INFORMATION RETRIEVAL AND VISUALIZATION BASED ON DOCUMENTS' GEOSPATIAL SEMANTICS

Sallaberry Christian
LIUPPA
UPPA,
64013 PAU Cedex, France
Christian.Sallaberry@univ-pau.fr

Etcheverry Patrick
LIUPPA, UPPA,
IUT de Bayonne
64100 BAYONNE, France
Patrick.Etcheverry@iutbayonne.univ-pau.fr

Marquesuzaà Christophe
LIUPPA, UPPA,
IUT de Bayonne
64100 BAYONNE, France
Christophe.Marquesuzaa@iutbayonne.univ-pau.fr

***Abstract*—Local cultural heritage document collections are characterized by contents strongly attached to a territory and its history. Our contribution aims at enhancing such a content retrieval process efficiently each time a query includes geographic criteria.**

We propose a core model for a formal representation of geographic information. It takes into account the characteristics of different expression modes: written language and captures of drawings, maps, photographs, etc. We have developed a prototype fully implementing geographic Information Extraction (IE) and geographic Information Retrieval (IR) processes. We approach geographic IE from semantic processings additionally to classic IE approaches. This paper focuses on IR and Information Visualization (IV) proposals relying on the geospatial characteristics of documents.

Index Terms — Geographic Model, Geographic Information Retrieval and Visualization, Non-Structured Documents, Digital Libraries, Cultural Heritage.

I. INTRODUCTION

Smart spatial information retrieval and visualization is the main goal of the work presented in this paper. Although GISs (Geographic Information Systems) contain high-level spatial operators that are uncommon in conventional RDBMSs (Relational Data Base Management Systems), they are not sufficient for queries in which the semantics of the search criteria concerns spatial relations [9]. The results are also unsatisfying if we consider EDMSs (Electronic Document Management Systems) or LMSs (Library Management Systems) that usually implement statistical approaches to answer such queries.

In a study of a log of the Excite search engine, [27] found that about one fifth of all queries were geographical, as determined by the presence of a geographical term such as a place name, a post code, a type of place or a directional qualifier such as north. The purpose of the Virtual Itineraries in the Pyrenees[1] (PIV) project[2] consists in managing a repository of digitalized books, newspapers, lithographs, postcards of the XIXth and XXth century. The MIDR media library[3] supporting this project aims at the diffusion of these resource collections: information is mainly textual and presents many Pyrenean territorial aspects [18]. We can say that about two fifth of our digital library queries contain geographical criteria. Hence, the PIV system proposes to upgrade basic services of existing LMSs with new services dedicated to geographic information extraction, retrieval and visualization. It uses a specific open architecture based on web services, a core model describing geographic information, and XML indexes to better manage geographic marks. The originality of our approach lies in the core model allowing to formalize any geographic information whatever its expression mode (*i.e.* text, image) is. To complete LMSs' statistical and full-text processes, we propose a more accurate semantic approach to analyze geographic information contained in such a corpus (or query) [19].

We present related works in the second section and the PIV geographic core model in the third section. The PIV geographic content-based information retrieval and visualization are presented in sections four and five.

II. SPATIAL INFORMATION MANAGEMENT WITHIN HETEROGENEOUS DOCUMENTS COLLECTIONS

Information Extraction (IE) generally organizes indexes for a better support Information Retrieval (IR). IE and IR used together have the potential to create powerful new tools in information processing [14]. This section describes IE, IR and Information Visualization (IV) approaches which are combined for specific geographic requirements.

A. Information Extraction

IE may be described as the activity of populating a structured information repository from an unstructured information source [14]. In a collection of documents, the result of IE constitutes what is called an index. It is generally made of a list of terms linked to each document [5]. These terms have to describe as precisely as possible the contents of the documents. Automatic IE processes extract either the whole information of a document or, specific parts of it. For

[1] Mountains of the south west of France

[2] Acknowledgements to the Pau metropolitan council and media library

[3] MIDR: Médiathèque Intercommunale à Dimension Régionale

example, in the first case, textual processes generally use statistical approaches (each term of a document is treated) [5] to associate a weight to each term while, in the second one, they use predefined rules in order to find out specific information [14].

