Journal of Big Data

METHODOLOGY Open Access

Cross-domain graph based similarity measurement of workflows

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

*Correspondence:
tahere.koohi@gmail.com
[1] International Campus of Kharazmi, Shahrood University of Technology, Shahrood, Iran
Full list of author information is available at the end of the article

Abstract

The aim of this article is to analyze search and retrieval of workflows. It represents workflows relatedness based on transfer learning. 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. This paper proposes a cross-domain concept extraction by similarity measurement and has a new research effort at the intersection of workflow domains. It deals with the huge amount of structured and unstructured data (Big Data) that is a demanding task when working on real-life event logs. The proposed method in this paper gives a general solution in the sense that it can be coupled to any Process Aware Information System.

Keywords: Scientific workflows, Business process, Big Data, Transfer learning

Introduction

Workflows are sequences of process components that can be categorized in different domains including scientific workflows and business workflows. 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 and software issues (see Fig. 1 for example), business process models describe the processes of companies or other organizations focusing on the sequences of activities, roles, and events. Business workflows are the automated parts of business processes. Most research works prefer to have contributions focusing either on scientific workflows or business workflows (for example see [1–3]).

Workflow usage nowadays grows in many parts of computer science world. Workflows might be used in software development, coding, web developing, etc. Now large amounts of efforts grow to make workflow proper, easy to use and understandable in computer science world. Software scientists are interested in forming workflows as exchangeable classes of tools or objects to tackle process flows. In line with this effort the usage of workflow in public domain also grows; workflow can be executed in different implementation styles depended on the process context, e.g. e-commerce, bioinformatics, etc.

Springer Open

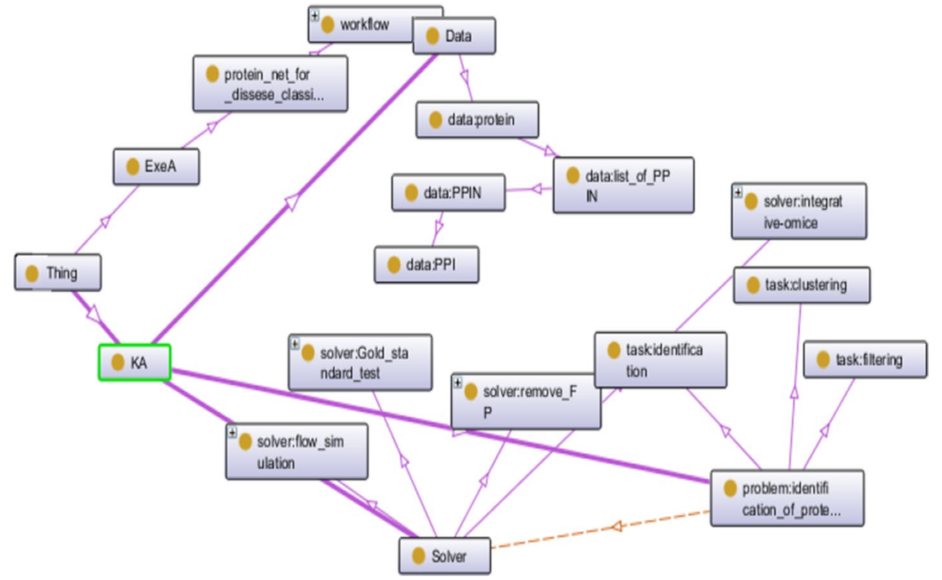

Fig. 1 Example of a scientific workflow graph

Figure 1 shows a scientific workflow graph. The tool used was Protégé 4.x[1]. Protégé is mainly used for ontology demonstration but it also provides some features to workflow demonstration. This workflow describes a simple process for bioinformatics experiment, and it demonstrates the interaction between proteins (protein–protein interaction: PPI). We designed this workflow based on the protein sequence interaction in a disease described in [4]. The granulized components of this workflow are in two major areas: Knowledge Area (KA) and Execution Area (ExeA). Each area has various types. In many workflow cases things may be grouped in different types to better describe the real process. For instance in Fig. 1 KA consists of three major types of components: Solver components, Problem component and Data component. In our example in Fig. 1, the Problem component is connected to the Solver component via a dashed red line. The nature of things used in workflow is limitless. Because describing parts of workflows are words and the words has no limitation and so many tags might be used in one simple workflow. This makes workflow fresh, full of variety and hence heavy to execute. Such burden may not be carried with only hardware. Software solutions might also be handle workflow usage. Some of the proposed software solutions prefer to avoid of useless parts before the workflow becomes heavy. Other software solutions may try forecasting what the designer wants and provide some template workflows. These templates are pre-executed and are easier to run even when the designer changes them or adds some things to them. Prediction of workflows needs enormous data in many fields. Classification of this enormous data in proper classes might be the first step to forecast what the designer wants. In the case of workflows the data classification might be applied to form templates. Workflows may use some Domain-Specific (DS) and Domain-Independent (DI) templates. In addition, some DS and DI templates may co-occur in workflow

[1] https://protege.stanford.edu/.

