Tracking User Activities and Marketplace Dynamics in Classified Ads

Muhammad Waqar [1], Davood Rafiei [2]

Department of Computing Science, University of Alberta
Edmonton, AB, Canada
[1] mwaqar@ualberta.ca
[2] drafiei@ualberta.ca

***Abstract*—Tracking users' posting activities in online classified ads and understanding the dynamics of their behavior is a topic of great importance with many implications. However, some of the underlying problems associated with modeling users and detecting their behavioral changes due to temporal and spatial variations have not been well-studied. In this paper, we develop a probabilistic model of user behavior based on the ads the user posts and the categories in which the ads are posted. The model can track some of the temporal changes in behavior, as revealed by our experiments on two classes of users monitored over a period of almost a year. We study the association between post categories and user groups, and show how temporal and seasonal changes can be detected. We further investigate a generative model for ad posts, based on user locations, and provide some evidence showing that the model is promising and that some interesting relationships can be identified.**

I. Introduction

The tremendous growth of the World Wide Web has been driving individuals away from traditional print ads towards online classified advertising. Tracking the users behaviour in posting classified ads and possibly predicting their actions can provide important information for both such sites and the applications that are built on top; it may also help in better understanding of the marketplace dynamics in classified ads.

There is a growing body of work on tracking topics [1], quotes, and news pieces [2], [3] as they evolve or spread over time. There is also past work on user modeling in general and activity tracking in particular, such as navigation between pages [4], [5]. However, classified ads are somewhat different from both information networks (e.g. the Web) and social networks (e.g. Facebook, Twitter, etc.) in that there is no direct tie between social actors (e.g. users or ads), and the social acts (if they can be called so) are in the form of weak ties (e.g. users who post to the same category). We are not aware of much study of the aforementioned issues in the context of classified ads; more specifically we could only find a limited number of work studying the social and economic impact of online classified ads (e.g. [6]).

In this paper, we present a probabilistic model of user behavior based on the interactions between users, ads and post categories. The user behavior can change over time and those changes are often triggered by external forces; we show how changes can be tracked and the collective behavior of users with similar interests can be identified. We further study the changes in behavior due to spatial variations and that how a generative model may predict ad production based on the location from which the ad is posted. We evaluate our models through experiments conducted on data collected from a real classified ad site, and report some of our results and findings.

Dataset The experiments in this paper use a dataset of ad postings from Edmonton Kijiji (http://edmonton.kijiji.ca) collected over a period of 9 months from May 2013 to January 2014. Each user account in Kijiji is associated with a unique id, which makes it easy to link each ad to the user who posted it. The granularity of our crawl was set to a day. Figure 1 shows the distribution of collected ads in different categories. Unless explicitly stated otherwise, all experiments in this paper use data from the two largest categories *buy and sell* and *cars & vehicles*. Each user in our dataset is also classified into *business* and *non-business* groups depending on the nature of their use of the classified ad medium (see [7] for details of the classification); we use these classes when we study the behavior of a group of users. We plan to make this data available to the research community.

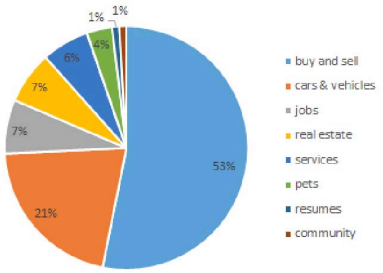

Fig. 1: Distribution of the ads in various categories

II. Temporal Changes in User Behavior

In a classified ad networks, a user can have multiple ads each listed under different categories. Hence the set of categories in which a user posts defines a probability distribution.

