Guiding Knowledge Graph Embeddings with Information Content: An Exploratory Study of Symbolic–Subsymbolic Alignment

Verdiana Schena[1,2], Ricardo M.S. Carvalho[2], Catia Pesquita[2], Simona Colucci[1] and Floriano Scioscia[1]

[1]*Polytechnic University of Bari, Bari, Italy.*
[2]*University of Lisbon, Lisbon, Portugal.*

Abstract

Knowledge graph embeddings map symbolic graph structures into continuous vector spaces that can be processed by machine learning models. However, it remains unclear to what extent the resulting subsymbolic representations preserve meaningful aspects of the symbolic structure encoded in the original knowledge graph. This work investigates whether symbolic information grounded in information theory can guide embedding generation and improve the alignment between symbolic and subsymbolic representations. We consider two Information Content measures that quantify entity informativeness from complementary structural perspectives, based respectively on occurrence frequency and on the rarity of local graph features. Information Content is used as a symbolic relevance signal to control which graph contexts are more prominently considered during embedding generation. We introduce two weighting strategies: one following the conventional information-theoretic view by emphasizing highly informative contexts, and another shifting the focus toward moderately informative ones. The weights produced by these strategies are integrated into RDF2Vec by biasing random-walk transition probabilities, thereby controlling which symbolic graph contexts exert greater influence on the learned representations. We evaluate five embedding configurations on FB15K-237, NELL-995, and WN18RR by comparing neighborhoods induced by graph-based features with those emerging in the embedding space. Their agreement is assessed using Overlap and Rank-Biased Overlap at multiple cutoff values. The results show that Information Content-based weighting can improve the alignment between symbolic and subsymbolic neighborhoods, although its effectiveness depends on the dataset and weighting strategy.

Keywords

Knowledge Graph Embeddings, Information Content, Biased Graph Walks, Symbolic–Subsymbolic Alignment, Representation Learning

1. Introduction

Knowledge graphs (KGs) provide structured and machine-interpretable representations of entities and their semantic relationships. Knowledge graph embeddings (KGEs) complement this symbolic representation by mapping entities into continuous vector spaces that can be processed by machine learning models [1, 2]. A central challenge in this symbolic-to-subsymbolic transformation is ensuring that the resulting embedding space reflects meaningful aspects of the symbolic organization and relational structure encoded in the original graph. However, recent studies [3, 4, 5, 6, 7] have questioned the extent to which KGEs reflect the symbolic organization of the original graph and whether proximity in the embedding space can reliably be interpreted as semantic relatedness.

A key consideration in semantic similarity measures in the context of KGs and ontologies is the information content (IC) of a concept or entity. IC has long been used in taxonomies and ontologies to quantify the specificity and informativeness of concepts, with several measures proposed over the years [8, 9, 10]. More recent approaches [11, 12] have extended the use of IC beyond purely taxonomic

Third Workshop on Knowledge Graphs and Neurosymbolic AI (KG-NeSy 2026), in ISWC 2026 Workshops Joint Proceedings, October 25–26, 2026, Bari, Italy
v.schena@phd.poliba.it (V. Schena); rmscarvalho@fc.ul.pt (R. M.S. Carvalho); clpesquita@ciencias.ulisboa.pt (C. Pesquita); simona.colucci@poliba.it (S. Colucci); floriano.scioscia@poliba.it (F. Scioscia)
0009-0009-1660-5462 (V. Schena); 0009-0001-6605-7603 (R. M.S. Carvalho); 0000-0002-1847-9393 (C. Pesquita); 0000-0002-8774-4816 (S. Colucci); 0000-0002-7859-9602 (F. Scioscia)
© 2026 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 ISSN 1613-0073
published 2026-11-05

CEUR-WS.org/Vol-4313/kgnesy-paper17.pdf

structures, adapting it to the richer topology of Linked Data and KGs. The intuition is simple: concepts that share more information are considered more similar.

This work investigates whether symbolic information about an entity's informativeness can guide embedding generation and improve the ability of the resulting vector representations to reflect graph-induced symbolic structure. More specifically, it addresses the following research question (RQ):

> ***RQ**: To what extent does incorporating Information Content into KGE improve agreement between graph-induced and embedding-based entity neighborhoods?*

To answer the RQ, from a neuro-symbolic perspective, we propose two strategies that use explicitly computed symbolic properties of the KG to guide embedding generation. Specifically, IC acts as a symbolic relevance signal that controls which graph contexts are more likely to be sampled and presented to the model. The two strategies modulate the influence of symbolic graph contexts in different ways. The first follows the conventional information-theoretic view, assigning greater importance to contexts involving highly informative resources, under the hypothesis that they provide more discriminative evidence for learning distinct entity representations. Conversely, contexts involving common and less informative resources are assigned lower importance, since broadly shared information may contribute similarly to many entities and reduce the distinctiveness of their embeddings. The second strategy challenges the assumption that higher informativeness always implies greater relevance by investigating whether representation learning may benefit more from contexts that balance specificity and structural sharing. Rather than simply prioritizing the most informative contexts, this strategy seeks an informativeness balance point, emphasizing contexts that are sufficiently distinctive to discriminate among entities while remaining sufficiently shared to capture recurring graph patterns. Accordingly, low-informativeness contexts are assigned lower weights because of their limited discriminative value, whereas extremely informative contexts are also given less emphasis because they may correspond to highly localized or insufficiently shared structural evidence.

