BLTJ132_20301.qxd 6/16/08 10:51 PM Page 31

♦ Intuitive Network Applications: Learning User Context and Behavior

Nilton Bila, Jin Cao, Robert Dinoff, Tin Kam Ho, Bharat Kumar, Daniel Lieuwen, and Paulo Santos

This paper presents some details on the Intuitive Network Applications (INA) project, which aims to monitor and analyze a user's behavior in a semiautomated fashion and utilize the learned information to provide personalized services to the user. An aspect of this framework is how user input can be used to direct the learning process. We discuss some scenarios that illustrate the use of the INA framework in the context of advertising and show how a user's learned habits and intentions can be used to provide targeted and timely advertisements to the user. We also provide some preliminary results on applying some of the learning techniques in a pilot study conducted among a group of volunteers whose activities and movements were monitored over time. © 2008 Alcatel-Lucent.

Introduction

Driven by a need to provide ubiquitous communication and data services to users, the telecom networks are evolving to an all-Internet Protocol (IP) core, enabling convergence across the wireless telecom, wireline telecom, and the Internet domains. This enables subscriber profile knowledge (e.g., presence, location, address books, and preferences) to be shared across these domains, thus enabling rich converged services that can take advantage of the full span of subscriber information available.

In some of our past work, we had assumed that the user would provision preferences (targeted toward a single application), which along with other subscriber data could then be interpreted at run time by a high-speed Alcatel-Lucent Vortex® [11] rules engine to personalize the application behavior. Though useful, this puts an undue burden on the user; the user needs to provision his preferences explicitly and keep them up-to-date as his habits and intentions change. In our current research, we are working toward an initiative called Intuitive Network Applications (INA), which combines human factors and automated learning techniques for acquiring user profile preferences, which can then be shared across multiple applications.

There are various dimensions involved in this approach: 1) machine learning techniques to gather user habits and preferences in an automated fashion; 2) human factors-directed approaches to explicitly gather additional information from the user, while minimizing the (perceived) inconvenience to him; 3) providing the user with an easy mechanism to view and control the acquired information (whether learned or explicitly gathered); and 4) ensuring that the overall information monitoring and analysis are done in a robust, scalable fashion, and the acquired information can be used at run time in a near-real time fashion. In this paper, we focus primarily on an overview of the INA architecture and present some details on machine learning directed at an advertising

Bell Labs Technical Journal 13(2), 31–48 (2008) © 2008 Alcatel-Lucent. Published by Wiley Periodicals, Inc. Published online in Wiley InterScience (www.interscience.wiley.com) • DOI: 10.1002/bltj.20301

Alcatel·Lucent

BLTJ132_20301.qxd 6/16/08 10:51 PM Page 32

Panel 1. Abbreviations, Acronyms, and Terms

CALO—Cognitive Assistant that Learns and Organizes
GPS—Global Positioning System
GSM—Global System for Mobile Communications
GUI—Graphical user interface
HTML—Hypertext Markup Language
ID—Identifier
IM—Instant messaging
INA—Intuitive network applications
IP—Internet Protocol
IPTV—Internet Protocol television
PLIANT—Preference Learning through Interactive Advisable Nonintrusive Training
SMS—Short message service
XML—Extensible Markup Language
XSLT—Extensible stylesheet language transformation

scenario. We assume that the user has provided the necessary permissions for the network to monitor his information (or a subset thereof).

The INA project aims at utilizing a holistic view of the user, irrespective of the devices he employs or the networks he accesses. The information available about a user via a given device/network can be collected, analyzed, and then utilized to provide a value-added service on a different device/network. Thus INA is well positioned to take advantage of the converged network, and to provide personalization capabilities that span across the various parts of the converged network. As an example, on the basis of the genre of music a user listens to via an (audio-only) Internet Protocol television (IPTV) channel, similar kinds of ringtones and other content can be promoted via his mobile device.

A previous paper on INA [7] had discussed some preliminary thoughts on the use of INA in lifestyle assisting services, targeted toward the interactions of a user within his social network(s). In this paper, we show how INA can be leveraged by an advertising solution. We show how we can analyze a user's movement over time to build up knowledge of the various locations he visits and utilize that—along with additional information acquired explicitly from the user—to offer targeted advertisements to the user. We do not focus on the advertisement matching problem itself, but on how a user profile suitable for advertising can be built up.

As mentioned earlier, a key aspect of INA is the notion of user interaction for building up parts of the user profile. One incentive that can be offered to users is free or discounted service, either for answering questions posed to them or for viewing ads. A user might be more amenable to answering questions truthfully if he feels that the ads he gets would be more relevant/useful. Note that the notion of ad-supported services is not new and has been in the Web domain for quite some time. It has gradually gained the attention of telecom operators as well, who view it as a complementary revenue stream since traditional voice services are generating less and less revenue. Recently, a number of mobile operators have begun investigating a complete switch to ad-supported subscriptions. For example, Blyk [5] provides a certain number of voice minutes and text messages free in exchange for viewing ads, though users have the option to purchase more minutes if needed. Virgin Mobile's* Sugar Mama [20] offers one free minute of voice service for every minute that the user spends in watching ads or answering survey questions. Other scenarios include displaying ads based on user accessed/purchased content or user-initiated searches. The INA framework can be applied to all such scenarios and provides a critical piece of the overall infrastructure required.

This paper is organized as follows. We first present a motivating example that illustrates some aspects of acquiring user profiles in an advertising context. We also present some results from a small survey done on privacy concerns in tracking a user's habits and preferences, which shows that users may not mind allowing the network to track such information, as long as there is some value to them. We then present details on a pilot study done with some Bell Labs staff members where the participants' movements and habits were tracked and analyzed and use that to

BLTJ132_20301.qxd 6/16/08 10:51 PM Page 33

illustrate the issues involved in automatically learning users' habits. Next, we present the overall framework that we are building to gather a user's profile, focused toward the advertising space. We end with discussions on related work and some conclusions.

Motivating Example

We start by describing the types of information we are interested in gathering in an advertising context. We focus on acquiring information about a user's movement patterns—the various locations he frequents, the types of locations (e.g., shopping mall), and the typical activities he performs there (e.g., eating at a restaurant). **Figure 1** provides a relevant subset of the user profile schema dealing with location information. The full profile also contains other aspects of behavior/preferences (determined partly via explicit querying to the user), and a summary of the user's social networks (determined by examining whom he communicates with). The schema for that part of the user profile is not presented in this paper.

The changes in a user's location over time are tracked via a small client application on his mobile device and sent to a logging server every 15 minutes. These logs are then periodically (e.g., every day) analyzed to determine the regions that the user frequents during that period. This logging and analysis are repeated to build up a more global picture of a user's movement pattern, such as the frequency with which he visits a particular region, the day(s) of the week, start times, and durations—along with the standard deviation of both—during which he is in that region.

We did not use the Global Positioning System (GPS) capabilities on the mobile device; rather we utilized only the knowledge of the cell IDs that the device is connected to at any given point in time. There are a number of reasons for not using GPS, including lack of indoor coverage, high battery drain, and the fact that GPS-enabled devices are not yet prevalent. The use of cell IDs does however cause the frequented region determination to be coarse-grained. Since cell towers in suburban/rural environments might be spread out by a distance of several miles, this type of region might encompass more than one location that the user visits on a regular basis. An example is that the user visits a shopping mall which is very close to his home. Hence, the user is also explicitly queried over time to determine whether there is more than one frequented location within a given frequented region.

Each location can be tagged with one of a set of prespecified location types. A subset of the location types is given in Figure 1. The user can also be prompted to find out what activity he performs within a given location, or in general. The user chooses an activity from a set of prespecified activities. This set of location and activity types is maintained as a simple ontology, which also captures the potential relationship between a given location type and the set of activities possible in that location. The ontology can be expanded as necessary. This can then be used to filter out irrelevant activities when posing questions to the user. The location activity types are organized in the form of a hierarchy. This allows users to provide information of coarser granularity if they are not comfortable in giving more detailed answers, since even data of coarse granularity might be still useful for advertising purposes.

