Association Analysis of Location Tracking Data for Various Telematics Services*

In-Hye Shin and Gyung-Leen Park**

Dept. of Computer Science and Statistics, Jeju National University
690-756, Jeju Do, Republic of Korea
{ihshin76,glpark}@jejunu.ac.kr

Abstract. This paper proposes an approach that extracts the association information from the location data obtained from the real fields but ignored so far. We provide and apply the approach to the real-life location tracking data collected from the Taxi Telematics system developed in Jeju, Korea. The analysis aims at obtaining taxies' meaningful moving patterns for the efficient operations of them. The proposed approach provides the flow chart which would not only take a glance around the overall analysis process but also help save temporal and economic costs required to employ the same or similar data mining analysis to similar services such as public transportations, distribution industries, and so on. Especially, we perform an association analysis on both of refined data and interesting factors extracted from the elementary analysis. The paper proposes the refined association rule mining process as follow: 1) obtaining the integrated dataset through the data cleaning process, 2) extracting the interesting factors from the integrated dataset using frequency and clustering method, 3) performing the association analysis, 4) extracting the meaningful and value-added information such as moving pattern, or 5) returning the feedback to adjust inappropriate factors. The result of the analysis shows that the association analysis makes it possible to detect the hidden moving patterns of vehicles that will greatly improve the quality of Telematics services considering the business requirements.

1 Introduction

The Taxi Telematics system has operated an efficient taxi dispatch service since 2006 in Jeju island, Korea [1]. Each taxi equipped with a GPS receiver reports its location to the central server every minute. The server is responsible for keeping such reports from each taxi and handles the call request from a customer by finding and dispatching a taxi closest to the call point. The tracking data obtained from the Taxi Telematics system can be utilized for various analysis on both research and business areas [2]. We have been doing our research to develop a data processing framework

* This research was supported by the MKE (The Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program supervised by the IITA (Institute for Information Technology Advancement). (IITA-2009-C1090-0902-0040).

** Corresponding author.

D. Taniar et al. (Eds.): ICCSA 2010, Part III, LNCS 6018, pp. 107–117, 2010.

for an efficient analysis [3-9]. It is the exact fact that the empty taxi ratio can be reduced by means of guiding a taxi to the spot where many passengers are waiting for. In the project, the empty taxi ratio is around 80 % according to a survey in [6]. Thus, many taxi businesses are interested in taxis' moving patterns in order to decrease the consumption of fuels as well as increase the income. Today the fuels consumption is more important to protect the environments. The real pattern analysis using data mining is an interesting topic due to importance of the value-added information, as shown in [10]. Especially, [4] has proposed the framework and perform the clustering analysis for the location recommendation. The given analysis framework consists of both of the data processing framework and the analysis processing framework for the refined clustering. However, it does not provide association rules effectively so far. In this regard, this paper is to propose the association-specific analysis framework, perform the detailed analysis for association rule mining, and provide concrete pattern information. The proposed analysis processing framework consists of both of the given data processing framework [4] and a novel analysis processing framework for the refined association analysis aiming at detecting the frequent or meaningful moving patterns. This framework can help perform even a sophisticated and a quick analysis. Particularly, the paper is to develop the refined association analysis by means of taking into account interesting factors such as the driving type of taxi, boarding hour, pick-up/drop-off time, pick-up/drop-off area, boarding rate, and so on.

The result of analysis shows that the approach employed in the paper can be extended to similar areas such as developing public transportation systems, distribution systems, and so on.

The paper is organized as follows. Section 2 proposes the analysis process framework for the location tracking data collected from the Taxi Telematics system. Section 3 exhibits the summary of the results obtained from the refined association analysis to search the taxis' moving patterns, based on whole data, short driving data, and long driving data, respectively. Finally, Section 4 concludes the paper.

