CPL: Enhancing Mobile Phone Functionality by Call Predicted List

Santi Phithakkitnukoon and Ram Dantu

Dept. of Comp. Sci. & Eng., University of North Texas, Denton, TX 76203, USA
{santi,rdantu}@unt.edu

Abstract. In this paper, we present a concept of a new advanced feature for a mobile phone that provides its user functionality for predicting future calls. The feature is envisaged as a Call Predicted List (CPL) which makes use of the user's call history to build a probabilistic model of calling behavior based on the caller's calling patterns and reciprocity. The calling behavior model is then used to generate a list of numbers/contacts that are the most likely to be callers in the next hour. The performance of the CPL is evaluated with the real-life call logs and it shows promising results in accuracy.

Keywords: Context-aware computing, Call prediction, Mobile phone.

1 Introduction

With the rapid development of telecommunication technologies and the fast-growing number of users on the networks, the mobile phone has moved beyond being a mere technological object and has become an integral part of many people's lives. The mobile phone is gradually becoming the ubiquitous computing device at this early stage of the pervasive-computing era where handheld devices are precursors to a phase of ambient computing that is always on, personalized, context-sensitive, and highly interactive.

Mobile phones record the history of our lives in the form of the call logs. Utilizing these call logs in computing human (user)'s behaviors can indeed enhance the capability of the mobile phone as it is becoming more than just a communication device but also an intelligent assistant to its user.

In this paper, we present a novel model for predicting future callers using calling patterns. In this way, the mobile phone becomes more responsive and sensitive to the user's context and needs. With our proposed model, the personal phone will become more intelligent as it learns the user's behavior over time as well as the behavior of those who call the user in order to provide the most accurate prediction possible of the future incoming caller for the user upon his/her request. The rest of this paper is structured as follows: Section 2 presents the concept of the Call Predicted List (CPL), Section 3 presents the CPL's framework which describes the behavior learning model, Section 4 discusses the performance of the CPL, and the paper is concluded in Section 5 with a summary and an outlook on future work.

R. Meersman, Z. Tari, and P. Herrero (Eds.): OTM 2008 Workshops, LNCS 5333, pp. 571–582, 2008.

2 Call Predicted List

The Call Predicted List (CPL), described here, is intended to provide a phone user with an ability to predict future incoming calls as well as an improvement over the "last received calls" functionality that is often provided on today's phones and communication clients (e.g. VoIP soft phones).

Quite often in our daily lives, we find ourselves in a situation where we wish to know who will be calling in the next hour so we could schedule (plan) things out accordingly. In many occasions that we know for certain that we will be unavailable to accept any incoming calls over the next hour (e.g. having a flight, attending a class, having a meeting) thus we wish to know who will be calling during the next hour so we could perhaps make a call to the persons to inform of our next-hour schedule as we do not wish to miss any important future calls which could be too important calls to miss.

The user interface on a today's mobile phone normally provides easy access to a list of recently received numbers (contacts). The list provided in this case is insensitive to the user's context. It only shows the most recently received numbers and therefore takes no account of other call related information (e.g. time, day of week, frequency, etc) to provide a better guess of the numbers that the user will find most useful.

Our CPL makes use of the user's call history, i.e. call numbers received, time of call received, day of call received, frequency of call, and last dialed numbers, to build a probabilistic model of calling behavior. The calling behavior model is then used to generate a list of numbers/contacts that are the most likely to be the callers for the next hour. The list can be presented to the user in a number of different ways for different purposes. We envisage the CPL as an "intelligent call predicted list," i.e. a list that anticipates the numbers/contacts that the user will receive in the next hour and gives these numbers (potential callers) higher precedence on the list. Figure 1(a) shows an example of the CPL where the most likely callers are listed higher on the list.

3 Call Predicted List Framework

In our daily life, when we receive a phone call, at the moment of the first phone ring and right before looking at the caller ID, we often guess who the caller might be. We often base this estimation on the caller's calling pattern and our past communications to the caller.

