This article has been accepted for publication in Computer but has not yet been fully edited.
Some content may change prior to final publication.

Social and Community Intelligence

Daqing Zhang[1], Bin Guo[1], Zhiwen Yu[2]

[1] Institut TELECOM SudParis, France

{daqing.zhang, bin.guo}@it-sudparis.eu

[2] School of Computer Science, Northwestern Polytechnical University, P. R. China
zhiwenyu@nwpu.edu.cn

Abstract

This paper introduces an emerging research area – Social and Community Intelligence (SCI). It aims at revealing the individual/group behaviors, social interactions as well as community dynamics by mining the digital traces left by people while interacting with cyber-physical spaces. The digital traces are generated mainly from three information sources: Internet and Web applications, static infrastructure, mobile devices and wearable sensors. The paper discusses the evolution, general framework, major applications, and research issues of Social and Community Intelligence.

Keywords: Community Intelligence, Social Computing, Mobile Phones, Sensor Network

1. Introduction

With the phenomenal growth of Internet and social network services, the recent explosion of sensor-equipped (e.g., accelerometer, GPS, Bluetooth, camera, and so on) mobile phones, the prevalence of GPS-equipped cars, taxis, and buses, and the large deployment of sensor network (e.g., Wi-Fi, surveillance cameras) in public facilities, private buildings and outdoor environments, the digital traces left by people while interacting with cyber-physical spaces are accumulating at an unprecedented breadth, depth and scale, those digital traces are also called "**digital footprints**".

Leveraging the capacity to collect and analyze the "digital footprints" at community scale, a new research field called "**social and community intelligence (SCI)**" is emerging that aims at revealing the patterns of individual/group behaviours and community dynamics. The scale and richness of the multimodal, mixed data sources

Digital Object Indentifier 10.1109/MC.2011.65 0018-9162/$26.00 2011 IEEE

present us an opportunity to compile the digital footprints into a comprehensive picture of individual's daily life facets, transform our understanding of our lives, organizations and societies, and enable completely innovative services in areas like human health, public safety, city resource management, environment monitoring, and transportation management. To illustrate the rough idea of SCI and its potential impact on our lives, we can look at a possible use scenario in a university campus as follows:

In university campus, students often face the problem of finding partners to do sports in a certain free time slot, searching if there are spaces available in library or a classroom, etc. When a pandemic like H1N1 occurs, how to quickly identify who has been contacted by a suspect person, when and where the contact takes place is crucial to avoid further spread of the disease. There are also queries like when will the next bus reaches the Bus Stop near Library, how many people are waiting in the bus stop, etc. In today's campus environment, it is still difficult to answer these questions about the individual, about the group interaction and about the society dynamics merely based on the state-of-the-art technologies. However, all those community services in university campus can be enabled by analyzing the pervasive data streams collected from personal mobile phone sensors, GPS from buses, WLAN or Bluetooth gateways inside the building, social relationship from the web, etc. In the case of pandemic, for example, the distance and contact time with the suspect, the logical places for the meeting (e.g., office, bus), the relationship with the suspect (e.g., family, friend, colleague) are all important clues and contexts affecting the probability of disease spread.

Different from the closely related research areas such as social computing [1], reality mining [2], and Internet of things, the unique characteristics of this new SCI area can be embodied in the following aspects:

(1) *Infrastructure*. The scale of the SCI system goes beyond single smart space and reaches the level of a community. Real-life, real-time data collection and inference is a key system feature. An infrastructure is required to integrate large-scale and heterogeneous devices, software, and spaces, and provide systematic support for rapid application development, deployment, and evaluation.

(2) *Data*. The data sources are multi-modal and heterogeneous. The social and community intelligence can be inferred from three main data sources: *the mobile/wearable sensor data about the individual and moving objects*, *the*

Digital Object Indentifier 10.1109/MC.2011.65 0018-9162/$26.00 2011 IEEE

infrastructure-bound sensor data about the environment, and *the social data about the individual's preference and relationship with others from social network and Internet interaction services*. While each data source can independently show one facet of the user's daily life, the combination of the three data sources can reveal unforeseen social and community behaviours.

(3) *Technology*. The core technologies for SCI are data mining, machine learning and AI. And the objective of data processing and inference goes from recognizing the individual's physical activity and environmental context to extracting higher-level community and social behaviours (from talking to meeting; from driving slowly to traffic jam, there exist semantic gaps between individual activities and social/community behaviours).

(4) *Application*. It aims to enable innovative services in society level like community healthcare, public safety, city resource and transportation management.

