The Utility of Tweeted URLs for Web Search

Vasileios Kandylas
Yahoo! Inc.
Sunnyvale, CA
kandylas@yahoo-inc.com

Ali Dasdan*
Ebay Inc.
San Jose, CA
ali_dasdan@yahoo.com

ABSTRACT

Microblogging as introduced by Twitter is becoming a source of tracking real-time news. Although identifying the highest quality or most useful posts or tweets from Twitter for breaking news is still an open problem, major web search engines seem convinced of the value of such posts and have already started allocating part of their search results pages to them. In this paper, we study a different aspect of the problem for a search engine: instead of the value of the posts, we study the value of the (shortened) URLs referenced in these posts. Our results indicate that unlike frequently bookmarked URLs, which are generally of high quality, frequently tweeted URLs tend to fall in two opposite categories: they are either high in quality, or they are spam. Identifying the quality category of a URL is not trivial, but the combination of characteristics can reveal some trends.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]: Search Process

General Terms: Experimentation, Human Factors, Measurement

Keywords: content quality, microblogging, shortened URLs, Twitter

1. INTRODUCTION

Since its appearance on the web, microblogging has become an increasingly popular form of blogging. One reason for its popularity is that the short post length requirement demands little time commitment for the users. But microblogging is not only used like regular blogs. Many new uses have come up that were not originally intended [2]. For example, microblogging sites like Twitter are being used to recommend popular articles in real-time [5], to track breaking news stories [4], for work-related communication [6], or for brand marketing [1].

Regardless of the type of use, one common element is the frequent link exchange that occurs through posts and the shortening of the posted links to conform to the maximum allowed post length. In this paper we look at the quality of these URLs and their value for a web search engine. We focus on Twitter posts and specifically the bitly URLs within them and examine their properties: their frequency of appearance, the length of time that they show up in posts, their presence in the index of a state-of-the-art search engine, their quality and spam characterization, their click to view ratio in search results and the correlations between these properties.

Unlike social bookmarking sites that have been shown to provide URLs of high value for search engines [3], we find that shortened URLs from Twitter are a mixed bag: some are of high quality and some are spam. On average, the quality is better than a that of a random set however. Using tweet counts and the number of days that a URL was posted is not enough to fully separate good from bad URLs, but it allows to identify some bad URLs. For example, the most frequently tweeted URLs are of low quality. More work is needed to identify additional features (for example followers count) that can help evaluate the quality with more accuracy.

2. METHODOLOGY

Dataset. We used the Twitter firehose to extract bitly URLs that appeared in every tweet for the month of October 2009. The reason we only looked at bitly URLs from Twitter is two-fold: (1) the majority of URLs that appear in Twitter are shortened URLs and of those most use bitly, which is the default shortening service (about 50% of the short urls), and (2) bitly provides a public API to expand the short URLs to their long form. We ended up with 35 million unique URLs and a total of 60 million URLs. About 29 million URLs were tweeted only once. URLs that were tweeted 5 times or less were removed from our dataset, which left us with 780,000 unique URLs. We expanded these to their long form using the bitly API and computed their *count* (frequency of appearance) and *lifetime* (the number of days they appeared in posts).

Quality scores. We used classifiers, similar to those used by search engines, to classify the expanded links. The classifiers are sophisticated machine learning algorithms that compute quality, spam and adult scores for a web page. They employ many carefully selected features derived from the link structure, the page content and the host. Thresholds on the generated scores are used to classify the pages as good/bad, spam/ham and adult/safe.

Clicks and views. To evaluate user satisfaction we compared the bitly URLs with a set of randomly sampled URLs. Both samples were of the same size and were presented to users as search results. From the user query logs we measured how URLs are distributed, first, according to the total number of clicks they generate and second, by their click-

*Work done while at Yahoo! Inc.

WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

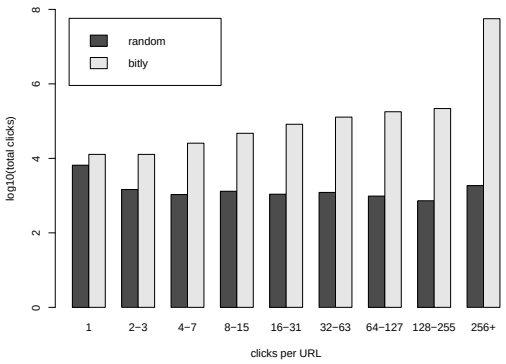

Figure 1: Total clicks per URL of random and bitly URLs, shown as organic search results.

