

Analyzing neuroimaging data in Grid environments using relational databases and a dedicated workflow language

Hasson, U., Andric, M., Kenny, S., Wilde, M. J., & Small, S. L.
Departments of Neurology and Psychology and Computation Institute
The University of Chicago

Neuroimaging research imposes substantial demands on computational infrastructures. Already, these infrastructures need to support management of massive amounts of data while at the same time affording rapid analysis, access to highly specific subsets of data, and secure remote access for collaborators. We have recently described an architecture that is used in practice to achieve these goals, which relies on distributed database management systems (DBMS; Hasson et al, 2008). A central component of this system relies on DBMSs to store neuroimaging data and metadata in a relational database. This allows for extracting highly specific subsets of data via SQL queries, which are then statistically analyzed via GRID accessible computing nodes. Here we present two recently introduced and central components of this system: (1) an extension of this architecture that utilizes the SWIFT system to enable analysis of DBMS-stored neuroimaging data on GRID sites (e.g., TeraGrid), and (2) an interface between this system and commonly used neuroimaging GUIs (AFNI, SUMA) that makes it possible to immediately graphically depict the results of databases queries.

Because the system stores neuroimaging data in DBMSs, these data can be accessed via standard Internet Protocols, thus enabling distributed analysis of remotely stored data. GRID-based computing can leverage this capacity for parallel computing as GRID sites host tens of thousands of computing nodes. However, such a scale of computation calls for special mechanisms to identify and submit jobs to GRID sites, establish provenance for each analysis, and return the results to the user. We will demonstrate how the SWIFT system achieves these goals in the context of distributed analysis of neuroimaging time-series analysis. In particular, a researcher can use SWIFT to issue multiple database queries from multiple GRID sites, extract time-series of interest, analyze those data, and retrieve the results.

For neuroscientists, the information that results from such analyses is often most meaningful when displayed graphically in the form of statistical parameter maps (SPMs). In collaboration with the AFNI development group we have recently established mechanisms for plotting the results of SQL queries. We will show how this interface allows researchers to visualize highly specific subsets of data without needing to retrieve the entire dataset.

Concrete example: A scientist wishes to establish the impact of a certain filtering parameter on the results of a neuroimaging time series analysis. The scientist therefore establishes a parameter sweep where the effect of various filter combinations are determined. Each job (defined as a certain filter specification) has a processing time of 2 hours, and there are 50 filter combinations. Using the SWIFT system, each job is submitted to a computing node located at a Grid sites. The nodes retrieve data from a database via SQL queries, process it using the assigned filtering parameters, plot the results graphically on a cortical (2D) representation, and save the result as an image. The result images are returned to the user via SWIFT mechanisms.

Reference

Hasson, U., Skipper, J. I., Wilde, M. J., Nusbaum, H. C., & Small, S. L. (2008). Improving the analysis, storage and sharing of neuroimaging data using relational databases and distributed computing. *Neuroimage*, 39(2), 693-706.