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

CrossMark

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

ARTICLE INFO

ABSTRACT

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%.

© 2013 Elsevier Ltd. All rights reserved.

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

---

* Corresponding author. Tel.: +34 985182512.
E-mail addresses: maytega@uniovi.es (M.T. González-Aparicio), garciaroberto@uniovi.es (R. García), brugos@uniovi.es (J.L. Brugos), xabiel@uniovi.es (X.G. Pañeda), melendi@uniovi.es (D. Melendi), cabrerosergio@uniovi.es (S. Cabrero).

"Superdeporte" (www.superdeporte.es) from Valencia and "La Nueva España" (www.lne.es) from Asturias. Despite their regional scope, all these digital newspapers have some important features, such as a great diversity of users, a daily introduction of news items and a wide variety of videos related to news classified by different categories (national, international, sports, etc.). The length of the period analyzed was one year, from 1st of January to 31st of December 2010.

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 θ 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 ($\theta = 0.56$ or $\theta = 0.668$). 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 ($\theta = 1.2$). 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 θ 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 ($\theta = 1.3$; $k = 20.85$) 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 ($\theta = 0.71$), 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 θ (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 θ 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