Chapter 4

Concept-Based Mining Model

Shady Shehata
University of Waterloo, Canada

Fakhri Karray
University of Waterloo, Canada

Mohamed Kamel
University of Waterloo, Canada

ABSTRACT

Most of text mining techniques are based on word and/or phrase analysis of the text. Statistical analysis of a term frequency captures the importance of the term within a document only. However, two terms can have the same frequency in their documents, but one term contributes more to the meaning of its sentences than the other term. Thus, the underlying model should indicate terms that capture the semantics of text. In this case, the model can capture terms that present the concepts of the sentence, which leads to discover the topic of the document. A new concept-based mining model that relies on the analysis of both the sentence and the document, rather than, the traditional analysis of the document dataset only is introduced. The concept-based model can effectively discriminate between non-important terms with respect to sentence semantics and terms which hold the concepts that represent the sentence meaning. The proposed model consists of concept-based statistical analyzer, conceptual ontological graph representation, and concept extractor. The term which contributes to the sentence semantics is assigned two different weights by the concept-based statistical analyzer and the conceptual ontological graph representation. These two weights are combined into a new weight. The concepts that have maximum combined weights are selected by the concept extractor. The concept-based model is used to enhance the quality of the text clustering, categorization and retrieval significantly.

INTRODUCTION

Due to the daily rapid growth of the information, there are considerable needs in extracting and discovering valuable knowledge from the vast amount of information found in different data sources today such

DOI: 10.4018/978-1-60566-908-3.ch004
as World Wide Web. Data mining in general is the field of extracting useful information, and sometimes high-level knowledge, from large sets of raw data. It has been the attention of many researchers to find efficient ways to extract useful information automatically from such information sources.

Text Mining is the process of deriving high quality information from text by discovering patterns and trends through different written resources. Text mining is generally considered more difficult than traditional data mining. This is attributed to the fact that traditional databases have fixed and known structure, while text documents are unstructured, or, as in the case of web documents, semi-structured. Thus, text mining involves a series of steps for data pre-processing and modeling in order to condition the data for structured data mining. Text mining can help in many tasks that otherwise would require large manual effort. Common problems solved by text mining include, but not limited to, searching through documents, organizing documents, comparing documents, extracting key information, and summarizing documents. Methods in information retrieval, machine learning, information theory, and probability are employed to solve those problems.

Natural Language Processing (NLP) is both a modern computational technology and a method of investigating and evaluating claims about human language itself. NLP is a term that links back into the history of Artificial Intelligence (AI), the general study of cognitive function by computational processes, with an emphasis on the role of knowledge representations. The need for representations of human knowledge of the world is required in order to understand human language with computers. Text mining attempts to discover new, previously unknown information by applying techniques from natural language processing and data mining.

The problem introduced by text mining is that natural language was developed for humans to communicate with one another and to record information. Computers are a long way from understanding natural language. Humans have the ability to understand the meaning of text and humans can easily overcome obstacles that computers cannot easily handle such as spelling variations and contextual meaning. However, although human mind can understand the meaning of unstructured data, human lacks the computer’s ability to process text in large volumes or at high speeds. Herein lays the key to create a new technology called concept-based mining that combines the human way of understanding with the speed and accuracy of a computer.

**MAIN FOCUS**

Most of text mining techniques are based on word and/or phrase analysis of the text. Statistical analysis of a term frequency captures the importance of the term within a document only. However, two terms can have the same frequency in their documents, but one term contributes more to the meaning of its sentences than the other term. Thus, a mining model should indicate terms that capture the semantics of text. In this case, the model can capture terms that present the concepts of the sentence, which leads to discover the topic of the document. Research suggests that there are important concepts either words or phrases in each sentence that can truly present the meaning of a sentence.
Related Content

Constrained Cube Lattices for Multidimensional Database Mining
Alain Casali, Sébastien Nedjar, Rosine Cicchetti and Lotfi Lakhal (2010). *International Journal of Data Warehousing and Mining* (pp. 43-72).
[www.igi-global.com/article/constrained-cube-lattices-multidimensional-database/44958?camid=4v1a](www.igi-global.com/article/constrained-cube-lattices-multidimensional-database/44958?camid=4v1a)

Identifying and Analyzing Popular Phrases Multi-Dimensionally in Social Media Data
Zhongying Zhao, Chao Li, Yong Zhang, Joshua Zhexue Huang, Jun Luo, Shengzhong Feng and Jianping Fan (2015). *International Journal of Data Warehousing and Mining* (pp. 98-112).

A Holistic View of Big Data
Won Kim, Ok-Ran Jeong and Chulyun Kim (2014). *International Journal of Data Warehousing and Mining* (pp. 59-69).
[www.igi-global.com/article/a-holistic-view-of-big-data/116893?camid=4v1a](www.igi-global.com/article/a-holistic-view-of-big-data/116893?camid=4v1a)

Two Case-Based Systems for Explaining Exceptions in Medicine
Rainer Schmidt (2009). *Data Mining and Medical Knowledge Management: Cases and Applications* (pp. 227-249).
[www.igi-global.com/chapter/two-case-based-systems-explaining/7535?camid=4v1a](www.igi-global.com/chapter/two-case-based-systems-explaining/7535?camid=4v1a)