Categorized Documents," Journal of Intelligent Information systems, vol. 20, no. 3, pp. 77-92, Sep. 2014.
[2] I. H. Witten, *Text Mining, Practical Handbook of Internet Computing*, CRC Press, 2004.
[3] J. Hong, H. Choi, H. Han, J. Kim, E. Yu, S. Lim, and N. Kim, "A Data Analysis-based Hybrid Methodology for Selecting Pending National Issue Keywords," *Entrue Journal of Information Technology*, vol. 13, pp. 97-111, Jun. 2014.
[4] R. J. Mooney, and R. Bunescu, "Mining Knowledge from Text Using Information Extraction," *ACM SIGKDD Explorations*, vol. 7, pp. 3-10, Jun. 2006.
[5] S. Song, J. Yu, and E. Kim, "Offering System For Major Article Using Text Mining and Data Mining," *Proceedings of the 32th annual conference on Korea Information Processing Society*, pp. 733-734, 2009.
[6] E. Yu, J. Kim, C. Lee, and N. Kim, "Using Ontologies for Semantic Text Mining," *The Journal of Information Systems*, vol. 21, pp. 137-161, Sep. 2012.
[7] D. Metzler, Y. Bernstein, W. B. Croft, A. Moffat, and J. Zobel, "Similarity Measures for Tracking Information Flow," *Proceedings of CIKM, Bremen*, Germany, 2005.
[8] C. J. V. Rijsbergen, *Information Retrieval, 2nd edition*, Butterworth, 1979.
[9] F. Sebastiani, Classification of Text, Automatic, The Encyclopedia of Language and Linguistics 14, 2nd edition, Elsevier Science Pub, 2006.
[10] G. Salton, A. Wong, and C. S. Yang, "A Vector Space Model for Automatic Indexing," *Communications of the ACM*, vol. 18, pp. 613-620, Nov. 1975.
[11] R. Albright, "Taming Text with the SVD," *SAS Institute Inc*., 2006.
[12] G. Salton, and M. J. McGill, *Introduction to Modern Information Retrieval*, McGraw Hill, 1983.
[13] C. Apte, and F. Damerau, "Automated Learning of Decision Rules for Text Categorization," *ACM Transactions on Information Systems*, vol. 12, pp. 233-251, Jul. 1994.
[14] J. Han, and M. Kamber, *Data Mining: Concepts and Techniques, 3rd ed.*, Morgan Kaufmann Publishers, 2011.
[15] H. Lim, and K. Nam, "Computer Science: Improving of KNN - Based Korean Text Classifier by Using Heuristic Information," *The Journal of Korean Association of Computer Education*, vol. 5, pp. 37-44, Jul. 2002.
[16] Y. Yang, "Expert network: Effective and Efficient Learning from Human Decisions in Text Categorization and Retrieval," *Proceedings of the 17th International Conference on Research and Development in Information Retrieval*, SIGIR 94, pp. 13-22, 1994.
[17] D. D. Lewis, and M. Ringuette, "Comparison of Two Learning Algorithms for Text Categorization", *Proceedings of the 13rd Annual Symposium on Document Analysis and Information Retrieval*, pp. 81-93, 1994.

[18] E. Weiner, J. O. Pedersenm, and A. S. Weigend, "A Neural Network Approach to Topic Spotting," *Proceedings of the 14th Annual Symposium on Document Analysis and Information Retrieval*, 1995.
[19] T. Joachims, *Text Categorization with Support Vector Machines: Learning with Many Relevant Features*, Springer Berlin Heidelberg, pp. 137-142, 1998.
[20] J. In, J. Kim, and S. Chae, "Combined Feature Set and Hybrid Feature Selection Method for Effective Document Classification," *Journal of Internet Computing and Services*, vol. 14, pp. 49-57, Oct. 2013.
[21] H. Lim, and D. Kim, "Using Mutual Information for Selecting Features in Multi-label Classification," *Journal of KIISE: Software and Applications*, vol. 39, pp. 806-811, Oct. 2012.
[22] J. Yun, J. Lee, and D. Kim, "Feature Selection in Multi-label Classification Using NSGA-II Algorithm," *Journal of KIISE: Software and Applications*, vol. 40, pp. 133-140, Mar. 2013.