Information Mediation Across Heterogeneous Government Spatial Data Sources

Amarnath Gupta Ashraf Memon Joshua Tran Rajiv P. Bharadwaja Ilya Zaslavsky

San Diego Supercomputer Center
University of California San Diego
{gupta, amemon, trantj, zaslasvsk}@sdsc.edu rajiv@cs.ucsd.edu

Project URL: http://www.sdsc.edu/DICE/

Abstract. Spatial information is a common component of databases maintained by governments at different hierarchical levels, from local to federal. Even within a single municipality, information about particular spatial objects may differ in measurement and representation type, content of attributes, data quality, amount and structure of metadata, currency, etc. Such heterogeneity becomes a serious obstacle in applications that require querying across multiple spatial databases. To address this problem, we extend the principles of conventional information mediation to processing of spatial queries against multiple heterogeneous spatial data sources and services. The spatial mediation infrastructure we are developing, includes a set of spatial wrappers which expose capabilities of spatial data sources and services in a homogeneous way, spatial mediators for query planning and orchestrating query execution, and spatial presentation services. This paper focuses on the specifics of spatial information mediation, presenting a model of source spatial capabilities for mediation. Using a running example, we demonstrate generation of feasible query evaluation plans based on source spatial coverage declaration and function signatures for all supported spatial data interchange and spatial data transformation functions.

1. Introduction

Databases maintained by government agencies at different hierarchical levels, commonly contain geographic identifiers or can be linked to geographic objects. Ability to integrate these databases "on demand" and pose complex queries against database collections is an important component of modern e-government operations. However, representational and semantic heterogeneity among spatial databases makes on-demand integration a challenging task. Even within one municipality, different administrative departments commonly maintain a variety of databases differently representing the same set of spatial objects. Depending on the resolution of the data, and the data collection mandate of a custodian agency, spatial objects can be differently measured and depicted as points, lines or polygons. For example, a street object may be represented differently in databases maintained by engineering, environmental planning, and transportation departments. Similarly, a network of streams will be represented with different spatial types by agencies responsible for hydrologic engineering and water quality, and by land use planners. The *representational heterogeneity* of semantically equivalent objects in different databases creates interesting problems for *information mediation* across multiple spatial databases, as the different representations will trigger different query evaluation scenarios.

Information mediation is a query evaluation strategy where user queries against multiple heterogeneous data sources are communicated via a middleware (*mediator*) which is responsible for query rewriting, dispatching query fragments to individual sources, and assembling individual query results into a composite query response (Levy et al. 1996, Papakonstantinou et al. 1996, Tomasic et al 1998, Wiederhold 1992). Within this framework, source heterogeneity is managed by *source wrappers* which export a *common data model* view of the data at each source, along with information about the source *schema* and metadata, and a description of the *supported queries*. In MIX (*Mediation of Information using XML*), the mediator architecture being developed at SDSC (Baru et al. 1999), the common data model is XML; and the integrated views over heterogeneous sources are presented as virtual XML sites.

In this paper, we extend our notion of information mediation in two directions. Firstly, we propose spatial capability descriptions for sources within a wrapper-mediator system, to enable spatial query planning at the mediator. Secondly, we consider spatial web services (computational services implementing spatial operations) as part of the mediation system. The goal of this paper is to explore specific features of spatial mediation and demonstrate the generation of spatial query evaluation plans based on source capabilities, as it is implemented in our prototype spatial services developed in the course of NSF-funded "I2T: Information Integration Testbed for Digital Government" project at San Diego Supercomputer Center.

2. What is special about spatial information mediation

Spatial information mediation, broadly speaking, is application of wrapper-mediator approaches to spatial data sources and services. This includes the ability to handle spatial data types (*points, lines, areas, grids*, etc.) and use spatial operations and predicates (*inside, disjoint, intersects*, etc.) for querying or generating spatial objects. In a more narrow sense, we can consider spatial mediation as supporting *transformations across spatial types*, and relying on *spatial integration models* to relate distributed heterogeneous data sources.

Spatial information mediation follows the mediation architecture for "traditional" sources – i.e. it includes: (1) spatial wrappers for spatial information sources that resolve source heterogeneity by expressing spatial data sets in a common model, using various spatial data exchange and metadata standards, (2) spatial mediators that support generation and orchestration of spatial query plans, and (3) specialized clients for navigating and querying distributed spatial sources, and presenting mediation results as maps. However, specifics of spatial data and spatial integration models have important ramifications for mediation architecture. Some of the main issues are:

- Creating a "fat" mediator (i.e. a full spatial information processor with a superset of source functionality) is challenging and possibly inefficient, due to diverse functionality of spatial information processing software, variety of spatial data types, and the lack of commonly accepted spatial query mechanisms. Additionally, shipping all data to the mediator for processing may be inefficient due to typically large volumes of data in geographic databases. Instead, a "lean" mediator, which lacks spatial processing capabilities and delegates spatial data handling to one or more sources, may be more appropriate for complex spatial queries. How much spatial processing the mediator can support will strongly impact the way query plans are generated.

