HELSINKI UNIVERSITY OF TECHNOLOGY
Laboratory of Computer and Information Science

Describing Rich Content: Future Directions for the Semantic Web

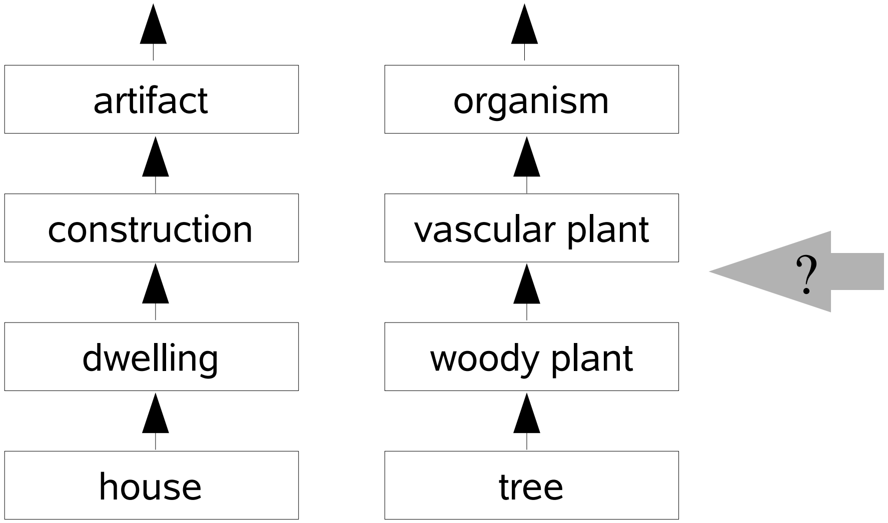
Timo Honkela and Matti Pöllä

Helsinki University of Technology

Laboratory of Computer and Information Science

Subjective and complex paths from data to metadata

- 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, often assign index terms for a given document differently.
- Human beings perceive, act and interact within complex environments.
- Basic assumption: 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
- Each conceptual system is subjective to some degree.



artifact

construction

dwelling

house

organism

vascular plant

woody plant

tree

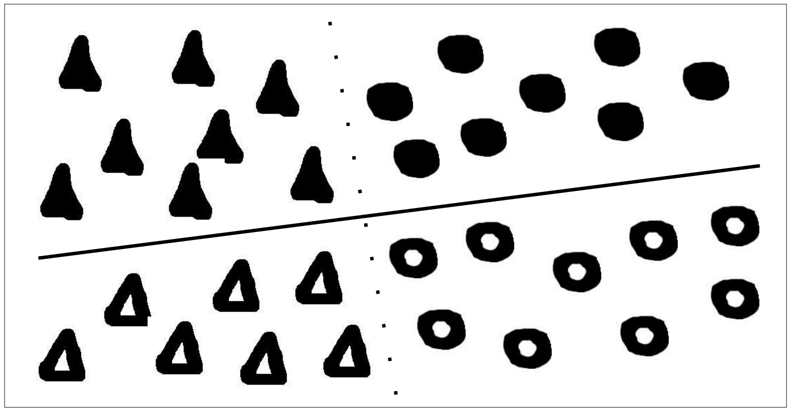
?

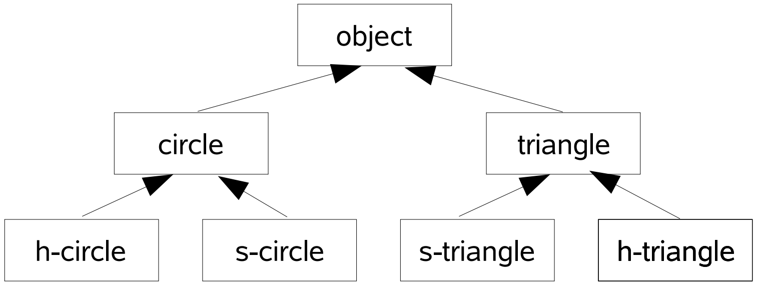
Insight by Vygotsky (1934)

- “... 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.”
- “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.

Formalization within rich contexts

- 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.





object

circle

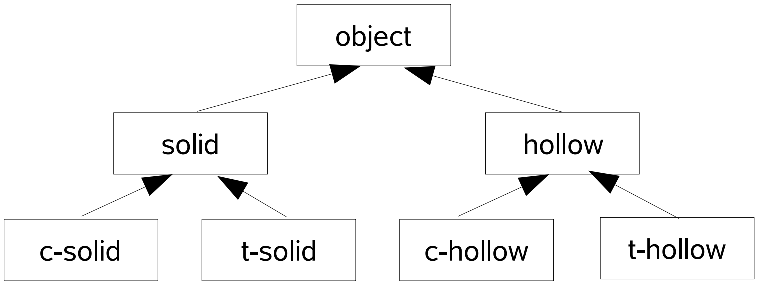
triangle

h-circle

s-circle

s-triangle

h-triangle



object

solid

hollow

c-solid

t-solid

c-hollow

t-hollow

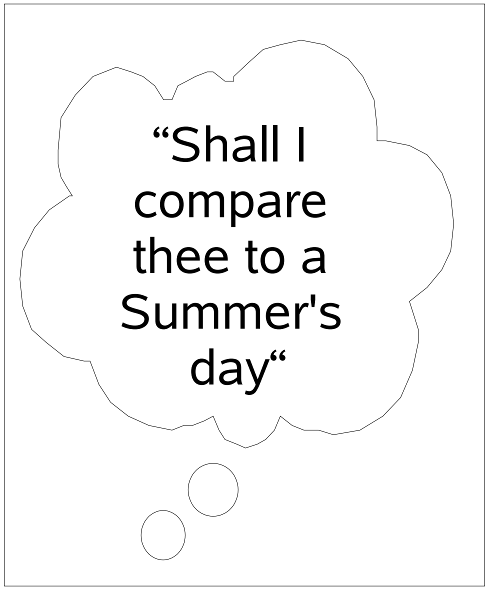
From two-valued logic to adaptive continuous-valued models

- 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).
- 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.

