Koohi-Var and Zahedi *J Big Data (2018) 5:26*
https://doi.org/10.1186/s40537-018-0135-6

Journal of Big Data

METHODOLOGY Open Access

Cross-domain similarity assessment for workflow improvement to handle Big Data challenge in workflow management

Tahereh Koohi-Var[1,2*] and Morteza Zahedi[1,2]

*Correspondence:
tahere.koohi@gmail.com
[2] CE Department, Shahrood University of Technology, Shahrood, Iran
Full list of author information is available at the end of the article

Abstract

With the increasing of using workflow management systems workflow improvement becomes a new emerging problem. Many issues must be considered to handle all aspects of the workflow improvement. Workflows might become quite complex, especially when we move to Web3 (ubiquitous computing web). Workflows from different domains (e.g., scientific or business) have similarities and, more important, differences between themselves. Some concepts and solutions developed in one domain may be readily applicable to the other. In ubiquitous computing, multi-domain workflow data analysis might cause Big Data challenge. This paper investigates the problem of workflow improvement having an observed behavior (i.e., event logs). It proposes a cross-domain concept extraction by similarity assessment to solve some aspects of workflow improvement problem, and it has a new research effort at the intersection of workflow domains. Besides, the proposed technique is evaluated with the benefit of using Deep learning and Transfer learning. One of the greatest assets to use these both learning methods is analyzing a massive amount of data. Our results show that our proposed technique is effectively applicable for analyzing real-life huge data in workflow improvement.

Keywords: Workflow management systems, Ubiquitous computing, Big Data, Deep learning, Transfer learning

Introduction

Ubiquitous computing is a concept in engineering where the computing is made to appear anytime and everywhere. This paradigm is also described as pervasive computing, or ambient intelligence. Considering ubiquitous computing perspectives can be useful to manage workflows, because running workflows usually needs large-scale computing resources and massive storages [1] and considering the relatedness between data. Ubicomp is a concept in engineering where the computing is made to appear anytime and everywhere. This paradigm is also described as pervasive computing, or ambient intelligence. Figure 1 represents Big Data challenges in workflow management systems (WfMSs). By creating workflows, skilled and non-skilled designers can manage and control their tasks. It is necessary to gather all the data and to consider various aspects of information in the design of a clear and helpful workflow, unless the workflow might cusses destructive effects on experiment or activities. Therefore, the WfMSs

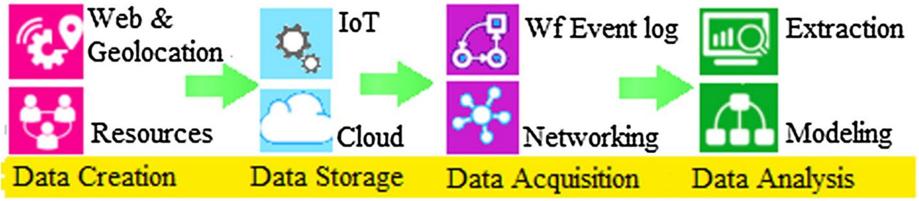

Fig. 1 Big Data challenge aspects in WfMSs

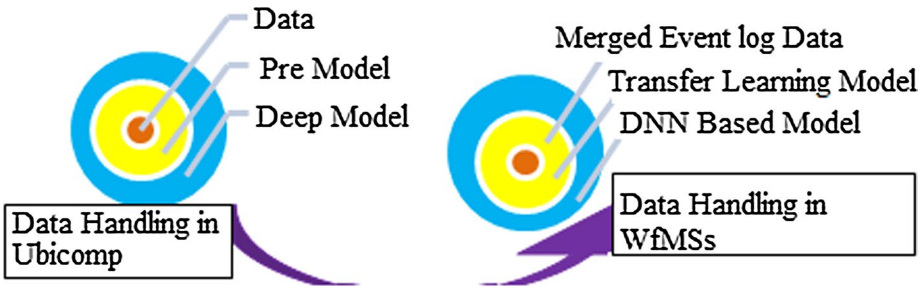

Fig. 2 Our motivation

development is an on demand effort. On the other hand, in real world data, challenges come not only from the creation and designing phase of workflows, but also from the analysis, storage management, and acquisition phases. In ubiquitous computing, we can also see such a variety of data. Developers and scientists in the last decades introduce new technologies and approaches to deal with Big Data challenges in ubiquitous computing.

Two critical technologies for growing of the ubiquitous computing infrastructure are: Cloud Computing and the Internet of Things (IoT) [2]. With the introduction of Cloud, one can rent some storage from somewhere that offers the hardware. Cloud Computing is also proposed as a distributed system to decrease the processing cost. IoT often specifies the user interaction with data, and how a system can handle data (e.g. Cloud or local storage). IoT might be at a same time both solution and problem in the Big Data management. It can decrease the Big Data storing cost because it allows users to utilize accessible things as a data storage device. On the other hand, it might produce new real time data from various devices, and the new various data in a united management system might become a new processing problem.

Motivation

Nowadays, Big Data handling in Ubicomp is our requirement when we propose a technique for workflow improvement. This requirement prevents the existence of having a holistic view to workflows that day to day they are created in context aware domains. Hence, for having a consistent view to the workflow improvement, we proposed a method based on Deep artificial Neural Networks (DNNs) [3] and Transfer learning for WfMSs data challenge management (Fig. 2). Recently, DNN has won numerous contests in pattern recognition and machine learning [3]. Deep learning in a DNN based method can be used to make discriminating tasks of Big Data analytics easier [4]. DNN is similar neural networks but it has:

1. More layers
2. Separate computations for each layer
3. Learning from unlabeled data

Workflows that we apply our method on them are sequences of process components that reusing them presents several advantages: allowing for principled attribution of established methods, improving quality through evolutionary workflow development, and making processes more efficient and deal with workflow complexity (e.g. by removing useless links in workflows). They can be categorized in two groups of scientific workflows and business workflows and executed in different implementation styles depended on the process context, e.g. e-commerce, Bioinformatics, etc. In this paper, we use some known process contexts to our method be context aware. While scientific workflows describe the setup of scientific experiments, by enabling scientists to focus on domain-specific aspects of their work (e.g. in astronomy, biology, etc.), and not dealing with complex data management, business process models describe organizational processes focusing on the sequences of activities, roles, and events. Workflows can be the automated parts of business processes. An important aim of WfMSs is to save machine cycles by optimizing workflow execution on available resources, and deal with its complexity. There are many approaches that deal with workflow complexity and improve it [5]. However, there is not a united and multi-disciplinary approach to deal with the complexity of workflows. Most research works prefers to have contributions focusing either on scientific workflows or business workflows (for example, see [6] and [7]). Hence, as a new contribution, this paper addresses a solution that has a workflow context alignment from the relatedness aspect of workflows. It works through an analysis of the common concepts in workflow developments, both in scientific workflows and business workflows, pursuing the following aims:

1. To find simple workflow abstractions (workflow composed by one or more sub-workflows [8]) that would ease understandably and therefore effective reuse.
2. To Prefer potential information for suggestpng workflow abstractions for creating simple improved workflows.
3. To deal with Big Data challenges.