2. Evolution of SCI Research

For a long time, our understanding of human behaviors, social interactions, and city dynamics has generally relied on data collected via individual observations and surveys, where the observations are usually sparse and the survey results are often incomplete with significant time delay. With the technological advances in computing, storage, Internet, wireless communication, and sensing, the human behaviours, social interactions, and city dynamics can be monitored and analyzed at a large-scale, in nearly real-time. Firstly, *Internet* contents were used as premier data sources for understanding large-scale human interaction. Then recognizing the human activities in physical environment become a reality with the emergence of static sensing infrastructure. Recently, the prevalence of sensor-enriched mobile devices brings forward unprecedented opportunities to observe human behaviour, social interaction, and community dynamics. All the three above-mentioned information sources: *Internet and Web, static infrastructure*, and *mobile devices*, have contributed to the evolution of SCI research.

During the last two decades, we have observed an explosive growth of Internet services such as e-mail, instant messaging, Web, etc., which have changed the way that people share/get information and communicate with each other. Leveraging on those

Digital Object Indentifier 10.1109/MC.2011.65 0018-9162/$26.00 2011 IEEE

services, a large body of work on information retrieval, information extraction, and human interaction analysis springs up, such as news recommendation, person/organization profile extraction, e-mail network analysis, and so on. More recently, as the Internet steps into the era of Web 2.0, researchers turn their attention to the online social utilities, such as social networking sites, wikis, and Blogs. A lot of work has been done on social behavior study and user-generated content analysis. For example, Domingos investigates how to mine social networks to study customer behaviors [3]. Xiang *et al.* develop an unsupervised model to estimate relationship strength from interaction activity and user similarity on a social website [4]. Amit Sheth's research group terms Web 2.0 service users as "citizen sensors" and has done much work on social event detection from user-contributed contents [5]. Twitter, a popular micro-blogging site, has been reported to support real-time mining of natural disasters such as earthquakes [6]. In their seminal paper, Wang *et al.* termed the social study based on Internet and Web as *social computing*, which aims at studying and extracting human social dynamics (e.g., human interaction patterns) from online human interactions [1].

With the prevalence of static sensing infrastructure, such as surveillance cameras, environment sensors, indoor positioning sensors, and RFIDs, monitoring and detecting real world events becomes possible. In the early stage, sensors are mainly used for environment monitoring in significant places. Surveillance camera can be regarded as the first sensing device that is widely deployed in public and critical spots to detect abnormal events. Other kinds of sensors, such as temperature sensors, light sensors, and humidity sensors, have also been widely used for environment monitoring (e.g., fire detection in the forest). With the development of sensing techniques, massive cheap and tiny sensors like RFID and switches are deployed to augment our daily living/working environments (i.e., the so-called smart spaces). The Active Bats is an early system that uses ultrasonic sensors and the triangulation location-sensing technique to locate indoor objects [7], which enables location-based services like lost-object finding (e.g., finding a lost key). Philipose *et al.* explore techniques to recognize human activities by analyzing people's interaction with RFID-equipped everyday objects [8]. Static sensing infrastructure brings opportunities to infer environmental and human contexts in smart spaces [9]. However, it is bounded to the sensor-enriched physical environments.

Digital Object Indentifier 10.1109/MC.2011.65 0018-9162/$26.00 2011 IEEE

The defect of static infrastructure is remedied with the presence of wearable sensors, which transform people into "mobile" sensors for both personal and ambient environment monitoring. Wearable sensors, such as accelerometers, pedometers, heart rate sensors, wireless webcams, microphones, are worn on different parts of human body to enable various human-centered services, including human behavior detection, health status monitoring, and social context recognition (e.g., in a meeting, talking with a friend). Although wearable sensors are portable and promising, they are still not viewed as a "personal companion". Things change with the recent prevalence of sensor-enhanced mobile phones, where a number of sensors such as GPS receivers, Bluettoth/WiFi, accelerometers, and cameras are embedded. The huge amount of multi-modal data collected from people's daily use of smart phones opens a new window to study large-scale human behavior patterns and community dynamics. For example, Real Time Rome (http://senseable.mit.edu/realtimerome/), initiated by MIT from 2006, is one of the pioneering projects that explicitly use mobile phone data to understand the dynamics of cities (e.g., movement patterns of people, spatial and social usage of streets and neighborhoods). Reality mining (http://reality.media.mit.edu/), on the other hand, is another effort that collects and analyses mobile phone data (e.g., physical proximity) to identify predictable patterns of social behavior (e.g., friendship). Human-centric sensing, a concept proposed by Dartmouth, explores how to link personal mobile sensing to mobile social networks and public environment monitoring (e.g., air pollution distribution in a city) [10].

Table 1: Definition of SCI and related research areas

Research Area	***Definition***
Social Computing	Computational facilitation of social studies and human interaction analysis as well as the design and use of technologies that consider social context.
Reality Mining	Reality mining is the collection and analysis of mobile sensing data pertaining to human social behavior, with the goal of characterizing human interaction and behavior patterns.
Human-Centric Sensing	Using mobile sensing data to derive people's daily patterns, interactions, and characteristics of public environments.
Urban Computing	Urban computing studies the interaction between humans and environments using technology in public environments such as cities, parks, forests and suburbs.
Social and Community Intelligence	Reveal individual/group behaviors, social interaction as well as community dynamics, leveraging the aggregated power of three information sources: Internet and Web, static infrastructure, mobile devices and wearable sensors.

