Mobile Netw Appl (2010) 15:237–252
DOI 10.1007/s11036-009-0186-1

Resource Management Strategies for the Mobile Web

Claudia Canali · Michele Colajanni · Riccardo Lancellotti

Published online: 7 July 2009

Abstract The success of the Mobile Web is driven by the combination of novel Web-based services with the diffusion of advanced mobile devices that require personalization, location-awareness and content adaptation. The evolutionary trend of the Mobile Web workload places unprecedented strains on the server infrastructure of the content provider at the level of computational and storage capacity, to the extent that the technological improvements at the server and client level may be insufficient to face some resource requirements of the future Mobile Web scenario. This paper presents a twofold contribution. We identify some performance bottlenecks that can limit the performance of future Mobile Web, and we propose and evaluate novel resource management strategies. They aim to address computational requirements through a pre-adaptation of the most popular resources even in the presence of irregular access patterns and short resource lifespan that will characterize the future Mobile Web. We investigate a large space of alternative workload scenarios. Our analysis allows to identify when the proposed resource management strategies are able to satisfy the computational requirements of future Mobile Web, and even some conditions where further research is necessary.

C. Canali (✉) · M. Colajanni · R. Lancellotti
Department of Information Engineering,
University of Modena and Reggio Emilia,
Modena, Italy
e-mail: claudia.canali@unimore.it

M. Colajanni
e-mail: michele.colajanni@unimore.it

R. Lancellotti
e-mail: riccardo.lancellotti@unimore.it

Keywords Mobile Web · multimedia resources · content adaptation · performance evaluation · server infrastructure

1 Introduction

The advent of mobility is significantly changing the Web scenario: a continuously increasing number of users will produce and require any type of content at any time, from any location and through any class of devices. According to recent research reports, the mobile browsing will grow over 900% by 2014 [18]. We should consider that other major changes are occurring in the Web in terms of novel forms of traffic and services. For example, photo and video sharing services (e.g., YouTube, Flickr) are causing an explosion of demand for multimedia content. These trends will determine a future Mobile Web scenario characterized by a large amount of heterogeneous contents, mainly consisting of multimedia resources (e.g., [4, 9, 18]), that will have to be tailored on-the-fly to user preferences and device capabilities at the moment of the client request [7, 17, 32].

By exploiting some evolutionary trends [9], we consider the workload characteristics of the next years and we evaluate whether or not the technological improvements at the server and client level can support the requirements of future Mobile Web in terms of computational and storage capacity. We identify some possible bottlenecks of the server infrastructure that are likely to represent a challenge for the deployment of future Mobile Web-based services. Performance issues can be addressed at the architectural and/or data management level, but in this paper we focus on resource management strategies for the future Mobile Web. In

particular, we consider the most expensive operations that is, adaptation of multimedia resources, and we exploit offline pre-adaptation strategies on the subset of the most popular resources. This idea has been applied in other contexts (e.g., [6, 23]) where, as in the case of the Mobile Web, resource popularity follows a skewed Zipf-like distribution [10, 38]. However, the effectiveness of pre-adaptation strategies strongly depends on their accuracy in estimating which are the most popular resources. We should consider that in the future Mobile Web any estimation about resource popularity will become a real challenge because of the following workload characteristics: heterogeneous resource composition and size; novel access patterns, including user upload operations; workload intensity characterized by different orders of magnitude; short resource lifespan and fast changes in resource popularity.

The workload dynamics of future Mobile Web make useless most available mechanisms for resource popularity estimation that are based on quite different workload characteristics [25, 29]. For this reason, we propose a novel class of algorithms that estimates the resource popularity by adopting different predictive techniques. Through extensive simulations we determine under which scenarios and conditions a pre-adaptation based on predictive algorithms is effective to guarantee adequate performance. We demonstrate that our proposal represents a promising solution to support future scenarios of Mobile Web thanks to the ability of the predictive algorithms to face highly dynamic characteristics of future workloads, short resource lifespan, and continuous upload of novel resources. Our study also identifies the limits of the proposed pre-adaptation strategy. When the computational demand is extremely high, pre-adaptation based on predictive algorithms does not guarantee satisfactory performance unless the resource popularity is highly skewed. This limitation opens new spaces of future investigation, such as the possibility to integrate other forms of content management strategies with data and server replication. These extensions are beyond the scope of this paper.

The remainder of this paper is organized as follows. Section 2 describes the services for the Mobile Web of interest for this paper, their expected evolution in the next future, and the impact of such evolution on the future server infrastructures. Section 3 considers resource management strategies based on content pre-adaptation. Section 4 evaluates the performance of the proposed solutions. Section 5 discusses the related work. Section 6 concludes the paper with some remarks and open issues.

2 Evolution of services and systems

2.1 Services of the Mobile Web

For the analysis of Mobile Web evolution and for the evaluation of possible system bottlenecks, we consider two classes of relevant sites [8]: *online-news* and *social-multimedia* sites that will be accessed through mobile Web-enabled devices. These sites are expected to be among the most popular and to provide the most challenging services from a server point of view.

The *online-news* category includes information portals, such as online newspapers and news broadcasting sites, that offer online information including events, stock quotes, and sports results. These sites typically deliver news in the form of textual resources and images. Text accounts for almost 60% of the requests, while requests for images represent the 35%. Furthermore, there is a growing tendency to deliver also audio and video content, although today it is limited to 5% of the resources (www.stateofthenewsmedia.org/2007).

The *social-multimedia* category includes sites representing a new form of user communication and interactivity that is a qualifying characteristic of the so called Web 2.0. Typical examples are represented by forums, blogs and content sharing sites where users exchange opinions, stories and files (e.g., MySpace, Flickr, Youtube). In this category of sites about half of the resources is textual and half is multimedia as many user communications involve exchanges of images, audio or video files [8, 10].

Each user request may involve upload or download of resources. Users upload novel contents in the form of comments to news, articles, polls, and even of multimedia resources including images and videos. The service of upload operations is and will remain a network-bound task especially if we consider mobile users. Content providers supporting upload services only need to receive the user supplied information, store it in some disk cache and then insert it in a database.

Download operations are more heterogeneous and will remain more critical from the server point of view, even because much more numerous. They may involve a simple retrieval of some static files, but also a dynamic generation of the content that is becoming quite common in most Web-based services, up to onerous operations for the resource adaptation to user preferences and/or client device capabilities. Downloading a resource without a dynamic generation or adaptation will be usually characterized by a service time in the order of few milliseconds, where the real time depends

on the resource size and on the I/O bandwidth of the server subsystem. On the other hand, the service time of a request involving some form of processing depends heavily on the resource content type, that may be textual (e.g., HTML) or multimedia (e.g., images, audio, video). The most common type of textual resource is dynamically generated on the basis of content stored in database(s), where page layout, insertion of banners and proposed links for navigation are derived from the user profile. The typical service time for the on-the-fly generation of a textual resource is in the order of hundreds of milliseconds, while the most expensive operations, such as data mining techniques to associate recommendations to specific user preferences, are carried out offline and do not contribute to the service time [9, 35].

Multimedia content is retrieved from some storage repository and its service time does not represent an issue if it has to be delivered as it is. The problems arise if this multimedia content has to be adapted to match the characteristics of the mobile devices in terms of computational power, rendering capabilities and network connection. To this end, we should consider the expected evolution of the mobile device capabilities. For example, future mobile devices are expected to experience an increase in processing power and storage space that would allow them to consume larger resources and perform some adaptation at the client side. On the other hand, other characteristics, such as display size and battery power, are likely to experience minor improvements. This client technology evolution has a positive effect on the server side because resources will not have to be tailored for every type of client device. Although we can expect that different devices will be able to consume similar versions of a multimedia resource with possible local adjustments, the client evolution does not avoid the need for server-side adaptation. Limitations on battery power and wireless bandwidth prevent transmission of large resources and adaptation at the client-side only.

