Chapter 19
Deep Model Framework for Ontology-Based Document Clustering

U. K. Sridevi  
Sri Krishna College of Engineering and Technology, India

P. Shanthi  
Sri Krishna College of Engineering and Technology, India

N. Nagaveni  
Coimbatore Institute of Technology, India

ABSTRACT

Searching of relevant documents from the web has become more challenging due to the rapid growth in information. Although there is enormous amount of information available online, most of the documents are uncategorized. It is a time-consuming task for the users to browse through a large number of documents and search for information about the specific topics. The automatic clustering from these documents could be important and has great potential to improve the efficiency of information seeking behaviors. To address this issue, the authors propose a deep ontology-based approach to document clustering. The obtained results are encouraging and in implementation annotation rules are used. The work compared the information extraction capabilities of annotated framework of using ontology and without using ontology. The increase in F-measure is achieved when ontology as the distance measure. The improvement of 11% is achieved by ontology in comparison with keyword search.

INTRODUCTION

The increase in the growth of text documents in the Web is a great challenge to information retrieval system. The searching and indexing systems are available for accessing the information but the retrieval of relevant information is still a problem. One current problem of information retrieval is that it is not really possible to extract relevant documents automatically. An information retrieval system uses index-
Deep Model Framework for Ontology-Based Document Clustering

ing and the system’s performance depends on the quality of the indexing. The two main challenges in indexing are to create representative internal descriptions of documents and to organize these descriptions for fast retrieval. Descriptions of documents in information retrieval are supposed to reflect the documents content and establish the foundation for the retrieval of information when requested by users. The documents are marked with the description in indexing for easy retrieval.

Ontology has good conceptual structure representation and can be combined with the knowledge representation. The model makes use of annotation and indexing. The ontology model depends on the semantic index terms but the vector space model depends on the keyword index. The semantics of the concepts are used to build a concept term representation. The ontology similarity measure improves the concept relevance score. The semantically related terms gain more weights and it will improve the term importance in indexing process. The semantic analysis should somehow recognize concepts in the documents and then map them into the ontologies. The indexing process maps information found in documents into the ontology, identifying concepts and their positions in the ontology. Information in queries can similarly be mapped into the ontology and thus in addition to retrieving the exact match, the structure of the ontology can be used to retrieve semantically related documents. Semantic similarity and indexing focuses on the similarity measure using ontology. It also compares the vector space model with semantic information retrieval model. The methods are integrated to find the concept relation information, while these concepts are considered to be independent in the term vector space method. Using the ontology similarity method given in Euzenat and Shvaiko (2007), the cosine similarity between concept are measured. The term reweighting approaches based on ontology is used in information retrieval applications (Varelas et al., 2005). The semantic annotation process includes the creation of domain ontology and the ontology maps into the concept terms of the documents. In this model, the weight of the concepts is computed using their semantic similarity to other concepts in the document. The concept vector is generated in the document annotation process and the concept index is built. To improve the recognition of important indexing terms, it is possible to weight the concepts of a document in different ways (Valkeapaa et al., 2007).

Text mining algorithm can handle the real-world data that come in a diversity of forms and can be tremendously bulky (Pankaj et al., 2015). The work provides ontology framework based on text analytics and social media analytics. Social tagging system improves the personalized document clustering. The knowledge gained from social tagging system should be tremendous assets for conducting and improving various business intelligent applications (Yang et al., 2015).

In text clustering there exist some issues to tackle such as feature extraction and data dimension reduction. To overcome these problems, Yi et al (2017) presented a novel approach named deep-learning vocabulary network. Deep learning is used to extract the features of the text document and in the process of clustering and extract features of text documents. Yan et al. (2015) has used semantic representation and deep belief network for document classification and retrieval. However, there are very few publications addressing semantic indexing with deep learning. Yan et al. (2016) included the semantic indexing in biomedical literature by including a vast amount of semantic labels from automatically annotating MeSH terms for MEDLINE.

Chen et al. (2016) proposed a deep learning based framework to identify and categorize topics related to diabetes in online Chinese articles. The experiments use the datasets with over 19,000 online articles showed that the framework achieved a higher effectiveness and accuracy in categorizing diabetes related topic. Deep learning techniques are applied in the classification of texts for the enhanced construction regulated documents (Zhou & Ei-Gohary, 2015). They have developed domain specific ontology to