

Project3 CFKG Algorithm

- In this task, you need to construct a Recommender System by the given user-item interaction data and knowledge graph (KG).
- The key to the Recommender System is a score function $f(u, w)$
- For the samples in the given user-item interaction data
 - While inputting a positive sample, f should return a high score
 - While inputting a negative sample or a sample not in the given data, f should return a low score.

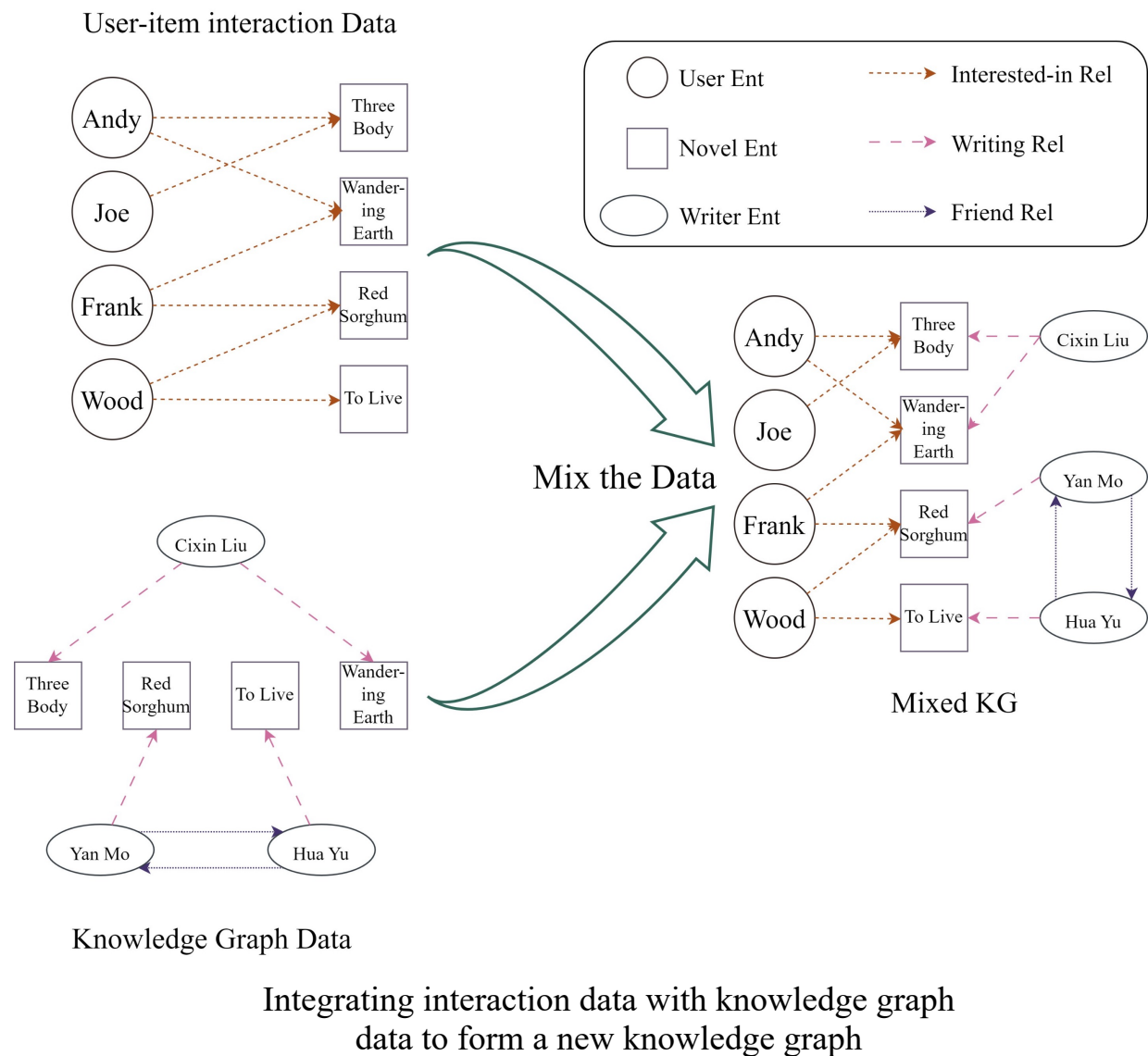
How to score a sample

- The input of the $f(u, w)$, u and w is only the index of the user/item.
- There is no information in the index of the user/item, it's impossible to give a significative score only based on the index.
- For the user u and item w , we need to find their feature vectors, and record as \mathbf{u} and \mathbf{w} .
- The calculation process of $f(u, w)$ depends on \mathbf{u} and \mathbf{w} .

- An intuitive method is designing some statistical metric as the feature, such as the correlation coefficient
 - Designing the statistical metric is difficult.
 - It's hard to use the information in the KG.
- Another method is representing the user/item as a low-dimensional real value vector (also called Embedding)
 - How to get the embedding?
 - Firstly, we define the calculation process of $f(u, w)$
 - Get the embeddings of all the users and items by the Gradient Descent method, let the return value of $f(u, w)$ for all the combinations of users and items are accordance with the given interaction data
 - Each element in the embedding can represent some information about the user/item, but we cannot explain them.
 - The information of KG can also be represented in the embedding.

