Extracting and visualizing geo-social semantics from the user mention network on Twitter

Caglar Koylu
Department of Geographical and Sustainability Sciences
University of Iowa
316 Jessup Hall, Iowa City, IA, 52242
caglar-koylu@uiowa.edu

This paper introduces an approach for extracting and visualizing geo-social semantics of user mentions on Twitter. The approach consists of three steps. First, data filtering and processing is performed to construct a directed area-to-area mention network in which tweets are aggregated into flow bins between geographic areas. Second, using flow bins as documents, a probabilistic topic model is employed to detect a collection of topics, and classify each area-to-area flow into a mixture of topics with differing probabilities. Third, for each topic, a modularity graph of mentions is obtained and visualized using a flow map and a topic cloud to infer semantics from the set of frequently co-occurring words for each topic. To demonstrate, a dataset of 19 million geo-tagged mentions during the primary elections (Feb-Jun, 2016) in the U.S. were analyzed. The results highlight changing patterns of symmetry, distance and clustering of flows by the topic of content.

1 Introduction

Previous studies have utilized user generated textual content such as geo-tagged tweets and messages exchanged in location-based social networks (LBSN) to study the effect of geographic proximity on social interactions [1-3]; the influence of information diffusion and social networks on real-world geographic events, such as demonstrations, protests, and group activities [4]; and the structural and geographic characteristics of the communication network [5-7]. Such studies use information flows to model social interactions, but often ignore the content of the information exchanged between the individuals of the network.

A variety of computational and semantic analysis techniques have been developed to infer human behavior, ideological and attitudinal similarity between individuals [8], common topics and way of speaking [9], group identities [10], demographic and socio-economic characteristics [11] from large volumes of user-generated textual data. Latent Dirichlet Allocation (LDA) [12, 13] has been successfully employed to detect geographic events, and recommend places and friends based on user location and similarity of shared content between users in LBSNs. Despite these efforts, there has been little work that focuses on understanding of geographic patterns of interpersonal communication and how the common topics of information vary based on the geographic distance and characteristics of locations [14].

We introduce an approach for extracting and visualizing geo-social semantics from the big data of user mentions on Twitter. The approach consists of three steps. First, data filtering and processing is performed and a bi-directional area-to-area mention network is constructed. In an area-to-area mention network, the original locations of tweets are aggregated into a small set of areas (e.g., counties) and mentions are combined into flow bins between these geographical areas. Second, flow bins are used to train a probabilistic topic model which generates a collection of topics and classify each flow bin into a mixture of topics with differing probabilities. Third, for each topic, a modularity graph of mentions is obtained and visualized using a flow map and a topic cloud to infer semantics from the set of frequently co-occurring words for each topic.

2 Data processing and filtering

To demonstrate, geo-tagged tweets within the Contiguous U.S. during presidential primaries and caucuses between February 1, 2016 and June 14, 2016 were collected through Twitter's streaming API. This specific time period was selected for the intent to capture election related conversations in addition to other themes from the Twitter corpora. The data consisted of 284,868,345 tweets, and 4,571,070 million distinct users. We removed tweets from non-personal accounts (e.g., weather, emergency, and job ads), and external sources such as pictures and check-ins (e.g., Instagram, Foursquare); users with more than 3000 followers; and users whose velocities (i.e., equals to the distance divided by time between two consecutive tweets) are above (1,000km/h) which would indicate spam users and bots. After the initial filtering, the number of tweets decreased to 75.0% (213,649,745) and the users to 88% (4,050,523).

A Twitter user can reply or mention other users by including their @username in her tweet. When a user A (sender) mentions user B (recipient), the tweet include only the location of the sender. The location of the recipient in a mention can be derived only if the recipient has a geo-tagged tweet in the sample. Among the filtered tweets, 45% (95,855,784) include a user mention, and 65.0% (2,632,840) of the users mentioned another user in a tweet at least once. The recipient's location was successfully extracted in 19.80% (19,046,949) of all mention tweets to form a network of geo-tagged user mentions.

