Demographic Prediction Based on User's Mobile Behaviors

Josh Jia-Ching Ying, Yao-Jen Chang, Chi-Min Huang and Vincent S. Tseng*
Department of Computer Science and Information Engineering
National Cheng Kung University
No.1, University Road, Tainan City 701, Taiwan, R.O.C.
{ jashying, autek.roy, qulvmp6}@gmail.com, *Correspondence: tsengsm@mail.ncku.edu.tw

ABSTRACT

In this paper, we propose a novel prediction framework for predicting end users' demographic by taking into account the users' behavior and environments at the same time. The core idea of our proposal is to extract key features to represent end users' behaviors in each location related to the users' demographic. To achieve this goal, we define 45 features to represent end users' behaviors and environment for capturing the key properties of locations recorded in MDC Data Set. In our framework, we propose a novel model, namely *Multi-Level Classification Model*, to solve the imbalanced class problem existing in the data. Based on the *Multi-Level Classification Model*, we make demographic prediction of an end user by combining several classification models. To our best knowledge, this is the first work on predicting end users' demographic by combining several classification models into a multi-level structure.

Categories and Subject Descriptors

H.2.8 [Database Management]: Database Applications – *Data Mining,*

General Terms

Experimentation, Performance, Human Factors

Keywords

Demographic Prediction, User Behavior, Feature Extraction, Multi-Level Classification Model

1. INTRODUCTION

With the increasing availability of smart phones, rapid development of location-based services, and growing interests in web 2.0 services such as Gowalla, Foursquare and Facebook have emerged. These services allow users to explore places, search other users, and share their experiences with others. The number of users of smart phones is growing continuously. Many users have registered account with their demographic such as *occupation* or *gender*, which are crucial for assisting users in searching and exploring new users as well as for developing friend recommendation services. However, based on our observation, most of users are lacking any meaningful textual descriptions. To address this problem, we develop a novel technique to automatically and precisely predict users' demographic.

The problem of demographic prediction of MDC can be formulated as two prediction problem and three classification problems for a given data of users. In MDC Data Set, an end user has 5 kinds of demographic attribute which are gender, job, marital status, age group, and number of people in the household. Hence, demographic prediction in MDC may be addressed as a *multi-target-attribute forecasting* problem [4]. While multi-target-attribute forecasting 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 multi-target-attribute forecasting problem by learning several classification models. To do so, a fundamental issue is to identify and extract a number of descriptive features for each location in MDC. Selecting the significant features is important because those features have a 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 location and times. Therefore, we explore the user behaviors and seek unique features of locations 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.

To realize our observation into our classification model, we extract features of locations in two aspects: 1) Users' Behavior and 2) Environment. We seek the importance of feature and use cross validation to find best feature set for each classification model. Based on these validation results, we adopt the decision tree forecasting model to fuse these models' results to make prediction. Besides, based on our observation, MDC Data Set is a class-imbalanced data [3] except for the attribute of gender. This issue is also dealt with in our work.

The contributions of our research are three-fold:

- We define several key features to represent end users' behaviors in each location related to their demographic, including two aspects: 1) Users' Behavior and 2) Environment.
- We develop a new classification framework, namely Multi-Level Classification Model, which is insensitive to the imbalanced data problem.
- In our Multi-Level Classification Model, we fuse several existing classification model's result by decision tree forecasting model.

The remaining of this paper is organized as follows. In Section2, we present related work. We describe the Feature Extraction and Feature Selection from MDC in Section 3 and Section 4, respectively. The proposed Multilayer Modeling is detailed in section 5. The experimental results are shown in section 6 to prove out our idea. Finally the conclusion and future work are described in Section 7.

2. RELATED WORKS

Research on demographic prediction area mainly focuses only on modeling the linguistics writing and speaking styles. Some research worked on classifying the user's gender by the spoken

This material was prepared for the Mobile Data Challenge 2012 (by Nokia) Workshop; June 18-19, 2012; Newcastle, UK.

language difference including phonological, intentional and conversational cues [8][9].Some research worked on the male and female writing styles in formal contexts such as books and articles. Berryman-Fink [10] and Simkins-Bullock [11] investigated these writing style and found there is no significant difference between male and female writing styles. Biber studied the difference between male and female in language structure using on correspondence corpus [6]. Palander investigated the male and female styles [7].

