Describing Rich Content: Future Directions for the Semantic Web

Timo Honkela and Matti Pöllä
Helsinki University of Technology
Laboratory of Computer and Information Science
P.O. Box 5400 FI-02015 TKK, Finland
{timo.honkela, matti.polla}@tkk.fi

Abstract

In this position paper, we discuss some problems related to those semantic web methodologies that are straightforwardly based on predicate logic and related formalisms. We also discuss complementary and alternative approaches and provide some examples of such.

1 Introduction

It is clear that the use of standardized formats within computer science is beneficial. For instance, the widespread use of the World Wide Web would not have been possible without the adoption of HTML. Similarly, there are serious attempts to create standards for metadata, data about data, so that a piece of art stored in electronic form would include information about, e.g., its creator, source, identification and possible access restrictions. Moreover, metadata usually includes also a textual summary of the contents, a content description that provides information for the organization and search of the data.

1.1 Subjective and complex paths from data to metadata

Especially if pictorial or sound data is considered, an associated description, metadata, is useful. The description may be based on a pre-defined classification or framework, for instance, like an ontology within the current semantic web technologies. However, even if something like the identity of the author or the place of publishing can rather easily be determined unambiguously, the same is not true for the description of the contents. In the domain of information retrieval and databases of text documents, Furnas et al. (1987) already found that in spontaneous word choice for objects in five domains, two people favored the same term with less than 20% probability. Bates (1986) has shown that different indexers, well trained in an indexing scheme, might assign index terms for a given document differently. It has also been observed that an indexer might use different terms for the same document at different times. The meaning of an expression (queries, descriptions) in any domain is graded and changing, biased by the particular context. Fig. 1 aims to illustrate the challenging aspects of the overall situation. Human beings perceive, act and interact within complex environments. It has occasionally been assumed that much of the underlying conceptual structure within human mind is readily given by some way, even already when we are born. There is a vast body of literature related to this question which we do not touch upon here. It suffices to note that the discussion in this paper is based on the assumption that the conceptual systems are mainly emergent: they are created, molded and shared by individuals in interaction with each other and the rest of the accessible part of the world.

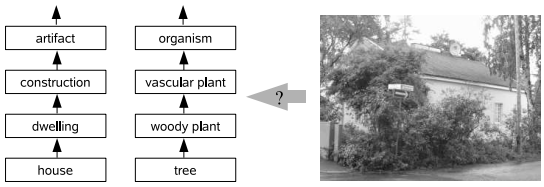

Figure 1: An illustration of the basic problem of symbol grounding: what is the process that is capable of generating ontological structures based on complex raw perceptions and activities.

Vygotsky (1986, originally published in 1934) has stated that "... the world of experience must be greatly simplified and generalized before it can be translated into symbols. Only in this way does communication become possible, for the individual's experience resides only in his own consciousness and is, strictly

speaking, not communicable." Later, he continues: "The relation of thought to word is not a thing but a process, a continual movement back and forth from thought to word and from word to thought. In that process the relation of thought to word undergoes changes which themselves may be regarded as development in the functional sense." This means in practice that conceptualization is a complex process that takes place in a socio-cultural context, i.e., within a community of interacting individuals whose activities result into various kinds of cultural artifacts such as written texts.

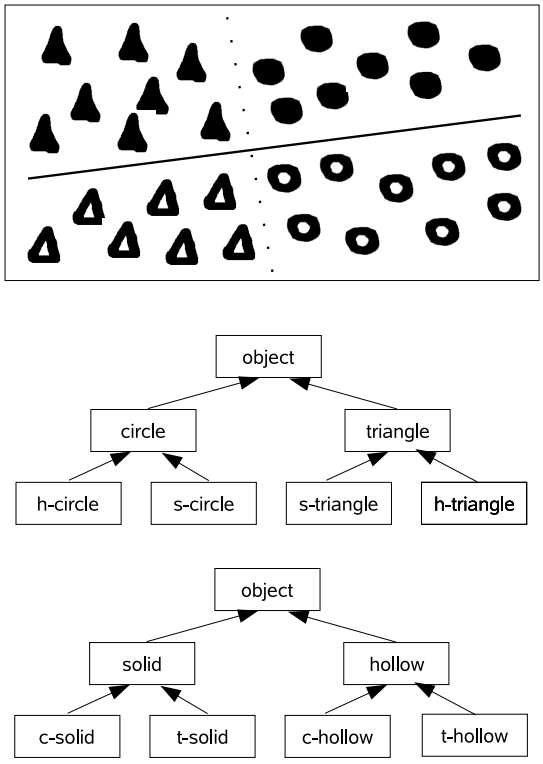

Figure 2: An illustration of two conflicting conceptual systems on the surface level. These systems prioritize the distinctive features differently.

