The Challenge of Geospatial Big Data Analysis

Authors

- **Teerayut Horanont**, University of Tokyo, *Japan*
- **Apichon Witayangkurn**, University of Tokyo, *Japan*
- **Shibasaki Ryosuke**, University of Tokyo, *Japan*

KEYWORDS : Pervasive geosensing, Spatial-temporal data, Big data analytics, Open source solutions, Cloud computing

Abstract - This paper discusses a new potential use of massive call phone location data for analyzing the urban system. Increasingly, large amount of spatial temporal information is becoming available with the help of technologies such as satellite-based positioning technologies, sensor networks and the penetration of the mobile phones. The mobile GPS data collected during year 2010-2011 and the special event of 311 Great Japan Earthquake is used to demonstrate the use case. These data are concerning as "Geospatial Big Data" where commercial or proprietary GIS software could not offer a suitable support for this particular large data set. We developed a prototype platform using all open source solutions as new approaches to handling, processing and analysis of the data. The results show that open source software and libraries are capable and are promising solution to cope with very large-scale geospatial data challenge.

1. Introduction

Communication network enables cities to gather more high-quality human mobility data in a timely fashion than ever before. Thinking in term of "scales of networks", mobile communication provides us an ideal solution to create a huge urban fabric in which urban populations simply become part of a network. These networks can be determined as telematic, physical, or even social interaction when people are all engaged. This research underscores the critical need and the opportunistic of these territories as well as the way to construct a new approach to manage of very large-scale spatial-temporal data by using open-source solutions.

In principle, telecommunication infrastructure may be seen as providing a service to people. This means people and infrastructure interact, making new interfaces and ways of representing the existence of any entity that can be seen by the network. The footprints from this interaction clearly become a new source to unambiguously identify people in the real world and that of create "space-time travel" data. Another reality is that this new spatial temporal data generated from telecommunication domain are rapidly increasing at a speed surpassing the capacities of ordinary computer's storage and computing capacity. How to handle such a large-scale geospatial dataset? This is a critical issue especially in spatial analysis domain. The ability to gain speed in data processing, data mining and data analytical support is a big concern of today.

In this study, a new type of mobile-GPS data has been collected for 1-year period from August 1, 2010 to July 31, 2011. Approximately 9.2 billion of GPS records from 1.56 million registered users in Japan are input into our system for analysis. The mobile-GPS is a service provided by a leading mobile phone operator in Japan. Technically, the mobile-GPS enabled handset position is measured every 5 minutes, the information is then sent through the network to perform specific analysis and provide services for the registered users. This time series GPS data is completely anonymous and could not be identified the individual.

2. Reviews

Cloud computing platform is suggested for processing such large data in range of terabytes to petabytes with dynamically scalable and virtualized resources [1,2]. Hadoop is an open source large-scale distributed data processing that are mainly designed to work on commodity hardware [3] meaning it does not require high performance server-type hardware. Hive is a data warehouse running on top of Hadoop to serve data analysis and data query by providing SQL-like language called HiveQL [4,5]. Hive allows users familiar with SQL language to easily understand and able to query data.

Our previous study, Witayangkurn et al [6] provided a comparative assessment of very large scale spatial data processing by comparing performance of three techniques specific for mobile dataset which are PostgreSQL with PostGIS, Java application with Java Topology Suite (JTS), and Cloud computing platform using Hadoop. In this implementation, the Hadoop platform with customized spatial function on Hive, an SQL-like language for Hadoop, give a drastically increase of performance boost from the traditional spatial database.

