Inferring Gender from Twitter Data

Xinyi Liu [*1] **Mingren Shen** [*1] **Faust Shi** [*1]

Abstract

1. Introduction

Social media are popular platforms which can record a variety of users' personal information. Among this information, demographic characteristics (Sloan et al., 2013) and space-time travel patterns (Hasan et al., 2013) are two of great interesting areas to geographers. Demographic characteristics include gender, age, occupation and so on. Space-time travel patterns refer to the law of people's travel trajectories along both space and time (e.g., periodicity of visiting a restaurant, time of leaving home on weekdays). Traditionally, these two pieces of information are gathered from surveys and interviews (Chen et al., 2011). As the Information and Communication Technology (ICT) develops, they can be mined from cell phone records and other electronic GPS records (Siła-Nowicka et al., 2016). However, these methods either cost too much money and too large resources or encroach privacy. Nowadays, social media data could take part in these works with little expense and free of privacy infringement (Preoţiuc-Pietro & Cohn, 2013). Researchers have created robust framework and methodologies to mine space-time travel patterns from the geo-tagged online messages with temporal stamps. However, some direct demographic records(e.g., gender and occupation information on Twitter) is missing for some online platforms. It is important to design and implement methods to mine them from online messages (Rao et al., 2011).

It is important to restore user demographic information and link them with space-time travel patterns to unveil their relationships (Ahn et al., 2016). The results indicate social segmentation and contribute to the development of public facilities for different subgroups (**?**). However, few studies are conducted to investigate their relationships. Thus the travel patterns are discussed in terms of a general group of people, which is insufficient for human mobility studies because diverse travel patterns of different subgroups have been studied using other data sources (Kang et al., 2010). To fill the research gap, this paper tries to develop a machine learning classifier to study the characteristic space-time travel pattern features of different subgroups, specifically different genders (male/female/others). Test results of our classifier are compared with an online word-based Bayes network classifier and a first-name-based classifier. Our classifier shows improvement in either test accuracy or in test F1 scores.

2. Relate Work

2.1. Twitter Gender Inference

Automatically inferring user gender from Twitter is heavily investigated by both academic socitey and industry because gender is one of the most important demographic property of the user. Generally, there are 3 main approaches used for deriving Twitter users gender information: (1) profile-based (2) content-based (3) hybrid (Beretta et al., 2015).

2.1.1. Profile-based Gender Inference

Profile-based methods use the meta-data of the user's account in Twitter to help determine the gender of the users (Sloan et al., 2015). In Twitter, a user can show his name, description, location, followers and friends publicly. Although Twitter does not check the authenticity of profile information, several studies have proven that most Twitter users provide their real name and real gender in their public profile (Cesare et al., 2017). The simplest and best feature of profile information is users' first name. Previous studies have shown that by comparing name record from nation demographic survey, first name based gender classifier can achieve real good performance (Sloan et al., 2013; Mislove et al., 2011) .There are several mature services like genderize.io and packages [1] [2] inferring gender using only first name. For example `genderize.io` provides API that can be used to determine gender of a first name with the help of a database contains 216286 distinct names across 79 countries and 89 languages. Generally, the profile-based method is considered as the benchmark of gender inference

[*]Equal contribution [1]University of Wisconsin,Madison , USA. Correspondence to: Xinyi Liu <xliu636@wisc.edu>.

CS 760 Final Project, UW-Madison, 2017.

[1]https://github.com/tue-mdse/genderComputer

[2]https://github.com/muatik/genderizer

due to the high efficacy it can achieve. For example, Liu et al. use first name as the main feature to infer gender in Twitter and they obtain the accuracy around 85% (Liu & Ruths, 2013).

