ARTICLE IN PRESS

Pervasive and Mobile Computing ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Pervasive and Mobile Computing

journal homepage: www.elsevier.com/locate/pmc

Frequency and recency context for the management and retrieval of personal information on mobile devices

Vassileios Stefanis [a,b,*], Athanasios Plessas [a,b], Andreas Komninos [c], John Garofalakis [a,b]

[a] Department of Computer Engineering & Informatics, University of Patras, Patras, Greece
[b] Computer Technology Institute and Press "Diophantus", Rion, Patras, Greece
[c] Glasgow Caledonian University, Cowcaddens Road, Glasgow, UK

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Context
Mobile personal information management
Call prediction

ABSTRACT

As users store increasingly larger amounts of personal information on their mobiles, the task of retrieving such items (e.g., contacts) becomes more difficult. We show that users can be categorized by their communication patterns and that each category benefits differently from supporting contact management applications. By examining mobile user call logs, we show that it is possible to aid retrieval tasks using relatively simple heuristics and algorithms that describe usage context, using solely the dimensions of contact use frequency and recency. We compare and discuss the results of the proposed method applied on two different mobile datasets: a large dataset from NOKIA and a smaller dataset collected by ourselves.

1. Introduction

Technological advances of the last decade have turned mobile phones into small multi-purpose personal computers being equipped with a camera, GPS receiver, accelerometer, Bluetooth, and other sensors. These devices are now used, among others, to access the World Wide Web and transfer files, and as email clients and calendar reminders. However, mobile phones are still primarily considered as communication devices. As such, some of the most common tasks for their owners include searching for a contact in a phonebook or selecting one from a recent call list [1]. As the contact lists become increasingly bigger, and since a significant number of contacts are never actually used [2], the cognitive load on the user increases while trying to retrieve a contact from an expanding dataset. This effort is further obstructed by the limitation of the relatively small screen that mobile phones are equipped with, which restricts the information presentation space. Furthermore, since call logs impart information about use and not *lack* of use, mobile devices have become good at supporting communication but provide little support for the task of managing social relationships (i.e., deciding who to contact and how frequently), leaving decisions entirely to the users.

Mobile devices can collect a significant amount of data and information about the user's context, including location, current date and time, the orientation of the device, whether the user of the device is on move and his/her speed, the user's current task (e.g., on the phone, messaging), whether vibration or silent mode is enabled, etc. [3]. Devices can capture a lot of personal information related to the user's social environment [4], generated either automatically by the device (e.g., a phone list saves the calls that have been made, the time of day for each call, and the duration) or by the user (e.g., SMS/MMS

* Corresponding author at: Department of Computer Engineering & Informatics, University of Patras, Patras, Greece. Tel.: +30 2610960423.
E-mail addresses: stefanis@ceid.upatras.gr (V. Stefanis), plessas@ceid.upatras.gr (A. Plessas), Andreas.Komninos@gcal.ac.uk (A. Komninos), garofala@cti.gr (J. Garofalakis).

1574-1192/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.pmcj.2013.08.002

and multimedia files, browser's history, calendar events, etc.). Therefore, a mobile device could also be aware of the social environment of the user (social context). The combination of social and mobile context results in a dynamically defined social context, termed the *mobile social context* [5].

In this paper, we deal with the problem of personal information management as applied to the users' contact lists, and, more specifically, the task of retrieving the desirable contact when performing an outgoing call. Our hypothesis is that this task can be aided by a mobile interface that can predict which contact is the most probable to be called at any time, and thus offer speedier access to it. We believe that a solution to the retrieval problem can be informed by mobile social context, as mobile users seem to adopt different behaviour patterns under different contexts. As an example, consider the following scenarios derived from our experiment participants that involve two different context dimensions, frequency and recency.

Scenario 1. *George is from Greece but works and lives in the UK. George's contact list is in both English and Greek, and the constant switching between languages makes searching for contacts quite bothersome. Thus, he relies on his call log for retrieving contacts, but, given the large number of calls he receives each day, he frequently has to scroll up and down a lot before he can find the contact he wants to call. Though not optimal, he prefers this style of interaction, as he perceives it to be less annoying than switching languages and searching.*

Scenario 2. *Maria is a Ph.D. student at the University. She rarely calls her supervisor on his mobile phone; however, today he called her to arrange a meeting. After the meeting, Maria tried hard to remember the name of a paper her supervisor recommended but she could not. She had to phone him again, as he was out of the office for the whole day.*

By modelling user "*socialness*" using actual call logs, we show later in the paper that the majority of users are significantly suboptimally served by existing contact management applications, such as the recent call list and the frequently used contacts list. As a first step in our research, we consider only two context dimensions, frequency and recency of use of each contact, as demonstrated by the scenarios above. The paper focuses on examining the effect of each dimension on the success of predicting the likelihood of a contact being called. We show that the combination of these context dimensions provides better prediction results than traditional access modes (list of recent calls, list of more frequently used contacts), something that has implications in the design of better interfaces for communication support and contact list management. Moreover, by comparing the results of our approach to those of the algorithms presented in the section on related work (directly applying two of them on the same dataset), we show that our algorithm, although simpler, outperforms existing more complex solutions. By presenting a preliminary analysis on the effect of adding a third dimension (time of day) in our algorithm, we argue that incorporating more dimensions in a prediction model does not necessarily provide better results, and that a thorough analysis of each candidate dimension seems to be required.

The rest of the paper includes a review of related work, followed by a section on methodology that justifies our context dimension choices, describes the datasets used, introduces the "*socialness*" metric for users and details our prediction algorithm. We present our findings in detail in Section 4. Section 5 presents a preliminary experiment with the addition of a third dimension (temporal context), while in Section 6 we compare our approach with other algorithms. Finally, in Section 7, we discuss these findings and provide our conclusions and suggestions for further work.

2. Related work

Users daily create, receive, and store significant amounts of personal data, whose organization and retrieval becomes a difficult task due to their increasing size. Personal information management (PIM) is an important research area not only for the case of desktop computers, but also for mobile devices. One of the primary reasons is that users are reluctant to remove old items from their computing devices, resulting in an ever-expanding collection that hinders search and retrieval [6]. Handheld devices are considered trusted devices in which several types of personal information item are stored (contacts, photos, music, videos, notes, tasks, etc.). They also impose an additional burden to PIM due to limitations such as screen size, input/interaction modes, and navigation [7]. Myers et al. [8] stress the need for mobile users to access quickly the right information at the right time and highlight how important it is for PIM tools to help users accomplish their tasks efficiently. According to Zhou et al. [7], existing mobile PIM tools require extensive involvement of human users, and as a result managing personal information such as to-dos and contacts consumes more time than needed. However, as the information needs of users highly depend on their context while on the go [9], context can be adopted in order to enhance PIM systems [10].

To the best of our knowledge, although the idea of taking advantage of context to provide adaptive services to mobile users is not new, and while contact lists are one of the most frequently used applications on mobile devices [11], little research has been conducted on predicting the next call a mobile user is going to make and providing a rearranged contact sublist to replace traditional methods of contact repository access.

In [1], an algorithm that builds an adaptive speed-call list based on call logs is presented. Based on the observation that outgoing communication follows a periodical pattern, five dimensions (day of week, weekend/weekday spans, time of day, dayparts of a day, 1-hour slots of a day) are proposed as recommendation conditions. Whenever a user presses the call button, the algorithm computes the Bernoulli probabilities of each dimension for each contact, sorts all contacts according to the respective conditions' maximum probabilities, and creates the speed-call list. However, the probabilities of the proposed dimensions are considered separately, and are not combined as in our approach.

