GRiD: Gathering Rich Data from PubMed Using One-class SVM

Junbum Cha
Department of Computer Science,
Yonsei University
Seoul, Korea
khanrc@yonsei.ac.kr

Jeongwoo Kim
Department of Computer Science,
Yonsei University
Seoul, Korea
jwkim2013@yonsei.ac.kr

Sanghyun Park*
Department of Computer Science,
Yonsei University
Seoul, Korea
sanghyun@yonsei.ac.kr

***Abstract*—The Medical Subject Headings (MeSH) term search is typical data-gathering method in biomedical text mining. However, it has two problems: the allocation delay of the MeSH term and missing valuable literature sources. Since MeSH term allocation is performed by a human being, the allocation process has delay. In addition, even if a literature source was allocated with a MeSH term, there is a still the problem that valuable literature sources are missed during the data-gathering process. There are literature sources that are not indexed to the MeSH term of a keyword, even though it contains valuable information related to the MeSH term. The MeSH term search misses these valuable literature sources.**

In order to resolve these problems, we propose a novel method to gather rich data using a one-class support vector machine (SVM) and relevance rule. The term frequency–inverse document frequency (TF-IDF) and paragraph vector are examined as text vectorization methods with various parameters and relevance factors. We apply our method to lung cancer, prostate cancer, breast cancer, and Alzheimer's disease. As a result, up to 26% of keyword data and 35% of target data are gathered with high quality (a C-score of at least 0.948).

I. Introduction

Text mining is data mining using text data and is used to analyze unstructured text for identifying valuable knowledge. The concept of text mining was first introduced in the 1980s and rapidly grew in the 1990s. During the same period of time, the Human Genome Project was started and rapidly generated biomedical data. Naturally, biomedical text mining was born.

As shown in Figure 1, biomedical text mining has rapidly increased. The biomedical text data used for biomedical text mining is typically gathered from the MeSH term search in PubMed. PubMed is a search engine that provides free access to the MEDLINE database of citations and abstracts on life science and biomedical topics. The database is developed and maintained by the National Center for Biotechnology Information (NCBI) at the National Library of Medicine (NLM) [1]. A MeSH term is controlled vocabulary used for indexing literature in the life sciences. As the term is maintained by the NLM and the literature consists of academic papers, it provides reliable information. Therefore, many biomedical text mining researchers use the MeSH term to gather text data for biomedical text mining.

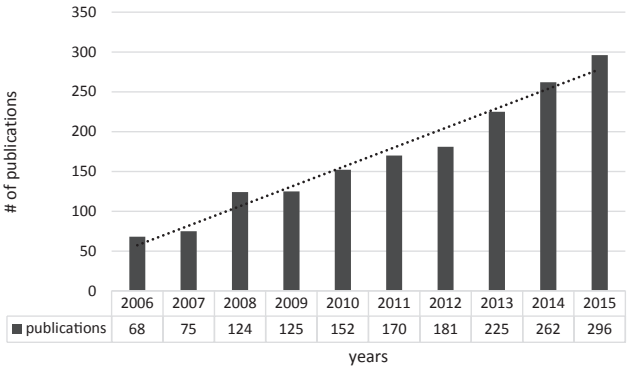


Fig. 1: Trend in biomedical text mining publications over the past 10 years

A. Main Challenge: Gathering Rich Data

A MeSH term search is a highly reliable search method, but it has two problems. The first problem is the allocation delay of the MeSH term. A MeSH term is allocated by human experts at the NLM. As these experts are human beings, the allocation process has some delay. Rodriguez [2] evaluated the time delay required for PubMed indexing by a MeSH term for articles published in major pharmacy practice journals in 2014. According to his work, indexing required an average of 114 days from 2010 to 2011.

Figure 2 shows the trend in the MeSH index ratio for "lung cancer" for five years prior to the search date 2015-07-07. The black bars indicate the accumulated MeSH index ratio, and the gray bars indicate the MeSH index ratio every two months.

The MeSH index ratio in this figure was lower than that found by Rodriguez. Literature sources registered within two months were rarely indexed by MeSH, and only approximately 10% of literature sources registered within four months had been indexed. The MeSH index rate was slightly less than 50% 10 months after the date on which the literature source was registered. Furthermore, approximately 10% of literature sources registered more than two years prior remained

*Corresponding author. Tel.: +82 2 2123 5714; fax: +82 2 365 2579; E-mail address: sanghyun@cs.yonsei.ac.kr.

