Measuring temporal redundancy in sequences of video requests in a News-on-Demand service

María Teresa González-Aparicio *, R. García, J.L. Brugos, Xabiel G. Pañeda, D. Melendi, S. Cabrero

Computer Science Department, University of Oviedo, Gijón, Asturias, Spain

A R T I C L E  I N F O

Article history:
Received 16 February 2013
Accepted 27 October 2013
Available online 7 November 2013

Keywords:
News-on-Demand
Video requests
Temporal redundancy
Forecast

A B S T R A C T

Streaming media is becoming one of the major components of Internet traffic. Therefore, a better understanding of users' video request patterns is essential, in order to design an effective and efficient video distribution system (caching, storage capacity, bandwidth, etc.). In this paper, the core issue will be the analysis and modeling of video requests temporal redundancy. The study will be centered on a News-on-Demand (NoD) service, which provides support to a wide variety of digital newspaper editions from different regions of Spain. Specifically, six digital newspapers with a high number of requests were analyzed during a period of one year. The level of redundancy has been measured by a global (gR) and a partial redundancy (pR) method, which is new in this type of services. As a result, the main contribution of our paper is a global and partial redundancy model for each digital newspaper, which would allow us to forecast the level of video requests likely to be repeated in the near future. The model turned out to be user independent and with a time-less effect. The validation process shows that all the models successfully pass the hypothesis test, which means that there were no significant differences between the model and the real data. The pR models could predict between 1% and 6% of video requests temporal redundancy with a level of accuracy which varies between 88% and 100%.

1. Introduction

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 (Kang and et al., 2010). Many technologies have emerged to manage this type of content and to reduce the impact on the different resources, such as 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.

The majority of the former studies are oriented to education, research or social fields on the Internet. However, this paper will be focused on an NoD service, which provides support to six Spanish digital newspapers among others, namely “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...
The main contribution of our paper is the proposal of a global and partial redundancy model for each digital newspaper, which would allow us to forecast the level of video requests likely to be repeated in the near future. In fact, the percentage of requests repetitions is important when a decision needs to be made, both in the short-term and in the long-term, to choose the best content distribution in time. Indeed, if a high number of requests are for videos which have been released 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 (Figueiredo and et al., 2011), where different alternative solutions have been analyzed in the literature, such as the combination of caching and replica placement (Bakiras and Loukopoulos, 2005) or a centralized decision-maker for content distribution (Caviglione and Cervellera, 2011). After studying the six on-line newspapers, it could be proved that in digital newspapers with a high level of repetitions per day it would be possible to make predictions between sequences of requests close in time to each other. In conclusion, we propose that our study provides relevant results in the design field of user access patterns focused on a news-on-demand service, which is characterized for the content distribution in a wide diversity of digital newspapers.

The rest of the paper is organized as follows. Section 2 reviews previous work. In Section 3 a case study is presented related to six Spanish digital newspapers, where a characterization of temporal consistency in sequences of video requests has been carried out. The validation results are shown in Section 4. Finally, conclusions and future work are proposed in Section 5.

2. Related work

In a broad variety of media services (file sharing, media broadcast, video-on-demand or live streaming), the search for video request patterns has helped media researchers to reach a better understanding of how the efficiency of video distribution from the server to the user could be improved. In the literature, many studies on video requests are based on video popularity characterization, where some statistical functions have been considered as the most appropriate. The common objective of all the previous statistical distributions is the search for a reliable model capable of predicting the level of requests for a video, according to its rank of popularity. To date, Zipf-like function (G. K., 1949) has been one of the most applied in this context. In (Chesire and al., 2001) a workload of one week was analyzed in a university environment, with streaming-media sessions from 4786 clients to 866 servers on the Internet, who requested 23,738 different streaming-media objects, where 78% were requested 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 $\theta$ equal to 0.47. The conclusion was that requests 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, (Kang and et al., 2010; Gill and al., 2007; Cheng and al., 2008) 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 ($\phi = 0.56$ or $\phi = 0.668$). In news-on-demand contexts, (T. Johnsen and al., 2007) analyzed a Norwegian on-line newspaper over a period of two years, with 4.6 million requests and 3500 videos, where only popularity for the most popular videos was adjusted ($\phi = 1.2$). Moreover, in Pañeda and et al. (2007) a regional on-line newspaper was studied, where 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). It was corroborated that the value of $\theta$ decreased as the time scale became bigger. In (García and et al., 2009) 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 for over 1500 videos, where content popularity was characterized with the Mandelbrot function ($\phi = 1.3$; $k = 20.85$) and a weak correlation between file duration and file popularity was found.

