Combining artificial intelligence, ontology, and frequency-based approaches to recommend activities in scientific workflows

Adilson Lopes Khouri [1]
Luciano Antonio Digiampietri [1]

Abstract: The number of activities provided by scientific workflow management systems is large, which requires scientists to know many of them to take advantage of the reusability of these systems. To minimize this problem, the literature presents some techniques to recommend activities during the scientific workflow construction. In this paper we specified and developed a hybrid activity recommendation system considering information on frequency, input and outputs of activities and ontological annotations. Additionally, this paper presents a modeling of activities recommendation as a classification problem, tested using 5 classifiers; 5 regressors; and a composite approach which uses a Support Vector Machine (SVM) classifier, combining the results of other classifiers and regressors to recommend; and Rotation Forest, an ensemble of classifiers. The proposed technique was compared to related techniques and to classifiers and regressors, using 10-fold-cross-validation, achieving a Mean Reciprocal Rank (MRR) at least 70% greater than those obtained by classical techniques.

1 Introduction

The number of research projects using intensive computing has been growing in areas such as biology, physics, and astronomy. One of the tools to assist in the management and construction of intensive computing experiments are the workflows management 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 ([21]). These systems facilitate the creation of new experiments, sharing of results and reuse of existing activities. Workflows are models to represent a flows of interrelated activities which execution leads to a goal. In this paper the term *workflow* is used as a synonymous of *scientific workflow*.

There is also another type of workflows, the business workflows, whose area of study is known as Business Process Management (BPM). Considering business workflows, the process mining aims to discover, monitor or improve process based on event logs [20].

Typically, scientific workflows have an intensive use of computational resources and

[1]Escola de Artes Ciencias e Humanidades, EACH
{adilson.khouri.usp@gmail.com, luciano.digiampietri@gmail.com}

are *data flow* oriented. On the other hand, business workflows are, typically, *control flow* oriented. In this paper, we focus only on scientific workflows..

Nowadays, there are a large number of activities available in repositories such as *myExperiment*[2] which stores more than $2,500$ workflows and *BioCatalogue*[3], 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 ([21]).

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 and/or creating workflows (without the aid of a recommender system): knowing a great number of available activities to avoid recreate them.

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, several workflows are built, automatically by a computer program combining the input and output of activities, and the user should select which one most satisfies him/her need. In the literature this second process is called: *workflow composition* [6].

Although there are already some approaches for recommending activities in workflows, in general, all have some limitations. For example, many of the approaches require a very large data set to enable a frequency based or a machine learning approach. Thus, they do not deal well with sparse data (i.e., the existence of a large set of activities each one of them used only in few workflows). Other solutions need detailed information about authors, workflows, and activities that typically are not available in public repositories of activities and workflows.

This paper presents a hybrid approach for recommending activities in scientific workflows based on the frequency of activities combined with the use of semantics, considering datasets with no provenance information, and without reliability information 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 Regres-

[2]http://www.myexperiment.org/
[3]https://www.biocatalogue.org/

sion Trees (CART), 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 (BR).

In the section *Relate Work* describes the techniques proposed in the related literature. In the section *Materials and Methods* presents the data source, selected sample, and the modeling of the problem as a classification problem. In the section *Proposed Solutions* describes the proposed solution: our algorithm that combines different characteristics in the recommendation and the recommendation of activities modeled as an artificial intelligence classification problem. In section *Results* we present and discuss a performance comparison of our approach and the approaches from the related literature is presented. In the section *Conclusion* the final considerations are presented.

2 Related Work

The related literature presents several techniques to recommend activities in scientific workflows. They will be briefly described in this section, for a complete systematic review we suggest the work of Khouri and Digiampietri, 2015. The works of [14] and [15], which consider the sequential mining of activities as *itemsets*, ignore the order of activities and their semantics. The proposal of [13] disregards only the semantics of activities. The present approach considers the order of activities as an important factor in the recommendation because in data flows applications (such as in scientific workflows) the results depends on the order of the activities. Previous activity could be used to predict the next activity when there is a high frequency of both in the same order. Previous activity could be used to predict the next activity when there is a high frequency of both in the same order.

