Journal of the Knowledge Economy manuscript No.
(will be inserted by the editor)

CityPulse: A Platform Prototype for Smart City Social Data Mining

Maria Giatsoglou · Despoina Chatzakou · Vasiliki Gkatziaki · Athena Vakali · Leonidas Anthopoulos

Received: date / Accepted: date

Abstract Cities experience radical shifts from conventional areas of fragmented services and interactions, to whole-of-service and end-to-end providers, while their citizens are empowered primarily via social networking applications with geotagging capabilities. This work is motivated by the fact that the exploitation of a (smart) city's social networking and collective awareness can lead to improvements in the citizens' daily life and assist city's crowd-wise policy and decision making. This challenging objective requires appropriate platforms which will not only offer analytics of the city's social networking data threads, but also aggregation and visualization of these data for revealing and highlighting latent information in terms of the city's emerging topics and trends. The proposed CityPulse is a modular platform for offering smart city services based on social data analysis in the context of a city. CityPulse is based on the main principle that a carefully designed backend system supports appropriate data storage, aggregation and analysis methodologies, while the derived results are exposed through web service interfaces to ensure interoperability with various smart city applications that serve the needs of various city stakeholders. Here, we indicatively describe a generic mobile frontend interface that demonstrates the functionalities that can be implemented based on CityPulse results derived by geolocated social data mining. We also demonstrate the results of CityPulse's application on an representative smart city

This research has been co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: "ARCHIMEDES III. Investing in knowledge society through the European Social Fund", as well as by the European Commission through the SmartSantander FP7-ICT project.

M. Giatsoglou · D. Chatzakou · V. Gkatziaki · A. Vakali
Informatics Department, Aristotle University of Thessaloniki, Greece
E-mail: {mgiatsog,deppych,gkatziaki,avakali}@csd.auth.gr

L. Anthopoulos
Department of Business Administration, TEI of Thessaly, Greece
E-mail: lanthopo@teilar.gr

case study which indicate that it can effectively capture and summarize social media user activities within the city, and deliver useful latent information to interested city communities in an comprehensive, flexible manner.

Keywords social data mining · smart city applications · social analytics · data analytics · software architecture

1 Introduction

Today's cities are shifting and transforming into testbeds where solutions driven by Information and Communication Technologies (ICT) impact people interactions and vice versa. The realization of a *smart city* vision, where (mainly ICT-based) urban innovation is applied for enhancing urban life in terms of people, economy, governance, mobility, living and environment (Anthopoulos, 2015), is now closer than ever. This is demonstrated by the growing number of applications and services which utilize sensor measurements in terms of several urban environment conditions (e.g. air pollution, traffic), or human-contributed information about the city through "social" or "crowd" sensing (Anthopoulos and Fitsilis, 2015). Information shared in social media may easily be connected to the physical space, i.e. the location of the user. As a result, although social media are known for "breaking location barriers" at a global level, in parallel, they serve as important, real-time sources of local information. Social media location awareness enables prioritizing topics or problems at different locations and therefore social media content and interactions can be exploited according to the topic, time, and location dimensions.

In the context of a city's daily routine, topics of interest emerge and problems are addressed around specific areas (locations), with citizens expressing their opinions in social media. In parallel, local government's decision making processes can be supported by knowledge derived from citizens' social networking activities. The analysis of social networking activities within a city will result in rich information about the conditions pertaining to different city areas that can be communicated to various stakeholders, and can also be used as a common reference base that will assist the collaboration of citizens and authorities for the common good and open smart city decision making.

Designing and implementing new platforms that utilize geolocated social media data from several sources for revealing city-wide activity profiles and the citizens' opinions / views can have a large impact on assisting city stakeholder communities in their everyday needs. It would be, thus, necessary to propose flexible methodologies that will enable local social media content's collection and analysis with respect to specific criteria (primarily topic, time, and location) in order to design and implement appropriate smart city applications.

This work views the city's locations as emerging virtual spaces which offer dynamic activity summaries, topics shifts, and social media analytics. The proposed ideas cross-cut Future Internet technologies along with societal "pulse" detection, and their focus towards assisting the needs of various city stakeholders. According to the authors' knowledge, up to now, each of the concepts

of "social media and ICT" and "civic empowerment", have been tackled independently in city applications and platforms, with no mature prototypes or tools integrating such processes.

The aim of this paper is to deal with the above observation and answer the following two related research questions:

RQ1 *how can social media be leveraged as a source for deriving crowd-sourced knowledge on a city?*

RQ2 *how can a social data analytics platform for smart cities be designed and implemented?*

In an attempt to answer RQ1, we harvest different types of social media applications and propose methodologies that enable detection of city-related topics and trends derived from social media, and visualizing them on top of the given city's areas. At the same time, we address RQ2 by proposing a platform whose design principles enable the implementation of these methodologies in an approach that safeguards the flexibility and functionality of the platform. Such a design builds on the fact that people who are active as social media users act as "prosumers" since they may contribute, share and interact with city spots and locations (by tagging, commenting, adding posts, etc.), thus exposing city problems is crowdsourced. Moreover, informing municipalities on such citizen interactions in a comprehensive, summarized way will allow them to understand the city "dynamics", and thus effectively respond on problems in relation to public spaces.

Building on the experiences of "architecting" smart city applications from the EADIC project[1] and the SEN2SOC smart city experiment[2] outcomes, this work's contribution is the proposal of CityPulse, a new platform prototype that aims to bridge social media activities with the needs of a smart city in terms of people's open informing and awareness. The proposed platform is flexible and adaptive, since it can be easily applied to different geographic regions, such as cities. More specifically, CityPulse

- *analyzes online social data* for revealing latent information on activities, discussion threads and expressed opinions in the context of a city;
- *exposes discovered information to third-party applications* so that it can be further exploited in smart city services in relation to *smart tourism* (also citizen entertainment), *safety and emergency* and *e-government* (Anthopoulos, 2015);
- *jointly exploits multiple, diverse social media types* to derive insights and latent knowledge about the "pulse of the city" and the citizens' opinions;
- *analyzes user generated content (UGC) at a geographic area-level* for deriving fine-grained information about the city and its relation with citizens.

The remainder of the paper is structured as follows. Section 2 discusses related work in smart city platforms / applications and geolocated social data

[1] The EADIC project: `eadic.teithessaly.gr`

[2] The SEN2SOC experiment: `smartsantander.eu/index.php/sen2soc`

mining approaches. Section 3 presents the basic design principles of our proposed platform and its targeted audiences, while Section 4 describes the platform's architecture and functionalities. Section 5 presents the results of the application of our platform in a selected city case study, and demonstrates the platform's frontend. Finally, Section 6 summarizes and discusses our experiences with CityPulse, and Section 7 concludes the paper.

2 Background

Here, we first outline widely used geolocated social networks capabilities, and focus on related approaches which analyze geotagged data for revealing useful insights about a given geographic region or a city. Next, we present some basic background information about smart cities, with special focus on existing smart city projects and applications. Finally, we highlight current gaps in incorporating insights from social content analysis in smart city applications.

2.1 Social networks as geolocated sources for urban dynamics discovery

UGC in social media is often accompanied by metadata, such as tags, references to external sources, geographic location, etc. In particular, the declaration of users' location by geotagging their shared content can be exploited to obtain insights about their social activities in relation to their surrounding, and harvest them for offering better recommendations about activities available around them. There is a variety of social media applications offering geotagging capabilities including among others applications that emphasize on: *news dissemination* (e.g. Twitter), *location sharing* (e.g. Foursquare and Facebook Places) and *multimedia content sharing* (e.g. Flickr and Instagram).

Social data mining can reveal insights on current city conditions as "reported" by people moving around the city areas. For instance, social media users often refer to topics pertaining to their local community ranging from local activities / gatherings / initiatives to problems in relation to the urban environment, and even critical events (e.g. floods, fires). Automatic detection of such topics, through social data mining, would facilitate the on time provision of information services to the public in terms of interesting activities in their region and, more importantly, about critical, unexpected events.

Critical event detection based on social data analysis was addressed in (Sakaki et al, 2010) through the classification of Twitter posts (i.e. *tweets*) in terms of various features, such as keywords, number of terms and their context, to identify the occurrence of earthquakes and the trajectory of typhoons. Another relevant effort by Kumar et al (2011) resulted in TweetTracker, a tool for Humanitarian and Disaster Relief respondents, which detects crisis events based on specific keywords, by monitoring and analyzing Twitter posts in near real time.

