User Profiling for Urban Computing: Enriching Social Network Trace Data

Andrea Ferracani, Daniele Pezzatini, Alberto Del Bimbo
Università degli Studi di Firenze - MICC
Firenze, Italy
[name.surname]@unifi.it

ABSTRACT

Location-Based Social Networks (LBSNs), with their huge amount of geo-located user generated content, are providing a lot of semantics on human mobility and behaviour as well as on users' interests and activities in cities. In this paper we propose an innovative approach to detect city zones and reveal city dynamics which exploits clustering techniques based on an original feature selection. We also present the results in LiveCities[1], a web application designed adopting new information visualisations paradigms in order to easily get cities' insights. Recommendation of city zones and venues close to user's interests, based on semi-automatic user profiling, is also provided exploiting semantic similarity algorithms. Results, validated by a case study on the city of Florence (Italy) through an online questionnaire filled out by residents, show that our feature performs better than traditional approaches.

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

Mobility patterns, social analysis, collective intelligence, urban computing

Keywords

Venues classification, clustering, recommendation, location-based services

1. INTRODUCTION

The large volume of information produced in realtime by LBSNs is becoming an essential means to depict and model the geographical space people move in. Furthermore, LBSNs provide a lot of contextual semantics about users who document, in social media streams, their momentary mood, activities and behaviours. The 'check-in' action as well as the possibility to add a location to a media content have become a typical online activity that reflects a real interaction between the user and the real world. These online activities enable data scientists to build an interpretative model of reality.

1.1 Related Work

There is a considerable number of works that address the problem of modelling geographical information analysing data from the most popular LBSNs such as Facebook, Twitter, Foursquare, Gowalla. It has already been demonstrated how user's check-in activity can be analysed and exploited in recommendation systems as it can provide information about the user's interests distribution, for example detecting check-ins latent topics [1]. Anagnostopoulos et al. [2] also emphasize how the social circles are a primary factor to be taken into account in modelling geo-localised activities since human movements are commonly influenced by their social context. In this respect Gao et al. propose a social and historical model that integrates the two constraints (social circles and check-ins time series) with the aim to understand and measure the relationship and the impact of the social user activities on the 'check-ins' pattern [3]. Hong et al. [4] mine the text of geo-located tweets in order to predict the location of the next tweet. For this purpose, a 'sparse modelling' approach is used which exploits global, regional and typical user's topics to define and geo-reference the probability of occurrence of the same topics on a geographical map. Entities extracted automatically from Twitter messages are used to geo-locate temporary events that are occurring in some places by Cano et al. [5]. Tweets are categorised on the basis of the detected entities using DBPedia categories structure. The work focuses the attention on the importance of categorising venues dynamically over the time analysing media streams. Foursquare's categories are employed by Noulas et al. [6] to characterize regions and users by means of spectral clustering. Cheng et al. [7] use check-ins from several social networks to understand the patterns of mobility (user displacement, radius of gyration and probability of return) and how these are influenced by the user's social status, the sentiment and the geographical constraints. Cranshaw et al. [8] cluster, via k-means, Foursquare venues using their 'social proximity'. The authors introduce a new algorithm of similarity between venues where each venue is represented by the vector containing the number of check-ins

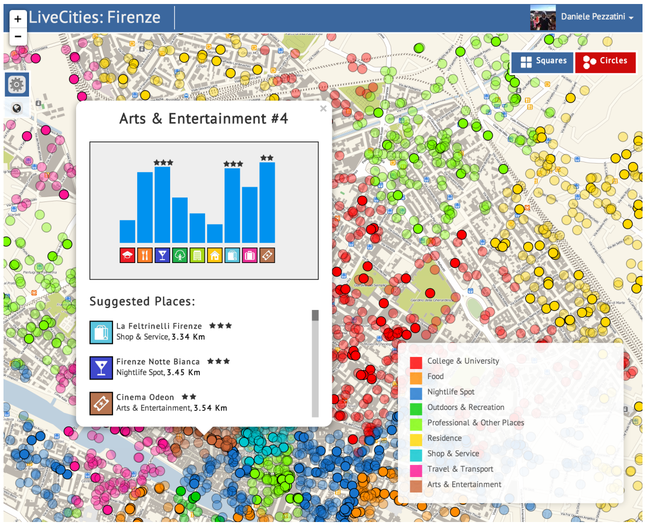
[1] Video available at http://vimeo.com/miccunifi/livecities

