



大规模分布式知识图谱 表示推理模型及应用

ECNUICA

杨燕



CONTENT

01. 知识图谱概述

02. 现有解决方案

03. KGPro模型

04. 知识图谱应用

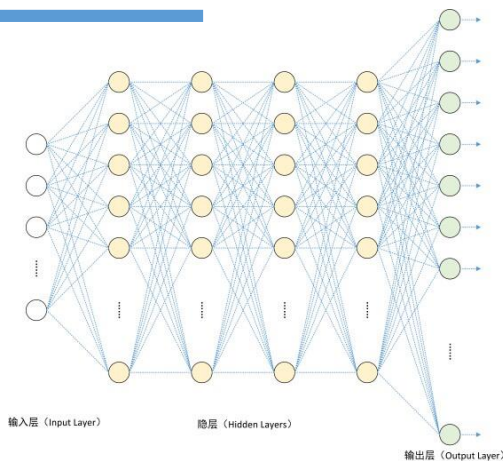
智慧是怎么来的？



计算机智能需要知识吗？

大数据

标注多



我们缺少知识吗？
什么样的知识？



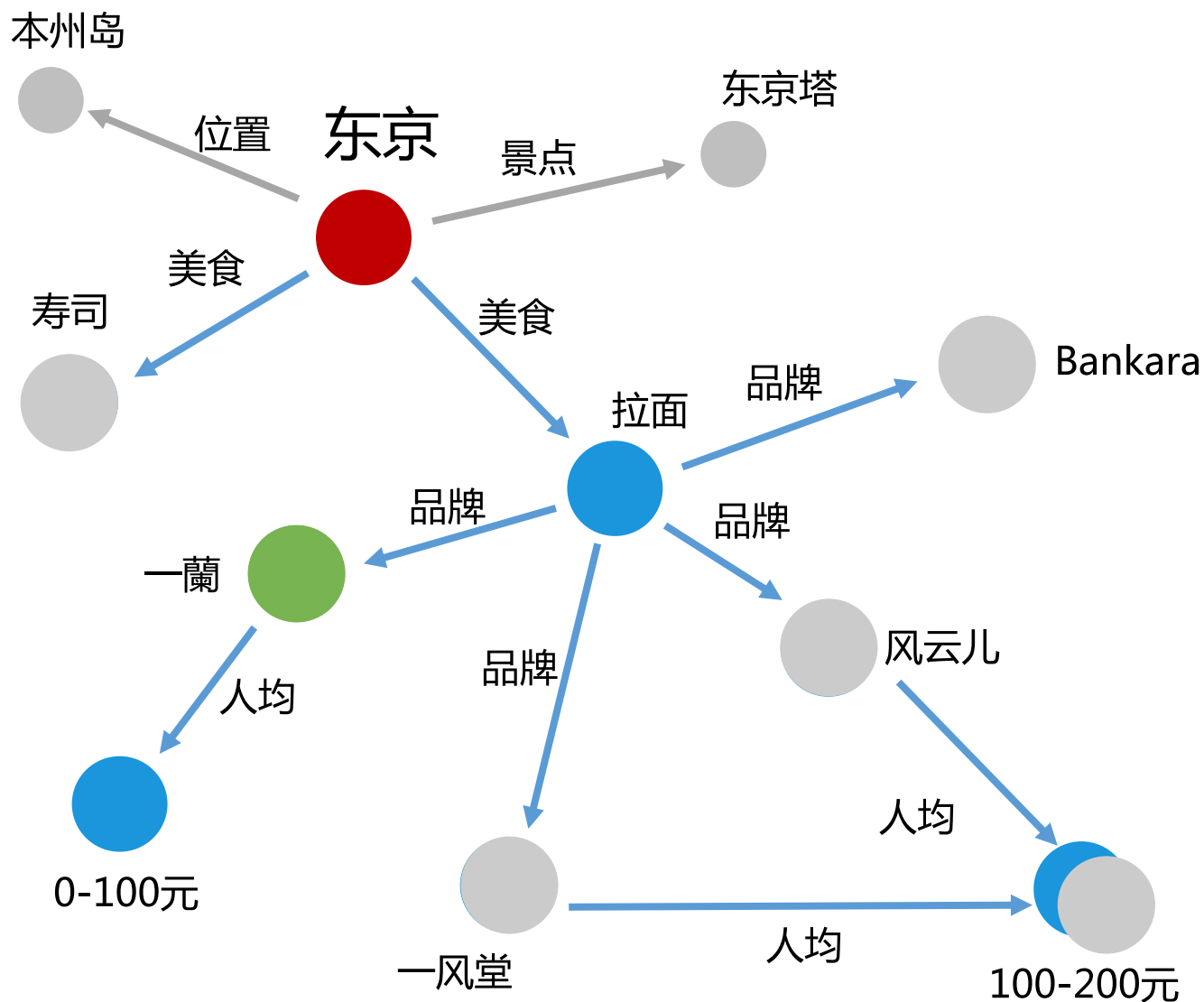
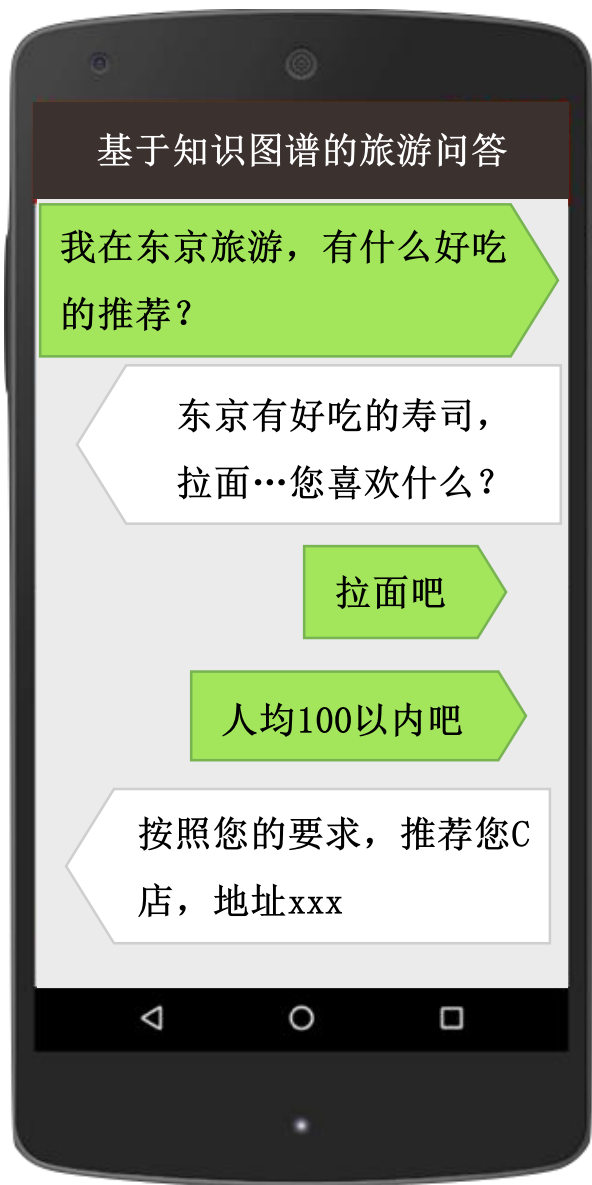
少标注

