Tweet URL Analysis

Guoxin Sun, Kehan Lyu, Liyan Li

CS 4624 Multimedia, Hypertext, and Information Access
Dr. Edward Fox

Virginia Tech, Blacksburg VA 24061
2018/05/01
Overview

- Recap
- Issues
- Results
- Future plan
- Acknowledgement
- References
Analyze the characteristics of URLs embedded in tweets.

Figure 1: Architecture of the URL Analysis System [1]
Issues

1. Bad separator for long URL files
   
   http://www.theictm.org/big-diabetes....|paugustine|lekb3vfg74m|thei....

2. Halt caused by using articleDateExtractor library
URL Characteristic Analysis

Percentage of the URL(s) with Keyword per year
Percentage of Tweets with URL(s) per year
Tweets with Different Number of URL(s)
Percentage of Unique URL(s) in Tweet Collections
Percentage of Unique URL(s) with different status code
Percentage of successful retrieved URL(s) per year
Time interval between Tweet Post Date and Webpage Date
Time interval between Tweet Post Date and Wayback Machine Archive Date
Top 10 Domains in Tweets/Retweets
Top 10 URLs in Tweets/Wayback Machine
Percentage of Tweets with URL(s) per year

Statistics:

- 50% of Tweets have URLs on average
- People are more interested in embedding URLs in Tweets from 2013~2015
- The Interest faded away from 2015~2017
Tweets with Different Number of URL(s)

Statistics:

- 90% of Tweets have 1 URL
- 10% of Tweets have 2 URLs
- Less than 1% of Tweets have 3 or more URLs
Percentage of Unique URL(s) with different status code

Statistics:

- 55%~70% of URLs have status code 2xx
- 25%~42% of URLs have status code 4xx
- Around 1% of URLs have other status codes
Percentage of successful retrieved URL(s) per year

Statistics:
URLs in earlier Tweets have higher chance to be archived by Wayback Machine.
Time interval between Tweet Post Date and Webpage Date

Statistics:

Most of Tweet posted on the same day of Webpage posted.
Time interval between Tweet Post Date and Wayback Machine Archive Date

Statistics:
Most of archived URLs were archived within 5 days of Tweets post date
Future Plan

- Finalizing the report
- Analyzing more collections
Future Plan - Possible Improvement

- Utilizing idle machines

<table>
<thead>
<tr>
<th>Node #</th>
<th># of URL converted:</th>
<th>percentage finished</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>187295</td>
<td>77.4100 %</td>
<td>[OK]</td>
</tr>
<tr>
<td>2</td>
<td>186773</td>
<td>77.1900 %</td>
<td>[OK]</td>
</tr>
<tr>
<td>3</td>
<td>224405</td>
<td>92.7400 %</td>
<td>[OK]</td>
</tr>
<tr>
<td>4</td>
<td>241960</td>
<td>100.0000 %</td>
<td>[FINISHED]</td>
</tr>
<tr>
<td>5</td>
<td>241960</td>
<td>100.0000 %</td>
<td>[FINISHED]</td>
</tr>
<tr>
<td>6</td>
<td>241961</td>
<td>100.0000 %</td>
<td>[FINISHED]</td>
</tr>
</tbody>
</table>
Acknowledgement

Liuqing Li

Graduate Research Assistant in DLRL (Digital Library Research Laboratory)

Ph.D. candidate, Department of Computer Science, Virginia Tech

Thanks go to NSF for support by grant IIS-1619028.
References

2. https://github.com/internetarchive/wayback/tree/master/wayback-cdx-server, access date: 10 April 2018
3. https://archive.org/help/wayback_api.php, access date: 10 April 2018
4. http://urlex.org, access date: 10 April 2018
Thank you!

Questions?