

Unveiling Relations in the Industry 4.0 Standards Landscape based on Knowledge Graph Embeddings

Ariam Rivas¹, Irlán Grangel-González², Diego Collarana³, Jens Lehmann³,
Maria-Esther Vidal^{1,4}

¹ L3S, Leibniz Univ. of Hannover
ariam.rivas@tib.eu

² Robert Bosch Corporate Research GmbH, Renningen, Germany
irlan.grangelgonzalez@de.bosch.com

³ University of Bonn & Fraunhofer IAIS
diego.collarana.vargas@

{diego.collarana.vargas|jens.lehmann}@iais.fraunhofer.de

⁴ TIB Leibniz Information Centre for Science and Technology
maria.vidal@tib.eu

Abstract. Industry 4.0 (I4.0) standards and standardization frameworks have been proposed with the goal of *empowering interoperability* in smart factories. These standards enable the description and interaction of the main components, systems, and processes inside of a smart factory. Due to the growing number of frameworks and standards, there is an increasing need for approaches that automatically analyze the landscape of I4.0 standards. Standardization frameworks classify standards according to their functions into layers and dimensions. However, similar standards can be classified differently across the frameworks, producing, thus, interoperability conflicts among them. Semantic-based approaches that rely on ontologies and knowledge graphs, have been proposed to represent standards, known relations among them, as well as their classification according to existing frameworks. Albeit informative, the structured modeling of the I4.0 landscape only provides the foundations for detecting interoperability issues. Thus, graph-based analytical methods able to exploit knowledge encoded by these approaches, are required to uncover alignments among standards. We study the relatedness among standards and frameworks based on community analysis to discover knowledge that helps to cope with interoperability conflicts between standards. We use knowledge graph embeddings to automatically create these communities exploiting the meaning of the existing relationships. In particular, we focus on the identification of similar standards, i.e., communities of standards, and analyze their properties to detect unknown relations. We empirically evaluate our approach on a knowledge graph of I4.0 standards using the Trans* family of embedding models for knowledge graph entities. Our results are promising and suggest that relations among standards can be detected accurately.

1 Introduction

The international community recognizes Industry 4.0 (I4.0) as the fourth industrial revolution. The main objective of I4.0 is the creation of *Smart Factories* by combining the

Internet of Things (IoT), Internet of Services (IoS), and Cyber-Physical Systems (CPS). In smart factories, humans, machines, materials, and CPS need to communicate intelligently in order to produce individualized products. To tackle the problem of interoperability, different industrial communities have created standardization frameworks. Relevant examples are the Reference Architecture for Industry 4.0 (RAMI4.0) [1] or the Industrial Internet Connectivity Framework (IICF) in the US [17]. Standardization frameworks classify, and align industrial standards according to their functions. While being expressive to categorize existing standards, standardization frameworks may present divergent interpretations of the same standard. Mismatches among standard classifications generate semantic interoperability conflicts that negatively impact on the effectiveness of communication in smart factories.

Database and Semantic web communities have extensively studied the problem of data integration [9,15,21], and various approaches have been proposed to support data-driven pipelines to transform industrial data into actionable knowledge in smart factories [13,23]. Ontology-based approaches have also contributed to create a shared understanding of the domain [16], and specifically Kovalenko and Euzenat [15] have equipped data integration with diverse methods for ontology alignment. Furthermore, Lin *et al.* [18] identify interoperability conflicts across domain specific standards (e.g., RAMI4.0 model and the IICF architecture), while works by Grangel-Gonzalez *et al.* [10,11,12] show the relevant role that Descriptive Logic, Datalog, and Probabilistic Soft Logic play in liaising I4.0 standards. Certainly, the extensive literature in data integration provides the foundations for enabling the semantic description and alignment of "similar" things in a smart factory. Nevertheless, finding alignments across I4.0 requires the encoding of domain specific knowledge represented in standards of diverse nature and standardization frameworks defined with different industrial goals. We rely on state-of-the-art knowledge representation and discovery approaches to embed meaningful associations and features of the I4.0 landscape, to enable interoperability.

We propose a knowledge-driven approach first to represent standards, known relations among them, as well as their classification according to existing frameworks. Then, we utilize the represented relations to build a latent representation of standards, i.e., embeddings. Values of similarity metrics between embeddings are used in conjunction with state-of-the-art community detection algorithms to identify patterns among standards. Our approach determines relatedness among standards by computing communities of standards and analyzing their properties to detect unknown relations. Finally, the *homophily* prediction principle is performed in each community to discover new links between standards and frameworks. We assess the performance of the proposed approach in a data set of 249 I4.0 standards connected by 736 relations extracted from the literature. The observed results suggest that encoding knowledge enables for the discovery of meaningful associations. Our contributions are as follows:

1. We formalize the problem of finding relations among I4.0 standards and present *I4.0RD*, a knowledge-driven approach to unveil these relations. *I4.0RD* exploits the semantic description encoded in a knowledge graph via the creation of embeddings, to identify then communities of standards that should be related.

