Research Article

Heterogeneous Information Network-Based Scientific Workflow Recommendation for Complex Applications

Yiping Wen,1,2 Junjie Hou,1,2 Zhen Yuan,1 and Dong Zhou1,2

1School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan 411201, China
2Key Laboratory of Knowledge Processing and Networked Manufacture, Hunan University of Science and Technology, Xiangtan 411201, China

Correspondence should be addressed to Yiping Wen; ypwen_0@qq.com

Received 3 December 2019; Accepted 24 January 2020; Published 19 March 2020

Copyright © 2020 Yiping Wen et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Scientific workflow is a valuable tool for various complicated large-scale data processing applications. In recent years, the increasingly growing number of scientific processes available necessitates the development of recommendation techniques to provide automatic support for modelling scientific workflows. In this paper, with the help of heterogeneous information network (HIN) and tags of scientific workflows, we organize scientific workflows as a HIN and propose a novel scientific workflow similarity computation method based on metapath. In addition, the density peak clustering (DPC) algorithm is introduced into the recommendation process and a scientific workflow recommendation approach named HDSWR is proposed. The effectiveness and efficiency of our approach are evaluated by extensive experiments with real-world scientific workflows.

1. Introduction

Scientific workflow is an effective and important means to deal with data-intensive, computation-intensive, and collaboration-intensive scientific issues in many large-scale complex systems or applications from domains such as physics, astronomy, chemistry, bioinformatics, and life sciences [1–3]. In practice, many scientific workflows have been successfully deployed and executed on clouds. Recently, with the quick development of smart user devices and edge computing, a number of studies have been carried out to construct and execute workflows in a cloud-edge collaborative manner [4, 5].

Scientific workflow modelling plays an important role in complex scientific workflow applications, which is a not only complex but also error-prone process. In recent years, more and more scientific workflows have been published onto the Web and shared in some repositories such as CrowdLabs, SHIWA, Galaxy, and the myExperiment [6, 7]. People can leverage and repurpose a part of existing scientific workflows for specific complex applications, rather than constructing new ones from scratch. However, with the growth of the amount of scientific workflows, finding suitable scientific workflows from a sea of candidates becomes a new problem for scientists and engineering personnel. Though process retrieval methods can help to handle this problem by retrieving similar scientific workflow fragments from repositories, much manual work is still required. Consequently, to provide better automatic support, it is necessary to build effective scientific workflow recommendation techniques, which is fundamental for the reuse and repurposing of current scientific workflows.

In scientific workflow repositories, various types of data can be used for recommendation, including scientific workflow structure and annotation. However, the tags of scientific workflows are usually neglected by existing scientific workflow recommendation methods. In fact, the tags of scientific workflows contain much valuable information and different underlying logical relations among scientific workflows which can be explored via them. For example, many tags in the myExperiment repository are substantially shared by multiple scientific workflows and there exist partial similarity relations among these scientific workflows. Therefore, integrating tags and other information of
scientific workflows is promising to generate more accurate recommendations.

On the other hand, heterogeneous information network (HIN) has been proved to be a powerful modelling method to incorporate various heterogeneous types of information and it has been successfully applied in recommender systems [8, 9]. Motivated by the HIN-based recommendation idea and data characteristics of scientific workflow repository, we plan to integrate multiple types of scientific workflow data into the form of HIN and use a metapath-based technique to measure similarity and calculate distance between scientific workflows, by which multiple metapaths can be combined with the semantic description information of scientific workflows and more accurate similarity computation results would be obtained.

With these observations, in this paper, we propose a heterogeneous information network-based approach for recommending scientific workflows to scientists and engineers. In our approach, different data objects and underlying logical relations on scientific workflows are organized as a HIN, according to which the similarity between scientific workflows is evaluated. In addition, to facilitate the reuse and repurposing of current scientific workflows, the density peak clustering (DPC) algorithm [10] is introduced and used to group candidates into clusters. Our main contributions are summarized as follows:

1. We propose a new representation form of scientific workflow based on HIN, which is enriched through incorporation of multiple types of data including tags and logical relations of such data.

2. We build a metapath-based method to assess the similarity between scientific workflows, where the similarity is calculated according to objects of tag, description, activity, and subscientific workflow involved in scientific workflows.

3. We present a HIN- and DPC-based scientific workflow recommendation approach named HDSWR to generate more accurate recommendations and, on the basis of it, to facilitate the reuse and repurposing of current scientific workflows for scientists and engineers.

4. We provide two real-world datasets with tags on scientific workflows for experiments.

The remainder of this paper is organized as follows. Section 2 describes the related studies. Section 3 introduces some notations and basic definitions used in the paper. Section 4 presents the scientific workflow similarity computation method. In Section 5, we propose the HDSWR approach. Then, we evaluate our method in Section 6. Section 7 concludes this paper.

2. Related Work

In this section, we briefly review related work on the workflow models, workflow recommendations, and HIN.

A workflow model is fundamental for various workflow applications. In practice, workflows can be modelled by different tools such as directed acyclic graphs (DAGs), Petri nets, event-driven process chains (EPCs), the business process execution language (BPEL), or the fairly complex business process modelling notation (BPML) language which has over 100 symbols [11]. However, modelling workflows is always a knowledge-intensive and laborious task. To improve workflow modelling, methods such as workflow mining [12] have been proposed to discover workflow models from event logs. However, similar to process retrieval, much manual work is still involved.

In recent years, some workflow recommendation approaches have been proposed. Current techniques can be mainly classified into two types: business workflow (process) recommendation and scientific workflow recommendation.

In the business process management domain, business workflow is usually modelled with block structures including sequential structures, alternative structures, parallel structures, and iterative structures. So far, only a limited number of business workflow recommendation methods have been proposed to serve different purposes, which can be classified into complete process recommendation and process fragments (nodes) recommendation [13]. For example, Zhang et al. [14] leveraged workflow provenance to recommend a set of nodes for a partial workflow. Li et al. [15] adopted minimum depth-first-search codes and string edit distances for representing and recommending business workflow fragments. Deng et al. [13] developed a recommendation system to generate a sorted candidate node sets, which used a subgraph mining method to extract patterns from process repositories. Wang et al. [16] utilized the properties of business process repositories and proposed a representation-learning-based recommendation method.