B. Information Retrieval

IR deals with models, techniques, and procedures to extract information that has already been treated, organized and stored (databases, files, XML files, etc.). [3] explains that satisfying user information requirements is not trivial: "*The user first specifies a user need which is then parsed and transformed. Then, query operations might be applied before the actual query, which provides a system representation for the user need, is generated. The query is then processed to obtain the retrieved documents. Fast query processing is made possible by the index structure previously built.*". [3] shows the importance of the query validation and/or reformulation stages to improve the interpretation of the user need.

C. Information Visualization

The retrieval of search engine results is a recurring problem [12]. Here is a summary of the different restitution modes that pinpoints the four most used approaches:

1) List representation: It is a representation method used by most search engines (*e.g.* Google[4]). Lists allow to present indefinite result numbers and also provide a simple access mode to the items. However, they do not offer a synthetic overview of the results and the browsing is often limited to the first results [26].

2) Topic representation: A topic gathers the elements around same concept. The Ujiko search engine[5] uses this type of representation. [25] showed that, in comparison with a non-hierarchical classification, such a classification could give better performances in terms of user satisfaction level. This representation offers a synthetic overview but choosing the topics to be represented is subjective and difficult to computerize.

3) Graph or Tree representation: It is close to topic classification but it integrates the semantics carried by the edges connecting the vertices (which can represent topics or documents). The Kartoo search engine[6] or topic maps [20] with tools such as TM4J[7] use this form of representation. The semantic links between concepts allow guided navigation between topics and documents. However, there is a limit relative to the number of concepts that can be represented with this approach.

4) Cartographic representation: The principle consists of representing a space (area, country, building, room, device layout) and associating on this space the data elements related to various points highlighted on this space. An example of geographical application is SPIRIT - Spatially-Aware Information Retrieval on the Internet[8] - which proposes a search engine whose results are Web pages geolocalized on a map. This form of representation integrates the space dimension but raises problems when the results are geolocalized on very close places, or on the same place.

D. Spatial Information

1) Linguists' works: They explain our specific manner of representing spatial information in written language. According to [7], we can link a place to a category and associate it with a natural or with an artificial boundary. We consider three main categories: land parcels, expanses of water, dwelling places. Referring to such places involves several elements. In written language, one might define spatial information by referring to a better known position. We thus understand sentence 1 perfectly while sentence 2 seems to be unusual to us:

- sentence 1: "the car is near the house"
- sentence 2: "the house is near the car"

[23] studied this assumption for textual documents and explained the concept of target/site couple. Our objective is to extend this hypothesis to any other expression modes.

2) GIS works: They present a Geographic Feature (GF) as a user-defined geographic phenomenon that can be modeled or represented using geographic data sets. Examples of geographic features include streets, sewer lines, manhole covers, accidents, lot lines, parcels[9]. Important related works address models of spatial relations [9], qualitative spatial representation and reasoning [10] [21] and spatial queries processing [9]. Other interesting works [17] concern digital gazetteers (Alexandria Digital Library[10]) which support important related dictionaries of geographic names and references [2]. GIS literature mainly represents a geographic feature by its name and its location. The location covers many facets:

- topographical coordinates, with geometric possibilities (the point or the polygon coordinates to position the building on a map);
- topological, direction or metric relations with other GFs (a direction relation to detail the position of the building within the village) [11] [15];
- conceptual links with topics of spatial theory within a specific ontology [16].

As IE, IR and I$\mathcal{V}$ approaches are rather generic, spatial information accurate management is yet a great challenge. Moreover, LMSs can neither take into account the geographic semantics of documents, nor the users' specific geographic requirements. Semantics processing seems to be an interesting way of spatial information management within IE, IR and I$\mathcal{V}$.

[4] www.google.com
[5] www.ujiko.com
[6] www.kartoo.com
[7] http://tm4j.org
[8] www.geo-spirit.org
[9] www.webgis.net/cms.php/glossary.htm
[10] www.alexandria.ucsb.edu - www.alexandria.ucsb.edu/gazetteer/

E. *Semantics processing*

It allows specific information extraction, *i.e.* focusing on the spatial information.