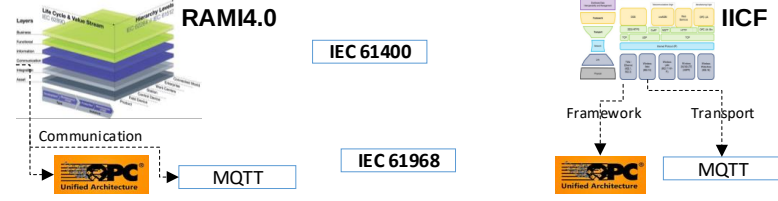


Fig. 1: **Motivating Example.** The RAMI4.0 and IICF standardization frameworks are developed for diverse industrial goals; they classify standards in layers according to their functions, e.g., OPC UA and MQTT under the communication layer in RAMI4.0, and OPC UA and MQTT in the framework and transport layers in IICF, respectively. Further, some standards, e.g., IEC 61400 and IEC 61968, are not classified yet.

2. We evaluate the performance of *I4.0RD* in different embeddings learning models and community detection algorithms. The evaluation material is available ⁵.

The rest of this paper is organized as follows: Section 2 illustrates the interoperability problem presented in this paper. Section 3 presents the proposed approach, while the architecture of the proposed solution is explained in Section 4. Results of the empirical evaluation of our methods are reported in Section 5 while Section 6 summarizes the state of the art. Finally, we close with the conclusion and future work in section 7.

2 Motivating Example

Existing efforts to achieve interoperability in I4.0, mainly focus on the definition of standardization frameworks. A standardization framework defines different layers to group related I4.0 standards based on their functions and main characteristics. Typically, classifying existing standards in a certain layer is not a trivial task and it is influenced by the point of view of the community that developed the framework. RAMI4.0 and IICF are exemplar frameworks, the former is developed in Germany while the latter in the US; they meet specific I4.0 requirements of certain locations around the globe. RAMI4.0 classifies the standards OPC UA and MQTT into the Communication layer, stating this, that both standards are similar. Contrary, IICF presents OPC UA and MQTT at distinct layers, i.e., the framework and the transport layers, respectively. Furthermore, independently of the classification of the standards made by standardization frameworks, standards have relations based on their functions. Therefore, IEC 61400 and IEC 61968 that are usually utilized to describe electrical features, are not classified at all. Figure 1 depicts these relations across the frameworks RAMI4.0 and IICF, and the standards; it illustrates interoperability issues in the I4.0 landscape.

Existing data integration approaches rely on the description of the characteristics of entities to solve interoperability by discovering alignments among them. Specifically, in the context of I4.0, semantic-based approaches have been proposed to represent standards, known relations among them, as well as their classification according to existing

⁵ <https://github.com/i40-Tools/I40KG-Embeddings>

frameworks [4,6,18,19]. Despite informative, the structured modeling of the I4.0 landscape only provides the foundations for detecting interoperability issues.

We propose *I4.0RD*, an approach capable of discovering relation over I4.0 knowledge graphs to identify unknown relations among standards. Our proposed methods exploit relations represented in an I4.0 knowledge graph to compute the similarity of the modeled standards. Then, an unsupervised graph partitioning method determines the communities of standards that are similar. Moreover, *I4.0RD* explores communities to identify possible relations of standards, enhancing, thus, interoperability.

3 Problem Definition and Proposed Solution

We tackle the problem of unveiling relations between I4.0 standards. We assume that the relations among standards and standardization frameworks like the ones shown in Figure 2(a), are represented in a knowledge graph named I4.0KG. Nodes in a I4.0KG correspond to standards and frameworks; edges represent relations among standards, as well as the standards grouped in a framework layer. An I4.0KG is defined as follows:

Given sets V_e and V_t of entities and types, respectively, a set E of labelled edges representing relations, and a set L of labels. An I4.0KG is defined as $\mathcal{G} = (V_e \cup V_t, E, L)$:

- The types Standard, Frameworks, and Framework Layer belong to V_t .
- I4.0 standards, frameworks, and layers are represented as instances of V_e .
- The types of the entities in V_e are represented as edges in E that belong to $V_e \times V_t$.
- Edges in E that belong to $V_e \times V_e$ represent relations between standards and their classifications into layers according to a framework.
- *RelatedTo*, *Type*, *classifiedAs*, *IsLayerOf* correspond to labels in L that represent the relations between standards, their type, their classification into layers, and the layers of a framework, respectively.

3.1 Problem Statement

Let $\mathcal{G}' = (V_e \cup V_t, E', P)$ and $\mathcal{G} = (V_e \cup V_t, E, P)$ be two I4.0 knowledge graphs. $\mathcal{G}'$ is an *ideal* knowledge graph that contains all the *existing relations* between standard entities and frameworks in V_e , i.e., an oracle that knows whether two standard entities are related or not, and to which layer they should belong; Figure 2 (b) illustrates a portion of an ideal I4.0KG, where the relations between standards are explicitly represented. $\mathcal{G} = (V_e \cup V_t, E, P)$ is an *actual* I4.0KG, which only contains a portion of the relations represented in $\mathcal{G}'$, i.e., $E \subseteq E'$; it represents those relations that are known and is not necessarily complete. Let $\Delta(E', E) = E' - E$ be the set of relations existing in the ideal knowledge graph $\mathcal{G}'$ that are not represented in $\mathcal{G}$. Let $\mathcal{G}_{\text{comp}} = (V_e \cup V_t, E_{\text{comp}}, P)$ be a *complete* knowledge graph, which includes a relation for each possible combination of elements in V_e and labels in L , i.e., $E \subseteq E' \subseteq E_{\text{comp}}$. Given a relation $e \in \Delta(E_{\text{comp}}, E)$, the problem of discovering relations consists of determining whether $e \in E'$, i.e., if a relation represented by an edge $r = (e_i \text{ } p \text{ } e_j)$ corresponds to an existing relation in the ideal knowledge graph $\mathcal{G}'$. Specifically, we focus on the problem of discovering *relations* between standards in $\mathcal{G} = (V_e \cup V_t, E, P)$. We are interested in

