

Chapter Two - problems of combining spatial data

27th November 2004

1 Introduction

Through the 1970s and early 1980s most GIS applications were considered as islands of information. They were self contained independent systems where spatial data were digitally captured, stored, analysed, and displayed (Bishr, Y. 1998). The need for re-use or share the existing geographical data due to the expensive, time consuming and the difficulty or sometimes impossible of recollection of the same specific data (e.g. due to the specific time domain or weather condition etc.) emerge the notion of sharing data within different organizations, without effecting the autonomy of these organizations. Many researchers provide definition for this notion as system interoperability (Goodchild, M. et al 1997; Sonhim, K. 1998; Bishr, Y. 1998). They define the 'interoperability' as the ability to move data or information from one system to another.

The term interoperability comes from the Open GIS Consortium (OGC) under the umbrella of 'Open GIS' (Albrecht, J. 1999) whose development relies on concepts such as interoperability and portability, however the term interoperability and Open GIS are often used interchangeably (Voisard, A. and Schwpper, H. 1998).

Interoperability in GIS means that software components work with each other to overcome tedious batch conversion tasks, import/export obstacles, and distributed resource access barriers imposed by heterogeneous processing environments and heterogeneous data ([OGC 1996]). The Geographic information infrastructure (GII) introduces the institutional and technical guidelines for combining and integrating data from distributed data sources. The special case of geographic information is of particular interest in sharing and infrastructure building, due to the high cost of producing it, its potential for widespread re-use (Evans and Ferreira, 1995).

The problems in combining these spatial data arises from the heterogeneity in terms of data format, the hardware used for storing data, and the software used to acquire the data and prepare it for further use. Since 1990, the U.S. Federal Geographic Data Committee has devised and promoted a National Spatial Data Infrastructure (NSDI) defined (Tosta, 1994) as:

- Standards to facilitate data collection, documentation, access, and transfer;
- A basic framework of digital geospatial data that meets the minimum needs of large numbers of data users over any given geographic area;
- A clearinghouse to serve, search, query, find, access, and use geospatial data;
- Education and training in the collection, management, and use of geospatial data.

The following sections will focus on illustrating the elements of GII and discussing the technical challenges in combining heterogeneous spatial data and discuss solutions under the umbrella of Geographic Information Infrastructure.

2 Geographic Interoperability

Interoperability of information systems has been a topic of discussion for more than two decades within the IT industry generally and that some aspects have been implemented outside the GIS arena for many years. Interoperability is not new to GIS. One of the earliest examples of the concepts of interoperation for GIS is the conversion between different map projections. In order to perform, say, an overlay of two data sets, the coordinates in the two data sets are expected to be in the same coordinate reference system; otherwise, the numerical processes of calculating line intersections will yield surprising results. Intelligent data sets would know about their map projections, and intelligent operations would know how to make themselves compatible.

The earliest attempts at making GISs work together with other software modules go back to a 1988 paper, when (Johnston et al. 1988) described their efforts to perform an allocation problem by integrating GIS software with other software pieces, demonstrating the difficulties system developers and users had with sharing geographic data across different computational processes. The integration, called Orpheus, was not a GIS software package, but a methodology of how to use a suite of product from different vendors to accomplish a complex task in an integrated fashion. It included, of course, a GIS package, and other software for image processing, CAD, surface modeling (then not integrated with GIS); and architectural, engineering, and construction software. All products were installed on the same machine running under the same operating system, and transfer of data was done through file systems. Besides the observation that the various software pieces could be used in sequence, the most important aspect of this work was the fact that the team thought they had come up with an informed decision for which the integration was seen as the critical component.

The idea of coupling GIS and other software was formalized by (Goodchild 1987), (Nyerges 1993), and others. Two packages were said to be tightly coupled when the user was presented with a single interface, and the two packages interacted with a common database. Loose coupling merely required the exchange

of data between the two packages, often with a third software component for format conversion. Finally, functions were embedded in GIS when they were executed within the GIS, using the GIS user interface.

