Homework 2:
构建小型领域知识图谱

大数据原理与技术 (Spring 2025)

22336226 王泓沣
Lectured by: Changdong Wang
Sun Yat-sen University

1 问题描述

- 自选数据集，说明数据来源。
- 使用 NLP 工具（如 spaCy）提取实体和关系
- 用 Neo4j 构建图谱并可视化关键节点

2 数据描述

数据由 ChatGPT 4.5 产生，共 47 条记录，内容是关于科技公司的自然语言信息，由三元组构成，包含 47 个实体（人物、公司、年份、地点），5 种关系（CEO、创始人、联合创始人、成立年份、总部所在地）

3 Method

3.1 spaCy 提取实体

使用 SpaCy 库中的英文小型预训练模型 `en_core_web_sm` 进行 nlp 任务，将原始的三元组数据归类成不同类型的实体和关系。

```
nlp.add_pipe("entity_ruler", before="ner").add_patterns(patterns)

for index, row in df.iterrows():
    doc1, doc2 = nlp(row["Entity1"]), nlp(row["Entity2"])
```

3.2 Neo4j 构建图谱

Neo4j 是基于图数据库的 NoSQL 数据库，使用节点（Node）和关系（Relationship）构建数据模型。使用 Cypher 语言进行操作。在此实验中通过计算节点度数评估节点的重要性，并赋予不同颜色。

```
def create_node(tx,name,label):
    tx.run(f"MERGE(:{label}{{name:$name}})",name = name)

def create_relationship(tx,ent1,rel,ent2):
    tx.run("""
        MERGE (a:Entity {name:$entity1})
        MERGE (b:Entity {name:$entity2})
        MERGE (a)-[:RELATION {type: $relation}]->(b)
    """, entity1 = ent1,entity2=ent2,relation = rel)

with driver.session() as session:
    for index, row in df.iterrows():
        ent1,rel,ent2 = row["Entity1"], row["Relationship"],row["Entity2"]
        session.execute_write(create_node,ent1,"Entity")
        session.execute_write(create_node,ent2,"Entity")
```

```
        session.execute_write(create_relationship,ent1,rel,ent2)

driver.close()

graph.run("""
    MATCH (n)
    SET n.degree = COUNT{(n)--()}
""")

query = """
    MATCH (a)-[r]->(b)
    RETURN a.name AS source, a.degree AS  source_degree, type(r) AS relation, b.
        name AS target, b.degree AS target_degree
"""
data = graph.run(query).data()

G = nx.DiGraph()

for item in data:
    G.add_node(item['source'], degree=item['source_degree'])
    G.add_node(item['target'], degree=item['target_degree'])
    G.add_edge(item['source'], item['target'], relation=item['relation'])

colors = []
for node in G.nodes(data=True):
    degree = node[1]['degree']
    if degree >= 4:
        colors.append('red')
    elif degree >= 2:
        colors.append('orange')
    else:
        colors.append('blue')
```

4 Result

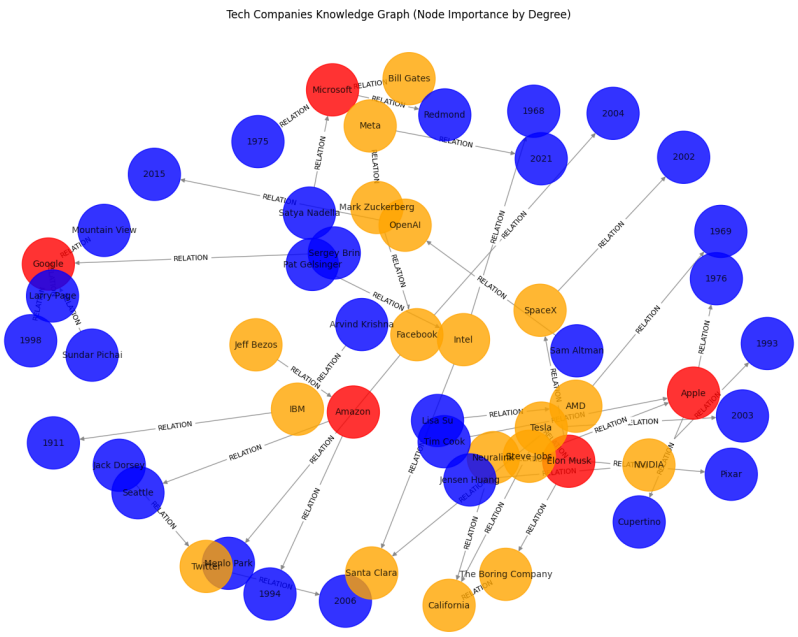

Figure 1: 知识图谱可视化结果