Scientific workflows are based on the automation of scientific process which is typically composed of multiple scientific programs or Web services. Compared with business workflows, scientific workflows have a strong focus on the dataflow to sufficiently support a variety of data-intensive applications, in which the control structure just simply describes the partial ordering of tasks. Therefore, scientific workflows are usually modelled with unstructured DAGs, which conceptually use a set of nodes and edges instead of complex block structures. However, similar to business workflow recommendation, there are two kinds of work in scientific workflow recommendation. For instance, Zhang et al. [17] used the term of unit of work (UoW) to represent a collection of services (i.e., fragments of a scientific workflow) chained together, based on which a UoW-driven scientific workflow recommendation framework and three algorithms for UoW mining and recommendation are proposed. Cheng et al. [18, 19] converted a scientific workflow into a layered hierarchy in terms of a tree style, where the hierarchical relations specify the links between a scientific workflow, its subworkflows, and activities. Based on it, a semantic similarity computation algorithm considering the lay hierarchy and description of scientific workflows is proposed for clustering and recommending appropriate scientific workflows. Krzywucki and Polak [20] utilized semantic-type comparison to evaluate the similarity of scientific workflows. Bergmann et al. [21] proposed a
semantic workflow graph-based method for modelling scientific workflow similarity and developed an A* search-based algorithm for workflow similarity computation. Starlinger et al. [7] presented a layer decomposition approach for the comparison and similarity search of scientific workflow. Mohan et al. [22] developed several folksonomy-based scientific workflow recommendation strategies and implemented them in a prototype system.

HIN is a newly emerging direction in recommender systems and a good candidate for improving the accuracy of recommendations. However, to the best of our knowledge, HIN is normally neglected in the workflow recommendation literature. So far, most of the HIN-based recommendation methods consider the metapath-based similarity. For example, Sun et al. [8] investigated a similarity search problem in HIN and introduced the concept of metapath-based similarity. Zhao et al. [23] introduced the concept of metapath to incorporate more complex semantics for HIN-based recommendation. Shi et al. [24] developed a meta-graph to incorporate more complex semantics for HIN and introduced the concept of metagraph-based recommendation. The other hand, scientific workflows in repositories have rich tag information, which are seldom exploited by existing workflow recommendation methods. Some research work related to tags has been done in the domain of Service Computing [25] and other related research work on service recommendation was carried out in [26]. Our previous work in [27] has preliminarily utilized scientific workflow tags for recommendation. In this paper, we further organize scientific workflows and their relations as a HIN to calculate the similarity of scientific workflows and generate more accurate recommendations.

3. Preliminaries

To make our approach well understood, we first introduce HIN and relevant concepts in this section. The notations we will use throughout this paper are summarized in Table 1.

**Definition 1** (Scientific Workflow [18]). A scientific workflow $sw$ is a tuple $(nm, sw_dsc, sw_D, sw_A, sw_L, and sw_T)$, where $nm$ and $sw_dsc$ are the name and text description of $sw$, respectively. $sw_D$ is the set of subscientific workflows that $sw$ invokes. $sw_A$ is the activity set of $sw$. $sw_L$ denotes a set of links connecting activities and subscientific workflows in $sw$. $sw_T$ is a set of tags on $sw$.

Generally, a subscientific workflow can be regarded as a scientific workflow [7]. For example, in the myExperiment repository, a subscientific workflow is stored as an independent scientific workflow.

**Definition 2** (Heterogeneous Information Network [24, 28]). A heterogeneous information network is defined as a direction graph $G = (V, E)$ with an object-type mapping function $\phi : V \rightarrow B$ and a link-type mapping function $\psi : E \rightarrow R$, satisfying $|B| + |R| > 2$.

**Definition 3** (HIN-Based Scientific Workflow Representation). The scientific workflow can be organized and represented as a heterogeneous information network, which contains five object types: scientific workflow (denoted as $SW$), tag (denoted as $T$), activity (denoted as $A$), subscientific workflow (denoted as $D$), and description (denoted as $dsc$). Each scientific workflow can link with a set of tags, a set of activities, and a set of subscientific workflows, and a description.

**Example 1.** An example of HIN-based scientific workflow representation is shown in Figure 1, which consists of two real-world scientific workflows named Chemical2URLs (https://www.myexperiment.org/workflows/97.html) (denoted as $sw_1$) and DF CUAM (https://www.myexperiment.org/workflows/4700.html) (denoted as $sw_2$).

The $sw_1$ links with a text description ($dsc_1$), three tags (annotation, chemspider and cheminformatics), two activities (REST_Service and Xpath_Service), and two subscientific workflows (CNTCI and workflow40).

Besides, $sw_1$ and $sw_2$ are linked by two tags (cheminformatics and chemspider), which are shared by $sw_1$ and $sw_2$. Similarly, if some objects of subscientific workflow, activity, or description are shared by two scientific workflows, there exists some link relation between these two scientific workflows.

**Definition 4** (Network Schema [24, 28]). The network schema is a meta template for a heterogeneous information network $G = (V, E)$ with the object-type mapping function $\phi : V \rightarrow B$ and the link-type mapping function $\psi : E \rightarrow R$, which is a directed graph $S = (B, R)$ defined over object types $B$ and link types $R$.

According to Definition 4, we can construct a HIN-based scientific workflow representation schema, which is shown in Figure 2. There are five types of objects: scientific workflow ($SW$), tag ($T$), activity ($A$), subscientific workflow ($D$), and description ($dsc$). Besides, there exist four types of links between objects to represent different relations:

1. A link relation between a scientific workflow and a tag.
2. A link relation between a scientific workflow and an activity.
3. A link relation between a scientific workflow and a subscientific workflow.
4. A link relation between a scientific workflow and a description. Such link relation is single-way, because a specific text description belongs to a specific scientific workflow.

**Definition 5** (Metapath [8, 24]). A metapath $p$ is a path defined on a network schema $S = (B, R)$ and is represented in the form of $B_1 \rightarrow B_2 \rightarrow \cdots \rightarrow B_{n+1}$ and thus defines a composite relationship $R = R_1 \circ R_2 \circ \cdots \circ R_k$ between two object types $B_1$ and $B_{n+1}$, where $\circ$ denotes the composition operator on relations $R$. 
According to Definition 5 and the HIN-based scientific workflow representation schema, we can construct four types of metapaths, which are shown in Figure 3:

1. Metapath $p_1$: if a tag is shared by two scientific workflows $sw_1$ and $sw_2$, we can use the metapath $\text{SW} \xrightarrow{T} \text{Tag} \xrightarrow{\text{SW}}$ to indicate a cotag relation between $sw_1$ and $sw_2$.

2. Metapath $p_2$: if an activity is shared by two scientific workflows $sw_1$ and $sw_2$, we can use the metapath $\text{Annotation} \xrightarrow{\text{metabolomics}} \text{Label}$ to indicate an activity relation between $sw_1$ and $sw_2$.