没数据

少计算



知识图谱可以带给我们什么？——深度挖掘和精确回答

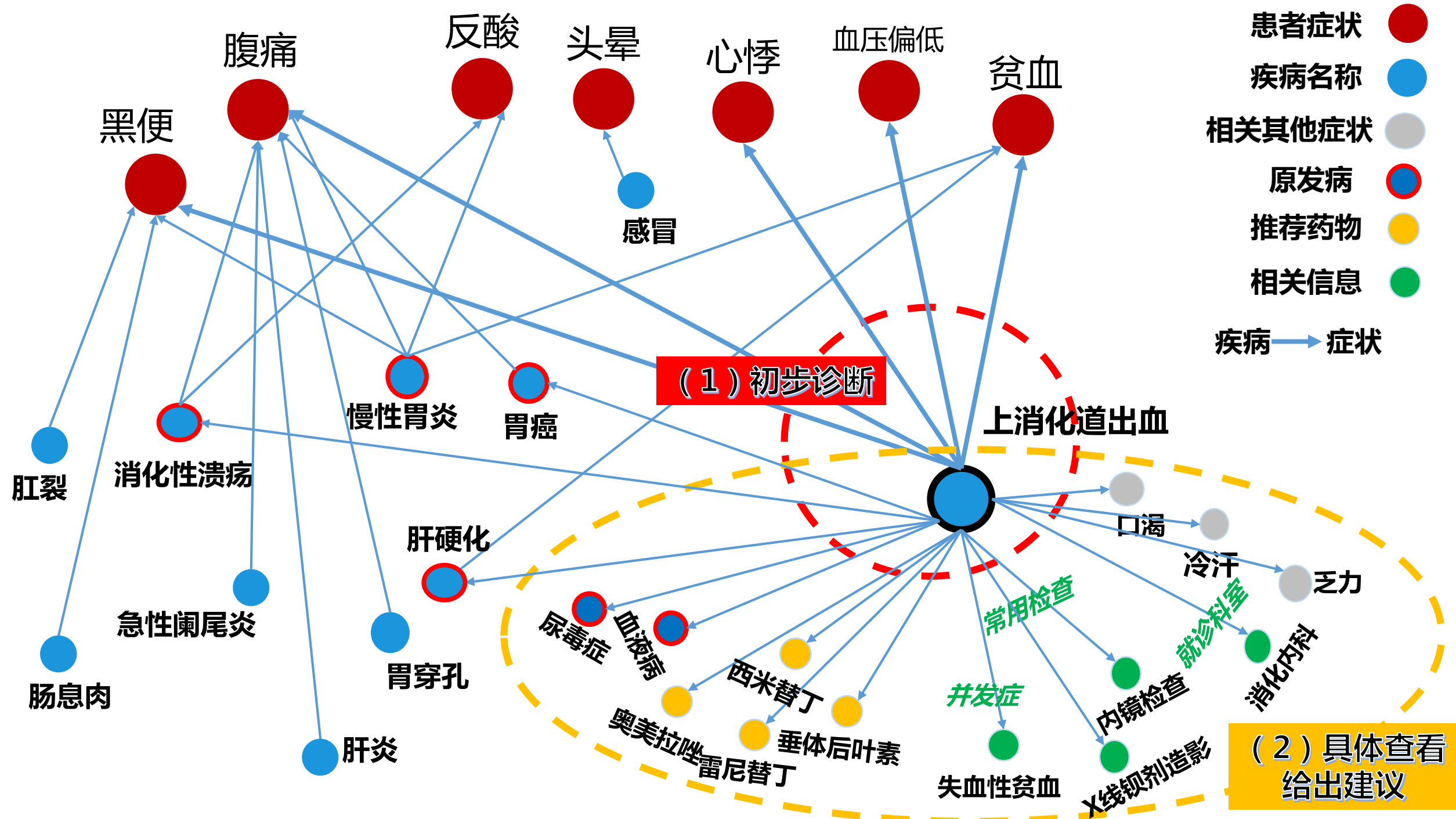


知识图谱可以带给我们什么？——深度挖掘和精确回答

智能 医疗

病历摘要

王力，男，32岁，司机。反复上腹疼痛三年，【黑便】一天。【腹痛】伴【反酸】，多空腹发作。昨又发作并解柏油糊状便4次，量约1000ml，便后【头晕】、【心悸】。查体：【血压80/50mmHg】（注：血压偏低），心率110次/分，无肝掌、蜘蛛痣。上腹剑突下有深压痛，无反跳痛。肝、脾肋下未触及。肠鸣音6次/分钟。【血Hb70g/L RBC3.0X10¹²/L】（注：贫血）
Pt230X10⁹/L ALT 40IU/L A/G 1.3 粪隐血
(+++)



现有的方案



知识表示



数据存储



查询推理



Apache Jena

A free and open source Java framework for building [Semantic Web](#) and [Linked Data](#) applications.

OWL/RDF

TDB

SPARQL



RDF

图数据库

SPARQL+子图匹配



节点关系/RDF

图数据库

Cypher

知识表示

RDF 语义网中的资源描述框架

三元组 <subject, predicate, object>

- **主语**是一个被描述的资源，由统一资源描述符**URI**来表示
- **谓词**可以表示主语的属性，或者表示主语和宾语之间的某种关系，但表示属性时，**宾语**就是属性值。通常是字面值，否则宾语是另外一个URI表示的资源。

```
<rdf:Description rdf:about=" http://example.org/Bob#me" >  
  <fofa:topic_interest  
    rdf:resource=http://www.wikidata.org/entity/Q12418" />  
</rdf:Description>
```



URI命名成本高、可维护性和可读性上有问题



三元组表示灵活，但支持的推理能力很弱

知识存储



- 1 基于内存的RDF存储方案：Sesame、DBLink
- 2 基于文件系统的存储方案：Native RDF、System II
- 3 基于关系数据库的存储方案：3store、Rstar
- 4 基于NoSQL的存储方案：



列式数据库：Hbase



文档型数据库：MongoDB、CouchDB

图数据库的存储方案：Neo4j、Taitan

查询和推理



查询的目的是将存储的知识呈现给用户



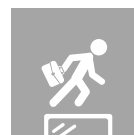
推理的主要任务则是为用户提供隐含的更深层次的知识。

SPARQL

```
SELECT ?thing
WHERE {
  {
    ?thing x:name ?name .
    FILTER regex(?name, "%s", "i" )
  } UNION {
    ?thing x:featurename ?name .
    FILTER regex(?featurename, "%s", "i" )
  }
}
```



