Geekbang. 极客邦科技 InfoQ

QCon 全球软件开发大会【北京站】2016

大数据和人工智能在P2P中的应用 – 风控

李文哲， 普惠金融（爱钱进）
liwenzhe@puhuifinance.com
wechat: liwenzhe595675

Intemational Software Development Conference

关于我

爱钱进 IQIANJIN.COM 靠谱的互联网金融平台

普惠金融 PUHUI FINANCE

amazon.com

Goldman Sachs

USC University of Southern California

Universiteit van Amsterdam

南開大學 Nankai University

TEXAS A&M UNIVERSITY

关注：人工智能、机器学习、深度学习

普惠金融&爱钱进

国内 TOP10 P2P 平台

- 互联网金融（P2P）公司，总部在北京
- 创立于2013年7月份
- 2014年12月 5000万美金A轮
- 5500+ 员工（包括线下销售），100+ 线下门店

数据驱动策略（DDS）

- **大数据驱动决策**
  - 解决方案的决策需要数据支持，而非仅通过数据知道问题或*insight*
  - 模型不是万能的，模型要和策略相结合
  - 数据不是万能的，数据分析要有目的性

如何建立数据驱动策略的能力？

数据

- 公司内部积累的数据
- 客户散落在各个地方的非结构化数据
- 其他机构积累的数据

数据处理能力

- 机器学习能力
- 模型能力
- 数据抓取能力

IT架构

- 系统性的存储数据
- 抓取客户散落在互联网的数据
- 存储海量数据

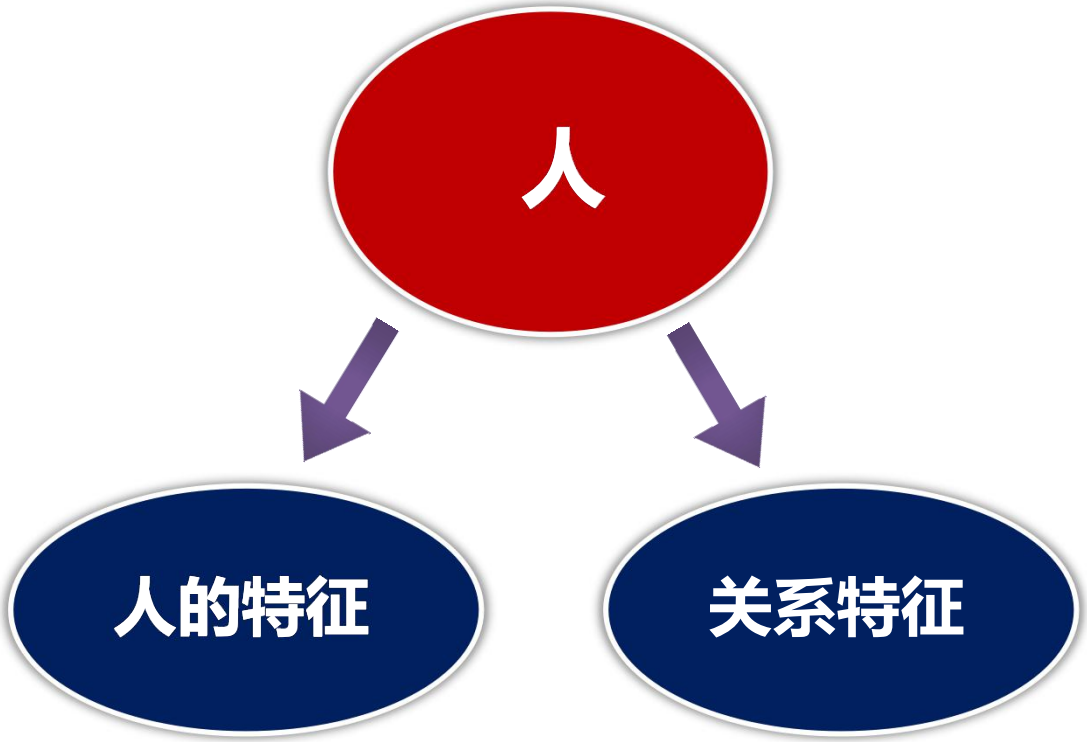
大数据风控是核心

为什么需要大数据风控？

- 国内的征信体系不完善
- 服务的人群特性
- 很多网上的痕迹
- 更高效的审核

风控的核心是“人”

以人为“中心”



- 年龄
- 工资
- 工作单位
- ....

- 他/她的朋友有没有触碰黑名单？
- 他/她同事的朋友.....

传统风控 VS 大数据风控

	人的特征	关系特征
传统风控	基于人的基本信息比如年龄、工资、工作单位等，这种特征一般少于 < 50	基本上不包包括关系特征
大数据风控	除了上面提到的基本信息、可以包括从行为数据里提取出来的特征、还有更细化的特征。特征数量 > 1000	包括一度、二度甚至高维度关系的特征

