Automatic Discovery of Concepts from Text

Ong Siou Chin
Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia.
scong@fit.unimas.my

Narayanan Kulathuramaiyer
Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia.
nara@fit.unimas.my

Alvin W. Yeo
Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia.
alvin@fit.unimas.my

Abstract

Existing mechanisms for concept discovery tend to pick up all possible relationships between terms in a document based on roles of terms identified [3]. The proposed work aims to enhance this discovery process by employing machine learning and semantic modelling. We explore a framework for automatically discovering labeled clusters from a large collection of documents. The aim of this framework is to enable the extraction of concepts and to structure these into labeled concepts for use by text processing applications such as text summarization and text categorization. We have developed a mechanism for automatically inducing a set of words that captures the meaning of a collection of documents. The WordNet lexical database is used to extract root meanings and to determine relationships amongst these terms.

1. Introduction

The Semantic Web is an extension of the World Wide Web (WWW), where semantic approaches will be integrated into WWW to improve the human-computer cooperation. These semantic approaches are required as keyword matching techniques for dealing with textual data which has become a bottleneck of the WWW. These approaches employ ontology as a data model to represent a domain and to reason about the objects with connections between them.

Manually crafting of the ontologies will require a great deal of effort and it can be extremely time consuming. Therefore the Ontolgy Learning has become an important area for acquiring and organising knowledge.

In the next section we discuss related works in concept discovery. The framework for automatic discovery of labeled clusters together with the design of the concept discovery is presented in Section 3. Section 4 describes the results followed by a section on some discussion.

Concept discovery is an important part in supporting ontology learning. Concept discovery or concept formation is the process of exploring new concepts by making connection or relationships between objects. A concept is defined as "a perceived regularity in events or objects, or records of events or objects, designated by a label", while propositions or rules are "statements about some object or event in universe, either naturally occurring or constructed" [5]. Concepts linked by propositions will be the fundamental unit of ontology. The knowledge structure employed is WordNet. We also proposed a novel approach of automatically resolve word sense ambiguity and labelling of clusters.

2. Related work

OntoLearn is an ontology population method, which is based on machine learning and text mining approach which learns domain concepts and taxonomic relations from input data [4]. OntoLearn combines natural language processing and statistical techniques for terminology extraction. Their Structural Semantic Interconnections (SSI) approach employs a syntactic pattern-matching algorithm to perform word sense disambiguation. A compositional interpretation technique is applied whereby the meaning of complex terms can be derived from its components [4]. Semantic relations between the components of a complex concept are determined using Context grammar rules. After a complex term is semantically interpreted, it is then integrated into the initial ontology (WordNet) and linked to a suitable parent node.

The Mo'K workbench [2] explores the learning of semantic classes and conceptual clustering for the purpose of ontology learning. In learning a semantic class, syntactic analysis is performed to describe the syntactic

Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06)
0-7695-2747-7/06 $20.00 © 2006 IEEE