The previous scenario describes a small set of the full range of possibilities. There are various other types of user activities that can also be logged and automatically analyzed. We can analyze a user's phone/short message service (SMS)/instant messaging (IM) records to determine whom he frequently communicates with and generate a profile of the social networks in which he participates. The contents of SMS/IM messages can also be analyzed (e.g., for keywords) to determine the type of information he might be interested in. In addition, the user can be explicitly queried for other information that cannot be automatically determined, or for validating what has already been learned. In general, a combination of human-factors-driven user interaction is coupled with automatic analysis of logged user activity to acquire the full user profile.

Once the user profile has been acquired, an advertising application can query the current state of the user profile on demand, in order to perform ad matching, or to determine the potential target audience for a given ad campaign. In the future, we can also envision a trigger-based mechanism to send relevant ads. For example, if the framework knows that the user frequents a particular location for lunch

BLTJ132_20301.qxd 6/16/08 10:51 PM Page 34

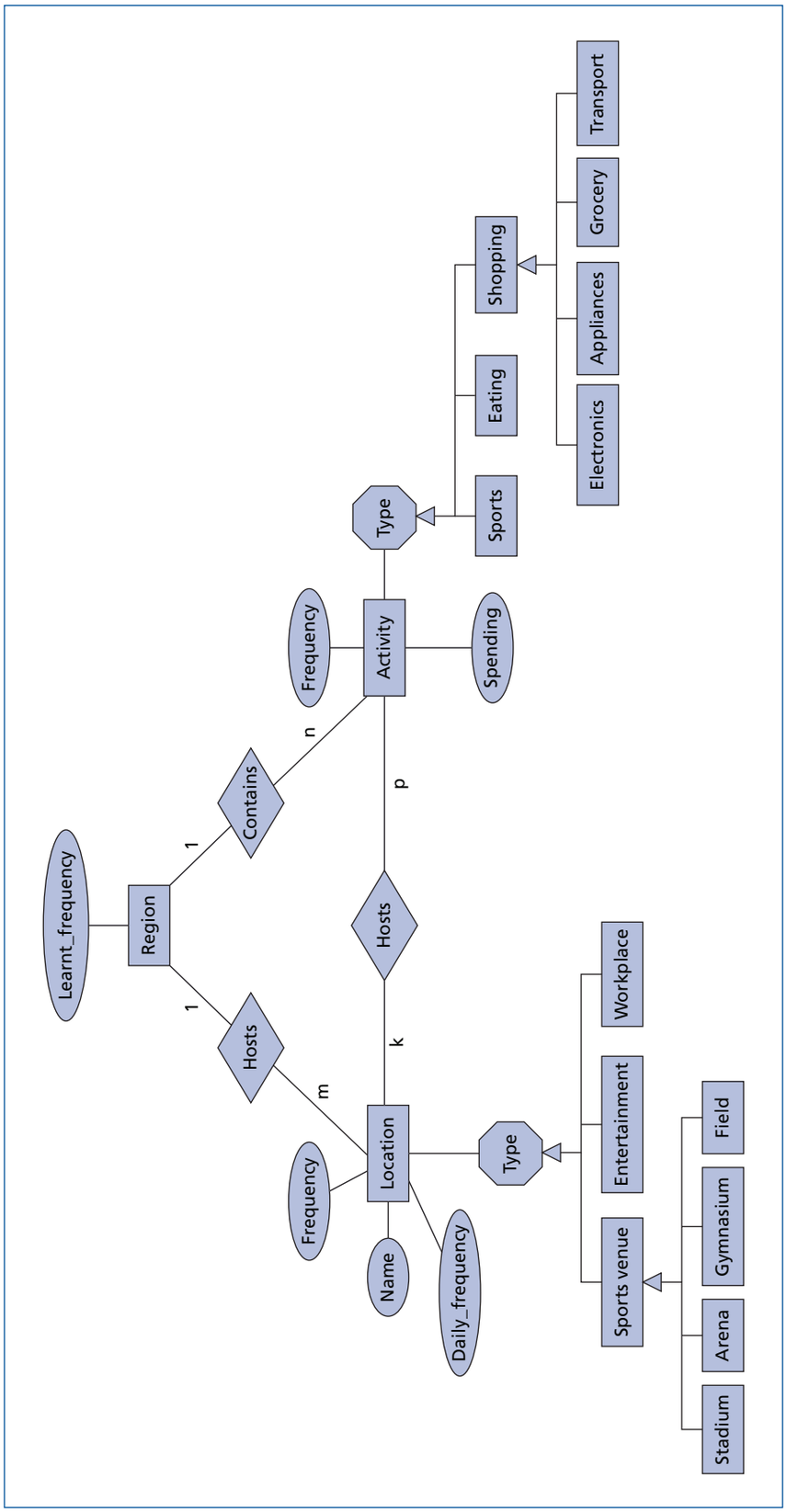


Figure 1.
Subset of the user profile schema.

BLTJ132_20301.qxd 6/16/08 10:51 PM Page 35

every Friday, an advertisement can be triggered that sends lunch coupons to the user for restaurants near that location a half-hour before his usual lunch time on Friday. Similarly, since the user's location is also being monitored continuously, the INA framework might be able to determine whether the user is starting to move toward a particular destination, and a coupon targeting that location can be sent to the user. Note that advertising is simply one context in which the user profile can be utilized. Other applications can also be envisioned that use this profile information to provide additional value-added services to the user.

Privacy Considerations

In order to collect user preferences that can be used to provide personalized services, such as the targeted advertising described in the previous section, users must consent to the collection and use of their personal information. In a traditional system where users set up their profiles and personal preferences explicitly, they know exactly what information they are disclosing and often can review and modify that information at will. They have both visibility of and control over their private information. Simply having them provide the information directly to the system implies their consent. However, in a system that learns about the user's habits and preferences through monitoring and learning, combined with only sporadic user interaction to confirm some hypothesis or obtain a specific piece of information, the user is never fully aware of everything that the system may know about him. Even though we plan to provide users with an interface that would allow them to view and manipulate their personal information, the system's knowledge about each user is always changing as the system monitors users and makes inferences. Therefore, users are never 100 percent certain of everything that the system knows about them, of what they might have intentionally or inadvertently disclosed, or of what may have been learned about them. They lose some of the visibility and control that are key components of privacy assurances. It was therefore necessary to investigate whether the level of privacy intrusion that the proposed system inflicts upon users would be acceptable. Users would have to trade off some privacy perception in order to enable these services. Would they accept? We conducted a study to find out.

We developed a questionnaire in which we presented users with five different scenarios, illustrating the type of applications that the INA framework would enable. Each scenario consisted of the following:

- A commonly understood day-to-day situation: going to the movies, eating at restaurants, taking a bus to school or work, shopping for groceries, getting books from the library,
- A summary of the observations a system would make to understand the user's habits and preferences: monitoring locations and travel habits, phone calls, purchases, and similar activities, to infer routine habits and preferences,
- An example of the personalized service that would be provided to the user to assist in day-to-day situations as a result of the learned habits and preferences: provide movie or restaurant recommendations with reviews, provide discount coupons, or alert the user to schedule changes.

For each scenario, we asked the user a series of questions, slightly different for each scenario but always following a common theme:

1. *Are you the kind of person described in the scenario?* If they answered no, they were instructed to proceed anyway with the questionnaire by trying to pretend they were such a person.
2. *Would you like to have the service described in the scenario provided to you?* The questionnaire further probed whether they would want it to be free, whether they would pay for it, or whether they would not want it at all.
3. *Would you agree to have a service provider or some trusted third party collect this type of information about yourself, in order to provide this service?*
4. *How should the information be collected?* We provided respondents with a range of options, from fully automated monitoring and learning, to user-provided preferences, to no information collected at all.

The first two questions are interest questions, assessing the respondent's interest in the service, and the last two questions are privacy questions, assessing

BLTJ132_20301.qxd 6/16/08 10:51 PM Page 36

the respondent's privacy concerns and acceptable means to obtain the necessary information.