3 Topic Modeling

LDA has been commonly used to identify topics from large collections of textual data. Given a collection of textual documents (e.g., books, articles, emails), LDA models a collection of k topics as a multinomial distribution over words within these documents.

$$P(Z|W,D) = \frac{W_{Z+\beta w}}{total\ tokens\ in\ Z + \beta} * D_{Z+\propto}$$

The probability that word W came from topic Z, is calculated as the normalized product of the frequency of W in Z $(W_{Z+\beta w})$ and the number of other words in document D that already belong to Z $(D_{Z+\propto})$. β and β_w are hyper-parameters that represent the chance that word W belongs to topic Z even if it is nowhere else associated with Z.

Training a topic model with short documents (i.e., individual tweets) results in unstable classifications with increased uncertainty due to the severe data sparsity [15]. To alleviate the problem, previous work combined tweets into temporal [16], spatio-temporal [17], and user bins [18]. In order to capture the semantics of interpersonal conversations between distinct pairs of geographic areas we combined tweets into county-to-county flow bins by assigning each user to a home county, and adding tweets into a flow bin based on the home counties of the users in the mentioned relationship. Each tweet exists in an origin-destination flow bin only once to avoid duplicate content when a tweet includes multiple mentions between the same county pair.

The density of geo-tagged mentions by distance (Figure 1) support the findings of previous research that the probability of a user mention is high between users with closer geographic proximity [19]. 46% of the geo-tagged mentions had both the recipient and sender within the same county, 70% were within the same state, and only 30% were between states. In contrast to the short-text problem, the documents in a topic model must be small enough that the proportions between topics could vary significantly between documents. Great variation in volume between flow bins within states and between states distort the results of LDA and undermine the topical heterogeneity of the model. Thus, we separated the flow bins into two groups: within state, and between state; and constructed a separate LDA model for each. In this paper, we report the results of our analysis on user mentions in which the recipient and sender are from different states, in order to capture the geographical variation of the topical content among long-distance interpersonal communication.

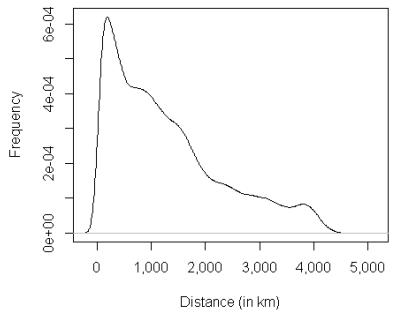

Figure 1 Density of geo-tagged user mentions by distance (in km)

In order to alleviate the bias from small bins, we further filtered out the flow bins that have less than 100 tweets and 10 users before training the topic model. While the number of flow bins (or county pairs) decreased from 350,295 to 49,436; the volume of mentions was reduced by only 13%.

We used the Mallet toolkit [20] to implement the LDA model and trained a set of topic models with differing number of topics (20, 50, and 100), number of iterations

(2000), and evaluated the topical overlap using cosine similarity. We selected 50 topics as it produced the less overlapping topics than 100, and more distinct topics than 20.

5 Mapping Modularity Flows

The topic model classifies each flow bin with a mixture of topics with differing probabilities. For example, a flow from county A to county B may be comprised of 50% topic 5, 30% topic 2, 10% topic 32, and 10% in other topics. Among a set of fuzzy classification of flow bins, one may isolate each topic to create a separate graph and estimate the weight of a link by multiplying the probability of isolated topic by the volume of user mentions on that link.

Although the number of mentions have been significantly reduced after excluding within-state mentions, county-to-county pairs still form a complex graph with a large number of links that requires further simplification. One can reduce the graph by graph partitioning and regionalization [21, 22] which combines unit areas into a smaller set of natural regions where there are more flows within regions than across regions. In order to ease the interpretation of our results, and reveal user mention patterns at the state level, we aggregate county pairs into state-to-state user mention flows for each topic and calculate a modularity measure to select the flows that are above expectation [22]. Expected number of mentions on a link is calculated as:

$$EM\ (O, D) = F_O\ F_D\ f\ (O, D)\ /\ (F_S{}^2 - \sum_{i=0}^{n} F_i{}^2)$$

where F_O is the number of mentions originated from state O, F_D is the number of times that state D is mentioned, f (O,D) is the number of mentions from state O to state D, F_S is the number of mentions between all states, and $\sum_{i=0}^{n} F_i{}^2$ is used to remove within-state expectations. Finally, modularity of a link O-D for topic Z is calculated as:

$$MOD_Z\ (O, D) = P_Z\ (AM - EM)$$

where P_Z is the probability of topic Z, AM is actual number of mentions, and EM is expected number of mentions on the link O-D.

