What is the Role of the Semantic Layer Cake for Guiding the Use of Knowledge Representation and Machine Learning in the Development of the Semantic Web?

Randy Goebel[1], Sandra Zilles[1], Christoph Ringlstetter[1], Andreas Dengel[2], Gunnar Grimnes[2]

[1]Alberta Ingenuity Center for Machine Learning
University of Alberta, Edmonton, Alberta T6G 2E8, Canada
{goebel, zilles,kristof}@ualberta.ca
[2] Knowledge Management Research
German Research Centre for Artificial Intelligence
D-67663 Kaiserslautern, Germany
{Andreas.Dengel, Gunnar.Grimnes}@dfki.de

Abstract

The World Wide Web Consortium (W3C) has been the consolidator for many ideas regarding the evolution of the World Wide Web (WWW), including the promotion of the so-called "Semantic Layer Cake" model for the development of the semantic web. The semantic layer cake provides a framework to discuss a variety of approaches to an integrated view of the meta-data that will support a broad range of machine and human manipulation of digital information.

Our goal is to develop a deeper understanding of the potential role of the semantic layer cake, by investigating some of the detailed relationships between the components. We also attempt to articulate some of the issues regarding the development of the semantic layer cake for real application. The path to this goal covers many fundamental issues underlying all of knowledge representation research. Paramount in this discussion is the goal of identifying the potential role of machine learning within the semantic layer cake, and finding the tradeoffs between the extremes of a completely automated construction of the semantic web via machine learning, versus one that is completely hand-engineered using the tools emerging from each layer of the semantic layer cake.

Introduction

The World Wide Web Consortium (W3C) has been the clearing house for many ideas regarding the evolution of the World Wide Web (WWW), including the promotion of the so-called Semantic Web, often simplistically illustrated by the "Semantic Layer Cake" model (e.g., see [2]). The semantic layer cake diagram of Figure 1 derives from several articles from the W3C consortium, and in this instance, from a summary paper of Dengel and Wahlster [1] on the evolution of the WWW.

While the idea of the semantic layer cake has captured a lot of attention as a convenient intellectual abstraction to frame discussions on various aspects of the semantic web (e.g., [3, 7]), there is considerable uncertainty about its exploitation. For example, the semantic layer cake level labeled "trust" is easily interpreted as a component that somehow manages the vital issue of information trust with respect to WWW digital information. Similarly, the level labeled "logic & rules" suggests a meta component that combines some kind of knowledge representation and reasoning capability, without any constraints on the scope or role.

Our goal, however, is not a detailed analysis of trust or logic (or any other level), but rather to develop at least an outline of the potential role of knowledge representation and machine learning ideas within the general semantic web framework, using the semantic layer cake as a guide.

Perhaps the most important point about the semantic web is the underlying principle that *both* computers and people should be able to reason about web content, which provides the motivation for the articulation and use of meta-data. We provide a hint at connections to knowledge representation and machine learning by considering their role in answer the following questions:

1 Who builds the meta-data?

2 How is the meta-data used?

3 Is the meta-data static, or is it modified by use?

4 How can the value of meta-data be assessed?

Question 1 begs the issue of how meta-data is constructed, within the spectrum of doing it by hand with some special authoring tools (e.g., for RDF), or by using some collection of machine learning methods (e.g., as in

[4, 10]). In addition, where the expected application is in information retrieval, one can consider Question 3 as asking how user input is considered in improving the quality of meta-data with respect to some evaluation measure. Evaluation is critical, as it determines the answers to Questions 2 and 4, which cannot be addressed without knowing how to assess the value of that meta-data.

One way to establish a stronger role for the semantic layer cake is by investigating some of the relationships between individual semantic web technologies and their role for the semantic interpretation of legacy WWW content. In this regard, we identify some of the issues regarding the development of the semantic layer cake for *real use*. We hope a deeper understanding of the potential guiding role of the semantic layer cake for the treatment of initially non-structured or imperfect annotated data will reveal the necessary preliminaries for efficient input-output relations between its components beyond a closed world of semantic web technologies.

Note that whatever role the semantic layer cake has in the development of the semantic web, there will always be the need to understand *both* the use *and* construction of semantic web content within this framework. This means that the use of meta-data will be vital in the creation of new content, as well as the improved organization and use of existing content.

The magnitude and breadth of the issues is enormous: some research has concentrated on the construction of meta-data within the semantic layer cake to interpret content (e.g., [3]), while others (e.g., [4]) have focused on inducing meta-data. Still others have noted the use of meta-data "standards" for guiding the creation of content (e.g., [1, 2]). Within these perspectives remains the common goal of using the abstraction of the semantic layer cake to guide the creation and interpretation of WWW content by *both* machines and humans.

