Hindawi
Complexity
Volume 2020, Article ID 4129063, 16 pages
https://doi.org/10.1155/2020/4129063

WILEY | Hindawi

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 Zhou**[1,2]

[1]*School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan 411201, China*
[2]*Key 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

Guest Editor: Yuan Yuan

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 engineering personnel. 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 engineering personnel

(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 (BPMN) 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 lay 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 metagraph to incorporate more complex semantics for HIN-based recommendation. Shi et al. [24] developed a metapath-based random walk strategy and proposed a HIN embedding-based recommendation algorithm. On 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 $\phi: V \longrightarrow B$ and a link-type mapping function $\psi: E \longrightarrow R$, 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 *Chemical2URIs* (https://www.myexperiment.org/workflows/97.html) (denoted as sw_1) and *DFCUAM* (https://www.myexperiment.org/workflows/4700.html) (denoted sw_2).

The sw_1 links with a text description ($ds\,c_1$), three tags (*annotation*, *chemspider* and *cheminformatics*), two activities (*REST_Service* and *Xpath_Service*), and two subscientific workflows (*CNTCI* and *workflow40*).

The sw_2 links with a text description ($ds\,c_2$), three tags (*cheminformatics*, *chemspider*, and *metabolomics*), and two activities (*SearchByMass* and *GetCompoundDetails*).

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 $\phi: V \longrightarrow B$ and the link-type mapping function $\psi: E \longrightarrow R$, 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 $B_1 \longrightarrow B_2 \longrightarrow \cdots \longrightarrow B_{l+1}$ and thus defines a composite relationship $R = R_1 \circ R_2 \circ \cdots \circ R_l$ between two object types B_1 and B_{l+1}, where $\circ$ denotes the composition operator on relations R.

Table 1: Notations and explanations.

Notation	Explanation
SW	A list of scientific workflows
dsc	A description type of object
D	A subscientific workflow type of object
A	An activity type of object
T	A tag type of object
p_1, p_2, p_3, p_4	Different types of metapaths
SWT, SWA, SWD	Adjacent matrices on the objects of tag, activity, and subscientific workflow, respectively
v_i^T, v_i^A, v_i^D	Feature vectors of scientific workflow sw_i in the adjacent matrices SWT, SWA, and SWD, respectively
$C_{i,j}^{p1,T}$, $C_{i,j}^{p2,A}$, $C_{i,j}^{p3,D}$	The similarity strength of scientific workflows sw_i and sw_j on metapaths p_1, p_2, and p_3, respectively
α, β, γ, δ	Weight coefficients

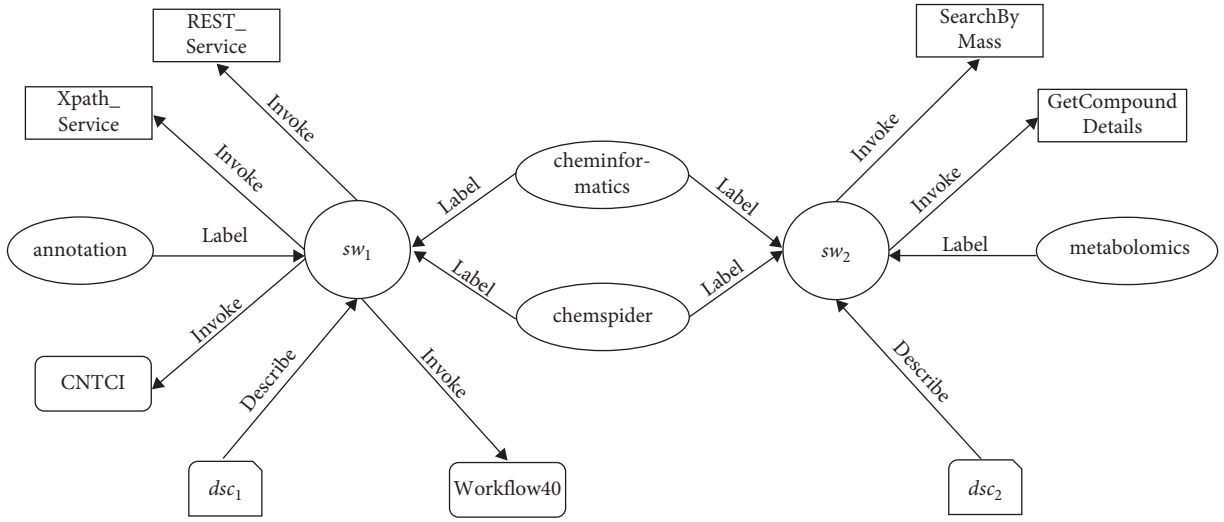


Figure 1: An example of HIN-based scientific workflow representation.

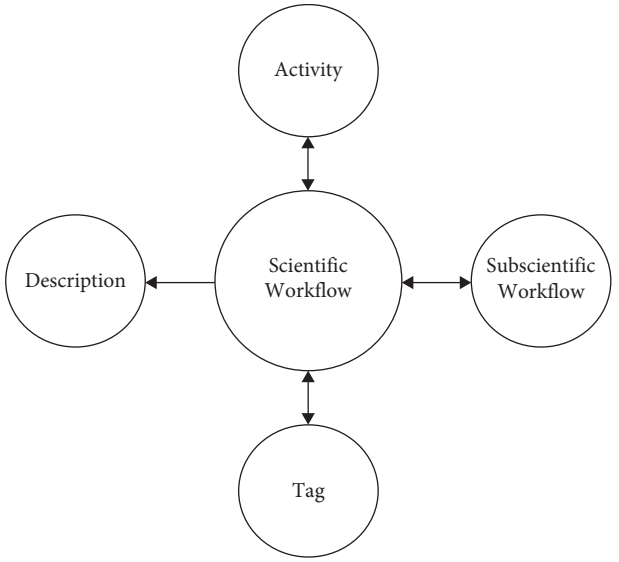


Figure 2: HIN-based scientific workflow representation schema.