Figure 2 illustrates modularity flows of two topic graphs: (a) NBA finals (b) democratic primaries. The mentions on NBA finals have symmetrical flows which suggest on-going conversations between pairs of states. A distinct asymmetrical flow is observed from Ohio to Florida, which suggests Cleveland fans mentioning Heat fans, but their tweets are not replied. On the other hand, the mentions on democratic primaries are dominated by asymmetrical flows of mentions. The three big states CA, NY and TX receive more mentions than expected. One distinct difference between the two topics is that mentions on democratic primaries reach out to longer distance states without reciprocating conversations, whereas NBA is being discussed between close by states with on-going conversations.

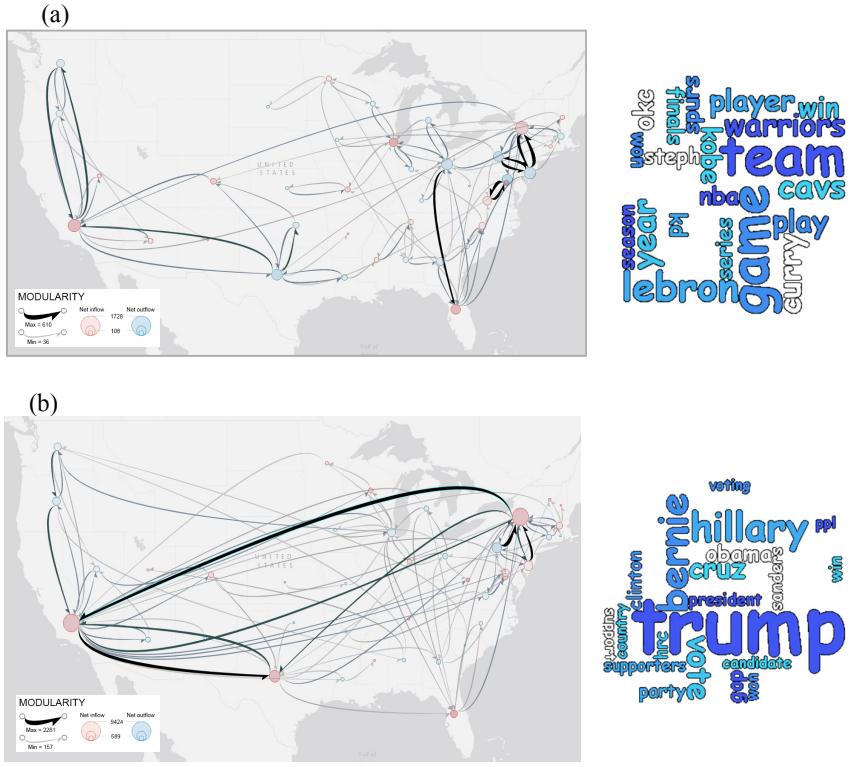

Figure 2. State-to-state modularity flows of user mentions for two distinct topics: (a) NBA Finals (Prob: 0.12) (b) primaries and caucuses (Prob: 0.15). The width and color value of each flow is proportional to its modularity value. Node size illustrates the total modularity; and blue circles depict negative net flow whereas red circles depict positive net flows. Word clouds illustrate the set of frequently co-occurring words for each topic where the size of each word is proportional to its probability of co-occurrence or popularity within that topic.

5 Conclusion and Future Work

We introduced a novel approach for extracting and visualizing geo-social semantics from the bi-directional user mention network on Twitter. The results highlighted distinct geographic patterns of symmetry, distance and clustering for user mentions. A major limitation of this study is that the temporal variation in topics is ignored. Future work is needed to incorporate temporal flow binning, and train the topic model to extract temporally varying patterns of mention topics between origin-destination pairs.

