

第六讲：知识工程与知识图谱

目录

- 6.1 知识表示与知识图谱
- 6.2 知识图谱的生命周期
- 6.3 知识图谱应用

2

6.1 知识表示与知识图谱

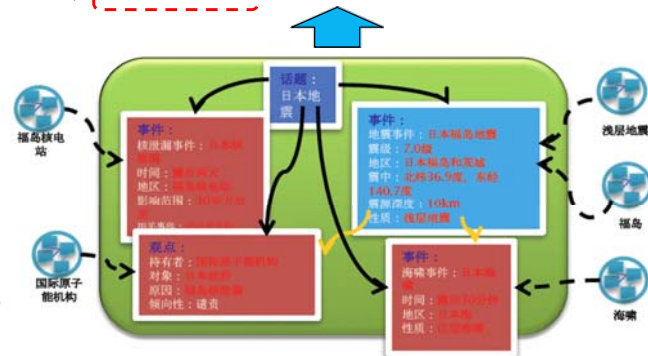
人类理解信息需要知识的支撑



4

人类理解信息需要知识的支撑

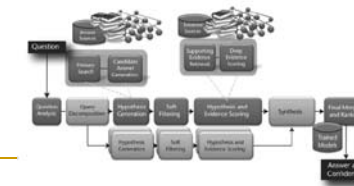
2011年4月11日17点16分，日本东北部的福岛和茨城地区发生里氏7.0级强烈地震（震中北纬36.9度、东经140.7度，即福岛西南30公里左右的地方，震源深度10公里，属于浅层地震）。当局已经发布海啸预警。震后约30分钟后在日本海地区发生巨型海啸，同时造成福岛核电站出现核泄漏。震后第十天，国际原子能机构对于日本政府反应迟钝进行了谴责。



5

机器智能化也离不开知识

- 沃森(IBM Watson): 2011年，IBM研发的超级计算机“沃森”在美国知识竞赛节目《危险边缘Jeopardy!》中上演“人机问答大战”，战胜人类选手Ken和Brad



辅助医疗



金融辅助决策



企业服务

6

机器智能化也离不开知识



蒲熠星，南京大学学生，《一站到底》五年巅峰会的获胜者，获“战神之神”称号

在IBM沃森与搜狗旺仔中，知识图谱起到了很关键的作用

7

知识工程 (Knowledge Engineering)

- 知识工程是运用信息技术手段高效率、大容量的获得并利用知识与信息的技术
- 知识工程主要包括：知识表示、知识获取和知识应用



费根鲍姆
美国斯坦福大学教授，
图灵奖获得者

Knowledge is the Power in AI
1977
第一个专家系统DENDRAL
1968



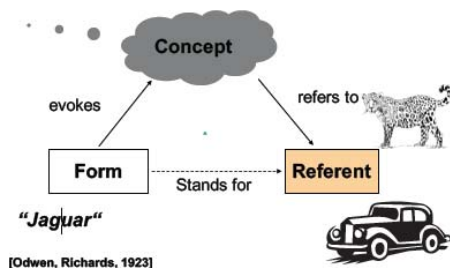
陆汝钤
中科院数学院研究员，
中国科学院院士

知识工程语言TUILI
1982
专家系统开发环‘天马’
1987

8

知识表示

- 知识表示是对**事物本身的替代**，使我们可以通过思考而不是行动来确定事物的来龙去脉、前因后果
 - 吃饭会饱（不需要每次都吃来确认）
 - 冰岛是一个国家（虽然没去过）
- 知识表示是一个本体约定 (ontological commitment) 集合，解决如下问题：**我们该用什么术语 (terms) 来思考这个世界？**



9

知识表示的主要方法

- 自然语言(对人而言最方便的表示和传播手段)
- 符号表示方法
 - 谓词逻辑
 - 语义网(Semantic Net)
 - 框架
 - 脚本
 - 语义网(Semantic Web)
 - 知识图谱
- 其他方法（如分布式表示）

10

谓词逻辑

- **谓词(Predicate)**是一个词组，用于描述**对象的属性**，或是不同对象之间的关系
- **命题(Proposition)**是谓词加应用于该谓词的一个term元组，表示一个属性或objects之间的关系
 - Brother(John, Fred)
 - Left-of(Square1, Square2)
 - GreaterThan(plus(1,1), plus(0,1))
 - 命题的语义是在特定interpretation中的真假值
- **复杂命题**可以通过逻辑连词($\neg \vee \wedge \Rightarrow \Leftrightarrow$)来构建
 - Owns(John, Car1) \vee Owns(Fred, Car1)
 - Sold(John, Car1, Fred) \Rightarrow Owns(John, Car1)

11

谓词逻辑--量词

- 通过量词机制，允许声明关于一个集合的对象的知识，而不需要一个个枚举它们
- **全称量词(Universal quantifier): \forall**
 - $\forall x \text{ Loves}(x, \text{FOPC})$
 - $\forall x \text{ Whale}(x) \Rightarrow \text{Mammal}(x)$
- **存在量词(Existential quantifier): \exists**
 - $\exists x \text{ Loves}(x, \text{FOPC})$
 - $\exists x (\text{Cat}(x) \wedge \text{Color}(x, \text{Black}) \wedge \text{Owns}(\text{Mary}, x))$

12

语义网(Semantic Net)

- 启发Idea:
 - 人脑记忆的一个重要特征是人脑中不同信息片段之间高度连接
 - 高度相关的概念能够比不太相关的概念更快的回忆起来

- 语义网是一个通过语义关系连接的概念网络
- 语义网将知识表示为相互连接的节点和边模式
 - 节点表示实体、属性、事件、值等
 - 边表示对象之间的语义关系

M.Quillian (1968). Semantic Memory, in M. Minsky (ed.), Semantic Information Processing, pp 227-270, MIT Press

J. F. Sowa (1987). Semantic Networks, in Stuart C Shapiro. Encyclopedia of Artificial Intelligence. Retrieved 2008-04-29.

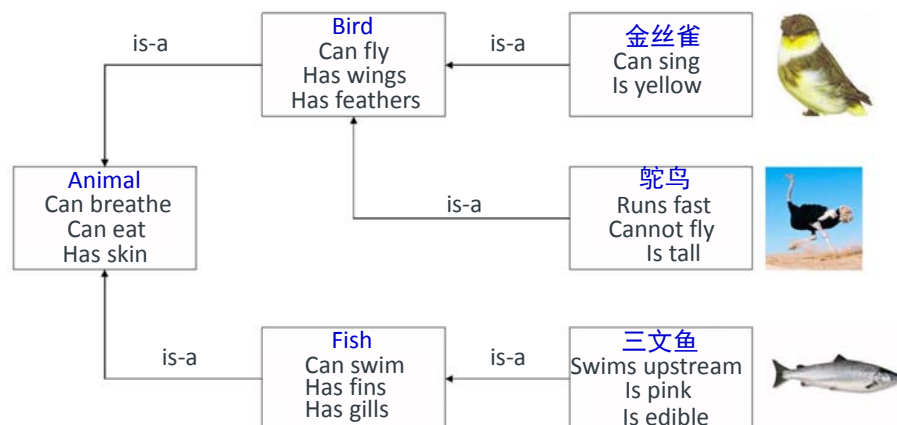
13

通常使用的语义关系

- Is-A
- Instance-Of 实例
- Part-Of
- Modifiers: on, down, up, bottom, moveto, ...
- 领域特定的关系类型
 - 医疗: 症状、治疗、病因...
 - 金融: 收购、持有、母公司...

14

语义网示例



15

框架理论

- 框架理论认为人们对现实世界中各种事物的认识都是以一种类似于框架的结构存储在记忆当中的，当面临一个新事物时，就从记忆找出一个适合的框架，并根据实际情况对其细节加以修改补充，从而形成对当前事物的认识

Proposed in 1968 by Marvin Minsky <http://web.media.mit.edu/~minsky>

16

框架表示

- 知识通过Frame来表示，**每一个Frame表示一种典型的知识**
- Frame包含**槽名(slot names)**和**槽值(slot fillers)**
 - 一个Frame的slot集合能够表示与该框架相关的对象
 - Slot可以指向其他的Frame, Procedure, Slot
- 两类Frame
 - **类Frame**: 类似于面向对象编程里面的Class
 - **实例Frame**: 类似于面向对象编程里面的Object
 - Slots 类似于OO里面的variables/methods
- **不同的Frame通常被组织成一个层次体系结构**
 - Instance Frame - *instance_of*-> Class Frame
 - Class Frame - *subclass_of*-> Class Frame

17

框架表示示例

DOG	COLLIE (牧羊犬)
Fixed	Fixed
legs: 4	breed of: Dog
	type: sheepdog
Default	Default
diet: 肉食	size: 65cm
sound: bark	
Variable	Variable
size:	colour:
colour:	

子类可以从父类继承属性和默认属性值

18

脚本定义

- 脚本与框架类似，由**一组槽组成**，用来表示特定领域内一组事件的发生序列，包含了一组**紧密相关的动作及改变状态的框架**[Winston, 1992]
- 一个脚本是一个描述**特定上下文中的原型事件序列**的结构化表示[Luger, Stubblefield, 1998]

19

脚本的组成元素

- **进入条件**
 - 给出在脚本中所描述事件的前提条件
- **角色**
 - 是一些用来表示在脚本所描述事件中可能出现的有关人物的槽
- **道具**
 - 是一些用来表示在脚本所描述事件中可能出现的有关物体的槽
- **场景**
 - 用来描述事件发生的真实顺序。一个事件可以由多个场景组成，而每个场景又可以是其他的脚本
- **结局**
 - 给出在脚本所描述事件发生以后所产生的结果

20

脚本示例

下面以夏克的“餐厅”脚本为例来说明各个部分的组成。

(1) 进入条件: ① 顾客饿了, 需要进餐; ② 顾客有足够的钱。

(2) 角色: 顾客, 服务员, 厨师, 老板。

(3) 道具: 食品, 桌子, 菜单, 钱。

(4) 场景:

场景1: 进入—— ① 顾客进入餐厅; ② 寻找桌子; ③ 在桌子旁坐下。

场景2: 点菜—— ① 服务员给顾客菜单; ② 顾客点菜; ③ 顾客把菜单还给服务员; ④ 顾客等待服务员送菜。

场景3: 等待—— ① 服务员告诉厨师顾客所点的菜; ② 厨师做菜, 顾客等待。

场景4: 吃饭—— ① 厨师把做好的菜给服务员; ② 服务员把菜送给顾客;

③ 顾客吃菜。

场景5: 离开—— ① 服务员拿来账单; ② 顾客付钱给服务员;

③ 顾客离开餐厅。

(5) 结果: ① 顾客吃了饭, 不饿了; ② 顾客花了钱; ③ 老板赚了钱;

④ 餐厅食品少了。

21

数据万维网

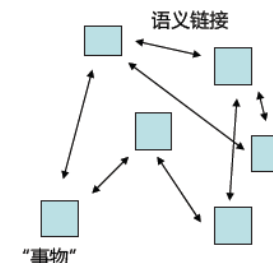
■ 使万维网成为全球开放的知识共享平台

■ 使用语义网(Semantic Web)技术

- 在Web上发布结构化数据
- 在不同数据源中的数据之间建立连接

■ 特征

- Web上的事物拥有唯一的URI
- 事物之间有链接关联(如人物、地点、事件、建筑物)
- 事物之间链接显式存在并拥有类型
- Web上数据的结构显式存在



22

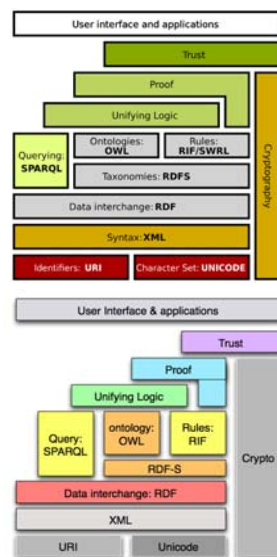
语义网信息描述语言

■ 语义网(Semantic Web)提供了一套为描述数据而设计的表示语言和工具, 用于形式化地描述一个知识领域内的概念、术语和关系

■ HTML描述文档和文档之间的链接

■ RDF, RDFS, OWL和XML能够描述事物和事物之间的关系, 如人和会议、飞机和飞机零件

- Use URIs as names for things
- When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)



23

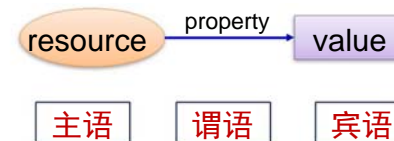
RDF

(Resource Description Framework)即资源描述框架

■ RDF是一种表述对象 (web resources) 和对象之间关系的简单语言

■ 使用(subject, predicate, object)三元组的形式来陈述关于对象的知识, 也就是两个对象之间带类别的链接

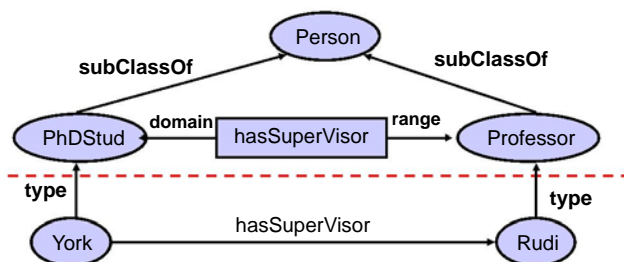
■ RDF是一个通用模型, 可以用各种不同的格式表示, 如XML、N-Triples、N3、JSON-LD等



(<http://...isbn...6682>, <http://.../original>, <http://...isbn...409X>)

RDF Schema

- **RDFS** 是 **RDF** 的一个扩展，提供了一个术语表用于描述 RDF 资源的 **属性**(properties) 和 **类别**(classes)
- 上述术语表被组织成一个带类别的层次体系结构
 - **Class**, **subClassOf**, **type**: 描述类别子类别
 - **Property**, **subPropertyOf**: 属性层次体系结构
 - **domain**, **range**: 定义新术语



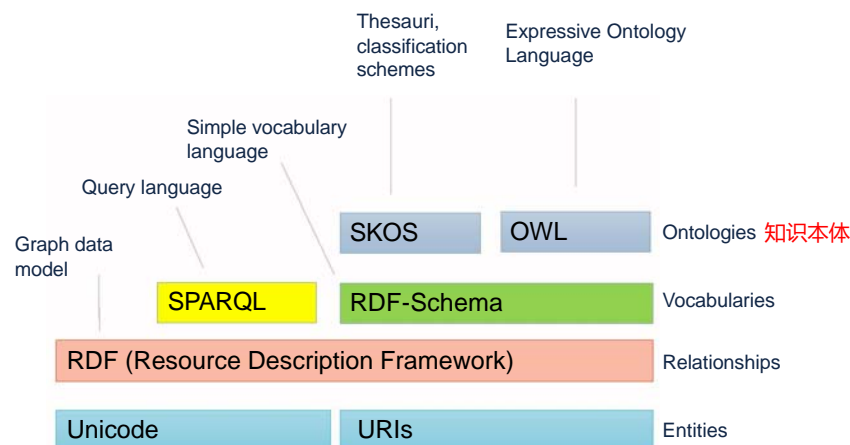
25

Web Ontology Language = OWL

- **OWL** 进一步提供了更多的术语来描述属性和类别
 - 类别之间的关系 (e.g. disjointness)
 - 基数 cardinality (e.g. "exactly one")
 - equality
 - richer typing of properties
 - characteristics of properties (e.g. symmetry)
 - 枚举类
 - ...

26

语义网知识描述语言体系



27

知识图谱 (Knowledge Graph)

- 知识图谱本质上是一种 **语义网络 (Semantic Net)**，其结点代表 **实体 (entity)** 或 **概念 (concept)**，边代表实体/概念之间的 **各种语义关系**；

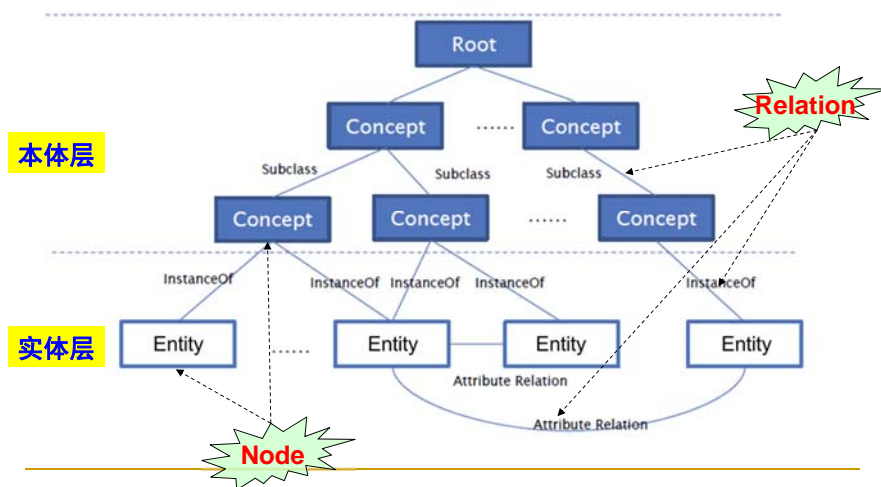


Google, 2012

- A Knowledge Graph (KG) is a system that understands facts about people, places and things and how these entities are all connected;
- 知识图谱把不同来源、不同类型的信息连接在一起形成关系网络，**提供了从关系的角度去分析问题的能力**

28

什么是知识图谱？

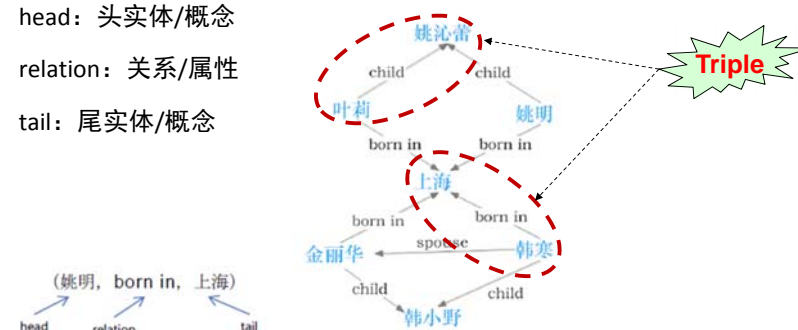


29

知识图谱中的知识表示：三元组

- 三元组Triple: (head, relation, tail)

- head: 头实体/概念
- relation: 关系/属性
- tail: 尾实体/概念



30

知识图谱的基本概念

- Node: 概念 (Concept)



31

知识图谱的基本概念

- Node: 实体/实例 (Entity/Object/Instance)

Yao Ming (Q58590)
Chinese basketball player
+edit

[In more languages](#)

[Configure](#)

Language	Label	Description	Also known as
English	Yao Ming	Chinese basketball player	
Chinese	姚明	中国篮球运动员	
Wu Chinese	No label defined	No description defined	
Cantonese	No label defined	No description defined	

All entered languages

Statements

instance of	* human	+edit
	-1 reference	
		+add value
image	* YaoMingonoffense2 crop.jpg	+edit
	-0 references	
		+add reference
		+add value
sex or gender	* male	+edit

Wikipedia (50 entries)

ar	ياو مينج
az	Yao Min
bcl	Yao Ming
bg	Яо Мин
ca	Yao Ming
cs	Jao Ming
da	Yao Ming
de	Yao Ming
el	Γιάο Μινγκ
en	Yao Ming
es	Yao Ming
et	Yao Ming
fa	یائو مینگ
fi	Yao Ming
fr	Yao Ming
gl	Yao Ming
he	יאו מינג
hr	Yao Ming
hu	Jao Ming
hy	Յաո Մինգ
id	Yao Ming
it	Yao Ming
ja	姚明
jv	Yao Ming
ka	იანო მინგა

知识图谱的基本概念

■ Node: 值 (Value)

