Acquiring Semantic Sibling Associations from Web Documents

Marko Brunzel, German Research Center for Artificial Intelligence (DFKI GmbH), Germany
Myra Spiliopoulou, Otto-von-Guericke-Universität Magdeburg, Germany

ABSTRACT

The automated discovery of relationships among terms contributes to the automation of the ontology engineering process and allows for sophisticated query expansion in information retrieval. While there are many findings on the identification of direct hierarchical relations among concepts, less attention has been paid on the discovery sibling terms. These are terms that share a common, a priori unknown parent such as co-hyponyms and co-meronyms. In this study, we present our results on the discovery of pairs or groups of sibling terms with XTREEM-SA (Xhtml TREE mining for sibling associations), an algorithm that extracts semantics from Web documents. While conventional methods process an appropriately prepared corpus, XTREEM-SA takes as input an arbitrary collection of Web documents on a given topic and finds sibling relations between terms in this corpus. It is thus independent of domain and language, does not require linguistic preprocessing, and does not rely on syntactic or other rules on text formation. We describe XTREEM-SA and evaluate it toward two reference ontologies. In this context, we also elaborate on the challenges of evaluating semantics extracted from the Web against handcrafted ontologies of high quality but possibly low coverage.

Keywords: co-hyponymy; ontology learning; mining semi-structured data; semantic associations; sibling relations

INTRODUCTION

Applications relying on text categorization, analysis, retrieval, and understanding benefit largely from background knowledge on semantic relations among terms. For example, semantically related terms are used for query expansion in IR and for context disambiguation in text understanding. While there is much research on the discovery of subordination relations (specialization vs. generalization, aggregation), there are less studies on the discovery of coordination relations such as co-hyponyms and co-meronyms.

In this article, we present our research and experimental results on the discovery of sibling terms by incorporating association rules to the XTREEM (Xhtml TREE Mining) approach, extending our work of (Brunzel & Spiliopoulou,
2006b) for the identification of sibling pairs. We denote the method as “XTREEM-SA” (XTREEM for sibling associations); it encompasses a frequent itemset discovery for binary and for n-ary associations among terms.

In ontology engineering, there are different approaches for the discovery of semantic relations. Many use as input unstructured plain or semi-structured text after converting it to plain text (Buitelaar, Cimiano, & Magnini, 2005; Faure & Nedellec, 1999; Maedche & Staab, 2001). There are also approaches that exploit resources like dictionaries, glossaries, or database schemata (Kashyap, 1999; Stojanovic, Stojanovic, & Volz, 2002), but are limited to the rare case when such resources are available. In contrast, XTREEM uses a set of documents from the Web as the result of one or more domain queries: This set of documents forms an ad hoc collection for the discovery of sibling terms. This approach has several advantages: There is no demand for linguistic resources, nor for a prepared corpus. There is no dependence on language nor domain, except of the demand for a domain-specific query to retrieve the documents from the Web.

The article is organized as follows: We discuss related work in the next section. Then we present briefly the “XTREEM” approach for the establishment of a collection of Web documents and the extraction of path structures from them. XTREEM exploits the tendency of Web authors in using similar markup conventions for the same type of content, thus delivering a good basis for the identification of similar terms. We then describe XTREEMSA, which uses XTREEM as component and applies frequent itemset mining upon its output. Then we present our evaluation methodology, followed by our experiments. The last section concludes our study.

**RELATED WORK**

The related work can be partitioned into the research topics (1) ontology learning, also covering methods for the acquisition of semantics (2) methods that use XML structure, and (3) the XTREEM approaches, our own prior work.

**Ontology Learning Methods**

Ontology learning as a research field emerged with the advent of the Semantic Web and its lack of existing ontologies. Ontology learning refers to those methods, which tackle the problem of acquiring semantics from various data sources. Early work on ontology learning was mainly performed by Maedche (Maedche, 2002; Maedche & Staab, 2000, 2001). Comprehensive overviews on ontology learning are given by Cimiano (2006), Gomez-Perez & Manzano-Macho, 2003; Shamsfard & Barforoush, 2003). On one side, ontology learning can be divided into subfields according to the used input data. One can therefore separate between ontology learning from unstructured text (Buitelaar et al., 2005), ontology learning from structure (Kashyap, 1999; Stojanovic et al., 2002), and ontology learning from semi-structure.

Most of the existing ontology learning methods are used to perform ontology learning from text. For an overview on this topic, see Buitelaar et al. (2005). Even if those methods are processing semi-structured data (e.g., HTML Web documents), they are removing the markup and process plain text.

Upon plain text, *Hearst patterns* (Hearst, 1992) are used to find relations among terms in text collections by means of matching patterns. Also *co-hyponym relations* can be found with this approach. However, the disadvantage is that such patterns are rare and the coverage is low, even on big document collections. Cimiano et al. (2004, 2005) discover (co-)hyponym relations by finding and analyzing examples of Hearst patterns on the WWW. Cimiano et al. did not exploit the markup of Web documents, Web documents are treated like regular plain text.

On the other side, there are those methods (Kashyap, 1999; Stojanovic et al., 2002) where highly structured data such as dictionaries, database schemas, or UML schemas are used as input. Those approaches have the disadvantage that such highly structured data is rare.

A rather recent subfield of ontology learning uses semi-structured documents as input data (Nayak & Zaki, 2006). There the added value, which is given by nowadays omnipres-
Query Recommendations for OLAP Discovery-Driven Analysis
www.igi-global.com/article/query-recommendations-olap-discovery-driven/53037?camid=4v1a

Web Mining for the Integration of Data Mining with Business Intelligence in Web-Based Decision Support Systems
Marcos Aurélio Domingues, Alípio Mário Jorge, Carlos Soares and Solange Oliveira Rezende (2015). *Integration of Data Mining in Business Intelligence Systems* (pp. 120-142).
www.igi-global.com/chapter/web-mining-for-the-integration-of-data-mining-with-business-intelligence-in-web-based-decision-support-systems/116812?camid=4v1a