J. Parallel Distrib. Comput. 119 (2018) 162–171

Contents lists available at ScienceDirect

J. Parallel Distrib. Comput.

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

A data-driven approach of performance evaluation for cache server groups in content delivery network

Ziyan Wu [a], Zhihui Lu [a,*], Wei Zhang [a], Jie Wu [a], Shalin Huang [b], Patrick C.K. Hung [c]

[a] School of Computer Science, Fudan University, Shanghai 200433, China
[b] Wangsu Science & Technology Co., Ltd., Shanghai, China
[c] Faculty of Business and IT, University of Ontario Institute of Technology, Canada

HIGHLIGHTS

- We frame CDN performance evaluation problem as a sequence learning problem.
- We use representation learning by LSTM auto-encoder to extract useful features from CDN monitoring log data.
- We use a deep neural network to predict the reach rate of CDN service, and we compare our methods with state-of-arts methods which show ours is superior by empirical studies.

ARTICLE INFO

Article history:
Received 31 January 2018
Received in revised form 4 April 2018
Accepted 16 April 2018
Available online 27 April 2018

Keywords:
Edge computing
Deep learning
Content delivery network
Sequence learning
Predictive analysis
High dimensional data

ABSTRACT

In industry, Content Delivery Network (CDN) service providers are increasingly using data-driven mechanisms to build the performance models of the service-providing systems. Building a model to accurately describe the performance of the existing infrastructure is very crucial to make resource management decisions. Conventional approaches that use hand-tuned parameters or linear models have their drawbacks. Recently, data-driven paradigm has been shown to greatly outperform traditional methods in modeling complex systems. We design a data-driven approach to building a reasonable and feasible performance model for CDN cache server groups. We use deep LSTM auto-encoder to capture the temporal structures from the high-dimensional monitoring data, and use a deep neural network to predict the reach rate which is a client QoS measurement from the CDN service providers' perspective. The experimental results have shown that our model is able to outperform state-of-the-art models.

1. Introduction

There is a trend [7,11,13,18,20] that both academia and industry use data-driven methods to model complex networked systems. Traditional approaches typically use some simple heuristics. These methods have several drawbacks. They cannot accurately reflect the complex systems due to the lack of knowledge of the real-world environment. Driven by the opportunity to collect and analyze data (e.g., application quality measurement from end users), many recent proposals have demonstrated the promise of using deep learning to characterize and optimize networked systems. Drawing parallel from the success of deep-learning on pattern recognition, instead of using an empirical analytical model to describe the complex interaction of different features, we use deep learning methods and treat networked systems as a black-box.

Uploading all data or deploying all applications to a centralized cloud is infeasible because of the excessive latency and bandwidth limitation of the Internet. A promising approach to addressing centralized cloud bottleneck is edge computing. Edge computing pushes applications, data and computing power (services) away from centralized points to the logical extremes of a network. Edge computing replicates fragments of information across distributed networks of web servers, which may spread over a vast area. As a technological paradigm, edge computing is also referred to as mesh computing, peer-to-peer computing, autonomic (self-healing) computing, grid computing, and by other names implying non-centralized, nodeless availability [5]. CDN (content delivery network or content distribution network) is a typical representative of edge computing. A CDN is a globally distributed networked system deployed across the edge of Internet. Composed with geographically distributed cache servers, CDNs deliver cached content to customers worldwide based on their geographic locations. Extensively using cache servers, content delivery over CDN has

* Corresponding author.
E-mail address: lzh@fudan.edu.cn (Z. Lu).

https://doi.org/10.1016/j.jpdc.2018.04.010
0743-7315/

low latency, reliability, supports better quality of experience and security.

The CDN Service providers are increasingly using data-driven mechanisms to build the performance model of their service-providing systems. To build a model to accurately provide an understanding of the performance of the existing infrastructure such as the health of cache groups and network status, is very crucial to make resource management decisions including content placement, network traffic scheduling, and load balance of the CDN network. Modeling all available physical resources, we can maximize a resource utilization in terms of service quality, cost, profit, etc.

Generally, CDN providers do not have direct measurement from the clients (the logs from video players, web browser that can show the QoE of clients). Instead, they use the indirect measurement reach rate, the ratio of requests that meet the minimum standard, which is collected and calculated offline from the log of the HA proxy of CDN cache groups. In order to enable themselves make resource management decisions in real time, the CDN providers have to use the metrics that can be collected in the real time to infer the reach rate.

