Mining users' Behaviors and Environments for Semantic Place Prediction

[1]A.Aqueel Ahmed, [2]P.Rizwan Ahmed
Assistant Professor, Head & Asst. Professor
Department of Computer Applications and PG Department of Information Technology
Mazharul Uloom College, Ambur.

ABSTRACT

In this work, we propose a novel prediction framework, which takes into account the spatial property, temporal property, users' behavior and environment at the same time, for semantic place prediction. The core idea of our proposal is to extract features to represent end users' behaviors in each place related to its semantic. To achieve this goal, we define 54 features to represent end users' behaviors to capture the key properties of places recorded in MDC Data Set. In our framework, we propose a novel model, namely Multi-Level Classification Model , to solve the imbalanced data problem. Based on the Multi-Level Classification Model, we make semantic prediction of a place by combining several classification models. To our best knowledge, this is the first work on predicting semantic label of places through integrating sub-classification models into a multi-level structure.

Keywords - *Semantic Prediction, User Behavior, Feature Extraction, Multi-Level Classification Model.*

1. INTRODUCTION

With the increasing availability of smart phones, rapid development of location-based services [3], and growing interests in Web 2.0 services such as Foursquare (https://foursquare.com/) and Gowalla (http://gowalla.com/) have emerged. These services allow users to explore spatial information, search other users, and share their experiences with others. The amount of user-generated spatial information of smart phones is growing continuously. A lot of spatial information has been labeled with some semantic tags such as *sightseeing (tagged on trajectory)* or *restaurant (tagged on place)*. Many such tags can be seen on the website of EveryTrail and Foursquare like Figure 1 shows, which are crucial for assisting users in searching and exploring this massive spatial information as well as for developing place or trip recommendation services. However, based on our observation, most of spatial information lacks meaningful textual descriptions. To address this problem, we develop a novel technique for automatically and precisely semantic place prediction. Here the notion of "semantic place prediction" is a process to predict the semantic meaning of these places for a number of users.

The problem of semantic place prediction can be formulated as predicting appropriate semantic label for a given place. In the MDC Data Set [11], there are 10 possible semantic tags which are home, home of a friend, relative or colleague, my workplace/school, location related to transportation, the workplace/school of a friend, relative or colleague, place for outdoor sports, place for indoor sports, restaurant or bar, shop or shopping center, and holiday resort or vacation spot. To resolve the semantic place prediction, we will build a model to label the most possible tag on the place. Hence, semantic place prediction in MDC may be addressed as a *classification* problem [1]. While classification techniques have been developed for many applications, the problem has not been explored previously under the context of cell phone data, where we can only operate over user's cell phone logs such as MDC Data Set.

We propose to address the semantic place prediction problem by learning a several classification models. In order to precisely classify semantic places, a fundamental issue is to identify and extract a number of descriptive features for each place in MDC. Selecting the significant features is important because those features have direct impact on the effectiveness of the classification task. As mentioned earlier, the only data resource we have is the user's cell phone logs at various place and times. Therefore, we explore the user behaviors and seek unique features of places captured in the cell phone logs, which are stored in MDC Data Set. Fortunately, human behaviors usually follow several rules, e.g., people usually stay home for rest at around night, moving continuously when doing sport, charging their phones at indoor environment

To realize our observation into our classification model, we extract features of places in four aspects: 1) Spatial Property, 2)

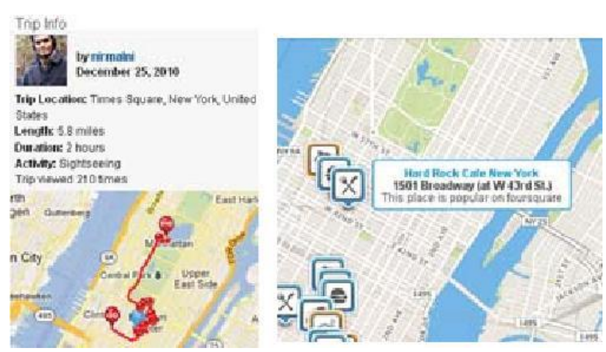

Figure 1. An example of semantic tag of spatial information.