2 Proposed Framework

This section proposes both of the data processing framework and the analysis processing framework, in order to show effectively how to analyze the moving patterns of taxis using association analysis. The proposed frameworks provide the flow chart which would take a quick look around both of the data flow and the overall analysis process. Also, it helps quickly perform the same or similar analysis, while sparing the temporal and economic costs.

2.1 Proposed Data Processing Framework

Figure 1 outlines the overall procedure for location data processing. Taxis report their location records to the central call server every minute. Each record includes the basic GPS data such as timestamp, latitude, longitude, direction, speed taxi ID as well as status fields (related with pick-up and drop-off). First of all, an analyzer transforms the type of such records into data type readable in SAS then stores them with SAS dataset types. We can analyze the location dataset using SAS analysis engine including data mining analysis tool, Enterprise Miner (or E-Miner).

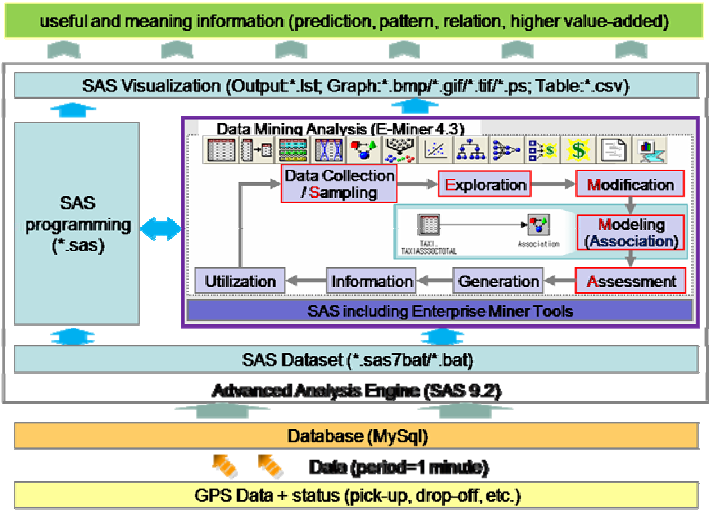

Fig. 1. Data processing framework including SAS Data Mining

To obtain more useful information, data mining analysis process consists of the data cleansing process, the modeling process which analyzes the refined data using a proper method, the assessment process, the utilization process, and feedback process, as shown in Fig. 1. As shown in Fig. 1, Modeling, fourth step of SEMMA [10], fits a predictive model; modeling a target variable using a cluster analysis, an association analysis, a regression model, a decision tree, a neural network, or a user-defined model. The paper exploits an association method for data modeling. Association searches interesting association or correlation relationships among a large set of data items.

2.2 Proposed Processing Analysis Processing for the Association Analysis

Fig. 2 shows the analysis processing framework for the refined association rule miming based on the taxi location records to inform their valuable moving patterns and moreover recommend the best location and time to the taxi drivers. The framework shows an overall process of the association analysis.

The processing framework could be expansively adapted to another transport industry, similar to a taxi business, such as delivery by changing some variables or generating new variables, and repairing the flow chart partially. As shown in Fig. 2, a data analyzer should load source records first, restore them as the SAS data type, modify the SAS dataset through the variable transformation and data filtering for data cleaning. We can create new data extracting the paired data by a pick-up and drop-off record using data split/merge. Then we select the interesting or significant variables form the pared dataset by the result of clustering. We adjust each variable's category by the result of frequency method, resulting in TaxiType (driving type of Taxi), SArea/EArea (pick-up area/drop-off area), SDay/EDay (pick-up day/drop-off day), STimeR/ETimeR (pick-up time/drop-off time). TaxiUseRatioR (boarding rate in terms of hour) variable, which means the boarding rate according to a driving hour, is obtained by dividing the number of pick-up report into the number of total report. The

analyzer completes the integrated dataset by merging original dataset and extracted interesting factors. It also needs new dedicated dataset for association analysis consisting of ID and Target variable. We divide association dataset into two by TaxiType variable with ShortDriveTaxi (short-driving) and the others (long-driving). The analyzer can return the feedback to the previous step (extraction of interesting factors), if not detect strong association rules. There is a recurrence of such a refined data mining analysis process, adjusting the feedback factors which mean modification of the given factor category and detection of a new factor until obtaining the good association rules.

