"All-about" Diaries: Concepts and Experiences

Laura Ferrari
DISMI
University of Modena and Reggio Emilia
ITALY
laura.ferrari@unimore.it

Marco Mamei
DISMI
University of Modena and Reggio Emilia
Italy
marco.mamei@unimore.it

Franco Zambonelli
DISMI
University of Modena and Reggio Emilia
Italy
franco.zambonelli@unimore.it

ABSTRACT

Smart phones and pervasive computing technologies enable the vision of all-about diaries: tools for recording, in a browsable and machine-processable format, the everyday activities and events of people, communities, objects and places. Diaries offer a wealth of opportunities for consumers and industries. Yet, while proposals exist indicating promising approaches to implement parts of them, several challenges still have to be faced to produce fully-edged working systems. In this paper we discuss opportunities and technologies that enable such diaries to be created. Then, we present a prototype of a diary based on location data.

Keywords

Life Logging, Pervasive Computing, Context Awareness, Pattern Analysis Location-based services

Categories and Subject Descriptors

I.5 [**Computing Methodologies**]: Pattern Recognition; I.2.3 [**Artificial Intelligence**]: Probabilistic Reasoning

1. INTRODUCTION

Keeping track of personal activities and events (i.e., keeping a diary) has always been a popular activity, which is now assuming a digital and public dimension via social networking tools such as Facebook and Twitter. Pervasive and wearable computing devices (e.g., wireless sensors and smart phones) are further enriching the scenario by making it possible to automatically collect, in a sharable, browsable and machine-processable format, the everyday activities of people, communities, objects and places. Soon, it will be possible to produce, merge and browse expressive digital diaries by everything and about everything, i.e., all-about diaries. Beside considerations related to the possible sociological impacts of their diffusion, use, and misuse (we leave these to appropriate experts), all-about diaries undoubtedly have the potential for opening several opportunities both for consumers and industries, paving the way for the development of innovative services and for the creation of novel markets. Indeed, the road towards the realization of all-about diaries is already paved and strongly pushed by such opportunities. Yet, although some basic technologies to implement the vision are already available and several proposals exist heading towards the all-about diary vision, several technological and architectural challenges still have be addressed to turn the vision into reality.

Our work is a step towards making such a vision into reality. The contributions of this paper lie in two aspects:

1. **Concepts.** In the first part of the paper we review opportunities and challenges of "all-about" diaries to help understanding such emerging systems, contributing to their shaping and - if needed - defending from them. Moreover, we present our vision and conceptual architecture of future "all-about" digital diaries. In particular, we describe the key modules and components comprising such an architecture.

2. **Experiences.** In the second part of the paper we present a first prototypical example of an "all-about" diary focusing on location-based data. In particular, the proposed diary service records location logs and applies data mining techniques to extract patterns and routine behaviors from the user whereabouts. Eventually, an external smart phone app connects to the diary to visualize the resulting life patterns. The goals of this diary prototype are twofold. We first want to test the efficacy of such an architecture and then gain a preliminary idea of its implementation.

The remainder of this paper is organized as follows. Section 2 illustrates our vision for "all-about" digital diaries. Section 3 describes the opportunities and markets supported by such a vision. Section 4 describes existing technologies and early-implementations of diary applications. Section 5 presents an architecture to frame the key components of future "all-about" diaries. Section 6 introduces a case study to show a preliminary prototype of such conceptual architecture focusing on location-based data. Section 7 presents related works in the area. Finally, Section 8 concludes and presents some future work.

COMSWARE 2011, July 04-07, Verona, Italy

DOI 10.4108/comsware.2011.4

2. THE VISION

Imagine. After a day of work Marco wears its novel GPS-enabled training shoes and goes for an hour of jogging. After 20 minutes it starts raining and he hurries home. The shoes automatically upload his jogging performances to his diary. The "sport tracking" app used by Marco fetches such an information from the diary, it realizes he has ran for a shorter time than average, but assumes this is due to the rain and not accounts this on Marco's jogging statistics. Now it's time for a drink outside. The "people like me" app that Marco downloaded, aggregating data from the diaries of people with matching characteristics, shows that "The Fox" pub will be probably crowded by interesting people in about an hour. Zooming-in the diary of "The Fox", it also appears that many of the usual customers are acquainted to Marco. So, he decides to go. While looking on his phone for a friend to call, the "friend tracking" app connected to Marco's diary reports that the social link between him and Franco is fading away: a graph with the number of recent calls between Marco and Franco shows a clear trend that Marco didn't noticed so far! He decides to ask Franco to join him for the night. At 9 p.m. Marco is at "The Fox" with Franco. Marco's phone buzzes: his "friend tracking" app, supported by proximity sensors, has recognized that a woman two tables away usually jogs in the same park as him. While Marco approaches her, the wall-mounted display close to them recognizes their common interest (as from their respective diaries) and starts advertising about training shoes. After a while, Marco and the woman agree on sharing their diaries, so as to meet for jogging next time. In this little story, we have synthesized many characteristics of the envisioned all-about diary, i.e., a software tool that:

1. Collects data from different sensing devices (e.g., GPS, body-worn accelerometers and environmental sensors), software services (e.g., geocoders and weather forecast), and user input.

2. Executes pattern-analysis algorithms to extract high-level information from the acquired data (e.g., accelerometer data turns into "jogging"), and exploit commonsense to extract coherent stories out of the available knowledge (e.g., "you stopped jogging earlier because of rain").

3. Compiles an overall diary/log with the obtained information. Diaries can be created for different entities (people, objects, places and communities), and these can also be aggregated towards social, statistical, and commercial purposes.

4. Provides views on its data to several smart phone apps that can offer innovative services.

From the technology viewpoint, the all-about diary vision will be enabled by the increasing diffusion of pervasive sensing devices embedded in places and in objects possibly worn by people. In fact, future infrastructures and the emerging *internet of things* will make immense data streams coming from thousands of sensors a reality. From the users' viewpoint, the future acceptance of such a tool is very likely, at least by the digital native generation (think at how young people eagerly share personal information in social networking sites). From the social and political viewpoint, we are aware that the implications of all-about diaries - sort of Big Brothers- can be disruptive. Whatever, the road towards all-about diaries is already paved, since solid economic opportunities exist for pushing their realization forward.

3. OPPORTUNITIES

The construction and diffusion of all-about diaries offers several opportunities to those actors interested in contributing to their shaping (i.e., consumer electronic manufactures, telecoms and Internet service providers, and application providers) and to provide, via them, a variety of novel services (Figure 1).