Recently, a number of works consider that users with similar demographic information would visit similar WebPages.Hu et al studied and propose an approach predict users' gender and age from their Web browsing behaviors [5]. Based on the facts of WebPages visited by similar users, Hu et al, use the WebPages visited to predict demographic tendency. However, all of above-mentioned works do not capture individual movement behavior for demographic prediction. As far as we know, there is no any work on predicting users' demographic attributes by extracting useful information from the data of smart phone, like MDC Data Set.

3. Feature Extraction

In this section we will introduce the feature we extract from MDC Data Set. To represent the each user's property, we argue that user's demographic attributes always reflect to her behavior and environment. For example, movement of young people may be more fluctuating than old people. Therefore we extract and categorize the features we utilize for demographic prediction in two aspect, behavior and environment.

3.1 Behavior Feature

Actually, we can observe three kinds of behavior in MDC Data Set. First is end users' movement behavior, second is phone usage behavior, and third is communication behavior. To reflect users' movement behavior, we extract the features as shown in follows.

- **Maximum Movement in a Location:** mean of maximum distance of movement in a location
- **Average Movement in a Location:** mean of average distance of movement in a location
- **Average Movement Change in a Location:** mean of average velocity of movement in a location
- **Maximum Distance from Home Location:** maximum value of geographical distance from the most visited place, (Here the Home Location is the place where the user most frequently visits)
- **Average Distance from Home Location to other Locations**: Average value of geographical distance from the most visited place,

To reflect users' phone usage behavior, we extract the features as shown in follows.

- **Kinds of Application Usage per Day**: average total kinds of application is performed by the end user per day
- **Number of Application Usage per Day:** average number of applications is performed by the end user per day
- **Kinds of Process Usage per Day**: average value of total kinds of process is performed by the end user per day
- **Number of Process Usage per Day**: average number of processes is performed by the end user per day.
- **Calendar Event Creation per Day**: average number of created events per day
- **Calendar Usage per Day:** average number of events, including created event and modified event, per day
- **Kinds of Music Listening per Day:** average value of total kinds of songs is played by the phone per day
- **Number of Music Listening per Day**: average number of songs is played by the phone per day
- **Average Playing Time per Music**: the ratio of total playing time to the number of played music
- **Number of Media in Phone**: number of kinds of media stored in phone
- **Average Size of Media:** ratio of total size of all the media to the number of media

To reflect users' communication behavior, we extract the features as shown in follows.

- **Text Usage per Day**: Average number of sent and received text
- **Number of Call per Day**: Average number of calls, including incoming call, outgoing call and missed call, per day
- **Number of Call-in per Day**: Average number of incoming call per day
- **Number of Call-out per Day**: Average number of outgoing call per day
- **Number of Miss-call per Day**: Average number of missed call per day
- **Proportion of Miss-call (# of Miss-calls / # of Calls)**: the ratio of number of all missed calls to number of all calls
- **Number of Text with Contacts per Day**: similar to the Text Usage per Day, but we just count the phone numbers are appear in the phonebook. (Here, the contacts means the phone numbers appear in the phonebook)
- **Number of Call with Contacts per Day**: Average number of calls of contacts in phonebook per day
- **Number of Call-in with Contacts per Day**: Average number of incoming calls of contacts in phonebook per day
- **Number of Call-out with Contacts per Day**: Average number of outgoing calls of contacts in phonebook per day
- **Number of Miss-call with Contacts per Day**: Average number of missed calls of contacts in phonebook per day
- **Proportion of Miss-call with Contacts**: the ratio of number of all missed calls of contacts in phonebook to the number of all calls of contacts in phonebook
- **Number of Text with Top 1 Contact per Day:** Average number of text, including sent and received text, with Top 1 Contact (Here, the top 1 contact means the phone number is appear in the phonebook of the cell phone and the most frequently contact to him)
- **Number of Call with Top 1 Contact per Day**: Average number of calls, including incoming call, outgoing call and missed call, with the top 1 contact per day
- **Number of Call-in with Top 1 Contact per Day**: Average number of incoming call with the top 1 contact per day

