Advanced Science and Technology Letters
Vol.66 (Networking and Communication 2014), pp.24-27
http://dx.doi.org/10.14257/astl.2014.66.07

Collecting Twitter Data using PlanetLab

Seunghun Lee[1], Sungcheon Lee[2] and Hyun-chul Kim[1*],

[1] Dept. of Computer Software Engineering, Sangmyung University,
300, Anseo-dong, Dongnam-gu, Cheonan, 330-720, South Korea
[2] Dept. of Stock and Finance, Sangmyung University,
300, Anseo-dong, Dongnam-gu, Cheonan, 330-720, South Korea
mr.leesh90@gmail.com, sc.l@me.com, hkim@smu.ac.kr
*Corresponding author

Abstract. As an increasing number of people are now using online social network services, there has recently been much interest in collecting and analyzing online social network data, particularly in a large scale. This paper proposes a methodology for collecting twitter dataset using PlanetLab, an open platform for developing, deploying, and accessing planetary-scale network services.

Keywords: twitter, data, PlanetLab

1 Introduction

As an increasing number of people are now using online social network services, there has recently been much interest in collecting and analyzing online social network data, particularly in a large scale. While Twitter provides an API with which we can collect users' twitted data, the API poses the following limitations: (i) the service limits the number of collectible tweets only to the amount of less than 1% of the total number of tweets, and (ii) a single (IP) machine often can not continuously collect even the 1% of the sampled tweets, because they enforce restrictions on data collection on a per-IP basis based on certain criteria; the number of tweets crawled, the collection period, etc. To overcome the limitations, we propose a system for collecting twitter dataset using PlanetLab[1], an open platform for developing, deploying, and accessing planetary-scale network services.

This paper is structured as follows: In Section 2, we review the related work. Section 3 proposes our own methodology and system. We conclude this paper in Section 4.

2 Related Work

While Online Social Networks such as Twitter, Weibo, Facebook, and Google+ provide APIs to help us get their data, they typically enforce restrictions on data

ISSN: 2287-1233 ASTL

collection (e.g., per-IP-based rate limiting, etc.) in order to defend their systems from offensive and aggressive data crawling which might affect and degrade the system performance. This in turn makes researchers who are interested in collecting and analyzing such data, particularly in a large scale, unable to collect a necessary amount of enough data for their research activities. To overcome such a limitation posed by those service providers, we propose a large-scale data collection method and build a system.

2.1 Crowd Crawling: Towards Collaborative Data Collection for Large-Scale Online Social Networks

Crowd Crawling [2] proposed a data collection system using the PlanetLab. The architecture of the Crowd Crawling consists of two major system modules: TAM (Task Assignment Module) and RCM (Result Collection Module). TAM is the controller responsible for managing component computers, while RCM is responsible for aggregation of the collected data through TAM's commands. Using the proposed system, their system successfully collected 2.2 Millions of Weibo users just in a 24 hour period. According to the authors, their system was able to collect the site data in a large scale, while minimizing the load placed on the social network system itself during the data collection process. We develop a similar system, which we plan to open with a publicly available front-end system through a simplified web interface, which we believe will help researchers interested in collecting such data from social networking services like Twitter.

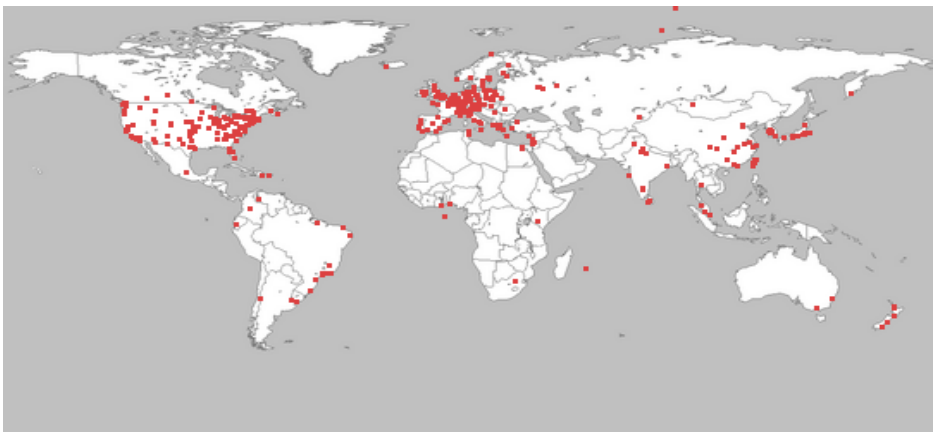

Fig. 1. PlanetLab nodes across the globe (as shown in http://www.planet-lab.org).