3.1 Consumer Electronics Manufacturer

Consumer electronics manufacturers, producing peripherals like sensors, networking tools, and visualization devices, can create add-on tools for the diary (to be possibly integrated into everyday objects - as the training shoes in our little story). Even sensors and devices of small intrinsic value, when integrated in the diary, can acquire a high extrinsic value, and can transfer a high value to the diary too. Similarly to what happens in the profitable iPod ecosystem [13] (where the vast number of add-on devices and applications overall bring a notable added value to the music player), the all-about diary can create a similar ecosystem with regard to sensing devices. For example, consider a Galvanic Skin Response sensor, which is a simple indicator of the emotional arousal. This sensor might not be useful and commercially appealing on its own. However, if embedded in an object of everyday use (e.g., a wristwatch) and integrated in the diary, the sensor could report information about your emotions and stress level possibly associated to your location and activity (*http://www.biomapping.net*), which can add expressiveness to the diary and can help the interpretation of existing data. As from the first column of Figure 1, various classes of service opportunities can thus be opened for this class of actors.

3.2 Telecoms and Internet Service Providers

Telecoms and ISPs provide the infrastructure and the data centers to run distributed services. They can participate in the diary by offering the computational/networking infrastructure to collect and organize data and provide access to (and run) all-about diary services. They would have several advantages in contributing to the diary as it could attract users or re-enforce their loyalty. Moreover, hosting and contributing to diaries implies having access to large and finer-grained users' data. With this data, they can: *(i)* create further possibilities to engage "three-party market" systems - a third party (e.g., advertisement) pays to participate in a market created by a free exchange between other two parties (Telecoms/ISPs and diary users) - and *(ii)* further engage "cross-subsidies" policies - where the cost of offering free user services is repaid by the selling of commercial services enabled by the all-about diary. The second column of Figure 1 sketches the main service opportunities for Telecoms and ISPs.

		Who can be interested in contributing to the diary (and making money)		
		Consumer electronics manufacturer	Telecom/ISPs providers	Application producers
What kind of services can be provided	Personal diary services	– Production of novel sensors and gadgets to detect user activities. – Use for integrate and enrich the personal diary in an automatic way, making it more expressive. This is useful also for objects' and places' diaries	– Hosting of diaries and provisioning of data analysis tools, summaries, trends, and notifications of events. – Use to effective self-building, storage and access to personal diary and basic diary services.	– Production and release specific applications for peculiar analysis of user data. – Use to personalize diaries with specific tools and annotation (e.g., for data visualization and extraction).
	Social network services	– Production of novel sensors and gadgets to detect proximity and social/collective activities. – Use to connect and socialize with friends, mates, colleagues, and also familiar strangers around.	– Collection and production detailed view on social activities and interactions. – Use to facilitate socialization, for the self-construction of global social diaries or special interest groups diaries, as well as for social sciences.	– Production and release of applications for special interest groups, to become reference for affiliation of specific groups and to be of use to specific groups. – Use by several classes of social communities for social matchmaking.
	Commercial services	– Production of sensors to detect activities and interactions with objects/places. – Use to deliver detailed and contextualized information on products and places usage.	– Production and selling of aggregated data on users' and social groups' behaviors for market research. – Use to deliver data on commercial and social trends, so as to enable targeted personalized advertisement and marketing analysis.	– Production and release of applications tailored for specific objects and places, or to the need of specific commercial or industrial sectors. – Use by specific industrial and commercial actors (e.g., for product recommendation, logistic and transport and planning).

Figure 1: Opportunities opened by the all-about diary vision

3.3 Application Producers

Application producers can be companies developing software, or skilled users creating applications to be distributed via suitable application stores (as in Apple store and Android market). These actors might contribute to the diary to use the special-purpose applications they create, possibly for selling them, but also for engaging in "three-party" markets and "cross-subsidies" policies, as already described. Application producers can create services using and taking advantage of the information contained in the diary to enact personalization and context-awareness (e.g., a messaging application that self-configure on the basis of the data in the diary). Moreover, application producers can create plug-ins for the diary to enrich it with tools for data analysis, aggregation and visualization. This can create an ecosystem of services and applications (at the personal, social, or commercial level, as from the third column of Figure 1) whose value, again, goes beyond the mere intrinsic value of each service/application.

In addition, diary-based systems are likely to have a profound and disruptive impact on several specific (but also extremely important) markets and scenarios. For example, the recent book "The Decision Tree" discusses the implications of diary-based systems on general public health. Chronic diseases (such as heart diseases, cancer, stroke) are the most important causes of mortality in western countries. They are sometimes referred as "diseases of life style" in that they are strongly affected by our everyday activities (e.g., dietary habits). Applications to collect and share in a reliable way information about such kind of daily habits would provide invaluable data on which to build new public health policies to prevent chronic diseases. Diary-based applications have also the power of completely revolutionizing habits and practices in the work environment. As discussed in [30] diary-based systems can automatically infer the social network and roles within an organization (e.g., Who talk with whom? Who is the expert in a give topic? Who is actively participating in meetings?). In most organizations, such kind of critical questions are answered only though episodic memories. Instead, an automatic and reliable log of what is happening within an organization, could be extremely useful for the management. These special purpose diary applications, together with the applications described in Fig. 1 further motivate the diary vision.

4. EXISTING TECHNOLOGIES

Starting from Vannevar Bush's classic paper "As We May Think" and from some pioneering approaches [27, 17], the interest in diary-based services is testified by the increasing number of applications moving the first steps in the towards them. Figure 2 shows the architecture of most of the available systems, which typically:

- collect data from few pre-specified sensors typically embedded in a smart phone (e.g., GPS), and/or by accepting explicit user inputs.
- provide some limited security and privacy features in the form of access control lists for sharing collected data.
- provide a Web front end to visualize collected data, and a simple API to forward data to third-party sites so as to use it as blog entries or mash-ups.

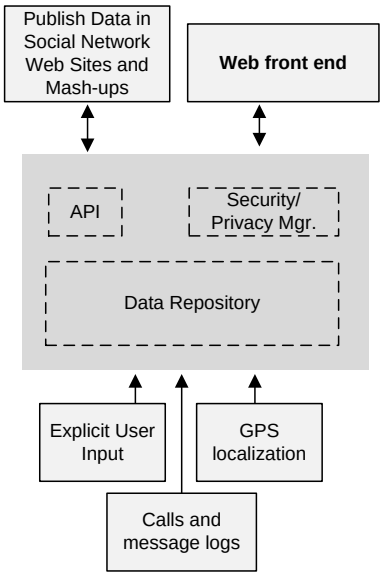

Figure 2: Present diary architectures.

