Jordi Nin
Daniel Villatoro (Eds.)

LNAI 7685

Citizen in Sensor Networks

First International Workshop, CitiSens 2012
Montpellier, France, August 2012
Revised Selected Papers

Springer

Lecture Notes in Artificial Intelligence 7685

Subseries of Lecture Notes in Computer Science

Jordi Nin Daniel Villatoro (Eds.)

Citizen in Sensor Networks

First International Workshop, CitiSens 2012
Montpellier, France, August 27, 2012
Revised Selected Papers

Springer

Volume Editors

Jordi Nin
Universitat Politècnica de Catalunya
Department of Computer Architecture
c. Jordi Girona 1-3
08034 Barcelona, Spain
E-mail: nin@ac.upc.edu

Daniel Villatoro
IIIA – Artificial Intelligence Research Institute
CSIC – Spanish Scientific Research Council
Campus Universitat Autonoma de Barcelona
08193 Bellaterra, Spain
E-mail: dvillatoro@iiia.csic.es

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-36073-2 e-ISBN 978-3-642-36074-9
DOI 10.1007/978-3-642-36074-9
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2012955347

CR Subject Classification (1998): I.2.9, H.2.8, I.2.6, I.2.11, H.3.4-5, H.4.1-3, H.5.3, J.2

LNCS Sublibrary: SL 7 – Artificial Intelligence

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Foreword from the CitiSens 2012 Program Chairs

The current volume constitutes the revised proceedings of the First International Workshop on Citizen Sensor Networks (CitiSens 2012), which includes revised versions of the papers presented at the workshop. The aim of CitiSens is to promote and stimulate the international collaboration and research exchange on novel smart cities and sensor networks topics. This first edition of the workshop was co-located with the ECAI 2012 conference in Montpellier (France).

The program of this year's workshop consisted of the presentation of the six accepted full papers (out of 16 submitted papers), and one keynote lecture. The accepted papers deal with topics such as crowd sourcing, smart cities, multi-agents systems, privacy in social networks, data anonymity, or smart sensors. Each paper was reviewed by at least three reviewers.

We would like to acknowledge and thank all the support received from the Program Committee members, external reviewers, and the organizing Committee of ECAI 2012. We would like to warmly thank Josep Domingo-Ferrer for his support through the UNESCO Chair in Data Privacy.

Last, but definitely not least, we would like to thank all the authors who submitted papers, all the attendees, and the keynote speaker, Victor Muntés-Mulero, who accepted our invitation to give a talk entitled "Crowdsourcing for Industrial Problems" for all the attendants of CitiSens 2012.

Jordi Nin
Daniel Villatoro

First International Workshop on Citizen Sensor Networks – CitiSens 2012

Organizing Committee

Jordi Nin	Universitat Politècnica de Catalunya, Spain
Daniel Villatoro	IIIA-CSIC, Spain

Program Chair

Jordi Nin	Universitat Politècnica de Catalunya, Spain

Program Committee

Albert Bifet	University of Waikato, New Zeland
Rosaria Conte	ISTC - CNR, Italy
Josep Domingo-Ferrer	Rovira i Virgili University, Spain
Elena Ferrari	University of Insubria, Italy
Sébastien Gambs	Université de Rennes 1 - INRIA, France
Lluis Godo	IIIA-CSIC, Spain
Vicente Julian	Universidad Politécnica de Valencia, Spain
Michael Luck	King's College London, UK
Stan Matwin	Ottawa University, Canada
Katarzyna Musial	King's College London, UK
Pascal Poncelet	LIRMM-CNRS, France
Marc Pous	Barcelona Digital Technology Center, Spain
Ramon Sanguesa	Politechnical University of Catalonia, Spain
Nishanth Sastry	King's College London, UK
Sandip Sen	University of Tulsa, USA
Jetzabel Serna	Universitat Politècnica de Catalunya, Spain
Marc Solé	Universitat Politècnica de Catalunya, Spain
Vicenç Torra	IIIA-CSIC, Spain
Meritxell Vinyals	University of Verona, Italy

Table of Contents

Citizen Sensor Networks

Citizens Sensor Networks 1
Daniel Villatoro and Jordi Nin

Crowd-Sourcing

Crowdsourcing for Industrial Problems 6
Victor Muntés-Mulero, Patricia Paladini, Jawad Manzoor, Andrea Gritti, Josep-Lluís Larriba-Pey, and Frederik Mijnhardt

Co-ordination and Data Mining in Citizen Sensor Networks

Multiagent Co-ordination of Wireless Sensor Networks 19
Maria del Carmen Delgado-Roman, Marc Pujol-Gonzalez, and Carles Sierra

On the Protection of Social Network-Extracted Categorical Microdata 33
Jordi Marés and Vicenç Torra

The TweetBeat of the City: Microblogging Used for Discovering Behavioural Patterns during the MWC2012 43
Daniel Villatoro, Jetzabel Serna, Víctor Rodríguez, and Marc Torrent-Moreno

Smart Cities

A Platform for Citizen Sensing in Sentient Cities 57
Fernando Koch, Carlos Cardonha, Jan Marcel Gentil, and Sergio Borger

Incorporating Mobility Patterns in Pedestrian Quantity Estimation and Sensor Placement 67
Thomas Liebig, Zhao Xu, and Michael May

Users as Smart Sensors: A Mobile Platform for Sensing Public Transport Incidents 81
Cristian Tanas and Jordi Herrera-Joancomartí

Author Index 95

Citizens Sensor Networks

Daniel Villatoro[1] and Jordi Nin[2]

[1] Barcelona Digital Technology Centre
Barcelona, Catalonia, Spain
dvillatoro@bdigital.org
[2] Department of Computer Architecture
Technical University of Catalonia - BarcelonaTECH (UPC)
Jordi Girona 1-3, 08034 Barcelona, Catalonia, Spain
nin@ac.upc.edu

Abstract. This introductory paper serves as an overview about the rest of the papers that are contained within this volume. In this article it is presented our vision of the Citizen Sensor Networks as twofold: one where the citizens are passive entities that need to be tracked to understand and optimize better the SmartCities, and the second where the citizens, motivated by their common sense and using their mobile device to communicate the sensed sample.

Also in this abstract we will introduce the concept of crowd sourcing or crowd compution and its industrial applications.

1 Introduction

Catalyzed by the Industrial Revolution, cities have become the acting scenario for the economic trading and exchanges that have derived into the modern economic system. Due to several factors (such as a reduction of commuting times and increase the interaction capabilities), the population within the cities suffered an outstanding growth and becoming such organizational structure the predominant and preferred for human interactions.

This growth has not only represented a significant growth in the population within the cities but also in the developed infrastructures to make such population increase sustainable. The infrastructures we refer to range in several typologies: transportation (roads, highways, underground trains, bus, trams), communications (phone lines, internet, post offices), health (hospitals, gyms, infirmary houses) or education (school, universities, libraries) amongst many others.

Trying to improve the control and optimization of urban behavior, city managers uniformly decided to invest in the improvement of the sensing infrastructures within the city. The penetration of mobile technologies have resulted in a massive data provision from their users, continuously sharing information anytime anywhere.

The advantages of this continuous sensing capability are twofold:

- *Proactive sensing*: The ability to share information through the mobile devices combined with the humans' "common sense" transforms any device-holder in a potential proactive intelligent sensor, which complements the classical continuous-sensing sensors installed in a certain location with an specific scope, whose performance is based on the detection of anomalous performances. Some successful

J. Nin and D. Villatoro (Eds.): CitiSens 2012, LNAI 7685, pp. 1–5, 2013.

applications within this line are SeeClickFix (where users can report about any type of infrastructure in the city) or Waze (this is a crowd-sourced navigator where the own users update the state of the traffic on the road in real time).
– *Passive sensing*: In the advent of the SmartCity Paradigm, one of the most challenging tasks to be solved is understanding the population distribution in real time. Such information would be of crucial importance for city managers as some of the services offered to the citizens could be tailored with complete information and in real time. Classical methods have profited from domain experts in the city or polling citizens to understand the needs of the citisens.

2 Crowd Sourcing

The generalized use of the Internet and social network platforms has changed the way human beings establish relations, collaborate and share resources. In this context, crowd sourcing (or crowd computing) is becoming a common solution to provide answers to complex problems by automatically coordinating the potential of machines and human beings working together [5,3,4]. Several challenges still separate crowd sourcing from its generalized acceptance by industry. For instance, the quality delivered by the workers in the crowd is crucial and depends on different aspects such as their skills, experience, commitment, etc. Trusting the individuals in a social network and their capacity to carry out the different tasks assigned to them becomes essential in speeding up the adoption of this new technology in industrial environments. Capacity to deliver on time, cost or confidentiality are just some other possible obstacles to be removed. In the main talk of this workshop, the plenary speaker Victor Muntés-Mulero, will discuss some of these issues, provide solutions to improve the quality in systems based on the use of crowd sourcing and present a real industrial problem where we use the crowd to leverage the work capacity of geographically distributed human beings.

3 Sensing the Citizens to Create a Smart Sensor Network

Nowadays, many people have become Internet *citizens* or Web-enabled social citizens; the use of smartphones enables such citizens to easily connect to the Internet to upload lot of data, this fact gives to such devices the ability to act as sensors. Thus, the term citizen-sensor network refers to an interconnected network of people who actively observe, report, collect, analyze, and disseminate information via text, audio, or video messages [6]. A possible example of social sensing is depicted in Figure 1.

This combination of human-in-the-loop sensing, Web 2.0, and mobile computing has led to the emergence of several citizen-sensor networks [7]. In particular, Web 2.0 fostered the open environment and applications for tagging, blogging, wikis, and social networking sites that have made information consumption, production, and sharing so incredibly easy. However, two significant developments in mobile computing helped enable citizen-sensor networks as we know them today: enhanced features such as GPS capability and cameras became a standard part of most mobile devices, and large companies created open mobile operating systems, such as Apple's OS X for the iPhone and Google's Android.

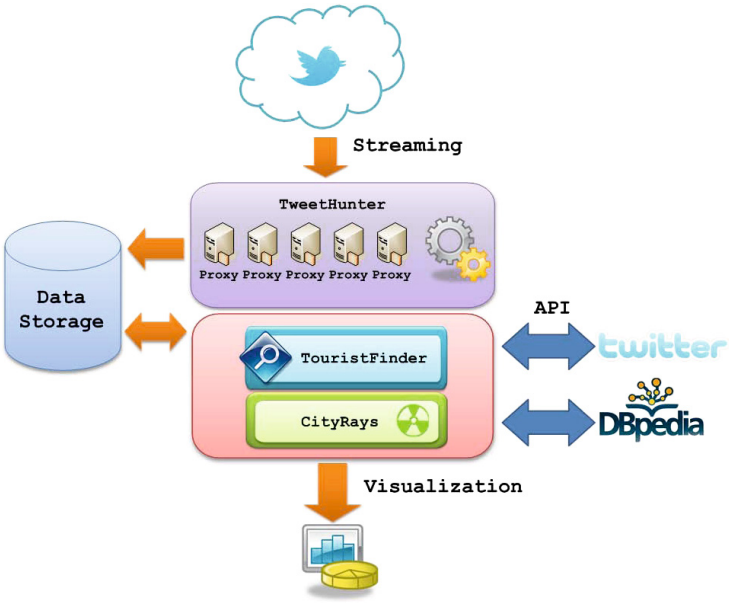


Fig. 1. Social Sensing Platform Example

For instance, Microblogging, in which users share short messages and pictures, typically over the Web, it is of particular interest to citizen-sensors [8,3]. This relatively new technology emerged on the Web in 2006 and achieved widespread adoption extremely quickly. This technology allows us to discover citizens behavioral patterns have been analyzed using several sources of information. In the work *"The TweetBeat of the City: Microblogging Used for Discovering Behavioural Patterns during the MWC2012"*, the authors present how this behavioral patterns can be analyzed using Twitter as a source of data. Social Media platforms act as an incentive where users share their content. This act of sharing in these types of platform has become more immediate with the usage of mobiles devices and users are more prone to share "while it happens". Profiting from the public availability of the information published in Twitter, and the possibility to attach GPS positions to the metadata of the tweets, authors analyze the different patterns of activity in the city of Barcelona with and without the presence of an event. Authors analyze how we can profit from the information shared in those platforms to use citizens as sensors of their own activity.

However, and even though we can profit from the information publicly shared by the users in Social Media platforms, this information has to ensure the anonymity of the users. In the work *"On the Protection of Social Network-Extracted Categorical Microdata"*. Their specific case of study is to extract and protect microdata datasets from a real social network such as Twitter. Authors analyze how different levels of k-anonymity ensures that the information can be publicly released for research without compromising the identity of the users and without having a very large information loss on the protected microdata file.

In many cases the usage of a single sensor cannot be enough to obtain all the information about the object/process that is sensorized. In such cases, more than one sensor needs to be deployed to capture all the information [2,1]. In the work *"Multi-agent coordination of Wireless Sensor Networks"*, authors propose a self-organization algorithm, called Coalition Oriented Sensing Algorithm (COSA), for sensors to correctly coordinate while maximizing the life span of the overall sensor network. The main contribution of this algorithm is extending the life span of the different sensors while guaranteeing their good performance.

Apart from these more theoretical works, more practical examples of how to apply citizen sensors have been also described. For instance, one example of a multi-sensor architecture for sensing the citisens is the one presented in the work *"Incorporating Mobility Patterns in Pedestrian Quantity Estimation and Sensor Placement"*. In this work, authors present an analysis of sensing the visitors attending to a soccer stadium, using a Bluetooth approach. The empirical analysis on real world data collected with Bluetooth tracking technology during a soccer event at a soccer event at Stade des Costieres in Nimes (France). This work represents a real-example of the problem of sensing the citizens even though at a lower scale than the city. The most important contribution of this paper is a novel method to determine where a fixed number of automatic pedestrian quantity sensors is to be located during a mass event in order to get an adequate estimate on the presence of the visitors within the site.

4 Citizens: The Smartest Sensor

The high level of citizen participation in disseminating information during last years demonstrates the growing power of citizen influence on real life events. Using Flickr or Twitter ordinary people can share their views of the events as they unfolded.

This is the case of sensing public transport incidents as it was described in work *"Users as Smart Sensors: A Mobile Platform for Sensing Public Transport Incidents"*. Here, the authors present us with a mobile app developed to notify and be informed about the incidences of the public rail network in the Barcelona metropolitan area. This type of application is a clear example of the proactive sensing performed by the citizens. The success of this type of application is on the critical mass necessary to obtain a representative sample of the reality that is trying to sensorize. Authors have been successful obtaining such critical mass thanks to a combination of a clear presentation of the benefits of the application to the own users and a publicity campaign. The other challenge that authors have proposed as future work is the research on methods to maintain the fidelity of users.

This type of works develops the concept of *Sentient City*, a city that can remember, correlate, and anticipate future events and needs. To do that, it is necessary to aim technologies to interconnect people, allowing them to actively observe, report, collect, analyse, and disseminate information about urban events. This is the job done in *"A Platform for Citizen Sensing in Sentient Cities"*, authors present an unified framework that bases its functioning in the citizens as sensors. This platform is an effort of *IBM*

specifically targeted towards the next events in Brazil: the 2014 World Cup and the 2016 Olympic Games. Three examples presented in the paper cover the detection of garbage in the streets, as well as the detection of areas of inundation and finally, the identification of traffic jams.

5 Conclusion

In this brief overview, we have presented three different scenarios where citizen sensor networks largely impacts in our daily life. Several issues on sensing aspects, interconnection and /or integration of sensors and humans still require from considering many computational aspects in order to improve the performance and quality of the actual solutions for smart cities. Distributed computation, a smarter use of network communications, the reduction of the energy required for sensing, new applications, etc. are just some of the topics that might feed this workshop.

Acknowledgments. We thank the Spanish MEC for its partial support through the project ARES-CONSOLIDER INGENIO 2010 CSD2007-00004. Also, this research is partially supported by the Spanish Centre for the Development of Industrial Technology under the INNPRONTA program, project IPT-20111006, "CIUDAD2020" (www.innprontaciudad2020.es).

Finally, we would like to acknowledge and thank all the support received from the program committee members, external reviewers, and the organization committee of ECAI 2012. We would like to warmly thank Josep Domingo-Ferrer for his support through the UNESCO Chair in Data Privacy.

References

1. Clare, L.P., Pottie, G.J., Agre, J.R.: Self-organizing distributed sensor networks. In: Proc. SPIE 3713, Unattended Ground Sensor Technologies and Applications (1999)
2. Ganeriwal, S., Balzano, L.K., Srivastava, M.B.: Reputation-based framework for high integrity sensor networks. ACM Transactions on Sensor Networks 4(3) (2008)
3. Geiger, D., Seedorf, S., Schulze, T., Nickerson, R.C., Schader, M.: Managing the crowd: Towards a taxonomy of crowdsourcing processes. In: AMCIS (2011)
4. Howe, J.: Wired 14.06: The Rise of Crowdsourcing
5. Lease, M., Yilmaz, E.: Crowdsourcing for information retrieval. SIGIR Forum 45(2), 66–75 (2012)
6. Nagarajan, M., Gomadam, K., Sheth, A.P., Ranabahu, A., Mutharaju, R., Jadhav, A.: Spatio-Temporal-Thematic Analysis of Citizen Sensor Data: Challenges and Experiences. In: Vossen, G., Long, D.D.E., Yu, J.X. (eds.) WISE 2009. LNCS, vol. 5802, pp. 539–553. Springer, Heidelberg (2009)
7. Sheth, A.: Citizen sensing, social signals, and enriching human experience. IEEE Internet Computing 13(4), 87–92 (2009)
8. Weng, J., Yao, Y., Leonardi, E., Lee, F.: Event Detection in Twitter. Technical report, HP Labs (2011)

Crowdsourcing for Industrial Problems

Victor Muntés-Mulero[1], Patricia Paladini[1], Jawad Manzoor[2], Andrea Gritti[2], Josep-Lluís Larriba-Pey[2], and Frederik Mijnhardt[3]

[1] CA Technologies, Spain
{Victor.Muntes,Patricia.PaladiniAdell}@ca.com
[2] Universitat Politècnica de Catalunya, Spain
{jawad,larri}@ac.upc.edu, andrea.gritti@est.fib.upc.edu
[3] University of Utrecht, The Netherlands
a.f.mijnhardt@uu.nl

Abstract. The generalized use of the Internet and social network platforms has changed the way human beings establish relations, collaborate and share resources. In this context, crowdsourcing (or crowd computing) is becoming a common solution to provide answers to complex problems by automatically coordinating the potential of machines and human beings working together. Several challenges still separate crowdsourcing from its generalized acceptance by industry. For instance, the quality delivered by the workers in the crowd is crucial and depends on different aspects such as their skills, experience, commitment, etc. Trusting the individuals in a social network and their capacity to carry out the different tasks assigned to them becomes essential in speeding up the adoption of this new technology in industrial environments. Capacity to deliver on time, cost or confidentiality are just some other possible obstacles to be removed. In this paper, we discuss some of these issues, provide solutions to improve the quality in systems based on the use of crowdsourcing and present a real industrial problem where we use the crowd to leverage the work capacity of geographically distributed human beings.

1 Introduction

For many years, we have seen a huge increase in the use of sophisticated systems that have enabled world-wide collaboration through our home PCs and, more recently, through ubiquitous hand-held devices. The purposes are as numerous as they are varied: content sharing, whether through blogs or many well-known peer-to-peer (P2P) applications; collaborative computation, starting from the early SETI@home project (setiathome.berkeley.edu) that was one of the first large-scale grid computing instances, and other examples.

People are gaining awareness of the power of collaborating through the network. We have recently seen national revolutions, like in Egypt, where people organized themselves using digital platforms. The *crowd* is becoming aware of its power, and the next natural step is to enhance the tools and modalities for collaborative computing. Powerful devices, like smartphones and tablets, are able to carry out an impressive amount and array of computation. P2P computing has shown to be feasible and efficient. We have some examples, such as Skype, that show the model is valid and can challenge serious cloud-based competitors, such as Google Voice.

J. Nin and D. Villatoro (Eds.): CitiSens 2012, LNAI 7685, pp. 6–18, 2013.

Not only machines but also real people are connected to the network combining their computing and thinking capacity. Trends seem to be pointing to this model as gaining the position to complement (or perhaps substitute) cloud computing: connecting people and machines in a single network. Nowadays, millions of people are asynchronously analyzing, synthesizing, providing opinion and labelling and transcribing data that can be automatically mined, indexed and even learned. Therefore, there is not much difference between this and classical computing: the crowd is working online, taking digital data as input and yielding digital data as output. The main difference is that human brain-guided computation is able to perform tasks that computers can hardly do, at overwhelming speeds. Tagging a picture or a video based on their content or answering questions in natural language, are just a couple of examples.

In this paper, we discuss about the typology of problems that can be solved through crowdsourcing at an industrial level and the main challenges to be solved to make this technology generally adopted. We present two elements that we consider essential for quality. First, we propose the Action-Verification Unit (AV-Unit), a quality control mechanism that helps organizing the crowd to not only perform actions, but also evaluate the quality of the results during the process. Second, we propose to use rewarding systems based on the quality delivered by each single worker. Also, we present a real industrial example: a crowdsourcing platform for translation that it is being developed by CA Technologies[1], using AV-Units. The system we present is novel in the sense that it proposes a model to perform a large number of tasks in parallel, and uses multiple levels of verification to ensure industrial quality standards. To our knowledge, it is also the first proposal that adds a quality-aware rewarding system that rewards workers based on a ranking that measures the quality of their work. Finally, our system allows sizing tasks at our convenience in such a way that we avoid the problem of the lack of context in translations based on isolated sentences.

This paper is organized as follows. Section 2 describes the state-of-the-are in crowdsourcing and presents some industrial applications that use crowdsourcing in order to solve non-trivial problems. In Section 3, we present a method to strengthen the capacity of the system to provide quality in the results. We also discuss about rewarding methods and scalability. Section 4 presents an example of a crowdsourcing platform developed at CA Technologies. Finally, Section 5 concludes and draws some future research lines.

2 Crowdsourcing

The term crowdsourcing was used for the first time by Jeff Howe in 2006 [8], referring to the increasing practice of outsourcing task to internet as an open call over a variety of users. Since then, crowdsourcing has evolved to exploit the work potential of a large crowd of people remotely connected through the Internet. For instance, recent work studies different typologies and uses of crowdsourcing and proposes a possible taxonomyin [6]. They categorized crowdsourcing depending on different methodologies and processes divided according to several dimensions that are shown to impact

[1] CA Technologies is a worldwide software and solutions provider that helps customers to make ICT management more agile, secure and flexible.

the behaviour of workers within the crowd, and the tasks that can be outsourced to the crowd.

As this idea increases in terms of popularity, several general purpose crowdsourcing platforms have appeared in the last years. For instance, Amazon Mechanical Turk (mturk.com) is a crowdsourcing marketplace that enables companies or individuals to utilize the human intelligence to perform tasks that are difficult for computers. The requesters post tasks known as Human Intelligence Tasks (HITs) that can be viewed by workers. Other examples like CrowdFlower (crowdflower.com) or ClickWorker (clickworker.com), extend Mechanical Turk capabilities offering a variety of crowdsourcing services. They improve quality by using gold standard units, redundant reviews of each data unit, etc. Their workflow management system divides complex tasks into smaller units and distributes them among the crowd based on the profile of individuals.

Quality control is a key point in crowdsourcing and it changes depending on the nature of the task which is crowdsourced. For instance, on the one hand, Lease and Yilmaz [10] show how the results obtained from the crowd are more inaccurate compared to laboratory participants. On the other hand, Yan et al. [14] present *CrowdSearch*, a system to search images on mobile phones using the crowd. In their work they show that workers are able to achieve over 95% precision. Other lines of research study the effect that different rewarding systems have on quality. For instance, Harris [7] shows that financial incentives actually encourage quality.

Crowdsourcing markets are traditionally used for simple and independent tasks. For example, labeling an image or finding relevance between search results. In [9], the authors present a framework that enables solving complex and interdependent tasks using crowdsourcing markets. The authors follow an approach similar to MapReduce for breaking down a complex problem into a sequence of simpler subtasks. The subtasks are solved in parallel by the crowd and the results are combined to form the final solution.

In the use case presented in this paper, we focus on the use of crowdsourcing for translation and, especifically, on software localization. Several research works are based on the use of Amazon Mechanical Turk for translation [4,5,11,16]. Zaidan et al. [16] propose some factors to select a good translation among a set of different versions. These factors include the workers country of residence, native language, etc. and each factor has a weight. In this way, the total score is calculated for a specific translated text. Experimental results show that some translations turn out to be very close in quality to the ones made by professional translators. Two other studies [2,4] investigate the use of crowdsourcing to evaluate the output of machine translated natural language. Gao and Vogel [5] present a case study of word alignment tasks performed by crowd on Amazon Mechanical Turk and Matteo et al. [11] use crowdsourcing to create corpora to feed and enrich Statistical machine translation. Other studies focus on using crowdsourcing for post-editing tasks. Bernstein et al. [3] propose the Find-Fix-Verify pattern to perform tasks like text shortening and proofreading. They split the tasks into a series of stages that utilize independent agreement and voting to produce reliable results.

However, these systems might not be suitable for implementing an industrial software localization process for two main reasons. Firstly, these systems are based on the

resolution of very small tasks, mostly at sentence level, and their entire quality assurance methodology is based on this reduced amount of information. Secondly, many of them use automatic evaluation methods like BLEU [12] and METEOR [1] to evaluate the quality of translation. These methods however rely on the existence of pre-computed golden translations. Golden translations are rarely available for unpopular languages and cannot be used for new texts. Thus requiring human intervention to decide which translations are to be accepted and which not, adding an additional managerial layer to the process.

2.1 Is the Crowd a Universal Solution for Industry?

Not any problem in industry is suitable to be crowdsourced. First of all, many processes are and can be automated by machines instrumented with intelligent AI-guided algorithms. Secondly, many processes will always require an onsite workforce to manage and monitor the operations. However, between these two a wide variety of applications will be suitable to be crowdsourced. To determine whether a process can be crowdsourced we detect a set of proprietary characteristics that must be taken into account:

- ***Activities must not be easy to automate:*** crowd computing activities usually require to perform actions that are better solved by the human brain than by currently available algorithms. Usual examples are tasks where creativity is essential, such as the proposal of new designs; tasks where the geographic distribution of the individuals provides a higher quality access to information; or tasks where the complexity of the problems posed are so high that there are not predefined mechanisms to solve them. A more complete survey on the type of tasks suitable for crowdsourcing may be found in [15].
- ***Information involved in the process must not be confidential:*** since data must be sent to the crowd, processes that involve sensitive information are not suitable in general for this type of solution, unless data is preprocessed first, for instance, applying anonymization techniques. On the positive side, crowdsourcing also mitigates concerns about loss of privacy, since a single provider does not have a global view of anyone's data. Later on in this paper, we will see a real example that benefits from crowd distributing data so that nobody has a complete picture of all the documents.
- ***Training must be simple or highly automatic:*** complex training processes are not suitable for crowdsourcing since this would imply training thousands or even millions of people, which would be unaffordable.

If these conditions hold, the use of crowdsourcing has multiple advantages. First, crowdsourcing delivers elasticity. Analogously to cloud computing, by working with the crowd, we have a virtually infinite number of resources that may be allocated and deallocated depending on the workload. Therefore, crowdsourcing offers flexibility in processes that include human beings. Secondly, it eliminates middlemen costs. By building a platform to manage the crowd automatically, we gain direct access to the final workers, eliminating intermediate vendors that increase the cost of services.

Real Industrial Applications of Crowd Computing. Crowd computing has been successfully used for industrial applications. Some tasks which are difficult to automate, such as those based on innovation and design, are good candidates for crowdsourcing. NamingForce (namingforce.com) or Threadless (www.threadless.com) are just two examples. Other examples are focused on technical areas. For instance, InnoCentive (innocentive.com) is a problem solving marketplace that brings together solution seeking companies and problem solvers dispersed all over the world. TopCoder (topcoder.com) uses the crowd so solve complex programming challenges. There exist many other examples such as testing, like in the case of uTest (www.utest.com), or journalism, like OpenFile (openfile.ca).

3 Quality, Motivation and Scalability

Three main challenges when developing a crowdbased system can be found in quality, motivation and scalability. This section proposes an approach to deal with the questions: (i) how to deliver quality when working with the crowd; (ii) how to motivate the crowd to participate in industrial processes; and (iii) how to make systems scalable.

3.1 Quality Based on Organized Verification

Crowdsourcing is sometimes associated to low quality. The participation of a massive amount of people who are geographically distributed is intuitively assumed to be a barrier for quality. In general, it is difficult to establish automatic mechanisms in order to monitor quality in crowd-based systems, since the lack of automatic methods to solve problems usually imply a poor definition of quality and a high complexity in order to establish a universally accepted quality measure. How do we measure the degree of innovation of an idea, the beauty of a proposal or the best style for a text? These are measurements that are highly dependant on subjectivity. Nevertheless, many previous proposals are still based on methods to monitor quality based on automatic measures or golden solutions, which might not be realistic in many different scenarios. In this section, we propose a general mechanism to crowdsource the evaluation of quality in the crowd. In order to build a trustworthy crowdsourcing system effectively, we need to work on two essential aspects: mechanisms to coordinate workers to guarantee the proper evaluation of quality, and a reliable mechanism to monitor the skills of workers.