Paramount in our discussion is the goal of identifying the potential role of machine learning within the semantic layer cake, and exploring the tradeoffs between the completely automated construction of the semantic web via machine learning, versus one that is completely hand-engineered using the tools emerging from each layer of the semantic layer cake.

The rest of this brief document is organized as follows: first provide a summary of semantic layer cake components, and give a simple description of their connection. Then follows some speculation on their potential role to their targeted users. We then present our high level speculation on the respective roles of knowledge representation and machine learning. In the case of the latter, we focus on ontology construction and use, and consider the challenges of that segment of the semantic layer cake. We conclude with a an appeal for research that improves the precision with which semantic layer cake components and ideas are developed, especially in not ignoring the strong inevitable connection to the more mature fields of knowledge representation and machine learning.

Parts Analysis: The Semantic Layer Cake

One of the difficulties in analyzing the semantic layer cake is that neither the individual components nor their relationships are very clearly described. So a first analysis here tries to isolate each component, and consider its role in the WWW, especially the relationship between a WWW user (e.g., via a web browser and search engine) and that component. A significant motivation for each semantic layer cake level is the formal articulation of meta-data for use by *both* human and machine. But for our immediate purposes, we begin with a more intuitive analysis based on human user expectations, and hope that in moving that agenda forward, the requirements for machine exploitation of meta-data will be clearer.

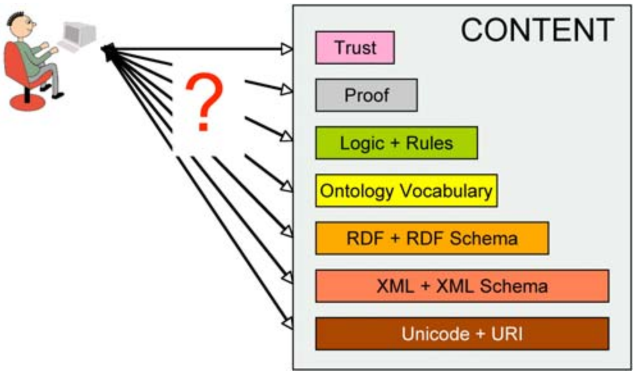

Figure 1. The Semantic Layer Cake

To initiate the component analysis, consider an elaboration of the role of the top of the layer cake: he "trust" component. Setting aside assumptions on the lower layer levels, one can presume that the trust component has responsibility to establish a trust relationship between the user and the information content delivered from the lower levels. But the concept of trust, especially within the evolution of community-based WWW (e.g., [1]) is complex, and practical comprehensive models of trust remain elusive. Existing trust instruments like authentication certificates, secure information exchange, and the heuristic combination of trust relationships represent some broad range of trusted information exchange (e.g., [11, 12]), but there are more innovative ideas for trust (e.g., [13]), and suggestions that connect trust to the notion of proof and explanation (e.g., [14]). This idea is related to that of providing "explanation," as in the 1980's research on expert systems. But note that, despite the suggested decomposition of information structure within the semantic layer cake, the issue of trust (and spam) exists at *every* level. The point is that the interpretation of the semantic layer cake labels can be arbitrarily broad, and that both knowledge representation and machine learning will have fundamental roles in *any* foundation of a trusted semantic web.

It is clear that trust is a complicated idea, and that there is a lot to do before a generally acceptable framework for semantic layer cake trust is achieved. Similarly, all of the semantic layer cake components require further articulation to be practically viable. This elaboration will fill volumes, so we set aside further elaboration here, but provide a speculative summary the broad responsibilities for each semantic layer cake level in Tables 1 and 2.

Overall, there are at least three different roles for each component, which can be summarized as

1. Scenarios of use between a WWW user and semantic layer cake component,
2. Relationship between the semantic layer cake component and the base content of the WWW, and,
3. The description of any possible machine learning opportunities, which could be deployed on WWW content in support of each semantic layer cake component.

Both 1 and 2 rely on the support of a variety of knowledge representation methods, which are required to provide support for both the capture and use of meta-data that enables the semantic web. A summary is provided in Table 1 below.

Any real instance of the basic semantic layer cake architecture will provide the scaffolding within which machine learning methods can be deployed to augment meta-data in some way, to improve the value of the WWW content. Therefore it is also clear that somehow measuring value improvements is important, as will be noted below.

