Utilizing Rich Bluetooth Environments for Identity Prediction and Exploring Social Networks as Techniques for Ubiquitous Computing

Jukka Perkiö and Ville Tuulos
Complex Systems Computation Group
Helsinki Institute for Information Technology
P.O. Box 68 FIN 00014
Finland

Marion Hermersdorf, Heli Nyholm,
Jukka Salminen and Henry Tirri
Nokia Research Center
P.O. Box 407 FIN-00045
Finland

Abstract

Personal identification and using that information is in the heart of many ubiquitous systems. We present two complementary techniques, namely personal identification without directly observing the subject, and using that information for understanding the social relations between the subjects.

We show that with certain presumptions it is possible to predict one's identity with reasonable certainty only by observing one's Bluetooth neighborhood without the need to directly observe the subject. We also show how this information can be used for exploring the social relations between the subjects.

1. Introduction

Ubiquitous systems require many complementary techniques to be used in seamless manner. The very aim of such systems is to be all pervasive but not noticeable. It is – among other things – these two almost contradicting requirements, that dictate the the design of these systems. We concentrate on one very basic requirement of many ubiquitous systems, namely personal identification and the use of that information.

The motivation for our research is following: One can imagine a situation where we have an ubiquitous system, that is not able to directly observe the presence of some individual in a some space, but we know with high certainty, that somebody – somebody, that we can not observe directly – is present in that space. However our system may be able to observe the space itself, i.e. it is able to observe the space in vicinity of that individual. We are interested in predicting person's identity based on the observed space in her vicinity. After we know the identity we are also interested in the relations between different identities.

We view the mobile devices as wearable sensors and the detected Bluetooth neighborhood as the observable space. The huge penetration of Bluetooth enabled devices makes our choice appropriate, since the sales of mobile phones in 2005 was 825 million [1] and the sales of Bluetooth enabled devices in general is predicted to be over 500 million in 2006 [2].

We argue that given some presumptions, it is possible to predict somebody's identity just by knowing her Bluetooth neighborhood even if this person is not carrying any Bluetooth enabled device herself. The presumptions are:

1. The space is crowded by other people who have Bluetooth enabled devices with them.
2. The space contains several stationary Bluetooth enabled devices.

It is enough if either one of the above is true, but the prediction accuracy may vary. The rationale is that in today's environment one defines her Bluetooth neighborhood through three different ways:

1. Through the personal mobile devices.
2. Through the stationary devices, personal or shared.
3. Through the social relations, i.e. people with whom one spends time.

The case 1 above is trivial as we can simply detect the presence of personal mobile devices, but cases 2 and 3 are more interesting. In those cases we assume that one has a specific Bluetooth fingerprint, that is defined by the combination of turned on devices in that specific space and by the devices of people with whom one spends time. It is obvious that the robustness of this method may vary a lot depending how predictable people's behavior is.

Proceedings of the 2006 IEEE/WIC/ACM International Conference
on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06)
0-7695-2747-7/06 $20.00 © 2006 IEEE

Recently there has been interesting research in using Bluetooth scanning as means for social proximity sensing [11] and understanding complex social systems [7, 8]. The latter is known as *Reality Mining*. The main emphasis in that research has been on understanding the social relations and group dynamics between and with the participants.

Our research in the first part concentrates on identifying the identity of an unobserved individual, and as such, is different from the research mentioned in previous paragraph even though detecting the Bluetooth devices as a methodology is similar. The second part of our research concerns detecting social relations between different identities, and as such, is quite closely related to the aforementioned research. Our methodology differs though as we use a statistical component model to model the relations.

The very important question of privacy is left out of the scope of this paper, but we acknowledge its importance, and it is one of the questions that have to be further researched.

The rest of this paper is organized as follows: In Section 2 we explain our data set and how it was obtained, in Section 3 we explain our identification models and evaluate them, in Section 4 we explain our component model for social relations and in Section 5 we present our conclusions.

2 Data

The Bluetooth class 2 radio has a maximum range of 10 meters and therefore provides information of the Bluetooth devices in the proximity of the scanning device. An application was running on the mobile phone, which scanned approximately every thirty seconds for other Bluetooth devices in the proximity. The scanning results were sent via GPRS to a back-end server.

Fourteen mobile phones with this application were distributed to selected users working in an office building. The users were encouraged to carry this Bluetooth scanning phone, additional to their personal phone, with them while being in the building. We also asked the users to leave the phone over night in their office for recharging.

We placed fifteen passive Bluetooth beacons in the building. These beacons can be used for positioning even though in this paper we do not do positioning. We use this information in the explorative part though as we investigate how our component models relate to different locations. The beacons were standard off-the-shelve USB Bluetooth dongles for PCs. Once started up, by connecting to a PC, the Bluetooth adapters were disconnected from the PC without interrupting the power supply (use of a battery pack), and placed at defined locations.