978-1-5090-1897-0/16/$31.00

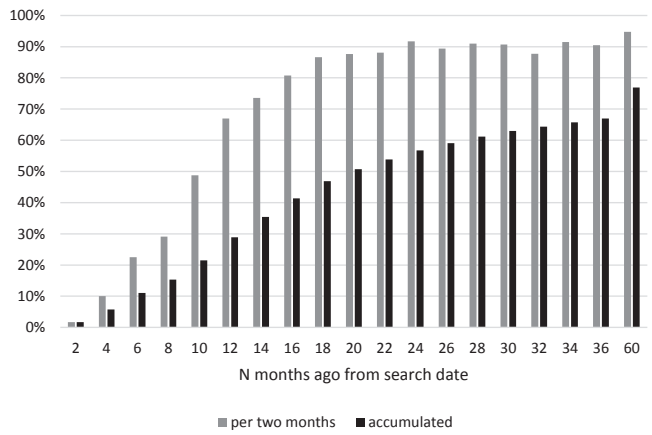


Fig. 2: Changes in the MeSH indexed ratio in the "lung cancer" literature over the last five years

unindexed. Thus, a user of PubMed cannot utilize the latest literature sources using a MeSH term search.

Example 1. The literature contains valuable information about lung cancer but was not indexed to the "lung neoplasms" MeSH term ("lung neoplasms" is a MeSH term for lung cancer).

PMID 23321468. In addition, significant association was maintained in prostate cancer (rs28360071), *lung cancer* (rs6869366) and bladder cancer (rs1805377) subgroups analysis.

PMID 23155281. Malignant pleural effusion (MPE) is common in most patients with advanced cancer, especially in those with *lung cancer*, metastatic breast carcinoma and lymphoma.

The second problem is related to missing valuable literature sources. A missing valuable literature sources is a source that is not indexed to the MeSH term of a keyword, even though it contains valuable information related to the MeSH term. As shown in Example 1, there are literature sources that do not contain the "lung cancer" MeSH term, but they contain valuable information for "lung cancer." This is not a problem for a human who wants to search for a particular subject, but this is a problem when collecting text data for text mining. Therefore, the main challenge of this study is to overcome these problems and gather rich data from PubMed.

B. Taxonomy

In this study, the keyword search results in PubMed are classified, and the name of each class is defined. This taxonomy is described in Figure 3.

- A keyword is a search keyword such as "lung cancer."
- Keyword data are keyword search results.
- MeSH-allocated data are literature sources that have the MeSH term.
- MeSH-keyword data are data in which at least one of MeSH terms is related to the keyword.

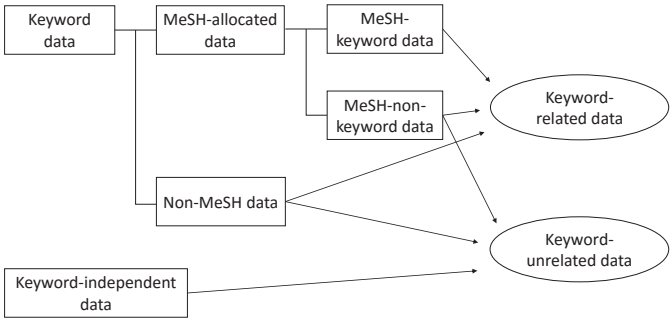


Fig. 3: Taxonomy of this study

- MeSH-non-keyword data are literature sources whose MeSH terms are not related to the keyword at all.
- Non-MeSH data are literature sources that are waiting for an indexing MeSH term.
- Target data, which are not defined in Figure 3, are MeSH-non-keyword data and non-MeSH data. Since these data consist of both related data and unrelated data, they are the target of our system.
- Keyword-independent data are negative data sets that are not related to the keyword at all.
- Keyword-related and keyword-unrelated data are data that is literally related and unrelated to a keyword, respectively.

We assume that all of MeSH-keyword data are related to a keyword because the MeSH term is allocated by the NLM, which is professional institution.

