2012 Fourth International Conference on Computational and Information Sciences

A Study of Spatial Query Optimization Based on Semantics in Data Integration

Fengyuan Zhong, Tingshan Zhang and Chengwu Wang
State Key Laboratory of Oil and Gas Reservoir Geology and Exploitation
Southwest Petroleum University
Chengdu, China
e-mail: fengyuanzhong@hotmail.com

***Abstract*—Data integration of geospatial data in distributed and heterogeneous environment involves the use of semantic ontologies. In this kind of integration system, semantic technologies play an important role in improving performance and effectiveness of spatial queries. This paper focuses on methods of query optimization based on spatial semantics at the top level of semantic layer in central data integration systems. After analyzing the hybrid approach for spatial data integration, two categories of query optimization strategies are proposed based on detailed examination of special characteristics of spatial data. With spatial knowledge explicitly specified in ontologies and associated rules, spatial queries can be optimized intelligently.**

Keywords- spatial semantics; spatial query; query optimization; data integration; spatial characteristics

I. INTRODUCTION

As traditional strategies for data integration merely integrate heterogeneous information at structural level and syntactic level, they do not address the issue of semantic heterogeneity in data sources. The new range of semantic technologies based on ontologies has been introduced into the systems for data integration. Formal ontology approaches are being developed for a number of reasons including the sharing of common understanding of information structure, enabling reuse of domain knowledge, making domain assumptions explicit and mediating between information resources (semantic translation) [1]. Five uses of ontologies in data integration have been identified: metadata representation, global conceptualization, declarative mediation, mapping support, and support for high-level queries [2].

Geospatial data sharing in distributed, heterogeneous environment is a typical case of data integration and the challenging problem of improving performance and effectiveness of spatial queries is yet to be solved. In the integrated environment, optimization of a spatial query at its top level is preferred for top-level decisions impact the consequent queries greatly. In this research, several methods are proposed to optimize spatial queries at the top level of the semantic layer in central data integration systems. Having given full consideration to spatial characteristics, these methods aim to reduce computational complexity and/or data traffic with the use of semantic technologies.

The rest of this paper is organized as follows. Section II describes the semantic-based data integration environment for spatial data, the underlying system architecture that is used throughout the paper. The optimization strategies based on spatial semantics are presented in Section III. Finally, Section IV summarizes our research.

II. SEMANTIC-BASED DATA INTEGRATION ENVIRONMENT

A. Ontologies for Data Integration

Semantic ontologies are the structural framework for organizing information. In theory, an ontology is a "formal, explicit specification of a shared conceptualization"[3]. An ontology renders shared vocabulary and taxonomy and represents knowledge as a set of standardized concepts within a domain and the relationships between these concepts. That is, it models a domain with the definition of objects and/or concepts, and their properties and relations. Common components of an ontology model include:

1)Individuals: instances or objects.

2)Classes: sets of concepts or kinds of things.

3)Attributes: aspects, properties that objects (and classes) can have.

4)Relations: ways in which classes and individuals can be related to one another.

Ontologies are commonly encoded using ontology languages. The ontology language, OWL(Web Ontology Language), became a W3C (World Wide Web Consortium) Recommendation in February 2004.

Ontologies for data integration have been used in one of the three ways: single ontology approach, multiple ontology approach and hybrid ontology approach. The hybrid ontology approach is a combination of the two preceding approaches. First, a local ontology is built for each source schema, which, however, is not mapped to other local ontologies, but to a global shared ontology[2]. Obviously, these ontologies are built at different semantic levels. Thus, one of its advantages is that it is suitable for building a multi-layer system, which is extensively used in distributed data integration. Furthermore, new sources can be easily added with no need for modifying existing mappings and shared semantic vocabulary. Another advantage is its compatibility of heterogeneous data sources. Hence, the hybrid approach is appropriate for building central data integration systems, especially large-scale systems with great complexity.

