Journal of Multidisciplinary Engineering Science and Technology (JMEST)
ISSN: 3159-0040
Vol. 2 Issue 5, May - 2015

Computation Analysis of Social Networks: Number of Kazakhstani Users in Twitter

Yerezhepov Nursultan
Faculty of Information Technology
Kazakh-British Technical University
Almaty, Kazakhstan
nursultan.yerezhepov@gmail.com

Akhmetzhanov Askhat
Faculty of Information Technology
Kazakh-British Technical University
Almaty, Kazakhstan
a.axmetzhanov@gmail.com

***Abstract*— This paper describes how information taken from social networks like Twitter can be used to identify is user is from Kazakhstan or not. It describes step by step our algorithm for identifying Kazakhstani users. Crawled most popular twitter accounts in Kazakhstan and tried to analyze and find Kazakhstani user in their followers. Do this recursively for each user from Kazakhstan.**

***Keywords*— twitter, social networks**

I. INTRODUCTION (*Heading 1*)

Twitter, a microblogging service less than nine years old, commands more than 500 million users as of December 2014 and is growing fast. Twitter users tweet about any topic within the 140-character limit and follow others to receive their tweets. The goal of this paper is to study the number of Kazakhstani users in Twitter according to tweets, retweets and personal information.

The term 'social network' has become a prominent part of modern day discourse, and in recent years there has been rapid growth in the field of social network studies. [1] Yet a world in which individuals are connected to one another in multifarious ways—spanning time, place, institutional affiliation, and other social boundaries—is not just a modern phenomenon. In the early modern period, neighborhoods, villages, cities and continents were crisscrossed with relationships and ties of obligation, through which passed friendship, as well as animosity; money, ideas, information, material goods, and more. The concepts and methodologies of social network analysis, together with new digital technologies, provide the tools to uncover the nature of these communities in the past. [2]

II. CRAWLER

Twitter has two own API that are comfortable to use: REST APIs and Streaming APIs.

The REST APIs provides programmatic access to read and write Twitter data. Author a new Tweet, read author profile and follower data, and more. The REST API identifies Twitter applications and users using OAuth; responses are available in JSON. [4]

The Streaming APIs give developers low latency access to Twitter's 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. [4]

A. Differences between Streaming and REST

Connecting to the streaming API requires keeping a persistent HTTP connection open. In many cases this involves thinking about your application differently than if you were interacting with the REST API. For an example, consider a web application which accepts user requests, makes one or more requests to Twitter's API, then formats and prints the result to the user, as a response to the user's initial request:

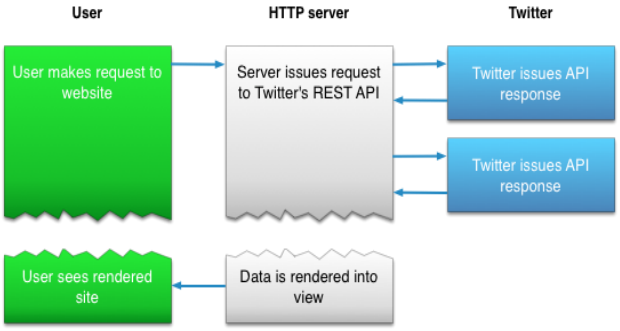

Fig. 1. *Process1.*

An app which connects to the Streaming APIs will not be able to establish a connection in response to a user request, as shown in the above example. Instead, the code for maintaining the Streaming connection is typically run in a process separate from the process which handles HTTP requests:

