Available online at www.sciencedirect.com
ScienceDirect
Procedia Computer Science 35 (2014) 1495 – 1504
Procedia Computer Science

18th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems - KES2014

Identifying types of staying facilities from traffic behavior log data

Yu Sugawara[a]*, Takeshi Morita[b], Hidenao Abe[c], Shuichi Matsumoto[c], Takahira Yamaguchi[a]

[a]*Keio University 3-14-1 Hiyoshi, Kohoku-ku, Yokohama-shi, 223-8522 Japan*
[b]*Aoyama Gakuin University 5-10-1 Fuchinobe, Chuo-ku, Sagamihara-shi, 252-5258 Japan*
[c]*Bunkyo University 1100 Namegaya, Chigasaki-shi, 253-8550 Japan*

Abstract

Traffic behavior surveys by hand require both a lot of money and human resources. Recently, traffic behavior surveys using information technology have been carried out. In this study, we propose a method to extract staying points from GPS-based positional data and identify the types of staying facilities by using Google Places API, a facility ontology, the regularity which is analyzed from trip chains about traffic behavior. This method could identify 68.5% types of staying facilities correctly in the evaluation using GPS location data from the Traffic Behavior Survey in Nagasaki.

Keywords: Traffic Behavior Survey, Person Trip Survey, Positional Data, Log Data, Automation

1. Introduction

Recently, “Person Trip(PT) Surveys”, which are one type of traffic behavior survey by hand are often carried out by local governments in Japan. Trip, or person trip, is a single or one-directional movement of one person from one point (origin) to a second point (destination), for a single purpose[1]. PT Surveys collect information including times and places of departure and destination, staying facilities, the types of staying facilities, means of transportation and purposes of traffic behaviors. General traffic behavior surveys including PT Surveys have three issues; the cost of filling out a survey form by hand, the loss of data by missing out and the cost of digitizing survey results. These issues are attributed to those traffic behaviour surveys have many items requiring notation and the fact that paper forms are used.

* Corresponding author. Tel.: +81-45-566-1614.
E-mail address: y.sugawara@z7.keio.jp.

doi:10.1016/j.procs.2014.08.232

To solve these problems, the automations of PT Surveys have been conducted. One of them is the Probe Person(PP) Survey which is a survey that uses GPS mobile phones or other mobile communication tools along with Web Diary on the Internet to record trip information of person and vehicle[2]. However, PP Survey records staying facilities and the purposes of traffic behaviors with manually operated methods and there is little research about identifying staying facilities and the purpose of traffic behaviors.

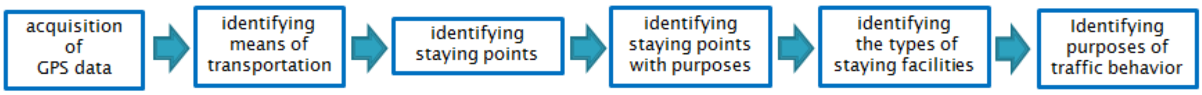

Fig. 1. Issues for the automation of PT Surveys

As shown in Fig 1, there are six issues for the automation of PT Surveys. In these 6 points, there is much research about the acquisition of GPS data[3], identifying of means of transportation[4] and the identifying of staying points[5]. Our research focuses on identifying the types of staying facilities. We suggest a method which uses GPS data to identify the types of staying facilities. In addition, we applied this method to the Traffic Behavior Survey in Nagasaki and evaluated the utility of this method.

On the other hand, there are a lot of problems with using large-scale GPS data such as big data relating the great east Japan earthquake and there is much research in relation to this[6]. The automation of PT Surveys is an example of using large-scale GPS data.

2. Related work

In this section, we explain the research relating to our study. [3] has shown an acquisition method of GPS data which uses smart phones. It is possible to lower cost than PT Surveys. However, the research[3] does not focus on acquisition other information such as transportation types, facility types and purpose of traffic behavior automatically. [4] has shown a method identifying means of transportation using data collected smartphones. In the research[4], Distinguishing eight different transportation modes, the classification results have ranged from 65% (train, subway) to 95% (bicycle). This research focuses on only transportation mode. [5] is a study which propose a method extracting stay, pass-by and potential stay areas. In addition, the research[5] has mentioned the relation between human activities and land use types.

Explained above, there is little research focusing on identifying types of staying facility and purposes of traffic behavior. Our research focuses on identifying the types of staying facilities. And then, our research identifies the staying points in order to acquire information about facilities surrounding staying point for identifying facility type.