Please cite this article in press as: V. Stefanis, et al., Frequency and recency context for the management and retrieval of personal information on mobile devices, Pervasive and Mobile Computing (2013), http://dx.doi.org/10.1016/j.pmcj.2013.08.002

In [12], a similar approach to predict outgoing calls analyzing mobile phone historical call log data is described. Three dimensions are proposed to capture the frequency and regularity of communication behaviour. An important finding from this research is that combining factors rather than exploiting each one independently leads to better results. The proposed algorithm analyses historical data from a period of two years, a decision that adds computational load to the device and seems to be unnecessary, since only a recent portion of communication history is needed to predict future behaviour [13]. Moreover, the weights that are assigned to each dimension seem to be arbitrarily decided, and the success rate is quite low (below 40%) for the period of five weeks that the experiment was running, following however an upward trend.

Another attempt to predict outgoing calls of mobile users is described in [14]. The researchers have implemented a call predictor for both incoming and outgoing calls. The outgoing call predictor constructs a probabilistic model capturing the user's behaviour based on call departure and inter-departure times. The prediction algorithm provides good results (for example, a success rate around 70% for a prediction list with five entries). Although the researchers prove that only recent history is needed to predict future communication behaviour, it is not clear whether the predictor takes into account all historical call data or only a recent portion of the call log. The probabilistic approach presented is promising; nevertheless, it seems difficult to incorporate other mobile and social contextual dimensions, such as location or personal preference. The same researchers also propose another approach to predict outgoing calls [15], based on a naive Bayesian classifier, considering as important factors of the calling pattern the time period of day, the day of each call, and the "reciprocity" (call interaction between the user and the caller). This approach seems to provide slightly worse results than the one presented in [14].

3. Methodology

Our research is based on two different mobile datasets: a large dataset from the NOKIA Lausanne Data Collection Campaign [16] and a preliminary dataset that we have collected during an unrelated experiment that was organized separately. This section first presents the mobile handset-based data collection method from our experiment and introduces the data collected, as well as some preliminary observations. Subsequently, the larger dataset from the NOKIA campaign is described. Then, we present in detail the context-based prediction procedure.

First, we present our rationale for choosing the context dimensions that will be employed in our methodology. When considering context dimensions that might influence retrieval needs, it is easy to imagine that such dimensions could include location, frequency, recency, time of day, day of week, and personal preference (user indicated favourite contacts). Obviously, location was not available in our preliminary dataset, as the Android OS does not record user location for phone call events. While location is available in the NOKIA dataset, this is recorded sporadically (hence not all calls could be associated with a location). Additionally, with many locations comes inherent uncertainty, as they are captured with a variety of methods with variable granularity and accuracy, depending on which sensors were available at the time on a user's mobile (e.g., Wi-Fi, Cell Tower, and GPS). Our work would hence need to incorporate an uncertainty management component, which is beyond the scope of this paper. Thus, we chose not to employ location in order to be able to directly compare between the two datasets.

In addition, in previous work [17], we found that the role of personal preference ("starred" contacts in Android) is ambiguous, since there is no correlation between the "starred" status of a contact and the probability of a call being made to that contact; hence we discarded this dimension. Temporal context (time of day and day of week) might depend on other dimensions such as social relationship between user and contact, cultural norms, and user activity, and can be an indicator of user or contact location etc.; thus it cannot safely be examined on its own without knowledge of these other types of context, which of course are not available in our dataset. To support this argument, we present in Section 5 a preliminary experiment that shows that the ad hoc addition of further dimensions does not result in improvements. For all the above reasons, the safest dimensions that could be used as a first step to establish a baseline performance for a predictive system were frequency and recency of communication, the effect of which in calling behaviour is extensively studied in [17].

3.1. Preliminary mobile dataset and users

In order to extract real communication data from mobile phones, we developed an Android application that extracts the contact list and the call log from the mobile device. The application was delivered to 42 subjects (all of them located in Greece) with Android smartphones; however, only 25 datasets were considered as valid, since some were incomplete (e.g., extremely small number of records in call log, coverage period of log being too short, too few contacts in their contact list). Concerning the 25 subjects that we take into account in our analysis, 22 of them were male and 3 female, while their age ranges were from 19 to 39 years old, and they were from varied backgrounds, though most were Computer Science students. In total, the participants' contact lists contained 4185 entries. We found that, on average, each contact list contained 167.4 entries (mean $= 167.4$, stdev $= 87.60$, min $= 33$, max $= 344$). The extracted logs covered a different time period in days for each mobile phone (mean $= 52.80$, stdev $= 35.23$, min $= 18$, max $= 170$). On average, each user made 449.88 calls (stdev $= 98.12$, min $= 182$, max $= 500$). We should stress here that the Android platform limits the call log history to 500 calls.

3.2. NOKIA mobile dataset

The NOKIA Lausanne Data Collection Campaign is a large-scale initiative that took place in the Geneva area in Switzerland, from October 2009 to March 2011. The number of participants reached 185 (38% female, with two thirds of the population

Please cite this article in press as: V. Stefanis, et al., Frequency and recency context for the management and retrieval of personal information on mobile devices, Pervasive and Mobile Computing (2013), http://dx.doi.org/10.1016/j.pmcj.2013.08.002

aged between 22 and 33 years old) and data related to location, motion, proximity, communication behaviour, application usage, etc. were collected, turning it into a rich mobile usage dataset. In this period, more than 240,000 calls (incoming/outgoing/missed) were logged.

In our analysis, we focused on communication data, and specifically on users' call logs. However, as previously, some users participating in the campaign had to be excluded, as they were not considered valid for call log analysis. Some of them had a call log covering an extremely short period of time, while others had too few call records or a very sparse call usage of their phone. This is natural, since, as described in the documentation provided by NOKIA, there were users who left the experiment early or others who often decided to turn off the recording software. After removing those with the undesirable extreme characteristics (call log period < 30 days, call records < 100 and calls per day < 2.25—the minimum value for the dataset from our experiment) we ended up with a dataset that consisted of 106 users. We cannot report on the average size of the contact list, since the NOKIA dataset is an evolving log and not a "snapshot" as our preliminary dataset, meaning that the number of contacts per user varies through the period covered. As in our experiment, in this dataset also the logs covered a different period in days (mean $= 374.98$, stdev $= 136.78$, min $= 33.96$, max $= 608.25$). On average, each user made 1928.75 calls (stdev $= 1036.11$, min $= 214$, max $= 5101$).

3.3. The "socialness" metric

Lee et al. [1] found that users in their study fell within two groups based on their perceived "*socialness*". Having selected a suitable sample of users, we then wanted to see if they could be organized into clusters, based on their communication behaviour and perceived "*socialness*". The need for this categorization is made greater because of the nature of the call logs, which have varying lengths and densities. Since it is not desirable to dilute the call logs by massaging the raw data (e.g., by normalizing), arranging the users into categories is a step towards ensuring the integrity of our conclusions. The obvious first step would be to look at the sparsity of communication, i.e., the number of calls made per day; however, this would not show emergent behaviour in terms of the "*socialness*" of a user. For the rest of the paper we use the term "*socialness*" not literally but to express the pattern of incoming and outgoing communication from the user's mobile device. A user could make many calls per day; however, these could be to a very few distinct contacts. Other, more "social" users could also make many calls per day, communicating with many distinct contacts, while there are users that could make only few calls, communicating only with a small number of contacts. Our purpose was to define a metric that reflects for each user this social pattern of communication. More specifically, we were interested in examining how the calls to contacts were distributed on two dimensions, i.e., the *number* of called contacts and *relative frequency* of calling for each contact, using a single value to represent this pattern. As a result, we needed the value of this metric to get influenced by the number of contacts called and the relative frequency, in a way that represents the aforementioned notion of "*socialness*". We thus analyzed each user's call log to examine the percentage of calls made to each of their contacts, as follows. For each user, we took a list of all contacts to whom calls had been placed for their entire call log duration. For each of these contacts, we calculated the percentage of calls made over the total calls made by the user. Having sorted these contacts and their percentages in descending order, we calculated for each contact the difference from the previous contact (e.g., if there were three contacts with a percentage of outgoing calls of 80%, 15%, and 5%, the differences would be 80%, 65%, and 10%). We then calculated the means of all differences and used that as a metric to determine "*socialness*". The lower the metric, the more "social" a user is, since this means the user calls more people and with less frequency. As an example, consider the following case: user A for a given period of time communicates with only three contacts (40%, 40%, and 20% respectively), while for the same period user B communicates with six contacts (with relative frequencies 40%, 20%, 15%, 10%, 10%, and 5%) and user C with ten contacts (20%, 20%, 15%, 15%, 10%, 8%, 5%, 3%, 2%, 2%). Then, the "*socialness*" metric S shows that user C ($S = 3.8$) is the more "social", user B ($S = 12.5$) follows, and user A ($S = 20$) is the least "social" user.