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:

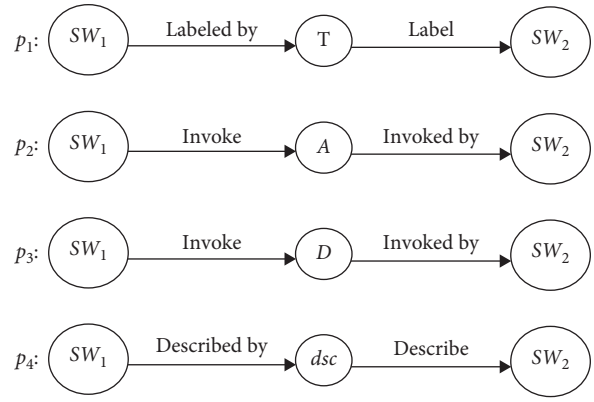


Figure 3: Four types of metapaths under the HIN-based scientific workflow representation schema.

(1) Metapath p_1: if a tag is shared by two scientific workflows sw_1 and sw_2, we can use the metapath $SWTSW$ (Scientific Workflow ⟶ Tag ⟶ 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

SWASW (Scientific Workflow ⟶ Activity ⟶ Scientific Workflow) to denote a coactivity relation of 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 *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_4: if a description is shared by two scientific workflows sw_1 and sw_2, we can use the meta-path *SWdscSW* (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 sub-scientific 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_i^T, 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 matric *SWT*. Likewise, we use the feature vector v_i^A and v_i^D to represent the relations between the scientific workflow sw_i and the objects of activity and sub-scientific 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:

$$C_{i,j}^{p1,T} = v_i^T \cdot \left(v_j^T\right)^t. \tag{1}$$

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

$$C_{i,j}^{p2,A} = v_i^A \cdot \left(v_j^A\right)^t, \tag{2}$$

$$C_{i,j}^{p3,D} = v_i^D \cdot \left(v_j^D\right)^t. \tag{3}$$

Based on equation (1), we can also obtain the values of $C_{i,i}^{p1,T}$ and $C_{j,j}^{p2,A}$. To normalize the similarity strength effectively, we utilize the ratio between the $C_{i,j}^{p1,T}$ and the max one in the $C_{i,i}^{p1,T}$ and $C_{j,j}^{p1,T}$ to represent the similarity between scientific workflows sw_i and sw_j with respect to metapath p_1, which is described as follows:

$$sim_{p1}(i,j) = \frac{C_{i,j}^{p1,T}}{\max\left(C_{i,i}^{p1,T}, C_{j,j}^{p1,T}\right)}. \tag{4}$$

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

$$\begin{aligned} sim_{p2}(i,j) &= \frac{C_{i,j}^{p2,A}}{\max\left(C_{i,i}^{p2,A}, C_{j,j}^{p2,A}\right)}, \\ sim_{p3}(i,j) &= \frac{C_{i,j}^{p3,D}}{\max\left(C_{i,i}^{p3,D}, C_{j,j}^{p3,D}\right)}. \end{aligned} \tag{5}$$

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 $\mathbb{v}_{swi}$ and $\mathbb{v}_{swj}$ for the descriptions of scientific workflows sw_i and sw_j, respectively. Besides, the normalized cosine similarity between $\mathbb{v}_{swi}$ and $\mathbb{v}_{swj}$ is calculated as the similarity value on the descriptions of scientific workflows sw_i and sw_j, which is described as:

$$sim_{p4}(i,j) = \frac{\left(\mathbb{v}_{swi} \cdot \mathbb{v}_{swj}\right)/\left(\left\|\mathbb{v}_{swi}\right\| \cdot \left\|\mathbb{v}_{swj}\right\|\right) + 1}{2}. \tag{6}$$

In equation (6), the notations $\mathbb{v}_{swi}$ and $\mathbb{v}_{swj}$ represent the norm of the paragraph vectors $\mathbb{v}_{swi}$ and $\mathbb{v}_{swj}$, 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:

$$sim(i,j) = \alpha \times sim_{p4}(i,j) + \beta \times sim_{p1}(i,j) + \gamma \times sim_{p2}(i,j) + \delta \times sim_{p3}(i,j). \quad (7)$$

In equation (7), α, β, γ, and δ are the weight coefficients satisfying $\alpha + \beta + \gamma + \delta = 1$.

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 *SW*, which may come from some scientific workflows repository. All the scientific workflows in the list *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 *SW* into multiple different clusters, where the similarity values in the matrix are used as the distances between scientific workflows and *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., *requirement.dscs*, we search and choose appropriate objects of activity and subscientific workflow involved in the list *SW*, where D_{smp} and A_{smp} denote a set of subscientific workflows and a set of activities, respectively. Then, a HIN-based sample scientific workflow sw_{smp} can be constructed (Line 4).

Step 4 (line 5): according to the sample scientific workflow sw_{smp}, we firstly select an appropriate group of scientific workflows in the set *Clusters* by the similarity values between sw_{smp} and different clusters. Then, a list SW_{rec} is generated for recommendation, where the number of scientific workflows in the list SW_{rec} is related to the parameter of *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 *ComputeSimilarity* is described as Algorithm 2.

In Algorithm 2, three adjacent matrices on the scientific workflow list *SW* are constructed first (lines 1–3). Then, the feature vector of scientific workflows sw_i and sw_j is used to compute the similarity strengths on metapaths by equations (1)–(3) (lines 6-7), based on which the similarity between sw_i and sw_j with respect to metapaths can be obtained by equations (4)–(6) (line 8). Finally, the similarity values are obtained by equation (7) (line 9) and stored in the matrix *Matrix* for further clustering and recommendation (line 10).

Example 2. The scientific workflows sw_1 and sw_2 in Figure 1 can be used as an example. As illustrated by Figure 1, there are four tags (*annotation*, *chemspider*, *cheminformatics*, and *metabolomics*) involved in the scientific workflows sw_1 and sw_2. Therefore, as shown in Figure 4(a), the corresponding value on these four tags in the adjacent matrix *SWT* is 1 or 0 with respect to the sw_1 and sw_2, where the value of 0 denotes that such tag does not belong to some scientific workflow. Similarly, the matrix *SWA* in Figure 4(b) shows the corresponding values on the activities of the scientific workflows sw_1 and sw_2, and the matrix *SWD* in Figure 4(c) shows the corresponding values on the subscientific workflows. Besides, the feature vectors of v_i^T, v_i^A, and v_i^D 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], by which the scientific workflows are grouped and divided into different clusters for further recommendation. Different from the work in [10, 30], 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 *DPCClustering* can be described as Algorithm 3.

