Chapter XLII
Text Mining in Bioinformatics:
Research and Application

Yanliang Qi
New Jersey Institute of Technology, USA

ABSTRACT

The biology literatures have been increased in an exponential growth in recent year. The researchers need an effective tool to help them find out the needed information in the databases. Text mining is a powerful tool to solve this problem. In this chapter, we talked about the features of text mining and bioinformatics, text mining applications, research methods in bioinformatics and problems and future path.

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 sequences. 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 (Aaron M. Cohen & William R. Hersh, 2005). Figure 1 shows the exploding number of articles available from Medline over the past 65 years (data retrieved from the SRS server at the European Bioinformatics Institute; www.ebi.ac.uk/). (Dietrich et al., 2005)

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 this application (Pharmabiz.com, 2002). 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.

In this chapter, 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 TM AND BIOINFORMATICS**

Text mining 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).

The 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 (M. 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 searching within databases, whether relational stand-alone databases or hyper-textually-networked databases such as the World Wide Web (Wikipedia.org, 1999). The process of IR is to find out the needed information.
8 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the product's webpage: www.igi-global.com/chapter/text-mining-bioinformatics/21755?camid=4v1


Related Content

Review of Data Mining Techniques and Parameters for Recommendation of Effective Adaptive E-Learning System
www.igi-global.com/chapter/review-of-data-mining-techniques-and-parameters-for-recommendation-of-effective-adaptive-e-learning-system/159492?camid=4v1a

Routing Attribute Data Mining Based on Rough Set Theory
www.igi-global.com/article/routing-attribute-data-mining-based/1769?camid=4v1a

Cluster-Based Input Selection for Transparent Fuzzy Modeling
www.igi-global.com/article/cluster-based-input-selection-transparent/1771?camid=4v1a

Cancer Pathway Network Analysis Using Cellular Automata
Kalyan Mahata and Anasua Sarkar (2015). *Improving Knowledge Discovery through the Integration of Data Mining Techniques* (pp. 145-159).
www.igi-global.com/chapter/cancer-pathway-network-analysis-using-cellular-automata/134536?camid=4v1a