Proceedings of the Information Access in Smart Cities Workshop (i-ASC 2014)

Editors:
M-Dyaa Albakour,
Craig Macdonald,
Iadh Ounis,
Charles L. A. Clarke,
and
Veli Bicer

Amsterdam, the Netherlands
April 13th, 2014

Preface

These proceedings contain the papers of the Information Access in Smart Cities (i-ASC) 2014 Workshop held in conjunction with the ECIR 2014 conference in Amsterdam, the Netherlands, on the 13^{th} of April 2014. Six technical papers and one positional paper were selected by the programme committee. Each submitted paper was reviewed by at least three members of an international programme committee. In addition to the selected papers, the workshop features two keynote speeches and two invited contributions. Keynote speeches are given by Pól Mac Aonghusa "A Content, Connection and Context: From Data to Insight in Smarter Cities", and Frank Kresin " Smart Cities, Smart Citizens and the case for the CitySDK". The invited contributions are given by Raffaele Perego "Mining digital footprints for smart tourism" and Jon Oberlander "Tourists in Smart Cities: Data mining for hidden treasures".

We would like to thank ECIR for hosting us. Thanks also go to the keynote and invited speakers, the program committee, the paper authors, and the participants, for without these people there would be no workshop.

M-Dyaa Albakour,
Craig Macdonald,
Iadh Ounis,
Charles L. A. Clarke,
and
Veli Bicer

Workshop Organisation

Organisation Committee

M-Dyaa Albakour (University of Glasgow)
Craig Macdonald (University of Glasgow)
Iadh Ounis (University of Glasgow)
Charles L. A. Clarke (University of Waterloo)
Veli Bicer (IBM Research Ireland)

Programme Committee

Fernando Diaz (Microsoft Research)
Jaap Kamps (University of Amsterdam)
Paul Thomas (CSIRO)
Daqing He (University of Pittsburgh)
Omar Alonso (Microsoft Bing)
Freddy Lecue (IBM Research)
Raffaele Perego (ISTI CNR)
Cathal Gurrin (Dublin City University)
Franco Maria Nardini (ISTI CNR)
Suzan Verberne (Radboud University Nijmegen)

Table of Contents

Searching Smart Cities

City Data

Context and Recommendation

HCI and Applications in Smart Cities

Content, Connection and Context: From Data to Insight in Smarter Cities

Keynote

Pól Mac Aonghusa
Smarter Cities Technology Centre, IBM Dublin Research Laboratory, Ireland

ABSTRACT

Big Data has been popularised in the media as "the new currency", fuelling a future vision of contextual systems that will transform our world. However the reality is we are only beginning to recognise significant research challenges across a spectrum of topics; from information retrieval, to knowledge representation & reasoning, and user experience, that must be addressed to realise the vision. Drawing on our research into real-world use cases in Urban Systems and Integrated Care, this talk will discuss a number of research challenges, and show early results and prototypes. The talk will be at a general level to stimulate discussion and identify possible areas for future research collaboration.

Biography

Pól Mac Aonghusa is a Senior Manager at the IBM Research, Ireland. He leads research teams focused on next generation High Performance Supercomputing (Exascale) and Big Data and Open Data for Urban Systems (City Fabric). Much of the research activity in the Big and Open Urban Data research program is being actively used with the Dublin Region to create their Open Data Portal. Dr. Aonghusa has led several emerging technology projects exploiting Big and Open Data, including Smartbay, an environmental sensing research program, collecting live sensor data from the Atlantic Coast of Ireland for research purposes. Smartbay was named as one of the 100 Icons of Progress in 2011. Dr. Aonghusa joined IBM 25 years ago, at the IBM Software Development Laboratory, Dublin. During his IBM career, he has held key Management, Technical and Consultant positions in IBM's Software, Services and Industry and Research Divisions.

ECIR'14 Information Access in Smart Cities Workshop (i-ASC 2014).
April 13, 2014, Amsterdam, the Netherlands.

From Smart Cities to Smart Neighborhoods: Detecting Local Events from Social Media

Yang Li
CLARITY: Centre for Sensor Web Technologies
Dublin City University
Glasnevin, Dublin 9, Ireland

Alan F. Smeaton
Insight Centre for Data Analytics
Dublin City University
Glasnevin, Dublin 9, Ireland
alan.smeaton@dcu.ie

ABSTRACT

There are several examples of work which uses data from social media to detect events which occur in our real, physical world. Our target for event detection is to partition a large geographic region, a whole city in our case, into smaller districts based on geotagged Tweets and to detect smaller local events. We generate a language model for Tweets from each district and measure the KL divergence on incoming Tweets to detect outliers. When these reach a sizable volume or intensity and are consistent, this indicates an event within that district. We used Tweets drawn from Dublin city and we describe experiments on partitioning the city into districts and detecting local events within districts.

1. BACKGROUND AND RELATED WORK

Much research work is reported in the literature utilizing the characteristics revealed by Twitter features, including the realtime detection of live events. Event detection has long been a research topic across many application areas and using many sources of data or information [7]. Early work leveraged natural language processing tools, such as named-entity extraction for online news event identification. Such tools work well on well-structured text like newspaper articles and TV transcripts, but do not perform well over some forms of social media such as Twitter. To address this, other methods have been proposed. Twitterstand [5] gathers and disseminates breaking news from Twitter using an online clustering method to cluster similar Twitter messages. Sakaki et al. [4] classify Twitter contents using a Support Vector Machine. Twitcident [1] enables filtering, searching, and analyzing Twitter information streams during incidents as they are happening as well as providing a faceted search interface to dive deeper into these Tweets. Other works [6] also reports real-time event detection from Twitter based on temporal and textual features of Tweets.

These previous works successfully detect breaking news or live events in Twitter streams globally, and their methods are sensitive to large-scale events which attract a large number of possibly global Tweets, such as the Presidential inauguration in the USA. This is because their target events generate significant boosts to the mainstream of Twitter and a significant volume of event Tweets which can be detected. Yet Twitter users often post information about local, community-specific events such as a local flood, a fire, or a road closure because of a tree falling, where traditional news coverage at a regional or national level is non-existent and indeed it is quite difficult to confirm if such events have actually happened. We illustrate some of these later in Table 2. The motivation for our work is examine whether Tweets, localised to a small geographical region, can be used to detect unusual events happing at a *local* level within a city. Our contribution is to use Tweets from Dublin city to partition the city into smaller regions, model the typical Twitter content for each region and then use a sufficiency of outlier Tweets to indicate the likelihood of local-level events in areas of Dublin city.

2. EVENT DETECTION IN SOCIAL MEDIA

We work on a relatively small data-set, Tweets from Dublin city. For the purpose of the detection of unusual socio-geographic events, we first determine normal crowd behavior in a geographical region of the city in terms of Twitter activity. After mapping geo-tagged Tweets onto defined partitions on a map, we focus on sudden increases or decreases in the number of Tweets happening in a geographical partition or the topics of discussion, which can be clues to an unusual event happening. Our assumption is that local events can be reported on Twitter and the content of such Tweets is a semantic irregularity to the typical Twitter behaviour of a region, i.e. people do not normally Tweet about floods, fires or road closures unless there are such events happening.

To detect unusual local events for a given large area we first partition the city area into sub-areas by establishing socio-geographic boundaries. We adopt a clustering-based space partition method that reflects geographical distribution of a dataset and better deals with heterogeneous regions. Some research works divided their target area into equally sized grids with different granularities. We chose not to use this approach because an appropriate cell size is difficult to determine and does not consider the geographical distribution of Tweets.

We adopt the K-means clustering method based on the geographical occurrences of our Tweets. The K-partitioned regions are demonstrated in different colors on a unit graph, as shown in Figure 1. As a result, we achieve an appropriate socio-geographic boundary setting for the target region by distributing the actual occurrences of Tweets. We partition Dublin into 25 regions, a number which is a guesstimate as to what would be best. When we compare the partition results to the actual population distribution of Dublin city area according to the Central Statistical Office data, as in Figure 1, we see the partition results are acceptable, so 25

ECIR'14 Information Access in Smart Cities Workshop (i-ASC 2014).
April 13, 2014, Amsterdam, the Netherlands.

seems to have been a reasonable choice. Hotspots can easily be identified, such as the city center where there is high population density and a high volume of Tweets, as well as some low population areas with a high Tweet volume such as Dublin Airport, and the Phoenix Park.

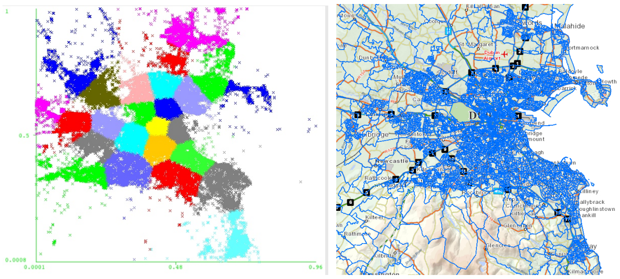


Figure 1: Geo-social partitioning of Dublin into 25 clusters and population distribution of Dublin

We make a major assumption that for each location there is consistency and periodicity in Twitter activity, such as appearances of regular users in regular locations and perhaps Tweeting about regular topics of interest. While some deviations outside usual or regular activities are caused by, for example, holidays, or visits from friends, these are mostly restricted to individuals. However when the same deviations are picked up by multiple users at the same time, same location, same topic, then this leads us to believe that we can recognize local events from inconsistencies in Twitter users' behavior at a regional level, including a change in the topic of Tweets, a so-called semantic irregularity.

We now explain how we set up the measurements of regularity. Within each partition of the city, there are Tweets generated over time, and in our work we analyze weekday and weekend days differently. This is because partitions have different activities for weekday vs. weekends such as offices which will be relatively quiet during weekends whereas shopping areas will be more active. The regularity of the total amount of Tweets are calculated using the average of each day during a rolling one month period, and with ± 1.0 standard deviation, assigned into hourly bins and any number outside the 1.0 standard deviation are considered as unusual activity, as shown in Figure 2.

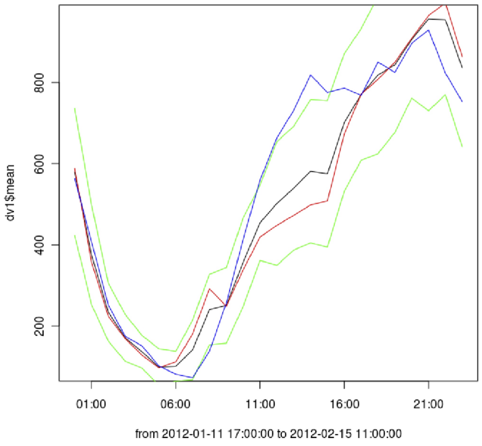


Figure 2: Twitter occurrences in hourly bins

For every partition we store a set of regular active Twitter users. If there are many visiting Twitter users sending Tweets from the partition, we consider this as another clue of irregular Twitter activity.

Measuring semantic regularity of Tweets in partitions is more complex. For each geo-tagged Tweet in our collection, we use all of the texts in each partition to build a language model that represents the semantic consistency of the partition. In order to preserve the semantics of Tweet contents we do not apply any stop-word filtering, and special characters such as "#" and "@" are not removed.

We use a language modeling approach to build individual models for each of the 25 partitions in the city allowing us to estimate the probability that a new Tweet issued from a given partition can be ranked by the probability that it was "generated" by the model. More concretely, given a set of locations L, and a Tweet T, our goal is to rank the locations by $P(L|T)$. Rather than estimate this directly, we use Bayesian inversion:

$$P(L|T) = \frac{P(T|\theta_L)P(L)}{P(T)} \tag{1}$$

where L is the model of the location. Assuming independence between terms:

$$P(T|\theta_L) = \prod_i P(t_i|\theta_L) \tag{2}$$

The probability of a term, given a location, $P(T_i|\theta_L)$, is estimated with Dirichlet smoothing [8]:

$$P(t|\theta_L) = \frac{c(t, L) + \mu P(t|\theta_C)}{|L| + \mu} \tag{3}$$

where μ is a parameter, set empirically, *c(t,L)* is the term frequency of a term t for partition L, $|L|$ is the number of terms in partition L. In this work we assume the prior probability of the partitions, *P(L)*, is distributed uniformly. We ignore $P(T)$, since it is the same for all partitions, and thus does not affect the ranking. partitions can be ranked directly by the probability of having generated the Tweet, or they can be ranked by comparing the model yielded by the Tweet, to the model of the partition, using Kullback-Leibler (KL) divergence. When ranking by KL divergence, we let θ_T be the language model for the Tweet T and L be the language model for the partition L. We use the Lemur Toolkit [2] for building our language models and carrying out our experiments.

Our aim is to detect geo-social events that result in unusual Twitter user behavior. For this, we define a socio-geographic boundary as under unusual conditions when its indicators, Number of Tweets (NT), Number of Users (NU) and Semantic Regularity (SR) satisfy the following function:

$$F = \alpha NT + \beta NU + \gamma SR \tag{4}$$

In function (4), F is a measure for the scale of an unusual event, α, β, and γ are coefficients for normalizing the measurements of each regularity. If the F is over a threshold, we predict that it is an indication that an unusual event is happening.

3. EXPERIMENTS

We crawled geo-tagged Twitter messages through the Twitter Streaming API. We setup a bounding box which covers the Dublin area and from 24/Jan/2013 to 19/Mar/2013 we crawled English-only Tweets with exact geo-locations attached. This yielded 387,800 Tweets in total from 14,533 unique users, each of which we mapped to one of our 25 city regions. To test how well our language model represents the consistency of partitioning, we compared our predicted locations for Tweets to actual locations. We used location accuracy (Acc), which calculates the percentage of correct predictions over all test examples and we obtained an Acc value of 0.3347. We also used Mean Reciprocal Rank (MRR), obtaining a figure of 0.4290. Based on our experimental results we find that with our identified city partitions, the language models generated from the contents created inside each of the partitions provide good consistency for defining the regularity of each partition.

4. USER TWEETING BEHAVIOUR ANALYSIS

We now focus on two aspects of users' Tweeting behaviour: geographic (where we Tweet) and temporal (when we Tweet).

4.1 Analysis of Geographical Behaviour

One would expect that people typically exhibit strong periodic behaviour in their movement as they move back and forth between their homes and workplace [3]. We observed this pattern in our users' Tweeting locations using the 25 partitions into which the Dublin city area was partitioned. We identified 5,875 unique users from our dataset who generated 95% of the overall Tweets, which reduced our total number of Tweets to 368,476 and we eliminated users who only generate 1 or 2 Tweets within a month as these are possibly visitors to the city.

We observed strong periodic behaviour in the distribution of locations from where Tweets were sent. In Table 1, we see that almost 44 % of users sent Tweets from only 1 or 2 of 25 different zones across the city during this one-month period. It is reasonable to assume that these locations are the users' homes, workplaces or leisure places. We also found that 23% of users generated Tweets across at least 5 seemingly random zones during this period and Tweets sent from these non-regular locations are of particular interest to our event detection task. If people only Tweet from their regular locations, their contents can be expected to be similar. Thus if we want to find irregular, unexpected event-related content, Tweets sent from non-regular locations should be of use.

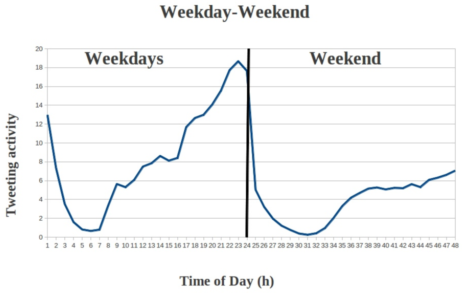
Table 1: User Tweeting in different zones

Number of zones	% of overall users
1	21.8%
2	22.7%
3	18.8%
4	13.7%
5-25	23%

4.2 Analysis of Temporal Dynamics

The volume of Tweets generated over time exhibits characteristics which potentially represent, in some way, each user's daily living patterns. Through studying temporal Tweeting behaviour, we can group users with similar daily life patterns. We aggregate the number of Tweets into hourly bins for each 24 hours, for weekdays and for weekends. Figure 3 shows trends from users' Tweeting patterns for weekdays and weekends in terms of the average number of Tweets generated per hour. Users are much less active during the weekend than weekdays, and the boost in volume starts much later in the weekend, 2pm as compared to 8am during weekdays.