Informativeness is defined using two different measures that capture distinct structural perspectives on Linked Data: an occurrence-based measure [11], which estimates entity informativeness from its frequency in the KG, and Partitioned Information Content (*PIC*) [12], which derives it from the rarity of local incoming and outgoing structural features. Each measure is applied under both weighting strategies, resulting in four IC-based embedding configurations, which are compared with standard uniform-weighting strategy.

The evaluation is conducted on three benchmark KGs by comparing the neighborhood structure induced by symbolic graph features with that emerging in the embedding space. The aim is to determine whether incorporating IC through the two proposed weighting strategies leads to a stronger alignment between symbolic and subsymbolic representations. For each entity, symbolic neighborhoods are obtained from rankings based on Jaccard similarity [13] over graph-based features, while subsymbolic neighborhoods are derived from cosine similarity and Euclidean distance between entity embeddings. Agreement between the resulting rankings is assessed using *Rank-Biased Overlap* (*RBO*) and top-k *Overlap* at different cutoff values.

2. Related Work

Several studies investigated how knowledge graph embeddings can be enhanced for different purposes. Some works address this issue from the perspective of interpretability, some focus on improving task-specific performance, while others aim to incorporate external or semantic relevance signals. Despite their different goals, they share the idea that embedding representations can be adapted to better suit specific objectives or application contexts. Along this line, Domingues et al. [4] propose InterpretE, a neuro-symbolic approach that derives interpretable vector spaces whose dimensions correspond to human-understandable entity features. Their evaluation considers both the quality of the resulting clusters in the derived vector space and the semantic similarity of the corresponding entities. A second line of research focuses on modifying the walk generation process used by RDF2Vec [14]

by incorporating weights into the sampling strategy. Cochez et al. [15] propose twelve different edge-weighting functions based on information internal to the KG, mainly relying on node and edge frequencies as well as PageRank-based signals. The proposed strategies are evaluated on classification, regression, and document similarity tasks. Extending this idea, Taweel and Paulheim [16] guide RDF2Vec walks using external edge weights derived from Wikipedia clickstream data. Their evaluation considers entity classification, regression, and a content-based recommender system, where cosine similarity between embedding vectors is used to identify similar items. Similarly, previous studies [17, 7, 18] incorporate information external to the KG to generate embeddings whose geometric similarity better reflects a user-oriented notion of semantic similarity. They qualitatively compare the similarity induced by the embeddings with the amount of relevant information that can be obtained through logical reasoning, thus investigating whether the embedding space reflects the information that is meaningful for the user. A further line of work adopts an information-theoretic perspective to assign different importance to graph statements. Mai et al. [19] argue that treating all triples equally overlooks differences in their informativeness. Their intuition is that triples that can be easily inferred from the remainder of the graph are more redundant and therefore less informative. To address this issue, they combine rule mining and PageRank to identify less redundant triples and assign them greater importance during embedding learning. The resulting weights are incorporated into TransE, TransR, and HolE, and the approach is evaluated on link prediction tasks. Although these studies show that embedding generation can be guided by structural, external, or information-theoretic signals, our work differs from them in both the guidance signal and the evaluation objective. Previous biased RDF2Vec approaches rely mainly on frequency-, centrality-, or externally derived weights and evaluate their effects on downstream tasks such as classification, regression, recommendation, or document similarity. Information-theoretic weighting approaches, instead, have mainly been applied to triple-based embedding models and evaluated through link prediction. In contrast, we use entity-level Information Content as an intrinsic symbolic relevance signal to bias RDF2Vec random walks, and evaluate whether this guidance improves the agreement between graph-induced and embedding-based entity neighborhoods. To the best of our knowledge, prior work has not examined whether IC-guided RDF2Vec walks improve neighborhood-level symbolic–subsymbolic agreement.

3. Methodology

3.1. Informativeness Measures

We consider two informativeness measures that capture different structural perspectives on Linked Data: the Information Content measure proposed by Nunes et al. [11], here denoted as IC_N to distinguish it from the general notion of IC, and Partitioned Information Content (*PIC*), proposed by Meymandpour and Davis [12].

IC_N. Given a knowledge graph (G), IC_N of an entity (e) is defined as:

$$IC_N(e) = -\log \frac{deg(e)}{|G|}, \tag{1}$$

where $deg(e)$ denotes the number of RDF triples in which e occurs, and $|G|$ is the total number of RDF triples in the KG. Thus, IC_N estimates entity informativeness from its occurrence frequency in the KG: entities occurring in fewer triples receive higher values.

PIC. *PIC* estimates entity informativeness by aggregating the information content of its local structural features. For an entity e, let F_e denote its set of features, where each feature is represented as a tuple $\langle p, e', D\rangle$, consisting of a predicate p, a neighboring entity e' and a direction D, which indicates whether the corresponding edge is incoming or outgoing. The IC of a feature f_i is computed as:

$$IC(f_i) = -\log_2 \frac{\phi(f_i)}{N}, \tag{2}$$

where $\phi(f_i)$ is the number of entities sharing feature f_i, and N is the total number of entities in the KG. *PIC* of an entity is then computed as the sum of the IC values of its features:

$$PIC(e) = \sum_{f_i \in F_e} IC(f_i). \tag{3}$$

Therefore, while IC_N captures informativeness through an entity's overall frequency in the KG, *PIC* captures it through the rarity of the incoming and outgoing structural features in its local neighborhood.

