國立交通大學

資訊科學與工程研究所

碩士論文

動態通訊錄 ： 行動電話撥打方的智慧型導引

Dynamic PhoneBook: An Intelligent Guide Phone Caller

研究生　：唐明

指導教授 ：彭文志 教授

中華民國 102 年 7 月

Dynamic PhoneBook: An Intelligent Guide Phone Caller

研究生　：唐明　　Student : Bustami

指導教授 ：彭文志　　Advisor : Wen-Chih Peng

國立交通大學

資訊科學與工程研究所

碩士論文

A Thesis
Submitted to Department of Computer and Information Science
Institute of Computer Science and Engineering
National Chiao Tung University
in partial Fulfillment of the Requirements
for the Degree of
Master
in

Computer Science

July 2013

Hsinchu, Taiwan, Republic of China

中華民國 102 年 7 月

動態通訊錄 ： 行動電話撥打方的智慧型導引

學生 ： 唐明　　指導教授 ： 彭文志

國立交通大學資訊科學與工程研究所

摘　　要

現代生活中的智慧型行動裝置或手機的使用量快速成長促進了行動應用程式開發的蓬勃發展。而提供使用者重要且實用性質高的資訊也是行動應用程式的主要關鍵之一。這些行動裝置應用程式的主要目的不外乎讓使用者認為智慧型手機能夠拓展與人群的連結性、提高生活周遭的機能性、使得國際社會資訊唾手可得。電話撥打行為的預測成為了使行動裝置更加貼近使用者生活的特徵之一。就我們所知目前的知識技術上，有兩篇近期的文獻探討了這些問題[11, 12]。即使此兩篇文獻已在既有的電話通訊系統中發現了具前瞻性的成就，但我們仍相信可以在前人的研究中發現更多隱含且少被關注的特徵。本論文中我們研究發現了許多能夠提升精確度結果的特徵。具體而言，給予使用者歷史通訊紀錄活動，我們可以發現以下四個主要特徵－頻率性(*Frequency*)、持續時間性(*Duration*)、近期性(*Recency*)與通話方向性(*Direction*)。頻率性(*Frequency*)參考了使用者與電話接受方(callee)的通訊次數。而使用者與電話接受方的每次通話時間則為持續時間性(*Duration*)。近期性(*Recency*)考慮了使用者與電話接受方(callee)每次通話的發生時間遠近。最後，通話方向性(*Direction*)參照了通話發起方的重要性並以預定義之權重值代表。根據此四項特徵，我們建立了以下三種機率等級序模型: *Probability General-Frequency* (PGF)， *Probability General-Duration* (PGD) 以及 *Probability Recency* (PR)。再者我們應用此三種模型在真實的通話詳細記錄(Call Detail Record, CDR)在 Reality Mining 以及中華電信資料庫得到相關結果。最終我們與相關文獻的比較顯示出了我們的機率模型擁有更高的精確度衡量。

Dynamic PhoneBook: An Intelligent Guide Phone Caller

Student : Bustami Advisor : Dr. Wen-Chih Peng

Institute of Computer Science
National Chiao Tung University

ABSTRACT

The rapid growth of smartphone uses in today's modern life encourages the development of useful application that provides a batch of useful information to it's users. The main aim of this development is to make smartphone smarter. Phone call prediction is one of applications that serve as an important feature to achieve smarter smartphone. To the best of our knowledge, there were two most recent works on the development of telephone call [11, 12]. Even though those studies have shown the promising achievements over the basic congenital telephone system, we are still confident to explore many basic features that seem that received little attention in their research. In this paper, we investigate more conservative features that can subscribe as same accuracy result or even better. More specifically, given the user historical call activities, we explore four major features: *frequency*, *duration*, *recency*, and *direction*. The *frequency* feature refers the number of interaction calls between the user and callee. The period of time that the user and callee spent in each of their communication call defines as *duration* feature. While, *recency* feature is the weight of each connection call between user and callee according to the recentness of that call. Lastly, the feature of *direction* describes the importance of call initiator between user and callee by giving pre-defined weight. According to these features, we develop three probability ranking models: *Probability General-Frequency* (PGF), *Probability General-Duration* (PGD), and *Probability Recency* (PR). Moreover, we train these models in two real Call Detail Record (CDR) datasets, Reality Mining and Chunghua Telecom dataset to gain depiction result. Finally, we compare these models with the existing works and demonstrate that our conventional models can reach same and even better accuracy prediction.

Acknowledgements

Foremost, I would like to express my sincere gratitude to my advisor Prof. Wen-Chih Peng for the continuous support of my Master study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Master study.

Besides my advisor, I would like to thank Dr. Meng-Fen Chiang (Young).As my senior, my mentor, my motivator, thanks for your encouragement, insightful comments, and many zsuggestions.

My sincere thanks also goes to ITRI-Groupmate: Fankai and Jordan, for sharing me many knowledge and information during working on ITRI's project.

I thank my fellow labmates in Advance Databases Syatem (ADS) lab: Ted, Eric, Liong-Shang, Yanti, and Jiejie Paipai, for many interesting discussions. All my juniors: Gunarto, Andreas, Alex, Thai, Hai, Yuting, Aggie, Wangting, Xinxin, and Yu-Hsiang for many endorsements.

In particular, I am grateful to my respectful senior Gege Demension, Barry, Oshin, Chen, and Wen-Yuan, for grateful advices.

Furthermore I would like to thank the Government of Aceh, for offering me the opportunity and founding during my master study. All Acehness friends, Basrul, Intan, Pak Den, Nuri, Nia, Malya, and Pak Nayan, thank for any advice and reminding me in many aspects.

Last but not the least, I would like to thank my big family: my mother Siti Dahlia and My late father Muhammad Yusoef Ubit, for giving birth to me at the first place and supporting me spiritually throughout my life, my brothers and sisters: Amrillah Yusoef, Yuslinda Yusoef, Idrus Yusoef, Asnawi Yusoef, and my younger sister Ikhwana Yusoef, thank for all of yours support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Over the last two decades, mobile phone has become the most common and popular communication tool in everyday life. Its advantages such as portability, convenience and affordability, make a mobile phone a prevalent reference to interactive communication. Although in recent years advanced technologies have been implemented on modern cell phone namely as **"smartphone"**, but they do not change the essential function of mobile phone itself as a communication tool.

As a substantial interface tool in the daily uses, each user historical communication logs (*e.g. voice call, messager, and data transfer*) represents the behavior how, when, and to whom the mobile phone is used to contact with. By evaluating and learning from this behavior, we can acquire the users pattern that can be used for various of purposes, for example detection and prediction of user's phone call activity.

Prediction problem and power consumption are the warmest associate issues that have been discussed in todays research. On mobile phone call (outgoing call) prediction, the issue heavily concentrated in the way how to make a call. Ordinarily, there are three particular ways to make a telephone call. The first way is typing directly on screen the exact number that we are going to call. This way does not require any logger time, but it requires a good memory of user to memorize all the numbers and it also not common way for most of users to make a telephone call. Another way is using default phone's **Current List**. Current List is the list of the numbers (at least 20 numbers) that ordered according to the recent dialed

number by user. The most recent dialed number will occupy the top position on the list and the least recent dialled number is at the bottom of the list. This way provides an efficient time i making a telephone call if the number that will be dialled appear on the list, otherwise the Current List will not useful. The last way is going through to the phone book to search/find prospective number on the list. Conventionally, **phone book** on the mobile phone is listed in an alphabetical order. In the worst scenario if we do not remember the number or even the name the prospective callee, we need to scroll the entire phone page until the desired number is found. This way not only requires a significant amount of time than two previous ways but also results in more energy uses. An illustration example of phonebook is shown on Fig. 1.1.

To deal with these issues, we believe that an automatic phonebook system is increasingly becoming important to predict a potential callee that will be dialed by the user in given query time. This paper propose an outgoing prediction model for the foundation of an intelligent phone book system based on the users calling behavior. Clearly, we begin this work by exploring 4 principal features: $frequency, duration, recency$, and $direction$. Subsequently, by using these features we construct three probability predictive models: $Probability\ General - Frequency(PGF)$, $Probability\ General - Duration(PGD)$, and $Probability\ Recency(PR)$. Those models will be implemented as an intelligent phone book system called **Dynamic Phonebook**(DP). Basically, DP dynamically rotates the most potential callees to the top of phonebook list in each given time slot according to users past communication information. DP not only provides an efficient way in making call, but also can be used as a reminder for the user in specific time to call somebody that he used to contact with. In contrast, the main goal of our work is not so different with previous researches, but we restrict our work only on outgoing call prediction in order to get maximum impact. We explore more baseline features that apparently acquired little attention on existing works. We also implement on either various existing ranking model and our adopted/modified ranking techniques.