We commonly think of geographic information as somehow homogenous, but in reality very different concepts are required to understand the distinction between a set of points sampling variation that is conceived as a single field, versus a collection of points representing the locations of outbreak of a disease, for example. Attempting to establish interoperability across this vast range of distinct concepts may be doomed from the outset—instead, it may be necessary to identify domains within which interoperability can reasonably be achieved, but between which interoperability is practically impossible.

3 Problem in Combining Spatial Data

There are several problems in combining distributed heterogeneous spatial data that can be categorised into two categories. The first category is the institutional problems that related to the rules and regulations that providing a control over sharing the spatial data between those organisations providing these kind of data. The second category is the technical problems which related to the technical issues for providing tools for a better querying such these systems in order to maintaining the transparency for the users of these kind of systems, resolving the heterogeneity of the spatial data, and the designing a communication protocols to enable the data to be transferred from one system to the other.

In this research we will focus only on the technical issues, were a prototype framework has being developed in order to resolve some of technical issue problems.

4 Distributed Query Process

In the client-server approach, the databases is persistently stored by server machines and the queries are initiated at client machines. Three approaches can be determined in order to execute the query namely the query shipping, the data shipping, and the Hybrid shipping approach.

4.1 Query shipping approach

The principle of the query shipping approach is to execute queries at the servers, i.e., at the lowest level possible in the hierarchy of the sites. The server performs all of the query-processing effort, and the answer is returned in the form of a stream of tuples.

4.2 Data shipping approach

In contrast, the Data Shipping approach does no query processing at the server. Instead, the query processing is all done at the client machines. As data is

needed by client query processing algorithms, it is requested from the server. The data shipping approach makes much better use of client resources than query shipping, but it can also substantially increase communication costs; further, it can lead to under-utilization of server resources.

4.3 Hybrid shipping

This approach allows the system to choose to perform some query processing on the server, and some portion of the processing on the client machine. Provided that an effective optimizer can be developed to choose the appropriate alternative, this approach has the ability to always perform at least as well as either of the two pure shipping approaches, and often much better.

The problem of finding efficient techniques for processing complex queries has been of keen interest in query optimization. In a way, decision support systems provide a testing ground for some of the ideas that have been studied before. We will only summarize some of the key contributions. There has been substantial work on "unnesting" complex SQL queries containing nested subqueries by translating them into single block SQL queries when certain syntactic restrictions are satisfied [17, 18, 19, 20]. Another direction that has number of invocations and batching invocation of inner subqueries by semi-join like techniques [21, 22]

5 Characteristy of the spatial data

The georeferenced spatial data has four major components: the geographic position component, the attribute component, the spatial relationship component, and the time component (Arnof 1993). The geographic position and the relationship components represents the geometric dicription of the object, and the attribute component represent the thematic dicription of the object. There are two principle structures to represent the spatial objects and linking their thematic and geometric data (raster data structure and vector data structure) :

- Raster data structure: a collection of points or cells distributed in a regular grid. Each cell in a grid is assigned to a thematic value which refers to a feature in the real world. Raster structure can be either single-value or multi-value type. In single value type, the raster has one attribute representing one particular thematic aspect of the terrain. In the multi value type, each raster has several attributes for the same terrain segment. The topology of the raster structure is based on the adjacency of the raster point or cell.
- Vector data structure: objects in the real world represented by point and lines that define their boundaries. The position of each object is defined by its placement in a map space that organized by a coordinate reference system. Points, lines and polygons are used to represents irregularly distributed geographic objects in the real world.

5.1 Heterogeneity Problem

Heterogeneity refers to the differences existing in all levels of participating systems. There is platform related heterogeneity (network, hardware, operating system, and application development tools such as distributed computing platform), and there is application related heterogeneity (data structure differences, data schema differences in the case of distributed database system).

The heterogeneity problem occurs when two different data organisation agreed to share their data with each other. And they need to keep their own autonomy over their system. This problem is related to the difference to how they abstract the real world phenomena, the difference in modelling schemas, the difference of the tools they use to display, store, process, and managing their data. (Ubbo Visser, et al 2002; Bishr, Y. et al 1999; Bishr 1997) describe these heterogeneity issues as syntactic, schematic, and semantic heterogeneity

5.2 Syntactic heterogeneity

Syntactic heterogeneity refers to the differences in software and hardware platforms, database management system, and the representation of the geospatial object (raster or vector, coordinate system, geometric resolution, quality of geometric representation, methods of data acquisition, etc.)

