e-Science and Data Mining: Workshop Introduction

Bob Mann

Institute for Astronomy & NeSC
University of Edinburgh

Outline

- The groups behind this workshop
  - e-Science Data Mining Special Interest Group
  - SC4DEVO “international sister project”
- Workshop goals
- Workshop programme

e-Science Data Mining SIG: Motivation

- Many disciplines have a data deluge
- Data integration major part of e-science
- What to do with data once integrated?
  - Many standard analyses won't scale
  - Much new science made possible
- Look to data mining as a way of separating wheat from chaff

"Scientific Data Mining, Integration & Visualization"

- NeSC, October 2002
- www.nesc.ac.uk/talks/sdmiv/report.pdf
- Participants from wide range of domains
  - astronomy, atmospheric science, bioinformatics, chemistry, digital libraries, engineering, environmental science, experimental physics, marine sciences, oceanography, and statistics…plus CS researchers and software engineers
- But a common set of problems

Problems from SDMIV

- Lots of DM packages, lots of data formats
- How to mine distributed data sources?
- How to mine large data volumes?
  - and high-dimensional datasets
- How to do data exploration?
  - Coupling data mining and visualization
- How to work iteratively & interactively?
  - Tracking provenance, building workflows...

Goals of the SIG

- Forum for e-science data miners
  - Application scientists, algorithm writers, infrastructure developers
- Identify requirements on infrastructure and algorithms from science drivers
  - Are there generic problems to solve?
    - Can there be an OGSA-DAI for data mining?
- Stimulate/foster R&D where needed

SIG activities to date

- Set up steering group
  - N. Adams (ICTSM), J. Austin (York), Y. Guo (ICSTM), R. Mann (Edin.), R. Nichol (Port.), A. Shepherd (Birkbeck), A. Storkey (Edin.)
- Mini-workshop at All Hands Meeting
- Review of data mining in UK e-Science
  - Questionnaire [Amos Storkey talk]
  - This workshop

Service Composition For Data Exploration in the Virtual Observatory (SC4DEVO)

- e-Science “international sister project”
  - One of first four funded
- Data exploration in astronomy
  - Data mining and visualization
- Organise four workshops in 2004/2005
  - Two astronomy-specific, two general
  - SC4DEVO-1: Caltech in July 2004
  - SC4DEVO-2: now (Tue-Thu)

Workshop Goals

- Forum for discussion of issues relating to data mining in an e-science context
- Contribution to *esdm-sig* review
  - Supplementing questionnaire results
- Plan for *esdm-sig* operations
  - What should it do?
  - How often should it meet?
  - Etc.

Workshop Programme

- Introduction & context
- Application drivers
- e-Science data mining issues:
  - Scalability
  - Web/Grid service implementations

Tuesday

  - Mining distributed data sources
  - Text mining
- Discussion
- Report back and wrap-up

Wednesday