Crowd-Based Quality Evaluation. In this section, we propose a general mechanism in order to guarantee quality when working with the crowd. The basic idea of our proposal is that human beings are the best quality evaluation method in many situations. Therefore, we propose to subdivide any complex task in a series of subtasks that we call *Action-Verification Units* (AV-Unit). An AV-Unit establishes a relationship pattern between the workers of the crowd to help them work collaboratively to provide a higher degree of quality.

Figure 1 depicts an AV-Unit. An AV-Unit is divided into two phases: Action and Verification. In the Action phase a single worker performs a specific action. In the Verification phase a set of workers verify the quality of the output generated in the previous

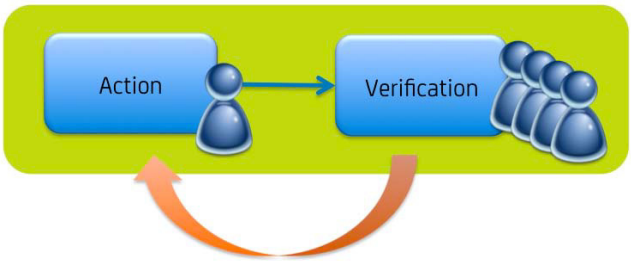


Fig. 1. Action-Verification Unit (AV-Unit)

Action phase. If the workers in the Verification phase consider that the quality provided is below a certain threshold, they might ask the first worker to repeat or improve the action. This process may be repeated iteratively until the output has reached a certain level of quality or the workers in the Verification phase decide to substitute the initial worker (or the worker is not available anymore). In practice, the Verification phase in the AV-Unit acts as a quality filter barrier, that does not allow to proceed with the process until the quality of each step in this process is approved by a set of human evaluators working collaboratively.

Figure 2 presents an example where a complex process is defined as the composition of AV-Units. It is important to remark that AV-Units do not necessarily have to be organized sequentially, and it is possible to create a complex network of interconnected AV-Units to solve complex processes.

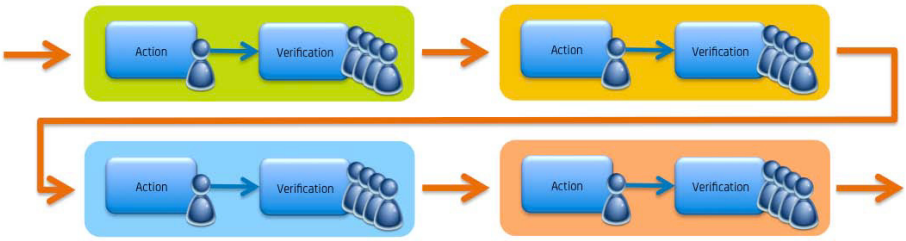


Fig. 2. Example of a use of AV-Units for general and complex problems

Measuring Trustworthiness with Worker Ranking. The success of AV-Units may depend highly on the profile of the workers. Involving many workers with poor skills in an AV-Unit, might have a negative impact on quality. In many industrial processes, quality standards are high and trusting the individuals in the crowd and their capacity to carry out the different tasks assigned to them becomes essential. Because of this, a primary concern of the system is to monitor workers in order to evaluate their skills.

Beyond AV-Units, we propose to develop crowdsourced industrial applications that take into account the quality produced by the actions done by the workers in the system and, their behaviour in general. For this, we propose to use ranking systems that are dynamically modified as the worker yields results and interacts with other workers in the system. Specifically, there are several aspects that might influence a ranking system:

- The quality measured from the output of the work produced by the workers in the crowd: it is necessary to establish rewarding and penalty measures that modify the ranking of the workers in the crowd. In general, the actions with a higher impact for workers are those performed in the Action phase of an AV-Unit. However, it would be also possible to modify the ranking of workers based on their activity when they are acting in a Verification phase.
- The behaviour of the workers in the crowd: other aspects might influence the ranking. For instance, a worker might click very fast in order to get solutions quickly and get an economical reward. Although, improper behaviours will lead with high probability to bad quality, taking into account behaviour patterns may help to multiply the positive or negative impact of action in the corresponding worker's ranking and speeding up the detection of incorrect behaviours.

The main idea behind using ranking systems is that a worker with a higher rank will be more trustworthy than other workers with lower ranks. As we mention at the beginning of this paper, establishing methods to automatically measure quality is complex, specially if we take into account that problems suitable for crowdsourcing are those which are not easy to automate.

3.2 Pay-per-Quality Rewarding System

There might be many different motivations for people to participate in a crowd-based process, ranging from their willingness to participate in a collaborative process to build something new, their will to help the community or their interest to be rewarded economically. Most industrial applications, if not all, pursue lucrative objectives. Because of this, industrial applications based on the crowd are quite more constraint in terms of motivating the crowd and tend to reward workers economically. The work presented in [13], for instance, confirms the importance of money compared to other motivations. We may still classifiy the different crowd systems depending on the rewarding model they use:

- ***Best-gets-paid systems:*** usually, in this type of systems, only the best workers get rewarded. In general, the system provides the tools to present ideas or solutions to a specific problem and a voting system for the crowd to decide the best proposals. This philosophy usually allows to reduce costs drastically and obtain very good quality, although it is in general unfair for workers, given that most of them work and are not rewarded, potentially becoming a source for lack of motivation.
- ***Pay-per-Work systems:*** in this case, workers are in general rewarded by the amount of work done. This is for instance the philosophy of Amazon's Mechanical Turk, where workers execute Human Intelligence Tasks (HITs) and get a predefined amount of money for it.

However, these two systems do not take quality into account. In this section, we propose to couple quality and the rewarding system. With this purpose, we propose to build crowd-computing systems based on a variant of a Pay-per-Work system. We call them Pay-per-Quality systems. The fundamental idea is that workers get paid for their work, but the amount that each worker receives depends on the profile of the worker. In other

words, the rewarding system depends on the rankings of the workers. In this way, a trustworthy worker will be better rewarded than an unexperienced worker, or a worker with lower skills in general. In the following section, we present an example.

3.3 Scalability

Since one of the key aspects of crowd computing is elasticity, we must provide crowd-based system with mechanism in order to cope with large problems as fast as possible. Because of this, parallelization becomes essential. There exist several paradigms in order to parallelize the execution of a task in a set of distributed computers. Among those MapReduce has become one of the most popular because of its simplicity and its capacity to scale. In [9] authors discuss MapReduce as an interesting alternative to parallelize tasks in a crowd-based system.

4 A Practical Industrial Example

For over ten years, CA Technologies has been developing and using machine translation technologies and tools to support software localization activities. Like most large software vendors, CA Technologies continuously improves its processes to reduce the cost of translation and increase the number of languages it supports. The most expensive and time-consuming phase of the localization process is the post-editing performed by human translators of the output produced by automatic machine translation, especially when this step has to be outsourced to external translation service providers.

In this section, we present a new semi-automatic management platform that allows the integration of crowd computing for reducing the cost of post-editing phase in software localization. Following the ideas presented in previous sections, in our system quality will be monitored and stored in individual records which will depend on the quality of the texts translated by each worker. Note that this problem fits the three prerequisites described at the beginning of this document: (i) machine translation does not deliver sufficient quality, (ii) translating a user guide or a user interface does not require complex training process and these training processes can be automated, and (iii) the information handled is only partially confidential.

4.1 Limitations of the Current Localization Process

Before describing the platform, we summarize the main motivations that lead to the proposal of a crowd-based system. The main limitations in the current localization process are:

- ***Long time-to-market periods for non-English versions of our products:*** products translated to languages different from English are usually released several months after the English version release because the localization process is time-consuming.
- ***Changing workload management:*** the translation workload is heterogeneous and there are some peaks during the year when a large number of products are released together. Hence, the localization teams cannot cope with the situation forcing the outsourcing of part of the work to external translation service providers.

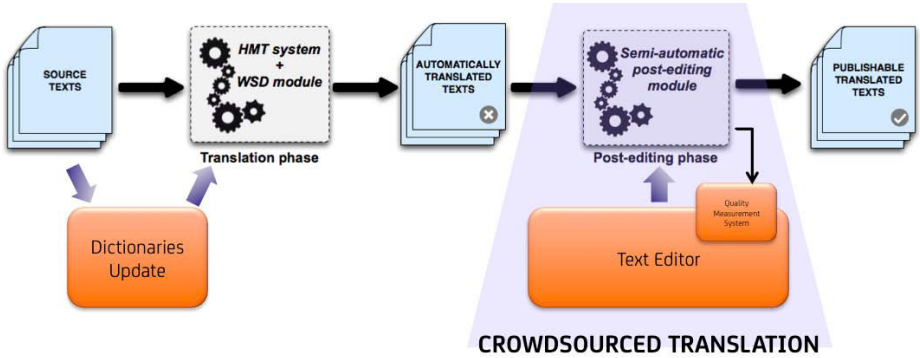


Fig. 3. Use of crowdsourcing for software localization

- ***High cost of extending our market to countries speaking languages that are currently not translated:*** in order to translate to a large number of languages, and open the possibilities for the company to explore new markets, the current approach does not work. Firstly, it is not easy to find translators for all languages. Secondly, it is economically expensive and thus unfeasible in general to hire a team of translators for every language, especially for emerging markets in some countries where the number of products sold is not expected to be huge.
- ***High cost of software localization for languages we currently translate:*** companies invest several million dollars in localization per year both in internal and outsourced localization.

4.2 Building a Crowdsourcing Platform for Software Localization

The objective of our platform is to overcome the above-mentioned issues. In Figure 3, we describe the usual process followed to localize a user guide. The source texts go through the machine translation engine and a first automatic translation is produced. Usually, the original and the machine-translated texts are sent to human translators that post-edit the text written in the target language. After this, the text is ready to be published. When the amount of work exceeds the capacity of the translators, the localization has to be outsourced. With our proposal, our goal is to crowdsource work instead of outsource. Our trustworthy crowdsourcing platform consists of:

- A model to divide a task into subtasks to be distributed among the crowd.
- A ranking function to evaluate the quality of the translations generated by a user.
- A model to organize tasks for its parallel execution based on the MapReduce philosophy.
- A quality-aware rewarding system to remunerate workers based on the quality of their work and other aspects that may range from their training to their location and native language.
- A quality-aware task sequence organization system that guarantees a minimum level of quality independently of the quality of translators, based on AV-Units.

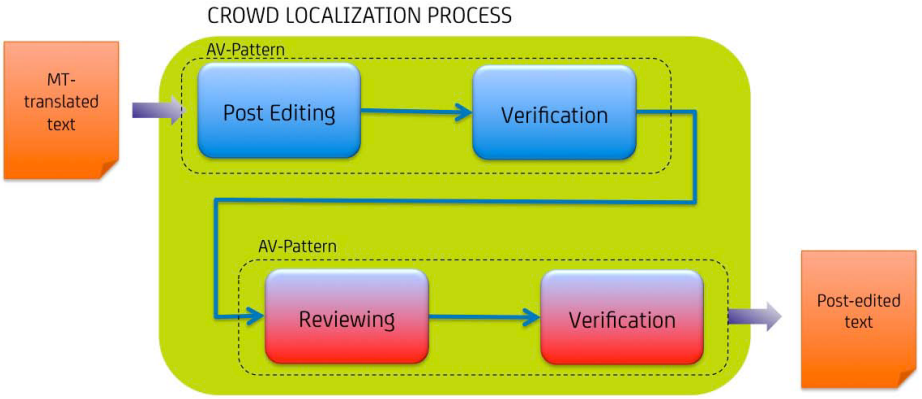


Fig. 4. Example of the use of AV-Units for the localization problems. Blue: bilingual. Red: monolingual.

We divide a task into smaller subtasks that can easily be solved by an individual worker. E-g a user guide containing roughly 100,000 words is divided into subtasks of 2000 words each. This is different from other existing approaches that divide tasks into sentence level or paragraph level subtasks thus losing the context of the text. Our selected subtask size is both convenient for a worker to solve in reasonable time and also preserves the context. Workers work on different jobs like post-editing, fixing errors, verifying, etc. We have developed a quality control system, which comprises two AV-Units to provide quality in the final translation. An important aspect of this system is that, since the quality of their work directly affects the reward obtained after finishing a task, this motivates workers to emphasize on quality. Also, the quality control algorithm adapts to the availability of workers. If high ranked workers are available, a small number of them are assigned to a task. However, if only low ranked workers are available, this number is increased.

The platform manages workers automatically. The software localization process is carried out by professional and non-professional from around the world, while the quality of translations is being maintained by the systems itself.

Quality and Scalability. In Figure 4, we show a possible division of the post-edition process into two AV-Units. In the first AV-Unit, the action consist in post-editing the text given the original text in English and the machine translation in the target language. The second AV-Unit is designed in order to review fluency and naturality in the target language of the translated text. The workers participating in the first unit are bilingual, and those participating in the second can be monolingual speakers of the target language.

Also, in order to achieve a large degree of scalability, we use a MapReduce approach. In the mapping phase, texts are run through the different phases described before. In the reduce phase, post-edited text are merged back to a single document. Figure 5, describes this process visually.

Advantages of Using Crowdsourcing in This Example. Besides the obvious benefits of crowdsourcing, namely scalability, elasticity, etc., the use of this type of crowdsourced platform reduces the cost of the software localization process significantly. It

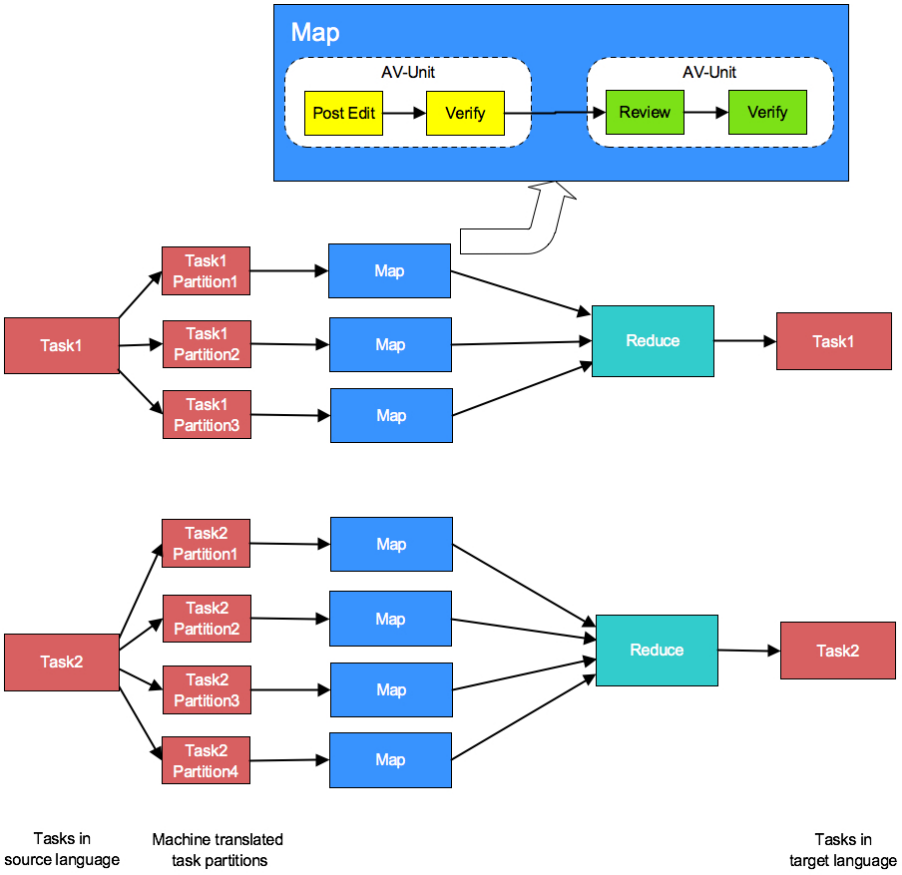


Fig. 5. Crowdsourced tasks are organized following a MapReduce strategy

solves the problem of finding translators for less popular languages because it allows for the participation of any remote translator around the world, and it improves quality. Specifically:

- ***We leverage the capacity provided by the crowd to interact one-to-one with translators, increasing quality and agility.*** Third-party systems are blackboxes, even if they are based on crowdsourcing strategies. When we work with localization service providers, quality issues usually arise. Many times this is due to the fact that translators are not familiarized with our products or even with the IT domain. When this happens, we have to send back translations to the vendor and this is very time-consuming. With this platform, we are able to make the process much more agile, since workers are exclusively trained through our products and documents and we can resubmit work directly to them, if necessary.
- ***We eliminate middleman costs.*** Existing translation service providers get a margin for their services.
- ***We have full control of all the real costs, specifically of the per-word rate paid to translators.*** Due to the socioeconomic situation of the countries speaking a certain

target language, it is a common practice that the per-word rate varies from language to language. For example, the per-word cost for a translation into German is higher than the per-word rate for a translation into Russian.

– ***We gain control in terms of confidentiality.*** Since we are sending our information to the network, security is one of the issues of crowdsourcing. By using the proposed platform, it is possible to establish your own security measures. For instance, we decide how to partition the documents and who we send information to. On the contrary, with third-party services, we rely on security measures that are not under our control, but information is still sent to the people distributed around the world.

By using this system costs are reduced, CA Technologies will gain immediate capacity to translate to any language in the world and time-to-market of CA Technology products will be significantly reduced.

5 Conclusions and Future Work

It is time for industry to start thinking about real-time and real-world crowd computing. Managing human beings automatically and helping them to collaborate with parallel machine processes will be a way to reduce costs and a competitive advantage for all those companies adopting the power of the crowd. In our dynamic world, elasticity becomes essential, human intervention is unavoidable and quality a key requirement. These three components converge to make crowdsourcing one of the most promising ideas to leverage the power of social networks. Nevertheless, new ethical issues arise. In existing systems workers are vulnerable to the whims of employers. Therefore, new legislations will have to be developed in order to create a fair work marketplace.

In the near future, new algorithms will have to be devised and more sophisticated ranking systems will be designed in order to improve the quality provided by the crowd. Several aspects such as confidentiality preservation in crowd computing systems will also become essential. In the specific area of localization, new methods have to be devised in order to preserve the style coherence on large documents.

Acknowledgements. We would like to thank Invest in Spain society, the Ministerio de Economa y Competitividad of Spain and the EU through FEDER funds for his support through grant C2_10_18.

References

1. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp. 65–72. Association for Computational Linguistics, Ann Arbor (2005), http://www.aclweb.org/anthology/W/W05/W05-0909
2. Bentivogli, L., Federico, M., Moretti, G., Paul, M.: Getting expert quality from the crowd for machine translation evaluation. In: MT Summit XIII, pp. 521–528 (2011)
3. Bernstein, M.S., Little, G., Miller, R.C., Hartmann, B., Ackerman, M.S., Karger, D.R., Crowell, D., Panovich, K.: Soylent: a word processor with a crowd inside. In: Procs. of the 23nd Annual ACM Symposium on User Interface Software and Technology, UIST 2010, New York, USA, pp. 313–322 (2010), http://dx.doi.org/10.1145/1866029.1866078

4. Denkowski, M., Lavie, A.: Exploring normalization techniques for human judgments of machine translation adequacy collected using amazon mechanical turk. In: Procs of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, CSLDAMT 2010, pp. 57–61. Association for Computational Linguistics, Stroudsburg (2010), http://dl.acm.org/citation.cfm?id=1866696.1866705
5. Gao, Q., Vogel, S.: Consensus versus expertise: a case study of word alignment with mechanical turk. In: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, CSLDAMT 2010, pp. 30–34. Association for Computational Linguistics, Stroudsburg (2010), http://dl.acm.org/citation.cfm?id=1866696.1866700
6. Geiger, D., Seedorf, S., Schulze, T., Nickerson, R.C., Schader, M.: Managing the crowd: Towards a taxonomy of crowdsourcing processes. In: AMCIS (2011)
7. Harris, C.: You're Hired! An Examination of Crowdsourcing Incentive Models in Human Resource Tasks. In: Lease, M., Carvalho, V., Yilmaz, E. (eds.) Proceedings of the Workshop on Crowdsourcing for Search and Data Mining (CSDM) at the Fourth ACM International Conference on Web Search and Data Mining (WSDM), Hong Kong, China, pp. 15–18 (February 2011)
8. Howe, J.: Wired 14.06: The Rise of Crowdsourcing, http://www.wired.com/wired/archive/14.06/crowds.html
9. Kittur, A., Smus, B., Khamkar, S., Kraut, R.E.: Crowdforge: crowdsourcing complex work. In: Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology, UIST 2011, pp. 43–52. ACM, New York (2011), http://doi.acm.org/10.1145/2047196.2047202
10. Lease, M., Yilmaz, E.: Crowdsourcing for information retrieval. SIGIR Forum 45(2), 66–75 (2012), http://doi.acm.org/10.1145/2093346.2093356
11. Negri, M., Bentivogli, L., Mehdad, Y., Giampiccolo, D., Marchetti, A.: Divide and conquer: crowdsourcing the creation of cross-lingual textual entailment corpora. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, pp. 670–679. Association for Computational Linguistics, Stroudsburg (2011), http://dl.acm.org/citation.cfm?id=2145432.2145510
12. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL 2002, Stroudsburg, PA, USA, pp. 311–318 (2002), http://dx.doi.org/10.3115/1073083.1073135
13. Silberman, M.S., Irani, L., Ross, J.: Ethics and tactics of professional crowdwork. XRDS 17(2), 39–43 (2010), http://doi.acm.org/10.1145/1869086.1869100
14. Yan, T., Kumar, V., Ganesan, D.: CrowdSearch: exploiting crowds for accurate. In: Intl. Conf. on Mobile Systems, Applications, and Services, pp. 77–90. ACM, New York (2010), http://dx.doi.org/10.1145/1814433.1814443
15. Yuen, M.C., King, I., Leung, K.S.: A Survey of Crowdsourcing Systems. In: Proceedings of the IEEE Third International Conference on Social Computing (SocialCom), pp. 766–773. IEEE (October 2011), http://dx.doi.org/10.1109/PASSAT/SocialCom.2011.203
16. Zaidan, O.F., Callison-Burch, C.: Crowdsourcing Translation: Professional Quality from Non-Professionals. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon, USA, pp. 1220–1229 (June 2011), http://www.aclweb.org/anthology/P11-1122.pdf

Multiagent Co-ordination of Wireless Sensor Networks

Maria del Carmen Delgado-Roman, Marc Pujol-Gonzalez, and Carles Sierra

Artificial Intelligence Research Institute (IIIA), Spanish Scientific Research Council (CSIC), Universitat Autònoma de Barcelona, Bellatera E08193, Barcelona, Spain
{delgado,mpujol,sierra}@iiia.csic.es

Abstract. Wireless Sensor Networks (WSNs) are generally composed of a large number of battery operated nodes with limited capacities. Therefore, a main challenge in the management of a WSN is how to reduce the energy consumption while maintaining a good quality of the sensed data. Artificial intelligence techniques like multiagent coalition formation can help on this. In this paper we propose an algorithm called *Coalition Oriented Sensing Algorithm* and test it in a realistic scenario. We experimentally show how this new algorithm allows nodes to self-organise: nodes perform a good monitoring of the environment while maximising the life span of the overall sensor network.

Keywords: Wireless Sensor Networks, Sensor Coalitions, Resourse Saving Strategies.

1 Introduction

Wireless Sensor Networks (WSNs) are networks formed by a large number of battery-operated sensing nodes to develop monitoring tasks in different environments. Each node is a low-cost, low-consumption device of limited capabilities, yet able to sense its environment and communicate wirelessly. As the nodes are cheap and easy to deploy, this technology allows to perform surveillance tasks in very large physical spaces. Moreover, the large numbers of nodes make these networks very robust to individual node failures, enabling them to operate in remote, hazardous environments. These characteristics, plus their non invasive nature, make WSNs appropriate for a great range of monitoring applications. As a result, WSNs have been applied to a number of different domains, such as environment monitoring, security control, military surveillance, and traffic control.

Depending on the application environment and its accessibility, the challenges posed by these systems can be more or less acute, especially those referred to the limited energy availability. Multiagent System (MAS) technologies can help in alleviating such constraints by introducing coordination mechanisms between sensors.

In MAS approaches, the nodes are understood as agents that can coordinate among themselves to improve their efficiency. This paper exploits that multiagent

J. Nin and D. Villatoro (Eds.): CitiSens 2012, LNAI 7685, pp. 19–32, 2013.

viewpoint to develop energy-saving data treatment strategies for WSNs. This is, nodes will coordinate to extend the life span of the network while maintaining a certain quality of the information transmitted (the main purpose of the network). In a generic scenario, the task of a sensor is to sense the environment and relay the collected data to a server node, the *sink*, where this information is further processed.

The main contribution of this paper is the Coalition Oriented Sensing Algorithm (COSA). COSA aims at exploiting the periods of invariance in (parts of) the environment. It implements a strategy for (not necessarily optimal) coalition formation in WSNs. Thereafter, only coalition leaders have to sense and transmit information, allowing the rest of the nodes to save energy. This is, COSA implements a mechanism that provides a trade-off between *information accuracy* and *energy consumption*.

As a result, the network's life span is increased at the expense of reporting less data to the sink. However, coalitions are made in such a way that the non-transmitted data do not cause a deterioration in the system performance. COSA is fully distributed in the network and robust to failures in individual nodes. Also, it assumes that the nodes are fully cooperative, as WSNs are built to serve the owner's goal.

Thereafter, we demonstrate the benefits of COSA by means of an empirical evaluation. Since deploying a full sensor network requires big investments, the experiments have been carried out in a simulation environment. Therefore, we modelled a scenario where the sensors are deployed along the course of a river, with the objective of monitoring it to detect sources of pollution. The simulation has been implemented using RepastSNS, a simulator especially designed to test sensor networks from a multiagent perspective. Further, we also run simulations where the sensors do not cooperate, sensing and transmitting data independently. The obtained results show that COSA is able to significantly extend the network's lifetime, without losing accuracy of the information received at the sink.

The rest of the paper is organised as follows. In Section 2, we revise some important contributions in the area of WSN and coalition formation in MAS. Section 3 is dedicated to the presentation and characterisation of COSA. The simulation model that we have used to test it is described in Section 4. Section 5 presents the experimental results obtained and finally, conclusions and future work are discussed in Section 6.

2 Related Work

From a MAS perspective, coalitions represent a fundamental form of organisation, as it allows the agents to organise themselves in coalitions. Agents then cooperate within the coalition in order to share resources or reach shared goals that cannot be achieved individually. Agents' association to perform a task has been considered almost from the initial conception of the MAS paradigm. The approach taken for the design of these coalition or groupal strategies have evolved

as the MAS application environments diversified. Therefore, a whole range of different coalition formation (CF) mechanisms exist depending on the conditions and characteristics of the application scenario and the nodes composing the network.

The application of CF techniques to distributed sensor networks has been investigated by numerous researchers, as it is the case of [1]. In this work, a negotiation process and individual utility calculations lead the agents to discover their organizational relationships and, according to them, to group establishment for tracking tasks.

As typically deployed in dynamic scenarios, sensor networks should be inherently adaptive. Based on this idea, the Dynamic Regions Theory was proposed in [2]. According to this theory, the network partionates itself into several regions based on the individual nodes' current circumstances and the system global policy.

The influence of the network topology structure in a MAS perfomance for task solving has also been considered in different approaches, [3,4,5]. In these cases, the system divides itself into disjoint groups in order to accomplish the demanded tasks.

In the work of [3], agents can rewire their connections to their neighbours to form better coalitions. This can be done according to their degree of connectivity or a performance-based policy. The decision factor for rewiring in [4] is the similarity among neighbours and some task and group success indicators. Finally, the work of [5] enriched the previous one by considering a more realistic coalition model. However, none of these three approaches takes into account the energy consumption and the cost derived from the rewiring pocilies.

Saving energy is one of the main objectives pursued by clustering algorithms proposed for WSNs, such as LEACH [6], EEHC [7] and HEED [8]. All these algorithms divide the sensor network distributedly into a set of non-overlapping clusters, each of them with a cluster head which is in charge of sending the collected data in the group to the sink. Our approach differs from these works in the way the cluster head is chosen, as the characteristics of the own node, its state and the perception their neighbouring nodes have of it are taken into account. A more recent approach to this problem is presented in [9], where a cluster based routing algorithm is introduced. In this case, the base station determines which the cluster heads are and implements also a centralised predictive filtering algorithm to decrement the amount of transmitted data. In contrast, we propose an approach in which the nodes make autonomous decisions without any centralised control.

In the same vein of reducing the number of transmissions, but far from the coalition/group perspective presented above,the work in [10] proposes an algorithm for individual node adaptive sampling that tries to extend the network lifetime of a glacial sensor network. This same goal is also pursued in the work of [11], in which a real deployment of an automated wildlife monitoring system is presented. In contrast to these previous works, we propose a CF strategy for homogeneous nodes in a sensor network scenario that allows to extend the useful

life time of the network by avoiding redundant sensing and transmission. This group formation strategy is based on the nodes' state and the conditions of the environment. There is no intervention of any central authority and the algorithm is fully distributed and embedded in the nodes' behaviour. The main objective of the algorithm is achieved by allowing nodes in a coalition to delegate their sensing tasks to other neighbouring nodes, while restricting the maximum information loss, therefore the initial purpose of the system —faithfully monitoring the environment— is not missed.

3 Algorithm Description

The *Coalition Oriented Sensing Algorithm* (COSA) has been designed considering an scenario composed of a set $A = \{a_1, \ldots, a_N\}$ of cooperative and homogeneous agents (the network's nodes). We do not consider that agents can be competitive or selfish as in this kind of problems there are neither resources to fight for nor rewards to be won by the agents.

The basic behaviour of an agent a_i is to sense the environment and relay the observed measures to a server or *sink*.

