SAMPLING TECHNIQUE FOR SREAMING DATASET USING SENTIMENT ANALYSIS

Dr. U.M. Fernandes Dimlo[1] **Mr.P.Sai Yadav** [2] **P.Suhasini**[3]

[1] Professor and HOD, Department of CSE, Narsimha Reddy Engineering College, Maisammaguda, Dhulapally, Secunderabad, Telangana, India

[2] Assistant Professor , Department of CSE, Narsimha Reddy Engineering College, Maisammaguda, Dhulapally, Secunderabad, Telangana, India

[3] M. Tech Student, Department of CSE, Narsimha Reddy Engineering College, Maisammaguda, Dhulapally, Secunderabad, Telangana, India.

Abstract **-** Researchers have begun studying content obtained from microblogging services such as Twitter to address a variety of technological, social, and commercial research questions. The large number of Twitter users and even larger volume of tweets often make it impractical to collect and maintain a complete record of activity; therefore, most research and some commercial software applications rely on samples, often relatively small samples, of Twitter data. For the most part, sample sizes have been based on availability and practical considerations. Relatively little attention has been paid to how well these samples represent the underlying stream of Twitter data. To fill this gap, this article performs a comparative analysis on samples obtained from two of Twitter's streaming APIs with a more complete Twitter dataset to gain an in-depth understanding of the nature of Twitter data samples and their potential for use in various data mining tasks.

I INTRODUCTION

Microblogging is an increasingly popular form of lightweight communication on the Web. Twitter as a typical and quickly emerging Microblogging service has attracted much attention. Millions of Twitter users around the world form a massive online information network by initiating one-way "following" relationships to others. Twitter users post brief text updates, which are commonly known as tweets, with at most 140-characters. The tweets posted by a user are immediately available to his direct followers, and can be quickly disseminated through the network via retweeting. Different from traditional blog platforms, where users write long articles with low update frequency, Twitter generates short and real-time messages in large volume daily. Some studies of the Twitter network reveal a variegated usage including daily chatter, conversation, information sharing, news reporting [Java et al. 2007], and a diverse topic
coverage such as arts, family and life, business, travel, sci-tech, health, education, style, world, and sports [Zhao et al. 2011]. Many researchers have analyzed Twitter content and made interesting observations with real business value. For example, Sakaki et al. [2010], utilize Twitter to detect earthquakes; Bakshy et al. [2011] study different methods of identifying influential Twitter users, which may be useful for online marketing and targeted advertising; and Bollen et al. [2011] analyze Twitter user sentiment to predict the stock market.

One obstacle to using Twitter data is its huge size, as measured by the size of the user base, the volume of tweets, and the velocity of updates. The number of registered user profiles on Twitter reached half a billion in 2012 [Semiocast 2012], and collectively, Twitter users now send over 400 million tweets every day [Bennett 2012]. These numbers keep growing rapidly. It is challenging for third-party researchers and developers to collect and manage such a huge amount of data.

Twitter provides API functions to facilitate third-party users to access the data (https://dev.twitter.com/docs/). There are two main types of Twitter APIs: the REST API and the stream API. The REST API supports queries to Twitter user accounts and tweets, and it usually has very strict limits on the query rate (e.g., 150 requests per hour). Although the REST API provides flexible access to Twitter data from almost every angle, the rate limits make it not suitable for collecting large amounts of Twitter data and monitoring updates. On the other hand, the stream API provides almost realtime access to Twitter's global stream of public tweets. Once the connection is built, tweet data are pushed into the client without any of the overheads incurred by pulling data from the REST API. The stream API produces near real-time samples of Twitter's public tweets in large amounts. Owing to the advantages of the Twitter stream API, it is used as the data source for many applications and mining tasks, for example, topic modeling [Hong et al. 2012; Pozdnoukhov and Kaiser 2011], disease outbreak surveillance [Sofean and Smith 2012], and popular trend detection [Mathioudakis and

Koudas 2010]. The convenience and immediacy of the stream API makes it a common source of Twitter data for a variety of research tasks. However, prior research has not addressed the issue of how well the sample data provided by the stream API represent the original data, and if do not, toward which properties the sample data might be biased.

