Video popularity characterization centered on news-on-demand

M. González-Aparicio, R. García, J. L. Brugos, X. G. Pañeda,
D. Melendi, S. Cabrero

Computer Science Department, University of Oviedo
Gijón, Asturias, Spain
{maytega, garciaroberto, brugos, xabiel, melendi,
cabrerosergio}@uniovi.es

ABSTRACT

Video popularity characterization, in a News-on-demand service, is the core issue of this paper. It will concentrate on the digital edition of six Spanish regional newspapers, where the different news articles are classified into a wide variety of issues. Request video data has been analyzed, including how accesses were distributed along the different topics, during a period of nine months, where the number of requests fluctuated between 85,758 and 586,398 over a minimum of 519 videos and a maximum of 4,221. On the one hand, the statistical function from which video popularity would come from is likely to be unknown. The problem has been tackled with some statistical functions, such as Zipf-like, Mandelbrot and Stretched, which are well-known in the literature. Another method called Box-Cox transformation has been looked into for this specific field. On the other hand, a brief overview has been carried out about the popularity growth tendency, that is how many days a video takes to reach the totality of its requests, and if requests are focused on a small number of videos or they are spread, known as temporal locality. In the end, after an inference and validation process, Box-Cox transformation turned out to be the best choice for the majority of the on-line newspapers. Moreover, the topic related to regional news captured the highest number of requests, so highlighting their regional nature. Finally, new videos achieved 80% of their requests in their first day of life in most of the newspapers, and in four out of six on-line newspapers a high percentage of requests were concentrated on a low percentage of videos (Principle of Pareto).

KEYWORDS

News-on-demand, popularity characterization, Box-Cox transformation, Mandelbrot, Stretched, Zipf-like.

1. INTRODUCTION

Nowadays the presence of streaming media on the Internet is quite popular, especially in social networks such as YouTube, Yahoo or Facebook and web sites dedicated to news, sports, entertainment, education and even in the business world for marketing purposes. As a result, system designers have to face the new features of streaming media content, such as more computing power, an increase of bandwidth and storage requirements or a long-lived nature in order to supply good Web services [1]. Many technologies have emerged to manage this type of content and to reduce the impact over the different resources, among which it could be mentioned

multicast/unicast delivery, encoding formats or complex cache replacement policies, some of which are being improved steadily. However, more multimedia workloads have to be analyzed to achieve a well-known user access understanding.

This paper will be focused on video popularity characterization in a news-on-demand service, while most of the former studies are oriented to education, research or social fields. Specifically, it will be centered on six Spanish on-line digital newspapers with a regional scope, such as "La Opinión A Coruña" (www.laopinioncoruna.es) and "Faro de Vigo" (www.farodevigo.es) from the region of Galicia, "La Provincia-Diario de Las Palmas" (www.laprovincia.es) from the Canary Islands, "Levante-EMV" (www.levante-emv.com) and "Superdeporte" (www.superdeporte.es) from Valencia and "La Nueva España" (www.lne.es) from Asturias. These services have some remarkable features such as a daily introduction of news items, classification of the videos by categories and a wide variety of content. The length of the period analyzed was nine months, from 1st of January to 30th of September 2009. The analysis was developed with data obtained from log servers.

Video requests will be analyzed with a per-day granularity, because it is the period of time where the underlying set of videos does not change. On the one hand, one of the main contributions of this paper is the characterization of video popularity by a statistical function called Box-Cox transformation. To the best of our knowledge, it has not been applied before in this field. On the other hand, other statistical functions, which have been mentioned in former studies of streaming media workloads, have been taken into account in this paper, namely Zipf-like, Mandelbrot and Stretched. Therefore, a wide range of statistical functions in this field has been analyzed at the same time, in order to determine the most suitable for this type of service.

The concentration of requests could also be important when making a decision about using caching or content distribution systems. Indeed, if a high number of requests are for videos which have been published recently, a caching solution could be less efficient, because of the time involved in propagating the content to the caches. Therefore, it is a problem of a cost-effective design [2, 3]. After studying the six on-line newspapers, it could be observed that requests were concentrated in a low percentage of videos. However, this type of services show the user a high number of videos [4, 5]. Therefore, it seems to follow the Principle of Pareto. Indeed, in four on-line newspapers out of six, the mean number of the most popular videos was between 3% and 9% with a percentage of daily requests between 81% and 90%, and the service "Superdeporte" was the only one where 5% of the most popular videos received more than 90% of requests. However, none of the videos in "La Opinión A Coruña" or "La Nueva España" received more than 80% of requests. In conclusion, we believe that our study provides relevant results in the design field of user access patterns focused on news-on-demand services, which are characterized by the diversity of regions they belong to, a wide variety of videos classified by different issues and a high percentage of news centered on their own region.

The rest of the paper is organized as follows. Section 2 reviews previous work. Section 3 presents a case study related to six news on-line video-on-demand services from Spain. A characterization of video popularity and a first approach to temporal locality have been carried out along the six news services in Section 4. Finally, conclusions and future work are proposed in Section 5.

2. RELATED WORK

The presence of videos in a wide variety of media services (file sharing, media broadcast, video-on-demand or live streaming) has led to media researchers analyzing how the video nature per se could impact the efficiency of its distribution from the server to the user. Some video features have been analyzed, such as size, duration, type of content, user interactions and so on. Specifically, this paper focuses on video popularity characterization and temporal locality. In the literature, some statistical functions have been considered as the most appropriate to model video popularity. To date, Zipf-like function [6] has been one of the most applied in this context. In [7] a workload of one week was analyzed in a university environment, with streaming-media sessions from 4,786 clients to 866 servers on the Internet, who accessed 23,738 different streaming-media objects, where 78% were accessed only once, 1% ten or more times, and the 12 most popular objects more than 100 times each. The popularity distribution was modeled with Zipf-like with θ equal to 0.47. The conclusion was that accesses to streaming-media objects were less concentrated on the popular objects. Moreover, popularity has been studied in social networks where videos are classified by categories. In YouTube, [1, 5, 8] analyzed video requests during a period of three months, where the level of video popularity depended on the category and requests were focused on a specific number of videos ($\theta = 0.56$ or $\theta = 0.668$). In news-on-demand contexts, [9] analyzed a Norwegian on-line newspaper during a period of two years, with 4,6 million requests and 3,500 videos, where only popularity for the most popular videos was adjusted ($\theta = 1.2$). Moreover, in [10] a regional on-line newspaper was studied, where the new content was introduced every day and the workload was analyzed at five different time scales (full time, one year, three months, a fortnight and one day). In this paper it was corroborated that the value of θ decreased as the time scale became bigger. Therefore, it was concluded that the introduction of new content seemed to have an influence on θ, and an algorithm was suggested for working out θ every day, due to the fact that the content remained stable during that time. In [11] the regional on-line newspaper "La Nueva España" was studied during a period of six months, from January to June 2007, with more than 300,000 requests over 1,500 videos, where content popularity was characterized with the Mandelbrot function ($\theta = 1.3$; $k = 20.85$) and a weak correlation between file duration and file popularity was found.

