Knowledge Discovery using the SAND Spatial Browser*

Hanan Samet
Adam Phillippy
Jagan Sankaranarayanan
Department of Computer Science
Center for Automation Research
Institute for Advanced Computer Studies
University of Maryland at College Park
{hjs,amp,jagan}@umiacs.umd.edu

ABSTRACT

The use of the SAND Internet Browser as a knowledge discovery tool for epidemiological cartography is highlighted by recreating the results of Dr. John Snow's study of the 1854 Cholera epidemic in Soho, London.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Spatial databases and GIS*

General Terms

Algorithms,Design

Keywords

Knowledge Discovery, SAND Database System, Snow Cholera Map, Distance Semi-Join

1. INTRODUCTION

The recent introduction of web-based mapping services such as those provided by Google Maps and Microsoft Virtual Earth have greatly increased the public's awareness of the importance of location thereby making the spatial component a key part of the normal search experience. This can be seen by the increasing trend to incorporate spatial data into conventional databases. In particular, today, people are increasingly becoming used to going to the web to get answers to all of their queries that require some form of information lookup such as, for example, finding the nearest restaurant to their home that serves a particular cuisine as well as searching for homes, schools, and so on. In essence, what we are seeing is a trend for the map to take on the form of a normal information relation, using database terminology. In this paper we briefly illustrate how current mapping techniques can be used to revisit some old queries thereby yielding a form of knowledge discovery in epidemiological cartography which may be useful in other applications.

2. THE SAND SPATIAL BROWSER

Over the past years we have been engaged in the development of the SAND Spatial Browser [2], one of whose key features is the ability to produce its results incrementally where the increment is a form of a ranking by distance from a query object which can be measured as the crow flies or constrained to lie on a spatial network. Potentially, ranking has wide applicability for spatial queries using web-based mapping services where presumably there is a premium on obtaining partial results so that the most relevant ones are delivered first. In particular, in our example scenario, relevance is measured by spatial proximity.

The SAND Spatial Browser provides more power to users of databases by enabling them to define and explore the specific spatial region of interest graphically. The SAND Spatial Browser allows users to form either purely spatial or mixed spatial/nonspatial queries intuitively which can present information to users that might have been missed if only a textual interface was available. The SAND Spatial Browser is built on top of the SAND Spatial Database System which provides a server that facilitates organization (i.e., indexing) of spatial and nonspatial data to support efficient query processing. This database system handles any two or higher dimensional data with extent (e.g., country boundaries, river paths), as well as point data (e.g., city locations). It facilitates the response to queries involving this data such as finding the closest hazardous waste site to the border of a particular state. The SAND Spatial Database System is positioned somewhere between a conventional database management system (DBMS) and a Geographic Information System (GIS). It is similar in spirit to a GIS but does not have the full functionality of a GIS in the sense that its spatial analysis capabilities are limited.

Users access and manipulate spatial and nonspatial data using the SAND Spatial Browser in a manner similar to that used in spreadsheets where the map plays the same role as a relation in a relational database management system. In particular, operations can be specified as compositions of maps with the output of one or more operations serving as input to other operations which can be saved for use as input to future operations. In addition, in many applications there is no need for the operation to run to completion in order to obtain the desired results. Thus the incremental nature of the SAND Spatial Browser permits users to proceed in a pipelined fashion where the first results of an operation are fed as inputs to subsequent operations.

One of the SAND Spatial Browser's features is the *distance semi-join* operation [1], which yields what we call a *discrete Voronoi diagram*. It can be used to provide a clustering where the result is that given two data sets A and B, we can determine for each element a of A, the closest element b of B to a. For example, we have used it with a pair of data sets corresponding to locations of nuclear facilities and monitoring stations to discover which monitoring stations are the closest to each nuclear facility. Similarly,

*This work was supported in part by the US National Science Foundation under Grant EIA-00-91474 and CCF-05-15241, as well as the Office of Policy Development & Research of the Department of Housing and Development (HUD PD&R)

given a set of locations of distribution centers (e.g., warehouses) and a set of locations of stores, we can determine which stores should be served by which distribution centers. This is done by the simple addition of the capability to rank results by spatial distance to the database, and applying the ranking in such a way that the closest pairs of elements (a, b) are returned in order of increasing distance and elements from A are not permitted to be repeated. It enables database users to perform a variant of knowledge discovery as we show in the following epidemiological cartography example.

3. APPLICATION TO EPIDEMIOLOGICAL CARTOGRAPHY

In 1854 Dr. John Snow, a pioneer in epidemiological research, showed that the cholera outbreak in Soho, London was a result of the contamination of a water pump on Broad Street [3]. He used a *spot map* that showed the number of deaths in each household, which he then overlayed on a map of Soho as shown in Figure 1. Upon looking at the map, it is immediately clear that the water pump, labeled X on the map, is responsible for the cholera deaths in Soho as the majority of the deaths due to cholera are clustered around this pump. The result was revolutionary in many ways. First of all, it established water as the carrier medium of the cholera disease. This was counter to the widely, yet mistaken, belief at that time that "bad air" was the culprit responsible for the spread of the cholera disease. Secondly, it laid the foundation for the field of epidemiological research. Incidentally, maps and GIS technologies are widely being used even to this date in epidemiological studies.