3.2. Information Content-based Context Weighting Strategies

We define two weighting strategies for controlling the contribution of symbolic graph contexts to the learned vector representations. Although each weight is associated with an entity, it is intended to modulate the relevance of the graph contexts in which that entity occurs during embedding generation. Consequently, contexts involving higher-weighted entities are expected to exert a stronger influence on the resulting embedding space. For both strategies, the informativeness score $s(e)$ of an entity e is instantiated either as $IC_N(e)$ or as $PIC(e)$.

3.2.1. Emphasizing Highly Informative Contexts

This first strategy follows the information-theoretic intuition that more specific entities provide more informative and discriminative evidence than generic ones. It therefore assigns progressively higher weights to entities with higher informativeness values. The objective is to give greater importance to resource-specific information, so that it can exert a stronger influence on the learned representations, while reducing the contribution of generic information shared by many entities, which may otherwise make their embeddings less distinguishable.
To obtain weights that preserve the relative informativeness of entities while remaining comparable and bounded, each raw score $s(e)$ is transformed as follows:

$$w_{\text{high}}(e) = \sigma\left(\frac{\log(1 + s(e)) - \mu}{\mathit{std}}\right), \tag{4}$$

where σ is the sigmoid function, and μ and *std* are the mean and standard deviation of the log-transformed informativeness scores within the corresponding dataset. The logarithmic transformation is applied to reduce the impact of very large values in the strongly skewed IC_N and *PIC* distributions. The resulting scores are then standardized to express entity informativeness relative to the distribution observed in each dataset. Finally, the sigmoid function maps the standardized values to the interval $(0, 1)$, producing bounded weights and further limiting the influence of extreme scores. Since all transformations are monotonically increasing, the original ordering of entity informativeness is preserved: entities with higher IC_N or *PIC* values receive progressively larger weights.

3.2.2. Emphasizing Moderately Informative Contexts

This strategy investigates whether entities with intermediate informativeness provide more useful information for constructing semantically meaningful representations. The underlying hypothesis is that moderately informative entities may offer a balance between generic information, which is broadly shared but weakly discriminative, and extremely specific information, which may describe isolated or highly local characteristics. To model this hypothesis, each raw score $s(e)$ (IC_N or *PIC*) is first transformed using $\log(1 + s(e))$ and then mapped to a Gaussian-shaped weight:

$$w_{\text{mid}}(e) = \exp\left(-\frac{(\log(1 + s(e)) - \mu)^2}{2(\alpha * \mathit{std})^2}\right), \tag{5}$$

where μ and *std* are the mean and standard deviation of the log-transformed scores $\log(1 + s(e))$ computed within each dataset, while α controls the width of the Gaussian weighting function. In

our experiments, we set $\alpha = 0.5$. This transformation assigns the highest weights to entities whose informativeness is close to the dataset-specific mean, while progressively reducing the weights of entities with very low or very high informativeness. As a result, information associated with moderately informative entities is given greater importance and is expected to be more prominently reflected in the learned embedding space.

3.3. Integrating Informativeness Weights into Embedding Generation

To incorporate the informativeness measures and weighting strategies introduced in Sections 3.1 and 3.2, we rely on RDF2Vec [14], which generates sequences of entities and relations through graph walks and uses them to learn vector representations. Since these sequences constitute the corpus from which the embeddings are learned, biasing the walks affects which graph information is more prominently represented in the resulting vector space. To this end, we adapt the node-centric Object Frequency Weight strategy proposed by Cochez et al. [15], in which each node is assigned a score derived from its frequency in the KG. We retain this mechanism but replace the object-frequency score with the informativeness-based weights introduced in Sections 3.2.1 and 3.2.2. Following the original formulation [15], suppose that a walk reaches a vertex v with outgoing edges $vo_1, \ldots, vo_d$, where vo_l denotes the edge connecting v to the destination node o_l. The probability of following edge vo_l is defined as:

$$\Pr[\text{follow edge } vo_l] = \frac{\text{weight}(vo_l)}{\sum_{i=1}^{d} \text{weight}(vo_i)}. \tag{6}$$

The normalized edge weights are thus directly interpreted as transition probabilities. In our formulation, the weight of an edge vo_l, denoted by $\text{weight}(vo_l)$ in Equation 6, is obtained by applying either the High or Mid weighting strategy to the $IC_N(o_l)$ or *PIC*(o_l) score of the destination entity, following the distributions illustrated in Figure 1. Consequently, edges leading to higher-weighted entities are more likely to be followed, affecting which graph contexts are captured in the generated sequences and, therefore, their influence on the learned embeddings.

4. Experiments

4.1. Datasets

Experiments are conducted on three benchmark KGs commonly used to evaluate knowledge graph embedding methods: FB15K-237 [20], WN18RR [21], and NELL-995 [22][1].

