IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

An AI approach to Collecting and Analyzing Human Interactions with Urban Environments

E. FERRARA[1], L. FRAGALE[2], G.FORTINO[2], W. SONG[3], C.PERRA[4], M. DI MAURO[5], A. LIOTTA[6]

[1]Data Science Research Centre, University of Derby, Derby, UK (e-mail: E.Ferrara@derby.ac.uk)
[2]Department of Informatics, Modeling, Electronics and Systems, University of Calabria, Rende, IT (e-mail: FragaleLuigi@gmail.com, G.Fortino@unical.it,)
[3]College of Information Technology, Shanghai Ocean University, Shanghai, China (e-mail: WSong@shou.edu.cn)
[4]Department of Electrical and Electronic Engineering, UdR CNIT, University of Cagliari, IT (e-mail: cperra@ieee.org)
[5]Department of Information and Electrical Engineering and Applied Mathematics, University of Salerno, IT (e-mail: mdimauro@unisa.it)
[6]School of Computing, Edinburgh Napier University, Edinburgh, UK (e-mail: A.Liotta@napier.ac.uk)

Corresponding author: E. Ferrara (e-mail: E.Ferrara@derby.ac.uk).

ABSTRACT Thanks to advances in Internet of Things and crowd-sensing, it is possible to collect vast amounts of urban data, to better understand how citizens interact with cities and, in turn, improve human well-being in urban environments. This is a scientifically challenging proposition, as it requires new methods to fuse objective (heterogeneous) data (e.g. people location trails and sensors data) with subjective (perceptual) data (e.g. the citizens' quality of experience collected through feedback forms). When it comes to vast urban areas, collecting statistically significant data is a daunting task; thus new data-collection methods are required too. In this work, we turn to artificial intelligence (AI) to address these challenges, introducing a method whereby the objective, sensor data is analyzed in real-time to scope down the test matrix of the subjective questionnaires. In turn, subjective responses are parsed through AI models to extract further objective information. The outcome is an interactive data analysis framework for urban environments, which we put to test in the context of a citizens' well-being project. In our pilot study, each new entry (objective or subjective) is parsed through the AI engine to determine which action maximizes the information gain. This translates into a particular question being fired at a specific moment and place, to a specific person. With our AI data collection method, we can reach statistical significance much faster, achieving (in our city-wide pilot study) a 41% acceleration factor and a 75% reduction in intrusiveness. Our study opens new avenues in urban science, with potential applications in urban planning, citizen's well-being projects, and sociology, to mention but a few cases.

INDEX TERMS Data Science, Social Science, Smart City, Data Analysis, Urban Analytics, Artificial Intelligence, Crowd Sensing

I. INTRODUCTION

PROGRESS in smart sensing, Internet of Things (IoT), crowd sensing, and artificial intelligence (AI) have made it possible to carry out multidisciplinary studies on a vast scale. Our focus herein is on urban analytics, particularly fusing objective data collected by smartphones with subjective (citizens') responses to pursue statistical significance efficiently. To deeply understand the citizens' interaction with cities, it is necessary to attain a holistic understanding of urban environment quality [1], in relation to sustainable urban and human well-being development, with minimal intrusion. It is therefore essential to develop new methods to collect, analyze and fuse heterogeneous data in real-time and at scale [2], to help administrations and competent authorities in better using the public infrastructure [3], operating it with minimal budget [4]. Research in this area arouses a great interest, with a pressing demand to develop new methods, algorithms and tools to manage smart cities and the complex processes around the citizens' well-being [5], [6], [7]. A challenging proposition is to manage quality of life in urban environments through targeted interventions based on objective and subjective data, at a time when data is becoming a commodity but is increasingly hard to make sense of. Urban analytics inherits methods from social sci-

ence and psychology to understand the human perception of the environment, measure human satisfaction, and pinpoint intervention methods [8]. It is, however, necessary to move forward from conventional methods whereby the subjective studies are based on static questionnaires, which suffer from poor statistical significance and do not make good use of the vast amount of objective data at hand. Conventional methods suffer from different problems. First, it is hard to collect sufficient samples, particularly when vast geographical areas and populations are involved, such in the mega cities. What is worse, getting the subjects' psychophysical state per se is often of little meaning if taken in isolation from the context. Another common problem is the intrusiveness of tests, whereby the interviewees' commitment (and accuracy) decreases as the test time increases. In this work, we turn to artificial intelligence (AI) to address these challenges, introducing a method through which the objective sensor data can be analyzed in real-time to scope down the test matrix of the subjective questionnaires. In turn, subjective responses are parsed through AI models to extract further objective information.

With our interactive, data analysis framework, we can determine (at any moment and point) which questions maximize the information gain, reducing the intrusiveness of subjective studies. In turn, we can extend the scope of urban studies well beyond the state-of-the-art. The idea is to use an AI chatbot to mediate the interaction between the citizen and the pot of questions we want to test.

The chatbot is integrated in the same smartphone app that is simultaneously collecting objective data from the phone sensors. The chatbot takes into account the overall context of the study (the statistical significance of each question in tandem with the context) to decide which question to fire, when to fire it, and in which location. This represents a novel approach, as it adopts AI methods to accelerate the collection of subjective data (from the citizen), through online fusion of subjective and objective data (i.e., the smartphone sensors data). This is a new way to reduce the intrusiveness of a subjective study, whilst improving its statistical significance.

The outcome is an interactive data analysis framework for urban environments that we put to test in the context of a citizens' well-being project, which we have previously investigated through static data collection methods and bulk data analysis [9], [10], [11]. In comparison with our earlier works, herein the whole subjective data extraction process is dynamically adapting to context (e.g. location), objective data (e.g. sensor data), and subjective data (e.g. citizens' feedback), in such a way as to first fire those questions that lead to maximum information gain. Our framework addresses the shortcomings of static subjective studies, whose scope is generally limited by the inability to reach statistical significance at scale.

In our pilot study, we aim to identify how urban green areas affect well-being, particularly the urban features that have the most impact. That requires collecting subjective responses from a sufficiently vast population, considering the extra dimensions of space and context, which would be impossible to achieve with conventional methods. With our AI data collection method, we can pursue statistical significance much faster, achieving (in our city-wide pilot study) a 41% acceleration factor and a 75% reduction in intrusiveness. Our study opens new avenues in urban science, with potential applications in urban planning, citizen's well-being projects, and sociology, to mention but a few cases.

