ISSN 2186-7437

NII Shonan Meeting Report

No. 2016-6

Big Data: Challenges and Opportunities for Disaster Recovery

Takahiro Hara
Sanjay Madria
Cyrus Shahabi
Calton Pu

March 28–31, 2016

National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-Ku, Tokyo, Japan

Big Data: Challenges and Opportunities for Disaster Recovery

Organizers:
Takahiro Hara (Osaka University)
Sanjay Madria (Missouri University of Science and Technology)
Cyrus Shahabi (University of Southern California)
Calton Pu (Georgia Tech)

March 28–31, 2016

1 Introduction

1.1 Motivation

Tsunamis in Japan, earthquakes in Haiti, floods in Colorado, and many such other disasters pose a significant threat to people and a nation's critical infrastructures. To minimize human loss, and protect critical infrastructures more effectively, we need a well-coordinated emergency management and response system that involves situation-awareness, timely information dissemination, coordination to move people from affected areas, guiding the medical team to the most appropriate locations, and real-time planning and decision-making. This is possible due to large amount of Big data is constantly generated by millions of smart phones and heterogeneous sensors in various formats, granularity, dynamicity, and quality. Thus, the challenge is to overcome the problem of interoperability, and support for the extraction of knowledge from large-scale dynamic data which can be used. This will help in providing more and more services for city officials, utility services, and citizens, a necessary requirement for disaster management.

1.2 Goal

The main purpose of this NII Shonan meeting was to bring together researchers from the multidisciplinary fields of data management and analytics; mobiles, sensors and pervasive computing; geography and urban-planning; and disaster response and recovery with public agencies and commercial entities towards using big data for better decision-making and problem solving in the event of a disaster. To do so, we need to close the gaps between those who collect the data (data providers), those who could benefit from using the data (domain experts), and those who are capable of developing the methods for storing/managing/processing the data (technology enablers).

1.3 Big Data

So-called "Big Data" began when the Enterprise era generated the first wave of data through various software applications such as inventory management or human resource applications. Soon the field of computer science realized that there were commonalities in how the data was being stored and accessed, which led to the development of databases. As the size of data grew due to broad adoption by many enterprises (*Volume*), new research fields emerged to deal with efficient access (parallel databases), integration (data warehouses) and analysis (data mining) of large datasets. However, the second wave of data, Human-generated data (the Web), exposed the fundamental challenges resulting from data heterogeneity (*Variety*); this data is semi-structured (text documents) or non-structured (pictures and videos) and is growing at a much higher rate. The rapid growth of web applications left academics with little opportunity to identify commonalities of data usage, leading to many independent tools that focus on a narrow aspect of data preparation for a given application type and requiring human in-the-loop data extraction and preparation. This worked to some extent, as human data creation processes led to a natural gap between data generation and data consumption. Machine-generated data represents the newest wave as they are generated continuously at a high rate (*Velocity*) from various sensors in the physical world, starting with sensor instrumentation, e.g., pavement traffic loop detectors, SCADA industrial automation sensors, CCTV cameras, satellite- or plane-based LIDAR sensors, and continuing with inexpensive sensors in our mobile phones, refrigerators, watches, and soon, everything we wear. These three waves of data gave rise to numerous approaches benefiting from data use in critical decision-making (Big Data).

1.4 Challenges

The time is ripe to embark on a fundamental approach to Big Data challenges by assembling stakeholders to review case studies, design and develop several prototypical end-to-end systems, identify the commonalities, and develop lessons learned stories. This was exactly the goal of our proposed meeting with a focus application of disaster response and recovery. This is because efficient and thorough data collection and its timely analysis are critical to any disaster response and recovery system in order to save people's lives during disasters. However, access to comprehensive data in disaster areas and their quick analysis to transform the data to actionable knowledge are major data science challenges. Moreover, the effective presentation of the collected knowledge to human decision-makers is an open problem. Therefore, the proposed meeting is to study and share experiences in Big Data research, Education and Training as well as discuss challenges and disseminate solutions, blueprints, and prototypes focusing on the disaster recovery application domain.

1.5 Results

Twenty researchers attended the meeting. They came from various disciplines (e.g., Computer Science, Geography and Engineering) and organizations (e.g., academic, industry and government). During the four-day meeting, various research challenges and potential solutions were discussed. The summary of the

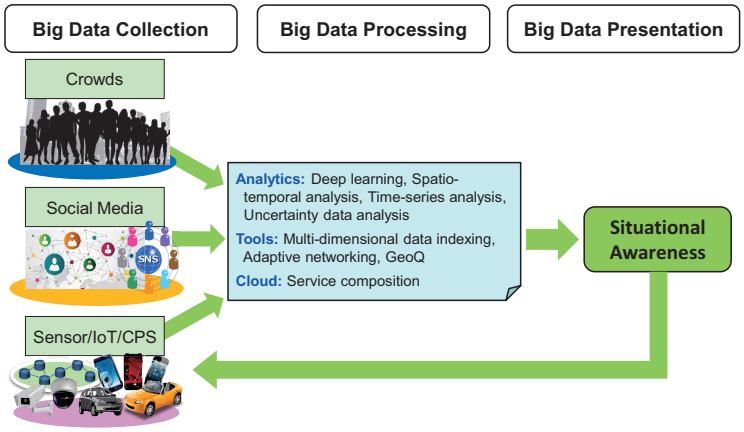

Figure 1: Summary of Discussed Topics

topics discussed is presented in Figure 1. Basically, the discussion topics can be grouped into three categories: Data Collection, Processing and Presentation.

Under data-collection, we discussed various sources of data available before, during and after disaster. The data can be obtained from all sorts of devices instrumented in the physical world, such as from sensors, Internet-of-Things and Cyber-Physical systems, and also from various kinds of social media such as Twitter and flickr. The data would then need to be stored and managed (discussed under "tools"), processed (e.g., in "cloud") and analyzed (under "analytics") using various techniques. Finally, the results should be made available to decision makers to become aware of the situation (i.e., situational awareness) and perhaps ask for more data to be collected, stored, processed and analyzed. In the remainder of this report, we discuss these topics in more details.

2 Overview of Talks

2.1 Big Data Collection based on Crowdsourcing

Crowd Sensing from Heterogeneous Sensors for Disaster Mitigation

Teruo Higashino, Osaka University, Japan

Recently, sensing technology and IoT (Internet of Things) have much attention for designing and developing affluent and smart social systems. Although huge sensing data are collected in cloud, generally cloud systems are facing poor scalability and difficulty of real-time feedback. Here, we focus on geospatial sensing data welled out continuously everywhere and consider how we can treat such huge sensing data. We introduce the notion of "Edge Computing" for this purpose, and discuss how we can apply this notion for designing scalable social systems.

We also introduce our recent research work about crowd sensing from heterogeneous sensors for disaster mitigation. In some areas, cameras cannot be used

for privacy problems. Smartphones have potentially powerful sensing abilities for crowd sensing Thus, first, we introduce Laser Range Scanner (LRS) based crowd sensing technique. We have realized the technique at Grand Front Osaka, which is a commercial complex near JR Osaka station, Japan.

We also introduce smartphone-based crowd sensing techniques using accelerometers, microphones and Bluetooth signals. Then, we apply our human mobility generation technique to create a realistic human mobility called Urban Pedestrian Flow (UPF) Mobility using sensing data from heterogeneous sensors. In order to create UPF mobility, we observe node densities at multiple observable points, enumerate pedestrians' moving routes for the target area such as routes from a station to a department store, and use a linear programming technique. Then, we can estimate the number of pedestrians for each route and create UPF mobility so that the observable errors become minimum. We create techniques to reproduce passages, add normal and emergency pedestrian flows and check efficiency of evacuation plans on 3D map so that local governments can make efficient evacuation plans (e.g., evacuation planning for fire and flooding of target underground malls). Depending on date/time and real-time situations of disaster victims, we provide suitable evacuation routes and disaster information to their smartphones.