Layer Roles and Knowledge Representation

Table 1 provides a sketch of the relevant knowledge representation and reasoning ideas related to each level of the semantic layer cake. Overall, given the anticipated role of the relevant layer, the role of knowledge representation ideas is naturally focused on modeling the appropriate knowledge believed to be required to support that role, as well as a repertoire of reasoning methods.

For example, both the notions of trust and proof, given the assumption that the role is to provide evidence for the quality of related WWW content, require the application of knowledge representation systems that can describe concepts of trust, and how to reason to confirm those concepts, e.g., in terms of abductive reasoning.

In general, the role of knowledge representation at each level is noe of modeling and reasoning about the meta-data that captures the concepts at that level. In general, as one moves further down the layer cake, there is increased emphasis on more specific information modeling, and perhaps even domain specific reasoning methods. A frequently-used example is meta-data and companion use about travel scenarios, as typified by the organization of WWW content within applications like Expedia or Travelocity. Knowledge about typical travel scenarios are like the scripts and frames of knowledge representation research (e.g., [??]), with attendant reasoning issues that suggest a combination of top down (hypothesis-driven) and bottom up (data-driven) reasoning is required to support effective use of relevant domain content.

Component	**Anticipated Role & Responsibility**	**Potential Knowledge Representation Concepts**
Trust	• create user perception of quality and credibility of content • measure value only by contrast within different content creation communities (e.g., Wikipedia, Citizendium) • not independent of proof, logic, and rules components (cf. “explanation)	• models of trust and trust relations • abductive reasoning
Proof	• confirming connection between queries and content, using inference • providing structure of explanations • partial dynamic integration of disparate data'/information (e.g., consistency approximation)	* representation of common sense concepts • models of similarity and case-based reasoning • high level presentation of explanations
Logic + Rules	• reasoning-based interaction about data/information merging	• models for domain-specific knowledge * efficient domain-specific reasoning
Ontology Vocabulary	• query refinement • content capture and maintenance • user guided ontology merging	• models for ontology authoring • reasoning support for ontology maintenance
RDF + RDF Schema	• user/ community tagging • support for query refinement and incremental classification	• database tools
XML + XML Schema	• creation of shareable ASCII-based content	• content authoring tools
Unicode + URI	• uniform language representation	• content authoring tools
CONTENT	• basic information repository	• all forms of manual and automatic content generation

Table 1. Component Roles & Knowledge Representation Concepts

The development of traditional knowledge representation tools for semantic web development has a strong thread within applied knowledge representation communities from which descriptions logics and even ontology languages like OWL have been developed. In this regard, there is already a relatively good connection between the development of tools for modeling at least the lower levels of the layer cake. As evidence, consider the use of RDF is as a first level of meta-data. There is a clear and obvious connection between RDF data, the semantic dictionaries of the computational linguistics communities, and even the object schemas and relational schemas of the object and relational database communities.

But the danger is that further elaboration of knowledge representation methods above the RDF layer will be driven only in an ad hoc manner, without principled consideration for the role of a variety of more and more sophisticated methods, including probabilistic methods (e.g., Bayes nets, hidden Markov models and Markov decision processes), and discrete logical methods, including constraint reasoning, incremental optimization, and non-monotonic reasoning and belief revision). In fact, as one ascends the layer cake, it seems that the knowledge representation requirements become more demanding, leading implicitly to considerations on inductive modeling and machine learning.

In this regard, an important distinction going forward will be that of using sophisticated reasoning in the application of WWW content (e.g., learning user retrieval patterns to improve targeted search), versus learning explicit meta-data that improves the quality of the semantic descriptions of WWW content. The latter leads naturally to consider how machine learning can help construct meta-data.

Layer Roles and Machine Learning Opportunities

There has been considerable research on learning in the semantic web context (e.g., [4, 7, 8, 10]), but so far no systematic/holistic approach. While it is unlikely that progress in this regard can be made without more detailed development of semantic web evaluation methods, it is possible to at least consider the kinds of meta-data that might be learned. Table 2 summarizes some potential opportunities for the application of machine learning to the automated construction of meta-data for each level, but it is clear that, as one ascends the layer cake, the potential learning outcomes are increasingly vague. It is the case that, at each level, the representation and reasoning support to exploit any learning outcome is also necessary (cf. Table 1).

So far, machine learning techniques related to the semantic web have been developed mainly for the purpose of

a) ontology learning (e.g., [4, 7]), and

b) ontology alignment, particularly entity resolution which can be seen as instance-level ontology alignment (e.g., [15]).

