SEARCH ON THE WEB WITH SPATIAL CRITERIONS
Improving a Search Engine with Spatial Queries

Jose E. Córcoles, Pascual González and Marcos Rodriguez
LoUISE Research Group. Universidad de Castilla-La Mancha 02071 Albacete, Spain

Keywords: Spatial Query, Search Engine, Spatial Search Engine, Ontology.

Abstract: In tackling the Semantic Web, a rich domain that requires special attention is the Semantic Geospatial Web. In order to achieve the *Semantic Geospatial Web*, an approach to query non-spatial resources using spatial criterions is presented in this paper. The non-spatial resources are well-known web resources represented by HTML, GIF, PDF, etc. However, to use spatial criterions a set of spatial resources with spatial information represented with GML (geographic Markup Language) is necessary. In this way, queries with spatial operators (touch, within, near, etc.) can be carried out over resources. In this approach, the relations between spatial and non-spatial resources are carried out "on the fly" in each query, i.e., they are not manually pre-stored. So, the main aim of this approach is to boost the capacity of Search Engines (such as Google, Yahoo!, etc) in order to include spatial criterions in the queries and not only string matches.

1 INTRODUCTION

With the growth of the World Wide Web has come the insight that currently available methods for finding and using information on the Web are often insufficient. A rich domain that requires special attention is the *Semantic Geospatial Web* (Egenhofer, 2002). The enormous variety of encoding of geospatial semantics makes it particularly challenging to process requests for geospatial information. In the future, the *Semantic Geospatial Web* will allow the returning of both spatial and non-spatial resources to simple queries, using a browser. For example, a query "*lakes in Maine"* should return all relational resources with lakes in Maine (pictures, text, ...) in different formats (XML, HTML, JPG, PDF, References, ...) (Egenhofer, 2002).

Actually, the Web is riddled with spatial information. For example, any web page about a restaurant or hotel has spatial information about its location, such as city, street, number, etc. In the same way, any document about a lake has relative location, such as near a particular mountain or in the northern part of a country. However, this spatial information is usually described with a natural language. Therefore, it is difficult to apply spatial operators to this information in order to obtain documents on the basis of spatial criterions.

Thus, in order to reach the *Semantic Geospatial Web*, an approach to query non-spatial resources using spatial criterions is presented in this paper. The non-spatial resources are well-know web resources represented by HTML, GIF, PDF, etc. However, in order to use spatial criterions, a set of spatial resources (with spatial information) is necessary. These spatial resources are represented by *Geographical Markup Language* - GML. In this way, queries with spatial operators can be carried out over these resources. In this approach, the relations between spatial and non-spatial resources are carried out "on the fly" in each query, i.e., they are not manually pre-stored.

The main aim of this approach is to boost the capacity of Search Engines (such as Google, Yahoo!, etc) in order to include spatial criterions in the queries and not only string matches. It provides the infrastructure for formulating structured spatial queries by taking into consideration the conceptual representation of a specific domain in the form of an ontology.

In our approach the spatial information is represented by GML because it is an XML encoding for the transport and storage of spatial/geographic information, including both spatial features and non-spatial features. The mechanisms and syntax that

E. Córcoles J., González P. and Rodriguez M. (2007).
SEARCH ON THE WEB WITH SPATIAL CRITERIONS - Improving a Search Engine with Spatial Queries.
In *Proceedings of the Third International Conference on Web Information Systems and Technologies - Web Interfaces and Applications*, pages 382-385
DOI: 10.5220/0001270303820385

GML uses to encode spatial information in XML are defined in the specification of OpenGeospatial Consortium (Open Geospatial Consortium, 2003). Thus, GML allows a more homogeneous and flexible representation of the spatial information.

Our system follows a classical architecture based on *mediator*. Thus, we have used a *mediator* layer, which is responsible for integrating GML resources and executing the spatial queries. It processes the user queries and sends full or partial queries to the GML sources. The sources return alphanumeric values, and these values are sent to a Search Engine to search for the associated resources.