The questionnaire also contained questions to collect demographic information about the respondent, information about his or her general privacy protection preferences, and how the respondent manages personal data. The questionnaire was administered over the Web to 236 respondents: 171 in the United States and 65 in Hong Kong. (In Hong Kong, respondents could choose to answer the questionnaire in either English or Chinese.) Respondents were from all ages, with a strong concentration in the 19 to 23 age bracket, accounting for 133 of the 236 respondents. Respondents were recruited over the Web, with targeted advertising placed on FaceBook* boards for eight colleges in the United States and two colleges in Hong Kong. The encouragement for participating was an entry into a drawing for a prize with a value of $500. The prize was awarded to a randomly selected participant.

Table I summarizes the interest in each of the proposed services.

The first column indicates the fraction of all respondents who said they would want the service, either as a free service or as a paid service. The second column indicates the fraction of respondents who self-identified as being the type of person described in the scenario. The third column indicates the fraction of respondents who self-identified with the scenario that said they would want the service. In every case except the library scenario, there is a significant difference between the desire for the service between the people who self-identify with the scenario and those who do not. This indicates to us that the responses of those who do not self-identify with the scenarios are likely not to be as reliable as the responses of those who do. Therefore, we concentrated our analysis on the responses of the users who self-identified with each scenario. For future studies, we will recruit subjects who are more likely to self-identify with the proposed application scenarios.

In analyzing the responses to the privacy questions, we found that over 51 percent of respondents would agree to provide personal information to the service provider in order to enable the provisioning of the service. When we restrict the analysis to the responses of those who self-identify with the scenario, the number jumps to 67 percent. And when we consider only those who both self-identified with the scenario and said they would want the service, the acceptance rate further increases to 71 percent. Thus we find that the service itself and the desire for such a service are stronger predictors of whether someone is willing to share personal information than the person's individual privacy concerns.

The responses regarding the methods used to collect information were unreliable. A large number of respondents indicated that they would like to fill out a survey with their preferences. This finding led us to suspect that the responses may have been influenced by the method of administration of the test—a survey itself—and the fact that the respondents were self-selected. Because of the confounding effect of the test method, we are designing and deploying additional test methods to obtain answers as to the preferred information collection methods that will not suffer from bias due to the experiment design. We will report those results separately.

Table I. Interest level in each service.

Scenario	Wants the service (all respondents)	Self-identifies	Wants the service (among self-identified only)
Movies	0.76	0.64	0.85
Library	0.70	0.74	0.70
Groceries	0.58	0.20	0.78
Restaurant	0.68	0.72	0.74
Bus	0.76	0.62	0.96

BLTJ132_20301.qxd 6/16/08 10:51 PM Page 37

Although the survey was preliminary, with only 236 respondents in only two countries, we are encouraged by the results. The types of services enabled by INA are interesting enough to get people to share personal information in order to get the services. Furthermore, those most likely to be early adopters of the service—those who already self-identified and indicated they would want the service—are more likely to disclose their information, enabling service launch.

Pilot Study: Learning From Location and Context Logs

We performed a pilot study where a few volunteers agreed to carry a smartphone (HP iPAQ* Pocket PC) with an added client that tracked the Global System for Mobile Communications* (GSM*) cell IDs that the device was in contact with over time. The client also provided a simple graphical user interface (GUI) interface that enabled the user to enter additional context information about what he or she was doing at a given time. The user could enter the following pieces of information: current location type (from a fixed hierarchy of location types, e.g., office, home, or mall), current primary context (also from a fixed hierarchy, e.g., at work or in transit), and an optional secondary context from the same set (e.g., the user could be in transit but working). This information was periodically uploaded onto a server, and annotated where possible. If the exact coordinates of a given location were known, these could be entered after uploading. The study was performed over a period of about four months. The uploaded information was then analyzed to determine what could be automatically learned about the user behavior, with or without taking the explicit user entries into account.

We now provide examples of the types of information one can infer from a user's location logs and what additional knowledge can be gained if the user provides annotations of the location and the associated activity context. Each of these learning processes results in some characterization of the user's lifestyle that is revealed through his movement patterns.

In particular, we present a method for inferring the spatial distribution of visited locations, a method to segment a cell ID log into a user's journeys and the places he visited, and a description of the user's activity contexts that are associated with the places he visited. These methods all start with minimal data, i.e., time-stamped logs of cell IDs and user-entered context descriptors. The statistical algorithms chosen are applied with temporal distances (or a shared portion of time) that are converted to similarity measures between different locations. The inferred spatial distribution provides a visualization that approximates the true geographical distribution of the user's activities. It can aid understanding of a user's travel patterns, provide a frame of reference upon which inferred journeys can be displayed, and lead to quantitative measures of the location concentration that differentiates among the gross patterns of different users' lifestyles. The inferred journeys and visited places and the associated distribution of activity contexts give additional detailed description of the user's behavior at specific moments. Jointly, these can be used to classify users into different types and to predict their behavior in the near future. This will be useful in behavior-sensitive applications such as advertising. For example, advertisements can be better targeted to potentially interested users both statically (by user classification) and dynamically (by behavioral prediction).

Inference on the Spatial Distribution of Visited Locations

A gross characterization of a user's movement patterns can reveal some essential aspects of his lifestyle. One way to derive such a characterization is by analyzing the spatial distribution of his visited locations. This can be done by a statistical method known as multidimensional scaling. The method takes as input an $N \times N$ matrix whose entries are distances between each pair of N points. It then computes the position of each point in an embedding space (e.g., a two-dimensional plane) to construct a map, such that the distances between the mapped points are proportional to their input distances. This method is applicable to distances that are not necessarily geographical. In our application to the time-stamped cell ID logs, we use the smallest gap between the visit times to each pair of locations to approximate their

BLTJ132_20301.qxd 6/16/08 10:51 PM Page 38

distance. A map obtained this way is more a representation of the perceived distances than of the physical distances. For example, a location in California could be at the same "distance" from central New Jersey as one in Washington, D.C., because both are about five hours away (one by air and the other by road). Arguably, for many purposes, the perceived distances are just as useful.

In **Figure 2** we show an example use of this method to lay out the cell IDs visited by four individuals on a two-dimensional plane. Immediately one can see obvious differences between the individuals: some of them have traveled to several widely separated destinations from a central area, whereas one of the individuals followed a very restricted trajectory between only two groups of destinations. A richer characterization is obtained when we mark each location with the earliest hour in which the individual ever visited the location.

Inference on the User's Frequented Regions

Our goal here is to infer the regions that a user frequents and when he frequents them, using only the cell ID logs. Notice the inferred regions are not pinpoint locations because of the coarse resolution of cell IDs, and the user may visit one or several locations within a single region.

For illustration, **Figure 3** shows a plot of a sample cell ID log for a telecommuter in the pilot study from a Sunday afternoon to Wednesday morning. The cell IDs are labeled in the order they appear (i.e., 1 represents the first cell ID seen during the time window, 2 represents the second cell ID seen, and so on). Points on the plot are colored on the basis of the location annotation that the user inputs. The plot depicts the movement pattern of the user: Sunday at home, Monday work from home during the day and visit the child's school in the evening, and Tuesday a conference seminar.