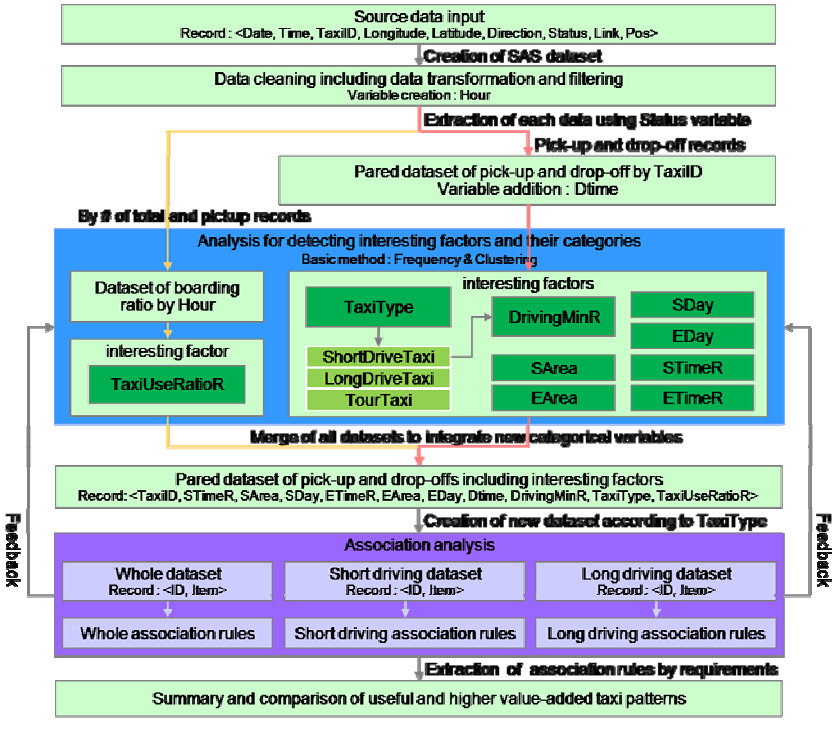

Fig. 2. The analysis processing framework for the association rule mining of taxi records

2.3 Extracted Interesting Factors

Table 1 shows the extracted interesting factors, or categorical variables, obtained from clustering and frequency method using the refined and integrated SAS dataset. Note that each frequency should be alike within each variable's category, or not biased. As shown in Table 1, we omit EArea, EDay, ETimeR variables (related with drop-off), which correspond to SArea, SDay, and STimeR (related with pick-up), due to space limitation of the paper as well as similar frequency distribution as pick-ups. Note that STimeR variable provides additionally the average frequency per hour in parenthesis, due to non-equaled time range.

Table 1. Summary of each interesting factor

Variable	Category	Description	Frequency	Percent (%)
TaxiType	ShortDriveTaxi	~ 30minutes	77447	95.51
	LongDriveTaxi	30minutes ~ 2hours	2309	2.85
	TourTaxi	2hours ~	1331	1.64
DrivingMinR	5MinDriving	1~5minutes	44507	54.89
	5-10MinDriving	5~10minutes	22254	27.44
	10-15MinDriving	10~15minutes	8787	10.84
	15~30MinDriving	15~30minutes	5539	6.83
TaxiUseRatioR	LowUse	10~15%	33082	40.80
	MiddleUse	15~20%	23389	28.84
	HighUse	20~25%	24616	30.36
SArea	SAirport	Ariport	3727	4.60
	SOldTown	Old town	43644	53.82
	SNewTown	New town	22929	28.28
	SEtc.	The others	10787	13.30
SDay	SMonday	Monday	9932	12.25
	STuesday	Tuesday	11405	14.07
	SWednesday	Wednesday	12335	15.21
	SThursday	Thursday	11582	14.28
	SFriday	Friday	12996	16.03
	SSaturday	Saturday	12033	14.84
	SSunday	Sunday	10804	13.32
STimeR	SOfficeStart	07~10o'clock	10469 (3490)	12.91
	SOfficeHour	10~18o'clock	30673 (3834)	37.83
	SOfficeEnd	18~21o'clock	17434 (5811)	21.50
	SEvening	21~23o'clock	8626 (4316)	10.64
	SNight	23~01o'clock	6906 (3453)	8.52
	SDawn	01~07o'clock	6979 (1163)	8.61

