DISCOVERING GEOGRAPHIC KNOWLEDGE IN DATA RICH ENVIRONMENTS: A REPORT ON A SPECIALIST MEETING

Harvey J. Miller
Department of Geography
University of Utah
260 S. Central Campus Dr. Rm 270
Salt Lake City, Utah USA 84112-9155
harvey.miller@geog.utah.edu

Jiawei Han
School of Computing Science
Simon Fraser University
Burnaby BC
Canada V5A 1S6
han@cs.sfu.ca

ABSTRACT

On 18-20 March 1999, a Specialist Meeting on "Discovering geographic knowledge in data-rich environments" was convened under the auspices of the Varenius Project of the National Center for Geographic Information and Analysis (NCGIA). This workshop brought together a diverse group of researchers and practitioners with interests in developing and applying new techniques for exploring large and diverse geographic datasets. The interaction prior to, during and after the three-day workshop resulted in the identification of research priorities and directions for continued development of "geographic knowledge discovery" (GKD) theory and techniques.

Keywords

Geographic data mining, spatio-temporal data mining, geographic information systems, geographic research.

1. INTRODUCTION

Similar to many scientific and applied research fields, geography has moved from a data-poor and computation-poor to a data-rich and computation-rich environment. The scope, coverage and volume of digital geographic datasets are growing rapidly due to new high-resolution satellite systems, initiatives such as the U.S. National Spatial Data Infrastructure and the automated collection of point-of-sale, logistic and behavioral data. In addition, the types of geographic data collected are expanding from the traditional vector and raster data models to include georeferenced multimedia data. This trend is likely to continue if not accelerate in the foreseeable future.

Geographers and others concerned with analyzing geospatial processes have a full set of spatial analytical techniques to bear on these problems. However, traditional spatial statistical and spatial analytical methods were developed in an era when data collection was expensive and computational power was weak. The increasing volume and diverse nature of digital geographic data easily overwhelm mainstream spatial analysis techniques that are oriented towards teasing scarce information from small and homogenous datasets. Traditional statistical methods, particularly spatial statistics, have high computational burdens. These techniques are confirmatory and require the researcher to have *a priori* hypotheses. Therefore, traditional spatial analytical techniques cannot easily discover new and unexpected patterns, trends and relationships that can be hidden deep within very large and diverse geographic datasets.

On 18-20 March 1999, a Specialist Meeting was convened in Kirkland, Washington, USA, on "Discovering geographic knowledge in data-rich environments." This meeting was sponsored by Project Varenius of the National Center for Geographic Information and Analysis (NCGIA) with additional support from Microsoft Corporation. The specialist meeting brought together a diverse group of researchers and practitioners with interests in developing and applying new techniques for exploring large and diverse geographic datasets. The interactions prior to, during and after the three-day workshop resulted in the identification of research priorities and directions for continued development of "geographic knowledge discovery" (GKD) theory and techniques.

2. WORKSHOP BACKGROUND

A key objective of the GKD workshop was to bring together the diverse communities with interests in GKD and geographic data mining. In the academic realm, these include:

i) geographers and other researchers interested in domain-oriented research questions;
ii) geographic information scientists formulating new digital geographic representations and geocomputational techniques;
iii) computer scientists formulating computational techniques for knowledge discovery from non-geographic and (increasingly) geographic databases; and,
iv) other scientists developing analytical techniques for data analysis (e.g., statisticians).

Also critical was bringing together academic researchers and scientists working in industry and other private sector settings.

A general Call for Participation was issued in December 1998. The response to this call was strong and the Steering Committee was forced to cull the submissions to a much smaller group of participants. Of the 30 final participants, 24 were from academic backgrounds, with an approximately 50-

50 split between geography and computer science/information science. One academic statistician also participated. The career stage of these participants ranged from graduate students through full Professors. The remaining six participants were from private sector companies, including Microsoft and Environmental Systems Research Institute, Inc. In brief, a wide-range of researchers and practitioners participated in the meeting.

