Targeted Interest-Driven Advertising in Cities Using Twitter*

Aris Anagnostopoulos
Sapienza University of Rome
aris@dis.uniroma1.it

Fabio Petroni
Sapienza University of Rome
petroni@dis.uniroma1.it

Mara Sorella
Sapienza University of Rome
sorella@dis.uniroma1.it

Abstract

Targeted advertising is a key characteristic of online as well as traditional-media marketing. However, it is very limited in outdoor advertising, that is, performing campaigns by means of billboards in public places. In this work we propose a methodology for performing targeted outdoor advertising by leveraging the use of social media. In particular, we use the Twitter social network to gather information about users' degree of interest in given advertising categories and about the routes that they follow. Given an advertising category, we estimate the most promising areas to be selected for the placement of an ad that can maximize its targeted effectiveness.

1 Introduction

A famous quote by John Wanamaker, a pioneer in advertising, states: "Half the money I spend on advertising is wasted. The trouble is, I don't know which half." Although social networks and media have been of tremendous help for online advertising, leading it to a high degree of targeting and tailoring, its outdoor counterpart has only relied on traffic data and rough demographic estimates: other than allowing for limited targeting power, these strategies select highly crowded areas for billboards, leading to an over-cluttering effect where the attention of customers, exposed to a high number of co-occurring ads, is lost. This lack of verified data on audience characteristics, has reportedly (Shimp and Andrews 2013) limited the growth of the outdoor-advertising industry, preventing many advertisers from investing heavily in it. Yet, there should be ways that one could perform targeting, something which is currently most probably performed manually by outdoor advertising companies. We tackle the problem proposing a new technique, which leverages public information from Twitter: we collect tweets' geotags to obtain information about user trajectories and then we perform user profiling to identify the degree of interest of each user towards different topics corresponding to a predefined set of advertising categories. Intuitively, interests drive the way in which users are influenced by an ad: we combine this information with the collective mobility patterns of users sharing the same interests, to estimate, for each category, the most promising areas to place a relevant ad. To assess the quality of the solution we perform validation on a test portion of the users to verify if those users, interested in some topics, will or not pass by the corresponding identified zones (thus having a chance to see the targeted ad). Furthermore, we use mobile communications usage data to measure how crowded is each zone, using it both as baseline, and to understand the difference between the zones found by our algorithms and the simply crowded areas. Our results show that even with a low budget in terms of the number of zones in which we can place an ad, we are able to cover a consistently higher portion of the users with respect to top crowded areas, for all the categories. Furthermore, we found some anecdotal evidence of the targeted interestingness of the discovered zones, both suggesting a possibly higher influencing effect, and giving insights on the applicability of this approach.

2 Related Work

User profiling in Twitter A key element of our approach is understanding the interests of Twitter users. Many past works rely either on the the text of the tweets issued by a user himself, or the users he follows. Early work of this kind are based on bag-of-words and statistical approaches (Chen et al. 2010), while more recent works use topic modeling techniques such as Latent Dirichlet Allocation (LDA) and its derivatives. In (Wagner et al. 2012) the authors test different types of user-related information to test if they convey interest-specific information and found that bio and list membership are the most discriminative to identify topical interests/expertise. In (Bhattacharya et al. 2014) the authors exploit lists meta-data to find experts and interests.

Urban computing using geotagged data Among the works that leverage geotagged data of particular interest are the tasks of finding local experts on Twitter (Cheng et al. 2014), and characterizing city areas such as neighborhoods in terms of the local activities (Cranshaw et al. 2012; Le Falher, Gionis, and Mathioudakis 2015). The latter two works are the most relevant to ours, and exploit data from location-based services. Nevertheless it may be the case that

*This research was partially supported by the Google Focused Research Award "Algorithms for Large-Scale Data Analysis" and by the EU FET project MULTIPLEX 317532.

some people can frequent an area for latent reasons that cannot simply be captured by the venues or point of interests contained therein. Furthermore, data from location-based services is generally not public.

