Database and Representation Issues in Geographic Information Systems (GIS)[⋆]

Hanan Samet

Computer Science Department, Center for Automation Research, Institute for Advanced
Computer Studies, University of Maryland, College Park, Maryland 20742
hjs@cs.umd.edu, www.cs.umd.edu/~hjs

Abstract. A review is provided of some database and representation issues involved in the implementation of geographic information systems (GIS).

The increasing popularity of web-based mapping systems such as Microsoft Virtual Earth and Google Earth and Maps, as well as other software offerings that are coupled with portable devices, such as the iPhone, has led to a proliferation of services that are characterized as being location-based. The data provided by these services is differentiated from other offerings by the presence of a locational component. In the past, this type of data was found primarily in geographic information systems (GIS). The available technology led to a focus on the paper map as the output device for responses. Since anything is better than drawing by hand, there was little emphasis on efficiency measures such as minimization of execution time.

However, the emergence of of display devices has changed the mode of operation to one of expecting answers relatively quickly. This has had a number of effects. First, the paper medium supports relatively high resolution output while the display screen is usually of a more limited resolution, thereby enabling the use of less precise algorithms. For example, a spatial range query such as a buffer/corridor (e.g., a query that seeks all cities of population in excess of 20,000 within 100 miles of the Mississippi River) often takes quite a bit of time to compute (i.e., the extent of the spatial region in question) when using the Euclidean distance. Instead, we can use a Chessboard distance (L_∞) to approximate the Euclidean distance coupled with a quadtree representation to yield algorithms that are several orders of magnitude faster [1] but, of course, may not be as precise.

Second, the ability to obtain answers quickly led to an increase in the volume of data that is input to the system. Such high volumes of data need to be organized and the obvious next step was to make use of databases, which meant incorporating this technology in the GIS thereby transitioning the field to one that is also known as spatial databases [2–7]. A natural outgrowth of this transition, which will eventually emerge, is the use of spatial spreadsheets [8, 9] as they traditionally enable users to ask what-if questions and see the results instantaneously. There are several possible innovations here.

[⋆] This work was supported in part by the National Science Foundation under Grants IIS-08-12377 CCF-08-30618, and IIS-07-13501, as well as NVIDIA Corporation, Microsoft Research, Google, the E.T.S. Walton Visitor Award of the Science Foundation of Ireland, and the National Center for Geocomputation at the National University of Ireland at Maynooth.

One innovation lies in the fact that the map plays the role of the spreadsheet with the the different attributes being the columns and the rows playing the roles of the individual tuples. The map can be used to visualize spatial variability by using sliders to select ranges of values of the attributes and show the tuples that satisfy them.

A second innovation lies in the fact that as compositions of operations are performed, data is generated with a locational component [10, 11]. The results can be viewed as maps and are stored as relations containing the tuples that satisfy the queries. For example, suppose that we want to find the locations of bridges and tunnels. One way to do so, assuming the existence of a roads and a rivers relation, is to take the spatial join of the two relations. The result would be a set of pairs of the form (road$_i$,river$_j$) where the spatial attribute would consist of the location(s) that road$_i$ and river$_j$ have in common. Notice that the input spatial attribute of both the roads and rivers relation is a line segment or a collection of piecewise linear line segments.

In addition to speed, there has also been much activity in expanding the range of queries for which answers are expected. In particular, spatial queries can be broken down into the following two classes: location-based and feature-based. A location-based query takes a location, traditionally specified using lat/long coordinate values, as an argument, and returns a set of features that are associated with the location, while a feature-based query takes a feature as an argument and returns the set of locations with which the feature is associated. The queries can also be characterized as functions where one function is viewed as the inverse of the other. Feature-based queries have also become known as spatial data mining [12, 13].

Although features are usually properties of the data such as crops, soil types, zones, etc., they can be much more diversified. In particular, they can correspond to collections of unstructured data such as documents in which case the queries reduce to finding all documents that mention location X, and possibly others, or finding all locations mentioned in document Y, and possibly others, that could be related in some manner. STEWARD [14] is an example of a system that supports such queries on a collection of documents on the hidden web, while NewsStand [15] is an example of a system that supports such queries on a collection of news articles where the relationship between the documents is that they are on the same or related set of topics. The two systems are also distinguished by the fact that the collection of documents in the former is relatively static, while in the latter it is very dynamic in the sense that its composition is constantly changing as new articles are processed and old ones fade away in importance.

The inclusion of documents in the range of features reminds us that although spatial data is usually specified geometrically, in this case it can also be specified using collections of words of text that can be (but are not required to be) interpreted as the names of locations Textual data that corresponds to spatial data are called *toponyms* and its specification invariably involves some ambiguity. This ambiguity has both advantages and disadvantages. The advantage of the ambiguity is that, from a geometric standpoint, the textual specification captures both point and spatial extent interpretations of the data (analogous to a polymorphic type in parameter transmission which serves as the cornerstone of inheritance in object-oriented programming languages). For example, geometrically, a city can be specified by either a point such as its centroid, or a region corresponding to its boundary, the choice of which depends on the level of zoom with which the query interface is activated. On the other hand, the disadvantage of the ambiguity is that we are not always sure if a term is a geographic location or not (e.g., does "Jordan" refer to a country or is it a surname as in "Michael Jordan"?). Moreover, if it is a geographic location, then which, if any, of the possibly many instances of geographic locations with the same name is meant (e.g., does "London" refer to an instance in the UK, Ontario, Canada, or one of many more others?).