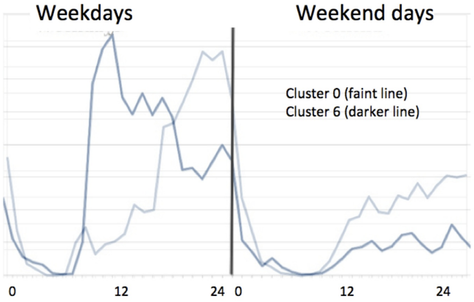
Figure 3: Overall Tweeting behaviour



We focused on 805 users who sent more than 100 Tweets in a month, and clustered these users by their temporal Tweeting features. For each user, there are 48 features, each representing the average number of Tweets per one-hour window for weekdays and for for weekend days. We used the EM algorithm clustering from WEKA to assign these 805 users to 10 clusters. Within each cluster we detect instances where users have noticeably unique characteristics in their temporal Tweeting patterns, as shown below.

Figure 4 shows the aggregated activities of two groups. Cluster 0 consists of very active users, 10 times more active than average in terms of hourly Tweeting volume and we consider these people as general Twitter users, who are just more active than others. By contrast, users in cluster 6 show completely different Tweeting patterns and we infer that these people are typical office workers, their Tweeting times peaking mostly during their lunch breaks, and after dinner, and they don't stay out late at night socialising.

Figure 4: Tweet distributions for Clusters 0 and 6



5. DISCUSSIONS AND FUTURE WORK

Unlike other areas of multimedia information retrieval, there are no standardised test collections of content, and limited standard tasks to execute on harvested Twitter content.

Event and Date	Time	GPS Coordinates	Related Twitter Content
Local flooding in Glencree Valley Jan 25, 2013	16:45:10	53.1809595,-6.1887448	The flooding around #Glencreevalley #Enniskerry is crazy! Watch out drivers! #Aaroadwatch
	16:50:08	53.182842,-6.191808	my car is like a floating boat #Enniskerry #flooding
Car crash on O'Connell Street caused by heavy rain, Jan 25, 2013	17:28:32	53.1809595,-6.1887448	@aaroadwatch bus and car collision on o'Connell street sb
	17:30:32	53.348604,-6.2597	@RobbieH46 slowly....it's a fecking car crash!!!!
	17:30:50	53.347887,-6.259207	Poor man or women in car crash.. #sayapray dangerous driving in this weather #5wordweather @spin1038
Heavy traffic jam Blanchardstown, Mar 09, 2013	17:17:11	53.3948484,-6.3912147	massive traffic jam in blanch won't be home till Christmas
	17:21:49	53.394718,-6.389326	traffic freaks me out!!!
	17:05:01	53.393323,-6.393317	Caught in a traffic jam
Pipe burst, cut off water supply Clongriffin Jan 07, 2013	14:22:16	8. 53.404341,-6.158719	@DonnieWahlberg its raining we have no water because of a burst pipe I am bogged down in housework but I am happy and having fun anyway :-)
	22:32:06	53.2853,-6.22825	@seanm91 apparently while attempting to fix the water pipe they damaged the gas line #incompetence

Table 2: Examples of Detected Real-time Events

For event detection on a city-wide or national scale, like Presidential elections, international sports matches, major concerts or other major social occasions, there is a groundtruth against which event detections can be compared. But who knows if there really was slow traffic on the M50 near the Blanchardstown exit on the morning of 5th March 2013. Instead we point to anecdotal examples of four local events which occurred and were detected by our method and which are shown in Table 2.

6. CONCLUSIONS

In this paper, we examined a way to comprehend the dynamics of small, local areas within a city through social media based on consistencies across Twitter users' behaviour, covering location, time and content which does not form part of a language model for each of our 25 regions. We ran a series of experiments which showed consistency across these and we demonstrated detecting events at a local level.

An algorithm for detecting local events in real time based on location, time, and content, of Tweets has not been presented before and our method provides good classification performance at a local, almost parochial level. Although event detection from social media, especially Twitter, has been studied for some time there are still many challenges, especially for processing information at a fine-grained local level and we believe that such information, when relayed or forwarded (re-Tweeted) automatically to the right person, will be of use. Our next challenge is detecting the Twitter users to notify about such locally detected events.

7. REFERENCES

[1] F. Abel, C. Hauff, G.-J. Houben, R. Stronkman, and K. Tao. Twitcident: fighting fire with information from social web streams. In *Proceedings of the 21st international conference companion on World Wide Web*, pages 305–308. ACM, 2012.

[2] J. Allan, J. Callan, K. Collins-Thompson, B. Croft, et al. The LEMUR toolkit for language modeling and information retrieval. *The Lemur Project. http://lemurproject. org (accessed 25 January 2012)*, 2003.

[3] N. Eagle and A. S. Pentland. Eigenbehaviors: Identifying structure in routine. *Behavioral Ecology and Sociobiology*, 63(7):1057–1066, 2009.

[4] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World Wide Web*, pages 851–860. ACM, 2010.

[5] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. Twitterstand: news in tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 42–51. ACM, 2009.

[6] H. Sayyadi, M. Hurst, and A. Maykov. Event Detection and Tracking in Social Streams. In *The International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2009.

[7] Y. Yang, T. Pierce, and J. Carbonell. A study of retrospective and on-line event detection. In *Proc. 21st annual international ACM SIGIR conference on Research and Development in information retrieval*, pages 28–36. ACM, 1998.

[8] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proc. 24th annual international ACM SIGIR conference on Research and Development in information retrieval*, pages 334–342. ACM, 2001.

Modeling the Web of Things from an IR approach

Héctor Cristyan Manta Caro
Faculty of Engineering, District University of
Bogotá Francisco José de Caldas, 110311,
Colombia
hcmantac@udistrital.edu.co

Juan M. Fernández-Luna
Department of Computer Science and AI,
CITIC-UGR, University of Granada, 18071,
Spain
jmfluna@decsai.ugr.es

ABSTRACT

Internet and Web technologies have evolved remarkably from their conceptualization. Nowadays, the origin of two novel paradigms have been triggered by the possibility of interconnecting not only traditional devices, smart phones, and wearable computing but also any object in the real world, and publishing Web-based services with dynamic content and data in real time. They are called the *Internet and the Web of Things*, respectively. The emergence of such paradigms implies a redefinition of the systems which they interact with, such as Information Retrieval systems. Thereby, it is essential to develop abstract models of Web representation, and simulation in order to establish new approaches in Information Retrieval for the Web of Things. A proposal for modeling the Web of Things based on a structured XML representation is described in this paper. This model has been designed with flexibility and modularity to allow the representation of multiple scenarios, being the conceptual source for future IR Systems development.

Categories and Subject Descriptors

H.1 [**Models and Principles**]: Miscellaneous;
H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Data Sharing, Web-based services*

Keywords

e-Things, Information Retrieval, Internet of Things, Search Engines, Sensor Web, Web of Things, Wisdom Web of Things (W2T)

1. INTRODUCTION

Nowadays, only 1% of real-world objects are connected to the Internet, but the expected number of connected devices is 50 to 100 billion by 2020 [6]. This new paradigm is referred to as the *Internet of Things (IoT)*, which describes technologies and research disciplines that enable the Net to adopt some intelligence and to venture into the real world of physical objects that are interconnected [4]. In addition, if we enable advanced Web access through virtual elements, which are abstract representation of things in the real world, we can create intelligent spaces which appears as the new paradigm called the *Web of Things (WoT)* [1]. At present, these paradigms bring new perspectives and challenges to systems interacting with them. In the context of Information Retrieval (IR), the new paradigms introduce dynamic factors to consider in detail: the WoT will abstract a huge amount of objects in the real world continuously producing a vast amount of information. Then its Web representation will incorporate status information or critical variables of interest that must be updated in real time and with frequent state changes, leading to highly dynamic and very large information sources.

This strong dynamism has not been well explored or evaluated in conventional retrieval systems. One of the most important open research topics, is the design, development, implementation and adaptation of real-time search engines that allow finding things, and information on variables of these things, as well as the features and services provided by them. This paper presents an approach for modeling the WoT as a basis for subsequent research in the development of new IR systems dealing with the dynamism and amount of ”documents”. However, for this purpose it is necessary to firstly develop an abstract model and a structured Web representation. In the next section, we present a state of the art in searching on the Internet of Things, and the Web of Things. The proposals are compared from different perspectives, conceptualizations and ordered according to sophistication. From simple search of embedded information on micro-devices attached to objects to elaborated search of things on semantic Web. We describe our model of the Web of Things from an IR perspective in Section 3, and our proposal of a structured WoT representation. Finally, Section 4 sums up and concludes.

2. SEARCHING ON THE WEB OF THINGS

The search for real-world entities (people, places, tangible and intangible things) will become an important and crucial service similar to current Web content search for media and documents [9]. There is also an increase in the relevance and worth of the information captured by sensors, the state, the properties, capabilities, functions and services that things may provide in the real world through their Web abstraction. The use of IR systems on the Web of Things is an issue of considerable complexity that imposes large demands on the design as the information is highly dynamic, inherently

ECIR’14 Information Access in Smart Cities Workshop (i-ASC 2014).
April 13, 2014, Amsterdam, the Netherlands..

distributed and with potentially colossal number of expected interconnected things[3]. In the rest of this section, several existing approaches are discussed, based on their area scale: i) only information found on the device connected to the thing, ii) things on a personal, iii) local, iv) metropolitan, as smart cities, or v) global area. Another distinction can be made by the range or possible results of search: sensor data, only sensors, sensor groups, things or physical objects based on constraints, or smart spaces.

Regarding IoT, Gander [2] presents a conceptual model of search engine for personalized networked spaces based on sampling of expressiveness and protocols responsive of space and time, focusing on the design of queries and data models resolved *in-situ*. The main contribution is its focus on the here and now, putting into consideration the high dynamism of information classifying the data as ephemeral, but with a low scalability factor. A similar hybrid approach is proposed in [7], where its main contribution is the novel and robust architecture that takes into account the dynamic collection and content. In this proposal, there is no Web representation of entities or a formal abstract model for the WoT.

The creation of a Web infrastructure to facilitate the integration, search, and interaction with smart things is presented in [10]. The proposal treats the location of a smart thing as the main property, structured hierarchically according to location identifiers. Searching space is larger than a personal smart space, considering the entire Web infrastructure. Another important contribution is the spatial hierarchy of things and their changing spatial association. The search engines return results at level of things, services or interfaces. However, the sensor level is not considered, nor are the dynamic changes on Web content. A line of approaches, we classified as State-based Searching of Entities on the WoT, propose the retrieval of things, which are in a particular state at the query time. For example, [9] introduces a method of ranking sensors formally modeled as random variables. A key contribution is the fact that each entity in the system is represented by a virtual counterpart with its own URL. The query language is not only based on keywords but also on the dynamic properties of entities captured by the associated sensors. The problem of searching for entities with dynamic content in real-time is addressed by different design dimensions.

On the other hand, other contributions introduce Web ontologies or semantic enrichment, mostly aimed at searching only at the level of a thing. For example, [8] presents a discovery system of RFID objects in smart spaces using a domain composed of two ontologies: one of general knowledge and another of specific domain user knowledge. In relation, [1] introduces a process and tools that allow users or applications to find connected objects that match a set of requirements and expectations. This work is based on the creation and use of semantic profiles of the connected objects, the establishment of similarities between the profiles to gather objects in groups, and a way of calculating a ranking for associating the context to incoming queries, also allowing for the selection of the most appropriate search algorithms. The limitation of this approach is that the possible outcomes are aimed at things, regardless of the sensor level, or spatial extent. In [5] Guinard introduces a vision and architecture of the Semantic WoT based on vocabularies, an abstraction of things and their high-level state, and semi-automatic sensor description generation. Searching of sensors and things is based on the high-level state of them. Both sensor and entities have well-defined semantic representations and considerations for their retrieval. The contribution of this work is its integration proposal based on Linked Sensor Data (RDF dataset of US weather stations sensor data), as well as the inclusion of semantics to the architecture of the WoT. However, the Semantic Web representation described adds complexity to the model, and the spatial-temporal context is not clearly defined in terms of the mobility of things and belonging to a space.

3. ABSTRACT MODEL OF THE WOT

We propose a WoT model, which begins with the abstraction of the real world, which mainly consists of two elements: things (tangible or intangible type), and spaces in which these things are contained or have a certain relationship. The model of the IoT involves the physical infrastructure that interconnects these two elements of the real world with the Net. Thus, there is a sensor layer in order to obtain real-time information on the properties of things in the real world, also on spaces. This information from multiple sensors is added to a nano or micro electronic device, called a data node. This by itself can get connectivity to other data nodes, or through gateway nodes, which would perform protocol conversion functions or provide connection to the Internet (see Fig 3). Our model of the WoT is comprised of five levels of abstraction involving the entire universe of elements that we consider relevant. Compared to other proposals, our model achieves completeness and balance by considering the spatial context on three levels, together with formal models of a virtual sensor and a virtual thing.

3.1 Description of the WoT Model

Given that the main and final motivation to model the WoT is publishing and establishing the information of smart things to Web services as a basis for developing future IR systems, our work proposes the representation of the WoT based on three main components of abstraction: Virtual Sensor, Virtual Thing, and Smart Space, and more important their corresponding hierarchical relationship. In contrast to [2], where there is neither a formal abstract model nor a hierarchy of elements in the proposals of searching on personalized networked spaces. However, to provide flexibility to the representation model, we add two additional spatial components: the possibility that a space is formed by sub-spaces, and further by sub-sub-spaces, and the ability to federate smart spaces in so-called Intelligent Zones. The proposed representation simplifies the lower layers, allowing to focus on a greater extend of the application and composition layers of the WoT. Therefore, the sensor node is associated in the model of WoT with an abstraction called Virtual Sensor, whose function is to allow its Web representation, composition features and high-level information to be viewed from the data collected by sensor nodes, that performs aggregation functions or information fusion. We propose to balance the model taking into account the spatial context and also creating a web abstraction of the sensor level. Some unbalanced models exist like [10] in the spatial search for smart things, where the model does not include a sensor level although does so with a spatial-oriented model. Each sensor as an abstraction model element has a URI that identifies its dynamic XML document, containing its description, properties and data. Things in the real world

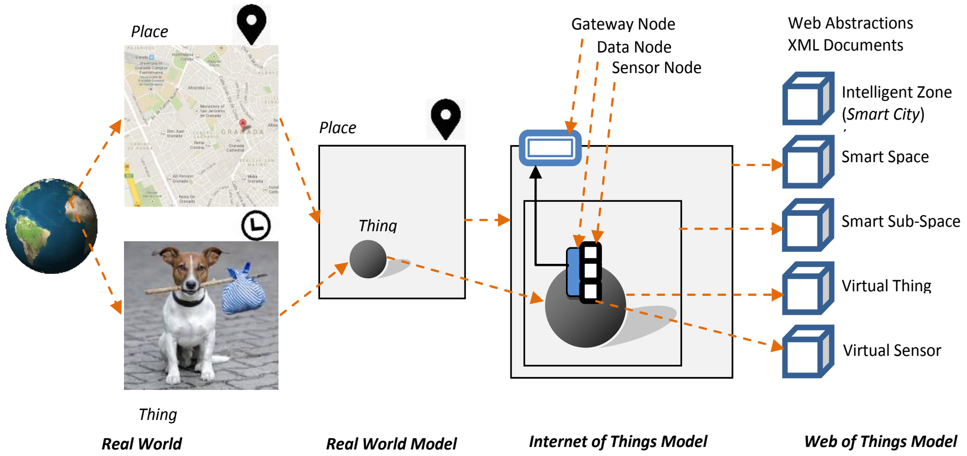


Figure 1: Proposed models of real world, IoT, WoT and their relationships

of tangible and intangible nature are modeled by the Web abstract component called Virtual Thing. Virtual things not only consolidate the information available at the virtual sensors linked to them, but also contain features, functionality or services that things through their Web abstractions can provide. Similarly, it has been decided that each element of the model, in this case the Virtual Thing, is uniquely identified by a URI, related to a dynamic XML document containing the real time information.