2.1.2. CONTENT-BASED GENDER INFERENCE

Content-based methods focus more on the content posted by Twitter users online. Twitter allows the users to post 140-character tweets(280-character after Nov.7th, 2017 [3]) on their personal account. Early researches have proven that user of different genders have different word choices and writing styles. For example, Rao et al. tries to processes text generated by Twitter user to extract unigram and bigram features using a Support Vector Machine (SVM) algorithm to determine latent user attributes like gender information (Rao et al., 2010). Several other studies have been using similar n-gram features in combination with logistic and linear regression models to infer more demographic information of the user like gender (Burger et al., 2011), age (Nguyen et al., 2013a;b) ,politic attitudes (Pennacchiotti & Popescu, 2011).

In addition to n-grams features, stylistic features in Twitter text have also been used for user gender and other demographic information inference. For instance, several approaches describe methods to determine gender based on the usage of gender specific words like he, she or his, her, abbreviations, punctuation (Fink et al., 2012), smileys, repeated letters, pronouns, EMOJI (Wolf, 2000) , hashtags and other grammatical features (Cheng et al., 2011; Ito et al., 2013).

However, content based methods require background and domain specific knowledge of natural language processing and can not be easily extended to other languages. Some studies have made some progresses on this direction but more efforts are needed (Mozetič et al., 2016).

2.1.3. HYBRID GENDER INFERENCE

Hybrid approaches try to combine both profile based methods, content based methods and other source features or information to improve the accuracy of results. Many efforts have been devoted for this methods, for example, Orlandi et al. tries to use information both from other sources like Facebook and Twitter to infer user profile (Orlandi et al., 2012). Li et al. tries to use online social networks to help user identification in Twitter (Li et al., 2017). Other directions like using hierarchical knowledge base (Kapanipathi et al., 2014), Twitter account the user following (Chamberlain et al., 2016) or migration patterns (Zagheni et al., 2014) etc.

In this paper, we want to use the third approach, the hybrid method to infer the user gender information. To be more specific, we want to combine Twitter users' mobility features like travel pattern with the content based information to design a new system to infer Twitter user gender. Previous studies have already shown that there are differences in Twitter user geo-temporal distribution between different genders (Graham et al., 2014; Mahmud et al., 2014; Weber & Garimella, 2014; Longley & Adnan, 2016). But few work has been done to utilize the geo-temporal feature of Twitter accounts to infer the gender information.

2.2. User Temporal and Spatial Travel Patterns

Human trajectories show a high degree of temporal and spatial regularity (Gonzalez et al., 2008). Two important statistical properties of such mobility patterns are displacement length (i.e., distance between a person's positions at consecutive locations) and radius of gyration (i.e., characteristic distance traveled by a person within a specific period of time)(Gonzalez et al., 2008). Besides, three kinds of entropy are calculated to quantify the degree of predictability of a person's travel trajectories (Song et al., 2010). Respectively, they measure location diversity, heterogeneity of visitation patterns along time and the full temporal and spatial order of the person's mobility pattern. Moreover, the number of co-location records are summarized within different time period to measure the size of the set of co-locations in social network studies (Cranshaw et al., 2010).

Although different variables characterizing detailed temporal and spatial travel patterns have been proposed , they are based on dense cell phone record data or verbose survey data. For the rising social media data, temporal and spatial travel patterns are usually visualized for interactive investigations (Yin et al., 2016), while lacking numeric parameters for further statistical analysis. Different from previous dense data study, social media data like Twitter is sparse, which only captures noncontinuous segments of users' travel trajectories. So it is hard to get enough useful features from Twitter data. To use the travel patterns stored in social media data like Twitter, our project tries to design new temporal and spatial features to represent travel patterns and weigh their relative importance to the classifier through several tests.

Acknowledgements

We would like to sincerely thank Prof.Liang, Depratment of Computer Science at UW-Madison, for valuable suggestions to implement this project and Prof.Huang, Department of Geography at UW-Madison, for the help of collecting Twitter data-set and annotations of user genders.

[3] https://blog.twitter.com/official/en_us/topics/product/2017/Giving-you-more-characters-to-express-yourself.html

Software and Data

We provide all our data and program in Github and you can check them online `https://github.com/iphyer/cs760_TwitterDemographics`.