In textual expression mode, data processing sequence used for highlighting spatial markers is composed of four main steps [1]:

- lemmatization carries out a segmentation of the words;
- lexical and morphological analysis proceeds to a word recognition;
- syntactic analysis, based on grammars, allows to find the bonds between words;
- "semantic" analysis carries out a more specific analysis allowing the extracted syntagms to be interpreted.

Some systems like Brill[11], Cordial[12] or Tree-Tagger[13] - morphosyntactical analysers- are dedicated to a specific part of such sub-processes. Other systems like Linguastream[14] [24], SPIRIT[15] or GATE[16], [13] support the whole process.

In graphic expression mode, semantic processes consider that an image is not represented in single pixels but in meaningful image segments and their mutual relations. [22] proposes semantics definition to represent spatial data. [4] presents fuzzy methods implementing expert spatial knowledge and describes a workflow from remote sensing imagery to GIS. eCognition system provides a new technology for image analysis[17].

III. A Geographic Core Model

In this model, according to the linguistic hypothesis, a GF is recursively defined from one or several other GFs and spatial relations are part of the GFs' definition. The target/site principle [23] can approximately but reasonably be defined in a recursive way.

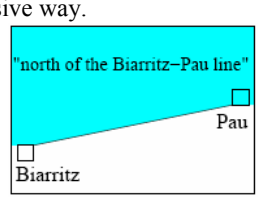


Figure 1. A GF expressed in a text or a schema.

For instance, the GF "*north of the Biarritz-Pau line*" (graphically and textually expressed in figure 1):

- is first defined by sites "*Biarritz*" and "*Pau*" that are well known named places,
- then, term "*line*" creates a new well known geometrical object linking the two sites and cutting the space into two sub-spaces,
- finally, an orientation relation creates a reference on the target to focus.

[11] www.cs.jhu.edu/brill/
[12] www.synapse.com
[13] www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
[14] www.linguastream.org
[15] www.spiritengine.com/
[16] http://gate.ac.uk/
[17] www.pcigeomatics.com/products/definiens.html

In Figure 2 it appears that a GF has at least one representation (A) with a natural or artificial boundary; it can be specialized (B) into an absolute feature (AGF) *i.e.* a named place or into a relative feature (RGF). A RGF is defined with a reference *i.e.* a relation linking at least one other GF (C). The cycle represents the recursive definition.

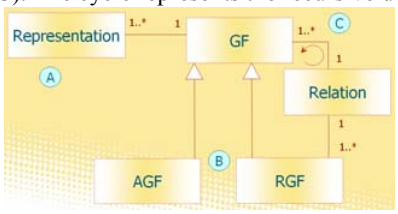
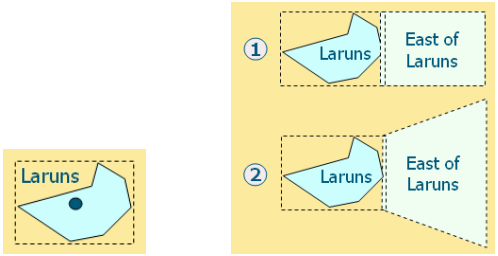


Figure 2. Geographic core model simplified schema.

Therefore, a GF can be either:

- an Absolute Geographic Feature (AGF) if it only consists in a well known named place *i.e.* a toponym with its geocode,
- or a Relative Geographic Feature (RGF) if it is defined using a spatial relation (generally topological) linking at least a GF (that can be an AGF or another RGF).

For textual IE, this approach has been adapted into a recursive grammar presented in [6]. A GF spatial relation is an adjacency, an inclusion, a distance, a geometric form or an orientation [6]. GFs representations rely on their relations description in the Geographic Model as well as on external gazetteers. So, a GF's representation is a set of given or computed geocoded data [6]: description attributes, precise and/or approximated shapes that may be Minimum Bounding Rectangles (MBR) *e.g.* Figure 3. For instance, a RGF ("*East of Laruns*") representation size is computed according to its AGF ("*Laruns*") size and its spatial relations semantics ("*East*").



1/ possible simplification
2/ more realistic interpretation

Figure 3. Representations of "Laruns" and "east of Laruns" village.

