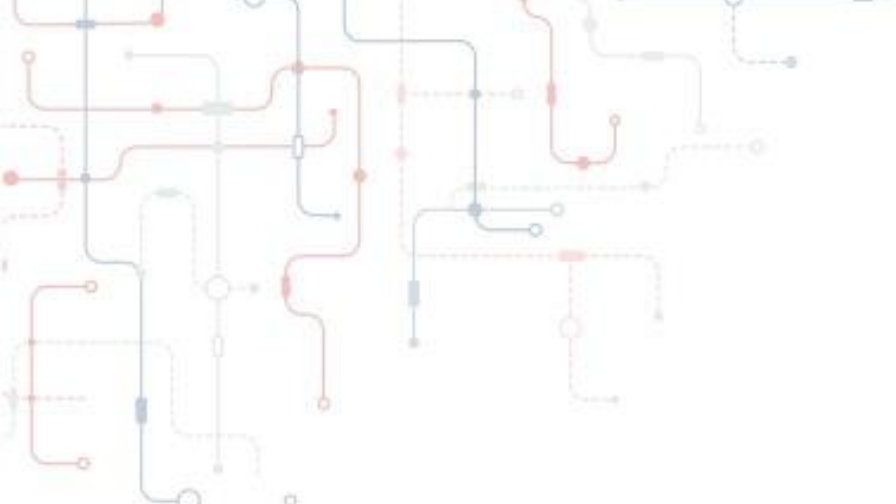


金融知识图谱的应用与探索

主讲人 张秋剑

联系方式 189-3092-8475





目录

知识图谱的实现基础

知识图谱的理论基础

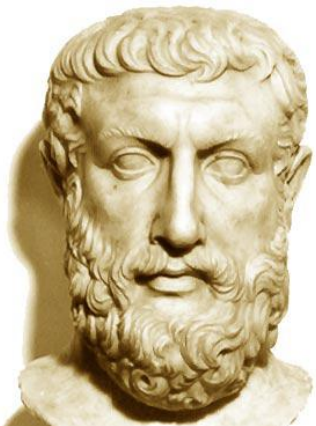
知识图谱的技术基础

案例分享

引用



知识图谱的实现基础



本体论-Ontology

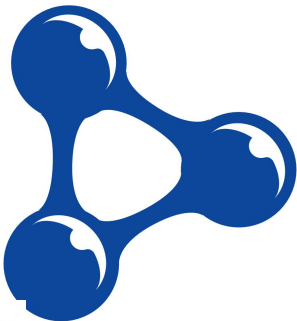
哲学概念“本体”被引入
到人工智能领域用来刻画

1960

语义网络

Semantic Networks

语义网络作为知识表示的一种方法被
提出，主要用于自然语言理解领域



知识

1980

1989

Web

Tim Berners-Lee发明了万维网
Linked Information System



WIKIPEDIA



Wikipedia is hosted by the Wikimedia Foundation, a non-profit organization that also hosts a range of other projects.

The Semantic Web
从超文本链接到语义链接

1998

链接数据Linked Data

Tim突出强调语义网的本质是
要建立开放数据之间的链接

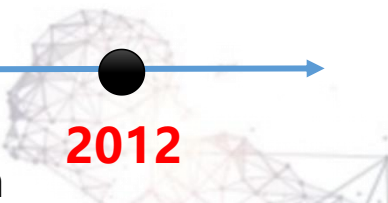


知识图谱

Knowledge Graph

谷歌发布了基于知识图谱
的搜索引擎产品

2012



2018 AI先行者大会

知识图谱的意义 — 洞察语义

- 2012年5月16日，Google为了提升搜索引擎返回的答案质量和用户查询的效率，发布了知识图谱（Knowledge Graph）
- 有知识图谱作为辅助，搜索引擎能够洞察用户查询背后的语义信息，返回更为精准、结构化的信息，更大可能地满足用户的查询需求
- Google知识图谱的宣传语 “**things not strings**” 给出了知识图谱的精髓，即，**不要无意义的字符串**，而是获取字符串背后隐含的对象或事物

贵州茅台

全部

图片

新闻

购物

更多

设置

工具

找到约 12,800,000 条结果 (用时 0.50 秒)

贵州茅台(600519)股票股价,行情,新闻,财报数据_新浪财经_新浪网

finance.sina.com.cn/realstock/company/sh600519/nc.shtml ▼

新浪财经为您提供**贵州茅台**(600519)股票实时行情走势,实时资金流向,实时新闻资讯,研究报告,股吧互动,交易信息,个股点评,公告,财务指标分析等与**贵州茅台**(600519) ...

个股资讯 · 年度报告 · 公司简介 · 公司章程

A 贵州茅台集团

中国上海市徐汇区

网站

路线

B 贵州茅台酒股份有限公司

中国上海市嘉定区

网站

路线

C 贵州茅台集团

中国上海市闵行区

网站

路线

更多地点

貴州茅台酒公司

贵州茅台酒股份有限公司, 1999年, 由中国贵州茅台酒厂有限责任公司、贵州茅台酒厂技术开发公司、贵州省轻纺集体工业联社、深圳清华大学研究院、中国食品发酵工业研究所、北京糖业烟酒公司、江苏省糖烟酒总公司、上海捷强烟草酒公司等八家公司共同发起。 [维基百科](#)

股价: 600519 (SHA) ¥650.97 +6.97 (+1.08%)
 9月7日 GMT+8 下午3:00 - 免费查询

总部: 中華人民共和國

收入: 526.6 亿人民币

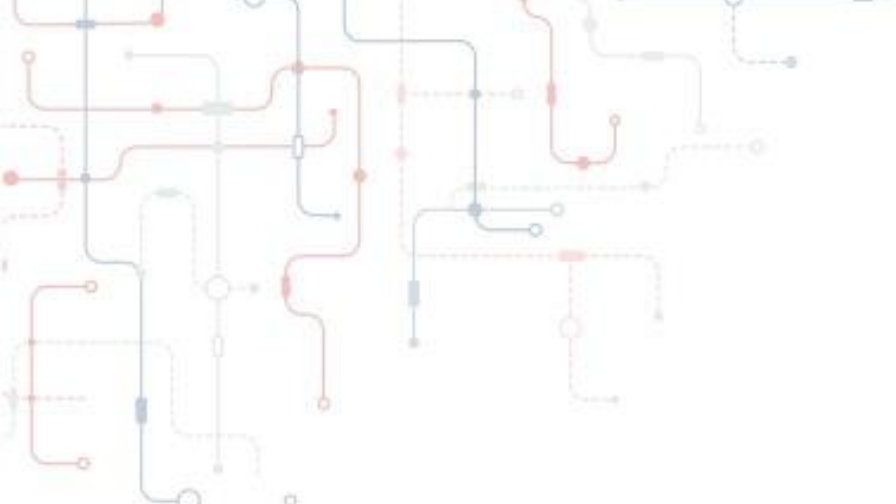
公司類型: 國有企業

股票代码: 上交所: 600519 (2001年8月上市)

子公司: Kweichow Moutai Distillery Group Technological Development Co., Kweichow Moutai Group Finance Co., Ltd.

总组织: CHINA KWEICHOW MOUTAI DISTILLERY (Group) CO., LTD

用户还搜索了 [查看另外 10+ 项](#)