An approach to detect local events, i.e. gatherings of people with common interest in a specific place and time, was proposed by Watanabe et al (2011)

based on the real-time analysis of geotagged and non-geotagged Twitter posts. Georgiev et al (2014) presented an event (i.e. observable anomalous activity with respect to a place) participation prediction model, based on Foursquare checkinss and by jointly considering various social, spatial and temporal factors. A possible application scenario of such an approach could be the provision of personalized event recommendations. Papadopoulos et al (2010) developed a methodology for analyzing photos from multimedia content-sharing social media (e.g. Flickr). According to this, photos are initially clustered based on visual and semantic similarity (the latter based on tag annotations), and then such clusters are classified as either objects or events, and further labeled based on tags frequently associated with them. Chen and Roy (2009) proposed an event detection approach again based on annotated Flickr photos that distinguishes between periodic and aperiodic events. Vakali et al (2012) presents another relevant work that addresses city-wide trending topic detection.

2.2 Smart cities: concept, platforms and applications

Several definitions have been given for smart cities over the years. According to IEEE[3], a smart city *brings together technology, government and society to facilitate the economy, mobility, environment, people, living and governance.* Hollands (2008) has published a critical review about the smart city concept, highlighting the fact that a smart city must be focused onto the needs of its people. The adoption of innovative technologies by a city does not automatically make it smart, as a smart city should rather pursuit the qualitative exploitation of technology for economic growth in balance with people's needs. Hollands also mentioned that smart cities could lead to more democratic communities by encouraging the citizens' participation in the decision making process. Finally, according to Anthopoulos (2015), smart city concerns "*innovation -not necessarily but mainly based on ICT- that enhances urban life in terms of people, economy, environment, mobility, governance and living*".

Several recent research projects (such as SmartSantander[4], Peripheria[5], EPIC[6]), have specifically focused on experimenting with ICT within cities, providing platforms and infrastructures for smart city management, as well as facilitating the development and operation of smart city applications and services. Other smart city projects (such as PEOPLE[7] and Smart-Islands[8]) have specifically focused on the development and provision of such services and applications for facilitating the daily life of citizens, the experiences of tourists and the decision making process. Securing successful smart city management was a core objective of research project EADIC.

[3] IEEE smart cities definition: `smartcities.ieee.org/about`
[4] The EU SmartSantander project: `smartsantander.eu`
[5] The EU Peripheria project: `peripheria.eu`
[6] The EU EPIC project: `epic-cities.eu`
[7] PEOPLE project `people-project.eu`
[8] Smart-Islands project: `smart-islands.eu`

There have been limited efforts on harvesting geotagged social media content for the provision of services to people moving around a city. Komninos et al (2013) proposed HotCity, a service that aggregates checkins and Point of Interest (POI) relevant information and ratings from several social media and other online sources (Foursquare, Facebook, Google Places and Wikipedia) to estimate the average popularity of POIs in different days of week and times of day. The service was made available to tourists / citizens through public displays that serve as an interactive guide for more effective activity planning. More specifically, the HotCity service presents a city map including POIs and their descriptions, with a superimposed heatmap layer that indicates the intensity of user activity in the different areas / POIs depending on the selected day / time. The map also depicts currently "trending" places identified based on real-time checkins.

2.3 Advancing smart city applications and services

One of the smart city objectives, according to the People dimension (Nam and Pardo, 2011), is the engagement of citizens in the city building and planning process and citizen informing about their surroundings, to improve the provided services and their quality of life in general. Thus, the exploitation of the abundance of information provided by people moving around a given city in social media can contribute to their transformation from passive to active city entities. The analysis of such UGC will not only enable the establishment of a continuous stream of collectively sourced information, but could also lead to the uncovering of hidden clues for people's opinions in relation to the effectiveness of policy making in important (everyday) city issues.

The previously discussed smart city applications and projects focus on providing specific services to citizens / city visitors, based mainly on static content (and in most cases, POIs), while they do not take advantage of the wealth of social media content which is attached to the various city-wide geographic locations. Cities, though, can be viewed as living organisms that constantly change in a dynamic fashion based on their people activities, and this is also reflected in social media. Thus, smart city applications should exploit the latent, evolving information that can be mined from geotagged social media content for improving and keeping their offered services up-to-date.

Up to now, several research efforts have focused on the extraction of useful insights about a specific geographic region, e.g. a city, based on geotagged content from social media. Such approaches, though, mainly targeted at addressing a specific, unilateral problem at a research study level, thus did not deal with the requirements and operational specifics of a live, smart city application. Moreover, applications like HotCity, which aimed at leveraging social media content for assisting citizens / city visitors, could be extended by the inclusion of data mining methodologies for extracting more detailed insights about the city and the citizens' perception (e.g. emerging topics and events, content generation, etc.). A more comprehensive approach is deemed necessary,

to fully exploit social media of multiple types (as diverse sources of information) and jointly analyze them via state-of-the-art data mining approaches for extracting useful insights for smart city applications. Such an approach will significantly support the smart city's commitment to position people at its highest priority, while the city dynamics sensing will be made possible.

The proposed CityPulse addresses this need by combining social media content from various sources to address the needs of city-wide active communities involving citizens, city visitors and the city administration. Unlike existing citizen reporting applications (such as, e.g., Improve My City[9] (Tsampoulatidis et al, 2013) and City Sourced[10]) that rely on explicit citizen feedback to identify the conditions pertaining to a city (mainly problems), CityPulse was designed under the concept that there is useful information about the city that can be unobtrusively derived from publicly-available, geolocated social data. Insights from social media analysis are used for providing information in relation to various important axes, such as: (i) city's emerging popular topics and areas, and (ii) popular POIs as they emerge from users' activities. CityPulse is adaptive to any city and can complement typical smart city infrastructures / services (e.g. based on physical sensors) or citizen reporting applications, by offering an additional "human"-driven layer of automatically derived information.

3 Towards a social content-aware smart city platform: principles and targeted audiences

This work aims at identifying the main principles and characteristics needed to design and implement a flexible, adaptive and collective awareness platform. Such a platform can offer city insights to several stakeholders in an automated, real-time manner. To achieve this goal, social media analysis is employed within the city's geographic span. Next, we describe the main design principles and characteristics of the proposed platform and outline the ways in which it can assist the needs of different city audiences.

3.1 Basic Platform Design Principles

Since this work aims at enabling urban dynamics detection, summarization and visualization, designing a platform to cover such needs requires data collection and processing methodologies, as well as appropriate user-oriented interfaces. In the context of smart city architecture standards (International Telecommunications Union, 2015), the proposed platform can be considered as part of an *n-tier* smart city meta-architecture, positioned at the *Services* layer. Moreover, the proposed platform is independent to city form (new or existing city) as well as smart city form (virtual, mobile, broadband, ubiquitous etc.).

[9] Improve My City: `improve-my-city.com`
[10] City Sourced: `citysourced.com`

The platform typically comprises a back-end and a front-end module (the *Backend* and the *Frontend*, respectively). The Backend is responsible for collecting geotagged social media content (e.g. from Twitter, Flickr, Foursquare) and aggregating posts and content (e.g. photos) based on the time and geographic location, under a user privacy-preserving approach. This component also implements data mining methodologies so that popular topics and content clusters can be detected. The Backend follows a flexible, modular architecture, as proposed in (International Telecommunications Union, 2015), while to ensure reusability of the derived social media insights about the city, it exposes such results to third-party services and applications via appropriate Web service APIs (Application Program Interfaces). The Frontend includes Web and mobile application(s), designed for city communities, and offering various services, such as the next ones: place recommendations based on social media information; visualization of the popularity of city areas and POIs; trending topics based on discussions in social media and a dynamic organization of multimedia UGC in clusters. Here, we specifically describe the functionality of a generic mobile application that can match several needs of people moving around the city, however by leveraging the platform's Web service APIs, additional, more focused mobile and Web applications can be developed.

Next, we provide more details for the basic principles based on which the Backend and Frontend of the platform are designed. This list is by no means exhaustive for a social data mining-driven smart city platform, and it largely depends on the geolocated social data analysis tasks that the platform will support. CityPulse demonstrates some well-known and widely-studied social data analysis operations (content clustering, trending topic detection, event detection), as discussed in the Background Section, which were not combined before in a smart city context. Their results, combined with the area- and POI-based popularity indicators that can be extracted based on social activities constitute the set of insights that CityPulse can offer in the context of a city. CityPulse was conceived as part of the SEN2SOC experiment, an approach to combine UGC (including content from social media) with environmental sensor readings in order to capture a richer view of the conditions pertaining to the city of Santander, Spain. SEN2SOC was selected based on its aim and its expected outcome / impact, as a pilot smart city experiment in the context of the SmartSantander project, which involved both local research groups and the municipality of Santander. This consortium overviewed the requirement analysis and design phases of the experiment as well as its progress, and thus CityPulse had constant feedback from people who were part of the envisioned city stakeholders.