In this work, we focus on characterizing the sample data from the Twitter stream API, studying possible sampling bias, if any, and understanding the implications of the findings to related applications. The Twitter stream API has different access priorities.

II RELATED WORK

The huge volume of user-generated content in modern online social networks presents challenges to researchers for collecting and analyzing these data. A common practice to deal with this problem is to generate and analyze a representative sample of the complete dataset. There are two main issues for generating the sample: What is a good sampling strategy, and what is a good sampling ratio. In the case of the Twitter streaming API, the sample data are generated by some unknown strategies designed by Twitter with approximately fixed sampling ratios. Therefore, our focus in this work is on the unresolved question of whether the sample data generated by the Twitter streaming API are good enough for various mining and analysis tasks.

Very recent work by Morstatter et al. [2013] studies the same problem; however, there are important differences between their work and ours. The main difference is that they use a sample dataset collected from the Twitter stream API that focuses on a particular event: The Syria conflict from December 2011 to

January 2012. We analyze a dataset that is not event-specific to provide more general observations. Their work also does not address the issue of sampling ratio, whereas we study two different sampling ratios and discuss their effects on the quality of the data obtained. In terms of methodology, Morstatter et al. measure the daily sampling ratio, whereas we also study the retweet ratio and the user tweet frequency distribution to provide a more comprehensive analysis. When studying the tweet content, they analyze the correlation of the ranks of the top hashtags and compare the topic distribution of the sample data with that of the complete data. We do not compare the topic distributions because we consider topic alignment across unlabeled datasets to be difficult, subjective, and unreliable. In this work, we study a rich set of terms in the tweet content including text terms, hashtags, URLs, and URL domains, and discuss the similarity of the sample data using these content terms to the complete data based on vocabulary coverage and frequency correlations. We also perform a sentiment classification task to compare the results obtained from the sample datasets and the complete dataset. In order to study user relationships, their work focuses on the user retweet network, whereas we study not only the user retweet relationships but also the mention relationships. Finally, their work analyzes the geolocation distribution of the tweets. However, because our dataset is based on Singapore Twitter users, the tweets are mainly located in Singapore; thus, geolocation distribution adds no new information.

Several other works discuss different Twitter data sampling methods. For example, Ghosh et al. [2013] study an expert generated tweet set and compare it with a random Twitter sample. They find that each dataset has its own relative merits. The expert tweets are significantly richer in information, more trustworthy, and capture breaking news marginally earlier. However, the random sample preserves certain important statistical properties of the entire dataset and captures more conversational tweets. Choudhury et al. [2011] propose a diversity-based sampling approach to generate topic-centric tweet set. Our work does not study new sampling approaches, rather it investigates the characteristics of the existing and widely used Twitter samples. We focus on understanding whether the quality of the samples is good enough for various mining and analysis tasks.

The topic of network sampling and the effect of the imperfect data on the common network measurements have been widely studied. The earlier work of Granovetter proposes a network sampling algorithm that allows estimation of the basic network properties [Granovetter 1976]. Later, many common network sampling techniques are studied such as snowball sampling, random-walk–based sampling, node sampling, and link sampling [Lee et al. 2006; Yoon et al. 2006; Leskovec and Faloutsos 2006; Maiya and Berger-Wolf 2011]. These works compare the structural properties of the sample networks obtained by different methods with those of the original network and address the sampling bias of different methods. There are also works that discuss the effect of data errors and missing data on common network measures (e.g., centrality) [Kossinets 2003; Borgatti et al. 2006; Costenbader 2003]. Recently, William and Samuel [2010] study the forest fire network sampling method with different seed user selection strategies, and discuss their impact on the discovery of information diffusion on Twitter. However, this prior research does not apply to our problem because we study data that are sampled from the Twitter public tweet stream, not the Twitter user

network. The (unknown) sampling mechanisms used by Twitter to generate data are presumably different from the network sampling methods discussed in this prior research.