Actually, there are some related works on study of telephone call prediction. For example, the works in [2][4][3][5] studied forecasting arrival telephone call in a call center. The main goal of these studies is to predict the volume of calls that will arrive at a given specific slot time. These works can contribute for many purposes, for instance staff scheduling, budgeting, improving sales marketing, etc.. The other works [13][6] concentrated on modelling telephone call prediction, where they proposed models on future communication activity and annoying

telephone call prediction. The most recent research has been conducted to predict incoming and outgoing call based on a nave Bayesian algorithm [11] which is closely related to our work. Although this machine learning approach has improved the performance of incoming/outgoing call prediction in comparison to the default current call list, prediction performance either high when the length of number of predictions was choosing high, or quite low when the length was low. Their works also pointed little comparison with either existing approaches or baseline approaches.

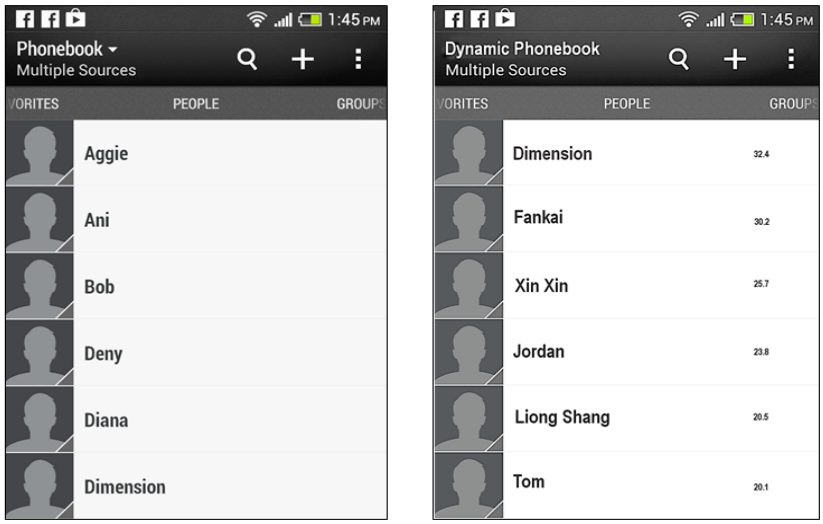


Figure 1.1: Example of Phonebook: (left)Conventional phonebook; (right) an intelligent phonebook

The rest of this paper is organized as follows. Chapter 2 presents some detailed literature review on closely related work. In Chapter 3, we describe the preliminaries of our work and our proposed system architecture. We present the phone call prediction models in Chapter 4. In Chapter 5, we provide comprehensive experimental results. We finally conclude our work in Chapter 6.

Chapter 2

Related Work

Even though there is an enormous amount research about forecasting based mobile phone data, the number of works that very exclusively peeling about telephone call prediction is not very impressive. In this section we will summarize briefly some previous works in predicting issues using mobile phone dataset.

2.1 Mobile device log analysis

With the rapid growth of the mobile devices in todays modern life, such as smart phone and tablet computer, almost all of users activities can be able to be captured. These activities data that are commonly stored on mobile devices storage or connected server has attracted many researchers to explore the unseen users behavior and mining many unexpected knowledge that involved in these dataset [9][16][8][7][15][1][17]. In [9][16], the authors analyzed and predicted the user location based on their SMS (Short Message Service e.g. text messaging) and cellular call information. They found that the user frequently used SMSs while they were moving and call pattern between two users is highly correlated with their co-location. By using the some part of Reality Mining dataset[8], Choujaa and Dulay [7] analyzed the history of human behavior based the specific points of a user where and when she made a call and predicted the user take place at given particular time in the coming time . Leman and Dalvi, in [1] also used large collected phone and SMS dataset in their work. They explored three main problems,

analyzed users relationship strengths, predicted the users tie persistence, and finally developed a change-point detection that can be used to identify the important day (event) that occurred among the users. Soto et al. [15] studied information that derived from cell phone data record to identify the socioeconomic status of a population or an individual. Meanwhile, Zhang et al. used ARIMA (Autoregressive Integrated Moving Average) model on phone-call detail record to predict social-ties strength among the users [17]. While these studies have successfully on mining neither Call Detail Record (CDR) or text messaging data collected and provided some predicting model, but none of these models can be used on call prediction. However, our work observes the features of user calling behaviour from CDR and text messaging and provides a call forecasting model to predict the users calling intention.

2.2 Telephone Call Forecasting

A number of studies have attempted to forecast telephone call. Some of them focused on call center forecasting [2][4][3][5], in order for efficiently staff scheduling, planning, budgeting, and improving sales marketing. Andrews and Cummings [2] developed two forecasting models, retailer of high-quality outdoor goods and apparel, to forecast incoming call (order, e.g. buying merchandize and inquiries, e.g. checking order status) at L. L. Beans call center. They used Transfer Function model/ARIMA on modeling these time series data and they found that the ARIMA model can be improved by using independent variables that can be directly effect to efficiency staff scheduling. Bianchi et al. in [4][3] tried to compare ARIMA model with their adaption the Holt-Winters model that were call NAMES in telemarketing center call forecasting. They found that ARIMA models can be used to account for outliers and performed significantly better than either of the Holt-Winters or additive and multiplicative versions of Holt-Winters models. Similarly, in [5] Boulin also used combination of ARIMA model and Judgmental forecasting to predict sales call volume. He found that the proposed method improves the accuracy both of weekly forecasted call volume and daily volume, respectively from 23% to 46% and from 27% to 41%. In contrast, our work is very different with these existing studies. Our work focus on forecasting in terms of when and to whom the call will be made, but most of their work proposed the model to predict the volume of arrival call in

some coming days.

2.3 Call Prediction Application

There are also some studies worked on call prediction modeling. Harless and Kowalski in [10] introduced a system that can be used for future communication activity prediction based on collected users previous communication information. The system consists of two main tasks, correlation logic and prediction logic. Correlation logic analyze the past communication event to consider whether the correlation exist in the past communication event information, while prediction logic examines a current communication event and predicts the future communication event according to the current communication event and the correlation. In 2010 Sing et al. [14] field a patent on predictive annoying telephone call. They proposed a method that can be used by the called party to consider whether a telephone call is annoying or not. Their proposed model was adopted from some mechanisms of spammer email detective model. For example a telephone call will be considered as annoying call by examining the characteristics responded by called party of previous calls from the same caller (e.g. when a called party receives an unwanted solicitation call, she will hung up within the first minute).

To the best of our knowledge, there are two studies that are the closet to our work and the most recent works on telephone call prediction [11][12]. Phitakkitnukoon et al. developed a **Call Predictor** (CP) [11] based on the user's call behavior and reciprocity. Callers calling behavior is measured by his previous calling pattern, whereas reciprocity is measured by number of outgoing call divided by number of incoming call. The CP itself consists of two main sections, **Probability Estimator** (PE) and **Trend Detector**(TD). The PE is the section that enumerates the probability of receiving a call based on the caller calling behavior and reciprocity. The detection of recent trend of caller calling pattern and computational of the adequacy of historical data in term of reversed time is occurred in TD section. Meanwhile, the author in [12] proposed **Call Prediction List** (CPL), a telephone call prediction system based on user's history call log. They used the Naive Bayesian Classifier with some independence parameters to forecast next incoming call and outgoing call. The differences of CP and CPL is that CP only can be used to predict when a specific caller/callee probably will make/receive

the call in the next-24 hour, while CPL capable generate the list of potential callers/callees that the user will receive/dial a call in the next hour. Even the most part of these works dedicated for incoming calls prediction, with the same system also suitable using for outgoing call prediction.

While these previous approaches which leverages machine learning algorithm and having a bunch of parameters, have been shown the promising enhancement of calling prediction model over the basic congenital telephone system, but these approaches were not evaluated enough with the simplest features that seems less contributes. In contrast, our work tries to pay more attention in exploring a number of base line features that can be included to achieve higher accuracy in call prediction.

Specifically, the objective that we want to reach is with the smallest number of prediction we can achieve the maximum accuracy. This goal may seem somewhat exaggerated, but this is the big challenges in prediction issue. Here we elaborate the illustration of this promising challenge. Mostly smartphone screen can hold maximum 10 lists of callees in one open screen. If our system can predict one per 10 callees at a specific timeslot, the user only need to show a single screen whenever she want to make a call without spend more time in scrolling down the screen until to find the intention callee. To be specific, the research questions that we want to address on this work are:

- Which context is highly related to outgoing call prediction?
- Will the simplest feature reach higher accuracy?
- Which existing predictive model work very well on call prediction?
- How well does our adopted propose model improve call prediction?

Chapter 3

Dynamic Phonebook Framework

3.1 Framework Overview

The architecture of our work is shown in Fig. 3.1. the mainframe of our system consists of two main sections: Phone call pattern discovery and Outgoing call prediction. In the section of Phone call pattern discovery, we investigate users calling behaviour and mine phone call pattern of the user. There are four features that we discover in this section, frequency, duration, recency, and direction. The second section is the key part of our system, which works based the features that generated in first section and input query time. This section, we formulate the features as probability model for generating top K callees prediction.

3.2 Preliminaries

To analyze the calling behaviour of users, the calling log of each user is first transformed into Calling-history. The purpose of transformation is to collect all calling log of a user and record the number of calls in each time slot for analyzing of calling behaviour and later processing. The calling logs and the calling-history are formally defined as follow:

Definition 1. Calling log: A calling log is a tuple $< callee, t >$, where a particular callee *callee* is called at a specific time t.

