SCIENCE CHINA
Information Sciences

• LETTER •

May 2020, Vol. 63 159202:1–159202:3
https://doi.org/10.1007/s11432-018-9534-5

GotU: leverage social ties for efficient user localization

Zidong YANG, Shibo HE & Jiming CHEN*

College of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China

Received 18 May 2018/Accepted 19 June 2018/Published online 16 October 2019

Citation Yang Z D, He S B, Chen J M. GotU: leverage social ties for efficient user localization. Sci China Inf Sci, 2020, 63(5): 159202, https://doi.org/10.1007/s11432-018-9534-5

Dear editor,
In recent years, there has been a rapid increase in online social networks (OSNs), which produces huge amounts of user-generated content (UGC) at an unprecedented rate every day. For example, it is reported that 58 million tweets are posted every day on Twitter. This vast data reservoir has provided opportunities for solving many problems, such as link modeling, user privacy, and community detection [1].

By doing so, we can link posted content to locations and thus enable applications such as effective advertisement. Owing to its importance and broad applications, dynamic localization has attracted much attention recently [2].

Although OSNs (e.g., Twitter) allow users to share their location, only a few users post their location on OSNs owing to privacy concerns. It is reported in [3] that only 16% of Twitter users provide their location in their profiles (city level) and very few people (0.5%) attach GPS tags to their tweets. Previous studies [2, 4] exploited the content of microblogs posted by the target users to infer their instant locations. They were not successful as location information in a single user's microblogs is typically sparse and thus not sufficient for dynamic localization.

System design. We propose a novel approach named GotU, which can efficiently leverage social ties for dynamic user localization in OSNs. An essential notion used in GotU is statistical location, which characterizes the most likely cities where the target OSN users are located. It can be observed that all historical data from the user and his/her friends can be utilized for the estimation of statistical locations. Furthermore, we introduce the concept of co-location friends in GotU to identify the set of social friends who are likely to stay in the same city as the target user. Thus, we can judicially choose social friends for a better estimation of statistical locations. Based on the statistical localization, we further design dynamic localization to estimate the instant locations of the target user by extracting and matching the point of interest (POI) names in the target user's microblogs. The primary contributions of this study are as follows: (1) We propose a hierarchical structure to solve the localization problem. Thus, we solve the utilization problems of friends' microblogs. (2) We devise an algorithm for detecting co-location friends by utilizing the information from social ties topology. The framework of GotU is depicted in Figure 1(a), which consists of two steps.

Statistical localization. In this step, our goal is to estimate k cities in which the target user is most likely to live. The main idea is to leverage both user content information and social ties information to improve estimation accuracy. Notably, for a target user who posts limited tweets, it is difficult to localize him/her only according to his/her content. Therefore, we leverage his/her social ties to retrieve more tweets to refine the localization. Further, we introduce the concept of co-location friends – social friends who live near the target

* Corresponding author (email: cjm@zju.edu.cn)

user. Finally, we use all the tweets from the target user and co-location friends to estimate the target user's location.

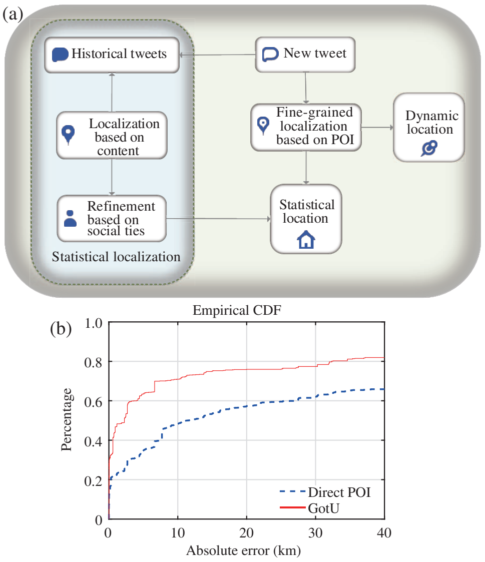

Figure 1 (Color online) (a) System framework; (b) localization results.