In (Guo and al., 2008) 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 2 years), and different types of contents. The video request pattern could be fitted with a Stretched Exponential function along all workloads. However, some factors have been taken into account that may affect media request patterns such as extraneous traffic (Yu and et al., 2006), caching or “fetch-at-most-once” (Cha and al., 2007). 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 ($\phi = 0.71$), and the same happens in Chesire and al. (2001). However, without this type of traffic the different 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 (González and et al., 2006) 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 were distributed according to a certain algorithm. If the value of $\phi$ (Zipf-like parameter) is low, the percentage of copies chosen had a great influence on the waiting time, but only slight otherwise. In (Wauters and al., 2005) a decentralized architecture network was studied, and particularly if the service was broadcast, the total cost of the architecture decreased when $\phi$ grew. In (Figueiredo and et al., 2011) 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, an alternative method has been studied in order to detect temporal consistencies and temporal redundancies in news video access series in time. Firstly, temporal consistency determines the level of regularity on different news video access sequences. Secondly, temporal redundancy refers to how accesses are concentrated along different news videos, varying from accesses concentrated on few news videos to a more dispersed behavior (a lot of different news videos). For this purpose, a global and partial redundancy method has been proposed, which has not been used before in this type of services. Throughout the literature, it could be found that these methods have been applied in Lancieri and Durand (January 2006) to analyze, during a period of 17 months, the level of regularity on how often a web site (URL) was demanded, and how often the same keywords were used. The conclusion was that temporal consistency in web sequences of queries was linked to the length of the sequence and the gap from one sequence to another. In the end, some statistical distributions, such as Zipf-like, Mandelbrot or Stretched, have been applied to characterize video popularity in many studies, but the main contribution of our paper is the proposal of a global and partial redundancy model for each digital newspaper. These metrics would allow to forecast the level of video requests likely to be repeated in the near future, where factors like user behavior and time would not influence on the temporal consistency prediction.

3. Digital newspapers in analysis

In this paper, video request redundancies have been analyzed in a digital edition of six Spanish newspapers, which are well-known in the region they belong to, namely “La Opinión A Coruña”, “Faro de Vigo”, “Levante-EMV”, “La Provincia”, “La Nueva España”, and “Superdeporte”. It is worth mentioning that all of their video web pages showed the same visual structure during the analysis. This was due to the fact that all of them were controlled by the same management content designer. Table 1 shows the number of requested videos and the number of requests, from 1st January to 31st December 2010. As a first approximation, these features could stress the importance of one newspaper to another. Indeed, the number of requests in the newspaper “Superdeporte” was 14%, 8% and 4% higher than “La Opinión A Coruña”, “Faro de Vigo” and “Levante-EMV” respectively. How-ever, it would be necessary to look into this in depth, as this could be due to many factors such as social, cultural or demographic. These digital newspapers are undoubtedly centered on specific communities of Spain and many items of news are related to local information, which could explain their level of user attention inside the Spanish community per se. Moreover, they offer a wide variety of other types of news items related to international, sports or culture news, among others.

The source of the information comes from a log file, where a record is added for every user’s video request. However, only two fields are considered in each record, the “time stamp” and the “ID” of the video. In this paper, it is important to consider the different requests sorted out chronologically by their “time stamp”, in order to analyze the video request redundancies as time moves forward.