3 Association Analysis Results

This section shows the results obtained from the refined association analysis. In the previous section, we have split the taxi association datasets into three categories, whole (81,087 records), short driving (77,447 records), and long driving (3,640 records). 2. Thus we present three association results according to each dataset. Association has a support level and a confidence level, two main measures of rule interestingness. In addition, lift is a measure of strength of the association rule. Typically, association rules are considered interesting if they satisfy both a minimum support threshold and a minimum confidence threshold. Such thresholds can be set by users or domain experts. Association rule mining is a two-step process: 1) finding all frequent itemsets; 2) generating strong association rules from the frequent itemsets. Note that we set minimum confidence to 10% by using the default association setting.

3.1 Analysis Results for Whole Taxis

Table 2 shows the meaningful and interesting association rules extracted from the association analysis based on the whole location data records. Stronger associations

Table 2. Association rules obtained from whole taxi records

Confidence	Support	Lift	Rule
100.00	10.64	3.29	SEvening ==> HighUse **(rule 1)**
70.04	15.06	2.31	SOfficeEnd ==> HighUse **(rule 2)**
83.53	31.60	2.05	SOfficeHour ==> LowUse **(rule 3)**
59.62	7.70	2.07	SOfficeStart ==> MiddleUse **(rule 4)**
100.00	5.69	3.29	SOldTown & SEvening ==> HighUse **(rule 5)**
69.94	8.22	2.30	SOldTown & SOfficeEnd ==> HighUse **(rule 6)**
74.85	8.79	1.46	SOldTownTown SOfficeEnd & ==> EOldTown
83.40	17.61	2.04	SOldTown & SOfficeHour ==> LowUse **(rule 7)**
72.40	8.50	1.46	SOldTown & SOfficeEnd ==> ShortDriveTaxi & EOldTown
83.17	8.61	2.04	SOfficeHour & SNewTown ==> LowUse **(rule 8)**
33.80	5.42	1.11	SFriday ==> HighUse
45.03	5.52	1.10	SMonday ==> LowUse
39.38	5.84	0.97	SSaturday ==> LowUse
41.99	5.60	1.03	SSunday ==> LowUse
40.46	5.78	0.99	SThursday ==> LowUse
41.87	5.89	1.03	STuesday ==> LowUse
40.56	6.17	0.99	SWednesday ==> LowUse
82.41	4.80	2.02	SOfficeHour & SFriday ==> LowUse
84.48	5.03	2.07	SWednesday & SOfficeHour ==> LowUse
71.78	6.31	1.40	SFriday & SOldTown ==> EOldTown
69.82	6.14	1.41	SFriday & SOldTown ==> ShortDriveTaxi & EOldTown

