Mining Machine-Readable Knowledge from Structured Web Markup

Ran Yu

GESIS - Leibniz Institute for the Social Sciences
50676 Köln, Germany
ran.yu@gesis.org

Abstract. The World Wide Web constitutes the largest collection of knowledge and is accessed by billions of users in their daily lives through applications such as search engines and smart assistants. However, most of the knowledge available on the Web is unstructured and is difficult for machines to process which leads to the lowered performance of such smart applications. Hence improving the accessibility of knowledge on the Web for machines is a prerequisite for improving the performance of such applications.

Knowledge bases (KBs) here refers to RDF datasets contains machine-readable knowledge collections. While KBs capture large amounts of factual knowledge, their coverage and completeness vary heavily across different types of domains. In particular, there is a large percentage of less popular (long-tail) entities and properties that are under-represented.

Recent efforts in knowledge mining aim at exploiting data extracted from the Web to construct new KBs or to fill in missing statements of existing KBs. These approaches extract triples from Web documents, or exploit semi-structured data from Web tables. Although the extraction of structured data from Web documents is costly and error-prone, the recent emergence of structured Web markup has provided an unprecedented source of explicit entity-centric data, describing factual knowledge about entities contained in Web documents. Building on standards such as RDFa, Microdata and Microformats, and driven by initiatives such as *schema.org*, a joint effort led by Google, Yahoo!, Bing and Yandex, markup data has become prevalent on the Web. Through its wide availability, markup lends itself as a diverse source of input data for KBA. However, the specific characteristics of facts extracted from embedded markup pose particular challenges.

This work gives a brief overview of the existing works on mining machine-readable knowledge from both structured and unstructured data on the Web, and introduces the KnowMore approach for augmenting knowledge bases using structured Web markup data.