3. Proposed method

In this section, we explain the overview and data flow of the method which identifies the types of staying facilities. Fig 2 shows the overview of our method.

This method consists of four main sub-processes; process for identifying staying points, map information base, process for converting facility types and process for determining the type of staying facility. This method identifies the types of staying facility using GPS data as input by communicating with each of the four sub-processes. Fig 3 shows the data flow of this method. In Fig 3, ovals represent the processes, rectangles represent data, columns represent ontology, one-way arrows indicate input and output and two-way arrows indicate references.

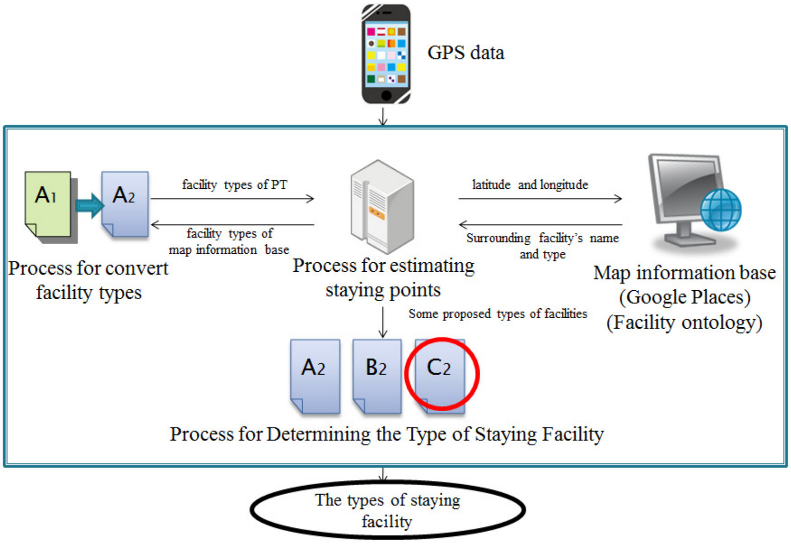

Fig. 2. Method overview

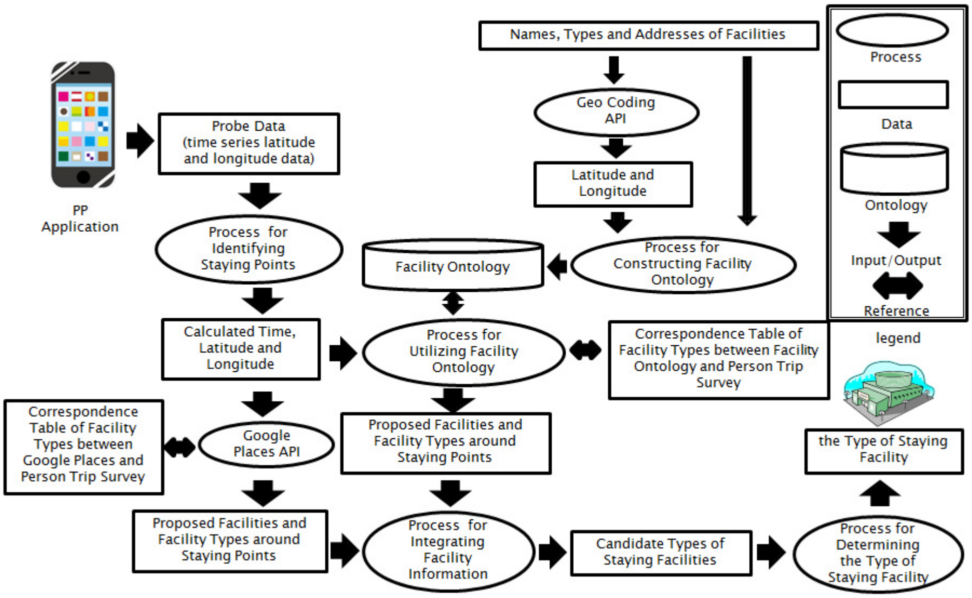

Fig. 3. Data flow of method

In this method, staying points are first identified in the Process for Identifying Staying Points by using GPS data based on a rule mentioned later. Next, the latitude and longitude of staying points are input into to Google Places API[7] and Facility Ontology, and facility information around staying points is obtained. Google Places API is a free web API offered by Google. At that time, facility types obtained from Google Places and Facility Ontology is converted to those used in PT Surveys using correspondence tables. Information regarding these facilities located around staying points is integrated and only facility types are extracted from the information in the Process for Integrating Facility Information. Finally, the types of staying facilities are identified by narrowing down the Candidate Types of Staying Facilities in the Process for Determining the Type of Staying Facility. Details of each process are explained below.

```
FOR counter = 1 to amount of GPS data DO
    IF velocity <= 3 km/h THEN
        Store latitude, longitude and time to data array
    ELSE
        IF length of data array >= 2 THEN
            IF last time in data array – first time in data array >= 5 minutes THEN
                staying point[ i ].start time = first time in data array
                staying point[ i ].average of latitude and longitude = Calculate the average of latitude and longitude
                staying point[ i ].standard deviation of distance = Calculate the average of latitude and longitude
                i++
            ENDIF
        ENDIF
        Initialize data array
    ENDIF
ENDFOR
```

Fig. 4. Pseudo code for identifying staying points

3.1. Process for identifying staying points