Schematic heterogeneity

Schematic heterogeneity refers to the differences in database models or schemas, e.g. a particular feature may be classified under different object classes in different databases, or an object in one database may be considered an attribute in another. The classes, attributes, and their relationships can vary within or across disciplines.

5.3 Semantic heterogeneity

Semantic heterogeneity is the way the same real world entity may have several meanings in different databases. This will also influence the geometrical representation of objects, because abstraction of the world is based on the semantics of each disciplines. It is intimately tied to the application context or discipline for which the data is collected and used.

6 Geographic Information Infrastructure

GII is a mechanism to provide GIS applications with access to geospatial data from distributed sources with different levels of details within institutional framework. Thus GII is important for anyone who works in GIS and remote sensing application areas. GII also provides inter-operability interfaces that are necessary for communication between different geospatial databases that have different spatial reference systems, different internal data formats, different units of measurements, and different geometric representation methods (Groot and

McLaughlin,2000). The global objectives of the geographic information infrastructure is to provide a political, institutional, economical, and technical platform to share geospatial informations (Bishr, Y. 1996). This can be further discribed as a set of institutional, economical, and technical arrangements to support the availability of relavent, up-to-date, and integrated geoinformation, timely and at an affordable cost (Radwan et all 1996).

The problems related to the information sharing in GII may be subsumed under two manin categories:

- Organizational problems: which include institutional organization and access ruls, legislation (e.g. copyright), pricing schema, and standards.
- Technical problems: which include inventory of the available data, mechanizem for seamlessly sharing data, data consistency and data quality, data exchange, and updating data.
- Data interchange is an essential element in achieving interoperability between heterogeneous systems. A lot of effort is being put into the field of interoperability research (Kim, W. 1995; Bishr, Y. 1998; Sheth, A.P, 1999). Three fundamental interoperability issues are addressed here. The first issue, semantic interoperability, is to define and agree on the semantics of the content and logical structures of data. The second issue is the schematic or structure interoperability, is to define unified hierarchies and attribute for two different independent database schema. The third issue, syntactic interoperability, is to define a system and platform independent data structure that can represent data corresponding to the application schema. XML and GML have been chosen as a data interchange format because of its simplicity, Web conformance, and extensively tool support.

6.1 Architecture of the GII

From the previous section, one may conclude that the GII can be formulaized in a Client-Server architecture. This architecture includes three main components, namely client interface, server or service provider, and data provider

Client: which provides a means of interacting between the user and the service provider or the server. It provides the user of the system with the nessesary interface to query and display the final results.

Server or service provider: which provides the transparency of the system to the user, so that the user does not need to know what to search and from where. The server also has to have the ability to integrate the retrived information in response to the user query. Also provide communication protocol to communicate with the client and the data provider.

Data Provider who provides the facility for querying the local data and wrp-ping the data to the format required by the Server.

7 Summary

The problems of combining spatial data arises when one has to deal with different autonomic data sources. The deferency arises from the way that the data collected, modelled, stored, and even presented in each independent databases. In this chapter three major problems can be concluded when data need to be combined. The first problem is the transparency of the query process, were a query like "what if " may need to be addressed to more than one data sources. The second problem is the characteristy of the spatial data which reflect the autonomy and the independency of the data provider, this problem is a sub-set of the last and the major problem, which is the hetrogeneity problem. The hetrogeneity problem is the main problem in combining data from heterogenous data suorces.

In chapter five, this type of problems will be discussed in the implementation of the prototype of the system.