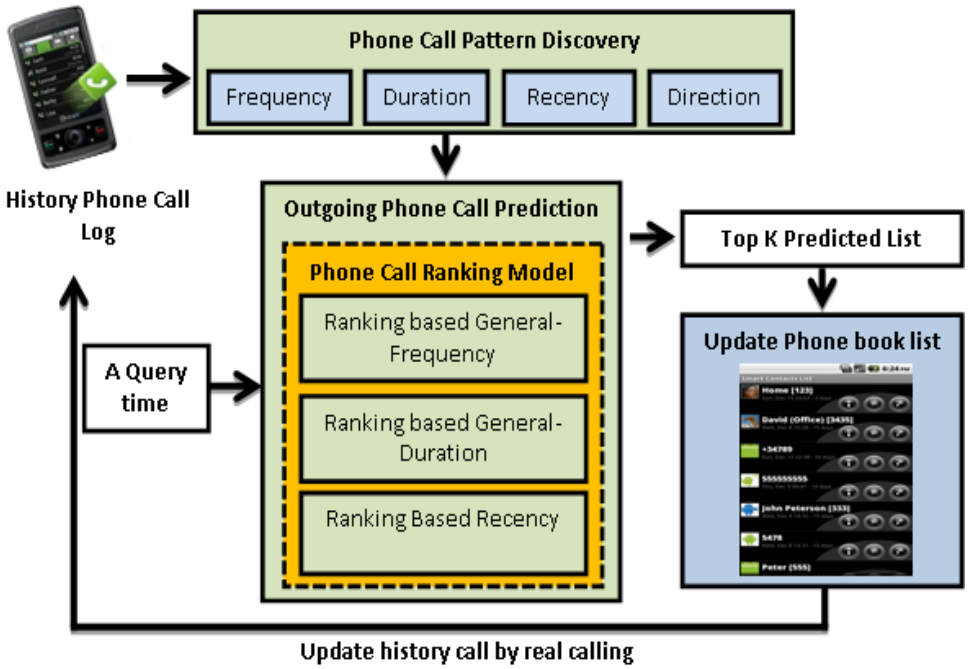


Figure 3.1: Overview of our system

Definition 2. Calling-history: Given a starting time of collection t_s and the length of time slot t, the Calling-history of callee callee is defined CH(callee) $= < n_1, n_2, ... >$ where n_i denote the number of user call callee callee at i^{th} time-slot.

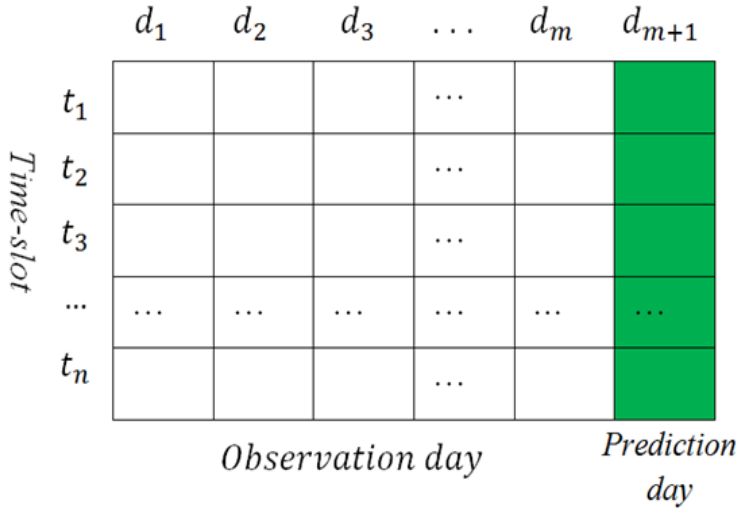


Figure 3.2: Illustration of an telephone call prediction schema

The most common way in prediction problem is by exploring previous/current information of predictable item to find a number of patterns and then use them to forecast the potential things that could happen in coming time. By adopted the same way, we apply this particular

procedure in outgoing call prediction model. Each of users history call will be plotted as nxm matrix, which n is number of divided time of the day (which we call $time - slot$) and m is the number observation days. During the observation day d_m, we accumulate all possible information of users calling behaviour and visualize them on each cell of the matrix accordingly to the slot time t_n (when the communication between the user and callee had occurred) and then use it to predict probably callee that will be dialed by the user in coming day $d_(m + 1)$ at given time t_n. As shown in Fig. 3.2., each cell of the matrix consist list of $< callee_i, (f_1, f_2, >$, where $callee$ is list of callees and f_j is considered features. Detail of used features will be explained in next chapter.

3.3 Dataset

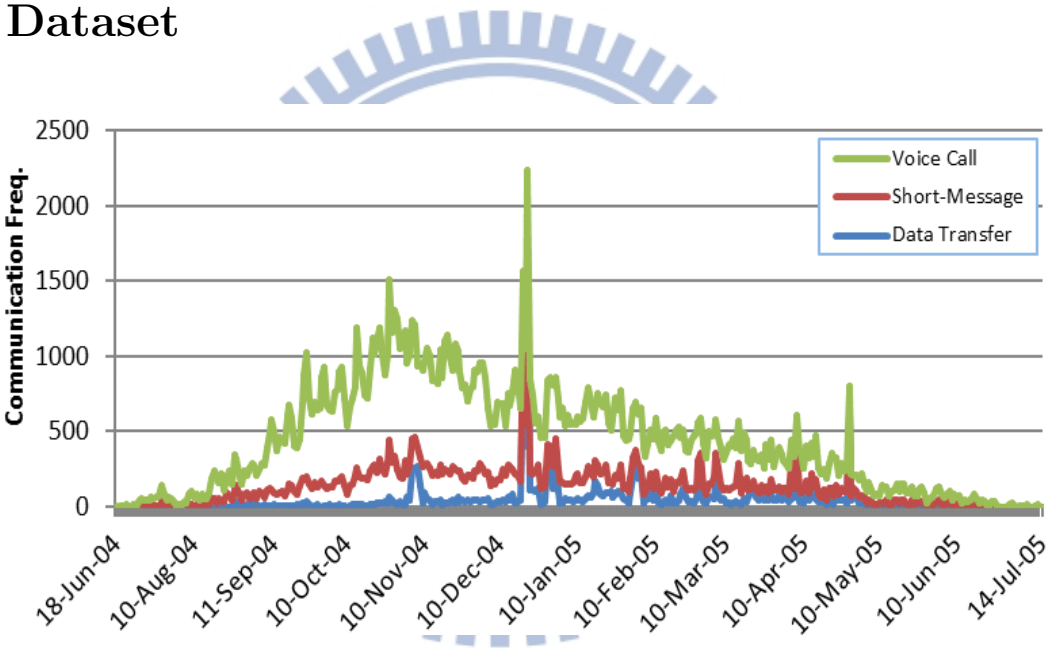


Figure 3.3: A plot of all type of communication frequency for each day of data collection

To evaluate all the methods for outgoing call prediction, we extensively conduct all the experiment on two real CDRs, Chunghua Telecom dataset and Reality Mining dataset. Both CDR are used 60% for training data and 40% for testing data. Generally, each raw CDR data consist of 5-knot of information as follows where an example of call detail record is shown in table (**??**).

- *Caller_ID* : Who make the call

- *Callee_ID* : Who receive the call
- *Start_time* : Start time of call (in unix-timestamp)
- *End_Time* : End time of call (in unix-timestamp)
- *Duration* : the period time of call (usually in second)

Table I: Example of Call Detail Record

Caller_ID	Callee_ID	Start_time	End_time	Duration
0001	00075	10915194	10787611	20
0006	00048	10915194	10787610	10
0025	0001	10915194	10787611	10
...	...	...	...	...

3.3.1 Chunghwa Dataset

Chunghua Telecom (CHT)[1] dataset is not public dataset. It was collected during August-September 2010 from Chunghwa telecom operator in Taiwan. Originally, this dataset is corresponds approximately 10 million Call Detail Records (CDR) from 85 thousand unique phone numbers, but in our experimental we only randomly select 100 users which have high frequency calls. For personal privacy issue, each of phone numbers was anonymized by hash random number. Each call record consists of caller, collee, start time, and duration of call.

3.3.2 Reality Mining Dataset (RMD)

The Reality Mining dataset were collected for a period of 9 months from ninety-four observation users at MIT and Sloan Business School[13]. The users were, professors, staff members, and students from both Media lab (MIT) and Sloan Business School. There have tree type of communications were recorded among the observation users, $voice-call$, $sort-massage$, and $Packet-data$. Each of communication types consist of callers and callies (ID user), unique time-stamp, direction (incoming or outgoing), description (missed or accepted), talk time (only for voice call type).

[1]The CHT data is not in public, and Chunghwa Telecom's website is http://www.cht.com.tw/

To determine the effect each kinds of communication type for dynamic phone book, we divide the RMD to 3 different conjunction matrix, *Voice call Only*, *Voice call* and *Sort massage*, and *All* of type of communication. See Fig. 3.3. for a concrete graph of frequency distribution of the difference the three types of user communication. We find Voice call to be the most common communication among the users compared with short massage and data packet sharing.

