Harnessing geo-located Tweets as a new geographical data source: Extracting mobility data from massive collection of geo-located Tweets

April 1, 2016

Abstract

geo-located Twitter data is attracting increasing attention from research communities

the increasing popularity of accessing today's pervasive social media platform has enabled a new geographical data.

In this paper, we present an efficient approach for geo-located data management and processing

1 Introduction

Understanding human mobility patterns is of great importance for a broad range of applications from urban planning [1], traffic management [2], and even the spatial spread of epidemic diseases [3]. Earlier research efforts relied on low resolution mobility data to understand human mobility patterns, such as using census records [4] or questionnaires by asking volunteers to report the trace of bank notes [5]. However, the lack of detailed human movements with fine-grained spatial and temporal granularity limits the findings to capture mobility patterns of individuals [6, 7]. More recent approaches to collecting detailed mobility data GPS trackers [**?**, **?**] and mobile phone call records [**?**, **?**, **?**],

emerging as a new source for mobility data, today's pervasive Location Based Social Media (LBSM) platforms (e.g., Twitter and Foursquare) offer continuous spatial Big Data streams with massive amount of detailed and frequently updated user digital traces in the form of real-world trails and footprints [8]. One significant advantage of LBSM data streams is the large spatial coverage, for example, researchers have used geo-located Twitter data for studying global mobility patterns [**?**], which is otherwise impossible by using other mobility datasets (e.g., GPS traces and mobile phone call records). In addition, the publicly available LBSM data streams offer unique

opportunities for conducting reproducible scientific findings regarding the concerns of infringement on individual privacy, such as using mobile phone call records [**?**, **?**, **?**].

However, there are some limitations and complexities in directly using the LBSM data for study human mobility patterns.

In this connection, it becomes increasingly popular for researchers to exploit the publicly accessible mobility data captured from today's pervasive Location Based Social Media (LBSM) platforms (e.g., Foursquare and Twitter). LBSM enables users to attach their current location as a geo-tag to the message they will post, which is derived from either the GPS or Wi-Fi positioning with a high position resolution down to 10 meters [**?**]. A LBSM Big Data emerges when millions social media users constantly posting messages.

Although Twitter data is accessible from its dedicated APIs, the collected data cannot be shared among Furthermore, as the data is collected in a continuous fashion, the accumulated data

In this regard, this paper provides a scalable computing framework for efficient data processing. In particular, this framework utilizes

Although Twitter allows users to gain access to the data streams, the obtained data cannot be shared directly with since the content of each Twitter message concerns user privacy. Therefore, when conducting research using Twitter data, one must download the data

The remainder of this paper is organized as follows. Section describes the

Data collection

In general, Twitter (`http://www.twitter.com`) provides two kinds of search APIs for retrieving Twitter messages: REST API and Streaming API (`http://dev.twitter.com/overvi`
Twitter REST API is used to search tweets that belong to a particular user account and its followers

If your intention is to conduct singular searches, read user profile information, or post Tweets, consider using the REST APIs instead.

an alternative way to collect large is through the Twitter Streaming API (application programming interface)

Geo-located tweets can be downloaded using

REST APIs responses are available in JSON. limitation

Streaming APIs, also known as Twitter Fire

The Streaming APIs give developers low latency access to Twitters global stream of Tweet data. A proper implementation of a streaming client will be pushed messages indicating Tweets and other events have occurred, without any of the overhead associated with polling a REST endpoint.

If your intention is to conduct singular searches, read user profile information, or post Tweets, consider using the REST APIs instead.

Twitter offers several streaming endpoints, each customized to certain use cases.

2 Data management and processing

2.1 Accommodating massive Twitter Dataset with Hadoop

Apache Hadoop is an open source software framework designed to facilitate data intensive computing on large commodity clusters. It combines a distributed file system, namely Hadoop Distributed File System(HDFS) with MapReduce programming paradigm, which can be applied to wide range of data-intensive problems. MapReduce is a programming model and an associated software framework designed to process massive data in a distributed fashion. MapReduce breaks the entire computation into small tasks and schedule them among different computing nodes. MapReduce consists of two main stages: map and reduce. In the map stage, input data is converted into series of intermediate $\langle key, value \rangle$ pairs. Customized computation will be performed in the reduce stage based on the same intermediate key. This framework provides parallelization since both map and reduce tasks are considered independent and can be done in parallel. More importantly, it provides scalability in relation to the growth of data size, where Hadoop can scale to more computing nodes in the cluster to maintain the performance.