Digital Object Indentifier 10.1109/MC.2011.65 0018-9162/$26.00 2011 IEEE

The convergence of the above-mentioned research directions leads to the birth of Social and Community Intelligence (SCI) research. SCI aims at *revealing individual/group behaviors, social interaction as well as community dynamics, leveraging the aggregated power of the three information sources: Internet and Web, static infrastructure, mobile devices and wearable sensors*. There are several closely related research areas that are interleaving with SCI as illustrated in Table 1, i.e., social computing, reality mining, human-centric sensing, and urban computing. Different from those areas that generally rely on one of the data sources for information extraction, SCI explores the fusion of the three data sources to infer intelligence at the group and community level, ranging from human interaction, group behaviours within a community, to dynamics of a whole community (e.g., traffic jams, hot spot detection, and significant location).

Compared to SCI, social computing emphasis mainly on the analysis of human interaction and social behaviours using Web data, it does not target at the study of large-scale physical community. Similar to social computing, reality mining also devotes to social interaction analysis, but primarily relies on the data gathered from mobile devices. While urban computing studies the relationship between human and environment at the city scale, SCI extends its scope from urban design to large-scale analysis of personal, group, and community dynamics. Human-centric sensing is the closest research area to SCI, they share the same research goals, but the underlying sensing mechanisms are different. Instead of using merely mobile phones for human-centric sensing, SCI aggregates the information from Internet services, static infrastructure as well as mobile phones.

In summary, SCI shares many things in common with the aforementioned four areas, yet it goes beyond all those areas in terms of scope and data origins. Breakthroughs in any of these four areas will contribute to the further progress of SCI research. In the following sections, we will discuss multi-source data aggregation, architecture support, major application areas, as well as research issues of SCI.

Digital Object Indentifier 10.1109/MC.2011.65 0018-9162/$26.00 2011 IEEE

This article has been accepted for publication in Computer but has not yet been fully edited.
Some content may change prior to final publication.

3. Aggregated Effects of Heterogeneous Data Sources

SCI aims to extract individual/group behaviors, and community dynamics from three important data sources: Internet and Web services, static sensing infrastructure, mobile devices and wearable sensors. The three sources have different attributes and strengths:

- *Internet and Web service* is a major source to extract static or slowly changing information, such as user profile, organization structure, user relationship in a community.
- *Static infrastructure* enables the detection of indoor and urban user activities, group activities, and space context in sensor-enriched environments.
- *Mobile devices and wearable sensors* are always user-centric, thus great at sensing individual activities, interpersonal interactions, significant user locations, and public environment contexts.

Due to the diverse features, aggregation and fusion of data from those three different sources provides unique opportunities to social and community intelligence extraction. Figure 1 illustrates three distinct examples to showcase the aggregated effects among the three data sources, yet there are many more that can be explored.

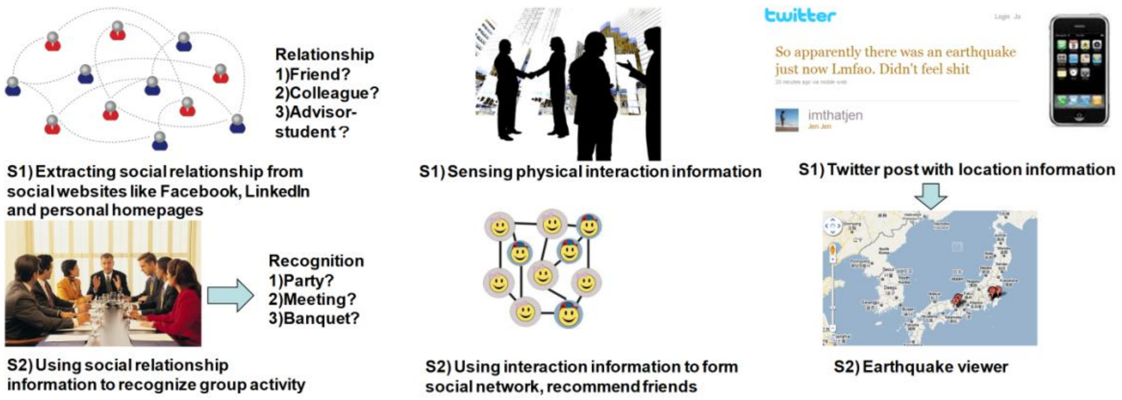

Fig. 1. Three examples to showcase the aggregated power of different data sources

(1) Sensor-based activity recognition enhanced by Web-mined knowledge. Social relationship from the Web can be used to assist social activity recognition in the physical world. For example, for a detected social gathering, if it is in the evening and the participants are all friends, it is more likely to be a party; if it occurs in a weekday morning and the participants are managers and subordinates, it is more likely to be a meeting.

