International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 03 | Mar-2017 www.irjet.net p-ISSN: 2395-0072

Web of Short URL's

Prerna Yadav [1], Pragati Patil [2], Mrunal Badade [3], Nivedita Mhatre [4]

[1]*Student, Computer Engineering, Bharati Vidyapeeth Institute of Technology, Maharashtra, India*

[2]*Student, Computer Engineering, Bharati Vidyapeeth Institute of Technology, Maharashtra, India*

[3] *Student, Computer Engineering, Bharati Vidyapeeth Institute of Technology, Maharashtra, India*

[4] *Student, Computer Engineering, Bharati Vidyapeeth Institute of Technology, Maharashtra, India*

Abstract: *Short URLs have become universal. Especially popular in social networking services, short URLs have been seen a compelling increase in their usage over the past years of URL shortening services, that are replaced long URLs with shorter ones and redirect it for the shortened URL to the original long URL. As the usage of shortened URLs is convenient, shortened URLs have become common. The study is based on the traces of short URLs from two different perspectives use of a large-scale crawl of URL shortening services use of a crawler for twitter messages. The different focus we will provide a more in depth analysis in the distribution, stability, lifespan and overall use of the short URLs.*

Key Words: Short URLs, Large scale crawls, Twitter, Stability.

1. INTRODUCTION

URL shortening has emerged into one of the leading methods for the easy propagation and allocation of URLs. URL shortening assistance provides their users with a less identical of any grant long URL, and switch consecutive visitant to the expected source. Despite the first well-known URL shortening assistance, namely tiny URL [2], dates back to 2002, currently users can appoint from a large selection of such assistance. The latest popularity of shortening assistance is an outcome of their large-scale usage in Online Social Networks (OSNs) assistance. Short URL accesses serve as small chunk of the "web hits" a site obtains; they are swiftly growing by as enough as 10% per month according to Alexa (Keyword Research, Competitor Analysis, & Website Ranking) [1]. Even with this speedy growth, there is a best of our observation, no other large-scale study in the history that sheds light onto the peculiarity and management guidance of short URLs. We sense that understanding their management has become crucial for certain reasons, along with:

Short URLs are extensively used in functional communities and services such as Twitter, as well as in several Online Social Networks and Instant Messaging (IM) systems. A study of URL shortening assistance will implement judgment into the significance of such association likewise an improved understanding of their peculiarity related to the broader web browsing community. Some URL shortening services, such as bit.ly have grown so much in popularity, which they now account for as much as one percent of the total web population per day.

If this trend continues, URL shortening services will become part of the web's critical infrastructure, posing challenging questions regarding its performance, scalability, and reliability. We believe that answering these questions and defining the proper architectures for URL shortening services without understanding their access patterns is not feasible. To understand the nature and impact of URL shortening services, we perform the first large-scale crawl of URL shortening services and analyze the use of short URLs across different applications.

Our study is depend on fragments of short URLs as examine from two distinct aspect they are collected over an extensive clamber of URL shortening assistances, and collected by crawling Twitter news. The 1st fragment administer insights for an ordinary characterization on the management of short URLs. The 2nd fragment changes our target onto how convinced association operate shortening assistances. The features of our task can be encapsulated as pursue we study the operations that operate short URLs and present that more connections to short URLs appear from IM Schemes, email applicants and OSN media operations, suggesting a "word of mouth" URL dissemination. This dissemination involves that short URLs comes frequently in short-lived media, with serious chattels on their reputation, period, and connection arrangements.