Query mediation has been extensively studied in the literature for different kinds of mediation models such as *Tsimmis* (Papakonstantinou et al., 1995), *YAT* (Cluet et al., 2001), (Levy et al., 1996) and (Boucelma et al., 2002). More directly concerned with the spatial XML integration, the approaches developed by (Gupta et al., 1999), (Córcoles and González., 2003) and (Córcoles et al., 2003) stand out. (Gupta et al., 1999) extends the MIX *wrapper-mediator* architecture for integrating information from spatial information systems and searchable databases of geo-referenced imagery. MIX is focused on integrating geo-referenced imagery but our approach is focused on spatial geometries. On the other hand, (Córcoles et al., 2003) designed a novel approach for integrating GML resources. The proposed architecture uses a Catalog expressed by RDF to relate the GML resources. Also, (Córcoles and González., 2003) describes a mediation system based on RDF for querying spatial and non-spatial information. Finally, (Córcoles, 2005) shows an architecture for integrating spatial resources and non-spatial resources on the web. It has the same aims as this paper, but it does not use commercial Search Engines to obtain non-spatial resources.

With regard to the use of current Search Engines for searching for resources based on spatial conditions, they are now starting to take their first steps. So, for example, *Google Maps* (Maps, 2006) offers an approach to provide information about businesses following spatial conditions. However, it offers an ad-hoc solution and not a relation between spatial and non-spatial resources on the fly.

This paper is structured as follows: An overview of the architecture is shown in Section 2 and conclusions and projected future work are shown in Section 3.

2 AN OVERVIEW

Figure 1 shows the basic architecture of our approach. It has two main sections: *mediator* and sources (Wrappers). *Mediator* stores Spatial resources (named GML resources) and has the main logic for executing the spatial queries to obtain non-Spatial resources (non-GML resources) related to the GML resources. *Mediator* has two aims: on the one hand, it relates GML resources and makes it possible to search for them, and on the other hand, it carries out the search for non-GML resources.

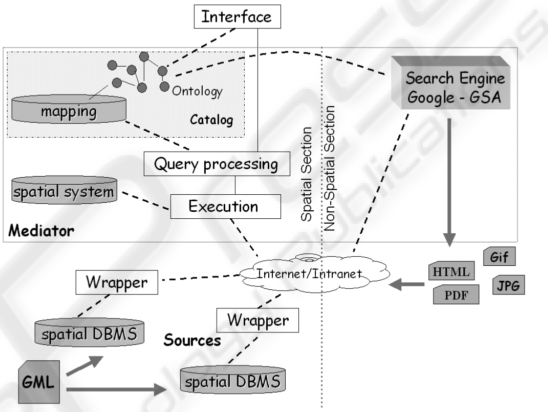
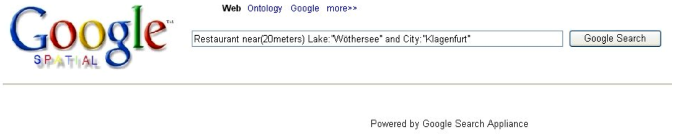


Figure 1: Architecture of the approach.

Non-GML resources (web pages, gif, pdf, etc) are stored on "Internet", though we suppose a copy is stored in the Search Engine. In the implementation of this approach we have based our work on the Google Search Engine. To be precise, we have used the Google Search Appliance (GSA). It is an integrated hardware and software product designed to give businesses the productivity-enhancing power of Google search (Google, 2006). However, though our implementation has been limited to Google Search Appliance, every feature of this approach can be applied over Google Search Engine or other Search Engines. For this reason, in this paper we use the term *Search Engine* in a general way.