3.1.1 The Backend component

Specification of geographic areas. The segmentation of the city into geographic areas (e.g. neighborhoods) constitutes a mandatory element of the platform that supports various functions, such as area / place recommendations and visualization of social data insights to the Frontend. Geographic areas are

represented as polygons specified by the geographic coordinates (latitude and longitude) of their vertices (which is a typical approach for spatial data).

Provision of information on critical situations / events per geographic area. The platform should be able to identify situations worth reporting to the citizens via the real-time analysis of geotagged social data (e.g. traffic) and make them available to the CityPulse users. For the identification of anomalous situations / events, NLP (Natural Language Processing) techniques and appropriate lexicons and event detection algorithms can be employed. Whenever such events occur, alert messages are synthesized and posted to social media (e.g., Twitter and Facebook) and to the Frontend. The posting of alerts to social media is a feature that establishes CityPulse as a vital information source for citizens, regardless of whether they use the proposed Frontend.

Social media data collection. The Backend should collect recent UGC from social media, which are geotagged within a given city's region. The collection of social media content is necessary in order to perform analysis and identify popular areas and topics of interest. To support this feature, focused data collection services have been developed that make use of the APIs of the corresponding social media applications.

Popular topic detection based on social media analysis. The Backend should analyze textual content from social media (e.g. Twitter: users posts, Flickr: photo metadata) in order to identify the most popular topics discussed.

Identification of city areas' and POIs' popularity based on geotagged social media activities. The Backend should analyze UGC extracted from Foursquare in terms of POI visiting information and users tips on POIs to identify the most popular places in different days of week. With respect to the city areas' popularity, CityPulse relies on social media geotagged content (e.g., posts in Twitter, photos in Flickr) to infer the level of activity observed in each geographic area depending on the time of day and day of week.

Organization of multimedia UGC in coherent groups. CityPulse's Backend should follow a clustering approach for organizing the plethora of multimedia content shared within a city / city area in semantically and geographically coherent groups to assist visualization and comprehension. This involves the detection of clusters of related photos (Twitter, Flickr) and posts (Twitter).

3.1.2 The Frontend component

Presentation of city areas' popularity. The Frontend should provide users with a comprehensive heatmap that visualizes the popularity of the different city areas based on social media geotagged content. Users should be able to easily parameterize the heatmap based on temporal constraints (e.g. day-time, night-time), or by selecting only the activities on selected social media.

Suggestion of areas and POIs to users. The Frontend should suggest areas and POIs to users based on people's activities reported in social media. Users are recommended with nearby areas based on the number and / or popularity of the enclosed POIs in the respective areas. Suggestion of areas is accompanied by a summary of the enclosed POIs (e.g., museums, monuments, etc.).

Presentation of social data analysis results. The Frontend should present to users the latest insights derived from the geotagged social content analysis, such as trending topics and photo cluster summaries.

Presentation of alerts on identified events. The Frontend should present to users a list of the current events identified from the geotagged social content analysis.

Authentication mechanisms for user login. The Frontend supports registration of users to a native user database and / or connectivity with their Facebook or Twitter accounts (through which CityPulse's content can be further propagated to popular social media to reach a wider audience).

3.2 CityPulse Users

CityPulse targets at assisting several user communities that can be encountered in an urban context. The needs of such user communities will be addressed either directly through CityPulse's Frontend, or indirectly via third-party applications that may utilize the analysis results generated by CityPulse's Backend. Next, we present three main user categories: (a) citizens, (b) city visitors and (c) city administration, and describe how they could benefit from the CityPulse platform.

Citizens. Citizens get informed about topics that are discussed and concern other fellow citizens and critical events through online or offline news agencies. With the advent of social media, more and more people tend to post information about emerging events and topics of interest via their social media accounts. CityPulse aims to assist citizens by offering real-time critical event detection and popular topic detection through social media analysis. It also provides a dynamically generated overview of the city that informs users about the profile and popularity of the various city areas and the (type of) activities that take place within them.

City visitors. City visitors and tourists usually seek information about a city they visit in tourist guides or in city maps. With the emergence of ICT, online city guides such as TripAdvisor[11] and Wikivoyage[12] have risen, offering an interactive experience by allowing users to share their own opinions about places.

[11] Tripavisor: `tripadvisor.com`
[12] Wikivoyage: `wikivoyage.org`

However, such guides focus specifically on the provision of information about POIs, rely on content provided only via their applications, and lack a real-time mechanism for updating the current events, topics and activity within the city, and presenting them to users. By taking advantage of social media UGC, city visitors can better understand the city's vibe and the activities of people within it. CityPulse aims to offer a better city visiting experience by providing place recommendations based on social media analysis to visitors and informing them on popular topics discussed by citizens in social media.

City administration. It is important for local authorities to understand the needs of citizens and the topics that they are concerned about with respect to the city. Driven by the requirement of putting the citizens' needs at the center of the decision making process, the platform aims to assist the city administration by offering a better understanding about the city, providing information about critical events that occur in the city and emerging topics that are discussed between its inhabitants based on social media analysis. Moreover, the profiling of the different city areas can assist city administration in tasks such as city planning and event organization.

4 CityPulse: A collective awareness platform focusing on city dynamics

Based on the outline of the proposed CityPulse provided in the previous section, this section proceeds with a more in-depth analysis of the functionality and structure of such a platform's components. Since the platform's objective is to provide city stakeholders with a better sensing on urban dynamics, here we propose the exploitation of information from three social media of different types, namely Twitter, Flickr and Foursquare. The specific social media platforms were selected due to their popularity and support for geotagged content sharing, as well as for their complementarity in terms of the platform's needs. This complementarity is justified by the fact that Twitter users generate (mainly) textual content focused on real time reporting, Flickr is a source of annotated photos that usually correspond to landmarks or events, whereas Foursquare offers aggregated information (in terms of time-dependent visiting frequency, activity type, and individual POI preferences, etc.) about trails of the users within a geographic area.

Next, we initially outline the CityPulse's Backend and, then, we present the proposed Frontend.

4.1 Backend processes and functionality

The Backend is structured in a three-tier architecture, comprised by the following components: (a) the Data Collector which is responsible for gathering geotagged data from social media, (b) the Data Aggregator which aggregates

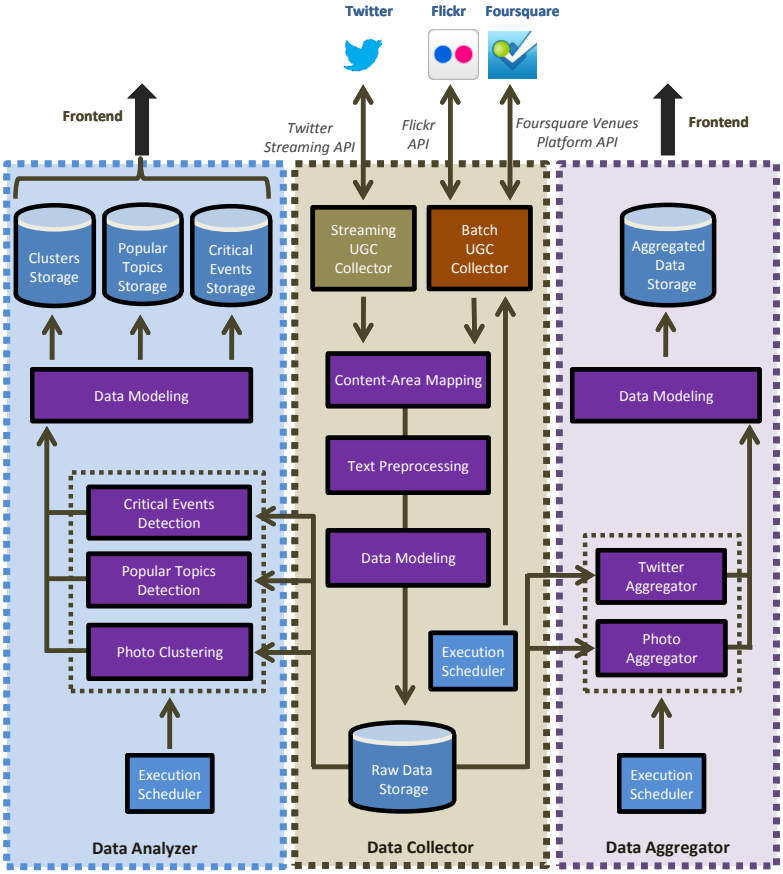


Fig. 1: Backend Architecture

the collected data and produces summaries, and (c) the Data Analyzer which is responsible for the analysis of the collected data and the extraction of useful information. More specifically, the Social Data Collector performs data collection under either a streaming or batch mode, while the Data Aggregator runs once a day to generate and store summarized reports regarding the activity observed within the geographic areas / sections of a city. Finally, the Data Analyzer component operates on the collected data to detect localized trending topics / events discussed in social media and organize the shared photos in clusters based on a location-wise and topic-wise approach. The results of the aggregation and analysis are stored and indexed in a lightweight, document-oriented database and further leveraged by the Frontend for presentation purposes. The following subsections, describe the three main components of the Backend in terms of their functionalities, scope, data, structures, and methodology. An overview of the Backend architecture is depicted in Figure 1.