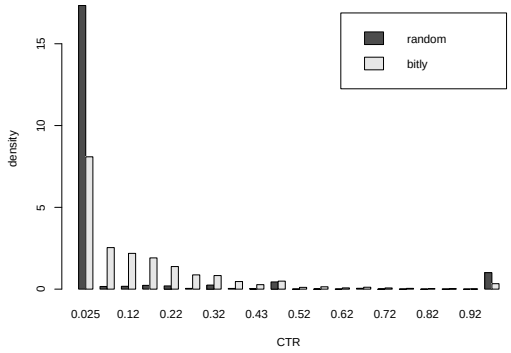

Figure 2: Clicks per view (CTR) of random and bitly URLs, shown as organic search results.

through ratio (CTR). The CTR of a URL is defined as the number of clicks divided by the number of views or impressions it received.

3. RESULTS

Discoverability. Three weeks after the last URLs were extracted from Twitter, we checked the search engine index and found that only about 10% of the unique, expanded URLs appeared there as organic search results (without demarcation as URLs from tweets). This shows that the majority of the URLs are new for the search engine.

Basic statistics. The tweet count distribution followed the power law. Human evaluation of the top 10 URLs showed that they were simple online web games, adult or spam. URLs with high PageRank, such as the CNN home page link, were also tweeted, but not frequently. For almost half of the URLs the lifetime was one day, while about 9% had lifetimes greater than 25 days. Compared to a random set, the bitly URLs had higher average quality score and lower adult score. Their average spam score was about the same, but concentrated in the middle with fewer URLs receiving extremely low or high scores.

Quality vs. frequency and lifetime. Looking at the correlations between the measured data, we found that the URLs that are most frequently tweeted tend to be of low quality. However, URLs that were tweeted few times are not necessarily of high quality either. The URLs with the most tweets have moderate spam scores, but below the threshold to be considered spam. URLs with few tweets cover the whole range of spam scores. Looking at the lifetime correlations, the quality score distribution is not affected by the URL lifetime. Non-spam URLs have longer lifetime than spam and they are tweeted more on average than spam URLs. URLs with a lifetime of 30 days are tweeted the most, but the relationship is not proportional. For example, URLs with a lifetime of 29 days are tweeted about the same as those in the 1 to 28 days range.

Clickability. Compared to the randomly selected set, the bitly URLs generate orders of magnitude more total clicks (Fig. 1). They also generate more clicks per view (Fig. 2). However, the bitly URLs tend to improve the lower CTR buckets more than the higher ones. In fact, the highest CTR bucket for bitly contains fewer URLs than the random set. The reason is that the URLs with the highest CTR in the random set are very popular web pages, such as Facebook, but in the Twitter set these URLs are not tweeted frequently.

4. CONCLUSIONS

We found that extracting bitly URLs from Twitter can be useful for a web search engine. The average URL quality is higher than that of a randomly selected set. One must be careful however, because the quality has a two-mode distribution. One mode is centered on high quality scores and another is around lower values and corresponds to spam URLs. The combined use of URL tweet count and lifetime provides insights into some of the URLs, but is not enough to filter out a significant portion of bad or spam URLs. Future work is needed to discover additional features that would permit a more efficient filtering of bad/spam URLs and a greater positive effect on the search engine.

5. REFERENCES

[1] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Micro-blogging as online word of mouth branding. *Proc. Int. Conf. on Human Factors in Computing Systems*, pp. 3859–3864. ACM, 2009.

[2] A. Java, X. Song, T. Finin, and B. Tseng. Why we Twitter: understanding microblogging usage and communities. *Proc. WebKDD and SNA-KDD Workshop on Web Mining and Social Network Analysis*, pp. 56–65. ACM, 2007.

[3] S. Kolay and A. Dasdan. The value of socially tagged URLs for a search engine. *Proc. Int. Conf. on WWW*, pp. 1203–1204. ACM, 2009.

[4] O. Phelan, K. McCarthy, and B. Smyth. Using Twitter to recommend real-time topical news. *Proc. Conf. on Recommender Systems*, pp. 385–388. ACM, 2009.

[5] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. Twitterstand: news in tweets. *Proc. Int. Conf. on Advances in Geographic Information Systems*, pp. 42–51. ACM, 2009.

[6] D. Zhao and M. B. Rosson. How and why people Twitter: the role that micro-blogging plays in informal communication at work. *Proc. Int. Conf. on Supporting Group Work*, pp. 243–252. ACM, 2009.