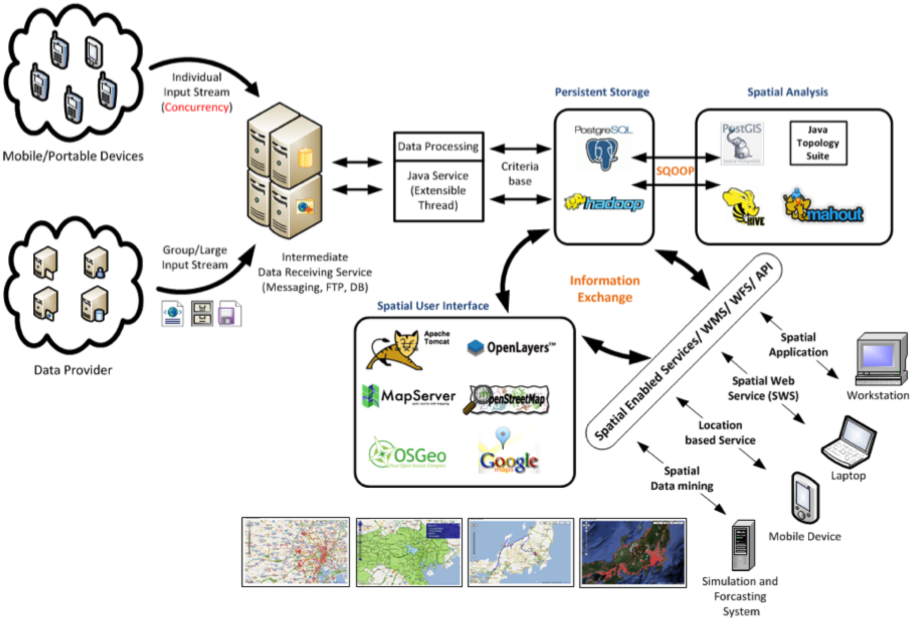
3. Implementation and System Overview

The GPS trajectory data of about 9.2 billion records and 600GB in size had been prepared to perform the analysis. In this research, the Hadoop Distributed File System (HDFS) and Hive is used to store and process the data. Generally, Hadoop/Hive did not support spatial query. However, Hive allows developers to create User-Defined Function (UDF) that could be any function based on the user requirement. We developed a custom function on top of Hive to manipulate and process the spatial temporal data. Figure 1 illustrates the overall system and basic idea in creating open source geospatial platform that do support the big data analysis and visualization. We have experience in conducting a system, which utilizes only spatial database (PostgreSQL/PostGIS) Horanont et al [7]. The previous system has several limitations and mainly from the constraint of high processing cost and calculation time regarding size of the data set. The proposed cloud computing technique dramatically increase the performance of spatial query [6] and therefore promising approach to address a new way of very large-scale geospatial processing.

4. Applications and Discussions

The March 11 Great Japan Earthquake event in 2011 was taken as our first analysis and visualization using the proposed platform. After a 9.0

FIGURE 1

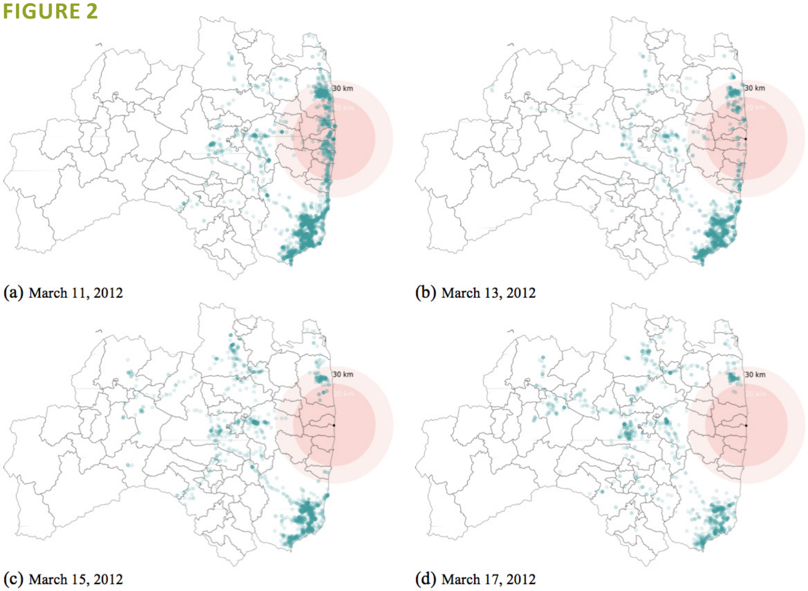


magnitude earthquake struck the coast of Japan, 46 minutes later a massive tsunami flattened the Fukushima Daiichi nuclear power plant (DNPP). This has resulted in large-scale radiation leaks and eventually forced everyone living within 20 km to evacuate their homes. By the early time of evacuation period, it was not even known where most of the evacuees were. The local government in Fukushima said they didn't know where 40 percent of the residents around the Fukushima DNPP went [8]. We utilized this platform to demonstrate a near real-time monitoring of people movement in disaster areas since it is crucially important for developing an evacuation plan.

