PREDICTING DATA POPULARITY USING BAYESIAN NETWORKS OVER ATLAS GRID SITES

M. Titov, G. Záruba, A. Klimentov, K. De on behalf of the ATLAS Collaboration
Data replication over ATLAS grid sites that provides efficient data utilization by minimizing the number of unpopular datasets and increasing the number of demanded datasets for the certain grid sites.

\[ \text{Popularity}(D, S, t) = \frac{\text{NumJobs}(D, S, t)}{\text{NumJobs}(D, t)} \]
- Set of datasets - \( \{D_i\} \)
  - job input
    - Dataset
    - Container
    - Data pattern

- Set of users - \( \{U_j\} \)
  - job owner
    - Certificate DN
    - Email

- Set of sites - \( \{S_k\} \)
  - Panda Site
  - DDM Endpoint
The Model of Datasets Interactions (MDI) is a directed acyclic graph that shows interactions between dataset objects: how a particular dataset object relates to other dataset objects.

It represents the way of how dataset objects are transformed through time (from RAW to NTUP dataset objects).

*ATL-COM-GEN-2007-003, ATLAS Dataset Nomenclature*
ATLAS data can be classified into one of the following categories based on their importance and popularity (i.e., the Data Temperature Scale or the Data Importance State): \{hot, warm, cold, frozen, obsolete\}.

\[
\text{DIS} = \alpha \times D_{GP} = \alpha \times \left( \frac{U_D}{U} \right) \times \left( \frac{J_D}{J} \right)
\]

where \(U_D\) is the number of users who submitted jobs with the dataset \(D\), \(U\) is the total number of users, \(J_D\) is the total number of jobs with the input dataset \(D\), and \(J\) is the total number of jobs.
A Bayesian network is a graphical model that encodes probabilistic relationships among variables of interest. Condition parameters (set of variables $X=\{x_1\ldots x_n\}$) are represented as a directed acyclic graph by encoding assertions of conditional independence. Based on the chain rule for Bayesian networks, the probability for $X$ is defined as:

$$p(X) = \prod_{i=1}^{n} p(x_i | pa_i)$$

where $x_i$ denotes both the variable and its corresponding node, $pa_i$ denotes the variables representing the parents of node $x_i$. 
Belief-Network Structure

User → Dataset’s parameters → Job’s parameters → GRID Site → Site Probability Coefficient

Job’s parameters with discrete values:
- userId
- processing type

Dataset’s parameters with discrete values:
- project
- dataType
- configTags
Consider the problem of determining a belief-network structure $B_S$ that maximizes $P(B_S \mid D)$.

For a given database $D$, $P(B_S, D) \propto P(B_S \mid D)$, and therefore finding the $B_S$ that maximizes $P(B_S \mid D)$ is equivalent to finding the $B_S$ that maximizes $P(B_S, D)$.

$$P(B_S, D) = \int_{B_P} P(D \mid B_S, B_P) \times f(B_P \mid B_S) \times P(B_S) \times dB_P$$

where $D$ is a database of cases, $X$ is the set of variables represented by $D$, $B_S$ is a belief-network structure containing exactly those variables that are in $X$, $B_P$ is a vector whose values denote the conditional-probability assignments associated with belief-network structure $B_S$, and $f$ is the conditional-probability density function over $B_P$ given $B_S$.
Study of Bayesian Networks helps to identify relationships among a certain number of variables.

Further investigation of belief-network structures for the definition of the probability of data popularity will give more accurate results and will lead to further understanding of the data distribution behavior.
References


- D. Heckerman, “Bayesian Networks for Data Mining”, Journal of Data Mining and Knowledge Discovery, 1997
Thank you!
Plots (data from PandaDB for 2011 year)