Each caller tends to have a unique calling pattern. This pattern can be observed through history of *time of calls*, i.e. we normally expect a call from someone who has history of making several calls at some particular time of day. For example, your spouse likes to call you while you drive to work in the morning therefore when your phone rings while you are on the way to work you are likely to guess that it is a phone call from your spouse. We also base our estimation on day of calls, for example, your close friend has made several calls to you on every Tuesday because it is his day off

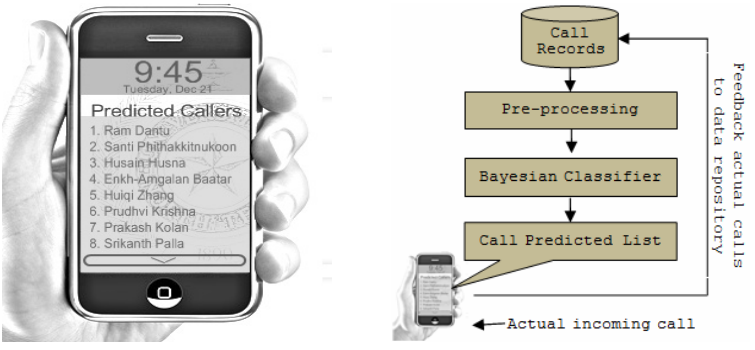

Fig. 1. (a) CPL user interface, **(b)** Basic system overview

therefore when your phone rings on Tuesday, the first person that comes to mind is your close friend. Similarly, the person who has made the most calls (*total call count*) to you (regardless of time and day) among other callers is also the person whom you most anticipate the calls from. Receiving a call is also influenced by the *reciprocity* or call interaction between user and caller. For example, you may expect a call from your friend based on your last phone conversation with him/her (e.g. "call me when you get home" or "call me same time tomorrow" or "I'm busy right now, call me back in an hour"). This reciprocity may sequentially lead to later receiving calls from that caller caused by your initiative. For example, you make a call to a friend to whom you have not called for a long time, and then you later receive calls back from this friend. Another example, you make a call to your mother to get some advice during the night (assume that you do not normally make or receive calls from her during this time), and then you receive calls from your mother later on during that night.

These are examples that actually happen in our everyday life for most of us who are phone users. Understanding the actual human behavior towards phone usage gives CPL an intelligence to assist its user effectively.

To predict the future incoming calls, the behavior learning model must be used. This model should incorporate mechanism for capturing caller's calling behavior. Calling behavior of the caller can be observed via the call logs which can be obtained from a variety of sources. For example, they may be collected by a network or service operator for billing purposes or they may be captured directly on device such as a mobile phone or on a software application such as a VoIP softphone.

In our current implementation, we use a set of actual call logs collected from 20 mobile phone users at the University of North Texas. These 20 individuals are faculty, staff, and students. We are in process of collecting several more call logs and make them publicly available for other researchers who have interests. This call logs collecting process is a continuation of the Nuisance Project [2], where Kolan et al. studied the nuisance level associated with each phone call. The details of the data collecting process are given in [3].

As part of the data collecting process, each user downloaded three months of detail telephone call records from his/her online accounts on the mobile phone service provider's website. Each call record in the dataset had 5-tuple information as follows.

- Date – data of call
- Start time – start time of call
- Type – type of call, i.e. "Incoming" or "Outgoing"
- Call ID – caller/callee identification
- Talk Time – duration of call (in minutes)

The call record is subject to pre-processing to extract features or information about *time of calls* (day and hour), *total call count*, and *reciprocity*.

The pre-processed call records are eventually fed into the classifier to be ingested. Classifier then outputs a list of phone numbers ordered by the predicted likelihood of the number being the next-hour caller given time of calls, day of calls, total call count, and reciprocity. The basic system overview is shown in Fig. 1(b).

Classifier has two modes of operation; training and predicting. During the training, classifier ingests the pre-processed call logs and constructs four hash tables which primarily contain call counts and corresponding callers. The first table maps each unique telephone number (or caller identifier) to a count of calls received for each day of the week. The second table maps each unique telephone number to a count of calls received for each hour of the week. The third table maps each unique telephone number to the total number of calls received.

It is not trivial to quantify the *reciprocity*. Having no knowledge about the context of the phone calls from the user to the callers, it is difficult to identify which outgoing calls may influence future incoming calls. Nevertheless a received call can be linked to user's calling behavior which is recorded in the "last dialed calls" list (normally a list of last 20 outgoing calls) where the lower order corresponds to more recent dialed number (e.g. "1" is the most recent dialed number, "20" is the least recent dialed number). Thus the same number/contact can occupy in more than one position on the list. Clearly the numbers/contacts on the list are pushed down one position when new call is received.

