# Social Smart Destination: A Platform to Analyze User Generated Content in Smart Tourism Destinations

Andrea Cacho[1], Mickael Figueredo[2], Arthur Cassio[3], Maria Valeria Araujo[4], Luiz Mendes[1], José Lucas[2], Hiarley Farias[2], Jazon Coelho[3] Nélio Cacho[3]and Carlos Prolo[3]

[1] Department of Tourism, [2] Metropole Digital Institute, [3]Department of Informatics and Applied Mathematics, [4]Department of Management
Federal University of Rio Grande do Norte, Natal, Brazil
{deiacacho, mickaelfigueredo, arthurcassio, valeriaaraujoufrn, luiz.mendesfilho, joselucas-20, hiarley.farias, falecomjazon, neliocacho, carlosprolo}@gmail.com

**Abstract.** The purpose of this paper is to present a platform that uses social media as a data source to support the decisions of policymakers in the context of smart tourism destinations initiatives. The proposed platform was implemented and tested during the 2014 FIFA World Cup. In total were analyzed 7.5 million tweets. The results show that it is possible to identify the nationality, language of the posts, points of agglomeration and concentration of visitors. Overall the initial results suggest that data collected from Twitter posts can be applicable to the effective management of smart tourism destinations.

**Keywords:** Smart Tourism Destination; Smart City, Social Media; Twitter.

## 1 Introduction

The concept of smart city arises due to the complexity and management challenges faced by the authorities to deal with the rapidly urban population growth. The authors on [4] argue that a city can be defined as 'intelligent', when there is investment in human and social capital, as well as in information and communication technology (ICT) infrastructure. Smart city initiative may comprise many different areas of the city administration. For instance, smart tourism destination concept emerged from the development of smart cities [3]. Smart tourism destination can be perceived as places utilizing the available technological tools and techniques to enable demand and supply to co-create value, pleasure, and experiences for the tourist and wealth, profit, and benefits for the organizations and the destination [2].

Smart city incorporates a large number of systems, which represent the most basic infrastructure for integrating the real and virtual worlds. One of the great challenges of deployment of smart cities is the extraction of relevant information from the ICT infrastructure of cities. Such extraction usually relies on the use of sensors that are installed to capture the flow of vehicles, water and energy consumption, thus requiring high public investment for the development of smart cities [7]. To overcome such difficulty, some studies suggest using social media to identify the perception of residents and visitors about a particular city [1, 5]. For example, social media can be

Á. Rocha et al. (eds.), *New Advances in Information Systems and Technologies*,
Advances in Intelligent Systems and Computing 444,
DOI 10.1007/978-3-319-31232-3_77

used to obtain relevant information on the situation of public transport, traffic and environmental conditions, public safety and general events in cities. In this sense, the purpose of this paper is to present a platform that uses social media as a data source to support the decisions of policymakers in the context of a smart tourism destination.

## 2 Smart Destination Initiative of Natal

Natal is located on the northeast of Brazil by the Atlantic Ocean. The capital city of the state of Rio Grande do Norte is home of approximately 862.000 thousands people. The city and the surround area are well known due its sandy beaches and natural resources which attract thousands of tourists every year. Although Natal was not the location for the World Cup knockout stage, Natal hosted 4 games in the group stage with an average attendance of 40,000 fans at each game. In total, Natal received around 173,000 tourists during the World Cup period. According to a study performed by Forward Data together with Pires & Associados in Brazil [11], Natal presented the highest growing number of bookings among all host cities when compared to the same period in 2013, for which bookings have grown by more than 1000%.

The high number of tourists puts severe pressure on the urban infrastructure and services related to transportation, safety and water consumption. In order to handle such pressure, the Natal city council in partnership with public and private sector have engaged in an initiative to create a smart tourism destination. In fact, smart city concept covers a variety of industries, including the tourism industry [12]. Caragliu et al. [4] claimed that cities can be defined as smart "when investments in human and social capital and traditional (transport) and modern (ICT) communication infrastructure fuel sustainable economic growth and a high quality of life, with a wise management of natural resources, through participatory governance". The development of Smart City facilitates seamless access to value-added services such as access to real-time information on public transportation network, enriches tourist experiences and enhances destinations competitiveness [3].

