An algorithm to recommend activities in scientific workflows:

An ontology, artificial inteligence and frequency-based approach

PhD Candidate Adilson Lopes Khouri &
Associate Professor Luciano Antonio Digiampietri

adilson.khouri.usp@gmail.com
digiampietri@usp.br

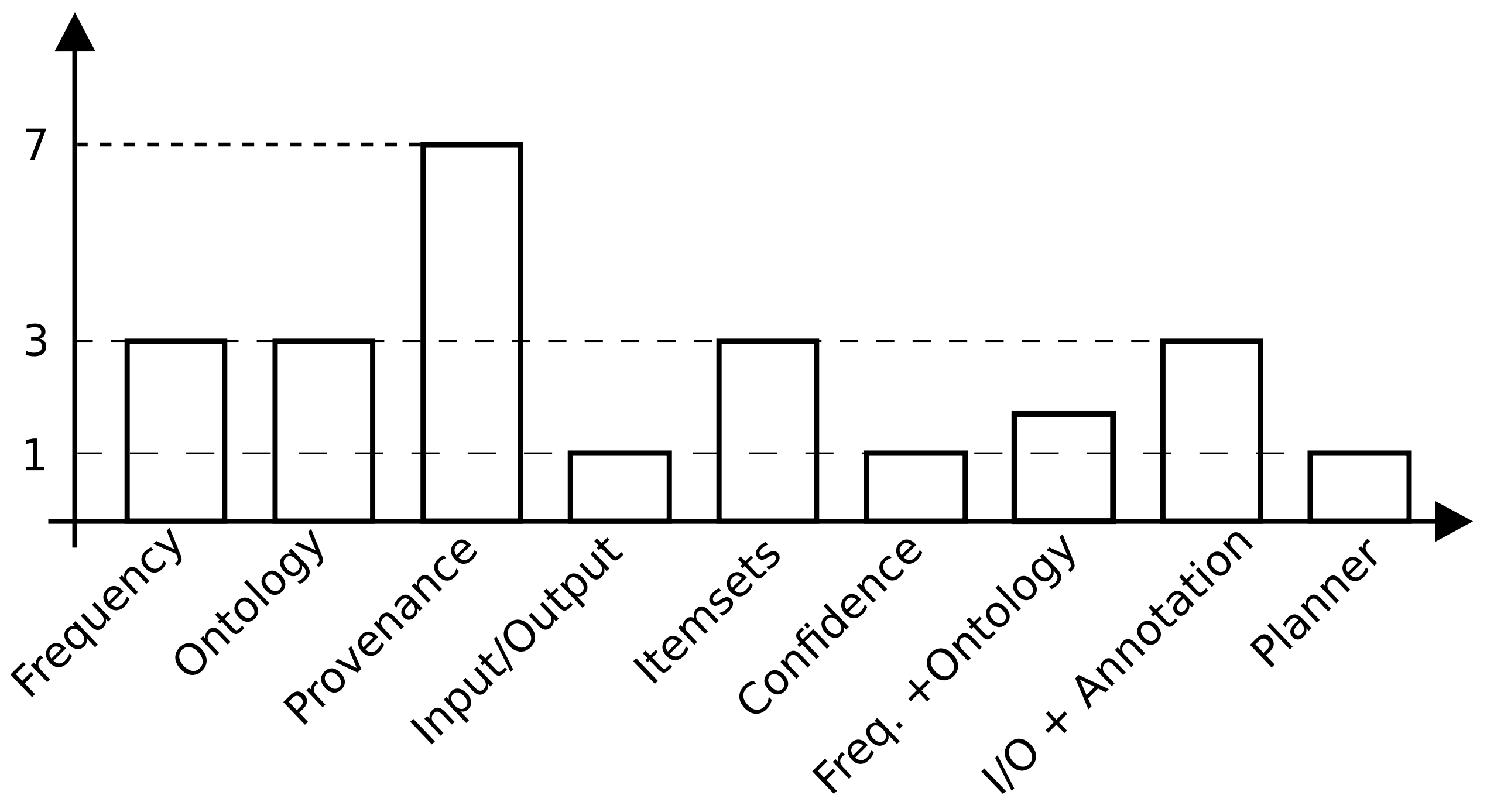
Introduction

The number of research projects using intensive computing has been growing in areas in which scientists 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 is the workflow management system. *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 2500 workflows and *BioCatalogue* which provides more than 2464 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 dataset 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 needs.



