Searching for Better Knowledge Graph Completion Evaluation Metrics : Recent Discoveries

2025-02-13

presenter

Sooho Moon

DMAIS

INDEX

- **Introduction**
- **Related works**
- **Unified metric framework**
- **Questions for the future**

Part 1

Part 2

Introduction

■ Background of knowledge graph(KG)

- □ A KG contains triplets in the form of *(subject entity, relation, object entity)*
- □ Useful in variety of domains(health care, information retrieval, RAG, recommender system, etc.)
- □ **Properties of KG**
  - ■ **Incompleteness**(link prediction proposed to solve this problem)
  - ■ **Long tail distribution of entities**

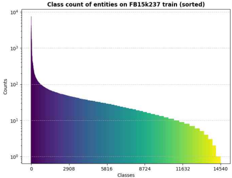
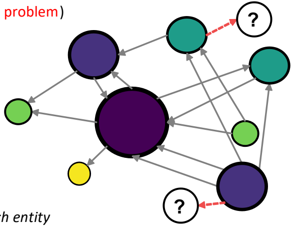


Fig 1.
Number of triplets for each entity on FB15k237 dataset(log scale)

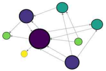


* The heat map color is based on sorted class index number for better visualization and interpretability.
** We only think about entity prediction(link prediction) in this work.

Introduction

- **Evaluation of knowledge graph link prediction**
  - □ Calculate ranks(relative confidence among other entities) for evaluation
  - □ **M**ean **R**ank, **M**ean **R**eciprocal **R**ank, **Hits@k** are the most used rank based metrics

Training data(KG) → **model** → ***Evaluation***



DJ Trump $\xrightarrow{isFrom}$ **?**

rank	score	entity
1	0.99	USA
2	0.43	North America
3	0.12	Canada
....	....	....
|E|	0.001	Orange

$$\mathrm{MR}(\downarrow) = \sum_i rank_i$$

$$\mathrm{MRR}(\uparrow) = \sum_i \frac{1}{rank_i}$$

$$\mathrm{Hits@k}(\uparrow) = \sum_i \mathbb{I}[rank_i \leq k]$$

Introduction

- **When KGC and long tail comes together...**
  - □ Training models to find correct links for frequently seen entities(head class) were easy
  - □ However, **results on tail class were not satisfactory**

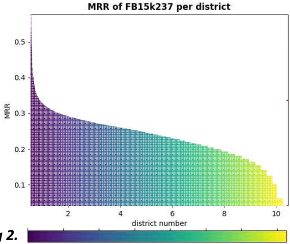
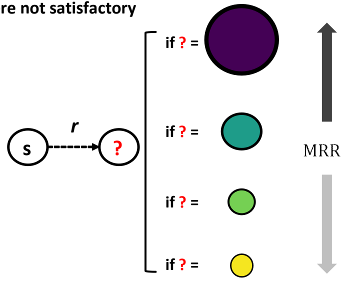


Fig 2.
Link prediction results of RotatE



$$\mathrm{MRR} = \sum_i \frac{1}{rank_i}$$

* Each district contains an equal number of entity classes. Classes that appear in smaller number districts have more training instances than bigger number districts. The dots represent the average MRR for each district.

Introduction

- **When KGC and long tail comes together...**
  - □ Vanila mean based ranking metrics **bias** toward the performance of high populated head class
  - □ **This is extremely unfair**

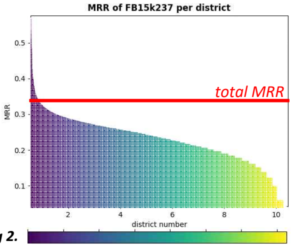


Fig 2.
Link prediction results on RotatE(ICLR' 19)

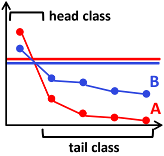
10% of entity classes take over almost **50% of the total training population**

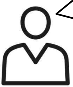
the other **90%** of entity classes are below the total metric

But, why should we care about metrics producing unbalanced evaluation?

