Capturing and Classifying Ontology Evolution in News Media Archives

Albert Weichselbraun
Vienna Univ. of Economics
1090 Vienna, Austria
aweichse@ai.wu-wien.ac.at

Arno Scharl
MODUL Univ. Vienna
1190 Vienna, Austria
scharl@modul.ac.at

Wei Liu
Univ. of Western Australia
Crawley WA 6009, Australia
wei@csse.uwa.edu.au

Abstract

Ontology evolution is an intrinsic phenomenon of any knowledge-intensive system, which can be addressed either implicitly or explicitly. This paper describes an approach to capture and visualize implicit data-driven ontology evolution using ontologies semi-automatically generated by extending small seed ontologies. This process captures ontology changes reflected in large document collections. Visualizing of these changes helps characterize the evolution process, and distinguish core, extended and peripheral relations between concepts. Finally, the paper presents an example of ontology evolution by monitoring and analyzing online media coverage on "energy sources" over a period of ten months.

1 Introduction

Ontologies are a key technology for the Semantic Web. By formally describing vocabularies and business processes, they provide the means for a common understanding. Domain knowledge evolves continually in dynamic environments, requiring regular updates of the underlying ontologies. *Ontology evolution* refers to "the process of adaptation of an ontology to the arisen changes in a domain while maintaining both the consistency of the ontology itself as well as the consistency of depending artifacts" [11].

Related fields like software and database engineering provide valuable methods towards addressing certain aspects of ontology evolution. As an inherently different phenomenon, however, ontology evolution cannot be described by these approaches completely. Noy and Klein [10] pointed out that the traditional distinction between versioning and evolution is not applicable to ontology evolution. This is due to the crucial differences between ontologies and database schemata - ontologies are themselves data, they incorporate semantics, are more often reused, offer richer data models and are de-centralized by nature. Based on this insight, Noy and Klein [10] distinguish between (i) information preserving changes, (ii) translatable changes, and (iii) information-loss changes, where the preservation of all instance data cannot be guaranteed.

Ontologies can be defined as a formal specification of a conceptualization of a domain [5]. Based on this definition, Flouris [4] and Noy [10] identify three possible causes for ontology evolution: (i) changes in the domain, (ii) modifications to the conceptualization of the domain and (iii) changes in the specification (for instance the use of another ontology language).

The last factor is dealt with by ontology translation. Ontology evolution leads to new ontology versions with possible differences in (i) terminology, (ii) scope, (iii) encoding and (iv) context [14].

Stojanovic et al. [12] distinguish between (i) *explicit*, *usage driven* changes, where changes are incorporated into the ontology due to a user's/ontology engineer's request and (ii) *implicit*, *data-driven* changes, reflecting changes in the system described by the ontology. Such changes have to be extracted by analyzing the affected system.

In this research we focus on *data-driven* changes, induced by the evolution of the usage and the importance of vocabulary. Changes are detected using our ontology extension system through analyzing text collected from 150 online media sites between November 2005 and August 2006.

Section 3 introduces the ontology extension architecture that is used to build ontologies automatically based on a given seed ontology. Section 3 then looks into the temporal aspects of ontology evolution by using this architecture to generate different versions of the same ontology at different points in time. Section 3.2 discusses our observations when using highly volatile and independent online media as a source for tracking ontology changes. The paper concludes with a summary and outlook in Section 4.

2 Related Work

Ontology integration identifies similarities and differences between different versions of an ontology to create a new ontology that resolves these problems. In a first

step, an alignment of the ontologies is necessary. This can be done manually or semi-automatically, as for instance demonstrated in Castano et al. [3]. It is important to note that it is not possible to automatically determine whether changes occur on a conceptual level or in the specification [6]. There are many steps in the ontology evolution process, where the ontology development system may require user input to correctly interpret suggested changes. Stojanovic et al. [12] propose that such situations can be resolved by *evolution strategies* containing definitions for the handling of such resolution points.

Antoniou and Kehagias [2] suggest the notion of conservative extensions, guaranteeing that changes to the ontology remain local. Other authors [6] suggest the use of a *versioning schema*, inspired by software library versioning to deal with different ontology versions.