Temporal Property, 3) Users' Behavior, and 4) Environment. Moreover, in order to seek best effectiveness, we utilize χ^2 statistic [5][7] to represent the importance of feature and cross validation to find best feature set for each classification model. Based on these validation results, we adopt the decision tree forecasting model to fusion these models' results to make predictions.

Besides, based on our observation, MDC Data Set is with class-imbalanced data problem, which is critical as discussed in [2][8]. Although there are 10 kinds of semantic label, most places are tagged with Home, Home of a friend, and my workplace/school. Thus the classification model will tend to predict a place as these three kinds of semantic label. To solve this problem, we propose a Multi-Level Classification Model, which divides original classification problem into several sub-problems for classification. For example, if a dataset consist of 10 raw data in which 5 are belong to class A and the remaining 5 are belong to class B, C, D, E, and F, respectively, the Multi-Level Classification Model will build a classification model to classify data into class A and "not A". Then, the Multi- Level Classification Model will build another classification model to classify data into class B, C, D, E, and F. In the testing step, the Multi-Level Classification Model will first classify testing data into class A or "not A". If the testing data is classify into class "not A", the low -level model will classify the testing data into class B, C, D, E, or F. By this way, we divide the original classification problem into several sub-classification problems and the main problem of class-imbalanced data is resolved through the sub-classification approaches. Accordingly, the accuracy of each sub -classification is improved and the total accuracy of classification is enhanced significantly.

Although semantic data mining in mobile data have been addressed in many our previous works [9][10], to our best knowledge, this is the first work that exploits both i) Behavior Features and ii) Environment Features in mobile data for semantic place prediction. The contributions of our research are three-fold:

We define 54 features to represent end users' behaviors in each place related to semantic labels, which consist of four aspects: 1) Spatial Property, 2) Temporal Property, 3) Users' Behavior, and 4) Environment.

We develop a new classification framework, namely Multi-Level Classification Model, which will not be affected by class-imbalanced data problem.

In our Multi-Level Classification Model, we fusion several existing classification model's result by decision tree forecasting model.

The remaining of this paper is organized as follows. We describe the Feature Extraction and Feature Selection from MDC in Section 2 and Section 3, respectively. The proposed Multilayer Modeling is detailed in section 4.

2. Feature Extraction

In this section we will introduce the features we extract from MDC Data Set and also show how we select features by χ^2 statistic. To represent each place's property, we argue that a place's type always reflects the environment of the place and user's behavior. For example, a place where people always have been at midnight is always home. Therefore, we extract and categorize the features we utilize for semantic place prediction in two aspects, behavior and environment.

2.1 Behavior Feature

We can observe four kinds of behavior in MDC Data Set. First, is end users' movement behavior, followed by phone usage behavior, then communication behavior, and finally temporal behavior. To reflect users' movement behavior, we extract the features as shown in follows.

Relative Visit Frequency: ratio of place visited times to all places visited times, indicating whether this place is always visited by the user.

Distance from Potential Home Location: geographical distance from the most visited place, indicating whether this place is far from the area where user lives in.

Average of Movement: average proper acceleration through this place, indicating whether the user changes his/her body posture frequently in this place.

Average of Movement Change: variation of proper acceleration.

Statistical Feature of Movement: this feature is adopted from Yan et al.'s work [6]. We use 3-axis value of accelerometer to extract the statistical features and 21 features are used here.

Calendar-time Frequency: times of visits time of the place that matches start time of calendar entry.

To reflect users' phone usage behavior, we extract the following features:

Application Usage Frequency: interaction times with application per hour, catching the interaction with phone, indicating whether this place is suitable for using the phone.

Kinds of Application Usage: total kinds of applications are used in this place, indicating the diversity of application usage.

Mediaplay Usage Frequency: media played times per hour, indicating whether this place is suitable for using the media files.

Process Usage Frequency: executed process number per hour.