Enabling the Crowd for Emergency Crisis Management

Stephen Jones, The MITRE Corporation, USA

On the afternoon of May 20, 2013, an EF5 tornado struck Moore, Oklahoma and adjacent areas with peak winds estimated at 210 mph (340 km/h), killing 24 people (plus one indirect fatality) and injuring 212 others. In order to coordinate an immediate response, analysts used a new application called GeoQ to sort through the open source video, pictures, social media, and overhead imagery that was available immediately after the event.

As an open source project sponsored by the National Geospatial Intelligence Agency (NGA), GeoQ's capabilities include visualizing disparate sources of data, incorporating a crowd-sourcing workflow for analysis, and providing a standards-based method of sharing this information. This presentation will describe how GeoQ provides these capabilities and lessons we learned using the tool during crisis events.

As described within the talk, GeoQ's main emphasis was the introduction of a common workflow for analysts and supervisors tasked to respond to a crisis. Data in the form of imagery, social media, and videos is often extensive during these situations which leads to analysts performing work on individual data, leading to an extensive amount of redundancy. GeoQ addresses this inefficiency by geospatially segmenting an impacted area into individual work cells. This work cell is drawn onto the user's map, with the size and shape scaled so the area covered can be analyzed in a few hours. This concept allows users to discover and overlay appropriate data into that area, building a situational picture from which they can draw appropriate conclusions. In discussions with analysts using GeoQ, it was reported that they saw a 20% increase in efficiency using this technique.

In addition to the geospatial segmentation, it was noted in discussions with analysts that there was no structure in tracking and completing work. Some

information was collected on spreadsheets, however most was done ad hoc, with little rigor in the process. GeoQ changes that approach by implementing a 6-step process (Unassigned, Assigned, In Work, Awaiting Review, In Review, Complete) that not only provided a framework for doing the work, but also provided a means for collecting metrics on the time to complete each step, providing valuable information to those managing the resources required to respond within a short time frame.

As seen through numerous sites online, valuable information can be found not only from commercially available imagery and systems, but also pictures and videos shared across the Internet. In fact, it has been noticed in our events that data mined from social media can provide a clearer picture than other sources, especially if weather conditions are severe. It has been said that every disaster event is a local event, so we want to be able to leverage the information from these sources. Naturally, the amount of information from these sources can be overwhelming, and often irrelevant to the task at hand. We don't offer a solution to this particular problem, however we want to maintain an open standards interface to allow us to integrate with systems that can mine important information from social media and allow GeoQ to display these results in conjunction with other data sources. Additionally, after analysts have created a common operating picture in near real-time, this information can be shared through a number of different API interfaces to other systems participating in the event. We have not designed GeoQ to be the required interface to any organization; instead our aim was to be a facilitator between a wide variety of data sources and emergency management systems already in place.

A significant advantage to structuring the workflow process when analyzing a crisis situation is that the process can be parallelized significantly. Within GeoQ, we combined this process with a web interface to allow a large number of users to work on a crisis event simultaneously. Within our environment, while we have tested with a relatively small number of users, we did receive positive feedback on this form of crowd-sourcing for doing analysis over a large area.

We believe a couple of factors will need to be considered when assembling a larger crowd to perform these tasks. Users will need an incentive in order to participate. Although it is likely that a number of people want to help during a crisis event, they may quickly become demotivated if the event does not impact them directly. We have experimented with the notion of gamification, awarding points and badges for those performing this work. We have seen positive effects, such as users competing with each other to accrue the most points, and we have also seen negative reactions to gamification as some analysts believe this makes the application 'childish'. However, a strong motivation for staying with gamification was found in rewarding users for achievable skills. By presenting users with badges that represent a skill gained, such as a Residential Damage Assessment Expert, they could not only be working towards a tangible goal, but it is something they can further advertise as a skill for other events requiring assistance. Likewise, groups organizing a crowd to provide help on an event of their own will have an indication of individuals qualified to perform critical tasks by assessing their users' skills. We have been experimenting with The Mozilla Foundation's Open Badges API as a way to provide this capability.

The decision to release GeoQ as an Open Source project had a number of motivations. First, we were looking to provide transparency for the work being done by our organization and allow outside organizations to freely use

the software and the techniques used by our analysts. The hope was that we would help to promote a community of crisis analysts, from local and state organizations through the national level. To this end, we have fielded GeoQ to some partner organizations during crisis event training exercises and have found a strong acceptance of the system.

Effectively Crowdsourcing the Acquisition and Analysis of Visual Data for Disaster Response

Seon Ho Kim, University of Southern California, USA

Efficient and thorough data collection and its timely analysis are critical for disaster response and recovery in order to save peoples lives during disasters. However, access to comprehensive data in disaster areas and their quick analysis to transform the data to actionable knowledge are challenging. With the popularity and pervasiveness of mobile devices, crowdsourcing data collection and analysis has emerged as an effective and scalable solution.

This talk introduces a new way of data collection (specifically mobile images and videos) with spatiotemporal metadata which can be used for collecting, organizing, sharing, and crowdsourcing data in disasters. On top of an existing mobile media management platform named MediaQ developed by USC IMSC (Integrated Media systems Center), this study addresses the problem of, and provides a solution for crowdsourcing mobile videos for disasters by identifying two unique challenges of 1) prioritizing visual- data collection and transmission under bandwidth scarcity caused by damaged communication networks and 2) analyzing the acquired data using a space decomposition technique which can evenly distribute human analyst's load.

MediaQ system has been described as the base for acquiring and analyzing the mobile videos utilizing fine granularity spatial metadata of videos including Field of View model in utilizing spatial metadata (i.e., camera location from GPS receiver and viewing direction from digital compass on a smartphone), novel queries such as directional query, and spatial crowdsourcing. An analytical model has been also explained to quantify the visual awareness of a video based on its metadata, specifically proposed the visual awareness maximization problem for acquiring the most relevant data under bandwidth constraints. So, recorded videos are ranked based on the amount of visual awareness and transmitted to server to maximize the visual awareness for a given network bandwidth. Then, the collected videos are evenly distributed to off-site analysts by space decomposition by kd-tree to collectively minimize their efforts for analysis. A cyclic data collection using urgency map (cycle of data crowdsourcing, human analysis, updated urgency map from analysis result, new crowdsourcing) can be applied during disaster time. Simulation results have been presented to demonstrate the effectiveness and feasibility of the proposed framework.

As a part of real application, USC IMSC's collection with US NGA's (National Geospatial-intelligence Agency) GeoQ has been presented. MeidaQ system is linked to GeoQ platform by a layer named MediaQ layer so that GeoQ users can access and search mobile videos collected by MediaQ system. Also, a new way of mobile video collection using drones has been introduced by NII, USC IMSC's Japanese partner in a joint international project between US NSF

and JST. A collaboration is going on between IMSC and NII to collect geo-tagged videos from multiple drones in disasters and utilize MediaQ for effectively archiving and efficiently searching them with geospatial metadata.

Privacy issue of the people in the collected images was questioned and discussed. We believe that people understand and sacrifice some privacy in disaster situation to help others. And at the same time a new privacy policy mechanism needs to be studied for such applications. Visual analysis of collected media can be considered to enhance situation awareness. And this approach is orthogonal to the presented spatiotemporal solution for big visual data.

2.2 Big Data Processing

2.2.1 Analytics

Spatio-Temporal Data Retrieval for Disaster Estimation

Hideki Hayashi, Hitachi, Ltd., Japan

Presentation:

Japan is located in the so-called "Circum-Pacific Mobile Belt," where seismic and volcanic activities occur constantly. Moreover, owing to its geographical, topographical, and meteorological conditions, Japan is subject to frequent natural hazards such as typhoons, torrential rains, and heavy snow. Consequently, such natural disasters can cause a great loss of life and significantly damage property. Accordingly, it is a national priority to protect citizens' lives, livelihoods, and property from large-scale natural disasters. When a large-scale natural disaster occurs, damage information is collected in the "first action" period. After that, disaster-relief operations and support over a wide area (depending on the scale of the disaster) are requested. At present, to collect damage information, municipalities, cities, districts, towns, and villages report damage information to prefectural governments, which then report to the national government by e-mail or fax. Disaster-related organizations need a few days to grasp the damage situation of a large-scale natural disaster because they collect the damage information manually. Thus, the time taken to grasp damage situations of large-scale natural disasters needs to be reduced.