Based on the position on the list and its corresponding number of times that actual incoming caller was listed on that position, the likelihood of receiving a call can be estimated. For example, suppose currently statistic (hash table) shows that position "3" of the list has the most counts, it implies that the number/contact that is on position "3" of the current "last dialed calls" list has the highest likelihood of being the next caller. Therefore the fourth hash table maps each position on the "last dialed calls" list to a count of calls received.

Once the input call records have been ingested and the hash tables generated, the classifier is considered trained. With the classifier trained on a set of representative call records, it is then ready to be used in predicting mode. The classifier is given a target day of week, hour of day, total call count, and current "last 20 dialed calls" list, and uses the calling behavior model to estimate the likelihood of the user receiving each of the telephone numbers (or caller identifiers) seen in the training data. Clearly the classifier can only make predictions for numbers that it has already seen.

A likelihood metric is calculated for each number known to the classifier and the numbers are then sorted in descending order of likelihood of being received. If the

caller's behavior has a degree of temporal predictability (i.e. they tend to make calls to user at a certain time of the day, or in a particular day of the week, or after some number of calls from the user), then it is expected that the number is likely to be listed towards the top of the list. When several numbers end up with the same value of likelihood, they are listed in alphanumeric order.

The classifier itself is of a type known as a Naïve Bayesian Classifier. In our case, we wish to compute the likelihood of each number (T_n) being received given that the day of the week (D_x), hour of the day (H_y), the current "last 20 dialed calls" list (L_z), and total call count (F_n).

Bayes rule [1] of conditional probability is given by Eq. (1).

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}, \tag{1}$$

where *P(A|B)* is the *posterior* probability which is the probability of the state of nature being *A* given that feature value *B* has been measured. The *likelihood* of *A* with respect to *B* is P*(B|A)* which indicates that other things being equal, the category *A* for which *P(A|B)* is large is more "likely" to be the true category. *P(A)* is called *prior* probability. The *evidence* factor, *P(B)*, can be viewed as a scale factor to guarantee that the posterior probabilities sum to one.

We use this rule to obtain the probability of a number being received given a specific hour of the day, day of the week, current "last 20 dialed calls" list, and total call count, as given by Eq. (2).

$$P(T_n \mid D_x, H_y, L_z, F_n) = \frac{P(D_x \mid T_n)P(H_y \mid T_n)P(L_z \mid T_n)P(F_n \mid T_n)P(T_n)}{P(D_x, H_y, L_z, F_n)}, \tag{2}$$

A known issue with the Naïve Bayesian classifier occurs if a particular attribute value doesn't occur in conjunction with every class value in the training data. The attributes in our case are D_x, H_y, and L_z. The class values are the incoming telephone numbers. The computed probability of a number being received at a particular time will be zero if the training data has no instance of that number being received during either the specified hour or the specified day.

A solution to this problem is to start all the call counts in the Hash tables for day-of-week and hour-of-day at one instead of zero and introducing some normalizing factors in the resulting computations.

This is not an issue for the F_n since there must be at least one call count for any seen incoming call. For L_z, this is sort of an issue since only those numbers/contacts that are on the current "last-20-dialed-calls" list are considered. A solution for this case is to assign the lowest call count of the position on the last-20-dialed-calls list (hash table) to those phone numbers that are not on the current last-20-dialed-calls list. Therefore, those numbers that are not on the current last-20-dialed-calls list will have the same probability of being received as the lowest probability of the number on the current list being received. There is also a possibility of one telephone number occupies more than one position on the current last-20-dialed-calls list. In this situation, the highest call count among all positions occupied by that telephone number is assigned to it.

Adopting this approach, we compute the likelihood of a number T_n being received, given D_x, H_y, L_z, and F_n, by Eq. (3).

$$L(T_n \mid D_x, H_y, L_z, F_n) = \left(\frac{C(T_n D_x)+1}{C(T_n)+7}\right)\cdot\left(\frac{C(T_n H_y)+1}{C(T_n)+24}\right)\cdot\left(\frac{C(T_n L_z)}{C(L)}\right)\cdot\left(\frac{C(T_n F_n)}{C(T_n)}\right), \quad (3)$$

where $C(T_nD_x)$ is the call count from caller T_n on day D_x (x = 1, 2, 3, …, 7), $C(T_nH_y)$ is the call count from caller T_n during hour H_y (y = 0, 1, 2, …, 23), $C(T_nL_z)$ is the call count from caller T_n when T_n's position on the current last-20-dialed-calls list is L_z (z = 1, 2, 3, …, 20), $C(T_nF_n)$ is the total call count from caller T_n (n = 1, 2, 3, …, N, where N is the total number of callers that have made at least one call to the user), $C(L)$ is the total call count of all position on the list (sum of the second column of hash table in Fig. 7), and $C(T_n)$ is the total call count from caller T_n over the whole training data.