Definition 1: **(User Profile)**: Given a set of categories C and a set of users U, let $p_{u,c}$ be the probability that user $u \in U$ posts an ad in category $c \in C$. The posting profile of u can be defined in terms of the distribution of his ads in different categories, i.e.

$$P(u) = \{(c, p_{u,c}) \mid c \in C\}$$

978-1-5090-4470-2/16 $31.00 © 2016 IEEE
DOI 10.1109/WI.2016.87

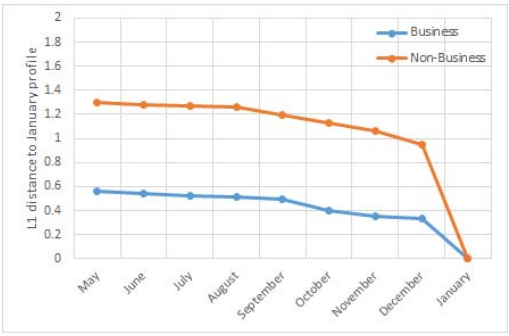

Fig. 2: Temporal changes in user profiles

To quantify temporal changes in user behavior, we place a sliding window over user postings and construct a profile for each window. Treating a profile as a population distribution, we measure the evolutionary distance between two profiles in terms of the change that is necessary to transform one distribution into another. We measure this change in terms of the mean absolute difference of the two distributions; this corresponds to the Manhattan (or L_1) distance which is used in similar contexts [8], [9]. It is computed between two distributions $\vec{x}$ and $\vec{y}$ as

$$d(\vec{x}, \vec{y}) = \sum_i |x_i - y_i|.$$

To validate the proposed model, we studied the changes in the profiles of business and non-business users. For each user (who posted at least one ad every month from May 2013 to January 2014), we assessed the change in profile for each month to that of the last month of our dataset (January 2014). The results are shown in Figure 2. As expected, the distance is zero for the last month since the profiles that are being compared are the same. We notice that non-business users exhibit continuously stronger changes in their profiles than businesses. This conforms to our intuition that private individuals tend to use the classified ad network only when a particular need arises, and as a result, their ads are likely to be scattered in different categories based on the nature of their needs at the time. On the other hand, businesses use the network for advertising their products (or services). Hence, we would expect them to be consistent in their postings in the categories related to their enterprise. We also observe that for both user groups, the L_1 distance gradually decreases over time; this is simply because the user needs (and consequently their posting behaviors) are less likely to change in shorter time intervals.

III. Distinctive Categories for User Groups

Given a set of users with possibly some commonalities, we want to know some of the distinctive features in terms of the posting behavior (in this case, the ad categories) shared by the members. Suppose l denotes a class label, U is the set of all users and $U_l \subset U$ is all users with label l. The probability of a post in a category c by a user can be expressed as

$$P(c \mid U) = \frac{P(c)P(U \mid c)}{P(U)}$$

where $P(c)$ is the probability of a user post falling into category c, and can be estimated as the mean of the fraction of user posts into category c. Similarly, $P(U \mid c)$ is the fraction of users who have listings under c. Likewise, we can express the probability of a post in a category c by a user in U_l as

$$P(c \mid U_l) = \frac{P(c)P(U_l \mid c)}{P(U_l)}.$$

Again, we can estimate $P(c)$ using the mean of the fraction of user posts into category c over users U_l, $P(U_l \mid c)$ with the fraction of users in U_l who have listings under c, and $P(U_l) = |U_l|/|U|$.

The importance score of c to group U_l, as determined by its contribution to the KL-divergence score between the two probability distributions $P(c \mid U_l)$ and $P(c \mid U)$, is:

$$I(c, U_l) = P(c \mid U_l) \ln \frac{P(c \mid U_l)}{P(c \mid U)}.$$

The categories having larger such values are the most distinctive for the respective user group (U_l).

Table I lists the top-12 distinctive categories thus obtained for the two classes of users: *business* and *non-business*. We observe that there are significant and meaningful differences between the two user groups. Particularly, for businesses, we notice the presence of a large number of *service-oriented* categories such as "home renovation services", "computer services" and "cell phone services" and other *business-oriented* categories including "heavy equipment", "business/industrial" and "cargo trailers".

It can also be observed that most of distinctive categories for businesses and non-businesses belong to *cars & vehicles* and *buy and sell* respectively. This is because of an overwhelming presence of non-business users in *buy and sell* category. Due to this fact, the only such categories which are determined as most distinctive for business users are the ones that are inherently service-oriented or business-oriented in nature (as mentioned previously).

