Principal Researcher & email	Farrukh Nadeem. abdullatif@kau.edu.sa
Co-researchers & emails	Thomas Fahringer
Affiliation	National University of Computers and Emerging Sciences, Lahore, Pakistan; University of Innsbruck, Austria.
Research Title	Predicting the execution time of grid workflow applications through local learning
Research Topic	Performance Modeling of Scientific Workflow Applications
Publisher	ACM
Publishing Year	2009
ISBN	978-1-60558-744-8
Key Words	Scientific workflow applications, Performance modeling, Grid
Journal Name, or (Conference + place and date being held)	SC '09: International Conference on High Performance Computing Networking, Storage and Analysis, November 14-20, 2009, Oregon, Portland, USA.
Volume No.+ Issue No. and the Number of Pages in case it has been published in a scientific journal	pages = {1--12},
Research Abstract	Workflow execution time prediction is widely seen as a key service to understand the performance behavior and support the optimization of Grid workflow applications. In this paper, we present a novel approach for estimating the execution time of workflows based on Local Learning. The workflows are characterized in terms of different attributes describing structural and runtime information about workflow activities, control and data flow dependencies, number of Grid sites, problem size, etc. Our local learning framework is complemented by a dynamic weighing scheme that assigns weights to workflow attributes reflecting their impact on the workflow execution time. Predictions are given through intervals bounded by the minimum and maximum predicted values, which are associated with a confidence value indicating the degree of confidence about the prediction accuracy. Evaluation results for three real world workflows on a real Grid are presented to demonstrate the prediction accuracy and overheads of the proposed method.

Predicting the Execution Time of Grid Workflow Applications through Local Learning*

Farrukh Nadeem[1,2] and Thomas Fahringer[2]

[1]Department of Computer Science, National University of Computer and Emerging Sciences, Lahore, Pakistan

[2]Institute of Computer Science, University of Innsbruck, Austria

farrukh.nadeem@nu.edu.pk, {farrukh,tf}@dps.uibk.ac.at

ABSTRACT

Workflow execution time prediction is widely seen as a key service to understand the performance behavior and support the optimization of Grid workflow applications. In this paper, we present a novel approach for estimating the execution time of workflows based on Local Learning. The workflows are characterized in terms of different attributes describing structural and runtime information about workflow activities, control and data flow dependencies, number of Grid sites, problem size, etc. Our local learning framework is complemented by a dynamic weighing scheme that assigns weights to workflow attributes reflecting their impact on the workflow execution time. Predictions are given through intervals bounded by the minimum and maximum predicted values, which are associated with a confidence value indicating the degree of confidence about the prediction accuracy. Evaluation results for three real world workflows on a real Grid are presented to demonstrate the prediction accuracy and overheads of the proposed method.

1. INTRODUCTION

Grid workflows from scientific and business domains typically consist of several different activities (executables, services, etc.) with complex control flow and data flow dependencies among them. Execution of such workflows in large scale computational Grids, like Grid5000 [25], EGEE [4], etc., is commonly accomplished through a workflow composition and runtime environment like ASKALON [5] for distributed execution of workflow activities. The workflow runtime environment depends on online workflow execution time predictions to guide the performance-oriented optimization of the workflows.

Predicting the execution time of a workflow in the Grid is a complex problem and has been largely ignored so far due to the execution of workflow activities in a distributed fashion, involvement of several Grid resources (multiple Grid sites, LAN/WAN, etc.), external load, dynamic behavior of the Grid, inherent architectural and functional heterogeneity of Grid resources, and different structures of the workflows.

In this paper we introduce a *Local Learning Framework* for workflow execution time prediction, which is based on static information (number of activities in the workflow, control and data flow dependencies among workflow activities, etc.) and dynamic information about the execution of the workflows (through execution traces). This information about workflows is stored in a repository whose data (referred as *workflow data set* or simply *data set*) is used for *local learning* (*LL*). In the course of this paper, we refer to each instance of the data in the *data set* as *data instance*. The workflows are parameterized in terms of *attributes* (determined from the repository) defining workflow static and dynamic information (Section 2.2). The importance of different attributes w.r.t. their impact on execution time of the workflow is determined by attribute weights, which are dynamically determined through an *evolutionary algorithm* (Section 3). These weights are optimized considering the entire data set (to generalize effects of different values of attributes) as well as data subsets (to include effects of specific values of the attributes) and the best weights are selected adaptively (Section 3.2). Our local learning framework (*LLF*) employs hybrid metrics to find similarities in different workflows. The workflows identified to be similar (Section 2.1) are selected for *LL*, and the data set corresponding to the selected workflows is named as *local data set*. One instance of the *local data set* is referred as *local data instance*. We introduce a notion of *distance class* (Section 4) to dynamically select the size of *local data* such that the overall prediction error is minimized. We employ three induction models (Section 5) to predict workflow execution times (called point predictions) from the selected *local data*. A confidence value (ranging between 0 and 100) is associated with each prediction to indicate the degree of confidence about the prediction accuracy. A confidence value 100 means that the prediction is accurate, and a confidence value 0 means that the prediction is unreliable. To indicate possible variations in the predicted execution of a workflow, the minimum and maximum predicted execution times are provided as an interval

*The work described in this paper has been partially funded by the "Higher Eduction Commission" (HEC) of Pakistan, the EC funded edutain@grid project and the project "Parallel Computing with Java for Manycore Computers" funded by the Tiroler Zukunftsstiftung".

SC09 November 14-20, 2009, Portland, Oregon, USA (c) 2009 ACM 978-1-60558-744-8/09/11... $10.00.