- Spatial query plans often require generating and persisting intermediate results as new spatial datasets. For example, a query "return area of land parcels that fall within wetlands" would require polygon overlay of parcels and wetland layers, and storing overlay results as a new polygon layer, so that a *compute_area* operation may be applied to it. Databases accessible over the Web are seldom capable of storing intermediate results created by user queries. Since materializing intermediate results is often necessary for implementing complex spatial query plans, the storage capabilities become a critical component of capability description of spatial sources, mediators and clients.

- Sources must expose their schemas, which, in case of spatial sources, additionally include *spatial coverage declaration*, and function signatures for *spatial data interchange* and *transformation* functions. Spatial query plans are first formulated by the spatial mediator in terms of *mediated spatial objects*, which are then instantiated using the spatial capabilities of the sources. The complete sequence of spatial query planning steps is described below.

- Measurement considerations and data quality are inherent components of geographic information suggesting accuracy-based source selection in the process of query evaluation. Instantiating abstract query evaluation plans with data sources, spatial mediator takes into account source data quality, accuracy specification of the query (if any), and error propagation functions characteristic of data transformations outlined in a query plan. A

particular query may request the most accurate constellation of data sources, or deliver query results within user-specified accuracy bounds (Zaslavsky et al. 2000)

- The result of a spatial query is often intended to be viewed graphically, as a map. Therefore, a spatial mediator should be able to retrieve client's spatial capabilities (map rendering capabilities, in particular) and prepare query results to the client's specification. Often, query results make no sense without a *context*, or collection of spatial objects and map elements that help users to situate and better understand the result, encourage further queries and analysis, etc. The purpose of result presentation plans generated by a spatial mediator is to situate query results in semantic, spatial and temporal contexts. While specific design issues of XML-based mapping clients (Zaslavsky 2000) are beyond the scope of this paper, the approach adopted in this paper allows to consider mediation clients within a common spatial capability description framework.

3. Spatial Capability Descriptions within a Wrapper-Mediator System

Analysis of the extent of "digital divide", the gap in computer ownership, literacy, and Internet access in different sections of the society, provides a motivating example. Suppose that we are interested in answering a query *"Find schools and their school districts within 3 miles of a given point, such that the city they are in has a high level of computer literacy."* Further assume that the query is executed over three spatial sources: a GIS system (source S1), a spatially enabled relational database that allows SQL queries extended by geometric operations (source S2), and an internet mapping source that can be accessed through an XML-based query language (source S3). Spatial capabilities of the sources include the following:

- Each spatial source represents a collection of objects with a single geometry attribute (**geom**).
- Each collection of spatial objects has a *Spatial Coverage Declaration***,** specified as:
  covers (source_name.exported_object_name, spatial_data_type, namespace, coverage_name, coverage_extent).
- *Spatial Data Interchange Functions* describe how a spatial source can exchange information with external software such as the mediator: by returning an image with or without non-spatial relation; a handle to spatial data residing at the source; a specific geometric data structure; or a composite source-specific data structure interleaving spatial and non-spatial information (e.g. a shapefile). A general form of function signature is:
  source_name.exported_operation({arguments}) returns (result.spatial as domain:type, result.nonspatial as relation)
  where "domain" is {image|handle|geometry|composite}
- *Spatial Data Transformation Functions* reference transformations between data types supported by a particular source, in the following general form:
  transform (source_name.exported_function_name, namespace_declaration, domain0: type0, description, parameters0, parameters1) returns result.spatial as domain1:type1

Following this outline, partial source schemas can be specified as follows:

Source S1 (city polygons for "northern" San Diego county):

Covers (S1.cities, 'polygon', 'namespace:administrative'.'county', 'San Diego' ('northern') [coordinate list])
S1.cities.geom(type:polygon;projection:geographic; units:DDMMSS; date:1/1/2000)
S1. cities.name (type:string; namespace:cities; lang:en; date: 1/1/2000)
S1. cities.comp_literacy (type:float; namespace:digital_divide; date: 1/1/2000)
Covers (S1.streets, 'polygon', 'namespace:administrative'.'county ,'San Diego' ('northern') [coordinate list])
S1.spatial_union(type1:geom;type2:geom)
S1.transform(geometry:multiline,'polyline to polygon', "{format:shapefile}","{format:

shapefile}") returns (result.spatial as geometry:polygon)