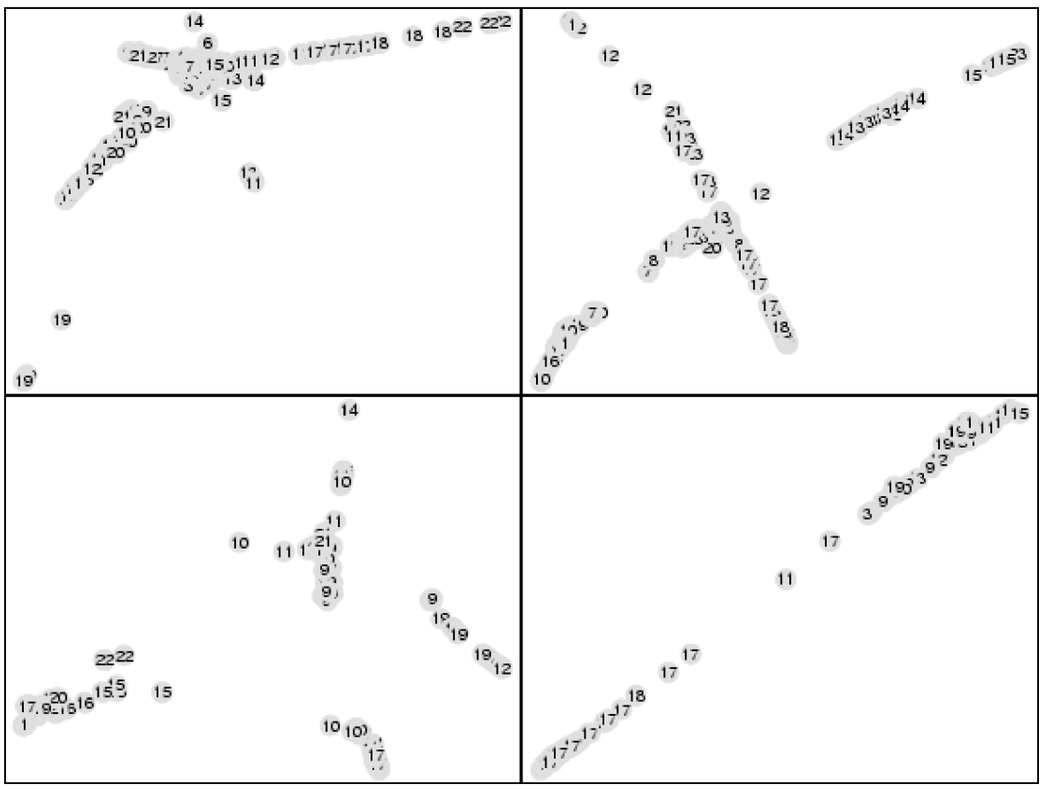


Figure 2.
Recovered spatial distribution of visited locations for four individuals.

BLTJ132_20301.qxd 6/16/08 10:51 PM Page 39

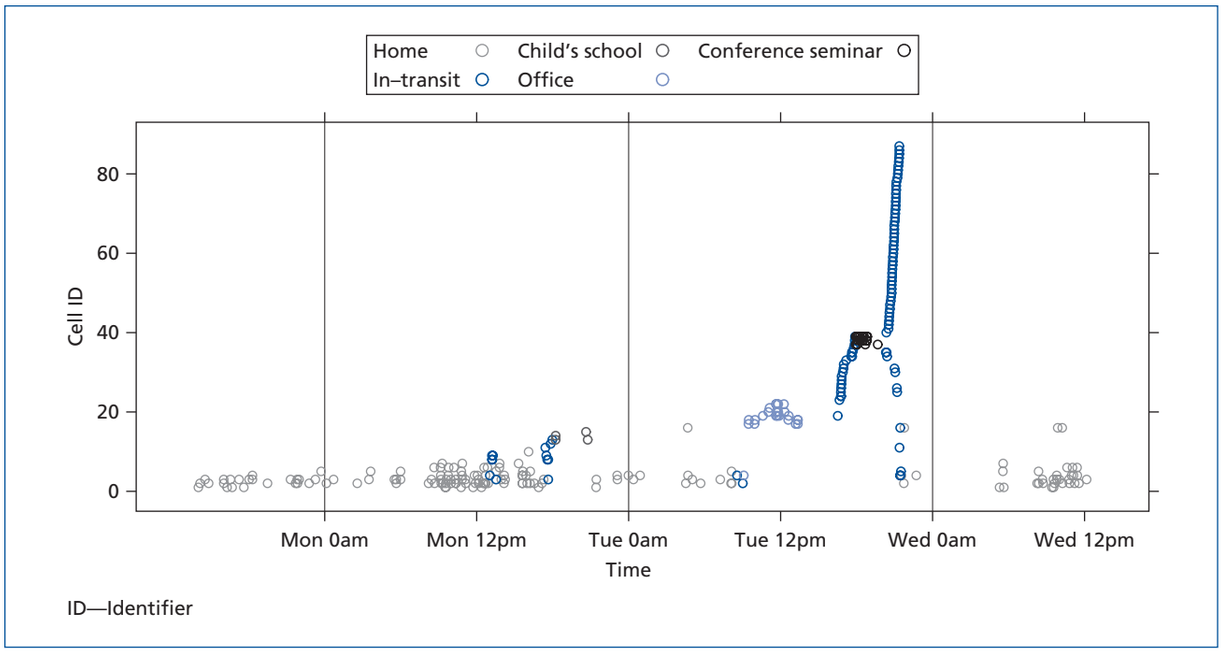


Figure 3.
A sample cell ID log for a telecommuter from a Sunday afternoon to Wednesday morning.

There are a few interesting observations here. First, from the cell ID logs alone, it is possible to locate the traveling periods by monitoring cell ID changes over time. When the user is not in transit, the observed cell IDs oscillate, and the oscillation pattern changes from time to time. However, when the user is in transit, new cell IDs that have not been observed in the past time window (say 30 minutes) appear. In fact, long-distance travel will typically result in more new cell IDs, which makes it easier to distinguish from the cell oscillation. For example, travel from the office to the conference seminar and back to home on Tuesday is more visible than the local travel from home to the child's school. Therefore, by locating traveling periods, we can identify cell clusters, i.e., a sequence of cell IDs that represents a possible destination where the user remained for a certain duration and performed some activities. Second, notice that two different cell ID clusters may represent the same region because of the random oscillation pattern. We can actually identify frequented regions by coalescing the cell clusters. For example, one cell cluster may consist of a sequence of cell IDs (1, 2, 3, 2, 4, 2), and another cell cluster at a different time may consist of (1, 3, 2, 4, 5, 2), but both represent the home location. Using statistical clustering techniques, we can coalesce both of these cell clusters into a single frequented region.

Our inference procedure on the frequented regions consists of the two steps discussed. First, we identify the cell clusters by monitoring cell ID changes. We define a statistic based on the cell ID change at each time instance and use statistical change detection methods [3] to identify the traveling periods. Next, we extract frequented regions by coalescing these cell clusters, where each frequented region is represented by a list of cell IDs associated with some probabilities for each of the cell IDs. Note that as a by-product, we can also make inferences on the traveling pattern between these frequented areas. However, we will leave this discussion to another paper as it is not the focus of this study.

The statistical approach we undertake here to uncover the frequented regions is quite different from the deterministic approach of [12]. In [12], the cell clusters where the user stayed for a certain duration are obtained through subgraphs of cells that satisfy a certain oscillation pattern on a cell-transition graph, while here

BLTJ132_20301.qxd 6/16/08 10:51 PM Page 40

they are obtained as a by-product of identifying transit periods by using change point detection. In addition, since our approach is probabilistic, two frequented regions can share cells, as long as the user spends an insignificant amount of time in the overlapping cells. However, because of the deterministic nature of the algorithm, two locations (equivalent to regions in this paper) in [12] must contain nonoverlapping cells; that may not be realistic as a cell may serve a geographical area of many miles in sparse areas. Furthermore, the parameters in defining cell clusters and locations in [12] are harder to train when the cell tower densities vary. In our case, since the detection of transit will be harder in sparse regions than it is in dense regions, we only need to target the worst case scenario for training.

After the frequented regions and their visits are identified, for advertising purposes, we derive a succinct statistical summary of the visit patterns to these frequented regions. Our summary is based on a statistical method called mixture modeling [14], where we model the joint distribution of the visit start times and durations using two-dimensional Gaussian mixtures. **Figure 4** shows an example of this procedure for a user's visits to his home. In this example, the joint distribution of the visit time of day and duration can be well approximated by a Gaussian mixture distribution with three components. Each point represents a visit and is colored by its classified Gaussian component. To capture the day of week variations in the visit patterns, for each day of the week, we also use a Gaussian mixture model for the joint distribution with the same Gaussian mixture components, but with different mixing probabilities. Each ellipse indicates the 85 percent confidence interval (at one standard deviation) of a Gaussian component centered at the mean.

