AAAI Technical Report WS-12-04
When the City Meets the Citizen

Using Social Media to Infer Gender Composition of Commuter Populations

Wendy Liu and Faiyaz Al Zamal and Derek Ruths
School of Computer Science
McGill University
Montreal, Quebec Canada

Abstract

In order for a municipality to effectively service and engage its constituency, it must understand the composition of the communities within it. Up to the present, such demographic estimates for target populations have been obtained largely from census data or expensive, time-intensive surveys.

In this paper, we use Twitter microblog content to estimate the gender makeup of commuting populations using different modes of transportation (cars, public transportation, and bikes) in Toronto, Canada. We apply a demographic inference algorithm to $33,215$ public Twitter accounts that follow one of three popular transportation-related Twitter-based news feeds (one for traffic, one for public transportation updates, and one for bicycling). Recent census data provides ground truth against which to compare the estimates we derive from Twitter.

We find that, for all three communities (car drivers, public transport users, and bicyclists), the estimates obtained from Twitter reflect the majority-minority relationships between genders reported in census data. This provides preliminary, but compelling evidence that Twitter may be a platform that can go beyond simply signaling the presence of physical communities to actually measure their compositions.

Introduction

An important objective for city administrators is to tune services and support to suit the citizens who use them. Among other things, this requires an understanding of the composition of communities that engage with municipal resources. Beyond adapting to gradual shifts in demographics, the idea of the "smart city" embodies the notion that a municipality needs to adapt its services and resources to even the hourly changes in community composition, distribution, and needs. Thus, city service administrators seek to be able to continually reassess the composition of communities present within the city. Unfortunately, in most urban centers administrators do not having access to such real-time information feeds, relying instead on figures from census data or costly, time-intensive surveys. New approaches are needed.

Here we investigate the use of social media data (in particular microblog content) to characterize different urban populations that engage directly with municipal resources. The merits of such an approach are three-fold. First, microblog data can provide up-to-date information: Twitter microblog data is generated continuously and, therefore, makes it possible to quickly update demographic estimates. Second, microblog data can provide relatively inexpensive measurements: compared to surveys which require personnel to collect and process the data, microblog content is collected and processed computationally, which significantly diminishes personnel and resource costs. Third, microblog data can provide estimates with geographical specificity: as geotagging becomes increasingly prevalent, the insights gained from mining microblog data will gain an increasingly precise spatial component.

This paper presents a first effort to understand whether such an approach can be successful. We use Twitter microblog data to estimate the gender breakdown of different types of commuter populations: car drivers, public transportation users, and bicyclists. The overarching approach is to obtain three sets of Twitter accounts belonging to users who (with high-likelihood) commute using cars, public transport, and bikes, respectively. We then apply a demographic inference algorithm to these users' microblog data and assess the overall gender breakdown of each set (Zamal, Liu, and Ruths 2012).

In the current study, we applied this approach to Toronto's commuting population. We identified three Twitter accounts dedicated to broadcasting information about traffic, public transportation issues, and bicycling and then used the community of Twitter followers for each account as the sets of users for demographic inference. Census data from 2006[1] provides ground truth against which to compare our microblog-based estimates (Statistics Canada 2007). We find that for each commuter group of Twitter users, the gender in the majority corresponds to the majority gender reported in census data.

These results demonstrate that microblog data may be fruitfully used to assess the general breakdown of certain demographic features within urban populations.

[1]The complete data from the 2011 census has not yet been released to the public.

Group	Twitter Account	Profile Description	# Followers	# NP Followers
Car	@680trafficnews	*Get up-to-the-minute Toronto & GTA traffic updates with 680News.*	6,947	4,290
Public	@ttcnotices	*The official Twitter page of the Toronto Transit Commission, including service advisories. Questions or comments about the TTC are welcome!*	27,522	19,393
Bike	@bikeunion	*a voice for all Toronto cyclists. Safe streets, a healthy city, a vibrant voice*	3,193	2,334

Table 1: Basic statistics on the three Twitter news accounts used to obtain online proxy populations for car (@680trafficnews), public transport (@ttcnotices), and bike (@bikeunion) commuters. The final column, *# NP Followers*, provides the number of *Non-Protected* followers with at least 100 tweets. These were the users whose data we used in this study.