We ran a k-means clustering algorithm on these users, and found that users cluster optimally in three groups, for both datasets. In the preliminary dataset, Group 1 (3 users) is the "least social" users, i.e., people who tend to make most of their calls to just a handful of numbers, and we hypothesized that they are therefore most likely to exhibit regular predictable behaviour. Group 2 (9 users) is the "averagely social" group, while Group 3 (13 users) is the "most social" users, in constant communication with a variety of contacts, and thus likely to be most difficult to predict. In the NOKIA dataset, the "least social" group (Group 1) consists of 4 users, while the "averagely social" (Group 2) and "most social" (Group3) groups consist of 20 and 82 users, respectively (see Table 1).

3.4. Prediction procedure

In order to evaluate the role of the frequency and recency metrics in predicting the likelihood of placing a call to a contact, we used the extracted datasets from our users to perform a series of predictions, using the concept of a sliding *training window* within the datasets, which is used to make predictions regarding the next call. In [13], it is shown that a small subset of *recent* data from the historical dataset for training could be adequate to predict future behaviour, and hence we do not need to use a cumulatively expanding training set. As such, we define the training window t to be of a fixed temporal size, measured in *t-days*, i.e., temporal periods of 86 400 s (1 day). All calls made within this time window are used as training data, upon which we attempt to predict the person to be called next. Within this training window t, we also define a fixed *recency window* r measured in hours, which contains all the calls made in a fixed time period from the start of the training

Please cite this article in press as: V. Stefanis, et al., Frequency and recency context for the management and retrieval of personal information on mobile devices, Pervasive and Mobile Computing (2013), http://dx.doi.org/10.1016/j.pmcj.2013.08.002

Table 1
"*Socialness*" characteristics for the two datasets.

	Preliminary dataset			NOKIA dataset		
	Calls/day	Number of users	Average "*socialness*"	Calls/day	Number of users	Average "*socialness*"
Group 1	13.73	3 (12%)	0.09983	5.74	4 (3.8%)	0.09916
Group 2	11.85	9 (36%)	0.03155	5.66	20 (18.9%)	0.01956
Group 3	10.7	13 (52%)	0.01232	5.33	82 (77.3%)	0.00579

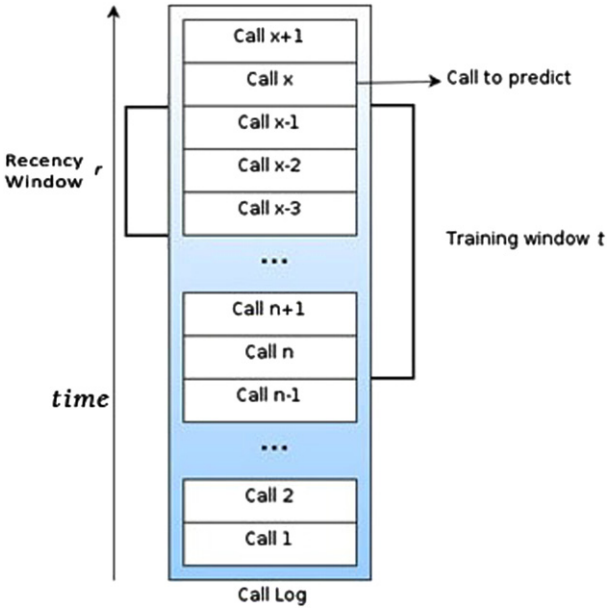

Fig. 1. Operation of the sliding training window.

window. All calls within this recency window r are used to give temporal significance to contacts, in contrast with training window t, which is used to capture the overall (historical) significance of a contact.

The prediction procedure works as follows. Suppose that we wanted to see if we could predict which contact was called at position x in the call log (Fig. 1), which must represent an outgoing call. Since we intend to support the retrieval task, we do not make predictions for incoming calls. We note here that predictions are made only for numbers that correspond to contacts, since there is no point in providing support for numbers that are not in the contact list. We then pick up all the calls (incoming and outgoing) in the call log that took place within the specified number of t-days from the timestamp of call x. We use both the incoming and outgoing calls to capture the effect of reciprocity as described in [18]. This training window t can thus have a varied number of calls, which are used as training data on which the prediction is made. Once a prediction has been made, we record the outcome, and move to the next call. This way, we work through the user's call log, call by call, and try to predict each one (obviously the earliest calls are only used as training data and not for predictions).

As can be seen, our algorithm's performance does not depend on the actual size of the user's contact list, but instead on the user's "*socialness*", i.e., the number of contacts that the user actually communicates with, and the temporal pattern of interactions with these contacts.

3.5. Calculation of prediction score

In our approach, personal information items such as contacts are represented as context augmented vectors $(x_1, x_2, \ldots, x_n)$, where x_i is the value of a context dimension i that characterizes the item. Our technique is based upon the context dimensions of contact use frequency and recency. Other dimensions of contact context can, of course, be incorporated in a predictive algorithm, and, in fact, in our conclusions we discuss how further context dimensions can possibly help with specific user and task conditions. However, the purpose of this paper is to investigate the role of these two dimensions of context; thus we focus solely on these. For each contact in the user's contact list, we assign a score, comprised of the sum of a weighted score $F(c)$ that reflects the frequency with which the contact has been used in the given training window, and a further weighted score $R(c)$ that reflects the temporal distance of the latest use of that contact within the training window. The equation used to assign a score to each contact is

$$\Pi(c) = w_f \times F(c) + w_r \times R(c),$$

where $\Pi(c)$ is the score assigned to the contact c, and w_f and w_r are the weights for the frequency score $F(c)$ and recency score $R(c)$, respectively. $F(c)$ is calculated as the percentage of the communications (incoming and outgoing) within the training window between the user and the contact. $R(c)$ is calculated as the percentage of the time interval between the start of a defined recency timeframe until the most recent communication between the user and the contact over the entire duration of the recency timeframe. In the case that there is no contact between them within the recency timeframe, $R(c)$ is zero. Thus, for each call, we can pick the top n contacts based on this score and offer these as likely candidates for our prediction.

4. Experimental results

4.1. Experimental considerations

Prior to proceeding with our experiments, we needed to determine an appropriate length for the training window that would be used, as well as to find a suitable temporal threshold for the recency score. Given the fact that our preliminary dataset concerned a significantly smaller period of time than the NOKIA dataset, we used the preliminary dataset as the determinant for these variables, since we wanted to run our experiments using the same parameters on both datasets. The recency threshold is desirable, as we have empirically found that recency of communication is influential, with a decayed effect, for a period of 6 h. Experimenting more with our preliminary dataset, we observed marginally better results for a timeframe for 12 h, so the temporal threshold was set to 12 h.

In [14], the researchers demonstrated that the accuracy of their call predictor did not improve in line with the size of the training data in their prediction work. This is a reasonable outcome, since, as people change behavioural patterns and perhaps interact more closely with different social groups during the course of time, older interaction data becomes not only redundant, but can be detrimental to the success of predictions. In our case, because the length of our preliminary dataset call logs was 52 days on average, and we needed to compare performance with the significantly longer NOKIA call logs on the same terms, we could not experiment with too large a training window. However, we did experiment further on the training window length using the NOKIA dataset alone, and we present our findings later. The training window should be long enough to provide adequate data but, at the same time, a training window of more than two weeks would likely fail to capture dynamic changes in a user's calling behaviour (e.g., taking a week off to go on holiday). Additionally, a training window of less than seven days would fail to include behavioural changes that could be attributed to weekends (and thus a change of social circumstances). We thus examined the performance of our technique, using equal weights for the frequency and the recency components, and compared the success of the technique with all possible combinations of a training window of 10 and 15 days and a recency threshold of 6 and 12 h. By examining the success means for all users, we found that our technique gave optimal scores with a ten-day training window and a recency threshold of 12 h, though the performance was not much better than in other combinations.