知识图谱可以用来有效地分析关系的特征

什么是知识图谱？

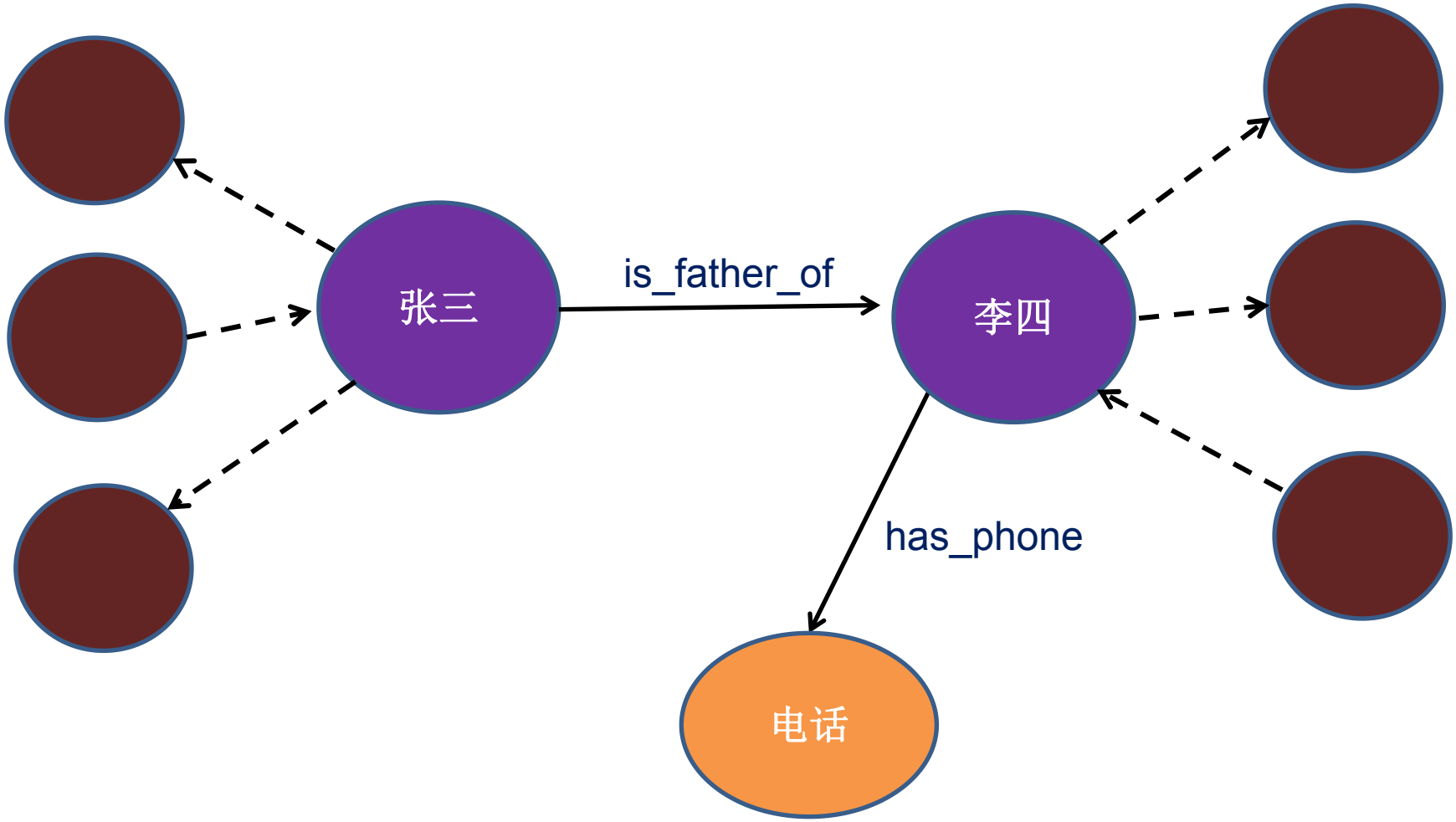
- 它是语义网络
- 一种基于图的数据结构，由节点(Point)和边(Edge)组成。每个节点表示“实体”，每条边为实体与实体之间的“关系”。通俗地讲，知识图谱就是把所有不同种类的信息（Heterogeneous Information）连接在一起而得到的一个关系网络。

