# A method for obtaining rich data from PubMed using SVM

Junbum Cha, Jeongwoo Kim, Yunku Yeu and Sanghyun Park*
khanrc@yonsei.ac.kr, {jwkim2013, yyk, sanghyun}@cs.yonsei.ac.kr
Yonsei University
Seoul, Korea

## ABSTRACT
As text mining advances rapidly in the biomedical field, the importance of text data is increasing. Most text data is obtained through a Medical Subjects Headings (MeSH) term search; in this process, a large amount of valuable data is missed because the data is not indexed yet with MeSH terms. In this paper, we propose a method for obtaining additional text data in addition to that obtained using a conventional MeSH term search.

In order to obtain additional data, we used the Support Vector Machine (SVM) as the data mining method for classifying documents to related or unrelated. We evaluated the results using a frequency-based text mining approach measuring the quality of data in study of lung cancer. This was confirmed that the data extracted using our method provided as much valuable information as searching using MeSH terms. Further, we found that the amount of information found was increased by 40% using additional extracted data.

## CCS Concepts
• **Applied computing ~ Bioinformatics** • *Applied computing ~ Document analysis* • *Computing methodologies ~ Supervised learning by classification*

## Keywords
Bioinformatics, Text Mining, Document Classification.

## 1. INTRODUCTION
Text mining is conducted to identify valuable knowledge by analyzing unstructured text and then presenting it as refined results. This concept was first introduced in the 1980s and developed rapidly in the 1990s. After the success of the Human Genome Project in the 1990s, many high-throughput biological data generating technologies such as Next-Generation Sequencing (NGS) have been developed. Consequently, an amount of biological data and a number of related biomedical text mining studies have grown rapidly.

With the growth of biomedical text mining, the importance of text data has increased, as it is the main resource used for text mining. In general, text data is obtained from the Medical Literature Analysis and Retrieval System Online (MEDLINE) database of PubMed using a MeSH term search. However, this method may miss valuable documents.

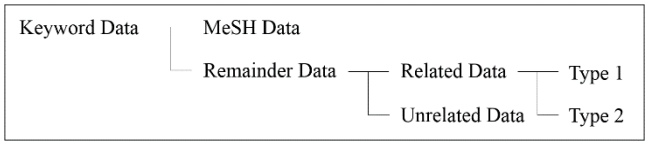


**Figure 1. Categories of PubMed search result**

The keyword search results of the PubMed are classified and defined, as shown in Figure 1. The keyword data refers to keyword search result in PubMed. The MeSH data indicates a MeSH term search result and remainder data indicates the keyword data except the MeSH data. The remainder data is also divided into related data and unrelated data. Related data refers to data related to the keyword and unrelated data refers to data unrelated to the keyword. In addition to the MeSH data commonly used in biomedical text data, the related data contains valuable information relevant to the keyword. Furthermore, there are documents that was not yet indexed by the MeSH. Document was not automatically assigned MeSH terms; rather, this process was carried out manually by the National Library of Medicine (NLM). Therefore, the assigning operation requires some time, and the recent literatures that has enough valuable information are missed in MeSH term search. In order to solve this problem, we propose a method for extracting additional data from the remainder data that is not conventionally used in text mining.

## 2. METHODS
### 2.1 Data sources
A data source is one of the most important prerequisites for performing data mining. Data used in this study included documents containing valuable information about the keyword and documents that did not contain keywords for a given topic. The documents containing valuable information belonged to the positive data as an object that we are looking for. In contrast, documents that are not containing valuable information belonged to the negative data as the object that subjected to filtering. Learning of the classifier requires both types of data. We obtained these data from the upcoming three PubMed search results.

***(A) MeSH Data.*** MeSH is a subject heading determined by NLM. The appropriate 10–15 MeSH terms are given showing the

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


*SAC 2016*, April 04-08, 2016, Pisa, Italy
 ISBN 978-1-4503-3739-7/16/04…$15.00
DOI: http://dx.doi.org/10.1145/2851613.2851866

contents for each document. The data obtained in the MeSH term search is reliable for positive data, as it is verified by NLM.

***(B) Non-target Data.*** Some of the rest of the keyword data. It is suitable for negative data as it is not related to the keyword. Since the amount of the rest of the keyword data is too larger than other data (A) or (C), we used only a portion of the data.

***(C) Remainder Data.*** This data may include related documents and is obtained by subtracting the MeSH data from the whole keyword data. The aim of this study was to classify this remainder data.

## 2.2 Experimental method

***Select Keyword & Data Download.*** We established the keywords of positive data and negative data, searched for these keywords in PubMed, and downloaded the data. When we specified a keyword, we established that the positive data and negative data were unrelated, taking into account the MeSH hierarchy and characteristics of the disease. PubMed provided search result data represented in an XML file.