To form an abstraction of workflows this paper uses workflow motifs [9] to enable improving workflows and relate workflows from one domain to other workflows by different domains. Workflow motifs are common structures or labels of workflow conceptual steps. Workflow motifs add a layer of abstraction that generalizes the functionality of each step or set of steps, helping users to understand the main functionality of the workflow [8]. Finding the layer of abstraction of workflows provides a deep insight that helps our DNN based technique to make a model to analysis workflow data. But, first we need to extract important data of workflows to have a holistic view of workflows. Based on our previous research in [9] we realize that cross-domain concept extraction might obtain important data of workflows that come from different domains. This paper targets at finding an effective approach to the cross-domain concept extraction problem. The approach is based on dense representations using spectral feature alignment (SFA) [9] and DNNs. The idea is to represent the entire workflow collection by a motif by-workflow matrix U whose rows correspond to workflows and columns to motifs [9]. However, the use of feature vectors implicates two limitations. First, as vectors always represent a predefined set of features, all vectors in a given application have to keep the same length despite the size or complexity of workflows. Second, there is no direct possibility to describe relationships often exist among different parts of a workflow. These two drawbacks are severe, particularly when patterns under consideration are characterized by complex structural relationships and not the statistical distribution of a fixed set of pattern features [10]. An alternative structural approach to represent each workflow can be based on graphs as basic specifications for workflow structures.

The remainder of this paper is organized as follows: a review of the relevant works is conducted in "Related works" section. "Problem setting" section introduces the problem setting. The new contributions reported in this paper are an extension of the related works which is described in "Similarity assessment" section. The section describes the similarity assessment part of the method. "DNN based workflow improvement" section introduces the DNN based workflow improvement part. "Method overview" section proposes the method overview. Results and experiment settings are mentioned in "Results and discussion" section. Finally, "Conclusion" section concludes the paper.

Related works

Workflow improvement and dealing with workflow complexity relate to different streams of research. Related work can be grouped under three major topics: workflow common feature extraction, workflow similarity assessment, and workflow improvement. This section first reviews the works about extraction of common features. Then, since our approach to workflow improvement is highly related to similarity assessment it reviews this topic. Finally a review about workflow improvement is done.

Workflow common feature extraction

Fiannaca et al. [11] found that the user cannot do the best design and it is necessary to provide automated systems tailored to the user's preferred workflow. They suggested not to produce a workflow exclusively by the automated system and away from the user. The main idea of their approach is that the steps of decision making of the system must be represented to user. Hence, the user can intervene in the generation of its desired workflow.

Ludascher et al. in [12] compared features of scientific workflows and business workflows. Finally, they concluded that the integration of workflow analysis methods based on data flow and on control-flow can yield new results and insights for both scientific workflows and business workflows. However, because of the most differences between scientific workflows and business workflows, authors prefer to work either on scientific workflows or business workflows. For example, Garijo et al. [13] presented an empirical analysis performed over 260 scientific workflow descriptions. They defined a catalog of domain independent (*DI*) and domain specific (*DS*) conceptual abstractions for workflow steps called scientific workflow motifs. Sample workflow motifs that they manually identified in their catalog were contained data preparation, data cleaning, data moving, data retrieval and workflow overloading. They defined the motif similar the term "pattern" refers to the established best practices solving recurring problems, instead of trying to prescribe a best practice. Finally, they discussed the distribution of the abstractions across different workflow systems. Then they emphasized that different workflow systems share a common core of workflow abstractions. They compared the maximum, minimum and average number of steps within workflows per domain and preferred a scientific workflow motif catalog for abstracting workflows. They showed data preparation motifs are the most common type of motifs and then they intuited that most of the motifs will be found in other domains and in other workflow systems. Finally, they suggested their identification of workflow oriented motifs to be acting as a set of heuristics for automatically creating abstractions over workflows [13]. The technique proposed in [13] aimed at understand groupings of scientific workflow steps that form a meaningful high-level data manipulation operation. However, the technique was not generalized to cover business workflows, too. Besides, the authors did not suggest an automated way for extraction of workflow motifs.

Alper et al. in [14] presented a solution for automated creating workflow description summaries. In fact, they proposed a method to workflow summarization, by analyzing scientific workflows. They used a rule-based approach that acts on patterns of semantic annotations on workflow graphs. The proposed approach in [14] rewrites the workflow with the well defined primitives, namely Composition, Collapse and Elimination. Although the effectiveness of their proposed automated summarization has shown better results than user summarization, but it suffers from time-consuming and expensiveness of workflow motif labeling.

More interestingly, other types of scientific workflow motif discovery can be targeted. For instance, using state transition information between workflow motifs, frequent patterns can be mined from the repository to provide both functional subunits to be used in workflow design, and best-practice patterns to guide workflow designers [15]. In [15] workflows achieve visually compressed workflows by replacing recurring motifs with macros. Replacing recurring motifs with macros can provide hierarchical concept abstraction, visual compression, improved readability and cost-effective task performance [15].

In another relevant work the approach presented in [8] aimed at the automatic abstraction creation over workflows. It defined two metrics of Minimum Description Length (MDL) and Size to find the best matching motifs based on grammatically finding ones. MDL gets the best structure as the one that minimizes the description length of the entire dataset [8]. The graph size metric aimed at finding a structure that best reduces the overall collection graph size. Garijo et al. in [8] proposed a classification algorithm based on manually labeled textual data. Labeling might be time-consuming and expensive. Moreover, users often use some different words when they express the sentiment in different domains.

Workflow similarity assessment

Similarity assessment is an important part of our approach to workflow improvement. To design a repaired and well-ordered similar workflow based on a complex or improper workflow model the paper assesses workflows similarity. Similarity in the present paper is a measure indicating equivalence of workflow.

Workflow similarity assessment can be structure based or text based [5, 6]. Each of them has its weakness and strengths: text based approaches are independent of the workflows' formats and can be used to compare workflows both across different systems, and across multiple repositories [6]. Yet, they need pre-processing step, e.g. for stop-word removal. Approaches for structural workflow comparison, can be applied without backing human-provided textual knowledge. Yet, they have to assess workflow functionality from the information in the graph structure and the composed modules. For example, Starlinger et al. in [16] presented Layer Decomposition (LD) approach that structurally compares workflows. The basic idea behind LD is to focus on the order in which modules are executed in both workflows by only permitting mappings of modules to be used for similarity assessment which respect this order.