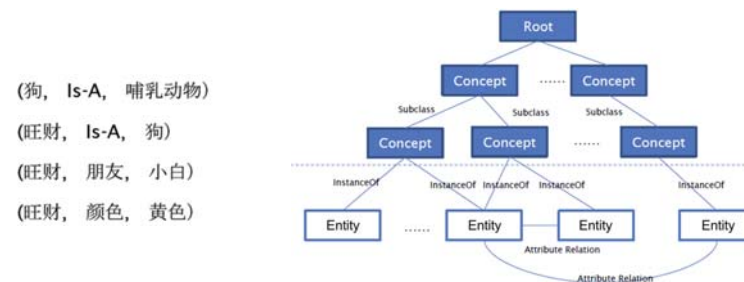
- 实体 (Entity)
 - (姚明, 出生地, 上海市)
- 字符串 (String)
 - (北京大学, 学术传统, 兼容并包、思想自由)
- 数字 (Number)
 - 平方公里: (北京市, 面积, 1.641万)
 - 公斤: (姚明, 体重, 140公斤)
 - 米: (姚明, 身高, 2.29米)
 - ...
- 时间 (Date)
 - (姚明, 出生年份, 1981年)
- 枚举 (Enumerate)
 - (姚明, 性别, 男)
-

33

知识图谱的基本概念

■ 边: 关系

- Subclass
- Type
- Relation
- Property、Attribute



34

知识图谱的基本概念

分类关系与非分类关系

■ 关系: Taxonomic Relation vs. Non-taxonomic Relation

- Taxonomic Relation: is-a/Hypernym-Hyponym
- Non-taxonomic Relation: 概念之间的相互作用
 - Part-whole 部分整体
 - Thematic role 论旨角色
 - Attribute 属性
 - Possession 领属
 - Causality 因果
 -

35

知识图谱的基本概念

■ Node: 高阶三元组

- 与时间、地点相关
 - (〈美国, 总统, 特朗普〉, 开始时间, 2017)

□ 事件

■ Compound Value Type (CVT)

- A CVT is a type within Freebase, which is used to represent data where each entry consists of multiple fields.
- CVTs are used in Freebase to represent complex data.

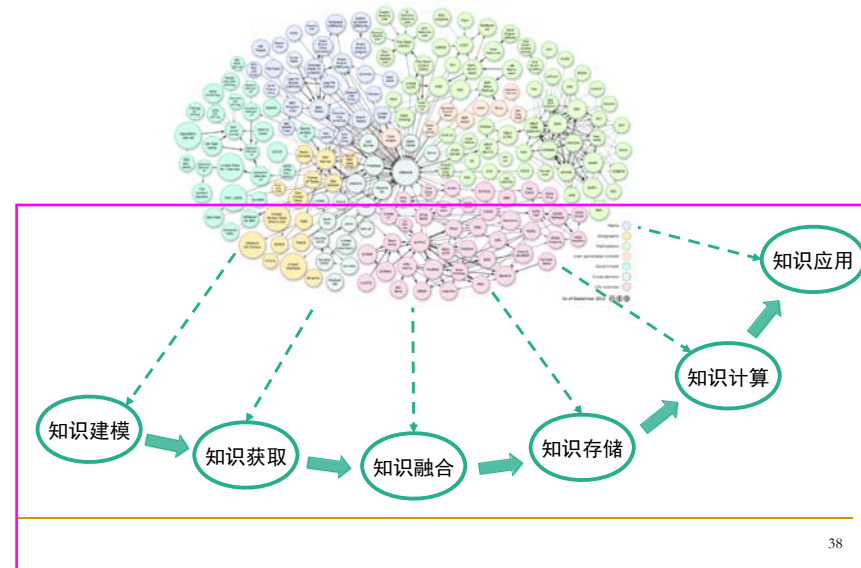
CVT是Freebase中的一种类型,
用于表示每个实体由多个字段组成的数据。

cvT在Freebase中用于表示复杂的数据。

36

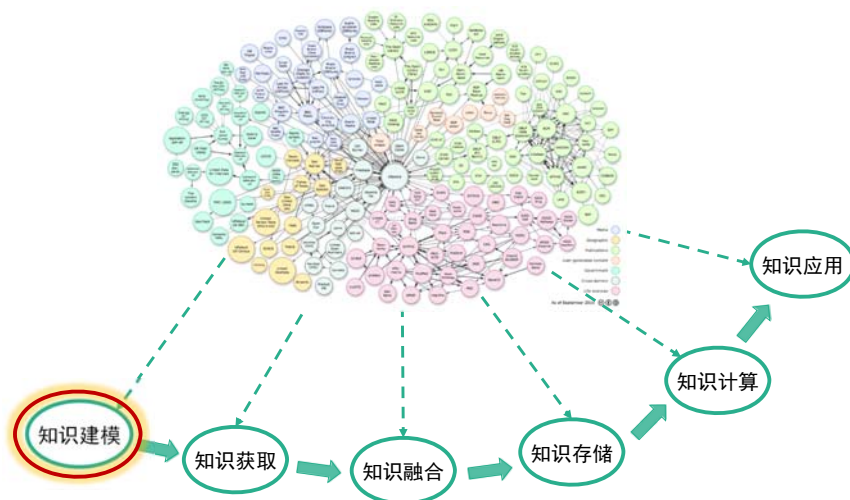
6.2 知识图谱的生命周期

知识图谱的生命周期



38

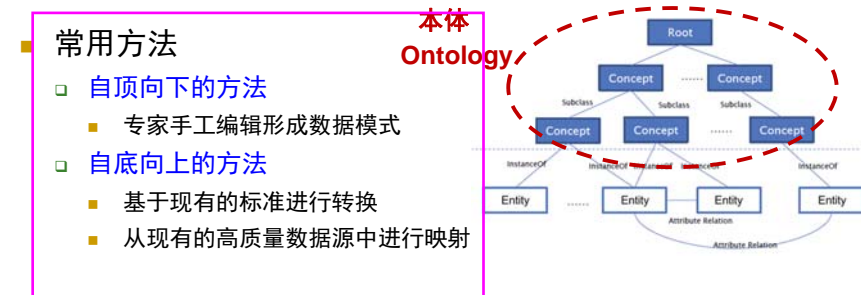
知识图谱的生命周期



39

知识建模

- 建立知识图谱的**数据模式（本体）**，包括**概念/实体的类型**，以及概念/实体之间的**关联关系**
- 数据模式定义了整个知识图谱的结构，因此需要**保证可靠性**



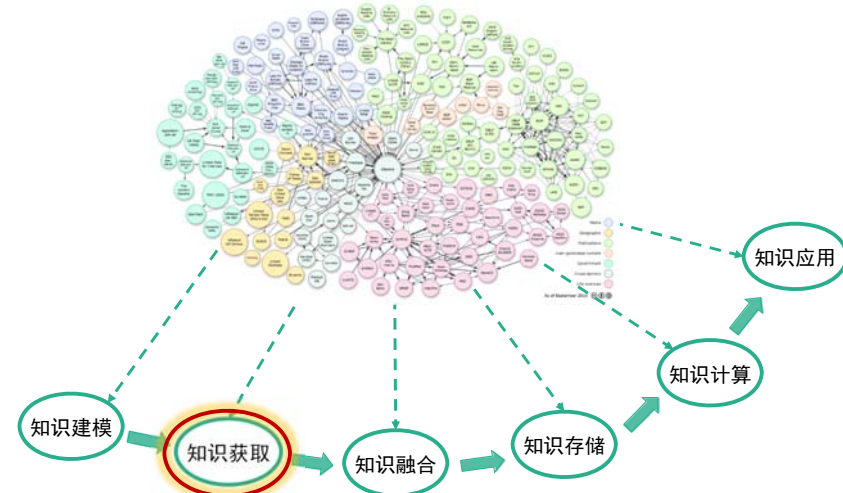
40

使用知识图谱对数据进行抽象建模

- 以**实体/概念**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**来描述各类被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**）

41

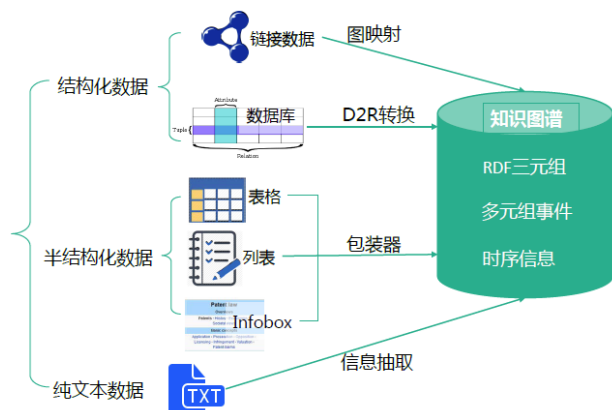
知识图谱的生命周期



42

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



43

知识获取关键技术与难点

- 从**结构化**数据库中获取知识：**D2R**
 - 难点：复杂表数据的处理
- 从**半结构化**（网站）数据中获取知识：**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**非结构化**文本中获取知识：**信息抽取**
 - 难点：结果的准确率与覆盖率

44

结构化知识获取工具：D2RQ

- D2RQ: 将关系数据库转换为虚拟的RDF数据库的平台, 主要包括:
 - D2RQ Mapping Language: 定义将关系型数据转换成RDF格式的 Mapping 规则;
 - 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;
 - 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;

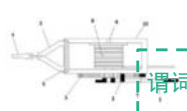
45

半结构化数据源解析

- 大多网站是通过模板来生成的, 因此通常使用包装器来进行解析
- 包装器可以自动学习, 但为了保证准确度, 通常使用人机结合的方法
- 由于网站的高度可变性, 数据源的解析尚没有统一的工具
- 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

46

包装器示例：专利知识抽取



【发明公布】一种便携式紧线装置 **主语**

申请公布号: CN105337221A 申请公布日: 2016.02.17

申请号: 201510712888 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沙坪镇沙坪村沙坪76号

分类号: H02G1/04(2006.01)I

摘要: 本发明公开了一种便携式紧线装置, 包括结构, 所述结构的右端活动连接有拉杆, 所述拉杆的右端通过销与机柄活动连接, 所述机柄的中部活动安装有轴, 所述轴的中部固定安装有收线盘, 所述机柄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>

...