For these systems, applications exploiting the API typically access to raw data and further data analysis - when needed - is embedded within specific applications. In this section we overview only diary-oriented systems that are already available to a wide audience, while we describe pioneering research approaches in Section 7.

Location-based diaries. Google Latitude (*http://www.google.com/latitude*) is a paradigmatic example of this architecture (see Figure 3-left). The application collects location data from user phones and uploads it to a server where information can be accessed and shared among selected friends, based on simple access control lists. Yahoo Fire Eagle/Friends on Fire (*http://fireeagle.yahoo.net*) is a similar proposal by Yahoo to share location data on Facebook. Although location will be a basic component for the all-about diary vision, these services are so far very limited in scope, dealing only with location data and being limited at presenting location data and providing APIs to access it.

Web self-tracker based on direct user input. Some Web sites (often accessible via mobile phones) allow users to keep a journal of (some aspects of) their life. These sites, called Web self-tracker, allow users' to enter information about specific activities/issues, in order to foster self-thinking of oneself life. Sites exist to monitor oneself diet (*http://www.fitday.com*), trips and whereabouts (*http://brightkite.com*), driving cars routine (*http://mymilemarker.com*), daily habits (*http://feltron.com*, see Figure 3-right), or even sex life (*http://www.bedposted.com*). The site (*http://www.quantifiedself.com*) collects several proposals in this direction. These sites, unlike traditional blogs and Web tools, allow to conduct some limited analysis on the data being inserted, to create statistics and summaries. Still, they mostly miss integrating with sensors and portable devices (i.e., to automatically fill-in diaries) and are not able to extract expressive and meaningful information from available data.

Mobile phone loggers. By exploiting the increasing versatility of smart phones, some applications are being proposed to log and organize the data generated by phones and the activities being performed with them. The Nokia Life Blog application (*http://www.nokia.com/lifeblog*) automatically collects and organizes text messages, notes and pictures taken from the phone in a diary-like service, and to possibly post them on an external blog. This could be complemented by Nokia viNe (*http://www.nseries.com/nseries/nokiavine*) that, similar to Google Latitude, could enrich the data with location information. Similar objectives characterize Context Watcher (*http://portals.telin.nl/context watcher/*), which adds the possibility to connect with external sensors. Again, and despite more closely approaching our all-about diary vision, these proposals still lack the automatic extraction of expressive information and the flexible and dynamic integration of data sources.

5. FUTURE WEB DIARIES

The conceptual architecture of Figure 4 frames the key components and functionalities of future all-about diaries. While currently available applications are developed in silos (i.e., a single or few sensors are logged by a single Web application that also presents the collected results, e.g., Google Latitude), future all-about diaries will be open platforms to develop a wide range of applications. Diaries will be fed by several pervasive and environmental sensors, user inputs and Web resources. They will aggregate and extract patterns from this data, and provide the results to external apps via open standards. Several smart phone apps will access the diary to create specific functionalities like those illustrated in the little story in Section 2. In the following we better overview the key components of the envisioned diary platform.

5.1 Plug-and-play Component Architecture

The diary should be structured as a flexible architecture able to host sensors and software modules in a plug-and-play way. Should the user buy a new sensor or a new object embedding some sensors, the diary should be able to automatically incorporate the new devices and the information. In addition, the diary should be robust to unexpected situations like user switched-off devices and sensors unavailability (bottom layer of Fig. 4). Several research proposals in diary-like systems are facing the issue of integrating in a flexible and uniform way diverse data sources. Microsoft MyLifeBits [1, 17] aims at creating a user diary accessing and indexing diverse information sources ranging from wearable sensors to continuous audio and video streams. The Betelgeuse project [23] goes further in facilitating the integration of new sensors by employing a microkernel architecture allowing components and sensors to be plugged-in dynamically. To design this kind of platform, it is also worth noticing that user perception plays a fundamental role. The recent failure of the Google Wave platform, trying to integrate users' Web experience in a unified architecture, shows that the idea of a coherent and unified platform may clash with the user need of well-defined, simple and self-contained services.

5.2 Inference Modules

Diaries have to extract high-level information from raw sensor readings. As examples: from "location" to "place" (e.g.,

Figure 3: Two representative examples of current diary-oriented systems. (left) Google Latitude recording user whereabouts. (right) *http://feltron.com* **recording user daily habits**

from GPS coordinates to "Parents' House"; from "date and time" to "event" (e.g., from "Dec 25, 8.20pm" to "Xmas Dinner"); from "sensing" to "activity" (e.g., from "accelerometer data" to "jogging"). Without these fundamental mechanisms, it is impossible to effectively exploit the acquired data [31]. Accordingly, the central layer of Fig. 4 is devoted to hosting these modules.

Several researches aim at analyzing data from pervasive devices and at extracting high-level information about users' activities and life. These include: *(i)* proposals to analyze GPS traces and extract a diary of the list of the places visited by a user [16]. *(ii)* Techniques to effectively recognize user activities from body-worn accelerometers [3]. *(iii)* Activity recognition systems fed by data coming from embedded cameras (e.g., SenseCam [28]). *(iv)* Data mining techniques to infer the activities of a person based on his/her interactions with RFID-tagged objects [6]. *(v)* Data mining techniques that, based on proximity sensors (e.g., Bluetooth) and/or on phones logs, can infer the social relationships between people (e.g., friend vs. colleague) [14]. In the specific area of diary-like research proposals, AniDiary [19] uses Bayesian networks to summarize mobile phone data into a cartoon-style diary of daily life. The Object-Blog System [25] - whose peculiarity, relevant for the all-about diary vision, is to focus on diaries of (sensor-enriched) objects - adopts inference modules to generate understandable, user-friendly reports about events associated to objects' usage.

5.3 Diaries Aggregators

Several services can be realized by extracting aggregated information from a collection of diaries. Aggregation enables to infer additional information than that inferred by the specific inference modules (e.g., it might be possible to infer the average clientele of a bar, by aggregating all the customers diaries). Aggregators are hosted in the top layer of the architecture in Fig. 4.

A number of techniques have been developed on the Web to aggregate together information produced by multiple users (e.g., collaborative filtering and recommendation systems). In particular, some recent projects (*http://www.citysense.com*) started dealing with these issues in the context of diary-like applications, with the goal of extracting features from a collection of mobility patterns [20].