According to Definition 5 and the HIN-based scientific workflow representation schema, we can construct four types of metapaths, which are shown in Figure 3:

### Table 1: Notations and explanations.

<table>
<thead>
<tr>
<th>Notation</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$SW$</td>
<td>A list of scientific workflows</td>
</tr>
<tr>
<td>$dsc$</td>
<td>A description type of object</td>
</tr>
<tr>
<td>$D$</td>
<td>A subscientific workflow type of object</td>
</tr>
<tr>
<td>$A$</td>
<td>An activity type of object</td>
</tr>
<tr>
<td>$T$</td>
<td>A tag type of object</td>
</tr>
<tr>
<td>$p_1, p_2, p_3, p_4$</td>
<td>Different types of metapaths</td>
</tr>
<tr>
<td>$\text{SWT, SWA, SWD}$</td>
<td>Adjacent matrices on the objects of tag, activity, and subscientific workflow, respectively</td>
</tr>
<tr>
<td>$v_T^i, v_A^i, v_D^i$</td>
<td>Feature vectors of scientific workflow $sw_i$ in the adjacent matrices $\text{SWT}, \text{SWA}$, and $\text{SWD}$, respectively</td>
</tr>
<tr>
<td>$C_{ij}^{p_1T}, C_{ij}^{p_2A}, C_{ij}^{p_3D}$</td>
<td>The similarity strength of scientific workflows $sw_i$ and $sw_j$ on metapaths $p_1$, $p_2$, and $p_3$, respectively</td>
</tr>
<tr>
<td>$\alpha, \beta, \gamma, \delta$</td>
<td>Weight coefficients</td>
</tr>
</tbody>
</table>

Figure 1: An example of HIN-based scientific workflow representation.

Figure 2: HIN-based scientific workflow representation schema.

Figure 3: Four types of metapaths under the HIN-based scientific workflow representation schema.

(1) Metapath $p_1$: if a tag is shared by two scientific workflows $sw_1$ and $sw_2$, we can use the metapath $\text{SW} \xrightarrow{T} \text{Tag} \xrightarrow{\text{SW}}$ (Scientific Workflow $\xrightarrow{T}$ Tag $\xrightarrow{\text{SW}}$ Scientific Workflow) to indicate a cotag relation between $sw_1$ and $sw_2$.

(2) Metapath $p_2$: if an activity is shared by two scientific workflows $sw_1$ and $sw_2$, we can use the metapath $\text{Annotation} \xrightarrow{\text{metabolomics}} \text{Label}$ (Annotation $\xrightarrow{\text{metabolomics}}$ Label) to indicate an activity relation between $sw_1$ and $sw_2$. 

(3) Metapath $p_3$: if a subscientific workflow is shared by two scientific workflows $sw_1$ and $sw_2$, we can use the metapath $\text{Activity} \xrightarrow{T} \text{Subscientific Workflow} \xrightarrow{\text{SW}}$ (Activity $\xrightarrow{T}$ Subscientific Workflow $\xrightarrow{\text{SW}}$ Scientific Workflow) to indicate a subscientific workflow relation between $sw_1$ and $sw_2$.

(4) Metapath $p_4$: if a description is shared by two scientific workflows $sw_1$ and $sw_2$, we can use the metapath $\text{Description} \xrightarrow{T} \text{Subscientific Workflow} \xrightarrow{\text{SW}}$ (Description $\xrightarrow{T}$ Subscientific Workflow $\xrightarrow{\text{SW}}$ Scientific Workflow) to indicate a description relation between $sw_1$ and $sw_2$. 

(5) Metapath $p_5$: if a tag is shared by two subscientific workflows $sw_1$ and $sw_2$, we can use the metapath $\text{Tag} \xrightarrow{T} \text{Subscientific Workflow} \xrightarrow{\text{SW}}$ (Tag $\xrightarrow{T}$ Subscientific Workflow $\xrightarrow{\text{SW}}$ Scientific Workflow) to indicate a subscientific workflow relation between $sw_1$ and $sw_2$. 

(6) Metapath $p_6$: if an activity is shared by two subscientific workflows $sw_1$ and $sw_2$, we can use the metapath $\text{Annotation} \xrightarrow{\text{metabolomics}} \text{Label}$ (Annotation $\xrightarrow{\text{metabolomics}}$ Label) to indicate an activity relation between $sw_1$ and $sw_2$. 

(7) Metapath $p_7$: if a description is shared by two subscientific workflows $sw_1$ and $sw_2$, we can use the metapath $\text{Description} \xrightarrow{T} \text{Subscientific Workflow} \xrightarrow{\text{SW}}$ (Description $\xrightarrow{T}$ Subscientific Workflow $\xrightarrow{\text{SW}}$ Scientific Workflow) to indicate a description relation between $sw_1$ and $sw_2$. 

(8) Metapath $p_8$: if a tag is shared by two descriptions $dsc_1$ and $dsc_2$, we can use the metapath $\text{Tag} \xrightarrow{T} \text{Description} \xrightarrow{\text{SW}}$ (Tag $\xrightarrow{T}$ Description $\xrightarrow{\text{SW}}$ Scientific Workflow) to indicate a description relation between $dsc_1$ and $dsc_2$. 

(9) Metapath $p_9$: if an activity is shared by two descriptions $dsc_1$ and $dsc_2$, we can use the metapath $\text{Annotation} \xrightarrow{\text{metabolomics}} \text{Label}$ (Annotation $\xrightarrow{\text{metabolomics}}$ Label) to indicate an activity relation between $dsc_1$ and $dsc_2$. 

(10) Metapath $p_{10}$: if a description is shared by two activities $act_1$ and $act_2$, we can use the metapath $\text{Description} \xrightarrow{T} \text{Activity} \xrightarrow{\text{SW}}$ (Description $\xrightarrow{T}$ Activity $\xrightarrow{\text{SW}}$ Scientific Workflow) to indicate an activity relation between $act_1$ and $act_2$. 

(11) Metapath $p_{11}$: if a tag is shared by two activities $act_1$ and $act_2$, we can use the metapath $\text{Tag} \xrightarrow{T} \text{Activity} \xrightarrow{\text{SW}}$ (Tag $\xrightarrow{T}$ Activity $\xrightarrow{\text{SW}}$ Scientific Workflow) to indicate an activity relation between $act_1$ and $act_2$. 

(12) Metapath $p_{12}$: if a description is shared by two tags $tag_1$ and $tag_2$, we can use the metapath $\text{Description} \xrightarrow{T} \text{Tag} \xrightarrow{\text{SW}}$ (Description $\xrightarrow{T}$ Tag $\xrightarrow{\text{SW}}$ Scientific Workflow) to indicate a tag relation between $tag_1$ and $tag_2$. 

