User Behavior in Mass Media Website

Manuel Gómez Zotano[a], Jorge Gómez-Sanz[b] and Juan Pavón[b]

[a]Corporación Radio Televisión Española, Alcalde Sainz de Baranda 92, 28007 Madrid, Spain

[b]Universidad Complutense de Madrid, Facultad de Informática, 28040 Madrid, Spain

manuel.gomez@rtve.es, jjgomez@fdi.ucm.es, jpavon@fdi.ucm.es

KEYWORD

Web traffic analysis; Zipf distribution; Web user patterns; Mass Media Website

ABSTRACT

Mass media websites can be worthy to understand user trends in web services. RTVE, the National Broadcaster in Spain is a sample of such kind of service. The analysis of trends points to a shorter user interaction over the last three years, and a more straight access to content. Besides the number of pages consumed in a visit is becoming smaller as well. This article reviews these trends with data obtained from public sources, and analyze the distribution of web pages in the client layer and the corresponding distribution observed in the server layer. These two distributions can be characterized by Zipf-like distributions and α, the degree of disparity in the popularity distribution, is calculated for both. In all cases α is higher to one implying a huge concentration of popularity on a few objects.

1. Introduction

Mass media websites offer a huge volume of information to users in the form of news, media, images and text. The contents are usually consumed in different devices, using different formats and sometimes in a dif-ferent way. For instance, the consumption of content in a *Smart TV* is different than the one in smart phones since a user perceives a TV Set as a place to consume video, while text and images are more demanded in a smart phone.

Website's usage and users' consumption patterns have evolved over the years. Gathering and analyzing this information should be done by considering the different devices, products, etc. However, business restrictions make difficult to collect such kind of data: it is very sensitive and therefore difficult to be shared. In this article, some public sources have been found, but only for a global behavioral analysis and not for a device-detailed study.

This paper presents evidences obtained from a mass media company, *RTVE* (Radio Televisión Española), the public Spanish broadcaster that accounts for around 300 million daily access to multimedia information through web services. This is relevant to analyze the empirical users' behavior while accessing to pages, applications, and diverse multimedia content. Data is gathered from *RTVE* and analyzed to understand the evolution of the web usage patterns. RTVE offers news, video, audio, galleries and polls in the form of websites, 20 mobile applications, *HbbTV* and *Smart TV* applications, and console applications. As it can be seen, *RTVE* content can be consumed in a very diverse set of devices, from smart phones through web applications to applications running in *TV* sets.

The methodology used in this article starts gathering data from public sources. The first one is the audience reports audit by OJD, which are published by the ojdinteractiva website (OJD, 2016). Next, this

©Ediciones Universidad de Salamanca / cc by-nc-nd

data is compared to private data from *RTVE* obtained from two sources: web analytic tools like Adobe Omniture and Apache-like logs from a *CDN* (Content Delivery Network). A similar approach can be found in (Mahanti et al., 2009).

The former is used to measure web pages, that is, pages a user browses to. A web page usually consists of a piece of news, a video page, a distribution page, and other elements. In native applications, pages also refer to the corresponding content, identified by an unique *URL*. In this way, no matter where content is consumed, the same identifier represents it.

CSS files, images, video files, audio files or JavaScript are not considered as web pages since they are auxiliary resources needed for the right page consumption. A web page requested from a browser causes many auxiliary web objects to be requested. In this article, we use the term ”web objects” to identify web pages and auxiliary objects. Data related to web objects that are served from a website is obtained by analyzing the Apache logs, our second source of information. This data is obtained from *CDN*s, and it represents the server layer view. This approach is widely used in other studies (Breslau et al., 1999; Urdaneta et al., 2009).

The analysis of the above data sources brings different perspectives from one website: web pages allow a user centric analysis while web objects give a more technical perspective of the website consumption.

The main conclusion of this article points to a more direct access to contents by the *RTVE* users and to a lower influence of webmasters on users' consumption over the years. This conclusion is supported by the fact that visits (see section 2) are shorter in terms of time and web pages, the huge increase in the number of visits (up to 50% increase in two years) and the wider disparity in the popularity distribution of web pages found from the evidences. It points to a poor inter linked ratio in *RTVE* website and a compulsive content usage as a result of some of the digital marketing techniques (sharing contents in social networks or content optimization for search engines). On average, only 3,86 pages are consumed per visit, 20% smaller than in 2014. Results can be applicable in other mass media websites since these marketing techniques are widely used.

