LiveCities: Revealing the Pulse of Cities by Location-Based Social Networks Venues and Users Analysis

Alberto Del Bimbo, Andrea Ferracani, Daniele Pezzatini, Federico D'Amato, Martina Sereni
Università degli Studi di Firenze - MICC
alberto.delbimbo, andrea.ferracani, daniele.pezzatini@unifi.it

ABSTRACT

It would be very difficult even for a resident to characterise the social dynamics of a city and to reveal to foreigners the evolving activity patterns which occur in its various areas. To address this problem, however, large amount of data produced by location-based social networks (LBSNs) can be exploited and combined effectively with techniques of user profiling. The key idea we introduce in this demo is to improve city areas and venues classification using semantics extracted both from places and from the online profiles of people who frequent those places. We present the results of our methodology in LiveCities[1], a web application which shows the hidden character of several italian cities through clustering and information visualisations paradigms. In particular we give in-depth insights of the city of Florence, IT, for which the majority of the data in our dataset have been collected. The system provides personal recommendation of areas and venues matching user interests and allows the free exploration of urban social dynamics in terms of people lifestyle, business, demographics, transport etc. with the objective to uncover the real 'pulse' of the city. We conducted a qualitative validation through an online questionnaire with 28 residents of Florence to understand the shared perception of city areas by its inhabitants and to check if their mental maps align to our results. Our evaluation shows how considering also contextual semantics like people profiles of interests in venues categorisation can improve clustering algorithms and give good insights of the endemic characteristics and behaviours of the detected areas.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services* ; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Clustering, Information filtering*

General Terms

Algorithms, Human Factors, Experimentation, Urban Computing

Keywords

Smart Cities, venues classification, recommendation systems, location-based services

[1]A video of the application is available at this url: `http://vimeo.com/miccunifi/livecities`

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2577035.

1. INTRODUCTION

An analysis capable to convey to a realistic and truthful representation of a city and of the activities that take place in its areas must necessarily take into account not only human mobility but also users' socio-economic characteristics and interests distribution. Emerging social realtime systems offer an opportunity for the computation in the field of spatial data mining due to the huge amount of geo-localised data they continuously produce and that can be considered real human sensor data.

There exist a considerable number of works addressing geographical modelling of information derived from widespread LBSNs like Twitter and Foursquare. Some recent studies analyse social media streams to obtain contextual semantics for city zones and venues whilst others focus more on human mobility. In [4] user's positions are observed predicting the locations of new tweets. A sparse modelling approach is exploited which uses global, regional and user dependant topics and terms distribution in order to geo-reference topics on areas. Resources detected from geo-localised Twitter messages are also utilized to infer transient representation of volatile events happening at venues in [1]. Foursquare places categories are used to create footprints of areas and users in [5] by means of spectral clustering. At the other hand, as regard to more focused works on urban computing, in [2] check-ins are used to understand mobility patterns and how these are influenced by users' social status, sentiment and geographic constraints. In the Livehoods project Cranshaw et al. [3] cluster Foursquare venues using spatial and social proximity introducing a new user-based 'bag-of-chekins' similarity algorithm. Although their approach is effective in capturing the social dynamics of cities according to people movements, it is completely lacking in considering who those people are and which are their motivations.

The key idea we propose in LiveCities instead is that city venues are characterisable both by static features, i.e. categories assigned by LBSNs on the basis of their type of service, and by dynamic features, i.e. the distribution of the interests of the people who checked-in there, which can

change over time. To accomplish this we extract users' profiles of interests and users' geo-localised media automatically from Facebook, then we categorise detected venues using Foursquare APIs and, finally, we weights these features on the basis of semantic similarities and interests distribution. The main contribution of the work is to present our clustering module for city areas identification and classification based on our features selection approach and to show the web application developed for clusters visualisation and venues recommendation.