Source S2 (community boundaries as polylines and school points for "southern" and "eastern" San Diego county):

Covers (S2.communities, 'polygon', 'namespace:administrative'.'county', 'San Diego' ('southern', 'eastern') [coordinate list])
S2.communities.geom (type:polyline; projection:geographic; units:DDMMSS; date:1/1/2000)
S2. communities.name (type:string; namespace:communities; lang:en; date: 1/1/2000)
S2. communities.comp_literacy (type:string; namespace:digital_divide; date: 1/1/2000)
Covers (S2.schools, 'polygon', 'namespace:administrative'. 'county, 'San Diego' ('southern', 'eastern') [coordinate list])
S2.schools.geom (type:point; namespace:education; projection:geographic; units:DDMMSS;date:1/1/2000)
S2.schools.name(type:string; namespace: education; date:1/1/2000)
S2.spatial_union(type1:geom;type2:geom)
S2.ship(result, destination)

Source S3 (school districts as polygons and school points for all of San Diego county):

Covers (S3.school_districts, 'polygon', 'namespace:administrative'.'county', 'San Diego' [coordinate list])
S3.school_districts.geom (type:polygon;projection:geographic, units:DDMMSS;date:1/1/2000)
S3. school_districts.name (type:string;namespace:communities;lang:en;date: 1/1/2000)
Covers (S3.schools, 'polygon', 'namespace:administrative'. 'county, 'San Diego' [coordinate list])
S3.schools.geom (type:point; namespace:education;projection:geographic; units:DDMMSS;date:1/1/2000)
S3.schools.name(type:string; namespace: education; lang:en date:1/1/2000)
S3.schools.school_district(type:string; namespace: education; lang:en date:1/1/2000)

Note that *comp_literacy* is available by cities and communities but not by school districts (in fact, this attribute is extracted from a San Diego county telephone survey where respondents were asked "Where do you live?" and "Do you have a computer in your household?" (Zaslavsky et al, 2001).

Initially, the query is formulated in SQL-like syntax against "mediated" objects:

```
select      MS.name, MS.geom, MD.name, MD.geom as map
with coverage = 'San Diego':county
from        med_cities MC, med_schools MS,
            med_districts MD
where       inside(MS.geom, MC.geom) and
            inside (MS.geom, MD.geom) and
            MC.comp_literacy = 'high' and
            distance(MS.geom, user_point.geom) < 3.
```

Here *med_cities, med_schools,* and *med_districts* are the mediated objects that the user can query on. Notice that the user expects the result of the query as a map, implying that there is a client that can accept and display suitably encoded geographic information. The client's and the mediator's own capability of accepting spatial data and performing spatial computations follow the same specification.

4. Query Plan Generation using Spatial Capabilities

The mediator formulates spatial query plans in the following steps:

4.1 Translating the SQL-like query to one or more tentative query plans.

Assuming that at view definition time the mediator has coverage relation:

coverage('View1':view,'namespace:administrative'.'county', 'San Diego')=
coverage(_,_,['San Diego'('northern'), ['San Diego'('southern'), ['San Diego'('eastern'),]),

we develop a tentative query plan *P0* for our example as follows:

$V1=$\pi_{geom}$ ($\sigma_{comp\_lit='high'}$(MC))
$V2= θ-join $_{inside(\$1.geom, \$2.geom)}$($\pi_{geom,name}$(MS), $\pi_{geom,name}$(MD))
$V3= θ-semijoin $_{inside(\$1.geom, \$2.1)}$($V1, $V2)[1]
$result= $\sigma_{dist(\$V3.geom.,address)<3}$($V3)

4.2 Unfolding the definitions of the mediated layers, and the coverage constraint in the user query

Considering **MC** a union type over **S1.CITIES** and **S2.COMMUNITIES,** we can rewrite the first statement of plan *P0* as:

$V11=$\pi_{geom}$ ($\sigma_{comp\_lit='high'}$(S1.CITIES))
if covers(S1.CITIES,_,_,_,G), covers(query, 'med_ns',_,Name,G1)), overlaps(G,G1).
$V12=$\pi_{geom}$ ($\sigma_{comp\_lit='high'}$(S2.COMMUNITIES))
if covers(S2.COMMUNITIES,_,_,_,G), covers(query, 'med_ns',_,Name,G1)), overlaps(G,G1).
$V1=$V11 spatial_union $V12

where *med_ns* is the mediator's namespace queried to retrieve the spatial extent of user's query. Using the coverage declaration from step 1, a simpler plan for the first statement of *P0* is:

$V11=$\pi_{geom}$ ($\sigma_{comp\_lit='high'}$(S1.CITIES))
$V12=$\pi_{geom}$ ($\sigma_{comp\_lit='high'}$(S2.COMMUNITIES))
$V1=$V11 **spatial_union** $V12