- There are already examples in which some soft computing approaches, including fuzzy logic, probabilistic modeling and statistical machine learning, are applied.
- Even a collection named “Soft Computing in Ontologies and Semantic Web” has recently been published.
- In the collection, a related topic is presented by Holi and Hyvönen who consider modeling uncertainty in semantic web taxonomies in particular in the domain of geographical information.
- Nikravesh presents an approach which is based on the use of, e.g., fuzzy logic, evolutionary computation and the self-organizing map.

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.
- 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.
- 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.
- 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.



“Shall I compare thee to a Summer's day“

Case studies

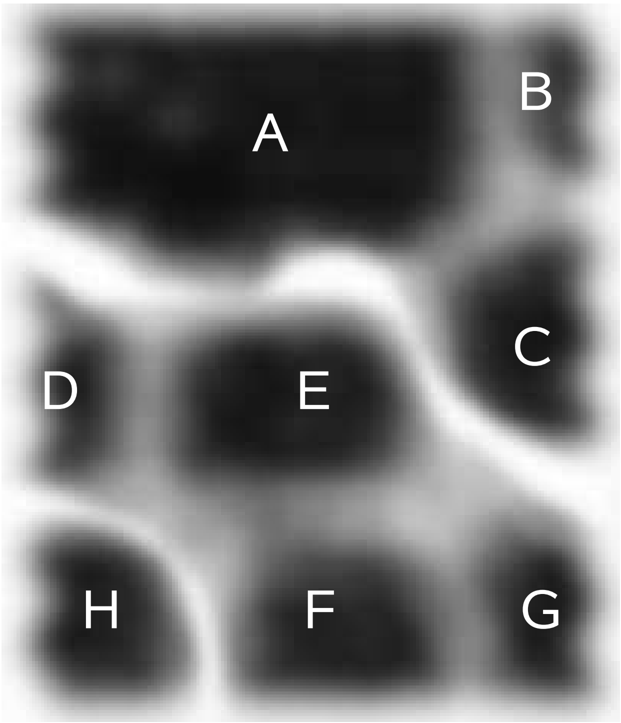
- Color labeling
- WEBSOM and PicSOM
- Quality assessment of medical web sites

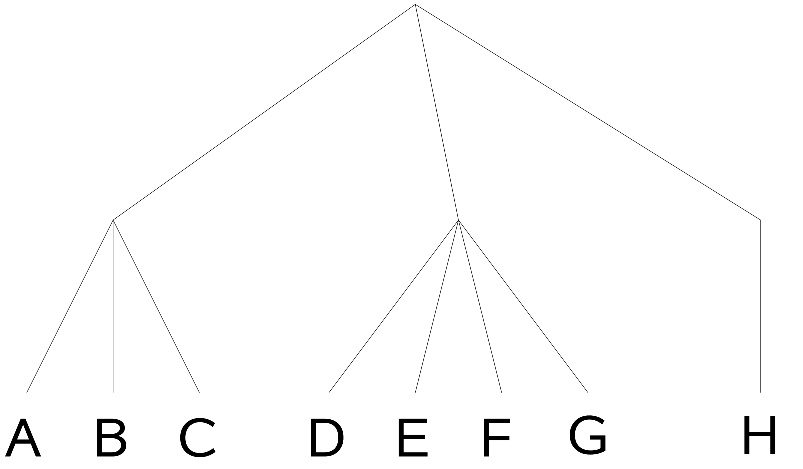
Color labeling

- Surprisingly complex phenomenon involving linguistic, cognitive, biological, physical and philosophical aspects
- 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.
- The tag ’red’ results in a reasonably narrow distribution centered around the point (1; 0; 0) in the RGB space.
- Other tags, however, can have much more variation in the way people place them in the color space (e.g. ’skin’ or ’hair’)
- 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 (redness of skin, book, hair, wine or soil).

WEBSOM and PicSOM

- WEBSOM: based on the Self-Organizing Map 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 (Honkela et al. 1996, etc.).
- The PicSOM method (Laaksonen et al. 1999, etc.) was developed for for content-based image retrieval.
- The PicSOM system uses a combination of several types of statistical features.
- The retrieval process with the PicSOM system is an iterative process utilizing relevance feedback from the user.
- The WEBSOM and PicSOM methods are a means for content-driven emergence of conceptual structures.
- Consider also Laaksonen and Viitaniemi (SCAI 2006).


A
B
C
D
E
H
F
G


A
B
C
D
E
F
G
H

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.
- Quality labeling of medical web sites is done by various organizations of medical professionals.
- The EU funded project MedIEQ (http://www.medieq.org/) 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.
- Some kinds of contents are difficult to analyze using the structured approach (e.g. target audience: laypersons or medical professionals, adults or children).

Conclusions

- 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.