How Much Data Do You Need? Twitter Decahose Data Analysis

Quanzhi Li, Sameena Shah, Merine Thomas, Kajsa Anderson, Xiaomo Liu, Armineh Nourbakhsh, Rui Fang

Research and Development
Thomson Reuters
3 Times Square, NYC, NY 10036

{quanzhi.li, sameena.shah, merine.thomas, kajsa.anderson, xiaomo.liu, armineh.nourbakhsh, rui.fang}@thomsonreuters.com

Abstract. Twitter generates between 500-700 million tweets a day. It is expensive, unnecessary and almost impossible to process the entire tweet data set for any application. Twitter's 1% Streaming API and Search API have their own limitations. In this paper, we present our findings on an alternative source, the 10% Decahose, to help researchers and businesses decide how much tweet data they need. This paper reports on the following analysis for the Decahose data: entity and metadata distribution; entity coverage and novelty evolution from 1% to 10% of Decahose; the amount of information change from 1% to 10%, as measured by recall of test tweets; and statistical comparison between Twitter's 1% streaming data and the Decahose data.

Keywords: twitter decahose · twitter streaming API · tweet metadata · social data analysis · twitter social media

1 Introduction

Twitter makes its data available through several mechanisms. One is through the 1% real-time Streaming API, which is 1% of Twitter data. This data is available at no cost. The Streaming API has limitations; in addition to the volume limit, it also has query limits if its filtering function is used. Another approach for accessing tweet data is through the Search API. The Search API's results are not real time, and it has limitations on the number of queries per user, per application, and per time (15 minute windows) [11].

Neither the Search API nor the Streaming API is designed for enterprise access, which usually needs high coverage of the data. Enterprise access options are the Decahose and Halfhose, which are 10% and 50% randomly sampled data streams respectively [2]. Researchers are very interested in the Decahose, but because of its non-trivial costs, access is usually elusive. It has already been applied to verticals [9].

In this paper, we report on statistical properties of the Decahose. We hope the analysis will provide helpful information for both the research community and businesses.

In this paper, we analyze: (1) Entities and metadata distribution, available through the Decahose, such as hashtags, Urls and tweet topics. (2) Entity coverage and novelty evolution as data sample size increases from 1% to 10%. (3) How the amount of information changes from 1% to 10% sample size. We measure this by computing the recall for a set of test tweets, based on each tweet's textual content. (4) We investigate whether there are any differences between Twitter's 1% streaming data and Decahose data. These analyses provide insight into the characteristics of the Decahose, and help users understand how much Twitter data they may need. Our analyses also show how significant and representative the different-sized samples are.

2 Related Work

In [7], the authors used Twitter's Sample API to test if its 1% Streaming API is biased. They took one hashtag and observed its trend line over one day in Sample API and streaming API. They claimed that there was a small bias because they observed two spikes and the spikes were not identical between the two APIs.

In [6], the authors questioned whether data obtained through Twitter's sampled Streaming API is a sufficient representation of activity on Twitter as a whole. They compared Firehose to the Streaming API, in terms of top n hashtags and geo-tagged tweets. They found that the Streaming API data estimates the top n hashtags for a large n well, but is often misleading when n is small. For geo-tagged tweets, they found that the Streaming API almost returns the complete set of the geo-tagged tweets despite covering a sub-sample of the Firehose.

Although the previous two studies [6, 7] are related to Twitter data, they focus on comparing streaming data to Firehose or sam ple data. To the best of our knowledge, our study is the first one exploring how information changes gradually as data size changes. No previous study has focused on our research questions 2 and 3 introduced in the next section. There are several studies that analyze the entire network sample of Twitter [3, 4, 12], but none of them is about analyzing the effect of data size.

3 Research Questions and Data Set

Following each question, we briefly introduce the methodology for answering it. The detailed research methodology will be described in the result analysis sections.

Q1. What is the metadata distribution across the Decahose data?

