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

[1] Institute for Information Business, Vienna University of Economics and Business, Vienna, Austria, e-mail: {aweichse,wohlg}@ai.wu.ac.at

[2] Department 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 extend 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 & Garcia-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'Aquin, 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 exoerts 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/sweoig/taskforces/communityprojects/linkingopendata
[8] swoogle.umbc.edu
[9] www.sindice.com
[10] iws.seu.edu.cn/services/falcons
[11] swse.deri.org

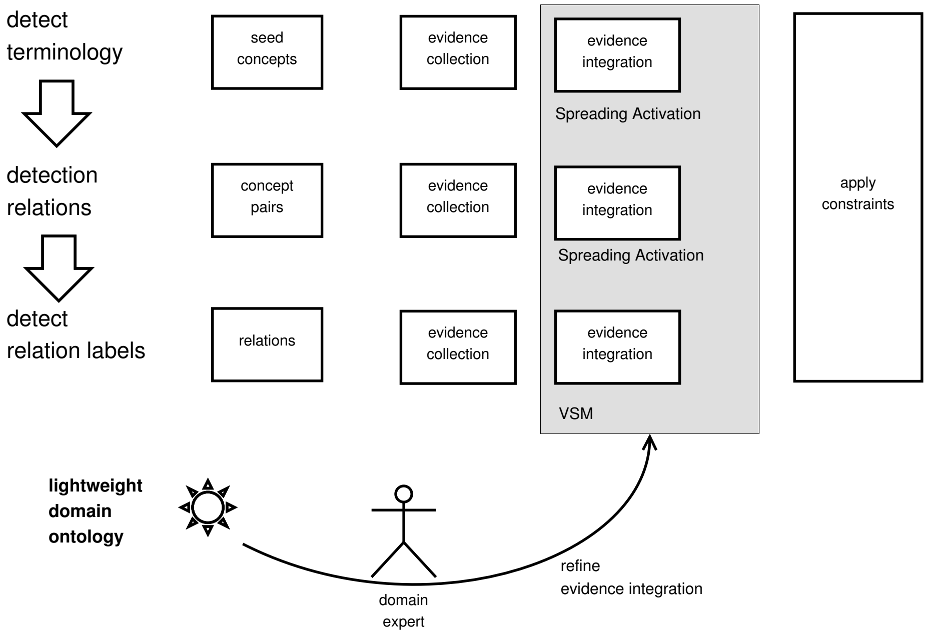


Figure 1: A generic ontology learning process

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 methdo, 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.

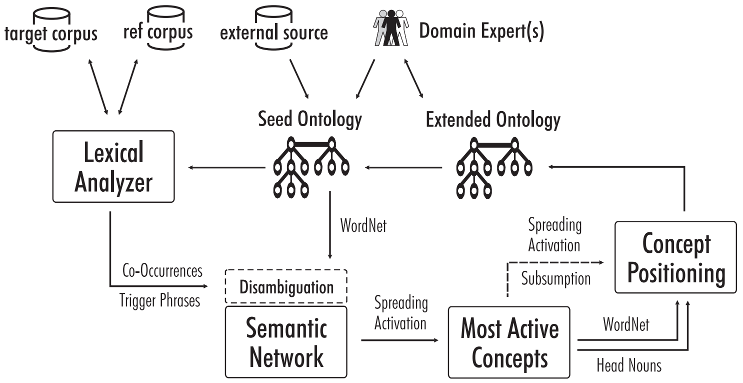


Figure 2: Ontology learning framework using spreading activation for evidence integration (Liu et al., 2005)

seed term	evidence	candidate term
climate change	wl:coOccurs	carbon dioxide
energy sources	wl:meronym	oil
energy	wl:hyperonym	renewable energy
climate change	wl:delicious	gas

