2010 Workshops on Database and Expert Systems Applications

Augmenting Lightweight Domain Ontologies with Social Evidence Sources

Albert Weichselbraun
Institute for Information Business
Vienna University of Economics
Vienna, Austria
Email: aweichse@ai.wu.ac.at

Gerhard Wohlgenannt
Institute for Information Business
Vienna University of Economics
Vienna, Austria
Email: wohlg@ai.wu.ac.at

Arno Scharl
Department of New Media Technology
MODUL University Vienna
Vienna, Austria
Email: scharl@modul.ac.at

***Abstract*—Recent research shows the potential of utilizing data collected through Web 2.0 applications to capture changes in a domain's terminology. This paper presents an approach to augment corpus-based ontology learning by considering terms from collaborative tagging systems, social networking platforms, and micro-blogging services. The proposed framework collects information on the domain's terminology from domain documents and a seed ontology in a triple store. Data from social sources such as Delicious, Flickr, Technorati and Twitter provide an outside view of the domain and help incorporate external knowledge into the ontology learning process. The neural network technique of spreading activation is used to identify relevant new concepts, and to determine their positions in the extended ontology. Evaluating the method with two measures (PMI and expert judgements) demonstrates the significant benefits of social evidence sources for ontology learning.**

***Keywords*-Evidence Source Integration; Ontology Learning; Spreading Activation; Social Evidence Source; Web 2.0**

I. Introduction

By conceptualizing an application domain (1), ontologies facilitate a common understanding of domain concepts and relations among different stakeholder groups. Such ontologies require ongoing updates and refinements to keep track with evolving domain knowledge, tasks that are both labor-intensive and costly. By supporting and guiding these processes, automated ontology learning improves productivity, reduces the human input required, and paves the way for applying semantic technologies to real-world problems.

Ontology learning includes a number of subtasks such as the detection of synonyms, concepts, taxonomies, relations and axioms, which partly build on each other. Cimiano presents an extensive overview of ontology learning methods in (2). Methods that rely on corpus statistics such as term co-occurrence (3), association rules (4), Latent Semantic Indexing (LSI) based methods for synonym and concept detection, and the application of kernel methods to classify semantic relations (5) are common techniques in various subfields of ontology learning. In recent years, blueprints to combine these approaches emerge. However, with notable exceptions such as the work of Correndo et al. (6), the field of ontology learning has not fully exploited the increased availability of structured and social sources (7). As a result, it lags behind other fields such as ontology matching and ontology-based question answering, where encouraging results were obtained by reusing structured evidence sources (i.e., ontologies) from third parties (8). This paper suggests to adapt novel paradigms of knowledge reuse to ontology learning. Many important concepts are never mentioned in textual data as they represent the common ground between readers and authors in a given community, so referring to them in an explicit manner is not necessary (2). Non-textual resources such as online ontologies and collective intelligence (social) in the form of folksonomies (9) represent rich sources of complementary data (when using selected parts of third-party ontologies, for example, or detecting relations between these ontologies (8). Integrating unstructured and social sources bridges the gap between knowledge expressed in textual form, and knowledge captured in social sources such as tagging systems, social networking and micro-blogging services.

Mika (10), Heymann (11) and Schmitz (12) retrieve data from social sources to build ontologies solely based on this information. In contrast, the approach presented in this paper focuses on learning domain-specific ontologies based on a corpus of domain documents by extracting relevant terminology and relations from these documents. The present work integrates social evidence sources into this process (i) to capture ontology evolution processes, and (ii) to retrieve external background knowledge on the domain's terminology. Mika (10) notes that social sources are likely to complement well-established but slowly evolving ontologies by revealing emergences from user actions. This observation is also backed by Angeletou et al. (13), who note that folksonomies tend to reflect the latest terminology within a domain due to their high update frequency.

The remainder of this paper is structured as follows: Section II presents a method for learning domain ontologies which considers social evidence sources in the ontology building process and describes the social sources used by the architecture. Section III presents an evaluation of our approach which has been performed using an evaluation measure and domain experts. The paper closes with an outlook and conclusions drawn in Section IV.