MM'14, November 3–7, 2014, Orlando, Florida USA.
Copyright 2014 ACM 978-1-4503-3127-2/14/11 ...$15.00.
http://dx.doi.org/10.1145/2661118.2661122 .

of each user at the venue. Although this approach is effective in capturing the dynamics of the city considering the movement of people, however, it does not address the problem of the classification and labelling of the clusters of venues and completely ignores who these people are and which are their motivations in moving.

As evidenced in the related works, in order to improve the quality of the features for clustering city venues, a good strategy is to exploit the additional semantics explicitly provided in LBSNs check-ins. The main contribution of this work is to show how also implicit contextual semantics, extracted from Facebook, can be used in order to improve this quality and, consequently, to better capture what is going on in city zones and to understand city evolution in the short term. In the field of urban computing a common venues' clustering approach is the use of places' categories, extracted from user's check-ins, as features in order to identify communities and urban neighbourhoods within cities. The key idea we present instead is that the interests of people who attend a place over the time have also to be considered in characterising venues. Thus, venues can be connoted both by the categories assigned by LBSNs on the basis of the type of service they provide as well as by the profiles of interest of the place goers. Results are shown in LiveCities, the web-based tool we provide, which is a fully-fledged application for smart cities [9]. In addition to identify and classify city zones, LiveCities includes advanced visualisation tools to give detailed insights on city regions, venues categories, and people interests distribution; furthermore, it exposes a search view, optimised routing mechanisms, and an accurate recommendation system based on semantic measures.

The rest of the article is organised as follows: Section 2 describes how data was collected, the socially aware feature we propose and the clustering algorithm. Section 3 presents the LiveCities web application used to visualised data insights and results. System evaluation is provided in Section 4. Conclusions are drawn in Section 5.



Suggested Places:

La Feltrinelli Firenze ★★★
Shop & Service, 3.34 Km

Firenze Notte Bianca ★★★
Nightlife Spot, 3.45 Km

Cinema Odeon ★★
Arts & Entertainment, 3.54 Km

Figure 1: Clusters of venues visualised as points on the map. The cluster interactive tooltip show the histogram of venues' categories and the recommended categories and venues (if the user is logged-in)

2. DATA, FEATURE SELECTION AND CLUSTERING

Users' profiles and geo-localised data have been collected from Facebook in three months from September to December 2013. Users' interests have been extracted analysing demographic infos and explicit or implicit interests. These are inferred analysing categories of page 'likes'. Venues have been detected in status updates, posts, events or photos where users tagged themselves or were tagged in by friends, and have been categorised using the Foursquare APIs. The dataset counts a total of 398884 'likes' distributed in 216 Facebook categories. 8839 user profiles have been collected, 124790 check-ins extracted and 52767 venues identified. Data have been filtered in order to keep only the check-ins localised in the italian territory and restricted to the major cities for which we had a sufficient number of check-ins to be clustered. City regions have been identified using k-means clustering on the venues detected in the check-ins dataset. The novelty of our approach is not the similarity measure exploited to group places but the choice and the computation of the subset of relevant features to be used in model definition. People profiles and semantic similarities are used to refine venues categorisation. Clustering is run with three different feature selection in order to have a means to validate the results and to highlight visual differences in relating views: 1) geographic: lat and long; 2) Foursquare based: lat, long, Foursquare venue's category; 3) socially aware: lat, long, and a special feature which considers the semantic similarity between Foursquare venue's category and a weighted vector of interests of place goers. To compute this socially aware feature each category of 'likes' for a venue has been weighted considering three factors:

- total 'likes' percentage of goers' interests,
- normalised probability of a 'like' to belong to a category,
- semantic relatedness between 'likes' category and the place's Foursquare category calculated using the Wikipedia Link-based Measure (WLM) [10].

