Chapter 55
Domain Knowledge Embedding Regularization Neural Networks for Workload Prediction and Analysis in Cloud Computing

Lei Li
https://orcid.org/0000-0001-7782-1876
School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China

Min Feng
21CN Co., Ltd., Guangzhou, China

Lianwen Jin
School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China

Shenjin Chen
School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China

Lihong Ma
School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China

Jiakai Gao
Xidian University, Xi'an, China

ABSTRACT

Online services are now commonly deployed via cloud computing based on Infrastructure as a Service (IaaS) to Platform-as-a-Service (PaaS) and Software-as-a-Service (SaaS). However, workload is not constant over time, so guaranteeing the quality of service (QoS) and resource cost-effectiveness, which is determined by on-demand workload resource requirements, is a challenging issue. In this article, the authors propose a neural network-based-method termed domain knowledge embedding regularization neural networks (DKRNN) for large-scale workload prediction. Based on analyzing the statistical properties of a real large-scale workload, domain knowledge, which provides extended information about workload changes, is embedded into artificial neural networks (ANN) for linear regression to improve prediction accuracy. Furthermore, the regularization with noisy is combined to improve the generalization ability of artificial neural networks. The experiments demonstrate that the model can achieve more accuracy of workload prediction, provide more adaptive resource for higher resource cost effectiveness and have less impact on the QoS.

DOI: 10.4018/978-1-7998-5339-8.ch055

1. INTRODUCTION

Cloud computing (Buyya, Yeo, Venugopal, Broberg, & Brandic, 2009) is now a critical internet infrastructure. The emergence of "every resource as a service" (XaaS) (Serrano et al., 2016) has significantly altered the traditional internet resource use. XaaS can be divided into three main categories: Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS) and Software-as-a-Service (SaaS). These three service types are similar to meter services such as those used to measure water, electricity and gas usage (i.e., pay-per-use). Using cloud computing platforms to deploy online services differs from traditional hosting in three principal respects. First, they are provided on-demand, typically by the minute or hour; second, they are elastic, and platform users can adjust resources according to the requirements at any time; third, the service is fully managed by the service provider (Garrison, Wakefield, & Kim, 2015). According to these three requirements, cloud computing can deliver the pay-per-use model to platform users to optimize the resource cost-effectiveness and more enterprises and users use cloud computing for task handling.

End users access SaaS, PaaS or the web deployed on IaaS, with the workload varying according to the time of day, week, month and year. As a result, static resource allocation can make applications less effective, since there is an excess of resources available during periods of low demand, incurring unnecessary costs by the resource provider. Alternatively, there are insufficient resources to guarantee the quality of service (QoS) during high-workload periods. Poor QoS may risk the loss of end-consumers. Therefore, enterprises or platform users generally require efficient resource utilization that minimally impacts on the QoS to the end-consumers of PaaS or SaaS.

This problem can be solved using the "dynamic elastic resources provision" method, which adapts the resource size to usage at different times. A cloud computing resource scheduler uses prior experience or knowledge of the workload, such as request change trends and time correlations, to increase the resource size during periods of high workload and release resources during low workload periods to achieve efficient on-demand resource usage (Garrison, Wakefield, & Kim, 2015).

In this paper, we use and analyze a real large-scale workload dataset to reveal its structural and statistical properties. We show that large-scale workloads are not stationary stochastic signals, so we design a neural network termed Domain Knowledge embedding Regularization Neural Networks (DKRNN) as the workload prediction model to predict the future workload on a cloud computing platform. Time-series domain knowledge is embedded into ANN to provide extended structural information of workload changes to reduce prediction error. To demonstrate the effectiveness of the proposed method, we compare the results of our model, the Auto Regressive Integrated Moving Average (ARIMA), multiple linear regression (MLR) and ANN model with a regression evaluation criterion and the queuing model.

The remainder of the paper is organized as follows. Related works are presented in Section 2. Section 3 introduces the system architecture and system models based on our workload prediction method. Details of the prediction model are described in Section 4, and Section 5 reports the results of experiments that evaluate the accuracy of our workload prediction method. In Section 6, a cost model based on queuing theory is presented to analyze the impact on the QoS. Finally, we conclude in Section 7.