SAM: Smart Assistant for Earth Science Data Mining

PI: Rahul Ramachandran
Co-I: Peter Fox, Chris Lynnes, Robert Wolf, U.S. Nair

itsc UAH The University of Alabama in Huntsville

1. Abstract

Scientific data mining is a very powerful means for automated knowledge extraction from the ever-increasing volumes of science observations and model output data available. NASA's Second Data Mining Workshop found that maturing data mining techniques show "potential for significantly expanding the scientific understanding of NASA's Earth science data." However, this type of tool has generally been difficult for domain scientists and students to fully exploit without extended learning curves. And even data mining specialists may not be familiar with the full range of components in a mining toolkit, so potentially useful mining strategies may be ignored. To facilitate exploitation of these promising techniques by the increasingly IT-sophisticated NASA Earth science community, the University of Alabama in Huntsville is leading a collaborative team to leverage Semantic Web technologies to build a **Smart Assistant for Mining (SAM)** and to deploy it for use at two data centers. This project will *reuse an existing toolkit of data mining web services* designed specifically for the analysis of NASA data in a web-based, service-oriented architecture. It will also *leverage and extend an initial ontology describing data mining* services, with links to other ontologies describing the Earth science problem domain and relevant data sets. The new *SAM user interface tool, which integrates semantic reasoning into a traditional workflow composer*, will allow users to discover available data and services, assist users in composing mining workflows, and invoke them to perform the desired analysis. SAM will provide a useful tool to assist researchers in creating data analysis and mining workflows for targeted Earth science problems in the *Climate Variability and Change* Science Focus Area. It will also position these services for integration with many other science data services in the Semantic Web Services context, pointing the way toward increased science return from NASA data.

2. Motivation

2.1 Science Need

- Study the impact of natural iron fertilization processes such as dust storms on plankton growth and subsequent DMS (dimethyl sulfide) production
  - Plankton plays an important role in the carbon cycle
  - Plankton growth is strongly influenced by nutrient availability (iron and phosphorus)
  - Dust deposition is an important source of iron over oceans
  - Satellite data is an effective tool for monitoring the effects of dust fertilization
- Analysis entails
  - Mine MODIS L1B data for dust storm events and identify the swath of area influenced by the passage of the dust storms.
  - Examine correlations between fertilization, plankton growth and DMS production

2.2 Current Analysis Process

- Problem: MODIS aerosol products don't provide speciation
- To address this, scientists must:
  - Locate and download all the data to their local machine
  - Write code to classify and detect dust accurately [3-4 month effort]
  - Write code to classify and detect other dust aerosols [3-4 month effort]
  - Write code to segment the detected region in order to account for advection effect and correlation coefficient [2 months effort]

2.3 Analysis with SAM

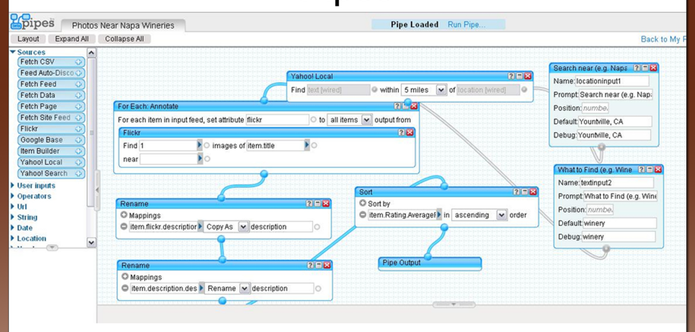
- Create a workflow to perform classification using many different state of the art classifiers (services) on distributed data
- Create a workflow to segment detected regions using image processing services on distributed data

Bottom line:

- Scientist does not have to write all the code to perform the analysis
- Can compose workflows that utilize distributed data/services
- Can share the workflow with others to collaborate, reuse and modify

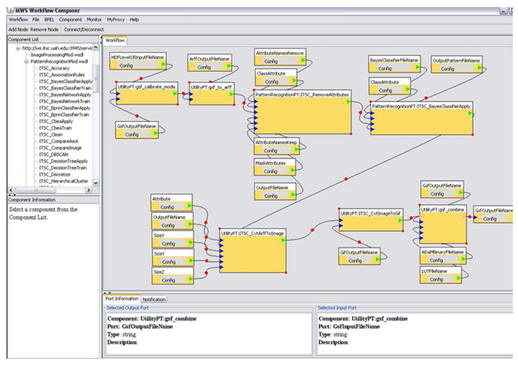
3. Conducting Science using the Internet: Service Mashups

3.1 Mash-ups Example: Yahoo Pipes

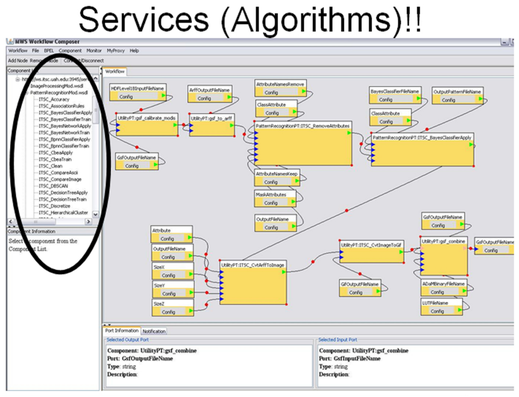


The use of mash-ups to process and filter information is common in the private industry

