

知识图谱在问答系统中的应用和挑战

韦克礼
图灵机器人

问答系统概述

- 问答系统能够更为准确地理解以自然语言形式描述的用户提问，并通过检索异构语料库或问答知识库返回简洁、精确的匹配答案。
- 相对于搜索引擎，问答系统能更好地理解用户提问的真实意图，同时更有效地满足用户的信息需求。

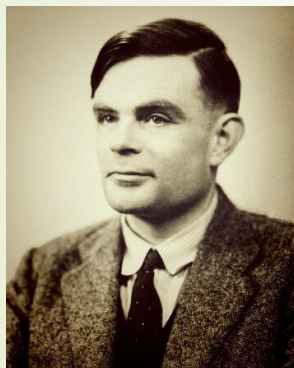
问答系统的演变

图灵测试

专家系统

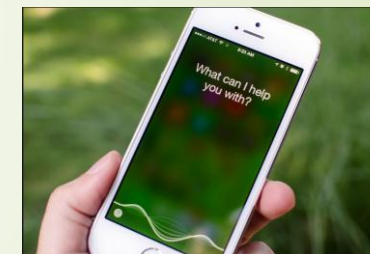
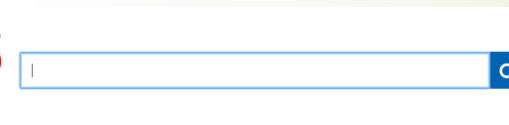
检索式问答