Related Work

Attribute inference. Our study leverages work on the problem of user latent attribute inference in online social media. Twitter users, in particular, have frequently been analyzed using the content of their microblogs (Rao et al. 2010; Pennacchiotti and Popescu 2011; Conover et al. 2011; Burger, Henderson, and Zarrella 2011; Rao and Yarowsky 2010; Zamal, Liu, and Ruths 2012). Recently there has been work that augments microblog content with neighbor information (Zamal, Liu, and Ruths 2012). Our work here uses rather than innovates on these existing methods.

Mapping online to physical populations. We know of two studies which have studied the connection between the features of online and physical communities (Mislove et al. 2011; Quercia et al. 2012). Mislove et al. (2011) estimated the gender and ethnic composition of Twitter users in the United States by county and compared their results to census data. While their study indicated some correlation between the two, their analysis as concerning the present study is limited because their estimates were based on single indicators of demography (i.e., names with strong gender or ethnic association) and their analysis covered the entire United States rather than specific urban regions, making it difficult to apply their results to understanding particular regional populations. Quercia et al (2012) considered the relationship between mood indicators for Twitter communities and their physical extensions. Noteworthy correlations were found.

Measuring urban populations through online social media. To our knowledge, this study is the first to approach the question of quantitatively studying specific urban communities through large-scale, computational analysis of online social media content. Even broadening our search to qualitative studies of this question, we were unable to find closely related work. In light of this and the significance of the central question, we consider this paper to be an important first step towards characterizing the promises and challenges in this area.

Data and Methods

In this section, we first describe the process by which Twitter data was collected for each commuting population. We then briefly describe the demographic inference method used.

Data

For each commuter population, we identified popular Toronto-specific Twitter accounts dedicated to broadcasting news pertaining to interests specific to that commuter population (see Table 1); note that while we would have processed multiple accounts for each population of interest, in all three cases we found only one large Twitter account that performed the desired purpose. @680trafficnews is a Twitter feed of updates on traffic conditions around Toronto; we made the assumption that individuals who follow traffic news tend to commute using a car. @ttcnotices provides updates on public transportation issues in the Toronto area; here, as with @680trafficnews, we assumed that followers are regular users of public transport (making them likely to use public means of getting to and from work). Finally, @bikeunion is an information hub for the cycling community of Toronto. Admittedly, @bikeunion has a weaker tie to its intended commuting population (bicyclists) than the others since bicycle enthusiasts (who commute by other means) may also follow this Twitter account. Nonetheless, of the available Twitter feeds we surveyed, it had the best fit to our target population. Furthermore, results reported later support this choice.

For each account in Table 1, the profile and most recent 1000 tweets were obtained for each follower with a public account. As seen in the final column of Table 1, the overwhelming majority of users had unprotected accounts.

In each commuter Twitter data set, the gender of 400 users was manually coded (200 males and 200 females) using user profile, profile picture, username, and tweet content. These manually coded users were used to train and test the accuracy of the demographic inference classifier described next.

Census data. Data from the 2006 Canadian Census provides figures for the gender breakdown of car, public transport, and bicycle commuters in the greater Toronto area. Data from the 2011 census would have been preferred; unfortunately the figures of interest have not been published yet. The figures used as ground truth in this study are shown in Table 3.

Demographic Inference

We used a recently published support vector machine-based demographic inference classifier to perform all gender inferences used in this study (Zamal, Liu, and Ruths 2012). Like most inference methods, this approach constructs a feature

Group	Accuracy	Acc. (F)	Acc. (M)
@680trafficnews	0.810 ± 0.019	0.81	0.80
@ttcnotices	0.847 ± 0.021	0.86	0.83
@bikeunion	0.738 ± 0.088	0.77	0.71

Table 2: The performance of the gender inference method on the 400 labeled users from the three commuter datasets. The reported accuracy is the average of the 5-fold cross-validation performed over all labels, female labeled users, and male labeled users, respectively.

vector for each user consisting of a set of features derived from the user's microblog content and profile. This method also has the ability to incorporate neighborhood information, though we did not use this capability as neighbor details were shown to not significantly improve gender inference accuracy. The method was selected because it subsumes feature sets considered in prior methods and demonstrates equal or better performance on all demographic attributes considered (including gender).

