Journal of Universal Knowledge Management, vol. 0, no. 1 (2005), 50-58
submitted: 7/2/05, accepted: 18/4/05, appeared: 28/6/05 © J.UKM

Semi-Automatic Ontology Extension Using Spreading Activation

Wei Liu, Albert Weichselbraun
University of Western Australia, School of Computer Science and Software Engineering
wei@csse.uwa.edu.au, weichselbraun@ecoresearch.net

Arno Scharl
Vienna University of Economics and Business Administration,
Information Systems Department, Austria
scharl@ecoresearch.net

Elizabeth Chang
Curtin University of Technology, Curtin Business School, Australia
elizabeth.chang@cbs.curtin.edu.au

Abstract: This paper describes a system to semi-automatically extend and refine ontologies by mining textual data from the Web sites of international online media. Expanding a seed ontology creates a semantic network through co-occurrence analysis, trigger phrase analysis, and disambiguation based on the WordNet lexical dictionary. Spreading activation then processes this semantic network to find the most probable candidates for inclusion in an extended ontology. Approaches to identifying hierarchical relationships such as subsumption, head noun analysis and WordNet consultation are used to confirm and classify the found relationships. Using a seed ontology on "climate change" as an example, this paper demonstrates how spreading activation improves the result by naturally integrating the mentioned methods.

Keywords: Ontology Extension, Disambiguation, Co-occurrence Analysis, Semantic Network, Spreading Activation
Categories: H.3.1, H.3.3, I.2.4

1 Introduction

Ontologies support shared understanding of domains of interest [Uschold and Grüninger 1996] by eliminating conceptual and terminological confusion among members of a virtual community. Although the crucial importance of ontologies in open environments is widely recognized, creating *specific domain ontologies* is still a laborious process for knowledge engineers and domain experts alike.

This paper describes ontology extension and ontology refinement by analyzing a large sample of international online media. The project drew upon the *Newslink.org*, *Kidon.com* and *ABYZNewsLinks.com* directories to compile a list of 156 news media sites from five English-speaking countries: United States, Canada, United Kingdom, Australia and New Zealand. A crawling agent mirrored their Web sites twice, once in March 2005 and once in April 2005. The agent followed the sites' hierarchical structure until reaching 50 megabytes of textual data. The system then identified and

removed redundant copies of headlines and non-contextual navigational elements, whose appearance on multiple pages may distort frequency counts.

This research uses spreading activation over weighted graphs to integrate methods of discovering hierarchical relationships such as trigger phrases, subsumption analysis, head noun analysis and lexical dictionary consultation. The prototype suggests additional, domain-specific terms and their ontological positioning, and serves as a test bed for evaluating heuristic rules to identify semantic relations. Domain experts are consulted to evaluate the validity of the resulting hierarchical structure.

The following [Section 2] describes the system architecture. [Sections 3 and 4] detail the syntactical and lexical analysis, as well as the process of discovering hierarchical relations. [Section 5] summarizes the results and outlines future research.

2 Ontology Extension System Architecture

[Fig. 1] presents a conceptual view on the ontology extension system architecture. A small set of terms from domain experts or from known ontology repositories is first selected as a seed ontology. The seed ontology terms are then fed into the *Lexical Analyzer*. Co-occurrence analysis at both the sentence and the document level limits the influence of popular terms not related to the domain [Roussinov and Zhao 2004]. Terms are selected according to a threshold value on the co-occurrence significance. Lexical analysis is done through consulting the WordNet lexical dictionary [Fellbaum 1998], and through analyzing the Web corpus for terms that are connected by *trigger phrases*. A trigger phrase is a phrase that matches a fragment of text that contains a parent-child description [Joho et al. 2004].

