Contribution ID: **40** Type: **Poster**

A study on the applicability of Recommender Systems for the Production and Distributed Analysis system PanDA of the ATLAS Experiment

Thursday, 24 August 2017 16:30 (15 minutes)

Scientific computing has advanced in a way of how to deal with massive amounts of data, since the production capacities have increased significantly for the last decades. Most large science experiments require vast computing and data storage resources in order to provide results or predictions based on the data obtained. For scientific distributed computing systems with hundreds of petabytes of data and thousands of users it is important to keep track not just of how data is distributed in the system, but also of individual user's interests in the distributed data (reveal implicit interconnection between user and data objects). This however requires the collection and use of specific statistics such as correlations between data distribution, the mechanics of data distribution, and mainly user preferences.

This work focuses on user activities (specifically, data usages) and interests in such a distributed computing system, namely PanDA (Production ANd Distributed Analysis system). PanDA is a high-performance workload management system originally designed to meet production and analyses requirements for a data-driven workload at the Large Hadron Collider Computing Grid for the ATLAS Experiment hosted at CERN (the European Organization for Nuclear Research). In this work we are going to investigate whether data collection that was gathered in the past in PanDA shows any trends indicating that users could have mutual interests that would be kept for the next data usages (i.e., data usage patterns), with using data mining techniques such as association analysis, sequential pattern mining, and basics of the recommender system approach. We will show that such common interests between users indeed exist and thus could be used to provide recommendations (in terms of the collaborative filtering) to help users with their data selection process.

Primary authors: TITOV, Mikhail (National Research Centre Kurchatov Institute (RU)); ZARUBA, Gergely (University of Texas at Arlington (US)); DE, Kaushik (University of Texas at Arlington (US)); KLIMENTOV, Alexei (Brookhaven National Laboratory (US)); JHA, Shantenu (Rutgers University (US))

Presenter: TITOV, Mikhail (National Research Centre Kurchatov Institute (RU))

Session Classification: Poster Session

Track Classification: Track 2: Data Analysis - Algorithms and Tools