Additionally a second conclusion obtained from the evidences is that Zipf is everywhere: web objects' frequency follows a Zipf-like distribution, as well as the web pages do. Given the relation described above between web pages and web objects, one question arises from the results: web object Zipf-like behavior is perhaps a consequence of the Zipf-like distribution found in web pages. The evi-dences found in this article point to that finding as it can be seen in section 7.

A third evidence found involves the redesign of *RTVE*: the website becomes responsive using *HTTP* redirects (response code 301) based on the user agent in order to adapt the content. When a *URL* is requested, depending on the user agent, a response code 200 happens pointing to the most adequate resource. The effect found from evidences is a smaller α in the Zipf function, due to more *URL*s to compute the same frequencies.

The rest of the paper is structured as follows. Section 2 explains the concepts involved in web analytics and log analytics. Section 3 introduces Zipf and power law distributions. Once Zipf is properly presented, section 4 presents and discusses previous works. Section 5 presents the evidences from the client layer while in section 6 the *CDN* logs are analyzed. Section 7 elaborates more on this knowledge and correlates data from client and server. Finally, section 8 presents the conclusions and discusses about the applicability of this analysis in a wider scope.

©Ediciones Universidad de Salamanca / cc by-nc-nd

2. Client-side vs. server-side

Traditionally web analysis has been made using Apache-like logs (Breslau et al., 1999). Some analysis gath-ers the information at intermediate routers while others from server-side logs. Apache-like logs are a col-lection of files that registers every request reaching the server. Among other information, the *URL*, a timestamp, the request method or the request response is saved for further analysis. The information saved is based on the browser interactions and therefore brings a technical perspective from the us-er/website interaction.

Using web analytics from the client-side can do a more sophisticated analysis. Every time a user accesses to a web page, a message is sent to a server containing information like the *URL* and the timestamp. This data is useful for a content-centric analysis, because data collected is only based on the pages and does not include all the resources that user interactions generate. This method is also called page tagging.

Web analytics tends to measure three metrics: unique visitor/user, visits and page views (Mahanti et al., 2009). The mechanism used for calculating the metrics is based on two identifiers: session id and user id. The first time a user access to a web page, a cookie is created containing an unique id. This is used to identify the user in every access. Additionally a session id is created for that user and stored in the cookie as well. This value is removed if the user does nothing for a period of time of 30 minutes. This action never happens if the user is watching a video, because it is assumed that while a video is consumed a user does not interact with the page.

The two ids are sent to the server in charge of web analytics for every page that a user visits. The request includes the time, the page's *URL* and the cookie. This data is stored for further processing, and using the session id and user id the metrics can be calculated.

Many websites usually use more than one website measurement service. Among others, *RTVE* uses Comscore, Adobe Omniture and Google Analytics depending on business rules. In the Spanish market, a company OJD is in charge of auditing the web analytics data and the audit reports are available for public access (OJD, 2016).

In this article some definitions are used (Adobe, 2016; Mahanti et al., 2009; Calzarossa et al., 2016):

Page View: when a page is requested from a client, a page view is counted. It includes every web page the client browses to.

Unique visitor: a unique visitor refers to a visitor who visits a site for the first time within a specified time period.

Visit: a series of resource requests from a unique visitor that are temporally clustered. After 30 minutes of inactivity, a page view by the same visitor is counted as a new visit. Visits are sometimes referred to as sessions.

Each data collection approach has advantages or disadvantages (Mahanti et al., 2009). Client-side analytics are not able to measure the traffic volume of a website because only web pages are sent to the analytics service, while the server-side approach stores all the web objects served. Client-side services track visitors and their visits using cookies allowing an easy identification of returning visitors, while in server-side an unique visitor is identified by the IP address. Server-side analysis can be affected by web caching and it is hard to distinguish between robots and human visitors. Finally, information like the operating system, browser capabilities, traffic sources, etc., can only be obtained from client-side.