(13) Metapath $p_{13}$: if an activity is shared by two tags $tag_1$ and $tag_2$, we can use the metapath $\text{Annotation} \xrightarrow{\text{metabolomics}} \text{Label}$ (Annotation $\xrightarrow{\text{metabolomics}}$ Label) to indicate an activity relation between $tag_1$ and $tag_2$. 

(14) Metapath $p_{14}$: if a description is shared by two tags $tag_1$ and $tag_2$, we can use the metapath $\text{Description} \xrightarrow{T} \text{Tag} \xrightarrow{\text{SW}}$ (Description $\xrightarrow{T}$ Tag $\xrightarrow{\text{SW}}$ Scientific Workflow) to indicate a tag relation between $tag_1$ and $tag_2$.
In this section, it mainly consists of four steps.

(3) Metapath $p_1$: If a subscientific workflow is shared by two scientific workflows $sw_1$ and $sw_2$, we can use the metapath $SWDSW$ (Scientific Workflow → Sub-Scientific Workflow → Scientific Workflow) to denote a relation between $sw_1$ and $sw_2$ on a sub-scientific workflow.

(4) Metapath $p_2$: If a description is shared by two scientific workflows $sw_1$ and $sw_2$, we can use the meta-path $SwdsSw$ (Scientific Workflow → dsc → Scientific Workflow) to denote a relation between $sw_1$ and $sw_2$ on a description.

4. Similarity Computation for Scientific Workflows

Based on the basic definitions mentioned above, we propose a novel scientific workflow similarity computation method in this section. It mainly consists of four steps.

Step 1: Construct three adjacent matrices on the objects of tag, activity, and subscientific workflow.

According to the objects of tag, activity, and subscientific workflow involved in the scientific workflows, we can construct three adjacent matrices, respectively, denoted as $SWT$, $SWA$, and $SWD$. A row of the adjacent matrices corresponds to a specific scientific workflow. A column of the adjacent matrices $SWT$, $SWA$, and $SWD$ corresponds to a specific object of tag, activity, and subscientific workflow, respectively. The values in these three adjacent matrices can be 1 or 0, which denotes whether a specific object belongs to a specific scientific workflow.

Besides, for computational convenience, we use the feature vector $v^T_{i,j}$ to represent the relation between the scientific workflow $sw_i$ and all the objects of tag involved, which corresponds to a row in the adjacent matrix $SWT$. Likewise, we use the feature vector $v^A_{i,j}$ and $v^D_{i,j}$ to represent the relations between the scientific workflow $sw_i$ and the objects of activity and subscientific workflow involved, respectively, which correspond to a row in the adjacent matrices of $SWA$ and $SWD$, respectively.

Step 2: Calculate the similarity on the metapaths.

As mentioned in Section 3, there exist four types of metapaths. Therefore, the similarity strength of $sw_i$ and $sw_j$ on meta-path $p_1$ can be calculated by the following equation:

$$C_{i,j}^{p_{1,TA}} = v^T_i \cdot (v^T_j)^t.$$  \hfill (1)

In equation (1), $v^T_i$ and $v^T_j$ are two feature vectors of scientific workflow $sw_i$ and $sw_j$ on tags, respectively. $(v^T_j)^t$ is the transpose of the feature vector $v^T_j$. The higher the number of common tags between $sw_i$ and $sw_j$, the greater the inner product of the $v^T_i$ and $(v^T_j)^t$.

and thus, the more the similarity between $sw_1$ and $sw_2$ on the tag. The meaning of notations in equations (2) and (3) is similar to these in equation (1).

Likewise, the similarity strength of $sw_i$ and $sw_j$ on metapaths $p_2$ and $p_3$ can be obtained by equations (2) and (3), respectively, where the meaning of notations is similar to these in the following equation:

$$C_{i,j}^{p_{2,TA}} = v^A_i \cdot (v^A_j)^t,$$ \hfill (2)

$$C_{i,j}^{p_{3,DA}} = v^D_i \cdot (v^D_j)^t.$$ \hfill (3)

Based on equation (1), we can also obtain the values of $C_{i,j}^{p_{1,TA}}$ and $C_{i,j}^{p_{2,TA}}$. To normalize the similarity strength effectively, we utilize the ratio between the $C_{i,j}^{p_{1,TA}}$ and the max one in the $C_{i,j}^{p_{1,TA}}$ and $C_{i,j}^{p_{2,TA}}$ to represent the similarity between scientific workflows $sw_i$ and $sw_j$ with respect to metapath $p_1$, which is described as follows:

$$sim_{p_1}(i, j) = \frac{C_{i,j}^{p_{1,TA}}}{\max(C_{i,j}^{p_{1,TA}}, C_{i,j}^{p_{2,TA}})}.$$ \hfill (4)

Analogously, the similarity between scientific workflows $sw_i$ and $sw_j$ with respect to metapaths $p_2$ and $p_3$ is described as follows:

$$sim_{p_2}(i, j) = \frac{C_{i,j}^{p_{2,TA}}}{\max(C_{i,j}^{p_{2,TA}}, C_{i,j}^{p_{2,TA}})}.$$ \hfill (5)

$$sim_{p_3}(i, j) = \frac{C_{i,j}^{p_{3,DA}}}{\max(C_{i,j}^{p_{3,DA}}, C_{i,j}^{p_{3,DA}})}.$$ \hfill (6)

Step 3: Calculate the similarity value on the descriptions of scientific workflows.

The doc2vec model can learn the fixed-length feature from the variable-length text [29]. Therefore, we utilize the doc2vec model to form the paragraph vectors $v_{sw_i}$ and $v_{sw_j}$ for the descriptions of scientific workflows $sw_i$ and $sw_j$, respectively. Besides, the normalized cosine similarity between $v_{sw_i}$ and $v_{sw_j}$ is calculated as the similarity value on the descriptions of scientific workflows $sw_i$ and $sw_j$, which is described as:

$$sim_{p_4}(i, j) = \frac{(v_{sw_i} \cdot v_{sw_j})((\|v_{sw_i}\| \cdot \|v_{sw_j}\|) + 1)}{2}.$$ \hfill (6)

In equation (6), the notations $v_{sw_i}$ and $v_{sw_j}$ represent the norm of the paragraph vectors $v_{sw_i}$ and $v_{sw_j}$, respectively.

Step 4: Summarize different similarity values.