5.4 Commonsense Reasoner

A diary based on the outcomes of several inference modules would represent the user life in terms of rather spotty set of events ("you ran 20 minutes", "it rains"). Accordingly, a module involving the recognition of complex situations from the individual activities would be useful. With this regard, commonsense reasoner systems (e.g., *http://www.cyc.com*) combine a large-scale commonsense database and reasoning techniques to create stories ("your ran only 20 minutes because it started raining") out of individual facts. Indeed, preliminary attempts to integrate commonsense reasoning in diaries exist [19].

As illustrated in Fig. 4, commonsense reasoning is composed of cross cutting elements that can impact on the diary at multiple levels. At the level of diaries aggregation, there could be benefit in finding way to exploit commonsense reasoning to build sound collective stories out of a set of facts pertaining to different diaries (e.g., service aggregating diaries of people can infer by using commonsense reasoning what are the "trendy" bars and what, instead, are frequented by families).

5.5 Security and Privacy Manager

Security and privacy are of primary importance for diary-like systems. Security implies preventing unregulated access to diaries and proliferation of mock-up data biasing collective analysis. Privacy implies that users (and secondary stakeholders - i.e., people observed by the someone else diary) can control what information to share and with whom. While security issues rely on rather assessed techniques (e.g., access control lists and Captcha-based solutions - controlling un-regulated diary proliferation), privacy-related issues are much harder. Current solutions (mainly relying on simple data anonymization) do not fully account for the many complex issues that can possibly arise. Since data managed by the diary assumes multiple meanings as inference and aggregation operations are applied, different privacy considerations apply at different levels of the diary architecture (see Fig. 4). For instance, while sharing blurred GPS coordinate

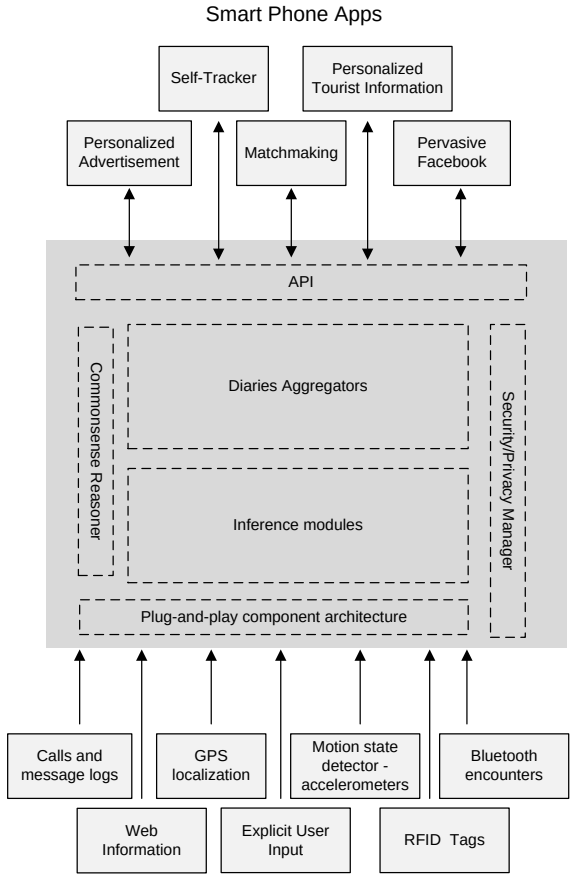

Figure 4: Future all-about diary architectures.

may be fine ("Marco was near latX,lonY at 10.00am") sharing the inferred place may be not (Marco was at "The Fox" pub at 10.00am). Furthermore, even when some inferred information can be shared (Marco was at "Mr. Brown" at 9.30pm), some aggregated information should be not (Marco was at "Mr. Brown" at 9.30pm dancing with Laura). Vice versa, it is possible to think at situations in which raw data should not be shared (e.g., a biosensor capable of identifying health problems) while the knowledge inferred from it could be (e.g., if the same biosensor is used only to infer happiness levels). As a paradigmatic example of the above subtleties and complexity, the DARPA LifeLog project, striving to the all-about diary vision, was cancelled in 2004 because of privacy concerns.

6. A SIMPLE PROTOTYPE OF A WHEREABOUTS DIARY

To experiment with the above ideas, we have built a preliminary prototype of the diary architecture focusing on location-based data. Localization devices and location-based services constitute one of the first pervasive technology going mainstream, so it is rather natural to ground diary prototyping on this data. In particular, the proposed diary service records location logs and applies data mining techniques to extract patterns and routine behaviors from the user whereabouts. Eventually, an external smart phone app connects to the diary to visualize the resulting life patterns. The goals of this diary prototype are twofold. First, we wanted to code an application that might benefit from our vision of all-about diaries to test the efficacy of the architecture. Second, we wanted to gain a preliminary idea of its implementation.

6.1 Accessing Data Sources

As a preliminary implementation, we based our prototype on a single source of data, in particular on location data. In this prototype we adopted Google Latitude (see Fig. 3-left) as an optimal way to acquire high-fidelity location data. We chose such application because it solves most of the data acquisition issues. On the one hand, it supports multi-modal localization automatically switching between GPS, WiFi and GSM localization based on their availability. Thus, the unavailability of a given sensor (e.g., GPS) does not preclude the functioning of the application. On the other hand, it is designed to use minimal battery power (e.g., by decreasing how often user location is updated when the phone's battery is low or when user location has not changed recently). Google Latitude is robust to service interruptions (e.g., the user switching off the phone) and can flexibly restart data acquisition when the device restarts. This allows collecting user location continuously in background in an effective way.

From the perspective of our architecture, we developed a component accessing Google Latitude site and fetching the location history of the users. The architecture makes fetched data available to other services to process it. For the sake of simplicity and flexibility, the architecture design is based on the idea of a tuple space [29]. The component accessing data from Google Latitude creates and inserts tuples associated to the past user locations. The tuple space acts as a shared infrastructure via which other components can access data.

6.2 Inference Modules

In our prototype we developed three inference modules. These components are connected to the diary architecture to exchange data with each other and to access the acquired data sources. The tuple space design flexibly enables such kind of interactions in that it completely decouples the components that access the shared space.

6.2.1 MODULE 1: Place Identification

The first step to process data is to identify the places most visited by the user. Mainstream approaches are either based on segmenting and clustering GPS-traces to infer what are the places relevant to the user [2], or on detecting places and mobility on the basis of nearby RF-beacons such as WiFi and GSM towers [14].

Since we use continuous GPS tracks, we adopted the first kind of approach. More in detail, we created a grid with cells of 500 meters square side over the area visited by the user. Cells in which on average the user spends more than half an hour per week are marked as relevant. Although this kind of place identification introduces a fixed half an hour threshold, it allows to extend the automatic check-in function recently introduced by Google Latitude, in that it does not require that the user explicitly checks-in a place the first time.

