Abstract

Ontologies are key components in the development of the SemanticWeb, as they provide a shared understanding of a domain. However, acquiring the knowledge needed to construct ontologies is a costly task that requires much time and many resources. This task can be supported by ontology learning methods, which are often based on machine learning techniques, natural language processing or clustering algorithms. This work describes the OntOAIr method (Ontologies from Open Archives Initiative Repositories to Support Information Retrieval), a semi-automatic construction method of lightweight ontologies called *ontologies of records*. The OntOAIr method allows human and software agents to organize and retrieve groups of documents from multiple collections. This method uses simplified representations of documents, an adaptation of the Frequent Itemset-based Hierarchical Clustering algorithm (FIHC), and ontological engineering techniques. Schemas and namespaces are defined to formalize the constructed ontologies; and different levels of expressivity are explored by means of mappings using the XML, RDF and OWL languages. Two applications of the OntOAIR method are described: (1) a keyword-based retrieval model, and (2) an ontology-based exploration model. An agent-based architecture supports these models in a prototypical system called *OntoSIR*. We present positive experimental results that use similar collections to those provided by the Open Archives Initiative (OAI) as a testbed. The OntOAIr method can be used to support the manual construction of ontologies, to cluster the responses of search engines, or as the basis to support reasoning in the Semantic Web. This method is a step towards the formal encoding of the contents of multiple document collections.