Introduction to the Special Issue of the Journal of Web Semantics: *Bridging the Gap – Data Mining and Social Network Analysis for Integrating Semantic Web and Web 2.0*

Bettina Berendt, Andreas Hotho, Gerd Stumme

The last years have seen increasing collaboration of researchers from the Semantic Web, Web 2.0, social network analysis and machine learning communities. Applications that use these research results are achieving economic success. Data now become available that allow researchers to analyze the use, acceptance and evolution of their ideas.

Highly popular user-centered applications such as Blogs, social tagging systems, and Wikis have come to be known as “Web 2.0” or the “Social Web”. A major reason for their immediate success is the high ease of use of new Web 2.0 services. These sites do not only provide data, but also generate an abundance of weakly structured metadata. A good example is tagging. Here, users add keywords from an uncontrolled vocabulary, called tags, to a resource. Such metadata are easy to produce, but lack any kind of formal grounding, as used in the Semantic Web.

The Semantic Web complements the bottom-up effort of the Web 2.0 community in a top-down manner. Its central point is a stronger knowledge representation, based on some kind of ontology with a fixed vocabulary and typed relations. Such a structure is typically something users implicitly have in mind when they provide their information in Web 2.0 systems. However, for further use, this structure is hidden in the data and needs to be extracted. Techniques to analyze network structures or weak knowledge representations, such as those found in the Web 2.0, have a long tradition in different other disciplines, like social network analysis, machine learning and data mining. These kinds of automatic mechanisms are necessary to extract the hidden information and to reveal the structure in a way that the end user can benefit from it. Using established methods to represent knowledge gained from unstructured data will also be beneficial for the Web 2.0 in that it provides Web 2.0 users with enhanced Semantic Web features to structure their data.

For this special issue, we invited submissions that show how synergies between Semantic Web and Web 2.0 techniques can be successfully used. Since both communities work on network-like data structures, analysis methods from different fields of research could form a link between those communities, and we encouraged submissions that show potential links. We encouraged authors to illustrate how techniques from social network analysis, graph analysis, machine learning and data mining methods, as well as other fields, can provide productive synergies between Web 2.0 and Semantic Web. From a total of 41 submissions that underwent a rigorous three-step reviewing process, three contributions were chosen that exemplify the above-mentioned connections between Social and Semantic Web in excellent ways.

All three papers describe how data from the Semantic Web can profit from a semantic approach by structuring these data, and how the produced semantic representation of the actors and contents of Social-Web platforms can be fed back to users to enable them to better utilize the Web. The example utilization/application areas studied in detail comprise the whole gamut of today's Web: the Social Web, the Semantic Web, and even applications that transcend the traditional Web and reach out into mobile and ubiquitous devices. The three papers also investigate different types of starting data – texts relating to activities, tags relating to multimedia resources, and (social-)media information of all kinds relating to a person. The techniques they use are likewise complementary, originating from machine learning, data/text mining and natural-language analysis, and social network analysis. Finally, ontologies for knowledge representation show a range of knowledge to be represented: from properties of an entity (person) via properties of situations and actions to subjective facets of entities.

The first paper, *Bridging the Gap Between Tagging and Querying Vocabularies: Analyses and Applications for Enhancing Multimedia IR* by Kerstin Bischoff, Claudiu Firan, Wolfgang Nejdl and Raluca Paiu, is motivated by one of the key goals of the Web 2.0 activity, social tagging: to help other users to find resources. The authors observe that this goal is impeded by the fact that users who are searching for content employ different keywords/tags than those set by tag creators. This difference between querying vocabulary and tag space leads to suboptimal results in search. The solution approach is to develop methods for recommending tags to creators that are more likely to be used in search. The input data are an increasingly relevant type of Web 2.0 data: tags attached to multimedia resources such as music and pictures. The semantic facets extracted from these data are usage (theme) and opinion (mood) characteristics of the annotated items; usage and opinion being what people tend to search for in such data. Bischoff et al. develop multi-class classifiers for recommending tags the annotators which are more likely to be used in subsequent queries. Their evaluation shows that the recommended annotations are of high quality, both compared to expert ground truth and rated by user judgments.

The second paper, *Automatic Construction of a Large-Scale Situation Ontology by Mining How-to Instructions from the Web* by Yuchul Jung, Jihee Ryu, Kyung-min Kim and Sung-Hyon Myaeng, is motivated by another key goal of the Web 2.0 and its precursors like UseNet: to share common problems and let others profit from solutions one has found. The authors observe that "HowTo"s on the Web are a rich source of data and knowledge, but are barely semantically structured today, which makes search for previous successful solutions to a problem like how to change a flat tire or how to make an airport check-in fast and safe depend on keyword search with all its well-known problems. The presented solution is to learn an ontology from the Web 2.0 material in two steps. First, action mining that extracts pairs of a verb and its ingredient (i.e., objects, location, and time) from individual instructional steps in the HowTo text is performed to form goal-oriented situation cases using the results. In the second step, the situation cases are normalized and integrated to form a situation ontology. Jun et al. evaluate the two steps of their approach and show its high quality by measuring accuracy of the action-mining method and by comparing the coverage of their situation ontology with existing large-scale ontology-like resources constructed manually. They discuss two application scenarios in detail: the detection of situation characteristics (and the ability to recommend actions) for context-aware services in ubiquitous computing, and the automatic composition of semantic web services.

The third paper, *Disambiguating Identity Web References using Web 2.0 Data and Semantics* by Matthew Rowe and Fabio Ciravegna, is motivated by a key characteristic of the Web 2.0: that the entities about which there is abundant information are no longer (only) resources such as products or problem-solving instructions, but the users themselves. In the decentrally organised Web, such information is typically not bundled at one place; rather, information in many places may refer to the same "physical" person. This may be wanted (for example, to achieve a separation of one's different social roles and identities), but it also harbours risks. For example, identity theft may occur: person A may be misrepresented and defrauded by someone else who poses as A and creates information on A somewhere on the Web without A's knowledge. A software monitoring what information the Web (2.0) as a whole contains about a person is an application that could be used to address that risk (as well as other goals). The authors therefore propose to monitor the web presence of a given individual by obtaining background knowledge to support automated disambiguation processes. They generate this background knowledge by exporting data from multiple Web 2.0 platforms as RDF data models and combining these models together for use as seed data. They present two disambiguation techniques. The first uses the semi-supervised machine learning technique of Self-training; the second uses the graph-based technique of Random Walks. The semantics of data support the intrinsic functionalities of these techniques. Rowe and Ciravegna evaluate their approach by comparing their disambiguation techniques against several baseline measures including human processing; both their techniques outperform several baselines measures.

Thus, together these three papers cover a comprehensive range of application areas, data types, and techniques, that together can be regarded as representative of current problems and solution approaches at the interface between the Social and the Semantic Web. New insights provided by machine learning and social network analysis techniques will lead to a new type of knowledge. We envision that research in this area will be of growing interest, as the automatic extraction of knowledge from weakly structured sources contributed by a huge mass of users and the combination with structured knowledge will be an important basis for the Semantic Web. It will lead to a broad range of new applications, which allow for combining knowledge of different types, levels and from different sources to reach their goals. The upcoming ubiquitous web is one target application area which will benefit from the newly integrated knowledge.