47

文本信息抽取：主要任务



48

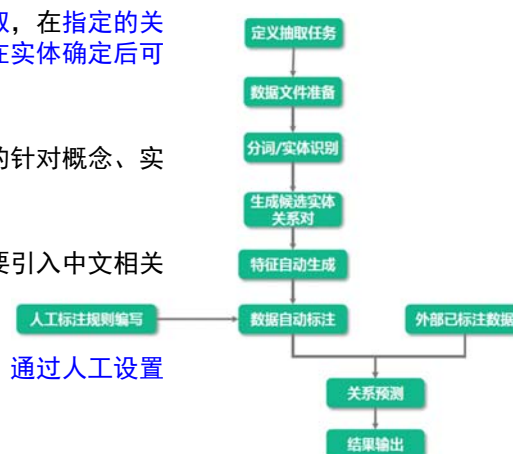
ClosedIE和OpenIE

ClosedIE	OpenIE
<ul style="list-style-type: none">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	<ul style="list-style-type: none">面向开放领域抽取信息关系模型事先未知基于语言模式抽取规模大精度比较低

49

ClosedIE 典型工具：DeepDive

- 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**
- 数据文件准备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**



50

OpenIE

- OpenIE的典型代表
 - CMU的**NELL**项目
 - 华盛顿大学的**ReVerb**、**TextRunner**
- OpenIE由于**准确率比较低**，在知识图谱构建中**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

51

NELL

- 2009年开始的CMU项目
- 输入**:
 - 初始本体（约800类别和关系）
 - 每个谓词的一些实例（约10-20个种子实例）
 - web（约10亿页面，ClueWeb）
 - 不定期的人工干预
- 任务**:
 - 24 x 7 持续运行（从2010年开始）
 - 每天
 - 抽取更多知识来补充给定本体
 - 学习如何更好的构建抽取模型
- 结果**：超过9千万实例（不同置信度）





















52

NELL抽取结果示例

Recently-Learned Facts

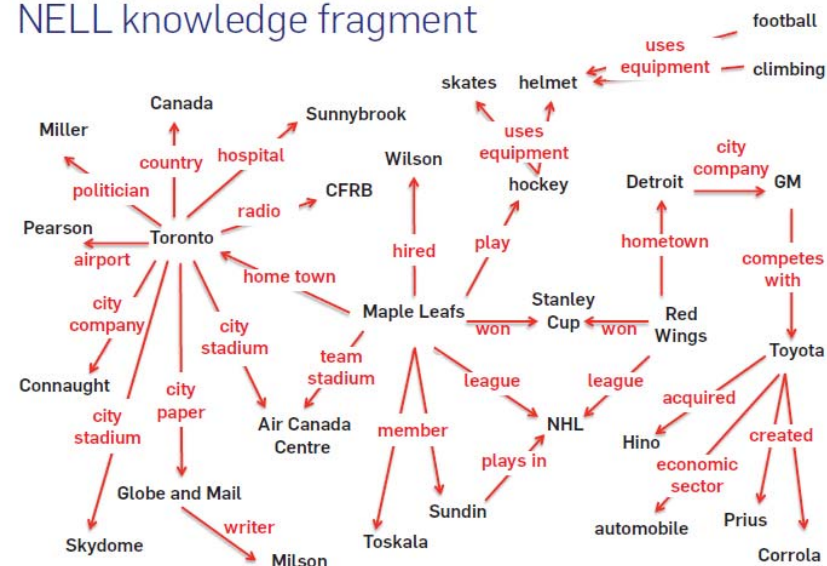
[twitter](#)

Refresh

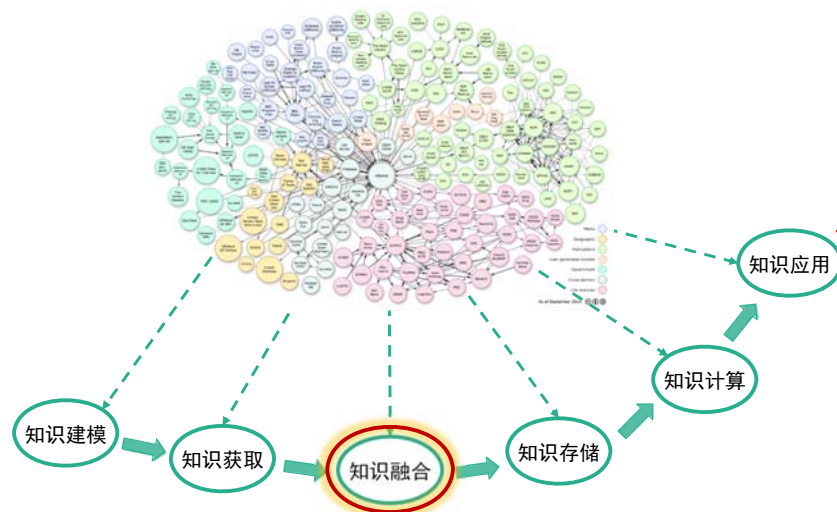
instance	iteration	date learned	confidence	
<u>mimizan</u> <u>plage</u> is a <u>visualizable scene</u>	1102	22-feb-2018	100.0	 
<u>neurofibromatosis_1</u> is a <u>disease</u>	1100	02-feb-2018	90.1	 
<u>almonds</u> is a <u>nut</u>	1103	18-mar-2018	100.0	 
<u>eyesight_requirements</u> is a <u>political issue</u>	1100	02-feb-2018	98.0	 
<u>teaching_learning_process</u> is a <u>cognitive action</u>	1100	02-feb-2018	94.3	 
<u>st_joseph_s_health_centre</u> is a hospital <u>in the city toronto</u>	1105	31-mar-2018	100.0	 
<u>pat_borzi</u> is a journalist that <u>writes for</u> the publication <u>new_york_times</u>	1102	22-feb-2018	96.9	 
<u>george_lucas directed</u> the movie <u>star_wars</u>	1103	18-mar-2018	99.2	 
<u>will contributed to</u> the creative work <u>commandments</u>	1100	02-feb-2018	99.8	 
<u>barry_sanders</u> is an athlete who <u>wins heisman trophy</u>	1105	31-mar-2018	100.0	 

<http://rtw.ml.cmu.edu/rtw/>

NELL knowledge fragment



知识图谱的生命周期



知识融合

- 知识融合分数据**模式层**与**数据层**两个层次进行融合
- 数据模式通常采用**自顶向下**和**自底向上结合**的融合方式，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

■ 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



57

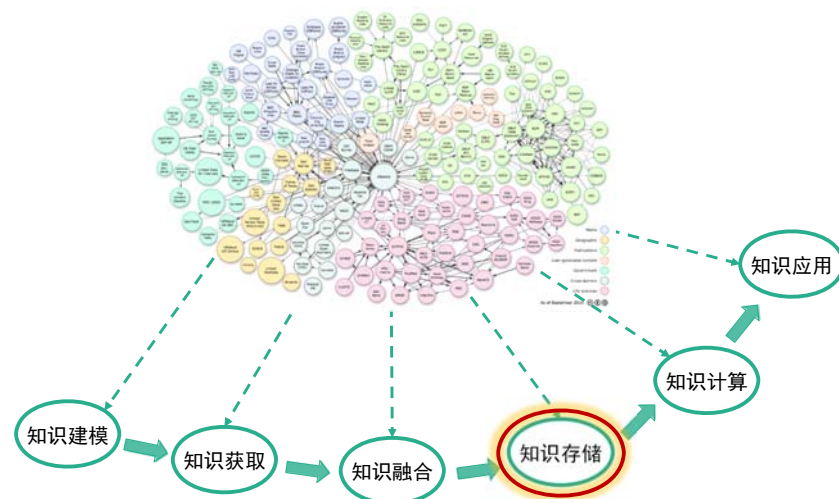
知识融合关键技术与难点

- 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



58

知识图谱的生命周期



59

知识存储关键技术与难点

- 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

→ 图数据库

60

常见的图数据存储：Graph DBMS

Rank	DBMS			Database Model	Score		
	May 2018	Apr 2018	May 2017		May 2018	Apr 2018	May 2017
1.	1.	1.	1.	Neo4j	Graph DBMS	40.58	-0.32 +4.44
2.	2.	4.	2.	Microsoft Azure Cosmos DB	Multi-model	17.54	+0.35 +12.70
3.	3.	3.	3.	Datastax Enterprise	Multi-model	7.38	-0.09
4.	4.	2.	4.	OrientDB	Multi-model	5.25	-0.39 -0.49
5.	5.	5.	5.	ArangoDB	Multi-model	3.70	-0.10 +0.75
6.	6.	6.	6.	Virtuoso	Multi-model	1.79	-0.01 -0.27
7.	7.	7.	7.	Giraph	Graph DBMS	0.98	-0.06 -0.11
8.	8.	8.	8.	Amazon Neptune	Multi-model	0.71	+0.02
9.	9.	8.	8.	AllegroGraph	Multi-model	0.58	+0.00 -0.02
10.	10.	9.	9.	Stardog	Multi-model	0.51	-0.02 +0.00
11.	11.	10.	10.	GraphDB	Multi-model	0.46	-0.00 -0.04
12.	14.	19.	19.	JanusGraph	Graph DBMS	0.41	+0.12 +0.29
13.	12.	16.	16.	Graph Engine	Multi-model	0.36	-0.04 +0.18
14.	13.	11.	11.	Sqrrl	Multi-model	0.33	-0.06 -0.13
15.	15.	22.	22.	Sparksee	Graph DBMS	0.19	-0.02 +0.14
16.	16.	17.	17.	TigerGraph	Graph DBMS	0.17	-0.01
17.	20.	14.	14.	Blazegraph	Multi-model	0.14	+0.01 -0.13
18.	18.	12.	12.	Dgraph	Graph DBMS	0.14	+0.00 -0.15
19.	17.	17.	17.	HyperGraphDB	Graph DBMS	0.14	-0.01 -0.02
20.	19.	15.	15.	FlockDB	Graph DBMS	0.13	+0.00 -0.06

