PARALLEL SEARCH IN SPATIAL DATA DISTRIBUTION

Mark L. Palmer
Digital Equipment Corporation
2 Iron Way MRO3-1/E13
Marlboro, MA U.S.A. 01752

ABSTRACT

GIS systems need to be able to distribute vast collections of spatial data so that the task of finding data that satisfies user queries is accomplished most expediently. The most desirable characteristics of such query systems are completeness, distributability, generality, mutability, and simplicity. A plausible general paradigm for distributing search is composed of a network of concurrent processes which cooperate in answering user queries. Certain aspects of this type of architecture are interdependent, and trade-offs between them must be considered. In particular, larger networks require better heuristics with which to constrain search. An approach to designing such heuristics consists of treating data points "probabilistically" in order to represent the uncertainty present in a generalization of a data population. These approaches to distribution and search satisfy the named criteria.

INTRODUCTION

This paper proposes an approach to distributing spatial data, a capability which is lacking in current GIS systems. This capability is needed because of the immense volumes of data typically associated with GIS processing. A plausible architectural paradigm is discussed as well as an approach to generalizing and searching spatial data which is based on probability and well suited for use in a distributed environment.

CRITERIA FOR SEARCH MECHANISMS

Before propounding any particular means of searching it is useful to identify what characteristics an ideal mechanism would have.

The mechanism will be composed of a generalized representation of data such as an index, tree, etc. (hereafter referred to as a "data sketch") which purports to be simpler and faster to search than the raw data itself, and an algorithm which uses the sketch to identify data that satisfies constraints specified by a user.

Each data set and its data sketch will form a "node" which, together with other nodes, form a "search network". This network may have heterarchical or hierarchical structure. For the purposes of this discussion, a search network is assumed to be bounded at depth of 1, fully connected, and heterarchical.

The ideal search mechanism should have the following characteristics:

Completeness
The search mechanism must guarantee that all available data will be found. Given the value of most spatial data, any mechanism which may fail to find data because of "type 2" uncertainty (Frank, 1985) is unacceptable.

Distributability
The search algorithm should lend itself to functional decomposition for parallel execution. Decomposition of any algorithm is constrained by dependencies resulting from the data flow required. Unit operations which rely on the output of another operation can only be executed after the first operation provides a result. The examination of one node to find data satisfying a query can be considered a unit operation. By definition, search does not modify the data sought. If the search operation is the same for any node, and the results of searching any node are not dependent upon the output from the search of any other node, it is plausible to execute the search operator concurrently on all nodes and allow each search to return results to a user or collection point independently.

Generality
Adaptation for 3D data should be trivial. Locational and attribute information should be treated similarly. Whether the search tree is local or spread over a communications network should not matter. Current approaches to data sketching require exponentially more resource upon adaptation to a 3D world, and do not adequately represent certain data, e.g. regular grid representations of reverse faults.

Mutability
The data sketch should be "refreshed" quickly to reflect changes in data at a node. Some current methods require hours to produce a data sketch which is expensive to modify dynamically when changes are made.

Simplicity
The location of data in the search network must not be restricted. It is unacceptable to place artificial restrictions on where data actually resides in the search network in order to enable effective search. Simpler types of representation are "refreshed" more quickly and are therefore more mutable.

PROPOSED PARADIGM

The characteristics of a model for parallel search which addresses the above criteria are:

Query Manager
The model consists of an arbitrary number of spatial data populations organized in an heterarchical network such that one independent process is dedicated at each node to manage queries. This process maintains a sketch of the data for which it is responsible, shares its data sketch with other nodes in the network, and answers queries posed by user applications by coordinating search of local and other nodes.

Data Clustering
The clustering of data into populations is initially assumed to be arbitrary. Ultimately, the query manager processes would be automatically assigned to particular populations of data to maximize the quality, or "representativeness" of the data sketches.

Process Communication
There is an application-layer protocol capable of establishing the shared universe of discourse (ISO 1982) for all nodes and of propagating queries to any other node in the network.

Session Layer
Each node can communicate with any other node without regard to whether the other node is on the same machine or connected by a computer network. This service is provided by a "session layer".

Bounding Constraints
The search network is assumed to be of depth 1 from the point of view of any of its nodes; removal of this constraint requires having a method for eliminating "cycles" in the path of a query through the network.

It is assumed that queries are propagated to nodes which are accessed sequentially rather than simultaneously, although nodes process the queries concurrently after receiving them.

Search Algorithm

The search operation may be informally decomposed into:

1. Identify nodes which may provide positive results and rank them.

2. Initiate search at all candidates by rank.

3. Examine local data via its sketch and return answers to requester.

Step 1 must happen before step 2, but both may be done concurrently with step 3 to increase the degree of parallelism. For a search network of depth 1, steps 1 and 2 are done only at the node first receiving the query.