Formally, we compute the weight w of each category of user 'likes' c for each venue i_V as follows:

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

$P(c)$ denote the probability of a like category to appear on the basis of the distribution of 'llikes' in the dataset. The semantic correlation between 'like' category and venue's Foursquare assigned category is obtained using WLM. The WLM is a measure to compute the semantic relatedness of two Wikipedia articles comparing their in-going or out-going links. To this end, every resource's category in the system has been associated with the corresponding Wikipedia article (there are 216 Facebook categories representing users' interests and 397 types of Foursquare venues). The association has been automatically achieved using MediaWiki APIs[2] in order to discover possible page matches. Since the MediaWiki APIs can return multiple and ambiguous search results, these have been filtered exploiting Latent Semantic Analysis (LSA) to identify the most relevant occurrence.

[2] http://bit.ly/RywgdI

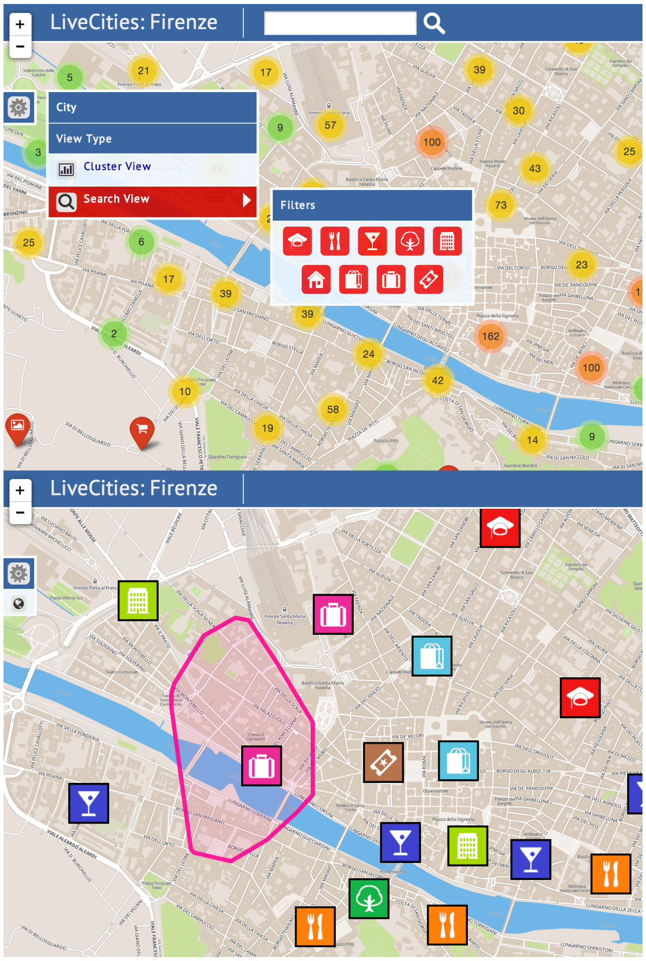


Figure 2: The search view and the cluster view in LiveCities

3. THE LIVECITIES WEB INTERFACE

LiveCities is a web application, which consists of an interactive map-based visualisation of clusters and venue insights that allow users to identify city's areas and to give them a quick understanding of their characteristics and dynamics. Data was initially collected for some italian cities, although the system exposes a Facebook Login through which new user data and check's can be extracted. Thus, new cities are identified automatically by the system with the increase of the application's users. As soon as city's data are retrieved and city's check-ins reach an adequate size, the relative daily clustering job for venues is added. The application provides two modalities of visualisation: a search view and a cluster view, cfr. Fig. 2.

