知识图谱可视化以及简单的问答系统

主要技术栈;

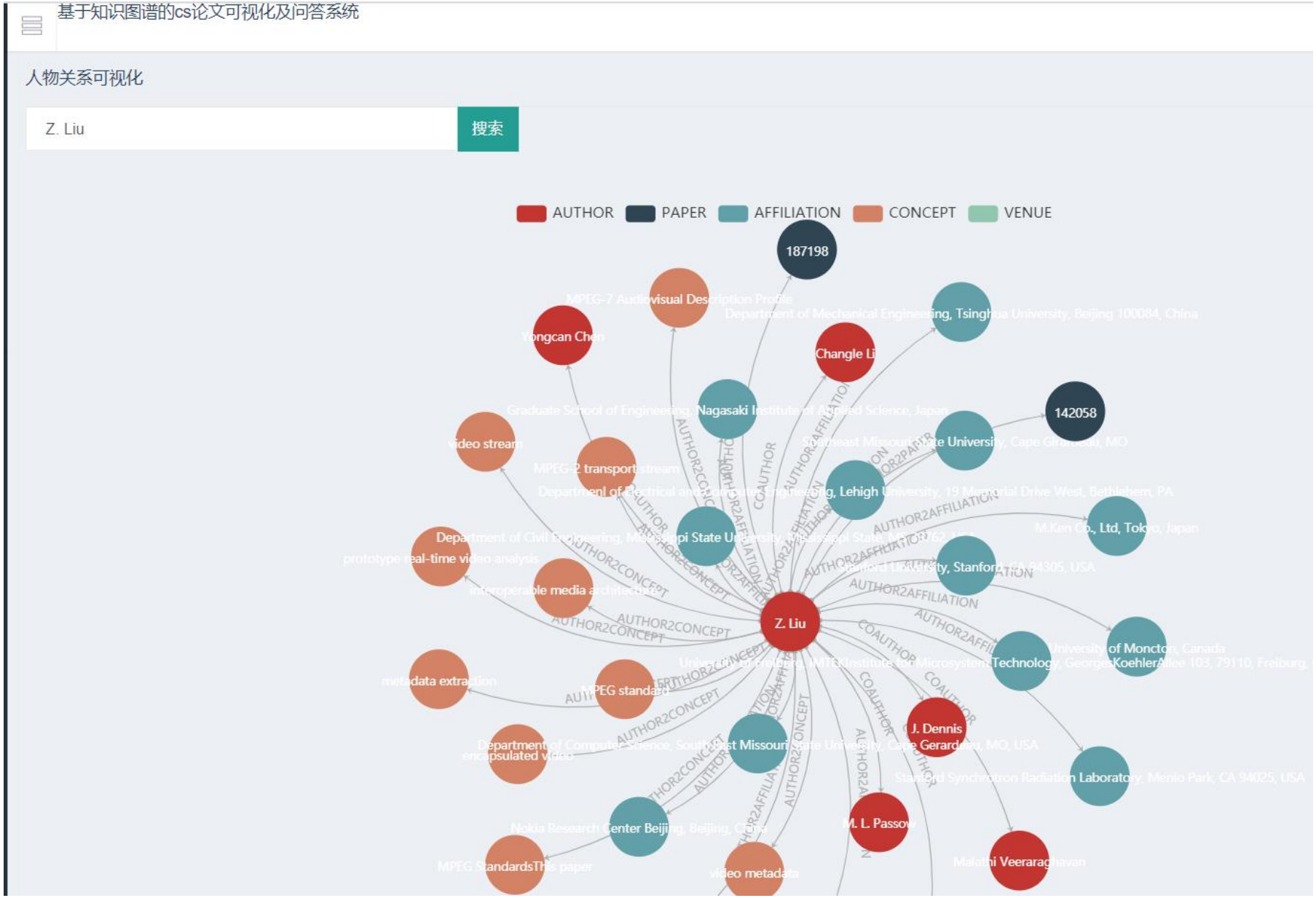
Flask
Neo4j
echart


基于知识图谱的CS论文可视化及问答系统
开启探索

合作关系


NLP课程作业
基于知识图谱的CS论文 可视化及问答系统
人物关系可视化
检索合作关系
检索作者信息
检索论文信息
红楼梦人物关系全貌
问答系统
人物关系问答
人物关系可视化
Z. Liu
搜索
1
AUTHOR
PAPER
AFFILIATION
CONCEPT
VENUE
John M. Dennis
P. V. Sarwate
Malathi Veeraraghavan
Yongcan Chen
Changle Li
Z. Liu
J. Dennis
Weihua Yi
M. L. Passow
M. Rais-Rohani
COAUTHOR

信息查询

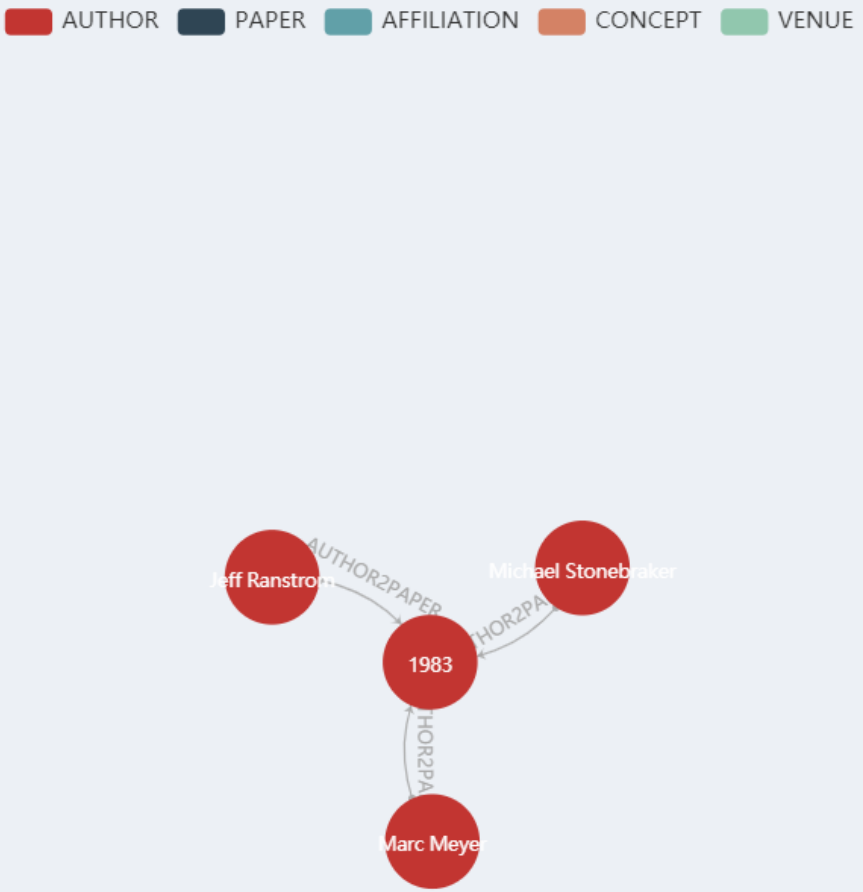

基于知识图谱的cs论文可视化及问答系统
人物关系可视化
Z. Liu
搜索
AUTHOR
PAPER
AFFILIATION
CONCEPT
VENUE
187198
142058
Z. Liu
Changle Li
Yongcan Chen
Nagasaki Institute of Applied Science, Japan
Department of Mechanical Engineering, Tsinghua University, Beijing 100084, China
video stream
MPEG-2 transportstream
Lehigh University, 19 Memorial Drive West, Bethlehem, PA
M.Ken Co., Ltd, Tokyo, Japan
prototype real-time video analysis
interoperable media architecture
AUTHOR2CONCEPT
AUTHOR2AFFILIATION
COAUTHOR
metadata extraction
MPEG standard
J. Dennis
Stanford Synchrotron Radiation Laboratory, Menlo Park, CA 94025, USA
encapsulated video
Nokia Research Center Beijing, Beijing, China
M. L. Passow
Malathi Veeraraghavan
MPEG StandardsThis paper
video metadata