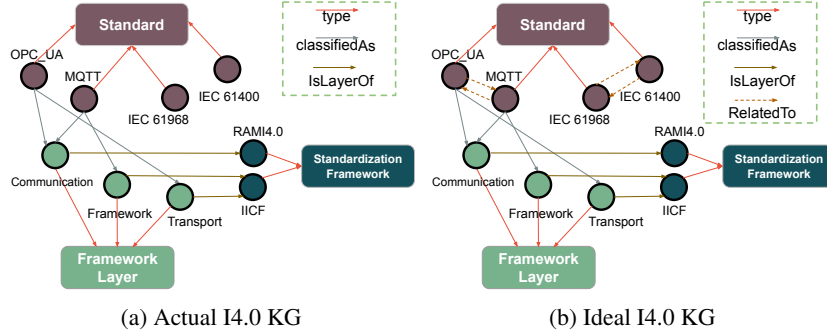


Fig. 2: **Example of I4.0KGs.** Figure 2a shows known relationships among standards to Framework Layer and Standardization Framework. While Figure 2b depicts all the ideal relationships between the standards expressed with the property `relatedTo`. Standards OPC UA and MQTT are related, as well as the standards IEC 61968 and IEC 61400. Our aim is discovering relations `relatedTo` in Figure 2b.

finding the maximal set of relationships or edges E_a that belong to the ideal I4.0KG, i.e., find a set E_a that corresponds to a solution of the following optimization problem:

$$\operatorname{argmax}_{E_a \subseteq E_{comp}} |E_a \cap E'|$$

Considering the knowledge graphs depicted in Figures 2 (a) and (b), the problem addressed in this work corresponds to the identification of edges in the ideal knowledge graph that correspond to unknown relations between standards.

3.2 Proposed Solution

We propose a relation discovery method over I4.0KGs to identify unknown relations among standards. Our proposed method exploits relations represented in an I4.0KG to compute similarity values between the modeled standards. Further, an unsupervised graph partitioning method determine the parts of the I4.0KG or communities of standards that are similar. Then, the *homophily* prediction principle is applied in a way that similar standards in a community are considered to be related.

4 The $I4.0\mathcal{RD}$ Architecture

Figure 3 presents $I4.0\mathcal{RD}$, a pipeline that implements the proposed approach. $I4.0\mathcal{RD}$ receives an I4.0KG $\mathcal{G}$, and returns an I4.0KG $\mathcal{G}'$ that corresponds to a solution of the problem of discovering relations between standards. First, in order to compute the values of similarity between the entities an I4.0KG, $I4.0\mathcal{RD}$ learns a latent representation of the standards in a high-dimensional space. Our approach resorts to the Trans* family of models to compute the embeddings of the standards and the cosine similarity

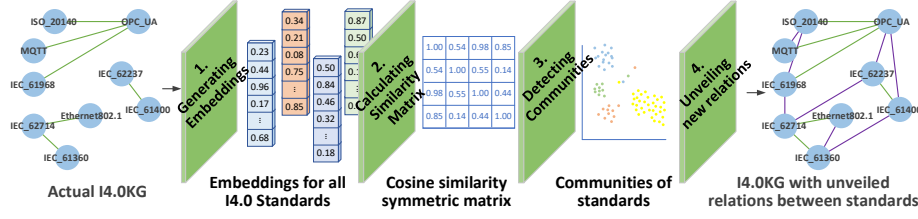


Fig. 3: **Architecture.** *I4.ORD* receives an I4.0KG and outputs an extended version of the I4.0KG including novel relations. Embeddings for each standard are created using the Trans* family of models, and similarity values between embeddings are computed; these values are used to partition standards into communities. Finally, the homophily prediction principle is applied to each community to discover unknown relations.

measure to compute the values of similarity. Next, community detection algorithms are applied to identify communities of related standards. METIS [14], KMeans [3], and SemEP [24] are methods included in the pipeline to produce different communities of standards. Finally, *I4.ORD* applies the *homophily* principle to each community to predict relations or alignments among standards.

4.1 Learning Latent Representations of Standards

I4.ORD utilizes the Trans* family of models to compute latent representations, e.g., vectors, of entities and relations in an I4.0 knowledge graph. In particular, *I4.ORD* utilizes TransE, TransD, TransH, and TransR. These models differ on the representation of the embeddings for the entities and relations (Wang et al. [26]). Suppose e_i , e_j , and p , denote the vectorial representation of two entities related by the labeled edge p in an I4.0 knowledge graph. Furthermore, $\|x\|_2$ represents the Euclidean norm.