In [12] sixteen workloads have been analyzed with different delivery methods (streaming, pseudo streaming, multicast, P2P and so on), different sizes of media file and duration (from 5 days to more than two years), and different types of contents. The video access pattern could be fitted with a Stretched Exponential function along all workloads. However, some factors have been taking into account that may affect media access patterns such as extraneous traffic [4], caching or “fetch-at-most-once” [13]. Indeed, the presence of extraneous traffic (31% of requests), such as ad and flag media clips, means that the different reference rank distributions were fitted with a Zipf-like function ($\theta = 0.71$), and the same happens in [7]. However, without this type of traffic the distinct workloads could be well fitted with a Stretched exponential model.

The concentration of user requests along the different videos, known as temporal locality, is another factor with a big influence on the selection of a delivery technology (multicast or caching). In [14] an algorithm called “Popularity and Partial Replication Load Sharing” was proposed, where a percentage of the most popular videos were copied in all servers, and the rest would be distributed according to a certain algorithm. If the value of θ (Zipf-like parameter) is low, the percentage of copies chosen had a great influence on the waiting time, and slight otherwise. In [15] a decentralized architecture network was studied, and particularly if the service

was broadcast, the total cost of the architecture decreased when θ grew. In [2] the growth pattern of video popularity was characterized, since the video was uploaded, in three different video datasets of YouTube, namely, videos that appear in the top lists, videos removed due to the lack of copyright and videos from random queries. This analysis highlights that popularity behavior was different along the three datasets. Indeed, top videos received a large fraction of their views on a single peak day, while the other two datasets experienced multiple smaller popularity peaks. This information is quite relevant in order to improve the effectiveness of content recommendation and search tools.

In our proposal, video popularity has been characterized with some statistical functions, such as Zip-like, Mandelbrot and Stretched. Moreover, another statistical function called Box-Cox transformation has been analyzed, which turned out to be the best in the majority of the on-line newspapers. In relation with the temporal locality, it could be confirmed that in four out of six on-line newspapers a low percentage of videos received a high percentage of requests.

3. NEWS-ON-DEMAND SERVICES

On-line newspapers inform readers about the latest news, mainly with texts and images, although certain news items are combined with a video as well. Speaking in general terms, some newspapers present a link located on their home site related to a news video section. Specifically, Spain presents three well-known on-line national newspapers with this feature, namely “El país” (www.elpais.es), “El mundo” (www.elmundo.es) and “ABC” (www.abc.es), where their link is called “Multimedia” or “Vídeos”. Moreover, each newspaper presents its videos in its own way according to a certain classification, perhaps due to its on-line publication policy. Currently, they seem to have in common two possible groups, namely the most watched and the latest videos. A brief glimpse of some international newspapers reveals that in U.K. "The Metro" (www.metro.co.uk) presents its videos classified into nine categories. In France "Le Monde" (www.lemonde.fr) classifies its videos out in descending order according to their publication date and in U.S.A. "New York Times" (www.nytimes.com) divides its videos into 24 categories, with a special one dedicated to local news called "N.Y./Region". In this paper, six on-line newspapers have been studied with a regional scope, but well-known in the region they belong to, namely “La Opinión A Coruña”, “Faro de Vigo”, “La Provincia", "La Nueva España”, "Levante-EMV", and "Superdeporte”. It is worth mentioning that all of their video webpages showed the same visual structure during the analysis, unlike the national and international newspapers previously mentioned. This was due to the fact that all of them were controlled by the same management content designer. Furthermore, videos were classified by categories, which were shown on the left side of the webpage, and when a category was chosen the most recent twelve videos were displayed and the remainder of the videos were displayed on different web pages, also in groups of twelve. So, videos were shown to the reader in descending order according to their publication date, from the most recent to the oldest.

3.1. Workload features

Some features could stress the importance of one service to another, such as the number of published videos per day or the number of requests that receive videos which have been released recently, which are referred to as new videos throughout the paper, as opposed to former videos. In Table 1 some of the characteristics of the news on-line services which have been studied in this paper can be observed. Firstly, “La Nueva España” (#586,398 requests) received approximately 7 times more requests than “La Opinión A Coruña” (#85,758 requests). Secondly, the number of

new videos in "La Nueva España" (#2,620 videos) stands out above the rest of the newspapers, where at most there are almost 10 times more than "Faro de Vigo" (#270 videos) and at least the number is nearly twice as high as "Levante-EMV" (#1,561 videos). However, many new videos receive none or a low percentage of requests. In fact, "Faro de Vigo" was the worst service with only 4% ((11/270)*100) of new videos demanded and "Levante-EMV" was the best with 97% ((1,517/1,561)*100). Surprisingly, the service "La Nueva España" with the highest number of new videos only 22% ((593/2,620)*100) were requested. Moreover, it is worthy of mention that in "La Provincia", "Superdeporte" and "Levante-EMV" roughly 97% of requests were directed towards new videos, while in "La Nueva España" the number reduced to 21%, and in "Faro de Vigo" and "La Opinión A Coruña" was only 3%. It would be necessary to look into this in deep, as it could be due to many factors such as social, cultural or demographic. These services are undoubtedly centered on specific communities of Spain, where many items of news were focused on local information, which could explain their high level of acceptance inside the Spanish community per se.

Table 1. Some workload features.

On-line newspaper	Number of new videos	Number of requested videos	Number of requests	Number of requests to new videos	Number of new requested videos
"La Opinión A Coruña"	469	764	85,758	2,982	22
"Faro de Vigo"	270	519	153,199	4,775	11
"La Provincia"	350	562	217,285	210,643	342
"Superdeporte"	336	752	238,114	233,334	311
"Levante-EMV"	1,561	2,345	383,031	375,108	1,517
"La Nueva España"	2,620	4,221	586,398	121,632	593

3.2. News video issues

News is displayed to the user through videos classified by issues. Some issues are common such as "Sports" or "International", others have a relation with the region that the on-line newspaper is linked to and the rest are quite different. It has to be highlighted that each on-line newspaper has its own classification. Therefore, in order to make a global comparison between the different newspapers, it was decided to group the main issues into nine sections according to the topic they are related to, namely "Sports", "Economy", "Local", "National", "International", "Society", "Culture", "Events" and "Weather", where the "Events" section includes news related to any kind of incidents. However, three sections are still missing in some newspapers, which are the sections "National" in "Faro de Vigo", and "Economy" and "Events" in "La Nueva España".

