

链接数据 · 洞察价值



发现数据之美

-- 大规模行业知识图谱的构建和应用

丁军 上海海翼知信息科技有限公司

■ 现有大数据应用面临的挑战

■ 行业知识图谱解决方案

■ 案例：企业知识图谱的构建与应用

■ 总结和展望



挑战1、非结构化数据计算机难以理解



Web of Documents



人脑



正确获取文本内信息



计算机



非结构化数据计算机难以正确理解

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

挑战2、多源异构数据难以融合



新闻网站、公司研报、公司公告、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!

挑战3、数据模式动态变迁困难

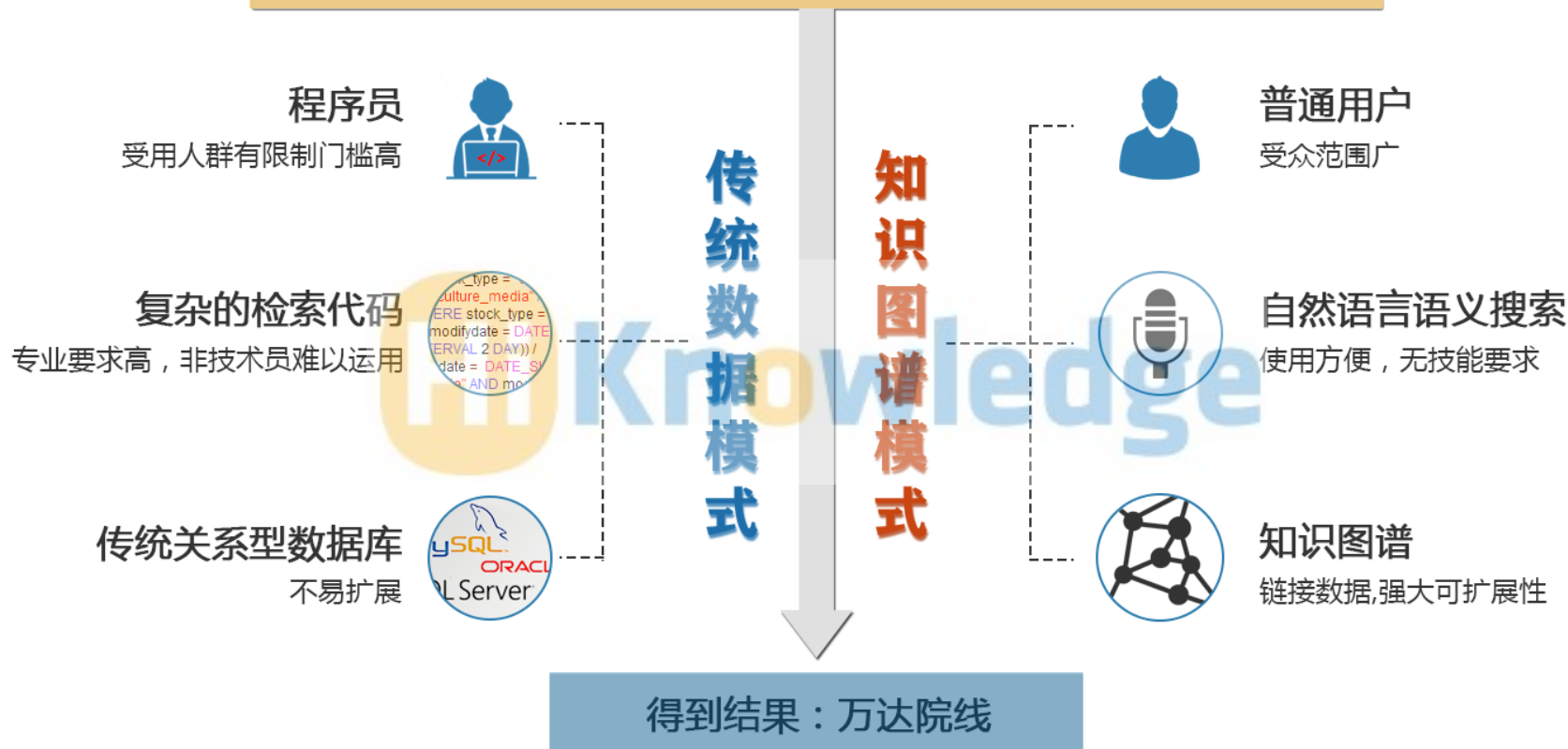


当前数据模式动态变迁困难,当客户新需求、业务新认知时程序员需痛苦的修改数据结构及业务逻辑,带来扩展性差、对客户响应慢、维护成本高等不良情况。

我们需要：可自由扩展的数据模式！

挑战4、数据使用专业程度过高

需求举例：了解最近连续两天涨幅大于9%的文化传媒股票



行业智能问答大幅降低数据使用门槛

- 现有大数据应用面临的挑战