Results

Prior to performing the gender inference on the entire commuter data sets, we assessed the expected inference accuracy of the method by performing cross-validation on the labeled users only. We first describe this validation stage and then the results of the full analysis.

Validation

Since the validity of our full analysis depended on the accuracy of the gender inference method, we first evaluated the accuracy of this method on each of the commuter populations considered. As described in the *Data* section above, for each commuter population the gender of 400 users was manually coded. To measure the accuracy of inferences on this labeled dataset, we performed 5-fold cross-validation on this labeled subset of each commuter data set. The result of this analysis are shown in Table 2.

The accuracy we observed is consistent with gender inference accuracies reported in prior work (Zamal, Liu, and Ruths 2012; Rao et al. 2010; Pennacchiotti and Popescu 2011; Conover et al. 2011; Rao and Yarowsky 2010). The only exception is (Burger, Henderson, and Zarrella 2011) which reported upwards of 95% accuracy; this result, however, was obtained on a highly specialized dataset of Twitter users who maintained a public, self-advertised blog — a detail which selected for users with greater signal-to-noise ratios in their Twitter content.

Overall, the results obtained indicate that, while non-trivial errors will be present in inferred gender labels, large differences in inferred gender composition will be significant: consider that accuracy values < 1 lead the majority gender to mislabel some fraction of users as the opposite gender. When an accommodation is made for these mislabelings (using the accuracy estimates in Table 2), the disparity between the majority and minority genders becomes more pronounced.[2]

[2]In the interest of space, we have not used the accommodation alluded to. These would not affect our overall results, however, since we consider only agreement in major-minor gender identification.

Commuter Gender Breakdown

For each commuter population, the gender classifier was trained on the labeled users from that population and applied to the remaining users. The inferred gender breakdown for each of these commuter groups is shown in Table 3.

We call the reader's attention to several features of the census data and inferred results.

First, in the census data, in each commuter population there is a gender that is in the significant majority (i.e., males for cars and bikes; females for public transport).

Second, in the inferred results, there is also a gender in the majority for each commuter group. Furthermore, given the 95% confidence intervals derived from the standard deviations computed in the prior section, these gender majorities are all significant — meaning that the majority reported is, with greater than 95% confidence, the majority in the commuter group itself.

Finally, note that for each commuter group the majority gender in both census data and the inferred results is the same (e.g., males form the majority of bicyclists in both census and Twitter populations). Because the gender majorities reported in the inferred results are significant, it follows that, for these three urban communities, a Twitter-based measurement has yielded quantitative insights into physical group demographics. This is the central result of this paper.

Discussion

The core finding of this paper is that the biases in gender makeup present in real-world commuter populations are also present in related Twitter populations as well. Foremost, this indicates that Twitter is capable of providing information about the demographic composition of some physical, urban populations. Of course, the scope of Twitter's demonstrated applicability is limited at present to commuter populations in Toronto. Nonetheless, in this context our approach delivers the benefits highlighted in the introduction.

Performance. The Twitter data used to train the data set was collected by six computers working in parallel in less than one hour. Furthermore, feature extraction, modeling building, and gender inference (assigning the gender labels) also require on the order of 15 minutes. Were bringing compute time down an important matter, there are numerous speed-ups possible.

Cost. The data used for this analysis was obtained entirely from a crawl of public Twitter accounts - which are free resources. Thus, the cost of running such an analysis requires only the provisioning of the bandwidth, power, and computing hardware needed.

Our work has also identified several of the challenges in building such systems.