The search view adopts a standard visualisation of markers on a map, where each marker corresponds to a venue and is represented with an icon suggesting its category. In order to reduce the number of markers shown on the map, small clusters of venues are visualised for each zoom level. Users can perform text-based search and filter venues by category or by their own interests in the profile. By clicking on a venue, users can visualise some details such as the category of the place and the average distribution of interests of the users who checked-in the venue. Furthermore, the application exploits Mapquest APIs[3] to provide routing directions from user's current position (obtained via browser capabilities) to the selected venue. The cluster view visualises the results obtained by the k-means algorithm. It's possible to choose between three different clustering approaches provided by the application and based on the different selection of features: geographic, Foursquare-based and socially aware. Clusters can be visualised as typed squared icons or as a set of points. The squared-based visualisation uses category-related icons as representative of clusters' centroids and allows a more intuitive visual access to the information, whilst the points-based view shows on the map all the venues in the dataset. Clusters and points are characterised by different colours, assigned correspondingly to the estimated category of the cluster. 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. Users can obtain detailed insights on clusters by an interactive tooltip. Cluster's insights present the histogram of venues' categories in the cluster and, for each column, which acts as a filter on the corresponding category, the list of the geo-referenced venues. These can be ordered by distance or, if the user is logged in, by the automatically assigned number of stars. Stars represent the affinity of the venue with the user's profile of interests (cfr. Fig. 1). Facebook Login is exploited in order to profile users, evaluating their Facebook 'likes' on pages, obtained with the Facebook Graph APIs [4]. When a logged user visualises cluster's informations, stars (from 1 to 3) are also shown above the category columns to emphasise the level of relatedness of each category with the user's interests. Recommendation, information filtering and similarity estimation between city zones, venues and user profiles are carried out exploiting the WLM measure.

4. RESULTS AND EVALUATION

We conducted a preliminary evaluation of the results obtained by the clustering algorithm in order to asses if the proposed feature selection method performs better than other common approaches. Briefly, the adopted method consisted of asking citizens of the city of Florence, IT how they would 'label' city areas according to their perception and then to compare the results with the clustering output for each of the proposed feature (geographical, Foursquare based, Socially aware). To collect the data we created an online questionnaire. The questionnaire shows a map of the city of Florence divided into 15 numbered cells corresponding to city areas. For every area, participants are asked to assign up to three different labels selected among the categories used by the system. We acquired answers to the questionnaire from 28 participants, aged between 20 and 56 years old, most of them claiming to have a good knowledge of the city (less than 5% of the participants declared to have an insufficient knowledge).

City areas corresponding to the grid's cells in the online questionnaire are defined as A_n with $n \in [1, 15]$. Areas are predefined and cannot exactly match the city zones identified by the clustering system, since the clustering produce k areas of polygonal shape, where k is the input pa-

[3] http://www.mapquestapi.com/

[4] http://bit.ly/1mAJRfT

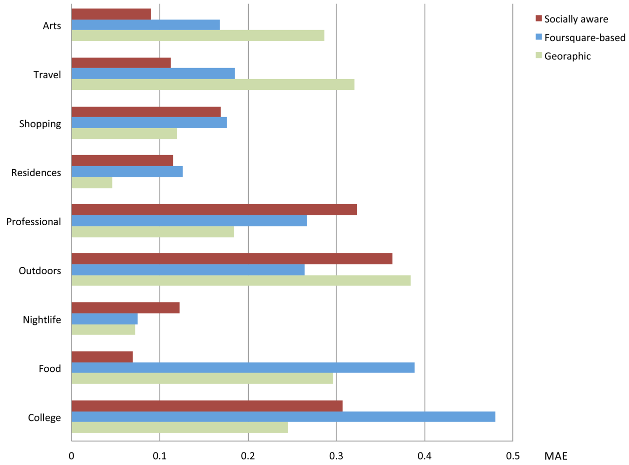


Figure 3: Comparison of the MAEs for each feature selection approach in every category.