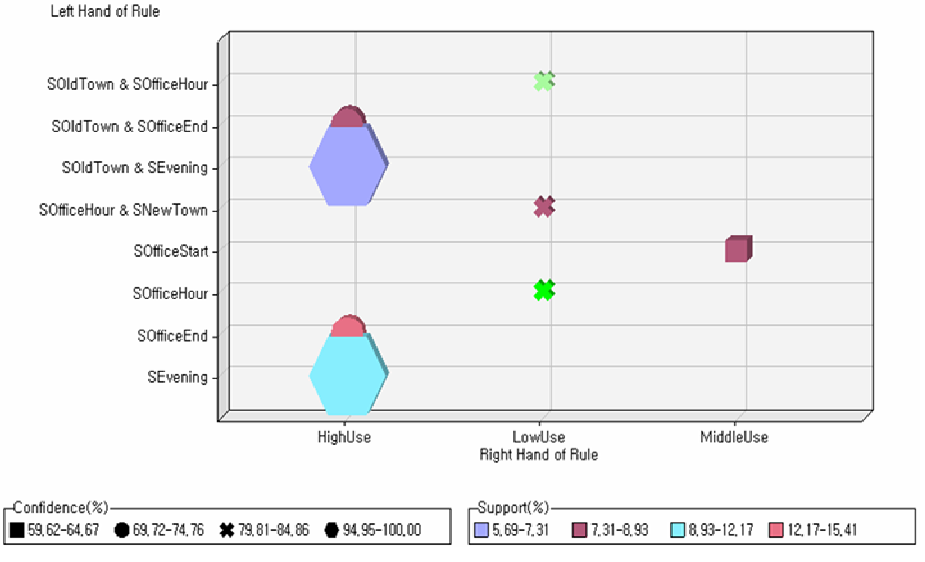

Fig. 3. The main association rules of whole taxi location records

are selected according the support level and the confidence level. The table shows that passengers use mostly taxi at the closing time of the office hours with 15.06% of the support level and 70.04% of the confidence level while the support level and the confidence level is 10.64% and 100%, respectively, for the evening hours (21~23 o'clock). Also it shows that only Friday is associated with HighUse of the taxies with 5.42% of the support level and 33.80% of the confidence level.

We depict graphs to compare those strong rules in Fig. 3 and Fig. 4. In Fig. 3, each polygon's shape, its color, and its size are regarded as confidence, support, lift, respectively. Fig. 4 shows the rule comparison in terms of two important factors, the confidence level and the support level.

3.2 Analysis Results for Short Driving Taxies

Table 3 shows the associations based on the short-driving taxi records. Eight strong associations are selected. Fig. 5 and Fig. 6 show the comparative graph among those

Fig. 4. Confidence and support of the main association rules

Table 3. Association rules obtained from short-driving taxi records

Confidence	Support	Lift	Rule
100.00	10.67	3.31	SEvening ==> HighUse **(rule 1)**
54.66	4.69	1.81	SNight ==> HighUse **(rule 2)**
45.34	3.89	1.57	SNight ==> MiddleUse **(rule 3)**
69.86	14.87	2.31	SOfficeEnd ==> HighUse **(rule 4)**
83.58	31.63	2.05	SOfficeHour ==> LowUse **(rule 5)**
69.85	8.18	2.31	SOldTown & SOfficeEnd ==> HighUse **(rule 6)**
70.45	4.46	2.33	SOfficeEnd & SNewTown ==> HighUse **(rule 7)**
100.00	3.11	3.31	SNewTown & SEvening ==> HighUse **(rule 8)**
75.89	13.61	1.44	SNewTown & ENewTown ==> 5MinDrving
68.69	27.33	1.30	SOldTown & EOldTown ==> 5MinDrving
76.03	8.90	1.47	SOldTown & SOfficeEnd ==> EOldTown
46.59	5.46	1.56	SOldTown & SOfficeEnd ==> EOldTown & 5MinDrving