Digital Object Indentifier 10.1109/MC.2011.65 0018-9162/$26.00 2011 IEEE

(2) Online social network enhanced by sensor-detected human interaction. Online social network still relies on user inputs to infer social relationship among users. As the user inputs only contain partial information about themselves and their friends, the predicted social connection is often inaccurate. By tracking real world user interactions via sensors and mapping the detected relationship onto online social network, we can significantly improve the quality of social network services. For example, if two people are found hanging out after work, they are probably close friends. If they meet only at work, they are merely colleagues. In [11], Pentland *et al.* use Bluetooth-enabled mobile phones to scan other devices in the user's proximity, which is then used to verify and better characterize relationships in an online social network system.

(3) Merging mobile sensing and Web data to better characterize a situation. Data from different sources often characterizes the specific facet of a situation, thus the fusion of several distinct data sources can often draw a better picture of the situation. For example, by integrating the mined theme from user posts and the revealed location information from GPS-equipped mobile phones, Twitter has been exploited to support near real-time report of earthquakes in Japan [6].

4. A General System Framework for SCI

As a large-scale community sensing system, the infrastructure for SCI needs a system framework to integrate large-scale and heterogeneous devices, software, and spaces, and provides systematic support for rapid application development, deployment, and evaluation. In Fig. 2, we present the general system framework for SCI, which consists of five layers.

Digital Object Indentifier 10.1109/MC.2011.65 0018-9162/$26.00 2011 IEEE

This article has been accepted for publication in Computer but has not yet been fully edited.
Some content may change prior to final publication.

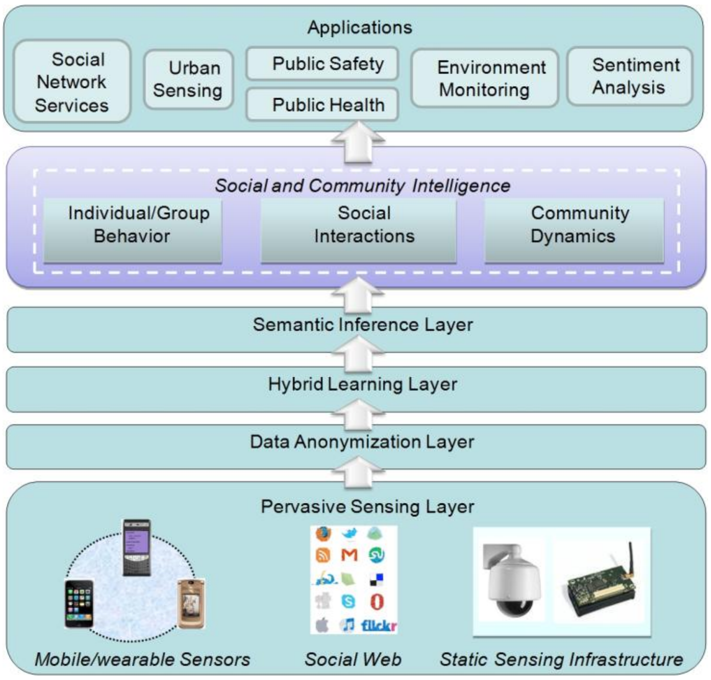

Fig. 2. A general SCI system framework

The *pervasive sensing* layer manages the three major information sources. As privacy is a major concern for both private and organizational data sharing, the proposed framework also incorporates a *data anonymization* layer before the data releasing and processing. The *hybrid learning* layer applies diverse machine learning and data mining techniques to convert the low-level single-modality sensing data into high-level features or micro-context, the focus is to mine the frequent data patterns to derive the individual's behavior and single space context, before extracting the complete social and community intelligence. The *semantic inference* layer is needed when different features or micro-context need to be aggregated using logic-based inferences, it is complementary with statistical learning approach and often very effective to process the explicit rules describing the logical relationship between hybrid learning layer outputs and expected SCI, based on expert's domain knowledge. Finally, the *application layer* includes a variety of potential services that can be enabled by the availability of SCI.

5. Major Application Areas

SCI applications are driven by the needs to (1) develop socially-aware services to facilitate interaction and communication among groups of people; (2) monitor the real-

Digital Object Indentifier 10.1109/MC.2011.65 0018-9162/$26.00 2011 IEEE

time change of physical world for public good; (3) track and predict specific events to benefit society. Here we present six major application areas.

5.1. Social Network Services

By logging various aspects of physical interactions and communication among people (e.g., co-location, conversations, call logs) and mining user behavior patterns (e.g., place of interests), SCI nurtures the development of many social network services, such as friend recommendation and augmented online interaction. For example, the FriendSensing application [12] can recommend friends to its users by monitoring one's activities with mobile phones, including text messages, phone calls, and encounters. The CenseMe project (http://www.cenceme.org/) exploits off-the-shelf smart phones to automatically infer people's presence (e.g., walking on the street, dancing at a party with friends) and then shares this presence through social network portals such as Facebook and Twitter.