4 Performance Analysis

In this section, the CPL is tested against the actual call logs of 20 mobile phone users as described in Section 3. The first two months (approximately 60 days) of call logs are used to train the CPL and the rest of the call logs are assumed to be the future observed call activities to test the performance of the CPL by observing for each call received what position that actual caller has in the predicted list.

Clearly, if the CPL performed perfectly, one would expect the actual caller to be at the top of the predicted list. Generally, such performance is not achievable, but one might expect that the actual caller would tend to appear earlier rather than later in the list.

The overall performance of the CPL based on these 20 users is shown in Fig. 2 where the its accuracy is measured by the average percentage of the actual callers listed within the predicted list as the length of the list varies from 1 to 20 comparing with the accuracy of the conventional "last 20 received calls" list. Figure 8 shows that the CPL outperforms the "last 20 received calls" list with roughly 20% better accuracy.

If there was only one caller, the CPL would always predict the caller correctly. In general, the population of the callers increases (e.g. meeting new friends, signing up for a new phone list, telemarketers gain access to your phone number, etc.). This increasing number of caller population may affect the accuracy of the CPL, i.e. it becomes harder to select a correct number out of a larger sample space.

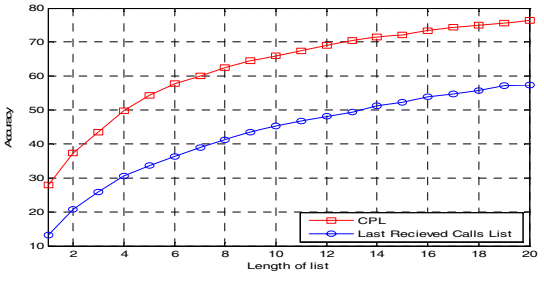

Fig. 2. Overall performance of the CPL comparing to the conventional "Lat 20 Received Calls" list

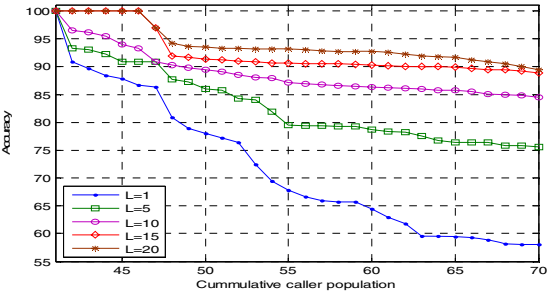

Fig. 3. Relationship between the accuracy of the CPL and the cumulative caller population

Figure 3 shows the relationship between the caller population and the accuracy of the CPL by selecting phone user #20 as an example where the vertical axis represents the accuracy of the CPL, and horizontal axis represents the cumulative caller population which continues to increase from 41 callers to 70 callers. Figure 3 shows that the accuracy decreases dramatically as the caller population becomes larger for different length of the list (L = 1, 5, 10, 15, 20). The accuracy drops with relatively higher rate for shorter length of the list as one may expect.

At the same time, the new callers or first-time callers (whose call received for the first time) also degrade the performance of the CPL. This may be an issue for those users who are more social and those who are unfortunately on telemarketers' lists. This is a voice spam problem which is expected to increase especially in the VoIP networks where the cost of communication is relatively low and with the absurdly large IPv6 address (can support up to 2^{128} addresses).

To demonstrate the impact of the new callers, we examine the accuracy of the CPL without considering the new callers, i.e. if the caller is the first-time caller then it is not taken into account for the accuracy computation. However, after the first call, the caller will be recognized and taken into account for accuracy computation as normal.

It can be seen from Fig. 4 that the accuracy of the CPL is indeed improved about 8% as the new callers are not considered.

