An ontology, artificial inteligence and frequency-based approach to recommend activities in scientific workflows

Adilson Lopes Khouri & Luciano Antonio Digiampietri
University of São Paulo EACH
adilson.khouri.usp@gmail.com
luciano.digiampietri@gmail.com

Introduction

The number of research projects using intensive computing has been growing in areas that lack advanced computer skills such as biology, physics, and astronomy. One of the tools to assist in the management and construction of intensive computing experiments are the workflows manager systems. *Scientific Workflows* represent structured and ordered processes, constructed manually, semi-automatically or automatically to solve scientific problems using activities, which can be: i) source code blocks; (ii) services; or iii) finished workflows [**?**]. These systems facilitate the creation of new experiments, sharing of results and reuse of existing activities.

Nowadays, there are a large number of activities available in repositories such as *myExperiment* which stores more than 2,500 workflows[1] and *BioCatalogue* Which provides more than 2,464 services [**?**]. The large number of activities and the low reuse of some activities and workflows motivate the construction of techniques to recommend activities to the scientists during the composition of workflows [**?**].

In the workflow management systems, activities are typically represented as graphical icons with drag and drop functionality. Thus, it is possible to construct computational experiments by dragging icons and filling in input parameters. Most of these systems provide sets of basic activities that can be used in different domains, for example, an activity that calculates the average value of a set of data is applicable in biology, physics, astronomy, and other areas. However, there is a precondition for reusing andor creating workflows: knowing the available activities.

In order to minimize the problem of knowing a large number of activities, several techniques were proposed to recommend activities or to compose workflows. In the first case, which aims to serve an expert user in these systems, during the construction of the workflow, activities are recommended to help to complete the workflow. In the second case, whose goal is to serve a less expert user on these systems, several workflows are built and the user should select which one most satisfies himher need.

This paper presents a hybrid approach for recommending activities in scientific workflows based on the frequency of activities combined with a knowledge-base ontology for data sets without provenance information, and without reliability data about the authors of the services and workflows. We also propose a modeling of the problem of recommending activities in scientific workflows to be used by classifiers such as: Support Vector Machine (SVM), Naive Bayes (NB), K-Nearest-Neighbor (KNN), Classification and Regression Trees, and Neural Network (MLP). The following regressors were also used: Support Vector Regression (SVR), CART, Neural Network, Multivariate Adaptive Regression Splines (MARS) and Binomial Regression (RB). At last, a comparison of our approach and the approaches from the related literature is presented.

Main Objectives

The 73 bioinformatics' workflows together with their 280 activities were converted into a matrix $M_{i,j}$. In this matrix, each line corresponds to a workflow and each column to an activity. $M_{i,j} = 1$ means that the workflows *i* has the activity *j*. Otherwise, $M_{i,j} = 0$ means that the workflow *i* does not have the activity *j*. Table **??** presents an fictitious example of a matrix M. To perform the evaluation of the approach, an activity is removed from each row of the table **??**, and a list of possible activities is recommended. The goal of the recommendation system is to correctly identify which activity is missing in the workflow (i.e., the one that was removed).

Workflow	Activ01	Activ02	...	Activ280
01	1	0	...	0
02	1	1	...	1
03	1	0	...	1
⋮	⋮	⋮	⋮	⋮
73	1	0	...	0

Figure 1: Figure caption

In order to use classification and regression techniques, some changes were proposed in the original dataset (exemplified in the table **??**), which can be viewed in the table **??**. Each workflow was replicated 118 times. 59 of these correspond to identical copies of the original workflow, while in the other 59 one activity was removed from original workflow and a new activity was added representing a possible recommendation. Thus, for each original workflow, there will be 59 correct instances and 59 incorrect instances and this type of information will be used to train the classifiers or regressors.

[1] http://www.myexperiment.org/

#	Workflow	Activ01	...	Activ280	Class
1	01	1	...	0	T
2	01	1	...	0	T
⋮	⋮	⋮	⋮	⋮	⋮
59	01	1	...	0	T
1	01	0 (removed)	...	0	F
2	01	0 (removed)	...	0	F
⋮	⋮	⋮	⋮	⋮	⋮
59	01	0 (removed)	...	1 (added)	F
	⋮				
1	73	1	...	0	T
2	73	1	...	0	T
⋮	⋮	⋮	⋮	⋮	⋮
59	73	1	...	0	T
1	73	1 (added)	...	0	F
2	73	1	...	0	F
⋮	⋮	⋮	⋮	⋮	⋮
59	73	1	...	1 (added)	F

Figure 2: Figure caption

Results

#	Approach	S@1	S@5	S@10	MRR
1	Random	0.0037	0.0260	0.0280	0.033
2	*Apriori*	0.0037	0.0385	0.0559	0.037
3	KNN_C	0.0037	0.0685	0.0959	0.040
4	Neural $Network_C$	0.0137	0.1507	0.1781	0.089
5	$CART_C$	0.0274	0.1233	0.3699	0.113
6	$CART_R$	0.1370	0.1370	0.2603	0.114
7	Naive $Bayes_C$	0.0274	0.1507	0.3425	0.114
8	$Binomial_R$	0.0822	0.1918	0.2055	0.136
9	Neural $Network_R$	0.1096	0.2603	0.2603	0.154
10	$MARS_R$	0.1233	0.2055	0.2192	0.167
11	FES	0.1474	0.2603	0.3699	0.196
12	SVM_R	0.1233	0.3151	0.4932	0.238
13	SVM_C	0.2425	0.4658	0.4932	0.244
14	composed SVM_C	0.2515	0.4458	0.5232	0.314
15	Rotation $Forest_C$	0.2925	0.4558	0.5432	0.324
16	FESO	0.3425	0.4658	0.5932	0.334

Figure 3: Figure caption

Conclusions

- This work developed a hybrid technique for recommending activities in scientific workflows, which uses syntactic compatibility, frequency, and domain ontologies to recommend activities, called FESO
- The developed technique is better than every other technique from literature
- We have modeled the recommendation problem as a regression and classification problem in artificial intelligence
- The main idea of the project was to add structured semantic information to the recommendation system
- The activities were not independent
- The problem is not linearly separable

Forthcoming Research

Vivamus molestie, risus tempor vehicula mattis, libero arcu volutpat purus, sed blandit sem nibh eget turpis. Maecenas rutrum dui blandit lorem vulputate gravida. Praesent venenatis mi vel lorem tempor at varius diam sagittis. Nam eu leo id turpis interdum luctus a sed augue. Nam tellus.

Acknowledgements

We thank the University of São Paulo (USP) and the CAPES agency wich provided scholarships for the student. Allowing to complete this master with publications in the area of computer science.