IV. Geographic Content-Based Information Retrieval

The PIV system aims non-expert users (teacher, learner, or tourist) to access to territorial-oriented digitized corpora. [6] exposes the geographic information detection and marking process (IE in an electronic documents corpus) as well as these markers operating within an IR process (user query processing). PIV system operates a semantic processing of geographic information within the corpus indexing stage and the query analyzing stage [6]. In the same

way, the information visualization process exploits the geographic characteristics of the results.

During IE stage instances of the geographic model are created and stored into index files. Any instance is constituted by the name of the feature, its interpretation (AGF or RGF, relations) and a corresponding geometric object (representing the concerned area).

A free-text querying interface supports the IR stage. Any query is analyzed with the same process than for the creation of the indexes of the corpus documents. The IE sequence is processed and every GF of a query is extracted. Then, all the validated GFs are geo-localized and a MBR is attached to each one of these GFs.

Our search technique is based on a spatial mapping between the query's GFs and the documents' GFs. This mapping processes the MBRs created dynamically for the query and the geographic representations stored in index files of the corpus.

For example, Figure 4 illustrates a query ("*I want documents dealing with places which are near Laruns village.*") and its corresponding MBR (the biggest one). The other shapes represent GFs (Pyrenean villages and roads extracted from documents of our corpus) that may match the query. Indexed documents' RGFs are represented by MBRs whereas AGFs are represented by more precise geometrics shapes. The relevance of a document is computed from its GF and the query's GF intersecting surface rate.

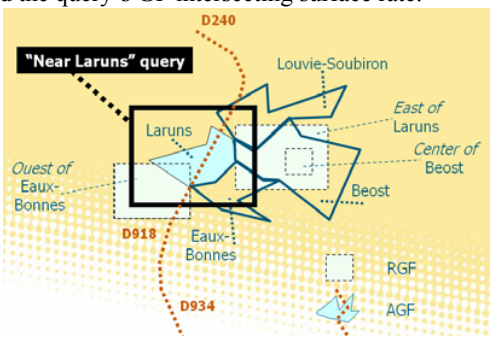


Figure 4. A query example.

eXist XML DBMS[18] supports indexes management and relevant documents content access.

V. Geographic Content-based Information Visualization

The visualization of territorial information is taken into account on two levels within our PIV prototype: on the one hand, that of graphical interpretation of the user request and, on the other hand, that of result presentation. Like in SPIRIT (*cf.* II-D/E), we chose a cartographic visualization mode which allows taking into account the territorial specificity of our documents.

[18] http://exist.sourceforge.net/

A. Graphical interpretation and refinement of the request

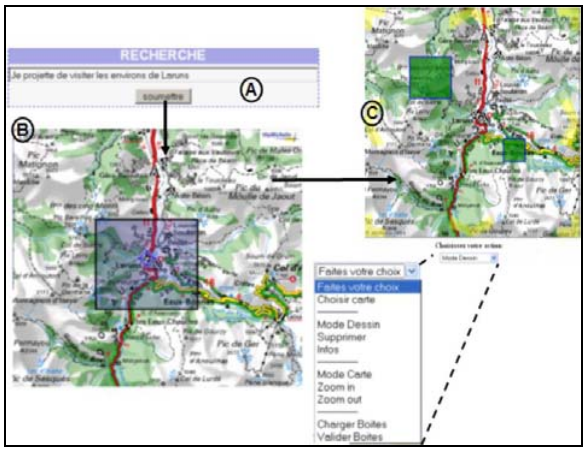


Figure 5. Querying territorial information in PIV.

At this stage, the system interprets the needs textually expressed by the user in order to provide a visual representation of the request which will be performed. In order to favor access to the documents according to space criteria, the system identifies the GFs which are present in the request (Figure 5 - A) and tries to represent them on a map (Figure 5 - B). The user thus visualizes the interpretation of the system according to its geographical interest. He/she can consequently validate this interpretation and launch a search, or, refine this interpretation by indicating more precise areas of interests (Figure 5 - C).

