Chapter 4
Interactive Summaries by Multi-Pole Information Extraction for the Archaeological Domain

Angela Locoro
University of Genova, Italy

Massimo Ancona
University of Genova, Italy

ABSTRACT
Understanding and describing past or present societies is a complex task, as it involves a multi-faceted analysis of the norms, interactions, and evolutions that characterize them. This serves as the motivation for developing a tool, named Herodotus, aiming at supporting domain experts, such as historians or archaeologists, in the reasoning tasks over complex interactions characterizing a society in order to explain why some event took place and, possibly, to predict what could happen when some factors change. An important part of Herodotus is the text mining module that is responsible for the extraction of knowledge from written sources, such as books and scientific papers. Machines cannot always help users in dealing with natural language, because of the variety, ambiguity and non-rigidity that language shows in its use; they can only try to process information in a meaningful way for users. Information Extraction (IE) is the technology that pulls specific information from large volumes of unstructured texts and stores this information in structured forms. Users can then consult, compose, and analyze them. Domain-based IE should focus on an analysis of a specific state of affairs and, in this way, it can obtain more precise and detailed results. This helps domain experts to deal with the complexity of their everyday objects and environments. This chapter is centered on the Interactive Summary Extractor tool, whose scope is to organize, in a partially automated but substantially interactive way, text summaries for archaeological and historical documental sources. The texts so analyzed will help domain experts to collect data, viewing a synthesized version of it, compose such summaries in units of sense for the particular archaeological study or research that is in place, and so on. Summaries can then be modified, stored, retrieved and managed for later elaboration.

DOI: 10.4018/978-1-4666-5019-0.ch004
INTRODUCTION AND MOTIVATIONS

Searching and structuring information for archaeologists is a complex task: information has to be collected from disparate sources in different languages and environments, often apparently unrelated, and structured into an organic (multimedia) form. For example, documentation on the mechanism of “Antikythera” (“Antikythera Mechanism,” 2013) has an unexpected, but fundamental relation with Cicero, mentioning that two of such devices have been constructed by Archimedes. This fact, paired with the sophistication of the device technology, suggested to the authors of the research in (Edmunds & Freeth, 2011), the assertion: “even if only one device of this kind is so far known to have survived, it was based on the tradition of a long line of precursors.”

Another example is represented by the description of the embalming process performed by Ancient Egyptians and described in detailed form by Herodotus in volume II of his book “Histories,” from which the archaeologists could formulate ideas on how this ancient practice was performed. These examples point out how wide is the environment and how disparate are the sources to be explored by archaeologists for collecting information about specific historic facts, techniques, objects and instruments.

Our system, named Herodotus, is going to be an evolution of a previous system (Ancona et al., 2005), which will serve as a support for making plans for the preservation of the natural and cultural heritage represented in such scenarios. In addition, the same tool used by the domain experts might be exploited for presenting the cultural and natural heritage and for training people in knowing, appreciating, respecting and, possibly, helping to safeguard it.

More specifically, the aim of Herodotus is to semi-automatically extract knowledge from collections of ancient books, scientific publications, paintings, images and documents reporting findings of archaeological excavations, to create a formalized electronic institution, or e-institution, (Bogdanovych, 2007) modeling an open multi-agent system (MAS) that mimics norms and relations of the real life relative to a specific archaeological area and historical time.

In this Chapter, we will describe the Interactive Summary Extractor Tool (ISET from now on), which will be part of the Herodotus text mining module and whose prototype is under implementation and evaluation. Text Mining may span from simple keyword matching (KM), which only superficially determines meaning, to a complete parsing, structuring and reasoning task, which attempts to understand the whole semantics of texts. The term information extraction (IE) usually defines activities similar to the last example. Nonetheless, KM may provide and support more granularities when trying to combine atomic methods to locate sentences of importance to the user’s subject area and application. There are several structures significant for archaeologists that can be searched and extracted from plain texts, ranging from simple keywords and key-phrases to text summaries, object properties and relations, as well as rules and even ontologies or taxonomies.

ISET is an interactive tool for the summarization of single documents. The tool functionalities range from keywords and phrases extraction to text summaries. In this version, all the intelligent actions are intended to be committed to the care of a human user: the system computes keywords frequencies and key-phrases related to them, but the task of organizing or structuring all the concepts discovered is demanded to the user. In particular, for ad-hoc texts, summarization will have the purpose of highlighting all portions of text that could be relevant to a specific theme. The originality of our tool lays in two main factors:

- The interactivity and lightness of the approach, resulting in more flexibility, user-oriented capacities, and easy access where a low speed connectivity may be the only available.
Related Content

A Taxonomy for Distance-Based Spatial Join Queries
www.igi-global.com/article/a-taxonomy-for-distance-based-spatial-join-queries/185656?camid=4v1a

Spatial OLAP and Map Generalization: Model and Algebra
Sandro Bimonte, Michela Bertolotto, Jérôme Gensel and Omar Boussaid (2012). *International Journal of Data Warehousing and Mining* (pp. 24-51).
www.igi-global.com/article/spatial-olap-map-generalization/61423?camid=4v1a

A Clustering Rule Based Approach for Classification Problems
Philicity K. Williams, Caio V. Soares and Juan E. Gilbert (2012). *International Journal of Data Warehousing and Mining* (pp. 1-23).
www.igi-global.com/article/clustering-rule-based-approach-classification/61422?camid=4v1a

Social, Ethical and Legal Issues of Data Mining
www.igi-global.com/chapter/social-ethical-legal-issues-data/7611?camid=4v1a