知识图谱 – 搜索优化

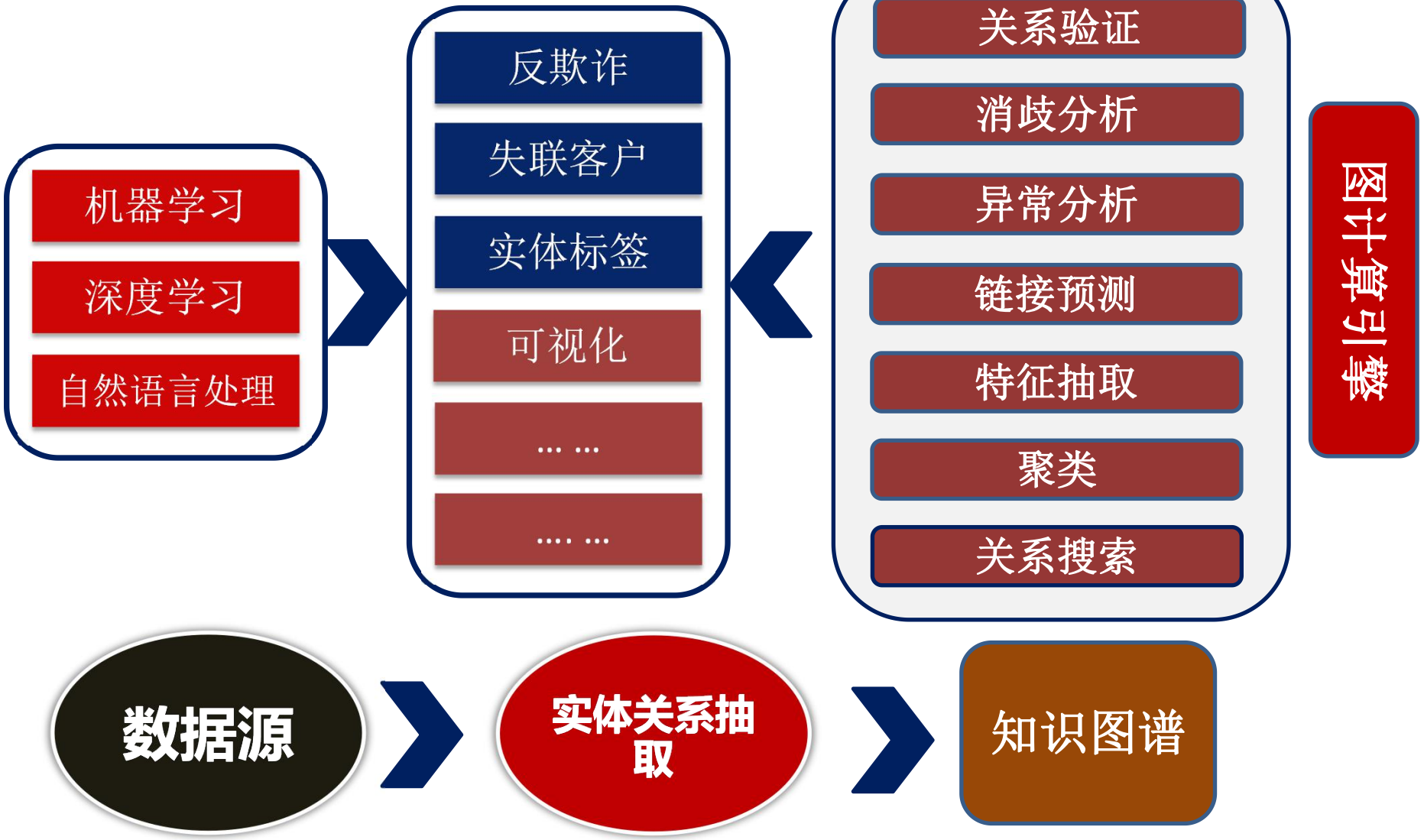
问答系统

推荐商品

知识图谱的表示： RDF, 属性图



知识图谱生态系统



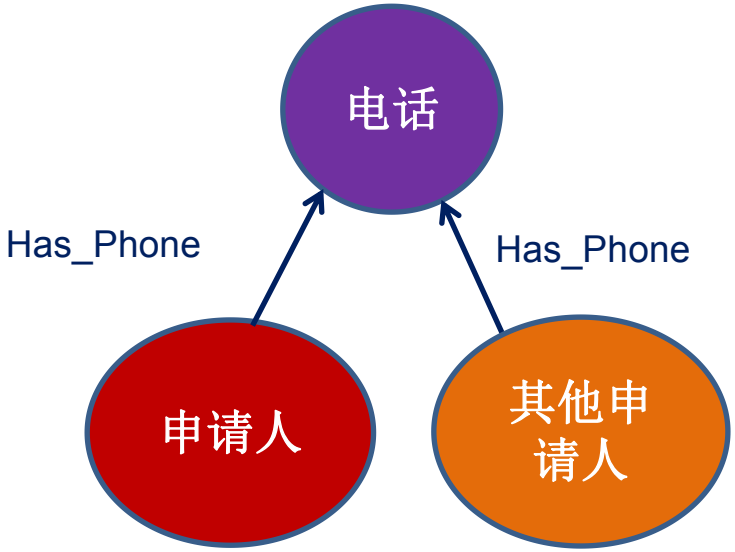
金融知识图谱

- ~10+ 种实体类型
- ~50+ 关系类型
- 上亿个实体和关系

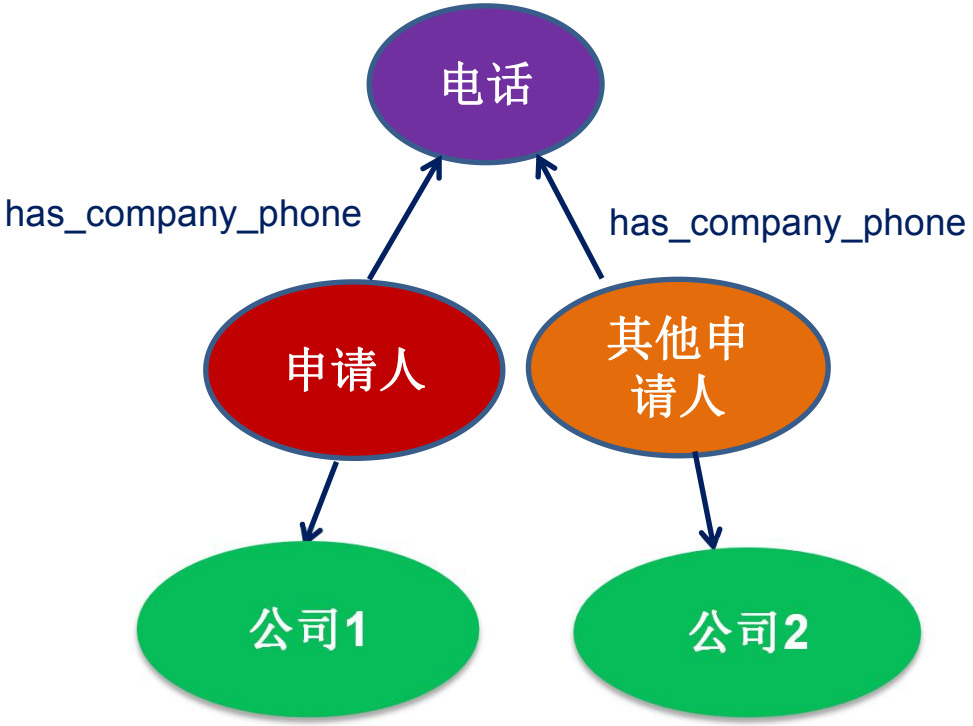
知识图谱案例

1. 反欺诈