智能交互式问答

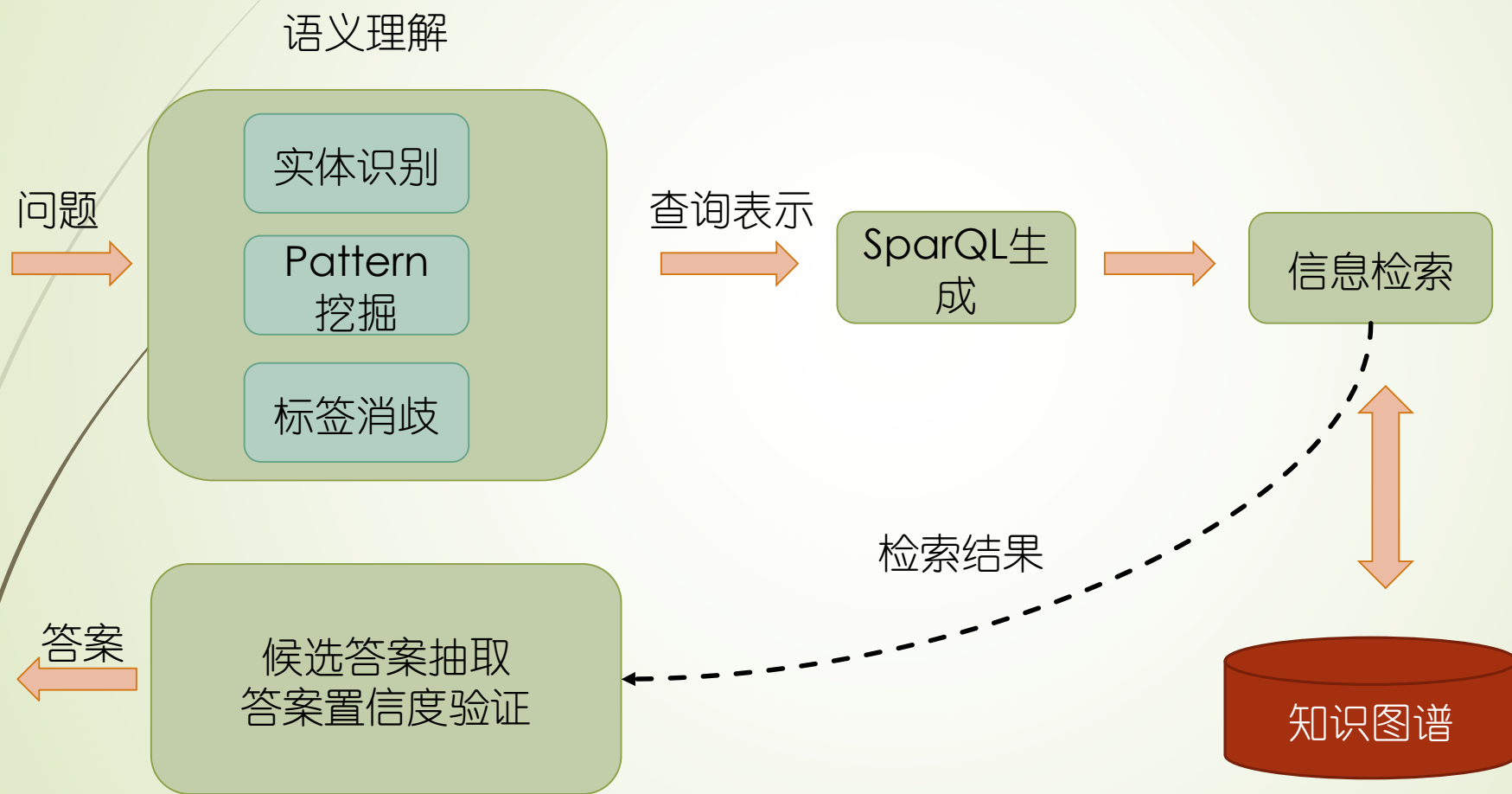


STUDENT

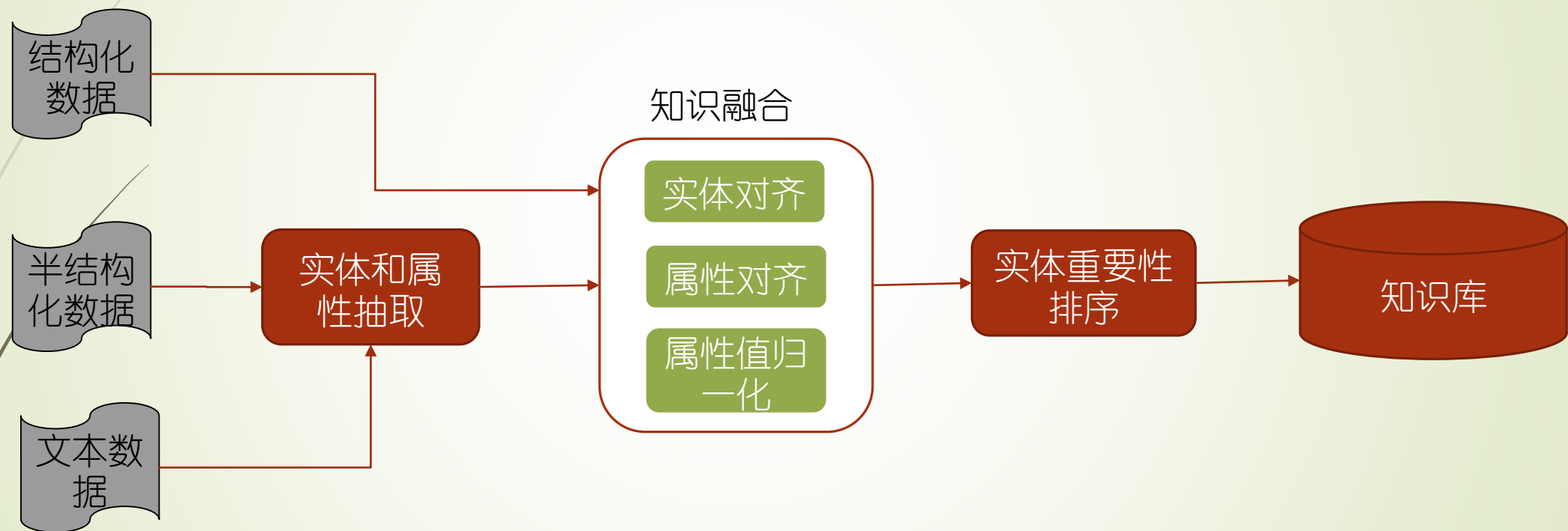
LUNAR



整体架构图



知识图谱构建的关键技术

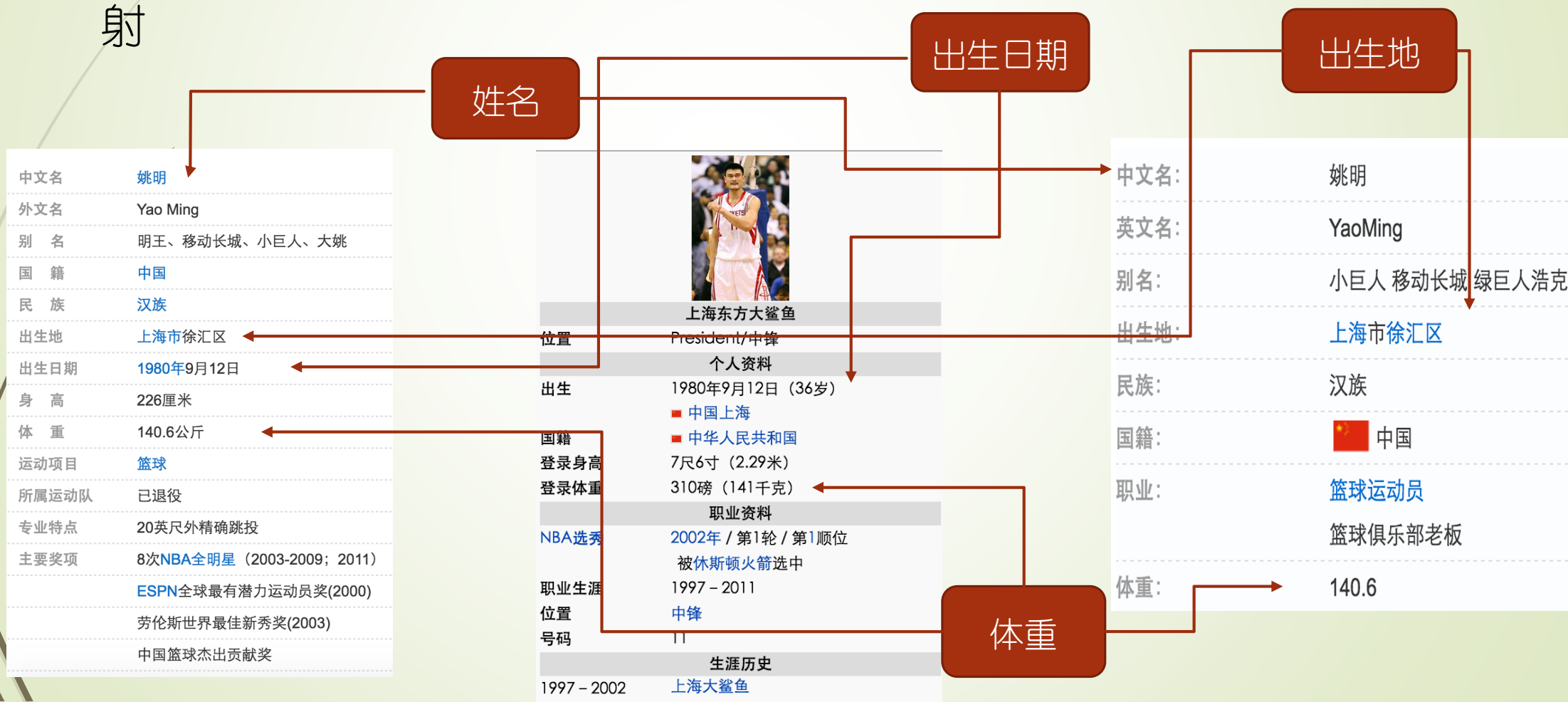


数据来源



实体属性对齐

- 不同的数据源有自己的标签和分类体系