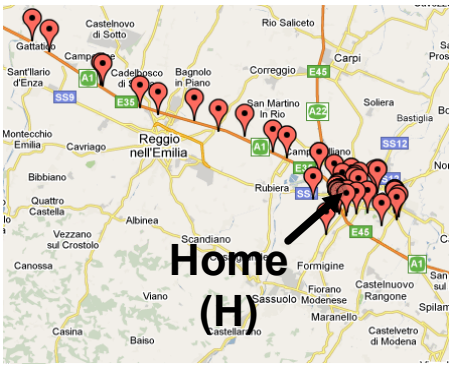

Figure 5: Relevant places are identified by clustering GPS traces. Places are annotated either by user input or by reverse geo-coding.

Once places have been identified, it is important to give a name to the discovered place. Asking labels to the user is a practical way to get concise and understandable names to label the places (this is basically the approach used by Foursquare). However, it could be also possible to reverse geocode a given location to discover what is in there, and use that information for labeling. This latter approach tends to produce noisier labels (due to GPS errors), but would be completely unobtrusive to the user. We adopted the first mechanism for data we collected with Google Latitude. In addition, we automatically name Home "H" the place most visited at night and Work "W" the place most visited during the day (see Fig. 5).

Information about the discovered places can be injected in the tuple space to serve as inputs for the following components.

6.2.2 MODULE 2: Bag of Words Creation

Following the approach proposed in [15, 14], a second component organizes the dataset into a sequence of days each consisting of 48 *time-slots* lasting 30 minutes each. For each *time-slot*, if the associated GPS records fall within 500 meters from a place identified in the previous step, then the algorithm marks the *time-slot* with that place. In addition we defined the following general labels: no reception(N), near elsewhere(NE) and far elsewhere(FE). "N" is a label used if there is missing data for a person for a given time, for instance when the phone is off. "NE" and "FE" are labels used to indicate any other location respectively within or out of 30 kilometers from the home of the given user. Following this process, each day is then represented by a string of 48 symbols. To capture transitions between locations, in the final step of the pre-processing phase, we run a sliding window over each day. The sliding window takes 3 consecutive symbols and concatenates them with another label capturing the time of the day where these locations have been visited. In particular, we considered the following 8 time labels: 0-3am(1), 3-6am(2), 6-9am(3), 9am-12pm(4), 12-3pm(5), 3-6pm(6), 6-9pm(7), 9pm-12am(8). These time segments were chosen to capture common events in daily life, such as lunch time, dinner time, or morning and afternoon work times. The result is a set of words each containing 3 location letters and a time label. For example if the day of the user starts with the following symbols "HHHHWWWWW", we will obtain the following words: "HHH1", "HHH1", "HHW1", "HWW1", "WWW1", "WWW1", "WWW2", etc.

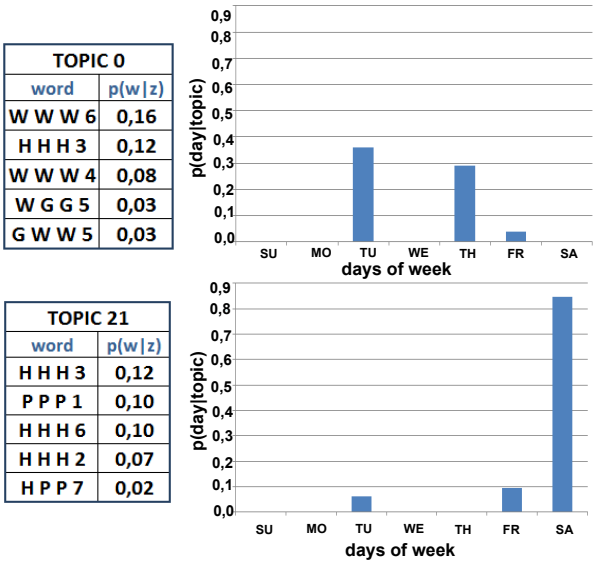

Figure 6: Routines extracted from LDA analysis. Each routine (i.e., topic) is represented by a set of places, a time frame in which the places are visited and the probability of that happening. Routines are probabilistically associated to specific days of the week. Visited places are described by means of the label H = Home, W = Work, G = Gym, P = Pub, NE = Near Elesewhere (a new place close to where the user lives)

The resulting *bag of words* summarizes the original dataset and is the input data structure for the algorithm to extract routine behaviors (i.e., topics). The bag of words representation is also compatible with the proposed tuple space architecture. Words can be encoded as tuples to be shared with the following inference modules.

6.2.3 MODULE 3: Discovering Routine Behaviors

While identifying relevant places is an important first step to create a diary of user whereabouts, it would be even more meaningful to extract the patterns with which such places are visited. This would allow to describe the day of the user in terms of routines such as "pub after work", "going to the gym during lunch break" routine. Discovering such routine behaviors is an important step to add semantics to users' whereabouts. On the one hand, patterns and routines represent a step further in describing information about the user. On the other hand, they represent also information about how a place is "used" by a given user. So that the "Fox pub" can be a place where to go after work for some users while it is the work place itself for the bar tender.

LDA is a probabilistic generative model [4] used to cluster documents according to the topics (i.e., word patterns) they

contain. LDA has two main characteristics that make it suitable to our pattern discovery task. On the one hand, it is an unsupervised approach: it does not require to define classes (i.e. topics) *a priori* and it does not require difficult-to-be-acquired labeled data. On the other hand, topics represent meaningful probabilistic distributions over words and documents. This allows to analyze and understand the routine behavior they stand for.

In LDA, a word w is the basic unit of data, representing in our case a user location at a given *time-label*. A set of N words defines a day of the user (i.e. a document). Each user has a dataset consisting of M documents. Each day is viewed as a mixture of topics z, where topics are distributions over words (i.e., each topic can be represented by the list of words associated to the probability $p(w|z)$). For each day i, the probability of a word w_{ij} is given by $p(w_{ij}) = \sum_{t=1}^{T} p(w_{ij}|z_{it})p(z_{it})$, where T is the number of topics. $p(w_{ij}|z_{it})$ and $p(z_{it})$ are assumed to have multinomial distributions.

Once the model is trained, Bayesian inference allows to identify the words comprising each topic $p(w_{ij}|z_{it})$ and to extract the topics best describing the routines of a given day $p(z_{it})$.

In order to test the effectiveness of the proposed prototype, we recorded the daily whereabouts of three persons over the period of almost one year. One subject is the one of the authors of this paper, the other two subjects are not part of our research group. Subjects recorded data by continuously running Google Latitude in background.