Fig. 1 depicts the percentage of the number of requests distribution from 1st of January to 30th of September 2009 of all the sections previously mentioned. It can be observed that "Local" is the section which receives the greatest number of requests within each on-line newspaper, with the minimum value being 47,628 (31%) requests in "Faro de Vigo" and the maximum value 112,372 (19%) requests in "La Nueva España". However, in "La Nueva España" the number of requests of the section "National" is 110,906 (18%), which is almost on a par with its "Local" section. On the other hand, the newspaper "Superdeporte" is a special case, due to the fact that it is solely dedicated to the sports world, for this reason it was not included in the graph, and its videos received 238,038 requests in all. On the one hand, it was superior to the rest of the on-line newspapers in comparison with the number of requests of their "Sports" section. In fact, the service with the largest number of requests in the "Sports" section was "La Nueva España" with 70,175 requests, almost three and a half times below "Superdeporte". On the other hand, the requests in the "Sports" section in the remaining on-line newspapers varied from 2,900 requests in "La Opinión A Coruña" to almost 34,000 requests in "Levante-EMV".

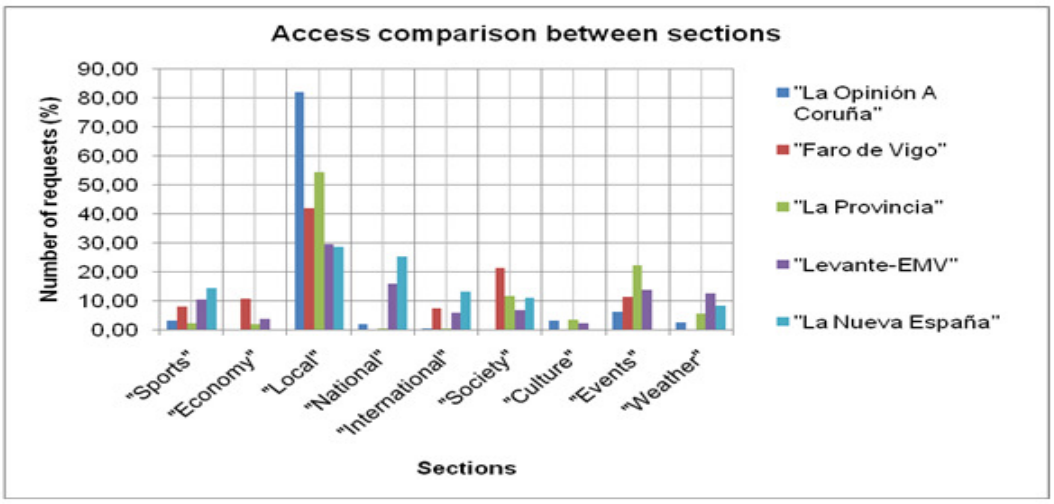

Figure 1.Percentage of the number of requests per section

4. DATA ANALYSIS

Video requests have been analyzed in six news-on-demand services, over a period of nine months, from 1st of January to 30th September 2009, in order to analyze the most recent requests available. Characterizing video popularity has been studied throughout the literature with different statistical functions. Some of them have been considered in this paper, such as Mandelbrot, Stretched and Zipf-like. Due to the fact that each function has its specific parameters, the problem is to find the right values in order to provide a high quality fit to the data. Moreover, another statistical method is proposed known as Box-Cox transformation. In all statistical functions, videos were classified in descending order according to their number of requests. As a result, a rank order distribution function was obtained.

4.1. Popularity growth patterns

A brief view of the popularity growth patterns could provide a first signal of some resource needs (bandwidth, caches, storage capacity and so on). This analysis is based on the number of days a

video takes to achieve from 10% of its requests totality to 100%. For this reason, it is important to establish the concept of a video's lifetime [2], which in this context will be understood as the number of days a video was receiving requests since it was released. The speed with which a video gains popularity during its first day of life was analyzed in all the different services. Fig. 2 depicts what percentage of new videos received between 10% and 100% of their total requests on their first day of life. A glance at Fig. 2 shows that the best percentage of new videos within each on-line newspaper turned out to be the following: in "La Opinión A Coruña" 30% (6 out of 20) of videos reached 30% of requests, in "Faro de Vigo" 30% (3 out of 10) of videos received 90% of requests, in "La Provincia" 17% of videos received 10%, in "Levante-EMV" 15% of videos received 70% of requests, in "La Nueva España" 18% of videos received 100% and finally "Superdeporte" was the most balanced because each percentage of requests in [10, 30, 40, 50, 60, 80] had 12% of videos. Moreover, the services "La Provincia", "Levante-EMV", "La Nueva España" and "Superdeporte" had a high percentage of videos which received more than 80% of requests on their first day of life, which were 24%, 24%, 38% and 23% respectively. Therefore, it seems that not all services would have the same level of resource consumption.

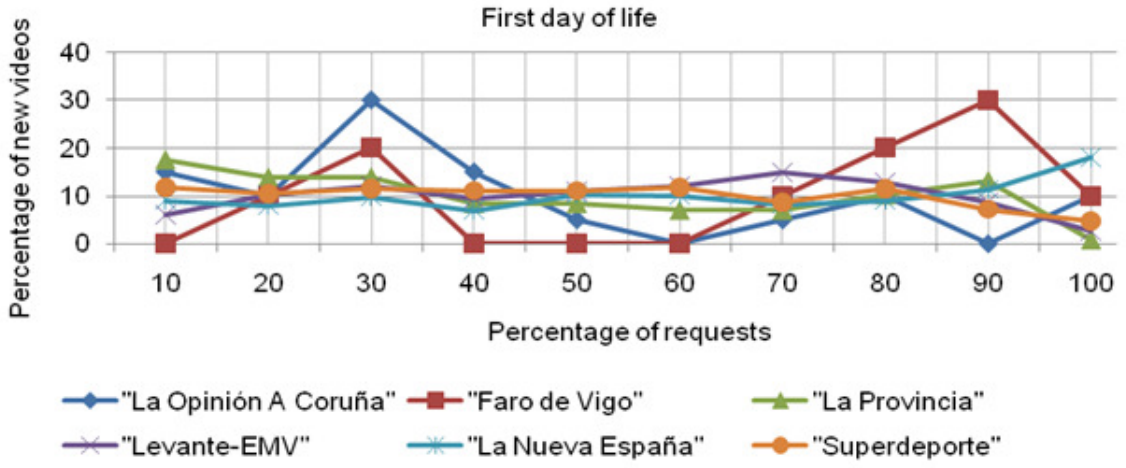

Figure 2. Percentage of new videos with a percentage of requests on their first day of life

4.2. Process for video popularity characterization

In general terms, a model has been proposed for characterizing the video popularity of each on-line newspaper. The length of the period of time, which has been studied for inferring the models was six months (182 days), from 1st of January to 1st of July 2009, where video popularity behavior has been analyzed day by day. For this reason, daily requests could be considered as a sample in statistical terms. On the one hand, each sample has been fit with Zipf-like, Mandelbrot, Stretched and Box-Cox transformation. On the other hand, techniques for measuring the goodness of fit, such as Chi-square and Kolmogorov-Smirnov test have been applied in order to discard those models which could not be considered good enough. In this paper, Kolmogorov-Smirnov test results have been chosen as the most reliable, as explained in Section 4.5 below. Eventually, daily requests could be adjusted to all statistical functions, some or none of them. In Table 2 the number of days where daily demand could be fitted successfully with the statistical method selected is shown.