SPARQL无法轻易查询出较为复杂的逻辑蕴含。



数据库对于SPARQL支持比较少。



查询和推理

传统一阶谓词逻辑

- 描述逻辑
- Horn 子句

不确定性推理

- 粗糙集推理
- 基于概率逻辑推理

我们的需求



轻量化

扬长避短，保留推理
逻辑等优势同时摆脱
OWL/RDF繁重框架



高性能

针对不同领域的场景
设计知识库架构以达
到高效查询需求



易用性

学习曲线低，易于
解释并对用户查询友好



可扩展

良好的结构设计易于
扩展，可适应日益增
加的知识图谱规模

模型考虑的元素



术语

用户输入的是一
个或多个术语



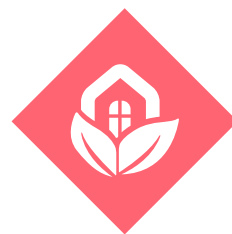
对应语义单元

根据输入的术语，
找到所对应概念
或者实体



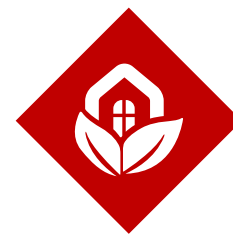
概念实体单元

根据找到的实体或
者概念，去查询该
实体或者概念的概
念实体单元



约束检验规则

用户要编辑知识
图谱是要符合一
定的规范和约束



推理规则

用户需要得到最
大的信息量，即
需要提供给用户
隐含的知识

我们的模型

应用层

知识管理分布式
应用

知识推理

问答/搜索

中间层（算法）

存储模型（统一的API
和中间件实现）

推理机

模型层

存储模型

推理模型
（基于Horn子句）

知识表示模型（KGPro-Schema）



KGPro_Schema知识表示

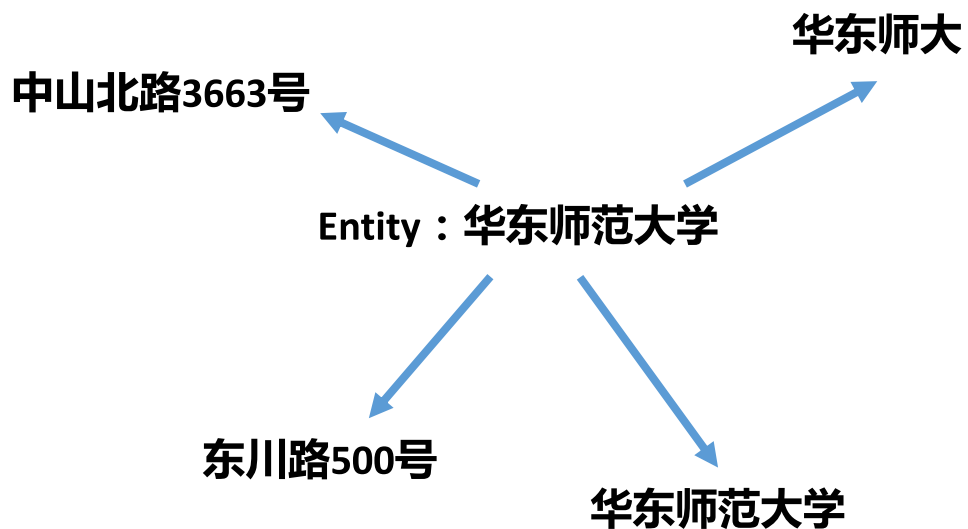
三元组知识表示



语义描述单元

语义单元，表示描述同一抽象概念或者实体的术语和同一术语描述不同的实体或者概念，术语和抽象概念以及实体以多对多的关系存在。

同一个实体对应不同描述



同一个描述对应不同的实体

Entity : 苹果

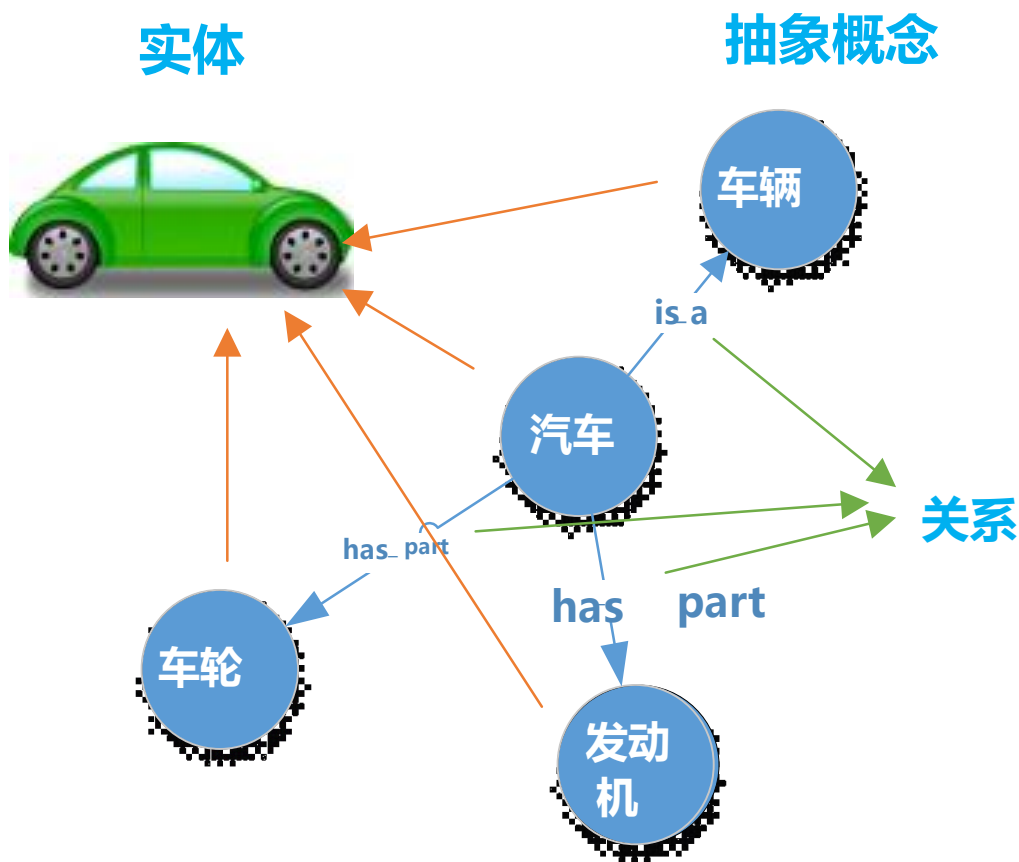


Entity : 苹果



抽象概念实体描述单元

抽象概念实体单元，同一个概念或者实体以及描述它的实体和其他抽象概念和它们之间的关系。



P关系分类法

人物

性别

男、女

职业

体育、教育

按职业属性分类，分类后职业确定

体育人物

性别

男、女

职业

体育

教育人物

性别

男、女

职业

教育

分类方法一

按性别属性分类，分类后性别确定

男人物

性别

男

职业

体育、教育

女人物

性别

女

职业

体育、教育

分类方法二



KGPro知识存储

MongoDB



灵活易用

文档型数据库采用JSON格式存储，支持更加自由和灵活的实体属性存储

支持索引

支持基本索引，全文索引等，可以针对性对查询进行优化

扩展性好

MongoDB支持分布式存储，文档型结构也易于水平分表等优化

存储模型设计

多个记录的集合：一张表



{cid: 009, 202 : [001,002,003,004,005],synsets:实体}

{cid: 001, 204 : 101, synsets:学校 }

{cid: 002, 204 : 104, 201: 009, synsets:娱乐人物 }

{cid: 003, 204 : 103, synsets:水果 }