©Ediciones Universidad de Salamanca / cc by-nc-nd

A deeper essay comparing techniques for both client and server side analysis is found in (Calzarossa et al., 2016).

3. Introduction to Zipf and power-law distributions

Zipf is a power-law distribution, which was first used in linguistics, for analyzing words occurrence. Zipf is a discrete distribution defined in the rank-frequency domain, which states that when items are ranked (R) in descending order of their popularity, then the frequency (F) of the item is inversely propor-tional to the rank item (Mahanti et al., 2013). The Zipf formula can be written as:

$$f(r) \sim k * r^{-\alpha} \quad (1)$$

where k and α are constants and α is close to 1. By applying logarithm in both sides, the formula is trans-formed to:

$$\log\big(f(r)\big) \sim -\alpha * \log(r) + \log(k) \quad (2)$$

As it can be seen, the formula 2 corresponds to a straight-line shape with slope −α. Therefore a power-law function, once transformed to log-log, is plot as a line. α is used to measure the concentration of frequency. The higher it is, the more concentration of frequency is found in the first rank items. A smaller α means a more uniform frequency distribution.

Depending on the value of α, a Zipf distribution is called strict when α is almost equals to 1, and it is called Zipf-like when α is different to one.

In the web domain, Zipf is interpreted as the distribution of the user requests accessing the servers. As the requests are distributed among a finite number of objects, and the client requests are considered as independent events, the popularity of a web object is characterized by its probability. Therefore, the web object popularity model (Shi et al., 2005) is established based on a redefinition of a Zipf in the form of $P_i = C / i^{\alpha}$, where C is a constant, i the i_{th} most popular object and α the Zipf parameter see$^{\alpha}$n before.

Although power laws are defined in the domain of real numbers, natural numbers are used for web sites. Also the straight line is only an approximation where values at the bottom (e.g., those objects requested only once) and values in the firsts ranks (e.g., the most popular web objects) can appear outside the straight line. These two features can be found in previous traces, as it will be shown in section 5. There is a rank from where the distribution behaves as a Zipf. This value is noted as x_{min} (Clauset et al., 2009).

4. Related work

Web analytics services and Apache logs are useful sources of information to understand web consumption re-spectively in the client and in the server side. There are many examples of its usage in the bibliography but this article does not pretend to be an exhaustive enumeration of such evidences and results.

The first articles started to analyze logs obtained from proxies or from servers in the form of Apache logs. The main conclusion obtained after the corresponding analysis was that file popularity followed a

Zipf distribution. Differences were noticed in the results depending on the place where data was obtained (Breslau et al., 1999).

(Krashakov et al., 2006) presented a summary of traces found up to that moment. The authors includ-ed a set of proxy data sets and introduced a variation to the Zipf formula as follows: $fr = b / (c + r)^{\alpha}$. The variation consisted in adding a constant value c to the rank in the Zipf general formula. The authors concluded that the introduction of this variation makes α more stable, independent of the time over the years, and obtained a value of α of 1,02 ± 0,05.

There were a consensus in the bibliography around the fact that Zipf was bound between 0,6 and 1 regard-less the type of content, the size of the sample or the position of the log (Breslau et al., 1999; Podlipnig et al., 2003; Shi et al., 2005; Nair et al., 2010).

Indeed, new evidences were found in recent articles (Zotano et al., 2015b): 14 websites were analyzed using server-side logs, and all of them show a Zipf-like distribution but with $\alpha > 1$. No matter the sector, the underneath technology or the website's purpose, Zipf-like always happens. The result implies a higher frequency/popularity concentration on a few object compared to previous results. The pop-ularity is more important now than before in these websites, and cache policies on some of them were revis-ited for an update (Zotano et al., 2015a).

Regarding the client-side analysis, many different techniques are used. In (Calzarossa et al., 2016) and (Singal et al., 2014) there is an extensive analysis of evidences, methodologies, techniques as well as the state of the art in video analysis, social networks analysis, etc. Most of these works are based on Google Analytics and they are applied on the web. Google Analytics is free and therefore is convenient for test-ing. Additionally it is easy to be included in a website and results can be obtained very quickly (Google, 2016; Plaza, 2011).

