論文の内容の要旨

論文題目　A STUDY ON HUMAN ACTIVITY ANALYSIS WITH LARGE SCALE GPS DATA OF MOBILE PHONE USING CLOUD COMPUTING PLATFORM

（クラウドコンピューティング環境を用いた大量のGPS携帯電話ログによる人々の活動分析に関する研究）

氏　　名　ウィタヤンクーン　アピチョン

Understanding urban mobility patterns are important aspects to explicitly express a current situation and allow an improvement in facilities and infrastructure to support people for a better life. Recently, mobile phones have become very common and used by large numbers of people in the worldwide. Moreover, with the advancement of mobile phone technology, mobile phones are mostly embedded with GPS and Wi-Fi module which could be used for identifying current location of people. Such dynamic positioning-data have gained a lot of benefits such as a location-based service application and navigation. In the viewpoint of mobility analysis, GPS-based trajectory data from mobile phone reflects the real movement of people which are much better than traditional survey data. It is also possible to increase a number of participants from small scale to large scale with less effort than survey. However, mobility analysis has a lot of challenges and issues especially when processing on specific dataset like GPS trajectories from mobile phones. Therefore, in this research, three main issues have been focused and taken into account. It included the issues on very large scale data, a time consuming aspect of spatial data processing and the scalability of mobility analysis techniques. The primary dataset used for testing and evaluation are large-scale GPS trajectory dataset collected from 1.5 million individual mobile phone users in Japan accumulated for one year, which constitutes of approximately 9.2 billion records.

Regarding very large data issue, large-scale data management for mobility analysis has been proposed by using a cloud computing platform named Hadoop as a core component. It provided the scalability both in data storage and parallel data-processing because the overall performances and storages can be increased by adding additional nodes to the cluster without requiring modification of software. Together with Hive, a data warehouse service on top of Hadoop, it allows using SQL-like language for easily and smoothly processing of data on Hadoop. Several techniques are applied to optimize the overall performances such as configuration tuning, HiveQL tuning and array-based data structure.

For spatial data processing, spatial extension on Hive has been proposed to provide spatial processing support on Hive. Several components and techniques including Java Topology Suite

(JTS), User-Defined Function (UDF, UDAF, and UDTF), Map-Join and Lateral-View are combined for developing of an extension. Spatial Index, Hadoop Distributed Caches and Static Objects are utilized to improve the processing speed. With performance comparison of Hive and ordinary methods such as PostgreSQL database, Hive with spatial extension outperformed other methods by processing more than 1000 times faster on the only four nodes of the cluster.

Considering the scalability of mobility analysis techniques, mobility analysis library has been developed on Hive with some improvements to accelerate processing speed on the large-scale data. The library included frequently used algorithms such stay point extraction, location clustering, trip segmentation and features calculation. Additionally, four essential mobility algorithms have been accommodated for more accurate analysis on the mobile phone dataset. Those algorithms comprised of life pattern extraction, significant places extraction, trip reconstruction and anomaly event detection.

For life pattern extraction, an unsupervised learning method using user active-period has been proposed to discover life pattern of people. It involves four main steps: active-period calculation, data grouping, normalization and data clustering. Iterative K-mean clustering was employed to automatically identify an optimum number of cluster results. Cosine Similarity was also used as the distance measurement function of the clustering method. As the result, it was able to cluster the activities of people into 24 common patterns including some important patterns such as ordinary-activity people, nighttime worker and shift-based worker. Those patterns allow efficiently selecting sampling data for the detail and for labeling which cover all kinds of people and behaviors in the dataset.

For significant places extraction, an approach to discover user important places and to derive types of locations especially home and work place has been proposed. The process involves several techniques including stay point extraction, spatial clustering of stay points (DBSCAN), and inference model (Random Forest). In addition, new classification features for inferring home and work are introduced such as a total days that point appear, a total hour periods that point appear and a percentage of stay point in appearing in low active period of a user. A web-based home/work labeling tool using Google Map is developed for creating validation data as well as training data for the classification. The proposed-features can accommodate inferring Home/Work of people especially unusual-behavior people such as nighttime worker, shift-based-worker.

For trip reconstruction, a framework based on supervised learning method was proposed to reconstruct user's trip information from low data rate GPS data of mobile phones. The approach consisted of three main steps including trip reconstruction, location labeling and route matching. Outlier detection and removal technique was employed to increase the accuracy of stay point extraction. Random Forest classifier was used for classifying transportation mode including stay, walk, bike, car and train. GIS information such as spatial train network and spatial road network was employed and used as classification features to improve segmentation and classification effectiveness. In addition, to validate the results, web-based trip visualization and labeling has been developed.

For anomaly event detection, large-scale mobile phone dataset were explored for a possibility of detecting anomaly events. A framework based on hidden Markov model (HMM) to detect anomalous area-based events has been proposed. The HMM was used as the main algorithm to construct a population-movement pattern in each grid. The used of local scoring, the difference in probability as compared with previous instances, has also been introduced. Through the local score, it is possible to detect an event down to the level of the event period, rather than only the date level such as detecting time period that the event occurs. The proposed system clearly distinguishes periods of anomalous events from other periods particularly for the large public event that attract people longer than 2.5 hours.