The rest of the paper is organized as follows. Section II captures the related work, for the different research topics involved. Section III introduces the proposed framework, including the key modules and explaining how we have prototyped them. Section V puts the general framework in context of a specific case study, to better illustrate the value of our approach in a practical setting. Section IV gives an account of the benefits that can be achieved, providing a comparative evaluation of our approached in comparison to a static data collection method. Conclusions and future work indications are finally drawn in Section VI.

II. RELATED WORK

Thanks to the significant technological advances in data sensing and analysis, it is now possible to carry out large-scale, multidisciplinary studies aimed at the interaction between humans and cyber-physical systems [12]. This can substantially aid local authorities in finding new ways to improve the utilization of the public infrastructure and the human well-being [4]. Ultimately, we need to explore integrated approaches to managing data-intensive systems, in planning and decision making, to offer better services and quality of life [5]. This is typically the goal of social science studies that are, however, broadly based on static questionnaires and typically suffer from poor statistical significance [8]. To counter these limitations, research has been increasingly focusing on automated data collection (e.g. through smart phones) and intelligent methods (e.g. using chatbots).

A. CHATBOTS

To reduce the negative effects that digital data collection platforms have on subjective tests, the interaction with users should be as simple and fast as possible. That is why chatbots (and particularly intelligent chatbots) have increasingly been used. In [13], the authors use a smart chatbot to answer students' frequently asked questions. This technology is also used in the medical field for different tasks. Authors in [14] introduce a chatbot that helps elderly people in the process of recalling past memories. The chatbot in [15] connects to diagnostics tools to formulate preliminary questions to ponder whether hospital admissions are required.

In social sciences, an app named Mappiness is used as an intervention tool to improve happiness as an element of well-being [16]. The participants are invited to report their well-being at random times of the day, while their position is constantly monitored. Urban Mind [17], is another app designed to examine in real time how exposure to green spaces affects mental well-being. Starting from the two previous apps,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2943845, IEEE Access

Shmapped [18] was developed, with a double objective. On the one hand, it wanted to be an intervention tool to improve the well-being, by encouraging people to interact more with the nature. On the other hand, it was a data collection tool useful for research. It is clear that the interaction of chatbots is still not fully effective. Currently, chatbots do not have the ability to correctly handle conversations based on the social context. Authors in [19] propose a chatbot model that can choose suitable dialogues, according to what the sociological literature refers to as "social practice".

There is a fundamental difference between the aforementioned literature and our approach. Typically, data analysis is a post-collection mechanisms, with all data being analyzed as a bulk process. By contrast, we carry out online data analysis, that is while the data is being collected. Thus, while the state-of-the-art is aiming at collecting as much data as possible (for big-data processing), we analyze data at run-time with the aim to guide the following steps of data collection. By fusing data in real-time, we aim at gaining information at each step. This has multiple benefits: 1) reducing the circulation of redundant (unnecessary) data; 2) accelerating the process of gaining statistical significance of data; and 3) reducing the intrusiveness of technology during the subjective data collection.

B. ANALYTIC FRAMEWORKS

The Internet of Things is a prominent framework for the collection of heterogeneous sensor data, which offers a unique opportunity to develop smarter and more efficient cities. City data may feed to planning, management and decision-support systems, revolutionizing the way cities operate [20].

There are currently several frameworks for analyzing smart city data. An example is "CityPulse" [21], which can perform semantic discovery, data analytics, near-real-time interpretation of large-scale IoT data, and social media data streams. Another interesting study has been done in [22], where the framework can perform real-time processing on data collected from different applications, to help in real-time decision-making.

State-of-the-art frameworks typically focus on data analysis, but there is little attention to the issue of how to collect data at scale. The authors in [23] introduce an efficient IoT framework for smart cities, which offers mechanisms to mitigate a variety of smart city challenges, using co-operative crowd sensing coupled with a data-centric approach. Another interesting framework is presented in [24], which analyzes the level of waste in the city waste bins (equipped with sensors). The systems maintains a prediction model, which determines the optimal route for the waste collection.

Existing methods tend to be strongly dependent on technology-specific solutions to improve automation and process efficiency. Yet, they use objective data, mostly from homogeneous sources, but fail to capture the citizen's point of view and quality of experience (subjective data). This is in fact the aim of our work, where we aim to extract as much subjective information as possible, putting it in connection with objective IoT data.

C. COMPLEXITY OF DATA COLLECTION

Most of the definitions relating to big data in urban studies are limited to the attribute "volume". A simplistic, yet widely adopted, definition of "big data" refers to any amount of data that would not fit into an Excel spreadsheet or could not be archived in a single machine. For example, the study in [25] analyzed half-million waste routes to identify inefficiencies in the collection. In [26], the authors analyzed the text streams included in 8 million tweets in the San Francisco metropolitan area.

In our study, although we are not aiming to reach such levels of data volumes, there are other difficulties typical of big data sets. We have a significant variety of data, ranging from objective to objective data, and including both structured and unstructured data. We are also facing data variability, since the interpretation of similar data values is sensitive to the context and time in which it is collected. We are also dealing with data uncertainty and bias, particularly, since we deal with subjective data [27].

D. OBJECTIVE AND SUBJECTIVE DATA

Most technological developments have been focusing on automating and accelerating the collection of objective data, such as those coming from sensors, smart phones and other IoT devices. Collecting subjective data in a reliable and statistical significant fashion, poses more serious hurdles. An example of objective data collection is presented in [28], where the study is based on users' activity detection through wearable accelerometers and, in turn, adopts gamification to encourage physical activity. Another interesting pilot study is presented in [29], an urban mobility project based on real-time traces of both traffic conditions and pedestrians. The data is objective in nature, since it is collected through GPS, smart phones, buses and taxis.

The importance of making computational inference on objective data is highlighted in the literature. Particularly, the interaction among intelligent objects and humans is crucial in the study of Smart Cities [30]. Yet, objective data is still affected by unreliability, inconsistency and uncertainty, for instance, in connection to sensor accuracy and missing data due to faults or communication issues [31].

On the other hand, subjective data collection (as in our study) has been targeted somewhat less frequently. In this case, the problem is that i) collecting subjective responses is less prone to automation; ii) it is difficult to collect statistically sufficient data; iii) data is inherently unreliable and suffers from human bias. As matter of fact, social networks have made it possible to collect subjective data, such as tweets about local events [32]. Yet, social networks data is not immune from errors, bias and uncertainty, especially when data is subject to interpretations demanded to inferential engines. In the context of smart city, the process of collecting subjective data can make the analysis richer, more diversified

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2943845, IEEE Access

and complementary [33]. And this is the specific target of our investigation.