We use scikit-learn (Pedregosa et al., 2011) as our machine learning program library and Pandas (McKinney, 2015) for data processing.

References

Ahn, Young-joo, Wooten, Dr, Marian, H, Norman, Dr, William, C, McGuire, Dr, et al. Gender differences in travel constraints and changed travel pattern after a senior travel program introduces: A case study of the florence county senior travel program. 2016.

Beretta, Valentina, Maccagnola, Daniele, Cribbin, Timothy, and Messina, Enza. An interactive method for inferring demographic attributes in twitter. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, pp. 113–122. ACM, 2015.

Burger, John D, Henderson, John, Kim, George, and Zarrella, Guido. Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1301–1309. Association for Computational Linguistics, 2011.

Cesare, Nina, Grant, Christan, and Nsoesie, Elaine O. Detection of user demographics on social media: A review of methods and recommendations for best practices. *arXiv preprint arXiv:1702.01807*, 2017.

Chamberlain, Benjamin Paul, Humby, Clive, and Deisenroth, Marc Peter. Probabilistic inference of twitter users age based on what they follow. *arXiv preprint arXiv:1601.04621*, 2016.

Chen, Jie, Shaw, Shih-Lung, Yu, Hongbo, Lu, Feng, Chai, Yanwei, and Jia, Qinglei. Exploratory data analysis of activity diary data: a space–time gis approach. *Journal of Transport Geography*, 19(3):394–404, 2011.

Cheng, Na, Chandramouli, Rajarathnam, and Subbalakshmi, KP. Author gender identification from text. *Digital Investigation*, 8(1):78–88, 2011.

Cranshaw, Justin, Toch, Eran, Hong, Jason, Kittur, Aniket, and Sadeh, Norman. Bridging the gap between physical location and online social networks. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pp. 119–128. ACM, 2010.

Fink, Clayton, Kopecky, Jonathon, and Morawski, Maksym. Inferring gender from the content of tweets: A region specific example. In *ICWSM*, 2012.

Gonzalez, Marta C, Hidalgo, Cesar A, and Barabasi, Albert-Laszlo. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

Graham, Mark, Hale, Scott A, and Gaffney, Devin. Where in the world are you? geolocation and language identification in twitter. *The Professional Geographer*, 66(4): 568–578, 2014.

Hasan, Samiul, Zhan, Xianyuan, and Ukkusuri, Satish V. Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. In *Proceedings of the 2nd ACM SIGKDD international workshop on urban computing*, pp. 6. ACM, 2013.

Ito, Jun, Hoshide, Takahide, Toda, Hiroyuki, Uchiyama, Tadasu, and Nishida, Kyosuke. What is he/she like?: Estimating twitter user attributes from contents and social neighbors. In *Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on*, pp. 1448–1450. IEEE, 2013.

Kang, Chaogui, Gao, Song, Lin, Xing, Xiao, Yu, Yuan, Yihong, Liu, Yu, and Ma, Xiujun. Analyzing and geo-visualizing individual human mobility patterns using mobile call records. In *Geoinformatics, 2010 18th International Conference on*, pp. 1–7. IEEE, 2010.

Kapanipathi, Pavan, Jain, Prateek, Venkataramani, Chitra, and Sheth, Amit. User interests identification on twitter using a hierarchical knowledge base. In *European Semantic Web Conference*, pp. 99–113. Springer, 2014.

Li, Yongjun, Zhang, Zhen, and Peng, You. A solution to tweet-based user identification across online social networks. In *International Conference on Advanced Data Mining and Applications*, pp. 257–269. Springer, 2017.

Liu, Wendy and Ruths, Derek. What's in a name? using first names as features for gender inference in twitter. In *AAAI spring symposium: Analyzing microtext*, volume 13, pp. 01, 2013.

Longley, Paul A and Adnan, Muhammad. Geo-temporal twitter demographics. *International Journal of Geographical Information Science*, 30(2):369–389, 2016.

Mahmud, Jalal, Nichols, Jeffrey, and Drews, Clemens. Home location identification of twitter users. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(3):47, 2014.

