Journal of Zhejiang University-SCIENCE C (Computers & Electronics)
ISSN 1869-1951 (Print); ISSN 1869-196X (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Personal View:

Semantics and the crowd

Mark GREAVES
Vulcan Inc., Seattle, WA, USA
E-mail: markg@vulcan.com

doi:10.1631/jzus.C1101003

One of the principal scientific challenges that drives my group is to understand the character of formal knowledge on the Web. By formal knowledge, I mean information that is represented on the Web in something other than natural language text—typically, as machine-readable Web data with a formal syntax and a specific, intended semantics. The Web provides a major counterpoint to our traditional artificial intelligence (AI) based accounts of formal knowledge. Most symbolic AI systems are designed to address sophisticated logical inference over coherent conceptual knowledge, and thus the underlying research is focused on characterizing formal properties such as entailment relations, time/space complexity of inference, monotonicity, and expressiveness. In contrast, the Semantic Web allows us to explore formal knowledge in a very different context, where data representations exist in a constantly changing, large-scale, highly distributed network of loosely-connected publishers and consumers, and are governed by a Web-derived set of social practices for discovery, trust, reliability, and use. We are particularly interested in understanding how large-scale Semantic Web data behaves over longer time periods: the way by which its producers and consumers shift their requirements over time; how uniform resource identifiers (URIs) are used to dynamically link knowledge together; and the overall lifecycle of Web data from publication, to use, integration with other knowledge, evolution, and eventual deprecation. We believe that understanding formal knowledge in this Web context is the key to bringing existing AI insights and knowledge bases to the level of scale and utility of the current hypertext Web.

Technically, the scalability of the Semantic Web is rooted in a large number of independently-motivated participants with a shared vision, each following a set of carefully-designed common protocols and representation languages (principally dialects of the Resource Description Framework (RDF), the Web Ontology Language (OWL), and the SPARQL Protocol and RDF Query Language (SPARQL)) that run on top of the standard Web server and browser infrastructure. This strategy builds on the familiar hypertext Web, and has been incredibly successful. The Semantic Web now encompasses more than 50 billion Semantic Web assertions (triples) shared across the world via large numbers of autonomous Web servers, processed by situation-specific combinations of local and remote logic engines, and consumed by a shifting collection of software and users. However, this kind of loosely-coupled scalability strategy comes at a technical price: the Semantic Web is by far the largest formal knowledge base on the planet, and certainly one of the broadest, but also one of the messiest. Semantic coherence can be guaranteed only locally if at all, performance is spotty, data updates are unpredictable, and the raw data can be problematic in many ways. These problems impact the overall scalability of the Semantic Web; beyond simply exchanging large quantities of data, we also want the Semantic Web to scalably support queries, integration, rules, and other data processing tools. If we can solve these problems, though, the Semantic Web promises an exciting new kind of data Web, with practical scaling properties beyond what federated database technology can achieve. In the full Semantic Web vision, massive amounts of partially-integrated data form a dynamically shifting fabric of on-demand information, able

© Zhejiang University and Springer-Verlag Berlin Heidelberg 2012

to be published and consumed by clients around the world, with transformational impact.

Our current work is inspired by two properties of the Semantic Web: how existing Internet social ('crowd') phenomena can apply to data on the Semantic Web, and how we can use these social Web techniques to improve the dynamic scalability of the Semantic Web. Most data currently published on the Semantic Web is originally sourced from existing relational databases, either via front-end systems like the D2R server (http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/), or by offline loading of the relational data into an associated high-performance triple-store to support Semantic Web access and processing. In each case, the core information is usually acquired by conventional means, cleansed and structured into a relational store by a database administrator, and imbued with a particular data semantics that is eventually reflected in the Semantic Web republication. Thus, much of the data presently on the Semantic Web relies heavily on the traditional computer science discipline of database construction.

Because the Semantic Web exists within a framework of other Web practices, it supports additional techniques for data acquisition, structuring, and logical characterization—namely, those that derive from human social behavior on the Web. In particular, we have been exploring 'semantic wikis', a new type of wiki that combines semantic technology with wiki-based social mechanisms. The specific wiki we are developing is called Semantic MediaWiki+ (SMW+), and is based on the popular open-source MediaWiki software framework. (Complete documentation and source code for SMW+ can be found at http://www.smwplus.com. SMW+ relies heavily on the family of Semantic MediaWiki extensions, found at http://www.semantic-mediawiki.org.) Semantic wikis marry the structural discipline and query power of a database with the crowdsourcing power of an ordinary text wiki. In particular, they take advantage of the consensus property that many popular text wikis exhibit: the revision stream of the wiki articles with a sufficiently large and diverse number of authors will tend towards a (temporally individuated) set of fixpoints that Wikipedia calls the neutral point of view (NPV). Essentially, the NPV represents the effective consensus of all the wiki authors at a particular time. In MediaWiki-based wikis like Wikipedia, the factors that promote convergence to the NPV in article text include:

1. Articles are trivial to edit, all changes are tracked, and anyone can perform a rollback to an earlier version of the article.