IV. Temporal Changes in the Categories for User Groups

In this section, we study how distinctive categories for user groups changes over time. Let $P(c \mid U_l)$ be the marginal probability of an ad category c over the entire duration of the dataset and $P_x(c \mid U_l)$ be the same probability over a time interval x. These probabilities can be derived as discussed in Section III. An importance score of category c at interval x can then be computed as

$$I_x(c, U_l) = P_x(c \mid U_l) \ln \frac{P_x(c \mid U_l)}{P(c \mid U_l)} \quad (1)$$

giving the degree at which postings in c for a particular user group U_l differ for interval x.

We apply the model to determine the most popular categories for each month covered by our dataset as compared to its entire time span for both our user groups. Table II shows the top three categories for each month, projected over the categories shown. The results reveal interesting trends. Specifically, for non-business users, we observe that books

S. No.	Business	Non-Business
1.	(bs, tickets)	(cv, auto parts/tires, tires/rims)
2.	(cv, cars & trucks)	(bs, books)
3.	(bs, business/industrial)	(bs, electronics)
4.	(cv, RVs/campers/trailers, cargo/utility trailers)	(bs, art/collectibles)
5.	(cv, ATVs/snowmobiles, ATV parts/trailers/accessories)	(bs, toys/games)
6.	(cv, home renovation materials, cabinets/countertops)	(bs, other)
7.	(bs, phones, cell phone services)	(bs, phones/tables)
8.	(cv, RVs/campers/trailers, RVs/motorhomes)	(bs, phones, cell phones)
9.	(cv, heavy equipment, other)	(bs, jewellery/watches)
10.	(bs, furniture, beds/mattresses)	(bs, clothing, women's - tops/outerwear)
11.	(cv, motorcycles, motorcycle parts/accessories)	(bs, home - indoor, home decor/accents)
12.	(bs, computer accessories, services (training/repair))	(cv, auto parts/tires, other parts/accessories)

TABLE I: Distinctive categories for user groups. "bs" and "cv" denote *buy and sell* and *cars & vehicles* respectively.

Categories	M	J	J	A	S	O	N	D	J
Business									
(bs, tickets)	-	-	-	-	*	*	*	*	*
(cv, auto parts/tires, other parts/accessories)	-	-	-	-	-	-	-	-	*
(cv, ATVs/snowmobiles, ATVs parts/trailers/accessories)	-	*	-	-	-	-	-	-	-
(cv, RVs/campers/trailers, RVs/motorhomes)	-	-	*	-	-	-	-	-	-
(cv, RVs/campers/trailers, cargo/utility trailers)	-	-	-	-	*	-	-	-	-
(cv, RVs/campers/trailers, travel trailers/campers)	*	-	-	-	-	-	-	-	-
(cv, auto parts/tires, tires/rims)	-	-	-	-	-	*	*	-	-
Non-Business									
(bs, books)	-	-	-	*	*	-	-	-	*
(bs, clothing, costumes)	-	-	-	-	-	*	-	-	-
(bs, clothing, women's - tops/outerwear)	-	-	-	-	-	-	*	-	-
(bs, garage sales)	-	*	-	*	-	-	-	-	-
(bs, jewellery/watches)	-	-	-	-	-	-	-	*	-
(bs, tickets)	-	-	-	-	-	-	-	*	-
(cv, ATVs/snowmobiles, snowmobiles)	-	-	-	-	-	-	-	-	*
(cv, RVs/campers/trailers, travel trailers/campers)	*	*	-	-	-	-	-	-	-
(cv, auto parts/tires, tires/rims)	-	-	-	-	-	*	*	-	-

TABLE II: Distinctive categories over time. Column headers M-J represent months from May 2013 to January 2014. "bs" and "cv" denote *buy and sell* and *cars & vehicles* respectively. The ranks are reported only for top few categories in each interval.