{cid: 004, 204 : 102, synsets:公司 }

{cid: 005, 202 : 006, synsets:作品 }

{cid: 006, 201 : 005, synsets:音乐作品}

{cid: 007, 201 : 005, synsets:影视作品}

{cid: 101, 203 : 001, synsets:[华东师大, ECNU]}

{cid: 102, 203 : 004, synsets:苹果 }

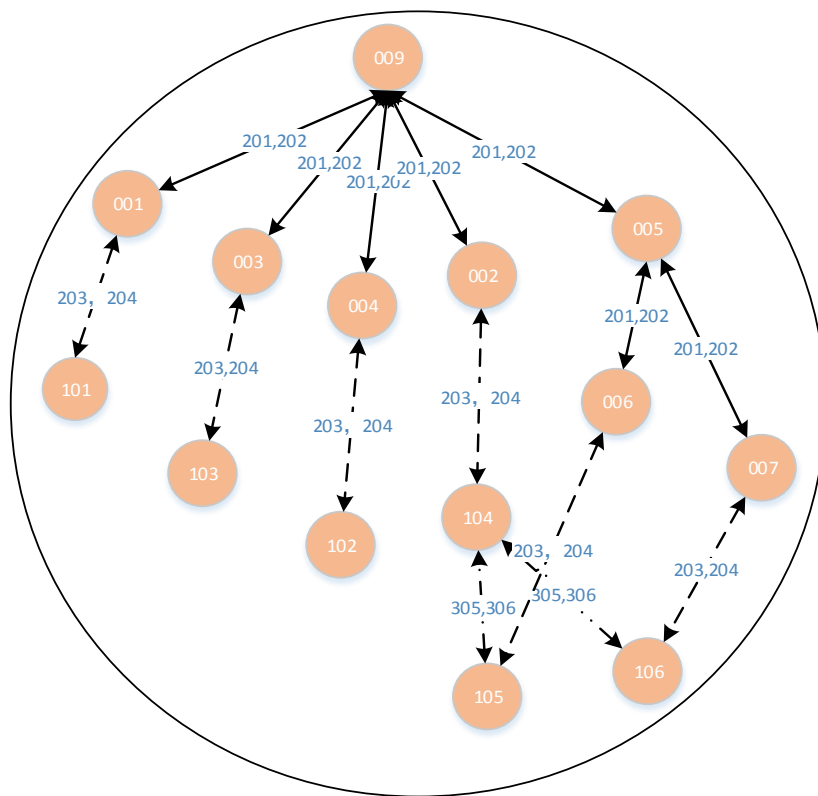
{cid: 103, 203 : 003, synsets:苹果 }

{cid: 104, 203 : 002, 305 : [105,106], synsets:王菲}

{cid: 105, 203 : 006, 306 : 104, synsets:红豆}

{cid: 106, 203 : 007, 306 : 104, synsets:大城小事}

多个子图并：一张图





KGPro分布式中间件

KGPro知识图谱中间件

客户端

中间件客户端模块

负载均衡和查询处理模块

中间件服务端模块

知识图谱异构存储框架中间件

数据库驱动模块

知识图谱抽象驱动模块

多元数据库支持组件

MangoDB

数据库驱动模块

知识图谱抽象驱动模块

多元数据库支持组件

Taitan

数据库驱动模块

知识图谱抽象驱动模块

多元数据库支持组件

....数据库

KGPro知识图谱中间件



多元数据库支持组建主要负责将数据库操作封装，抽象成统一的接口，起到了数据库驱动的作用。



数据库驱动模块将数据服务方式统一，将数据库操作封装。提供三元组的增、删、改、查操作API服务。



知识图谱抽象驱动模块将基础的三元组操作接口组合，封装成图操作接口，查询节点距离为 n 的所有节点，三元组任意位置查询，路径查询。



知识图谱分布式中间件模块主要提供了查询和推理功能，并在此基础上加入了负载均衡和多种知识图谱操作接口

存储性能分析

验证存储模型的有效性



数据源：

- 100万、1000万、1亿规模
- 在给定相同机器的情况下，通过KGPro-Schema、Jena、Titan进行查询操作，并记录相同查询下的平均响应耗时。

硬件环境和软件环境：

- 硬件环境采用酷睿i5，4核系列CPU，内存8G
- 软件开发环境为Java。
- 存储方案简单查询性能对比
- 存储方案插入性能对比

存储性能分析

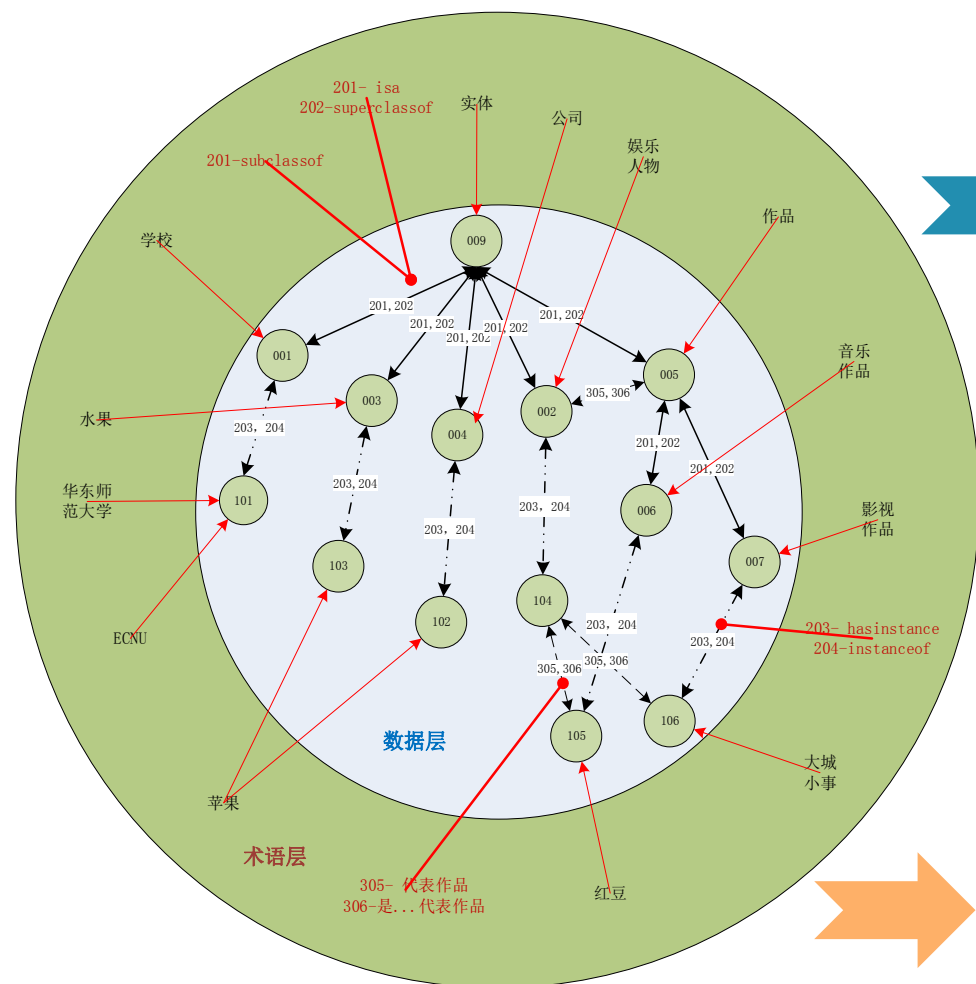
查询性能对比

	100W(Triple)	1000W(Triple)	10M(Triple)
KGPro-Schema	70ms	110ms	200ms
Titan	510ms	0.6s~1.5s	1s~3s
Jena	30ms	1.6s	5s

平均插入耗时对比

	100W(Triple)	1000W(Triple)	10M(Triple)
KGPro-Schema	9ms/Triple	10ms/Triple	11ms/Triple
Titan	10ms/Triple	13ms/Triple	17ms/Triple
Jena	<1ms/Triple	1ms/Triple	2ms/Triple