目录

知识图谱的实现基础

知识图谱的理论基础

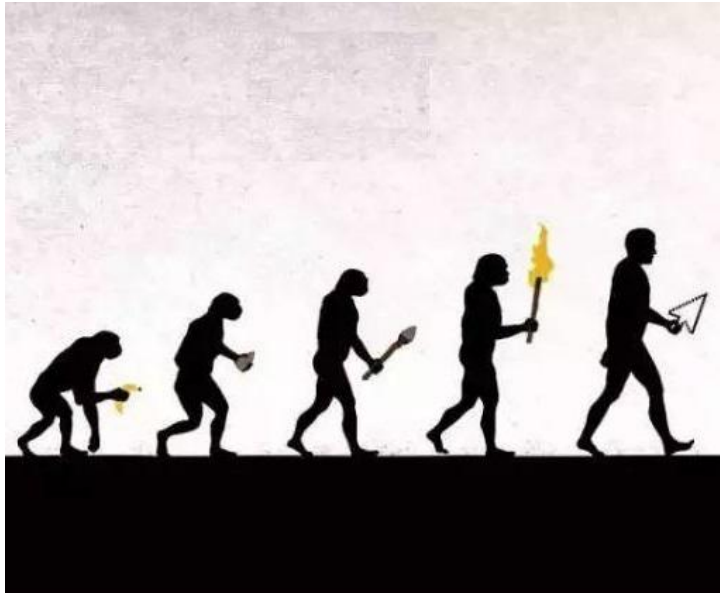
知识图谱的技术基础

案例分享

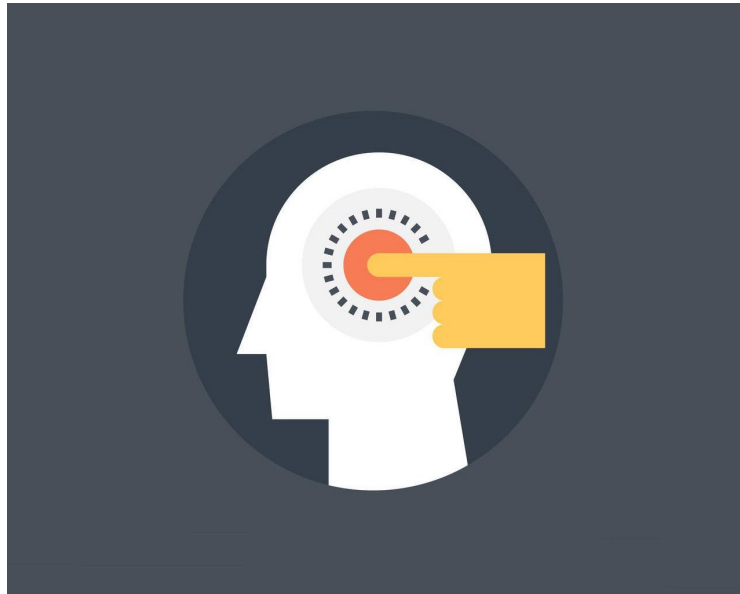
引用



知识从何而来？



进化



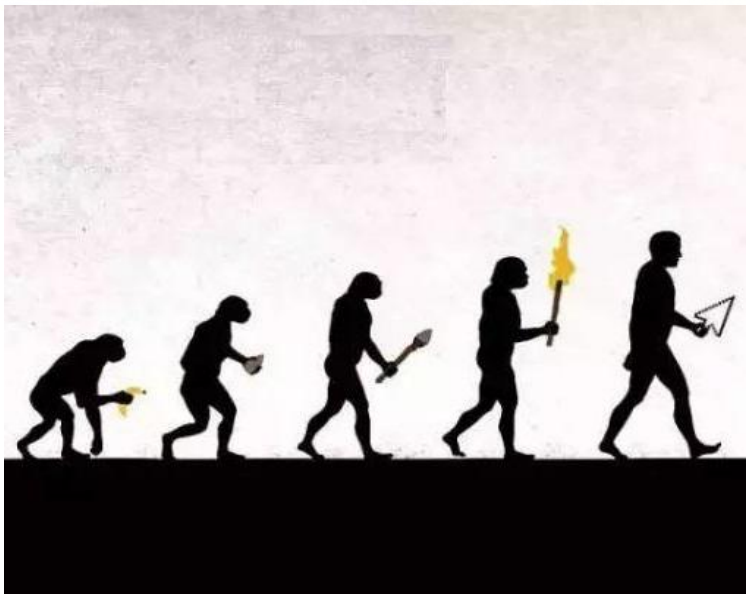
经验



文化

知识从何而来？

进化



文化



经验



计算机

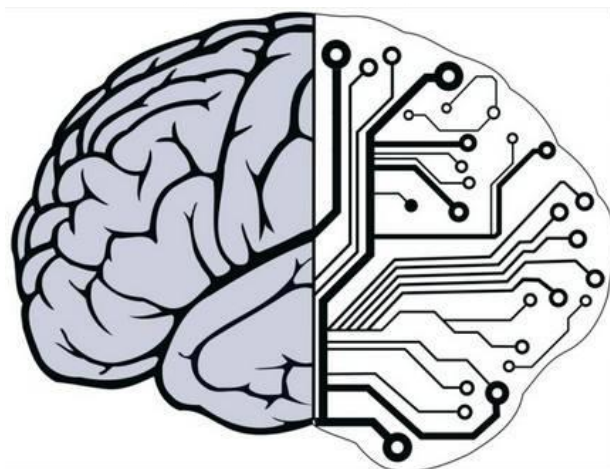


2018 AI先行者大会

计算机如何发现新知识?



填补现有知识空白



模仿大脑

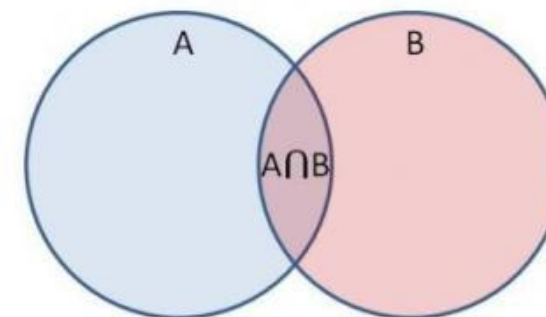


模拟进化

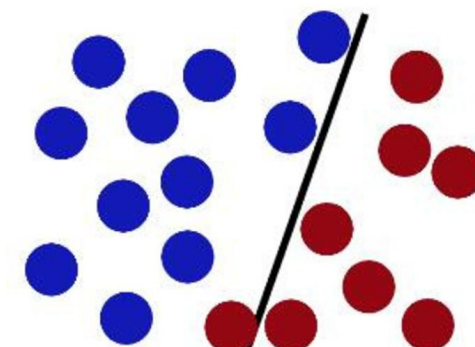
A已发生的条件下B发生的概率:

$$P(B | A) = P(A \cap B) / P(A)$$

$$P(A \cap B) = P(B | A) \times P(A)$$



系统性的减少不确定性



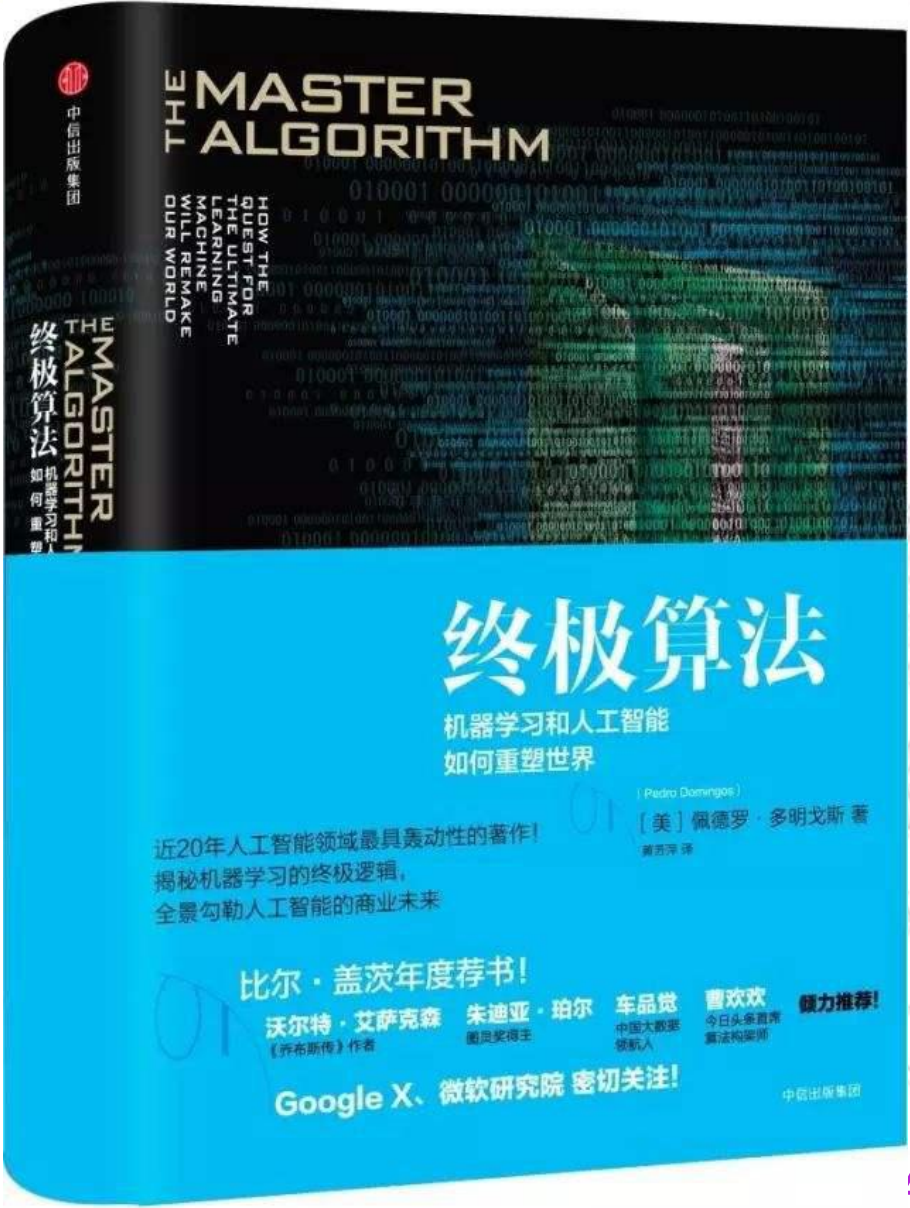
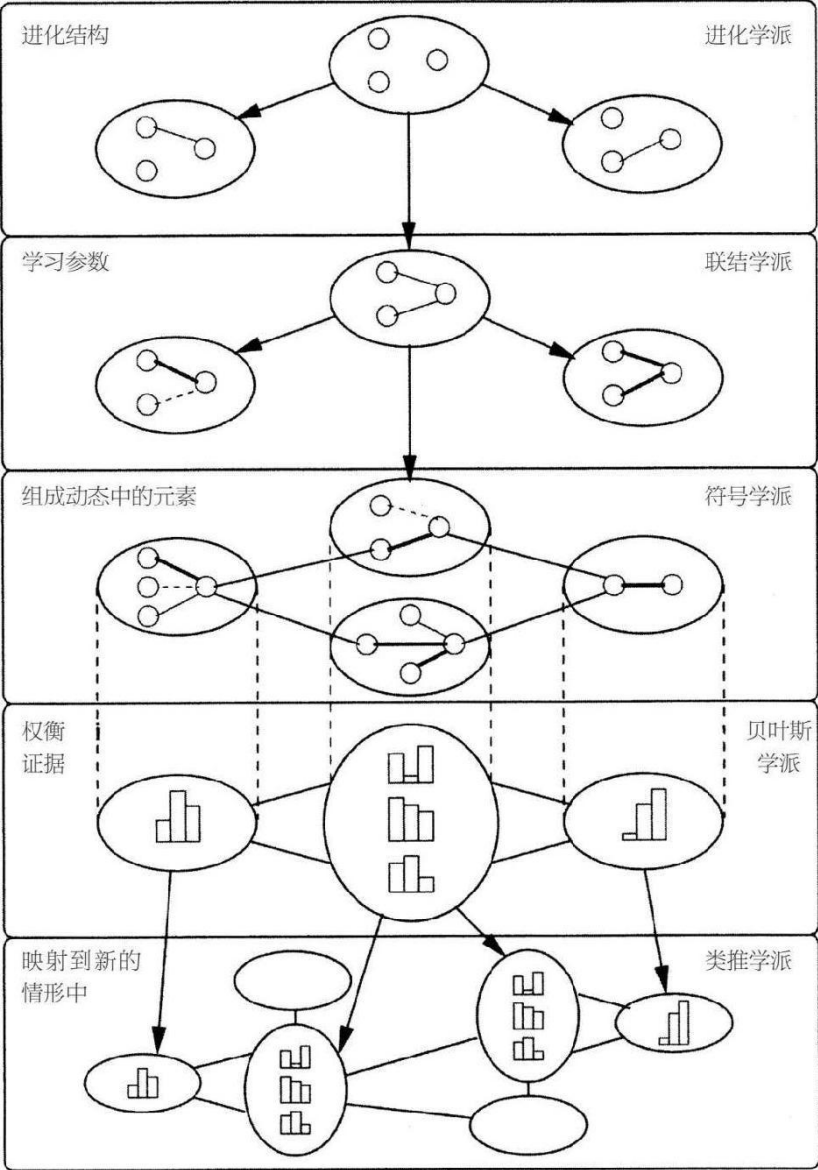
注意新旧知识之间的相似性