Virtual things, like their real counterparts, are confined in smart spaces that correspond to abstractions of places and sites of the real world that have been endowed with intelligence. Virtual things through their virtual location sensors have the potential to change not only their state, but also the smart space where they are. Thus, there will be a change in links between documents, and belonging to a place. Environments, sites and places in the real world are modeled using an abstract component called Smart Space, which condenses the characteristics of the environment in which they are located, bordering one or more virtual things. In comparison with other models, [9] proposes for state-based search of entities to have a stochastic sensor model and space considerations where the possibility of search results are limited to the level of entities, where information of sensors is given no relevance, and the constrained search for spaces are not contemplated, in our models the possible results can be in all the levels: data, sensors, things, and spaces. We propose to extend the modeling to allow a wider range, to enable searching also smart spaces that meet certain restrictions and/or contain certain things or types of sensors or data/states.

For example, a pet can be modeled as a virtual thing, which is contained somewhere within a smart city, in our model an Intelligent Zone. As illustrated in Fig. 1: the Dog is connected to Internet by means of an IoT infrastructure (Data and Gateway Nodes). Additionally, real-time information as Dog's location can be collected with the sensor node and published to the Web abstraction for further access. The model includes the possibility that a smart space can be composed of one or more sub-smart spaces, so the place where the Dog is, establish a sub-space as well as any other particular location for example a free car parking closest to the Dog. A Real-Time Search Engine for the WoT can be resolved the query to found the closest free car parking to the current Dog's location. In the search for entities in the semantic web, some proposals does not considered the sensor level nor the spatial context [8], [1], however [5] has a balanced model that includes both levels so similar to ours, even though the spatial context is not clearly defined. These works explore the semantic enrichment by means of ontologies. Our proposed Web representation model is presented in the next section. Our main motivation is feeding an IR system with a collection of dynamic documents originating from the abstract model, with the aim of retrieving the relevant ones, given a query. The representation of elements of the model can be as simple as using the associated metadata of the physical objects, or as complex as using ontologies and semantic profiles, however following the philosophy of the WoT, the reuse of web technologies, our proposal is to employ XML as a simple, structured, semantically enriched vehicle, containing the information of the items that can be retrieved in the model and, secondly, to consider advances based on XML sensor networks.

3.2 Structured WoT Representation

Each of the components proposed in each of the model layers corresponds to a XML schema, that together define a dynamic collection of XML documents for the WoT, as illustrated in Fig. 2.

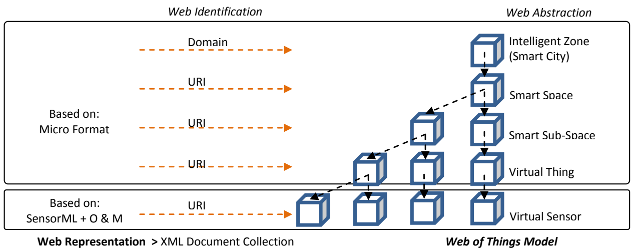


Figure 2: Dynamic collection of XML documents on the Web of Thing

The proposed virtual sensor XSD schema is composed of a group of general information tags, which contains elements of keywords, identification and classification. A group of references with contact elements, with the possibility of specifying and elements of role-based documentation specified in the XML implementation of model OM Observations and Measurements, Modeling Language SensorML. The next section of the XML schema contains the group of properties allows the characterization of the virtual sensor, the description of the sensor capabilities, the high level state, along with a membership element to associate the virtual sensor to the virtual thing sensed. Also, there is a history and events field, and Observing elements of the virtual sensor XML schema, are based scheme O&M of OpenGIS, with an element of sampling time, the time of the result, the feature of interest and the result. The Virtual Thing component should capture the information of the observed physical phenomenon, so it is proposed to use elements of the SWE scheme. The proposed scheme follows the structure of the XML representation of the virtual sensor using groups and elements: general info, references, properties and history. The property group has been enriched with an element of availability of the object, and the field of membership is associated with higher hierarchical level, indicating the smart space where the virtual object. An element location and a list of attached sensors are added. The XML schema of the Smart Space component is built based on microdata Place scheme in schema.org. This contains the same elements of the above components with a list of: virtual things, and subspaces. The Intelligent Zone follows a similar scheme with Web domain identification, and the list of smart spaces that compose it.

4. CONCLUSIONS

The WoT imposes a different dynamic to consider in the design and development of applications and systems for IR, given the change in the location of things in the physical world, the change in the collection of documents on behalf of inserting new sensors or recording new stuff, or depletion of life-time connectivity and thus removal of these documents, and in the same update real-time data collected by sensors. Most models of the WoT located in the first layer to the sensors, and from there a series of overlapping layers, according to the vision and purpose of each investigation. As core layer of most models is a abstraction layer of entities or things in the real world, with multiple alternatives for description and representation of both: non-Web, or using Web technologies like metadata, microformats, microdata, or ontologies. Given the flexibility, simplicity and use of XML standards has been selected for the construction of representation schemes for the WoT. The abstract model of the WoT, and the proposed dynamic representation take elements, and oriented considering the efforts of organizations such as the OpenGIS, W3C, ISO to standardize technologies that point to the interconnection of the real world. Our proposed model of the WoT, consider a real-world view that gives importance to the spatial context, adding relations between things and spaces. In addition the temporal context is added via elements of historical events. Future work is planned to study different conventional and semi-structured indexing information methods, focusing on XML, and real-time to assess their suitability for the WoT. The model is being used to build a Web of Things discrete event simulation with XSD schemas as inputs and to marshal an extend collection of XML documents.

Acknowledgments. This work has been funded by the Spanish Ministry of Ciencia e Innovación, and the Regional Government of Andalusia through the TIN2011-28538-C02-02 and TIC-04526 projects, respectively. Special thanks to the Iberoamerican University Association for Postgraduate AUIP, for their support to the implementation, progress and development of this work.

5. REFERENCES

[1] C. Benoit, V. Verdot, and V. Toubiana. Searching the web of things. In *Fifth IEEE International Conference on Semantic Computing Proceedings*, pages 1–8. IEEE, September 2011.

[2] Z. Ding, X. G. J. Dai, and Q. Yang. A hybrid search engine framework for the internet of things. In *Ninth Web Information Systems and Applications Conference Proceedings*, pages 57–60. IEEE, November 2012.

[3] B. M. Elahi, K. Romer, B. Ostermaier, M. Fahrmair, and W. Kellerer. Sensor ranking: A primitive for efficient content-based sensor search. In *International Conference on Information Processing in Sensor Networks Proceedings*, pages 217–228. IEEE, April 2009.

[4] M. A. Feki, F. Kawsara, M. Boussard, and L. Trappeniers. The internet of things: The next tehnological revolution. *Computer*, 46(2):24–25, February 2013.

[5] D. Guinard. *A Web of Things Application Architecture - Integrating the Real-World into the Web.* TH EidgenÃűssische Technische Hochschule ZÃijrich, Zurich, Switzerland, 2011.

[6] S. Hodges, S. Taylor, N. Villar, J. Scott, D. Bial, and P. Fischer. Prototyping connected devices for the internet of things. *Computer*, 46(2):26–84, February 2013.

[7] S. Mayer, D. Guinard, and V. Trifa. Searching in a web-based infrastructure for smart things. In *3rd International Conference on the Internet of Things Proceedings*, pages 119–126. IEEE, October 2012.

[8] D. Pfisterer, K. Romer, D. Bimschas, and et. al. Spitfire: Toward a semantic web of things. *IEEE Communications Magazine*, 49(11):40–48, November 2011.

[9] K. Romer, B. Ostermaier, F. Mattern, M. Fahrmair, and W. Kellerer. Real-time search for real-world entities: A survey. *Proceedings of the IEEE*, 98(11):1887–1902, November 2010.

[10] M. Ruta, T. D. Noia, E. D. Sciascio, F. Scioscia, and E. Tinelli. A ubiquitous knowledge-based system to enable rfid object discovery in smart environments. In *2nd International Workshop on RFID Technology - Concepts, Applications, Challenges Proceedings*, pages 87–100. IWRT, June 2008.

Smart Cities, Smart Citizens and the case for the CitySDK

Keynote

Frank Kresin
Waag Society, Amsterdam, the Netherlands

Biography

Frank Kresin is Research Director at Waag Society, institute for Arts, Science and Technology, based in Amsterdam. Waag Society develops and researches creative technology for social innovation, putting people and their needs at the center. It involves artists, scientists and entrepreneurs in early stages to come up with truly useable systems and services. Frank's background is in Artificial Intelligence and film making, and his interest is in developing technology for societal goals. He was involved at the start of many international innovation programmes, amongst them Apps for Europe, City SDK, CineGrid Amsterdam and Code 4 Europe. Frank has spoken, written and lectured on Smart Citizens, Open Innovation, Open Data & Open Design, Users-as-Designers, Living Labs and Fablabs. He is a regular moderator at the PICNIC Festival, as well as at design and innovation workshops in the Netherlands and abroad.

ECIR'14 Information Access in Smart Cities Workshop (i-ASC 2014).
April 13, 2014, Amsterdam, the Netherlands.

The Influence of Indoor Spatial Context on User Information Behaviours

Yongli Ren[1], Martin Tomko[2], Kevin Ong[1], Yuntian Brian Bai[1], Mark Sanderson[1]
[1]School of Computer Science and Information Technology, RMIT University, Melbourne, Australia
[2]Department of Computing and Information Systems, the University of Melbourne, Melbourne, Australia
yongli.ren@rmit.edu.au, tomkom@unimelb.edu.au, kevin.ong@rmit.edu.au,
yuntianbrian.bai@rmit.edu.au, mark.sanderson@rmit.edu.au

ABSTRACT

Through analysing a large data set of Web logs collected at a shopping mall, this study shows that the indoor spatial context significantly influences the information contents users search for and access on the Web. Specifically, this study shows that (1) at different locations of a large-scale indoor retail space, users tend to access different kinds of Web pages; (2) at indoor locations with similar context, users tend to request similar Web pages. These findings support a range of research questions in the context of information behaviour research, from a fresh understanding of mobile Web usage in indoor spaces to new applications of mobile surfing that matches users' dynamic indoor spatial context.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Experimentation, Measurement

Keywords

Indoor spatial context, indoor information behaviours

1. INTRODUCTION

Visiting large-scale indoor spaces, such as shopping malls, airports, and museums, has become a pervasive part of modern life. For example, Algethami describes a mall in Dubai, which attracted 75 million visitors in 2013 [1]. The Palace Museum in Beijing attracts approximately 12 million visitors each year [9]. All such buildings are designed to serve particular purposes: shopping malls, for example, are more than just a collection of retail stores [8].

One aspect of such spaces that does not appear to have been studied on a large scale, is to what extent does the context of indoor location affect the information users need? We hypothesize that users information needs can be identified based on their Web activity and consequently, in this paper, we investigate the following research question:

Does the spatial context of a structured indoor space implicitly influence a user's information behaviours on the Web?

By analysing an anonymised data set containing 18 million Web accesses from 12 thousand users collected over a 1 year period, it is found that the users' Web information behaviour significantly changes with their indoor spatial context.

Specifically, users at different locations tend to access different Web content, while users at locations with similar spatial context tend to access similar content. To the best of our knowledge, this is the first research concerning the relationship between the context of physical indoor spaces and users' Web surfing behaviours conducted on a dataset of a significant size.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 describes the collected data and Section 4 presents our methodology. We analyse the dataset and describe the identified association between of indoor spatial context on user information behaviours in Section 5. Section 6 concludes the paper.

2. RELATED WORK

Previous research focussed on either indoor spaces or mobile Web searching/browsing information behaviour, but rarely in connection, investigating the influence of indoor spatial context on user Web behaviour. For example, Biczok et. al. analysed the users' indoor spatial mobility through MazeMap, a live indoor/outdoor positioning and navigation system [2]. They found strong logical ties between different locations in users' spatial mobility. Church and Smyth focused on the differences between mobile browsing and mobile searching, showing that browsing was more common than searching, though mobile searching was increasingly popular [5]. Their other work [4] analysed the intent behind mobile information needs through a diary study. Church and Oliver noticed the shift that users are using mobile internet in more stationary and familiar settings, and explored the popularity of mobile usage in different contexts [3]. Similar research focused on the popularity of mobile searching in different contexts [7].

However, all this previous work only analysed context in the general forms, e. g. "at home/work", "travelling abroad", "with friends/family", "in transit/commuting". In our work, we focussed on the influence of specific contexts on user information behaviours, rather than the popularity of mobile usage in different general contexts.

3. DATA ACQUISITION

In this paper, we study a dataset of Web accesses gathered from a publicly available Wi-Fi network at a large inner-city shopping mall. The mall has over 200 stores and is covered

ECIR'14 Information Access in Smart Cities Workshop (i-ASC 2014).
April 13, 2014, Amsterdam, the Netherlands.

Table 1: The statistics of the query log data

Feature	Value
Number of users:	120,548
Number of access point association:	907,084
Number of Web accesses:	18,088,018
Number of days covered:	406

Table 2: Sample shop categories

Category	Category
Women's Fashion	Men's Fashion
Fine Jewellery	Music/Videos/DVDs
Furniture/Floor Coverings	Hair & Beauty
Fruit & Vegetable	Groceries

by between 50-100 Wi-Fi access points. The stores belong to 34 shop categories as defined by the mall operator. The data was collected between September 2012 and October 2013.

To ensure user privacy, identifying information is not stored in the data set we use. Such identifying information gathered by the operator are hashed in a non-invertible way. Table 1 shows the statistics of the collected data.

The data includes user spatial behaviour and the user Web information behaviour. Specifically, the users' spatial behaviour is characterised by the following parameters (1) users' location in the mall defined through by the location of the Wi-Fi access point with which the user's mobile device is associated; (2) timestamp and duration of users' association with the access point; (3) a computed convex area served by the access point (computed as a Voronoi cell) and related to the physical stores within this area. The users' information behaviour is characterised by: (1) timestamp of the Web request. (2) what Web page is requested, as defined by the uniform resource locator (URL); (3) the location of the users at the time of the request.

4. METHODOLOGY

We explore the associations between users' physical spatial context and their information behaviours in a large-scale indoor space. We investigate such correlation by integrating the spatial context of access points in terms of shop categories and the user information behaviours of Web accesses through Wi-Fi access points. The shop categories were made available to us from the mall owners, and the Web page categories were generated through a public Webroot Content Classification Service (*Brightcloud*[1]). Some sample shop categories are shown in Table 2.

We then define the spatial indoor context for each access point as a vector of shop categories, and the users' information behaviours are defined as a vector of *Brightcloud* categories, as follows:

Definition 1. *The indoor context of access point a_i is defined as a vector of shop categories $\mathcal{C}_s$,*

$$\mathbf{E}_i = [e_{i1}, \cdots, e_{ik}, \cdots, e_{im}],$$

where $\mathcal{C}_s = \{c_s^1, \cdots, c_s^m\}$, e_{ik} is the number of shops, which are located in the Voronoi cells of a_i and belong to $c_s^k \in \mathcal{C}_s$. This vector can be computed for each access point through a spatial overlay operation between the Voronoi cells and the outline of shop footprints from the mall floor layout, although in this case it has been executed manually for quality control.