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

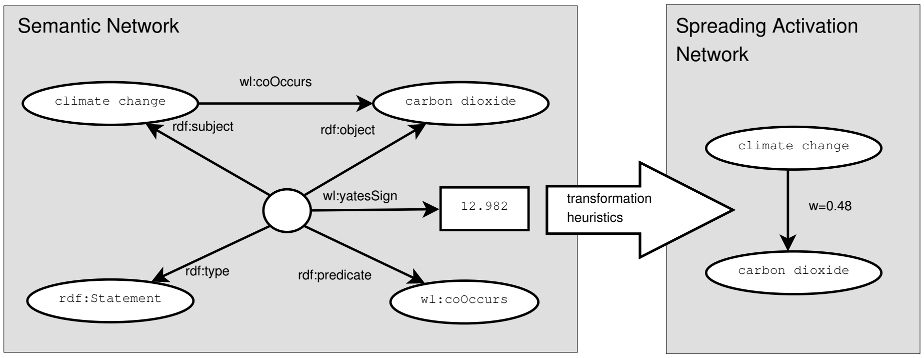


Figure 3: Transformation of RDF statements to spreading activation weights.

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 (Sanderson & 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

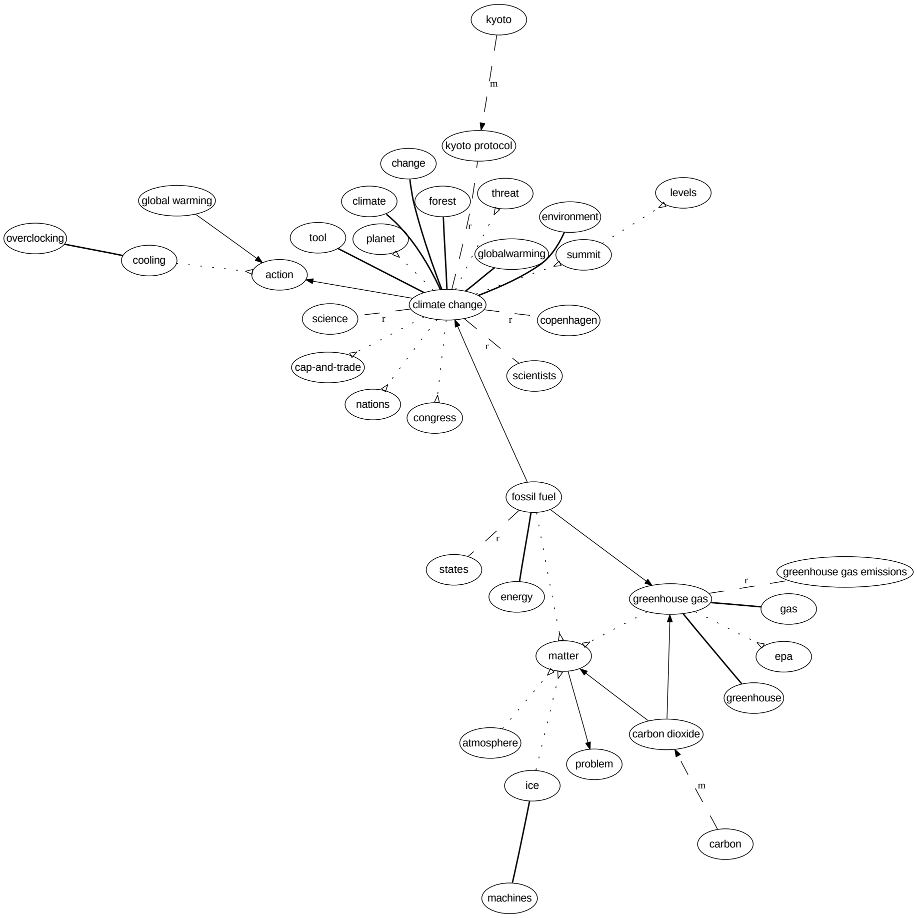


Figure 4: Extended ontology based on new terms from Delicious as an example of social evidence sources.

might be limited to seed ontology concepts, or links between new terms might be promoted.

Currently, the relation detection component only detects the *strongest* relation between the candidate term and the other terms in the ontology. Using cut-off levels and additional heuristics will allow the detection of multiple relations and provide a more fine-grained control over the relation detection process.

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 relation 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).