2018 AI先行者大会

机器学习的五大学派

知识发现	学派	学科依据	方法论	主要算法	应用场景
系统性的减少不确定性	贝叶斯学派	统计学	概率推理	朴素贝叶斯、分类器	风控
注意新旧知识之间的相似性	类推学派	心理学	行为类推主义	支持向量机、内核机	推荐
模拟进化	进化学派	进化生物学	遗传算法	遗传编码	机器人
模仿大脑	联结学派	神经科学	希望从大脑运行方式得到启发	反向传播	深度学习
填补现有知识空白	符号学派	逻辑学、哲学	相信填补现有知识的空白的	逆向演绎	专家系统

机器学习的五大学派



AI先行者大会

构建“知识图谱”的理论基础

Inverse Deduction

Deduction

Socrates is human
+ Humans are mortal
= ?

Induction

Socrates is human
+ ?
= Socrates is mortal

语义理解

Inverse Deduction

Addition

2
+ 2
= ?

Subtraction

2
+ ?
= 4

知识表示

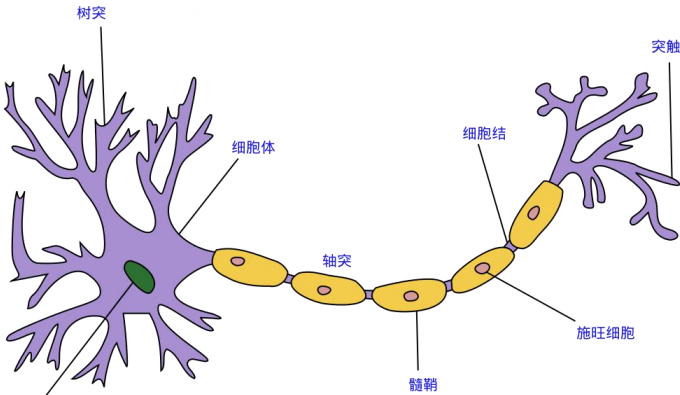
模仿人的“心”

模仿人的大脑结构

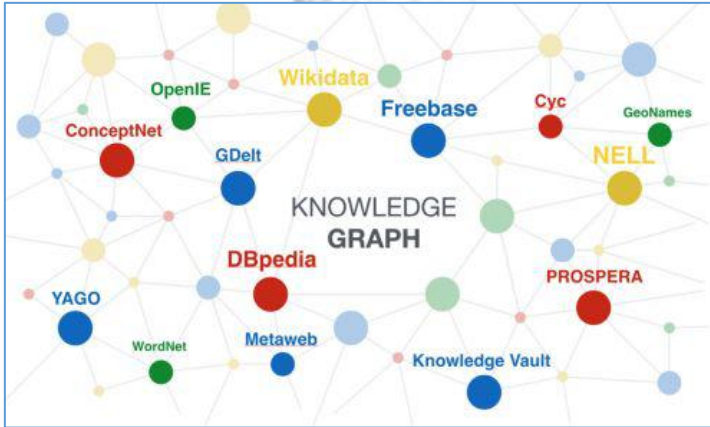


模仿人的“智”

