PlantData

行业知识图谱构建与应用101

链接数据 洞察价值

LINKED DATA & SMART DATA

王昊奋
深圳狗尾草智能科技有限公司
Wang_haofen@gowild.cn

胡芳槐
上海海翼知信息科技有限公司
hufh@hiekn.com

Tutorial 简介

PlantData

- 以行业知识图谱为主
- 偏向于行业知识图谱应用及相关的技术
- 结合行业应用的一些最佳实践及相关的组件
- 针对行业知识图谱在行业中的应用进行实战演示

Tutorial 主要内容

PlantData

- 行业知识图谱概述，包括行业图谱简介，行业知识图谱的应用及挑战，以及行业知识图谱生命周期管理。

- 行业知识图谱关键技术，包括行业知识图谱生命周期中各过程的相关技术、现有可用的工具，以及各过程中的最佳实践及相关组件。

- 行业知识图谱应用实战，以金融证券行业应用为例，演示知识图谱从知识建模、知识抽取到行业应用的全过程。

Tutorial 目标听众

PlantData

- 知识图谱学习者，对知识图谱在行业应用感兴趣的技术人员。
- 各行业应用中想引入知识图谱相关技术的知识及数据管理人员，尤其是有行业知识库构建及上层问答搜索等有需求的。
- 希望了解知识图谱如何在行业中应用的管理决策者。

Tutorial 预期目标

PlantData

- 了解行业知识图谱相关概念及其在行业中的现有应用，理解其给行业应用带来的价值。

- 理解知识图谱在行业中应用的相关挑战与生命周期，理解生命周期各过程的基本目标及相关组件。

- 对行业知识图谱应用相关的技术进行熟悉，了解有哪些现有的工具可以使用和相关注意事项、以及一些行业应用的最佳实践。

Tutorial 听众的知识基础

PlantData

- RDF：资源描述框架
- OWL： RDF Schema 的扩展
- SPARQL：RDF查询语言

PlantData

行业知识图谱概述

行业知识图谱简介

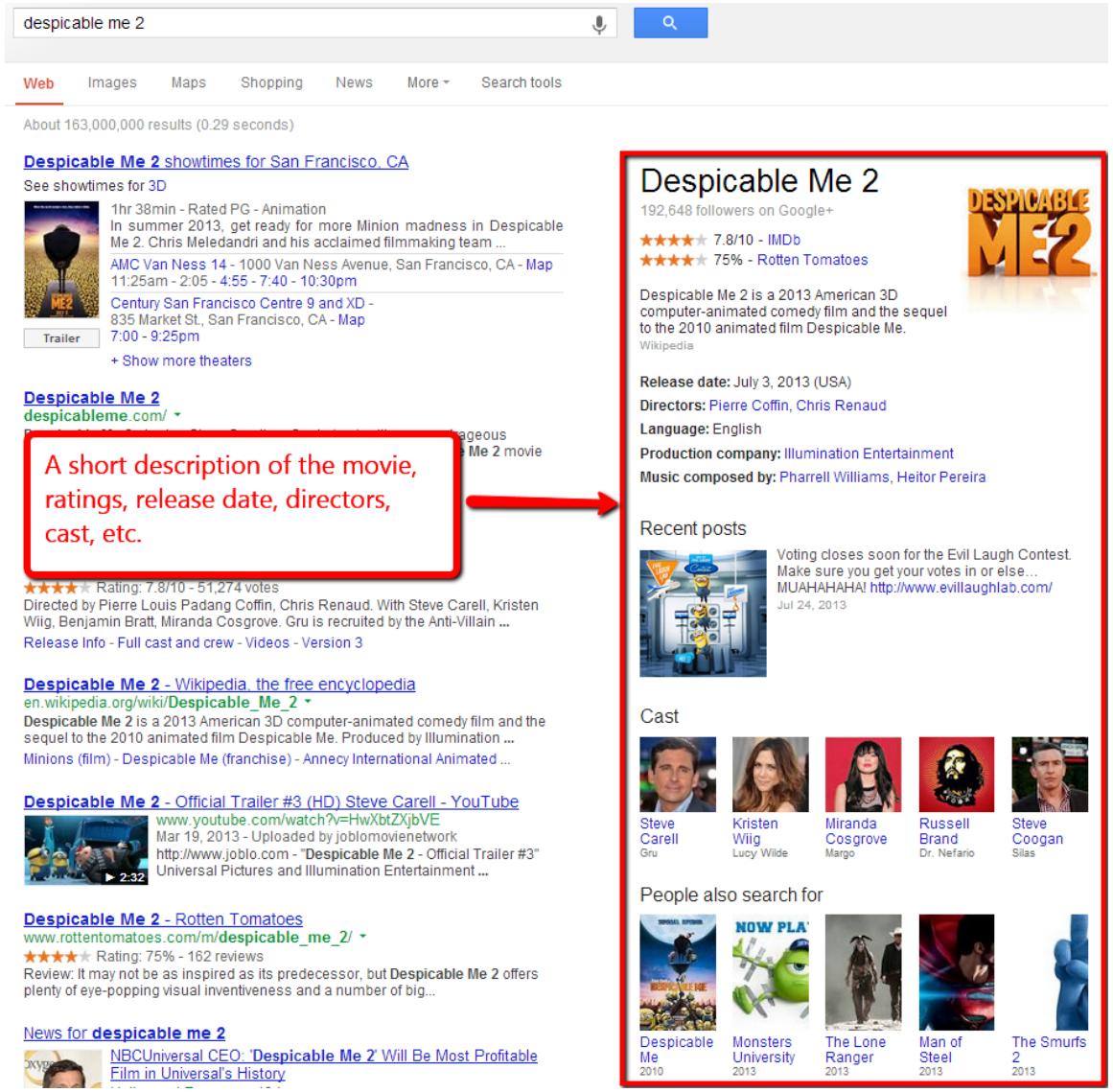
行业知识图谱应用

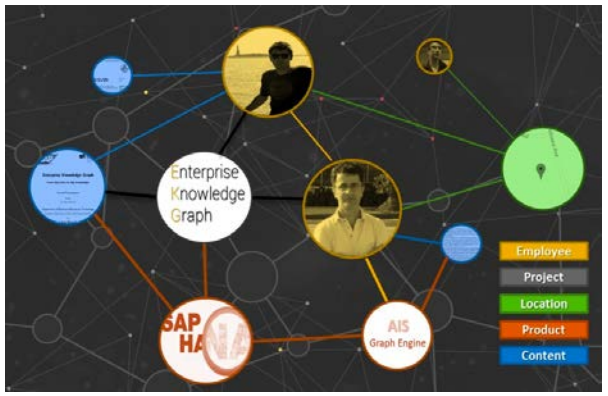
KG应用挑战

行业知识图谱生命周期

谷歌知识图谱: Thins not strings

PlantData





知识图谱助力人工智能应用

PlantData

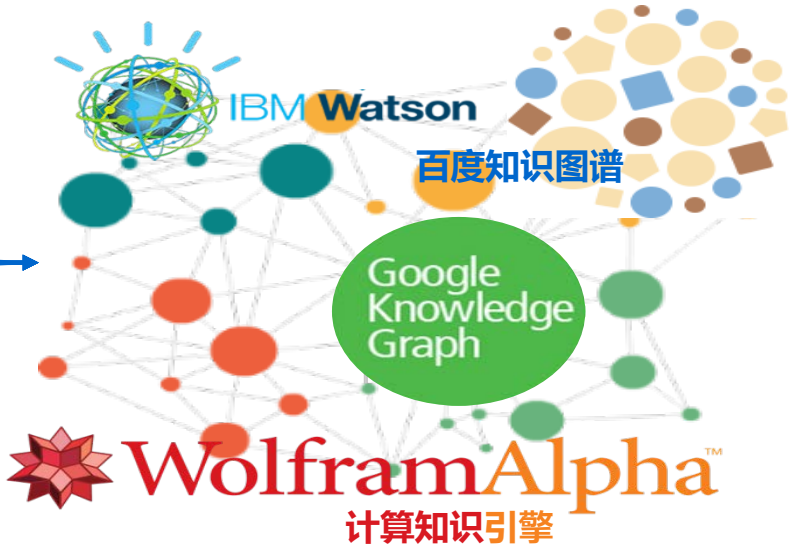
- Google
- Bing
- 百度

- 微软小冰
- 公子小白

聊天机器人

- IBM Watson Health

决策支持



- Siri
- Google Now
- 微软小娜
- 百度度秘

私人助理

- Apple Watch
- Ticwatch

智能硬件

- 智能家居
- 智能厨房

智能硬件

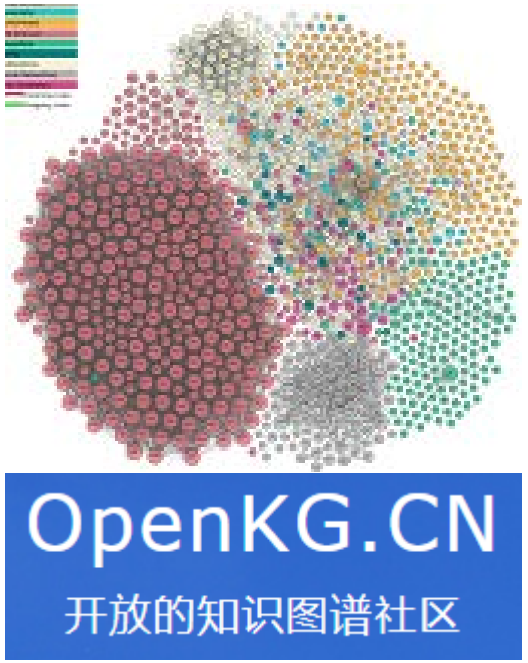
通用知识图谱

PlantData

- Google所提出的知识图谱是面向全领域的通用知识图谱。
- 通用知识图谱主要应用于面向互联网的搜索、推荐、问答等业务场景。
- 通用知识图谱，它强调的是广度，因而强调更多的是实体，很难生成完整的全局性的本体层的统一管理。

通用知识图谱相关项目

PlantData



语言学类：

WordNet

MIT - ConceptNet5的中文部分

汉语开放词网(Chinese Open WordNet)

百科类：

DBpedia

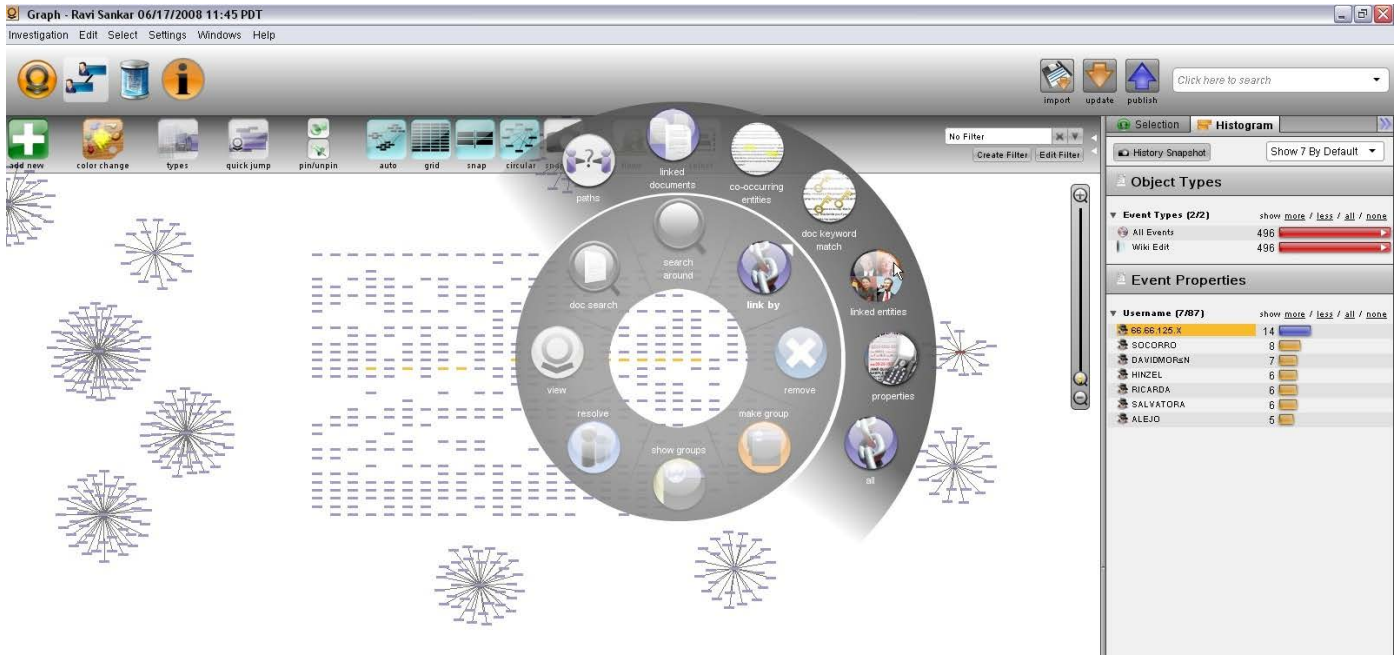
中文通用百科知识图谱（CN-DBpedia）

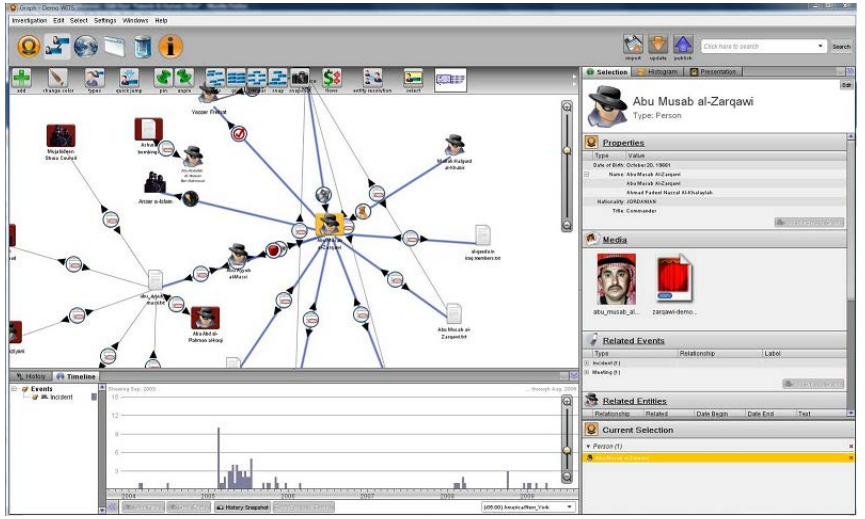
Zhishi.me

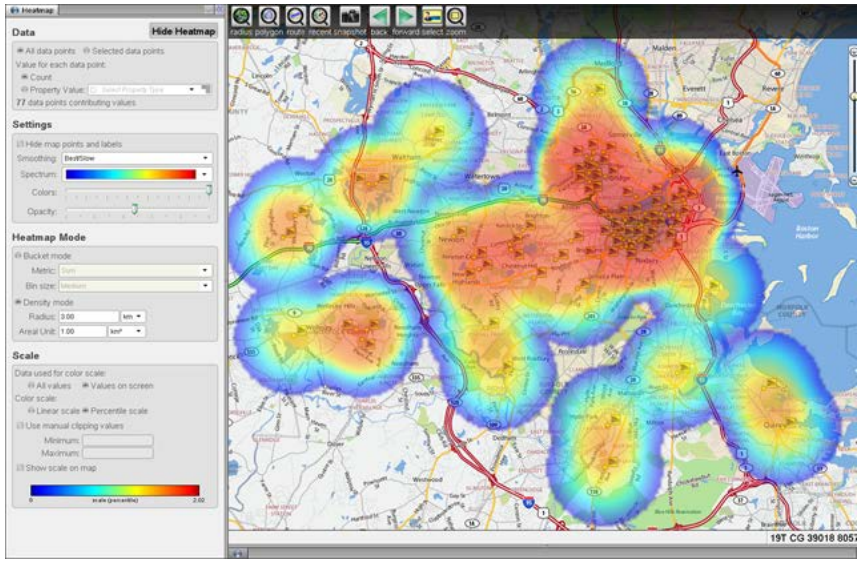
PKU-PIE 知识库

行业知识图谱：Palantir

PlantData







行业知识图谱

PlantData

- 行业知识图谱指面向特定领域的知识图谱。
- 用户目标对象需要考虑行业中各种级别的人员，不同人员对应的操作和业务场景不同，因而需要一定的深度与完备性。
- 行业知识图谱对准确度要求非常高，通常用于辅助各种复杂的分析应用或决策支持。
- 有严格与丰富的数据模式，行业知识图谱中的实体通常属性比较多且具有行业意义。

行业知识图谱数据的特点

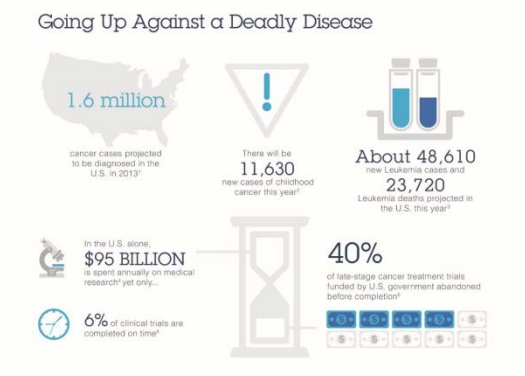
PlantData

- 数据来源多：内部数据、互联网数据、第三方数据
- 数据类型多：包含结构化、半结构化、非结构化数据，且后两者越来越多
- 数据模式无法预先确定：模式在数据出现之后才能确定；数据模式随数据增长不断演变
- 数据量大：在大数据背景下，行业应用的数据的数量通常都以亿级别计算，存在通常在TB、PB级别甚至更多

行业知识图谱应用一览

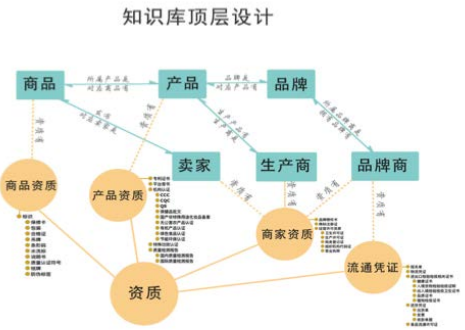
PlantData

金融证券

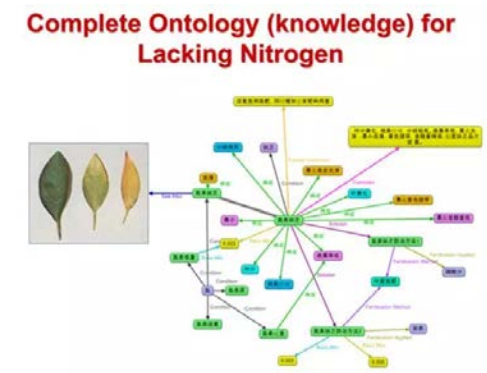


生物医疗

图书情报

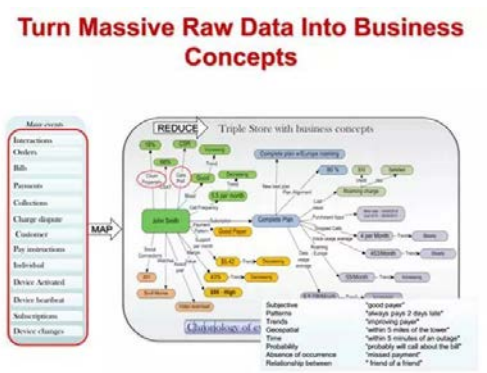


电商



农业

政府

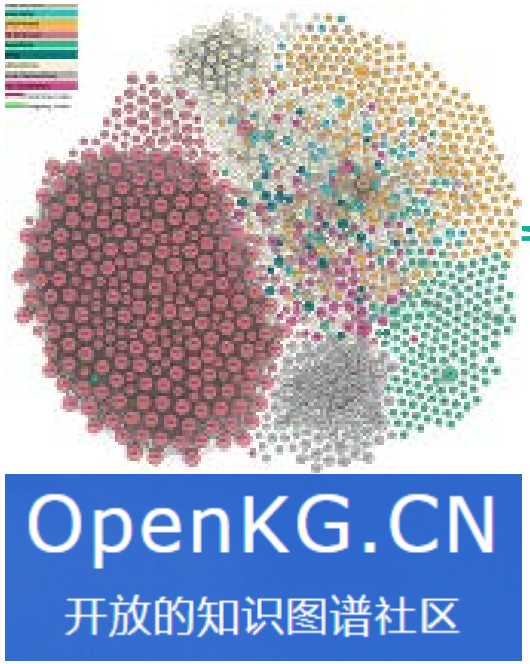


电信

出版

行业知识图谱相关项目

PlantData



- Linked life data
- Open Government Data
- Geonames
- 有色行业产业链图谱
- 中医医案知识图谱
- 华人家谱关联数据集

通用知识图谱 vs 行业知识图谱

PlantData

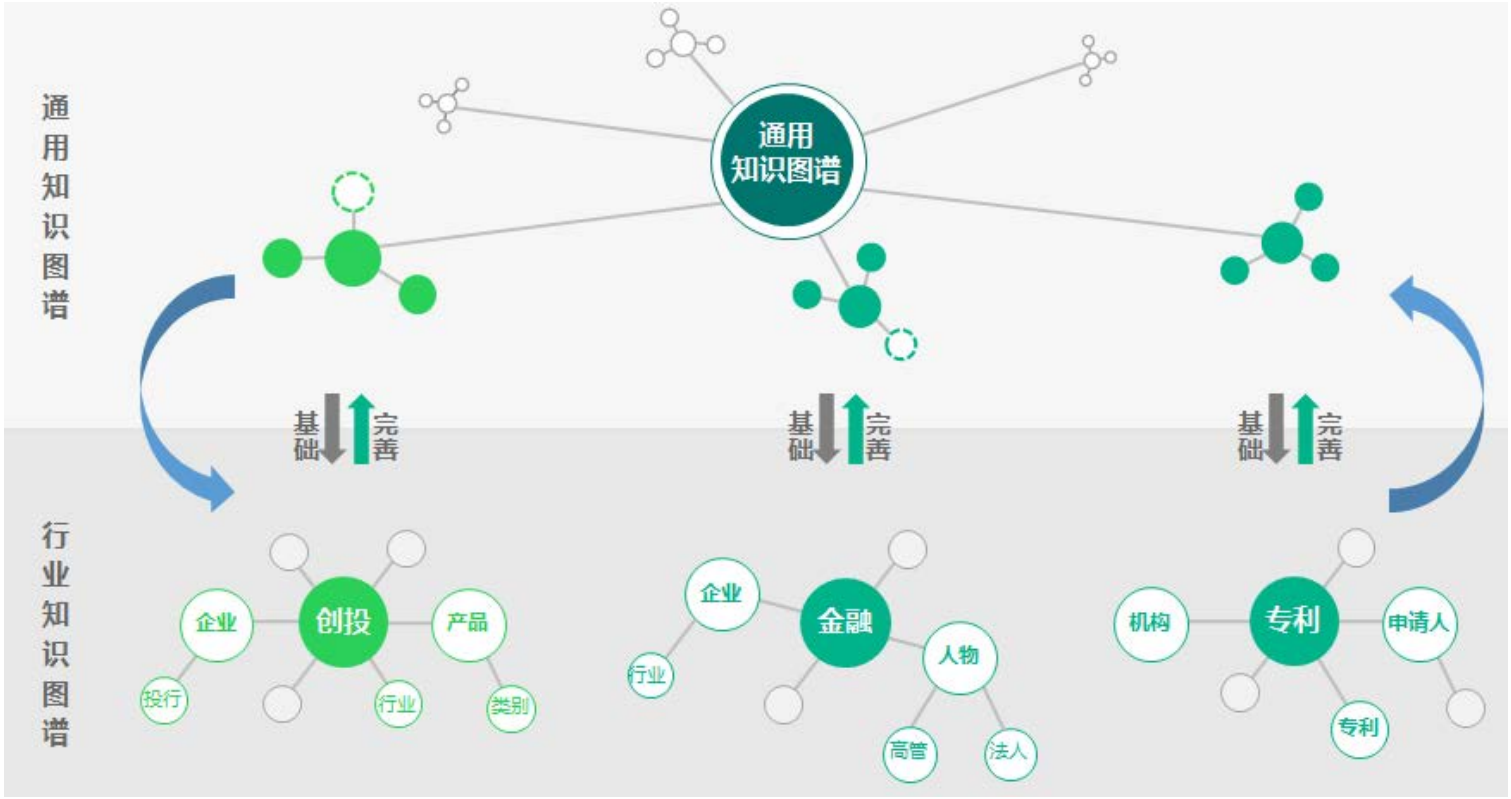
- ✓ 面向通用领域
- ✓ 以常识性知识为主
- ✓ “结构化的百科知识”
- ✓ 强调知识的广度
- ✓ 使用者是普通用户

- ✓ 面向某一特定领域
- ✓ 基于行业数据构建
- ✓ “基于语义技术的行业知识库”
- ✓ 强调知识的深度
- ✓ 潜在使用者是行业人员

通用知识图谱 + 行业知识图谱

PlantData

- 通用知识图谱的广度，行业知识图谱的深度，相互补充，形成更加完善的知识图谱。
- 通用知识图谱中的知识，可以作为行业知识图谱构建的基础；而构建的行业知识图谱，再融合到通用知识图谱中。