The Bluetooth scanning data of users were collected on the back-end server for approximately two weeks. It was possible for us to monitor the data reception to ensure that all phones and beacons were operating properly. The data set consists of 2 867 167 rows of

$$(PhoneID, timestamp, MAC)$$

tuples. Some of detected MACs correspond to known locations (beacon MACs).

Description	Quantity
Total number of BT scans	73 588
Total number of BT devices detected	854 926
Number of individual BT addresses detected	1 299
Average number of BT devices detected on one scan	11.6
Max. number of BT devices detected on one scan	52

Table 1. Bluetooth Data Summary

Even though our test site might not correspond to a typical office environment of today, the trend is obvious: Urban areas are becoming increasingly covered by various uncontrolled short-range radios, such as Bluetooth, WiFi and ZigBee. Table 1 summarizes our data set, collected within around two weeks by fourteen people in one building. One should note especially the remarkable number of unique BT addresses (1299) and the high average number of detected devices (11.6). We see that rich environments like this provide fruitful ground for data-intensive tasks, such as probabilistic modeling.

There are some anomalies in the data, that resulted mainly from the battery running out of charge or a test subject forgetting to carry the phone. Mostly these anomalies appear in the form of missing data and different test subjects producing different amounts of data.

3 Identification based on one's Bluetooth neighborhood

We have test subjects, that can be thought to form a set of predefined classes. In our setting we consider these classes mutually exclusive, i.e. we are interested in predicting one's identity, and not a group to which one belongs. We have also a set of observations of the classes in the form of one's recorded Bluetooth neighborhood. These observations are time-dependent, i.e. form a time series of vectors containing Bluetooth MAC addresses. Given the classes, we are interested in assigning each observation to one of the classes, i.e. we want to predict one's identity based on her Bluetooth neighborhood. This problem is a typical classification problem.

We use statistical classifiers, namely Bayesian ones. From our training data we are able to estimate a prior probability distribution $p(O \mid C_i)$ for observation O given the class C_i. In our case the observation O is a vector containing MAC addresses and the class C_i is the test subject who collected the data.

The Bayes rule states that the posterior probability for test subject i being the one who made the observation O is

$$p(C_i \mid O) = \frac{p(O \mid C_i)\, p(C_i)}{\sum_{j \in |\mathcal{C}|} p(O \mid C_j)\, p(C_j)}, \qquad (1)$$

where C_j is any class of all classes $\mathcal{C}$.

3.1 Identification models

We present two models with some simplifications based on the Eq. 1. For the first model we assume that in each observation vector O the individual MAC addresses o are mutually independent. This is not very realistic assumption, but in practice these *Naive Bayes* models perform well. We also do not consider time in this model, i.e. our model is static.

The static model is as below

$$p(C_i \mid O) = \frac{1}{Z} p(C_i) \prod_{o \in O} p(o \mid C_i), \qquad (2)$$

where O is the observation, o is a single item in the observation, i.e. a single MAC address, C_i is the class, i.e. the test subject whose identity we are predicting and Z is a scaling factor.

Our second model is also a Naive Bayes model, but here we also consider the time, hence producing a time-dependent dynamic model. The model above (Eq. 2) is modified only so that we learn the probabilities time dependently from our training data, and we get

$$p(C_i \mid O, t) = \frac{1}{Z} p(C_i \mid t) \prod_{o \in O} p(o \mid C_i, t), \qquad (3)$$

where the additional variable t is the time of the day and rest of the variables are the same as in Eq. 2.

With these models we use MAP (maximum a posteriori) decision rule, which means that the observation is assigned to the class, that gets the highest posterior probability.

3.2 Evaluation

We evaluated both of the models using our data set that is explained in Section 2. Due to previously mentioned anomalies, we did not use the whole data set, but only seven test subjects were included. We also restricted the evaluation to working days.

The models were trained using a distinct training set, and test subjects' possible personal phones and other Bluetooth devices were separated from the data set so that we were able to evaluate the models reliably. We used five-fold stratified cross-validation. Table 2 shows the results for both of the models. It can be seen that the accuracy is quite high. The lowest accuracy with the static model is 84.9% and with the dynamic model 94.8%. This tells us, that our office environment is quite predictable and people seem to have reasonably established daily routines.

Class	Static model accuracy	Dynamic model accuracy
1	99.0	99.2
2	94.3	98.9
3	100.0	100.0
4	97.7	98.6
5	94.5	97.0
6	100.0	100.0
7	84.9	94.8

Table 2. Classification accuracy for both the models.

It can be also seen, that – as one would assume – using a time-dependent dynamic model has an positive effect on the results. That can be seen with all the test subject, but that is especially clear with the test subject 7. The classification accuracy increased 10%.