4.1.1 Data Collector

This component, which is depicted in the upper-middle part of Figure 1, acts as a connector between the targeted social media and the platform. It primarily operates under an online / streaming mode to collect Twitter content in real time, via the Streaming UGC Collector, while it also runs periodically to collect photos recently uploaded to Flickr, as well as to gather POIs within a city based on Foursquare activity, via the (Offline) Batch UGC Collector. In the following paragraphs we will elaborate on these two types of data collection, the Streaming UGC Collector and the Batch UGC Collector, and the phases that follow them, namely: the Content-Area Mapping, the Data Processing, and the Data Modeling phases.

The *Streaming UGC Collector* is a constantly running program which communicates with the Twitter platform via the Twitter Streaming API. To collect geotagged Twitter posts within a city, the "locations" request parameter of the API is used. This parameter specifies a bounding box, which is used to filter the Twitter posts streamed to the application making the request based on their geolocation. The response comprises Twitter posts in a JSON format that are either geotagged within the defined bounding box or the post is annotated with a place mapped by Twitter to a geographic area that overlaps with it. Posts of the second category are filtered out, since they cannot be mapped to a specific geographic point.

The *Batch UGC Collector* is an offline process which is triggered periodically based on a predefined time interval and involves two different tasks: (i) the collection of UGC from Flickr which is geotagged within a city (fired at a daily interval) and (ii) the communication with Foursquare for the retrieval of the recommended and popular places within a city (fired at a sparser time interval under the hypothesis that a certain amount of time is required for the overall POI popularity to be affected). More specifically, the Flickr data collection is conducted by using the Flickr API[13] to collect photos posted by users in a given time period within a predefined geographic area. The result of each query to the Flickr API is a list of photos satisfying the request's parameters (time interval and geolocation bounding box) represented in a JSON format, and including information corresponding to the user who posted the photo, the title of the photo, the URL (Uniform Resource Locator), and tags, etc. The Popular place data collection is conducted by communicating with the Foursquare Venues Platform API[14] to obtain a list of the recommended and popular POIs within a city. These POIs are identified based on the Foursquare rating mechanism in relation to a range of place categories which conform to Foursquare's classification (food, drinks, coffee, shops, arts, outdoors, sights). In fact, the retrieval of popular places from Foursquare is repeated on each different weekday, since according to the Foursquare Venues API description, the recommended places differ depending on the weekday. Additional infor-

[13] Flickr API: `flickr.com/services/api`

[14] Foursquare Venues Platform: `developer.foursquare.com/overview/venues`

mation can be acquired for each retrieved POI, such as a list of related photos and tips submitted by users with respect to the POI.

The *Content-Area Mapping* phase follows the arrival of each new post / photo / popular POI information, and involves the identification of whether the new geotagged information is within the span of one of the given city areas used as a reference basis throughout the CityPulse.

The *Text Preprocessing* phase involves the *URL Expansion* and the *Text Cleaning* processes. The first one pertains only to the Streaming UGC Collector (Twitter posts). Twitter posts often contain URLs that appear in a shortened form derived by dedicated Twitter or third-party services, i.e. the length of the URL becomes shorter, but it still links to the same web page. To enable the identification of multiple references to a given URL and also multimedia content shared in Twitter, this process expands the shortened URLs into their actual form, identifies their domain, and examines whether they correspond to a photo or video shared via a relevant third-party social media application. Regarding the Text Cleaning process, it is applied not only on the Twitter posts, but also on the titles and tags of Flickr photos to remove all the uninformative content and prepare them for the following data mining step. This procedure involves the removal of common (stop) words, URLs and special notation in Twitter (user mentions, represented with "@username", or prefixes added to rebroadcasts of tweets having the form "RT @username") and Flickr (e.g., machine-tags which are irrelevant to the photo's theme), or very common terms, e.g. the name of the city.

The *Data Modeling* phase is responsible for the modeling and storage of the collected data. Essentially the platform has separate databases with appropriate representation models for Twitter posts, Flickr photos and Foursquare POIs. The adoption of the *Raw Data Storage Models* (depicted in the top-left part of Figure 2) in CityPulse was based primarily on the kind of information required for the provision of the platform's services (considering both analysis of data and delivery of results), combined with the available information offered by the selected social media APIs. A common representation form was followed for all types of collected content, wherever possible, that allows maintaining connection to the original social media application (for collecting additional information in the future and supporting attribution to the source), and also positioning the content both at a specific geographic location as well as at a sparser geographic area (according to the city division into areas). Thus, all data models have the following fields in common: the geolocation of the content and its mapped geographic area, a local ID and its unique ID in the corresponding social media application, and user-contributed tags. The data model for Twitter posts additionally includes: the text, the user who created it and its creation time, as well as an array containing information on referenced URLs. Each array element contains: the shortened version of the URL, its expanded form, its type (either "simple" web page, photo, or video), and its domain. URLs that correspond to photos are aggregated together with content from Flickr by the Data Aggregator. For each collected Flickr photo, we also maintain its timestamp, title and URL, along with the user who up-

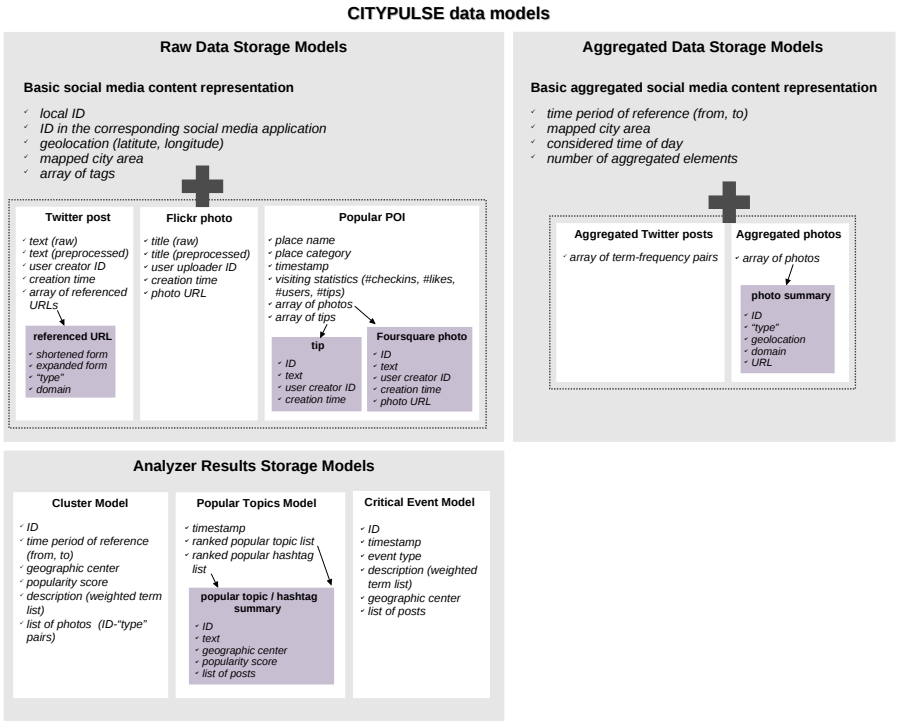


Fig. 2: CityPulse data models

loaded it. The semantic analysis of posts and photo descriptions is assisted by the additional storage of the preprocessed versions of the post texts and the photo titles. Finally, the data model for each popular POI additionally includes: its name, category, related photos and tips (if any), statistics with respect to its popularity (number of checkins, number of checked-in users, number of tips and number of likes), and the time on which it was retrieved from Foursquare. The information maintained for POIs allows their ranking, in terms of popularity, and their comprehensive presentation in smart city applications in relation to tourism and entertainment.