At the time of a large-scale natural disaster, disaster-related organizations obtain insufficient information because the infrastructure such as the power networks and the communications network is damaged. Even so, they still have to estimate the damage situation concerning a large-scale natural disaster from that insufficient information. Disaster-related organizations especially need to quickly obtain the coverage of the natural disaster. So, they can plan where to dispatch rescue teams and disaster relief. Recently developed simulation technologies for simulating tsunamis, spreading of fires, heavy rain, and so on have made it possible to estimate the coverage of a natural disaster.

Here, after a large-scale natural disaster occurs, a natural disaster simulation is considered to be run to estimate the disaster's effect. An input condition is necessary to run the natural disaster simulation. However, there are some cases where the input condition is not collected after a large-scale natural disaster occurs. For example, a fire spread simulation is run for estimating the damage caused by fire after a large earthquake right under an urban area. Information on fire outbreak spots and directions and velocities of the wind is necessary as

the input condition, but the information on fire outbreaks spots is difficult to collect. Even if an input condition is collected, there are some cases where the processing time for simulating a natural disaster is long. For example, tsunami simulators take much time to simulate tsunami coverage because they calculate complicated tidal wave propagation.

To solve these problems, we proposed a fast spatio-temporal similarity search method that searches a database storing many scenarios of disaster simulation results represented by time-series grid data for some scenarios similar to insufficient observed data sent from sensors. The proposed method efficiently processes spatio-temporal intersection by using a spatio-temporal index to reduce the processing time for the spatio-temporal similarity search. The results of the performance evaluation showed that the proposed method achieves a shorter response time for the spatio-temporal similarity search than two conventional methods that use a temporal index and a spatial index.

Discussion:

Applications of spatio-temporal similarity search were discussed. As applications, we gave a disaster estimation system and a disaster warning system at the time when a large-scale disaster occurs. Especially, in a case where an urgent response of a disaster such as tsunami is necessary, these systems allow disaster-related organizations and residents to do quick actions. In such a situation, we recognized issues of improving estimation accuracy and translating of the probabilities into information that people can understand. Additionally, spatio-temporal indexing of spatio-temporal similarity search was discussed. The reduction of simulation result data and the the shortening of response time are important challenges.

Acknowledgement:

This work is supported by consignment research and development of techniques about use and application of real-time information in G-space platform from the Ministry of Internal Affairs and Communications, Japan.

Traffic Flow Analysis for Urban Railway Networks using Self-learnt Cellular Handoff Patterns

Hiroki Ishizuka, KDDI R&D Laboratories, Inc., Japan

Increasing traffic congestion at commute times has significantly degraded the efficiency of the transportation infrastructure and increased travel time, air pollution, and fuel consumption. As a result, urban areas are faced with the necessity of improving public transport to reduce car use. Urban areas (New York, London, Paris, Beijing, and Tokyo) have multimodal public transport systems consisting of railways and buses or light rail. In particular, railway services can carry massive amount of people at once are expected to be effective during commute times.

The Tokyo metropolitan area experienced a shaker in the upper 5 during the Great East Japan Earthquake. As a result, almost all railways temporarily stopped their service. Since the earthquake occurred on a weekday afternoon, a large number of commuting workers and students were in their offices and schools in Tokyo. The stoppage of railways made it impossible for such people to return home. During the Great East Japan Earthquake, there were 3.5

million as commuters stranded in the Tokyo metropolitan area. For the temporary evacuation of such stranded commuters, the government should estimate a number of potential candidates for the stranded commuters by analyzing railway commuters in daily life.

In order to roughly understand the behavior of railway commuters, we can use the results of a metropolitan transportation census provided by Japan's Ministry of Land, Infrastructure, Transport, and Tourism's (MLIT) and local governments. However, the census has been conducted every five years., e.g., the latest result was summarized in 2010. Because the environment of a transport infrastructure changes every year, understanding the behavior of railway commuters in detail is difficult using such old data. As another approach to find the behavior of railway commuters, we may use ticket sales from railway operators; however, railway operators have not provided such data to the public because of privacy issues.

On the other hand, mobile phones have become a key device for pervasive computing with users carrying them at almost all times. The ubiquity of these platforms has transformed mobile phones into one of the main surveyors of human behavior. However, conventional work using sensors on a mobile phone might not be suitable for large-scale, sustainable data collection in the analysis of human behavior. For example, participants (users) must install to analyze user activities. Moreover, the application wastes energy from the mobile phone battery with the unintended sensing operations. Though participatory sensing is one of the attractive solutions, collaborative users have been limited.

In order to cope with the problem, understanding human behaviors using call detail records (CDRs) that automatically recorded the location of the user and without the user's intentional operation have attracted attention. In fact, every time a subscriber makes or receives a phone call, a short message service (SMS), a multimedia messaging service (MMS), or data communication regarding the interaction and the location of the user (in the location of the base station used for the communication) is logged as CDR for billing purposes. CDRs have automatically and routinely been recorded without the user's intentional operation. Even if the user doesn't operate the mobile phone intentionally, some social networking applications on the mobile phone periodically are sending keep alive messages in several minutes. Although large-scale data are accumulated from CDRs through everyday use of mobile phones, the spatial and temporal resolution of CDRs is lower than that of existing positioning technologies.

To overcome the weak points of CDRs, some existing works have focused on the cellular handoff patterns as a substitute for the sparse location of CDRs. Characteristics of the handoff pattern when people get on the same rail line are likely to be similar. Thus, the handoff pattern of base stations can be useful data to analyze railway commuters. To manage a large numbers of rail lines, however, the conventional works are not suitable at the point of scalability, because the works have to learn the handoff patterns for each route manually. Moreover, the works don't adapt to be rebuilt the correct cellular handoff pattern including the modification when a new rail line or a new base station has been laid after once building the pattern.

To solve the issue, we propose the system that enables to maintain a large number of the cellular handoff patterns automatically. Our proposed system self-learns all cellular handoff patterns in urban complex railway networks using a huge amount of anonymous sparse location data of CDRs. In this talk, we

evaluate the three kinds of self-learning methods that include the Voronoi diagram based approach, the mahalanobis distance based approach and the static range based approach using CDRs from a millions of subscribers.

As the result, the accuracy of classifying the rail lines from the individual cellular handoff pattern achieved 81% against accurate 7600 GPS trajectories. In addition, the correlation about the traffic flow of urban railway networks between our output and the latest census data was 0.768. As the contribution of this talk, our proposed system indicates the possibility of understanding the traffic flow of the urban railway networks more often using only CDRs as substitute for the census.

A Deep Learning-Based Forecasting Tool for Big Data

Kuo-yi Lin, Asia University, Taiwan
Jeffrey J.P. Tsai, Asia University, Taiwan

The disaster response provides immediate assistance in several areas such as healthcare, operation and manufacturing to manage the occurring suddenly to avoid the loss of life and resources damage. Therefore, the forecasting tool should be developed to plan in advance to minimize the negative effects of a disaster and allow the organizations to recovery the critical functions. In the big data era, the real time collection and analysis of data help managers to search, define, evaluate and rescue resources and facilitates immediately. However, to forecast situation and support decision for disaster response in real setting are challenging.

We developed a deep learning-based forecasting tool to extract the relationship between the inputs and outputs factors via historical data for capturing disaster response insights. The convolutional neuron networks (CNN) and convolutional deep belief networks (CDBN) were proposed. We validated the proposed tool with realistic data collected from a leading medical information company, and perform scenario analysis to evaluate its performance under different conditions. The results have shown the practical viability of this tool to allow the decision makers to identify the solution among a set of non-dominated solutions within a reasonable time. Due to the high competitively among different types of fields in the artificial intelligence, these trends are forcing them to constantly improve their strategy to enhance the efficiency for disaster response. While this research are focused on the accuracy of the predicting model that was built based on data extracted from the environment data and life style data.

