User Profiles Based on Revisitation Times

Philipp Pushnyakov, Gleb Gusev
Yandex
16 Leo Tolstoy St., Moscow, 119021 Russia
{pushnyakov, gleb57}@yandex-team.ru

ABSTRACT

Our work is devoted to Web revisitation patterns of individual users. Everybody revisits Web pages, but their reasons for doing so can differ. We analyzed Web interaction logs of millions users to characterize how people revisit Web content. We revealed that each user have its own distribution of revisitation times. This distribution follows Power Law with some exponent, which captures specific user peculiarities.

Categories and Subject Descriptors

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

General Terms

Experimentation, Theory

Keywords

User behavior, User profiles, Revisitation, Power Law

1. INTRODUCTION

Revisiting Web pages is common, but reasons for doing that can be diverse. For example, a person may revisit a travel agency web page every couple of months to check for new world tours. Another example is that a person may revisit a shopping site to check for sales every couple of days. Both provide us with examples of important revisitation pattern, namely, revisitation time elapsing between consecutive visitations of one user to the same page. This pattern is the main object of our study. Previous studies on revisitation have demonstrated that 50 percent to 80 percent of all visits to Web pages are previously visited pages. The current work distinguishes itself from previous studies in a number of significant ways. While [1] is focused on Web pages, [3] is focused on tags, our study is focused on user engagement with web pages of his interest. To the best of our knowledge, this is the largest study of revisitation behavior ever done. Our study involves millions users, while previous works considered no more than million people.

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2577380.

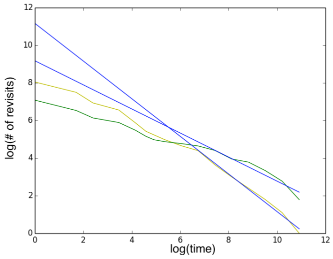


Figure 1: Examples of revisitation curves (green and yellow) with lines approximating their tails.

By widely sampling pages and users, we are able to understand the characteristics of users that are associated with specific types of revisitation behavior.

The main goal of the current paper is to investigate the *user revisitation curve* associated to a given user as follows. First, we introduced sixteen exponential time bins as in [1], namely, $\{t_i\}_{i=1}^{16}$, where $t_1 = 1$ minute and $t_{16} = 55000$ minutes. Formally, revisitation curve is a vector $\{y_i\}_{i=1}^{16}$, where y_i is the count of revisits done by the user in time τ after his previous visit to that page with $\tau \in (t_{i-1}, t_i]$ (we put $t_0 = 0$). We are interested to look at the curve in log-log axes to see if it decreases by the Power Law. Examples of two different curves are given in Fig. 1. As one can see, the ”green” user engages more revisits after a significant time than ”yellow” one.

To the best of our knowledge, we are the first to introduce the notion of user revisitation curve. Our approach is similar to [1] in that we also consider the distribution of revisitation times, but we focus on the complete visitation behavior of a given user instead of following the incoming traffic of a fixed web page. In next sections we approximate the curve by Power Law and conclude that the parameter of the power-law distribution reflects a specific user property. Our analysis can enable search engines to better support revisitation, so we discuss the applications of our study at the end of the paper.

2. DATA

To understand how different people revisit Web pages we analyzed data from the logs of an add-on toolbar for Web browsers distributed by a major search engine [1]. The toolbar provides augmented search features and reports anonymized browsing behavior to a central server. Our analysis makes use of data from a sample of millions users for the 41 days period starting September 1, 2013. Users were identified by an anonymized ID associated with each toolbar. We sorted the users by the total number of their visits to pages recorded

[1] yandex.com

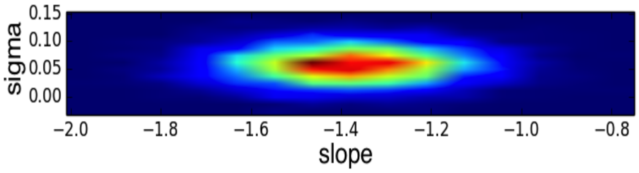


Figure 2: Heat map of (β, σ), where β is slope and σ is standart deviation in bootstrap distribution for that slope σ

in the data. Top 1 percent and bottom 10 percent of users were removed to prevent the data from possible noise.

3. CURVE ANALYSIS

It can be seen that a user revisitation curve is formally a sixteen-dimensional vector $\{y_i\}_{i=1}^{16}$, where each y_i corresponds to the i-th bin. We analysed revisitation curves by considering their tails. Namely, we hypothesized that for every user U, there is a positive number $\beta = \beta(U)$ such that the tail of his revisitation curve follows the power-law distribution $y_i = t_i^{1-\beta(U)}$. The hypothesis emerged from [2], where it was formulated in a different way for the first time. The authors considered the pairs of users and pages while we consider users. We show that the tail of a user revisitation curve in log-log axes can be approximated by a line whose slope equals to $-\beta(U)$ (see Fig. 1).

