Stigmergy-based modeling to discover urban activity patterns from positioning data.

Antonio L. Alfeo[1], Mario G. C. A. Cimino[1] Sara Egidi[1], Bruno Lepri[2], Alex Pentland[3], and Gigliola Vaglini[1]

[1] University of Pisa, largo Lazzarino 1, Pisa, Italy
luca.alfeo@ing.unipi.it, mario.cimino@unipi.it,
s.egidi1@studenti.unipi.it, gigliola.vaglini@unipi.it
[2] Bruno Kessler Foundation, via S. Croce, 77, Trento, Italy
lepri@fbk.eu
[3] M.I.T. Media Laboratory, Cambridge, MA 02142, USA
pentland@media.mit.edu

Abstract. Positioning data offer a remarkable source of information to analyze crowds urban dynamics. However, discovering urban activity patterns from the emergent behavior of crowds involves complex system modeling. An alternative approach is to adopt computational techniques belonging to the emergent paradigm, which enables self-organization of data and allows adaptive analysis. Specifically, our approach is based on stigmergy. By using stigmergy each sample position is associated with a digital pheromone deposit, which progressively evaporates and aggregates with other deposits according to their spatiotemporal proximity. Based on this principle, we exploit positioning data to identify high-density areas (hotspots) and characterize their activity over time. This characterization allows the comparison of dynamics occurring in different days, providing a similarity measure exploitable by clustering techniques. Thus, we cluster days according to their activity behavior, discovering unexpected urban activity patterns. As a case study, we analyze taxi traces in New York City during 2015.

Keywords: Urban mobility, stigmergy, emergent paradigm, hotspot, pattern mining, taxi-GPS traces.

1 Introduction

The increasing volume of urban human mobility data arises unprecedented opportunities to monitor and understand crowd dynamics. Identifying events which do not conform to the expected patterns can enhance the awareness of decision makers for a variety of purposes, such as the management of social events or extreme weather situations [1]. For this purpose GPS-equipped vehicles provide a huge amount of reliable data about urban human mobility, exhibiting correlation with people daily life, events, and city structure [2]. The majority of the methods approaching the analysis of vehicle traces can be grouped into three categories: *cluster-based*, *classification-based*, and *pattern mining-based*; whereas the main

application problems include the hotspot discovery, the extraction of mobility profiles, and the detection and monitoring of big events and crowd behavior [3]. For example, in [4] the impact of a social event is evaluated by analyzing taxi traces. Here, the authors model typical passenger flow in an area, in order to compute the probability that an event happens. Then, the event impact is measured by analyzing abnormal flows in the area via Discrete Fourier Transform. In [5] GPS trajectories are mapped through an Interactive Voting-based Map Matching Algorithm. This mapping is used for off-line characterization of normal drivers' behavior and real-time anomaly detection. Furthermore, the cause of the anomaly is found exploiting social network data. In [6] the authors use a Multiscale Principal Component Analysis to analyze taxi GPS data in order to detect traffic congestion.

One of the main issues concerning the analysis of this kind of data is their dimensionality. Many approaches handle it by focusing on specific areas (*hotspots*) whose high concentration of events and people can summarize mobility dynamics [7]. As an example, in [8] a density-based spatial clustering is employed to perform spatiotemporal analysis on taxi pick-up/drop-off to find seasonal hotspots. Authors in [9] use OPTICS algorithm in order to detect city hotspots as density-based clusters of taxi drop-off positions. Recently, in [10] an Improved Auto-Regressive Integrated Moving Average algorithm is proposed; it is aimed to detect urban mobility hotspots via taxi GPS traces and analyze the dynamics of pick-ups in dense locations of the city. However, due to the complexity of human mobility data, the modeling and comparison of their dynamics over time remain hard to manage and parametrize [11]. In this paper, we present an innovative approach based on *stigmergy* [12] that aims to handle both complexity and dimensionality of these data, providing an analysis of urban crowds dynamics by exploiting taxi GPS data. Specifically, our investigation covers the city hotspots identification, the characterization of their activity over time and the unfolding of unexpected activity pattern.

The paper is structured as follows. In Section 2 the architectural view of our approach is described. In Section 3 the experimental studies and results are presented. Finally, Section 4 summarizes conclusions and future work.