An individual tweet has multiple metadata. We study two types of metadata, those derived from the tweet's sourcing and those derived from its content. Tweet's sourcing include whether the author is a verified user, a news agency user, or an influential user. Content-related metadata include tweet's topic, whether it is a retweet, its hashtags, mentions, Urls, and proper nouns (named entities). This question is answered in Section 4.

Q2. As tweet data is increased by 1% at each step going from 1% to 10%, how does entity coverage change? How many new entities are available at each increment?

What does the novelty curve look like, that is, how many new entities are available at each percentage point addition?

We answer these questions in Section 5, where we analyze the changes of the following entities: hashtag, Url, user, verified user, proper noun and mention.

Q3. Given a tweet as a target, what percentage of tweet data do we need in order to find other tweets relevant to the target? At each increment from 1% to 10%, how many relevant tweets are there?

We set this up as a recall experiment, where a set of test tweets are computed to answer this question. Tweet text is used to measure the similarity between tweets for computing recall. The recall was computed separately for news tweets and non-news tweets.

Q4. Are there any differences, in terms of metadata distribution, between Twitter's 1% streaming data and Decahose data?

To answer this question, we compare Twitter's 1% streaming data to the 1% Decahose sampling. The comparison is based on both user related and tweet content related metadata. The result of this comparison is reported in Section 7.

Decahose Dataset. The Decahose is 10% of the whole Twitter data randomly sampled. We obtained one month of Decahose data covering the entire month of Oct. 2015. In total, there are 1.04 billion tweets in the dataset. Among them, there are 280 million English-language tweets. The data used in the experiments are all English tweets; non-English tweets are removed. Some experiments in this study require the Decahose data to be split into 10 parts, each representing 1% of Twitter data. In order to do this, each tweet was randomly assigned a sequence number, from 1 to 10, when it was ingested into our storage.

Test Dataset for Recall Study. For the recall analysis, we need a set of tweets as the standard dataset, on which the recall metrics are computed. These tweets need to be from the same time period as the Decahose data, i.e. October 2015. We have two types of test tweets - tweets from news organization accounts, and tweets from other tweet accounts, such as politicians and sports. They will be called news tweets and non-news tweets, respectively, hereafter. The reason for this distinction is that news tweets are more related to important events, and we want to see if there is any difference between these two types of tweets, in terms of recall. The news tweets were collected from 108 news organization accounts, such as CNN and Reuters. 3,875 news tweets were collected through Twitter's search API. The non-news tweets were collected from 547 non-news accounts, and 2,704 non-news tweets were collected.

4 Tweet Metadata Distribution Analysis

This section tries to answer research question Q1.

4.1 Metadata Generation

- **Tweet topic.** Each tweet is marked with one topic based on its content. We used OpenCalais for tweet topic classification as previous studies [10].
- **Url.** This is the link presented in a tweet. A short link is resolved to its absolute address.

- **Verified user.** This identifies whether a user is verified by Twitter.
- **Influential user.** Following previous studies [1, 5], we use the number of followers to measure a user's influence level. We define two types of influential users by using two thresholds. number of followers greater than 5,000, and 10,000, respectively.
- **News organization user**. These are user accounts that belong to news organizations, such as CNN. 2,040 news accounts are used in this study.
- **Proper noun.** Proper noun (named entity) refers to the name of an organization, person, or other types of entities. A tweet with proper nouns usually conveys more meanings. The TweetNLP package [8] was used to identify proper nouns from tweet text.
- **Other metadata**. Retweet – if it is a retweet; Hashtag – the hashtags in the tweet; Mention - mentions in each tweet; Media - if the tweet contains media content.

4.2 Metadata Distribution

Table 1 presents the metadata distribution. The result is based on one week's data (Oct. 1-7). It presents the results for both the 10% and 1% datasets. The total number of tweets for the one-week dataset is 63,699,142 in the 10% Decahose, and 6,369,072 in the 1% portion. This table shows that, in terms of distribution of metadata, there is no major difference between 1% and 10% data, which is expected because the 1% dataset is already very large from a statistical point of view.