B. Visualization of request results

At this stage, the system presents the documents likely to answer the user expectations. We suppose that, if the user searches for documents according to spatial criteria, his/her interest (partly) relates to space. As any document containing at least a GF is geolocalizable, the resulting documents are presented on a map according to the places they mention (Figure 6).

Figure 6. Visualization of the results.

The geolocalized representation of the results entails a spatio-documentary navigation. Space becomes the access criterion to the documents since the user will not consult a document according to its title or author but according to the place which it mentions. This new way of browsing within a document corpus requires to propose new browsing marks which are built at the same time on spatial information (for example, to consult the documents close to the current place) and on documentary information (for example, to know the already consulted documents). We are currently studying the feasibility of proposing the user to browse within the corpus as if he/she was on a map, while passing from one document to another as if he/she went from one place to another.

VI. CONCLUSION

We focus our work on restricted corpora such as local cultural heritage collections of documents. This specific context allows implementing sensible scans which take into account the document contents. Our contribution is complementary to library traditional search methods. Our objective is to process the geographic semantics of such collections of documents and users' queries in a more accurate way. The PIV prototype implements and combines original geographic semantics Information Extraction (IE), Information Retrieval (IR) and Information Visualization (IV) approaches. Its experimentation with heterogeneous (texts and images) documents collections shows that this approach enhances the relevance of geographic query results. A first IE process evaluation [28] has just been finished and an IR process evaluation is on the go.

The key features of our proposal are:

- the geographic core model description: it supports a formal description of every Geographic Feature (GF) detected in collections of text or image documents. Complex GFs (Relative GFs) include other GFs: they are recursively defined from Absolute GFs and/or Relative GFs;
- any GF's geo-reference is found out in gazetteers or approached using relation semantics interpretation and the Minimum Bounding Rectangle (MBR) approach: the IR process is based on the intersections of geographic representations extracted from a user's query and representations of documents collections within indexes;
- a spatio-documentary navigation approach supports information visualization: it relies on the geographic characteristics of retrieved information within visualization and navigation scenarios.

Further work concerns the enhancement of information visualization and navigation. We will therefore design new scenarios to take into account the context and the user requirements both during the querying step and the visualization one. We are also planning to manage the features of a new topic. PIV document collections are characterized by contents referring to a territory and its history. This is the reason why time dimension management would obviously improve our IR process.

REFERENCES

[1] M. Abolhassani, N. Fuhr, and N. Govert, "Information Extraction and Automatic Markup for XML documents", Springer, p. 159–174. 2003.

[2] F. B. Zhan, "How much is region q covering region r 'a little bit,' 'somewhat,' or 'nearly completely'?". M. Cristani and B. Bennett (Eds.), SVUG01: The First COSIT (Conference on Spatial Information Theory) Workshop on Spatial Vagueness, Uncertainty and Granularity, Morro Bay, CA, September 2001.

[3] R. A. Baeza-Yates and B. A. Ribeiro-Neto, "Modern Information Retrieval". ACM Press / Addison-Wesley, 1999.

[4] U. C. Benz, P. Hofmann, G. Willhauck, I. Lingenfelder and M. Heynen, "Multi-resolution, object-oriented fuzzy analysis of remote sensing data for GIS ready information". International Workshop - Semantic Processing of Spatial Data - Geopro, 2002.

[5] Zipf. "Human Behaviour and the Principle of Least Effort". Addison., 1949.

[6] J. Lesbegueries, M. Gaio, P. Loustau and C. Sallaberry, "Geographical information access for non-structured data", ACM SAC ASIIS, pp. 83-89, 2006.

[7] A. Borillo, "L'espace et son expression en français". L'essentiel. Ophrys, 1998.

[8] T. Charnois, Y. Mathet, P. Enjalbert, and F. Bilhaut. "Geographic reference analysis for geographic document querying". Workshop on the Analysis of Geographic References, Human Language Technology Conference (NAACL-HLT), Association for Computational Linguistic, 2003.

[9] E. Clementini, J. Sharma, and M. Egenhofer, "Modeling topological spatial relations: Strategies for query processing". Computers and Graphics 18 (6): 815-822, 1994.

[10] A. G. Cohn and S. M. Hazarika, "Qualitative spatial representation and reasoning: An overview". Fundamenta Informaticae, 46(1-2):1–29, 2001.