In recent years, many business processes and scientific workflow matching algorithms and similarity measures have been developed [6, 7]. For example, Starlinger et al. [6] proposed domain agnostic similarity measurement methods. They provided a comparative study of workflow similarity measures for a set of scientific workflows from Taverna [17] and Galaxy [18]. Although they collected a subset of basic techniques that they are used for business process model similarity measurement, but they focused on scientific workflows. As another example, the technique proposed in [19] mapped motifs for workflow comparison and categorizing based on Copula theory [20]. It has mapped workflow motif features in a probability density function (PDF) to have a finite set based similarity assessment. For a cross-domain based similarity assessment, the proposed method in [9] measured the similarity of workflows from different domains by a feature clustering using a Sparse coding (SC) based clustering method that is the basis of the present work.

An overview of some related works reflecting the diversity of approaches taken so far can be found in Table 1.

Table 1 shows existing approaches to similarity assessment of workflows. It also shows goals associated with similarity assessment used in the papers. Table 2 shows the abbreviations used in the overview presented in Table 1.

Dijkman et al. in [21] studied similarity measures of label matching similarity, and Graph Edit Distance (GED). The label matching metric is based on pair-wise comparisons of component labels, and the GED metric takes into account both the node labels and the topology of the process models [21].

There are two main similarity assessment granularity levels. Some research is interested in the relation of singular templates (single-components) [6, 14], while other research is interested in the relation of the common sub-workflows (multi-component) [8, 15, 19] for similarity analysis. The work proposed here is relevant to multi-component similarity level analysis. Our similarity assessment is based on MM and structural. Motifs are already known as predefined pattern structures that they can be found by the machine. Motifs may also include the unseen structures that the

Table 1 Existing approaches to similarity assessment in workflow domains

Ref.	Text based	Structure based			Goals	Data scope	Level
		Label motifs	Topological motifs	Topology			
[35]	BW	No	No	No	SR	S	F1
[23]	Frequent tag sets	No	No	Frequent MS	Data exploration	S	F1
[15]	No	No	Common motifs, ML	No	Wf visual compression	S	F2
[19]	No	No	Common motifs, MM	No	Organizing and grouping	S	F2
[8]	No	Semantic annotations	MDL, size	No	SR	S	F2
[21]	No	No	No	Label matching, GED	SR	B	F1, F2
[6]	BT, BW	No	No	MS, GED, PS	Facilitating reuse	S	F1, F2
[14]	No	Semantic annotations	No	No	Summarization	S	F1
[22]	No	No	No	PS	Improve Wf design	S, B	F2
[9]	No	No	MM	No	SR	S, B	F2

F1 single-component, *F2* multi-component, *S* scientific, *B* business

Table 2 Algorithm shorthand notation overview

Notation	Description
Wf	Workflow
MS	Module sets
PS	Path sets
GED	Graph Edit Distance
BW	Bag of words
BT	Bag of tags
ML	Motif label
MM	Motif mapping
SR	Search and retieval
MDL	Minimum Description Length [8]

machine memorizes while learning by the increase in repetition, and the machine adds them to its earliest known structures.

A similarity assessment method that used feature selection techniques based on text is [22]. It treated workflows as a group of words (BWs). In [22] for each workflow, a pre-processing component counted the number of occurrences of each term. The method presented in [22] by using latent semantic analysis (LSA) considered the shared occurrence of terms. It then produced similarity values between pairs of workflows. The most common feature selection step in text based approaches is the pre-processing step (the removal of stop-words and stemming (returning the word to its stem or root e.g. flies→ fly) step) [24]. Another work also explored the most frequently used modules (MS) and frequent tag sets [23].

Workflow improvement

Mendling et al. in [25] presented a guideline to create a good process model. Guideline suggestions are as follows:

1. Model as structured as possible
2. Decompose a model with more than 50 elements
3. Use as few elements in the model as possible
4. Use verb-object, activity labels
5. Minimize the routing paths per element
6. Use one start and one end event
7. Avoid OR routing elements

Based on these suggestions improved alternative models can be constructed. This guideline relates many efforts such as repair [26, 27], improve [22], and simplify [19] workflow designs. Based on the guideline a technique proposed in [28] for the automatic removal of infrequent behavior from process execution logs is an improvement to process model creation. In [28] the dependencies are detected and removed from an automaton built from event log, and then the original log is updated accordingly, by removing individual events using alignment-based replay of processes [28].

Another approach to workflow improvement is proposed in [22]. The approach mines reusable tasks and identifies task sequences (i.e., paths) that frequently occur, to improve workflow design. By mining these concepts from the business domain and the scientific experiment domain, Tosta et al. in [22] adapted the data mining problem to the process mining. Tosta et al. concluded that their [21] concluded that their proposed approach is general in the sense that it can be coupled to any scientific workflow management systems. Although it has presented a general solution on workflows from Bioinformatics and e-commerce domains, but similar the technique of [6] it has focused on scientific workflows.

This work develops a general solution to workflow improvement by extracting concept when having no labels in a domain of interest (hereafter, called the target domain) but having some labeled data in a different domain, regarded as the source (or auxiliary) domain. Our work is similar to the process model repair which has discovered a control-flow model [29] based on event data and conformance checking that takes a predefined model as the norm. It is similar to the case of the proposed technique in [26]: having a process model does not conform to reality, one can suggest a model such that the observed behavior can be fully explained by the model. Fahland and Aalst in [26] repaired workflows by minimizing misalignments such that the parts of the model that were not invalidated by the event log were kept as was.

Problem setting

Previous section was about our proposed method overview. This section, like our previous work at [9] gives a problem setting for further analysis. Before giving a formal definition of the problem we first present some definitions.

Definition 1 *(Domain)* A domain D denotes a class of entities in a space or a semantic concept.

For example, different types of workflows, such as e-commerce, hospital, and Bioinformatics, can be regarded as different domains. As another example, computer science, mathematics and physics can be also regarded as different domains.

Definition 2 *(Type)* Given a specific domain D, type data shows the type of workflows correspond to a sequence of tags $t_1 t_2 ... t_n$ where t_i is a conceptual abstraction of workflow step tag from a Lexicon T. In this work, type data is appended to specify type of tags in a workflow for a given domain D.

Based on the definitions described above, now the problem is defined.

Problem Definition Having a set of labeled data (e.g. tag sets) from a source domain, to train a model for a target domain, we leverage some unlabeled data from the target domain to help improving a given workflow. In detail, first for the dataset we create a workflow graph from workflow descriptions, and we apply a function mapping on each trace of the workflow event log. We discover workflow motifs from all the stored workflows, and then we improve a given workflow *G*.