Introduction

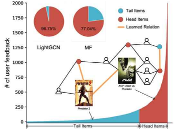
- **When a model is evaluated solely on biased conventional metrics**





Model A is better!! I should use model A.

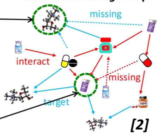
KG for Recommendation



can't recommend proper items to users, "popularity bias" "overfit to head class" [1]

Medical KGC

Medical Knowledge Graph

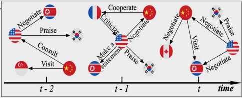


predictions on sparsely linked drugs, diseases, materials will be poor

[2]

Temporal KGC

[3]

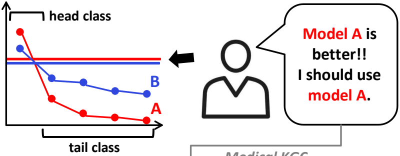


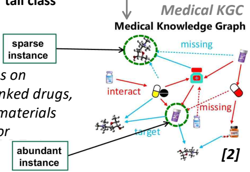
completion on newly added entities in Temporal KG will be poor

[1] Wei, C., Liang, J., Liu, D., Dai, Z., Li, M., & Wang, F. (2023). Meta Graph Learning for Long-tail Recommendation. Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining,2023,1–10.
[2] F. Gong, F., Wang, M., Wang, H., Wang, S., & Liu, M. (2021). SMR: Medical Knowledge Graph Embedding for Safe Medicine Recommendation. Big Data Research, 24, 100174.
[3] Xu, Y., Ou, J., Xu, H., & Fu, L. (2023). Temporal Knowledge Graph Reasoning with Historical Contrastive Learning. Proceedings of the AAAI Conference on Artificial Intelligence, 37(4), 4765–4773.

Introduction

- **When a model is evaluated solely on biased conventional metrics**



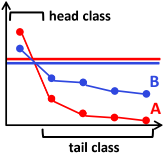


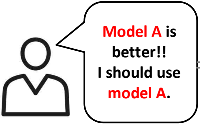
predictions on sparsely linked drugs, diseases, materials will be poor

[1] Wei, C., Liang, J., Liu, D., Dai, Z., Li, M., & Wang, F. (2023). Meta Graph Learning for Long-tail Recommendation. Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining,2023,1–10.
[2] F. Gong, F., Wang, M., Wang, H., Wang, S., & Liu, M. (2021). SMR: Medical Knowledge Graph Embedding for Safe Medicine Recommendation. Big Data Research, 24, 100174.
[3] Xu, Y., Ou, J., Xu, H., & Fu, L. (2023). Temporal Knowledge Graph Reasoning with Historical Contrastive Learning. Proceedings of the AAAI Conference on Artificial Intelligence, 37(4), 4765–4773.

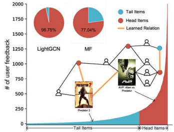
Introduction

- **When a model is evaluated solely on biased conventional metrics**





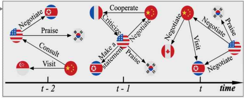
KG for Recommendation



can't recommend proper items to users, "popularity bias" "overfit to head class" [1]

Temporal KGC

[3]



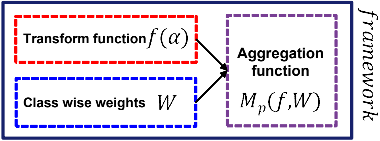
completion on newly added entities in Temporal KG will be poor

[1] Wei, C., Liang, J., Liu, D., Dai, Z., Li, M., & Wang, F. (2023). Meta Graph Learning for Long-tail Recommendation. Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining,2023,1–10.
[2] F. Gong, F., Wang, M., Wang, H., Wang, S., & Liu, M. (2021). SMR: Medical Knowledge Graph Embedding for Safe Medicine Recommendation. Big Data Research, 24, 100174.
[3] Xu, Y., Ou, J., Xu, H., & Fu, L. (2023). Temporal Knowledge Graph Reasoning with Historical Contrastive Learning. Proceedings of the AAAI Conference on Artificial Intelligence, 37(4), 4765–4773.