In Algorithm 3, we first initiate the matrix dist according to the matrix *Matrix* (line 1) and initiate the value of cutoff distance *dc* according to the rule of thumb introduced in [10] (line 2). Then, we calculate the local density values of scientific workflows (lines 3–10) and their relative distances values (lines 11–18). Finally, we can apply the DPC algorithm to divide scientific workflows into different clusters (line 19), where each cluster in the *Clusters* can be denoted as a group of scientific workflows with a scientific workflow as its cluster center.

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 *GetActivity_SubWF*, which is described as Algorithm 4.

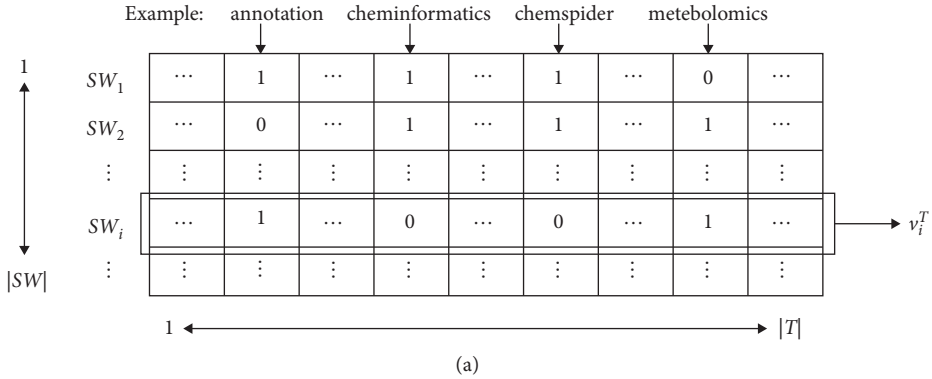
In Algorithm 4, because the descriptions *requirement.dscs* provided in the requirement are related to activities or subscientific workflows, the best matching result on each description in *requirement.dscs* may be an activity or a subscientific workflow. Therefore, we calculate the similarity values on activities and subscientific workflows,

Input:
(i) SW: a list of scientific workflows.
(ii) $requirement$: a modelling requirement, denoted as (dsc_{smp}, T_{smp}, $dscs$).
(iii) α, β, γ, δ: parameters for similarity computation.
(iv) rec_K: a parameter on the number of recommended scientific workflows.
Output:
(i) SW_{rec}: a list of recommend scientific workflows.
(1) $Matrix \longleftarrow ComputeSimilarity\ (SW, \alpha, \beta, \gamma, \delta)$
(2) $Clusters \longleftarrow DPCClustering\ (Matrix, SW)$
(3) $D_{smp}, A_{smp} \longleftarrow GetActivity\_SubWF\ (requirement.dscs)$
(4) $sw_{smp} \longleftarrow$ Construct a sample scientific workflow with D_{smp}, A_{smp}, $requirement.\ dsc_{smp}$ and $requirement.\ T_{smp}$
(5) $SW_{rec} \longleftarrow RecommendSWs\ (sw_{smp}, Clusters, rec\_K, \alpha, \beta, \gamma, \delta)$
(6) **return** SW_{rec}

Algorithm 1: HDSWR (SW, $requirement$, rec_K, α, β, γ, and δ).

Input:
(i) SW: a list of scientific workflows.
(ii) $\alpha, \beta, \gamma, \delta$: weight coefficients.
Output:
(i) $Matrix$: the final similarity matrix of SW.
(1) $SWT \longleftarrow$ construct the adjacency matrix of SW on tag objects
(2) $SWA \longleftarrow$ construct the adjacency matrix of SW on activity objects
(3) $SWD \longleftarrow$ construct the adjacency matrix of SW on sub-scientific workflow objects
(4) **for** each scientific workflow sw_i in SW **do**
(5) **for** each scientific workflow sw_j in SW **do**
(6) obtain v_i^A, v_i^T, v_i^D, v_j^A, v_j^T, v_j^D *from* SWT, SWA, SWD
(7) calculate $C_{i,j}^{p1,T}$, $C_{i,j}^{p2,A}$, $C_{i,j}^{p3,D}$
(8) calculate $sim_{p1}(i, j)$, $sim_{p2}(i, j)$, $sim_{p3}(i, j)$, $sim_{p4}(i, j)$
(9) calculate $sim(i, j)$
(10) $Matrix_{i,j} \longleftarrow sim(i, j)$
(11) **end for**
(12) **end for**
(13) **return** $Matrix$

Algorithm 2: Function $ComputeSimilarity$ (SW, α, β, γ, δ).



(a)

Figure 4: Continued.