- 需要建立一个统一的schema，将不同数据源中的数据与全局schema进行映射



属性对齐和值的规范化

- 同名属性

出生日期=出生时间=出生年月、金牌=第一名

- 属性的包含关系

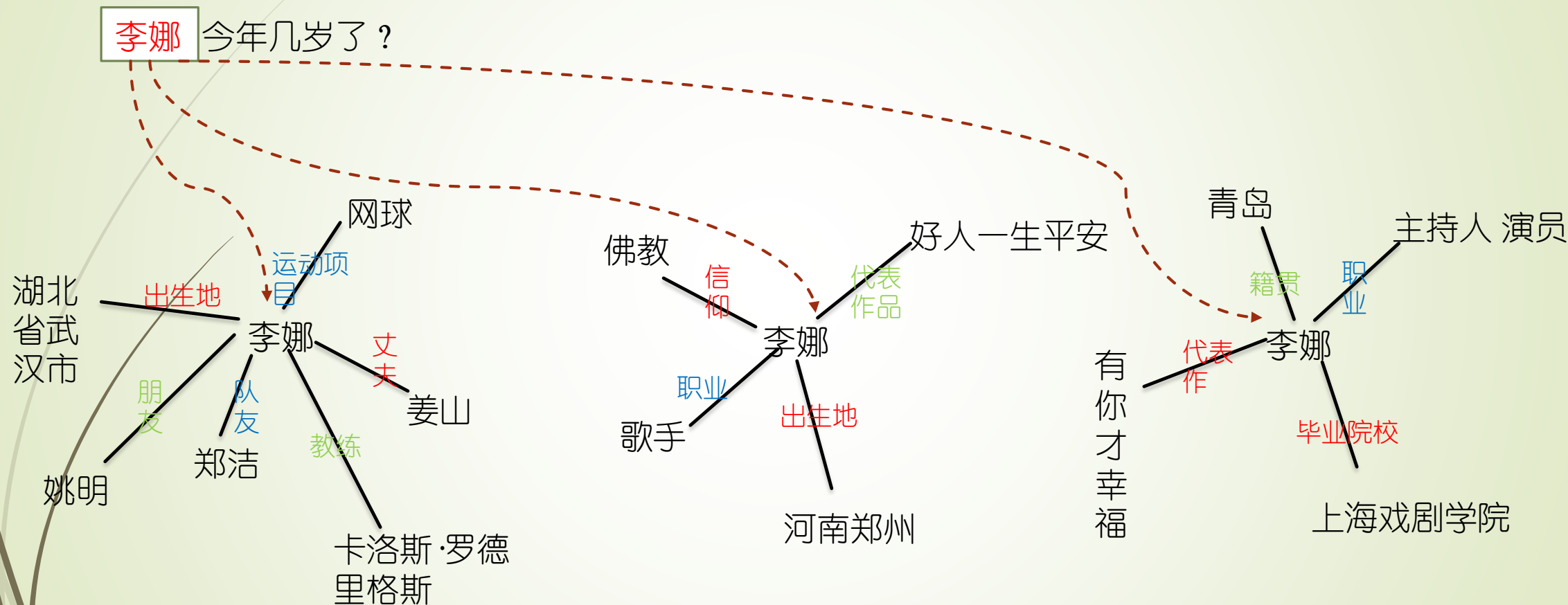
出生=出生地点 + 出生日期

- 属性值的规范化

体重：310磅 ~ 141公斤

