Ontology Engineering: Co-evolution of Complex Networks with Ontologies

Francesca Arcelli Fontana[1], Ferrante Raffaele Formato[2] and Remo Pareschi[3]

[1]Università degli Studi di Milano Bicocca, viale Sarca 336, 20126 Milano, Italy

[2]Università del Sannio, Benevento, Italy

[3]Università del Molise, Italy

Abstract. Our assumption here is that the relationships between networks of concepts (ontologies) and people networks (web communities) is reciprocal and dynamic. ontologies identify communities and communities through practice define ontologies. Ontologies describe complex domains and therefore are difficult to create manually. Our investigation aims at building tools and methodologies to drive the process of ontology building. In particular we define a model by which ontologies evolve through Web community extraction. In this paper, we observe that *tags* in Web 2.0 are mathematical objects called *clouds* and studied in [11]. And we introduce *NetMerge,* an algorithm for transforming an ontology into a complex network.

1 Introduction

Ontology design is actually performed by a panel of experts. The organization of design follows the structure of the ontology. Ontology engineering is defined as the set of methods used for building from the scratch, enriching or adapting an existing ontology in a semi-automatic fashion, using heterogeneous information sources ([15]). This data-driven procedure uses text, electronic dictionaries, linguistic ontologies and structured information to acquire knowledge.

Recently, with the enormous growth of the Information Society, the Web has become a valuable source of information for almost every possible domain of knowledge. This has motivated researchers to start considering the Web as a valid repository for Information Retrieval and Knowledge Acquisition. However, the Web suffers from problems that are not typically observed in classical information repositories: human oriented presentation, noise, untrusted sources, high dinamicity and over-whelming size. Even though, it also presents characteristics that can be interesting for knowledge acquisition: due to its huge size and eterogeneicity it has been assumed that the Web approximates the real distribution of the information in humankind. The present research aims to introduce a novel approach for ontology design and learning, presenting new methods for knowledge acquisition from the Web. The adaptation of several well known learning techniques to the web corpus, the exploitation of particular characteristics of the Web environment and the search for

Arcelli Fontana F., Raffaele Formato F. and Pareschi R. (2009).
Ontology Engineering: Co-evolution of Complex Networks with Ontologies.
In *Proceedings of the 1st International Workshop on Ontology for e-Technologies OET 2009*, pages 33-40
DOI: 10.5220/0002222300330040

Web community unsupervised by a semi-automatic and domain independent approach distinguishes the present proposal from previous works ([18], [.22]).

In [1],[2] we sketched a model in which ontologies are generated by human expertize and then they evolve according to laws of complex networks in Web 2.0. We built also an algorithm that extracts some Web communities using a combination of HITS ([17]) and a visual Web Crawler called TouchGraph. In that paper we started by giving some definitions and propositions in order to show how to merge semantic web into complex networks and then we showed how to interpret ontologies with a network of Web communities. Then we described the dynamic process of the interaction between an ontology and Web communities through a max-flow based approach to Web communities discovery.

In this paper, we refine our model by which ontologies evolve through Web community extraction and we observe that *tags* in Web 2.0 are mathematical objects called *clouds* (studied in [11]). We introduce *NetMerge,* an algorithm for transforming an ontology into a complex network, by associating a Web community to some nodes of a network. We have analysed several other tools and applications that assist experts in the generation of ontologies, that help the knowledge engineer to build a taxonomy or enrich an existing one, among them we analysed OntoGen [14], KAON and Text2Onto [7], OntoLearn, [8] Jatke and Ontobuilder [21].

The paper is organized through the following sessions: in Section 1 we introduce some aspects and concepts related to ontologies engineering and learning; in Section 3 we discuss some problems on ontologies learning through the web; in Section 4 we formulate our hypothesis that an ontology can be interpreted by a complex and dynamic network and we introduce the prototype we have developed, called NetMerger, based on focus crawling and able to merge an ontology and a significant portion of the web. Finally in Section 5, we conclude and outline some future directions of this research, in particular relating to an an algorithm for cloud extraction.