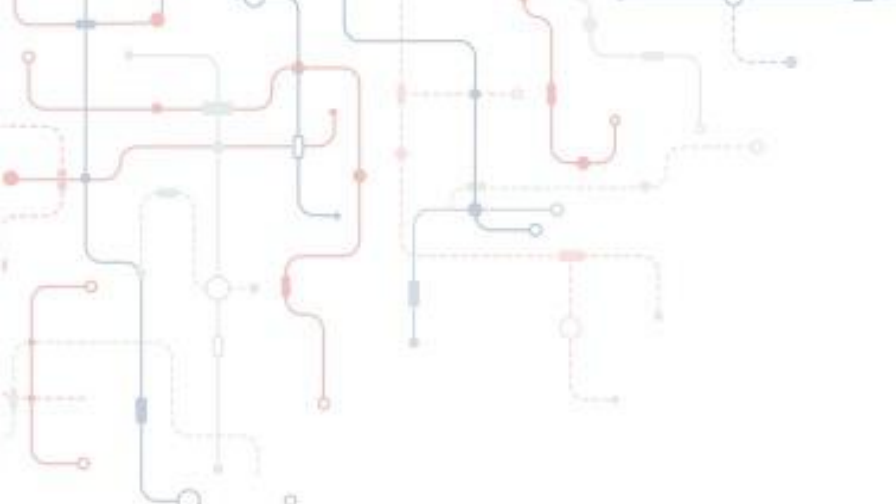
语言、思考、推理



深度学习



知识图谱



目录

知识图谱的实现基础

知识图谱的理论基础

知识图谱的技术基础

案例分享

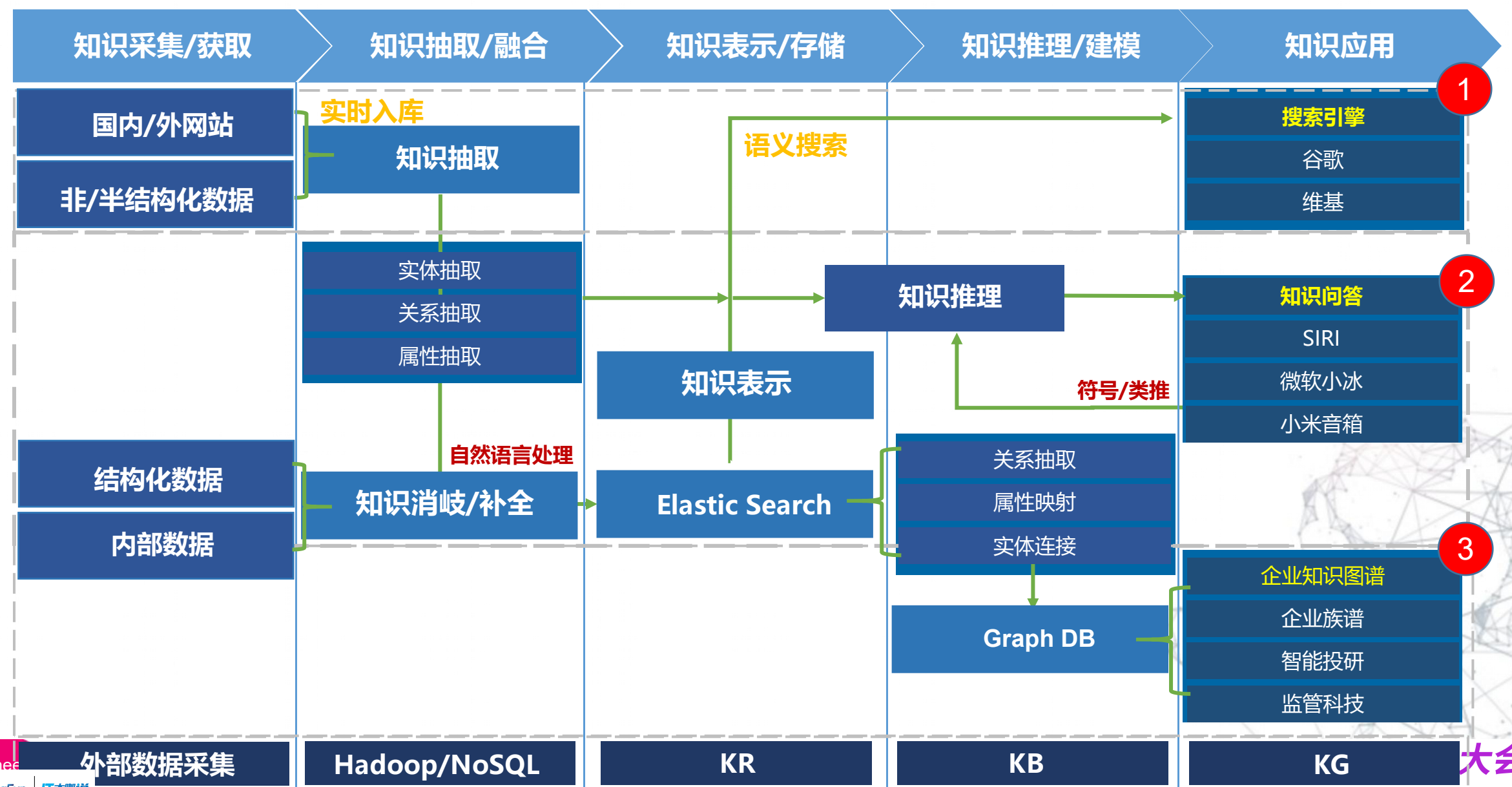
引用



知识图谱的“技术栈”

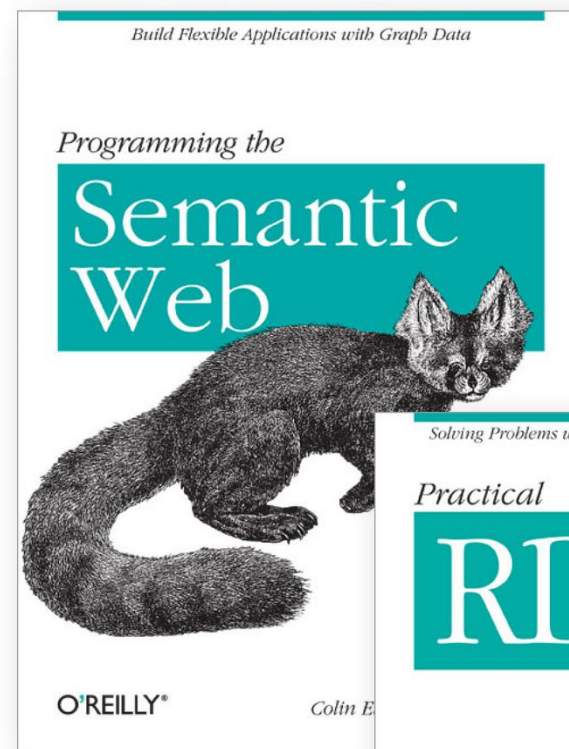


知识图谱解决方案



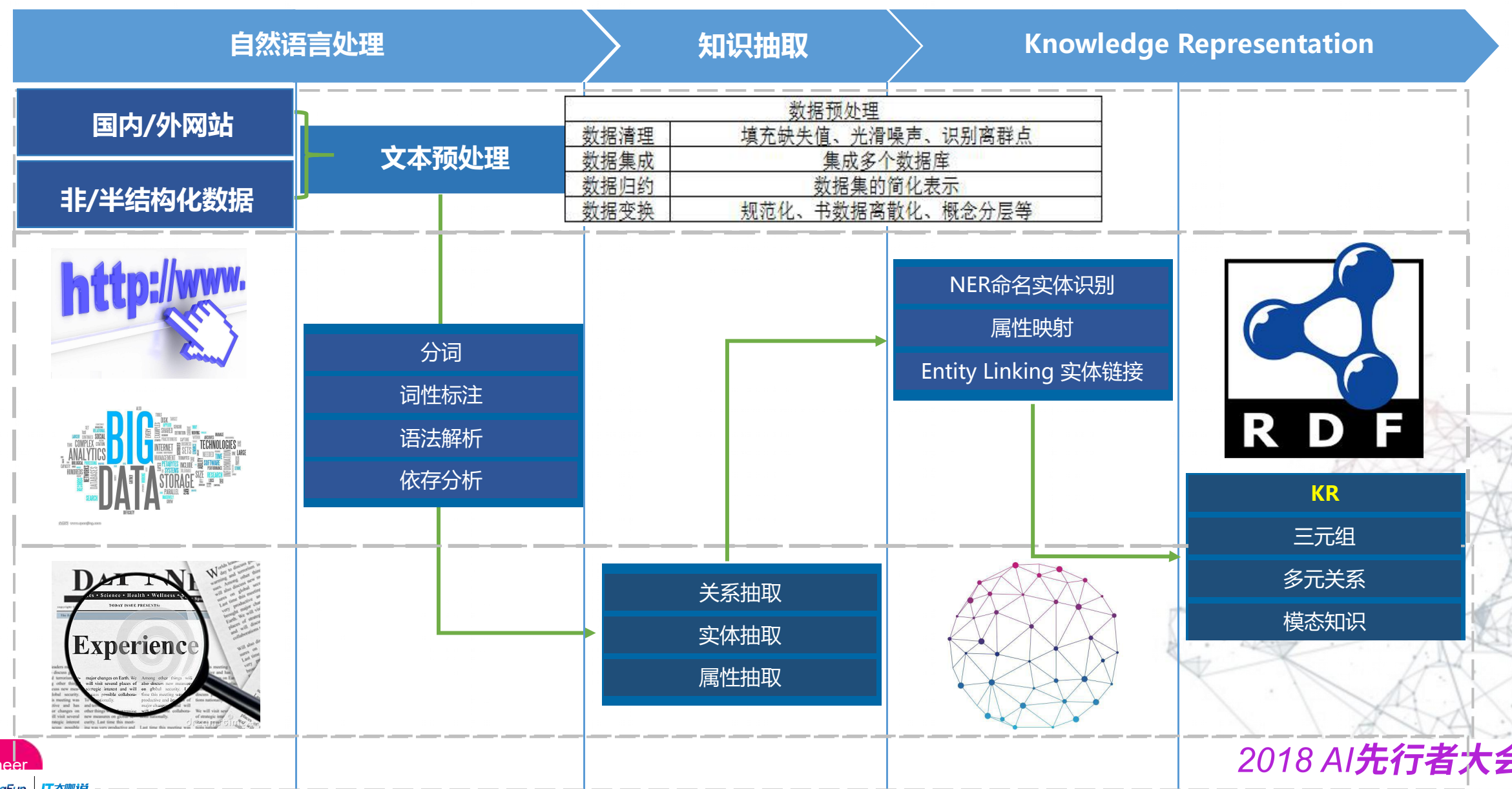
知识图谱的数据源

知识图谱库名称	机构	特点、构建手段	应用产品
FreeBase	MetaWeb(2010年被谷歌收购)	• 实体、语义类、属性、关系； • 自动+人工：部分数据从维基百科等数据源抽取而得到；另一部分数据来自人工协同编辑 • https://developers.google.com/freebase/	Google Search Engine, Google Now
Knowledge Vault (谷歌知识图谱)	Google	• 实体、语义类、属性、关系； • 超大规模数据库；源自维基百科、Freebase、《世界各国纪实年鉴》 • https://research.google.com/pubs/pub45634	Google Search Engine, Google Now
DBpedia	莱比锡大学、柏林自由大学、OpenLink Software	• 实体、语义类、属性、关系 • 从维基百科抽取 •	DBPedia
维基数据(Wikidata)	维基媒体基金会 (Wikimedia Foundation)	• 实体、语义类、属性、关系,与维基百科紧密结合 • 人工(协同编辑)	Wikipedia
Wolfram Alpha	沃尔夫勒姆公司(Wolfram Research)	• 实体、语义类、属性、关系,知识计算 • 部分知识来自于Mathematica；其它知识来自于各个垂直网站	Apple Siri
Bing Satori	Microsoft	• 实体、语义类、属性、关系,知识计算 • 自动+人工	Bing Search Engine, Microsoft Cortana
YAGO	马克斯·普朗克研究所	• 自动：从维基百科、WordNet和GeoNames提取信息	YAGO
Facebook Social Graph	Facebook	• Facebook 社交网络数据	Social Graph Search
百度知识图谱	百度	• 搜索结构化数据	百度搜索
搜狗知立方	搜狗	• 搜索结构化数据	搜狗搜索
ImageNet	斯坦福大学	• 搜索引擎 • 亚马逊 AMT	计算机视觉相关应用