Inference on the Activity Context Associated With the Locations

In the case where a user provides additional information associated with the location logs, a much richer characterization of the user's lifestyle can be inferred. We illustrate this with examples from the pilot study, where some users recorded their activity context whenever the context is changed.

As in the case of frequented regions, one way to characterize a user's lifestyle is to identify his most frequently performed activities, or those that occupy the largest portions of his time, such as home, work, or meals. Similarly to the frequented regions, we can also derive a statistical summary of his frequent activities in terms of time and duration. Besides the frequent activities, the temporal distribution of the activities provides another perspective on the user's lifestyle. This includes the typical sequence of activities he performs in a day, and the allocation of time to each major activity over the course of a day and over different days of

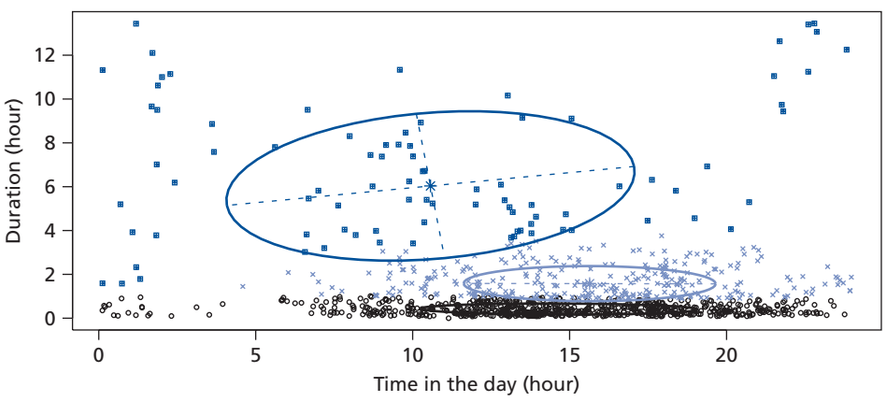


Figure 4.
A Gaussian mixture model of three components for the joint distribution of the time in the day and duration of a user's visits to his home.

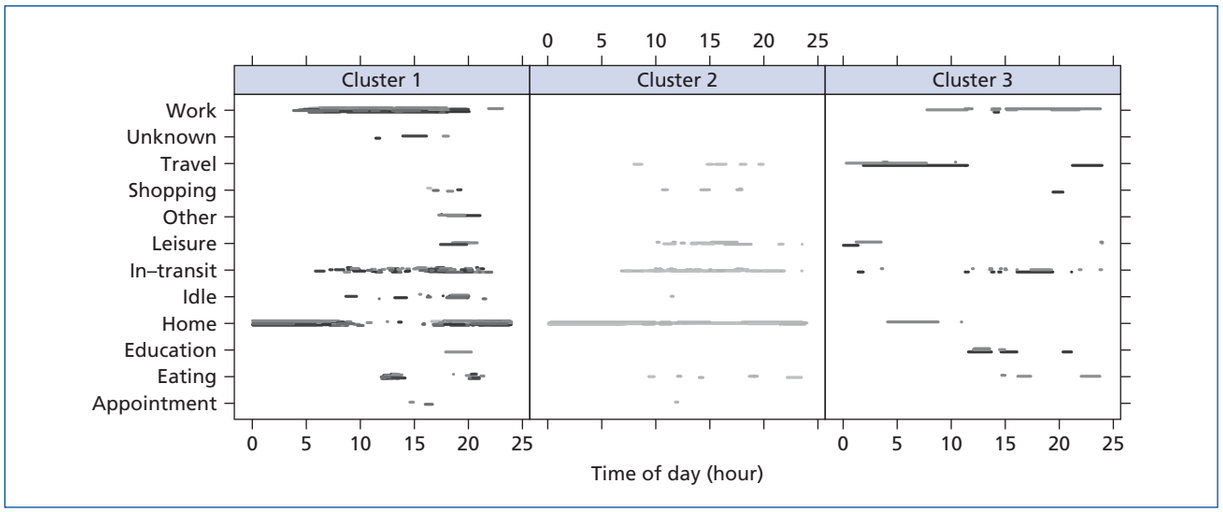


Figure 5.
The daily activity patterns of a user in the pilot study, grouped by clusters.

a week. As an example, **Figure 5** shows the result of a clustering analysis on the sequence of daily activities for a user over a one-month period. Each panel represents a single cluster of the user's daily activities where different colors indicate different days. It is interesting to see that cluster 1 is the user's normal workday routine (home?office?home), cluster 2 is the users' normal weekend routine (home?leisure or in transit), and cluster 3 presents the user's occasional travels. Besides these dominant activity patterns, deviations from these patterns may also be of interest. For instance, a user may be found on vacation, while his location log suggests that he is out of town. This may imply his potential interest in future lodging or vacation/entertainment packages.

INA Architecture

We now present some details on the INA framework, in the context of an advertising scenario. The overall framework is presented in **Figure 6**.

We assume that a user's location is monitored via an application on his mobile device, which internally logs time-stamped information about the mobile cell towers to which the device is connected at any point in time. Periodically (e.g., every 15 minutes) this log is transmitted to an INA server for storage and further analysis. In our current scenario, we focus on the GSM network and log cell IDs on the mobile device, though analogous techniques can be applied to other network architectures as well. Note that the GSM cell ID that the device is connected to changes over time, even if the device is not in use and is physically stationery: i.e., the device may still bounce between cell IDs of a few cell towers that are close to the device. We term such changes *oscillations*, and these need to be filtered out when performing the location analysis.

As mentioned earlier, we do not use GPS capabilities on the device (though that would make the analysis simpler) for a number of reasons:

- GPS-enabled mobile phones are not yet prevalent,
- GPS signals do not work well indoors,
- GPS signal-lock can sometimes take a long time, and
- Continuous GPS usage on a mobile phone drains the battery rapidly.

On the other hand, obtaining cell ID information on the mobile device is relatively inexpensive, since this information is received by the mobile device continuously in a passive mode. The disadvantage of working with cell IDs is that the granularity of location information is reduced. To alleviate this problem, we assume that we can get information about the mapping between cell IDs and their corresponding latitude-longitude coordinates. This information,