Another interesting point is shown in Figures 1 and 2. The classification accuracy during a normal working day is not constant, since people have their own daily routines in the work place. During the office hours, when there are more people around, it is hardest to do the predictions. People are interacting with other people, having meetings etc. This results higher overlap between the Bluetooth neighborhoods of different people, thus the predictions are harder to do. These figures also show very clearly the benefits of time-dependent dynamic model. The overall classification accuracy is higher and the variation is also much lower.

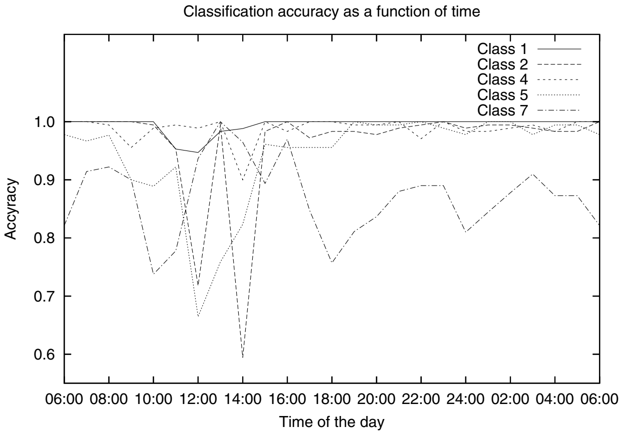

Figure 1. Classification accuracy for the static model over 24 hours during a normal working day. Classes 3 and 6 are omitted, since the accuracy for them is 100%.

One has to remember that in our test setting the phones

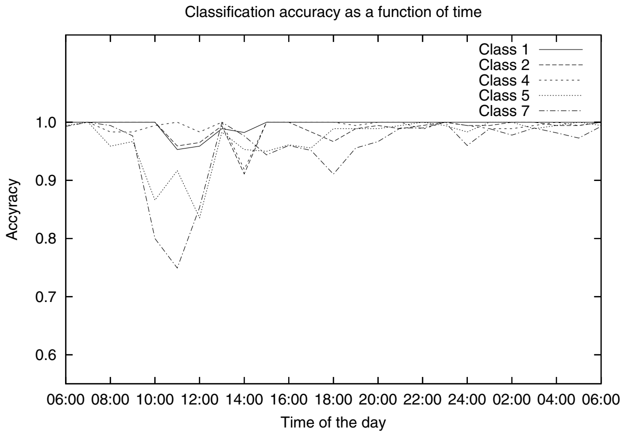

Figure 2. Classification accuracy for the dynamic model over 24 hours during a normal working day. Classes 3 and 6 are omitted, since the accuracy for them is 100%.

were carried only while one was in the office. Our work environment is such that most people work during the normal office hours and occasionally also during the night. There are also some people that have little less conventional working hours staying in the office during the nights. In table 3 we show the classification accuracy during the working hours between 8:00 – 18:00.

Class	Static model accuracy	Dynamic model accuracy
1	98.7	98.8
2	92.1	98.1
3	100.0	100.0
4	98.1	98.6
5	88.1	93.5
6	100.0	100.0
7	89.5	92.1

Table 3. Classification accuracy for both the models during working hours between 8:00 and 18:00.

4 Component models of social relations

After we are able to identify persons using the models presented in the previous section, we can go on, and start exploring more about their social relations. We use component models to model the social relations between people. The component model that we are interested in is the MPCA [5] model.

When using component models, one assumes that there are some latent components, that are more characteristic to the data than the observed data itself. In social relations there are many very fine grained variables, that may be hard to detect. Our aim is not to do deep psychological or sociological analysis, but use simple means to collect rather coarse data the hypothesis being: The more one spends time with somebody else the more important the relationship is.

4.1 MPCA component model

Suppose that we have n l-dimensional vectors of discrete values each value being a count of some feature. Our data $\mathbf{D}$ is then $n \times l$ matrix. We want to reduce the dimensionality of the data to $k \ll l$. The reason for this is two-fold. First the lower dimensional data is computationally cheaper, and second we are interested in the latent structures of the data in the form of latent components. This can be formulated as

$$\mathbf{D} \simeq \mathbf{m} \times \Omega, \qquad (4)$$

where Ω is the component loading matrix, and $\mathbf{m}$ contains the latent components of the data. Thus $\mathbf{m}$ is $n \times l$ and Ω is $k \times l$.

There are many different methods, that could be used. Among those are PCA (principal component analysis), ICA (independent component analysis) [9]. NMF (non-negative matrix factorization) [10] and LDA (latent dirichlet allocation) [3] could also be used.