- We present that the short URL snap dissemination can be firmly near by a log-normal arch, validating the order that a limited number of URLs have a very huge number of connections, while the bulk of short URLs has very small connections.
- We present the connection density of short URLs and examine that a huge proportion of short URLs are not short-lived. 50% of short URLs alive for likewise three months. In addition, we examine large business in the connection of short URLs extra short URLs turn into trendy acutely rapid implying a "twitter effect", which may build compelling service deluges and may pose attractive architecture objections for web sites.
- We present that the biggest attractive web sites changes deliberately extra, while having a robust fundamental of web sites which remains stable completely the tested duration. Our analysis also recommend that the web sites which are attractive in the short URL association alter completely from the sites which are attractive among the large-scale web association.
- We examine the achievement indication of the operation of short URLs. We treasure trove that in likewise 90% of the cabinets, the emerging short URL shorten the bulk of bytes required for the URL by 95%. This emerge recommends that URL shortening utilities are intensely adequate in field expanding. On the other hand, we examine that the introduce alternation of URL shortening utilities raises the web page connection times by a supplementary 54% corresponding overhead. This consequence should be taken into examination for the architecture of subsequent URL shortening utilities.

2. METHODOLOGY

We operate two ways to gather short URLs:

i) Crawling, in which we hunt Twitter to treasure trove tweets which encompass URLs and
ii) Brute-Force, in which we drag two URL shortening utilities, that is bit.ly and ow.ly, by organizing hashes of contrasting sizes and exploring which of them earlier occurs. As specified in the earlier section, bit.ly preserves an information page for every generated short URL.

This page implements accurate analysis respecting the bulk of hits a short URL collected, its HTTP assigns and the geographical positions of its visitors. The regular bulk of hits since the formation of the short URL is also document. Information regarding the number of hits from every referrer and country is granted as well. For every bit.ly short URL in our fragments we also gather the followed information pages. Information pages for short URLs generated by certified users still encompass a quotation to the global short URL for this long URL.

For the well-being of integrity, our analysis combines the information hand over by the global hash. Unfortunately, ow.ly does not grant any such information.

2.1. Twitter dragging

Using the 1st approach, we hunt for HTTP URLs that were placed on Twitter. Using the Twitter hunt functionality [4], we gather tweets that encompass HTTP URLs. Twitter establish rate limiting in the number of hunt demands per hour from an obsessed IP address [3]. To regard this policy we check our crawler to one hunt demand every 5 minutes. Every search request retrieves up to 1500 results, going no more than 7 days (max) back in time. During our collection period we handled to collect more than 20 million tweets consist of HTTP URLs. A small portion of the HTTP URLs (13%) collected was not shortened by any URL shortening service. Among the HTTP URLs collected from Twitter, 50% were bit.ly URLs. The second most popular shortening service was tl.gd with 4%, while tiny URL compare to 3.5% and ow.ly become to 1.5% of the overall URLs. Hence, part of our analysis target on bit.ly URLs.

2.2. Brute-Force

Using the second method, we intensely search the available key space for ow.ly and bit.ly hashes. While the Twitter crawling access returns links recently "gossiped" in a social network, this access acts as an alternate source of collection, providing hashes irrespective of their published medium and regency. In the bit.ly case, we searched the entire key space [0-9a-zA-Z] for hashes of up to 3 characters in length presently, the shortening service returns 6-character hashes, indicating a significant exhaustion of shorter combinations. In the case of ow.ly, the system does not disseminate random hashes of the user's long URL, but serially iterates over the available short URL space; thus, if the same long URL is submitted multiple times, it will result in multiple different hashes. Considering this deterministic registration

mechanism, we collected the full set of short URLs created for a period of 9 days. During that time, we monitored the evolution of the key space by creating a new short URL of our own every hour and measuring the distance from the one we had created the previous hour. Using this heuristic, we were able to determine which and how many short URLs were created during that timeframe with a granularity of one hour.

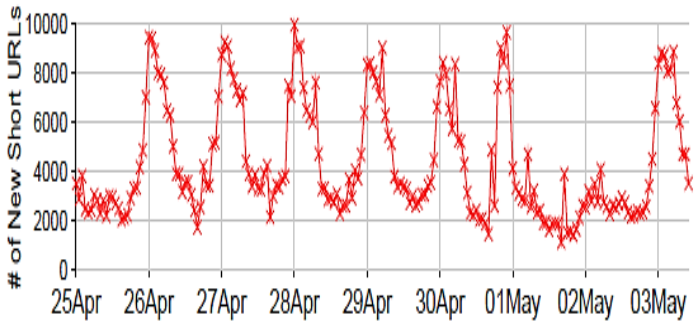

Chart -1: Number of ow.ly short URLs created as a function of time.