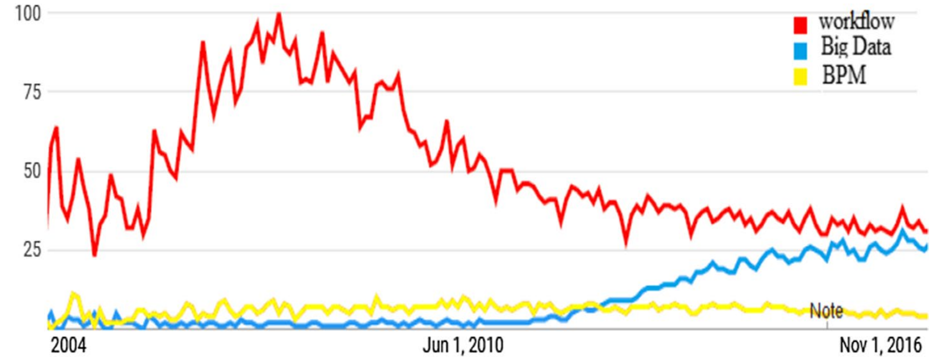

Fig. 2 Interest over time. Workflow vs. Business Process Management (BPM) vs. Big Data

domains frequently, which mean there may be a correlation between them. This observation motivates us to propose a Spectral Feature Alignment (SFA) based method to align workflow DS templates from different domains in a latent space by modeling the correlation between the DI and DS workflow templates in a bipartite graph and using DS features as a bridge for cross-domain Big Data classification.

As a new contribution, we address a classification algorithm that has a workflow context alignment from the relatedness aspect of workflows. In our classification, we use some part of former workflows and reuse them to make a classifier. Today, scientific workflows have become the workhorse of Big Data analytics for scientists [5]. Aggregating these workflows with other workflows in other domains may cause challenges of Big Data [6–8]. The problem of enormous data of workflow is not only arises of the context of workflow but also arises of the order of its context. Hence attempt to solve workflow enormous data problem is not avoidable.

According to Fig. 2 workflow interest researches have stated at a balanced point. On the other hand, Big Data interest has increased over time. The tool used was Google Trends.[2] We target at finding an effective approach for the cross Big Data domain concept extraction problem. Concept extraction can help us to handle the problem of dealing with enormous workflow data and to predict what the designer wants. The approach is based on transfer learning using Latent Semantic Indexing (LSI) [9]. Transfer learning [10] is commonly used when we have several domains of source and targets with different data to make classifier have better recognition. We have two domains of workflows with big volume. Hence, we try to adapt our classifier to use transfer learning settings. Some researches [11] mention that transfer learning makes some crashes to recognition or to classification when the source and target data of several domains has large difference. Actually, in workflow domain the main problem is the large difference between enormous data. To manage the transfer learning crashes we benefit from LSI when we use some unlabeled data from background knowledge.

LSI is a collection of matrix algorithm and some probabilistically analyses on it. The weakness of this collection when it is used in workflow concept extraction is that it

[2] https://trends.google.com/.

measures similarities between workflows that all of them are in a same domain. In this paper, we deal with this issue by addressing our cross-domain feature extraction. The basic idea is (a) to extract unique content-bearing workflow motifs from the set of workflows treating these motifs as features and (b) to then represent each workflow as a vector in this feature space. Thus, the entire workflow collection may be represented by a motif-by-workflow U matrix in which rows correspond to workflows and columns are motifs. By extracting workflow motifs we obtain frequent tags of each domain and the vector representation helps to avoid complexity of workflow structures. 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 [12].

An alternative structural approach which we use to represent each workflow is based on graphs as basic specifications for workflow structures. With the graph based representations of workflows we can benefit from the previous graph based search and analyze methods. Collective classification model is proposed in this paper to overcome the granularity of workflow components and this model is used in different workflow domains not only scientific workflow or business workflow. Main contributions of this paper are as follows. First, it shows how to adapt knowledge acquired from structured and unstructured data of different domains. Finally, it gives an algorithm learning common features; namely, concepts, as domain independent conceptual abstractions for workflow steps. For the purpose of this paper we embed 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 to obtain similarities between workflows. We train a model that learns representational embeddings for motifs from a large collection of unlabeled data using a generative model. We view the training method as a problem of cross-domain concept extraction of different workflows.

In the present paper some recent approaches to workflow similarity are reviewed. Particularly, this article describes a novel approach to workflow analysis using similarities. In the experimental evaluation, we involve several data sets with diverse properties as Big Data as a great source. According to results, one can consider the proposed cross domain workflow motif extraction as a potentially useful alternative to traditional approaches focusing just on one domain.

The paper has organized as follows: A review of relevant works is conducted in "Related works" section. In the next section we introduce the "Problem setting". The new contributions reported in this paper are an extension of the related works which is described in "Similarity measurement" section. The section describes the similarity measurement part of the method. Section proposes the "Method overview". Results and experiment settings are mentioned in "Results and discussion" section. Finally, section "Conclusions" concludes the paper.

Related works

PAISs [13] must continually adjust and evolve the performance to ensure the capability of modeling, enacting, and supporting real-world processes. Recently, Workflow Recommenders (WRs) [14] in PAISs have proposed recommendation services that aim at suggesting frequent combinations of workflow tasks for reuse. Some of these recommenders apply data mining techniques such as similarity measurement to help users find items to improve their workflow designs by prediction [14, 15]. For instance Wang et al. [15] has proposed and evaluated a design space of four different WR algorithms based on data mining, which can be used to recommend new workflows and their associated videos to software users. One of the limitations of the algorithms proposed in [15] is that they used heuristics. Besides, the authors in [15] did not evaluate their proposed algorithms with different datasets of different domains. Tosta et al. [14], as another example proposed a WR method that works potentially in different domains. However, they just focused on scientific workflow domains in practice.

