Contribution ID: **77** Type: **plenary reports**

Data Mining Techniques to Estimate User Preferences in Data Usage for the PanDA Production and Distributed Analysis System

For distributed computing systems with hundreds of petabytes of data and a large user base it is important to keep track not just of data distribution, but also of individual users'interests in distributed data. This motivates the collection of correlation statistics between data distribution mechanisms with information about user preferences. Beyond providing a popularity matrix per data collection, this approach can also deliver an explicit understanding of reasons for the popularity coefficients. 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. An earlier study of data popularity with probabilistic methods based on information from PanDA (jobs'parameters of processed data from the ATLAS experiment) required a highly complex model design; yet it provided low correlation to actual user interests. Thus it resulted in slow reaction times to changes in system (e.g. the popularity of derived data and/or data summary objects is limited by the time of data processing cycle). Data mining techniques are usually employed to reveal relations between objects in large data collections. In our case, we consider two main classes of objects as part of the study: users and items (datasets in terms of PanDA/ATLAS). Data mining processes, in general, consist of three steps: data preprocessing, data analysis, and result interpretation. We will focus on data mining techniques that are used in so-called recommender systems. Recommender systems are aimed to explore user activity, distinguish item features or group of features that are significant to individuals, reveal similar items and users, and as result estimate a user's preference to items that the user has never used before. Thus recommender system techniques could help in designing a user model that considers user preferences and preferences of group of users (with similar areas of interest). It includes methods to reduce space dimensionality, to measure similarity between objects, classification methods, and other methods. In this work we will investigate the applicability of such techniques in the context of PanDA.

Primary author: Mr TITOV, Mikhail (The University of Texas at Arlington)

Co-authors: Dr KLIMENTOV, Alexei (Brookhaven National Laboratory); Dr ZÁRUBA, Gergely (The University of Texas at Arlington); Dr DE, Kaushik (The University of Texas at Arlington)

Presenter: Mr TITOV, Mikhail (The University of Texas at Arlington)

Track Classification: Section 3 - Technology for storaging, searching and processing of Big Data