As explained above, COSA's objective is to save system resources by allowing agents to form coalitions of agents that are perceiving very similar measures. Thereafter, a single agent can act as a representative for the whole coalition, avoiding redundant sensing and saving resources. To find an appropriate distribution of the agents in coalitions, we take into account the similarity of the individual measurements and the topology of the neighbourhood structure, which determines the neighbourhood relationships to be established among agents.

A *coalition structure* $c = \{g_k\}_{k:1..K}$ is defined as a partition of A in K groups. The criterion that guides the formation of the different coalition structures is to find (in a distributed manner) the best partition so that the energy consumption of the system is somehow minimised, while the accuracy of the information sent to the sink is constrained to a certain range. COSA appears as a tuneable algorithm thanks to the definition of a set of parameters p (to be explained later) whose values drive the agents' behaviour. Depending on p, agents take different kinds of sampling and *transmission actions*, represented as $m^j \in M_p$, where M_p is the set of existing actions available for that p configuration.

The objective of minimising the system's energy consumption is formally expressed in Equation 1. According to this equation, we try to find an optimal set of parameters $p*$, where m_i^j is the action j taken by agent i and E_j represents the energy consumption associated to that action. Measurements' accuracy is guaranteed through an adequate p parameters election.

$$p* = \arg\min_{p \in P} \Delta E = \arg\min_{p \in P} \sum_{m^j \in M_p} \sum_{a_i \in A} \#m_i^j E_j \tag{1}$$

Group formation among agents is based on a peer-to-peer negotiation protocol by means of which agents exchange information about their measurements and

their adequacy to represent their neighbours. As a consequence of this negotiation, agents assume one of two possible roles : *leader*, if it is the representative of its coalition (where it may be the only member); or *follower*, if it joined a coalition lead by another agent. The main concepts that drive this negotiation are *adherence degree* and *leadership attitude*.

The *adherence degree* of an agent i to an agent j is a measure that indicates how much agent i intends to form part of a group led by agent j. The higher the degree, the higher the intention. The *adherence degree* is defined as the product of two factors. To evaluate those factors, we assume that the variable under observation follows a Normal distribution, $\mathcal{N}$. On the one hand, the first factor in the adherence equation (2) captures the similarity between the measurements of agents i and j. To avoid unproductive calculation, this factor is only defined for neighbour agents whose measurements verify that $\|x_j - x_i\| \leq d_{max}\sigma_j$, where d_{max} is a parameter and x_i, σ_j are the corresponding sample and deviation of agents i and j. On the other hand, the second factor captures the *goodness* of the neighbour's distribution and avoids obtaining high adherence values to neighbours with wide distributions. To achieve that, this factor restricts the evaluation to those neighbours whose σ belongs to the interval $(\sigma_{min}, \sigma_{max})$, through the evaluation of the distribution's entropy normalized on that range.

As a result, the evaluation of the degree to which an agent a_i may be interested in being led by one of its neighbours a_j is calculated as follows:

$$adh(a_i, a_j) = \frac{p(x_i, \mathcal{N}_j(\bar{x}_j, \sigma_j))}{p(\bar{x}_j, \mathcal{N}_j(\bar{x}_j, \sigma_j))} \cdot (1 - \frac{e^{H_j} - e^{H_{min}}}{e^{H_{max}} - e^{H_{min}}}) \tag{2}$$

Note that the set of p parameters, as presented previously, can be identified now as $p = \langle d_{max}, \sigma_{min}, \sigma_{max} \rangle$ defined over the space $p \in \Re^3$. As previously stated, the set of values to which these parameters are set influences the actions that a node can take.

When an agent receives an adherence value from a neighbour, it has to decide whether it is interested in becoming the leader of this agent or not. Let us call $P(a_i)$ (potential group) the group formed by a_i and the agents willing to become part of a group led by a_i. The attitude of a_i as a leader of this group depends on different factors that can be identified in (3). The first factor is called *prestige* and it is an average of the adherence level of the group's members. The *capacity* factor indicates the available energy of the node to act as a leader. This value is derived from the current energy level of the node minus the security energy level (E_{sl}) divided by the maximum energy level available E_{max}. E_{sl} defines the minimum energy that the node has to keep to ensure sending one last message before completely depleting its battery.

Finally, the last factor in (3), *representativeness*, indicates how well the potential leader's measurement fits as a representative of the potential group agents' measurements. So, a_i characterizes the set of data received together with its own data, that is, the set $\{x\}_{P(a_i)}$, with their mean and standard deviation, noted as $(\bar{x}_{P(a_i)}, \sigma_{P(a_i)})$. To encourage the formation of groups with very similar measurements, an exponential function establishes the divergence growing ratio.

Those potential groups whose measurement distribution is very disperse are also penalized through the inclusion of the Pearson's coefficient in the equation.

A good group leader is an agent who has enough energy and whose measurements are similar enough to the measurements of the other group members. In summary, the leadership capacity of an agent a_i for its potential group $P(a_i)$ is calculated as follows:

$$\frac{\sum_{a_j \in P(a_i)} adh(a_j, a_i)}{N} \cdot \frac{E(a_i) - E_{sl}}{E_{max}} \cdot \frac{1}{e^{|x_i - \bar{x}_{P(a_i)}| CV_{P(a_i)}}} \tag{3}$$

This information exchange takes place following a certain operational protocol according to which, every agent involved in a negotiation with a neighbour goes through these phases:

- *Sample information exchange.* This process corresponds to the variable sampling and measurements broadcast.
- *Adherence graph construction.* Once the agent has calculated the adherence degrees to its neighbours, it communicates the maximum adherence value to the corresponding most preferred neighbour.
- *Leadership information exchange.* Based on the current adherence relationships, the agent calculates and communicates its attitude as a leader towards the agents willing to adhere to it.
- *Group definition.* Depending on the information available for an agent at a certain moment, it decides wether to stay in its current group (as a leader or dependant of a leader node), to leave this group to join a different one or to constitute its own group.

The set of performatives that the agents use to complete these stages are:

- *inform*: to indicate the transmission of data (measurements, maximum adherence and leadership values).
- *firmAdherence*: to express the desire of the sending node to adhere to the addressee node.
- *ackAdherence:* to express the acknowledgment to a previously received firmAdherence message.
- *break*: for a leader node to break a leadership relationship.
- *withdraw*: for a dependant node to break a leadership relationship.

The CF protocol is embedded in the agent generic behaviour. Agents behave in a proactive and reactive way. Proactive because the core behaviour of an agent is the continuous process of looking for the best group of neighbours that matches with its measurement and its state. To achieve this objective, an agent exchanges messages asynchronously with its neighbours. Reactive because their acts and decisions are triggered by the observation of the environment and the information they receive.

4 Simulation Model

Our experiments run over RepastSNS [12], an event-based simulator especially designed to model sensor networks as multi-agent systems. In RepastSNS, all the environment objects are modelled as agents that communicate via message passing. The platform is open source and developed in Java over Repast. It is designed as a two-layered structure with an object layer and a network layer. The object layer defines the behaviour of the individual sensors and the network layer defines the topology and the relationships among the sensors. The simulation platform is an extension of Repast classes, so the program structure of RepastSNS fits into this known MAS simulation engine.

All these characteristics make RepastSNS a general purpose simulation environment, that allows different application domains for WSN to be tested over it. This can be done without too much effort, as the environment provides a scalable and extensible infrastructure to build up networks of basic WSN components. The main task that a programmer has to do over this environment is just to configure and adapt its pre-defined elements (observable phenomena, sensors, agents, communication mechanisms) to the specific domain being modelled.

The advantage of using RepastSNS instead of any other network simulator such as ns2, OMNeT or J-SIM is twofold: (i) it provides for a more abstract level description than these other simulators, allowing the programmer to concentrate on the actual agent behaviour instead of dealing with hardware details; and (ii) it brings with it a convenient basic implementation of all the components needed to model wireless sensor networks.

To avoid confusions, from now on we will use the term *node* when we refer to a wireless sensor device and the term *sensor* when we refer to the specific physical device that measures a parameter of environment.

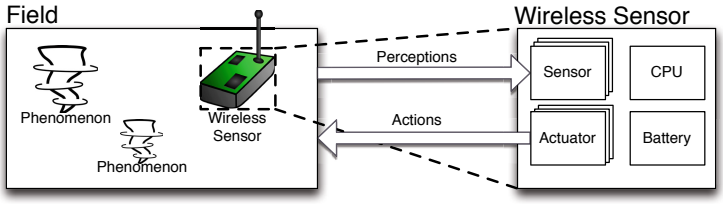


Fig. 1. RepastSNS simulation architecture

Figure 1 outlines the architecture of a sensor network simulation on RepastSNS. In RepastSNS all the observable phenomena are contained inside a *field* that includes the nodes themselves. Furthermore, the nodes are composed of multiple modules: a cpu, a battery, and any number of sensors and actuators. Sensors are those devices that allow the node to perceive the field's phenomena and their properties. Analogously, actuators allow the cpu agent to modify existing phenomena or produce new ones. This very simple model is surprisingly sound, as any phenomena or agent behaviour needed in a system can be easily

modelled and incorporated. For instance, wireless radio interfaces can be modelled as an actuator that generates wireless waves (a phenomenon), plus a sensor that detects them.

5 Experimental Results

To demonstrate the proposed MAS algorithm, the experimental evaluation compares a WSN performance when it implements COSA and when it behaves according to a random sampling policy. Our main aim is to compare the energy usage of both approaches, as well as the accuracy of the data reported in each case. The two approaches deliver different data as the random sampling scheme implies that measurements taken from the environment are directly send to the general server, while the characteristic grouping imposed by COSA translates in association data sending.

5.1 Experimental Setting

COSA algorithm aims at faithfully monitoring the state of a dynamic environment and extending the life time of the network as much as possible. To test this, the scenario considered is that of a river, whose state is to be monitored. Different state variables and phenomena that could be sensed in this domain are water temperature, salinity or hydrocarbon presence. The deployment of the sensor network in such an environment can rely on the buoys and signalling elements deployed along the course of the waterway.

To correctly fit this application domain into the simulation platform, we start by defining a river phenomenon. This phenomenon represents a river section of 50 kilometers long by 2 kilometers wide, and it is composed of a grid of water cells. In order to mimic the effects of water flowing through the river, we define a simple river movement schedule that will cause that any phenomenon appearing in the river will be displaced by the current of the water. The model used to implement this functioning considers a drift component, a sedimentation component and a solvent component, i.e. the general intensity of the phenomenon reduces in time and a part of it remains in its origin, while the rest flows according to the strength of the current. Therefore if any contaminant is poured in a water cell, it will spread to its downward cells through time according to the following model $River(x,y) = (1-\rho)River(x,y)+\rho(\alpha(River(x-1,y-1))+\beta(River(x,y-1))+\gamma(River(x+1,y-1)))$. During each simulation run, three different contamination sources appear at random locations, but at specific times and keep spewing contaminant between 30 and 60 weeks.

The surveillance nodes are the ones responsible of monitoring the river's condition and informing the sink about their observations. As explained before, they are formed by a CPU, battery, sensor and radio. These components are modeled after Waspmote ones, real wireless sensor devices whose specifications are summarized in Table 1.

Two different kinds of surveillance nodes are considered according to the two sampling approaches proposed. Regardless the sampling policy nodes implement,

Table 1. Node components specifications

Component	Specification
Battery capacity	13000mAh@3.7V
CPU active consumption	9000uAh@3.3V
CPU sleep consumption	62uAh@3.3V
CPU hibernate consumption	1uAh@3.3V
Radio transmission consumption	210000uAh@3.3V
Radio reception consumption	80000uAh@3.3V
Radio bandwidth	156Kbps
Radio Sensing radius	1.5km
Radio sleep consumption	60uAh@3.3V
Sensor consumption	6uAh@3.3V
Sensor sampling time	1.63s

both kind of nodes have to send the collected data to a sink node. This sink node represents the central monitoring station to which the nodes deployed along the river are reporting to. In our setting, this agent has the ability to obtain the actual contamination values at any time at every point in the river, and can therefore determine the differences between node-reported values and the real ones. Differently from sensing nodes, the sink node does not take samples from the environment, neither is it constrained to low power or low processing capacity, as it acts as a server in the system, being part of the network control unit.

As previously explained, the system basic functioning consists of monitoring the environment through periodical samples collection. The way nodes deployed in the scenario behave to satisfy this purpose is what defines the applied sampling policy. The base case considered for the experimentation set is that of a random policy. The random setting presents a set of nodes (called random nodes) that take a sample from the environment at a random moment within the sampling period specified for the network.

The so called cf nodes (coalition formation) act according to the COSA algorithm presented in previous sections. Therefore, these nodes sample the environment periodically as expected, but instead of directly sending every individual measurement to the sink, these measurements are used to establish peer to peer negotiations with neighbourghs so that sensing groups can be formed. Consequently, only one node for each group (the leader) senses and sends information to the sink on behalf of the others, which delegate their tasks on it for a certain period of time.

The sink node in both scenarios receives the information collected by active nodes. In the random scenario, this information corresponds to every single measurement periodically collected by all the random nodes. The sensing task delegation among cf nodes may cause the sink to receive a group measurement representing the information associated to a set of nodes in a group. Assuming a common sample for a set of nodes may cause the loss of some pieces of information and consequently, some noise is added to the reported data.

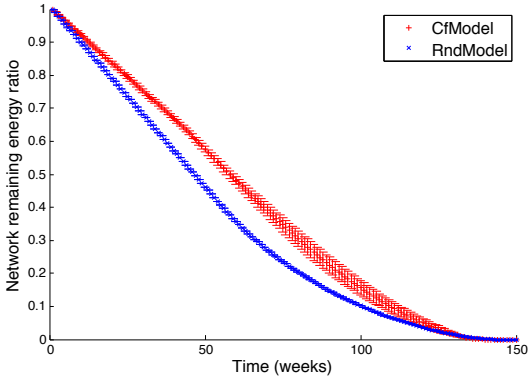


Fig. 2. Network remaining energy ratio

Finally, to completely define the experimental setup, Table 2 presents the values assigned to the COSA algorithm parameters as well as the sampling frequency demanded to the system. The number of nodes considered is set to 50 and their deployment along the river course is assumed to follow a regular chain distribution. Every node is situated in the middle of the river section considered and evenly spaced.

Table 2. Parameters' values

Parameter	Value
Sampling frequency	10min
d_{max}	1.75
σ_{min}	0.0005
σ_{max}	6
Node asleep time	1day

5.2 Results

To test if COSA achieves the objectives that inspired its definition, we study its performance in terms of the energy consumption and the quality of the reported information. The reference base for these two gauges are supplied by the random nodes' behaviour, as they follow a dummy sampling policy.

All the experiments have been run until every node in the network has completely depleted its battery, that is, for our experimental setting, 140 weeks. Figure 2 shows the ratio of network remaining energy for both kind of nodes, cf nodes and random nodes.

In this figure, as initially expected, we can observe how the COSA algorithm allows the network to keep a higher level of global energy than the random policy during most of its life time; however, both sampling policies lead to a very similar network death time.

The energy consumption curve obtained for the random nodes follows a stable pattern, whereas cf nodes present more variability, especially by the end of the experiments. This phenomenon is because COSA causes different group configurations to appear in the network over time. The influence of the group configuration reached in the network grows as the global energy in the system decreases. At these middle-end stages, the leader node situation and the available energy of the set of nodes still alive have a severe impact on the global energy level.

In contrast, the results obtained for random nodes clearly show the effect of their independent sampling behaviour. This is neither affected by their neighbours' state, nor by the dynamics of the environment. The decreasing energy curve, therefore, shows the effect of nodes dying, dependent on their distance to the sink. The extension of the useful life span of the network can be better identified in Figure 3.

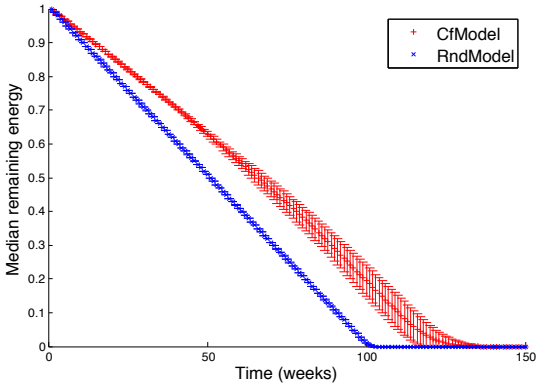


Fig. 3. Network median remaining energy

This figure shows the median of the nodes' energy values per week. We observe that half of the random nodes are already disabled by week 101, whereas this same value is reached over 30 weeks later for cf nodes (specifically by week 134). This result translates directly into better system performance during the network lifetime. COSA causes nodes' death to be evenly distributed, which guarantees that the network is going to get a fairly good representation of the whole environment, for most of its life time.

This result relates to the previous figure, as the nodes that deplete their battery first are the more distant ones to the sink. This means that the sink is blind to this area. Random nodes situated there are the first to die, while the ones situated near the sink keep most of their battery power and are the last to die, but are only able to sample the sink's surroundings. The even battery depletion in cf nodes causes a more simultaneous node death phenomenon, but in terms of global energy in the system, both policies reach zero level at the same time.

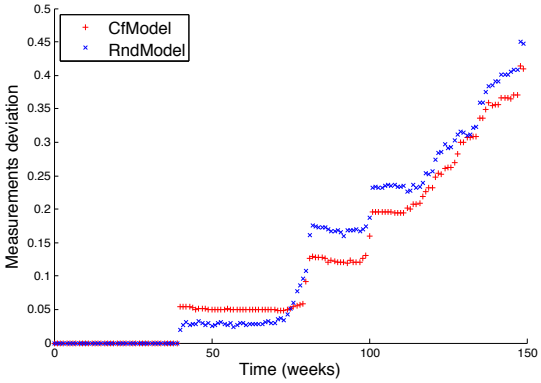


Fig. 4. Reported information deviation

Figures 4 and 5 represent the deviation of the information reported to the sink by random and cf nodes. When no pollution phenomenon has appeared yet, there is no deviation from the real environment state, but as a first pollution phenomenon appears (at week 40), the deviation value of the reported information changes. As both models have all their nodes fully operational by this time, we see that the resulting deviation of both models' reported samples is quite low (in fact, the lowest deviation from real phenomenon values is reached during this period).

Between weeks 40 and 80, the deviation corresponding to the data reported by cf nodes is slightly higher than the corresponding to the random nodes in the same period, with the maximum difference between them of only 0.035. Although both models provide really good representations of the phenomenon, cf nodes data is a little more deviated from reality due to the characteristic group sampling of COSA, which allows a leader node to send a sample on behalf of its dependent nodes. However, as random nodes begin depleting their batteries and becoming unable to sense, the deviation of the information they report quickly deteriorates. Therefore, by the time the second pollution event takes place (week 80) the deviation of random nodes' information suffers a bigger increase than the corresponding leap that cf nodes' deviation takes.

The same behaviour repeats for the third pollution stain, appearing in week 100. Again, the deviation of the data reported by cf nodes is lower than that offered by random nodes. Moreover, the smaller leaps in the deviation resulting from cf nodes' data indicates that they are also more stable, and therefore, more robust to nodes' failure or exhaustion.

This property can also be observed in Figure 5, which shows mean and dispersion values corresponding to the deviation reported by both kinds of nodes. The fact that the standard deviation associated to this gauge increases as the number of nodes decreases supports the hypothesis of robustness for cf nodes for different environmental conditions and lower number of nodes.

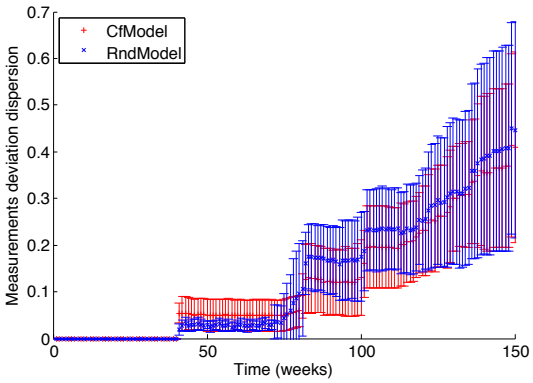


Fig. 5. Mean and standard deviation information accuracy

6 Conclusions

In this paper we have presented the COSA algorithm and have given experimental evidence of its computing properties. This algorithm is aimed at extending the life span of WSNs while guaranteeing their good performance. In contrast to previous approaches that tried to save energy using adaptive sampling schemes, COSA innovates by reaching this objective via a peer to peer negotiation protocol. This negotiation protocol enables nodes to interact and generate groups that produce a network-wide benefit. To attain a good group configuration the algorithm relies on the node local information about its environment state and neighbouring nodes. This local information together with the appropriate COSA algorithm parameter configuration leads to the formation of groups of nodes that act as a single entity, avoiding redundant sensing and transmissions efforts.

The improvements obtained by this algorithm have been shown in a simple scenario representing the section of a river where different pollutant phenomena appear in random positions. Simulating the scenario required the development of the simulation platform RepastSNS, which represents a powerful tool for WSN simulation from a MAS perspective.

The results obtained for the experimentation showed how a sensor network whose nodes implement COSA guarantees a better use of the network energy and a more homogeneous system energy depletion than the ones offered by a sensor network whose nodes follow a simple random sampling policy. Achieving this more regular system exhaustion reverts in the extension of the useful life of the network, as the whole monitoring area can be sampled for longer periods, therefore, getting a more accurate view of the environment.

As future work, we plan to test the behaviour of COSA in different scenarios and for different network topologies. We believe that the COSA parameter configuration highly depends on the dynamics of the phenomena being observed and the distribution of the nodes in the environment. Getting to know the impact of these two factors in the algorithm performance will allow us to fully characterise

it and to be able to identify the set of cases for which its use would result beneficial. Even more, being able to assess the improvement expected, which would result in better WSN deployment planning.

Acknowledgments. This work has been supported by the Agreement Technologies project (funded by CONSOLIDER CSD 2007-0022, INGENIO 2010).

References

1. Sims, M., Goldman, C.V., Lesser, V.: Self-organization through bottom-up coalition formation. In: Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS 2003, pp. 867–874. ACM, New York (2003)
2. Mac Ruairí, R., Keane, M.T.: The dynamic regions theory: Role based partitioning for sensor network optimization. In: Proceedings of the Sixth International Joint Conference on Autonomous Agents and Multiagent Systems (2007)
3. Gaston, M.E.: desJardins, M.: Agent-organized networks for dynamic team formation. In: Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS 2005, pp. 230–237. ACM, New York (2005)
4. Barton, L., Allan, V.H.: Methods for Coalition Formation in Adaptation-Based Social Networks. In: Klusch, M., Hindriks, K.V., Papazoglou, M.P., Sterling, L. (eds.) CIA 2007. LNCS (LNAI), vol. 4676, pp. 285–297. Springer, Heidelberg (2007)
5. Glinton, R., Scerri, P., Sycara, K.: Agent-based sensor coalition formation. In: 2008 11th International Conference on Information Fusion, pp. 1–7 (July 2008)
6. Heinzelman, W.R., Chandrakasan, A., Balakrishnan, H.: Energy-efficient communication protocol for wireless microsensor networks. In: Proceedings of the 33rd Hawaii International Conference on System Sciences, HICSS 2000, vol. 8, p. 8020. IEEE Computer Society, Washington, DC (2000)
7. Bandyopadhyay, S., Coyle, E.J.: An energy efficient hierarchical clustering algorithm for wireless sensor networks. In: Proceedings of IEEE INFOCOM 2003, pp. 1713–1723 (April 2003)
8. Younis, O., Fahmy, S.: Heed: A hybrid, energy-efficient, distributed clustering approach for ad hoc sensor networks. IEEE Transactions on Mobile Computing 3, 366–379 (2004)
9. Cordina, M., Debono, C.J.: Maximizing the lifetime of wireless sensor networks through intelligent clustering and data reduction techniques. In: Proceedings of the 2009 IEEE Conference on Wireless Communications & Networking Conference, WCNC 2009, pp. 2508–2513. IEEE Press, Piscataway (2009)
10. Padhy, P., Dash, R.K., Martinez, K., Jennings, N.R.: A utility-based sensing and communication model for a glacial sensor network. In: Proceedings of the Fifth International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS 2006, pp. 1353–1360. ACM, New York (2006)
11. Dyo, V., Ellwood, S.A., Macdonald, D.W., Markham, A., Mascolo, C., Pásztor, B., Scellato, S., Trigoni, N., Wohlers, R., Yousef, K.: Evolution and sustainability of a wildlife monitoring sensor network. In: SenSys, pp. 127–140 (2010)
12. IIIA-CSIC: Repast sensor network simulation toolkit (2012), http://www.iiia.csic.es/~mpujol/RepastSNS/

On the Protection of Social Network-Extracted Categorical Microdata

Jordi Marés[1] and Vicenç Torra[2]

[1] Artificial Intelligence Research Institute (IIIA),
Spanish Council of Scientific Research (CSIC),
Universitat Autònoma de Barcelona (UAB), Spain
jmares@iiia.csic.es
[2] Artificial Intelligence Research Institute (IIIA),
Spanish Council of Scientific Research (CSIC), Spain
vtorra@iiia.csic.es

Abstract. Social networks have become an essential part of the people's communication system. They allow the users to express and share all the things they like with all the people they are connected with. However, this shared information can be dangerous for their privacy issues. In addition, there is some information that is not explicitly given but is implicit in the text of the posts that the user shares. For that reason, the information of each user needs to be protected.

In this paper we present how implicit information can be extracted from the shared posts and how can we build a microdata dataset from a social network graph. Furthermore, we protect this dataset in order to make the users data more private.

1 Introduction

With the continuous growing amount of public available data, individual privacy has become a very important issue to deal with because several agencies are collecting a huge amount of data from people daily. This data is very valuable for the knowledge of our society status but it is also dangerous in terms of privacy. Data privacy field tries to protect all the public data sources in order to allow the data extraction but taking into account the individuals privacy. Until a few years ago, the major part of the data was collected via surveys. However, nowadays there is a new place to take data much more easy and much less controlled: the online social networks.

Social networks have become a very important part of the people's communication system and, as most sociologists agree, this online social interaction will not fade away [18]. People use these networks to express all their feelings, emotions or simply to meet people who have the same hobbies or interests. It can be seen that all this information is sensible and is related to a single user profile. Therefore it can be dangerous to collect this kind of data and publish it without protection. An example of the need to protect social networks can be found in [19] where it says that epidemiology researchers use social networks to study the social network structure and epidemic phase in sexually transmitted disease. In addition it should be noticed that not all the information is explicitly given by the user profile. There is some information that is implicitly hidden into the posts the user shares in his profile such as the main topics of interest of the user.

J. Nin and D. Villatoro (Eds.): CitiSens 2012, LNAI 7685, pp. 33–42, 2013.

Although there have been several approaches to protect the user anonymity modifying the social graph structure adding or removing edges [12][23], there are less approaches to deal with the privacy in the semantic data included in the graph nodes [4]. The most well known model to protect social graphs is k-anonymity which is a very popular model for microdata datasets protection [14] and it has been adapted to graphs [9] and relies on the property that every node will be indistinguishable with at least $(k-1)$ nodes.

In this paper we present a way to protect a real online social network-extracted microdata dataset with explicit and implicit information about Twitter users using a k-anonymity protection method. Several approaches have been developed to protect microdata datasets [2][17][20] in order to achieve enough protection to prevent attacks to the confidential information about individuals from the disseminated data.

Regarding the data in the microdata datasets, there exist two types: categorical and continuous. In our case, we focus on categorical data. The problem of categorical data over continuous data is that there are less actions to perform in the protection process because arithmetic operations are not allowed here, so the only actions allowed with categorical data are the exchange of categories by others that already exist, suppression of category, and generalizations of some categories into new ones. This lack of possible operations makes the protection a difficult task.

Protection methods are typically evaluated using two measures: information loss and disclosure risk. Information loss [17] checks the quantity of data that has been harmed during the protection process and therefore is no longer useful. Disclosure risk [6][21][22] measures the quantity of original data that can be discovered through the protected data.

The remaining of this paper is structured as follows. In Section 2 we explain the methodology followed to go from a real social network like Twitter to obtaining a microdata dataset with explicit and implicit information about users. Section 3 contains the description of the protection method used in this work to protect the microdata dataset: the microaggregation. In Section 4 we present the measures used to evaluate the quality of the protection. Section 5 shows the results of our experiments comparing privacy and utility in the original microdata dataset and the generated protections. In Section 6 we make some concluding remarks. Finally, in Section 7 we describe our next steps to do as a future work.

2 Social Network-Extracted Microdata Generation

In this section we describe the methodology we used in order to extract a microdata dataset from a real online social network like Twitter.

2.1 Crawling Algorithm

The first step to take is to build a crawler in order to get information about connected users in the social network. Algorithm 1 shows the steps followed by our crawler.

Algorithm 1. Twitter Profiles Crawling Algorithm

Input: uID Initial user id, $numUsers$ Maximum number of user to crawl, $numTweets$ Number of tweets to get from each user.
Output: Y List of public available data for each user.
id $\Leftarrow$ uID
actualUser $\Leftarrow$ $getDataFromUser$(id,numTweets)
unvisited$\Leftarrow$ $getFollowingUsers$(actualUser)
visited $\Leftarrow$ [id]
$Y \Leftarrow$ [actualUser]
while ($|unvisited| > 0$) and ($|visited| < numUsers$) **do**
  id $\Leftarrow$ $getRandomId$(unvisited)
  actualUser $\Leftarrow$ $getDataFromUser$(id,numTweets)
  unvisited.$remove$(id)
  newRemaining $\Leftarrow$ $getFollowingUsers$(actualUser)
  unvisited.add(newRemaining)
  visited.add(id)
  Y.add(actualUser)
end while
return Y

The algorithm is started with a given initial user id as the starting node in the social network, a maximum number of users we want to get information from, and a number of tweets we want to get from each user. Then, we use the Twitter API [3] to get user data such as location, hashtags, urls, following users, and tweets posted by the user. Three lists are used: *unvisited* contains the ids of the not yet crawled users connected to the already crawled ones, *visited* contains the ids of the already crawled users, and *Y* contains the data structures containing all the information about each crawled user.