Definition 2. *The user information Behaviour at access point a_i is defined as a vector of Web page categories $\mathcal{C}_w$,*

$$\mathbf{B}_i = [b_{i1}, \cdots, b_{ik}, \cdots, b_{in}],$$

where $\mathcal{C}_w = \{c_w^1, \cdots, c_w^n\}$, b_{ik} is the average number of URLs, which users issued through a_i and which belong to $c_w^k \in \mathcal{C}_w$.

At the level of Wi-Fi access points, the influence of spatial context on users' information behaviours can be viewed as the correlation between $\mathbf{B}_i$ and $\mathbf{B}_j$ for every two access points. We use the Pearson Correlation Coefficient (PCC) to test this association, defined as follows:

$$r(\mathbf{B}_i, \mathbf{B}_j) = \frac{\sum_{c_w^k \in \mathcal{C}_w} (b_{ik} - \bar{b}_i)(b_{jk} - \bar{b}_j)}{\sqrt{\sum_{c_w^k \in \mathcal{C}_w} (b_{ik} - \bar{b}_i)^2 \sum_{c_w^k \in \mathcal{C}_w} (b_{jk} - \bar{b}_j)^2}}, \quad (1)$$

where $\mathcal{C}_w$ is the set of URL categories, $\bar{b}_i$ and $\bar{b}_j$ are the average numbers of issued URLs at a_i and a_j, respectively. Results are shown in Fig. 2 and Table 4, and detailed discussion are shown in the following section.

5. INDOOR SPATIAL CONTEXT & USER INFORMATION BEHAVIOURS

5.1 Basic Indoor Information Behaviours

We start by investigating common indoor information behaviour patterns by analysing the distribution of the URLs over URL categories. It is observed that around one fifth of URLs are associated with *Social Networking* (e.g., Facebook.com). *Content Delivery Networks* (e.g., akamaihd.net) and *Computer and Internet info* (e.g., apple.com) take roughly the same proportion, around 13%. *Search Engines* are the fourth most popular category at 11%, and followed by *Business and Economy* with 10.6%. However, the users' indoor information behaviours is different from general mobile surfing, as reported by Church and Smyth [4]. Specifically, they reported that there are only 3.2% data for *Email and Social Networking* in general mobile information needs, but this category is much more represented (at 23.1%) in our dataset study. It is possible that, either the indoor context leads to a different information behaviour, or that the information behaviour of mobile users has shifted since the publication of the study of Church and Smyth [4].

To show common information behaviours, we identify the commonality of URL categories by measuring its access entropy. For a URL category c_w, its access entropy $H(c_w)$ is defined as:

$$H(c_w) = -\sum_{v \in S(c_w)} p(v|c_w) \log p(v|c_w), \quad (2)$$

where $S(c_w)$ is the set of visits when users accessed URLs in category c_w, $p(v|c_w)$ is the percentage of accesses to c_w during a visit v out of all visits, and a visit is defined as a single device per day in the mall. A high access entropy $H(c_w)$ means that c_w is a common category among all users; a low entropy means a category is accessed by a sub-set of users.

[1] http://brightcloud.com/resourcecenter/categories.php

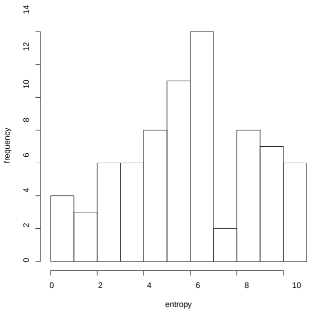


Figure 1: Distribution of $H(c_w)$

Table 3: Top-5 common URL categories and example URLs

Rank	category	example URL
1	Computer and Internet Info	*apple.com*
2	Social Networking	*facebook.com*
3	Search Engines	*google.com*
4	Business and Economy	*kakao.com*
5	Personal Storage	*icloud.com*

For example, *Computer and Internet Info*, *Social Networking* and *Search Engines* are very common URL categories with an entropy 10.75, 10.72 and 10.50, respectively. Table 3 lists the top-5 common categories based on their access entropy value and some corresponding example URLs. Fig. 1 shows the distribution of $H(c_w)$. It is observed that (1) there are some categories of websites that are more commonly visited than others, and (2) around 50% of the categories have an entropy smaller than 6.00. We will investigate this phenomenon of user information behaviour on these common and uncommon categories of websites in our future work.

5.2 Influence from Different Locations

There are differences in the types of shops served by different Wi-Fi access points. The collection of shop categories served by a single access point is what described our indoor context. We have hypothesized that the proximity of different types of shops will lead to a different Web information behaviour of the mall visitors. To test this hypothesis, we analyse the average PCC value r for every pair of access points, as defined in Eq. 1. The overall average of r reflects the general similarity of Web activity throughout the space. A small r indicates that different locations in the mall lead to different user information behaviours.

When using all URL categories, the average value of r is 0.9619, which seems to indicate that there is little difference between the information behaviours at different access points. In fact, the correlation is caused by the large proportion of common Web requests pointing to a small subset of URLs, of well defined categories. The top 5 common URL categories takes over 57.8% of the overall URL records and thus dominate the dataset. This significantly skewed Web behaviour introduces a bias in the r value.

Thus, we conduct another experiment to isolate the influence of these frequent Websites. We remove top common URL categories identified by Eq. 2 based on $p(v|c_w)$, and the r value is calculated by Eq. 1 based on $\mathbf{B}_i$. Thus, the calculation of r is independent from the identification of URL commonality. Namely, there is no logical influence between the calculation of r and the URL elimination based on $p(v|c_w)$. To show the influence of indoor location on user information behaviours, we calculate the r value by progressively eliminating common URL categories. Specifically, we select c_w based on its access entropy, $H(c_w)$ with a threshold, and we vary the threshold from $max(H(c_w))$ to 5 with a unit step[2]. Fig. 2 shows the r value over various thresholds. It is observed that when common URLs are removed from the calculation of r, differences in information behaviours at different access points appear. The more common URL categories we remove, the more substantial a difference we observe indicating that there is an influence from the local context of access points on user information behaviours.

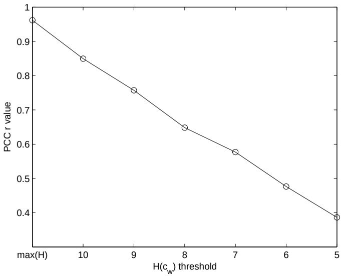


Figure 2: PCC r value without common c_w

5.3 Influence of Indoor Context

To show the influence of indoor context, we apply a clustering algorithm to group similar access points into clusters based on their indoor context. From *definition* 1, the surrounding indoor context information for an access point a_i is represented by a vector $\mathbf{E}_i$ of shop categories. We apply the k-means clustering algorithm to cluster $\mathcal{E}$ by treating each $\mathbf{E}_i \in \mathcal{E}$ as an instance. We set $k = 6$ because it achieves a relatively low value of the Davies-Bouldin index [6].

The k-means algorithm groups similar spatial contexts into clusters. If the users' information behaviour is influenced by their indoor context, the users' information behaviours *within* a cluster should be *similar* and the users' information behaviours *between* clusters should be *different.* To verify this association, we apply PCC, from Eq. 1, to measure the similarity between the information behaviours at two access points. The *intra-cluster* similarity (*within*) and the *inter-cluster* similarity (*between*) are defined as follows:

$$within = \frac{1}{k} \sum_{x=1}^{k} \left(\frac{2}{|t_x|(|t_x|-1)} \sum_{\mathbf{B}_i \in t_x} \sum_{\mathbf{B}_j \in t_x, i \neq j} r(\mathbf{B}_i, \mathbf{B}_j) \right), \tag{3}$$

where k is the number of clusters, t_x denotes the x-th cluster, and $|t_x|$ denotes the size of t_x.

$$between = \frac{1}{k} \sum_{x=1}^{k} \left(\frac{1}{|t_x|(|\mathcal{B}|-|t_x|)} \sum_{\mathbf{B}_i \in t_x} \sum_{\mathbf{B}_j \notin t_x} r(\mathbf{B}_i, \mathbf{B}_j) \right), \tag{4}$$

[2] When $H(c_w) \leqslant 4$, some $\mathbf{B}_i$ become empty, which renders the calculation of PCC r undefined. So, we analysed in the cases when $H(c_w) > 4$.

Table 4: Correlation of user information behaviours in groups of access points with similar spatial context

	$H(c_w)$	PCC r value based on $\mathcal{B}$				
		k-means		random		average
		within	between	within	between	
Groups of Access Point based on $\mathcal{E}$	$H(c_w) \leqslant max(H(c_w))$	**0.9659**	0.9623	0.9609	0.9617	0.9619
	$H(c_w) \leqslant 10$	**0.8601**	0.8509	0.8493	0.8501	0.8498
	$H(c_w) \leqslant 9$	**0.7721**	0.7599	0.7564	0.7573	0.7573
	$H(c_w) \leqslant 8$	**0.6817**	0.6572	0.6493	0.6473	0.6483
	$H(c_w) \leqslant 7$	**0.6410**	0.5966	0.5767	0.5750	0.5770
	$H(c_w) \leqslant 6$	**0.5045**	0.4778	0.4755	0.4751	0.4763
	$H(c_w) \leqslant 5$	**0.4107**	0.3942	0.3821	0.3848	0.3863

where $\mathcal{B}$ denotes the set of user information behaviours, and $|\mathcal{B}|$ denotes the size of $\mathcal{B}$. We emphasize that the groups of access points are clustered based on their physical context information $\mathcal{E}$, but the r value is defined based on user's information behaviours $\mathcal{B}$. Hence, the user's information behaviour is isolated from the clustering process.

We vary $H(c_w)$ from $max(H(c_w))$ to 5 with a unit step. We apply a *random* clustering method as a baseline to show the influence of indoor context[3]. The mean r for all $\mathbf{B}_i$ pairs is also applied as another baseline, and is defined as:

$$average = \frac{2}{|\mathcal{B}|(|\mathcal{B}|-1)} \sum_{\mathbf{B}_i} \sum_{\mathbf{B}_j, i \neq j} r(\mathbf{B}_i, \mathbf{B}_j). \qquad (5)$$

Table 4 shows the results of the experiment and Table 5 the results of the analysis, where a two-tailed, paired t-test is applied to evaluate whether the observed influence is significant or not. We observe: (1) the *within* of k-means is significantly larger than the *between* of k-means. (2) the *within* of k-means is significantly larger than the *within* of *random* method. (3) the *within* of k-means is significantly larger than the *average*. (4) the *within* of *random* is not significantly different from its *between* value. (5) the *within* of *random* is not significantly different from the *average*. As shown in the first row of Table 4, even when no common URL categories are removed, the *within* value of k-means 0.9659 is larger than the corresponding *between* value 0.9623, and is also larger than that of *random* 0.9609 and the *average* 0.9619.

Table 5: Paired t-test results

Methods	t	p-value
within(k-means) VS *between*(k-means)	3.7962	0.0090
within(k-means) VS *within*(*random*)	3.5871	0.0115
within(k-means) VS *average*	3.4126	0.0143
within(*random*) VS *between*(*random*)	0.2526	0.8090
within(*random*) VS *average*	1.6007	0.1606

The results show that the observed influence is statistically significant (see paired-t statistics in Table 5). This indicates that there is an influence from indoor spatial context on users' information behaviours.

6. CONCLUSION

Based on a large data set collected through the public Wi-Fi system of a large-scale shopping mall, we present an analysis of the influence of indoor spatial context on users' Web information behaviours in large-scale retail indoor spaces. We have found that the users' indoor information behaviour manifests a significant location-based bias when the baseline, common information behaviour is excluded. Furthermore, this location-based element captured by the indoor spatial context leads to similar information behaviours between indoor locations with similar contexts. In other words, users in similar indoor contexts tend to access similar categories of Web pages, while users in dissimilar indoor contexts tend to request dissimilar Web pages. This study has raised many new research questions: 1) what are the specific differences in user Web behaviours in two kinds of indoor contexts? 2) can the differences in information behaviours help identify and recommend different spatial indoor locations that can satisfy these needs? We leave the analysis of these possible implicit effects for future work.

[3] Both *random* and k-means are run 10 times, then averaged.

7. REFERENCES

[1] S. Algethami. Dubai Mall welcomes more than 200,000 shoppers a day. *Gulfnews*, 2014.

[2] G. Biczok, S. Martinez, T. Jelle, and J. Krogstie. Navigating MazeMap: indoor human mobility, spatio-logical ties and future potential. *CoRR*, arXiv:1401, 2014.

[3] K. Church, P. Ernest, and N. Oliver. Understanding Mobile Web and Mobile Search Use in Today's Dynamic Mobile Landscape. In *MobileHCI'11*, pages 67–76, 2011.

[4] K. Church and B. Smyth. Understanding the intent behind mobile information needs. *IUI*, pages 247–256, 2009.

[5] K. Church, B. Smyth, P. Cotter, and K. Bradley. Mobile information access: A study of emerging search behavior on the mobile Internet. *ACM TWEB*, 1(1), May 2007.

[6] D. L. Davies and D. W. Bouldin. A Cluster Separation Measure. *IEEE TPAMI*, 1(2):224–227, 1979.

[7] J. Teevan, A. Karlson, S. Amini, a. J. B. Brush, and J. Krumm. Understanding the importance of location, time, and people in mobile local search behavior. In *MobileHCI '11*, pages 77–80. ACM Press, 2011.

[8] J. D. Vernor, M. F. Amundson, J. A. Johnson, and J. S. Rabianski. *Shopping Center Appraisal and Analysis*. 2009.

[9] C. G. Wayne. A Better Space. *Smithsonian Magazine*, 2011.

On Mining Mobile Users by Monitoring Logs

Dmitry Namiot
Lomonosov Moscow State University
119991, GSP-1, 1-52, Leninskiye Gory
Moscow, Russia
+7-495-9392359
dnamiot@gmail.com

ABSTRACT

This paper considers a new model of data analysis for monitoring of mobile devices. Passive monitoring of mobile devices is based on ideas of network proximity and uses network protocol analysis for Wi-Fi and Bluetooth to gather presence information on mobile visitors. This is a direct analogue for web log and web site usage data, but we can deal with real visitors (with their mobile devices), rather than with abstract requests for web pages. In this paper, we propose a new model for processing of these data, which can detect some form of relationships between mobile users.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]

General Terms

Algorithms, Experimentation.

Keywords

Mobile monitoring, Wi-Fi, clustering, data mining.

1. INTRODUCTION

Monitoring the presence of mobile users (subscribers) is one of the most interesting and useful sources of information in smart cities data processing [1]. Monitoring of mobile users (in fact, it is monitoring for mobile devices) supplies data to evaluate the mobility of residents, planning transport routes, etc. [2]. On the lower level, we can talk, for example, about retail applications where analysis of the presence of mobile subscribers can be used to improve service, evaluation of marketing campaigns, planning, etc. [3]. At this moment, we can list well known and commonly used methods for determining the location of mobile devices based on the location of Wi-Fi access points [4]. Mobile operating systems (mobile applications) can use the information about the objects of the network infrastructure for verifying (or even determine) the true state of the subscriber. By analyzing the signal strength and visibility of access points we can build various metrics about the location of mobile devices (mobile users) [5, 6].

Passive Wi-Fi monitoring is one of the commonly approaches [7]. It lets anonymously collect data about mobile users (mobile devices) in proximity of some metering device [8]. This paper presents a new model for processing data collected during the passive monitoring for Wi-Fi (Bluetooth) devices. The rest of the paper is organized as follows. In Section 2 we describe the mobile monitoring and collected datasets. In Section 3 we describe exiting approaches for data processing as well as our data mining approach.

2. MOBILE MONITORING

Collecting traces of Wi-Fi beacons is the well-know approach for getting the locations of mobile devices. Beacon frames are used to announce the presence of a Wi-Fi network. As a result, an 802.11 client receives the beacons sent from all nearby access points. The client receives beacons even when it is not connected to any network. In fact, even when a client is connected to some particular Access Point (AP), it periodically scans Wi-Fi channels to receive beacons from other nearby APs [9]. It lets clients keep track of networks in its vicinity. But at the same time a Wi-Fi client periodically broadcasts an 802.11 probe request frame. The client expects to get back an appropriate probe request response from Wi-Fi access point. As per Wi-Fi spec, a station (client) sends a probe request frame when it needs to obtain the information from another station [10]. Figure 1 illustrates data flow for Probe Requests.

