Learning Concepts and Relations for Incremental Ontology Learning

S. Thenmalar, B. Sathiya and Dr. T.V. Geetha

1,2 Research Scholar, Dept of CSE, Anna University, Chennai, India.
3 Professor, Dept of CSE, Anna University, Chennai, India.

ABSTRACT

Ontologies have been well-known to be the advanced knowledge representation model providing the key for the success of semantic web. The domain ontology is widely used in various applications like Natural Language Processing, Information Retrieval and Information Extraction. However, the manual construction and updation of ontologies with up-to-date information is expensive and requires domain experts. Hence in this paper, we propose to update or enrich the existing ontology with new concepts and taxonomical relations by Incremental Ontology Learning process using text mining approach. The incremental process is automated by learning and populating the elements of ontology (concepts, taxonomical relations and instances) from the given documents. The enriched ontology has been evaluated with ontology profiling metrics and results have proved the effectiveness of our methodology.

INTRODUCTION

Ontologies are the rich semantic knowledge representation of a real world domain. “Ontology is defined as a formal and explicit specification of a shared conceptualization of a domain” (Ushold and Gruninger, 1996; Thomas, 1995; Parekh et al., 2004; Studer et al., 1998). Domain ontology conveys vocabulary of concepts and their relationships for that given domain, thus describing the domain semantics. The reusability of the domain knowledge is the added benefit from the use of ontologies (Noy and McGuinness, 2001). The development of the Semantic Web has reached an additional stage in the ontology research field. The Semantic Web provides the meaning for the given information enabling the support of the users (Parekh et al., 2004) and (Berners-Lee et al., 2001). Ontologies provide support and form the important component for the success of semantic web. These ontology based domain models are shared and are understandable to humans and machines. The significant benefits of ontologies in the Semantic Web world are knowledge sharing, ontology based reasoning, information integration and interoperability (Parekh et al., 2004). Thus, evolving domain specific ontologies is necessary for semantic web.

However, manual construction of ontology is a tedious task and requires domain experts. Ontology learning is a methodology of ontology building from external knowledge sources based on the text mining and machine learning techniques. The approach of Ontology Learning uses several methods of knowledge acquisition from structured (database), semi-structured (knowledge base) and unstructured data sources (texts). The important tasks for ontology learning include ontology evolution, reasoning and evaluation (Steffen and Studer, 2010). The dynamic process in ontology evolution (Haase and Stojanovic, 2005) and (M’adche, 2003) is essential for ontology learning. The ontologies need to be updated along with the domain knowledge which evolves over time. The reuse of existing ontology by updating it with incremental ontology learning process would be more effective than relearning the entire ontology from scratch (Zhou 2007). Our main objective is to explore the ontology learning process using text mining approach in an incremental manner.

In this paper, we explore to learn new concepts, instances and taxonomical relations by the Incremental Ontology Learning process. For the given set of documents, the concepts are identified through the process of clustering of terms and mapping with UNL KB. The taxonomical relations and the position where the concept need to be included in the existing ontology are identified through the UNL Ontology. Finally, the new
concepts, instances and their relations are updated to the existing domain ontology resulting in an enriched ontology.

This paper is organized as follows: Section 2 presents the related work to ontology learning approaches from the Web and the Incremental Ontology learning process. In section 3, describes the proposed process of incremental Ontology learning approach. Section 4 discusses about the evaluation carried out. Finally, we conclude and give some future perspectives for this research work.

1. Related Work:

Agirre et al., (2000) proposed the enrichment of concepts for large ontologies (Word Wide Web (WWW), The authors overcome the issues in WordNet in which the ‘absence of links among concepts’ is identified by constructing topic signatures and ‘proliferation of senses’ is resolved by clustering the concepts. For each word, this method uses the WWW or the Internet for constructing topic signatures.

Parekh, V. and Gwo, J (2004) proposed the text-mining approach to evaluate and enrich domain ontologies using glossaries or dictionaries and contextual information in the text. The domain expert or the knowledge engineer is required to evaluate and update the concepts in the ontology. The ontology learning framework (Fortuna et al., 2006) which uses supervised method (active learning) for concept discovery along with updating new instances to ontology based on users comments. In an incremental ontology learning and population framework (Loos and Schwarten, 2008) new learned instances of the ontological concepts are stored in a separate database rather than integrating them directly into the system’s knowledge base. The incorrect knowledge is stored aside. A combination of using ontologies and textual content was presented by Mustapha et al., (2009). This framework describes the integration of semantic search approaches and ontology learning from web documents to enable the engineering of web ontology and semantic indexing of web documents by means of case based reasoning. For enriching concepts of the ontology using WWW and corpus (Gharib et al., 2012), the semantics for the target word is identified through the WordNet. The semantically similar Fine-Grained senses are merged to yield Coarse-Grained sense. The query for each Coarse-Grained sense of the word is created. Then searches for the related documents in WWW and Corpus, dividing the retrieved documents into collections (one collection per word sense). Each collection of words along with frequencies is determined to create the topic signature from words having distinctive frequency. The need for domain expert or glossaries of terms from the WordNet is required for the learning new concepts to enrich the Ontology. However, we explore the concepts and relations with the use of external source( UNL KB and UNL Ontology) which provides semantically richer representation of concepts.