Chapter 4

Method

In this section we describe the key point in our work. First, we will discover four potential features from historical call log, which are *frequency, duration, recency, and direction*. Based on these features, we propose two call pattern based prediction models, one adoptive model, and also we describe one ranking based algorithm.

4.1 Feature Discovery

4.1.1 Frequency

Given the historical call log of a user, the frequency of each callee, *freq*, is counted by capturing all direction call from the user to the callee. With simple word, frequency of a callee is the number of calls he received during observation day from the user. Consideration of frequency feature is the basic principle of the callee who receive frequent call from the user consider as more important than the callees with which the user interacts infrequently. The following definition is the way how we define frequency of a $callee_i$ by given the historical call log of a user:

$$freq(callee_i) = \sum U(callee_i) \tag{1}$$

where $U(callee_i)$is the direction call from the user U to callee $callee_i$.

4.1.2 Duration

Same as the frequency, the feature of duration, *duration*, is discovered with assumption the callee which the user spend longer time in every interaction time consider more important than those callee where the user only stay for a while in each communication. To discover the feature of duration, is the overall time consuming in each conversation time between the user and callee. So that, given the historical log call of user U, the *duration* of each callee $callee_i$ as defined below:

$$duration(callee_i) = \sum U(callee_i) \tag{2}$$

where $UT(callee_i)$is the period of time that user U spend in each his phone call interaction to a callee $callee_i$ during observation day.

4.1.3 Recency

Regarding the feature of recency, *recency*, we observe that the callee which the user very actively interacting with recently is more important than the callee which the user made a call a couple months ago. Compared to two previous features, the recency is calculated in the same ways, but the recent call will contribute more weight according to decay function. The recency model was introduced by Carvalho and Cohen [6]. In simple word, given the historical log call of user U, the recency of each callee $callee_i$ is defined by following equation:

$$recency(callee_i) = \sum UT(callee_i) e^{(\frac{-timeCall(callee_i)}{\lambda})} \tag{3}$$

where $timeCall(callee_i)$ is the the rank according to chronological order of time when user U made a call to a callee $callee_i$, while setting parameter λ relate to number of observation day.

4.1.4 Direction

Basically, the feature of direction explore as a required feature by the prediction based ranking model (Interaction Rank [13]) to generate list of potential candidate callees which will detail explanation in next following subsection. To clearly imagination, the callee who initiates the

interaction with the user is considered more significant that the callee only receiving call from the user without any respond back to call the user.

4.2 Call Predictor Based Probability Model

According to users historical call log pattern, in this subsection we formulate these discovered features as probability model of a user dialing call. There are three measurements according to these feature, general-frequency, general-duration, and Recency. We also attempt to use an existing interaction rank model [13] as a predictor in term of generating the list of potential callee that will be dialed by the user in given time-slot.

4.2.1 General-Frequency Probability

To formulate the probability of $general - frequecy$, the user frequency call to a specific callee and the frequency call to all callees are considered. The probability of $general - frequecy$ of a callee represents that the frequency calling in receiving calls under all population call that made by the user. For a given a historical log call of the user U, and time-slot t, the $general - frequecy$ probability can be defined as the following formula:

$$PGF(callee_i)^t = \frac{freq(callee_i)^t}{\sum_{p=1}^{C} freq(callee_p)^t} \tag{4}$$

where $freq(callee_i)^t$ is the total number of call that $callee_i$ receive from user U at specific time-slot t. While C is set of callees who receive at least one call from the during observation day. For example, if the callee $callee_0 01$(e.g. 001 is callees ID) receive 25 calls from user U during time-slot 15pm 14pm and the total call that user made during that time-slot is 100 calls. So, the probability of $callee_0 01$ of feature general-frequency is $25/100 = 0.4$.

4.2.2 General-Duration Probability

Consideration duration call time to a specific callee and total duration call to all callee is needed be to formulate the probability general-duration. Basically, the general-duration is

period of time the user spent to communicate with the callee. If the duration time of a callee is calculated by $duration(callee_i)$ (Eq.2), and $\sum_{p=1}^{C} duration(callee_p)$ is summarization of period time that he spent to all callees, so the general-duration probability of a callee, $callee_i$, at time-slot t we define as given in the following equation:

$$PGD(callee_i)^t = \frac{duration(callee_i)^t}{\sum_{p=1}^{C} duration(callee_p)^t} \tag{5}$$

For example, if the feature duration duration at time slot t of callee $callee_0 01$ is 100 seconds and total period users communication during time slot t throughout observation days to all callees is 1500 seconds, so that the general-duration of $callee_0 01$ is $100/1500 = 0.0666$.

4.2.3 Recency Probability

As we mention in subsection of feature discovery, the feature of recency can be calculated by Carvalho and Cohen Recency model [6](Eq. 3). In this model each of call that received by callee will be ordered according to when the call is made and the score of each call will be weighted by an exponential decay function computed over its ordinal rank. In this case, it will have a problem on distinguishing between two calls that happen in two days sequentially and two calls that happen very far away each other. For example, you have an interaction with your friend Jhoen by phone today and two days ago, and you also made twice calls to your friend Tom, today and a week ago. According to Carvalho and Cohen Recency model, the recency score for both Jhoen and Tom are the same although your rencentness time interaction with them are different.

Meanwhile, in 2010 Roth et al. proposed an Interaction Rank model [13] (in the entire paper we call this model as **Gmail-Approach**). Their model is constructed by considering three major components: frequency, recency, and direction. They begin the model by summarize the number of interaction between the user and his opponent, then weight each of interaction with decay exponential over the time defined as its recency. To clarify the importance direction of the interaction, they give an additional parameter to one of parties, so that the interaction direction of one party will be more significant than others. The following

equation is Gmail-Approach in telephone call interaction case:

$$GA(callee_i) = \psi_{out} \sum_{d \epsilon D_{out}} (\frac{1}{2})^{\frac{T_{now} - T_{(d)}}{\lambda}} + \sum_{d \epsilon D_{in}} (\frac{1}{2})^{\frac{T_{now} - T_{(d)}}{\lambda}} \tag{6}$$

where $D = D_{out}, D_{in}$ set of call between user and callee, D_{out} is set of calls from user to callee, D_{in} is set of calls from the callee to user, T_{now} is the current time,$T_{(d)}$ is the timestamp where the call occurred (where $d \ \epsilon \ D$), λ is the decided observation period, while ψ_{out} is defined weight for each of call from the user. According to $Gmail - Approach$ equation, there is an opportunity to use this model as recency calculation. Given a users historical call logs and query slot-time t, the recency of a callee $callee_i$ can be redefined as a following equation:

$$PR(callee_i)^t = \sum_{c \epsilon CU^t_{out}} (\frac{1}{2})^{\frac{T_{now} - T_{(d)}}{\lambda}} \tag{7}$$

Chapter 5

Evaluation Result

In this section we show our experiment result of our proposed methods and some comparison with existing work. Section (5.1) explains the methodology of evaluation measurement. One competitor and some default variables setting will be described in section (5.2) and (5.3) consecutively. Finally, section (5.4) investigates the result of both proposed method and competitor. The whole of our system are implemented in Perl language on a Linux and window operating system.

5.1 Evaluation Measurement

We use $HitRate$ (HR) to evaluate our system. HR associate with the number of Hit and Miss. Hit is the correct prediction that the system can truly predict, while Miss the number of incorrect prediction, so that the HR is number of Hit over by the number of summarize Hit and $Miss$ and it is defined as follows:

$$HitRate(HR) = \frac{\sum Hit}{\sum Hit + \sum Miss} \tag{1}$$

For example, given a set of prediction list PL of the user $user001$ at specific $slot - time$ 2 pm, $PL_{t=2pm} = 004, 002, 006$, and in the next day he only make the call to $callee004$ and $callee006$ in the same slot time, then the hit rate will be $HitRate(user001_{t=2pm}) = 2/2 = 1$.