4.1.2 Data Aggregator

This component, depicted in the upper-right part of Figure 1, is responsible for generating aggregated summaries of the retrieved social data from Twitter and Flickr, at a regular time basis, to enable fast retrieval of the information needed by the Frontend. It performs the aggregation of (i) Twitter posts and (ii) Photos which are either extracted from Flickr or correspond to URLs referenced in Twitter posts. The execution is triggered similarly to the process

responsible for the initiation of the (Offline) Batch UGC Collector, while the aggregation is performed at a daily granularity.

The Data Aggregator is based on the axes of geographic location and time and, to this end, it groups posts / photos which correspond to the same area of the city and also have been posted / uploaded on the same day. In addition, it performs more fine-grained aggregation in terms of time, by generating different summaries for the content posted during the day-time and night-time periods of each given date (the time limits for the day and night time intervals are set manually). Therefore, for each date and time period three summaries are generated with respect to the observed posts and photos: the day-time, night-time, and overall summaries. These are represented as database records in the **Data Modeling** phase and stored at dedicated instances.

In the case of the collected Twitter posts, apart from the time- and location-related information, the corresponding summaries contain a term-frequency list of all terms that appear in the aggregated posts, and the number of posts. The photo aggregation results in a similar representation, with the exception that instead of a term-frequency list, it contains a list of some basic information about the aggregated photos. In specific, for each photo we keep its id, type ("flickr", "native twitter", or "Web photo via twitter", depending on whether it has been acquired from Flickr, uploaded directly to Twitter, shared in Twitter via a third-party social media application, respectively), domain (of the social media application in which it was originally uploaded), URL and taken date. The adopted *Aggregated Data Storage Models*, depicted in the top-right part of Figure 2, offer convenient, basic summaries of the posts and photos produced by social media users in a city context, which can easily be adapted to any temporal and spatial granularity. The number of elements contained in each such "summary" can be used by smart city applications as a numeric indicator of the overall activity at a given time and location, whereas the more detailed information kept for the corresponding posts and photos offers additional context the observed activity.

4.1.3 Data Analyzer

The Data Analyzer component analyzes social data to obtain insights on the activities of people within a city. It performs three main tasks, namely: (i) Popular Topics Detection, (ii) Critical Event Detection, and (iii) Photo Clustering, which are implemented in different sub-components, as depicted in the upper-left part of Figure 1. The results of each sub-component are stored in dedicated databases based on the models depicted in the bottom part of Figure 2. Overall, the adopted models allow positioning the discovered insights in the temporal and spatial dimension and offer their summarized descriptions to facilitate both the Frontend as well as third-party applications that may consume the CityPulse results.

The *Popular Topics Detection* sub-component is responsible for the recognition of the most popular topics discussed by citizens. The detection of new topics is held on nearly real-time analysis from the stream of Twitter posts.

For popular topics detection, a lightweight streaming data processing approach has been developed, which applies an additional filtering process on the collected posts (for spam removal), and maintains a sorted collection of the most frequent n-grams encountered in them over a rolling time-window, as well as a sorted collection of the most frequent hashtags (based on Apache Storm[15]). The top-K topics (both n-grams and hashtags) in terms of popularity are regularly exported to their dedicated database so that at any given time, the most recent popular topics can be retrieved along with their related posts.

The **Critical Event Detection** sub-component is responsible for the detection and the reporting of critical events, such as earthquakes, fires, storms, car accidents, etc., that are discussed through social media by citizens. Based on Sakaki et al (2010), the platform follows a near real-time methodology for critical event detection. At first, a list of target events is defined including a set of keywords typically encountered in the context of each event. Then, each streamed Twitter post is semantically analyzed to identify possible connections to target events due to the existence of indicative keywords. A post classifier is required to identify whether a candidate post is indeed relevant to an on-going critical event, which can be implemented as a *Support Vector Machine* using as features: all terms in the post, their number, and the context of the discovered event-related keywords. A probabilistic spatiotemporal model is employed for detecting the occurrence of critical events and estimating their central location. Apart from the event's type, time of discovery and estimated geographic center, a list of associated posts and a summarized description of the event (containing most frequent terms) are stored in the dedicated database. Obviously, since all unsupervised approaches may result in a number of false positives, due to the sensitive nature of this sub-component (since it involves informing humans on critical issues), possible events should be first validated by a human operator before alert posting.

The *Photo Clustering* sub-component is responsible for detecting groups of similar photos (based on the Data Aggregator component), shared in social media, based both on semantic and on geographic information. This is triggered at a fixed time interval taking into account the most recent photos under a rolling time-window scheme. Regarding the semantic information, clustering relies on the metadata derived either from Flickr or Twitter photos, and in particular, it is based on the TF-IDF vector representation of each textual description of a photo (derived by the Text Preprocessing process). To address the high dimensionality of the vectors and improve performance, we apply the LSI (Latent Semantic Indexing) dimensionality reduction algorithm (Deerwester et al, 1990). Having initially calculated the pair-wise photo geographic distance (taken as their spherical distance based on the Haversine formula) and their mapping onto a number of the most significant semantic dimensions (LSI results), photos are clustered via an algorithm that combines both the geographic and semantic types of similarity (Masutani and Iwasaki, 2007). For each cluster, the platform generates and stores a model contain-

[15] Apache Storm: `storm.apache.org`

ing: a list of the photos assigned to the cluster (their numeric identifiers and "type" which are required for their retrieval), a summarized description containing terms and their estimated "significance" for the cluster (based on their frequency), the cluster's estimated geographic center, and a cluster popularity score (for ranking purposes) in terms of the number of photos which belong to them and the number of unique users-photographers.

4.2 Frontend

The CityPulse Frontend targets at presenting the data analysis results derived from the Backend. To better support interactivity and live updating needs, a mobile application is proposed to reach city stakeholders in a continuous and real time manner. It is comprised of four services: (i) Recommendation of Popular Places, (ii) Imprinting of City Areas' Social Activity, (iii) Presentation of Popular Topics, Events and Photo Clusters. Indicative screenshots of the mobile application are provided in Section 5.

4.2.1 Recommendation of Popular Places

Based on POIs collected from Foursquare by the Data Collector component (Batch UGC Collector), this service provides information about popular places around a user's position (based on the device GPS) and relevant to her specific preferences (i.e. food, drinks, coffee, shops, arts, outdoors, and sights). The retrieved data is populated to the user's mobile device screen through a list of scrollable POIs, which are sorted by increasing distance from the user's location. Result filtering is enabled using a free text search mechanism which allows users to search for popular nearby POIs. Additional information is provided for each POI, including statistics (i.e. number of checkins, number of unique users, number of tips and likes), tips, relevant photos and its exact location, while direct attribution to the corresponding POI web page in Foursquare is also available.

4.2.2 Imprinting of City Areas' Social Activity

This service provides information about the popularity of city areas with respect to social media activity, which is measured based on two axes: (a) the number of posts in Twitter, and (b) the number of photos uploaded to Flickr or shared with Twitter. The user can zoom in on an area to explore the related shared photos in a gallery mode or Twitter posts in a tag cloud form. Tag clouds are generated based on the terms observed in the corresponding posts and their frequencies in the context of the given area.

4.2.3 Presentation of Popular Topics, Events, and Photo Clusters

These services demonstrate the results of the Data Analyzer component to the mobile application users in three axes: (i) visualization of clusters of semanti-

cally and geographically similar photos shared in Twitter and Flickr (i.e. photo spots that may correspond to a popular place or event), (ii) presentation of the most recent popular topics discussed in the city, and (iii) provision of the currently active events. The derived *photo clusters* are presented on a map as pins positioned at the geolocation of each cluster's center (calculated based on the geocoordinates of its photo members). By zooming in on each cluster, a user can view detailed information, including: related tags and photos (i.e. photo gallery view mode and geolocation of each photo). Finally, a user can retrieve a list with the most popular photo clusters based on the popularity score (as defined in the Photo Clustering sub-component). A similar visualization approach is followed for the presentation of the current *popular topics* which are discussed within the city. The application presents to users a ranked list of topic summaries (generated based on each topic's most indicative terms), while users can select a list item to view details for a particular topic (related posts and their geolocation). Identified critical events are presented separately in a list prioritized by the event's proximity to the given user, where each event is described with a short summary of relevant keywords (similarly to popular topics). Again, by selecting an event in the list, users can view details on the event (similarly to popular topics).