This is executed in a loop until we reach the maximum number of users we wanted to crawl or until we have no more users in the *unvisited* list.

After this step we have a collection of structures containing information about each user.

2.2 User Profiles Generation

The second step to do is to use the data structures collected by the crawler in order to get a profile for each user containing his location, his connected users and, his three most relevant topics of interest. In order to do this it should be noticed that information is not always explicitly given in the social networks. That is, using the Twitter API we can get the location but it is not possible to get the topics that a user is interested about because they are not specified nor described anywhere. However, these topics can be extracted using natural language processing techniques on the text of the tweets shared by the user.

In order to process the information contained in the tweets we used Web services provided by OpenCalais [15], which allow for the extraction of entities such as people, organizations or events and moreover assign topics to a piece of text. In this work we only used the topics categorization capacities of OpenCalais. The 18 possible topic output

values are: Business_Finance, Disaster_Accident, Education, Entertainment_Culture, Environment, Health_Medical_Pharma, Hospitality_Recreation, Human Interest, Labor, Law_Crime, Politics, Religion_Belief, Social Issues, Sports, Technology_Internet, Weather, War_Conflict and, Other.

Our first approach was to apply directly the OpenCalais Web services to the tweets text. However, as tweets are very short pieces of text (maximum of 140 characters) it was very difficult to extract topics and we got a very high percentage of users without any topic of interest found. Then, as a second approach, we used the urls within the tweets texts to enhance their semantics following the approach described in [1].

In this work, we do not use the hashtags because most of the times they are written in a useless form such as $\#ToMyFutureKids$. This forms do not provide any information to us and therefore we decided to not use hashtags but use the web pages shared in the tweets, which are much more rich semantically.

To do this, we executed two times the OpenCalais Web service to check the topics found in the tweet text and also in the text of the website shared inside the tweet. Then, the topics found in both executions were merged. At the end of processing all the tweets from a given user, the three most frequent topics were the ones taken as a result. By doing this we obtained a higher number of topics per user. The final topics also kept the level of interest for each topic because we took the most frequent one as the main topic of interest for the user, the second most frequent is the second main topic of interest, and the same happens for the third.

At the end of this profiles generation step we have a set of user profiles containing the location of a user, the users who is connected with, and the sorted three major topics of interest. So, as a result we obtained profiles combining explicit information given by the Twitter API calls and implicit information extracted from the tweets shared by the user using natural language processing tools.

2.3 Graph Generation

As a third step, after generating the users profiles, we generated the social graph connecting all the users with the ones they are following in the real social network. Figure 1 shows the resulting graph representing the relations between users.

It can be seen that there are more density of edges in the center of the graph than in the borders. This is because when we crawled the social network we kept a list of remaining users to crawl which were connected users to already crawled users. This fact gives higher probabilities to the first crawled users to expand more their neighbors than to the last crawled users.

Then, as the initial user we crawled is in the center, all the users near to him had much more attempts to expand their neighbors than the users in the borders which are the newest ones.

2.4 Microdata Dataset Construction

Finally, once we have a social graph where each node has information about a single user and each user is connected to his real following users, it is possible to extract all this information from the nodes and generate a microdata dataset.

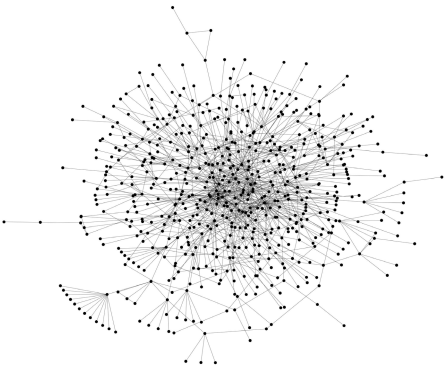
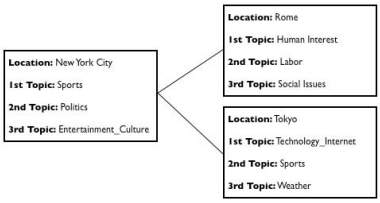


Fig. 1. Graph generated from the crawled users profiles

In order to do this we extracted the information of each node placing it in a single row of the dataset. Then, the resulting dataset file has one row per user and one column per attribute. In our case we used five attributes per user: the degree of the user node, the location of the user, the main topic of interest, the second main topic of interest, and the third main topic of interest. Figure 2 shows an example of microdata dataset construction from a graph with three user profiles.



Degree	Location	1st Topic	2nd Topic	3rd Topic
2	New York City	Sports	Politics	Entertainment_Culture
1	Rome	Human Interest	Labor	Social Issues
1	Tokyo	Technology_Internet	Sports	Weather

Fig. 2. Example of microdata construction

At the end of this step we have a real social network-extracted microdata dataset with either explicit and implicit information about the users. This kind of datasets would be very interesting for research purposes but they must be protected before publishing it.

3 The Microaggregation Protection Method

In this section we present the microaggregation protection method that is the one we have used to protect the microdata dataset in our approach.

In microaggregation [5][16][8], records are clustered into small aggregates or groups of size at least k. Then, instead of publishing an original variable V_i for a given record, the median of the values of V_i over the cluster to which the records belongs to is published.

To define the microaggregation procedure we need to define how to compute the distance between two categories when we create the clusters. This distance is defined in a different way when the variable is nominal than when it is ordinal because of the possibility of sorting the categories in the ordinal case, what is not possible in the nominal case.

For a nominal variable V the distance between two categories is defined as follows

$$d_{nominal}(c, c') = \begin{cases} 0 & \text{if } c = c' \\ 1 & \text{if } c \neq c' \end{cases} \tag{1}$$

and for an ordinal variable

$$d_{ordinal}(c, c') = \frac{|c'' : (c, c') \leq c'' \leq max(c, c')|}{|D(V)|} \tag{2}$$

where c is a category in the original dataset and c' is the category corresponding to c in the masked dataset, and $D(V)$ is the domain of variable V. Then, the ordinal distance will be the computed as the number of categories between c and c', divided by the total number of categories for the attribute V.

There exist several approaches for the microaggregation clustering. In this work we used the MDAV-generic described in [8] because it can work with any type of attribute, aggregation operator and distance. Algorithm 2 shows the algorithm of this method. Basically, MDAV create clusters of size k around the two most distant records in the dataset, leaving a final cluster with at least k records.

4 Protection Evaluation Measures

After protecting a microdata dataset it must be evaluated in order to assess the quality of the protection. In this paper we used the two main measures used in the microdata protection field: the information loss and the disclosure risk.

Information loss is known as the quantity of harm that is inflicted to the data by a given masking method. This measure is small when the analytic structure of the masked dataset is very similar to the structure of the original dataset. Then, the motivation for preserving the structure of the dataset is to ensure that the masked dataset will be analytically valid and interesting. In this work we used the *contingency table-based information loss* (CTBIL)[17], the *distance-based information loss* (DBIL)[17], and the *entropy-based information loss* (EBIL)[10].

Assessment of the quality of a protection method cannot be limited to information loss because disclosure risk has also to be measured. Disclosure risk measures the information can be obtained about the individuals from the protected data set. This measure is small when the masked dataset values are very different to the original values. In this work we used the *interval disclosure* (ID)[6], the *distance-based record linkage* (DBRL)[7], the *probabilistic record linkage* (PRL)[7], and the *rank swapping record linkage* (RSRL)[13].

Algorithm 2. MDAV-generic microaggregation algorithm

Input: X dataset, k level of anonymity.
Output: X' protected dataset.
while $(|X| > 3k)$ **do**
  Compute the average record $\bar{x}$ of all records in X. The average record is computed attribute-wise
  Consider the most distant record x_r to the average record $\bar{x}$ using appropriate distance
  Find the most distant record x_s from the record x_r
  Form two clusters c_r and c_s around x_r, and x_s where $|c_r| = k$ and $|c_s| = k$
  Take as a new dataset X the previous dataset X minus the records in c_r and c_s
end while
if there are between $3k - 1$ and $2k$ records in X **then**
  compute the average record $\bar{x}$ of the remaining records in X
  Find the most distant record x_r from $\bar{x}$
  Form a cluster c_r containing x_r and the $k - 1$ records closest to x_r
  Form another cluster containing the rest of the records
else
  Form a cluster with the remaining records
end if
return Y

The problem here is that both measures are inversely related so the higher information loss the lower disclosure risk, and the inverse. In order to perform a good protection there must be a minimized and balanced combination of both measures.

5 Experimental Results

In this section we present the results obtained for the protection of the Twitter-extracted microdata dataset using the microaggregation protection method.

The microdata dataset we used in our experiments contained 621 Twitter users profiles but only 324 users have an associated topic of interest. As all the users without associated topics of interest will be directly aggregated into a single cluster, we just focused on the protection of the ones that have some associated topics.

It should be noticed that our method is sensitive to the choice of the initial node in the sense of that each generated graph will be different. However, in order to just make an initial test of our method we used one single graph to run the experiments.

To protect this dataset, we generated 10 different microaggregation protections with different levels of k-anonymity. Then, we evaluated the original microdata dataset without protection and all the 10 protections in order to assess the lack of privacy in the original microdata dataset and to determine which would the best protection. The results are shown in Table 1.

As we explained in Section 4 it can be seen that information loss and disclosure risk measures are inversely related. Then, as we want to obtain good protections and the quality of these protections is described by two inverse related measures, the best protections will be the ones that have minimum values in both measures and that these values are balanced. Having this into account, it can be seen that the original microdata

Table 1. Results of the original and protected microdata evaluation

Dataset	Information Loss	Disclosure Risk
Original	0.00	99.54
Protected K=2	26.19	52.44
Protected K=3	32.13	43.58
Protected K=4	35.44	36.36
Protected K=5	41.50	31.63
Protected K=6	42.21	30.43
Protected K=7	46.05	28.02
Protected K=8	48.08	26.30
Protected K=9	49.88	24.26
Protected K=10	51.84	23.55

dataset obtained, as expected, a very bad results with a 99% of disclosure risk and a 0% of information loss. This is very bad because it means that almost all the users are exposed to the disclosure of their sensible information. However, if we take a look at the different protections results we can see that measures are more reduced and balanced.

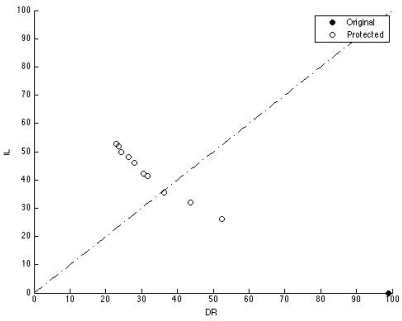


Fig. 3. Dispersion plot of the protected and original microdata evaluation results

Figure 3 shows the obtained results graphically. The dotted line represents the perfect balance of the measures so, the closest to the line and to the (0,0) point, the better protection. It can be seen that the original microdata is too far away from both. However, there is a protection that has the almost perfectly balanced values in both measures. Taking a look at Table 1 it can be seen that this is the case of the protection with k=4 (4-anonymity).

Comparing the results obtained in the original microdata and in the K=4 protection evaluations it can be seen that we have been able to decrease 63 points the risk of sensible information disclosure, but at the cost of increasing 35 points the analytically useful information. Then, we can be much more confident to publish this protected dataset than the original one in terms of individuals privacy.

Finally, it should be noticed that, as we are protecting a set of nodes attributes that include degree of each node, we are getting as a result a k-anonymous graph following the definition proposed by [11] that says that a graph is k-anonymous if every different

node degree appears at least in k nodes. Then, we can conclude saying that our protection approach could be used to perform this kind of k-anonymity protections.

6 Conclusions

In this paper we presented an approach to extract and protect microdata datasets from a real social network such as Twitter.

We have demonstrated that there is information that is not explicitly given in the social network user profile, but is implicit inside the posts the user shares. In order to get this kind of information we used the OpenCalais Web services to categorize the posts and extract the topics of interest from each user. In addition, in order to enrich the semantic content of the shared posts, we used the url's contained in the posts text.

We also have shown how to build a graph from the user extracted profiles, and how to convert it into a microdata dataset by taking the users profiles in the graph nodes.

Finally, we presented a way to protect this microdata dataset in order to be able to publish it for research purposes without violating the privacy of the contained users. We protected the dataset using the microaggregation method with different levels of k-anonymity. As a result we compared the evaluation of the privacy in the original dataset, and the protected ones. We demonstrated that the original dataset was violating the privacy of almost all the users, while using the microaggregation with 4-anonymity we obtained the best protection results reducing the risk of sensible data disclosure by 63 points but with the cost of increasing 35 points the loose of analytically useful information.

Then, we can conclude that microdata datasets can not only be extracted via surveys or statistical studies. They can also be extracted from the real social networks or graphs and, in this case, they may contain more information than the one explicitly described by the user in his social network profile. Then, they need also to be protected in order to publish them.

7 Future Work

In this work we only protected the dataset once it has been extracted from the social network. However, our main goal of the future work is to be able to protect the social graph information to get an already protected microdata dataset when it is extracted from the graph.

In addition, we also would like to consider the l-diversity rather than the k-anonymity since it has been proven that sometimes k-anonymity is not enough to protect a dataset.

Acknowledgments. This work has been done under the PhD in Computer Science program of the Universitat Autònoma de Barcelona (UAB). It is also partially supported by the Spanish MEC ARES-CONSOLIDER INGENIO 2010 CSD2007-00004.

References

1. Abel, F., Gao, Q., Houben, G.-J., Tao, K.: Semantic Enrichment of Twitter Posts for User Profile Construction on the Social Web. In: Antoniou, G., Grobelnik, M., Simperl, E., Parsia, B., Plexousakis, D., De Leenheer, P., Pan, J. (eds.) ESWC 2011, Part II. LNCS, vol. 6644, pp. 375–389. Springer, Heidelberg (2011)

2. Aggarwal, C., Yu, P.: Privacy-Preserving Data Mining: Models and Algorithms. Springer Publishing Company, Incorporated (2008)
3. Twitter API, https://dev.twitter.com
4. Campan, A., Truta, T.M.: Data and Structural *k*-Anonymity in Social Networks. In: Bonchi, F., Ferrari, E., Jiang, W., Malin, B. (eds.) PinKDD 2008. LNCS, vol. 5456, pp. 33–54. Springer, Heidelberg (2009)
5. Domingo-Ferrer, J., Mateo-Sanz, J.: Practical data-oriented microaggregation for statistical disclosure control. IEEE Trans. Knowl. Data Eng. 14(1), 189–201 (2002)
6. Domingo-Ferrer, J., Torra, V.: A quantitative comparison of disclosure control methods for microdata, pp. 111–133. Elsevier (2001)
7. Domingo-Ferrer, J., Torra, V.: Distance-based and probabilistic record linkage for re-identification of records with categorical variables. Butlletí de l'ÀCIA 28, 243–250 (2002)
8. Domingo-Ferrer, J., Torra, V.: Ordinal, continuous and heterogeneous *k*-anonymity through microaggregation. Data Min. Knowl. Discov. 11(2), 195–212 (2005)
9. Stokes, K., Torra, V.: Reidentification and k-anonymity: a model for disclosure risk in graphs. CoRR, abs/1112.1978 (2011)
10. Gouweleeuw, J., Kooiman, P., Willenborg, L.: Pram: A method for disclosure limitation of microdata. CBS research paper 9705 (1998)
11. Liu, K., Terzi, E.: Towards identity anonymization on graphs. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, SIGMOD 2008, pp. 93–106. ACM, New York (2008)
12. Nettleton, D.F., Sáez-Trumper, D., Torra, V.: A Comparison of Two Different Types of Online Social Network from a Data Privacy Perspective. In: Torra, V., Narakawa, Y., Yin, J., Long, J. (eds.) MDAI 2011. LNCS, vol. 6820, pp. 223–234. Springer, Heidelberg (2011)
13. Nin, J., Herranz, J., Torra, V.: Rethinking rank swapping to decrease disclosure risk. Data and Knowledge Engineering 64, 346–364 (2008)
14. Samarati, P., Sweeney, L.: Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report (1998)
15. OpenCalais Web Services, http://www.opencalais.com/calaisAPI
16. Torra, V.: Microaggregation for Categorical Variables: A Median Based Approach. In: Domingo-Ferrer, J., Torra, V. (eds.) PSD 2004. LNCS, vol. 3050, pp. 162–174. Springer, Heidelberg (2004)
17. Torra, V., Domingo-Ferrer, J.: Disclosure control methods and information loss for microdata, pp. 91–110. Elsevier (2001)
18. Tse, H.: An ethnography of social networks in cyberspace: The facebook phenomenon. The Hong Kong Anthropologist 2, 53–57 (2008)
19. Ward, H.: Prevention strategies for sexually transmitted infections: the importance of sexual network structure and epidemic phase. Sex Transm. Infect. (2007)
20. de Waal, T., Willenborg, L.: Elements of statistical disclosure control. Lecture Notes in Statistics. Springer (2001)
21. Winkler, W.: Re-identification methods for masked microdata (2004)
22. Yancey, W.E., Winkler, W.E., Creecy, R.H.: Disclosure Risk Assessment in Perturbative Microdata Protection. In: Domingo-Ferrer, J. (ed.) Inference Control in Statistical Databases. LNCS, vol. 2316, pp. 135–152. Springer, Heidelberg (2002)
23. Zheleva, E., Getoor, L.: Preserving the Privacy of Sensitive Relationships in Graph Data. In: Bonchi, F., Malin, B., Saygın, Y. (eds.) PInKDD 2007. LNCS, vol. 4890, pp. 153–171. Springer, Heidelberg (2008)

The TweetBeat of the City: Microblogging Used for Discovering Behavioural Patterns during the MWC2012

Daniel Villatoro, Jetzabel Serna, Víctor Rodríguez, and Marc Torrent-Moreno

Barcelona Digital Technology Centre, Spain
{dvillatoro,jserna,vrodriguez,mtorrent}@bdigital.org

Abstract. Twitter messages can be located in a city and take the pulse of the citizens' activity. The temporal and spatial location of spots of high activity, the mobility patterns and the existence of unforeseen bursts constitute a certain Urban Chronotype, which is altered when a city-wide event happens, such as a world-class Congress. This paper proposes a Social Sensing Platform to track the Urban Chronotype, able to collect the Tweets, categorize their provenance and extract knowledge about them. The clustering algorithm DBScan is proposed to detect the hot spots, and a day to day analysis reveals the movement patterns. Having analyzed the Tweetbeat of Barcelona during the 2012 Mobile World Congress, results show that a easy-to-deploy social sensor based on Twitter is capable of representing the presence and interests of the attendees in the city and enables future practical applications. Initial empirical results haven shown a significant alteration in the behavioural patterns of users and clusters of activity within the city.

1 Introduction

In the last decades, the number of inhabitants in urban spaces has enormously increased. In addition to being a place where people dwell, cities have become the center of human activities, the place where people move, work, play, learn, buy, sell and experience emotions. From a high-level perspective, a certain *beat* of the city can be perceived, and its realization as measurable figures and precise facts has always been of the highest interest for sociologists, entrepreneurs, urban planners, policy makers, or just for mere observers.

To achieve the understanding of citizens' behaviour in the cities, information has been traditionally gathered via random polling and indirect measures, when not directly from the mere intuition. These approaches however, imply the main disadvantage of obtaining biased information (i.e. polled participants not providing real information, having partial sampling), and often require expensive procedures that often require long execution time.

Fortunately, in the last few years, mobile technologies have gained massive spread among the citizens, and these can now be used as proactive sensors. Device capabilities improve constantly, prices drop and wireless connectivity infrastructures are becoming each time more universal. Indeed, device-holders have now the facility to continuously share all kind of information anytime and anywhere, and much of this information is public. Moreover, the proliferation of social networking has generated an incomparable

J. Nin and D. Villatoro (Eds.): CitiSens 2012, LNAI 7685, pp. 43–56, 2013.

and incentivizing framework for users to share any type of information about friendships (e.g. Facebook), work-colleagues (e.g. Linkedin), pictures (e.g. Flickr) or even the favourites dishes in the restaurants around (e.g. Foodspotting).

In this paper, we propose to capture the beat of the city and its alterations (provoked by the visitors activity) in the face of a public event by exploiting the public data offered by Twitter (i.e. a microblogging platform with a relative abundance of messages, geolocation capabilities and good temporal stamp for each point). Moreover, the portmanteau *Tweetbeat* conveys the idea that the beat of the city is reflected in the beat of Twitter.

Therefore, the scope of this research is twofold: (1) to develop a hardware-infrastructure-less social sensor whose observation target is the city (fed with the microblogging information individually provided by their users), and (2) to evaluate the viability of the newly developed sensor to build urban-behavioural models (e.g. HotSpots identification, mobility patterns and unforeseen events detection, etc), which from now on will be referred as *Urban Chronotypes (UC)*. To achieve this, Twitter seemed to be the ideal candidate because Twitter-users can attach the GPS position of the device from where they *tweet*, allows information public access, and this access is obtained in "near" real-time. Thus, we highlight the role of Twitter as a Social Sensor for Urban Chronotypes identification. Gathered information will allow us to identify the average UC, and therefore, be able to detect potential disturbances within a city.

Our case of study focuses on the city of Barcelona, host of the international event Mobile World Congress (MWC). To cover the necessary phases in the UC creating process, we implemented the Social Sensing Platform (SSP), which gathered all tweets in Barcelona during 3 weeks: one week before to the MWC, the MWC week and one week after the event. Once data was captured, the framework was capable of performing several types of knowledge extraction through statistical analysis, and clustering techniques, using the DBScan algorithm. The SSP allowed us to observe the average Chronotype of the city of Barcelona, and how the UC was affected by a major event such as the MWC.

The rest of the paper is organized as follows: Section 2 reviews the state of the art and similar contributions; in Sec. 3 we state the problem, and the case of study is determined in Sec. 4. In Sec. 5 we describe the modular architecture of our social sensor, and some initial statistical results are presented in Sec. 6. Later in Sec. 7 we present the results obtained after applying clustering techniques, and finally we draw some conclusions and sketch the future work in Sec. 8.

2 Related Work

The idea of examining the mobility patterns in a city during a certain event by observing microblog posts has not been directly considered, but a number of related experiments have been described.

The first piece of information about an event is its mere existence. Microblogging has actually been used as a sensor to detect both natural phenomena (like earthquakes [1], levels of pollen in the air [2] or even weather events [3]) and events of human nature (crime and disaster events [4], those that gather crowds [5] or general events in [6]).

Beyond the mere detection, these events have been also further characterised in order to extract useful information. Among this information, the spatial and temporal

coordinates of the users have been a key aspect. The temporal description of microblogging posts or social media in general, together with a sentiment analysis can be used to anticipate events of any sort, like the commercial success of a movie [7] or even the stock market [8]. The spatial description of social media can be used to detect points of interest, like geolocated *flickr* photos for the tourist case in [9], which actually reaveal trajectories when it combines this information with the time stamps. The joint analysis of temoporal and spatial description of Twitter messages produces indeed richer results and it has been used to anticipate music popularity [10], political alignment [11] or general Twitter themes [12].

The rich information provided by telecommunication networks used to characterize the city dynamics in front of a certain event, as in [13], can be thus replaced by *public* information and far more sparse; as an example the regular beat of New York was well characterized from Tweets collected by [14].

From another point of view, [15] characterized the nature of different regions regarding mobility by proposing a first step of identification of the relevant areas (applying a k-means clustering algorithm to Tweets), and then a second step tracking the movement pattern with new users coming into a cluster or existing users leaving it. These movements were shown to be predictable with a semi-Markov model by [16], although in this case data was acquired from Mobile Social Networks where the posting frequency was much higher than that of microblogging.

3 Problem Statement

The Smartcity paradigm has recently received an ever-increasing level of attention from the scientific community. Optimizing urban processes is a research goal that gathers different scientific areas such as Policy-making, Computer Science, Urbanism and Sociology. Opposed to the classical passive continuous sensing approach (where system owners had to study special parameters such as where to locate the sensor or the sensing-frequency, and then detect anomalies in the normal behavior of the observed object/phenomenon), in this work we plan to profit from the human pro-active sensing capabilities (enabled by their mobile devices such as smartphones).

As we have seen in Sec. 2 other researchers have profitted from Social Media platforms to obtain information related to the urban behavior of users. Although, until now, experiments done with Twitter as a source of information have been performed by advantaging of the full Twitter support (i.e. not considering only publicly available information via the standard API, but a full opened connection which is not commonly granted to the general public). Thus, in this work we emphasize in, how this Social Sensor can provide us information to observe and detect alterations in the Urban Chronotype of the city, and, the viability of achieving it with a given dataset that is based on information publicly available without the requirement of any private partnership or special Social Domain participation. Moreover, we demonstrate that, the developed platform can be considered as a low-cost social sensor. Finally, by applying intelligent analyses (such as geospatial clustering), our main goal is to support decision making processes and identify important changes in the city in near real-time.

To achieve this goal, we profit from our knowledge and expertise in Smartcities and apply it to the city of Barcelona, together with the celebration of a worldwide event such

as the MWC. This specific case of study is interesting, as it provides us of a controlled event (such as the MWC) where we have already a background knowledge inferred from previous editions (e.g. the amount of participants or location of the venue), and makes it an ideal candidate for the SSP initial experimentation. The controlled situation that the MWC offers, serves as a control test, allowing us to evaluate the performance of the platform even when facing unexpected results.

4 Case of Study

The city of Barcelona has been the host for the Mobile World Congress for six editions celebrated yearly (from 2006 to 2012) at *Fira de Barcelona - Plaza España Pavilion*. From the previous editions, we know that this event brings to the city a numerous amount of people (an average of 50.000 people in the last 6 editions, and 65.000 in the 2012 edition) from all over the world, with a common interest: mobile devices. Because of this common technological interest, we hypothesized that the infiltration level of Social Media applications within this community should be high. Therefore, the MWC attendees (as any other visitor) have an effect on the average Urban Chronotype of the city, with the slight difference that this effect might be reflected on the activity recorded by Twitter with a higher impact.

In order to observe this variation on the Barcelona Urban Chronotype, first we need to obtain the average Urban Chronotype that would work as a control case for us. The city of Barcelona was the host for the Mobile World Congress from February 27th to March 1st 2012, and as control cases we have decided to use the week before and after the event.

5 Social Sensing Platform Architecture

The proposed social sensing architecture is composed of 3 independent but interacting elements, as shown in Figure 1, each of them described below.

This platform receives as an input the City to be "sensed". Other parameters are optional, such as the hashtags of the potential event to track.

5.1 Tweet Hunter

This module is in charge of the information acquisition of the City, which in this specific case is gathered from the microblogging site Twitter (further versions of this platform will include other social media sources). Twitter allows us to query via a single streaming connection per user and IP by using the public Streaming API, obtaining near real-time information. Queries to the Twitter Streaming API can be of several types, but we focused on the geospatial type of queries, where, given a bounding box (delimited by the south-west and north-east coordinates) Twitter will return 'all' the Tweets generated within that area. Moreover, according to the Twitter streaming API Documentation[1], it

[1] How are rate limits determined on the Streaming API? on https://dev.twitter.com/docs/faq, Accessed on May 31st 2012

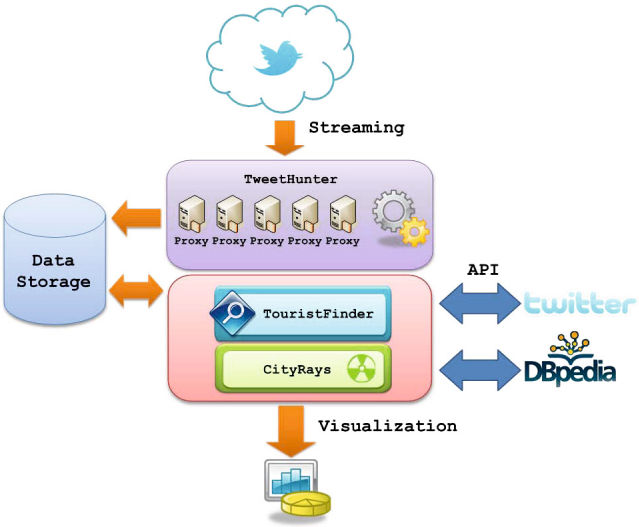


Fig. 1. Social Sensing Platform Architecture

provides a maximum of 1% of the global Twitter streaming (returning a message if this limits is exceeded), meaning that we could risk information loss, if the query response overpasses this limit. Thus, to reduce the probability of information loss considering the specific case of Barcelona, we have setup five proxies, each of them with an opened stream connection, targeting particular area of the city and configured with the targeted area related parameters. We have strategically selected four different and highly important spots in Barcelona (i.e. airport, main train station, tourist center and Fira Barcelona - epicenter of the MWC2012 event) and setup four of the collectors with each of the spot's bounding box. Additionally, a fifth collector with a bounding box that covered the whole city of Barcelona. By taking advantage of the bounding box configuration, we assumed that, *1)* the targeted spots (bounding boxes) were small enough and therefore not able to produce more than 1% of the global twitter streaming, and, *2)* information not captured by the fifth streaming - targeting the whole city, could be completed with the other four streaming connections (targeting the most crowded places), and ultimately the data loss could be neglected.