Due to the complexity to define a smart city initiative that may comprise many different areas of the city administration, the Natal initiative decided to focus on some specific areas [17, 18]. This paper describes a smart city initiative to create a smart tourism destination. Hence, the initiative to create a smart tourism destination is a first step towards the creation of a smart city initiative for Natal. A smart tourism destination is perceived as places utilizing the available technological tools and techniques to enable demand and supply to co-create value, pleasure, and experiences for the tourist and wealth, profit, and benefits for the organizations and the destination [2]. Guo, Liu, and Chai[12] argue that Smart Tourism Destination is a relevant part of the construction of the smart city's application system since it depends on the infrastructure of the smart city, utilization of information resources, and development of the intelligence industry.

Among the many benefits of defining a smart tourism destination, gathering the tourist's perceptions about the city is considered one of the most important ones [3]. When tourists use the internet to express their perception by means of words, terms and phrases that form their spoken language, this perception is said to be a user generated content (UGC). A UGC has been described as creative work that is published on publicly accessible websites and is created without a direct link to monetary profit or commercial interest [14]. UGC websites have evolved into multiple forms [14]: virtual communities (e.g., LonelyPlanet), consumer reviews (e.g., Yelp), personal stand-alone blogs, blog aggregators (e.g., LiveJournal) and microblogging platforms (e.g., Twitter), as well as social networks (e.g., Facebook), and media sharing tools (e.g., Flickr, YouTube). This paper focus on microblogging UGC since it is carried out usually by mobile phone text messages and is currently restricted to just 140 characters. Twitter is possibly the best known microblogging site where users post messages ("tweets") sharing views on topics and news stories and ask advice and help [15]. Twitter produce millions of posts being broadcast over time. These posts need to be analyzed in order to extract situation awareness about tourists behavior in a smart tourism destination. Litvin et al. [16] claim that UGC on social media is a substantial source of strategic information which can be used for the development of a number of business strategies—including enhancing visitor satisfaction through product improvement, solving visitor problems, discovering visitors' experience, analyzing competitive strategies as well as monitoring image and reputation of a tourism destination.

Based on the importance of the UGC to the tourism industry [14, 15, 16], the Natal initiative decided to define and implement a platform to support the monitoring of tourist perception and movement during their visits in Natal. The Natal initiative comprises many actions [17, 18] in terms of management and organization, governance, policy and technology. Due to space constraints, next section presents only the technology solutions implemented to support the creation of a smart tourism destination.

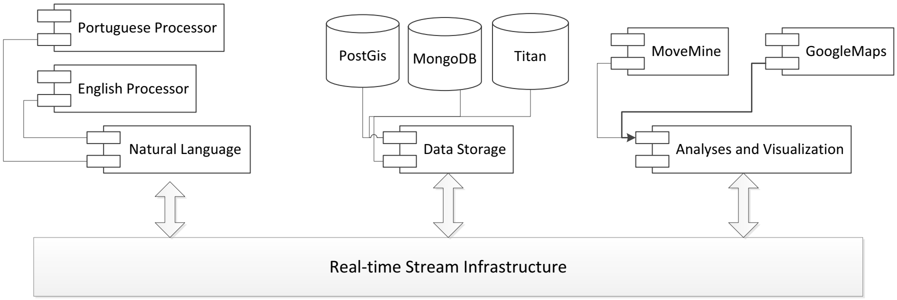

**Fig. 1.** SSD Platform Architecture.

## 3 Social Smart Destination Platform

This work has adapted and optimized a set of machine learning and natural language processing techniques to deal with real-time and high-volume text streams. These techniques are packaged in a proposed software platform named Social Smart Destination (SSD). This platform provides capabilities that include identifying early indicators of tourism industry, exploring the presence of tourism issues and monitoring tourist's behavior. In order to provide such technologies, SSD encompasses four components (Figure 1): a real-time stream infrastructure, a natural language component, a data storage component, and an analyses and visualization component. Next subsections describes in details each of these aforementioned components.

### 3.1 Real-time Stream Infrastructure

The SSD platform uses real-time processing infrastructure as a central component. This infrastructure is implemented through the use of *Apache Storm* [13]. *Storm* is a free and open source stream-processing framework capable of processing one million 100 byte messages per second per node [13]. A *Storm* cluster is formed by a distributed network of processing nodes that process a set of data compartmentalized in tuples. For this, three components are defined: (i) *zookeeper*, (ii) *Nimbus* and (iii) the *supervisor*. *Zookeeper* is a high-performance service that coordinates distributed applications through configuration management, appointment and work services group synchronization. On *storm's* architecture, it stores the synchronization of data and the processing state of *tuples* that will be performed at the nodes *supervisor*. The *Supervisor* is the nodes of the cluster *Storm* responsible for the data processing. Finally, *Nimbus* is the primary node of the cluster *Storm*, responsible for distribution of code to be processed, assigning tasks to nodes *supervisors* and fault monitoring.

