Journal of Network and Computer Applications 35 (2012) 2081–2091

Contents lists available at SciVerse ScienceDirect

Journal of Network and Computer Applications

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

A structural approach for modelling the hierarchical dynamic process of Web workload in a large-scale campus network

Yi Xie [a], J. Hu [b,*], S. Tang [c], X. Huang [d]

[a] School of Information Science and Technology, Sun Yat-Sen University, Guangzhou 510275, China
[b] School of Engineering and Information Technology, University of New South Wales at the Australian Defence Force Academy (UNSW@ADFA), Canberra, ACT 2600, Australia
[c] Department of Engineering Technology, Missouri Western State University, St. Joseph, MO 64507, USA
[d] Network and Information Technology Center, Sun Yat-Sen University, Guangzhou 510275, China

ARTICLE INFO

Article history:
Received 21 December 2011
Received in revised form
14 July 2012
Accepted 20 August 2012
Available online 6 September 2012

Keywords:
Structural approach
Hierarchical dynamic process
Workload

ABSTRACT

A new structural approach based on hidden Markov model is proposed to describe the hierarchical nature of dynamic process of Web workload. The proposed approach includes two latent Markov chains and one observable process. One of the latent Markov chains is called macro-state process which is used to describe the large-scale trends of Web workload. The remaining latent Markov chain is called sub-state process which is used to describe the small-scale fluctuations that are happening within the duration of a given macro-state. An efficient parameter re-estimation algorithm and a workload simulation algorithm are derived for the proposed discrete model. Experiments based on a real workload of a large-scale campus network are implemented to validate the proposed model.

1. Introduction

Network traffic analysis and modelling are very important in performance analysis and network security (Bossardt et al., 2007; Mahmood et al., 2010; Su, 2011). Without a correct model of the Internet workload, it is difficult to give an accurate prediction of performance metrics. A validated model is the basis of performance test, capacity planning and attack defending. Therefore modelling Internet workload represents a critical task in both the design of Internet architectures and the research area.

In order to derive an accurate workload model, workload characteristics have to be understood. During the past two decades, properties of workload have been the subject of considerable interest. However, although numerous models have been proposed for modelling workload and some important characteristics of workload have been discovered, e.g., self-similar and long-range dependent (Paxson and Floyd, 1995; Leland et al., 2002), it is still not easy for most existing workload models to accurately describe the Internet workload, due to the heterogeneous and complex nature of workload running through the network.

Statistics model, structure model and time series model are three main approaches for workload modelling. The significant advantage of the statistical models is that they can accurately capture and represent the statistical characteristics of any type of observed data. However, one of their disadvantages is that they miss the information of time domain and hence lose the accuracy of fitting such models to the empirical workload. Moreover, statistics models can only present the global (large-scale, or said average) characteristics, but fail to cover much local details of the workload behaviour, especially those relevant to small probability events. The main contribution of structure models is that they can describe the self-similarity properties of most practical signals. However, their main drawback is that almost all existing models focusing on self-similarity properties are only defined and controlled by the same parameter, *Hurst*. It is well known that workload has many various features, e.g., the diurnal variation and the weekly variation in addition to the self-similarity (Li and Yu, 2008). Since it is unlikely that the wide variety of scaling encountered in real workload can be modelled by a process with a single parameter, such a model becomes overly rigid in some cases. Another drawback of structure models is that they also ignore the dynamic process of time domain which is similar to the statistics models. Time series models typically provide a good fit to the temporal dynamic of the observed workload. However, most existing time series approaches only consider one-step and flat models. The existing studies on network measurement indicated that the real workload is irregular, diversified and fractal. Thus, performance of these flat models tends to degrade when the variation range of workload becomes wider and more complex. A general way for solving this issue is to introduce more parameters for a given model, which may increase the computational complexity of the model.

* Corresponding author. Tel.: +61 3 99259793; fax: +61 2 62688580.
E-mail address: J.Hu@adfa.edu.au (J. Hu).

1084-8045/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jnca.2012.08.010

Another common drawback of most current workload models is that it is not easy for a flat model to accurately describe the multi-resolution dynamic time process, e.g., the large-scale slowly varying process and the small-scale fast-varying process. As a result, these models only capture the large-probability events, but ignore the small-probability events which are important for profiling the local changes of workload. These limitations lead to the inflexibility and inaccuracy of most existing workload models and motivate the continuous research on the Internet workload modelling.

In contrast with those previous works, we focus on modelling the hierarchical dynamic time process of Web workload. We achieve this goal by improving the structure of workload model instead of increasing the number of the model parameters. In this paper, we consider that the general Web workload is controlled by two nested time processes: the large-scale process and the small-scale process. The large-scale process sketches the coarse profile of the Web workload. It can be considered as a slowly varying process, e.g., the diurnal variation, the weekly variation or the seasonal variation. The small-scale process is a fast-varying process which describes the small fluctuations process or details of each phase of the large-scale process. The main contribution of this paper is twofold: (i) A structured and discrete time series model is proposed to profile the hierarchical dynamic time process of Web workload. (ii) The proposed model is applied to workload analysis and simulation. The main novelty of this work is that the proposed scheme enables us to capture workload's hierarchical dynamic processes which vary over time at different scales, e.g., large-scale/slowly varying process and small-scale/ fast-varying process. Thus, the proposed model not only can capture the static statistical properties, but also can describe the details of multi-scale dynamic process of workload.

