

# Learning and the Semantic Web (Extended Abstract)

Steffen Staab

Institute AIFB, University of Karlsruhe  
sst@aifb.uni-karlsruhe.de

<http://www.aifb.uni-karlsruhe.de/WBS/sst>

## Abstract

The talk will report on the basic idea of the Semantic Web. We will show how the Semantic Web vision may integrate nicely with core capabilities that Machine Learning may provide. As an example we show results from Ontology Learning and Ontology-based Clustering and lay out a preliminary roadmap of further requirements.

This extended abstract is only about sketching the general picture and referring to some (strongly biased choice of) work in this area.

## 1 The Semantic Web

In spite of the many successes and the proliferation of the World Wide Web, many tasks in the Web are still cumbersome and take a lot of time or even remain impossible. For instance, it is nearly impossible to find information with a keyword search when the information is not contained on a single Web page, but rather spread over two or more resources. For instance, on a company intranet one may query for people who know about Data Mining, but the only links available may be between people and projects as well as between projects and project topics. The rule that people in a project know about the topics of the latter cannot be applied to produce useful results.

The vision for the Semantic Web is to repeat the success of the current, syntactic Web by transferring its core principles to machine understandable content rather than “only” human understandable content such that tasks like in the example just mentioned may be solved by a machine agent. Thus, the overall goal of the Semantic Web is to make the current Web smarter. Because of the intricacies involved — and well-known from AI — there are several

complementary strategies for the Semantic Web to achieve its objectives.

### 1.1 Produce Useful Results at Each Step

The Web has been built around some simple principles (cf. (Berners-Lee, 1999)), like

- Start simple.
- Let the underlying techniques be extensible.
- Make it easy to create and use information.
- Connect information.

In order to transfer these principles to the Semantic Web, one should start with the easier tasks first, viz. provide

1. a common syntax for machine understandable statements;
2. common vocabularies;
3. a logic language;
4. for the exchange of proofs; and
5. for the verification of proofs.

For these increasingly difficult tasks, the Semantic Web builds on a layered architecture (enumerated from basic to more sophisticated):

1. URIs and Unicode.
2. XML
3. RDF and RDFS
4. Ontology
5. Logics
6. Proof
7. Trust

The underlying idea of this architecture is that every step may already produce many useful benefits. For instance, when we employ RDF and RDFS, we start from a common data model which has proved useful for data integration tasks (Wiederhold and Genesereth, 1997). When we formalize ontologies, we increase the expressiveness taking advantage of core reasoning capabilities (Hotho et al., 2001b). When we can exchange and verify proofs, we can build a Web of agents that may talk to each other about trust.

## 1.2 Let the Users provide Semantics

One basic strategy in the Semantic Web is that Semantic Web developers should think about applications for which it is worth that people provide machine understandable statements rather than natural language.

Naturally, this will not be worthwhile for all types of applications. If it is in one's interest that his Web page is found, because he offers Web services or because he is a researcher and wants to be famous, it may be worthwhile for him to create the ontologies and the semantic annotation by the user himself — otherwise he will not invest the efforts.

## 1.3 Connect Knowledge

One constant source of failure of semantic descriptions is that many, many people must redo them again and again. Just imagine the many hundreds of thousands of bibtex entries that are typed in and maintained in the machine learning community world wide.

Currently, we have the situation that a lot of such useful information is available in single databases, but it is not possible to exploit it directly because every database has its own syntax and semantics and cannot directly be reused.

The purpose of the Semantic Web is to gather knowledge from individual descriptions like [www.informatik.uni-trier.de/~ley/db/](http://www.informatik.uni-trier.de/~ley/db/), be able to provide mappings *once* and use it ever after. If DBLP would provide RDF in Trier, some nice person would have defined a mapping to bibtex in New York, one could use both in order to produce good citations. The conceptual basis for such integration tasks are common basic languages, such as sketched above.

## 1.4 New Possibilities

One driver for the Semantic Web will be its potential to new, so far often unrecognized, benefits. For instance, domain ontologies on the Web will allow to construct portals more easily (Hotho et al., 2001b) or to better describe Web services (McIlraith et al., 2001).

## 2 Machine Learning

Now, where does machine learning enter the picture? In fact, in all but the first of the strategies just sketched machine learning may eventually play a pivotal role for the success of the Semantic Web. Let us consider the different types of interactions (also cf. Table 1).

### 2.1 Machine Learning for the Semantic Web

Machine learning may contribute to the construction of the Semantic Web. The users of the Semantic Web will have to produce ontologies and eventually populate these ontologies by semantic descriptions.

The first step takes place at the conceptual level, where different types of ontology learning mechanisms may be applied (cf. (Maedche and Staab, 2001; Maedche, 2002)). These mechanisms consider various existing information resources, like text, databases, machine readable dictionaries and integrate the results of several ontology learning steps.

The second step takes place at the instance level, where various means for learning the information extraction system may be applied in order to produce instances from existing concepts (cf., e.g., (Ciravegna, 2001)).

The third step integrates the two aspects and moves toward an integrated system that may bootstrap itself semi-automatically.

Neither of these three tasks will eventually take place in the absence of the human, however they will be very significant steps in order to *cheaply* produce semantic descriptions for ontology definitions (Maedche and Staab, 2001; Maedche, 2002), semantic annotations (Handschuh et al., 2001; Erdmann et al., 2001), and integrated bootstrapping-based systems (Maedche et al., 2002b).