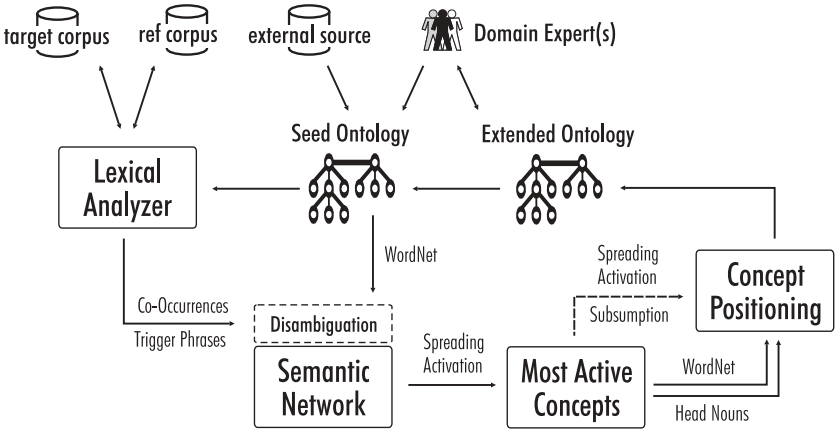

Figure 1: Ontology Extension System Architecture

The generated terms are connected with the seed ontology via directed weighted links. Once the network is established, spreading activation identifies the terms most relevant within the domain and suggests their incorporation into the seed ontology.

WordNet, head nouns and subsumption analysis are then used to confirm the semantic relationship. For terms not confirmed automatically, the domain experts are consulted, or another iteration of spreading activation over newly acquired terms is triggered to gather additional evidence.

3 Using Co-Occurrence Analysis to Discover Related Terms

Domain terminologies describe the "aboutness" of documents, the surface appearance of embedded concepts [Navigli and Velardi 2004]. Such terminologies may consist of single-word terms such as *ice* or *water*, or multi-word phrases such as *kyoto protocol* (noun compound), *department of meteorology* (prepositional phrase) and *global warming* (adjective-noun phrase).

Keyword analysis locates words in a given text and compares their frequency with a reference distribution from a usually larger corpus of text. A chi-square test of significance with Yates' correction for continuity, for example, can be used to determine over-represented terms and list them in order of decreasing significance. Extending the keyword algorithm, the *term co-occurrence module* used for this research relies on a pattern matching algorithm based on regular expressions to identify text fragments frequently appearing within the same sentences or documents. When formulating regular expressions, analysts have to enumerate common inflections of a term while excluding general terms with multiple meanings.

Co-occurrence analysis assumes that two semantically related terms regularly co-occur in the same text segments. This research uses the *Log Likelihood Algorithm* [Dunning 1994] to analyze the significance of co-occurrence with the target term at both the sentence level and the document level. A term frequency threshold filters rare words with less than five occurrences in the reference corpus, after a part-of-speech tagger limits the consideration set of terms to nouns.

4 Generating the Concept Hierarchy

Co-occurrence analysis is a popular method in corpus-based analysis. It establishes whether retrieved terms are related, but cannot tell *how* they are related. Statistical lexical analysis is therefore often criticized as "knowledge poor" [Grefenstette and Hearst 1992]. Moving towards a detailed semantic analysis – e.g., determining the hierarchical relation of two terms – is far from trivial. The following sections review reported heuristics for identifying hypernyms and building concept hierarchies [Caraballo 1999, Joho and Sanderson 2000, Joho et al. 2004, Barriere 2005], and describe the suggested spreading activation approach. [Fig. 2] shows the seed ontology on CLIMATE CHANGE, which represents the basis for all subsequent computations.

