Research: Diversity and the Semantic Web

Denny Vrandečić, Elena Simperl, Rudi Studer – KIT

Twenty years after its introduction, the Web provides a platform for the publication, use and exchange of information, on a global scale, on virtually every topic, and representing an amazing diversity of opinions, viewpoints, mindsets and backgrounds. The success of the Web can be attributed to several factors, most notably to its principled scalable design, but also to a number of subsequent developments such as smart user-generated content, mobile devices, and most recently cloud computing. The first two of these have dramatically lowered the last barriers of entry when it comes to producing and consuming information online, leading to an unprecedented growth and mass collaboration. They are responsible for hundreds of millions of users all over the globe creating high-quality encyclopedias, publishing Terabytes of multimedia content, contributing to world-class software, and lively taking part in defining the agenda of many aspects of our society by raising their voices, and publicly expressing and sharing their ideas, viewpoints and resources.

The other side of the coin in this unique success story is, nevertheless, the great challenges associated with managing the sheer amounts of information continuously being published online, whilst allowing for purposeful use and leveraging the diversity inherently unfolding through global-scale collaboration as an asset. These challenges are still to be solved at many levels, from the infrastructure to store and access the information, through the methods and techniques to make sense out of it, to the paradigms underlying the processes of Web-based information provision and consumption.

Current information management methods and techniques at the core of essentially every channel one can use when attempting to interact with the vast ocean of information available on the Internet or elsewhere – be that Web search engines, news sites, eCommerce portals, online marketplaces, media platforms, the blogosphere or corporate intranets – are based on principles that do not reflect, and cannot scale to, the plurality of opinions and viewpoints captured in this information.

The term Linked Data refers to a set of principles and best practices for publishing and interlinking structured data on the Web, leading to a creation of a “global data space” called the Web of Data (also referred to as Linked Open Data). This global data space already consists of interlinked data covering diverse topics such as geographic locations, scientific knowledge, statistical data, and, perhaps most importantly from the perspective of diversity, people, social networks, and opinion statements like books, blogs, and tweets. One of the unique technical features of Linked Data is that it is directly published on the Web, so that the data is machine-readable, its meaning is explicitly defined and can in turn be linked to external data sets. Berners-Lee outlines the rules of how to publish data into the single, interlinked global space and these have become known as the Linked Data principles.

Several major ICT trends underlie the explosive growth expected in data provision in the coming years: the transition of closed enterprise systems to open Web-based models, the uptake of Linked Data principles by large data publishers, user-generated content published in an open way, the Internet of Services with the increased provision of data through services, and the Internet of Things and the parallel explosive growth in the amount of data being created by those devices.

Current modes of browsing and visualizing Linked Open Data are displaying the graph structure in a browseable table or form-like structure (see, e.g. Tabulator, Sig.ma, or VisiNav). Although these structures provide a viable method to access a small set of entities, the size and diversity of structured data in the large often leads such naïve representations of the RDF graph to become too complicated for most users to be processed and understood. Beyond simple provenance information, they do not provide any ways to detect bias or to easily recognize partial data.

Methods and approaches for visualizing diversity and selecting manageable sets of data from the Web of Data are becoming increasingly necessary. The smooth integration of data and data analytics into any Website, including the provenance trail of the data, and the ability to understand the data within its history and environment, will provide crucial in order to enable a Web where the consumer can always drill down to the raw facts, and fully follow the argumentation throughout the Web. This will keep the publishers accountable for the usage of the data, and for explaining how it links into their own argumentation.