As per results, in (Mahanti et al., 2009) a site was analyzed using the server-side and the client-side logs. It is interesting highlighting that Zipf-like distributions were found in the visits per visitor distribu-tion and in the number of pages per visitor distribution. Both have an $\alpha > 1$. That study, which last for one year, showed that a half of the users are coming from search engines.

Other metrics have been analyzed for a deep understanding of the user behavior. As an example some models appears in articles to explain the visit duration: one study points to demographics as a key factor to determine the duration time since female and older users visit the web for longer (Danaher et al., 2006).

Regarding the news sector, many reports are found containing global statistics focused on the kind of contents and usages from news websites (Cherubini et al., 2016), as well as tools and techniques. The main trends found are the higher relevance of Internet consumption in broadcasters' websites, and the higher smart phones' consumption.

After this analysis of related works, we could not find an analysis as deep, involving such a huge vol-ume of data, different types of usages and devices, analyzing trends over the years in the different layers, as the one we have made in this article. A key difference is the usage of *RTVE* as a reference: it includes a huge volume of data in the client and in the server sides, and this article can use internal and private in-formation. This makes this study more thorough and detailed than in previous works.

©Ediciones Universidad de Salamanca / cc by-nc-nd

5. Web analytics evidences

RTVE uses several analytic services, Adobe Omniture among others. A third party audits this data, in particular for the Spanish market, the company in charge of such auditing is OJD. The majority of the data attached in table 1 are public and can be consulted in their website.

But some pieces of data are not public, like the percentage of time spent per visit or the web pages requested from the client layer. Own analysis makes use of both public and private data.

Table 1: Web analytics figures

Metric	(July) 2014	(Feb) 2015	(Feb) 2016
Unique Users	17.348.642	19.492.087	19.193.320
Visits	65.453.758	77.741.268	93.937.776
Pages	327.884.09	364.712.83	363.249.62
Pages Per Visit	5,01	4,69	3,87
Average Visit duration	11:52	09:52	08:48
Average Page duration	02:15	01:52	01:59
% Visits smaller 1 minute	28,9	28,8	29,4
% Visits between 1-10 min	30	30,2	30,8
% Visits between 10-30	23,0	23,8	24,3
% Visits between 31-60 %	17,3	15,9	14,3
Visits higher 1 Hour	1,8	1,3	1,2
α	1,08	1,12	1,23

Table 1 shows that *RTVE* has around 19 millions unique users, that is, users that reached *RTVE* services at least once per month. So this analysis involves a huge amount of population. As per visits, it has grown very quickly, with a 50% of increase from 2014 to 2016, while the pages seem to be correlated to the unique users: the more users, the more pages. Regarding the pages per visit, a decrease is shown from 5,01 to 3,87 as well as the average visit and page duration. The comparison between 2015 and 2016 is relevant due to the similar amount of users, and in general, trends are consistent: more pages, more visits but less pages per visit and less visit duration.

Table 2: Logs analyzed

Period	Avg Requests per day	α
July	130 millions	1,41
Fe b	288 millions	1,43
Fe b	330 millions	1,33

As per duration ranges, data shows that the visits are becoming shorter, and visits smaller than 30 minutes are more frequent. A visit spending less than one minute is the most usual behavior, while the least usual one is spending more than one hour. More than 50% of users spend less than 5 minutes in *RTVE* services despite of *RTVE* being primarily a media website.

Regarding web analytics, a question to be answered is the relation between the visit profile, that is, users' behavior in a website, and web pages that are being requested from clients. To do so, the source of information used in this analysis is Adobe Omniture, where the web pages requested from users are saved. We have collected all the *URLs* stored in the web analytic tool for the months shown in table 1.

After some attempts, a model of web pages' popularity has arisen from the evidences: the shape of the page-frequency function follows a Zipf-like distribution (see section 3). The estimator by (Clauset et al., 2009) has been used for a positive matching and to determine the α value. The results appear in ta-ble 1. The three sets of *URLs* follow Zipf-like distributions that are increasingly more skewed through the years. It means that some contents are very popular, and they receive the majority of the requests. In some way, users are massively focused on only a certain set of pages, while the majority of them are not very demanded. But over the years this trend is increasing from 1,08 to 1,25, i.e., a 15,75% of increase in only two years.