C. Existing Method

Cha [3] proposed the first state-of-the-art method for this challenge of gathering rich data. The method learns a classifier using lung cancer MeSH data as positive data and occupational disease as negative data. The target data are classified into keyword-related or keyword-unrelated data using the learned classifier. However, this method has two problems. The first problem is the selection of negative data. In order to learn a classifier, the system needs negative data that are independent to search keyword. In Cha's paper, this selection is performed by a researcher. However, the ultimate purpose of the challenge is an automated system. Therefore, the automatic selection of negative data is another problem of this method. Even if this problem is overcome and successfully chooses negative data, another problem remains, which is inappropriate data selection. The classifier learned a difference between the positive data and negative data. However, our target data are independent of the negative data. Therefore, the difference is not suitable for the target data.

D. Contributions

In order to overcome the above problems, we propose a novel method that does not use negative data through a one-class SVM. Our contributions are fourfold:

1) Richer data than that using only the MeSH term search are gathered

2) The latest data are available, which are not available in the MeSH term search
3) Negative data are not required
4) A higher accuracy than the existing method is achieved

II. Background

In this study, we carried out experiments with two types of vectorization methods in the preprocessing step. Vectorization transforms unstructured natural text into a fixed size vector. We describe the vectorization methods used in this study in this section. Furthermore, the one-class SVM used in the classification step is described below.

A. TF-IDF Vectorization

The TF-IDF vectorization method applies a TF-IDF analysis to bag-of-words vectors [4]. The TF indicates the importance of a term in a document, while the IDF indicates the importance of a term in a set of documents. The TF can be obtained by calculating the emergence percentage of a term in a document, and the IDF can be calculated from the DF, which is the emergence percentage of a term in the entire documents set. For a term i in a document j,

$$w_{i,j} = tf_{i,j} \times \log(\frac{N}{df_i})$$

$tf_{i,j}$ is the number of occurrences of i in j, df_i is the number of documents containing i, and N is the total number of documents.

Therefore, a high TF-IDF value for a term means that the term appears frequently in the document and does not appear frequently in other documents.

B. Paragraph Vector

Recent advances in neural networks represented by deep learning have had a profound impact on literature vectorization. In 2013, Mikolov et al. proposed a novel word vectorization method using a neural network called word2vec [5]. The main difference between word2vec and the conventional TF-IDF or bag-of-words method is the grouping of vectors of similar words in a vector space. That is, it preserves the semantics of the words.

In a similar context, Le and Mikolov proposed a literature vectorization method called a paragraph vector or doc2vec [6]. It considers the semantics of words as does word2vec and the orders of words. These two properties are the major differences with TF-IDF vectorization.

C. One-class Classification

As mentioned above, our novel approach does not use negative data but only positive data. Thus, this situation corresponds to one-class classification [7], which is also known as unary classification, outlier detection, or novelty detection. The classification identifies objects of a specific class, which is called an outlier or a novel object in each context, by learning only one-class labeled data.

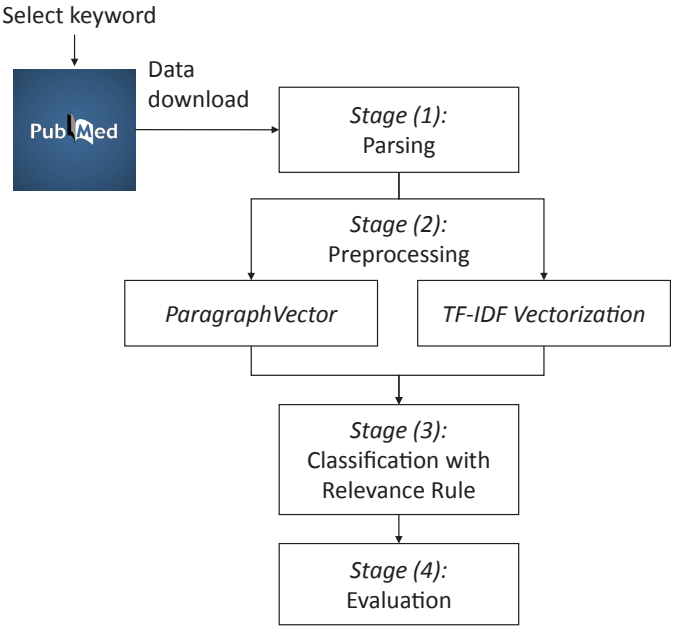


Fig. 4: Overview of the experiment

Among one-class classification algorithms, we selected a one-class SVM [8] that finds a hyperplane that separates a labeled data region from a non-labeled data region, maximizing the distance from the origin.

