The Value of Socially Tagged URLs for a Search Engine

Santanu Kolay
Yahoo! Inc.
Sunnyvale, CA 94089, USA
santanuk@yahoo-inc.com

Ali Dasdan
Yahoo! Inc.
Sunnyvale, CA 94089, USA
dasdan@yahoo-inc.com

ABSTRACT

Social bookmarking has emerged as a growing source of human generated content on the web. In essence, bookmarking involves URLs and tags on them. In this paper, we perform a large scale study of the usefulness of bookmarked URLs from the top social bookmarking site Delicious. Instead of focusing on the dimension of tags, which has been covered in the previous work, we explore social bookmarking from the dimension of URLs. More specifically, we investigate the Delicious URLs and their content to quantify their value to a search engine. For their value in leading to good content, we show that the Delicious URLs have higher quality content and more external outlinks. For their value in satisfying users, we show that the Delicious URLs have more clicked URLs as well as get more clicks. We suggest that based on their value, the Delicious URLs should be used as another source of seed URLs for crawlers.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search Process

General Terms

Experimentation, Human Factors, Measurement

Keywords

Content quality, Delicious, social bookmarking

1. INTRODUCTION

In recent years, social bookmarking sites have seen tremendous growth. For example, popular sites like Delicious is now reporting to have more than 5 million users and 180 million unique URLs tagged. These collaborative bookmarking sites allow users to save and tag URLs with related keywords and to share them with the public or their friends. With the growing popularity of social bookmarking, these tagged URLs are becoming a growing source of human generated content.

A few studies have tried to analyze various social bookmarking systems from the "tags angle": the quality of tags [3], the impact of social bookmarking on information organization [2], tag popularity as a ranking feature [5], taxonomies of tags [4], and growth patterns [1]. In contrast, we approach social bookmarking from the "URLs angle". Specifically, we try to quantify the value of tagged URLs for a major search engine.

In our study, we focus on the most popular social bookmarking site, Delicious. We try to answer these questions: Can the bookmarked URLs lead to new or quality content on the Web? Can the bookmarked URLs satisfy users when presented in search results? We answer both questions in the affirmative. In particular, we show that bookmarked URLs serve as good seed URLs for crawling, lead to the discovery of many new URLs, have higher quality content, and finally receive relatively more clicks from users. Since we had privileged access into the Delicious data, we were able to perform a full scale study of bookmarked URLs, far larger than done in the previous work.

WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

2. METHODOLOGY

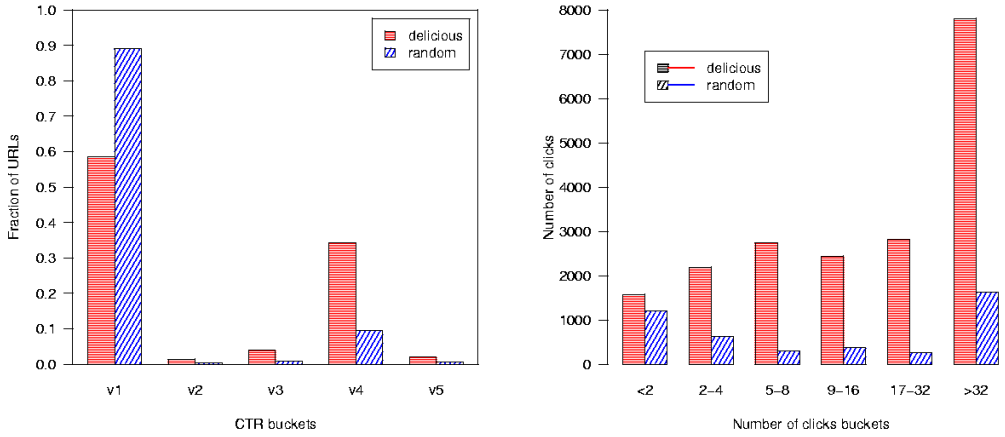
Crawling. Search engines mainly acquire content from crawling the web. Crawlers start crawling from seed URLs and follow links (outlinks) from the crawled pages. Among important areas of focus during crawling are the discovery of new content and the discovery of high quality content. The easiest way to utilize the Delicious URLs is to feed them into the crawler as seed URLs.

Quality and spam scores. Search engines use sophisticated machine learning techniques with many features to compute quality and spam scores for a given page. Features include various link, site, and page features. Each score is mapped to a finite range, say, zero to 100. For each range, we defined a threshold to classify the good and bad content.