The examples that we have outlined serve to show many of the database and representation issues involved in geographic information systems. Some additional issues include:

1. How to integrate spatial with nonspatial data in a seamless manner.
2. Retrieval is facilitated by building an index (e.g., [16–18]). There is a need to find a way to sort the data [19]. The index should be compatible with the data being stored. We need an implicit rather than an explicit index as it is impossible to foresee all of the possible queries in advance. For example, if we sort all cities with respect to their distance from Chicago, we can find the nearest city to Chicago with population in excess of 200,000. However, this sort will not help in finding the closest city to Denver with population in excess of 200,000.
3. There is a need to identify the possible queries and to find their analogs in a conventional database [20–22]. For example, a map in a spatial database is like a relation in a conventional database. However, the difference is the presence of input spatial attributes and also the presence of output spatial attributes as in the rives/roads example. Another example, is the combination of ranking and the distance semijoin [23] to yield a discrete Voronoi diagram (e.g., [24]) and the ability to do clustering. A recent example, is the ability to perform queries on spatial networks using simple SQL commands [25].
4. How to interact with the database. SQL may not always be easy to adapt. It may be desirable to make use of a graphical query language.
5. Determining what functionality users really desire and need, and providing it.
6. How to ensure the spatial integrity of the data such as that the edges of a polygon link to form a complete object, that line segments do not intersect except at vertices, that contour lines do not cross, etc.
7. Develop a strategy for answering a query that mixes spatial data with nonspatial data. This implies a need for query optimization strategies [26], which in turn calls for the definition of selectivity factors. This depends on whether or not an index exists on the spatial data. If not, then select on the nonspatial data first. Otherwise, the situation is more complex as we perform the spatial selection first only if there is high spatial selectivity (e.g., the range in a spatial range query is small).
8. How to incorporate time-varying data as well as deal with the fact that temporal data, as well as spatial data, is also continuous rather than being restricted to discrete which is the case when the valid time and transaction time primitives are being used. This will enable the handling of rates.
9. Processing data lying on a spatial network using network distance [25, 27–31].
10. How to incorporate imagery into the database.
11. Interoperability.
12. How to make use of advanced computing architectures such as GPUs (e.g., [32]).

13. Resolving ambiguities in the textual specification of spatial data with no errors (or almost none).
14. Determining the geographic focus of a set of documents on a related topic.

As the above show, many database and representation issues are involved in geographic information systems that need resolution, thereby forming a vibrant area of research.

References

1. Ang, C.H., Samet, H., Shaffer, C.A.: A new region expansion for quadtrees. IEEE Transactions on Pattern Analysis and Machine Intelligence **12** (1990) 682–686 Also see *Proceedings of the Third International Symposium on Spatial Data Handling*, pages 19–37, Sydney, Australia, August 1988.
2. Günther, O.: Environmental Information Systems. Springer-Verlag, Berlin, Germany (1998)
3. Güting, R.H.: An introduction to spatial database systems. VLDB Journal **3** (1994) 401–444
4. Laurini, R., Thompson, D.: Fundamentals of Spatial Information Systems. Academic Press, San Diego, CA (1992)
5. Rigaux, P., Scholl, M., Voisard, A.: Spatial Databases with Application to GIS. Morgan-Kaufmann, San Francisco (2002)
6. Shekhar, S., Chawla, S.: Spatial Databases: A Tour. Prentice-Hall, Englewood-Cliffs, NJ (2003)
7. Worboys, M.: GIS A Computing Perspective. Taylor & Francis, London, United Kingdom (1995)
8. Iwerks, G.S., Samet, H.: The spatial spreadsheet. In Huijsmans, D.P., Smeulders, A.W.M., eds.: Proceedings of the 3rd International Conference on Visual Information Systems (VISUAL99), Amsterdam, The Netherlands (1999) 317–324
9. Iwerks, G.S., Samet, H.: The internet spatial spreadsheet: enabling remote visualization of dynamic spatial data and ongoing query results over a network. In Hoel, E., Rigaux, P., eds.: Proceedings of the 11th ACM International Symposium on Advances in Geographic Information Systems, New Orleans, LA (2003) 154–160
10. Samet, H., Alborzi, H., Brabec, F., Esperança, C., Hjaltason, G.R., Morgan, F., Tanin, E.: Use of the SAND spatial browser for digital government applications. Communications of the ACM **46** (2003) 63–66
11. Brabec, F., Samet, H.: Client-based spatial browsing on the world wide web. IEEE Internet Computing **11** (2007) 52–59
12. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann, San Francisco (2000)
13. Aref, W.G., Samet, H.: Efficient processing of window queries in the pyramid data structure. In: Proceedings of the 9th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS), Nashville, TN (1990) 265–272 Also in *Proceedings of the Fifth Brazilian Symposium on Databases*, pages 15–26, Rio de Janeiro, Brazil, April 1990.
14. Lieberman, M.D., Samet, H., Sankaranarayanan, J., Sperling, J.: STEWARD: architecture of a spatio-textual search engine. In Samet, H., Schneider, M., Shahabi, C., eds.: Proceedings of the 15th ACM International Symposium on Advances in Geographic Information Systems, Seattle, WA (2007) 186–193
15. Teitler, B., Lieberman, M.D., Panozzo, D., Sankaranarayanan, J., Samet, H., Sperling, J.: NewsStand: A new view on news. In Aref, W.G., Mokbel, M.F., Samet, H., Schneider, M., Shahabi, C., Wolfson, O., eds.: Proceedings of the 16th ACM SIGSPATIAL International