To effectively fuse different similarities of scientific workflows obtained by the above steps, we introduce the weighting mechanism, which is described as:
\[ \text{sim}(i, j) = \alpha \times \text{sim}_{p1}(i, j) + \beta \times \text{sim}_{p2}(i, j) + \gamma \times \text{sim}_{p2}(i, j) + \rho \times \text{sim}_{p3}(i, j). \] (7)

In equation (7), \( \alpha, \beta, \gamma, \) and \( \rho \) are the weight coefficients satisfying \( \alpha + \beta + \gamma + \rho = 1. \)

5. HDSWR Approach

To improve the accuracy and efficiency of scientific workflow recommendation, we propose an approach named HDSWR. In this section, we provide an overview of the HDSWR and introduce its related function algorithms in detail.

5.1. Overview of the HDSWR Approach. The proposed HDSWR approach is shown in Algorithm 1, which consists of four steps:

Step 1 (line 1): we construct a matrix to denote the similarity values between scientific workflows in the list \( \text{SW} \), which may come from some scientific workflows repository. All the scientific workflows in the list \( \text{SW} \) are organized as a HIN for similarity computation.

Step 2 (line 2): we adopt the density peak clustering (DPC) algorithm [10] to group all the scientific workflows in the list \( \text{SW} \) into multiple different clusters, where the similarity values in the matrix are used as the distances between scientific workflows and \( \text{Clusters} \) denotes a set of clusters on scientific workflows.

Step 3 (lines 3–4): according to textual description in the requirement of scientists and engineering personnel, i.e., \( \text{requirement.dscs} \), we search and choose appropriate objects of activity and subscientific workflows involved in the list \( \text{SW} \), where \( \text{D}_{\text{amp}} \) and \( \text{A}_{\text{amp}} \) denote a set of subscientific workflows and a set of activities, respectively. Then, a HIN-based sample scientific workflow \( \text{sw}_{\text{amp}} \) can be constructed (Line 4).

Step 4 (line 5): according to the sample scientific workflow \( \text{sw}_{\text{amp}} \), we firstly select an appropriate group of scientific workflows in the set \( \text{Clusters} \) by the similarity values between \( \text{sw}_{\text{amp}} \) and different clusters. Then, a list \( \text{SW}_{\text{rec}} \) is generated for recommendation, where the number of scientific workflows in the list \( \text{SW}_{\text{rec}} \) is related to the parameter of \( \text{rec}_K \).

5.2. Similarity Computation. Assessing workflow similarity is important for workflow recommendation. Its main purpose is to measure the distances between workflows. Based on the scientific workflow similarity computation method introduced in Section 4, the function \( \text{ComputeSimilarity} \) is described as Algorithm 2.

Example 2. The scientific workflows \( \text{sw}_1 \) and \( \text{sw}_2 \) in Figure 1 can be used as an example. As illustrated by Figure 1, there are four tags (\( \text{annotation}, \text{chenspider}, \text{cheminformatics}, \) and \( \text{metabolomics} \)) involved in the scientific workflows \( \text{sw}_1 \) and \( \text{sw}_2 \). Therefore, as shown in Figure 4(a), the corresponding value on these four tags in the adjacent matrix \( \text{SWT} \) is 1 or 0 with respect to the \( \text{sw}_1 \) and \( \text{sw}_2 \), where the value of 0 denotes that such tag does not belong to some scientific workflow. Similarly, the matrix \( \text{SWA} \) in Figure 4(b) shows the corresponding values on the activities of the scientific workflows \( \text{sw}_1 \) and \( \text{sw}_2 \), and the matrix \( \text{SWD} \) in Figure 4(c) shows the corresponding values on the subscientific workflows. Besides, the feature vectors of \( v_1^A, v_1^D, \) and \( v_1^T \) are also illustrated by Figure 4.

5.3. DPC-Based Clustering of Scientific Workflows. To improve the efficiency of recommendation, we introduce the clustering strategy proposed in [10, 30, 31], by which the scientific workflows are grouped and divided into different clusters for further recommendation. Different from the work in [10, 30, 31], we choose the density peak clustering (DPC) algorithm [10] as our clustering method, because it can effectively identify clusters with different distribution shapes and it is rarely affected by noise points. Based on the DPC algorithm, the function \( \text{DPCClustering} \) can be described as Algorithm 3.

5.4. Retrieval of Appropriate Activities and Subscientific Workflows. According to the modelling requirement of scientists and engineering personnel, we can search in the scientific workflow list and get appropriate activities and subscientific workflows, which can be used to construct a sample scientific workflow and guide the recommendation process. Such procedure is performed by the function \( \text{GetActivitySubWF} \), which is described as Algorithm 4.

In Algorithm 4, because the descriptions \( \text{requirement.dscs} \) provided in the requirement are related to activities or subscientific workflows, the best matching result on each description in \( \text{requirement.dscs} \) may be an activity or a subscientific workflow. Therefore, we calculate the similarity values on activities and subscientific workflows,
Input:
(i) SW: a list of scientific workflows.
(ii) requirement: a modelling requirement, denoted as \((dsc_{imp}, T_{imp}, dscs)\).
(iii) \(\alpha, \beta, \gamma, \delta\): parameters for similarity computation.
(iv) \(rec_K\): a parameter on the number of recommended scientific workflows.

Output:
(i) \(SW_{rec}\): a list of recommended scientific workflows.

ALGORITHM 1: HDSWR (SW, requirement, rec_K, \(\alpha, \beta, \gamma, \delta\)).

ALGORITHM 2: Function ComputeSimilarity (SW, \(\alpha, \beta, \gamma, \delta\)).

Example:

<table>
<thead>
<tr>
<th></th>
<th>annotation</th>
<th>cheminformatics</th>
<th>chemspider</th>
<th>metabolomics</th>
</tr>
</thead>
<tbody>
<tr>
<td>SW_1</td>
<td>...</td>
<td>1</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>SW_2</td>
<td>...</td>
<td>0</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>SW_i</td>
<td>...</td>
<td>1</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>[SW]</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

(a) Figure 4: Continued.
Input:
(i) Matrix: a similarity matrix of scientific workflows.
(ii) SW: a list of scientific workflows.

Output:
(i) Clusters: the set of generated scientific workflow clusters.