The work of [11, 4, 23, 26, 17, 2, 5] and [7, 25] consider the order of activities, input, output and data provenance. Their limitations are the need of provenance data, since not all Scientific Workflow Management System (SWMS) stores this information. Besides, they do not use semantic information of workflows and activities. Our approach does not require provenance information and considers the semantics of the information using an ontology.

The work of [1] uses only a mapping between activities and ontology, disregarding the input and output, which potentially generates inefficient recommendations. In our approach, the inputs and outputs of each activity are considered, in addition to the use of a domain ontology. The match between input and output is important to ensure that the output data type of an activity are syntactically compatible with the input of another.

Wang et al, 2008 and Leng et al, 2010 use only the posteriori probability of occurrence of a new activity. For example, if there was a workflow where a service *b* call service *c* and *c* call service *d*, during the construction of a new workflow, if the user added the service *b*

the system will suggest *c* and *d*. The authors do not consider the use of semantics or even the order of pairs of services (or activities).

The work of [24] requires calculating the confidence of users and of their workflows. Repositories like *myExperiment* do not require users to fill in this data, thus much of the information related to this aspect is not filled by users. In addition, the authors disregard the semantics of activities and workflows.

The works of [18, 3] and [27] disregard the use of semantics to recommend, which is a limitation as discussed by [8, 16]. In our approach, the frequency is considered in conjunction with the domain ontology.

The works of [8] and [16] consider the use of frequency and ontology, as in this approach, but they recommend *subworkflows* which limits the recommendations of activities. Only activities used in common fragments of workflows may be recommended. In other words, if the activity is in the "middle" of a *subworkflow* it will never be recommended individually. In our approach, all activities can be recommended even at the end of the recommendation list. In addition, it presents a more comprehensive recommendation, as it deals with activities, *subworkflows* (more than one activity working together as one workflow inside another workflow) and *shims* (data type converters and/or adapters).

3 Materials and Methods

The workflows were obtained from the *myExperiment* repository, using the program *wget*[4]. After downloading the $2,481$ workflows in *xml* format, the *BeautifulSoup*[5] code analyzer was used to organize the dataset in a relational database

The data is exported to a simple matrix used for techniques that do not use the order of the activities. And also in an array adapted to modelling the recommendation problem as artificial intelligence binary classification and regression problems. These representation will be described in the following sections.

3.1 Simple matrix

The 2481 contains 73 bioinformatic's workflows related to genome assembly and annotation. These workflows were used in the study case for this paper and they are composed of 280 activities. The 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 workflow *i* contains the activity *j*. Otherwise, $M_{i,j} = 0$ means that the workflow *i* does not contain the

[4]https://www.gnu.org/software/wget/
[5]https://www.crummy.com/software/BeautifulSoup/bs4/doc/

activity *j*. Table 3.1 presents an fictitious example of a matrix M. To perform the evaluation of the approach, an activity is removed from each row of Table 3.1, 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).

Table 1. Input matrix example.

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

3.2 Adapted Array

In order to use classification and regression techniques, some changes were proposed in the original dataset (exemplified in Table 3.1), which can be viewed in the table 2. 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 or regressors. The choice of 59 activities to be recommended was made for two reasons. The first is to select the 59 activities most frequently used in the database. The second is the computational limitation: replicating the 280 possible recommendations might be impractical in terms of training. Thus, the number 59 was chosen empirically after some exploratory tests. We have replicated 59 instances of identical workflows considered correct, i.e. with the correct activity not removed, to ensure inter-class balancing. The last change was to add a column indicating whether the recommendation of the proposed activity is correct, that is, the one belonging to the respective workflow (*T*) or not (*F*).