Component	**Anticipated Role & Responsibility**	**Potential Learning Outcomes**
Trust	• create user perception of quality and credibility of content • measure value only by contrast within different content creation communities (e.g., Wikipedia, Citizendium) • not independent of proof, logic, and rules components (cf. “explanation)	• tracking evolution of content creation communities • individual and community models of trust relations
Proof	• confirming connection between queries and content, using inference • providing structure of explanations • partial dynamic integration of disparate data’/information (e.g., consistency approximation)	• abstracting explanations by analogy/case-based reasoning
Logic + Rules	• reasoning-based interaction about data/information merging	• adapting to user/community inference rule practices • maintaining status of conflicting information
Ontology Vocabulary	• query refinement • content capture and maintenance • user guided ontology merging	• adaptation of ontologies adjustment, merging, etc.) by user/ community practices
RDF + RDF Schema	• user/ community tagging • support for query refinement and incremental classification	• classification of vertical domains • semi-automatic classification with community-based editorial oversight
XML + XML Schema	• creation of shareable ASCII-based content	• template learning and deployment for constraint-guided content creation
Unicode + URI	• uniform language representation	
CONTENT	• basic information repository	• all forms of manual and automatic content generation

Table 2. Component Roles & Machine Learning Opportunities

The most obvious place for the deployment of learning is at the RDF level, where one can imagine that even simple classification in large RDF repositories would provide structure that can aid in improved semantic clarity of WWW user access. It is less clear about the potential value of machine learning as one ascends the layer cake, partly because what can be induced is more abstract, and the description of any component above the RDF level remains vague without more experimental development.

We can illustrate some of the lower level opportunities for machine learning with a few examples. Consider earning OWL-Class descriptions from existing RDF data. Legacy data extracted from an RDBMS into RDF will result in shallow ontologies based on the database schema, and methods like inductive logic programming can be used to learn more expressive OWL class expressions for such data. See here for instance [8]. However, it remains unclear how such approaches scale in terms of runtime performance when dealing with a huge amount of data. Quality and runtime complexity of learning algorithms will have to be improved — where at this point it is still very unclear in which way quality of a learning algorithm or its output should be defined in the context of learning ontologies.

Further examples arise in within the context of entity resolution and ontology alignment. A simple technique used for entity resolution in the semantic web is “smushing,” which refers to replacing all references to resources which share the same object in an inverse functional relation by a single reference, since they must all refer to the same resource, see for instance [15].

This is not the only way entity resolution is dealt with in the semantic web research community. For example, in systems that use so-called PIMO (personal information model) and other semantic desktop systems, a third entity is created: the “pimo thing” and multiple entities are linked from that “thing” and considered as “occurrences” of the “thing.” More details can be found in [8]. However, we think that more complex machine learning techniques, like for instance relational clustering, are required for enhancing entity resolution and ontology alignment for the semantic web, see, e.g., [4] and the references therein. Envisioning large-scale applications, different techniques from on-line learning and incremental learning have to be given consideration and that also affects the methods to be used for entity resolution and ontology alignment. Current work focuses on the trade-off between the accuracy and the run-time efficiency of entity resolution using query-time

Evaluation: putting it altogether

A crucial aspect in the deployment of machine learning techniques for the semantic web will be the evaluation of their performance. For instance, considering learning ontologies and ontology alignment, there is no straightforward quality measure. If there are test data for which target ontologies are known, it still remains to compare ontologies learned to target ontologies. This involves not only the design of similarity measures for ontologies and the definition of quality criteria for learning algorithms, but essentially also the use of visualization techniques. We are aware of the need for good evaluation methods but do not see it as a challenge rather than an obstacle to deploying machine learning for the construction of a semantic web.

In the broader picture, the evaluation of any knowledge representation or machine learning technology will be determined by how effective meta-data resources are in efficiently supporting a user’s (human or machine) intentions when using WWW content. Naturally, much of the literature on evaluation begins with traditional information retrieval measures, where precision and recall form the basis for differentiating various semantic web methods. This includes both manually coded meta-data (e.g., [4]), as well as that augmented by machine learning (e.g., [5]).

The bottom line is that it sesms that design and deployment of machine learning techniques for the semantic web has to be accompanied with the design and deployment of special evaluation techniques.

That there is a tradeoff between a hand-engineered semantic web and an automatically constructed (e.g., by continuous online learning) semantic web is taken as an assumption. This is partly because there is already too much WWW content to be hand-engineered to an acceptable level of semantic richness, and that the development of tools (e.g., ontology management, proof and explanation construction, etc.) will not ensure the veracity of hand-engineered metadata.