5.2 Competitor

As we mentioned in previous chapter the work of Phithakkitnukoon *etal.* [11]is the most current work on incoming and outgoing call prediction. They proposed a system which they called **Call Prediction List**(CPL). Based on its original model, given a users historical call and specific time-slot t at particular day of the week, CPL generates list of potential callers or callees based on estimating the likelihood of user receiving/making each of telephone numbers. The likelihood of user is computed by using probabilistic classifier based on Bayes theorem with four independence assumptions. More obviously, to obtain the probability of a number (U_n) being the callee (in term of receiving call) at given a specific hour of the day (H_t), day of the week (D_w), current last-20-dialed-calls list (C_l), and total call count (S_n), as shown on following definition:

$$P(U_n|H_t, D_w, C_l, S_n) = (\frac{P(U_nH_t)+1}{P(U_n)+24})x(\frac{P(U_nD_w)+1}{P(U_n)+7}) \\ x(\frac{P(U_nC_l)}{P(C)})x(\frac{P(U_nS_n)}{P(U_n)}) \tag{2}$$

where $P(U_nH_t)$is the total call count from the user to number U_n during hour H_t(where $t = 1, 2, 3,...\ 24$), $P(U_nD_w)$ is the total call count that number U_n receive at specific day D_w (where $w = Monday, Tuesday, Wednesday,...\ Sunday$), $P(U_nC_l)$ is the total call count number U_n receive when U_ns position on the current last-20-dialed-calls list C_l (where $l = 1, 2, 3,...\ 20$), $P(U_nS_n)$ is the total call count that each number U_n(where $n = 1, 2, 3,...\ N$, where N is total number of callees), $P(C)$ is the total call count of all position on the last-20-dialed-calls list, and $P(U_n)$ is the total call count from the user to all entire callees. Starting all call count in each parameters (especially for day-of-week and hour-of-day) with 1 (one) instead of 0 (zero) and defining normalizing factor (by added 24 and 7 into the total call count $P(U_n)$) are the way to avoid the empty space in conjunction matrix (it mean no any call occurred during the targeting day of the week or the targeting hour of the day) that will affect probability computation to be zero.

By considering the number of calls during $hour - of - day$ and $day - of - week$, CLP clearly expect that the user have specific schedule on repeating the call in the same day at the same hour slot time in his next calling activity to the same callee. But in actual life, there have variety behaviour of callers, some of callers only have a specific either day time or slot hour time to have an interaction to his partner. For example, a young couple usually used to call each other every day in the whole week with some specific $slot - hour - time$ (*e.g.* in morning, lunch time, dinner time, or even before going to bed). These repeating call happen every day. So that, considering the number of call at specific $day - of - week$ will not affect much on predicting these kinds of users. Therefore, we offer a new CPL without considering day-of-week variable that we named it CPL(-d) to deal with this mentioned problem. The remaining formula for CPL(-d) to compute the likelihood of callee U_n by giving H_t, C_l and S_n is shown in following equation.

$$F(U_n|H_t, C_l, S_n) = (\frac{F(U_n H_t) + 1}{F(U_n) + 24})\ x(\frac{P(U_n C_l)}{P(C)})x(\frac{P(U_n S_n)}{P(U_n)}) \tag{3}$$

All methods are listed in the following table:

5.3 Variables

Table (**??**) lists all the variables that are used throughout the experiments. All the variables are set to be the default values unless for specified explicitly. In all phone call ranking models,

Table I: List of methods and abbreviation

Abbreviation	Method
PGF	Probability General Frequency
PGD	Probability General Duration
PR	Probability Recency
GA	Gmail Approach
CPL	Call Prediction List
CPL(-d)	Call prediction List without considering day-of-week

the variables t, λ, and K are used. t is number of dividing slot time in a day, λ is duration day of data observations, while K is the number of predicted callees. Meanwhile, ψ_{out} is the weighted variable that only used in $Gmail-Approach$. This variable is the relative importance of outgoing and contra incoming call.

5.4 Evaluation Result

5.4.1 Prediction Performance

First, we compare the performance of our prediction models with the models used in the previous studies. We selected CPL and $Gmail - Approach$ as the baseline models. We trained these models with two real datasets, Reality Mining and Chunghua Telecom dataset, by setting the number observation day $\lambda = 21$ days as training data and the rest of data as the testing data. Figure 6.2 shows the result, where x-axis represents the number of callee candidate prediction ranging from 1 to 10, while the y-axis represents the accuracy (Hit-rate) of each models. Recall, that our main goal is to have the highest accuracy with the simplest way (i.e. feature).

In general, prediction accuracy varied according to the number of predictions and type of predictors. Fig. 5.1. shows the prediction result for all predictor models on reality mining dataset, where PGF and PR almost dominant in every number of callee candidates. It is clearly shown when prediction candidates for 1 to 8, PGF and PR models outperformed than other models. The only when up to 9 prediction candidates (i.e. selecting the 9 most probable callee) the accuracy distribute almost relatively same for PGF, PR, and CPL(-d) models. In the case of CPL, its accuracy was slightly lower than CPL without consideration $day - of - week$

Table II: List of variables: it's default values and description

Variables	Default value	Description
t	24	Number of time-slot in one dividing day
λ	60	Number of observation days
K	10	Number of predicted callees
ψ_{out}	5	The outgoing call weight

parameter or CPL(-d) for all number of callee candidates (0.5 %). While, in the case of *Gmail – Approach*, its accuracy always left far behind RCs accuracy. Overall, the existing model did not perform as well as these three models (PGF, PR, and CPL(-d) (i.e. in term of prediction candidates when up to 5)), but as the number of candidates increased, all models performed reasonably well except PGD. Meanwhile, Fig. 5.2. shows the prediction result for all predictors on Chunghua Telecom dataset. Similar with previous experimental result, GF and RC models are very dominating in all number of prediction candidates. Surprisingly, *Gmail – Approach* almost reaches the accuracy of PR model in every number of prediction candidates. This caused by there are almost no any incoming call information that included in Chunghua Telecom dataset.

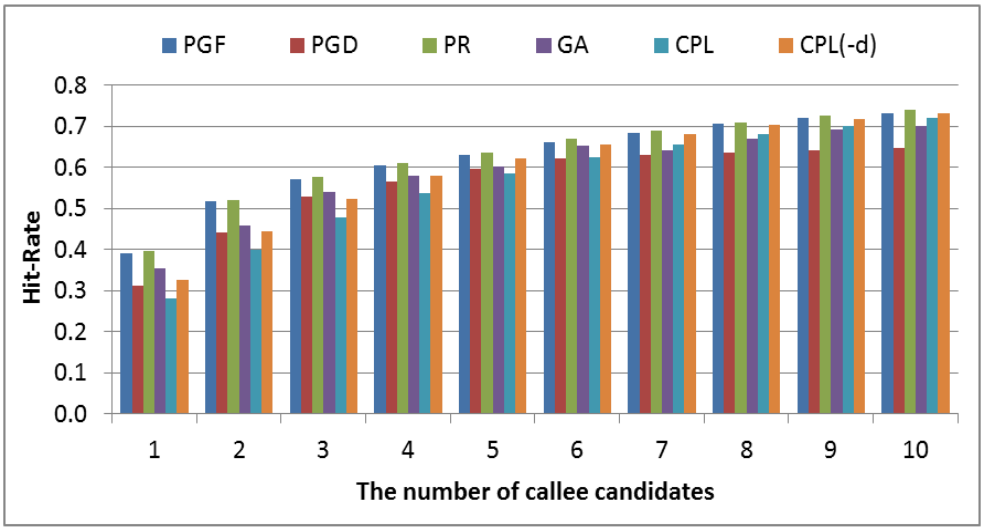


Figure 5.1: Accuracy of prediction model in Reality Mining dataset

We also examine the effectiveness of all prediction models with different number of observation days (λ). We set $\lambda = 7, 14, 21, 30, 45, 60, 70, 80, 90, and 100$ days for Reality Mining dataset, while we only set $\lambda = 7, 14, 21, 30, and 45$ days for Chunghua Telecom dataset. By increasing the number of observation days, give an impact enlargement on population of prediction (e.g. increase number of callees). Fig. 5.3. shows CPL and CPL(-d) predict more accurate (10%) than other models when the number of observation day for 7 to 21, but both of them decrease dramatically as the number of observation days becomes larger. In contrast, the accuracy of the rest of models PGF, PR, and *Gmail – Approach* except PGD are slightly increasing as the observation day expanded and stabilized in $\lambda = 60$. However, the accuracy PGD model getting increasing from beginning until λ reach 45, then it get down slightly as

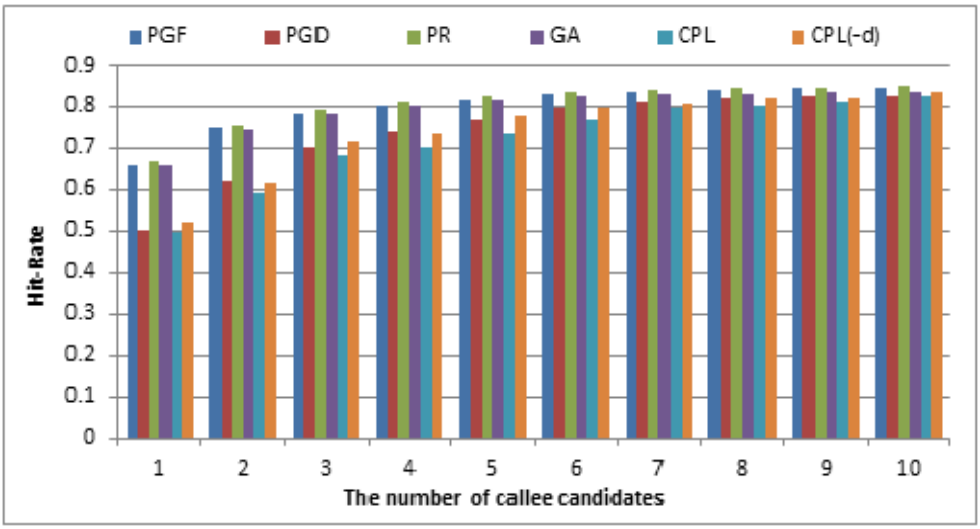


Figure 5.2: Accuracy of prediction model in Chunghua Telecom dataset

the increasing the number of observation days. Overall, PR model significantly outperform than other models from 45 to 100 observation day and this result also implies that either CPL or CPL(-d) model seem cannot handle bigger population of callees.