Our framework benefits from using Hadoop in both data management and processing. First, since the input data is large it is desirable to store it on multiple machines. Second, by using Hadoop we can parallelize the computational tasks and make the data processing faster and more efficient. Further, by adopting the MapReduce computing paradigm, we can model each individual user's trajectory by treating user's unique ID as key, and summarize the movement flows among different spatial units by utilizing the ID of the spatial unit as key. The details will be introduced in the following section.

2.2 Data extraction based on geographical regions

2.3 Constructing Twitter user space-time trajectories

3 Case study

In this study, the geo-located tweets were downloaded using the Twitter Streaming API, where we specified a geographical bounding box as an area-of-interest to retrieve all the geo-located tweets that fall within it. To ensure the complete coverage over Great Britain, we have set the bounding box to British Isles (lower left coordinates in (latitude, longitude): (49.497, -14.854)

and upper right coordinates: (61.186, 2.637)), which also includes the whole area of Ireland and a part of France. We have implemented a data crawler and continuously collected 7-month data (1st June 31st December, 2014) with over 101.8 million tweets and 60 GB in size. During the data collection phase, the data crawler did not encounter any issue regarding whether it exceeds the data quota by the 1% policy mentioned in (Hawelka et al., 2014). It means we have managed to download all the geo-located tweets for the given bounding box. In particular, to showcase the overall spatial coverage of the collected geo-located tweets, a density map of the Twitter user locations in British Isles for July 2014 is shown in Figure 4, where the collected points alone visually reveal the geography of cities and countries, e.g. the clusters with high density of tweets (in color red) correspond to the skeleton of major cities.

Conclusion

References

1. Zheng, Y., Li, Q., Chen, Y., Xie, X., and Ma, W.-Y., 2008. Understanding mobility based on gps data. In Proceedings of the 10th international conference on Ubiquitous computing, pp. 312-321. ACM.

2. Jiang, B., Yin, J., and Zhao, S., 2009. Characterizing the human mobility pattern in a large street network. Physical Review E, 80(2):021136.

3. Belik, V., Geisel, T., and Brockmann, D., 2011. Natural human mobility patterns and spatial spread of infectious diseases. Physical Review X, 1(1):011001.

4. Greenwood, M. J., 1985. Human migration: Theory, models, and empirical studies. Journal of regional Science, 25(4), pp. 521-544.

5. Brockmann, D., Hufnagel, L., and Geisel, T., 2006. The scaling laws of human travel. Nature, 439(7075), pp. 462-465.

6. Gonzalez, M. C., Hidalgo, C. A., and Barabasi, A.-L., 2008. Understanding individual human mobility patterns. Nature, 453(7196), pp.779-782.

7. Jurdak, R., Zhao, K., Liu, J., AbouJaoude, M., Cameron, M., Newth, D., 2015. Understanding Human Mobility from Twitter. PLoS ONE 10, e0131469. doi:10.1371/journal.pone.0131469

8. Thatcher, J. (2014). Living on fumes: Digital footprints, data fumes, and the limitations of spatial big data. International Journal of Communication, 8, 1765-1783.

9. Frank, M.R., Mitchell, L., Dodds, P.S. and Danforth, C.M., 2013. Happiness and the patterns of life: A study of geolocated tweets. Scientific reports, 3. doi:10.1038/srep02625

10. Sakaki, T., Okazaki, M., Matsuo, Y., 2010. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors, in: Proceedings of the 19th International Conference on World Wide Web, WWW 10. ACM, New York, NY, USA, pp. 851860. doi:10.1145/1772690.1772777

11. Liu, J., Zhao, K., Khan, S., Cameron, M., Jurdak, R., 2014. Multi-scale Population and Mobility Estimation with Geo-tagged Tweets. ArXiv14120327 Phys.

12. Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., Ratti, C., 2014. Geo-located Twitter as proxy for global mobility patterns. Cartography and Geographic Information Science, 41, 260271. doi:10.1080/15230406.2014.890072