2. THE SYSTEM

2.1 Dataset

Through a Facebook app we have collected and gained access to 8839 user profiles, from which we extracted 124790 checkins and identified 52767 venues. Location information is available on Facebook from 2010. Facebook Places started out as a mobile application for people to check into business locations, then it was integrated in Facebook featuring a location tagging tool. People on Facebook can tag specific locations in status updates, image posts, or video posts. Others members can also tag their Facebook friends in specific locations within their updates and posts. Since the most part of the people registered in the application is resident of Florence and its surroundings we chose to conduct our evaluation on this city. The data used for the tests consists of 24031 check-ins and 5321 venues in Florence. Considered that Florence population counted 366443 in January 2013[2] this is a large amount of information. Places were identified in updates, post and events in which the users participated and photographs they were tagged in. Each place has been categorised using the Foursquare API to assign a static label representing the venue's macro-category. As for profiling, users' interests were extracted by retrieving the categories of Facebook pages for which users expressed a 'like'. There are total 398884 'likes' distributed in 216 Facebook categories. User's data is the main reason for which we chose the Facebook APIs to build our dataset instead of the Foursquare or Twitter APIs, commonly used by works in the field[6] [1] [4] [3] [2]. In this respect we can say that Facebook offers, in addition to check-ins data, a higher degree of contextual awareness and an 'environment' exploitable to enrich check-ins semantics.

2.2 Clustering module

LiveCities uses k-means clustering to partition the venues dataset into k groups. We run the algorithm on the features selected on the basis of the main idea of this work that people semantics and semantic distances can be exploited to refine places categorisation. Clustering was performed for each city with similarity distances based on different features:

- **Geographic**: latitude and longitude;
- **Foursquare based**: latitude, longitude, Foursquare venue's category;
- **Socially aware**: latitude, longitude, Foursquare venue's category, a weighted vector of interests of the users who checked-in.

These three modalities of features selection have been essential in order to conduct the evaluation and to measure the improvements of our approach (i.e. socially aware). One of the very first problem we have to tackle in our data is that Facebook 'likes' categories show an unbalanced distribution. The reason is that some interests like "music" or "sport" are more commonly shared between users than others and that Facebook pages in these categories are more widespread.

To solve this issue, we calculate the weight of a category of 'likes' on a venue considering three factors: 1) percentage of 'likes' in each category for all the people who checked-in, 2) probability of a generic 'like' to belong to a category, 3) semantic distance between each 'likes' category and the assigned Foursquare category. Formally, supposing we have a vector F of i_F Facebook places and also a set of L users' 'likes' for each venue, denoting as c a 'likes' category, we can compute the weight w for each $c \in i_F$ as follows:

$$\mathrm{w}(c, i_F) = \mathrm{percentage}(c, i_F) \cdot \log_{10}\left(\frac{10}{P(c)}\right) \cdot \mathrm{correlation}(c, i_F)$$

The function uses *de facto* a TF-IDF approach. With $P(c)$ we mean the probability in 2) calculated and normalised on the basis of the distribution of the category 'likes' in all the dataset 'likes'. The correlation function instead uses a semantic distance to compute the affinity between 'likes' categories and the Foursquare venues. Distances are pre-calculated and obtained using the Wikipedia Link-based Measure (WLM) by Milne et al. [7]. WLM is a measure for the estimation of the semantic relatedness of two Wikipedia articles through the comparison of their links. In our dataset there is a total number of 216 Facebook categories for pages and 397 types of Foursquare venues, this means that it was necessary to calculate 85752 correlations. To accomplish this, every resource (Facebook category or venue type) has been associated to a corresponding Wikipedia article. We experimented two approaches: 1) manual association, 2) using the MediaWiki API to retrieve possible articles' matching titles and filtering the results using Latent Semantic Analysis (LSA). Both gave almost the same accuracy. There are two version of the WLM algorithm, the first considers in-bound links and is modeled after the Normalized Google Distance, and the other uses out-bounds links and is defined by the angle between the vectors of the links found within the two articles calculated with the cosine similarity. In LiveCities we re-implemented the algorithm in the latter version because less computationally expensive. To improve the correlation measure, we also observed that when two resources have an high semantic relatedness, often one of the two article contains a link to the other. When this condition occurs, we add a *bonus* to the correlation value.