In the next step we use a SFA algorithm to find a new representation in cross-domain process data, such that the gap between domains can be reduced [9]. SFA uses some *DI* conceptual abstractions for workflow sequences (i.e. workflow motifs) as a bridge to construct a bipartite graph to model the co-occurrence relationship between the *DS* tags and *DI* ones. The idea is that if two *DS* tags have connections to more common *DI* ones in the graph, they tend to be aligned together with higher probability. Similarly, if two *DI* tags have connections to more common *DS* tags in the graph, they tend to be aligned together with higher probability. We aim to improve a given workflow graph *G* by learning instance models after embedding tag sets based on DNN. For this purpose, our proposed method first searches for the best match of *G* with the other stored workflow graphs in the dataset. Then it decides about the best possible order of workflow by embedding workflow motifs based on DNN. Workflow Motifs, which occur frequently in workflow domains, can be reused as fragments of workflows. They can be extracted after analyzing different workflow domains.

To achieve the aims of our approach follows the following steps:

1. Learn higher-level features from source and target adaptation,
2. Use the learned higher-level features to represent workflow motifs,
3. Provide a training model from the new representations of workflows with corresponding labels based on cross-domain concept extraction,
4. Search among learned workflow models,
5. Reconcile the complex test workflow based on detecting workflow model.

The method is divided into two sections of workflow similarity assessment and DNN based workflow improvement. Following the method is explained in detail.

Similarity assessment

Process Management Systems (PMSs) are tools focusing on the management of process execution quality. They offer tools for having a good visualization of processes by workflow management means. An empirical study has shown that PMSs are central objects [30] in the many conceptual modeling (e.g., to support documentation, improvement and automated enactment processes). The emerging workflow management means have made publishing, and finding workflows more easily [13], but users still face the challenges of understanding and reusing available workflows. For example, while a data visualization can combine logical and structural features of a system to lead to a good understanding of the system structure, it itself may be an obstacle for workflow understandably [5]. Especially if the system is complex and relatively unfamiliar, a visualization that can show a good structure might help. Besides, providing a huge amount of structured and unstructured data in workflow repositories can be considered as a great source of decision making in understanding system structures. A major difficulty in understanding workflows is their complex nature. A workflow may contain several significant analysis steps in a special network describing a process that consists of repetitive patterns of sub-processes. The structure of this network has control and data dependencies. Such a network may combine with the other activities. Due to the huge number of tasks and their configuration parameters, these activities may become heavily error prone and have complex nature.

Sometimes functional results of various workflow designs are similar. In other words, some workflows are equal such that there is a workflow model with greater simplicity and simplest workflow is more understandable and efficient. In the simplified and improved workflow can be fewer loops, fewer nodes, or fewer links [25]. For example, in tk the simplifying of workflow visualization makes a better understanding of complex and relatively unfamiliar systems.

If the workflow management tools reuse the knowledge of the other workflows to let the user create and manage complex workflows, this difficulty in understanding stands also in the way of reusing workflows. After specifying same behavior of different workflow models, workflow management means can use an equivalent simpler workflow than a complex one. In this way goals associated with similarity measurements in workflows are raised. The similarity measurement also helps to find some useful fragments of workflows to analysis and design workflows with reusing fragments. Recently, workflow recommenders (WRs) [22] have proposed recommendation services that aim of suggesting frequent combinations of workflow tasks for reuse. These recommenders apply data mining techniques to help users find items to improve their workflow designs by prediction. We similar Tosta et al. in [22] do not present recommendations to indicate whether an especial fragment of a workflow is better or worse, but only if this especial fragment is more common or not.

Common features of workflows are important properties of workflows. They are effective in their potentially reusability [31]. We use similarity assessment of workflows to automate workflow improvement. Hence, features are extracted based on LSA in a transformation step as Transfer learning. Transfer learning [32] may automate learning of features across workflows, from different domains and of very different natures. It is a process mining method used to deal with Big Data in transferring workflow information like our previous work at [9]. Same workflows that have been already improved before can be grouped into one group in a feature database (Fig. 3). Finally, a similarity assessment with features extracted from the workflow and from the feature database is done to catch the important useful fragments. Based on the orders extracted by a DNN based method the workflow is improved. The source data of our work are workflow event logs which represent traces (sequence of events) of workflows.

The workflow improvement process can use previous knowledge of previous workflows, i.e. after reconciling and improving a workflow, it can be reserved to reuse as a proposed sample solution for the same workflow, later. Important properties of workflows are effective in their potentially reusability [31]. We use similarity assessment of workflows to automate workflow improvement. To do this, features are extracted based on LSA in a transformation step as Transfer learning. Transfer learning [32] is used to automate learning of features across workflows, from different domains and of very different natures. It is used to deal with Big Data in transferring workflow information like our previous work at [9]. By Transfer learning it is possible to have a classification task in one domain of interest, but only having sufficient training data in another domain of interest, where the latter data may be in a different feature space or follow a different data distribution [32].

Same workflows that have been already improved before can be grouped into one group in a feature database (Fig. 4). Finally, a similarity assessment with features extracted from the workflow to be improved and from the feature database is done to reuse the best detected workflow. Based on the orders extracted by a DNN based method the workflow is improved. The Source data in our presented approach are workflow event logs which represent traces (sequence of events) of workflows.

By using Transfer learning our concern is to learn motifs, even only a few labeled data are given. It supposes that there are some higher-level features can help training the