(1) \( \text{dist} \leftarrow 1 - \text{Matrix} \)
(2) \( \text{dc} \leftarrow \) select a value from the dist so that the number of values below it is around 1 to 2% of the total number of values in the dist
(3) \( \text{for each scientific workflow } sw_i \text{ in } SW \text{ do} \)
(4) \( \text{ld}_i \leftarrow 0 \)
(5) \( \text{for each scientific workflow } sw_j \text{ in } SW \text{ do} \)
(6) \( \text{if dist}_{i,j} < \text{dc} \text{ then} \)
(7) \( \text{ld}_i \leftarrow \text{ld}_i + 1 \)
(8) \( \text{end if} \)
(9) \( \text{end for} \)
(10) \( \text{end for} \)
(11) \( \text{for each scientific workflow } sw_i \text{ do} \)
(12) \( \text{rd}_i \leftarrow \max(\text{dist}) \)
(13) \( \text{for each scientific workflow } sw_j \text{ do} \)
(14) \( \text{if ld}_i < \text{ld}_j \text{ and rd}_i < \text{dist}_{i,j} \text{ then} \)
(15) \( \text{rd}_i \leftarrow \text{dist}_{i,j} \)
(16) \( \text{end if} \)
(17) \( \text{end for} \)
(18) \( \text{end for} \)
(19) \( \text{Clusters} \leftarrow \text{clustering scientific workflows by the DPC algorithm with the local density values such as ld}_i \text{ and relative distances values such as rd}_i \)
(20) \( \text{return Clusters} \)

Algorithm 3: Function DPCClustering (Matrix, SW).

Figure 4: An example of three adjacent matrices and feature vectors. (a) SWT. (b) SWA. (c) SWD.
respectively, where the working procedure of the function \( \text{cosine\_sim} \) in lines 6 and 13 is similar to that of equation (6).

Besides, for each description in \( \text{requirement\_dscs} \), we search the best matching activity (lines 5–11) and the best matching subscientific workflow for it (lines 12–18), then we choose the better one for constructing a sample science workflow (lines 20–24).

5.5. Generation of Scientific Workflow Candidate List. Once a sample science workflow is constructed, we can generate a list of scientific workflows that are most relevant to it, the whole procedure of which is described as Algorithm 5.

The Algorithm 5 mainly consists of three steps.

Step 1 (line 1): as introduced before, we can construct the feature vectors of the sample scientific workflow \( sw_{\text{amp}} \) on the objects of activity, subscientific workflow, and tag.

Step 2 (lines 3–14): we compute the similarity between the sample scientific workflow \( sw_{\text{amp}} \) and the cluster center scientific workflow first (lines 4–8), where the procedure is performed according to the method introduced in Section 4. Then, a cluster is selected as \( \text{cluster}_{\text{amp}} \) if the similarity value between its cluster center and \( sw_{\text{amp}} \) is the largest among all the clusters (lines 9–13).

Step 3 (line 15): after the cluster \( \text{cluster}_{\text{amp}} \) is determined, the \( \text{rec}_K\% \) scientific workflows of the \( \text{cluster}_{\text{amp}} \), which are most related to \( sw_{\text{amp}} \) in similarity values are chosen as candidate scientific workflows and recommended in a list.

5.6. An Example on Textual Descriptions. So far, research studies for recommending whole scientific workflows typically adopt the scientists’ requirements for recommendation. For example, Cheng et al. [18] used a layer hierarchy with respect to the scientist’s requirement. In our approach, we mainly adopt textual descriptions with respect to the scientist’s requirement. For ease of illustration, the scientific workflow \( sw_1 \) in Figure 1 is used as an example on textual descriptions.

Example 3. As illustrated by Figure 1, there exists a subscientific workflow named \( \text{CNTCI} \), which is short for \( \text{Chemical\_Name\_To\_Chemspider\_ID} \), and a subscientific workflow named \( \text{Workflow40} \) in the scientific workflow \( sw_1 \). We can get the textual descriptions of the \( sw_1 \), i.e., “This workflow will map a chemical name or identifier to uniform
Input:
(i) $sw_{imp}$: a sample scientific workflow.
(ii) $Clusters$: the set of scientific workflow clusters.
(iii) rec$_K$: a hyper-parameter to control the number of recommend scientific workflows.
(iv) $\alpha, \beta, \gamma, \delta$: weight coefficients.

Output:
(i) $SW_{rec}$: a list of recommended scientific workflows.

\begin{algorithm}
\begin{align*}
\text{(1)} & \quad v_{w1}^{imp}, v_{w2}^{imp} \leftarrow \text{construct the feature vector on the activities, tags and sub-scientific workflows of } sw_{imp}. \\
\text{(2)} & \quad sim_{imp} \leftarrow 0 \text{ and } cluster_{imp} \leftarrow \emptyset \\
\text{(3)} & \quad \text{for each cluster } c \in Clusters \text{ do} \\
\text{(4)} & \quad sw_{ct} \leftarrow \text{choose the cluster center scientific workflow of the cluster} \\
\text{(5)} & \quad v_{ct1}^{imp}, v_{ct2}^{imp} \leftarrow \text{construct the feature vector on the activities, tags and sub-scientific workflows of } sw_{ct}. \\
\text{(6)} & \quad C_{imp,ct}^{p1,2}, C_{imp,ct}^{p2,3}, C_{imp,ct}^{p3,0} \\
\text{(7)} & \quad \text{calculate sim}_{p1}(sw_{imp}, ct), sim_{p2}(sw_{imp}, ct), sim_{p3}(sw_{imp}, ct), sim_{p4}(sw_{imp}, ct) \\
\text{(8)} & \quad \text{calculate sim}(sw_{imp}, ct) \\
\text{(9)} & \quad sim_{imp} \leftarrow \text{sim}(sw_{imp}, ct) \\
\text{(10)} & \quad \text{if } sim_{imp} < sim_{imp} \text{ then} \\
\text{(11)} & \quad sim_{imp} \leftarrow sim_{imp} \\
\text{(12)} & \quad cluster_{imp} \leftarrow \text{cluster} \\
\text{(13)} & \quad \text{end if} \\
\text{(14)} & \quad \text{end for} \\
\text{(15)} & \quad SW_{rec} \leftarrow \text{choose the top } rec_K \% \text{ most similar scientific workflows in } cluster_{imp} \\
\text{(16)} & \quad \text{return } SW_{rec}
\end{align*}
\end{algorithm}

6. Experiments

In this section, a series of experiments are performed to answer two questions: (1) Compared with the state-of-the-art scientific workflow recommendation techniques, does our approach have better performance? (2) What is the performance of our HDSWR approach in the presence of different parameters and datasets used for recommendation?

All experiments are performed on a computer with Intel (R) Core (TM) i5-7300HQ CPU@ 2.50 GHz 2.50 GHz and 8 GB memory running Window 10, JDK 1.8.0 and python 3.5. Next, we focus on experimental evaluations of these two questions.