The paper is organized as follows: Section 2 summarizes the existing related work. Section 3 briefly introduces the preliminaries to be used in the design of the proposed model. Section 4 introduces the proposed approach and the related algorithms. Section 5 shows the experiment results. Section 6 discusses some important issues of the proposed scheme. Conclusion and future outlook are presented in Section 7.

2. Related work

Our literature survey has noted that researchers attempt to model workload from three different ways: pure statistical methods, structure methods and time process methods. From all these perspectives, researchers are mining various underlying characteristics of workload. Here, we survey representative research from each perspective.

The objective of statistical methods is to capture the underlying statistical nature of workload, e.g., statistical distributions of workload. Such statistical features are regarded as the intrinsic mechanisms which control the behaviour of workload, and then they can be applied to improving link performance, generating simulation workload and classification (Lin et al., 2009). One of the most widely used and oldest statistical mode for workload is the Poisson Model which is memoryless and is the predominant model used for analysing workload in traditional telephony networks (Frost and Melamed, 2002). In a Poisson process the inter-arrival times are exponentially distributed with a rate parameter $\lambda : P\{A_n \le t\} = 1 - e^{-\lambda t}$. The potential assumptions of it include: (i) the arrivals are from a large number of independent sources, referred to as Poisson sources; (ii) the number of sources is infinite; and (iii) the workload arrival pattern is random. Then, the distribution has a mean and variance equal to the parameter λ. Heavy-tailed distributions are another main ways to profile the arrival rate of workload. In Adas (2002), the Pareto distribution process is used to produce independent and identically distributed (IID) inter-arrival times; the Weibull distributed process is applied to modelling the fixed rate in ON period and ON/OFF period lengths. The main disadvantages of the statistical methods are twofold: (i) they miss the information of time domain and lose the accuracy of fitting such models to the empirical traffic trace; (ii) they can only present the global (large-scale/slowly varying) characteristics, but cover much local (or said fast-varying) details of the workload.

The early work regarding the structure properties of workload was experimentally studied by Leland et al. (2002), Beran et al. (1995), and Crovella and Bestavros (1997), where authors discovered that workload is bursty and correlated over a wide range of time scales. That is, workload is both self-similar and long-range dependent. These two important characteristics are not captured by traditional Poisson or Markov based models. However, they are naturally modelled in the framework of self-similar or fractal processes. After that, there exist a wide range of mathematical models of self-similar and long-range dependent traffic each with its own idiosyncrasies, e.g., classical ON/OFF model (Leland et al., 2002), $M/G/\infty$ model (Paxson and Floyd, 1995), wavelet (Riedi et al., 2002), Fractional Brownian motion model (Norros, 2002). However, almost all existing models focusing on self-similarity properties are only defined and controlled by the same parameter, *Hurst*. Since it is unlikely that the wide variety of scaling encountered in real workload can be modelled by a process with a single parameter, such a model becomes overly rigid in some cases. Another drawback of structure models is that they also ignore the dynamic process of time domain which is similar to the statistics models. Thus, the workload generated by structural models is quite different from the real workload.

A common way to providing a good fit to the real workload is using time series models. Such models consider the workload as time series and use observed data to train the model parameters. These parameters are then used for network performance analysis or network traffic rebuild. The typical time series methods include: Markovian family models, e.g., Markov arrival rate model (Frost and Melamed, 2002), Markov modulated source models (Elwalid et al., Kulkarni, 1997; Ren and Kobayashi, 2002; Muscariello et al., 2005), hidden Markov model (HMM) (Garcia et al., 2010; Dainotti et al., 2006; Maia, 2010; Domańska et al., 2011), and Autoregressive Integrated Moving Average (ARIMA) model series (Beran et al., 1995; Cox, 1984). The main disadvantage of these models is that most of them only consider one-step and flat models. In this type of models, parameters only reflect the global average of a given workload, but cannot describe the details of local changes. As a result, the small-probability events of workload may be covered by those large-probability events, and cannot be captured by these models. Thus, they are not suitable for the irregular and wide-varying workload. Using more parameters of a model, e.g., increasing the number of model's parameters, may be an option to solve this issue. However, for most mathematical models, the relationship between the number of model parameters and its computational complexity is not linear. Thus, the computational complexity of the model may also be greatly raised with the increment of the number of model parameters.

"Hierarchical method" has been used in many prior works, e.g., Nuzman et al. (2002), Menascé et al. (2003), and Hariri et al. (2008). However, these works are quite different from the issues concerned in this paper. In Nuzman et al. (2002) and Menascé et al. (2003), authors only focused on workload's statistical properties of different TCP/IP layers instead of the hierarchical dynamic time process of workload. In Hariri et al. (2008), the underlying state processes were used to model the game states