TransE, TransH, and TransR represent the entity embeddings as $(e_i, e_j \in \mathbb{R}^d)$, while TransD characterizes the entity embeddings as: $(e_i, w_{e_i} \in \mathbb{R}^d - e_i, w_{e_j} \in \mathbb{R}^d)$. As a consequence of different embedding representations, the scoring function also varies. For example, TransE is defined in terms of the score function $\|e_i + p - e_j\|_2^2$, while $\|M_p e_i + p - M_p e_j\|_2^2$ defines TransR⁶. Furthermore, TransH score function corresponds to $\|e_{i\perp} + d_p - e_{j\perp}\|_2^2$, where the variables $e_{i\perp}$ and $e_{j\perp}$ denote a projection to the hyperplane w_p of the labeled relation p , and d_p is the vector of a relation-specific translation in the hyperplane w_p . To learn the embeddings, *I4.ORD* resorts to the PyKeen (Python KnowlEdge EmbeddINgs) framework [2]. As hyperparameters for the models of the Trans* family, we use the ones specified in the original papers of the models. The hyperparameters include embedding dimension (set to 50), number of epochs (set to 500), batch size (set to 64), seed (set to 0), learning rate (set to 0.01), scoring function (set to 1 for TransE, and 2 for the rest), margin loss (set to 1 for TransE and 0.05 for the rest). All the configuration classes and hyperparameters are open in GitHub⁷.

⁶ M_p corresponds to a projection matrix $M_p \in \mathbb{R}^{d \times k}$ that projects entities from the entity space to the relation space; further $p \in \mathbb{R}^k$.

⁷ <https://github.com/i40-Tools/I4.0KG-Embeddings>

4.2 Computing Similarity Values Between Standards

Once the algorithm–Trans* family–that computes the embeddings reaches a termination condition, e.g., the maximum number of epochs, the I4.0KG embeddings are learned. As the next step, *I4.0RD* calculates a *similarity symmetric matrix* between the embeddings that represent the I4.0 standards. Any distance metric for vector spaces can be utilized to calculate this value. However, as a proof of concepts, *I4.0RD* applies the Cosine Distance. Let u be an embedding of the Standard-A and v an embedding of the Standard-B, the similarity score, between both standards, is defined as follows:

$$\text{cosine}(u, v) = 1 - \frac{u \cdot v}{\|u\|_2 \|v\|_2}$$

After building the *similarity symmetric matrix*, *I4.0RD* applies a threshold to restrict the similarity values. *I4.0RD* relies on percentiles to calculate the value of such a threshold. Further, *I4.0RD* utilizes the function Kernel Density Estimation (KDE) to compute the probability density of the cosine similarity matrix; it sets to zero the similarity values lower than the given threshold.

4.3 Detecting Communities of Standards

I4.0RD maps the problem of computing groups of potentially related standards to the problem of community detection. Once the embeddings are learned, the standards are represented in a vectorial way according to their functions preserving their semantic characteristics. Using the embeddings, *I4.0RD* computes the similarity between the I4.0 standards as mentioned in the previous section. The values of similarity between standards are utilized to partition the set of standards in a way that standards in a community are highly similar but dissimilar to the standards in other communities. As proof of concept, three state-of-the-art community detection algorithms have been used in *I4.0RD*: SemEP, METIS, and KMeans. They implement diverse strategies for partitioning a set based on the values of similarity, and our goal is to evaluate which of the three is more suitable to identify meaningful connections between standards.

4.4 Discovering Relations Between Standards

New relations between standards are discovered in this step; the *homophily* prediction principle is applied over each of the communities and all the standards in a community are assumed to be related. Figure 4 depicts an example where new relations are computed from two communities; unknown relations correspond to connections between standards in a community that did not existing in the input I4.0KG.

5 Empirical Evaluation

We report on the impact that the knowledge encoded in I4.0 knowledge graphs has in the behavior of *I4.0RD*. In particular, we assess the following research questions:

RQ1) Can the semantics encoded in I4.0KG empower the accuracy of the relatedness between entities in a KG?

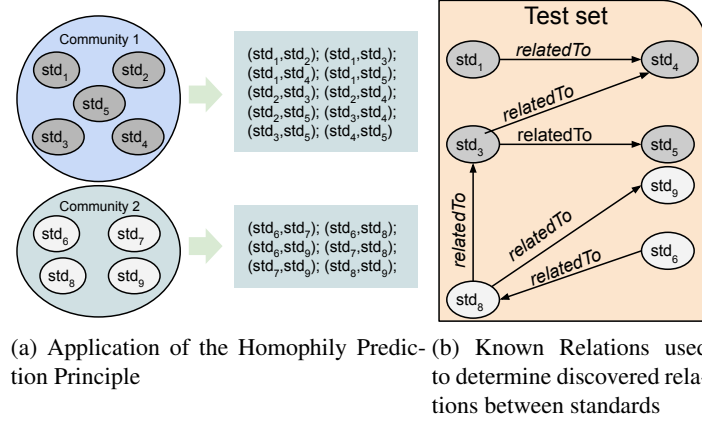


Fig. 4: Discovering Relations Between Standards. (a) The homophily prediction principle is applied on two communities, as a result, 16 relations between standards are found. (b) Six out of the 16 found relations correspond to meaningful relations.

RQ2) Does a semantic community based analysis on I4.0KG allow for improving the quality of predicting new relations on the I4.0 standards landscape?

Experiment Setup: We considered four embedding algorithms to build the standards embedding. Each of these algorithms was evaluated independently. Next, a similarity matrix for the standards embedding was computed. The similarity matrix is required for applying the community detection algorithms. In our experiments, three algorithms were used to compute the communities. That means twelve combinations between embedding algorithms and community detection algorithms to be evaluated. To assure statistical robustness, we executed 5-folds cross-validation with one run.