6.1. Datasets. The myExperiment is a widely used scientific workflow repository supporting the publication and sharing of scientific workflows. It also allows scientists to search scientific workflows related to their research and then reuse and repurpose scientific workflows according to their distinct needs [31]. There are various types of scientific workflows in the myExperiment, such as Tarvena1 and Tarvena2. We crawled related data on the Tarvena2 type of scientific workflows from the myExperiment and created two datasets named SW#80 and SW#236 accordingly. The datasets used in our experiments are publicly accessible from GitHub via the website: https://github.com/yixinxunwu/myExperiment.

As Table 2 shows, the SW#80 dataset includes 80 scientific workflows with 229 activities, 125 tags, and 85 subscientific workflows, where the number of activities contained in each scientific workflow is in the range of 3 to 20. The SW#236 dataset includes 236 scientific workflows with 430 activities, 310 tags, and 243 sub-scientific workflows, where the number of activities contained in each scientific workflow is in the range of 2 to 30.

6.2. Evaluation Metrics. To evaluate the efficiency of scientific workflow recommendations, we adopt the precision and recall measures used in [18] and the $F_1$ score used in [16] as our evaluation metrics, which are described as equations (8)–(10), respectively:

"resource identifiers (URIs). First the ChemSpider web service is used to map the chemical name to a ChemSpider identifier, then the ChemSpider identifier is mapped to URIs via the Open PHACTS platform."

According to the textual descriptions of the $sw_{w1}$, we can use the doc2vec model to learn the sequence relationship between the subscientific workflows of CNTCI and Workflow40. Furthermore, by this way, similar structural information involved in scientific workflows can also be obtained and used for retrieval of appropriate activities and subscientific workflows, some of which can be performed with the function $cosine\_sim$ in Algorithm 4. Similarly, logical relationships involved in the components of scientific workflows can also be clearly described in the scientist’s requirement. Therefore, though these structural features are not explicitly expressed in the form of HIN, they are implicitly considered and used in our proposed approach for generating more accurate recommendations.
6.4. Comparison with Related Scientific Workflow Recommendation. As described in Section 6.2, the evaluation metrics are based on \( \text{SW}_{\text{rec}} \) and \( \text{SW}_{\text{opt}} \), which are affected by parameters \( \text{rec}_K\% \) and \( \text{exc}_K\% \) for our approach. Therefore, we study the impact of \( \text{rec}_K\% \) and \( \text{exc}_K\% \) on different recommendation methods with the SW#80 dataset.

To investigate the impact of \( \text{rec}_K\% \) on scientific workflow recommendation precision and recall, the \( \text{exc}_K\% \) is set to 10% and \( \text{rec}_K\% \) is set to 4%, 6%, 30%, respectively (step size is 2%). As shown in the Figures 5(a) and 5(b), methods HDSWR and LHWT perform higher precision and recall than LH. This is due to some functions being implemented in some scientific workflows, which does not mention in the description of scientific workflows, but in tags [27]. As a result, it is challenging for these scientific workflows to gather into the appropriate clusters. When tag information is considered, these scientific workflows are reaggregated into the appropriate cluster. This demonstrates that function semantics of tags have a great impact on scientific workflow recommendation. Besides, we also discover that HDSWR is superior to LHWT in precision and recall because the HDSWR approach applies metapaths to capture the weak semantics between scientific workflows and thus achieves high-level semantics recommendation, compared to the LHWT method.

When \( \text{rec}_K\% \) is set to be a relatively small value (e.g., 4%, 6%), we detect that the precision and recall of several methods are extremely close. This indicates that these scientific workflows particularly similar to the sample scientific workflow are recommended to scientists naturally, whatever recommendation methods they are. When \( \text{rec}_K\% \) sets to a relatively large value, the precision of several methods is reduced greatly in Figure 5(a). This is due to the fact that many unrelated scientific workflows are recommended, which do not exist in \( \text{SW}_{\text{opt}} \). Meanwhile, the recall of several methods is relatively stable in Figure 5(b), for \( \text{SW}_{\text{opt}} \) determined by the \( \text{exc}_K\% \), and \( \text{exc}_K\% \) is a fixed value. Furthermore, when the \( \text{rec}_K\% \) is 14%, the recall of HDSWR is stable. This manifests that most expected scientific workflows in \( \text{SW}_{\text{opt}} \) were identified and recommended to scientists through HDSWR. When the \( \text{rec}_K\% \) is 18%, the recall of LHWT is stable, and the recall of LH is stable until the \( \text{rec}_K\% \) is 22%.

Studying the impact of \( \text{exc}_K\% \) on scientific workflow recommendation precision and recall, the \( \text{rec}_K\% \) is set to 10%, \( \text{exc}_K\% \) is set to 4%, 6%, 30%, and 30%. In Figures 5(c) and 5(d), we discovered that the precision and recall of HDSWR are higher than LH and LHWT. Due to the above reason, when \( \text{exc}_K\% \) sets a relatively large value, the scientific workflows in \( \text{SW}_{\text{opt}} \) are abundant, while scientific workflows in \( \text{SW}_{\text{rec}} \) are fixed. Therefore, the precision of several methods is stable. However, due to the increasing discrepancy between \( \text{SW}_{\text{opt}} \) and \( \text{SW}_{\text{rec}} \), the recall of all methods has been declining.

To display the difference in scientific workflow recommendation efficiency intuitively, \( F_1 \) is applied to achieve this target. Studying the impact of \( \text{rec}_K\% \) or \( \text{exc}_K\% \) on the recommendation efficiency in Figures 6(a) and 6(b), the

---

### Table 2: Statistics of datasets.

<table>
<thead>
<tr>
<th>#Datasets</th>
<th>#Scientific workflows</th>
<th>#Activities</th>
<th>#Tags</th>
<th>#Sub-scientific workflows</th>
<th>#Activities per workflow</th>
</tr>
</thead>
<tbody>
<tr>
<td>SW#80</td>
<td>80</td>
<td>229</td>
<td>125</td>
<td>85</td>
<td>3–20</td>
</tr>
<tr>
<td>SW#236</td>
<td>236</td>
<td>430</td>
<td>310</td>
<td>243</td>
<td>2–30</td>
</tr>
</tbody>
</table>

To calculate precision and recall:

\[
\text{precision} = \frac{|\text{SW}_{\text{opt}} \cap \text{SW}_{\text{rec}}|}{|\text{SW}_{\text{rec}}|}, \quad (8)
\]

\[
\text{recall} = \frac{|\text{SW}_{\text{opt}} \cap \text{SW}_{\text{rec}}|}{|\text{SW}_{\text{opt}}|}, \quad (9)
\]

\[
F_1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}, \quad (10)
\]

In equations (8)–(10), the notation \( \text{SW}_{\text{rec}} \) represents a list of scientific workflows which are generated by recommendation algorithms, and the notation \( \text{SW}_{\text{opt}} \) represents an expected list of scientific workflows. Similar to the work in [18], we adopt a means to generate \( \text{SW}_{\text{opt}} \), by which the top \( \text{exc}_K\% \) most similar scientific workflows involved in a dataset are selected. Besides, the symbols \( |\text{SW}_{\text{rec}}| \) and \( |\text{SW}_{\text{opt}}| \) denote the numbers of scientific workflows in the \( \text{SW}_{\text{rec}} \) and \( \text{SW}_{\text{opt}} \), respectively.