3.1. Methods for the evaluation of temporal consistency

Temporal consistency has been looked into video request redundancies, that is to say how the concentration of requests is over the different videos. For example, requests could be concentrated on a few videos (strong redundancy) or on the contrary, requests could be more dispersed (weak redundancy). In consequence, this study will be focused on the search of a level of regularity in the sequence of videos, in order to look for temporal consistencies. On the one hand, the ratio between the number of unique videos and the total number of requests during the entire period would suggest how the requests are spread over the different videos. Therefore, this measure would provide an idea of the redundancy as a whole. This metric is called global redundancy ($gR$) and it is expressed in the Eq. (1). Therefore, there is no redundancy when all requested videos are different ($gR = 0$), but there is a high redundancy when all requested videos are the same.

\[
gR = 100 - \frac{\text{# unique videos}}{\text{# total requests}} \times 100\%
\]  

On the other hand, a set of requests are ordered chronologically and are split into sequences with an equal size $T$, where $T$ will correspond to a specific percentage of the total number of requests in the set. It would be possible to look for repetition

<table>
<thead>
<tr>
<th>Digital newspapers</th>
<th>“La Opinión A Coruña”</th>
<th>“Faro de Vigo”</th>
<th>“Levante-EMV”</th>
<th>“La Provincia”</th>
<th>“La Nueva España”</th>
<th>“Superdeporte”</th>
</tr>
</thead>
<tbody>
<tr>
<td># Videos</td>
<td>529</td>
<td>520</td>
<td>1106</td>
<td>819</td>
<td>2494</td>
<td>1514</td>
</tr>
<tr>
<td># Requests</td>
<td>29,653</td>
<td>48,449</td>
<td>97,804</td>
<td>211,754</td>
<td>409,607</td>
<td>418,332</td>
</tr>
</tbody>
</table>
of elements from one sequence to another. Therefore, for each pair of sequences \((Q_i, Q_j)\), the ratio is obtained of how many unique videos (items) from the first sequence \(Q_i\) are requested in the second sequence \(Q_j\). This metric is called partial redundancy (pR) and it is expressed in the Eq. (2).

\[
pR(Q_i, Q_j) = 100 \left( \frac{\# \text{items} Q_i \cap Q_j}{\# \text{items} Q_i} \right)
\]  

The whole process for partial redundancy (pR) calculation will consist in analyzing pairs of sequences according to whether or not they have sequences between them, as illustrated in Fig. 1. The sequence gap between the two sequences to be analyzed will be represented by the parameter delta \((\Delta)\). Firstly, pR will be calculated between pairs of consecutive sequences \((pR(Q_1, Q_2), pR(Q_3, Q_4), ...; \Delta = 0)\). Secondly, a gap of one sequence will be left between each pair \((pR(Q_1, Q_3), pR(Q_2, Q_4), ...; \Delta = 1)\). Then, pR will be studied with a gap of two sequences \((pR(Q_1, Q_4), pR(Q_2, Q_5), ...; \Delta = 2)\), and so on. The maximum length of a sequence gap (delta) between a pair of sequences and the maximum sequence length \(T\) will have to be established at the beginning of the experiment.

In the end, a set of requests will be tested with different sequence lengths \(T\) (\% items) and different number of sequence gaps. For each pair of values \((T, \Delta)\) a final pR value will be calculated, which will come from the mean of all pR values obtained from the pairs of sequences with the same size \(T\) and the same distance \(\Delta\) between them. For instance, the partial redundancy for sequences with size \(L\) and delta equal to 0 will be \(pR(T, 0) = \text{mean}(pR(Q_1, Q_2), pR(Q_2, Q_3), ..., pR(Q_{S-1}, Q_S))\), if delta is equal to 1 will be \(pR(T, 1) = \text{mean}(pR(Q_1, Q_3), pR(Q_2, Q_4), ..., pR(Q_{S-2}, Q_S))\) and so on. The procedure will be led by the function represented in Algorithm 1.