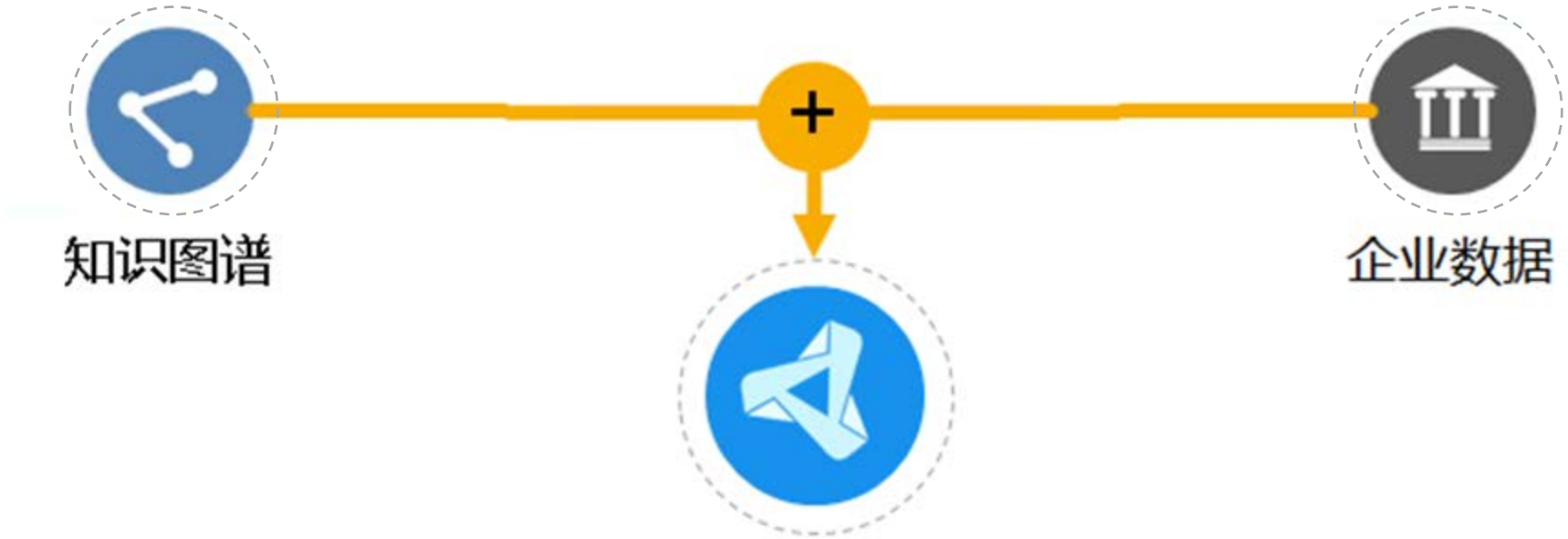
PlantData

行业知识图谱概述

- 行业知识图谱简介
- 行业知识图谱应用
- KG应用需求与挑战
- 行业知识图谱生命周期

金融证券——企业知识图谱

PlantData



企业知识图谱

企业基础数据
投资关系
任职关系

企业专利数据
企业招投标数据
企业招聘数据

企业诉讼数据
企业失信数据
企业新闻数据

企业知识图谱应用——企业风险评估

PlantData

基于企业的基础信息、投资关系、诉讼、失信等多维度关联数据，利用图计算等方法构建科学、严谨的企业风险评估体系，有效规避潜在的经营风险与资金风险。

用户群体

银行、担保、投行、政府……

应用环节

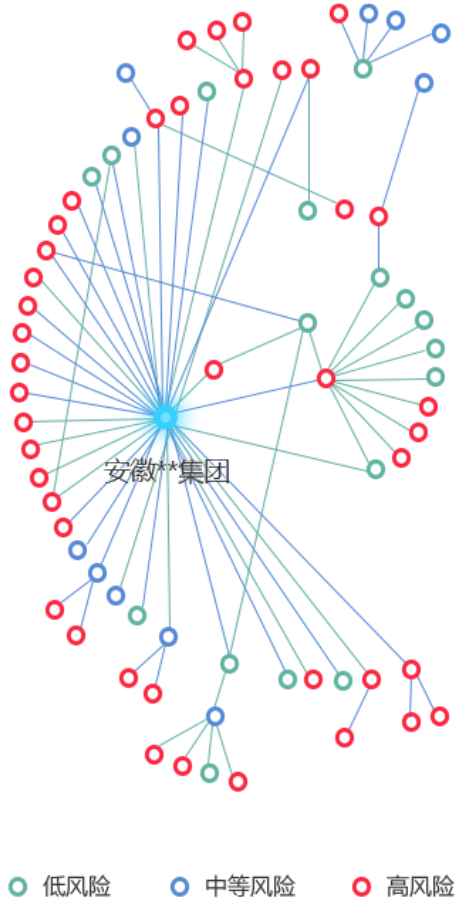
客户资源分类管理

信贷前期风险评估

采购企业风险审核

招投标企业资质评级

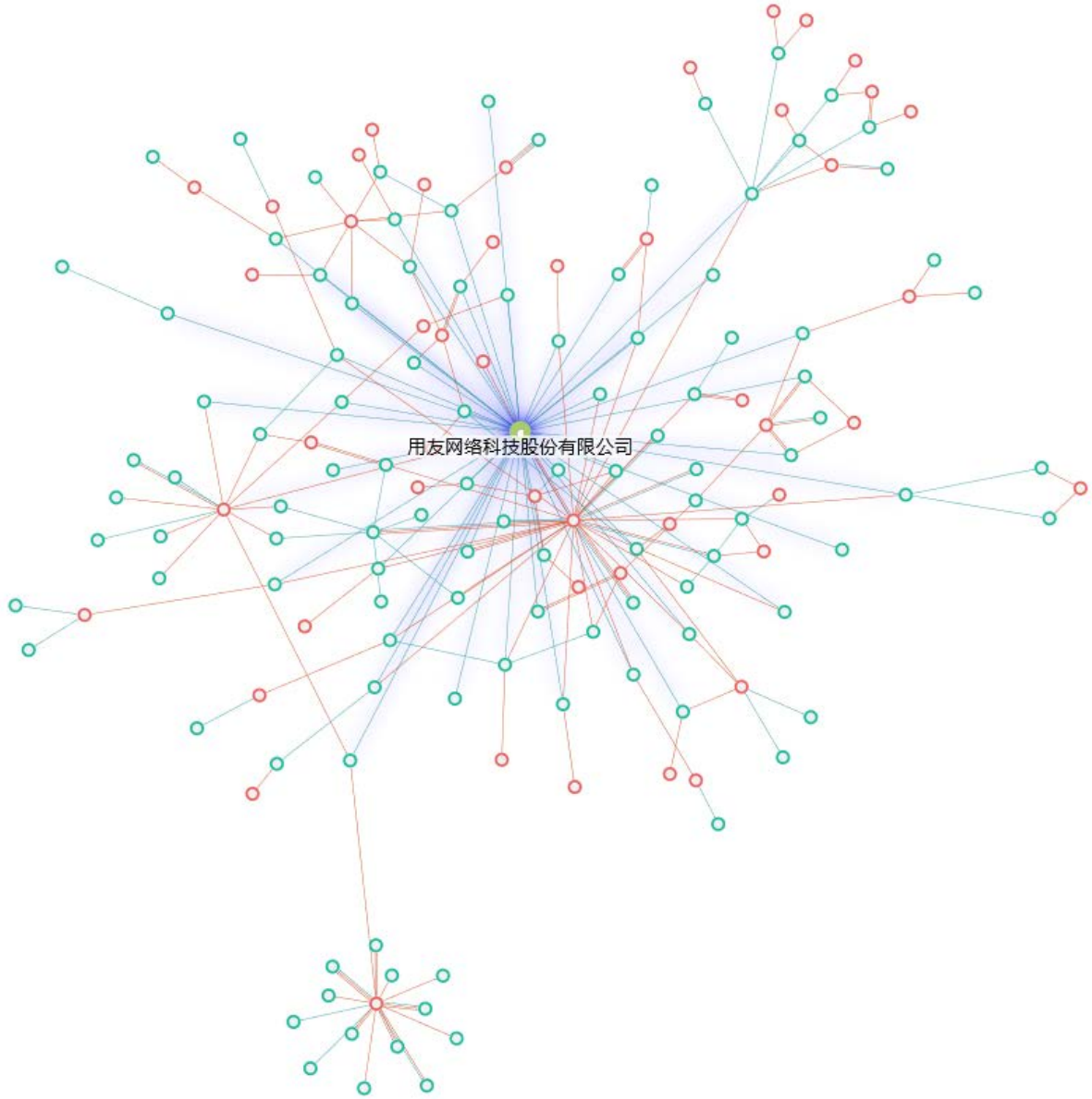
……



企业知识图谱应用——企业社交图谱查询

PlantData

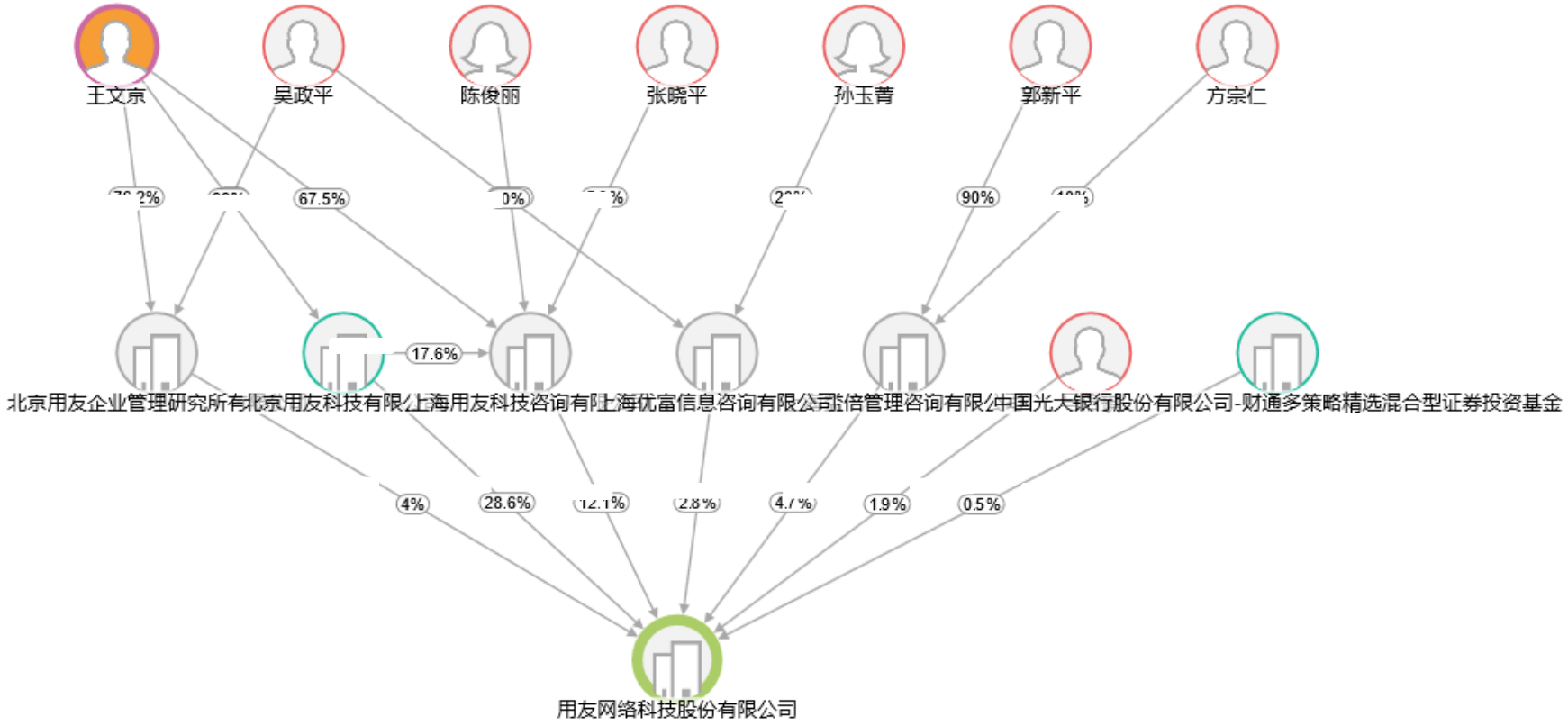
- 基于投资、任职、专利、招投标、涉诉关系以目标企业为核心向外层层扩散，形成一个网络关系图，直观立体展现企业关联。



企业知识图谱应用——企业最终控制人查询

PlantData

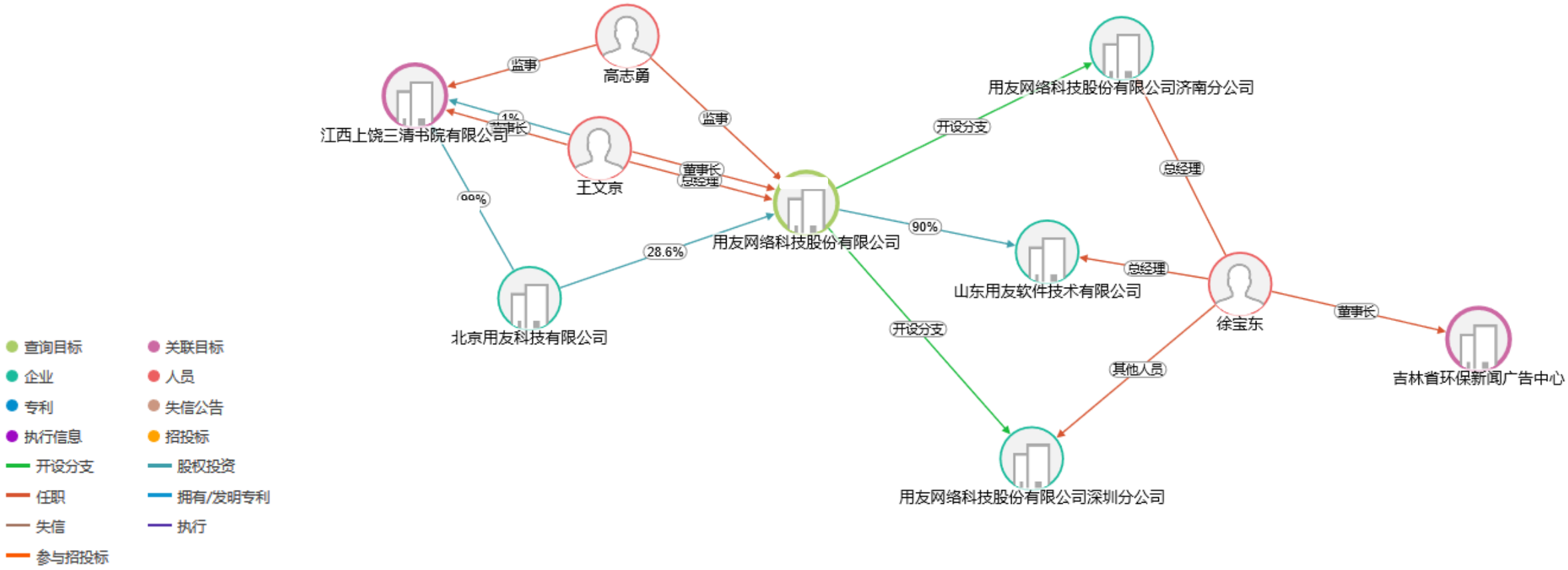
- 基于股权投资关系寻找持股比例最大的股东，最终追溯至自然人或国有资产管理部门。



企业知识图谱应用——企业之间路径发现

PlantData

- 在基于股权、任职、专利、招投标、涉诉等关系形成的网络关系中，查询企业之间的最短关系路径，衡量企业之间的联系密切度。



企业知识图谱应用——初创企业融资发展历程

PlantData

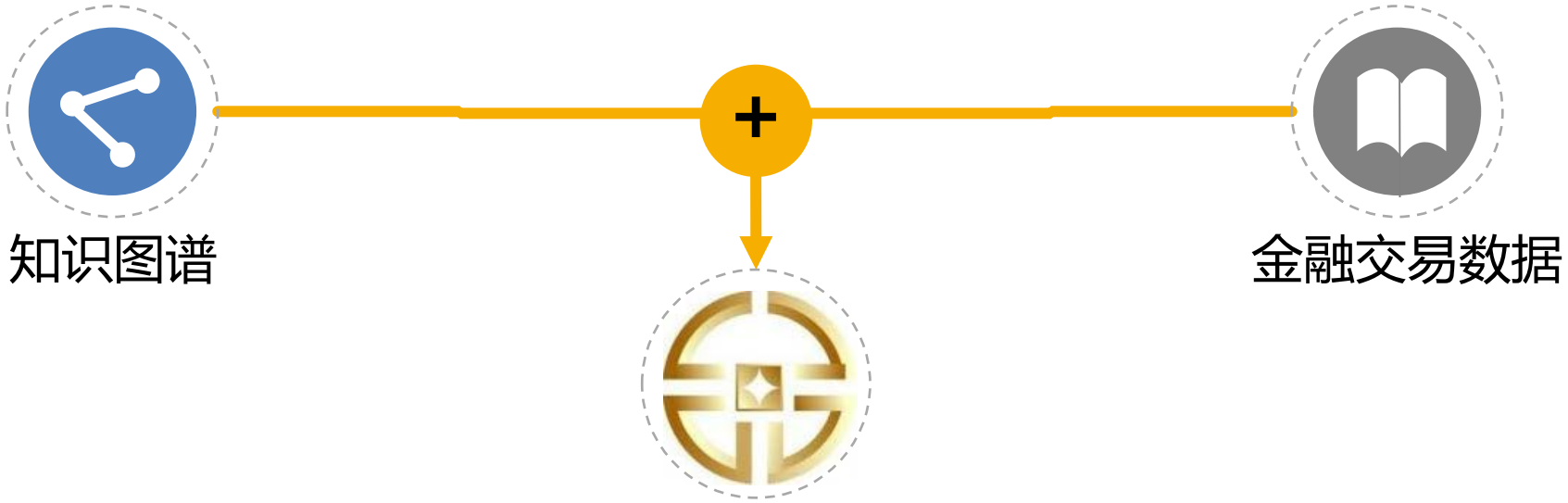
- 基于企业知识图谱中的投融资事件发生的时间顺序，记录企业的融资发展历程。

企业知识图谱应用——上市企业智能问答

PlantData

金融证券——金融交易知识图谱

PlantData



金融交易知识图谱

企业知识图谱

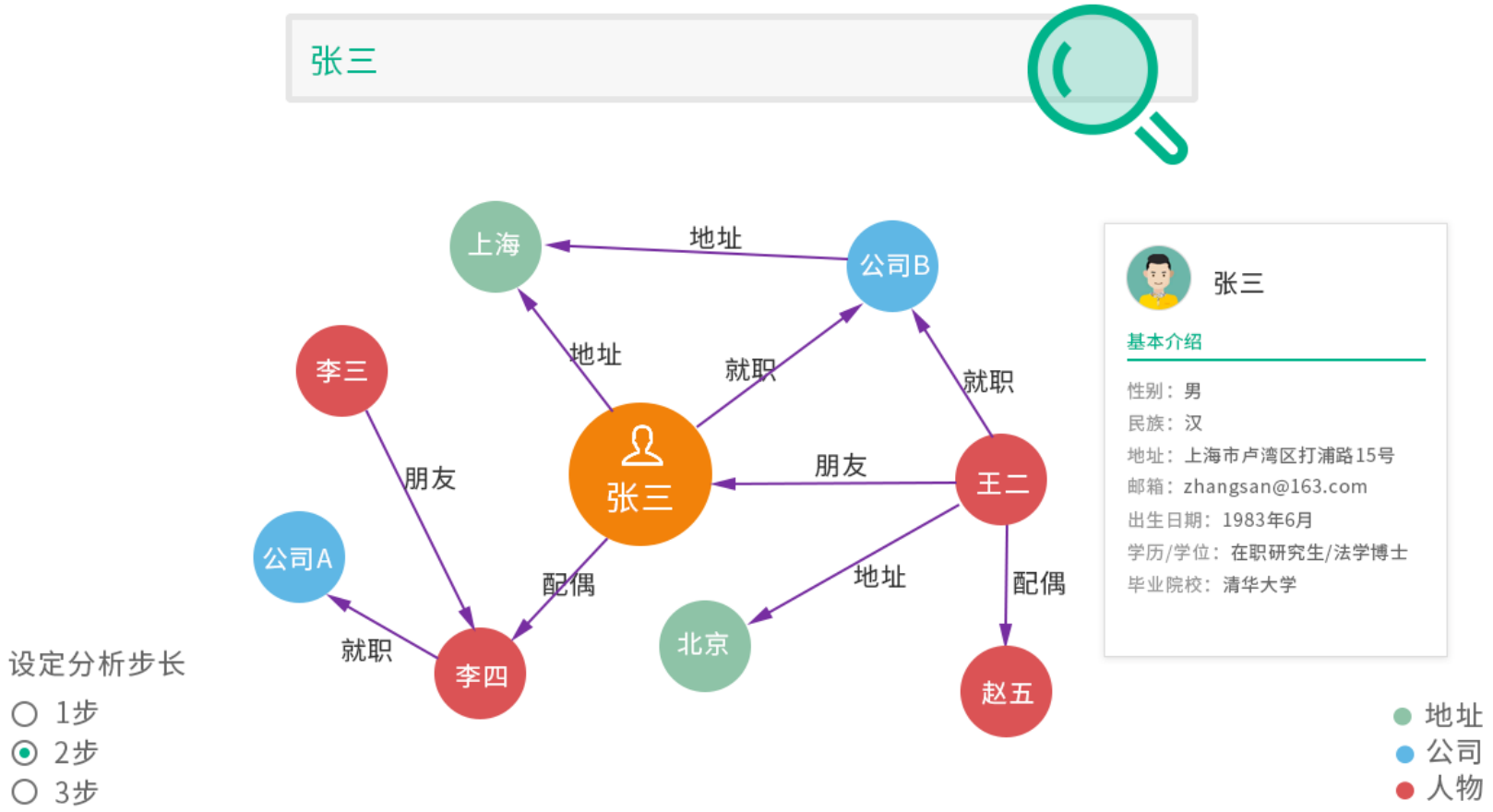
交易客户数据
客户之间的关系

交易行为数据

金融交易知识图谱应用——辅助信贷审核

PlantData

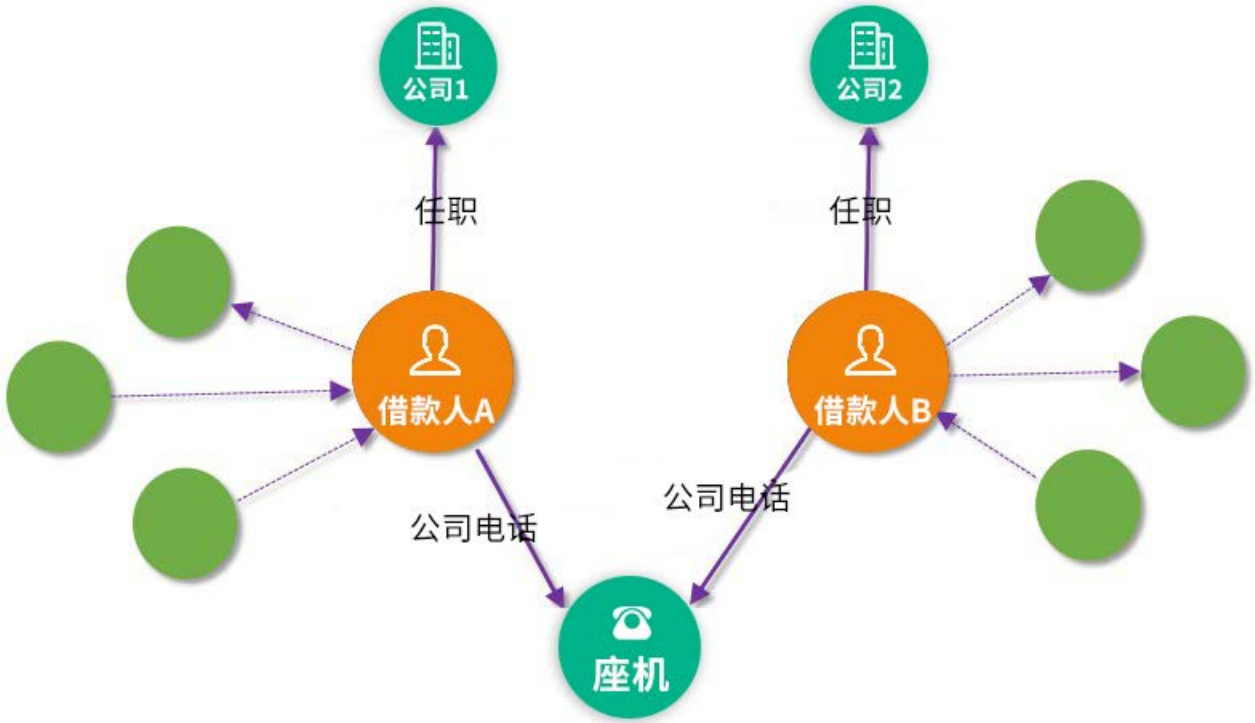
- 基于知识图谱数据的统一查询，全面掌握客户信息；避免由于系统、数据等孤立造成的信息不一致造成信用重复使用、信息不完整等问题。



金融交易知识图谱应用——反欺诈（1）

PlantData

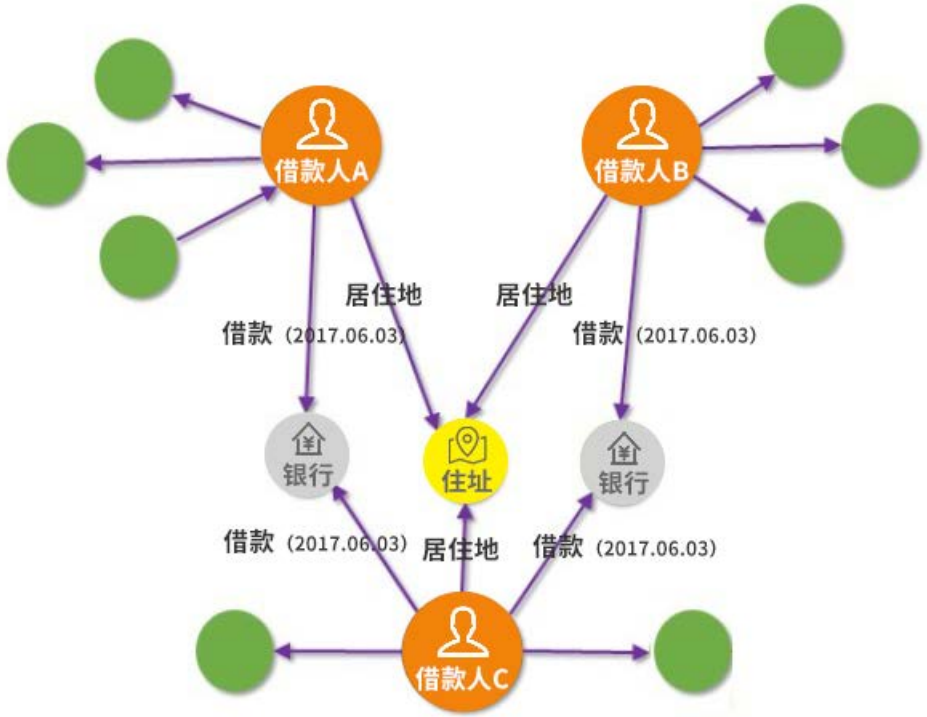
不一致性验证可以用来判断一个借款人的欺诈风险，类似交叉验证。比如借款人A和借款人B填写的是同一个公司电话，但借款人A填写的公司和借款人B填写的公司完全不一样，这就成了一个风险点，需要审核人员格外的注意。



金融交易知识图谱应用——反欺诈（2）

PlantData

组团进行欺诈的成员会用虚假的身份去申请贷款，但部分信息是共享的。如下图可以看出贷款人A、B和C之间没有直接的关系，但通过知识图谱可以很容易的看出这三者之间都共享着某一部分信息，存在一定的组团骗贷风险。



金融证券——其它应用场景

PlantData

- 异常分析（异常交易、异常客户）
- 失联客户管理
- 精准营销
- 智能投研
- 智能公告
- ……

生物医疗——医疗知识图谱

PlantData

医疗知识图谱

医疗专业知识
医疗文献
医疗常识

电子病历大数据
医案

现有医疗资源
疾病库
指南与规范

医疗知识图谱应用——中医药知识平台

PlantData

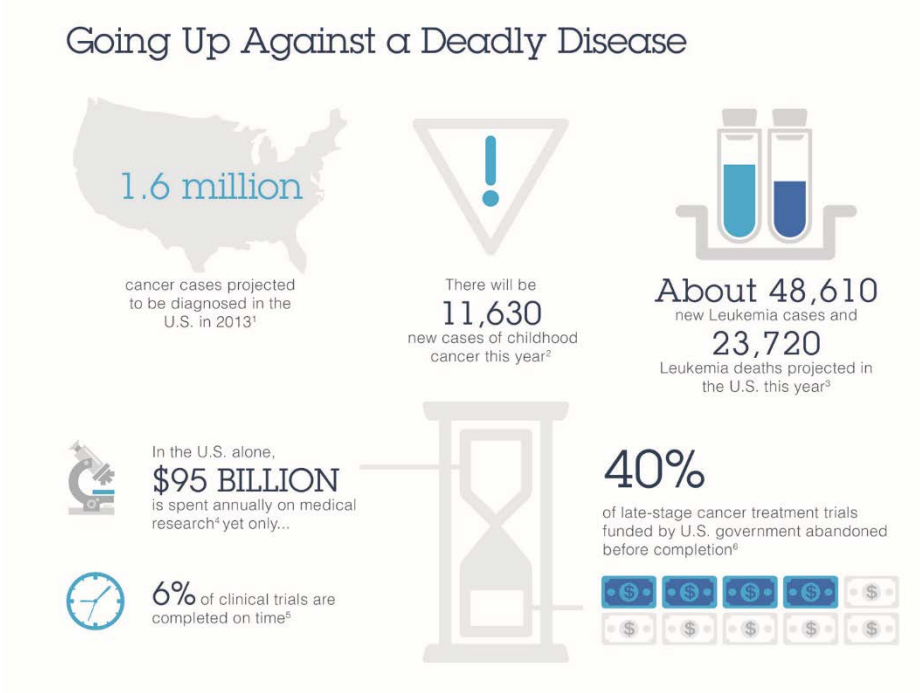
- 针对中医药知识体系系统梳理、建模和展示
- 以图形可视化方式展示核心概念之间的关系
- 辅助中医专家厘清学术发展脉络，浏览中医知识，发现知识点之间的联系。
- 与阅读文献等手段相比，可大幅度节约知识检索获取时间。

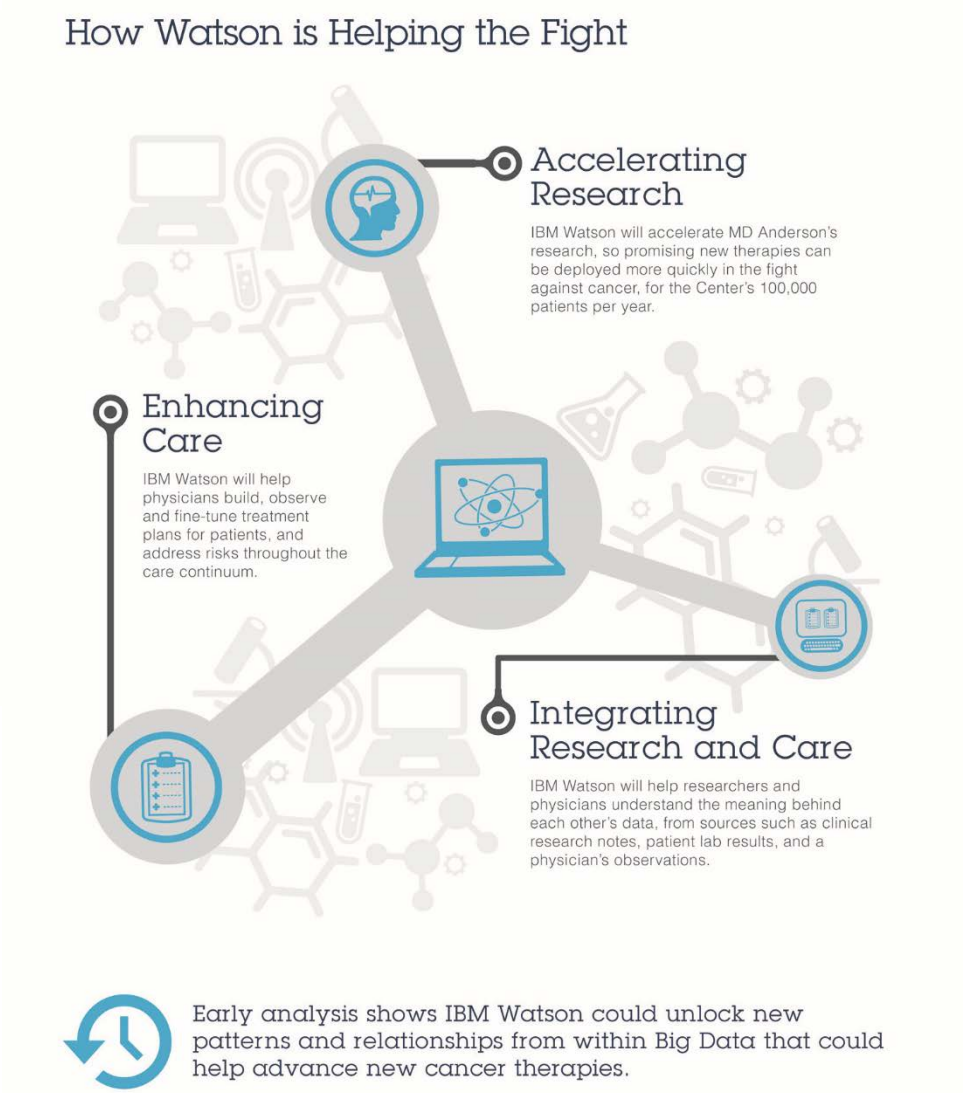
http://www.tcmkb.cn

生物医疗——Watson辅助诊断与治疗

PlantData

安德森癌症中心联合IBM Watson开展终结癌症的任务。





生物医疗——Open PHACTS 新药物发现

PlantData

Bringing together **pharmacological data resources** in an integrated, interoperable infrastructure

Explore.