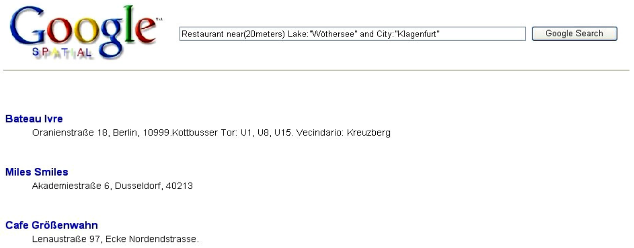
GML resources are stored in the sources (in the same or different locations), whose purpose is that of executing the queries received from the *mediator* and returning the results.

The main task of an integration *mediator* is to provide users with a unique interface for querying the data, independently of its actual organisation and location (Levy, 2000). This interface, or global schema, is described as an *ontology*.



Restaurant near(20meters) Lake:"Wöthersee" and City:"Klagenfurt"

Figure 2: Spatial Query.



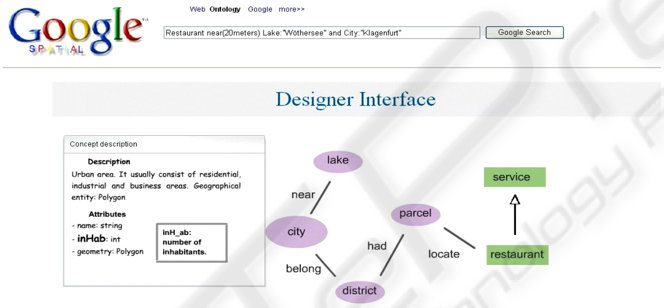
Restaurant near(20meters) Lake:"Wöthersee" and City:"Klagenfurt"

Bateau Ivre
Oranienstraße 18, Berlin, 10999.Kottbusser Tor: U1, U8, U15. Vecindario: Kreuzberg

Miles Smiles
Akademiestraße 6, Dusseldorf, 40213

Cafe Größenwahn
Lenaustraße 97, Ecke Nordendstrasse.

Figure 3: Results.



Restaurant near(20meters) Lake:"Wöthersee" and City:"Klagenfurt"

Concept description

Description

Urban area. It usually consist of residential, industrial and business areas. Geographical entity: Polygon

Attributes

- name: string
- **inHab**: int
- geometry: Polygon

inH_ab: number of inhabitants.

Figure 4: Ontology in the User Interface.

To evaluate a user query expressed in terms of the ontology, the *mediator* translates it into one or more queries on the GML sources. For this purpose, we need to establish a correspondence between each source and the global ontology. This correspondence is described by a *mapping*.

When the correspondence between each source and the global ontology has been carried out, the system can execute spatial queries over the GML resources, and then over non-GML resources. In order to execute the queries, we need a spatial query language over the ontology and an execution plan in order to know what GML or non-GML resources satisfy the query. If a source *s* can only partially answer a query, then the query is decomposed into two parts: one to be fully answered by *s*, the other part being sent to the other sources. In this case, it needs a variable binding algorithm and a query execution plan that includes query decomposition for searching all sources for which there exists a full/partial binding. This process is discussed in (Córcoles and González., 2004).

Therefore, the description of the global schema in terms of the ontology allows users to formulate structured queries, without being aware of the source specific structure.

Figure 2 shows the main query web pages defined in a prototype that implements the concepts defined in this paper. Because we have used a Google Search Appliance - GSA, we have used a similar web page design to *Google*. (Note that the logotype *GoogleSpatial* is not a *Google* trademark. It is only a play on words used in the prototype).

A result page is shown in Figure 3 Here we group the resources by each result returned for the GML resources. Finally, Figure 4 shows an example of our ontology which a user can look at to understand the available concept in the community.

3 CONCLUSIONS

We have used a system based on a *mediator* to execute spatial queries on the web. We distinguish between non-spatial resources and spatial resources. The non-spatial resources are represented by HTML, GIF, PDF, etc., and the spatial resources are represented by Geographical Markup Language - GML.