Algorithm 1. Partial redundancy calculation.

```plaintext
Function Calculate Partial Redundancy (Requests): matrix [[T1, T2, ..., Tm], [0, \Delta1, \Delta2, ..., \Deltaq]]
//Input: video requests sequence 'Requests'
//Output: Matrix 'mean Partial Redundancy' of 'mxq', where a cell (i,j) represents the mean of all pR calculated with pairs of sequences with 'i' length and a distance of 'j' sequences between them.

Requests = {Item1, Item2, ..., ItemN} // Video requests sequence, where Itemi represents the chosen video in the i-th position
N = Number of requests in a certain period of time

mean Partial Redundancy[[T1, T2, ..., Tm], [0, \Delta1, \Delta2, ..., \Deltaq]] = 0

For each sequence T \in [T1, T2, ..., Tm], [0, \Delta1, \Delta2, ..., \Deltaq]] = 0

\(L = N/T\) //Number of items in the sequence

S = N/L //Number of sequences with length L

\(\text{Building S sequences with length } L\) each
initial Position = 1;

For i from 1 to S
Q(i) = Requests[initial Position ... (initial Position + L - 1)];
initial Position = initial Position+L;
End For i

For each gap delta \((\Delta) \in [0, \Delta1, \Delta2, ..., \Deltaq]

\(\text{For j from 1 to (S-delta-1)}\)

Partial Redundancy(j) = Compute Partial Redundancy(Qj, Q(j+delta+1))

End For j

Mean Partial Redundancy(T, \Delta) = mean(Partial Redundancy(1 ... (S-delta-1))

End For \Delta

End For T

return Mean Partial Redundancy

End Function
```
The whole period of time, which represents a sequence of video requests ordered chronologically, has been divided into blocks with the same size $B$ (specific for each digital newspaper), and numbered from one onwards. For each block the function from Algorithm 1 was applied, and a matrix with information related to partial redundancies for pairs $(T_i, D_j)$ was obtained. The odd blocks together were used to build a unique model and the even ones were used to build another unique model for the validation of the first one. In Algorithm 2 is represented how to build the two final models.

Algorithm 2. Final partial redundancy calculation.

```algorithm
Requests = \{Item_1, Item_2, ..., Item_M\} //Video requests sequence from the whole period of time with length M
P = M / B; //Number of Blocks with length B
//Building P blocks with length B
initialPosition = 1;
For i from 1 to P
    B(i) = Requests[initialPosition ... (initialPosition + B - 1)];
    initialPosition = initialPosition+B;
End For i
//Calculation of $pR$ for each block
For i from 1 to P
    partial Redundancy Block,i(m, q) = Calculate Partial Redundancy(Bi)
End For i
//Calculation of a final model
For i from 1 to P
    If (i is odd)
        odd Final Partial Redundancy(m,q)=odd Final Partial Redundancy(m, q)+partial Redundancy BlockM_i(m, q)
    else
        par Final Partial Redundancy(m, q)=par Final Partial Redundancy(m, q)+partial Redundancy BlockM_i(m, q)
    End
End For i
middle = P/2
par Final Partial Redundancy(m, q) = par Final Partial Redundancy(m, q) / middle
If(P is odd)
    odd Final Partial Redundancy(m, q) = odd Final Partial Redundancy(m, q) / (middle+1)
else
    odd Final Partial Redundancy(m, q) = odd Final Partial Redundancy(m, q) / middle
End
```

3.2. Analysis of temporal consistency in the six digital newspapers

It is a fact that the number of requests, from one day to another or from service to service, suffers fluctuations, which could be more acute with the appearance of peaks or lack of requests at certain periods of time (holidays, weekend, working day, etc.). In order to reach a first approximation to a timeless effect and to be less vulnerable to fluctuations of requests, a uniform distribution of daily requests $N$ (#requests/365) has been considered in the six digital newspapers. In consequence, a homogeneous analysis within each service is proposed. Moreover, these services have a dairy incorporation of new content, making it desirable to make predictions of the percentage of video requests repetition within short periods of time. Therefore, the whole set of video requests, from the 1st January to the 31st December 2010, has been divided into blocks with the same size $B$ ($N/2$). The $B$ size value for each digital newspaper is shown in Table 2.

The results for global redundancy ($gR$) in the six digital newspapers are shown in Table 3. According to the $B$ size division, $gR$ was calculated for each block $B$ individually. Then, the mean $gR$ is calculated for the odd blocks set and another one for the even blocks set. It can be observed that the two digital newspapers with the lowest number of requests per block, “La Opinión A Coruña” and “Faro de Vigo”, reached the highest $gR$. On the contrary, the digital newspaper with the highest number of

<table>
<thead>
<tr>
<th>Digital newspapers</th>
<th>“La Opinión A Coruña”</th>
<th>“Faro de Vigo”</th>
<th>“Levante-EMV”</th>
<th>“La Provincia”</th>
<th>“La Nueva España”</th>
<th>“Superdeporte”</th>
</tr>
</thead>
<tbody>
<tr>
<td>$B$ Size</td>
<td>162</td>
<td>264</td>
<td>534</td>
<td>1,160</td>
<td>2,244</td>
<td>2,292</td>
</tr>
</tbody>
</table>
requests per block B, "Superdeporte", reached the lowest gR. Speaking in probability terms, since we have used the average gR, it is likely that in smaller blocks the number of repetitions is lower than in bigger ones. However, the number of available news videos in the different digital newspapers should be another factor to be aware of.