2018 AI先行者大会

知识抽取 – NLP+KR



知识表示 – 数据结构 – RDF

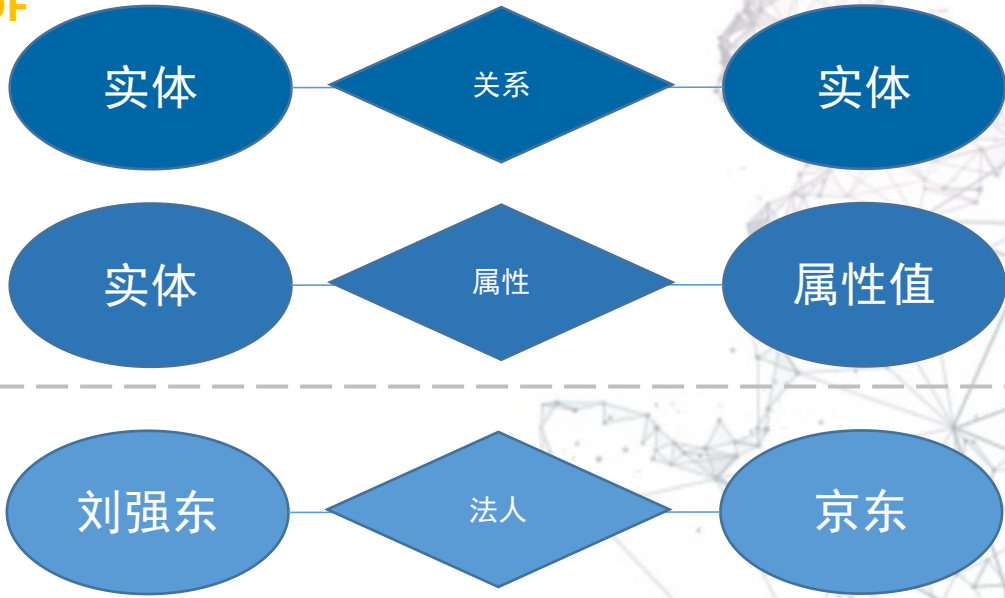
- 实体：指的是具有可区别性且独立存在的某种事物。如某一个人、某一个城市、某一种植物等、某一种商品等等。世界万物由具体事物组成，此指实体。如图1的“中国”、“美国”、“日本”等。，实体是知识图谱中的最基本元素，不同的实体间存在不同的关系。
- 语义类（概念）：具有同种特性的实体构成的集合，如国家、民族、书籍、电脑等。概念主要指集合、类别、对象类型、事物的种类，例如人物、地理等。
- 内容：通常作为实体和语义类的名字、描述、解释等，可以由文本、图像、音视频等来表达。
- 属性：从一个实体指向它的属性值。不同的属性类型对应于不同类型属性的边。属性值主要指对象指定属性的值。如图1所示的“面积”、“人口”、“首都”是几种不同的属性。
- 属性值：主要指对象指定属性的值，例如960万平方公里等。
- 关系：形式化为一个函数，它把 k k个点映射到一个布尔值。在知识图谱上，关系则是一个把k k个图节点(实体、语义类、属性值)映射到布尔值的函数。

SPO



RDF	数据类型	释义	引申
Subject	IRI/blank node	URI/匿名资源	空资源
Predicate	IRI	URI	
Object	IRI/blank node/literals	URI/字面量	文本

RDF



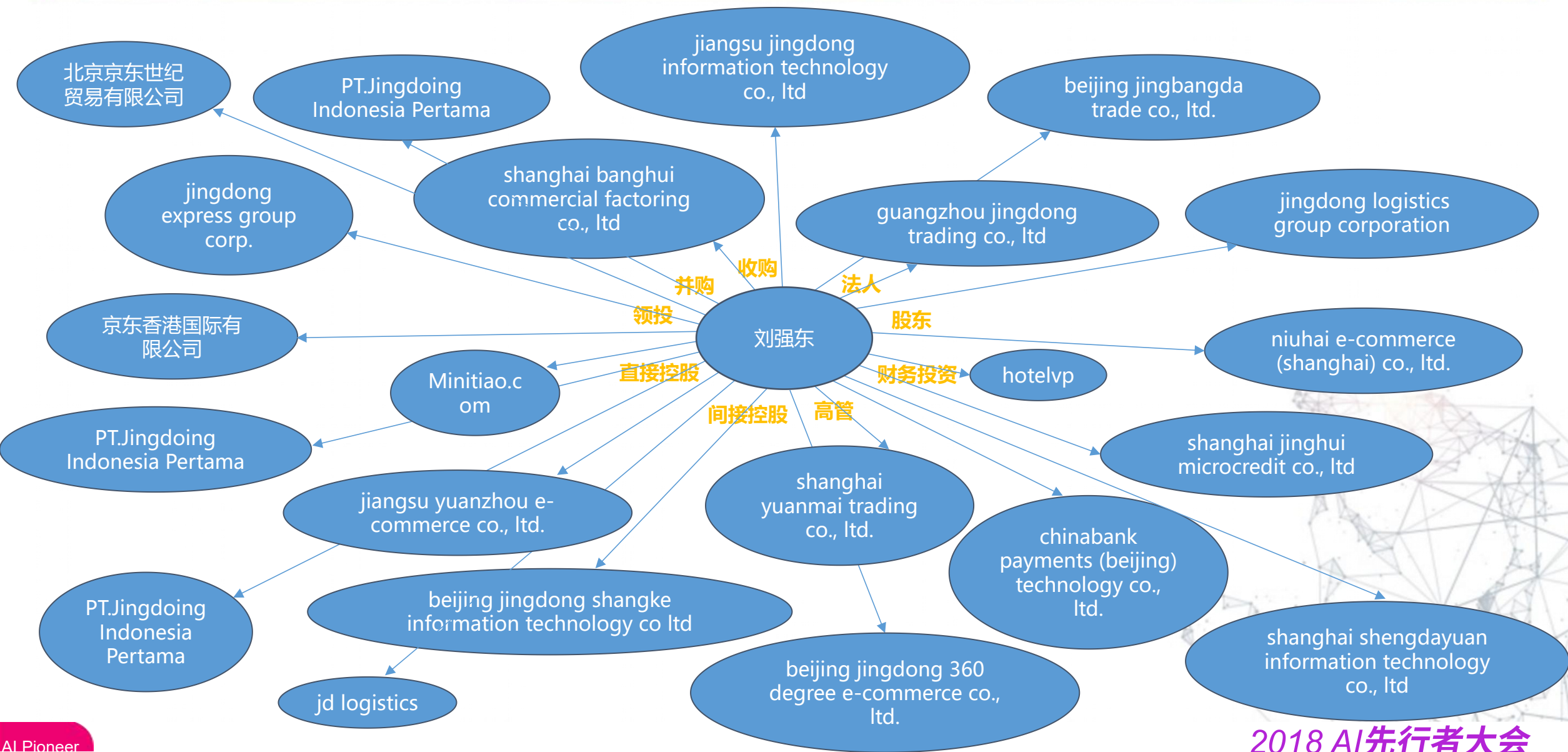
RDF举例 — 某证券“公司与现任高管三元组”

数据建模

- 1、从任职关系表中抽出(group by)高管信息作为图数据库中高管的节点
- 2、将上市公司信息表所有信息作为图数据库中公司的节点
- 3、将任职关系表中的高管个人信息与高管节点进行关联，将任职关系表中的公司代码与公司节点进行关联，构造图数据库中的任职关系