Reference

1. Albrecht, J. (1999) Geospatial information standards. A comparative study to approaches in the standardisation of geospatial information. *Computers and Geosciences*, 25, 9-24.
2. Bishr, Y. A., Pundt, H., Kuhn, W., and Rdwan, M. (1999) Probing the Concepts of Information Communities - A First Step Toward Semantic Interoperability. in: M.Goodchild, M. Egenhofer, R. Fegeas, and C. Kottman, (Eds.), *Interoperating Geographic Information Systems*. pp. 55-70, Kluwer, Norwell, MA.
3. Sheth, A.P., Changing focus on interoperability in information systems: From system, syntax, structure to semantics, in *Interoperating Geographic Information Systems*, M.F. Goodchild, et al., Editors. 1999, Kluwer Academic Pub. ISBN 0792384369.
4. Kim, W., ed. *Modern Database Systems: The Object Model, Interoperability, and Beyond*. 1995, Addison-Wesley. ISBN 0-201-59098-0.
5. Bishr, Y., Overcoming the semantic and other barriers to GIS interoperability. *International Journal of Geographic Information Science*, 1998. 12(4): p. 299-314.
6. Nyerges, T. (1993) In M.F. Goodchild, B.O. Parks, and L.T. Steyaert, editors, *Environmental Modeling with GIS*. New York: Oxford University Press.
7. Goodchild, M.F. (1987) A spatial analytic perspective on geographical information systems. *International Journal of Geographical Information Systems* 1(4): 327-334.
8. Buehler K. and McKee L. (1996) *The OpenGIS Guide*. <http://www.opengis.org>, The Open GIS Consortium, Inc.

9. OGC, 1996. "The OpenGIS® Guide - Introduction to Interoperable Geoprocessing. Part 1 of the Open Geodata Interoperability Specification (OGIS)". OGIS Project Technical Committee of the Open GIS Consortium, Inc, 1996. Eds.: Buehler, K. and L. McKee. Wayland, Massachusetts, USA. <http://www.opengis.org/techno/guide/guide.doc>
10. Groot, R. and McLaughlin, J., 2000, Geospatial Data infrastructure concepts: cases and good practice, Oxford university press: New York , United States.
11. Evans John D., and Ferreira Joseph Jr., 1995. "Sharing spatial information in an imperfect world: interactions between technical and organizational issues," Chapter 27 of Onsrud, Harlan J., and Rushton, Gerard (eds.), Sharing Geographic Information. New Brunswick, NJ: Center for Urban Policy Research, Rutgers University.
12. Tosta, Nancy, 1994. Continuing evolution of the National Spatial data Infrastructure. In Proceedings, GIS/LIS '94. Bethesda: ACSM-ASPRS-AAG-URISA-AM/FM. Vol. 1, pp. 768-776.
13. Ubbo Visser, Heiner Stuckenschmidt, Christoph Schlieder, 2002. Interoperability in GIS - Enabling Technologies 5th AGILE Conference on Geographic Information Science, Palma (Balearic Islands, Spain) April 25 th -27 th 2002. Also can be found on the internet
14. http://agile.isegi.unl.pt/Conference/Mallorca2002/Papers/pdf/dia26/Session_7/s7_Visser.pdf
15. Sondheim, M. Gardels, K. and Buehler, K, (1998) GIS interoperability. In D. J. Maguire, M. F. Goodchild, and D. W. Rhind, (eds). Geographical Information Systems: Principles and Technical Issues. Vol 1. John Wiley & Sons, Chichester, 347-358.
16. Vckovski, A. (1998) Special Issue: Interoperability in GIS. International Journal of Geographical Information Science, 12(4), 297-298
17. Gardels, K., (1996) The Open GIS approach to Distributed Geodata and Geoprocessing. Proceedings, Third International Conference/Workshop on Integrating GIS and Environmental Modelling, Santa Fe, NM, January 21-25.
18. Kim W. "On Optimizing a SQL-like Nested Query" ACM TODS, Sep 1982.
19. Ganski,R., Wong H.K.T., "Optimization of Nested SQLQueries Revisited " Proc. of SIGMOD Conf., 1987.
20. Dayal, U., "Of Nests and Trees: A Unified Approach to Processing Queries that Contain Nested Subqueries, Aggregates and Quantifiers" Proc. VLDB Conf., 1987.

21. Murlaikrishna, "Improved Unnesting Algorithms for Join Aggregate SQL Queries" Proc. VLDB Conf., 1992.
22. Seshadri P., Pirahesh H., Leung T. "Complex Query Decorrelation" Intl. Conference on Data Engineering, 1996.
23. Mumick I.S., Pirahesh H. "Implementation of Magic Sets in Starburst" Proc.of SIGMOD Conf., 1994.