2 Approach Description

In this section, we present our approach, based on the principle of *stigmergy*. Stigmergy is an indirect coordination mechanism used in social insect colonies [12]. It is based on the release of chemical markers (*pheromones*), which aggregate when subsequently deposited in proximity with each other. This mechanism can be employed in the context of data processing, providing self-organization of data [13] while unfolding their spatial and temporal dynamics [14]. By exploiting stigmergy, we discover city hotspots, characterize their activity dynamics (i.e. presence of people over time) and assess unexpected activity patterns. In order to focus on activity dynamics, we employ New York City taxi positioning data,

considering the amount of passengers together with the GPS position of each pick-up/drop-off.

2.1 Hotspot Detection

At the beginning, data samples are transformed in digital pheromone deposits, allowing the progressive emergence of city hotspots (i.e. the most high-density areas within the city). Firstly, data are treated by the smoothing process (Fig.1), in order to remove insignificant activity levels and highlight relevant dynamics. This process is implemented by applying a sigmoidal function to the samples. Then, a mark is released in correspondence of each smoothed sample in a three-dimensional virtual environment. Marks are defined by a truncated cone with a given width and intensity (height) equal to data sample value. The trailing process aggregates marks, forming a *stigmergic trail*, which is characterized by evaporation (i.e. temporal decay δ) and defined as $T_i = (T_{i-1} - \delta) + Mark_i$.

As an effect, isolated marks tend to disappear, whereas the arrival of new marks in a given region counteracts the evaporation. Thus, aggregation and evaporation can act as an agglomerative spatiotemporal clustering with historical memory. Hotspots are identified as the city areas corresponding to the overlapping of the most relevant trails obtained by processing data in early morning (i.e. 3am-8am), morning (i.e. 9am-2pm), afternoon/evening (i.e. 3pm-8pm), and night (i.e. 9pm-2am) time slots. As an example, Fig. 1 shows the hotspots identified in Manhattan (New York City). Their locations correspond to: East Harlem - Upper East Side (A), Midtown East (B), Broadway (C), East Village - Gramercy - MurrayHill (D), Soho - Tribeca (E), Chelsea (F) and Time Square - Midtown West - Garment (G).

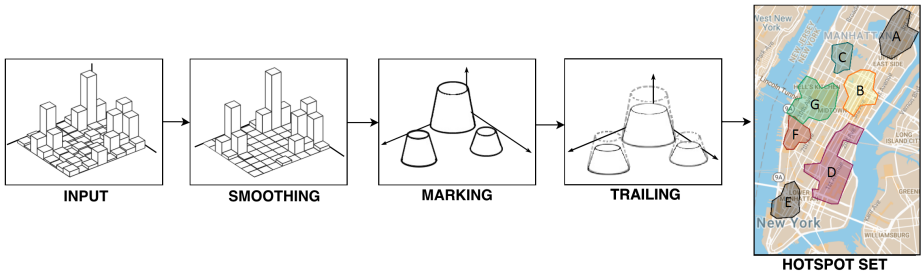

Fig. 1. The stigmergy-based process of hotspot discovery.

2.2 Hotspot Activity Characterization

For each identified hotspot, we generate the activity time series, by periodically collecting the amount of activity occurred in the hotspot during a day. Let us

consider an activity time series; what is actually interesting is not the continuous variation of the activity over time, but the transition from one type of behavior to another.

Generally, given a time window each hotspot behavior can be characterized by an ideal time series segment of hotspot activity representing that specific behavior. More formally, we define it as an *archetype*. An example of an archetype is *asleep* behavior, which usually occurs during the night, between the calming down of the nightlife and the arrival of the workers; here the city exhibits its lowest activity level.

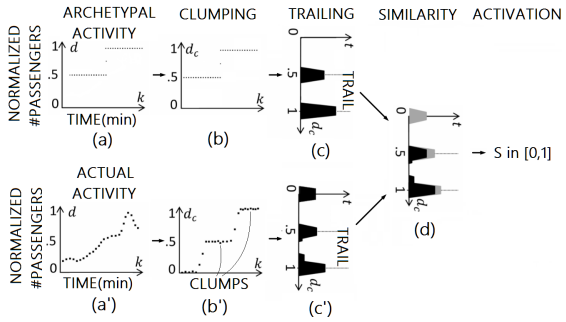

Fig. 2. The architecture of a SRF.

In order to detect an archetypal behavior in hotspot activity time series, we design a processing schema called Stigmergic Receptive Field (SRF), because it is receptive to a specific archetype and it processes samples employing the principle of stigmergy. Specifically, SRF computes a degree of similarity between a specific archetype (Fig. 2a) and an activity time series (Fig. 2a'), by subsequently processing their samples, which are assumed to be normalized between 0 and 1.