Chart 1 shows the number of ow.ly URLs certified as a function of time. We observe a clear diurnal and weekly cycle, with about 70,000 new short URLs created each day.

3. URL'S POINT TO?

Having observed that short UR smartly arise in non-browser types of applications, we now plan at understanding the type of web pages that are very popular through bit.ly links. We casually classified the content of the 100 most accessed domains in the twitter trace. Similarly, we classified the links of the owly trace, which was gained via the Brute-Force method. In this case of ow.ly, the number of approach per short URL is not available so we selected the most popular domain that are more popular in the webs.

trace name	service	number of URLs	accesses
twitter	bit.ly	887,395	101,739,341
twitter2	bit.ly	7,401,026	2,202,442,600
owly	ow.ly	674,239	not available
bitly	bit.ly	171,044	15,096,722

Chart -2: Summary of data collected

twitter		*owly*	
Category	% Sites	Category	% Sites
news (inc. portals)	25	news (inc. portals)	51
info / edu	18	various	17
various	13	info / edu	10
entertainment	10	social networking	5
personal	9	media sharing	5
twitter-related	9	shorten urls	4
commercial	6	commercial	4
media sharing	4	twitter-related	2
social networking	4	sharing articles	1

Chart -3: Most popular types of content.

Chart 3 represents the top most categories for each case. One may notice that news and informative content come first. This observation corroborates the finding of Kwak et al. [9], which suggested that Twitter acts more as an information-relaying network rather than as a social networking site. However, while this study suggests that trending topics are related to news by as much as 85%, the fraction of news related short URLs is significantly lower in our case (25% and 51% for the two traces). A surprising finding is that 4 of the most accessed URLs in the owly trace were shortening services. Such cases reflect short URLs packed inside other short URLs to avoid exposure of the long URLs from tools that unwrap the first level of redirection. Spammers use such techniques to avoid detection, as mentioned by Grier et al. in [16]. Manually examining a number of these URLs confirmed this suspicion with a large number of short URLs pointing to spam content. We plan further investigation of this phenomenon as future work.

4. RELATED WORK

Interest in online social networks and services has been significant over the past years. Several measurement studies have examined basic graph properties such as degree distributions or clustering coefficients [8, 11] or their particular structure [17]. While part of our traces originates from Twitter, our work significantly differs from these

studies as we focus on the use of short URLs and their presence within a social network, rather than network itself.

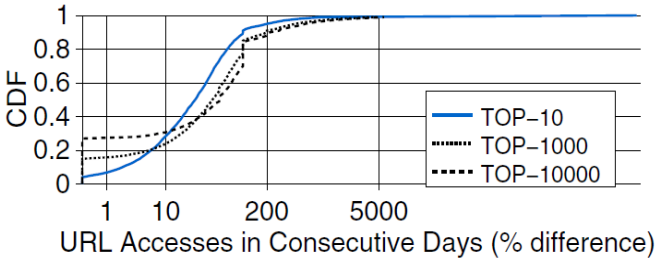

Chart -4: Cumulative Distribution Function for the daily click differences for the TOP-10/1000/10000 short URLs.

Chart 2 is a part of our analysis relates to the evolution of content popularity, information propagation through social links, as well as popularity of objects and applications in social networks. For example, in [6, 7] the authors study how Flickr images evolve and how information propagates through the Flickr social graph. Lerman and Ghosh [10] in investigated the information spread in Twitter and Digg and showed that although Twitter is a less solid network and spreads information slower than Digg, information continues to spread for longer and penetrates further the social graph. In a spirit similar to these studies, we examine how content becomes popular over time.

