

Chapter 1.27

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 auto-

matically 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

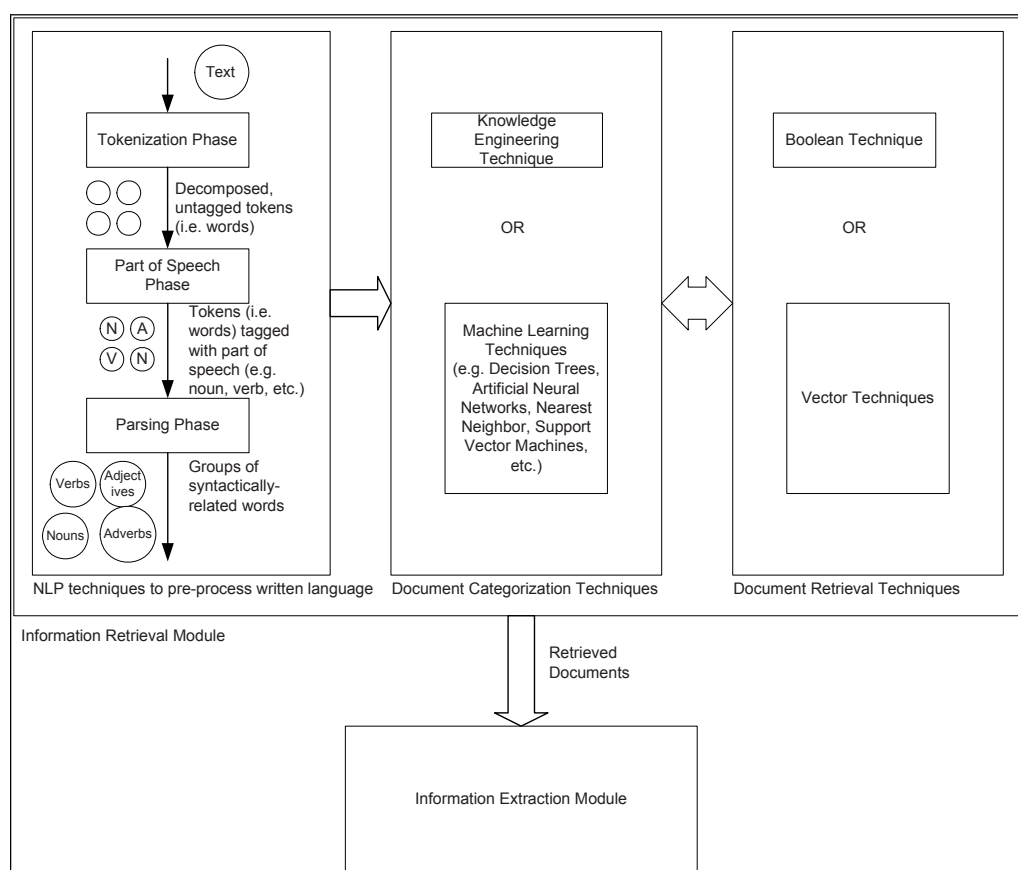
Text mining can be viewed as a modular process that involves two modules: an information retrieval module and an information extraction

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.

Figure 1. Overview of text mining process



5 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/chapter/data-mining-medical-digital-libraries/26227

Related Content

Application of Text Mining Methodologies to Health Insurance Schedules

Ah Chung Tsoi, Phuong Kim To and Markus Hagenbuchner (2009). *Medical Informatics: Concepts, Methodologies, Tools, and Applications* (pp. 944-963).

www.irma-international.org/chapter/application-text-mining-methodologies-health/26272

High-Throughput Data Analysis of Proteomic Mass Spectra on the SwissBioGrid

Andreas Quandt, Sergio Maffioletti, Cesare Pautasso, Heinz Stockinger and Frederique Lisacek (2009). *Handbook of Research on Computational Grid Technologies for Life Sciences, Biomedicine, and Healthcare* (pp. 228-244).

www.irma-international.org/chapter/high-throughput-data-analysis-proteomic/35696

Technology Enablers for Context-Aware Healthcare Applications

Filipe Meneses and Adriano Moreira (2009). *Mobile Health Solutions for Biomedical Applications* (pp. 260-269).

www.irma-international.org/chapter/technology-enablers-context-aware-healthcare/26775

Arabidopsis Homologues to the LRAT a Possible Substrate for New Plant-Based Anti-Cancer Drug Development

Dimitrios Kaloudas and Robert Penchovsky (2018). *International Journal of Biomedical and Clinical Engineering* (pp. 40-52).

www.irma-international.org/article/arabidopsis-homologues-to-the-lrat-a-possible-substrate-for-new-plant-based-anti-cancer-drug-development/199095

Development of an Affordable Myoelectric Hand for Transradial Amputees

Alok Prakash and Shiru Sharma (2020). *International Journal of Biomedical and Clinical Engineering* (pp. 1-15).

www.irma-international.org/article/development-of-an-affordable-myoelectric-hand-for-transradial-amputees/240742