- **行业知识图谱解决方案**

- 案例：企业知识图谱的构建与应用

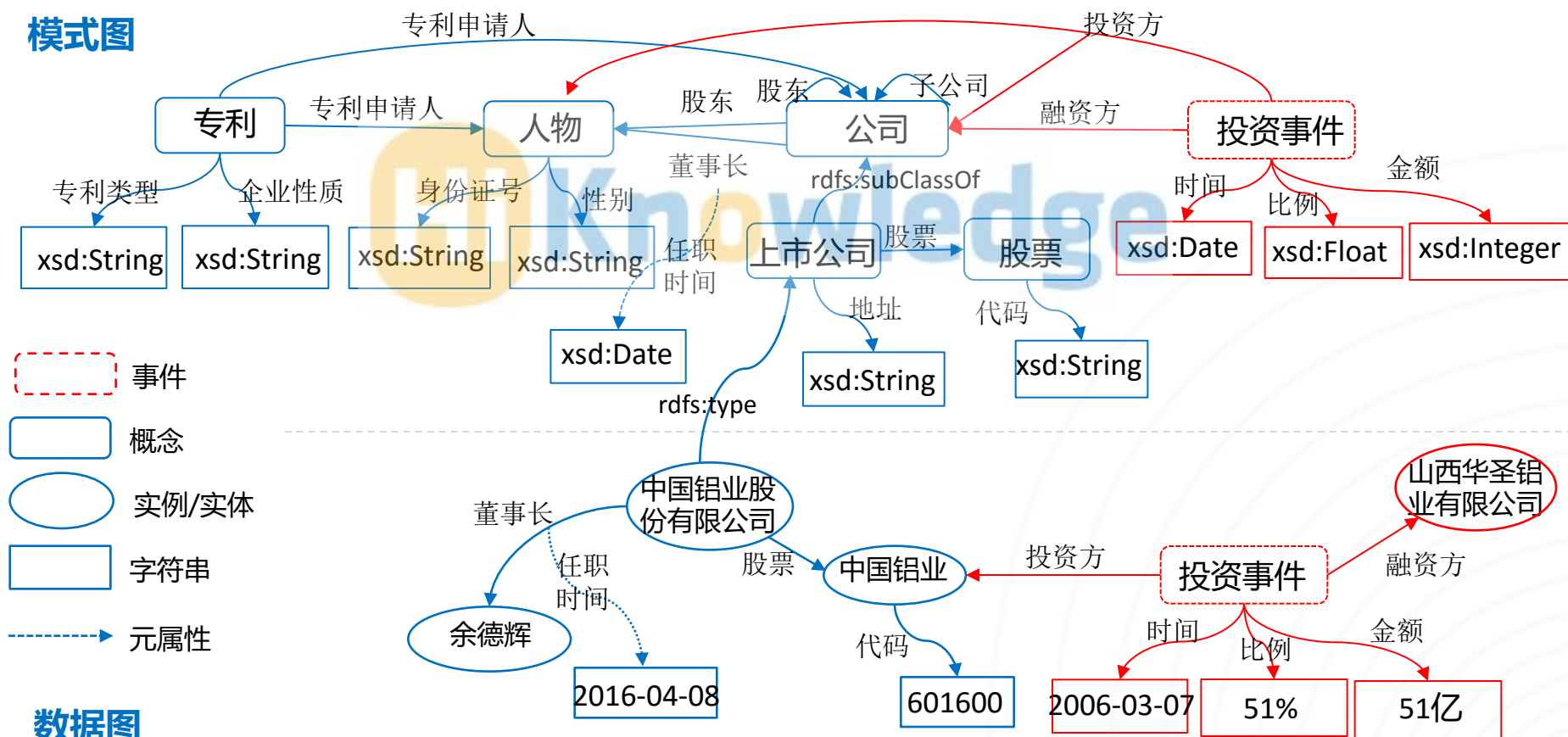
- 总结和展望



什么是知识图谱？

- 知识图谱本质上是一种语义网络
- 结点代表实体或概念，边代表实体/概念之间的各种语义关系

模式图



渐增式数据模式设计

初始设计的时候，很难清楚所有的概念，而知识图谱的动态可扩充性以及“无模式”特性使得用户很容易增加或修改模式。

数据集成更轻松

本体的语义互操作特性以及“链接数据”原则，使得来自不同供应商的数据集成更为方便。

现有标准支持

有RDF(S),OWL, SPARQL等标准，可以逐渐要求内容供应商支持。

语义搜索

用户可以查询具有某类特征的某类实体，比起基于关键词的搜索，更为精准。

我们提供行业知识图谱解决方案！

知识图谱图书馆应用案例

- ✓ 可视化的知识图谱编辑器
- ✓ 知识抽取与学习
- ✓ 近4000w书籍论文实体链接

专利图谱应用案例

- ✓ 海量专利文档图谱语义化处理
- ✓ 专利大数据语义检索
- ✓ 专利图谱关联分析

企业图谱应用案例

- ✓ 全国3000w企业360°全息画像
- ✓ 完整的企业社交谱系
- ✓ 客观的企业风险评价和财务实力洞察
- ✓ 已应用于券商、银行、P2P等需要追踪企业动态的场景