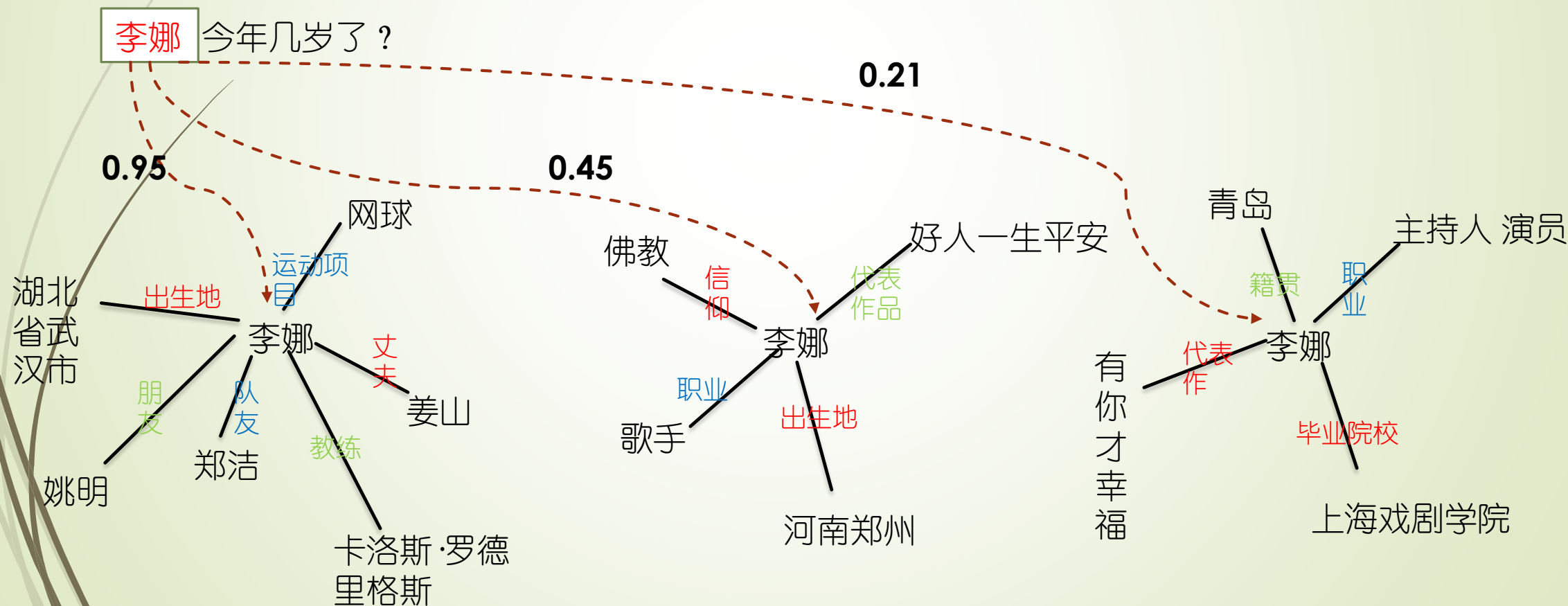
出生日期：1982年2月26日 ~ 1982.2.26

实体的重要性排序



实体重要性排序算法

- 按照实体的信息丰富程度、实体间的关联关系给予初始权重
- 使用带偏的PageRank算法，不断迭代计算，直到收敛



数据存储（图数据库）



- 传统的关系数据库难以表示实体之间灵活的关系结构
- 图数据库则以节点之间的关联关系为中心，方便复杂的关联查询

实体识别

长江是世界第三长的河流

黄河的支流有哪些？

中文名称	黄河	发源地	青藏高原巴颜喀拉山脉
英文名称	Yellow River	主要支流	汾河、洮河、渭河等
别 称	中国母亲河、河水、浊河	河 长	约5464公里
所属水系	黄河水系	河流面积	约752443平方公里
地理位置	中国北部	平均流量	2571立方米/秒
流经地区	青海、四川、甘肃、宁夏、内蒙古、陕西、山西、河南、山东	注入海洋	渤海

