Chapter 3

SODA: A Service Oriented Data Acquisition Framework

Andreea Diosteanu
Bucharest Academy of Economic Studies, Romania

Armando Stellato
University of Roma Tor Vergata, Italy

Andrea Turbati
University of Roma Tor Vergata, Italy

ABSTRACT

In this chapter, the authors present Service Oriented Data Acquisition (SODA), a service-deployable open-source platform for retrieving and dynamically aggregating information extraction and knowledge acquisition software components. The motivation in creating such a system came from the observed gap between the large availability of Information Analysis components for different frameworks (such as UIMA [Ferrucci & Lally, 2004] and GATE [Cunningham, Maynard, Bontcheva, & Tablan, 2002]) and the difficulties in discovering, retrieving, integrating these components, and embedding them into software systems for knowledge feeding. By analyzing the research area, the authors noticed that there are a few solutions for this problem, though they all lack in assuring a great level of platform independence, collaboration, flexibility, and most of all, openness. The solution that they propose is targeted to different kinds of users, from application developers, benefiting from a semantic repository of inter-connectable information extraction and ontology feeding components, to final users, who can plug and play these components through SODA compliant clients.

DOI: 10.4018/978-1-4666-0188-8.ch003
INTRODUCTION

While the Semantic Web (Berners-Lee, Hendler, & Lassila, 2001) is finally becoming a concrete reality, thanks to bootstrapping initiatives such as Linked Open Data (Bizer, Heath, & Berners-Lee, 2009) and assessment of W3C standards for expressing, querying, and accessing distributed knowledge, a large part of the information available from the web is still made available by traditional means: web pages and multimedia content.

To be able to cope with this huge volume of information, Information Extraction (IE) engines are allowed to lift relevant data from heterogeneous information sources and project it towards predefined knowledge schemes, thus enabling higher-level access based on semantic rather than textual indexing.

The purpose of such systems is actually twofold: if documents (and media in general) are the focus, then these systems may support systems for document management, advanced semantic search, smart document tracking, etc. by identifying references to entities already available in knowledge bases and indexing these documents with them; on the contrary, if the focus is on knowledge production, they may extract the information which is contained inside information sources and compose it into semantic compound resources that can then be fed to knowledge bases.

The success of semantic search engines such as Eqentia and Evri and Information Extraction and services such as OpenCalais and Zemanta show that there is large demand for this kind of solutions. However, all of them—while promising to break the old-fashioned concept of knowledge-silos by providing services and API for producing knowledge modeled according to open standards—still represent silos in their own offer, as users are not allowed to participate in the definition of new content lifters, nor can they access the code (or runnable instances) of the engines implicitly available through the provided services.

In this chapter we present a novel framework which aims to overcome the above limitation, being completely based on standard technologies (such as UIMA for Unstructured Information Management) and models (from RDF to Web Ontology Language [W3C, 2004] and Simple Knowledge Organization Systems [W3C, 2009]) and offering an open architecture and platform for provisioning of IE components, which may help in composing systems for semi-automatic development and evolution of ontologies by lifting relevant data from unstructured information sources and projecting it over formal knowledge models.

BACKGROUND

Related Works

Applications such as the ones described above are becoming more and more popular because they provide many advantages like: facilitating communication between computers, creating standardized metadata, and facilitating the management and processing of various document formats. However, the development of such applications requires lot of time and development effort.

Development-free solutions are based on services: this is the case of Thomson Reuters’ OpenCalais. OpenCalais offers valid and robust services for tagging documents. However, the main drawbacks of its offer consist in the limitations of a completely web-based approach: it does not assure user document privacy and also it does not offer any possible vertical customization to different domains. Under a certain point of view, we could say that Open-Calais is much closer than what its name would suggest.

Other approaches have been proposed in the scientific literature, describing less-opaque frameworks for composition of NLP components. In Wilkinson, Vandervalk, & McCarthy (2009), SADI—Semantic Automated Discovery and Inte-
Related Content

Shadow Sensitive SWIFT: A Commit Protocol for Advanced Data Warehouses
www.igi-global.com/chapter/shadow-sensitive-swift/46894?camid=4v1a

Preview: Optimizing View Materialization Cost in Spatial Data Warehouses
www.igi-global.com/chapter/preview-optimizing-view-materialization-cost/39587?camid=4v1a

A Framework for Examining Data and Taking Action to Increase Student Retention and Graduation
www.igi-global.com/chapter/framework-examining-data-taking-action/60854?camid=4v1a

An Overview of Shallow and Deep Natural Language Processing for Ontology Learning
Amal Zouaq (2011). Ontology Learning and Knowledge Discovery Using the Web: Challenges and Recent Advances (pp. 16-37).
www.igi-global.com/chapter/overview-shallow-deep-natural-language/53879?camid=4v1a