Clustering Job Accounting Data

Huyen Nguyen

CS5352 Course Research Project

Mentor: Jie Li

Professor: Dr. Yong Chen

Outline

1. Problem & Research activities
2. Current progress
3. Plan for final deliverables

Problem & Research activities

PROBLEM STATEMENT

Tasks

DATA

Job accounting data

RESEARCH ACTIVITIES

The process

Problem statement

High Performance Computing (HPC) serves researchers and domain experts on numerous computing tasks.

It is crucial to understand the current HPC workloads

- Assisting informed scheduling research
- Enabling efficient scheduling in future HPC systems

How:

Clustering data to analyze usage behavior

Data

- Two months of job accounting data from a Quanah cluster
- 457,486 records
- 27 attributes, including:
  - start time, end time
  - cpu usage
  - memory usage

data.csv

Delimiter: ,

	qname	hostname	group	owner	job_name	job_number	submission_time	start_time
1	omni	compute-4-29.localdomain	phys	msanati	MPI_Test_Job	1925113	2020-07-30 07:15:06	2020-07-30 07:15:04
2	omni	compute-9-44.localdomain	phys	msanati	MPI_Test_Job	1925113	2020-07-30 07:15:06	2020-07-30 07:15:04
3	omni	compute-2-54.localdomain	chem	juandomi	phi_1-2.6	1927135	2020-08-01 02:08:08	2020-08-01 03:37:09
4	omni	compute-8-3.localdomain	chem	juandomi	phi_1-2.1	1927130	2020-08-01 02:08:02	2020-08-01 03:16:27
5	omni	compute-2-56.localdomain	ME	hge	Gm15ctc3	1927227	2020-08-01 05:09:45	2020-08-01 05:09:45

Research activities

1. Examine the job accounting data
2. Understand the metric in the job accounting information and extract features from it
3. Select and apply appropriate clustering algorithms on the data
    a. Partitioning clustering
    b. Hierarchical clustering
4. Determine the appropriate number of behavior categories
    a. Data-intensive jobs and
    b. Computational-intensive jobs
5. Discuss findings from the cluster analysis. By exploring the usage behavior, we can gain insights into the system and then work out the strategy to use the resource productively.

Current process

RESEARCH ACTIVITIES

About the current process

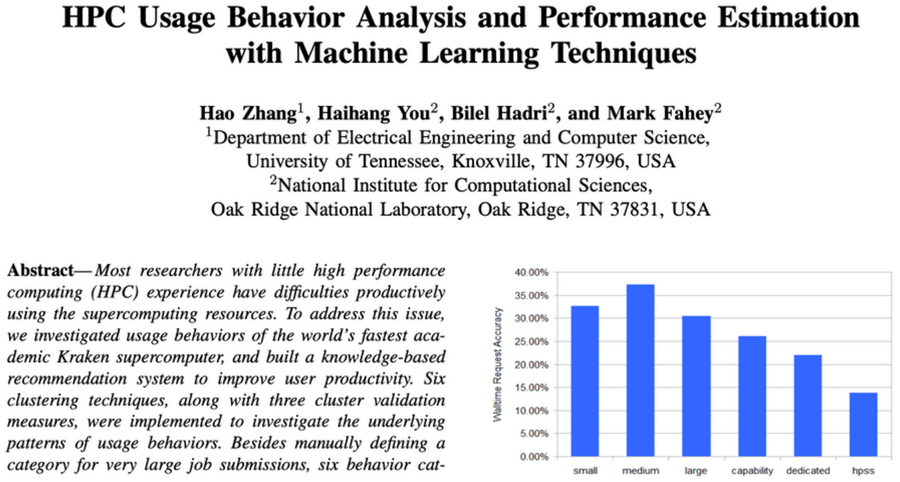
ADDRESSING COMMENTS

From the proposal

Reference

Zhang, Hao, et al. "HPC usage behavior analysis and performance estimation with machine learning techniques." Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2012. [1]

- Six clustering methods
- Three cluster validation measures
- Kraken supercomputer



HPC Usage Behavior Analysis and Performance Estimation with Machine Learning Techniques

Hao Zhang[1], Haihang You[2], Bilel Hadri[2], and Mark Fahey[2]
[1]Department of Electrical Engineering and Computer Science,
University of Tennessee, Knoxville, TN 37996, USA
[2]National Institute for Computational Sciences,
Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA

Abstract—*Most researchers with little high performance computing (HPC) experience have difficulties productively using the supercomputing resources. To address this issue, we investigated usage behaviors of the world's fastest academic Kraken supercomputer, and built a knowledge-based recommendation system to improve user productivity. Six clustering techniques, along with three cluster validation measures, were implemented to investigate the underlying patterns of usage behaviors. Besides manually defining a category for very large job submissions, six behavior cat-*