Finally, we needed to consider suitable suggestion list lengths for our experiments. In [14], several sizes of prediction suggestions (up to 20) are considered, though we felt that a mobile interface that would offer quick access to a likely desirable contact should not display more than eight suggestions, i.e., the approximate maximum list entries that can fit on a single mobile screen as shown by various modern devices [17], as this would force the user to further interact with the interface by scrolling, thus detracting from the usability of such a system. We decided to perform experiments for 1 (straight hit/miss), 3, 5 and 8 suggestions.

4.2. Baseline experiments

Apart from performing a search on the alphabetic phonebook, two other methods are usually available on a typical smartphone when its user wants to retrieve a contact in order to start a phone call: using the list with the *most frequently called contacts* or the list with the *most recent calls*. A simulation of these two methods is possible using executions of the prediction algorithm with pairs of weights ($w_f = 1$, $w_r = 0$) and ($w_f = 0$, $w_r = 1$), respectively. We consider these executions as a baseline experiment, the results of which, compared with the results of the executions where both dimensions are combined, reveal the improvement of our approach over the traditional contact retrieval methods. For the preliminary dataset, the average performance of the two methods for each user group and for different suggestion list lengths are shown in Fig. 2(a) and (b).

If we equate these performances to the use of the call log screen and the frequently used contacts screens on a real device, we would conclude that most "social" users (Groups 2 and 3) in the preliminary dataset could only expect to find the person they actually want to call at the very top of the call log around 40% of the time. Generally, success rates are between 53% and 61% for suggestion list size up to 8 entries. Presenting fewer suggestions seems to have a very small impact on the likelihood of retrieval. With regard to using a frequently called contacts list, users can find their desired contact at the very top of the list around 17–32% of the time, while an almost linear increase in performance takes the chance of finding the said contact to 66–80% somewhere in that list, if it contains up to 8 contacts. Group 1 (the least "social" users) exhibits a similar behaviour, but with much better performance (60–77% on the call log and 59–96% on the frequently called contacts list respectively, depending on the list size), as expected.

Please cite this article in press as: V. Stefanis, et al., Frequency and recency context for the management and retrieval of personal information on mobile devices, Pervasive and Mobile Computing (2013), http://dx.doi.org/10.1016/j.pmcj.2013.08.002

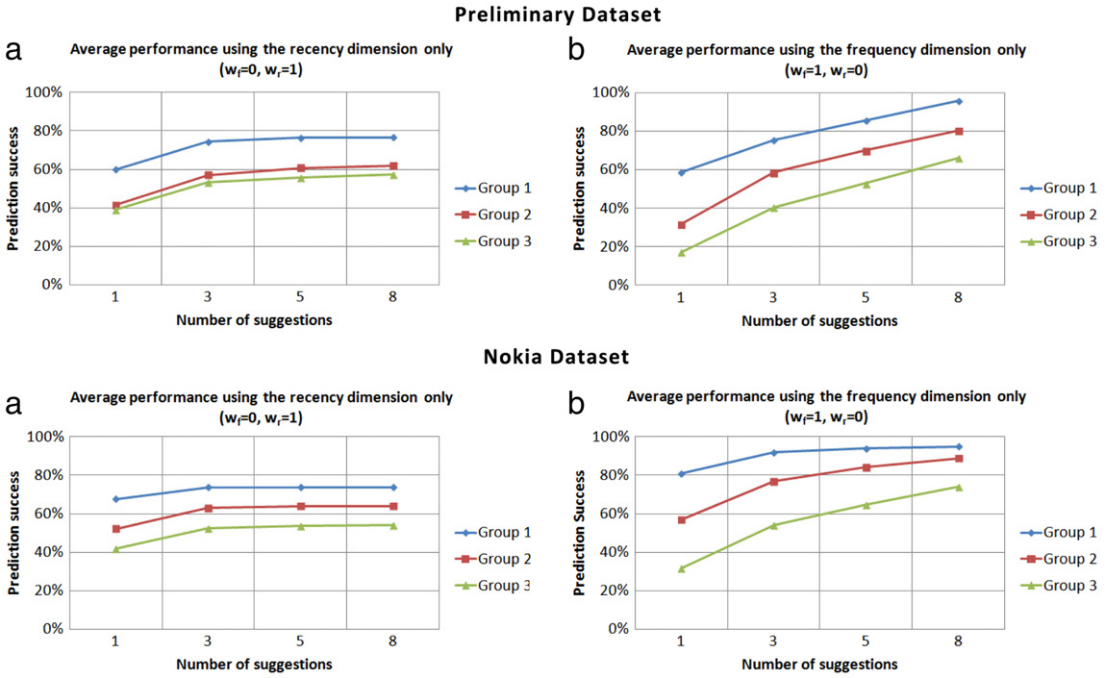

Fig. 2. (a, c) Performance using only the frequency dimension (equivalent to a list of most frequently used contacts). (b, d) Performance using only the recency dimension (equivalent to using the standard call log).

When performing the same baseline analyses in the NOKIA dataset (Fig. 2(c) and (d)), we observe that, though the figures are not quite the same, the behaviour overall is similar, and it confirms the findings of the analyses in the preliminary set. The least "social" group (Group 1) exhibits better performance that approaches 74% and 95% with just the recency or frequency weighting enabled, for 8 suggestions. Similarly, for Group 3 (most "social"), which is the largest group, the prediction scores are significantly worse in both cases. The results from these baseline analyses are revealing. First, when looking at the most familiar of contact retrieval tools (i.e., the call log), we notice that, for the largest groups and thus most users, even if the number of available suggestions (8) is large, there is, at best, only a 54% chance that the desired contact will be found there. In fact, it seems to matter little if more than three recent calls are present in the list, as the prediction success rates do not improve significantly with more than three candidates. We observe that using the "frequently called" list could yield better results, with the possibility of finding the actual desired contact therein increasing with the number of suggestions offered by the list. Even for Group 3, the success rate approaches 74% (NOKIA users). Interestingly enough, while the call log is a feature that is quickly accessible in most phones, the frequently called numbers list, which is likely to yield better results, is often less accessible, as it is hidden deeper in the contact list application structure. In any case, we show that the interfaces currently meant to improve access times to desired contacts and minimize the information retrieval problem are not optimal, and leave considerable room for improvement.

4.3. Actual experiments

Our experiments are divided into two distinct sets that explore the relationship between the importance of the frequency and the recency criteria, as discussed above. The methodology of the experiments remains precisely the same, except that in the first run (Set A) we are only interested in knowing whether the actual called contact is in the list of suggestions (a "hit"), while in second run (Set B) we keep a track of the position in the suggestion list that the contact is found, in the case of a hit. In this case, a suitable score to rate the quality of the prediction is given, which ranges in increments of one unit between [1..n] that reflect the number of positions available within the suggestion list (higher is better). For the calculation of the scores, we take into consideration just those circumstances where a hit has been achieved; thus the scoring reflects the quality of the "hits" and not a scored performance of the algorithm overall.