Kinds of Process Usage: total kinds of processes are used in this place, indicating the diversity of executed processes.

To reflect users' communication behavior, we extract the following features:

Text-out Frequency: times of sending short messages per hour.

Call-out Frequency: times of making phone calls per hour.

Miss-call Frequency: number of missed call per hour.

To reflect users' temporal behavior, we extract the following features:

Relative Visit Frequency in Holiday: ratio of visited times on holiday to visited times on weekday. This feature is helpful in indicating workplace/school because workplace/school usually has low value in this feature.

Relative Visit Frequency in Hour of a Day: split one day into twelve time slot and count the place visited times of each slot. A total of 12 features are used here.
Average Stay Time: average stay time of this place.

2.2 Environment Feature

In fact, there are two kinds of environment feature in MDC Data Set. One is actively detecting environment, and another is inactively detecting environment. To reflect actively detecting environment, we extract the following features:
Bluetooth Count: ratio of numbers of Bluetooth devices seen by the user to place visited times. If user meets different people contiguously, this value will be very high because the user may encounter many Bluetooth devices carried by other people. This kind of phenomenon is significant in shop or shop center.

Diversity of Bluetooth: We can obtain a set of Bluetooth devices every time when user visits this place. For every two different visits. We compute the ratio of intersection to union of Bluetooth devices and average all the values as feature.
WLAN Count: ratio of numbers of WLAN devices seen by the user to place visited times.
Diversity of WLAN: We also can obtain a set of WLAN devices every time when user visits this place. For every two different visits. We compute the ratio of intersection to union of WLAN devices and average all the values as feature.
To reflect inactively detecting environment, we extract following features:
Proportion of Charging Time: we use system information of this place to compute the ratio of charging record to all record as the value of this feature. This feature is helpful in indicating the indoor type place.
Proportion of Mute Time: we also use system information of this place to compute the ratio of silent mode record to all record as the value of this feature.

3. Feature Selection

After extracting features of each place, 54 features are used in our work. The next step is to determine what kind of features should be used in our classification model. Based on Yang et.al.'s observation [7], the χ2 has excellent effectiveness for measuring importance of textual features. Since the semantic place prediction also focus on textual information, such as semantic label, we adopt the χ2 statistic to evaluate the association between features and class labels, and rank features according to their associations. Table 1 lists top 15 features in the ranking list as an example. In the ranking list of features, the first feature is considered the best feature for classification and the 54th feature is considered the worst one for classification. Due to this relation, we can use the ranking list to select what features should be kept or not. Our selection process is shown in Table 2.
In first step, we use first feature in ranking list to build a classification model, verifying by cross validation and record accuracy of this model. In the second step, we use the first and second feature to build a model, verifying and record accuracy. In

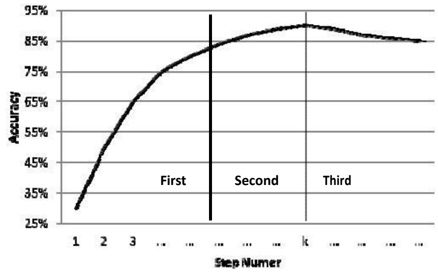

Figure 2. An example of feature selection.

the following step, we add 3rd feature and …nth feature to do the same thing and record accuracy on every step until all of 54 features are used in building model. After all, we use the feature composition of the highest accuracy from previous step to build the classification model. For example, in Table 2, if we have the highest accuracy on third step, then we use first, second and third feature in the ranking list to build the classification model.