A full list of participants, the final report in PDF format and other material related to the specialist meeting can be found at:

```
http://www.ncgia.ucsb.edu/varenius/discove
ring/description.html
```

3. RESEARCH DIRECTIONS IN GKD

The three-day specialist meeting consisted of a structured series of plenary presentations and panel discussions by leaders in their respective fields combined with a free-form set of breakout sessions and open discussions. The following transcendental, cross-cutting issues in GKD emerged from this interaction.

i) What are the new questions for geographic research? A fundamental question for the geographic and related research communities is "What questions do we want to answer that we could not answer previously?" These communities need to form well-structured (but possibly open-ended questions) in order to guide the computer science and related communities in their tool and algorithm development.

ii) Better spatio-temporal representations in GKD. Current GKD techniques use very simple representations of geographic objects and geographic relationships, for example, point objects and Euclidean distances. Other geographic objects (including lines, polygons and more complex objects) and geographic relationships (including non-Euclidean distances, direction, connectivity, attributed geographic space such as terrain and constrained interaction structures such as networks) should be captured by GKD techniques. Time needs to be more completely integrated into geographic representations and relationships. This includes a full range of conceptual, logical and physical models of spatio-temporal objects. Finally, we need to capture multiple representations (in particular, robust geographic concept hierarchies) and granularities in spatio-temporal representation in order to manage the complexity of GKD.

iii) GKD using richer geographic data types. Geographic datasets are rapidly moving beyond the well-structured vector and raster formats to include semi-structured and unstructured data, in particular, georeferenced multimedia. GKD techniques should be developed that can handle these heterogeneous datasets.

iv) User interfaces for GKD. GKD needs to move beyond the technically-oriented researcher to the broader geographic and related research communities. This requires interfaces and tools that can aid these diverse researchers in the GKD process. These interfaces and tools will likely be based on useful metaphors that can guide the search for geographic knowledge and make sense of discovered geographic knowledge.

v) Proof of concepts and benchmarking problems for GKD. There is a strong need for some "examples" or "test cases" to illustrate the usefulness of GKD. This includes a demonstration of GKD techniques leading to new, unexpected knowledge in key geographic research domains. Also important is benchmarking to determine the effects of varying data quality on discovering geographic knowledge. A related issue is research and demonstration projects that illustrate the usefulness of GKD techniques in forecasting and decision support for the public and private sectors.

vi) Building discovered geographic knowledge into GIS software. Current geographic information system (GIS) software uses simple representations of geographic knowledge. Discovered geographic knowledge should be integrated into GIS, possibly through inductive geographic databases or online analytical processing (OLAP)-based GIS interfaces.

vii) Developing and supporting geographic data warehouses. A glaring omission from current research in GKD techniques is the development and supporting infrastructure for *geographic data warehouses* (GDW). To date, a true GDW does not exist. This is alarming since data warehouses are central to the knowledge discovery process. Creating true GDWs requires solving issues in geographic and temporal data compatibility, including differences in semantics, referencing systems, geometry, accuracy and precision. Supporting GDWs may also require restructuring of transaction-oriented spatial databases systems

4. POST-MEETING ACTIVITIES

Post-meeting activities included the funding of seed grant proposals for new collaborative efforts emerging from the meeting and an edited volume of readings based on selected position papers. Participants also expressed an interest in a regular GKD conference and/or workshop.