Researchers can use Open PHACTS to access vast amounts of pharmacological data, all from a single, simple interface

Build.

Developers get free access to the Open PHACTS API, to query the pharmacological data resources in our integrated triple store

Join.

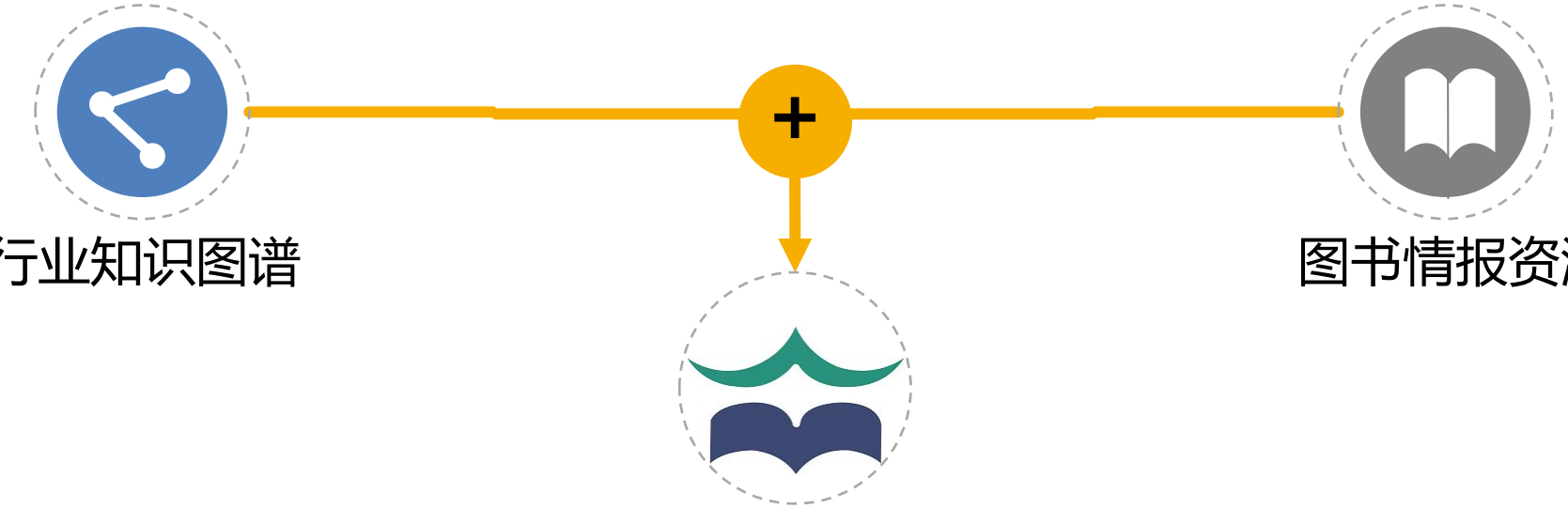
Members of the Open PHACTS Foundation get prioritised access to data, support and updates, as well as training opportunities

欧盟重大联合攻关项目

面向药物研发的开放数据访问平台开发，其核心技术就是采用语义技术为有关研究人员提供高效的数据访问技术环境的支持。

图书情报——图情资源知识图谱

PlantData



图情资源知识图谱

图书馆分类学体系
特定方向的知识体系

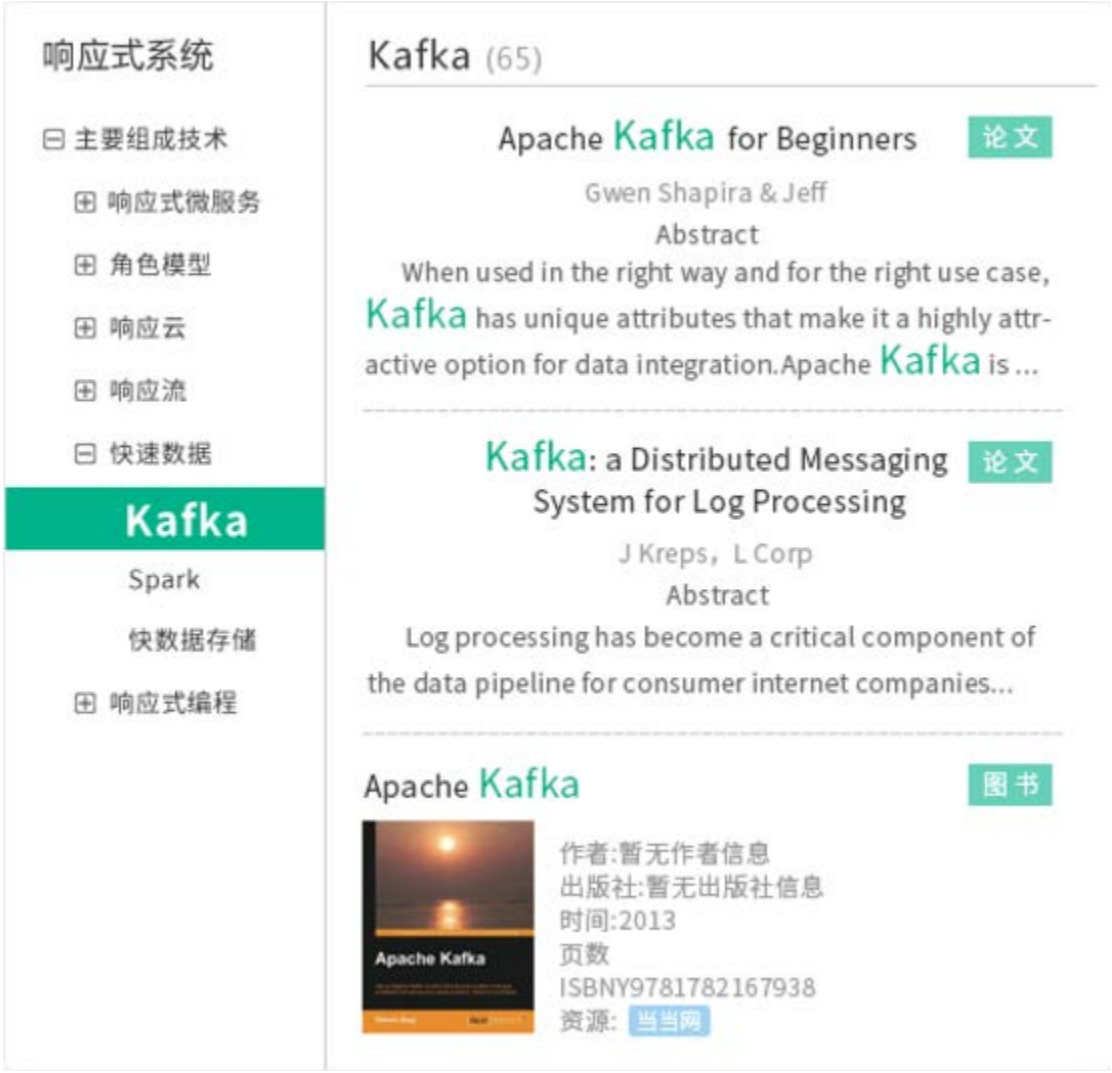
图书
期刊
论文
专利
报刊

百科数据
行业网站数据

图情资源知识图谱应用——知识导航与资源展示

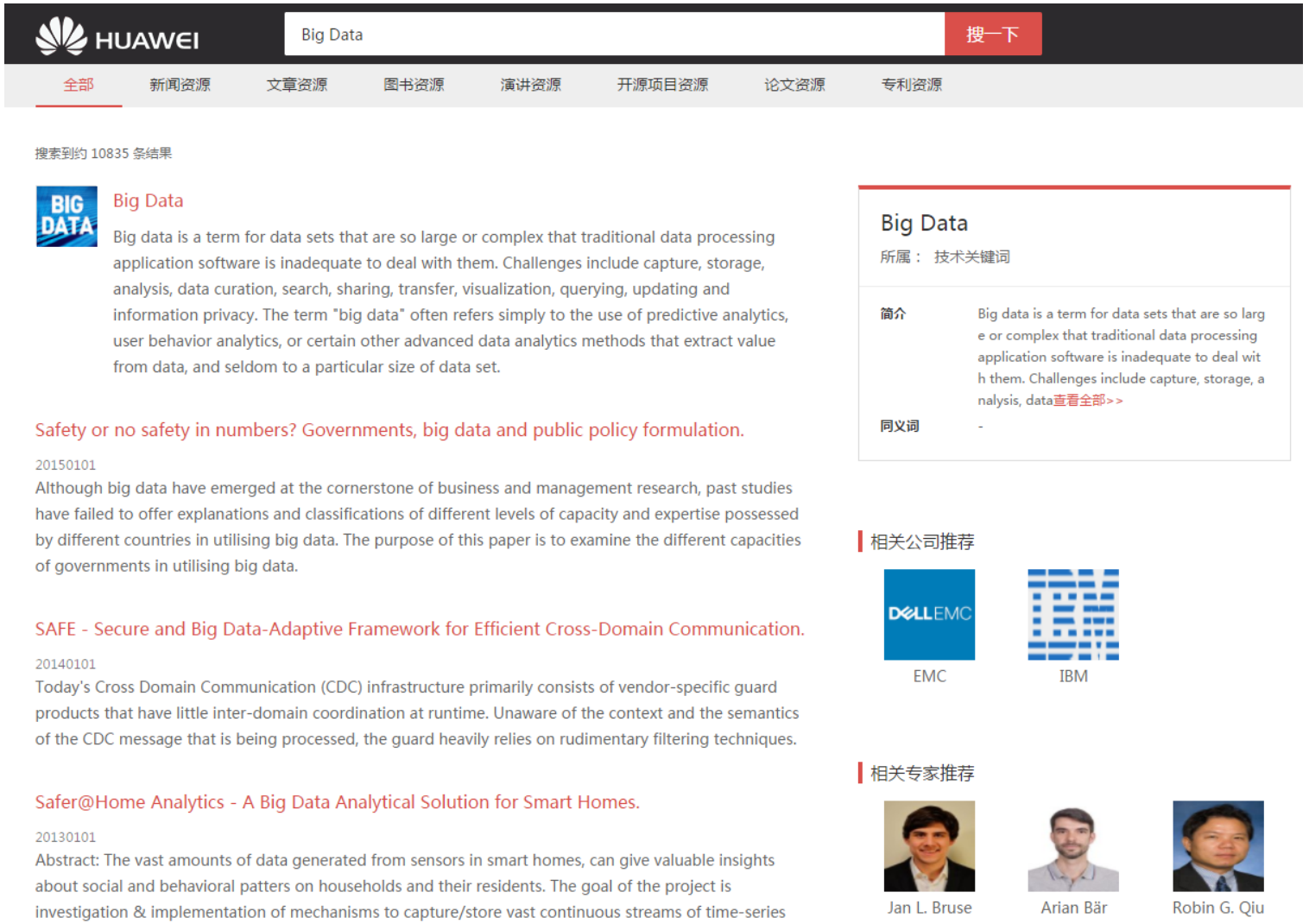
PlantData

- 使用知识图谱中的知识体系进行知识导航，引导用户学习知识体系，以及通过实体链接所关联的资源。



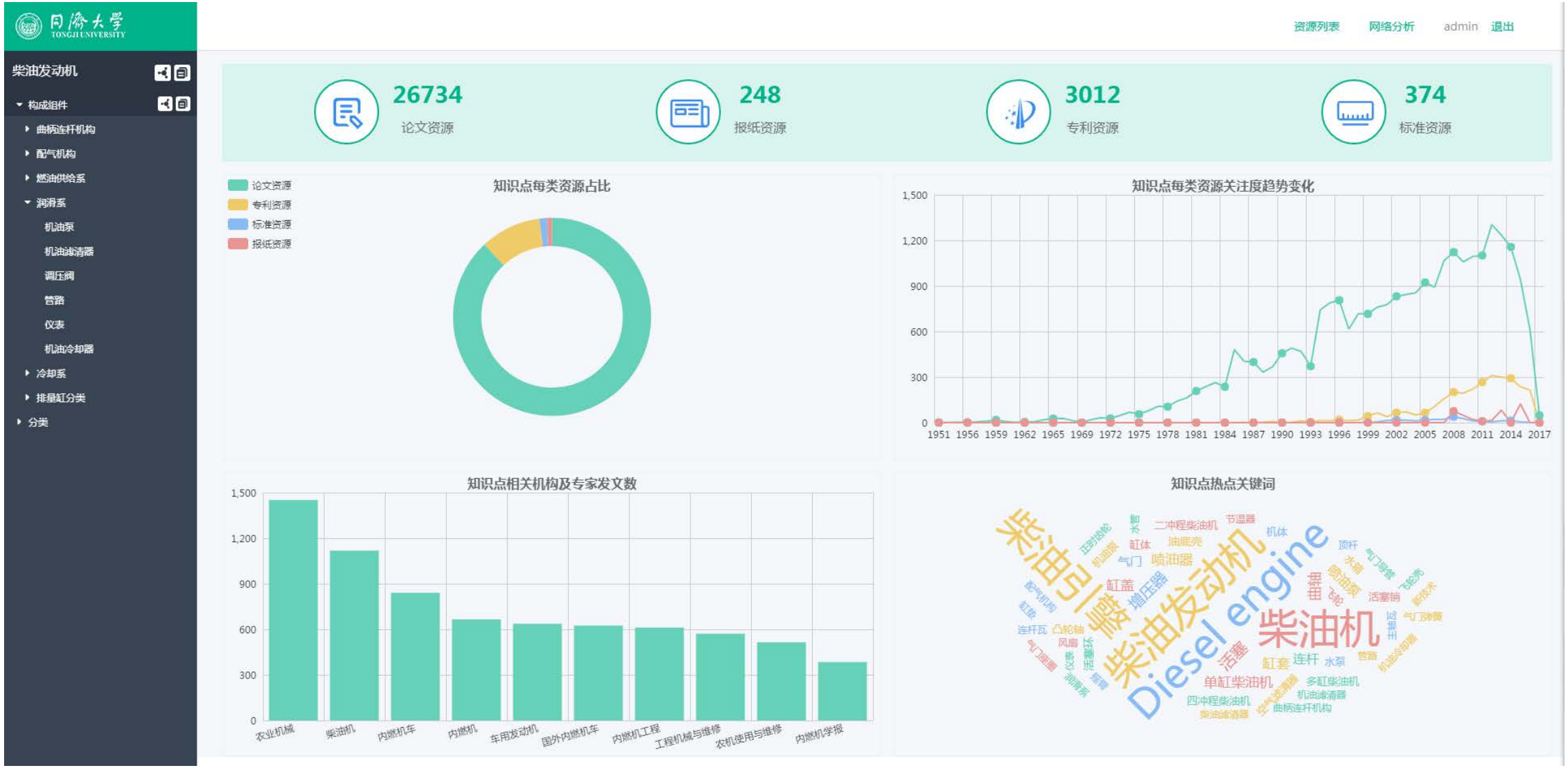
图情资源知识图谱应用——知识点推荐与搜索

PlantData



图情资源知识图谱应用——图情资源统计

PlantData



知识图谱行业应用——其它行业

PlantData

- 农业
  - 识别作物危害
- 政府行业
  - 政府大数据管理
- 客服系统
  - 基于知识图谱的智能客服系统
- ……

PlantData

行业知识图谱概述

- 行业知识图谱简介
- 行业知识图谱应用
- KG应用挑战
- 行业知识图谱生命周期

从数据库发展到大数据时代，企业希望融合使用全量数据

PlantData

DB → BD

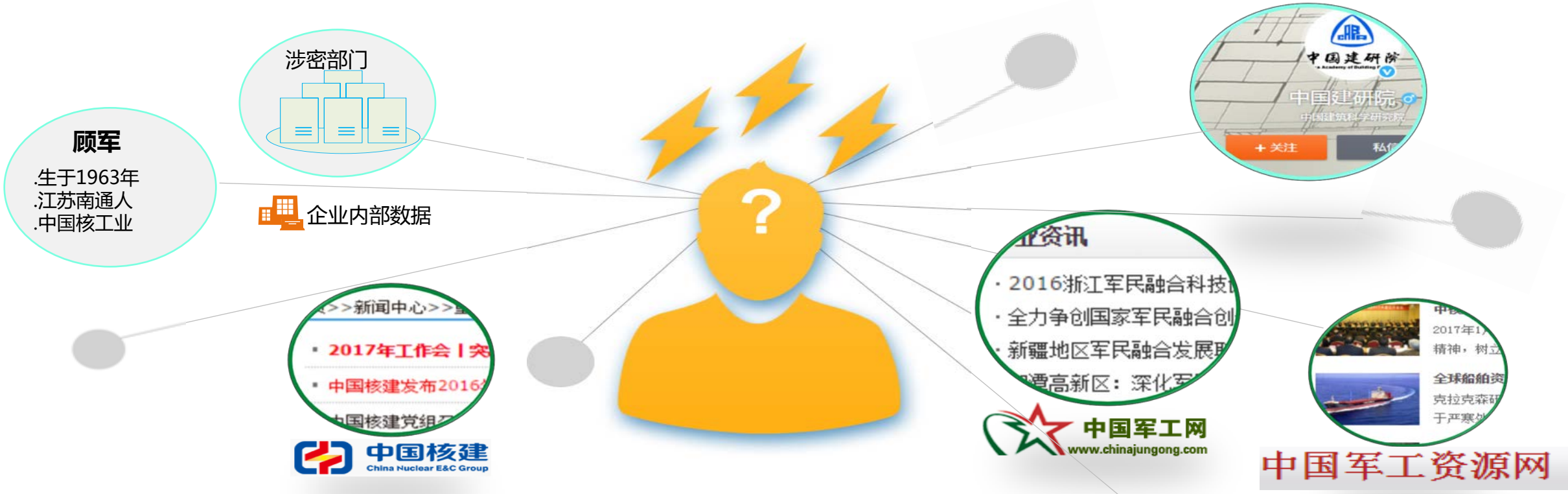
融合使用全量数据会
遇到哪些挑战呢？

	数据库时代	
数据规模	小 MB/GB	大 TB/PB/ZB
数据类型	少 结构化数据为主	多 包含结构化、半结构化、非结构化数据，且后两者越来越多
数据模式	可预先确定 先有数据模式后产生数据； 数据模式相对固定；	无法预先确定 模式在数据出现之后才能确定； 数据模式随数据增长不断演变
处理方法	One Size Fits All	No Size Fits All

来源：开源技术社区 Nock(洛神)

企业全量数据应用挑战1： 多源异构数据难以融合

PlantData

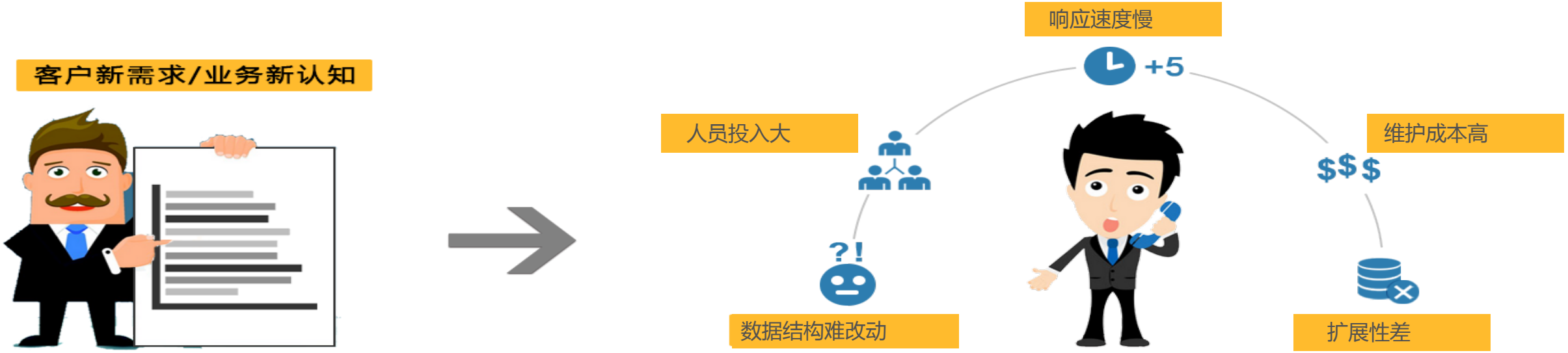


公司内部数据、新闻网站、论坛帖子、微博 …
多源异构数据难以融合

信息聚合、数据融合需求迫切！

企业全量数据应用挑战2：数据模式动态变迁困难

PlantData



当前数据模式动态变迁困难,当客户新需求、业务新认知时程序员需痛苦的修改数据结构及业务逻辑，带来扩展性差、对客户响应慢、维护成本高等不良情况。

我们需要：可自由扩展的数据模式！

企业全量数据应用挑战3：非结构化数据计算机难以理解

PlantData

计算机无法理解非结构化数据的语义

企业迫切需要将非结构化数据结构化

企业大数据应用挑战4：数据使用专业程度过高

PlantData

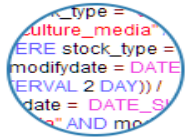
需求举例：了解最近连续两天涨幅大于9%的文化传媒股票

程序员

受用人群有限制门槛高

复杂的检索代码

专业要求高，非技术员难以运用



传统关系型数据库

不易扩展

传统数据模式

知识图谱模式

普通用户

受众范围广

自然语言语义搜索

使用方便，无技能要求

知识图谱

链接数据,强大可扩展性

得到结果：万达院线

行业智能问答大幅降低数据使用门槛

企业大数据应用挑战5： 分散的数据难以统一消费利用

PlantData

✗业务系统繁多

✗使用方式各异

✗难以全局把握

✔统一使用入口

✔可视化

✔匹配度高

基于知识图谱数据存储、融合、分析统一平台，为用户提供统一的消费入口，以不同的形态（检索、可视化、分析等）展示给用户。

解决方案：基于行业知识图谱进行数据融合使用

PlantData

- 挑战1：使用知识图谱（本体）对各种类型的数据进行抽象建模，基于可动态变化的“概念—实体—属性—关系”数据模型，实现各类数据的统一建模。
- 挑战2：使用可支持数据模式动态变化的知识图谱的数据存储，实现对大数据及数据模式动态变化的支持。
- 挑战3：利用信息抽取技术，对非结构化数据及半结构化数据进行抽取和转换，形成知识图谱形式的知识。
- 挑战4、5：在知识融合的基础上，基于语义检索、智能问答、图计算、推理、可视化等技术，提供统一的数据检索、分析和利用平台。