- 现有大数据应用面临的挑战

- 行业知识图谱解决方案

- **案例：企业知识图谱的构建与应用**

- 总结和展望



构建企业知识图谱的挑战

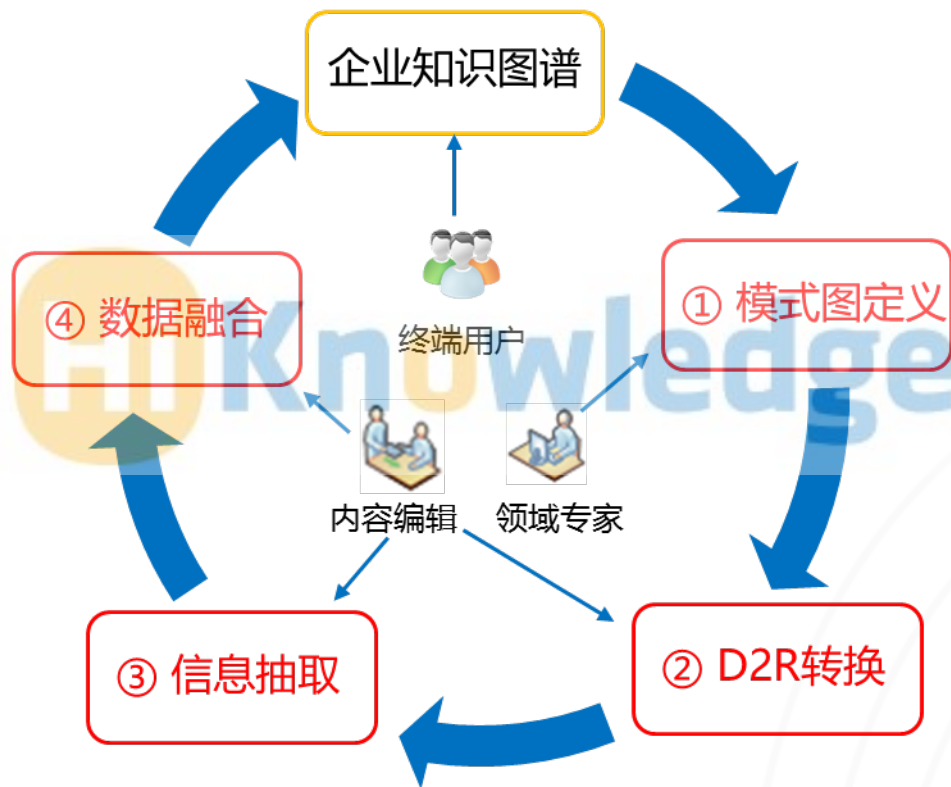
- 商业挑战



- 技术挑战



- 数据驱动的增量式企业知识图谱构建流程



- 根据数据源类型选择数据抽取方式: D2R转换或信息抽取
- 新的数据源触发新一轮迭代

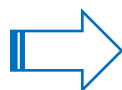
经过两轮迭代构建企业知识图谱

- 第一轮：分别构建基础企业知识图谱、专利知识图谱



中国工商数据

- 40,000,000企业
- 8,000,000诉讼信息
- 60,000,000人
- 1,000,000信用信息



基础企业知识图谱



专利检索及分析

Patent Search and Analysis of SIPO

- 5,000,000专利信息



专利知识图谱

- 第二轮：融合两个知识图谱，并整合其他数据源



中国政府采购网
www.ccgp.gov.cn

- 3,000,000企业招投标信息



互动百科
baike.com



- 上市公司的股票信息



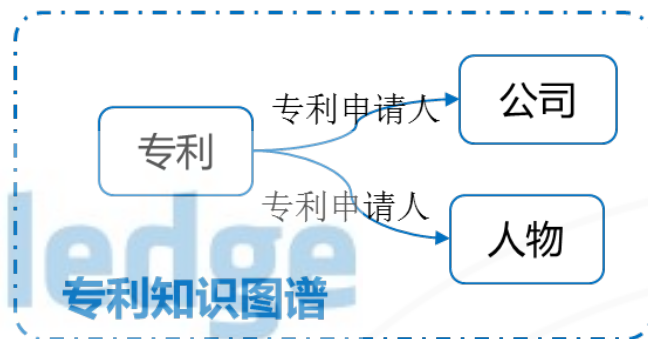
维基百科
自由的百科全书

- 竞争关系
- 并购事件

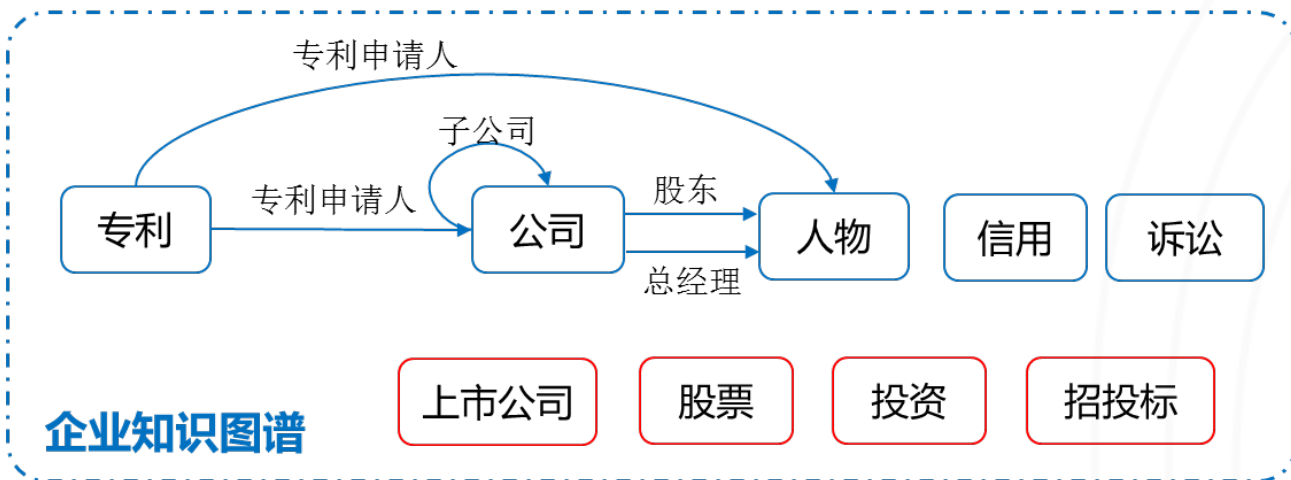
1、模式图定义

- 在迭代构建知识图谱的过程中，动态更新、扩展模式图
- 保证数据质量和模式图的严谨性

第一轮迭代：



第二轮迭代：



2、D2R转换

难点：

- 数据表不规范
 - 同一列包含不同类型的数据
- 元属性映射问题
- 现有D2R工具无法解决复杂映射关系



解决方案：

- 数据表切割
- 基础D2R转换
- 后处理

• 数据表切割

原始数据表

个人信息表

ID card_Number	Name	Postition	Regist_Num	Enterprise_Name	Entry_Time	...
11111xx	熊维平	董事长	100000000035734	中国铝业股份有限公司	2009-05	...

企业信息表

Enterprise_Name	Regist_Num	Industry_Categories	...
中国铝业股份有限公司	100000000035734	3100	...

股票信息表

Stock_Code	Regist_Num	Stock_Price	...
601600	100000000035734	958052.19	...

人物表

ID card_Number	Name	Postition	...
11111xx	熊维平	董事长	...