Table 1. Chi-Square Statistic of Features

Rank	Feature Name	Chi-Square Statistic
1st	Statistical feature of movement_VH_Correlation	63.72906746
2nd	Statistical feature of movement_XZ_Correlation	9.421775751
3rd	Statistical feature of movement_Y_Mean	0.596468881
4th	Statistical feature of movement_YZ_Correlation	0.52044518
5th	Statistical feature of movement_XY_Correlation	0.470582807
6th	Statistical feature of movement_Y_Horizontal	0.364203378
7th	Statistical feature of movement_Z_Horizontal	0.333590205
8th	Statistical feature of movement_Z_Vertical	0.122604209
9th	Statistical feature of movement_Z_Mean	0.09815563
10th	Miss-call frequency	0.093157322
11th	Text-out frequency	0.088965829
12th	Statistical feature of movement_Y_Vertical	0.079930426
13th	Bluetooth count	0.04283207
14th	Relative Visit Frequency in 0~2 o'clock	0.030720924
15th	Relative Visit Frequency in 22~24 o'clock	0.030509035

Table 2. Feature selection process

Step Number	Use 1st feature	Use 2nd feature	Use 3rd feature	Use …th feature	Use 54th feature	Accuracy
1	O					50%
2	O	O				65%
3	**O**	**O**	**O**			**80%**
…	O	O	O	O		…%
54	O	O	O	O	O	60%

We argue that an excellent feature selection process will always have "log-like" curve of accuracy as shown in Figure 2, which can be described in three parts. In the first part of the curve, since the features on the top of ranking list are effective in classification, the curve of accuracy will rise rapidly in this part. The second part of the curve is rising slowly. This is because the features in the middle of list could only provide a little help for accuracy. Finally, in the third part of the curve, we can observe that the accuracy slightly starts to descend, so we can find out that step k has the highest accuracy. It means that features from 1 to k are effective and features from k+1 onwards could be noise. We take features from 1 to k as our best feature composition to classification.

We select features individually instead of select features in set. This is because even when a set of features performs well on classification, it could still has some noisy features, so we treat every feature equally and use feature selection to find out what are good features.

4. Multi-Level Classification Model

In this section, we propose a multi-level classification model to handle multi-class classification problem. Doing multi-class classification may be hard by using a single model, especially when the characteristic of each class label are not distinguishable. Nevertheless, it is easier to do classification when the characteristic of each class label have significant differences. Therefore, the main idea of our approach is that only one model is used at a time in dealing with one easy classification problem. To realize our idea, we split the complex classification problem into several easier classification problems, conquering all these easier problems and combined all the results to achieve higher accuracy of multi-class classification. In that case, what is important on multi-level classification model is the way to split the multi-class classification problem.

4.1 Model Building

The way to split the multi-class classification problem is determined by the characteristic of each class. We group ten class labels in a hierarchical way based on their characteristic and then build models on every level. The result is shown in Figure 3. As shown in Table 3 & 4, we manually build the model to make each label of training data with balanced size and different characteristic. In this way, the imbalance problem can be resolved for better classification result.

Table 3. Description of "Root" model

"Root" model	Size	Characteristic
Home	130	High "Relative Visit Frequency"
Workplace/school	102	High "Relative Visit Frequency" Low "Relative Visit Frequency in Holiday"
Other 7 labels	104	Low "Relative Visit Frequency"

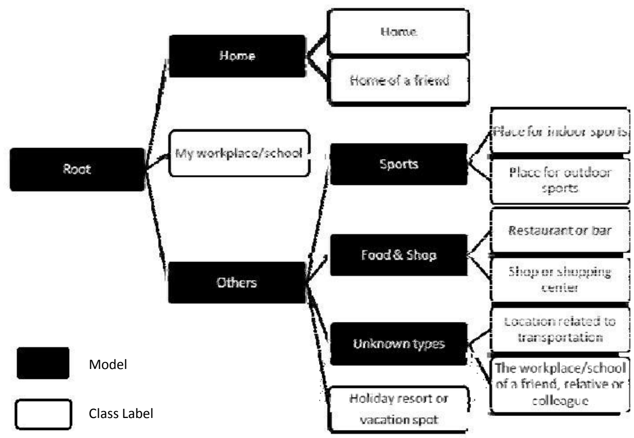

Figure 3. Multi-level classification.

