Text Mining in Bioinformatics: Research and Application

Yanliang Qi, Millennium Pharmaceutical, Inc., Cambridge, MA, USA

ABSTRACT

Biomedical literatures have been increased at the exponential rate. To find the useful and needed information from such a huge data set is a daunting task for users. Text mining is a powerful tool to solve this problem. In this paper, we surveyed on text mining in Bioinformatics with emphasis on applications of text mining for bioinformatics. In this paper, the main research directions of text mining in bioinformatics are accompanied with detailed examples. This paper suited the need for the state-of-the-art of the field of text mining in Bioinformatics because of the rapid development in both text mining and bioinformatics. Finally, the problems and future way are identified at last.

Keyword: Bioinformatics, Biomedical, Gene Annotation, Gene Expression, Micro Array Analysis, Named Entity Recognition, Systems Biology, Text Mining

INTRODUCTION

In recent years, there has been an exponential increase in the research of biological area. The biological studies have been transformed from an “information-poor” to an “information-overload” environment. For example, GENBANK release 122 (2/01) contains 11,720,120,326 bases in 10,896,781 sequences11,720,120,326. There is also wealth of online information. MEDLINE 2004 database contains over 12.5 million records, and the database is currently growing at the rate of 500,000 new citations each year (Cohen & Hersh, 2005). Figure 1 shows the exploding number of articles available from Medline over the past 60 years (Corlan, 2012).

So, it is obvious that problem faced by the biological researchers is how to effectively find out the useful and needed documents in such an information overload environment. Traditional manual retrieval method is impractical. Furthermore, online biological information exists in a combination of structured, semi-structured and unstructured forms (M.Ghanem et al., 2005). It is impossible to keep abreast of all developments. Computational methodologies increasingly become important in research (G. Black & P. Stephan, 2004). Text mining techniques, which involve the processes of information retrieval, information extraction and data mining, provide a means of solving this (Ananiadou et al., 2006).

The interactivities of computational methodologies and life science formed a new research area—Bioinformatics. Bioinformatics is where the information sciences meet the life sciences. Bioinformatics is the application of information technologies to biological structures and processes, and the information generated by

DOI: 10.4018/ijirr.2013040102
The goal of text mining in bioinformatics is to help researchers to identify needed information more efficiently, uncover relationships from the vast amount of information.

This paper is distinguished from other survey papers in two aspects:

1. It discusses the difference between text mining, information retrieval and information extraction. They look like similar but play the different role;
2. This paper gives users an up-to-date survey on this area. Both text mining and bioinformatics have developed rapidly in these two years. Researchers need an up-to-date survey in this area. Focused on this, we present this updated survey paper.

In this paper, we will talk about the role of text mining in bioinformatics. Firstly, we will elaborate the features of text mining and bioinformatics. Then, we will talk about the application of text mining in the bioinformatics area. The third part is the discussion of the research method in this area. Discussion about the problems and future way in this field will be the last part.

FEATURES OF TEXT MINING AND BIOINFORMATICS

Text mining (TM) is a technology that makes it possible to discover patterns and trends semi-automatically from huge collections of unstructured text. It is based on technologies such as Natural Language Processing (NLP), Information Retrieval (IR), Information Extraction (IE), and Data Mining (DM) (N. Uramoto et al., 2004).

From the search-centric point, the above several technologies, IR, IE and DM look very similar as Text mining. But actually, Text Mining (TM) is different from IR, IE and DM. The main difference is that whether there is novel produced in the process (Hearst, 1999).

Information retrieval (IR) is the science of searching for information in documents, searching for documents themselves, searching for metadata which describe documents, or
Survey of Clustering: Algorithms and Applications
www.igi-global.com/article/survey-of-clustering/100038?camid=4v1a