Contribution ID: **125** Type: **not specified**

Interactive Information Extraction based on Distributed Data Management for D-Grid Projects

Tuesday, 27 March 2012 14:30 (30 minutes)

URL

www.wisnetgrid.org

Overview (For the conference guide)

The current D-Grid infrastructure primarily covers resource management and exchange at the data level supporting mainly technical resources such as computational capacity, data transport networks, storage resources, and management
software.
The WisNetGrid project(www.wisnetgrid.org) aims to broaden the focus of resources exchange towards the actual content, such as research and production data, to enable interdisciplinary usage. To achieve this goal, resource sharing is supported on different abstraction layers. First, we create an information layer by providing an universal interface to access data on the grid independent of the underlying grid storage system. Second, at the knowledge layer, we offer highly interactive knowledge extraction and management tools that can also take advantage of a community's grid resources.

Description of the Work

The current D-Grid infrastructure primarily covers resource management and exchange at the data level for individual communities and comprehends computational capacity, data transport networks, storage resources, and management software. The actual content, such as research and production data, documents, images, domain specific know-how etc., are generated and maintained within the individual communities for specific tasks. This limits their interdisciplinary usability for research and business ventures.
The WisNetGrid project(www.wisnetgrid.org) aims to overcome these limitations by supporting resource sharing on different abstraction levels. To enable
transparent access to the sources, we establish an uniform information layer on top of widely used grid middlewares providing access by a generalized method
to distributed heterogeneous data across grid resources. This allows the constitution of a knowledge layer containing services for knowledge generation and management on top of the information layer by applying information extraction methods to gather higher semantic knowledge from the information available.
Typical domain-independent information extraction methods, which go beyond named entity recognition, are either imprecise or computationally expensive, and even the most precise systems need human supervision for critical applications. Additionally, extraction methods often work iteratively. Thus, catching mistakes early on has significant positive impacts on the performance and can reduce the overall need for human corrections. Our goal is to provide extraction services that tightly integrate user feedback into the extraction process while minimizing human effort needed to obtain optimal results. To achieve this goal, an interactive

extraction system is provided, which learns during the extraction process from user feedback. Our approach also enables the use of grid middleware to transfer computation intensive tasks to the grid.

Conclusions

We provide tools to enable content oriented resource sharing on two levels.
First on the data level, by providing uniform access to different grid-based data storage backends. Second on the knowledge level, by providing interactive grid-based information extraction methods. Our information extraction methods can work automatically, given adequate domain knowledge, yet they can also incorporate human feedback directly into the iterative extraction process, thus learning globally by human feedback. As large scale information extraction is computationally very expensive, our architecture allows to use grid resources for computational tasks, if such grid resources are available to the user community, while preserving the user's control over the extraction process.

Impact

While we feel that many grid communities can benefit from a cross-community data and knowledge exchange in general, the WisNetGrid project works closely together with two particular communities from different domains to gather first user experiences during development. One of these partner communities is represented by the TextGrid project (www.textgrid.de), another grid project within the D-Grid community, which aims to support German scientists from the humanities. First, the TextGrid data repository is integrated into WisNetGrid's
information access layer. Second, the humanities community can make advantage by using extraction tools that support daily work, such as text annotations,
but still allow a strong integration of manual adjustments into the process. The second project partner is from the landscape architecture community. While the TextGrid community mainly deals with different text forms, the data of this community is far more diverse. Hence, our uniform data access methods may be of help even within the community. As the community works with
various information sources of different kind and needs to connect information from several sources for particular tasks, our information extraction methods
can help establish these links and generate some higher level understanding of the vast data collections at hand.

Primary author: Dr JÄKEL, René (Technische Universität Dresden)

Co-authors: Dr SCHULLER, Bernd (Jülich Supercomputing Centre); Mr HÜNICH, Denis (Technische Universität Dresden); Mr DAIVANDY, Jason Milad (Jülich Supercomputing Centre); Dr HOSE, Katja (Max-Planck-Institute for Informatics); Mr JUNGHANS, Martin (Karlsruhe Institute of Technology); Mr HARMS, Patrick (SUB Göttingen); Dr SCHENKEL, Ralf (Saarland University and MPI for Informatics); Mr METZGER, Steffen (Max--Planck-Institute for Informatics); Dr AGARWAL, Sudhir (Karlsruhe Institute of Technology)

Presenters: Dr JÄKEL, René (Technische Universität Dresden); Mr METZGER, Steffen (Max-Planck-Institute for Informatics)

Session Classification: Community-tailored Services