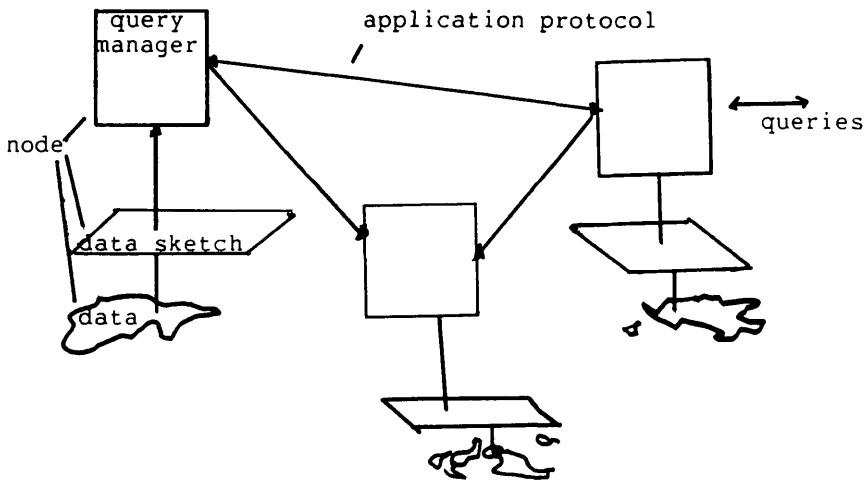

Figure 1: A Search Network of 3 Nodes

DESIGN TRADE-OFFS

Interdependencies between various aspects of the system's performance will exist in any design derived from this architectural paradigm. These should be considered for any proposed design.

- Increasing the representativeness of a data sketch generally increases its complexity and increases the time required to search it. As more information is lost through generalization, more uncertainty is present in the sketch, and queries against it have less chance of succeeding.

- As the amount of independently controllable access to storage devices used increases, greater benefit can be realized by introducing parallelism.

- As node access costs increase, time available to examine data sketches decreases, therefore the desired amount of representativeness of the sketches increases.

- Search cost increases with overall network size, total amount of data, and node access costs.

PROBABILISTIC SKETCHING AND SEARCH HEURISTICS

For larger networks especially, the scope of search must be narrowed by using computationally inexpensive heuristics. Such heuristics are useful whether or not the mechanism is distributed. For distributed mechanisms, however, the need to eliminate nodes from consideration becomes less important than the need to order search. If the "best" nodes can be set in action first, the time between issuing a query and receiving the first responses is minimized, while nodes which are least likely to produce results may still be examined with minimal delay to the entire operation. Therefore, a "best-first" heuristic approach which doesn't necessarily eliminate lots of nodes but can reliably direct search first towards nodes where it is most likely to find data will be very useful.

The question of "where to look first" can be treated as a matter of "reasoning under uncertainty", a concept borrowed from the field of AI. Spatial data locality can be represented in probabilistic terms, and the uncertainty of a data sketch can be made made explicit for heuristic use with the approach described below.

The "certainty factor" in the MYCIN system (Winston, 1977) was an approximation of the cumulative degree of belief in a deduction. During search for spatial data, an analogous "certainty factor" can be calculated which approximates the

chance that the query being processed will produce positive results by thorough search at the node in question. The "certainty factor" for any given node and locational query is a function of the degree of commonality between the query and the data sketch and of the representativeness of the data sketch. The derivation of this factor is as follows:

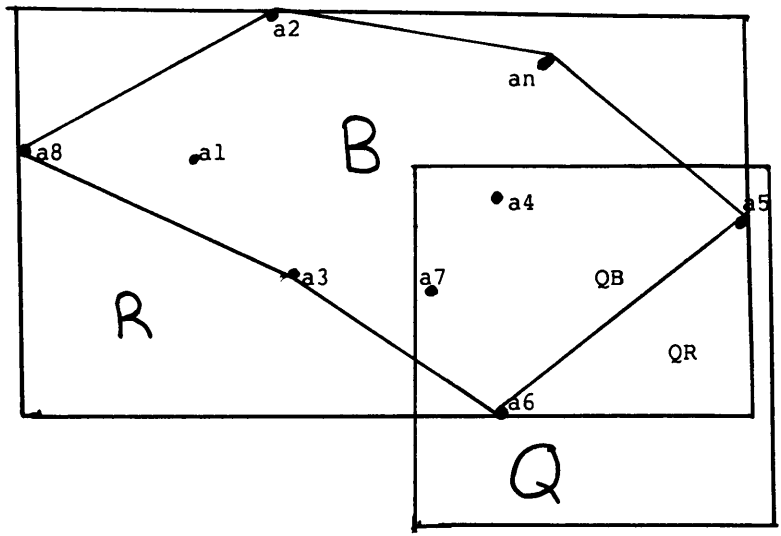

Figure 2: Points, Bounding Polygon and Rectangle, Query

Firstly, make the convenient and pragmatic assumption that there is a finite number, p, of "points" (which may be arbitrarily but not infinitely small) per unit of measure. If p = 1 "points" become the unit of measure. Think of areas and volumes as having a number of these points which is in constant proportion to, or equal to, the measure.