Cache server groups can be characterized as multi-dimensional, highly non-linear, time variant vectors. The metrics collected from members of the CDN cache server groups are sequence data that are measured every minute, which have hundreds of dimensions. The state-of-art methods are typically simple heuristics which are oversimplified and biased due to the human experience, or linear models, which cannot characterize the complex relationship between multiple metrics. Deep learning is a branch of machine learning based on a set of algorithms and attempt to model high-level abstractions in data by using artificial neural network architectures composed of multiple non-linear transformations [17]. They have a lot of successful applications in sequence modeling [15]. Compared to other machine learning techniques, a lot of work shows that it can detect complex relationships among features, and extract a hierarchical level of features from high-dimensional data, including monitoring data.

We frame our problem as a sequence learning problem, which consists of three stages: (1) feature engineering, (2) representation learning by LSTM auto-encoder to extract useful features, (3) a feed forward neural network black-box machine learning algorithm to output the predictions.

Our main contributions are listed below:

- We present a data-driven feasible approach to evaluating the performance of cache server groups in Content Delivery Network.
- We frame performance evaluation problem as a sequence learning problem.
- We use representation learning by LSTM auto-encoder to extract useful features from data.
- We compare our methods with state-of-arts methods to show that ours is superior by empirical studies.

The remainder of this paper is organized as follows. In Section 2, we first introduce the related concepts as our research background. In Section 3, we formulate our performance evaluation problem as a sequence learning problem and we also compare the baseline methods. In Section 4, we introduce our method of feature engineering to reduce the dimensionality for the high-dimensional data. We also introduce our reach rate prediction model based on LSTM auto-encoder. In Section 5, we show our experiment setting and demonstrate performance improvements of our methods over baseline models. Section 6 is discussion for related work. We provide concluding remarks in Section 7.

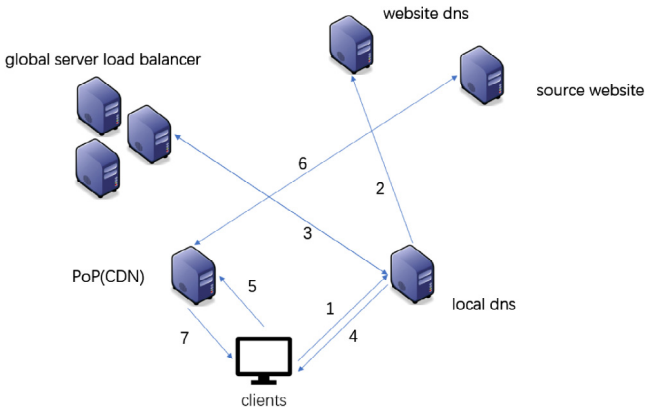

Fig. 1. The basic working procedure of CDN.

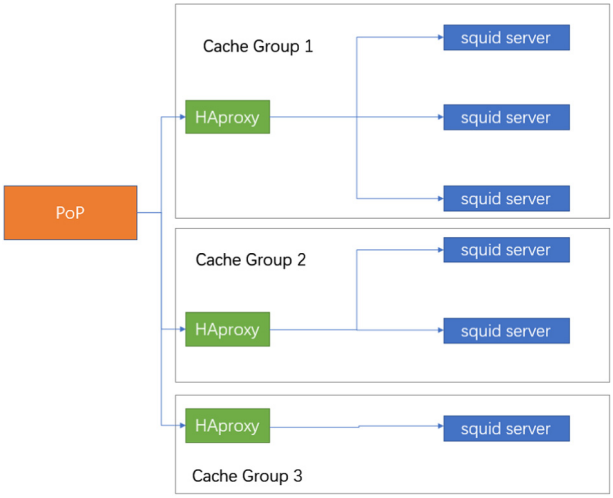

Fig. 2. The structure of cache server groups.

2. Background

A content delivery network or content distribution network (CDN) (Fig. 1) is a geographically distributed network of cache servers. CDN helps content provider to deliver web pages and other multimedia content to the clients, based on the locations of the clients and cache servers nearby the clients. The basic working procedure is as follows. Step 1: a client sends a request to local DNS. Step 2: the DNS finds the CNAME and redirects the request to the GSLB (global server load balance). Step 3: the local DNS server sends a request to the GSLB and GSLB returns the ip address of CDN servers based on the scheduling policy. Step 4: the local DNS returns the ip address to the clients. Step 5: clients request to fetch content from the selected PoP. Step 6: the cache server groups will pull the content from source website if the content does not exist locally. Step 7: content is sent to the clients.

A CDN cache server group (Fig. 2) is the basic resource scheduling unit for CDN. A CDN cache group is a load balanced cluster that consists of interconnected cache servers. A typical implementation consists of HAproxy and squid servers. HAproxy distributes the requests from clients to cache servers. HAproxy can be set to use different algorithms to maximize the utilization of every server. Round-robin algorithm distributes the load equally to each server in a homogeneous cluster. In a heterogeneous cluster, weighted round-robin algorithm is used. A weight was assigned to the server