Adaptations that are commonly performed on multimedia content involve transformations such as scaling, cropping and color reduction for images [30] or recoding at a different bit rate for audio and video [9]. Content adaptation may also be driven by specific settings of the user preferences: for example, a color-blind user may require specific enhancements on images and videos to match his/her impaired vision [22]. The typical service time for online adaptation of a multimedia resource depends on the resource size. It may reach up to one or more seconds when complex operations are carried out on audio and video clips of several megabytes [7, 12]. For download of audio and video resources, we consider a play-while-downloading approach, that is gaining popularity over the more traditional download-and-play approach [21]. In this case we consider that download and adaptation of the resource occur in a chunk-by-chunk way. Furthermore, we should take into account services related to the resource upload that determines and will determine even more the fast popularity changes that reduce the resource lifespan of the future workload. We refer to the social-multimedia category of sites, that represents the major challenge, and we consider that 5% of the user interactions with the server involve some content upload [16].

When we refer to the service times of the future Mobile Web, we should consider the technological improvements of the server technology in the next 5 years in terms of computational power and storage capacity. We can assume that no disruptive technology will appear and that the server CPU will continue to improve according to the Moore Law that is, the computational power doubles every 18 months. On the basis of this assumption, we may expect that the service times for the future Mobile-based services will be reduced by a factor of 8 with respect to the current values. Table 1 outlines current and future service times for the considered services in the context of Mobile Web. Although each parameter is supported by some experimental evidence or some literature result, the reported data should be interpreted in their order of magnitude and not as precise values. We recall that the service time for the generation of a textual resource includes an interaction with one or multiple databases and a generation of HTML code (on the basis or not of some user preferences). The adaptation time for multimedia resource includes the retrieval of the original resource from some storage device and its adaptation on the basis of user preferences and/or client device characteristics.

2.2 Expected trend of the workload characteristics

In this section we present an analysis about the evolutionary trends of the workload of the Mobile Web for the next 5 years (2009–2014). The analysis is extrapolated from a previous study of the same authors [9], in which we consider the workload characteristics of the future Mobile Web in terms of *workload composition* and *workload intensity*. We anticipate that the 5 years trends concerning the workload composition

Table 1 Service times (median)

	Current value	Future value	References
Download of a static resource	10 ms	5 ms	[9]
Upload of a static resource	20 ms	10 ms	[9]
Textual resources generation	220 ms	27.5 ms	[9]
Image adaptation	730 ms per MB	91.3 ms per MB	[7]
Audio/video adaptation	1054 ms per MB	131.8 ms per MB	[12]

are defined rather well in literature, while the assumptions regarding the growth of the workload intensity are less clearly defined, hence we consider a scenario of low growth and one of high growth.

Table 2 reports the workload composition for the two classes of *online-news* and *social-multimedia* sites over the next 5 years. We consider the *workload mix*, that is the mix of the resource types, and the *resource sizes*. The workload mix for both sites will be characterized by an increasing amount of multimedia content, especially video and audio resources [9, 18]. This trend implies more adaptation services and a consequent increase of the computational demand.

Besides the increase in the amount of multimedia resources in future workloads, we must also consider that the resource size is likely to increase in the next 5 years. We focus our analysis on multimedia resources because their size affects the computational cost of the adaptation [12], while the size of textual resources is not correlated to the costs of content generation. The median resource size is expected to grow per year of about 12% for images and 16% for audio and video resources [9]. As regard multimedia video resources, we should also consider the advent of content in High Definition (HD) quality. Some popular Web sites belonging to the online-news and social-multimedia classes, such as YouTube, Vimeo, CNN, have recently introduced the support for HD video, and the presence of this type of resource is likely to grow in the next years as in other contexts, such as video-on-demand services [13]. We assume that the percentage of video resource in HD quality is expected to grow in the next 5 years up to 13% and 20% for online-news and social-multimedia, respectively. For this reason, for online-news and social-multimedia Web sites we consider two possible values for future median size of the video resources, as shown in Table 2: a lower value for a scenario that does not consider the presence of HD content and a higher value that accounts for a percentage of video in HD quality. Finally, the last row of the table reports the number of adapted versions of each resource that are necessary to satisfy the mobile device requirements. The main difference between the current and the future scenario is due to the evolution of the mobile devices towards more powerful devices (see Section 2.1). We should also consider that for some multimedia resource such as videos, we could generate only one adapted version of the content by means of Scalable Video Codecs (SVC format), from which it is possible to have a suitable version for any client device [36].

The workload intensity denotes the frequency of client requests to the Mobile Web sites in a defined interval, that is typically referred to one second. Determining the exact evolution of the workload intensity is not straightforward: while we have information about the growth of client population [18], no information is available on the evolution of the user behavior (e.g., how many users will actually exploit the features of their client devices, the duration of their sessions, and the request frequency of each user), that may contribute to the growth of the workload intensity [9]. We consider three future scenarios for each Web site category:

– *Low growth future scenario*
– *High growth future scenario*
– *HD growth future scenario*

The low growth scenario follows a more conservative approach under the assumption of a moderate growth

Table 2 Workload composition

	Online-news class		Social-multimedia class		References
	Current	Future	Current	Future	
Percentage of textual resources	60%	49%	54%	41%	[9, 37]
Percentage of image resources	35%	38%	38%	40%	[21, 37]
Median image size	900 KB	1.6 MB	900 KB	1.6 MB	[11]
Median audio size	3 MB	6 MB	3 MB	6 MB	[21]
Median video size—no HD	8 MB	17 MB	8 MB	17 MB	[21]
Median video size—with HD	8 MB	22 MB	8 MB	26 MB	[13]
Number of resource versions	10	1–5	10	1–5	[7, 9]

of workload intensity for the two Web sites categories, while the high growth and HD growth scenarios assume a more challenging increase of the Mobile Web workload intensity.

The increments of the workload intensity for the three considered scenarios are reported in Table 3 and motivated below. The future low growth scenario for the online-news sites assumes an increase of 20% per year of the workload intensity that is mainly related to the presence of mobile users [18], while traditional Web accesses are not expected to contribute to a growth of the workload intensity. For the future high growth and HD growth scenarios, we consider that the workload intensity increases by 35% per year as for the most popular sites (e.g., CNN, MSNBC, Google News, Yahoo!—www.stateofthenewsmedia.org/2007). Social-multimedia sites have achieved a more recent popularity and the number of mobile users accessing these sites is expected to grow extremely fast in the next 5 years. We assume that for the future high growth and HD growth scenarios the requests will augment at a rate of about 55% per year [24, 26]. Other studies are more conservative and expect a fast increment in the next 2 years and a stabilization afterwards. This trend is represented by the future low growth scenario where the workload intensity is expected to augment about 40% per year.

2.3 Storage technology

It is clear that the future scenarios will place a strain on the server computational capacity, and it is interesting to evaluate the impact on storage devices. Here two opposite forces will emerge.

The requirement for storage space is expected to grow for two reasons. As pointed out in Section 2.2, we expect that the resource size can increase up to 16% per year in the next 5 years. Moreover, the presence of heterogeneous clients will require multiple versions of the same resource that may be pre-adapted or cached in some adapted version. Up to now, a version of each resource for every type of client device must be available and this would cause a severe strain on storage.

On the other hand, the impact on storage will be reduced by other two factors. Future client devices will be more powerful in terms of computational and memory capacity and they will be able to accept a broader range of contents [9]. This will reduce the number of adapted versions of each resource to 5 and for some video resources encoded through SVC even to 1 copy. Moreover, the technological evolution guarantees a fast growth of storage space: the disk density has increased from 60% to 100% every 12 months during the last decade [20] and we may assume that the storage capacity will follow the same pattern in the next 5 years. Even in the hypothesis that the working set of the Mobile Web will be increased by the combination of larges multimedia resources stored in multiple versions, we can easily assume that the storage capacity will not represent a real issue for the future Mobile Web-based services. We expect that the working set size explosion will be limited more by the costs for guaranteeing consistency among multiple versions of each resource than by technological limits at the level of storage capacity. These reasons motivate our focus on computational issues.

