Chapter VI

A Comparison of Semantic Annotation Systems for Text-Based Web Documents

Lawrence Reeve, Drexel University, USA
Hyoil Han, Drexel University, USA

Abstract

The Semantic Web promises new as well as extended applications, such as concept searching, custom Web page generation, and question-answering systems. Semantic annotation is a key component for the realization of the Semantic Web. The volume of existing and new documents on the Web makes manual annotation problematic. Semi-automatic semantic annotation systems, which we call platforms because of their extensibility and composability of services, have been designed to alleviate this burden for text-based Web documents. These semantic annotation platforms provide services supporting annotation, including ontology and knowledge base access and storage, information extraction, programming interfaces, and end-user interfaces. This chapter defines a framework for examining semantic annotation platform differences based on platform characteristics,
surveys several recent platform implementations, defines a classification scheme based on information extraction method used, and discusses general platform architecture.

Introduction

The Semantic Web, as described in Berners-Lee (1998), is the next generation of the Web providing machine-understandable information that is based on meaning. One way to provide meaning to Web information is by creating ontologies, and then linking information on a Web page to specifications contained in the ontology using a markup language (Berners-Lee et al., 2001). A key problem for the realization of the Semantic Web is providing these annotations for both existing and new documents on the Web. Semantic annotation is the process of mapping instance data to an ontology. Ontologies are conceptualizations of a domain that typically are represented using domain vocabulary (Chandrasekaran, Josephson, & Benjamins, 1999). Benefits of adding meaning to the Web include: query processing using concept-searching rather than keyword-searching (Berners-Lee et al., 2001); custom Web page generation for the visually-impaired (Yesilada, Harper, Goble, & Stevens, 2004); using information in different contexts, depending on the needs and viewpoint of the user (Dill et al., 2003); and question-answering (Kogut & Holmes, 2001).

It is not yet possible to automatically identify and classify all entities within source documents with complete accuracy (Popov et al., 2003). Manual annotation can be done using tools such as Semantic Word (Tallis, 2003), which provides a single interface for authoring as well as document markup. Manual approaches, however, suffer from several drawbacks. Human annotators can provide unreliable annotation for many reasons: complex ontology schemas, unfamiliarity with subject material, and motivation, to name a few (Bayerl, Lüngen, Gut, & Paul, 2003). It is expensive to have human annotators markup documents (Cimiano, Handschuh, & Staab, 2004), and the human annotator may not consider using multiple ontologies (Dingli, Ciravegna, & Wilks, 2003). Documents and ontologies can change, requiring new or modified markup, which leads to document markup maintenance issues (Dingli et al., 2003). Finally, the volume of existing documents on the Web can lead to an overwhelming task for humans to manually complete (Kosala & Blockeel, 2000). For all these reasons, manual efforts have been identified as a “knowledge acquisition bottleneck” (Maedche & Staab, 2001). Semi-automatic annotation platforms offer advantages over manual efforts, primarily document volume scalability through reduction of the human workload (Dill et al., 2003).
Related Content

Knowledge-Assisted Image Analysis Based on Context and Spatial Optimization
[www.irma-international.org/article/knowledge-assisted-image-analysis-based/2822/](www.irma-international.org/article/knowledge-assisted-image-analysis-based/2822/)

Web Data Quality: Current State and New Challenges
[www.irma-international.org/article/web-data-quality/116449/](www.irma-international.org/article/web-data-quality/116449/)

Forum Summarization to Support Tutor and Teacher in group interaction management
[www.irma-international.org/chapter/forum-summarization-support-tutor-teacher/52863/](www.irma-international.org/chapter/forum-summarization-support-tutor-teacher/52863/)

MERA: A Musical Entities Reconciliation Architecture Based on Semantic Technologies
[www.irma-international.org/article/mera/189764/](www.irma-international.org/article/mera/189764/)

SPORTAL: Profiling the Content of Public SPARQL Endpoints
[www.irma-international.org/article/sportal/160175/](www.irma-international.org/article/sportal/160175/)