知识图谱助力企业商业智能

PlantData

业务需求

技术方案

数据挑战

语义理解

数据关联探索

业务动态扩展

智能检索与问答

数据结构化

数据融合

自由扩展数据模式

行业智能问答

非结构化数据计算机难以理解

多源异构数据难以融合

数据模式动态变迁困难

数据使用专业程度过高

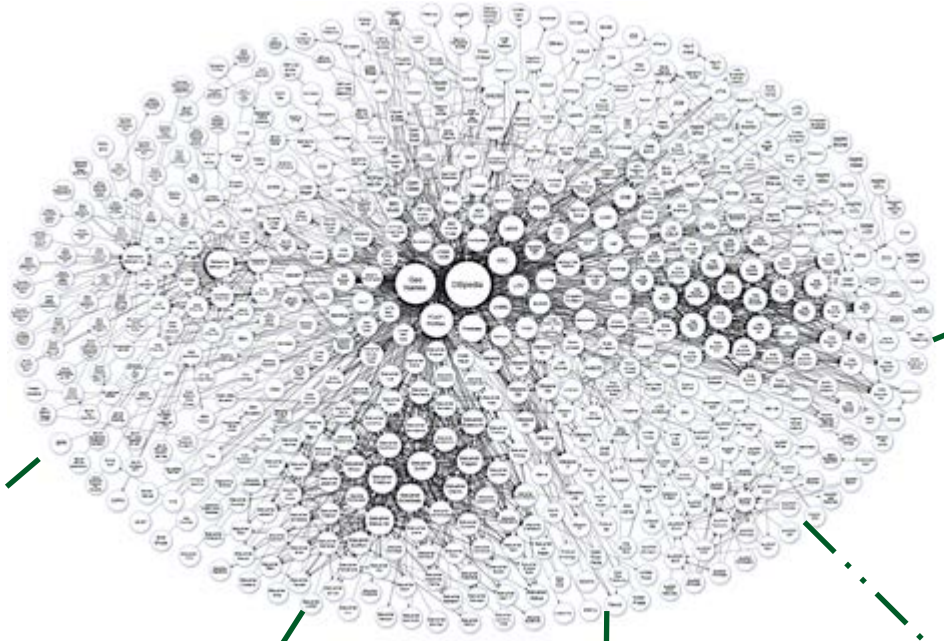
PlantData

行业知识图谱概述

- 行业知识图谱简介
- 行业知识图谱应用
- KG应用挑战
- 行业知识图谱生命周期

行业知识图谱生命周期

PlantData



知识应用

知识计算

知识存储

知识融合

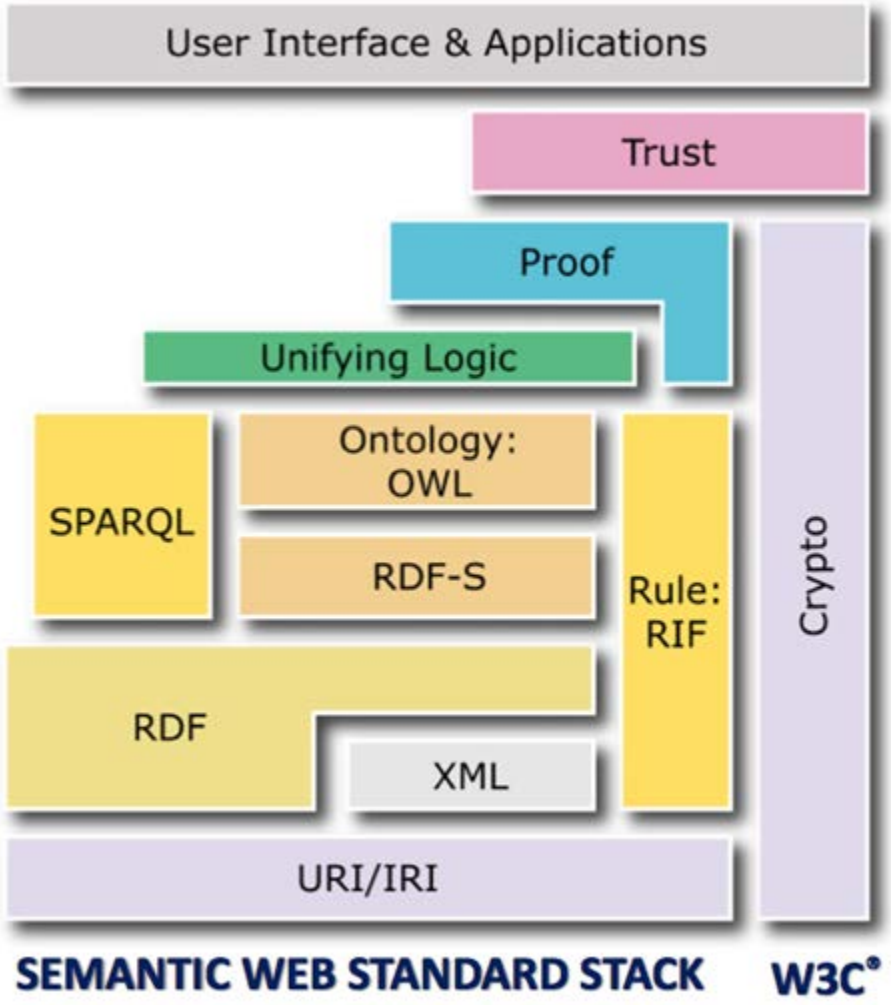
知识获取

知识建模

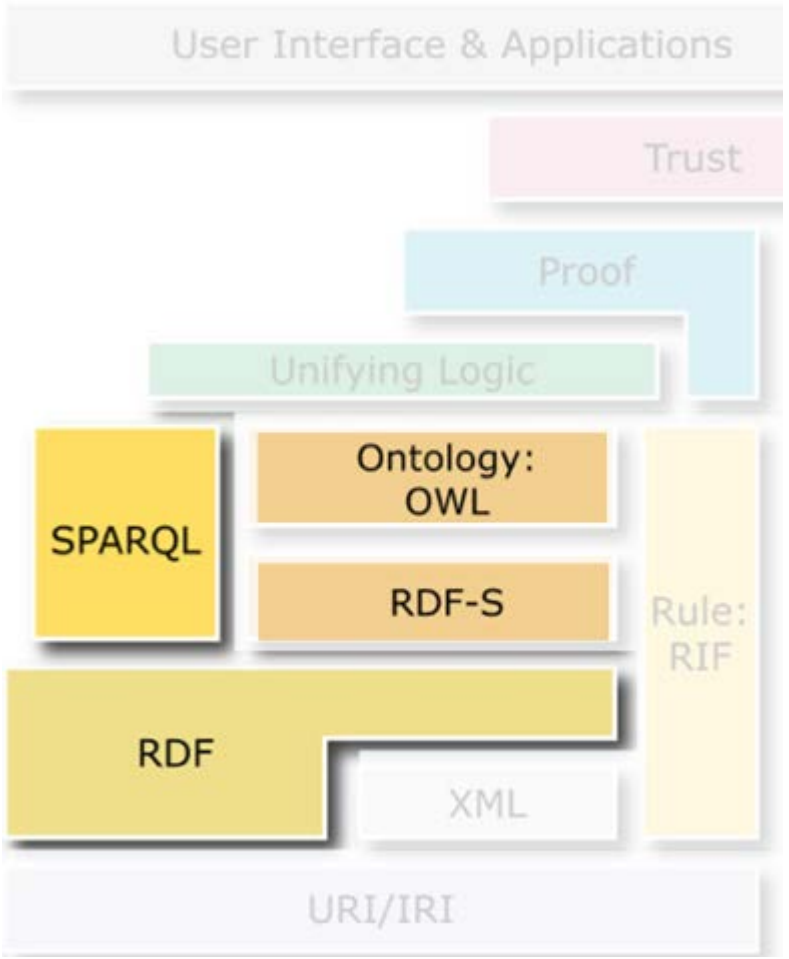
知识图谱基础技术规范

PlantData

W3C推荐的语义网标准栈



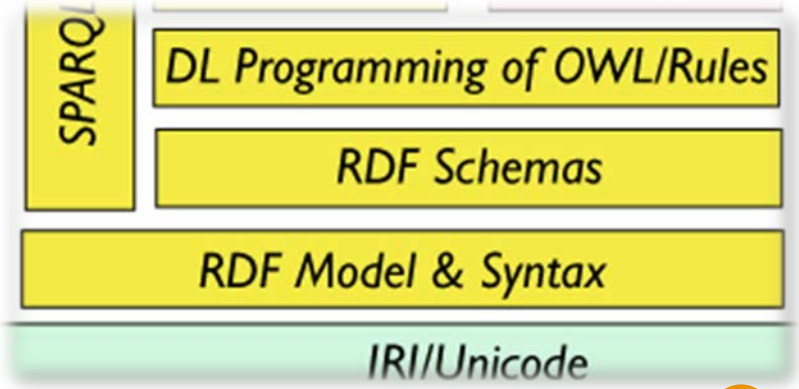
行业知识图谱主要技术标准



RDF

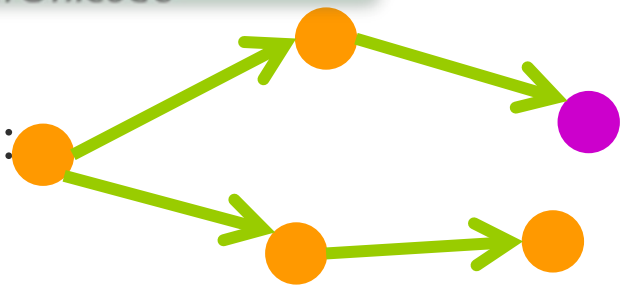
PlantData

- RDF是语义网标准中的第一层
- RDF 代表
  - Resource: 页面、图片、视频等任何具有URI标识符
  - Description: 属性、特征和资源之间的关系
  - Framework: 模型、语言和这些描述的语法



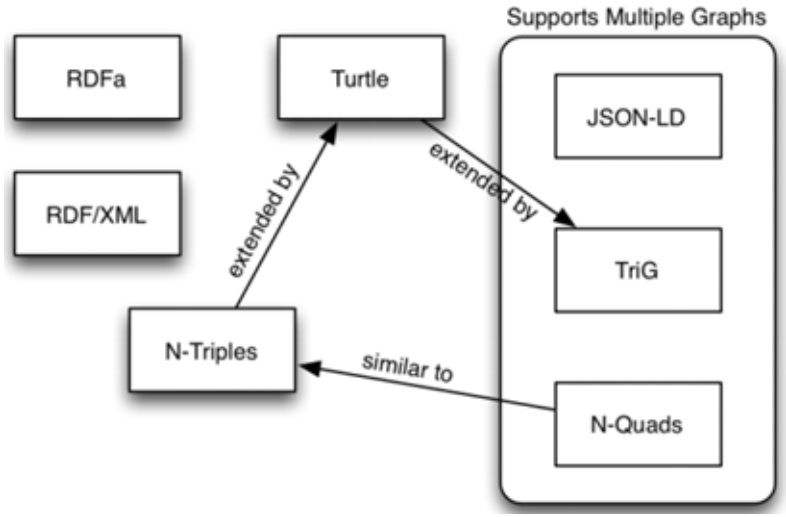
- RDF是一个三元组 (triple) 模型，即每一份知识可以被分解为如下形式：

（subject（主），predicate（谓），object（宾））



（顶点，边，顶点）

- RDF 是一个链接资源描述的图模型，其三元组可看作图中的弧。
- RDF其它语法：Turtle、TriG
  N-Triples、N-Quads、
  JSON、RDFa



OWL：RDF Schema 的扩展

PlantData

- 复杂类：交、并、补
- 属性约束：存在量化、全称量化
- 基数约束：最大基数约束、最小基数约束
- 属性特征：反、对称、非对称、不相交、自反
- 属性链

复杂类

```
:Mother  owl:equivalentClass  [
  rdf:type                owl:Class ;
  owl:intersectionOf  ( :Woman :Parent )
] .
```

复杂类

```
:Parent  owl:equivalentClass  [
  rdf:type                owl:Restriction ;
  owl:onProperty          :hasChild ;
  owl:someValuesFrom      :Person
] .
```

对称属性

```
:hasSpouse  rdf:type  owl:SymmetricProperty .
```

传递类

```
:hasAncestor  rdf:type  owl:TransitiveProperty .
```

属性链

```
:hasGrandparent  owl:propertyChainAxiom  ( :hasParent  :hasParent ) .
```

SPARQL简介

PlantData

- RDF的查询语言：基于RDF数据模型
- 可以对不同的数据集撰写复杂的连接 (joins)
- 由所有主流图数据库支持
- SPARQL Protocol and RDF Query Language

A SPARQL query comprises, in order:

- *Prefix declarations*, for abbreviating URIs
- *Dataset definition*, stating what RDF graph(s) are being queried
- A *result clause*, identifying what information to return from the query
- The *query pattern*, specifying what to query for in the underlying dataset
- *Query modifiers*, slicing, ordering, and otherwise rearranging query results

```
# prefix declarations
PREFIX foo: <http://example.com/resources/>
...
# dataset definition
FROM ...
# result
                    clause
SELECT ...
# query pattern
WHERE {
    ...
}
# query modifiers
ORDER BY ...
```

A quick introduction to SPARQL, the de-facto query language for RDF. Those familiar with SQL will have an easy time understanding SPARQL… at the beginning…

A Simple select query

```
PREFIX dc: <http://purl.org/dc/elements/1.1/>
SELECT ?title
WHERE { <http://example.org/book/book1> dc:title ?title }
```

A Join

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT ?name ?mbox
WHERE { ?x foaf:name ?name . ?x foaf:mbox ?mbox . }
```

An Outer Join

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT ?name ?mbox
WHERE { ?x foaf:name ?name . OPTIONAL { ?x foaf:mbox ?mbox } }
```

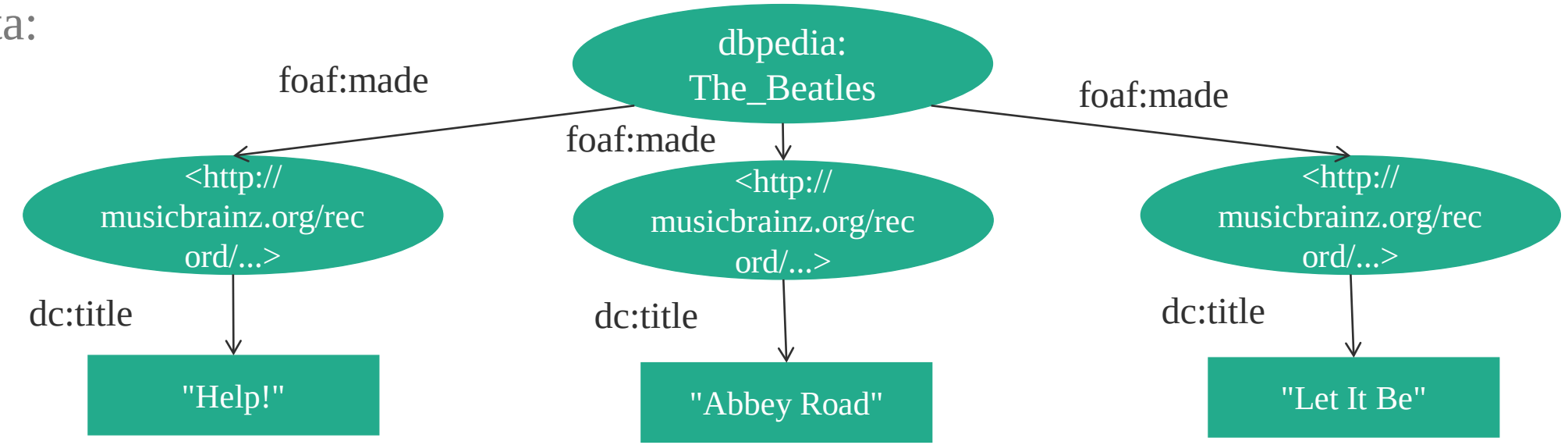
Like “Like”

```
PREFIX dc: <http://purl.org/dc/elements/1.1/>
SELECT ?title
WHERE { ?x dc:title ?title . FILTER regex(?title, "^SPARQL") }
```

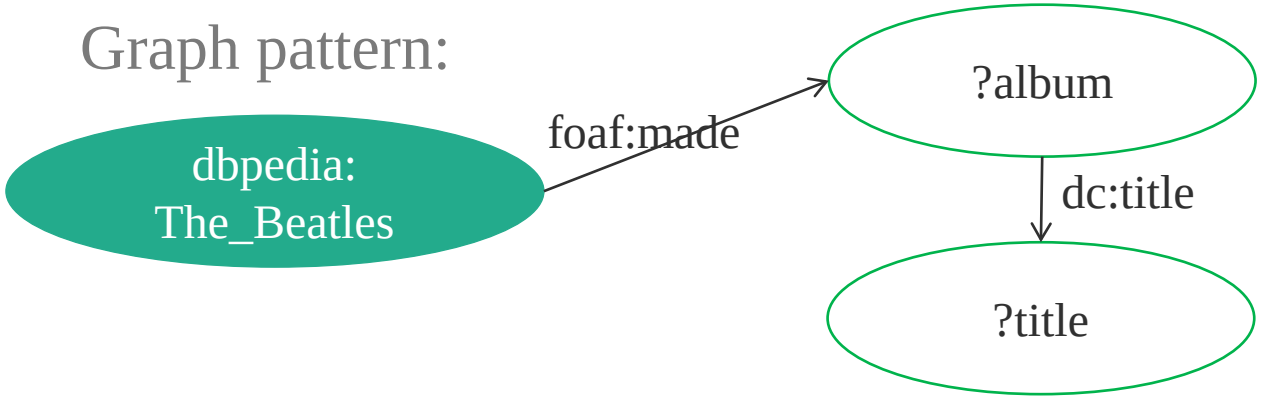
SPARQL Query – graph visualization

PlantData

Data:



Graph pattern:



Results:

?album	?title
<http://...>	"Help!"
<http://...>	"Abbey Road"
<http://...>	"Let It Be"

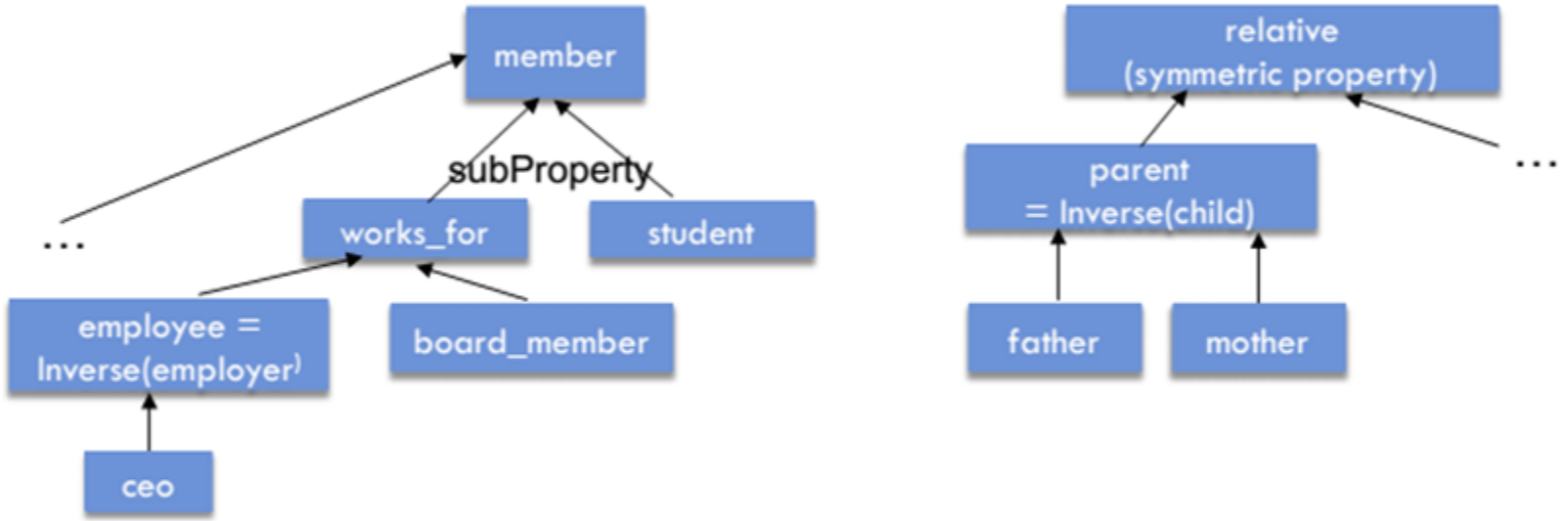
本体 (ontology)可以填充知识与查询之间的语义间隙

PlantData

Let's revisit our simple conceptual query

"Find people who are members of an organization founded by a relative."

Assume some mechanism to specify relationships between predicates as follows:



Simpler precise query against the RDF data enriched with the above semantic knowledge (ontology)

```
SELECT* WHERE {
  ?x member ?y .
  ?z founder ?y .
  ?z relative ?x
}
```

1. 知识建模

PlantData

- 即建立知识图谱的数据模式。行业知识图谱的数据模式对整个知识图谱的结构进行定义，因此需要保证可靠性。

- 常用方法
  - 自顶向下的方法：专家手工编辑形成数据模式
  - 自底向上的方法：
    - 基于行业现有的标准进行转换
    - 从现有的高质量行业数据源（如业务系统数据库表）中进行映射

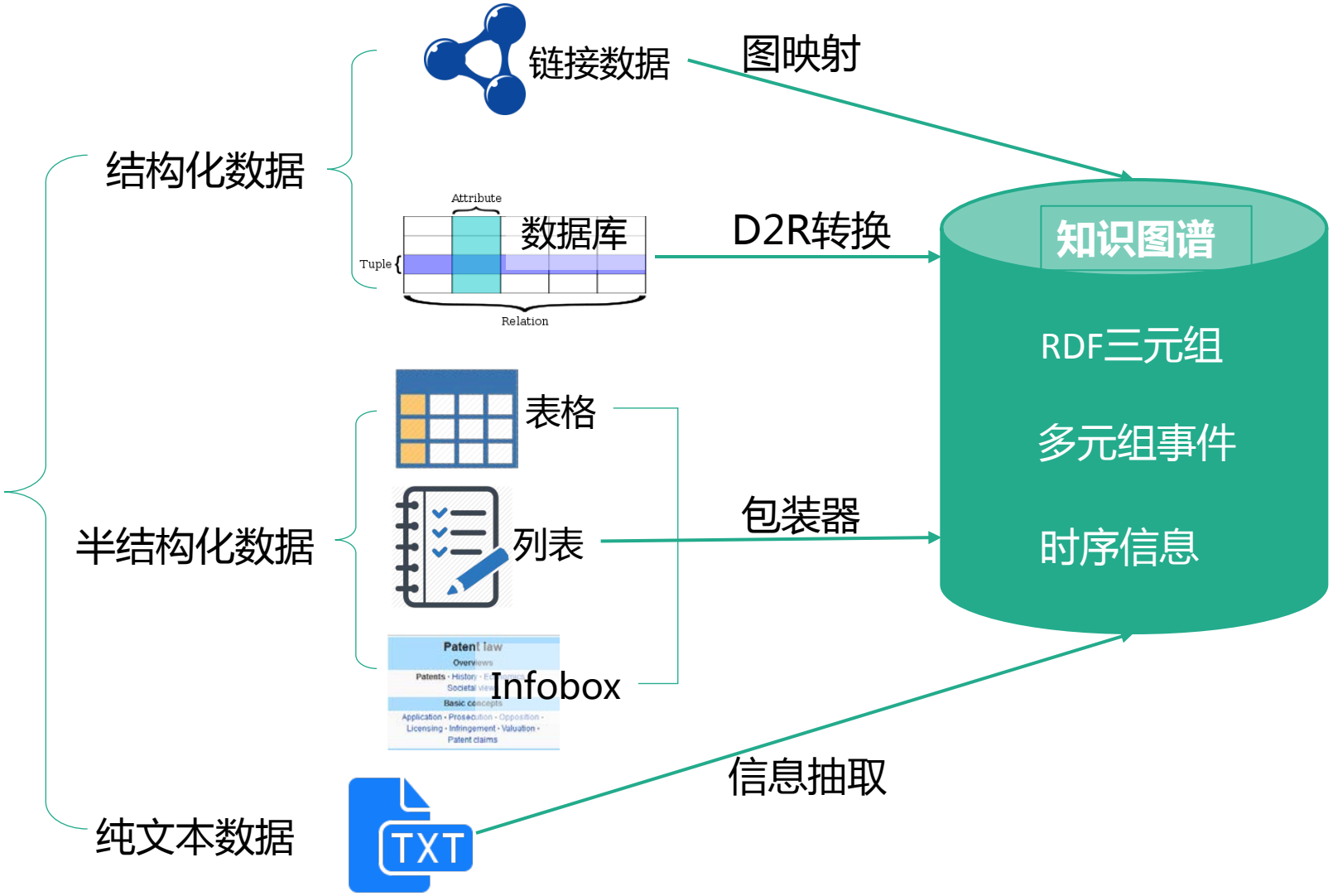
知识建模关键技术与难点

PlantData

- 多人在线协同编辑，并且实时更新
- 能够导入集成使用现有的（结构化）知识
- 支持大数据量
- 能够支撑事件、时序等复杂知识表达
- 可以与自动算法进行结合，避免全人工操作

2. 知识获取

从不同来源、不同结构的数据中进行知识提取，形成知识存入到知识图谱。



知识获取关键技术与难点

PlantData

- 从结构化数据库中获取知识：D2R
  - 难点：复杂表数据的处理
- 从链接数据中获取知识：图映射
  - 难点：数据对齐
- 从半结构化（网站）数据中获取知识：使用包装器
  - 难点：方便的包装器定义方法，包装器自动生成、更新与维护
- 从文本中获取知识：信息抽取
  - 难点：结果的准确率与覆盖率

3. 知识融合

PlantData

- 数据模式层融合
  - 概念合并
  - 概念上下位关系合并
  - 概念的属性定义合并
- 数据层融合
  - 实体合并
  - 实体属性融合
  - 冲突检测与解决

- 行业知识图谱的数据模式通常采用自顶向下和自底向上结合的方式，因此基本都经过人工的校验，保证了可靠性；因此，知识融合的关键任务在数据层的融合。

知识融合：跨语言融合

PlantData

Unified Medical Language System ®

- Abdominal distension
- Abdominal mass
- Abdominal pain
- Abnormal dermatoglyphics
- Abnormal heart sounds
- Abnormal reflexes
- Abnormal splitting of heart sounds
- Abnormal sputum
- Absence of uvula
- Acneiform lesions
- Acute confusional state
- Adrenal cortex hyperplasia
- Akathisia
- Alopecia
- Amaurosis fugax

咳嗽 症状 [症状起因] [诊断详述] [检查鉴别] [问医生] [找医院]

咳嗽是呼吸系统疾病最常见的症状之一，它是一种保护性神经反射，通过咳嗽产生呼气性冲击动作，能将呼吸道内的异物或分泌排出体外。[详细]

相关疾病 夏季感冒 | 病毒性支气管炎 | 病毒性喉炎 | 肾咳 | 痰饮

阴茎短 症状 [症状起因] [诊断详述] [检查鉴别] [问医生] [找医院]

小阴茎(micropenis)是指阴茎外观正常长度与直径比值正常，但阴茎体的长度小于正常阴茎长度平均值2.5个标准差以上。阴茎的长度是指用手提阴茎头尽量拉直即相当于阴茎充分勃起时从阴茎顶到耻骨联合的距...[详细]

相关疾病 男性生殖器畸形 | 小阴茎 | 类固醇5α-还原酶2缺乏综合征 | 先天性睾丸发育不全

阴道出血 症状 [症状起因] [诊断详述] [检查鉴别] [问医生] [找医院]

阴道出血是女性生殖器官疾病常见的症状。出血可来自外阴、阴道、子宫颈和子宫内膜，但以来自子宫者为最多。阴道出血量固然可以危及生命，但如良性疾病所致者，预后良好；而出血量少的，也可能是恶性肿...[详细]

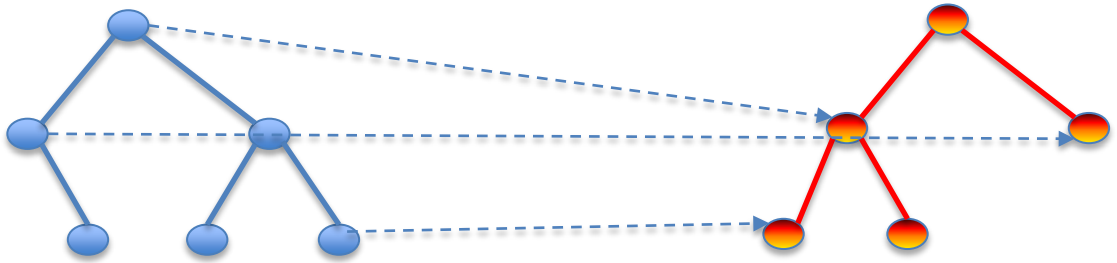
相关疾病 重度宫颈糜烂 | 胎热 | 血热崩漏 | 妊娠热病 | 原发性纤维蛋白溶解症

关节疼痛 症状 （别名：关节痛） [症状起因] [诊断详述] [检查鉴别] [问医生] [找医院]

关节炎或关节病牵涉范围非常广泛，种类繁多，病因各异，然而，普遍的临床症状均可表现关节疼痛，因此关节疼痛的鉴别诊断至关重要，但是有的病因已知，有的病因尚未明了，总之，明确诊断才是合理治疗疾...[详细]

中文体系结构

英文体系结构



R30 Pain associated with micturition
Excludes: psychogenic pain (F45.3)
小便疼痛
不包含：心因性小便疼痛 (F45.3)