<https://db-engines.com/>

61

分析型图数据库系统SQLGraph



62

SQLGraph技术亮点

查询和分析速度快

- 单机支持亿级顶点、十亿级边图计算
- 图查询比传统数据库快百倍
- 复杂图计算秒级响应
- 图计算比Neo4j快10倍以上

4000万节点，15亿条边

单位：秒	SQL Graph	Ligra
建图	4分钟	40分钟
PageRank	2.3(一次迭代)	3.58
BFS	0.43	0.64
最大连通子图	6.08	9.97

```

SELECT <results> FROM
  (initial vertex set SQL)
LOOP[(num)] JOIN
  graphAlgorithm(
    (graph construction SQL),
    params...)
[WHERE ...]
[GROUP BY ...]
[ORDER BY ...]
[LIMIT ...]
    
```

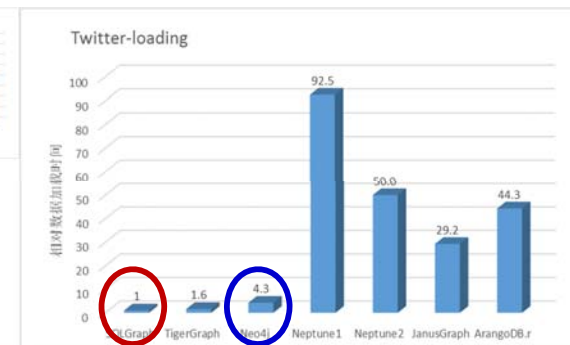
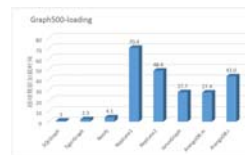
内置丰富图分析算法

- 与SQL引擎紧密结合，类SQL接口
- 可订制可扩展
- 查询结果可视化

63

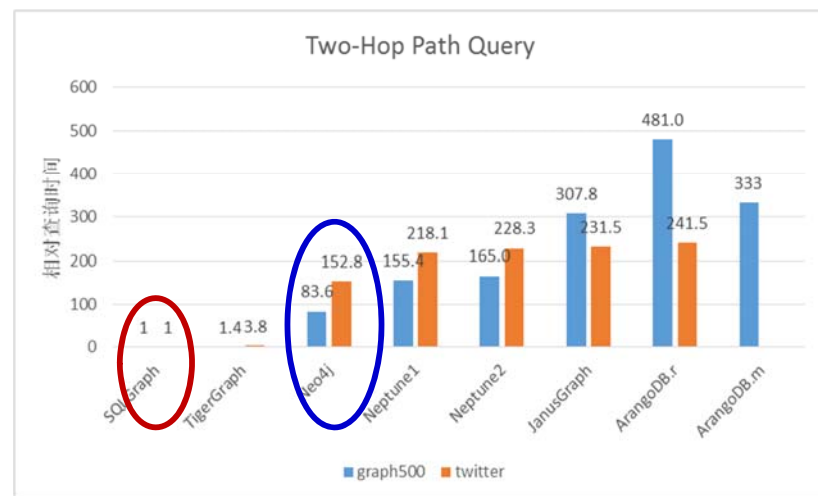
性能对比：数据加载

Name	Description and Source	Vertices	Edges
graph500	Synthetic Kronecker graph http://graph500.org	240万	6700万
twitter	Twitter user-follower directed graph http://an.kaist.ac.kr/traces/WWW2010.html	4160万	14.7亿



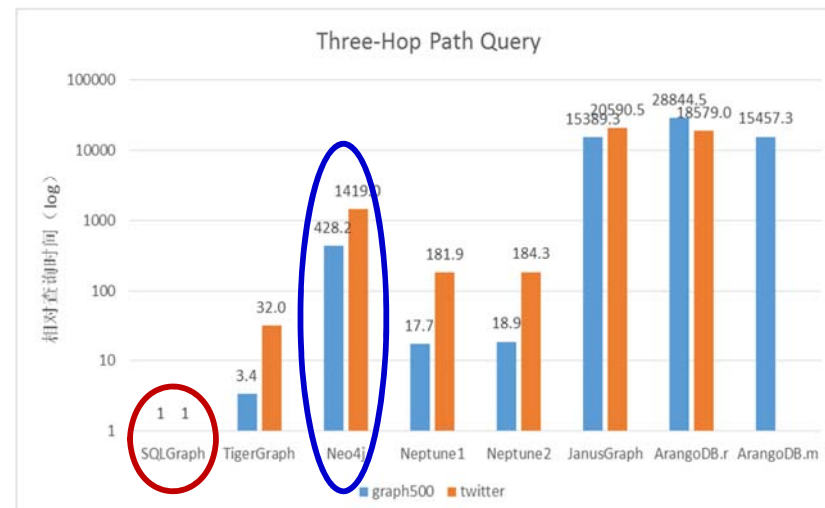
64

性能对比：路径查询



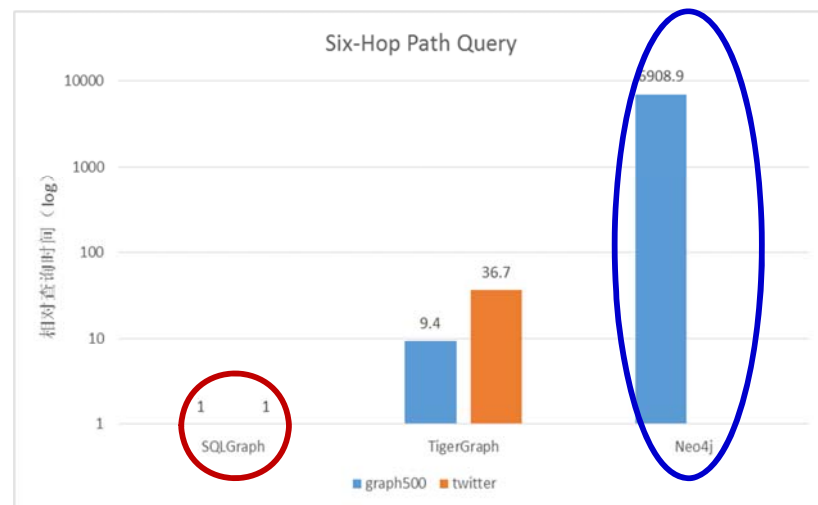
65

性能对比：路径查询



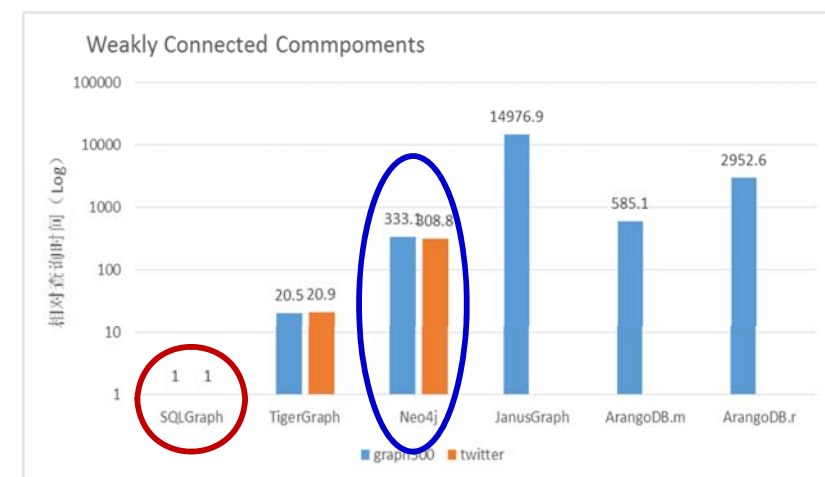
66

性能对比：路径查询



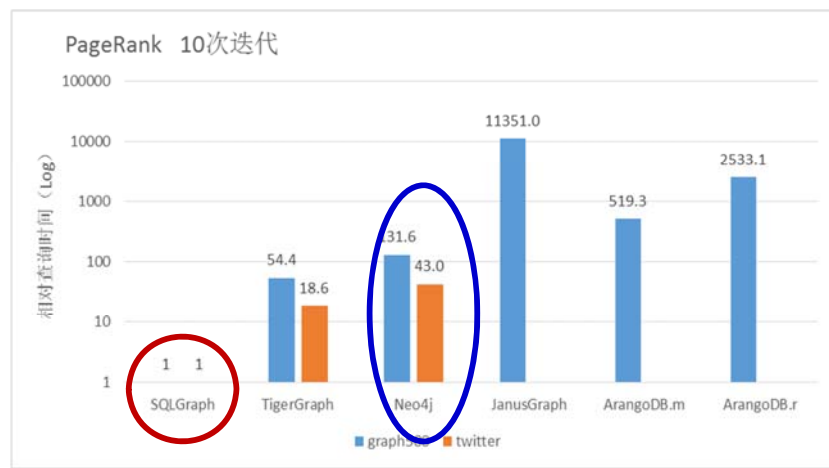
67

性能对比：弱联通子图查询



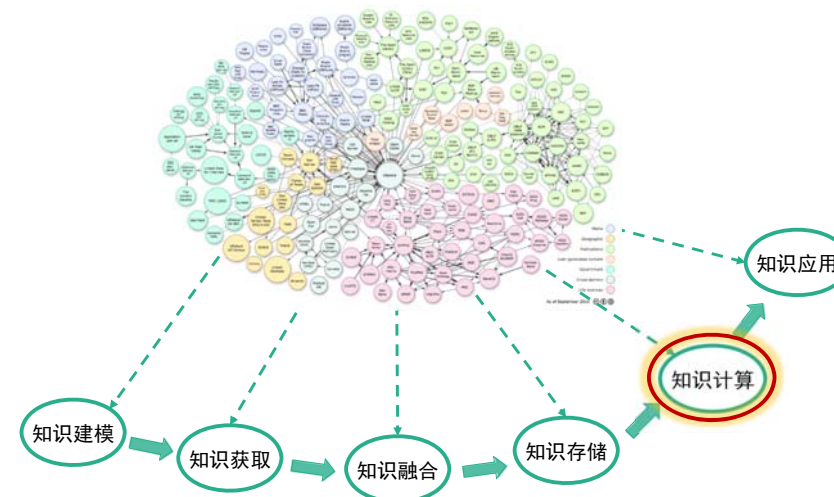
68

性能对比：PageRank分析



69

知识图谱的生命周期



70

知识计算

基于产生式

- 基于**规则**的推理：编写相应的业务规则，通过使用规则引擎实现推理
- 基于**分布式表达**的推理：利用机器学习算法将知识嵌入低维连续空间，通过**向量计算**，实现**知识推理**

71

基于产生式规则的推理

产生式系统

- 一种**前向推理**系统，可以按照一定机制**执行规则**从而达到某些目标，与一阶逻辑类似
- 应用
 - 自动规划
 - 专家系统
- 产生式系统的组成
 - 事实集合
 - 产生式集合（规则集合）
 - 推理引擎

Feigenbaum研制的化学分子结构专家系统DENDRAL
Shortliffe研制的诊断感染性疾病的专家系统MYCIN
...