Thresholds for Computing Values of Similarity Figure 5 depicts the probability density function of each fold for each embedding algorithm. Figures 5a and 5b show the values of the folds of TransD and TransE where all the similarity values are close to 0.0, i.e., all the standards are different. Figure 5d suggests that all the folds have similar behavior with values between 0.0 and 0.5. Figure 5c shows a group of standards similar with values close to 1.0 and the rest of the standards between 0.0 and 0.6. The percentile of the similarity matrix is computed with a threshold of 0.85. That means all values of the similarity matrix which are less than the percentile computed, are filled with 0.0 and then, these two standards are dissimilar. After analyzing the probability density of each fold (cf. Figure 5), the thresholds of TransH and TransR are set to 0.50 and 0.75, respectively. The reason is because the two cases with a high threshold find all similar standards. In the case of TransH, there is a high density of values close to 1.0; it indicates that for a threshold of 0.85, the percentile computed is almost 1.0. the values of the similarity matrix less than the threshold are filled with 0.0; values of 0.0 represent that the compared standards are not similar.

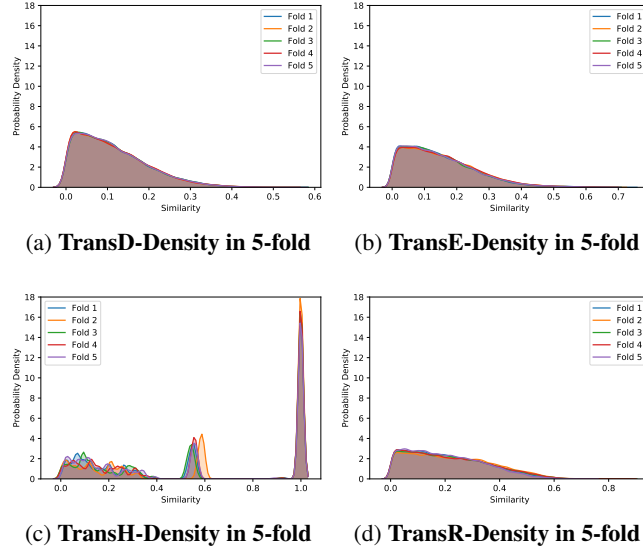


Fig. 5: Probability density of each fold per Trans* methods. Figures 5a, 5b, and 5d show that all folds have values close to zero, i.e., with embeddings created by TransD, TransE, and TransR the standards are very different from each other. However, TransH (cf. Figure 5c), exploits properties of the standards and generates embeddings with a different distribution of similarity, i.e., values between 0.0 and 0.6, as well as values close to 1.0. According to known characteristics of the I4 standards, the TransH distribution of similarity better represents their relatedness.

Metrics: the following metrics are used to estimate the quality of the communities from the I4.0KG embeddings.

- a) **Conductance (InvC):** measures relatedness of entities in a community, and how different they are to entities outside the community [7]. The inverse of Conductance is reported: $1 - \text{Conductance}(K)$, where $K = \{k_1, k_2, \dots, k_n\}$ the set of standards communities obtained by the cluster algorithm, and k_i are the computed clusters.
- b) **Performance (P):** sums up the number of intra-community relationships, plus the number of non-existent relationships between communities [7].
- c) **Total Cut (InvTC):** sums up all similarities among entities in different communities [5]. The Total Cut values are normalized by dividing the sum of the similarities between the entities. The inverse of Total Cut is reported as follows: $1 - \text{NormTotalCut}(K)$
- d) **Modularity (M):** is the value of the intra-community similarities between the entities divided by the sum of all the similarities between the entities, minus the sum of the similarities among the entities in different communities, in case they are randomly distributed in the communities [22]. The value of the Modularity is in the range of $[-0.5, 1]$, which can be scaled to $[0, 1]$ by computing: $\frac{\text{Modularity}(K) + 0.5}{1.5}$.

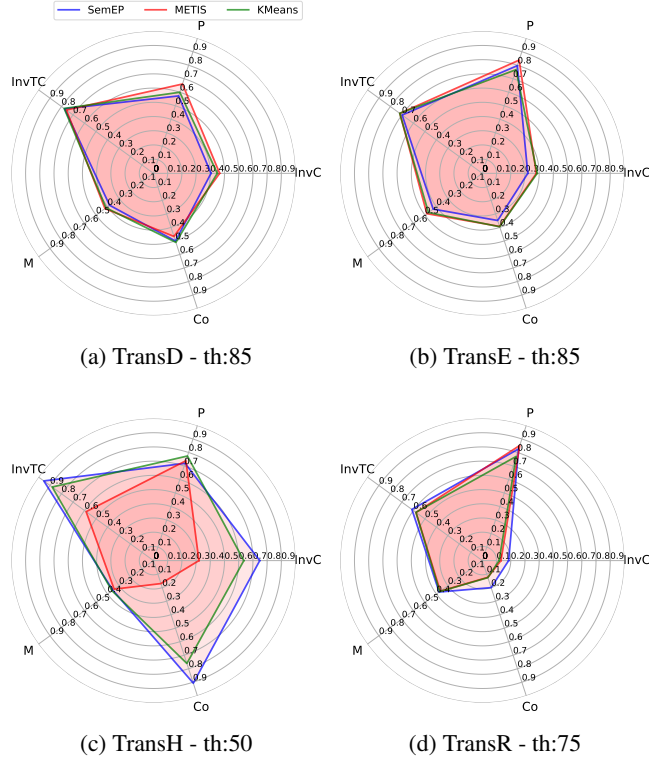


Fig. 6: **Quality of the generated communities.** Communities evaluated in terms of prediction metrics with thresholds (th) of 0.85, 0.50, and 0.75 using the SemEP, METIS, and KMeans algorithms. In this case higher values are better. Our approach exhibits the best performance with TransH embeddings and a threshold of 0.50 for computing the similarity matrix, i.e., Figure (c). SemEP achieves the highest values in four of the five evaluated parameters.

