Chapter 4

Applying Semantic Relations for Automatic Topic Ontology Construction

Subramaniyaswamy Vairavasundaram
SASTRA University, India

Logesh R.
SASTRA University, India

ABSTRACT

The rapid growth of web technologies had created a huge amount of information that is available as web resources on Internet. Authors develop an automatic topic ontology construction process for better topic classification and present a corpus based novel approach to enrich the set of categories in the ODP by automatically identifying concepts and their associated semantic relationships based on external knowledge from Wikipedia and WordNet. The topic ontology construction process relies on concept acquisition and semantic relation extraction. Initially, a topic mapping algorithm is developed to acquire the concepts from Wikipedia based on semantic relations. A semantic similarity clustering algorithm is used to compute similarity to group the set of similar concepts. The semantic relation extraction algorithm derives associated semantic relations between the set of extracted topics from the lexical patterns in WordNet. The performance of the proposed topic ontology is evaluated for the classification of web documents and obtained results depict the improved performance over ODP.

INTRODUCTION

The unbridled growth of World Wide Web (WWW) has made a huge amount of information and resources available over the Internet. This rapid growth of information has resulted in searching for information on the web a challenging task (Sridevi & Nagaveni, 2011). In order to access web resources, a large number of standard web mining algorithms and information retrieval techniques have been developed based on simple keyword based matching. Yet, in a large corpus of documents, the users are unable to retrieve the desired information because these techniques do not consider semantic concepts in the
web contents (Fortuna, Grobelnik, & Mladenic, 2005). To overcome this challenge, a semantic web is evolved with ontologies to describe the conceptual relationship between entities in a specific domain. Ontologies are simply defined as the taxonomy of the hierarchy of concepts. It is mainly constructed to provide the knowledgeable representation that can describe the web resources using intelligent techniques for human understanding and machine processing (David & Antonio, 2004). In ontology, concepts in a specific domain are formulated using a proper encoding mechanism that can support efficient information retrieval and reduced information load due to the large corpus of documents (Nicola, 1998). An Ontology creation methodology for domain experts should be efficient and easy to learn (Nikolai, 2011). Ontology represents a set of concepts and the relationships among them for a particular domain (Jongwoo & Veda, 2011).

Topic ontology is defined as a hierarchy of a set of topics that are interconnected using semantic relations (Xujuan, Yuefeng, Yue & Raymond, 2006). It is denoted as a graph in which each node represents the specific topic that forms a topic hierarchy. Further, a group of relevant topics is related to the specific concept in the topic ontology by maintaining a hierarchical semantic relationship among the concepts in topics. The construction process of topic ontology involves extracting keywords using standard text mining and information retrieval techniques. The construction is purely based on semantic relevance of the keywords. However, the keyword based construction approach is not efficient as it is not possible to construct ontology from the large corpus of web documents (Ana, Rocio, Carlos & Filippo, 2010).

Due to the shortcomings of keyword based construction, we propose the Open Directory Project (ODP), a multilingual open content directory of World Wide Web links (Dengya & Heinz, 2009). The ODP works on the principle of listing out the set of categories related to a specific concept. We propose a hyperlink based approach, wherein ontology is constructed through exploring and discovering the semantic concepts related to the categories associated in ODP. The main advantage of this approach is to allow the user to extend the categories according to their perspective to construct topic ontology. This approach merely requires the users to have a basic knowledge of the topic that they are searching to enrich the existing ontology. Hence, we deploy knowledge-based web resources, such as Wikipedia and WordNet to obtain the background semantic knowledge about the categories in the ODP.

Wikipedia, most commonly used as online encyclopedia by the Internet users, is a multilingual online encyclopedia which contains articles and categories that are related to a particular concept. In Wikipedia, each web page defines a set of topics related to the specific concept and is more formally organized in hierarchical fashion according to the general knowledge. In addition, unambiguity between similar words is removed by providing web pages according to the different word sense. Wikipedia uses a straightforward approach for searching; usually a set of keywords associated with each web page can be used to construct topic ontology. Constructing topic ontology based on keywords alone has limitations as keywords are just an arrangement of a set of relevant topics related to the specific concept and does not render any semantic relation among the set of concepts (Huan, Xing, Liang-Tien & Ah-Hwee, 2010).

WordNet is a lexical reference system that consists of a set of related usual language terms within the same lexical concepts. Each word possesses one or more synonyms depending upon the parts of speech, such as nouns, verbs, adjectives and adverbs. A collection of words that represent the same sense is known as synset and is organized hierarchically (Vaclav, Pavel & Jaroslav, 2005). A set of related terms in synset possess Hyponym/Hypernym (subclass/superclass, represented as is–a relation) and Meronym/Holonym (part-whole relationships) which is used to facilitate the efficient construction of topic ontology. Figure 1 shows the partial hierarchical structure of the tag “Java” based on the relations of WordNet.