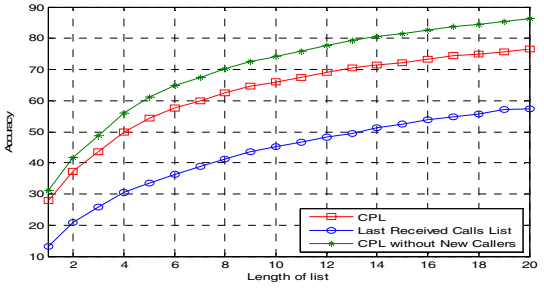

Fig. 4. Overall performance of the CPL without considering first-time callers comparing to the original CPL and the conventional "Lat 20 Received Calls" list

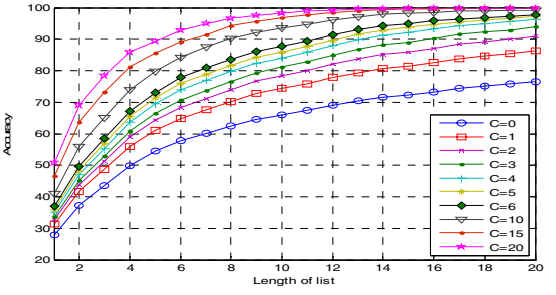

Fig. 5. The impact of the new callers to the accuracy as the criterion of new caller (*C*) varies from 0 to 20

If we modify our definition or criterion for a new caller by defining a new caller to be a caller who has called *C* times in the past, then we observe that as variable *C* increases the accuracy of CPL also increases accordingly, as can be seen in Fig. 5. This tells us that CPL can predict more accurately for the callers whose behaviors have been learned for a longer period of time.

We can further extend the concept of the new callers by using variable *C* to infer the *social closeness*. It is reasonable to assume that the callers who have made higher number calls to the user are more socially connected to the user. Thus, we can classify callers into two groups based on the number of calls received.

For any given phone user, let $\overline{C}$ be the average number of calls received per caller during one particular time. For any callers who have made less than $\overline{C}$ calls to the user, such callers are classified as *socially distant callers* (*SDC*) e.g. telemarketers, wrong-number callers, and voice spam, which are normally unwanted calls. On the other hand, for any callers who have made at least $\overline{C}$ calls to the user, such callers are classified as *socially close callers* (*SCC*) e.g. family members and friends.

$$Caller = \begin{cases} SDC, & C(T_n F_n) < \overline{C} \\ SCC, & C(T_n F_n) \geq \overline{C} \end{cases}, \tag{4}$$

Based on our 20 phone users, the users received an average of six calls per caller during the first two months (learning period). According to Eq. (4), the callers who have made at least six calls are considered socially close callers and the rest of the callers are socially distant callers.

Table 1 shows the accuracy of the CPL for callers who have made at least six calls to the user (*C*=6) at the different length of the list (1, 5, 10, and 20) for each user.

Table 1 shows the comprehensive result which reflects the genuine character of the CPL whose mechanism driven by Bayes rule of conditional probability where the future events conditioned by the past. Hence CPL needs input of historical call logs to learn calling behavior. In fact, it only needs at least six calls for each caller to be an effective predictor. In addition, *SCC* are normally family members and friends who are more desired callers than *SDC* who are normally telemarketers and voice spam.

From Table 1, if the list is only allowed one entry, the CPL would have correctly predicted the socially close callers on average of 40% of the time. If the list has five

Table 1. The performance of the CPL for all 20 users for different length of the predicted list (1, 5, 10, 15, and 20)

Phone User	Accuracy of CPL as length of the list (L) varies (%)				
	L=1	L=5	L=10	L=15	L=20
1	23.68	60.53	84.21	94.74	100.00
2	16.15	51.55	66.77	83.85	91.93
3	62.00	98.00	100.00	100.00	100.00
4	30.95	92.86	100.00	100.00	100.00
5	42.71	97.92	98.96	100.00	100.00
6	30.42	70.28	91.96	96.50	99.30
7	33.33	100.00	100.00	100.00	100.00
8	39.17	68.66	84.33	93.09	97.24
9	12.90	45.16	77.42	98.39	100.00
10	48.51	90.10	96.04	100.00	100.00
11	10.56	38.73	71.83	90.85	99.30
12	35.71	92.86	100.00	100.00	100.00
13	11.11	35.90	64.96	81.20	87.18
14	74.25	94.31	98.66	99.67	100.00
15	14.29	49.21	76.19	92.06	98.41
16	13.31	45.04	67.99	75.35	82.72
17	68.82	91.25	98.86	100.00	100.00
18	52.28	76.14	89.15	95.44	98.92
19	43.75	69.17	88.75	98.75	100.00
20	73.53	93.38	100.00	100.00	100.00

