Data Mining Medical Digital Libraries

Colleen Cunningham
Drexel University, USA

Xiaohua Hu
Drexel University, USA

INTRODUCTION

Given the exponential growth rate of medical data and the accompanying biomedical literature, more than 10,000 documents per week (Leroy et al., 2003), it has become increasingly necessary to apply data mining techniques to medical digital libraries in order to assess a more complete view of genes, their biological functions and diseases. Data mining techniques, as applied to digital libraries, are also known as text mining.

BACKGROUND

Text mining is the process of analyzing unstructured text in order to discover information and knowledge that are typically difficult to retrieve. In general, text mining involves three broad areas: Information Retrieval (IR), Natural Language Processing (NLP) and Information Extraction (IE). Each of these areas are defined as follows:

- **Natural Language Processing:** a discipline that deals with various aspects of automatically processing written and spoken language.
- **Information Retrieval:** a discipline that deals with finding documents that meet a set of specific requirements.
- **Information Extraction:** a sub-field of NLP that addresses finding specific entities and facts in unstructured text.

MAIN THRUST

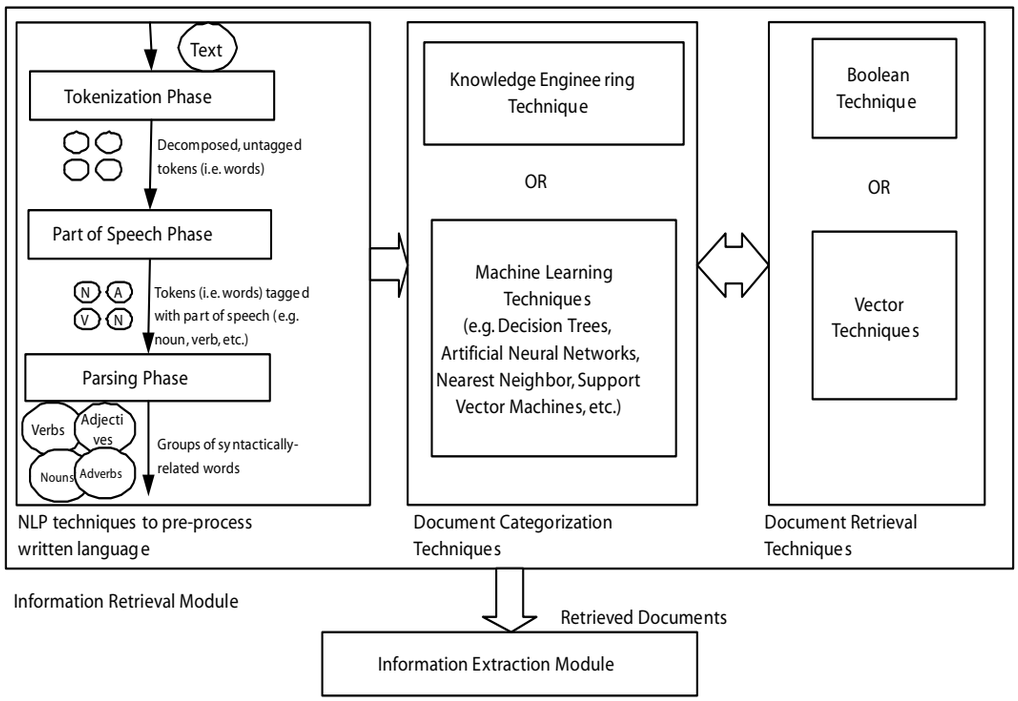
The current state of text mining in digital libraries is provided in order to facilitate continued research, which subsequently can be used to develop large-scale text mining systems. Specifically, an overview of the process, recent research efforts and practical uses of mining digital libraries, future trends and conclusions are presented.

Text Mining Process

Text mining can be viewed as a modular process that involves two modules: an information retrieval module and an information extraction module. presents the relationship between the modules and the relationships between the phases within the information retrieval module. The former module involves using NLP techniques to pre-process the written language and using techniques for document categorization in order to find relevant documents. The latter module involves finding specific and relevant facts within text. NLP consists of three distinct phases: (1) tokenization, (2) parts of speech (PoS) tagging and (3) parsing. In the tokenization step, the text is decomposed into its subparts, which are subsequently tagged during the second phase with the part of speech that each token represents (e.g., noun, verb, adjective, etc.). It should be noted that generating the rules for PoS tagging is a very manual and labor-intensive task. Typically, the parsing phase utilizes shallow parsing in order to group syntactically related words together because full parsing is both less efficient (i.e., very slow) and less accurate (Shatkay & Feldman, 2003). Once the documents have been pre-processed, then they can be categorized.

There are two approaches to document categorization: Knowledge Engineering (KE) and Machine Learning (ML). Knowledge Engineering requires the user to manually define rules, which can consequently be used to categorize documents into specific pre-defined categories. Clearly, one of the drawbacks of KE is the time that it would take a person (or group of people) to manually construct and maintain the rules. ML, on the other hand, uses a set of training documents to learn the rules for classifying documents. Specific ML techniques that have successfully been used to categorize text documents include, but are not limited to, Decision Trees, Artificial Neural Networks, Nearest Neighbor and Support Vector Machines (SVM) (Stapley et al., 2002). Once the documents have been categorized, then documents that satisfy specific search criteria can be retrieved.

Figure 1. Overview of text mining process



There are several techniques for retrieving documents that satisfy specific search criteria. The Boolean approach returns documents that contain the terms (or phrases) contained in the search criteria; whereas, the vector approach returns documents based upon the term frequency-inverse document frequency (TF x IDF) for the term vectors that represent the documents. Variations of clustering and clustering ensemble algorithms (Iliopoulous et al., 2001; Hu, 2004), classification algorithms (Marcotte et al., 2001) and co-occurrence vectors (Stephens et al., 2001) have been successfully used to retrieve related documents. An important point to mention is that the terms that are used to represent the search criteria as well as the terms used to represent the documents are critical to successfully and accurately returning related documents. However, terms often have multiple meanings (i.e., polysemy) and multiple terms can have the same meaning (i.e., synonyms). This represents one of the current issues in text mining, which will be discussed in the next section.

The last part of the text mining process is information extraction, of which the most popular technique is co-occurrence (Blaschke & Valencia, 2002; Jenssen et al., 2001). There are two disadvantages to this approach, each of which creates opportunities for further research. First, this approach depends upon assumptions regarding sentence structure, entity names, and etcetera that do not always hold true (Pearson, 2001). Furthermore, this approach relies heavily on completeness of the list of gene names and synonyms and summarizes the modular process of text mining.

Research to Address Issues in Mining Digital Libraries

The issues in mining digital libraries, specifically medical digital libraries, include scalability, ambiguous English and biomedical terms, non-standard terms and structure and inconsistencies between medical repositories (Shatkay & Feldman, 2003). Most of the current text mining research focuses on automating information extraction (Shatkay & Feldman, 2003). The scalability of the text mining approaches is of concern because of the rapid rate of growth of the literature. As such, while most of the existing methods have been applied to relatively small sample sets, there has been an increase in the number of studies that have been focused on scaling techniques to apply to large collections (Pustejovsky et al., 2002; Jenssen et al., 2002). One exception to this is the study by Jenssen et al. (2001) in which the authors used a predefined list of genes to retrieve all related abstracts from PubMed that contained the genes on the predefined list.