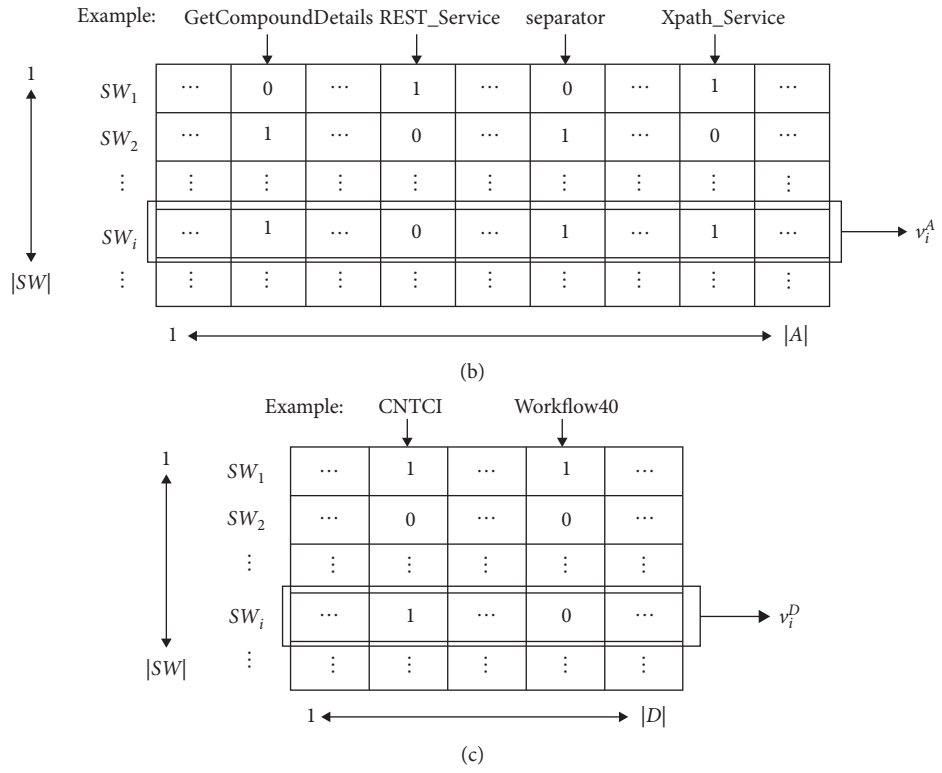


Figure 4: An example of three adjacent matrices and feature vectors. (a) *SWT*. (b) *SWA*. (c) *SWD*.

Input:
(i) *Matrix*: a similarity matric of scientific workflows.
(ii) *SW*: a list of scientific workflows.
Output:
(i) *Clusters*: the set of generated scientific workflow clusters.
(1) dist $\longleftarrow 1 - Matrix$
(2) $dc \longleftarrow$ 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) **for** each scientific workflow sw_i in *SW* **do**
(4) $ld_i \longleftarrow 0$
(5) **for** each scientific workflow sw_j in *SW* **do**
(6) **if** $\text{dist}_{i,j} < dc$ **then**
(7) $ld_i \longleftarrow ld_i + 1$
(8) **end if**
(9) **end for**
(10) **end for**
(11) **for** each scientific workflow sw_i **do**
(12) $rd_i \longleftarrow \max(\text{dist})$
(13) **for** each scientific workflow sw_j **do**
(14) **if** $ld_i < ld_j$ and $rd_i < \text{dist}_{i,j}$ **then**
(15) $rd_i \longleftarrow \text{dist}_{i,j}$
(16) **end if**
(17) **end for**
(18) **end for**
(19) *Clusters* $\longleftarrow$ clustering scientific workflows by the DPC algorithm with the local density values such as ld_i and relative distances values such as rd_i
(20) **return** *Clusters*

Algorithm 3: Function *DPCClustering* (*Matrix*, *SW*).

Input:
(i) *requirement.dscs*: a list of descriptions on activities and subscientific workflows.
(ii) *SW*: a list of scientific workflows.
Output:
D_{smp}: a set of subscientific workflows
A_{smp}: a set of activities
(1) $D_{smp} \longleftarrow \emptyset, A_{smp} \longleftarrow \emptyset$
(2) **for** each *dsc* in *dscs* **do**
(3) $sim_{tmp1} \longleftarrow 0, sim_{tmp2} \longleftarrow 0$
(4) **for** each *sw* in *SW* **do**
(5) **for** each activity *a* in *sw* **do**
(6) $sim \longleftarrow cosine\_sim$ (doc2vec (*dsc*), doc2vec (*a*))
(7) **if** $sim > sim_{tmp1}$ **then**
(8) $sim_{tmp1} \longleftarrow sim$
(9) $a_{tmp} \longleftarrow a$
(10) **end if**
(11) **end for**
(12) **for** each sub-scientific workflow *d* in *sw* **do**
(13) $sim \longleftarrow cosine\_sim$ (doc2vec (*dsc*), doc2vec (*d*))
(14) **if** $sim > sim_{tmp2}$ **then**
(15) $sim_{tmp2} \longleftarrow sim$
(16) $d_{tmp} \longleftarrow d$
(17) **end if**
(18) **end for**
(19) **end for**
(20) **if** $sim_{tmp1} > sim_{tmp2}$ **then**
(21) append a_{tmp} to A_{smp}
(22) **else**
(23) append d_{tmp} to D_{smp}
(24) **end if**
(25) **end for**
(26) **return** A_{smp}, D_{smp}

Algorithm 4: Function *GetActivity_SubWF* (*requirement.dscs*, *SW*).

respectively, where the working procedure of the function *cosine_sim* in lines 6 and 13 is similar to that of equation (6).

Besides, for each description in *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_{smp} on the objects of activity, subscientific workflow, and tag.

Step 2 (lines 3–14): we compute the similarity between the sample scientific workflow sw_{smp} 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 $cluster_{smp}$ if the similarity value between its cluster center and sw_{smp} is the largest among all the clusters (lines 9–13).

Step 3 (line 15): after the cluster $cluster_{smp}$ is determined, the *rec_K%* scientific workflows of the $cluster_{smp}$ which are most related to sw_{smp} 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 *CNTCI*, which is short for *Chemical_Name_To_Chemspider_ID*, and a subscientific workflow named *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_{smp}: 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.
(1) $v^A_{smp}, v^T_{smp}, v^D_{smp} \longleftarrow$ construct the feature vector on the activities, tags and sub-scientific workflows of sw_{smp}.
(2) $sim_{smp} \longleftarrow 0$ and $cluster_{smp} \longleftarrow \varnothing$
(3) **for** each $cluster \in Clusters$ **do**
(4) $sw_{ct} \longleftarrow$ choose the cluster center scientific workflow of the *cluster*
(5) $v^A_{ct}, v^T_{ct}, v^D_{ct} \longleftarrow$ construct the feature vector on the activities, tags and sub-scientific workflows of sw_{ct}.
(6) calculate $C^{p1,T}_{smp,ct}$, $C^{p2,A}_{smp,ct}$, $C^{p3,D}_{smp,ct}$
(7) calculate $sim_{p1}(smp, ct)$, $sim_{p2}(smp, ct)$, $sim_{p3}(smp, ct)$, $sim_{p4}(smp, ct)$
(8) calculate $sim(smp, ct)$
(9) $sim_{tmp} \longleftarrow sim(smp, ct)$
(10) **if** $sim_{smp} < sim_{tmp}$ **then**
(11) $sim_{smp} \longleftarrow sim_{tmp}$
(12) $cluster_{smp} \longleftarrow cluster$
(13) **end if**
(14) **end for**
(15) $SW_{rec} \longleftarrow$ choose the top *rec_K*% most similar scientific workflows in $cluster_{smp}$
(16) **return** SW_{rec}

Algorithm 5: Function *RecommendSWs* (sw_{smp}, *Clusters*, *rec_K*, α, β, γ, δ).