3.3 Results evaluation

The 10-fold-cross-validation technique was used to evaluate the proposed approach. In this technique, the dataset is divided into 10 subsets (*folds*) and ten executions are performed. In each, 10% of the workflows are separated for testing and 90% for training. Thus, for each run, the system trains with 90% of the data and the training result is tested for the

Table 2. Input matrix used by classifiers and regressors

#	WF	Act01	Act02	...	Act280	Class
1	01	1	0	...	0	T
2	01	1	0	...	0	T
⋮	⋮	⋮	⋮	⋮	⋮	⋮
59	01	1	0	...	0	T
1	01	0 (removed)	1 (added)	...	0	F
2	01	0 (removed)	0	...	0	F
⋮	⋮	⋮	⋮	⋮	⋮	⋮
59	01	0 (removed)	0	...	1 (added)	F
	⋮					
1	73	1	1	...	0	T
2	73	1	1	...	0	T
⋮	⋮	⋮	⋮	⋮	⋮	⋮
59	73	1	1	...	0	T
1	73	1 (added)	0 (removed)	...	0	F
2	73	1	0 (removed)	...	0	F
⋮	⋮	⋮	⋮	⋮	⋮	⋮
59	73	1	0 (removed)	...	1 (added)	F

remaining 10%.

It is worth noticing that 100% of the data set is labeled (that is, it makes explicit to the system which activity was removed) and, thus, it is possible to verify the performance of each of the runs. The test presents the 10% workflows, without informing the labels (the activity removed), for the recommendation systems that have already been trained. At the end of the ten executions, the averages of the metrics are calculated: i) *Success at rank k* ($S@k$); and ii) Mean Reciprocal Rank (MRR) ([9]).

The metric $S@k$ calculates the probability of an item of interest being recommended in the k first positions in the list of recommended activities. Its value lies in between zero and one. The results of this metric are cumulative for increasing values of k, this occurs because if an activity of interest is in the top five of the list of recommendations, it is also in the top ten positions. At the limit, the activity will always be in the L first positions, where L is the total size of the recommendation list. Thus, high values for $S@k$ are considered good, especially

for low values of k. These metric are calculated using the following equation:

$$S@k = \frac{1}{N}\sum_{i=1}^{N}(I(n_i \leq k)) \quad (1)$$

N is the number of recommending lists; n_i is the position of the required item in the list i,k is the input parameter which determines the last position that will be considered in the equation 1; and function *I* indicates if the activity n_i occurs in a position smaller or equal to k.

The metric Mean Reciprocal Rank (MRR) is the inverse position an item of interest being recommended, one is the best value of this metric and zero is the worst. These metric are calculated using the following equation:

$$MRR = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{n_i}\right) \quad (2)$$

4 Proposed Solutions

In this paper we proposed two types of solution, in the first, the recommendation of activities is modeled as an artificial intelligence classification problem. Where you train a statistical model to learn patterns using the workflows' data and validate it with a cross validation strategy to evaluate the model.

In the second type of solution, the proposed solution recommends activities using three important concepts in the area of scientific workflows: i) frequency of activities; ii) compatibility between input and output; and iii) semantics of activities. We called it FIOO (Frequency Input Output and Ontology). To explain this proposal, Figure 1 is be used as an example. It is possible to observe six workflows with their annotations, which simulate a database of scientific workflows.

The FIOO solution begins by 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 of the figure 1, using the concepts of the domain ontology (see Figure 2). This step was performed manually which is a limitation of the work. Finally, the algorithm annotates all activities with the same annotations of their respective workflow; i.e., if the *X* activity (Figure 1) is inside two workflows with distinct annotations, then this activity will be related to two different

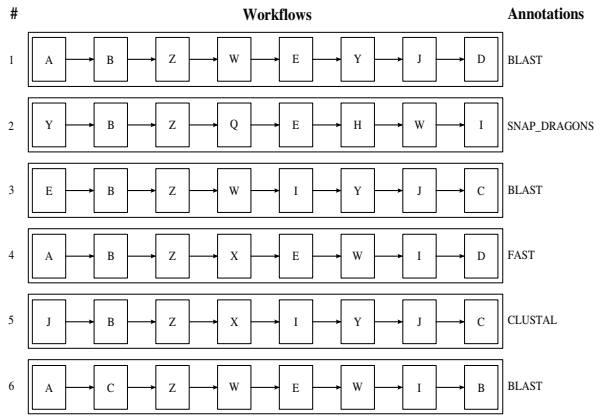

Figure 1. An example of a database of scientific workflows. Scientific workflows with ontological annotations used to exemplify the proposed solution.

concepts from the ontology. The final result is presented in Table 3, which contains the activities' frequencies and annotations.

