Short summary of IG-KDD activity

G. Longo on behalf of many others

Poone, India, October 2011

- KDD has been recognised as a priority by several national projects (VAO, VO-India, etc)
- No VO funded activity for KDD at the moment
- Activity is proceeding mainly outside of IVOA and IVOA IG-KDD benefits from sporadic contributions from national projects (VO.it, VAO, VO-India, CANFAR, etc) and also from completely independent (outside of astronomy) efforts
- We are currently writing the document containing a list of recommendations for the IVOA WG's and this should close the exploratory part of our committments.

- N. Ball has completed an introductory guide to Data Mining which is available on the WIKI site. It is especially tailored for beginners, and we believe it will prove useful for people who are approaching DM methodologies for the first time

- A set of template data sets (quite large and complex enough) has been identified and uploaded to the WEB site …

- A VAO group led by A. Mahabal and C. Donalek (with contributions from many other people also outside of VAO) has completed an extensive benchmarking of the most commonly used DM packages and WAs.

- ORANGE, Rapid Miner, WEKA (AstroWEka) do not scale to very large data sets, even though some of them are «user friendly». DAME is the only truly scalable package (not so user friendly though)

- In order to deal with MDS, WAs are by far to be preferred with respect to client side solutions.

- There are however some problems which need to be urgently solved (they are more difficult to sampify, problems with visualization, etc.)

- M. Brescia, the VAO group and others, have identified the main requirements which a DM package (or WA) needs to match in order to be useful for the Vobs and have agreed on a series of priorities:

  1. Asynchronous access to data
  2. scriptable (STILTS like) in order to implement and preserve workflows.
  3. Built in pre-processing facilities
  4. Interfaceable with VO-Space
  5. SAMPified (problems in SAMPIfying WA's)

- In a data-centric environment, we need to minimize massive data flows on the network and move applications towards the data centers, especially if they are organized as KDD application warehouses.

- This of course requires a well-defined standardization process, in order to organize applications and SWFs in a fully interoperable way. This step is still “far from the coast” (literal translation from the Italian “*in alto mare* !”)

The end