1529-4188/10 $26.00 © 2010 IEEE
DOI 10.1109/DEXA.2010.53

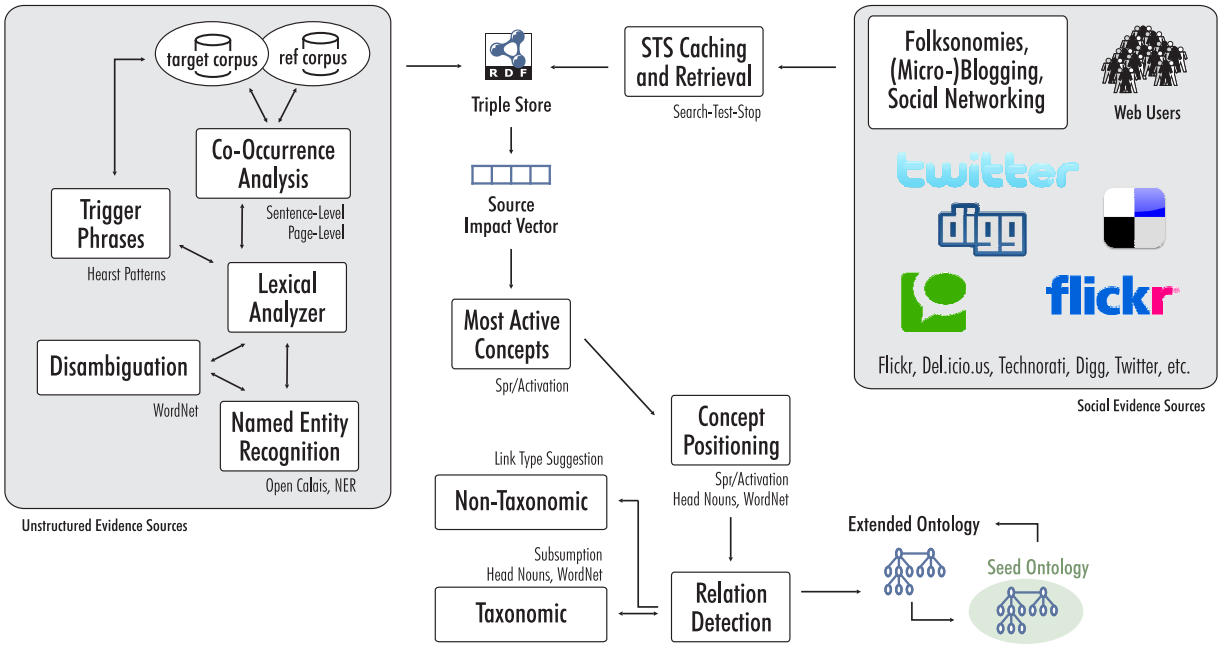

Figure 1. Ontology Extension Architecture System Diagram.

II. Ontology Learning

The method presented in this article extends existing seed ontologies based on a corpus of domain documents. Figure 1 outlines the ontology extension process. Unstructured sources (Section II-A) such as Web documents are enriched by data from social evidence sources such as Del.icio.us, Flickr, Technorati and Twitter (Section II-B). Evidence sources (e) process the seed ontology and corpus text to yield relations between seed ontology concepts (C_s) and candidate concepts (C_c). The evidence source determines the kind of this relation (e.g., "co-occurs", "related Twitter tag", "related Delicous tag", etc.). An RDF triple store collects the evidence, as outlined in Table I.

seed ontology concept concept (C_s)	evidence source (e)	candidate concept (C_c)
climate change	oe:coOccurs	greenhouse gases
climate change	oe:twitterTag	environment
climate change	oe:deliciousTag	fuel

Table I
Example evidence entries in the triple store.