For raw-text analysis, a dedicated content analyzer can perform rough user-location estimation only based on the literal content. Researchers have proposed many approaches to localize a user according to the text. Here, we adopt a location predictor based on the Bayesian classification proposed in [5], which provides city-level predictions along with their confidential scores. By feeding raw tweet texts into this analyzer, a list of estimated cities can be obtained[1]. For example, after analyzing a user's tweets, we may find that the scores for his/her possible location being San Francisco, Portland, and Seattle are -1145.49, -1167.66, and -1179.09, respectively.

To fully utilize social ties information, we must distinguish co-location friends from others. It is observed that people living in the same city have a higher probability of becoming OSN friends owing to spatial approximation. Based on the above observation, it is possible to infer the labels of nodes (co-location or not) according to the topology. In other words, we aim at assigning each node with a proper label (local or nonlocal), which maximizes the total probability of forming the observed topology structure.

Let $G(V,E)$ denote the graph of the immediate social network of the target user and $U \subset V$ is the set of co-location users. We first derive the probability of producing an existing topology given U. For simplicity, different links are assumed to be formed independently. Let us consider the existing links between all local nodes. Let $P_{\rm local}$ denote the probability to establish a connection between two local nodes. Then, the probability of forming this kind of links is $\prod_{(i,j)\in E, i\in U\wedge j\in U} P_{\rm local}$.

Moreover, the missing links between two co-location nodes also provide opposing evidence for the current assignment of labels and therefore should be taken into consideration. The probability of observing this kind of missing links is $\prod_{(i,j)\notin E, i\in U\wedge j\in U} 1-P_{\rm local}$.

Similarly, consider the links between two nodes, at least one of which is not a co-location node. The probabilities of observing the presence and absence of these links are $\prod_{(i,j)\in E, i\notin U\vee j\notin U} P_{\rm nonlocal}$ and $\prod_{(i,j)\notin E, i\notin U\vee j\notin U} 1-P_{\rm nonlocal}$, respectively.

The probability of producing the entire network can be obtained by calculating the product of all four probabilities above, which is shown as follows:

$$P(U)=\prod_{\substack{(i,j)\in E\\ i\in U\wedge j\in U}} P_{\rm local}\prod_{\substack{(i,j)\notin E\\ i\in U\wedge j\in U}} 1-P_{\rm local}\prod_{\substack{(i,j)\in E\\ i\notin U\vee j\notin U}} P_{\rm nonlocal}\prod_{\substack{(i,j)\notin E\\ i\notin U\vee j\notin U}} 1-P_{\rm nonlocal}. \quad (1)$$

Our goal is to maximize $P(U)$ by choosing a proper U, which is shown as follows:

$$\arg\max_{U} P(U). \quad (2)$$

In practice, we adopt the log form of probability $P(U)$ to avoid underflow in computation. Please refer to supporting information for more details.

After the detection, all tweets from the target user and detected co-location friends will be fed into the content analyzer and the top k cities with the highest probability will be selected. The result will be regarded as a refined statistical location.

The computational complexity in this step is $O(|V|)$, and hence, we use a greedy method for determining the label of a friend: we invert his/her label if the probability in Eq. (1) increases. Moreover, because we must compute the probability for all possible cities, the computational complexity is not related to k.

Fine-grained localization. For fine-grained localization, we attempt to determine the real-time location from where the user posts a specific tweet. In this step, the main goal is to extract the POI information from the tweets and compare the obtained POI with records in the existing database (gazetteer). As POI names are not unique (e.g., chain stores) and may appear in different cities, it is difficult to determine the exact user location only according to the POI names. We propose to exploit the statistical locations obtained in the

[1] We use top five cities in the evaluation.

first step to narrow down the search area. It is observed that this approach can significantly reduce the computational complexity and increase the accuracy of dynamic localization.