2. 失联客户管理
3. 给实体打标签

反欺诈 – 不一致性验证

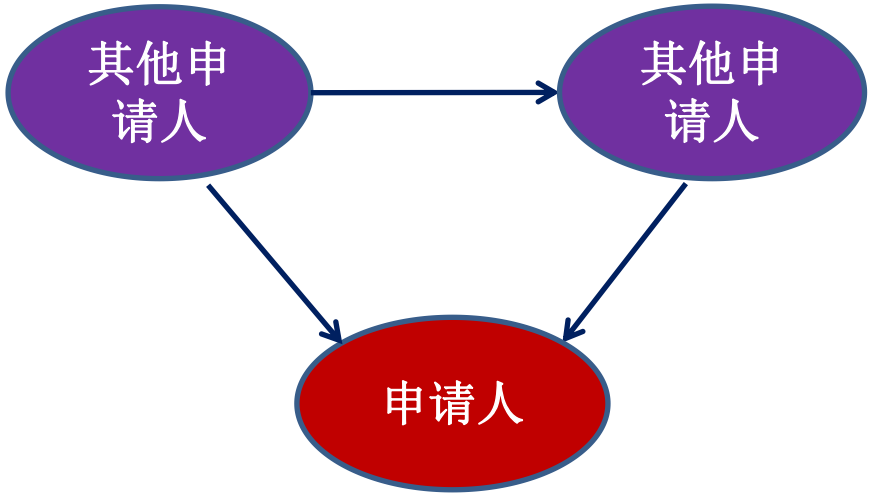


申请人与其他申请人拥有同样的电话

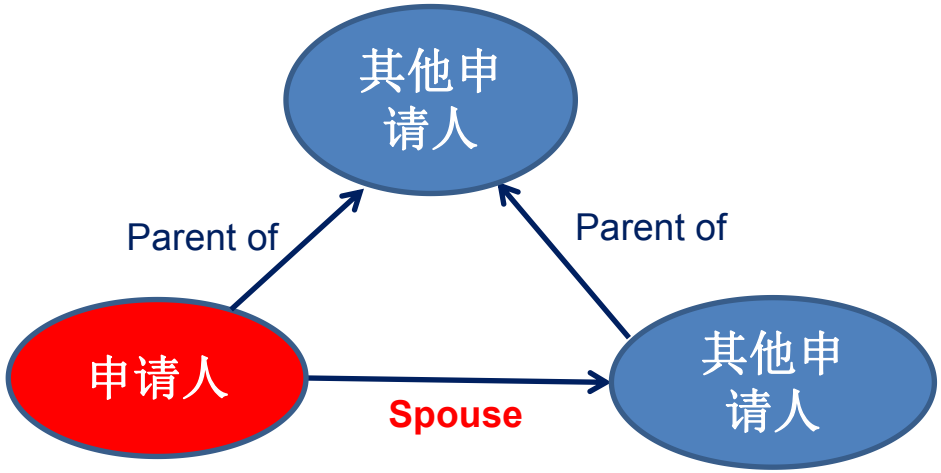


申请人与其他申请人填写了同样的公司电话号，但填写了不同的公司名

反欺诈 – 三角关系

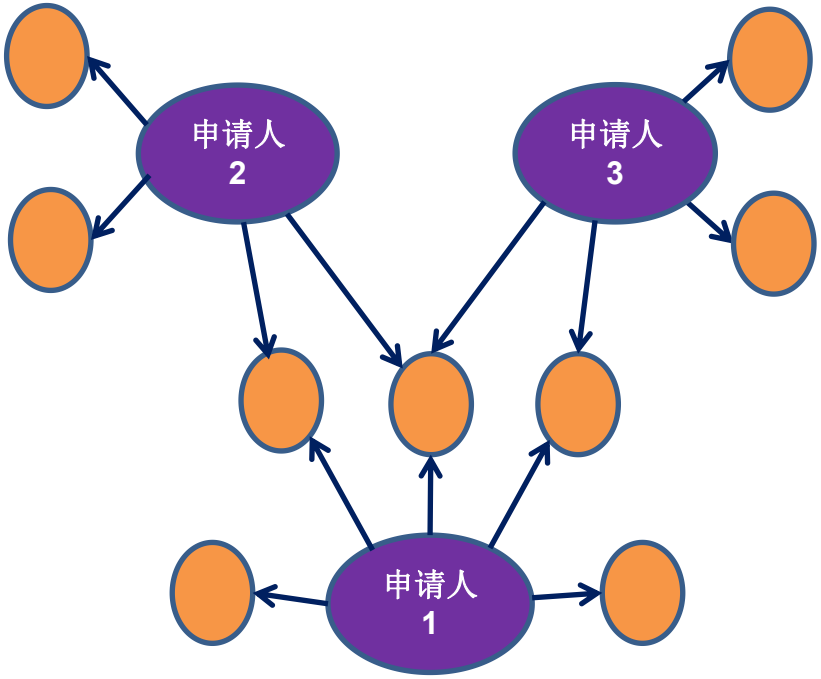


三角关系