First, samples undergo the *clumping process* (Fig. 2b and Fig. 2b'), which acts as a sort of soft discretization creating clumps of samples. Clumps arrangement can be parametrized allowing to fit the analysis over the archetype's levels of interest. The clumping can be implemented as a double sigmoidal function. Second, the marking process (Fig. 2c and Fig. 2c') enables the release of a mark in a bi-dimensional virtual environment in correspondence of the sample value. The mark can be implemented by a trapezoid with given intensity (height) and width ϵ. Third, the trailing process accumulates marks creating the trail structure, whose intensity decays (i.e. evaporates) of a given rate δ at each step of time. As an effect, evaporation rate and mark width allow the trail to capture coarse spatiotemporal structure in data, handling micro-fluctuations. Fourth, current T_{act} and archetypal trails T_{arc} are compared by the similarity process (Fig. 2d), by using the Jaccard coefficient $S = |T_{arc} \cap T_{act}|/|T_{arc} \cup T_{act}|$ [15]. This coefficient provides a measure of similarity between 1 (identical trails) and

0 (non-overlapping trails). Finally, the activation process is applied to enhance only relevant similarity values and remove insignificant values according to the activation thresholds α_a, β_a. This process can be implemented by using the already mentioned sigmoidal function, i.e. $f(x, \alpha_a, \beta_a) = 1/(1 + e^{-\alpha_a(x-\beta_a)})$.

In order to provide an effective similarity, the SRF's parameters have to be properly tuned. With this aim, the Adaptation process uses the Differential Evolution (DE) to adapt the structural parameters of the SRF: (i) the clumping inflection points α, β, γ, λ; (ii) the mark width ϵ; (iii) the trail evaporation δ; (iv) the activation thresholds α_a, β_a. The aim of DE is to minimize the mean square error (MSE), considering the error as the difference between the target $\hat{S}$ and the computed S similarity values over a set of M labeled time series, i.e. $Fitness = \sum_{i=1}^{M}(|S_i - \hat{S}_i|^2)/M$. The target similarity value is 1 if the current time series exhibits the archetypal behaviour, 0 otherwise.

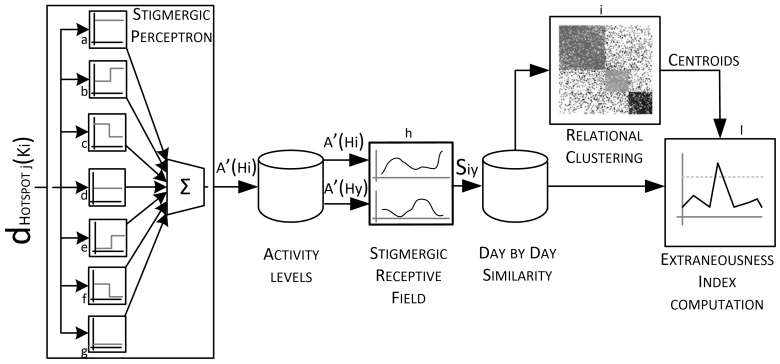

Fig. 3. The overall processing of activity samples.

Since any real signal is usually similar to more than one archetype, a collection of SRFs, specialized on different archetypes and ordered for increasing activity, is arranged in a connectionist topology to make a Stigmergic Perceptron (Fig. 3). Specifically, adopted archetypes are: Asleep (Fig. 3g), i.e. the hotspot at its lowest activity level; Falling (Fig. 3f), i.e. the flow just before the city activity calms down; Awakening (Fig. 3e), i.e. the waking up of urban life after a calm phase; Flow (Fig.3d), i.e. the hotspot at its operating capacity, usually exhibited during working hours; Chill (Fig. 3c), which usually occurs after a rush hour, when people leave work and take taxis to return home; Rise (Fig. 3b), i.e. the hotspot transition to its most intense activity level; and Rush-Hour (Fig. 3a), which usually occurs in early morning and late afternoon, when people movement is at its highest rate. A perceptron computes a single output from multiple inputs, by forming a linear combination of them. Similarly, the stigmergic perceptron (SP) combines linearly SRFs' outcomes by computing their weighted mean, using the provided similarities S_i as weights, i.e. $ActivityLevel = \sum_{i=1}^{N}(S_i * i)/\sum_{i=1}^{N}(S_i)$. The resulting value is called activity