Main Contributions

Recommendation as a classification problem

We have downloaded 73 bioinformatic's workflow's with their 280 activities (from *myExperiment*) were converted into a matrix $M_{i,j}$. In this matrix (we call it *original dataset*), 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*. To evaluate the approach, an activity is removed from each row of the original dataset, 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).

In order to use classification and regression techniques, some changes were proposed in the original dataset, which can be viewed in the next 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 the 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 and the regressors.

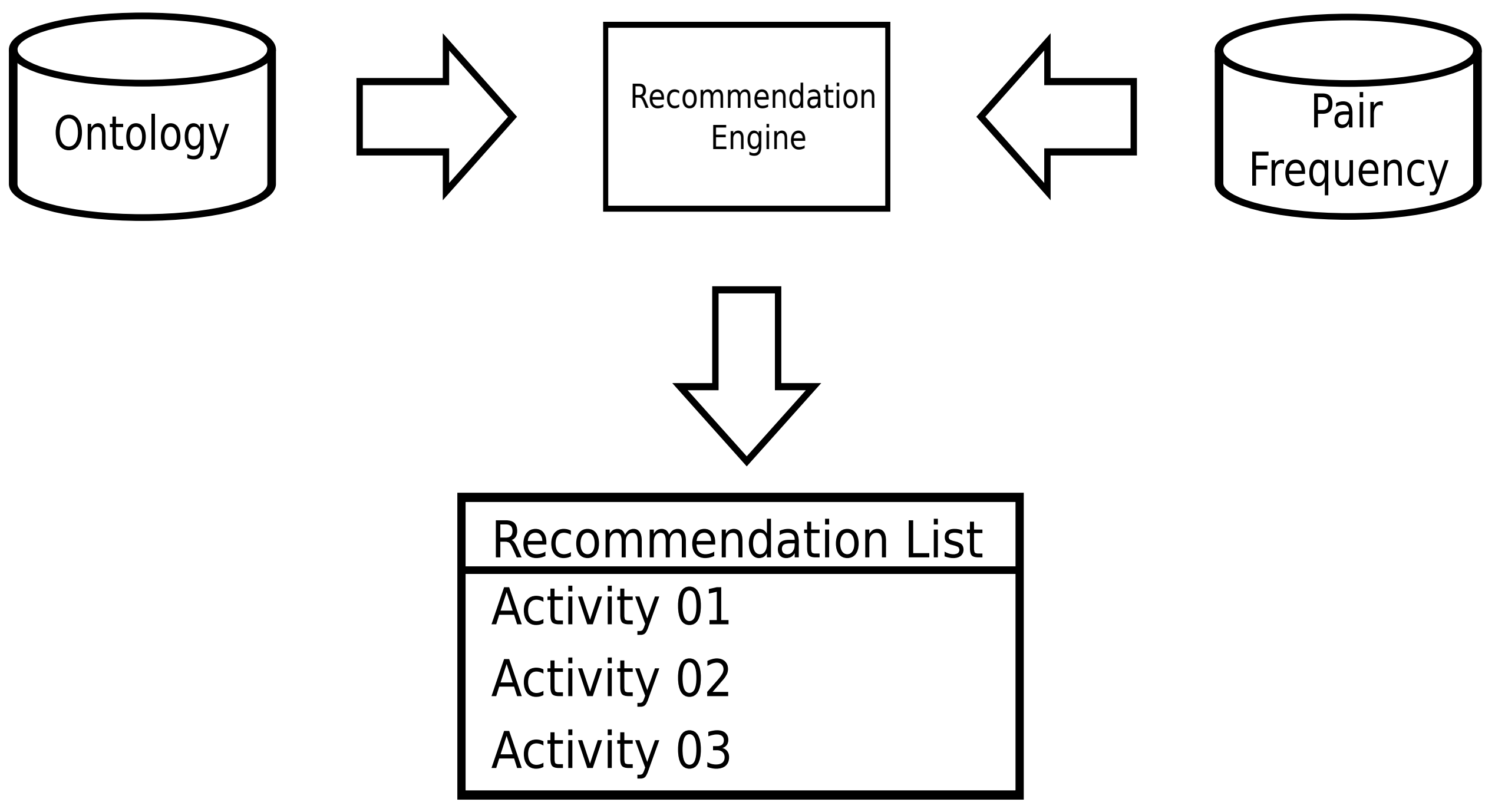
#	Workflow	Activ01	. . .	Activ280	Class
1	01	1	. . .	0	T
2	01	1	. . .	0	T
⋮	⋮	⋮	⋮	⋮	⋮
59	01	1	. . .	0	T
1	01	0 (removed)	. . .	0	F
2	01	0 (removed)	. . .	0	F
⋮	⋮	⋮	⋮	⋮	⋮
59	01	0 (removed)	. . .	1 (added)	F
1	73	1	. . .	0	T
2	73	1	. . .	0	T
⋮	⋮	⋮	⋮	⋮	⋮
59	73	1	. . .	0	T
1	73	1 (added)	. . .	0	F
2	73	1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