4.2. Knowledge Graph Embeddings

Knowledge graph embeddings are generated using pyRDF2Vec [23]. pyRDF2Vec allows the walk-generation process to be easily customized through biased sampling strategies, enabling us to incorporate informativeness-based transition probabilities. Following the experimental configuration adopted by Portisch et al. [24]: we use Skip-gram Word2Vec with an embedding dimensionality of 200, a context window of 5, and 25 training epochs. For the random walk generation, we set the maximum depth to 4 and generate at most 2,000 walks per entity. The random seed is fixed to 42 to support reproducibility. For each dataset, we generate five embedding configurations that differ exclusively in the strategy used to assign weights to the outgoing edges during walk generation:

- ***Uniform***: assigns equal transition probabilities to all outgoing edges;
- ***IC$_N$-High***: uses IC_N to favor more informative contexts;

[1]All datasets are obtained from the OpenKE benchmark repository: https://github.com/thunlp/OpenKE/tree/OpenKE-PyTorch/benchmarks. The *train2id.txt*, *valid2id.txt*, and *test2id.txt* files are merged to construct the complete KG, since neither RDF2Vec-based embedding generation nor the neighborhood-based evaluation requires predefined data splits.

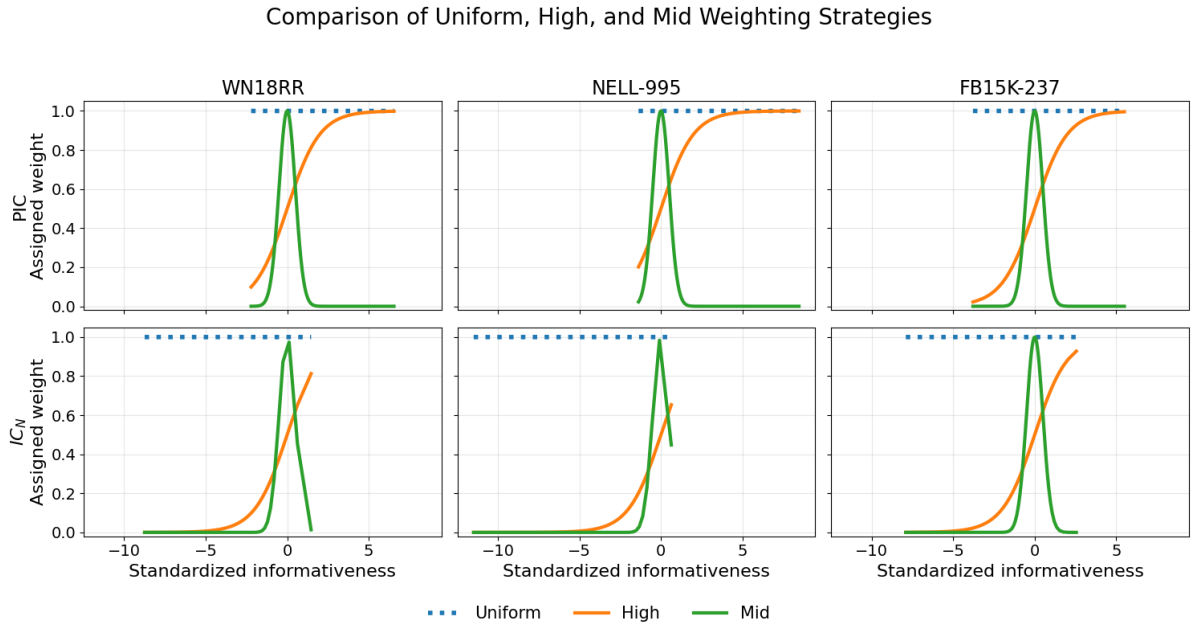


Figure 1: Entity-weight distributions produced by the *Uniform*, *High*, and *Mid* strategies from *PIC* and IC_N values on WN18RR, NELL-995, and FB15K-237. Standardized informativeness denotes log-transformed scores, $\log(1 + s(e))$, normalized through dataset-specific z-scores. The *Uniform* strategy is represented by a dashed line at weight 1 for visual comparison; any positive constant value would produce the same uniform transition probabilities after normalization. The resulting weights bias random-walk sampling and, consequently, the contexts used to learn the embeddings.

- ***IC_N-Mid***: uses IC_N to favor moderately informative contexts;
- ***PIC-High***: uses *PIC* to favor more informative contexts;
- ***PIC-Mid***: uses *PIC* to favor moderately informative contexts.

All remaining hyperparameters are kept fixed, allowing differences between the resulting embeddings to be attributed to the walk-weighting configuration.

4.3. Evaluation

To assess whether informativeness-weighted embedding generation affects the alignment between symbolic and subsymbolic representations, we compare the four informativeness-based configurations with the *Uniform* baseline. More specifically, we investigate whether the neighborhood structure induced by graph-based entity features is reflected in the learned embedding space. For each entity, the symbolic neighborhood is obtained by ranking the remaining entities according to *Jaccard similarity* between their respective sets of one-hop incoming and outgoing graph features. The subsymbolic neighborhood is instead obtained from the corresponding entity embeddings using *cosine similarity*[2] and *Euclidean distance*[3]. Accordingly, entities are ranked in descending order of Jaccard and cosine similarity, and in ascending order of Euclidean distance. To compare the symbolic and subsymbolic neighborhood rankings, we use *Overlap* and *Rank-Biased Overlap*[4] [25].

For each entity, let S_k and T_k denote the sets of the first k entities in the symbolic and subsymbolic rankings, respectively. Top-k *Overlap* is defined as:

$$\text{Overlap@}\, k = \frac{|S_k \cap T_k|}{k}. \tag{7}$$

[2]https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.cosine_similarity.html
[3]https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise_distances.html
[4]https://github.com/changyaochen/rbo/

It measures the proportion of entities shared by the two rankings within their first k positions.