It is a very basic problem in knowledge management that different words and phrases are used for expressing similar objects of interest. Natural languages are used for the communication between human beings, i.e., individuals with varying background, knowledge, and ways to express themselves. When rich contents are considered this phenomenon should be more than evident. Therefore, if the content description is based on a formalized and rigid framework of a classification system, problems are likely to arise. Fig. 2 shows a simple example of two conflicting formalizations.

Natural languages have evolved to have a certain degree of compositionality to deal with such situations. Vogt (2006) has developed a simulation model that considers the transition of holistic languages versus compositional languages. Holistic languages are languages in which parts of expressions have no functional relation to any parts of their meanings. It is to be noted, though, that the compositionality is a matter of degree in natural languages. One cannot assume that, for instance, each noun would refer to one concept and the conceptual structures would follow isomorphically the composition of the syntactical structures. For instance, the existence of collocations complicates the situation.

1.2 From two-valued logic to adaptive continuous-valued models

In addition to the different ways of creating conceptual hierarchies discussed above, the inherent continuous nature of many phenomena makes it impossible to determine exactly, in a shared manner the borderlines between some concepts or how some words are used. Actually, we prefer to consider concepts as areas in high-dimensional continuous spaces as suggested by Gärdenfors (2000). Fig. 3 illustrated the basic problem of dividing continuous spaces into discrete representations. Various approaches have been developed to deal with this phenomenon, fuzzy set theory as a primary example Zadeh (1965).

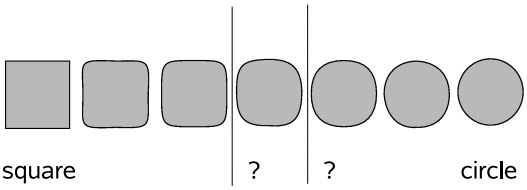

Figure 3: The obvious difference between continuous and discrete. The continuous can be discretized but there is a certain cost associated. The cost relates to the need to learn and deal with an increasing number of explicit symbols.

The basic semantic web formalisms are based on predicate logic and other symbolic representations and are subject to most of those problems that earlier AI formalisms have. In order to alleviate the problems related to the traditional approach, there are already examples of research projects in which some soft computing approaches, including fuzzy logic, probabilistic modeling and statistical machine learn-

ing, are applied. Even a collection named "Soft Computing in Ontologies and Semantic Web" has recently been published Ma (2006). In the collection, a related topic to the discussion above is presented by Holi and Hyvönen (2006). They consider modeling uncertainty in semantic web taxonomies in particular in the domain of geographical information. Nikravesh (2006) presents an approach which is based on the use of, e.g., fuzzy logic, evolutionary computation and the self-organizing map.

In this position paper, we outline some complementary and sometimes even alternative approaches to the core methodologies currently applied within the semantic web research and development.

2 Alternatives to highly formalized metadata

It may be useful not to define any artificial limitations for the descriptions. For instance, when the domain develops into directions which did not exist when the classification system was developed, problems arise.

Moreover, if the content it described using large enough body of text the for better recall, i.e., higher likelihood for finding the information is greater. However, tools for ensuring precision are needed. Precision refers to the number of relevant retrieved documents over the total number of retrieved documents.

If a word or an expression is seen without the context there are more possibilities for misunderstanding. Thus, for human reader the contextual information is often very beneficial. The same need for disambiguation can also be relevant within information systems. As the development of ontologies and other similar formalizations are, in practice, grounded in the individual understanding and experience of the developers and their socio-cultural context, the status of individual items in a symbolic description may be unclear. Fig. 4 illustrates the influence of varying sociocultural contexts. Namely, if one considers the pragmatic meaning (if not semantic meaning) of "summer's day" in a Shakespeare sonnet, it is obvious that there is a great difference between the contexts that are being referred to, for example, in Scandinavia and in Sahara.

