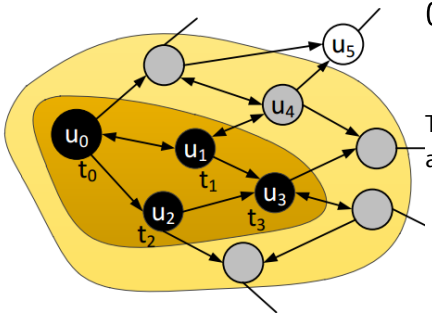
Home

TWITTER CASCADE DATASET



OVERVIEW

Submit Request

This dataset comprises a set of information cascades generated by Singapore Twitter users. Here a cascade is defined as a set of tweets about the same topic.

This dataset was collected via the Twitter REST and streaming APIs in the following way. Starting from popular seed users (i.e., users having many followers), we crawled their follow, retweet, and user mention links. We then added those followers/followees, retweet sources, and mentioned users who state Singapore in their profile location. With this, we have a total of 184,794 Twitter user accounts. Then tweets are crawled from these users from 1 April to 31 August 2012. In all, we got 32,479,134 tweets.

To identify cascades, we extracted all the URL links and hashtags from the above tweets. And these URL links and hashtags are considered as the identities of cascades. In other words, all the tweets which contain the same URL link (or the same hashtag) represent a cascade. Mathematically, a cascade is represented as a set of user-timestamp pairs. **Figure 1** provides an example, i.e. cascade C = {< u1, t1 >, < u2, t2 >, < u1, t3 >, < u3, t4 >, < u4, t5 >}.

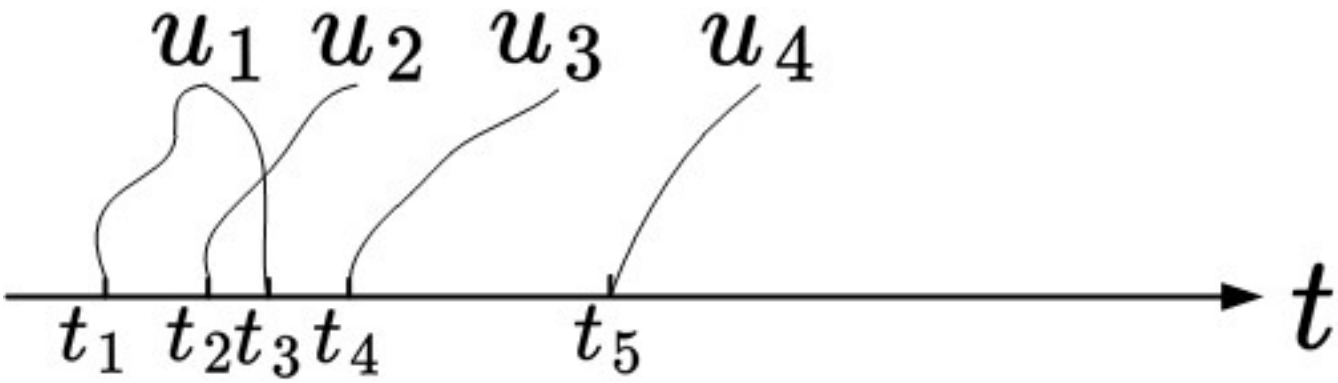


Figure 1. An example of cascade C.

DESCRIPTION

For evaluation, the dataset was split into two parts: four months data for training and the last one month data for testing. **Table 1** summarizes the basic (count) statistics of the dataset. Each line in each file represents a cascade. The first term in each line is a hashtag or URL, the second term is a list of user-timestamp pairs. Due to privacy concerns, all user identities are anonymized.

Table 1. Statistics of the dataset

Dataset	Type	Training	Testing
Twitter Cascades	URL	6,452,732	1,657,145
	hashtag	540,115	190,420

CITATION

Coming soon

Disclaimer: The bot labels in this dataset were obtained from the observation period of 1 January - 30 April 2014. Given the rapidly changing nature of bot behavior, however, these labels may no longer be relevant today. As such, when analyzing the labels, you are advised to use the corresponding tweet posts/contents from the observation period mentioned above.

Last updated on **27 Dec 2017** .

SMU
SINGAPORE MANAGEMENT UNIVERSITY

WHERE TO FIND US

Living Analytics Research Centre

School of Information Systems
Singapore Management University
80 Stamford Road
Singapore 178902

Tel: 65 6808 5227 | larc@smu.edu.sg

LOOKING FOR SOMETHING?

Search SMU...