Indeed, disaster response decisions are a nonlinear assignment problem with the two objectives for minimizing the negative effects and minimizing the gap between management preferences including the allocation ratios among the resources. The conducted leading factors provide insights to adjust the manage settings based on the strategies. In particular, the environment data and life style data were collected. Deep learning utilized to execute significant analysis and produce reliable predictions to build forecasting model based on environment data as the prediction parameter. As a result of the information obtained by this research, recommendations concerning the management insight were made based on the domain knowledge.

Deep Learning played a fundamental role been crucial in the building of the prediction model through the artificial neural network based on environment

parameters to disaster forecasting. The information obtained through the research benefits the administrative policies bases, in aspects such as, allocations of resources and enhances the rescue efficiency. Although this tool is validated towards medical information industry, it is expected for it to be applicable on different area. The research sets the bases in an administrative medical information matter to continue developing different models for them to provide more accurate actual and future information that can enlighten the managers. Further research should be done to classify the factors regarding different scenario such as earthquake and typhoon via rolling prediction method to forecast further information.

Reliable Spatial and Spatio-Temporal Pattern Analysis to Support Decision Making in Disaster Management Applications

Matthias Renz, George Mason University, USA

Geo-spatial- related data has a tremendous impact in disaster recovery applications and spatial applications are critical in pre-, during, and in post-disaster management and response. An inherent property of any spatial and spatio-temporal dataset is uncertainty due to various sources of imprecision. In addition to the fact that measurements are naturally imprecise, uncertainty is also a consequence of insufficient sensor coverage and the quest for energy- and (network-) communication savings. All of these issues introduce the notion of uncertainty in the context of spatial and spatiotemporal data management - an aspect raising an imminent need for its scalable and flexible management.

In my talk, i discussed two ways of coping with uncertain data for data query processing and analysis aiming with focus on achieving reliable results: 1) data cleaning and 2) probabilistic data processing. Data cleaning is the most common approach. The idea is to estimate the true data values by applying data cleaning methods (e.g. expected value, most probable value, or other statistic methods) and taking the estimated ("non-uncertain") values as true data basis which allows to apply standard query processing and data analysis methods. Since the data cleaning approach goes hand in hand with information loss and missing information about the trustability of the query and data analysis results that are returned to the users, the probabilistic data processing is presented as better alternative. The idea of the probabilistic data processing approach is to incorporate the potentially available knowledge about the uncertainty of the input data into the entire query process yielding probabilistic query and data analysis results. The advantage of the probabilistic query processing strategy is that the user gets a more reliable query and analysis output in from of a set of possible true result alternatives associated with confidences that might help to make better decisions. On the other hand, probabilistic data processing is much more expensive and requires special techniques and methods that are scalable and meet the performance requirements of disaster management and recovery applications. Two generic approaches are introduced that addresses the aforementioned problems, (1) the paradigm of equivalent worlds and (2) the representative result oriented approach based on Monte Carlo sampling. Example applications are sketched for both approaches, including probabilistic

co-location mining that follows the paradigm of equivalent worlds and representative clustering as example of the second generic approach.

One important and critical issue that has been discussed most is that all approaches addressed in this talk make the assumption that the knowledge about the uncertainty, i.e. the uncertainty model, of the data is available. The derivation of the uncertainty model is a huge research field by itself. One possibility to get the information is to learn the model based on multiple observations (if available) or based on historic data, or using approaches from statistics, e.g. Kalman filter or Kriging, the latter is an approach used in geographics and makes estimations based on distance to the origin of a measurement. The question here is how accurate these models have to be to get sufficient reliable query and analysis results. Correlations between the (input) data quality and the corresponding result quality for specific query and data mining tasks is an open and quite interesting future research topic. Another question that has been discussed was the applicability of probabilistic query or data analysis results in the context of disaster recovery. I belief that the diversity of possible results is a very important parameter for decision making. However, the applicability highly depends on how the probabilistic results are represented, reported to the user, and how they are interpreted, which is also another huge research field for the future on itself.

Mining and Forecasting of Big Time-series Data

Yasushi Sakurai, Kumamoto University, Japan

The increasing volume of online, time-stamped activity represents a vital new opportunity for data scientists and analysts to measure the collective behavior of social, economic, and other important evolutions in the real world. Time-series data occur naturally in many online applications, and the logging rate has increased greatly with the progress made on hardware and storage technology. One big challenge for time-series data mining is to handle and analyze such large volumes of data (i.e., "big" time-series data) at a very high logging rate. Time-series data comes in various types of formats including co-evolving numerical sequences (e.g., IoT device data, video, audio), complex time-stamped events (e.g., web-click logs of the form ¡user-ID, URL, time¿), and time-evolving graph (e.g., social networks over time). Data variety imposes new requirements to data mining, therefore recent studies has revealed some new directions for research on time-series analysis, which include:

- Large-scale tensor analysis:

  Given huge collections of time-evolving online activities such as Google search queries, which consist of multiple attributes (e.g., keywords, locations, time), how can we analyze temporal patterns and relationships among all these activities and find location-specific trends? Time-evolving online activities and many other time-series data and can be modeled as tensors, and tensor analysis is an important data mining tool that has various applications including web-click logs for multiple users/URLs, IoT data streams, hyperlinks and social networks over time.

- Non-linear modeling:

Non-linear models are widely used in a variety of areas, such as epidemiology, biology, physics and economics, since the nature of real data sets suggests that non-linear models are appropriate for describing their dynamics. In the data mining field, analyses of social media and online user activities have attracted considerable interest, and recent studies have focused on non-linear time-series analysis to understand the dynamic behavior of social networks (e.g., information diffusion, influence propagation).

- Automatic mining:

  We also emphasize the importance of fully-automatic mining. Most of the existing time-series tools basically require parameter settings and fine tuning, such as the number of coefficients, and the reconstruction error thresholds, and they are very sensitive to these parameters. So, what if we have to handle very large datasets with millions, billions (or even trillions of time-series? In fact, faced with "big data", fully automatic mining is even more important: otherwise, the data scientists and analysts would have to try several parameter tuning steps, each of which would take too long (e.g., hours, or days). Namely, as regards real big data analysis, we cannot afford human intervention.

Recent work has revealed a huge demand for research on natural problems, including natural disasters. Since the environmental IoT data is insufficient to discover all important properties of natural issues. One of our future goals is to combine environmental IoT and social sensing data and examine the influence of natural phenomena (e.g., earthquakes in Kumamoto) to the society, along with the data of the number of patients/victims. We will integrate the above approaches (i.e., large-scale tensor analysis, non-linear modeling, automatic mining), which helps us to achieve this goal.

Estimation of People Movement from Mobile Phone Data using Data Assimilation Technology

Yoshihide Sekimoto, University of Tokyo, Japan

Recently, an understanding of mass movement in urban areas immediately after large disasters, such as the Great East Japan Earthquake (GEJE), has been needed. In particular, mobile phone data is available as time-varying data. However, much more detailed movement that is based on network flow instead of aggregated data is needed for appropriate rescue on a real-time basis. Hence, our research aims to estimate real-time human movement during large disasters from several kinds of mobile phone data. In this paper, we simulate the movement of people in the Tokyo metropolitan area in a large disaster situation and obtain several kinds of fragmentary movement observation data from mobile phones. Our approach is to use data assimilation techniques combining with simulation of population movement and observation data. The experimental results confirm that the improvement in accuracy depends on the observation data quality using sensitivity analysis and data processing speed to satisfy each condition for real-time estimation.

First, the system starts to collect mobile phone CDR data when it detects that a large-scale disaster has occurred. At the same time, it starts to collect

disaster information from mobile phones owned by some people. Fifteen minutes after the disaster occurs, it displays fragmentary disaster information collected from mobile phones. Thirty minutes after the disaster occurs, it displays the estimated movement of people and disaster status. After that, it updates the estimated results every 30 minutes. As a result, local governments can then examine the situation to ensure the safety and security of residents based on fragmentary disaster information 15 min after the disaster occurs. After 30 min, they start to set up their emergency headquarters and request support from neighboring prefectures based on the estimated results of the movement of people and the disaster situation. At the same time, the national government establishes the Headquarters for Major Disaster Management and consults with the local government to allocate resources such as relief supplies and rescue teams to the prefectures based on the estimated results.