Table 2. Number of days modeled with success

	Statistical functions			
On-line newspapers	**Zipf-like**	**Mandelbrot**	**Stretched**	**Box-Cox**
"La Opinión A Coruña"	158 (87%)	178 (98%)	165 (91%)	106 (58%)
"Faro de Vigo"	123 (68%)	145 (80%)	125 (69%)	165 (91%)
"La Provincia"	117 (67%)	162 (89%)	123 (68%)	127 (73%)
"Levante-EMV"	170 (97%)	172 (95%)	173 (95%)	39 (22%)
"La Nueva España"	182 (100%)	175 (96%)	182 (100%)	60 (33%)
"Superdeporte"	66 (36%)	133 (73%)	67 (37%)	140 (86%)

Then, in order to establish a representative value for each parameter inside each statistical function, the set of values associated to the parameter was checked under different tests. Firstly, the Shapiro test was applied to confirm if the set of values comes from a normal distribution. Secondly, the homoscedasticity was tested, if the Shapiro test was successful F test would be used, otherwise the Levene's test. Moreover, due to the fact that news items were published daily, it has been considered whether there were differences between demand behavior along the different days of the week. Therefore, models were distributed in seven groups, from Monday to Sunday, according to the day of the week they were associated with. Finally, the seven groups of models were compared through them through a confidence interval, with the F-Anova test of one factor, in order to detect any difference. If F-Anova test was successful, a representative value for each parameter could be proposed both for each group and for the totality. Many times, F-Anova test of one factor failed for the seven groups as a whole, which implied that one model was not enough to explain the video requests behavior. For this reason, it was necessary to look for more than one model for the service, and the group of seven days was split into two groups, which included those days of the week compatible between them (F-Anova test of one factor was positive for each of them). The division in two groups was enough. Then, for each parameter only the valid models were taking into account, and the mean of its set of values along the different models was chosen as its final datum. In the end, each service obtained a global model for characterizing video popularity, one for each day of the week and one or two general ones. This paper will focus only on the global models.

4.3. A brief view of the statistical distributions applied

Box and Cox (1964) [16] introduced a family of power transformation for a non-negative random dependent variable *y*. The method turns *y* into y(λ) where the family of transformations indexed by λ is expressed in the Equation (1).

$$y(\lambda) = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda} & if \quad \lambda \neq 0 \\ \log(y) & if \quad \lambda = 0 \end{cases} \tag{1}$$

In the end, the transformation chosen gives the best λ in order to obtain the best fit to the data. The parameter λ is calculated with the maximum likelihood method [17]. Moreover, a confidence interval 100*(1- α)% is defined for λ [16], for a given significance level α, and it is expressed in the Equation (2), where $\hat{\lambda}$ represents the maximum value that λ could achieve.

$$\left\{ \lambda : \lambda > \hat{\lambda} - \frac{1}{2} \chi_1^2 (1 - \alpha) \right\} \tag{2}$$

Box-Cox would offer a general linear model prediction such as Equation (3), which could allow the prediction of future demand based on independent past samples.

$$y(\lambda) = \beta_0 + \log(x)\beta_1 \qquad with \quad 1 \le x \le n \tag{3}$$

However, if one sample is fitted with a linear regression and R^2 is high, it does not mean that the same linear regression model could be applied to another sample of the population, although it could be good enough for that sample. For that reason, the parameters β_0 and β_1 must have significance in the model. In other words, statistically they must be checked with a hypothesis test if $\beta_0 = \beta_1 = 0$ (p_value $>$ 0.05), which we accordingly did.

Table 3. Statistical functions.

Statistical functions	Expressions
Zipf-like	$y = \dfrac{C}{x^{\theta}} \quad 1 \le x \le n, \quad with \quad C = \dfrac{1}{\sum_{x=1}^{n} \frac{1}{x^{\theta}}}$
Mandelbrot	$y = \dfrac{C}{(x+k)^{\theta}} \quad 1 \le x \le n \quad y \quad k \ge 0, \quad with \quad C = \dfrac{1}{\sum_{x=1}^{n} \frac{1}{(x+k)^{\theta}}}$
Stretched	$y_x^c = -a * \log(x) + b \quad with \quad 1 \le x \le n$

The remaining distributions applied in this paper have the expressions shown in Table 3.

4.4. The best statistical models for each on-line newspaper

Many times it was not possible to establish a global model at all, as was mentioned in Section 4.2. Indeed, any statistical functions could provide a unique model for each service. For instance, the distribution Zipf-like in the services "La Opinión A Coruña", "La Provincia" and "Superdeporte" obtained a unique model, but not for the rest. Indeed, in "Faro de Vigo" and "Levante-EMV" obtained one model for Monday to Friday and another for Saturday and Sunday together. However, "La Nueva España" obtained one for Monday to Saturday and another for Sunday. Another example is the Box-Cox distribution, where after applying the process explained in the Section 4.2, the results were the following. First of all the Shapiro test was tested, but it failed

along the different samples. Secondly, the Levene test was checked, and it was successful in all cases. Thirdly, from the point of view of F-Anova test, each parameter did not have significant differences (p_value > 0.05) between the models associated to the days of the week. However, this test failed in two on-line newspapers, “La Nueva España” and “Faro de Vigo” in relation to the parameter “λ” and parameter “a” respectively, with a p_value below 0.05 in both cases. In particular, video popularity on Saturday in “La Nueva España” seems to be different to Tuesday, and with a low permissible degree along the pairs (Tuesday, Friday), (Tuesday, Monday), (Thursday, Saturday) and (Wednesday, Saturday). The F-Anova test result for the parameter "λ" is represented in Fig. 3.

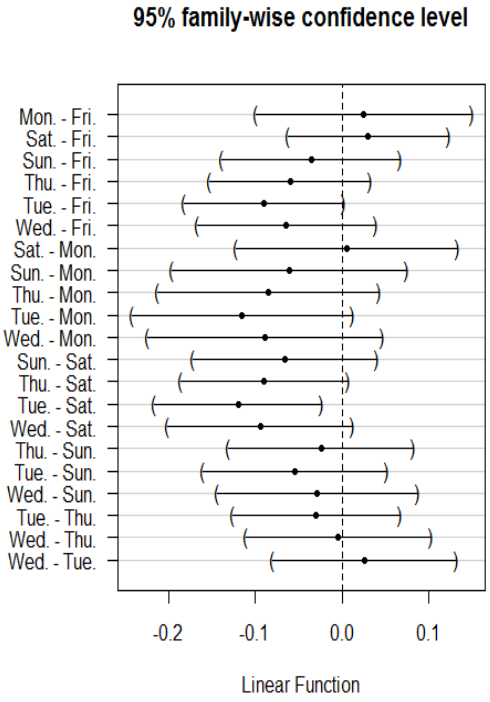

Figure 3. F-Anova test in "La Nueva España"

Finally, video requests patterns have been modeled through the statistical distributions Zipf-like, Mandelbrot, Stretched and Box-Cox transformation, whose results are expressed in Tables 4, 5, 6 and 7 respectively. The variable x represents the ranking of the different multimedia videos, and each parameter is presented with its standard deviation (s.d.) in parenthesis.

