Journal of Spatial Information Science
Number 20 (2020), pp. 51–55

doi:10.5311/JOSIS.2020.20.665

Invited Article

Mining urban perceptions from social media data

Yu Liu[1], Yihong Yuan[2], and Fan Zhang[3]

[1]Institute of Remote Sensing and Geographical Information Systems, School of Earth and Space Sciences, Peking University, Beijing, China
[2]Department of Geography, Texas State University, San Marcos, TX, USA
[3]Senseable City Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA

Received: February 29, 2020; accepted: April 13, 2020

Abstract: This vision paper summaries the methods of using social media data (SMD) to measure urban perceptions. We highlight two major types of data sources (i.e., texts and imagery) and two corresponding techniques (i.e., natural language processing and computer vision). Recognizing the data quality issues of SMD, we propose three criteria for improving the reliability of SMD-based studies. In addition, integrating multi-source data is a promising approach to mitigating the data quality problems.

Keywords: social media data, urban perceptions, place, data quality, natural language processing, computer vision

1 Introduction

With the rapid development of the mobile Internet, it is feasible to harvest large volumes of social media data (SMD) with spatio-temporal tags. Such geo-tagged big data provide a new approach to measuring the perceptions of different places at a collective level [12,14]. This new research field is evolving rapidly due to the increasing visibility of social media platforms, such as Facebook, Twitter, Weibo, Foursquare, Flickr, and Yelp. Most geolocated social media posts are from densely-populated urban areas, making it possible to measure urban perceptions and represent human cognition of intra-urban spatial heterogeneity with the support of SMD [15,18]. These studies provide valuable support in urban planning and policy making, and consequently, the implementation of sustainable development goals (SDG) [9].

In addition to spatial information (e.g., locations or place names), geo-tagged social media posts usually contain contextual and semantic information, such as text descriptions of a specific place, activities, or feelings of the users, as well as photos taken at a place [3,17]. Two data mining techniques are widely used to mine SMD. The first is natural language processing (NLP), which extracts activities and emotions of a massive volume of users from textual information [1]. Second, thanks to the rapid development of computer vision supported by deep convolutional neural networks (DCNNs), we can conduct content-based image analyzes for photos posted on social media.

NLP provides an effective measure to not only extract the spatial extent of human activities, but also to analyze the emotions and experiences associated with a place. This is in line with the definition of a "place" in human geography, where places are defined based on both the physical and human characteristics of a location [2]. Given that most place boundaries are vague, previous studies introduced fuzzy sets to represent the spatial extents of places. With the support of SMD, it is feasible to collect posts with particular place names (e.g., "Harvard Yard") and compute the kernel density surface of this place name. The resulting density surface can be viewed as a two-dimensional membership function of a place [6]. In addition, the latent Dirichlet allocation (LDA) model is widely used to analyze textual information and extract the spatial distribution of topics (e.g., "tourism" or "sport") from SMD [4]. For example, Ilieva and McPhearson [10] investigated the spatial distribution of topics on Twitter to represent the geography of urban functional regions in London. Areas with a high density of topics like "shopping" or "sales" are likely to be a shopping centre. Researchers also developed methods to extract emotional expressions and place sentiments from SMD. Some studies used non-negative matrix factorization to measure feelings such as depression or happiness. Yang and Mu [19] quantified the degree of depression using Twitter data collected in New York City and revealed the relationship between depression levels and demographics.

Computer vision, especially driven by DCNNs, makes it possible to automatically and efficiently extract and "understand" the content in an image. In addition to photos on social media, street view provides an alternative imagery data source to quantitatively represent the physical and human characteristics of a place, such as demographics, land uses, vibrancy, and place sentiments. Hence, we can measure certain physical characteristics of environments, such as the greenness and openness of urban spaces, from geo-tagged images collected either from social media platforms or street view services [11]. A new research direction along this line is to build end-to-end models that can predict particular socioeconomic characteristics of interested places. Using street view images, Zhang et al. [21] quantified the perceptions of urban spaces across six dimensions: how safe, lively, beautiful, wealthy, depressing, and boring the spaces were. As an example, Figure 1 presents a "safety map" of Houston derived from Google street view images and deep convolutional networks.

2 Challenges and future research directions

Although SMD provide an unprecedented opportunity to mine urban perceptions, this line of research encounters several challenges.