Outdoor advertising Outdoor advertising forms a crucial part of marketing science, and naturally it has attracted a very large attention by researchers in the area (e.g., (Woodside 1990; Osborne and Coleman 2008)). To the best of our knowledge our work is the first to make use of social media data to perform targeted outdoor advertising.

3 Interest-Driven Urban Zone Ranking

We have as input a set of geotagged tweets T made by a set of Twitter users U during a given time period and a fixed set of categories I to which both user interests and ads conform. For instance, we may have $I = \{\mathit{Food}, \mathit{Cinema}, \mathit{Sports}, \ldots\}$. Furthermore, we assume that we have enough budget to select k zones for targeting (e.g., $k = 10$) meaning that we may target 10 city zones for a given category by placing billboards. We partition the area spanned by the tweets into a set of n squared, non-overlapping city zones $Z = \{z_1,\ldots,z_n\}$, such that each tweet's coordinates included in the geotags belong to a single zone. Denote by $Z_u \subseteq Z$ the set containing all the zones where user $u \in U$ has issued at least one tweet, which we refer to as the *trace* of the user. For each category $i \in I$, our goal is to compute a ranked list of zones, and to provide the top-k zones Z_i^* that will be the candidates for targeted advertising. Intuitively, the rank of a zone for a given category should reflect the expected effectiveness (in terms of interested users reached) of an ad placed in that zone.

3.1 Methodology

Our approach consists of two components: (1) a method to identify the interest of the users towards the identified category set I and (2) a procedure to find, for each category, the ranking over the city zones. We describe them next.

Inferring user interests To get information about the interests of Twitter users, we use a technique similar to the one of Ghosh et al. (Bhattacharya et al. 2014), which we explain next. We exploit Twitter *lists*, an organizational feature of Twitter, which allows users to create and manage curated lists of other users. Each list is characterized by a name and an optional description. Lists are mainly used to group followed or simply popular accounts under topical themes. For instance, a user can create a list called "*Music and Bands*," and add accounts such as @YahooMusic, @radiohead, or @katyperry. Given a target user $u \in U$, we obtain the set F_u of all the users he follows. The objective is to categorize each followed user $f \in F_u$ into some *topics*, using the lists in which f was (possibly) added by some other Twitter user. To this end, for each user $f \in F_u$ we gathered all lists containing f: we refer to this set as L_f. We consider as topic all unigrams and bigrams, composed by only nouns and adjectives[1], found in all the descriptions and names of each list $l \in L_f$, rejecting all topics that do not appear in at least 10 lists. Furthermore, we keep only the top 100 most frequent topics for user f, and we manually classify them in the categories set I. We then associate to f a set of categories $I_f \subseteq I$, such that for each $i \in I_f$ there is at least a topic classified in the corresponding category i. We therefore informally consider user f as *expert* (or *authority*) in each category $i \in I_f$ (see Figure 1).

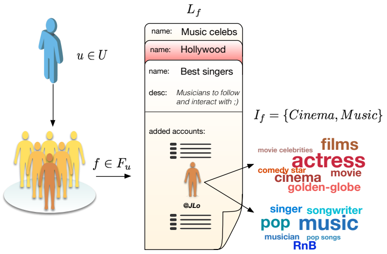

Figure 1: First step of the process of interest inference for a user u. For each followed user f we gather all the list containing f and look for top occurring topics (right).

We make the assumption that, the more experts u follows on a certain category, the more he is likely to be interested in that category. We denote as $E_u^i = \{f \in F_u : i \in I_f\}$ the set of users followed by u who are expert on category i. Finally, we associate with the original user u an $|I|$-dimensional vector $interest_u$ of interest scores, one for each considered interest category. The score of each user relative to a specific category j will be the fraction of experts on category j she follows, normalized over all followed experts:

$$\mathit{interest}_u[i] = \frac{|E_u^i|}{\sum_{j \in I} |E_u^j|}.$$

Top-k Zone Ranking In this phase we compute a ranking over the considered zones for each category so as to select the most promising locations for advertising. Intuitively, our approach is to use the user traces and project the amount of users' interest towards the different interest categories on the various city zones, thus exploiting the power of this collective signal to drive the ranking. Let $U_z = \{u : z \in Z_u\}$ identify the users u who passed through zone z and $\mathit{Freq}(u,z)$ be the number of geotagged tweets issued by user u in zone z.