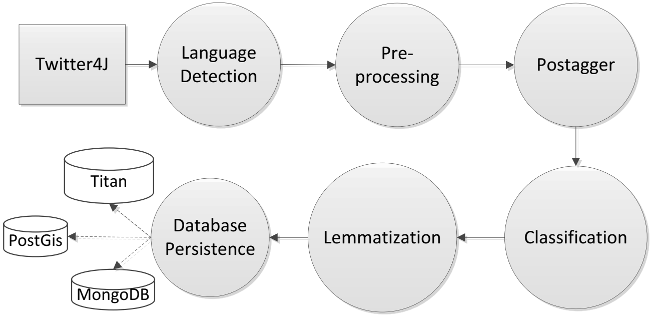

**Fig. 2.** Natural Language processing Topology.

*Storm* uses *topologies* as mechanisms for computational organization. *Topologies* are defined as a processing graph where each node in a topology contains a logic processing and links between them. The links indicate that data can be exchanged between nodes. The data stream is represented by *Streams* elements. *Stream* is a sequence of tuples that can be affected by *Spouts* components and *Bolts*. For instance, Figure 2 shows a topology to process natural language. *Spouts* and *bolts* have interfaces that can be used by developers to implement the program logic. *Spouts* (rectangles in Figure 2) are elements that receive a data stream and organize them in tuples. *Bolts* (circles in Figure 2) consume stream elements from a spout or a bolt. Each topology can still be seen as a spouts and bolts package.

### 3.2 Natural Language Component

Due to the great variety of spoken languages that visitors in a smart tourist destination may use to create UGC, the Natural Language Component (NLC) was design with extension points, which are well-defined places where other language processing components can be added. For this version, Figure 1 shows that SSD support two languages: English and Portuguese. Moreover, the Natural Language component is structured around the Storm abstractions. Figure 2 shows how Storm frames the natural language processing techniques used in this work. The top-left of this topology takes inputs from a public streaming Twitter API (Twitter4J). This *spout* creates a *stream* with tuples that contain the following data: name, age and city of Twitter users, tweet ID, latitude, longitude, date and time of the post, and body of the tweet. Some of these data, such as the user's city, are only obtained when the user allows access to your profile on Twitter.

Tuples created by *spout* are passed onto the Language Detection Library[1] *bolt* which detects the language of the post. The Language Detection bolt allows performing specific actions for the feelings of tourists from different countries that may visit a smart destination. The next bolt performs a preprocessing that includes the removal of special terms (RT, via, etc.) and hashtags treatment according to the language. The Postagger *bolt* performs the *Part-Of-Speech Tagger* (POS Tagger) using the Apache OpenNLP API[2]. The *Postagger bolt* receives a tuple with the tweet text in some language and assigns parts of speech to each word, such as noun, verb, adjective, etc. For instance in the post "I really loved this city," the *Postagger bolt* creates a tuple with the words properly classified with its context, like "I" being a personal pronoun, 'really' an adverb and "loved" a verb. Then the *Classification bolt* implements a Naïve Bayes classifier based on [9], reaching accuracy levels of about 82% with only the positive, neutral and negative classes. The *bolt Lemmatization* takes into account the language of the post to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form (lemma). For instance, in English, if confronted with the token “saw”, *Lemmatization bolt* would attempt to return either “see” or “saw” depending on whether the use of the token was as a verb or a noun.

