

Universal Preprocessing Operators for Embedding Knowledge Graphs with Literals

Patryk Preisner¹, Heiko Paulheim^{1,*}

¹Data and Web Science Group, University of Mannheim, Germany

Abstract

Knowledge graph embeddings are dense numerical representations of entities in a knowledge graph (KG). While the majority of approaches concentrate only on relational information, i.e., relations between entities, fewer approaches exist which also take information about literal values (e.g., textual descriptions or numerical information) into account. Those which exist are typically tailored towards a particular modality of literal and a particular embedding method. In this paper, we propose a set of universal preprocessing operators which can be used to transform KGs with literals for numerical, temporal, textual, and image information, so that the transformed KGs can be embedded with any method. The results on the kgbench dataset with three different embedding methods show promising results.

Keywords

Knowledge Graph, Embedding, Representation, Literal Information

1. Introduction

Knowledge graphs have become a common means to represent information across various domains. [1, 2] They are comprised of entities and their relations, but many also contain literal information, like textual descriptions of entities, numerical values, or even images. For example, the following is an excerpt of the representation of the entity *Mannheim* in DBpedia [3]:

```
dbr:Mannheim dbo:country dbr:Germany .
dbr:University_of_Mannheim dbp:city dbr:Mannheim .
dbr:Mannheim dbo:populationMetro "2362046"^^xsd:nonNegativeInteger .
dbr:Mannheim dbo:foundingDate "1607-01-24"^^xsd:date .
dbr:Mannheim dbo:abstract "Mannheim [...] officially the University City of
Mannheim (German: Universitätsstadt Mannheim), is the second-largest city
in the German state of Baden-Württemberg..."@en .
dbr:Mannheim foaf:depiction
<http://commons.wikimedia.org/wiki/Special:FilePath/
NUB_Mannheim_2014-03-13.jpg> .
```

Deep Learning for Knowledge Graphs (DL4KG) 2023


*Corresponding author.

✉ patryk.konrad.preisner@students.uni-mannheim.de (P. Preisner); heiko@informatik.uni-mannheim.de (H. Paulheim)

🌐 <http://www.heikopaulheim.com/> (H. Paulheim)

🆔 0000-0001-6950-4459 (P. Preisner); 0000-0003-4386-8195 (H. Paulheim)

© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

Most embedding approaches only consider relations between entities when computing numeric representations for entities. In the above example, when learning a representation for the entity *Mannheim*, they would use only the first two statements, but neglect the latter three, containing textual, numerical, and image information. However, those also contain relevant information about the entity, which could lead to a better latent representation if they were used by the embedding approach.

While a few embedding approaches have been proposed which take into account literal information, they have a few shortcomings: most of them (1) target only one modality (e.g., text, numbers, or images), and (2) are adaptations of a particular embedding method and hence cannot be used in conjunction with arbitrary embedding methods.

In this paper, we propose a set of knowledge graph preprocessing operators for textual, numeric, and image literals which can be used to create a KG with only relations from one containing literal information. The resulting knowledge graph can then be processed by any arbitrary embedding method.

The rest of this paper is structured as follows. Section 2 positions our approach in the light of existing research. Section 3 introduces our approach, followed by a set of experiments described in section 4. We conclude with a summary and an outlook on future work.

2. Related Work

Many standard benchmarks for knowledge graph embeddings, especially in the link prediction field, do not come with literals. Hence, the topic has not drawn as much attention as knowledge graph embeddings for purely relational KGs for quite some time.

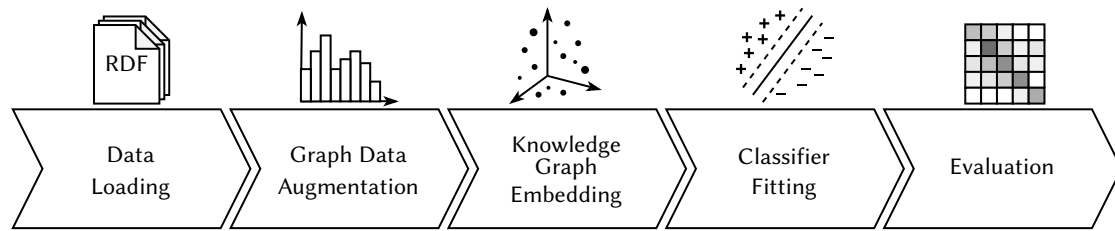
A survey from 2021 [4] lists a number of approaches, which mostly are extensions of existing knowledge graph embedding models, mostly classic models like TransE. Those approaches usually change the loss function of the underlying model and hence are bound to that model alone. An exception is LiteralE [5], which has been applied to different embedding algorithms like TransE, ComplEx, or DistMult. Moreover, most approaches focus only on one modality of literals. A more recent survey from 2023 [6] confirms that picture.