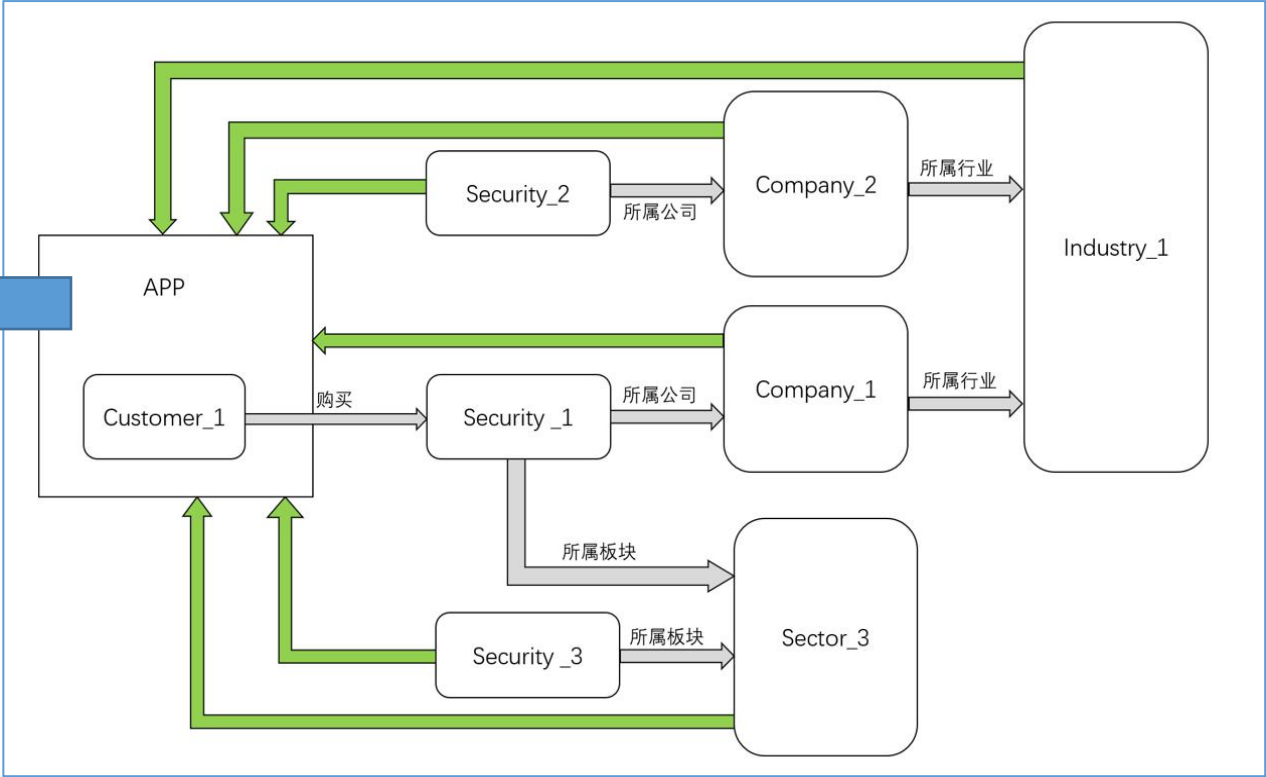
图中含义	节点：高管				关系：任职				节点：公司			
关系数据库中表	lc_manager				lc_leaderposition				lc_stockarchives			
字段及对应关系	Field	Type	Null	抽取+关联	Field	Type	Null	关联	Field	Type	Null	
	1 id	int(11)	NO		1 ID	bigint(20)	NO		1 ID	bigint(20)	NO	
	2 LeaderName	varchar(100)	NO		2 CompanyCode	int(11)	NO		2 CompanyCode	int(11)	NO	
	3 BirthYM	datetime	YES		3 LeaderName	varchar(50)	NO		3 State	int(11)	NO	
					4 BirthYM	datetime	YES		4 SecretaryBD	varchar(20)	YES	
					5 Position	varchar(50)	YES		5 SecuAffairsRepr	varchar(30)	YES	
					6 PositionCode	int(11)	YES		6 AuthReprSBD	varchar(20)	YES	
					7 PositionType	int(11)	YES		7 ContactTel	varchar(60)	YES	
					8 BeginningDate	datetime	YES		8 ContactFax	varchar(60)	YES	
					9 EndDate	datetime	YES		9 ContactEmail	varchar(50)	YES	
					10 ChangeReason	varchar(500)	YES		10 RegAddr	varchar(100)	YES	
					11 ChangeType	int(11)	YES		11 RegZipCode	varchar(6)	YES	
					12 IfPosition	int(11)	YES		12 OfficeAddr	varchar(100)	YES	
					13 UpdateTime	datetime	NO		13 OfficeZipCode	varchar(6)	YES	
					14 JSID	bigint(20)	NO		14 ContactAddr	varchar(100)	YES	
					15 BirthYMInfo	varchar(20)	YES		15 ConatactZipCode	varchar(6)	YES	
					16 CandidacyDate	datetime	YES		16 Email	varchar(50)	YES	
									17 Website	varchar(100)	YES	
									18 DisclosureWeb...	varchar(160)	YES	
									19 DisclosurePap...	varchar(160)	YES	
									20 Establishment...	datetime	YES	

知识表示 – 数据结构 – RDF



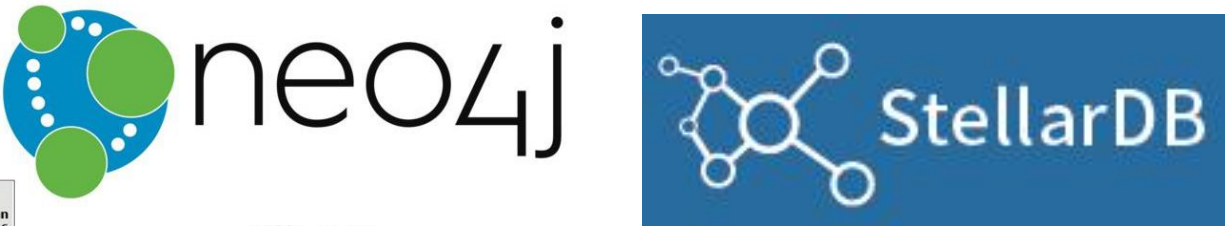
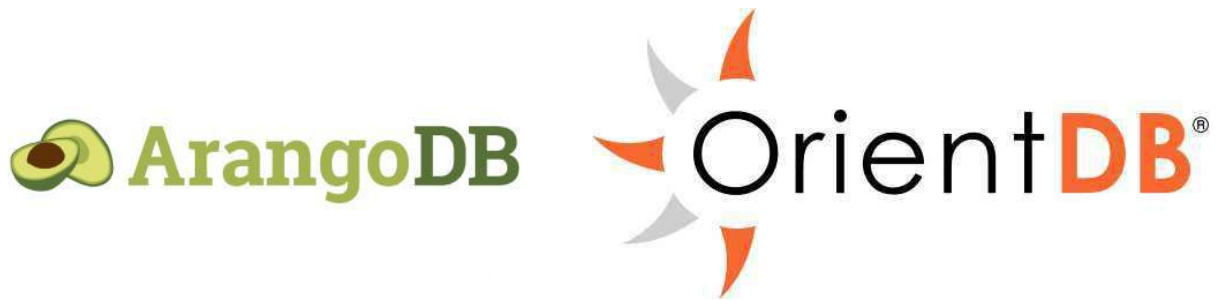
知识表示举例 — 某证券 “消息精准推送”

以某客户购买某只证券为例，通过构建的三元组关系可以查询到与该证券关系深度为n的所有节点及其关系，并选择性向客户推送查询到的节点的更新消息



图计算/图检索

- 兼容SQL，采用SQL语法的简单扩展实现图计算
- 大规模拓扑图分析(100亿+点，10000亿+边)，线性扩展
- 高度容错的计算引擎
- 支持点、边 CRUD
- 秒级图检索查询