72

基于产生式规则的推理

- 事实集合
 - 用于存储当前系统中所有事实
- 事实
 - 描述对象
 - 形如(*type attr₁:val₁ attr₂:val₂ ... attr_n:val_n*), 其中*type,attr_i,val_i*均为原子(常量)
 - 例如: (*student name: Alice age: 24*)
 - 描述关系(Refication)
 - 例如: (*basicFact relation: olderThan firstArg: John secondArg: Alice*)简记为(*olderThan John Alice*)

73

基于产生式规则的推理

- 产生式集合
 - 产生式规则的集合
- 产生式 LHS RHS
 - IF *conditions* THEN *actions*
 - 例如:
 - IF (*Student name:x*)
 - THEN ADD (*Person name:x*)
 - 亦可写作(具体语法因不同系统而异)
(*Student name:x*) ⇒ ADD (*Person name:x*)

如果有一个学生名为? *x*, 那么向事实集中加入一个事实, 表示有一个名为? *x*的人

74

基于产生式规则的推理

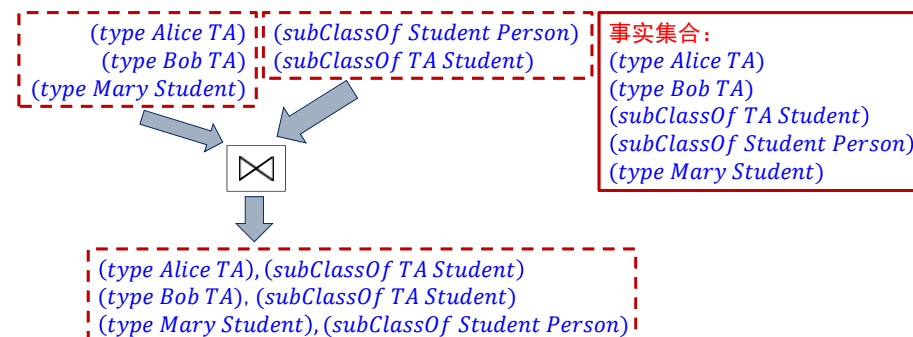
- 推理引擎
 - 控制系统的执行
 - 模式匹配
 - 用规则的条件部分匹配事实集中的事实, 整个LHS都被满足的规则被触发, 并被加入议程 (agenda)
 - 解决冲突
 - 按一定的策略从被触发的多条规则中选择一条
 - 执行动作
 - 执行被选择出来的规则的RHS, 从而对事实集合进行一定的操作

产生式系统 = 事实集合 + 产生式集合 + 推理引擎

75

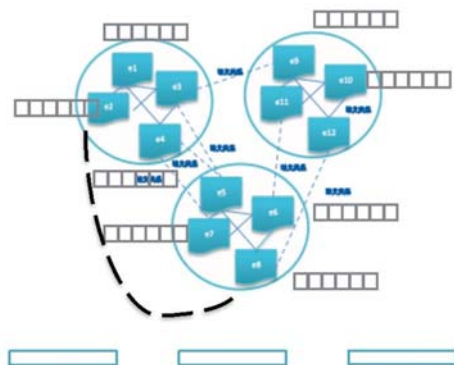
基于产生式规则的推理

- 模式匹配
 - 用每条规则的条件部分匹配当前事实集合
(*type x y*), (*subClassOf y z*) ⇒ ADD(*type x z*)



76

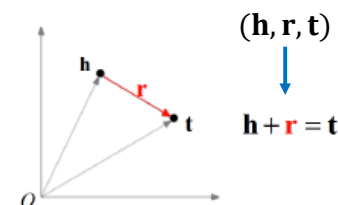
基于分布式表达的推理



77

基于分布式表达的推理：TransE

- 关系事实=(head, relation, tail)，其对应的向量表示为 $(\mathbf{h}, \mathbf{r}, \mathbf{t})$
- 基本思想：把关系看作是头尾实体之间的平移(翻译)操作



向量加法的三角形法则：

中国 + 首都 = 北京

法国 + 首都 = 巴黎

俄罗斯 + 首都 = 莫斯科

Bordes, et al. Translating embeddings for modeling multi-relational data. In Advances in Neural Information Processing Systems, 2013 (pp. 2787-2795).

78

基于分布式表达的推理：TransE

■ 势能函数

- 对于真实事实的三元组 (h, r, t) ，要求 $\mathbf{h} + \mathbf{r} = \mathbf{t}$ ；而对于错误的三元组则不满足该条件

$$f(h, r, t) = \|\mathbf{h} + \mathbf{r} - \mathbf{t}\|_2$$

$$f(\text{姚明, 出生于, 北京}) > f(\text{姚明, 出生于, 上海})$$

79

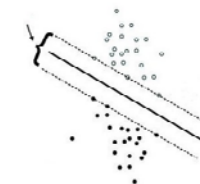
基于分布式表达的推理：TransE

■ 损失函数：

$$L = \sum_{(h, r, t) \in \Delta} \sum_{(h', r, t') \in \Delta'} \max(0, f_r(h, t) + M_{opt} - f_r(h', t'))$$

正例三元组集
负例三元组集

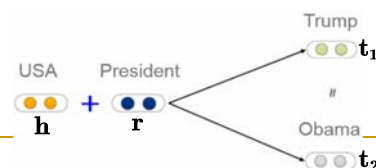
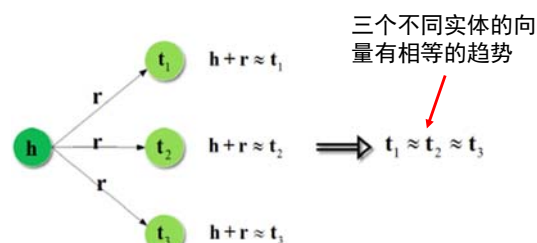
最优Margin超参



80

关系多样性问题

- 知识图谱中关系有1-1、1-N、N-1、N-M多种类型



81

Trans系列

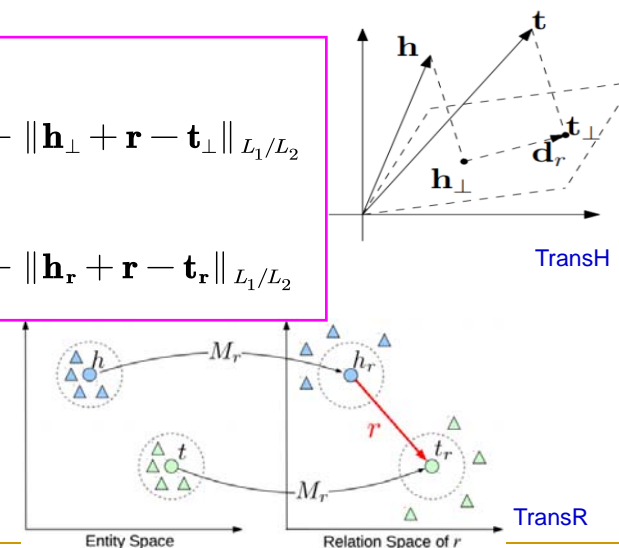
- TransH

$$f_r(h, t) = - \|\mathbf{h}_\perp + \mathbf{r} - \mathbf{t}_\perp\|_{L_1/L_2}$$

- TransR

$$f_r(h, t) = -\|\mathbf{h}_r + \mathbf{r} - \mathbf{t}_r\|_{L_1/L_2}$$

- ●



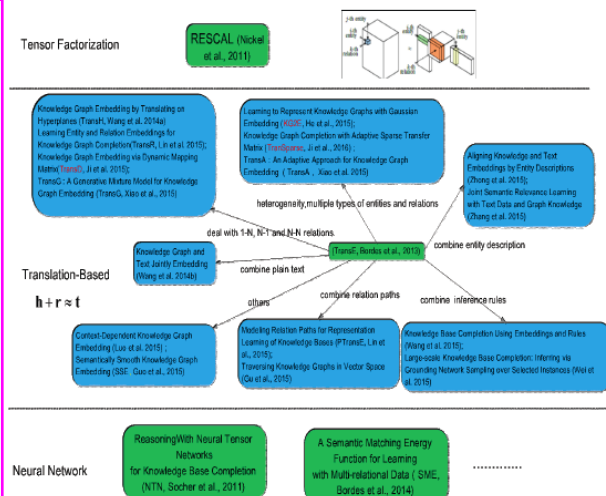
82

其他基于分布式表达的推理

张量分解方法

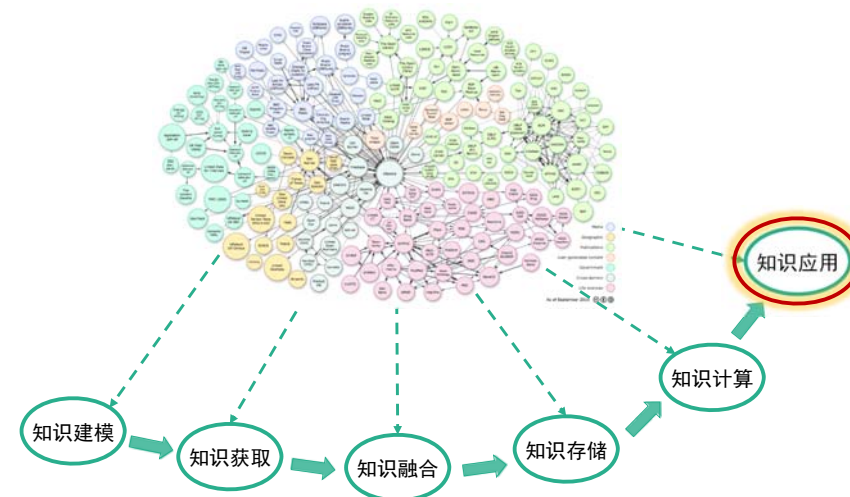
基于翻译的方法

神经网络方法



83

知识图谱的生命周期



84

6.3 知识图谱应用

通用知识图谱

- 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
- 通用知识图谱强调的是**知识的广度**，因而关注更多的是**实体**，**很难生成完整的、全局性的本体层，也很难统一管理**