In contrast, the work presented in this paper proposes to preprocess a KG with literals in a way that the information in the literals is represented in a KG with only relational information. We investigate a number of preprocessing techniques for various modalities, which can be applied together with arbitrary embedding models.

The pyRDF2vec [7] implementation of RDF2vec [8] has a functionality to extract literals directly as features. This creates a heterogeneous representation of an entity (consisting of an embedding plus an additional vector of literal values), which is similar to the *Data Properties* strategy described in [9]. In contrast, the approach in this work targets a uniform embedding representation.

An alternative is to alter the knowledge graph upfront, aiming at transforming information in encoded literals into relational statements. Such approaches would not be bound to a particular embedding method, and, if developed for literals with different modalities, could also be combined to exploit. However, approaches based on preprocessing are still rare. One exception is [10], who propose the use of binning of numerical values. We reuse some of

Figure 1: Overall Framework



their approaches in our work in this paper. Another paper [11] also proposes three strategies preprocessing literals, one of which is used as a baseline in this paper.

3. Approach

Our approach relies on graph preprocessing. Instead of changing the embedding approach per se, we augment the graph with additional nodes and edges encoding some of the information encoded in the literals. Fig. 1 shows the overall framework. Specifically, the embedding step is decoupled from the augmentation step. The last two steps (classifier fitting and evaluation) are concerned with evaluation. For the experiments in this paper, we consider node classification problems, but other downstream tasks (such as link prediction, node regression, or node clustering) would also be possible.

3.1. Baselines

For all approaches, we employ three simple baselines. The first, tagged EXCLUDE, simply excludes all literals. Since most embedding approaches ignore literals, this should not have an impact.

The second, tagged TRANSFORM, creates an entity for each combination of a literal value and a property. In the example above,

```
dbr:Mannheim dbo:populationMetro "2362046"^^xsd:nonNegativeInteger .
```

would be transformed to¹

```
dbr:Mannheim dbo:populationMetro new:populationMetro2362046 .
```

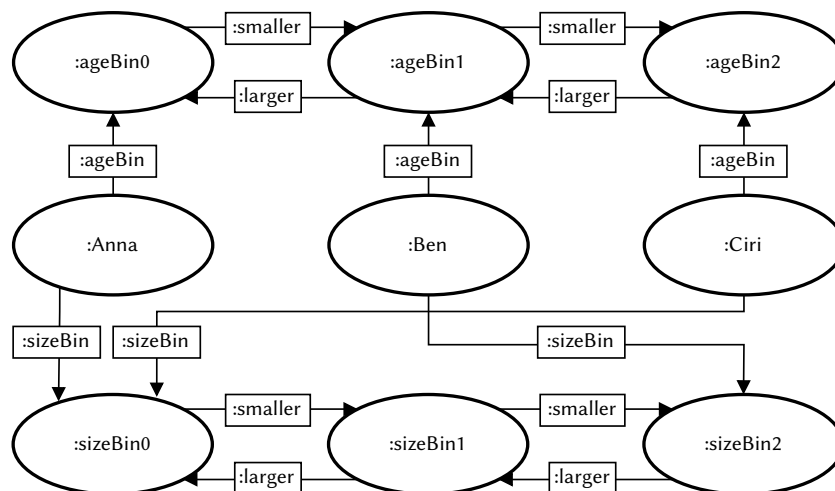
This strategy is identical with the method called *Literal2Entity* in [11].

The third and final baseline, tagged ONEENTITY, creates one single entity for each relation. The idea is to capture any information that is indicated only by the presence or absence of a datatype property (such as `dbo:populationMetro`), regardless of the actual literal value, similarly to the relation strategy in [9]. This strategy would transform the above triple to

```
dbr:Mannheim dbo:populationMetro new:populationMetroAnyValue .
```

¹Note that all of the approaches technically turn an `owl:DatatypeProperty` into an `owl:ObjectProperty`. If this is not wanted, e.g., since the ontology should be further reused, this can trivially be changed, e.g., by moving the property into a different namespace.

Figure 2: Illustration of the nBINS Approach



3.2. Handling Numeric Literals

Creating a single entity for each literal value may not be a good strategy for capturing the semantics of that value. Besides scalability issues, two very similar literal values are indistinguishable from two very dissimilar ones. To counter those issues, we employ a number of additional techniques for representing numeric literals, based on binning.