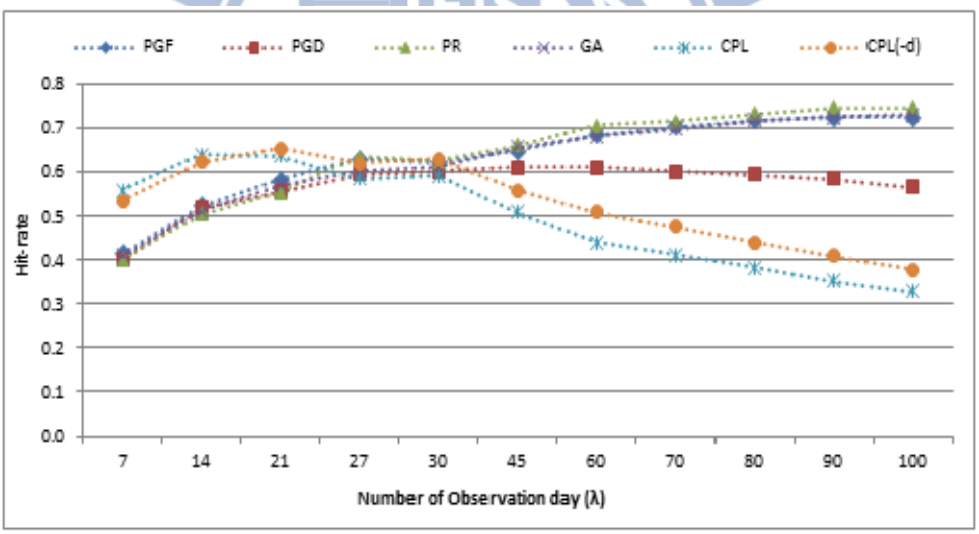


Figure 5.3: Prediction accuracy comparison under Reality Mining dataset with different number of observation day(λ)

5.4.2 Sensitivity Analysis

In this section we evaluate some sensitive factors that might contribute either negative or positive impact to outgoing call predictors. The diversity of users connectivity, users call-

ing population, regularity users calling behaviour, and users plurality on interconnectivities behaviour are the factors that we are going to explore in this coming subsection.

Impact of calling diversity

To better understand the context of each observation user, it would be useful to have information of users calling diversity. We use the concept of Entropy measurement to quantify the diversity of users calling behaviour. The main objective of this measurement is to be able to distinguish the users based on the callee population that user used to contact with. So that, we will be able to see the impact of callee population on prediction accuracy, where the user with high entropy clarifies he has many callees that he used to contact with, otherwise he has only a few callees that he usually to interact. To measure the $UserEntropy$, let begin with some formal definitions. Let U be a user and C_U is set of callee that at least received one call from the user U. For a callee $c \epsilon C_U$, let to define Z_c as the number calls that callee c received from the user U and Z_U is the total calls user U made. Therefore, the probability a callee c receive the call from user U is $P_c(U) = \frac{|Z_c|}{|Z_U|}$, where $P_c(U)$ is the total fraction of all receiving calls to callee c from the user U. So, the computation of $UserEntropy$ accordingly to the following equation:

$$UserEntropy(U) = -\sum_{c \epsilon C_u} P_c(U) log P_c(U) \tag{4}$$

For concrete illustration of the difference levels of users calling diversity, see Fig. 5.4.. It shows the difference between three users with difference levels of entropies. The diverse of shape of point represents the difference callees and the number of point represents frequency call from each user to the callees. Fig. 5.4(a). illuminates the user with low level entropy which he only interacts with one callee and Fig. 5.4(b). illustrate the user with medium entropy. The user with high entropy, where he has many callees, is indicated by Fig. 5.4(c).. With simple word, a user will have high entropy if he made many calls to many difference callees. Conversely, the user will have low entropy if he interacts frequently focus on few callees.

To illustration the impact of increasing of user calling diversity on all prediction methods, we randomly select 25 users from reality mining dataset as a sample. Fig. 5.5. shows the diversity level of observation users, where the horizontal axis represents the observation users

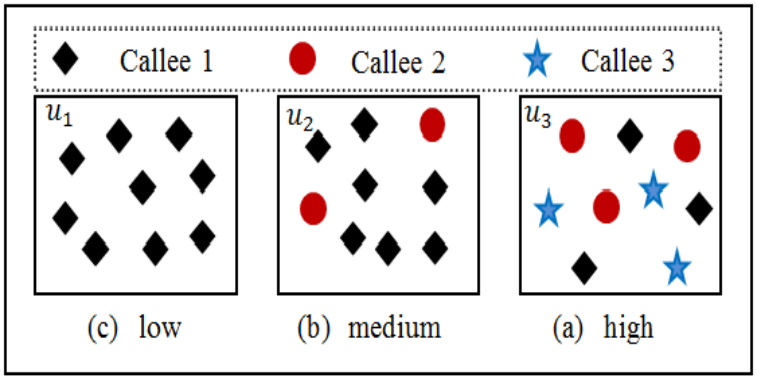


Figure 5.4: Illustration of user calling diversity

and their frequency calls, and vertical axis represents the level of user entropy. The user entropy is not affected by the frequency call, but it strongly influenced by interaction of user to many callees. The user with highest entropy (0.705) represented by user33, while the lowest entropy (0.345) represented by *user*60. According to the level entropy on Fig. 5.5., those sample users are divided to 3 categories users: first, $low-level$ are the users which their entropy value place in between 0.3 0.5. Second, $medium-level$ where the entropy value of user be in between 0.5 0.6, finally the users who have entropy value up to 0.6 is categorized as $high-level$.

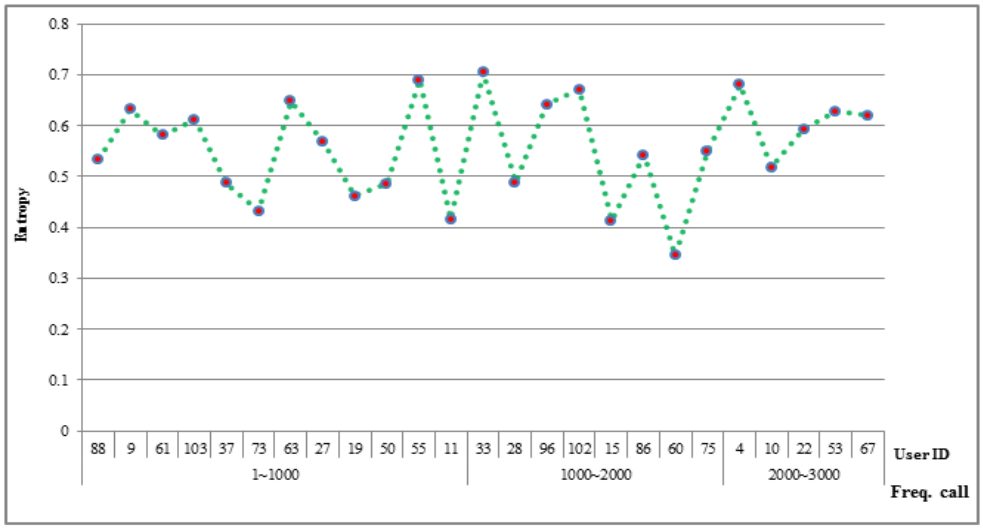


Figure 5.5: Entropy level each of observation user

Fig. 5.6. shows the prediction accuracy comparison under difference level of user entropies with difference setting value of λ ((a)λ=21 days, λ=30 days, and λ=60 days). In Fig. 5.6(a), CPL and CPL(-d) models perform better accuracy than our proposed models on low-level

and middle-level data. But on high level-data, our proposed models perform as well as CPL and CPL(-d) model. CPL and CPL(-d) still outperform than our proposed models when the observation day (λ) is increased to 30 days in low-level data (Fig. 5.6(b)), but only CPL(-d) model perform better than others model for middle-level data. Fol high-level data, our proposed models achieve better accuracy than existing models. Meanwhile, our proposed models PGF and PR, outperformed than other models in every level of data. For the middle-level users, all models getting decrease in average 10% except CPL and CPL(-d), where both of these model decline significantly (approximately 40% for CPL and 30% for CPL(-d)) and getting even worst on high-level user. Once again, this result implies inability of CPL or CPL(-d) in case for handling a lot of prediction. For PGF and PR in case for $high - level$ users, even getting more decreasing, but they still outperformed than other models. Both of these models almost reach 60% of accuracy. Overall, PGF and PR are still reliable in many situations of datasets.