Introduction

- **Creating a new metric is non trivial**
  - □ Absence of general framework makes metric creation, comparison subjective and ambiguous
    - ■ Which metric is more sensitive to rank change?
    - ■ Which metric is more(less) swayed to outlier ranks?
    - ■ Which metric is more(less) swayed to head class performance?
  - □ Thus establishing a general metric framework that can also represent conventional metrics will enable more reliable and consistent analysis across different metrics

$\{\text{MR, MRR, Hits@k, ...}\} \subset$



Introduction

- **The need for more fair evaluations**
  - □ Not only for practical issues, but to guide KGC research to a better way
  - □ To acquire the solutions, we investigate the below research questions

- **Research questions**

  RQ1 “Will our metric produce good evaluation according to the desired evaluation objective?”

  RQ2 “Can we calculate how sensitive framework parameters are?”

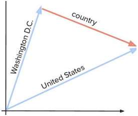
  RQ3 “Can we efficiently reduce calculation overhead under the framework”

Related works

■ Knowledge graph completion

□ **Embedding based models**

- ■ Learn individual embeddings for entities, relations($\mathbf{h} \times \mathbf{r} \approx \mathbf{t}$)
- ■ Translation based : TransE(NeurIPS' 13), RotatE(ICLR' 19), HousE(ICML' 22)
- ■ Tensor decomposition based : ComplEx(ICML' 16), TuckER(EMNLP' 19)



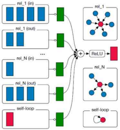
□ **Logical rule based models**

- ■ Mine sequence of relations for completion($r_h = r_i \rightarrow r_j \rightarrow r_k$)
- ■ NeuralLP(NeurIPS' 17), DRUM(NeurIPS' 19), RNNLogic(ICLR' 21)

$$X \xrightarrow{\texttt{brotherOf}} Z \xrightarrow{\texttt{sisterOf}} Y \quad \Rightarrow \quad X \xrightarrow{\texttt{brotherOf}} Y$$

□ **GNN based models**

- ■ Utilize GNNs for KGC
- ■ R-GCN(ESWC' 18), KBGAT(ACL' 19), AdaProp(KDD' 23)



Related works

KGC evaluation

- □ Apart from building models, a wide range of concerns for KGC model evaluation emerged
  - ■ Dataset size invariant metric and proposing general metric framework
  - ■ How different tie breaking protocals will affect the ranking evaluation
  - ■ Shortcomings of MRR under the OWA

KGC evaluation(works addressing long-tail evaluation problem)

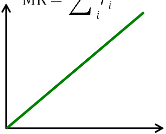
- □ Works specifically for creating metrics to tackle biased entities, relations
  - ■ Incorporating popularity of entities, relations for unbiased metrics(strat-MRR, strat-Hits@k)

Unified metric framework

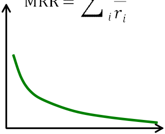
Rank based metrics in the context of OWA(Open World Assumption)

- KGs adopt OWA to address the incompleteness
  - **present link = true fact**, **absent link = don't know whether it is true or false**
- Thus F1, ROC-AUC are not confidently calculable due to the absent **TN** & **FN**, which led to exclusive use of **rank based metrics** in KGC literature [1]
- Most commonly used conventional metrics(**Mean Rank, Mean Reciprocal Rank, Hits@k**)

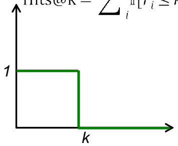
$$\mathrm{MR} = \sum_i r_i$$



$$\mathrm{MRR} = \sum_i \frac{1}{r_i}$$



$$\mathrm{Hits@k} = \sum_i \mathbb{I}[r_i \leq k]$$



x-axis : rank
y-axis : transformed rank

[1] Berrendorf, M., Faerman, E., & Tresp, V. (2022). A unified framework for rank-based evaluation metrics for link prediction in knowledge graphs. In WWW '22 (pp. 1923–1933). ACM.