This process inputs GPS data and outputs the average and the standard deviation of latitudes and longitudes of identified staying point and start and end time of each stay.

To identify staying points, we defined a stay as a state of having velocity below three kilometers per hour (km/h) for a period of five minutes. This definition was determined based on the fact that it is said the average speed of walking is about four km/h and that short time stays such as visiting a convenience store are covered by PT Surveys. According to this definition, we developed an algorithm for extracting staying points from GPS data as shown in Fig. 4.

This algorithm determined data whose velocity was under three km/h as a trigger and then found the data whose velocity was over three km/h. The elapsed time was calculated using this data. If the elapsed time was over five minutes, this algorithm outputted the average and standard deviation of latitudes and longitudes during this stay in addition to start and end time of the stay as staying point data.

3.2. Google Places API

Google Places API is a web API which inputs latitude, longitude and radius and outputs names, addresses and types of facilities which are registered in Google Places. Most information used in this method is acquired by Google Places API. In this method, the latitude and longitude were that which is calculated in the Process for Identifying Staying Points and the radius was calculated as follows. The basic value was defined as 50 meters and standard deviation of latitude and longitude were added to the basic value. When Google Places API inputted the calculated radius, latitude and longitude and outputted less than five surrounding facilities, the basic value was defined as 100m and the radius was recalculated. Because the types of facilities in the Google Places API were different from those in the PT Surveys, conversion of these was necessary.

Table 1. Facility types correspondence table

Facility Types in Google Places API	Facility Types in PT Survey
accounting	office and company
airport	others
amusement_park	other commercial facility
aquarium	other commercial facility
art_gallery	public facility
atm	others
bakery	other commercial facility
bank	bank
bar	restaurant
beauty_salon	others

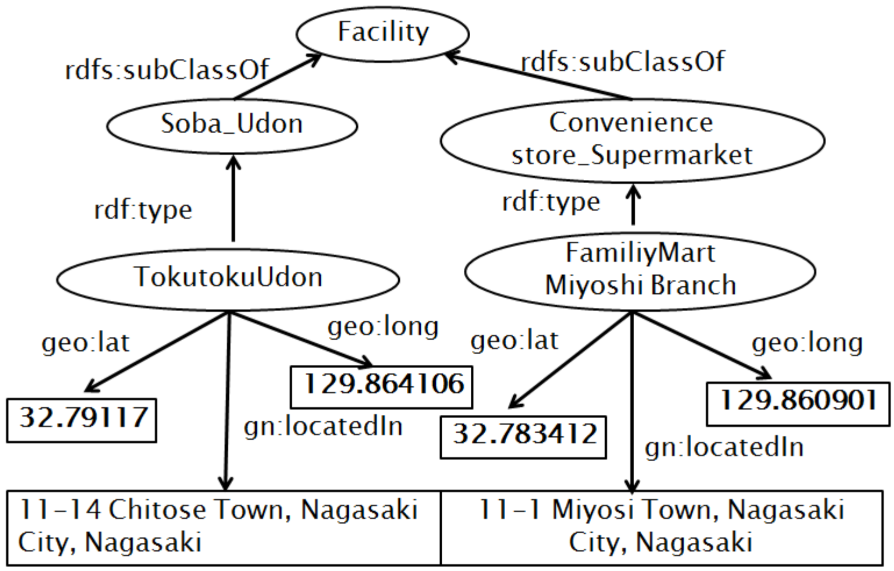

Fig. 5. Nagasaki Facility Ontology

In this research, the types of facilities were converted using correspondence tables like that of Table 1, which was a part of a correspondence table between facility types in the Google Places API and facility types in PT Survey named Traffic Behavior Survey in Nagasaki explained in section 4.