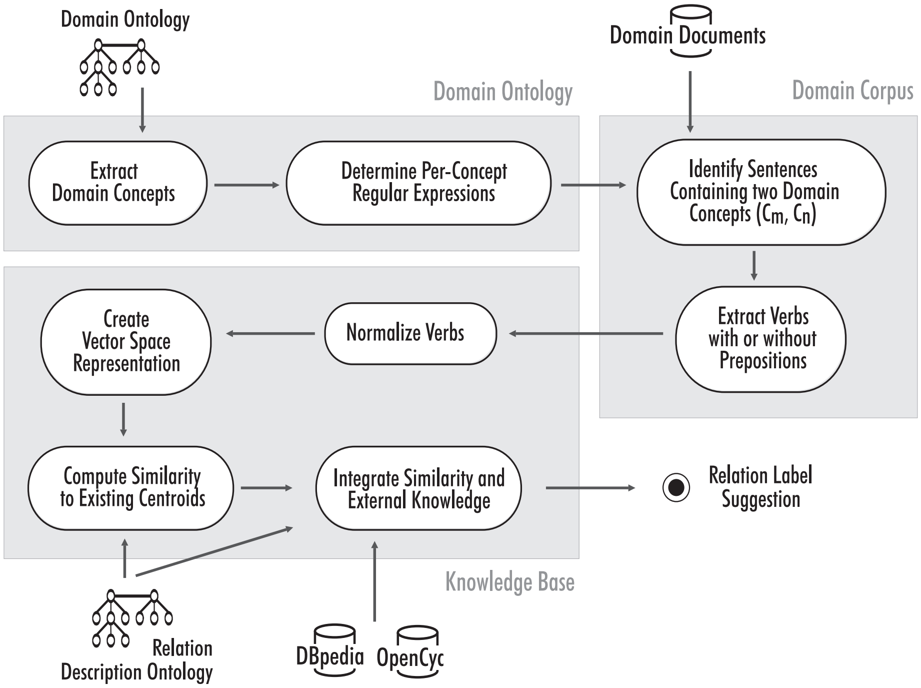

Domain Ontology
Domain Documents
Domain Ontology
Domain Corpus
Extract Domain Concepts
Determine Per-Concept Regular Expressions
Identify Sentences Containing two Domain Concepts (C_m, C_n)
Create Vector Space Representation
Normalize Verbs
Extract Verbs with or without Prepositions
Compute Similarity to Existing Centroids
Integrate Similarity and External Knowledge
Relation Label Suggestion
Knowledge Base
Relation Description Ontology
DBpedia
OpenCyc

Figure 5: Learning relation labels (Weichselbraun et al., 2010).

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 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* → *handcart*, *air travel* → *snowcam* which where 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

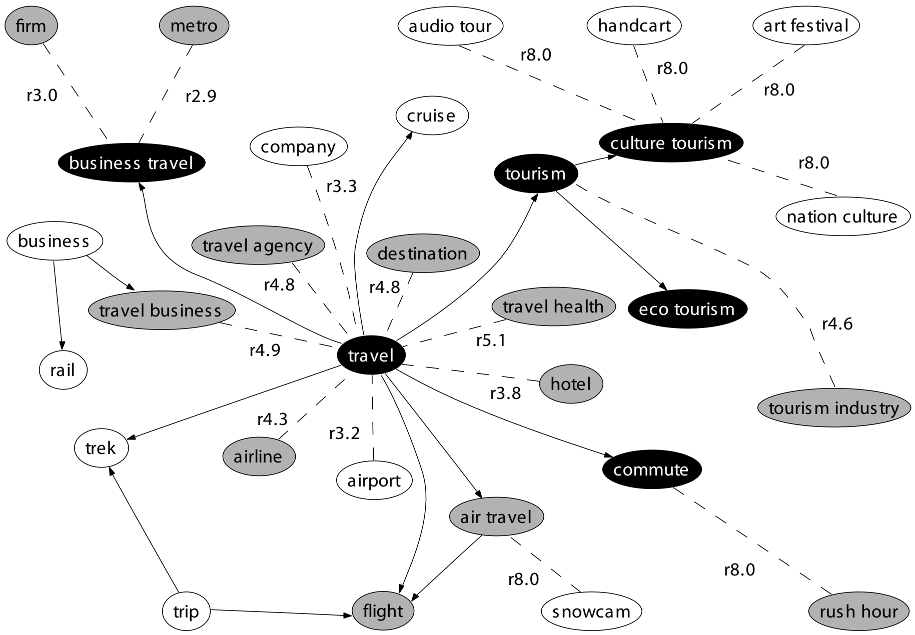


Figure 6: Tourism ontology (Dickinger et al., 2008).