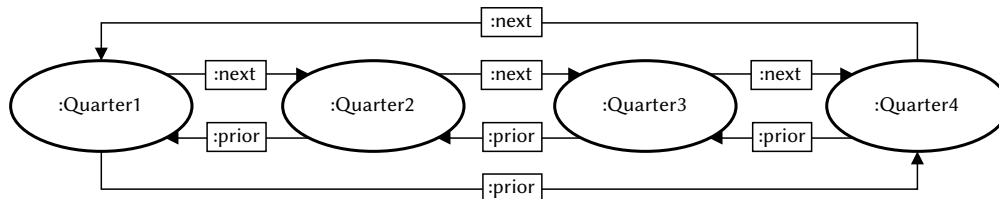
The most basic one, tagged nBINS, is similar to the one proposed in [10]. We create n bins from the set of literal values for each predicate. Furthermore, the entities representing the bins are connected to each other. Fig. 2 shows the idea of this approach. While nBINS requires setting a fixed value for n , p%BINS lets the user set a percentage of unique values. For example, for a datatype property with 1,000 occurrences, and 200 unique values, 10%BINS would create 20 bins (10% of 200). Moreover, we also adapt the idea of *overlapping bins* and *hierarchical binning* from [10], which allows for literal values to be contained in more than one bin, and therefore extends the expressivity of the entities representing bins.

Since outliers can distort the bins created, we also combine the binning with a preceding outlier detection step. Specifically, we use the local outlier factor (LOF) method [12] to first discard outliers, then perform a binning.

Finally, we adopt an idea from [13], which is based on the observation that the same property may be used for multiple types of objects, hence resulting in different blended value distributions. For example, the property `height` may be used for people and buildings, but binning should be conducted on values from both classes separately, since the bin `high` would have a different span for people and buildings.

Since many knowledge graphs do not come with an extensive type system, we alter the original approach in [13] to use either sets of relations for identifying similar and dissimilar entity types (in the example above, people and buildings would come with different sets of relations), and sets of relations and entities. The two approaches are coined KL-REL and KL-RELENT. Both approaches build a lattice of entities with the datatype at hand, and compute

Figure 3: Date Nodes Encoding Quarter of Date



the KL-divergence of the set of relations (or the set of relations and the connected entities, respectively) and split the population of values until it falls below a certain threshold (in our experiments, we use 300 values as a threshold). Then, the binning is performed individually for each subpopulation.

All the approaches create one entity per bin and relation. Hence, the population statement in our example would be transformed to a statement like

```
dbr:Mannheim dbo:populationMetro new:populationMetroBin02 .
```

3.3. Handling Temporal Literals

For temporal literals, i.e., literals typed with `xsd:date`, we follow a different strategy. The first strategy for handling dates, coined DATBIN, turns the date into a UNIX timestamp and applies the nBINS strategy above. In the above example, the statement

```
dbr:Mannheim dbo:foundingDate "1607-01-24"^^xsd:date .
```

would be replaced by a statement like

```
dbr:Mannheim dbo:foundingDate new:foundingDateBin14 .
```

This strategy, however, does not capture the entire information in a date. For example, a similarity of two people with the same birthday (in different years) might not be captured with such an approach. Therefore, to handle temporal literals, we propose a second strategy coined DATFEAT and extract five new features from a date literal.

In the above example, this would yield the statements

```
dbr:Mannheim dbo:foundingDate
  new:wednesday ,
  new:day24 ,
  new:month1 ,
  new:quarter1 ,
  new:year1607 .
```

As shown in Fig. 3, the new entities for days, months, and quarters can again be connected in order to also capture interrelations between them.

3.4. Handling Text Literals

Many knowledge graphs contain rich textual information, but this cannot be represented as easily as the information in numbers and dates. In order to represent textual information, we use *topic modeling*, which assigns each text literal a certain number of topics [14]. Each of those topics is then represented as a node in the graph.