Similarity measurement is a common task in workflow retrieval [16] and the related fields in data mining. A vast number of workflow representations have been designed for similarity measurement of workflows given about feature vectors [17]. However, workflow representation by graphs has several advantages over feature vectors. First, graphs are able to represent not only the values of object properties, i.e. features, but can be used to explicitly model relations that exist between different parts of an object. The graph based representation of workflows is quite general as it allows representing different workflow types, including business workflows and scientific workflows. Moreover, graphs do not suffer from the constraint of fixed dimensionality [12]. Workflows, however, require a graph based representation to appropriately cover control- and dataflow [18]. For an example of representing workflows by graphs in similarity assessment, the work at [18] contributed to the core problems of process oriented case-based reasoning, particularly to the representation of semantically annotated workflows. Although the authors in [18] presented algorithms that work on the two domains of scientific and business, the semantic workflows rely heavily on ontology to provide, and it is one of the major limitations of their work.

Some workflow similarity measurements are text based [19], and some others consider different aspects of workflows (such as workflow motifs [17], or workflow structures [1, 20]). For example a workflow similarity measurement method is the work at [19] that used feature selection techniques based on text. It treated workflows as group of words (BWs). In [19] for each workflow, the pre-processing component counts the number of occurrences of each term. The work used Latent Semantic Analysis (LSA) to consider the shared occurrence of terms. It then produced similarity values between pairs of workflows. The most common feature selection step in text-based similarity measurements is pre-processing (the removal of stop-words and stemming [returning the word to its stem or root e.g. flies→ fly)] [21]. Another example is [22] in which the most frequently used modules and frequent tag sets were explored.

Structure based workflow similarity measurements have the advantage of considering relatedness of workflow elements, over text based workflow measurements. To introduce the structural aspects of workflows in the similarity assessment, several graph algorithms have been proposed for similarity measurement such as sub-graph isomorphism, maximal common sub-graphs, or edit-distance measures [18–23], which usually cause problems because of their high computational complexity. To avoid these problems, the present paper relies on motifs as common steps of workflows to be extracted. We suggest to workflows be in their graphical model and so the structure of workflows is considered in our proposed method. Our method finds common motifs automatically, to improve workflow search and retrieval. We use workflow motifs [17], to enable relating workflows from one domain to other workflows by different domains. So far techniques were focused on discovering workflow motifs in one domain [17, 24–26]. The present paper addresses extraction of workflow motifs at intersection of different domains causing large amount of data. The extraction of useful information out of large amount of data is the main challenge of the data mining field. To tackle this challenge, we propose a process mining [27] method based on transfer learning [28]. Process mining aims at discovering, analyzing, and extending formal models of process revealing the real process in a system, using the event log containing the footprints of real process executions [29]. Transfer learning has been proposed to allow domains, tasks, and distributions to be different. The advantage of transfer learning is that it can intelligently apply knowledge learned earlier to solve new problems faster [28].

Traditionally, tasks such as classification and clustering of workflow are solved by defining an applicable distance measure (such as Cosine, or Euclidean [17]) and then using an algorithm that is based on distances exclusively. Recently, a prominent alternative class of graph embedding methods based on spectral clustering has been used [12]. Spectral clustering theory is concerned about understanding how the structural properties of graphs can be characterized using eigenvectors of the connection or Laplacian matrix. Although graph spectral theory has the advantage of removing the matching step to bring graph nodes into correspondence, this approach remains somewhat limited. For instance, spectral methods are not fully able to cope with larger amounts of noise. This stems from the fact that the Eigen decomposition is very sensitive to structural errors, such as missing or spurious data known as contradictory data [7]. To address this issue, the present paper uses a technique based on Point-wise Mutual Information (PMI) extraction like [30] and discovers knowledge from provenance graphs like [31].

Problem setting

Previous section was about related works. In this section, we give a problem setting for further analysis. Before giving a formal definition of the problem, the paper first presents 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 *(Motif)* Given a sequence of tags $t_1t_2...t_n$ where t_i is a conceptual abstraction of workflow step tag from a Lexicon T, a Motif M in workflows is a small functional unit that occurs significantly more frequently than expected. It adds a layer of abstraction that generalizes the functionality of each step or set of steps, helping understanding main functionality of workflow. Here, Domain Independent (*DI*) sets are defined as *M* Motifs.

Definition 3 *(Type)* Given a specific domain D, type data shows the type of workflows correspond with 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*.

Problem definition (cross-domain workflow alignment)

Based on the definitions described above, now the problem is defined. Having a set of labeled data (e.g. Tag sets) from a source domain, to train a model for a target domain, this work leverages some unlabeled data from the target domain. In detail, first for the dataset it creates workflow graph from workflow descriptions, and applies a function mapping on each trace of the workflow event log. Source and target domains are the event logs which represent traces (sequence of events) of workflows. We propose Spectral Feature Alignment (SFA) [32] based algorithm to find a new representation for cross-domain process data, such that the gap between domains can be reduced. SFA uses some *DI* conceptual abstractions for workflow steps (i.e. Workflow motifs) as a bridge to construct a bipartite graph to model the co-occurrence relationship between *DS* tags and *DI* ones (*U* matrix). As such it bridges the gap between domains. The idea to find the relatedness between workflows is that if two DS tags have connections to more common DI ones in the graph, they tend to be aligned 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.