5 CityPulse demonstration

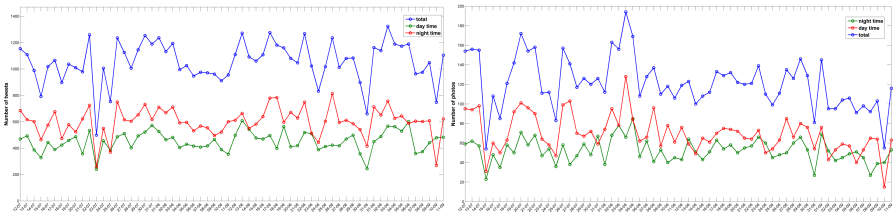
Next, we demonstrate the CityPulse platform in terms of the results derived by its operation in a selected city case study, and present the CityPulse mobile application.

5.1 Social analytics for a city scenario

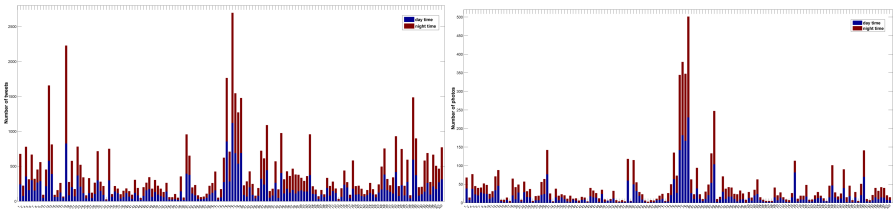
CityPulse was applied in a smart city case study, i.e. the city of Santander in Cantabria, Spain, in the context of the SmartSantander research project (within the SEN2SOC experiment). The Santander case study demonstrates the usefulness of CityPulse in terms of understanding the activities of people within the city. We present the results derived by the CityPulse application on the city of Santander for a specific time period based on the Social Data Aggregation, Photo Clustering and Popular Topics Detection functionalities that were available at the moment. The derived results reveal interesting insights about the city and its people's activities. Such insights can be useful for the city administration; for instance, city administration can easily spot "neglected" city areas (i.e. of very low popularity) and plan supporting actions, or popular ones that can serve as potential places for targeted events. Moreover, indicatively, by observing the identified popular topics, they can be readily informed about what happens within the city and how it affects its people.

5.1.1 Case study: Social data insights for Santander city

This section presents some statistical analysis and mining results derived from the operation of CityPulse on the Santander case study, covering the time period from 12/07/2013 to 12/09/2013 (Vakali et al, 2013). During this period, the CityPulse Backend monitored Twitter, Flickr and Foursquare for UGC geolocated in Santander, and applied the CityPulse's analysis pipeline. A division of the city in 148 geographic regions provided by the National Statistics Institute of Spain[16] was adopted for the segmentation of Santander in areas.



(a) Twitter daily activity timeline (b) Photo daily uploads timeline



(c) Twitter posts per city area (d) Photo uploads per city area

Fig. 3: Social media activity summarization. (a) and (b) depict the intensity of users' activity within Santander for each day of the monitoring period based on Twitter posts and shared photos, respectively. (c) and (d) illustrate the number of posts and photos geolocated within each geographic area of the city of Santander, accordingly. (Best viewed in color)

Data Collector and Aggregator In the 2-month monitoring period, the Data Collector retrieved in total 65,348 Twitter posts, 43% of which were posted during the day-time, and 57% during the night-time, respectively. Regarding the collection of photos (from both Twitter and Flickr), CityPulse retrieved 7,548 photos in total, with similar day-time versus night-time posting ratios.

The activity observed in Twitter for each day of the monitoring period is depicted in the diagram of Figure 3(a), where the different lines represent the number of posts published during the day-time, night-time, and whole day. A

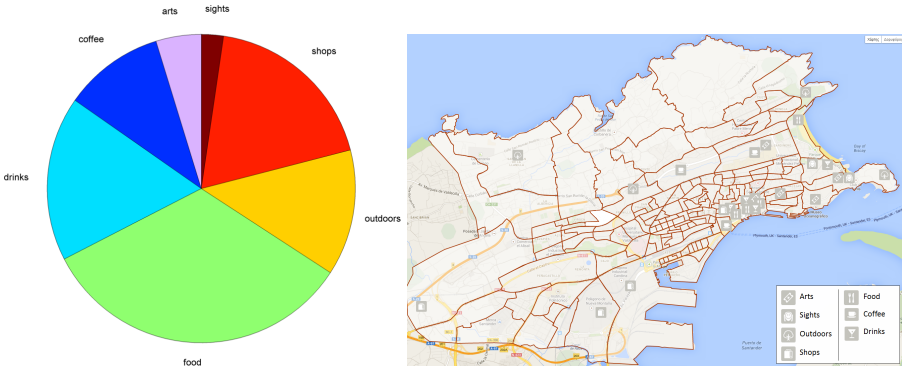
[16] This geographic division is reviewed every 10 years and sections of Santander are adapted according to various operational criteria, such as the constraint of 2000 inhabitants per section.

Table 1: Frequency of URL domains in the Twitter dataset

Domain	**Frequency**
pbs.twimg.com	5138
instagram.com	1644
foursquare.com	1086
youtube.com	250
ask.fm	216
sylodium.com	150
endomondo.com	48
esentialcine.blogspot.com	36
spreaker.com	32
twitter.com	28
elmundo.es	18
eldiariomontanes.es	16
elpais.com	14
twitpic.com	13
pediatriabasadaenpruebas.com	13
europapress.es	13
publico.es	12
path.com	12
eldiario.es	11
rtve.es	11
20minutos.es	11
armakdeodelot.blogspot.com	11
antena3.com	11
eltomavistasdesantander.com	10
other	544

similar diagram is depicted in Figure 3(b) for the number of photos shared in social media within Santander. Figures 3(c) and 3(d) depict the distribution of all collected posts and photos, respectively, across the geographic areas of Santander. There is great diversity within the activity observed in each area, with many photos geolocated in two areas, in the middle of the diagram, which cover the El Sardinero bay area and the Magdalena Peninsula, respectively. Regarding the activity in Twitter, on the other hand, the two most *active* areas cover in accordance: the La Pereda neighborhood; the university area which also includes the stadium of the Real Racing Club Santander team. To understand the type of content shared in Twitter in the area of Santander, we processed the URLs referenced in the collected Twitter posts, by taking the *expanded* URLs derived from the URL Expansion process and extracting their domain. We found that 16.5% of the 65,348 collected posts contained URLs, with their domain distribution being presented in Table 1. This table contains domains which correspond to more than 10 URLs, whereas the frequency of the "other" domain stands for the number of URLs which belong to all other (infrequent) domains. The results indicate that the most frequent domain in shared hyperlinks in Twitter corresponds to the native Twitter photo sharing mechanism ("pbs.twimg.com"), whereas the majority of URLs belong to

multimedia sharing sites such as Instagram and YouTube, and other social networking applications, such as Foursquare.



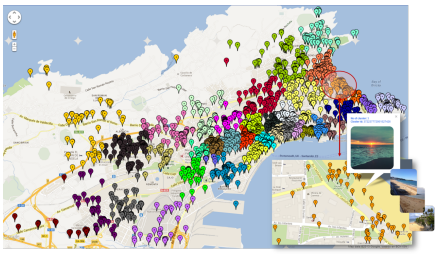
(a) Distribution of Santander popular places in terms of activity

(b) Preferred Santander areas depending on the activiy type

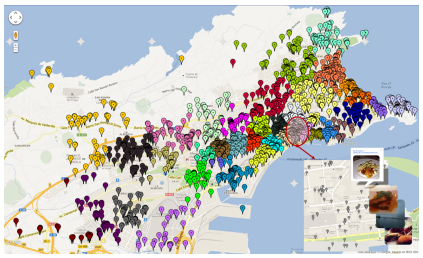
Fig. 4: Results from the analysis of Foursquare activities

The collection of popular places within Santander from Foursquare resulted in 865 places in total, over all weekdays. Popular places span different categories (activities), as presented in the legend of Figure 4(b). The distribution of the retrieved popular places in the different activities is presented in the pie chart of Figure 4(a). The retrieval of popular places from Foursquare for each distinct weekday led to relatively similar results. On the other hand, differences were observed in the types of popular places which were retrieved in the context of different geographic areas. Figure 4(b) depicts the areas with the most popular places for each activity. Each activity is represented in the map with the corresponding category symbol, while e.g. a "Food" symbol in a given area "Calle Muelle de Calderón" means that after retrieving the recommended "Food" places for each area from Foursquare, "Calle Muelle de Calderón" was in the top-3 areas regarding the retrieved number of places. It can be observed that people in Santander prefer different types of areas for each type of activity, whereas some areas are preferred for combinations of activities, such as e.g. for coffee and arts.