Our approach utilizes two kinds of data obtained from mobile phones. For the first type, we utilize aggregated mesh-based population data from the CDR data of a mobile telecommunication company. These CDR data are automatically and routinely recorded without the user's intention. In fact, every time a user makes or receives a phone call, short message service (SMS), multimedia messaging service (MMS), or data communication regarding the interaction and location of the user (to locate a base station for communication), this data is logged as CDR for billing purposes. Even if the user does not operate the mobile phone intentionally, some applications on the mobile phone periodically send "keep-alive" messages every few minutes. Using these data, we are able to determine an individual user's location every several minutes using triangulation over the round trip delays of each base station to which the mobile phone of the user is connected. The provider of this data, as a mobile operator, has obtained a clear opt-in permission for the use of CDRs from each user for this research.

In this work, we generated an aggregated mesh-based population for privacy issues. The spatio-temporal resolution of each mesh is 500 m2 and 5 min When a large-scale disaster occurs, some base stations will immediately stop service because of damages caused by the disaster. Further, other base stations will gradually stop because of power supply loss within three to six hours after the disaster occurs. Therefore, the population data from the CDR will also decrease over time. We created the observation population data of our experiment to consider the shutdown rate of base stations using an actual base station spatial distribution according to the disaster level.

In this experiment, we showed that the sensitivity of the results depends on the observation data quality and quantity in the Tokyo metropolitan area. Moreover, the processing time was well within the 30 min assimilation step duration. However, some problems still remain. The accuracy of the estimated network-based population movement volume is still not sufficient, although we estimate population movement. Additional improvement will be needed using network-based volume observation data.

Mining Social Media to Support Disaster Management

Kyoungsook Kim, National Institute of Advanced Industrial Science and Technology, Japan

Social media data provide insight into people's opinions, thoughts, and reactions about real-world events such as hurricanes, infectious diseases, or urban crimes. In particular, the role of location-embedded social media (for short, geo-social media) for innovative disaster management systems is being emphasized to monitor surrounding situations and predict future effects by the geography of data shadows, as evidenced by experience from recent natural disasters such as the Tsunami and earthquake in Japan and Hurricanes in the USA (Sandy and Katrina). For example, tweets mentioning flooding from Oct. 29th to 30th, 2012 concentrated in the path of Hurricane Sandy and reflected experiences of the storm. Even though, there is high potential of geo-social media for disaster management, it brings big challenges to find meaningful information about dynamic social phenomena from the mountains of fragmented, noisy data flooding.

In this talk, I introduce Sophy framework to create a spatiotemporal knowledge in real time and RendezView system as an interactive visual data mining tool. People in disaster areas need to continuously receive disaster data that are related to local areas (such as flooded areas nearby and collapse of the nearest freeway). The spatiotemporal knowledge about proximity and relationships plays a basic role in localizing the information during the disaster. The Sophy framework constructs spatiotemporal relations among morphological features (i.e., geometric shape and size) of the distribution of geo-tagged Twitter messages (for short, geo-tweets) regardless of the identification of users in the real time. Then, it estimates basic topological relationships between features by using differential measurements in spatial, temporal, and semantic dimensions and stores the extracted features and relations into a graph-based database. Through the demonstration, I show how to discover an interesting pattern of topic flocks through our framework by using a typhoon case of HALONG in 2014.

Also I demonstrate another visualization tool, called RendezView, composed of a three-dimensional (3D) map, word cloud, and Sankey flow diagram. RendezView allows a user to interactively and intuitively discern spatiotemporal and semantic contexts of local social flock phenomena and their co-occurrence relationships. In the 3D map represented in the X-Y dimensions as a geospatial space and the Z dimension as a time line, cubes indicate local social flock phenomena (spatiotemporal clusters) and their color represent the measure of data concertation. When the user selects boxes in the map, the word cloud appears and displays a group of keywords as semantic contexts. Moreover, it represents the most common keyword co-occurrences through the Sankey diagram. I expect RendezView can be used by researchers and scientists to investigate social phenomena over a geographic region, such as information flow during disasters, patterns in work and hiring, or trends in political discourse. As an example, a social scientist can use the 3D map to compare the flock pattern of job postings between the east and west coasts in the US by searching on the keyword "work." The researcher then identifies keywords in the word cloud with a similar pattern, such as "apply" and "hire." The Sankey diagram informs the next keyword search based on strong spatiotemporal co-occurrence frequency with the

original keyword, and the user iterates with this process to thoroughly analyze nationwide work and hiring patterns.

2.2.2 Tools

Big Data Framework for Integrating Sensor Data and Information Extracted from Social Media

Takahiro Hara, Osaka University, Japan

Recent studies have revealed that messages posted on SNSs (Social Network Services) such as twitter and Facebook can be used for detecting various kinds of facts in the real world such as events, trends, and people's sentiment, which can be considered as kinds of social sensor data. Social sensor data are very useful for Big Data analysis because these tell many things representing the real world, which cannot be known by only analyzing traditional Big Data using general sensor data such as GPS (location), temperature, seismometer, and water-level gauge data. In particular, in disaster situations, real-time messages from ordinary people are quite useful for disaster situation awareness and decision making. However, in the current situations such as the cases of Hurricane Sandy and in Haiti earthquake, over 20M messages posted, only 100,000 messages were actually processed by the government agencies mainly by manual. On the other hand, while many research projects have proposed various techniques for extracting information from SNSs, most of them do not assume to reuse the output results for other applications as social sensor data, i.e., these are sole applications.

To tackle these problems and encourage reuse of SNS analytical results as social sensor data, we have just started a new project to develop a framework to share social sensor data, which can handle both (real) sensor data and social sensor data. In this position talk, we briefly present this on-going study. Our framework aims to share not only social sensor data (i.e., analytical result of SNS data) but also definitions of social sensor data and procedures to generate (analyze) them, which, we hope, are useful for application developers to reuse the existing definitions and procedures of social sensor data for developing new applications or generating new types of sensor data. To this end, we have developed a prototype where the social sensor definition, procedures to generate social sensor data, and database configuration are separately written in three different files, and the framework converts these files into a runtime program and executes it to generate social sensor data. Then, other users can search and download both the generated sensor data and its source files.

During the discussion in the meeting, some open issues were discussed including how to share SNS data stored by different groups/organizations because making stored SNS data is basically illegal in many SNS applications. The issues discussed also include the ownership problem for social sensor data generated from SNS data whose owner is not the one made analytics.

M-Grid : A Distributed Framework for Multidimensional Indexing and Querying

Sanjay Madria, Missouri University of Science and Technology, USA

The widespread use of mobile devices and the real time availability of user-location information is facilitating the development of new personalized, location-based applications and services (LBSs). Such applications require multi attribute query processing, scalability for supporting millions of users, real time querying capability and analyzing large volumes of data. Cloud computing aided a new generation of distributed databases commonly known as key-value stores. Key-value stores were designed to extract value from very large volumes of data while being highly available, fault-tolerant and scalable, hence providing much needed features to support LBSs. However, complex queries on multidimensional data cannot be processed efficiently as they do not provide means to access multiple attributes.

In this talk, I will present MGrid, a unifying indexing framework which enables key-value stores to support multidimensional queries. We organize a set of nodes in a modified P-Grid overlay network which provides fault-tolerance and efficient query processing. We use Hilbert Space Filling Curve based linearization technique which preserves the data locality to efficiently manage multidimensional data in a key-value store. We propose algorithms to dynamically process range and k nearest neighbor (kNN) queries on linearized values. This removes the overhead of maintaining a separate index table. The approach is completely independent from the underlying storage layer and can be implemented on any cloud infrastructure. Experiments on Amazon EC2 show that MGrid achieves a performance improvement of three orders of magnitude in comparison to MapReduce and four times to that of MDHBase scheme.