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_1, 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. 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 subscientific 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:

Table 2: Statistics of datasets.

#Datasets	#Scientific workflows	#Activities	#Tags	#Sub-scientific workflows	#Activities per workflow
SW#80	80	229	125	85	3–20
SW#236	236	430	310	243	2–30

$$\text{precision} = \frac{\left(\left|SW_{\text{ept}} \cap SW_{\text{rec}}\right|\right)}{\left|SW_{\text{rec}}\right|}, \tag{8}$$

$$\text{recall} = \frac{\left(\left|SW_{\text{ept}} \cap SW_{\text{rec}}\right|\right)}{\left|SW_{\text{ept}}\right|}, \tag{9}$$

$$F_1 = \frac{2 \times \text{precision} \times \text{recall}}{(\text{precision} + \text{recall})}. \tag{10}$$

In equations (8)–(10), the notation SW_{rec} represents a list of scientific workflows which are generated by recommendation algorithms, and the notation SW_{ept} represents an expected list of scientific workflows. Similar to the work in [18], we adopt a means to generate SW_{ept}, by which the top *exc_K%* most similar scientific workflows involved in a dataset are selected. Besides, the symbols $|SW_{\text{rec}}|$ and $|SW_{\text{ept}}|$ denote the numbers of scientific workflows in the SW_{rec} and SW_{ept}, 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 $\delta = 0.05$.

6.4. Comparison with Related Scientific Workflow Recommendation. As described in Section 6.2, the evaluation metrics are based on SW_{rec} and SW_{ept}, which are affected by parameters *rec_K%* and *exc_K%* for our approach. Therefore, we study the impact of *rec_K%* and *exc_K%* on different recommendation methods with the *SW#80* dataset.

To investigate the impact of *rec_K%* on scientific workflow recommendation precision and recall, the *exc_K%* is set to 10% and *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 *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 *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 SW_{ept}. Meanwhile, the recall of several methods is relatively stable in Figure 5(b), for SW_{ept} determined by the *exc_K%*, and *exc_K%* is a fixed value. Furthermore, when the *rec_K%* is 14%, the recall of HDSWR is stable. This manifests that most expected scientific workflows in SW_{ept} were identified and recommended to scientists through HDSWR. When the *rec_K%* is 18%, the recall of LHWT is stable, and the recall of LH is stable until the *rec_K%* is 22%.

Studying the impact of *exc_K%* on scientific workflow recommendation precision and recall, the *rec_K%* is set to 10%, *exc_K%* is set to 4%, 6%, ..., 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 *exc_K%* sets a relatively large value, the scientific workflows in SW_{ept} are abundant, while scientific workflows in SW_{rec} are fixed. Therefore, the precision of several methods is stable. However, due to the increasing discrepancy between SW_{ept} and SW_{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 *rec_K%* or *exc_K%* on the recommendation efficiency in Figures 6(a) and 6(b), the