- **Number of Call-out with Top 1 Contact per Day**: Average number of outgoing call with the top 1 contact per day
- **Number of Miss-call with Top 1 Contact per Day**: Average number of missed call with the top 1 contact per day
- **Proportion of Miss-call with Top 1 Contact**: the ratio of Number of Miss-call with Top 1 Contact to the number of all missed call with Contacts
- **Proportion of Call-in with Top 1 Contact**: the ratio of Number of Call-in with Top 1 Contact to the number of all incoming call with Contacts
- **Proportion of Call-out with Top 1 Contact**: the ratio of Number of Call-out with Top 1 Contact to the number of all outgoing call with Contacts
- **Proportion of Text-in with Top 1 Contact**: the ratio of number of received text with Top 1 Contact to the number of all received text with Contacts
- **Proportion of Text-out with Top 1 Contact**: the ratio of number of sent text with Top 1 Contact to the number of all sending text with Contacts

3.2 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 features as shown in follows.

- **Kinds of Bluetooth Device Detected per Day:** Average value of total kinds of Bluetooth Device Detected per day
- **Number of Bluetooth Device Detected per Day:** Average number of Bluetooth Device Detected per day
- **Average Similarity of Bluetooth Device between Home Location and other Location**: 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 this feature.
- **Kinds of Wireless Device Detected per Day:** Average value of total kinds of Bluetooth Device Detected per day
- **Number of Wireless Device Detected per Day:** Average number of Wireless Device Detected per day
- **Average Similarity of Wireless Device between Home Location and other Location**: We can obtain a set of Wireless 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 this feature.

To reflect users' phone usage behavior, we extract the features as shown in follows.

- **Average Mute Time per Week**: Average value of total mute time, which means the phone is in silent mode, per week
- **Average Staying Time per Location**: the ratio of total duration record in data to the number of location

4. Feature Selection

After extracting features of each place, a total of 45 features are used in our work. The next step is to determine what kind of features should be used in our classification model. In order to seek the best effectiveness, we utilize χ^2 statistic [1][2] to represent the importance between features and class labels and cross validation to find best feature set for each classification model. Then, we rank features according to their associations for 5 kinds of demographic attribute. In the ranking list of features, the 1st feature is considered to be the best feature for classification and the 45th feature is considered to be the worst one for classification. Due to this relation, we can use the ranking list to select what features should be kept or not, and there is a unique ranking list in each attribute.

In first step, we use 1st feature in ranking list to build a classification model, verifying by cross validation and record accuracy of this model. In second step, we use 1st and 2nd feature to build a model, verifying and record accuracy. In the following step, we add 3rd feature and the feature to do the same thing and record accuracy on every step until all of 45 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. Finally, we can find the number of features in the building model with best performance.

5. Multi-Level Classification Model

In this section, we propose a multi-level classification model to handle multi-class classification problem of MDC task 3. Doing multi-class classification may be hard for a model, especially when the characteristic of each class label are not distinguishable. But it's easier to classify when the characteristic of each class label have significant differences. So the main idea of our approach is that one model only deal with one easy classification problem at one time. To fit our idea, we split the complex classification problem of MDC task 3 into several easier classification problems, conquering all these easier problems and combined all the results to achieve higher accuracy of multi-class classification. So what's important on multi-level classification model is the way how to split the multi-class classification problem.

For example, in the demographic "age group", most people belong to group "2" and "3". It will lead the classification model tense to predict the answers in these two age groups. To solve this problem, we propose a Multi-Level Classification Model which divides original classification problem into several classification sub problem. For example, if a data set 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, our model will work on imbalanced data, like MDC Data Set.

5.1 Model Building

The way how to split the multi-class classification problem is determined by the similarity between each class. For each demographic attribute, we group class labels of MDC task 3 in a hierarchical way based on their characteristic, then building models on every levels. We manually build the model to make each label of training data with balanced size. It is obvious the training data is imbalance except the "gender". In this way, the imbalance problem can be resolved for better classification result.

To build multi-level classification models, in each level we test several models (i.e., SVM, J48, etc) and use the cross validation to find the best performance with the on each model we tested. Using the method described in the feature selection section, we can find the best feature set in each model, so the different level may use different model with different number of features. Finally, the way to integrate all the models can be a classification problem using multi-level classification.