Table 4. Global Zipf-like models.

"La Opinión A Coruña"	"Faro de Vigo"	"La Provincia"
$\Theta = 1.5971$ (s.d. 0.2242) $R^2 = 0.9457$ (s.d. 0.02)	**Monday to Friday** $\theta = 1.6273$ (s.d. 0.3659) $R^2 = 0.9532$ (s.d. 0.0175) **Saturday and Sunday** $\theta = 1.3687$ (s.d. 0.2023) $R^2 = 0.9518$ (s.d. 0.0162)	$\theta = 1.6084$ (s.d. 0.3521) $R^2 = 0.9338$ (s.d. 0.0307)
"Levante-EMV"	**"La Nueva España"**	**"Superdeporte"**
Monday to Friday $\theta = 1.8162$ (s.d. 0.1907) $R^2 = 0.9527$ (s.d. 0.0187) **Saturday and Sunday** $\theta = 1.6896$ (s.d. 0.2364) $R^2 = 0.9550$ (s.d. 0.0197)	**Monday to Saturday** $\theta = 1.557$ (s.d. 0.1797) $R^2 = 0.9661$ (s.d. 0.0168) **Sunday** $\theta = 1.4429$ (s.d. 0.1293) $R^2 = 0.9638$ (s.d. 0.0161)	$\theta = 1.7668$ (s.d. 0.4219) $R^2 = 0.9348$ (s.d. 0.0218)

Table 5. Global Mandelbrot models.

"La Opinión A Coruña"	"Faro de Vigo"	"La Provincia"
$\Theta = 2.7659$ (s.d. 2.1426) $k = 4.4033$ (s.d. 11.6168)	$\Theta = 4.9714$ (s.d. 7.0705) $k = 16.6942$ (s.d. 31.3946)	$\Theta = 2.2536$ (s.d. 1.2076) $k = 2.7468$ (s.d. 8.28)
"Levante-EMV"	**"La Nueva España"**	**"Superdeporte"**
Monday to Friday $\Theta = 3.0356$ (s.d. 1.0124) $k = 5.7733$ (s.d. 5.2116) **Saturday & Sunday** $\Theta = 2.5341$ (s.d. 0.5207) $k = 3.4642$ (s.d. 2.1455)	**Monday to Friday and Sunday** $\Theta = 2.361$ (s.d. 0.2986) $k = 4.7341$ (s.d. 3.1228) **Saturday** $\Theta = 2.4603$ (s.d. 0.3733) $k = 6.5764$ (s.d. 3.9299)	$\theta = 2.3028$ (s.d. 1.2625) $k = 2.0357$ (s.d. 4.9789)

Table 6. Global Stretched models.

"La Opinión A Coruña"	"Faro de Vigo"	"La Provincia"
$a = -0.1354$ (s.d. 0.2635) $b = 1.3463$ (s.d. 0.6823) $c = 0.0628$ (s.d. 0.0986)	**Monday to Sunday except Friday** $a = -0.0364$ (s.d. 0.0833) $b = 1.1991$ (s.d. 0.15) $c = 0.0194$ (s.d. 0.0403) **Friday** $a = -0.1094$ (s.d. 0.1486) $b = 1.2716$ (s.d. 0.3794) $c = 0.0558$ (s.d. 0.0707)	$a = -0.0385$ (s.d. 0.117) $b = 1.1012$ (s.d. 0.3063) $c = 0.0207$ (s.d. 0.0592)

"Levante-EMV"	"La Nueva España"	"Superdeporte"
Monday to Saturday a = -0.1954 (s.d. 0.2399) b = 1.5127 (s.d. 0.6626) c = 0.0982 (s.d. 0.1065) **Sunday** a = -0.0631 (s.d. 0.0914) b = 1.1569 (s.d. 0.2272) c = 0.0346 (s.d. 0.0489)	**Monday to Sunday except Saturday** a = -0.0583 (s.d. 0.1112) b = 1.1604 (s.d. 0.2549) c = 0.0390 (s.d. 0.0601) **Saturday** a = -0.1124 (s.d. 0.1112) b = 1.3188 (s.d. 0.3265) c = 0.0782 (s.d. 0.0779)	a = -0.0209 (s.d. 0.0719) b = 1.0448 (s.d. 0.1491) c = 0.0091 (s.d. 0.0288)

Table 7. Global Box-Cox models.

"La Opinión A Coruña"	"Faro de Vigo"	"La Provincia"
B_0 = 3.5112 (s.d. 0.1673) B_1 = -1.5154 (s.d. 0.0669) R^2 = 0.9478 (s.d. 0.0262) λ = -0.1	**Monday to Friday** B_0 = 3.0364 (s.d. 0.1655) B_1 = -1.4315 (s.d. 0.0715) R^2 = 0.9435 (s.d. 0.0306) λ = -0.1 **Saturday and Sunday** B_0 = 3.0116 (s.d. 0.1432) B_1 = -1.2036 (s.d. 0.0632) R^2 = 0.9415 (s.d. 0.0297) λ = -0.1	B_0 = 2.5546 (s.d. 0.1707) B_1 = -1.396 (s.d. 0.0634) R^2 = 0.9294 (s.d. 0.0381) λ = -0.3
"Levante-EMV"	**"La Nueva España"**	**"Superdeporte"**
B_0 = 4.3750 (s.d. 0.7288) B_1 = - 1.8259 (s.d. 0.2503) R^2 = 0.9703 (s.d. 0.0112) λ = 0.1	**Monday to Sunday except Saturday** B_0 = 3.8341 (s.d. 0.2699) B_1 = - 1.5525 (s.d. 0.0242) R^2 = 0.9769 (s.d. 0.0068) λ = 0.02 **Saturday** B_0 = 4.0959 (s.d. 0.0884) B_1 = - 1.5326 (s.d. 0.024) R^2 = 0.9775 (s.d. 0.0053) λ = 0.02	B_0 = 2.3401 (s.d. 0.3292) B_1 = - 1.9462 (s.d. 0.1458) R^2 = 0.9120 (s.d. 0.0481) λ = -1

In the end this procedure has allowed us to find a global model for each service with the aim of being able to make popularity estimations for the different services. In Fig. 4 it can be observed (log-log scale) how the global models that we have calculated for the Box-Cox, Mandelbrot, Stretched and Zipf-like functions fit the popularity on Sunday 6th of September 2009 for the service "La Nueva España". Actually, all of them seem to provide a high adjustment for that specific day.