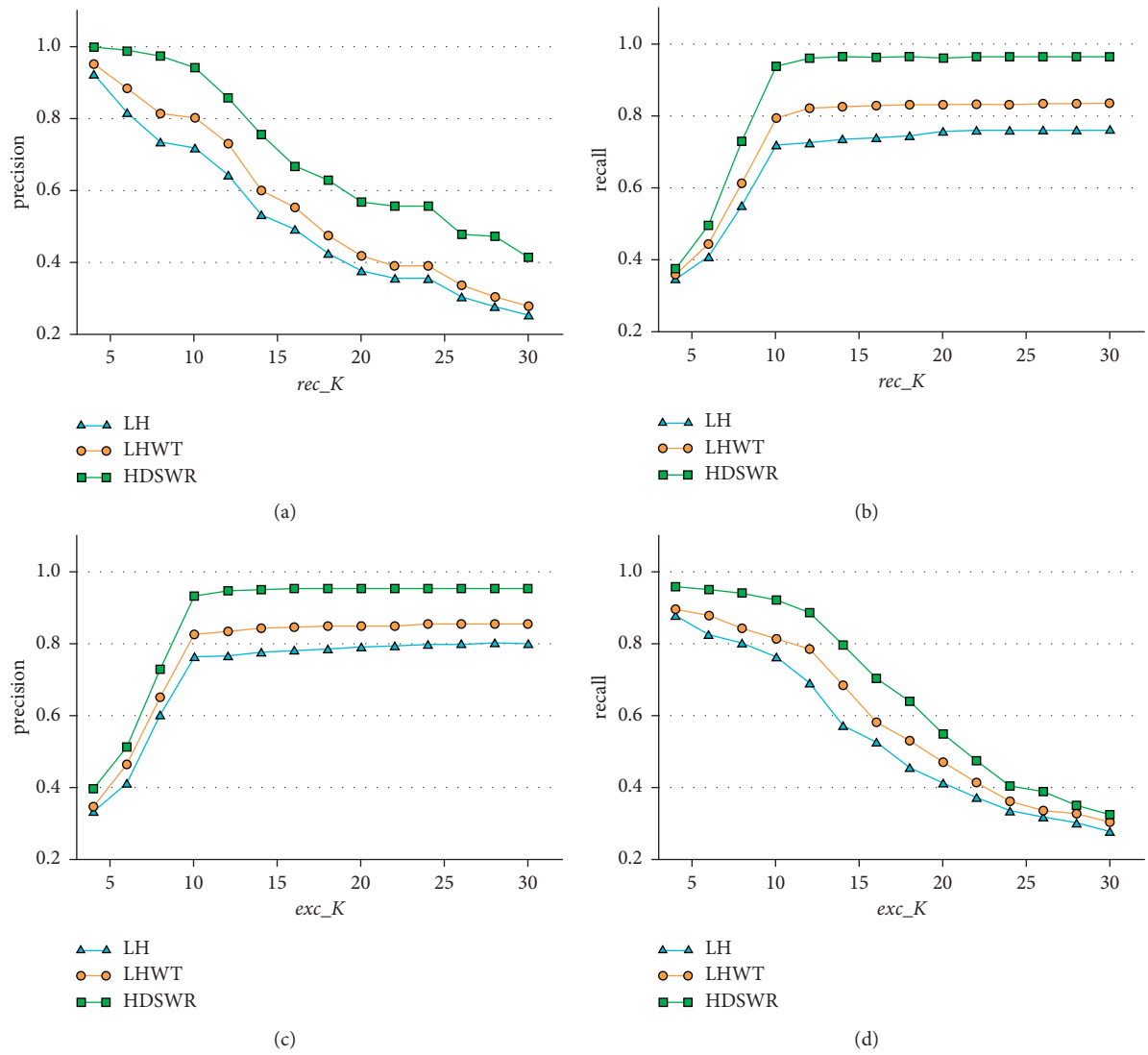


Figure 5: Precision and recall of different recommendation methods on the dataset *SW#80*.

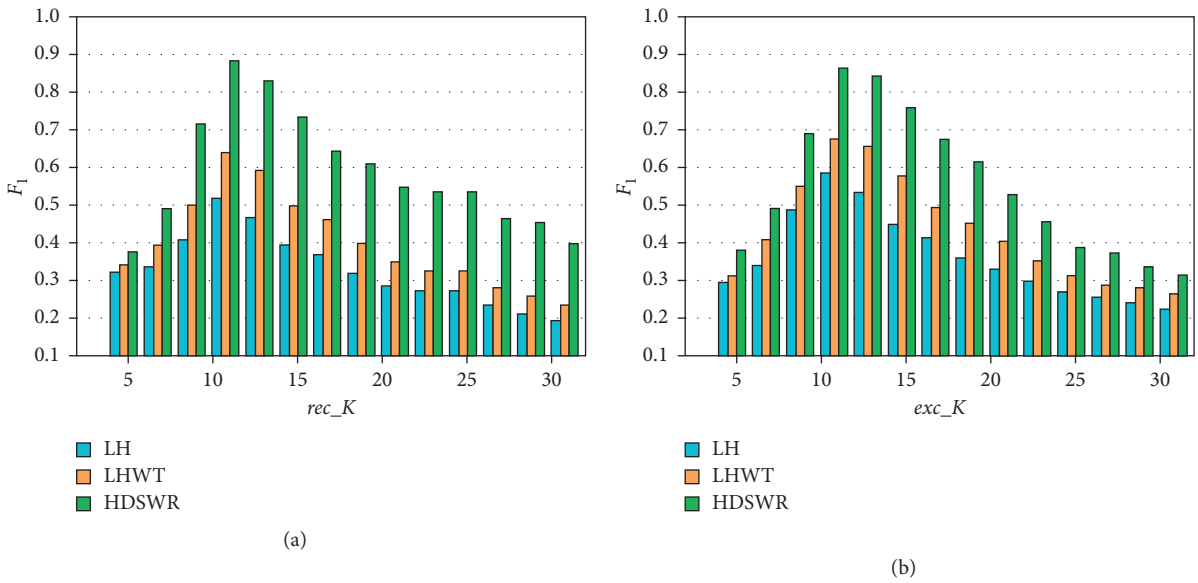


Figure 6: Efficiency comparison of different recommendation methods on the dataset *SW#80*.

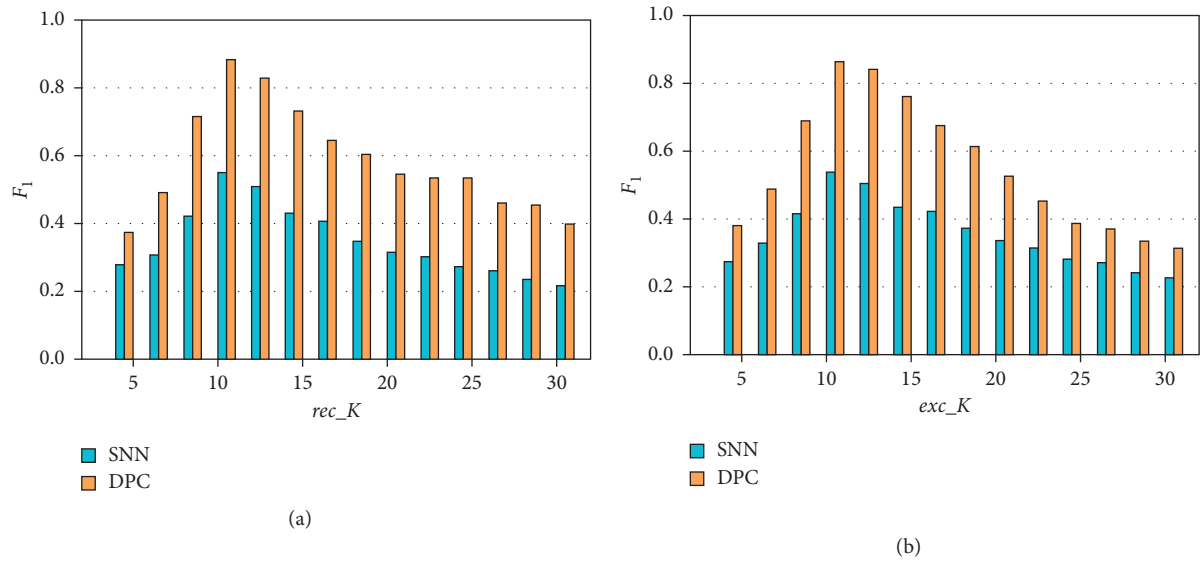


Figure 7: Efficiency comparison of different clustering methods on the dataset *SW#80*.