Next, think of each point as a binomial event. That is, for each population of points, we can state for each point whether member of a subset of points within the population which share a certain property, for example the property of belonging to the same chain, area, or volume.

Now, let A be the number of points randomly distributed in subset A within a population of B points. The probability that a point known to be in B is also in A is

$$P(A) = A/B \quad (1)$$

If A is a number of spatial data points we wish to generalize, and B is the area of the bounding polygon for A, we can represent the probability that a point within area B will also be in A by [1].

The bounding polygon and the total point count are all that is explicit in the sketch of these points, a situation reminiscent of a "Heisenberg" enclosure. The points are treated as if they are as likely to be distributed in any way inside of the border. The chance of finding a point in A at any given point is a function of the point density. If the "representativeness" of B relative to A is taken to be P(A), the "uncertainty" inherent in B is expressed as

$$U(B) = 1 - P(A) \quad (2)$$

The probability, P(QA), that Q will "find" one point which is a member of A, given P(A) and QB, is approximated by a Hypergeometric probability distribution:

$$P(QA) = \frac{C_1^A \, C_{QB-1}^{B-A}}{C_{QB}^{B}} \quad (3)$$

where QB is the sample size; B is the population size, A is the number of "successes" in the population. [3] is solving for the probability of one success in the sample; substituting greater numbers would lead to a higher confidence.

Suppose P(A) = 1. This would be true if the datum to be represented, A, were itself a polygon. This polygon is generalized using its bounding rectangle, R. Then P(A) would be A/R, and P(QR) could be approximated as above, using the intersection of Q and R, QR, as the sample size. In particular, we know that if A + QR > R then P(QA) = 1. If P(A) < 1 because of uncertainty present in the polygon sketch B, P(QA) is the compound probability that a point in QR is in B AND is a point in A:

$$P(QA) = P(QB) \; P(A) \quad (4)$$

Between any representation of a particular set of data and its next-most-general representation some specific information is lost; the amount of specificity lost may be represented by the ratio of total coverages in the two representations, limited by the uncertainty of the more specific representation. The uncertainty present in any

data sketch is a function of the cumulative uncertainty of lower-level sketches and the degree to which the entity being "sketched" instantiates its sketch.

Using these principles, data sketches with accurate information about their representativeness can be composed simply, using area ratios. This allows a certainty factor to be calculated and used as an heuristic for any query against the sketch, allowing nodes to be searched in "best-first" order.

SUMMARY

Within the framework described for a search network, data sketches, and spatial heuristics, how are the criteria initially stated addressed?

The "completeness" requirement discussed above can be satisfied by ensuring that all data populations are represented by an all-inclusive sketch. As the search narrows in on a particular item of data, it is always directed to that item by finding that an all-inclusive sketch of the item is applicable to answering the query.

Because the search network can exist on many different resources, and given the definition of the search operation, the work of searching is distributable.

The principles proposed for heuristics and data sketching are general. They are easily extended for use with attribute data if the attribute nomenclature is unambiguous, shared, and hierarchical. They can be easily extended for use with 3-D data.

Any design using this approach should be mutable since the elements used are commonly found in current-day representations of spatial data (e.g. points, bounding polygons, bounding rectangles etc.) and do not take long to calculate or recalculate to reflect changes.

This approach is simple; it does not require that data be arranged in any regular or particular way. Extensive interpolation and extrapolation procedures are not required to fit values into a rigid representational system. Sketches may be arranged to represent "clusterings" of data within the common coordinate system so as to maximize the collective representativeness of the sketches; it adapts to whatever the existing distribution of data collections is.

ACKNOWLEDGMENTS

The author wishes to thank Barbara Higgins for editorial assistance.

REFERENCES

Peuquet, D. and Chen, Z. (1985) "Quad Tree Spatial Spectra Guide: A Fast Spatial Heuristic Search in a Large GIS", Auto-Carto 7 Proceedings, 75 - 82

Frank, A. and Robinson, V. B., (1985) "About Different Kinds of Uncertainty in Collections of Spatial Data", Auto-Carto 7 Proceedings, 75 - 82

Winston, P.H. (1977) Artificial Intelligence, 243 - 246 Addison-Wesley Publishing

Pearl, J. (1984) Heuristics - Intelligent Strategies for Computer Problem Solving, 124-140 Addison-Wesley Publishing

ISO Working Group 3 (1982) Concepts and Terminology for the Conceptual Schema and the Information Base, publication ISO/TC97/SC5 - N 695