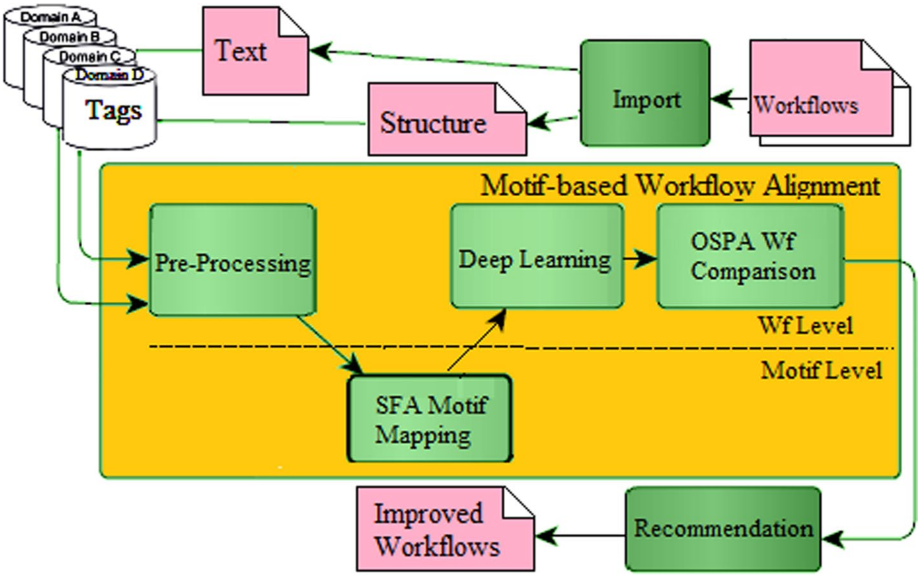

Fig. 3 The proposed method

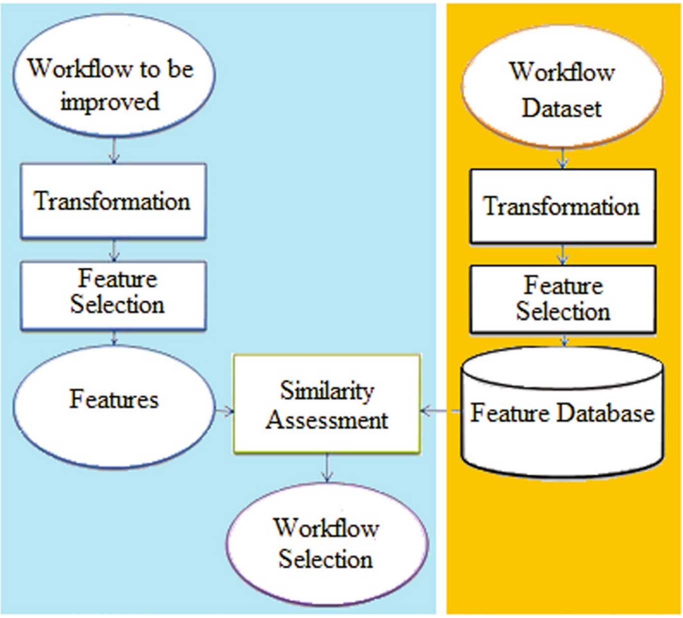

Fig. 4 Similarity assessment overview

model of the classifier. It aims to *align DS* (non-pivot) features from different domains by workflow domain adaptation.

Two solutions for higher-level feature construction are *Sparse coding* (*SC*), and *Deep learning* [33]. Our paper benefits from both of these two methods. The *SC* is an unsupervised *feature construction* method and Deep learning is learning directly from raw inputs. *SC* learns basis functions that capture high-level features in the data given only unlabeled input data [34].

Tags assigned to workflow steps in a repository can be used for similarity assessment, as done in [35]. The tags assigned to steps in a workflow are treated as a BT and the workflow similarity can be calculated in the same way that the approach described in [35] used BW. Our presented approach like [35] has used a BT. The difference is that we align tags without the need of to be specifically pre-selected by the workflow designer, and we select tags of workflow steps with the most frequency as workflow motifs among different domains. We align tags automatically by spectral clustering and feature generation tasks with a special setting of Transfer learning [32], which aims at transferring knowledge across workflow domain motifs. Another advantage of our approach over the proposed technique in [35] is that no stop-word removal or other pre-processing of the tags is performed in this paper. Our approach Deep learning may extract nonlinear complicated features in Big Data analytic. The next section is about it.

DNN based workflow improvement

This work is on both finding an appropriate co-occurrence matrix of workflow motifs and a DNN embedding method. By DNN, feature vectors are constructed based on the structure of workflows and processing the tags in the context. Based on DNN the tags that are repeated more closely together have similar vectors. Using discovery of workflow motifs by co-occurrence frequency of tags within workflows, prediction of an improved workflow becomes more convenient. For example, a reasonable order of workflow motifs can be derived based on extracted concepts. As such the work unites small data into enough big data to benefit from DNN.

Given a labeled or an unlabeled workflow by type, the workflow is tagged into unique motifs by DNN method. Tagging unique motifs by DNN helps to improve workflows by determining workflow motif relations among workflows. After searching the most similar workflow in the dataset with the given complex workflow, DNN can give an insight about the relation of the given workflow and the detected workflow. For finding similarity of workflows this work uses a set based distance function called OSPA [19]. OSPA is a systematic alternative of the squared error [19].

In remainder of this section, a cost function that helps picking alignments based on DNN is introduced.

Cost function

The best alignment can be obtained by adjusting a cost function with a weight for component places in the workflow and with a frequency. The idea is to find a cost function N in which a component tag that occurs very rarely has approximately high costs (and hence is avoided), and a component tag that occurs very frequently has low costs (and hence is preferred with $(1-\alpha)\%$):

$$N = (1-\alpha)\mathit{freq} + \mathit{weight} \tag{1}$$

where *freq* is the frequency of workflow motifs, and *weight* is the weight assigned for component levels.

By computing the cost function N for each neuron in DNN a multi-component improvement is applied on workflow.

Method overview

Our contribution shows how to adapt knowledge acquired from a repository of a dataset that come from different domains. It gives an algorithm learning common features, namely concepts, as domain independent conceptual abstractions for workflow steps. Then after some steps it predicts the most conceptual workflow that matches to the given complex workflow. Hence, our method embeds sample workflow graphs in vector spaces to benefit from both the universality of graphs for workflow representation and the convenience of embedded features for pattern recognition. It learns representational embedding for motifs from a large collection of unlabeled data using a generative model. It views the model as a problem of cross-domain concept extraction of different workflows. As a result, it can suggest the best order of workflow motifs in a preferred order based on embedded feature sets learned by DNN. For this it applies an Optimal Sub-Pattern Assignment (OSPA) Distance based similarity measure to identify the 'best' connecting paths between the workflow motifs. Fig. 3 shows an overview of our proposed technique. As a summarization, steps of the multi-domain workflow similarity assessment and workflow improvement are as follows:

Algorithm 1 Workflow Improvement

```
1:  Input: $(L, data)$, $L$ is a log and or a workflow specification over $T$, and data is a function mapping each
    $tracete \in L$ to a set of pairs $data(te) = (t1, x1), \ldots, (tq, xq)$ such that $xi \in type(ti)$ for each
    $i \in 1, \ldots, q, andt1, \ldots, tq = t \in T|task(t) = t[j]$.
2:  Load dataset for source and target domains
3:  for each source and target data do
4:      Remove outlier data with a given threshold
5:      Assign type labels to workflow steps
6:      Separate data in two sections of $DS$ and $DI$ data
7:      Pivot feature selection: selection of the meaningful features that have relation to other features
8:      Specific feature selection that does not have relations with other features
9:      Compute co-occurrence matrix that shows relations between different data in the dataset
10:     Compute the bipartite graph that has relations between different motifs in the dataset
11:     Find $k$ largest Eigen Values (Construction of $U$ matrix)
12:     Embed DNN based unique features of workflow motifs to $U$ matrix
13: Search the best match of the $k$ largest Eigen Values of a given workflow with others
14: Suggest the best workflow motif order for a given workflow
15: Output: Improved workflow
```

Algorithm 1 describes the workflow improvement algorithm.