Research Questions and Comments:

Q. improve multiple queries or single query?
A: multiple queries, with load balancing.
Q: does Hilbert ordering preserve locality?
A: table per share and table per node preserve data locality, not Hilbert encoding.
Q: Do you handle updates?
A: update is an insertion.
Q: indexing updates may cause data movements.
A: we are not considering updates, but only insertions (append).
Addition comments: No concern for consistency problems.

Big-Data-Driven Research on Disaster Response and Management

Calton Pu, Georgia Tech, USA

Critical infrastructures from the backbone of modern civilization, and they can be damaged or stressed when disasters strike. As a representative example, transportation networks such as roads are often affected by earthquakes and floods, causing stress in traffic due to reduced transportation capacity. Similar situation happens with healthcare. When hurricane Katrina flooded part of

New Orleans, medical facilities and personnel became stressed and the quality of healthcare in some hospitals became sub-optimal. Our hypothesis is that better information (both from the past and on the present) will enable better scheduling of limited resources. A successful example in transportation that supports our hypothesis is the Waze mobile app, which uses real-time crowdsourcing to help drivers get around traffic congestion.

Although the actual tasks to repair physical infrastructures are engineering in nature, optimized use of limited resources will depend fundamentally on improved information on the events and the current state of infrastructures affected. We will integrate two complementary kinds of sensor information. The first kind of information come from physical sensors, which supply precise information on a narrow area (or electromagnetic band), but they are often limited by the small scale of sensor network deployment. The second kind of information come from social media, which provide a large amount of data on a wide range of topics, but their value is often limited by noise, where real information is mixed with misinformation and disinformation. The strengths and weaknesses of physical sensors and social media complement each other, providing more information on current events and up-to-date state of critical infrastructures.

Within the general areas of resilient infrastructures and big data, we will focus on the integration of heterogeneous big data and real-time analytics that will improve the adaptive management of resources when critical infrastructures are under stress. The integration of heterogeneous data sources is needed because we are working with many kinds of physical sensors and social media when trying to determine the status of infrastructures and demands under stress. Real-time analytics is needed when we apply the algorithms and software tools to manage the resources smartly and adaptively in a stressed environment.

We outline an illustrative example, the LITMUS project that demonstrates the feasibility and usefulness of integrating heterogeneous data sources for real-time big data analytics to collect and filter data on landslides, an example of small and medium disasters that do not have reliable physical sensor detection networks). Reliable and timely detection of disasters such as landslides is an important building block towards resilient infrastructures in smart cities. Good information on disasters can help focus the efforts to delimit the damages caused by disasters, and improve our knowledge of transportation networks, which in turn can lead to faster and more precise emergency response and recovery.

LITMUS is a landslide information service based on the integration of multiple, independent sensor information providing high quality information on landslides by integrating data from several physical sensors (rain and earthquakes, the two major causes of landslides) and social media such as Tweeter, Instagram, YouTube, and Facebook. LITMUS uses several filtering techniques to make social media data useful for landslide information, starting from simple and effective positive and negative keyword filters, geo-location information synthesis, machine learning algorithms, and reduced explicit semantic analysis (ESA).

The current version of LITMUS consists of a front-end implemented as a Web application and a back-end, which is the core of the system. The front-end is a live demonstration that runs on Apache web server. It uses Google Maps JavaScript API to render all feeds, including detected landslides, and PHP to access LITMUS' back-end. The back-end is developed in Python, except for binary classification for which we used Weka's library implemented in Java. All

data from social and physical sensors are currently stored and the social media data from the 4 sources have been collected since August 2013 and it takes up about 2GB every 6 months. The LITMUS software tools have been released on github.

Information services such as LITMUS provide the situational awareness information needed for adaptive resource scheduling during disasters and large scale events. Although the current implementation of LITMUS is focused primarily on the disaster aspects, we will start a parallel effort to collect and filter situational awareness information on critical infrastructures, by applying techniques similar to the outlined above for LITMUS. Achieving high quality real-time situational awareness information will enable effective adaptive resource allocation algorithms to manage limited infrastructures in an effective and resilient way.

Resilience and Dependability in Challenged Environments – A Middleware Approach

Nalini Venkatasubramanian, University of California, Irvine, USA

Advances in technology mobile computing, wireless communications, cyber-physical systems, Internet-of-Things (IoT) and cloud computing technologies are making available new modalities of information and new channels of communication. It has enabled the interconnection of objects and data to provide novel services and applications that will improve and enrich our lives. During large scale disasters and unexpected events, such technologies can be brought to bear to gain improved situational awareness and better decision support for response personnel and agencies. Mechanisms to integrate any and all available technologies and information in real-time to support reliable and timely sensemaking in resource challenged situations where the computing and communication infrastructure may be limited, unavailable and/or partially damaged (e.g. in human or manmade disasters). Dependability needs vary across applications; and that challenges and solutions exist at multiple system layers.

We discuss challenges in leveraging heterogeneous and emerging networking technologies to create a resilient and flexible infrastructure for both information collection and information dissemination. Middleware technologies play and important role in supporting cross-layer interactions amongst (a) the devices that collect and receive information, (b) multiple networks that communicate the content, (c) the platforms that process and store the information and (d) applications that use the information for diverse purposes. Drawing on recent IoT deployments, we will show how IoT and big data can drive semantic middleware to incorporate diverse sensors and inputs in a structured manner to generate situational awareness. The ability to combine novel technologies at multiple layers will open up new possibilities for resilient, adaptive and scalable societal scale systems of the future.

2.2.3 Cloud

Great Service: Benefits for Disasters Big Data?

Jianwei Yin, Zhejiang University, China
Yan Tang, Zhejiang University, China

Calton Pu, Georgia Tech, USA

Through the applications of Big Data software techniques and tools that can handle the "4V" challenges (Volume, Velocity, Variety and Veracity), many fields have reaped the benefits of Big Data, e.g., healthcare, manufacturing, and disasters preparation. One of the earliest and best known successes of Big Data was in service industry: through integrating data from different areas enabled Wal-mart to increase the inventory of lumber and other disaster preparation materials in Florida stores several days before a hurricane arrives. Customers benefit from higher quality of service enabled by Big Data. Service providers get higher profits from more precise control of costs and accurate knowledge of customer needs. We define the next generation high quality services as ***Great Services***, characterized by "4P" Quality-of-Service (QoS) dimensions: *Panorama*, *Penetration*, *Prediction* and *Personalization*.

Panorama QoS indicates such an essence that Great Services is built on much more comprehensive observations and descriptions towards objective substances. As a product atop Big Data, Great Services inherit the ability of gathering diversified data and sources from multi-domains and focus on synthetic and integrated views rather than isolated facts. Panorama QoS represents the integration of knowledge and facts from multi-domains. Panorama profiles Great Services from the dimension of horizontal space. As an orthogonal view to Panorama, **Penetration** stands for the evolution and innovation that services can achieve in the context of Big Data. Great Services with Penetration QoS achieves deeper and more accurate insights towards objective things, as well as innovative experience and services that do not exist before. Penetration profiles Great Services from the dimension of vertical space. **Prediction** is the most perceptual QoS of Great Services. Based on the rules and facts derived from massive volumes of historic information, service providers are capable of predicting the future and establishing more targeted services — knowing past well to predict the future better. Prediction profiles Great Services from the dimension of Time. **Personalization** is the embodiment of the diversity that Big Data introduces to services industry. Compared to general-purpose services, Great Services personalize specific business logic to specific customers, making different users feel that such services are designed exclusively for themselves. Personalization profiles Great Services from the dimension of User.

Nevertheless, challenges exist in the process of achieving and measuring "4P" QoS. Data heterogeneity is the biggest challenge for Panorama. The efficiency of processing various types of data from totally different backgrounds directly determines "how panoramic can Great Services be." Challenges for Penetration QoS is establishing an unified evaluation system and a set of specified standards to access the "Level of Integration" among services. Major challenges of evaluating Prediction QoS are the diversity among different services and the instability of some services, which makes it sophisticated to estimate how close to the optimal predictability can a service achieve. For Personalization QoS, the major focus is about how to avoid or alleviate the over-enlargement of single individual behavior, and how to control and detect positive and negative changes of behavior that go beyond historical prediction.