In the EU FP7 SOCIETIES project (http://www.ict-societies.eu/), we plan to support the creation and management of different social communities in pervasive computing environments. A community can be formed by people co-located in a physical space (via environment sensing infrastructure); it can be built by gathering people with common interests/expertise through extracting information from home page or social web; or it can be created by linking those who have followed the similar routine doing exercise, which are detected by analyzing the traces recorded by wearable or mobile sensors. Apparently, social communities can take different forms and have different goals, they can be highly dynamic. The more information we can get from different data sources about people, the better we can support and manage social communities.

5.2. Urban Sensing

With wireless sensor platforms in the hands of masses, we can leverage community sensing to address urban-scale problems, such as city ambient monitoring, traffic planning, and better use of public utilities.

MIT's Real Time Rome project (http://senseable.mit.edu/realtimerome) uses aggregated data from cell phones, buses and taxis in Rome to better understand urban dynamics in real-time. The Biketastic project (http://biketastic.com) improves bike commuting in Los Angeles by collecting and mining biker-contributed data (using

Digital Object Indentifier 10.1109/MC.2011.65 0018-9162/$26.00 2011 IEEE

mobile phones). It enables bikers in the area to plan routes with the least probability of traffic accidents and the best air quality. Zheng et al. extract interesting locations and travel sequences from multiple user's GPS trajectories, and provide travel recommendations for new visitors of a city [13].

5.3. Environment Monitoring

The nomadic, participatory, and in-situ nature of community sensing provides new opportunities for environment monitoring and natural resource protection.

(1) Nature preservation. With the help of human volunteers, the Great Backyard Bird Count project reports the cumulative counts of birdwatchers from across American in its website (http://www.birdsource.org/gbbc/). The MIT Owl project (http://web.mit.edu/newsoffice/2008/tracking-0822.html), on the other hand, leverages the network of smart phones equipped with GPS, compasses, and directional microphones, to reduce human efforts in assessing owl populations.

(2) Pollution measurement. With the aid of portable pollution sensing devices, there have also been several projects targeting environment pollution measurement. The BikeNet application measures several metrics to give a holistic picture of the cyclist experience, including the CO2 level along the path. It facilitates public sensing and sharing by letting multiple users merge their individual data, for example, to create pollution and noise maps of their city [10]. In the PEIR project (http://peir.cens.ucla.edu/), GPS-enabled phones are used to detect user transportation mode (e.g., driving, walking), which is then used to assess an individual's environmental impact and exposure, like carbon footprints and exposure to air pollution.

5.4. Public Health

SCI can facilitate the anticipation and tracking of disease outbreaks across populations. For example, Epidemics of seasonal influenza are a major public health concern, causing tens of thousands of deaths worldwide each year. This problem can be remedied by early detection of the disease. The Google researchers have shown that by mining indirect signals from millions of geographically localized health-related search queries, one can estimate the level of influenza-like illnesses in regions of the United States with a reporting lag of just 1 day [14]. It is faster than the estimates provided by government agencies, which publish regional data weekly based on virology and clinical statistics.

Digital Object Indentifier 10.1109/MC.2011.65 0018-9162/$26.00 2011 IEEE

Besides public health monitoring, SCI also brings new opportunities for personal well-being management. With community sensing, we can log personal physical activity trajectory, track the food intake, sense the mental status in real-time, and record the social activities we attend each day, which can be used to improve human well-being management. For example, the Neat-o-Games system detects human movements (e.g., walking, running) by using a wearable accelerometer, and motivates users doing more exercises by showing the avatar of the user in a virtual community race game [15].

5.5. Sentiment Analysis

Sensing of user sentiments is important in context-aware computing. However, using physical sensors to directly sense personal sentiments is not an easy thing. Researchers have been exploring indirect ways to deal with this, one attempt is to collect or mine user-generated Web data. For example, Emotional City (http://www.emotionalcities.com/) and D-Tower (www.d-toren.nl) collect citizen moods through daily Web surveys, and display the emotions of the city through the change of lighting colors of a building or a public sculpture.

5.6. Public Safety

Public safety involves the prevention of and protection from events that could endanger the safety of the general public, these events can be crimes or disasters. Public video surveillance systems have assisted a lot to city-wide event sensing and safety monitoring. Recently, the Boston police department has embraced user contributed sensor data to assist in crime prevention [5].

Although only one or two types of data sources are used to support the above-mentioned applications, we can always find ways to enhance those applications, or even build new applications by incorporating more and heterogeneous data sources. For example, both mobile devices and infrastructure can help improving the applications in urban planning, environmental monitoring, well-being management and public safety.

