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.

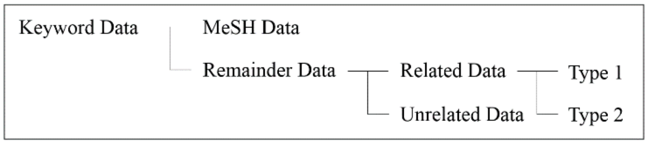

Figure 1. Categories of PubMed search result

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
© 2016 ACM. ISBN 978-1-4503-3739-7/16/04…$15.00
DOI: http://dx.doi.org/10.1145/2851613.2851866