There are 96 types of facilities in the Google Places API and each of them corresponded to one of 14 categories in the PT Survey.

3.3. Process for utilizing facility ontology

In order to compensate for the lack of facilities in the Google Places API, facility ontology was built such as the Nagasaki facility ontology, which is for the Traffic Behavior Survey in Nagasaki and part of this is shown in Fig. 5.

Nagasaki Facility Ontology uses four name spaces; “rdf”, “rdfs”(RDF Schema), “geo” and “gn”(Geo Names). “rdf” and “rdfs” are name spaces providing data-modeling vocabulary for RDF and RDFS data[8,9]. “geo” is a name space providing vocabularies about location information such as latitude and longitude. “gn”(Geo Names) is a name space providing vocabularies about geographical information. “geo” and “gn” are provided by [10].

Nagasaki Facility Ontology has types of facilities as class and facilities as instances which belong to each class. Each instance has an address, latitude and longitude. In this research, types of facilities, facilities belonging to each type, address, latitude and longitude were obtained by taking from “i Town Page”[11] and Nagasaki Facility Ontology was constructed automatically using this information. Because the types of facilities in the facility ontology were different from those in the PT Surveys as well as the Google Places API, those were converted using the correspondence table.

Latitude, longitude and radius were calculated as well as the way explained in Section 3.2. A central point was determined as the point which has the calculated latitude and longitude. SPARQL query that selected facilities which were in the radius from the central point was made and performed.

3.4. Process for integrating facility information

This process focused on only the types of facilities and integrated those in the Google Places API and facility ontology removing duplicates.

There are some types of facilities which cannot be categorized with only information in the Map Information Base such as residential buildings, homes, places of employment and places of study. This is because residential buildings are not registered in the Google Places API and it is difficult to obtain the addresses of residential buildings to construct the ontology. This is also because homes, places of employment and places of study cannot be categorized with common information such as information from the Map Information Base as it is unique to each person. Due to the aforementioned reasons, this process added these types of facilities.

The way to add residential buildings, homes, places of employment and places of study is explained bellow. When the number of facilities surrounding staying points is less than five, residential buildings are added to the types of staying facilities. This is because a residential street tends to have less commercial facilities than other places. When homes, places of employment and study are added to the types of staying facilities, long time stays are focused on. Namely, extracted staying points are divided as group so that staying points whose distances are close are part of the same group. And then, in these groups, the number of long time stays, whose sojourn time is longer than two hours is counted and the group whose number of long time stay is largest of all is named as the home group, and other groups whose numbers of long time stays are large these are named as a place of employment and study. If a staying point belongs to the groups named, the corresponding facility type is added to the types of staying facilities.

3.5. Process for determining the type of staying facility

In this process, a type of staying facility is chosen from the types of staying facilities integrated in the Process for Integrating Facility Information. The degree of relative priority is used to choose this at that time. The degree of relative priority is acquired by analyzing the trip chain, consecutive data of the move from departure place to place of arrival. The degree of relative priority of facility types in the Traffic Behavior Survey in Nagasaki are explained below as an example.

Table 2 shows the types of staying facilities and degrees of relative priority in the Traffic Behavior Survey in Nagasaki. A smaller number indicates a higher priority facility.

Homes, places of employment and study were determined with high reliability because they were identified based on rules in section 3.4. The degree of relative priority of residential buildings was judged to be high because residential buildings were not recognized if the degree of relative priority was set lower than the others.

Table 2. Priority of facility types in the Traffic Behavior Survey in Nagasaki

Priority	Type of Facility
1	home
2	place of employment and study
3	residential building
4	department store and shopping mall
5	supermarket and convenience store
6	restaurant
7	other commercial facility
8	school
9	hospital
10	office and company
11	public facility
12	bank
13	hotel
14	others

Regarding the other types, the lower priority the type of facility had, the more frequently it appeared in candidate types of staying facilities, especially from department stores and shopping malls to other commercial facilities. In other words, department stores and shopping malls had high rarity. The lower priority the type had, the lower rarity the type tended to have. In addition, when there was a type whose rarity was high, it was usually the type of facility where the examinee actually stayed. These were the reasons the priority was set as per Table 2.