2 Ontology Engineering and Ontology Learning

The set of activities that concern the ontology development process, the ontology life cycle, the principles, methods and methodologies for building ontologies, and the tool suites and languages that support them, is called *Ontology Engineering* ([8,15,18]). With regard to methodologies, several proposals have been reported for developing ontologies manually.

Considering Guarino's classification ([16), philophical ontologists and artificial intelligence logicians are usually involved in the task of defining the inalterable basic kinds and structures of concepts (objects, properties, relations and axioms) that are applicable in every possible domain. Those basic principles are contained in the mentioned *Top-level ontologies*, also called *Fondational or Upper ontologies.*

On the contrary, *Application ontologies* have a very narrow context and limited reusability as they depend on the particular scope and requirements of a specific application. These ontologies are typically developed ad hoc by the application designers.

At an intermediate point, *Task* and *Domain ontologies* are the most complex to develop: on one hand, they are general enough to be required for achieving consensus between a wide community of users and, on the other hand, they are concrete enough to present an enormous diversity with many different and dynamic domains of knowledge and millions of possible concepts to model.

A global initiative such as the Semantic Web ([5, 6]) relies heavily on domain ontology. The Semantic Web tries to achieve a semantically annotated Web in which search engines could process the information contained on web resources from a semantic point of view, increasing drastically the quality of the information presented to the user. This approach requires a global consensus in defining the appropriate semantic structures, -i.e. the domain ontologies- for representing any possible domain of knowledge. As a consequence, there is a wide agreement that a critical mass of ontologies is needed for representing semantics on the Semantic Web.

The construction of domain ontologies relies on domain modellers and knowledge engineers that are typically overwhelmed by the potential size, complexity and dynamicity of a specific domain. As a consequence, the construction of an exhaustive domain ontology is a barrier that very few projects can overcome.

It turns out that, although domain ontologies are recognized as crucial resources for the Semantic Web, in practice they are not available, and when available they are used outside specific research environments.

Due to all this reasons, nowadays there is a need of methods that can perform, or at least ease, the construction of domain ontologies. In this sense, Ontology Learning([18.22]) is defined as the set of methods and techniques used for building from scratch, enriching, or adapting an existing ontology in a semi-automatic fashion using distributed and heterogeneous knowledge and information sources. This allows a reduction in the time and effort needed in the ontology development process.

3 Ontology Learning through the Web

In the last years, as well known, the Web has become a valuable source of information for almost every possible domain of knowledge. However, the Web suffers from many problems that are not typically observed in the classical information repositories. Those sources even written in natural language, are often quite structured in a meaningful organisation or carefully selected by information engineers and as a consequence, one can assume that the trustiness and validity of the information contained in them is reliable and valid. In contrast, the Web raises a series of new problems that need to be tackled:

- Web resources are presented in human oriented semantics –natural language- and mixed with a huge amount of information about visual representations. This adds a lot of noise over the really valuable information and makes difficult a machine-based processing approach. There have been several attempts to improve the machine interpretability of the Web content like using XML notation to represent concepts and hierarchies, or the definition

of some HTML extensions –like SHOE- to include tags with semantics information, but none of them has been widely accepted.

- All kinds of documents for almost every possible domain coexist. Some of them offer valuable –up-to-date information from reliable sources; others are simply spam that even tries to confuse the user. Everyone can post any kind of information –fake or real- without any control and, in consequence, the Web becomes a completely untrustable environment.
- It presents a highly dynamic and uncontrolled changing nature. Web sites are rapidly modified, updated or deleted, making difficult and outdating any attempt of structuring the information.
- The amount of available resources on one hand, can overwhelm the final user or information engineer that tries to search specific data; on the other hand, it makes nonviable a complex machine-biased processing for extracting data in an automated way.

Ontology learning is performed by defining and maintaining two levels in the ontology: a lower one that is scale-free and connects the ontology to the web sites, a top one with a random distribution that is the proper ontology.