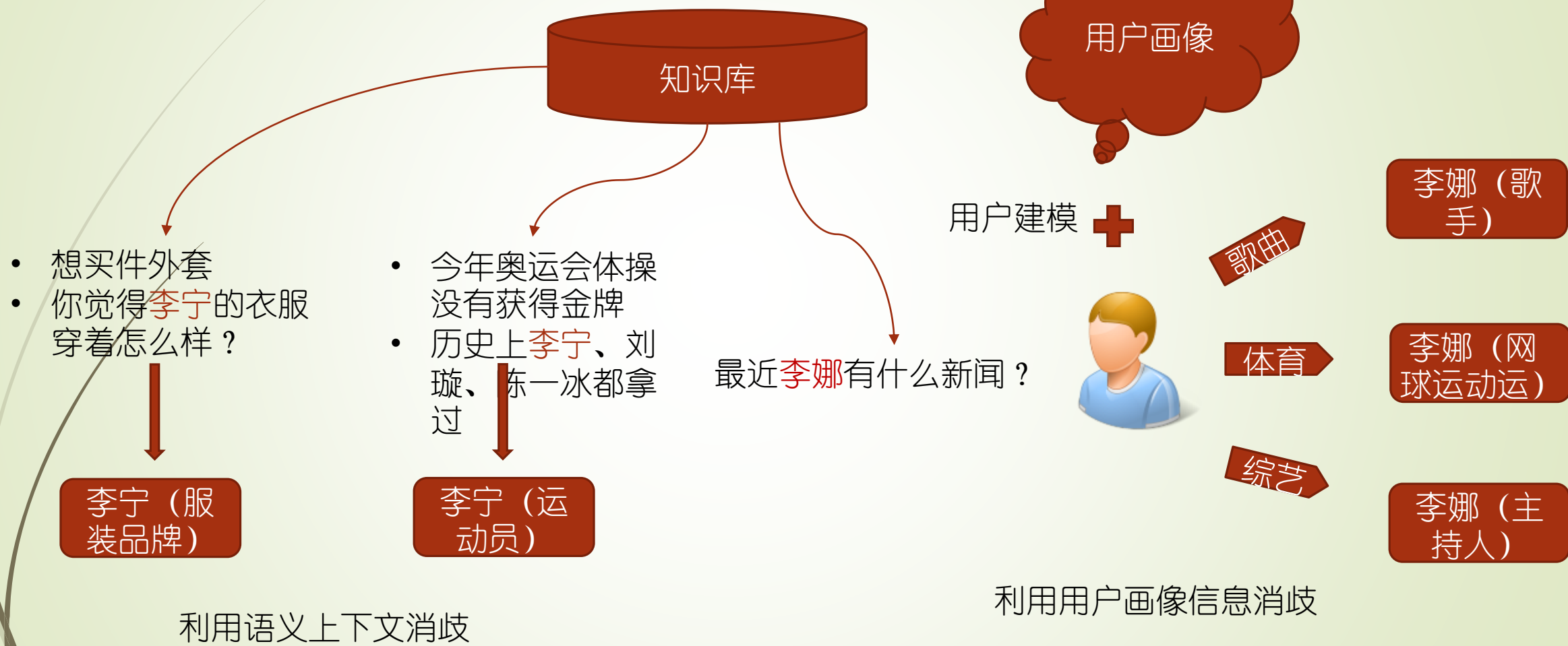
长江
扬子江
河流



瞿塘峡

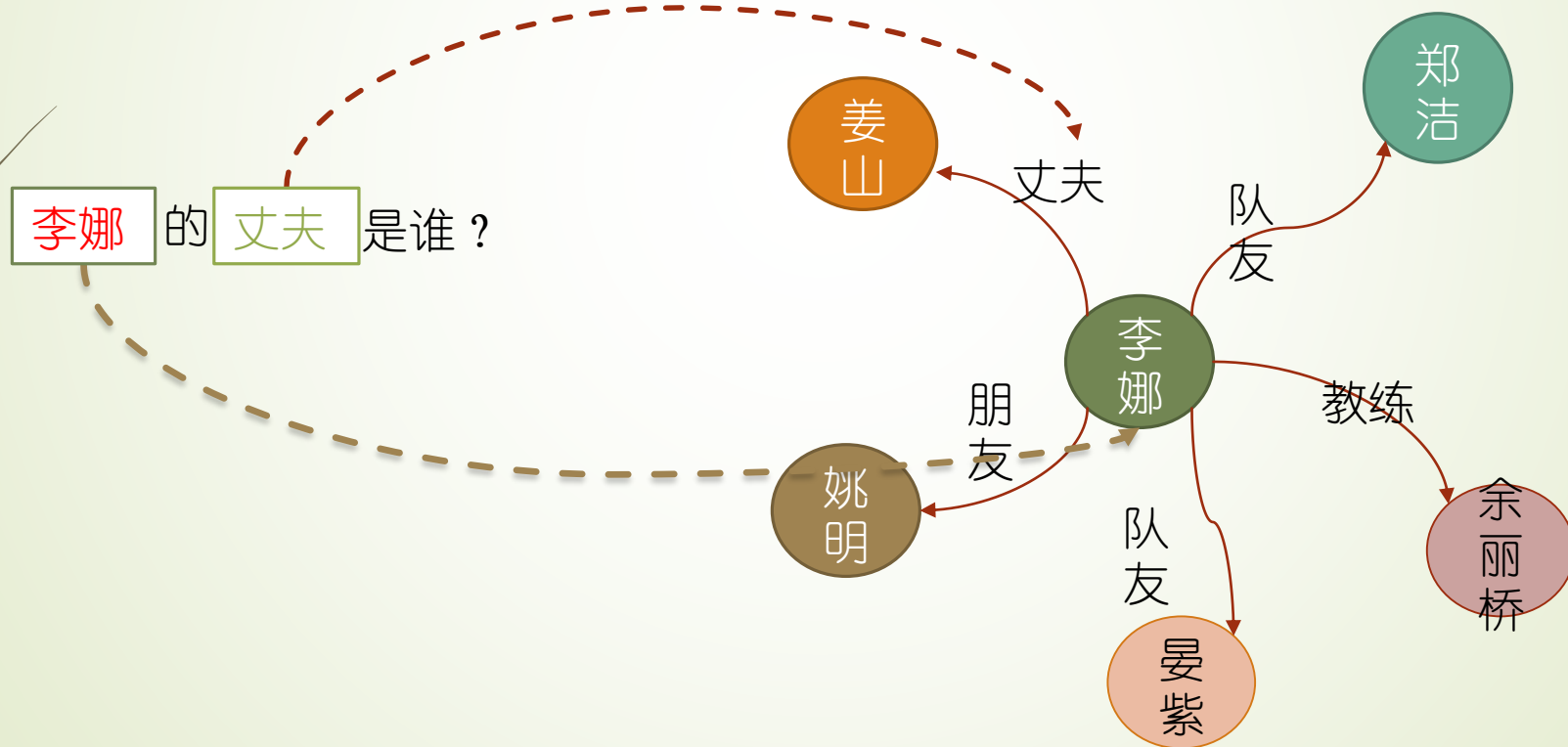
国家 省/州/邦	<div>中国</div> <div>青海, 西藏, 云南, 四川, 重庆, 湖北, 湖南, 江西, 安徽, 江苏, 上海</div>
支流 - 左侧支流 - 右侧支流 城市	<div>雅砻江, 岷江, 嘉陵江, 汉水 乌江, 沅江, 湘江, 赣江</div> <div>宜宾 泸州 重庆 涪陵 万州 宜昌 荆州 岳阳 武汉 黄冈 鄂州 黄石 九江 安庆 池州 铜陵 芜湖 马鞍山 南京 镇江 扬州 南通 上海</div>
源头 - 位置 - 海拔 - 坐标 河口 - 位置 - 海拔 - 坐标	<div>各拉丹冬峰 青海唐古拉山 5,042 m (16,542 ft) 33° 25′ 44″ N 91° 10′ 57″ E</div> <div>东海 江苏南通及上海 0 m (0 ft) 31° 23′ 37″ N 121° 58′ 59″ E</div>
长度 流域面积 流量 - 平均流量 - 最大流量 - 最小流量	<div>6,300 km (3,915 mi) [1]</div> <div>1,808,500 km² (698,266 mi²) [2]</div> <div>30,166 m³/s (1,065,302 ft³/s) [3] 110,000 m³/s (3,884,613 ft³/s) [4] 2,000 m³/s (70,629 ft³/s)</div>

实体消歧



语义解析 Semantic Parsing

- 问句到知识图谱中关系和实体的映射



传统方法

- 基于关键字匹配或者人工模板的方法

<person>的<丈夫|女儿|朋友>是谁？

<person>的<身高|体重|年龄>是多少？



SparQL语句生成

帮我查询我去年从北京发放上海的快递有哪些？

快递
Pattern
挖掘

Type	快递
时间	去年
From	北京
To	上海

Sparql
生成

```
select distinct ?x
where {
  ?x from "北京"; ?x to "上海";
  ?x 时间 ?t;
  FILTER xsd:dateTime(?t) >= "2015-01-01T00:00:00Z"^^xsd:dateTime && xsd:dateTime(?t) < "2016-01-01T00:00:00Z"^^xsd:dateTime
}
```

结构化查询 (sparql)