$$\min_{w,\ \xi_i,\ \rho} \frac{1}{2}\|w\|^2 + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i - \rho$$

subject to:

$$(w \cdot \Phi(x_i)) \geq \rho - \xi_i \text{ for all } i = 1, \ldots, n$$
$$\xi_i \geq 0 \text{ for all } i = 1, \ldots, n$$

where x_i is a data point, w is a normal vector of the hyperplane, n is the number of data points, ξ_i is a slack variable, and ρ is a bias term. Therefore, $\xi_i - \rho$ is the degree of misclassification. Furthermore, ν is a parameter that controls the upper bound on the fraction of outliers. That is, a high value of ν creates a small and strict decision boundary, and a low value of ν creates a large and rough decision boundary.

III. Method

Figure 4 shows the structure of the experiment. First, the keyword data downloaded from PubMed has all of the information about a literature source in the extensible markup language (XML) format. The data were parsed into the PubMed identifier (PMID), literature type, title, and abstract, which are used in the classification stage. The literature type is either MeSH-keyword data, MeSH-non-keyword data, non-MeSH data, or keyword-independent data. The extracted data are vectorized during the preprocessing stage. As mentioned earlier, two vectorization methods were examined. Each method has a similar but different preprocessing step before vectorization. The difference is described in Section III-A.

Next, classification was performed in the third stage. We used a one-class SVM with a relevance rule, which is described in Section III-B. Finally, the classification results need to be evaluated. However, they are difficult to evaluate because there is no answer set. In order to overcome this evaluation problem, we evaluated the results in various ways, as described in Section III-C.

A. Preprocessing

Silva [9] examined the effectiveness of three text refining methods including stop-word removal, low-frequency word removal, and stemming. The results showed that stop-word removal was the most effective method and that the application of all three methods provided the best results. Moreover, Cooley [10] showed that the use of the full text provided better results compared to performing feature reduction for text mining. Therefore, we used all three text refining methods and the full text without feature reduction for TF-IDF vectorization.

Specifically, TF-IDF vectorization has five steps in this experiment. The first step is to split the unstructured text into sets of words. In order to split the text into words well, we removed nonletters such as punctuation and then split the text by spaces. Next, we removed the stop words and low-frequency words that are useless for document classification. The remaining useful words are transformed into word stems through a Porter stemmer [11]. The fourth step is to vectorize these word stems on the basis of the bag-of-words scheme. Lastly, the bag-of-words vectors are transformed into TF-IDF vectors using a TF-IDF analysis, which is a method used to extract the significant terms in a literature source.

In contrast, the paragraph vector method does not adopt stop-word removal and stemming. We utilized the original preprocessing method discussed in the paper on the paragraph vector method [6] and only adjusted the text splitting step to the properties of biomedical text data.

B. Classification with a Relevance Rule

An exploratory data analysis (EDA) is a method for understanding data. It is helpful when little or no hypothesis for the data exists or when a specific hypothesis exists, but the evidence to support the hypotheses is lacking [12].

In order to understand biomedical text data, we observed the properties of data through an EDA. The results of an EDA showed that it is very important that a keyword is contained in the title of the literature source. Therefore, we assigned a relevance factor (RF) to each literature source according to the title. The RF value determines how strict the decision boundary is through $\nu = 1 - \text{RF}$. Consequently, the optimization equation of the one-class SVM is rewritten as follows:

$$\min_{w,\ \xi_i,\ \rho} \frac{1}{2}\|w\|^2 + \frac{1}{(1-\text{RF})n}\sum_{i=1}^{n}\xi_i - \rho$$

We examined various RF values for each case of the title containing and not containing a keyword and chose $\text{RF} = 0.9$ for the case in which the title contains the keyword and $\text{RF} = 0.3$ for the case in which the title does not contain the keyword. Since the precision is more important than the recall in this challenge, the RF value pair that creates a very strict decision boundary was chosen. In the classification stage, different classifiers were trained and used on the basis of the relevance rule.