Unified metric framework

- **Overall view of unified metric framework(UMF)**
  - □ Three parts(rank transformation function, class wise weights, aggregation function)

$r_1, r_2, \vdots, r_n \rightarrow f \rightarrow f(r_1), f(r_2), \vdots, f(r_n)$

$g \rightarrow w_c \in W$

$$\sqrt[p]{\sum_{c}^{\mathbb{C}} w_c \mu_c} \rightarrow M_p(f, W) \in \mathbb{R}$$

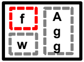
generalized weighted mean

$$\mu_c = \frac{1}{n_c} \sum_{i}^{n_c} f(r_i)^p$$

$$c \in \mathbb{C} \qquad \sum_{c}^{\mathbb{C}} w_c = 1$$

* For p=0, the power mean equals to weighted geometric mean

Unified metric framework



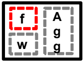
Transformation function(f)

- □ Input : raw rank // Output : mapped rank
- □ Roles : normalize raw ranks, controls the metric's range
- □ Properties : **bound**, **focus-on-top rate(FOTR)**
- □ **f(x)** can be written as power over rank

$$f(x = rank_i) = x^{\alpha} \quad \left[\begin{array}{l} \mathrm{MR}(\alpha = 1) \\ \\ \mathrm{MRR}(\alpha = -1) \\ \\ \cancel{\mathrm{Hits@k}}(\text{non discriminative}) \end{array}\right.$$

what $\boldsymbol{\alpha}$ *should we use and when?*

Unified metric framework



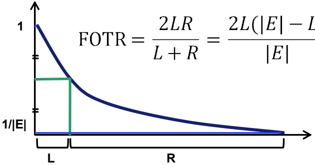
Transformation function(f) and FOTR

- □ [1] mathematically proved that under the OWA, FOT **f(x)** under-evaluates true power of KGC models(**criticized MRR, sugested using less FOT f(x)**)
- □ e.g., **MRR(α=-1)** change **0.5** for rank 2 → rank 1**(1 improvement)**, change **0.1999** for rank 10000 → rank 5**(9995 improvement)**
- □ 'less FOT **f(x)**' accomodates more non-top rank improvement than FOT **f(x)** such as **MRR**

[1] Rethinking Knowledge Graph Evaluation Under the Open-World Assumption (NeurIPS' 22)

Unified metric framework

- **Transformation function(f) and FOTR**
  - □ [1] mathematically proved that under the OWA, FOT **f(x)** under-evaluates true power of KGC models(**criticized MRR, sugested using less FOT f(x)**)
  - □ Value of **MRR(α=-1)** change **0.5** for rank 2 → rank 1(**1 improvement)**, change **0.1999** for rank 10000 → rank 5(**9995 improvement)**
  - □ 'less FOT **f(x)**' accomodates more non-top rank improvement than FOT **f(x)** such as **MRR**
    But how much is 'less'?



$$\mathrm{FOTR} = \frac{2LR}{L+R} = \frac{2L(|E|-L)}{|E|}$$

bigger FOTR indicates less FOT

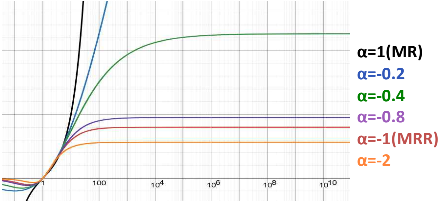
smaller FOTR indicates more FOT

[1] Rethinking Knowledge Graph Evaluation Under the Open-World Assumption (NeurIPS' 22)

Unified metric framework

- **FOTR of f(x) on different α**

*** Further investigation on “which α suits best for the dataset” will be done in the future**



Unified metric framework

- **Properties over f(x)**

Lemma 1 *For non negative real numbers with* $a \le b,\ f(x) \in [a, b]$ *implies* $M_p(f, W) \in [a, b]$ *for* $\forall p \in \mathbb{R}$

Corollary 1.1 *Range of* $M_p(f, W)$ *is only determined by range of* $f(x)$

As long as $\sum_c^{\mathbb{C}} w_c = 1$ *is satisfied, the above holds.*

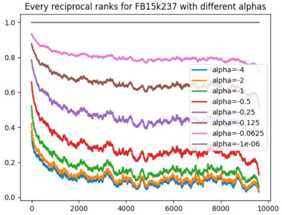
Thus selection of W *and* p *are independent to the range of the metric.*