The NCGIA, acting on the recommendations of the meeting's Steering Committee, funded two Seed Grant proposals resulting from the workshop. These projects are:

i) *Ontologies for spatial data mining and geographic knowledge discovery in large multimedia spatial datasets*, Jonathan Raper (City University London) and Myke Gluck, (Florida State University). This research addresses the deep ontological and epistemological aspects of discovering geographic knowledge. It also intends to design interfaces and build systems to support the application of these constructs to real world problems. The investigators will observe, catalog and organize user interactions with information retrieval systems. These processes will be translated into meta statements on geographic conjunctions and/or patterns which can then be converted into computable statements capable of being used to mine/retrieve data in spatial databases.

ii) *Multi-Scale Tools for Modeling Flows in Geographic Databases*, Eric Kolaczyk (Boston University). This project will conduct preliminary work adapting recently developed partition-based, multi-scale statistical models to the analysis of large flow datasets. It is anticipated that this initial investigation will proceed in two stages, the first involving model development and the second focusing on implementation and testing of the resulting methods. The investigator's recent work on hierarchical, partition-based multi-scale models will be extended to accommodate the unique nature and indexing of flow data, with its spatially complex "start-point/end-point" structure. These models will use factorizations of statistical likelihoods to effectively de-couple the information in the data at a given location across various scales. The second stage of this project will develop a prototype software implementation of the relevant modeling framework. The resulting tools will allow for adaptive response to user queries regarding flow data, in a manner that is sensitive to both spatial and scale variations.

As a follow-up to the meeting, the co-organizers are also editing a volume of readings based on expanded versions of selected position papers. This volume, entitled *Geographic Data Mining and Knowledge Discovery*, will be published by Taylor and Francis as part of its Research Monographs in Geographic Information Science series. Chapter themes include overviews of research issues and paradigms in geographic knowledge discovery, geographic data warehouses and repositories, techniques for geographic knowledge discovery and applications of geographic data mining and knowledge discovery. Publication is expected during 2000.

Finally, most of the participants at the specialist meeting perceived a strong need for a regular GKD conference and/or workshop. A key objective of this meeting should be to follow up on the fruitful cross-disciplinary interaction initiated during the specialist meeting. This will require participation by a similar, diverse group of researchers and practitioners.

ACKNOWLEDGMENTS

The efforts of many individuals were invaluable in making the "Discovering geographic knowledge in data-rich environments" workshop a success and pleasure. These include:

i) The National Center for Geographic Information and Analysis-Project Varenius (Michael Goodchild, Director) for funding the workshop. Funding for the Varenius project is provided by the National Science Foundation under NSF grant number SBR-9600465.

ii) Max Egenhofer (University of Maine) and LaNell Lucius (University of California – Santa Barbara) for their outstanding organization efforts.

iii) Steve Smyth (Microsoft), Charles Roche (Mobile GIS LTD) and Raina Smyth for excellent local arrangements (and a special thanks to Steve for use of his beautiful summer home on Lake Washington).

ABOUT THE AUTHORS

Harvey J. Miller is Associate Professor of Geography at the University of Utah where he also directs the geographic information science program. He received the Ph.D. in Geography from The Ohio State University in 1991. His research interests include geographic information science, transportation and computational regional science. A continuing theme in his work is using geocomputational methods to improve theory and decision support in transportation and locational analysis. He has recently been investigating the use of visualization and data mining techniques for exploring spatio-temporal data and the results from simulating complex transport systems. Dr. Miller is a recent recipient of the Hewings Award for Outstanding Young Scholar from the North American Regional Science Council. As of January 2000, he will be North American Editor of the *International Journal of Geographical Information Science.*

Jiawei Han is Professor in the School of Computing Science and Director of Database Systems Research Laboratory, Simon Fraser University, Canada. He obtained the Ph.D. in Computer Sciences at the University of Wisconsin - Madison in 1985. Prior to joining SFU, he was an assistant professor in Northwestern University from 1986 to 1987. He has conducted research in the areas of data mining (knowledge discovery in databases), data warehousing, spatial databases, multimedia databases, deductive and object-oriented databases, and logic programming. Dr. Han has published over 100 articles in journals and conference proceedings on these topics. He has served as a program committee member for over 30 international conferences and workshops and as the guest co-editor of special issues on data mining for several top-rated journals.