Formally, the tail of curve $\{y_i\}_{i=1}^{16}$ is vector $\{y_i\}_{i=N}^{16}$ for some N which was founded in the following way: we took the minimal value of N such that the determination coefficient of the best linear approximation of the tail was close to 1. Specifically, we chose $N = 9$ in our experiments.

To make a linear regression for each user‘s revisitation curve, we considered the cumulated curve $\{u_i\}_{i=1}^{16}$, where $u_i = \sum_{j=i}^{16} y_j$ as it was done in [5]. Clearly, if a tail of the cumulated curve is well approximated by the Power Law with an exponent β_{int}, then the initial curve is approximated by the Power Law with the exponent $\beta_{int} - 1$.

We obtained that revisitation asymptotic behavior is best approximated by the Power Law $n(t_i) \sim t_i^{-\beta}$ with $\beta = 1.35 \pm 0.2$ (see Fig. 1). In Fig. 4 it can be seen distribution of users over slopes which is Gaussian distribution with the mean value -1.35 and the standard deviation 0.18. It is easy to see that the more exponent is, more likely the user to revisit a page after a short period of time. Also we calculated the deviation of curves from the fitted lines to examine if the approximation is valid.

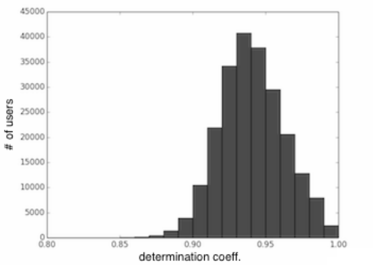
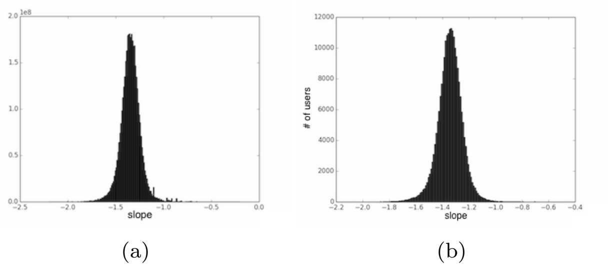


Figure 3: Distribution of determination coefficients for users.



(a)

(b)

Figure 4: Distribution of (a) number of users and (b) revisits done by them over slopes

In Fig. 3 we can see that for most users, determination coefficient is more than 0.9 which means that our approximation is great. To estimate confidence interval for every possible slope, bootstrapping was done. Bootstrap distribution was performed by taking repeated bootstrap replications of re-sampling visitation data, rebuilding curves and re-estimating exponents. It turned out to be a normal distribution whose deviation equals to 0.05. Moreover, distribution of slopes and corresponding deviations (σ) is shown in Fig. 2. It is seen that the majority of users have slopes about -1.35 with deviation approximately 0.05.

4. APPLICATIONS AND CONCLUSIONS

Previous studies of revisitation times were aimed at web sites ([1], [4]), at tags ([3]), but not the users. A novel aspect of our work is that we are aimed at users. We shown that the number of revisits made by user decreases with revisitation time as a Power Law. This result could be used to personalize search results shown to user. As we have seen, the exponent in the power-law distribution of revisitation times of a user follows the Gaussian distribution with a density $\frac{1}{\sqrt{2\pi}\sigma} \exp - \frac{(a-a_0)^2}{2\sigma^2}$, where σ and a_0 were introduced before. So, by knowing the exponent of a user, we can estimate probability of a revisiting after a period of time τ for that user. It can be applied by search engines obviously. The results with high probability should be re-ranked higher, as well as results with lower probability should be re-ranked lower.

5. REFERENCES

[1] E. Adar, J. Teevan, and S. T. Dumais. Large scale analysis of web revisitation patterns. In *CHI*, 2008.

[2] Z. Dezsö, E. Almaas, A. Lukács, B. Rácz, I. Szakadát, , and A.-L. Barabási. Dynamics of information access on the web. In *Phys. Rev. E 73*, 2006.

[3] H. Kawase. Classification of user interest patterns using a virtual folksonomy. 2011.

[4] L. Tauscher and S. Greenberg. Revisitation patterns in world wide web navigation. In *Proceedings of the ACM SIGCHI*, CHI '97, New York, NY, USA, 1997. ACM.

[5] M. Zhukovskiy, D. Vinogradov, Y. Pritykin, L. Ostroumova, E. Grechnikov, G. Gusev, P. Serdyukov, and A. Raigorodskii. Empirical validation of the buckley-osthus model for the web host graph: Degree and edge distributions. CIKM '12, New York, NY, USA, 2012. ACM.