First, social media data suffer from various data quality issues. In addition to accuracy and precision problems, sampling biases are inevitable from sociodemographic, spatiotemporal, and semantic perspectives [20]. Without a statistically-sound sampling strategy, the findings derived from SMD can be questionable. Additionally, social media data

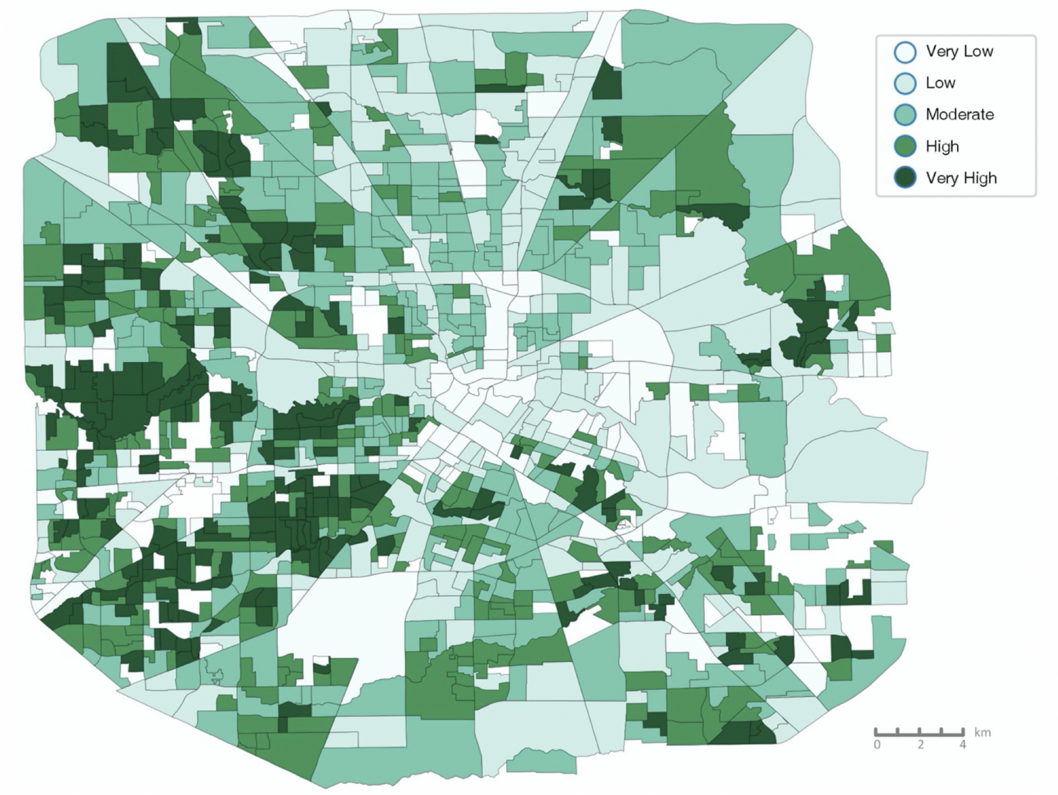

Figure 1: Safety map of Houston. The darker green areas indicate a higher safety score inferred by a deep learning model. The model was trained using millions of Google Street View images rated by individuals worldwide.

are considered a type of "thin" data, meaning that although the volume is large, the data are missing details, such as the background information of social media users [13]. This further contributes to the difficulty of obtaining a well-balanced sample set from different population groups, leading to ecological fallacy when aggregating social media posts using different spatial units. Second, most SMD are used to extract metrics that cannot be directly measured or lack ground truths (e.g., the degree of happiness [16]). Hence, it is difficult to validate the results obtained from social media analysis. Third, the use of mathematical models introduces uncertainty when dealing with social media data [5]. We often have to balance the generality and the accuracy of models. If a model is too narrowly defined, the rules for one place may not be generalizable to another place if their socio-economic characteristics differ too much.

To address these concerns, three criteria, from the least to the most challenging, may help researchers design a sound SMD-based study. First, the data mining methods should be reasonable. For example, LDA is a widely used method to model textual topics, so applying it to analyzing the textual content on Twitter is justifiable [8]. Second, the results should be consistent with common sense. Although a certain indicator value (e.g., happiness=0.8) cannot be directly validated, one can easily determine whether the results follow common sense based on the ordinal relations of the metrics between places. If the degree of happiness of place A is greater than that of B, in general place A should be viewed as more emotionally positive. Third, a good validation strategy is to use the derived indicator to predict another indicator whose actual values can be easily collected or observed. For instance, Gebru et al. [7] extracted car makes and models from street view images and used this information to predict if a voter is more likely to vote for the Democratic or the Republican Party. The results are verifiable using election data.

To mitigate the data quality issues and achieve the above-mentioned criteria, a possible solution is to integrate multi-source data. Given that single-source social media data may be biased, combining various data sources provides complementary information from multiple perspectives and reduces the representativeness biases. Traditional survey-based data can be very useful in this case because these data are collected from well-designed samples and contain rich individual-level background information; therefore, they can complement SMD to build more solid models. It is also necessary to develop a statistical framework to integrate data from different sources and with various quality issues.