### 3.3 Data Storage Component

The data storage is executed by the bolt *Database Persistence*. This bolt stores the processed tweets in three databases: MongoDB[3], PostGIS5[4] and Titan6[5]. *MongoDB* is an open-source document database that provides high performance, high availability, and automatic scaling. *MongoDB* was used to create two data collections. The first

[1] https://code.google.com/p/language-detection/

[2] https://opennlp.apache.org/

[3] https://www.mongodb.org/

[4] http://postgis.net/

[5] http://thinkaurelius.github.io/titan/

collection stores all tweets, creating a log for tweets. In turn, the second collection stores the most commonly used terms, their common base form (lemma) and the time they were collected. In order to understand visitor behavior, the platform uses the *PostGIS* database to create a cross-reference among the spatial information (latitude, longitude and time) generated by the Twitter posts and the list of georeferenced tourist attractions obtained from *Google Places*. This cross-reference allows identifying: (i) which are the most or the least visited places, (ii) what are the tourist perception about the attractions, (iii) and which kind of attractions are preferred by different target groups. Finally, *Titan* is used to store, in the form of a graph, the relationship between users and tweets. *Titan* is a distributed graph database optimized for storing and querying graph structures. Like *Storm* and *MongoDB*, *Titan* databases can run as a cluster and can scale horizontally to accommodate increasing data volume and user load. The graph structure of *Titan* facilitates the implementation of algorithms to discover the underlining rules governing the behavior of people in a social network, such as centrality and closeness.

### 3.4 Analyses and Visualization

The analysis and visualization component also uses bolts and functions to implement the data analysis. These elements have not been described in Figure 2 for the sake of simplification of the figure. The spouts are used to retrieve data from databases and bolts are used to generate the results that will be displayed by the graphical interface of the platform. The Analyses and Visualization components leverages the *MoveMine* [17] tool to perform sophisticated moving object data mining. MoveMine integrates many data mining functions including moving object pattern mining and trajectory mining based on state-of-the-art methods. MoveMine has many application scenarios. For example, it can automatically detect an approximate period in movements; it can reveal collective movement patterns like flocks, followers, and swarms; and it can perform trajectory clustering, classification and outlier detection for geometric analysis of trajectories. The graphical interface is a web interface in the form of a dashboard, implemented in HTML / JavaScript using the Google Maps API V3 and Google Charts library to generate maps and graphics, respectively.

## 3 Assessment Methodology

In order to assess the SSD, the platform has been run to collect and process the Twitter posts during the 2014 FIFA World Cup, which took place in Brazil. The platform ran from the 10th of June 2014 to the 15th of July 2014. The initial date was two days before the opening ceremony and the closing date was two days after the closing ceremony. The decision to collect tweets during the World Cup was primarily based on the fact that during the same period many organizations perform surveys to identify the tourist perception, nationality, etc. Hence, the intention is to compare our results with other researches performed during the same period. The platform collected automatically all tweets containing at least one of the following terms:

Salvador, Manaus, Natal, RiodeJaneiro, Recife, SaoPaulo, BeloHorizonte, PortoAlegre, Fortaleza, Cuiaba, Brasilia and Curitiba. These terms corresponds to the World Cup host cities names. The platform processed during the assessment period approximately 7.5 million tweets.

# 4 Results

The SSD platform allows performing different analysis and use different filters to be applied to the collected dataset. For instance, when only the tweets with latitude and longitude are taking into account, the number of tweets drops from 7.5 million to 286,000 tweets. Based on this subset (tweets with latitude and longitude), the SSD platform shows that 81.16% of all posts were originated in the Brazilian territory. Other South America countries accounted for 5.69% of the posts whereas the North and Central American accounted for 4.89% and 2.75%, respectively. Despite being a continent with great tradition in football, Europe appears only with 3.05% of the posts.

These data shows an implicit feature of Twitter, i.e., to be used to share information and describe day-to-day activities of people lives [6]. According to [10], 80% of users use Twitter to update their followers on what they are doing, while the remaining 20% use Twitter to send general background information. The small number of posts outside Brazil helps to confirm this notion that Twitter users have little reciprocity in the exchange of messages among users [8], unlike other social media, suggesting that the main goal of Twitter is not maintaining relationships, but disseminating personal news.

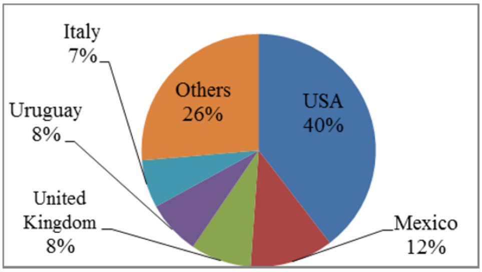

**Fig. 3.** Tourist nationality.

In order to analyze the tourism demand, it has been defined two additional filters. The first selected only the posts of Twitter users whose original location was not in Brazil (based on the user profile information), and the second restricted to tweets posted within the perimeter of the metropolitan region of Natal (i.e., posts sent from Natal, but from foreign users). These two filters have generated a subset that comprises 7,465 posts. Based on this subset, it was found that Natal received during the FIFA World Cup tourists from 25 different nationalities (see Figure 3), such as Americans (39.60%), Mexicans (11.56%), British (8.38%), Uruguayans (7.51%), and Italians (6.65%).