股票表

Stock_Code	Stock_Price	...
601600	958052.19	...

企业表

Enterprise_Name	Regist_Num	Industry_Categories	...
中国铝业股份有限公司	100000000035734	3100	...

企业-股票表

Stock_Code	Regist_Num	...
601600	100000000035734	...

人物-企业表

ID card_Number	Regist_Num	Entry_Time	...
11111xx	100000000035734	2009-05	...

原子实体表

原子关系表

复杂关系表

复杂实体表

2、D2R转换

- 基础D2R转换

利用D2RQ工具将原子实体表、原子关系表转化成RDF三元组。

人物表

ID card_Number	Name	Postition	...
11111xx	熊维平	董事长	...

股票表

Stock_Code	Stock_Price	...
601600	958052.19	...

企业表

Enterprise_Name	Regist_Num	Industry_Categories	...
中国铝业股份有限公司	100000000035734	3100	...

企业-股票表

Stock_Code	Regist_Num	...
601600	100000000035734	...

原子实体表

原子关系表

D2RQ

表名 → 概念
列名 → 属性
单元格值 → 属性值

- 后处理

对复杂关系表、复杂实体表进行

- 元属性映射
- 属于不同上下位概念的实体映射
- 属于不同类别的同列数据映射

3、信息抽取

难点：

- 多数据源： 专利检索及分析
Patent Search and Analysis of SIPO

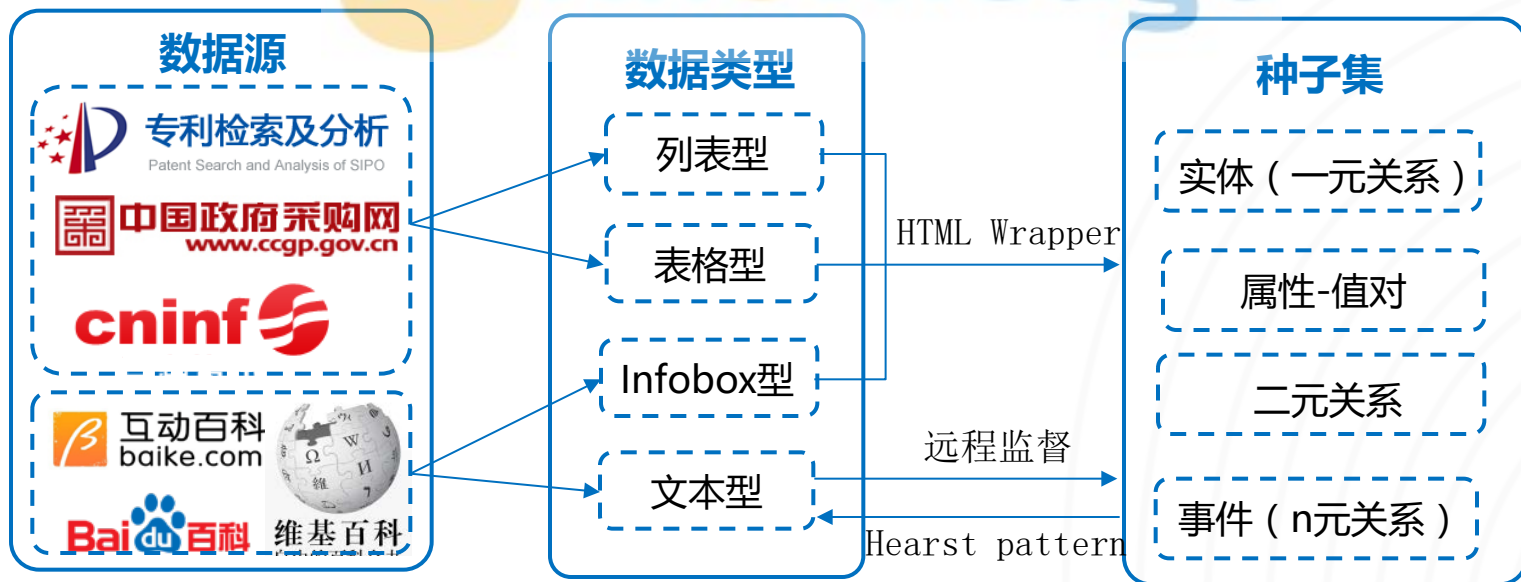


- 目标数据类型多样：

不同类型实体（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



4、数据融合

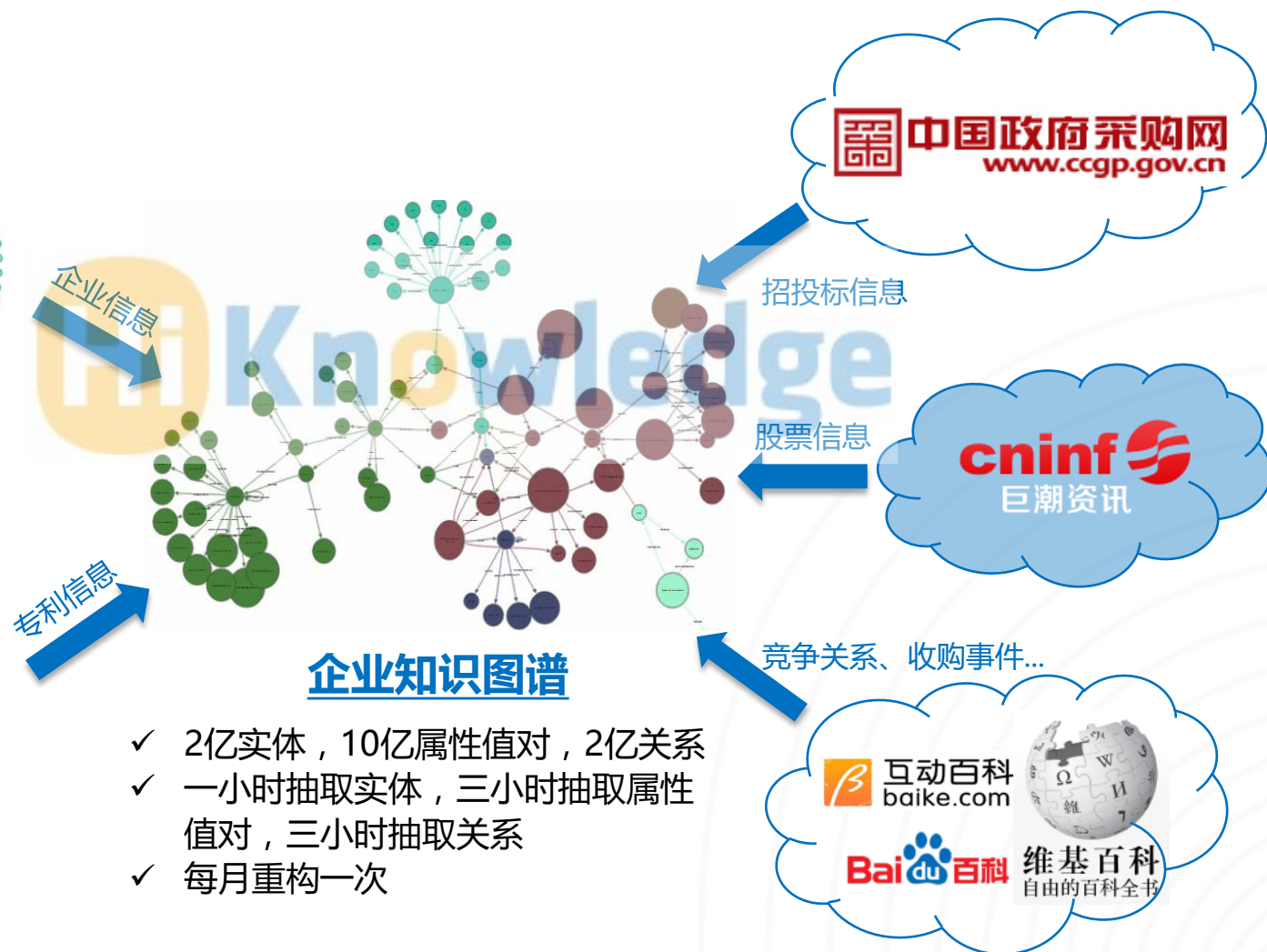
- 公司、人物等实体对齐
- 数据冲突问题



基础企业知识图谱



专利知识图谱



5、存储设计和查询优化

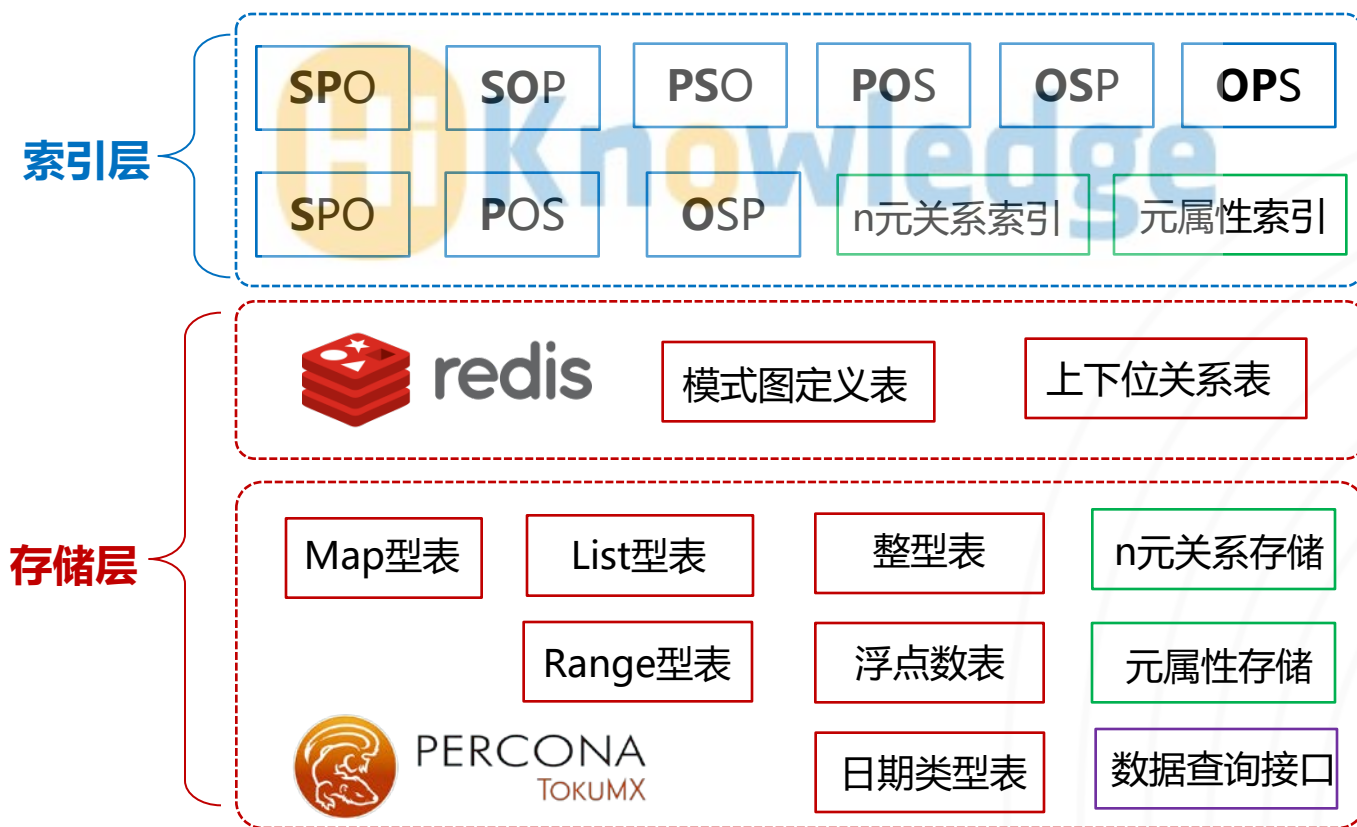
- 现有RDF存储方案对比

存储方案	数据库示例	缺点
基于关系数据库的RDF存储方案	MySQL等关系型数据库	<ul style="list-style-type: none">• 大量自连接操作的开销巨大• RDF灵活性&属性未定查询• 大量数据表• 删除属性代价巨大