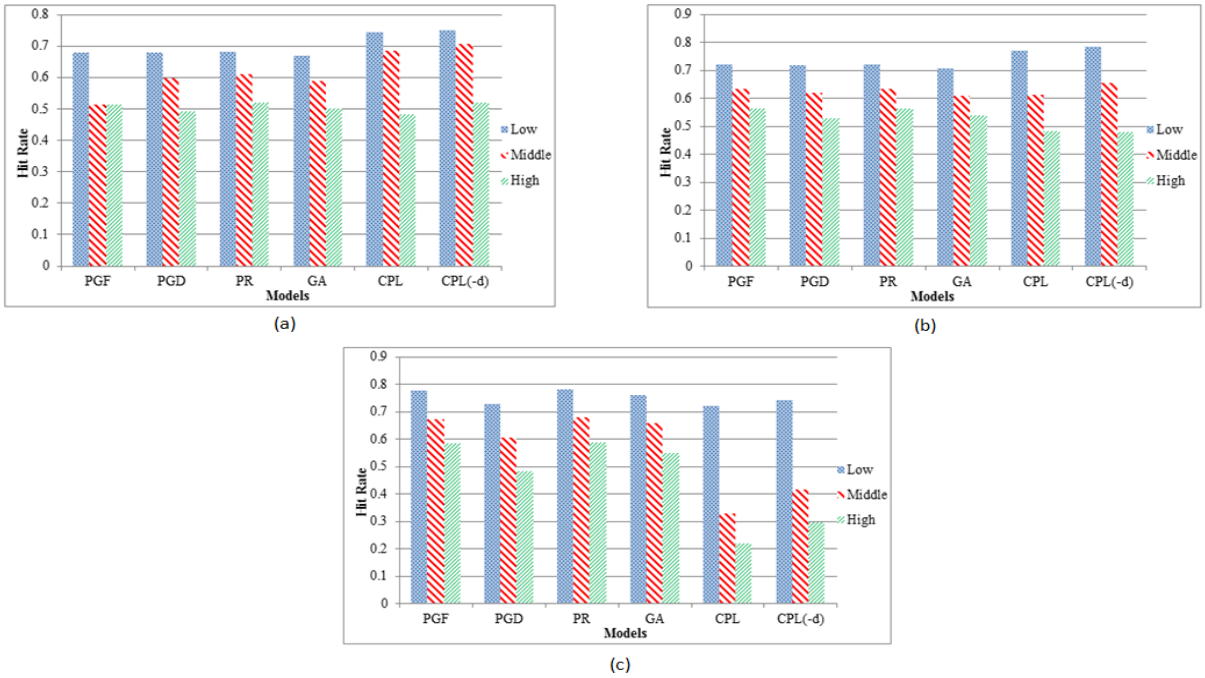


Figure 5.6: Prediction accuracy for difference categories of users with difference setting of observation day(λ): (a)λ=21 days, (b)λ=30 days, and (c)λ=60 days

Impact of the regularity of users routine

In studying the properties of user calling behaviour that related to accuracy portion, we explore the attribute of the regularity of user schedule. The regularity of user schedule is calling density of a user to specific callee at exactly time slot in hour of a day time. To be more clearly, let begin with an example. Suppose you used to call your colleges during day time and your parents every night before going to bed, so the density of your calling activity will be high during day time and low at night time. Since the number of callees is heavily at the day time, the regularity of your calling activity will be detected which we define this situation as irregular schedule calling time. Otherwise, your night time we define your regular schedule calling due to vary specific callee that you used to contact with. To measure the regularity users calling schedule again we use the concept of entropy. Given time-slot the historical call log of the user and number of time-slot t (where $t = 1, 2, 3, , 24$), the $userSchedule$ defines as in the following equation:

$$UserEntropy(U) = -\sum_{c \epsilon C_u} P_c^t(U) log P_c^t(U) \tag{5}$$

where $P_c^t(U)$ is the probability of user U call callee c ($c \epsilon C_U$), $where C_U$ is all callee who receive at least one call during observation day) at given slot-time t. To the best of description of user schedule, we select one user from each category ($user60$ from $low - level$, $user75$ from $medium - level$, and $user96$ from $high - level$) that mentioned in previous subsection as a sample which shown in Fig. 5.7.

Fig. 5.7. gives vary clear description of three differences users in calling activities. Where, $user60$ we define as the user who has more regularity in calling behaviour during 2 pm until the entire night time and has very heavy calling activity during 9 oclock to 12 oclock. $User75$ seem has not so regular schedule in calling activity since his entropy value distributes evenly in the whole day. Meanwhile, $user96$ seem has tightly call activity during 11 oclock in evening until 6 oclock in morning and he has less call activity in slot time of 11am and 3 pm.

To clarify the accuracy prediction in each time slot, Fig. 5.8. show the accuracy of all model prediction on each time slot of $user96$. In general, we can say that almost all accuracy of all the models seem reasonable, where when the entropy value is high the accuracy down,

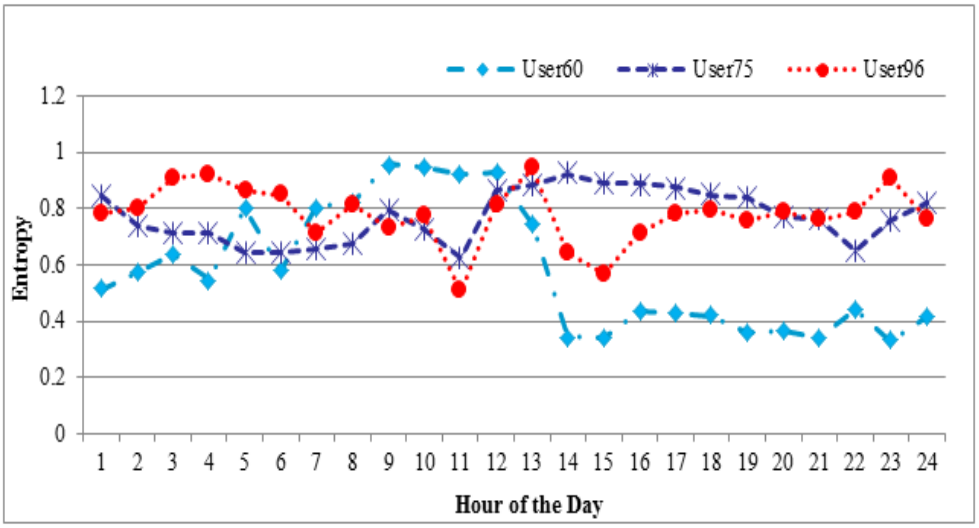


Figure 5.7: Entropy value of three selected users ($user60$, $user75$, and $user96$) in each time-slot

otherwise the accuracy is increasing. Overall, the model of PGF, PR and GA look dominant in every slot time.

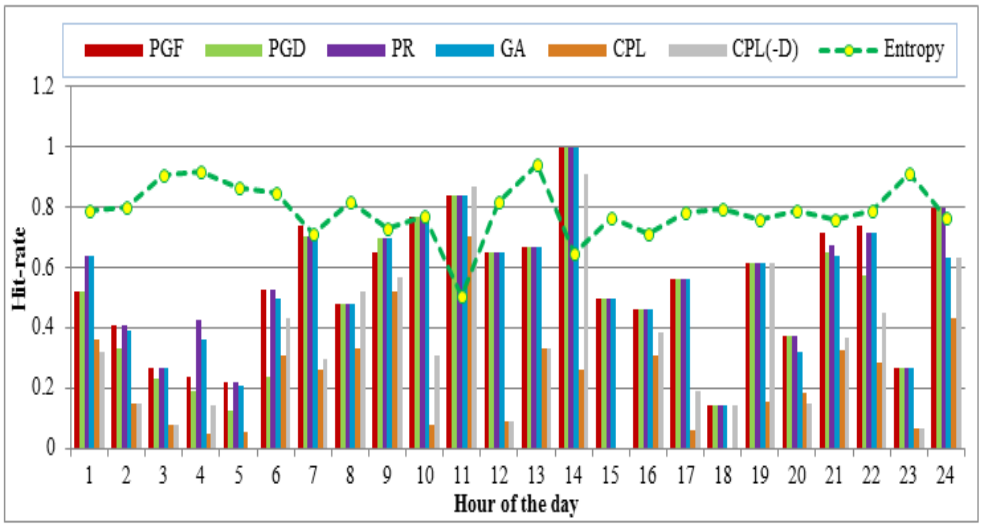


Figure 5.8: Accuracy prediction of all model of users $user96$ in each time-slot

Impact of the variety of communication

As we mentioned in the beginning of this chapter the Reality Mining dataset consist of three difference types of user communication: $Voice-call$, $short-message$, and $data-transfer$, where the volume portion of $voice-call$ type higher than two other types. In this subsection we want to find out the impact of using variety of communication ways to the outgoing

call prediction. According to the result that we show on Fig. 5.9., $short - message$ and $data - transfer$ give a positive contribution on outgoing call prediction, even very slightly on PGF, PR, and GA. Otherwise, with the increasing of prediction volume by considering short massage and data transfer, give a negative impact for CPL and CPL(-d). Again, we prof that CPL and CPL(-d) get negative impact on dealing with bigger data prediction.