level and is defined between zero and N, where N is the number of SRFs. An important aspect concerning hotspot activity level computation is to train each SRF inside a SP in order to prevent multiple activations of SRFs. Let us consider the most sensitive SRF parameter, i.e. the evaporation δ. High evaporation prevents marks aggregation and pattern reinforcement, while low evaporation causes the saturation of the trail. In order to handle this sensitivity, the adaptation of each SRF inside a SP is twofold: (i) the Global Training phase is aimed to determine an interval for the evaporation rate of each SRF. The interval $[\delta_{min}, \delta_{max}]$ is obtained considering the narrowest interval including the fitness values above its 90th percentile, while the intervals for the other parameters can be statically assigned on the basis of application domain constraints; and (ii) the Local Training phase aims to find the optimum values for every module of each single SRF, by using the interval generated in the Global Training phase. As a result, a proper trained Stigmergic Perceptron provides the characterization of hotspot activity, by transforming a given time series of activity samples in a new time series of activity levels. In order to compute the overall similarity between hotspot activity levels gathered in two different days, we employ a further SRF (Fig. 3h) which uses one activity level time series just like it was an archetype. The adaptation in this specific SRF tunes mark width ϵ, trail evaporation δ, and activation thresholds α_a and β_a. As fitness function, we use the Mean Squared Error (MSE) between computed and ideal similarity over a set of labeled pairs of activity time series (i.e. the training set).

2.3 Unexpected patterns detection

Exploiting the mechanism described above, we generate the similarity matrix, that is the collection of similarities obtained by matching with each other the activity level time series of the training set. Provided similarity matrix can be processed by a fuzzy relational clustering technique, grouping days according to their daily activity similarity. Specifically, we employ Fuzzy C-Mean to compute the clusters centroid. The number of clusters corresponds to the number of daily activity behaviors taken into account in the analysis. Based on these centroids, the membership degrees of further daily activity level time series can be computed. The membership degrees are between 0 (not belonging to the cluster) and 1 (completely belonging to the cluster). By exploiting the membership degrees u_n as a distance, we measure the extraneousness of current activity level with respect to its expected cluster. The Extraneousness Index (EI) is defined as the Manhattan Distance between current daily activity level series d and the centroid of the cluster in which current day is assumed to belong. In Eq. 1, the computation case with 3 clusters is shown.

$$EI(d) = (|u_1(d) - u_1(C_2)| + |u_2(d) - u_2(C_2)| + |u_3(d) - u_3(C_3)|)/2 \quad (1)$$

We define as an Unexpected Pattern a day characterized by an activity level whose EI exceeds the maximum EI computed over the training set.

3 Experimental Studies and Results

We have analyzed a dataset of taxi traces provided by the Taxi and Limousine Commission of New York City, which contains information about all medallion taxi trips from 2009 to 2016 [16]. We focus our investigation on dynamics occurred during 2015 in Manhattan considering that it attracts the most of the taxi trips in New York City. A pre-processing step has been performed to remove missing values and discretize data in spatiotemporal bins defined as a squared area 10-foot- wide with duration of 5 minutes. Then, the min-max normalization is applied. In order to search for hotspots characterizing every possible city routine (i.e. summer and winter ones), the hotspot discovery procedure has been performed comprising data gathered in working days and week-ends of February 2015 and June 2015.

Since archetypes are assumed to be general, the training set for the SP's global and local phases is generated by using the pure archetype time series as seeds and applying spatial noise and temporal shift.

In order to validate the SP archetypal behavior detection, a set of time series have been manually labeled and the difference with the actual results of the SP is used to evaluate detection error. Each label corresponds to the expected SP result according to the archetypal behavior visually detected in current time series (i.e. 1 if Asleep, 2 if Falling, and so on). To this purpose, 35 time series (i.e. 5 for each archetype) have been provided to the SP. The obtained MSE is shown in Table 1. By considering the activity level operative range (i.e. [1 7]) and the provided MSE values, the system shows good detection performances, proving the functional effectiveness of the SRF and the SP.

Table 1. Mean Square Error in Archetypal Behavior Detection via SP.