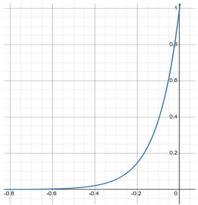
Corollary 1.2 $M_p(f, W) \in (0, 1]$ *iff* $\alpha \le 0$

Thus $M_p(f, W)$ *gains a fixed optimum and fixed pessimum(almost zero).*

Unified metric framework

- **So, use bigger α for 'less-FOT'?**
  - □ The transformation function was able to generalize MR and MRR
  - □ However, **in the case for α<0**, the lower bound of **f(x)** keep rising as α approaches 0
  - □ **This results in unavoidable optimistic evaluation for bigger α**





Applied dataset = FB15k237
|E|=14541

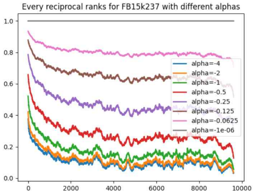
x-axis : α
*y-axis : lower bound of **f(x)***
=|E|^α

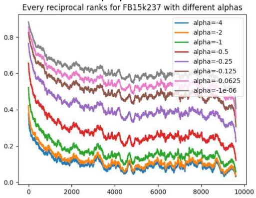
* In the left figure, x-axis corresponds to entity classes sorted by training instance count. Each y point corresponds to the MRR of class(x point).

Unified metric framework

- **Alter f(x) so that point (1, 1), (|E|, 0) are crossed**
  - □ Intuitively, a metric should be 0 if the predicted rank is the **worst possible** (rank=|E|)
  - □ **No other works pointed out this problem**, we can newly define **f(x) for α<0**

$$f(x) = x^{\alpha} \longrightarrow f(x) = \left(\frac{1}{1-|E|^{\alpha}}\right)(x^{\alpha}-1)+1$$





Unified metric framework

- **With the new f(x), hinder previous properties?**
  - No

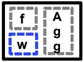
$$f(x) = \left(\frac{1}{1-|E|^{\alpha}}\right)(x^{\alpha}-1)+1 \approx x^{\alpha} \text{ for } \alpha < 0$$

$$\mathrm{FOTR}\left(\left(\frac{1}{1-|E|^{\alpha}}\right)(x^{\alpha}-1)+1\right) = \mathrm{FOTR}(x^{\alpha}) \text{ for } \alpha \neq 0$$

Corollary 1.2 $M_p(f, W) \in [0, 1]\ iff\ \alpha \leq 0$

Thus $M_p(f, W)$ *gains a fixed optimum and fixed pessimum*~~(almost zero)~~.

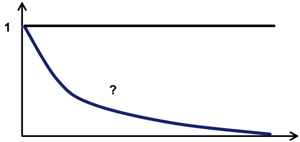
Unified metric framework



Class wise weights(W)

- □ Input : entity class wise information // Output : weights of each class
- □ Role : define the reletive importance of each class
- □ Property(?) : **expected entropy, cumulative gain**
- □ Sufficient amount of analysis was not done in this part of the framework...

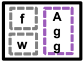
conventional weighting protocals can be seen as CDF on uniform distribution



$$\mathrm{w}_c = \frac{P(X \leq n_c)}{n_{eval}}$$

count of entity class in evaluation data

Unified metric framework

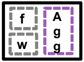


- **Final aggregation(weighted power mean)**
  - □ Input : **f, W** // Output : single real value
  - □ Role : define the reletive importance of each class
  - □ Property : **degree of emphasis on individual f(x)**

$$\sqrt[p]{\sum_{c}^{\mathbb{C}} w_c \mu_c} \quad \left[\begin{array}{l} (p=\infty)\ max_i f(r_i) \\ \quad\vdots \\ (p=1)\ arithmetic\ mean \\ (p=0)\ geometric\ mean \\ (p=-1)\ harmonic\ mean \\ \quad\vdots \\ (p=-\infty)\ min_i f(r_i) \end{array}\right.$$

