Evidence Sources, Methods and Use Cases for Learning Lightweight Domain Ontologies

Albert Weichselbraun,* Gerhard Wohlgenannt* and Arno Scharl†

Abstract

By providing interoperability and shared meaning across actors and domains, lightweight domain ontologies are a cornerstone technology of the Semantic Web. This chapter investigates evidence sources for ontology learning and describes a generic and extensible approach to ontology learning that combines such evidence sources to extract domain concepts, identify relations between the ontology’s concepts, and detect relation labels automatically. An implementation illustrates the presented ontology learning and relation labeling framework and serves as the basis for discussing possible pitfalls in ontology learning. Afterwards, three use cases demonstrate the usefulness of the presented framework and its application to real-world problems.

1 Introduction

Ontologies, which are commonly defined as explicit specifications of shared conceptualizations (Gruber, 1995), provide a reusable domain model which allows for many applications in the areas of knowledge engineering, natural language processing, e-commerce, intelligent information integration, bio-informatics etc.

Not all ontologies share the same amount of formal explicitness (Corcho, 2006), nor do they include all the components that can be expressed in a formal language, such as concept taxonomies and various types of formal axioms. Therefore, ontology research distinguishes lightweight and heavyweight ontologies (Studer et al., 1998). The creation of such conceptualizations for non-trivial domains is an expensive and cumbersome task, which requires highly specialized human effort (Cimiano, 2006). Furthermore the evolution of domains results in a constant need for refinement of domain ontologies to ensure their usefulness.

Automated approaches to learning ontologies from existing data aim at improving the productivity of ontology engineers. Buitelaar et al. (2005) organize the tasks in ontology learning in a set of layers. Especially in ontology learning from text, lexical entries $L_C$ are needed to link single words or phrases to

1Institute for Information Business, Vienna University of Economics and Business, Vienna, Austria, e-mail: {aweichse,wohlg}@ai.wu.ac.at
2Department of New Media Technology, MODUL University Vienna, Austria, e-mail: scharl@modul.ac.at
concepts $C$. Synonym extraction helps to connect similar terms to a concept. Taxonomies $H^C$ provide the ontology's backbone. Non-taxonomic relations $R$ supply arbitrary links between concepts. Finally, axioms are defined or acquired to derive additional facts.

Data sources for ontology learning typically include unstructured, semi-structured and structured data (Cimiano, 2006). Ontology learning from structured data consumes information sources such as database schemas or existing ontologies – it is also called lifting as it lifts or maps parts of existing schemas to new logical definitions. Since most of the available data is unstructured or semi-structured, a major research focus over the last two decades has been the extraction of domain models from natural language text through a variety of methods. Cimiano (2006) presents an extensive overview of ontology learning methods from unstructured data. Many of the methods involve corpus statistics, such as co-occurrence analysis (Liu et al., 2005), association rules (Maedche et al., 2002), latent semantic analysis based techniques for the detection of synonyms and concepts (Landauer & Dumais, 1997), or the application of kernel methods for example to classify semantic relations (Giuliano et al., 2007).

A lot of work in the field, especially for tasks that involve term clustering, exploits Harris’ distributional hypothesis (Harris, 1968), which states that terms or words are similar to the extent that they occur in syntactically similar contexts. Besides corpus statistics, many authors apply linguistic parsing and linguistic patterns in ontology learning. Building on the seminal work of Hearst (Hearst, 1992), patterns support taxonomy extraction (Liu et al., 2005), the detection of concepts and labeled relations in combination with the application of Web statistics (Sánchez-Alonso & García, 2006), or Web-scale extraction of unnamed relations (Etzioni et al., 2008).

The integration of Semantic Web resources has become quite popular in ontology learning in the recent years. In the presented modular and extensible framework, we use structured information to apply semantic constraints on learned ontological elements, for example in the task of detecting non-taxonomic relations where the system penalizes suggested relation label candidates conflicting with the constraints defined. Gracia et al. (2006) describe an unsupervised approach that dynamically uses online ontologies for word-sense disambiguation. d’Aquin, Motta, et al. (2008) provide the Scarlet service for discovering relations between two concepts by harvesting the Semantic Web. Similarly, Aleksovski et al. (2006) extract relations between terms in background knowledge. Alani (2006) proposes a method for ontology construction by cutting and pasting ontology modules from online ontologies.