III. PROPOSED FRAMEWORK

A. OVERVIEW

Our framework realizes an information gathering and analysis tool, with emphasis on real-time collection of objective and subjective data in urban environments. The proposed tool relies on a typical software architecture, based on client-server communications (Fig.1). The prototype has been developed in Java for the server side. The client-side is a JavaScript, React-Native application. We have not reused any specific libraries, since all necessary routines were custom-made, as we further detail in the next sections. The client devices are smart phone apps, equipped with intelligent chatbots whose key goal is to handle a seamless interaction with the citizens. The subjective questions fired by the app are first pondered by the chatbox to minimize annoyance and maximize the information gain achieved with each question.

The system is designed to be generic and adaptive to the context through simple customizations. The chatbots are provided with a list of questions (collectively, the subjective questionnaire under investigation), along with the geographical boundaries of the study, or "entities", which may include also specific points of interests. Examples in our study are green areas, buildings, roads, parks and so on. These are the objective of the study, that is the context in which we wish to better understand human interactions and their effects on well-being.

All the information collected by the chatbots is collected into the server, which has sufficient resources to store and process data, fusing subjective data with objective data, and contextualizing the statistical significance of each data point. Through data fusion and inference, a feedback is provided back to the chatbots, in a way that further trigger points and questions can be raised. In this way, the questions that are fired by each chatbot are those that are associated with maximal information gain. Fig. 1 gives a snapshot of the chatbot interaction diagram. Fig. 2 shows an example of the textual interaction between chatbot and user.

B. CLIENT SIDE

Although each chatbot is fed with exactly the same set of subjective questionnaires, individual questions are fired at different times, picking first the questions that have minimum statistical significance. This, in turn, will depend upon the context of the question, considering both the internal context (within the individual chatbot) and the external one (an aggregate of all contexts experienced by the chatbots across the system). Specifically, the internal context accounts for things such as the level of intrusiveness reached at a given point (how many questions the same user has been asked before); the position of the user (if they are on an area that is missing subjective responses or not); and other objective data from the sensors (e.g. the level of motion, type of activity, etc). On the other hand, the external context takes into account the statistical significance of each question, aggregated across all users.

The chatbot query system emulates a normal conversation that dynamically chooses the questions to ask (the one with minimal statistical significance), taking into account the overall progress of the analysis. This methodology aims to optimize the information gain compared to what would be obtained with conventional statistical methods. Moreover, a more accurate choice of questions allows the chatbot to ask fewer questions, leading to a reduced intrusiveness of the app. At the same time, we have the added benefit of collecting subjective responses, which goes well beyond what may be achieved by making simple inference on sensor data. About issues of privacy and ethical collection of data, which always remains a critical point with subjective studies [34], we should note that all data is anonymized and aggregated, using it only for statistical purposes [35].

A key component of the chatbot is in charge of handling the questions. It determines which question to ask, in which moment to fire it, and which user will see it. These variables depend on both the internal context (within the chatbot) and the external one (on the server side). Three different situations may occur:

- The server communicates the questions that need to be answered;
- The server communicates a new question, generated based on the data already analyzed;
- The server provides an external context that does not affect the question ordering.

C. SERVER SIDE

Our framework follows the client-server architecture and is ultimately compatible with cloud-based service provisioning. Next, we describe the server-side overall structure and key modules.

The basic server functions are:

- Storing the users' data (both subjective answers and objective sensed data);
- Analyzing the data received from all clients, to extract valuable insights and make inference;
- Producing the overall (aggregated) external context, which is pushed to the individual client-side chatbots.

The whole system is organized in modules, with individual components being independent from each other and having internal functions that can be extended independently. Next, we introduce the modules that handle the web server, data storage, data analysis, and external context management.

1) Web server

This module is a Web Server Application, providing various services that can be reached through the http/https protocol via POST and GET calls. It collects users' information by means of the client-side chatbots, including both subjective (observations) and objective data. The latter includes sensory data (e.g. user activity) and geo-location information col-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2943845, IEEE Access

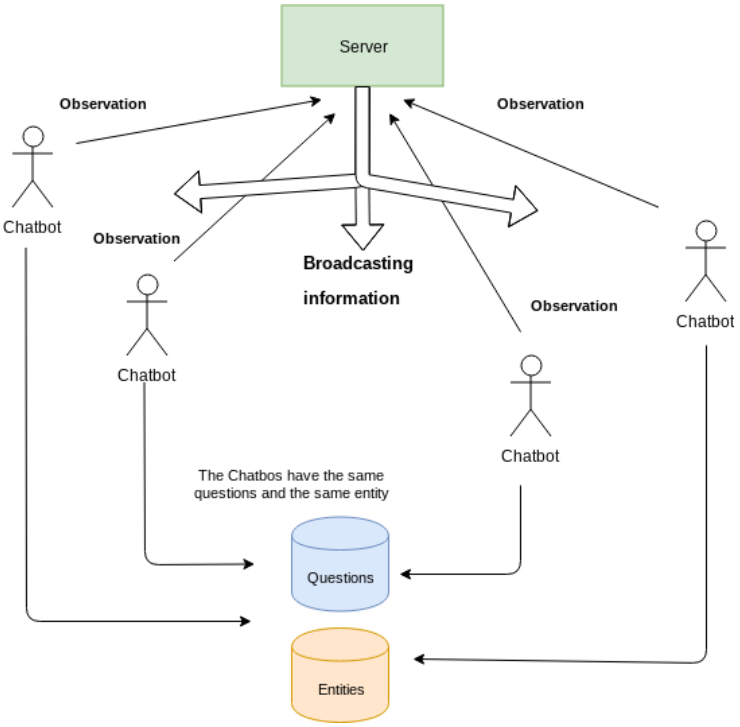

FIGURE 1. Interaction diagram between clients, chatbots, and server.