Table 1. Metadata distribution (1% and 10%)

	Metadata	Decahose data 1%	Decahose data 10%
User related	Verified	0.51%	0.51%
	News Organization	0.02%	0.02%
	Influential (followers >5k)	5.94%	5.95%
	Influential (followers >10k)	3.24%	3.25%
Tweet related	Is retweet	37.99%	37.98%
	Has hashtag	19.27%	19.25%
	Has mention	63.89%	63.88%
	Has url	22.65%	22.62%
	Has media	17.47%	17.46%
	Has > 1 proper noun	35.95%	35.92%
	Has > 2 proper nouns	12.36%	12.36%

5 Entity Coverage and Novelty Analysis

This analysis tries to answer the research question Q2. When we talk about coverage in this study, it is based on the 10% Decahose. The dataset used for this experiment is

similar to the one used in the last section: one week of Decahose data ranging from 6.37 million tweets for 1% to 63.7 million for the 10% Decahose.

Table 2. Entity coverage change from 1% to 10% Decahose

Decahose data (%)	Hashtag	Url	User	Verified User	Proper Noun	Mention
1%	20.87%	13.57%	26.86%	35.00%	17.72%	24.25%
2%	33.63%	25.05%	42.57%	52.28%	30.12%	39.04%
3%	44.40%	35.69%	54.30%	63.85%	40.96%	50.60%
4%	54.03%	45.75%	63.79%	72.39%	50.81%	60.29%
5%	62.83%	55.43%	71.80%	79.14%	60.00%	68.68%
6%	71.01%	64.79%	78.74%	84.54%	68.68%	76.15%
7%	78.80%	73.90%	84.89%	89.14%	76.98%	82.91%
8%	86.17%	82.78%	90.40%	93.24%	84.92%	89.07%
9%	93.21%	91.47%	95.41%	96.79%	92.58%	94.72%
10%	100%	100%	100%	100%	100%	100%

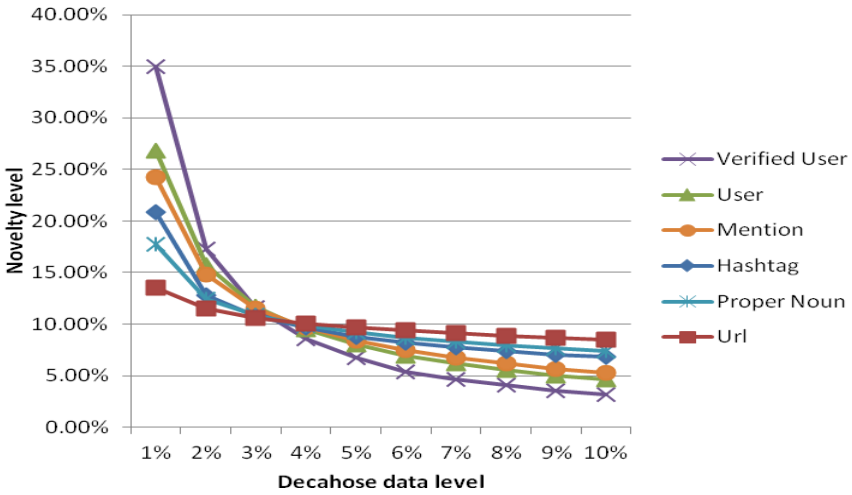

Fig. 1. Entity novelty change rate over data level

Table 2 presents the coverage result. It shows what percentage of an entity-type can be found in each data level. This table shows some interesting findings. For user information, at the 5% level, we can find about 80% of the verified users, and 72% of all users. And it also shows that at just the 2% level, we can find more than half of the

verified users present in the 10% Decahose data. One explanation is that users, especially the verified users, usually author or retweet multiple tweets during this period of time. In terms of Url, only 55% of them are discovered at the 5% data level, which means the emerging rate of new Urls is nearly linear to the data increase level, and Urls are less repeated in different tweets.