基于知识图谱的cs论文可视化及问答系统

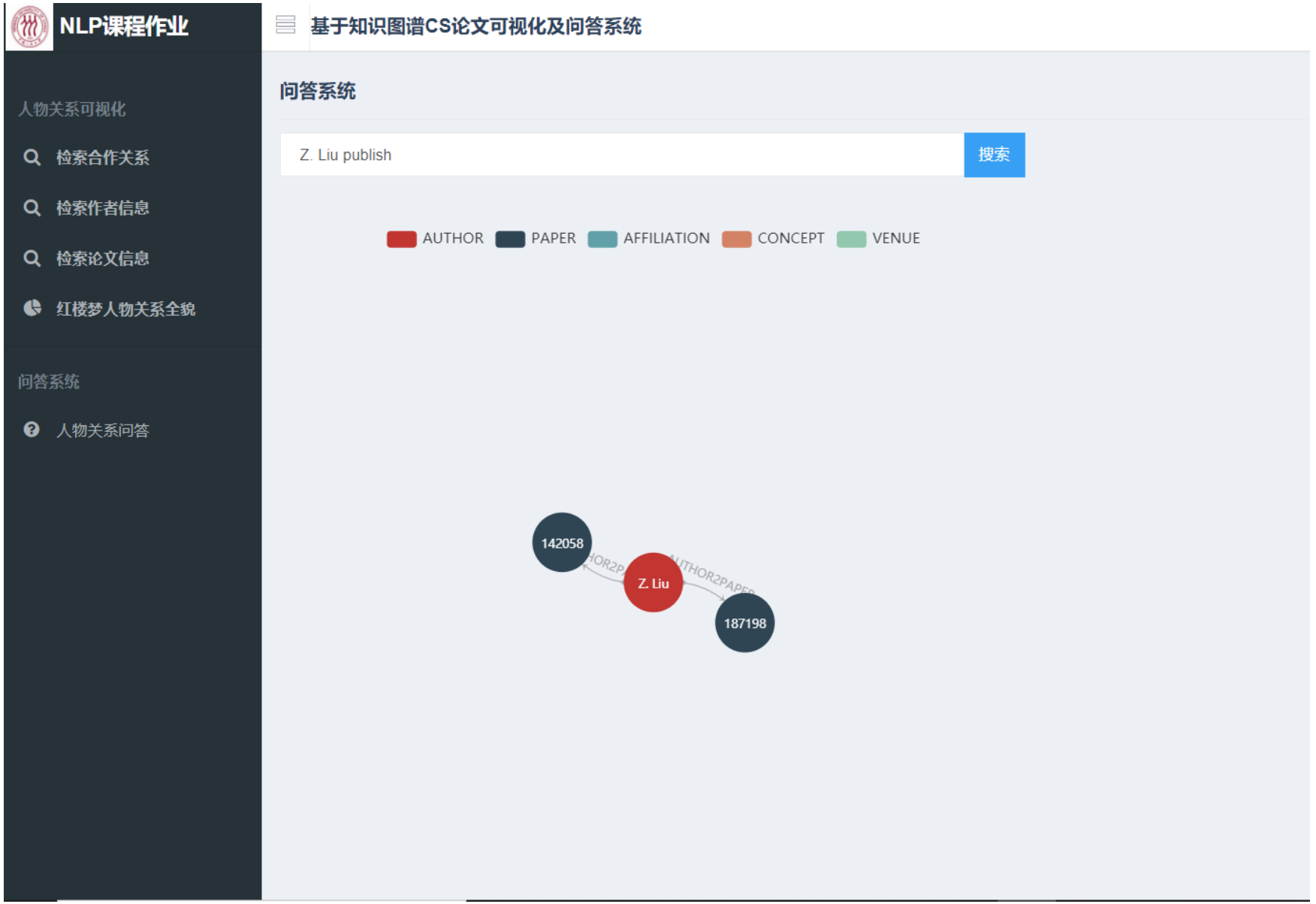
人物关系可视化

1983

搜索



问答系统


NLP课程作业
基于知识图谱CS论文可视化及问答系统
人物关系可视化
检索合作关系
检索作者信息
检索论文信息
红楼梦人物关系全貌
问答系统
人物关系问答
问答系统
Z. Liu publish
搜索
AUTHOR
PAPER
AFFILIATION
CONCEPT
VENUE
142058
Z. Liu
187198