Next section will analyze the server side perspective from *CDN* evidences.

6. Log analysis

We have collected data from *RTVE* website as a set of Apache-like access logs on the same days of the analysis in section 5. A major issue that makes this analysis different is the volume of files to be processed. In the case of *RTVE* logs involve up to 6.000 millions *URLs*.

An Apache-like log shows information for every single request a server receives, with the response code, user-agent, response size, *URL* and a timestamp. For legal restrictions, the *IP* addresses and other personal data are omitted from the logs. But not all the collected requests are needed for this analysis: only those with response method GET and response code equals to 200, 203, 206, 300, 301 or 410 are used (Breslau et al., 1999).

The data is processed using some commands and the frequencies are calculated. Then, the estimator by (Clauset et al., 2009) is used again to determine if the distribution follows a Zipf-like and to estimate α. Table 2 shows the results and includes the α estimation as well as the average number of requests per day.

It is important to highlight that α was becoming higher through the time, but in 2016 something changed: *RTVE* website moved to a responsive design, so the same page is consumed in a different way depending on the device. The technique used to adapt the content is based on 301 *HTTP* redirects depending on the user agent. It makes the same *URL* to be redirected to a different one. As a consequence, a smaller α is obtained.

The three samples in table 2 show an α value higher than one, implying a high frequency of a few objects. Remind that the higher α, the higher the popularity is concentrated. The lower α means more uniformity.

7. Analysis of the results

RTVE offers video and audio from the broadcast emission. The average duration for a radio emission is around one hour and for *TV* is around 50 minutes, but prime time emissions have an average duration of 90 minutes. It is important to note, that prime time contents are usually the most demanded by users in televi-

©Ediciones Universidad de Salamanca / cc by-nc-nd

sion. If we analyze the average duration shown in table 1, full episode are not so widely consumed by users due to the fact that only the 1,2% of visits have a duration higher than one hour, and therefore the upper bound of visits that watch a prime time program is only 1.116.000. These data highlight that online users' behavior is very different to broadcast consumption.

As per the average consumers in *RTVE*, they spend a shorter time compared to 2014. It was 11'52" in 2014 and 08'48" in 2016 which implies a decrease of 25,84%. Actually, almost 50% of the visits in 2016 are shorter than 5 minutes, and almost a 30% of visits are shorter than one minute. It implies that when users arrive to the website, they consume the content and it is more likely to leave the website compared to past years. It means that the consumption is more direct than in 2014 or 2015. All the segments higher than 30 minutes show a decrease but the others an increase.

The above behavior must be reviewed with the fact that pages per visit have decreased from 5,01 to 3,87: the conclusion is that users are more focused on the content they want to consume and that they spend less time to move around the contents of websites.

In *RTVE* we have detected a higher relevance of social networks and search engine. Social network users' behavior consist in accessing to the content and then living the page to share more contents or to comment about that content in the social network but not in the *RTVE* page. Users from search engines usually check the pages and, from the results, spend less time on additional content than before. One piece of data that seems to confirm this hypothesis is the fact of the decrease in the time spent for page and, especially the huge increase in the visits, up to 50% comparing 2014 with 2016.

Another fact to explain the consumption in *RTVE* is the more intensive consumption from smart phones. Data from OJD (OJD, 2016) shows that mobile users in *RTVE* spend only 03'40" per visit, with 12,5 millions visits on 2016, while in 2015 it was 11,5 millions. Take in mind that mobile users spend almost one-third of the average visit duration in the website. It implies a very direct access to content from mobile.

Regarding the correspondence between visits and web pages visited, table 1 shows that web pages visits follow a Zipf-like distribution. The distribution is driven by α as discussed in section 3. It can be observed that popularity is becoming more concentrated along the years. This result is consistent with the con-sumption pattern discussed above: a more compulsive consumption is happening. Actually, almost one-third of users leave the website and do not consume more contents, which is a bad news for the *SEO* (Search Engines Optimization) managers. One conclusion arises from these results: webmasters must pay attention to this facts since it points to a more relevance of external facts, that is, a good social network management or good positioning on search engines, than doing an effort for internal linking.