Table 3. Recall result using cosine similarity threshold of 0.75

Decahose data (%)	News account tweets		Non-news account tweets	
	Recall	Average number of matches	Recall	Average number of matches
1	47.0%	3.4	23.3%	7.2
2	63.3%	5.1	34.1%	9.9
3	71.9%	6.8	41.2%	12.2
4	77.4%	8.4	46.1%	14.5
5	77.9%	7.4	50.3%	16.6
6	84.6%	11.5	54.0%	18.6
7	86.5%	13.2	56.8%	20.7
8	88.5%	14.7	58.4%	22.9
9	89.9%	16.3	60.7%	24.8
10	91.0%	17.8	62.4%	26.9

Figure 1 shows the result from the novelty point of view: how many new entities emerge at each data level? We can see that the verified user line drops very fast from 1% to 10% level, which means very few new verified users emerge when the sample size reaches a certain level. In contrast, the Url line is almost a straight line, which means it keeps at a high novelty level as the volume of data increases.

6 Tweet Content Recall Analysis

This section tries to answer question Q3. As described before, there are two types of test tweets, news tweets and non news tweets. Each test tweet was compared to all Decahose tweets in this 3-day time range: the day the test tweet was created, 1 day before, and 1 day after. A tweet event usually lasts for a couple of days, and we think a 3-day window is a reasonable time period for finding relevant tweets. Expanding this window may increase the recall value, but the increase is small based on our testing. On average, each test tweet was compared to about 27 million tweets. Cosine similarity is used to measure the similarity between two tweets. Cosine similarity is a popular measure for computing the similarity between two sets of text, and has been used by many previous studies; a value of 1 means the two text segments are the same and 0 means totally different. Before the calculation, some basic pre-processing is applied to the tweet text, such as stopword removal. Different applications may choose different cosine values, usually greater than 0.5, as the threshold for computing recall. The recall result in Table 3 is based on 0.75.

From Table 3 we can see that news tweets have a much higher recall than non-news tweets, since the tweets from news agencies are usually about important events, there are usually more tweets talking about them. In contrast, non-news tweets usually attract less attention. One interesting observation is that although non-news tweets have lower recall, the average number of matches is higher than news tweets. This means that a non-news tweet either has no related tweets, or if it does, it may have a large amount of tweets. For example, an event about Justin Bieber may go viral on Twitter. Another observation is that news tweets' recall is already close to 0.5 at the 1% data level, and when it is at 10%, its recall is 0.91. This means if one is only interested in tweets related to news, the 10% Decahose will provide coverage very close to the 100% Firehose.

Table 4. Metedata comparison of Twitter 1% streaming data with Decahose data

Metadata		Distribution	
		1% Decahose data	1% Twitter streaming data
User related	Verified	0.51%	0.42%
	News Organization	0.024%	0.020%
	Influential (followers >5k)	5.95%	6.40%
	Influential (followers >10k)	3.25%	3.71%
Tweet related	Is retweet	37.98%	39.62%
	Has hashtag	19.24%	18.65%
	Has mention	63.89%	64.55%
	Has url	22.63%	22.35%
	Has media	17.45%	18.37%
	Has > 1 proper noun	35.92%	35.18%
	Has > 2 proper nouns	12.36%	11.82%

7 Comparison of 1% Twitter Streaming Data with Decahose Data

We try to address the research question Q4 in this section. Twitter claims that the 1% streaming data is randomly sampled from the 100% Twitter data in real time, but how exactly that is done is not clear. People may wonder if there is any difference between the 1% streaming data and the Decahose data. We have both the 1% streaming data and the Decahose data from the same period of time, which makes the comparison possible. The 1% Decahose sample used in this study was generated as follows: a tweet from Decahose was randomly assigned to one of ten buckets; after all tweets were processed, one bucket was randomly selected as the 1% sampling of the Decahose.