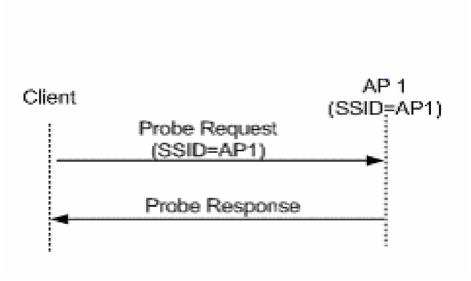


Figure 1. Wi-Fi Probe request/response

Technically, probe request frame contains the following information:

- source address (MAC-address)
- SSID
- supported rates
- additional request information
- extended support rates
- vendor specific information

Our metering device (it could be a Wi-Fi router, for example) can analyze received probe requests. Obviously, any new request (any new MAC-address) corresponds to a new wireless customer

ECIR'14 Information Access in Smart Cities Workshop (i-ASC 2014).
April 13, 2014, Amsterdam, the Netherlands.

nearby. Note, that Bluetooth devices could be monitored by the same principles.

Wi-Fi based device detection uses only a part of the above mentioned probe request. It is a device-unique address (MAC address). This unique information lets us re-identify the devices (mobile phones) across our monitors. The sequence of sequential requests (records) with the same MAC-address forms a session (similar to HTTP session in web applications).

Technically, data collected during passive Wi-Fi monitoring is similar to data collected in web statistics. Web statistics (web logs mining) are based on the standard format (formats) for log-files. The common standard is provided by W3C [11]. An extended log file contains a sequence of lines containing ASCII characters terminated by either the sequence LF or CRLF. Each line there corresponds to one request. Each line may contain either a directive or an entry [12]. The typical records contain the following fields: host address (IP address), user name, date, time, time zone information, URI (request), HTTP protocol version, status code, size of response in bytes.

For mobile monitoring, we can use the following fields: MAC-address for the device (hash-code for the privacy replacement), date, time, time zone info, signal strength (RSSI), name of the access point. The key missed point is the request (URI). It is obvious, that there are simply no requests for presence records. It is a key point, because many of exiting processing models can use URI data (e.g., for clustering).

Also, we should note, that passive Wi-Fi detection is not 100% reliable. Mobile phones (mobile OS, actually) can actually transmit probe requests at their discretion. Our own experiments with commercially available Wi-Fi probe scanners confirm data from [13]. The monitor detects in average about 70% of passing smartphones.

3. DATA PROCESSING

Technically, most of the monitoring systems for mobile devices treat collected data as some form of web log and provide appropriate statistics. The typical explanation of the existing systems is something like "Google Analytics for the real world" [14]. The typical analytical issue contains the number of visitors during the period, their timing, the number of unique visitors, the estimate for the number of regular visitors, etc. Figure 2 demonstrates a distribution of visits by type of mobile devices.

Extracting information from a Web log is fairly well-known research topic [14, 15], and consequently, the different software products. Usually, the study (analysis) can be classified into the following categories: content analysis, analysis of the structure and usage analysis. Analysis of the usage, in turn, may include personalization system, recommendations for modification sites, and business intelligence.

From the analysis of different patterns allocated when analyzing Web logs, we have identified one direction, which is almost not covered in this context. It is the mining of user groups. Actually, it is explainable for web statistics. Stable group of users for web access is several visitors browsed the site in parallel (approximately parallel) mode. This may make sense if we are talking, for example, about search bots. Yes, they can demonstrate sometimes correlation in time visit. For routine visits such grouping is rather artificial. Vice versa, for mobile monitoring, where each hit (each record in the log file) is some real visit, time based grouping makes sense.

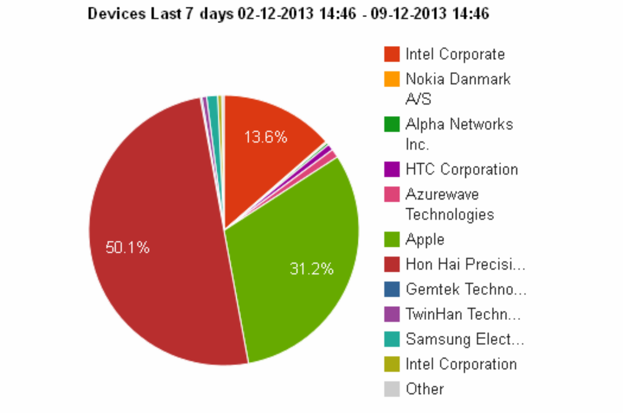


Figure 2. Mobile devices for visitors

There are some papers describing grouping for moving objects (for trajectories) [16, 17] Yes, it could be reproduced for proximity data too. For example, our paper [18] describes relationships mining for proximity data and models like Spotex [19]. But for such kind of tasks we need several metering devices. In this paper, we deal with the classical schema – one metering device and one log file.

For the typical web statistics, frequent visitors, for example, are IP addresses recorded (logged) every day for 7 (week) or 30 days (month). We want to extend this pattern to groups. Let us see a practical example. There is some group of friends, which occurs within a certain time in a cafe (co-working space, etc.). Not all of the members are present at each meeting, not all of them, as usually, arrive simultaneously (Figure 3). Can we discover such a group (groups) by proximity log?

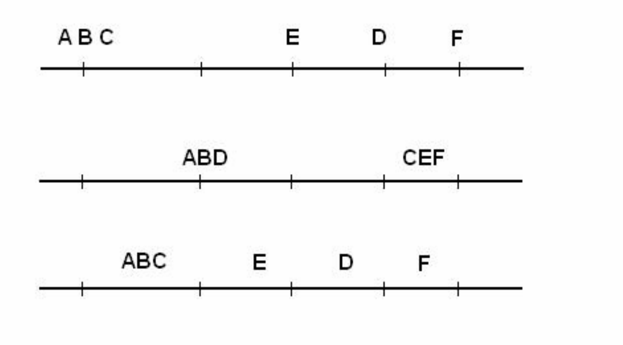


Figure 3. Visitors (A B C D E F) for 3 days.

One of the possible approaches for time-based analysis and events clustering is presented in [20]. It is based on the temporal similarity matrix. For two events *i* and *j* with timestamps t_i and t_j similarity for time interval *K* is:

$$S_k(i,j) = \exp\left(-\frac{\left|t_i - t_j\right|}{K}\right)$$

Authors present a method that first calculates the temporal similarity between all pairs of events (originally – photographs).

The calculated values are stored in a chronologically ordered matrix. And cluster boundaries are determined by calculating novelty scores for each set of similarity matrices. The authors assume that the events (in the original paper – photos) at cluster boundaries (in the original paper – event boundaries) separate two adjacent groups of events with high intra-class temporal similarity and low inter-class similarity.

In our research, we've followed to another approach. As it is mentioned in [21], time based clustering could be different from the traditional K-means clustering [22]. K-means clustering might find cluster centers with an idea to minimize some cost function. A traditional clustering algorithm (K-means might) find clusters and cluster–centers for the given K. As a basic point it uses the fact the cost function would change if those cluster–centers are moved. Cost function is the distance of data points to cluster centers. For our time stamped events we are not concerned with finding cluster centers at all. Really, the exact value for time any group is collected should be irrelevant. Our algorithm should only assign collected points to clusters and as long as the segmentation remains the same. We need segmentation or splitting of the time sequence. The key question is how to split our events in time, so that intra-cluster variance is reduced.

And our idea of mining groups is based on two sequential steps:

- find clusters for the each day
- detect the sequences of clusters across all days with some minimum set of common members

It is illustrated on Figure 4.

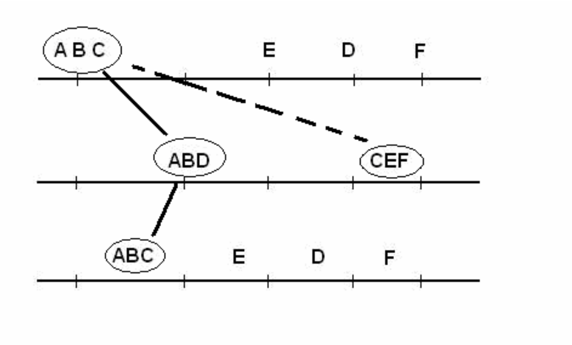


Figure 4. Clusters and groups

For getting clusters we followed to the algorithm originally developed for clustering photos [21]. It offers dynamic clusters with automatically detected boundaries (splitting points). It is based on two assumptions.

- Increased (long) time interval without registrations usually marks the end of cluster (all members of the group are in already). "Long" is defined as being either large relative to the extent of the cluster currently being examined (collected), or large relative to the average inter-group interval.
- Changed frequency of registrations corresponds to the start of a new group

As soon as clusters are detected, we can examine them for common elements (MAC-addresses), presented in the majority of per-day clusters. It means that group detection is always associated with predefined percentage of visits. E.g. visitors participated at least in 75% of meetings.

For this examination let us present each group as a string, where each element corresponds to the unique MAC-address (see Figure 4). Now we need to find a common subsequence of strings (groups) across all days (see solid line on Figure 4).

String C is a common subsequence of strings A and B if C is a subsequence of A and also a subsequence of B. String C is a longest common subsequence of string A and B if C is a common subsequence of A and B of maximal length. It means that there is no common subsequence of A and B that has greater length [23]. The typical algorithm for finding the longest common subsequence could be obtained from papers [23, 24].

The proposed system has been implemented in connection with Wi-Fi scanner from Libelium. During the testing stage we've successfully identified 8 groups from 11 (café in office building).

4. CONCLUSION

In this paper, we propose a new model for analyzing web-logs collected by the mobile phones monitoring systems. From the analysis of different patterns of web logs mining, we have identified one direction, which is almost not covered in this connection. It is the mining of user groups. In our paper, we propose two step algorithm for grouping mobile visitors. It could be used in Smart City projects as well as in retail information systems.

5. REFERENCES

[1] Murty, R., Gosain, A., Tierney, M., Brody, A., Fahad, A., Bers, J., & Welsh, M. (2008, May). CitySense: A vision for an urban-scale wireless networking testbed. In Proceedings of the 2008 IEEE International Conference on Technologies for Homeland Security, Waltham, MA.

[2] Cornelius, C., Kapadia, A., Kotz, D., Peebles, D., Shin, M., & Triandopoulos, N. (2008, June). Anonysense: privacy-aware people-centric sensing. In Proceedings of the 6th international conference on Mobile systems, applications, and services (pp. 211-224). ACM.

[3] Ryder, J., Longstaff, B., Reddy, S., & Estrin, D. (2009, August). Ambulation: A tool for monitoring mobility patterns over time using mobile phones. In Computational Science and Engineering, 2009. CSE'09. International Conference on (Vol. 4, pp. 927-931). IEEE.

[4] Namiot, D., and Sneps-Sneppe, M. (2012, April). Proximity as a service. In Future Internet Communications (BCFIC), 2012 2nd Baltic Congress on (pp. 199-205). IEEE. DOI: 10.1109/BCFIC.2012.6217947.

[5] Lassabe, F., Canalda, P., Chatonnay, P., & Spies, F. (2009). Indoor Wi-Fi positioning: techniques and systems. Annals of telecommunications-Annales des télécommunications, 64(9-10), 651-664.

[6] Zàruba, G. V., Huber, M., Kamangar, F. A., & Chlamtac, I. (2007). Indoor location tracking using RSSI readings from a single Wi-Fi access point. Wireless networks, 13(2), 221-235.

[7] Labiod, H., Afifi, H., & De Santis, C. (Eds.). (2007). Wi-Fi, Bluetooth, Zigbee and WiMAX. Springer.

[8] Namiot D. and Sneps-Sneppe M. Geofence and Network Proximity. In Internet of Things, Smart Spaces, and Next Generation Networking, Lecture Notes in Computer Science. Volume 8121, 2013, pp. 117-127, DOI: 10.1007/978-3-642-40316-3_11.

[9] Dmitry Namiot and Manfred Sneps-Sneppe. "Local messages for smartphones". Future Internet Communications (CFIC), 2013 Conference on (pp. 1-6). IEEE. DOI: 10.1109/CFIC.2013.6566322.

[10] M.Gast 802.11 Wireless Networks: The Definitive Guide O'Reilly Media, Inc., 2005, 654 p.

[11] Jansen, Bernard J., Amanda Spink, and Isak Taksai. Handbook of research on web log analysis. London: Information Science Reference, 2009.

[12] W3C log: http://www.w3.org/TR/WD-logfile.html Retrieved: Jan, 2014

[13] A. Musa and J.Eriksson, “Tracking Unmodified Smartphones Using Wi-Fi Monitors”, SenSys’12, November 6–9, 2012, Toronto.

[14] Yang, Q., Zhang, H. H., & Li, T. (2001, August). Mining web logs for prediction models in WWW caching and prefetching. In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 473-478). ACM.

[15] Grace, L. K., Maheswari, V., & Nagamalai, D. (2011). Analysis of web logs and web user in web mining. arXiv preprint arXiv:1101.5668.

[16] M. Andersson, J. Gudmundsson, P. Laube, and T. Wolle. Reporting leaders and followers among trajectories of moving point objects. GeoInformatica, 12(4):497–528, 2008.

[17] Li, Z., Ding, B., Wu, F., Lei, T. K. H., Kays, R., & Crofoot, M. C. (2013). Attraction and Avoidance Detection from Movements. Proceedings of the VLDB Endowment, 7(3).

[18] Namiot, D. (2013). Mining Relationships in Proximity Movements. Applied Mathematical Sciences, 7(144), 7173-7177.

[19] Namiot, D. (2012, September). Context-Aware Browsing--A Practical Approach. In Next Generation Mobile Applications, Services and Technologies (NGMAST), 2012 6th International Conference on (pp. 18-23). IEEE. DOI: 10.1109/NGMAST.2012.13

[20] Cooper, M., Foote, J., & Girgensohn, A. (2003, September). Automatically organizing digital photographs using time and content. In Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on (Vol. 3, pp. III-749). IEEE.

[21] Gargi, U. (2003). Consumer media capture: Time-based analysis and event clustering. HP-Labs Tech Report.

[22] Hartigan, John A., and Manchek A. Wong. "Algorithm AS 136: A k-means clustering algorithm." Journal of the Royal Statistical Society. Series C (Applied Statistics) 28.1 (1979): 100-108.

[23] Hirschberg, Daniel S. "Algorithms for the longest common subsequence problem." Journal of the ACM (JACM) 24.4 (1977): 664-675.

[24] Hunt, J. W., & Szymanski, T. G. (1977). A fast algorithm for computing longest common subsequences. Communications of the ACM, 20(5), 350-353.

Mining digital footprints for smart tourism

Invited contribution

Raffaele Perego
ISTI CNR, Pisa, Italy

Biography

Dr. Raffaele Perego is a senior researcher at ISTI, where he leads the High Performance Computing Lab carrying out research on algorithms and information systems addressing computational-, and data-intensive problems arising in scientific, business, social, and knowledge-based applications scenario. His main research interests include: Web information retrieval, data mining, machine learning, parallel and distributed computing. Raffaele Perego coauthored more than 120 papers on these topics published in journals and in the proceedings of peer reviewed international conferences.

ECIR'14 Information Access in Smart Cities Workshop (i-ASC 2014).
April 13, 2014, Amsterdam, the Netherlands.

