Chapter 2
Managing Large Volumes of Interlinked Text and Knowledge With the KnowledgeStore

Francesco Corcoglioniti
Fondazione Bruno Kessler, Italy

Marco Rospocher
Fondazione Bruno Kessler, Italy

Roldano Cattoni
Fondazione Bruno Kessler, Italy

Bernardo Magnini
Fondazione Bruno Kessler, Italy

Luciano Serafini
Fondazione Bruno Kessler, Italy

ABSTRACT

This chapter describes the KnowledgeStore, a scalable, fault-tolerant, and Semantic Web grounded open-source storage system to jointly store, manage, retrieve, and query interlinked structured and unstructured data, especially designed to manage all the data involved in Knowledge Extraction applications. The chapter presents the concept, design, function and implementation of the KnowledgeStore, and reports on its concrete usage in four application scenarios within the NewsReader EU project, where it has been successfully used to store and support the querying of millions of news articles interlinked with billions of RDF triples, both extracted from text and imported from Linked Open Data sources.

DOI: 10.4018/978-1-5225-5042-6.ch002

INTRODUCTION

The last decades achievements in Natural Language Processing (NLP) and Knowledge Extraction (KE) have enabled the large-scale extraction of structured knowledge about world entities from unstructured text (Weikum & Theobald, 2010; Grishman, 2010; Vossen et al., 2016; Corcoglioniti, Rospocher, & Palmero Aprosio, 2016). As a result, new application scenarios are appearing where large amounts of information are available in different interlinked forms: text, the knowledge extracted from it, and the NLP annotations involved in the KE process. To support applications having to jointly store, access, and process all this information, there is an increasing need for scalable frameworks that seamlessly integrate structured and unstructured knowledge, providing the necessary scalability (e.g., up to millions of documents and billions of RDF triples) and data access and manipulation methods.

This chapter describes the latest achievements on the KnowledgeStore (http://knowledgestore.fbk.eu) extending the work previously reported by Corcoglioniti, Rospocher, Cattoni, Magnini, and Serafini (2015). The KnowledgeStore is a scalable, fault-tolerant, and Semantic Web (SW) grounded open-source (Apache License v2.0) storage system to jointly store, manage, retrieve, and query interlinked text and RDF knowledge extracted from it, e.g., using KE tools such as PIKES (Corcoglioniti, Rospocher, & Palmero Aprosio, 2016), or coming from Linked Open Data (LOD) resources. Conceptually, the KnowledgeStore acts as a data hub populated by KE systems and queried by end users and applications, whose contents are organized according to three representation layers: Resource, Mention, and Entity. To illustrate the interplay of these layers in the KnowledgeStore, and the capabilities it offers, 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 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 knowledge about entities of the world – e.g., the fact of being a US president and the event of making a statement – either extracted from text or available in LOD/RDF datasets such as DBpedia (Lehmann et al., 2015) and YAGO (Hoffart, Suchanek, Berberich, & Weikum, 2013). And last, through the notion of mention, it enables linking an entity or fact of the world to each of its specific occurrences in documents – e.g., a US president to the documents mentioning him – allowing also the storage of additional mention attributes, typically extracted while processing the text, such as the explicit way the entity or fact occurs – e.g., "commander-in-chief" – and the sentiment of the article writer on that entity – e.g., positively mentioned. Besides supporting the scalable storage and management of this content, through an architecture compliant with the deployment in distributed hardware settings like clusters and cloud computing, the KnowledgeStore provides a ReST API and a user interface supplying query and retrieval mechanisms that enable accessing all its contents, and thus answering the example 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