Similarly, the methods that are used to manage data should be able to deal with contextual information, or even in some cases provide context. The Self-Organizing Map by Kohonen (2001) can be considered as an example of such a method. Often it is even possible to find relevant features from the data

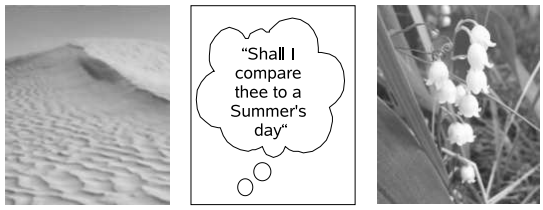

Figure 4: Different contexts potentially related to the expression given.

itself Kohonen et al. (1997). However, a computerized method – using a some kind of autonomous agent – does not provide an "objective" classification of the data while any process of feature extraction, human or artificial, is based on some selections for which there most often are some well-grounded alternatives.

Below, we describe some examples in which the ideas and principles outlined above have been applied.

3 Case studies

We consider three cases which exemplify complementary and alternative approaches to the use of (first order) logic-based formalisms in knowledge representation.

3.1 Color labeling

The concept of color cannot be adequately studied only by considering the logico-semantic structure of color words. One has to take into account the color as a physical phenomenon. Color naming also requires consideration of the qualities of the human color perception system. A thorough study of color naming, thus, would require consideration of at least linguistic, cognitive, biological, physical and philosophical aspects Hardin (1988).

Subjective naming of color shades has been studied in a demonstrative web site where users can pick color shades and give name labels (or 'tags') to the colors. In this case color shades are represented as a RGB tuple indicating the intensities of red, green and blue color.

As the database of tagged color shades grows, interesting properties can be found by looking at the distributions of individual tags. For example, the tag 'red' would result in a reasonably narrow distribution centered around the point (1.0; 0.0; 0.0) in the RGB space. Other tags, however, can have much more variation in the way people place them in the color space.

For example, the tag 'skin' or 'hair' is an example of a highly subjective name for a color shade due to various skin and hair color conceptions. A related theme is the fact that the domain for which a color name is an attribute has a clear influence on which part of the RGB space the color name refers to. Gärdenfors (2000) lists as an example the differences between the quality of redness of skin, book, hair, wine or soil. Fig. 5 below gives an example how the color space can be divided in a substantially different manner in two languages.

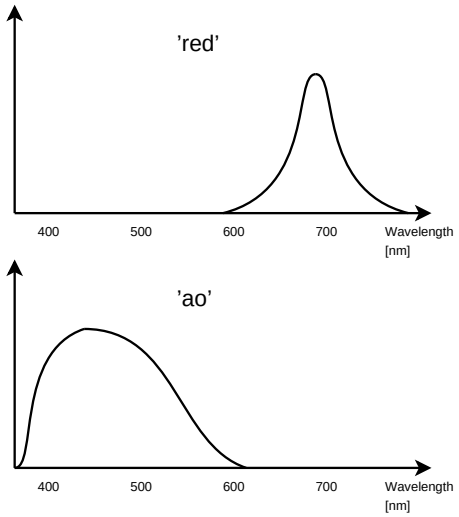

Figure 5: Illustration of fuzzy color definitions. The word 'red' is usually referred to light wavelengths close to 700 nm while the Japanese term 'ao' is associated with a wider scale of wavelengths. In English the term 'ao' would cover both 'blue' and 'green'.

3.2 WEBSOM and PicSOM

The WEBSOM method was developed to facilitate an automatic organization of text collections into visual and browsable document maps (Honkela et al., 1997; Lagus et al., 2004). Based on the Self-Organizing Map (SOM) algorithm (Kohonen, 2001), the system organizes documents into a two-dimensional plane in which two documents tend to be close to each other if their contents are similar. The similarity assessment is based on the full-text contents of the documents. In the original WEBSOM method (Honkela et al., 1996) the similarity assessment consisted of two phases. In the first phase, a word-category map (Ritter and Kohonen, 1989; Honkela et al., 1995) was formed to detect similarities of words based on the contexts in which they are used. The Latent Semantic Indexing (LSI) method (Deerwester et al., 1990) is nowadays often used for similar purpose. In the second phase, the document contents were mapped on the word-category map (WCM). The distribution of the words in a document over the WCM was used as the feature vector used as an input for the document SOM. Later, the WEBSOM method was streamlined to facilitate processing of very large document collections (Kohonen et al., 1999) and the use of the WCM as a preprocessing step was abandoned.

