

OntoGen Semi-automatic Ontology Editor

Blaz Fortuna, Marko Grobelnik, Dunja Mladenic

Department of Knowledge Technologies,

Institute Jozef Stefan, Ljubljana, Slovenia

1. Ontology editors

The rapid growth of documents, web pages and other types of textual content pose a great challenge to modern content management systems. Ontologies offer an efficient way to reduce the amount of information overload by encoding the structure of a specific domain and offering easier access to the information for the users. However, all major ontology editors (such as Protégé[2], OntoStudio[3], ...) are fully manual and offer little support to the users for structuring domains.

OntoGen (previous version was introduced in [1]) is a semi-automatic and data-driven ontology editor focusing on editing of topic ontologies (a set of topics connected with different types of relations). The system combines text-mining techniques with an efficient user interface to reduce both: the time spent and complexity for the user. In this way it bridges the gap between complex ontology editing tools and the domain experts who are constructing the ontology and not necessarily having skills of ontology engineering. The two main characteristics of the system are:

Semi-Automatic – The system is an interactive tool that aids the user during the ontology construction process. It suggests: concepts, relations between the concepts, names for the concepts, automatically assigns instances to the concepts and provides a good overview of the ontology to the user through concept browsing and various kind of visualization. At the same time the user is always full in control of the systems actions and can fully adjust all the properties of the ontology by accepting or rejecting the system's suggestions or manually adjusting them. This lets the user to establish a trust towards the system in a way that he has a full control over all the modifications to the edited ontology.

Data-Driven – Most of the aid provided by the system is based on the underlying data provided by the user typically at the beginning of the ontology construction. The data reflects the structure of the domain for which the user is building an ontology. The system supports automatic extraction of instances (used for forming concepts) and co-occurrences of instances (used for forming relations between the concepts) from the data.

2. Main features

2.1. Interaction with the ontology

The system enables to the user multiple views on the ontology in construction. It supports a tree-view on the ontology, as it is usually intuitive for most users and presents a natural way to represent a topic hierarchy. This view is exposed as a standard Windows control and as a visualization offering a one-glance view of the whole ontology. Each topic from the ontology is exposed by the set of the most

informative keywords for the target topic being automatically extracted using text-mining techniques. Topics are further exposed with relations to other topics in other ontologies and through a topic-map of the documents belonging to the topic. We use Document Atlas [4] tool/component to construct the topic-maps. Figures 1 and 2 show the main parts of the system.

Topic suggestions play a central part in the system. We provide unsupervised and supervised methods for generating suggestions. Unsupervised learning methods automatically generate a list of sub-topics for a currently selected topic by using k-means clustering and latent semantic indexing (LSI) techniques [1, 5] to come up with a list of possible topics. Supervised learning methods on the other hand require the user to have a rough idea about a new topic – this is identified through a query returning the documents. The system automatically identifies the documents that correspond to the topic and the selection can be further refined by the user-computer interaction through an active learning loop [1] (active learning being a machine learning technique for semi-automatic acquisition of the knowledge from the user).

Prototype of the system was already successfully applied in case studies of several commercial projects from the domain of business, legislation and digital libraries. The users participating in the case studies were domain experts with limited experience in ontology construction. The feedback we got from the user was positive and we have used it to further improve the user interface. In particular, the system enabled the users to model ontologies which would be significantly more difficult/expensive to model otherwise.

2.2. Collaborative editing and user profiling

The system also offers collaborative editing of ontologies where the user can use topics and relations which were previously constructed by other users. The system supports the user by suggesting similar topics/relations from the collection of ontologies.

User profiling is also used to tune the human ontology interaction based on the previous work of the users. This is done by recording previous choices that the user made when constructing ontologies and using them as an extra input to provide a personalized view on the underlying data and the ontology constructed so far through “personalized word weighting schemas” (instead of using predefined schemas such as TFIDF [5]).

Acknowledgments

This work was supported by the Slovenian Research Agency and the IST Programme of the European Community under SEKT Semantically Enabled Knowledge Technologies (IST-1-506826-IP), NeOn Lifecycle Support for Networked Ontologies (IST-4-027595-IP) and PASCAL Network of Excellence (IST-2002-506778).

References

- [1] Fortuna, B., Grobelnik, M., Mladenic, D. *System for Semi-automatic Ontology construction*. Proceedings of the 3rd European Semantic Web Conference ESWC-2006, June 11-14, 2006, Budva, Montenegro
- [2] More information on Protégé (Stanford University) and the download of the latest version is available from <http://protege.stanford.edu>.

- [3] Information about OntoStudio (Ontoprise GmbH) is available through http://www.ontoprise.de/content/e1171/e1249/index_eng.html.
- [4] Fortuna, B., Grobelnik, M., Mladenic, D. *Visualization of Text Document Corpus*. Informatica 29 (2005), 497-502.
- [5] Fortuna, B., Grobelnik, M., Mladenic, D. *Background Knowledge for Ontology Construction*. Proceedings of the 15th International World Wide Web Conference WWW 2006, May 23.26, 2006, Edinburgh, Scotland
- [6] C. D. Manning and H. Schütze, *Foundations of statistical Natural Language Processing*, MIT Press.