III DESIGN OF THE WORKFLOW:

In order to study sampling bias, we need the complete Twitter dataset to serve as the baseline, with which the sample datasets can be compared. However, collecting the complete Twitter stream is not practical for our study because of its cost. Instead of considering the full set of more than 500 million Twitter users, we focus on the complete set of Singapore Twitter users, which is a smaller group. We used all the tweets posted by these Singapore Twitter users within a 1-month period as the complete dataset. We also gathered all tweets by these Singapore users that appeared in the Spritzer and

Gardenhose Twitter streams during the same time span to create two sample datasets. The complete dataset was collected with the help of the social network mining research group of Singapore Management University.1 To locate the Singapore Twitter users, a set of 58 popular Singapore Twitter users were manually selected as seeds.

Initially, the user set only contained these seed users. The user set was then expanded by exploring the follower and friend lists of users in the set. A follower or a friend of a current user was added to the user set if either he specified his location to be “Singapore” or he followed at least three of the known Singapore users. In this way, a set of 151,041 Singapore Twitter users in 2012 was identified, which we believe covered the majority of the Singapore Twitter users.

After the set of users was constructed, the Twitter REST API was invoked to crawl the tweets generated by these users for a 1-month period beginning on May 1, 2012 and ending on May 31, 2012. The collected tweets formed the complete dataset, referred as Complete.

We collected two sample datasets at the same time period via the Twitter stream API using Spritzer and Gardenhose access priorities respectively. The Spritzer and Gardenhose streams output samples of the entire public tweet stream with different sampling ratios. According to Twitter, Spritzer provides an approximately 1% sample of the complete public tweets, whereas Gardenhose generates a larger sample with the sampling ratio around 10%. Twitter does not provide any description of the algorithms that generate the samples nor does it guarantee the sampling ratios to be stable.

From the sampled tweets, we extracted the subsets that were posted by the identified Singapore users. In this way, two samples of the complete dataset were obtained, referred as SampleSpritzer and SampleGardenhose, respectively. Table I provides some basic information about the datasets.

IV ANALYSIS OF RESULTS

In this section, we perform detailed comparative analysis of the collected sample and complete datasets. Specifically, we compare them in terms of the tweet statistics, content representativeness, user coverage, and user interactions. Through the comparison, we try to understand the nature of the sample datasets, for which properties the sample datasets are representative of the complete dataset, and for which properties the sample datasets are not representative, and discuss the implications of our findings for certain mining tasks.

Tweet Statistics

We first study the sampling ratio and the basic tweet statistics in this section. We perform the analysis on the datasets collected over the 1-month time period and also present results on daily bases.

We begin the analysis by examining the actual sampling ratios of the two sample datasets from the Twitter stream API and present the average daily sampling ratios and standard deviations in Table II. As shown in Table II(a), the Singapore users generate around a half million tweets a day, on average. The Spritzer and

Table II. Average Daily Sampling Ratios

(a) Daily sampling ratios for tweets.

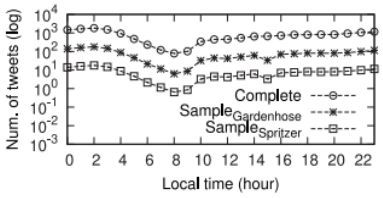
Daily statistic	*Complete*	*Sample*$_{Gardenhose}$		*Sample*$_{Spritzer}$	
	tweet#	tweet#	sampling ratio	tweet#	sampling ratio
Daily avg.	481,024	46,332	9.62%	4,634	0.96%
Std. dev.	67,446	6,637	0.15%	664	0.014%

(b) Daily sampling ratios for users.