6. Research Issues

We now turn our attention to key SCI research issues. Many of these are directly motivated by the SCI application discussed earlier, they are also in line with the functional layers in the SCI system framework described in Fig. 2, i.e. *sensing*, *data*

Digital Object Indentifier 10.1109/MC.2011.65 0018-9162/$26.00 2011 IEEE

anonymization, *data processing*, *social context learning*, and *SCI intelligence extraction*.

6.1. Sensing: Participatory or Opportunistic?

The first research issue to be considered is what roles people should play in community sensing. For example, in the case of using mobile phone as a sensing device, should the users be interrupted to control the status (e.g., accept, stop) of a sensing task-? There are two extreme cases for sensing:

- *Participatory sensing*. It incorporates people into significant decision making process of the sensing system, deciding which application request to accept, what data to share, and to what extent privacy mechanisms should be allowed to impact data fidelity. That's to say, it allows participants to retain control over their raw data. The Personal Data Valut system is based on this idea, which seeks to provide easy-to-use toolkits to support data control [16].
- *Opportunistic sensing*. It shifts the burden of users by automatically determining when devices can be used to meet application's sensing requests. Instead of requiring human intervention to actively and consciously participate in the sensing, opportunistic sensing requests that a sensing device is automatically used whenever its state (location, user activity, and so on) matches an application's requirements.

Obviously there exists a tradeoff between participatory sensing and opportunistic sensing. Participatory sensing places demands on user involvement, which restricts the pool of willing participants. Also people's tolerance in interruptions limits the number of applications that can be supported. These factors will limit both an application's scale and the diversity of applications that a purely participatory network could support [10]. For the opportunistic approach, one major concern is the potential leak of personally sensitive information during self-managed sensing process. Another issue is that opportunistic sensing takes on more resources for decision-making, such as determination of the sampling context (indicates when sampling should be started and stopped), adapting to the device's changing resource availability, and so on. More work needs be done to balance users' involvement and proper control while integrating proper protection mechanisms on data privacy.

6.2. Privacy, Data Quality, and Trust

Digital Object Indentifier 10.1109/MC.2011.65 0018-9162/$26.00 2011 IEEE

Sharing and revealing personal digital data could have a number of risks on user privacy. Compared with personal data (e.g., user profile, IDs), data gathered in community can reveal much more information about individual and organization's behaviors. For example, your location might reveal your interests; the health data about an organization might suggest environmental problems for the staff. The impact is obvious: if personal data cannot be anonymized and under the control of data owners, people may be less likely to share their data.

Privacy protection involves many elements, including identity (who is asking for the data?), granularity (how much does the data reveal about people? does it reveal one's identity?), and time (how long will the data be retained?). There are two main research areas that deal with these needs: *data anonymization* and *user control*.

(1) Data anonymization techniques. The objective of *data anonymization* is not revealing the identity of users when they contribute their data. Several methods have been proposed. For instance, MetroSense uses *k*-anonymous method when users contribute location data to a server, where a user's position is generalized to a region containing at least *k* users [10]. Another promising approach based on secure multiparty computation allows mining data from many different organizations without ever aggregating these data into a central data repository. Each organization performs part of the computation based on its privately held data, and uses cryptography to encode intermediate results that must be communicated to other organizations performing other parts of the computation [17]. Other privacy-preserving methods are also being explored, such as sharing only statistical summaries of the individual data sets, and inserting random perturbations into individual data records before sharing them.

(2) Enhancing user control and decision making. User control is very important in personal data sharing as it is about what one wants to reveal and to whom one allows the system to reveal. For example, you might want to track your heart rate each day, but there is no reason to share that information with anyone but your doctor. Researchers in this field are exploiting methods that enable users to manage their data by tailoring access-control and data-management tools [16].

The second issue is data quality. The data quality from the web can be very different, ranging from authorized, to inaccurate and even fake ones. The data quality from mobile phones and infrastructure also vary a lot. For example, some people put

Digital Object Indentifier 10.1109/MC.2011.65 0018-9162/$26.00 2011 IEEE

their mobile phones in the pocket, some put in the handbag. Thus for the same user activity like walking or running, the data quality is very different. Therefore, it is better to train different classifiers that work in different contexts. However, both data collection and context identification are challenging issues.

In addition to data privacy and quality issue, trust of data sources is another big thing. To mine social and community behaviors, we often need to collect data from many anonymous participants. If there lacks the control to ensure the source is valid and information is accurate, this can lead to data trust issue. For example, Twitter data is sometimes unreliable; mobile phone users may send incorrect or even faked data to the data centre. Therefore, trust and abnormal data detection methods should be developed to ensure the trustworthiness and quality of the collected data.