86

行业知识图谱

- 行业知识图谱指面向**特定领域**的知识图谱
- 目标用户对象需要考虑行业中各种不同级别的人员，而不同人员的业务场景不同，**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**严格且丰富的数据模式**，行业知识图谱中的**实体通常属性比较多且具有行业意义**

87

行业知识图谱数据的特点

- **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，行业应用数据的体量通常都以亿级别计算，存在通常在TB、PB级别甚至更多

88

通用知识图谱 VS 行业知识图谱



- ✓ 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

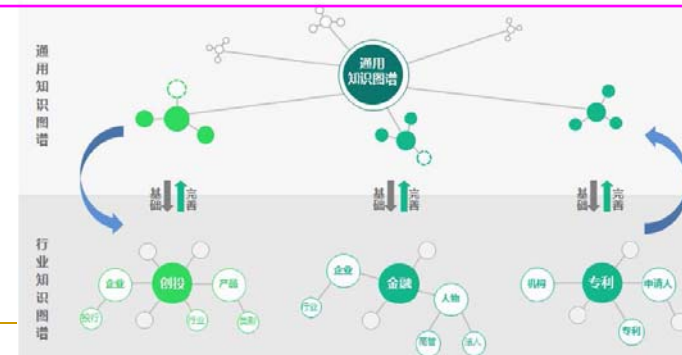


- ✓ 面向某一特定领域
- ✓ 基于行业数据构建
- ✓ 基于语义技术的行业知识库
- ✓ 强调知识的深度
- ✓ 潜在使用者是行业人

89

通用知识图谱 + 行业知识图谱

- 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



90

6.3.1 通用知识图谱的应用

智能问答
精准搜索
关系搜索
分类浏览
智能推荐
知识推理

91

智能问答

92

知识推理

<p>梁启超的儿子的妻子的情人 ✕</p> <p>网页 新闻 微信 知乎 图片 视频 地图 英文 问问 学术 更多</p> <p>搜索已为您找到10,124条相关结果 查看全部</p> <div>  <div> <p>梁启超儿子妻子情人 徐志摩</p> <p>徐志摩，(1897年1月15日—1931年11月19日)，现代诗人、散文家。原名章铤，字季真，留学美国时改名志摩。曾用过别的笔名：南湖、详情>></p> <p>推理说明：梁启超的儿子是梁思成。梁思成的第一任妻子是林徽因。林徽因的情人是徐志摩。</p> <p>徐翰立知 反馈</p> </div> </div>	<p>梁启超的儿子的妻子的情人的父亲 ✕</p> <p>网页 新闻 微信 知乎 图片 视频 地图 英文 问问 学术 更多</p> <p>搜索已为您找到11,764条相关结果 查看全部</p> <div>  <div> <p>梁启超的儿子妻子的情人的父亲 徐申如</p> <p>徐申如 (1872—1944)，名光甫，字曾筠，海盐石山人，实业家，诗人徐志摩之父。交游甚广，与南唐张謇尤为友善，深受其“实业救国”思想的影响。徐氏世代经商，早年继承祖业，独资经营徐报丰基园。清光绪二十三年(1897。详情>></p> <p>推理说明：梁启超的长子是梁思成。梁思成的第一任妻子是林徽因。林徽因的前男友是徐志摩。徐志摩的父亲是徐申如。</p> <p>徐翰立知 反馈</p> </div> </div>
<p>梁启超的儿子的妻子的情人 搜狗知识 (全部约1585条结果)</p> <ul style="list-style-type: none"> • 梁启超的儿子的太太的情人的太太是谁? 梁启超的儿子梁思成的太太——林徽因的情人——徐志摩——的太太——陆小曼!!!! 来自 匿名用户 9个回答 2012-01-12 • 梁启超的儿子的太太的情人的老婆是谁 梁启超的儿子呢是中国的著名建筑学家梁思成 梁思成的老婆呢叫林徽因看过《人面四月天》的人应该都知道 那么林徽因的老公梁思成太太的情人的老婆是谁呢 那么林徽因的老公梁思成的太太的情人的老婆是谁呢 	<p>梁启超的儿子的妻子的情人的父亲 搜狗知识 (全部约235条结果)</p> <ul style="list-style-type: none"> • 梁启超的儿子的太太的情人的父亲是谁 <input checked="" type="checkbox"/> 林徽因的情人去世了-徐志摩他爹叫徐申如 详情>> • 梁启超的儿子的妻子的情人的老婆是谁? 提问时间:2014-04-21 梁启超的儿子的太太的情人的父亲是谁 提问时间:2014-02-01

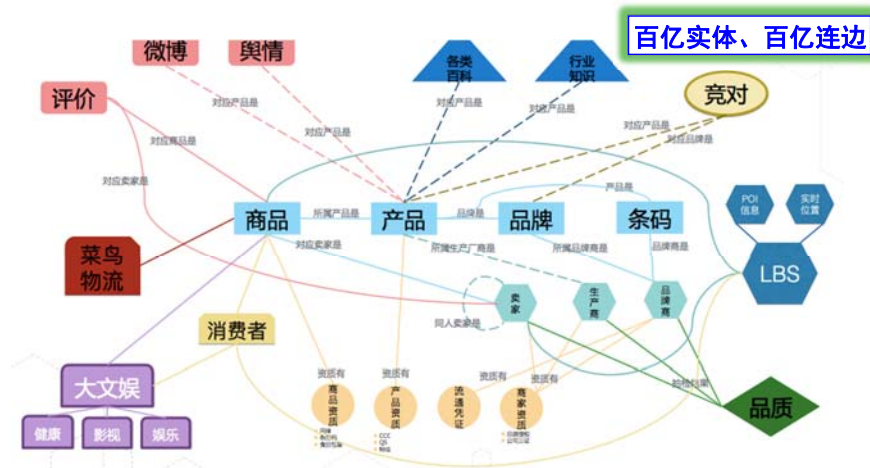
97

98

- 电子商务：阿里电商知识图谱
- 电商知识图谱：在管控中的应用，让判断更智能
- 电商知识图谱：在搜索中的应用，让搜索更智能
- 企业社交图谱查询
- 企业最终控制人查询
- 企业之间路径发现
- 辅助信贷审核
- 反欺诈
- 信息分析：人物群体分析

6.3.2 行业知识图谱的应用

电子商务：阿里电商知识图谱



在线服务：**毫秒级响应**

分级存储: 图数据库, 在线关系数据库 (搜索引擎, 缓存), 全量离线关系数据库

99

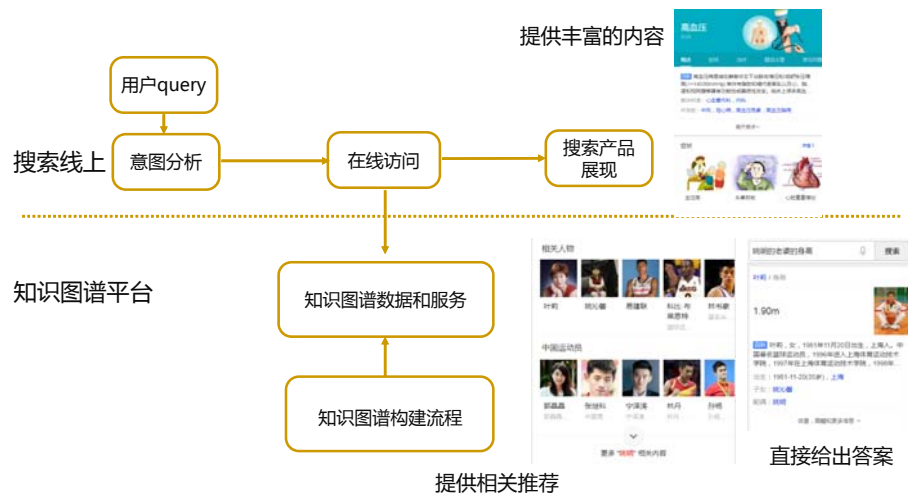
电商知识图谱：在管控中的应用，让判断更智能



判断出“真皮”和“含量30%及以下”描述有冲突

100

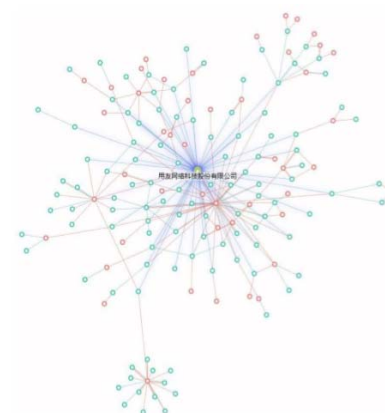
电商知识图谱：在搜索中的应用，让搜索更智能



101

企业社交图谱查询

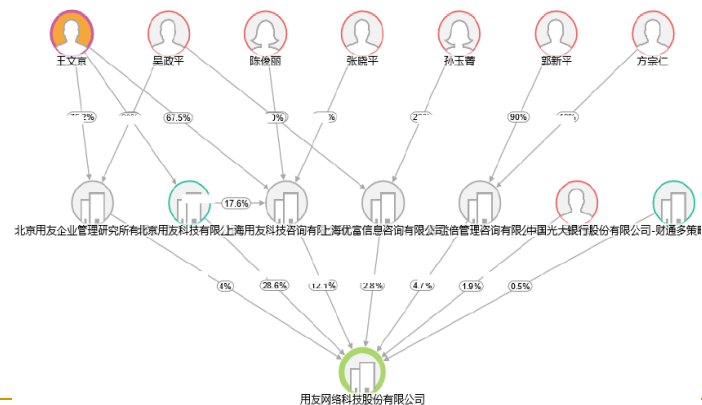
- 基于投资、任职、专利、招投标、涉诉关系以目标企业为核心向外层层扩散，形成一个网络关系图，直观立体展现企业关联