***Stage (1): Parsing.*** The XML file provided in PubMed contained a large amount of data that was non-useful in this study. Therefore, we refined and extracted the PMID, title, abstract.

***Making Test Set.*** In order to evaluate the classifier, a test set is required. To obtain the test set, we manually read part of the remainder data and then determined whether the document was related to the keyword to form the test set.

***Stage (2): Preprocessing.*** Vectorization of the text data is required for text mining. In this study, we represented a document as a vector. Therefore, we first performed stop word removal and low frequency word removal to delete meaningless words. Next, we integrated inflected or derived words into word stems using Porter stemmer [6] and then transformed this refined text to a vector of bag-of-words form. We used TF-IDF analysis to identify key words in the document.

***Stage (3): Classification.*** We learned text mining classification models using documents represented as vectors. We examined a total of three algorithms, including SVM, random forest, and Naïve Bayes. The results showed that the input documents were classified into related documents or unrelated documents. We evaluated the performance of the classifiers using the test set described above, and finally selected an algorithm.

***Stage (4): Evaluation.*** We evaluated the performance of the selected classifier. Since the test set prepared manually is less reliable because of the small quantity of samples, it is used only to select the classification algorithm. Thus, we applied other approaches for evaluation.

## 2.3 Evaluation Method

As described above, the reliability of test data prepared manually is limited, and thus a new evaluation approach is required. In this study, the results were evaluated indirectly using a frequency-based text mining approach. This approach counts the appearance frequency of the genes and chooses the top N genes of appearance counting, ranking the genes related to the search keyword. We evaluated the data by calculating the genes actually related to the keyword of the extracted genes. We obtained human gene names from HGNC [3], lung cancer related genes from KEGG [2] and El-Telbany's study [1].

### *2.3.1 MeSH vs. Related vs. Unrelated*

We compared the MeSH data used as input data in the general biomedical text mining using the related and unrelated data extracted by the classifier. When the frequency-based text mining approach was applied for each data, we conducted repeated tests, increasing N from 10 to 100 to obtain reliable results. We evaluated the accuracy of the classifier in this experiment.

### *2.3.2 MeSH vs. MeSH+Related*

Additionally, we compared the MeSH data with the MeSH + related data. First, in the same manner as described in 2.3.1, we examined whether our method could extract only valuable data without noise data from the remainder data. In addition, we compared the amount of information by counting the total number of different genes for each data.

# 3. RESULTS AND DISCUSSION

## 3.1 Data Descriptions

**Table 1. Final collected data information**

| **Keyword** | **Lung cancer** |
|---|---|
| *# of Keyword data* | 56,833 |
| *# of MeSH data* | 29,992 |
| *# of Non-target data* | 15,295 |
| *# of Remainder data* | 26,841 |

This study is generalized for any keyword, but a specific keyword was required for the experiment, so we used "lung cancer" as keyword. Since both diseases are genetic diseases associated with specific genes, we selected "occupational disease", which is not related to the genes, as negative data (non-target data). The search result did not overlap with lung cancer data in the search option.

To use the most recent data, we used documents published in the last 5 years for lung cancer. Since the number of documents for occupational disease was less than those available for lung cancer, we used documents published in the last 10 years.

## 3.2 Evaluation

**Table 2. Experiment results**

| **Keyword** | **Lung cancer** |
|---|---|
| *# of Remainder* | 26,841 |
| *Related ratio* | 69.9% |
| *Unrelated ratio* | 30.1% |
| *Related/MeSH ratio* | 62.6% |

We extracted related data, which was 70% of remainder data and 63% of MeSH data (Table 2). This means that 70% of the data that was not used originally contained valuable information and the amount of the valuable data was greater than half of the data originally used. Furthermore, because much of the originally discarded data consisted of recent data, related data extracted using our classifier was more important than the above value, as described in the introduction.

### 3.2.1 MeSH vs. Related vs. Unrelated

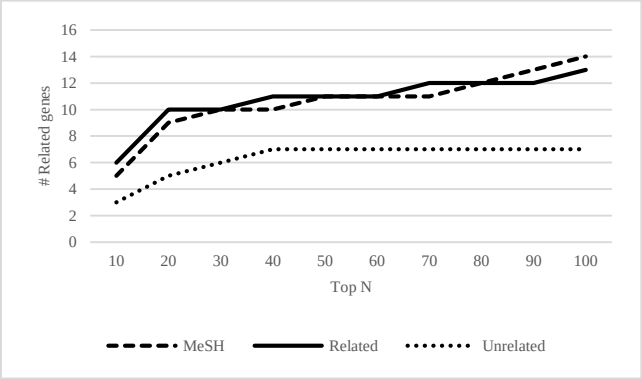


**Figure 2. MeSH, related, and unrelated data comparison in lung cancer**

Figures 2 shows the trend in the number of extracted genes associated with the keyword, with increasing N from 10 to 100. MeSH and related showed similar trends, while only a small number of genes were extracted in unrelated data. Thus, related data extracted using our method was similarly valuable to MeSH data and showed a significant difference compared to unrelated data.