2.3 Data Presentation

Situational Awareness by Social Solutions: Textual Geo-Tagging and Photographic Crowd-Reporting

Asanobu Kitamoto, National Institute of Informatics, Japan

In the case of disasters, we need to know what is happening where and what time. This requires massive data collection schemes which cannot be achieved by traditional schemes such as hierarchical governmental information network or mass media with limited number of reporters. Our challenge is to establish an alternative data collection scheme, namely social solutions, such as social network and crowd-sourcing. Hence we propose two methods, textual geo-tagging and photographic crowd-reporting, to solve the problem of information communication and sharing during disasters using social solutions. The former is a solution for information overload, in which too much information should be divided into spatio-temporal units, while the latter is a solution for information scarcity, in which too little information should be enhanced by crowd reporting to fill spatio-temporal gaps.

First, to solve the problem of information overload, we developed software called GeoNLP, which aims at integrating geographic information processing (Geo) and natural language processing (NLP). GeoNLP has three components, namely GeoNLP data, GeoNLP software and GeoNLP application. GeoNLP data is the collections of place names (toponyms) with attributes that help GeoNLP software perform disambiguation. GeoNLP software deals with the task of extracting and disambiguating place names from natural language text. Finally GeoNLP application provides infrastructure to build geo-tagging application on Japanese natural language text such as news articles and social media like Twitter. One example of the application is social weather monitoring to improve situation awareness on weather during snow. We showed that the collection of tweets can estimate the boundary between snow and rain by counting the number of tweet mentions at each place.

For GeoNLP, the feasibility of using machine translation services was suggested. Geo-tagging software is normally designed for a specific language because linguistic difference much more significant than commonality across languages. A machine translation service may solve the problem because geo-tagging software can be designed for one language accepting translated text from other languages. This idea, however, has a number of problems. It may work on famous place names, but local names cannot be treated properly due to the lack of gazetteer in translated languages and the insufficient performance of machine translation services. Second, to solve the problem of information scarcity, we developed a mobile app called "Snowlog," which is based on the idea of "active viewfinder" to help users take photographs of the same scene in different times. The idea is to use a mobile phone as a fixed-point camera for everywhere to observe the change of scenes over time. Change detection is the fundamental task in situation awareness, because it differentiates abnormal from normal. In fact, the idea of fixed point observation was studied in several recent papers, such as time-lapse mining, and re-photography, but our app focuses more on the idea of spatial crowd sourcing, or crowd reporting that asks people report the situation using a smartphone camera with annotated text.

This app was applied to snow monitoring at Sapporo city, where snow monitoring and removal is a critical societal problem for all citizens. Our app was experimentally used by several Sapporo city citizens during the winter season of 2015, and collected time-series photographs of several points over several days. The result shows that our app is helpful to collect the situation of snow at multiple points. We also found, however, that the smart phone app was not easy to manipulate due to the complexity of finding the best matching under the six degree of freedom.

We explained a typical scenario to use this app during disasters. We first prepare the list of images as crowd reporting tasks, and ask people to take pictures from the same point with the same scene. This method is good for controlling the behavior of crowd workers because photographs can be chosen to avoid dangerous places for monitoring. It was also suggested that a requirement for taking the same scene may be overly restrictive. We emphasize, however, that this requirement may contribute to the gamification of the app, because a difficult task may be perceived as a challenging task to achieve the goal. This hypothesis was empirically proved by observing the behavior of children. Some of them were strongly attracted by the task and highly motivated so that they spent much more time than adults to find the same scene and improve the quality of matching. It was also suggested that the network of smartphone-based fixed point cameras may become more powerful by integrating it with other fixed point cameras, such as cameras on the road, or other network of sensors such as weather monitoring. Enhancement by other types of information is always promising, but this is left for future work, and our immediate goal is to improve the system of crowd reporting to assess the effectiveness of this idea during disasters.

A Situational Awareness Platform for Disaster Response based on the Supervision of Multiple Unmanned Aerial Vehicles

Helmut Prendinger, National Institute of Informatics, Japan

The ultimate goal of our work is to create a dynamic digital map that provides real-time or near real-time sharable information of all objects in some relevant geographical areas, such as large-scale event or disaster site. Our approach is to use multiple Unmanned Aerial Vehicles (UAVs) under the supervisory control of an offsite operator. This involves important challenges such as (1) the shared control of the UAVs by the operator and an algorithm that allocates the UAVs to waypoints, (2) estimation of the operator's attentional state, and (3) scene understanding by the UAVs using Deep Learning methods. As part of the JST-NSF project, we focus on the effectiveness of different interfaces for the offsite commander in Search and Rescue operations.

2.4 Visionary Talks

Perspectives on Smart and Connected Communities and Cyber-Physical Systems

Gurdip Singh, Kansas State Univ. and NSF, USA

Development of Smart and Connected Communities (SCC) will require novel approaches to design reliable and robust infrastructure systems. In addition, to provide resilient services, the interactions and interdependence of infrastructure systems in different domains (e.g., energy, transportation, and public health) must be addressed. This is also resulting in accumulation of large amounts of data, which can be analyzed, interpreted, and appropriately leveraged. When multiple systems are interacting with each other, and closed-loop control is implemented, real-time analysis of the large amount of cross-device data becomes a critical requirement. A number of programs at NSF such as the Cyber-Physical Systems program, the Critical Resilient Interdependent Infrastructure Systems and Processes program, and the Partnership for Innovation: Broadening Innovation Capacity program are supporting the development of technologies to support SCC and big-data analytics in real-time. In this presentation, we provide an overview of these programs, and focus on their multidisciplinary nature. We will discuss synergies between these programs, and provide perspectives on techniques for reliable and robust software infrastructure systems. Some of the key challenges faced in advancing resilience research include:

- ***Development of Test-beds:*** Conducting experiments to study interdependence of infrastructure systems require availability of test-beds that have two or more infrastructure systems, which are currently lacking. Cities as living labs to conduct such evaluation appears to be a promising solution.
- ***Availability of datasets:*** This is a challenge faced by researchers in academia. Datasets to drive experimental evaluations must be made available to academic researchers.
- ***Academia-Industry-Government partnerships:*** Partnership among representatives from these three sectors is important to advance research in smart cities, disaster response and infrastructure resilience. Each brings unique resources and perspectives that are essential for advancing research in the areas discussed above.

The Challenges and Opportunities Accompanying Geospatial Big Data

John P. Wilson, University of Southern California, USA

The traditional geospatial workflow incorporates five sets of tasks: (1) the collection and/or acquisition of data; (2) the preparation, reconciliation, and integration of data; (3) the completion of a series of spatial analysis and/or modeling tasks; (4) the interpretation of the results; and (5) the transformation of these results into "actionable" information. The volume, velocity, variety,

and veracity of "big" data provides new opportunities for knowledge discovery and modeling for disaster response and recovery so long as we can solve three challenges: (1) the need to work with populations in place of samples; (2) the need to work with messy as opposed to clean data; (3) the need to work with correlations as opposed to causality; and (4) the need to fuse the information gathered with these new methods and data sources with the theory and empirical knowledge of existing domains (Miller and Goodchild 2014).

The talk next used two examples to illustrate the production and use of "actionable" information. The first example described how spatial modeling has been used with a variety of inputs to delineate critical habitat areas for threatened & endangered species. The verification of the model outputs constitutes the greatest challenge with this kind of application and several examples (i.e. the Audubon Christmas Bird Count and the eBird Program) were used to illustrate how volunteered geographic information (VGI) might be used to gather ground-truth data and help with model calibration and/or verification. The second example described how spatial analysis was used to support Operation Smile, a volunteer organization that provides free medical services for children around the world with facial deformities. Geographic data were used with a series of spatial analysis tools to first identify the number of children by geographic unit in Vietnam and second, provide a series of travel itineraries to maximize the opportunities to locate and identify children in need in the field. There are numerous opportunities to add additional geospatial variables to refine these searches and provide better guidance in future campaigns.