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).

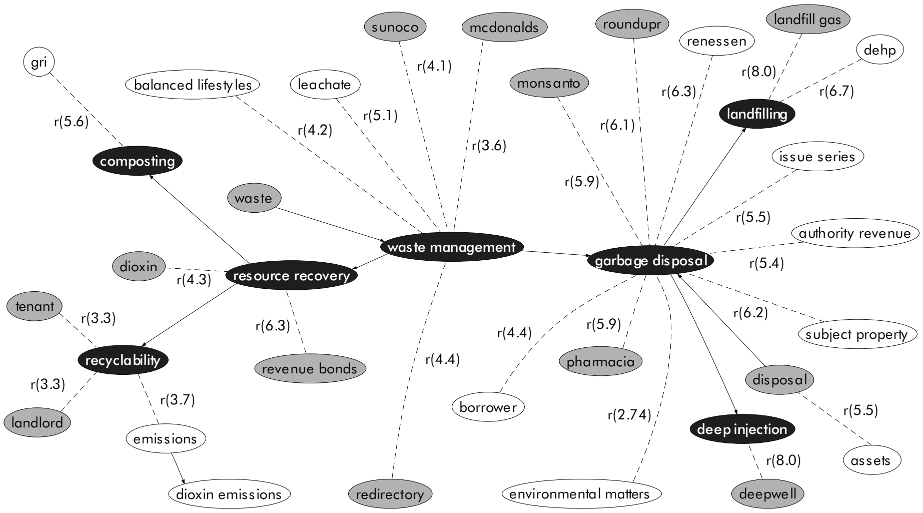


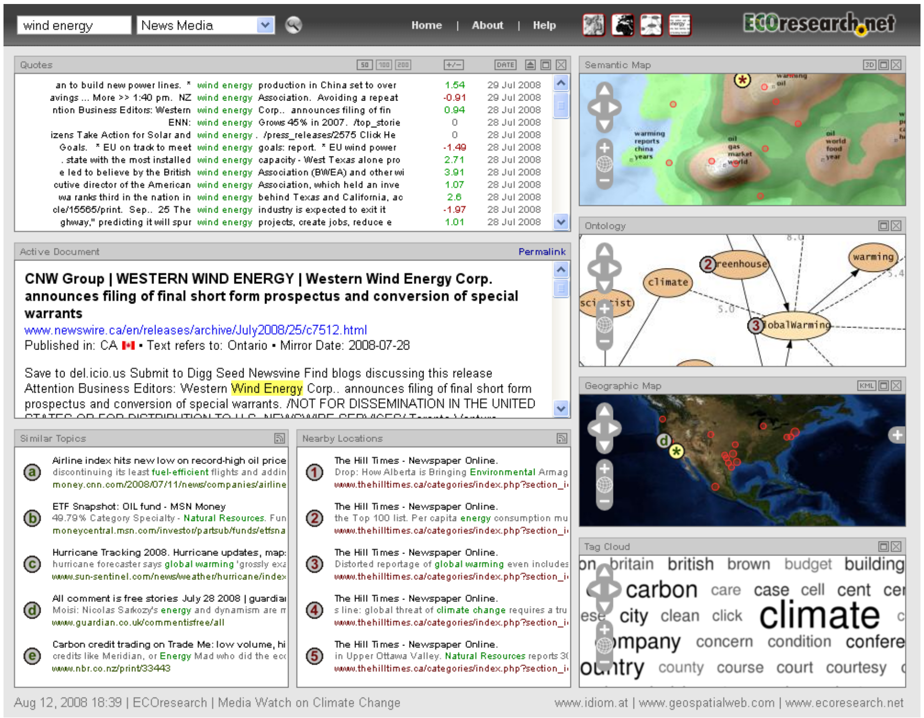
Figure 7: Solid waste disposal ontology (Pollach et al., 2009).