However, in this work, we focus on how this popularity is reflected by the hit rate of short URLs. Cha et al [5] also deal with content popularity by performing a study of user generated content via crawling the YouTube and Daum sites. The authors observed the presence of the Pareto principle. Our analysis confirms that this is also the case in the popularity of short URLs. Our observations on the dispersion of the hit rates of short URLs are consistent with the well-documented findings on the existence of Zipf's Law and heavy-tailed distributions in WWW. However, our work further highlights that a web site's popularity does not necessarily translate in an equivalent popularity in the "web of short URLs". Information propagation in Twitter has been studied in. The authors have crawled the Twitter network and analyzed the temporal behavioral of trending topics. The authors suggested that Twitter is mostly a news propagation network, with more than 85% of trending topics reflecting headline news. Indeed, this observation is also confirmed by our study. A large fraction of short URLs points to news-related domains; however, the percentage of news related URLs appears lower in our study, 7 out of the top-100 URLs. Information propagation in Twitter has been studied in [18]. The authors have crawled the Twitter network and analyzed the temporal behavioral of trending topics.

5. CONCLUSIONS

We have presented a large-scale study of URL shortening services by exploring traces both from the services themselves and from one of the largest pools of short URLs, namely the Twitter social network. To our knowledge, this paper presents the first extensive characterization study of such services. Specifically, we provided a general characterization on the web of short URLs, presenting their main distribution channels, their user community and its interests, as well as their popularity. Furthermore, we explored their lifetime and access patterns showing an activity period of more than a month with an increased popularity over the first days of their life. We explored the publishers of short URLs, and show a possibility of increased popularity when short URLs are accessed through Twitter.

Additionally, a publisher of such URLs is more likely to be considered a spammer and enjoy decreased popularity when operating at an aggressive rate. Finally, we quantified the performance of URL shortening services, showing a high space gain in terms of bytes used, but also increased overhead in the web page transfer times when accessed through short URLs. This overhead increases web page access time by more than 54% in 50% of the cases, implying that alternative shortening archive.

6. REFERENCES

[1] Alexa Traffic Stats. http://www.alexa.com/ siteinfo/bit.ly#trafficstats.

[2] TinyURL.com. http://tinyurl.com/.

[3] Twitter Rate Limit. http://apiwiki.twitter.com/Rate-limiting.

[4] Twitter Search. http://search.twitter.com/.

[5] M. Cha, H. Kwak, P. R. P., Y.-Y. Ahn, and S. Moon. I Tube,

You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System. In ACM IMC '07, San Diego, CA, USA, pages 1–14, 2007.

[6] M. Cha, A. Mislove, B. Adams, and K. Gummadi. Characterizing Social Cascades in Flickr. In ACM SIGCOMM Workshop on OSNs, 2008.

[7] M. Cha, A. Mislove, and K. P. Gummadi. A Measurement-driven Analysis of Information Propagation in the Flickr Social Network. In Proc. of the 18 Intl. World Wide Web Conference (WWW), 2009.

[8] H. Chun, H. Kwak, Y. Eom, Y. Ahn, S. Moon, and H. Jeong. Comparison of online social relations in volume vs interaction: a case study of cyworld. In IMC '08: Proc. of the ACM SIGCOMM conference on Internet measurement.

[9] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In WWW '10: Proceedings of the 19th international conference on World wide web, pages 591–600, New York, NY, USA, 2010. ACM.

[10] K. Lerman and R. Ghosh. Information contagion: n empirical study of the spread of news on digg and twitter social networks. In Proceedings of the 3th AAAI Conference on Weblogs and Social Media (ICWSM'10), pages 90–97, 2010.

[11] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and Analysis of Online Social Networks. In Proc of the 5th ACM/USENIX Internet Measurement Conference (IMC'07), 2007.

7. BIOGRAPHIES

Prerna Yadav, Student of Bharati VidyaPeeth college of Engineering, completing last year of my diploma in Computer Technology department.

Pragati Patil, Student of Bharati VidyaPeeth college of Engineering, completing last year of my diploma in Computer Technology department.

Mrunal Badade, Student of Bharati VidyaPeeth college of Engineering, completing last year of my diploma in Computer Technology department.

Nivedita Mhatre, Student of Bharati VidyaPeeth college of Engineering, completing last year of my diploma in Computer Technology department.