Poster K-28

Mining Provenance Logs from Bioinformatics Workflows

Authors:
Sérgio Manuel Serra da Cruz (*NCE-UFRJ*)

Short Abstract: Mining bioinformatics logs allow e-scientists to gain insight into the data and come up with new scientific hypotheses. To explore this opportunity we present PM-BioWSLogA, a framework which relies on mining provenance logs of scientific workflows based on Web Services log data captured from bioinformatics in silico experiments

Long Abstract:
Introduction

Scientific workflows generate huge amounts of data, mandating rich and descriptive metadata about the data in order to make sense of it and reuse it. One kind of metadata is provenance, it can be used for various purposes, such as: estimate data quality, audit trail, copyright ownership. For the purposes of this paper, we define data provenance as information that helps determine the derivation history of a data product, starting from its original sources.
In bioinformatics environments, program composition is a frequent operation, requiring complex management. Scientist faces many challenges when building an experiment: finding the right programs to use, the adequate parameters to tune, managing external data, and last but not least, building or reusing Web services workflows.

Mining Web Services Provenance Logs

Recording the execution context of Bioinformatics Web services experiments into proper logs is an important phase of a scientific workflow execution. A concrete usage of a Web services log is to help e-scientists to avoid redundant efforts when repeating experiments. Bioinformatics tools have several input parameters, which can modify the behavior of their algorithm, and consequently modify its results. Another issue regarding service monitoring in bioinformatics environments is that e-scientists need efficiency. Some queries consist of hundred of sequences at a time and can take several days to run. Besides, due to the confidential nature of experiments, may be useful to keep of track of errors and securities issues. Finally, e-scientists should also keep track of data sources and program inputs and outputs they have used if they are willing to reproduce the same experiments in future occasions. In order to solve those issues, we previously presented an architecture named BioWSLogA which supports at runtime a flexible log generation without changing the code of existing services, generating XML logs repositories about services execution[1]. It can store complex and heterogeneous data structures and, at the same time, describe them through encapsulated metadata.

In spite of this facility, e-scientists still need to understand the results of bioinformatics experiments and design/manage scientific workflows, so we propose the use of process

mining techniques to help them to extract knowledge from provenance logs. This approach plays two important roles in exploratory data analysis, where analytical processing can help them to build up an understanding of the content of their in silico experiments and also supports them to perceive discrepancies between workflows processes.
The contributions of PMBioWSLogA (Provenance-Mining Bio Web Services log Architecture) are: It provides perception of biological experiments towards e-scientist auditing, supporting new ranges of experimental strategies; it supports refined investigations about services and data quality, addressing administrative issues like performance, security and services availability.

A Web Services Provenance Log Mining Architecture

Our proposal outlines scientists needs for gathering knowledge about their in silico experiments. The architecture's prototype can be used to amplify the perception of rules, patterns, regularities and behaviors. It aids scientists to visualize four different aspects of Bioinformatics experiment data sets, such as: the usage of suitable experiment parameters; a simple view of Bioinformatics Web services composition; an easy way to audit and track the Web services and data utilization; a feasible way to keep track of data provenance.

PMBioWSLogA act as plugin of BioWSLogA framewrork which can uses both Web server and XML Web services logs. It is a multilayered architecture capable to deal with provenance data. The integration layer is a set of programs used to prepare data for further processing. For instance: extraction, cleaning, transformation and loading. This layer uses XML Schemas to feed the data repository. The sessionization layer is used to tie the instances of Web services to Web sessions and to workflow user. This layer is important to investigate the usage of the Web services composition used through users sessions.
The Database layer is a repository which encode the relationships between the workflow input and output constituents of a bioinformatics experiments and Web services invocations. It also stores provenance data with different granularities, from preprocessed logs files, scientists sessions, to information's about the Web services execution times. We also defined a metadata model used to store attributes about data and services provenance some of which is created manually beforehand, and some of which is generated by the workflow invocation.
The ProvenanceMiner Engine Layer is a data mining engine and is in charge of bulk loading XML data from database, executing SQL commands against it and execute the provenance mining algorithms, such as the ones used to compare the results of Bioinformatics parameters, collate actual services composition patterns to expected patterns, or more generally, to compare and track services and data products utilization.

Conclusion

As far as we are concerned, there are few initiatives of mining bioinformatics web services provenance data and services composition logs. PMBioWSlogA allows easily provenance retrieval, it is being tested with data originated by a collection of real world bioinformatics Web services. We are involved in refining the architecture, which was implemented as java prototype using TomcatAxis as SOAP engine and PostgresSQL as data repository. Future work will involve tight integration with traditional techniques from disciplines like process mining and text mining. Our ultimate goals lies on the enhancement of mining algorithms in

order to bring the power of visualization technology to e-scientists desktops and allow a better, faster and more intuitive and cognitive exploration of experimental biological dataset resources. We are also enhancing the provenance metadata model in order to allow manual and automatic annotation with additional semantic information from different workflows enactment engines.

References

Cruz, S. M. S.,et al.. “Monitoring Bioinformatics Web services Requests and Responses through a Log Based Architecture”,2005 In XXXII SEMISH-SBC, São Leopoldo, 2005 pp.1787-1801.

Aslst, W. M. P, et al. “Workflow Mining: Discovering process models from event logs”. IEEE Transactions on Knowledge and Data Engineering. 2004, Vol.16(9), pp.1128-1142.

Bose, R., Frew, J. Lineage Retrieval for Scientific Data Processing:A SurveyACM Computing Surveys, Vol. 37(1), March 2005, pp.1–28.