To achieve aims of this alignment the method follows these steps:

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

For modeling, our classifier is based on the following formulation, Given D_{Src} for source domain and D_{Tgt} for target domain we learn *f* as classifier function

$$D_{Src} = \{t_{Src_i}, t_{Src_j}\}_{i,j=1}^{n_{Src}},$$
$$D_{Tgt} = \{t_{Tgt_i}, t_{Tgt_j}\}_{i,j=1,\mathrm{n_{Tgt}} << \mathrm{n_{Src}}}^{n_{Tgt}},$$

$$f = \sum_{Tgt} \sum_{i=1,j}^{n_{Tgt}} l(f_{Tgt}(t_{Tgt_i}), t_{Tgt_j}) + \lambda \Omega(U)$$

where λ and Ω are transfer learning parameters, and f_{Tgt} has good generalization on unseen tags t_{Tgt}. We obtain the *U* matrix for each domain. Following the method is explained in detail.

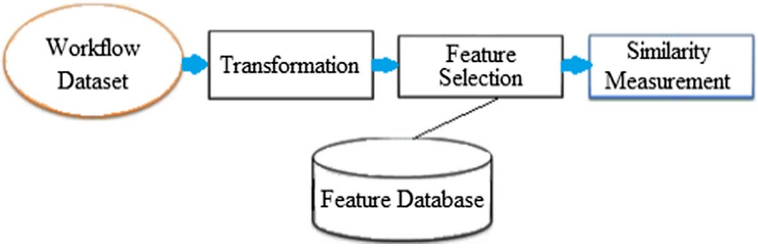

Fig. 3 Similarity measurement overview

Similarity measurement

In the present paper features are extracted based on LSA in a transformation step as transfer learning [33]. Transfer learning automates learning of features across workflows, from different domains and of very different natures. 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 [33]. Same workflows can be grouped into one group in a feature database (Fig. 3).

This work concerns with learning motifs even only a few labeled data are given. It supposes that there are some higher-level features that can help training the model of classifier. It aims to *align* DS (non-pivot) *features* from different domains by workflow domain adaptation. Two solutions of higher-level feature construction are *Sparse Coding* (*SC*), and *Deep learning* [34, 35]. The paper benefits from SC that is an unsupervised *feature construction* method. SC learns basis functions that capture higher-level features in the data have given only unlabeled input data [36]. Here, the co-occurrence matrix construction is based on SC and we assumed each observation is a sparse combination of the lexicon elements. Tags assigned to workflow steps in a repository can be used for similarity measurement, as done in [19]. The tags assigned to steps in a workflow are treated as a bag of tags and the workflow similarity can be calculated in the same way as in the Bag of Words approach described in [19]. The presented work is similar the work at [19] in the sense that a bag of tags is used. 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 in this work align tags automatically with spectral clustering and feature generation tasks by a special setting of transfer learning [33], which aims at transferring knowledge across workflow domain motifs. Another advantage of our proposed method over the work at [19] is that no stop-word removal or other pre-processing of the tags is performed in this paper.

We present a cross-domain concept extraction setting for domain adaptation. In the proposed setting, to bridge the information gap between domains, SFA algorithm aligns DS tags from different domains into unified clusters, with the help of DI tags as a bridge. In this way, clusters can be used to cut the gap between DS tags of domains, which can be used to train classifiers in the target domain accurately. Compared to previous approaches, SFA can discover a robust representation for cross-domain data by fully exploiting the relationship between DS and DI tags (workflow motifs) via simultaneously co-clustering them in a common latent space [32]. Hence, the paper experiments with a clustering method for similarity measurement of workflows called spectral clustering. Moreover, it removes the fraction of observations to be discarded as outliers to enhance the clustering performance. Outliers are observations that generated by a different mechanism [37]. If actual *outliers* are not discarded from the data set, they corrupt the results.

SFA for concept extraction uses Eigen-decomposition of the co-occurrence matrix, e.g. Latent Semantic Analysis (or sometimes Indexing) (LSA or LSI) [32]. Construction of the concept space (also called embedded or latent space) is done by applying LSA. LSA is a straightforward statistical method that projects data onto a lower dimensional space, by an Eigen-decomposition of the tag co-occurrence matrix. In this cross-domain extraction work, derived co-occurrence patterns called concepts, define the dimensions in the new concept space. Estimating correlations to measure dependencies between DS data and DI data helps to select good pivot features. The method selects a subset of frequent tags that occur in domains called pivots or DI tags. Pivots are DI conceptual abstractions for workflow steps.

There are some approaches to estimate correlations between pivot and DS features. Examples of such methods are Structural Correspondence Learning (SCL) [38] and SFA [32]. SCL is similar to SFA, with this difference that it heuristically selects pivot features and this might not guarantee the best performance on target domains.