Reification is used to add meta-data on the extracted relations such as significance levels, dice coefficients, and counts of the relations in the triple store. Applying transformation heuristics and spreading activation integrates the collected evidence into the ontology building process. Per evidence source heuristics transform the relation between the seed and candidate concepts into a numerical weight used in a spreading activation network. These weights reflect the expected contribution of the evidence source to the final ontology and consider source-specific annotations (Section II-C). The neural network method of spreading activation integrates the typically heterogeneous data from multiple evidence sources and computes a ranked list of candidate concepts for inclusion in the extended ontology. A combination of noun phrases, subsumption analysis and spreading activation then determines the new concepts' positions in the extended ontology (3). This work focuses on terminology and therefore only determines relations between concepts, but not the relation type.

A major challenge of this approach is balancing internal and external sources. The impact of external sources should be limited in order not to jeopardize the creation of domain ontologies, which reflect the terminology used in the domain corpus (13). This might be achieved by assigning relatively high weights to in-corpus data while still including external data (e.g., from social evidence sources) reflecting the latest trends in the field. The following sections outline the balancing of evidence sources and the composition of the source impact vector in greater detail.

A. Unstructured Evidence Sources

Candidate terms are extracted from text documents by applying information extraction and text mining techniques such as significant phrase detection, co-occurrence analysis and trigger phrases. Liu et al. (3) provide a more detailed description of this automated approach to ontology learning from unstructured evidence sources.

B. Social Evidence Sources

Querying Web 2.0 services (e.g., tagging, social networking and micro-blogging applications) with seed ontology terms provides candidate concepts for the extended ontology. The TagInfoService interface of the easy Web Retrieval Toolkit (eWRT; www.semanticlab.net/index.php/eWRT) helped capture these social evidence sources. The interface allows determining a tag's popularity, and retrieving tags related to the input tag.

Del.icio.us and Flickr provide an Appliation Programming Interface (API) to retrieve the number of entities which have been labeled with a specific tag (= tag popularity), and to determine related tags. Technorati does not offer such an API. Therefore, we had to implement a method which computes related tags based on the tags in the top 100 blog entries returned for a target tag. The same strategy has been applied to Twitter.

A transformation function t transforms ontology concepts C into tags T for comparing tag popularities. We apply the dice coefficient to determine the similarity (s_d) between these tags:

$$s_d(T_s, T_c) \quad = \quad \frac{2 \cdot n_{T_{sc}}}{n_{T_s} + n_{T_c}} \tag{1}$$

$n_{T_{sc}}$ denotes to the number of times the tags T_s and T_c have been used together to tag a blog or a Web site, (n_{T_s} + n_{T_c}) refers to the number of times a Web site has been tagged using any of these tags.

An inherent problem of consulting external evidence sources is the introduction of unrelated terms due to linguistic ambiguities. Currently, we utilize WordNet data (wordnet.princeton.edu) for word sense disambiguation in order to minimize the negative impact of such ambiguities.

C. Evidence Integration

The evidence vector $\vec{r}(C_s, C_c)$ contains evidence sources e which indicate relations $r_e(C_s, C_c)$ between seed concepts C_s and candidate concepts C_c. Heuristic per-evidence-source translation rules s_e transform these relations using the source impact vector $\vec{S} = (s_{e_1}, s_{e_2}, ...s_{e_n})^T$ into a numerical weight

$$w(C_s, C_c) \quad = \quad |\vec{S}\big(\vec{r}(C_s, C_c)\big)| \tag{2}$$

for the spreading activation network. The translation heuristics reflect the evidence source's importance in the extension process and have proven their usefulness in related research on ontology extension and evolution (3; 14; 15; 16). The following example illustrates the integration process. Three evidence sources suggest a relation between the terms climate change (cc) and fuel, resulting in the following evidence vector:

$$\vec{r}(\text{cc,fuel}) \quad = \quad \begin{pmatrix} (oe : coOccurs, sign = 3.2) \\ (oe : deliciousTag, dice = 1.59) \\ (oe : triggerPhrase) \end{pmatrix}$$

Equation 2 uses the source impact vector

$$\vec{S} = \begin{pmatrix} 0.1 + 0.5 \cdot sign \\ 0.2 \cdot dice \\ 0.3 \end{pmatrix}$$

to compute the weight w(cc, fuel) $= 2.318$ for this data. Applying this process to all seed candidate concept pairs (C_s, C_c) yields the spreading activation network used to determine the candidate concepts to include in the domain ontology.