Our method has to be able capture descriptions of complex workflow models. So, it uses an ISA-Tab handler for scientific workflows that allows the linkage of a single sample to multiple analyses employing various assays [22]. For other workflows in other domains a XML-based workflow Event Logging Mechanism has been used in the XES format [36].

Algorithm 1 uses SFA. SFA can fully exploit the relationship between *DI* and *DS* tags via co-*aligning* them on the bipartite graph to learn a more compact and meaningful representation of space. So our method uses the automatically created lexicons to expand feature sets in a model learned (*U* matrices) at train time by introducing related workflow motifs. For this, different domains of workflows are co-aligned. Finally a unique DNN based feature set is embedded in the training matrices of workflow models for further analysis. Operations of the algorithm are applied for a given complex workflow that should be reconciled.

The most similar *U* matrix and the related DNN based embedded feature set to the test complex workflow is detected using lexicon. Lines 4–11 of the algorithm describe the discovery step of workflow motifs.

In the second step the sequential motif information of workflows is considered with a DNN based feature embedding. Based on this the method decides about the order of workflow motifs related to different domains and the best conceptual order of workflow motifs is suggested. For this purpose, at first the work constructs a BT by giving sources and targets. Then we compute the belonging measures based on the Eqs. (2), (3).

$$S\_percent = \frac{A_{S_i} \times 100}{DS\_size} \times \frac{A_{S_i}}{S\_size}, \quad \{\mathrm{i} = 0, \ldots, \mathrm{n}\} \tag{2}$$

where A_{S_i} is the type label count of a source domain, *S_size* is the count of Source tags, *DS_size* is the count of *DS* Source tags, and *S_percent* shows the percentage of belonging measure. *S_percent* shows the percentage of belonging of common motifs to the source.

$$T\_percent = \frac{A_{T_i} \times 100}{DS\_size} \times \frac{A_{T_i}}{T\_size}, \quad \{\mathrm{i} = 0, \ldots, \mathrm{n}\} \tag{3}$$

where A_{T_i} is the type label count of a target domain, T_size is the count of Target tags, DS_size is the count of DS Target tags, and $T\_percent$ shows the percentage of belonging measure. $T\_percent$ shows the percentage of belonging of common tags to the target.

A_{S_i} and A_{T_i} are obtained by $c(t_i, x_j)$ denotes the frequency of tags t_i in relation to x_j which x_j specifies the type.

Such categorization helps to identify the best match of a given complex test workflow. For example, if the given workflow type is Bioinformatics, the method decides based on the results of the Eqs. (2), (3), and selects the group of motifs that have the most matching to Bioinformatics.

Results and discussion

This section performs extensive experiments on some workflow datasets from different domains, and shows that a universal function to improve workflows can be applied based on Transfer learning and distributional semantics. We focus on cross-domain concept extraction to operate on every workflow domain, e.g. scientific and business workflows.

The experiment setup is presented in the remainder of this section. Then the individual analysis for cross-domain concept extraction is done. Finally, we show the results of the workflow improvement.

Workflow datasets

Experiment datasets for the train are from different domains containing of hospital, repair, financial and scientific workflows. In our experimental work, four datasets are used to train model and a dataset is used to test as follows:

The real-life event logs of an academic hospital originally intended for use in the first Business Process Intelligence (BPI) Challenge [37] is used in the present paper. It contains of 1143 workflow traces. Another dataset is the financial dataset that contains of 4366 traces used for *BPI Challenge* 2012 [38].

For scientific workflows, we extracted information from a resource holding over 70,665 experimental design workflows (ArrayExpress) [39]. We used a subset of this collection, comprising of 29 scientific workflows.

Finally, a simulated dataset with 4123 traces of a telephone repair process has been used for train [40].

The test dataset contains of an event log pertains to a loan application process of a Dutch financial institute used for BPI Challenge 2017 [41]. The data and the process under consideration is the same as [38]. However, the system supporting the process has changed in the meantime. In particular, the system now allows for multiple offers per application. The data contains all applications filed through an online system in 2016 and their subsequent events until 2017. The Number of traces in this dataset is 31,509.

Cross-domain concept extraction

The frequency by which the motifs appear depends on the differences among the workflow environments and differences in domains. In the present paper different types of domains are used (e.g. Repair, Bioinformatics, Financial, and Hospital). At first the method computes the percent of dependent of some tags for a special domain (the financial data used for BPI Challenge 2017), as SFA step.

Table 3 shows the results of measuring the tag sets dependence percent of the financial domain (i.e. BPI Challenge 2017) as source in a BT containing of source and different target domains.

Results from Table 3 show the belonging measures of different workflow domains in percent with the case study of the financial BPI Challenge 2017 domain. Table 3 shows that almost (in average) 63.03% of the motifs in BT belong to the financial domain. The financial domain contains of the BPI Challenge 2017 source domain and different target domains. The results of the Table 3 show that most common tags are correlated with the type of execution environment for which the workflow is designed. The work identifies this based on the average of the results. As a result, the knowledge transfer has been done successfully, and the performance of learning has been improved by avoiding much expensive data labeling efforts. In the remainder of this section the results of the workflow improvement are presented.

Workflow improvement

After measuring similarities of workflows and finding frequent parts of them we can decide what domains might have a given unseen workflow. Then, we need to know what might be a better form of this unseen given workflow. In other words, how we can improve this given workflow. To achieve the best form of workflow we need to know the other better workflow model in its determined domain. We select some candidate of any domains such that the candidate is the best formatted in order and semantic. We select this candidate based on repetition and based on high score resulted from the bipartite graph. Then we extract U matrix for this candidate. In other hand, we construct U matrix for the unseen given workflowthat its content is similar to the U matrix that is selected from the candidate. By using the OSPA distance we compute the distance of U matrix to the given workflow. For example the distance of U matrix computed for the complex workflow event log from other financial workflows is computed. The result of this step is shown in Fig. 5.

The nearest workflow model to the test workflow is the model that helps to determine the order of the improved workflow model. From the results of SFA step we identify that the workflow test is more similar to financial type domain. Therefore, based on the results of Fig. 5 the DNN vector of 5th financial workflow stored in the dataset will be used to identify improved workflow order.