References

1. Backstrom, L., Sun, E., Marlow, C.: Find me if you can: improving geographical prediction with social and spatial proximity. In: Proceedings of the 19th international conference on World wide web, pp. 61-70. ACM, (Year)
2. Mok, D., Wellman, B., Carrasco, J.: Does distance matter in the age of the Internet? Urban Studies 47, 2747-2783 (2010)
3. Garc, R., #237, a-Gavilanes, Mejova, Y., Quercia, D.: Twitter ain't without frontiers: economic, social, and cultural boundaries in international communication. Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing, pp. 1511-1522. ACM, Baltimore, Maryland, USA (2014)
4. Vasi, I.B., Suh, C.S.: Protest in the Internet Age: Public Attention, Social Media, and the Spread of "Occupy" Protests in the United States. (2013)
5. Takhteyev, Y., Gruzd, A., Wellman, B.: Geography of Twitter networks. Social Networks 34, 73-81 (2012)
6. Park, P., Weber, I., Mejova, Y., Macy, M.: The mesh of civilizations and international email flows. WebSci 2013 Proceedings. ACM (2013)
7. Kylasa, S.B., Kollias, G., Grama, A.: Social ties and checkin sites: Connections and latent structures in Location Based Social Networks. In: Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, pp. 194-201. ACM, (Year)
8. Adamic, L.A., Lento, T.M., Adar, E., Ng, P.C.: Information evolution in social networks. arXiv preprint arXiv:1402.6792 (2014)
9. McCallum, A., Wang, X., Corrada-Emmanuel, A.: Topic and role discovery in social networks with experiments on enron and academic email. Journal of Artificial Intelligence Research 30, 249-272 (2007)
10. Tamburrini, N., Cinnirella, M., Jansen, V.A., Bryden, J.: Twitter users change word usage according to conversation-partner social identity. Social Networks 40, 84-89 (2015)
11. Lansley, G., Longley, P.A.: The geography of Twitter topics in London. Computers, Environment and Urban Systems 58, 85-96 (2016)
12. Pozdnoukhov, A., Kaiser, C.: Space-time dynamics of topics in streaming text. In: Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks, pp. 1-8. ACM, (Year)
13. Hu, B., Ester, M.: Spatial topic modeling in online social media for location recommendation. In: Proceedings of the 7th ACM conference on Recommender systems, pp. 25-32. ACM, (Year)
14. Yardi, S., Boyd, D.: Tweeting from the Town Square: Measuring Geographic Local Networks. In: ICWSM. (Year)
15. Yan, X., Guo, J., Lan, Y., Cheng, X.: A biterm topic model for short texts. In: Proceedings of the 22nd international conference on World Wide Web, pp. 1445-1456. International World Wide Web Conferences Steering Committee, (Year)
16. Malik, S., Smith, A., Hawes, T., Papadatos, P., Li, J., Dunne, C., Shneiderman, B.: TopicFlow: visualizing topic alignment of Twitter data over time. Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 720-726. ACM, Niagara, Ontario, Canada (2013)

17. Gerber, M.S.: Predicting crime using Twitter and kernel density estimation. Decision Support Systems 61, 115-125 (2014)
18. Hong, L., Davison, B.D.: Empirical study of topic modeling in twitter. In: Proceedings of the First Workshop on Social Media Analytics, pp. 80-88. ACM, (Year)
19. Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P., Tomkins, A.: Geographic routing in social networks. Proceedings of the National Academy of Sciences of the United States of America 102, 11623-11628 (2005)
20. McCallum, A.K.: {MALLET: A Machine Learning for Language Toolkit}. (2002)
21. Andrienko, G., Andrienko, N., Mladenov, M., Mock, M., Pölitz, C.: Discovering bits of place histories from people's activity traces. In: Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on, pp. 59-66. IEEE, (Year)
22. Guo, D.: Flow Mapping and Multivariate Visualization of Large Spatial Interaction Data. vol. 15, pp. 1041-1048. IEEE, IEEE Transactions on Visualization and Computer Graphics (2009)