Chapter XXXI
Automatic Data Enrichment in GIS Through Condensate Textual Information

Khaoula Mahmoudi
High School of Communications – Tunis (SUPCOM), Tunisia

Sami Faïz
National Institute in Applied Sciences and Technology (INSAT), Tunisia

INTRODUCTION

Geographic Information Systems (GIS) (Faïz, 1999) are being increasingly used to manage, retrieve, and store large quantities of data which are tedious to handle manually. The GIS power is to help managers make critical decisions they face daily. The ability to make sound decisions relies upon the availability of relevant 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"*. Hence, it is of primary interest to provide other data sources to make these systems rich information sources.

To meet these information requirements, data enrichment strategies have been undertaken, broadening the amount of information the users can access. To enrich data stored in the geographic database (GDB), we extract knowledge from online textual documents corpora. This is accomplished by using multi-document summarization (Barzilay et al., 2005). To obtain complementary data in a reasonable time, we propose a distributed multi-document summarization approach. We do so, because the summary generation among a set of documents can be seen as a naturally distributed problem. Hence, each document is considered as an agent, working towards a concise representation

Copyright © 2009, IGI Global, distributing in print or electronic forms without written permission of IGI Global is prohibited.

of its own document, which will then be included as part of the corpus summary. In conformity with the multi-agent paradigm (Ferber, 1999), we use a set of cooperating autonomous agents: an *Interface* agent, *Geographic* agents, and *Task* ones. These agents collaborate to jointly lead the system to an optimal summary. The approach we propose is modular. It consists of: thematic delimitation, theme identification, delegation, and text filtering. Besides these main steps, we propose a refinement process that can be executed in the case of unsatisfactory results.

The reminder of the chapter is organized as follows. In section II, we present a review of the enrichment literature. Section III is dedicated to our data enrichment process. In section IV, we describe the refinement process. The implementation of our approach is reported in section V. Finally, in section VI, we present the evaluation of our system.

BACKGROUND

Typically the database enrichment can be classified as; spatial enrichment and semantic one.

Spatial enrichment is interested to the spatial aspect of the GDB. One use of this enrichment is its application within the overall generalization process. In this context, the newly acquired information are used to provide geometrical knowledge and procedural knowledge in terms of generalisation algorithms and operations order, to guide the choice of the generalization solution (Plazanet, 1996).

For the semantic enrichment, it aims to link unstructured data to the already available structured thematic data (also referred as aspatial, descriptive or semantic) stored in the GDB. The works classified under this category are: Metacarta (MetaCarta, 2005), GeoNode (Hyland et al., 1999), Persus (David, 2002).

MetaCarta's technology provides a bridge between document systems and GIS systems. MetaCarta, allows the semantic enrichment through the Geographic Text Search (GTS). GTS allows to link textual documents to geographic entities localised in digital maps to add supplementary data to GDB. GTS is offered as an extension to the GIS *ArcGIS*.

For GeoNode (Geographic News On Demand Environment), given a sequence of news stories, GeoNode, can identify different events that happen at particular time and place. GeoNode makes use of the Information Extraction technique and more precisely the *Alembic* system to accomplish the enrichment. The latter allows to determine the named entities for geospatial visualizations. The GIS *ArcView* supports GeoNode.

Persus is initially conceived as Digital Library focusing 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 estimate the co-occurrence of the dates and emplacements to determine eventual events. This mean of enrichment is explored by the *TimeMap* GIS.

What distinguish the data enrichment approach we propose, is that it goes beyond merely providing and linking the relevant documents related to geographic entities, we instead process the documents to locate the gist of information embedded into them. Besides, our approach exploits the spatial relations inherent to GIS to refine the enrichment results.

MAINTHRUST: OUR DATA ENRICHMENT PROCESS

The data enrichment process we propose (Mahmoudi & Faïz, 2006$_a$, Mahmoudi & Faïz, 2006$_b$, Faïz & Mahmoudi, 2005) emanate from informational requirements claimed by GIS users. Hence,

7 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/chapter/automatic-data-enrichment-gis-through/20712

Related Content

On the Versatility of Fuzzy Sets for Modeling Flexible Queries
P Bosc, A Hadjaliand O Pivert (2008). *Handbook of Research on Fuzzy Information Processing in Databases (pp. 143-166).*
www.irma-international.org/chapter/versatility-fuzzy-sets-modeling-flexible/20352

ONTOMETRIC: A Method to Choose the Appropriate Ontology
Adolfo Lozano-Telloand Asunción Gomez-Perez (2004). *Journal of Database Management (pp. 1-18).*
www.irma-international.org/article/ontometric-method-choose-appropriate-ontology/3308

Fuzzy Querying Capability at Core of a RDBMS
Ana Aguilera, José Tomás Cadenasand Leonid Tineo (2011). *Advanced Database Query Systems: Techniques, Applications and Technologies (pp. 160-184).*
www.irma-international.org/chapter/fuzzy-querying-capability-core-rdbms/52301

Evolutionary Algorithms in Supervision of Error-Free Control
Bohumil Sulcand David Klimanek (2010). *Soft Computing Applications for Database Technologies: Techniques and Issues (pp. 39-48).*
www.irma-international.org/chapter/evolutionary-algorithms-supervision-error-free/44381

Keyword-Based Queries Over Web Databases
Altigran S. da Silva, Pável Calado, Rodrigo C. Vieira, Alberto H.F. Laenderand Bertheir A. Ribeiro-Neto (2003). *Effective Databases for Text & Document Management (pp. 74-92).*
www.irma-international.org/chapter/keyword-based-queries-over-web/9206