Challenges in Recommending Venues within Smart Cities

Romain Deveaud, M-Dyaa Albakour, Craig Macdonald, and Iadh Ounis
University of Glasgow, UK
firstame.lastname@glasgow.ac.uk

ABSTRACT

Recommending venues to a user within a city is a task that has emerged recently with the growing interest in location-based information access. However, the current applications for this task only use the limited and private data gathered by Location-based Social Networks (LBSNs) such as Foursquare or Google Places. In this position paper, we discuss the research opportunities that can arise with the use of the digital infrastructure of a smart city, and how the venue recommendation applications can benefit from this infrastructure. We focus on the potential applications of social and physical sensors for improving the quality of the recommendations, and highlight the challenges in evaluating such recommendations.

Keywords

social and physical sensors, location-based social networks, contextual evaluation

1. INTRODUCTION

With the wide use of Internet-connected and sensor-enabled smartphones, users can now share their location while performing online tasks, such as searching for information. The collected location data allows to model the behaviour of the users, such as their travel preferences, thereby enabling the emergence of new tasks such as the recommendation of venues that might be of interest to the user. This task encompasses a wide variety of sub-tasks, ranging from the recommendation of venues that the user have never visited before [4] to the prediction of the next location of the user [3].

Currently, the existing venue recommendation systems heavily rely on data extracted from Location-based Social Networks (LBSNs) [7], such as Foursquare[1], Yelp[2], or Google Places[3]. In these LBSNs, users can broadcast their location to their friends (or to other users), and rate and comment on venues. In the literature, the standard approach for recommending venues is to identify users that are similar to the current user, and to recommend venues that these similar users have rated highly [8, 9].

Since the aforementioned approaches rely on the profiles of the users, which are composed of private data, the only existing applications, thus far, are industrial and are provided by the LBSNs. Such applications examine the history of the visits of the users, and recommend venues that might be of interest to them. However, the recommendations only rely on a few signals (i.e. the derived preferences of a user through the venues she/he visited before) to estimate the relevance of venues. The privacy of the user data also raises questions about the generalisation and the reusability of such venue recommendation approaches. We argue that the digital infrastructure provided by smart cities can overcome these problems. In this position paper, we discuss the new possibilities and the underlying challenges of using new types of data for recommending venues. In Section 2, we discuss the use of different signals and indicators found in smart cities deployments that might improve the representation of venues, while we focus in Section 3 on the problems related to the evaluation of such recommendations.

2. ON THE USE OF SENSORS

One of the main characteristic of smart cities is the abundance of sensors connected to the Internet of Things [5], which allow to gather information on what is currently happening in the city. While we may think primarily of physical sensors (e.g. CCTV cameras), we may also consider social sensors (e.g. Twitter [1]). Currently, only social sensors are used to perform venues recommendation. Indeed, the collaborative filtering recommendation methods that are typically used in the literature do not use other indicators than the profiles of the users. Moreover, the preferences inferred from these profiles cannot be shared amongst different LBSNs [4]. Combining data from different social sensors could improve the representation of venues, and eventually overcome the sparsity problem that can arise for venues that have few associated ratings and comments. Since some tweets are geolocated, it is possible to analyse the sentiments [6] expressed in the tweets that were emitted at the locations corresponding to the venues, and use these sentiments as a sensor of the quality of the venue.

In addition to a combination of different social sensors, using physical sensors allows to derive further information about the venues. For example, CCTV cameras and microphones, along with some audio/video processing, can be used to estimate crowd densities, thus helping to identify in real-time popular areas (including the corresponding venues) that might be of interest to a user. Such physical observations may also help to better model the context of the users by identifying their behaviours (e.g. a venue recommendation system should not suggest outdoor venues to someone who appears to be feeling cold while walking in the street). The GPS sensors on public transports can also help to detect traffic jams (and even predict them, with sufficient training

[1] http://foursquare.com/

[2] http://yelp.com

[3] http://maps.google.com/

ECIR'14 Information Access in Smart Cities Workshop (i-ASC 2014). April 13, 2014, Amsterdam, the Netherlands.

data), which might prevent users to promptly reach certain venues. Environment sensors (e.g. rain-related) can also help to determine areas that might not be suitable to recommend (e.g. outdoor venues), or that might be difficult to reach for certain types of users (e.g. elderly persons).

We argue that having rich representations of real-world entities, such as venues, is essential for providing high-quality and accurate recommendations to the users. While social sensors are essential sources of subjectivity (e.g. opinions/ratings about a venue), physical sensors can provide valuable additional objective indicators that can help to identify the context of the venue.

However, there is a need for new effective methods that can interpret all of the raw data extracted from these physical sensors, in order to generate useful information. There is also a need for agreed standards for storing this data (e.g. RDF). One can imagine that all sensors could feed a dedicated knowledge base of the smart city that holds all the records of the different entities of the city. Internet-connected sensors can then update the attributes of the entities in real-time. Such knowledge bases will need however to evolve from a static representation of the information to a time-aware representation, allowing to track the evolution of each entity's attributes and to eventually forecast them. Finally, it is of note that special considerations should be paid to the privacy and ethical issues arising from the storage of such data. Such issues still need to be explored so as to ensure an adequate balance between added-value and privacy.

3. EVALUATING RECOMMENDATIONS

Although a variety of new ideas can be imagined to improve the recommendation of venues, they need to be tested and properly evaluated. However, we face here again a challenge that is related to the nature of the task itself. Indeed, a venue recommendation is contextual to (among other parameters) the time of the day, the current location of the user, and her/his preferences. The TREC Contextual Suggestion track [2] explored such research questions, aiming to develop a test collection that can support the venue recommendation task. In its first year (2012), the track explored the relevance of recommendations with respect to a description of the venue generated by the systems, the geographical relevance, the adequacy of the website linked to the venue, and the temporal relevance. Several evaluation measures combining these aspects were also proposed. In its second year (2013) however, the time aspect was dropped, thus highlighting the difficulties in evaluating venue recommendations with respect to all the contextual parameters at stake. Moreover, other parameters could be considered. We raised the issue of privacy in the previous section; some users might want to pay the price of giving their personal information in order to benefit from highly relevant recommendations, while others might prefer to receive only "good" enough recommendations by sharing less personal information. Such evaluation frameworks will then need to take a wide range of parameters into account in order to accurately estimate the relevance of recommendations to users.

Building an entirely reusable test collection for evaluating venue recommendations is also a challenge. Indeed, we argued that the relevance of a venue recommendation is highly contextual and can change depending on a wide range of parameters, which may be difficult to reproduce in a controlled setting. While simplified settings are required to understand how the proposed systems perform, they do not necessarily reflect real use cases. Using a smartphone application, and analysing the feedback provided by the users, is a possible way to evaluate the quality of the recommendations. This feedback can be of two different types: explicit or implicit. An explicit feedback takes the form of a questionnaire, asking the user if the recommendation is interesting. This questionnaire can also be presented before and after the visit, in order to analyse if the experience actually changed the opinion of the user about the venue. An implicit feedback can be gathered from the GPS data of the smartphone: if users actually went to a venue that the application recommended, then they must have found the recommendation interesting given their context. The implicit evaluation techniques used by commercial search engines, such as A/B testing, could also be used as an implicit feedback.

4. CONCLUSION

The task of recommending venues to users is an emerging task that presents many challenges. In this position paper, we argued that this task can benefit from the digital infrastructure deployed by smart cities. More particularly, the use of physical sensors offers advantages that have currently remained unexplored. We have also argued that sensing the behaviours of users when they interact with their mobile devices is also a promising direction for accurately evaluating the quality of recommendations.

5. ACKNOWLEDGMENTS

This work has been carried out in the scope of the EC co-funded project SMART (FP7-287583).

6. REFERENCES

[1] M.-D. Albakour, C. Macdonald, and I. Ounis. Identifying Local Events by Using Microblogs As Social Sensors. In *Proc. of OAIR*, 2013.

[2] A. Dean-Hall, C. L. Clarke, J. Kamps, P. Thomas, N. Simone, and E. M. Voorhees. Overview of the TREC 2013 Contextual Suggestion track. In *Proc. of TREC*, 2013.

[3] A. Monreale, F. Pinelli, R. Trasarti, and F. Giannotti. WhereNext: A Location Predictor on Trajectory Pattern Mining. In *Proc. of KDD*, 2009.

[4] A. Noulas, S. Scellato, N. Lathia, and C. Mascolo. A Random Walk Around the City: New Venue Recommendation in Location-Based Social Networks. In *Proc. of SOCIALCOM-PASSAT*, 2012.

[5] J. Soldatos, M. Draief, C. Macdonald, and I. Ounis. Multimedia Search over Integrated Social and Sensor Networks. In *Proc. of WWW*, 2012.

[6] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas. Sentiment in Short Strength Detection Informal Text. *J. Am. Soc. Inf. Sci. Technol.*, 61(12):2544–2558, 2010.

[7] M. Ye, P. Yin, and W.-C. Lee. Location Recommendation for Location-based Social Networks. In *Proc. of SIGSPATIAL GIS*, 2010.

[8] M. Ye, P. Yin, W.-C. Lee, and D.-L. Lee. Exploiting Geographical Influence for Collaborative Point-of-interest Recommendation. In *Proc. of SIGIR*, 2011.

[9] Q. Yuan, G. Cong, Z. Ma, A. Sun, and N. Magnenat-Thalmann. Time-aware Point-of-interest Recommendation. In *Proc. of SIGIR*, 2013.

Smarter Cities, Safer Travels: Intergrating Contextual Suggestion

Adriel Dean-Hall
University of Waterloo
adeanhal@uwaterloo.ca

Jack Thomas
University of Waterloo
j26thoma@uwaterloo.ca

ABSTRACT

Contextual suggestion is meant to provide meaningful venue recommendations tailored to a personal profile, but sometimes a user's information need goes beyond their taste in restaurants. This paper seeks to demonstrate that integrating information concerning crime reports and public safety alerts into suggestions allows for more informed decision-making by users. We will explore existing sources of this public safety data, demonstrate its potential use and examine the implications – positive and negative – implied by the approach.

1. INTRODUCTION

A common issue for travelers visiting a new city is not knowing the lay of the land. Finding restaurants, shops and entertainment venues compatible with their interest is an obvious concern, but this can also be extended to the public safety geography. A street map is not enough to tell visitors which roads are well-lit at night or which neighborhoods are currently experiencing higher-than-average crime rates. This information can be just as important to a user's decision as any part of their personal preferences.

Where once tourists might read guidebooks to identify high-risk neighborhoods, the modern traveler uses websites such as Trip Advisor[1] or Virtual Tourist[2]. The internet brims with reviews and anecdotes about cities, as well as users looking to make informed decisions while abroad.

The development of smart cities provides an opportunity to meet this information need. Municipal governments collect vast reams of data concerning public safety, from crime statistics to accident reports to the status of infrastructure. As the paradigm shifts toward making this data open and accessible to users, contextual suggestion developers can draw on this information to improve the suggestions they make.

"Contextual suggestion" covers an ongoing field of research at the border of search and recommendation meant to provide personalized recommendations to users for "points of interest", such as venues and other attractions. The subject has garnered interest[2] for providing travelers with a convenient way to explore an unfamiliar city using their personal preferences as a guide.

These early forays have been marked by limited scope, concerned solely with the user's tastes. Now that smart cities are expanding the data horizon to include new subjects, the time has come to consider safety's potential as a recommendation feature.

2. RELATED WORK

Recommender systems for contextual suggestion have already been the topic of considerable research, such as Garcia et al.'s paper in 2011 [4]. The authors make great strides in producing a system which adapts its recommendations according to the tastes, demographics and history of the user. Baltrunas et al.'s paper from the same year[1], emphasizes the importance of the user's context when making recommendations, such as the season, available transport, who they're traveling with and so on.

However, these papers do not consider the safety of an attraction, nor is any data relevant to that conclusion provided for the user. This is not to say that the value of public safety data has not been explored. As early as 2001, Estivill-Castro et al.'s work[3] demonstrated how crime statistics, one part of public safety, can be mined and mapped.

Smart city projects have made great strides in building safety databases. In 2011, the city of St. Louis participated in IBM's Smarter Cities Challenge[3] in an effort to improve synergy between separate law enforcement agencies within the city. At IBM's recommendation, the city created a plan for a centralized crime database accessible by all agencies so that a unified view of all offenders and offenses could be made. The focus so far has been on enhancing communication within government, but these same data sources could be made available to the public.

[1] `http://www.tripadvisor.ca/Travel-g28970-s206/Washington-Dc:District-Of-Columbia:Health.And.Safety.html`

[2] `http://www.virtualtourist.com/travel/North_America/United_States_of_America/Washington_DC/Warnings_or_Dangers-Washington_DC-TG-C-1.html`

[3] `http://smartercitieschallenge.org/city_st_louis.html`

ECIR'14 Information Access in Smart Cities Workshop (i-ASC 2014).
April 13, 2014, Amsterdam, the Netherlands.

Some smart cities infrastructure is already being used by other applications and fields of study to enhance their results. Even the Van Gogh museum in Amsterdam has found uses for smart city data to modify recommendations made to tourists. Through the European Union's CitySDK project[4], sensors have been installed in the museum to detect the presence of lines and crowds[5], letting prospective visitors know when the museum is busy and how long the expected wait will be. This museum system is a textbook application for how smart cities data is already being used to enhance the user experience.

Between existing research into contextual suggestion, the potential of public safety data and the success of similar smart cities initiatives, integration seems a natural avenue to explore. In the next section we will begin this exploration by taking profiles and data drawn from the TREC 2011 track on contextual suggestion and integrating it with real, publicly available crime statistics.

3. DEMONSTRATION

The 2013 Contextual Suggestion TREC track gathered, from several systems, personalized point-of-interest suggestions for multiple users and cities. In this track each system made ranked suggestions to users and the users rating the suggestions based on their interest in visiting the attraction [2]. One of these cities, the one which we will focus on, is Washington D.C., in this section we will combine crime incident reports with the suggestions that the systems made for the area.

District of Columbia Data Catalog [5] contains crime reports for a 10 km radius around the Washington, D.C. area. This information is provided as a live feed as well as reports that contain all the crimes for a given year. For this paper we use the 2013 data, a full implementation will probably find it valuable to use the most up to date information.

Our strategy is to provide an indicator to users about how many crimes have occurred in the area around the attraction they are considering visiting. A simple first attempt at doing this would be to count how many crime incidents occurred within a certain distance of the attraction:

$$S(y) = |x \in X | D(x, y) < \Theta|. \tag{1}$$

Here X is our set of crimes, y is our attraction, D is the distance, in kilometers, between two locations calculated using the Haversine formula, and Θ is our threshold for how far the crime has to be for it to be considered. So, $S(y)$ is the safety score of the attraction where lower numbers mean the attraction has had fewer crime incidents nearby and we assume are safer.

However, we also want to consider how long ago the crime occurred because crimes that occurred more recently are more likely to be an indication of safety, we can incorporate this into our equation:

TREC Rank	Title	Score
1	Zaytinya	1104.86
2	Bistrot Du Coin	610.58
3	Old Ebbitt Grill	406.15
4	Blue Duck Tavern	353.16
5	Founding Farmers	484.29

Table 1: Safety score for suggestions.

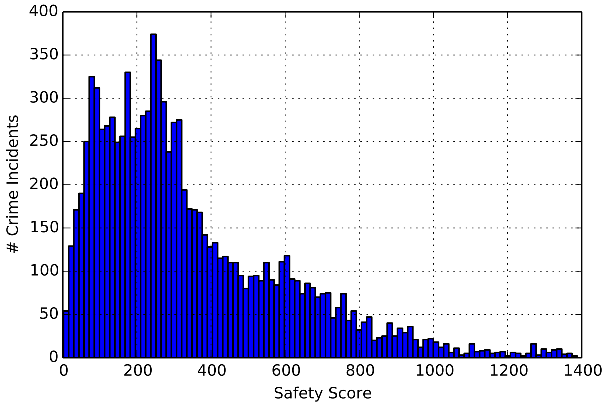


Figure 1: Safety score (equation 2) of locations where a crime occured.

$$S(y) = \sum_{x \in X \,|\, D(x,y) < \Theta} \frac{1}{\log(1 + \epsilon + T(x))}. \tag{2}$$

Here T is how many days have passed since the crime occurred divided by 365, so crimes that occured more recently are weighted higher. This gives us a safety score that warns users more heavily about crimes that occurred more recently. In this equation $\epsilon = 1.0$ and prevents division by zero. For this paper we pick a distance threshold $\Theta = 0.5$. The reason we choose this equation is that it gives more weight to crimes that have occurred more recently.