The PicSOM method (Laaksonen et al., 1999, 2002; Koskela et al., 2005) was developed for similar purposes than the WEBSOM method for content-based image retrieval, rather than for text retrieval. Also the PicSOM method is based on the Self-Organizing Map (SOM) algorithm (Kohonen, 2001). The SOM is used to organize images into map units in a two-dimensional grid so that similar images are located near each other. The PicSOM method brings three advanced features in comparison with the WEBSOM method. First, the PicSOM uses a tree-structured version of the SOM algorithm (Tree Structured Self-Organizing Map, TS-SOM) (Koikkalainen and Oja, 1990) to create a hierarchical representation of the image database. Second, the PicSOM system uses a combination of several types of statistical features. For the image contents, separate feature vectors have been formed for describing colors, textures, and shapes found in the images. A distinct TS-SOM is constructed for each feature vector set and these maps are used in parallel to select the returned images. Third, the retrieval process with the PicSOM system is an iterative process utilizing relevance feedback from the user. A retrieval session begins with an initial set of different images uniformly selected from the database. On subsequent rounds, the query focuses more accurately on the user's needs based on their selections. This is achieved as the system learns the user's preferences from the selections made on previous rounds.

WEBSOM and PicSOM methods are a means for content-driven emergence of conceptual structures. Fig. 6 illustrates how the clustering structure discernable on a self-organizing map can correspond to a hierarchical structure. This basic principle can be applied in multiple ways to provide a bridge between the raw data directly linked with some phenomenon and the linguistic and symbolic description of its conceptual structure. Fig. 6 also shows why the SOM is gaining popularity as a user interface element replacing, e.g., traditional menus. With suitable labeling of the map, the user can easily browse the map zooming into details when necessary. The structure of the map

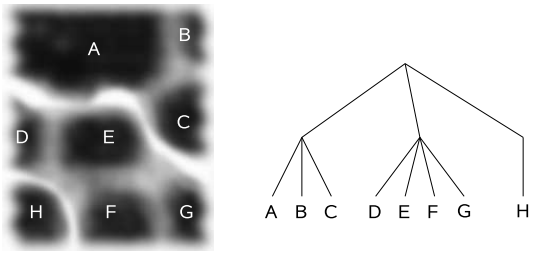

Figure 6: The emergent structure of an organized self-organizing map reflects the structure of the underlying data. The clustering that is visible on the left can be interpreted as the tree diagram on the right.

depends on the data and its preprocessing. The preprocessing can be used to bring up relevant structures. For instance, it may be useful to increase or decrease the weight of some variables. If we consider a map of an e-mail collection, one can decrease the weight of those variables that are commonly related to unsolicited e-mail messages. This leads into the situation in which the SOM can allocate more resources on modeling useful e-mail. The same idea applies to many application areas in which a particular point of view is more relevant than some other.

3.3 Quality assessment of medical web sites

The web has become an increasingly important source of medical information replacing much of the area of medical self-help literature. The consequences of relying on medical information found on web sites can be crucial in terms of making decisions about treatment. Hence there is need for assessing the quality of these web sites to help the layman decide which information he/she should rely on.

Currently, quality labeling of medical web sites is done by various organizations of medical professionals. The assessment process is usually done completely by hand requiring a large amount of manual work by trained physicians in searching web sites for the required elements (including, for example, proper contact information). The EU funded project MedIEQ[1] aims to develop tools to facilitate the process of web site quality labeling.

The MedIEQ project applies mostly the current semantic web technologies to describe the web site contents. However, already in the early stages of the project, it has become apparent that some kinds of contents are difficult to analyze using the structured approach. For instance, the target audience of a medical web site is a property by which the site is being labeled in the assessment process. As it turns out, it is less than trivial to automatically detect whether a site is intended to be read by laymen or medical professionals. Further, the division of the target audience types into crisp categories is often subjective by nature.