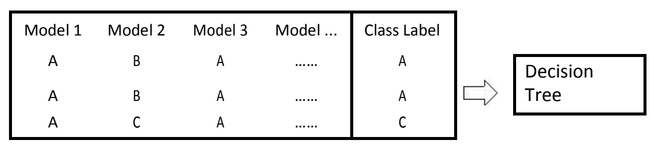

Figure 4. An example of Fusion Model.

In the Table 3, the "root" model first classifies places into three types that have significant differences on the feature of "Relative Visit Frequency" and "Relative Visit Frequency in Holiday". Home and workplace/school both tend to have high "Relative Visit Frequency" and workplace/school always have low "Relative Visit Frequency in Holiday". Therefore, we can easily get the right results on this model. If a place is classified to be "My workplace/school" on this model, then we take it as our answer of classification. Otherwise, we forward the place to the next level's model.

Table 4. Description of "Other" model

"Other" model	Size	Characteristic
Sports	39	Features of "Movement behavior"
Food&Shop	28	Features of "Actively detecting environment"
Unknown types	31	Unknown
Holiday resort or vacation spot	5	High "Relative Visit Frequency in Holiday" High "Distance from potential home location"

In the Table 4, if a place is forwarded to the "other" model, then it will be classified into four types. The differences between these four types are also significant. The sport type labels have the characteristic about the movement features. The food and shop type labels have the characteristic about the environment features. The Holiday resort or vacation spot has characteristic about "Relative Visit Frequency in Holiday" and "Distance from Potential Home Location". Two labels belong to unknown type; this is because it is hard to find out characteristic about them on this level.

The other four models are deal with 2-class classification. Therefore, it becomes an easier classification problem and we believe our features are good enough to perform well classification on these four model. Multi-level classification model is built by using the feature selection method in Section 3. In this way, we can achieve best accuracy on classification.

4.2 Fusion Model

Every type of classification models has its characteristic. They have advantages on classifying different kind of data. To integrate all their advantages, we use a fusion model, which has the ability to combine several kind of model. Figure 4 shows the way in which we build the fusion model.

First, we use several models (e.g., SVM, J48, etc) to build multi-level classification models. Second, we classify every training instance by these models and record their results combined with the original class label of instance into a table just like the left part of Figure 4. Finally, the way to integrate all the models can be a classification problem. We use this table as training data to build a classification model, discovering the association between the results of models and the real answer of place. Note that any classification model can be used for this process.

For example, on the first two row of table in Figure 4, when Model 1 and 3 think the answer is A and Model 2 think the answer is B, the real answer is A. However, on the row 3, Model 1 and 3 think the answer is A, this time Model 2 think the answer is C and the real answer is C. Therefore, we can use the association between the models to find out the real answer. This kind of association can be discovered by decision tree and improve accuracy of classification.

5. EXPERIMENTS

In this section, we conduct a series of experiments and use 10-fold cross validation to evaluate the effectiveness for the proposed model using MDC Data Set in terms of Accuracy and F-measure. The reason we consider F-measure as a measurement to evaluate our model is that it can reflect the ability to deal with class-imbalanced data problem. All the experiments are implemented in Java JDK 1.6 on an Intel i7 CPU 3.40GHz machine with 4GB of memory running Microsoft Windows 7. We present our results followed by discussions.

5.1 Effectiveness of Multi-Level Classification

Here we evaluate the F-measure of the single-level classification model and multi-level model classification. In column 1 and 3 of Table 5, we can see the different performance of 10 labels between two kind of approach and multi-level model outperform single-level model.