category experiences a strong surge during months that coincide with the beginning of Fall and Winter terms in the universities. Likewise, costumes category sees a higher than usual traction during the month of October, which concurs with the Halloween season. Similarly, some other clothing categories became popular with the user group around the same period. Likewise, some seasonal trends can be identified. For example, garage sale category experiences great activity at the onset of the summer season in Edmonton; jewellery and watches become especially popular during Christmas; tires/rims category witnesses abnormally high number of postings at beginning of winter in Edmonton (October/November). Similarly, recreational categories also reveal interesting seasonal trends. We observe that the ticket category becomes active in winter months most probably due to interest in Edmonton Oilers games. We can also note that travel trailers, campers, RVs attract a lot of attention during summer while snowmobiles become popular with the users during winter.

Many of the trends described above can also be noticed for the business users. For example, licensed ticket sellers are most active during the Oilers hockey season; automobile technicians and businesses post a large number of ads in tires/rims category at the commencement of winter season offering their services; activity in RVs, motorhomes, travel trailers and campers categories sees an upward jump during summer.

V. Distinctive Categories for Locations

Our goal, in this Section, is to identify the categories that are the most distinctive or unusual for a particular location.

Following the approach taken by Backstrom et al. [10] for search engine queries, we model the ads posted in various categories as Bernoulli trials; the trial is a success if an ad is posted in a category c. Let p be this probability, computed as the fraction of overall ads posted in c, and t_x be the total number of ads posted from a specific location x. Considering that the posting of ads in categories is independent of each other, the probability of s_x successes (i.e. the probability of seeing that many posts in category c from location x) is:

$$P(X = s_x) = \binom{t_x}{s_x} p^{s_x} (1-p)^{t_x - s_x}$$

A low probability value for a location x and category c indicates that the observed frequency significantly differs from the global background distribution of c, making c a distinctive category for location x.

For the experiment, we used all categories in Kijiji. We divided the city into 38 neighborhoods (Figure 3), with each neighborhood starting with a specific 3-digit postal code. To obtain the appropriate neighborhood for each ad, we utilized the *address* attribute in each ad; ads not having postal code (nearly 24%) were ignored.

Table III shows the top few categories for some neighborhoods. We observe that those in locality T6G are the ones

Categories	T5H	T5J	T5K	T5S	T5V	T6C	T6G	T6P	T6S
(bs, books)	-	-	*	-	-	-	*	-	-
(bs, furniture, beds/mattresses)	-	-	-	-	*	-	-	-	-
(bs, phones, cell phones)	-	-	-	-	-	-	*	-	-
(bs, tickets)	*	*	*	-	-	*	-	-	-
(re, apartments/condos, 1 bedroom)	*	-	*	-	-	-	-	-	-
(re, apartments/condos, 2 bedroom)	*	-	-	-	-	-	-	-	-
(re, house rental)	-	-	-	-	-	*	-	-	-
(re, houses for sale)	-	-	-	-	-	-	-	*	-
(re, room rental, roommates)	-	-	-	-	-	-	*	-	-
(jobs, bar/food/hospitality)	-	*	-	-	-	-	-	-	-
(jobs, construction/trades)	-	-	-	*	*	-	-	*	*
(jobs, customer service)	-	*	-	-	-	-	-	-	-
(jobs, driver/security)	-	-	-	*	-	-	-	*	*
(jobs, general labour)	-	-	-	*	*	-	-	*	*
(jobs, office mgr/receptionist)	-	*	-	-	-	-	-	-	-
(jobs, sales/retail sales)	-	*	-	-	-	-	-	-	-

TABLE III: Distinctive categories for various neighborhoods. “bs”, “cv” and “re” stand for *buy and sell*, *cars & vehicles* and *real estate* respectively. The ranks are reported only for the top few categories in each neighborhood.