Specifically, we run all values of a text literal (e.g., `dbo:abstract`) through a Latent Dirichlet Allocation (LDA) algorithm, and connect each entity to all topics exceeding a certain threshold (in our experiments in this paper, we use a threshold of 10%). With this strategy coined TXTLDA, the statement

```
dbr:Mannheim dbo:abstract "Mannheim [...] officially the University City of
Mannheim (German: Universitätsstadt Mannheim), is the second-largest city
in the German state of Baden-Württemberg..."@en .
```

could be replaced, e.g., by

```
dbr:Mannheim dbo:abstract
  new:abstractTopic04, new:abstractTopic17 .
```

3.5. Handling Image Literals

For images, we use a similar technique. We reuse a large-scale neural image classification model, which predicts tags for images (e.g., whether the building is showing a person or an animal). Those are then represented as nodes, which are then used to describe the image contents.

In our experiments, we use the pre-trained VGG16 model [15], which computes probabilities for 1,000 classes of images. For each image, we classify it with VGG16 and use the most likely class for each image. In our example above, the triple

```
dbr:Mannheim foaf:depiction
  <http://commons.wikimedia.org/wiki/Special:FilePath/
  NUB_Mannheim_2014-03-13.jpg> .
```

could be replaced by

```
dbr:Mannheim foaf:depiction new:VGG_building .
```

Table 1 depicts the size changes of a knowledge graph for the individual strategies. It can be observed that the number of statements equals the number of original literal statements, and the number of entities is also changing only moderately.

4. Experiments

We test all of the approaches above on the node classification benchmark `kgbench` [16], which contains four heterogeneous datasets, as shown in table 4. As embedding methods, we use TransE [17] and DistMult [18] using the `pyKeen` library [19], and RDF2vec [20] using the `pyRDF2vec` library [7]. As classifiers, we use kNN and SVM using the `scikit-learn` library [21].

Strategy	δE	δS
EXCLUDE	-	-
TRANSFORM	$V * R$	S
ONEENTITY	R	S
nBINS	$n * R$	S
DATBIN	$n * R$	S
DATFEAT	$DW+DD+DM+DQ+DY$	$5 * S$
LDA	T	$T * S$
VGG16	1,000	S

Table 1

Maximum size changes to the knowledge graph in number of entities (δE) and statements (δS). Variables used: number of distinct literal values (V), number of relations (R), number of literal assignment statements (S), number of distinct weekdays (DW), days (DD), months (DM), quarters, (DQ), and years (DY), topics in LDA (T).

Dataset	amplus	dmgfull	dmg777k	mdgenre
Classes	8	14	5	12
Relations	33	62	60	154
Nodes	1,153,679	842,550	341,270	349,344
Triples	2,521,046	1,850,451	777,124	1,252,247
objects thereof...				
...IRIs	1,464,871	593,291	288,379	1,001,791
...blank nodes	256,515	-	-	-
...literals	799,660	1,257,160	488,745	250,456
thereof...				
...numbers	160,959	88,168	10,706	14,352
...dates	202,304	-	-	113,463
...text	377,542	834,244	329,987	54,838
...images	58,855	58,846	46,108	67,804
...others	-	275,902	101,944	-

Table 2

The kgbench dataset

Using the Adam optimizer, the two pyKeen embedders DistMult and TransE were trained in 100 epochs for TransE and 150 epochs for DistMult, using the LCWA train loop. We use a batch size of 75,000 for DistMult and 2,000 for TransE. For all additional parameters, the default parameters provided by pykeen were used. Hereby the pykeen selects the parameters used in the original paper that introduced the selected embedder as default parameters [19]. RDF2vec was trained using a maximum walk depth and 500 walks per node, and 50 training epochs for word2vec. For all additional parameters, the default parameters of pyRDF2vec are used.

For the classifiers, we use a grid search for parameter optimization. For kNN, the parameters in the search space are $k = \{2, 4, 7, 9, 15\}$, for SVM, the parameters in the search space are $C = \{0.01, 0.1, 1, 10, 100\}$. For all other parameters, we use the default values defined by scikit-learn [21].²

²The code for all experiments is available online at <https://gitlab.com/patryk.preisner/mkga/>

Table 3 shows the experiment results. For each literal type, we show the ones which got the best results overall, in addition to the three baselines.³ These are KL-REL with LOF for numeric literals, DATBIN for dates (however, only *amplus* and *mdgenre* contain dates), LDA for text, and VGG16 for images. Moreover, we report results of a combined approach using the combination of the five aforementioned strategies.

From the table, we can observe that in three out of four cases, the best baseline can be outperformed by a few percentage points (0.779 vs. 0.708 on *amplus*, 0.676 vs. 0.606 on *dmg777k*, 0.726 vs. 0.662 on *dmgfull*), whereas for *mdgenre*, none of the approaches yields an advantage over the best baseline excluding literals (RDF2vec+SVM).