Table 5. Comparison of different strategy to build the model

	Fusion Single-Level	Multi-Level Classification Model	
		None fusion	fusion
Holiday resort or vacation spot	0	0	0
Home	0.84210526	0.847458	0.843931
Home of a friend, relative or colleague	0.55670103	0.555556	0.6
Location related to transportation	0.45833333	0.509804	0.5
My workplace/school	0.81690141	0.84058	0.84058
Place for indoor sports	0.25	0.275862	0.384615
Place for outdoor sports	0.2	0.358209	0.382353
Restaurant or bar	0	0	0.090909
Shop or shopping center	0.35714286	0.4	0.32
The workplace/school of a friend, relative or colleague	0	0	0.153846

5.2 Performance of Fusion Model

We tried several existing models to build multi-level classification model and preserved top four accurate multi-level classification models. In Table 6, we using different compositions from these four models to build the fusion model and find out the best composition on accuracy. We can see the best result is the composition of SMO and Simple Logistic. When we only use SMO for classification, the accuracy is 64.58% and Simple Logistic is 61.01%. However, when we use the fusion model to combine both models, it reaches a higher accuracy of 65.77%. This result show our fusion model is working. It can really combine advantages of both models to perform a better classification. The improvement of fusion model on all labels can be seemed in column 2 and 3 of Table 5. Finally, we use the best 5-fusion model to be our final fusion models for MDC task 1.

We also tried several classification models to fusion. In Table 7, we find out that tree-based model has better performance and at last, we adopt the decision tree model, REPTree, to fusion all the models.

Table 6. Accuracy of fusion models

SMO	J48	PART	SimpleLogistic	Accuracy
O				64.58%
	O			55.06%
		O		56.25%
			O	61.01%
O	O			55.06%
O		O		56.25%
O			O	65.77%
	O	O		56.55%
	O		O	55.06%
		O	O	56.25%
O	O	O		56.55%
O	O		O	55.06%
O		O	O	56.25%
	O	O	O	56.55%
O	O	O	O	56.55%

Table 7. Analysis of accuracy of different model on fusion

Fusion Model	REPTree	J48	RF	LMT
Accuracy	65.77%	64.58%	64.29%	63.99%
Fusion Model	SimpleLogistic	PART	SMO	
Accuracy	63.99%	63.39%	63.10%	

5.3 Effectiveness of all features on all labels

Table 8 shows the F-measure of six set of features on all labels. Every feature has its effective on different kind of labels except “Holiday resort or vacation spot”. Table 9 shows the confusion matrix for all 10 labels. An entry in the i^{th} row & j^{th} column denotes the fraction of $label_i$ instances which the classifier predicted as $label_j$. For example, if we predict 100 instances of home, 87 instances will be predict as home and 2 instances will be predict as My workplace/school.

6. CONCLUSIONS

In this paper, we propose the Multi- Level Classification model, a new approach for semantic place prediction. Meanwhile, we tackle the problem of users’ behavior and environment features extracted from MDC Data Set, which is a crucial prerequisite for effective prediction of semantic place. The core of task of semantic place prediction is a classification problem which classifies place into a semantic label by learning a classifier. In the proposed Multi-Level Classification model, we explore i) Behavior Features and ii) Environment Features by exploiting the MDC Data Set to extract descriptive features. To our best knowledge, this is the first work that exploits both i) Behavior Features and ii) Environment Features in mobile data for semantic place prediction. Through a series of experiments, we validate our proposal and show that the proposed semantic place prediction has excellent performance under various conditions. And we use the top 5 performance models to obtain the uploaded testing result.

Table 8. F-measure of 6 set of features for 10 semantic labels

	Movement behavior	Phone usage behavior	Communication behavior	Temporal behavior	Actively detecting environment	Inactively detecting environment	All the features
Holiday resort or vacation spot	0	0	0	0	0	0	0
Home	0.616161616	0.660377	0.4848485	0.823529	0.602041	0.704082	0.843931
Home of a friend, relative or colleague	0.126315789	0.055556	0	0.506667	0.323529	0	0.6
Location related to transportation	0.259259259	0.193548	0	0.377358	0	0	0.5
My workplace/ school	0.372670807	0.41	0.519337	0.855769	0.496552	0.478689	0.84058
Place for indoor sports	0.1	0	0.1052632	0.08	0	0	0.384615
Place for outdoor sports	0.106666667	0.066667	0	0.342857	0	0	0.382353
Restaurant or bar	0	0	0	0	0	0	0.090909
Shop or shopping center	0.461538462	0	0	0	0.230769	0	0.32
The workplace/ school of a friend, relative or colleague	0	0.166667	0	0	0	0	0.153846
Accuracy	35.12%	36.31%	38.10%	61.90%	43.15%	42.26%	65.77%