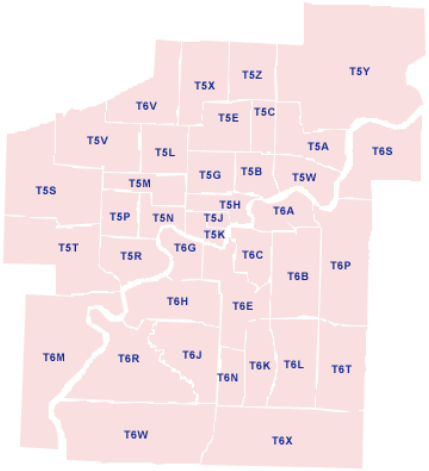

Fig. 3: Edmonton’s Postal Code Map (Canada Post, 2001)

in which we would expect great activity by the students (for example phones, books, tablets and room rental). This is not surprising since due to presence of University of Alberta here, the area is inhabited by many students. Similarly, books category is also popular in the nearby T5K community due to the presence of MacEwan University in T5J neighborhood.

Moreover, T5J area experiences many ads from jobs categories. This is because it represents Edmonton Downtown, the hub of city’s business activities. Thus, as it can be easily imagined, there is always a high demand for accommodations around this neighborhood. Appropriately, we find that the nearby T5H, T5K and T6C areas attract a lot of postings from room rental, aparments/condos and house rental categories.

Finally, we observe that the nature of jobs required in Edmonton’s downtown area (T5J), which is also the center of the city, is office-oriented i.e., desk jobs. On the contrary, as we move towards the outskirts of the city (T6P, T6S, T5S, T5V), we notice that the jobs become more physically demanding (e.g. construction and general labor) showing construction and new housing activities.

VI. Conclusions

We have presented models of user behavior in an online classified ad network based on user postings, and have shown that the models can track temporal and spatial variations in user behavior. We have also experimented with our models in a real setting, with the ads posted in the Kijiji network, showing some of the patterns in user behavior.

This work can be extended in a few directions. One direction is to study nearby spatial or temporal patterns and detect boundaries that may better describe the behavior of the users. Another direction is to include other ad attributes in the study; one particularly important one is the price. Finally detecting other groupings of users and tracking their behavior may reveal areas where our models can be augmented.

Acknowledgments

This research is supported by the Natural Sciences and Engineering Research Council of Canada.

References

[1] J. Allan, *Topic detection and tracking: event-based information organization*. Springer-Verlag, 2012.

[2] J. Leskovec, L. Backstrom, and J. Kleinberg, “Meme-tracking and the dynamics of the news cycle,” in *Proc. of the KDD Conference*, 2009, pp. 497–506.

[3] S. Myers, C. Zhu, and J. Leskovec, “Information diffusion and external influence in networks,” in *Proc. of the KDD Conference*, 2012, pp. 33–41.

[4] C. Chen, K. Wu, V. Srinivasan, and X. Zhang, “Battling the internet water army: Detection of hidden paid posters,” in *Proc. of the ASONAM Conference*, 2013, pp. 116–120.

[5] R. Atterer, M. Wnuk, and A. Schmidth, “Knowing the users every move user activity tracking for website usability evaluation and implicit interaction,” in *Proc. of the WWW Conference*, 2006, pp. 203–212.

[6] K. Kroft and D. G. Pope, “Does online search crowd out traditional search and improve matching efficiency? evidence from craigslist,” *Journal of Labor Economics*, vol. 32, no. 2, pp. 259–303, 2014.

[7] M. Waqar and D. Rafiei, “Characterizing users in an online classified ad network,” in *Proc. of the Web Intelligence, Mining and Semantics Conference (WIMS)*, 2016, pp. 28:1–28:9.

[8] G. R. Carvalho, *Advances in Molecular Ecology*. IOS Press, 1998.

[9] F. Abel, Q. Gao, G.-J. Houben, and K. Tao, “Analyzing user modeling on twitter for personalized news recommendations,” in *User modeling, adaptation, and personalization*. Springer, 2011, pp. 1–12.

[10] L. Backstrom, J. Kleinberg, R. Kumar, and J. Novak, “Spatial variation in search engine queries,” in *Proc. of the WWW conference*, 2008, pp. 357–366.