2.3 User interface, personalization and recommendation

LiveCities features a web application based on the principles behind visual analytics for dynamically exploring time-varying, localised and multivariate attribute data relative to city venues and venues customers. LiveCities provides a map based interface and exposes advanced visual components intended to maximise 1) explorative data analysis and 2) service targeting and personalisation.

The application provides two main views, a search view and a clusters view. The search view has been designed as a traditional geographic search interface for venues and it allows users to efficiently filter data by categories or by people interests on the map. The cluster view instead visualises

[2] `http://demo.istat.it/bilmens2013gen/index.html`. Istat data, January 2013

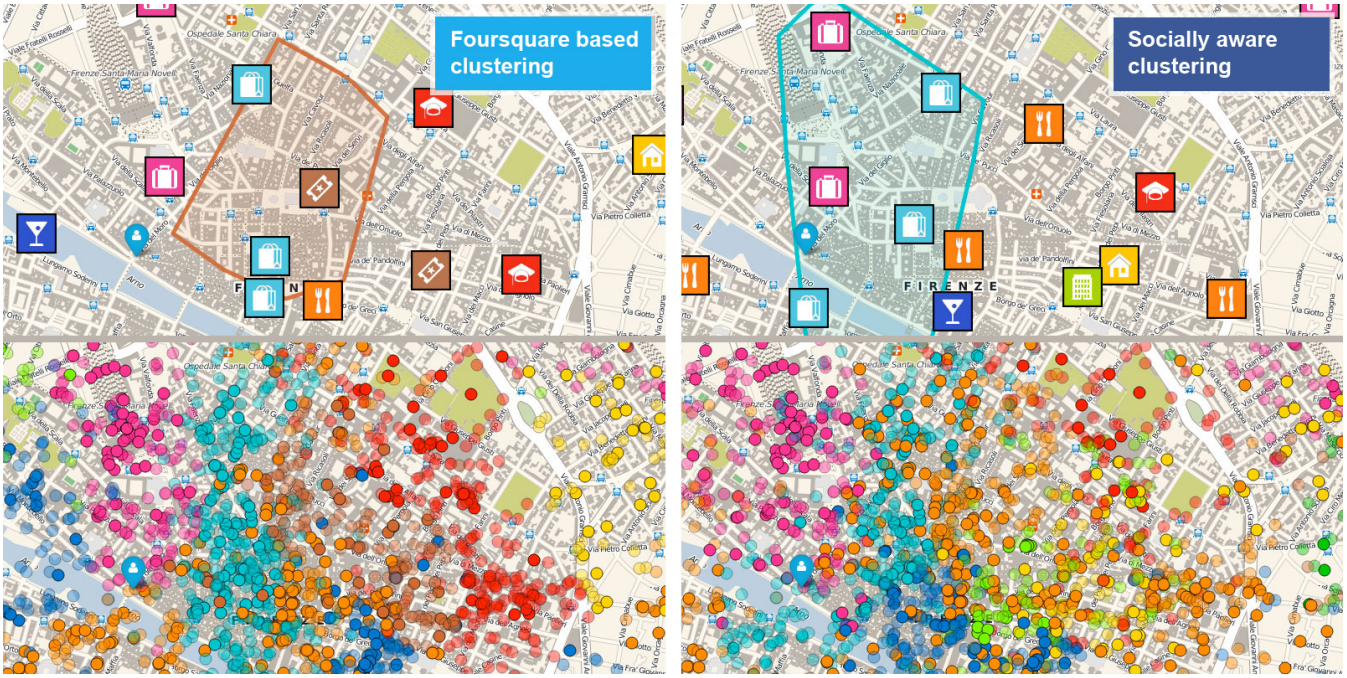

Figure 1: LiveCities clusters visualisation of Florence, IT. The figure shows a comparison between the clustering visualisation based on Foursquare categories and the results of our methodology that considers people interest distribution (Socially aware clustering).