In addition, it was possible to identify the languages used in Twitter posts. In total, it was identified 11 different languages: English being the most used with 41% of the posts, followed by the Spanish (25%), and Italian (18%). For example, it was observed a majority presence of posts in English on the date of the match between USA and Ghana (i.e., 16th June).

Real-time identification of tourist's behavior is another information relevant to the context of smart destinations. With such information, the public managers of the tourism sector can improve the tourist infrastructure in those areas while the public safety managers can optimize the distribution of vehicles to cover the places of greatest tourist presence.

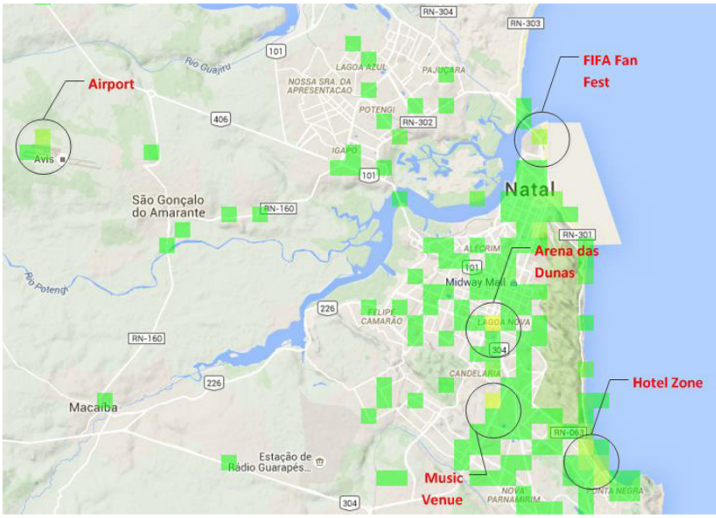

**Fig. 4.** Places where the crowds were concentrated

According to Figure 4, the area of greatest concentration of posts was in the Arena das Dunas Stadium, then a live music venue near by the stadium, followed by the hotel zone, the airport, and lastly, FIFA Fan Fest. These results revealed that the location of this music venue close to the Stadium probably was favored in comparison with FIFA Fan Fest (official music venue). Finally, it was observed that 72% of posts in English were positive, 18% neutral and 10% negative. Most of the negative posts were related to the bus service, since during the competition in Natal, bus drivers went on strike.

As described in Section 3, the purpose to analyze tweets during the world cup was to compare our results with other researches that usually performed in the same period. This comparison aims to assess the validity and reliability of the data provided by the SSD platform. One of the studies was performed by the Spanish company Forward Data [11] that in partnership with Pires & Associados in Brazil researched the nationalities of foreigner tourists to Brazil. This research looked at five billion reserves issued by 180,000 online travel agencies around the world. These results showed that for the city of Natal, 29% of bookings were made by North Americans, 14% of Uruguayans and 7% by Italians. Moreover, the greatest concentration of post in the Arena das Dunas Stadium was expected for a football competition and may suggest that the SSD can be used to identify crowd concentration in other circumstances.

## 6 Related Work

Social networks are widely used to collect information about people and events. For example, [1] and [5] describe approaches that capture tweets on a certain radius starting from an application point. Then those approaches used probabilistic models to identify problems relating to traffic and on other events. Our approach is based on the solutions defined by [1, 5] and describes additional contribution as a platform that supports the treatment and identification of events in real time using multiple languages and supporting analysis and geo-referenced data visualization. There are many researches[14, 15, 16] in the tourism and hospitality area that analyse the use of UGC as a data source. For instance, Lua and Stepchenkovab [14] surveyed 122 peer-reviewed journals articles and conference proceedings to investigate among many other thing the use of software that has been used to collect and extract information from UGC. According to [14], there is no tool with the same capabilities of the SSD platform.