In Fig. 6-left, we illustrate two exemplary topics extracted from one of the user of the Google Latitude dataset (*topic 0* and *topic 21*). In particular, we present each topic by listing the top words (ranked by $p(w|z)$).

- Topic 0 captures the *home-work-gym-work* routine. The most probable words for such topic are WWW6 and HHH3, which are respectively being at work in *time-slot* 6 (3-6pm) and being at home in *time-slot* 3 (6-9am). They are followed by working in *time-slot* 4 (9am-12pm) and going from work to gym and then back to work in *time-slot* 5 (12-3pm). From the distribution of the routine over the days of the week we can see that it corresponds to a weekdays trend (Tuesdays and Thursdays in particular).
- Topic 21 captures the *pub-home-pub* routine. The corresponding top words illustrate that the user is in a pub from 0am to 3am (PPP1), then he remains at home during the day (HHH2, HHH3, HHH6) and finally he moves from home to the same pub in *time-slot* 7 (6-9pm). From the distribution of the routine over the days of the week we can see that it corresponds to a weekends trend.

This result illustrates that the LDA model applied to GPS data successfully reveals different types of patterns. In addition, based on such method, we are able to answer to several interesting questions such as "Are there specific patterns occurring on weekends versus weekdays?" or "How do the topics characterize the days in the dataset?".

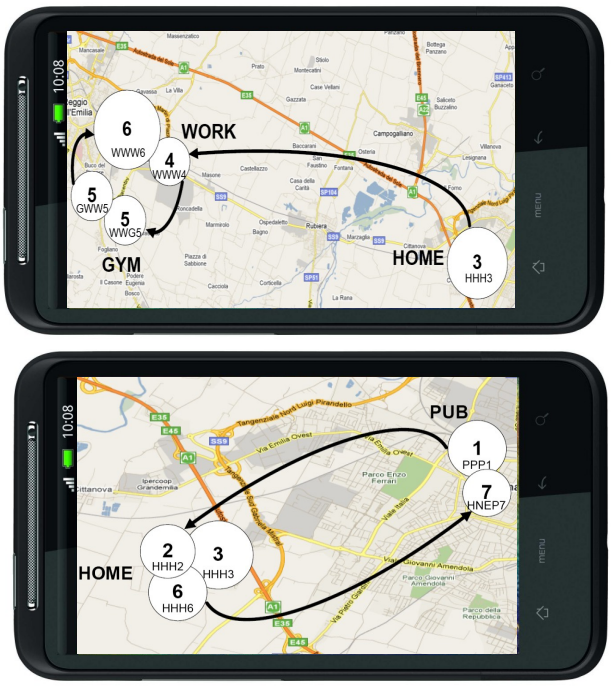

Figure 7: Graphic representation of two discovered routines: spatial distribution of the routine displayed on a map. The larger the circle, the higher is the probability associated to the given place. Time labels: 0-3am(1), 3-6am(2), 6-9am(3), 9am-12pm(4), 12-3pm(5), 3-6pm(6), 6-9pm(7), 9pm-12am(8). Results can be easily integrated in social network sites to create an automatic log of the user whereabouts)

One of the strengths of the proposed diary approach is that the proposed inference module can be plugged into the diary architecture seamlessly. They are just processes accessing the tuple space in a completely decoupled way.

Finally, the diary makes available the extracted topics to third party applications that could take context-aware decisions on the basis of users' routine behaviors.

6.3 External App

The resulting topics can be used in a number of applications: location-based services and location-based information retrieval could provide resources that are relevant and practically accessible given the user's routine whereabouts. Life-log applications could readily use the extracted patterns to automatically create an entry in the user blog. Advertisement applications could take into account user's typical whereabouts and recommend businesses and shops in her typical route [22]. Social network applications could take advantage of user's routines to recommend friends, events and happenings that the user might like. As an example, Fig. 7 presents a simple life-logging visualization where users can see their actual and past routine behaviors. A map illustrating the user whereabouts (LDA topics) is displayed. The map has circles associated to the top words,

and arcs connecting the places being visited. Circle diameter is proportional to the probability of the associated word, the number in the circle is the *time-label* of the word. This kind of visualization can be easily integrated in social network sites (such as Facebook) to create an automatic log of what the user is up to. This result shows the potential of diary applications. Of course, the integration with information sources other than mobility will further strength diary effectiveness.

7. RELATED WORK

In this section we review the state of the art in all-about diary and life-logger research systems. Beyond the existing technologies already available shown in Section 4, there are lots of pioneering research approaches that go in the directions depicted in Section 5 and Section 6. In particular, we focus on those systems that go beyond the idea of just collecting past experiences, but add mechanisms to infer high-level information from captured data.

In the first part we review research approaches related to the first goal of this paper, i.e., the creation of a flexible architecture for future "all-about" diaries. In the second part we review research approaches related to our prototype of a whereabouts diary presented in Section 6. In particular, we focus on the issue of automatically discovering people's whereabouts.

7.1 "All-about Diaries" Approaches

The mobile applications Context Watcher (*http://portals.telin.nl/contextwatcher/*) and IYOUIT [5] are designed to run on smart phones, connect to a predefined set of sensors, log information, and generate daily summaries about user's location and activities. Moreover, they allow to post retrieved information to third-party blog sites. These applications are very close in spirit to our idea of the all about diary. However, the use of inference techniques is still limited. To overcome this limitation, CenceMe [26] is a people-centric sensing application. It exploits sensor-enabled mobile phones to automatically infer people's sensing presence (e.g., dancing at a party with friends) and then shares this presence through social network portals such as Facebook. The already introduced Betelgeuse platform [23] is a similar proposal in this direction.

The Mobile Sensing Platform (MSP)[9, 21] is a small wearable device designed for embedded activity recognition with the aim of broadly supporting context-aware ubiquitous computing applications. The platform integrates data collection and inference modules to capture high-level information about the user's daily activities.

The above platforms are at the state of the art in this kind of diary and life-logging systems. However, they are almost developed in silos: rather than creating open platforms to combine sensors and software components together, they specialize in a single application running on smart-phones. Future all-about diaries will have to be open platform capturing data from social network sites, Web resources, environmental sensors other than the user's smart phone.

In [10, 24] and [28] video and audio feeds obtained from wearable cameras and microphones are analyzed to infer what the user is doing. These works are similar to the ones discussed above, but they deal with feature-rich multimedia sensors. Again, although video and audio feeds are of primarily importance for diary-based systems, these approaches ar built in silos an not on open extensible platforms to be possibly shared among many applications.