This work adapts a spectral clustering algorithm, similar the work of [30], on the bipartite graph to co-align DS and DI features. We obtain the co-occurrence matrix based on bipartite graph. If DS features co-occur with DI features in some feature vector, then an edge is formed between the two features in the bipartite graph. Finally, a spectral clustering is performed on the bipartite graph and a value is assigned to each of the features in DS and DI to show how much they are similar. that represents the two sets of features. k principal components extracted by LSA can automatically apply the clustering in the subspace spanned. This implies that a mapping function constructed from k principal components can cluster original data and map them to a new space spanned by the clusters simultaneously. To measure how much more often tags occur together than expectation, if they were independent, Eq. (1) is used [30].

$$f(s,t) = \log\left(\frac{\frac{c(s,t)}{N}}{\frac{\sum_{i=1}^{n} c(i,t)}{N} \times \frac{\sum_{j=1}^{m} c(s,j)}{N}}\right) \tag{1}$$

where f (s, t) is PMI between two tags s and t, and c (s, t) denotes the number of motifs in which two tags s and t co-occur, n and m respectively denote the total number of s tags and t tags, and $N = \sum_{i=1}^{n} \sum_{j=1}^{m} c(i,j)$.

Next, for two tags s and v (represented by feature vectors s and v, respectively), the relatedness τ (v, s) of the tag v to the tag s is computed as follows [30]:

$$\tau(v,s) = \frac{\sum_{t \in \{x|f(v,x)>0\}} f(s,t)}{\sum_{t \in \{x|f(s,x)>0\}} f(s,t)}. \tag{2}$$

The relatedness score τ (v, s) can be interpreted as the proportion PMI weighted features of the tag s that are shared with tag v. Note that PMI values can become negative in practice even after discounting for rare occurrences. To avoid considering negative PMI values, only positive weights in Eq. (2) is considered [30]. Using Eq. (2) is for overcoming the bias towards infrequent elements and features [30]. To compute the PMI values in feature vectors, the co-occurrence information between numerous tags should be stored. For this, a lexicon matrix from a large set of DS tags and DI tags can be used. Particularly, workflow motifs can be avoided that are likely to have very small relatedness scores thus are unlikely to become neighbors of a given s tag by using approximate vector similarity computation techniques.

The proposed method to deal with contradictory data trivially chooses the best θ to optimize performance. It tries to discard the (1-θ) features that have little connectivity in the co-occurrence matrix, thus outliers are eliminated. For this purpose, it has two level outlier removals. First, it removes outliers by applying a threshold on the source data domains. Second, it applies a threshold based on PMI weights obtained by Eq. (2). It removes the data that is below the PMI values. For this a threshold limits the connections. After applying the threshold, transformation of the tag-workflow co-occurrence matrix indicates the relation between various tags in the data. U matrix of workflow tag relations from source domains and target domains is constructed based on co-occurrence matrix by finding k largest Eigen Values introduced in [32]. The relatedness measure defined in Eq. (2) is used to construct a motif sensitive lexicon in which, for each tag element s lexical elements t is listed up that co-occur with t (i.e. $f(s, t) > 0$) in the descending order of the relatedness values τ (v, s). Next section is about the method overview.

Method overview

As a summarization, steps of the multi-domain workflow alignment are described in Algorithm 1.

Algorithm 1 Workflow Similarity Measurement

```
Input: $(L, data)$, $L$ is a log and or a workflow specification over $T$, and data is a function mapping each $tracez \in L$ to a set of pairs $data(z) = (t1, x1), \ldots, (tq, xq)$ such that $xi \in type(z)$ for each $i \in 1, \ldots, q, andt1, \ldots, tq = t \in T | task(t) = t[j]$.
Load dataset for source and target domains
for each source and target data do
    Remove outlier data with a given threshold
    Assign type labels to workflow steps
    M ← select Pivot features (DI)
    DS ← select Domain Specific features
    Compute co-occurrence matrix that shows relations between different data in dataset
    Compute the bipartite graph that has relations between different motifs in dataset.
    Find $k$ largest Eigen Values (Construction of $U$ matrix)
Return a classifier $f$ trained
Search the best match of $k$ largest Eigen Values of a given workflow with others
Output: Extracted Concepts
```

The method has to be able capture descriptions of complex workflow models. So, it uses ISA-Tab handler for scientific workflows that allows the linkage of a single sample to multiple analyses employing various assays [14]. For other workflows in other domains a XML-Based workflow Event Logging Mechanism has been used in XES format [39]. Algorithm 1 uses the proposed SFA based method. 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. It 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 purpose, different domains of workflows are co-aligned. The most similar U matrix is detected using lexicon. Lines 4–10 of the algorithm describe the discovery step of workflow motifs.

Results and discussion

This section performs experiments on some workflow datasets from different domains, and shows that a universal function to model workflow classification can be applied based on SFA and distributional semantics. It gives an approach to cross-domain concept extraction (CDCE) and so it can work on every workflow domain, e.g. scientific and business workflows. In reminder of this section we first present the "Experiment setup", and then we analyze the results. We describe our experiments on four real world datasets and show the effectiveness of our SFA based method for cross-domain alignment in terms of handling Big Data when the number of workflow Tags increases. Finally we discuss about the results.

