From Text to Semantic Geodata Enrichment

Khaoula Mahmoudi, Laboratory of Remote Sensing and Information Systems for Spatial Reference, Tunis, Tunisia

Sami Faiz, Laboratory of Remote Sensing and Information Systems for Spatial Reference, Tunis, Tunisia

ABSTRACT

Geographically referenced information are captured, managed, analyzed and displayed by the geographic information system (GIS). This information will help to better understand and analyze a host of human activities in an attempt to reach a more sustainable future. Moreover, these information are sometimes hardly available due to the cost entailed by the acquisition process and the maintenance. Hence, efforts are devoted to bring the interesting information with less cost. In this context, the authors propose an approach dedicated to the enrichment of the semantic aspect of geographic database (GDB). The approach emerges from the agents’ interaction in conformity with the multi-agents framework. It deals with an Interface agent, Task agents and Geographic ones. These agents are to process a corpus of documents related to the geographic entity at hand. This processing consists of extracting the essence of information from the managed corpus to enrich the GDB. These agents follow a pipelined approach consisting of the following stages: text fragmentation, thematic annotation, delegation and text filtering. Besides these basic stages, a refinement is to be performed in case of user unsatisfaction and this by exploring the spatial relationships.

Keywords: Geographic Information System (GIS), Multi-Agents System, Semantic Data Enrichment, Spatial Relationships, Text Filtering, Text Segmentation, Thematic Identification

INTRODUCTION

Designed to capture, store and manage geographical data, Geographic Information System (GIS) (Faïz & Krichen, 2013) is being widely used to solve problems in a host of domains. As a decision making system, the GIS power is to help managers make critical decisions they face daily. Indeed, conducting an adequate analysis requires detailed and high level information. Typically, spatial databases do not contain much information that could support the decision making process in all situations. Besides, Jack Dangermond (President of a private GIS software company) argued that “The application of GIS is limited only by the imagination of those who use it”.

In front to such situation, the enrichment of the GDB incarnated to the GIS was proposed as a preponderant solution for consistent analy-
sis. In this context, we propose an approach (Mahmoudi & Faïz, 2012) which is based on processing textual data to extract the gist of information relative to the geographic entity at hand. The idea underpinning our approach is to provide a condensate view from a corpus of documents relative to the geographic entity. Since the information are distributed among different documents, a distributed processing is the best tailored resolution way. This distribution is materialized via a multi-agents (Ferber, 1999, Mihaylov, 2012) modeling. Hence, three classes of agents are involved to the data enrichment process, it deals with: Interface agent, Task agents and eventually Geographic agents. The main stages of the proposed approach are: segmentation, thematic identification, delegation and text filtering. A subsidiary stage may be solicited once the primary results are not satisfactory enough, it deals with the refinement. The latter relies on best locating the entity under test to best target the required information. This stage is accomplished by exploring the spatial relationships.

The rest of the paper is organized as follows: In the next section, a special emphasis is put on the works devoted to semantic data enrichment of GDB. In the section after that, we outline the approach we propose to perform this enrichment. The section following that is dedicated to the text segmentation to identify different segments that encompass the whole text. Section 5 reports the theme identification to label the detected segments. The delegation to distribute equitably the filtering task among the system agents is detailed in the next section. The section following is devoted to the filtering of the segments to maintain only the most salient information. In the section after that, we describe the refinement stage executed eventually in case of unsatisfactory results reported by the basic data enrichment stages. The implementation of our approach is reported in the next section. Finally, in the last section, we present the evaluation of our system.

BACKGROUND

In the context of GIS, the enrichment process falls into two basic classes: the semantic enrichment and the spatial one.

For the spatial enrichment it may be operated in the context of the generalization process (Plazanet, 1996). The information to complement the GDB spatial aspect is explored to supply procedural and geometrical knowledge. The latter are to guide the choice of the adequate generalization solution.

Regarding the semantic enrichment, it operates on the semantic data (also referred to as aspatial, descriptive or thematic) stored in the GDB. A set of works may be classified under this data enrichment category.

MetaCarta (2005) may be seen as a bridge between the GIS systems and a set of documents. This is materialized by applying the Geographic Text Search (GTS). GTS allows to link textual documents to geographic entities localized in digital maps to add supplementary data to GDB. GTS is offered as an extension to the GIS ArcGIS.

The GeoNode (Geographic News On Demand Environment) (Hyland et al., 1999) adopts the MITRE’s Alembic system (Day et al., 1997) to carry out the enrichment. Alembic system allows extracting the named entities highlighted in each story. It consists of exploring as input news stories to depict events happening at particular time and place. The GIS ArcView supports GeoNode.

Initially conceived as a Digital Library, Persus (David, 2002) focuses on historical documents relative to past events. Persus has incorporated tools allowing to GIS to make use of the historical collection to bring out knowledge to enrich the GDB. The processing of the collection consists in determining the terms, the toponyms, the dates and estimating the co-occurrence of the dates and emplacements to determine eventual events. This mean of enrichment is explored by the TimeMap GIS.