Chapter 9
Ontology Based Feature Extraction From Text Documents

Abirami A.M
Thiagarajar College of Engineering, India

Askarunisa A.
KLN College of Information Technology, India

Shiva Shankari R A
Thiagarajar College of Engineering, India

Revathy R.
Thiagarajar College of Engineering, India

ABSTRACT

This article describes how semantic annotation is the most important need for the categorization of labeled or unlabeled textual documents. Accuracy of document categorization can be greatly improved if documents are indexed or modeled using the semantics rather than the traditional term-frequency model. This annotation has its own challenges like synonymy and polysemy in the document categorization problem. The model proposes to build domain ontology for the textual content so that the problems like synonymy and polysemy in text analysis are resolved to greater extent. Latent Dirichlet Allocation (LDA), the topic modeling technique has been used for feature extraction from the documents. Using the domain knowledge on the concept and the features grouped by LDA, the domain ontology is built in the hierarchical fashion. Empirical results show that LDA is the better feature extraction technique for text documents than TF or TF-IDF indexing technique. Also, the proposed model shows improvement in the accuracy of document categorization when domain ontology built using LDA has been used for document indexing.

DOI: 10.4018/978-1-5225-4044-1.ch009

INTRODUCTION

Necessity of annotating the text documents has become increased for analyzing the large amount of documents existing in the World Wide Web. But most of the documents are in unstructured format and the machines cannot simply process them. People who buy/sell the products give their comments, feedback, additional features needed, etc., in the form of text which is mostly unstructured. It becomes necessary to categorize these voluminous texts to make business intelligent solutions. The huge data available in the internet has to be modeled, analyzed and then the decision has to be taken. Retrieving the information from the unstructured text is the difficult task. Document annotation with added semantics enables the information or knowledge extraction from the repository in an intelligent way.

Feature extraction is the process which starts from an initial set of measured data and builds features intended to be informative and non-redundant. It involves reducing the amount of resources required to represent a large set of data. Many algorithms are used for identifying the features from the textual data that requires grouping or classifying the entities based on their similar property.

Some of the problems faced with feature extraction by traditional methods are: (i) existing techniques aren't compatible with the current Web size and growth rate and hence automated techniques are essential if practical and scalable solutions are to be obtained (ii) absence of semantic relations between concepts in feature search processes (iii) imperfections in classifying the feature reviews into more degrees of polarity terms and (iv) misinterpretation of textual features due to lack of prior knowledge.

Ontology is a set of concepts and categories in a subject area or domain that shows their properties and the relations between them. Domain specific Ontology represents the particular meanings of terms as they apply to that domain. The semantic web technologies can be used to model the textual data to represent domain vocabularies and their relationships through Ontologies, RDF, etc. The analysis has to be done in such a way that the context has to be matched both between the writer and the reader. All these challenges can be well handled by representing the different vocabularies for the domain, and their relationship between the concepts. Ontology-based information extraction is the use of ontologies and their specifications to "drive" or inform the information extraction process. The terms and concepts in the source Ontology form the basis for term matching when tagging text documents. Difficulties in feature extraction problems can be overcome if the text document can be modeled using the Ontology representation along with the use of topic modeling techniques. The objective of this proposed work is set to build domain Ontology for the set of documents with relevant features extracted from the text documents.

BACKGROUND

Use of Ontology in Information Extraction

Ontology is an explicit description of a domain. It defines a concept by describing its properties, attributes, and constraints. It defines a common vocabulary and it gives a shared understanding. UML diagrams are used to identify and classify biological entities and interactions between proteins and genes using Ontology (Rindfleisch, 2000). Ontology provides a formal conceptualization of a particular domain that is shared by a group of people. In the context of the Semantic Web, Ontology describes domain theories for the explicit representation of the semantics of the data (Maedeche, 2003). Ontology-based