Moreover, in order to take into account temporal consistency of requests it was considered variations of the sequence length inside the block B, as mentioned in Section 3.1. It was interesting to choose a relative value for $T$ and not an absolute one, so a $T$ value would represent a percentage of the total number of requests inside the block B. In this paper, $T$ values belonging to $[1\%, 2\%, ..., 50\%]$ have been considered. The pR results in the six digital newspapers for $\Delta = 0$ and variations of the sequence length $T$ are shown in Fig. 2. The abscissa axis represents the sequence length $T \in [1\%, 2\%, ..., 50\%]$, and the ordinate axis represents the average of pR of all odd blocks, where both axes are in log–log scale.

Moreover, in Fig. 2 it can be observed that the decrease in the pR moves linearly according to the increase in the sequence length $T$, with an excellent coefficient of regression up to 0.96. This linear behavior seems to be logical. In fact, when a sequence length $T$ is short, requests to the same video are closer to each other in time, especially if those videos have been released quite recently. As the sequence length grows, the distance between requests to the same video increases. Therefore, the number of requests to the same video drops significantly from one sequence length $T$ with respect to any of the previous sequence lengths $T' \ (T' < T)$, in comparison with the total number of requests in $T$. This fact would become sharply with the soaring sequence lengths. Consequently, partial redundancy goes up as the sequence length decreases. Moreover, users preferences in digital newspapers could change with the arrival of new videos into the website, because the most recently and/or the most popular videos could be the likelihood of getting the user’s attention in a short period of time. Therefore, depending on their preferences, users may prefer to watch videos that have been recently released.

Table 3
<table>
<thead>
<tr>
<th>gR</th>
<th>Digital newspapers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&quot;La Opinión A Coruña&quot;</td>
</tr>
<tr>
<td>Odd blocks</td>
<td>11.2856 (s.d. 13.7464)</td>
</tr>
<tr>
<td>Even blocks</td>
<td>11.1247 (s.d. 13.3415)</td>
</tr>
</tbody>
</table>

Table 4
<table>
<thead>
<tr>
<th>Parameters</th>
<th>Digital newspapers</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a$</td>
<td>&quot;La Opinión A Coruña&quot;</td>
</tr>
<tr>
<td></td>
<td>-0.4454</td>
</tr>
<tr>
<td>log ($b$)</td>
<td>1.3568</td>
</tr>
<tr>
<td>$\kappa^2$</td>
<td>0.9739</td>
</tr>
<tr>
<td>Standard error</td>
<td>0.0488</td>
</tr>
</tbody>
</table>
on the sequence length $T$, the partial redundancy ($pR$) could be expressed according to the linear equation expressed in (2) in logarithmic scale.

$$\log(pR(T)) = -a \log(T) + \log(b)$$

In linear scale the Eq. (2) is equivalent to the Eq. (3).

$$pR(T) = b \cdot T^{-a}$$

The $b$ value is a constant for a scale adjustment, which represents how much the partial redundancy is when $T$ is equal to the unity, and $a$ is the slope of the straight line in the logarithm representation. The results for each of the six digital newspapers according to the Eq. (2) are represented in Table 4.

As explained in Section 3.1., the gap between pairs of sequences $(Q_i, Q_j)$ would be another parameter to take into account in partial redundancy calculation. It makes sense to think that as the gap increases, the pair of sequences will have a lower number of video requests in common. The partial redundancy results obtained in “La Nueva España” for $\Delta \in \{0, 1, 2, 3, 5, 10\}$ in log–log scale can be observed in Fig. 3. On the one hand, these results reflect that when two sequences were close in time the partial redundancy was higher. Indeed, the maximum value for $pR$ was obtained when delta was equal to 0. On the other hand, if we observed a particular $T$ length value, the partial redundancy decreased as the delta value increased. Indeed, the
The difference between the slope from the linear model associated to one $D$ value to the next ($D + 1$) went down at a constant rate, and the same happens with its origin value. This fact could be observed in Fig. 4, where the sets of slope and origin values, which are associated to the different $D$ models, follow a linear model with an excellent coefficient of regression higher than 0.97 in both cases.