Finally, we want to note that mapping and merging of semantic descriptions may be considered a task strongly related to the three categories just mentioned. First examples for the

Table 1: Machine Learning and the Semantic Web — A preliminary, exemplary classification

	Concept Level	Instance Level
ML for the Semantic Web	<p>Ontology Learning (Maedche and Staab, 2001),(Pekar and Staab, 2001)</p> <p>Ontology Mapping and Alignment (Stumme and Maedche, 2001)</p> <p>Ontology-based Clustering (Hotho et al., 2002; Hotho et al., 2001a) (Hotho et al., 2001b)</p>	<p>ML for Information Extraction (Ciravegna, 2001)</p> <p>(needed by (Handschuh et al., 2001; Erdmann et al., 2001))</p> <p>Re-Classification (Agrawal and Srikanth, 2001; Doan et al., 2001)</p> <p>Input for multi-relational mining (content, structure, usage mining)</p>
Semantic Web for ML		
Bootstrapping	Information Extraction System Bootstrapping (Maedche et al., 2002b)	

latter may e.g. be found in (Stumme and Maedche, 2001; Agrawal and Srikant, 2001; Doan et al., 2001).

## 2.2 The Semantic Web for Machine Learning

The Semantic Web has a lot to offer for machine learning because it provides a rich description of background knowledge that may be exploited for means like Web content, structure and usage mining.

At this point in time, the potential of the Semantic Web may not even be roughly estimated. However, just consider WordNet as an example. In spite of its many benefits, WordNet exhibits many disadvantages for machine learning tasks, too, as it is not domain specific. Nevertheless, it has triggered a lot of fruitful work that may become much better exploited with the arrival of the Semantic Web because the latter makes domain specific semantic descriptions much more widely spread and available.

We have shown that domain specific ontologies may enhance clustering results with regard to mathematical properties as well as with regard to explainability of the results — to mention but one example (cf. (Hotho et al., 2002; Hotho et al., 2001a)).

## 3 Conclusion

We believe that Machine Learning may drive the Semantic Web and the Semantic Web will bring about intriguing Machine Learning problems. Most germane to the Semantic Web is the idea of having semantics intrude from many different types of interactions — ideally in a way such that semantics is a byproduct of normal software use (cf. (Maedche et al., 2002a)).

## References

- R. Agrawal and R. Srikant. 2001. On integrating catalogs. In *Proceedings of WWW 2001*, pages 603–612. ACM Press.
- T. Berners-Lee. 1999. *Weaving the Web*. Harper.
- F. Ciravegna. 2001. Adaptive information extraction from text by rule induction and generalisation. In *Proceedings of 17th International Joint Conference on Artificial Intelligence (IJCAI 2001)*, Seattle, August 2001, San Francisco/CA. Morgan Kaufmann.
- A. Doan, P. Domingos, and A. Halevy. 2001. Reconciling schemas of disparate data sources: A machine-learning approach. In *Proceedings of SIGMOD Conference 2001*. ACM Press.
- M. Erdmann, A. Maedche, H.-P. Schnurr, and S. Staab. 2001. From manual to semi-automatic semantic annotation: About ontology-based text annotation tools. 6.
- S. Handschuh, S. Staab, and A. Maedche. 2001. CREAM — creating relational metadata with a component-based, ontology-driven annotation framework. In *K-CAP 2001 - Proceedings of the First International ACM Conference on Knowledge Capture. October 21-23, 2001, Victoria, B.C., Canada*. ACM Press.
- A. Hotho, A. Maedche, and S. Staab. 2001a. Ontology-based text clustering. In *Proceedings of the IJCAI-2001 Workshop "Text Learning: Beyond Supervision"*, August, Seattle, USA.
- A. Hotho, A. Maedche, S. Staab, and R. Studer. 2001b. SEAL-II — the soft spot between richly structured and unstructured knowledge. *Journal of Universal Computer Science (J.UCS)*, 7(7):566–590.
- A. Hotho, A. Mdche, and S. Staab. 2002. Text clustering based on good aggregations. (3).
- A. Maedche and S. Staab. 2001. Ontology learning for the semantic web. *IEEE Intelligent Systems*, 16(2):72–79.
- A. Maedche, F. Nack, S. Santini, S. Staab, and Luc Steels. 2002a. Emergent semantics. *IEEE Intelligent Systems (Trends and Controversies)*, 17(1).
- A. Maedche, G. Neumann, and S. Staab. 2002b. Bootstrapping an ontology-based information extraction system for the web. In P.S. Szczepaniak, J. Segovia, J. Kacprzyk, and L.A. Zadeh, editors, *Intelligent Exploration of the Web*, Studies in Fuzziness and Soft Computing. Springer/Physica-Verlag, Heidelberg.
- A. Maedche. 2002. *Ontology Learning for the Semantic Web*. Kluwer.
- S. McIlraith, T. Son, and H. Zeng. 2001. Semantic web services. *IEEE Intelligent Systems*, 16(2):46–53.
- V. Pekar and S. Staab. 2001. Learning taxonomies also from rare word occurrences. Technical report, University of Karlsruhe.
- G. Stumme and A. Maedche. 2001. FCA-Merge: A bottom-up approach for merging ontologies. San Francisco/CA. Morgen Kaufmann.
- G. Wiederhold and M. Genesereth. 1997. The conceptual basis for mediation services. *IEEE Expert*, 12(5):38–47.