基于模板和机器学习的问答

1、中文 or 英文
2、同义词语料库 wordnet
3、词性标注和ner nltk or others
4、机器学习 or 深度学习

基于模板的方法

模板定义
模板生成
模板匹配

参考tbsl论文

Step 1: 模板生成 – Linguistic processing

1. 首先，获取自然语言问题的POS tags信息
2. 其次，基于POS tags, 语法规则表示问句
3. 然后利用domain-dependent词汇和domain-independent词汇辅助分析问题
4. 最后，将语义表示转化为一个SPARQL模板

模板匹配与实例化–实体识别与属性检测

- 有了SPARQL模板以后，需要进行实例化与具体的自然语言问句相匹配。即将自然语言问句与知识库中的本体概念相映射的过程。
  - 对于resources和classes,实体识别常用方法:
    - 用WordNet定义知识库中标签的同义词
    - 计算字符串相似度 (trigram, Levenshtein 和子串相似度)
  - 对于property labels,将还需要与存储在BOA模式库中的自然语言表示进行比较
  - 最高排位的实体将作为填充查询槽位的候选答案

排序打分

1. 每个entity根据string similarity和prominence获得一个打分

2.一个query模板的分值根据填充slots的多个entities的平均打分

3. 另外,需要检查type类型:

--对于所有的三元组?x **rdf:type <class>**,对于查询三元组?x **p e**和 **e p ?x**需要检查**p**的domain/range是否与<class>一致

4. 对于全部的查询集合，仅返回打分最高的.

TBSL主要缺点

- 创建的模板结构未必和知识图谱中的数据建模相契合

- 考虑到数据建模的各种可能性，对应到一个问题的潜在模板数量会非常多，同时手工准备海量模板的代价也非常大

Nltk 和standfordnlp

比较两种方式，我们可以发现，NLTK下的命名实体识别更加倾向于分词和词性标准，虽然它也会将组织名，人名，地名等标注出来，但由于它把文件中的谓语，宾语等成分也标注了出来，造成了输出文本的冗余性，不利于读者很好的识别命名实体，需要我们对文本做进一步处理。NLTK下的命名实体识别的有点时，可以使用NLTK下的treebank包将文本绘制为树形，使结果更加清晰易读。相较而言，我更加倾向于Stanford的命名实体识别，它可以把Time, Location, Organization, Person, Money, Percent, Date七类实体很清晰的标注出来，而没有多余的词性。但由于NER是基于java开发的，所以在用python实现时可能由于jar包或是路径问题出现很多bug。

基于朴素贝叶斯的机器学习分类方法

```
What papers have nr publishes  author2paper  1
Which organization does nr belong to  author2affiliation  2
Who has nr worked with  coauthor  3
What areas is nr interested in  author2concept  4
Which papers have been cited in this paper  citation  5
Who is included in this organization  affliation2author  6
Who are studying this concept  concept2author  7
What are the papers related to this concept  concept2paper  8
Which papers A and B have worked together  coauthor2paper  9
```

```python
    # 训练并测试模型-NB
    def train_model_NB(self):
        X_train, y_train = self.train_x, self.train_y
        self.tv = TfidfVectorizer()

        train_data = self.tv.fit_transform(X_train).toarray()
        clf = MultinomialNB(alpha=0.01)
        clf.fit(train_data, y_train)
        return clf
```

To_do_list

1、增加机器学习的样本量
2、尝试使用其他机器学习算法以及深度学习算法