Dataset. We had an initial sample of 100 million unique URLs from Delicious. We then filtered out the URLs that have been bookmarked only once. Using a crawler, we fetched the pages of these URLs from the Web. We also computed their quality and spam scores. Finally, we joined the URLs with 2 days of user query logs to compute their click frequency.

3. RESULTS AND DISCUSSION

Stability and depth. After one month of crawling, 95% of the URLs in our dataset were crawlable. Around 2% of them were not crawlable due to the robots exclusion protocol and 1% resulted in a 404 error (page not found). These results show that the Delicious URLs are not only crawlable but also stable on the web. Moreover, out of the crawled URLs, only 20% of them were the domain roots. In other words, most of them reach deep into the sites.



(a) Click-through rate (CTR).

(b) Total number of clicks.

Figure 1: Clickability of the Delicious URLs compared to a random sample of URLs of the same size.

Discovery. The external outlink distribution of the Delicious URLs is similar to that of a random set of URLs; their medians were the same at 4 but their means were a little different: the mean for the Delicious URLs was about 60% more than that for the random URLs. We noticed that this was due to the presence of some higher-degree Delicious URLs. When we compared the discoverability using the URLs that were one hop away from the seed URLs, we found out that the average external outdegree for the neighbor URLs of the Delicious URLs was more than 3 times larger that that for the neighbor URLs of the random URLs. These results support the high value of the Delicious URLs for discovering new content.

Content quality. Assuming 100 buckets for the quality score range, we have found out that more than 90% of the Delicious URLs fell in the top bucket and more than 95% of them fell among the top three buckets. In other words, almost all of these URLs were among the highest quality URLs. From the point of view of spam and adult content, we have found that less than 4% of the Delicious URLs had spam score larger than what we considered spam content. Moreover, less than 1% of them were flagged as adult content. These observations support the claim that the Delicious URLs are of high quality.

Clickability. For the clickability analysis, we created one sample from the Delicious URLs and another one from a random set of URLs. Both samples had the same size and were presented to users as search results. We then measured two metrics out of these samples: (1) the histogram of the click-through rate (CTR) and (2) the histogram of the total number of clicks. These histograms are shown in Fig. 1(a) and Fig. 1(b), respectively. Note that the CTR of a URL is the ratio of the number of clicks to the number of views or impressions that the URL receives. For Fig. 1(a), the URLs were distributed to the buckets on the x-axis based on the ctr values (five values, linear increases); and for Fig. 1(b), the distribution to the x-axis buckets were on the basis of the number of clicks the URLs receive.

It is evident from these figures that compared to the random URLs, the Delicious URLs not only receive significantly more clicks but they also get more clicks per view. These findings are even more important given that the number of URLs that get clicked decreases as the number of clicks they receive increases, i.e., as we move towards the right on the x-axis of Fig. 1(b). All in all, these findings indicate the the higher clickability of the Delicious URLs.

4. CONCLUSIONS

Our study shows that the Delicious URLs provide significant value to web search engines from the perspective of content discovery and user search satisfaction. Seeding a crawler with these URLs leads to faster discovery of good quality content. Moreover, when they are served to users in response to queries, the number of them receiving clicks as well as the total number of clicks they receive are significantly high.

5. REFERENCES

[1] U. Farooq, Y. Song, J. M. Carroll, and C. L. Giles. Social bookmarking for scholarly digital libraries. *IEEE Internet Computing*, 11(6):29–35, 2007.

[2] P. Heymann and H. Garcia-Molina. Can tagging organize human knowledge? Technical report, Stanford University, 2008.

[3] P. Heymann, G. Koutrika, and H. Garcia-Molina. Can social bookmarking improve web search? In *Proc. Int. Conf. on Web Search and Web Data Mining (WSDM)*, pp. 195–206. ACM, 2008.

[4] D. Ramage, P. Heymann, C. D. Manning, and H. Garcia-Molina. Clustering the tagged web. In *Proc. Int. Conf. on Web Search and Web Data Mining (WSDM)*. ACM, 2009.

[5] Y. Yanbe, A. Jatowt, S. Nakamura, and K. Tanaka. Can social bookmarking enhance search in the web? In *Proc. Joint Conf. on Digital Libraries (JCDL)*, pp. 107–116. ACM/IEEE, 2007.