4.3.1. Experiment set A—hit or miss

The Figs. 3 and 4 show the success rates for all users. The average success for each suggestion list size for all users is shown in Fig. 3(a), while Fig. 3(b) shows the performance for all suggestion list sizes per group, and, finally, Fig. 4 shows the precise breakdown for each list size for all groups. The first conclusion that is immediately obvious is that using the frequency or the recency dimensions alone offers worse performance than any combination of weights. This indicates that

Please cite this article in press as: V. Stefanis, et al., Frequency and recency context for the management and retrieval of personal information on mobile devices, Pervasive and Mobile Computing (2013), http://dx.doi.org/10.1016/j.pmcj.2013.08.002

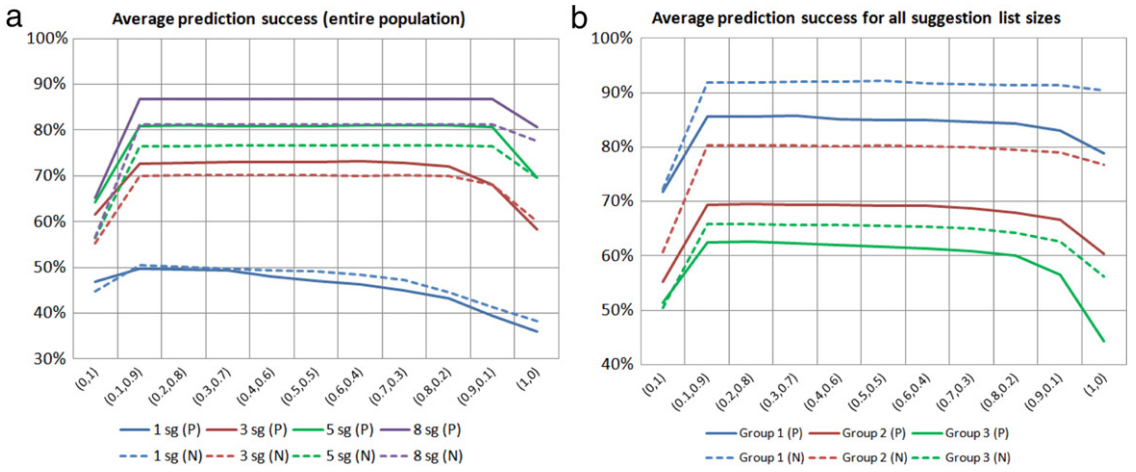

Fig. 3. (a) Average prediction success of the entire population, broken down by suggestion list size. (b) Average prediction success for all list sizes, broken down by group, for the preliminary (P) and NOKIA (N) datasets, for all (w_f, w_r) weight combinations.

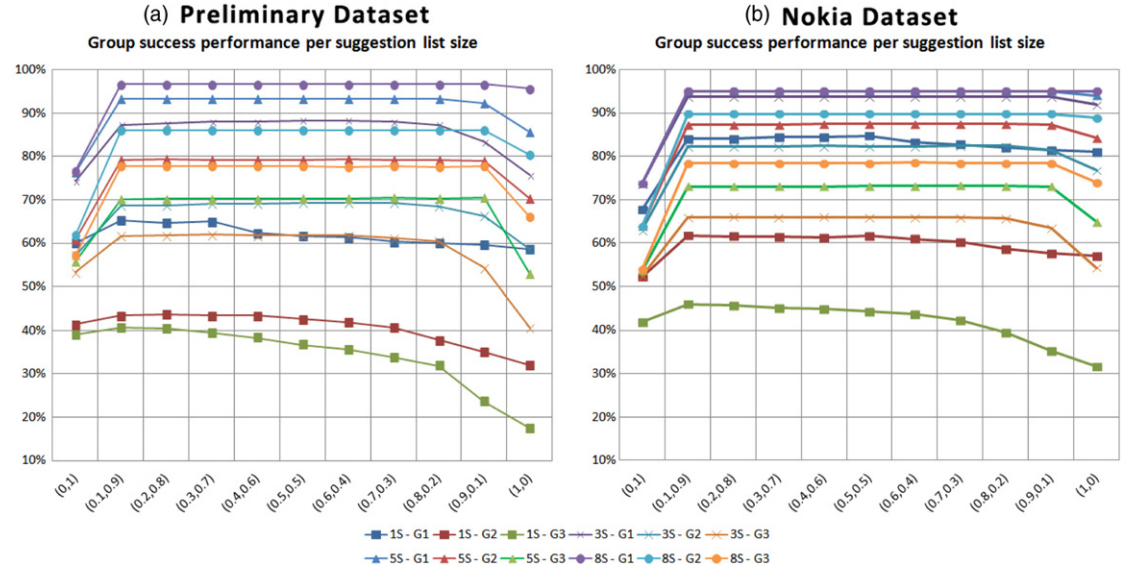

Fig. 4. Breakdown of each group's (Gx) performance for all suggestion list sizes (xS), for all (w_f, w_r) weight combinations. (a) Preliminary dataset. (b) Nokia dataset.

the standard mobile device screens that provide a call log and a most frequently used contacts view are less than optimal, and that an interface that would provide call suggestions based on both metrics is much more effective. It is clear also from these results that Group 1 has consistently the best performance, while Groups 2 and 3 follow. This confirms our hypothesis, as Group 1 exhibits the most predictable behaviours (frequent calls to a limited number of users). Additionally, we observe that, as the size of the suggestions list grows, the role of the weights becomes less important. For a small suggestion list (1–3 suggestions), the weight of the recency dimension seems to play a more important role in obtaining a "hit", which is a clear indicator that call recency is more important than call frequency for determining the importance of a contact.

4.3.2. Experiment set B—scored performances

Our first experiment set showed that, for list sizes greater than 3, the weighting balance of the frequency and recency dimensions is practically immaterial, as the performance remains more or less constant. To investigate the quality of the suggestions (i.e., how close was the contact that was actually called to the top of the suggestion list, and thus likely to be seen sooner by the user), we performed the second set of experiments as previously described, for all suggestion list sizes apart from the one (as this is equivalent to the one suggestion hit-or-miss experiment reported earlier) (Figs. 5 and 6). In this case, we notice that, generally, the algorithm offers good placement of the actual correct predictions within the suggestion list (Fig. 5(a)), which is, on average, quite close to the top in each case. Again, we note that the recency dimension seems to offer better performance when weighed favourably over the frequency dimension. We note also (Fig. 5(b)) that users of Group 1

Please cite this article in press as: V. Stefanis, et al., Frequency and recency context for the management and retrieval of personal information on mobile devices, Pervasive and Mobile Computing (2013), http://dx.doi.org/10.1016/j.pmcj.2013.08.002

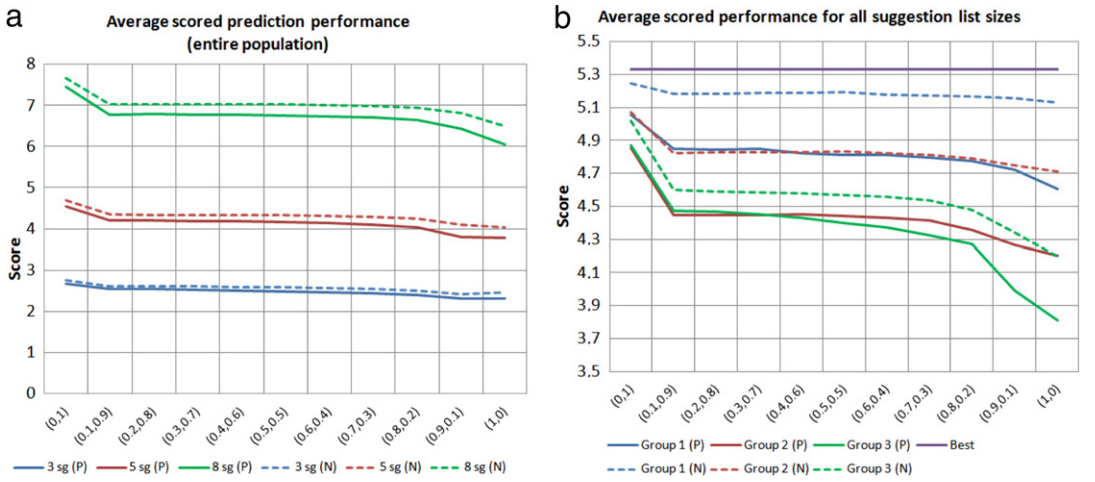

Fig. 5. (a) Average scored prediction performance for the entire population. (b) Average scored performance for all suggestion list sizes (xsg) (the purple line shows the theoretic optimal average score of 5.33), for the preliminary (P) and NOKIA (N) datasets, for all (w_f, w_r) weight combinations.