differences between HDSWR, LHWT, and LH are small in the first two groups (i.e., the value of *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 *rec_K%* or *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 *rec_K%* exceeds 24% and *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 LHWT 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 LHWT method is proposed based on the LH method, which simply appended extra label information for similarity computation. Therefore, the LHWT method needs more running time than the LH method. In contrast, the similarity computation operation

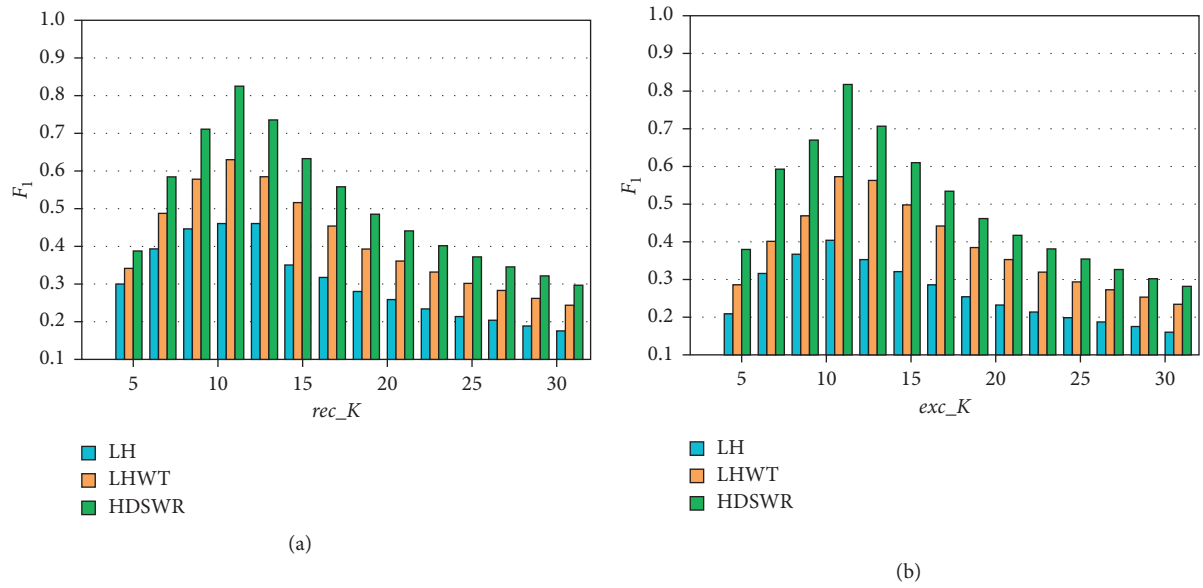


Figure 8: Efficiency comparison of different recommendation methods on the dataset *SW#236*.

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

Methods	SW#80	SW#236
LH	**806.68**	**42436.2**
LHWT	**852.7**	**42923.61**
HDSWR	**34.82**	**292.72**

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.

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. 2017JJ4036 and 2018JJ2139), and Research Foundation of Hunan Provincial Education Department of China (Nos. 17K033 and 19A174).

References

[1] W. Song, F. Chen, H.-A. Jacobsen, X. Xia, C. Ye, and X. Ma, "Scientific workflow mining in clouds," *IEEE Transactions on Parallel and Distributed Systems*, vol. 28, no. 10, pp. 2979–2992, 2017.

[2] W. Song and H.-A. Jacobsen, "Static and dynamic process change," *IEEE Transactions on Services Computing*, vol. 11, no. 1, pp. 215–231, 2016.

[3] X. Song, W. Dou, and J. Chen, "A workflow framework for intelligent service composition," *Future Generation Computer Systems*, vol. 27, no. 5, pp. 627–636, 2011.

[4] X. Xu, X. Zhang, H. Gao, Y. Xue, L. Qi, and W. Dou, "Become: blockchain-enabled computation offloading for IOT in mobile edge computing," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 6, pp. 4187–4195, 2019.

[5] X. Xu, C. He, Z. Xu, L. Qi, S. Wan, and M. Z. A. Bhuiyan, "Joint optimization of offloading utility and privacy for edge computing enabled IoT," *IEEE Internet of Things Journal*, 2019.

[6] D. De Roure, C. Goble, and R. Stevens, "The design and realisation of the virtual research environment for social sharing of workflows," *Future Generation Computer Systems*, vol. 25, no. 5, pp. 561–567, 2009.

[7] J. Starlinger, S. Cohen-Boulakia, S. Khanna, S. B. Davidson, and U. Leser, "Effective and efficient similarity search in scientific workflow repositories," *Future Generation Computer Systems*, vol. 56, pp. 584–594, 2016.

[8] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu, "Pathsim: meta path-based top-k similarity search in heterogeneous information networks," *Proceedings of the VLDB Endowment*, vol. 4, no. 11, pp. 992–1003, 2011.

[9] C. Shi, Y. Li, J. Zhang, Y. Sun, and S. Y. Philip, "A survey of heterogeneous information network analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 1, pp. 17–37, 2016.

[10] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, 2014.

[11] M. Weske, "Business process management architectures," in *Business Process Management*, pp. 333–371, Springer, Berlin, Germany, 2012.

[12] L. Wen, J. Wang, W. M. P. van der Aalst, B. Huang, and J. Sun, "Mining process models with prime invisible tasks," *Data & Knowledge Engineering*, vol. 69, no. 10, pp. 999–1021, 2010.

[13] S. Deng, D. Wang, Y. Li et al., "A recommendation system to facilitate business process modeling," *IEEE Transactions on Cybernetics*, vol. 47, no. 6, pp. 1380–1394, 2016.

[14] J. Zhang, Q. Liu, and K. Xu, "FlowRecommender: a workflow recommendation technique for process provenance," in *Proceedings of the Eighth Australasian Data Mining Conference-Volume 101*, Australian Computer Society, Inc., Melbourne, Australia, pp. 55–61, December 2009.