Data Analyzer results Photo clustering can contribute to the organization of UGC shared within Santander by providing comprehensive summaries and assisting the presentation of photos in maps. Here, we present a set of clusters derived from geolocated photos in Santander covering a 1-month time period. The geographic distribution of photos is presented in Figure 5, where it can be observed that pins (photos) of the same color (cluster) are located nearby. Figures 5(a) and 5(b) depict two randomly selected clusters of photos detected by our approach: the cluster of Figure 5(a) is located near the El Sardinero

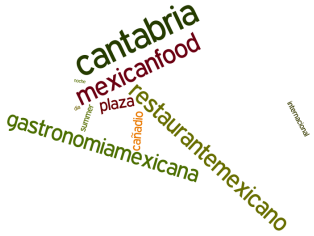


(a) "Beach cluster" situated near the El Sardinero beach area



(b) "Food cluster" situated in the city center

(c) Tag cloud for the "Beach cluster"



(d) Tag cloud for the "Food cluster"

Fig. 5: Clusters of photos geolocated in Santander. Each pin represents a photo, while the pin's color denotes the cluster in which the photo was assigned to. (Best viewed in color)

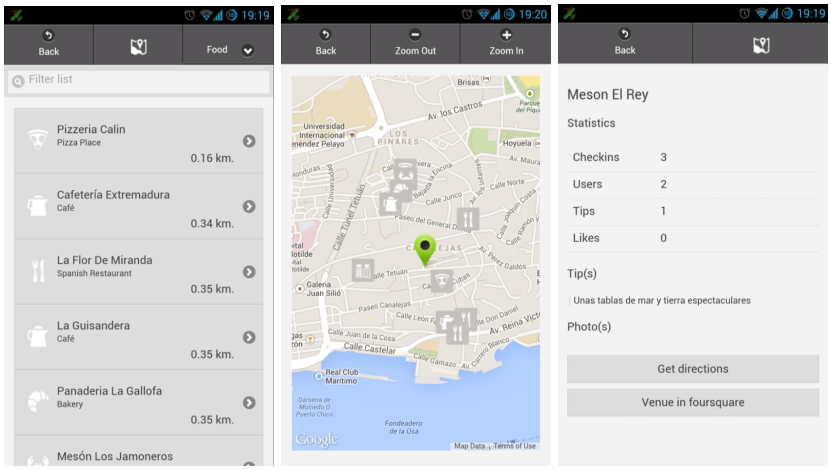
beach area and contains mainly photos with beach sceneries ("Beach cluster"), whereas the cluster of Figure 5(b) is located in Santander's center and contains mainly photos of restaurants and food ("Food cluster"). The difference in the topic aspect of each cluster is also evident in Figure 5(c) and Figure 5(d) which present the tag clouds for the top-10 most representative (significant) terms for the corresponding clusters, respectively. In more detail, the terms included in the tag clouds are the top-10 most representative terms derived for each cluster by our photo clustering method, while the size of each term is proportional to its significance score.

5.2 CityPulse mobile application

In the following paragraphs, we describe the main components and services of the CityPulse Mobile Application[17]. The Mobile Application aims to satisfy the users' needs in terms of the provision of dynamic, city-relevant information.

With respect to the *Recommendation of Popular Places* feature (described in Section 4.2.1), the mobile application users are presented with a list of POIs, which is populated based on the current day of week and the users' geographic location, as depicted in Figure 6(a). Users have the option of viewing the POIs

[17] All figures presented in this section correspond to the Santander case study.



(a) Popular POIs populated as list (b) Popular POIs on the city map (c) Details on a selected POI

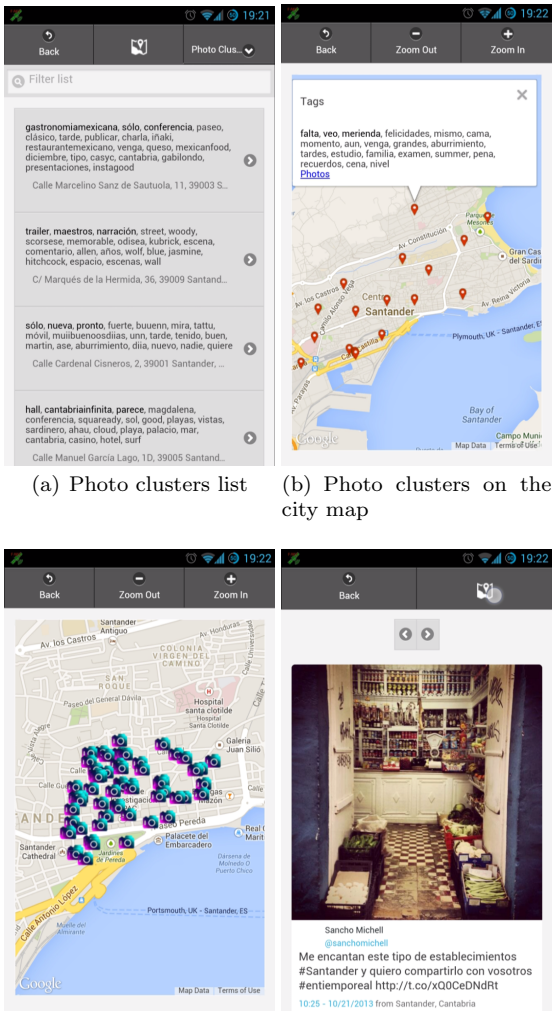
Fig. 6: Popular places based on social activities

(a) Area selection on the chromatic activity map (b) Tag Cloud view for a selected area

Fig. 7: Social activity in city areas

on the city map (Figure 6(b)) or selecting a specific POI to access additional information about it (Figure 6(c)). Additionally, users can view the location of the selected place, with respect to their location, depicted on the map by clicking on "Show on Map" button.

The *Imprinting City Areas' Social Activity* service (see Section 4.2.2) is provided to the mobile application users via the Social Activity Map, which



(a) Photo clusters list

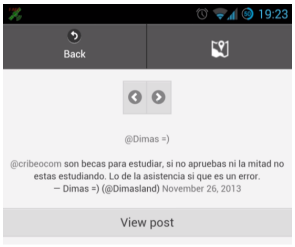
(b) Photo clusters on the city map

(c) Photos in a selected cluster

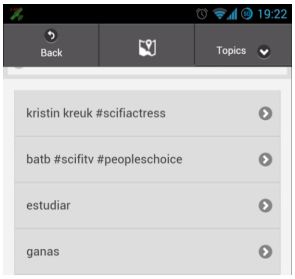
(d) Presentation of a selected photo

Fig. 8: Presentation of photo clusters

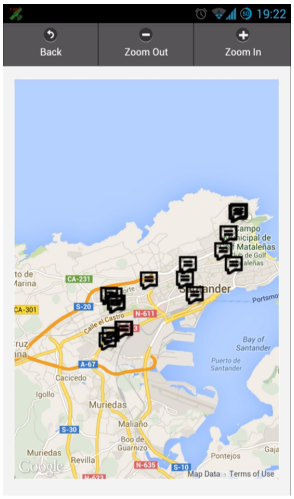
provides information about the popularity of city areas based on social media activity. Social media activity is depicted in terms of either (a) the number of Twitter posts, or (b) the overall number of shared photos. A user can zoom in on an area to explore the related shared photos or posted tweets. In the case of photos, users navigate to a screen where the related content is presented in a gallery mode. Similarly, when users select an area on the map that is based on Twitter activity (Figure 7(a)), they are presented with a tag cloud, generated



(a) Popular topics list



(b) Presentation of a selected post



(c) Topic-relevant posts on the city map

Fig. 9: Presentation of popular topics

by the terms observed in the corresponding posts and their frequencies in the context of the given area (as displayed in Figure 7(b)).