In the future, the emergence of new communication techniques such as 5G will significantly increase the network bandwidth and broaden the content that can be shared on social media platforms. With the support of high-speed mobile networks, users can easily share video clips and three-dimensional point cloud data. These data contain richer information about the geographical environment compared to texts and images. We need to develop more powerful tools to extract features from such data and enrich the perceptions of urban places.

References

[1] AGGARWAL, C. C., AND WANG, H. Text mining in social networks. In *Social network data analytics*. Springer, 2011, pp. 353–378.

[2] AGNEW, J. *Space and place*. The SAGE handbook of geographical knowledge. Sage London, 2011.

[3] BAO, J., ZHENG, Y., AND MOKBEL, M. F. Location-based and preference-aware recommendation using sparse geo-social networking data. In *Proceedings of the 20th international conference on advances in geographic information systems* (2012), pp. 199–208.

[4] BLEI, D. M., NG, A. Y., AND JORDAN, M. I. Latent Dirichlet allocation. *Journal of machine Learning research 3*, Jan (2003), 993–1022.

[5] BOX, G. E., AND DRAPER, N. R. *Empirical model-building and response surfaces*, vol. 424. Wiley New York, 1987.

[6] GAO, S., LI, L., LI, W., JANOWICZ, K., AND ZHANG, Y. Constructing gazetteers from volunteered big geo-data based on Hadoop. *Computers, Environment and Urban Systems 61* (2017), 172–186.

[7] GEBRU, T., KRAUSE, J., WANG, Y., CHEN, D., DENG, J., AIDEN, E. L., AND FEI-FEI, L. Using deep learning and google street view to estimate the demographic makeup of

neighborhoods across the United States. *Proceedings of the National Academy of Sciences 114*, 50 (2017), 13108–13113.

[8] HONG, L., AND DAVISON, B. D. Empirical study of topic modeling in Twitter. In *Proceedings of the first workshop on social media analytics* (2010), pp. 80–88.

[9] ILIEVA, R. T., AND MCPHEARSON, T. Social-media data for urban sustainability. *Nature Sustainability 1*, 10 (2018), 553–565.

[10] LANSLEY, G., AND LONGLEY, P. A. The geography of Twitter topics in London. *Computers, Environment and Urban Systems 58* (2016), 85–96.

[11] LI, X., ZHANG, C., LI, W., RICARD, R., MENG, Q., AND ZHANG, W. Assessing street-level urban greenery using Google Street View and a modified green view index. *Urban Forestry & Urban Greening 14*, 3 (2015), 675–685.

[12] LIU, Y., LIU, X., GAO, S., GONG, L., KANG, C., ZHI, Y., CHI, G., AND SHI, L. Social sensing: A new approach to understanding our socioeconomic environments. *Annals of the Association of American Geographers 105*, 3 (2015), 512–530.

[13] LONGLEY, P. A., AND ADNAN, M. Geo-temporal Twitter demographics. *International Journal of Geographical Information Science 30*, 2 (2016), 369–389.

[14] MACEACHREN, A. M. Leveraging big (geo) data with (geo) visual analytics: Place as the next frontier. In *Spatial data handling in big data era*. Springer, 2017, pp. 139–155.

[15] MARTÍ, P., SERRANO-ESTRADA, L., AND NOLASCO-CIRUGEDA, A. Using locative social media and urban cartographies to identify and locate successful urban plazas. *Cities 64* (2017), 66–78.

[16] MITCHELL, L., FRANK, M. R., HARRIS, K. D., DODDS, P. S., AND DANFORTH, C. M. The geography of happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place. *PloS one 8*, 5 (2013).

[17] ROICK, O., AND HEUSER, S. Location based social networks—definition, current state of the art and research agenda. *Transactions in GIS 17*, 5 (2013), 763–784.

[18] SHELTON, T., POORTHUIS, A., AND ZOOK, M. Social media and the city: Rethinking urban socio-spatial inequality using user-generated geographic information. *Landscape and urban planning 142* (2015), 198–211.

[19] YANG, W., AND MU, L. GIS analysis of depression among Twitter users. *Applied Geography 60* (2015), 217–223.

[20] YUAN, Y., LU, Y., CHOW, T. E., YE, C., ALYAQOUT, A., AND LIU, Y. The missing parts from social media-enabled smart cities: Who, where, when, and what? *Annals of the American Association of Geographers 110*, 2 (2020), 462–475.

[21] ZHANG, F., ZHOU, B., LIU, L., LIU, Y., FUNG, H. H., LIN, H., AND RATTI, C. Measuring human perceptions of a large-scale urban region using machine learning. *Landscape and Urban Planning 180* (2018), 148–160.