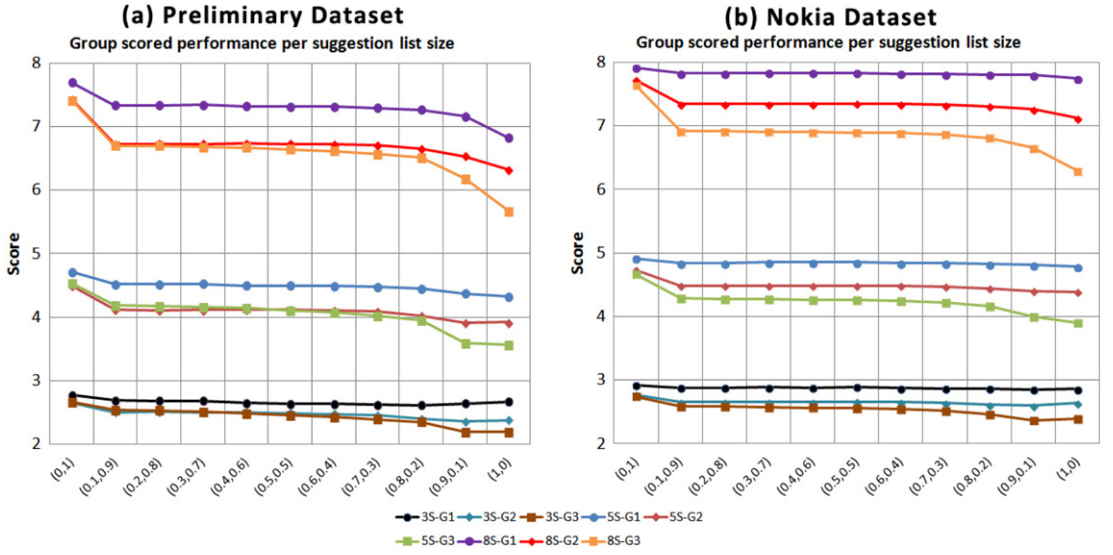

Fig. 6. Breakdown of each group's (Gx) scored performance for all suggestion list sizes (xS), for all (w_f, w_r) weight combinations. (a) Preliminary dataset. (b) Nokia dataset.

enjoy the best performance, which is followed by the performance experienced by Group 2 and Group 3, again confirming our earlier hypothesis.

4.3.3. Experiment set B—training window (NOKIA dataset)

Since data from our preliminary dataset were restricted to a relatively short period of time, the NOKIA dataset provided us with the opportunity to investigate the effect that different training window lengths would have on the prediction results. As we have already explained in a previous paragraph describing our experimental considerations, the length of the training window should be such that at the same time it represents accurately the communication pattern of the user and captures changes in this pattern.

As the prediction results for Group 3 (the most social group) were lower compared to those for the other two groups and Group 3 has by far the most members, we performed some more tests with changing training window lengths only for this group. We ran again our prediction algorithm for Group 3 with the same suggestion list (1, 3, 5, 8) and recency window (12 h) lengths, $w_f = 0.1$ and $w_r = 0.9$ (the best weight combination for Group 3), while the training window length was made a variable with the values of 5, 10, 30, or 90 days.

Fig. 7 shows how the prediction success rate changes for different training window lengths for each suggestion list size. We note that the variance in success rates is negligible for the different training window lengths. Fig. 8 shows how the average prediction score changes for different suggestion list sizes for each training window length. As expected, again,

Please cite this article in press as: V. Stefanis, et al., Frequency and recency context for the management and retrieval of personal information on mobile devices, Pervasive and Mobile Computing (2013), http://dx.doi.org/10.1016/j.pmcj.2013.08.002

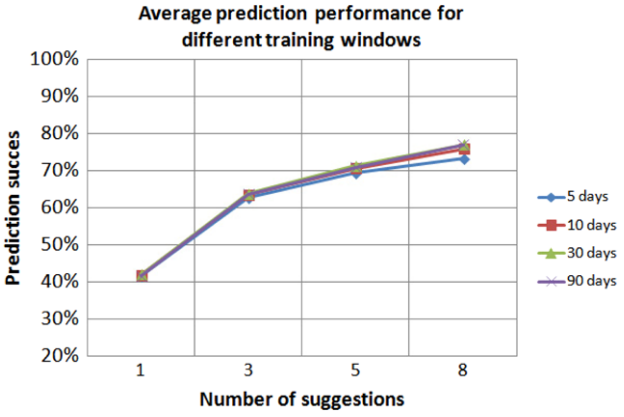

Fig. 7. Average prediction performance on the NOKIA dataset for different training windows.

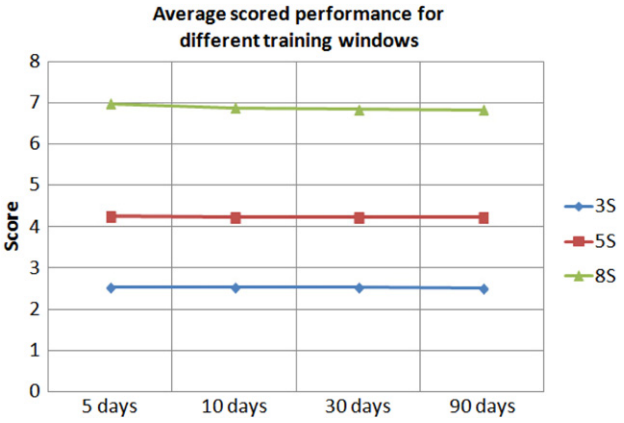

Fig. 8. Average scored performance on the NOKIA dataset for different training windows.

there are not significant differences regarding the average score for the different training window lengths. We thus see that communication behaviour does not change significantly through time for the NOKIA users, and that it can effectively be captured by a period of just five days. This is contradictory to our previous assumption that a training window would need to include a weekend period; however, the differences in performance can be attributed to the nature of the users in the dataset. By contrasting the average number of calls made over weekends to those made during the weekdays, we can observe that there is a difference of just 1.06 calls per day (4.65 cpd and 5.71 cpd, respectively). In any case, we show that the computational load can be safely decreased by choosing a smaller training window without detrimentally affecting performance, something that is important when considering applying these techniques to resource-constrained mobile devices.

5. Preliminary investigation of temporal context

Although the scope of this paper is to assess frequency and recency of use as context dimensions in a predictive model for contact retrieval, in this section we present a preliminary analysis on the effect of adding a third dimension in our model. Since frequency and recency of use describe behavioural context, we wanted to investigate the effect of a non-behavioural context dimension. In Section 3, we discussed the difficulties in extracting other contextual information (e.g., location, task) from our datasets. However, temporal context was available, and according to literature it can be used to detect patterns of movement [19] and semantic information about location [20]. More specifically, since people tend to visit places at specific times of day (e.g., home at night, office/work in the morning, restaurants at lunch time, etc.) [21], the significance of locations may vary with the time of day [22].