[15] Y. Li, B. Cao, L. Xu et al., "An efficient recommendation method for improving business process modeling," *IEEE Transactions on Industrial Informatics*, vol. 10, no. 1, pp. 502–513, 2013.

[16] H. Wang, L. Wen, L. Lin, and J. Wang, "RLRecommender: a representation-learning-based recommendation method for business process modeling," in *Proceedings of the International Conference on Service-Oriented Computing*, pp. 478–486, Springer, Hangzhou, China, November 2018.

[17] J. Zhang, M. Pourreza, S. Lee, R. Nemani, and T. J. Lee, "Unit of work supporting generative scientific workflow recommendation," in *Proceedings of the International Conference on Service-Oriented Computing*, pp. 446–462, Springer, Hangzhou, China, November 2018.

[18] Z. Cheng, Z. Zhou, P. C. Hung, K. Ning, and L.-J. Zhang, "Layer-hierarchical scientific workflow recommendation," in *Proceedings of the 2016 IEEE International Conference on Web Services (ICWS)*, pp. 694–699, IEEE, San Francisco, CA, USA, June 2016.

[19] Z. Zhou, Z. Cheng, and Y. Zhu, "Similarity assessment for scientific workflow clustering and recommendation," *Science China Information Sciences*, vol. 59, no. 11, Article ID 113101, 2016.

[20] M. Krzywucki and S. Polak, "Workflow similarity analysis," *Computing and Informatics*, vol. 30, no. 4, pp. 773–791, 2012.

[21] R. Bergmann and Y. Gil, "Similarity assessment and efficient retrieval of semantic workflows," *Information Systems*, vol. 40, pp. 115–127, 2014.

[22] A. Mohan, M. Ebrahimi, and S. Lu, "A folksonomy-based social recommendation system for scientific workflow reuse," in *Proceedings of the 2015 IEEE International Conference on Services Computing*, pp. 704–711, IEEE, New York City, NY, USA, June 2015.

[23] H. Zhao, Q. Yao, J. Li, Y. Song, and D. L. Lee, "Meta-graph based recommendation fusion over heterogeneous information networks," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 635–644, ACM, 2017.

[24] C. Shi, B. Hu, W. X. Zhao, and S. Y. Philip, "Heterogeneous information network embedding for recommendation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 2, pp. 357–370, 2018.

[25] L. Chen, Y. Wang, Q. Yu, Z. Zheng, and J. Wu, "WT-LDA: user tagging augmented LDA for web service clustering," in *Proceedings of the International Conference on Service-Oriented Computing*, Springer, Kauai, HI, USA, pp. 162–176, 2013.

[26] L. Qi, X. Xu, W. Dou, J. Yu, Z. Zhou, and X. Zhang, "Time-aware IoE service recommendation on sparse data," *Mobile Information Systems*, vol. 2016, Article ID 4397061, 12 pages, 2016.

[27] J. Hou and Y. Wen, "Utilizing tags for scientific workflow recommendation," in *Proceedings of the International Conference on Applications and Techniques in Cyber Security and Intelligence*, Springer, Huainan, China, pp. 951–958, 2019.

[28] Y. Sun, Y. Yu, and J. Han, "Ranking-based clustering of heterogeneous information networks with star network schema," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 797–806, ACM, Paris, France, June 2009.

[29] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proceedings of the International Conference on Machine Learning*, pp. 1188–1196, Beijing, China, June 2014.

[30] L. Ertöz, M. Steinbach, and V. Kumar, "Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data," in *Proceedings of the 2003 SIAM International Conference on Data Mining*, pp. 47–58, San Francisco, CA, USA, May 2003.

[31] C. A. Goble, J. Bhagat, S. Aleksejevs et al., "My experiment: a repository and social network for the sharing of bioinformatics workflows," *Nucleic Acids Research*, vol. 38, no. suppl_2, pp. W677–W682, 2010.

[32] P. Wang, J. Huang, Z. Cui, L. Xie, and J. Chen, "A Gaussian error correction multi-objective positioning model with NSGA-II," *Concurrency and Computation: Practice and Experience*, vol. 32, no. 5, 2019.

[33] X. Cai, Y. Niu, S. Geng et al., "An under-sampled software defect prediction method based on hybrid multi-objective cuckoo search," *Concurrency and Computation: Practice and Experience*, vol. 32, no. 5, 2019.

[34] M. Hassan, M. Rehmani, and J. Chen, "Differential privacy techniques for cyber physical systems: a survey," *IEEE Communications Surveys and Tutorials*, vol. 22, no. 1, pp. 746–789, 2019.

[35] M. Hassan, M. Rehmani, and J. Chen, "DEAL: differentially private auction for blockchain based microgrids energy trading," *IEEE Transactions on Services Computing*, 2019, In press.

[36] L. Qi, R. Wang, C. Hu, S. Li, Q. He, and X. Xu, "Time-aware distributed service recommendation with privacy-preservation," *Information Sciences*, vol. 480, pp. 354–364, 2019.

[37] L. Wen, W. M. P. van der Aalst, J. Wang, and J. Sun, "Mining process models with non-free-choice Constructs," *Data Mining and Knowledge Discovery*, vol. 15, no. 2, pp. 145–180, 2007.

[38] L. Wen, J. Wang, W. M. P. van der Aalst, B. Huang, and J. Sun, "A novel approach for process mining based on Event Types," *Journal of Intelligent Information Systems*, vol. 32, no. 2, pp. 163–190, 2009.

[39] S. Deng, L. Huang, and G. Xu, "Social network-based service recommendation with trust enhancement," *Expert Systems with Applications*, vol. 41, no. 18, pp. 8075–8084, 2014.

[40] S. Deng, L. Huang, G. Xu, X. Wu, and Z. Wu, "On deep learning for trust-aware recommendations in social networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 5, pp. 1164–1177, 2016.

[41] L. Qi, X. Zhang, W. Dou, C. Hu, C. Yang, and J. Chen, "A two-stage locality-sensitive hashing based approach for privacy-preserving mobile service recommendation in cross-platform edge environment," *Future Generation Computer Systems*, vol. 88, pp. 636–643, 2018.