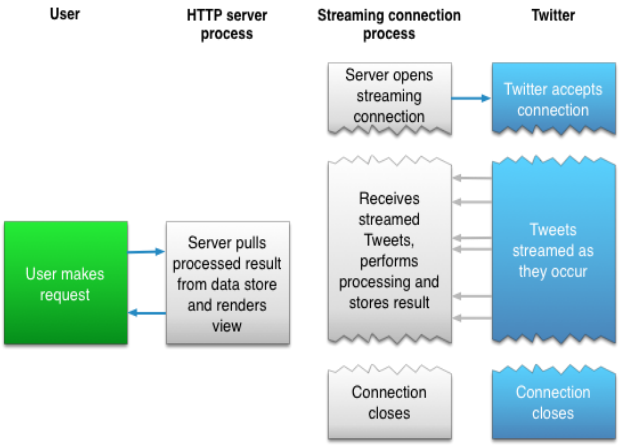

Fig. 2. *Process2.*

The streaming process gets the input Tweets and performs any parsing, filtering, and/or aggregation needed before storing the result to a data store. The HTTP handling process queries the data store for results in response to user requests. While this model is more complex than the first example, the benefits from having a realtime stream of Tweet data make the integration worthwhile for many types of apps. [3]

B. Own Crawler

So, after analyzing the Twitter APIs we decided to write own crawler. One of the most reasons was the rate limiting: only 180 requests per 15 minutes. Where using our crawler we can make about 45000 requests per 15 minutes.

III. ALGORITHM

The algorithm was defined as following:

We have a queue where we have stored our checked users. As the first users we take our famous celebrities such as Bayan Yessentayeva, Alisher Yelikbaev and other. Bayan had for about 221,000 followers. So, on each iteration we take first user from a queue and run through all his followers. [5] Analyzing his followers for following criteria:

•Location if it is pointed. In location field we tried to find matches with cities of Kazakhstan and Kazakhstan in Russian, English languages and in transliteration of English letters.

•Site if it is pointed. In site field we tried to find matches with the domain ".kz".

•Hashtags. In hashtags field we tried to find matches with the set of keywords of words which are applies to Kazakhstan. For example: "yvision, zakon, kazakh".

•Tweets. In tweets field we tried also to find matches with the set of keywords of words which are applies to Kazakhstan.

•Retweets. In retweets field we tried also to find matches with the set of keywords of words which are applies to Kazakhstan.

Finally analyzing each followers for above criteria we give 100 points for location and site for concurrence. And for hashtags, tweets and retweets we give 3 points to each concurrence.

After this if points scored by followers exceed 20 we add him to our queue. We repeat these steps until queue is not empty.

IV. CORRECTNESS

To check our algorithm we took randomly 1000 users and divided them to 10 groups where in each group were 100 users. So we run algorithm for 10 times and get following results:

TABLE I. ITERATIONS

Iteration	Number of users from Kazakhstan
1	81
2	56
3	52
4	63
5	72
6	58
7	64
8	87
9	52
10	73

The average of correctness of our algorithm was 65.8%.

V.Results

Using our crawler we checked following number of users, tweets, URLs, hashtags and retweets covering a 7 days period from May 3 2015 to May 10 2015.

TABLE II. DATASET STATISTICS

Dataset Statistics	
Number of users	348030
Number of tweets	9716175
Number of URLs	3238725
Number of Hashtags	1068452
Number of retweets	1646181

Using our algorithm we found 18320 users from Kazakhstan.

VI.CONCLUSION

We have crawled the Twittersphere and obtained 348030 user profiles and 9716175 tweets. And using our algorithm which correct for 65.8% found 18320 users from Kazakhstan. Using these numbers we can see that Twitter is very popular in Kazakhstan. Our work is the first step towards exploring the great potentials of this platform in Kazakhstan.

REFERENCES

[1] Jeff Huang, Katherine M. Thornton, Efthimis N. Efthimiadis , Conversational tagging in twitter (2010)

[2] Strapparava, C., Mihalcea, R.: Learning to identify emotions in text. In: ACM Symposium on Applied Computing, pp 1556-1560, Fortaleza, Brazil (2008)

[3] Esuli, A., Sebastiani, F.: SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. In: 5th International Conference on Language Resources and Evaluation (LREC 2006), pp. 417-422, Genoa, Italy (2006)

[4] https://dev.twitter.com/overview/documentation

[5] Larry Hoyle, Institute for Policy and Social Research, University of Kansas, Implementing Stack and Queue Data Structures with SAS® Hash Objects, SAS Global Forum 2009