Geospatial data integration is, more often than not, in need of the mechanism supporting a large-scale distributed

978-0-7695-4789-3/12 $26.00 © 2012 IEEE
DOI 10.1109/ICCIS.2012.47

system. For example, many governments have developed or been developing "critical infrastructure" databases. In these projects, integration of large amount of geospatial data produced by variety of data sources is doubtless a tough challenge. Therefore, the hybrid approach is a reasonable choice for geospatial data integration.

B. Proposed Framework for Data Integration

Fig. 1 shows the main components of the proposed framework using the hybrid approach. At the top level of the semantic layer, the global ontology defines the shared vocabulary of all information sources, usually the basic terms of the domain. It consists of the union of concepts, but not the terminology, from the local ontologies, and a set of axioms that define inter-ontology properties. Each data source has its corresponding local ontology at the intermediate level, which describes the source database schema, like relational, XML (eXtensible Markup Language), or RDF(Resource Description Framework). A local ontology is a conceptualization of the elements and relationships between elements in each source schema. For the sake of correct query processing, the structure of source schemas and the integrity constraints (e.g., relational foreign keys) expressed on the schemas should be preserved in the local ontology[2].The local ontology is built independently, that is, without taking into account the other data sources. At the bottom level, XML schema is used to "wrap" data in XML format for output. Between the top level and the intermediate, mapping is built to connect the concepts defined in the global ontology and in the local ontologies. The mapping rules are used to solve the semantic heterogeneity problems [2-6].

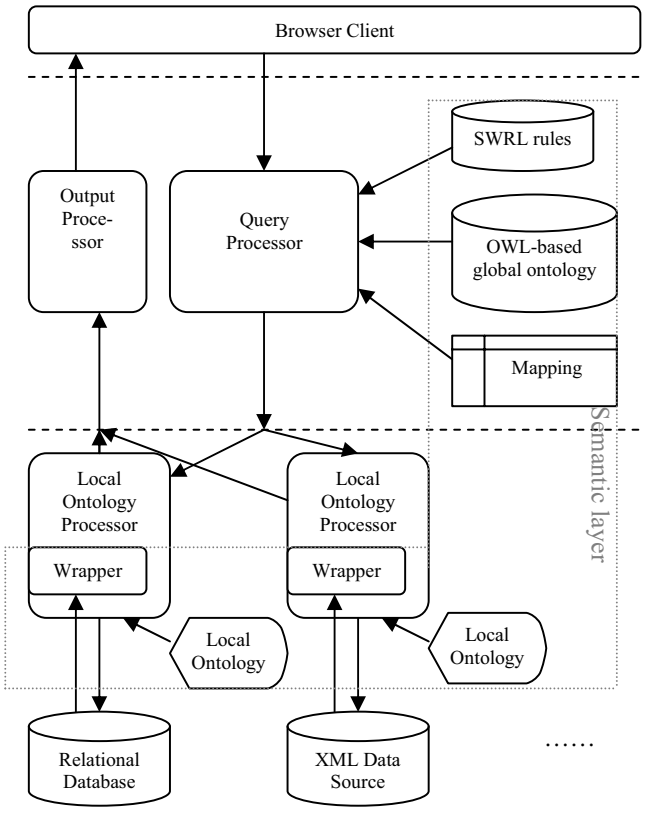

Figure 1. Framework for spatial data integration.

III. SPATIAL QUERY OPTIMIZATION

A. Special Characteristics of Spatial Data

Spatial data is a term used to describe data that pertains to the space occupied by objects in a database. Special characteristics of spatial data include:

1) Spatial dependency. The 1st law of geography states, "Everything is related to everything else but near things are more related than distant things".

2) Heterogeneity. Heterogeneity results from the unique nature of each place, indicating that spatial data very rarely presents stationary characteristics.

3)Fuzziness. Geographic space is continuous and infinite (or at least compact). The discrepancies between the real world and the representations of the real world used as the basis of analysis can also affect the quality of the analysis [7].