Experiment setup

To analyze the proposed method four datasets are considered from different domains of hospital, financial and scientific workflows. The used datasets to train model are as follows:

1. A real-life event log of an academic hospital originally intended for use in the first Business Process Intelligence (BPI) Challenge [40] is used. It contains 1143 workflow traces.
2. Other dataset is the financial dataset that contains of 4366 traces used for *BPI Challenge* 2012 [41].
3. For scientific workflows the work extracted information from a resource holding over 70,665 experimental design workflows (ArrayExpress) [42]. It used a subset of this collection, including 29 scientific workflows.
4. Another financial dataset used in this paper contains of an event log containing of 31,509 traces pertains to a loan application process of a Dutch financial institute used for BPI Challenge 2017 [43]. The data and the process under consideration is the same as [41].

We construct the bag of tags by given sources and targets. With these datasets from different domains, a huge amount of Big Data can be integrated into the decision making process. Such a Big Data can be considered as a great source and target.

Discussion

The frequency by which the motifs appear depends on the differences among the workflow environments and differences in domains. In this paper different types of domains are used (e.g. Bioinformatics, Financial, and Hospital). To construct the U matrix with four domains, we obtained 16 bags of tags. Any bag of tags has a correspondence U

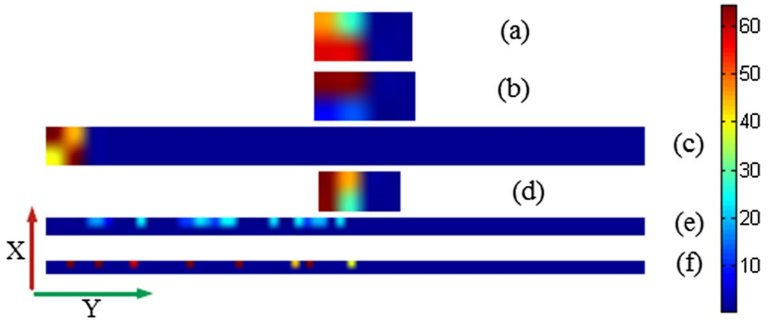

Fig. 4 *U* matrix values, k = 2 (a) hospital and BPI-2012 dataset, (b) scientific and hospital dataset, (c) BPI-2017 and scientific dataset, (d) scientific and BPI-2012 dataset, (e) hospital and BPI-2017 dataset, (f) BPI 2012 and BPI 2017 dataset. The X axis shows feature vectors and the Y axis represents workflow motifs

matrix that obtained by computing the K largest Eigen Values of the bipartite matrix. The parameter of K is important to achieve more accurate distinctions between related workflows. We illustrated the U matrix of the six bags of tags in Fig. 4. For example Fig. 4a is a U matrix based on hospital and BPI-2012 datasets. Figure 4 shows the values of U matrix when the method has $K = 2$ largest Eigen Values in 2D representation of the matrix.

In Fig. 4, the colored values in the U matrix represent workflows relatedness between various data of different domains. The greater length of the U matrix in Fig. 4 represents the more varied data being examined. Non-zero values are shown with warmer color points. The zero values represent the hidden data that are dropped by threshold because of their different mechanism. These values indicate that there is a relationship between the data obtained from the source and target domains. Therefore, as the number of color points and colors are more varied, more data can be distinguished. As can be seen in Fig. 4, in the U matrices (a) (b), and (d), the matrix length is shorter than other matrices. This shows that the matrix can scan a small range of data. Also, the non-zero values of the matrices (a) (b), and (d), are centered around and nearby one segment. Although in Fig. 4c has studied more data, but still non-zero values are concentrated around one segment. So, based on that, we will have almost the same behavior. The best output about the data distinction is related to (e) and (f) matrices. Both matrices of (e) and (f) cover a larger field of values. This is caused by the color points that are scattered uniformly.

As can be seen in figures, when the domain has various values or enormous data the U matrix has varied values and becomes more complex and extended. Because the execution time of the mentioned method is constrained only once when the classifier is trained, this complexity causes increasing in run time, and might not be crucial. Because the execution time of the method is constrained only once in the system when the classifier is trained. Here the U matrices (e) and (f) have more data distinction capability.

It can be observed from Table 1 that the execution time increases when the datasets become big and have more tag of tasks. On the other hand as can be seen from Fig. 5 it is not essential to have Big Data to increase the processing time of CDCE method, and in most of times the method deals with Big Data. Figure 5 datum shows that in each implementation of the source domains, the proposed method can decrease the runtime even if the target domain includes enormous data. In the more detailed explanation, the volume

Table 1 CDCE on Big Data

Source	Target	Number of tags	Time (ns)
BPI-2012	BPI-2017	1,542,600	161942430734
BPI-2012	Scientific	2,950,89	47502683874
BPI-2012	Hospital	454,923	51955451599
BPI-2017	Scientific	1,247,887	64293109780
BPI-2017	Hospital	1,407,721	57628940065
Scientific	Hospital	160,210	44093711337

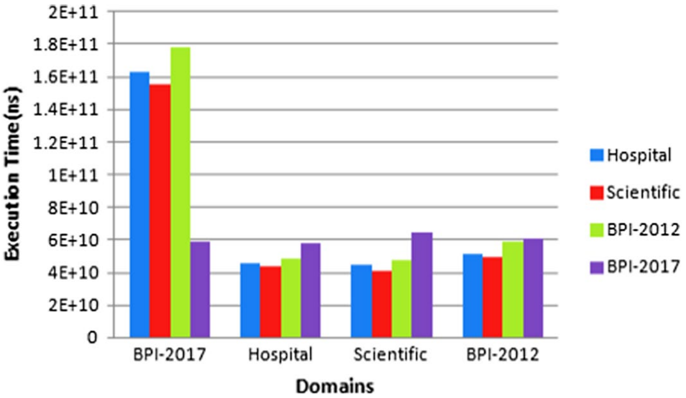

Fig. 5 Processing time of CDCE for different dataset sizes.Vertical axes shows time (ns)

of the BPI-2017 data is the most, and the other domains have less data volume than this domain. Clearly when the BPI-2017 data is added to each of the source domains, the Big Data problem finds a heavier form.