3.2 Data Mining in the 'new' Distributed Data/Services Paradigm



3.3 Problem: Too many choices in Services (Algorithms)!!



- And that's only part of the toolkit
- ADaM-IVICS toolkit has over 100+ services

4. Use of Semantics

4.1 What is Semantics? And How can it Help?

- Semantics (from SAM's context) is capturing valuable information from user/scientist/algorithm developer in a machine *readable, process-able and interpretable* form!
- Information captured will include:
  - Function of algorithms - what they do
  - Context of use
  - Input and output data requirements
  - Pre- and post- processing requirements
- Semantics are put in the interfaces, between the layers in the processing, in the architecture to assist the user in during the analysis process in a seamless manner

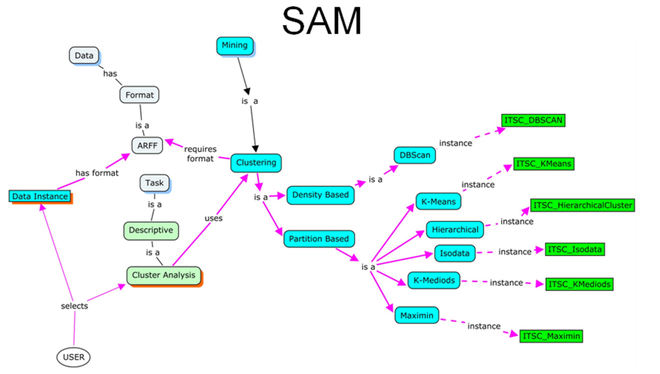
4.2 SAM Objectives

- Improve usability of Earth Science data by existing data mining services for research, by incorporating semantics into the *workflow composition* process.
  - Semantic search capable of mapping a conceptual task to specific services
  - Assistance in mining workflow composition
  - Verification that services are connected in a semantically correct fashion

4.3 Ontology and Knowledge Engineering

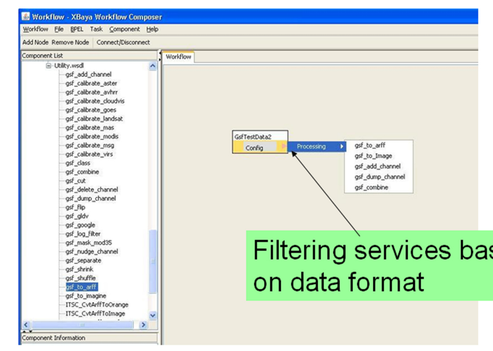
- An ontology is a formal, explicit specification of a shared conceptualization. An ontology can be viewed as a collection of multiple taxonomies and thesauri in a formal structure used by machines to infer and reason.
- Capturing the knowledge into an ontology use cases to bound the domain and detailed analysis to include only the useful concepts. Knowledge engineers and domain experts iterate until there is convergence.

4.4 Envisioned Ontology Use in SAM

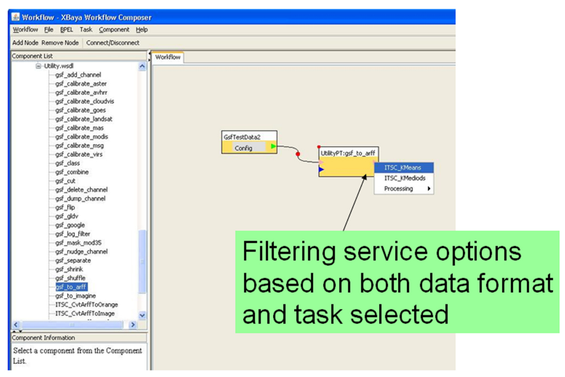


5. Mining Workflow Composition Example using SAM

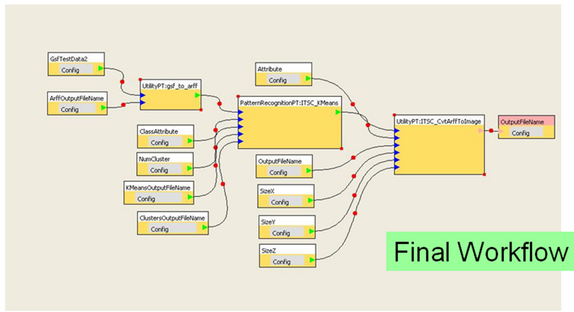
Semi-automated Workflow Composition



Semi-automated Workflow Composition



Semi-automated Workflow Composition



6. Impact and Relevance

- The Second NASA Data Mining Workshop found that *"Current data analysis methods employed in Earth science are no longer adequate for dealing with the complexity, size, and novelty of NASA's 21st century data resources,"* but that maturing data mining techniques show *"potential for significantly expanding the scientific understanding of NASA's Earth science data."* However, the mining process can also be complicated, even for mining experts.
- This project will infuse semantic web technologies into a data mining environment to provide intelligent assistance to the user, thus greatly increasing the usability of data mining web services. The Smart Assistant for Mining developed here will provide a significant shift for data mining in the direction of a "simple and intuitive" application.
- While SAM will target a specific science research domain, the toolkit of data mining web services, associated interoperable service ontologies, and user interface tool are more generally applicable beyond Earth science, and can be used by other NASA communities and any others requiring automated methods of knowledge extraction