存储性能总结



(1) 数据库设计简单，NoSQL数据库的数据存储模式能够方便的存储数据的语义信息

(2) 将符号语言与事物ID分离存储，有效解决了事物术语和事物本身间经常混淆的问题

(3) 能够适用于任何领域的知识图谱的存储，也兼容现有的其它格式的本体

(4) 具有较强的表达能力，该存储方式除了能够存储静态的事物外，还可以存储动态的规则

——> 表示 isa 与superclassof 关系对
- - -> 表示 hasinstance 与instanceof 关系对
...-> 表示 “代表作品” 与 “是..代表作品” 关系对



KGPro查询推理

查询和推理

NO

NoSQL的知识图谱产品中，还没有对推理功能的完整实现



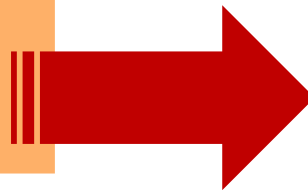
查询
推理
描述
语言

要具有丰富而直观的表达能力

具备良好定义的语法和语义

要具有一定的推理能力

Horn子句逻辑



查询和推理

基本查询：

1. 查实体：(tom , father , X)
2. 查关系：(China , X , Beijing)

推理：

规则1：(X, grandfather , Y):- (X,father, Z),(Z , father , Y)

查询：Tom的祖父是谁？

→ (tom, grandfather, Y)

事实库

father(tom,jack)
father(jack, john)

规则库

(X, grandfather , Y):-
(X,father, Z),(Z , father , Y)

查询和推理

本体
计算

推理：

推理规则为： $(X, \text{老乡}, Y) : - (X, \text{籍贯}, Z), (Y, \text{籍贯}, Z)$

规则的语义为：如果X和Y的籍贯都是一个地方，那么X和Y就是老乡。

查询：张三的老乡有哪些？

➡ (张三, 老乡, Y)

对于籍贯，在知识图谱中是已存在的，而老乡则是在籍贯的基础上新生成的一个关系。

1. 将X实例化为“张三”，然后对规则体中事实进行匹配，得到变量Z的值
2. 再由第二个事实得到Y的值。

查询：张三和李四是不是老乡？

➡ (张三, 老乡, 李四)

1. 找到规则体中的第一个事实，得到Z的值。
2. 根据第二个事实得到李四的籍贯，这时需要通过本体计算来处理，即将“张三”和“李四”的“籍贯”进行计算，求其交集。
3. 如果有交集则是老乡，如果交集为空，则不是老乡关系。



KGPro知识图谱应用

KGPro知识图谱管理系统

本体管理系统

用户:

用户管理平台 本体 概念及实例查询 规则 计算 请输入

我的本体 我的本体 人物本体

概念 实例 插入子概念 兄弟概念 删除概念 重命名 属性分类 选中展开两级节点

My Tree Panel

- 出生日期
- 籍贯
- 人物
 - 体育人物
 - 娱乐人物
 - 政治人物
 - 历史人物
 - 军事人物
 - 教育人物
 - 经济人物
 - 社会人物

My Tree Panel

添加关系及值 删除关系及值

- Root
 - is_a
 - 出生日期
 - 籍贯
 - superclass_of

关系 属性 添加关系 删除关系 重命名

My Tree Panel

关系

My Tree Panel

添加关系性质及值 删除关系性质及值

关系体

The knowledge graph visualization shows a central node '人物' (Person) in red. It is connected to several other nodes via 'superclass_of' and 'is_a' relationships. The nodes include: '科学人物' (Scientist), '教育人物' (Educator), '政治人物' (Politician), '历史人物' (Historian), '虚拟人物' (Virtual Person), '其他人物' (Other Person), '体育人物' (Sports Person), '文化人物' (Cultural Person), '经济人物' (Economic Person), '社会人物' (Social Person), '娱乐人物' (Entertainment Person), '军事人物' (Military Person), '热点人物' (Hotspot Person), and '籍贯' (Hometown). The graph also shows '出生日期' (Date of Birth) as a property of '人物'.

KGPro知识图谱管理系统

规则管理

新规则名:

师兄弟

(Y,Z)

关系:

有弟子



(X,Y)

关系:

有弟子



(X,Z)

规则:

添加规则

推理输入:

X:

颜回

颜回->>儒家人物

Y:

确定概念X

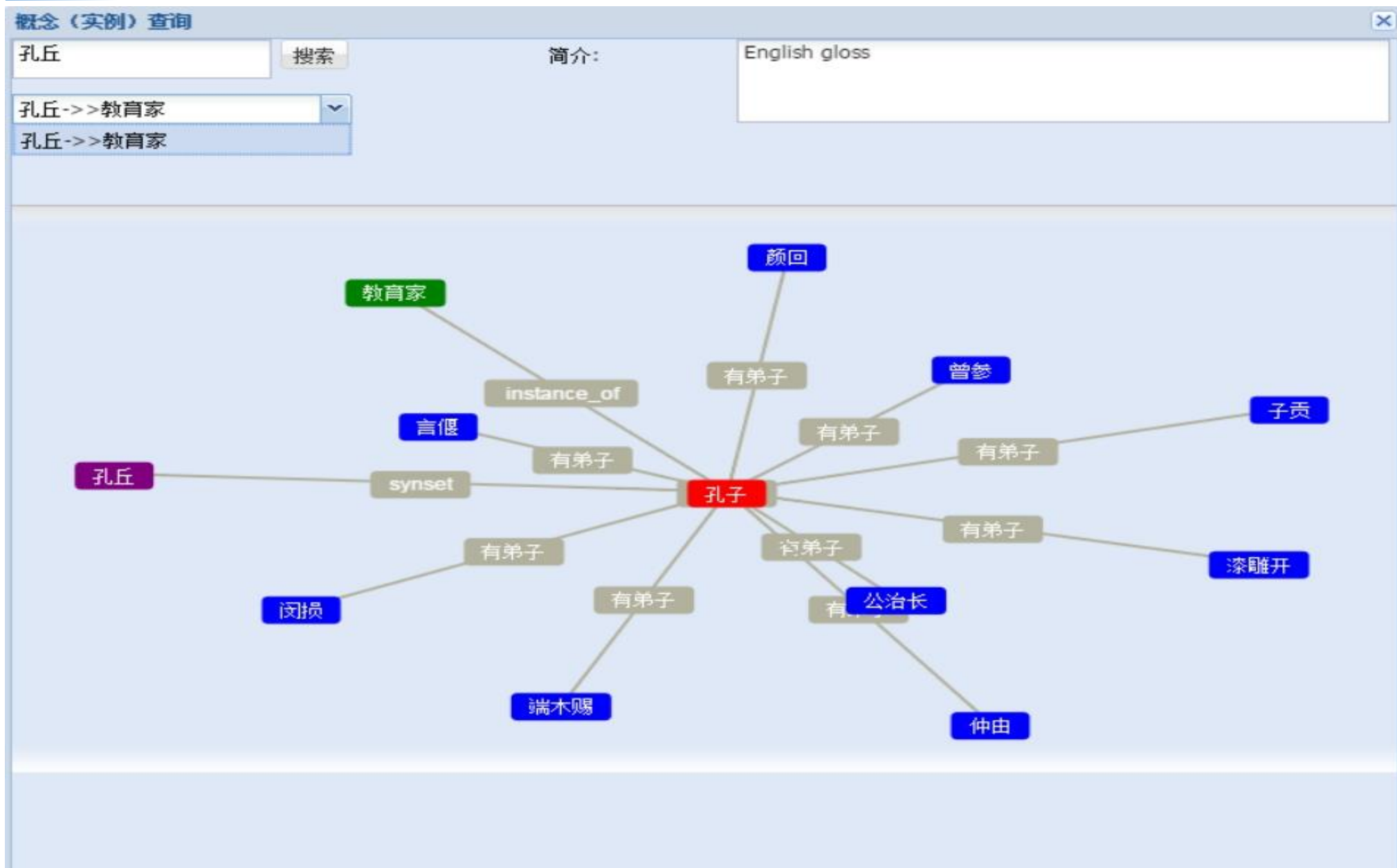
确定概念Y

推理

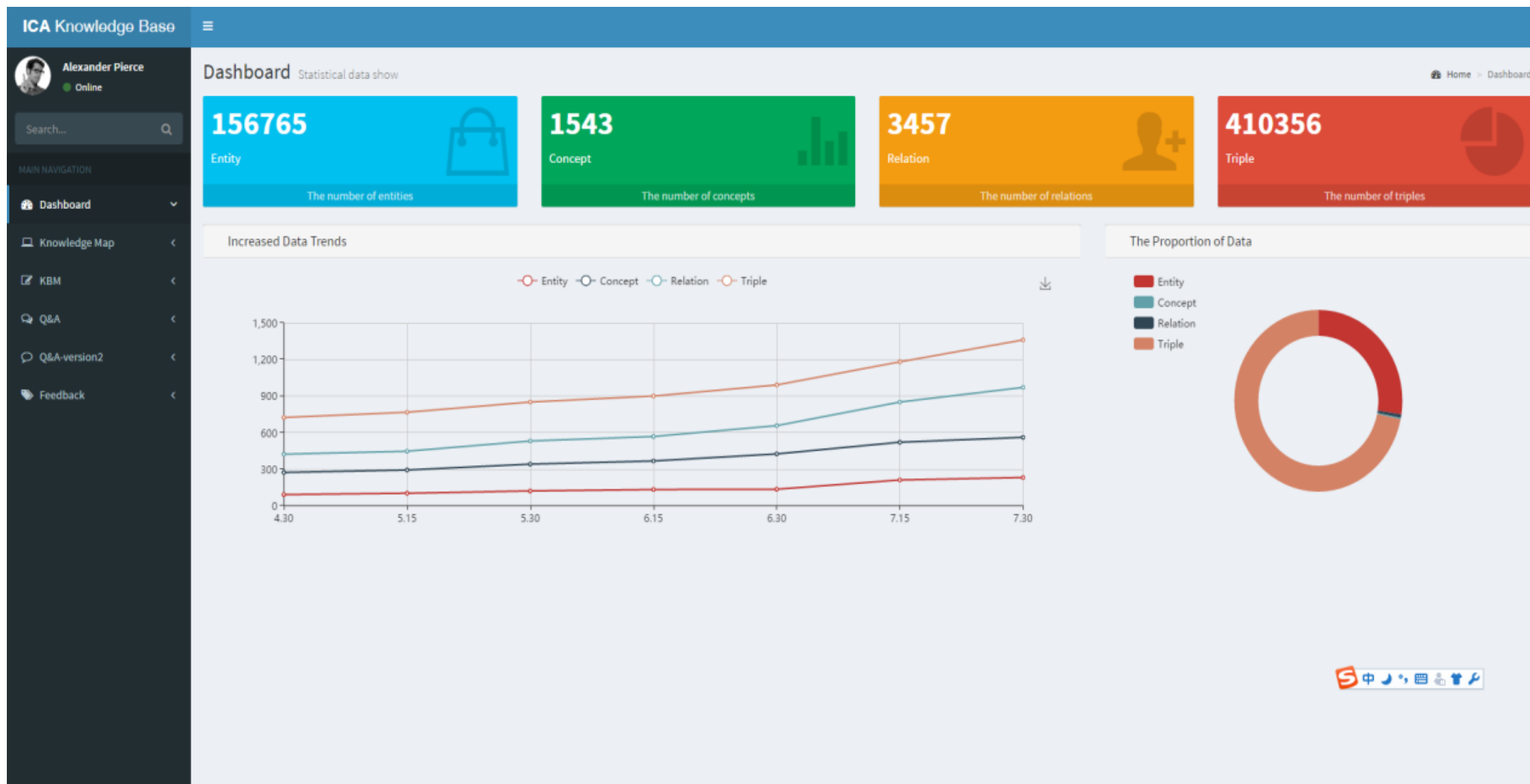
推理结果:

曾参
子贡
公冶长
端木赐
仲由
闵损
高偃
漆雕开

KGPro知识图谱管理系统



KGPro知识图谱管理系统



KGPro知识图谱管理系统

ICA Knowledge Base

Alexander Pierce

Online

Search...

MAIN NAVIGATION

Dashboard

Knowledge Map

KBM

Q&A

Q&A-version2

Feedback

Knowledge Base Management panel

+添加...

Search

刷新

列表

网格

分享

<input type="checkbox"/>	Name	Description	Is_a	Operation
<input type="checkbox"/>	霍华德·舒尔茨	霍华德·舒尔茨: 性别男, 出生年月 1953年7月19日, 籍贯 纽约的布鲁克林区, 是一位CEO 昵称有: Howard Schultz,	root	删除 修改 查看关系
<input type="checkbox"/>	Howard Schultz	霍华德·舒尔茨: 性别男, 出生年月 1953年7月19日, 籍贯 纽约的布鲁克林区, 是一位CEO 昵称有: Howard Schultz,	root	删除 修改 查看关系
<input type="checkbox"/>	男	English gloss	root	删除 修改 查看关系
<input type="checkbox"/>	1953年7月19日	English gloss	root	删除 修改 查看关系
<input type="checkbox"/>	纽约的布鲁克林区	English gloss	root	删除 修改 查看关系
<input type="checkbox"/>	裴熙亮	裴熙亮: 性别男, 出生年月未知, 籍贯 未知, 是一位CEO 昵称有:	root	删除 修改 查看关系
<input type="checkbox"/>	英德拉·努伊	英德拉·努伊: 性别女, 出生年月 1957年, 籍贯 印度晨奈, 是一位CEO 昵称有:	root	删除 修改 查看关系
<input type="checkbox"/>	女	English gloss	root	删除 修改 查看关系
<input type="checkbox"/>	1957年	English gloss	root	删除 修改 查看关系
<input type="checkbox"/>	印度晨奈	English gloss	root	删除 修改 查看关系

Showing 1 to 20 of 134618 rows

20 records per page

<

1

2

3

4

5

...