However, there were some phenomena which could not be dealt with using only the priorities. Therefore, two additional rules were made:

- If there are residential buildings and schools, schools have a higher priority than residential buildings
- If a staying point is one which is visited right after homes or places of employment and study, other commercial facilities have a higher priority than supermarkets and convenience stores.

The first rule was set in order to avoid the situation that the chosen type was residential building although the type of facility where the examinee actually stayed was a school. Because facilities belonging to schools had large area, there were few facilities in the search radius registered in the Map Information Base if a staying point was in a school. In this case, residential buildings were added to the candidate types of staying facilities since there were few facilities surrounding the staying point. Compared to the priority of these types, residential buildings had a high priority and so, the type of staying facility was identified as a residential building. Avoiding this situation, if the priority of the school is set higher, the type of staying facility is identified as a school in error when an examinee utilizes commercial facilities in order to go shopping or a restaurant in order to have a meal in a downtown area near school buildings. Considering these factors, the first rule was set.

The second rule was set in order to correctly identify other commercial facilities, which were frequently visited. Other commercial facilities had a wide range; 40 types of facilities out of the 96 types in the Google Places API corresponded to other commercial facilities. As a result, the number that the examinee stayed at other commercial facilities is large and the frequency that there is another commercial facility in the candidate types of staying facilities is quite high. If the priority of other commercial facilities is set high, the number of failures of identifying increases. According to analysis of the trip chain, it was discovered that other commercial facilities tended to be visited right after a stay at home or to visiting a place of employment and study. These factors were reflected in the second rule.

Based on the priorities and the rules explained above, the type of staying facility is identified by choosing a type from some candidate types.

When a stay follows another one whose group is the same as the one before, these are considered to be the same stay and the data of these stays is combined in this process.

4. Discussion

4.1. Traffic Behavior Survey in Nagasaki

In order to evaluate this method, a PP Survey was held in Nagasaki; Traffic Behavior Survey in Nagasaki. This survey had 10 people as examinees and was held for ten days. The obtained data is GPS location data based on smart phones and trip chains.

For the purpose of acquisition of GPS data, the "Probe Person System"[12] offered by Transfield Co. was used and to acquire trip chains, Web Diary was used. Web Diary is a system on computers to enter information which can be obtained in PT Surveys - staying facilities, the type of staying facilities, time of start and arrival, means of transportation and the purpose of traffic behavior.

According to this GPS data, the types of staying facilities in this survey were identified and they were checked to be right or wrong by referring to the trip chains. Accuracy of identifying was verified to be either correct or incorrect. In this verification, GPS data of which large parts was missing due to smart phones running out of buttery, etc., was excluded.

4.2. Verification of the accuracy of identifying and observation

In this research, precision and recall were used for the verification. Our definitions of these indicators were shown (1) and (2).

$$precision = \frac{the\ number\ of\ staying\ facility\ types\ identified\ correctly\ by\ this\ method}{the\ number\ of\ staying\ facility\ types\ identified\ by\ this\ method} \tag{1}$$

$$recall = \frac{the\ number\ of\ staying\ facility\ types\ identified\ correctly\ by\ this\ method}{the\ number\ of\ t\arg et\ staying\ facility\ types\ of\ this\ evaluation} \tag{2}$$

Precision reflects the percentage this method could identify correctly out of all identified types, so this indicator is deteriorated if this method recognizes stays incorrectly. Recall reflects the percentage this method could identify correctly out of all facility types actually visited, so this indicator shows how much all target stays were covered. Table 3 shows the result of identifying the types of staying facilities. As a result of this verification, precision is 54.4% and recall is 68.5% in this survey. The recall and failure situation of each type of staying facility is shown in Table 4 and Table 5. The failure situations of each type of staying facility in Table 4 corresponds to Table5; the number shows the failure code in Table5 and the number between parentheses shows the number of the failure in Table 5.

Table 3. Result of identifying the types of staying facilities

		Is this identified correctly?		Recall
		Yes	No	
Is this a target staying point?	Yes	174	80	68.5%
	No	146	N/A	
Precision		54.4%		

Table 4.Failure situation of each type of facility

Type of Facility	the Number of Facilities	Recall	Failure Situation
home	100	88.0%	1(6),5(4),6(2)
place of employment and study	60	78.3%	1(4),4(1),5(1),6(7)
residential building	6	66.7%	1(1),4(1)
department store and shopping mall	20	65.0%	1(4),5(2),8(1)
supermarket and convenience store	14	35.7%	1(1),2(6),3(1),5(1)
restaurant	8	12.5%	2(5),3(1),10(1)
other commercial facility	27	40.7%	1(3),2(8),11(1),12(4)
school	7	57.1%	2(2),9(1)
hospital	4	0.0%	1(2),2(1),5(1)
office and company	0	N/A	N/A
public facility	4	25%	2(2),5(1)
bank	0	N/A	N/A
hotel	0	N/A	N/A
others	4	0.0%	2(3),3(1)