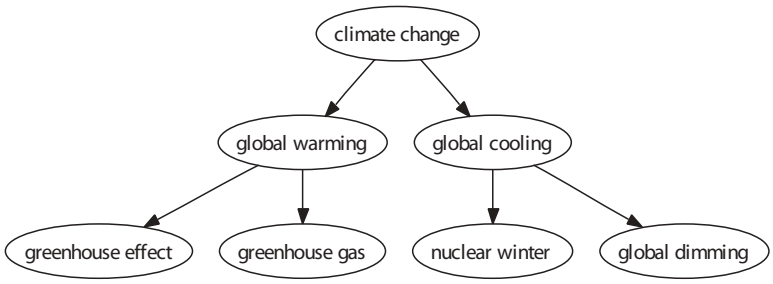

Figure 2: Seed ontology on CLIMATE CHANGE

4.1 Hierarchical Relationship Discovery

The *subsumption approach* [Sanderson and Croft 1999] automatically generates concept hierarchies by assuming that if two terms co-occur, general terms should co-occur more frequently than specific terms. This implies that the documents containing specialized terms are a subset of the documents containing general terms. Given two terms *x* and *y*, *x* is said to subsume *y* if the following condition holds:

$$P(x|y) \geq 0.8 \text{ and } P(y|x) < 1$$

A value of 0.8 was chosen through informal analysis of hypo-/hypernym pairs identified through subsumption analysis in order to relax the initially strong condition $P(x|y) = 1$ (term *x* occurs whenever term *y* occurs).

The alternative *trigger phrase approach* [Joho and Sanderson 2000] relies upon the heuristic that certain common phrases (e.g. SUCH AS, AND OTHER, INCLUDING, etc.) often link hypo-/hypernym pairs. Regular expressions can identify such trigger phrases to suggest possible semantic relations between the terms linked by the phrase. Similarly, head nouns in multi-word phrases (e.g. DIOXIDE in CARBON DIOXIDE) often super-ordinate the containing phrases [Joho et al. 2004].

4.2 Spreading Activation over Weighted Graphs

Spreading activation is a search technique inspired by the human brain's cognitive models, where neurons fire activations to adjacent neurons. Connectionist (as opposed to symbolic) artificial intelligence often uses spreading activation for retrieving hidden network information. Spreading activation is also widely used in associative information retrieval [Crestani 1997]. A spreading activation design involves the creation of a *network data structure*, and selecting the *processing technique*. The network structure typically consists of nodes connected by weighted links.

The methods outlined in [Section 4.2.1] generate a semantic network and connect its nodes via annotated links according to the specific type of the analysis. Weights are calculated based on the link types [Section 4.2.2]. Terms that acquire high activation levels via *spreading activation* are then suggested to the domain expert as candidate terms to extend the ontology.

4.2.1 Establishing the Network Data Structure

The current system assigns a term or a concept to each node. A *concept* is obtained after disambiguation when consulting WordNet for the correct sense of a *term*. Each link indicates a directed relationship between a pair of terms. Four types of analysis are carried out to discover candidate terms to add to the seed ontology:

(i) Co-occurrence analysis. For each seed term, the system determines co-occurring terms at the sentence and document level, ranked according their significance. A significance threshold determines whether to include the identified term.

(ii) WordNet hyponyms, hypernyms and synonyms. Each seed term is first disambiguated into *concepts* using *vector space models* [Bernstein 2005], representing concepts as vectors of features in a k-dimensional space and computing the cosine or Euclidean distance as a similarity measure. Each seed term vector consists of selected co-occurring terms. Each WordNet sense of a seed term is represented by a vector of relevant keywords found through WordNet. The seed term vector is then measured against WordNet sense vectors to select the most plausible sense. After sense disambiguation, WordNet provides hyponyms, hypernyms and synonyms of the seed terms.

(iii) Trigger phrase analysis. Regular expressions identify likely synonyms and hyponyms based on heuristic rules. In the sentence "*Methane is a greenhouse gas"*, for example, the trigger phrase "IS A" probably connects a hypo-/hypernym pair. Suggested synonyms and hyponyms are then incorporated in the semantic network.

(iv) Head noun analysis. Head nouns that often subsume noun compounds are added to the network as potential hypernym.