4)Spatial scale. Geographic objects and phenomena have different structures at different spatial scales.

While typical databases can understand various numeric and character types of data, additional functionality needs to be added for databases to process spatial data types. The additional efforts are focused to solve the following two problems.

1)Unstructured data. Since the data is geometric and varied, including points, lines, polygons etc., coordinates can not be treated as additional attributes in a tuple.

2)Spatial relationships. The data is multi-dimensional and spatial relationships such as spatial distribution and topology must be specially treated.

Spatial data is usually used in conjunction with what is known as attribute or non-spatial data. As long as non-spatial data is involved in spatial queries, the mixed queries contain both spatial predicates and non-spatial predicates. In some spatial database systems (eg.GEOQL) , the non-spatial subqueries are processed by the SQL backend and the spatial subqueries by the spatial processor. Some (eg.Paradise) use basic relational operators and optimization techniques for both spatial and non-spatial operations. The methods like ordering and merging of spatial and non-spatial operations are adopted by SAND and GRAL [8]. Semantic information stored in databases as integrity constraints are proposed for query optimization [9-11].

B. Query Optimization Strategies

Most of these research findings have focused on optimization through the query optimizer in DBMS (Database Management System). However, this paper presents query optimization strategies which exploit both the spatial characteristics and the ontological model at the top level of the semantic layer. Our strategies are divided into two categories, according to the aim of processing: definition specification and extent division. In the integrated data environment, the semantic layer actually forms a virtual database with its own schema. The definition of the schema is the starting point of query optimization. In the first

```
<owl:Class rdf:ID="Clity">
  <rdfs:subClassOf rdf:resource="#Settlement"/>
  <rdfs:comment xml:lang="en"
    >a relatively large and permanent settlement</rdfs:comment>
</owl:Class>
<owl:DatatypeProperty rdf:ID="hasDimension">
  <rdfs:domain rdf:resource="#Clity"/>
  <rdfs:range>
    <owl:DataRange>
      <owl:oneOf>
        <rdf:List>
          <rdf:first rdf:datatype="&xsd;int">0</rdf:first>
          <rdf:rest>
            <rdf:List>
              <rdf:first rdf:datatype="&xsd;int">2</rdf:first>
              <rdf:rest rdf:resource="&rdf;nil"/>
            </rdf:List>
          </rdf:rest>
        </rdf:List>
      </owl:oneOf>
    </owl:DataRange>
  </rdfs:range>
</owl:DatatypeProperty>
```

Figure 2. Dimension definition of class *City* written in OWL.

strategy, not only mappings to the original sources are constructed in the virtual schema, rules or specifications that make spatial queries more efficient are also incorporated. Here, object dimension and topological rules are specified for their fundamental importance in a spatial query. After queries are submitted, the query processor uses another strategy, that is, extent division, to narrow down the spatial sphere of candidate objects. Spatial extent is typically fuzzy at the semantic level. Experience rules of spatial scale are used to provide heuristic guidance.

1) Definition Specification

As entities in the real world are represented in a database through the process of abstraction, what spatial dimensions they have are generally decided by the domain of interest and the plotting scale. Thus, a same concept or a same ontological class can have different dimensions. For example, a city is often represented as a point in a national map, but in the city map it is naturally a polygon. It means that a same ontological class using different dimensions has different applicable spatial functions. So it is absolutely necessary to specify dimension information in the ontology for query optimization. For instance, in the global ontology the dimension definition of the class, City, is specified in Fig.2.

SELECT City WHERE WithinDistance（City， LpDL，“Distance=100km”）AND LpDL =“Lp-x”

Figure 3. Query of the cities.

SELECT City WHERE Contain (City, LpDL) AND LpDL =“Lp-x”

Figure 4. Query of the city.

It indicates that the class, City, has the dimension of 0 or 2, which means points or polygons. So we actually have both point data and polygon data for the class at database layer.