Moving to the technical point of view, that is the server analysis, web pages pattern should modify the web objects' frequency distribution. According to the trickle down effect (Doyle et al., 2002), the more cache levels implies a smaller α in the server. The contrary also happens. A key factor that must be considered is caching at client side: browser caches helps to soften the skewness of frequencies observed in server. If we compare the α value between 2014 and 2015, an increase happens due to two facts: in 2015 more visits and more users exists, while the pages in a visit becomes smaller. It could explain the increase in the α value. Another factor that support this increase is that the web pages distribution is more skewed in 2015 than in 2014, where α moves from 1,41 to 1,43.

Another issue appears from evidences to the latter hypothesis: when comparing 2015 and 2016, this increasing does not happen. In 2016 α decreases to 1,33 (7%). Our hypothesis is that this is due to the effects of a responsive redesign made in *RTVE* in June 2015: the same page behaves in a different way depending on the device. Pages do not change their *URL*s and this explains why web pages patterns keep a consistent

©Ediciones Universidad de Salamanca / cc by-nc-nd

trend. But the server, as different web objects are served, notices a different frequency pattern, resulting in a decline of α from 2015 to 2016. This hypothesis is supported by a huge increase in the average of daily requests, from 288 millions in 2015 to 330 millions, despite of *RTVE* maintains around 19 million users a day. In any case, the responsive design has not changed the highly skewed frequency distribution profile since α is still higher than one.

8. Conclusions

This work aimed to discover the user behavior from 2014 to 2016 in a mass media service. It also pretended to analyze how content popularity has evolved through the years and its impact in the client and the server side of a website. The results are relevant to understand how digital marketing techniques (sharing contents in social networks or content optimization for search engines) and mobility have affected a mass media website's usage. Trends obtained are applicable in many broadcasters like BBC or Mediaset since these marketing techniques are widely used in mass media websites.

The higher relevance of web objects' popularity were pointed in (Zotano et al., 2015b) and its effects were analyzed in (Zotano et al., 2015a), but a deeper analysis was needed: real evidences of how client side trends changes the server workloads in a highly demanded website offers a new perspective to webmasters to forecast scalability issues.

For the client layer, Adobe Omniture has been used. The user trends discovered points to a more di-rect access to the content, and a less webmaster's influence over user navigation. It is supported by some facts: the visits are shorter, the user consume less contents per visit and finally web pages frequency is more skewed, that is, popularity of web pages is becoming more unfair. We think this is a consequence of the increase in mobility, with its consumption patterns. Contents are also quickly consumed from social networks or from search engines, and the linked contents in that web page are not so consumed as before.

Trends in the client layer changes the server perception: the wider disparity in the client popularity distribution implies a higher α w h e n calculated in the server layer. There is a positive empirical correla-tion between these two facts.

Another conclusion obtained is the effect of the responsive design. No effect was detected in client layer trends, but they exist in the server layer. *URL*s requested in the client layer are redirected using a 30X response to other resources that are more suitable for the devices. This makes the server to split one request in others, and then, the effect is a smaller α.

For a future work, we want to work on the future trends: we plan to verify the trends with empirical data obtained in 2017, and we want to check the impact of a more social activity in terms of popularity and server workload. *RTVE* is working hard on a better search engine positioning and on social net-works, but also in improving the internal linking. It is interesting to analyze whether the latter technique can help to increase the visit duration.

Another issue we want to analyze empirically is the relation between the web pages distribution and the server distribution. We plan to create a set of tools to generate different client patterns and to analyze the real impact in a server. For that analysis we plan to build agent based tools to simulate users' behav-ior like those appearing in this study.

©Ediciones Universidad de Salamanca / cc by-nc-nd

9. References

Adobe, 2016. Metric Descriptions. https://marketing.adobe.com/resources/help/enUS/reference/metrics.html [accessed 26 April 2016]

Breslau, L., Cao, P., Fan, L., Phillips, G., and Shenker, S., 1999. Web caching and Zipf-like distributions: Evidence and implica- tions. In INFOCOM'99. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE, volume 1, pages 126–134. IEEE.

