Methodology for Semi-automatic Ontology construction using Ontology learning : A Survey

Pradnya Gotmare
Department of Computer Engineering,
K.J. Somaiya College of Engineering,Vidyavihar
Mumbai-77(India)

ABSTRACT
Modern information system is moving from data processing towards concept processing. Semantic web Technologies offer a new approach to manage information and processes by adding meaning to the data. Ontology is a structure where knowledge about a particular domain is described by relevant concepts and relations between them.

Ontology learning refers to the task of automatically creating ontology by extracting concepts and relationships from the given data set. Manual method of ontology construction is expensive and time consuming. So the aim of this research is to automate the process of ontology building by using the techniques of ontology learning .

The various issues addressed regarding the automated learning include use of Semantic annotations and use of Controlled Language for Information Extraction (CLIE) which is a subset of natural language. Natural language processing and machine learning techniques can be useful in order to build ontologies in semiautomatic way.

Keywords
Semantic Web Technology, Ontology learning , Natural language processing ,Controlled Language Information Extraction (CLIE).

1. INTRODUCTION
An ontology is a formal specification of a conceptualization of a domain of interest. In other words, it is the meaningful representation of knowledge . Ontologies are used for organizing knowledge in a structured way in many applications. Ontology learning consists of various tasks. The different tasks includes identifying the concepts, identifying the relationships among concepts, populating the ontology with instances and ontology updating. In other words, Ontology learning refers to the task of automatically creating an ontology by extracting concepts and relationships from the given data set . The data can be in the form of text document, web page, relational form etc. Different approaches have been used for building ontologies, most of them are mainly manual methods. It involves various stages like identifying the purpose of ontology, utilization of it and the range of users. In this process user identifies the key concepts and relationships among them. Manual method of ontology construction is expensive and time consuming.

In semiautomatic ontology learning some of the classes and relationships are given and some are missing. So it is necessary to find out the missing one and add those in the preconstructed ontology. Ontology learning algorithm helps the user to understand how to represent the knowledge by using different concepts.. It can be used to provide the user with an aggregated view of the knowledgebaseewhich contains data, concepts, instances, relations with them.

The various issues addressed regarding the automated learning include use of Semantic annotations , use of knowledge discovery techniques, and use of Controlled Natural Languages for information extraction. Natural language processing and machine learning techniques can be useful in order to build ontologies in an automatic or semiautomatic way.

The remainder of paper is organized as given below: Section 2 presents current knowledge management techniques, Section 3 outlines the proposed system overview.. Section 4 gives a summary of research work and future scope is mentioned.

2. RELATED WORK
An ontology learning is a formal representation of knowledge. It consists of different components like concepts, relationships, instances and axioms about classes and properties. The OWL language provides the mechanism for creating all the components of ontology. There are two types of properties as object properties and data properties[4]. Object properties relate instances to instances and data type properties relate instances to data type values. The construction of ontology is a time consuming and expensive process. It requires the knowledge of ontology engineering and knowledge about domain of interest. There are various methodologies for building ontologies. All the methodology includes following steps.

1. To identify the purpose of thr domain for which ontology needs to be build.
2. To capture the concepts and relationships between these concepts
3. To capture the terms used to refer to these concepts and relationships and
4. To code the ontology.

Methontology [11] is a methodology for building ontologies either from scratch , or reusing other ontologies as they are, or by a process of reengineering them. A knowledge engineer defines an initial ontology. It is then extended and changed with the feedback taken from a panel of domain experts.

The problem under study is how to apply ontology learning methods to automate the ontology building process. Another problem is to integrate the information from various data sources and various data forms.