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 1 (see Figure 2) a scientist inserts the *Z* activity and asks for a recommendation. The system will look at the list of activities that occurs after *Z* sorted by frequency and ontological concept and will return the recommendation list presented in table 3. The sorting considering the ontological concepts serves as a tiebreaker criterion when two activities have the same frequency. In this example, according to the recommendation list of Table 3, the *W* activity would be first recommended to the user.

Table 3. Recommendation for activity *Z* sorted by frequency and ontological concept

Position	Activ	Frequency	Annotation
1	W	3	BLAST
2	X	2	FAST, CLUSTAL
3	Q	1	SNAP DRAGONS
⋮	⋮	⋮	⋮
280	⋮	⋮	⋮

The activities are annotated with the same annotation of the workflows that contain

them. Thus, it is possible that there is at least one activity with more than one annotation. This creates a new recommendation case to consider. Suppose both *W* and *X* activities contains in their annotation lists the concept *BLAST*. In this case, the activity with a lower number of annotations would be recommended, since it is considered more specific for the experiment in question. If both activities have the same number of annotations, the alphabetical order of concepts is used as the tie-breaking criterion. If a new tie occurs a random selector is used.

4.1 Ontology construction

The ontology was developed using the Skeletal methodology [19] which define the steps: (a) Identify the objective; (b) Capture of ontology; (c) Code; (d) Merge with other ontologies; and (e) Validation.

All these steps were followed in the construction of the ontology. The objective of this ontology was to standardize annotations for workflows and to create an hierarchy to be used in the recommendation step. During the Capture of ontology step, authors studied the bioinformatics area to understand the ontology domain. For the Code step, the Protégé tool[6] was used. No other ontology has found, thus there was no merge step and a domain specialist validated the ontology produced.

5 Results

This sections show all the sixteen experiments that were done, each one was a typical machine learning experiment trained by a 10-fold cross validation for each parameter change. Each approach has a typical set of parameters to be optimized during the training step. Table 4 shows all the variations of parameters.

Table 4 displays the results of each recommendation system used. FIOO corresponds to our Frequency Input Output and Ontology approach. The techniques that have the letter *C* in subscript are classifiers; the ones that have letter *R* in subscript are regressors; and those that have nothing are from related literature. Each system makes its recommendations according to its own criteria in an recommendation list. Then the activities not recommended are added to the end of this list. Thus, the correct activity will always be found, and the factor that differentiates the recommendation systems is the position the activities are in the recommendation list, which contains 280 positions.

The *Random* approach did not require training. The algorithm only randomly selected activities, forming a list of recommended activities. This system recommended less than 3% of the correct activities in the top ten positions. Most of the correct activities were rated close

[6]https://protege.stanford.edu/

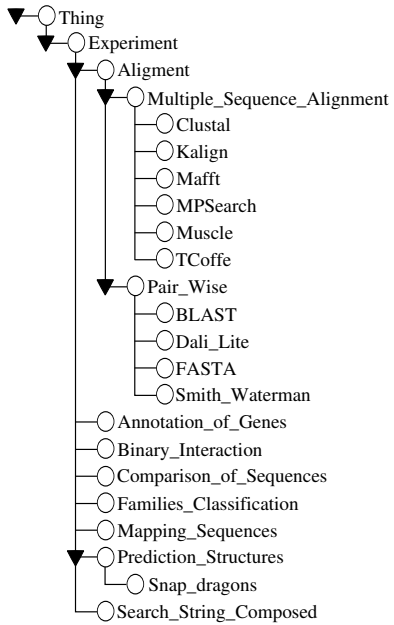

Figure 2. Ontology. Ontology built to annotate scientific workflows with ontological concepts.

to position 140 which is the average position of the recommended lists. The metric values $S@280 = 1$ and $S@100 = 0.04$ indicate that most of the correct items were found after the hundredth position. This system was used to calculate the simplest baseline.