In conclusion, the parameter delta should be part of the Eq. (3), and the final expression should be the Eq. (4).

The initial value for delta is always 0, but its last value will be associated to the last valid model ($p$ value >0.05). It would vary from one digital newspaper to another. The different $\gamma$, $\beta$ and the range of valid $D$ values for each digital newspaper are shown in Table 5.

Table 5
$\Delta$, $\gamma$ and $\beta$ values in the six digital newspapers.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Digital newspapers</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta$ values</td>
<td>“La Opinión A Coruña” “Faro de Vigo” “Levante-EMV” “La Provincia” “La Nueva España” “Superdeporte”</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>[0, 5]</td>
</tr>
<tr>
<td>$\beta$</td>
<td>$-0.1103$ ($R^2 = 0.95$)</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>$-0.0381$ ($R^2 = 0.93$)</td>
</tr>
</tbody>
</table>

$pR(T, \Delta) = e^{(b-\gamma\Delta)} \cdot T^{-(a/\beta\Delta)}$  \hspace{1cm} (5)

In Figs. 5–10 are depicted (log–log scale) how the models are similar to the real data for each digital newspaper. In all cases the coefficient of regression was up to 0.95. The model ensures with a high probability that for a specific $D$ value, $pR$ will fall along the plot as $T$ increases.

4. Validation

As mentioned in Section 3.1 the even blocks have been used for the validation process. The different models, which have been proposed for each of the six digital newspapers in Section 3.2 had a high adjustment degree with the real data. Indeed, all the models pass successfully the hypothesis test ($p$ value >0.05), which means that there were no significant differences between the model and the real data. Therefore, the forecast would reach a high reliability. The quality of the adjustment could be observed in Figs. 11–16 for the different digital newspapers (log–log scale), where apparently slight differences between the model and the real data seem to be higher due to the logarithmic scale.

Finally, the validation results predict a video requests temporal redundancy of between 1% and 6% with a level of accuracy which varies between 88% and 100%. The minimum level of repetitions was reached in “Superdeporte”, of 1% with 89% of accuracy, and the maximum value was reached in “Faro de Vigo” with 6% of repetitions and 100% of accuracy. This fact highlights that in digital newspapers with a high number of videos, such as “Superdeporte”, each video is less likely to be selected several times. On the contrary, in digital newspapers with a low number of videos, such as “La Opinión A Coruña” and “Faro de Vigo”, videos have less competence between each other, and they are more likely to be chosen several times. Bearing some of the hallmarks of a video distribution system, it seems that a combination between caching and replica placement...
Partial redundancy in “Faro de Vigo” - B size = 264

Fig. 6. Partial redundancy in “Faro de Vigo” with $B = 264$.

Partial redundancy in “Levante-EMV” - B size = 534

Fig. 7. Partial redundancy in “Levante-EMV” with $B = 534$.

Partial redundancy in “La Provincia” - B size = 1160

Fig. 8. Partial redundancy in “La Provincia” with $B = 1160$. 
should be established in order to reduce time latency (Bakiras and Loukopoulos, 2005). In digital newspapers with a high number of videos the caching tends to be more suitable, while in digital newspapers with a low number of videos a replica placement seems to be better.

5. Comparison with other statistical distributions

Many statistical distributions have been applied in video popularity characterization, as it was mentioned in Section 2. Mainly, the most common statistical distribution is Zipf-like, followed by Stretched and Mandelbrot. Their mathematical expressions are shown in Table 6.

In the six digital newspapers, the introduction of new content has a dairy rhythm. For this reason, video popularity has been analyzed with a dairy approach in this paper. However, it could be likely the presence of peaks or lack of requests at certain periods of time (holidays, weekend, working day, etc.), as it was said in Section 3.2. In order to reduce this effect, alternative days have been selected in order to reach a more precise model, which would represent a dairy video popularity behavior. Therefore, the odd days were considered to build the model and the even days for its verification. The same idea was used for the calculus of a partial redundancy model in Section 3.1, where the different blocks B were taken alternatively.