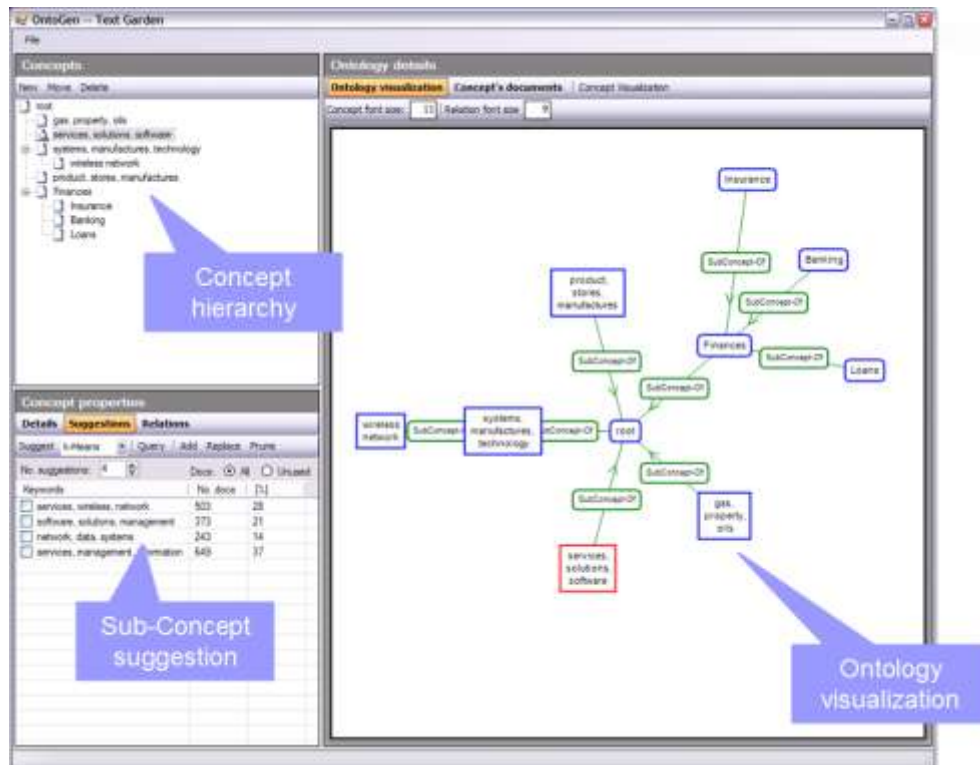


Figure 1 OntoGen interface when the user is getting suggestions for the sub-concepts of the selected concept. The ontology is visualized in textual mode as a concept hierarchy (left upper part) and in graphical mode (right central part).

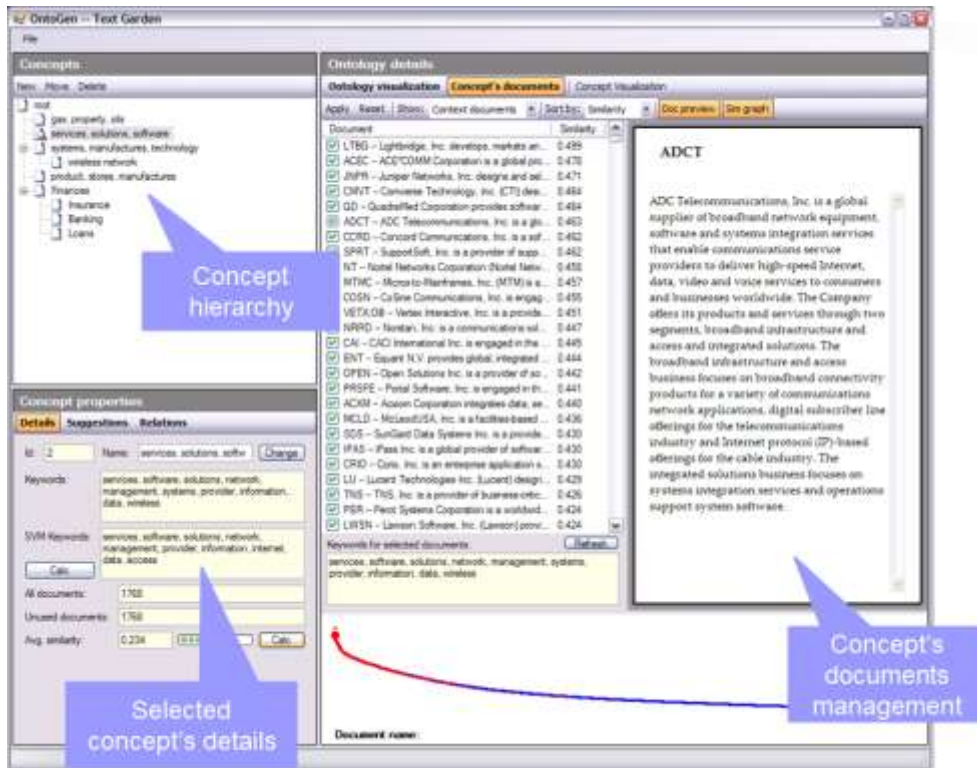


Figure 2 OntoGen interface when the user is investigating the selected concept by browsing through the concept statistics (bottom left), the concept documents including their content (upper right) and the graph of document similarity to the concept (bottom right).