$$\mu_c = \frac{1}{n_c} \sum_{i}^{n_c} f(r_i)^p$$

Unified metric framework



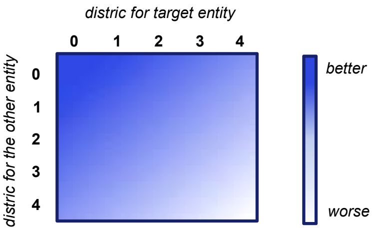
- **Final aggregation(weighted power mean)**
  - □ **p** controls how much attention should the metric give to larger **f(x)**
  - □ If W is for ‘class wise weight’, p is for ‘transformed rank weight’
  - □ Able to see this property intuitively...(more works needed for Agg part also...)

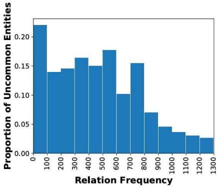
$$\sqrt[p]{\sum_{c}^{\mathbb{C}} w_c \mu_c} \quad \left[\begin{array}{l} (p=\infty)\ max_i f(r_i) \\ \quad \vdots \\ (p=1)\ arithmetic\ mean \\ (p=0)\ geometric\ mean \\ (p=-1)\ harmonic\ mean \\ \quad \vdots \\ (p=-\infty)\ min_i f(r_i) \end{array} \right.$$

$$M_i(f,W) \leq M_j(f,W)\ for\ i<j$$

Questions for the future

- Will other **two parts(entity, relation, target_entity) also effect the prediction quality?**





[1]

Figure 1: A histogram about relation frequencies and the corresponding proportions of uncommon entities in DBpedia.

[1] Tackling Long-Tailed Relations and Uncommon Entities in Knowledge Graph Completion(EMNLP' 19)

Questions for the future

□ **Cold start anomoly, worse than guessing?**

|E|= 14000, batch_size = 1024, max_step=100000, |train|=270000

0 step : MR[6969.577, 7115.239, ... , 6799.429, 6957.469]

10000 step : MR[**2546.239, 4430.743,** ... , **9175.899, 10824.321**]

Funny thing : From what we observed, **RotatE** only has this property (even with the best hyper-parameter setting). **This doesn't happen in ComplEx, TuckER, HousE**.

Questions for the future

□ **By tuning α, is it possible to create a dataset-size invariant metric [1, 2]?**

[1] 'Size invariant' desiderata for **evaluating the objective power of the model**

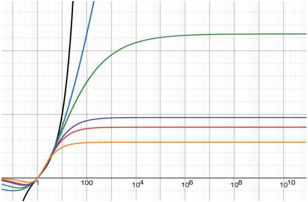


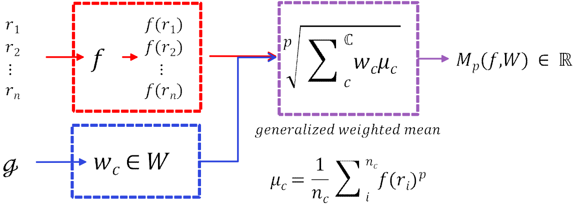
Table 1: Desiderata for rank-based metrics

Property	Constraint	MR	MRR	H_k
Non-negativity	$\forall r \in \mathbb{N} : f(r) \geq 0$	✔	✔	✔
Fixed optimum	$f(1) = c_{\text{opt.}}$	✘	✔	✔
Asymp. pessimum	$\lim_{r \to \infty} f(r) = c_{\text{pes.}}$	✘	✔	✔
Anti-monotonic	$r > r' \to f(r) < f(r')$	✘	✔	✘
Size invariant	$\mathbb{E}[f] \not\propto n$	✘	✘	✘

[1] A Unified Framework for Rank-based Evaluation Metrics for Link Prediction in Knowledge Graphs(WWW' 22)
[2] Interpretable and Fair Comparison of Link Prediction or Entity Alignment Methods(WI-IAT' 20)

Questions for the future

□ **Calculation overhead?**



$c \in \mathbb{C}$ $\quad \sum_c^{\mathbb{C}} w_c = 1$

Thank You!

Contact: Sooho Moon (Email: moonwalk725@cau.ac.kr)