2.4 Performance impact of Mobile Web evolution

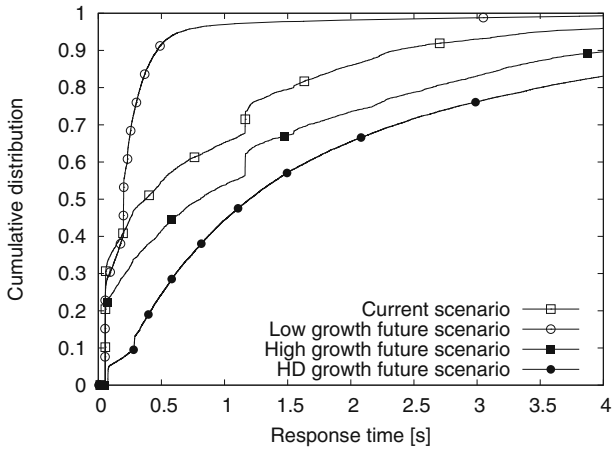
In order to evaluate whether, and to which extent, the future server infrastructure can support the evolution of Mobile Web-based services, we need to consider the aggregate effects of the workload trends and the concurrent improvements of the server technology. The goal is to understand under which conditions the computational requirements of the future services may overcome the capabilities of the next generation of server infrastructures. As we consider future load and technological scenarios, with heavy-tailed distributions, interdependency and non-linearity in the system models (memory, CPU, network), we have to refer to a simulation study. To this purpose, we use a discrete event simulator based on the Omnet++ framework [27].

For the experiments, we consider the workload represented by the *current* scenario, the *low growth*, the *high growth* and the *HD growth* future scenarios described in Section 2.2 and related Tables 1, 2, and 3.

The content provider system includes a server that receives and processes requests issued from multiple mobile clients, a data repository server and a server for on-the-fly adaptation. Client requests are processed

Table 3 Workload intensity

	Current	Future		References
		(Low growth)	(High/HD growth)	
Request rate (Online-news)	12 req/s	30 req/s	64 req/s	[9, 18]
Request rate (Social-multimedia)	12 req/s	54 req/s	106 req/s	[9, 24, 26]
Upload percentage (on total sessions)	2%	5%	5%	[16]



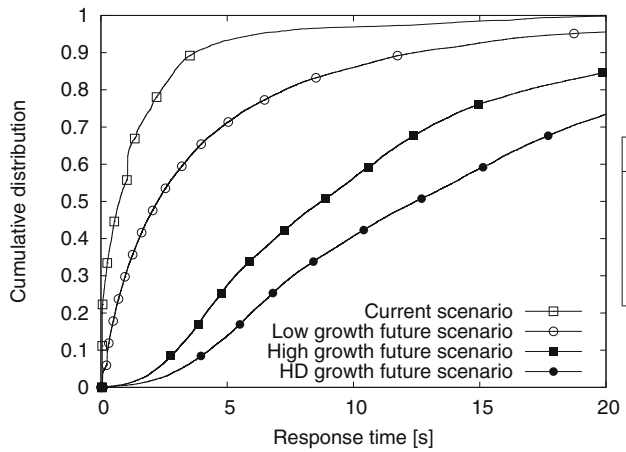
Scenario	CPU utilization
Current	0.45
Low growth future	0.43
High growth future	0.81
HD growth future	0.89

Fig. 1 Expected evolution of the response time and CPU utilization for online-news Web sites

in parallel. After entering the system, each request is assigned to a thread and a service time is determined for each required resource. The CPU is shared among all the active threads through a round-robin CPU scheduler. The service time accounts for the resource generation/adaption tasks and depends on the type of resource: textual, image or audio/video. Table 1 shows the considered service times for each resource type.

The performance evaluation considers two main metrics: the *response time* at the server side and the *CPU utilization* during the entire experiment. For textual and image resources, the response time is the time elapsed between the arrival of the client request, the retrieval or generation of the file(s) and the dispatch of the last byte of the response flow. Audio and video resources are delivered through HTTP pseudo-streaming, hence the response time is measured when the system ends to serve the first chunk of the resource (1.5MB represents a common buffer size for media players supporting HTTP streaming). Even content is adapted on a chunk-by-chunk basis of 1.5MB. We can model the play-while-downloading behavior typical of the widely used HTTP pseudo-streaming [21], but we should consider that the system continues to work also on the other chunks that contribute to the server load even if they are not included in the response times.

Figure 1 shows the cumulative distribution of the response time for the online-news sites in the four considered scenarios: current, low growth future, high growth future and HD growth future. The future scenarios have quite different impact on performance. In the low growth scenario, the more powerful CPU of the future servers can satisfy a higher number of requests and can guarantee even a lower response time than that of the current scenario. On the other hand, in the high growth and HD growth future scenarios, the



Scenario	CPU utilization
Current	0.57
Low growth future	0.85
High growth future	0.92
HD growth future	0.99

Fig. 2 Expected evolution of the response time and CPU utilization for social-multimedia Web sites

technological improvements cannot counterbalance the increase of computational requirements of the Mobile Web to the extent that the performance can become really critical. These results are confirmed by the average CPU utilization that for the high growth and HD growth future scenarios is close to saturation.

Figure 2 shows analogous results for the social-multimedia sites. In this case, all future scenarios are affected by a severe performance degradation. We can conclude that the on-the-fly adaptation of all contents is not a feasible alternative to support all future services for the Mobile Web.

3 Resource management strategies

The analysis of the evolutionary trends of the Mobile Web evidences severe performance problems for the future server infrastructures. These issues can be addressed through architectural solutions that replicate on a local or a geographical scale the servers and the resources and/or through resource management solutions that aim to reduce the service time experienced by each client request. This paper is focused on the latter alternative, although it should be clear that architectural and management approaches can be combined.

3.1 Pre-adaptation strategies

There is a wide space of possible resource management solutions, ranging from on-the-fly adaptation of every requested resource to offline pre-adaptation of all resource versions. In this paper, we focus on a resource management strategy based on an offline pre-adaptation of a "suitable" subset of the resource working set. The choice of this resource management strategy is motivated below, but it is important to evidence the research challenge related to the identification of the "suitable" subset in the context of future Mobile Web scenarios. Wrong decisions may cause waste of computational and storage resources, while right choices can half the response times.

Let us distinguish textual from multimedia resources. The generation of personalized textual resources is not critical for performance because the most computationally expensive tasks, such as data mining to gather user preferences, are carried out offline. Hence, if we want to reduce the computational cost of the Mobile Web-based services, we should avoid whenever possible on-the-fly adaptation of multimedia resources that are the most expensive operations.

A straightforward solution is to pre-adapt offline all multimedia resources for any class of device/connection. This approach is theoretically feasible from the point of view of the storage investment (see Section 2.3). Moreover, for some resources the adoption of techniques such as Scalable Video Codecs (SVC) would further reduce the storage requirement because every adapted version can be stored in one file. However, an offline adaptation of every multimedia content is unfeasible because it would impose a computational load similar to an on-the-fly adaptation due to the continuous upload of new resources that is typical of Mobile Web. Even the use of SVC cannot address this issue because original content is seldom available in an SVC format, hence an effort of pre-adaptation is required whenever a resource is added. Furthermore, a complete offline content adaptation would be not convenient due to consistency issues in a volatile context. Indeed, the increase of the working set size that is expected in the next future [4, 10] would require pre-adaptation for a huge amount of resources for online-news and social-multimedia workloads that are highly volatile. The resources, that may be uploaded at any time even by the users, are typically characterized by a short lifespan because they usually concern real-world events or popular hot topics for which the user interest rapidly subsides, thus determining sudden but short popularity surges. When multiple versions of each content are generated, or when the resources are replicated across a distributed infrastructure, the costs to maintain consistency among the pre-adapted content increases, up to the point where the cost of pre-adaptation and consistency enforcement overweight the benefits.