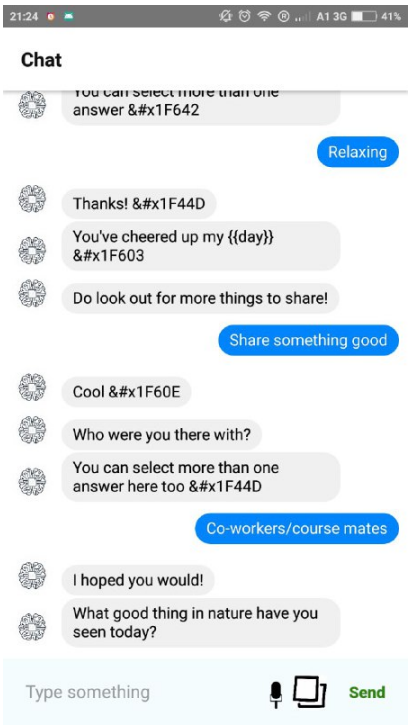

FIGURE 2. Chatbot Interface.

lected in the background. The server will also push updates of the external content to the chatbots.

2) Data storage

The database model used is NoSQL, which provides a flexible data model without fixed schemes that can support large volumes and the broad variety of data generated by modern applications. There are several implementations associated with the NoSQL concept. The one we use is the document-oriented model, in which each record is stored as an entity with its independent properties. The framework uses two collections: the "positions" in which all the geo-location information collected by the users are stored; and a separate "user" collection in which the user's information is stored together with all other associated data.

3) Data analysis

As previously explained, the whole framework (developed from scratch), is built on a typical server-client architecture, in which the client-side sends the necessary data to the server through a JSON file, including only the data that is strictly necessary to make data fusion and update the contextual information. The server side includes also a data pre-processing, integrity and verification step, in order to counter errors that may be due to transmission or data-entry errors. Classic checks are performed on outliers, missing values and inconsistent data. Data pre-processing algorithms remove those entries that have missing or incoherent data.

After this pre-processing phase, the data analysis module proceeds with extracting relevant data from the acquired client-side information. This is a customizable, application-dependent feature that requires the modification or introduction of new plug-in modules.

With regards to image and audio analysis, we are using a mix of third-party APIs and custom-made components.

4) External context management

This component operates in the background, on a separate thread, to create and process the external context. The various messages received from the chatbots are first analyzed through the data analysis modules. The resulting data are filtered and fused to produce the external context, as shown in the flowchart of Fig.3. We recall that the external context is used by individual chatbots in conjunction with the internal context, to determine the questioning order, as explained in section III-B.

The filters shown in Fig.3 are key to determining the level of statistical significance of each question (of the subjective study), and to compute the information gain attained at any moment/location by firing specific questions to specific users. In this way, we have a real-time status of the overall information level of the system, and can turn the system towards the statistically weak data. Thus, each filter contributes to the external context creation, which is saved in a JSON file that is then sent to the clients/chatbots. These can thus prioritize on the questions that area statistically weaker first.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2943845, IEEE Access

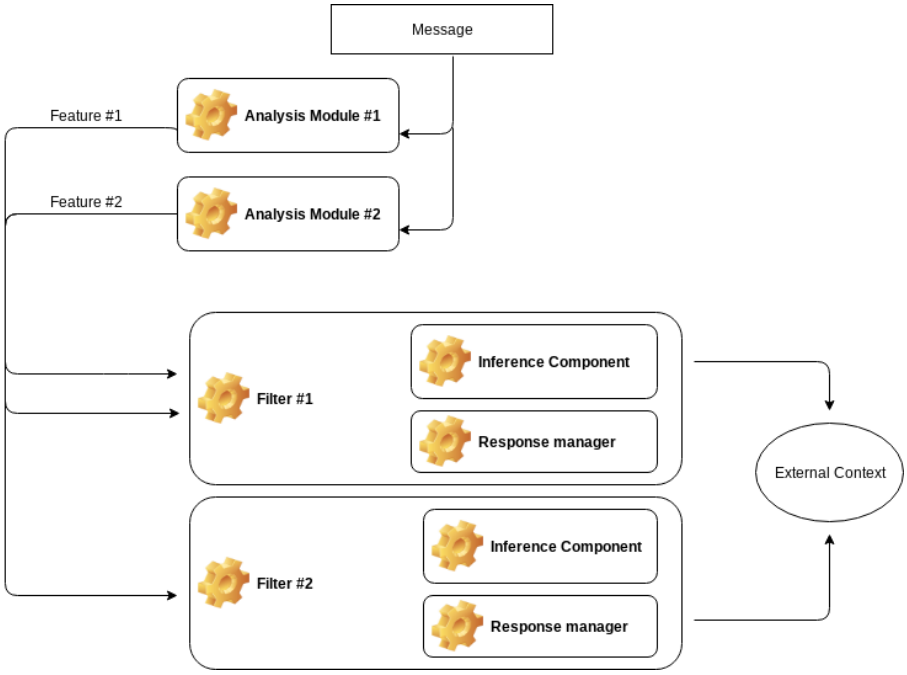

FIGURE 3. Analysis flow.

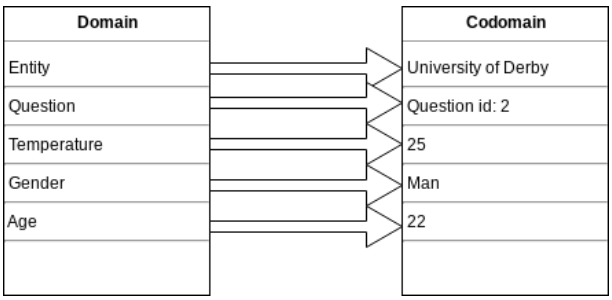

FIGURE 4. Mapping functions.

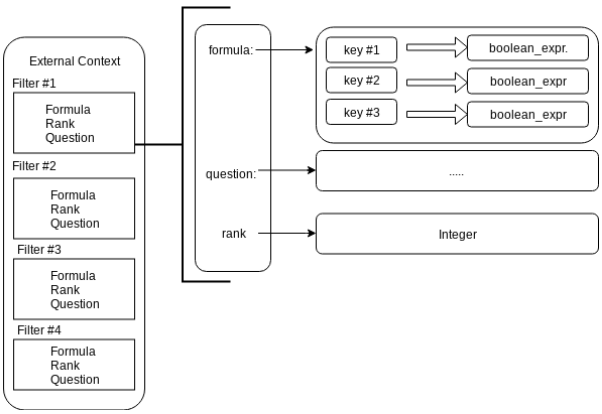

FIGURE 5. Context analysis and representation.