Plessers [11] introduces an approach for tracking ontology changes by using a *version log* keeping track of the changes of the ontology's representation, and an *evolution log* recording changes to the interpretation of the ontology. Klein and Noy [7] define an ontology for basic change operations. They suggest the use of heuristics to combine simple changes to more complex ones, which are more useful for specification of the consequences of a change.

Luong and Dieng-Kuntz [9] address explicit changes of the domain vocabulary by detecting the evolution of semantic annotations using a rule-based approach.

Stojanovic et al. [13] present an approach to model ontology evolution as a *reconfiguration-design problem*. They define graphs mapping different ontology evolution paths. Ontology evolution is reduced to a graph search where the nodes are evolving ontologies and the edges represent changes transforming the source ontology into the next node. Another approach suggested by Flouris [4] applies the principles of belief change to ontology evolution. The authors apply the ADM paradigm [1] to ontology evolution.

3 Ontology Evolution

This paper applies a methodology introduced by Liu et. al [8] for semi-automatically extending ontologies. The approach combines natural language processing techniques and spreading activation to extract domain concepts and relations from text documents covering the domain. The architecture uses a small set of seed terms from domain experts or from known ontology repositories to initialize the extension process.

Ontology evolution focuses on temporal phenomenon observed on concepts, relations and the ontology as a whole. The ontology extension architecture described above extends the seed ontology, based on the terminology and relations in the target corpus. Changes in the corpus are therefore directly reflected in the extended ontology.

This research studies 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.

Based on this sample, the ontology extension component built ontologies from corpora created between November 2005 and August 2006, reflecting the evolution of the domain's vocabulary - i.e., the rise and fall of the importance of domain concepts and their relations.

3.1 Types of Evolution

The architecture described above provided us with the means to monitor data-driven changes to ontologies more effectively. An ontology comprises a *conceptualization* of a given domain, and a *language* to express the entities and their relations in these domain. Explicit changes in the domain directly translate into modifications of the ontology's language and/or its conceptualization. It is important to note that this research only comprises changes to domain language. *Automatically detecting* modifications of domain *conceptualization* is far beyond the capabilities of today's natural language processing techniques.

Building upon the approaches described in Section 2, we define three levels at which the domain vocabulary referring to concepts and relations may evolve: *core*, *extended* and *peripheral*. Applying these levels to the *evolution of concepts* yields three types of terminologies:

(i) The *core domain terminology* comprises frequently used concepts which are constantly included into the domain's ontology, showing only small variations in their importance to the domain. Terms referring to these concepts are considered keywords for the given domain.

(ii) The *extended domain terminology* contains additional domain concepts of lower importance. These concepts are used for special topics within the domain (e.g. `nuclear power`, `oil prices`) but are not as universally used as the core domain terminology. As media coverage on these special topics increases (decreases), these concepts get included in (excluded from) the domain ontology.

(iii) *Peripheral terminology* is used in domain documents, but does not carry important domain concepts. These terms are therefore not included in the domain ontology.

One important factor, controlling whether a particular concept gets included into the core domain terminology or in the extended domain terminology is the ontologies' granularity. The more concepts an ontology comprises the more complex it gets, and the more concepts are considered part of the *core domain terminology*.

Corresponding to the distinction above, we can define

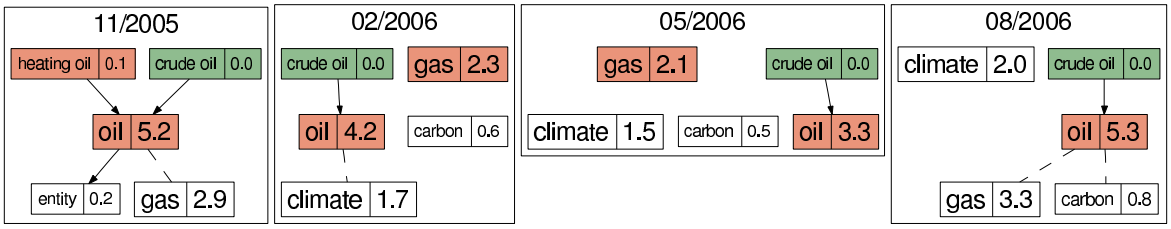

Figure 1. Evolution of the concept "oil" from November 2005 to August 2006.

the *evolution of relationships* distinguishing between *core domain relations* featuring essential relations between core domain vocabulary, *extended domain relations* comprising relations to extended domain vocabulary as well as non-essential relations between the core vocabulary, and *Peripheral domain relations* which do not carry enough weight to be included into the ontology.