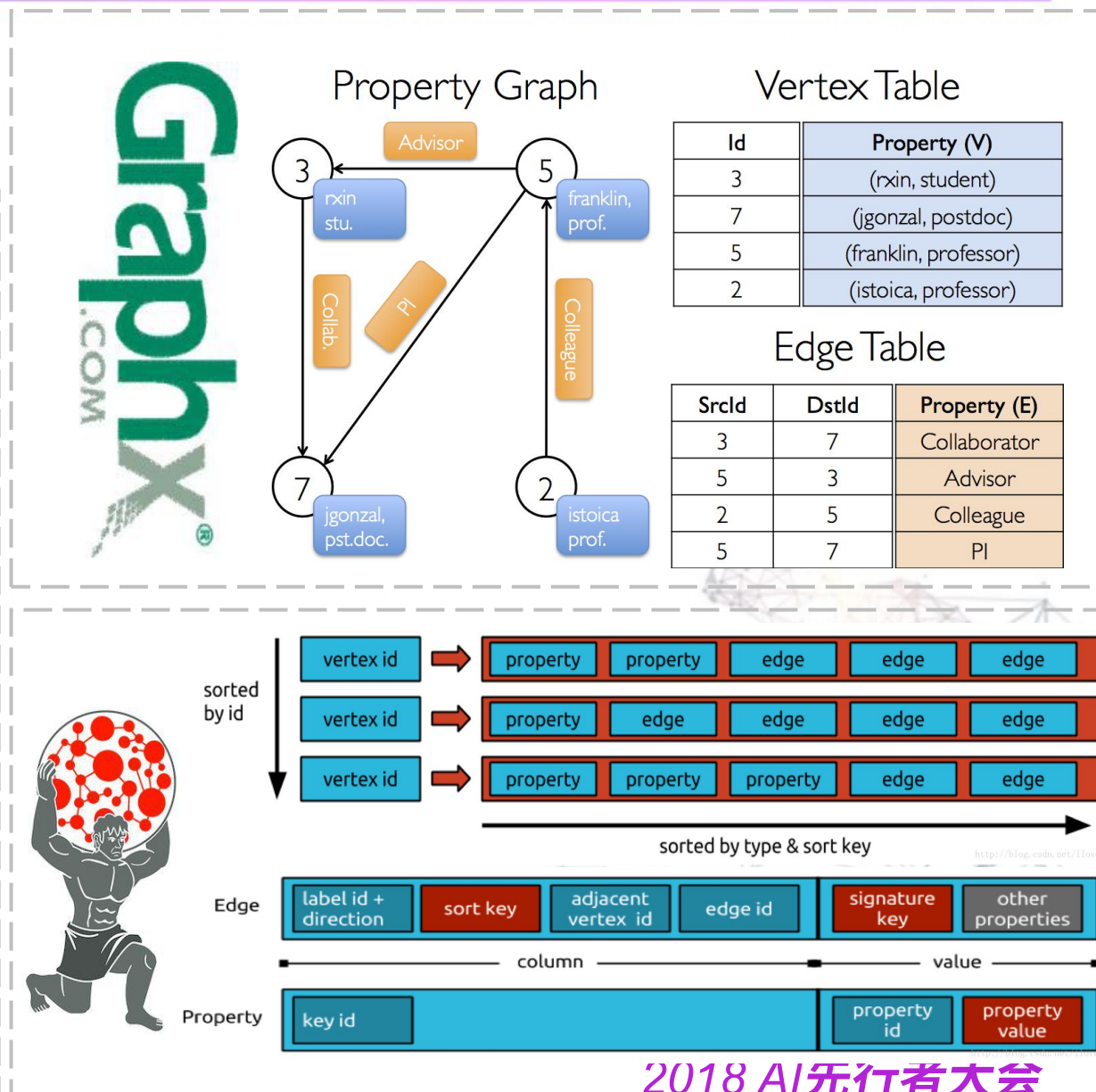
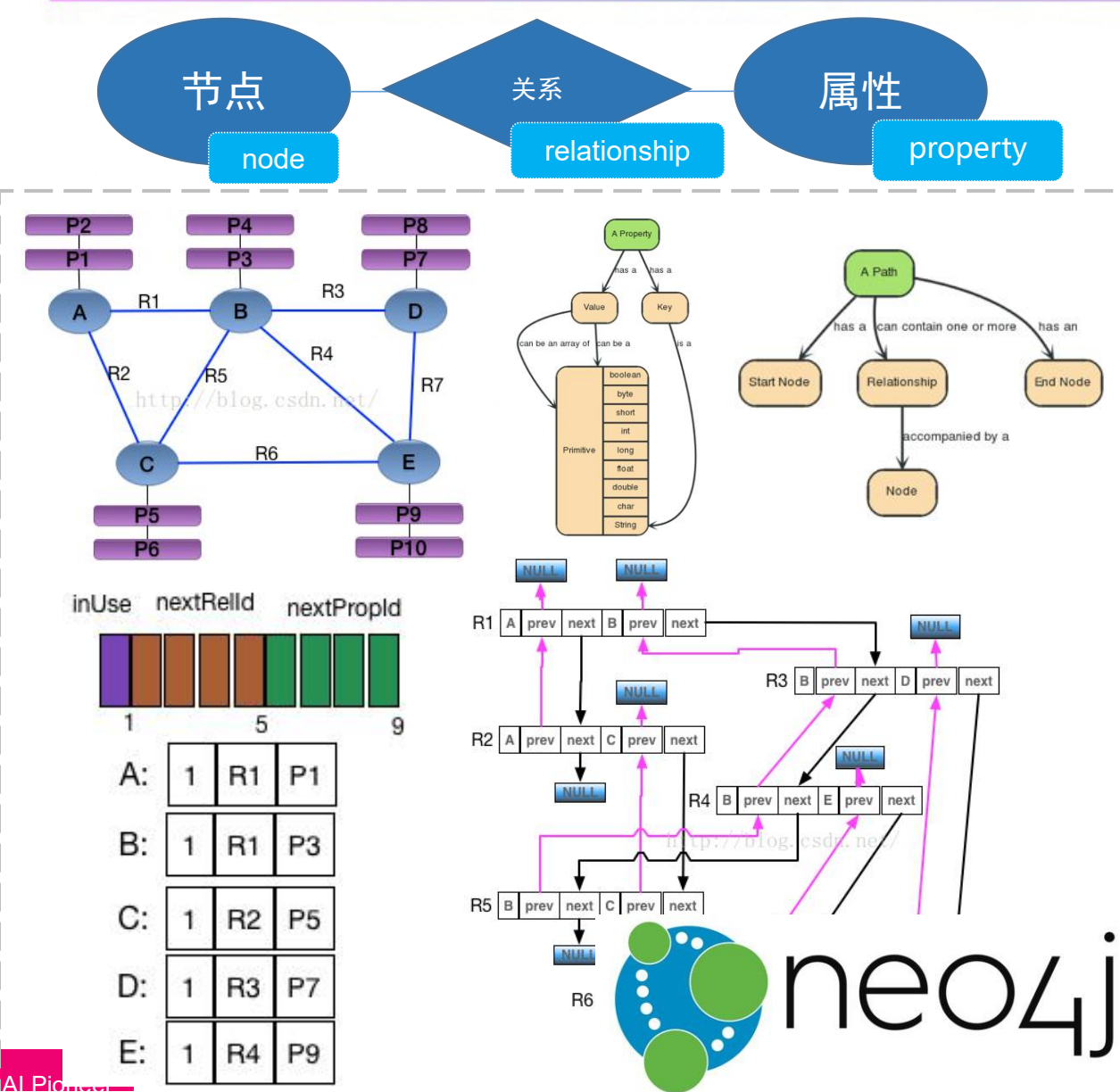
Rank			DBMS	Database Model	Score		
Jan 2017	Dec 2016	Jan 2016			Jan 2017	Dec 2016	Jan 2016
1.	1.	1.	Neo4j	Graph DBMS	36.26	-0.56	+3.26
2.	2.	2.	OrientDB	Multi-model	5.81	-0.06	-0.28
3.	3.	3.	Titan	Graph DBMS	5.50	+0.04	-0.15
4.	5.	5.	ArangoDB	Multi-model	2.41	+0.02	+1.00
5.	4.	4.	Virtuoso	Multi-model	2.37	-0.11	-0.66
6.	6.	6.	Giraph	Graph DBMS	1.04	+0.00	+0.11
7.	7.	8.	Stardog	Multi-model	0.52	+0.00	+0.02
8.	8.	7.	AllegroGraph	Multi-model	0.46	-0.01	-0.24
9.	10.	17.	GraphDB	Multi-model	0.31	+0.09	+0.21
10.	9.	9.	Sqrrl	Multi-model	0.25	-0.02	-0.07
11.	11.	10.	InfiniteGraph	Graph DBMS	0.18	-0.00	-0.12
12.	12.	15.	InfoGrid	Graph DBMS	0.16	+0.02	+0.03
13.	13.		Dgraph	Graph DBMS	0.13	+0.00	
14.	14.	14.	FlockDB	Graph DBMS	0.10	-0.00	-0.05
15.	15.	19.	Blazegraph	Multi-model	0.08	+0.00	+0.07
16.	16.	12.	GlobalsDB	Multi-model	0.06	-0.01	-0.10
17.	17.	13.	HyperGraphDB	Graph DBMS	0.06	-0.01	-0.10
18.	19.	18.	GraphBase	Graph DBMS	0.04	+0.00	-0.03
19.	18.	11.	Sparksee	Graph DBMS	0.03	-0.02	-0.18
20.	20.	16.	VelocityGraph	Graph DBMS	0.00	-0.02	-0.12
21.	21.		AgensGraph	Multi-model	0.00	±0.00	
21.	21.	20.	Amisa Server	Multi-model	0.00	±0.00	±0.00