We further perform a more fine-grained comparison using *RBO*. Following the truncated-ranking formulation adopted by Hubert et al. [5], *RBO* is computed as:

$$\mathrm{RBO@}\, k = \frac{1}{k} \sum_{d=1}^{k} \frac{|S_d \cap T_d|}{d}. \tag{8}$$

Unlike *Overlap*, which considers only the first k entities as sets, *RBO* compares the rankings at each depth d, giving greater relevance to agreement at higher positions.

We compute both measures at $k \in \{3, 10, 100\}$. For each dataset and embedding configuration, we report the average scores over all entities. Higher values indicate stronger alignment between the symbolic and subsymbolic neighborhood structures.

5. Results

Figure 2 provides a visual summary of the alignment between symbolic and subsymbolic neighborhood rankings obtained using cosine similarity, reporting mean *Overlap* and *RBO* at different cutoff values. The corresponding numerical values are reported in Tables 1 and 2. Both graphical and tabular representations are included to highlight complementary aspects of the results: the plots facilitate the identification of overall trends, while the tables expose exact values. Results based on Euclidean distance closely reproduce the same overall trends and do not lead to substantially different conclusions; therefore, the corresponding plots and numerical values are reported in *Appendix A* and *Appendix B*, respectively[5].

Mean Overlap@3					
Dataset	*Uniform*	IC_N*-High*	IC_N*-Mid*	*PIC-High*	*PIC-Mid*
NELL-995	0.09988	0.10908	0.10857	0.09779	0.10685
WN18RR	0.25146	0.25365	0.25301	0.25332	0.25318
FB15K-237	0.12427	0.14712	0.16111	0.12317	0.15998
Mean Overlap@10					
Dataset	*Uniform*	IC_N*-High*	IC_N*-Mid*	*PIC-High*	*PIC-Mid*
NELL-995	0.12678	0.13729	0.13776	0.12379	0.13673
WN18RR	0.27172	0.27324	0.27221	0.27147	0.27279
FB15K-237	0.17374	0.20521	0.21102	0.17474	0.20896
Mean Overlap@100					
Dataset	*Uniform*	IC_N*-High*	IC_N*-Mid*	*PIC-High*	*PIC-Mid*
NELL-995	0.16476	0.17475	0.17959	0.16318	0.17848
WN18RR	0.17454	0.17506	0.17473	0.17475	0.17383
FB15K-237	0.22233	0.24847	0.24558	0.22386	0.24534

Table 1
Mean *Overlap* for cosine–Jaccard neighborhood ranking alignment at different cutoff values.

The results shown in Figure 2 and reported numerically in Tables 1 and 2 can be discussed from several complementary perspectives:

Dataset perspective. The dataset-dependent behavior can be interpreted in light of the weight distributions shown in Figure 1. The plots illustrate how the standardized informativeness scores observed in each KG are transformed by the *High* and *Mid* weighting functions. When the scores span a broad region of a curve, the transformation produces more heterogeneous entity weights; when they fall within a narrow or relatively flat region, the resulting weights are more similar. This distinction

[5]The Appendices are provided in the Supplementary Material (https://github.com/researchprojects57-ui/Supplementary_Material_KG_Nesy_Workshop_ISWC_2026/tree/main