IV. ANALYSIS OF THE INTERNAL AND EXTERNAL CONTEXTS

We recall from Section III that the client-side receives the "external context" from the server-side, which allows chatbot to modify its behaviour, particularly in relation to which questions to submit next, and, then, the submission rules are updated accordingly. Also, both clients and server must agree on the domain/context representation and on the users and environmental features to use. The smartphone App will have to catch all those features, as exemplified in Fig.4, where the domain contains the required information.

The co-domain elements are periodically updated by the chatbots. Some elements may be known thanks to earlier questionnaires, such as age and gender. Some others may be collected automatically through sensors, e.g. position and temperature. The "Question ID" field identifies the next question to be sent to the user via the chatbot.

The external context, formed as shown in Fig. 5, is pushed to the client for continuous analysis. It is represented as a JSON file, containing several filters that indicate the conditions by which specific questions are fired to the user. Each filter has three fields: the "Formula" defines the conditions; the "Rank" defines the filter's priority with respect to the other filters; and the "Question" indicates the question or set of questions to be submitted to the user. This process is better specified as pseudocode in Algorithm 1.

The chatbots are also managing a parameter that determines the conversation naturalness degree. Once a filter is verified, the choice of the question is influenced by two values: the distance from where the question has previously been fired, and a multiplicative factor that determines the weight of the external context with respect to the naturalness

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2943845, IEEE Access

Algorithm 1

```
for var key in formula do
    booleanExpression=set(formula[key],key,mapping[k])
    booleanResult=SafeEvaluation(booleanExpression)
    if booleanResult==false then
        return false
    else
        return true
    end if
end for
```

of the conversation. These two values are used together to define the probability for submitting a question to the user. This mechanism is useful in situations where we have only one question to ask to complete the analysis because, at the same time, it is important to maintain a natural conversation between user and chatbot. Nevertheless, the chief aim of our study was to give priority to the information gaining process. Thus, although it is possible to play with the naturalness degree parameter to assess this aspect, we have kept it constant at this stage.

V. EVALUATION THROUGH A CASE STUDY

In this section we evaluate the proposed methodology and framework, applying it to a real-world case study. Our aim is to show that significant benefits, in terms of statistical significance and speed, can be achieved in urban data analysis.

A. EVALUATION METHOD

Our evaluation strives for generality. We take an existing dataset, particularly one that includes a mix of objective and subjective user data, collected over a broad geographical area. We treat the data as a time-series with the extra dimension of geo-location. We parse the dataset through our emulation environment, where we have the capability of replaying the very same events (i.e., the data collected), changing the events order at will. This is a key feature that allows studying the effects that a reordering of events has on information gain and, in turn, to verify the efficacy of our AI-based data collection method. We can change the order and target in which subjective questions are fired, deciding the specific moment in which a question is best asked, picking specific users, specific locations, and so forth.

The emulator is a software package written in Java, which relies on a database to retrieve the dataset to be replayed. Each message is taken in chronological order and analyzed through the same procedures explained in Section IV. The mapping function is created by setting the position and the question associated with each message. Then an analytical procedure is started to determine whether the question under scrutiny satisfies both the internal and external contexts. The emulator works in tandem with the context management processes (that run in the background on the server-side system), which allows evaluating the performance of the intelligent chatbots. The key performance indicators are the statistical significance of the subjective study and the time needed to complete the study. Thus, our aim is to show that statistical significance may be achieved more rapidly by means of intelligent chatbots.

TABLE 1. Gender distribution of the sample dataset.

Gender	Number of users	Percentage
Female	894	64.64%
Male	489	35.36%

B. THE IWUN DATASET

We have used the dataset generated by the IWUN project (Improving Well-being through Urban Nature - www.iwun.uk) [36]. The project used data science methods to understand the effects that urban nature has on the citizen and, in turn, identify the urban features that correlate directly with human well-being. Vast amount of data, in excess of one terabyte, has been collected in a series of pilot studies, involving a total of 1,870 participants. The citizens were tracked using the specially-made smartphone App Shmapped [18], as they entered any of the 760 geo-fenced locations (identifying the green spaces in the city of Sheffield in the UK). Also, questions were posed to the users through the App. They could provide feedback in terms of text or photos, which we parsed through AI automatic detection and recognition models to extract features. We could thus determine how much time people spent in green areas, the type of activity and find out the top interaction areas.

However, the initial studies based on this dataset have been based on a post-collection (offline) analysis of a wealth of objective and subjective data [9], [11], which proved difficult in terms of achieving statistical significance over broad geographical areas. Recently, we have reported on the benefit of urban analytics, for the purpose of understanding the interaction between citizen and city [10], which has motivated the new approached proposed herein (based on real-time, intelligent data collection/analysis) to pursue more effective urban analysis studies.

Starting from the complete IWUN dataset, we have created a curated version involving the most significant among a total of 5,626 observations. The original dataset has been cleaned through filtering and selections, using publicly available Python and Pandas libraries. The gender and age distribution of the new dataset are shown in Table 1 and Fig.6, respectively.

C. SETTING THRESHOLD GOALS

To appreciate the benefits linked to our real-time (online) approach using intelligent chatbots, we carried out a comparative evaluation, benchmarking against our earlier method that was using batch (offline) analysis and static chatbots [10].

In essence, the intelligent chatbots used the internal and external contexts to determine which questions to fire first and in which location, based on the information gain attain-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2943845, IEEE Access

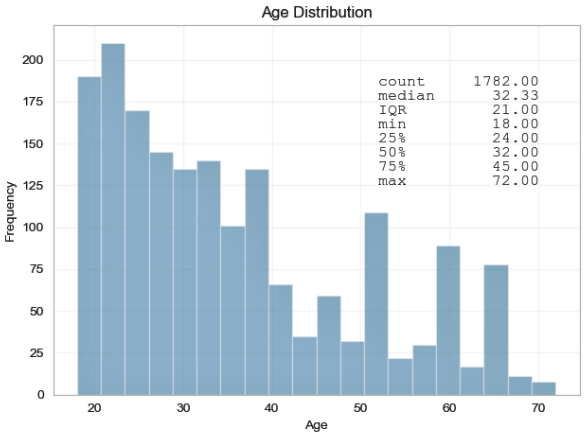

FIGURE 6. Age distribution of the sample dataset.