6731

>

Home > Knowledge Base

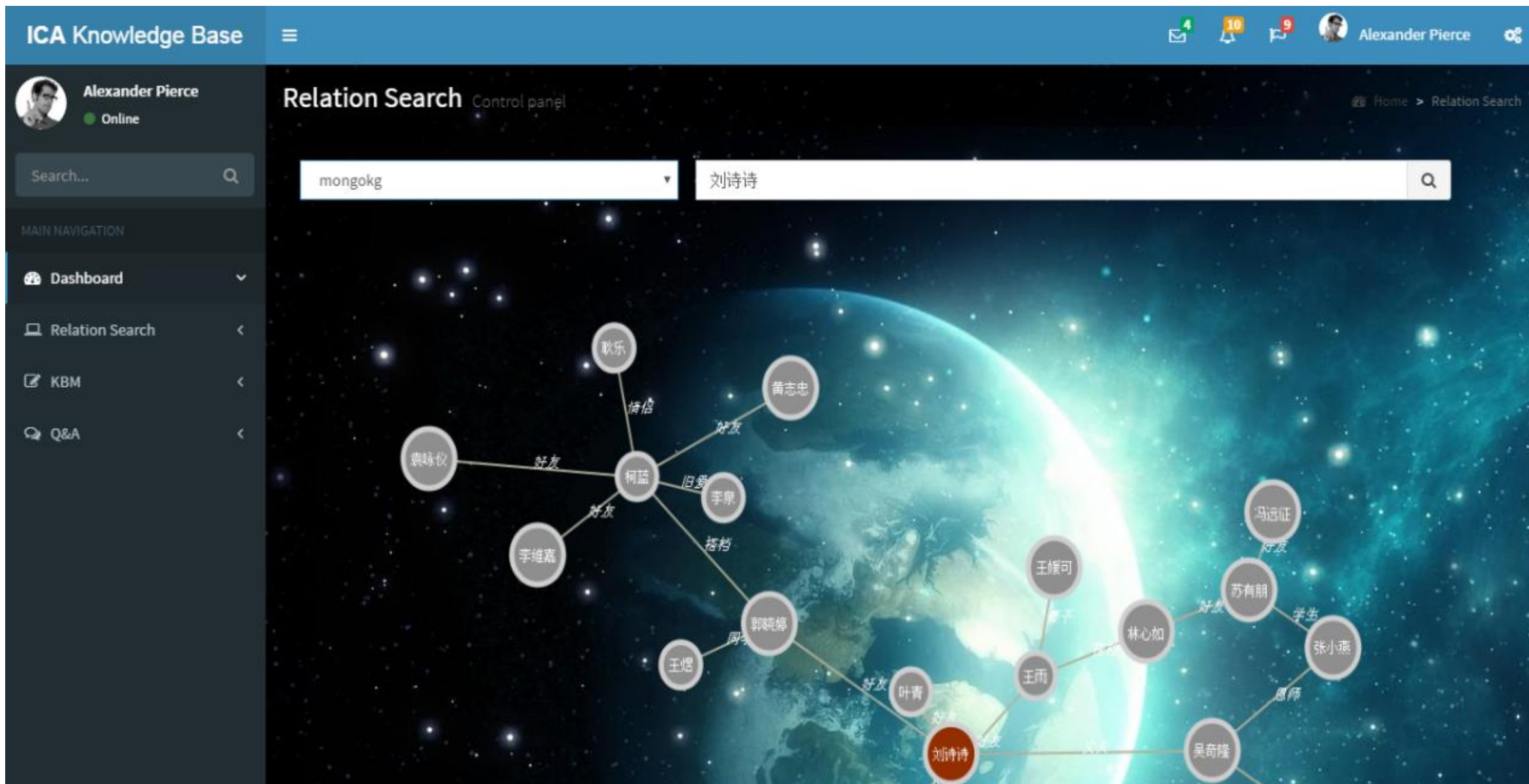
Entity

Concept

Relation


Triple

KGPro知识图谱管理系统



KGPro知识图谱管理系统

ICA Knowledge Base



Alexander Pierce
Online

Search...

MAIN NAVIGATION

- Dashboard
- Knowledge Map
- KBM
- Q&A
- Q&A-version2
- Feedback
- baba





知识图谱推理 Knowledge Map Reasoning based entity

Home > Knowledge Map Reasoning

(A,"父亲",&(B,"父亲",C)-->(A,"祖父",C)

+添加...

Search



Subject	Relation	Object
杨幂	丈夫	刘恺威
杨幂	好友	周笔畅
高圆圆	丈夫	赵又廷
高圆圆	好友	贾静雯
黄晓明	好友	赵又廷
邓超	饰演	包青天
邓超	妻子	孙俪
杨幂	女儿	小糯米
黄晓明	好友	赵又廷
邓超	饰演	包青天

Reason >>

Reasoning Result
(谢贤, 祖父, 谢振轩)

基于影视领域知识图谱的问答系统



ask me a question

梁朝伟和刘德华合作过几部电影？



文本回答 - 系统查询结果

Result Count: 6

无间道

无间道

五虎将之决裂

反斗马骝

中环英雄

无间道III

查询问句：
梁朝伟和刘德华合
作过几部电影？

结果展示

基于影视领域知识图谱的问答系统

问句解析 - 问句关键词解析标注

梁朝伟和刘德华合作过几部电影？

</> 带参数三元组 - 由问句解析生成查询三元组

```
triple: {  
  function: Count(var(x)),  
  condition:  
    [var(w), in, [导演, 演员表, 出演, 编剧]],  
    [var(z), in, [导演, 演员表, 出演, 编剧]],  
  rule:  
    [梁朝伟, var(w), var(x)],  
    [var(x), var(z), 刘德华],  
}
```