Conference on Advances in Geographic Information Systems, Irvine, CA (2008) 144–153 (2008 ACM SIGSPATIAL (ACMGIS08) Best Paper Award).

16. Samet, H.: Applications of Spatial Data Structures: Computer Graphics, Image Processing, and GIS. Addison-Wesley, Reading, MA (1990)
17. Samet, H.: The Design and Analysis of Spatial Data Structures. Addison-Wesley, Reading, MA (1990)
18. Samet, H.: Foundations of Multidimensional and Metric Data Structures. Morgan-Kaufmann, San Francisco (2006)
19. Samet, H.: A sorting approach to indexing spatial data. International Journal on Shape Modeling **14** (2008) 15–37
20. Aref, W.G., Samet, H.: An approach to information management in geographical applications. In: Proceedings of the 4th International Symposium on Spatial Data Handling. Volume 2., Zurich, Switzerland (1990) 589–598
21. Aref, W.G., Samet, H.: Extending a DBMS with spatial operations. In Günther, O., Schek, H.J., eds.: Advances in Spatial Databases—2nd Symposium, SSD'91. vol. 525 of Springer-Verlag Lecture Notes in Computer Science, Zurich, Switzerland (1991) 299–318
22. Samet, H., Aref, W.G.: Spatial data models and query processing. In Kim, W., ed.: Modern Database Systems, The Object Model, Interoperability and Beyond. ACM Press and Addison-Wesley, New York (1995) 338–360
23. Hjaltason, G.R., Samet, H.: Incremental distance join algorithms for spatial databases. In Hass, L., Tiwary, A., eds.: Proceedings of the ACM SIGMOD Conference, Seattle, WA (1998) 237–248
24. Samet, H., Phillippy, A., Sankaranarayanan, J.: Knowledge discovery using the SAND spatial browser. In: Proceedings of the 7th National Conference on Digital Government Research, Philadelphia, PA (2007) 284–285
25. Sankaranarayanan, J., Samet, H.: Distance oracles for spatial networks. In: Proceedings of the 25th IEEE International Conference on Data Engineering, Shanghai, China (2009) 652–663
26. Aref, W.G., Samet, H.: Optimization strategies for spatial query processing. In Lohman, G.M., Sernadas, A., Camps, R., eds.: Proceedings of the 17th International Conference on Very Large Databases (VLDB), Barcelona, Spain (1991) 81–90
27. Sankaranarayanan, J., Samet, H., Alborzi, H.: Path oracles for spatial networks. In: Proceedings of the VLDB Endowment PVDB: Proceedings of the 35th International Conference on Very Large Data Bases (VLDB). Volume 2., Lyon, France (2009) 1210–1221
28. Samet, H., Sankaranarayanan, J., Alborzi, H.: Scalable network distance browsing in spatial databases. In: Proceedings of the ACM SIGMOD Conference, Vancouver, Canada (2008) 43–54 Also see University of Maryland Computer Science Technical Report TR–4865, April 2007 (2008 ACM SIGMOD Best Paper Award).
29. Sankaranarayanan, J., Alborzi, H., Samet, H.: Efficient query processing on spatial networks. In: Proceedings of the 13th ACM International Symposium on Advances in Geographic Information Systems, Bremen, Germany (2005) 200–209
30. Sankaranarayanan, J., Alborzi, H., Samet, H.: Enabling query processing on spatial networks. In: Proceedings of the 22nd IEEE International Conference on Data Engineering, Atlanta, GA (2006) 163
31. Sankaranarayanan, J., Alborzi, H., Samet, H.: Distance join queries on spatial networks. In: Proceedings of the 14th ACM International Symposium on Advances in Geographic Information Systems, Arlington, VA (2006) 211–218
32. Lieberman, M.D., Sankaranarayanan, J., Samet, H.: A fast similarity join algorithm using graphics processing units. In: Proceedings of the 24th IEEE International Conference on Data Engineering, Cancun, Mexico (2008) 1111–1120