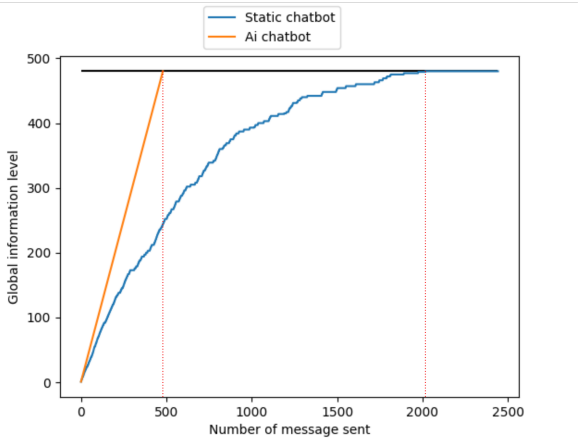

FIGURE 7. Comparative results between static and intelligent chatbots, when setting threshold goals.

able. Also the target users were chosen to minimize intrusion (e.g. users that had provided the least amount of feedback were asked first).

In a first type of experiments, we set as a goal the total number of responses we wanted to collect, for each of the questions included in the subjective questionnaire pot, and for each region under scrutiny (the urban green areas). We also set a constraint on the maximum number of responses that each user was asked to provide, to minimize intrusion.

For the sake of statistical significance, we have restricted our comparative evaluation to the top-20 most visited areas, chosen out of the whole dataset that originally included 760 green areas in the city. Our choice was driven by the dataset at hand, which had been collected prior to our proposed (intelligent collection) method, and was found not be be statistically significant over the whole city (an insufficient number of sample answers was available, despite the scale of the pilot study). For the same reason, and for the sake of simple visualization, we have also restricted the subjective test matrix to three different questions, setting a target number of responses to 8 per question and per region.

Fig. 7, shows the significant acceleration in information gain attainable with intelligent data collection/processing. The goal is reached within the first 500 interactions between the system and the user, compared to the 2,000 messages required with static chatbots. This is because of the fact that our approach can guide the question firing process based on global information, prioritizing on the least asked questions/areas first, which justifies the linear information gain graph.

Fig. 8 provides a different view of the same process, showing how the information gain evolves over time (days). Again, a significant acceleration factor is achieved thanks to the intelligent re-ordering of events. This has been computed as shown in (1), whereby α is the number of days that were required to reach a specific level of global information in the original pilot study. On the other hand, β is the number of days incurred to reach the same objective through our AI chatbot method.

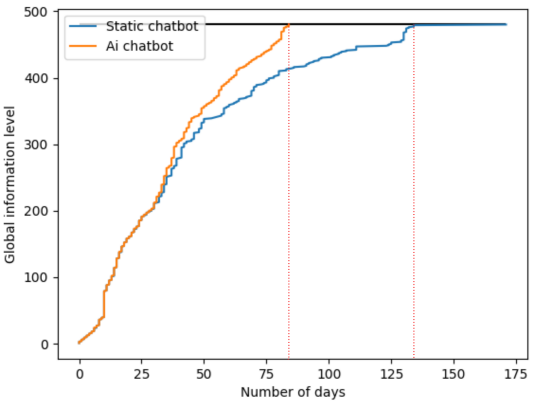

FIGURE 8. Information gain evolution over time (days), when setting threshold goals.

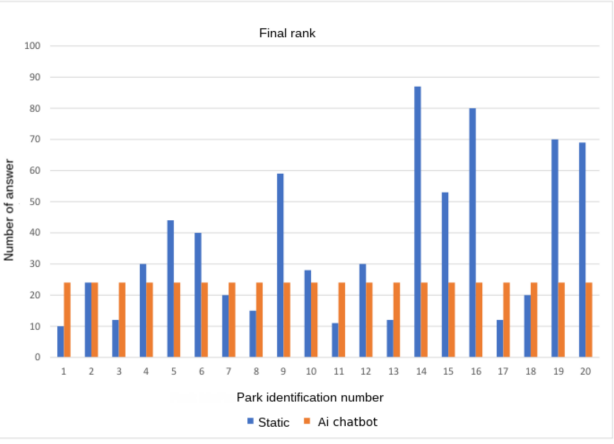

FIGURE 9. Distribution of user's feedback, when setting threshold goals.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2943845, IEEE Access

$$acceleration\ factor = \frac{\alpha - \beta}{\alpha} \quad (1)$$

The spread of user's feedback achieved in the original pilot study can be compared to our re-ordered set in Fig. 9. For simplicity we have empirically set the threshold to 25 answers (for each of the questions in the subjective questionnaire). The striking improvement (from scattered to uniform distribution) is a direct consequence of re-ordering the questions over time and location. It should be noticed that the actual threshold is meant to be study-dependent and would normally be set by the researchers of specific cases. However, our method will generally lead to significant gains in information at each step (i.e., higher information gain per question answered).

D. SETTING STATISTICAL GOALS

Having seen the benefits of intelligent data collection, in terms of reducing the user's interaction and duration of the experiment, we then moved into evaluating how far we could accelerate the overall statistical process. The goal now was set to reaching statistical significance for each question (of the subjective test matrix), for each of the regions under scrutiny, respectively. We restricted the experiment to the top-10 most visited regions, with a set of 5 target questions.

The results included in Fig. 10 show two step-wise functions, relating to the two methods under comparison. In this case, the information gain steps up as soon as any of the questions has received a statistically significant number of user's replies. To determine statistical significance we look at the confidence level. This is closely related to the P value, which indicates when a specific response can be considered statistically significant with respect to the others. We have set the confidence level parameter the default value of 95%, corresponding to a value of 5% for the p-value parameter. Other typical values of 3% and 1% would influence the time required to reach a statistically satisfactory outcome of the questionnaire. Setting a lower threshold will also have a negative impact on intrusiveness, since more user's feedback would be required.