After improving the workflow the coupling of the improved workflow is computed (Table 4). Simplified models of workflows can be defined as models which are not

Table 3 Cross domain belonging measures of different workflow domains

	Target			
	Bioinformatics	Financial	Hospital	Repair
Repair %	0.0	0.0	0.0	3.70
Bioinformatics %	8.30	0.0	0.0	0.0
Financial %	74.58	54.32	38.60	84.62
Hospital %	0.0	0.0	39.89	0.0

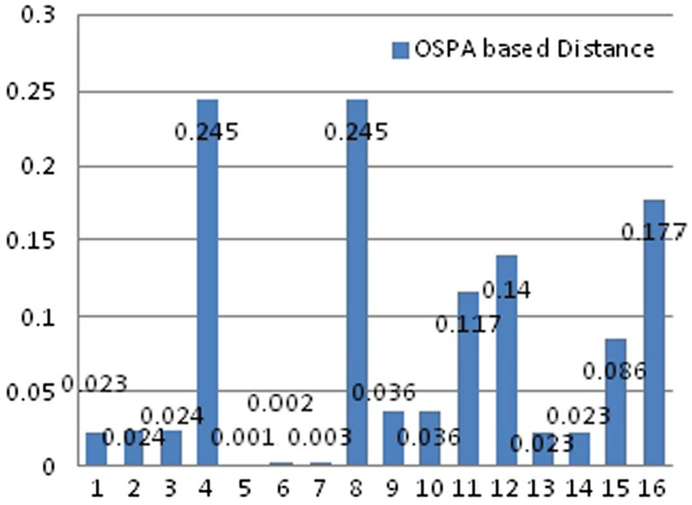

Fig. 5 OSPA based distance

Table 4 Workflow improvement analysis

Metric	Value	
	Not-improved	Improved
cp	0.5274	0.1860

overly complex, e.g., they are not extremely large and the density of arcs and coupling of elements is low [42, 43].

For a process that consists of a set of task tags (S) on the workflow structure the process coupling (cp) [42] is defined as follows:

$$cp = \begin{cases} \frac{\left\|\left\{(T_1,T_2)\in S\times S \| \overline{T_1}\neq\overline{T_2}\wedge(\widehat{T}_1 \bigcap \widehat{T}_2)\neq 0\right\}\right\|}{\|\overline{S}\|.(\|\overline{S}\|-1)}, & \text{for } \|S\| > 1 \\ 0, & \text{for } \|S\| \leq 1 \end{cases} \tag{4}$$

The results of the Table 4 show the coupling of the complex workflow after improving. The quality of the improved workflow is desirable in the sense that the cp value is lower after improvement. In the computation of cp, available resources could be involved to motifs. We assumed that resources are available for all activities. We did not restrict the motif extraction to resources.

Discussion

In the remainder of this section, we give a brief discussion from the perspective of the algorithm similarity measurement to explain the importance of Big Data analysis. In the age of Ubicomp and Big Data, most of the current computer systems will not be able to handle the whole dataset all immediately; thus, how to design a good data analytics framework [44] or platform and how to design analysis methods are both important things in the data analysis process. In the field of Big Data, we showed how the process mining can be helpful in bridging the gap between data and processes. In the process mining, our solution to handle Big Data was to use Transfer learning [9] and to benefit from using Deep learning. In the most of application domains, calculating process coupling (cp) would not be enough, and a more responsive (as soon as possible) detection of violations via e.g. Compliance monitoring [45] would be desirable. Compliance monitoring is a way to identify when a sequence of events deviates from the expected behavior. Because the management of business processes typically needs large interconnected environments, the concepts that need to be monitored become often very complex. It demands spreading the compliance monitoring task over a network of computing nodes, to achieve the desired scalability. However, most of the proposed monitoring approaches have limitations and they require responsible approaches that deal with Big Data and the IoT era [45]. Hence, as a future work we suggest to adopt compliance monitoring approaches that deal with distributed architectures [45] and services beside workflow improvement processes.

Conclusion

This paper proposed an algorithm for learning common motifs, a cross-domain similarity assessment and finally, workflow improvement. The first step of the algorithm used in the present paper was based on Transfer learning. It could discover an accurate representation for cross-domain data by fully exploiting the relationship between the domain-specific, and domain independent conceptual abstractions of workflow steps. This simultaneously was done via co-clustering in a common latent space. Then, it tried to improve a given complex workflow to the best known extracted workflow suggesting a unique relational model automatically. Improving the complex workflow was based on DNN. In fact, the proposed technique to Big Data transfer, and analysis was based on Transfer learning and Deep learning. Both of these algorithms lead us to handle Big Data challenge in WfMSs.

Authors' contributions

TK is the main author, she did all the research and explored different examinations. The idea was to workflow motif extraction between large amount of structured and unstructured data (Big Data). We used MZ experience to improve the process. Both authors read and approved the final manuscript.

Author details

[1] International Campus of Kharazmi, Shahrood University of Technology, Shahrood, Iran. [2] CE Department, Shahrood University of Technology, Shahrood, Iran.

Acknowledgements

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

Open source data (BPI Challenge 2012: http://dx.doi.org/10.4121/uuid:3926db30-f712-4394-aebc-75976070e91f). Open source data (BPI Challenge 2017: http://dx.doi.org/10.4121/uuid:5f3067df-f10b-45da-b98b-86ae4c7a310b). Open source data (BPI Challenge 2011: http://dx.doi.org/10.4121/uuid:d9769f3d-0ab0-4fb8-803b-0d1120ffcf54). Open source data (ArrayExpress: https://www.ebi.ac.uk/arrayexpress/). Open source data (Telephone Repair: http://www.processmining.org/prom/tutorials).

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Funding

Not applicable.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 31 May 2018 Accepted: 17 July 2018
Published online: 23 July 2018