Changing *relation types* is an important aspect of ontology evolution. Our current architecture detects three named relation types (hypernyms-hyponyms, synonyms and modifiers) and assigns unnamed relations to all other relations. This separation is rather coarse, but suffices to observe and classify changes in relation types.

We apply the definitions introduced in Section 3.1 to the evolution of single concepts through the evolutionary process. Figure 1 visualizes the evolution of the concept `oil` from November 2005 to August 2006 (font sizes and numbers indicate the concepts' importance in terms of its average frequency; arrows indicate hyponym-hypernym relationships and dashed lines unnamed relations between the concepts). Loose concepts have been identified as relevant and relate to the term `oil` in one of the other snapshots.

Based on the definitions in Section 3.1 we can identify the following change patterns for the term `oil` in the energy domain:

(i) The terms `oil`, `crude oil` and `gas` belong to the ontology's core domain terminology. The stable relation between `crude oil` and `oil` is a core domain relation, while all other visible relations are considered extended domain relations.

(ii) At the chosen granularity level, the concepts `entity`, `climate`, `carbon` and `gas` belong to the extended domain terminology, because these term's relative importance is too low for being included into all versions of the evolving ontology.

(iii) Examples for peripheral terminology include terms like `management`, `states` and `capital`, but also seemingly relevant concepts like `environment`, `greenhouse gases` and `wind turbines`. Imposed restrictions on the ontology's complexity prevented the inclusion of these terms.

3.2 Results and Discussion

When analyzing online media articles, one encounters heterogeneous textual material. Analysts are not dealing with *one* authoritative usage of the domain vocabulary, but with many articles, published in various newspapers and written by different authors which might use certain terms in slightly different ways. Therefore, the assembled domain ontology always produces a composite ontology, reflecting the terms' *average* usage and importance in all analyzed media.

As topics and coverage in online media change, the underlying ontological assumptions are modified. Online media with their highly volatile vocabularies are an excellent source for investigating ontology evolution. Given a seed ontology comprising seven basic concepts, the automated ontology extension process (Section 3) yields the ontology visualized in Figure 2, reflecting the changes in the domain during the observed period of time (seed ontology = concepts with darker shading). The following effects have been observed during our experiments:

(i) *Changes in a term's importance.* As the focus of media coverage shifts, certain topics and the associated vocabulary are used less frequently in online media. Therefore the concepts weights will be reduced or might even be removed alltogether (depending on the ontology's granularity).

(ii) *Change of the assigned concept.* Such changes might be very obvious or hardly noticeable, depending on the following distinction:

(ii.a) Change in *term focus.* Articles might use the term `oil` to cover the concepts `crude oil`, `gas` and `petrol` in one month, while only `gas` and `petrol` are relevant in another one. During the extension process these changes will be notable by the concepts attached to the *target concept.* According to Figure 1, media coverage in November 2005 used the term `oil` in conjunction with the terms `heating oil`, `crude oil` and `gas`. In the following months, the concept `heating oil` decreases in popularity, making way for other relations between the concept `oil` and its neighbor terms.

(ii.b) Change in *term assignment.* Back in the 60s, the term `fuel` in the context of cars referred to the concept