Recommendation Algorithm

The proposed solution recommends activities using three concepts in the area of scientific workflows: i) frequency of activities; ii) compatibility between input and output; and iii) semantics of activities. The solution begins calculating the frequency of occurrence of each pair of existing activities, which is the number of times that an activity *W* occurs immediately after another activity *Z*. By considering only activities that have already been connected (on the dataset of workflows), the output and input compatibility is guaranteed.

After calculating the frequency it is necessary to annotate all the workflows, using the concepts of the domain ontology, this step was performed manually (not automatically). Finally, the algorithm annotates all activities with the same annotations of their respective workflow; i.e., if the *X* activity is inside two workflows with distinct annotations, then this activity will be related to two different concepts from the ontology.

To understand the recommendation training mechanism, another example will be used to simulate a user interacting with the recommendation system. Let us assume that during the construction of the workflow a scientist inserts the *Z* activity and asks for a recommendation. The system will look at the list of activities after *Z* sorted by frequency and ontological concept and will return the recommendation list. The sorting considers the ontological concepts serves as a tiebreaker criterion when two activities have the same frequency.


Ontology
Recommendation Engine
Pair Frequency
Recommendation List
Activity 01
Activity 02
Activity 03

Results

Our ontology-based approach (FESO) achieved better (or at least equal) results than the previous approaches for almost all of the evaluated metrics. It considers the use of frequency, input, output and semantic information about the activities, in comparison to the other techniques, its results were higher for all calculated metrics. Concerning the FES technique, its results were superior. In particular, part of this improvement is justified by cases where the correct activity has zero frequency in the training set. Since FESO considers the ontology information it is able to recommend activities even if they have zero frequency in the train set. In addition, for the cases where there is a tie between two activities considering the input, output, and the frequency criteria, the proposed technique presents an additional factor to be used as a tie breaker.

We were able to identify some trends in these results. Increasing information on data in the recommendation improves their recommendation performance, as the results of the experiments: 11 and 16 show. A second trend is that the SVM classifier was the only one that obtained a better result than the regressors. indicating that solutions by maximizing space between data in high dimension may be a promising area of study. A third trend is the use of composite classifiers and ensembles which presented promising 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

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 *Frequency Input Output* (FESO, the number 16 in the table above). Moreover, we have modeled the recommendation problem as a regression and classification problem in artificial intelligence. As future work we intend to investigate the use of data provenance to increase the accuracy of the recommendations.

Acknowledgements

We thank the University of São Paulo and the CAPES, which provided scholarships for the student.