Chapter 30
The KnowledgeStore: A Storage Framework for Interlinking Unstructured and Structured Knowledge

Francesco Corcoglioniti
FBK-irst, Italy

Marco Rospocher
FBK-irst, Italy

Roldano Cattoni
FBK-irst, Italy

Bernardo Magnini
FBK-irst, Italy

Luciano Serafini
FBK-irst, Italy

ABSTRACT

Although the quantity of structured information on the Web and within organizations is increasing, the majority of information remains available only in unstructured form. While different in form, both unstructured and structured information sources provide information about entities in the world and their properties and relations; still, frameworks for their seamless integration have not been deeply investigated. In this paper the authors describe the KnowledgeStore, a scalable, fault-tolerant, and Semantic Web grounded open-source storage system for interlinking structured and unstructured data. They present the concept, design, function and implementation of the system, and report on its concrete usage in three application scenarios within the NewsReader EU project, where it stores and supports the querying of millions of news articles interlinked with millions of RDF triples extracted from text and imported from Linked Open Data sources. The authors report on data population and data retrieval performances of the system measured through a number of experiments, and they also discuss the practical issues and lessons learned from these experiences.

DOI: 10.4018/978-1-5225-5191-1.ch030

1. INTRODUCTION

With Semantic Web (SW) technologies coming of age and the public acclaim of the Linked Open Data (LOD) initiative, the last few years have seen a massive proliferation of structured data,[1] both on the Web and within organizations. Nonetheless, the majority of information remains available only in unstructured form.[2] While different in form, both unstructured and structured information sources provide information about entities in the world (e.g., persons, organizations, locations, events), their properties, and relations among them. Indeed, coinciding, contradictory, and complementary facts about these entities could be available in structured form, unstructured form, or both, and content available in one form may help in better interpreting the information contained in the other, something that may turn out to be crucial in applications where having "complete" knowledge is a requirement (e.g., situations where users have to make potentially critical decisions).

The last decades achievements in Natural Language Processing (NLP) now enable the large scale extraction of knowledge about world entities from unstructured text (Weikum & Theobald, 2010; Grishman, 2010), thus setting the basis to combine knowledge coming both from unstructured and structured content. However, the development of frameworks enabling the seamless integration and linking of knowledge available in structured and unstructured forms has only been partially investigated.

In this paper we present the KnowledgeStore, a scalable, fault-tolerant, and Semantic Web grounded storage system to jointly store, manage, retrieve, and query both structured and unstructured data. To illustrate the capabilities and peculiarities of the KnowledgeStore, let us consider the following scenario. Among a collection of news articles, a user is interested in retrieving all 2014 news reporting statements of a 20th century US president where he is positively mentioned as "commander-in-chief". On one side, the KnowledgeStore supports storing of *resources* (e.g., news articles) and their relevant metadata (e.g., the publishing date of a news article). On the other side, it enables storing structured content about *entities* of the world (e.g., the fact of being a US president, the event of making a statement), either extracted from text or available in LOD/RDF datasets (e.g., DBpedia[3], Yago[4]), in a *contextualized* fashion (e.g., someone is US president only for a certain period of time). And last, through the notion of *mention*, it enables linking an entity or fact of the world to each of its occurrences in documents, allowing also to store additional information (*mention attributes*, typically extracted while processing the text) for each specific occurrence in a document: to name a few, the position of the entity/fact in the text (e.g., between character 1022 to 1040), the explicit way it occurs (e.g., "commander-in-chief"), and the sentiment of the article writer on that particular occurrence (e.g., positively mentioned). Besides supporting the storage and management of this content, the KnowledgeStore provides query and retrieval mechanisms that enable to access all the information it contains and can be used to answer the user query presented above.

Thanks to the explicit representation and alignment of information at different levels, from unstructured to structured knowledge, the KnowledgeStore enables the development of enhanced applications, and favors the design and empirical investigation of information processing tasks otherwise difficult to experiment with. On the one hand, the possibility to semantically query the content of the KnowledgeStore with requests combining knowledge from structured sources and unstructured sources, similarly to the example previously discussed, allows a deeper exploration and analysis of stored data, a capability particularly useful in applications such as *decision support*. On the other hand, the joint storage of structured knowledge (both background and extracted knowledge), the resources it derives from, and mention information — all effectively accessible through a single API — provides an ideal scenario for developing, debugging, training, and evaluating tools for a number of NLP and knowledge processing