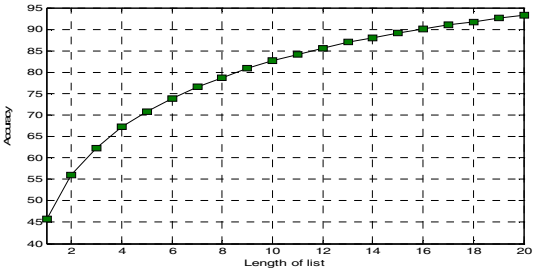

Fig. 6. The performance of CPL as outgoing call predictor (Intelligent Address Book)

entries, the CPL would have correctly predicted the callers 75% of the time. The accuracy would reach 90% for the list of only ten entries.

Since call logs represent human behavior associated with trends and changes over time, thus the accuracy of the CPL can also be impacted by the change of the caller's life schedule because it changes the calling pattern towards the user. For example, your friend changes job from working Monday through Thursday from 8AM to 5PM to working Friday through Sunday from 6PM to 3AM. This major change of your friend's life schedule may result in totally different calling pattern towards you, from receiving several calls at night and on weekends to several calls during the day and on weekdays, for instance. With change of calling pattern of several callers could degrade the performance of the CPL even more.

The concept of CPL can be extended to predicting outgoing calls. For any time the user attempts to make a call (e.g., unlock the keypad, flip up the phone, etc.), a list of the most likely contacts/numbers to be dialed is generated according to computed probability based on call history (day, hour, total call count, and reciprocity). This feature can be envisaged as an "Intelligent Address Book" to reduce the searching time and enable better life synchronization for the phone user. The performance of this Intelligent Address Book is shown in Fig. 6 where it can achieve average accuracy rate of 45%, 70%, and 85%, for the list of one, five, and ten entries, respectively.

5 Conclusion

In this paper, we present a novel concept of the Call Predicted List (CPL) that provides phone user an ability to predict future incoming calls as well as an improvement over the "last received calls" functionality that is often provided on today's phones and communication clients (e.g. VoIP soft phones). CPL makes use of the user's call history to build a probabilistic model of calling behavior based on the caller's calling patterns and reciprocity. The calling behavior model is then used to generate a list of numbers/contacts that are the most likely to be the callers for the next hour. To validate the performance of the CPL, the real-life call logs of 20 mobile phone users are used. The accuracy of the CPL is measured by the percentage of the actual callers listed within the predicted list as the length of the list varies from 1 to 20. The CPL shows 20% improvement in accuracy over the conventional "last 20 received calls" list. In addition, we infer the social closeness from number of calls received as we classify callers into two categories; socially distant callers (e.g. telemarketers, voice spam) and socially close callers (e.g. family members, friends). We believe that socially close callers are more desired callers than socially distance callers. Based on our call logs of 20 phone users, we find that callers who have made at least six calls to the user can be classified as socially close callers for which the CPL accurately predicts 40% if the length of the predicted list is one, 75% if the length is five, and 90% if the length is ten. We also discuss that the accuracy of the CPL can be also impacted by the increase of caller population, new callers, and change of caller's life schedule. We also show that with a simple modification in input variables, CPL can also be useful for predicting outgoing calls as an "Intelligent Address Book," by which for any time the user attempts to make a phone call, a list of the likely contacts/numbers to be dialed based on call history is generated. We believe that CPL helps pave the way for future pervasive computing research, which aims to improve quality of life. As our future direction, we will continue to investigate other parameters to characterize and detect the trends/changes in calling behaviors, and explore other prediction techniques to improve the accuracy of the CPL.

Acknowledgements. This work is supported by the National Science Foundation under grants CNS-0627754, CNS-0619871and CNS-0551694.

References

1. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. A Wiley-Interscience Publication, New York (2001)
2. Kolan, P., Dantu, R., Cangussu, J.W.: Nuisance Level of a Voice Call. In: ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP) (November 2008) (to appear)
3. Phithakkitnukoon, S., Dantu, R.: UNT Mobile Phone Communication Dataset (2008), http://nsl.unt.edu/santi/data_desc.pdf