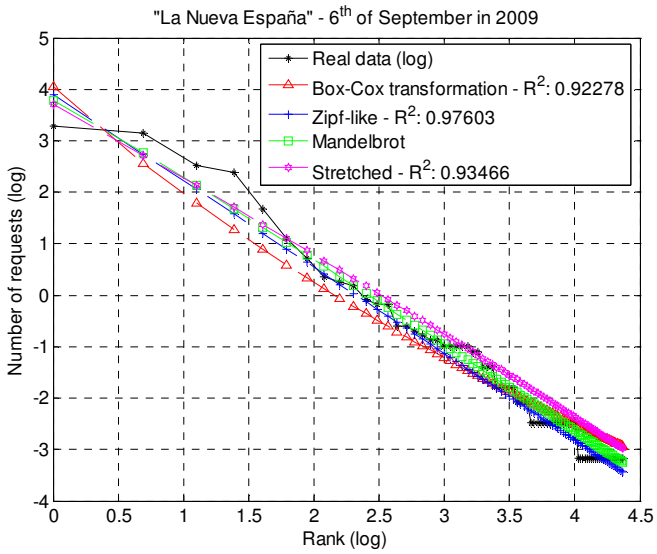

Figure 4. Fitting for video popularity.

Specifically, if only Box-Cox transformation is considered in Fig. 5 it could be observed an example of modeling popularity in the service "Levante-EMV", on Wednesday 11th of February 2009, with 1,220 requests distributed along 38 videos. The x-axis represents the reference rank of each video in log scale, and the y-axis represents the number of references to the corresponding video in log scale as well. This method of representation will be the same for the remaining figures of this paper. It has been fitted with the general model expressed in Table 5, the linear model which characterizes video popularity on Wednesday (a = -1.8633; b = 4.5613; λ = 0.1), and the model calculated for that day (a = -2.111; b = 5.3704; λ = 0.2), which has been called specific model. The result seems to be logical, where the specific model was the best (R^2 = 0.9699), and the Wednesday model (R^2 = 0.9379) almost drew with the general model (R^2 = 0.9320). Indeed, it could be observed that both models showed the same shape and they overlap.

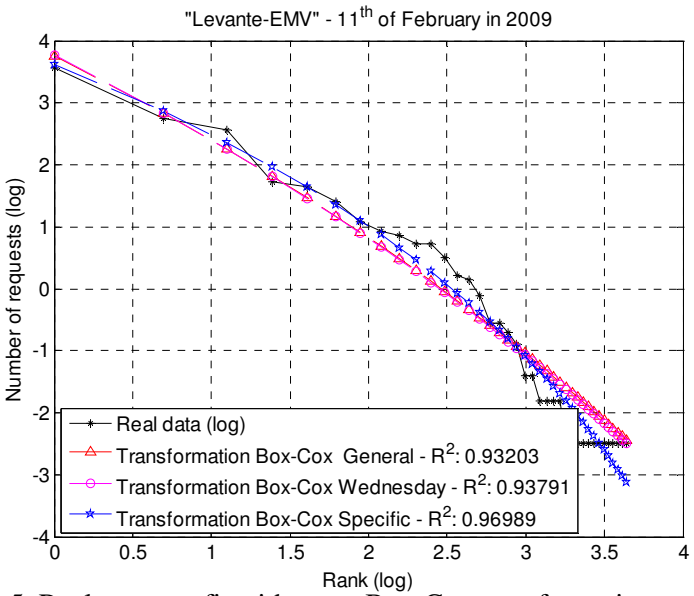

Figure 5. Real requests fit with some Box-Cox transformation models.

4.5. Model Validation

In order to evaluate the goodness-of-fit of the distributions Mandelbrot, Stretched, Zipf-like and Box-Cox transformation, the Kolmogorov-Smirnov (K-S) test has been used to decide if it is reasonable to assume if a sample comes from a population with a specific distribution. It is based on a comparison between the empirical cumulative distribution function (ECDF) and the theoretical one. This is checked with a hypothesis contrast where null hypothesis H_0 is if sample data comes from the stated distribution and the alternative hypothesis H_1 is that sample data does not come from the stated distribution. The hypothesis regarding the distribution form is rejected if the p value is lower than a significance level α. In this research the value chosen for α was 0.05. Kolmogorov-Smirnov test has two important features. On the one hand, it can be applied when the sample size is not big enough. On the other hand, it does not need to bin the data. This is an important detail to take into account because the result could change depending on how the data is binned.

Many papers have used Chi-square (χ^2) to measure the goodness-of-fit. The advantages mentioned before for the Kolmogorov-Smirnov test are the main drawbacks for χ^2 test. Indeed, in our study both methods have been applied. As a result, it could be proved that there were many cases where the χ^2 test passed when it should have failed. This phenomenon can be observed in Fig. 6, where video requests have been modeled with a Stretched distribution, and K-S test failed (p_value < 0.05), but χ^2 test passed (General model: p_value = 0.9995; Sunday model: p_value = 0.9853). In conclusion, the Kolmogorov-Smirnov test is more restricted and validation seems to be stronger.

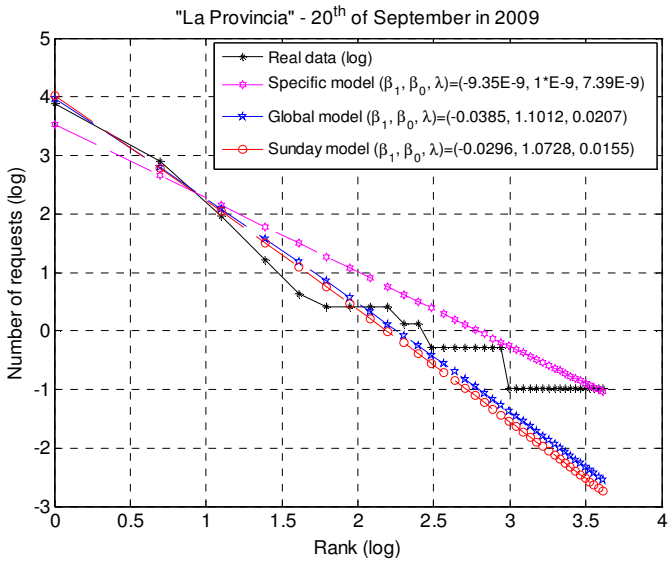

Figure 6. Fitting video popularity with Stretched distribution

Model validation has been made over the six on-line newspapers every day during a period of three months (91 days), from 2nd of July to 31st of September 2009. Global models have been validated with Chi-square and K-S tests. The results are shown in Table 8. Each box contains three numbers, the first represents the number of days with a positive result in Chi-square and K-S tests (p_value > 0.05), the second represents the number of days that Chi-square test was positive and the third the number of days that K-S test was positive. Although both validation tests have been calculated, due to the reason explained above, our analysis will be centered on K-S test.

