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

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 provided by these systems. To minimize this problem, the literature presents some techniques to recommend activities during the scientific workflow construction. This paper contribution is a hybrid activity recommendation system considering information on frequency, input and outputs of activities and ontological annotations. Additionally, this project presents a modeling of activities recommendation as a classification problem, tested using 5 classifiers; 5 regressors; a SVM classifier, which uses the results of other classifiers and regressors to recommend; and an ensemble of classifiers (Rotation Forest). The proposed technique was compared to other related techniques and to classifiers and regressors, using 10-fold-cross-validation, achieving a MRR at least 70% greater than those obtained by other techniques.

1 Proposed Solution

The number of research projects using intensive computing has been growing in areas that lack advanced computer skills such as biology, physics, and astronomy (Oliveira et al. (2015)). 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 Wang et al. (2010). These systems facilitate the creation of new experiments, sharing of results and reuse of existing activities.

This paper proposes an algorithm to recommend activities (Adomavicius and Tuzhilin (2005), Garijo et al. (2014)), during the construction of an workflow, using machine learning and three important concepts in the area of scientific workflows (Lin et al. (2009)): a) frequency of activities; b) compatibility between input and output; and c) semantics of activities. To explain this proposal, the figure 1 will be used as an example. It is possible to observe six workflows with their annotations, which simulate a database of scientific workflows.

The proposed 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. This step was performed manually, automatically annotate workflows is outside the scope of this paper (Heijst et al. (1997)). Finally, the algorithm annotates all activities with the same annotations of their respective workflow; i.e., if the *X* activity (of the figure 1) is inside two workflows with distinct annotations, then this activity will be related to two different concepts from the ontology.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

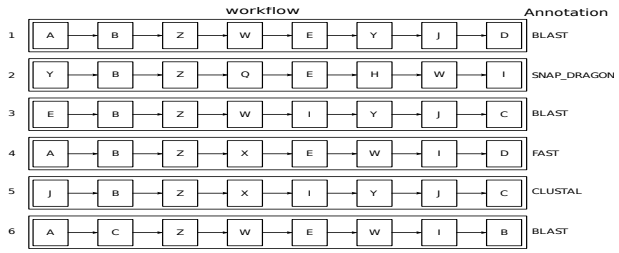

Figure 1: Example of scientific workflow database.

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 ontological 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 another new tie occurs then a random selector is used.

Table 1: Comparison between our approach (FESO) and the correlated literature

Approach	S@1	S@5	S@10	MRR
Apriori	0,0037	0,0385	0,0559	0,037
KNN_C	0,0037	0,0685	0,0959	0,040
Neural Network_C	0,0137	0,1507	0,1781	0,089
CART_C	0,0274	0,1233	0,3699	0,113
CART_R	0,1370	0,1370	0,2603	0,114
Naive Bayes_C	0,0274	0,1507	0,3425	0,114
Binomial_R	0,0822	0,1918	0,2055	0,136
Neural Network_R	0,1096	0,2603	0,2603	0,154
MARS_R	0,1233	0,2055	0,2192	0,167
SVM_R	0,1233	0,3151	0,4932	0,238
SVM_C	0,2425	0,4658	0,4932	0,244
composed SVM_C	0,2515	0,4458	0,5232	0,314
Rotation Forest_C	0,2925	0,4558	0,5432	0,324
FESO	0,3425	0,4658	0,5932	0,334

Our technique was compared with the literature (a systematic review of the literature have done by Khouri and Digiampietri (2015)) using a 10-fold cross-validation. In this technique, the dataset is divided into 10 subsets (*folds*) and ten executions are performed (Han et al. (2011)). 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 remaining 10%

It is worth noticing that 100% of the data set is labeled and one activity is removed from each workflow randomly (this activity will be the expected output by the recommendation system) 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 trained recommender system. After ten execution, the average of the metrics (metrics for recommendation system Harvey et al. (2010)) are calculated: i) Sucess at rank k $S@k$; and ii) Mean Reciprocal Rank (MRR).

The metric $S@k$ calculates the probability of an item of interest being located between the k first positions in the list of recommended activities. Its value lies between zero and one. The result 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. In the limit, the activity will always be between the L first positions, where L is the total size of the recommendations list. Thus, high values for $S@k$ are considered good, specially for low values of k. The results of the comparison could be seen in the table 1

References

Adomavicius, G. and Tuzhilin, A. (2005). Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on*, 17(6):734–749.

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

Han, J., Kamber, M., and Pei, J. (2011). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., 3rd edition.

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

Heijst, G. V., Schreiber, A., and Wielinga", B. (1997). Using explicit ontologies in {KBS} development. *International Journal of Human-Computer Studies*, 46(2–3):183 – 292.

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

Lin, C., Lu, S., Fei, X., Pai, D., and Hua, J. (2009). A Task Abstraction and Mapping Approach to the Shimming Problem in Scientific Workflows. In *2009 IEEE International Conference on Services Computing*, SCC '09, pages 284–291. IEEE Computer Society.

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

Wang, F., Deng, H., Guo, L., and Ji, K. (2010). 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.