Daily statistic	*Complete*	*Sample*$_{Gardenhose}$		*Sample*$_{Spritzer}$	
	user#	user#	sampling ratio	user#	sampling ratio
Daily avg.	35,316	15,769	44.55%	3,625	10.22%
Std. dev.	2,407	1,601	1.88%	484	0.85%

Table III. Daily Tweets and Retweets Ratios

Daily statistic	*Complete*		*Sample*$_{Gardenhose}$		*Sample*$_{Spritzer}$	
	tweet%	retweet%	tweet%	retweet%	tweet%	retweet%
Daily avg.	84.41%	15.59%	84.24%	15.76%	84.21%	15.79%
Std. dev.	0.56%	0.56%	0.54%	0.54%	0.76%	0.76%



Gardenhose Fig. 1. Average hourly tweet count of the Singapore Twitter users of a 1-month period.

samples return around 0.96% and 9.6% of them, respectively. The actual tweet sampling ratios are both slightly lower than what Twitter announced (i.e., 1% and 10%).

Table II(b) shows the sampling ratios on users each day. We find that, on average, there are around 35,000 Singapore users who generate tweets each day, and the Spritzer and Gardenhose samples capture around 10% and 45% of them, respectively. The sampling ratio for users is much higher than it is for tweets, which is not surprising; each tweet appears just once in the complete dataset, whereas a user may appear many times, thus increasing the likelihood that he will also appear in a sample.

V CONCLUSION

This article provides a descriptive study of Twitter data samples obtained from the Twitter stream API with two different access priorities (i.e., Spritzer and Gardenhose). These two data streams are data sources for a variety of research and commercial activities. By comparing the sample data with the corresponding complete dataset from different perspectives, we explore the nature of the sample data, its biases, and how well it represents the complete data stream. Our results provide insights about the sample data obtained from the Twitter stream API and provide incentives for people to use or not to use them for their research.

We find that the Twitter stream API with the Spritzer and Gardenhose access priorities provides samples of the entire public tweets with actual sampling ratios around 0.95% and 9.6%, respectively. The sample datasets truthfully reflect the daily and hourly activity patterns of the Twitter users in the complete dataset. Moreover, the sample datasets capture the approximate power-law property of the user tweeting frequency

distribution in the complete dataset but with smaller exponents. In other words, the sample datasets preserve the same scaling behavior of the user tweeting frequency distribution with the complete dataset, but tend to overestimate the proportions of low-frequency users. The overestimation is more serious when the sampling ratio is small. These observations indicate that the sample datasets, even with very small sampling ratios such as the Spritzer stream (i.e., 0.95%), are good for studying Twitter user activity patterns in general. However, researchers should be careful about the overestimation of the low-frequency users when trying to analyze users based on their activity levels (i.e., tweeting frequencies), and if possible, use a larger sample (e.g., Gardenhose) to reduce the estimation error.

VI REFERENCES

1. Yong-Yeol Ahn, Seungyeop Han, Haewoon Kwak, Sue Moon, and Hawoong Jeong. 2007. Analysis of topological characteristics of huge online social networking services. In Proceedings of the 16th International Conference on World Wide Web (WWW'07). ACM, New York, NY, 835–844.

2. Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. 2011. Everyone's an influencer: Quantifying influence on Twitter. In Proceedings of the 4th ACM International Conference on Web Search and Data Mining (WSDM'11). ACM, New York, NY, 65–74.

3. Fabr´ıcio Benevenuto, Tiago Rodrigues, Meeyoung Cha, and Virg´ılio Almeida. 2009. Characterizing user behavior in online social networks. In Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference (IMC'09). ACM, New York, NY, 49–62.

4. Shea Bennett. 2012. Twitter Now Seeing 400 Million Tweets per Day, Increased Mobile Ad Revenue, Says CEO@ONLINE. Retrieved from http://www.mediabistro.com/alltwitter/twitter-400-milliontweets_b23744.

5. Johan Bollen, Huina Mao, and Xiao-Jun Zeng. 2011. Twitter mood predicts the stock market. Journal of Computer Science 2, 1, 1–8.

6. S. Borgatti, K. Carley, and D. Krackhardt. 2006. On the robustness of centrality measures under conditions of imperfect data. Social Networks 28, 2 (May 2006), 124–136.