The system using the *Apriori* technique achieved its best performance when the *confidence* and *support* parameters were defined as *without limitation*, that means, no minimum confidence and support values were defined for the creation of association rules. All rules were considered valid. Even without restricting these values, the results of this technique presented results better than only the *Random* approach. Recommending less than 6% of correct activities between the 50 first positions, its accuracy is still low with the value of $MRR = 0.037$. The poor results of this technique happened due to the fact the technique disregards the order of the activities during the generation of the rules and, consequently, of the recommendation.

The *KNN* based approach has been trained considering different values of the parameter *k* (from 1 to 100) which represents the number of nearest neighbours (and using the Euclidean distance as the proximity metric). The best recommendation results for this ap-

Table 4. Recommendation results

#	Approach	S@1	S@5	S@50	S@100	S@280	MRR
1	Random	0.00	0.02	0.03	0.04	1	0.033
2	*Apriori*	0.00	0.03	0.05	0.05	1	0.037
3	KNN_C	0.00	0.06	0.50	1.00	1	0.040
4	N. Network_C	0.01	0.15	0.80	1.00	1	0.089
5	CART_C	0.02	0.12	0.76	1.00	1	0.113
6	CART_R	0.13	0.13	0.61	1.00	1	0.114
7	Naive Bayes_C	0.02	0.15	0.63	1.00	1	0.114
8	Binomial_R	0.08	0.19	0.84	1.00	1	0.136
9	N. Network_R	0.10	0.26	0.26	1.00	1	0.154
10	MARS_R	0.12	0.20	0.72	1.00	1	0.167
11	FIO	0.14	0.26	0.86	1.00	1	0.196
12	SVM_R	0.12	0.31	0.84	1.00	1	0.238
13	SVM_C	0.24	0.46	0.71	1.00	1	0.244
14	Comp. SVM_C	0.25	0.44	0.76	1.00	1	0.314
15	Rot. Forest_C	0.29	0.45	0.77	1.00	1	0.324
16	FIOO	0.34	0.46	0.81	1.00	1	0.334

proach were obtained with $k = 2$. Even so, less than 10% of the correct items were found among the top ten items in the list and 50% of items in the first 50 items. According to the MRR metric, the average position of the recommended items was far from the first position in the list ($MRR = 0.04$). These results indicate that classifying activities according to the distance between groups of neighbours is not an appropriate approach to this problem.

The approach which uses an MLP neural network as a classifier had results significantly better than the ones achieved by the *KNN* approach when considering the metric S@1 (0.0137 versus 0.0037). For the training of the network the parameters used were: i) number of neurons η (ranging from $1 : 40$); ii) learning rate α (ranging from $10^{-7} : 10$); iii) two hidden layers; and iv) fully connected architecture. The best classification results were achieved with $\eta = 18$ and $\alpha = 10^{-4}$, obtaining 17% of items ranked among the top ten positions in the list, and 80% among the 50 first positions, which represents an improvement of 30% when compared with the *KNN* approach. The metric value $MRR = 0.089$ was twice as high as the one from *KNN*, this growth in precision indicates the neural network generalization power to solve non-linear problems was more efficient than the previous approaches.

The approach which uses CART as a classifier, dealing with categorical data, presented a result superior to the one from neural network. The training used the parameters: i) minimum division value $\gamma = [0 : 30]$; ii) maximum final tree size $\delta = [0 : 10000]$; iii)

minimum variation value to perform a division $cp = [10^{-7} : 10]$; iv) division function (ξ) as Gini Index or Information Gain. The best result was achieved with $gamma = 0$, $\delta = 30$, $cp = 10^{-3}$, and $\xi =$ Information Gain.

The results of this approach were approximately twice as good as those of the neural network. This indicates approaches that deal with categorical data by nature has a potential for obtaining good results in problems such the one addressed in this paper. The MRR results were 26% better than the ones achieved by the MLP based approach, and this approach was able to recommend 12.3% of the searched items in the first position and 76% in the first 50 positions.

The approach which uses CART as regressor achieve its best results with $gamma = 2$, $\delta = 20$, $cp = 10^{-5}$, and $\xi =$Information Gain. The recommendation that used continuous values presented a result superior to $CART_C$ considering the metrics $S@1$ and $S@5$ and worse results for $S@10$ and $S@50$. The general precision (MRR) of $CART_R$ was slightly higher than the one achieved by $CART_C$.