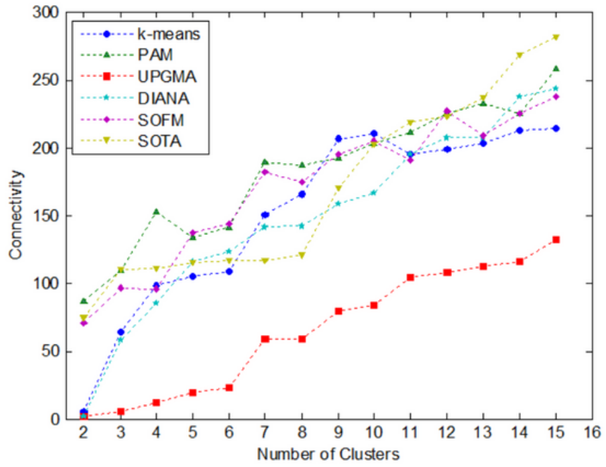
Understanding data

A feature is a distinctive characteristic of an object, such as a job, which is often represented as a function of the object's attributes. Four features are selected or extracted from job attribute space [1]

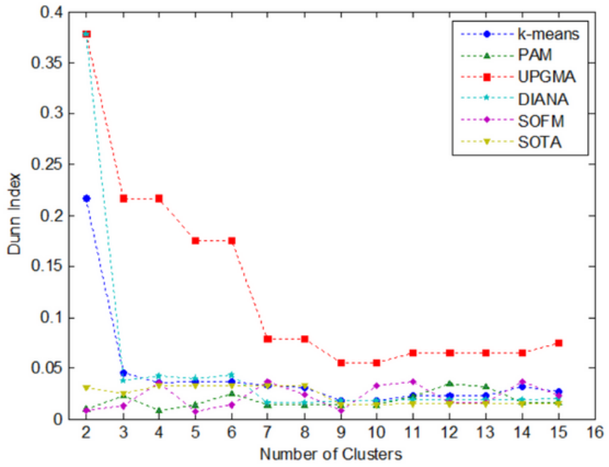
$N_n(j) = \log(\, j.nproc \,/\, C_n \,)$	number of allocated nodes
$M_u(j) = \log(\, j.mem\_used \,)$	memory used
$T_q(j) = \log(\, j.start\_time - j.submit\_time \,)$	job queue time
$T_r(j) = \log(\, j.end\_time - j.start\_time \,)$	job runtime

Clustering algorithms

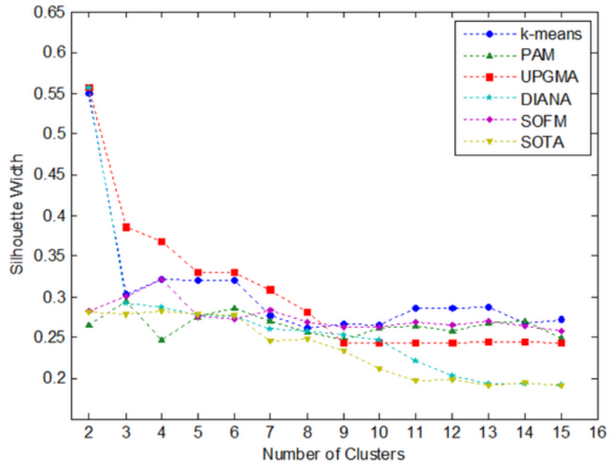
The comparison of six clustering algorithms on HPC data [1]



(a) Connectivity measure



(b) Dunn index measure



(c) Silhouette width measure

-> Selection: UPGMA for hierarchical clustering and k-means for partition clustering, # of clusters: 6

Addressing questions

1) Does the current job accounting data provide necessary metrics to differentiate these categories?

Answer: Yes. Although different systems may provide different encodings; but from the formula from [1], the current dataset provide necessary metrics:

$N_n(j) = \log(\, j.nproc \,/\, C_n \,)$	number of allocated nodes -> slots
$M_u(j) = \log(\, j.mem\_used \,)$	memory used
$T_q(j) = \log(\, j.start\_time - j.submit\_time \,)$	job queue time
$T_r(j) = \log(\, j.end\_time - j.start\_time \,)$	job runtime

In our dataset, the metric j.nproc is not available. The equivalent of such metric is slots, which is the total cpu cores used by a job.

Addressing questions

2) How do you plan to define these categories, i.e. what are your metrics to define these categories?

Answer: After getting the clusters, we can use Algorithmic mean of each job feature in different categories. For example from [1] with seven clusters. c3 can be considered as the set of data-intensive jobs: Jobs in this category use significant memory but relatively less number of compute nodes.

Feature	c1	c2	c3	c4	c5	c6	c7
$\mathrm{E}(T_q)$	4.24	2.62	2.87	1.89	4.65	3.74	4.27
$\mathrm{E}(T_r)$	4.19	3.63	3.22	2.38	4.15	2.64	3.22
$\mathrm{E}(M_u)$	4.05	4.03	5.11	4.09	4.17	4.12	5.07
$\mathrm{E}(N_n)$	0.19	0.26	1.01	0.98	1.51	1.51	3.90

Plan for final deliverables

Plan

1. Implement k-means and UPGMA on the Quanah dataset.
    a. See if the accuracy of UPGMA still holds in this dataset
2. Use, for example, a Dendrogram in accordance with the algorithm to identify the categories.
3. Determine job features with Algorithmic mean to define jobs belong to a category: data-intensive, computing intensive -> computing resources and well-tune jobs.

Thank you!