- e) **Coverage (Co):** compares the fraction of intra-community similarities between entities to the sum of all similarities between entities [7].

Implementation: Our proposed approach is implemented in Python 2.7 and integrated with the PyKeen (Python KnowlEdge EmbeddiNgs) framework [2], METIS 5.1⁸, SemEP⁹, and Kmeans¹⁰. The experiments were executed on a GPU server with ten chips Intel(R) Xeon(R) CPU E5-2660, two chips GeForce GTX 108, and 100 GB RAM.

⁸ <http://glaros.dtc.umn.edu/gkhome/metis/metis/download>

⁹ <https://github.com/SDM-TIB/semEP>

¹⁰ <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html>

RQ1 - *Corroborating the accuracy of relatedness between standards in I40KG.* To compute accuracy of *I4.0RD*, we executed a five-folds cross-validation procedure. To that end, the data set is divided into five consecutive folds shuffling the data before splitting into folds. Each fold is used once as validation, i.e., test set while the remaining fourth folds form the training set. Figure 6 depicts the best results are obtained with the combination of the TransH and SemEP algorithms. The values obtained for this combination are as follows: **Inv. Conductance** (0.75), **Performance** (0.77), **Inv. Total Cut** (0.95), **Modularity** (0.36), and **Coverage** (0.91).

RQ2 - *Predicting new relations between standards.* In order to assess the second research question, the data set is divided into five consecutive folds. Each fold comprises 20% of the relationships between standards. Next, the precision measurement is applied to evaluate the main objective is to unveil uncovered associations and at the same time to corroborate knowledge patterns that are already known. As shown in Figure 7, the best results for the property `relatedTo` are achieved by TransH embeddings in combination with the SemEP and KMeans algorithm.

The communities of standards discovered using the techniques TransH and SemEP contribute to the resolution of interoperability in I4.0 standards. To provide an example of this, we observed a resulting cluster with the standards *ISO 15531* and *MTConnect*. The former provides an information model for describing manufacturing data. The latter offers a vocabulary for manufacturing equipment. It is important to note that those standards are not related to the training set nor in I40KG. The membership of both standards in the cluster means that those two standards should be classified together in the standardization frameworks. Besides, it also suggests to the creators of the standards that they might look after possible existing synergies between them. This example suggests that the techniques employed in this work are capable of discovering new communities of standards. These communities can be used to improve the classification that the standardization frameworks provide for the standards.

5.1 Discussion

The techniques proposed in this paper rely on known relations between I4.0 standards to discover novel patterns and new relations. During the experimental study, we can observe that these techniques could group together not only standards that were known to be related, but also standards whose relatedness was implicitly represented in the I40KG. This feature facilitates the detection of high-quality communities as reported in Figure 6, as well as for an accurate discovery of relations between standards (cf. Figure 7). As observed, the accuracy of the approach can be benefited from the application of state-of-the-art algorithms of the Trans* family, e.g., TransH. Additionally, the strategy employed by SemEP that allows for positioning in the same communities highly similar standards, leads our approach into high-quality discoveries. The combination of both techniques TransH and SemEP allows discovering communities with high quality.

To understand why the combination of TransH and SemEP produces the best results, we analyze in detail both techniques. TransH introduces the mechanism of projecting the relation to a specific hyperplane [27], enabling, thus, the representation of relations with cardinality many to many. Since the materialization of transitivity and symmetry of

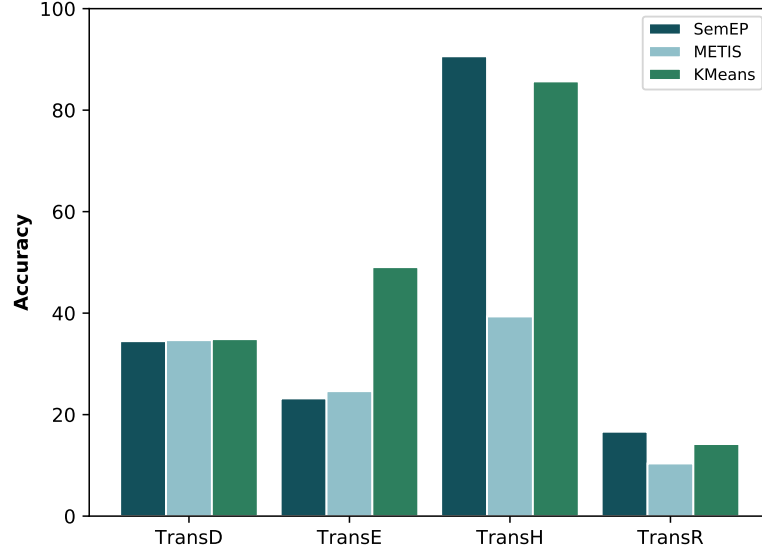


Fig. 7: **I4.0RD accuracy**. Percentage of the test set for the property `relatedTo` is achieved in each cluster. Our approach exhibits the best performance using TransH embedding and with the SemEP algorithm reaching an accuracy by up to 90%.