ARCHETYPE	Asleep	Falling	Awakening	Flow	Chill	Rise	Rush-hour	TOT
MSE	0.215	0.029	0.029	0.028	0.166	0.020	0.143	0.633

In the next processing phase, a further SRF is aimed to assess the similarity between daily activity levels. It is provided with a training set obtained by selecting a set of pairs of daily activity levels. In order to supply a clustering process, such SRF is trained to distinguish similar and dissimilar signals, according to the behavioral class of daily activity levels, namely: (i) Working days (expected to fall between Monday and Tuesday), when crowd movements are mainly caused by working routines; (ii) Entertainment days (expected to fall on Friday and Saturday), in which people tend to spend the night out; (iii) Leisure days (expected to fall on Sunday), which are characterized by limited transportation usage. Their target similarity is 1 if days belong to the same behavioral class, 0

otherwise. Since the defined classes refer to the cyclical sequence of week days, our ground truth can be provided by the calendar itself. The 10% of computed daily activity levels have been used to create these pairs (i.e. 1296 pairs overall).

The Fuzzy C-Mean algorithm is used to group days according to their stigmergy-based similarity in order to arrange them among the three provided clusters, namely: Working, Entertainment and Leisure days. Upon this, we exploit the Extraneousness Index in Eq. 1 to assess unexpected patterns.

We show results obtained analyzing hotspot D, since it is characterized by multiple usages [17] allowing the displaying of every activity level behavioral class. Interestingly, this area is also found to be an hotspot by [9].

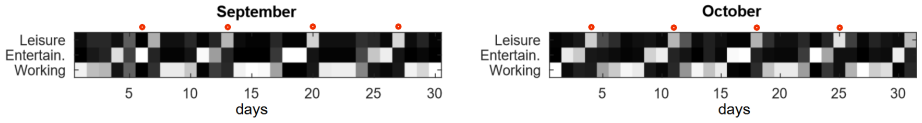

Fig. 4. Membership degrees of days in September and October. The whitest, the higher.

Fig. 4 shows the computed membership degree for each cluster, obtained with days in September and October. The whitest the box, the higher the degree. Clearly, the stigmergy-based characterization of hotspot daily activity allows to cluster days according to their behavioral class which corresponds to the arrangement we assumed. Indeed, most of the Sundays (highlighted by a circle in Fig. 4) exhibit their highest membership degree with Leisure day cluster. The same happened with days in Entertainment and Working cluster. It is worth noting that provided approach allows the mapping of daily behaviors to emerge from data instead of being explicitly injected into the system.

However, some days does not confirm this behavior. Indeed, by comparing their EI with the maximum EI in the training set (red line in Fig. 5), they are recognized as an unexpected pattern (red spot in Fig. 5).

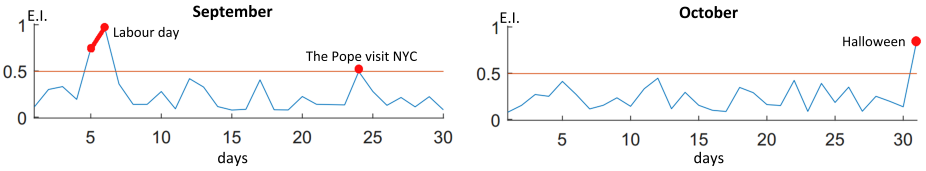

Fig. 5. Extraneousness Index computed over days in September and October.

Table 2 shows the most relevant unexpected patterns detected by analyzing the whole year 2015. Each unexpected pattern date is shown together with their most probable cause, such as an occurred social event. EI provides a continuous measure of the magnitude of unexpected patterns, allowing the comparison of

their impact on hotspot activity dynamics. As an example, Easter affects the activity in hotspot D much more than the NYC Half Marathon. Indeed, the greatest Easter celebrations in NYC are kept by the St. Patrick Cathedral, which is located in the area corresponding to hotspot D, whereas this area was not directly involved in the NYC Half Marathon 2015. By repeating the analysis in the same date on hotspot C, the computed EI results roughly 60% higher (i.e. 0.96); indeed the zone corresponding to hotspot C was directly crossed by NYC Half Marathon 2015.

Table 2. Most relevant unexpected patterns detected all over 2015.