[11] M. J. Egenhofer, "Toward the semantic geospatial web". In GIS '02: Proceedings of the 10th ACM international symposium on Advances in geographic information systems, pp. 1–4. ACM Press, 2002.

[12] N. Bonnel, F. Moreau, "Quel avenir pour les moteurs de recherche ?" In Proceedings of the workshop Manifestation des Jeunes Chercheurs francophones dans les domaines des STIC, MajecSTIC 2005, Rennes, France, pp. 291-299, november 2005.

[13] R. Gaizauskas, T. Wakao, K. Humphreys, H. Cunningham and Y. Wilks, "University of sheffield: Description of the lasie system as used for muc", 1995. http://acl.ldc.upenn.edu/M/M95/M95-1017.pdf

[14] R. Gaizauskas and Y. Wilks, "Information extraction: Beyond document retrieval". Journal of Documentation, 54(1):70–105, 1998.

[15] N. Gotts and J. Goodday, "A connection based approach to common-sense topological description and reasoning". The Monist. pp. 51-75. 1996. citeseer.ifi.unizh.ch/gotts95connection.html

[16] N. Hernandez, "Ontologies pour l'aide à l'exploration d'une collection de documents". In Ingénierie des Systèmes d'Information, vol. 10, pp. 11–31. Hermès Sciences, 2005.

[17] L. Hill, "Core elements of digital gazetteers: Place names, categories, and footprints". In ECDL '00: Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries, pp. 280–290. Springer-Verlag, 2000.

[18] J. Casenave, C. Marquesuzaà, P. Dagorret and M. Gaio, "La revitalisation numérique du patrimoine littéraire territorialisé". In Colloque International EBSI-ENSSIB, Montréal, Octobre 2004. www.ebsi.umontreal.ca/rech/ebsi-enssib/ebsi-enssib-programme.html.

[19] C. Marquesuzaà, P. Etcheverry, and J. Lesbegueries, "Exploiting geospatial markers to explore and resocialize localized documents". In Proceedings of the first International Conference on GeoSpatial Semantics, GeoS 2005, Mexico City, nov. 29-30, Lecture Notes in Computer Science, Vol. 3799, pp. 153-165, 2005.

[20] J. Cossanel, J.P. Cahier, M. Zacklad, J. Charlet, "Les Topic Maps sont-ils un bon candidat pour l'ingénierie du Web Sémantique ? " In Proceedings of conférence Ingénierie des Connaissances IC2002, Rouen, Mai 2002

[21] P. Muller, "Topological spatio-temporal reasoning and representation. Computational Intelligence", 18(3):420–450, 2002.

[22] M. Torres, "Semantics definition to represent spatial data. International Workshop - Semantic Processing of Spatial Data" - Geopro, 2002.

[23] C. Vandeloise, "L'espace en français". Travaux Linguistiques. Seuil, 1986.

[24] A. Widlocher and F. Bilhaut, "La plate-forme linguastream : un outil d'exploration linguistique sur corpus". In Actes de la 12e Conférence Traitement Automatique du Langage Naturel, 2005.

[25] B. Kules and B. Schneiderman, "Categorized graphical overviews for web search results: An exploratory study using U.S. government agencies as a meaningful and stable structure". Proceedings of the Third Annual Workshop on HCI Research in MIS, Washington, D.C., December 10-11, 2004

[26] A. Leuski and J. Allan, "Lighthouse: Showing the way to relevant information". In Steven F. Roth and Daniel A. Keim, editors, *Proceedings of IEEE Symposium on Information Visualization (InfoVis'00)*, pp. 125-130, Salt Lake City, Utah, USA, October 9-10, 2000. IEEE Computer Society.

[27] M. Sanderson and J. Kohler, "Analyzing geographic queries". In Proceedings of the Workshop on Geographic Information Retrieval, SIGIR 2004, www.geo.unizh.ch/~rsp/gir/

[28] C. Sallaberry, M. Gaio, J. Lesbegueries and P. Loustau, "PIV: a Geographic Content-Based Documents Management System". LIUPPA Report DocEng2006.