AniDiary [8, 19] is a research application aiming at summarizing user's daily life with a cartoon-style diary based on information collected from mobile devices such as smart phones. AniDiary uses modular Bayesian networks to detect and visualize landmarks (relevant or novel events) and transform numerous logs into user-friendly cartoon images. This work is similar in spirit to the above ones, but raises the important point of data presentation/visualization: it is important to identify means to effectively render diary data to users so that to let them understand and take actions. Given the multi-faceted nature of the data collected by the diary this is an open research issue [11].

The works [2, 34] address the problem of constructing a diary of user whereabouts and run inference operations on the acquired data. As already discusses, whereabouts information are one of the primary data to be acquired by diary-systems. Novel mechanisms to process effectively location-based data will be one of the main workhorses of future diaries. In addition, while several research to extract patterns from mobility data do exist, one problem related to all the above researches is that they are often conducted as stand-alone data-mining exercises, and seldom exploited in a synergetic way. Instead, the real challenge is to integrate several techniques and have them cooperate

The Affective Diary [33] is a wearable computing platform consisting of a mobile phone, body sensors, and a Tablet PC. During the day, the sensor armband collects sensor data indicating movement and arousal levels. This allows to collect a diary comprising also some psychological characteristics of the user. A similar system, aggregating individual diaries of this kind, allow to create an emotion map of the city (*http://www.emotionmap.net*). This kind of advanced sensing system is very interesting in that it highlight the vast range of modalities to capture different aspects of the user life.

The above systems are some of those specifically addressing the vision of a diary. However, several other approaches, as already introduced in Section 5, could be easily casted to diary systems.

7.2 "Discovering Whereabouts" Approaches

Several researches extract and identify those places that matter to the user. Mainstream approaches are either based on segmenting and clustering GPS-traces to infer what are the places relevant to the user [2], or on detecting places and mobility on the basis of nearby RF-beacons such as WiFi and GSM towers [14]. Our approach tries to go further by extracting routine behaviors other than relevant places.

The CitySense project (*http://www.citysense.com*) uses GPS and WiFi data to summarize hotspots of activity in the city area, which can then be used to make recommendation to people regarding, e.g. popular bars and clubs. In a similar

work based on extremely large anonymized mobility data coming from Telecom operators authors were able to extract the spatio-temporal dynamics of the city, highlighting where people usually go during the day. Authors were able also to identify the most visited areas by tourists during the day and the typical time of the visit (see for example [7, 18]). Eagle and Pentland [14], use Principal Component Analysis (PCA) to identify the main components structuring daily human behavior. The main components of the human activities, which are the top eigenvectors of the PCA decomposition are termed *eigenbehaviors*. Similarly, the work presented in [32] compares different data mining techniques to extract patterns from mobility data. In particular, they found Principal Component Analysis (PCA) and Independent Component Analysis (ICA) best suited to the task of identifying daily patterns. In [15] authors propose the use of probabilistic topic models to capture human routines from cell tower connections. The latter method, which is also the one proposed in this paper, has the advantage of capturing characteristic trends occurring over part of the day (such as lunch time only), whereas eigenbehaviors capture features over the entire day. In comparison with [15] our work uses a more complex dataset, thus allowing to analyze the topic models method at a finer-grain scale with a higher number of places. As above mentioned, the geographic coordinates provided by Google Latitude allows to enrich the location vocabulary with a higher number of places (in contrast with the 'home', 'work' and 'elsewhere' label used in [15]). In addition, Google Latitude provides data at a finer-grain scale.

8. CONCLUSIONS AND FUTURE WORK

All-about diaries promise to be a key component of our future networked society, and a gold mine for a number of useful services and business opportunities. However, although the number of systems and Web sites going in this direction is increasing, the road towards the realization of effective, expressive, and usable diaries still entails addressing several challenges.

Our future work in this area will comprise the design and development of a complete architecture for diary-based systems. Far from being a niche issue of diary systems, the need for architectures capable of dealing with the dynamic integration of heterogeneous components and devices is a general issue of modern networked scenarios. And so are the associated challenges of defining means to release self-describing components capable of self-advertising their services; identifying general-purpose data and meta-data models able to capture the main characteristics of sensors and data sources; finding the right trade-off between the need to adopt common ontologies for data and meta data and the necessity to pragmatically account for emerging ontologies [12]. In all-about diary systems, these issues are exacerbated by the need to account for dynamic component integration not only int terms of sensing and network devices, but also of high-level data sources (user inputs and Web information). Also, diary systems should enable to uniformly deal with information at different granularity levels, ranging from raw atoms of data to high-level information generated by inference engines and aggregators.

From the development point of view, the implementation of all-about diaries and services requires its components to be allocated somewhere. Deploying diaries as stand-alone applications running on personal devices seems not suitable due to the complexity of the overall architecture and because it clashes with the need to aggregate diaries. Fully centralized solutions based on data centers accessible via Web interfaces can be conceived, also in consideration of the progress made in the processing (other than in the storing) of large amounts of data (*http://hadoop.apache.org*). However, since diary data is generated in a decentralized environment and often locally used, and since modern sensors and devices can effectively act as small data centers to store and pre-process local data, intermediate flexible solutions to distribute and/or replicate some components of the diary could be conceived and engineered.

The presented issues to create such a flexible architecture are one of the objectives of the EU-funded project "SAPERE: Self-aware Pervasive Service Ecosystems" (*www.sapere-project.eu*). Furthermore, the integration of several mobile data sources in a flexible architecture, is one of the main objectives of the project "Mr.Typ: Mobile and Real-Time Yellow Pages"(*www.mr-typ.com*), funded by Telecom Italia, the main Italian telecom operator.

Acknowledgments. Work supported by the FP7 project SAPERE (Grant No. 25874) funded by the FET Program of the European Commission and by Telecom Working Capital.

9. REFERENCES

[1] G. Bell and J. Gemmell. *Total Recall: How the E-Memory Revolution Will Change Everything.* Dutton, 2009.

[2] N. Bicocchi, G. Castelli, M. Mamei, A. Rosi, and F. Zambonelli. Supporting location-aware services for mobile users with the whereabouts diary. In *International Conference on MOBILe Wireless MiddleWARE, Operating Systems, and Applications*, Innsbruck, Austria, 2008.

[3] N. Bicocchi, M. Mamei, and F. Zambonelli. Detecting human activities from body worn accelerometers via instance-based algorithms. *Pervasive and Mobile Computing Journal*, 6(4):70–77, 2010.