Domain text often misses some of the terms important to a particular domain, since those terms and associated concepts are assumed to be common ground shared by the authors and readers of documents. Additional resources, such as collective intelligence in the form of folksonomies (Specia & Motta, 2007), social networking or micro blogging systems, as well as online ontologies are rich sources to augment knowledge expressed in textual resources. Some authors (Mika, 2007; Heymann & García-Molina, 2006; Tang et al., 2009; Schmitz, 2006) build ontologies solely based on information gathered from social sources.
The presented architecture uses data from social sources together with other evidence with the intention to capture the latest terminology of evolving domains (Angeletou et al., 2007) and to integrate background knowledge about the domain from external data sources.

The remainder of this chapter is structured as follows: Section 2 introduces the three evidence sources utilized in the presented ontology learning framework. Section 3 presents the major steps and methods applied in the ontology building process for (i) extracting terms, relations, and relation labels as well as (ii) applying ontological constraints. Section 4 demonstrates the potential of the ontology learning architecture by means of real-world use cases in three different domains (tourism, waste management, climate change). The chapter closes with an outlook and conclusions in Section 5.

2 Data Sources

Methods for ontology construction rely on evidences gathered from relevant data sources such as domain documents, online communities and ontology repositories. Generally speaking, one can distinguish between (i) in-corpus evidence sources which mostly rely on unstructured data such as domain relevant text and Web documents (Section 2.1) and (ii) external sources which provide an outside view of the domain by including social (Section 2.2) and structured (Section 2.3) data in the ontology learning process.

2.1 Unstructured Evidence Sources

From unstructured evidence sources (e.g., relevant Web documents), automated ontology learning systems can extract candidate terms by means of information extraction and text mining techniques - e.g., significant phrase detection, co-occurrence analysis and trigger phrases.

*Significant phrase detection* determines bi- and trigram terms in the domain corpus by comparing the number of a term’s observed occurrences to the number of expected occurrences under the hypothesis of independent terms using the log likelihood ratio (Hubmann-Haidvogel et al., 2009). *Co-occurrence analysis* locates these terms and unigrams in the domain corpus and compares their frequency in sentences and documents containing seed ontology concepts with their general distribution in the corpus. A chi-square test with Yates’ correction for continuity (Yates, 1934) suggests a ranked list of terms, which occur significantly more often with seed ontology concepts, for inclusion into the domain ontology. *Trigger Phrases* (Grefenstette & Hearst, 1992; Joho et al., 2004) yield concept candidates and relations by matching text fragments that indicate a particular relationship (e.g., hyponym, hypernym and synonym) between terms in the domain corpus.
2.2 Social Evidence Sources

Social evidence sources query Web 2.0 applications such as tagging systems, social networking and micro-blogging services to retrieve candidate concepts for the extended ontology based on a set of given seed ontology terms.

Delicious\(^1\) and Flickr\(^2\), for example, provide an API to retrieve the number of entities which have been labeled with a specific tag (= tag popularity) and to determine related tags. Technorati\(^3\) does not offer such an API. Therefore, we had to implement a method to compute related tags based on the tags in the top 100 blogs returned for a target tag. The same strategy has been applied to Twitter\(^4\).

Comparing tag popularities by applying similarity measures such as the dice coefficient or pointwise mutual information yields suggestions for relations between tags.

2.3 Structured Evidence Sources

Structured evidence sources include repositories such as DBpedia (Bizer et al., 2009), Freebase\(^5\) and OpenCyc\(^6\), which provide ontological data including concepts, relations and instance data; their integration is the goal of the Linking Open Data\(^7\) initiative. Several search engines such as Swoogle\(^8\) specialize in sharing ontologies via standardized formats, others like Sindice\(^9\) concentrate on providing triple-based instance data from RDF and microformats. Many engines offer both conceptual data as well as instances; e.g., Watson (d’Aquín, Sabou, et al., 2008), Falcons\(^10\), and SWSE\(^11\).

3 Method

Figure 1 outlines the process of constructing lightweight ontologies. In the initial step, domain experts identify seed terms or a seed ontology. The system then detects relations between these terms, and identifies labels for these relations. These steps are independent of each other and can be performed using different extension frameworks that use the process outlined in Figure 1 to learn concepts, relations and relation labels.