2. Methodology - Process Of Incremental Ontology Learning:

The overall process of the proposed incremental ontology learning is shown in Fig. 1. First, the given documents are pre-processed which includes stop words removal and finding out the occurrence of words in the documents. The documents are represented as Bag-of-words. The similarity measures are used to identify documents having similar terms and the documents are clustered using K-Means algorithm. The centroids obtained from the K-Means algorithm of the clustered documents represent the set-of-terms. The repetition of terms across the centroids is eliminated. The centroid terms are mapped to the UNL Knowledge Base (UNL KB) for concept identification. The UNL KB is a semantic representation of concepts (UW-Universal Words and its SC-Semantic Constraint) with the binary related relations. The UW and the SC are extracted for those centroid terms. The UW word would form the concept. We determine the taxonomical relations for those obtained concepts using the UNL Ontology. The UNL Ontology is a semantic and hierarchical representation of concepts. The existing domain ontology is updated using the identified concepts and relations. The document processing and the clustering of terms are discussed as follows.

2.1 Document Processing and Clustering Of Terms:

The given documents are pre-processed with the removal of stopwords. We identify the importance of words that occur in a document and across the document based on the TF-IDF measure (Chung et al., 2006). The frequently occurring words in the document and the document length describes the TF measure and is given as

\[
\text{tf}(w) = \frac{f_w}{n}
\]

where \(f_w\) denotes the number of times the word occurred in the document; \(n\) denotes the number of words in the document. The importance of word occurring across the document describes the IDF and is given as

\[
\text{idf}(w) = \log\left(\frac{\text{Total number of documents}}{\text{Number of documents where the word w occurs}}\right)
\]

The importance of words in the document is based on the TF-IDF measure and is given as

\[
\text{tfidf}(w) = \text{tf}(w) \times \text{idf}(w)
\]

The documents are represented as the bag-of-words with the calculated tfidf value and is given as

\[
d_i = ((w_1, \text{tfidf}_{w_1}), (w_2, \text{tfidf}_{w_2}), ..., (w_n, \text{tfidf}_{w_n})) \text{ where } i=1,2,...,n \text{ and } n \text{ is the total number of words in a document.}
\]

The similarity between the documents is calculated by the cosine similarity measure and is given as

\[
\text{sim}(d_i, d_j) = \frac{(d_i \cdot d_j)}{\|d_i\| \|d_j\|}
\]
Using the similarity value we cluster the documents using K-Means algorithm (Jain et al., 1999). The K-Means algorithm clusters the documents based on the occurring of similar words. The iteration of the algorithm continues till the centroids changes. The centroid describes the set of words obtained for a cluster (Fortuna et al., 2006). The words that occur in a centroid may be available in the other centroids; this type of redundant word is eliminated across the centroids. The words or the terms are the essence of concepts in the ontology. The terms in the centroids are to be represented as concepts.

![Diagram of Incremental Ontology Learning](image)

**Fig. 1:** Process of Incremental Ontology Learning.

### 2.2 Concept and Taxonomy Relation Identification:

We make use of the Knowledge Bases (UNL KB and UNL Ontology) for concepts and relation identification. The use of external sources provides additional conceptual information and richer representation to the terms. The terms in the centroids are mapped to the UNL KB which comprises of the UW and SC. For example the term city when mapped to the UNL KB the semantic representation “city (icl>place)” is obtained where the city represents the UW word and the icl>place represents the semantic constraint. We extract this information for each term in the centroid of each cluster. The UW word would represent the concept name. Then we map the concept with the UNL Ontology to obtain the hierarchical representation for the concept. For the given example the parent concept for the city would be of ‘Place’ and thus determines the ISA relationship for the concept city. The concept city would be added as a child concept to the concept ‘Place’.

### 2.3 Updating the existing ontology:

The obtained concepts are to be updated in the existing domain ontology. The obtained concepts that are not available in the existing ontology are considered thus providing the additional concepts to the existing domain ontology. With the SC information for the concept and the UNL Ontology we determine where to update the concept in the existing ontology. The parent concept obtained from the UNL ontology and the SC obtained from the UNL KB should be identical for semantic consistency. And the additional constraint is that the maximum common subsumers (ancestors) path for the parent concept till the root node in the existing domain ontology and in the UNL Ontology should be identical. The new concept is added as a child concept to the parent concept identified from the UNL ontology in the existing domain ontology. The concepts updated at the leaf level of the hierarchy in the existing ontology are the instances of the concepts. Thus the existing ontology is enriched with the new concepts, instances and the taxonomical relations.