102

企业最终控制人查询

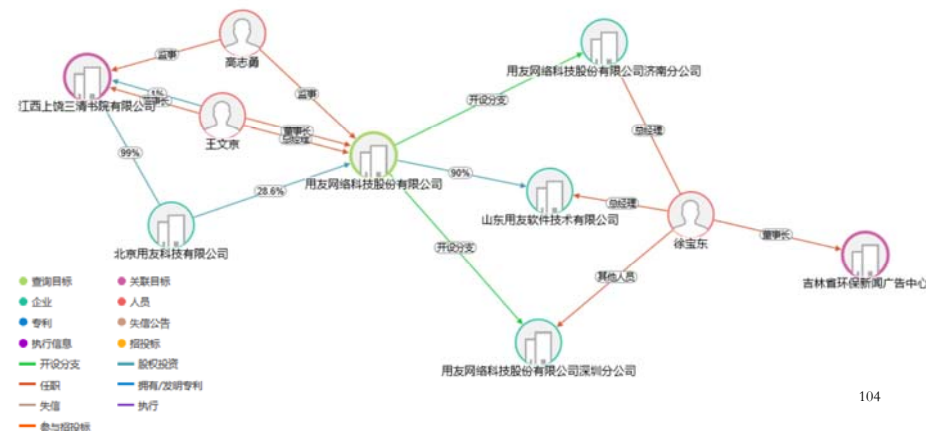
- 基于股权投资关系寻找持股比例最大的股东，最终追溯至自然人或国有资产管理部



103

企业之间路径发现

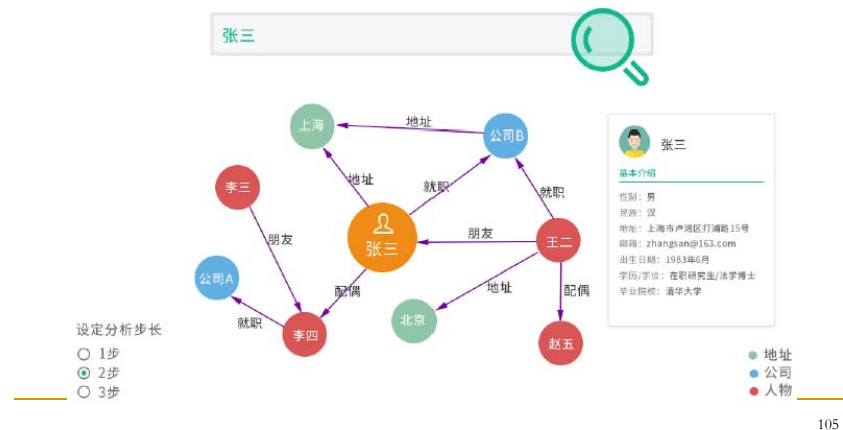
- 在基于股权、任职、专利、招投标、涉诉等关系形成的网络关系中，查询企业之间的最短关系路径，衡量企业之间的联系密切度



104

辅助信贷审核

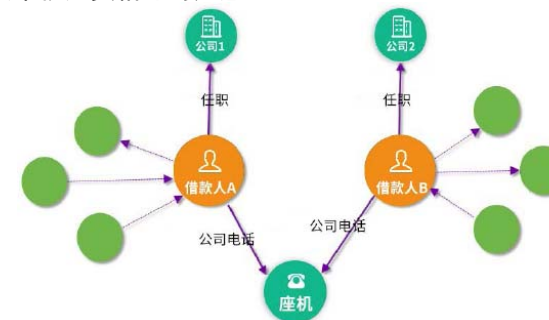
- 基于知识图谱，实现对**客户信息的全面掌握**；避免由于系统、数据等孤立造成的信息不一致，**避免信用重复使用、信息不完整**等问题



105

反欺诈 (1)

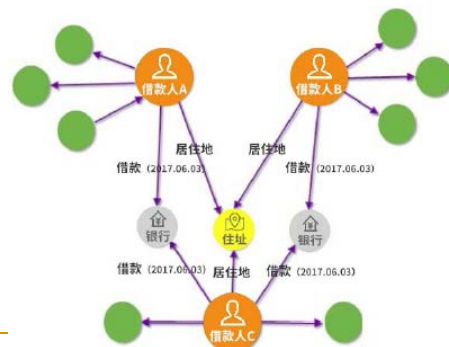
- 基于知识图谱的**不一致性验证**可以用来判断一个借款人的欺诈风险，类似交叉验证
 - 比如借款人A和借款人B填写的是同一个公司电话，但借款人A填写的公司和借款人B填写的公司完全不一样，这就成了一个风险点，需要审核人员格外的注意。



106

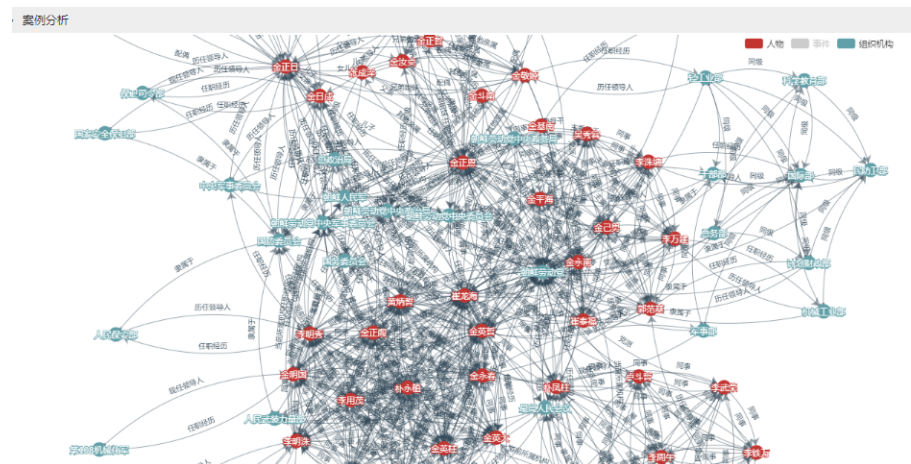
反欺诈 (2)

- 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**



107

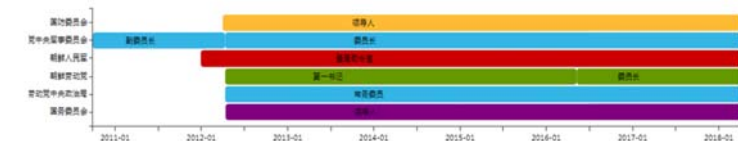
信息分析：人物群体分析



108

信息分析：人物群体分析

任职经历



近期活动



活动地点



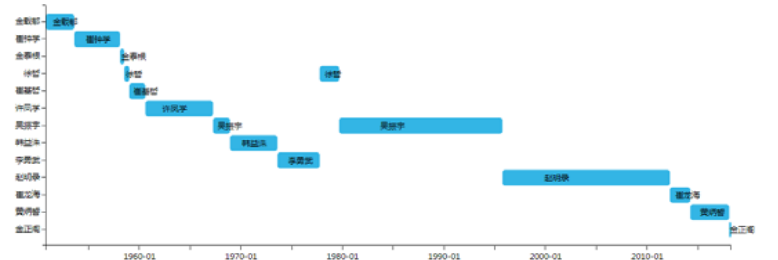
109

信息分析：人物群体分析

关系网络

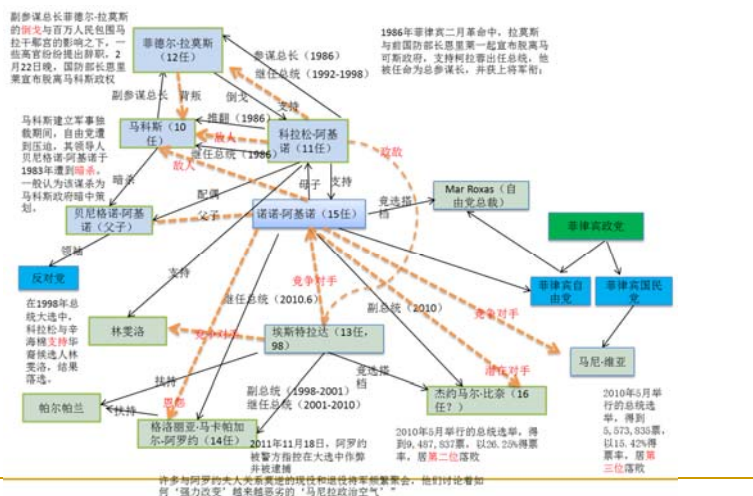


历届领导人



信息分析：菲律宾政治人物关系分析

菲律宾政治人物竞争关系（派系纷争）分析示意



111

当前知识图谱的限制和不足

■ 领域限制

- 一些知识库侧重于语言：WordNet, BabelNet
- 一些知识库侧重于Schema：Cyc, UMBEL
- 一些知识库侧重于Fact：DBPedia, Yago

■ 对时空属性的建模

- 对动态性的实体，如Event建模不足
- Yago 3在一定程度上考虑时间和地理属性

■ 完全自动构建

- 自动构建是维护和保持知识图谱质量和覆盖的核心技术

112

知识图谱展望

■ 新的知识表示模型

- Ontology engineering已经被用了超过15年
本体工程

■ 新类型的知识图谱

- 不再围绕实体和关系的存储
- 如Event-centric KG
事件为中心的知识图谱

■ 知识图谱自动构建技术

- 在Freebase中，71%的人没有出生日期
- 新技术：Distant Supervision, KG embedding, 知识集成（如Google的Knowledge Vault）
远程监控、KG嵌入

谢谢聆听！

