Contribution ID: **76** Type: **Poster**

Data Recommender System for the Production and Distributed Analysis System «PanDA»

Tuesday, 2 September 2014 08:00 (1 hour)

The Production and Distributed Analysis system (PanDA) is a distributed computing workload management system for processing user analysis, group analysis, and managed production jobs on the grid. The main goal of the recommender system for PanDA is to utilize user activity to build a corresponding model of user interests that can be considered in how data needs to be distributed. As an implicit outcome, the recommender system provides a quantitative assessment of users'potential interest in new data. Furthermore, relying on information about computer centers that are in users'activity zones, it provides an estimated list of computing centers as possible candidates for data storage. As an explicit outcome, the system recommends data collections to users by estimating/predicting the likelihood of user interest in such data.
The proposed recommender system is based on data mining techniques and combines two basic approaches: content-based filtering and collaborative filtering. Each approach has its own advantages, while their combination helps to increase the accuracy of the system. Content-based filtering is focused on creating user profiles based on data features and group of features including corresponding weights that show the significance of features for the user. Collaborative filtering can reveal the similarity between users and between data collections, thus such similarity measure may indicate how "close"objects are, i.e., how close pairs of users' preferences are to each other; or how close data-sets are to those that individual users have used previously.
Information about processed jobs taken from a PanDA database (in this study focusing on data coming from the ATLAS experiment) provides the recommender system with corresponding objects: users and input data (items in terms of recommender systems), and relations between them. This is the minimum required information to build the user-item matrix (i.e. utility matrix, where each element is an implicit rating of item per user). The herein proposed recommender system is not intrusive, i.e., it does not change any part of PanDA system but it can be used as an added-value service to increase efficiency of data management for PanDA.

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

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

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

Session Classification: Poster session

Track Classification: Computing Technology for Physics Research