Combining the methods described above yielded a semantic network comprising more than 1,200 nodes connected via annotated links – the link types include *co-occurrence*, *trigger phrase* {*hyponym, hypernym, synonym}*, *wordnet* {*hyponym, hypernym, synonym}*, *co-occurrence significance*, *hypernym of the original hierarchy* and *head noun*. The partial view of the associations shown in [Fig. 3] illustrates the magnitude of the semantic network (generated with the *IsaViz RDF Authoring Tool;* www.w3.org/2001/11/IsaViz).

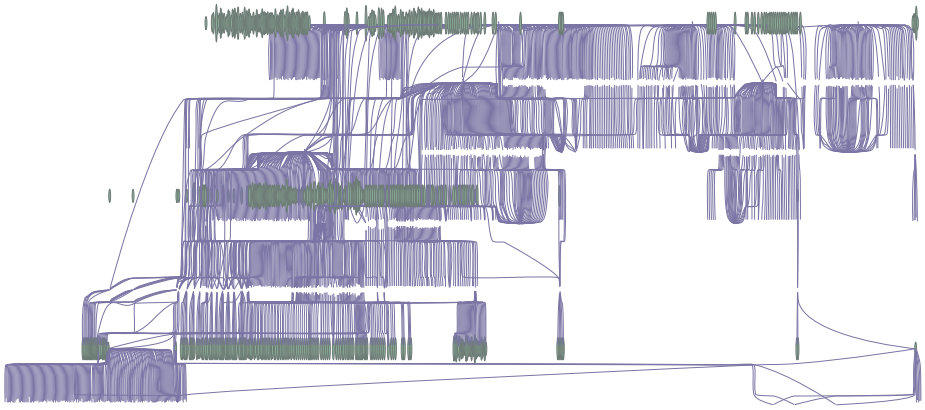

Figure 3: Partial view of the semantic network

4.2.2 Spreading Activation Processing over the Semantic Network

Data processing consists of iterations that contain a set of pulses including checks for termination conditions. In each pulse, the system distinguishes between three phases: pre-adjustment, spreading and post-adjustment. The pre- and post-adjustment phases are optional, typically used to avoid energy retention, and to control the activation of single nodes or the entire network [Crestani 1997]. In its simplest form, the activation level of a node in spreading activation is determined trough the following formula:

$$l_j = \sum o_i \omega_{ij} \quad (i = 1, \ldots, k \text{ for all nodes connected to node } j)$$

where

l_j is the activation level of node j, calculated as the total input to node j;
o_i is the output of unit i connected to node j;
ω_{ij} is a weight associated to the link connecting i and j.

Spreading activation is used over the augmented semantic network for identifying the network's most relevant keywords. ω_{ij} in our implementation represents a trustworthiness value between 0.0 (lowest) and 1.0 (highest) for each type of analysis.

Terms of the seed ontology receive the highest value of $\omega_{ij} = 0.9$, because they are either specified by a domain expert or taken from previously validated ontologies. Results from the head noun analysis also receive a value of $\omega_{ij} = 0.9$, as this linguistic rule performed exceptionally well according to domain experts.

A low value of $\omega_{ij} = 0.2$ is assigned to the trigger phrases, as these heuristics rules are not always effective. *"Likely"*, for example, can identify synonyms while *"cause"* suggests a hyper-/hyponym pair. When both terms occur in the same sentence, however (e.g., "global warming is the *likely cause*"), such rules produce wrong results.

WordNet terms also receive a value of $\omega_{ij} = 0.2$. The meaning of trustworthiness takes a slight turn here – the low value does not indicate skepticism regarding the correctness of WordNet, but limits its influence to promote the inclusion of domain-specific terms in the resulting semantic network.

For the terms identified through co-occurrence analysis, a range of $[\omega_{min}, \omega_{max}]$ with $\omega_{min} = 0.3$ and $\omega_{max} = 0.6$ is obtained based on linear adjustment over the significance values:

$$\omega_{ij} = \omega_{min} + (\omega_{max} - \omega_{min}) \cdot (\eta^{i}_{max} - \eta_{ij}) / \eta^{i}_{max}$$

where

η^{i}_{max} is the maximum significance of all co-occurring terms for target term i.
η_{ij} is the significance value of term *j* when term *i* is the target term.