The second assumption regarding the framework of the semantic layer cake has to do with existing thrusts in the development of domain specific semantically rich web content, largely vetted by small trusted communities of users (e.g., Wikipedia). The assumption is that the semantic web may well find that evolving tools based on semantic layer cake ideas will help contribute to the development of semantic content, but that will be done by individual communities of self-interest, rather than some uniform application of W3C (or any other) standards, including the semantic layer cake.

Development-Driven Layer Cake Role Evolution

It is apparent that the Semantic Layer Cake and its individual components address the structure and content of the emerging semantic web, and not necessarily external or ancilliary components (e.g., semantic web services, and external agents or brokers) that provide support for interpreting both the base and meta-content of the WWW. In this regard, for example, the interaction with a user and the trust component might be based on some external trust broker that is not a part of any of the semantic layer cake. Similarly, all the interaction and user support deriving from any of the components might be replaced or at least augmented by external agents (e.g., proof support, logic and reasoning support, ontology management, etc.)

Conclusions

Our collaboration consists of two research centres focusing on the role of fundamental work on knowledge representation and machine learning within the context of the semantic web. Our activities are directed at articulating the tradeoffs in the deployment of existing models and theories, and of the practical challenges in making semantic web development decisions, within the framework of the semantic layer cake.

We have only scratched the surface of the possible dissection of the semantic layer cake idea, and its role for elucidating research strategies for orchestrating the development of the semantic web. It is quite clear that more intensive and principled research on the combined role of knowledge representation and machine learning is necessary, and that development-driven evolution is only a part of what is required to integrate sound principles into tools that both help interpret current web content, and help guide future content development.

Acknowledgements

Discussions leading to this summary have been supported by the Alberta Ingenuity Centre for Machine Learning (AICML), the Alberta ICT Circle of Research Excellence (iCORE), and the German Research Centre for Artificial Intelligence (DFKI).

References

[1] W. Wahlster and A. Dengel and W. Wahlster (2006), Web 3.0: Convergence of Web 2.0 and the Semantic Web, Technology Radar Feature Paper, Edition 11/2006, June 2006, German Research Centre for Artificial Intelligence Research/Deutsche Telekom Laboratories

[2] Tim Berners-Lee, James Hendler, and Ora Lassila (2001) The Semantic Web, Scientific American, 279.

[3] V. Geroimenko and C. Chen (eds, 2006), Visualizing the Semantic Web, Springer

[4] P. Cimiano (2006), Ontology Learning and Population from Text: Algorithms, Evaluation and Applications, Springer.

[5] Y. Sure, P. Hitzler, A. Eberhart, R. Studer (2005), The Semantic Web in One Day, IEEE Computer, May/June 2005, 85-87.

[6] Z. Huang, F. van Harmelen, A. ten Teije, P. Groot, and C. Visser (2004) Reasoning with inconsistent ontologies: a general framework, Semantically Enabled Knowledge Technologies, Vrije Universiteit Amsterdam, June 30, 2004.

[7] H. Hu and D-Y Liu (2004), Learning OWL ontologies from free texts, Proceedings of the Third International Conference on Machine Learning and Cybernetics, Shanghai, PRC, August 2004, 1233-1237.

[8] G. Grimnes, P. Edwards, and A. Preece (2004), Learning Meta-Descriptions of the FOAF Network, Proceedings of the 3rd International Semantic Web Conference, Hiroshima, Japan, November 7-11, 2004, Lecture Notes in Computer Science 3298, 152-165, Springer.

[9] I. Bhattacharya and L. Getoor (2007), Query-time Entity Resolution, Journal of Artificial Intelligence Research 30:621-657.

[10] A. Maedche (2002), Ontology Learning for the semantic web, Kluwer Academic, Norwell, USA.

[11] H. Waguih (2006), A proposed trust model for the semantic web, Proceedings World Academy of Science, Engineering, and Technology, Volume 11, ISSN 1307-6884.

[12] Y. Zhang, H. Chen, Z. Wu, X. Zheng (2006) A reputation-chain trust model for the semantic web, 20th International Conference on Advanced Information Networking and Applications, 20th International Conference, ISSN: 1550-445X.

[13] J. Reagle Jr. (2002), Finding Bacon's key: does Google show how the semantic web could replace public key infrastructure, http://www.w3.org/2002/03/key-free-trust.html

[14] T. Roth-Berghofer, S Schulz, D. Bahls, D. Leake (2007) Explanation-Aware Computing, Papers from the 2007 AAAI Workshop, Vancouver, British Columbia, Canada, July 22-23, 2007 AAAI Press 2007.

[15] J. Euzenat and P. Shvaiko (2007), Ontology Matching, Springer 2007.