In the distribution Zipf-like, the parameter $\theta$ would reflect how requests would be concentrated on different videos. In fact, when the number of videos with the highest popularity shrinks, then the value of $\theta$ increases. In Table 7, the results for the six digital newspapers are shown. The newspaper “Superdeporte” got the highest concentration of requests on a few videos. This was due to the fact that only a few specific news sport events captured the majority of user’s attention.
Validation in "La Opinión A Coruña" - B size = 162

Validation in "Faro de Vigo" - B size = 264

Validation in "Levante-EMV" - B size = 534

Fig. 11. Validation of pR in “La Opinión A Coruña” with $B = 162$.

Fig. 12. Validation of pR in “Faro de Vigo” with $B = 264$.

Fig. 13. Validation of pR in “Levante-EMV” with $B = 534$. 
Fig. 14. Validation of $pR$ in “La Provincia” with $B = 1160$.

Fig. 15. Validation of $pR$ in “La Nueva España” with $B = 2244$.

Fig. 16. Validation of $pR$ in “Superdeporte” with $B = 2292$. 
According to the parameter $h$, the level of requests’ concentration in “Superdeporte” ($h = 1.7062$) turned out to be 50% higher than in “La Nueva España” ($h = 1.1297$) at the minimum and more than 100% in “Faro de Vigo” ($h = 0.7979$) at the maximum. Therefore, the number of videos with a high demand in “Superdeporte” is lower than in “La Nueva España” and much lower than in “Faro de Vigo”. In relation to the validation results, the level of accuracy varies between 98% and 100%.

In the distribution Mandelbrot, similar results were reached. Once again, in the newspaper “Superdeporte” the parameter $h$ ($h = 5.5209$) got the highest value and in “Faro de Vigo” got the lowest as it is reflected in Table 8. In the validation results, the level of accuracy varies between 83% and 98% in the parameter $h$, and between 67% and 93% in the parameter $k$.

Finally, in the Stretched distribution the parameter $c$ decreases when the request concentration increases. In the validation results, the level of accuracy varies between 84% and 98% in the parameter $c$, between 78% and 97% in the parameter $a$, and between 89% and 99% in the parameter $b$ (Table 9).

In the Fig. 17 is shown a comparison of a specific parameter from the different statistical methods in the six digital newspapers. It could be observed that the parameter $a$ in the partial redundancy seems to present small variations in its

Table 6
Statistical distributions.

<table>
<thead>
<tr>
<th>Statistical distributions</th>
<th>Expressions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zipf-like</td>
<td>$y = \frac{1}{x^h}$, $1 &lt; x \leq n$, with $C = \frac{1}{\sum_{x=1}^{n} x^h}$</td>
</tr>
<tr>
<td>Mandelbrot</td>
<td>$y = \frac{1}{(x+k)^h}$, $1 &lt; x \leq n$, with $C = \frac{1}{\sum_{x=1}^{n} (x+k)^h}$</td>
</tr>
<tr>
<td>Stretched</td>
<td>$y'_c = -a + \log(x) + b$, with $1 &lt; x \leq n$</td>
</tr>
</tbody>
</table>

Table 7
$h$ Value in the six digital newspapers.

<table>
<thead>
<tr>
<th>Zipf-like distribution</th>
<th>Digital newspapers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter $\theta$ (Odd days)</td>
<td>“La Opinión A Coruña”</td>
</tr>
<tr>
<td>$\theta$</td>
<td>0.8264 ($R^2 = 0.77$)</td>
</tr>
<tr>
<td>$\theta$ (Even days)</td>
<td>0.8412 ($R^2 = 0.78$)</td>
</tr>
</tbody>
</table>

Table 8
$h$ and $k$ values in the six digital newspapers.

<table>
<thead>
<tr>
<th>Mandelbrot distribution</th>
<th>Digital newspapers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter $\theta$ (Odd days)</td>
<td>“La Opinión A Coruña”</td>
</tr>
<tr>
<td>$\theta$</td>
<td>2.0965</td>
</tr>
<tr>
<td>$k$</td>
<td>2.5289</td>
</tr>
<tr>
<td>Even days</td>
<td>2.0464</td>
</tr>
<tr>
<td>$k$</td>
<td>2.1136</td>
</tr>
</tbody>
</table>

Table 9
c, $a$ and $b$ values in the six digital newspapers.