### 3.2.2 MeSH vs. MeSH+Related

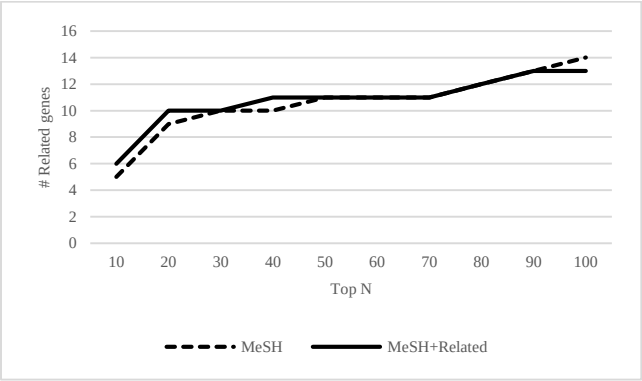


**Figure 3. MeSH and MeSH+related data comparison in lung cancer**

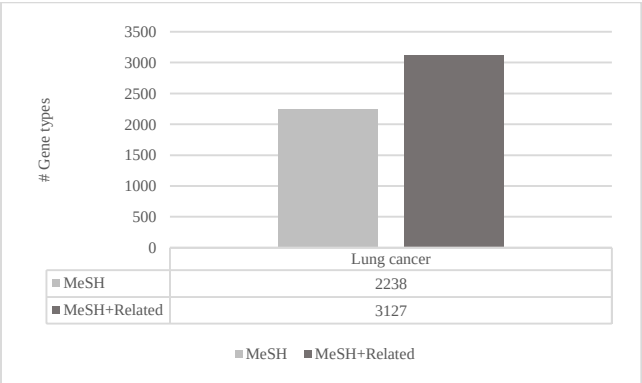


**Figure 4. Number of different genes for each data. The number of different genes increased to 39.7% for lung cancer**

MeSH+related data produced equal or better results than MeSH data as shown in Figures 3. The classifier successfully filtered the noise data that did not contain valuable information from the remainder data. This means that quantity of data was increased while quality of data was maintained. Moreover, the increase of data brought increase of information as shown in Figure 4. We found that the amount of information increased by 39.7% when we added related data to MeSH data. Therefore, we can utilize the more rich information and the quality of the data can be maintained by using further related data.

## 4. CONCLUSIONS AND FURTHER WORK

In this study, we proposed a method for obtaining additional PubMed data using remainder data, which is not typically used in biomedical text mining. We obtained additional data up to 63% of the original data by using our method. This method not only increases the amount of data, but also enables the use of recent data within six months of publication, which is not identified using traditional mesh term search. We expect that this method can be used to obtain additional data in future studies.

We developed a generalized method for extracting additional data for any keyword. Future studies will involve examining additional and different keywords. We will also attempt to improve the performance of the classification method for extracting data. As described in the Discussion, because it is difficult to identify appropriate negative data, we will consider one-class SVM [5], which does not require negative data. In addition, the method used to represent the data can be changed to improve performance. Recently, Le and Mikolov [4] proposed a method known as doc2vec for embedding of a document to a low-dimensional vector using deep learning. Performance can be improved by representing a document in a more effective manner using this method.

## 5. ACKNOWLEDGEMENTS

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

## 6. REFERENCES

[1] El-Telbany, A. and Ma, P.C. Cancer Genes in Lung Cancer. *Genes & cancer*, *3*, 7-8 (2012), 467–80.

[2] Goto, S. and Kanehisa, M. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, *28*, 1 (2000), 27–30.

[3] HUGO Gene Nomenclature Committee. HGNC Database. 2011. *http://www.genenames.org/*.

[4] Le, Q. and Mikolov, T. Distributed Representations of Sentences and Documents. *Proceedings of The 31st International Conference on Machine Learning*, (2014), 1188–1196.

[5] Manevitz, L.M. and Yousef, M. One-class svms for document classification. *The Journal of Machine Learning Research*, *2*, (2002), 139–154.

[6] Porter, M.F. An algorithm for suffix stripping. *Program: electronic library and information systems*, *14*, 3 (1980), 130–137.