A Machine Learning Perspective for the Semantic Web

SWWS Position Paper

Alexander Maedche

FZI Research Center on Information Technologies
at the University of Karlsruhe,
Haid-und-Neu-Str. 10-14, 76131 Karlsruhe, Germany
http://www.fzi.de/wim
maedche@fzi.de

According to [1] we consider the Semantic Web, as a **Meta-Web**, that is build on the existing WWW. We adopt the point of view that there are two main core **challenges for putting the vision of the Semantic Web in action**:

- *First*, one has to support the re-engineering task of "semantic enrichment" for building the Meta-Web. The success of the Semantic Web greatly depends on the proliferation of ontologies and relational metadata, which requires quick and easy engineering of them and the avoidance of knowledge acquisition bottlenecks.
  Additionally, we consider the task of merging and aligning ontologies for establishing semantic interoperability as an engineering artifact that may be supported by machine learning techniques (see below).

- *Second*, one has to provide means for maintaining and adopting the machine-processeable data that is the basic for the Semantic Web. Thus, we need mechanisms that support the dynamic nature of the Web.

The research area of Machine Learning has a long history as well on knowledge acquisition or extraction as knowledge revision or maintenance and provides a large number of techniques that may be applied to solve the challenges introduced above. Seminal work on combining machine learning with knowledge acquisition and maintenance has shown its practical usefulness (cf. [9]). The Semantic Web opens a wide range of new research challenges for the machine learning community.

We here give a short overview and research challenges with respect to the combination of machine learning research with Semantic Web research:

Extraction of ontologies from existing data on the Web. The task of extracting ontologies is a typical re-engineering task. In general one may roughly distinguish between existing ontologies (such as thesauri, lexical-semantic nets), schemata (such as relational database, web schemata), instances (in data- and knowledge bases), semi-structured data (e.g. in the form of XML documents), natural language documents. Each of these different kinds of data requires its specific import and processing techniques and learning algorithms. To derive ontologies from existing data on the Web a common picture and framework for re-engineering existing data as given in [8] is required. The integration of multiple resources seems to be a promising approach for the difficult task of extracting ontologies form the existing Web data.

Extraction of relational (meta-)data from existing data on the Web. There exists a number of approaches for (semi-)automatically generating relational (meta-)data from existing data on the Web. The active research field of machine learning for information extraction (cf. http://www.dcs.shef.ac.uk/~fabio/ecai-workshop.html) is concerned of the instantiation of so-called templates from natural language text of a restricted domain and structure. Along the same lines techniques for automatic generation of wrappers are researched.

Merging and mapping ontologies by analyzing extensions of concepts. Several approaches supporting the merging and mapping process simulating the behaviour of the ontology engineer have

been presented. Machine learning techniques may be used by analyzing extensions of concepts to derive overlapping intensional descriptions. In the area of ontology mapping an approach using supervised classification has been described by [7]. A technique taken from conceptual knowledge discovery (based on the theory of formal concept analysis) for deriving a lattice of concepts as structural description of the overall merging process has been introduced by [10]. Recently, techniques well known from the area of data mining (namely association rules) have been used for the task of catalogue integration [1].

Maintaining ontologies by analyzing instance data. The idea of automatically generating a "T-Box" from given relational instance data is not a new one, e.g. an approach for deriving concept descriptions from instances has been presented in [6]. Recently, the work in this area is described under the name of "A-Box-Mining". A first approach on deriving a taxonomy from a given set of RDF statements has been described in [4]. Similar research has been done by the database community, where the critical problem of discovery of the structure implicit contained in semi-structured data and, subsequently, the recasting of the raw data in terms of this structure has been researched.

Improving Semantic Web applications by observing users. Web Mining applies data mining techniques on the existing web (e.g. Web usage mining analyzes the user behavior or web structure mining explores the hyperlink structure). Machine learning in general may be applied on Semantic Web data to improve existing applications by observing users. We refer the interested reader to an ECML/PKDD'2001 workshop on Semantic Web Mining (http://semwebmine2001.aifb.uni-karlsruhe.de/)

In this position paper we have presented a machine learning perspective for the Semantic Web. Based on the old idea of applying machine learning for knowledge acquisition and maintenance, we have presented several core challenges for the Semantic Web that may be approached by machine learning techniques. An important aspect is that there is a wide range of techniques available from the machine learning community that may directly applied for solving existing problems. Another important aspect is that the Semantic Web opens new research questions for the machine learning community (e.g. dealing with multi-relational data) and provides real-world data (e.g. for inductive logic programming algorithms, cf. [5])

References

[1] R. Agrawal and R. Srikant : On Integrating Catalogs. In: Proceedings of WWW-11, Hong-Kong, 2001. (http://www10.org/cdrom/papers/076/index.html)

[2] T. Berners-Lee. J. Hendler, O. Lassila: The Semantic Web. *Scientific American*, 2001.

[3] S. Decker, S. Melnik, F. van Harmelen, D. Fensel, M. Klein, J. Broekstra, M. Erdmann, I. Horrocks, The Semantic Web: The Roles of XML and RDF. *IEEE Internet Computing*, 4 (5): 63-74, September/October 2000.

[4] A. Delteil, C. Faron-Zucker and R. Dieng: Learning Ontologies from RDF annotations. In: Procedings of the IJCAI'01 Workshop on Ontology Learning OL-2001, Seattle, USA, 2001.

[5] P.A. Flach. Inductive Logic Databases: from extensional to intensional knowledge. In F. Bry, R. Ramakrishnan, and K. Ramamohanarao, editors, 5th International Conference on Deductive and Object-Oriented Databases (DOOD'97), Springer-Verlag, December 1997.

[6] J.U. Kietz and K. Morik: A polynomial approach to the constructive induction of structural knowledge. In: Machine Learning, 1994.

[7] M.S. Lacher and G. Groh: Facilitating the exchange of explicit knowledge through ontology mappings. In: Proceedings of FLAIRS'2001, AAAI Press, 2001.

[8] A. Maedche and S. Staab: Ontology Learning for the Semantic Web. In: *IEEE Intelligent Systems, Special Issue on the Semantic Web,* 16(2), 2001.

[9] G. Webb, J. Wells and Z. Zheng: An Experimental Evaluation of Integrating Machine Learning with Knowledge Acquisition. In: *Machine Learning*, 35(1), 1999

[10] G. Stumme and A. Maedche: FCA-Merge: Bottom-Up Merging of Ontologies. In: Proceedings of the International Joint Conference on Artificual Intelligence IJCAI'01, Seattle, USA, August, 2001

[11] Visser, P.R.S & Tamma, V.A.M., 1999: An Experience with Ontology-based Agent Clustering. In: Proceedings of the IJCAI'99 Workshop on Ontologies and Problem-Solving Methods: Lessons Learned and Future Trends, Stockholm, Sweden, August 2, 1999