A study of the applicability of recommender systems for the Production and Distributed Analysis system PanDA of the ATLAS Experiment

M. Titov[1], G. Záruba[2], K. De[2], A. Klimentov[1,3], S. Jha[4], R. Mashinistov[1,2] on behalf of the ATLAS Collaboration

[1] National Research Centre «Kurchatov Institute» (RU), [2] University of Texas at Arlington (US), [3] Brookhaven National Laboratory (US), [4] Rutgers University (US)

ACAT 2017
21-25 August 2017
University of Washington, Seattle

Abstract

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 users' interests in the distributed data (implicit interconnection between user and data objects). This however requires the collection and use of specific statistics such as correlations between data distribution, user preferences and the mechanics of data distribution.

The goal of this work is to investigate whether data 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, using data mining techniques such as association analysis, sequential pattern mining, and basics of the recommender system approach.

PanDA

Production **an**d **D**istributed **A**nalysis system PanDA [1] is a high-performance pilot-based workload management system. This means that workload is assigned based on the feedback from successfully activated and validated *pilot jobs*, which are lightweight processes that probe the environment (on compute nodes and clusters) and act as *"smart wrappers"* for the payload.

In PanDA, an independent subsystem manages the delivery of pilot jobs to all worker nodes via a number of well-known cluster and grid scheduling systems (e.g., Condor-G). Pilot-based systems [2] like PanDA also enable the integration of non-grid based resources and thus can scale and work simultaneously on large clusters as well as individual (non-grid organized) computers. The current high-level global structure of PanDA is shown in Figure 1.

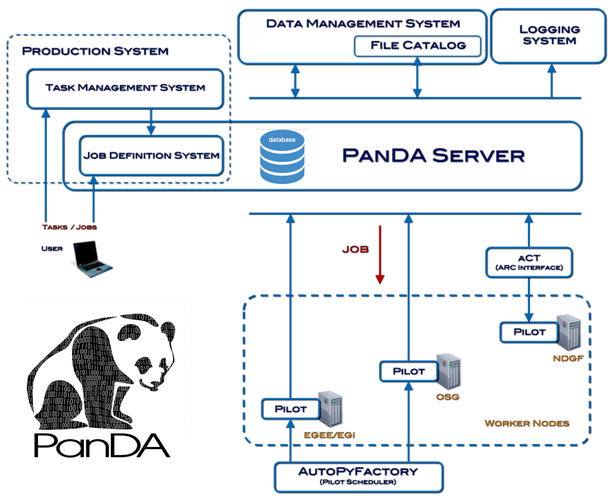


Figure 1. General structure of the PanDA system.

Recommender System

A recommender system uses a set of machine learning/data mining processes, that aim to guide users in a personalized way to interesting or useful items in a large space of possible options [3].

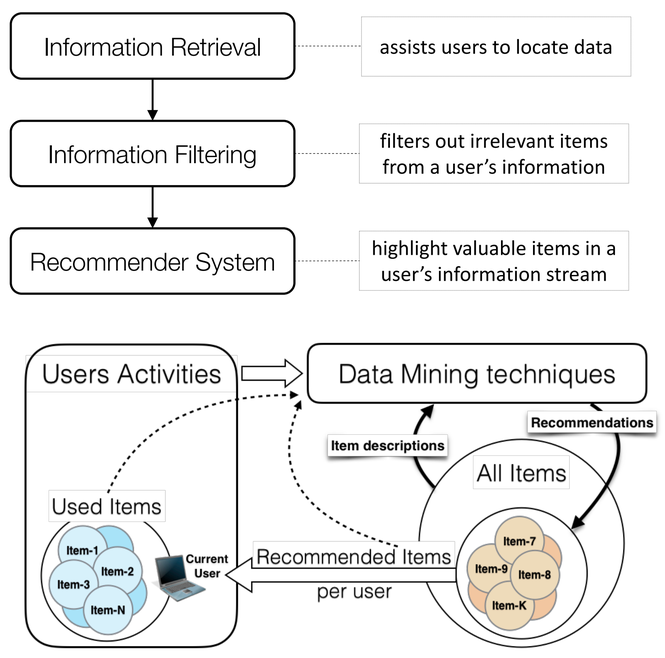


Figure 2. Recommender system overview.

Recommendation Simulation

Data Analysis on PanDA data using a simulated recommender system showed that some data that were marked as possibly interesting for peer users (based on popularity of data among similar users) were actually used later by the peers (Chart 1). Thus user activity might follow some usage pattern for the group of similar users, and could be correlated with specific user interests.

Estimated time difference between when an item was marked and its actual use is 28.7 days (avg.)

** Chart 1 shows the per day average (per user) number of items used (and its standard deviation), the average number of recommended items, and the true positive recommendations (related to the certain used object list size)*

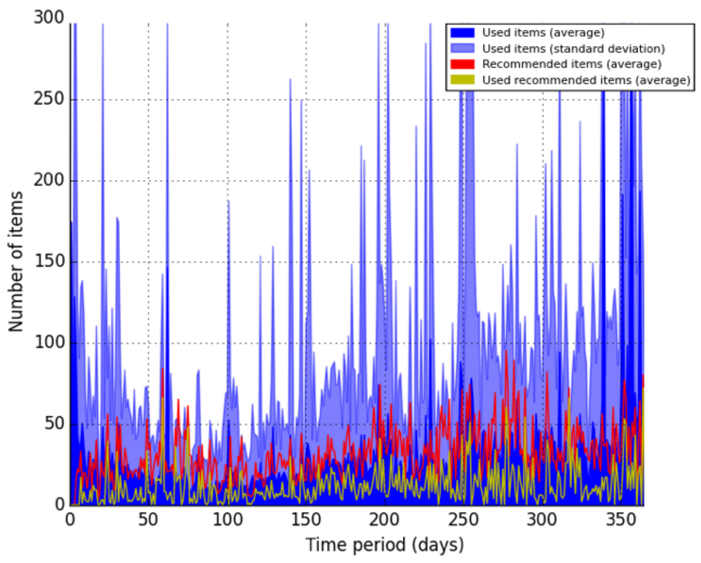


Chart 1. Number of items per day during the analysis period (y-axis is truncated to 300 items).