Table 8. Number of success for Chi-square and K-S test (α = 0.05).

On-line newspapers	Validation for general model			
	Zipf-like	Mandelbrot	Stretched	Box-Cox
"La Opinión A Coruña"	64/ 71/ 75	17 / 18 / 40	34 / 37 / 53	36 / 37 / 77
"Faro de Vigo"	59/ 69/ 66	14 / 16 / 31	68 / 75 / 74	74 / 72 / 74
"La Provincia"	34/ 52/ 51	12 / 18 / 32	35 / 49 / 46	57 / 66 / 57
"Levante-EMV"	53/ 66/ 67	35 / 59 / 37	68 / 81 / 76	0
"La Nueva España"	62/ 88/ 62	44 / 72 / 48	83 / 91 / 83	78 / 91 / 82
"Superdeporte"	42/ 55/ 54	34 / 46 / 52	34 / 45 / 40	55 / 52 / 54

As can be observed in Table 8, for the services "La Opinión A Coruña" and "La Provincia" the Box-Cox transformation was the best with 77 and 57 days respectively, although there was a slight difference with Zipf-like in the first case with 75 days. In "Levante-EMV" and "La Nueva España" the best distribution was the Stretched distribution with 76 and 83 days respectively, but in the second service there was almost a draw with Box-Cox transformation with 82 days. Finally, "Faro de Vigo" could use two distributions Stretched or Box-Cox with 74 days, and "Superdeporte" could apply Zipf-like or Box-Cox with 54 days. However, it is worth pointing out that in "Levante-EMV" it was a disappointment, because none of the 91 days could be validated with Box-Cox transformation. In conclusion, Box-Cox transformation turned out to be a good choice, except for one service.

4.5. Requests concentration

After studying how requests are concentrated along the different videos in the six on-line newspapers, it could be confirmed that requests were centered on a few videos. However, there was an exception in "La Opinión A Coruña" and "La Nueva España", where requests were more spread along the different videos. This effect could be due to the presence of videos with a high interest for the user in both services [4, 5, 18]. The results of the percentage of videos with a certain percentage of requests are depicted in Table 9. The highest percentage of videos with the lowest percentage of requests was observed in "La Nueva España", where 24% of videos received a number of requests equal to or lower than 10%. If a percentage of requests equal to or lower than 50% is taken into account (mean number of requested videos from [1-10] to [41-50]), "La Nueva España" was again the first, where requests went to 84% of the videos, in second position would be "Levante-EMV" with 82%, in third position was "La Opinión A Coruña" with 75%, "Faro de Vigo" and "La Provincia" drew in fourth position with 62%, and finally "Superdeporte" with 56%.

Table 9. Percentage of requests per group of requested videos.

	Mean number of requested videos per service (%)					
Number of requests (%)	"La Opinión A Coruña"	"Faro de Vigo"	"La Provincia"	"Levante-EMV"	"La Nueva España"	"Super-deporte"
[1-10]	19	17	15	18	24	13
[11-20]	15	12	13	18	17	11
[21-30]	14	11	12	16	15	11
[31-40]	14	11	10	16	15	10
[41-50]	13	11	12	14	13	11
[51-60]	10	11	12	11	8	10
[61-70]	9	8	12	4	4	11
[71-80]	6	10	9	0	4	10
[81-90]	0	9	5	3	0	8
[91-100]	0	0	0	0	0	5

In the end, some on-line newspapers seem to follow the Principle of Pareto, that is a high percentage of requests have to be concentrated in a small percentage of videos. The same conclusion has been observed in [18, 19]. In this context, it could be said that in four on-line

newspapers out of six, the mean number of the most popular videos would be between 3% and 9%, with a daily percentage of requests between 81% and 90%. Moreover, the on-line newspaper "Superdeporte" would be the only one where 5% of videos would receive between 91% and 100% of requests. On the contrary, "La Opinión A Coruña" and "La Nueva España" did not follow the Principle of Pareto with a percentage of requests lower than 81%.

5. CONCLUSIONS AND FUTURE WORK

Many distributions have been researched in order to look for the best fit to the popularity. At the moment, Stretched distribution seems to be the best distribution which fits a wide variety of popularity with different characteristics [12]. The rest of the distributions could be applied when videos have specific features [10, 20]. Although a specific workload might be fitted for a statistical distribution, the same type of distribution might fail to capture another workload [21]. For this reason, in our study, Box-Cox transformation is proposed as an alternative method for modeling popularity in a variety of news-on-demand. Moreover, it must be pointed out that a Zipf-like distribution is a particular case of Box-Cox ($\lambda = 0$).

Box-Cox provides good predictions with our data. It has to be taken into account that our videos have a short duration due to the fact that they are related to news. Therefore, a user is likely to watch videos only once [13, 22] due to their nature. Exceptionally, videos related to periodic events could be seen more times and perhaps follow a seasonal pattern [10]. This method provides a unified approach that covers the majority of the lifetime distributions to avoid facing model selection. Finally, a model is proposed for the whole service. Moreover, if it is necessary to raise the level of precision, there is a model associated to each day of the week, which would enhance the degree of adjustment.

News-on-demand services present their news items classified by categories to their readers, as our services do. In our analysis, it is concluded that not all categories or sections attracted the same percentage of requests. It is noteworthy that "Local" is the section which received the highest number of requests in each on-line newspaper. Therefore, this fact highlights their regional nature. Popularity of one section against others was corroborated in other services. For instance, videos in YouTube are classified by categories and accesses have a tendency towards a small number of sections [5, 8, 21].

Requests were more spread along the different videos published in the six on-line newspapers. Firstly, the percentage of requests was studied in intervals of ten, and in the majority of the cases a uniform distribution of the mean percentage of videos requested per group could be observed, when the percentage was lower than 81%, with slight variations. Only a small percentage of videos received more than 81% requests, between 3% and 13% in four out of six newspapers, but 0% in the remainder. Therefore, the concentration of requests followed the Principle of Pareto in four out of six on-line newspapers. Secondly, It is worth highlighting that in four on-line newspapers out of six, a high percentage of videos, 23% (65 out of 278) at least and 38% (193 out of 507) at most, obtained more than 80% of requests in their first day of life.

In conclusion, this study proposes the Box-Cox transformation as an alternative statistical distribution to characterize video popularity in digital editions of some local newspapers. To the best of our knowledge it has not been applied before in this kind of services. This method provided good results in the majority of the on-line newspaper workloads.It was confirmed that their local section was the most demanded, which highlighted their regional nature. Finally, in the

majority of the on-line newspapers the concentration of requests followed the Principle of Pareto. Future work would follow two lines of research. On the one hand, a wider variety of news on-line video-on-demand services should be tested in order to confirm the Box-Cox applicability. Moreover, if longer periods of time were considered the models should be more reliable. On the other hand, not all the services have presented the same degree of popularity and their sections have not been equally demanded. Therefore, the service administrator should plan all the network resource needs and schedule a better distribution of content, with the best cost-effective purpose. In conclusion, both lines should be combined in order to achieve a better accomplishment in the use of video-on-demand resources.