R30.0 Dysuria
Strangury
小便困难
尿铰痛

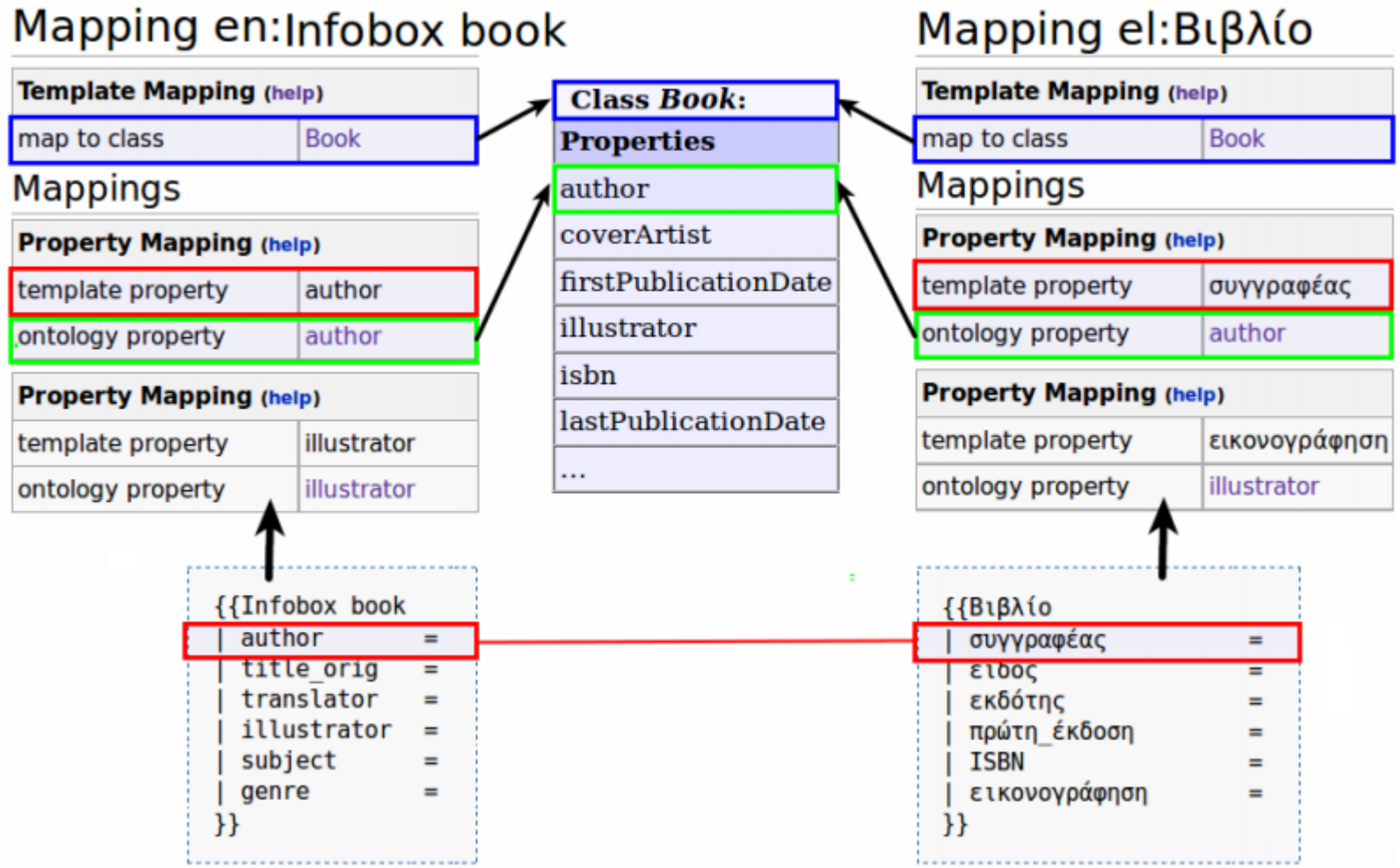
R30.1 Vesical tenesmus
膀胱里急後重

R30.9 Painful micturition, unspecified

ICD编码

Dbpedia Mapping

PlantData



http://mappings.dbpedia.org/

识别sameAs关系

PlantData

ZHISHI.me

owl:sameAs

- zhwiki:白癜风 (this)
- baidu:白癜风 (this)
- hudong:白癜风 (this)
- dbpedia:Vitiligo

MERGE PAGE

标 签

zhwiki:白癜风
hudong:白癜风
baidu:白癜风

Index

- zhishi:abstract
- infobox
- dcterms:subject
- zhishi:relatedImage

摘 要

zhishi:abstract

白癜风（vitiligo）一词来源于拉丁语vitium,意"缺损"，或vi-tellus,指"小牛白色斑片"。公元一世纪，罗马医生Celsus所写DeMedicina中，可见vitiligo一词。Vitiligo在许多古籍中常常出现。印度文献中，kilas(kil指白色，as意抛弃)和palita(pal含灰色、年老、老年之意)可追溯到公元前1500~1000年，指的是皮肤上的白色斑片。佛教圣书Vinay Pitak(624~544B.C.)中记载，患kilas的人不能做牧师。印度Manusmriti(200 B.C.)中记载，患svitra(白斑)不受尊重。《古兰经》中baras一词意指白色皮肤，常用来描述耶稣治愈后的状况。世界各地均有发生，印度发病率最高，中国约有1200万人发病，本病可以累及所有种族，男女发病无显著差别。近年白癜风发病率逐年上升，引起人们的普遍关注。 它是一种常见多发的色素性皮肤病。该病以局部或泛发性色素脱失形成白斑为特征，是一种获得性局限性或泛发性皮肤色素脱失症，是一影响美容的常见皮肤病，易诊断，治疗难。

信 息 框

dcterms:subject

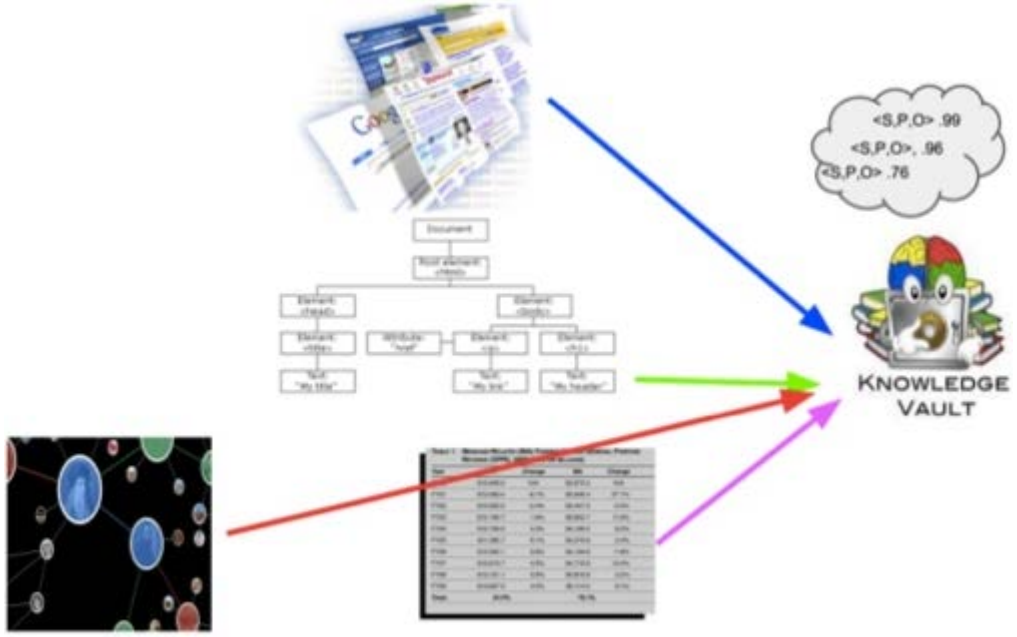
主 题

皮肤和皮下组织疾病 应用科学 白斑 科学 疾病大全 疾病 医学名词 皮肤病 疾病 常见疾病 白癜风预防

Knowledge Vault：Google 将建全球最大知识库

PlantData

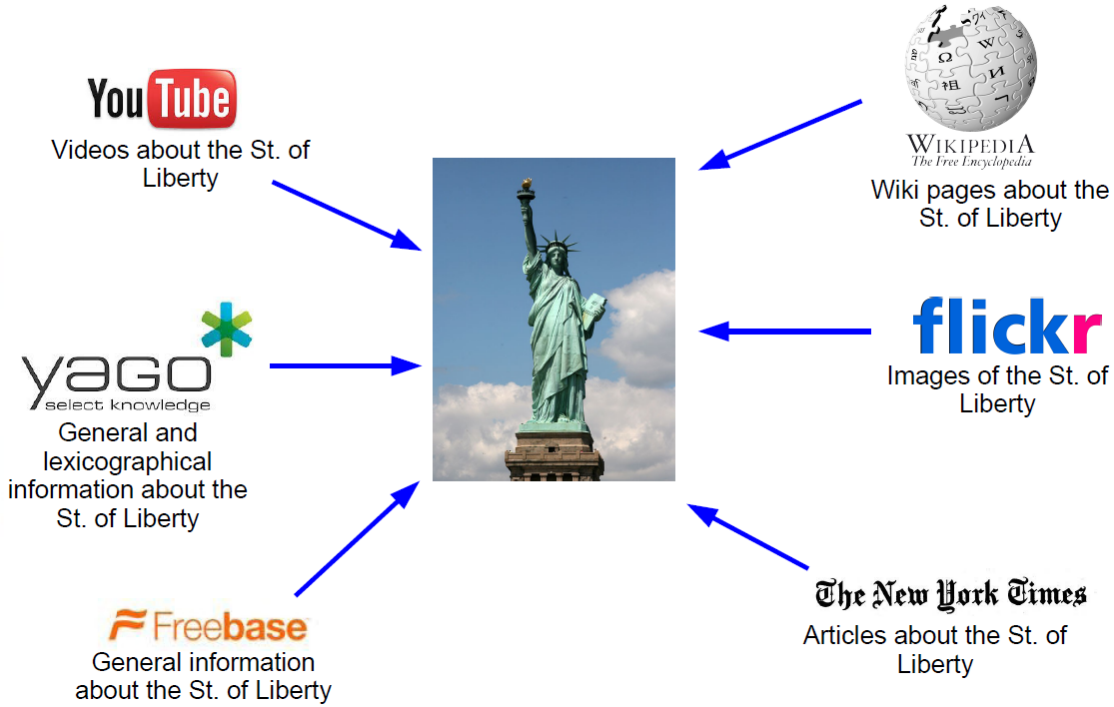
- Knowledge Vault是一种以互联网信息为基础的数据库，即“知识库”。
- 知识来源：Gmail、Google+、Youtube
- 通过特定算法自动搜集整编互联网信息，再将其存入数据库中。
- Knowledge Vault的入库信息已达16亿条（至2014年），其中2.7亿条内容为“事实”（真实性在90%以上）。
- Knowledge Vault 能够建立历史和社会的模型。



知识融合关键技术与难点

PlantData

- 实现不同来源、不同形态数据的融合
- 海量数据的高效融合
- 新增知识的实时整合
- 多语言的融合



4. 知识存储

PlantData

- 知识图谱数据存储需要完成的基本数据存储：
  - 三元组知识的存储
  - 事件信息的存储
  - 时态信息的存储
  - 使用知识图谱组织的数据的存储

- 知识图谱上层应用需要支持：
  - 知识推理
  - 知识快速查询
  - 图实时计算

知识存储关键技术与难点

PlantData

- 大规模三元组数据的存储
- 知识图谱组织的大数据的存储
- 事件与时态信息的存储
- 快速推理与图计算的支持

5. 知识计算

PlantData

- 图挖掘计算：基于图论的相关算法，实现对图谱的探索和挖掘。
- 本体推理：使用本体推理进行新知识发现或冲突检测。
- 基于规则的推理：使用规则引擎，编写相应的业务规则，通过推理辅助业务决策。

关键技术与难点

PlantData

- 图挖掘计算
  - 大规模图算法的效率
- 本体推理与规则推理
  - 大数据量下的快速推理
  - 对于增量知识和规则的快速加载

6. 知识应用

PlantData

- 语义搜索：基于知识图谱中的知识，解决传统搜索中遇到的关键字语义多样性及语义消歧的难题；通过实体链接实现知识与文档的混合检索。

- 智能问答：针对用户输入的自然语言进行理解，从知识图谱中或目标数据中给出用户问题的答案。

- 可视化决策支持：通过提供统一的图形接口，结合可视化、推理、检索等，为用户提供信息获取的入口。

关键技术与难点

PlantData

- 语义检索
  - 自然语言的表达多样性问题
  - 自然语言的的歧义问题
- 智能问答
  - 准确的语义解析
  - 正确理解用户的真实意图
  - 答案确定与排序
- 可视化决策支持
  - 通过可视化方式辅助用户模式快速发现
  - 高效地缩放和导航
  - 大图环境下底层算法（图挖掘算法）的效率

行业知识图谱关键技术

PlantData

知识融合

可视化

图挖掘

语义搜索

知识建模

行业知识图谱
更加规范的数据表示
更强的数据关联
更深邃的数据价值

知识推理

知识抽取

知识存储

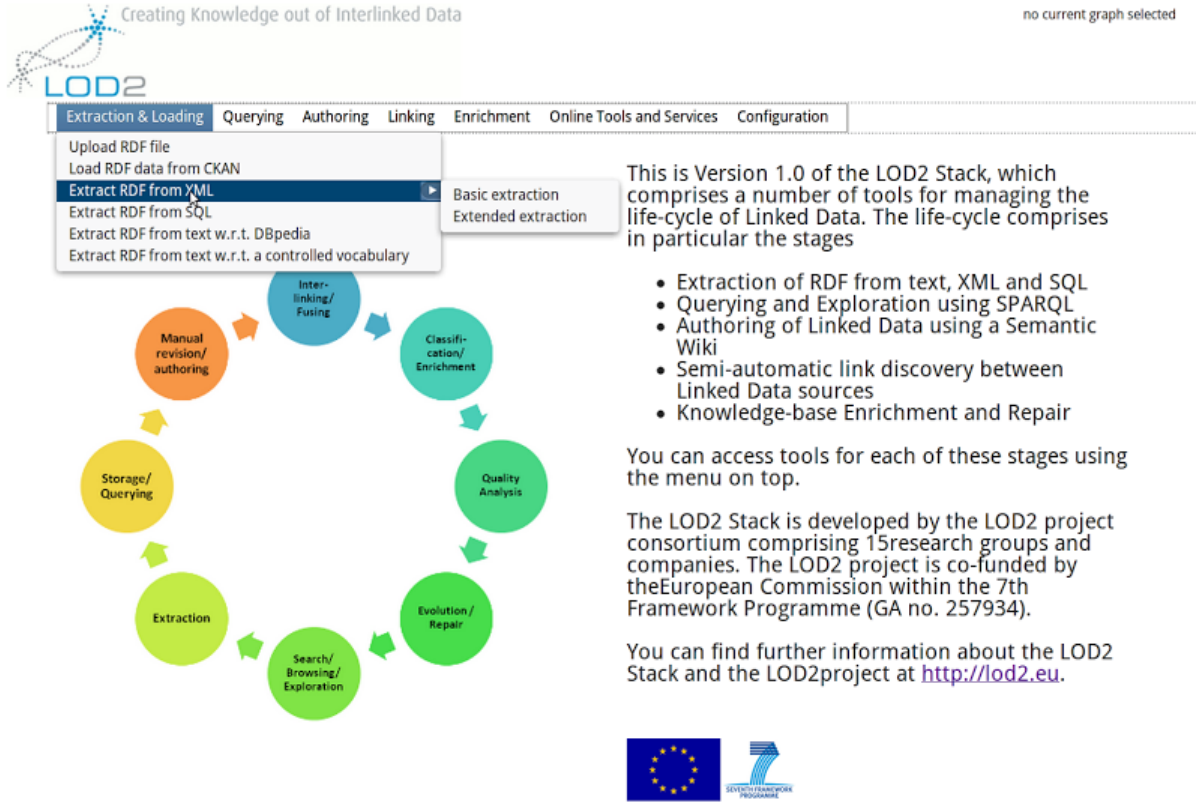
实体链接

LOD2

PlantData

- LOD2 项目的主要目标是构建结构化链接数据的企业级管理工具和方法学，提供一个搜索、浏览和生成链接数据的平台。

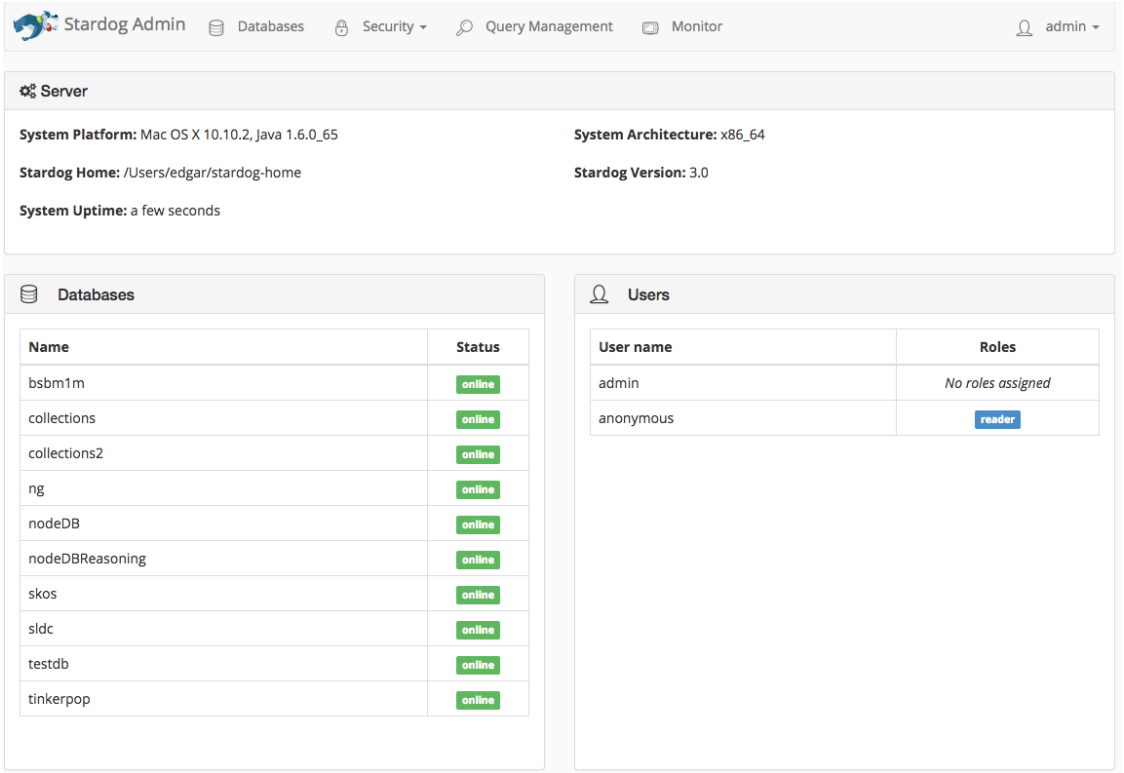
- LOD2 侧重于链接数据的生命周期管理，其它类型的数据需要首先转换成链接数据。
- LOD2 没有对中文处理的支持。



Stardog

PlantData

- Stardog 是一个企业级知识图谱平台，通过把数据转换成知识，使用知识图谱进行组织，对外提供查询、检索、分析服务。主要特点：
  - 把关系数据库映射成虚拟图
  - 支持OWL2的推理
  - 支持Gremlin
- 但 Stardog 仅包含对结构化数据（RDBMS、Excel等）的处理，没有针对非结构化数据的知识抽取，也没有包含知识融合功能。



知识图谱生命周期总结

PlantData

- 在行业应用中使用知识图谱，大致有如下几种方式：
  - 使用现有的套装工具（如 LOD2、Stardog）
  - 在现有套装工具的基础上进行扩充：
  - 使用各生命周期过程的相应工具并进行组合使用
  - 针对性开发或扩展生命周期中特定工具
  - 完全从零开始构建

PlantData

行业知识图谱关键技术

知识建模

- 知识获取
- 知识融合
- 知识存储
- 知识计算
- 知识应用

使用知识图谱对数据进行抽象建模

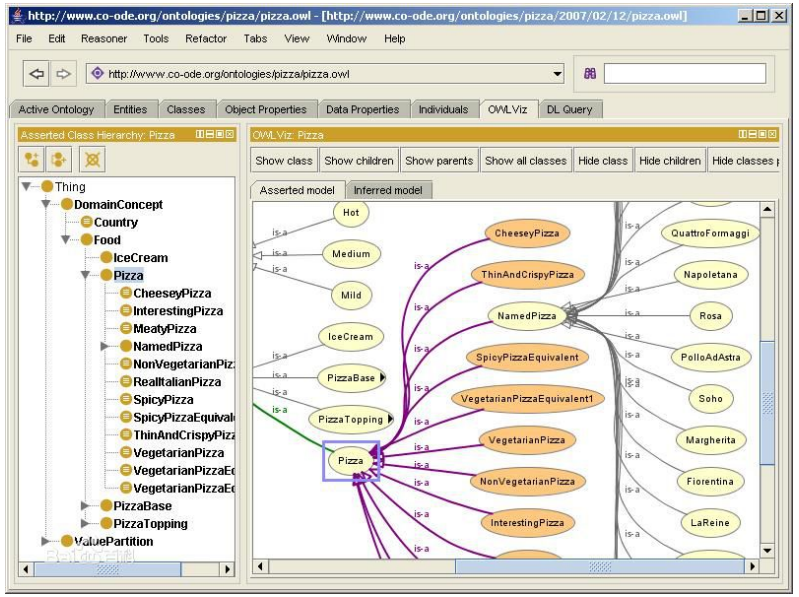
PlantData

- 以实体为主体目标，实现对不同来源的数据进行映射与合并。（实体抽取与合并）
- 利用属性来表示不同数据源中针对实体的描述，形成对实体的全方位描述。（属性映射与归并）
- 利用关系来描述各类抽象建模成实体的数据之间的关联关系，从而支持关联分析。（关系抽取）
- 通过实体链接技术，实现围绕实体的多种类型数据的关联存储。（实体链接）
- 使用事件机制描述客观世界中动态发展，体现事件与实体间的关联；并利用时序描述事件的发展状况。（动态事件描述）

知识建模工具——Protégé

PlantData

- 本体编辑器
- 基于RDF(S)，OWL等语义网规范
- 图形化界面
- 提供了在线版本——WebProtégé
- 适用于原型构建场景



Protégé 的不足：

- 基本只提供单人编辑，在线版本的并发功能支持也不完善；并发编辑时需要通过文件共享来实现；
- 因为基于单机构建，因此对大数据量支持不够，会出现内存溢出；
- 不支持复杂事件及时态的建模；
- 完全依靠人工，难以实现与知识图谱构建（半）自动化过程的交互。

构建一个适用的建模工具（1）

PlantData

- 在线并发编辑支持；
- 编辑的知识实时保存，当其它用户对当前用户正在编辑的内容有更新时，系统自动提示加载最新版本，因此能够有效地解决并发知识编辑冲突。

在线编辑

上下位关系定义

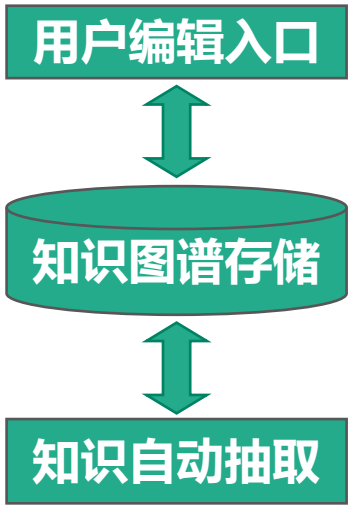
属性定义编辑

构建一个适用的建模工具（2）

PlantData

- 提供导入、集成功能，能够把现有的知识通过导入功能进行集成；
- 以存储为桥梁，可以对自动算法的结果进行编辑。

知识导入



构建一个适用的建模工具（3）

PlantData

- 支持对动态事件数据的建模，使用时态信息存储实现事件时间描述。

属性名	属性类型	定义域	操作选项
轮次	数值	融资事件	详情
融资额	数值	融资事件	详情
融资时间	数值	融资事件	详情
投资方	对象	融资事件	详情 边属性编辑
融资方	对象	融资事件	详情 边属性编辑
融资事件标签	对象	融资事件	详情 边属性编辑

融资事件的建模

构建一个适用的建模工具（4）

PlantData

- 支持大数据量的知识图谱编辑：编辑是基于底层的知识图谱存储的，每次编辑时加载到前端的仅为当前相关的数据，因此不会造成内存溢出等问题。

PlantData

行业知识图谱关键技术

- 知识建模
- **知识获取**
- 知识融合
- 知识存储
- 知识计算
- 知识应用

知识获取

PlantData

D2R

包装器

文本信息抽取

D2R工具——D2RQ

PlantData

- **D2RQ：将关系数据库转换为虚拟的RDF数据库的平台，主要包括：**
  - D2R Server：HTTP Server，提供对 RDF数据的查询访问接口，以供上层的 RDF 浏览器、SPARQL 查询客户端以及传统的 HTML 浏览器调用；
  - D2RQ Engine：利用一个可定制的 D2RQ Mapping 文件将关系型数据库中的数据换成 RDF 格式；
  - D2RQ Mapping Language：定义将关系型数据转换成 RDF 格式的 Mapping 规则。

- **存在的问题：**
  - 直接转换成RDF，难以与知识建模结果对应，也难以同其它知识进行融合
  - 新数据的增量映射
  - 海量数据映射

D2R工具构建（1）

PlantData

- D2R映射与知识建模结合，在数据模式的基础上进行映射；例如从数据库中的“企业信息表”中把记录映射成概念“企业”下的实体。

- 通过设置合并条件，把D2R的结果与知识图中的已有知识进行融合；例如对于企业，设置“如果企业名称相同则进行合并”的规则。

D2R工具构建（2）

PlantData

- 通过特定的关键词及规则来设置数据更新的标记，从而实现数据的增量映射；例如，对于企业，设置“若成立时间为上次更新时间之后的企业为新的企业”。

- 经过D2R映射的数据直接存储成为知识图谱中的知识，因而其数据量仅取决于存储的支撑量。

半结构化行业数据源解析

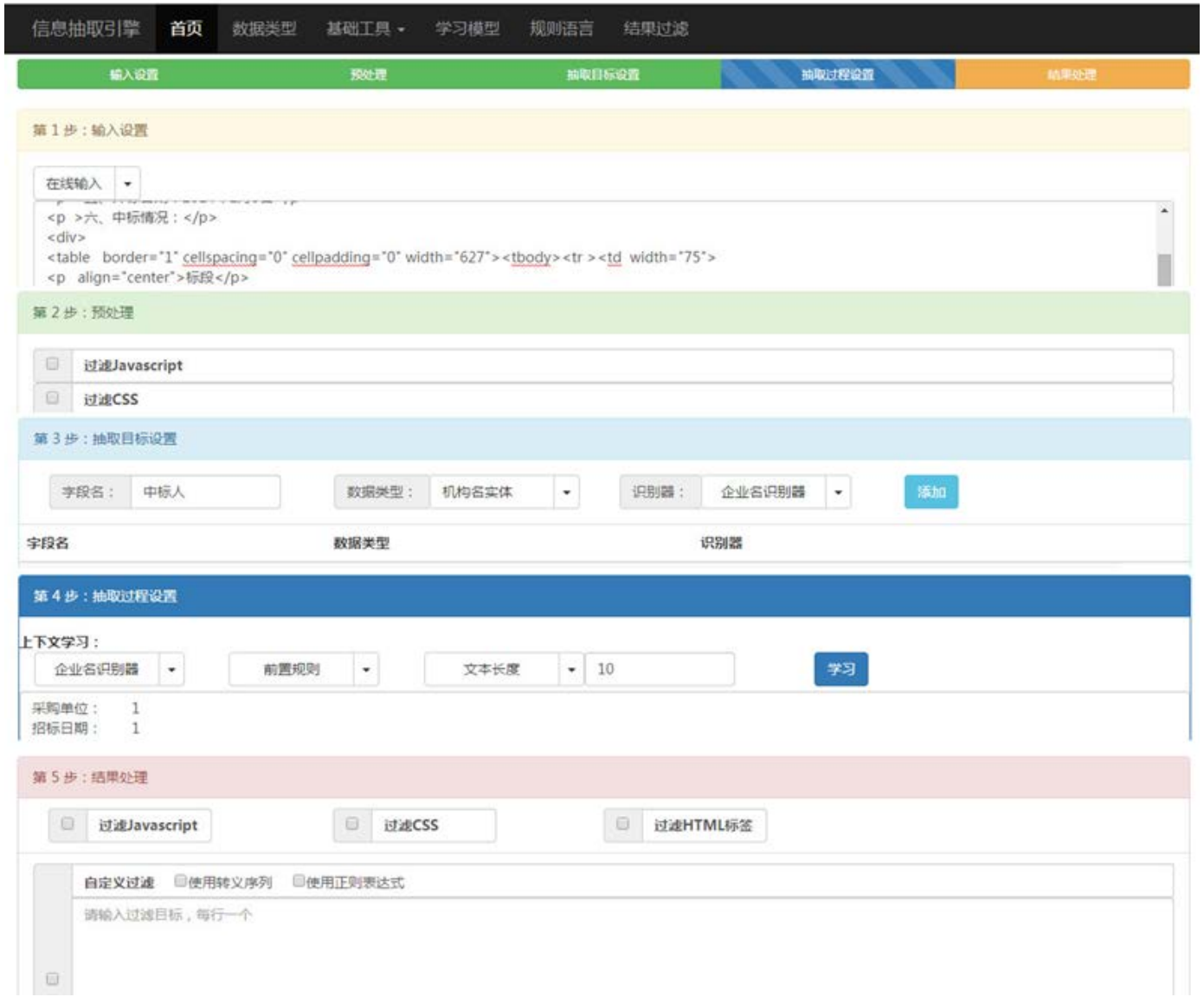
PlantData

- 行业网站大都是通过模板来生成的，因此通常使用包装器来进行解析；
- 包装器可以自动进行学习，但为了保证准确度，通常使用人机结合的方法；
- 行业数据源解析，由于网站的高度可变性，因此尚没有统一的工具。
- 在实际应用中，通常针对不同结构的数据配置相应的包装器，完成数据的解析。