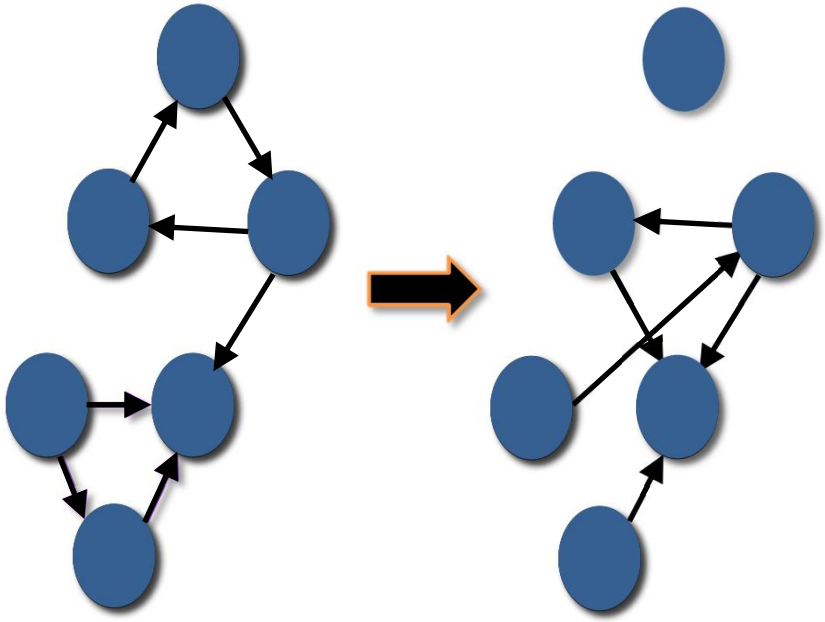


关系的不一致性

反欺诈 – 其他风险

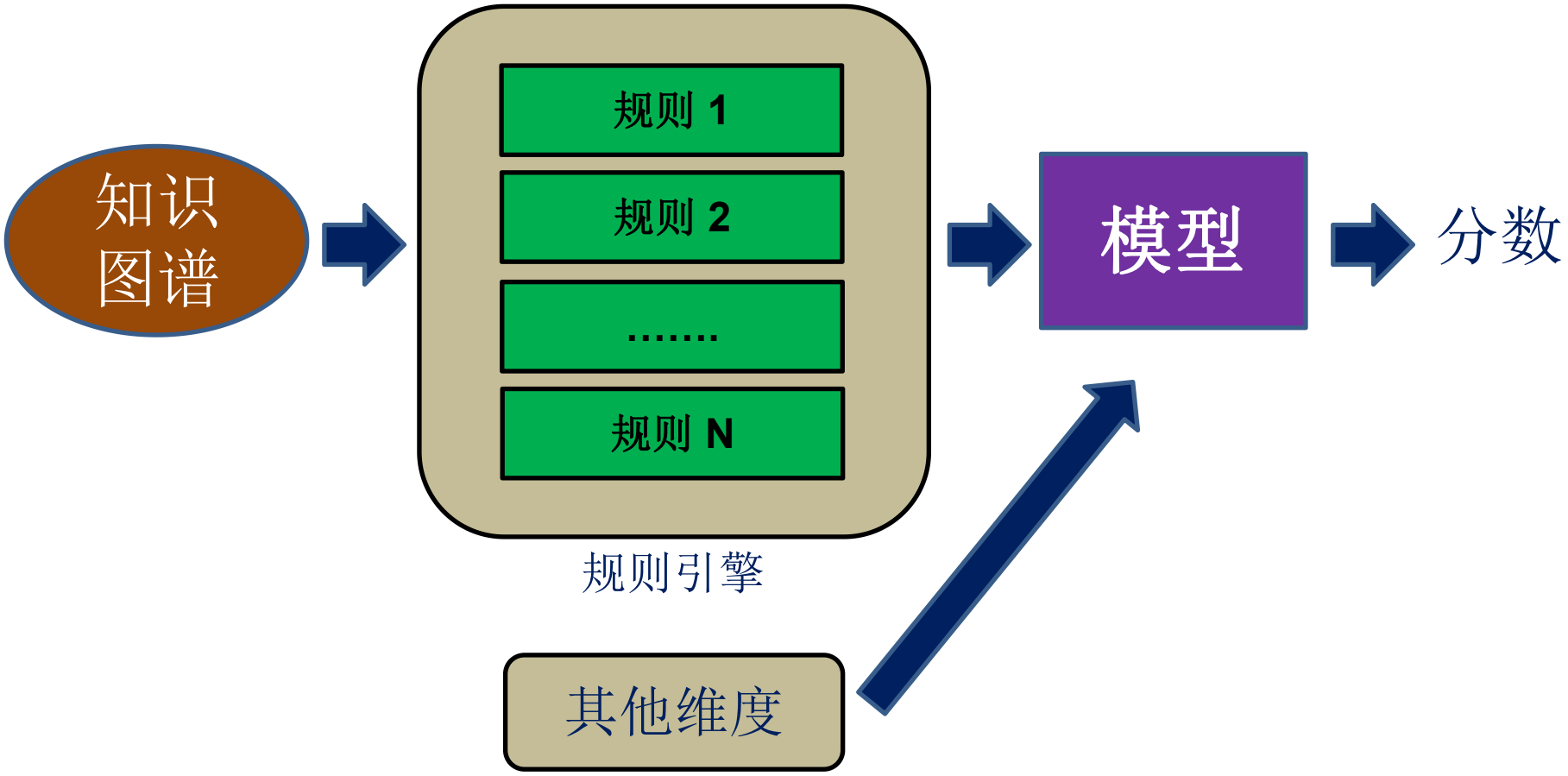


共享很多信息

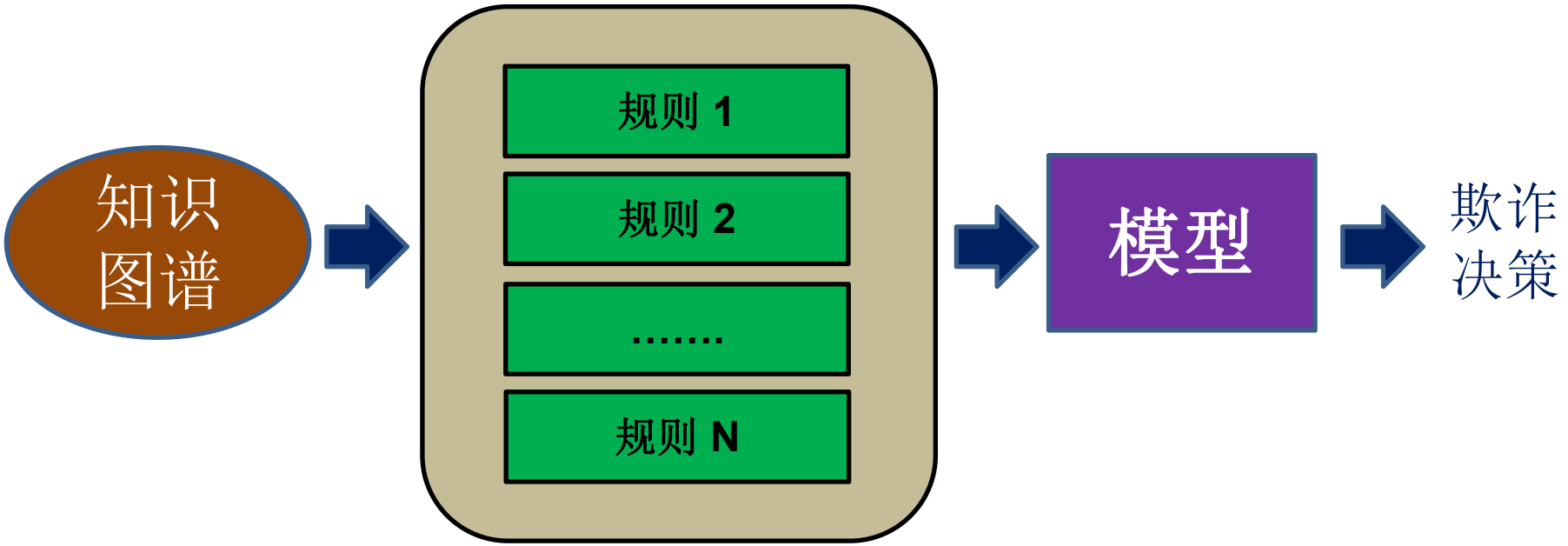


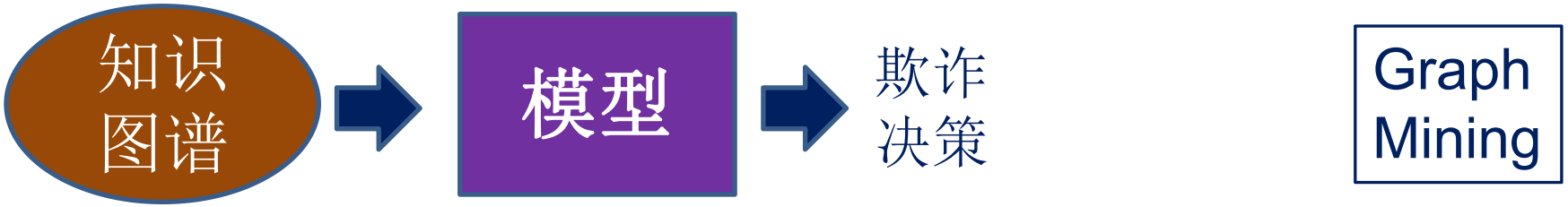
图的结构短时间内变化比较大

反欺诈分数



反欺诈分数 – what is next?