In this paper, we suggest an approach alternative to a total pre-adaptation, where this strategy is applied only to a limited subset of the working set, that is that corresponding to the most popular multimedia resources. The motivation for this proposal lies in the popularity of multimedia resources that is characterized by a Zipf-like distribution, with a skewness that is expected to be even higher in the Mobile Web than in the traditional Web [10, 38]: while the Zipf α parameter of the popularity distribution is in the range between 0.68 and 0.84 for the traditional Web, it appears to be between 0.84 and 1 for the Mobile Web. These characteristics allow a system for Mobile Web-based services to satisfy a high number of requests by pre-generating only a small fraction of popular resources. However, to identify the resources that will receive more requests in the near future is an open challenge especially in the context of the Mobile Web workload that is characterized by high volatility, download and upload of content, short

resource lifespan and sudden popularity peaks. The proposed resource management strategy relies on algorithms for resource popularity estimation that explicitly addresses the challenges related to the future Mobile Web workload.

3.2 Problem model

The identification of the most popular resources depends on the ability to predict the future accesses to each resource by means of information available at the server side.

We consider that resource pre-adaptation is a periodic task with period Δt that relies on a run-time evaluation of resource popularity. Let $R(t)$ be the working set of resources that can be required at time t. The goal of the algorithms for the identification of the most popular resources is to identify the subset $PR(t)$ containing the resources that are expected to receive the highest number of accesses in the future interval $[t, t + \Delta t]$. A typical algorithm estimates the popularity $p_r(t)$ for each resource $r \in R(t)$ in terms of number of expected accesses in the future interval $[t, t + \Delta t]$. Next, the resources are sorted according to their popularity $p_r(t)$ to determine the set $PR(t)$ of the most popular resources at the time t.

The problem of resource popularity estimation in the context of the Mobile Web requires novel algorithms that are able to address the issues of future workload characteristics. In particular, the set of the most popular resources will change quickly due to a twofold reason. On one hand, resources are likely to be frequently requested just in a short span of time, after which fewer people will access them [10]. This short lifespan leads to changes in the resource popularity during the period $[t, t + \Delta t]$. On the other hand, the continuous upload of newly generated resources in the period $[t, t + \Delta t]$ determines a growth of the whole working set such that $R(t) \subset R(t + \Delta t)$. Consequently, it is necessary to identify a new set of popular resources $PR(t + \Delta t)$. In this context, an algorithm for resource popularity estimation that is based on a simple observation of the aggregate number of accesses in a past interval [25, 29] does not offer a sufficiently reactive mechanism to detect sudden popularity changes. This motivates our proposal of predictive models that use information about the past accesses not only as a plain information but through some simple statistical elaboration.

Our algorithms exploit a model of the past accesses to each resource r based on time series that capture access pattern variations occurring at different time points. For example, let us consider the time series $D^r = \{d^r_t, d^r_{t-\Delta t}, \ldots, d^r_{t-(n-1)\Delta t}\}$, where d^r_t is the number of accesses to the resource r in the current time interval $[t - \Delta t, t]$, $d^r_{t-\Delta t}$ is the number of accesses in the interval $[t - 2\Delta t, t - \Delta t]$, and so on until $d^r_{t-(n-1)\Delta t}$ that is the last element of the time series. The predictive algorithms use the time series D^r to forecast the number of accesses $\hat{d}^r_{t+\Delta t}$ in the future time interval. The expected number of accesses $\hat{d}^r_{t+\Delta t}$ is used as the popularity metric for the resource r, that is $p_r(t)$.

3.3 Algorithms for the identification of the most popular resources

3.3.1 Predictive-EWMA algorithm

The Predictive-EWMA algorithm is a novel proposal that aims to estimate the number of accesses in the next future as a mean to identify the subset $PR(t)$.

We first consider that prediction is based on the Exponential Weighted Moving Average (EWMA) model [31]. This algorithm applies weighting factors that decrease exponentially for older data points, thus giving more importance to recent observations while still not discarding older observations. For each resource r, the prediction of the future value of the number of accesses is given by:

$$p_r(t) = \hat{d}^r_{t+\Delta t} = \gamma \hat{d}^r_t + (1 - \gamma) d^r_t$$

The considered algorithm begins with $\gamma = \frac{2}{n+1}$, where $n + 1$ is the length of the time series, which represents a typical choice for EWMA-based prediction. This algorithm is characterized by a low computational cost for prediction, that is an important feature for run-time contexts and huge numbers of resources. Furthermore, the adopted predictive technique is suitable for the Mobile Web, characterized by highly variable and dynamic scenarios, due to the great simplicity in the parameters' choice and robustness with respect to workload characteristics, oppositely to other approaches that may require long training time (e.g., neural networks [34]) or complex parametrization (e.g., Kalman filters and ARIMA models [3]).

We carried out some preliminary experiments on typical access patterns of Mobile Web-based services, characterized by popularity surges and burst of requests, with the aim of determining which is the minimum number n of values in the time series that is necessary to consider for our prediction. Figure 3 shows how the autocorrelation of the time series values varies for different n. From the histogram we observe that a time series of 10 or more elements (that is, $n \geq 10$) guarantees a high auto-correlation (> 0.8) for a representative experimental scenario, while using less

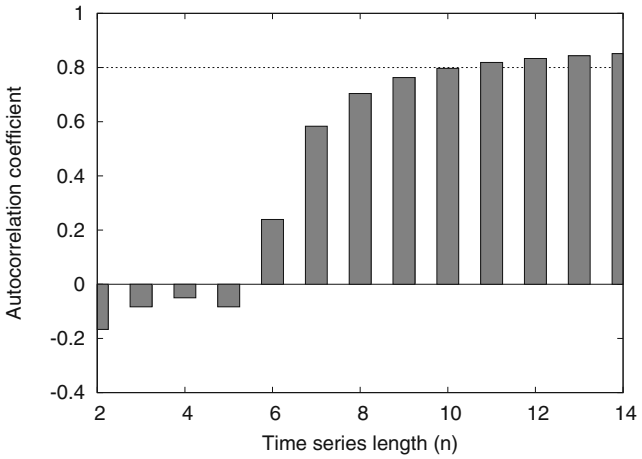


Fig. 3 Autocorrelation with respect to n

values does not guarantee sufficient correlation and predictions risk to be less accurate.

3.3.2 Predictive linear regression algorithm

The Predictive Linear Regression algorithm (Predictive-LR) estimates the number of accesses in the future by considering just the last value of the time series, as following [2]:

$$p_r(t) = \hat{d}^r_{t+\Delta t} = \alpha^r_t d^r_t + \beta^r_t, \tag{1}$$

where the coefficients α^r_t and β^r_t are dynamically chosen with the goal of minimizing the mean quadratic deviation over the time series D^r, that is:

$$\sum_{\tau=t}^{t-(n-1)\Delta t} [\hat{d}^r_\tau - d^r_\tau]^2 \tag{2}$$

The simplicity of the Predictive-LR algorithm guarantees a low computational cost. Its prediction quality is good when the data set is subject to long- or medium-term variations. On the other hand, when the data set is characterized by very short-term variations, this model tends to overestimate the changes of the time series with a possible degradation of the prediction quality.

3.3.3 Traditional algorithm

For the sake of comparing the performance of the novel predictive algorithms with other popular solutions, we consider an example of *Traditional* algorithm that estimates the resource popularity on the basis of past values of D^r with no elaboration. This approach is consistent with current techniques for resource management where the resource popularity is determined on the basis of metrics such as absolute number, frequency, or freshness of past accesses [25, 29]. Although the exact policies for estimating the resource popularity in modern Internet-based services are industrial secrets, many algorithms appear to rely on the number of accesses received by the resources. For example, the Flickr Web site rates the resource popularity on the basis of how many times an image is viewed or commented.

The considered Traditional algorithm evaluates the popularity of a resource as the request frequency that is computed over the time interval since the previous pre-adaptation step [25], that is:

$$p_r(t) = d^r(t) \tag{3}$$

Unlike the two proposed predictive algorithms, the estimation of the popularity follows a short memory approach that does not consider the whole resource history. It has the advantage of identifying the recently uploaded resources that are rapidly gaining popularity. As a consequence, it guarantees a fair comparison especially in a context of highly dynamic workload and short resource lifespan characterizing the future Mobile Web. On the other hand, the lack of historical information about the past accesses may lead to an over-reaction that hinders the efficacy of this algorithm.

