Journal of Network and Computer Applications 36 (2013) 219–227

Contents lists available at SciVerse ScienceDirect

Journal of Network and Computer Applications

journal homepage: www.elsevier.com/locate/jnca

Towards a predictive cache replacement strategy for multimedia content

Jeroen Famaey *, Frédéric Iterbeke, Tim Wauters, Filip De Turck

Ghent University – IBBT, Department of Information Technology, Gaston Crommenlaan 8 bus 201, B-9050 Gent, Belgium

ARTICLE INFO

Article history:
Received 6 April 2012
Received in revised form
14 July 2012
Accepted 20 August 2012
Available online 18 September 2012

Keywords:
Multimedia content delivery
Caching strategies
Popularity prediction

ABSTRACT

In recent years, telecom operators have been moving away from traditional broadcast-driven television, towards IP-based interactive and on-demand multimedia services. Consequently, multicast is no longer sufficient to limit the amount of generated traffic in the network. In order to prevent an explosive growth in traffic, caches can be strategically placed throughout the content delivery infrastructure. As the size of caches is usually limited to only a small fraction of the total size of all content items, it is important to accurately predict future content popularity. Traditional caching strategies only take into account the past when deciding what content to cache. Recently, a trend towards novel strategies that actually try to predict future content popularity has arisen. In this paper, we ascertain the viability of using popularity prediction in realistic multimedia content caching scenarios. The proposed generic popularity prediction algorithm is capable of predicting future content popularity, independent of specific content and service characteristics. Additionally, a novel cache replacement strategy, which employs the popularity prediction algorithm when making its decisions, is introduced. A detailed evaluation, based on simulation results using trace files from an actual deployed Video on Demand service, was performed. The evaluation results are used to determine the merits of popularity-based caching compared to traditional strategies. Additionally, the synergy between several parameters, such as cache size and prediction window, is investigated. Results show that the proposed prediction-based caching strategy has the potential to significantly outperform state-of-the-art traditional strategies. Specifically, the evaluated Video on Demand scenario showed a performance increase of up to 20% in terms of cache hit rate.

1. Introduction

The proliferation of interactive, personalized and on-demand television services is causing an increasing need for bandwidth in telecom operator networks. Obviously, broadcasting or multicasting cannot sufficiently reduce bandwidth consumption of on-demand multimedia services. Proxy caching, which had already been widely employed in the delivery of web content, has been proposed as a way of offloading bottleneck links (Ma et al., 2011) in on-demand scenarios. Caches are strategically placed throughout the network and store a subset of the available content. However, the size of such caches is usually limited, so they are only capable of storing a fraction of available content. Therefore, it is very important to accurately predict the future popularity of content, so that the most popular items, or item segments, can be offered closer to the end-users.

Over the years, many caching strategies have been proposed. Traditional strategies, such as least recently used (LRU) and least frequently used (LFU), assume that what was most popular in the past, will also be most popular in the future. However, the popularity of multimedia content is known to be highly dynamic (Devleeschauwer and Laevens, 2009). Consequently, caching efficiency can be further increased by taking these dynamics into account and actually try to predict future popularity instead of directly applying historical information.

Predicting the future popularity of individual multimedia content items can be reduced to a time series prediction problem (Makridakis, 1994). Several efforts have been made to apply this theory to the prediction of multimedia content popularity (Avramova et al., 2009; Szabo and Huberman, 2010). However, to our knowledge, these predictions have never been integrated into an actual cache replacement strategy. Additionally, the effect of important parameters, such as the prediction window size, has not yet been thoroughly evaluated.

This paper presents a generic popularity prediction algorithm. In contrast to existing algorithms, it is not tailored to a specific service. The prediction algorithm fits a set of functions to the cumulative request pattern of content items. These fitted curves are then used as an approximated model of the request pattern. Through extrapolation the future of the request pattern can then be estimated. Subsequently, we present a novel prediction-based

* Corresponding author. Tel.: +32 93314938; fax: +32 93314899.
E-mail addresses: jeroen.famaey@intec.ugent.be (J. Famaey), tim.wauters@intec.ugent.be (T. Wauters), filip.deturck@intec.ugent.be (F. De Turck).

1084-8045/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jnca.2012.08.014

cache replacement strategy. It uses the predicted request patterns to determine the subset of all available content to store in the cache. Additionally, to assess the theoretical maximal gain in caching efficiency that can be achieved using predictions, a theoretical variant is also presented. It assumes the future can be perfectly predicted.

The proposed cache replacement strategies are thoroughly evaluated and compared to traditional strategies that directly employ historical information. The goal of this evaluation is to determine both the theoretical and practical gain in caching efficiency that can be achieved using popularity prediction. Moreover, the effect of the prediction window parameter is assessed. This parameter is defined as the future time-frame that is predicted (i.e., the counterpart of the history window parameter of LFU). The effect of this parameter is influenced by the cache size. Therefore, the synergy between these parameters is thoroughly evaluated. In order to increase the applicability and validity of the presented results, all evaluations are performed using a trace of an actual deployed Video on Demand (VoD) service of a leading European telecom operator. This gives our evaluations more leverage and credibility than those performed on synthetically generated datasets. The ultimate goal of this study is to show that popularity prediction indeed improves caching efficiency and to determine under what circumstances it achieves the most optimal result.

