Chapter 9

A Semantic Approach for Geospatial Information Extraction from Unstructured Documents

Christian Sallaberry • Mauro Gaio • Julien Lesbegueries • Pierre Loustau

Abstract. Local cultural heritage document collections are characterized by their content, which is strongly attached to a territory and its land history (i.e., geographical references). Our contribution aims at making the content retrieval process more efficient whenever a query includes geographic criteria. We propose a core model for a formal representation of geographic information. It takes into account characteristics of different modes of expression, such as written language, captures of drawings, maps, photographs, etc. We have developed a prototype that fully implements geographic information extraction (IE) and geographic information retrieval (IR) processes. All PIV prototype processing resources are designed as Web Services. We propose a geographic IE process based on semantic treatment as a supplement to classical IE approaches. We implement geographic IR by using intersection computing algorithms that seek out any intersection between formal geocoded representations of geographic information in a user query and similar representations in document collection indexes.

9.1 Introduction

Smart spatial information extraction and retrieval in repositories of electronic documents is the main goal of the work presented in this chapter. The semi-structured and nonstructured data are supported by Electronic Document Management Systems (EDMS) or Library Management Systems (LMS). All these systems aim at providing fast and effective content-based access to a large amount of information. But if we consider that they usually implement statistical approaches to retrieve information, they are insufficient for queries in which the semantics of the search criteria concerns spatial relations (Clementini et al. 1994).

The Virtual Itineraries of the Pyrenees (PIV) project manages a repository of electronic versions of books, newspapers, postal cards and lithographs of the 19th and 20th centuries. It appears that the information is supported by heterogeneous documents but presents many local sources of cultural heritage and denotes various Pyrenean territorial aspects (Cazenave et al. 2004). This kind of repository is still quite unknown. Moreover, it is accessible only in local-area archives of museums and libraries. This is the reason why the regional media library (MIDR) supports this project and intends to diffuse these resources. Thereby the PIV project proposes a semantic approach to analyzing and interpreting geographic information contained in such a corpus or query (Etcheverry et al. 2005; Marquesuzàà et al. 2005). The PIV system proposes to extend basic services of existing LMSs to include new services dedicated to the marking and retrieval of geographic information. It relies on a specific open architecture based on Web services as well as a model describing
geographic information and XML indexes. The originality of our approach is the geographic core model that allows one to formalize any geographic information, regardless of its mode of expression (i.e., text, image). This approach is based on the incremental enrichment of electronic documents. It supports complex processing streams that involve various resource types. The results of each subprocess produce a new XML stream upon which the subsequent subprocesses can rely.

This chapter is dedicated to the marking and indexing of geospatial information. We present related work in Section 9.2 and the PIV geographic core model in Section 9.3. In Section 9.4, we present the PIV prototype; its spatial information extraction and retrieval processes are dealt with in Sections 9.5 and 9.6. Finally, we present an evaluation of this information extraction prototype in Section 9.

9.2 Approaches for Specific Geographic Needs

Information extraction (IE) generally organizes indexes in order to better support information retrieval (IR). Natural Language Processing (NLP) allows specific IE from textual documents, i.e., named entity recognition on diverse types of text (Maynard et al. 2003). Used together, these approaches have the potential to create powerful tools in content-based information systems (Gaizauskas and Wilks 1998).

9.2.1 Information Extraction and Retrieval

IE may be described as the activity of populating a structured information repository from an unstructured information source (Gaizauskas 2002). In a collection of documents, the result of an IE process constitutes what is called an index. It generally consists of a list of terms linked to each document (Tebri 2004). These terms have to describe as precisely as possible the contents of the documents. The automatic IE processes extract either the entire information of a document or only specific parts of it. For example, in the former case, textual processes generally use statistical approaches (all terms of a document are treated) to associate a weight to each term (Zipf 1949). However, in the latter one, they use predefined rules in order to find out specific information (Gaizauskas 2002).

IR deals with models, techniques and procedures to extract information that has already been treated, organized and stored – databases, files, XML files, etc. (Baeza-Yates and Ribiero-Neto 1999). As IE and IR approaches are rather generic, accurate management of spatial information is yet a great challenge.

9.2.2 Natural Language Processing

When dealing with textual documents, a "standard" NLP is based on a set of processing resources (Abolhassani et al. 2003) sequentially applied to all the textual flow: (1) a tokenizer is used for splitting the text into tokens; (2) a splitter is used to segment the text into logical substructures like sections, paragraphs and sentences; (3) a Part-of-Speech tagger produces a POS tag for each token in each sentence; (4) finally, a semantic tagger generally consists of grammar sets. Each grammar set contains a series of rules, which act on previously assigned tags in order to produce annotated sets of tokens and/or sentences.

Some systems, like Brill (Brill 1992), Cordial⁴ and Tree-Tagger (Schmid 1994), are dedicated resources of such subprocesses. Other ones, like GIPSY (Woodruff and Plaunt 1994), Linguastream (Bilhaut 2003; Widlocher and Bilhaut 2005), SPIRIT (Jones et al. 2004) and GATE (Gaizauskas et al. 1995; Cunningham et al.