The *Presentation of Popular Topics and Photo Clusters* service (described in Section 4.2.3) presents the identified (a) photo clusters and (b) current popular topics to the mobile application users. For the clusters' visualization, users retrieve a list of the most recent photo clusters sorted by the popularity score as displayed in Figure 8(a). Related tags are used as the title of each cluster of the list and the address of its geo-center location is used as a subtitle. Then, users can select the "Map View" to view all the clusters on the city map as depicted in Figure 8(b). By selecting a specific cluster, users can view additional information by clicking the "Details" button, including all the tags that are related to the cluster, and a photo gallery view mode to navigate through its photo members, as in Figure 8(c) and obtain details on photos of their choice. Moreover, by clicking on a photo, users are presented with details about it (Figure 8(d)).

Similarly, Figures 9(a), 9(c) and 9(b) depict the list of the most recent popular topics discussed within the city; information about a selected, topic-specific post; and a map-based view of the posts referring to a specific topic (the selection mechanism is the same as for the photo clusters), accordingly.

6 Discussion

The design and implementation of collective awareness platforms is crucial in today's smart cities, and closely associated with the need to aggregate and summarize existing social networking online and offline activities that are numerous, multi-type and fragmented. Being rich in geolocated UGC, social media have a lot of potential for serving as sources of real time information in the context of a smart city. In this work, we posed research question RQ1 in an attempt to identify the ways in which social media can be leveraged for deriving crowd-sourced knowledge in relation to the city. To answer this question, we identified the prevalent categories of social media applications and studied them in terms of their data type and availability. Based on our experience, in order to maximally benefit from the available UGC, a smart city platform should use as sources diverse social media applications that offer different types of data (e.g. focused on news dissemination, location sharing and multimedia content) and which additionally offer open APIs to provide access to them. For these reasons, we selected Twitter, Flickr and Foursquare as indicative social media applications of different focus, and proposed appropriate models and methodologies that can be applied on such data to derive results revealing latent information about urban activities and conditions (popular topic detection, area / place popularity estimation, event detection shared content summarization). The proposed data mining tasks are a small subset of the possibilities that social media analysis holds for urban dynamics discovery. However, we believe that they constitute a set of basic operations that can be performed on geolocated UGC to derive results that match a smart city application scenario that affects several stakeholders, as indicated by the demonstrated case study of CityPulse. Having a platform that is responsible for the collection, modeling and aggregation of social data in an urban context, and which provides access to such data through open APIs, will give rise to several third-party applications (as the presented Frontend) that exploit social data in different ways to benefit the city.

We also addressed the associated problem of implementing a platform that would perform social data analysis in a smart city context, as the one mentioned above (RQ2). Social data mining based on several parallel data threads from different social media is challenging due to the unpredictable data generation rates, the data heterogeneity, the different data access constraints, and the requirement for (near) real-time processing. Therefore, we emphasized the design architecture and implementation / deployment of CityPulse so that it can efficiently support the parallel and intensive data mining operations. To this end, we proposed a modular framework design that separates the processes of data collection, aggregation and analysis and described it in detail. The proposed framework is flexible since it can be applied on different geographic areas (e.g. cities) and can be adapted according to the specific city parameters and requirements (e.g. more frequent triggering of the different processes through the Execution Scheduler). CityPulse is the implementation of this framework and can be incorporated in a smart city architecture

as a service provider that will make available social data analysis insights, so that they can be leveraged by smart city applications. The unobtrusive operation of CityPulse in the context of the Santander case study for a 2-months period is a promising indicator for its suitability for smart cities of today. However, further experimentation with CityPulse on other smart city scenarios is required in order to more strongly ground such a claim.

7 Conclusions

In this work, we presented the CityPulse platform, an architectural prototype which exploits online social data in a flexible and adaptive manner, due to its design and capability to manage parallel data threads from different social media, and provides services to city communities. The platform's flexibility is supported by its modular architecture that separates the processes of data collection, aggregation and analysis. UGC monitoring and summarization techniques enable the detection of trends and scales of popularity, showcased via experiments on the city of Santander in Spain, which is one of the European major smart cities, over different social media data threads (Twitter, Foursquare, Flickr). It is worth mentioning that the platform has so far enabled the experimentation in this city, but can easily be customized to handle any city's such data threads. The CityPulse platform's demonstration in the selected city case study exhibits very interesting results for the summarization of social media. The revealing of the identified social networking activity's evolution can offer city administration the means to reach the public in an immediate manner and the citizens to realize their community collective activities, so that a mutual benefit for the city can be reached.

CityPulse platform and its case study offer intuition and motivation for proceeding with open community-driven platforms and collective awareness in smart cities of tomorrow. Here, we have shown that many latent and hidden opinions can be revealed for the benefit of the city administration's decision making and the citizens' common good. To safeguard such future platforms, an extension of our work is underway to involve the exploitation of people's sentiments, and more advanced models for unified social networking data representations. Another interesting future extension is the combination of CityPulse's automatically derived insights from social activities with explicitly provided user feedback, so as to improve and extend the offered services. Finally, approaches for incorporating additional layers of city-specific data need to be explored, which should appropriately address anonymity, privacy and proprietary concerns to increase trust and viability of collective awareness platforms towards their wide adoption by smart city communities.

References

Anthopoulos L (2015) Defining Smart City Architecture for Sustainability. In: 4th IFIP Electronic Government (EGOV) and 7th Electronic Participation

(ePart) Conference

Anthopoulos L, Fitsilis P (2015) Social Networks in Smart Cities: Comparing Evaluation Models. In: Proceedings of the First IEEE International Smart Cities Conference (ISC2-2015)

Chen L, Roy A (2009) Event detection from flickr data through wavelet-based spatial analysis. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, ACM, New York, NY, USA, CIKM '09, pp 523–532

Deerwester S, Dumais T S, Furnas W G, Landauer K T, Harshman R (1990) Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science 41(6):391–407

Georgiev P, Noulas A, Mascolo C (2014) The Call of the Crowd: Event Participation in Location-based Social Services. ArXiv e-prints 1403.7657

Hollands RG (2008) Will the Real Smart City Please Stand up? Intelligent, Progressive or Entrepreneurial? City 12(3):303–320

International Telecommunications Union (2015) Technical Specifications on "Setting the framework for an ICT architecture of a smart sustainable city (SSC-0345)". http://www.itu.int/en/ITU-T/focusgroups/ssc/Pages/default.aspx

Komninos A, Besharat J, Ferreira D, Garofalakis J (2013) HotCity: Enhancing Ubiquitous Maps with Social Context Heatmaps. In: Proceedings of the 12th International Conference on Mobile and Ubiquitous Multimedia, ACM, p 52

Kumar S, Barbier G, Abbasi M, Liu H (2011) TweetTracker: An Analysis Tool for Humanitarian and Disaster Relief. In: International AAAI Conference on Web and Social Media (ICWSM)

Masutani O, Iwasaki H (2007) BEIRA: A Geo-semantic Clustering Method for Area Summary. In: Proceedings of the 8th International Conference on Web Information Systems Engineering, Springer-Verlag, Berlin, Heidelberg, WISE '07, pp 111–122

Nam T, Pardo TA (2011) Conceptualizing smart city with dimensions of technology, people, and institutions. In: Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times, ACM, New York, NY, USA, dg.o '11, pp 282–291

Papadopoulos S, Zigkolis C, Kompatsiaris Y, Vakali A (2010) Cluster-based landmark and event detection for tagged photo collections. IEEE MultiMedia (1):52–63

Sakaki T, Okazaki M, Matsuo Y (2010) Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. In: Proceedings of the 19th international conference on World Wide Web, ACM, pp 851–860

Tsampoulatidis I, Ververidis D, Tsarchopoulos P, Nikolopoulos S, Kompatsiaris I, Komninos N (2013) ImproveMyCity: An Open Source Platform for Direct Citizen - Government Communication. In: Proceedings of the 21st ACM international conference on Multimedia, ACM, pp 839–842

Vakali A, Giatsoglou M, Antaris S (2012) Social networking trends and dynamics detection via a cloud-based framework design. In: Proceedings of the

21st International Conference companion on World Wide Web, ACM Press, pp 1213–1220

Vakali A, Angelis L, Giatsoglou M (2013) Sensors Talk and Humans Sense: Towards a Reciprocal Collective Awareness Smart City Framework. In: Proceedings of the 2013 IEEE International Conference on Communications Workshops (ICC), pp 189–193

Watanabe K, Ochi M, Okabe M, Onai R (2011) Jasmine: A real-time local-event detection system based on geolocation information propagated to microblogs. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, ACM, New York, NY, USA, CIKM '11, pp 2541–2544