3.3.4 Ideal algorithm

We consider also an Ideal algorithm that exploits a perfect knowledge of the resource popularity distribution. Resource popularity is modeled at the client-side and follows a Zipf distribution. Every time the Ideal algorithm is invoked, it collects information about the top ranked resources from the client model and uses this information to build the list of the most popular resources. Since the algorithm accesses the same information that are used by the clients to issue their requests, it achieves the best performance and represents an ideal bound for every other algorithm. The Ideal algorithm is used only as a comparison, and cannot be applied in a real system because it exploits knowledge of data that are not really available at the server side.

3.4 Resource management operations for the Mobile Web

We now describe how client requests are serviced and how the popularity estimation algorithm is integrated with the resource management scheme.

Figure 4 outlines the main operations that are carried out by the server infrastructure supporting the Mobile Web. Some tasks, such as the management of client requests and on-the-fly resource adaptation, are carried

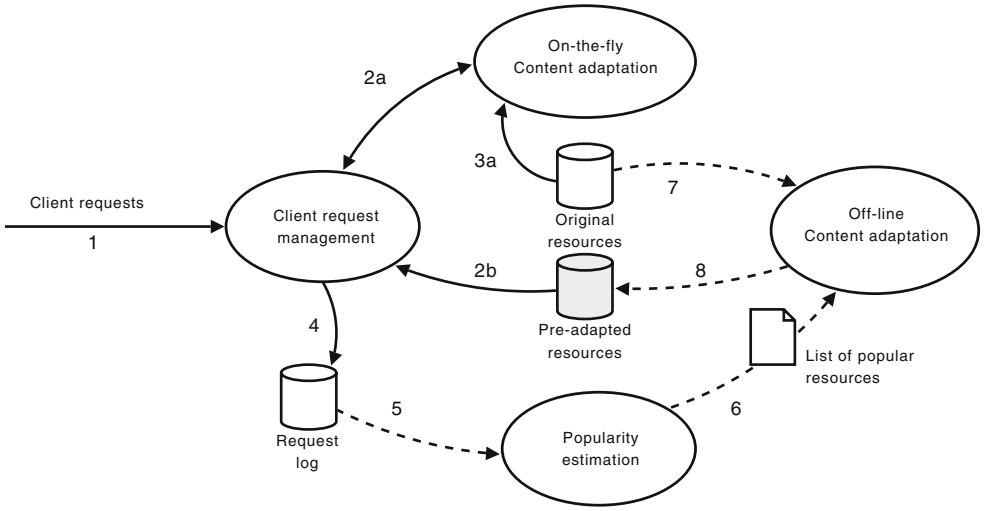


Fig. 4 Mobile Web-based service management through resource pre-generation

out as a response to client interaction, while other tasks, such as resource pre-adaptation and popularity estimation, occur offline.

When a server of the content provider receives a client request for a resource r in the time interval $[t, t + \Delta t]$ (Step 1 in Fig. 4), two possible cases may occur (identified by the letters a and b).

In the a case, we consider that the requested resource does not belong to the set of popular resources ($r \notin PR(t)$): after receiving the request, the client request management task must interact with a content adapter (step 2a) in order to generate on-the-fly the resource version that matches the client requirements. The content adapter accesses the repository of the original resources (Step 3a) to retrieve the content to be adapted.

In the b case, we suppose that the client request refers to a popular resource ($r \in PR(t)$). In this instance, the resource has already been pre-adapted: after receiving the client request, the management task consults the list of popular resources and determines that no on-the-fly operation is to be carried out. The available version of the resource is retrieved from the data storage and sent to the client (Step 2b).

In both instances, the client requests are stored in a request log file that is used for subsequent analyses on the resource popularity (Step 4). To this purpose, a periodic job with period Δt coordinates the operations related to the pre-adaptation of the resources that are expected to become the most popular. In our proposal we consider that the periodic job occurs every 20–30 min, to guarantee fast response to the changes in the workload characteristics. In order to pre-adapt popular resources, the system collects the access logs from every server and feeds the information to the popularity estimation task (Step 5 in Fig. 4—the solid lines represents interactions occurring as a response to a client request, while the dashed lines represent interactions occurring periodically and asynchronously with respect to client request-reply). The algorithms described in Section 3.3 are used to estimate the future popularity of each resource. The result is a list of popular resources (Step 6) that is passed to the offline content adaptation task. To this aim, the content adapter retrieves the original versions of the expected popular resources (Step 7), pre-adapt asynchronously the resource versions and save them in the repository of pre-adapted resources (Step 8). The new set of pre-adapted resources is used as a cache to serve the successive client requests.

4 Experimental results

In this section we evaluate to which extent and for which scenarios the proposed strategies for resource management may give a valid support to the future Mobile Web-based services. All considered resource management strategies have been implemented in the simulator described in Section 2.4 and evaluated in the context of future workloads and technological scenarios. The resource management task is represented by a periodic job ($\Delta t = 20$ min) that evaluates the sets of the most popular resources $PR(t)$, pre-adapts and caches six versions (five adapted plus the original) of these resources, or a single SVC version of the content. Requests for these resources are directly served by the cache server; the other requests may involve, if necessary, on-the-fly adaptation. Each simulation lasts for 10 h of simulated time, and the performance results are averaged over 10 runs.

We consider a large space of alternatives represented by various workload scenarios, different resource

popularity distributions and different percentages of the working set resources that can be pre-adapted. The six scenarios are derived from the combination of two categories of Web sites (that is, future online-news and social-multimedia), and three workload evolutions (that is, low growth, high growth and HD growth). The resource popularity is denoted by the α parameter of the Zipf distribution. The analysis with respect to different Zipf α parameters is important because the efficacy of pre-adaptation is highly affected by the skewness of the popularity distribution. Even if recent studies suggest that $\alpha \in [0.84, 1]$ for the Mobile Web [10, 38], we prefer to consider a broader range of values, that is $\alpha \in [0.4, 1]$, to analyze the entire space of feasible application of the proposed strategies. The percentage of popular resources in the working set to be pre-adapted may range from 5% to 35%.

4.1 Performance of the algorithms

The first analysis evaluates the efficacy of the algorithms for the identification of the most popular resources. To this purpose, we evaluate the set of popular resources identified by the Traditional, Predictive-EWMA and Predictive-LR algorithms with respect to the set identified by the Ideal algorithm. The efficacy is measured as the percentage of resources that are correctly identified by each algorithm as belonging to the set of popular resources. We carried out several experiments where the set of popular resources ranges between 5% and 35% of the working set. The Ideal algorithm identifies different sets of resources for different workload scenarios and parameters. However, it is interesting to observe that all algorithms show quite stable results with respect to the set of popular resources identified by the Ideal algorithm. We report some representative results in Table 4: the set of popular resources corresponds to 15% of the working set, and $\alpha \in \{0.68, 0.84, 1\}$, as shown in recent studies [10, 15, 19].

For any scenario, the Predictive algorithms outperform the Traditional solution: the amount of popular resources correctly identified by the Predictive-EWMA and Predictive-LR algorithms ranges between 80% and 85%, with the EWMA solution obtaining a slight gain over the LR alternative. On the other hand, the Traditional algorithm does not reach 70%. This result is important because higher percentages will allow the system to deliver more adapted resources from the cache server without the need of computationally expensive on-the-fly adaptations.