To demonstrate this scenario, the past 6 months of mobile-GPS data were used to calculate and find the numbers of people who live within the restricted area of Fukushima DNPP. The most often visited places during 0-6 am are determined as their home locations. By giving the same assumption, the data after March 11 were computed to estimate their evacuation places after the nuclear accident. Theoretically, monitoring of the evacuation activity can be performed in near real time or in daily basis. Figure 2 (a, b, c, d) explain daily evacuation situation after March 11. Figure 3 shows the area where the evacuees moved during the first month. The line thickness connected between two cities defines the evacuation ratio. This evacuation ratio is an estimate of the fraction of the population exiting from a particular city within the restricted area to the city where they seek for shelter or temporary housing, which is defined in black dot on the map. Please note that more than 200 cities were detected as temporary places for the refugees who live within the restricted area although only the cities where evacuation ratio is greater than 5 percent are plotted on the map.

This initial study demonstrates the first use of mobile-GPS as pervasive emergency information retrieval tool and the results disclose part of human movement during the crisis. Managing and Mining of large-scale participatory data of mobile systems are crucial importance for the emergency management, particularly in the collection of real time data to provide accurate and complete picture of the current situation to the decision makers.

FIGURE 2



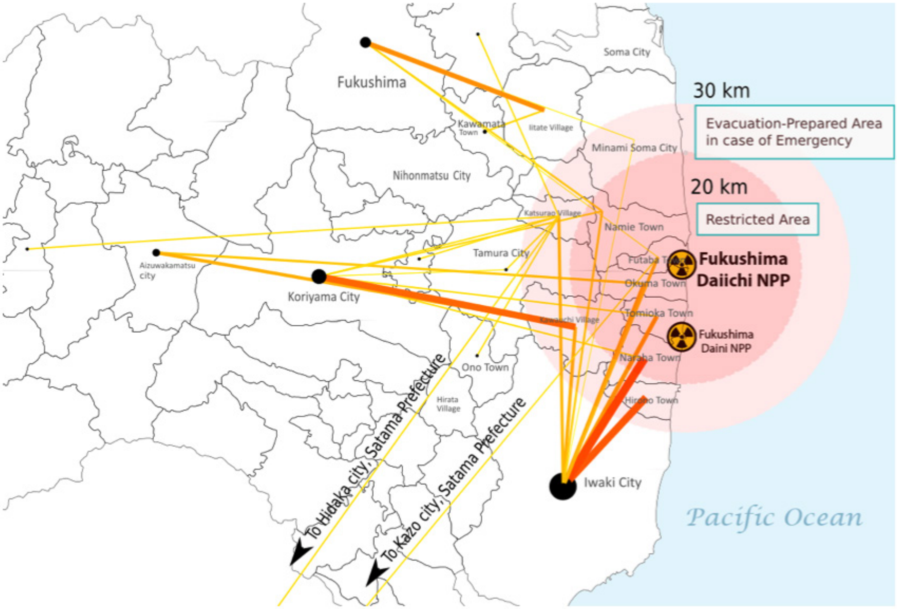
(a) March 11, 2012

(b) March 13, 2012

(c) March 15, 2012

(d) March 17, 2012

FIGURE 3



[1] YANG, J., AND WU, S. Studies on Application of Cloud Computing Techniques in GIS. In IITA-GRS 2010: *Proceeding of the IEEE Second IITA International Conference on Geoscience and Remote Sensing* (China, 2010), pp. 492-495.

[2] BUYYA, R., et al. 2008. Cloud Computing and Emerging IT Platforms: Vision, Hype, and Reality for Delivering Computing as the 5th Utility. *Future Generation Computer Systems archive 25* (2009), 599-616

[3] Hadoop Project. Retrieved March 11, 2012 from http://hadoop.apache.org/

[4] Hive Project. Retrieved March 11, 2012 from http://hive.apache.org/

[5] THUSOO, A., et al. Hive - A Petabyte Scale Data Warehouse using Hadoop. In ICDE 2010: *Proceedings of the 26th International Conference on Data Engineering* (California, CA, USA, 2010), IEEE Press, pp. 996-1005.

[6] WITAYANGKURN, A., HORANONT, T., SHIBASAKI, R. Performance Comparison of Spatial Data Processing Techniques for Large Scale Mobile Phone Dataset. In COM.Geo 2012: *Proceedings of the 3rd International Conference on Computing for Geospatial Research and Applications* (New York, NY, USA, 2012), ACM Press.

[7] HORANONT, T., AND SHIBASAKI, R. An Implementation of Mobile Sensing For Large-Scale Urban Monitoring. In UrbanSense'08: *Proceeding of the International Workshop on Urban, Community, and Social Applications of Networked Sensing Systems* (North Carolina, NC, USA, 2008).

[8] HAYS, J. 2012. Fukushima Evacuees. Retrieved March 11, 2012 from http://factsanddetails.com/japan.php?itemid=2288&catid=26&subcatid=162