5.2 Tourist Finder

Some information about the users is handled by the Tourist Finder, and its main task is to determine the origin of the captured Tweet's users and place them in any of the following categories: *Local*, *Tourist* or *Unknown*. The content of the Tweet, which in some occasions has been used to locate the message ([17]) has been neglected in our experiment.

The Tourist Finder performs two important interactions:

1. With the Twitter REST API: to query about user's location. As each gathered tweet is accompanied with the user id that have originated that tweet, we can query

Twitter (via REST API) about the user's specified location in his own Twitter profile. In that way, our platform obtains a list of users and their locations.
2. With the DBpedia REST API: to query the set of keywords that identifies a city or region of the world. With the origin's dataset, this module builds up an ontology based on *1)* automatic queries to identify a set of keywords related to a particular city/place of the world, this is done thanks to the DBpedia API (which for this specific case returned us a complete list of all the *Populated Places*, such as cities, towns or villages, in the Catalonia Region), and *2)* a semi-automatic classifier that automatically extracts a set of unrecognized keywords (not identified by the previous process), and that, needs manual interaction to be able to identify when a location input is clearly undefined (e.g. "somewhere in the world", "in this planet", "Gaga's Heart", etc). Note that, since the location parameter in Twitter is an input to be entered by the user, an inherent truthfulness error probability will always exist.

After these two steps the Tourist Finder will be able to classify all the users in one of the following categories:

1. Locals: all users with a location parameter within Catalonia region.
2. Tourists: all users with a location parameter outside Catalonia.
3. Unknown: all users with no location or location undefined (location not recognized).

5.3 City Rays

This module is in charge of actually analyzing the data obtained from the gathered Tweets and extract knowledge out of them. It is also in charge of analyzing temporal and geospatial analysis. It is capable of extracting average behaviours in different temporal ranges, combining the information obtained from the *Tourist Finder* module, and it also performs spatial clustering techniques. This module however can be easily extended with new functionalities (e.g. spatio-temporal analysis).

6 Geostatistical Analysis

As previously specified, the experiment lasted for three uninterrupted weeks from Feb 20th (00:00:00) 2012 to March 11th (23:59:59) 2012. Table 2 describes the date ranges covered by each of the experiment weeks and its corresponding identifier. Week 1 and Week 3 are the control cases, although we understand that both weeks might be slightly affected by the event.

During these three weeks we gathered around 250,000 tweets from the Twitter Streaming API (generated by 15,911 different users), where the 43.10% of them contained the GPS coordinates associated to the user's geolocation when posting those tweets.

Figure 2 shows the total number of different Twitter Users (that tweeted at least once) in Barcelona with respect to those whose Tweet's contained specific geopositions. It is easy to observe the behavioural variation during Week 2. We hypothesize that the peak reached the first day of Week 2 is affected by the studied event (the MWC). Moreover,

Table 1. Experiment Weeks Identifiers and Coverage

Week Id.	Initial Date	End Date	Exp. Category
Week 1	02/20/2012	02/26/2012	Control Case
Week 2	02/27/2012	03/04/2012	Subject of Study
Week 3	03/05/2012	03/11/2012	Control Case

we can observe that the ratio of Geolocated tweets remains constant at around 40% during the three weeks[2] as it can be seen in Fig.2. This is an interesting result that lead us to think that the individual behavioural trends with respect to geo-positioning are not affected by the external event.

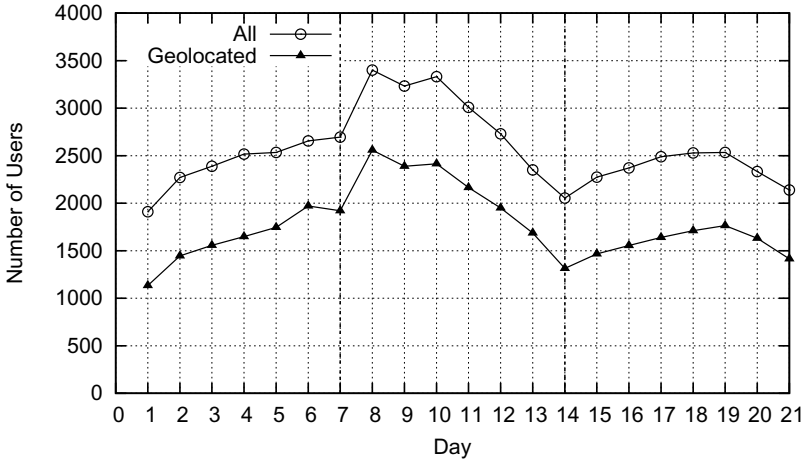


Fig. 2. Twitter Users and Geopositioned Users

However, we were interested in observing the specific trends of the event attendees. To do so, we extract the tweets that contained any information related to the event that for the sake of readability will referred as #MWC (although it refers to a set of hashtags and keywords related to the event such as, #MWC2012, #MWC12, MWC, "Mobile World Congress", etc).

Figure 3 compares the amount of different users per day that used at least once one of the #MWC related hashtags in a geolocalized tweet with those that did not geolocalize the tweet. We can observe that the event possesses a strong geolocalized facet wrt the average city trends shown in Fig. 2, since users actively provided their geoposition when they tweeted about this event. Moreover, we can observe that the curve reaches its peak during the event (as predicted by [18]), showing an initial activation effect two days previous to the event.

[2] We would like to remark that this ratio improves the the state of the art analysis performed in the literature, where the most successful identified case worked with a dataset that presented at geopositioning ratio of the 0.41%.

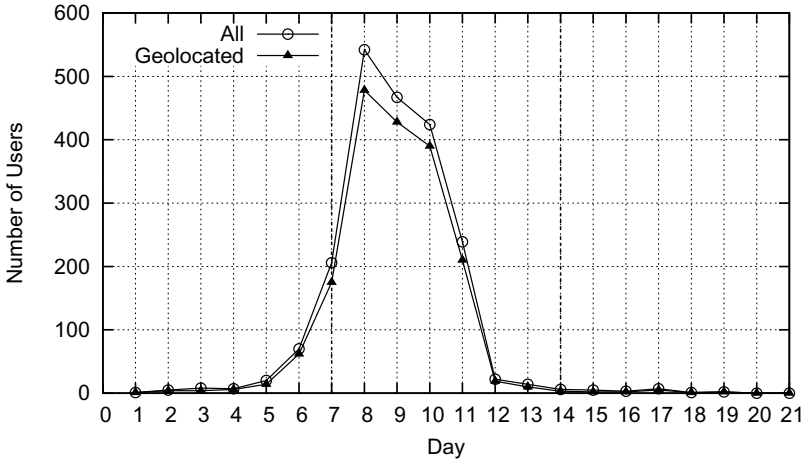


Fig. 3. #MWC User distribution

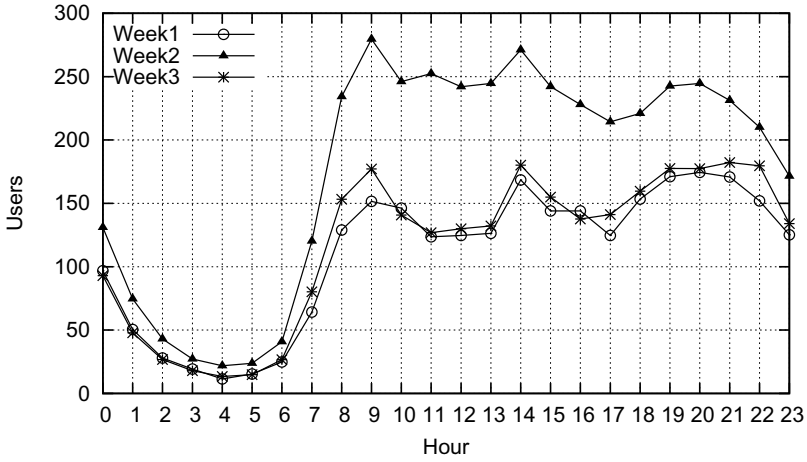


Fig. 4. Week comparison of Geopositioned Users

In order to assure the geoposition character of the event, we extract the average daily pattern of the amount of the different geopositioned twitters (users that tweeted with the specific geolocation at least once in a period of time) and compare the results of the three different weeks. In Figure 4 we can observe the substantial difference observed during Week 2. This result combined with the previous result lead us to hypothesize that users interested in the MWC event were also active geopositioning users, which can therefore provide us with relative information about their behavior within the host city.

In order to know more about the type of users that are influencing the city during the event, the *Tourist Finder* module help us to determine the origin of users. We focus only on the geopositioned users. Figure 5 shows how during the event, the number of *Tourists* increases up to the point of exceeding the number of *Locals* during the opening day.

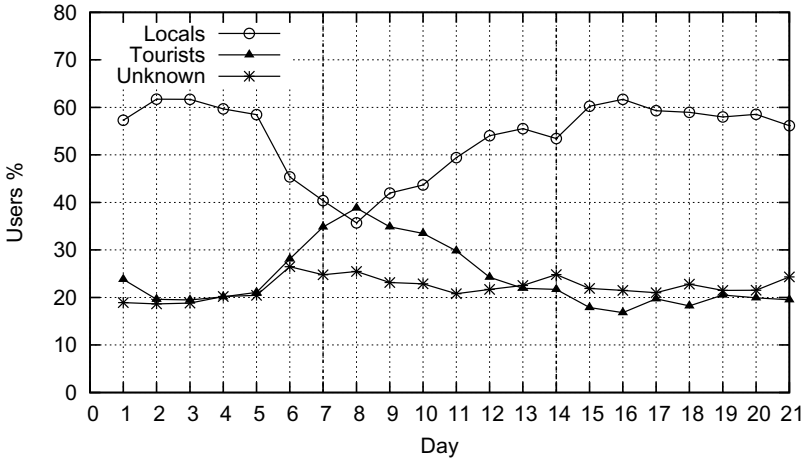


Fig. 5. Geopositioned Users Origin's Distribution

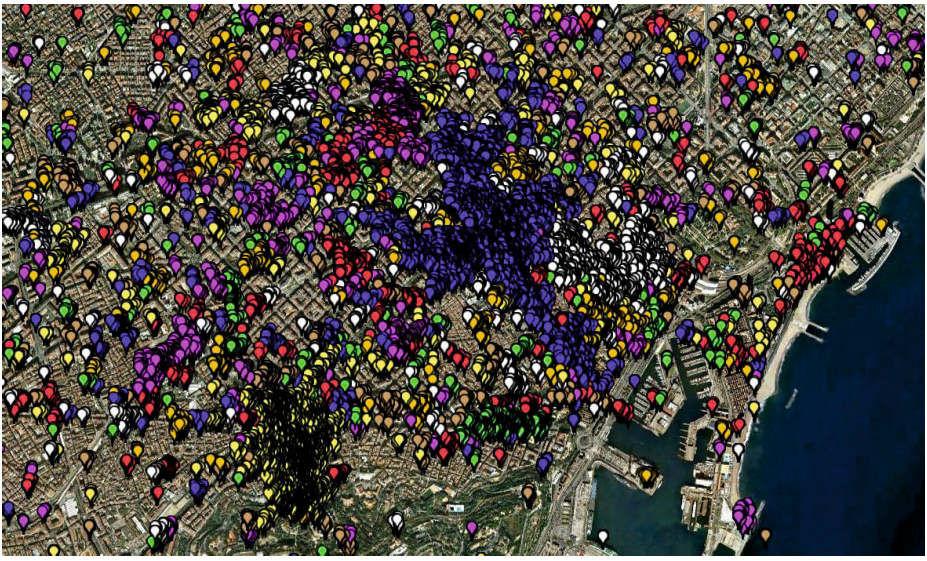
7 Geospatial Clustering Analysis

After a detailed comparative analysis of clustering algorithms with spatial data [19], we opt for DBScan, rather than other well-known clustering algorithms such as CLARANS, EM or *k*-means. The DBScan Algorithm [20] is a clustering algorithm that posses a number of characteristics that differentiates it from the other standard algorithms in the literature and makes it the ideal candidate for our scientific scope:

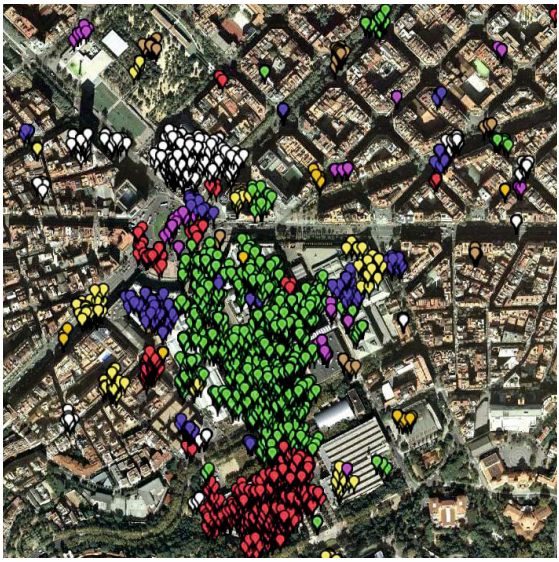
1. It is based in the concept of *density reachability*, producing satisfying results identifying arbitrarily shaped clusters.
2. The number of clusters is not given a priori.
3. The algorithm tolerates noise, allowing for some data points not to be assigned to any cluster.

Table 2. DBScan ε-sensitivity Results

Epsilon	m.	# Clusters	Noise
0.025	2000 m.	7	63
0.0125	1000 m.	46	229
0.00625	500 m.	273	1248
0.0025	200 m.	1385	6899
0.00125	100 m.	2773	17223
0.000625	50 m.	3452	31298



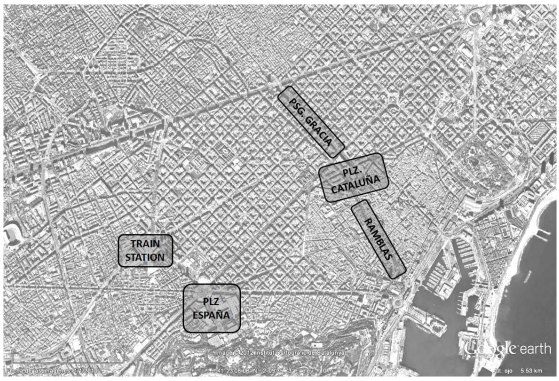
(a) $\epsilon = 0.00125$ **Downtown Barcelona Close-up**



(b) $\epsilon = 0.000625$ **Plaza España Close-Up**

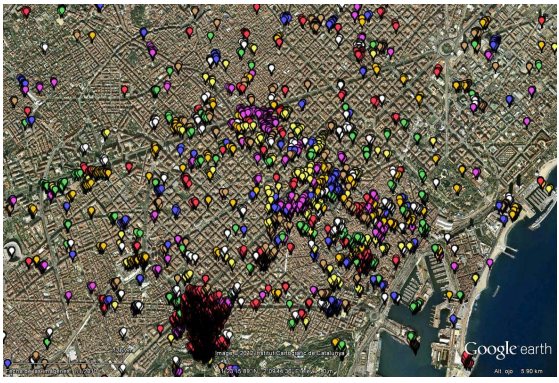
Fig. 6. ϵ-Sensitivity

The DBScan is sensible to two input parameters: the minimum amount of points (intuitively fixed to 5 for the results presented), and the ϵ value. The ϵ value determines the minimum distance amongst point to be part of a cluster, and therefore, also determines the granularity of the cluster (higher values of ϵ results in coarser clusters).



(a) **Downtown Barcelona Areas of Interest**

(b) **Clusters obtained in Week 1 and 3**



(c) **Clusters obtained in Week 2**

Fig. 7. Geospatial Clustering Results

Determining *a priori* the correct value of ϵ is problem dependent and almost unfeasible for the type of scenario we face. Therefore we perform a search space of this parameter. For the complete dataset gathered from the 3 weeks, we can observe in Table 2 that with lower values of ϵ the number of clusters increases, and therefore the number

of noise generated. As the bounding box that we defined included some surroundings towns of Barcelona, we can see how certain values of ϵ ($\epsilon = 0.0125$ or $\epsilon = 0.00625$) create clusters that differentiate cities, although the whole city of Barcelona remain as one unique cluster, being hard to determine the areas of interest. When we apply lower values of ϵ ($\epsilon = 0.00125$), we start detecting realistic clusters to be considered in a city such as Barcelona (e.g. in Figure 6(a) we can see how the Ramblas are perfectly clustered, as well as Plaza España). However, with smaller levels of ϵ ($\epsilon = 0.000625$) the clusters become very granular (e.g. in Figure 6(b), focused in Plaza España, we can see different clusters for the different pavilions and surrounding areas). Determining the correct value of ϵ becomes an interesting problem to be solved, although for our initial experimental set we use empirically obtained values.

Having empirically tested the effects of different values of ϵ in our dataset, we decided to use $\epsilon = 0.00125$. This value represents a distance of 100m, which in the city of Barcelona has special sense, as it is the regular measure of one block in Eixample neighborhood (dominating a substantial area of the city-center).

The clusters generated in each of the three weeks are substantially different in the city after executing the DBScan algorithm (with $\epsilon = 0.00125$). Figure 7 shows a snapshot of these results, accompanied with a reference guide of the city of Barcelona (in Fig. 7(a)). We can easily observe that in the control-case weeks the clusters generated in the city (partially shown in Fig 7(b)) would be part of the Urban Chronotype of the city, showing the clusters of activity in the city in its normal state. However, we can spot how a substantial cluster appears in the Plaza España area (host of the MWC) in Fig. 7(c).

8 Conclusions and Future Work

In this work we have presented a modular social sensing architecture for urban environments. This architecture is fed with information obtained from the Streaming Twitter API, and has resulted satisfactory since we have obtained a 40% of the generated tweets with specific coordinates. This architecture can be seen as a low-cost sensor of the city, and allow us to construct the urban chronotype. This urban chronotype serves to compare the current behavior of the city and try to detect anomalous behaviour in the city in near real-time. However, before taking a real-time approach, we have used a controlled scenario that have an impact in the city, such as the Mobile World Congress, which in the 2012 edition had around 65.000 participants.

Along this paper we have shown the behavioural patterns of the city in its normal state and during the event, where we can see trends in the amount of users, geospatial information generated or distribution of the population depending their origin. The differences between the control case weeks and the event week are substantial enough to determine the success of our urban social sensor.

Moreover, we have used clustering algorithms to extract the areas of high-activity in the city. Amongst the existing clustering algorithms we opted using the DBScan algorithm, even though, it is extremely sensible to the ϵ value. Unfortunately, there are no existing techniques to approximate the value of ϵ for specific problems.

As future work, we will extend the DBScan algorithm, by profiting from the nature and context of the information that we are handling and improve the efficiency of it.

Specifically, there is some urban information that is publicly available in the form of open data (such as the average cost of housing square meter or the population density per neighborhood). We plan to dynamically adapt the ϵ value with respect to the the average population density of the dataset; in that way, when one instance is selected to be evaluated, the adapted algorithm will obtain the neighborhood or city to which that instance pertains and then obtain the population density to dynamically adapt the value of ϵ: coarser when evaluating points out of the observed city and, fine grained when evaluating points within the city.

Moreover, we plan to evaluate the efficiency of our platform when facing real-time detection of anomalous behaviour within the city, which will imply the adaptation of the algorithms to perform in real-time.

Finally, and as part of our long-term research, we will evaluate the mobility patterns of those active users, considering each of their geolocalized tweets as digital footprints that can be evaluated as part of a track. Combined with the classification of the user's origin we will be able to extract mobility patterns within the city depending on the users origin.

Acknowledgements. This research is partially supported by the Spanish Centre for the Development of Industrial Technology under the INNPRONTA program, project IPT-20111006, "CIUDAD2020" (www.innprontaciudad2020.es).

References

1. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes twitter users: real-time event detection by social sensors. In: Proceedings of the 19th International Conference on World Wide Web, WWW 2010, pp. 851–860. ACM, New York (2010)
2. Takahashi, T., Abe, S., Igata, N.: Can Twitter Be an Alternative of Real-World Sensors? In: Jacko, J.A. (ed.) HCI International 2011, Part III. LNCS, vol. 6763, pp. 240–249. Springer, Heidelberg (2011)
3. Jeff Cox, B.P.: Improving automatic weather observations with the public twitter stream. Technical report, Indiana University Computer Science Program (February 2011)
4. Li, R., Lei, K.H., Khadiwala, R., Chang, K.C.-C.: Tedas: a twitter based event detection and analysis system. In: Proceedings of the IEEE International Conference on Data Engineering, ICDE (April 2012)
5. Fujisaka, T., Lee, R., Sumiya, K.: Detection of unusually crowded places through microblogging sites. In: 2010 IEEE 24th International Conference on Advanced Information Networking and Applications Workshops (WAINA), pp. 467–472 (April 2010)
6. Weng, J., Yao, Y., Leonardi, E., Lee, F.: Event Detection in Twitter. Technical report, HP Labs (2011)
7. Asur, S., Huberman, B.A.: Predicting the future with social media. In: Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT 2010, vol. 01, pp. 492–499. IEEE Computer Society, Washington, DC (2010)
8. Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. Journal of Computational Science 2(1), 1–8 (2011)
9. Girardin, F., Fiore, F.D., Ratti, C., Blat, J.: Leveraging explicitly disclosed location information to understand tourist dynamics: a case study. J. Locat. Based Serv. 2(1), 41–56 (2008)

10. Schedl, M.: Analyzing the potential of microblogs for spatio-temporal popularity estimation of music artists. In: Proceedings of the IJCAI 2011: International Workshop on Social Web Mining (2011)
11. Conover, M., Gonçalves, B., Ratkiewicz, J., Flammini, A., Menczer, F.: Predicting the political alignment of twitter users. In: Proceedings of 3rd IEEE Conference on Social Computing, SocialCom (2011)
12. Nagarajan, M., Gomadam, K., Sheth, A.P., Ranabahu, A., Mutharaju, R., Jadhav, A.: Spatio-Temporal-Thematic Analysis of Citizen Sensor Data: Challenges and Experiences. In: Vossen, G., Long, D.D.E., Yu, J.X. (eds.) WISE 2009. LNCS, vol. 5802, pp. 539–553. Springer, Heidelberg (2009)
13. Martino, M., Vaccari, A., Ratti, C.: Pulse of the city: Visualizing urban dynamics of special events. In: GraphiCon - Internation Conference on Computer Graphics and Vision (2010)
14. Ferrari, L., Rosi, A., Mamei, M., Zambonelli, F.: Extracting urban patterns from location-based social networks. In: Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks, LBSN 2011, pp. 9–16. ACM, New York (2011)
15. Fujisaka, T., Lee, R., Sumiya, K.: Discovery of user behavior patterns from geo-tagged micro-blogs. In: Proceedings of the 4th International Conference on Uniquitous Information Management and Communication, ICUIMC 2010, pp. 36:1–36:10. ACM, New York (2010)
16. Du, Y., Fan, J., Chen, J.: Experimental analysis of user mobility pattern in mobile social networks. In: 2011 IEEE Wireless Communications and Networking Conference (WCNC), pp. 1086–1090 (March 2011)
17. Cheng, Z., Caverlee, J., Lee, K.: You are where you tweet: a content-based approach to geo-locating twitter users. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, CIKM 2010, pp. 759–768. ACM, New York (2010)
18. Lehmann, J., Gonçalves, B., Ramasco, J.J., Cattuto, C.: Dynamical classes of collective attention in twitter. CoRR abs/1111.1896 (2011)
19. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proc. of 2nd International Conference on Knowledge Discovery and Data Mining, pp. 226–231 (1996)
20. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Simoudis, E., Han, J., Fayyad, U.M. (eds.) Second International Conference on Knowledge Discovery and Data Mining, pp. 226–231. AAAI Press (1996)

A Platform for Citizen Sensing in Sentient Cities

Fernando Koch, Carlos Cardonha, Jan Marcel Gentil, and Sergio Borger

IBM Research - Brazil
{fernandokoch,carloscardonha,jgentil,sborger}@br.ibm.com

Abstract. This work develops upon the concepts of *Sentient City* – living in a city that can remember, correlate, and anticipate – and *Citizen Sensor Networks*. We aim at technologies to interconnect people, allowing them to actively observe, report, collect, analyse, and disseminate information about urban events. We are investigating new methods and technologies to enhance administrators' capabilities in urban planning and management. We are proposing a platform to instrument citizens and cities, interconnect parties, analyse related events, and provide recommendation and feedback reports. The solution encompasses four types of elements: (i) mobile applications for intentional and non-intentional reporting of events; (ii) enhanced analytic models to centralize information, analyse the data, identify trends and operation patterns, and provide insightful information to decision makers; (iii) advanced social simulations to anticipate "what if" scenarios for infrastructure planning; and (iv) interfaces for monitoring, feedback, and recommendation. This research builds upon the IBM Smarter Cities project, part of the IBM Smarter Planet program. The outcomes of this research yield significant social contributions. By using it, administrators can make reliable decisions that will impact social services, traffic, energy and utilities, public safety, retail, communications, and economic development.

1 Introduction

Sentient City [12] promotes the concept of living in a city that can remember, correlate, and anticipate situations. Supporting this concept, *Citizen Sensor Networks*, described in [13], aims at technologies to interconnect people, allowing them to actively observe, report, collect, analyse, and disseminate information about urban events.

We are investigating new methods and technologies to enhance a capability in urban planning and management. We aim at a solution that will help city administrator's to answer questions like:

- How to manage the flow of people and things in a city?
- How to get citizens engaged to and aware of cities affairs?
- How to guide people in the city under extraordinary situations, such as large-scale events and natural disasters?
- How to identify economic activities and the process to boost their development?
- How to provide a useful feedback to citizens?

For instance, *IBM Intelligent Operation Centers*[1] support the centralization of critical information from multiple sensors that control the city, providing advanced collaboration,

[1] Ref: http://www-01.ibm.com/software/industry/intelligent-oper-center/

J. Nin and D. Villatoro (Eds.): CitiSens 2012, LNAI 7685, pp. 57–66, 2013.

deep investigations based on analytics tools, and executive dashboards support for city operations. *San Francisco 311*[2] provides a Twitter channel that accepts inbound requests for information, services, notifications, and feedback. *Rio de Janeiro's Citizen Support Central*[3] provides mobile and web solutions for on-spot reporting of issues like broken public lights, road potholes, damaged vegetation, irregular parking, request for trash removal, and others.

We are proposing a *Platform for Citizen Sensing in Sentient Cities* to instrument citizens and cities. It provides the tools to interconnect parties, analyse related events, provide reports, and simulate possible development scenarios. It also supports the interfaces to *recommendation systems* that will help in decision-making and coordination. The long-term goals established for this project are:

1. Development of *mobile applications* to provide the public interface for intentional and non-intentional reporting, monitoring, and feedback, as well as the back-end services required to support them.
2. Conception of *analytic models* for filtering reports based on correlation of observation data, variations of local context, and variations of user profile.
3. Formulation of *analytic models* for evaluating the level of impact and priority of reports based on sentiment analysis of the accompanying text, as well as image analysis ran over an optionally accompanying photograph.

This research builds upon the IBM Smarter Cities project, part of the IBM Smarter Planet program. The analytic models represent innovations being introduced in this research. They advance the state-of-the-art in analysing and correlating data in Citizen Sensing solutions.

This paper is structured as follows. Section 2 provides an overview of related technologies and previous works. Section 3 introduces our proposal for a *Platform for Citizen Sensing in Sentient Cities*. Section 4 presents results that demonstrate the support to interrelate data and decision making. The paper concludes with Section 5.

2 Background

In this section, we analyse the concepts and state-of-the-art in the topics of Citizen Sensing, Social Analytics, and Social Simulation.

Figure 1 provides an overview of the elements in generic Citizen Sensing solutions. First, there are the (i) *Sensors* to collect information from the real world. They are the entry points to "feel the pulse" of life in the city. The implementation of standard sensor platforms and open interfaces will greatly facilitate the deployment of citizen sensing solutions. The works in [3, 11] emphasise the importance of ubiquitous sensors elements, arguing that this technology supports major breakthrough in the areas of *Human Dynamics* and *Crowd Coordination*.

There are two types of sensors that can be applied to different scenarios: (a) *Intentional Sensors* require end-user intervention, providing an interface for data entry; for instance, users tweeting in *San Francisco 311* and reporting issues in *Rio de Janeiro's*

[2] Ref: http://sf311.org/
[3] Ref: http://www.1746.rio.gov.br/

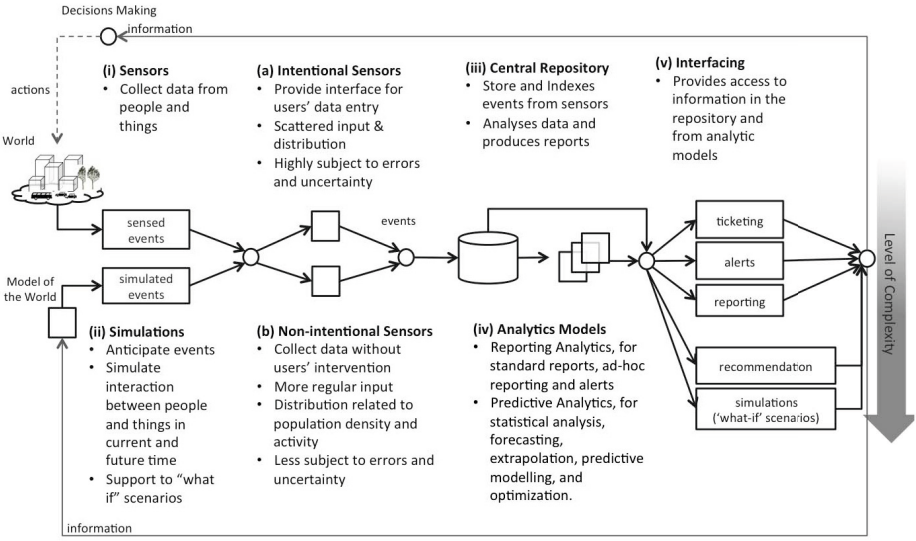


Fig. 1. Concept of Citizen Sensing Environment

Citizen Support Central; and (b) *Non-Intentional Sensors* collect data from the environment automatically; for instance, tracking systems that collect GPS information periodically, weather monitors, and others.