Table 4 presents the comparison results of the general metadata. One limitation in this comparison study is that the 1% Decahose data has fewer tweets than the 1% streaming data. The size difference between the two data sets is about 15%. The reason is that when Twitter handled the Decahose data to us, some tweets were already deleted either by their authors or by Twitter. Twitter deletes tweets that are considered spam by their off-line spam filter, or violate copyrights or other rules. This might be one main reason that some of the distributions are different between these two data sets. Table 4 shows that the Decahose data has a slightly higher percentage of verified users, while it has slightly lower percentages of influential users. One explanation is that verified users are more careful when they tweet and therefore it is less likely for them to delete their own tweets, and also it is rare that Twitter would delete tweets from verified users. In contrast, users with many followers are more likely to tweet more, and the chance of deleting their tweets is also higher than the ordinary users.

Table 5. Topic comparison of Twitter's 1% streaming data to Decahose data

Tweet Topic	Distribution	
	1% Decahose data	1% Twitter streaming data
Business/Finance	2.23%	2.18%
Technology/Internet	1.50%	1.48%
Politics	1.08%	0.98%
Sports	11.57%	11.21%
Entertainment	9.55%	9.66%
Health/Medical	1.55%	1.56%
Crisis/War/Disaster	1.61%	1.58%
Weather	0.47%	0.46%
Law/Crime	1.01%	0.98%
Life/Society	66.94%	67.28%
Other	2.48%	2.63%

In this analysis, in addition to the metadata used in previous sections, the topic of a tweet is also identified by the topic classifier described before. Table 5 shows the topic distribution for both data sets. The result tells us that the streaming data and Decahose have basically the same distributions. Because the majority of tweets are talking about people's daily lives, a large portion of tweets are classified as Life/Society.

8 Conclusion

In this paper, we analyze Twitter's Decahose dataset and report on the following analyses on the Decahose: the distribution of a rich set of metadata, how the volume of entities evolves when the Decahose data changes from 1% to 10%, the amount of information change at different data levels, and the potential difference between Twit-

ter's 1% streaming data and the Decahose. We hope the statistics and findings will provide insight and help interested parties decide the amount of Twitter data needed for their applications.

References

1. Castillo, C.; Mendoza, M.; and Poblete, B. 2011. Information credibility on twitter. In Proceedings of WWW 2011, 675–684.
2. Gnip. 2015. An overview of twitter's streaming API.
3. Java, A.; Song, X.; Finin, T.; and Tseng, B. 2007. Why we twitter: Understanding microblogging usage and communities. In Proceedings of 9th WebKDD, New York, NY
4. Kwak, H.; Lee, C.; Park, H.; and Moon, S. 2010. What is twitter, a social network or a news media? In Proceedings of WWW 2010.
5. Liao, Q., and Shi, L. 2013. She gets a sports car from our donation: rumor transmission in a chinese microblogging community. In Proceedings of CSCW 2013.
6. Morstatter, F.; Pfeffer, J.; Liu, H.; and Carley, K. 2013. Is the sample good enough? Comparing data from twitter's streaming API with twitter's firehose. In Proceedings of ICWMS 2013.
7. Morstatter, F.; Pfeffer, J.; and Liu, H. 2014. When is it biased? assessing the representativeness of twitter's streaming API. In Proceedings of WWW 2014 Companion.
8. Owoputi, O.; O'Connor, B.; Dyer, C.; Gimpel, K.; Schneider, n.; and Smith, N. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In Proceedings of NAACL 2013.
9. Pal, A., and Counts, S. 2011. Identifying topical authorities in microblogs. In Proceedings of WSDM 2011
10. Quercia, D.; Askham, H.; and Crowcroft, J. 2012. Tweetlda: Supervised topic classification and link prediction in twitter. In the 4th ACM Web Science Conference.
11. Twitter. 2015. Twitter api rate limits.
12. Wu, S.; Hofman, J. M.; Mason, W. A.; and Watts, D. J. 2011. Who says what to whom on twitter. In Proceedings of WWW '2011.