ClowdFlows: A Cloud Based Scientific Workflow Platform

Janez Kranjc[1,2], Vid Podpečan[1,2], and Nada Lavrač[1,2,3]

[1] Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia
[2] Jožef Stefan International Postgraduate School, Ljubljana, Slovenia
[3] University of Nova Gorica, Nova Gorica, Slovenia
{janez.kranjc,vid.podpecan,nada.lavrac}@ijs.si

Abstract. This paper presents an open cloud based platform for composition, execution, and sharing of interactive data mining workflows. It is based on the principles of service-oriented knowledge discovery, and features interactive scientific workflows. In contrast to comparable data mining platforms, our platform runs in all major Web browsers and platforms, including mobile devices. In terms of crowdsourcing, ClowdFlows provides researchers with an easy way to expose and share their work and results, as only an Internet connection and a Web browser are required to access the workflows from anywhere. Practitioners can use ClowdFlows to seamlessly integrate and join different implementations of algorithms, tools and Web services into a coherent workflow that can be executed in a cloud based application. ClowdFlows is also easily extensible during run-time by importing Web services and using them as new workflow components.

Keywords: cloud computing, data mining platform, service-oriented architecture, web application, web services, scientific workflows.

1 Introduction and Related Work

The paper presents ClowdFlows, an open cloud based platform for composition, execution and sharing of data mining workflows. It was designed to overcome deficiencies of existing comparable data mining platforms while retaining their useful features along with new features, not provided in comparable software. The presented platform is distinguished by these important features — a visual programming user interface that works in a Web browser, a service-oriented architecture that allows using third party services, a social aspect that allows sharing of scientific workflows, and a cloud-based execution of workflows. ClowdFlows is accessible online at http://clowdflows.org.

Tools for composition of workflows most often use the visual programming paradigm to implement the user interface. Notable applications that employ this approach include Weka [1], Orange [2], KNIME [3], and RapidMiner [4]. The most important common feature is the implementation of a *workflow canvas* where workflows can be constructed using simple drag, drop and connect

P. Flach et al. (Eds.): ECML PKDD 2012, Part II, LNCS 7524, pp. 816–819, 2012.

operations on the available components. This feature makes the platforms suitable also for non-experts due to the representation of complex procedures as sequences of simple processing steps (workflow components).

In order to allow distributed processing, a service oriented architecture has been employed in platforms such as Orange4WS [5] and Taverna [6]. Utilization of Web services as processing components enables parallelization and remote execution. Service oriented architecture supports not only distributed processing but also distributed software development.

Sharing of workflows is a feature already implemented at the *myExperiment* website [6]. It allows users to publicly upload their workflows so that they are available to a wider audience and a link may be published in a research paper. However, the users that wish to view or execute these workflows are still required to install specific software in which the workflows were designed.

Remote workflow execution (on different machines than the one used for workflow construction) is also employed by RapidMiner using the RapidAnalytics server [4]. This allows the execution of workflows on more powerful machines and data sharing with other users, with the requirement that the client software is installed on the user's machine, which is a deficiency compared to our ClowdFlows solution.

With the described features in mind, we designed ClowdFlows, a platform which implements these features, but facilitated by enabling their access from a Web browser. The advantage of this approach is that no installation is required and that workflows may be run from any device with a modern Web browser, while being executed on the cloud. Apart from software and hardware independence, the implementation as a cloud based application takes all the processing load from the client's machine and moves it to the cloud where remote servers can run the experiments with or without user supervision.

2 The ClowdFlows Platform

ClowdFlows consists of the workflow editor (the graphical user interface) and the server side application which handles the execution of the workflows and hosts a number of publicly available workflows. The editor is implemented in HTML and JavaScript and runs in the client's browser. The server side is written in Python and uses the Django Web framework[1].

The workflow editor shown in Figure 1 consists of a *workflow canvas* and a *widget repository*. The widget repository is a list of all available workflow components which can be added to the workflow canvas. The repository includes a wide range of default widgets. Default widgets include Orange's implementation of classification algorithms, which were imported seamlessly as Orange is also implemented in Python. Weka's implementations of algorithms for classification and clustering, which we have wrapped as Web services, are also included in the widget repository by default. The widget repository may also be expanded by anyone at any time by importing Web services as workflow components by