Usually, classical ontologies are designed by a panel of experts that gathers to negotiate classes and relations. But this is not the way in which knowledge is acquired. By so doing, Semantic Web is static and is being phagocyted by the complexity and homeostatic response of Web 2.0.

4 Using Complexity to Beat Complexity

The basic research hypothesis that we formulate is that an ontology can be interpreted by a complex and dynamic network and –at the same time- maintain the granularity level that make ontologies abstract enough to beat complexity ([19]). By so doing, ontologies are transformed into "meta-ontologies", an object that grows by the law of preferential attachment and at the same time maintains a "democratic" random distribution of concept sufficient to abstract complex knowledge.

To investigate our hypothesis, we formulate a computable model in which ontology classes are linked to the Web via a supervised classifier. We try to show BIB that –in our model- the distribution of arches is exponential, according to the asymptotic law formulated by [3,4].

In our model, ontology classes are linked to a scale-free dynamic network via focused crawling. Each concept-or class-generates a set of URLs that are the positive outcomes of a supervised/unsupervised classification process. Successively, these sites are used as *seeds* to discover a web community focused upon the concept of the class. Finally, the community is attached to the corresponding concept in a random way. By so doing, the uniform distribution of the ontology is preserved.

Finally, new concepts are added to the ontology by knowledge synthesis as follows: after a given time, the community is checked for connection and –if splitted- two new seeds are extracted and the corresponding concepts are added to the

ontology and the process continues cyclically. This approach grants the merging of Semantic Web into Web 2.0.

4.1 Preliminary Results

We used Gelsomino [13] a focused crawler developed according to Chakrabarty's architecture ([12]), Gelsomino has been used to find several communities through the Web. Gelsomino has four modules: the first one is a classic web crawler, the second is a Bayesian classifier, the third is the HIT algorithm and the fourth is a module for the extraction of Web communities. Web communities are strategic because a web site inside a web community is a highly connected site with high business potential. On the contrary, isolated sites are mostly ignored. Therefore, tools for extracting Web communities are the favourite candidates to replace search engines in Web 2.0.

A Web community is a subgraph of the Web such that for any node, the number of inner edges is greater than the number of outer edges. For example, Figure 1 illustrate a typical example of Web community Extracted from the Web through Gelsomino and TouchGraph.

4.2 Tokenizing Concepts through Web Community Trawling

Tags have been diffused in Web 2.0 after folksnomies [23]. In [11] a mathematical object was introduced that corresponds to tags:, called clouds.

Given a set S, a cloud is a finite subset of S. Clouds are given a geometric structure through a similarity. A similarity is an application R: $S \times S \rightarrow [0,1]$ such that R(x,x)=1, R(x,y)=R(y,x) and $R(y,z) \geq R(x,y)*R(y,z)$. Then we can associate to every subset X of S a number μ(X) expressing the "worst" degree of similarity between pairs of elements in X, i.e.

$$\mu(X) = \underset{x,x' \in X}{\wedge} R(x,x').$$

For instance, if S = {square, polygon, rectangle} And R(square,polygon) = 0.5, R(polygon,rectangle)=0.4 and R(polygon,square)=0.3, then μ({square, polygon, rectangle}) =0.3. We can regard a non-empty cloud X as a sparse point and the number μ(X) as a many valued evaluation of the claim that X is a point. Clouds have the following interesting geometrical properties:

$$\mu(Y) \geq \mu(X) \wedge Inc(X,Y).$$

$$\mu(X \cup Y) = \mu(X) \wedge \mu(Y) \wedge Over(X,Y).$$

Where $Over(X,Y) = \sup_{x \in X, y \in Y} R(x,y).$

$$Inc(X,Y) = \underset{x \in X}{\wedge} \underset{y \in Y}{\vee} R(x,y).$$

Once we have extracted a cloud of concepts through ontology learning, we must choose the concepts that are reflected into Web 2.0. To such an extent, we perform

Web community trawling upon each concept, and we discard the concepts that do not refer to any Web community. By so doing we obtain a scale-free meta-ontology.

