WP8: Data Mining

Bob Mann
University of Edinburgh

Recap: overview of WP8 data mining

- Pick up where VOTECH DS6 finished
  - Both temporally and thematically
- Thematically
  - DS6 studied client-side data exploration
  - AIDA will focus on server-side data exploration
- Temporally
  - Kick-off meeting in July 2009: after completion of VOTECH
- Two proposed strands of work
  - Prototyping server-side data exploration (STILTS, Weka)
  - Prototyping distributed data mining (with ADMIRE project)

Update: progress since Trieste

- Focus on second strand
  - Prototyping distributed data mining
  - Collaboration with ADMIRE (FP7 project)
- Testbed: photometric quasar detection
  - SDSS (optical) plus UKIDSS (near-infrared)
  - Training data: spectroscopic quasar catalogue
- VOTECH: Brian Walshe did this one way
  - Download data to local workstation and analyse
- AIDA: try and do this in a distributed infrastructure

ADMIRE

- FP7 project
  - Partners in UK, Austria, Spain, Slovenia, plus Fujitsu Europe
- To develop consistent and easy-to-use framework for integration and mining of distributed data
  - Builds on OGSA-DAI: leading data access and integration technology in e-Science/Grid world
  - Developing architecture *and* deploying testbed (using Weka)
- Assumed model for distributed data centres resembles VO

Progress to date

- Published SDSS and UKIDSS with OGSA-DAI services
  - Worked well: some minor configuration issues only
- Problems with setting up Distributed Query Processor service to effect join between the two databases
  - DQP tries to load one side of join in memory – won't scale!
  - DQP's subset of SQL not wholly consistent with ADQL
- Working with OGSA-DAI team to fix these two problems
- Also working on expressing quasar search operation in ADMIRE's data mining language

Plans

- Plan to complete quasar detection scenario
  - OGSA-DAI and ADMIRE teams very keen – MSc and PhD students also involved
  - Want to implement TAP within OGSA-DAI, to see if can deliver general distributed query service for VO
- Also trying to identify effort needed to complete other strand of proposed work
  - Prototyping server-side data exploration with Weka