[4] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(1):993–1022, 2003.

[5] S. Bohm, J. Koolwaaij, M. Luther, B. Souville, M. Wagner, and M. Wibbels. Introducing iyouit. In *International Semantic Web Conference*, Karlsruhe, Germany, 2008.

[6] M. Buettner, R. Prasad, M. Philipose, and D. Wetherall. Recognizing daily activities with rfid-based sensors. In *International Conference on Ubiquitous Computing*, Orlando (FL), USA, 2009.

[7] F. Calabrese, J. Reades, and C. Ratti. Eigenplaces: analysing cities using the space-time structure of the mobile phone network. *IEEE Pervasive Computing*, 9(1):78–84, 2010.

[8] S. Cho, K. Kim, K. Hwang, and I. Song. Anidiary: Daily cartoon-style diary exploits bayesian networks. *IEEE Pervasive Computing*, 6(3):66–75, 2007.

[9] T. Choudhury, S. Consolvo, B. Harrison, J. Hightower,

A. LaMarca, L. LeGrand, A. Rahimi, A. Rea, G. Borriello, B. Hemingway, P. Klasnja, K. Koscher, J. Landay, J. Lester, D. Wyatt, and D. Haehnel. The mobile sensing platform: An embedded system for activity recognition. *IEEE Pervasive Computing*, 7(2):32–41, 2008.

[10] B. Clarkson. *Life Patterns*. PhD Thesis, Massachusetts Institute of Technology, 2003.

[11] A. K. Clear, R. Shannon, T. Holland, S. Dobson, A. Quigley, and P. Nixon. Situvis: a visual tool for modeling a user's behaviour patterns in a pervasive environment. In *International Conference on Pervasive Computing*, Nara, Japan, 2009.

[12] N. Dalvi, R. Kumar, B. Pang, R. Ramakrishnan, A. Tomkins, P. Bohannon, S. Keerthi, and S. Merugu. A web of concepts. In *ACM Symposium on Principles of Database Systems*, Providence (RI), USA, 2009.

[13] D. Darlin. *The iPod Ecosystem*. The New York Times, February 3, 2006.

[14] N. Eagle, A. Pentland, and D. Lazer. Inferring social network structure using mobile phone data. *Proceedings of the National Academy of Sciences (PNAS)*, 106(36):15274–15278, 2009.

[15] K. Farrahi and D. Gatica-Perez. Discovering routines from large-scale human locations using probabilistic topic models. *ACM Transactions on Intelligent Systems and Technology*, 2(1), 2011.

[16] L. Ferrari and M. Mamei. Discovering daily routines from google latitude with topic models. In *IEEE Workshop on Context Modeling and Reasoning*, Seattle (WA) , USA, 2011.

[17] J. Gemmell, G. Bell, and R. Lueder. Mylifebits: a personal database for everything. *Communications of the ACM*, 49(1):88–95, 2006.

[18] F. Girardin, J. Blat, F. Calabrese, F. D. Fiore, and C. Ratti. Digital footprinting: Uncovering tourists with user-generated content. *IEEE Pervasive Computing*, 7(4):36–43, 2008.

[19] K. Hwang and S. Cho. Life log management based on machine learning technique. In *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, Seoul, South Korea, 2008.

[20] R. Ji, X. Xie, H. Yao, and W. Ma. Mining city landmarks from blogs by graph modeling. In *ACM international Conference on Multimedia*, Beijing, China, 2009.

[21] P. Klasnja, B. Harrison, L. LeGrand, A. LaMarca, J. Froehlich, and S. Hudson. Using wearable sensors and real time inference to understand human recall of routine activities. In *International Conference on Ubiquitous Computing*, Seoul, South Korea, 2008.

[22] J. Krumm. Ubiquitous advertising: The killer application for the 21st century. *IEEE Pervasive Computing*, 10(1):66–73, 2011.

[23] J. Kukkonen, E. Lagerspetz, P. Nurmi, and M. Andersson. Betelgeuse: A platform for gathering and processing situational data. *IEEE Pervasive Computing*, 8(2):49–56, 2009.

[24] H. Lu, J. Yang, Z. Lu, N. Lane, T. Choudhury, and A. Campbell. The jigsaw continuous sensing engine for mobile phone applications. In *ACM Conference on Embedded Networked Sensor Systems*, Zurich, Switzerland, 2010.

[25] T. Maekawa, Y. Yanagisawa, Y. Kishino, K. Kamei, Y. Sakurai, and T. Okadome. Object-blog system for environment-generated content. *IEEE Pervasive Computing*, 7(4):20–27, 2008.

[26] E. Miluzzo, N. Lane, K. Fodor, R. Peterson, H. Lu, M. Musolesi, S. Eisenman, X. Zheng, and A. Campbell. Sensing meets mobile social networks: the design, implementation and evaluation of the cenceme application. In *ACM Conference on embedded network sensor systems*, Raleigh (NC), USA, 2008.

[27] W. M. Newman, M. A. Eldridge, and G. M. Lamming. Pepys: Generating autobiographies by automatic tracking. In *European Conference on Computer-Supported Cooperative Work*, Amsterdam, NL, 1991.

[28] D. Nguyen, G. Marcu, G. Hayes, K. Truong, J. Scott, M. Langheinrich, and C. Roduner. Encountering sensecam: Personal recording technologies in everyday life. In *International Conference on Ubiquitous Computing*, Orlando (FL), USA, 2009.

[29] L. Nixon, E. Simperl, R. Krummenacher, and F. MartinRecuerda. Tuplespace-based computing for the semantic web: a survey of the state-of-the-art. *The Knowledge Engineering Review*, 23(2):181–212, 2008.

[30] D. Olguin and A. Pentland. Sensor-based organisational design and engineering. *International Journal of Organisational Design and Engineering*, 1(1/2):69–97, 2010.

[31] A. Sellen and S. Whittaker. Beyond total capture: A constructive critique of lifelogging. *Communications of the ACM*, 53(5):70–77, 2010.

[32] S. Sigg, S. Haseloff, and K. David. An alignment approach for context prediction tasks in ubicomp environments. *IEEE Pervasive Computing*, 9(4):90–97, 2010.

[33] A. Ståhl, K. Höök, M. Svensson, A. S. Taylor, and M. Combetto. Experiencing the affective diary. *Personal and Ubiquitous Computing*, 13(5):365–378, 2011.

[34] Y. Zheng, L. Zhang, Z. Ma, X. Xie, and W. Ma. Recommending friends and locations based on individual location history. *ACM Transaction on the Web*, 5(1), 2011.