Moreover, we can observe that there is no clear correlation between the amount of literals of a particular modality (see table 4) and the improvement achieved by including the corresponding literals. While this might seem counter intuitive, the sheer amount of literals does not reflect the utility of the information contained therein.⁴

The baselines TRANSFORM and ONENTITY are often strong competitors as well, indicating that in many of the cases, the presence of a literal is a strong signal, regardless of the actual literal value.

5. Conclusion and Future Work

We have shown that graph preprocessing is a promising strategy for representing literal information in knowledge graph embeddings, which can be combined with arbitrary embedding methods.

The set of preprocessing operators is not fixed, but can be extended. For example, for text or image representation, while we used basic models to demonstrate the effectiveness of our approach, newer representation models can also be easily plugged in. A staged approach would also be feasible, e.g., representing texts first by means of a BERT encoder and then binning the resulting dimensional values.

Most of the approaches used do not only create entities (e.g., for numerical bins, topics, or image labels), but also come with some score for those. For example, LDA assigns probabilities to topics, given a text. In the experiments in this paper, we used a simple thresholding mechanism to include and exclude the corresponding edges, but it would also be possible to pass the scores to the embedding model as edge weights. [22]

References

- [1] N. Heist, S. Hertling, D. Ringler, H. Paulheim, Knowledge graphs on the web-an overview., Knowledge Graphs for eXplainable Artificial Intelligence (2020) 3–22.
- [2] A. Hogan, E. Blomqvist, M. Cochez, C. d’Amato, G. D. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, et al., Knowledge graphs, ACM Computing Surveys (Csur) 54 (2021) 1–37.

³A full table with the results for all configurations can be found at <https://gitlab.com/patryk.preisner/mkga/>.

⁴As a thought experiment, imagine a numerical ID for each entity, which would greatly increase the number of numerical literals, but the literals would not contain any useful information.

Table 3

Experiment Results. The best results per dataset and embedding method are printed in bold, the best overall results per dataset are additionally underlined.

		amplus		dmg777k		dmgfull		mdgenre	
		KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM
DistMult	EXCLUDE	0.458	0.512	0.548	0.542	0.619	0.658	0.605	0.622
	TRANSFORM	0.477	0.546	0.593	0.611	0.560	0.576	0.575	0.610
	ONEENTITY	0.511	0.549	0.517	0.501	0.613	0.649	0.599	0.617
	KL-REL+LOF	0.564	0.608	0.528	0.523	0.634	0.673	0.616	0.632
	DATBIN	0.501	0.538	-	-	-	-	0.609	0.626
	LDA	0.464	0.504	0.564	0.582	0.652	0.665	0.612	0.623
	VGG16	0.485	0.528	0.549	0.552	0.553	0.642	0.606	0.621
	COMBINED	0.542	0.590	0.579	0.583	0.595	0.643	0.612	0.620
RDF2Vec	EXCLUDE	0.550	0.536	0.586	0.606	0.629	0.661	0.590	0.662
	TRANSFORM	0.616	0.616	0.626	0.628	0.635	0.685	0.583	0.658
	ONEENTITY	0.588	0.612	0.630	0.609	0.631	0.679	0.565	0.657
	KL-REL+LOF	0.536	0.564	0.594	0.594	0.626	0.660	0.584	0.662
	DATBIN	0.554	0.523	-	-	-	-	0.591	0.662
	LDA	0.606	0.610	0.620	0.636	0.631	0.663	0.591	0.660
	VGG16	0.584	0.575	0.623	0.676	0.627	0.651	0.591	0.660
	COMBINED	0.688	0.691	0.664	0.665	0.650	0.715	0.586	0.661
TransE	EXCLUDE	0.682	0.708	0.506	0.528	0.649	0.662	0.634	0.646
	TRANSFORM	0.727	0.761	0.602	0.611	0.665	0.688	0.639	0.649
	ONEENTITY	0.737	0.761	0.610	0.621	0.657	0.673	0.641	0.647
	KL-REL+LOF	0.716	0.726	0.476	0.512	0.643	0.664	0.638	0.644
	DATBIN	0.683	0.719	-	-	-	-	0.634	0.642
	LDA	0.670	0.701	0.554	0.578	0.653	0.672	0.631	0.648
	VGG16	0.723	0.735	0.560	0.588	0.656	0.671	0.632	0.644
	COMBINED	0.760	0.779	0.627	0.643	0.709	0.726	0.635	0.646