Most real workflows are complex in the sense that they present many non-trivial features. With ever increasing number of workflow repositories organizing and categorizing them to diverse need of user by manual means is a complicated job. For better organizing real workflows, we proposed a classification algorithm for aligning workflow context from the relatedness aspect of workflows. In our classification, we used some sub-workflows of former workflows as motifs. Constructing co-occurrence matrix of workflow motifs in different domains allows us to retrieve a similar flow from domain data when we reach an unseen workflow. Retrieving similar flows from previous workflows provides better organization for search results. We applied our method on real-life datasets, obtaining good results about handling Big Data when the number of workflow tags increases. The advantage of our algorithm on other methods present in literature, like LSA, or graph spectral theory, is that it removes outliers and considers data relatedness between domains, and can thus be effectively applied to a wide range of PAISs working on different domains like scientific or business. For more detail on comparative study of the methods see Table 2.

By the way, different proposals and implementations of techniques are proposed to improve the performance of systems that deal with Big Data challenge [8, 28, 44]. For example, except transfer learning used in the present paper some promising learning

Table 2 Overview of methods (for an explanation see text)

Method	Pivot features selection	Eigen value	Outliers sensitive
SCL	Heuristical	No	No
LSA	No	Yes	Yes
SFA	Non-heuristic	Yes	No
Graph spectral theory	No	Yes	Yes

methods in recent studies, such as representation learning, deep learning, distributed and parallel learning, active learning, and kernel-based learning are proposed to deal with Big Data [28]. An example of dealing with big data in workflow management systems is [44] which scales workflows to High Performance Computing (HPC) supercomputing systems. Another solution remains in workflow representations; one can encode workflows for workflow indexing to further reduce processing time and scale similarity search to sizes of current repositories [17].

Conclusions

This paper proposed an algorithm for aggregating large volume of data, and discovered common workflow motifs. The proposed method based on the relatedness of workflows can measure the similarity of unseen workflows. It can be easily adapted to any workflow domains. Cross domain alignment can give us a multi side view to domains, but it costs on time and when the datasets are as big as real-life workflows data it exhibits its expensiveness so we utilize Transfer learning to reduce the middle process sequences time of the cross domain workflow alignment. Transfer learning as our results show can give a solution to time consummation of checking data. The proposed technique to deal with Big Data was based on transfer learning. It can limit time consummation of multiple domains that caused Big Data, by constructing a co-occurrence matrix. The co-occurrence matrix constructed based on two domains per each run. Hence one of future works is to form the co-occurrence matrix based on more than two domains per each run, to reach a multi-cross domain extraction.

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 use MZ experience to improve our work. 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 (https://www.ebi.ac.uk/arrayexpress/).

Consent for publication

Not applicable.

Ethics approval and consent to participate
Not applicable.