GML resources can be stored in different sources. For this reason, we have defined an execution plan (QEP) to optimize the execution. QEP has two phases, the first one is executed over GML resources and the second phase is executed over the Search Engine. QEP processes the user queries and sends full or partial queries to the GML sources. The sources return alphanumeric values (*keys* of the concepts), and these values are sent to a Search Engine to search for the associated resources.

Non-GML resources are riddled with spatial information in an unstructured textual way, and resources have spatial information, i.e., geometries, and structured textual descriptions. Therefore, in order to relate GML resources to non-GML resources we try to find the value of the *keys* of the concepts involved in the spatial query in the Search Engine.

Future work foresees an approach for automatically loading the mapping rules in the Catalog. Moreover, we are performing a study in order to exhaustively validate the quality of the results. Furthermore, in order to increase the usability of the query language, adaptative techniques are being studied to help the user to understand the ontology and write the queries.

ACKNOWLEDGEMENTS

We are grateful to Sitesa – Grupo (Spain) for kindly allowing us to use its Google Search Appliance and its infrastructure. They contributed to making our work a reality.

REFERENCES

Boucelma, O., Essid, M., Lacroix., Z., 2002. A WFS-Based Mediation System for GIS Interoperability. In *Proceedings of ACM-GIS 2002. 10th ACM International Symposium on Advances in Geographic Information Systems*,. McLean, USA.

Cluet, S., Veltri, P., Vodislav, D., 2001. Views in a large Scale XML Repository. In *Proceedings of VLDB*, Rome, Italy.

Córcoles, J., González, P., 2003. Querying Spatial Resources. An Approach to the Semantic Geospatial Web. *In Proceedings of CAiSE'03 workshop "Web Services, e-Business, and the Semantic Web (WES)". Lecture Notes in Computer Science (LNCS)* Springer-Verlag, *LNCS 3095*, pp. 41-50.

Córcoles, J., González, P., López-Jaquero, V., 2003. Integration of Spatial XML Documents with RDF. *International Conference on Web Engineering (ICWE03). Spain. Lecture Notes in Computer Science (LNCS)* by Springer-Verlag. *LNCS 2722*, pp. 407-410..

Córcoles, J., González, P., 2004. Integrating GML Resources and Other Web Resources. *DEXA'04 Workshops. 1° International Workshop on Geographic Information Management, GIM 2004. IEEE Computer Society*, pp. 872-877.

Córcoles, J., 2005. Integración de Recursos Espaciales y No-Espaciales en la Web: Un Acercamiento a la Web Semántica Geoespacial. *Ph.D. Dissertation. UCLM University,* Spain, pp. 244.

Egenhofer, M., 2002. Toward the Semantic Geospatial Web. In *Proceedings of ACM-GIS 2002. 10th ACM International Symposium on Advances in Geographic Information Systems,* McLean, USA.

Google, 2006 www.google.co.uk/enterprise/gsa. Acceded in 2006

Gupta, A., Marciano, R., Zaslavsky, I., Baru, C., 1999. Integrating GIS and Imagery through XML based information Mediation. Integrated Spatial Databases: Digital Images and GIS. *Lecture Notes in Computer Science. Springer-Verlag*, 1737, pp. 211-234.

Levy, A.Y., Rajaraman, A. and Ordille, J. Querying Heterogeneous Information Sources Using Source Description. In *Proc. of the Int. Conference on Very Large Databases*, pp.25-262. India.

Levy, A.Y., 2000. Logic-Based Techniques in Data Integration. In *Jack Minker, editor, Logic Based Artificial Intelligence, Kluwer*, p.p 575-595.

Maps, 2006 Maps.google.com. Acceded in 2006

Open Geospatial Consortium, 2003 Open Geospatial Consortium, 2004. Geography Markup Language (GML) v3.1.1.

Papakonstantinou, Y., García-Molina, H. and Widom, J., 1992. Object Exchange Across Heterogeneous Information Sources. In *Proc. ICDE Conf. TSIMMIS project.*