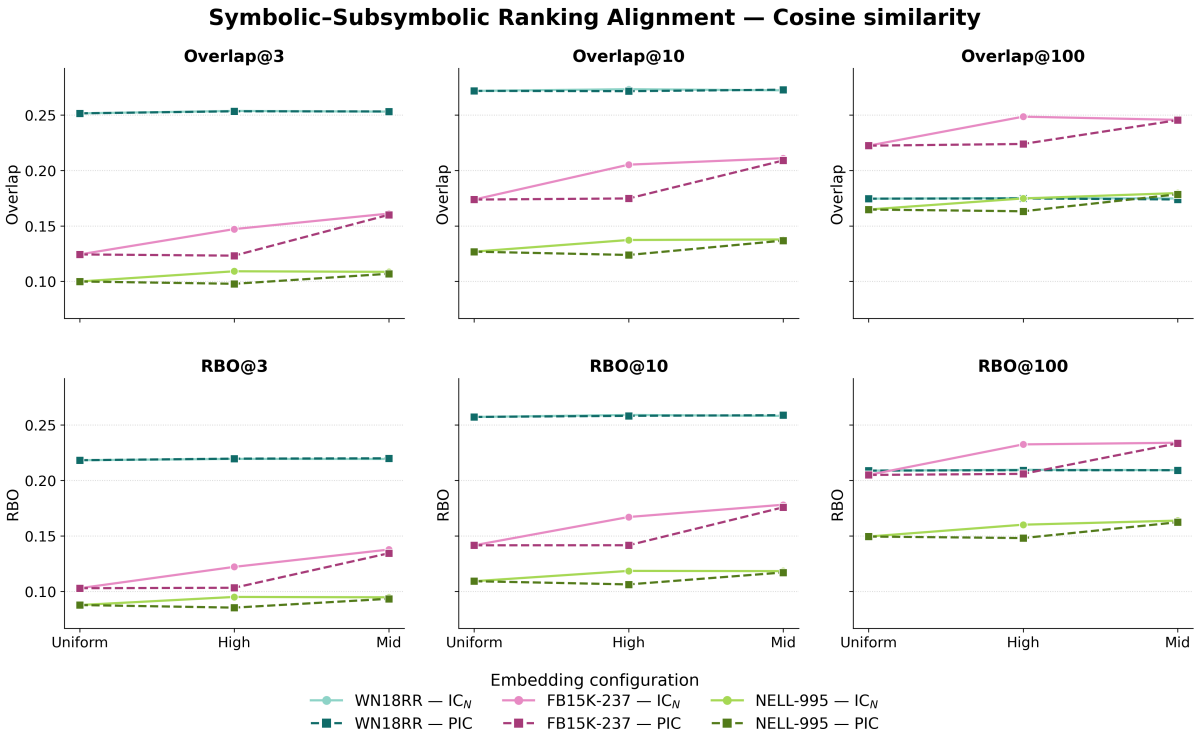


Figure 2: Alignment between symbolic and subsymbolic neighborhood rankings using cosine similarity. Mean *Overlap* (top row) and Mean *RBO* (bottom row) are reported at $k \in \{3, 10, 100\}$ for the *Uniform*, *High*, and *Mid* weighting strategies on WN18RR, FB15K-237, and NELL-995. Solid lines denote configurations based on IC_N, while dashed lines denote *PIC*-based configurations. Higher values indicate stronger alignment between the symbolic and subsymbolic neighborhood structures.

Mean RBO@3					
Dataset	*Uniform*	IC_N*-High*	IC_N*-Mid*	*PIC-High*	*PIC-Mid*
NELL-995	0.08771	0.09501	0.09469	0.08537	0.09333
WN18RR	0.21827	0.21947	0.21939	0.21969	0.21999
FB15K-237	0.10289	0.12212	0.13772	0.10326	0.13441
Mean RBO@10					
Dataset	*Uniform*	IC_N*-High*	IC_N*-Mid*	*PIC-High*	*PIC-Mid*
NELL-995	0.10919	0.11849	0.11829	0.10631	0.11707
WN18RR	0.25728	0.25907	0.25822	0.25830	0.25888
FB15K-237	0.14157	0.16708	0.17814	0.14163	0.17583
Mean RBO@100					
Dataset	*Uniform*	IC_N*-High*	IC_N*-Mid*	*PIC-High*	*PIC-Mid*
NELL-995	0.14947	0.16014	0.16384	0.14809	0.16235
WN18RR	0.20889	0.20963	0.20936	0.20927	0.20925
FB15K-237	0.20500	0.23255	0.23402	0.20598	0.23354

Table 2
Mean *RBO* for cosine–Jaccard neighborhood ranking alignment at different cutoff values.

matters because the weights are normalized locally during random-walk generation. At each step, only the candidate entities reachable from the current node determine the transition probabilities. Therefore, IC-based weighting can substantially alter the sampled contexts only when locally competing candidates receive sufficiently different weights. Otherwise, the resulting probabilities remain close to those of the *Uniform* strategy. This effect is most evident for FB15K-237. In particular, the IC_N values span a relatively broad portion of both weighting curves, producing clearer contrasts among entity weights and larger deviations from uniform sampling. This is consistent with the strongest improvements over

the *Uniform* baseline: IC_N-*Mid* achieves the highest score in five of the six cosine-based settings and in all six Euclidean-based settings, while IC_N-*High* performs best only for cosine-based *Overlap* at $k = 100$. *PIC-Mid* generally remains close to IC_N-*Mid*, whereas *PIC-High* yields limited improvements. For NELL-995, the informativeness scores occupy a less differentiated region of the weighting curves, producing smaller contrasts among entity weights and more moderate improvements. Both IC_N-based configurations outperform the baseline, and *PIC-Mid* also improves neighborhood alignment, whereas *PIC-High* remains slightly below the *Uniform* configuration. For WN18RR, especially under IC_N, the observed values produce limited differentiation among the assigned weights. Consequently, outgoing candidates may receive similar weights after normalization, leaving the transition probabilities close to those of the *Uniform* baseline. This is consistent with the negligible variations observed for both *Overlap* and *RBO*.

Neighborhood-size perspective. For FB15K-237, the improvements introduced by informativeness-based weighting are largest at the smallest cutoff values and decrease as k increases. Although this reduction is not always clearly visible in the plots, the numerical results show that the gains remain non-negligible at $k = 100$. Compared with the *Uniform* baseline, the best configurations improve *Overlap*@100 by approximately 10–12% and *RBO*@100 by approximately 14% under both cosine similarity and Euclidean distance. This indicates that the weighting strategies primarily affect the nearest subsymbolic neighborhoods, while still influencing broader neighborhoods. For NELL-995, the gains are more modest and relatively stable across cutoff values, whereas WN18RR remains largely unaffected by neighborhood size.

Informativeness-measure perspective. The IC_N-based configurations generally achieve higher scores than the corresponding *PIC*-based configurations. On FB15K-237 and NELL-995, IC_N outperforms *PIC* for both weighting strategies across all cutoff values and evaluation measures. On WN18RR, the comparison is less consistent: the differences between the two measures remain marginal.