The Naive Bayes classifier based approach obtained results very similar to the ones achieved by the CART regressor. The training occurred by ranging the *Laplace correction* attribute with values between $[0 : 100]$. The best result occurred with value zero for this parameter, achieving 34% of the recommended items in the top ten positions and 63% among the first 50 positions. In contrast, the value of MRR did not change much.

The binomial regressor approach presented better results when compared with the Naive Bayes and the neural network approaches. The training of this technique occurs using the maximum likelihood for a generalized linear model approximated by a binomial distribution. The results for $S@5$ and $S@50$ were higher than the achieved by the previous approaches and the value of the metric MRR improved by approximately 19% when compared to the one achieved by the Naive Bayes approach.

The approach which uses a neural network as regressor, considering the weight of the neural network as output, was trained in an analogous way to the neural network used as a classifier. The best result was obtained for the values of $\eta = 10$ and $\alpha = 10^{-2}$. It was able to recommend 26% of the correct items among the top ten positions in the list. System accuracy (MRR) improved 13% from the achieved by the binomial regressor. These results indicate that using a regressor instead of a classifier presents a better result for this kind of problem, at least when using neural networks.

The approach based on the MARS algorithm as regressor achieved a better result than the one from the neural network (used as regressor). The metric $S@1$ was improved in 12.5% and the overall precision (MRR) increased 8%. This result shows that the curves created by the various connected functions of the MARS obtained a better generalization than the neural network. The training of the parameters was performed using likelihood.

Among the systems proposed in the literature, the system based on input, output and frequency (FIO) ([22]) is the one with the best results. In the experiments performed, this system identified the correct item among the top ten positions of the recommendation list in 37% of the cases, and obtained a value of $MRR = 0.196$.

The SVM regressor presented results twice as good as the MARS algorithm for the metric $S@10$, since in 49% of the recommendations the correct item was among the top ten positions in the recommendations list. The MRR value was also higher (42%). The training was performed using margin optimization with the values of $c = [10^{-7} : 10^{2}]$, $\epsilon = [10^{-7} : 10^{2}]$, Tolerance values $\beta = [10^{-7} : 10^{2}]$, kernel functions: i) linear; ii) sigmoid; iii) polynomial; and (iv) radial. The tested values of the parameter of the polynomial kernel were $p = [1 : 10]$ which is the power of the function. The best results were achieved for $c = 1$, $\epsilon = 1$, $\beta = 10^{-4}$, and polynomial kernel with $p = 2$.

The approach based on the SVM algorithm for classification was the only classifier that surpassed the results of the regressors. His training was analogous to the SVM for regression. Its best results were achieved with $c = 10^{-1}$, $p = 10^{-4}$, and *linear kernel*. The value of the metric $S@1$ was 64% better than the one from FIO technique and the general precision value (MRR) increased 24%. This result indicates that the solution using *kernel* for high-dimensional mapping is an efficient approach in the case of classifiers.

The composed SVM system, which recommends items based on the results of the other recommendation approaches, achieved better results than the SVM. Its training was analogous to that of SVM_C and its best results were achieved with $c = 10^{-2}$, $p = 1$, and polynomial kernel. There was an improvement of 3% in the metric $S@1$ and 28% in the metric MRR, this improvement is due to the use of the result of other classifiers together with the sparsity reduction of the data set.

The system using *Rotation Forest* presented the second best result, its training used the parameters: i) minimum division value $\gamma = [0 : 30]$; ii) maximum final tree size $\delta = [0 : 10000]$; iii) minimum variation value to perform a division $cp = [10^{-7} : 10]$; iv) division function (ξ) using Gini Index and Information Gain; v) $K = [1 : 10]$ as the number of partitions; vi) $L = [1 : 10]$ as the number of classifiers; and (vii) cutoff values $0.25, 0.5, 0.75$. Is use of an ensemble classification technique was able to achieve better results, for example, $S@1 = 0.29$, $S@10 = 0.54$, and $MRR = 0.324$.