Given an interest category i, a zone z, and the set U_z of the users that have the zone in their traces, we evaluate the *targeted effectiveness* of the city zone z for category i in four different ways, leading to four different approaches:

- *All*: We sum the $\mathit{interest}_u[i]$ scores, of *all* users $u \in U_z$.
- *Primary*: Sum of $\mathit{interest}_u[i]$ scores, considering only users u for whom i is the category of *primary* interest (the highest score in the interest vector of the user corresponds to category i), namely, the set of users $\overline{U}_z = \{u \in U_z : \mathit{interest}_u[i] \geq \mathit{interest}_u[j], \forall j \neq i\}$.
- *AllFreq*: Sum of the product of $\mathit{interest}_u[i]$ scores and the number of geotagged tweets by each user u in zone z.

[1] as recognized by a standard part-of-speech (POS) tagger

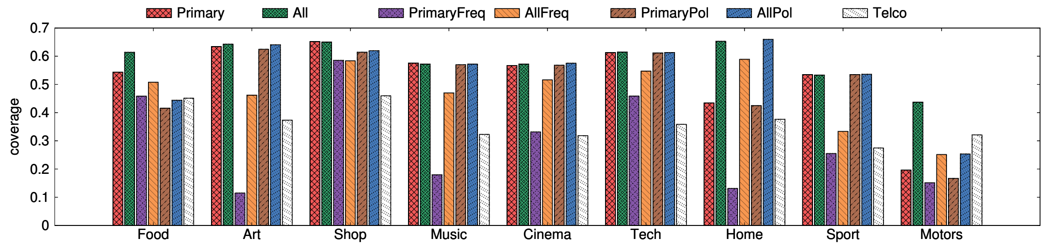

Figure 2: Coverage with $k = 10$.

- *PrimaryFreq*: Like *AllFreq*, but considering only the users for which i is the category of *primary* interest.

The first two algorithms don't take into consideration the actual frequency of the tweets in the cell, but just the number of distinct users. For each algorithm, we then output a ranked list of the top-k zones Z_i^*, for each category $i \in I$.

Polarization In the ranking, each category is treated as independent. Nevertheless, our goal of finding a ranking of the best zones could benefit from a criterion that also considers how the interests *overlap*, decreasing the score of zones frequented by users with many different interests. In other words, it can be interesting to consider the *polarization* of zone z towards category i, defined as the relative portion of users interested in category i that the zone z could attract in the future, given the observed data. We estimate it with a Beta distribution. Concretely, for each zone z and category i we keep a Beta(α_z^i,β_z^i), where the parameter α_z^i is the number of users $u \in U_z$ interested in category i (plus one) and β_z^i is the number of users $u \in U_z$ not interested in category i (plus one). We derive a polarization score for each zone (and each category) as the mean of the beta distribution, discounted by 4 standard deviations to penalize the scores of zones for which we do not have enough information. We refer to versions of the aforementioned algorithms that explicitly consider the polarization effect as *AllPol* and *PrimaryPol*.

4 Experimental Evaluation

In this section we describe the datasets used for the evaluation of the proposed solution, then we discuss our results and some interesting properties of the zones found.

4.1 Experimental Setup

Datasets Our main dataset is a collection of geotagged tweets gathered from the Twitter Firehose for the two-month period of November and December 2013 obtained specifying as boundary region (or *bounding-box*) the city of Milan, Italy. It consists of a total of 477,913 tweets by 31,356 users. By restricting the set of users to those having at least 10 tweets, we end up with 404,077 tweets and 5,086 total users. The city zones $z \in Z$ are composed by square cells, each of $235m^2$, for a total of 10,000 cells. For each zone we also possess mobile telecommunications usage data spanning the same two months observation period for each cell[2].