4.2 Overall system performance

We now evaluate the impact of the Traditional, Predictive and Ideal algorithms on the performance of the overall system for Mobile Web-based services. As performance measure, we consider the server response

Table 4 Efficacy of the algorithms for the identification of the most popular resources

Experimental setup			Popular resource identified with respect to the set of the ideal algorithm		
Category of web site	Workload intensity	Zipf α parameter	Traditional	Predictive EWMA	Predictive LR
Online-news	Low growth	0.68	68%	85%	82%
Online-news	Low growth	0.84	69%	84%	82%
Online-news	Low growth	1.00	69%	85%	83%
Online-news	High growth	0.68	68%	81%	80%
Online-news	High growth	0.84	69%	82%	80%
Online-news	High growth	1.00	66%	83%	81%
Online-news	HD growth	0.68	69%	80%	80%
Online-news	HD growth	0.84	68%	81%	80%
Online-news	HD growth	1.00	66%	82%	81%
Social-multimedia	Low growth	0.68	68%	82%	80%
Social-multimedia	Low growth	0.84	68%	82%	81%
Social-multimedia	Low growth	1.00	69%	85%	83%
Social-multimedia	High growth	0.68	67%	83%	81%
Social-multimedia	High growth	0.84	66%	83%	81%
Social-multimedia	High growth	1.00	65%	84%	82%
Social-multimedia	HD growth	0.68	67%	83%	81%
Social-multimedia	HD growth	0.84	67%	84%	81%
Social-multimedia	HD growth	1.00	66%	85%	82%

time for various sites and workload scenarios, as defined in Section 2.4. In particular, we focus on the 90-percentile of response time, that is a more significant metric than average values when heavy tailed distribution are involved, such in the case of Mobile Web. Due to the large space of alternatives, we present the most representative results aiming to identify the conditions where the pre-adaptation and predictive algorithms may provide a significant performance gain.

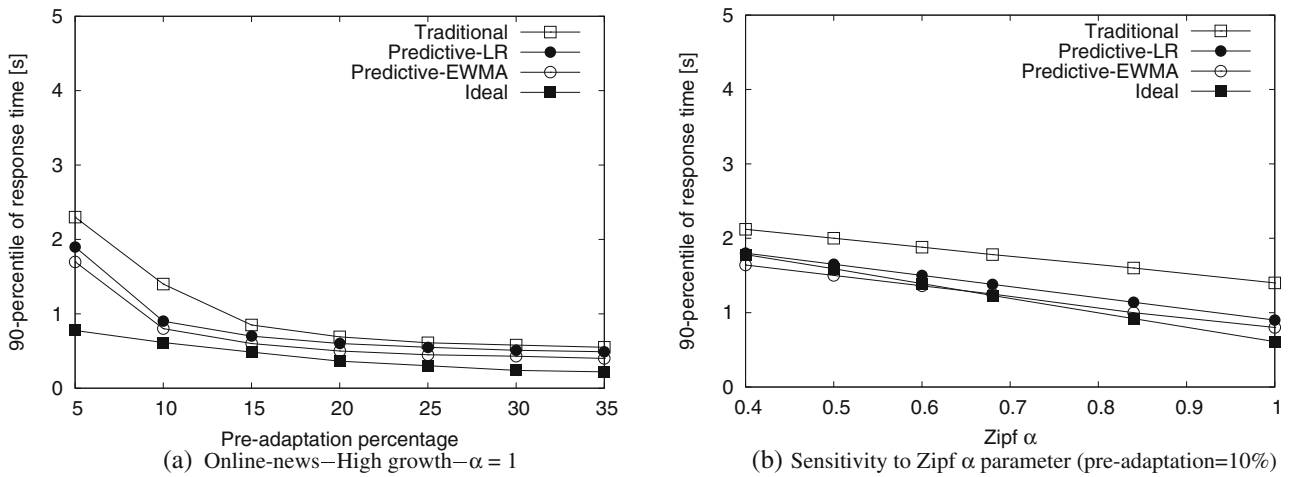
We initially consider the less challenging scenarios in terms of computational load. In these instances, the results are similar for any future scenario related to the online-news sites and for the low growth future scenario related to social-multimedia sites. In Fig. 5 we report only the results referred to the online-news high growth scenario as a representative case of workload characterized by a low computational demand.

Figure 5a represents the 90-percentile of response time as a function of the amount of pre-adapted resources for $\alpha = 1$. This figure shows that a pre-generation between 5% and 10% of the working set is sufficient to improve performance and to avoid system overload. Furthermore, the close curves related to the considered algorithms show that any pre-adaptation strategy can manage these kinds of scenarios. If we consider the sensitivity to the resource popularity skewness in Fig. 5b, we observe that, for every value of the Zipf parameter α in the range [0.4, 1], even a small amount of pre-adaptation (10% in the figure) is sufficient to guarantee good performance. These results are motivated by the low computational demand caused by the these workload scenarios on the system.

A more demanding scenario from a computational point of view is represented by social-multimedia sites with a high growth scenario. In this case the choice of which fraction of the working set must be pre-adapted, and the algorithm to estimate the resource popularity have a significant impact on system performance.

Figure 6a shows the performance of pre-adaptation in the social-multimedia high growth scenario for different percentages of pre-adapted resources. We observe that pre-adapting a large fraction of the working set (that is, 25%–35%) allows the system to achieve low response times for any algorithm. Furthermore, the low response time is achieved for a wide range of resource popularity skewness. In particular, if we pre-generate 35% of the working set, the 90-percentile of response time remains below 10 seconds for any value of $\alpha \geq 0.6$. However, due to the inherent dynamic nature of future Mobile Web scenarios, a large amount of pre-adaptation is likely to introduce consistency issues when multiple versions of the same content are created or when the content is replicated in more servers. On the other hand, if pre-adaptation is applied to a small fraction of the working set (that is, below 5%), most algorithms cannot guarantee adequate performance even for $\alpha = 1$.

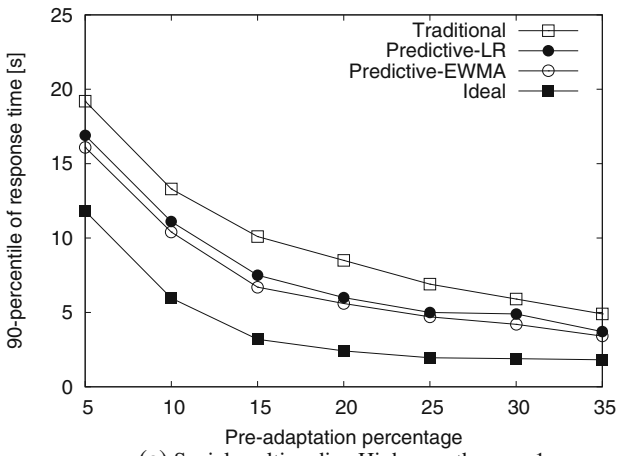
For pre-adaptation in the range from 10% to 20% the efficacy of the algorithm to identify popular resources plays a fundamental role. If we compare the Traditional algorithm with the Predictive alternatives in Fig. 6a, we observe a performance gain from 20% to 29% when the amount of pre-adaptation ranges between 10% and 20%. Furthermore, if we compare the Predictive-EWMA and the Predictive-LR algorithms we observe very similar performance, with the Predictive-EWMA outperforming the LR alternative by less than 7% considering the 90-percentile of the response time. This confirms that the approach of using



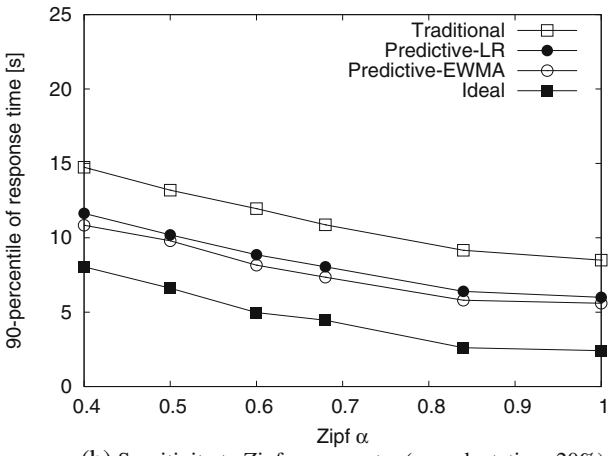
(a) Online-news–High growth–α = 1

(b) Sensitivity to Zipf α parameter (pre-adaptation=10%)