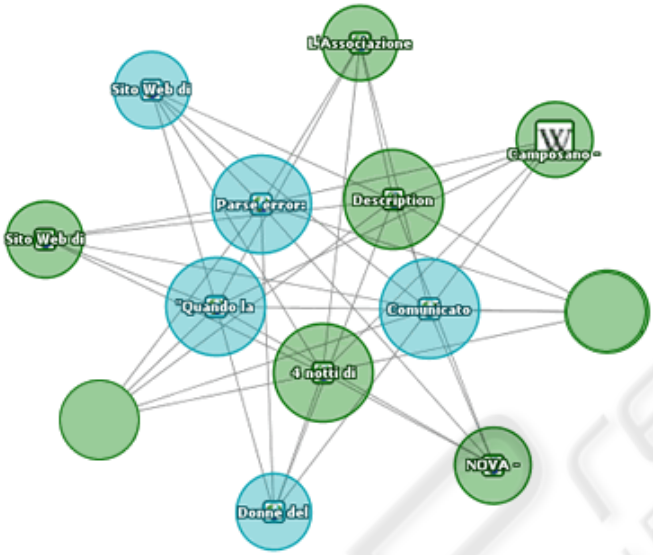

Fig. 1. The Web community around the event agency IsideNova discovered with Gelsomino and TouchGraph.

4.3 The Prototype

We are working on *NetMerger*: a software that is based on proven classification techniques (Bayesian inference, SVM) with the aim to merge two networks, given by, respectively, a domain ontology (which corresponds to a lattice or a random graph) and a significant portion of the Web, such as the portion of the Web induced by a corresponding domain directory of Web sites (which corresponds to a scale free network with hubs and preferential attachments).
NetMerger is a software based on the focused crawler Gelsomino and it is applied to an ontology. The architecture of NetMerger is described in Figure 2.

Here is a brief description of the architecture's modules:

Ontology Builder. A tool for building ontology –like Ontobuilder- with facilities for concept generation and relation estabilishment.

Focused Crawler. A module for extracting communities and associating them to concepts. It is based on focused crawling

Network Merger. A module for merging several Web communities focused on the same concept. In our implementation, a network merger is just a link-based graph merger.

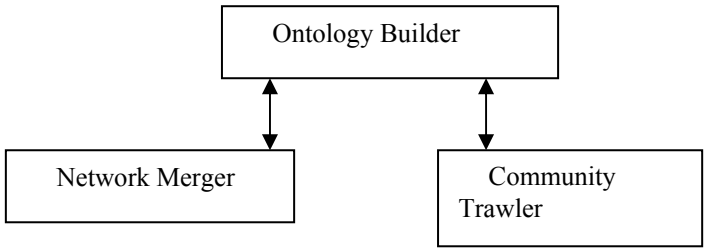

Fig. 2. NetMerger architecture.

The result is a scale-free meta-network (or a meta-Web), where the ontology adds to the sub-Web, a set of conceptual nodes, which provide an interpretation for the nodes in the scale free network, clustering them into "interest" groups, which can be used as channels for e-commerce and e-business agents.

4 Conclusions and Future Developments

This research aims at the integration of Semantic Web technologies into the theoretical framework of complex dynamic network. If successful, Web 2.0 will be open to semantic hulls that will make the Web 2.0 not only socially inhabitable, but also machine understandable. As a consequence, knowledge will be continuously negotiated between corporate standards of panels and Web community. Software agents will not be clutched to static knowledge, but can enjoy the flexibility of dynamic networks.
In the future research we aim to develop a tool, called *Concept-Seeder*: able to extracts new concepts from the Web by identifying "communities" (namely highly linked regions of the Web) and creates concepts identifiable with the content of the sites belonging to such communities. Thus these concepts come as already "trained" with the content they are identified with, and provide a way of evolving domain ontologies from the "bottom-up", by observing how the world effectively goes, as opposed to the traditional way of evolving them "top-down" through the decision of a committee of experts. On the other hand they provide in any case input to the experts for revising the general architecture of the ontology.