As an example, we can pick suggestions made by one of the systems (UDInfoCS1) for one of the profiles (554) as part of the contextual suggestion track. Here we we equation 2 to calculate the safety score of the top 5 suggestions, as ranked by the TREC system, for this user. These scores can be seen in table 1. In order to calculate the score we choose December 5th, 2013 as the date that the user is searching.

Now that we have a safety score we need to get context as to what this number means in terms of relative safety in the city. In order to do this we calculate the safety score at the location of every crime incident. Figure 1 shows the frequency of the safety scores given to locations where crimes have occurred in Washington, D.C.

In order to provide users with an easily digestible safety score we map the score given in equation 2 to 1-5 depending upon if the score is below the 20% percentile, between the 20% and 40% percentiles, between the 40% and 60% per-

[4] `http://www.citysdk.eu`
[5] `http://data.dc.gov`

Percentile	Score
0%	2.08
20%	180.55
40%	278.32
60%	428.66
80%	667.03
100%	1388.53

Table 2: Safety score percentiles.

TREC Rank	Title	Mapped score
1	Zaytinya	5
2	Bistrot Du Coin	4
3	Old Ebbitt Grill	3
4	Blue Duck Tavern	3
5	Founding Farmers	4

Table 3: Safety score for suggestions.

centiles, between the 60% and 80% percentiles or above the 80% percentile. of the scores given to areas where crimes have occurred Table 2 lists the percentiles for the scores in figure 1.

With this data we can now calculate our mapped safety scores for the sample suggestions as shown in table 3. Note that there are other possibilities that could be used to normalize safety scores. In particular, one possibility would be to factor in population density around the attraction location into the normalized scores.

Expanding on the use case of having safety scores to indicate the threat to personal safety we can note that the types of crimes have been annotated provide more detailed information about the type of crime in the area. The types of crimes are listed in table 4. One possible way to break down the type of threat is to differentiate between threat to personal safety and threat to the safety of a visitor's car. This information can be especially useful to visitors that are deciding where to park their car, or whether to leave their car at their hotel or not.

Using this information we can provide two safety scores: personal safety and vehicle safety. The only modification to equation 2 is to only consider certain types of incidents when calculating the score. In the types of incidents listed in table 4, “Motor Vehicle Theft” and “Theft From Auto” are the two that need to be considered when calculating the vehicle safety score, these two combined make up 35.7% of the incidents. All the other types of incidents are considered when calculating the personal safety score.

Type	# of Occurences	% of Occurences
Theft/Other	12453	35.42%
Theft From Auto	9917	28.21%
Robbery	4080	11.60%
Burglary	3346	9.51%
Motor Vehicle Theft	2634	7.49%
Assault With Dangerous Weapon	2289	6.51%
Sex Abuse	295	0.83%
Homicide	103	0.29%
Arson	35	0.09%

Table 4: Offense occurences.

Rank	Title	Vehicle	Personal
1	Zaytinya	4	5
2	Bistrot Du Coin	4	5
3	Old Ebbitt Grill	1	4
4	Blue Duck Tavern	3	3
5	Founding Farmers	1	5

Table 5: Personal and vehicle safety score for suggestions.

Following the same process of developing a general safety score we can develop the vehicle and personal safety scores, the score for our example can be seen in table 5. Now we can see that in these suggestions there is more of a concern with regards to personal safety than to vehicle safety.

The mean general safety score associated locations that crimes occurred shown in figure 1 is 426. In comparison to that the mean vehicle safety score is 145 and the mean personal safety score is 281. The purpose of reporting these two score separately is so that areas that travellers need to be aware of one type of safety warning but not the other can be identified. We note that the Kendall tau coefficient of vehicle vs. personal safety scores is $\tau = 0.5224$, which shows that these areas are somewhat correlated but there are still areas where reporting these metrics separately can be useful.

We calculate the safety scores for all (roughly 600) attractions in Washington, D.C. that were given an interest rating by users as part of the Contextual Suggestion Track. Here, the Kendall tau coefficient of interest rating vs. safety score is $\tau = -0.0044$. We can see that regardless of whether the user likes or dislikes the attraction the range of given safety scores is similar. Liked attractions don't have a tendency to be in either safe or unsafe areas and so providing a safety score to users will help them make decisions.

Other information could also be taken into account when calculating safety scores. Firstly, the time of day the user is searching could be taken into account: a visitor might not need to be concerned about an area when crimes occur during the night if they are visiting during the day. Secondly more severe crimes could be weighted more heavily, for example homicides or crimes with a gun could be given more weight in the safety score.

4. DIRECTIONS FOR FUTURE WORK

Our demonstration above has shown the power of just one data source in one city to allow users to make more informed decisions. This is just the beginning, and as more cities develop their information infrastructure these data sources will grow in number. Those cities leading the way will reap the earliest benefits by being the first to offer these services to citizens and visitors, which will in turn intensify advocacy for the creation and release of more databases elsewhere.

Making safety data collected by governments available to the public has been a cornerstone of many smart cities initiatives. Some sources, like the St. Louis one, remain closed, but data concerning public safety is not limited to the government, or to crime reports. The Open311 project, run by the nonprofit OpenPlans, effectively crowdsources the gathering of non-emergency public safety information, from accidents to utility breakdowns. Open311 collects data in a number of cities worldwide, even being integrated into other smart cities projects such as CitySDK to help cities communicate their data to developers. While standard adoption is always a delicate issue in the early days of any field, a savvy developer might be wise to keep their eye on Open311's progress.

Another possibility to consider is extending the personalizing of profiles beyond merely the user's taste in venues. By integrating their security needs and habits into their profile, a system can provide more personalized alerts. A user who owns a car will naturally be more concerned about reports of car theft in an area they intend to park, while a user who regularly spends time in high-crime areas of a city will be less likely to be discouraged by security warnings surrounding a venue. Security profiles can be built alongside the existing preference profiles, by looking at the safety scores for those attractions the user has already visited.

5. IMPLICATIONS

While we have focused so far on the positive potential of integrating public safety information, it must be acknowledged by anyone looking to implement these ideas that there are real hazards. The most immediate one is that recommending a restaurant according to someone's tastes and recommending a course of action for their personal security are two wildly different actions with different consequences to match.

A poor restaurant recommendation might ruin an evening, whereas a mistaken safety report could put someone in jeopardy. Even if the report is accurate, having that information might lull some users into a false sense of security when they should still be alert. A mere suggestion system cannot possibly take full responsibility for the safety of its users, nor can a developer guarantee that an area reported as safe could not be the site of a crime. Developers must be aware of this when designing systems, noting that they provide only supplementary information and do not take the place of caution and common sense.

Another unavoidable issue is the locality of crime. Neighborhoods and vicinities with a high crime frequency still have businesses who may not appreciate customers being "scared off". Residents of these neighborhoods may also not appreciate their homes being labeled a high-risk area, or the implications of tourists being steered away from them. Many databases also currently include personal information, such as the perpetrators or victims of crime, whose use should be avoided to protect the privacy of residents.

These are valid concerns, which is why the decision to act on available data must remain firmly with the user. Suggestion systems, like tourist guidebooks and websites, should remain focused on merely providing publicly available data to promote informed decision-making. Likewise, those who maintain smart cities databases should take care not to release unnecessary personal information to the public. The misrepresentation of a neighborhood and those who live in it can be damaging in many ways, and a thoughtful developer will consider this in the design of their system.

Having discussed these pitfalls, it might seem as though the risks of integrating public safety information outweigh the benefits. However, keeping public safety information outside of contextual suggestion does not mean that users won't factor in safety when making their decisions. If anything, by not presenting public data, users are more likely to make ill-informed decisions based on hearsay, prejudice or outdated information. While we can avoid responsibility by not making any pretense of speaking to safety, we cannot diminish the user's real information need.

6. CONCLUSIONS

In this work we have demonstrated the power and potential of public safety data when integrated with contextual suggestion. With just crime data from a single city, we can provide users with the information and analysis necessary to promote safer, smarter decision-making - the overarching goal of the smart cities paradigm. A greater number of more varied public safety databases are emerging all the time, but it will take far more work to go from talking about their integration to implementing it.

There can be no doubt that this work is necessary. As information technology infrastructure continues to grow, cities will continue to get smarter, and developers will need to keep up. It behooves us to take advantage of these opportunities in order to better serve users, despite the many risks and challenges involved. Even with these new sources of information being made available, tackling the safety of users is an enormous responsibility. To shirk it because we fear the consequences would be the greater failure.

7. REFERENCES

[1] L. Baltrunas, B. Ludwig, S. Peer, and F. Ricci. Context relevance assessment and exploitation in mobile recommender systems. *Personal and Ubiquitous Computing*, 16(5):507–526, 2012.

[2] A. Dean-Hall, C. L. Clarke, J. Kamps, P. Thomas, and E. Voorhees. Overview of the trec 2013 contextual suggestion track. In *To appear in 22st Text REtrieval Conference, Gaithersburg, Maryland*, 2013.

[3] V. Estivill-Castro and I. Lee. Data mining techniques for autonomous exploration of large volumes of geo-referenced crime data. In *Proceedings of the 6th International Conference on Geocomputation*, 2001.

[4] I. Garcia, L. Sebastia, and E. Onaindia. On the design of individual and group recommender systems for tourism. *Expert Systems with Applications*, 38(6):7683 – 7692, 2011.

[5] M. Groen, W. Meys, and M. Veenstra. Creating smart information services for tourists by means of dynamic open data. In *Proceedings of the 2013 ACM Conference on Pervasive and Ubiquitous Computing Adjunct Publication*, UbiComp '13 Adjunct, pages 1329–1330, New York, NY, USA, 2013. ACM.

Tourists in Smart Cities: Data mining for hidden treasures

Invited contribution

Jon Oberlander
School of Informatics, University of Edinburgh, UK

ABSTRACT

Tourism plays a large part of the cultural and economic benefits of Scotland; according to Deloitte, it accounts for just over 10% of the country's GDP. The SICSA Smart Tourism Programme brings together ScotlandâĂŹs informatics researchers, cultural heritage organisations, and technology companies to address some of the key challenges in the sector. Key common needs are: (i) Personalisation: Improving how visitors navigate the city/country and volume of events/sites and associated information overload; (ii) Elastic demand and greening: Making available more resources as efficiently as possible during times of peak demand for services (ticket booking, accommodation, food, transport, healthcare); (iii) New channels, new content: Increasing access, audiences and brand awareness for places, performances and events. The talk will start with an overview of the 15 projects we and our collaborators have tackled over the last three years. It will then focus in on two specific projects which are data-mining city archives (from museums and libraries) to reveal hidden treasures, and help venues attract visitors off the streets.

ECIR'14 Information Access in Smart Cities Workshop (i-ASC 2014).
April 13, 2014, Amsterdam, the Netherlands.

Biography

Jon Oberlander has been Professor of Epistemics at the University of Edinburgh since 2005. He works on getting computers to talk (or write) like individual people, so his research involves not only studying how people express themselves - face to face or online - but also building machines that can adapt themselves to people. He collaborates with linguists, psychologists, computer scientists and social scientists, and has long standing interests in the uses of technology in cultural heritage and creative industries. He was founder-Director of the Scottish Informatics and Computer Science Alliance, and is Co-Director of the Centre for Design Informatics.

“Search-the-City” – A versatile dashboard for searching and displaying Environment and User Generated Content in the context of the future Smart City

Athanasios Moralis George Perreas Anastasios Glaros Dimitrios Dres

Telesto Technologies
Imitou 62 Street Cholargos
Athens, Greece
T. +30-210-6541942
F. +30-210-6545782

amoral@telesto.gr perreas@telesto.gr tasosglaros@telesto.gr jdres@telesto.gr

ABSTRACT

The modern city incorporates a large amount of heterogeneous data, which are produced by diverse sources like sensors, cameras and social networks and present a new challenge: harnessing these data in a usable approach providing suitable views for the city’s numerous stakeholders. In this paper we shall present a Visualization Framework based on SMART FP7 project, that allows users and developers to build visual applications that empower these environmental- and user-generated data in a meaningful way, appropriate for the future Smart City.

Keywords

Future Internet, Sensors, GUI, Web 2.0, Social Networks, Smart City

1. INTRODUCTION

The Future Internet will include a large number of internet-connected sensors (including cameras and microphone arrays), which provide opportunities for searching and analyzing large amounts of multimedia data from the physical world, while also integrating them into added-value applications. Despite the emergence of techniques for searching physical world multimedia (including the proliferation of participatory sensing applications), existing multimedia search solutions do not provide effective search over arbitrary large and diverse sources of multimedia data derived from the physical world.

Future Internet in the context of a Smart City [1] defines an environment where a huge amount of data is produced from heterogeneous sources. These sources include real time environmental sensors (cameras and sensors), user generated content and content produced by the various city authorities. The volume of data and the diversity of sources render these data difficult to be consumed, by the end users, thus making them unusable for any practical application.

ECIR'14 Information Access in Smart Cities Workshop (i-ASC 2014).

April 13, 2014, Amsterdam, the Netherlands.

The numerous stakeholders increase the complexity of the problem. These include the municipal administration that aims at increasing the quality of life within the city, the citizens that work and live, the private contractors that provide services to the city and the visitors. These groups have different interests in the information produced. **The goal of the current contribution is to describe a holistic approach followed in the framework of SMART FP7 [2] project with specific emphasis placed on the visualization of data to suit the specific needs of the stakeholders**. The visualization framework offers the required tools to the **Web 2.0 Smart City application developer** to easily create new applications to consume these open data, thus enabling the open services vision and the development of an ecosystem.

In the following sections we first present (Section 2) the related work, we then proceed with the presentation of the enabling architecture (Section 3). In section 4 we present the proposed Visualization Framework and finally in section 5 we present the future developments of the framework.

2. RELATED WORK

The ability to explore and discover the city through the data gathered is essential for the Smart Cities of tomorrow. This is a common problem and various efforts have been made up-to now to tackle it. IBM with Intelligent Operations Center [3] provides an executive dashboard to help city leaders gain insight into various aspects of city management. Other efforts include the CityDashboard [4], a project of the University College of London (UCL). It is designed to offer an at-a-glance view of eight cities around Great Britain. It combines official, observational and social media data into a single screen, the dashboard, which updates continuously high level data.. Suakanto in [5] presents a city dashboard for the city of Bandung in Indonesia that summarizes the condition of the city in terms of traffic congestion, water supply, energy supplies, air quality and public health quality. Michael Batty in [6] defines a smart city and how existing infrastructure is merged with ICT using new digital technologies.

A common place of the above related work is to present city’s heterogeneous data on a mashup, i.e. a web site that combine information from various sources presenting them visually by using different “web widgets”[1]. These widgets are not connected

[1] http://en.wikipedia.org/wiki/Web_widget

and do not allow designers to create new interactions with displayed data. The designer just “drops” these widgets on a page thus offering the city’s dashboard. Based on the idea of the mashup and the flexibility that widgets offer to the designer, we propose a Visualization Framework that attempts to provide a similar visual solution, and at the same time add interactions between the different widgets, offering a more interactive environment that enables the city monitoring and allows the searching of specific information. The latter is enabled by the SMART FP7 infrastructure that gathers and annotates semantically the information generated from various data sources, by the use of the SMART Search Layer. The proposed Visualization Framework targets the developer of the city’s dashboard by providing him a) with the means to build an interactive mashup and b) the best practices to extend it functionality. In the following section we present the architecture drives the Visualization Framework.

