Response Letter

Dear editor,
We would like to thank you and the referees for the comments and suggestions on the paper. The suggestions were extremely helpful and we have incorporated them all in the revised manuscript. In the following, we list in red the comments and in black the changes we have made in response to each of the referees' comments.
Kind regards,
The authors.

Question 01

I suggest the authors to rethink the title of the paper that emphases the ontology aspect of the approach. In my opinion the paper has a strong emphasis in the results presented in section 5.
Answer - We reviewed the paper's title to: "Combining artificial intelligence, ontology, and frequency-based approaches to recommend activities in scientific workflows"

Question 02

The abstract presents acronyms that are not introduced (e.g. SVM, MRR). Moreover, you say that in this project you propose the modeling. It should be: in this paper
Answer - We reviewed the abstract according to the comment: we updated "in this project" to "in this paper" and added all acronyms.

Question 03

In the introduction you mention workflow management systems but currently there are business process management systems (BPMS) in the area of business process management (BPM). I suggest the authors to discuss the differences (if the case).
Answer - We have added the following text:

"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 [4].

Typically, scientific workflows have an intensive use of computational resources and are *data flow* oriented. On the other hand, business workflows are, typically, *control flow* oriented. In this paper, we focus only on scientific workflows. Therefore, the word workflow will be used as a synonym of "scientific workflow"."

Question 04

In the second paragraph of the introduction you refer to a repository of activities. In the text, in addition to the reference number, provide the name of the repository or the authors.
Answer - We have changed the text to:

"Nowadays, there are a large number of activities available in repositories such as *myExperiment*[1] which stores more than 2,500 workflows and *BioCatalogue*[2], 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 ..."

Question 05

In the introduction you say: However, there is a precondition for reusing and/or creating workflows: knowing the available activities... What do you mean?
Answer - We have changed the text to:

"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."

Question 06

Introduction:..the second case, whose goal is to serve a less expert user, several workflows are build and the user should select which one most satisfies him/her need. What do you mean in this sentence is that several workflows are dynamically built?
Answer - We have changed the text to:

"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

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

satisfies him/her need. In the literature this second process is called: *workflow composition* ...”

Question 07

There is a gap in the introduction (second page) when you introduce your approach. You should better motivate the problem addressed in the paper before presenting your proposal (more precisely, when you start the paragraph This paper present a hybrid (second page of the Introduction).
Answer - We added the following paragraph to the text:

“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.”

Question 08

In the introduction you say: In section Results: we present and discuss the results obtained with the experiment. But which experiment do you mean? Until this point of the text you haven't introduced (even shortly) the experiment
Answer - We have changed the text to:

“In section *Results* we present and discuss a performance comparison of our approach and the approaches from the related literature is presented.”

Question 09

In the Related Work Section you say: (a) The present approach considers the order of activities as an important factor in the recommendation. In this point you should explain why the order is important; (b) Moreover, this acronym SWMS was not introduced; (c) Why you say that: In our approach, the inputs and outputs of each activity are considered, in addition to the use of a domain ontology? Improve discussion; (d) The paragraph starting with [22, 11] do not consider... is too short. Include the last name of the authors in the text and provide more information: (e) What is the meaning of these terms: simple activities, subworkflows and shims? You should have a background section where you introduce the terms and concepts used in the paper. For example, the term workflow is not defined. (f) You defined scientific workflow in the Introduction (you should inform that by simplicity you will refer to scientific workflow as workflow along the text).
Answer - (a) We have changed the text to:

“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."

(b) The acronym SWMS was introduced (Scientific Workflow Management System).

(c) Explanation about input and output - we have changed the text to: "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."

(d) We have improved the paragraph as suggested: "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)."

(e) We explained all terms inside the sections: In the introduction we added: "Workflows are models to represent a flows of interrelated activities which execution leads to a goal." In the Related Work section, we added: "... 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)."

(f) We added the following text to the introduction: "In this paper the term *workflow* is used as a synonymous of *scientific workflow*."