### RESULTS AND DISCUSSION

This section discusses in detail the experiments conducted to prove the effectiveness of the proposed methodology to enrich the ontology in terms of concepts, taxonomical relations and population of the concepts. The ontology used for the evaluation belongs to a tourism domain (Thenmalar and Geetha, 2014) which has been manually constructed and checked for consistency using Protégé tool comprises of 600 tourism specific concepts and subconcepts. The number of object properties that describes relations are 24 out of which 15 are relations specific to tourism domain. Depth of this tourism ontology is 7 with 13168 instances. The type of relations available in the tourism ontology are “locatedIn hasAccomodation, nearBy, surroundedBy, hasActivity, provides, includes, stateOf, districtOf, headquartersOf, capitalOf, boundedBy, covers, hasContact, hasPart, hasRating, accessibleBy, offeredAt and reservedFor” (Thenmalar and Geetha, 2014). Jena API in java is used to enrich the existing tourism ontology with new concepts, relations and instances. The input to the Incremental Ontology Learning process also includes 50,000 tourism specific documents which were obtained by the focused crawling process (Thenmalar and Geetha, 2011).

Metrics like Number of concepts, Inheritance Richness (Samir et al., 2005), Average Depth of the ontology (Cruz et al., 2012), Class Richness (Samir et al., 2005) and Average Population (Samir et al., 2005) which are used in literature (Samir et al., 2005; Cruz et al., 2012) for ontology profiling are used for evaluation. The
Inheritance Richness (Samir et al., 2005) describes the IS-A relationship richness of the ontology. It is defined as the average number of subclasses per class and formally represented as follows.

\[
\text{Inheritance Richness} = \frac{\sum |SC(c_i)|}{|C|}
\]

where \(|SC(c_i)|\) represents the number of subclasses for the class \(c_i\) and \(|C|\) represents the total number of classes in the ontology. IR values will be a real number where greater value indicates a horizontal ontology and lesser value indicates a vertical ontology. Horizontal ontology represents a general wide range knowledge whereas the vertical ontology represents a detailed in depth knowledge about a particular domain of interest.

Average Depth (Cruz et al., 2012) is a metric to measure the average specificity of each class. Average Depth values will be a real number where more the Average Depth, more detailed/informative is the ontology. It is measured using the depth \((D)\) of each class in ontology and is formally defined as follows.

\[
\text{Average Depth} = \frac{\sum \text{Depth}(c_i)}{|C|}
\]

where \(\text{Depth}(c_i)\) represents the depth of the class \(c_i\).

Class Richness (Samir et al., 2005) metric is used to determine the average number of classes which have instances. It is formally defined as the ratio of the number of classes for which instance exist to the total number of classes in the ontology \(|C|\).

\[
\text{Class Richness} = \frac{|\text{Class with Instances}|}{|C|}
\]

The Class Richness value is a real number ranging from 0 to 1. The value 1 indicates that ontology possesses a rich instance or knowledge base and value 0 indicates a poor knowledge base.

Average population (Samir et al., 2005) metric is used to measure the average number of instances per class. It is formally defined as the ratio of number of instances to the number of classes in the ontology.

\[
\text{Average Population} = \frac{|\text{No. of Instances}|}{|C|}
\]

The Average population value is a real number where a larger value indicates that ontology possesses a rich instance or knowledge base and lesser value indicates a poor knowledge base.

The experimental results based on the above discussed metrics are listed in Table 1. As depicted by the results, the number of concepts is increased by 45.6% after the ontology was enriched by our methodology. The structural richness of the ontology is also increased by our methodology as shown by the metrics like Inheritance Richness and Average Depth of the ontology which is increased by 22.9% and 47.9%. The instance richness of the ontology is also increased by our methodology as shown by the metrics like Class Richness and Average Population of the ontology which is increased by 14.1% and 8.2%. The experimental results depict that our proposed measure enriched the ontology in terms of concepts, structural and instances.

<table>
<thead>
<tr>
<th>Table 1: Ontology Profiling Metrics.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metrics</td>
</tr>
<tr>
<td>-------------------------------</td>
</tr>
<tr>
<td>Number of concepts</td>
</tr>
<tr>
<td>Inheritance Richness</td>
</tr>
<tr>
<td>Average Depth of the ontology</td>
</tr>
<tr>
<td>Class Richness</td>
</tr>
<tr>
<td>Average Population</td>
</tr>
</tbody>
</table>

Conclusion:

In this paper, we have discussed our incremental ontology learning process for enriching the existing domain ontology through new concepts, instances and taxonomical relations. Concepts and instances are identified using semantically rich UNL KB. Similarly, taxonomical relations and position in the existing ontology where the new concepts and instances to be updated are obtained using UNL Ontology. Experimental results on the ontology profiling metrics proved that our incremental ontology learning process is effective in terms of new concepts, instances and taxonomical relation extraction.

This method limits to concept, instances and taxonomical relation identification and can further be extent to incorporate the non-taxonomical relation identification and generating axioms to the ontology. Future work also includes using other knowledge base such as Linked Open Data (LOD) in addition to UNL KB and UNL Ontology for incremental ontology learning process.
REFERENCES

Agirre, E., O. Ansa, E. Hovy and D. Martínez, 2000. Enriching Very Large Ontologies Using the WWW. In the Workshop on Ontology Construction of the European Conference of A.I., ECAI.


Viral Parekh, Jack Gwo and Tim Finin, 2004. Mining Domain Specific Texts and Glossaries to Evaluate and Enrich Domain Ontologies. in IKE, 533-540.