Calzarossa, M. C., Massari, L., and Tessera, D., 2016. Workload Characterization: A Survey Revisited. ACM Computing Surveys (CSUR), 48(3):48. http://dx.doi.org/10.1145/2856127

Cherubini, F. and Nielsen, R. K., 2016. Editorial Analytics: How News Media Are Developing and Using Audience Data and Metrics. Available at SSRN 2739328.

Clauset, A., Shalizi, C. R., and Newman, M. E., 2009. Power-law distributions in empirical data. SIAM review, 51(4):661–703. http://dx.doi.org/10.1137/070710111

Danaher, P. J., Mullarkey, G. W., and Essegaier, S., 2006. Factors affecting web site visit duration: a cross- domain analysis. Journal of Marketing Research, 43(2):182–194. http://dx.doi.org/10.1509/jmkr.43.2.182

Doyle, R. P., Chase, J. S., Gadde, S., and Vahdat, A. M., 2002. The trickle-down effect: Web caching and server request distribution. Computer Communications, 25(4):345–356. http://dx.doi.org/10.1016/S0140-3664(01)00406-6

Google, 2016. Google Analytics Web Site. https://www.google.com/analytics/ [accessed 17 April 2016].

Krashakov, S. A., Teslyuk, A. B., and Shchur, L. N., 2006. On the universality of rank distributions of website popularity. Computer Networks, 50(11):1769–1780. http://dx.doi.org/10.1016/j.comnet.2005.07.009

Mahanti, A., Carlsson, N., Arlitt, M., and Williamson, C., 2013. A tale of the tails: Power-laws in internet measurements. Network, IEEE, 27(1):59–64.http://dx.doi.org/10.1109/MNET.2013.6423193

Mahanti, A., Williamson, C., and Wu, L., 2009. Workload characterization of a large systems conference Web server. In Communication Networks and Services Research Conference, 2009. CNSR'09. Seventh Annual, pages 55–64. IEEE. http://dx.doi.org/10.1109/cnsr.2009.19

Nair, T. and Jayarekha, P., 2010. A rank based replacement policy for multimedia server cache using zipf-like law. arXiv preprint arXiv:1003.4062.

OJD, 2016. Evolución Audiencias RTVE.ES. http://www.ojdinteractiva.es [accessed 17-April-2016].

Plaza, B., 2011. Google Analytics for measuring website performance. Tourism Management, 32(3):477–481. http://dx.doi.org/10.1016/j.tourman.2010.03.015

Podlipnig, S. and Böszörmenyi, L., 2003. A survey of web cache replacement strategies. ACM Computing Surveys (CSUR),35(4):374–398. http://dx.doi.org/10.1145/954339.954341

Shi, L., Gu, Z., Wei, L., and Shi, Y., 2005. Quantitative analysis of zipf's law on web cache. Parallel and Distributed Processing and Applications, pages 845–852.

Singal, H., Kohli, S., and Sharma, A. K., 2014. Web analytics: State-of-art & literature assessment. In Confluence The Next Genera- tion Information Technology Summit (Confluence), 2014 5th International Conference-, pages 24–29. IEEE.

Urdaneta, G., Pierre, G., and Van Steen, M., 2009. Wikipedia workload analysis for decentralized hosting. Computer Networks, 53(11):1830–1845. http://dx.doi.org/10.1016/j.comnet.2009.02.019

Zotano, M. G., Gómez-Sanz, J., and Pavón, J., 2015a. Impact of traffic distribution on web cache performance. International Journal of Web Engineering and Technology, 10(3):202–213. http://dx.doi.org/10.1504/IJWET.2015.072349

Zotano, M. G., Sanz, J. G., and Pavón, J., 2015b. Analysis of Web Objects Distribution. In Distributed Computing and Artificial Intelligence, 12th International Conference, pages 105–112. Springer. http://dx.doi.org/10.1007/978-3-319-19638-1_12. Springe