Evidences from unstructured, structured and social sources help identify possible candidates for integration in the domain ontology. Methods such as spread-

\(^1\)www.delicious.com
\(^2\)www.flickr.com
\(^3\)www.technorati.com
\(^4\)www.twitter.com
\(^5\)www.freebase.com
\(^6\)www.opencyc.org
\(^7\)esw.w3.org/topic/sweig/taskforces/communityprojects/linkingopendata
\(^8\)swoogle.umbc.edu
\(^9\)www.sindice.com
\(^10\)iws.seu.edu.cn/services/falcons
\(^11\)swse.deri.org
ing activation (Crestani, 1997) or the vector space model (Salton et al., 1975) integrate these evidences and provide a ranked list of candidates. Applying domain constraints on the collected data penalizes entries violating ontological constraints. Domain experts help to refine and optimize the ontology learning process by providing feedback on the suggested concepts, relations, and relation labels (Figure 1). The following section will outline each step of the learning process in more detail and describe our implementation of the proposed ontology building method, which comprises (i) the framework introduced in Liu et al. (2005) for term extraction and relation detection, and (ii) the relation labeling component presented in Weichselbraun et al. (2010), which applies constraints to ensure that its suggestions are consistent with the domain model.

3.1 Term Extraction

Figure 2 presents an implementation of the first two steps in the ontology construction process outlined above which follows Liu et al. (2005).

The ontology extension architecture assembles evidences from unstructured data sources such as Web pages, blogs and media archives. Plugins extract evidences such as co-occurring terms, Hearst patterns and WordNet relations from this data and forward them to the evidence integration component. Social sources such as Delicious, Flickr, Twitter and Technorati could be integrated in this step as well.
Figure 2: Ontology learning framework using spreading activation for evidence integration (Liu et al., 2005)

<table>
<thead>
<tr>
<th>seed term</th>
<th>evidence</th>
<th>candidate term</th>
</tr>
</thead>
<tbody>
<tr>
<td>climate change</td>
<td>w1:cooccurs</td>
<td>carbon dioxide</td>
</tr>
<tr>
<td>energy sources</td>
<td>w1:meronym</td>
<td>oil</td>
</tr>
<tr>
<td>energy</td>
<td>w1:hyperonym</td>
<td>renewable energy</td>
</tr>
<tr>
<td>climate change</td>
<td>w1:delicious</td>
<td>gas</td>
</tr>
</tbody>
</table>

Table 1: Example evidences collected by the ontology learning framework

The system then collects all evidences as RDF statements in a semantic network as illustrated in Table 1. Reification adds relation meta data such as significance values, number of occurrences and weights to the suggested concepts.

The left site of Figure 3 shows an example entry for the term climate change, which co-occurs with carbon dioxide with a significance of 12.982 according to a Chi-squared test with Yates correction.

Liu et al. (2005) use spreading activation to transform the data collected in the semantic network into a ranked list of candidate terms for integration in the domain ontology. Per evidence source heuristics translate evidences into spreading activation weights and build a spreading activation network which will be used for the ranking process. The subjects of the statements collected in the semantic network are transformed into sources, the objects into sinks and evidence type and annotations into the appropriate weights (Figure 3, for details see Liu et al. (2005)). Activating the source nodes yields activation energy levels for the collected evidences which correspond to their ranking resulting from the evidence integration step.

Angeletou et al. (2007) note that the integration of structured and social evidences introduces new and evolving vocabulary into the domain ontology. External sources also cause the problem of including unrelated terms, or terms that are irrelevant in the context of the ontology. Figure 4 provides such an
example. Terms which are connected with bold lines to other concepts have been determined by a social source (Delicious). Most of the included vocabulary such as methane, environment, greenhouse, etc. is intuitive, but the relations cooling → overclocking and ice → machine clearly introduce terms that are irrelevant in this particular context. One potential strategy to prevent the inclusion of such concepts is the use of a disambiguation process, which includes additional context terms for ambiguous seed terms. The importance of a proper selection of social and structured evidence sources should not be underestimated. For instance, including Flickr into the extension process of an ontology which focuses on abstract concepts would probably not be an excellent choice, although the impact of a single source might be reduced by combining multiple social and structured sources.