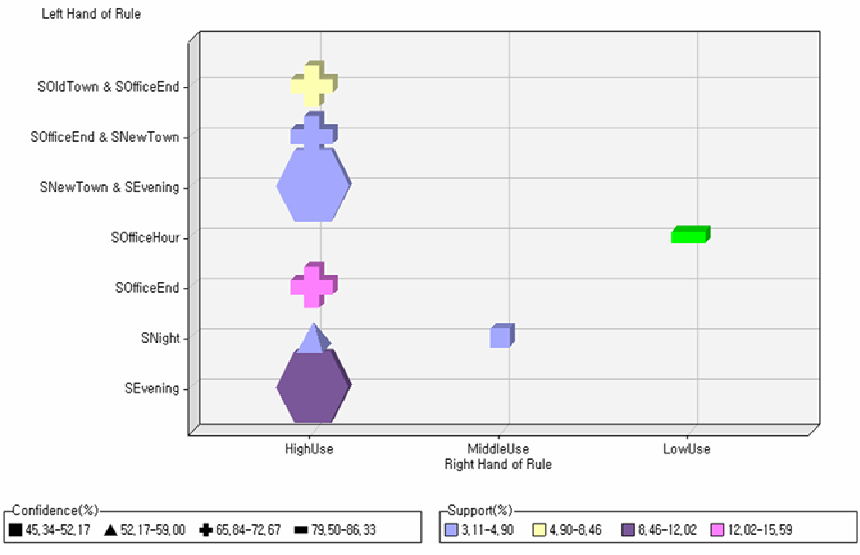

Fig. 5. The main association rules of short-driving taxies

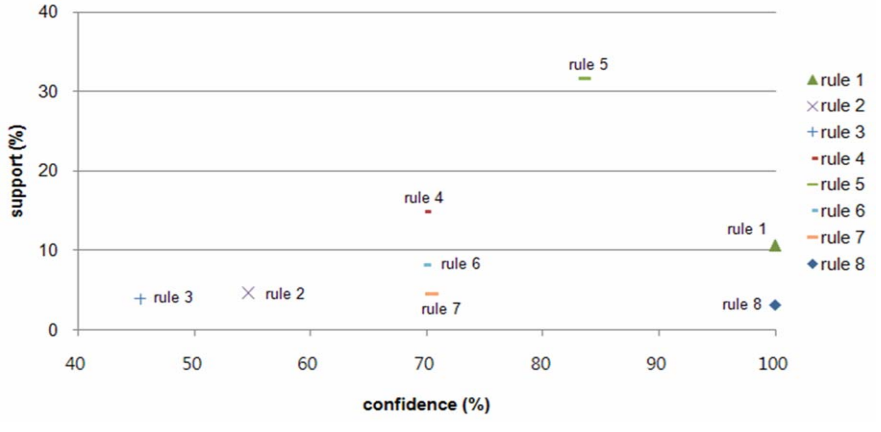

Fig. 6. Confidence and support of the main association rules for short-driving taxies

strong rules. The main association rules mean that short-driving taxis pick up many passengers at the close of office hours with support level 14.87% and confidence level 69.86%, respectively. Also mainly passengers move either within the old town or the new town within 5 minutes by a taxi. Especially, many taxies picking up a passenger move within the old town at the close of office hours.

3.3 Analysis Results for Long Driving Taxies

Taxi drivers are interested in the long driving pattern than the short due to the economic profit and the fuel cost, because the short driving consumes more fuel than

the long driving. Table 4 shows the meaningful and interesting associations based on the long-driving taxi records. Fig. 7 and Fig. 8 depict the main association rules. Finally, the main association indicates that long-driving taxis mainly move from old town to old town, especially at the close of office hours.

Table 4. Association rules obtained from long-driving taxi records

Confidence	Support	Lift	Rule
49.22	3.49	1.25	SDawn ==> LowUse **(rule 1)**
50.78	3.60	1.84	SDawn ==> MiddleUse **(rule 2)**
100.00	10.05	3.02	SEvening ==> HighUse **(rule 3)**
56.23	4.09	1.70	SNight ==> HighUse **(rule 4)**
43.77	3.19	1.59	SNight ==> MiddleUse **(rule 5)**
73.12	18.98	2.21	SOfficeEnd ==> HighUse **(rule 6)**
82.43	30.93	2.10	SOfficeHour ==> LowUse **(rule 7)**
40.00	4.84	1.02	SOfficeStart ==> LowUse **(rule 8)**
60.00	7.25	2.17	SOfficeStart ==> MiddleUse **(rule 9)**
100.00	4.56	3.02	SOldTown & SEvening ==> HighUse **(rule 10)**
100.00	3.05	3.02	SNewTown & SEvening ==> HighUse **(rule 11)**
71.87	8.98	2.17	SOldTown & SOfficeEnd ==> HighUse **(rule 12)**
82.27	14.53	2.10	SOldTown & SOfficeHour ==> LowUse **(rule 13)**
81.60	7.55	2.08	SOfficeHour & SNewTown ==> LowUse **(rule 14)**
61.18	2.86	2.22	SOldTown & SOfficeStart ==> MiddleUse **(rule 15)**
49.21	22.36	1.30	SOldTown ==> EOldTown
51.21	6.40	1.36	SOldTown & SOfficeEnd ==> EOldTown
28.62	6.40	1.10	SOldTown & EOldTown ==> SOfficeEnd