6.3. Managing Large-Scale Heterogeneous Data Sources

As in SCI system, the data producers can be very different in terms of modality (e.g., mobile phones, fixed cameras, Web services), their connectivity to the Internet (e.g., constant, intermittent, or affected by a firewall), their sharing willingness or privacy sensitivity, and resource capabilities for processing data locally. The information consumers are also heterogeneous in terms of running environments (applications that run locally or at community-level remotely), data needs (some might need only a high-level context information while others might need raw sensor data). The heterogeneity leads to several challenges on data management:

(1) Multi-modal. Different type of sensors have different attributes and capabilities, they might have different accuracy in sensing the physical and virtual world. Integrating information from diverse data sources adds difficulty to SCI mining. Raw data from different sensor sources need to be transformed to the same metrics and represented by a shared vocabulary/ontology to facilitate the learning and inference process [9].

(2) Temporal and continuous. The sensing data is recorded according to the time sequence, the system should consider multiple samples in the data stream while modeling the behaviors of individual and group, rather than consider each sensor reading in an isolated way. In addition, as the real world systems are all continuous, it’s important to build models catering for the discrete, sampled sensor state.

Digital Object Indentifier 10.1109/MC.2011.65 0018-9162/$26.00 2011 IEEE

(3) Large-scale data processing. The SCI applications that deal with real-time data collected from a large number of sensing nodes (e.g., computing and visualizing the real-time traffic condition in a city-scale) may suffer from the modeling and computational difficulties that exist in most data mining tasks. Further efforts have to be done on sampling optimism, problem decomposition, as well as adoption of advanced computational/learning models in terms of particular problem domain.

(4) Inconsistency. The same sensor may sense the same event under different conditions (for example, sensing one's voice in a quiet office or noisy restaurant). However, for the same event, user context often leads to different inference results (good or poor). Due to environmental differences, a group of co-located sensors running the same classification algorithm and sensing the same event in time and space could compute different inference results, and thus leads to the issue of system inconsistency. Miluzzo *et al.* have proposed a collaborative approach to dealing with this inconsistency problem [18] and more solutions are needed.

(5) Difficult to label all data. Asking human to label large amount of data set is often difficult since it is extremely time consuming to perform real-life experiments to collect data, it takes even more time to label all the data properly. Thus it is highly desirable to learn system models from relatively small amount of labeled data.

6.4. Extracting High-level SCI from Low-level Sensing Data

Social and community intelligence extraction considers the identification of a set of characteristics or behaviors associated with a social community based on the collection of user activity/preference/interest, group activity/preference/interest, social relationship and environmental context (those activities and relationships are learned or derived from raw sensor data). Such social communities can be flexibly formed by those people in the same organization, at same places, with same behaviors, of same interests, etc., depending on different social application requirements [11]. By pooling individual user's behavior traces together and mining the underlying social patterns, different social or group behaviors can be extracted [17]. The extracted social context can be a social event such as an open concert, can be a social pattern in daily activity, can be a relationship among a group of people, and can be socially significant locations.

Digital Object Indentifier 10.1109/MC.2011.65 0018-9162/$26.00 2011 IEEE

The key of the SCI pattern mining is to identify user similarity in the aforementioned social patterns with the objectives of offering social aware services. Unsupervised learning techniques, such as clustering, latent semantic analysis, matrix factorization, can be applied to achieve social context mining based on the user behavioral similarities. The process includes mining and discovery of common social contexts such as personal characteristics, cuisine preferences, eagerness of social participation, and also discovery of undefined social patterns for interest matching and social choice ranking.

In order to infer the social events based on the user context traces, the semantic gap between the low-level individual activities/spaces (e.g., walking/street, eating/restaurant, etc) and high-level social events (e.g., meeting, party, etc) should be bridged using machine learning and inference techniques. As highlighted previously, the analysis of the latent relations between the basic human activities and semantic social events is the research focus of this part, with the goal of learning an ontology describing the relationship between the basic human activities and semantic social events.

7. Conclusion

Social and Community Intelligence (SCI) represents a new interdisciplinary research and application field. With the rapid accumulation of "digital footprints" at community scale, we believe that the research scope of SCI will expand and its applications to multiply in next years to come. As an emerging area, the prevalence and development of SCI still face challenges ranging from multi-modal data gathering, heterogeneous data representation, to complex intelligence inference and privacy issues, which are expected to nurture many new research opportunities. Even though the existing practices on social and community intelligence mainly consider single type of information sources − static sensor infrastructure, mobile and wearable sensors, or Internet and Web − we expect to see the explosion of the research on utilizing the aggregated power of three information sources as well as innovative applications enabled by SCI.

Acknowledgment

This research is supported by the European Commission FP-7 SOCIETIES Project.

Digital Object Indentifier 10.1109/MC.2011.65 0018-9162/$26.00 2011 IEEE

References

[1] F. Wang et al., "Social Computing: From Social Informatics to Social Intelligence," *IEEE Intelligent Systems*, vol. 22, no. 2, 2005, pp.79–83.

[2] N. Eagle, et al., "Inferring Social Network Structure using Mobile Phone Data," *Proceedings of the National Academy of Sciences (PNAS)*, vol. 106, no. 36, 2007, pp. 15274-15278.