BLTJ132_20301.qxd 6/16/08 10:51 PM Page 42

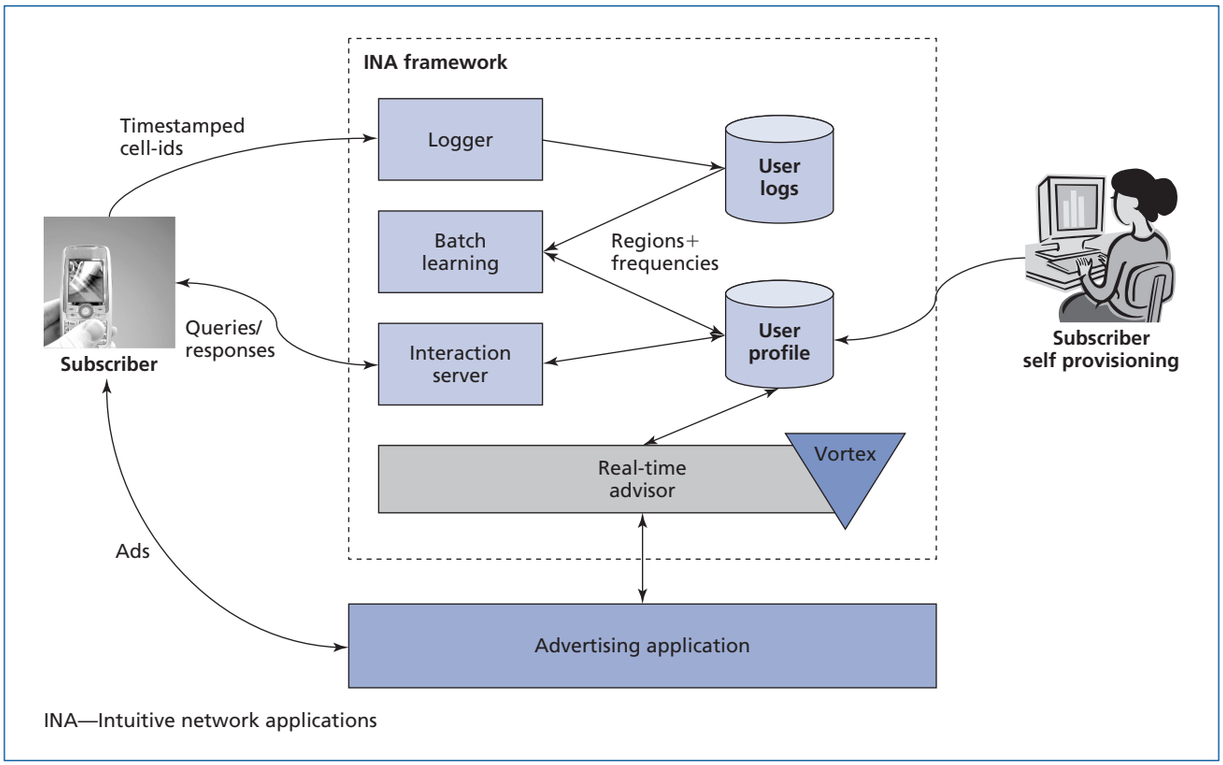


Figure 6.
High-level INA architecture applied to the advertising scenario.

though not publicly available, can be obtained in partnership with a mobile operator.

Periodically (e.g., once a day), statistical algorithms are run on the cell ID logs of each user, to determine the favorite regions (i.e., clusters of cell IDs) which he frequents, with what frequency, and during what times, using the statistical algorithm outlined previously. This information is then stored in the user's profile and can later be refined by sending explicit queries to the user. This summary data is obviously valuable for advertising purposes (i.e., location-aware ads), or for the advertiser to determine the approximate size of a target audience for future advertising campaigns (e.g., how many people might visit a given mall in the next two weeks).

We also have a mechanism for asking explicit user queries, which can either be of an ad hoc nature or be more directed and based on information that has been learned in the past. Typically, these queries will target information that cannot be completely automatically inferred. In some cases, though, such queries will try to confirm (or deny) whether the learned information is accurate or not.

Figure 7 presents our user interaction framework. In our current version, the INA framework requires that the user queries be created manually, though in the future, we might explore the use of description logic to create these queries semiautomatically. These queries are created as templates, which can later be instantiated on certain data in the user profile before being posed to the user. Each query template consists of a precondition that determines whether the query is ready to be asked, an input binding that specifies which parts of the user profile are used to instantiate the query template, the actual question to ask the user, and an output binding that determines how the user input is to be handled.

We use the Velocity [1] framework from Apache* to create and execute query templates. In

BLTJ132_20301.qxd 6/16/08 10:51 PM Page 43

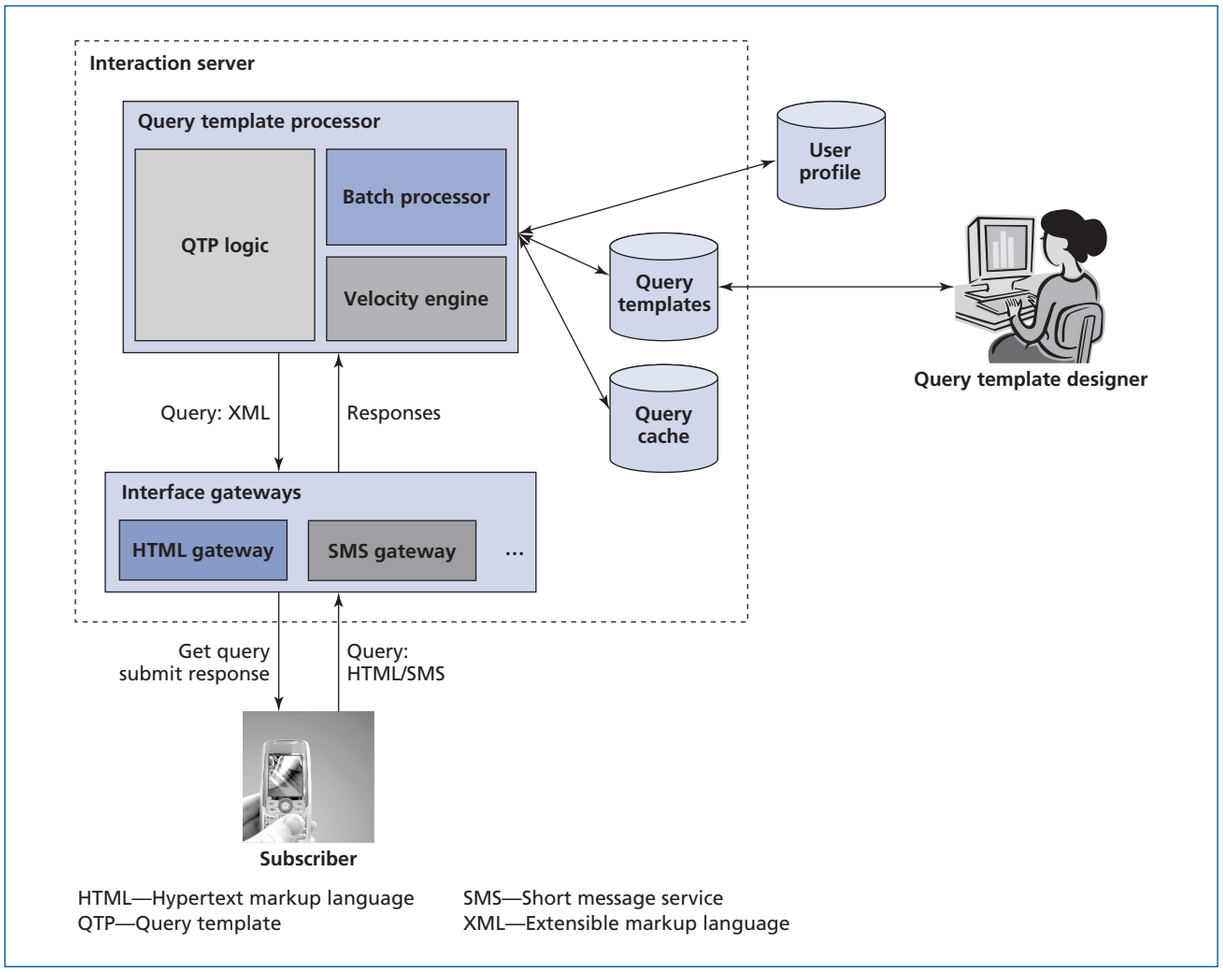


Figure 7.
High-level INA architecture applied to the advertising scenario.

the current prototype, the actual query is posed to the user via a Web form: i.e., the mobile device is assumed to support a modern Web browser. Note that the Velocity query templates do not emit Hypertext Markup Language (HTML) directly when processed. The INA framework is intended to support devices that do not understand HTML (e.g., the interaction can be purely over SMS). Hence, the processing of a Velocity template produces an abstract version of the instantiated query, in an Extensible Markup Language (XML) syntax. This XML document is then transformed via a mapping component (in this case, the HTML gateway) into HTML. The bulk of the HTML mapping functionality is handled via an Extensible Stylesheet Language Transformation (XSLT) script. Various other such mapping components (gateways) can be added in the future to handle other modes of interaction.

Figure 8 and **Figure 9** show two sample instantiated queries displayed via a user's Web browser. The first query, shown in Figure 8, aims to verify the location of the user's home region, determined via analysis of his location logs, for users whose home address is not known (e.g., for prepay mobile users). The second query, shown in Figure 9, tries to determine the name of a specific location that the user visits within a given frequented region, and the activity he usually performs at that location.