Triplestore	Native RDF : Hexastore、RDF-3X 图数据库 : Neo4j	<ul style="list-style-type: none">• 空间开销大• 更新维护代价高• 图查询复杂度高
分布式RDF存储库	商业 : AllegroGraph、Microsoft Trinity 、OpenLink Virtuoso、BigOWLIM 开源 : Bigdata	<ul style="list-style-type: none">• RDF图分割和维护的复杂度高

注：来自tutorial--大规模RDF图数据的存储、查询、检索和推理，by 天津大学 王鑫

5、存储设计和查询优化

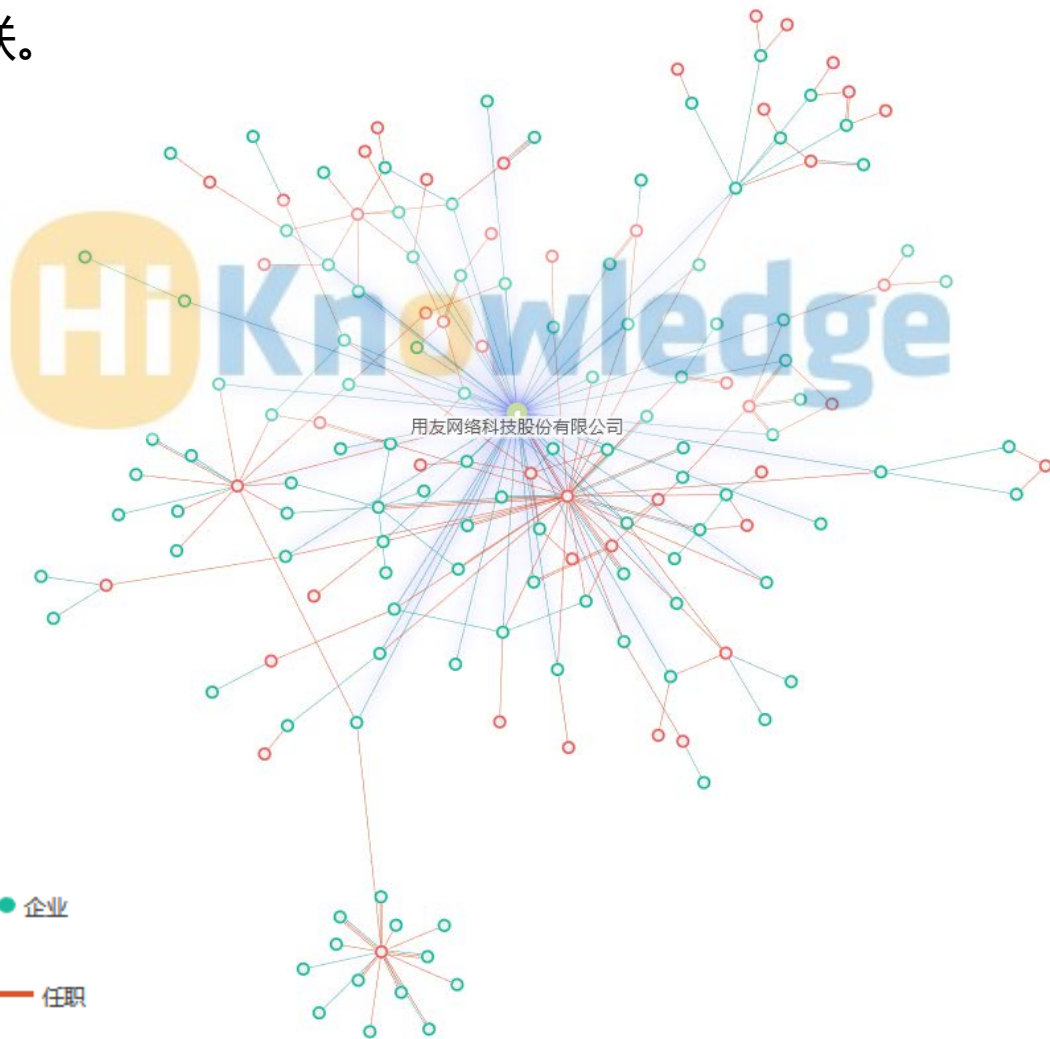
- 基于TokuMX存储多类型数据，建立九重索引，并提供不同类型数据的查询接口；
- 基于数据类型，对数据分表存储；
- Redis存储频繁访问的数据；
- 设计支持高效查询元属性和n元关系的存储结构。





社交图谱
Social Graph

- 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向欧盟和成交额美好一个网络关系图，直观立体展现企业关联。