This intelligent data-gathering method leads to a significant acceleration factor, since both user intrusiveness and overall execution time are reduced. The intrusiveness reduction has been computed as shown in (2), whereby γ is the number of messages required to achieve statistical significance in the original pilot, whereas θ is the number of messages required by the AI chatbot solution.

$$intrusiveness\ reduction = \frac{\gamma - \theta}{\gamma} \quad (2)$$

VI. CONCLUSIONS

Making insights into urban data is a daunting task, both in terms of collecting data and analyzing it. We set off with the even more complex goal of carrying out a mix subjective/objective study. Typically, subjective studies aim

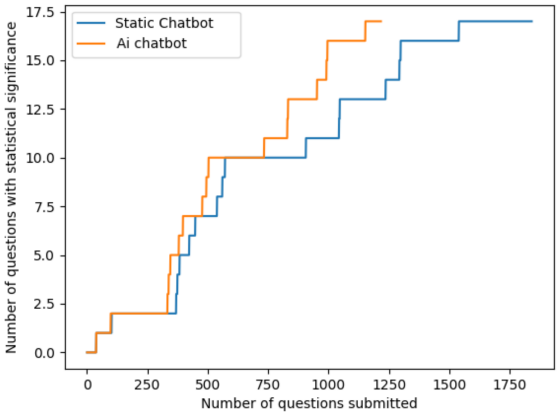

FIGURE 10. Comparative results between static and intelligent chatbots, when setting statistical goals.

to collect relatively few samples directly from people. Conventional questionnaire-based studies are improved with digital systems, e.g. using smartphone Apps. Nevertheless, if the subjective information we wish to collect concerns the citizen, there is the additional dimension of space. In urban subjective studies, the users' feedback needs to be contextualized in time and space, since in many cases an answer depends critically on a specific moment and location.

This is a general problem in urban science, where optimizing city processes requires the collection and analysis of subjective data (e.g., human behaviour, human perceptions, and human quality of experience), in conjunction with objective data (e.g. smart city data, Internet of Things data, and citizens' sensory data). Urbane science needs to go well beyond the analysis of objective data, since most value lays with human data (and their correlation with city data). Yet, collecting subjective data at scale poses significant challenges in terms of automation and statistical significance, which is a core element in this paper.

We take this challenge, adopting a use case to illustrate the scale of the problem. While the case is specific in trying to capture the interactions between citizens and green urban areas, our methodology is generic. Using the IWUN project dataset, we show how difficult it would be to collect a statistically significant data sample in a vast geographical area. Despite being a terabyte-large dataset, the IWUN data provides insufficient information to draw a complete picture, even for the relatively small city of Sheffield (UK), where 760 urban green areas have been scrutinized.

We argue that striving for statistical significance in urban science requires moving away from conventional methods, which typically separate the data collection phase from the data analysis one. By contrast, we perform data collection and analysis alongside, using intelligent processes in real-time (during data collection) to guide the subsequent steps of data collection. The analysis of users' feedback in real-

time (through AI-based feature extraction and text analysis) and the combination of feedback with context (location and information level of each question of the subjective test matrix), lead to a significant acceleration to the overall process. In the case under scrutiny, we achieved a 41% acceleration in reaching statistical significance and a 75% reduction in intrusiveness. Yet, we expect comparable improvements to translate to other analogous cases involving both citizens and cities.

Our work sets the scenes for integrating intelligent data collection and analysis processes in urban analytics, which is particularly useful in urban subjective studies. Establishing when a pilot study has reached statistical significance is essential to drawing reliable conclusions. With our method we are not yet able to anticipate the necessary duration of a complex subjective study - this comes out during the pilot. Next, we wish to work on this and other pilot design factors to help planning and budgeting large-scale city pilots.

We have demonstrated our method in the context of a citizen-to-city interaction project. Next we aim to carry out new pilots to identify other features that are critical to the citizen well-being. Even more ambitious will be to explore the mutual interactions between citizens and 'smart' cities, whereby more uncorrelated data need to come together to benefit the citizens.

ACKNOWLEDGMENTS

The dataset used for the case study comes from the IWUN project, supported by the Natural Environment Research Council, ESRC, BBSRC, AHRC & Defra [NERC grant reference number NE/N013565/1]. The emulation framework development and data analysis were performed in collaboration between the Data Science Research Centre (*www.derby.ac.uk/data-science*) and the Joint Intellisensing Lab (*www.thejil.com*)

REFERENCES

[1] T. Panagopoulos, J. A. G. Duque, M. B. Dan, Urban planning with respect to environmental quality and human well-being, Environmental Pollution 208 (2016) 137 – 144, special Issue: Urban Health and Wellbeing (2016). doi:https://doi.org/10.1016/j.envpol.2015.07.038.

[2] Z. Khan, A. Anjum, K. Soomro, M. A. Tahir, Towards cloud based big data analytics for smart future cities, Journal of Cloud Computing 4 (1) (2015) 2 (2015).

[3] C. Yin, Z. Xiong, H. Chen, J. Wang, D. Cooper, B. T. David, A literature survey on smart cities, Science China Information Sciences 58 (10) (Oct. 2015). doi:10.1007/s11432-015-5397-4.

[4] A. Zanella, N. Bui, A. Castellani, L. Vangelista, M. Zorzi, Internet of things for smart cities, IEEE Internet of Things Journal 1 (1) (2014) 22–32 (Feb 2014). doi:10.1109/JIOT.2014.2306328.

[5] J. Jin, J. Gubbi, S. Marusic, M. Palaniswami, An information framework for creating a smart city through internet of things, Internet of Things Journal, IEEE 1 (2014) 112–121 (04 2014). doi:10.1109/JIOT.2013.2296516.

[6] T. Bakıcı, E. Almirall, J. Wareham, A smart city initiative: the case of barcelona, Journal of the Knowledge Economy 4 (2) (2013) 135–148 (2013).

[7] J. H. Lee, M. G. Hancock, M.-C. Hu, Towards an effective framework for building smart cities: Lessons from seoul and san francisco, Technological Forecasting and Social Change 89 (2014) 80–99 (2014).

[8] J. Maas, R. A. Verheij, P. P. Groenewegen, S. De Vries, P. Spreeuwenberg, Green space, urbanity, and health: how strong is the relation?, Journal of Epidemiology & Community Health 60 (7) (2006) 587–592 (2006).

[9] E. Ferrara, A. Liotta, L. Erhan, M. Ndubuaku, D. D. Giusto, M. Richardson, D. Sheffield, K. McEwan, A pilot study mapping citizens' interaction with urban nature, in: 16th Intl Conf on Pervasive Intelligence and Computing (PiCom), IEEE, 2018, pp. 836–841 (2018).

[10] L. Erhan, M. Ndubuaku, E. Ferrara, M. Richardson, D. Sheffield, F. J. Ferguson, P. Brindley, A. Liotta, Analyzing objective and subjective data in social sciences: Implications for smart cities, IEEE Access 7 (2019) 19890–19906 (2019).