知识图谱案例

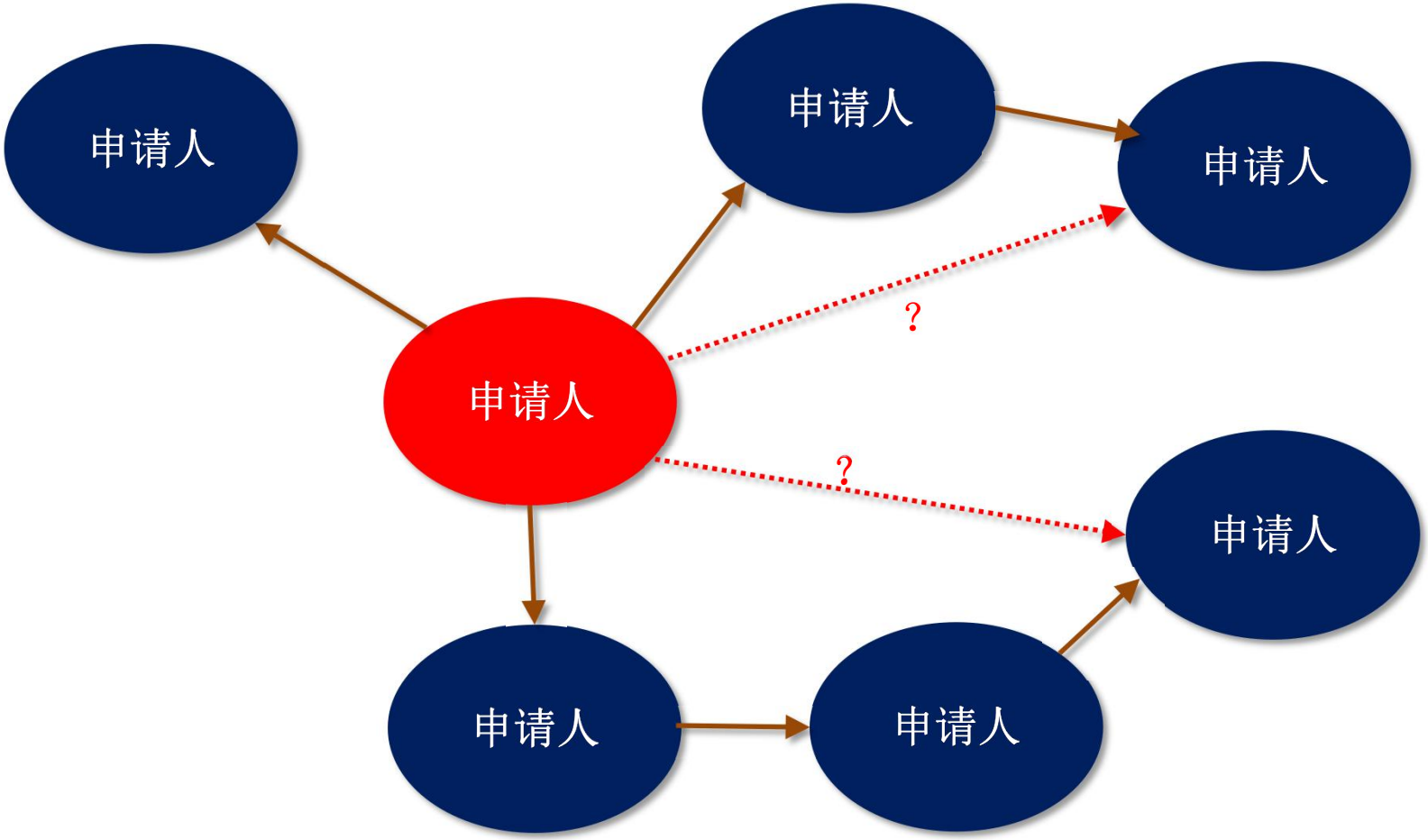
1. 反欺诈
2. 失联客户管理
3. 给实体打标签

什么是失联客户？

借款人借钱之后失联，怎么办？

挖掘更多的联系方式

链接预测（link prediction）



$$p\left(r_{ij}^{l}=1\middle|\phi(e_i),\phi(e_j),\phi(nei(e_i)),\phi(nei(e_j))\right)$$