<table>
<thead>
<tr>
<th>Stretched distribution</th>
<th>Digital newspapers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter $\theta$ (Odd days)</td>
<td>“La Opinión A Coruña”</td>
</tr>
<tr>
<td>$c$</td>
<td>0.2134</td>
</tr>
<tr>
<td>$a$</td>
<td>0.4027</td>
</tr>
<tr>
<td>$b$</td>
<td>2.3862</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.78</td>
</tr>
<tr>
<td>Even days</td>
<td>0.1881</td>
</tr>
<tr>
<td>$a$</td>
<td>0.3701</td>
</tr>
<tr>
<td>$b$</td>
<td>2.2196</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.78</td>
</tr>
</tbody>
</table>
behaviour, followed by the parameter $\theta$ in Zipf-like and parameter $c$ in Stretched. Finally, the worst case is the parameter $\theta$ in Mandelbrot.

6. Conclusions and future work

In this paper, the level of redundancy in sequences of video requests has been analyzed, belonging to six Spanish digital newspapers from different regions of Spain, with a wide variety of content and type of users. It is a fact that the number of requests, from one day to another or from service to service, suffers fluctuations, so in order to reach a timeless effect and reduce the vulnerability to daily request fluctuations, a uniform distribution of daily requests has been considered. In consequence, a homogeneous analysis has been proposed for each service. Moreover, these services have a daily incorporation of new content. Therefore, predictions have been made of the percentage of video request repetitions between pairs of sequences of video requests which were close in time.

The presence of regularity in sequences of video requests has been studied with a partial and global redundancy method. Firstly, the global redundancy results showed that when the number of requests per block ($B$) became smaller the number of different requested videos per block increases. Indeed, the two digital newspapers with the lowest number of requests per block $B$ and the lowest number of videos, ”La Opinión A Coruña” and “Faro de Vigo”, had the highest global redundancy. On the contrary, the digital newspaper with the highest number of requests per block $B$ and with a high number of videos, “Superdeporte”, had the lowest global redundancy. Therefore, it seems that the number of videos with a high probability to be requested from one sequence to the following increases as the number of requests per block and the number of videos decreases. Secondly, a partial redundancy has been studied between pairs of request sequences with a certain gap between them. From the analysis it could be concluded that the $pR$ value could be determined from the knowledge of the sequence length ($T$) and the sequence gap ($\Delta$). As a result, the behavior of $pR$ function ($pR = f(A, T)$) turned out to be independent from the user and time evolution, as it was in Lancieri and Durand (January 2006). It has to be pointed out that the maximum value for $pR$ was obtained when there was no gap between pairs of sequences ($\Delta = 0$) and where the temporal consistency reached its maximum value.

In conclusion, the main contribution of our paper is the proposal of a global and partial redundancy model for each digital newspaper, which would allow to forecast the level of video requests likely to be repeated in the near future. The validation results predict between 1% and 6% of video requests temporal redundancy with a level of accuracy which varies between 88% and 100%. Therefore, the different pR models obtained excellent results in the six digital newspaper workloads. Moreover, pR seems to have a low level of fluctuations along the different newspapers in comparison with the rest of statistical methods. Therefore, pR could be likely to provide video popularity predictions with more accuracy. Future work would follow two lines of research. On the one hand, a wider variety of NoD services should be tested in order to confirm the behavior of pR. On the other hand, another key factor is the prediction of the feasibility and the performance that the knowledge of video requests repetitions could provide, in order to detect user interests in videos related to specific categories of the newspapers (Legouix and et al., 2000; Lancieri and Durand, 2004). 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.
Acknowledgments

This work has been funded by different sources. On the one hand, the network operator Telecable from Asturias SAU and on the other, the Prensa Ibérica Editorial within the project UNOV-11-MA-03. 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" and "Superdeporte". This paper would not have been possible without their support.

References


P. Gill, et al., Youtube traffic characterization: a view from the edge, in Proc. 7th ACM SIGCOMM Internet measurement, San Diego, California, USA, 2007.


M. Cha, et al., I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system, in Proc. 7th ACM SIGCOMM Internet measurement, San Diego, California, USA, 2007.