BLTJ132_20301.qxd 6/16/08 10:51 PM Page 44

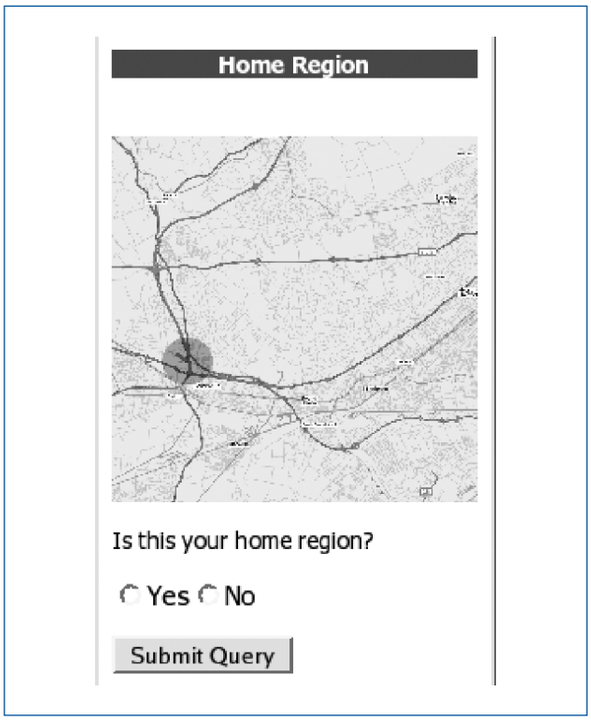


Figure 8.
Sample query for determining a user's home.

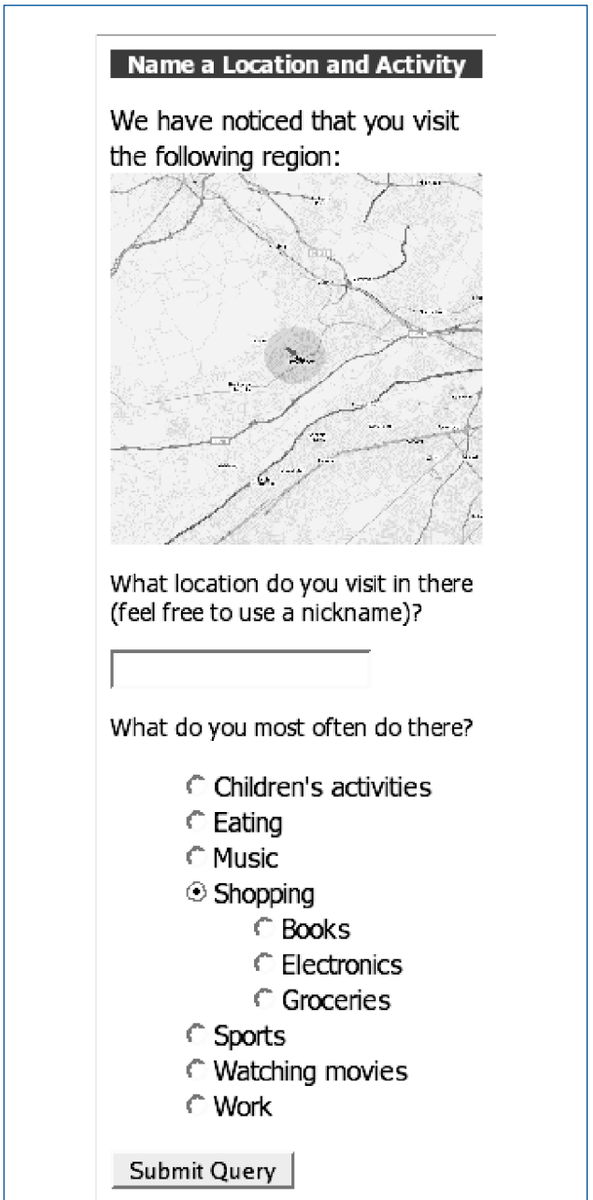


Figure 9.
Sample query for determining a user's activity at a specific location.

Related Work

There has been some prior work on combining user input with learning techniques, in the context of the Cognitive Assistant that Learns and Organizes (CALO) project. An instantiation of those techniques is found in the Preference Learning through Interactive Advisable Nonintrusive Training (PLIANT) system [9] where the system interacts with the user to propose an acceptable schedule for a given meeting, while also taking into account the history of the user's past scheduling preferences. However, the scope of this system is limited to calendar scheduling, and it is not clear whether it can be adapted to other domains as well.

Techniques for inference and prediction of user locations are described in [2, 4, 15]. An interesting effort in modeling human activities using location logs is the Reality Mining project at the Massachusetts Institute of Technology (MIT) [8, 13], where analysis of social interactions is a major theme. Modeling and inference of user context were also studied in [6, 10, 16]. There has also been some preliminary work trying to establish a link between user location and context [16, 17], as well as showing that location changes over time can be used to learn a user's activities and even predict future activities [15].

BLTJ132_20301.qxd 6/16/08 10:51 PM Page 45

Conclusions

This paper presents a framework for utilizing learning and human-factors-driven approaches to user profile acquisition and illustrates the use of this framework for advertising purposes. The framework is built in a modular fashion, allowing it to be rapidly customized to new information sources, or to support different mechanisms for interactions with the user. It can take advantage of the network convergence trend to span multiple access network technologies, thus enabling the overall habits and intentions of a user to be captured, rather than his interactions with a particular service/network only.

We are also exploring ways to entice users to reveal information about themselves, via either automatic monitoring or explicit interaction. If free/discounted service is the only perceived benefit, users may not have enough incentive to provide truthful information about themselves. However, there are ways to alleviate this by offering incentives with a more immediate perceived benefit. For example, if the user reveals information about his habits, he is entered into a raffle with prizes targeted toward that particular activity.

The INA framework is not limited to advertising only. It can serve as an enabler that allows a large number of personalized applications to take advantage of the acquired user profile, in order to provide a more personalized experience to the user. Future work involves determining how to "age" previously learned information at an appropriate rate since the user's behavior might evolve over time.

Acknowledgments

We would also like to thank our Bell Labs colleagues: Willem Romijn for allowing us to use his cell ID tracking software for the HP iPAQ Pocket PC and customizing it for our needs and Erik Meeuwissen for the numerous discussions we had on location tracking and analysis issues. We would also like to thank one of our summer interns, Charina Leung, for her help with the user study.

*Trademarks

Apache is a trademark of The Apache Software Foundation.
Facebook is a trademark of Facebook, Inc.
GSM and Global System for Mobile Communications are registered trademarks of the GSM Association.
iPAQ is a trademark of Compaq.
UNIX is a registered trademark of The Open Group.
Virgin Mobile is a registered trademark of Virgin Enterprises Ltd.

References

[1] Apache Software Foundation, "The Apache Velocity Project," <http://velocity.apache.org/>.
[2] D. Ashbrook and T. Starner, "Learning Significant Locations and Predicting User Movement With GPS," Proc. 6th Internat. Symposium on Wearable Comput. (ISWC '02) (Seattle, WA, 2002), pp. 101–108.
[3] M. Basseville and I. V. Nikiforov, Detection of Abrupt Changes: Theory and Application, Prentice Hall, Englewood Cliffs, NJ, 1993.
[4] A. Bhattacharya and S. K. Das, "LeZi-Update: An Information-Theoretic Approach to Track Mobile Users in PCS Networks," Proc. 5th Annual ACM/IEEE Internat. Conf. on Mobile Computing and Networking (MobiCom '99) (Seattle, WA, 1999), pp. 1–12.
[5] Blyk, <http://www.blyk.co.uk/>.
[6] T. Buchholz, A. Kuepper, and M. Schiffers, "Quality of Context: What It Is and Why We Need It," Proc. 10th Internat. Workshop of the HP OpenView University Association (HPOVUA '03) (Geneva, Switz., 2003).
[7] R. Dinoff, T. K. Ho, R. Hull, B. Kumar, D. Lieuwen, P. Santos, and H. Ren, "Intuitive Network Applications: Learning for Personalized Converged Services Involving Social Networks," J. Comput., 2:6 (2007), 72–84.
[8] N. Eagle and A. Pentland, "Reality Mining: Sensing Complex Social Systems," Personal and Ubiquitous Computing, 10:4 (2006), 255–268.
[9] M. T. Gervasio, M. D. Moffitt, M. E. Pollack, J. M. Taylor, and T. E. Uribe, "Active Preference Learning for Personalized Calendar Scheduling Assistance," Proc. 10th Internat. Conf. on Intelligent User Interfaces (IUI '05) (San Diego, CA, 2005), pp. 90–97.
[10] K. Henricksen, J. Indulska, and A. Rakotonirainy, "Modeling Context Information in Pervasive Computing Systems," Proc. 1st Internat. Conf. on Pervasive Computing (Pervasive '02) (Zurich, Switz., 2002), published in Lecture Notes in Comput. Sci. (LNCS 2414) (F. Mattern and M. Naghshineh, eds.), Springer, Berlin, New York, 2002, pp. 167–180.
[11] R. Hull, B. Kumar, D. Lieuwen, P. F. Patel-Schneider, A. Sahuguet, S. Varadarajan, and A. Vyas, "Enabling Context-Aware and Privacy-Conscious User Data Sharing," Proc.