McKinney, Wes. pandas: a python data analysis library. *see http://pandas. pydata. org*, 2015.

Mislove, Alan, Lehmann, Sune, Ahn, Yong-Yeol, Onnela, Jukka-Pekka, and Rosenquist, J Niels. Understanding the demographics of twitter users. *ICWSM*, 11:5th, 2011.

Mozetič, Igor, Grčar, Miha, and Smailović, Jasmina. Multilingual twitter sentiment classification: The role of human annotators. *PloS one*, 11(5):e0155036, 2016.

Nguyen, D, Gravel, R, Trieschnigg, D, and Meder, Theo. "how old do you think i am?": A study of language and age in twitter. In *Proceedings of the 7th International Conference on Weblogs and Social Media, ICWSM 2013*, pp. 439–448, 01 2013a.

Nguyen, Dong, Gravel, Rilana, Trieschnigg, Dolf, Meder, Theo, and Yeung, C-M Au. Tweetgenie: automatic age prediction from tweets. *ACM SIGWEB Newsletter*, 4(4), 2013b.

Orlandi, Fabrizio, Breslin, John, and Passant, Alexandre. Aggregated, interoperable and multi-domain user profiles for the social web. In *Proceedings of the 8th International Conference on Semantic Systems*, pp. 41–48. ACM, 2012.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Pennacchiotti, Marco and Popescu, Ana-Maria. Democrats, republicans and starbucks afficionados: user classification in twitter. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 430–438. ACM, 2011.

Preoţiuc-Pietro, Daniel and Cohn, Trevor. Mining user behaviours: a study of check-in patterns in location based social networks. In *Proceedings of the 5th Annual ACM Web Science Conference*, pp. 306–315. ACM, 2013.

Rao, Delip, Yarowsky, David, Shreevats, Abhishek, and Gupta, Manaswi. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pp. 37–44. ACM, 2010.

Rao, Delip, Paul, Michael J, Fink, Clayton, Yarowsky, David, Oates, Timothy, and Coppersmith, Glen. Hierarchical bayesian models for latent attribute detection in social media. *ICWSM*, 11:598–601, 2011.

Siła-Nowicka, Katarzyna, Vandrol, Jan, Oshan, Taylor, Long, Jed A, Demšar, Urška, and Fotheringham, A Stewart. Analysis of human mobility patterns from gps trajectories and contextual information. *International Journal of Geographical Information Science*, 30 (5):881–906, 2016.

Sloan, Luke, Morgan, Jeffrey, Housley, William, Williams, Matthew, Edwards, Adam, Burnap, Pete, and Rana, Omer. Knowing the tweeters: Deriving sociologically relevant demographics from twitter. *Sociological research online*, 18(3):7, 2013.

Sloan, Luke, Morgan, Jeffrey, Burnap, Pete, and Williams, Matthew. Who tweets? deriving the demographic characteristics of age, occupation and social class from twitter user meta-data. *PloS one*, 10(3):e0115545, 2015.

Song, Chaoming, Qu, Zehui, Blumm, Nicholas, and Barabási, Albert-László. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

Weber, Ingmar and Garimella, Venkata Rama Kiran. Visualizing user-defined, discriminative geo-temporal twitter activity. In *ICWSM*, 2014.

Wolf, Alecia. Emotional expression online: Gender differences in emoticon use. *CyberPsychology & Behavior*, 3 (5):827–833, 2000.

Yin, Junjun, Gao, Yizhao, Du, Zhenhong, and Wang, Shaowen. Exploring multi-scale spatiotemporal twitter user mobility patterns with a visual-analytics approach. *ISPRS International Journal of Geo-Information*, 5(10): 187, 2016.

Zagheni, Emilio, Garimella, Venkata Rama Kiran, Weber, Ingmar, et al. Inferring international and internal migration patterns from twitter data. In *Proceedings of the 23rd International Conference on World Wide Web*, pp. 439–444. ACM, 2014.