C. Evaluation

We can simply evaluate the result using MeSH-keyword data as positive data and MeSH-non-keyword data as negative data. However, a MeSH-non-keyword does not only consist of keyword-unrelated data, as shown in Figure 3. Therefore, we additionally used keyword-independent data, which only consist of keyword-unrelated data. The final classification score is calculated by average of two measurements:

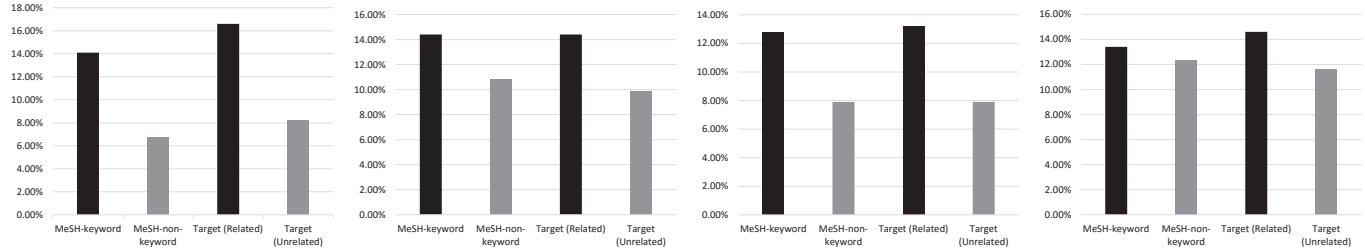
$$s_1 = pr(\text{MeSH-keyword}, \text{MeSH-non-keyword})$$

$$s_2 = pr(\text{MeSH-keyword}, \text{keyword-independent})$$

$$\text{C-score} = 2 \times \frac{s_1 \times s_2}{s_1 + s_2}$$

pr is the function that takes positive data and negative data as arguments and returns a precision. MeSH-keyword data were used as the positive data and MeSH-non-keyword and keyword-independent data were used as the negative data. Since the purpose of this study is to find additional data, the precision is much more important than the recall. Therefore, these equations are only composed of the precision.

The equation for s_1 refers to MeSH-non-keyword data as negative data. Since MeSH-keyword data and MeSH-non-keyword data are classified by the NLM, it is fairly reliable but not perfect. As described in Section I-A, there are valuable data in the MeSH-non-keyword data. Therefore, $s_1 \in [0, 1)$. On the other hand, keyword-independent data are completely independent of MeSH-keyword data. Thus, $s_2 \in [0, 1]$. If only one of either s_1 or s_2 is low, the overall accuracy is close to the low value. For this reason, the harmonic mean is adopted for the C-score equation, as in the F1-score [13]. The problem is that the classifier extracts keyword-related data from the MeSH-non-keyword data, but the question of how to evaluate this remains. It is very difficult to evaluate owing to absence of test data. Therefore, we used an indirect evaluation method—a frequency-based approach. This approach measures a fraction of the literature sources that contain keyword-related genes. The related genes are manually collected from various sources: lung-cancer-related genes from KEGG [14], [15], Genetics Home Reference (GHR) [16], and the study of A. El-telbany et al. [17]; prostate-cancer-related genes from PGDB [18], KEGG, and DDPC [19]; breast-cancer-related genes from GHR, Cancer Research UK [20], BreastCancer [21], and the study of M. De Jong et al. [22]; and Alzheimer's-disease-related genes from MalaCards [23], the study of M. Cruts et al. [24], and OMIM [25].



(a) Lung cancer (b) Prostate cancer (c) Breast cancer (d) Alzheimer disease

Fig. 5: Frequency-based approach

TABLE I: Data information. Each keyword indicates lung cancer, prostate cancer, breast cancer, and Alzheimer's disease.

Data type	Keywords			
	Lung	Prostate	Breast	Alzheimer's
# of whole keyword data points	61,041	36,360	80,550	21,713
# of MeSH-keyword data points	31,512	21,376	44,404	15,850
# of MeSH-non-keyword data points	14,389	6,034	15,212	2,569
# of non-MeSH data points	15,140	8,950	20,934	3,294
# of target data points	29,529	14,984	36,146	5,863
	Eye disease			
# of keyword-independent data points	31,850			