We use MPCA (multinomial principal component analysis). MPCA is quite recent statistical method, that has been used e.g. in analyzing large corpora of text documents, see e.g. [6]. It suits for all kind discrete data though. MPCA produces exactly that kind of factorization, that is presented in Eq. 4. For the estimation of MPCA model one should consult e.g. [4] or [5]. We use methods presented in [5].

4.2 Experiments

We built several models with different number of components to investigate our data set. In these experiments we were interested in the social relations between 21 individuals, that include the seven individuals, that we used in our classification experiments in Section 3.2, and we wanted also investigate how different known locations relate to the model.

Table 4 shows how nine most important locations are distributed in the components. In these experiments we considered 13 locations. Table 5 shows the identifiers of the most important persons within each component. Even though we limited our analysis to 21 persons in this experiment, the models were built from a much larger subset that contained 759 distinct Bluetooth MAC addresses. From the tables 4 and 5 above we can see many interesting details. First we can assume that the social proximity between persons within the same component is higher than between persons in different components. That is because we can think of

Comp.	Most important location	Second most important location
0	cafeteria	a3 coffee
1	1st sofas	cafeteria
2	main entrance	a5 lab
3	1st tables	3rd sofas
4	b5 lab	7th sofas

Table 4. Five component model and the two most important locations within each component.

Comp.	Number of persons with strong weight in component	IDs of persons with strong weight in component
0	6	0, 4, 5, 7, 8, 10
1	2	7, 14
2	4	6, 7, 15, 16
3	4	13, 14, 16, 20
4	4	7, 9, 11, 18

Table 5. Five component model and the most important persons within each component.

our experiment as clustering and remembering our hypothesis stated in Section 4.1. We can also deduce something about the relations between different components solely on the basis of the persons, that have strong weight in more than one component. Person with an identifier seven is present in four components, whereas twelve persons are present only in one component, and six persons do not have strong presence in any component. The three-way relationship between locations, components and persons is interesting. Adding the number of components adds the resolution of the model up to a certain point. In our data set reasonable number of components seem to be 5 – 10.

5 Conclusions

We have presented two complementary techniques, that are interesting per se, and can be used effectively together in certain ubiquitous systems.

The first model is for identifying one's identity, when we have no direct observations of her, but we know that the person is in some space, and we can observe the Bluetooth neighborhood of that space. Here we have the assumption, that the Bluetooth neighborhood is dependent on the person's presence in the form of turned on devices, and more importantly the company of other people that may carry Bluetooth enabled devices.

We showed that we can achieve high prediction accuracy in our office environment. We presented both static and time-dependent dynamic statistical model for this task. We showed that the time-dependent dynamic model improves the performance considerably. We investigated the prediction accuracy as a function of time, and showed that the predictions are harder to do during the office hours, which is natural.

The second model continues from the point where the first model left. Once we know the identities of subjects, we can explore their social relations. Our experiments support that assumption. We can identify social groups and hubs through which those group connect to each other. We can also estimate how different groups and persons relate to different locations.

6 Acknowledgments

This work was supported in part by the Academy of Finland, under the project Prose, and by the National Technology Agency, under the project Search-Ina-Box. This work was also supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views.

References

[1] http://www.idc.com/getdoc.jsp?containerid=prus20056906. *IDC Press release, 26 January 2006, verified 29 September 2006.*

[2] *Bluetooth: The Global Outlook.* ABI Research, 2006.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[4] W. Buntine and A. Jakulin. Applying discrete pca in data analysis. In *20th Conference on Uncertainty in Artificial Intelligence (UAI'04)*, pages 59–66. AUAI Press, 2004.

[5] W. Buntine. and A. Jakulin. Discrete components analysis. In *Subspace, Latent Structure and Feature Selection Techniques*. 2006.

[6] W. Buntine, J. Löfström, J. Perkiö, S. Perttu, V. Poroshin, T. Silander, H. Tirri, A. Tuominen, and V. Tuulos. A scalable topic-based open source search engine. In *IEEE/WIC/ACM Conference on Web Intelligence (WI 2004)*, pages 228–234. IEEE Computer society, 2004.

[7] N. Eagle. *Machine Perception and Learning of Complex Social Systems*. MIT, 2005.

[8] N. Eagle and A. S. Pentland. Reality mining: sensing complex social systems. *Personal Ubiquitous Comput.*, 10(4):255–268, 2006.

[9] A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9:1483–1492, 1997.

[10] D. D. Lee and S. S. Seung. Learning the parts of objects by non-negative matrix factorizat ion. *Nature*, 401:788–91, October 1999.

[11] T. Nicolai, N. Behrens, and E. Yoneki. Wireless rope: Experiment in social proximity sensing with bluetooth. In *Fourth Annual IEEE International Conference on Pervasive Computing*, 2006.