2.1 Methodology for semiautomatic ontology construction
For semiautomatic ontology construction different knowledge discovery techniques can be used. It is difficult to completely automate the process of ontology construction. At the same time manual method of ontology construction is expensive

and time consuming. Therefore the technology should provide the support in order to minimize human efforts required for ontology building. The existing ontology can be refined by considering the suggestions provided by machine. Thus technology should provide the support for identifying interesting information in order to minimize human intervention. From knowledge discovery point of view ontology learning tasks are about finding the different classes(concepts) , relationships among the classes. It is about finding the mappings between ontology components . Knowledge discovery technique based on machine learning can be applied to Ontology learning. Some of those techniques can be described as

1. Supervised learning
2. Unsupervised learning
3. Semi- supervised learning
4. Active learning

Ontology learning :

Ontology learning includes learning ontology concepts, learning ontology relationships between existing concepts, learning both the concepts and relations at a time, populating the existing ontology structure, dealing with dynamic data streams. It also include construction of ontologies giving different views on the same data. It is basically finding the mapping among ontology components. Data on the web is of heterogeneous type. It consists of text documents, images, data records etc. Another problem of web data is of dynamic nature, as the data on the web is continuously updated. By applying unsupervised learning algorithm such as clustering , similarity between the objects used within documents can be identified. In semiautomatic approach, ontology is constructed from the topics(concepts) from the documents data. Machine can provide suggestions for the topics appearing in the documents and can assists human by automatically assigning documents to the topics. It can suggests naming of the topics. From the set of documents possible concepts and relationships can be identified by applying document clustering or similarity measure based on semantics [2]. Latent Semantic Indexing is based on the concept of finding similarity by considering the words with similar meanings [2]. In semi-supervised and active learning methods, content categories are assigned to uncategorized documents from a large document collection. The large document collection can be news data from the web .Manually labeling the documents is costly and slow . Initially some labeled instances can be considered and then the new documents can be labeled, considering initial labeled instances.

In supervised learning a set of predefined topic categories are provided well in advance and the new document is classified according to the predefined classes. The classes can be concepts or relations in ontology. Ontology updating is needed as the underlying data and corresponding structures change with time. Stream mining addresses the issue of rapidly changing data and updating of ontology accordingly.

2.2 Semantic Annotations

In semantic annotations, the machine understandable data is added with the web resources. Adding semantic metadata with web resources is one of the important task in semantic annotation. The problem of automating annotations is one of the significant challenges in semantic web [3] . It is the process of combining semantic model and natural language processing together. Information Extraction (IE) is used to analyse the natural language and link the ontology concepts with the documents. There is a significant role of Human Language Technology (HLT) in the development of semantic metadata. Ontology is the formal knowledge representation about a domain, while Human Language Technology involves analysis, mining, and production of natural language. Human language Technology can be used to bring together the natural language and formal knowledge representation of semantic web.

Information Extraction is a technology based on analyzing natural language text to get the required information. Information Retrieval finds the relevant text documents according to user search query. Information Extraction is focused on finding the specific information from the text after analyzing it while Information retrieval finds relevant text documents. IE systems are knowledge intensive and difficult to build . IE is used to find out entities, relations between entities, entity reference, and events associated with the entities.

In Information Extraction metadata can be generated from the information discovered in the documents. This extracted information can be linked with the predefined ontology. With the linked information data is represented semantically in terms of ontology. in order to provide semantics and connectivity to the web. Maintaining semantic metadata about the web pages of different domains is another issue in managing semantic annotations.

2.3 CLIE approach

Natural language is easy for communication, but it has large degree of ambiguity. so it is difficult to process automatically.Machine can extract limited amount of information from natural language , formal data which is in terms of Ontology is rigidly structured , but it is easy for processing by machines. On the other side formal representation is difficult and unnatural for the people to use. So it's necessary to bridge the gap between natural language and formal language [4].

A Controlled language is a subset of natural language, which is usually less ambiguous than the complete language . If consist of certain vocabulary terms and grammar rules which are relevant to the specific task .

The limited number of allowed syntactic sentence structure, make the language easier to learn . It is easier to use than understanding OWL , RDF or SQL[10]..

The language analysis is done by applying information extraction , it includes English tokenize, part of speech tagger and finite state transducers , based on pattern matching linage. Transducers are used to search the pattern over annotations . it identifies tithe noun patterns which are likely to be classes , instances and other ontological objects . They also look for the specific patterns to extract the information.

CLIE can be used to create a new ontology or to add information to an existing one.