Evaluation Methodology We consider 9 different categories, namely, $I = \{$*Food, Art–Photography, Shopping–Fashion, Music, Cinema–TV, Technology, Home–Design, Sport, Motors*$\}$. To evaluate the rankings found by our techniques we identify, for each category $i \in I$, the set $U_i \subset U$ of representative users as the users whose corresponding interest score for interest category i is greater than 0 (i.e., users interested in that topic). We evaluate our approach performing a 5-fold cross-validation. The information about what zones are selected in practice for advertising is not publicly available. As a baseline to compare with our approach, we consider a strategy that selects the most crowded zones (e.g., train stations, main squares). To estimate the crowdedness we compute the average daily communication activity per zone. We refer to this baseline approach as *Telco*. As we mentioned previously, we consider the trace of a user as a proxy for his movements. To evaluate the performance of all algorithms described (*Primary*, *All*, *PrimaryFreq*, *AllFreq*, *PrimaryPol*, *AllPol*, and *Telco*), for each interest category i, we compute their *coverage* metric, the fraction of users in the test set that have a positive interest in category i who passed in at least one of the *top-k* zones in the solution, that is,: Cov $= \frac{1}{|U_i|} \sum_{u \in U_i} x_{u,i}$ where $x_{u,i}$ is 1 if $Z_u \cap Z_i^* \neq \emptyset$, and 0 otherwise. A high coverage indicates that the selected zones are good spots to place an advertising, because a high number of interested users can be potentially reached. Moreover, we also make use of a metric to compute the similarity (i.e., the overlap) between pairs of solutions of different Z_i^* and Z_j^* where i and j are different ad categories. In particular, we use the Jaccard similarity index, defined as ratio of the size of the intersection to the size of the union of two solutions.

4.2 Evaluation Results

Performance Figure 2 shows coverage values of the solutions for all the categories, for a fixed budget $k = 10$. All algorithms that ignore the *frequency* of the tweeting activity (i.e., *Primary*, *All*, *PrimaryPol* and *AllPol*) achieve a good coverage, outperforming the baseline solution *Telco* by a consistent margin in all the considered categories. Algorithms *PrimaryFreq* and *AllFreq*, instead, perform poorly.

[2] All the datasets were made available by the Telecom Big Data Challenge 2014 international competition.

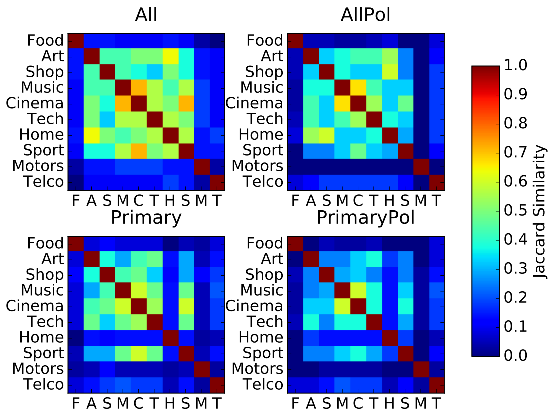

Figure 3: Jaccard similarity matrices of the rankings obtained by the frequency-agnostic algorithms ($k = 10$).

A possible explanation for this behavior is that these algorithms boost the ranking of users' *everyday zones* (e.g., work, gym, favorite bars), which are often peculiar to the specific user and may not be a good indicator for a global perspective. To mitigate this issue, we filtered the *home* location for each user (which we assume, for simplicity, to be the zone where she tweets the most). This doesn't completely solve the problem. Moreover, algorithms based on tweet frequency are affected by a common bias underlying the use of geotagged tweets as a proxy for user movements: some spots are more suitable for the tweeting activity (e.g., bars, parks, rest places). By taking into account frequency, these zones are even more privileged in the ranking. Next, we investigate the extent to which the zones in the top-k solutions (Z_i^*) overlap between different categories. We compute the Jaccard similarity index between pairs of top-k solutions (i.e., Z_i^* and Z_j^*, for $i,j \in I$) of frequency agnostic algorithms, for $k = 10$, shown in Figure 3. The results confirm our intuition: *Primary* is able to better differentiate the solutions among different categories—we can observe that the similarity between different categories is lower compared to *All*. The overlap can be further reduced by considering the polarization of the zones, i.e., *AllPol* and *PrimaryPol* algorithms), boosting the rank of zones where the interest in a specific category is significant with respect to the others.