We tried several type of existing model to build multi-level classification model and preserved top 2 accurate multi-level classification models on each level.

5.2 Description for each demographic Multi-Level Classification Model

Take the demographic "job" as an example. The Fig. 1 shows our proposed Multi-Level Classification Model, which consists of two classifiers. The model A is first classifying end users into three types, "Ph. D. Student", "Employee without executive function" and "OTHERS". This grouping strategy is based on the distribution of class of data.

As shown in the Tab.1 & Tab.2, we build the model with balanced size on each level. Hence we can easily get the right results on this model. If a user is classified to be "Ph. D. Student" or "Employee without executive function" on this model, then we take it as our answer of classification. Otherwise we forward the user to the next level's model. If a place is forwarded to the model B, then it will be classified into two types. The differences between these two types are also significant and the sizes of these two types are the same such that it is easily to classify well. By this way, the model could predict demographic attributes more precise and ignore the effect of imbalanced data problem.

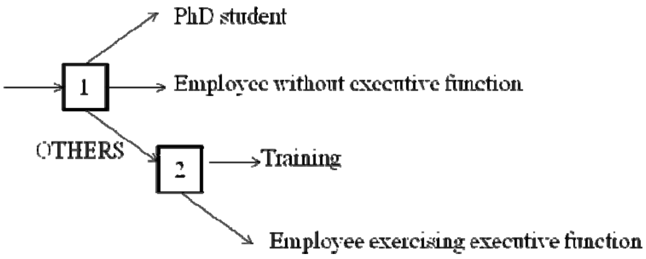
PhD student
1
Employee without executive function
OTHERS
2
Training
Employee exercising executive function

Fig. 1. Multi-Level Classification Model for "job"

Tab. 1. First Level model for "job"

class label	data size
PhD student	24
Employee without executive function	29
OTHERS	20

Tab. 2. Second Level model for "job"

class label	data size
Employee exercising executive function	10
Training	10

The other four models for remaining four demographic are built based on the same functionality and simply show in the Fig.2, Fig.3, Fig.4 and Fig.5. So 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 4. In this way, we can achieve best accuracy on classification.

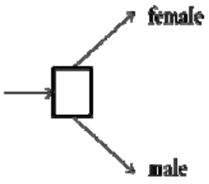

Fig. 2. Classification Model for "gender"

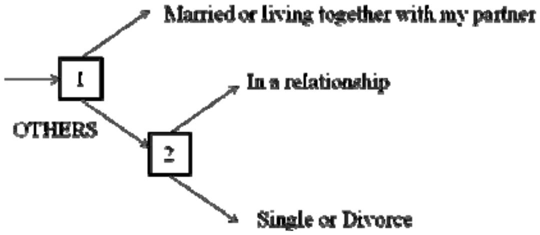

Fig. 3 Classification Model for "marital status"

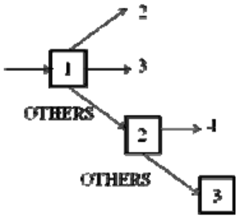

Fig. 4. Classification Model for "age group"

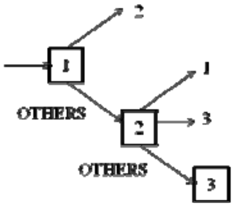

Fig. 5. Classification Model for "number of people"

6. Experimental Result

In this section, we conduct a series of experiments to evaluate the performance for the proposed Multi-Level Classification Model using MDC dataset. 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 first introduce the evaluation methodology and then we present our results followed by discussions. Due to the limitation of paper space, we can not put all of experiment results in this section. We will make deeply discussion on the demographic "marital status" in Section 6.1. The overview of performance of the remaining four models are given in Section 6.2

6.1 Experimental results of "marital status"

We discuss the experimental results of "marital status" in three aspects. First we focus on the effect of our proposed features and feature selection. Then we experiment the effectiveness of existing classification model (Single-Level) with our proposed features. Finally, we describe the effectiveness of our proposed Multi-Level Classification compared with Single-Level Classification.