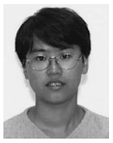
BLTJ132_20301.qxd 6/16/08 10:51 PM Page 46

5th IEEE Internat. Conf. on Mobile Data Management (MDM '04) (Berkeley, CA, 2004), pp. 187–198.
[12] K. Laasonen, M. Raento, and H. Toivonen, "Adaptive On-Device Location Recognition," Proc. 2nd Internat. Conf. on Pervasive Computing (Pervasive '04) (Vienna, Aus., 2004), published in Lecture Notes in Comput. Sci. (LNCS 3001) (A. Ferscha and F. Mattern, eds.), Springer, Berlin, Heidelberg, New York, 2004, pp. 287–304.
[13] Massachusetts Institute of Technology (MIT), "Reality Mining," <http://reality.media.mit.edu/>.
[14] G. McLachlan and D. Peel, Finite Mixture Models, Wiley, New York, 2000.
[15] E. Meeuwissen, P. Reinold, and C. Liem, "Inferring and Predicting Context of Mobile Users," Bell Labs Tech. J., 12:2 (2007), 79–86.
[16] N. T. Nguyen, D. Q. Phung, S. Venkatesh, and H. Bui, "Learning and Detecting Activities From Movement Trajectories Using the Hierarchical Hidden Markov Model," Proc. IEEE Comput. Society Internat. Conf. on Comput. Vision and Pattern Recognition (CVPR '05) (San Diego, CA, 2005), vol. 2, pp. 955–960.
[17] A. J. H. Peddemors, M. M. Lankhorst, and J. de Heer, "Presence, Location, and Instant Messaging in a Context-Aware Application Framework," Proc. 4th Internat. Conf. on Mobile Data Management (MDM '03) (Melbourne, Austral., 2003), published in Lecture Notes in Comput. Sci. (LNCS 2574) (M.-S. Chen, P. K. Chrysanthis, M. Sloman, and A. Zaslavsky, eds.), Springer, Berlin, Heidelberg, New York, 2003, pp. 325–330.
[18] J. W. Sammon, "A Nonlinear Mapping for Data Structure Analysis," IEEE Trans. Comput., C-18:5 (1969), 401–409.
[19] W. N. Venables and B. D. Ripley, Modern Applied Statistics with S-Plus, 2nd ed., Springer, New York, 1997.
[20] Virgin Mobile, "Sugar Mama," <http://www.virginmobileusa.com/stuff/sugarmama.do>.

(Manuscript approved April 2008)



NILTON BILA is a graduate student at the University of Toronto. His research interests include mobile systems, ubiquitous computing, and computer networks. The overarching theme of his work is enabling pervasive access to information for everyone. He has a M.Sc. in computer science from the University of Toronto, Ontario, Canada.



JIN CAO is a member of technical staff in Bell Labs' Statistics, Learning, and Computing Research Group in Murray Hill, New Jersey. She conducts basic and applied research in statistical sciences. She received a Ph.D. from the Department of Mathematics and Statistics, McGill University, Montreal, Quebec, Canada. Her Ph.D. thesis was on the statistical analysis of brain images. Dr. Cao's current research focuses on statistical problems arising from data networks, for example, network tomography, traffic modeling and simulation, performance analysis, streaming data analysis, and network anomaly detection.



ROBERT DINOFF is a member of technical staff in Bell Labs' Computing and Software Principles Research Department in Murray Hill, New Jersey. He received his bachelor of technology (BT) degree in computer science from Rochester Institute of Technology, Rochester, New York. His interests early on were in UNIX kernel development and UNIX system administration. He then moved into work on directories where he was part of the team that developed POST, the Lucent corporate enterprise directory; LUCID, the Bell Labs directory; the Alcatel-Lucent Datagrid™ modeling tool; and the Privacy-Conscious Personalization Server.*



TIN KAM HO is the leader of Bell Labs' Statistics, Learning, and Computing Research Group in Murray Hill, New Jersey. She conducts basic and applied research in statistical sciences and scientific computing. Her interests are in pattern recognition, data mining, and computational modeling and simulation. She pioneered research in decision combination in multiple classifier systems, random decision forests, data complexity analysis, and many topics in image and text analysis. She has also led major efforts on modeling and monitoring large-scale optical transmission systems. Recently she worked on problems in user profiling, optical network diagnostics, and customer experience management. She received a Ph.D. in computer science from State University of New York (SUNY) at Buffalo. Dr. Ho is Editor-in-Chief of the journal Pattern Recognition Letters and an elected Fellow of the International Association for Pattern Recognition and the Institute of Electrical and Electronics Engineers (IEEE). She has over 80 publications and was granted 7 U.S. patents.

BLTJ132_20301.qxd 6/16/08 10:51 PM Page 47



BHARAT KUMAR is a member of technical staff in Bell Labs' Computing and Software Principles Research Department in Murray Hill, New Jersey. His areas of interest include Web technologies, policy management, and infrastructure for converged services, and he has numerous publications and several patents in these areas. He holds a Ph.D. in computer and information science from The Ohio State University in Columbus. Dr. Kumar was the main architect of the Alcatel-Lucent Vortex™ high-speed rules engine and is currently involved in the design of the Intuitive Network Applications framework.



DANIEL LIEUWEN is a member of technical staff at Google Inc. in New York City. At Bell Labs, his early research focused on implementing object-oriented databases, materialized views, main-memory databases, and active databases. He then moved into work on data integration, Lightweight Directory Access Protocol (LDAP) triggers, the Web videocassette recorder (WebVCR), and training the Lucent Technologies LambdaRouter micro-electro-mechanical systems (MEMS) optical cross-connect. During his Bell Labs career, he also was involved in both design and implementation of the Alcatel-Lucent Datagrid™ modeling tool, and on privacy-conscious personalization (PCP) and intuitive network applications (INA). He received a B.S. in mathematics and computer science from Calvin College, Grand Rapids, Michigan, and M.S. and Ph.D. degrees in computer science from the University of Wisconsin at Madison. Dr. Lieuwen has 40 published papers and 12 patents



PAULO SANTOS is a member of technical staff in Bell Labs' Network Data and Services Research Department in Murray Hill, New Jersey. He holds an M.S. in information and computer science and a Ph.D. in computer science from the Georgia Institute of Technology in Atlanta. He is also a Certified Human Factors Professional. In his current work, he is involved in defining the user experience for personalized interactive services. He is also involved in researching consumers' communications habits and preferences, as well as privacy, security, and other human factors affecting adoption of new services and selection of communication services. Dr. Santos's research interests span a broad area of human factors in communications services, including mobile interaction, ubiquitous computing and communication user experiences, accessibility and internationalization of user interfaces, and the processes for continuous improvement of user interfaces. ◆

BLTJ132_20301.qxd 6/16/08 10:51 PM Page 48