Second, there are (ii) *Simulations*, which are required to develop models to anticipate events and generate "what if" scenarios. The more information they entail in the *model of the world*, the more refined will be the simulation. We are seeking solutions in terms of traffic simulation, as e.g. [15], [6], and [9], including solutions focused on data extrapolation techniques such as [1] and [2].

Third, there are the (iii) *Central Repository* and (iv) *Analytic Models*. They provide the functionality to store, index, group, summarize, analyse, and cross-relate information. They deliver the solutions for *business intelligence*, making sense on the collected data beyond pure reactive and historical analysis. There is prior art showing the importance of such tools such as in [10] and [4]. These elements play central roles in commercial systems such as the IBM Intelligent Operation Centres [8].

Finally, there are the (v) *Interfacing Elements*, which provide access to data in the (iii) *Central Repository* and result from the (iv) *Analytic Models*. These solutions usually focus on visualization for ticketing systems, alerts, and reporting interfaces. A survey of existing information visualization is presented in [14]. More advanced systems provide recommendation interfaces, as for example [7]. Even more sophisticated, there are the interfaces to simulations required by solutions that provide "what if" scenarios.

In analysing the state-of-the-art, we concluded that current solutions focus on resolving only parts of the problem, such as on data collecting and analytic models. Therefore, opportunities exist to propose an all-encompassing solution for intelligent city monitoring and decision recommendation. This comprehensive solution must integrate the elements for sensing, simulation, central repositories, analytics, and visualization.

Next, we outline our proposal.

3 Proposal

We are proposing a *Platform for Citizen Sensing in Sentient Cities* that entails the elements for (i) sensors; (ii) data processing and augmentation; (iii) flexible data repositories; (iv) innovative methods for data analysis; and (v) extended interface. Figure 2 depicts the system architecture. The solution advances the state-of-the-art in urban planning and management by providing:

- *End-to-end Solution* for Citizen Sensing and social engagement, that can be scaled to cities of any size and applied to coordinate large events (e.g. crowd gatherings, sports games), emergency relief coordination, infrastructure planning, traffic monitoring, and others.
- *Enhanced Analytic Models* to identify trends and operational patterns in social interactions, anticipate emerging situations impacting society, and provide insightful information to decision makers.
- *Advanced Social Simulations* to be able to simulate interactions between the population and city infrastructure, supporting the anticipation and composition of "what if" scenarios.
- *Recommendation Systems* that leverage from these technologies and work closely with city leaders and deliver recommendations on how to make the city smarter and more effective.

The solution is being architectured to scale to cities of any size. It will be applied to deliver solutions to the complex scenarios involving the coordination of large events (e.g. crowd gatherings, sports games), emergency relief coordination, infrastructure planning, traffic monitoring, and others.

The elements of the solution are detailed next.

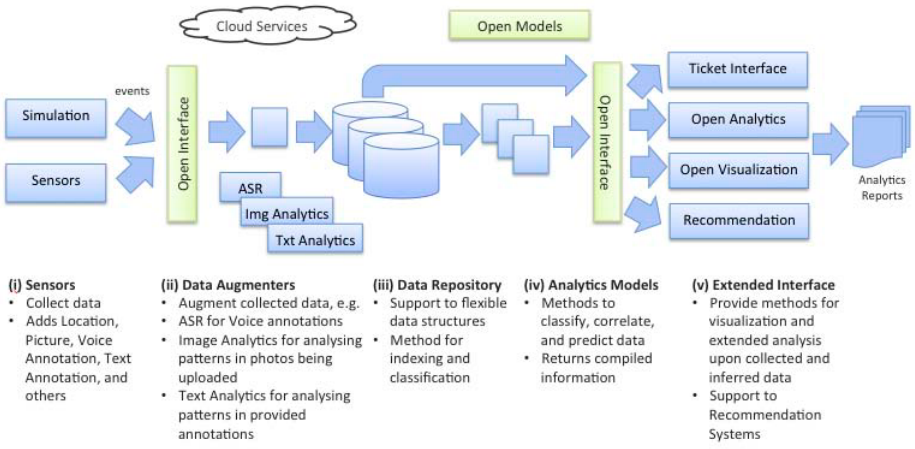


Fig. 2. System Architecture

Fig. 3. Mobile Citizen Sensors

Sensors Module. We are providing a configurable Mobile Sensor Application to support intentional and non-intentional Citizen Sensors. Figure 3 depicts the current prototype for the (i) *Sensors Module*. The *Citizen Watcher* application (in Portuguese, "Ver a Cidade") allows for citizens to report urban situations on the spot using their own smartphone devices. Examples of possible reports involve: aggression to urban vegetation, fauna, requests for gardening, reports of inaccessible areas, floods, traffic, road potholes and others.

The Mobile Sensor Application is freely available from regular mobile application stores[4] and connect to the (iii) *Data Repository* module through regular mobile Internet connection and open interfaces.

Using this platform, municipalities can easily promote the use of *Mobile Citizen Sensing*. The reports will be uploaded to customized Cloud Services providing open Web and mobile access to the data.

Data Modules. There are two sets of functional modules for (ii) *Data Augmenters* and (iii) *Data Repositories*, accessible through a open interface.

The (ii) *Data Augmenters* provide methods to augment received data as, for instance, by adding contextual information and converting data formats. In this module, we are incorporating methods for voice recognition to support voice annotation. Voice annotations are transcribed on the server by using IBM Voice Recognition technology.

Both voice and transcriptions are stored in the (iii) *Data Repository*, forming a database of rich annotations. This can be used to run text analytics on citizen's moods whilst reporting the events, allowing for extended analysis of the reports. Moreover,

[4] Note: only an iOS version is available at the moment. We intend to produce an Android version soon and, possibly, a Windows Mobile version in the short term.

we are including solutions for image recognition and image analytics that provide the pre-classification of the user generated annotations. This combination of technologies provide a key contribution to the overall solution, making it more pervasive and easy to use.

The (iii) *Data Repositories* implement the data store, indexing, and organization. In our architecture, this component is implemented upon IBM DB2 pureXML technology, providing methods for flexible data structures supporting data input from different sensors and simulations.

Analytics Modules. This model provides the methods to classify, correlate, and predict information upon the data stored in the (iii) *Data Repositories*. It is composed of analytic models based on data being collected and simulation for urban scenarios. In Section 4 we are presenting an illustrative scenario of the application of analytic models in the context of Smarter Cities.

Analytic models for Smarter Cities is a new concept and still field for research. For that, we are implementing a number of innovative models for both (i) *reporting analytics*, such as standard reports, *ad-hoc* reporting and alerts; and (ii) *predictive analytics*, for statistical analysis, forecasting, extrapolation, predictive modelling, and optimization.

In that regard, we are researching on methods and techniques for filtering reports based on correlation of observation data, variations of local context, and variations of user profile. Context can be defined by additional information that may be used to augment reports, such as time, location, and surrounding events, while user profiles are categorized in terms of end-users' demographics, psychographic information, and statistical analysis of previous contributions made to the platform[5].

Additionally, we are investigating on models for evaluating the level of impact and priority of reports based on sentiment analysis of the accompanying text, as well as image analysis ran over an optionally accompanying photograph[6]. This solution provides a pre-classification of the user generated annotations, in which the severity of certain observations can be recognized based on the pictures being uploaded along with the reports.

Finally, we are including open interfaces for analytics and visualization. It will allow for third-party developers and the public in general to have access to (public) data and create their own analysis and reports. We see openness and free access as pivotal in this platform. We believe that, by promoting free access, it will boost collaboration and analysis from different points of view. This approach will contribute to both (i) the sense of transparency and social inclusion and (ii) further development of analytic models for Smarter Cities.

Simulation Module. This module allows for the creation of prediction and "what if" scenario simulations. For instance, we are developing simulations to predict scenarios for early warning systems. This is based on previous work like [5], aiming at solutions for the coordination of large-scale events. The objective is to understand the likely problems that could occur and be prepared to respond. The environment encompasses the

[5] IBM patent pending.
[6] IBM patent pending.

description of city's infrastructure, population behaviour, crowd dynamics, law enforcement, emergency units, and others. We are implementing a bottom-up approach where we are modelling individuals' behaviour and their group attitude.

We are proposing agent-based simulations, where a population is instantiated with a number of general parameters according to culture and environmental conditions. These parameters will indicate which are the valued priorities of the people and, therefore, create certain preference orderings for actions given the current conditions. We will also model how the actions influence all the environment parameters and thus change the world and the context of the agents to make decisions.

In this line of research, we intend to establish how natural social groups, norms, and regulations can be used to provide some centrally controlled mechanisms to guide "realistic" crowd behaviour. For instance, a group of people with the purpose of going to a football stadium for a soccer match will not be easily deterred from its path, but people coming out from the match might be easily guided.

The simulations will be calibrated by using real data from areas where the population is well-known (using statistical data), e.g. what is the average income, unemployment rate, etc. All these parameters influence the value priorities. The outcomes of the simulation can then be compared with those in reality, e.g. how much garbage is produced, how do people dispose of their garbage, how much water is used, how many people own cars and use them every day, etc.

The inherent support to simulations in the composition allows for the creation of next-generation analytics solution that goes from "understand what has happening and inform" to "understand what is likely to happen and optimize".

Next, we describe case scenarios where we apply the proposed technology to coordinate resources and anticipate events.

4 Illustrative Scenarios

Let us consider a scenario where citizens using the *Mobile Sensor* module provide intentional reports for (i) garbage on the streets, (ii) areas of inundation, and (iii) traffic jams. The observations are collected by citizens and uploaded to the *Data Repository*, including the report, a text annotation, a photo, and the location. In this scenario, the proposed platform provides important and non-trivial insights that can be applied in the prediction of traffic conditions.

In this context, Figure 4 presents the data collection performed in a same area for the period of 15 days. Markers on the maps pinpoint the counters for each occurrence. The graphs on the right side present the variation of occurrence counters during the time period.

The analysis is as follows. On the third day, only a few garbage reports have been registered. Thus far, neither traffic jams nor floods have been reported. This observations suggests that garbage alone does not have an influence on congestions.

On the eighth day, the number of garbage occurrences increased. In addition, some spots of road inundation have been reported. In this circumstance, traffic jams started to be registered, specially around the road inundation locations. Hence, one can infer, based on both common sense and on analysis of the data, that the number of occurrences

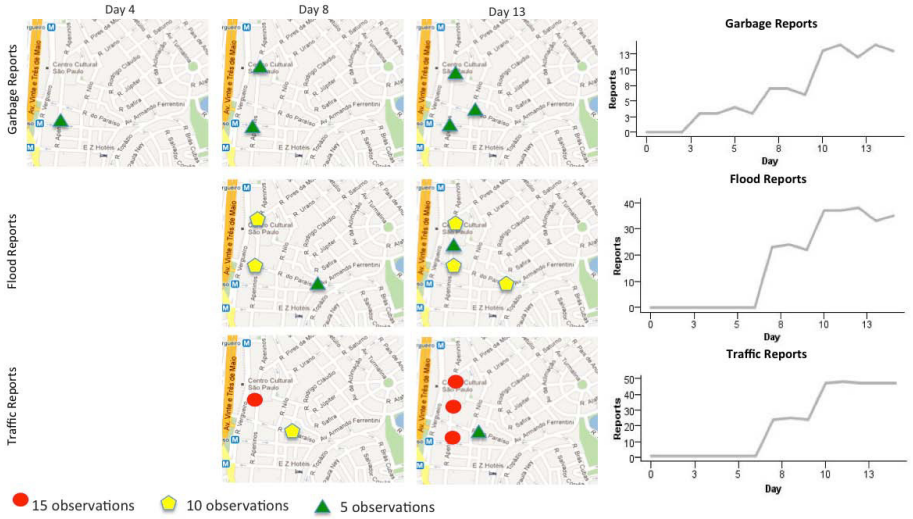


Fig. 4. Example of application based on the platform

of traffic jam is linearly proportional to reports of road inundation, but it is not related to garbage in the streets.

However, as we can see in the snapshots of the thirteenth day, the number x of flood observations increased and the number of traffic reports clearly had a superlinear growth in x. Therefore, the original conclusion is not precise. Modelling traffic as a function of rain may still be the correct approach, but if a decent formulation cannot be identified, cross relation of variables may bring some important insights. In particular, in the presented scenario, we can see that the number of garbage reports also grew significantly in the time period. If a large amount of garbage accumulated on the streets, it is possible that culverts have become clogged, leading to a scenario where floods were more likely to occur.

This fact suggests that garbage occurrence may be relevant in the context of traffic jam prediction, and therefore one should also consider the possibility of modeling traffic jams as a function of flood and garbage occurrences.

Finally, it is interesting to notice that, if data suggests a strong dependence between garbage, rain, and traffic jams in a certain area, it may indicate infrastructure issues in the regions (e.g. culverts need maintenance, garbage collection has not been adequately performed in the area, etc.). We conclude that the extended analytic capabilities provided by the *Platform for Citizen Sensing in Sentient Cities* provide useful information for city administrators to identify trends and operational patterns beyond the obvious. This allows for anticipating emerging situations impacting society, and provides insightful information to decision makers.

5 Conclusions

We are creating a decision-support system for urban planning and management that provided important tools and methods for city administrators. The proposed solution

encompasses sensors, data processing and augmentation, flexible data repositories, innovative methods for data analysis, extended interface, and simulation support. We are creating a solution that can be scaled to cities of any size, applied to the coordination of large events (e.g. crowd gatherings, sports games), emergency relief coordination, infrastructure planning, traffic monitoring, and others.

We presented the application of extended analytic capabilities to provide useful information in a scenario where citizens use the *Mobile Sensor* module to report (i) garbage on the streets, (ii) areas of inundation, and (iii) traffic jams. The analysis allows for understanding the cross-relation of events. It implied that there is a strong dependence between garbage, rain, and traffic jams in a certain area, that may indicate infrastructure issues in the regions. This sort of analysis is valuable for decision makers, justifying the implementation of the proposed platform by the municipality.

We foresee the application of the proposed technology to provide analysis in different areas of interest, as for example:

- *Accessibility Applications*: If people with disabilities are equipped with non-intentional sensors, it is possible to identify locations that were frequented by them and, more important, locations that were not. This data may be used by municipalities for the identification of inaccessible places.
- *City Occupation Analysis*: The solution can estimate the volume of people present in a certain area in a certain time span. Based on this information, it is possible to identify which regions are under-occupied at certain periods of time, what is the profile of the group of citizens that come to these places, and others.
- *Infrastructure Needs*: It is possible to determine the paths followed by citizens. If the platform is enriched with data describing which transportation vehicles the person used, it is possible to discover how people go from one place to the other, their origins and destinies and the timestamps of these events. With this information, governments may able to identify potential improvements in infrastructure and services related to public transportation.

As this work continues to develop, we are targeting to incorporate data available from external sources to the (iv) *Analytic Models*. For example, data can be imported from Open Data repositories[7] and social networks, such as Facebook and Twitter (e.g. a citizen posting a status about his neighbourhood being affected by road inundations whenever it continuously rains for more than a couple of hours).

Finally, we stress that the outcomes of this research yield significant social contributions. Being able to broadly coordinate resources and anticipate events, city administrators can make reliable decisions that will impact social services, traffic, energy and utilities, public safety, retail, communications, and economic development.

Acknowledgment. We would like to thank Christine Robson for the early implementation of IBM Creek Watch and her assistance in porting her project to the broader application being proposed. In addition, to Frank Dignum for his help in outlining the models to be applied for the simulations we intend to conduce in this work.

[7] Ref: http://datacatalogs.org/

References

1. Braxmeier, H., Schmidt, V., Spodarev, E.: Spatial extrapolation of anisotropic road traffic data. Image Analysis and Stereology 23, 185–198 (2004)
2. Braxmeier, H., Schmidt, V., Spodarev, E.: Kriged road-traffic maps. In: Proceedings of the International Conference StatGIS03, pp. 39–50 (2005)
3. Buchanan, B.: Behavioural science: Secret signals. Nature 457, 528–530 (2009)
4. Davenport, T., Harris, J.: Competing on Analytics: The New Science of Winning. Harvard Business School Press (2007)
5. Di Tosto, G., Dignum, F.: Simulating social behaviour implementing agents endowed with values and drives. In: Proceedings of the 13th International Workshop on Multi-Agent Based Simulation, Valencia, Spain (June 2012)
6. Dressler, D., Flötteröd, G., Lämmel, G., Nagel, K., Skutella, M.: Optimal evacuation solutions for large-scale scenarios. In: Operations Research Proceedings 2010, pp. 239–244 (2010)
7. Gretarsson, B., et al.: Smallworlds: Visualizing social recommendations. In: Eurographics/IEEE-VGTC Symposium on Visualization, vol. 29 (2010)
8. IBM. Intelligent operations center (May 2012), http://www-01.ibm.com/software/industry/intelligent-oper-center/
9. Malone, S., Miller, C., Neill, D.: Traffic flow models and the evacuation problem. UMAP Journal 22(3), 1–47 (2001)
10. Negash, S., Gray, P.: Business Intelligence, pp. 175–193. Springer (2008)
11. Pentland, A.: To signal is human. American Scientist 98(3), 204–211 (2010)
12. Shepard, M.: Sentient City: Ubiquitous Computing, Architecture, and the Future of Urban Space. The MIT Press (2011)
13. Sheth, A.: Citizen sensing, social signals, and enriching human experience. IEEE Internet Computing 13(4), 87–92 (2009)
14. Tegarden, D.: Business information visualization. Communications of the Association for Information Systems 1(4) (1999)
15. Thulasidasan, S., et al.: Designing systems for large-scale, discrete-event simulations: Experiences with the fasttrans parallel microsimulator. In: Proceedings of the 16th International Conference on High Performance Computing, HiPC 2009 (2009)

Incorporating Mobility Patterns in Pedestrian Quantity Estimation and Sensor Placement

Thomas Liebig, Zhao Xu, and Michael May

Fraunhofer IAIS
Schloss Birlinghoven, 53754 Sankt Augustin, Germany
{Thomas.Liebig,Zhao.Xu,Michael.May}@iais.fraunhofer.de

Abstract. Pedestrian quantity estimation receives increasing attention and has important applications, e.g. in location evaluation and risk analysis. In this work, we focus on pedestrian quantity estimation for event monitoring. We address the problem (1) how to estimate quantities for unmeasured locations, and (2) where to place a bounded number of sensors during different phases of a soccer match. Pedestrian movement is no random walk and therefore characteristic traffic patterns occur in the data. This work utilizes traffic pattern information and incorporates it in a Gaussian process regression based approach. The empirical analysis on real world data collected with Bluetooth tracking technology during a soccer event at Stade des Costières in Nîmes (France) demonstrates the benefits of our approach.

Keywords: Pedestrian Quantity Estimation, Trajectory, Gaussian Process Regression, Graph Kernels, Sensor Placement.

1 Introduction

Major public events such as soccer matches, concerts and festivals attract thousands or even millions of visitors. On the one hand this offers interesting business opportunities for event organizers, advertisement companies and street marketers. On the other hand it also creates a growing financial risk for the organizers due to huge expenses, and safety risks for the guests themselves. Understanding movement behaviour and identification of attractors and distractors gives insights on visitor preferences and motivations during a particular event. This can help in avoiding risks by better management of visitor flows. Various locations and attractions can be ranked by their popularity, safety or frequency, and measures against over-crowding can be taken immediately or for future events.

Sensor technologies that are currently in use to measure people quantities automatically are surveys, video surveillances, GPS, and Bluetooth scanners. Whereas the first solution (surveys) is expensive and hardly representative due to the non-random sampling among all visitors, the second one (video surveillance) depends on weather, brightness and density of the people and does sometimes require special scaffoldings to carry the cameras. GPS finally is not available everywhere, e.g. indoors and in urban canyons. In this paper we perform our tests

J. Nin and D. Villatoro (Eds.): CitiSens 2012, LNAI 7685, pp. 67–80, 2013.

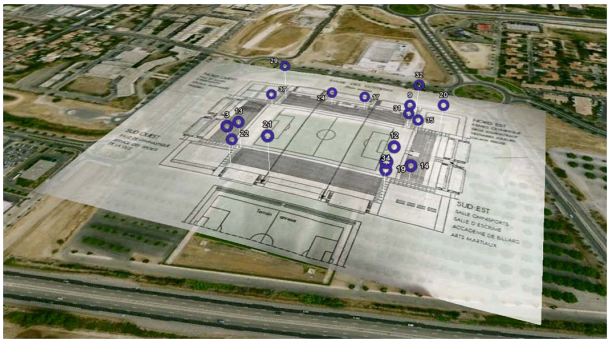


Fig. 1. 3D Sensor Placement at Stade des Costières, Nîmes (France) 05/08/2011

on a Bluetooth tracking dataset collected during a soccer match at the Stade des Costières, Nîmes (France) [1]. The data was collected using 17 Bluetooth beacons [2] at various locations in the stadium (Figure 1).

This work addresses the question where a fixed number of automatic pedestrian quantity sensors (i.e. Bluetooth beacons) is to be located during a mass event in order to get an adequate estimate on the movement of the visitors within the site. Our approach addresses the following questions:

- How can pedestrian quantities be estimated from a relatively small number of empirical measurements?
- At which places should a constrained number of pedestrian quantity sensors be located?

Often, available data for investigating these questions is limited to a small number of measurements and some prior knowledge, e.g., floor plan sketches or knowledge on preferred routes by local domain experts. Incorporating prior knowledge is thus essential to address the above challenges. However, so far there are few approaches that explicitly take into account the movement patterns, although pedestrians generally show some move preferences [3–6], especially in closed environments, e.g., sport stadiums.

In this paper we address both, pedestrian quantity estimation and sensor placement in the case where movement patterns are provided as background knowledge (Section 3) and the acquisition of movement patterns from Bluetooth observation data (Section 4). The paper is structured as follows. Section 2 discusses related work and gives an introduction to episodic movement data and its analysis. In Section 3 we introduce our Bayesian method (Gaussian processes). Section 4 highlights the application to the real world dataset. We conclude in Section 5.

2 Related Work

Bluetooth monitoring has found a number of interesting applications in recent years. Besides event monitoring, also other successful applications of Bluetooth

tracking technology are described in the literature. In [7] various scanners were placed at Dutch train stations to record transit travellers. Accurate location and tracking of objects within complex facilities is another important research topic [8]. Bluetooth tracking is also used to monitor a sample of visitors [1,8,9] and extract their route choices [1,10]. The work presented in [11] uses Bluetooth tracking to record people in a public transportation network, whereas [12] gives a general overview on possibilities using Bluetooth tracking technology. In a few works time-geography and movement patterns are addressed as well [9,13].

Bluetooth tracking is based on collecting episodic movement data (EMD) [9]. In GPS-less environments episodic movement data is the major representation of pedestrian mobility. Differently from outdoor pedestrian quantity estimation, continuous tracking technologies such as GPS cannot be used in many closed environments due to the lack of a GPS signal in buildings and/or expensive deployment of the hardware. Instead, recently developed alternative technologies such as light beams, video surveillance, and Bluetooth meshes record episodic movement data or its location-based-aggregate, presence counts, at low expenses. Episodic movement data is represented by tuples $< o, p, t >$ of moving object identifier o, discrete location identifier p and a time stamp t. The location-based-aggregate, presence counts, for time interval Δt, is also known as *number of visits*, *quantity* or *traffic frequency*. It is defined as $NV(p, \Delta t) = | < o, p, t >, t \in \Delta t|$. The *number of moves* among two locations p_i and p_j is similarly defined as $NM(p_i, p_j, \Delta t) = | < o, p_i, p_j, t >, t \in \Delta t|$. Other prominent examples of episodic movement data are spatio-temporal activity logs, geo-tagged photos, cell based tracking data and billing records.

Episodic movement data poses great challenges for existing data mining algorithms based on (linear) interpolation between data points. For example, speed and movement direction cannot be directly derived from episodic data; trajectories may not be depicted as a continuous line; and densities cannot directly be computed. The reason is that there are normally unmeasured locations between two measurements that cannot be reliably inferred by linear or other parametric interpolation.

Though this data is thus difficult to use for individual movement or path analysis, it still contains rich information on group movement on a coarser level. Our approach is to aggregate movement in order to overcome some of the uncertainties present at the individual level. Deriving the number of objects for spatio-temporal areas and transitions among them gives interesting insights on spatio-temporal behavior of moving objects. As a next step to support analysts, [9] proposes clustering of the spatio-temporal presence and flow situations (see Figure 3). In this figure the colour shading, which supports a visual understanding and analysis of the flows, results from Sammon's mapping [14]. To be more precise, the two-dimensional clustering of the flow situations (vector among all sensors) is mapped on a colour plane. As a result, similar flows get similar colours, and difference between flows corresponds to difference between colours. The different stages of the match are visible, and are subject for data partition in Section 4.

3 Pedestrian Quantity Estimation with Movement Patterns

Although pedestrians show systematic behavior and move preferences, especially in closed environments, e.g., stadiums, concert halls or trade fairs, few approaches systematically take into account the trajectory patterns for analysis. However, incorporating prior knowledge on pedestrian movement is essential to address the two questions posed above (see section 1). Existing traffic volume estimation methods, e.g., k-nearest neighbour [15,16] and standard Gaussian process regression [17], do not take into account this form of expert knowledge and thus may not effectively provide accurate estimations, e.g. in case of side corridors.

To estimate the traffic volume at unmeasured locations, we propose in [18] a nonparametric Bayesian method, Gaussian Processes (GP) with a random-walk based trajectory kernel. The method explores not only the commonly used information known from the literature, e.g. traffic network structures and recorded presence counts NV at some measurement locations, but also the move preferences of pedestrians (trajectory patterns) collected from the sensors. As firstly introduced in [18], we provide here a brief discussion on the GP approach for quantity estimation and sensor placement. Consider a traffic network $\mathcal{G}(\mathbf{V}, \mathbf{E})$ with N vertices and M edges. For some of the edges, we observe the pedestrian quantities, denoted as $\mathbf{y} = \{y_s := NV(\tilde{v}_s, \Delta t) : s = 1, \ldots, S\}$. Additionally, we have information about the major pedestrian movement patterns $\mathcal{T} = \{T_1, T_2, \ldots\}$ over the traffic network, collected from the local experts or the tracking technology (e.g. Bluetooth). The pedestrian quantity estimation over traffic networks can be viewed as a link prediction problem, where the predicted quantities associated with links (vertices) are continuous variables.

In the literature on statistical relational learning [19,20], a commonly used GP relational method is to introduce a latent variable to each vertex, and to model the values of edges as a function of latent variables of the involved vertices, e.g. [21,22]. Although these methods have the advantage that the problem size remains linear in the size of the vertices, it is difficult to find appropriate functions to encode the relationship between the variables of vertices and edges for different applications.

The observed pedestrian quantities (within a time interval Δt) are conditioned on the latent function values with Gaussian noise ϵ_i: $y_i = f_i + \epsilon_i, \;\; \epsilon_i \sim \mathcal{N}(0, \sigma^2)$. As mathematical form and parameters of the function are random and unknown, f_i is also unknown and random. For an infinite number of vertices, the function values $\{f_1, f_2, \ldots\}$ can be represented as an infinite dimensional vector. Within a nonparametric Bayesian framework, we assume that the infinite dimensional random vector follows a Gaussian process (GP) prior with mean function $m(x_i)$ and covariance function $k(x_i, x_j)$ [23]. In turn, any finite set of function values $\mathbf{f} = \{f_i : i = 1, ..., N\}$ has a multivariate Gaussian distribution with mean and covariances computed with the mean and covariance functions of the GP [23].

Without loss of generality, we assume zero mean so that the GP is completely specified by the covariance function. Formally, the multivariate Gaussian prior distribution of the function values $\mathbf{f}$ is written as $P(\mathbf{f}|\mathbf{X}) = \mathcal{N}(0, K)$, where K denotes the $N \times N$ covariance matrix, whose ij-th entry is computed in terms of the covariance function. If there are vertex features $\mathbf{x} = \{x_1, ..., x_N\}$ available, e.g., the spatial representation of traffic edges, a typical choice for the covariance function is the squared exponential kernel with isotropic distance measure.

Since the latent variables $\mathbf{f}$ are linked together into an edge graph $\mathcal{G}$, it is obvious that the covariances are closely related to the network structure: the variables are highly correlated if they are adjacent in $\mathcal{G}$, and vice versa. Therefore we can also employ graph kernels, e.g. the regularized Laplacian kernel, as the covariance functions:

$$K = \left[\beta(L + I/\alpha^2)\right]^{-1}, \tag{1}$$

where α and β are hyperparameters. L denotes the combinatorial Laplacian, which is computed as $L = D - A$, where A denotes the adjacency matrix of the graph $\mathcal{G}$. D is a diagonal matrix with entries $d_{i,i} = \sum_j A_{i,j}$.