知识图谱案例

1. 反欺诈
2. 失联客户管理
3. 给实体打标签

给每个实体打标签

怎么根据公网爬取的数据、对每个实体打个标签 （比如黑名单？）

实体标签 – 搜索电话号码

Anonymous +3

Received call from lady stating i will have charges filed against me for fraud and i needed to return call immediately to resolve matter before i end up served with legal action taken against me. This woman called from a boise Idaho based numbe; 208-392-1904 but instructed on my vm to return her call to above # (979-256-4754) this lady did have knowledge of some of my personal identifying info such as last 4 of ss # so beware of these scammers! Also was able to (possibly) trace these scammers to anthem security breach that recently took place. BEWARE...

The caller was identified as Cynthia erickson

Reported 1st, Aug. 2015.

实体标签 – 搜索邮箱

冒用"中交一航局第五工程有限公司"（骗子邮箱：zjhwren@foxmail.com，骗子手机：18674046551）

2013年7月20日更新骗子名单：

冒用"远大医药（中国）有限公司"（骗子手机：15623724794）

冒用"浙江大经建设集团股份有限公司"（骗子邮箱：hr_zjdjc@163.com，骗子手机：15557167017）

实体标签

- 电话号 Email
- QQ
- 其他信息

关键词

- 百度
- 360搜索
- 其他公网

搜索引擎 & 公网

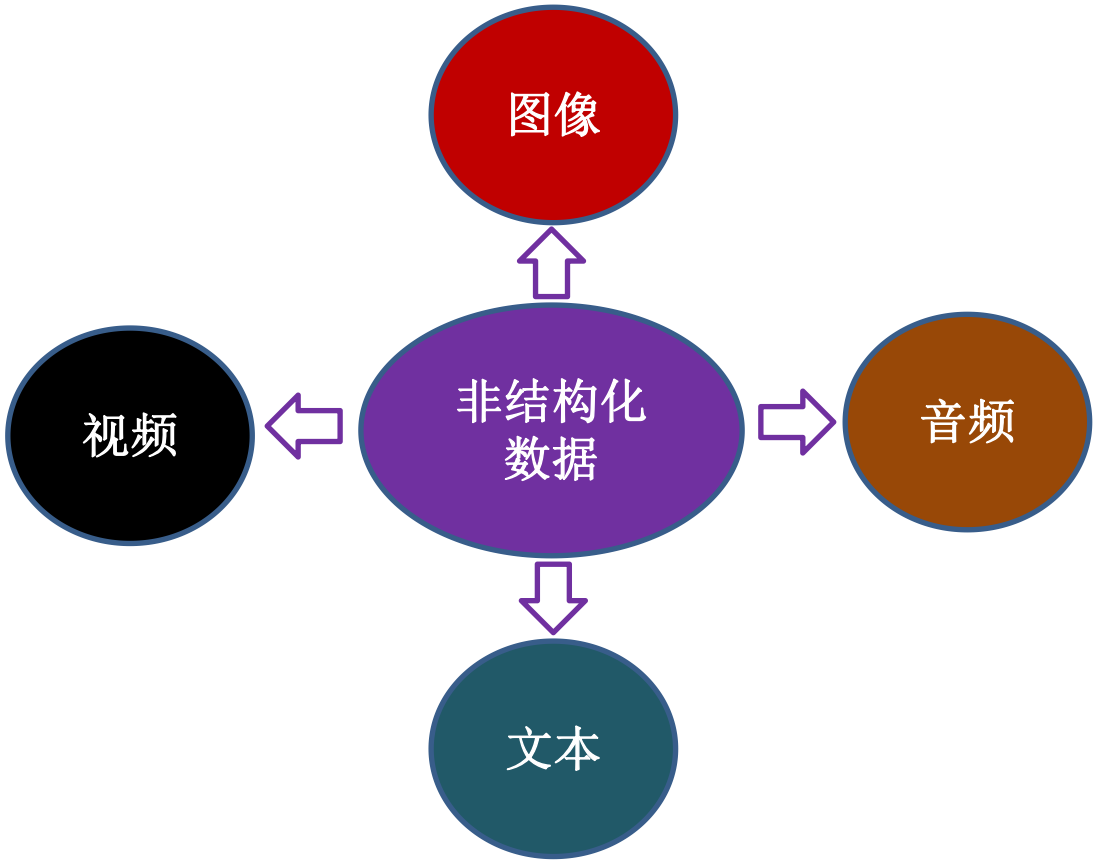
标签

$p(\text{标签}|doc_1, \ldots, doc_N)$

自然语言处理

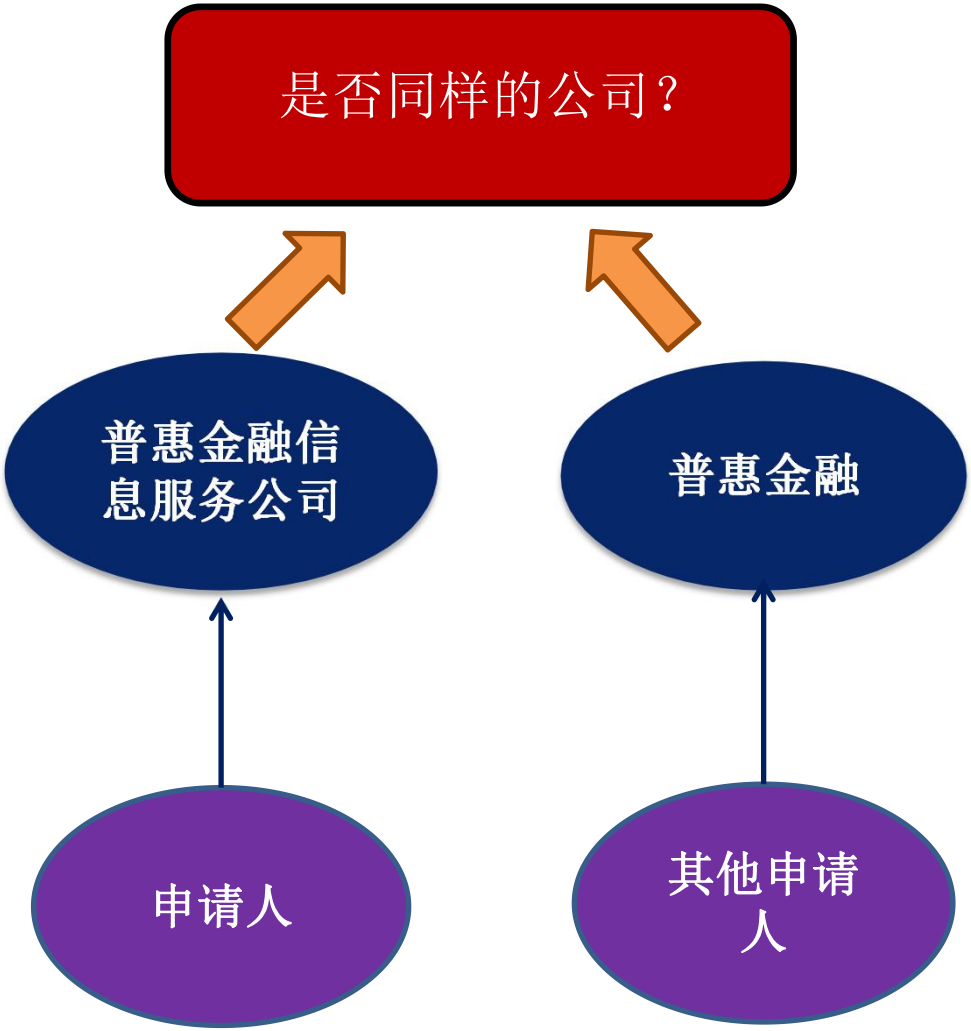
- 主题分析
- 依存分析
- 标签传播
- ... ...

挑战 – 非结构化数据



- 机器学习
- 自然语言处理
- 统计计算

挑战 – 消歧分析



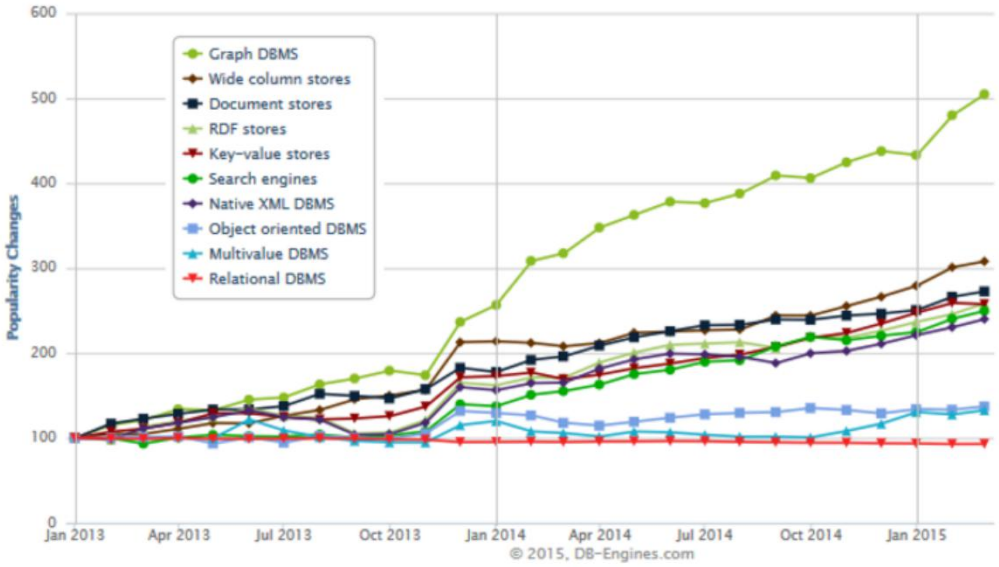
$$sim(e_i, e_j) > threshold$$

其他挑战

链接预测

大数据、小样本

讨论 - 存储知识图谱



Ranking	DBMS
21	Neo4j (Graph Database)
32	MarkLogic (XML)
42	Titan (Graph Database)
46	OrientDB (Graph Database)
61	Virtuoso (RDF)
80	Jena (RDF)
88	Sesmae (RDF)
90	ArangoDB (GraphDatabase)
120	AllegroGraph (RDF)

知识图谱提供了哪些好处？

- 更方便地整合和管理不同种类的数据源
- 方便地做关系的分析
- 在分析关系特征的效率上，有好几个数量级的提升
- 提供实时性服务
- ……

设计知识图谱时需要考虑的点

- 紧密结合业务逻辑
- 知识图谱需要轻便，只存跟关系推理有关的信息
- 尽量不要存属性
- 线上、先下分离
- 内存消耗比较大
- ……

谢谢！

爱钱进 IQIANJIN.COM
靠谱的互联网金融平台
普惠金融 PUHUI FINANCE

wechat: liwenzhe595675

加入我们

- 全栈架构师/运维架构师
- 安全架构师/算法架构师
- Java架构师
- Hadoop开发工程师
- 高级前端开发工程师
- 数据挖掘/算法工程师
- 数据科学家

邮箱： zhaopin@puhuifinance.com