To facilitate the selection of most active concepts and minimize computational requirements, a heuristic rule of $2 \cdot (n-1)$ restricts the number of terms selected after each activation pulse. Here *n* is the number of nodes in the seed ontology. The first spreading activation run over the network shown in [Fig. 3] resulted in the following most activated nodes: KYOTO PROTOCOL (1.26), CARBON DIOXIDE (1.24), CLIMATE (1.20), GREENHOUSE (1.19), WARMING (1.18), EMISSIONS (1.16), KYOTO (1.12), GAS EMISSIONS (1.10), CARBON (1.10), GASES (1.10), DIOXIDE (1.08), SCIENTISTS (1.06).

Comparing the activation levels with the results of the initial co-occurrence analysis outlined in [Section 3], the spreading activation results favor document-level over sentence-level associations. The analysis shows that less convincing candidate terms are replaced by more relevant alternatives because of the incoming energy they receive from other sources such as trigger phrase analysis and WordNet.

4.2.3 Confirming Semantic Relationships

Hierarchically positioning the most activated terms – i.e., those highly relevant to the domain and seed ontology – is the most challenging task. We propose the following steps: (i) accept semantic relations confirmed by WordNet and head noun; (ii) remove modifiers of a noun phrase that also appear in the activated list, as they do not represent the term's core meaning. CLIMATE and GREENHOUSE, for example, can be removed; (iii) trigger another round of spreading activation using the non-confirmed terms as seed terms to identify appropriate nodes for attaching these terms; use subsumption analysis to determine the type of relationship; (iv) consult domain experts for suggestions on how to position the remaining terms.

[Fig. 4] shows the extended ontology after two iterations of spreading activation. Such a semi-automated process cannot generate a completely accurate positioning of all activated terms in a complex network of hyper- and hyponym relations. Outliers include the terms SCIENTISTS, EPA (Environmental Protection Agency) and ENS (Environment News Service). Although our approach specifically targets hierarchical relationships, the complexity of natural languages and the lack of contextual meaning in co-occurrence analysis inevitably lead to the inclusion of relevant but not hierarchically related terms. Unidentified relations are labeled “relations”. Taking into account that hierarchical relationships only represent a small subset of an ontology’s possible relationship types, domain experts will specify additional relationship types such as "working in the domain of" (SCIENTISTS → GREENHOUSE GAS).

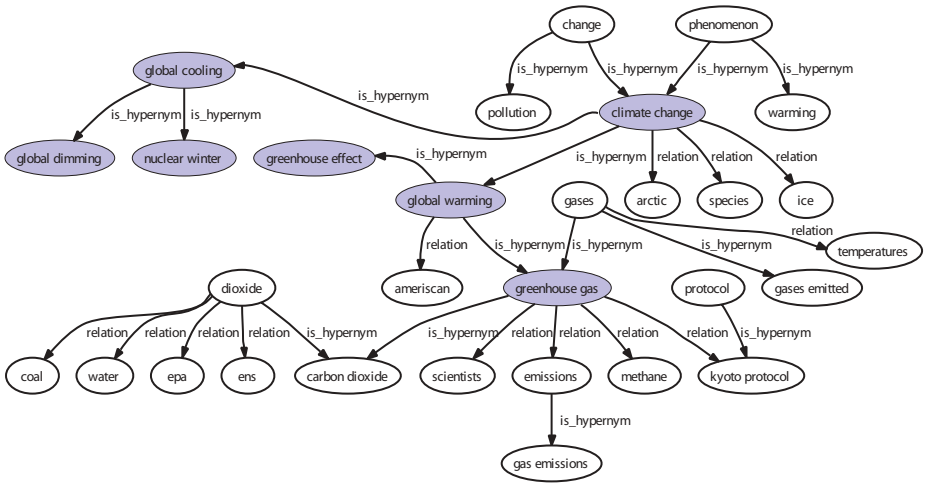

Figure 4: Concept hierarchy after two rounds of spreading activation

5 Conclusions

Heuristic rules common in text mining and natural language processing face a number of shortcomings in automatically discovering semantic relationships between domain concepts. Attempts to semi-automatically extend and refine domain-specific ontologies call for a more fine-grained processing of textual data. Based on a corpus gathered from a large sample of news media sites, a novel approach to finding semantic concept relations using spreading activation on weighted graphs has been proposed and combined with existing heuristics.

