Augmenting Knowledge Graphs for Better Link Prediction

Jiang Wang, Filip Ilievski, Pedro Szekely, Ke-Thia Yao

IJCAI 2022

报告人：石智超

2022 年 11 月 28 日

Background

目前 KG 上的 link prediction(LP) 方法大多只关注了 entity→entity，而对于 KG 中另一种事实陈述 entity→literal 缺乏考虑。

KG 中的 literal 字段，例如数据类型 quantities 和时间类型 dates 字段，对于 LP 任务建立情境十分重要。例如人物实体的出生日期字段、公司实体的成立时间字段等能有效限制 LP 情境；国家人口数量能一定程度暗示国家大小等。

现有的部分将 literal 嵌入 embedding 的方法，通过向打分函数添加 literal 相关项或者修改模型损失函数来平衡对 KG 结构和 literal 信息捕获程度。但这些方法要么引入额外参数，要么需要针对模型进行特殊修改，缺乏可伸缩性和泛化性。

Contributions

KGA:

提出方法 **K**nowledge **G**raph **A**ugmentation，引入数量型字段 Quantities 和时间型字段 Dates。该方法可作为现存任意 KGE 模型的预处理步骤，在不需要针对模型修改情况下提升模型 LP 任务上的表现。

DWD:

提出一个 LP 任务的比现有大几个数量级的 benchmark，解决 LP 评估阶段的大规模和过拟合挑战。

KGA 的离散化步骤包括两个部分：

区间间隔划分：包括 2 种划分策略：

- **fixed**：依据数值极差平均划分，不同区间极差相同
- **quantile**：依据实体数量平均划分，不同区间从属的实体数量相同。

区间层级划分：包括 3 种划分策略：

- **single**：仅创建一组不相交区间
- **overlapping**：逐个融合相邻区间，将原先不相交区间融合为存在重叠部分的区间。
- **hierarchy**：逐层划分为更细粒度的区间，第 l 层共有 b^l 个区间。

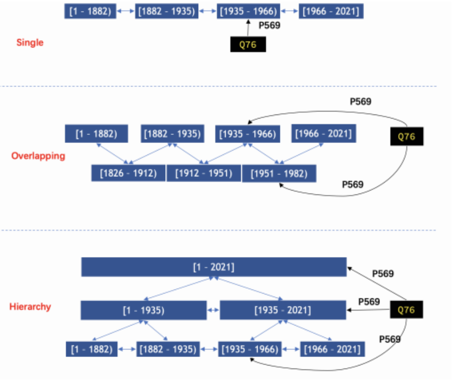

Figure 1: Single (top), overlapping (middle), and hierarchy (bottom) binning for the attribute triple (Q76, P569, “1961”), which specifies the year of birth of Barack Obama. We use $b = 4$ bins, quantile-based binning, and chaining in each level mode.

KGA: 图数据增强

KGA 离散化步骤得到的区间作为实体节点加入原始图谱 G。图数据增强步骤还通过两种操作向 G 添加新链接边：

- **区间链接**：增加区间 b_i 到 b_{i-1}、b_i 到 b_{i+1} 的链接。对于 hierarchy 策略划分的区间，除了同粒度区间链接，纵向层级上 b_i 与来源区间、b_i 与细分区间之间也增加链接。
- **区间赋予**：对于 G 中存在的三元组 (e, a, v)，若数值 $v \in b$，则添加新三元组 (e, a, b)。

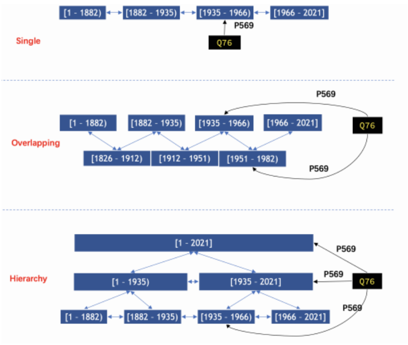

Figure 1: Single (top), overlapping (middle), and hierarchy (bottom) binning for the attribute triple (Q76, P569, "1961"), which specifies the year of birth of Barack Obama. We use $b = 4$ bins, quantile-based binning, and chaining in each level mode.