包装器配置工具

PlantData

1. 输入源设置
2. 预处理配置
3. 抽取目标配置
4. 抽取过程配置：为抽取的目标设置抽取规则
5. 结果后处理



包装器示例——专利知识抽取（1）

PlantData

- **包装器示例：**从半结构化数据中抽取专利知识

专利文本数据

谓词 宾语

主语

<REC>↓
<申请号>=CN01820110.5↓
<专利号>=CN01820110.5↓
<申请日>=2001.11.15↓
<公开（公告）日>=2015.05.06↓
<公开（公告）号>=CN1479810B↓
<名称>=生产金属间化合物的方法↓
<主分类号>=C25C5/04(2006.01)I↓
<分类号>=C25C5/04(2006.01)I;C25C7/00(2006.01)I↓
<地址>=英国剑桥↓
<国省代码>=英国;GB↓
<发明（设计）人>=D·J·弗雷;R·C·科普卡特;G·Z·陈↓
<专利代理机构>=中国国际贸易促进委员会专利商标事务所 11038↓
<代理人>=蔡胜有↓
<国际申请>=2001-11-15 PCT/GB2001/005034↓
<国际公布>=2002-05-23 WO02/40748 EN↓
<进入国家日期>=20030605↓
<优先权>=2000.11.15 GB 0027930.7↓
<申请（专利权）人>=剑桥企业有限公司↓
<摘要>=生产一种金属间化合物(M¹Z)的方法，涉及通过在使非金属物质溶解在熔融的盐中的条件下，与包含一种熔融的盐(M²Y)的熔体接触进行电脱氧，对包含三种或更多物质的固态前体材料进行处理，所述三种或更多物质包括第一和第二金属或准金属物质(M¹，Z)和一种非金属物质(X)。该第一和第二金属或准金属物质形成一种金属间化合物。该方法在一个包括前体材料(2)的阴极的电解槽中进行，所述电解槽被浸入到容纳在坩埚(6)中的熔体(8)中，用以进行电脱氧。↓
<参考文献>=GB 626636 A,1949.07.19,权利要求1-14和说明书第二页95-114行.;CN 1034964 A,1989.08.23,权利要求1-6.;WO 9964638 A,1999.12.16,权利要求1-25和实施例12.↓
<审查员>=朱峰↓
<专利类型>=8↓
<申请来源>=国际↓
<页数>=9↓
<主权项>=一种生产金属间化合物(M¹Z)的方法，包括：对包含三种或更多物质的固态前体材料进行处理，所述三种或更多物质包括第一金属或准金属(M¹)，第二准金属(Z)和一种非金属物质(X)，所述第二准金属为C或B，所述处理通过使该固态前体材料与包含一种熔融的盐(M²Y)的熔体接触，并且向该固态前体材料提供阴极电压以使非金属物质溶解或通过熔体移向阳极，其中，施加在前体材料的阴极电压小于在阴极表面上使阳离子(M²)从熔融的盐中沉积出来所需电压、或者如果熔体包含盐的混合物则小于在阴极表面上使任何一种阳离子(M²)从熔体中沉积出来所需电压。↓
<申请国代码>=CN↓
<发布路径>=BOOKS/SD/2015/20150506/01820110.5↓

<生产金属间化合物的方法，申请号，CN01820110.5>

<生产金属间化合物的方法，专利号， CN01820110.5 >

<生产金属间化合物的方法，申请日，2001.11.15>

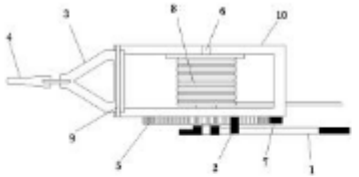
<生产金属间化合物的方法，公开（公告）日，2015.05.06>

…… ……

包装器示例——专利知识抽取（2）

PlantData

专利网站



主语

[发明公布]一种便携式紧线装置

申请公布号：CN105337221A　　申请公布日：2016.02.17

申请号：2015107128886　　申请日：2016.01.05

申请人：林蓉瑶　　发明人：林蓉瑶

地址：323699浙江省丽水市云和县云和镇沙溪村沙溪76号

分类号：H02G1/04(2006.01)I

宾语

谓词

摘要： 本发明公开了一种便携式紧线装置，包括挂钩，所述挂钩的右端活动连接有拉杆，所述拉杆的右端通过销与机架活动连接，所述机架的中部活动安装有轴，所述轴的中部固定安装有收线盘，所述机架的外侧设置有棘轮盘、且所述棘轮盘与轴固定连接，所述棘轮盘的下侧活动安装有把 全部

【发明专利申请】　　*事务数据*

<一种便携式紧线装置，申请公布号，CN105337221A>

<一种便携式紧线装置，申请公布日，2016.02.17>

<一种便携式紧线装置，申请号，2015107128886>

<一种便携式紧线装置，申请日，2016.01.05>

<一种便携式紧线装置，申请人，林蓉瑶>

……

文本信息抽取：主要任务

PlantData

实体识别

概念抽取

关系抽取

事件抽取

CloseIE 和 OpenIE

PlantData

CloseIE

- 面向特定领域抽取信息
- 预先定义好抽取的关系类型
- 基于领域专业知识抽取
- 规模小
- 精度比较高

OpenIE

- 面向开放领域抽取信息
- 关系类型事先未知
- 基于语言学模式进行抽取
- 规模大
- 精度相对较低

OpenIE

PlantData

- OpenIE 的典型代表工具有 ReVerb、TextRunner

- OpenIE 工具由于准确率比较低，在行业知识图谱构建中实用性不高，会增加知识融合的难度。（通常用于做第一轮的信息抽取探索，从它的结果中发现新的关系，然后在此基础上应用其它的信息抽取方法。）

CloseIE 典型工具：DeepDive

PlantData

- DeepDive基于联合推理的算法让用户只需要关心特征本身，要求开发者思考特征而不是算法，而其他机器学习系统要求开发者思考聚类算法、分类算法的使用等；

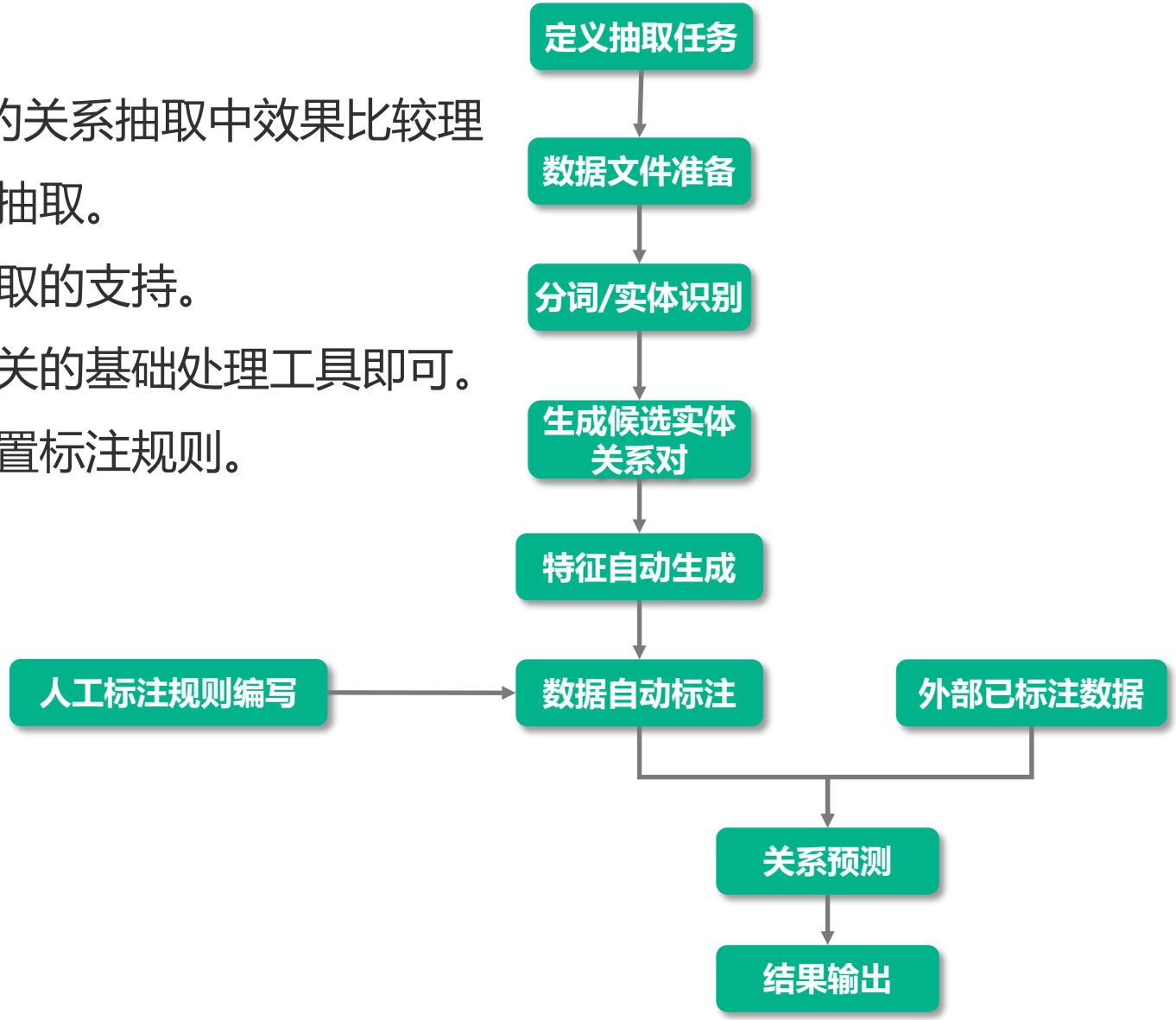
- DeepDive允许用户使用简单的规则来影响学习过程以提升结果的质量，也考虑用户反馈来提高预测的准确度；

- DeepDive使用机器学习算法训练系统来减少各种形式的噪音和不确定性，并为每一个决断进行复杂的可能性计算。

DeepDive 关系抽取过程与总结

PlantData

- DeepDive主要针对**关系抽取**，在指定的关系抽取中效果比较理想，在实体确定后可以很好地进行关系抽取。
- 未提供专门的针对概念、实体和事件抽取的支持。
- 支持中文关系抽取，仅需要引入中文相关的基础处理工具即可。
- 需要大量的标注语料支持，通过人工设置标注规则。



文本信息抽取实践方法

PlantData

- 目前还没有统一的实现各类信息抽取的现成工具。
- 把现有的工具进行集成，依据抽取任务使用不同的工具
  - NLP分词、命名实体识别工具：NLPIR、LTP、FudanNLP、Stanford NLP……
  - 关系抽取工具：DeepDive
- 对于行业抽取任务，需要针对性的方法来完成
- 通常基于已有的结构化知识进行远程监督学习

事件抽取

PlantData

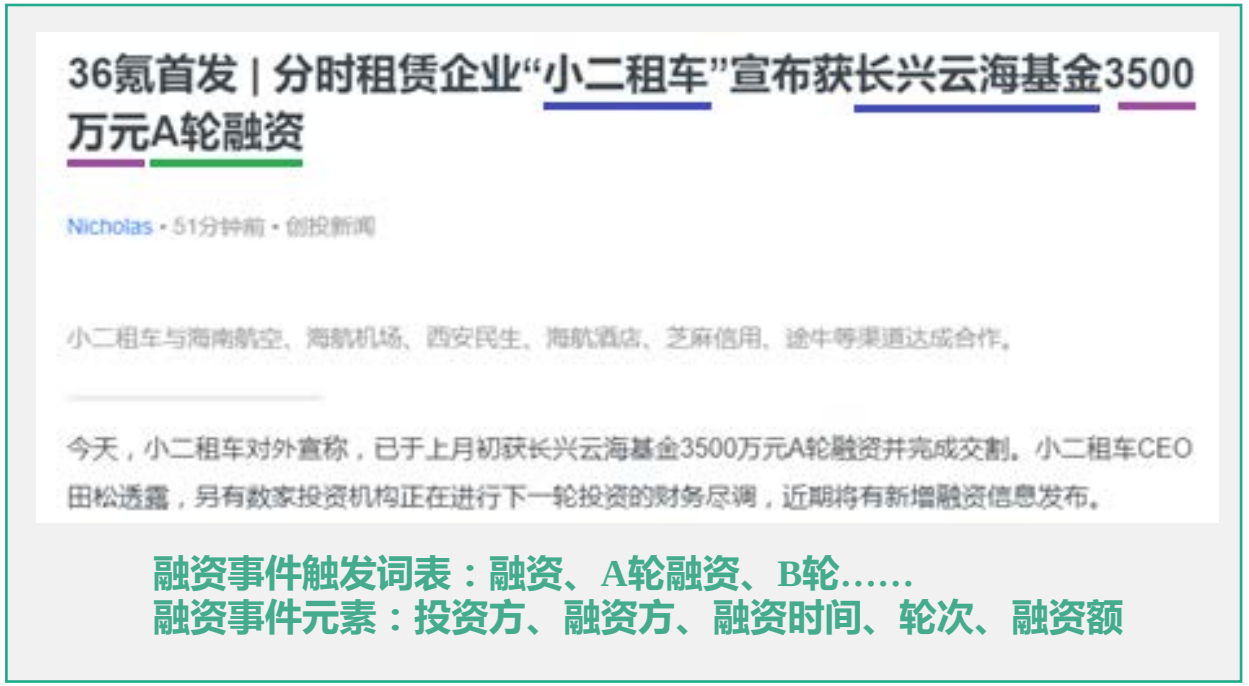
- 事件抽取可以分为预定义事件抽取和开放域事件抽取，行业知识图谱中主要为预定义事件抽取。

- 采用模式匹配方法，包括三个步骤：

① 准备事件触发词表

② 候选事件抽取：寻找含有触发词的句子

③ 事件元素识别：根据事件模版抽取相应的元素



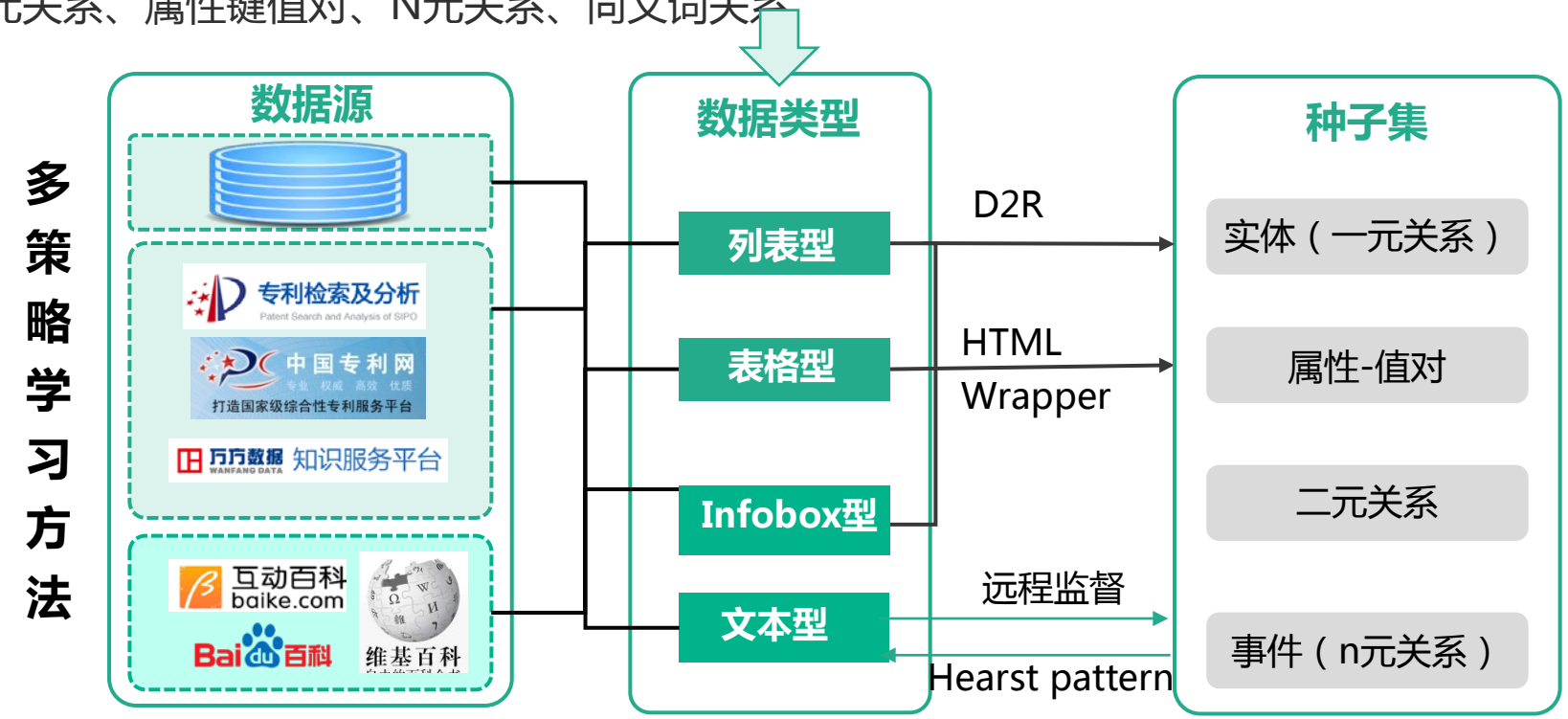
融资事件触发词表：融资、A轮融资、B轮……
融资事件元素：投资方、融资方、融资时间、轮次、融资额

知识抽取最佳实践：多策略学习方法

PlantData

利用不同数据源之间的冗余信息，使用较易抽取的信息（结构化数据库）来辅助抽取那些不易抽取的信息。

- **多数据源：结构化数据、半结构数据、文本数据**
- **目标数据类型：**
  - 不同类型目标数据（整数型、Map型、List型、Range型...）
  - 二元关系、属性键值对、N元关系、同义词关系



多策略学习方法示例（1）

PlantData

结构化数据

中国工商数据库

D2R

RDF

<深圳迈瑞生物医疗电子股份有限公司, 注册号, 2101332100187>

<深圳迈瑞生物医疗电子股份有限公司, 法人代表, 张平>

<深圳迈瑞生物医疗电子股份有限公司,注册资本, 50万人民币>

半结构化数据

专利检索及分析

Patent Search and Analysis of SIPO

一种便携超声诊断仪 【公开】

申请号：CN200710124611.7

申请日：2007.11.19

公开（公告）号：CN101438965A

公开（公告）日：2009.05.27

IPC分类号：A61B8/00；G12B15/00；H05K7/20；

申请（专利权）人：深圳迈瑞生物医疗电子股份有限公司；

发明人：姚海锋；彭国庆；高清山；

<一种便携超声诊断仪, 申请号, CN20071012461.7>

<一种便携超声诊断仪, 申请日, 2007.11.19>

<一种便携超声诊断仪, 申请人, 深圳迈瑞生物医疗电子股份有限公司>

文本数据

36氪首发 | 分时租赁企业“小二租车”宣布获长兴云海基金3500万元A轮融资

Nicholas · 51分钟前 · 创投新闻

小二租车与海南航空、海航机场、西安民生、海航酒店、芝麻信用、途牛等渠道达成合作。

今天，小二租车对外宣称，已于上月初获长兴云海基金3500万元A轮融资并完成交割。小二租车CEO田松透露，另有数家投资机构正在进行下一轮投资的财务尽调，近期将有新增融资信息发布。

<A轮融资[小二租车], investor, 长兴云海基金>

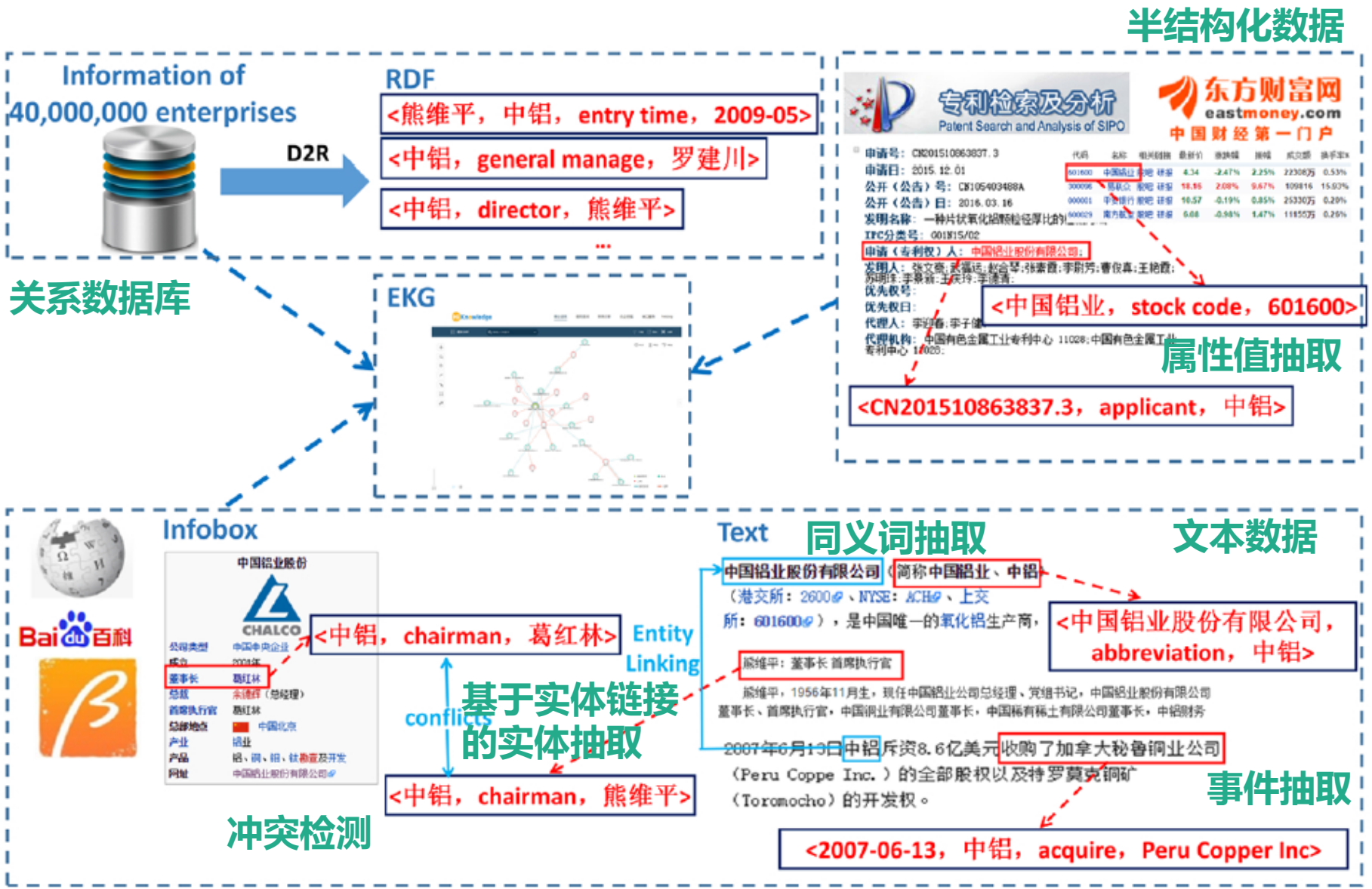
<A轮融资[小二租车], financed party, 小二租车>

<A轮融资[小二租车], round, A轮>

<A轮融资[小二租车], amount, 3500万元>

多策略学习方法示例（2）

PlantData



PlantData

行业知识图谱关键技术

- 知识建模
- 知识获取
- **知识融合**
- 知识存储
- 知识计算
- 知识应用

知识融合

PlantData

- 知识图谱中的知识融合是一件非常复杂的工作，包括数据模式层（概念、概念的上下位关系、概念的属性的融合与数据层的融合。

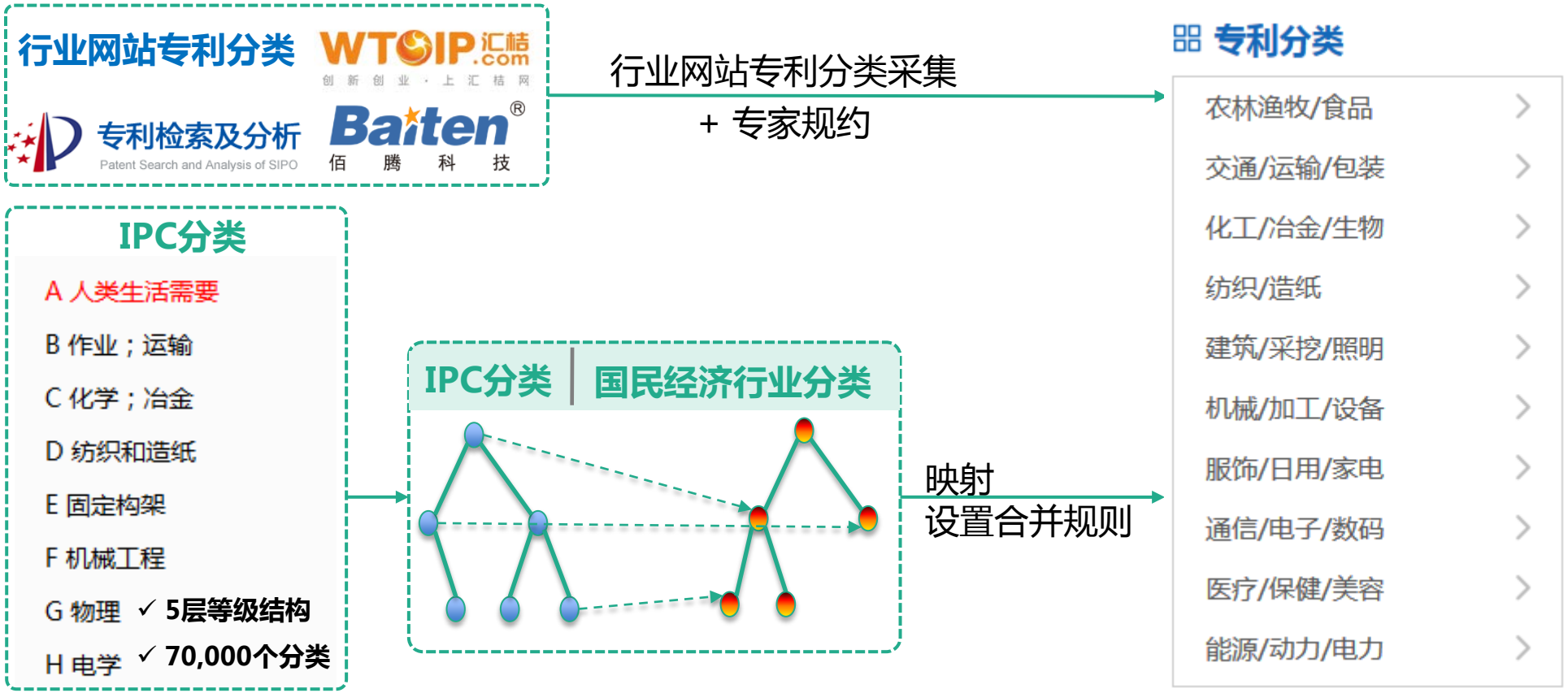
- 行业知识图谱的数据模式通常采用自顶向下和自底向上结合的方式，因此基本都经过人工的校验，保证了可靠性；因此，知识融合的关键任务在数据层的融合。

- 对于数据层的融合，为保证数据的质量，通常在知识抽取环节中进行控制，减少知识融合过程的难度。

知识融合实践 ——数据模式层融合

PlantData

- 数据模式层融合：行业知识图谱的数据模式层通常是由专家人工构建或从可靠的结构化数据中映射得到的，通常在映射时会通过设置融合的规则来确保数据的统一。



知识融合实践 ——数据层融合

PlantData

- **实体合并**
  - 在构建行业知识图谱时，实体优先从结构化的数据在获取；对于结构化的数据，通常有对实体进行唯一标识的主键，因此在进行知识抽取时即可设定实体合并的依据。
  - 从非结构化数据中抽取的实体，同样使用设置合并条件的规则来完成实体的合并；
    例如：
    - 企业合并：可以通过企业名称直接合并
    - 企业高管合并：人名相同 + 同一企业 （企业高管中同名的概念极低）

- **实体属性与关系的合并**
  - 具有时态特性的属性（如）：使用新的数据覆盖老的数据
  - 依据数据源的可靠性进行选取：结构化数据源中的质量通常较高

数据层融合示例——人物实体合并

PlantData

前提规则：

规则1：申请人为单一企业时，专利发明人都属于该企业，或与该企业有密切关联；

规则2：同名自然人属于同一专利的两个不同申请企业的概率忽略不计。

人物实体对齐步骤：

（1）对申请人为单一企业的组合，基于规则1，将发明人所属企业作为消歧标识，对发明人进行对齐；

（2）对申请人为多个企业的组合，利用（1）的结果和定义2进行人物实体对齐；另外为未合并的自然人数据添加已合并之外的申请人企业消歧标识；

（3）重复步骤（1）和步骤（2），直至自然人数据合并完成。

PlantData

行业知识图谱关键技术

- 知识建模
- 知识获取
- 知识融合
- **知识存储**
- 知识计算
- 知识应用

图数据存储

PlantData

知识图谱是基于图的数据结构，其存储方式主要有两种方式：RDF存储 和 图数据库(Graph Database)。

A triplestore or RDF store is a purpose-built database for the storage and retrieval of triples through semantic queries. A triple is a data entity composed of subject-predicate-object. [Wikipedia]

