Position paper by Bin Jiang

Division of Geomatics, The Royal Institute of Technology Research School, University of Gävle, SE-801 76 Gävle, Sweden, Email: bin.jiang@hig.se, Web: http://fromto.hig.se/~bjg/

Data-Intensive Geospatial Analysis and Computation

Over the past a few years, we have been carrying out research projects using massive geographic information, some of which is volunteered, so called volunteered geographic information (VGI) (Goodchild 2007). The data scale is at the level of gigabyte. In this position paper, I present a short introduction to the projects, and then add a few reflections which I believe can contribute to the workshop's discussions along the line of data-intensive geospatial analysis and computation.

(1) A universal pattern of urban street networks (**GB data examined)**
Triggered by an obvious morphological difference between European (being organic or self-organized) US cities (being more planned in nature), I started a study trying to illustrate the difference in a quantitative manner. This is my original study motivation. However, instead of the difference, I ended up with a similarity – a universal pattern of urban street networks (Jiang 2007). This pattern can be described by the 80/20 principle: about 80% of streets are less connected (below an average, e.g. 4 or 5), while 20% of streets are highly connected (above the average). This discovery is out of my surprise. In the course of investigation, I chosen 10 largest cities, 10 smallest cities, and 20 middle cities from over 4000 US cities, and all bear this pattern or regularity. This pattern holds true even for the smallest US town Duffield in Virginia, according to the Guinness Book. This study goes beyond US cities and includes some cities elsewhere from Israel, China and Germany. This pattern still remains valid. All these data have been archived on the web for free access, http://fromto.hig.se/~bjg/PhyAData/. In the study, the total size of the data examined is up to a couple of GB.

(2) Hägerstrand project (observation and simulation of mobility patterns, **GB data investigated**)
Hägerstrand project is funded by the Swedish Research Council FORMAS, and it is aimed to explore GIS-based mobility information for sustainable urban planning. In one of the studies with the project, we collaborated with a local taxi company, and collected a massive GPS data (6 GB) about human movement patterns. We analyzed over 72 000 people's moving trajectories, obtained from 50 taxicabs during a six-month period in a large street network involving four cities. We illustrated that human mobility exhibits Lévy flight behavior (Jiang, Yin and Zhao 2009). We further implemented some agent-based simulations (with both random and purposive moving agents) which took days in a state-of-art personal computer to get one result, in total a couple of weeks' computer time to get the simulations done. The study ends up with a surprising finding that purposive or random moving behavior has little effect on overall mobility pattern (Jiang and Jia 2009). This finding implies that given a street network, the movement patterns generated by purposive human beings and by random walkers are the same. We are still bound by a non-disclosure agreement, for the taxi company is reluctant to release the part of the GPS data for research purposes. I do understand that because there are two other local taxi companies involved in the business. However, we have made the massive simulated data public available (http://fromto.hig.se/~bjg/PREData/).

(3) FromToMap project (http://fromtomap.org/wiki, **GB data involved**)
This project aims to develop intelligent route services using VGI. OpenStreetMap (OSM) is a major data source for the project. OSM is a wiki-like collaboration to create a free editable map of the world, using data from portable GPS devices, aerial photography and other free sources. We aim to provide shortest-distance-yet-with-fewest-turns route directions or instructions between two or multiple locations for both desktop (route planning) and mobile users (personal navigation). So far, we have developed a method to derive the shortest-distance-yet-with-fewest-turns routes, and implemented with the entire European OSM data (www.fromtomap.org). In this project, we have conducted a data intensive computation for deriving the shortest-distance-yet-with-fewest-turns routes. For Europe,

there are over 30 GB OSM data, from which we extracted 17 GB street data. For the purpose of computing routes, we generated a huge graph involving 10 millions nodes and 17 millions links, about 30 GB memory occupied. Currently, FromToMap is hosted in a HP server with a configuration as such: Memory 42GB, Hard disk 1TB, CPU E5540 2.53GHz (dual core), OS Windows Server 2008 SP2, 64-bit Operating System. According to our estimation, we would need a server with memory up to 200 GB in order for FromToMap to host the entire world data for the computation of routes on a graph of 50 millions nodes and 70 millions links for the entire world.

Based on the projects introduced, let me add a few reflections along the line of data-intensive geospatial analysis and computation. Some points may be arguable, and can be further discussed and debated during the workshop with other participants.

(1) Geospatial research should go beyond the data scale of kilobytes and megabytes. If raw data is gigabyte, there is no need anymore to sample the data. Geographic theory and knowledge should be based on some massive data for verification. Current computing storage and processing capacity are fairly powerful in dealing with the data scale.

(2) Do not assume space is normally distributed (homogeneity), and space with a massive large data scale is likely to bear enormous heterogeneity. For example, some physicists have extracted massive trajectories data of registered dollar notes from wheresgeorge.com to study human mobility patterns which bear this heterogeneity. In this regard, I believe that geographers or geospatial researchers should have unique contributions to the research using the cutting edge geographic information technologies.

(3) OSM can be a benchmark data for geospatial research. OSM data is very rich in content. It contains streets, pedestrian and cycling paths, as well as public transports. In addition, points of interest and land use information are also embedded in the data. More importantly, there is NO constraint to get access to it, since it is owned by no one. This point seems applicable to other formats of VGI.

(4) Following up the third point, geospatial community should setup a data repository to archive data, algorithms and source codes which can be shared among researchers. This way geospatial research is based on some common datasets, becoming replicable, extendible and reusable in terms of algorithms and codes. A similar job has been done in many other disciplines such as biology, physics, and computer sciences.

(5) Many analytics tools or geovisualization tools we have developed do not meet the challenge of data-intensive computing. And many tools developed are still targeted to verify a given hypothesis, and find some presumed relations. Current tools are still rather weak in discovering hidden knowledge. Also we seem fond of building up new tools, while tend to forget the ultimate goal is to find knowledge. Knowledge is power, not the tools!

References:

Goodchild M. F. (2007), Citizens as sensors: the world of volunteered geography, *GeoJournal*, 69(4), 211 - 221.

Jiang B. (2007), A topological pattern of urban street networks: universality and peculiarity, *Physica A: Statistical Mechanics and its Applications*, 384, 647 - 655.

Jiang B., Yin J. and Zhao S. (2009), Characterizing human mobility patterns in a large street network, *Physical Review E*, 80, 021136, Preprint, arXiv:0809.5001.

Jiang B. and Jia T. (2009, under review), Agent-based simulation of human movement shaped by the underlying street structure, Preprint, http://arxiv.org/abs/0910.3055.