Our ontology-based approach (FIOO) 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 result was higher for all calculated metrics, except $S@50$ for some techniques. In relation to the FIO technique, its result was superior. In particular, part of this improvement is justified by cases where the correct activity has zero frequency in the train-

ing set. Since FIOO considers the ontology information it is able to recommend activities even if they have zero frequency in the train set. In addition, for the case 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 the recommenders performance, as the result of the experiments: 2, 12, and 14 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. In the case of the *ensemble*, there is a clue that techniques of this kind, which use thresholds to convert the mean values of the set result L into binary values, have promising results in recommending activities.

6 Conclusions

This work presented a hybrid approach to recommend activities in scientific workflows, called FIOO. It uses syntax compatibility, frequency, and domain ontology to recommend activities. We also modeled the problem of recommendation as an artificial intelligence classification and regression problem.

Our results were compared with the ones presented in the related literature which was previously identified through a systematic literature review. In this review, we identified techniques, their restrictions, their advantages and the forms that they were validated.

In order to perform the comparison, a relational database of workflows and their activities was constructed. It was also necessary to establish a methodology to compare different activity recommendation techniques for the same data set with the same validation metrics ($S@k$ and MRR).

When comparing all techniques, certain aspects of the data set were verified, such as the fact that the activities were not independent; the problem is not linearly separable, and that clustering techniques were not adequate to solve this problem. With the exception of SVM, regressors presented more accurate results than classifiers. Finally, adding information in the recommendation systems improved their accuracy.

The extension of this system to another domain is easy. The first step is to develop or find a domain ontology. After that, all the activities should be annotated according to this ontology. Finally, use the recommendation technique described in this paper. If someone does not want to annotate the activities, he/she can use the classification based approaches. In order to do this, it is necessary to only model the workflows as an input for the classifiers,

as presented in this paper.

As future work, we intend to investigate the use of data provenance to increase the accuracy of the recommendations. Moreover, we will investigate how to automatically annotated the workflows, the main limitation of this paper.

References

[1] E. Bomfim, J. Oliveira, J. de Souza, and J. Strauch. Thoth: improving experiences reuses in the scientific environment through workflow management system. In *Computer Supported Cooperative Work in Design, 2005. Proceedings of the Ninth International Conference*, volume 2, pages 1164–1170 Vol. 2, 2005.

[2] B. Cao, J. Yin, S. Deng, D. Wang, and Z. Wu. Graph-based workflow recommendation: on improving business process modeling. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, CIKM '12, pages 1527–1531. ACM, 2012.

[3] F. T. de Oliveira. Um sistema de recomendação para composição de workflows. Master's thesis, Universidade Federal do Rio de Janeiro, 2010.

[4] F. T. de Oliveira, L. Murta, C. Werner, and M. Mattoso. Using provenance to improve workflow design. In J. Freire, D. Koop, and L. Moreau, editors, *Provenance and Annotation of Data and Processes*, volume 5272 of *Lecture Notes in Computer Science*, pages 136–143. Springer Berlin Heidelberg, 2008.

[5] C. Diamantini, D. Potena, and E. Storti. Mining Usage Patterns from a Repository of Scientific Workflows. In *Proceedings of the 27th Annual {ACM} Symposium on Applied Computing*, SAC '12, pages 152–157. ACM, 2012.

[6] X. Fei and S. Lu. A dataflow-based scientific workflow composition framework. *IEEE Transactions on Services Computing*, 5(1):45–58, Jan 2012.

[7] D. Garijo, O. Corcho, and Y. Gil. Detecting Common Scientific Workflow Fragments Using Templates and Execution Provenance. In *Proceedings of the Seventh International Conference on Knowledge Capture*, K-CAP '13, pages 33–40, New York, NY, USA, 2013. ACM.

[8] D. Garijo, O. Corcho, Y. Gil, M. N. Braskie, D. Hibar, X. Hua, J. Neda, P. Thompson, and A. W. Toga. Workflow Reuse in Practice: A Study of Neuroimaging Pipeline Users. *Proceedings of the 2014 IEEE 10th International Conference on eScience*, pages 239–246, 2014.

