Using an Ontology Learning System for Trend Analysis and Detection

Gerhard Wohlgenannt and Stefan Belk and Matyas Karacsonyi and Matthias Schett

Vienna Univ. of Economics and Business, Welthandelsplatz 1, 1200 Wien, Austria
{gerhard.wohlgenannt,stefan.belk,matyas.karacsonyi}@wu.ac.at
http://www.wu.ac.at

Abstract. The aim of ontology learning is to generate domain models (semi-) automatically. We apply an ontology learning system to create domain ontologies from scratch in a monthly interval and use the resulting data to detect and analyze trends in the domain. In contrast to traditional trend analysis on the level of single terms, the application of semantic technologies allows for a more abstract and integrated view of the domain. A Web frontend displays the resulting ontologies, and a number of analyses are performed on the data collected. This frontend can be used to detect trends and evolution in a domain, and dissect them on an aggregated, as well as a fine-grained-level.

Keywords: trend detection, ontology evolution, semantic technologies, ontology learning

1 Introduction

Ontologies are a cornerstone technology of the Semantic Web. As the manual construction of ontologies is expensive, there have been a number of efforts to (semi-)automatic ontology learning (OL). The demo application builds upon an existing OL system, but extends the system to apply it as a Web intelligence, resp. a trend detection, tool.

As the system generates lightweight domain ontologies from scratch in regular intervals (ie. monthly), the starting point is always the same. This allows meaningful comparisons between ontologies, allowing to trace ontology evolution and general trends in the domain. The system captures an abundance of data about the ontologies in a relational database, from high-level to low-level (see below), which helps to analyze and visualize trends. The OL system generates ontologies from 32 heterogeneous evidence sources, which contain domain data from the respective period of time, so we can not only analyze the resulting ontologies but trace which sources support which ontological elements.

In summary, we use Semantic Web technologies as a Web intelligence tool by extending the system with visual and analytic components for trend detection. Trend detection is an major issue in a world that is changing rapidly. Timely detection of trends (and reaction to them) is important in many areas, eg. for success in business [2].

2 The Underlying Ontology Learning System

This section gives a brief introduction to the ontology learning (OL) system, as well as the sources of evidence used. We try to be as brief as possible, and include only information crucial to understand the trend detection application (for more details see the related work section and the referenced literature).

All trend detection analyses described in the following are based on a specific system for OL and ontology evolution. The system learns lightweight ontologies, more precisely taxonomies plus unlabeled non-taxonomic relations, from heterogeneous input sources. At the moment we use "climate change" as our test domain, and generate ontologies in monthly intervals. As the framework learns from scratch, it starts with a small seed ontology (two static concepts). For this seed ontology, we collect evidence from the evidence sources, and integrate the data (typically a few thousand terms including their relation to the seed concepts) into a spreading activation network. The spreading activation algorithm selects the 25 (current setting) most important new domain concept candidates. The only step which needs human assessment is a relevance check for the concept candidates done with crowdsourcing. A positioning step integrates the candidates into the existing seed ontology. This concludes the first "stage" of OL. We then use the extended ontology as new seed ontology, and start over. The system halts after three rounds of extension.

As already mentioned, the learning process relies on 32 heterogeneous evidence sources. Most of these sources are very dynamic and therefore well-fit for trend detection. The text-based sources include domain-specific corpora extracted from news media articles (segregated by country of origin), Web sites of NGOs and Fortune 1000 companies, domain-filtered postings from Facebook, Youtube, etc. We use keyword extraction and Hearst-style patterns to collect evidence, i.e. terms and relations. Furthermore, the system queries Social Web APIs (Twitter, Flickr) to get related and terms. We also use a few rather static sources, such as WordNet and DBpedia to help with taxonomy building.

3 Trend Detection and Analysis on Different Levels

Our demo system contains three main areas, namely (i) the ontologies, ie. the monthly snapshots of the domain model, (ii) high-level evolution, which include aggregated analyses on the characteristics of the evidence sources and ontologies, and (iii) low-level evolution, which trace the dynamics of concepts and single evidence on a fine-grained level. The demo portal can be found at `http://hugo.ai.wu.ac.at:5050`, a screencast presentation of the portal is available at `http://ai.wu.ac.at/~wohlg/iswc-demo.mp4`.

3.1 Ontologies

The *Ontologies* menu lists all ontologies computed per computation setting. The computation setting is simply a distinct system configuration. By clicking on an ontology, the system displays detailed information. This includes representations in OWL/Turtle syntax and as graph of the resulting ontology, as well as of intermediary results. A user also finds performance data and the list of concepts by extension level. For a more detailed analysis, one can take a look at all evidence collected and used in the learning process. Multiple viewpoints (by concept, by evidence source, ...) allow investigating the underlying data.

In a nutshell, the *Ontologies* menu facilitates the analysis of trends in the domain both on the level of ontologies and the underlying evidence data.