Consider the following query in an SQL-like syntax in Fig.3, which retrieves all the cities within 100km from the gas pipeline with the code “Lp-x”. Here, Low-pressure distribution pipelines are defined in the global ontology as the class “LpDL” with the dimension of 1, that is, line.

From the global ontology, the processor gets the dimensions available and chooses the dimension that saves time. Since a polygon has a number of points, the space complexity and time frequency of the distance algorithm of city points is much less than that of city polygons. The query processor chooses the faster plan with the applicable dimension. In this case, the faster choice is the dimension of 0.

But in another example query, given in Fig.4, which is to retrieve the city where the gas pipeline with the code ” Lp-x” is within, our choice is the dimension of 2 because the spatial operation “Contain” semantically implies the dimension of City is higher than that of LpDL. The global ontology provides the available dimensions and lays the basis for optimization process.

The reasoning process is based on explicit knowledge representation. OWL and its associated SWRL (Semantic

Rule1: LpDL(?p) ∧ hasStartpoint(?p, ?n)→hasRegulator (?p,?n)

Rule2: LpDL(?p) → IsIncity (?p,true)

Figure 5. Topological relations defined as SWRL rules.

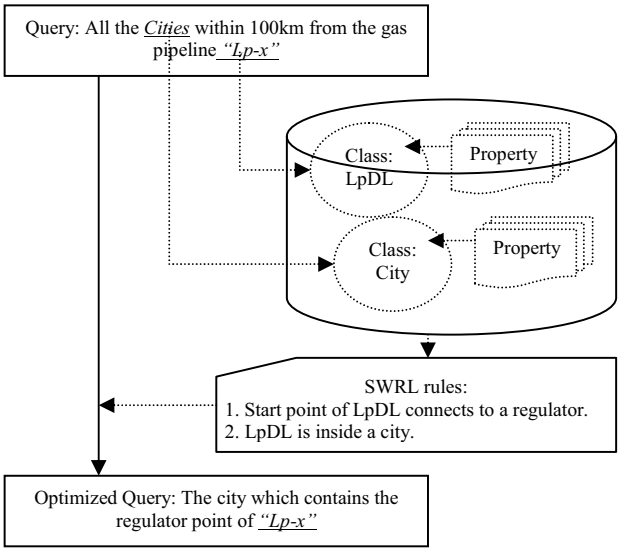

Figure 6. Optimized query with spatial semantics.

Web Rule Language) provide a powerful standardized approach for representing information and reasoning with it. SWRL allows users to write rules that can be expressed in terms of OWL concepts to provide more powerful deductive reasoning capabilities than OWL alone. As topological relations between spatial objects are foundation for qualitative spatial reasoning and spatial query, they contribute to query optimization. This is illustrated in the example query about City and LpDL as follows.

Before processing queries，topological relations should be represented as rules. Gas pipelines are generally divided into several categories, one of which is the low-pressure distribution pipeline. Low-pressure distribution pipelines start from regulator stations and then connect to low-pressure branch pipelines. Thus, a low-pressure distribution pipeline has a topological relation with its corresponding regulator station. Its topological relation is written as the following SWRL rules in Fig.5. It indicates that a low-pressure distribution pipeline with a start point has its regulator station at that point and it must be in a city.

Now for the query given in Fig.2, the gas pipeline with the code "Lp-x" belongs to the class "LpDL". The query processor checks the rules concerned and gets the information that a regulator station, which is also defined as a class, is at its start point. Then instead of calculating which city polygon the pipeline line is within，the polygon city with the regulator point in it is retrieved. So the complexity of the spatial operation is decreased by transforming line-in-polygon operations to point-in-polygon operations.

2) Extent Division

Different spatial scales are generally appropriate for different spatial tasks. Through building up experience rules of spatial scales as SWRL rules，the query processor draws on spatial knowledge to narrow down the search extent. In the gas pipeline example, city gas networks are constructed within a city and long-distance pipelines for natural gas transmission are not. At a national scale, long-distance transmission pipelines between cities are often studied, while city gas networks are more of interest at a city scale. For the query in Fig.3, if the pipeline is replaced by a long-distance line, which is away from cities, distance calculation should be carried out as said above. But it is a low-pressure pipeline, so after reasoning according to Rule 2 in Fig.5, which means it is inside a city, the query is actually rewritten to the query in Fig.4. This query is further optimized with the methods mentioned above and finally we have the optimized query showed in Fig.6.