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 of 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



an to build new power lines. *	wind energy	production in China set to over	1.54	29 Jul 2008
avings ... More >> 1:40 pm. NZ	wind energy	Association. Avoiding a repeat	-0.91	29 Jul 2008
ntion Business Editors: Western	wind energy	Corp.. announces filing of fin	0.94	28 Jul 2008
ENN:	wind energy	Grows 45% in 2007. /top_storie	0	28 Jul 2008
izens Take Action for Solar and	wind energy	. /press_releases/2575 Click He	0	28 Jul 2008
Goals. * EU on track to meet	wind energy	goals: report. * EU wind power	-1.49	28 Jul 2008
. state with the most installed	wind energy	capacity - West Texas alone pro	2.71	28 Jul 2008
e led to believe by the British	wind energy	Association (BWEA) and other wi	3.91	28 Jul 2008
cutive director of the American	wind energy	Association, which held an inve	1.07	28 Jul 2008
wa ranks third in the nation in	wind energy	behind Texas and California, ac	2.6	28 Jul 2008
cle/15565/print. Sep.. 25 The	wind energy	industry is expected to exit it	-1.97	28 Jul 2008
ghway," predicting it will spur	wind energy	projects, create jobs, reduce e	1.01	28 Jul 2008

CNW Group | WESTERN WIND ENERGY | Western Wind Energy Corp. announces filing of final short form prospectus and conversion of special warrants
www.newswire.ca/en/releases/archive/July2008/25/c7512.html
Published in: CA 🇨🇦 • Text refers to: Ontario • Mirror Date: 2008-07-28

Save to del.icio.us Submit to Digg Seed Newsvine Find blogs discussing this release Attention Business Editors: Western Wind Energy Corp.. announces filing of final short form prospectus and conversion of special warrants. /NOT FOR DISSEMINATION IN THE UNITED

a Airline index hits new low on record-high oil price
discontinuing its least fuel-efficient flights and addin
money.cnn.com/2008/07/11/news/companies/airline

b ETF Snapshot: OIL fund - MSN Money
49.79% Category Specialty - Natural Resources. Fun
moneycentral.msn.com/investor/partsub/funds/etfsna

c Hurricane Tracking 2008. Hurricane updates, map:
hurricane forecaster says global warming 'grossly exa
www.sun-sentinel.com/news/weather/hurricane/index

d All comment is free stories July 28 2008 | guardia
Moisi: Nicolas Sarkozy's energy and dynamism are n
www.guardian.co.uk/commentisfree/all

e Carbon credit trading on Trade Me: low volume, hi
credits like Meridian, or Energy Mad who did the ec
www.nbr.co.nz/print/33443

1 The Hill Times - Newspaper Online.
Drop: How Alberta is Bringing Environmental Armag
www.thehilltimes.ca/categories/index.php?section_i

2 The Hill Times - Newspaper Online.
the Top 100 list. Per capita energy consumption mu
www.thehilltimes.ca/categories/index.php?section_i

3 The Hill Times - Newspaper Online.
Distorted reportage of global warming even includes
www.thehilltimes.ca/categories/index.php?section_i

4 The Hill Times - Newspaper Online.
s line: global threat of climate change requires a tru
www.thehilltimes.ca/categories/index.php?section_i

5 The Hill Times - Newspaper Online.
in Upper Ottawa Valley. Natural Resources reports 3
www.thehilltimes.ca/categories/index.php?section_i

Aug 12, 2008 18:39 | ECOresearch | Media Watch on Climate Change

www.idiom.at | www.geospatialweb.com | www.ecoresearch.net

Figure 8: Media Watch on Climate Change (Hubmann-Haidvogel et al., 2009).

the balance between external and in-corpus evidences, and the integration of user feedback into this process.

References

Alani, H. (2006, May 23-26). Position paper: ontology construction from online ontologies. In L. Carr, D. D. Roure, A. Iyengar, C. A. Goble, & M. Dahlin (Eds.), *Proceedings of the 15th international conference on world wide web (www 2006)* (pp. 491–495). Edinburgh, Scotland, UK: ACM.