最终控制人
Ultimate Controller

- 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。

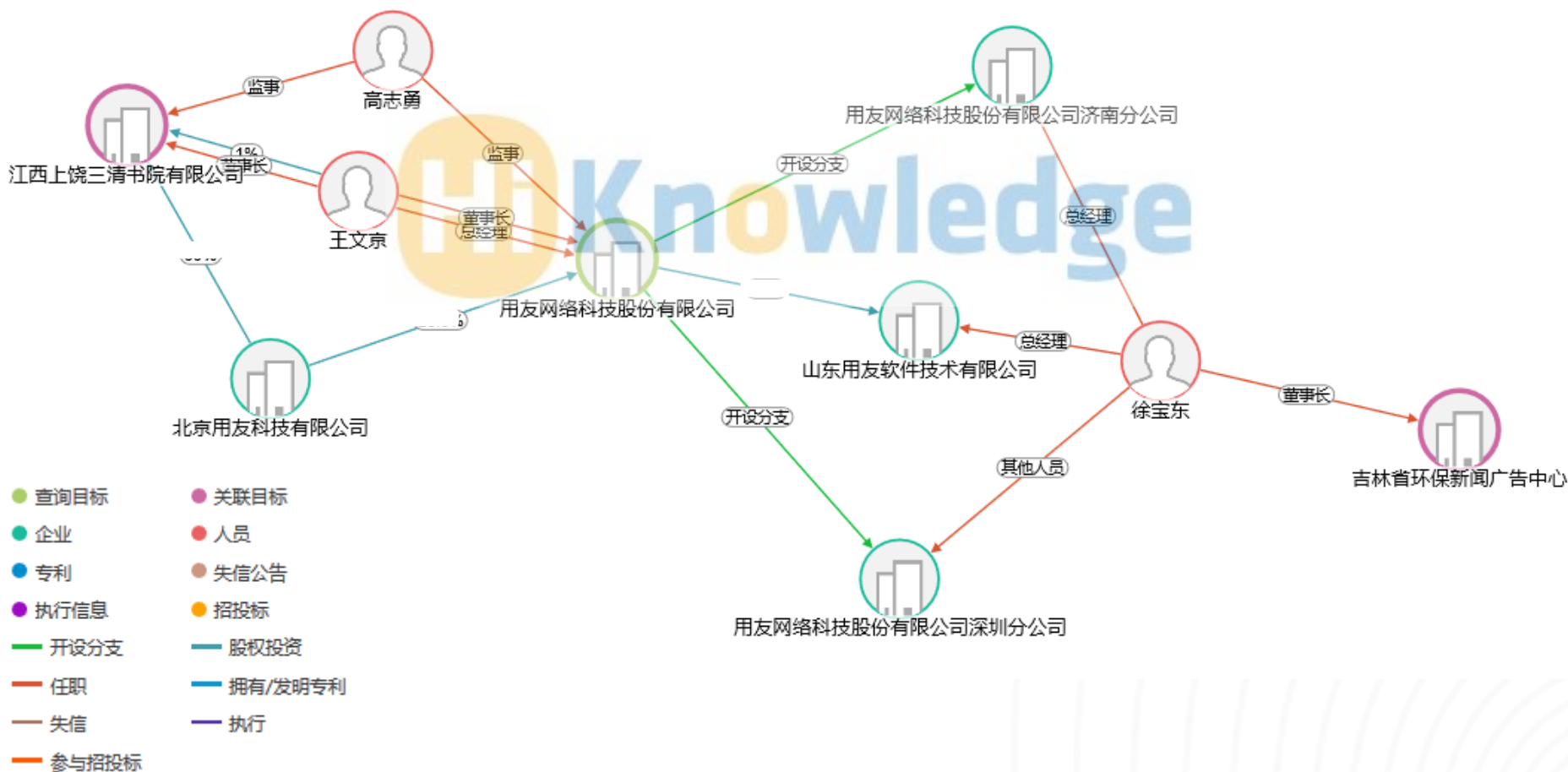


企业知识图谱的应用场景



路径发现
Shortest Path

- 在基于股权、任职、专利、招投标、涉诉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。



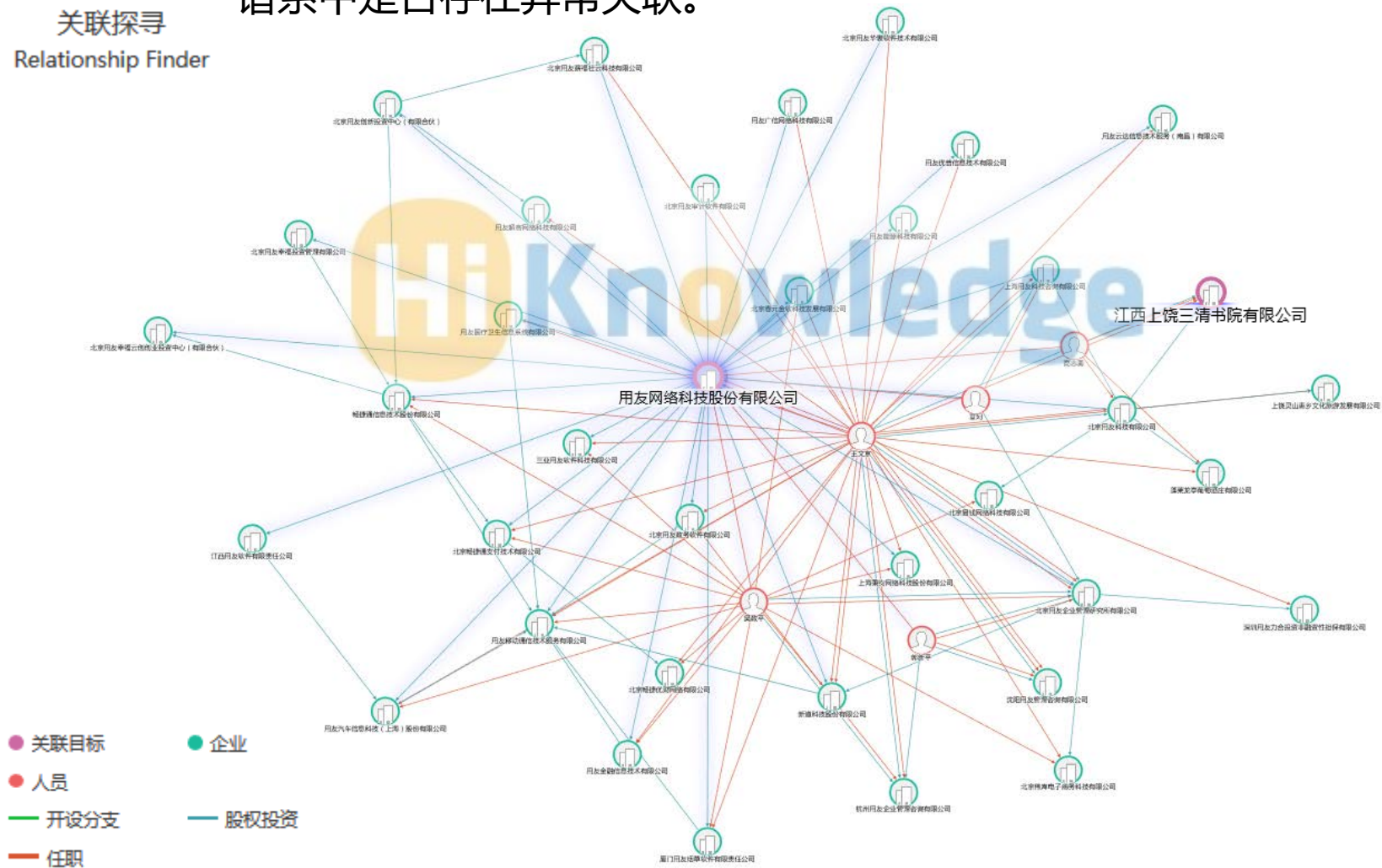
企业知识图谱的应用场景



关联探寻

Relationship Finder

- 查询企业之间的关系网络，探寻企业间的关联关系，挖掘目标企业谱系中是否存在异常关联。

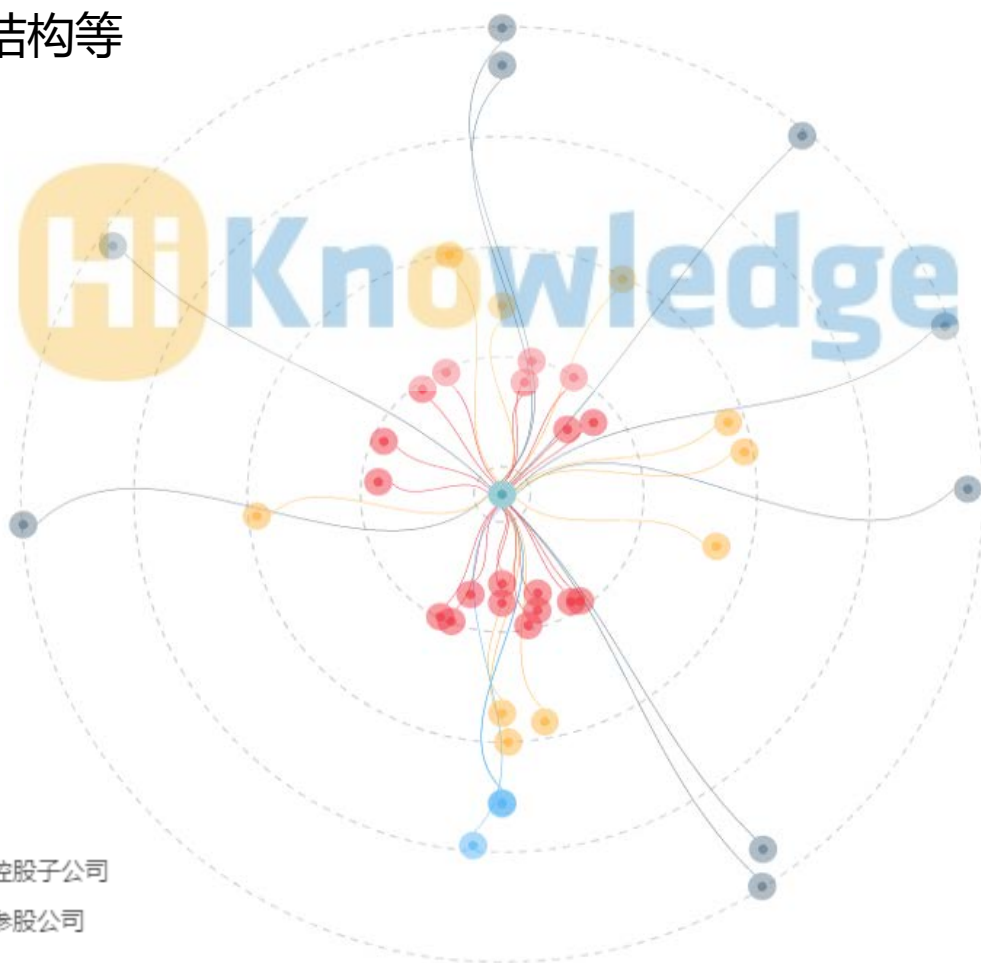




战略发展

Strategic Layout

- 以“信任圈”的展现形式将目标企业的对外投资企业从股权上加以区分，探寻其全资、控股、合营、参股的股权结构及发展战略,从而理解竞争对手和行业企业的真实战略,发现投资行业结构、区域结构、风险结构、年龄结构等



● 全资子公司

● 控股子公司

● 合营/联营

● 参股公司

- 现有大数据应用面临的挑战
- 行业知识图谱解决方案
- 案例：去也知识图谱的构建与应用

■ 总结和展望





从石墨到钻石，只是结构的重塑

同物质因组织结构不同产生巨大价值差异

HiKnowledge

致力于提供行业知识图谱构建及应用解决方案


现已有全国企业知识图谱，中外创投知识图谱，海洋鱼类知识图谱，全国专利知识图谱等行业应用

欢迎各种形式的数据合作、技术合作、业务合作！



 丁军

 13524277070

 dingjun@hiekn.com

 www.hiekn.com