the property `relatedTo` corresponds to many to many relations, the instances of this materialization are taken into account during the generation of the embeddings, specifically, during the translating operation on a hyperplane. Thus, even though semantics is not explicitly utilized during the computation of the embeddings, considering different types of relations, empowers the embeddings generated by TransH. Moreover, it allows for a more precise encoding of the standards represented in I4.0KG. Figure 5c illustrates groups of standards in the similarity intervals $[0.9, 1.0]$, $[0.5, 0.6]$, and $[0.0, 0.4]$. The SemEP algorithm can detect these similarities and represent them in high-precision communities. The other three models embeddings TransD, TransE, and TransR do not represent the standards in the best way. Figures 5a, 5b, 5d report that several standards are in the similarity interval $[0.0, 0.3]$. This means that no community detection algorithm could be able to discover communities with high quality. Reported results indicate that the presented approach enables – in average – for discovering communities of standards by up to 90%. Although these results required the validation of experts in the domain, an initial evaluation suggest that the results are accurate.

6 Related Work

In the literature, different approaches are proposed for discovering communities of standards as well as to corroborate and extend the knowledge of the standardization frameworks. Zeid *et al.* [28] study different approach to achieve interoperability of different standardization frameworks. In this work, the current landscape for smart manufacturing is described by highlighting the existing standardization frameworks in different

regions of the globe. Lin *et al.* [18] present similarities and differences between the RAMI4.0 model and the IIRA architecture. Based on the study of these similarities and differences authors proposed a functional alignment among layers in RAMI4.0 with the functional domains and crosscutting functions in IIRA. Monteiro *et al.* [20] further report on the comparison of the RAMI4.0 and IIRA frameworks. In this work, a cooperation model is presented to align both standardization frameworks. Furthermore, mappings between RAMI4.0 IT Layers and the IIRA functional domain are established. Another related approach is that outlined in [25]. Moreover, the IIRA and RAMI4.0 frameworks are compared based on different features, e.g., country of origin, source organization, basic characteristics, application scope, and structure. It further details where correspondences exist between the IIRA viewpoints and RAMI4.0 layers. Garofalo *et al.* [8] outline KGEs for I4.0 use cases. Existing techniques for generating embeddings on top of knowledge graphs are examined. Further, the analysis of how these techniques can be applied to the I4.0 domain is described; specifically, it identifies the predictive maintenance, quality control, and context-aware robots as the most promising areas to apply the combination of KGs with embeddings. All the approaches mentioned above are limited to describe and characterize existing knowledge in the domain. However, in our view, two directions need to be considered to enhance the knowledge in the domain; 1) the use of a KG based approach to encode the semantics; and 2) the use of machine learning techniques to discover and predict new communities of standards based on their relations.

7 Conclusion

In this paper, we presented the *I4.0RD* approach that combines knowledge graphs and embeddings to discover associations between I4.0 standards. Our approach resorts to I4.0KG to discover relations between standards; I4.0KG represents relations between standards extracted from the literature or defined according to the classifications stated by the standardization frameworks. Since the relation between standards is symmetric and transitive, the transitive closure of the relations is materialized in I4.0KG. Different algorithms for generating embeddings are applied on the standards according to the relations represented in I4.0KG. We employed three community detection algorithms, i.e., SemEP, METIS, and KMeans to identify similar standards, i.e., communities of standards, as well as to analyze their properties. Additionally, by applying the homophily prediction principle, novel relations between standards are discovered. We empirically evaluated the quality of the proposed techniques over 249 standards, initially related through 736 instances of the property `relatedTo`; as this relation is symmetric and transitive, its transitive closure is also represented in I4.0KG with 22,969 instances of `relatedTo`. The Trans* family of embedding models were used to identify a low-dimensional representation of the standards according to the materialized instances of `relatedTo`. Results of a 5-fold cross validation process suggest that our approach is able to effectively identify novel relations between standards. Thus, our work broadens the repertoire of knowledge-driven frameworks for understanding I4.0 standards, and we hope that our outcomes facilitate the resolution of the existing interoperability issues in the I4.0 landscape. As for the future work, we envision to have a more fine-

grained description of the I4.0 standards, and evaluate hybrid-embeddings and other type of community detection methods.

