Chapter 3
Mining Scientific and Technical Literature:
From Knowledge Extraction to Summarization

Junsheng Zhang
Institute of Scientific and Technical Information of China, China

Wen Zeng
Institute of Scientific and Technical Information of China, China

ABSTRACT

In this chapter, the authors study text mining technologies such as knowledge extraction and summarization on scientific and technical literature. First, they analyze the needs of scientific information services and intelligence analysis on massive scientific and technical literature. Second, terminology recognition and relation extraction are important tasks of knowledge extraction. Third, they study knowledge extraction based on terminology recognition and relation extraction. Fourth, based on terminology and relational network, they study the text summarization techniques and applications. Last, they give comments on current research and applications on text summarization and give their viewpoints for the possible research directions in the future.

DOI: 10.4018/978-1-5225-9373-7.ch003

INTRODUCTION

With the quick development of Internet and social media technologies, we have entered the era of information explosion. Online texts such as news, books, scientific and technical literature, microblogs, blog and production comments have generated and propagated quickly, which has formulate a huge corpus in the Web (Resnik and Smith, 2006). Massive texts make convenience for users to acquire information conveniently; however, information overloading becomes a new challenging problem. Although search engines such Google and Baidu have help users to search information quickly by providing a list of documents as the result, users still have to spend more time to browse and read the returned documents for seeking specific information and understanding the content, for example, the development of a specific event or the tracking of a specific object such as a person and an organization. How to provide brief and enough information for users becomes an urgent problem in big data era. Automatic summarization is a research direction for solving the information overloading problem.

Automatic summarization is to analyze one or more documents by machine and extract important information which is organized into a short and readable text (Gambhir and Gupta, 2017). Summarization is to compress the content of original documents and with the important content left. Users can grasp necessary information by reading the short summary generated by machine and save reading time. There are two ways of summarization generation, one is extractive summarization and the other is abstractive summarization (Gupta and Gupta, 2018).

Extractive summarization is to select important sentences from documents without changing the components of original sentences (Alguliyev et. al, 2018), while abstractive summarization is to generate new sentences based on understanding the content of documents. For extractive summarization, texts of documents have different information units such as section, paragraph, sentences, phrases and words. Extractive summarization approach assigns different weights to different information units via different weighting algorithms, and then select information units with high weights.

Abstractive summarization requires understanding of the text and produce summaries which fuse information from multiple sources effectively. Abstractive summarization needs syntactic and semantic analysis on texts, and then makes information fusion for generating sentences with natural language processing technologies.

Currently, most summarization systems adopt extractive summarization approach, and abstractive summarization technologies are harder than extracting sentences from texts. However, summarization research area has been gaining momentum with the shift towards semantic processing in recent years. An important distinction between