Datasets:

- **FB15K-237、YAGO15K**
- **DWD**：DARPA Wikidata，Wikidata 的一个子数据集，排除了 Wikidata 中具有显著领域特征的数据，例如综述文章、学术文章、化学化合物等。

dataset	FB15K-237	YAGO15K	DWD
# Entities	14,541	15,136	42,575,933
# Relations	237	32	1,335
# Triples	310,116	98,308	182,246,241
# Attributes	116	7	565
# Literals	29,220	23,520	31,925,813

Evaluation: MRR, Hits@K

Experiments: Entity Link Prediction Results

method	FB15K-237 MRR	FB15K-237 H@1	FB15K-237 H@10	YAGO15K MRR	YAGO15K H@1	YAGO15K H@10
TransE	0.315	0.217	0.508	0.459	0.376	0.615
+LiteralE	0.315	0.218	0.504	0.458	0.376	0.612
+KBLN	0.308	0.210	0.496	0.466	0.382	0.621
+KGA	0.321	0.223	0.516	0.470	0.387	0.623
DistMult	0.295	0.212	0.463	0.457	0.389	0.585
+LiteralE	0.309	0.223	0.481	0.462	0.396	0.587
+KBLN	0.302	0.220	0.470	0.449	0.377	0.581
+KGA	0.322	0.233	0.502	0.472	0.402	0.606
ComplEx	0.288	0.205	0.455	0.441	0.370	0.572
+LiteralE	0.295	0.212	0.462	0.443	0.375	0.570
+KBLN	0.293	0.213	0.451	0.451	0.380	0.583
+KGA	0.305	0.219	0.478	0.453	0.380	0.591
ConvE	0.314	0.226	0.488	0.470	0.405	0.597
+LiteralE	0.317	0.230	0.489	0.475	0.408	0.601
+KBLN	0.305	0.219	0.479	0.474	0.408	0.600
+KGA	0.329	0.239	0.512	**0.492**	**0.427**	0.616
RotatE	0.324	0.232	0.506	0.451	0.370	0.605
+LiteralE	0.329	0.237	0.512	0.475	0.400	0.619
+KBLN	0.314	0.222	0.500	0.469	0.393	0.613
+KGA	0.335	0.242	0.521	0.473	0.392	**0.626**
TuckER	0.354	0.263	0.536	0.433	0.360	0.571
+LiteralE	0.353	0.262	0.536	0.421	0.348	0.564
+KBLN	0.345	0.253	0.530	0.420	0.349	0.556
+KGA	**0.357**	**0.265**	**0.540**	0.454	0.380	0.592

Table 2: LP results on FB15K-237 and YAGO15K. We compare KGA to the original model (-), and the baselines LiteralE and KBLN. We report the reproduced results for all baseline methods, and provide the original results in the appendix. For KGA, we show the best results across discretization strategies (single, overlapping, hierarchy) and numbers of bins (2, 4, 8, 16, 32). We bold the best overall result per metric, and underline the best result per model.

Method	TransE MRR	TransE H@10	DistMult MRR	DistMult H@10	ComplEx MRR	ComplEx H@10
-	0.580	0.762	0.559	0.740	0.568	0.746
Quantity	0.582	0.764	0.564	0.744	0.571	0.748
Year	0.580	0.763	0.562	0.744	0.569	0.747
KGA	**0.583**	**0.764**	**0.566**	**0.746**	**0.574**	**0.751**

Table 3: LP results on DWD. We show the performance (MRR and Hits@10) of the vanilla embedding model (-), and KGA with binned quantities, with years, and the full KGA. We use 32-bin KGA with QOC (quantile, overlapping, and chaining) discretization.