III. Evaluation

This section outlines the experiments conducted to evaluate the performance of the ontology learning framework. We initiated the ontology extension processes with a small seed ontology that involves only two relations: *fossil fuels* $\xrightarrow{relatedTo}$ *climate change* and *fossil fuels* $\xrightarrow{relatedTo}$ *greenhouse gas(es)*. The extensions utilized five distinct domain-specific corpora collected between April and August 2009.

To contrast the impact of social evidence sources, each extension was performed (i) based on only the unstructured source (= text corpus), and (ii) based on the unstructured source in conjunction with related tags collected from social sources. Starting with the seed ontology, two iterations of ontology learning extended the ontology by 24 new concepts, which were chosen from the ordered list of automatically generated suggestions.

A. Domain Documents and Social Sources

To create the corpora for the evaluation, we mirrored 156 news media sites from the Newslink.org, Kidon.com and ABYZNewsLinks.com directories. The webLyzard suite of Web mining tools (www.weblyzard.com) crawls those sites in regular intervals, gathering around 200,000 documents per week. Domain detection based on regular expressions was used to compile domain-specific corpora with documents published between April and August 2009. Since the number of documents in each corpus was restricted to 1250, the domain corpora represent a broad overview of monthly media coverage on the seed concepts.

Table II presents the terms generated in the second extension step for the ontology based on the August 2009 corpus. The column *unstructured* includes terms generated solely based on corpus data, the remaining columns the first 17 of the related tags collected from social evidence sources.

The terms extracted from corpus data (column 1) vary from very relevant (e.g. "carbon dioxide emissions") to hardly relevant at all (e.g. "levels"). This also applies to terms gathered from social sources, where the percentage of non-relevant entries is even higher. The fact that some of the input terms generated in previous ontology learning steps are themselves not domain-relevant might help explain this observation, but even for domain-relevant input social evidence sources return terminology of different levels of

relevance. Social sources are still helpful when combined with evidence from unstructured sources as described in Section II, which decreases the influence of irrelevant terms and enforces relevant terminology.

unstructured	**social**	
	delicious	flickr
targets	animalcare	architecture
building	architects	art
coal	atmosphere	auckland
levels	award	beach
climate change policy	britney	bicycle
pact	carbonfootprint	brian
reduce greenhouse gas	...	...
pollution		
firm		
carbon dioxide emissions	technorati	twitter
ets	agile	aces
its carbon	apple	afghan
	architecture	afghanistan
	art	africa
	automotive	al_gore
	...	...

Table II
TERMS FROM UNSTRUCTURED AND SOCIAL EVIDENCE SOURCES.

B. Results

Table III outlines the method's performance based on the pointwise mutual information (PMI) measure, which determines how well the terms participating in a relation are associated to each other based on the counts of the source (n_{T_s}) and the candidate tag (n_{T_c}) and the number of times they occur with each other ($n_{T_{sc}}$):

$$n_z = n_{T_{sc}} + n_{T_s} + n_{T_c} \quad (3)$$

$$f(i) = \frac{n_i}{n_z} e^{-\frac{ni}{n_z}} \quad (4)$$

$$PMI(T_s, T_c) = f(n_{T_{sc}})/f(n_{T_s}) \cdot f(n_{T_c}) \quad (5)$$

The PMI measure is complemented by domain expert assessments. Four domain experts rated each relation identified by the ontology learning components on a discrete scale from not relevant (0), slightly relevant (1) and very relevant (2) to the domain. The average expert rating per relation serves as measure to evaluate relations learned from unstructured sources (column 2) and unstructured combined with social sources (column 3). We applied the measures to five ontologies based on corpus data from the time periods indicated in column 1. The values in parenthesis refer to the number of relations unique to the respective ontology and therefore omits relations shared between the learned ontologies as the evaluation focuses on the *differences* in the terminology yielded by both methods.

