Integrating Text Mining into Bioinformatics Workflows
Neil Davis, George Demetriou, Robert Gaizauskas, Yikun Guo, Ian Roberts
Department of Computer Science, University of Sheffield

Major application area

Text processing for bioinformatics, specifically the integration of text retrieval, clustering and terminology tagging into biomedical workflows.

Demonstration Details

In this demonstration we show how we have integrated text mining web services into the myGrid bioinformatics workbench [1]. myGrid supports the creation and execution of bioinformatics workflows via the Taverna workflow editor [2] and the associated Freefluo workflow enactor [3]. As steps in such workflows users may stipulate that literature related to BLAST or Swissprot records generated from sequence alignment operations be retrieved, clustered and made available for browsing as an integral part of the workflow. Workflows are initiated from a portal and text mining results are viewed via a separate portlet in the portal, from which users can carry out follow-on searches of various sorts. The demonstration illustrates:

1. Various text mining web services to cluster document collections and automatically annotate terminology. For example we have a web service to annotate multiword sequences in biomedical texts with term classes based on their occurrence in domain specific resources such as UMLS, HUGO, Uniprot and so on.
2. How such services can be integrated into bioinformatics workflows using the myGrid workbench. Specifically we demonstrate a workflow designed to support a scientist investigating the genetic basis of Williams syndrome.
3. How a richly functioned user interface deployed via a portlet within the myGrid portal can be used to deliver text mining results to an end user.

The net result is a flexible system whereby research biologists can have relevant literature delivered, clustered and annotated, as part of a standard homology search workflow. Furthermore not only is the overall demonstrator useful in reducing the effort in accessing the relevant literature, but its components are available separately for re-use/reconfiguration in other applications.

Platform

The text processing services at Sheffield University are implemented as platform independent web service applications. The web services are implemented using the PERL SOAP::Lite library [4] and run on a Sun-Fire-280R server. The web services have been extensively tested using clients coded using the Perl SOAP::Lite and Java AXIS [5] libraries on a variety of platforms. The portal is implemented using Gridsphere [6] and the client browser using Java Swing.

Availability

Use of the text processing web services is unrestricted, but subject to a fair use policy. Other software available from cited sources according to their policies.

Where or how to obtain the software

WSDL interfaces are available at http://wsdlinterfaces.don.shef.ac.uk. Portal interface available at http://don.dcs.shef.ac.uk:8098/gridsphere/. For further information on access to the text processing web services please contact Dr Neil Davis at ndavis@dcs.shef.ac.uk

References

1. myGrid : http://www.mygrid.org
2. Taverna: http://taverna.sourceforge.net/
3. Freefluo: http://freefluo.sourceforge.net/
4. SOAP::Lite: http://www.soaplite.com/
4. AXIS: http://ws.apache.org/axis/
5. Gridsphere: http://www.gridsphere.org/.