Group	Twitter Group	% Female	Census # Males	Census # Females	Census % Female
Car	@680trafficnews	23.1	360,345	267,625	42.6
Public	@ttcnotices	81.2	173,590	271,225	61.0
Bike	@bikeunion	27.8	11,400	6,545	36.5

Table 3: The results of the gender inference method. Figures are reported as *% Female*. The last three column provide the gender breakdown for different commuting groups taken from the 2006 census.

Building portable demographic models. We initially expected to be able to use gender inference classifier models built from other labeled data sets (Zamal, Liu, and Ruths 2012). Interestingly, these performed very poorly on the commuter data, requiring us to build new models from labeled users drawn from the commuter data itself. It remains unclear as to why these existing classifier models did not perform well. It is clear that, statistically speaking, the optimal signals of gender change when a Toronto-specific population is considered. Determining how to avoid requiring specialized classifier models for each population of interest will be important in building commercially viable systems.

Improving accuracy. Obtaining consistency between census and online data in the majority gender present in commuter populations is an important milestone. A clear next step is to understand the extent to which online social data can provide accurate estimates of demographic composition. Certainly, the inherent sampling bias present within online populations will be a major challenge to overcome. Nonetheless, our existing finding suggests that additional accuracy improvements may be possible.

Selecting correct online communities. Initially, we considered the @bixitoronto group as the online proxy for bicycle commuters. Interestingly, we found that the gender makeup of this group is markedly different from that observed in the real-world Toronto bicycling community. Preliminary evidence suggests that this is because the Bixi service (a bike-sharing service) attracts a subpopulation of bicyclists whose gender makeup differs from that of the broader bicycling community. Identifying when an online group is a good proxy for a real-world community presents both philosophical and technical challenges.

Modeling demographics and urban communities. Our work thus far has notably benefited from three factors. First, the commuter populations of a city, generally speaking, have a notable presence online (as evidenced by our findings in this paper). Second, the presence of ground truth census data made validation of our findings possible. Third, the computational inference of gender is a relatively well-understood problem, as opposed to other demographic features such as ethnicity, education, and income. In applying our approach to other urban communities and demographic distinctions, it will be important to assess the extent to which these three conditions are met.

Conclusion

As individuals increasingly incorporate online social platforms into their daily life, new opportunities emerge for using this data to improve the services and resources that cities provide to their citizens. By mining this data, city administrators can gain new insights into the composition and needs of the communities they serve. In this paper, we have presented a technique by which the gender makeup of commuter populations can be inferred from microblog data. While many research question remain, our findings indicate that there is significant potential in using such data to make cities smarter, more efficient, and more responsive to their constituents.

References

Burger, J.; Henderson, J.; and Zarrella, G. 2011. Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Conover, M.; Ratkiewicz, J.; Francisco, M.; Goncalves, B.; Menczer, F.; and Flammini, A. 2011. Political polarization on twitter. In *Proceedings of the International Conference on Weblogs and Social Media*.

Mislove, A.; Lehmann, S.; Ahn, Y.-Y.; Onnela, J.-P.; and Rosenquist, J. N. 2011. Understanding the Demographics of Twitter Users. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM'11)*.

Pennacchiotti, M., and Popescu, A. 2011. A machine learning approach to twitter user classification. In *Proceedings of the International Conference on Weblogs and Social Media*.

Quercia, D.; Ellis, J.; Capra, L.; and Crowcroft, J. 2012. Tracking "gross community happiness" from tweets. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, 965–968.

Rao, D., and Yarowsky, D. 2010. Detecting latent user properties in social media. In *Proceedings of the NIPS workshop on Machine Learning for Social Networks*.

Rao, D.; Yarowsky, D.; Shreevats, A.; and Gupta, M. 2010. Classifying latent user attributes in twitter. In *Proceedings of the International Workshop on Search and Mining User-generated Contents*.

Statistics Canada, S. 2007. Population and dwelling counts, for canada, provinces and territories, 2006 and 2001 censuses, 100% data (table). Statistics Canada Catalogue no. 97-550-XWE2006002.

Zamal, F. A.; Liu, W.; and Ruths, D. 2012. Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. In *Proceedings of the International Conference on Weblogs and Social Media*.