The remainder of this paper is structured as follows. Section 2 gives a more in depth description of existing work on popularity prediction of multimedia content. Section 3 presents our proposed generic popularity prediction algorithm and cache replacement strategy that uses it. Subsequently, Section 4 evaluates the proposed cache replacement strategy using simulation results. Finally, the paper is concluded in Section 5.

2. Related work

The large size and stringent sequential delivery demands of multimedia content have caused a push towards novel caching strategies. Traditional caching strategies have been adapted to operate on individual content segments instead of entire items (Wu et al., 2004; Chen et al., 2005). This allows the caches to better utilize the sequential nature of multimedia content demand patterns. Additionally, such techniques better map to the skewed internal popularity of multimedia content. Yu et al. (2006) argue that selecting a suitable segment size is a complex problem and therefore propose an alternative solution that models the internal popularity of multimedia streams independent of segment size. Kim and Das (2007) further extended the work on segment-based caching, by way of an analytical model that exploits the temporal locality and popularity of content. Guo et al. (2007) studied a prefix caching method that exploits the fact that the beginning of a movie is more popular than the ending. Their algorithm automatically determines the optimal number of segments from the beginning of each movie that should be cached, based on the internal popularity distribution of the movie. Certain IP-TV services have specific properties that can be exploited by caching strategies. For example, the use of sliding-window caches has been proposed in the context of time-shifted TV services (Wauters et al., 2008). In line with our work, these techniques aim to improve caching efficiency. Nevertheless, they focus on a different aspect, which falls outside the scope of this paper.

In the field of time series prediction, a wide range of techniques have been developed for forecasting all sorts of time series. Recently, machine learning techniques, such as support vector machines and artificial neural networks have been applied to this problem (Verstraeten et al., 2007; Samsundin et al., 2010). Recently, Wyffels and Schrauwen (2010) have used reservoir computing, a form of recurrent neural networks, for time series prediction. Additionally, time series often exhibit repeating trends and periodical effects. For example, multimedia content request patterns often show repeating effects on a daily and weekly basis. The use of wavelet decomposition has been proposed to decompose time series into signals with dynamics in different scales. This has been shown to simplify prediction with neural network based techniques (Soltani, 2002). This approach was also successfully combined with reservoir computing (Wyffels et al., 2007).

Recently, several studies have been conducted on modeling the popularity of multimedia content. These studies can be split into two types. A first type focuses on characterizing the popularity distribution among different multimedia objects, while the second type focuses on modeling the popularity evolution of individual multimedia files.

A popularity distribution among multiple multimedia objects models the static popularity relationship between the content items offered by a multimedia service. It can be used to derive the probability that the content item with a specific popularity index (e.g., the *X*th most popular item) will be requested. Many models have been proposed for modeling the popularity distribution of a multimedia service, including Zipf (Breslau et al., 1998), Zipf–Mandelbrot (Tang et al., 2007), stretched exponential (Guo et al., 2008), Zipf with exponential cut-off tail (Cha et al., 2007), power-law with exponential cut-off tail (Cheng et al., 2007), log-logistic (Abhari and Soraya, 2010) and Weibull (Abhari and Soraya, 2010).

Other research has focused on modeling the dynamic popularity evolution of individual content items. It thus allows the request evolution of content to be estimated, based on historical request information. This latter type of research is also the focus of our work. Most work on this topic was performed in the context of video-sharing services such as YouTube. Cha et al. (2007, 2009) found that there is a strong correlation between the popularity of a video after two days and after 90 days. These observations were supported by a study performed by Szabo and Huberman (2010). More recently, Chatzopoulou et al. (2010) studied the correlation between popularity and a wider range of metrics. They found that the popularity of a video is highly correlated with the amount of posted comments, ratings and favorites. Figueiredo et al. (2011) characterized the popularity evolution of YouTube videos. They found that several different categories exist that exhibit distinct popularity evolutions. For example, copyright protected videos got most of their views early in their lifetime. Additionally, they identified and quantified the main referrers that lead users to videos, as they are key mechanisms in attracting users and thus highly influence popularity evolutions. An alternative approach was proposed by Avramova et al. (2009). They found that YouTube video popularity traces follow several different distributions, such as power-law or exponential. An analytical model is devised that predicts the distribution associated with specific popularity traces. Jamali and Rangwala (2009) studied the evolution of popularity on the social news website Digg. Based on comment data and the co-participation network, they are able to accurately predict the future popularity of any news item shortly after it has been posted. In the context of VoD services, Devleeschauwer and Laevens (2009) propose a prediction method based on a generic user-demand model derived from traces of VoD and catch-up TV services. Wu et al. (2010) adapted the previously mentioned reservoir computing approach to the popularity prediction of multimedia content. Niu et al. (2011) employ time-series analysis techniques to predict future content popularity, online population, peer upload and server bandwidth consumption in peer-to-peer