The comparison in Table III shows that the results obtained with social evidence sources clearly outperform corpus-only data for both evaluation metrics, and for each of the ontologies evaluated. The differences in the average evaluation score of the two methods are significant, exceeding 99.9% for a Welch two sample t-test (for the PMI) as well as a Wilcoxon rank sum test with continuity correction (for the discrete expert ranking). The average standard deviation among expert assessments is 0.45. A substantial portion of disagreement was caused by a single evaluator, the standard deviation among the remaining experts amounts to 0.34.

avg. PMI	unstructured	unstr. & social
April 2009	0.694 (16)	0.833 (17)
May 2009	0.753 (15)	0.921 (10)
June 2009	0.569 (16)	0.544 (15)
July 2009	0.625 (8)	0.862 (8)
August 2009	0.493 (5)	0.874 (9)
Sum	0.503 (60)	0.646 (59)

expert eval.	unstructured	unstr. & social
April 2009	0.875 (16)	1.353 (17)
May 2009	0.883 (15)	1.550 (10)
June 2009	1.000 (16)	1.283 (15)
July 2009	1.469 (8)	1.563 (8)
August 2009	1.150 (5)	1.167 (9)
Sum	1.013 (60)	1.369 (59)

Table III
IMPACT OF SOCIAL EVIDENCE SOURCES ON ONTOLOGY LEARNING PERFORMANCE EVALUATED WITH PMI AS WELL AS DOMAIN EXPERT JUDGEMENT.

Table IV contrasts a selection of terms included and removed in accordance with social evidence sources. Most of the removed terms (as well as the added ones) are relevant, but n-grams are often removed due to the bias of tagging towards unigrams. User-generated tags typically consist of unigrams, although users often indicate n-grams by concatenating words or using underscores to separate them. Since there is no agreed notation for n-grams, such tags are rare and hard to extract. Many users also use abbreviations such as AGW (Anthropogenic Global Warming), CPRS (Carbon Pollution Reduction Schema) and EPA (Environmental Protection Agency).

terms removed	**terms added**
carbon dioxide emissions	agw
climate change policy	biomass
developing nations	cprs
kyoto protocol	cars
scientific assessments	epa
sulfur dioxide	ethanol
tom magliozzi	greenhouse-gas

Table IV
SELECTION OF TERMS REMOVED AND ADDED BASED ON EVIDENCE FROM SOCIAL SOURCES.

The included and removed concepts demonstrate the impact of social sources on the extension process, currently emphasizing unigrams and therefore causing the removal of some relevant n-grams such as *climate change policy* or *reduce greenhouse gas*. Depending on the actual application of the technology, deployed systems should consider this effect by compensating n-gram resources for the lack of support from tagged resources.

IV. Conclusions

The ontology extension framework presented in this paper draws upon unstructured evidence sources (e.g. archives of Web documents) and social evidence sources such as Del.icio.us, Flickr, Technorati and Twitter. The results demonstrate the benefits of integrating multiple evidence sources for ontology learning from a multi-stakeholder view. Two evaluation metrics (pointwise mutual information, domain expert assessments) measure the quality of ontological concepts as well as the impact of including social evidence sources in the ontology extension process.

Future work will address the issue of extracting n-grams from social sources by combining statistical approaches toward tag relatedness such as co-occurrence analysis, distributional measures and FolkRank (17) with thesauri and lexicons. Additional and refined evaluation metrics will be able to detect and assess shifts in terminology caused by integrating social evidence sources, for example the inclusion of implicit domain terminology or latest trends. Incorporating evidence structured sources such as Swoogle, DBpedia and Freebase in the extension process represents another promising research avenue.

V. Acknowledgment

The research project RAVEN (Relation Analysis and Visualiation; www.modul.ac.at/nmt/raven) is funded by the Austrian Ministry of Transport, Innovation & Technology and the Austrian Research Promotion Agency within the strategic objective FIT-IT Semantic Systems (www.fit-it.at). The authors would like to thank Heinz Lang for his help in implementing the necessary extensions to eWRT and providing the ontology visualizations and Syed Kamran Ali Ahmad for proofreading the manuscript.

