Chapter 17

Finding Explicit and Implicit Knowledge:
Biomedical Text Data Mining

Kazuhiro Seki
Kobe University, Japan

Javed Mostafa
University of North Carolina at Chapel Hill, USA

Kuniaki Uehara
Kobe University, Japan

ABSTRACT

This chapter discusses two different types of text data mining focusing on the biomedical literature. One deals with explicit information or facts written in articles, and the other targets implicit information or hypotheses inferred from explicit information. A major difference between the two is that the former is bound to the contents within the literature, whereas the latter goes beyond existing knowledge and generates potential scientific hypotheses. As concrete examples applied to real-world problems, this chapter looks at two applications of text data mining: gene functional annotation and genetic association discovery, both considered to have significant practical importance.

INTRODUCTION

With the help of high throughput gene analysis and increasing computing power, the amount of data produced in biomedicine is rapidly growing, making it one of the most attractive domains for the exploration of data mining techniques. Among various types of data, such as DNA sequences, medical images, and clinical records, that could be mined for interesting knowledge and discoveries, this chapter focuses on textual resources, specifically, the biomedical literature, and discusses two different types of **text data mining** (TDM).

Medline, the world largest bibliographic database in life science, currently contains bibliographies for over 17 million articles, and 2000−4000 new records are added each day. Both the volume and the pace of the publications exceed the capacity of any individuals, which calls for the aid of intelligent information processing techniques, i.e., TDM. There are roughly two types of TDM with respect to the

DOI: 10.4018/978-1-61520-757-2.ch017

goals they pursue. One focuses on the information explicitly stated in text and attempts to extract and organize it for better information access. This type of TDM has been extensively studied in recent years and includes, for example, information retrieval (IR) (Hersh, 2004), information extraction (IE) (Hobbs, 2002), and automatic summarization (Reeve et al., 2007). The other type of TDM is not bound to explicit or existing information but targets implicit information, or heretofore unknown knowledge (Hearst, 1999), that could be revealed by synthesizing fragments of information extracted from a large volume of textual data. This type of TDM is called **hypothesis discovery** and was initiated by Swanson (1986b) in the 1980's. We will see more concrete examples for each type of TDM throughout this chapter.

BACKGROUND

There have been numerous efforts in biomedical TDM dealing with explicit information (Ananiadou et al., 2006; Cohen and Hersh, 2005; Shatkay, 2005). One of the earliest and most successful attempts in this type of TDM is named-entity (NE) recognition, the first step to IE, mainly targeting genes and proteins (Fukuda et al., 1998; Seki and Mostafa, 2005b; Hsu et al., 2008). NE recognition in biomedicine is largely different from other domains tackled earlier, such as newspaper articles, in a sense that biomedical NEs have surprisingly many synonyms and writing variants. This issue is essential in dealing with biomedical text and heavily affects the performance of TDM systems as we will see in the next section.

For biomedical IR, the Genomics Track (Hersh and Bhuptiraju, 2003; Hersh et al., 2004, 2005, 2006, 2007) at the Text REtrieval Conference (TREC) was undoubtedly the most significant strides made in the history. The track was a five-year project held between 2003 and 2007 and tackled various types of IR tasks, including Ad Hoc retrieval and passage retrieval, as well as other IE oriented tasks. While the track was successful, having attracted the largest number of research groups world-wide among the TREC tracks, there is still much room for improvement, especially for the passage retrieval challenged in 2007 which, given a user query, required to return passages containing relevant named entities of a certain type within the context of supporting text.

Another task from the Genomics Track that is pertinent to TDM (in broader sense) is Gene Ontology (GO) annotation via automatic text analysis. The following provides some background of the task since this will be one of the main focuses in this chapter.

After the completion of the Human Genome Project, the major activities in molecular biology have shifted to understanding the precise functions of individual genes. The consequence in part is the increasing, large number of publications that one cannot digest. To provide direct access to the information regarding gene functions buried in natural language text, three model organism databases created controlled vocabularies, namely, the Gene Ontology (GO), to annotate genes with their functions. GO is structured as directed acyclic graph (DAG) under three top level nodes, Molecular Function (MF), Cellular Component (CC), and Biological Process (BP), as shown in Figure 1.

While GO annotation allows uniform queries across different databases, manually annotating GO terms is labor-intensive and costly due to the voluminous and specialized contents. This resulted in a demand to automate GO annotation, which was the main objective of the GO annotation task at the TREC Genomics Track and, independently, the BioCreative challenge (Blaschke et al., 2005), another important workshop targeting biomedical TDM. The next section will describe in detail how GO annotation can be effectively done in a standard categorization framework coupled with gene-centered document representation.

For biomedical hypothesis discovery, also known as **literature-based discovery** or LBD,