Based on these remarks, we decided to include the time of day (as an indicator of semantic location) as a third contextual dimension. In [1,15], temporal context is obtained either hourly or by segmenting the day into four non-equal parts; hence there does not seem to be a consensus on how to optimally represent this type of context. In our case, we split each day in three logical dayparts: workday (08–16), evening (16–24), and night (00–08). In order to compute the score of this dimension for a contact i at a datetime dt, we count all communications with this contact that were made during the training window

Please cite this article in press as: V. Stefanis, et al., Frequency and recency context for the management and retrieval of personal information on mobile devices, Pervasive and Mobile Computing (2013), http://dx.doi.org/10.1016/j.pmcj.2013.08.002

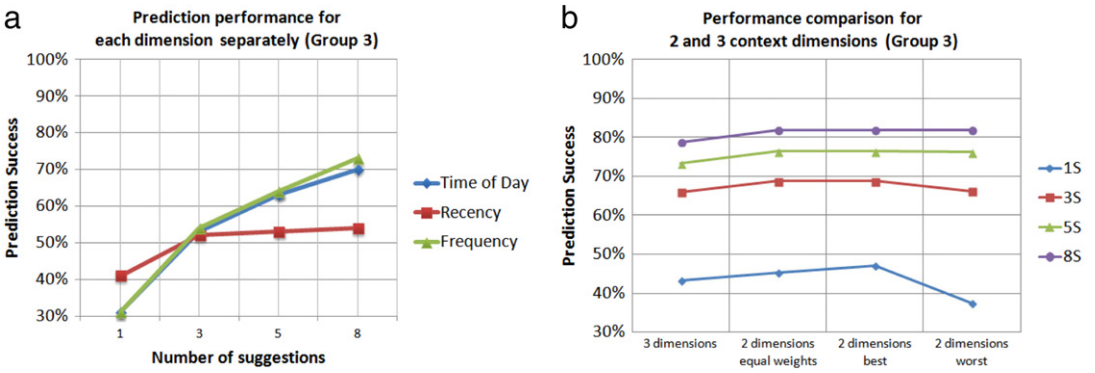

Fig. 9. (a) Prediction performance for each dimension separately. (b) Performance comparison for the three combined context dimensions (frequency, recency, time of day).

within the daypart that corresponds to *dt*, and we divide by the total count of communications within this daypart for the specified training window.

As we can see from Fig. 9(a), the dimension of time of day alone provides comparable performance results to the frequency dimension. Fig. 9(b) shows the algorithm's performance when adding the third dimension of time of day for Group 3, which has the largest margin for potential improvement. For these results, we tried several weight combinations, and found no significant differences in performance; hence we present the agnostic situation where we know nothing about the user and hence set all weights to an equal value. As we can see, adding a third dimension actually shows a slight decrease in average performance, despite our expectation that it would improve the prediction success rate; hence it seems that the inclusion of a new dimension in a predictive model requires a thorough investigation of its effect on users' calling behaviour.

6. Comparison with other algorithms

An indirect comparison of our results in predicting the next contact to be called can be contrasted against the findings of Lee et al. [1], Barzaiq and Loke [12], and Phithakkitnukoon et al. [14], particularly for predicting using five suggestions. Lee et al. [1] achieve performances greater than 75% for just one type of user (easily predictable ones) while the other two groups that emerge in their study do not exceed 50% and 30% average success, respectively. Barzaiq and Loke [12] achieve a 40% success on average for five suggestions after five weeks' worth of training and adapting their system. Finally, Phithakkitnukoon et al. [14] achieve a 70% average success rate for the five-suggestions prediction list.

The algorithms used in all cases are much more complex in nature than our own technique, which achieves an average success for the entire population of approximately 80%, while the performance even for Group 3 (who are the most social and thus unpredictable users) hovers around 70–73% (Fig. 10). Finally, the scored performance experiment set shows that the algorithm offers good ranking for the predictions within each suggestion list.

In order to perform a direct comparison between our model and the other algorithms, we decided to implement them and assess their success rate over the NOKIA dataset. We implemented the algorithms of [1,12], since for the one presented in [14] all the necessary details are not clearly described. As these algorithms provide five suggestions to the user, Fig. 11 shows a comparison with our algorithm's maximum, minimum, and average success rates for a suggestion list of five entries and the respective rates for the simulated frequently called contacts list ($w_f = 1$, $w_r = 0$) and call log list ($w_f = 0$, $w_r = 1$).

It is apparent that, apart from Group 1, where all approaches (except the call log list) work almost equally well, for Groups 2 and 3, that are more difficult to predict, our algorithm provides significantly better results. For these groups, the algorithms of Lee et al. [1] and Barzaiq and Loke [12] prove to be less efficient, even compared to traditional contact retrieval means.

7. Discussion and further work

In the previous sections, we have presented in detail the results of our experiments. A significant finding is that by combining the dimensions of frequency and recency we achieve better prediction results than by considering each dimension separately, which is in line with the findings of Barzaiq and Loke [12]. The performance for "least social" types exceeds 90%, showing that for such users further use of context is not required. For the "most social" users, significant room for improvement is displayed, which could be addressed by considering additional types of context such as location and time. However, as shown in Section 5, adding new dimensions requires extensive analysis in understanding how they affect retrieval tasks. Furthermore, as was expected, and as Phithakkitnukoon et al. [14] also note, the larger the suggestion list, the higher the prediction success rate is. However, this positive effect decreases as the suggestion list increases, and having in mind that small screens of mobile devices usually display only a few items of information, it seems that there is no point in providing more suggestions. Another interesting observation is that, as the size of the suggestion list increases, the role of the weights becomes less important.

Please cite this article in press as: V. Stefanis, et al., Frequency and recency context for the management and retrieval of personal information on mobile devices, Pervasive and Mobile Computing (2013), http://dx.doi.org/10.1016/j.pmcj.2013.08.002

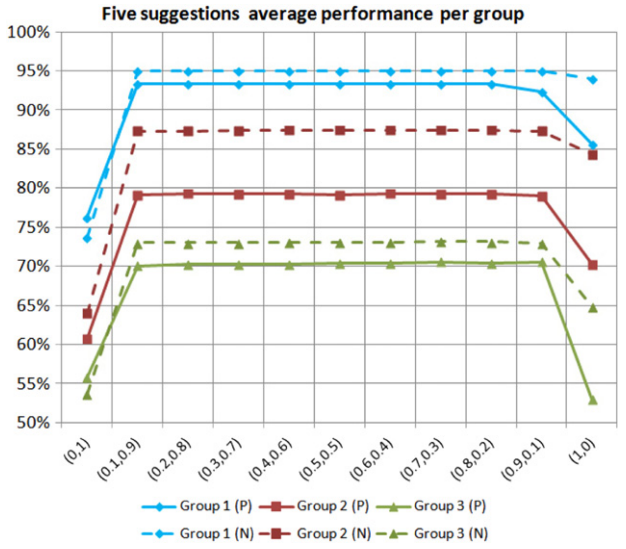

Fig. 10. Breakdown of each group's performance for a suggestion list size of 5, for the preliminary (P) and NOKIA (N) datasets, for all ($w_{\mathbf{f}}$, $w_{\mathbf{r}}$) weight combinations.

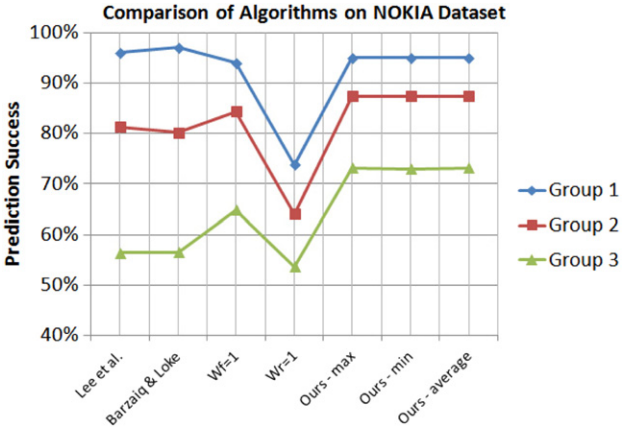

Fig. 11. Comparison of the results of Lee et al. [1], Barzaiq and Loke [12], frequently used contacts list, call log, and our algorithm (max, min and average performance) on the NOKIA dataset for five suggestions.

The observation of Lee et al. [1], that the existence of groups of users with different social communication behaviour influences the prediction performance, was also confirmed from our experiments (though we find three distinct groups instead of the two mentioned in that study). The variance of the results due to the different communication pattern of each group is a concrete indication that weights should not be static, but dynamic for each user. In addition to this, we believe that the weights should be dynamic even for the same user under different contexts.