Question 10

In Section 3 tell the name of the program in the sentence: myExperiment repository ([14]), using the program [15]

Answer - We have changed the text to:

"The workflows were obtained from the *myExperiment* repository, using the program *wget*[3]. After downloading the 2,481 workflows in *xml* format, the *BeautifulSoup*[4] code analyzer was used to organize the dataset in a relational database."

Question 11

- Did you considered the possibility of using process mining algorithms (such as Alpha algorithm) as an alternative to the Simple Matrix solution (Section 3.1). Not sure if it is possible because it is based on process logs, but would be interesting to think about.

Answer - We did not used Alpha algorithm because we do not have a large

[3] https://www.gnu.org/software/wget/

[4] https://www.crummy.com/software/BeautifulSoup/bs4/doc/

amount of workflows of the same type (to solve the same problem). In your institution there is a group working specifically with process mining, we already discussed the use of our approach to their problem and vice-versa, but, for now, the use of process mining algorithms to mining the scientific workflows was not considered promising.

Question 12

In Section 3.2 why 118 times? Is there any reason for this number or was just defined?
Answer - This number was chosen empirically after some exploratory tests. We added to the text this explanation: "...the number 59 was chosen empirically after some exploratory tests." The number 118 corresponds to twice this value (59 positive instances and 59 negative ones).

Question 13

In Section 3.3 you tell that you used 10-fold-cross-validation technique. Justify why did you use this technique in particular.
Answer - This technique was used because it is a well-known and widely used validation strategy for artificial intelligence problems.

Question 14

Second paragraph of Section 4 is too short. Improve discussion.
Answer - We have merged the three paragraphs and expanded the discussion:

"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 **??** is be used as an example. It is possible to observe six workflows with their annotations, which simulate a database of scientific workflows."

Question 15

In Section 4 you say that the annotation of all the workflows was manually. Don't you consider a limitation of your work that could be described in the Conclusions?
Answer - We agree with the reviewer. For the datasets were there is no semantic information it is recommended to use the *Rot. Forest*$_C$ approach. We have changed the conclusion to include this limitation: "As future works, we intend

to investigate the use of data provenance to increase the accuracy of the recommendations. Moreover, we investigate study how to automatically annotated the workflows, the main limitation of the proposed solution."

Question 16

Provide more information about the ontology building (Figure 2). How was it build or add references for further information. The title of the paper includes the ontology term. So, it is expected more information.
Answer - He added a new subsection to describe the ontology creation: *Ontology construction*

Question 17

In the Results section I suggest you to add a comparative table where you include some parameters and provide a more pragmatic view of your analyses.
Answer - Due to the fact the parameters of the algorithms are different in nature and do not fit elegantly in the resulting table, we have chosen to present them only in a textual way.

Question 18

- In the Conclusion Section you mention a systematic literature review. However, you didn't talked about this review when you discuss related work. Improve your Related Work section and add a reference to the systematic literature review.
Answer - In the related work we add the reference for the systematic literature review:

"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."

Question 19

The Conclusion Section must be improved. You should provide deeper analysis of your results. How it can be used in practice. How could you extend your computational solutions to other domains or problems. Also, the limitations of your approach should be discussed.
Answer - We have improved the conclusions adding two paragraphs:

"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."

Question 20

The paper should be proofread because it has many typos in its writing. E.g. (abstract: The proposed technique were; Introduction: Section Relate Work describes the techniques proposed in the related literature are described; as a artificial intelligence; Section Conclusion the final considerations are presented; Section 5, such the one treated (use the term addressed) in this paper. Conclusions, problemrecommendation

Answer - We have reviewed the text.

Question 21

Some references look incomplete: L. Richardson. Beautiful soup, 11 2015; C. G. Roure. myexperiment, 2015.

Answer - The reference are complete now.

Bibliography

[1] E. Bomfim, J. Oliveira, J.M. 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] Adilson Lopes Khouri and Luciano Antonio 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.

[3] Yan Leng, Mahmoud El-Gayyar, and Armin. 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.

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

[5] Jing Wang, Yanbo Han, Shuying Yan, Wanghu Chen, and Guang 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.