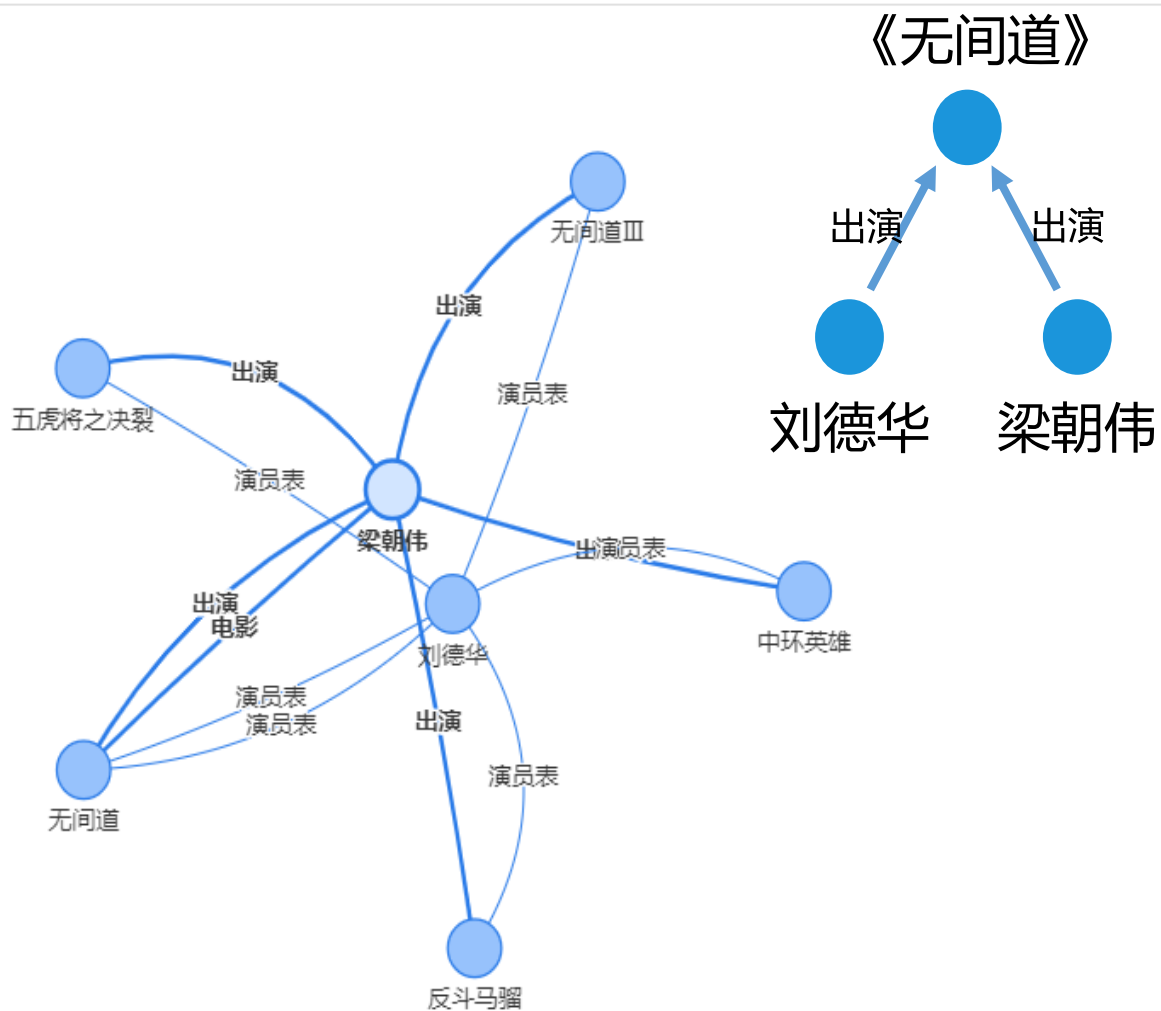
查询问句：
梁朝伟和刘德华
合作过几部电影？

问句解析流程

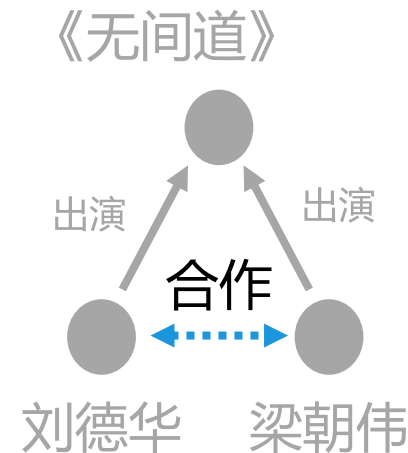
基于影视领域知识图谱的问答系统



实体图谱展示 - 问句中关键实体图谱关系



推理



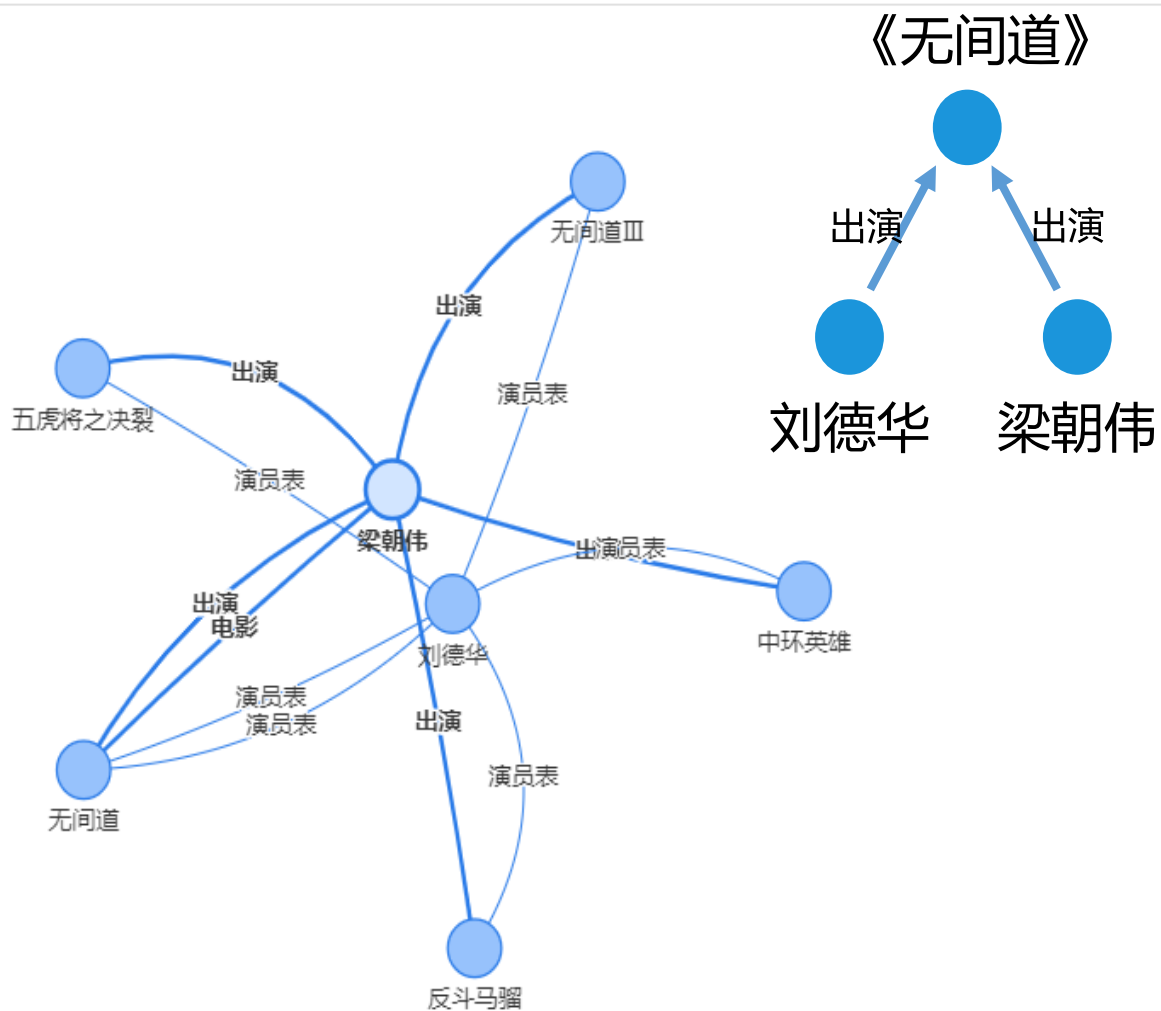
查询问句：
梁朝伟和刘德华
合作过几部电影？

实体图谱展示

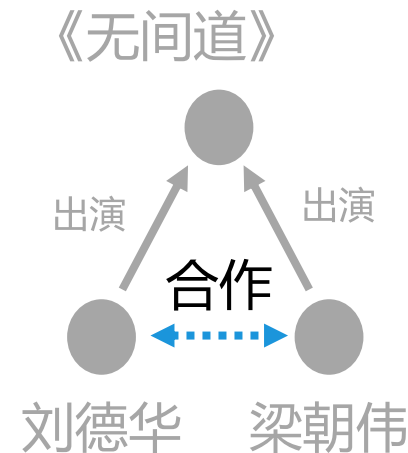
基于影视领域知识图谱的问答系统



实体图谱展示 - 问句中关键实体图谱关系



推理



查询问句：
梁朝伟和刘德华
合作过几部电影？

实体图谱展示



Q&A

Thank You !