Publisher's Note

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

Received: 14 March 2018 Accepted: 15 May 2018
Published online: 24 May 2018

References

1. Starlinger J, Brancotte B, Cohen-Boulakia S, Leser U. Similarity search for scientific workflows. In: Proceedings of the VLDB Endowment (PVLDB), VLDB Endowment. 2014;7(12):1143–54.
2. 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.
3. Dijkman R, Dumas M, Van Dongen B, Kaarik R. Similarity of business process models: metrics and evaluation. Inf Syst. 2011;36(2):498–516.
4. 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.
5. Alper P, Belhajjame K, Goble CA. Small is beautiful: summarizing scientific workflows using semantic annotations. In: IEEE 2nd international congress on Big Data. 2013.
6. Sohangir S, Wang D, Pomeranets A, Khoshgoftaar TM. Big Data: deep learning for financial sentiment analysis. J Big Data. 2018;5(1):3–28.
7. Nwagwu HC, Okereke G, Nwobodo C. Mining and visualising contradictory data. J Big Data. 2017;4(1):36–47.
8. Tan W, Blake MB, Saleh I, Dustdar S. Social-network-sourced big data analytics. IEEE Internet Comput. 2013;7(5):62–9.
9. Papadimitriou CH, Raghavan P, Tamaki H, Vempala S. Latent semantic indexing: a probabilistic analysis. In: Proceedings of the ACM symposium on principles of database systems. 1998.
10. Iqbal MS, Luo B, Khan T, Mehmood R, Sadiq M. Heterogeneous transfer learning techniques for machine learning. Iran J Comput Sci. 2018;1:31–46.
11. Xiang EW, Cao B, Hu DH, Yang Q. Bridging domains using world wide knowledge for transfer learning. IEEE Trans Knowl Data Eng. 2010; 22(6):770–83.
12. Bunke H, Riesen K. Recent advances in graph-based pattern recognition with applications in document analysis. J Pattern Recognit. 2011;44:1057–67.
13. Grambow G, Oberhauser R, Reichert M, editors. Advances in intelligent process-aware information systems: concepts, methods, and technologies. Cham: Springer; 2017.
14. Tosta FE, Braganholo V, Murta L, Mattoso M. Improving workflow design by mining reusable tasks. J Braz Comput Soc. 2015;21(1):1–16.
15. Wang X, Lafreniere B, Grossman T. Leveraging community-generated videos and command logs to classify and recommend software workflows. In: Proceedings of the 2018 CHI conference on human factors in computing systems. ACM. 2018. p. 285.
16. Bergmann R, Müller G. Similarity-based retrieval and automatic adaptation of semantic workflows., Synergies between knowledge engineering and software engineeringCham: Springer; 2018. p. 31–54.
17. Koohi T, Zahedi M. Scientific workflow clustering based on motif discovery. Int J Comput Sci Eng Inf Technol IJCSEIT. 2017;7(4):1-13.
18. Bergmann R, Gil Y. Similarity assessment and efficient retrieval of semantic workflows. Inf Syst. 2014;40:115–27.
19. Schoknecht A, Fischer N, Oberweis A. Process model search using latent semantic analysis. In: International conference on business process management. Springer. 2016. p. 283–95.
20. Starlinger J, Cohen-Boulakia S, Khanna S, Davidson S, Leser U. Layer decomposition: an effective structure-based approach for scientific workflow similarity. In: IEEE eScience conference. 2014.
21. Medhata W, Hassan A, Korashy H. Sentiment analysis algorithms and applications: a survey. Ain Shams Eng J. 2014;5:1093–113.
22. 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. ACM. 2010.
23. Dijkman R, Dumas M, García-Bañuelos L. Graph matching algorithms for business process model similarity search. In: International conference on business process management. 2009. p. 48–63.
24. Garijo D, Alper P, Belhajjame K, Corcho O, Gil Y, Goble C. Common motifs in scientific workflows: an empirical analysis. Future Gen Comput Syst. 2014;36:338–51.
25. Garijo D, Corcho O, Gil Y. Detecting common scientific workflow fragments using templates and execution provenance. In: The proceedings of the seventh international conference on knowledge capture. 2013. p. 33–40.
26. Maguire E, Rocca-Serra P, Sansone SA, 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.
27. Polato M, Sperduti A, Burattin A, de Leoni M. Time and activity sequence prediction of business process instances. Computing. 2018. https://doi.org/10.1007/s00607-018-0593-x.
28. Qiu J, Wu Q, Ding G, Xu Y, Feng S. A survey of machine learning for big data processing. J Adv Signal Process. 2016;2016(1):67–83.
29. de Lén HP, Nardelli L, Carmona J, van den Broucke SKLM. Incorporating negative information to process discovery of complex systems. Inf Sci. 2018;422 (2018):480–96.
30. D Bollegala, Weir D, Carroll J. Cross-domain sentiment classification using a sentiment sensitive dictionary. IEEE Trans Knowl Data Eng. 2013;25(8):1719–31 ISSN 1041-4347.

31. Chen P, Plale BA. Big data provenance analysis and visualization. In: Cluster, cloud and grid computing (CCGrid). 15th IEEE/ACM international symposium. 2015. p. 797–800.
32. Pan SJ, Ni X, Sun JT, Yang Q, Chen Z. Cross-domain sentiment classification via spectral feature alignment. In: WWW. 2010;2010:26–30.
33. Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng. 2009;22(10):1345–59.
34. Schmidhuber J. Deep learning in neural networks: an overview. J Neural Netw. 2015;61:85–117.
35. Aggarwal CC. Data classification: algorithms and applications. Boca Raton: CRC Press; 2014.
36. Lee H, Battle A, Raina R, Ng AY. Efficient sparse coding algorithms. In: Advances in neural information processing systems. 2007. p. 801–8.
37. Hawkins D. Identification of outliers. London: Chapman & Hall Reading; 1980.
38. Blitzer J, Dredze M, Pereira F. Biographies, bollywood, boom-boxes and blenders: domain adaptation for sentiment classification. In: ACL. 2007;2007:440–7.
39. Verbeek HMW, Gunther CW. XES standard definition 2.0. Technical report. BPMcenter.org, July 2014. http://bpmcenter.org/wp-content/uploads/reports/2014/BPM-14-09.pdf. BPM Center Report BPM-14-09.
40. van Dongen BF. Real-life event logs - Hospital log. Eindhoven University of Technology. Dataset. 2011. https://doi.org/10.4121/uuid:d9769f3d-0ab0-4fb8-803b-0d1120ffcf54.
41. van Dongen BF. Bpi challenge. Eindhoven University of Technology. Dataset. 2012. https://doi.org/10.4121/uuid:3926db30-f712-4394-aebc-75976070e91f.
42. 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.
43. van Dongen BF. BPI Challenge 2017. Eindhoven University of Technology. Dataset. 2017. https://doi.org/10.4121/uuid:5f3067df-f10b-45da-b98b-86ae4c7a310b.
44. Deelman E, Peterka T, Altintas I, Carothers CD, van Dam KK, Moreland K, Parashar M, Ramakrishnan L, Taufer M, Vetter J. The future of scientific workflows. Int J High Perform Comput Appl. 2017;32(1):159–75.