References

1. Adolphs, P., Auer, S., Billmann, M., Hankel, M., Heidel, R., Hoffmeister, M., Huhle, H., Jochem, M., Kiele, M., Koschnick, G., Koziol, H., Linke, L., Pichler, R., Schewe, F., Schneider, K., Waser, B.: Structure of the Administration Shell. Status report, ZVEI and VDI (2016)
2. Ali, M., Jabeen, H., Hoyt, C.T., Lehmann, J.: The keen universe: An ecosystem for knowledge graph embeddings with a focus on reproducibility and transferability. (in press)
3. Arthur, D., Vassilvitskii, S.: K-means++: The advantages of careful seeding. In: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms. SODA '07, Philadelphia, PA, USA, Society for Industrial and Applied Mathematics (2007) 1027–1035
4. Bader, S.R., Grangel-González, I., Tasnim, M., Lohmann, S.: Structuring the industry 4.0 landscape. In: 24th IEEE International Conference on Emerging Technologies and Factory Automation, ETFA, Zaragoza, Spain, September 10–13. (2019) 224–231
5. Buluç, A., Meyerhenke, H., Safro, I., Sanders, P., Schulz, C.: Recent advances in graph partitioning. In: Algorithm Engineering - Selected Results and Surveys. (2016) 117–158
6. Chungoora, N., Cutting-Decelle, A.F., Young, R., Gunendran, G., Usman, Z., Harding, J.A., Case, K.: Towards the ontology-based consolidation of production-centric standards. International Journal of Production Research **51**(2) (2013) 327–345
7. Gaertler, M., Erlebach, T. In: Clustering. Springer Berlin Heidelberg, Berlin, Heidelberg (2005) 178–215
8. Garofalo, M., Pellegrino, M.A., Altabba, A., Cochez, M.: Leveraging knowledge graph embedding techniques for industry 4.0 use cases. CoRR **abs/1808.00434** (2018)
9. Golshan, B., Halevy, A.Y., Mihaila, G.A., Tan, W.: Data integration: After the teenage years. In: Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, PODS 2017, Chicago, IL, USA, May 14–19, 2017. (2017) 101–106
10. Grangel-González, I., Baptista, P., Halilaj, L., Lohmann, S., Vidal, M., Mader, C., Auer, S.: The industry 4.0 standards landscape from a semantic integration perspective. In: 22nd IEEE International Conference on Emerging Technologies and Factory Automation, ETFA, Limassol, Cyprus, September 12–15. (2017) 1–8
11. Grangel-González, I., Collarana, D., Halilaj, L., Lohmann, S., Lange, C., Vidal, M.E., Auer, S.: Alligator: A deductive approach for the integration of industry 4.0 standards. In: 20th Int. Conf. on Knowledge Engineering and Knowledge Management, EKAW. (2016) 272–287
12. Grangel-González, I., Halilaj, L., Vidal, M., Rana, O., Lohmann, S., Auer, S., Müller, A.W.: Knowledge graphs for semantically integrating cyber-physical systems. In: Database and Expert Systems Applications - 29th International Conference, DEXA, Regensburg, Germany, September 3–6, Proceedings, Part I. (2018) 184–199
13. Hodges, J., García, K., Ray, S.: Semantic Development and Integration of Standards for Adoption and Interoperability. IEEE Computer **50**(11) (2017) 26–36
14. Karypis, G., Kumar, V.: A fast and high quality multilevel scheme for partitioning irregular graphs. SIAM J. Sci. Comput. **20**(1) (December 1998) 359–392
15. Kovalenko, O., Euzénat, J.: Semantic matching of engineering data structures. In: Semantic Web for Intelligent Engineering Applications, Springer (2016)
16. Lelli, F.: Interoperability of the time of industry 4.0 and the internet of things. Future Internet **11**(2) (2019) 36

17. Lin, S.W., Miller, B., Durand, J., Bleakley, G., Chigani, A., Martin, R., Murphy, B., Crawford, M.: The Industrial Internet of Things Volume G1: Reference Architecture. White Paper IIC:PUB:G1:V1.80:20170131, Industrial Internet Consortium (2017)
18. Lin, S.W., Murphy, B., Clauer, E., Loewen, U., Neubert, R., Bachmann, G., Pai, M., Hankel, M.: Reference Architectural Model Industrie 4.0 (RAMI 4.0). Technical report, Industrial Internet Consortium and Plattform Industrie 4.0 (2017)
19. Lu, Y., Morris, K.C., Frechette, S.: Standards landscape and directions for smart manufacturing systems. In: IEEE International Conference on Automation Science and Engineering, CASE, Gothenburg, Sweden, August 24-28. (2015) 998–1005
20. Monteiro, P., Carvalho, M., Morais, F., Melo, M., Machado, R., Pereira, F.: Adoption of architecture reference models for industrial information management systems. In: Int. Conf. on Intelligent Systems (IS), IEEE (2018) 763–770
21. Mountantonakis, M., Tzitzikas, Y.: Large-scale semantic integration of linked data: A survey. *ACM Comput. Surv.* **52**(5) (September 2019) 103:1–103:40
22. Newman, M.E.J.: Modularity and community structure in networks. *Proceedings of the National Academy of Sciences* **103**(23) (2006) 8577–8582
23. O'Donovan, P., L.K.B.K.: An industrial big data pipeline for data-driven analytics maintenance applications in large-scale smart manufacturing facilities. *Journal of Big Data* **2** **25** (2015)
24. Palma, G., Vidal, M.E., Raschid, L.: Drug-target interaction prediction using semantic similarity and edge partitioning. In: Proc. of the 13th Int. Semantic Web Conf. - Part I. ISWC '14, NY, USA, Springer-Verlag New York, Inc. (2014) 131–146
25. Velasquez, N., Estevez, E., Pesado, P.: Cloud computing, big data and the industry 4.0 reference architectures. *Journal of Computer Science and Technology* **18**(03) (2018) e29–e29
26. Wang, Q., Mao, Z., Wang, B., Guo, L.: Knowledge graph embedding: A survey of approaches and applications. *IEEE Trans. Knowl. Data Eng.* **29**(12) (2017) 2724–2743
27. Wang, Z., Zhang, J., Feng, J., Chen, Z.: Knowledge graph embedding by translating on hyperplanes. In: AAAI. (2014)
28. Zeid, A., Sundaram, S., Moghaddam, M., Kamarthi, S., Marion, T.: Interoperability in smart manufacturing: Research challenges. *Machines* **7**(2) (2019) 21