[1] http://www.medieq.org/

4 Conclusions

We have presented some motivation why the core technologies in Semantic Web should not solely rely on predicate logic and related formalisms. We have argumented for a certain data-driven approach in which the original data is analyzed automatically rather than relying on hand-crafted ontologies and their use as a basis for choosing descriptors in the metadata. We have given examples of such an approach mainly using the Self-Organizing Map as the core method.

References

M. J. Bates. Subject access in online catalog: a design model. *Journal of the American Society of Information Science*, 37(6):357–376, 1986.

S.C. Deerwester, S.T. Dumais, T.K. Landauer, G.W. Furnas, and R.A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41:391–407, 1990.

G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11):964–971, 1987.

P. Gärdenfors. *Conceptual spaces: The Geometry of Thought*. MIT Press, 2000.

C.L. Hardin. *Color for Philosophers - Unweaving the Rainbow*. Hackett Publishing Company, 1988.

M. Holi and E. Hyvönen. *Soft Computing in Ontologies and Semantic Web*, chapter Modeling Uncertainty in Semantic Web Taxonomies, pages 31–46. Springer, 2006.

T. Honkela, V. Pulkki, and T. Kohonen. Contextual relations of words in grimm tales analyzed by self-organizing map. In F. Fogelman-Soulié and P. Gallinari, editors, *Proceedings of ICANN'95, International Conference on Artificial Neural Networks*,

pages 3–7, Paris, France, October 1995. EC2 et Cie, Paris.

T. Honkela, S. Kaski, K. Lagus, and T. Kohonen. Newsgroup exploration with WEBSOM method and browsing interface. Technical Report A32, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland, 1996.

T. Honkela, S. Kaski, K. Lagus, and T. Kohonen. WEBSOM—self-organizing maps of document collections. In *Proceedings of WSOM'97, Workshop on Self-Organizing Maps*, pages 310–315. Espoo, Finland, 1997.

T. Kohonen. *Self-Organizing Maps*. Springer, 2001.

T. Kohonen, S. Kaski, and H. Lappalainen. Self-organized formation of various invariant-feature filters in the adaptive-subspace SOM. *Neural Computation*, 9:1321–1344, 1997.

T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, J. Honkela, V. Paatero, and A. Saarela. Self organization of a massive text document collection. In *Kohonen Maps*, pages 171–182. Elsevier, Amsterdam, 1999.

P. Koikkalainen and E. Oja. Self-organizing hierarchical feature maps. In *Proc. IJCNN-90, International Joint Conference on Neural Networks, Washington, DC*, volume II, pages 279–285, Piscataway, NJ, 1990. IEEE Service Center.

M. Koskela, J. Laaksonen, M. Sjöberg, and H. Muurinen. PicSOM experiments in TRECVID 2005. In *Proceedings of the TRECVID 2005 Workshop*, pages 267–270, 2005.

J. Laaksonen, M. Koskela, and E. Oja. Picsom: Self-organizing maps for content-based image retrieval. In *Proceedings of IEEE International Joint Conference on Neural Networks (IJCNN'99)*, pages 2470–2473, 1999.

J. Laaksonen, M. Koskela, and E. Oja. Picsom - self-organizing image retrieval with mpeg-7 content descriptions. *IEEE Transactions on Neural Networks, Special Issue on Intelligent Multimedia Processing*, 13(4):841–853, 2002.

K. Lagus, S. Kaski, and T. Kohonen. Mining massive document collections by the WEBSOM method. *Information Sciences*, 163:135–156, 2004.

Z. Ma. *Soft Computing in Ontologies and Semantic Web*. Springer, 2006.

M. Nikravesh. *Soft Computing in Ontologies and Semantic Web*, chapter Beyond the Semantic Web: Fuzzy Logic-Based Web Intelligence, pages 149–209. Springer, 2006.

H. Ritter and T. Kohonen. Self-organizing semantic maps. *Biological Cybernetics*, 61(4):241–254, 1989.

P. Vogt. Cumulative cultural evolution: Can we ever learn more? In S. Nolfi et al., editor, *Proceedings of SAB 2006, From Animals to Animats 9*, pages 738–749, Berlin, Heidelberg, 2006. Springer.

L. Vygotsky. *Thought and language*. MIT Press, 1986, originally published in 1934.

L. A. Zadeh. Fuzzy sets. *Information and Control*, 8: 338–353, 1965.