For identifying location names in tweets, we adopt a method that extracts the location name according to some language pattern rules. Based on observations from real-world data, POI names in tweets exhibit similar patterns. For example, "Going on an evening hike. @ Runyon Canyon Park" – such POI names usually begin with the symbol "@" and consist of some consecutive words with capital initial letters. This pattern provides an effective way for POI extraction.

The POI name mentioned in tweets is not necessarily identical to the name used in the gazetteer, even though they refer to the same place. Although the names may not be the same, they are quite similar. We use the Jaccard similarity as the similarity score between the tweet POI name and gazetteer POI name, which is computed using (3).

$$\mathrm{sim}(x,y)=\frac{|\mathrm{word}(x)\cap\mathrm{word}(y)|}{|\mathrm{word}(x)\cup\mathrm{word}(y)|},\tag{3}$$

where $\mathrm{word}(x)$ and $\mathrm{word}(y)$ are the sets of words used in the tweet POI name and gazetteer POI name, respectively. The score will be 1 if the two names are identical whereas it will be 0 if the two names share nothing in common.

To compare POI results among the k predicted residential cities, we adjust the Jaccard similarity by multiplying it with the confidential score of the corresponding city (Eq. (4)).

$$\mathrm{sim}_{\mathrm{adjusted}}(x,y,j)=\mathrm{sim}(x,y)\times|\mathrm{conf}(j)|,\tag{4}$$

where $\mathrm{conf}(j)$ is the confidential score of city j.

Thus, the POI with the highest adjusted similarity score is considered as the predicted place where the target user posted this tweet. The computational complexity in this step is $O(k)$ for that we must consider the k candidate cities one by one.

Evaluation. We test our algorithms in a real-world data set. The dataset used in our evaluation is collected from official Twitter API[2] during December 2015.

First, we evaluate the prediction results using statistical localization, which is also the first step of GotU. In this part, the system utilizes both tweets and social ties information of a user to identify the k most likely cities where he/she must be located. We use "ACC@K" as our evaluation metric, which is the accuracy of top-k predicted cities. We compare GotU with two baselines: (1) "User Only" (UO). This method only uses the target user's information; (2) "All Friend" (AF). This method treats all friends equally. When we consider "ACC@1", GotU achieves an accuracy of 42.7% whereas UO and AF achieve the accuracy of 38.7% and 40.4% respectively. Similar patterns can be observed for "ACC@3" and "ACC@5". Specifically, GotU achieves an accuracy of 67.6% for "ACC@3" and 69.3% for "ACC@5".

For dynamic localization, we compare our approach with "Direct POI", which localizes users using POI names directly. As shown in Figure 1(b), GotU can locate approximately 47% POIs within 1 km and 71% POIs within 10 km, thus outperforming the baseline approach (22% and 48% respectively).

Conclusion and future work. We studied the problem of identifying the dynamic locations of an OSN user. To achieve more accurate results, we designed a two-stage system, consisting of statistical localization and fine-grained localization. In future work, we will focus on localizing multiple users simultaneously through statistical localization and exploiting the spatial and temporal correlation (e.g., traces) for fine-grained localization.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant No. 61672458).

Supporting information Detection of nearby friends. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Gong J B, Gao X X, Cheng H, et al. Integrating a weighted-average method into the random walk framework to generate individual friend recommendations. Sci China Inf Sci, 2017, 60: 110104

2 Li G L, Hu J, Feng J H, et al. Effective location identification from microblogs. In: Proceedings of the 30th International Conference on Data Engineering, 2014. 880–891

3 Li R, Wang S J, Deng H B, et al. Towards social user profiling: unified and discriminative influence model for inferring home locations. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2012. 1023–1031

4 Cheng Z, Caverlee J, Lee K. You are where you tweet: a content-based approach to geo-locating twitter users. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, 2010. 759–768

5 Han B, Cook P, Baldwin T. A stacking-based approach to twitter user geolocation prediction. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, 2013. 7–12

[2] https://apps.twitter.com/.