Automatic Ontology Learning from Multiple Knowledge Sources of Text

B Sathiya, College of Engineering, Guindy, Anna University, Chennai, India
T.V. Geetha, College of Engineering, Guindy, Anna University, Chennai, India

ABSTRACT

The prime textual sources used for ontology learning are a domain corpus and dynamic large text from web pages. The first source is limited and possibly outdated, while the second is uncertain. To overcome these shortcomings, a novel ontology learning methodology is proposed to utilize the different sources of text such as a corpus, web pages and the massive probabilistic knowledge base, Probase, for an effective automated construction of ontology. Specifically, to discover taxonomical relations among the concept of the ontology, a new web page based two-level semantic query formation methodology using the lexical syntactic patterns (LSP) and a novel scoring measure: Fitness built on Probase are proposed. Also, a syntactic and statistical measure called COS (Co-occurrence Strength) scoring, and Domain and Range-NTRD (Non-Taxonomical Relation Discovery) algorithms are proposed to accurately identify non-taxonomical relations(NTR) among concepts, using evidence from the corpus and web pages.

KEYWORDS
Knowledge Engineering, Pattern Analysis, Semantic Web, Ontology Learning, Probase, Statistical measure

1. INTRODUCTION

As the requirements and use of the web increase, the quantum of available, assorted textual information has proliferated. This calls for a representation that semantically consolidates and organizes information in a conceptual hierarchy so as to store, retrieve and infer knowledge from a range of sources. An ontology is the best candidate for this sort of representation. According to Gruber (Gruber, 1993), “Ontologies are effectively formal and explicit specifications in the form of concepts and relations of shared conceptualizations.” Hence, the prime components of an ontology are concepts and their taxonomical and non-taxonomical relations. Taxonomical relations (hyponyms and hypernym) between concepts are to be discovered to construct a concept hierarchy. A hyponym is defined as a concept
with a specific meaning with respect to its super-concept. In contrast, a hypernym is defined as a concept with a generic meaning with respect to its sub-concept. However, non-taxonomical relations are domain-specific, because two concepts are related based on a domain-specific relation.

Taxonomical relations can be identified using different techniques such as clustering, syntactic/dependency structure analysis, and LSP. According to (Wu et al., 2012), LSP are one of the most prominent and valuable technique to discover taxonomical relations. Non-taxonomical relations can be discovered using different techniques such as syntactic structure analysis, LSP, semantic templates and association rule mining. LSP and semantic templates need a set of predefined, domain-specific relations. Further, association rule mining is unable to incorporate the huge masses of evidence available on the web. Therefore, the proposed system uses a lexical-syntactic pattern technique and a syntactic structure analysis to discover taxonomical and non-taxonomical relations respectively.

The construction of an ontology from unstructured textual sources, termed ontology learning from text, can be processed in three ways: manual, semi-automatic and automatic. In the manual method, an ontology is constructed from the scratch by domain experts and knowledge engineers using the most painstaking procedures (Maedche, 2013). In the semi-automatic method, the domain experts and trained users utilize semi-automatic prototype (Kim & Storey, 2011) and tool (Dahlem, 2011) to construct the ontology. However, both these methods are time consuming and require domain experts. Consequently, the automatic method of ontology learning is becoming a major trending and challenging task in this area of research.

In the automatic method, the following types of information are used – a static, limited and possibly outdated set of texts called a corpus, and/or the vast, dynamic and recent collection of information from web pages retrieved, based on queries to search engine. Initially, in the process of ontology learning, concepts from these textual sources are extracted and taxonomical and non-taxonomical relations among them are discovered. The quality of the constructed ontology chiefly depends on the completeness and correctness of the textual information contained (Rios-Alvarado et al., 2013). However, the corpora may lack completeness owing to the static and limited sources of text and web pages that could lack appropriateness themselves, given the uncertainty prevailing on the web.

To overcome the aforesaid problems with information sources, we have used multi-sources of information such as a domain-specific corpus, web pages and a rich, universal and probabilistic taxonomy called Probase (Wu et al., 2012). Although a handful of general-purpose taxonomies/ontologies do exist (Table 1), they have limited concept space (Wu et al., 2012) and hence Probase has been chosen. Probase consists of 2.7 million concepts (also multi-word terms) obtained from 1.8 billion web pages, and is constructed to handle the inconsistency, ambiguity and uncertainty of knowledge. An iterative learning algorithm was used in Probase to extract hypernym-hyponym pairs from the web and a probabilistic taxonomy construction algorithm was utilized to construct taxonmot from these pairs.

<table>
<thead>
<tr>
<th>Existing Taxonomies</th>
<th>Number of Concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freebase</td>
<td>1,450</td>
</tr>
<tr>
<td>WordNet</td>
<td>25,229</td>
</tr>
<tr>
<td>WikiTaxonomy</td>
<td>111,654</td>
</tr>
<tr>
<td>YAGO</td>
<td>352,297</td>
</tr>
<tr>
<td>DBPedia</td>
<td>259</td>
</tr>
<tr>
<td>ResearchCyc</td>
<td>~120,000</td>
</tr>
<tr>
<td>NELL</td>
<td>123</td>
</tr>
<tr>
<td>Probase (Wu et al., 2012)</td>
<td>2,653,872</td>
</tr>
</tbody>
</table>

Table 1. General purpose taxonomies (Wu et al., 2012)
Related Content

On Some Algebraic Structures of Type 2 Fuzzy Multisets
www.igi-global.com/article/on-some-algebraic-structures-of-type-2-fuzzy-multisets/179318?camid=4v1a

Issues for the Evaluation of Ambient Displays
www.igi-global.com/article/issues-evaluation-ambient-displays/3880?camid=4v1a

Overview of Type-2 Fuzzy Logic Systems
www.igi-global.com/article/overview-type-fuzzy-logic-systems/70754?camid=4v1a
Neutrosophic Sets and Logic
www.igi-global.com/chapter/neutrosophic-sets-and-logic/171900?camid=4v1a