GridR: Distributed Data Analysis using R

Dennis Wegener, Stefan Rüping and Michael Mock

Fraunhofer Institute for Intelligent Analysis- and Information Systems
Schloss Birlinghoven
53754 St. Augustin, Germany
{ dennis.wegener, stefan.rueping, michael.mock }
@iais.fraunhofer.de

Abstract. In the last couple of years, the amount of data to be analyzed in different areas grows rapidly. Examples range from natural sciences (e.g. astronomy or particle physics), business data (e.g. a high increase use data volume is expected by the use of RFID technology), life sciences (such as high-throughput genomics and post-genomics technologies) or data generated by normal users on the internet (see Google, Youtube, etc.). The enormous growth of the amount of data is complemented by advances in distributed computing technology enabling the data analyst to handle this amount of data in reasonable time. Two main streams of current distributed technology development and research are particularly useful in this respect: the grid technology is aiming at making data stores and computing facilities which are geographically widely spread available for a common, global data analysis. The other stream of development is cluster-based computing which transforms large amounts of standard computers into high-performance computing bases.
However, even if the above mentioned advances in distributed computing technology make available the computing and storage resources for handling large amounts of data, they introduce another level of complexity in the system, such that the traditional data analyst, with a strong background in statistics and application domain knowledge, might be overwhelmed by the complexity of the underlying distributed technology. For instance, an application developer using R might not be interested in any details of how web services are built. Therefore, ongoing research aims at bridging the gap between advanced distributed computing technology and traditional statistical software.
The Advancing Clinico-Genomics Trials on Cancer project (ACGT) aims at providing a data analysis environment that allows the exploitation of an enormous pool of data collected in European cancer treatments. In the context of this project, the GridR package was developed, which was one of the first attempts to connect R to a grid environment - to grid-enable R.

Keywords: R statistical language, Grid, GridR, ACGT

Acknowledgements. The authors gratefully acknowledge the support of the ACGT project that is funded by the European Commission (FP6/2004/IST-026996).

References

1. Dennis Wegener, Thierry Sengstag, Stelios Sfakianakis, Stefan Rüping and Anthony Assi. GridR: An R-based grid-enabled tool for data analysis in ACGT clinico-genomic trials. In: Proceedings of the 3rd International Conference on e-Science and Grid Computing (eScience 2007), Bangalore, India.
2. Stefan Rüping, Stelios Sfakianakis and Manolis Tsiknakis. Extending Workflow Management for Knowledge Discovery in Clinico-Genomic Data. In: From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences, Proceedings of HealthGrid 2007, pp. 183-193, IOS Press, 2007.
3. Vlado Stankovski, Martin Swain, Valentin Kravtsov, Thomas Niessen, Dennis Wegener, Joerg Kindermann, and Werner Dubitzky. Grid-enabling data mining applications with DataMiningGrid: An architectural perspective. Future Generation Computer Systems Journal, 2007.
4. Vlado Stankovski, Martin Swain, Valentin Kravtsov, Thomas Niessen, Dennis Wegener, Matthias Röhm, Jerney Trnkoczy, Michael May, Jürgen Franke, Assaf Schuster and Werner Dubitzky. Digging Deep into the Data Mine with DataMiningGrid. IEEE Internet Computing, accepted for publishing in 2007.
5. Dennis Wegener and Michael May. Extensibility of Grid-Enabled Data Mining Platforms: A Case Study. In Proc. of the 5th International Workshop on Data Mining Standards, Services and Platforms, KDD 2007, pages 13--22, San Jose, USA, August 2007. ISBN 978-1-59593-838-1.