[9] M. Harvey, I. Ruthven, and M. Carman. Ranking Social Bookmarks Using Topic Models. *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 1401–1404, 2010.

[10] A. L. Khouri and L. A. Digiampietri. A systematic review about activities recommendation in workflows. In *12ª Conferência Internacional sobre Sistemas de Informação e Gestão de Tecnologia (CONTECSI)*, page 14, 2015.

[11] D. Koop. Viscomplete: Automating suggestions for visualization pipelines. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1691–1698, nov 2008.

[12] Y. Leng, M. El-Gayyar, and A. B. Cremers. Semantics Enhanced Composition Planner for Distributed Resources. In *2010 Ninth International Symposium on Distributed Computing and Applications to Business, Engineering and Science*, pages 61–65. IEEE, aug 2010.

[13] F. T. d. Oliveira, V. Braganholo, L. Murta, and M. Mattoso. Improving workflow design by mining reusable tasks. *Journal of the Brazilian Computer Society*, 21(1):16, 2015.

[14] Q. Shao, M. Kinsy, and Y. Chen. Storing and Discovering Critical Workflows from Log in Scientific Exploration. In *2007 IEEE Congress on Services (Services 2007)*, pages 209–212. IEEE, jul 2007.

[15] Q. Shao, P. Sun, and Y. Chen. Efficiently discovering critical workflows in scientific explorations. *Future Generation Computer Systems*, 25(5):577–585, may 2009.

[16] K. Soomro, K. Munir, and R. Mcclatchey. Incorporating semantics in pattern-based scientific workflow recommender systems. 2015.

[17] W. Tan, J. Zhang, R. Madduri, I. Foster, D. De Roure, and C. Goble. Providing Map and GPS Assistance to Service Composition in Bioinformatics. In *2011 IEEE International Conference on Services Computing*, pages 632–639. IEEE, jul 2011.

[18] A. Telea and J. J. van Wijk. Vission: An object oriented dataflow system for simulation and visualization. In *PROCEEDINGS OF IEEE VISSYM*, pages 95–104, 1999.

[19] M. Uschold and M. King. Towards a methodology for building ontologies. In *In Workshop on Basic Ontological Issues in Knowledge Sharing, held in conjunction with IJCAI-95*, 1995.

[20] W. M. P. van der Aalst. *Process Mining - Discovery, Conformance and Enhancement of Business Processes*. Springer, 2011.

[21] F. Wang, H. Deng, L. Guo, and K. Ji. A Survey on Scientific Workflow Techniques for Escience in Astronomy. In *2010 International Forum on Information Technology and Applications*, volume 1, pages 417–420. IEEE, jul 2010.

[22] J. Wang, Y. Han, S. Yan, W. Chen, and G. Ji. Vinca4science: A personal workflow system for e-science. In *Internet Computing in Science and Engineering, 2008. ICICSE '08. International Conference on*, pages 444–451, 2008.

[23] Y. Wang, J. Cao, and M. Li. Change Sequence Mining in Context-Aware Scientific Workflow. In *2009 IEEE International Symposium on Parallel and Distributed Processing with Applications*, pages 635–640. IEEE, 2009.

[24] J. Yao, W. Tan, S. Nepal, S. Chen, J. Zhang, D. De Roure, and C. Goble. Reputationnet: A reputation engine to enhance servicemap by recommending trusted services. In *Services Computing (SCC), 2012 IEEE Ninth International Conference on*, pages 454–461, 2012.

[25] P. Yeo and S. S. R. Abidi. Dataflow Oriented Similarity Matching for Scientific Workflows. In *2013 IEEE International Symposium on Parallel & Distributed Processing, Workshops and Phd Forum*, pages 2091–2100. IEEE, may 2013.

[26] J. Zhang, Q. Liu, and K. Xu. Flowrecommender: A workflow recommendation technique for process provenance, 2009.

[27] J. Zhang, W. Tan, J. Alexander, I. Foster, and R. Madduri. Recommend-As-You-Go: A Novel Approach Supporting Services-Oriented Scientific Workflow Reuse. In *2011 IEEE International Conference on Services Computing*, pages 48–55. IEEE, jul 2011.