Aleksovski, Z., Kate, W. ten, & Harmelen, F. van. (2006). Ontology matching using comprehensive ontology as background knowledge. In P. S. et al. (Ed.), *Proceedings of the international workshop on ontology matching at ISWC 2006* (pp. 13–24). Athens, GA, USA: CEUR.

Angeletou, S., Sabou, M., Specia, L., & Motta, E. (2007). *Bridging the gap between folksonomies and the semantic web: An experience report* (Vol. 2).

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., et al. (2009). DBpedia - a crystallization point for the web of data. *Journal of*

Web Semantics: Science, Services and Agents on the World Wide Web, *7*(3), 154–165.

Buitelaar, P., Cimiano, P., & Magnini, B. (2005, 7). Ontology learning from text: An overview. In *Ontology learning from text: Methods, evaluation and applications/ frontiers in artificial intelligence and applications* (Vol. 123, pp. 3–12). Amsterdam: IOS Press.

Cimiano, P. (2006). *Ontology learning and population from text: Algorithms, evaluation and applications*. Springer.

Corcho, O. (2006). Ontology based document annotation: trends and open research problems. *IJMSO*, *1*(1), 47–57.

Crestani, F. (1997). Application of spreading activation techniques in information retrieval. *Artificial Intelligence Review*, *11*, 453–482.

d'Aquin, M., Motta, E., Sabou, M., Angeletou, S., Gridinoc, L., Lopez, V., et al. (2008). Toward a new generation of semantic web applications. *IEEE Intelligent Systems*, *23*(3), 20–28.

d'Aquin, M., Sabou, M., Motta, E., Angeletou, S., Gridinoc, L., Lopez, V., et al. (2008, December 15-17). What can be done with the semantic web? an overview of watson-based applications. In A. Gangemi, J. Keizer, V. Presutti, & H. Stoermer (Eds.), *Proceedings of the 5th workshop on semantic web applications and perspectives (swap2008)* (Vol. 426). Rome, Italy: CEUR-WS.org.

Dickinger, A., Scharl, A., Stern, H., Weichselbraun, A., & Wöber, K. (2008). Applying optimal stopping for optimizing queries to external semantic web resources. In P. O'Connor, H. Wolfram, & G. Ulrike (Eds.), *Information and communication technologies in tourism 2008, proceedings of the international conference in innsbruck, austria, 2008* (pp. 545–555). Vienna-New York: Springer.

Etzioni, O., Banko, M., Soderland, S., & Weld, D. S. (2008). Open information extraction from the web. *Commun. ACM*, *51*(12), 68–74.

Giuliano, C., Lavelli, A., & Romano, L. (2007). Relation extraction and the influence of automatic named-entity recognition. *ACM Transactions on Speech and Language Processing*, *5*(1), 1–26.

Gracia, J., Trillo, R., Espinoza, M., & Mena, E. (2006, July 11-14). Querying the web: a multiontology disambiguation method. In D. Wolber, N. Calder, C. Brooks, & A. Ginige (Eds.), *Proceedings of the 6th international conference on web engineering (ICWE 2006)* (pp. 241–248). Palo Alto, California, USA: ACM.

Grefenstette, G., & Hearst, M. A. (1992). A method for refining automatically-discovered lexical relations: Combining weak techniques for stronger results. In *Aaai workshop on statistically-based natural language programming techniques* (pp. 64–72). Menlo Park, CA: AAAI Press.

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing? *International Journal of Human-Computer Studies*, *43*(5-6), 907–928.

Harris, Z. (1968). *Mathematical structures of language.* John Wiley & Sons.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Coling* (p. 539-545). Nantes, France.

Heymann, P., & Garcia-Molina, H. (2006, April). *Collaborative creation of communal hierarchical taxonomies in social tagging systems* (Technical Report No. 2006-10). Computer Science Department.

Hubmann-Haidvogel, A., Scharl, A., & Weichselbraun, A. (2009). Multiple coordinated views for searching and navigating web content repositories. *Information Sciences*, *179*(12), 1813–1821.