6.3. Methods Used for Experiments. The scientific workflow recommendation methods used for experiments are as follows:

(i) LH [18]: this method converts a scientific workflow into a hierarchy incipiently, which manifested as the relationship between scientific workflows and subscientific workflows and activities. Thus, the similarity assessment between scientific workflows becomes the similarity assessment between the hierarchies.

(ii) LHWT [27]: this method transforms a scientific workflow into a hierarchy incipiently, as described in [18]. Considering tag information of scientific workflow enables labeling of the functional semantics of the scientific workflow in similarity computation. Hence, the tag information utilized the scientific workflow recommendation in this method.

(iii) HDSWR: it is our proposed recommendation approach. In our experiments, some parameters for HDSWR are set as follows: \( \alpha = 0.55, \beta = \gamma = 0.2 \), and \( \delta = 0.05 \).
Figure 5: Precision and recall of different recommendation methods on the dataset SW#80.

Figure 6: Efficiency comparison of different recommendation methods on the dataset SW#80.
differences between HDSWR, LHW, and LH are small in the first two groups (i.e., the value of \( \text{rec}_K \% \) is 4\% and 6\%, respectively); this indicates that scientific workflows most similar to the sample scientific workflow are recommended easily. With the increase of \( \text{rec}_K \% \) or \( \text{exc}_K \% \), the difference between several methods becomes distinct and the differences between HDSWR and other methods are obvious. Hence, this demonstrates that HDSWR can capture the similarity semantics between scientific workflows effectively and thus promote the reasonable clustering of scientific workflows. When \( \text{rec}_K \% \) exceeds 24\% and \( \text{exc}_K \% \) exceeds 22\%, the difference of several methods becomes stable, and this indicates that recommendation performance of all methods cannot play a role while excess scientific workflows are recommended.

6.5. Detailed Analysis of the Proposed Approach. In this part, we conduct a series of experiments to analyse the details of our proposed method.

6.5.1. Impact of Clustering Method. As described in Section 5.3, HDSWR requires the DPC clustering algorithm to group scientific workflows into appropriate scientific workflow clusters and assist the scientific workflow recommendation. Therefore, the impact of clustering algorithms on scientific workflow recommendation is worth studying. In our previous work [27], the SNN (Shared Nearest Neighbour) clustering algorithm [30] is used for the clustering of scientific workflows. In our study, the DPC clustering algorithm is utilized to cluster scientific workflows to the appropriate scientific workflow clusters. In Figure 7, the performance comparison of two clustering algorithms DPC and SNN on the dataset SW#80 is displayed.

The overall recommendation performance ranking is as follows: DPC > SNN, shown in Figure 7. SNN has poor performance, because it takes some data points below the density threshold and points within its domain as noise. Meanwhile, the DPC performs better recommendation performance than SNN.

6.5.2. Impact of the Size of Datasets. To study the impact of the size of datasets on the recommendation efficiency of several recommendation methods, we conduct a series of experiments with three methods on the dataset SW#236 which has a relatively larger amount of data. The experiment setting is the same with that of the dataset SW#80.

As shown in Figures 8(a) and 8(b), the HDSWR approach has better recommendation performance than other methods, both in the dataset SW#80 with a small amount of data and in the dataset SW#236 with a relatively large amount of data. This proves that the HDSWR approach has good robustness, and the recommendation performance can be effectively improved considering the attribute information of scientific workflows. Besides, we find that the distinction between the recommendation efficiency of the LHW and HDSWR approaches on the dataset SW#236 is lower than that on the dataset SW#80.

6.5.3. Comparison of the Time Efficiency. To evaluate the time efficiency of the HDSWR approach, we conduct a series of experiments with the datasets of SW#236 and SW#80. Table 3 shows the experiment results of three methods on their average running time (in seconds) with two datasets.

As shown in Table 3, the HDSWR approach has better running time performance than other methods. In fact, the operations of similarity computation occupy most of the running time of three methods, while their operations of clustering need little time. The LHW method is proposed based on the LH method, which simply appended extra label information for similarity computation. Therefore, the LHW method needs more running time than the LH method. In contrast, the similarity computation operation
adopted by the HDSWR approach is based on the HIN, which is totally different from that of other methods. Therefore, it effectively reduces the running time of handling various information for similarity computation.

7. Conclusion

In this paper, we aim to provide automatic support for the reuse and modelling of scientific workflows. Specifically, we utilize heterogeneous information network as a means of organizing and representing the relations between scientific workflows and consider the objects of tag, description, activity, and subscientific workflow for scientific workflow recommendation. We propose a novel scientific workflow similarity computation method based on metapath. In addition, we present a scientific workflow recommendation approach named HDSWR, where the density peak clustering algorithm is adopted for grouping scientific workflows into clusters and a list of scientific workflows is ranked and recommended according to the requirements of scientists and engineering personnel. As future work, we tend to consider how to apply machine learning methods to automatically tune some parameters on the [32–35] HDSWR and yield better performance. Furthermore, we will handle related privacy problems in view of the newest research studies [36–41].

Data Availability

The data sets of our experiments are publicly accessible via the following website: https://github.com/yixinxunwu/myExperiment.

Table 3: Experimental results on the comparison of average running time (in seconds).

<table>
<thead>
<tr>
<th>Methods</th>
<th>SW#80</th>
<th>SW#236</th>
</tr>
</thead>
<tbody>
<tr>
<td>LH</td>
<td>806.68</td>
<td>42436.2</td>
</tr>
<tr>
<td>LHWT</td>
<td>852.7</td>
<td>42923.61</td>
</tr>
<tr>
<td>HDSWR</td>
<td>34.82</td>
<td>292.72</td>
</tr>
</tbody>
</table>

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This research in this paper was supported by the National Key Research and Development Project of China (Nos. 2018YFB1702600 and 2018YFB1702602), National Natural Science Foundation of China (Nos. 61772193, 61402167, 61872139, and 61876062), Hunan Provincial Natural Science Foundation of China (Nos. 2017J4036 and 2018J2139), and Research Foundation of Hunan Provincial Education Department of China (Nos. 17K033 and 19A174).

References