Future research will attempt to parameterize the ω_{ij} values through machine learning. Using parts of a validated semantic structure as seed ontology, an adaptive process could adapt ω_{ij} values based on the similarity between the validated structure and the extended ontology suggested by the system. This method would optimize results within a given domain. Applying this method to several domains should allow formulating general strategies to extend ontologies without prior domain knowledge.

Acknowledgements

This work is an initiative of the *Research Network on Environmental Online Communication* (www.ecoresearch.net) and was funded by several research grants of the *University of Western Australia* (www.uwa.edu.au). The *Interactive Virtual Environments Centre* (www.ivec.org) has kindly provided the high performance computing infrastructure for running the experiments.

References

[Barriere 2005] Barriere, C.: "Building a Concept Hierarchy from Corpus Analysis", Terminology, 10, 2 (2005), 241-263.

[Bernstein 2005] Bernstein, A., Kaufmann, E., Burki, C. and Klein, M.: "How Similar Is It? Towards Personalized Similarity Measures in Ontologies", 7th International Conference Wirtschaftsinformatik (WI-2005), Bamberg, Germany, (2005).

[Caraballo 1999] Caraballo, S. A.: "Automatic Acquisition of a Hypernym-Labeled Noun Hierarchy from Text", 37th Annual Meeting of the Association for Computational Linguistics, (1999), 120-126.

[Crestani 1997] Crestani, F.: "Application of Spreading Activation Techniques in Information Retrieval", Artificial Intelligence Review, 11 (1997), 453-482.

[Dunning 1994] Dunning, T.: "Accurate Methods for the Statistics of Surprise and Coincidence", Computational Linguistics, 19, 1 (1994), 61-74.

[Fellbaum 1998] Fellbaum, C.: "WordNet An Electronic Lexical Database", Computational Linguistics, 25, 2 (1998), 292-296.

[Grefenstette and Hearst 1992] Grefenstette, G. and Hearst, M. A.: "A Method for Refining Automatically-Discovered Lexical Relations: Combining Weak Techniques for Stronger Results", AAAI Workshop on Statistically-based Natural Language Programming Techniques, AAAI Press, Menlo Park, CA, (1992), 64-72.

[Joho and Sanderson 2000] Joho, H. and Sanderson, M.: "Retrieving Descriptive Phrases from Large Amounts of Free Text", 9th International Conference on Information and Knowledge Management, (2000), 180-186.

[Joho et al. 2004] Joho, H., Sanderson, M. and Beaulieu, M.: "A Study of User Interaction with a Concept-based Interactive Query Expansion Support Tool", Advances in Information Retrieval, 26th European Conference on Information Retrieval, (2004), 42-56.

[Navigli and Velardi 2004] Navigli, R. and Velardi, P.: "Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites", Computational Linguistics, 30, 2 (2004), 151-179.

[Roussinov and Zhao 2004] Roussinov, D. and Zhao, J. L.: "Automatic Discovery of Similarity Relationships through Web Mining", Decision Support Systems, 35, (2003), 149-166.

[Sanderson and Croft 1999] Sanderson, M. and Croft, W. B.: "Deriving Concept Hierarchies from Text", 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, USA, (1999), 206-213.

[Uschold and Grüninger 1996] Uschold, M. and Grüninger, M.: "Ontologies: Principles, Methods, and Applications", Knowledge Engineering Review, 11, 2 (1996), 93-155.