Fig. 5 Analysis of online-news high growth scenario (**a**, **b**)



(a) Social-multimedia−High growth −α = 1



(b) Sensitivity to Zipf α parameter (pre-adaptation=20%)

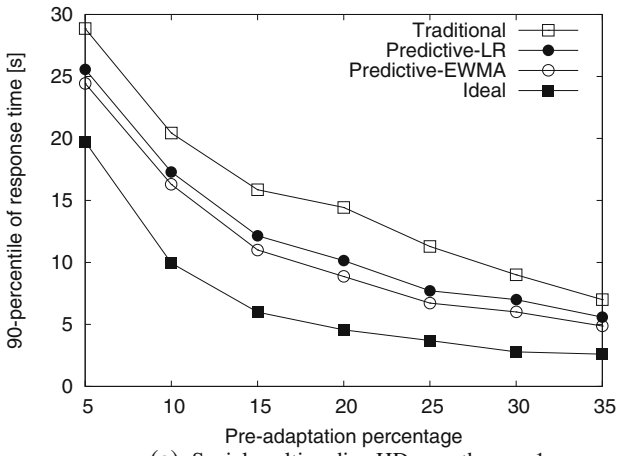
Fig. 6 Analysis of social-multimedia high growth scenario (**a**, **b**)

past access history to predict resource popularity is effective and scarcely dependent on the specific prediction algorithm.

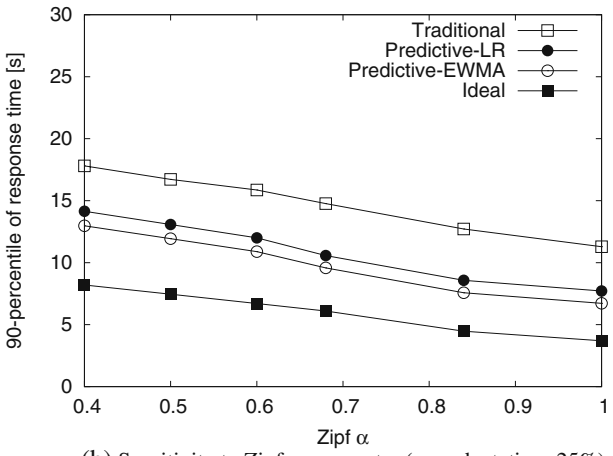
Although the Predictive algorithms outperform the Traditional algorithm, their response time is significantly higher than that achieved by the Ideal algorithm. We can conclude that in this scenario there is space for proposing novel algorithms for popular resource identification.

Due to the dependency of the response time on the resource popularity skewness, we consider important to evaluate the impact of the Zipf α parameter on the efficacy of the pre-adaptation strategies. Figure 6b shows the 90-percentile of the response time for the considered algorithms when pre-adaptation is set to 20% and α ranges between 0.4 and 1. We observe a performance degradation for every algorithm as α decreases. This result suggests that as the resource popularity becomes less skewed, the same level of performance may be achieved only by increasing the amount of pre-generation or by adopting more effective algorithms.

Finally, Fig. 7a shows the 90-percentile of the response times for the considered algorithms in the most computationally demanding scenario that is, social-multimedia HD growth for $\alpha = 1$. This figure shows that the Traditional algorithm is affected by poor results unless the amount of pre-adaptation is close to 35%. On the other hand, the Predictive-EWMA and Predictive-LR algorithms can still provide adequate performance with an amount of pre-adaptation close to 25%. In Fig. 7b we show the results as a function of less skewed popularity distributions. The efficacy of the proposed content management strategy is really critical in these instances. For example, when $\alpha <$



(a) Social-multimedia−HD growth− α =1



(b) Sensitivity to Zipf α parameter (pre-adaptation=25%)

Fig. 7 Analysis of social-multimedia HD growth scenario (**a**, **b**)

0.84, the system performance is unacceptable even for the Predictive and Ideal algorithms, hence the only alternative is to increase the amount of pre-adapted resources. On the positive hand, we can consider that the presence of High Definition content is likely to be related to video resources even for the next 5 years. As video on demand services are characterized by highly skewed popularity distribution [10], the results of pre-adaptation close to $\alpha = 1$ are the most probable, and the most acceptable from the performance point of view.

5 Related work

Recent research about the Mobile Web spans different areas, ranging from workload characterization and novel services to studies on strategies for content adaptation and generation, caching and delivery to mobile clients up to proposal of innovative hardware/software architectures for supporting the Mobile Web services. This paper considers workload evolution for the next future and fits in the area of resource management strategies based on selective pre-adaptation.

Several solutions to tailor contents to mobile devices are based on on-the-fly adaptation [30, 32]. These approaches are usually integrated with caching strategies [6, 7, 12] that exploit a sort of utility function to determine whether or not it is convenient to cache an adapted version of some resource(s). All these proposals consider a traditional Web scenario, with a limited amount of multimedia resources and a small fraction of requests coming from mobile devices and, consequently, requiring adaptation. In the context of future Mobile Web, the solutions that are based just on on-the-fly content adaptation may not represent a viable approach even if we consider technological improvements and content management strategies based on caching.

A preliminary analysis of how the performance requirements of the Mobile Web are expected to evolve in the next future was proposed by the same authors in [8, 9]. The paper in [9] represents the first attempt to point out the main issues related to the impact of future services on the server infrastructures. In [8] we propose a preliminary resource management strategy for the future Mobile Web that is based on an offline pre-adaptation of the resources. This paper represents a clear step ahead with respect to the previous studies for several reasons. We present an extensive analysis of the performance impact of future services where we take into account multiple components (CPU and disk) of the server infrastructure; the analysis carried out through a simulator platform allows us to evaluate the performance in the context represented by the future server infrastructures. Furthermore, we propose innovative algorithms based on prediction for the identification of the most popular resources of the working set that are specifically tailored to the workload characteristics of the future Mobile Web.

Although not so popular in the current Web, offline pre-adaptation of the working set is a strategy used by some portals, such as AvantGo [23], to deliver content to mobile users. This portal pre-adapts a set of defined resources to reduce the response time perceived by the users. This solution limits adaptation to few well defined Web sites, for which all multimedia resources are pre-adapted. On the other hand, we apply pre-adaptation to the most popular resources of the working set. This proposal represents a more affordable solution that allows us to serve contents belonging to several Web sites, because the number of required adaptations is limited to the restricted set of the most popular resources. Our idea takes advantage of recent results on the workload characterization of the Mobile Web [10, 19]. Specifically, we exploit the analysis on multimedia resource popularity in social-multimedia Web sites, that have been proved to follow a Zipf-like distribution [10]. Our results confirm that limiting pre-adaptation to a subset of popular resources can improve the overall performance of the Mobile Web services.

The effectiveness of the proposed pre-adaptation strategy strongly depends on the ability to estimate the future popularity of the working set resources. Several algorithms to identify the most popular resources were considered in the context of traditional Web-based services for replication and caching purposes [25, 29, 33]. In these studies, the resource popularity is mainly determined through simple measures on past resource accesses. When the resource popularity changes slowly, the mechanisms based on direct observation of the past request rates may be sufficient to identify the most popular resources of the working set with an acceptable accuracy. However, in the context of the Mobile Web, the workload characteristics are significantly different and require novel approaches. The algorithm for the identification of the most popular resources proposed in this paper exploits recent results on the workload characterization of the emerging Web context [16, 19] showing that the resource popularity in Mobile Web sites rapidly changes due to irregular access patterns, short resource lifespan and frequent resource uploads. These results motivate our innovative approach that exploits predictive techniques for the identification of the most popular resources of the working set.

