Chapter 2
Text Mining on Big and Complex Biomedical Literature

Boya Xie
East Carolina University, USA

Qin Ding
East Carolina University, USA

Di Wu
Drexel University, USA

ABSTRACT

Driven by the rapidly advancing techniques and increasing interests in biology and medicine, about 2,000 to 4,000 references are added daily to MEDLINE, the US national biomedical bibliographic database. Even for a specific research topic, extracting useful and comprehensive information out of the huge literature data pool is challenging. Text mining techniques become extremely useful when dealing with the abundant biomedical information and they have been applied to various areas in the realm of biomedical research. Instead of providing a brief overview of all text mining techniques and every major biomedical text mining application, this chapter explores in-depth the microRNA profiling area and related text mining tools. As an illustrative example, one rule-based text mining system developed by the authors is discussed in detail. This chapter also includes the discussion of the challenges and potential research areas in biomedical text mining.

INTRODUCTION

Text mining is a process that automatically derives quality information from text, also known as text analysis or data analytics. It was first introduced with labor-intensive manual approaches in the mid-1980s. Later, it has evolved to use intelligent approaches and algorithms to derive useful and quality information from text. Techniques involved in text mining come from multiple disciplines, such as information retrieval, data mining, machine learning, natural language processing, etc. With the big, complex, and fast-changing text data, it is challenging yet rewarding to perform text mining to extract useful and reliable information, which will greatly benefit researchers in various fields.

DOI: 10.4018/978-1-4666-6611-5.ch002

Biomedicine has been one of these fields where text-mining is utilized extensively. Some of the text mining applications in biomedical research include gene ontology, signal transduction pathway and gene mining, yeast metabolites extracting, disease specific mutation-gene pair identification, protein interactions, and microRNA profiling. From the application perspective, text-mining could be categorized as the following five major classes: Gene-centric, protein-centric, microRNA-centric, disease related, and pathway mining. From the text analysis objective perspective, text-mining could be categorized as: name entity recognition (NER), association extraction, and event extraction. A brief introduction of all these categories is presented in the background section. Because microRNA (miRNA) is used as the example throughout this chapter, the background information of miRNA as well as a comprehensive review of miRNA-centric text-mining applications is also provided in the background section.

MiRNA is a small sequence of nucleotides that plays an important role in many biological processes. Abundant research has been carried out to study miRNA functionalities, emphasizing on its profile in diseases. This chapter takes mining for miRNA profile in human cancer as an example, walks through steps that construct a typical text-mining system: data preparation, information extraction, and result validation. Questions such as where and at what granularity to gather source text, how to clean data for miRNA specific literature are discussed in the data preparation section. Information extraction section covers miRNA, cancer, and expression term recognition, followed by the discussion of co-occurrence-based, rule-based and machine learning dependent approaches that discover miRNA-cancer relationships. Result validation and system evaluation methods are briefly described. This chapter concludes with a discussion of challenges and future research directions in miRNA-centric text mining. Potential biomedical areas that may deal with big data as well as possible approaches are presented at the end of this chapter.

BACKGROUND

The knowledge and techniques in biomedicine have been advanced drastically. Collaboration among computer science, physics, mathematics, and engineering has enabled biomedical researchers to explore solutions to many of the world's most concerned health problems. Biomedicine research topics span from molecules to macro environment. Text mining has been applied to many of these data intensive areas. The most studied areas in biomedicine text-mining are discussed as follows.

Categorize by Applied Area

Gene-centric mining is the most popular area where research is carried out to study what is gene and how it works. The double helix molecule deoxyribonucleic acid (DNA) brings instructions for development and functioning for all living organisms. The entire DNA does not carry information at each section, and regions that are informative are called genes. Understanding gene greatly helps to explain many biological processes. Gene-centric mining consists of gene name recognition (R. McDonald & Pereira, 2005; Sasaki, Tsuruoka, McNaught, & Ananiadou, 2008), gene relationship mining (Chen & Sharp, 2004; Y. Liu et al., 2005), gene-disease association mining (Al-Mubaid & Singh, 2010; Bauer-Mehren, Rautschka, Sanz, & Furlong, 2010; Ozgur, Vu, Erkan, & Radev, 2008), gene-drug relationship mining (Garten, Tatonetti, & Altman, 2010), and gene annotation (Aerts et al., 2008; Haeussler, Gerner, & Bergman, 2011). After the completion of the Human Genome Project in 2003, the genome research direction has shifted from identifying genes towards understanding the human genome functionalities in both health and disease. The Gene Ontology project (Ashburner et al., 2000) provides a unified representation of gene and gene product attributes that becomes a standard reference for many other systems annotating genes and gene products.