Table 9. Confusion matrix for 10 semantic labels

Holiday resort or vacation spot	**0.00**	0.20	0.00	0.00	0.00	0.00	0.80	0.00	0.00	0.00
Home	0.00	**0.87**	0.05	0.00	0.02	0.00	0.02	0.02	0.01	0.00
Home of a friend, relative or colleague	0.02	0.20	**0.52**	0.04	0.02	0.07	0.13	0.00	0.00	0.00
Location related to transportation	0.00	0.00	0.09	**0.57**	0.17	0.00	0.09	0.04	0.00	0.04
My workplace/school	0.00	0.06	0.01	0.03	**0.85**	0.01	0.02	0.01	0.00	0.01
Place for indoor sports	0.00	0.00	0.00	0.14	0.14	**0.36**	0.29	0.00	0.07	0.00
Place for outdoor sports	0.00	0.00	0.08	0.12	0.00	0.08	**0.52**	0.16	0.00	0.04
Restaurant or bar	0.00	0.00	0.00	0.00	0.27	0.00	0.45	**0.09**	0.18	0.00
Shop or shopping center	0.00	0.00	0.06	0.18	0.12	0.06	0.24	0.12	**0.24**	0.00
The workplace/school of a friend, relative or colleague	0.00	0.00	0.00	0.33	0.44	0.00	0.11	0.00	0.00	**0.11**

REFERENCES

[1] J. Han and M. Kamber. "Data Mining, Conecpts and Techniques", Morgan Kaufmann Publishers, 2001..

[2] T. Jo and N. Japkowicz. "Class Imbalances versus Small Disjuncts."SIGKDD Explorations, 6(l):40-49, 2004.

[3] E. H.-C. Lu and V. S. Tseng. Mining Cluster-Based Mobile Sequential Patterns in Location-Based Service Environments. Proceedings of IEEE International Conference on Mobile Data Management (MDM), Taipei, Taiwan, May, 2009.

[4] A. Papoulis and S.U. Pillai. Probability, Random Variables and Stochastic Processes. McGraw-Hill, New York, NY, 2002.

[5] S.M. Ross. Introduction to Probability and Statistics for Engineers and Scientists. Wiley, New York, NY, 2004.

[6] Z. Yan, D. Chakraborty, A. Misra, H. Jeung, K. Aberer "Semantic Activity Classification Using Locomotive Signatures from Mobile Phones." *EPFL-REPORT*, 2012.

[7] Y. Yang and J. Pedersen, A comparative study on feature selection in text categorization. Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97), pp. 412-420, 1997.

[8] S.-J. Yen, Y.-S. Lee, C.-H. Lin and J.-C. Ying, Investigating the Effect of Sampling Methods for Imbalanced Data Distributions, *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics (SMC'2006)*, pp. 4163-1468, October 2006.

[9] J. J.-C. Ying, E. H.-C. Lu, W.-C. Lee, T.-C. Weng, V. S. Tseng. Mining User Similarity from Semantic Trajectories. In Proceedings of ACM SIGSPATIAL International Workshop on Location Based Social Networks (LBSN' 10), San Jose, California, USA, November 2, 2010.

[10] J. J.-C. Ying, W.-C. Lee, T.-C. Weng, V. S. Tseng. Semantic Trajectory Mining for Location Prediction. In Proceedings of ACM SIGSPATIAL International Workshop on Location Based Social Networks (LBSN' 11), Chicago, IL, USA, November 1, 2011.

[11] J. K. Laurila, D. Gatica-Perez, I. Aad, J. Blom, O. Bornet, T. Do, O. Dousse, J. Eberle, and M .Miettinen, "The Mobile Data Challenge: Big Data for Mobile Computing Research" in Proc. Mobile Data Challenge Workshop, Newcastle, Jun. 2012.