A graph database has a more generalized structure than a triplestore, using graph structures with nodes, edges, and properties to represent and store data. [Wikipedia]

常见的图数据存储 — Graph DBMS

PlantData

Rank Apr 2017	Rank Mar 2017	Rank Apr 2016	DBMS	Database Model	Score Apr 2017	Score Mar 2017	Score Apr 2016
1.	1.	1.	Neo4j	Graph DBMS	34.91	+0.59	+3.00
2.	2.	2.	OrientDB	Multi-model	5.44	+0.10	-0.87
3.	3.	3.	Titan	Graph DBMS	4.63	-0.24	-0.48
4.	4.	↑ 5.	ArangoDB	Multi-model	2.61	+0.23	+1.10
5.	5.	↓ 4.	Virtuoso	Multi-model	1.87	-0.11	-0.53
6.	6.	6.	Giraph	Graph DBMS	1.02	-0.04	+0.17
7.	7.	↑ 8.	AllegroGraph	Multi-model	0.49	+0.01	-0.09
8.	8.	↑ 19.	GraphDB	Multi-model	0.42	+0.02	+0.36
9.	9.	↓ 7.	Stardog	Multi-model	0.41	+0.03	-0.17
10.	10.	↓ 9.	Sqrrl	Multi-model	0.39	+0.06	+0.09
11.	11.	↓ 10.	InfiniteGraph	Graph DBMS	0.23	+0.01	+0.00
12.	12.		Dgraph	Graph DBMS	0.22	+0.01	
13.	13.	↑ 18.	Blazegraph	Multi-model	0.20	+0.03	+0.13
14.	14.	↑ 15.	FlockDB	Graph DBMS	0.13	+0.01	-0.01
15.	↑ 16.		Graph Engine	Multi-model	0.12	+0.03	
16.	↓ 15.	↑ 17.	InfoGrid	Graph DBMS	0.12	+0.01	-0.01
17.	17.	↓ 13.	HyperGraphDB	Graph DBMS	0.10	+0.01	-0.05
18.	18.	↓ 12.	VelocityDB	Multi-model	0.08	+0.01	-0.10
19.	19.	↓ 14.	GlobalsDB	Multi-model	0.06	-0.01	-0.09
20.	↑ 21.	20.	GraphBase	Graph DBMS	0.04	+0.04	+0.01

图数据存储的选用指标

PlantData

- 数据存储支持
- 数据操作和管理方式
- 支持的图结构
- 实体和关系表示
- 查询机制

数据存储支持

PlantData

图数据库	支持内存	支持外存	依赖外部存储	支持索引
Neo4j	•	•		•
Titan	•		•	•
Virtuoso	•	•		•
AllegroGraph	•	•		•
DEX	•	•		•
Filament	•		•	
G-Store		•		
HyperGraphDB	•	•	•	•
InfiniteGraph		•		•
Sones	•			•
vertexDB		•	•	

数据操作和管理方式

PlantData

图数据库	数据定义语言	数据操作语言	查询语言	API
Neo4j			•	•
Titan	•	•	•	•
Virtuoso	•	•	•	•
AllegroGraph	•	•	•	•
DEX				•
Filament				•
G-Store	•		•	•
HyperGraphDB				•
InfiniteGraph				•
Sones	•	•	•	•
vertexDB				•

支持的图结构

PlantData

图数据库	简单图	超图	嵌套图	属性图	节点标签	节点属性	关系有向	边标签	边属性
Neo4j				•	•	•	•	•	•
Titan	•			•	•	•	•	•	•
Virtuoso				•	•		•	•	
AllegroGraph	•				•		•	•	
DEX			•	•	•	•	•	•	•
Filament	•				•		•	•	
G-Store	•				•		•	•	
HyperGraphDB		•			•		•	•	
InfiniteGraph				•	•	•	•	•	•
Sones		•		•	•	•	•	•	•
vertexDB	•				•		•	•	

实体和关系表示

PlantData

	Schema			Instance					
图数据库	节点类型	属性类型	关系类型	对象节点	数值节点	复杂节点	对象关系	简单关系	复杂关系
Neo4j				•	•		•	•	
Titan			•	•			•	•	
Virtuoso					•			•	
AllegroGraph					•			•	
DEX	•		•	•	•		•	•	
Filament					•			•	
G-Store					•			•	
HyperGraphDB	•		•		•			•	•
InfiniteGraph	•		•	•	•		•	•	
Sones					•			•	•
vertexDB					•			•	

查询机制

PlantData

图数据库	查询语言	API	查询	推理	分析
Neo4j	•	•	•		•
Titan	•	•	•		
Virtuoso	•	•	•	•	•
AllegroGraph	•	•	•	•	•
DEX		•	•		•
Filament		•	•		
G-Store	•		•		
HyperGraphDB		•	•		
InfiniteGraph		•	•		
Sones	•		•		•
vertexDB		•	•		

neo4j

PlantData

- Neo4j is a highly scalable native graph database that leverages data relationships as first-class entities, helping enterprises build intelligent applications to meet today’s evolving data challenges.
- **特点**
  - 原生图存储和处理
  - 支持ACID事务处理
  - 不使用 Schema
- **不足**
  - 企业数据管理场景下不使用 Schema 会难以从整体把握数据
  - 不支持时态信息的存储
  - 非企业版本受数据量、查询速度等方面的限制

大规模知识图谱存储解决方案

PlantData

指导思想

数据思维

依据数据特点进行数据存储结构选择与设计

No Size Fits All

没有一种通用的存储方案能够解决所有问题

整体原则

- 基础存储支撑灵活
- 基础存储可扩展、高可用
- 按需要进行数据分割
- 适时使用缓存和索引
- 善于利用现有成熟存储
- 保持图形部分数据的精简
- 不在图中作统计分析计算

大规模知识图谱存储最佳实践（1）

PlantData

1. 基础存储

- 可按数据场景选择使用关系数据库、NoSQL数据库及内存数据库。
- 基础存储保证可扩展、高可用

2. 数据分割

- 属性表：依据数据类型划分
  基本类型：整数表、浮点数表、日期类型表、...
  集合类型：List型表、Range型表、Map型表、...
- 大属性单独列表：例如数量超过10M的属性单独列表

3. 缓存与索引

- 使用分布式 Redis 作为缓存，按需对数据进行缓存。
- 对三元组表按需进行索引，最多情况下可建立九重索引。

大规模知识图谱存储最佳实践（2）

PlantData

4. 善于使用现在成熟存储

- 使用 ElasticSearch 实现数据的全文检索
- 结构固定型的数据可使用关系数据库或NoSQL

5. 对于非关系型的数据尽量不入图存储，避免形成大节点

- 非关系型的数据，使用适合的数据存储机器进行存储，通过实体链接的方式实现与图谱数据的关联。

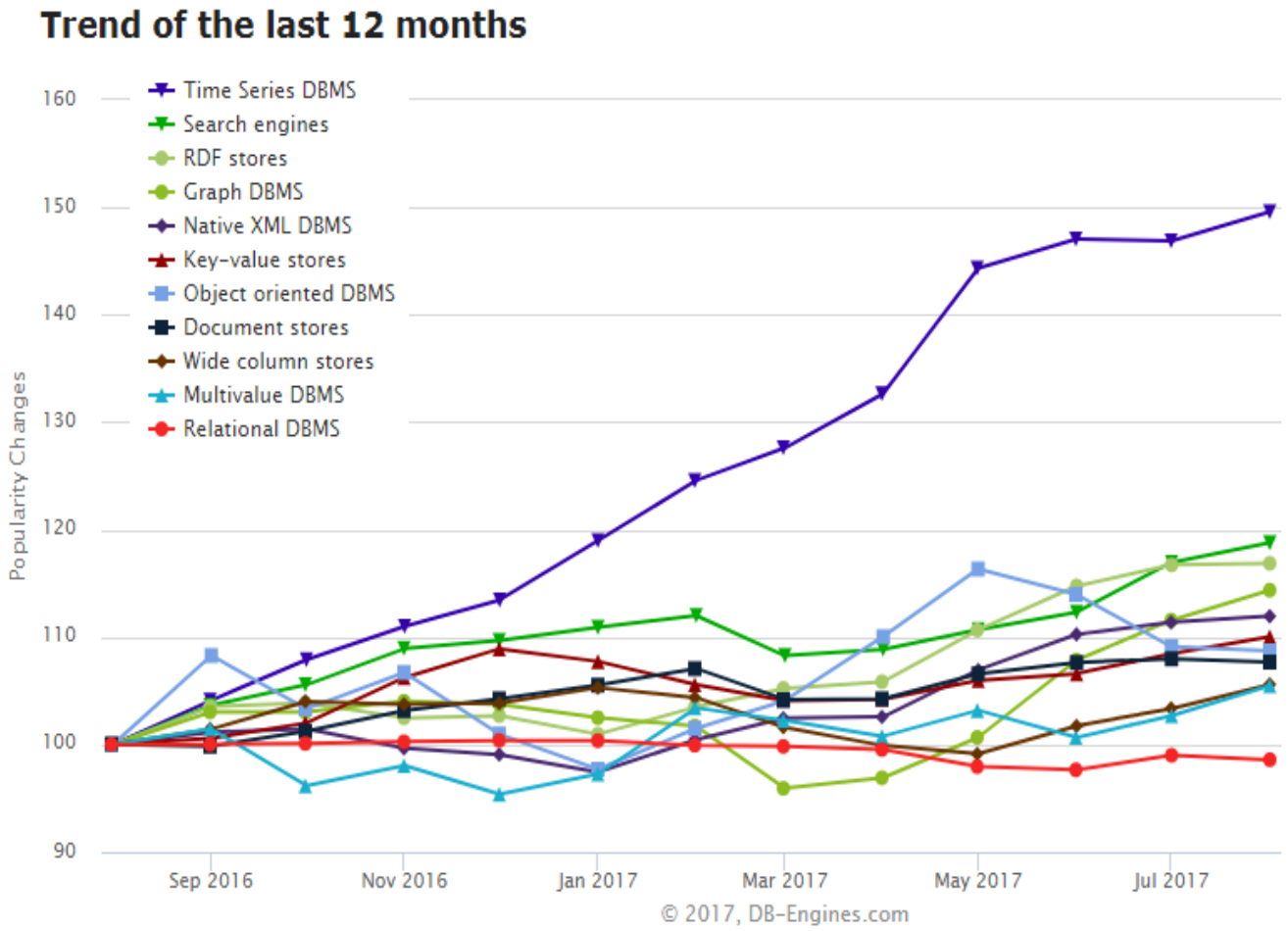
6. 不直接在图存储中进行统计分析计算

- 对于需要进行统计分析计算的数据，需要导出到合适的存储中进行。

知识图谱中的时态信息

PlantData

- 事实的生成时间
- 某事实的有效时间段
- 示例：融资事件的时间



知识图谱时态信息存储

PlantData

- 在知识图谱存储中应用的为历史数据库。

- **历史数据库**：记录事实的有效时间，用有限的数据冗余实现数据时态信息的应用。

- 实现原则
  - 在基础知识图谱的基础上，构建针对时态数据处理的中间件；
  - 对于特定类型的时序型数据，采用其它的存储机制进行存储。

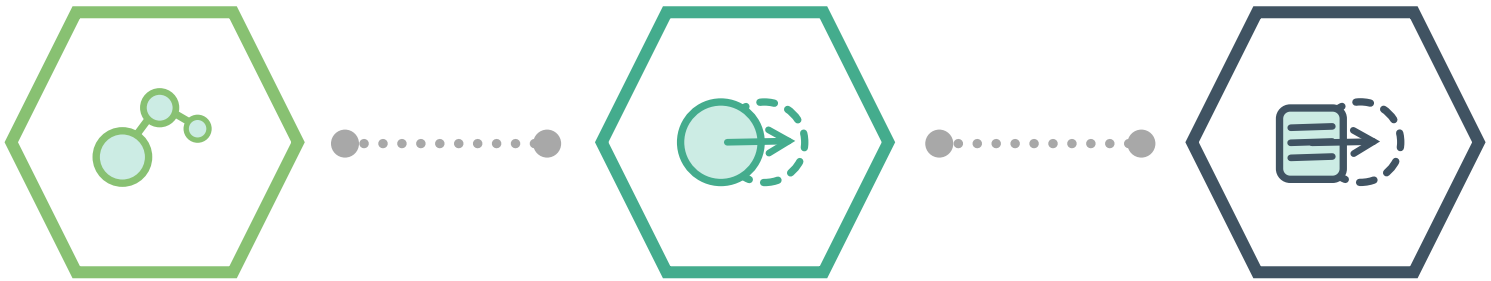
PlantData

行业知识图谱关键技术

- 知识建模
- 知识获取
- 知识融合
- 知识存储
- **知识计算**
- 知识应用

PlantData

知识计算关键技术



01 图挖掘计算

02 基于本体的推理

03 基于规则的推理

图挖掘计算

PlantData

- **集成实现基本图算法**
  - 图遍历：广度优先遍历、深度优先遍历
  - 最短路径查询： Dijkstra（迪杰斯特拉算法）、Floyd（弗洛伊德算法）
  - 路径探寻：给定两个或多个节点，发现它们之间的关联关系
  - 权威节点分析：PageRank算法
  - 族群发现：最大流算法
  - 相似节点发现：基于节点属性、关系的相似度算法

基于本体的知识推理

PlantData

- **本体推理基本方法**
  - 基于表运算及改进的方法：FaCT++、Racer、 Pellet Hermit等
  - 基于一阶查询重写的方法（Ontology based data access，基于本体的数据访问）
  - 基于产生式规则的算法（如rete）：Jena 、Sesame、OWLIM等
  - 基于Datalog转换的方法如KAON、RDFox等。
  - 回答集程序 Answer set programming。

本体知识推理工具——RDFox

PlantData

- **RDFox 特点**
  - 支持共享内存并行OWL 2 RL推理
  - 三元组数据可以导出为Turtle文件，规则文件可以导出为RDF数据记录文件；全部数据内容可以导出为二进制文件，可完全恢复数据存储状态。
  - 支持Java、Python多语言APIs访问，并且 RDFox 还支持一种简单的脚本语言与系统的命令行交互。

- **RDFox完全基于内存，对硬件的要求较高**

基于本体的知识推理应用示例——冲突检测

PlantData

花椒直播	亿元及以上 人民币	B轮
花椒直播	1000,000,00 人民币	B轮

基于规则的推理

PlantData

- 在知识图谱基础知识的基础上，专家依据行业应用的业务特征进行规则的定义。

- 引擎基于基础知识与所定义的规则，执行推理过程给出推理结果。

基于规则推理工具——Drools 规则定义

PlantData

```
1. import com.hiekn.ruleengine.bean.RiskBean;
2. rule highRiskRule1 //高风险企业规则一：企业注册资金每100万加1分，成立的时长每年加1分，每
   有失信记录减3分，有税务问题每次减少5分。
3.         when
4.                 $risk : RiskBean(eval(true))
5.         then
6.                 $risk.addOrMinusScore (702, 1);
7.                 $ risk.addOrMinusScore (703, 1);
8.                 $ risk.addOrMinusScore (711, -3);
9.                 $ risk.addOrMinusScore (712, -5);
10.end
```

PlantData

行业知识图谱关键技术

- 知识建模
- 知识获取
- 知识融合
- 知识存储
- 知识计算
- **知识应用**

知识应用关键技术

PlantData

01 **语义搜索**

02 智能问答

03 可视化辅助决策

语义搜索

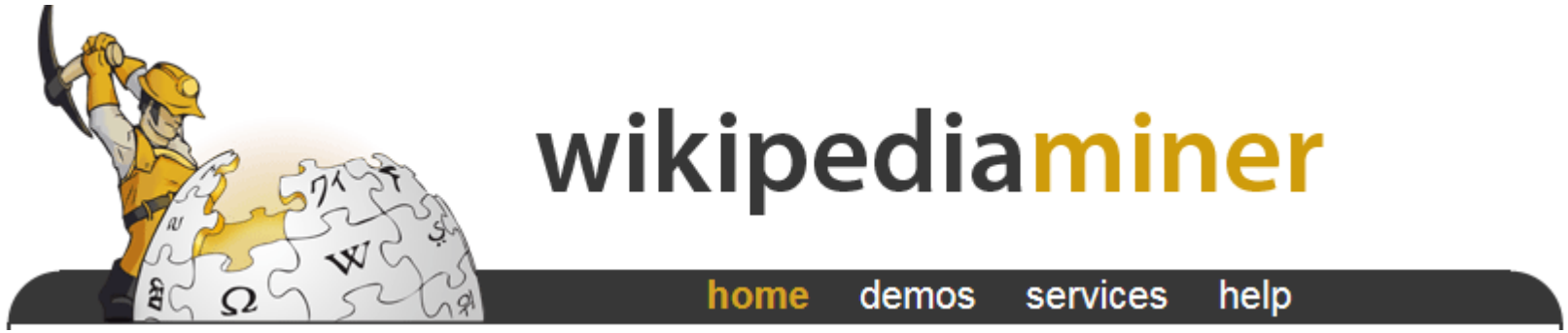
PlantData

- 知识图谱提出的初衷即为解决搜索的准确率问题；传统基于关键词的检索完全不考虑语义信息，主要面临两个难题：
  - 自然语言表达的多样性
  - 自然语言的歧义

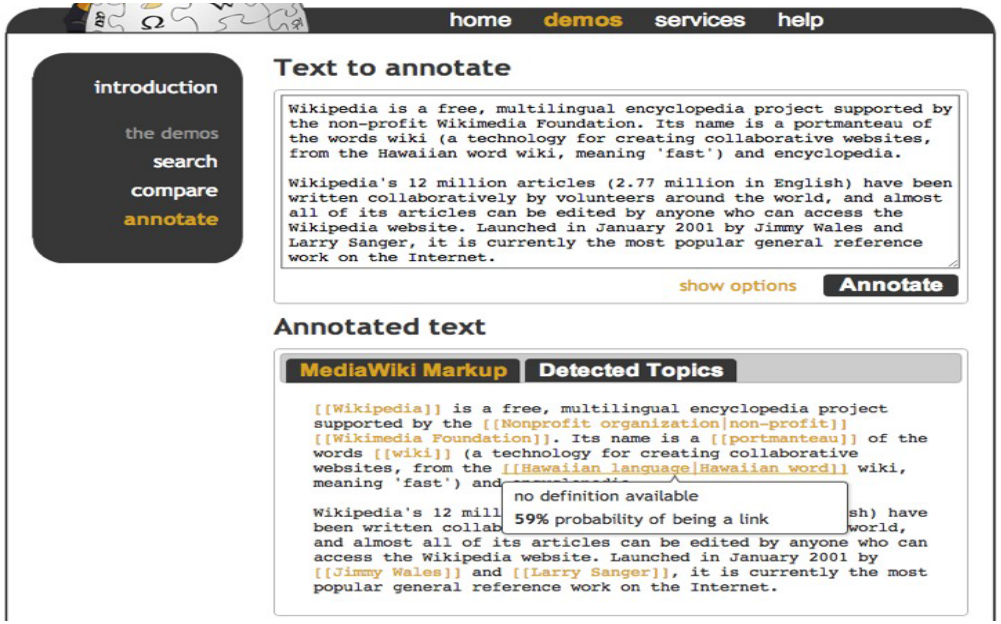
- 解决方案
  - 实体链接
  - 基于知识图谱的语义搜索

实体链接工具——Wikipedia Miner

PlantData



- Open source
- (Public) web service
  - Java
  - Hadoop preprocessing pipeline
- Lexical matching + machine learning
- Target KB: Wikipedia
- See http://wikipedia-miner.cms.waikato.ac.nz

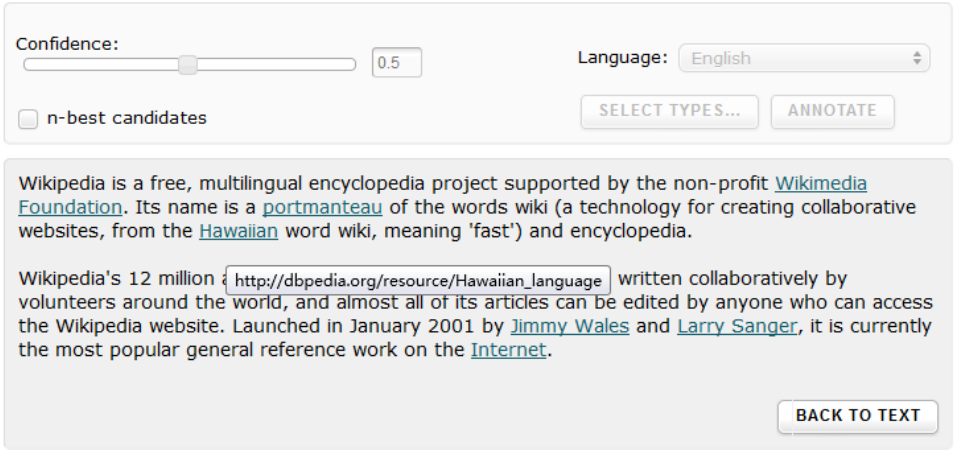


Entity Linking and Retrieval Tutorial
@ SIGIR 2013

实体链接工具—— DBpedia Spotlight

PlantData

- Open source
- Public web service
  - Disambiguation in local context
  - vector-space model using bag-of-words and cosine similarity
  - (actually, Lucene)
- Target KB: DBpedia
- See http://spotlight.dbpedia.org



Demo: http://dbpedia-spotlight.github.io/demo/

现有实体链接工具使用总结

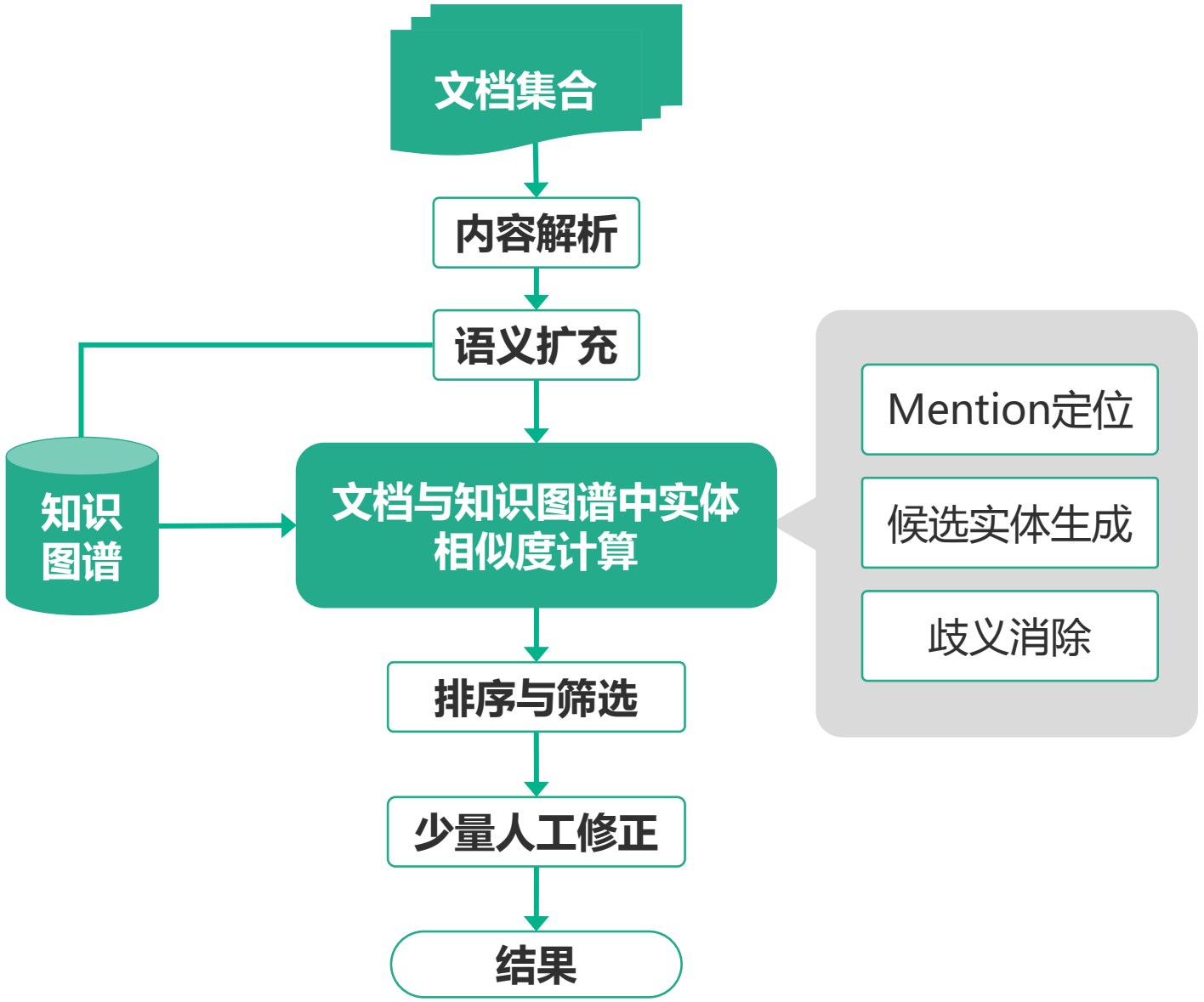
PlantData

- 大部分都是针对百科类的知识库工作的
- 基本不支持中文的处理

实体链接的基本方法过程

PlantData

- 基于向量模型相似度计算的实体链接方法
- 基于知识图谱语义扩展的实体链接方法
- 基于propagation计算相似度的实体链接方法



基于知识图谱的语义搜索

PlantData

传统搜索

- 基于关键词的搜索

语义搜索

- 基于知识图谱对用户输入进行理解，识别实体、概念和属性，并返回实体、关系、链接的数据等丰富的结果

智能问答

- 自然语言智能问答

语义搜索示例——实体搜索

PlantData

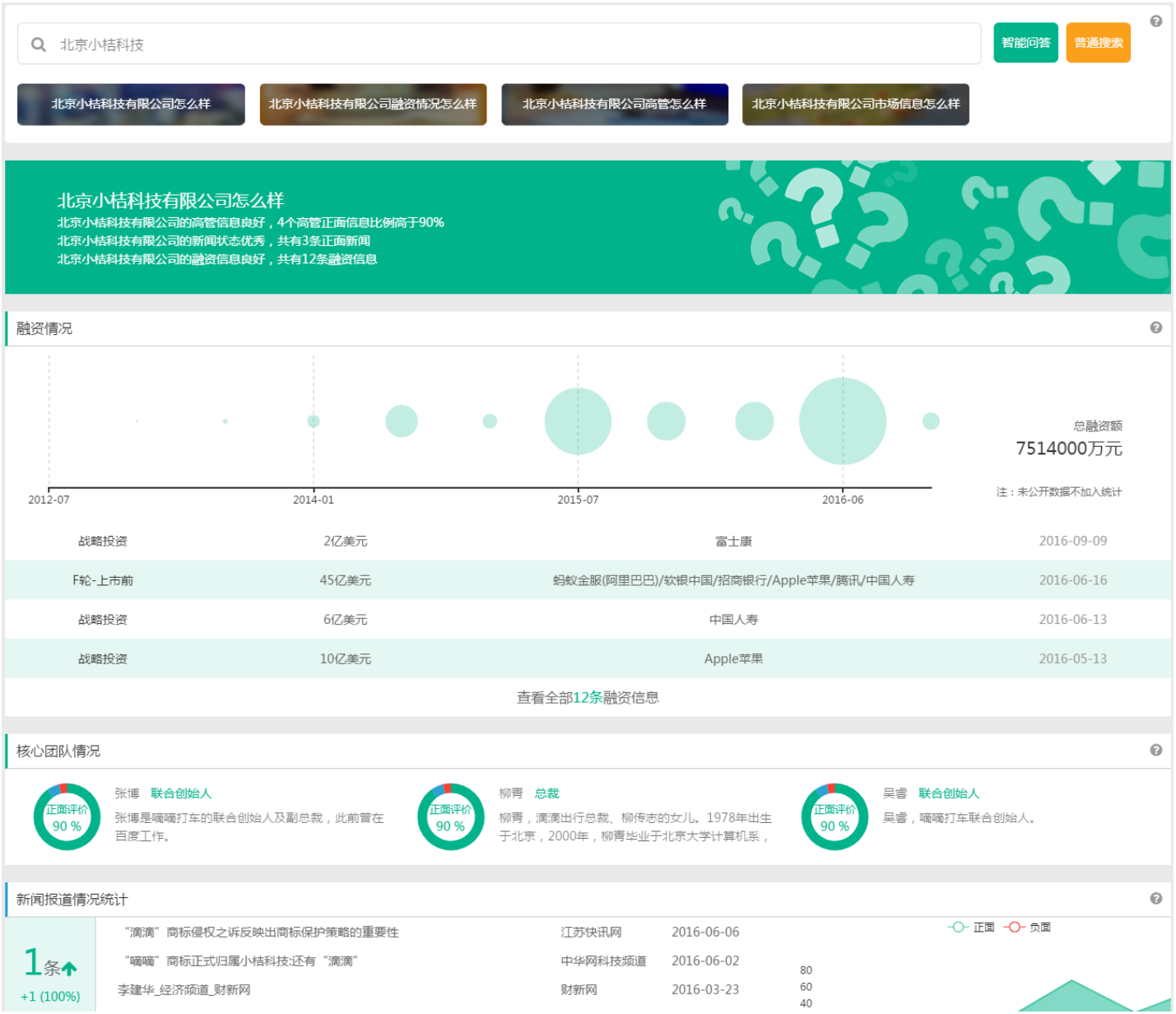
搜索输入 “北京小桔科技”