Anecdotal results Figure 4 shows the actual top-10 zones identified by *All* for each category in the city of Milan. We can see that such zones do not fall exclusively in the city center, but they span the entire considered area. To support a qualitative assessment, in Figure 4 are highlighted some zones, present in the solution and containing very important and relevant venues (e.g., Triennale exhibition for the Art category). By manual inspection we found other less obvious, yet relevant places (crosshatched squares); in clockwise order: (1) Technology - Computer Eng. building (Politecnico); (2) Food - InKitchen Loft (cooking lessons); (3) Sports: public sport camps of Via Dezza; (4) Music - Rock&Roll (live music pub).

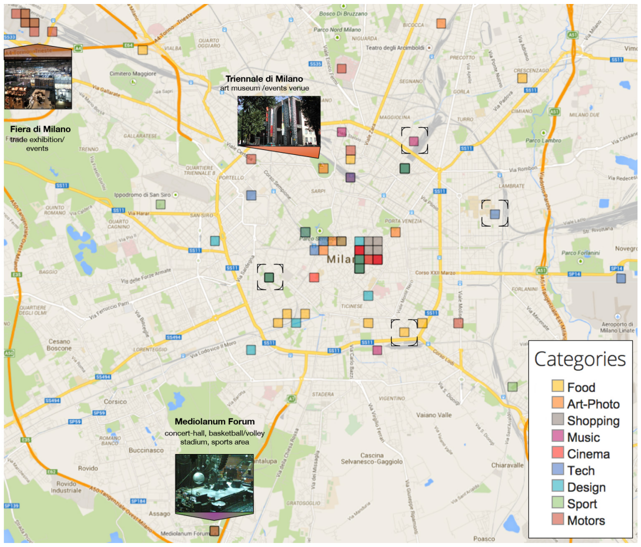

Figure 4: Actual top-10 zones identified by the *All* algorithm for each considered category in the city of Milan.

5 Conclusions

This work presents a novel method for offline targeted advertising in urban contexts, making use of public data from the Twitter social networks. Our technique indicates that we can potentially achieve a higher level of targeting with respect to a baseline method relying on crowd estimates: this potentially enables more effective advertising campaigns with better budget allocation.

References

Bhattacharya, P.; Zafar, M. B.; Ganguly, N.; Ghosh, S.; and Gummadi, K. P. 2014. Inferring user interests in the twitter social network. In *Proc. RecSys 2014*.

Chen, J.; Nairn, R.; Nelson, L.; Bernstein, M.; and Chi, E. 2010. Short and tweet: experiments on recommending content from information streams. In *Proc. SIGCHI 2010*.

Cheng, Z.; Caverlee, J.; Barthwal, H.; and Bachani, V. 2014. Who is the barbecue king of texas?: a geo-spatial approach to finding local experts on twitter. In *Proc. SIGIR 2014*.

Cranshaw, J.; Schwartz, R.; Hong, J. I.; and Sadeh, N. M. 2012. The livehoods project: Utilizing social media to understand the dynamics of a city. In *Proc. WWW 2012*.

Le Falher, G.; Gionis, A.; and Mathioudakis, M. 2015. Where is the soho of rome? measures and algorithms for finding similar neighborhoods in cities. In *Proc. ICWSM 2015*.

Osborne, A. C., and Coleman, R. 2008. Outdoor advertising recall: A comparison of newer technology and traditional billboards. *Journal of Current Issues & Research in Advertising*.

Shimp, T., and Andrews, J. C. 2013. *Advertising promotion and other aspects of integrated marketing communications*. Cengage Learning.

Wagner, C.; Liao, V.; Pirolli, P.; Nelson, L.; and Strohmaier, M. 2012. It's not in their tweets: modeling topical expertise of twitter users. In *PASSAT, SocialCom 2012*.

Woodside, A. 1990. Outdoor advertising as experiments. *Journal of the Academy of Marketing Science*.