Prediction has a long tradition in Web-based scenarios, but several predictive models are designed for offline operations and long-term forecasting. This is the case of Support Vector Machines [14], wavelet analysis [28] and neural networks [34] that may achieve a valid prediction accuracy only after a long learning time. These models cannot be exploited for short-term prediction in extremely variable contexts, such as the future Mobile Web. Other models, such as Kalman filters and ARIMA [3], require a careful choice of the model parameters, that is typically based on the evaluation of the auto-correlation and partial auto-correlation functions on a specific data set [5]. These predictive techniques may have serious difficulties to predict accurate values when the data set is extremely variable, as in the considered Mobile Web scenario, because they would require an update of their parameters at any significant change of system/workload state. For these reasons, we prefer to recur to simple yet robust linear algorithms that do not require a complex choice of their parameters. Similar models were applied to estimation of Internet traffic [31], server load [1], hot spots [2], and this paper represents a first attempt to apply linear prediction techniques to the context of the Mobile Web, characterized by extreme workload variability and short resource lifespan.

6 Conclusions

The advent of the Mobile Web will cause additional computational and storage requirements to the server infrastructure of the content provider that has to generate and tailor contents to user preferences and device capabilities. In this paper, we analyze the server requirements of present and future Mobile Web by taking into account technological improvements in terms of disk and CPU, and the main evolutionary trends of the workload characteristics for the period 2009–2014. Our evaluation demonstrates that the computational demand of the Mobile Web is likely to grow to the extent where for some classes of services it is impossible to support on-the-fly adaptation for every resource. To reduce the computational demand of the future Mobile Web we propose a selective resource pre-adaption strategy that pre-adapts offline the subset of the most popular resources. We combine this strategy with a class of novel predictive algorithms that estimate the resource popularity also in the challenging workload context of the Mobile Web, that is characterized by short resource lifespan, high variability and innovative user interaction patterns.

We explore the space of workload scenarios to evaluate the performance and the limits of the proposed pre-generation strategies and algorithms. We found that, when the computational demand of future scenarios does no exceed significantly the available computational power, pre-adapting just a small fraction of the most popular resources (5%–10% of the working set) is sufficient to guarantee adequate response times and any algorithm can be used to identify the set of popular resources. In the more demanding scenarios, as in the case where social services and multimedia content are to be delivered, the traditional algorithms for resource popularity estimation are unsuitable, and we have to recur to the proposed predictive algorithms. These strategies work finely even when the resource popularity is scarcely skewed, and when the pre-adaptation can involve up to 20% of the working set. On the other hand, when the workload is characterized by High Definition multimedia resources, no algorithm guarantees adequate performance, unless the resource popularity is really highly skewed. In these instances, novel algorithms and/or the integration with architectural solutions seem the only viable alternative for resource management.

References

1. Andreolini M, Casolari S, Colajanni M (2008) Models and framework for supporting runtime decisions in web-based systems. ACM Trans Web 2(3):1–43
2. Baryshnikov Y, Coffman E, Pierre G, Rubenstein D, Squillante M, Yimwadsana T (2005) Predictability of web-server traffic congestion. In: Proceedings of the 10th international workshop on web content caching and distribution (WCW 2005)
3. Basseville M, Nikiforov I (1993) Detection of abrupt changes: theory and application. Prentice-Hall, Englewood Cliffs
4. Berg Insight AB (2007) Mobile internet 2.0. Research report
5. Brockwell PJ, Davis RA (2001) Introduction to time series and forecasting. Springer, New York
6. Buchholz S, Buchholz T (2004) Replica placement in adaptive content distribution networks. In: Proc. of the 2004 ACM symposium on applied computing (SAC'04)
7. Canali C, Cardellini V, Lancellotti R (2006) Content adaptation architectures based on squid proxy server. World Wide Web J 9(1):63–92
8. Canali C, Colajanni M, Lancellotti R (2008) Resource management strategies for Mobile Web-based services. In: Proc. of 4th IEEE international conference on wireless and mobile computing (WIMOB'08)
9. Canali C, Colajanni M, Lancellotti R (2009) Performance evolution of mobile-web based services. IEEE Internet Computing 13:60–68
10. Cha M, Kwak H, Rodriguez P, Ahn YY, Moon S (2007) I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In: Proc. of the 7th ACM SIGCOMM conference on internet measurement (IMC'07), San Diego

11. Chandra S, Gehani A, Ellis CS, Vahdat A (2001) Transcoding characteristics of web images. In: Proceedings. of multimedia computing and networking (MMCN'01), San Jose
12. Chang CY, Chen MS (2003) On exploring aggregate effect for efficient cache replacement in transcoding proxies. IEEE Trans Parallel Distrib Syst 14:611–624
13. Comcast (2008) Comcast announces more than 1,000 HD choices available the most HD content anytime, anywhere. Press release
14. Cortes C, Vapnik V (1995) Support-vector networks. J Mach Learn 20(3):273–297
15. Crovella ME, Bestavros A (1997) Self-similarity in world wide web traffic: evidence and possible causes. IEEE/ACM Trans Netw 5(6):835–846
16. Duarte F, Mattos B, Bestavros A, Almeida V, Almeida J (2007) Traffic characteristics and communication patterns in blogosphere. In: Proc. of international conference on weblogs and social media, Seattle
17. Flesca S, Greco S, Tagarelli A, Zumpano E (2005) Mining user preferences, page content and usage to personalize website navigation. World Wide Web 8(3):317–345
18. Gartner (2007) Mobile Web trends 2007 to 2011. Research report
19. Gill P, Arlitt M, Li Z, Mahanti A (2007) YouTube traffic characterization: a view from edge. In: Proc. of internet measurement conference (IMC'07)
20. Grochowski E, Halem RD (2003) Technological impact of magnetic hard disk drives on storage systems. IBM Syst J 42(2):205–217
21. Guo L, Chen S, Xiao Z, Zhang X (2005) Analysis of multimedia workloads with implications for internet streaming. In: WWW '05: proc. of the 14th international conference on world wide web
22. Iaccarino G, Malandrino D, Percio MD, Scarano V (2006) Efficient edge-services for colorblind users. In: Poster proc. of WWW 2006
23. iAnywhere Inc. (2005) AvantGo. http://www.avantgo.com/
24. Juniper Research (2007) Mobile user generated content—dating, social networking and personal content delivery 2007–2012. Research report
25. Karlsson M (2005) Replica placement and request routing. In: Tang, Xu, Chanson (eds) Web content delivery. Springer, New York
26. e Marketer (2008) Everyone is talking about mobile social networking. Research report
27. OMNeT++ discrete event simulation system (2008) http://www.omnetpp.org
28. Percival DB, Walden AT (2000) Wavelet methods for time series analysis. Cambridge University Press, Cambridge
29. Rabinovich M, Spatscheck O (2002) Web caching and replication. Addison Wesley, Reading
30. Rodriguez J, Dakar B, Marras FL, Schreder C, Suzuki S, Tapera E (2001) New capabilities in IBM websphere transcoding publisher version 3.5 extending web applications to the pervasive world. IBM RedBooks
31. Sang A, Li SQ (2000) A predictability analysis of network traffic. In: Proc. of the 9th annual joint conference of the IEEE computer and communications societies (INFOCOM)
32. Singh G (2004) Guest editor's introduction: content repurposing. IEEE Multimed 11(1):20–21
33. Sivasubramanian S, Pierre G, van Steen M, Alonso G (2007) Analysis of caching and replication strategies for web applications. IEEE Intell Syst 11(1):60–66
34. Spooner JT, Ordonez R, Maggiore M, Passino KM (2001) Stable adaptive control and estimation for nonlinear systems: neural and fuzzy approximation techniques. Wiley, New York
35. Sung HH (2002) Helping online customers decide through web personalization. Inf Sci 17(6):34–43
36. Tan W, Chan E, Zalchor A (1996) Real time software implementation of scalable video codec. In: Proceedings of international conference on image processing, Lausanne
37. Williams A, Arlitt M, Williamson C, Barker K (2005) Web workload characterization: ten years later. In: Tang, Xu, Chanson (eds) Web content delivery. Springer, New York
38. Yamakami T (2006) A zipf-like distribution of popularity and hits in the Mobile Web pages with short life time. In: Proc. of the international conference on parallel and distributed computing, applications and technologies (PDCAT'06), Taipei