相关问题

融资状况

高管信息

链接的新闻数据



知识应用关键技术

PlantData

01 语义搜索

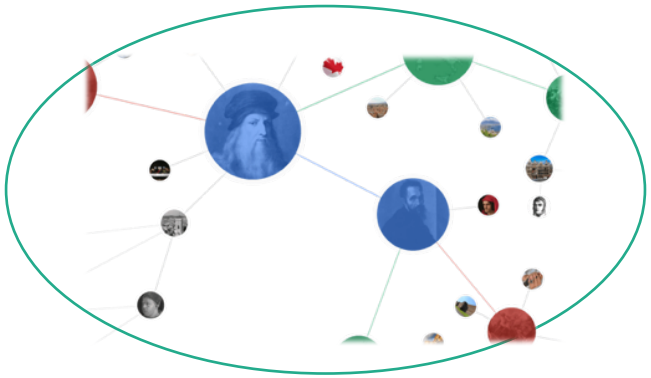
02 **智能问答**

03 可视化辅助决策

智能问答——基本过程

PlantData

基于知识图谱的自动问答系统的基本过程



智能问答——基本访求

PlantData

01基于信息检索的方法

02基于语义解析的方法

方法分类

03基于规则的专家系统方法

04 基于深度学习的方法

智能问答方法（1）

PlantData

01 基于信息检索的方法

- 首先利用中文分词、命名实体识别等自然语言处理工具找到问句中所涉及到的实体和关键词，然后去知识资源库中去进行检索。

- 优点：实现简单，应用面广，在大部分场景下均可得到结果。
- 缺点：要求答案中必须至少包含问句中的一个字或词，所以不如语义解析方法精确。

- 改进：基于知识图谱知识进行语义扩充，提高匹配率；基于知识图谱进行检索时的语义消歧。

智能问答方法（2）

PlantData

02 基于语义解析的方法

- 基于语义解析的方法非常符合人们的直觉，它将一个自然语言形式的问句，按照特定语言的语法规则，解析成语义表达式，在得到语义表达式之后，即可非常容易地将其转化为某种数据库的查询语言。

- 首先自然语言问句的词汇被映射到语义表达式中的词汇，然后按照特定的规则将汇组合起来，进而得到了最终的语义表达式。

- 改进方法：在特定的领域里边，基于知识图谱的实体、属性、概念等进行词法解析与映射，然后基于图结构进行语法规则匹配。（子图查询匹配）

智能问答方法（3）

PlantData

03 基于规则的专家系统方法

- 专家系统是一个具有大量的专门知识与经验的程序系统，它应用人工智能技术和计算机技术，根据某领域一个或多个专家提供的知识和经验，进行推理和判断，模拟人类专家的决策过程，以便解决那些需要人类专家处理的复杂问题。

- 优点：在限定的领域范围内准确度高。
- 缺点：通用性欠缺，不能覆盖很多的应用场景。

智能问答方法（4）

PlantData

04 基于深度学习的方法

- 近几年卷积神经网络（CNN）和循环神经网络（RNN）在NLP领域任务中表现出来的语言表示能力，越来越多的研究人员尝试深度学习的方法完成问答领域的关键任务，包括 问题分类（question classification），语义匹配与答案选择（answer selection），答案自动生成（answer generation）；即对用户输入解析、答案查询与检索等环节进行优化。

- 优点：实现“端到端”的问答：把问题与答案均使用复杂的特征向量表示，使用深度学习来计算问题与答案的相似度。

- 不足：不支持复杂的查询；需要比较长的训练过程，不适用于现实应用场景中的知识更新后的实时查询。

智能问答的最佳实践方法

PlantData

基于语义解析的方法 ✚ 基于信息检索的方法

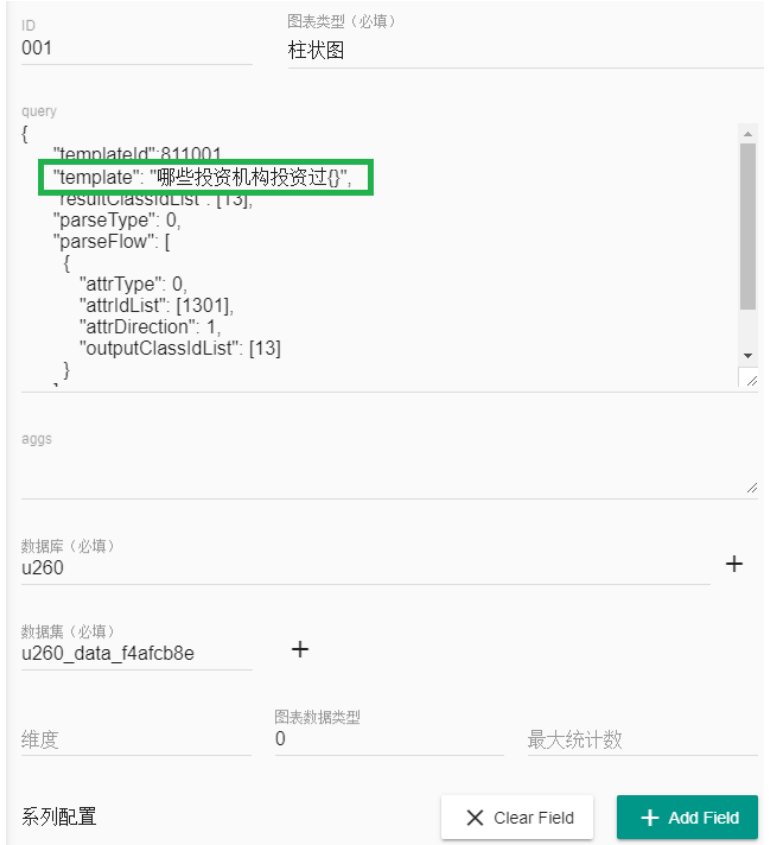
- 基于语义解析的方法可解释性强，并且能够方便地转换成知识图谱的查询，给出明确的答案；因此对于用户输入，首先使用基于语义解析的方法进行回答。

- 基于信息检索的方法应用面广，因此当语义解析方法无法给出结果时，则使用信息检索的方法进行回答。

基于语义解析的自动问答

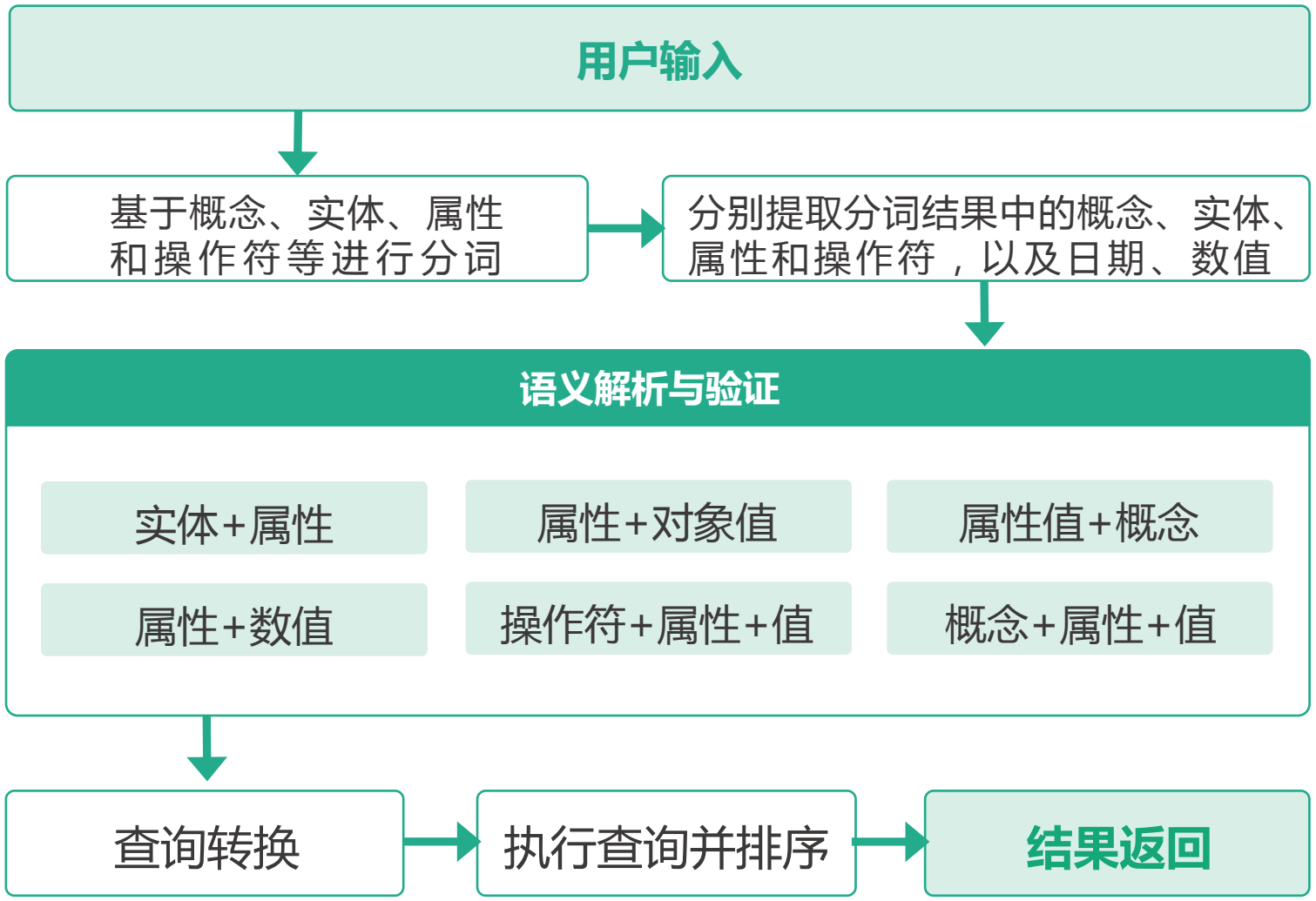
PlantData

- 人工配置语义解析模板
- 知识图谱通用的子图匹配模板



基于语义解析的自动问答

PlantData



北京小桔科技高管?

实体解析： 北京小桔科技

属性解析： 高管

语义解析结果：

北京小桔科技　高管

语义检索结果：

1、张博　2、柳青　3、吴睿

知识应用关键技术

PlantData

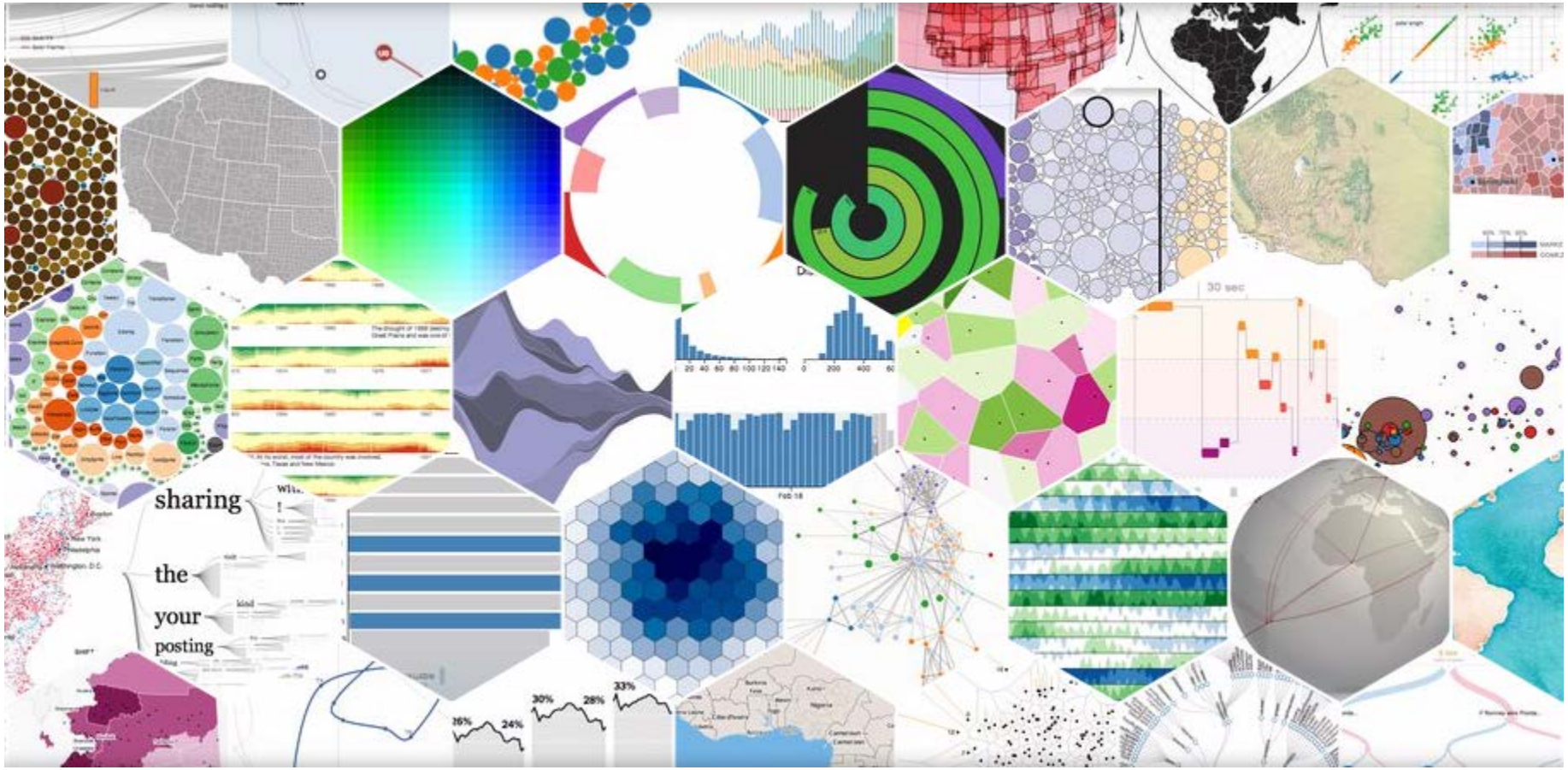
01 语义搜索

02 智能问答

03 **可视化辅助决策**

可视化工具——D3.js

PlantData



可视化工具——ECharts

PlantData

特性

- 丰富的图表类型
- 多个坐标系的支持
- 移动端的优化
- 深度的交互式数据探索
- 大数据量的展现
- 多维数据的支持以及丰富的视觉编码手段
- 动态数据
- 绚丽的特效

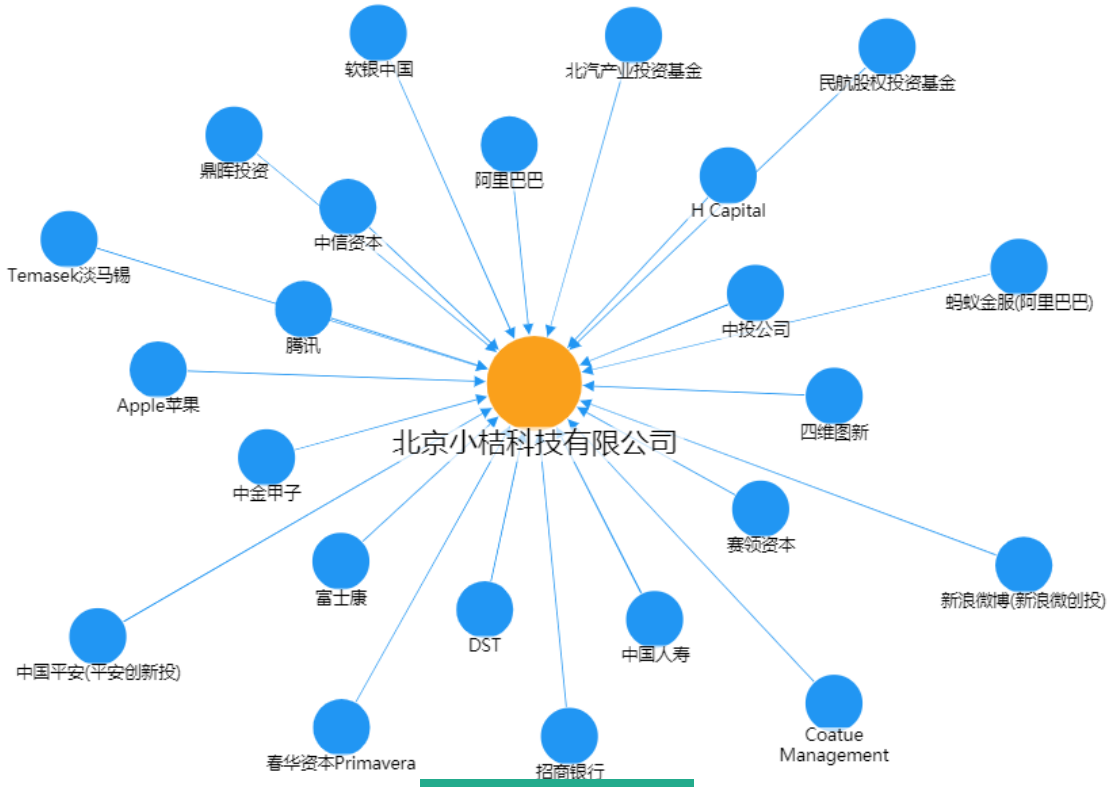
知识图谱可视化

PlantData

- 缺少面向知识图谱的可视化工具
- 需要思考：
  - 依托的设备及环境
  - 展现数据的什么特点
  - 数据量过大（小）时我该怎么做
- 实现方法：集成现有的可视化工具，实现知识图谱的可视化。

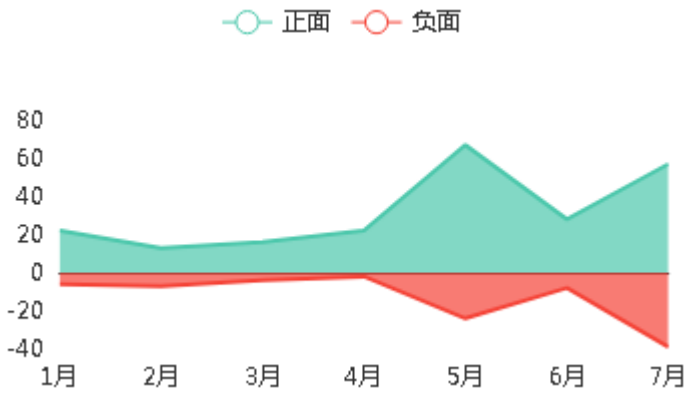
图谱可视化基本组件（1）

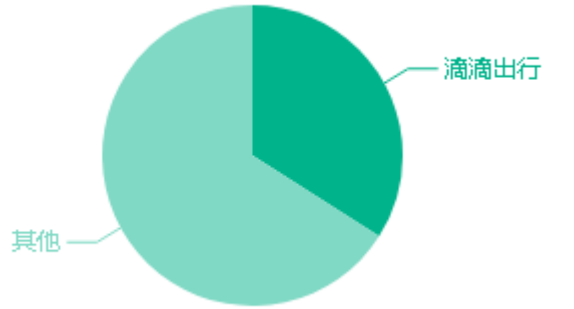
PlantData

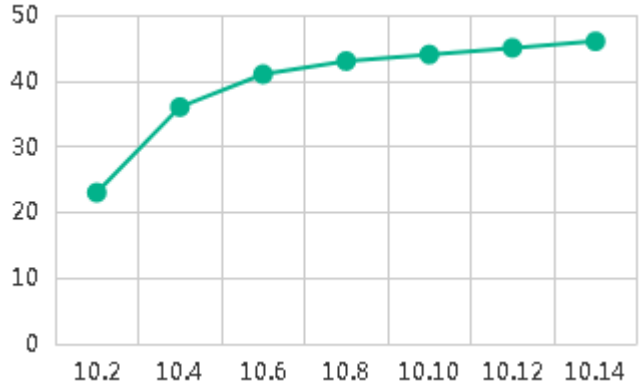


图谱展示

统计分析

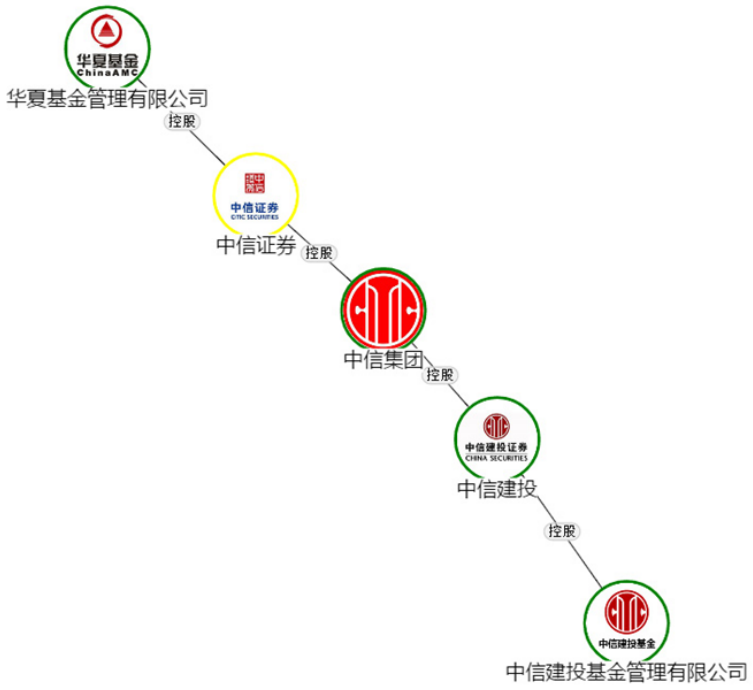




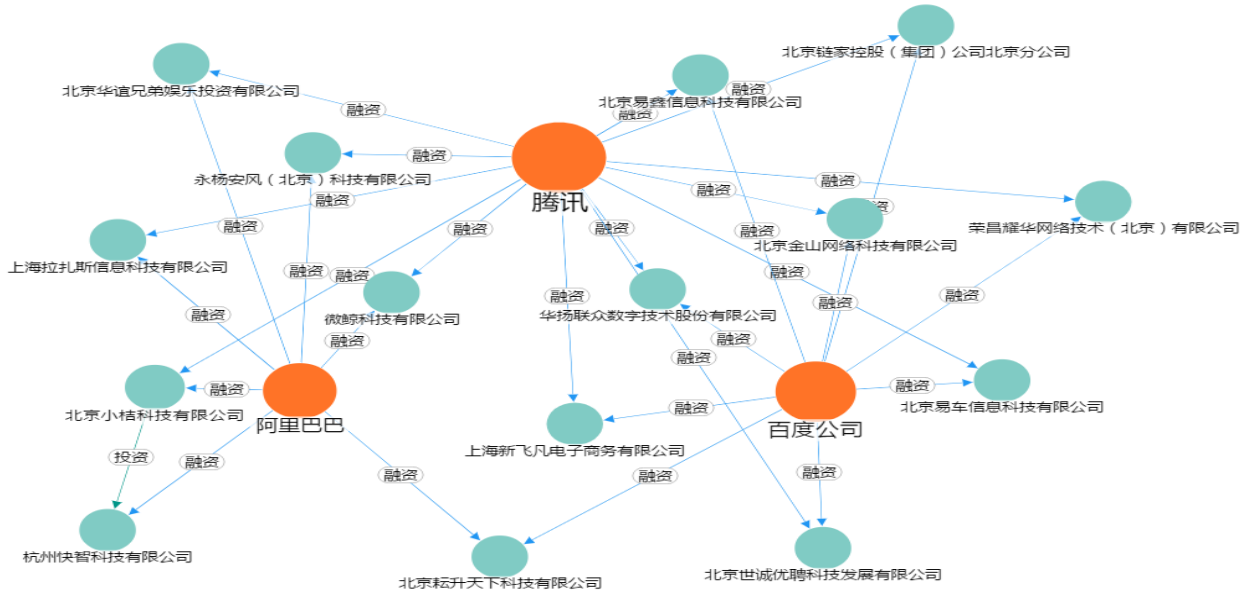


图谱可视化基本组件（2）

PlantData



最短路径发现



多结点关联探寻

图谱可视化演示

PlantData

基于知识图谱大数据的统一决策支持分析平台——PlantData

PlantData

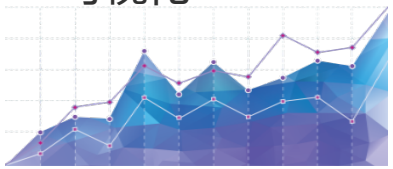
知识应用

知识整合

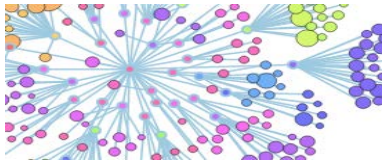
知识抽取

数据源

可视化

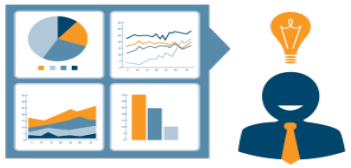


知识计算



智能问答 语义检索

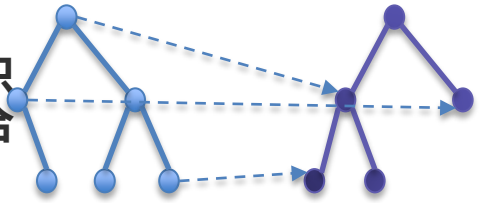
商业智能决策



数据智能平台

知识导入

知识融合



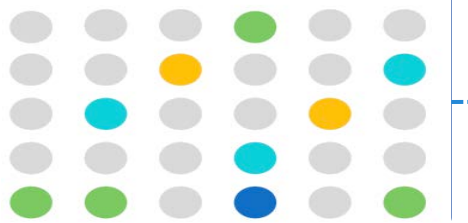
模式整合
实体对齐
冲突检测
冲突解决

行业知识图谱

知识抽取

- 结构化数据
- 半结构化数据
- 非结构化文本

D2R映射器
包装器
文本抽取引擎

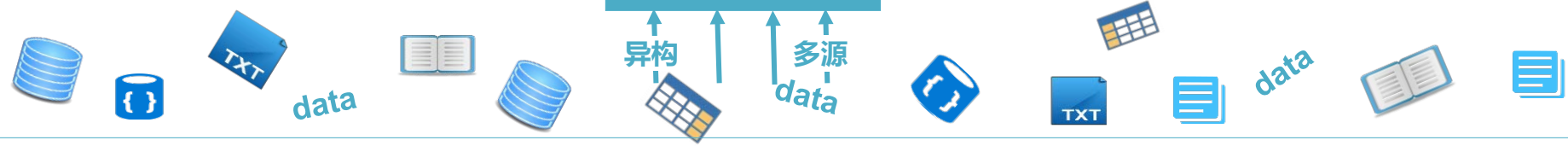


知 识 存 储

数据采集

异构

多源



PlantData 未来规划

PlantData

- 在更多的行业中应用知识图谱技术
- 研发更丰富的组件，使用户可以组合不同的组件实现应用场景
- 面向非商业用户的开放知识图谱平台，为开放知识图谱社区作贡献
  - 基础平台开放：让用户可以快速地在线体验知识图谱行业应用
  - 开放平台中的数据集
  - 提供面向用户的开发接口、集成接口