6.1.1 Impact of Feature Selection

First, we would to find the best features sets in these models. See the Fig.6 & Fig.7 we tests every feature set to find the best two performance models with the best feature. We can see the accuracy is not low at the beginning point and the accuracy tend to better and better so our feature extraction and feature selection does work. It is because that our feature selection is effective. Moreover, event we use only one feature the accuracy is greater than 40%. It shows our proposed features are effective too.

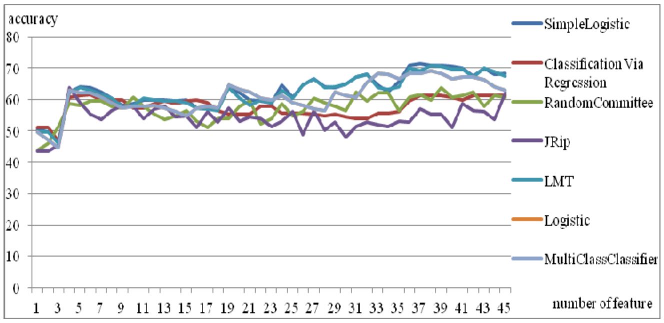

Fig.6. accuracy curve of feature selection in first level

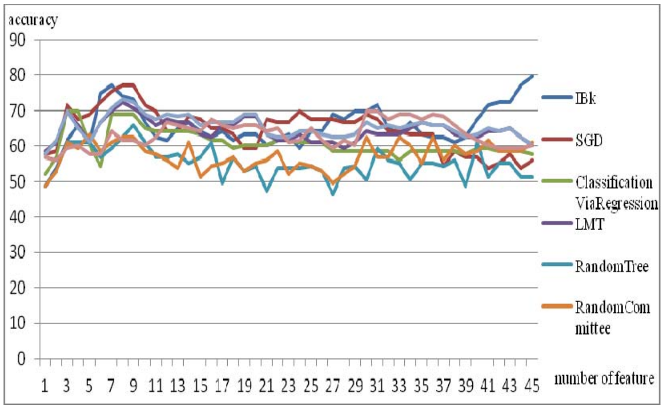

Fig.7. accuracy curve of feature selection in second level

6.1.2 Impact of Classification Model

We tried several existing models to build single-level classification model and preserved top four accurate single-level classification models. In Tab.3, we use 7 existing model on first level of "marital status" model and find out the best single on accuracy. We can see the best two results appear in Simple Logistic and Classification Via Regression. In Tab.4, we also use 7 existing model on second level of "marital status" model and find out the best single on accuracy. We can see the best two results appear in IBk and Random Committee.

Finally, we use the best 2 single-level models to construct our multi-level models for MDC task 3.

Tab. 3. best models for "marital status" in first level

Model	accuracy
Simple Logistic	68.35%
Classification Via Regression	67.08%
Random Committee	65.82%
Jrip	65.82%
LMT	65.82%
Losgistic	62.86%
Multi Class Classifier	62.86%

Tab. 4. best models for "marital status" in second level

Model	accuracy
IBk	79.67%
RandomCommittee	60.97%
LMT	60.16%
SimpleLogistic	60.16%
SMO	60.16%
ClassificationViaRegression	57.72%
SGD	56.09%
RandomTree	51.21%

6.1.3 Effectiveness of Multi-Level Classification

Finally, we show the difference between the Multi-Level Classification Model and directly classification model to prove out the effect of Multi-Level Classification Model. Comparing the two methods, we can see the different performance in Tab.5 & Tab.6. These two tables is created by the above series and show the best performance with the best feature sets. When we only use Logistic for classification, the accuracy is 58.22% and LMT is 57.38%. However, when we to make a two level model, it reaches a higher accuracy in each level. We can see the best accuracy is 71.3% in level 1 and 79.67% in level 2. This result show our Multi-Level Classification Model is working. It does really solve the imbalance problem.