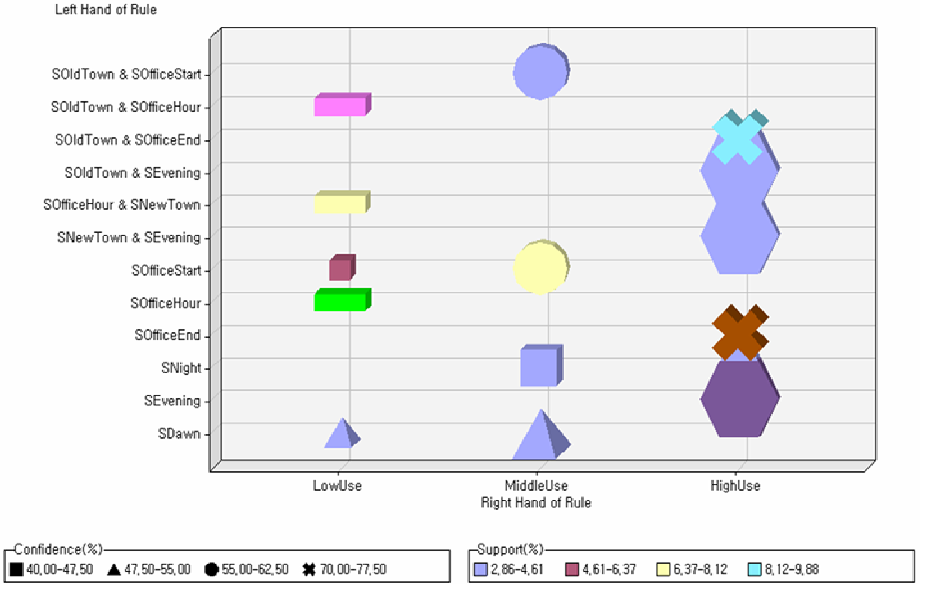

Fig. 7. The main association rules of long-driving taxies

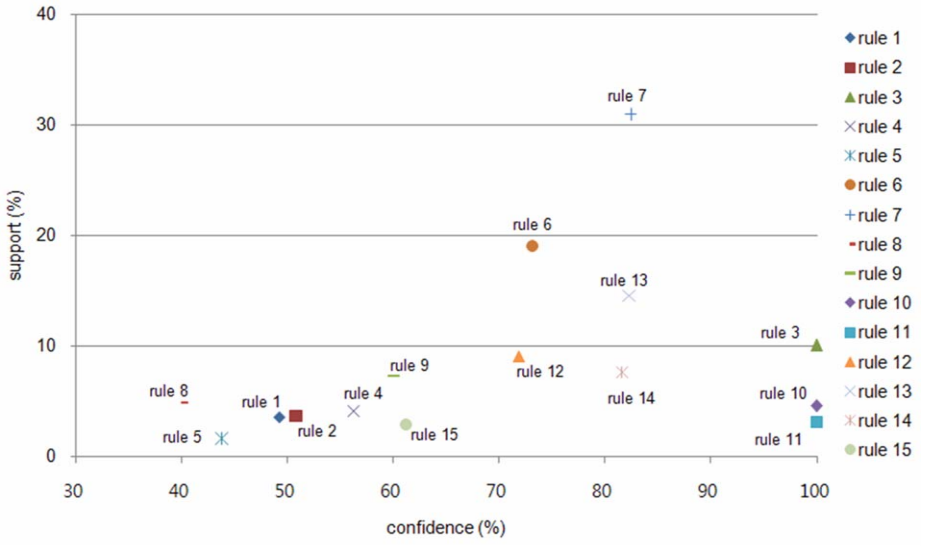

Fig. 8. Confidence and support of the main association rules for long-driving taxies

4 Conclusions

The paper has developed an analysis processing framework for detecting the hidden association rules of moving objects. In order to show the effectiveness of the framework, the framework has been applied to the real-life location history data accumulated from the Taxi Telematics system developed in Jeju Island. The proposed framework provides the flow chart which would have a glance around the overall analysis process as well as help save temporal and economic costs consumed to quickly address the same or similar mining analysis. The paper has extracted diverse interesting categorical factors such as taxi's driving type, driving time, driving area, boarding rate, and so on. The approach enables us to obtain meaningful and value-added association rules of moving objects by performing repeatedly the refined association analysis as follow: 1) obtaining the integrated dataset through the data cleaning process such as filtering and split-merge, 2) extracting the various categorical variables from the integrated dataset after performing the primary analysis such as frequency and clustering, 3) performing the method based on another dedicated dataset created for association, and 4) deducing the meaningful and value-added pattern information such moving pattern, or 5) returning the feedback in order to appropriately adjust the categories of extracted factors, if desired. The approach can be applied not only to taxi business but also to similar areas such as public transportation services and distribution industries.

References

