Estimating Community Composition in Twitter and the Real World

Derek Ruths, derek.ruths@mcgill.ca
School of Computer Science, McGill University

Position paper prepared for
Workshop on Mapping Ideas: Discovering and Information Landscapes
San Diego State University, San Diego, CA
August 1-2, 2012

The composition of a community may have a dramatic impact on its behavior. Consider, for example, that a group of conservatives and another of liberals will respond very differently to legislation on taxes, gun control, and abortion; the mean age of individuals in a group is strongly correlated with their use of certain kinds of technologies; the income of community members will influence a variety of group features including its capacity to mobilize resources. Thus, fundamental to the study of human populations and large-scale human behavior is the knowledge of what kinds of individuals comprise the population of interest.

Remarkably, this information is difficult, error-prone, and resource-intensive to obtain. For physical populations, traditionally, surveys have been used to obtain demographic information. This involves interacting directly (either in person or over the phone) with a large number of people. Besides being time-consuming and expensive, this method is also implicitly not real-time.

In virtual (online) populations, very little conclusive work has been done. Demographic information is typically limited to the details that individuals choose to share in their profile. This information is seriously limited by what attributes a platform asks an individual to complete (e.g., in Twitter a user cannot even indicate their gender), the sporadic completion of available fields, and the accuracy of the information provided (e.g., individuals reporting that they live "on the moon").

In our work we are investigating computational methods for (1) inferring the demographic attributes of Twitter groups without relying on explicit profile details (2) studying the distribution and clustering of demographic attributes within Twitter, and (3) extrapolating Twitter group demographic attributes to physical populations.

Inferring demographic attributes of Twitter users

In our work to date, we have developed methods for inferring gender, age, and political orientation [1]. Our published work in this area has focused on understanding the extent to which the social context of a user can be used to improve inference accuracy. Our current findings suggest that using even just the textual features of immediate neighbors can yield a dramatic and stable improvement in inference accuracy. This finding appears to be independent of machine learning technique used.

In this and related work, we discovered that all inference methods are highly sensitive to time differences between training data and testing data [2,3]. One direction of our current work is considering various techniques by which methods can be made more robust and insensitive to time. Another approach is considering the development of online learning algorithms that continually update the model used to make demographic attribute inferences – thereby rendering the time sensitivity problem moot.

Studying demographic attributes in Twitter

Our initial studies of demographic attributes in Twitter led us to consider political orientation of individuals and their neighbors [2]. Our findings here were quite striking. Looking at the differences between Republican and Democrat individuals revealed that the two groups have very different strategies in the selection of Twitter followees. Our findings may be interpreted to suggest that some strategies are more likely to create echo chambers, which have been a source of significant concern in light of evidence of increasing degrees of polarization and radicalization of political discourse.

Extrapolating online population demographics to physical populations

The ultimate goal of our research is to use our Twitter demographic inference methods to made estimates of the demographic attributes of populations in the real-world. Clearly, this is a non-trivial problem made more difficult by the fact that Twitter is not a uniform random sample of the human population. This suggests that any estimates must be adjusted by this degree of bias present in the Twitter dataset.

In our initial work, we have set the issue of bias-correction aside, and considered a very concrete case study: can Twitter groups be used to infer demographic attributes of online populations [3]. Our findings indicate that Twitter *can* be a source of information about physical populations. We use previously developed methods (e.g., [1]) to infer the gender composition of certain Twitter communities and find that the compositions match up with the compositions observed in certain commuter populations. This work suggests that there is much profitable work to be done in mapping online trends and population characteristics onto the online world.

References

[1] Zamal, Liu, Ruths. “Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors.” Proceedings of ICWSM 2012.
[2] Liu, Zamal, Ruths. “Characterizing Political Orientation on Twitter.” Submitted to SocialCom, 2012.
[3] Liu, Zamal, Ruths. “Using social media to infer gender composition of commuter populations.” When the City meets the Citizen Workshop, Proceedings of ICWSM 2012.