Chapter 6
Knowledge Extraction from Domain-Specific Documents

Ram Kumar
Barkatullah University, India

Shailesh Jaloree
SATI, India

R. S. Thakur
MANIT, India

ABSTRACT

Knowledge-based systems have become widespread in modern years. Knowledge-base developers need to be able to share and reuse knowledge bases that they build. As a result, interoperability among different knowledge-representation systems is essential. Domain ontology seeks to reduce conceptual and terminological confusion among users who need to share various kind of information. This paper shows how these structures make it possible to bridge the gap between standard objects and Knowledge-based Systems.

INTRODUCTION

In this paper, we proposed a technique for Knowledge extraction from domain specific documents using ontology and K-Means. Clustering for automatic extraction of Knowledge from the various sources (web, databases etc...) on (from) specific documents and for organize them. The documents collected according to (the) user specifications and are organized in form of ontology. The system offers support to the user during the extraction process by extracting keywords and groups them into classes & concepts by using K-Means Clustering. Web characteristics, such as dimension and dynamics (Levene, 2004) place many difficulties to users willing to explore it as an information source. Moreover, information retrieved from the Web is typically a large collection of documents. A query in Google for "Artificial Intelligence" gives, today, a list of 95.000.000 results. Organizing this information conveniently improves the efficiency of its exploitation. To take advantage of the value contained in this huge information system there is a need

DOI: 10.4018/978-1-5225-0536-5.ch006

for tools that help people to explore it and to retrieve, organize and analyze relevant information. It is also important to give the user the possibility of specifying how he or she requires the retrieved documents to be organized. When working with large corpora of documents it is hard to comprehend and process all the information contained in them. Standard text mining and information retrieval techniques usually rely on word matching and do not take into account the similarity of words and the structure of the documents within the corpus. We try to overcome that by automatically extracting the keywords covered within the documents in the corpus and helping the user to organize them into ontology.

BACKGROUND

Ontology, in its original meaning, is a branch of philosophy (specifically, metaphysics) concerned with the nature of existence. It includes the identification and study of the categories of things that exist in the universe. One scenario of ontology is in Artificial Intelligence, where it is defined as "An ontology is a formal, explicit specification of shared conceptualization. This definition is given by Gruber (Gruber, 1993) which is most commonly used by knowledge engineering community. Here Conceptualization is a "world view" that often present as a set of concepts and their relations. It is the abstract representation of a real world entity (view) with the help of domain relevant concepts (Bhowmick, 2010). Since the ontologist has huge amount of knowledge which is unstructured and it should be organized. Conceptualization helps to organize and structures the acquired knowledge using external representations that are independent of the implementation languages and environments (Arabshian, 2012). Ontology refers to the shared understanding of a domain of interest and is represented by a set of domain relevant concepts, the relationships among the concepts, functions and instances (Bhowmick, 2010).Ontology is used for representing the knowledge of a domain in a formal and machine understandable form in areas like intelligent information processing. Thus it provides the platform for effective extraction of information and many other applications (Choudhary, 2012). It is very useful for expressing and sharing the knowledge of semantic web.

Ontology is a set of concepts connected with different types of relations. Each topic includes a set of related documents. Construction of such ontology from a given corpus can be a very time consuming task for the user. In order to get a feeling on what the topics in the corpus is, what the relations between topics are and, at the end, to assign each document to some certain topics, the User has to go through all the documents. This system aims at assisting the user in a fast extraction of keywords and semi-automatic construction of the ontology from a large document collection. Domain Ontology refers to a specific vocabulary used to describe a certain reality. It presupposes the identification of the key concepts and relationships in the domain of interest. Domain ontology seeks to reduce or eliminate conceptual and terminological confusion among the members of a user community who need to share various kinds of electronic documents and information. It does so by identifying and properly defining a set of relevant concepts that characterize a given application domain, say, for travel agents or medical practitioners. An ontology specifies a shared understanding of a domain. It contains a set of generic concepts (such as "object," "process," "accommodation," and "single room"), together with their definitions and interrelationships. The construction of its unifying conceptual framework fosters communication and cooperation among people, better enterprise organization, and system interoperability. It also provides such system- engineering benefits as reusability, reliability, and Specification. Ontologies can have different degrees of formality, but they must include metadata such as concepts, relations, axioms, instances, or