[3] S. Staab et al., "Social networks applied," *IEEE Intelligent systems*, vol. 20, no. 1, 2005, pp. 80-93.

[4] R. Xiang et al., "Modeling relationship strength in online social networks," *Proc. of WWW 2010 Conf.*, 2010.

[5] A. Sheth et al., "Citizen Sensing, Social Signals, and Enriching Human Experience," *IEEE Internet Computing*, vol. 13, no. 4, 2009, pp. 87-92.

[6] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors," *Proc. WWW 2010 Conf.*, ACM, 2010.

[7] H. Andy et al., "The Anatomy of a Context-Aware Application", *Wireless Networks,* vol. 8, no. 2-3, 2002, pp.187-197.

[8] M. Philipose et al., "Inferring ADLs from interactions with objects," *IEEE Pervasive Computing*, vol. 3, no. 4, 2004.

[9] X. Wang, et al., "Semantic Space: An Infrastructure for Smart Spaces," *IEEE Pervasive Computing*, vol. 3, no. 3, 2004, pp. 32-39.

[10] A. T. Campbell et al., "The Rise of People-Centric Sensing," *IEEE Internet Computing*, vol. 12, no. 4, 2008, pp. 12-21.

[11] A. Pentland, "Socially Aware Computation and Communication," *IEEE Computer*, vol. 38 no.3, 2005, pp. 33-40.

[12] D. Quercia et al., "Nurturing Social Networks Using Mobile Phones," *IEEE Pervasive Computing*, 2010.

[13] Y. Zheng et al, "GeoLife: A Collaborative Social Networking Service among User, location and trajectory," *IEEE Data Engineering Bulletin*, vol. 33, no. 2, 2010, pp. 32-40.

[14] N.M. Ferguson et al., "Strategies for mitigating an influenza pandemic," *Nature*, vol. 442 no, 7101, 2006, pp.448-452.

[15] Y. Fujiki, et al., "NEAT-o-Games: Blending Physical Activity and Fun in the Daily Routine," *ACM Computers in Entertainment*, vol.6 no. 2, 2008.

[16] K. Shilton, "Four billion little brothers?: Privacy, mobile phones, and ubiquitous data collection," *Communications of the ACM*, vol. 52, no. 11, 2009, pp. 48–53.

[17] T.M. Mitchell, "Mining Our Reality," *Science*, vol. 326, no.5960, 2009, pp. 1644-1645.

[18] E. Miluzzo et al., "Darwin Phones: The Evolution of Sensing and Inference on Mobile Phones," *Proc. of MobiSys '10 Conf.*, San Francisco, USA, 2010.

Short Bios:

Daqing Zhang is a professor at Institute TELECOM SudParis, France. He obtained his Ph.D. from University of Rome "La Sapienza" and University of L'Aquila, Italy in 1996. His research interests include pervasive and human-centric sensing, context-aware systems, social and community intelligence, mobile social networks and pervasive healthcare. He is a member of IEEE and ACM. Contact him at daqing.zhang@ it-sudparis.eu.

Bin Guo is a post-doctorate researcher at Institute TELECOM SudParis, France. His research interest includes ubiquitous computing, context-aware systems, and community sensing. He received his Ph.D. degree in computer science from Keio University, Japan, in 2009. He is a member of IEEE. Contact him at bin.guo@it-sudparis.eu.

Digital Object Indentifier 10.1109/MC.2011.65 0018-9162/$26.00 2011 IEEE

This article has been accepted for publication in Computer but has not yet been fully edited.
Some content may change prior to final publication.

Zhiwen Yu is currently a professor at Northwestern Polytechnical University, P. R. China. He received his B.Eng, M.Eng and Ph.D degree of Engineering in computer science and technology in 2000, 2003 and 2005 respectively from the Northwestern Polytechnical University. His research interests cover pervasive computing, context-aware systems, human-computer interaction, and personalization. He is a member of ACM and IEEE. Contact him at zhiwenyu@nwpu.edu.cn.

Contact Information:

Dr. Daqing ZHANG
Network and Services Department
Institut TELECOM SudParis
9, rue Charles Fourier
91011 Evry Cedex, France
Email: Daqing.Zhang@it-sudparis.eu
Phone: (+33) 1 6076 4123
Fax : (+33) 1 6076 42 91

Dr. Bin GUO
Network and Services Department
Institut TELECOM SudParis
9, rue Charles Fourier
91011 Evry Cedex, France
Email: Bin.Guo@it-sudparis.eu
Phone: (+33) 1 6076 4435
Fax : (+33) 1 6076 4291

Dr. Zhiwen YU
Mailbox 886,
Northwestern Polytechnical University,
Chang An Campus, Xi'an, Shaanxi,
710129, P. R. China
Email : zhiwenyu@nwpu.edu.cn

Digital Object Indentifier 10.1109/MC.2011.65 0018-9162/$26.00 2011 IEEE