Another risk of external sources is that they might lead to shifts in the ontology’s focus. Therefore, it is extremely important to balance in-corpus sources and external sources and to include safeguards, such as rules which enforce a certain relationship between external and internal concepts, which ensure a proper focus of the extended ontology.

3.2 Relation Detection

The relation detection process (step 2 in Figure 1) takes concept pairs and populates a semantic network with evidences such as relation types suggested by certain patterns (Hearst, 1992; Joho et al., 2004), subsumption analysis (Sander-son & Croft, 1999) and grammatical relationships between the terms. It is even possible to use the semantic network from the concept detection step for this process.

Evidence specific transformation heuristics translate this data into spreading activation weights (Section 3.1). Subsequently, an iterative process activates new concepts and creates a relation to the concept with the semantically strongest relation (= the relation with the highest share of the activation energy from the new concept). Depending on the use case, specific preference relations
3.3 Relation Labeling

Figure 5 presents the relation labeling approach (step 3 in Figure 1) introduced in Weichselbraun et al. (2010) which follows the generic process illustrated in Figure 1.

Based on a set of candidate relations, which are formally described in a rela-
tion description ontology, the method starts extracting evidences from domain relevant documents which contain the subject and the object participating in the relation. The relation labeling prototype introduced in Weichselbraun et al. (2010) only considers verbs or verbs together with prepositions as evidences. Future versions might consider other part-of-speech tags in the evaluation as well.

A vector space model is used to integrate the data – every evidence collected corresponds to a position in the vector space model (Figure 5).

Applying the evidence collection process to known relations defined in the relation description ontology yields vector space representations (centroids) for those known relation labels. The label of newly acquired and therefore unlabeled relations is determined by choosing the label of the semantically closest centroid based on the vector space model with the cosine similarity measure.

### 3.4 Constraints

In a final step, the proposed process uses constraints to ensure the consistency of the generated ontology, and to refine the ranking of choices based on their conformance with these constraints.

For applying domain and range restrictions (as defined in the relation description ontology) to relation candidates, a concept grounding using an external

![Figure 5: Learning relation labels (Weichselbraun et al., 2010).](image-url)
ontology such as OpenCyc has proven to be beneficial as it allows constraints based on more general concepts such as organization, person, etc. Currently we verify domain, range and property restrictions by enforcing relation label suggestions which fulfill these constraints and penalizing elements violating them. The refined ranking is the base for deciding on the concepts, relations and relation labels to include in the domain ontology.

4 Use Cases

This chapter presents three real-world use cases which successfully applied the ontology learning framework introduced in this chapter. The use cases generate ontologies that reflect the knowledge contained in a given corpus. Since knowledge representation experts are not the primary target audience for the resulting structures, the ontologies only contain taxonomic relations indicated by arrows as well as an abstract “related” ($r$) type to indicate non-taxonomic relationships between terms.

4.1 Tourism Destinations

Dickinger et al. (2008) analyze news media coverage to acquire and structure tourism knowledge using ontology learning. They apply contextual filtering to differentiate between general and tourism-specific news media coverage.

Our ontology learning process extends an ontology of six seed concepts (the black terms in Figure 6) to a lightweight domain ontology which comprises 30 concepts, which were extracted from the input corpus (Figure 6).

Most of the terms and relations included by ontology learning are straightforward to interpret. Nevertheless, there are also a number of unexpected relations such as culture tourism $\rightarrow$ handcart, air travel $\rightarrow$ snowcam which were added due to a special coverage of certain Web pages (e.g. the CNN and USA Today coverage on “Mormon Hand Track” referring to a cultural tourism attraction). From the ontology engineering point of view, the inclusion of such relations is not necessarily a good thing and might be avoided by (i) using a larger input corpus which reduces the impact of singular events, or (ii) by adding additional evidence sources such as the ones suggested in Section 2.2 and Section 2.3.