Sequential Pattern Mining

Sequential pattern mining [4] is the mining of frequently occurring ordered events or subsequences as patterns.

User activities are represented as data-sequences (i.e., user transactions that are ordered by their start times); one of the goals is to discover sequential patterns of data usage between correlated users (i.e., users with similar activity).

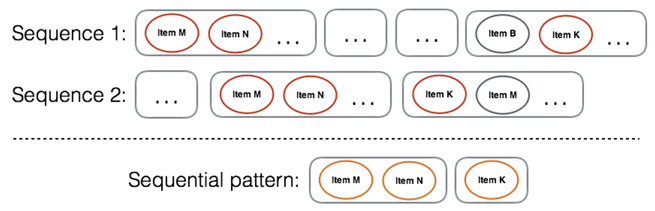


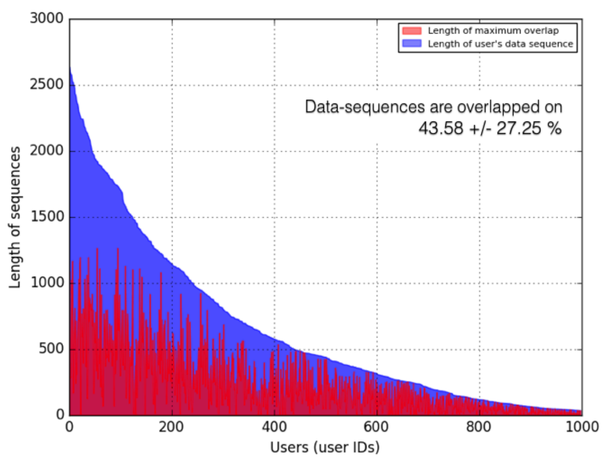
Figure 3. Sequential pattern representation.

User Activities

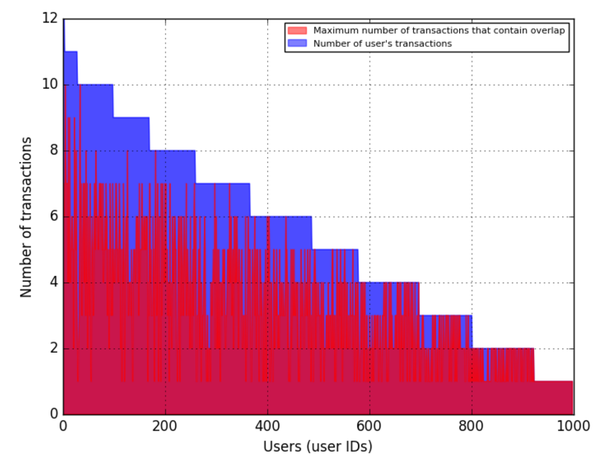
User data-sequences are significantly overlapped among 998 users (out of 1,267, that is 79% of users that have overlapping with others, and out of 1,597, that is 64% of total number of users), i.e., data-sequences share a significant portion of the timeline of used items with at least one other user.

Only sequences with a length greater than 5% of the average length (average sequence length is 700) were considered.

** The length of the sequence is a sum of lengths of its transactions.*



(a) Length of sequences per user



(b) Number of transactions per user

Chart 2. Maximum data usage overlap per user.

Conclusions

Initial analysis of data usage confirmed that a certain percentage of recommendations from similar users are actually followed despite the low quality of those recommendations. Thus user activity can follow some usage pattern of the group of similar users and be correlated with specific user interests. Deeper analysis, which included considerations for relationships between items in relation to users, presented correlations between users based on items' relations. Indeed data usage activity shows about 44% overlap for 64% of all relevant users, and about 17% of all relevant users had overlapping, but not significant data usage correlations. These findings indicate a strong correlation between users' data needs, validating our belief that a recommender system would enable users to find interesting data more readily and rapidly for their experiments.

References

1. F. Barreiro Megino, K. De, J. Caballero, J. Hover, A. Klimentov, T. Maeno, P. Nilsson, D. Oleynik, S. Padolski, S. Panitkin, A. Petrosyan, and T. Wenaus, "Panda: Evolution and recent trends in lhc computing," in Procedia Computer Science, vol. 66, pp. 439–447, 2015.
2. A. Luckow, M. Santcroos, O. Weidner, A. Merzky, P. K. Mantha, and S. Jha, "P*: A model of pilot-abstractions," CoRR, vol. abs/1207.6644, 2012.
3. F. Hernández del Olmo and E. Gaudioso, "Evaluation of recommender systems: A new approach," Expert Systems with Applications, vol. 35, no. 3, pp. 790–804, 2008.
4. T. Slimani and A. Lazzez, "Sequential mining: Patterns and algorithms analysis," International Journal of Computer and Electronics Research, vol. 2, no. 5, 2013.