CSC17103 – KHAI THÁC DỮ LIỆU ĐỒ THỊ

HOMEWORK 04: Link Prediction & Graph Embedding

Knowledge Graph

A knowledge graph is a structured representation of knowledge that captures relationships between entities in a particular domain. It is a way to organize and connect information to make it more accessible, useful, and meaningful for machines and humans alike. In a knowledge graph, information is represented as nodes and edges connecting these nodes.

Here are some key characteristics of a knowledge graph:

- Nodes (Entities): Entities are the individual pieces of information or concepts represented in the graph. These can be anything from people, places, organizations, products, to abstract concepts, events, and more.
- Edges (Relationships): Edges represent the connections or relationships between entities. These relationships describe how entities are related or linked to each other. For example, "born in," "works for," "is married to," etc.

Link Prediction

Link prediction in knowledge graphs is a technique used to predict missing or potential relationships (links) between entities in the graph. Since knowledge graphs often represent a subset of the real-world knowledge and relationships, they are rarely complete. The process of link prediction involves using the existing information in the knowledge graph to predict new edges (relationships) between nodes (entities) that are not explicitly present but are likely to exist. This can be valuable for various reasons, such as:

- Knowledge Completion: Link prediction helps in filling the gaps in the knowledge graph, thereby making it more comprehensive and informative.
- Recommendation Systems: Link prediction can help suggest potential connections between users and items to make personalized recommendations.
- Identifying Missing Data: Link prediction can be used to identify entities with potential missing attributes or properties based on their relationships with other.

Graph Embedding

Graph embedding, also known as network embedding or graph representation learning, is a technique that aims to learn low-dimensional vector representations (embeddings) of nodes, edges, or subgraphs in a graph. The goal of graph embedding is to transform the graph's complex structure into a continuous vector space while preserving essential information and capturing meaningful relationships between nodes. Graph embeddings have become popular because they enable the application of machine learning and data mining techniques on graph-structured data, allowing algorithms designed for continuous vector spaces to be used for tasks on graphs. By representing nodes as vectors in a lower-dimensional space, graph embeddings facilitate efficient and scalable processing of graph.

TransE

a) Learn and present the TransE model the way you understand it.
b) Suppose we consider a simpler loss function for the TransE model as follows:

$$\mathcal{L}_{simple} = \sum_{(h,r,t)\,\in\, S} d(\boldsymbol{h} +\ \boldsymbol{r}, \boldsymbol{t})$$

Whether the entity and relational embeddings become better than the original version after minimizing the loss function to zero. Give an example to support that statement.

c) Suppose we consider a simpler loss function for the TransE model as follows:

$$\mathcal{L}_{no\ margin} = \sum_{(h,r,t)\in S}\ \sum_{(h',r,t')\notin S} \max\left[0, d(\boldsymbol{h} +\ \boldsymbol{r}, \boldsymbol{t}) - d(\boldsymbol{h}' +\ \boldsymbol{r}, \boldsymbol{t}')\right]$$

Whether the entity and relational embeddings become better than the original version after minimizing the loss function to zero. Give an example to support that statement.

d) Do entities embedding in TransE need to be normalized to the same length? Why ?
e) Give an example of a simple graph for which no perfect embedding exists, i.e., no embedding perfectly satisfies h + r = t for all (h, r, t) ∈ S and h’ + r ≠ t’ for (h’, r, t’) ∉ S, for any choice of entity embeddings (e for e ∈ E) and relationship embeddings (r for r ∈ R). Explain why this graph has no perfect embedding in this system.