4.2 Communication in Waste Management

Pollach et al. (2009) investigate the Internet coverage of solid waste management on media sites, corporate Web sites and on the Web pages of non-governmental organizations. The authors extend a small seed ontology with the ontology framework introduced by Liu et al. (2005) (Section 3.1) to investigate how well the content of these Web sites corresponds to the perception of domain experts. A detailed analysis of the frequency and sentiment of terms identified reveals that there are significant differences regarding attention and attitude towards the respective topics among the Web coverage of actors such as news media
sites, NGOs and companies. Figure 7 visualizes the domain ontology obtained from the ontology extension process. Black terms indicate the seed ontology, gray terms were added in the first iteration of ontology extension, white concepts in the second iteration. The method extracted terms such as “landfill gas”, “emissions”, “dioxin emissions” which are clearly relevant to the domain. Companies addressing environmental issues (McDonalds, Monsanto, Pharmacia, Renessen, Sunoco) got included, as well as environmental programs (Balanced Lifestile, GRI, Redirectory) and chemical substances (DEHP, RoundupR). The usefulness of terms such as “borrower”, “issue series” and “assets” is less clear, but were included since they co-occur quite frequently with some of the seed terms. Including external evidence sources into the ontology learning process (Section 2) would reduce the impact of such relations derived from in-corpus evidence sources.

4.3 Media Watch on Climate Change

The Media Watch on Climate Change Hubmann-Haidvogel et al. (2009) builds contextualized information spaces by enriching documents with geospatial, semantic and temporal annotations. Ontology learning (relying on the frameworks presented in this chapter) is used to create lightweight domain ontologies for structuring the information in the contextualized information space, and to provide means for navigating the repository (Figure 8).
As in the previous use cases, the authors only consider taxonomic relations and a general non-taxonomic related type in the ontologies to prevent them from getting too complicated for (non-expert) users to read and understand. Geographic maps, semantic maps and tag clouds complement the ontology view and provide context information on documents and search queries.

5 Outlook and Conclusions

This chapter presented a blueprint for a generic ontology extension framework, together with a number of real-world applications. It sheds light on the different aspects of such a framework by (i) suggesting data sources which contain domain knowledge and might act as input for the ontology learning process, (ii) presenting a set of techniques to assemble the necessary components and achieve useful results for different application domains, (iii) stressing the balance necessary between in-corpus evidence and evidence from external sources to ensure a proper focus of the extended ontology, (iv) discussing actual implementations of these techniques and finally, (v) presenting use cases where these techniques have already been applied successfully.

The selected use cases demonstrate the importance and broad applicability of ontology learning. Simple lightweight ontologies help to structure knowledge and to navigate complex information spaces, and indicate how different actors perceive a domain. Future research will focus on the identification and inclusion of new data sources for ontology extension, the improvement of evidence plugins (e.g., by including more sophisticated information extraction and text mining algorithms), the optimization of transformation heuristics, the improvement of
the balance between external and in-corpus evidences, and the integration of user feedback into this process.

References


**Related Reading**


Glossary

Co-occurrence Analysis. Co-occurrence analysis determines whether terms are significantly over-represented in designated spans of text. The calculation of statistical significance compares the distribution of terms in a domain-specific target corpus with their distribution in a generic reference corpus to identify candidate terms of inclusion in the extended ontology.

Evidence. Evidence represents the input data for the ontology learning process. The presented framework relies on evidences from unstructured sources (domain text), social sources (for example APIs of Web 2.0 applications and tagging systems) and structured sources (online Semantic Web data and ontologies).

Evidence Integration. Integration of evidences from heterogeneous sources supports the ontology learning process. Combining in-corpus data with social sources, for example, will include an outside view of the domain into the learned ontology.

Sentiment. Sentiment is the emotional attitude towards abstract or real objects of their environment. Measures of individual or organizational bias that distinguish between positive, negative and neutral media coverage are important indicators for investigating trends and differing perceptions of stakeholder groups.

Spreading Activation. Spreading activation is a graph-based, interactive search technique inspired by cognitive models of the human brain. It is typically applied to various types of networks (e.g., associative, semantic or neural networks).

Trigger phrases. Trigger phrases rely on the heuristic that certain phrases (e.g., “renewable energy, especially solar energy ...”) often indicate hyponym, hypernym, and meronym relations. Trigger phrase analysis detects these constructs by using pattern matching via regular expressions combined with part-of-speech tags.

Vector Space Model. Vector space models are a common way to represent documents and queries in information retrieval systems, e.g. for computing similarity between documents, or between a query term and a document collection.