2.2 PlanetLab

PlanetLab is an open platform for developing, deploying, and accessing planetary-scale network services. It is a global research network that supports the fast

development and deployment of new network services. Since the birth of it in 2003, more than thousands of researchers at top academic institutions and industrial research labs have used PlanetLab to develop new technologies for distributed storage, network mapping, p2p systems, distributed has tables, and query processing. It currently consists of 1,333 nodes at 634 sites distributed across the globe, as shown in Figure 1.

3 Data Collection System

Our proposed system consists of the three components, as depicted in Figure 2; Twitter site, PlanetLab crawlers, and our data collection command and merge server where all the data collection commands and activities are orchestrated. Out of the 1,333 nodes in the PlanetLab system, we use 281 nodes that was available when we built our system.

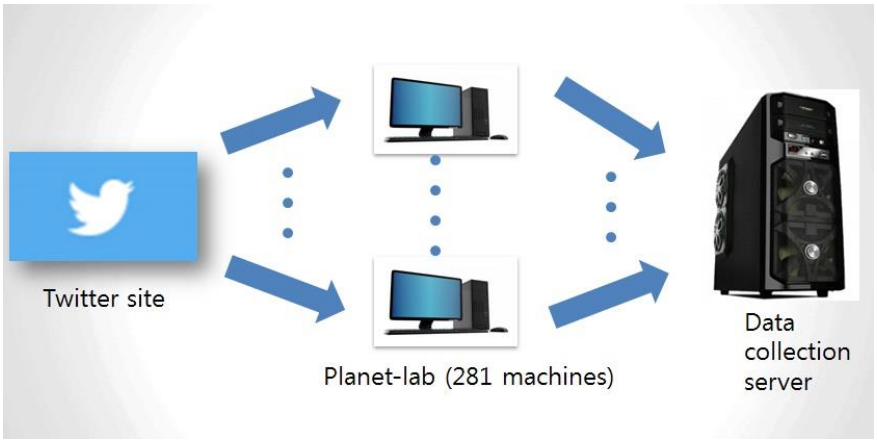

Fig. 2. System Architecture

Communications between PlanetLab machines (slice computers) and data collection servers are controlled and managed by our command server (which currently is co-located at the Data collection server), a slightly modified version of CoDeploy, which is a PlanetLab tool. As soon as a communication channel between the data collection server and PlanetLab slice machines are established, our commander server gives commands to the 281 target slice machines in order to crawl tweet data using the Twitter Streaming API, in parallel. As each slice has its own limitation on the amount of allocated storage, collected twitter data at each slice machine are automatically transferred to the data collection server at every hour. Once the collected data arrived at the collection server, the server first eliminates redundant tweet data and then stores the remaining data. We use MongoDB as well as text file format to collect, retrieve, and store the tweet data.

4 Conclusions

In this paper, we propose and present our own system for large-scale twitter data collection using PlanetLab. We will soon release our system publicly available to help researchers interested in collecting and analyzing datasets from online social networks like Twitter.

Acknowledgments. This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (No. 2013-R1A1A2010474).

References

1. PlanetLab, http://www.planet-lab.org (2014)
2. Ding, C., Chen, Y., Fu, X.: Crowd Crawling: Towards Collaborative Data Collection for large-scale Online Social Networks. COSN ’13 Proceedings of the first ACM conference on Online social networks, 183-188(2013)
3. Foster, I., Kesselman, C.: The Grid: Blueprint for a New Computing Infrastructure. Morgan Kaufmann, San Francisco (1999)
4. Czajkowski, K., Fitzgerald, S., Foster, I., Kesselman, C.: Grid Information Services for Distributed Resource Sharing. In: 10th IEEE International Symposium on High Performance Distributed Computing, pp. 181--184. IEEE Press, New York (2001)
5. Foster, I., Kesselman, C., Nick, J., Tuecke, S.: The Physiology of the Grid: an Open Grid Services Architecture for Distributed Systems Integration. Technical report, Global Grid Forum (2002)