[1] More information on Django can be found at `https://www.djangoproject.com/`

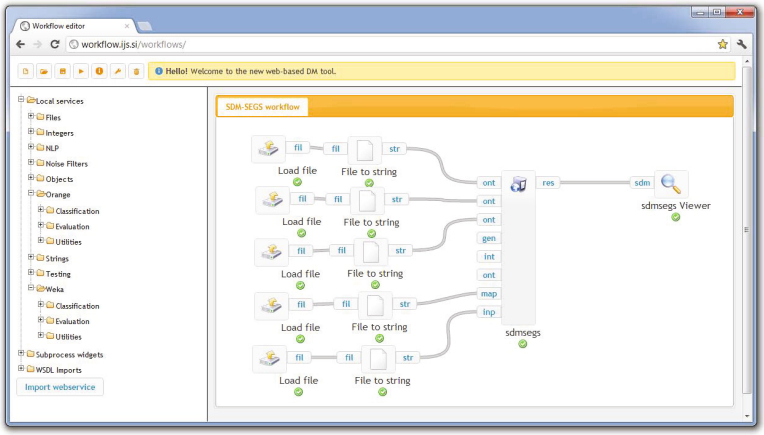

Fig. 1. A screenshot of the workflow editor with a semantic data mining workflow loaded [7]. This workflow can be accessed at the following address: http://clowdflows.org/workflow/104/.

entering a URL of a WSDL described Web service. The operations of the Web service are converted to widgets and their arguments and results are converted to inputs and outputs of these widgets. The repository also includes a set of process control widgets which allow the creation of meta-workflows (a workflow of workflows) and loops (meta-workflows that run multiple times), and a set of widgets for results visualization.

The server side consists of methods for the client side workflow editor to compose and execute workflows, and a relational database of workflows, widgets and data. The methods for manipulating workflows are accessed by the workflow editor using a series of asynchronous HTTP requests. Each request is handled by the server and executes widgets if necessary, saves the changes in the database and returns the results to the client. The server can handle multiple requests at a time and can simultaneously execute many workflows and widgets.

The data are stored on the server in the database. The platform is database independent, but MySQL is used in the public installation. The data can be passed as pointers or as the data itself, depending on the widget or Web service implementation.

A repository of public workflows which also serve as use cases for this demo is available in ClowdFlows and can be accessed athttp://clowdflows.org/existing-workflows/. Whenever the user opens a public workflow, a copy of that workflow appears in her private workflow repository in the workflow editor. The user can execute the workflow and view its results or expand it by adding or removing widgets. The user may again share her changes in the form of a new public workflow. Each public workflow can also be accessed through a unique address which is provided for the user to be shared through the workflow editor.

3 Conclusion and Further Work

We have developed an open and general cloud based platform for data mining, which employs service-oriented technologies, and is ready to be used in any data analysis scenario.

The social aspect of the platform provides a way for users to share their work easily as each workflow can be accessed through a simple address. This allows researchers to distribute their work and results with ease, since ClowdFlows provides cross-platform functionality. The platform is also suitable for non-experts and beginner data mining enthusiasts because of its intuitive and simple user interface.

We are currently working on adding support for mining continuous data streams from the Web (e.g. RSS feeds). We will also continue to add new widgets for specialized machine learning and data mining tasks, focusing on text mining.

Acknowledgments. This work was supported by the FP7 European Commission projects "Machine understanding for interactive storytelling" (MUSE, grant agreement no: 296703) and "Large scale information extraction and integration infrastructure for supporting financial decision making" (FIRST, grant agreement 257928).

References

1. Witten, I.H., Frank, E., Hall, M.A.: Data Mining: Practical Machine Learning Tools and Techniques, 3rd edn. Morgan Kaufmann, Amsterdam (2011)
2. Demšar, J., Zupan, B., Leban, G., Curk, T.: Orange: From Experimental Machine Learning to Interactive Data Mining. In: Boulicaut, J.-F., Esposito, F., Giannotti, F., Pedreschi, D. (eds.) PKDD 2004. LNCS (LNAI), vol. 3202, pp. 537–539. Springer, Heidelberg (2004)
3. Berthold, M.R., Cebron, N., Dill, F., Gabriel, T.R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., Wiswedel, B.: KNIME: The Konstanz Information Miner. In: Preisach, C., Burkhardt, H., Schmidt-Thieme, L., Decker, R. (eds.) GfKl. Studies in Classification, Data Analysis, and Knowledge Organization, pp. 319–326. Springer (2007)
4. Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., Euler, T.: Yale: Rapid prototyping for complex data mining tasks. In: Ungar, L., Craven, M., Gunopulos, D., Eliassi-Rad, T. (eds.) KDD 2006: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 935–940. ACM, New York (2006)
5. Podpečan, V., Zemenova, M., Lavrač, N.: Orange4ws environment for service-oriented data mining. The Computer Journal 55(1), 89–98 (2012)
6. Hull, D., Wolstencroft, K., Stevens, R., Goble, C.A., Pocock, M.R., Li, P., Oinn, T.: Taverna: a tool for building and running workflows of services. Nucleic Acids Research 34(web-server-issue), 729–732 (2006)
7. Lavrač, N., Vavpetič, A., Soldatova, L.N., Trajkovski, I., Novak, P.K.: Using ontologies in semantic data mining with segs and g-segs. Discovery Science, 165–178 (2011)