ACKNOWLEDGEMENTS

This work has been funded by different sources. On the one hand, the network operator Telecable from Asturias SAU and the Prensa Ibérica Editorial within the project UNOV-11-MA-03. On the other hand a National R+D Plan within the project TSI2007-60474. The authors are grateful for the data provided by the digital newspapers "La Opinión A Coruña", "Faro de Vigo", "La Provincia", "Levante-EMV", "La Nueva España" y "Superdeporte". This paper would not have been possible without their support.

REFERENCES

[1] X. Kang, et al., "Understanding internet video sharing site workload: a view from data center design," Journal of Visual Communication and Image Representation, Vol. 21, pp. 129-138, 2010.

[2] F. Figueiredo, et al., "The Tube over Time: Characterizing Popularity Growth of YouTube Videos," in Web Search and Data Mining Conference, Hong Kong, China, 2011.

[3] M. Dakshayini and T. R. Gopala Krishnan Nair, "Client-to-Client Streaming Scheme for VoD Applications," Journal of Multimedia & Its Applications, Vol. 2, No. 2, May 2010.

[4] H. Yu, et al., "Understanding user behavior in large-scale video-on-demand systems," SIGOPS Oper. Syst. Rev., Vol. 40, pp. 333-344, 2006.

[5] P. Gill, et al., "Youtube traffic characterization: a view from the edge," in the 7th ACM SIGCOMM Conference on Internet measurement, San Diego, California, USA, 2007.

[6] Z. G. K., Ed., Human Behavior and the Principle of Least-Effort. Cambridge, MA: Addison Wesley, 1949.

[7] M. Chesire, et al., "Measurement and analysis of a streaming-media workload," in Symposium on Internet Techonologies and Systems, 2001, pp. 1-12.

[8] X. Cheng, et al., "Understanding the characteristics of internet short video sharing: YouTube as a case study," ArXiv e-prints, Vol. 707, 2008.

[9] F.T.Johnsen, et al., "Workload characterization for news-on-demand streaming services," in Performance, Computing, and Communications Conference, New Orleans, LA, 2007, pp. 314-323.

[10] X.G.Pañeda, et al., "Popularity analysis of a video-on-demand service in a digital newspaper: influence of the subject, video characteristics and new content publication policy," Journal of Adv. Media Commun., Vol. 1, pp. 369-385, 2007.

[11] R. García, et al., "Probabilistic analysis and interdependence discovery in the user interactions of a video news on demand service," Journal of Computer Networks, Vol. 53, pp. 2038-2049, 2009.

[12] L. Guo, et al., "The stretched exponential distribution of internet media access patterns," in 27th ACM Symposium on Principles of distributed computing, Toronto, Canada, 2008.

[13] M. Cha, et al., "I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System," in Internet Measurement Conference, San Diego, California, 2007, pp. 1-14.

[14] S. González, et al., "A case study of load sharing based on popularity in distributed VoD systems," IEEE transactions on Multimedia, Vol. 8, pp. 1299-1304, 2006.

[15] T. Wauters, et al., "Optical network design for video on demand services," in 9th Conference on optical network design and modelling, Milan, Italy, 2005, pp. 251-260.

[16] Z. Yang, "Predicting a Future Lifetime through Box-Cox Transformation," Journal of Lifetime Data Analysis, Vol. 5, pp. 265-279, 1999.
[17] R. H. Myers, Ed., Classical and Modern Regression with Applications. United States of America: Duxbury Press, 1990.
[18] S. Mitra, et al., "Characterizing Web-Based Video Sharing Workloads," ACM Transactions on the Web, Vol. 5, pp. 1-27, 2011.
[19] Y. Borghol, et al., "Characterizing and modelling popularity of user-generated videos," Journal of Performance Evaluation, Vol. 68, pp. 1037-1055, 2011.
[20] W.Tang, et al., "Modeling and generating realistic streaming media server workloads," Journal of Computer Networks, Vol. 51, pp. 336-356, 2007.
[21] S. Acharya, et al., "Characterizing user access to videos on the world wide web," in Proceedings of Multimedia Computing and Networking San Jose, CA, 2000, pp. 130-141.
[22] K. P. Gummadi, et al., "Measurement, modeling, and analysis of a peer-to-peer file-sharing workload," Proceedings of the 19th ACM Symposium on Operating systems principles, Bolton Landing, NY, USA, 2003.

AUTHORS

María Teresa González-Aparicio is a Computer Science Engineer from the University of Oviedo. Nowadays, she is an Associate Professor in the Department of Computer Science at the University of Oviedo. Her current research interest is in the area of multimedia systems and services, specifically in video popularity characterization.

Roberto García has a Ph.D. degree from the University of Oviedo and a Telecommunication Engineering degree from The Technical University of Madrid. He is an Associate Professor in the Department of Computer Science and Engineering, University of Oviedo, and former Associate Professor with the Electronics Department at the University of Alcalá (Madrid). His current research interests are in the area of telecommunication networks and services, applied to the performance analysis, modeling and simulation of telecommunication systems and multimedia services. He is also taking part in several research projects on the national and European levels.

Jose Antonio López Brugos is a professor at the Unversity of Oviedo.His thesis was related to computation evolution in FORTRAN at the Valencia University. He made doctoral courses at the University of Valencia and Pierre et Marie Curie (UPMC, Paris VI). He was a teacher visitor at AII Edimburgh, Sussex University, Manchester Polithecnic. His studies are centered on Automatic Theory Proving for Logic Programming. He is a guest Professor at Pierre et Marie Curie with the specialization in Logic Programming & BDDeductives, TeleLearning, Multimedia, software designing in computer games.

Xabiel García Pañeda is Computer Science Engineer and PhD from the University of Oviedo. Nowadays, he is an Associate Professor in the Department of Computer Science of the University of Oviedo and member of the SYMM Working Group of the W3C. His current research interests are in the area of multimedia systems and services, in content distribution networks, digital interactive TV services and mobile ad-hoc networks. He is also taking part in several research projects on the national and European levels.

David Melendi is Computer Science Engineer from the University of Oviedo and PhD from the University of Oviedo. Nowadays, he is an Associate Professor in the Department of Computer Science of the University of Oviedo and member of the SYMM Working Group of the W3C. His current research interests are in the area of multimedia systems and services, in content distribution networks, digital TV services, mobile ad-hoc networks, performance analysis, modeling and simulation of telecommunication systems. He is also taking part in several research projects on the national and European levels.

Sergio Cabrero is a telecommunication engineer from the University of Oviedo and a Ph. D. student. He is also an Assistant Professor in the Department of Computer Science at the University of Oviedo. His current research interests are in telecommunication networks, digital TV, multimedia services, and mobile ad hoc networks.