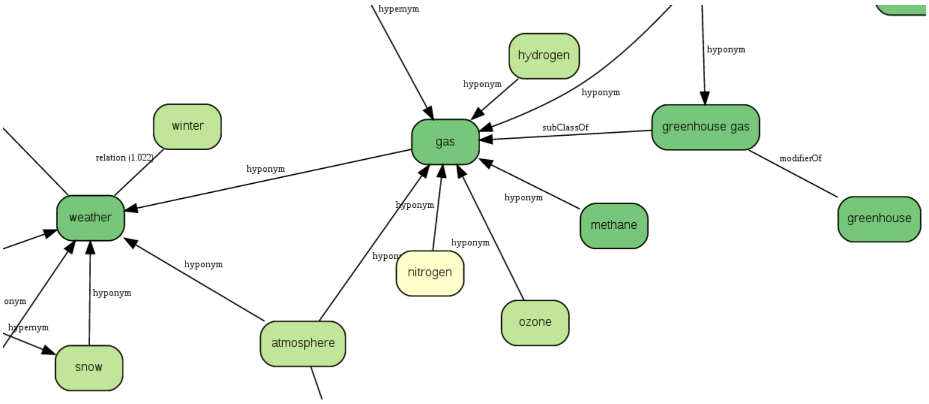

Fig. 1. Example snippet from an ontology (as graph) generated.

3.2 Low-Level Evolution

The *Concept History* shows which concepts have been added and removed from the ontology over time – for a specific system setting. For example, due to media coverage on hurricanes in October 2013 (see also Google trends), the concept hurricane was added to the ontology in November 2013 (in most settings). Entering "hurricane" as concept candidate in the *ECM* analysis presents the fine-grained development of evidence of the concept. Figure 2 shows which sources (US news media, UK news media, etc.) support the concept to what extend.

3.3 High-Level Evolution

The *High-Level Evolution* menu includes tools and visualizations to trace the evolution of evidence sources and the quality of the OL algorithms. For example, the source impact vector (SIV) graph shows the impact of the evidence sources on the system, which is computed according to the observed quality of suggestions from these sources. *Source evolution* displays the evolution of quality of concept candidates suggested by the source.

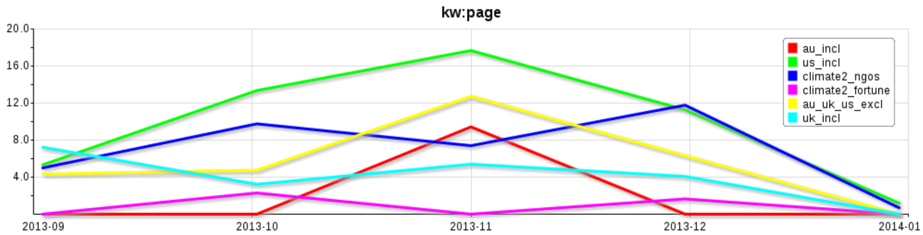

Fig. 2. (Keyword-generated) evidence for concept *hurricane* in various Web corpora (News Media, NGOs Websites, etc.)

4 Related Work

More information about the OL system used as foundation for the trend detection experiments and visualizations can be found in Weichselbraun et al. [3] and Wohlgenannt et al.[4]. A number of approaches have been proposed for trend detection from text data. For example, Bolelli et al. [1] first divide documents into time segments, then detected topics with a latent Dirichlet allocation model, and finally trace the evolution of the topics. In the realm of social media, TwitterMonitor [2] identifies trends on Twitter in real time.

5 Conclusions

The demo application uses Semantic Web (ontology learning) technologies to facilitate trend analysis and detection in a given domain. Users can trace change on different levels, (i) on the level of ontologies themselves, (ii) the aggregated level of quality of the system and impact of evidence sources, and (iii) the fine-grained level on concepts and single evidence. The fine-grained level is especially helpful to determine the reasons for trends in the sources of evidence. Future work will include the implementation of additional analyses and visualizations and the application of the tool in other domains, for example finance and politics.

References

1. Bolelli, L., Ertekin, S., Giles, C.L.: Topic and trend detection in text collections using latent dirichlet allocation. In: Proc. 31th European Conf. on IR Research. pp. 776–780. ECIR '09, Springer-Verlag, Berlin, Heidelberg (2009)
2. Mathioudakis, M., Koudas, N.: Twittermonitor: Trend detection over the twitter stream. In: Proc. of the 2010 ACM SIGMOD Int. Conference on Management of Data. pp. 1155–1158. SIGMOD '10, ACM, New York, NY, USA (2010)
3. Weichselbraun, A., Wohlgenannt, G., Scharl, A.: Refining non-taxonomic relation labels with external structured data to support ontology learning. Data & Knowledge Engineering 69(8), 763–778 (2010)
4. Wohlgenannt, G., Weichselbraun, A., Scharl, A., Sabou, M.: Dynamic integration of multiple evidence sources for ontology learning. Journal of Information and Data Management (JIDM) 3(3), 243–254 (2012)