[1] Lee, J., Park, G., Kim, H., Yang, Y., Kim, P., Kim, S.: A telematics service system based on the Linux cluster. In: Shi, Y., van Albada, G.D., Dongarra, J., Sloot, P.M.A. (eds.) ICCS 2007. LNCS, vol. 4490, pp. 660–667. Springer, Heidelberg (2007)

[2] Hariharan, R., Toyama, K.: Project Lachesis: Parsing and modeling location histories. In: Egenhofer, M.J., Freksa, C., Miller, H.J. (eds.) GIScience 2004. LNCS, vol. 3234, pp. 106–124. Springer, Heidelberg (2004)
[3] Lee, J., Park, G.: Design and implementation of a movement history analysis framework for the taxi telematics system. In: Asia-Pacific Conference on Communications, pp. 1–4 (2008)
[4] Shin, I., Park, G., Saha, A., Kwak, Y., Kim, H.: Analysis of Moving Patterns of Moving Objects with the Proposed Framework. In: Gervasi, O., Taniar, D., Murgante, B., Laganà, A., Mun, Y., Gavrilova, M.L. (eds.) ICCSA 2009. LNCS, vol. 5593, pp. 443–452. Springer, Heidelberg (2009)
[5] Lee, J., Hong, J.: Design and implementation of a spatial data processing engine for the telematics network. Applied Computing and Computational Science (2008)
[6] Lee, J.: Traveling pattern analysis for the design of location-dependent contents based on the Taxi telematics system. In: International Conference on Multimedia, Information Technology and its Applications (2008)
[7] Liao, Z.: Real-time taxi dispatching using global positioning systems. Communication of the ACM, 81–83 (2003)
[8] He, H., Jin, H., Chen, J., McAullay, D., Li, J., Fallon, T.: Analysis of Breast Feeding Data Mining Methods. In: Proc. Australasian Data Mining Conference, pp. 47–52 (2006)
[9] Madigan, E.A., Curet, O.L., Zrinyi, M.: Workforce analysis using data mining and linear regression to understand HIV/AIDS prevalence patterns. Human Resource for Health 6 (2008)
[10] Matignon, R.: Data Mining Using SAS Enterprise Miner. Wiley, Chichester (2007)