EI	Date and occurred city event
0.96	06-Sep, Labour Day
0.94	24-May, Memorial Day
0.86	31-Oct, Halloween
0.83	26-Nov, Thanksgiving
0.83	28-Jun, Gay Pride
0.82	25-Dec, Christmas
0.81	01-Jan, New Year's Eve
0.80	04-Apr, Easter (holy Saturday)
0.79	27-Jan, Winter Storm Juno [18]
0.74	05-Sep, Labour day celebrations
0.63	03-Jul, Independence Day
0.63	31-Dec, New Year's Eve
0.61	15-Mar, NYC Half Marathon
0.49	24-Sep, Pope Francis visit NYC

4 Conclusion

In this paper, we proposed a novel approach aimed to provide knowledge discovery in the context of human urban mobility data. In contrast with the literature in the field, our approach does not require the in-depth modeling of the dynamics under investigation since it relies on data self-organization provided by employing the principle of stigmergy. Indeed, by using stigmergy, the spatiotemporal density in data has been exploited to identify city hotspots and characterize their dynamics, allowing to generate data-driven prototypes of typical daily activity. By treating them via a clustering technique, we were able to discern expected patterns from unexpected ones, which were found to be usually related to various events. One of the most promising improvements for this investigation can be achieved by cross-checking results obtained via vehicle GPS data with other data sources (e.g. social media or car crash data). Indeed, by employing a more detailed ground truth, the system can be specialized to model and detect patterns characterized by a timescale shorter than a daily one.

References

1. Sagl, G., Loidl, M., Beinat, E.: A visual analytics approach for extracting spatio-temporal urban mobility information from mobile network traffic. ISPRS International Journal of Geo-Information, 1(3), 256–271 (2012).
2. Veloso, M., Phithakkitnukoon, S., Bento, C.: Urban mobility study using taxi traces. ACM, In Proceedings of the 2011 international workshop on Trajectory data mining and analysis, 23–30 (2011).
3. Mazimpaka J. D., Timpf S.: Trajectory data mining: A review of methods and applications. Journal of Spatial Information Science, 2016.13, 61–9, (2016).
4. Zhang, W., Qi, G., Pan, G., Lu, H., Li, S., Wu, Z.: City-scale social event detection and evaluation with taxi traces. ACM, Transactions on Intelligent Systems and Technology, 6(3), 40 (2015).
5. Pan, B., Zheng, Y., Wilkie, D., Shahabi, C.: Crowd sensing of traffic anomalies based on human mobility and social media. ACM, In Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, 344–353. (2013).
6. Kuang, W., An, S., Jiang, H.: Detecting traffic anomalies in urban areas using taxi GPS data. Mathematical Problems in Engineering, (2015).
7. Hu, Yujie, Harvey J. Miller, Xiang Li. : Detecting and analyzing mobility hotspots using surface networks. Transactions in GIS, 18.6, 911–935, (2014).
8. Lu, Yu: An Intelligent System for Taxi Service Monitoring, Analytics and Visualization. Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, (2016).
9. Keler, Andreas, Jukka M. Krisp.: Is there a relationship between complicated crossings and frequently visited locations? A case study with boro taxis and OSM in NYC. 13th International Conference on Location-Based Services, (2016).
10. Li X., Pan G., Wu Z., Qi G., Li S., Zhang D., Zhang W., and Wang Z.: Prediction of urban human mobility using large-scale taxi traces and its applications. Frontiers of Computer Science, 6(1), 111–121, (2012).
11. Castro, P. S., Zhang, D., Chen, C., Li, S., Pan, G.: From taxi GPS traces to social and community dynamics: A survey. ACM Computing Surveys, 46(2), 17. (2013).
12. Marsh, L., Onof, C.: Stigmergic epistemology, stigmergic cognition. Cognitive Systems Research, 9(1-2), 136–149, (2008).
13. Vernon, D., Metta, G., Sandini, G.: A Survey of Artificial Cognitive Systems: Implications for the Autonomous Development of Mental Capabilities in Computational Agents. IEEE Transactions on Evolutionary Computation, 11(2), 151–180, (2007).
14. Barsocchi P., Cimino M.G.C.A., Ferro E., Lazzeri A., Palumbo F., Vaglini, G.: Monitoring elderly behavior via indoor position-based stigmergy Pervasive and Mobile Computing, Elsevier Science, 23, 26–42, (2015).
15. Niwattanakul S., Singthongchai J., Naenudorn E., Wanapu S.: Using of Jaccard coefficient for keywords similarity. In Proceedings of the International MultiConference of Engineers and Computer Scientists, 1, 6, (2013).
16. NYC.gov, Taxi and Limousine Commission (TLC) Trip Record Data, http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml
17. Zola, interactive map for zone and land use of NYC. http://maps.nyc.gov/doitt/nycitymap/template?applicationName=ZOLA
18. Weather NYC: Thousands of transatlantic travellers face serious disruption caused by New York winter storm 'Juno'. The Independent. January 26, 2015.