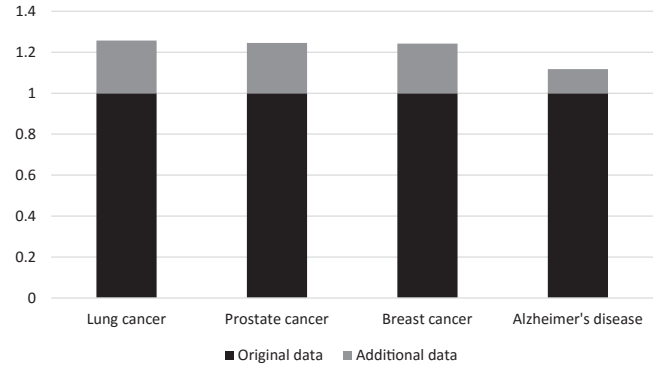


Fig. 6: Additional data ratio based on keyword data

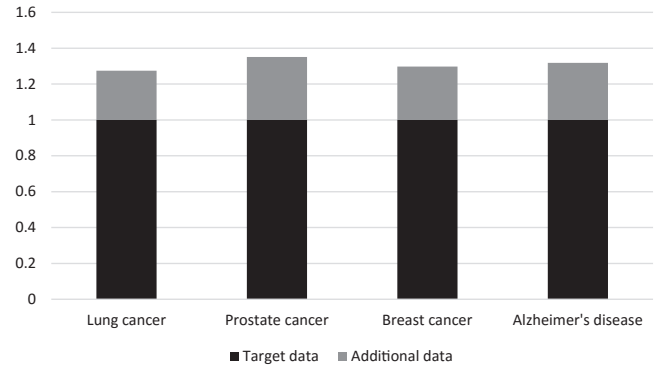


Fig. 7: Additional data ratio based on target data

IV. Results and Discussion

We implemented our method using Python 3.4 with several open source libraries. BeautifulSoup [26] was used for XML parsing, and the one-class SVM was implemented by Scikit-learn [27]. In the preprocessing step, NLTK [28] was used for word splitting and stemming, and Gensim [29] was used to implement the paragraph vector method.

A. Data Sources

In this experiment, we applied our method to lung cancer, prostate cancer, breast cancer, and Alzheimer's disease literature data. Moreover, eye disease data were used for the keyword-independent data. In order to ensure the independence of the eye disease data, literature sources that have the MeSH term or text related to the keywords were excluded through the PubMed search option.

All keyword data are obtained from PubMed and published within last five years so that the most recent data are used, except for the eye disease data, which was within the past three years. Since eye disease is a larger category than the other keywords, there is a large amount of eye disease data. Thus, only literature sources published within the past three years were used for the keyword-independent data. The details of the data size information are summarized in Table I.

B. Comparison of the Preprocessing Methods

Table II summarizes the results of a comparison of the two preprocessing methods. OCSVM is the one-class SVM, and RR is the relevance rule. The results demonstrate that the TF-IDF method is better than the paragraph vector in this study. Therefore, we used TF-IDF vectorization in the preprocessing step.

Furthermore, Table II indicates the high accuracy of our method. The method classified MeSH-keyword data and keyword-independent data with a precision of 100%, represented as s_2. The scores, represented as s_1, for the MeSH-keyword data and MeSH-non-keyword data are not 100%, but this is an expected result. As mentioned above, since a MeSH-non-keyword has valuable data, the scores cannot reach 100%.

C. Frequency-based Approach

Figure 5 summarizes the experimental results of an evaluation of the frequency-based approach. Each percentage rela-

TABLE II: C-scores of two preprocessing methods

Model	Measure	Keywords			
		Lung cancer	Prostate cancer	Breast cancer	Alzheimer's disease
OCSVM + RR + TF-IDF	s_1	0.946	0.957	0.902	0.980
	s_2	1.00	1.00	1.00	1.00
	C-score	**0.972**	**0.978**	**0.948**	**0.990**
OCSVM + RR + Paragraph Vector	s_1	0.781	0.882	0.774	0.934
	s_2	0.897	0.926	0.911	0.940
	C-score	0.835	0.903	0.837	0.937

TABLE III: Frequency-based approach

	Keywords			
	Lung cancer	Prostate cancer	Breast cancer	Alzheimer's disease
MeSH-keyword	14.1%	14.4%	12.8%	13.4%
MeSH-non-keyword	6.7%	10.8%	7.9%	12.3%
Non-MeSH (Related)	17.6%	14.3%	13.7%	14.6%
Non-MeSH (Unrelated)	10.9%	8.9%	8.3%	10.6%
MeSH-non-keyword (Related)	8.5%	15.3%	10.0%	14.4%
MeSH-non-keyword (Unrelated)	6.6%	10.5%	7.7%	12.2%
Target (Related)	16.6%	14.4%	13.2%	14.6%
Target (Unrelated)	8.2%	9.9%	7.9%	11.6%