References

[1] T. R. Gruber, "Toward principles for the design of ontologies used for knowledge sharing?" *International Journal of Human-Computer Studies*, vol. 43, no. 5-6, pp. 907–928, 1995.

[2] P. Cimiano, *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. Springer, 2006.

[3] W. Liu, A. Weichselbraun, A. Scharl, and E. Chang, "Semi-automatic ontology extension using spreading activation," *Journal of Universal Knowledge Management*, vol. 0, no. 1, pp. 50–58, 2005.

[4] A. Maedche, V. Pekar, and S. Staab, "Ontology learning part one - on discovering taxonomic relations from the web," in *Web Intelligence*, N. Zhong, J. Liu, and Y. Yao, Eds. Springer, 2002, pp. 301–322.

[5] C. Giuliano, A. Lavelli, and L. Romano, "Relation extraction and the influence of automatic named-entity recognition," *ACM Transactions on Speech and Language Processing*, vol. 5, no. 1, pp. 1–26, 2007.

[6] G. Correndo, H. Alani, and M. Salvadores, "Social support for ontological mediation and data integration," *International Journal of Virtual Communities and Social Networking*, vol. 1, no. 3, pp. 19–34, 2009.

[7] D. Sánchez and A. Moreno, "Learning non-taxonomic relationships from web documents for domain ontology construction," *Data & Knowledge Engineering*, vol. 64, no. 3, pp. 600–623, 2008.

[8] M. d'Aquin, E. Motta, M. Sabou, S. Angeletou, L. Gridinoc, V. Lopez, and D. Guidi, "Toward a new generation of semantic web applications," *IEEE Intelligent Systems*, vol. 23, no. 3, pp. 20–28, 2008.

[9] L. Specia and E. Motta, "Integrating folksonomies with the semantic web," in *The Semantic Web: Research and Applications, 4th European Semantic Web Conference (ESWC-2007)*, ser. LNCS, vol. 4519. Berlin: Springer, 2007, pp. 624–639.

[10] P. Mika, "Ontologies are us: A unified model of social networks and semantics," *Journal of Web Semantics*, vol. 5, no. 1, pp. 5–15, 2007.

[11] P. Heymann and H. Garcia-Molina, "Collaborative creation of communal hierarchical taxonomies in social tagging systems," Stanford University, Department of Computer Science, Technical Report 2006-10, April 2006.

[12] P. Schmitz, "Inducing ontology from flickr tags," in *Proc. of the Collaborative Web Tagging Workshop (WWW '06)*, Edinburgh, Scotland, May 2006.

[13] S. Angeletou, M. Sabou, L. Specia, and E. Motta, "Bridging the gap between folksonomies and the semantic web: An experience report," in *Workshop: Bridging the Gap between Semantic Web and Web*, vol. 2, 2007.

[14] A. Scharl, A. Dickinger, and A. Weichselbraun, "Analyzing news media coverage to acquire and structure tourism knowledge," *Information Technology and Tourism*, vol. 10, no. 1, pp. 3–17, 2008.

[15] A. Weichselbraun, A. Scharl, and W. Liu, "Capturing and classifying ontology evolution in news media archives," in *19th International Conference on Database and Expert Systems Applications (DEXA '08); Seventh International Workshop on Web Semantics*, A. M. Tjoa and R. R. Wagner, Eds. Turin, Italy: IEEE Computer Society Press, 2008, pp. 197–201.

[16] I. Pollach, A. Scharl, and A. Weichselbraun, "Web content mining for comparing corporate and third party online reporting: a case study on solid waste management," *Business Strategy and the Environment*, vol. 18, no. 3, pp. 137–148, 2009.

[17] C. Cattuto, D. Benz, A. Hotho, and G. Stumme, "Semantic Grounding of Tag Relatedness in Social Bookmarking Systems," in *The Semantic Web - ISWC 2008*, ser. Lecture Notes in Computer Science. Springer, 2008, vol. 5318, pp. 615–631.