Track A: Research

New Evidences of Power-Law Distributions for Modeling New Patterns of Information Consumption Habits

Manuel Gómez Zotano

RTVE

Jorge J. Gómez Sanz, Juan Pavón Mestras

UCM-GRASIA

It has been known for some time that content requests to web information systems do follow a power-law, in particular, a Zipf-like distribution. Zipf is a power-law parameterized by alpha, a parameter that determines the shape of the cumulative probability function that web objects show in the distribution.

The Zipf model has been very useful to generate workloads to test web information systems and also to setup web content caches. However, so far, Zipf was assumed to have an alpha less than 1. Some authors even claimed that there was no Zipf at all, when the alpha was too low. Literature has pointed out that alpha ought to change, and, in fact, recent studies by the authors have proven that the situation has changed. Authors evaluated using 16 web sites logs to find out that alpha is no longer less than one, but greater, up to 80%. This has a direct impact for engineers in charge of content cache systems and are of significance for massive media sites. Despite these evidences, it is necessary to know more of the reasons for the variance of the alpha parameter.

The talk contributes with new evidences that support the existence of Zipf distribution at the server side and analyses four factors affecting it: the impact of the domain of the information server and its technology, the impact of the considered time window, the changes in accessed content types, and the impact of non-human traffic generated by bots. These aspects have not been addressed before and may help to understand better why the change in the values observed in alpha in modern sites with respect to past studies.

The result has been that alpha is not as stable as one may expect. It varies depending on the considered hour of the day, but also can vary attending to the kind of content. Measurements show an increase of the application content types that favor concentration of popularity, as user access to less pages to do more. The impact of bots and crawlers has been measured too. The traffic they represent is a 0.69\% of the total, their requests also follow a Zipf, though with an alpha far from the total values observed (an alpha greater than one) and closer to older values observed in the literature (an alpha less than one). The hypothesis that would explain this variation is the fact that crawlers do not generate as much traffic per visited page as a human using a browser. Therefore, the behavior resembles the one expected years ago where embedded applications in the web pages were less common. To conclude, similar web frameworks also seem to return similar alphas, what reinforce the hypothesis that the existence of Zipf in web content requests is due to the structure of web pages, rather than the behavior of the users.