4.3 Assigning operations to sources and rewriting query fragments using source spatial capabilities

This step quickly proliferates the number of tentative plans, as sources may support large number of interchange and transformation functions with a variety of inputs. For example, if source **S1** can export both images and composite polygon or multiline objects in the form of "shapefiles", the plan generator will create alternate plan segments like:

$V11[1]=$\pi_{image}$ ($\sigma_{comp\_lit='high'}$(S1.CITIES))
$V11[2]=$\pi_{shapefile(polygon)}$ ($\sigma_{comp\_lit='high'}$(S1.CITIES))
$V11[3]=$\pi_{shapefile(multiline)}$ ($\sigma_{comp\_lit='high'}$(S1.CITIES))

4.4 "Gap analysis", and injecting, where possible, compatibility operators between disconnected parts of a query using a set of equivalence rules

In our example, the join-operation in

$V3′= θ-semijoin S1.inside($1.geometry(point), $2.1)($V1, $V2)

[1] $2.1 takes the first attribute of the second relation, namely, the geometry of the school

is not feasible because the *S1.inside()* operation has the signature: $point \times polygon \rightarrow Bool$ while $V2 contains the geometry in the form of multiline. To make the plan feasible, the system has to "complete" it by filling it in as:

$$\$V3\ [1] = \theta\text{-semijoin}_{S1.inside(\$1.geometry(point),\ S1.multiline2polygon(\$2.1))}(\$V1, \$V2)$$

provided some source contains the required transformation function.

4.5 Selecting an optimal plan from the feasible plans

At this step, the mediator can evaluate a query cost model or apply a set of heuristic rules (currently) to select the best available plan. Common heuristics would "commit all processing to minimal number of sources", "commit processing to source with the largest volume of data", etc.

5. Conclusion

Spatial mediation offers an intriguing strategy for on-demand integration of heterogeneous spatial information ubiquitous in government databases. This paper outlined a specification for spatial capabilities of sources and services in a wrapper-mediator system, and described a multi-step query planning based on the capability descriptions. Testing this model on a variety government datasets, development of efficient spatial query plans and evaluation of query cost models represent the immediate future directions of our research.

References

Baru C, Gupta A, Ludäscher B, Marciano R, Papakonstantinou Y, Velikhov P and Chu V. (1999) "*XML-Based Information Mediation with MIX*". Proc. of the ACM SIGMOD'99, June 1-3, Philadelphia, PA, USA, pp. 597-599.

Gupta A, Marciano R, Zaslavsky I and Baru C (1999) "*Integrating GIS and Imagery through XML-based Information Mediation*". In Agouris P and Stefanidis A (eds) Integrated Spatial Databases: Digital Images and GIS, Lecture Notes in Computer Science, Vol. 1737, pp. 211-234, Berlin, Springer-Verlag.

Gupta, I. Zaslavsky, R. Marciano (2000). "*Generating Query Evaluation Plans within a Spatial Mediation Framework*". Proceedings of the 9th International Symposium on Spatial Data Handling, August 2000, pp. 8a18 – 8a31.

Levy, A., A. Rajaraman and J. Ordille (1996) "*Querying Heterogeneous Information Sources Using Sources Descriptions*". Proceedings of VLDB, pp. 251-262.

Papakonstantinou, Y, S. Abiteboul, and H. Garcia-Molina (1996) "*Object Fusion in Mediator Systems.*" In Intl. Conf. on Very Large Data Bases (VLDB).

Tomasic, A., L. Raschid, P. Valduriez (1998). "*Scaling Access to Heterogeneous Data Sources with DISCO*". IEEE Transactions on Knowledge and Data Engineering, Vol. 10, No. 5, pp. 808-823.

Wiederhold G (1992) Mediators in the Architecture of Future Information Systems. IEEE Computer, 25, 3, 38-49.

Zaslavsky, I., B. Ludäscher, A. Gupta and R. Marciano (2000). "*Accuracy Mediation in a Spatial Wrapper-Mediator System*". 1st Geographic Information Science Conference (Savannah, GA, October 2000). Extended abstract.

Zaslavsky, I. 2000. "*A New Technology for Interactive Online Mapping with Vector Markup and XML*". Cartographic Perspectives, # 37 (Fall 2000), pp. 65-77.

Zaslavsky, I., Sensenig, P., Forkenbrock, D. (2001) "*Mapping a Future for Digital Connections: A Study of the Digital Divide in San Diego*". San Diego Regional Technology Alliance. (available at http://www.sdrta.org/sdrta/aboutsdrta/RTA_Report_0201.pdf)