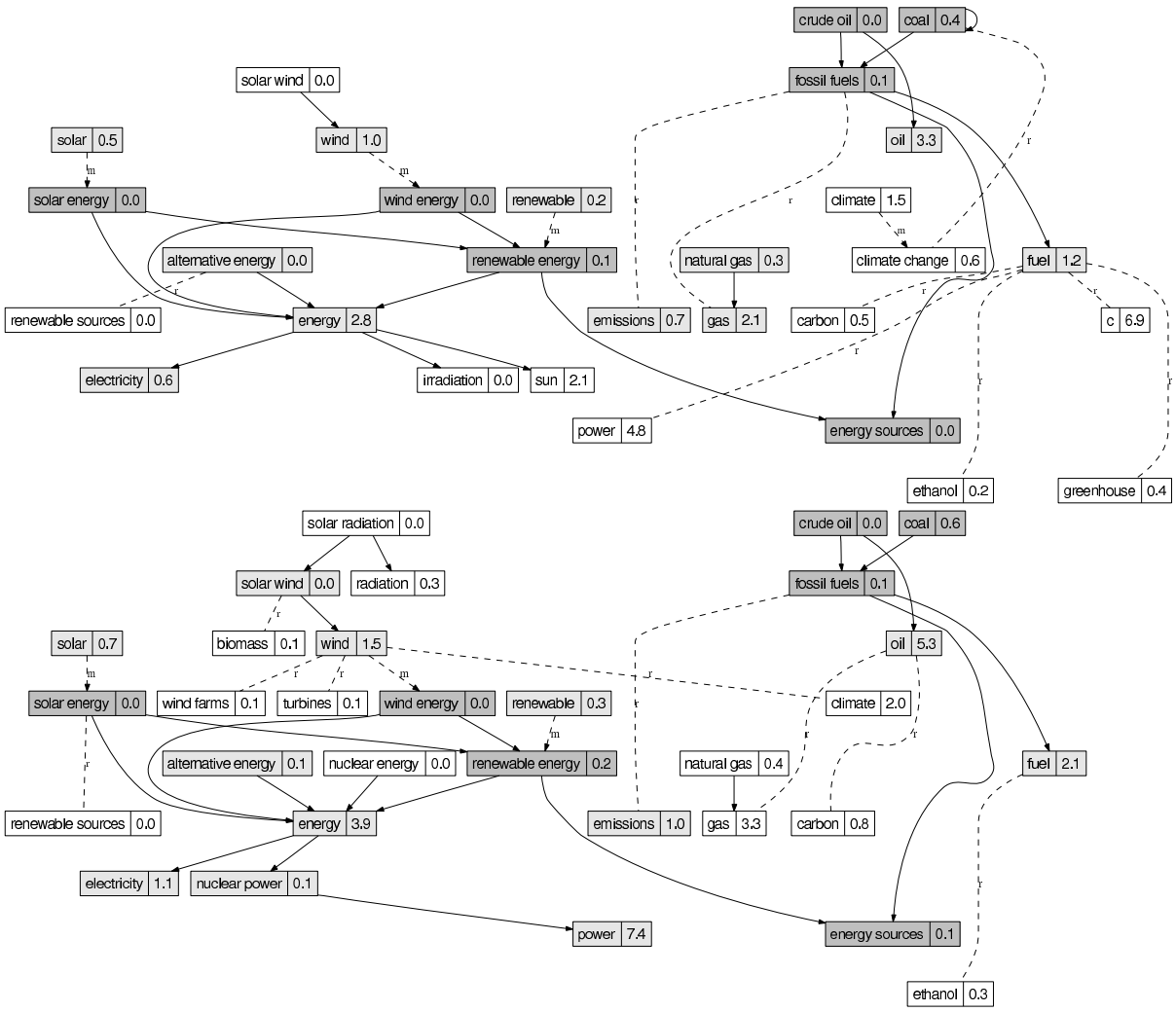

Figure 2. The extended ontologies between May 2006 and August 2006.

petrol. Nowadays, a significant share of automobiles use diesel engines or alternative fuels, which leads to a changed meaning of the term.

(ii.c) Change in *context*. Sometimes the meaning of a concept might remain stable, but the context of media coverage drastically changes. In Europe most articles concerning countries like Sri Lanka or the Maldives had been published in a tourism context. This changed drastically after these popular tourism destinations had been hit by a Tsunami in 2004. While the concepts remained identical, the context of media coverage has changed completely. In the underlying ontology, such an event is reflected by the replacement of a significant share of the neighbor concepts. Terms like beach and holidays are replaced by terms linked to the catastrophe. A similar shift of context can also be observed if homonyms or homographs are used in different meanings in media coverage.

From the effects described above, changes in a concept's importance as well as minor shifts in a concept's focus have been observed in our study. Due to the relatively short timespans involved, we were not able to observe changes of a concept's meaning. Shifts in a concept's context can be tracked by the current architecture, provided that media data before and after the change are available.