[11] E. Ferrara, A. Liotta, M. Ndubuaku, L. Erhan, D. D. Giusto, M. Richardson, D. Sheffield, K. McEwan, A demographic analysis of urban nature utilization, in: 2018 10th Computer Science and Electronic Engineering (CEEC), 2018, pp. 136–141 (Sep. 2018). doi:10.1109/CEEC.2018.8674206.

[12] N. Zhong, J. H. Ma, R. H. Huang, J. M. Liu, Y. Y. Yao, Y. X. Zhang, J. H. Chen, Research challenges and perspectives on wisdom web of things (w2t), The Journal of Supercomputing 64 (3) (2013) 862–882 (2013).

[13] S. Ghose, J. J. Barua, Toward the implementation of a topic specific dialogue based natural language chatbot as an undergraduate advisor, in: 2013 International Conference on Informatics, Electronics and Vision (ICIEV), 2013, pp. 1–5 (May 2013). doi:10.1109/ICIEV.2013.6572650.

[14] N. Svetlana, F. Daniel, B. Marcos, F. Casati, K. Georgy, Crowdsourcing for reminiscence chatbot design, in: Conference on Human Computation and Crowdsourcing (HCOMP 2018), 2018, pp. 1–5 (2018).

[15] L. Ni, C. Lu, N. Liu, J. Liu, Mandy: Towards a smart primary care chatbot application, in: International Symposium on Knowledge and Systems Sciences, Springer, 2017, pp. 38–52 (2017).

[16] G. Mackerron, S. Mourato, Happiness is greater in natural environments, Global Environmental Change 23 (2013) 992–1000 (10 2013). doi:10.1016/j.gloenvcha.2013.03.010.

[17] I. Bakolis, R. Hammoud, M. Smythe, J. Gibbons, N. Davidson, S. Tognin, A. Mechelli, S71. urban mind: Using smartphone technologies to investigate the impact of nature on mental wellbeing in real time, Biological Psychiatry 83 (2018) S374 (05 2018). doi:10.1016/j.biopsych.2018.02.962.

[18] The smart app "Shmapped" 2019 [Online] Available: http://iwun.uk/shmapped/ [Accessed: 31 July 2019].

[19] A. Augello, M. Gentile, L. Weideveld, F. Dignum, A model of a social chatbot, in: G. D. Pietro, L. Gallo, R. J. Howlett, L. C. Jain (Eds.), Intelligent Interactive Multimedia Systems and Services 2016, Springer International Publishing, Cham, 2016, pp. 637–647 (2016).

[20] O. Bates, A. Friday, Beyond data in the smart city: learning from a case study of re-purposing existing campus iot, IEEE Pervasive Computing 16 (2) (2017) 54–60 (6 2017). doi:10.1109/MPRV.2017.30.

[21] D. Puiu, et al., Citypulse: Large scale data analytics framework for smart cities, IEEE Access 4 (2016) 1086–1108 (2016). doi:10.1109/ACCESS.2016.2541999.

[22] K. K. Mohbey, An efficient framework for smart city using big data technologies and internet of things, in: Progress in Advanced Computing and Intelligent Engineering, Springer, 2019, pp. 319–328 (2019).

[23] S. K. Datta, R. P. Ferreira da Costa, C. Bonnet, J. Harri, onem2m architecture based iot framework for mobile crowd sensing in smart cities, in: 2016 European Conference on Networks and Communications (EuCNC), 2016, pp. 168–173 (June 2016).

[24] J. M. Gutierrez, M. Jensen, M. Henius, T. Riaz, Smart waste collection system based on location intelligence, Procedia Computer Science 61 (2015) 120–127 (2015).

[25] H. Shahrokni, B. Van der Heijde, D. Lazarevic, N. Brandt, Big data GIS analytics towards efficient waste management in Stockholm, in: Proceedings of the 2014 conference ICT for Sustainability, Atlantis Press; Paris, 2014, pp. 140–147 (2014).

[26] P. Anantharam, P. M. Barnaghi, K. Thirunarayan, A. P. Sheth, Extracting city traffic events from social streams, ACM TIST 6 (2015) 43:1–43:27 (2015).

[27] R. Kitchin, Big data, new epistemologies and paradigm shift, Big Data & Society 1 (2014) 1–12 (04 2014). doi:10.1177/2053951714528481.

[28] Y. Fujiki, K. Kazakos, C. Puri, P. Buddharaju, I. Pavlidis, J. Levine, Neat-o-games: blending physical activity and fun in the daily routine, Computers in Entertainment (CIE) 6 (2) (2008) 21:1–21:22 (2008).

[29] F. Calabrese, M. Colonna, P. Lovisolo, D. Parata, C. Ratti, Real-time urban monitoring using cell phones: A case study in rome, Intelligent Transportation Systems, IEEE Transactions on 12 (2011) 141 – 151 (04 2011). doi:10.1109/TITS.2010.2074196.

[30] B. Guo, D. Zhang, Z. Wang, Z. Yu, X. Zhou, Opportunistic IoT: Exploring the Harmonious Interaction Between Human and the Internet of

Things, J. Netw. Comput. Appl. 36 (6) (2013) 1531–1539 (Nov. 2013). doi:10.1016/j.jnca.2012.12.028.

[31] Y. Qin, Q. Z. Sheng, N. J. Falkner, S. Dustdar, H. Wang, A. V. Vasilakos, When things matter: A survey on data-centric internet of things, Journal of Network and Computer Applications 64 (2016) 137 – 153 (2016). doi:https://doi.org/10.1016/j.jnca.2015.12.016.

[32] P. d. Meo, E. Ferrara, F. Abel, L. Aroyo, G.-J. Houben, Analyzing user behavior across social sharing environments, ACM Trans. Intell. Syst. Technol. 5 (1) (2014) 14:1–14:31 (Jan. 2014). doi:10.1145/2535526.

[33] A. Sheth, Citizen sensing, social signals, and enriching human experience, IEEE Internet Computing 13 (4) (2009) 87–92 (July 2009). doi:10.1109/MIC.2009.77.

[34] Y. Li, W. Dai, Z. Ming, M. Qiu, Privacy protection for preventing data over-collection in smart city, IEEE Transactions on Computers 65 (5) (2016) 1339–1350 (May 2016).

[35] L. Van Zoonen, Privacy concerns in smart cities, Government Information Quarterly 33 (3) (2016) 472–480 (2016).

[36] "Improving Wellbeing through Urban Nature (IWUN)" 2019 [Online] Available: http://iwun.uk/ [Accessed: 31 July 2019].