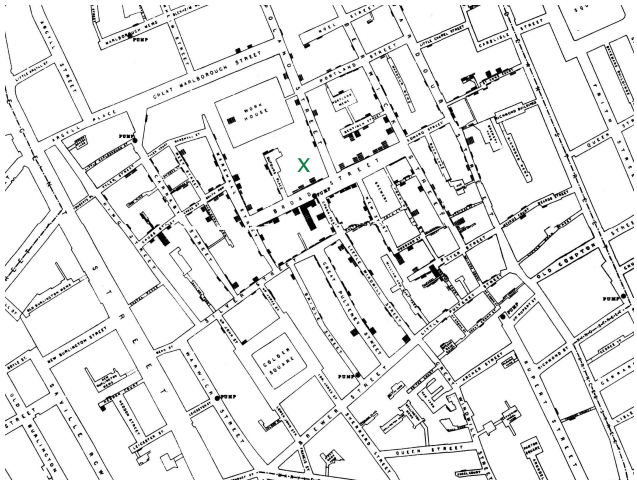

Figure 1: Result of overlaying the deaths caused by the 1854 London cholera outbreak on the map of the water pumps in London by Dr. John Snow.

In order to demonstrate the utility of some of the operations that we have developed in the course of our research on spatial spreadsheets and browsers, below, we recreate the 1854 experiment of Dr. John Snow using the SAND Internet Browser. We first created the road layer from a map image of Soho from around the same time period. Next, we overlayed the position of the water pumps on the map. Each death occurring in a Soho household at that time was recorded as a unique point. For example, if a household h recorded c deaths due to cholera, we placed c unique points at the position corresponding to the location of h on the map. The resulting setup is as shown in Figure 2. Dr. Snow used the map as a medium, looking at which an observer could deduce the strong clustering of deaths around pump X. Below we show how to make such a deduction using the SAND Internet Browser. We do this by computing the *distance semi-join* of the points in the *death* relation with the points in the *pump* relation. As we pointed out earlier, the distance semi-join operator uniquely associates each incidence of deaths with the nearest pump on the map. When this operator is executed to completion, the result is equivalent to a discrete Voronoi diagram on the points in the death relation. The thick polygonal subdivision in Figure 2 shows the result of drawing the discrete Voronoi diagram for this semi-join operation from which it is easy to see that the Voronoi cell that contains pump X as its Voronoi site, has the most incidents of deaths.

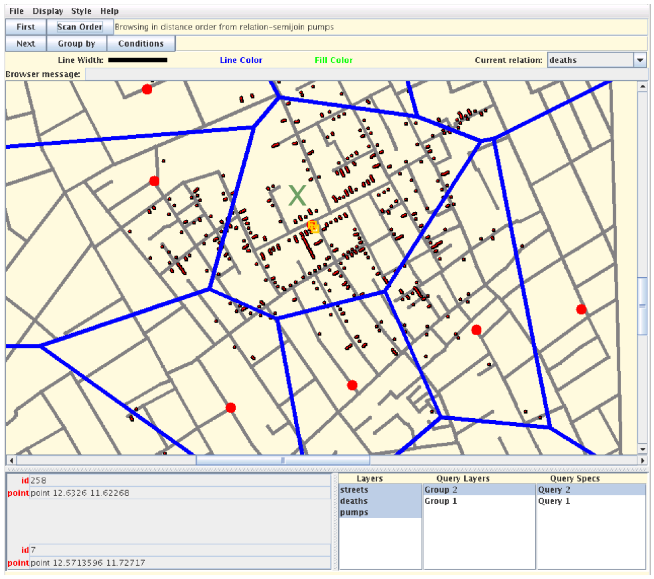

Figure 2: Discrete Voronoi diagram corresponding to the result of applying the distance semi-join of the death and pump relation using the SAND Spatial Browser.

4. CONCLUDING REMARKS

We have seen an example of the utility of applications such as the SAND Spatial Browser in knowledge discovery. We are confident that the increasing power afforded users to deploy spatial information will lead to an increased ability to address environmental causes of infection and disease such as Salmonella outbreaks, Legionnaires' disease, and so on.

5. REFERENCES

[1] G. R. Hjaltason and H. Samet. Incremental distance join algorithms for spatial databases. In L. Hass and A. Tiwary, editors, *Proceedings of the ACM SIGMOD Conference*, pages 237–248, Seattle, WA, June 1998.

[2] H. Samet, H. Alborzi, F. Brabec, C. Esperança, G. R. Hjaltason, F. Morgan, and E. Tanin. Use of the SAND spatial browser for digital government applications. *Communications of the ACM*, 46(1):63–66, Jan. 2003.

[3] J. Snow. *On the Mode of Communication of Cholera*. John Churchill, London, England, second edition, 1855.