2. Every article includes a 'Talk' page for discussion of proposed changes.

3. A general notification facility allows anyone to monitor articles of interest and to be automatically notified at every change.

4. Software bots can recognize certain kinds of vandalism to articles and auto-revert, or recognize articles that need work, and flag them for editors.

5. Certain pages are protected by a security system, and logins are often required for traceability.

6. (In Wikipedia) A hierarchy of administrators, gardeners, and editors helps enforce editing conventions.

SMW+'s MediaWiki heritage gives it all of the above capabilities, but it also contains a number of specialized software modules that supplement these NPV-promoting functions to extend the consensus property to cover user-entered data types and values. In this way, SMW+ merges graph-based data authorship (at roughly the RDF(S) level of expressivity) with the familiar socially-based edit/discussion/feedback features that are characteristic of text wikis. Additionally, the standard SMW+ package includes several modules that support data capabilities found in modern data environments, such as forms-based data entry, data and schema browsing, semantic query, data visualization, rules, Web service extensions, external procedures, and Linked Data mapping and integration tools.

The effect of blending a wiki with a database in this way allows us to experiment with social Internet phenomena and data authorship/publication within the user-friendly environment of a wiki. We have identified at least three distinct advantages. First, in contrast with the 'schema-first' design philosophy of traditional relational databases, semantic wikis are 'schema-last'. This means that, like the Semantic Web itself, the schema for data in SMW+ emerges from the actual patterns of properties and values that exist in the moment-to-moment RDF(S) data in the wiki. Individual data values, classes, and properties are wiki elements, represented on user-editable wiki pages, and subject to the same types of crowd-driven

consensus forces that govern the text-based elements of the wiki. Second, because all data elements in the wiki can be represented in familiar wiki pages, the explanatory metadata that goes with the data elements can be documented in ordinary text that is visually linked to the target data, promoting human readability, understandability, and comfort with the data. In this way, SMW+ serves as a type of collaborative semi-structured data authoring environment, enhanced by standard wiki text features. Third, the ability to embed powerful SPARQL-based query mechanisms into wiki articles allows users to easily assess wiki data scope and integrity, find conflicts and missing values in wiki-represented data, and locate areas where the emergent schema of the wiki data is not consistent or needs revision or refactoring. In this way, we can support rapid convergence on the NPV for data and schema in the wiki.

These three characteristics of SMW+ also allow us to start to address the dynamic scalability of the larger Semantic Web. One of the key difficulties that hinders the wide use of the Semantic Web is the large number of incompatible ontologies, data types, and identifiers that occur in Semantic Web data. This diversity makes it difficult to construct useful SPARQL queries that span multiple Semantic Web datasets, and in effect partitions the broad Semantic Web into individually-managed data islands of limited scope and range. Current methods to address this issue are derived from traditional data integration techniques, but the success of these relies on authority conditions and organizational relations that are designed for federated databases and do not scale to the size or diversity of the Web. Providing a tool that can leverage the 'wisdom of the crowd' to synthesize an NPV for Semantic Web style data is a powerful possibility. We look forward to expanding our work with SMW+ in this area, and believe that the crowd-sourcing techniques we are exploring will become a key approach for improving the dynamic scalability of the Semantic Web.

Recommended reading

Ankolekar, A., Krötzsch, M., Tran, T., Vrandecic, D., 2007. The Two Cultures: Mashing up Web 2.0 and the Semantic Web. Proc. 16th Int. Conf. on World Wide Web, p.825-834. [doi:10.1145/1242572.1242684]

Gruber, T., 2008. Collective knowledge systems: where the Social Web meets the Semantic Web. *Web Semant.*, **6**(1):4-13. [doi:10.1016/j.websem.2007.11.011]

Zaidan, F.H., Bax, M.P., 2011. Semantic wikis and the collaborative construction of ontologies: a case study. *J. Inform. Syst. Technol. Manag.*, **8**(3):539-554. [doi:10.4301/S1807-17752011000300002]

Hansch, D., Schnurr, H., Pissierssens, P., 2009. Semantic MediaWiki+ als Wissensplattform für Unternehmen. Proc. WM 2009: 5th Conf. on Professional Knowledge Management, p.211-215 (in German).

Vrandečić, D., Krötzsch, M., 2009. Semantic MediaWiki. *In*: Davies, J., Grobelnik, M., Mladenic, D. (Eds.), Semantic Knowledge Management. Springer, Heidelberg, p.171-179. [doi:10.1007/978-3-540-88845-1_13]