References

1. F.Arcelli, F.Formato, R.Pareschi,: Networks as interpretation of Networks, Tech.Report University of Milano Bicocca –TD 23/08, May 2008.
2. F.Arcelli, F.Formato, R.Pareschi,: Reflecting Ontologies into web Communi*ties, IEEE Proceedings of IAWTIC 2008*, Vienna, December 2008
3. A.L. Barabasi, A. Réka and Hawoong J.,: The diameter of the World Wide Web, *Nature*, Vol 401, September 1999, pp130-131.
4. A.L.Barabasi and A. Réka, Emergence of Scaling in Random Networks, *Science,n.* 286, October 1986, pp. 509-512.

5. T. Berners Lee, N. Shadboldt and W. Hall, The Semantic Web Revisited, *IEEE Intelligent Systems*, 21 (3), 2006, pp. 96-101.
6. A. Broder, R Kumar, F.Maghoul, R. Prabhakar, Rajagopalan, R. Stata, A. Tomkins and J. Wiener, The Web as a Graph, *Proc. of the 9th International Web Conference,* Amsterdam May 5h 1999. http://www9.org/w9cdrom/160/160.html.
7. P. Cimiano and J. V¨olker. Text2onto - a framework for ontology learning and data-driven change discovery. In *Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems* (NLDB'2005), 2005.
8. A. De Nicola, M. Missikoff, R. Navigli (2009). "A Software Engineering Approach to Ontology Building". Information Systems, 34(2), Elsevier, 2009, pp. 258-275.
9. G.W. Flake S. Lawrence and C.L. Giles, Efficient Identification of Web Communities, *Proc. Sixth ACM SIGKDD international conference on knowledge discovery and data mining*, Boston, MA, 2000.
10. L. R. Ford; D. R. Fulkerson, Maximal flow through a network. *Canadian Journal of Mathematics* 8, 1956. pp. 399-404.
11. Ferrante Fomato,: On similarity and extensions of logic Programming, PhD Thesis, University of Salerno, 1999.
12. Chakrabarty S. M. van den Berg and B. Dom Focused Crawling, A new approach to topic-specific Web resource discovery. *Proceeings of www8*, Toronto, May, 2008
13. Gelsomino Focused Crawler Diogene Project, Moma & DIIMA University of Salerno Tech. Report 2008.
14. B. Fortuna, M. Grobelnik, D. Mladenic. OntoGen: Semi-automatic Ontology Editor. *HCI International 2007*, July 2007, Beijing
15. Gruber (2008). "Ontology". To appear in the *Encyclopedia of Database Systems*, Ling Liu and M. Tamer Özsu (Eds.), Springer-Verlag, 2008.
16. Guarino, N. Formal Ontology, Conceptual Analysis and Knowledge Representation, *International Journal of Human-Computer Studies*, 43(5-6):625–640, 1995.
17. N. Imafuji,. M. Kitsuregawa Finding a web community by maximum flow algorithm with HITSscore based capacity, *Proceedings of the Eighth International Conference on Database Systems for Advanced Applications*, Kyoto, 2003.
18. Maedche, A. & Staab, S. (2001). "Ontology learning for the Semantic Web". In: *Intelligent Systems*. IEEE, 16(2): 72-79.
19. H.Morin, Restricted Complexity vs. General Complexity" *Philosophy and Complexity* Carlos Gershenson, Diederik Aerts (editors), World Scientific, 2005.
20. R. Navigli and P. Velardi. Learning domain ontologies from document warehouses and dedicated web sites. *Comput. Linguist.*, 30(2):151–179, 2004.
21. OntoBuilder: Fully Automatic Extraction and Consolidation of Ontologies fromWeb Sources using Sequence Semantics, EDBT 2006 Workshops
22. Gómez-Pérez, A., Manzano-Macho, D.: An overview of methods and tools for ontology learning from texts. *Knowledge Engineering Review* 19 (2005) 187–212
23. VanderWal T. Folsonomy at www.vanderwal.net/folksonomy.html