3. ENABLING ARCHITECTURE

As we mentioned above, the Visualization Framework is tightly coupled with the SMART Search Layer. Considering this fact, we give a brief presentation of the enabling architecture.

The SMART Search Layer is a core component responsible of providing effective and efficient real-time indexing and retrieval of social and sensor data streams. It is built with Terrier and a MapReduce (termed as “SmartReduce”) architecture using open source Storm stream processing framework to scale to a large volume of social and sensor streams produced by multiple edge nodes. The architecture has been designed as layered system consisting of three main layers, namely:

- **A layer of sensor edge servers**, which ensures the interfacing of the SMART with data stemming from the real world. The sensor edge server undertakes the tasks of acquiring physical world data via sensors (notably cameras and microphone) and accordingly structuring data based on standard formats (such as XML, RDF/Linked Data) to allow the upper layers to consume these data.

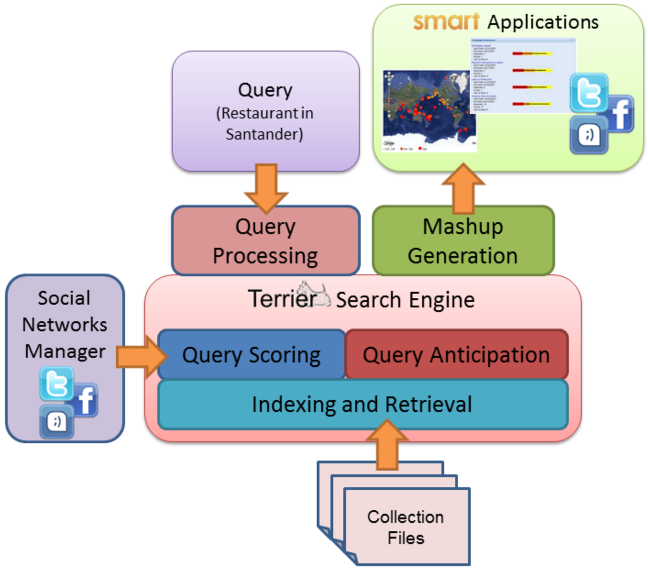


Figure 1. High-Level Overview of the SMART Architecture

- **A search layer** comprising a customized version of the Terrier search engine, adapted to the SMART project’s needs. This layer indexes and retrieves data from the layer of edge servers. At the same time it retrieves, parses, analyzes queries and uses them for the selection (or prioritization) of appropriate edge servers where data of interest can be found. It can also support event-driven queries («pull» mode), in which case the search engine subscribes to events and updates the upper layer upon the occurrence of these events (i.e. recurring queries).

- **A layer of end-user applications** (including text-based interactive queries), which submit queries to the search engine and visualize the results. In addition to interactive-text based queries, this layer supports the Web2.0 compliant visualization of information stemming from the physical world (according to a mash-up concept).

A high level overview that describes the general architecture is depicted in Figure 1. SMART promises also to augment query results on the basis of information stemming from the social networks, while also using social networks information (such as information stemming from facebook, twitter and foursquare) in order to personalize the query results. To this end, a Social Media Manager has been developed that removes the heterogenity of the various social networks and inserts their data as sensor data (feeds) to the edge sensor server layer ensuring that the SMART system can additinally process and consume data stemming for social networks, thereby allowing sensor networks and sensor networks processing algorithms (e.g., sentiment analysis [7], event detection [8], gender analysis) to be included in the SMART system. At the same time, passing personalized information to the search layer allows the acquisition of personalized information for the specific user on the basis of their personal social networks’ accounts and associated social graphs.

4. VISUALIZATION FRAMEWORK

The goal of the Visualization Framework is to offer components and describe the best practice for building applications based on intuitive interfaces, tailored to the needs of the city stakeholders by combining these components. Hence it is crucial to provide the whole functionality in visual components (widgets) that can communicate with each other and combine them on a dashboard that adheres user management and access control management capabilities, based on a Role Based Access Control (RBAC) approach. The RBAC enables the definition of roles that map on the different city’s stakeholders, permitting specific fine-grained view based on user’s role. Thus the same dashboard can have a public view that reveals the data that the citizen or the visitor expects to see and private views for authorized users such as the municipal administration or the private contractors that reveal a special view to satisfy the stakeholder’s requirements.

On the functionality aspect, an application composed by these visual components must provide the capability of adding new components to the dashboard (via a drag-and-drop functionality) and configuring their behavior, like the sources that provide their data. To this end we have identified (based on our experience on the SMART FP7 project), some standard functionality and developed a number of Web2.0 reusable widgets. These widgets (visual components) consume services and semantics that originate from the underlying A/V processing algorithms, sensors and social networks. The availability of such widgets, aims at facilitating embedding queries made and search results returned from the Search Engine into applications. These widgets are indicative of the functionality that the framework offers. Following their design principles and methodology, a developer can create new ones to meet specific use case requirements.

The Visual Framework offers the required functionality by utilizing the SMART Search Layer to query the results and get the required data in the form of events. An event is an aggregation of data. E.g. a demonstration that takes place in a city is an event that is linked with a place, date of event, data from sensors near the event (video feeds, audio feeds), and user-generated data (tweets, posts) from social networks.

The Visual Framework consists of two main components:

- The Dashboard
- Standard Visual Components (mash-ups)

The Visual Components (VCs) retrieve their data using the underlying Search Layer. They have the ability to provide data representation following their design and communicate with each other, sharing information via "software application events". For example, a VC, acting as a master, can upon users input, retrieve some results from the undelaying search layer and fire an software event that other VCs catch and process showing some details, thus, implementing a master-detail GUI functionality.

The stakeholders that have logged-in the system, may create their preferred view of the application by adding (as simple as drag-and-drop) the visual components that meet their own requirements on condition that they have the required privileges.

Besides, application developers can extend the visual framework by designing and developing their own VCs that exploit the services provided by the SMART Search Layer, following the frameworks guidelines.

4.1 The "Demonstration" Application

A demonstration of such Web Environment built on the Visual Framework and the standard Visual Components is presented in

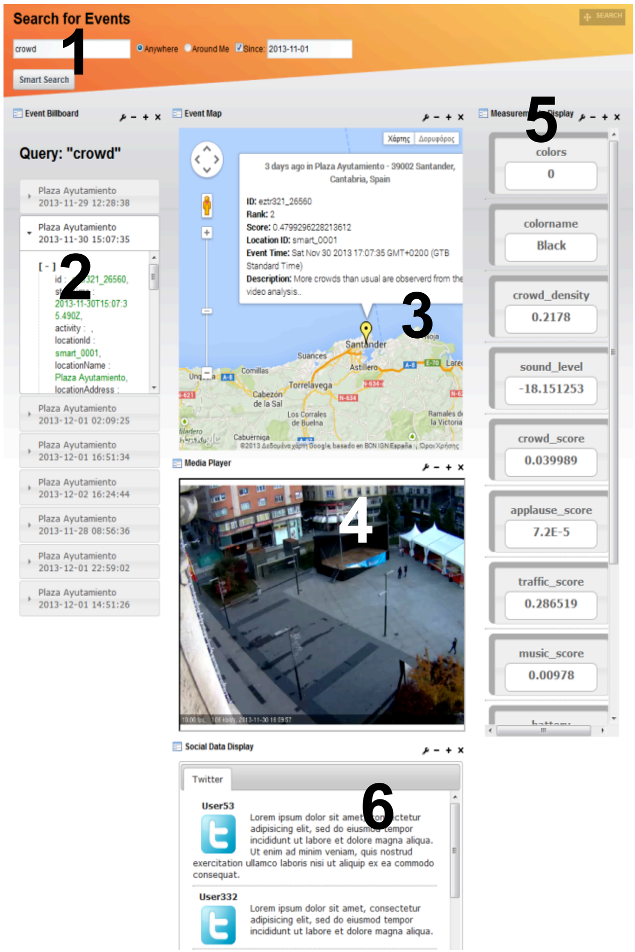
Figure 2. The application is the "Developer scenario" for the SMART FP7 project that demonstrates the functionality of an event search application. The numbers on the figure show the different VCs that compose the application as well as their position in a sequence of actions the user performs.

In the example a "Search" visual component (1) is used to fetch results according to the user's input. It makes a request to the SMART Search Layer and receives a list of results relative to events of interest for the specific search performed, ranked according to the score awarded by the search layer.

The results are displayed on an "Event Billboard" visual component (2). When a user selects a result, an event in our case, this event is shown on the "Event Map" visual component (3). The event as explained in section 4, may be accompanied with other data. These data are displayed according to their type to the suitable visual component. Images and videos from nearby cameras relevant to an event matching the search made are displayed in the "Media Player" (4). Crowd analysis of the event and the indicative measurements from non-Audiovisual sensors critical for the understanding of the event (e.g. heavy rainfall) are presented in a "Measurements Display" visual component (5). Finally, mentions in social media related to the event are shown on a "Social Data Display" visual component (6).

4.2 The dashboard

The dashboard of our visual framework is the foundation, as it provides user identification and access control. Moreover it allows the deployment of visual components by the users. It is based on the Liferay portal. Liferay is a free and open source enterprise portal project providing a web platform with features commonly required for the development of websites and portals. It differs from the mainstream Web Content Management System (CMS), as it employs Java-based technologies. It offers a user and role management system with single-sign-on support as well as a built-in Web CMS which offers the users with the familiar features required to build websites and portals, i.e. an assembly of themes, pages and portlets. The portlets supported adhere to the Java JSR-286 portlet specification.



Search for Events

crowd ● Anywhere ○ Around Me ☑ Since: 2013-11-01

Smart Search

Query: "crowd"

Plaza Ayutamiento 2013-11-29 12:28:38

Plaza Ayutamiento 2013-11-30 15:07:35

[-]
id : ...321_26560,
st...time :
2013-11-30T15:07:35.490Z,
activity : ,
locationId :
smart_0001,
locationName :
Plaza Ayutamiento,
locationAddress :

Plaza Ayutamiento 2013-12-01 02:09:25

Plaza Ayutamiento 2013-12-01 16:51:34

Plaza Ayutamiento 2013-12-02 16:24:44

Plaza Ayutamiento 2013-11-28 08:56:36

Plaza Ayutamiento 2013-12-01 22:59:02

Plaza Ayutamiento 2013-12-01 14:51:26

3 days ago in Plaza Ayutamiento - 39002 Santander, Cantabria, Spain

ID: eztr321_26560
Rank: 2
Score: 0.4799296228213612
Location ID: smart_0001
Event Time: Sat Nov 30 2013 17:07:35 GMT+0200 (GTB Standard Time)
Description: More crowds than usual are observerd from the video analysis..

colors	0
colorname	Black
crowd_density	0.2178
sound_level	-18.151253
crowd_score	0.039989
applause_score	7.2E-5
traffic_score	0.286519
music_score	0.00978

User53 Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

User332 Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

Figure 2. An example of Rich Application

The portal thus has been customized to be utilized as an Urban sensing / Urban monitoring platform, which may be customized to the user's needs through the use of reusable portlets developed specifically for the project's needs. The portlets are available for deployment in any way the user sees fit, thus personalizing the resulting dashboard to meet the needs of any scenario.

4.3 Standard Visual Components

The Visual Components described in this section are developed using Web 2.0 technologies, like the jChartFX and JQueryUI, which are open source Javascript libraries. The map-based Visual Component uses the Google Maps Javascript API V3. These VCs are developed as Java Portlets (JSR-286 Java API). Consequently, the VC-to-VC communication is implemented as inter-portlet communication using the event mechanism described in JSR-286, described as "software application events" in the previous sections so as to distinguish them from the real events that the search layer returns. A portlet (VC in our case) issues an event that encapsulates the data that need to be sent to other VCs. The VCs that are configured for listening for that specific event retrieve the data from the event, process and illustrate them according to the VC's designed functionality. In this section we shall present the visual components we have implemented.

4.3.1 Search Visual Component

The user is able to query the Search Engine VC about specific words or phrases of interest and define some search options such as the maximum radius of the event's location around the user's current location or events within specific dates. The Search VC retrieves the search results in a JSON format and fires an event with these results. Other VCs listening for this type of event receive the search results and process them according to their functionality.

4.3.2 Event Billboard Visual Component

The Event Billboard VC displays the top ten events provided by the search engines using accordion widget. The sorting of the results is based on the score awarded by the search engine. Upon clicking an event, the accordion expands providing the full event's details while it triggers the map VC to show the specific event on the map.

4.3.3 Event Map Visual Component

The Event Map VC features a map showing the selected event from the Event Billboard VC. The map moves to center around the event. Upon selecting an event, it provides the rank, score, location and date-time information of the event according to the classification returned from the Search Layer. Additionally the score of the event is visually indicated by its' map marker color. The red marker indicates a low score, the yellow color a medium score, whereas the green color indicates a high score. The score is computed by the Search Layer dynamically and it relates to the search terms given of the user.

4.3.4 Measurements Display Visual Component

This VC displays the various measurements recorded in proximity of the event selected on the Event Billboard VC at the date-time of the event. The measurements come from all sources from thermometer sensors in the area to the crowd density virtual sensors derived from a camera feed (if a camera is available at the event's location and a crowd density sensor has been defined for the location of the event taking place).

4.3.5 Media Player Visual Component

The Media Player VC provides audio and/or video playback for the clips recorded in the area of the selected event at the event's date-time.

4.3.6 Social Data Display Visual Component

Finally, the Social Data Display VC provides the social posts/tweets, which are geo-located around the area of the event and are relevant to the search term/phrase used in the search VC.

5. FUTURE WORK

In this paper we have presented a Visual Framework for building dashboards for Smart Cities using Future Internet technologies and demonstrated it's functionality by developing a number of Visual Components. Based on the experience gained, we will focus on the optimal combination of environment generated content originated from sensors and social networks, in order to provide new Visual Components that add value to a city's dashboard. Regarding the technical capabilities of the framework, we shall investigate the possibility to generalize the Visualization Framework by providing an abstract layer and an implementation that utilizes the search layer. The abstract layer will consist of an abstract data model that the VCs process and abstract services that acquire the data. This will allow the loosely coupling between the VCs and the underneath search layer, allowing the easy integration of new data sources.

6. ACKNOWLEDGMENTS

Part of this work has been carried out in the scope of the SMART project (FP7-287583) (http://www.smartfp7.eu/). The authors acknowledge help and contributions from all partners of the project.

7. REFERENCES

[1] Shapiro, J. M. 2006. Smart Cities: Quality Of Life. Productivity, And The Growth Effects Of Human Capital. *Review of Economics and Statistics. v88 (2,May 2006) 324-335*

[2] M.-D. Albakour, D, Macdonald, C., Ounis, I., Pnevmatikakis, A., Soldatos, J. 2012. SMART: An Open Source Framework for Searching the Physical World. *Proceedings of the ACM SIGIR 2012 Workshop on Open Source Information*.

[3] IBM Intelligent Operation Center - http://www-03.ibm.com/software/products/en/intelligent-operations-center/

[4] CityDashBoard project - http://citydashboard.org/london/

[5] Suakanto, S., Supangkat, S.H., Suhardi, Saragih, R. 2013. Smart city dashboard for integrating various data of sensor networks. *ICT for Smart Society (ICISS), 2013 International Conference on , vol., no., pp.1,5, 13-14 June 2013* DOI=http://doi: 10.1109/ICTSS.2013.6588063.

[6] Batty, M., Axhausen, K., Fosca, G., Pozdnoukhov, A., Bazzani, A., Wachowicz, M., Ouzounis, G. and Portugali, Y. Oct 2012. *Smart Cities of the Future*. UCL Working Paper, ISSN 1467-1298.

[7] Agarwal, A., Xie, B., Vovsha, I., Rambowod, O., Passonneau, R. 2011. Sentiment analysis of Twitter data. *In the Proceedings of the Workshop on Languages in Social Media (LSM '11), pp. 30-38.*

[8] Becker, H., Naaman, M., Gravano L. 2009. Event Identification in Social Media. *In WebDB, 2009.*