TABLE IV: Extracted source comparison

Data type	Measure	Keywords			
		Lung	Prostate	Breast	Alzheimer's
Non-MeSH	count	7158	4967	11011	1736
	extraction ratio	47.3%	55.5%	52.6%	52.7%
	data-type ratio	88.4%	94.5%	91.3%	92.8%
MeSH-non-keyword	count	939	290	1050	134
	extraction ratio	6.5%	4.8%	6.9%	5.2%
	data-type ratio	11.6%	5.5%	8.7%	7.2%

tively represents how each data set is related to the keyword. Every result demonstrates that the additional data extracted by our method have as much relevance as MeSH-keyword data, and the data classified into keyword-unrelated data by our method have little relevance and are as low as the MeSH-non-keyword data. In the case of Alzheimer's disease, there is little difference between the MeSH-keyword data and MeSH-non-keyword data relative to the other experiments. This may result from the small amount of target data. The amount of target data for Alzheimer's disease is up to six times and at least 2.5 times less than those for the other data. Nevertheless, the difference between the keyword-related data and the keyword-unrelated data increased for the target data. This means that the classification works well. The specific numerical values for the experiments are listed in Table III.

D. Additional Data Analysis

As shown in Figure 6, our method additionally gathered up to 26% more data with a high quality. The ratio of Alzheimer's disease data is less than that for the other data because of small quantity of the target data, as shown in Figure 7. The figure shows the additional data ratio based on the target data. Since the amount of target data for Alzheimer's disease is small even though the additional amount of extracted data is also small, the ratio is not small.

Table IV summarizes the sources of the extracted additional data. In the "Measure" column, the "count" indicates the number of extracted literature sources, the "extraction ratio" refers to the ratio of the extracted data to the each data of types, and the "data-type ratio" is the ratio of the each data of types to the entire target data:

$$\text{extraction ratio} = \frac{\text{extracted data}}{\text{data of types}}$$

$$\text{data-type ratio} = \frac{\text{data of types}}{\text{target data}}$$

For example, a extraction ratio of 47.3% for lung cancer means that 47.3% of non-MeSH data is extracted as keyword-related data, and a data-type ratio of 88.4% means that 88.4% of the extracted data is non-MeSH data.

From the results, most of the extracted data are non-MeSH data, which consists of literature sources that have not yet been allocated a MeSH term. That is, most of the extracted data are the latest data that were not available in the typical MeSH term search method.

V. Conclusions and Future Work

The gathering of richer data than a MeSH term search is a challenge. A MeSH term search is the traditional and typical method of gathering biomedical text data, but it cannot obtain the latest data that are registered within 10 months on average. In addition, there are valuable literature source that cannot be gathered by a MeSH term search, even if they do not contain the latest data. In order to resolve these issues, we proposed a novel method that gathers rich data from PubMed. The experimental results demonstrate that our method can extract additional high-quality data from target data that were not

previously available for biomedical text mining. The extracted data were up to 26% of the keyword data and 35% of the target data including the latest data.

In this paper, we used a one-class SVM, but there are many other methods for one-class classification. Thus, we will examine other one-class classification methods and develop a novel one-class classification method that is suitable for the challenge discussed in this paper. Furthermore, the utilization of web data has recently attracted attention [30], [31]. Web data contain a great deal of information but are unreliable. Therefore, we will apply our method to web data to extract reliable data in a future study. Our final goal is to create an application that receives any PubMed keyword from the user, learns the appropriate classifier, gathers rich data, and provides data to the researchers.

ACKNOWLEDGEMENTS

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) (NRF-2015R1A2A1A05001845).

REFERENCES

[1] National library of medicine. [Online]. Available: https://www.nlm.nih.gov/

[2] R. W. Rodriguez, "Delay in indexing articles published in major pharmacy practice journals." *American Journal of Health-System Pharmacy*, vol. 71, no. 4, 2014.

[3] J. Cha, "A method for obtaining rich data from pubmed using svm," in *Proceedings of the 31th Annual ACM Symposium on Applied Computing*, 2016.

[4] G. Salton and M. MacGill, *Introduction to modern information retrieval*. McGraw-Hill Education, 1983.

[5] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[6] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 2014, pp. 1188–1196.