Tab. 5. Multi-Level Classification Model for "marital status"

level 1 models	**number of using features**	**accuracy**
Simple Logistic	37	71.30%
LMT	38	70.88%
level 2 models	**number of using features**	**accuracy**
IBk	45	79.67%
SGD	8	77.23%

Tab.6.Single-Level classification model for "marital status"

model	**number of using features**	**accuracy**
Logistic	7	58.22%
LMT	7	57.38%

Tab.7. Model for "gender"

models	**number of features used**	**accuracy**
Classification via Regression	22	85.47%
Decision Stump	11	82.05%

6.2 Multi-Level Classifications Overview

The remaining four results are simply show in the Tab.5, Tab.6, Tab.7 and Tab.8. We can see the improved accuracy when we use the Multi-Level Classification Model in the Tab.6, Tab.7 and Tab.8.The Tab.5 shows the accuracy of the "gender", and it's the only one demographic attribute doesn't have the imbalance problem.

Tab.8. Multi-Level Classification Model for "job"

level 1 models	**number of features used**	**accuracy**
JRip	36	45.20%
Random Forest	33	42.92%
level 2 models	**number of features used**	**accuracy**
LWL	41	83.33%
Random Sub Space	10	78.33%

Tab.9. Multi-Level Classification Model for "age group"

level 1 models	**number of features used**	**accuracy**
J48	1	54.16%
Random SubSpace	1	50.83%
level 2 models	**number of features used**	**accuracy**
Attribute Selected Classifier	4	77.77%
Decision Table	4	77.77%
level 3 models	**number of features used**	**RMSE**
Additive Regression	5	2.036%
M5P	2	2.16%

Tab.10. Multi-Level Classification Model for "people"

level 1 models	**number of features used**	**accuracy**
Naïve Bayes	1	58.22222
Naïve Bayes	15	57.77778
level 2 models	**number of features used**	**accuracy**
Multilayer Perceptron	6	62.60163
SMO	33	62.60163
level 3 models	**number of features used**	**RMSE**
Linear Regression	33	0.385568
SMOreg	40	0.533976

The Tab.8 shows the accuracy of the "job". This Multi-Level Classification Model is similar to the above example for "marital" (See the Fig.1 & Fig.3). For the "age group" and "number of people in the household", Our goal is the value to do the RMSE. Otherwise, they both are three level (seeing Fig.4 & Fig.5) because they still have the imbalance problem in the second level. Therefore, the first and second level do the classification job and the third level would predict the last value.

7. CONCLUSION AND FUTURE WORK

In this paper, we propose the Multi-Level Classification model, a new approach for demographics 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 demographics. The core of task of demographic prediction is a classification or prediction problem which classifies user into a demographic group by learning a classifier. In the proposed Multi-Level Classification model, we explore i) Behavior Features and ii) Environment Features by exploiting the MDC Dataset 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 demographics prediction has excellent performance under various conditions. And we use the top 5 performance models to obtain the uploaded testing result.

8. ACKNOWLEDGMENTS

This research was supported by National Science Council, Taiwan, R.O.C. under grant no. NSC100-2218-E-006-017and by NCKU Top University project of Ministry of Education, Taiwan.

REFERENCES

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

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

[3] 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.

[4] M.-L. Zhang and Z.-H. Zhou. A k-nearest neighbor based algorithm for multi-label classification. In *IEEE International Conference on Granular Computing (GrC)*, pages 718–721, 2005

[5] J. Hu, H.-J. Zeng, H. Li, C. Niu, and Z. Chen. Demographic prediction based on user's browsing behavior. In Proc. 16th WWW, pages 151–160, 2007.

[6] Biber, D., S. Conrad, R. Reppen (1998). Corpus Linguistics Investigating Language Structure and Use, Cambridge University Press, Cambridge, 1998.

[7] Palander-Collin, M. (1999). Male and female styles in 17th century correspondence, Language Variation and Change 11, pp. 123-1

[8] Holmes, J. (1993). Women's talk: The question of sociolinguistic universals, Australian Journal of Communications 20, 3, 1993.

[9] Eckert, P. (1997). Gender and sociolinguistic variation, in J.Coates ed., Readings in Language and Gender, Blackwell,Oxford 1997, pp. 64-75.

[10] Berryman-Fink, C. L., J. R. Wilcox (1983). A multivariate investigation of perceptual attributions concerning gender appropriateness in language, Sex Roles 9, 1983.

[11] Simkins-Bullock, J. A., B. G. Wildman (1991). An investigation into the relationship between gender and language, Sex Roles 24, 19