中国有多少运动员
进入了NBA打球？

有五个人，
分别为：
王治郅
巴特尔
姚明
易建联
孙悦

句法分析

语法分析

```
select
distinct ?name
where {
  ?x
  foaf:name ?name;
  ?x country '中国';
  ?x isA '运动员';
  ?x playsAt 'NBA';
}
```

Sparql查
询语句

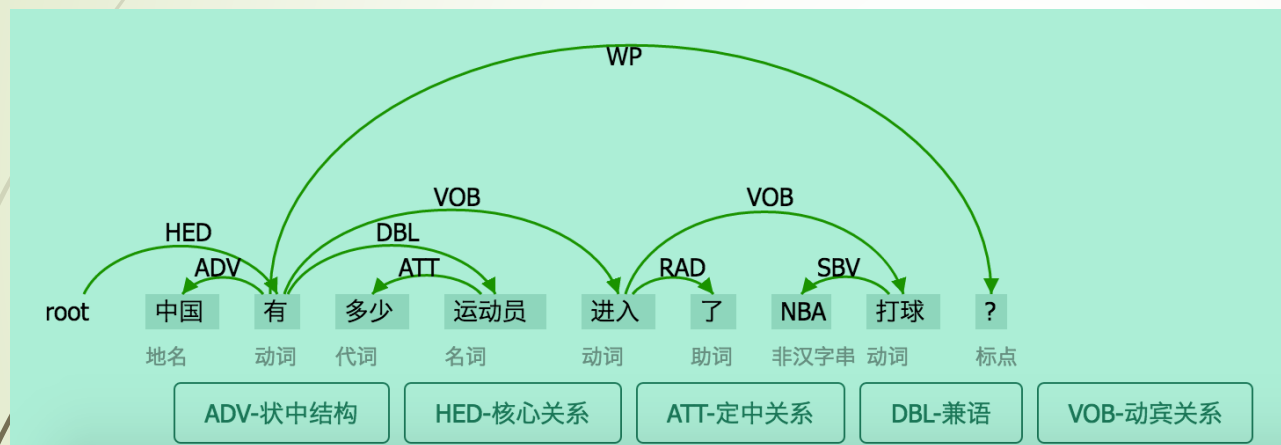
构造答案

知识库

王治郅
巴特尔
姚明
易建联
孙悦

SparQL语句生成

中国有多少运动员进入了NBA打球？



三元组语义表示:

<?x isA "运动员">

<?x country "中国">

<?x playsAt "NBA">



sparql生成:

```
select distinct ?name
```

```
where {
```

```
  ?x foaf:name ?name; ?x country '中国';
```

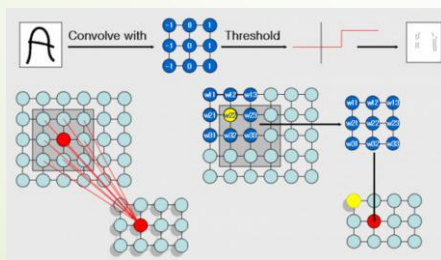
```
  ?x isA '运动员';
```

```
  ?x playsAt 'NBA';
```

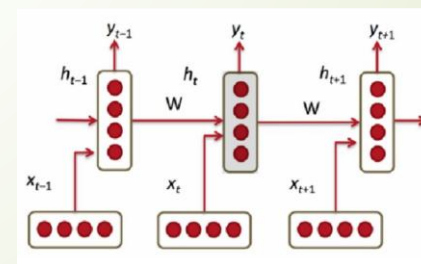
```
}
```

基于向量空间建模

- 使用向量空间描述自然语言问句以及知识库中的实体和关系
- 不需要人工设计规则、提取特征
- 利用收集的问题-答案对进行各向量表征的自动训练
- 通过比较问句和备选答案在向量空间中的距离实现对输入问题的回答




CNN



RNN

挑战

- 语义理解难度大（缺乏严格的句法信息，需要人工构造大量模板，结合上下文信息比较复杂）
- 三元组表示能力有限
- 如何推理及对新知识的验证



展望

- 知识图谱在众多垂直领域会得到更好的应用，如客服、医疗、咨询等。
- 深度的关系实体抽取可以极大的丰富现有知识库
- 问答系统依赖大量构建的垂直领域知识



谢谢！