Table 5.Correspondence between failure code and failure situation

Failure Code	Failure Situation	the Number of Failure
1	Identified point is far from correct point.	21
2	Wrong type is chosen in Process for Determining the Types of Staying Facility.	27
3	The type is a wrongly contained place of employment and study.	3
4	There is no residential building in the candidate types.	2
5	The stay is not recognized by this method.	10
6	The type is not correctly recognized as a home or place of employment and study.	9
7	Another stay is combined because the groups are the same	0
8	There are facilities in the building but there is not a building in surrounding facilities.	1
9	The facility is not in the radius because its size is too large.	1
10	The type is wrongly categorized as a residential facility	1
11	This method cannot distinguish the stay from another stay	1
12	The type is wrongly categorized as place of employment and study.	4

As shown in Table 4 and Table 5, the recall of each type of facility is different. Home and places of employment and study make up a high percentage of recall but the other types' recalls are less than the total recall. In the other words, this method can identify homes and places of employment and study well but other parts need improvement for practical use. Furthermore, the failure situation whose number is largest of all is the "Wrong type is chosen in Process for Determining the Types of Staying Facility". This shows Process for Determining the Type of Staying Facility has some points to improve upon. These results indicate that the identifying method based on priority and the additional two rules cannot identify complex traffic behavior, so an increase in the number of rules is needed. Additionally, it is thought that the identifying method using only data about latitude, longitude and time limits accuracy of identifying. With a view to improving accuracy of identifying, verification is needed whether other information such as means of transportation can improve accuracy. Furthermore, for consideration of the automation of PT Surveys, accuracy of identifying obtained from this method and how long this method will need to alter the priorities and rules for when this method is adapted to another area and other types of facilities, should be considered.

5. Conclusion

In this paper, PT Surveys were focused on as a case study of traffic behavior surveys and we proposed an identifying method for types of staying facilities, one of steps in the process of automation of PT Surveys. This method inputted GPS location data which was acquired by the PP Survey and outputted types of staying facilities using a Map Information Base. Moreover, we identified 68.5% types of staying facilities correctly in the evaluation using GPS location data obtained in the Traffic Behavior Survey in Nagasaki.

In the future, we will utilize not only GPS data but also information identifying other processes such as identifying of means of transportation. In addition, we will consider the costs and the possibility of being adapted to other types of facilities and other areas.

References

[1] Introdution in the 2011 Origin-Destination (O-D) survey, http://www.ncr-trans-rcn.ca/surveys/o-d-survey/o-d-survey-2011/ (2011)
[2] JSTE (Japan Society of Traffic Engineers) Probe Research Group, www.probe-data.jp/eng/index.html.
[3] Nohara K, Matsuda K , Maruyama T. Development and Application of Smartphone-based Travel Survey Method : Preliminary results in Kumamoto, Japan. *the 13the World Conference on Transport Research*, 2013, #3242.
[4]Philippe N, Peter W, Simon B, Norbert B, Peter M. Supporting large-scale travel surveys with smartphones – A practical approach. *Transportation Research Part C: Emerging Technologies,*2013, http://dx.doi.org/10.1016/j.trc.2013.11.005.
[5] Shan J, Gaston A. F, Yingxiang Y, Joseph F Jr., Emillio F, Marta C. G. A Review of Urban Computing for Mobile Phone Traces: Current Methods, Challenges and Opportunities. *Proceeding UrbComp '13 Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing Article No2*, 2013, http://dx.doi.org/10.1145/2505821.2505828 .
[6] Great East Japan Earthquake Big Data Workshop - Project 311 -, https://sites.google.com/site/prj311e/home.
[7] Google Places API, https://developers.google.com/places/documentation/.
[8] RDF/XML Syntax Specification (Revised),http://www.w3.org/TR/REC-rdf-syntax/.
[9] RDF Schema 1.1, http://www.w3.org/TR/rdf-schema/.
[10] GeoNames Ontology, http://www.geonames.org/ontology/.
[11] i Town Page, http://en.itp.ne.jp/
[12] Probe Person System, http://www.transfield.co.jp/pp.html.