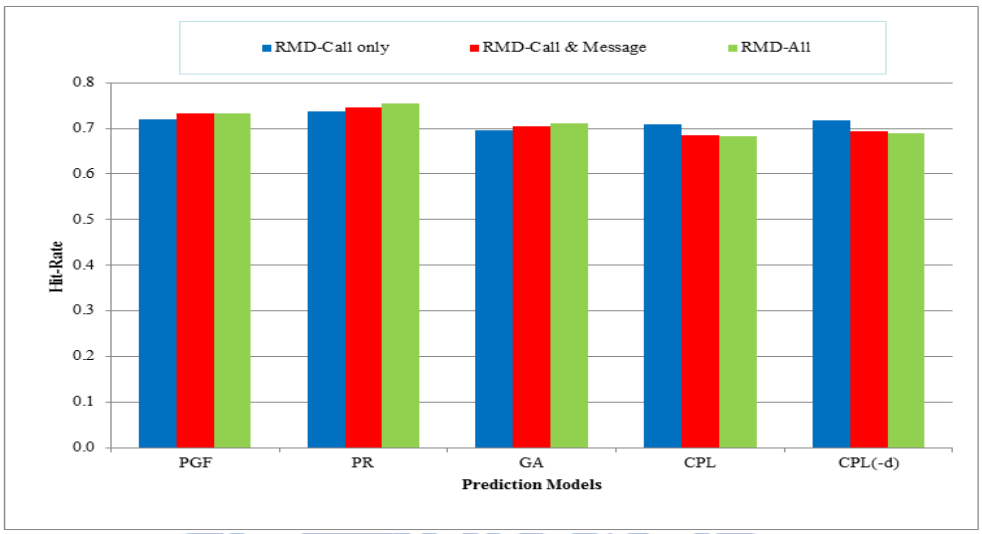
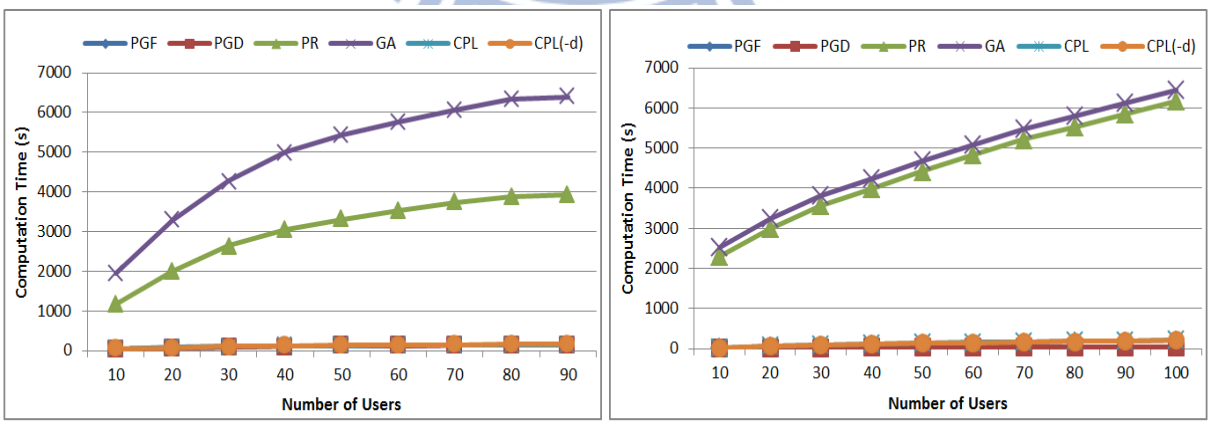


Figure 5.9: Impact of the variety of communication to outgoing call prediction. $Short - message$ and $data - transfer$ give a positive impact to our proposed model, but inversely proportional to the existing model

5.4.3 Time Computation



(a) Reality Mining Dataset (b) Chonghwa Telecom Dataset

Figure 5.10: Computation time for all models: (a)Reality Mining Dataset and (b) Chonghwa Telecom Dataset

In this subsection we investigate the computation time for all the models. Fig. 5.10 shows the computation time of all the models with Reality Mining Dataset5.10(a) and Chonghwa Telecom Dataset 5.10(b). In both of these datasets PR and GA models cost very significant computation time than the other models. Either PR or GA model require for checking the date of every single communication of user to ensure the weight of the recentness of the phone call. Meanwhile, the rest of models PGF, PGD, CPL and CPL(-d) entail almost the same computation time. Nevertheless, our proposed models PGF and PGD perform faster 10 seconds than CPL and CPL(-d) in every different number of users.

Chapter 6

Conclusion

In this paper, we have presented the model to predict users outgoing call intention at the query time slot. After observing on the historical call logs, we have discovered four features *e.g.* $frequency$, $duration$, $recency$, and $direction$ that are potential to be used in modelling call intention. Based on those features, we have proposed three conventional predictive based probability models: $Probability\ General - Frequency(PGF)$, $Probability\ General - Duration(PGD)$, and $Probability\ Recency(PR)$.

From this study, we have learned that the accuracy of each model strongly depends on the volume of historical communication log. In another word, when the number of prediction is increasing, the accuracy of predictors is declining. We also have founded that the regularity of user's calling activities assist a positive impact for prediction model. More regular the user call the same callees at the same slot time, more higher accuracy can be achieved by each models. In studying the correlation of using a variety of communication data, we have explored that another kind of communication data such as messaging and data-transfer can contribute a positive impact for our proposed model.

Finally, we have compared the effectiveness of our models with the existing model. We have demonstrated that our proposed models achieves the same accuracy as or better accuracy than the existing model. We also found that the existing model sensitively varies on handling

a bigger data.

Bibliography

[1] L. Akoglu and B. Dalvi. Structure, tie persistence and event detection in large phone and sms networks. In *Proceedings of the Eighth Workshop on Mining and Learning with Graphs*, pages 10–17. ACM, 2010.

[2] B. H. Andrews and S. M. Cunningham. Ll bean improves call-center forecasting. *Interfaces*, 25(6):1–13, 1995.

[3] L. Bianchi, J. Jarrett, and R. Choudary Hanumara. Improving forecasting for telemarketing centers by arima modeling with intervention. *International Journal of Forecasting*, 14(4):497–504, 1998.

[4] L. Bianchi, J. E. Jarrett, and R. Choudary Hanumara. Forecasting incoming calls to telemarketing centers. *Journal of Business Forecasting Methods and Systems*, 12:3–3, 1993.

[5] J. M. Boulin. *Call center demand forecasting: improving sales calls prediction accuracy through the combination of statistical methods and judgmental forecast.* PhD thesis, Massachusetts Institute of Technology, 2010.

[6] V. R. Carvalho and W. W. Cohen. Ranking users for intelligent message addressing. In *Advances in Information Retrieval*, pages 321–333. Springer, 2008.

[7] D. Choujaa and N. Dulay. Predicting human behaviour from selected mobile phone data points. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pages 105–108. ACM, 2010.

[8] N. Eagle, A. S. Pentland, and D. Lazer. Mobile phone data for inferring social network structure. In *Social computing, behavioral modeling, and prediction*, pages 79–88. Springer, 2008.

[9] J. Froehlich, M. Y. Chen, S. Consolvo, B. Harrison, and J. A. Landay. Myexperience: a system for in situ tracing and capturing of user feedback on mobile phones. In *Proceedings of the 5th international conference on Mobile systems, applications and services*, pages 57–70. ACM, 2007.

[10] C. E. Harless and T. J. Kowalski. System and method for correlating incoming and outgoing telephone calls using predictive logic, July 4 2000. US Patent 6,084,954.

[11] S. Phithakkitnukoon and R. Dantu. Towards ubiquitous computing with call prediction. *ACM SIGMOBILE Mobile Computing and Communications Review*, 15(1):52–64, 2011.

[12] S. Phithakkitnukoon, R. Dantu, R. Claxton, and N. Eagle. Behavior-based adaptive call predictor. *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, 6(3):21, 2011.

[13] M. Roth, A. Ben-David, D. Deutscher, G. Flysher, I. Horn, A. Leichtberg, N. Leiser, Y. Matias, and R. Merom. Suggesting friends using the implicit social graph. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 233–242. ACM, 2010.

[14] N. Singh, S. Bagchi, and Y.-S. Wu. Annoying telephone-call prediction and prevention, May 4 2009. US Patent App. 12/434,750.

[15] V. Soto, V. Frias-Martinez, J. Virseda, and E. Frias-Martinez. Prediction of socioeconomic levels using cell phone records. In *User Modeling, Adaption and Personalization*, pages 377–388. Springer, 2011.

[16] D. Zhang, A. V. Vasilakos, and H. Xiong. Predicting location using mobile phone calls. *ACM SIGCOMM Computer Communication Review*, 42(4):295–296, 2012.

[17] H. Zhang and R. Dantu. Predicting social ties in mobile phone networks. In *Intelligence and Security Informatics (ISI), 2010 IEEE International Conference on*, pages 25–30. IEEE, 2010.