The third and final part of the talk briefly described three disaster response and recovery applications that were focused on one or more of the preparation, response, recovery, and mitigation elements that comprise the disaster life cycle. The first consisted of an interactive seismicity and building response map of Los Angeles (Carlson et al. 2015). This map looks at the past, present, and future and includes specific building characteristics, instrumentally recorded and interpolated past ground motions, and earthquake scenarios. The second example recommends evacuation routes using a new and novel method to perform urban routing efficiently under capacity constraints (Shahabi and Wilson 2014). The third example collected spatial video following a tornado and used FEMA's damage assessment protocols to assess damage to individual residences and document the speed, accuracy and reliability of the damage estimates provided from this new data source (Lue et al. 2014). These three examples, taken as a whole, were used to show the kinds of work that will be needed to take advantage of the new data sources discussed in this meeting.

References

[1] Carlson, A., Moffett, B., Longcore, T., McPherson, K. 2015. Interactive Seismic Map of Los Angeles. Paper presented at the Second ATC and SEI Conference on Improving the Seismic Performance of Existing Buildings, San Francisco, California.

[2] Lue, E., Wilson, J.P., Curtis, A. 2014. Conducting disaster damage assessments with spatial video, experts and the inexperienced public. Applied Geography, 52, 46-54.

[3] Miller, H.J., Goodchild, M.F. 2014. Data-driven geography. GeoJournal, 80, 449-461.

[4] Shahabi, K., Wilson, J.P. 2014. CASPER: Intelligent capacity-aware urban routing. Computers, Environment and Urban Systems, 46, 12-24.

3 Open Issues

During the closing session, the participants have discussed various open issues which are not covered by the topics presented by the speakers. Some typical and significant issues are summarized as follows.

- How to get data at disaster sites

  In many disaster situations, it is difficult to get enough and right data for big data analysis. Crowdsourcing and Crowdsensing can be promising approaches to this end. However, more efforts are needed from both technical and practical perspectives to use such approaches in real disaster situations.

- How to get right samples

  Even if enough amount of data are available for big data analysis during disaster situations, the data generally include many messy (noisy) data or that are not suitable for the target analysis. So, we need some techniques to filter out unnecessary data (samples) and get only right samples.

- Big data directory

  Recently, many open data repositories which are mainly provided by governments and public organizations are available. However, it is not easy for big data application developers to fully make use of them because (i) it is difficult to horizontally search these repositories to find data of interest, and (ii) most of open data are provided in a format (e.g., PDF) that cannot be easily processed by machines for big data analysis. Therefore, we need a new big data directory to easily find data of interest from a big volume of available open data.

- Privacy preserving data analysis

  It has been known that someone's private information may be revealed by analyzing a big volume of data. Therefore, preserving privacy has been considered as one of the most significant issues in big data analysis. Here, since various data come from different data sources in disaster situations, it becomes more difficult to conduct data analysis while preserving privacy.

- How to integrate functionalities developed by different organizations and groups

  While different organizations and groups have been developing various analytical techniques, systems and tools for disaster big data management, these should be integrated into a single system to fully make use of them. In doing so, there must be many new technical challenges.

- How to integrate simulations and big data

  In disaster situations, it often happens that some data necessary for analysis are missing or not available. Data interpolation helps in such a case. To this end, simulations based on models developed by using historical data are useful. Therefore, some mechanisms for integrating simulations and big data are needed.

Participants

- Takahiro Hara, Osaka University, Japan
- Hideki Hayashi, Hitachi, Ltd., Japan
- Teruo Higashino, Osaka University, Japan
- Hiroki Ishizuka, KDDI R&D Labora-tories, Inc., Japan
- Stephen Jones, The MITRE Corporation, USA
- Kyoungsook Kim, National Institute of Advanced Industrial Science and Technology, Japan
- Seon Ho Kim, University of Southern California, USA
- Asanobu Kitamoto, National Institute of Informatics, Japan
- Kuo-yi Lin, Asia University, Taiwan
- Sanjay Madria, Missouri University of Science and Technology, USA
- Helmut Prendinger, National Institute of Informatics, Japan
- Calton Pu, Georgia Tech, USA
- Matthias Renz, George Mason University, USA
- Yasushi Sakurai, Kumamoto University, Japan
- Yoshihide Sekimoto, University of Tokyo, Japan
- Cyrus Shahabi, University of Southern California, USA
- Gurdip Singh, Kansas State University and NSF, USA
- Nalini Venkatasubramanian, University of California, Irvine, USA
- John P. Wilson, University of Southern California, USA
- Jianwei Yin, Zhejiang University, China

Meeting Schedule

Check-in Day: March 27 (Sun)

- Welcome Banquet

Day1: March 28 (Mon)

- Talks and Discussions

  Session 1: Tools and Frameworks for Big Data Analysis
  Chair: Sanjay Madria, Missouri University of Science and Technology, USA

  - Kuo-yi Lin, Asia University, Taiwan, A Deep Learning-Based Forecasting Tool for Big Data
  - Matthias Renz, George Mason University, USA, Reliable Spatial and Spatio-Temporal Pattern Analysis to Support Decision Making in Disaster Management Applications
  - Hideki Hayashi, Hitachi, Ltd., Japan, Spatio-Temporal Data Retrieval for Disaster Estimation
  - Yasushi Sakurai, Kumamoto University, Japan, Mining and Forecasting of Big Time-series Data
  - Sanjay Madria, Missouri University of Science and Technology, USA, M-Grid : A Distributed Framework for Multidimensional Indexing and Querying

  Session 2: Social Media Analysis for Disaster Management
  Chair: Takahiro Hara, Osaka University, Japan

  - Takahiro Hara, Osaka University, Japan, Big Data Framework for Integrating Sensor Data and Information Extracted from Social Media
  - Calton Pu, Georgia Tech, USA, Big-Data-Driven Research on Disaster Response and Management
  - Kyoungsook Kim, National Institute of Advanced Industrial Science and Technology, Japan, Mining Social Media to Support Disaster Management

Day2: March 29 (Tue)

- Talks and Discussions

  Session 3: Disaster Data Collection Using Crowdsourcing
  Chair: Cyrus Shahabi, University of Southern California, USA

  - Stephen Jones, The MITRE Corporation, USA, Enabling the Crowd for Emergency Crisis Management
  - Seon Ho Kim, University of Southern California, USA, Effectively Crowdsourcing the Acquisition and Analysis of Visual Data for Disaster Response
  - Asanobu Kitamoto, National Institute of Informatics, Japan, Situational Awareness by Social Solutions: Textual Geo-Tagging and Photographic Crowd-Reporting

– Nalini Venkatasubramanian, University of California, Irvine, USA, Resilience and Dependability in Challenged Environments – A Middleware Approach

Session 4: Big Data Analysis and Platform for Situational Awareness
Chair: Takahiro Hara, Osaka University, Japan

– Yoshihide Sekimoto, University of Tokyo, Japan, Estimation of People Movement from Mobile Phone Data using Data Assimilation Technology
– Hiroki Ishizuka, KDDI R&D Laboratories, Inc., Japan, Traffic Flow Analysis for Urban Railway Networks using Self-learnt Cellular Handoff Patterns
– Helmut Prendinger, National Institute of Informatics, Japan, A Situational Awareness Platform for Disaster Response based on the Supervision of Multiple Unmanned Aerial Vehicles

Group Photo Shooting

Day3: March 30 (Wed)

- Talks and Discussions

Session 5: Challenges on Big Data Research for Disaster Management
Chair: Calton Pu (Georgia Tech, USA)

– John P. Wilson, University of Southern California, USA, The Challenges and Opportunities Accompanying Geospatial Big Data
– Gurdip Singh, Kansas State Univ. and NSF, USA, Perspectives on Smart and Connected Communities and Cyber-Physical Systems
– Jianwei Yin, Zhejiang University, China, Great Service: Benefits for Disasters Big Data?
– Teruo Higashino, Osaka University, Japan, Crowd Sensing from Heterogeneous Sensors for Disaster Mitigation

Excursion and Main Banquet

Day4: March 31 (Thu)

- Discussions

Closing Session
Wrap up