Supporting personal information retrieval through adaptive UIs requires that "standard" behaviour is first understood, so that we can design an intervention that can have a predictable and desirable effect on it. The question thus arises of how we could utilize these findings to aid user behaviour, through designing interfaces informed by this knowledge. First, to our knowledge, there is no study about the effect of the availability of calling lists or frequently used number lists to the actual calling behaviour of users. Does this information affect the behaviour of users, i.e., does it exert influence on the maintenance of strong social links with other users by making access to calling them easier? Would a system that always predicts the right person to call next prevent users from making contact with other less important users? Perhaps the introduction of "false positives", particularly for users who are not very social, could encourage them to communicate more often with a wider variety of contacts. Perhaps, also, a user interface should not only help users find the next contact to call quickly, but also remind them of contacts that used to be important but have not been contacted for some time. And then, is the concept of "calling" the optimal means of contacting someone? Would an interface that suggested not only the person but also the mode of contact to something "more appropriate" than just text or talk (e.g., Facebook message) be desirable, or help increase participation in social networks rather than one-to-one communication? In this sense, the discovery of distinct groups of users in terms of their communication behaviour is fortunate, as, for example, the most social group's behaviour

Please cite this article in press as: V. Stefanis, et al., Frequency and recency context for the management and retrieval of personal information on mobile devices, Pervasive and Mobile Computing (2013), http://dx.doi.org/10.1016/j.pmcj.2013.08.002

could be used as a baseline, and further research could be undertaken on how close a system can bring to this behaviour users from other groups. We thus see our work of understanding and predicting communication patterns as an essential first step into designing persuasive user interfaces for users.

Although we start with the contact list and the task of facilitating contacts retrieval from it as a problem domain, we believe that this approach could extend and apply to other information management problems that involve context as well. As a first step, we intend to introduce more contextual dimensions to our algorithm, since in this work we focus on behavioural factors such as frequency and recency of communication. The same analysis could also extend to include SMS communications or other forms of social interactions (e.g., social networking).

To conclude, the previously discussed observation about the dynamic nature of dimension weights for different users and different contexts is an indication that a more generic approach that would not involve manual adjustment of weights is needed. In previous work [23], we proposed the application of a dimensionality reduction technique to context augmented personal information items, such as entries in a contact list, in order to extract a small number of features that could accurately represent the original items and their relationships. Our future experiments include the application of this technique to the available datasets for the problem of predicting the next contact to be called. We hope that this work could provide us with valuable insight and understanding of mobile users' behaviour, allowing us to proceed with the design and experimentation of novel persuasive mobile user interfaces that help users manage their personal information more effectively. We aim to test these longitudinally in the field as a replacement to traditional contact list access methods under real-life conditions.

Acknowledgements

We would like to thank J. Laurila, D. Gattica-Perez, and J. Blöm from NOKIA Research Centre Lausanne for providing access to the Lausanne Mobile Data Collection Campaign dataset.

References

[1] S. Lee, J. Seo, G. Lee, An adaptive speed-call list algorithm and its evaluation with ESM, in: Proceedings of ACM CHI 2010, ACM, 2010, p. 2019.
[2] O. Bergman, A. Komninos, D. Liarokapis, J. Clarke, You never call: demoting unused contacts on mobile phones using DMTR, Personal and Ubiquitous Computing 16 (6) (2012) 757–766.
[3] A. Komninos, A. Plessas, V. Stefanis, J. Garofalakis, Context dimensionality reduction for mobile personal information access, in: Proceedings of KDIR 2011, SciTePress, 2011, pp. 493–498.
[4] A. Toninelli, D. Khushraj, O. Lassila, R. Montanari, Towards socially aware mobile phones, in: Proceedings of SDoW 2008, Karlsruhe, Germany, 2008.
[5] P. Gilbert, E. Cuervo, L. Cox, Experimenting in mobile social contexts using JellyNets, in: Proceedings of the 10th Workshop on Mobile Computing Systems and Applications, HotMobile 2009, ACM, 2009. Article No. 16.
[6] R. Boardman, M.A. Sasse, Stuff goes into the computer and doesn't come out: a cross-tool study of personal information management, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI'04, ACM, New York, 2004.
[7] L. Zhou, A. Mohammed, D. Zhang, Mobile personal information management agent: supporting natural language interface and application integration, Information Processing and Management 48 (1) (2012) 23–31.
[8] R. Myers, E. Zapata, G. Singh, Linking information for mobile use, in: Proc. of Mobility'07, ACM, New York, 2007, pp. 607–613.
[9] K. Church, B. Smith, Understanding the intent behind mobile information needs, in: Proc. of IUI'09, ACM, New York, 2009, pp. 247–256.
[10] O. Bergman, R. Beyth-Marom, R. Nachmias, The user-subjective approach to personal information management systems design: evidence and implementations, Journal of the American Society for Information Science and Technology 59 (2) (2008) 235–246.
[11] A. Komninos, D. Liarokapis, The use of mobile contact list applications and a context-oriented framework to support their design, in: Proc. of MobileHCI'09, ACM, New York, 2009. Article No. 79.
[12] O. Barzaiq, S. Loke, Adapting the mobile phone for task efficiency: the case of predicting outgoing calls using frequency and regularity of historical calls, Personal and Ubiquitous Computing 15 (8) (2011) 857–870.
[13] S. Phithakkitnukoon, R. Dantu, Adequacy of data for characterizing caller behavior, in: Proceedings of the 2nd ACM SIGKDD International Workshop on Social Network Mining and Analysis, SNA-KDD 2008, ACM, 2008.
[14] S. Phithakkitnukoon, R. Dantu, R. Claxton, N. Eagle, Behavior-based adaptive call predictor, ACM Transactions on Autonomous and Adaptive Systems (TAAS) 6 (3) (2011) Article No. 21.
[15] S. Phithakkitnukoon, R. Dantu, Towards ubiquitous computing with call prediction, SIGMOBILE Mobile Computing and Communications Review 15 (1) (2011) 52–64.
[16] J.K. Laurila, D. Gatica-Perez, I. Aad, T.-M.-T.D. Jan Blom, O. Bornet, O. Dousse, J. Eberle, M. Miettinen, The mobile data challenge: big data for mobile computing research, in: Mobile Data Challenge by NOKIA Workshop, in Conjunction with Int. Conf. on Pervasive Computing, 2012.
[17] V. Stefanis, A. Plessas, A. Komninos, J. Garofalakis, Patterns of usage and context in interaction with communication support applications in mobile devices, in: Proceedings of ACM MobileHCI'12, ACM, 2012, pp. 25–34.
[18] S. Phithakkitnukoon, R. Dantu, Mobile social closeness and communication patterns, in: IEEE Conference on Consumer Communications & Networking Conference (CCNC 2010) Special Session on Social Networking, SocNets, 2010.
[19] N. Bicocci, G. Castelli, M. Mamei, A. Rosi, F. Zambonelli, Supporting location-aware services from mobile users with the whereabouts diary, in: Proc. of MOBILEWARE'08, Brussels, Belgium, 2007.
[20] N. Eagle, A. Clauset, J.A. Quinn, Location segmentation, inference and prediction for anticipatory computing, in: AAAI Spring Symposium on Technosocial Predictive Analytics, 2009.
[21] R. Montoliu, A. Martinez-Uso, J. Martinez-Sotoca, Semantic place prediction by combining smart binary classifiers, in: Mobile Data Challenge by NOKIA Workshop, in Conjunction with Int. Conf. on Pervasive Computing, 2012.
[22] X. Cao, G. Cong, C. Jensen, Mining significant semantic locations from GPS data, Proceedings of the VLDB Endowment 3 (1–2) (2010).
[23] A. Komninos, A. Plessas, V. Stefanis, J. Garofalakis, Application of dimensionality reduction techniques for mobile social context, in: Proceedings of the 13th ACM International Conference on Ubiquitous Computing, Ubicomp'11, ACM, 2011, pp. 583–584.

Please cite this article in press as: V. Stefanis, et al., Frequency and recency context for the management and retrieval of personal information on mobile devices, Pervasive and Mobile Computing (2013), http://dx.doi.org/10.1016/j.pmcj.2013.08.002