Discovering Disease-associated Drugs Using Web Crawl Data

Hyunjin Kim, Sanghyun Park*
Department of Computer Science, Yonsei University
50 Yonsei-ro, Seodaemun-gu, Seoul, South Korea
{chriskim, sanghyun}@cs.yonsei.ac.kr

ABSTRACT
The purpose of research on biomedical literature-based discovery is to bring out new knowledge from the existing biomedical information. Beginning with Dr. Swanson's ABC model, many studies extended or applied the ABC model to find new associations between biomedical entities. While the methods applied to data have advanced, in most cases biomedical literature has been used for the text data. Assuming that web crawl data is helpful in studying literature-based discovery as well as biomedical literature which is the existing but rather limited data source, we discovered new disease-drug associations using web crawl data in addition to biomedical literature. We also analyzed how helpful the additional use of web crawl data is for biomedical literature mining. Literature-based discovery using web crawl data has its significance as a pioneering work utilizing new data.

Categories and Subject Descriptors
J.3 [**Life and medical sciences**]: Biology and Genetics

General Terms
Experimentation, Measurement, Verification

Keywords
literature-based discovery; biomedical text mining; disease-associated genes; web crawl data; ABC model;

1. INTRODUCTION
Research on literature-based discovery has advanced a lot since Dr. Don R. Swanson's study on Raynaud's syndrome in 1986 [1]. Until recently text mining has been used to extract unpublished information from literature published in the field of biomedical research. The core concept is Dr. Swanson's ABC model. The ABC model is a rather simple concept that if A and B share a connection and B and C share also a connection, then A and C may have an implicit connection to each other. Since it was turned out that by using this concept new information can be obtained from literature mining without real biomedical tests [2], many researchers have joined literature-based discovery studies. Since then, most of subsequent methods either extended or applied the ABC model.

The final aim of the ABC model is to find new connections between biomedical entities using text data. Three steps are necessary to do this. First, names of bio-entities and text data should be obtained. Then, associations get extracted from the obtained text data using the names of the bio-entities and lastly, new association that has never existed should be found using the extracted associations. Names of bio-entities and text data are needed to extract a biomedical association from the text data. Bio-entities can be disease name, gene name, protein name, drug name, symptom name and miRNA name. And biomedical research papers are mostly used for text data. If text data and a name list of bio-entities are secured, it can be found from the text data whether the bio-entities have connections between them. The most accurate way is to read through and to check the connections in person but it is not feasible because there is too much biomedical literature which can be used for text data. Therefore, it is necessary to find a method of detecting whether a text describes a certain bio-entity association: if names of two bio-entities come up together in one literature or names of two bio-entities appear at the same time in one sentence of one literature, it is generally supposed that the two bio-entities have an association between them.

As mentioned earlier, most of studies on literature-based discovery have been on methods to extract new associations which have never existed. First of all, there is a conventional method, Dr. Swanson's ABC model. The ABC model searches for the associations between A and C which are not on the existing list when associations exist between A and B and between B and C. Petric [3] proposed an algorithm which advanced the ABC model: to use rare terms which come up infrequently in all the literature as B for a middle stage. This method is based on the assumption that if the rare terms appear together with A and C both, there is a high possibility of association between A and C. In case of Li's method [4], a set of MeSH (Medical Subject Headings: the National Library of Medicine's controlled vocabulary thesaurus, used for indexing articles as citing the content of the literature with 10-15 terms) terms are used to create a term document matrix. Mutual information value can be calculated based on the matrix, then the bigger the value, the more important the two terms become. The next step is to apply ABC model and to choose term B having the largest mutual information among associations between A and B. Then, to choose term C in order of mutual information among term C related to the chosen term B. Li's method has a significant point in creating term document matrix and using the mutual information based on the matrix. Tsuruoka [5] also proposed a method of expanded ABC model that utilizes ranking approach considering strength and reliability of direct association and indirect association between term A and term C to extract new associations that have never existed.

In literature-based discovery field, studies on discovering new biomedical information from biomedical literature have been broadly researched. However, it is hard to find approaches which use other text data sources. Finite and fixed data sources make finite and fixed results. If we use new data sources, we have a chance to identify new and undiscovered knowledge that have not been discovered from biomedical literature. Surely, there exist a few approaches which do not use other text data sources but utilize other data sources such as

* Corresponding author. Tel.: +82 2 2123 5714

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/10.1145/2851613.2851615