How to mix information of KG and interaction data?——CFKG Algorithm

- CFKG algorithm models the interaction records as a new type of relations in the KG, there relation type is "Interested-in".
- The interaction data and origin KG will be mixed into a bigger KG
- We also can define a score function $g(h, r, t)$ for the relation in the KG:
 - Input is a relation in KG: g returns a high score
 - Input is a relation not in KG: g returns a low score
- If we define $r' = \text{Interested-in}$:
 - For a positive sample, $g(u, r', w)$ returns a high score
 - For a negative sample or a sample not in the given data, $g(u, r', w)$ returns a low score.

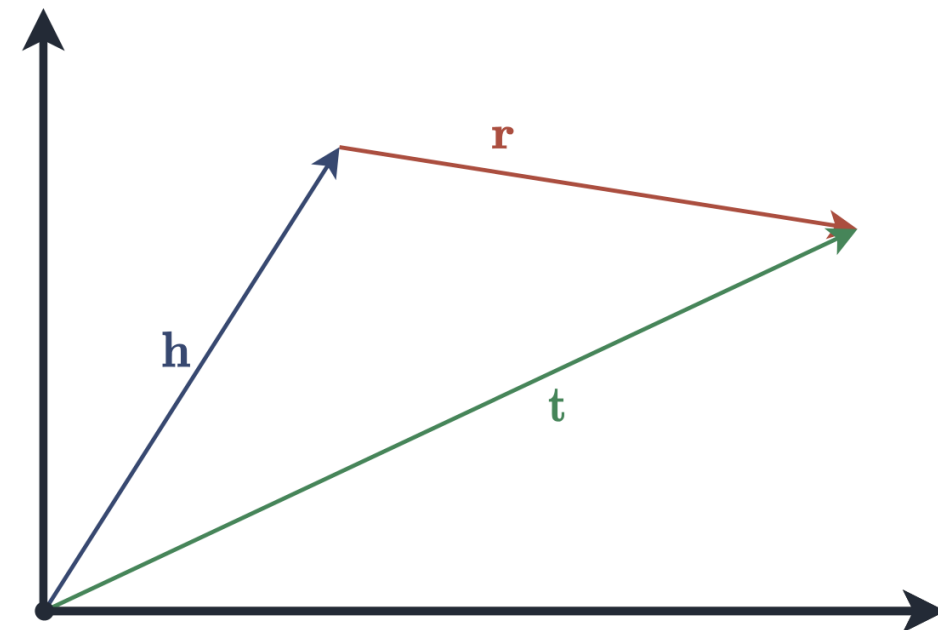


A specific design scheme of g — the TransE Algorithm

- TransE Algorithm is an example of $g(h, r, t)$
 - Entity and relation type will be represented as a vector in the d -dimensional Euclidean Space
 - Ideally, when the relational triplet (h, r, t) holds, its corresponding vector representation should satisfy :

$$\mathbf{h} + \mathbf{r} = \mathbf{t}$$

- However, the ideal situation often does not hold true.
Therefore, the likelihood of a relational triplet (h, r, t) being valid is positively correlated with $-\|\mathbf{h} + \mathbf{r} - \mathbf{t}\|$
- Therefore, $g(h, r, t) = -\|\mathbf{h} + \mathbf{r} - \mathbf{t}\|$



Schematic diagram of the calculation principle of the TransE relation model

- For the merged knowledge graph, the Embedding representations of all entities and relations should satisfy:
 - For a relation in KG: g returns a high score
 - For a relation not in KG: g returns a low score
- Then ideally, for the merged knowledge graph \mathcal{G}' , there is :

$$\forall (h, r, t) \in \mathcal{G}' \forall (h', r', t') \notin \mathcal{G}', g(h, r, t) > g(h', r', t')$$

- Which means : $\sum_{(h,r,t) \in \mathcal{G}'} \sum_{(h',r',t') \notin \mathcal{G}'} [g(h', r', t') - g(h, r, t)]_+ = 0$, where $[x]_+ = \max(x, 0)$
- There is a loss function : $\min \sum_{(h,r,t) \in \mathcal{G}'} \sum_{(h',r',t') \notin \mathcal{G}'} [g(h', r', t') - g(h, r, t)]_+$
- If we require that $g(h, r, t)$ be at least γ higher than $g(h', r', t')$, then the final form of the loss function is:

$$\min \sum_{(h,r,t) \in \mathcal{G}'} \sum_{(h',r',t') \notin \mathcal{G}'} [g(h', r', t') - g(h, r, t) + \gamma]_+$$

Reference

1. Y. Zhang, Q. Ai, X. Chen, and P. Wang, “Learning over Knowledge-Base Embeddings for Recommendation.,” *arXiv*, vol. abs/1803.06540, 2018, [Online]. Available: <http://arxiv.org/abs/1803.06540>
2. A. Bordes, N. Usunier, A. García-Durán, J. Weston, and O. Yakhnenko, “Translating Embeddings for Modeling Multi-relational Data.,” in *Advances in Neural Information Processing Systems 26*, 2013, pp. 2787–2795. [Online]. Available: <https://proceedings.neurips.cc/paper/2013/hash/1cecc7a77928ca8133fa24680a88d2f9-Abstract.html>