Figure 2: 1) Insights of a cluster, showing the histogram of venues categories and 2) the distribution of people interests on a venue.

the results of the k-means algorithm. There are three types of visualisation on the basis of three different features selections: 1) geographic, 2) Foursquare-based, 3) socially aware (our approach which takes into account people interests and semantic distances), cfr. Fig. 1. Clusters can be visualised as typed squared icons or as set of points. The squared based visualisation uses icons as representative of the 'centers of mass' of the detected clusters and allows a less bulky visual access to the information, whilst the points based view show on the map all the venues in the dataset.

Clusters are characterised by different colors, each one corresponding to 9 general Foursquare categories. Points transparency is directly proportional to the computed semantic affinity of the venue category to the cluster classification. In this way colour information is exploited in order to effectively depict points distribution *per* cluster. Clusters boundaries are visualized on user interaction hovering with the mouse over the map, and are calculated using the convex hull algorithm. Users can have statistic insights on clusters and venues through an interactive tooltip, cfr. Fig. 2. In particular cluster's insights present the histogram of venues categories in the cluster and, for each column, the actual geo-referenced venue's place. Venue's insights show the distribution of interests of people who checked-in and provide address details and routing. Stars (from 1 to 3) on columns and venues represent recommended resources. LiveCities provides Facebook Login and it profiles users evaluating their Facebook 'likes' on pages, obtained with the Facebook APIs. Recommendation of areas and venues in LiveCities tries to maximise an objective function

$$\max_{p \in places} \mathrm{f}(p, logged\_user)$$

The f estimates the correlation between the user profile of interests and the characteristics of city areas and venues. The semantic relatedness is computed using the WLM measure and weighting suggestions on the basis of users affinity with area's categories and individual venues.

3. RESULTS AND EVALUATION

A preliminary estimation of the results has been conducted for the city of Florence comparing outputs from the three different clustering procedures. We created an online questionnaire with the intent of receiving feedback from city residents about how they perceive the different areas of the city. The questionnaire shows users a map of the city, divided into 15 numbered cells. For each cell, we asked the

users to assign labels, according to their mental maps, selecting up to three different categories among those used by LiveCities. We collected answers from 28 users, among 20 and 56 years old and for the most part affirming to have sufficient, good or excellent knowledge of the city (only 4% of the interviewed declared to have an insufficient knowledge). Since clusters shapes are irregular, a single cell can comprehend one or more clusters. On this basis we evaluate how interviewed people labeling of city areas aligns with detected clusters measuring the displacement in the weights of its venues categories. Let A_n be the area of predefined cells adopted in the questionnaire, with $n \in [1, 15]$, we consider the set of clusters OC_n that have some overlapped area with A_n. Formally, for each geographical cluster C_i with $i \in [1, K]$, where K is the number of output clusters of k-means algorithm, $C_i \in OC_n$ only if $A_n \cap C_i \neq \varnothing$. Clusters are described with a multi-dimensional vector formed by weights w_{cat} for every category of the system, with $0 \leq w_{cat} \leq 1$. We define the vector that describe OC_n by computing mean values of the clusters contained in OC_n. We use the data obtained by the questionnaire, represented as a vector of categories weights for every area A_n, as testing data. We can so calculate the Mean Squared Error (MSE) between the expected values (weights in A_n) and the predicted values (weights in OC_n). As an example, figure 3 shows intra-categories MSE of each of the three clustering methods for the cell A_{14}. We repeat those steps for every n in order to obtain a global MSE of every clustering method (i.e. geographical, foursquare based and socially aware). The results are the following:

MSE_{geo}	0.059
$MSE_{foursquare}$	0.062
MSE_{social}	0.046

Results show that the MSE in the socially aware clustering approach is lower than with the other ones. Even if the conducted study is still preliminary, results may suggest that our method tend to reflect more correctly the perception that inhabitants have about the characteristics of city areas.