Although graph kernels have some successful applications to public transportation networks [17], there are probably limitations when applying the network-based kernels to the scenario of closed environments: the pedestrians in a train station or a shopping mall have favorite or commonly used routes, they are not randomly distributed on the networks. In a train station, the pedestrian flow on the main corridor is most likely unrelated to that on the corridors leading to the offices, even if the corridors are adjacent. To incorporate the information of the move preferences (trajectory patterns, collected from the local experts or tracking technology) into the model, we explore a graph kernel inspired with the diffusion process [24]. Assume that a pedestrian randomly moves on the edge graph $\mathcal{G}$. From a vertex i he jumps to a vertex j with $n^k_{i,j}$ possible random walks of length k, where $n^k_{i,j}$ is equal to $[A^k]_{i,j}$. Intuitively, the similarity of two vertices is related to the number and the length of the random walks between them. Based on diffusion process, the similarity between vertices v_i and v_j is defined as

$$s(v_i, v_j) = \left[\sum_{k=1}^{\infty} \frac{\lambda^k}{k!} A^k\right]_{ij}, \tag{2}$$

where $0 \leq \lambda \leq 1$ is a hyperparameter. All possible random walks between v_i and v_j are taken into account in similarity computation, however the contributions of longer walks are discounted with a coefficient $\lambda^k/k!$. The similarity matrix is not always positive semi-definite. To get a valid kernel, the combinatorial Laplacian is used and the covariance matrix is defined as [24]:

$$K = \left[\sum_{k=1}^{\infty} \frac{\lambda^k}{k!} L^k\right] = \exp(\lambda L)\ . \tag{3}$$

On a traffic network within closed environment, the pedestrian will move not randomly, but with respect to a set of trajectory patterns and subpatterns denoted as sequences of vertices, e.g.,

$$\left\{ \begin{array}{l} T_1 = v_1 \to v_3 \to v_5 \to v_6, \\ T_2 = v_2 \to v_3 \to v_4, \\ \quad \ldots \end{array} \right\}. \tag{4}$$

Each trajectory pattern T_ℓ can also be represented as an adjacency matrix in which $\hat{A}_{i,j} = 1$ iff $v_i \to v_j \in T_\ell$ or $v_i \leftarrow v_j \in T_\ell$. The subpatterns are subsequences of the trajectories. For example, the subpatterns of T_1 are $\{v_1 \to v_3, v_3 \to v_5, v_5 \to v_6, v_1 \to v_3 \to v_5, v_3 \to v_5 \to v_6\}$. Given a set of trajectory patterns $\mathcal{T} = \{T_1, T_2, \ldots\}$, a random walk is valid and can be counted in similarity computation, if and only if all steps in the walk belong to $\mathcal{T}$ and subpatterns of $\mathcal{T}$. Thus we have

$$\begin{aligned} \hat{s}(v_i, v_j) &= \left[\sum_{k=1}^{\infty} \frac{\lambda^k}{k!} \hat{A}^k \right]_{ij}, & \hat{K} &= \left[\sum_{k=1}^{\infty} \frac{\lambda^k}{k!} \hat{L}^k \right] = \exp(\lambda \hat{L}) \\ \hat{A} &= \sum_{\ell} \hat{A}_\ell, & \hat{L} &= \hat{D} - \hat{A}, \end{aligned} \tag{5}$$

where $\hat{D}$ is a diagonal matrix with entries $\hat{d}_{i,i} = \sum_j \hat{A}_{i,j}$.

For pedestrian quantities $\mathbf{f}_u$ at unmeasured locations u, the predictive distribution can be computed as follows. Based on the property of GP, the observed and unobserved quantities $(\mathbf{y}, \mathbf{f}_u)^T$ follows a Gaussian distribution

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_u \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \hat{K}_{\overline{u},\overline{u}} + \sigma^2 I & \hat{K}_{\overline{u},u} \\ \hat{K}_{u,\overline{u}} & \hat{K}_{u,u} \end{bmatrix}\right), \tag{6}$$

where $\hat{K}_{u,\overline{u}}$ is the corresponding entries of $\hat{K}$ between the unmeasured vertices u and measured ones $\overline{u}$. $\hat{K}_{\overline{u},\overline{u}}$, $\hat{K}_{u,u}$, and $\hat{K}_{\overline{u},u}$ are defined equivalently. I is an identity matrix of size $|\overline{u}|$. Finally the conditional distribution of the unobserved pedestrian quantities is still Gaussian with the mean m and the covariance matrix Σ:

$$\begin{aligned} m &= \hat{K}_{u,\overline{u}} (\hat{K}_{\overline{u},\overline{u}} + \sigma^2 I)^{-1} \, \mathbf{y} \\ \Sigma &= \hat{K}_{u,u} - \hat{K}_{u,\overline{u}} (\hat{K}_{\overline{u},\overline{u}} + \sigma^2 I)^{-1} \, \hat{K}_{\overline{u},u} \, . \end{aligned}$$

Besides pedestrian quantity estimation, incorporating trajectory patterns also enables effectively finding sensor placements that are most informative for traffic estimation on the whole network. To identify the most informative locations $\mathcal{I}$, we employ the exploration strategy, maximizing mutual information [25]

$$\arg\max_{\mathcal{I} \subset \mathbf{V}} H(\mathbf{V} \backslash \mathcal{I}) - H(\mathbf{V} \backslash \mathcal{I} \mid \mathcal{I}) \, . \tag{7}$$

It is equal to finding a set of vertices $\mathcal{I}$ which maximally reduces the entropy of the traffic at the unmeasured locations $\mathbf{V}\backslash\mathcal{I}$. Since the entropy and the conditional entropy of Gaussian variables can be completely specified with covariances, the selection procedure is only based on covariances of vertices, and does not involve any pedestrian quantity observations. To solve the optimization problem, we employ a poly-time approximate method [25]. In particular, starting from an empty set $\mathcal{I} = \emptyset$, each vertex is selected with the criterion:

$$v_* \leftarrow \arg\max_{v \in \mathbf{V}\backslash\mathcal{I}} \; H_\epsilon(v \mid \mathcal{I}) - H_\epsilon(v \mid \overline{\mathcal{I}}) \;, \tag{8}$$

where $\overline{\mathcal{I}}$ denotes the vertex set $\mathbf{V}\backslash(\mathcal{I} \cup v)$. $H_\epsilon(x|Z) := H(x|Z')$ denotes an approximation of the entropy $H(x|Z)$, where any element z in $Z' \subset Z$ satisfies the constraint that the covariance between z and x is larger than a small value ϵ. Within the GP framework, the approximate entropy $H_\epsilon(x|Z)$ is computed as

$$\begin{aligned} H_\epsilon(x \mid Z) &= \frac{1}{2} \ln 2\pi e \sigma^2_{x|Z'} \\ \sigma^2_{x|Z'} &= \hat{K}_{x,x} - \hat{K}^T_{x,Z'} \hat{K}^{-1}_{Z',Z'} \hat{K}_{x,Z'} \;. \end{aligned} \tag{9}$$

The term $\hat{K}_{x,Z'}$ is the corresponding entries of $\hat{K}$ between the vertex x and a set of vertices Z'. $\hat{K}_{x,x}$ and $\hat{K}_{Z',Z'}$ are defined equivalently. Given the informative trajectory pattern kernel, the pedestrian quantity observations at the vertices selected with the criterion (8) can well estimate the situation of the whole network. Refer to [18] for more details.

4 Real World Application

In this section, we test our approach on a dataset collected through Bluetooth tracking technology [1]. The analysis is inspired by the workflow presented in [13]. Instead of applying two phases we conduct our experiment in three consecutive phases.

- The *field study phase* is performed during (1) survey design and (2) data collection.
- The second *visual analysis phase* is conducted within the (3) data preparation, aggregation and visual analysis.
- In the *knowledge discovery phase* we conduct the (4) data mining step.

Next, each of the steps is described and experiments to the previously described sensor placement strategy are performed.

4.1 Field Study Phase

For data collection a mesh of 17 Bluetooth sensors has been deployed within a soccer stadium (Stade des Costières, Nîmes at France) during a soccer match on

05.08.2011. The three-dimensional sensor placement is depicted in Figure 1. All Bluetooth enabled devices (e.g. smartphones or intercoms) that pass at one of the sensors (more precisely its footprint) trigger the creation of a datalog entry consisting of the timestamp, the sensor identifier (which denotes the position), the radio signal strength and a hashed identifier for this particular device [26].

The range of the sensors is approximately 15 meters, thus there remain unobserved regions in the stadium as well as overlapping areas. Whenever a Bluetooth enabled (i.e. visible) mobile device traverses multiple sensors, it becomes re-detected. In this way, transition times as well as movement patterns can be reconstructed. However, the recorded data is episodic (see Section 2 for specifics on *Episodic Movement Data*) as it provides uncertainties on continuity, accuracy as well as coverage [9].

We recorded 47,589 data points from 553 different devices at 17 distinct locations. The average number of distinct visited sensor locations is 4.37, the median number is 2. The recorded movements have an average duration of 3 hours and 25 minutes. In total, about 14 percent of the visitors, 553 of 3898 (this official visitor number does not contain the people which worked there), have been recorded during the period of the match; thus we expect the dataset to be large enough to allow inferences from the sample to the whole population even for less frequent flows.

4.2 Visual Analysis Phase

The recorded Bluetooth tracking dataset contains sequence movement patterns, which can for example be extracted using the Teiresias algorithm [27], which was firstly applied to episodic movement data in [28]. Application of the algorithm reveals that the most frequent pattern with more than one location starts at the main entrance and ends at a tribune (depicted in Figure 2A). The movement in the stadium thus is not a random walk but aims at a target. These individual movement preferences cause correlations among the sensor readings. Next, we visually explore the correlations contained in the soccer dataset [1]. The visual analysis of movement dependencies among discrete regions is subject of our previous work presented in [29,29]. There, the contained dependencies are represented by a Spatial Bayesian Network which connects the different regions by directed edges and associated conditional probability tables. In result, queries for co-visits of spatial regions given arbitrary (positive or negative) evidences can be answered. Next, we apply this method [30] to the presented dataset and study the contained movement preferences in detail (Figure 2).

For visualization of the three-dimensional dependencies, we created a Voronoi Dirichlet tessellation of a three-dimensional stadium model. Materials to the resulting geometries (colour and opacity) are assigned according to the probability distribution computed by the Spatial Bayesian Network. Figure 2 depicts the results of the Spatial Bayesian Network for four different queries. Red colours indicate a high visit probability; blue colours indicate a low probability. The yellow arrows in the picture mark the points of positive evidence. The picture A (in the upper-left corner) depicts the probability distribution given the evidence

that the sensor at the ground floor (sensor 34 for comparison with Figure 1) has been visited. It is remarkable that the probability on this side of the stadium is high and low in most of the other parts. The places in the other tribunes (at the bottom of the pictures) that possess a relative high probability as well as the VIP rooms and thus visited by the catering staff and prominent visitors from all tribunes after the match ended.

In the next step we examine the impact of the staff and prominent guests by change of evidence to a restricted entry within the Spatial Bayesian Network. Results are depicted in picture B. All paths that have been used by the catering crew and safety deputies are inked in red which denotes a high probability of movement. The shops possess a relatively high probability. They were located in the uppermost floor of the two towers in the left side of the picture and also in the VIP lounges. Safety deputies helped us during data collection, thus it can be seen to the right that they visited sensor location three (top of the upper left tower, compare Figure 2) in order to check its presence. In the bottom of Figure 2 we combine multiple points of evidence within the query. To the left (picture C) is a visualization of the combined probability of the visitors at the entry to the major tribune and to the VIP entry. The visitors selected by this query distribute among the major tribune and within the VIP rooms. By further addition of evidence at sensor location three, the places considered so far reach their highest conditional probability. Most likely this untypical movement pattern depicted in picture D was our movement for maintenance of the sensors. The tribune to the left shows a very low probability as it could not be traversed. The tribune on the right was open for traversing before the match began. Thus, our visual analysis reflects these circumstances and helps to understand movement behavior contained in the dataset.

After visual analysis of the recorded spatial movement correlations our further visual analysis focuses on the temporal analysis and the preparation of the dataset for the data mining (i.e. sensor placement step). Since episodic movement

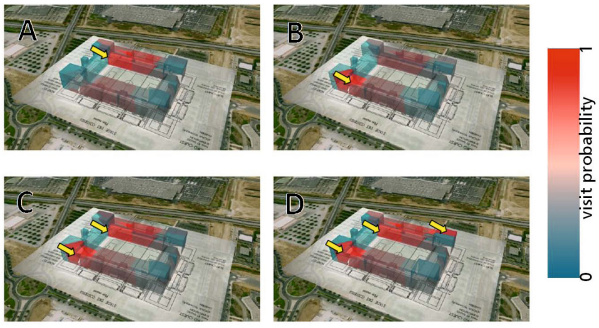
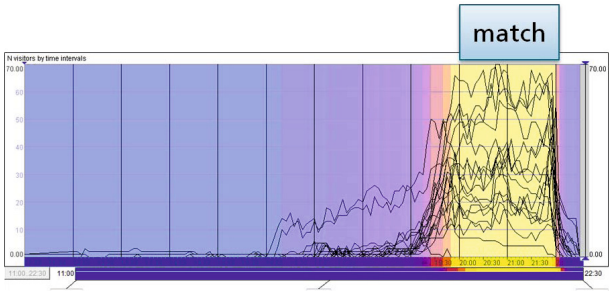
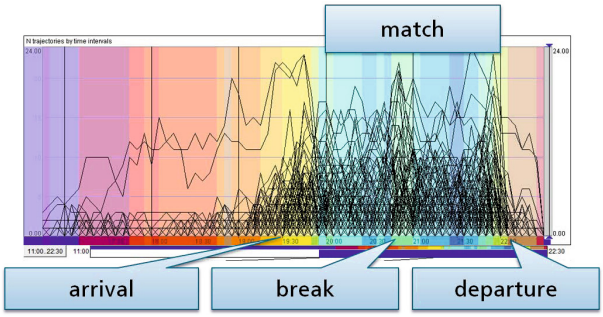


Fig. 2. Visual representation of the spatial correlations in the soccer dataset, yellow arrow denotes the evidence of the query



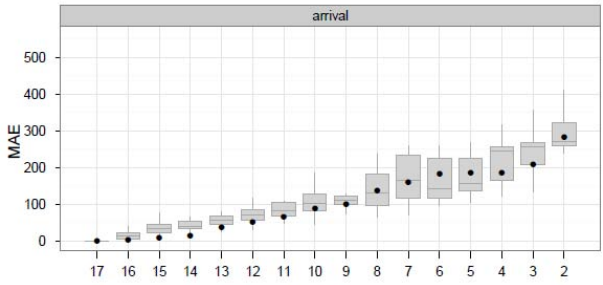
(a) Clustering of presence situations over time



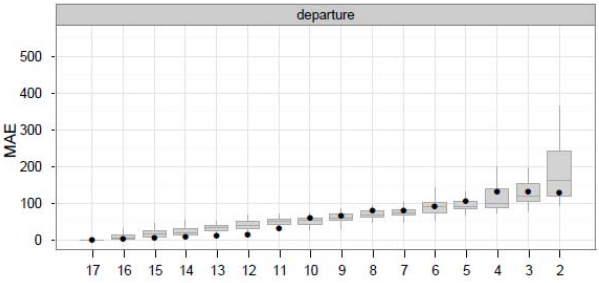
(b) Clustering of flow situations over time

Fig. 3. Temporal analysis of presence and flow situations

data contains uncertainties on individual movement, the proposed approach in [9] is the spatio-temporal aggregation of *presence* and *moves*. This results in presence and flow situations which denote for a time interval Δt the total number of *visits* for each discrete location as well as the total number of *moves* among pairs of locations. Thus, in contrast to the existing workflow for Bluetooth tracking data analysis, the soccer dataset [1] is divided into three consecutive time intervals (arrival, match, departure) derived from the clustering of presence and flows (Figure 3). In this picture the lines represent the number of persons per scanner (Figure 3a) or the numbers of persons per link among two locations (Figure 3b). The background colouring of the Figure utilizes Sammons mapping [14] and was discussed in Section 2. Based on the achieved visual analysis of the flow data (depicted in Figure 3) the time-stamps for splitting are (14:00, 20:00, 21:45, 22:00). These time intervals correspond to the three different consecutive phases of the match: arrival of the visitors, match and the departure after the match. Note that in Figure 3b (which analyses the moves of the visitors) even the break of the match is visible. Movement of the stadium visitors differs in each of these time spans from its successive time interval (indicated by different colours in Figure 3b). During the match there is very low movement of the visitors. Thus, we perform our sensor placement experiments for the safety critical phases of *arrival* and *departure*.



(a) arrival dataset



(b) departure dataset

Fig. 4. MAE for random (grey boxplots) and trajectory pattern kernel based sensor placement (black dots) for different number of sensors

4.3 Knowledge Discovery Phase

The Gaussian process based sensor placement algorithm (Section 3) is applied to the two previously separated datasets (arrival and departure of the visitors). Thus, the recorded movement sequences of the visitors (studied in Figure 2) are considered as movement patterns. All of the recorded patterns are treated equally (we remove duplicates) without any weighting. The recorded *counts of visits* per sensor are subject for quantity estimation. This is also an important difference from our work presented in [18] where we model *counts of flows*.

The quantity estimation is performed with different numbers of sensors, starting from 17 up to 2. In each test we apply our sensor placement algorithm among the predefined locations chosen in the given dataset. The performance of the placement is then compared to random sensor placement (run 35 times each). The quantity estimation error is measured in mean absolute error MAE. Figure 4 depicts the performance for different numbers of sensors. The placement of 17 sensors (to the left) equals to the case were all 17 previously placed sensors (contained in the dataset) are used. In the next step, one of the sensors is omitted. The grey boxplot denotes performance for its random selection, the black dot the performance of our kernel based placement strategy.

The tests show that when omitting up to 6 of the applied sensors (35%) in the sensor mesh, our placement still outperforms random placement and has an acceptable absolute prediction error of 80 persons (2% of the total number of 3,898 visitors[1]).

5 Conclusion and Summary

The paper addressed the visitor quantity estimation in an event monitoring scenario under constraints (i.e., a bounded number of sensors). Thus we tackled the following two challenges (1) pedestrian quantity estimation from a relatively small number of empirical measurements, and (2) placement of the constrained number of quantity sensors. We proposed a novel method to determine where a fixed number of automatic pedestrian quantity sensors is to be located during a mass event in order to get an adequate estimate on the presence of the visitors within the site. Note that we considered here *counts of presence*, instead of *counts of moves*, which is subject to [18].

Our proposed method incorporates trajectory patterns for automatic sensor placement and quantity estimation. Real world experiments at a soccer stadium dataset show that our method holds potential for automatically determined sensor number reduction.

Future work may focus on reduction of communication costs among the sensor network, inclusion of mobile sensors (e.g. mobile Bluetooth sensors [31]) and creation of a dynamic pedestrian model.

Acknowledgments. The work was supported by the European Project Emergency Support System (ESS 217951) and the Fraunhofer ATTRACT Fellowship STREAM.

References

1. Liebig, T., Kemloh Wagoum, A.U.: Modelling microscopic pedestrian mobility using bluetooth. In: Proc. of the Fourth International Conference on Agents and Artificial Intelligience - ICAART 2012, pp. 270–275. SciTePress (2012)
2. Bruno, R., Delmastro, F.: Design and Analysis of a Bluetooth-Based Indoor Localization System. In: Conti, M., Giordano, S., Gregori, E., Olariu, S. (eds.) PWC 2003. LNCS, vol. 2775, pp. 711–725. Springer, Heidelberg (2003)
3. Liebig, T., Xu, Z.: Pedestrian monitoring system for indoor billboard evaluation. Journal of Applied Operational Research 4(1), 28–36 (2012)
4. Liebig, T.: A general pedestrian movement model for the evaluation of mixed indoor-outdoor poster campaigns. In: Proc. of the Third International Conference on Applied Operation Research, ICAOR 2011, pp. 289–300. Tadbir Operational Research Group Ltd. (2011)

[1] Info to the match at http://www.foot-national.com/match-foot-nimesvannes-32912.html, last accessed 08/05/2012

5. Liebig, T., Stange, H., Hecker, D., May, M., Körner, C., Hofmann, U.: A general pedestrian movement model for the evaluation of mixed indoor-outdoor poster campaigns. In: Proc. of the Third International Workshop on Pervasive Advertising and Shopping (2010)
6. Li, M., Konomi, S., Sezaki, K.: Understanding and modeling pedestrian mobility of train-station scenarios. In: Sabharwal, A., Karrer, R., Zhong, L. (eds.) WINTECH, pp. 95–96. ACM (2008)
7. Pels, M., Barhorst, J., Michels, M., Hobo, R., Barendse, J.: Tracking people using Bluetooth. Implications of enabling Bluetooth discoverable mode. Technical report, University of Amsterdam (2005)
8. Hallberg, J., Nilsson, M., Synnes, K.: Positioning with Bluetooth. In: 10th International Conference on Telecommunications, vol. 2, pp. 954–958 (2003)
9. Andrienko, N., Andrienko, G., Stange, H., Liebig, T., Hecker, D.: Visual analytics for understanding spatial situations from episodic movement data. KI - Künstliche Intelligenz, 241–251 (2012)
10. Utsch, P., Liebig, T.: Monitoring Microscopic Pedestrian Mobility Using Bluetooth. In: Proceedings of the 8th International Conference on Intelligient Environments, pp. 173–177. IEEE Press (2012)
11. Hagemann, W., Weinzerl, J.: Automatische Erfassung von Umsteigern per Bluetooth-Technologie. In: Nahverkerspraxis. Springer, Heidelberg (2008)
12. Leitinger, S., Gröchenig, S., Pavelka, S., Wimmer, M.: Erfassung von Personenströmen mit der Bluetooth-Tracking-Technologie. In: Angewandte Geoinformatik 2010, 15th edn. Addison Wesley Longman Inc., New York (2010)
13. Stange, H., Liebig, T., Hecker, D., Andrienko, G., Andrienko, N.: Analytical Workflow of Monitoring Human Mobility in Big Event Settings using Bluetooth. In: Proceedings of the 3rd International Workshop on Indoor Spatial Awareness, pp. 51–58. ACM (2011)
14. Sammon, J.W.: A nonlinear mapping for data structure analysis. IEEE Transaction on Computers 18(5), 401–409 (1969)
15. Gong, X., Wang, F.: Three improvements on knn-npr for traffic flow forecasting. In: Proceedings of the 5th International Conference on Intelligent Transportation Systems, pp. 736–740. IEEE Press (2002)
16. May, M., Hecker, D., Körner, C., Scheider, S., Schulz, D.: A vector-geometry based spatial knn-algorithm for traffic frequency predictions. In: Data Mining Workshops, International Conference on Data Mining, pp. 442–447. IEEE Computer Society, Los Alamitos (2008)
17. Neumann, M., Kersting, K., Xu, Z., Schulz, D.: Stacked gaussian process learning. In: Proceeding of the 9th IEEE International Conference on Data Mining, ICDM 2009, pp. 387–396. IEEE Computer Society (2009)
18. Liebig, T., Xu, Z., May, M., Wrobel, S.: Pedestrian Quantity Estimation with Trajectory Patterns. In: Flach, P.A., De Bie, T., Cristianini, N. (eds.) ECML PKDD 2012, Part II. LNCS, vol. 7524, pp. 629–643. Springer, Heidelberg (2012)
19. De Raedt, L.: Logical and Relational Learning. Springer (2008)
20. Getoor, L., Taskar, B. (eds.): Introduction to Statistical Relational Learning. The MIT Press (2007)
21. Yu, K., Chu, W., Yu, S., Tresp, V., Xu, Z.: Stochastic relational models for discriminative link prediction. In: Neural Information Processing Systems (2006)
22. Chu, W., Sindhwani, V., Ghahramani, Z., Keerthi, S.: Relational learning with gaussian processes. In: Neural Information Processing Systems (2006)
23. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. The MIT Press (2006)

24. Kondor, R.I., Lafferty, J.D.: Diffusion kernels on graphs and other discrete input spaces. In: Proceeding of the International Conference on Machine Learning, pp. 315–322 (2002)
25. Krause, A., Guestrin, C., Gupta, A., Kleinberg, J.: Near-optimal sensor placements: maximizing information while minimizing communication cost. In: Proceedings of the 5th International Conference on Information Processing in Sensor Networks, IPSN 2006, pp. 2–10. ACM, New York (2006)
26. National Institute of Standards and Technology: Secure Hash Standard. National Institute of Standards and Technology, Washington, Federal Information Processing Standard 180-2 (2002)
27. Rigoutsos, I., Floratos, A.: Combinatorial pattern discovery in biological sequences: The teiresias algorithm. Bioinformatics 14(1), 55–67 (1998)
28. Kisilevich, S., Keim, D., Rokach, L.: A novel approach to mining travel sequences using collections of geotagged photos. In: Painho, M., Santos, M.Y., Pundt, H., Cartwright, W., Gartner, G., Meng, L., Peterson, M.P. (eds.) Geospatial Thinking. Lecture Notes in Geoinformation and Cartography, pp. 163–182. Springer, Heidelberg (2010)
29. Liebig, T., Körner, C., May, M.: Fast visual trajectory analysis using spatial bayesian networks. In: ICDM Workshops, pp. 668–673. IEEE Computer Society (2009)
30. Liebig, T., Körner, C., May, M.: Scalable sparse bayesian network learning for spatial applications. In: ICDM Workshops, pp. 420–425. IEEE Computer Society (2008)
31. Naini, F.M., Dousse, O., Thiran, P., Vetterli, M.: Population size estimation using a few individuals as agents. In: Proceedings of the International Symposium on Information Theory, pp. 2499–2503. IEEE (2011)

Users as Smart Sensors: A Mobile Platform for Sensing Public Transport Incidents*

Cristian Tanas and Jordi Herrera-Joancomartí

Universitat Autònoma de Barcelona
ctanas@deic.uab.cat, jordi.herrera@uab.cat

Abstract. Sensor networks may become a key element in a smart city in order to collect and provide information to its citizens. In this paper, we propose a new mobile phone sensing application, *Incidències 2.0*, that helps users notify and stay informed about the incidents of the public rail network in the Barcelona metropolitan area. The application takes advantage of the widespread use of smartphones combined with their sensing capabilities to gather sensory data from the environment and then send the sensed information back to a central data collection facility using cellular network technology. Data retrieved from the application provided by real users allows us to make a first analysis on the potentials of this new sensor network paradigm.

1 Introduction

Smart cities use technology and network infrastructure to improve economic and political efficiency and enable social, cultural and urban development. Though there are many factors involved within a smart city, their citizen engagement and the necessity of sensors to monitor the city's activities are becoming key concepts towards a successful deployment.

There is an undeniable need for sensor networks in a smart city to collect and provide information to its citizens. Sensor networks have become one of the most active areas in networking research over the last decade, providing overwhelming potential for information collection and processing in a wide range of environments. The state of the art approaches in sensor networking include a limited number of static devices, usually wirelessly connected, spanned over a pre-determined geographical area gathering evanescent information of the environment surrounding them.

Nevertheless, we can not overlook the increasing popularity and huge potential of smartphones to build a new generation of sensor networks, targeting daily life activities of individuals and the environment surrounding them. Indeed, modern smartphones besides being sophisticated computing platforms, include a wide range of capabilities, like computing (CPU, data storage,...), communication (UMTS, WiFi, Bluetooth) and sensing (positioning -GPS-, motion -accelerometer-, image -camera- , audio -microphone-). In addition, smartphones development is exploding, and competition between Apple and Google expands over 74% of the market share with approximately 149 millions of devices sold during the 4th quarter of 2011 according to Gartner, Inc. [1].

* This work has been partially supported by the Spanish Government through project TIN2010-15764 N-KHRONOUS and the UAB grant PIF 472-01-1/E2010.

J. Nin and D. Villatoro (Eds.): CitiSens 2012, LNAI 7685, pp. 81–93, 2013.

Therefore, we can take advantage of the widespread use of smartphones combined with their sensing capabilities to gather sensory data from the environment and then send the sensed data back to data collection facilities using cellular network technology. Furthermore, it might be useful to have individuals participating in the sensing tasks. Surrounding environment detection, information processing, or great communication skills are just some of the qualities that individuals posses. Therefore, we can take advantage of both available sensors in a smartphone and the smartphone's owner intelligence to acquire better knowledge on long-lasting features of the landscape. Users can provide additional information to sensor readings, such as natural language description of the environment or location-tagged images, thereby provisioning researchers with a substantial wealth of data. When relying on users to act as sensors we could refer to them as *smart sensors*, and we will refer to this type of sensor networks as *smart sensor networks* (SSN).

SSN can help overcome many of the limitations of existing proposals in wireless sensor networks, which require physical deployment and customized node management, in addition to complex communication protocols. However, the new opportunities and benefits offered by modern smartphones as sensing devices come at a price. Bringing together geographically and sociologically unrelated individuals to create a community that performs tasks for a greater good brings up front new challenges and security issues that might have a strong impact on the overall performance of the network. Sensor network managers, now have to deal with potential sabotage (intentional or unintentional) from the smartphones users. How to derive trust in the sensor readings provided by a crowd of volunteer individuals becomes an important research question in these environments. Moreover, to engage as many users to participate in the sensor network's sensing tasks is a major challenge since usually, device owners are reluctant to share their valued resources if no direct benefits are perceived.