IV. CONCLUSION

Semantic-based data integration environment is the development trend of spatial data integration. In this paper, the appropriate integration approach is examined in details and levels in the semantic layer are distinguished after analysis. In the interest of space, the focus of the paper is primarily on spatial query optimization at the top level of the semantic layer in central data integration systems. The emphasis is to take advantage of spatial characteristics. Based on the idea that spatial semantics helps query optimization at the semantic level, the methods of "definition specification" and "extent division" are proposed. Future research includes building a cost model for analyzing the suggested spatial query processing and optimization strategies as well as building a top-layer spatial query optimizer to experiment with such strategies.

REFERENCES

[1] Peter L. Pulsifer, "An Ontological Exploration of Antarctic Environmental Governance:Towards a Model for Geographic Information Mediation," Ph.D., Carleton University , 2008, 381 pages, NR43905. DAI-A 69/11, p. , May 2009.

[2] Isabel F. Cruz, Huiyong Xiao, "The Role of Ontologies in Data Integration," Journal Of Engineering Intelligent Systems, vol.13/2005,Apr.2005, pp. 245-252.

[3] Gruber T R, " A Translation Approach to Portable Ontology Specifications" , Knowledge System Laboratory KSL 92-71. 1993.

[4] Tian Zhao, Chuanrong Zhang, Mingzhen Wei and Zhong-Ren Peng, "Ontology-based Geospatial Data Query and Integration," Geographic Information Science, vol.5266,2008,pp. 370-392, doi: 10.1007/978-3-540-87473-7_24.

[5] Vânia M.P. Vidal, Eveline R. Sacramento, José Antonio Fernandes de Macêdo and Marco Antonio Casanova, "An Ontology-based Framework for Geographic Data Integration," Proc. ER 2009 Workshops CoMoL, Advances in Conceptual Modeling - Challenging Perspectives,vol. 5833/2009, 2009,pp. 337-346, doi: 10.1007/978-3-642-04947-7_40.

[6] S. Zlatanova, M.de Vries and P.J.M. van Oosterom, "Ontology-based Query of Two Dutch Topographic Datasets:an Emergency Response Case", Proc. Core Spatial Databases–Updating, maintenance and services from theory to practice. Haifa, Israel, ISPRS, vol.XXXVIII , Part 4-8-2-W9,pp. 193–198, Dec. 2010.

[7] Fernando Bacao, Victor Lobo and Marco Painho, "On the Particular Characteristics of Spatial Data and Its Similarities to Secondary Data Used in Data Mining", http://www.isegi.unl.pt, 2005.

[8] Ho-hyun Park, Yong-ju Lee and Chin-wan Chung, "Spatial Query Optimization Utilizing Early Separated Filter and Refinement Strategy", Information Systems, vol. 25, Jan. 2000, pp. 1-22, doi：10.1016/S0306-4379(00)00006-5.

[9] Matthias Jarke and Jurgen Koch, "Query Optimization in Database Systems," ACM Computing Surveys, vol. 16, Feb. 1984, pp. 111-143, doi: 10.1145/356924.356928.

[10] Hichul An and Lawrence J. Henschen, "Knowledge based semantic query optimization", Methodologies for Intelligent Systems, Volume 542/1991,1991,pp. 82-91, doi: 10.1007/3-540-54563-8_72.

[11] Walid G. href and Hanan Samet, "Optimization Strategies for Spatial Query Processing", Proc. of 17th International Conference on VLDB (Very Large Data Bases)，Barcelona, Spain, Sept. 1991, pp. 81-90, Key: citeulike:2715635.