4. CONCLUSIONS

LiveCities is a web application designed to provide users with a dynamic view of the social patterns characterising city areas and to facilitate resident and visitors in finding places and zones likely to be of interest. Urban computation can have a lot of applications, from marketing to trade area analysis, buildings design, urban planning, demographics, entertainments, or simply citizens' life practice. LiveCities offers pictorial depictions of cities and exploits information visualisation techniques in order to shed new light on cities inner workings and on the relationship between people and the environments which they inhabit. In turn it can help to reveal the real 'fabric' cities are woven out. In this demo we showed our methodology for features selection and clustering. We use k-means in order to group venues on the basis both of topological and sociological features. With sociological features we mean that venues are somehow representable not only by their static category assigned by LBSNs but also by the 'bag-of-interests' of the people who checked-in. We also presented the web interface as well as the recommendation and personalisation module. Finally we conducted a preliminary evaluation through an online questionnaire. Results are encouraging and show that our approach deserves to be deepened and that LiveCities can be an useful web tool to suggest to users how to enjoy the best of the places in which they live.

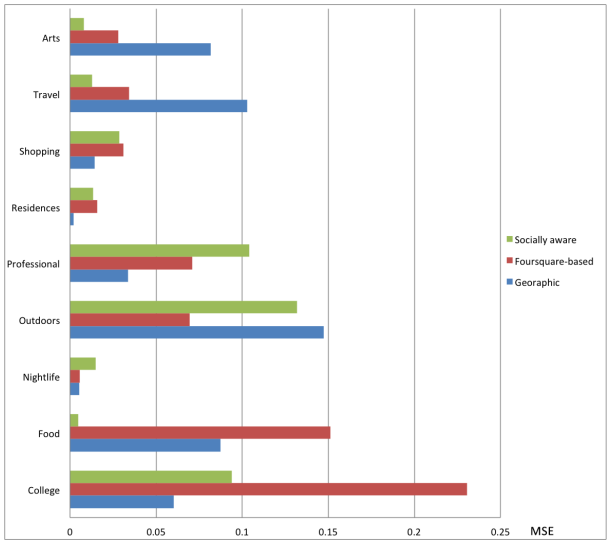

Figure 3: Comparison of the MSE in every category for each clustering approach in a case study area of the city.

5. REFERENCES

[1] A.-E. Cano, A. Varga, and F. Ciravegna. Volatile classification of point of interests based on social activity streams. In *In Proceedings of the 10th International Semantic Web Conference, Workshop on Social Data on the Web (SDoW)*, 2011.

[2] Z. Cheng, J. Caverlee, K. Lee, and D. Z. Sui. Exploring Millions of Footprints in Location Sharing Services. In *Proceedings of the Fifth International Conference on Weblogs and Social Media*, Menlo Park, CA, USA, July 2011. AAAI.

[3] J. Cranshaw, R. Schwartz, J. I. Hong, and N. M. Sadeh. The livehoods project: Utilizing social media to understand the dynamics of a city. In *Proceedings of the Sixth International AAAI Conference on Weblogs an Social Media (ICWSM 2012)*, 2012.

[4] L. Hong, A. Ahmed, S. Gurumurthy, A. J. Smola, and K. Tsioutsiouliklis. Discovering geographical topics in the twitter stream. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 769–778, New York, NY, USA, 2012. ACM.

[5] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil. Exploiting semantic annotations for clustering geographic areas and users in location-based social networks. In *The Social Mobile Web*, 2011.

[6] Y. Qu and J. Zhang. Trade area analysis using user generated mobile location data. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 1053–1064, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

[7] I. H. Witten and D. Milne. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, AAAI Press, Chicago, USA*, pages 25–30, 2008.