Weighting-strategy perspective. Qualitatively, Figure 1 shows that the two strategies emphasize different regions of the informativeness distribution. The *High* strategy progressively increases the weight assigned to entities toward the right tail of the informativeness distribution, systematically favouring the most informative contexts. In contrast, we introduce the *Mid* strategy to investigate a different hypothesis: moderately informative entities may provide more useful contexts than entities at either extreme of the distribution. Accordingly, *Mid* shifts the region of maximum weight toward the centre, prioritizing entities with intermediate informativeness while assigning lower weights to both poorly informative and highly informative ones. The evaluation results provide support for this alternative perspective, particularly on FB15K-237 and NELL-995. On FB15K-237, the newly introduced *Mid* strategy outperforms *High* in eleven of the twelve pairwise comparisons, with IC_N-*High* exceeding IC_N-*Mid* only for *Overlap* at $k = 100$. On NELL-995, *PIC-Mid* consistently outperforms *PIC-High*, whereas the two IC_N-based strategies alternate depending on the cutoff and evaluation measure. On WN18RR, all weighting configurations produce very similar scores, with only negligible variations across strategies and informativeness measures. The observed variations are too small to support a meaningful comparison between *High* and *Mid*.

Evaluation-measure perspective. *Overlap* and *RBO* exhibit highly consistent trends across datasets and weighting configurations. In most settings, *Overlap* produces numerically higher values than *RBO*, consistently with the fact that *Overlap* only considers whether the same entities occur within the two top-k neighborhoods, whereas *RBO* also accounts for differences in their relative ordering. An exception is observed for WN18RR at $k = 100$, where *RBO* is higher than *Overlap*. This suggests that the two rankings agree more strongly at the top.

6. Discussion

The results show that Information Content-based weighting can improve the alignment between symbolic and subsymbolic neighborhoods, although its effectiveness varies substantially across datasets. The clearest improvements are observed on FB15K-237 and, to a lesser extent, on NELL-995, whereas the differences from the *Uniform* baseline on WN18RR are negligible. This variability indicates that the usefulness of Information Content as a symbolic guidance signal is strongly influenced by the structural characteristics of the underlying knowledge graph. A possible explanation lies in the dependence of the adopted informativeness measures on the topology of the graph. Their score distributions, and consequently the weights assigned during walk generation, vary across datasets. When the weighting scheme sufficiently differentiates locally competing entities, it can substantially alter transition probabilities and the contexts presented to the embedding model. Conversely, similar weights result in sampling behavior that remains close to the *Uniform* baseline. The results also highlight the importance of the mapping from informativeness to context relevance. On FB15K-237, the *Mid* strategy is generally more effective than the *High* strategy. This suggests that systematically assigning greater relevance to the most informative contexts, as implied by a conventional information-theoretic interpretation, may not always produce representations that more closely reflect symbolic neighborhood structure. Moderately informative contexts may instead provide a more effective balance between discriminative evidence and structural patterns shared across entities.

7. Conclusions and Future work

This work presents an exploratory investigation into the use of Information Content to guide knowledge graph embedding generation toward representations that more closely reflect symbolic neighborhood structures. We combined two informativeness measures, IC_N and *PIC* with two weighting strategies: *High* and *Mid*, and integrated the resulting weights into RDF2Vec walk generation. The resulting embedding configurations are evaluated by measuring the agreement between graph-based and embedding-based entity neighborhoods, showing that IC-based guidance can improve their alignment, although its effectiveness varies across knowledge graphs. Future work will assess the stability of these findings across multiple random seeds and additional knowledge graphs. We also plan to investigate which graph properties influence the effectiveness of the weighting strategies.

Acknowledgments

This work is co-funded by the Italian aerospace technological district "Distretto Tecnologico Aerospaziale S.C.a.r.l.". This work was partially funded by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., under the fellowship 2022.10769.BD and under the LASIGE Research Unit, ref. UID/00408/2025, DOI
https://doi.org/10.54499/UID/00408/2025. It was also partially supported by the CancerScan project which received funding from the EU's HORIZON Europe research and innovation programme under grant agreement No 101186829.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] S. Ji, S. Pan, E. Cambria, P. Marttinen, P. S. Yu, A survey on knowledge graphs: Representation, acquisition, and applications, IEEE Transactions on Neural Networks and Learning Systems 33 (2022) 494–514. doi:10.1109/TNNLS.2021.3070843.

[2] J. Cao, J. Fang, Z. Meng, S. Liang, Knowledge graph embedding: A survey from the perspective of representation spaces, ACM Comput. Surv. 56 (2024). doi:10.1145/3643806.

[3] N. Jain, J.-C. Kalo, W.-T. Balke, R. Krestel, Do Embeddings Actually Capture Knowledge Graph Semantics?, in: The Semantic Web: 18th International Conference, ESWC 2021, Virtual Event, June 6–10, 2021, Proceedings, Springer-Verlag, Berlin, Heidelberg, 2021, p. 143–159. doi:10.1007/978-3-030-77385-4_9.

[4] A. Domingues, N. Jain, A. M. Peñuela, E. Simperl, Bringing Back Semantics to Knowledge Graph Embeddings: An Interpretability Approach, in: T. R. Besold, A. d'Avila Garcez, E. Jimenez-Ruiz, R. Confalonieri, P. Madhyastha, B. Wagner (Eds.), Neural-Symbolic Learning and Reasoning, Springer Nature Switzerland, Cham, 2024, pp. 192–203.