Joho, H., Sanderson, M., & Beaulieu, M. (2004, April 5-7). A study of user interaction with a concept-based interactive query expansion support tool. In S. McDonald & J. Tait (Eds.), *Advances in information retrieval, 26th european conference on ir research (ECIR 2004)* (Vol. 2997, p. 42-56). Sunderland, UK: Springer.

Landauer, T., & Dumais, S. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, *104*(2), 211-240.

Liu, W., Weichselbraun, A., Scharl, A., & Chang, E. (2005). Semi-automatic ontology extension using spreading activation. *Journal of Universal Knowledge Management*, *0*(1), 50–58.

Maedche, A., Pekar, V., & Staab, S. (2002). Ontology learning part one - on discovering taxonomic relations from the web. In N. Zhong, J. Liu, & Y. Yao (Eds.), *Web intelligence* (p. 301-322). Springer.

Mika, P. (2007). Ontologies are us: A unified model of social networks and semantics. *Journal of Web Semantics*, *5*(1), 5–15.

Pollach, I., Scharl, A., & Weichselbraun, A. (2009). Web content mining for comparing corporate and third party online reporting: a case study on solid waste management. *Business Strategy and the Environment*, *18*(3), 137-148.

Salton, G., Wong, A., & Yang, C. S. (1975, November). A vector space model for information retrieval. *Communications of the ACM*, *18*(11), 613-620.

Sánchez-Alonso, S., & García, E. (2006). Making use of upper ontologies to foster interoperability between skos concept schemes. *Online Information Review*, *30*(3), 263-277.

Sanderson, M., & Croft, W. B. (1999). Deriving concept hierarchies from text. In *22nd annual international acm sigir conference on research and development in information retrieval* (pp. 206–213). Berkeley, USA.

Schmitz, P. (2006). Inducing ontology from flickr tags. In *Collaborative web tagging workshop at www2006.* Edinburgh, Scotland.

Specia, L., & Motta, E. (2007). Integrating folksonomies with the semantic web. In *The semantic web: Research and applications, 4th european semantic web conference (eswc-2007)* (Vol. 4519, p. 624-639). Berlin: Springer.

Studer, R. R., Benjamins, R., & Fensel, D. (1998). Knowledge engineering: principles and methods. *Data and knowledge engineering*, *25*(1-2), 161–197.

Tang, J., Leung, H. fung, Luo, Q., Chen, D., & Gong, J. (2009). Towards ontology learning from folksonomies. In *Ijcai'09: Proceedings of the 21st international jont conference on artifical intelligence* (pp. 2089–2094). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Weichselbraun, A., Wohlgenannt, G., & Scharl, A. (2010). Refining non-taxonomic relation labels with external structured data to support ontology learning. *Data & Knowledge Engineering*, *69*(8), 763–778.

Yates, F. (1934). Contingency table involving small numbers and the χ^2 test. *Supplement to the Journal of the Royal Statistical Society*, *1*(2), 217–235.

Related Reading

Fellbaum, C. (1998). Wordnet - an electronic lexical database. *Computational Linguistics*, *25*(2), 292–296.

Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.* Prentice Hall.

Powers, S. (2003). *Practical RDF.* Sebastopol, CA, USA: O'Reilly & Associates, Inc.

Ruiz-Casado, M., Alfonseca, E., & Castells, P. (2007). Automatising the learning of lexical patterns: An application to the enrichment of wordnet by extracting semantic relationships from wikipedia. *Data & Knowledge Engineering*, *61*(3), 484 - 499.

Segaran, T. (2007). *Collective intelligence - building smart web 2.0 applications.* O'Reilly.

Shadbolt, N., Berners-Lee, T., & Hall, W. (2006). The semantic web revisited. *IEEE Intelligent Systems*, *21*(3), 96–101.

Tao, C., & Embley, D. W. (2009). Automatic hidden-web table interpretation, conceptualization, and semantic annotation. *Data & Knowledge Engineering*, *68*(7), 683 - 703. (Special Issue: 26th International Conference on Conceptual Modeling (ER 2007))

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, interative 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.