[7] D. M. J. Tax, *One-class classification*. TU Delft, Delft University of Technology, 2001.

[8] B. Schölkopf, R. C. Williamson, A. J. Smola, J. Shawe-Taylor, and J. C. Platt, "Support vector method for novelty detection," in *Advances in Neural Information Processing Systems*, 2000, pp. 582–588.

[9] C. Silva and B. Ribeiro, "The importance of stop word removal on recall values in text categorization," in *Neural Networks, 2003. Proceedings of the International Joint Conference on*, vol. 3. IEEE, 2003, pp. 1661–1666.

[10] R. Cooley, "Classification of news stories using support vector machines," in *Proc. 16th International Joint Conference on Artificial Intelligence Text Mining Workshop*. Citeseer, 1999.

[11] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.

[12] J. T. Behrens and C.-H. Yu, "Exploratory data analysis," *Handbook of psychology*, 2003.

[13] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation," in *AI 2006: Advances in Artificial Intelligence*. Springer, 2006, pp. 1015–1021.

[14] M. Kanehisa and S. Goto, "Kegg: kyoto encyclopedia of genes and genomes," *Nucleic acids research*, vol. 28, no. 1, pp. 27–30, 2000.

[15] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, "Kegg as a reference resource for gene and protein annotation," *Nucleic acids research*, vol. 44, no. D1, pp. D457–D462, 2016.

[16] Genetics home reference. [Online]. Available: https://ghr.nlm.nih.gov/

[17] A. El-Telbany and P. C. Ma, "Cancer genes in lung cancer racial disparities: are there any?" *Genes & cancer*, vol. 3, no. 7-8, pp. 467–480, 2012.

[18] L.-C. Li, H. Zhao, H. Shiina, C. J. Kane, and R. Dahiya, "Pgdb: a curated and integrated database of genes related to the prostate," *Nucleic Acids Research*, vol. 31, no. 1, pp. 291–293, 2003.

[19] M. Maqungo, M. Kaur, S. K. Kwofie, A. Radovanovic, U. Schaefer, S. Schmeier, E. Oppon, A. Christoffels, and V. B. Bajic, "Ddpc: Dragon database of genes associated with prostate cancer," *Nucleic acids research*, vol. 39, no. suppl 1, pp. D980–D985, 2011.

[20] Cancer research uk. [Online]. Available: http://www.cancerresearchuk.org/

[21] Breastcancer. [Online]. Available: http://www.breastcancer.org/

[22] M. De Jong, I. Nolte, G. Te Meerman, W. Van der Graaf, J. Oosterwijk, J. Kleibeuker, M. Schaapveld, and E. De Vries, "Genes other than brca1 and brca2 involved in breast cancer susceptibility," *Journal of medical genetics*, vol. 39, no. 4, pp. 225–242, 2002.

[23] N. Rappaport, N. Nativ, G. Stelzer, M. Twik, Y. Guan-Golan, T. I. Stein, I. Bahir, F. Belinky, C. P. Morrey, M. Safran *et al.*, "Malacards: an integrated compendium for diseases and their annotation," *Database*, vol. 2013, p. bat018, 2013.

[24] M. Cruts, J. Theuns, and C. Van Broeckhoven, "Locus-specific mutation databases for neurodegenerative brain diseases," *Human mutation*, vol. 33, no. 9, pp. 1340–1344, 2012.

[25] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick, "Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders," *Nucleic acids research*, vol. 33, no. suppl 1, pp. D514–D517, 2005.

[26] L. Richardson. Beautifulsoup. [Online]. Available: https://www.crummy.com/software/BeautifulSoup/

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *The Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[28] S. Bird, "Nltk: the natural language toolkit," in *Proceedings of the COLING/ACL on Interactive presentation sessions*. Association for Computational Linguistics, 2006, pp. 69–72.

[29] P. Sojka, "Software framework for topic modelling with large corpora," in *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Citeseer, 2010.

[30] J. Kim, H. Kim, Y. Yoon, and S. Park, "Lgscore: A method to identify disease-related genes using biological literature and google data," *Journal of biomedical informatics*, vol. 54, pp. 270–282, 2015.

[31] H. Kim and S. Park, "Discovering disease-associated drugs using web crawl data," in *Proceedings of the 31th Annual ACM Symposium on Applied Computing*, 2016.