The size of the news media archive and the availability of computational resources resulted in certain restrictions. To correctly interpret the presented results, therefore, it is important to note that the current implementation computes a concept's weight based on the total concept counts in the corpus, regardless of the context. This leads to higher counts for terms reflecting multiple concepts (e.g., power). Follow-up studies should address this shortcoming by incorporating sentence-based word sense disambiguation. Applying disambiguation techniques will also

address other issues like considering salience.

Term selection reflects the trade off between completeness (leading to complexity) and conciseness of the ontology. Concepts such as water disappear from the ontology during the evolution process. This *does not* mean that they are no longer relevant to the domain, but reflects a shift in priorities and interests of the media. Choosing a fine-grained visualization of the ontology with more concepts, some of these concepts would be re-selected for inclusion.

4 Outlook and Conclusions

Ontology evolution is an intrinsic phenomenon of knowledge-intensive systems that happens either implicitly or explicitly. This paper describes the use of an automatic system to track changes in the ontologies used by online media, which are an excellent data source for this type of investigation given their independent and often volatile content production. The automated system generates extended ontologies based on small seed ontologies by analyzing text corpora. The system is used to analyze online media data collected in a specific period of time to generate different versions of ontologies in order to reveal the changes in the context and the importance of the ontology.

Due to the sheer volume of the data to be analyzed, identifying change requires an ontology extension system that adds concepts automatically. We defined three levels to classify the importance of domain concepts and relations: core, extended and peripheral. This classification helps clarify the observed movement of the concepts and relations during the specified period. The presented approach allows identifying changes in the ontology context, and in the importance of ontology concepts and relations.

This paper used visual techniques to represent ontologies and enable human experts to manually identify the changes. In addition to the visual format, future research will use formal language to represent ontologies such that changes in both concepts and relations can be automatically detected and reasoned.

Acknowledgment

The research projects AVALON (Acquisition and Validation of Ontologies; kmi.tugraz.at/avalon) and RAVEN (Relation Analysis and Visualization for Evolving Networks) are 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).

References

[1] C. E. Alchourrón, P. Gärdenfors, and D. Makinson. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50(2):510–530, 1985.

[2] G. Antoniou and A. Kehagias. On the refinement of ontologies. *International Journal of Intelligence Systems*, 15:623–632, 2000.

[3] S. Castano, A. Ferrara, and S. Montanelli. A matchmaking-based ontology evolution methodology. In *Proc. of the 3rd CAiSE INTEROP Workshop On Enterprise Modelling and Ontologies for Interoperability (EMOI - INTEROP 2006)*, Luxembourg, June 2006.

[4] G. Flouris, D. Plexousakis, and G. Antoniou. Evolving ontology evolution. In *Proceedings of the 32nd Int. Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM 06)*, Merin, Czech Republic, 2006.

[5] T. R. Gruber. Toward principles of the design of ontologies used for knowledge sharing. *International Journal of Human and Computer Studies*, pages 907–928, 1995.

[6] M. Klein, D. Fensel, A. Kiryakov, and D. Ognyanov. Ontology versioning and change detection on the web, 2002.

[7] M. Klein and N. Noy. A component-based framework for ontology evolution, 2003.

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

[9] P.-H. Luong and R. Dieng-Kuntz. A rule-based approach for semantic annotation evolution. *The Computational Intelligence Journal*, 23(3):320–338, August 2007.

[10] N. F. Noy and M. Klein. Ontology evolution: Not the same as schema evolution. *Knowledge and Information Systems*, 6(4):428–440, July 2004.

[11] P. Plessers. *An Approach to Web-based Ontology Evolution*. PhD thesis, Vrije Universiteit Brussel, Departement Computer Wetenschappen, Web & Information System Engineering, 2006.

[12] L. Stojanovic, A. Maedche, B. Motik, and N. Stojanovic. User-driven ontology evolution management. In *EKAW '02: Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management.*, pages 285–300, London, UK, 2002. Springer-Verlag.

[13] L. Stojanovic, A. Maedche, N. Stojanovic, and R. Studer. Ontology evolution as reconfiguration-design problem solving. In *K-CAP '03: Proceedings of the 2nd international conference on Knowledge capture*, pages 162–171, New York, NY, USA, 2003. ACM Press.

[14] G. Wiederhold. An algebra for ontology composition. *Proceedings of 1994 Montery Workshop on Formal Methods*, pages 56–62, 1994.