References

1. Zeng L, Veeravalli B, Zomaya AY. An integrated task computation and data management scheduling strategy for workflow applications in cloud environments. J Netw Comput Appl. 2015;50:39–48.
2. Gubbi J, Buyya R, Marusic S, Palaniswami M. Internet of Things (IoT): a vision, architectural elements, and future directions. J Future Gen Comput Syst. 2013;29(2013):1645–60.
3. Schmidhuber J. Deep learning in neural networks: an overview. J Neural Netw. 2015;61:85–117.
4. Sohangir S, Wang D, Pomeranets A, Khoshgoftaar TM. Big Data: Deep learning for financial sentiment analysis. J Big Data. 2018;5(1):3–18.
5. Koohi-Var T, Zahedi M. Linear merging reduction: a workflow diagram simplification method. In: 8th international conference on iInformation and knowledge technology (IKT). Piscataway: IEEE; 2016. p. 105–10.
6. Starlinger J, Brancotte B, Cohen-Boulakia S, Leser U. Similarity search for scientific workflows. Proc VLDB Endow. 2014;7(12):1143–54.
7. Schoknecht A, Thaler T, Fettke P, Oberweis A, Laue R. Similarity of business process models—a state-of-the-art analysis. ACM Comput Surv. 2017;50(4):52–85.
8. Garijo D, Corcho Ó, Gil Y. Detecting common scientific workflow fragments using templates and execution provenance. In: the proceedings of the seventh international conference on knowledge capture. New York: ACM; 2013. p. 33–40.
9. Koohi-Var T, Zahedi M. Cross-domain graph based similarity measurement of workflows. J Big Data. 2018;5(1):18–34.
10. Bunke H, Riesen K. Recent advances in graph-based pattern recognition with applications in document analysis. J Pattern Recognit. 2011;44:1057–67.
11. Fiannaca A, Rosa ML, Rizzo R, Urso A, Gaglio S. An expert system hybrid architecture to support experiment management. J Expert Syst Appl. 2014;41:1609–21.
12. Ludäscher B, Weske M, McPhillips T, Bowers S. Scientific workflows: business as usual. In: International conference on BPM 2009. Berlin: Springer; 2009. p. 31-47.
13. Garijo D, Alper P, Belhajjame Kh, Corcho O, Gil Y, Goble C. Common motifs in scientific workflows: an empirical analysis. Future Gen Comput Syst. 2014;36:338–51.
14. Alper P, Belhajjame KH, Goble CA. Small is beautiful: summarizing scientific workflows using semantic annotations. In: IEEE 2nd international congress on Big Data. Piscataway: IEEE; 2013. p. 318–25.
15. Maguire E, Rocca-Serra Ph, Sansone S-A, Davies J, Chen M. Visual compression of workflow visualizations with automated detection of macro motifs. IEEE Trans Vis Comput Graph. 2013;19(12):2576–85.
16. Starlinger J, Cohen-Boulakia S, Khanna S, Davidson S, Leser U. Layer Decomposition: an effective structure-based approach for scientific workflow similarity. IEEE eSci Conf. 2014;1:169–76.
17. Wolstencroft K, Haines R, Fellows D, Williams A, Withers D, Owen S, Soiland-Reyes S, Dunlop I, Nenadic A, Fisher P, Bhagat J. The taverna workflow suite: designing and executing workflows of web services on the desktop, web or in the cloud. Nucleic Acids Res. 2013;41(W1):W557–61.
18. Goecks J, Nekrutenko A, Taylor J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 2010;11(8):R86.
19. Koohi-Var T, Zahedi M. Scientific workflow clustering based on motif discovery. Int J Comput Sci Eng Inf Technol (IJCSEIT). 2017;7(4):1–13.
20. Durante F, Sempi C. Principles of copula theory. Boca Raton: CRC press; 2015.
21. Dijkman R, Dumas M, Dongen BV, Kaarik R, Mendling J. Similarity of business process models: metrics and evaluation. Inf Syst. 2011;36(2):498–516.
22. Tosta FE, Braganholo V, Murta L, Mattoso M. Improving workflow design by mining reusable tasks. J Braz Comput Soc. 2015;21(1):16.
23. Stoyanovich J, Taskar B, Davidson S. Exploring repositories of scientific workflows. In: Proceedings of the 1st international workshop on workflow approaches to new data-centric science. New York: ACM; 2010. p. 7.
24. Medhata W, Hassan A, Korashy H. Sentiment analysis algorithms and applications: a survey. Ain Shams Eng J. 2014;5:1093–113.
25. Mendling J, Reijers HA, van der Aalst WMP. Seven process modeling guidelines (7 pmg). J Inf Softw Technol. 2010;52(2):127–36.
26. Fahland D, van der Aalst WMP. Model repair—aligning process models to reality. J Inf Syst. 2015;47:220–43.
27. Polyvyanyy A, van der Aalst WMP, Ter Hofstede AHM, Wynn MT. Impact-driven process model repair. ACM Trans Softw Eng Methodol. 2016;25(4):28.
28. Conforti R, Rosa ML, ter Hofstede AHM. Filtering out infrequent behavior from business process event logs. IEEE Trans Knowl Data Eng. 2017;29(2):300–14.
29. Augusto A, Conforti R, Dumas M, Rosa ML, Maggi FM, Marrella A, Mecella M, Soo A. Automated discovery of process models from event logs: review and benchmark. arXiv preprint arXiv:1705.02288, 2017.
30. Reichert M, Weber B. Enabling flexibility in process-aware information systems: challenges, methods, technologies. Berlin: Springer Science and Business Media; 2012.
31. Bergmann R, Müller G. Similarity-based retrieval and automatic adaptation of semantic workflows. Berlin: Springer; 2018. p. 31–54.
32. Pan SJ, Yang Q. A survey on Transfer learning. IEEE Trans Knowl Data Eng. 2009;22(10):1345–59.

33. Aggarwal CC. Data classification: algorithms and applications. Boca Raton: CRC Press; 2014.
34. Lee H, Battle A, Raina R, Ng AY. Efficient sparse coding algorithms. In: Schölkopf B, Platt JC, Hoffman T, editors. Advances in neural information processing systems. MIT Press; 2007. P. 801–8. https://papers.nips.cc/paper/2979-efficient-sparse-coding-algorithms.pdf.
35. Schoknecht A, Fischer N, Oberweis A. Process model search using latent semantic analysis. In: International conference on business process management. Berlin: Springer; 2016. p. 283–5.
36. Verbeek HMW, Gunther CW. XES standard definition 2.0. Technical report. BPM Center Report BPM-2014. http://bpmcenter.org/wp-content/uploads/reports/2014/BPM-14-09.pdf. Accessed 31 May 2018.
37. van Dongen BF. Bpi challenge. 2011. https://doi.org/10.4121/uuid:d9769f3d-0ab0-4fb8-803b-0d1120ffcf54. Accessed 31 May 2018.
38. van Dongen BF. Bpi challenge. 2012. http://dx.doi.org/10.4121/uuid:3926db30-f712-4394-aebc-75976070e91f. Accessed 31 May 2018.
39. Tikhonov A, Parkinson H, Petryszak R, Sarkans U, Brazma A. ArrayExpress update-simplifying data submissions. Nucleic Acids Res. 28(43):D1113–6, 2015. https://www.ebi.ac.uk/arrayexpress/. Accessed 11 Jan 2018.
40. http://www.processmining.org/prom/tutorials. Accessed 11 Jan 2018.
41. van Dongen BF. Bpi challenge. 2017. https://data.4tu.nl/repository/uuid:5f3067df-f10b-45da-b98b-86ae4c7a310b. Accessed 11 Jan 2018.
42. Vanderfeesten I, Reijers HA, van der Aalst WMP. Evaluating workflow process designs using cohesion and coupling metrics. Comput Ind. 2008;59(5):420–37.
43. Janssenswillen G, Donders N, Jouck T, Depaire B. A comparative study of existing quality measures for process discovery. J Inf Syst. 2017;71:1–15.
44. Tsai CW, Lai CF, Chao HC, Vasilakos AV. Big data analytics: a survey. J Big Data. 2015;2(1):21.
45. Loreti D, Chesani F, Ciampolini A, Mello P. A distributed approach to compliance monitoring of business process event streams. Future Gen Comput Syst. 2018;82:104–18. https://doi.org/10.1016/j.future.2017.12.043.