In this paper, we present *Incidències 2.0*, a SSN application[1] that allows users to notify and keep abreast of any incident that affects the rail public transport network nearby Barcelona. Although the application is built on top of a general framework that may allow more general sensing tasks, we would like to evaluate the correctness of our framework by developing a real, specific and useful application for that framework that will provide us with real user data in order to analyse the possibilities of a real deployed SSN.

The paper is organized as follows. In Sect. 2, we review the existing proposals in which users take part as sensors entities. Section 3 introduces a new sensing application, identifying its main functionalities. In Sect. 4 we present the architecture and modular design of the proposed framework. The data obtained from the proposed platform is analysed in Sect. 5. Finally, Sect. 6 concludes the paper.

2 State of the Art

Smart sensor networks have a large number of potential applications. However there are just a few proposals that leverage the idea of having sensing tasks relayed to consumer-owned smartphones, and the majority were developed for experimental purposes.

The SSN application spectrum ranges from CO_2 emission monitoring [7] to patient health monitoring systems where smartphones are used in combination with wireless

[1] http://www.incidencies.org

(bio) sensors to monitor a patient's vital signs [10], passing through a longer list of location-based services, such as traffic accidents detection and situational awareness provisioning to first responders [15], traffic conditions monitoring [13], or real-time trail network update for hikers and mountaineers [14]. In addition, a built-in GPS receiver and an accelerometer can be used to identify the transportation mode of an individual (i.e. walking, running, biking, or in motorized transport), as described by Reddy et al. [12].

Furthermore, smart sensor networks can provide support in emergency scenarios or environmental disasters as A. Gahran explained in an article on how citizens living in the Gulf Coast region could use their smartphones sensors, such as GPS and cameras, to enter data on the ecological impact of the Gulf oil spill, providing specialists with first hand information of this disaster [8]. This information was latter used to generate impact analysis and provide recommendations.

As for general purpose urban sensing network architecture, the MetroSense project [3] is worth mentioning. MetroSense offers a network architecture for urban-scale people-centric sensing, leveraging existing urban infrastructure and human mobility to opportunistically sense and collect data "about people and for people". Some applications include BikeNet [5] and SkiScape [4], developed as sample studies to demonstrate the usefulness of the platform.

Practically all proposals in people-centric sensing applications face the problem of data reliability. Although some approaches have been studied, such as game theory-based mechanisms, where data pollution detection is combined with punishment strategies [2,11], or entropy dynamics measuring in a descriptive distribution over the course of a game [6], they all assume that one can infer a relationship graph among the members of the sensor network, or there exist a tamper proof hardware providing a "ground truth" for data validation. Alternatively, reputation-based strategies seem to provide a promising solution. Nevertheless, all the sensing applications studies fail to provide a robust data validation mechanism.

3 Application Overview

All existing proposals in smart sensor networks and related areas, such as ad hoc networks, lack a model of users behaviour. Instead, they all assume that users will behave in a pre-determined manner, and they measure the network's performance, or make hypothesis based on this assumption. However, individuals are normally passionate and many times act in an unpredictable way. In order to break with this tradition, we developed a sensing application and a framework to facilitate the implementation of a real-time incident reporting system focused on the rail network services of the metropolitan area of Barcelona, where users in possession of a smartphone running Google's Android or Apple's iOS will act as smart sensors and information providers. Although the application is focused on the rail network services, it is developed on top of a framework that is highly scalable, allowing the implementation of an incident reporting system in other environments, such as traffic or urban furnishing damages.

Fig. 1. Incidències 2.0 common features

The application follows a client/server paradigm in which end-users will be provided with a smartphone sensing application that will provision them with a tool to easily notify the occurrence of a new event, such as a delayed train. Once the user enters all the information, it will be transmitted using the smartphone's network connection to a central data collection facility. Then, the information passes through a validation process, it is stored into the database and is made publicly available to all the devices having the sensing application installed.

3.1 Incident Notification

Incidències 2.0 allows users to notify new incidents and stay up to date of the incidents that are currently going on, through the client application installed on their smartphones.

From the main menu of the client application, users are allowed to report a new incident as shown in the left-hand side of Fig. 1. If a user chooses to report a new rail network incident, he or she must provide information about the rail network service and station where the incident is taking place, the event that caused it, an evaluation of the incident's severity. Optionally, a textual description can also be added. On the other hand, if the user chooses to report an incident of any other type, he or she will be asked introduce a description, which is now mandatory, and make an evaluation of the event's severity. Once the user introduced all the information, it is sent jointly with the current date and time and the user's GPS position, if available, to the Incident Management Center (IMC).

In addition to report a new incident, users can confirm incidents already reported by other participants in the sensor network. In the confirmation process, users are allowed to specify an optional comment or description, which is included in the incident information and can be later viewed in the detailed description of the incident.

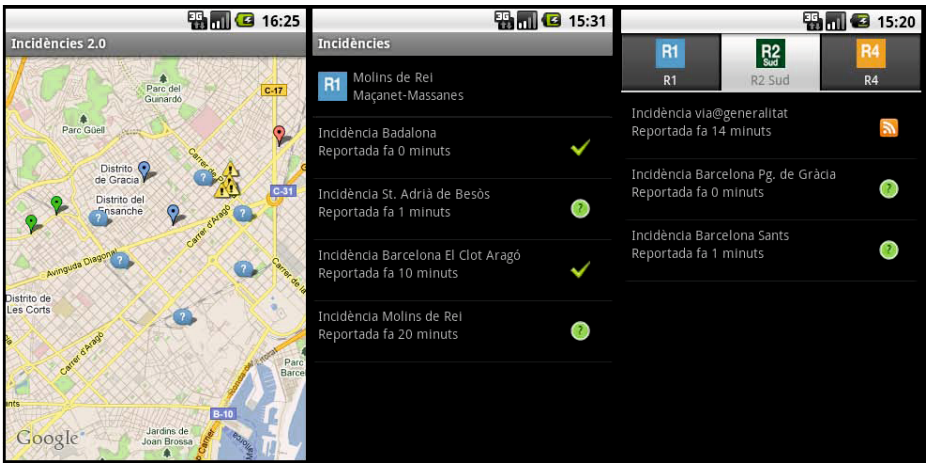


Fig. 2. Incidències 2.0 incidents visualisation options

3.2 Incident Visualisation

Incidències 2.0 provides to all application users the information of all the incidents that are taking place in a precise moment of time, offering various visualisation modes through the client application that the users have installed on their smartphones (Fig. 2).

Incidents can be seen as markers on a map where different colors and different icons are used to distinguish between different types of incidents. Users can select a map marker by tapping on that marker, which will open an alternate menu that offers the possibility to access more information about that incident, or to upload new information regarding it. This feature provides users with a quick and effective way of consulting current incidents in a given region. However, users that usually travel using the rail network system, tend to take just one railway line or a combination of a few of them. Thereby, they should be able to filter the existing incidents according to their preferences. To satisfy this requirement, *Incidències 2.0* allows users to select a specific railway line and see only the incidents affecting that line as a list (central part of Fig. 2). Tapping on one incident in the list will bring to front a detailed view about that incident, from which the user can upload new information or consult the incident's location on the map. If the user combines different railway services or lines, the application gives the possibility to define up to three favourite lines to follow (right-hand side of Fig. 2), so that the user can have a quick access only to the information that might have an impact on his or her quotidian travels.

4 Application Architecture

As we have already mention, *Incidències 2.0* has been developed through a more general framework that follows a client/server architecture with a smartphone sensing application on the client side and the Incident Management Center as the server application.

4.1 The Smartphone Sensing Application

We have developed a sensing application that runs on both Android and iOS-based devices. It implements a data gathering module, which relies on the smartphone's owner as a sensor using its GPS receiver, the current time and date, and a data visualisation module, which connects with the IMC to retrieve all existing incidents.

The Android version of the application was developed using the Android SDK and the Java programming language, while the iPhone version was developed as a native iOS application, based on the Cocoa Touch framework and using the Objective-C programming language. Both versions interact with the device's localisation services to retrieve the smartphone's current GPS position, and with the Google Maps API to offer a map-based visualisation service. Furthermore, the application takes advantage of the device's Internet connection to send the collected data back to a central server.

4.2 The Incident Management Center (IMC)

The IMC is the central part of the framework and it is responsible of processing, validating, and storing the incident notifications provided by end-users. It follows a modular design and it is composed of the following five modules (interrelated as shown in Fig. 3):

1. *Incident Definition Module (IDM)*
2. *Incident Reception and Triage Module (IRTM)*
3. *Data Validation Module (DVM)*
4. *Public Relationship Module (PRM)*
5. *Data Storage Module (DSM)*

Incident Definition Module (IDM). A type of incident is described by an XML codification schema and a set of attributes (i.e. available information regarding the incident) and actions associated with the incident. Although at present time, the application framework focuses on incidents in the rail network services, it is able to deal with new types of incidents through a request sent to the Incident Definition Module specifying the XML codification schema of the new type of incident, along with the attributes and actions associated with it.

Incident Reception and Triage Module (IRTM). The IRTM is the front-end interface of the IMC, allowing end-users to communicate with the central servers to report a new incident or retrieve the existing ones. It also performs a triage phase to check if the type of incident reported or requested is supported by the platform, that is, if it has been previously registered through the Incident Definition Module. In addition, this module is responsible for delivering information about on going incidents, as requested by end-users, and for managing confirmations.

After the received notifications passed the triage stage, the information is forwarded to the Data Validation Module, where it runs through a data validation process.

Data Validation Module (DVM). The DVM provides a validation scheme for the incoming notifications based on the users' reputation and collective knowledge. However, how to derive trust from the information collected by a crowd of volunteer individuals is a

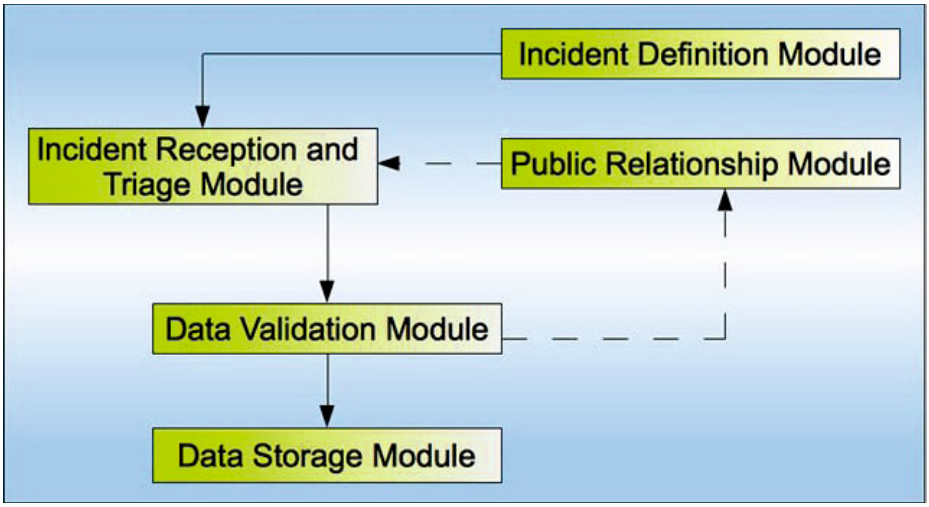


Fig. 3. Modular design of the Incident Management Center

major challenge in people-centric sensing applications. It is not straightforward to assess if a notification received from a user is valid and corresponds to a real event happening at that precise moment, or on the contrary, it provides dishonest or counterfeit information. In Subsect. 4.3 we provide some ideas on how to achieve such data validation.

Public Relationship Module (PRM). *Incidències 2.0* relies on end-users to provide incident notifications and this information is made publicly available for all the other users having the application installed. However, by nature, individuals are selfish and they are not inherently motivated to collaborate in the sensing tasks unless direct benefits are perceived from their participation. Then, selfish users will only consult currently active incidents, but will refuse to provide incident notifications. Therefore, the PRM must provide and manage the incentives that should be provided to end-users to stimulate their cooperation. The PRM could certainly exploit data from the IRTM or the DVM, such as reputation or user behaviour and query patterns to design cooperation protocols that better adapt to the user's needs.

Data Storage Module (DSM). After the incident notifications are processed and validated, the information has to be stored in a database, so it can be retrieved and used in the future by the other modules, or statistically analysed. The database scheme was implemented following a relational database model and using the MySQL database manager system. The design respects a modular scheme, so that new types of incidents can be easily incorporated and stored into the database.

The information stored into the database includes users' credentials and reputation value, the different incident notifications and confirmations, and also the queries performed by the users. In addition, the database must reflect the results of the validation process, results that are stored jointly with the incident notification scheme.

4.3 Reliable Data Readings

Data validation can be performed through a reputation system and collective knowledge to ensure the reliability of the incident notifications sent by the users. The user's reputation provides a measure of his or her credibility within the system. Thereby, the incident notifications received from users with a high reputation value, exceeding a given threshold, will be considered as valid. On the other hand, if the user does not benefit from high credibility within the community, different observations from different users of the same incident will help establish the validity of the incident if a pre-defined number of observations is accounted. Collective knowledge is handled by the application framework through the client application, offering the ability to confirm incidents previously notified by other users. Therefore, if a pre-define threshold of confirmations is reached, then the incident will be considered as valid, and at the same time the reputation of the users that observed that incident will be increased.

At the beginning, each user starts with a neutral reputation score. They can increase their reputation by participating in the validation process of an incident, either notifying the incident or confirming it.

Computing User's Reputation Scores. We need to quantify the reputation of a user and attune the threshold from which a user is considered to have a sufficiently high reputation value to be granted with total credibility. The past interactions with a given user determine up to a certain degree the future behaviour of the user. For example, if a user always reports valid incidents, it is most likely that the next time it reports an incident it would be a valid one.

We take a Bayesian system approach as described in [9]. The reputation score can be computed based upon the beta probability density function parameter tuple (α, β), where α and β represent the valid incident notifications sent by an user and the unconfirmed ones respectively. The beta PDF $f(p|\alpha, \beta)$ can be expressed using the gamma function Γ as:

$$f(p|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\alpha)} p^{\alpha-1}(1-p)^{\beta-1} \quad , \tag{1}$$

where $0 \leq p \leq 1,\ \alpha > 0,\ \beta > 0$.

The probability expectation value of the beta distribution is given by:

$$E(p) = \frac{\alpha}{\alpha+\beta} \quad . \tag{2}$$

When there is no information about the past action of a certain user, the a priori distribution is the uniform beta PDF with $\alpha = 1$ and $\beta = 1$. Then, if r valid incidents and s unconfirmed incidents are observed, the a posteriori distribution is the beta PDF with $\alpha = r + 1$ and $\beta = s + 1$. The modeled PDF expresses the uncertain probability that in the future the user will send valid incident notifications. For example, if a user sends 7 valid incident notifications and 1 that can not be verified, the probability expectation value according to (2) is $E(p) = 0.8$. This can be interpreted by saying that the relative frequency of reporting a valid incident notification in the future is somewhat uncertain, and that the most likely value is 0.8.

However, it is effortful to compute the threshold value for a high reputation score for one user since it strongly depends on context-dependant factors, such as the number of

previous incident notification received from the users or the total number of incident notification in the system. One possible solution would be to publish the incident together with a reputation index associated with the user who notified it. Nevertheless, this approach requires a further explanation for users about the meaning and usage of the reputation score.

Collective Knowledge Management. The validation process of an incident notification received from a user with a lower reputation value than the defined threshold, must combine confirmations from other users, and also notifications that refer to the same incident. Counting confirmations is straightforward since these refer to an existing incident in the database. However, in order to verify that different notifications correspond to the same incident we must verify that those notifications make reference to the same public transport service, the same railway or subway line, the same station, and they were caused by the same event. The necessary information to determine if two notifications correspond to the same incident is completely dependant on the type of incident, and the characteristics that make two notifications refer to the same incident object must be passed along with the incident type definition.

When a new incident notification is received, the first step is to check if it makes reference to an existing incident in the database. If the result from the previous query is positive, then we must verify if the notification corresponds to an unconfirmed incident or to a valid one. In the former case, we have to add the notification to the number of confirmations of this incident and validate it if the number exceed the pre-define threshold. In this case, we must also update the user's reputation accordingly. In the later case, we will increase the number of confirmations by 1 and update the user's reputation, but no change will take place in the incident's state. In addition, the validation scheme must take into account that all incidents have a limited lifetime or time-to-live (TTL). So, only those incidents that have a positive remaining TTL must be considered to review if an incoming notification corresponds to an existing incident in the database.

As in the reputation case, it is hard (if even possible) to determine a threshold value for the number of confirmations required to validate an incident since it strongly depends on context-dependant parameters, such as the total number of smart sensors or the density of users around the incident's area. Once again, a possible solution would be to publish the number of confirmations associated with an incident and let the user decide the validity of the information. As in the reputation score case, further explanation for users is required regarding the meaning and usage of the added information.

5 Application Data Analysis

Incidències 2.0 is currently in use in the Barcelona metropolitan area and the obtained data allows us to draw some analysis regarding the usage of such platform.

5.1 General Application Usage

At present time, the application has more than 3400 users registered in its database and they have performed more than 25000 queries since February the 1st, 2012. Although the application is freely available both in the Android market and in the Apple Store,

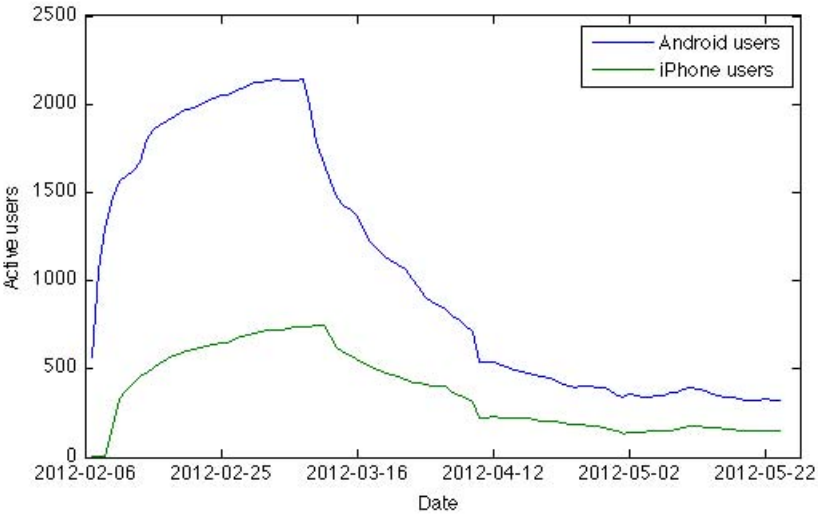


Fig. 4. Number of active users by operating system

the user distribution between platforms is not uniform, having a $74,7\%$ of users using an Android device and $25,3\%$ an iPhone.

In order to properly analyse the time evolution of the application usage, we define an *active user* at a particular time as the one that has performed at least one query in the previous month. Using this definition, Fig. 4 draws the time evolution of active users. We can see that the number of active users soared after the application was presented in a news conference, but then it fell steadily before stabilizing at around 500 active users per month. On the other hand, Fig. 5 shows the stability of the users in the sensor network. Notice that the drop out rate mainly affect new users, which means that the time users need to evaluate the utility of the application is short. This fact is important regarding the Validation Module, based on reputation, since the performance of this kind of measures improves for long term users.

5.2 Quantification Benefits of SSN

In this section, we provide some data retrieved from the application that demonstrate empirically the potential of Smart Sensor Networks.

One of the advantages of a SSN is the speed at which a sensor network can be deployed. In Table 1 we present the data deployment of *Incidències 2.0*. Notice that in less than 36 hours we were able to deploy more than 890 sensor nodes in the Barcelona metropolitan area without any economic cost[2].

On the other hand, SSN deployment allows a wide geographical spread following the patterns of population density. Figure 6 shows the geographical distribution of active users at the moment of writing this paper in the surroundings of Barcelona metropolitan area.

[2] Users were aware of the application through the mass media (after a news conference we did) plus the viral effect of social networks.

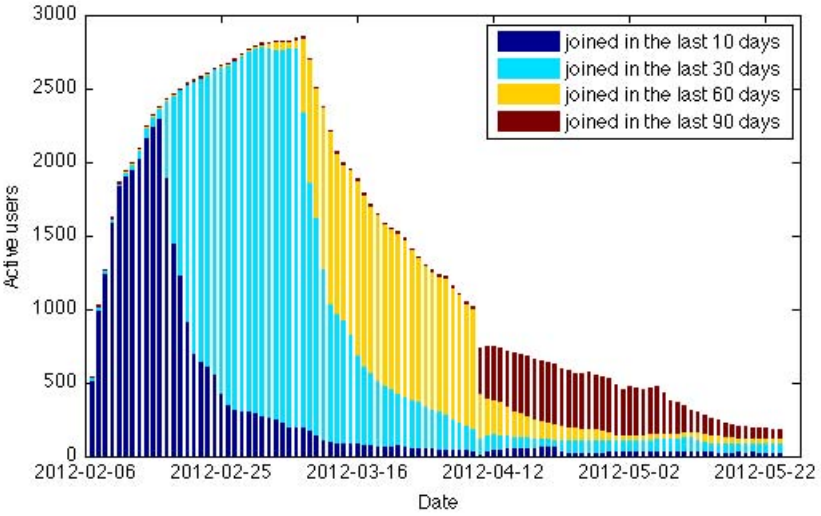


Fig. 5. Stability of the users in the sensor network

Table 1. Number of total accumulated installs by time intervals

	Days		
Hour	February the 5th	February the 6th	February the 7th
06:00	306	310	794
10:00	**306**	**326**	**890**
14:00	307	405	1001
18:00	307	571	1113
22:00	310	688	**1202**

Furthermore, sensor maintenance can be efficiently managed in a Smart Sensor Network. Traditional sensor networks entail a difficult process for software/firmware node's update process that is even harder in the case of wireless sensor networks due to the limitation of the transmission channel. However, SSN handle such process in a more simple way. For instance, 20 days after the massive deployment of our application, we add more features to allow users to sense new events. Such modification was deployed to the sensor nodes using the standard process of application upgrades defined in the Android market and Apple Store. Table 2 shows the update rate of the sensor nodes. It is to be mentioned that there were 982 active users at the time of the upgrade release. Notice that within less than a month more than 80% of the Android nodes and more than 69% of the iPhone nodes were updated and ready to use the new functionalities included for sensing new events. Even though the update process is straightforward, users can choose whether to upgrade the sensing application or not, or might not even be aware of the new release. Thereby, the latency of the update process might be higher compared to the latency of updating the nodes of a WSN.

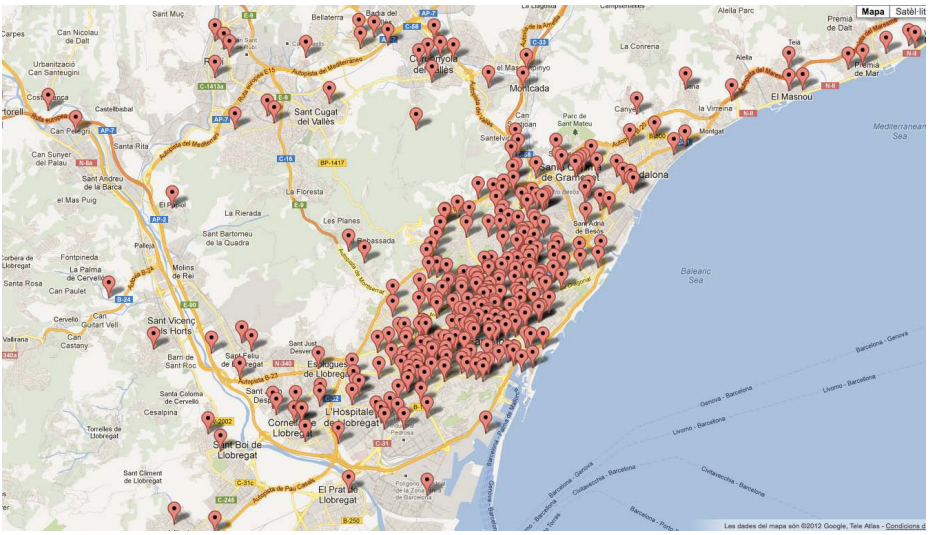


Fig. 6. Geographical deployment of sensor nodes (image showing aprox. 750 square kilometers around Barcelona)

Table 2. Percentage of updated sensor nodes by operating system

	Days			
OS	Feb. the 22nd	Feb. the 29th	Mar. the 14th	Mar. the 21st
Android	13,8%	47,5%	72,4%	80,2%
iPhone	0%	29,8%	64,7%	69,8%

6 Conclusion and Further Research

We believe that citizen-centring mobile sensing is becoming an important research area providing many interesting challenges from architectural to security and privacy specific. The wide spread and use of smartphones unfolds great potential to effectively map human-centring sensing tasks to end-user controlled smartphones. However, the architecture to support this kind of sensor networks bear little resemblance to the traditional wireless sensor network architecture discussed in the literature to date. In this paper, we have presented *Incidències 2.0*, a citizen-centric mobile sensing platform that allows individuals to report incidents in the railway transport services of the metropolitan area of Barcelona. The data obtained from the application shows us that it is possible to deploy a SSN in a very short period of time (almost 900 nodes in 36 hours) obtaining a wide geographic spread of nodes following the population geographic distribution. Nonetheless, this new vision of sensor networks raises complex privacy-related issues, that we intend to analyse and bring under discussion in further research.

References

1. Press Release. Gartner Says Worldwide Smartphone Sales Soared in Fourth Quarter of 2011 With 47 Percent Growth, Gartner Newsroom (February 2012), http://www.gartner.com/it/page.jsp?id=1924314 (last access March 26, 2012)
2. Abraham, I., Dolev, D., Gonen, R., Halpern, J.: Distributed computing meets game theory: robust mechanisms for rational secret sharing and multiparty computation. In: Proc. of the Twenty-Fifth Annual ACM Symposium on Principles of Distributed Computing, PODC 2006, pp. 53–62 (2006)
3. Campbell, A., Eisenman, S., Lane, N., Miluzzo, E., Peterson, R.: People-centric urban sensing. In: Proc. of the 2nd Annual International Workshop on Wireless Internet, WICON 2006. ACM, New York (2006)
4. Eisenman, S., Campbell, A.: SkiScape sensing. In: Proceedings of the 4th International Conference on Embedded Networked Sensor Systems, SenSys 2006, pp. 401–402. ACM (2006)
5. Eisenman, S., Miluzzo, E., Lane, N., Peterson, R., Ahn, G.S., Campbell, A.: BikeNet: A mobile sensing system for cyclist experience mapping. ACM Trans. Sen. Netw. 6(1), 6:1–6:39 (2010)
6. Eunkyung, K., Luyan, C., Yu-Han, C., Maheswaran, R.: Dynamics of Social Interactions in a Network Game. In: 2011 IEEE Third International Conference on Social Computing (Socialcom), Privacy, Security, Risk and Trust, pp. 141–148 (October 2011)
7. Froehlich, J., Dillahunt, T., Klasnja, P., Mankoff, J., Consolvo, S., Harrison, B., Landay, J.: UbiGreen: investigating a mobile tool for tracking and supporting green transportation habits. In: Proceedings of the 27th International Conference on Human Factors in Computing Systems, pp. 1043–1052. ACM (2009)
8. Gahran, A.: Reporting on the gulf oil spill from your cell phone (June 2010), http://articles.cnn.com/2010-06-11/tech/oil.spil.app_1_cell-phones-apps-geotagged?_s=PM:TECH (last access March 26, 2012)
9. Josang, A., Ismail, R., Boyd, C.: A survey of trust and reputation system for online service provision. Decision Support Systems 43, 618–644 (2007)
10. Leijdekkers, P., Gay, V.: Personal heart monitoring and rehabilitation system using smart phones. In: International Conference on Mobile Business, ICMB 2006, p. 29 (June 2006)
11. Lysyanskaya, A., Triandopoulos, N.: Rationality and Adversarial Behavior in Multi-party Computation. In: Dwork, C. (ed.) CRYPTO 2006. LNCS, vol. 4117, pp. 180–197. Springer, Heidelberg (2006)
12. Reddy, S., Mun, M., Burke, J., Estrin, D., Hansen, M., Srivastava, M.: Using mobile phones to determine transportation modes. ACM Trans. Sen. Netw. 6(2), 13:1–13:27 (2010)
13. Rose, G.: Mobile phones as traffic probes: practices, prospects and issues. IEEE Spectrum 38(1), 90–91 (2001)
14. Sayda, F.: Involving LBS users in data acquisition and update. In: Proceedings of the AGILE 2005, Conference on Geographic Information Science (2005)
15. Thompson, C., White, J., Dougherty, B., Schmidt, D.C.: Optimizing Mobile Application Performance with Model–Driven Engineering. In: Lee, S., Narasimhan, P. (eds.) SEUS 2009. LNCS, vol. 5860, pp. 36–46. Springer, Heidelberg (2009)

Author Index

Borger, Sergio 57

Cardonha, Carlos 57

Delgado-Roman, Maria del Carmen 19

Gentil, Jan Marcel 57
Gritti, Andrea 6

Herrera-Joancomartí, Jordi 81

Koch, Fernando 57

Larriba-Pey, Josep-Lluís 6
Liebig, Thomas 67

Manzoor, Jawad 6
Marés, Jordi 33
May, Michael 67
Mijnhardt, Frederik 6
Muntés-Mulero, Victor 6

Nin, Jordi 1

Paladini, Patricia 6
Pujol-Gonzalez, Marc 19

Rodríguez, Víctor 43

Serna, Jetzabel 43
Sierra, Carles 19

Tanas, Cristian 81
Torra, Vicenç 33
Torrent-Moreno, Marc 43

Villatoro, Daniel 1, 43

Xu, Zhao 67