Experiments: Ablation Study

KGA	TransE MRR	TransE H@10	DistMult MRR	DistMult H@10	ComplEx MRR	ComplEx H@10	ConvE MRR	ConvE H@10	RotatE MRR	RotatE H@10	TuckER MRR	TuckER H@10
-	0.315	0.508	0.295	0.463	0.288	0.455	0.314	0.488	0.324	0.506	0.354	0.536
FSC	0.317	0.509	0.301	0.471	0.291	0.459	0.320	0.494	0.328	0.513	0.354	0.536
FOC	0.318	0.512	0.306	0.482	0.296	0.466	0.319	0.494	0.327	0.511	0.354	0.536
FHC	0.318	0.511	0.304	0.478	0.299	0.469	0.320	0.495	0.327	0.510	0.354	0.535
FON	0.319	0.510	0.305	0.480	0.296	0.464	0.321	0.498	0.328	0.513	0.353	0.535
QSC	0.320	0.513	0.303	0.475	0.296	0.465	0.319	0.494	0.330	0.516	**0.357**	0.540
QOC	0.321	0.513	0.312	0.487	0.299	0.471	0.322	0.499	0.332	0.517	0.356	**0.542**
QHC	**0.321**	**0.516**	**0.322**	**0.502**	**0.305**	**0.478**	**0.329**	**0.512**	**0.335**	**0.521**	0.356	0.538
QON	0.320	0.514	0.309	0.480	0.299	0.468	0.321	0.498	0.332	0.516	0.355	0.536

Table 4: Ablation study on modes of graph augmentation with link prediction on FB15K-237. KGA variants: '-' represents the original graph (no augmentation), F = Fixed Size, Q = Quantile, S = Single, O = Overlapping, H = Hierarchy, C = Chaining, N = No Chaining. The best result for each column is marked in bold. We show the best results among the different numbers of bins (2, 4, 8, 16, 32).

model	**2**	**4**	**8**	**16**	**32**
TransE	0.321	0.320	0.321	**0.321**	0.321
DistMult	0.306	0.308	0.314	0.317	**0.322**
ComplEx	0.294	0.295	0.300	0.304	**0.305**
ConvE	0.321	0.320	0.325	0.325	**0.329**
RotatE	0.327	0.326	0.332	**0.335**	0.334
TuckER	0.354	0.356	0.355	0.356	**0.357**

Table 5: Effect of bin size on the performance of different models on FB15K-237. We show results for the best discretization strategy. We experiment with 2, 4, 8, 16, and 32 bins. Numbers indicate MRR.

Experiments: Numeric Link Prediction Results

		KGA	NAP++	MrAP
FB15K-237	date_of_birth	18.9	22.1	**15.0**
	date_of_death	20.6	52.3	**16.3**
	film_release	**4.0**	9.9	6.3
	org_founded	**49.0**	59.3	58.3
	location_founded	**76.0**	92.1	98.8
	latitude	2.1	11.8	**1.5**
	longitude	7.1	54.7	**4.0**
	area	**6.1e4**	4.4e5	4.4e5
	population	**4.0e6**	7.5e6	2.1e7
	height	**0.077**	0.080	0.086
	weight	**11.6**	15.3	12.9
YAGO15K	date_of_birth	**16.3**	23.2	19.7
	date_of_death	**30.8**	45.7	34.0
	date_created	**58.2**	83.5	70.4
	data_destroyed	**23.3**	38.2	34.6
	date_happened	**29.9**	73.7	54.1
	latitude	3.4	8.7	**2.8**
	longitude	7.2	43.1	**5.7**

Table 6: Performance of our numeric predictor with different choices of base model on graph augmented with 32-bin QOC, when compared to existing SOTA methods on the FB15K-237 and YAGO15K dataset. Numbers indicate MAE. Values of NAP++ and MrAP are taken from [Bayram *et al.*, 2021]. We show results for KGA with TransE for a fair comparison to NAP++.

attribute	Median	LR	KGA
Elo rating	119.03	86.09	**55.20**
declination (degree)	18.68	**9.83**	18.53
elevation above sea level	466.51	**366.64**	459.48
right ascension (degree)	82.98	**40.90**	82.51
apparent magnitude	3.02	**2.00**	2.37
date of birth	62.71	**49.70**	58.59
date of death	90.68	**78.10**	79.38
publication date	28.33	**17.37**	28.27
inception	72.84	**61.45**	72.27
point in time	88.76	**81.65**	83.70

Table 7: Performance of our numeric predictor KGA-QOC on DWD compared to a linear regression (LR) model and a median baseline. We use 32 bins for both quantities and years. Numbers indicate MAE reduction percentages against a median value baseline. We report results for the most populous 5 properties for both quantities and years, with identifiers: P1087, P6258|Q28390, P2044|Q11573, P6257|Q28390, P1215, P569, P570, P577, P571, and P585.

Thanks