[5] N. Hubert, H. Paulheim, A. Brun, D. Monticolo, Do Similar Entities Have Similar Embeddings?, in: The Semantic Web: 21st International Conference, ESWC 2024, Hersonissos, Crete, Greece, May 26–30, 2024, Proceedings, Part I, Springer-Verlag, Berlin, Heidelberg, 2024, p. 3–21. doi:10.1007/978-3-031-60626-7_1.

[6] G. Méroué, F. Gandon, P. Monnin, Link prediction or perdition: The seeds of instability in knowledge graph embeddings, in: M. Acosta, M. van Erp, S. Rudolph, O. Hartig, B. Spahiu, A. Rula, D. Garijo, F. Osborne (Eds.), The Semantic Web, Springer Nature Switzerland, Cham, 2026, pp. 198–216.

[7] V. Schena, S. Colucci, F. Scioscia, F. M. Donini, Human-in-the-Loop Refinement of Knowledge Graph Embeddings for Semantic Similarity, in: L. M. Camarinha-Matos, F. Ferrada (Eds.), Technological Innovation to Tackle Societal Challenges, Springer Nature Switzerland, Cham, 2027, pp. 358–369.

[8] P. Resnik, Using information content to evaluate semantic similarity in a taxonomy, in: Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1, IJCAI'95, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1995, p. 448–453.

[9] N. Seco, T. Veale, J. Hayes, An intrinsic information content metric for semantic similarity in wordnet, in: Proceedings of the 16th European Conference on Artificial Intelligence, ECAI'04, IOS Press, NLD, 2004, p. 1089–1090.

[10] Z. Zhou, Y. Wang, J. Gu, A new model of information content for semantic similarity in wordnet, in: 2008 Second International Conference on Future Generation Communication and Networking Symposia, volume 3, 2008, pp. 85–89. doi:10.1109/FGCNS.2008.16.

[11] S. Nunes, S. Badreddine, C. Pesquita, Rewarding explainability in drug repurposing with knowledge graphs, in: Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI '25, 2025. doi:10.24963/ijcai.2025/515.

[12] R. Meymandpour, J. G. Davis, A semantic similarity measure for linked data: An information content-based approach, Knowledge-Based Systems 109 (2016) 276–293. doi:https://doi.org/10.1016/j.knosys.2016.07.012.

[13] P. Jaccard, Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines, Bull Soc Vaudoise Sci Nat 37 (1901) 241–272.

[14] 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, Springer-Verlag, Berlin, Heidelberg, 2016, p. 498–514. doi:10.1007/978-3-319-46523-4_30.

[15] 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, WIMS '17, Association for Computing Machinery, New York, NY, USA, 2017. doi:10.1145/3102254.3102279.

[16] A. A. Taweel, H. Paulheim, Towards Exploiting Implicit Human Feedback for Improving RDF2vec Embeddings, 2020. arXiv:2004.04423.

[17] S. Colucci, F. M. Donini, V. Schena, F. Scioscia, E. D. Sciascio, LISE: A Logic-Based Interactive Similarity Explainer for Clusters of RDF Data, IEEE Access 13 (2025) 90109–90128. doi:10.1109/ACCESS.2025.3571518.

[18] S. Colucci, F. M. Donini, V. Schena, LISE: a logic-based interactive similarity explainer, in: SeWebMeDA-2025: 8th International Workshop on Semantic Web Solutions for Large-scale Biomedical Data Analytics, 2025.

[19] G. Mai, K. Janowicz, B. Yan, Support and Centrality: Learning Weights for Knowledge Graph Embedding Models, in: C. Faron Zucker, C. Ghidini, A. Napoli, Y. Toussaint (Eds.), Knowledge Engineering and Knowledge Management, Springer International Publishing, Cham, 2018, pp. 212–227.

[20] K. Toutanova, D. Chen, Observed versus latent features for knowledge base and text inference, in: A. Allauzen, E. Grefenstette, K. M. Hermann, H. Larochelle, S. W.-t. Yih (Eds.), Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality, Association for Computational Linguistics, Beijing, China, 2015, pp. 57–66. doi:10.18653/v1/W15-4007.

[21] T. Dettmers, P. Minervini, P. Stenetorp, S. Riedel, Convolutional 2d knowledge graph embeddings, in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI'18/IAAI'18/EAAI'18, AAAI Press, 2018.

[22] W. Xiong, T. Hoang, W. Y. Wang, DeepPath: A Reinforcement Learning Method for Knowledge Graph Reasoning, in: M. Palmer, R. Hwa, S. Riedel (Eds.), Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 564–573. doi:10.18653/v1/D17-1060.

[23] B. Steenwinckel, G. Vandewiele, T. Agozzino, F. Ongenae, pyRDF2Vec: A Python Implementation and Extension of RDF2Vec, in: The Semantic Web: 20th International Conference, ESWC 2023, Hersonissos, Crete, Greece, May 28–June 1, 2023, Proceedings, Springer-Verlag, Berlin, Heidelberg, 2023, p. 471–483. doi:10.1007/978-3-031-33455-9_28.

[24] J. Portisch, N. Heist, H. Paulheim, Knowledge graph embedding for data mining vs. knowledge graph embedding for link prediction – two sides of the same coin?, Semant. Web 13 (2022) 399–422. doi:10.3233/SW-212892.

[25] W. Webber, A. Moffat, J. Zobel, A similarity measure for indefinite rankings, ACM Trans. Inf. Syst. 28 (2010). doi:10.1145/1852102.1852106.