## 7 Conclusion

This paper presented a platform that uses social media as a data source to support the decisions of policymakers in the context of smart destinations initiatives. The Social Smart Destination platform aims to enhance tourists' travel experience through collecting and analyzing in real-time Twitter posts. This paper detailed the topology responsible for collecting, processing, and storing the Twitter posts. Some of the data gathered by the platform was analyzed to show how the platform can be used by a smart destination. The results showed that it is possible to identify the nationality, the language of the posts, the sentiment, and the points of agglomeration of visitors during a big event. Overall results suggest that data collected from Twitter posts can be applicable to the effective management of smart tourism destinations. As future work, we intend to use the comments obtained from of the *tripadvisor* web site to train our sentiment analysis component. We believe this will improve the precision of our approach since real comments from tourists in Natal will be used to better train the machine learning algorithms. Finally, we intend to add new domains to the platform, such as the support identify issues in the public safety of a smart city.

## Acknowledgments

This work was partially supported by the National Institute of Science and Technology for Software Engineering (INES), grants 573964/2008-4(CNPq) and APQ-1037-1.03/08 (FACEPE), and by the Smart Metropolis Project.

# References

1. Anantharam, Pramod, et al. "Extracting City Traffic Events from Social Streams." ACM Transactions on Intelligent Systems and Technology 9.4 (2014).
2. Boes, Kim, Dimitrios Buhalis, and Alessandro Inversini. "Conceptualising Smart Tourism Destination Dimensions." Information and Communication Technologies in Tourism 2015. Springer International Publishing, 2015. 391-403.
3. Buhalis, Dimitrios, and Aditya Amaranggana. "Smart tourism destinations." Information and Communication Technologies in Tourism 2014. Springer International Publishing, 2013. 553-564.
4. Caragliu, Andrea, Chiara Del Bo, and Peter Nijkamp. "Smart cities in Europe." Journal of urban technology 18.2 (2011): 65-82.
5. Doran, Derek, Swapna Gokhale, and Aldo Dagnino. "Human sensing for smart cities." Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. ACM, 2013.
6. Java, Akshay, et al. "Why we twitter: understanding microblogging usage and communities." Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis. ACM, 2007.
7. Komninos, Nicos, Marc Pallot, and Hans Schaffers. "Special issue on smart cities and the future internet in Europe." Journal of the Knowledge Economy 4.2 (2013): 119-134.
8. Kwak, Haewoon, et al. "What is Twitter, a social network or a news media?." Proceedings of the 19th international conference on World wide web. ACM, 2010.
9. Silva, Mário J., Paula Carvalho, and Luís Sarmento. "Building a sentiment lexicon for social judgement mining." Computational Processing of the Portuguese Language. Springer Berlin Heidelberg, 2012. 218-228.
10. Naaman, Mor, Jeffrey Boase, and Chih-Hui Lai. "Is it really about me?: message content in social awareness streams." Proceedings of the 2010 ACM conference on Computer supported cooperative work. ACM, 2010.
11. Forward Keys and Pires & Associados. (2014) . Fifa World Cup shakes Brazilian Tourism Trends.
12. Guo, Y., Liu, H., and Chai, Y. (2014). The embedding convergence of smart cities and tourism internet of things in China: An advance perspective. Advances in Hospitality and Tourism Research, 2(1), 54-69.
13. Apache Software Foundation. *Apache Storm.* Avalilable : https://storm.apache.org/.
    Lu, W. and Stepchenkova, S. User-Generated Content as a Research Mode in Tourism and Hospitality Applications: Topics, Methods, and Software. Journal of Hospitality Marketing & Management, Feb, 2015, Vol. 24, N0. 2, p.119-154
14. Akehurst, Gary, (2009), User generated content: the use of blogs for tourism organisations and tourism consumers, Service Business, 3, issue 1, p. 51-61.
15. Litvin, S. W., Goldsmith, R. E., & Pan, B. (2008). Electronic word-of-mouth in hospitality and tourism management. Tourism Management, 29(3), 458–468.
16. Wu, F., Lei, T., Li, Z., and Han., J. (2014). MoveMine 2.0: mining object relationships from movement data. Proc. VLDB Endow. 7, 13, 1613-1616.
17. Cacho, A. et al. A Smart Destination Initiative: the Case of a 2014 FIFA World Cup Host City. Proceedings of the IEEE International Smart Cities Conference. IEEE, 2015.
18. Coelho, J. et al. ROTA: A Smart City Platform to Improve Public Safety. Proceedings of the 4th World Conference on Information Systems and Technologies. Springer, 2016.