Chapter 3
Semantic-Based Indexing Approaches for Medical Document Clustering Using Cognitive Search

Logeswari Shanmugam
Bannari Amman Institute of Technology, India

Premalatha K.
Bannari Amman Institute of Technology, India

ABSTRACT

Biomedical literature is the primary repository of biomedical knowledge in which PubMed is the most absolute database for collecting, organizing and analyzing textual knowledge. The high dimensionality of the natural language text makes the text data quite noisy and sparse in the vector space. Hence, the data preprocessing and feature selection are important processes for the text processing issues. Ontologies select the meaningful terms semantically associated with the concepts from a document to reduce the dimensionality of the original text. In this chapter, semantic-based indexing approaches are proposed with cognitive search which makes use of domain ontology to extract relevant information from big and diverse data sets for users.

INTRODUCTION

The huge volume of biomedical text in the online repositories provides a rich source of knowledge for biomedical research. The rapid growth of data is a challenge for the modern society and the 80% percent of the current available data is not structured or indexed. As the cognitive computing becomes the current era, new searching techniques are evolved by combining powerful indexing technology with advanced Natural Language Processing (NLP) capabilities and machine learning algorithms in order to build an increasingly deep corpus of knowledge.

DOI: 10.4018/978-1-5225-7522-1.ch003

Text mining takes advantage of machine learning specifically in determining features, reducing dimensionality and removing irrelevant attributes. Text mining facilitates the researchers to extract information and mine knowledge from a pile of text and it is now extensively applied in biomedical research. Text mining outcomes are obtained with noisy information and false positives from natural language text. This is due to the ambiguities caused by semantics, syntax, sparsity of class specific core words and high dimensionality. In the recent researches many methods have been developed to facilitate discovering trends and patterns in medical documents. These researches proved that the new searching techniques developed are no longer based on just keyword matching; all these techniques are become cognitive with the ability to deliver the most relevant answers to search queries. It is observed from the literature that the inclusion of domain knowledge with the cognitive search during the mining process enhances the efficiency as well as the quality of the mined patterns.

The high dimensionality of the natural language makes the text data quite noisy and sparse in the vector space model. There is a possibility that mining may lead to inaccurate results in clusters if the input has noisy information. Hence the data preprocessing and feature selection are the important practices for the text mining. The transformation from unstructured text document into Bag-of-Words (BOW) representation is the compassion of document indexing. The preprocessing tasks which include tokenization, stop-word removal, Part-of-Speech (POS) tagging and weighting are incorporated during indexing.

The traditional term-based indexing suffers on semantic issues related to the synonymy and polysemy problems. Thus, the term-based methods are not appropriate for clustering the medical documents which involve with complex semantics. In order to deal with the issues, a concept-based indexing is proposed for medical document clustering using Medical Subject Headings (MeSH) ontology as the domain reference.

The MEDLINE contains a major entry point to biomedical research for biologists (Hersh 2008). The handling of biomedical domain is complex due to its ambiguous nature of terminologies. The cost of manual indexing of the biomedical documents is high; so many efforts have been prepared in order to offer automatic indexing. The MEDLINE database gives MeSH ontology for biomedical research articles. MeSH based representations cover the conceptual content of entire articles. Its representation has been shown to be consistent across different indexers (Funk & Reid 1983). Hence MeSH-based document representation gets more attraction in IR. The Mesh descriptors are not only sufficient for extracting information from the PubMed documents. Feature or term weighting is an important part in the process of IRS. Precise term weighting can greatly improve the process of finding index terms. The amount of influence of term in representing the document reflects on term weight. Hence a concept-based indexing is proposed for biomedical document clustering with concept weight which is computed using the frequency and weight of the semantic relation.

LITERATURE REVIEW

Medical document analysis is one of the innovative fields with remarkable research potential. It employs with the extraction of novel, significant information from the huge quantity of biomedical associated documents. The substantial amount of biomedical text offers a comfortable source of knowledge for biomedical research.

IR is involved with choosing from a group of documents, those that are probable to be appropriate to a user's information requirement expressed using a query. The objective of IR is to extend the users with documents that satisfy their requirements. The retrieval of documents involves with the indexing