- [3] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives, Dbpedia: A nucleus for a web of open data, in: *The Semantic Web: 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007+ ASWC 2007*, Busan, Korea, November 11-15, 2007. Proceedings, Springer, 2007, pp. 722–735.
- [4] G. A. Gesese, R. Biswas, M. Alam, H. Sack, A survey on knowledge graph embeddings with literals: Which model links better literal-ly?, *Semantic Web 12* (2021) 617–647.
- [5] A. Kristiadi, M. A. Khan, D. Lukovnikov, J. Lehmann, A. Fischer, Incorporating literals into knowledge graph embeddings, in: *The Semantic Web–ISWC 2019: 18th International Semantic Web Conference*, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part I 18, Springer, 2019, pp. 347–363.
- [6] N. Fanourakis, V. Efthymiou, D. Kotzinos, V. Christophides, Knowledge graph embedding methods for entity alignment: experimental review, *Data Mining and Knowledge Discovery* (2023) 1–68.
- [7] G. Vandewiele, B. Steenwinckel, T. Agozzino, F. Ongenae, pyrdf2vec: A python implementation and extension of rdf2vec (2022). URL: <https://arxiv.org/abs/2205.02283>. doi:10.48550/ARXIV.2205.02283.

- [8] P. Ristoski, H. Paulheim, Rdf2vec: Rdf graph embeddings for data mining, in: *The Semantic Web–ISWC 2016: 15th International Semantic Web Conference*, Kobe, Japan, October 17–21, 2016, Proceedings, Part I 15, Springer, 2016, pp. 498–514.
- [9] H. Paulheim, J. Fümkrantz, Unsupervised generation of data mining features from linked open data, in: *Proceedings of the 2nd international conference on web intelligence, mining and semantics*, 2012, pp. 1–12.
- [10] J. Wang, F. Ilievski, P. Szekely, K.-T. Yao, Augmenting knowledge graphs for better link prediction, arXiv preprint arXiv:2203.13965 (2022).
- [11] M. Blum, B. Ell, P. Cimiano, Exploring the impact of literal transformations within knowledge graphs for link prediction, in: *Proceedings of the 11th International Joint Conference on Knowledge Graphs*, 2022, pp. 48–54.
- [12] M. M. Breunig, H.-P. Kriegel, R. T. Ng, J. Sander, Lof: identifying density-based local outliers, in: *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 2000, pp. 93–104.
- [13] D. Fleischhacker, H. Paulheim, V. Bryl, J. Völker, C. Bizer, Detecting errors in numerical linked data using cross-checked outlier detection, in: P. Mika, T. Tudorache, A. Bernstein, C. Welty, C. Knoblock, D. Vrandečić, P. Groth, N. Noy, K. Janowicz, C. Goble (Eds.), *The Semantic Web – ISWC 2014*, Springer International Publishing, Cham, 2014, pp. 357–372.
- [14] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, *Journal of machine Learning research* 3 (2003) 993–1022.
- [15] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).
- [16] P. Bloem, X. Wilcke, L. van Berkel, V. de Boer, kgbench: A collection of knowledge graph datasets for evaluating relational and multimodal machine learning, in: *The Semantic Web: 18th International Conference, ESWC 2021, Virtual Event, June 6–10, 2021, Proceedings* 18, Springer, 2021, pp. 614–630.
- [17] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, O. Yakhnenko, Translating embeddings for modeling multi-relational data, *Advances in neural information processing systems* 26 (2013).
- [18] B. Yang, W.-t. Yih, X. He, J. Gao, L. Deng, Embedding entities and relations for learning and inference in knowledge bases, arXiv preprint arXiv:1412.6575 (2014).
- [19] M. Ali, M. Berrendorf, C. T. Hoyt, L. Vermue, S. Sharifzadeh, V. Tresp, J. Lehmann, PyKEEN 1.0: A Python Library for Training and Evaluating Knowledge Graph Embeddings, *Journal of Machine Learning Research* 22 (2021) 1–6. URL: <http://jmlr.org/papers/v22/20-825.html>.
- [20] H. Paulheim, P. Ristoski, J. Portisch, *Embedding Knowledge Graphs with RDF2vec*, Springer, 2023.
- [21] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, G. Varoquaux, API design for machine learning software: experiences from the scikit-learn project, in: *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013, pp. 108–122.
- [22] M. Cochez, P. Ristoski, S. P. Ponzetto, H. Paulheim, Biased graph walks for rdf graph embeddings, in: *Proceedings of the 7th International Conference on Web Intelligence, Mining and Semantics*, 2017, pp. 1–12.