rameter of k-means algorithm. Therefore, for each area A_n we have to consider a set of overlapping clusters OC_n. Let's name the geographical clusters obtained using the algorithm as C_i with $i \in [1, k]$. A cluster $C_i \in OC_n$ only if $A_n \cap C_i \neq \varnothing$. Both computed clusters and questionnaire's areas labelled by users are described with a multi-dimensional vector $\mathbf{w} \in \mathbb{R}^9$ formed by weights w_{cat} for each of the main categories 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.Given the vector $\mathbf{wa_n}$ of weights for the area A_n and the vector $\mathbf{woc_n}$ of weights of OC_n defined as above, we can then compute the global Mean Absolute Error (MAE) as follow:

$$MAE = \frac{1}{15}\sum_{n=1}^{15}\frac{1}{9}\sum_{c=1}^{9}|wa_{n_c} - woc_{n_c}|$$

Results show that the proposed feature selection for clustering city venues (socially aware), based on implicit semantics extracted by user profiles, has a lower error rate than the other two (geographic and Foursquare based):

MAE_{geo}	0.217
$MAE_{Foursquare}$	0.236
$\mathbf{MAE_{social}}$	**0.185**

Figure 3 shows MAEs for every category of the system in order to compare the three different approaches.

5. CONCLUSION

In this paper we present a contribution in the field of feature engineering. A new feature which exploits user profiling on Facebook and semantic distances is proposed and evaluated for modelling geographical information derived from LBSNs. City venues extracted from Facebook and categorised with the Foursquare's API are clustered using k-means for three different features based respectively on geographic infos, Foursquare labelling and the proposed adoption of social semantics. A qualitative evaluation is conducted through an online questionnaire with residents of Florence who have been asked to label several zones of their city. The evaluation shows that the proposed feature performs better than the other approaches. Results and data insights are also provided in LiveCities, a map-based web application which exploits infovis techniques in order to allow users to browse categorised city zones and venues and to obtain personalised recommendations.

Acknowledgments. The authors thank Federico D'Amato and Martina Sereni for their help in the project development.

6. REFERENCES

[1] Alberto Del Bimbo, Andrea Ferracani, and Daniele Pezzatini. Flarty: recommending art routes using check-ins latent topics. In *Proc. of ACM International Conference on Multimedia (MM)*, pages 457–458. ACM, ACM Press, 2013.

[2] Aris Anagnostopoulos, Ravi Kumar, and Mohammad Mahdian. Influence and correlation in social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 7–15, New York, NY, USA, 2008. ACM.

[3] Huiji Gao, Jiliang Tang, and Huan Liu. Exploring social-historical ties on location-based social networks. In *ICWSM*, 2012.

[4] Liangjie Hong, Amr Ahmed, Siva Gurumurthy, Alexander J. Smola, and Kostas 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-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.

[6] Anastasios Noulas, Salvatore Scellato, Cecilia Mascolo, and Massimiliano Pontil. Exploiting semantic annotations for clustering geographic areas and users in location-based social networks. In *The Social Mobile Web*, 2011.

[7] Zhiyuan Cheng, James Caverlee, Kyumin Lee, and Daniel Z. Sui. Exploring millions of footprints in location sharing services. In *ICWSM*, 2011.

[8] Justin Cranshaw, Raz Schwartz, Jason I. Hong, and Norman M. Sadeh. The livehoods project: Utilizing social media to understand the dynamics of a city. In John G. Breslin, Nicole B. Ellison, James G. Shanahan, and Zeynep Tufekci, editors, *ICWSM*. The AAAI Press, 2012.

[9] Gang Pan, Guande Qi, Wangsheng Zhang, Shijian Li, Zhaohui Wu, and L.T. Yang. Trace analysis and mining for smart cities: issues, methods, and applications. *Communications Magazine, IEEE*, 51(6):120–126, June 2013.

[10] Ian H Witten and David 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.