3. PROPOSED APPROACH

The idea behind proposed approach is to use the different methods to automate the process of ontology building. The aim is to start with the domain otology and to populate it automatically as and when require. The proposed methodology is divided into two phases.

1. Creation of the domain ontology (seed ontology) about the specific domain

2. To populate the knowledge base by combining the various approaches like information extraction, metadata generation

In this approach various tools and techniques can be used in order to automate the process of ontology learning.

3.1 Creation of Domain Ontology

Domain ontology can be created manually by considering the existing data related to a particular domain. Knowledge bases can be created by extracting relevant instances from information [9] . This information can be taken from various resources like web pages,word documents, or from any other forms of data representation. Control language information extraction (CLIE) is one of the approach to get the information about the required domain. Protégé framework can be used to design the starting ontology, which can be treated as seed ontology.

3.2. Populating the existing Domain ontology with ontology learning

Various semiautomatic ontology learning techniques can be applied to get the relevant information from the web resources. Various techniques of machine learning and natural language processing can be applied to automate the process of ontology building. Semantic annotation can be automated by adding machine understandable data with the web resources.

Preparation of labeled corpora for training learner's model can be used as one of the approach [4]. Different machine learning approaches like supervised, semi- supervised can be applied to automate the process of semantic annotations. Information extraction can be used to analyze natural language and link ontology concepts with the web documents .

4. CONCLUSION

Semantic web technologies offer a new approach to manage information and processes by adding meaning to the web data. Ontology learning is one of the phases for managing information in semantic web. In order to build the semantic web system, it is essential to identify the need of the system, data needed in the system, relationship of the data, and handlinginstances. Manual methods of knowledge management with semantic web system are time consuming and expensive. From this point of view, a methodology is suggested to automate the different tasks, like information extraction from web documents, adding semantic metadata with web data and liking it with Ontology . Combining machine learning techniques and natural language processing techniques can be useful to build the Ontology in a semi-automatic way.

5. REFERENCES

[1] hard Jesus Gil Herrera , Maria Jose Martin- Bautista "A novel Process-based KMS success framework empowered by Ontology learning technology " Elsevier journal of Engineering applications of artificial intelligence Vol 45,295-312 , 2015

[2] Carlos Vicient, David Sanchez, Antonio Moreno, " An automatic approach for ontology-based feature extraction from heterogeneous texual resources ", Elsevier journal of Engineering applications of artificial intelligence Vol 26, 1092-1106,2013.

[3] EfstrationsKontopoulos , Christos Berberidis , TheologosDergiades ,Nick Bassiliades, " Ontology – based sentiment analysis of twitter posts " Elsevier journalExpert Systems with Applications Vol 40 ,4065-4074,2013

[4] HamedHassanzadeh and Mohammed Reza Keyvanpour ," A Machine Learning Based Analytical Framework for Semantic Annotation Requirement " International Journal of web & semantic Technology (IJWest) Vol.2, No.2, April 2011 .

[5] A. Gyrard, "A machine-to-machine architecture to merge semantic sensor measurements," in Proceedings of the 22nd international conference on World Wide Web companion. International World Wide Web Conferences Steering Committee , pp. 371–376, 2014.

[6] A. Splendiani*et al.*, "Biomedical semantics in the Semantic Web," *Journal of Biomedical Semantics*, vol. 2, no. Suppl 1, pp. S1, 2011.

[7] A. Ruttenberg*et al.*, "Advancing translational research with the Semantic Web," *BMC Bioinformatics*, vol. 8 Suppl 3, pp. S2, 2007.

[8] Hevner, A.R., March, S.T., Park, J., and Ram, S. "Design Science in Information Systems Research," *MIS quarterly* (28:1), pp 75-105. 2004.

[9] Wilson Wong, Wei Liu , Mohammed Bennamoun, " Ontology Learning from Text : A look back and into the Future",ACM Computing Surveys, Vol 44, No 4, Article 20,Aug 2012

[10] Ian Horrocks , "DAML+OIL : A Description Logic for the semantic Web " , Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 2002.

[11] Gomez-Perez A, Fernandez –Lopez M, Corcho O. " Ontological Engineering, Advanced Information and Knowledge Processing, Springer,2003