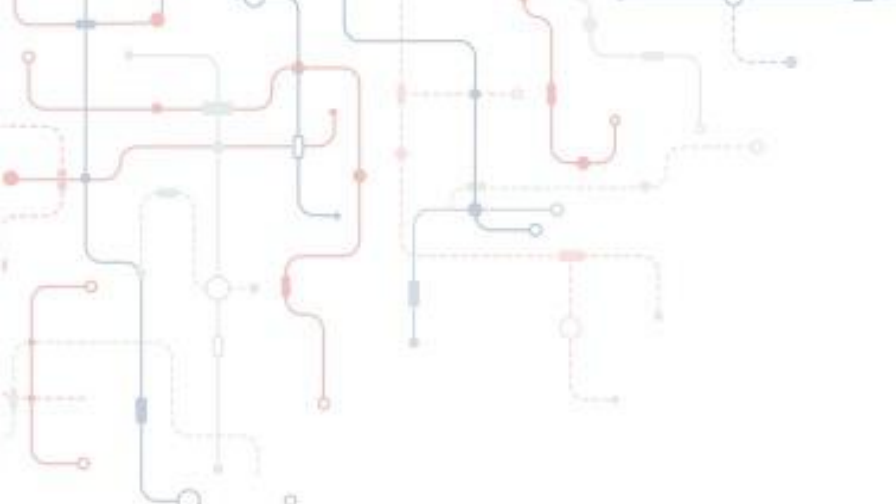
图2 比较流行的基于图存储的数据库排名（截止到2017年1月）



TITAN

知识存储- 图数据库





目录

知识图谱的实现基础

知识图谱的理论基础

知识图谱的技术基础

案例分享

引用

案例场景

搜索引擎

谷歌

百度

搜狗

知识问答

SIRI

微软小冰

小米音箱

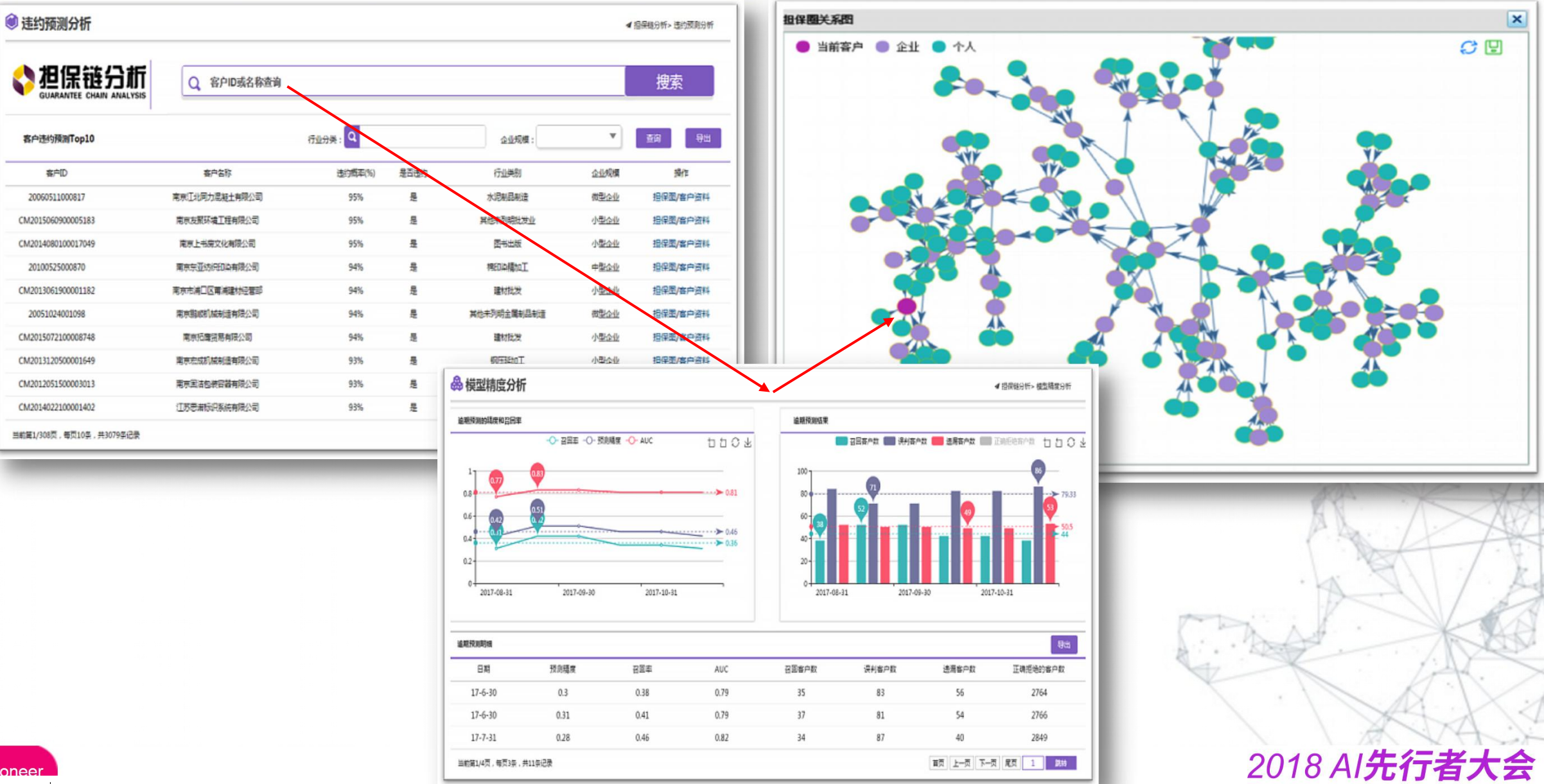
企业知识图谱

企业族谱

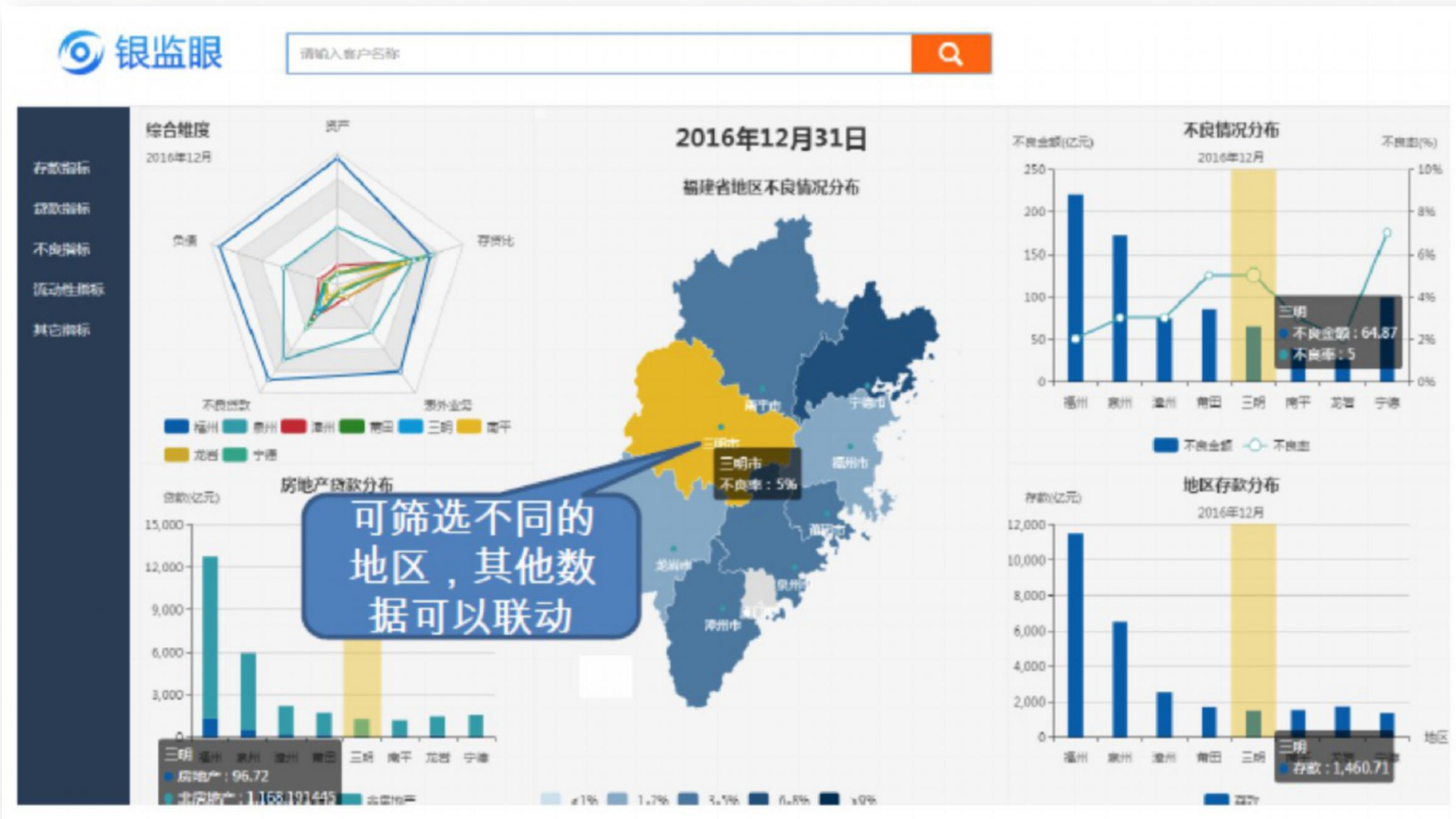
监管科技

智能投研

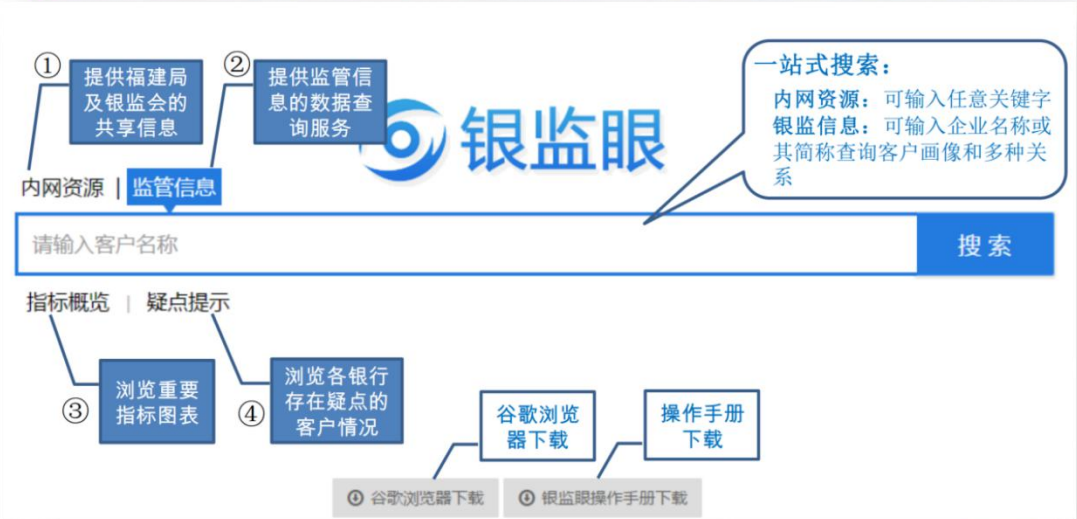
案例：银行对公信贷分析



案例：银监会 — 企业监管图谱



案例：银监会 — 企业监管图谱



客户列表						疑点说明
评分	日期	客户名称	地区名称	银行名称	疑点提示	
130	201612	泉州市丰泽新力鞋业有限公司	福州 泉州	民生银行 泉州银行 华夏银行	以贷还贷 欠息但分类正常 失信被执行	
	201612	福建...		泉州银行	以贷还贷 失信被执行	
	612	福建...		福建海峡银行	以贷还贷 失信被执行	
	612	中联文创控股有限公司	福州 泉州	泉州银行 交通银行 浦发银行	以贷还贷 失信被执行	
	612	英泰(福建)鞋业有限公司	泉州	泉州银行	以贷还贷 失信被执行	
	110	龙岩市创通汽车贸易有限公司	龙岩	泉州银行 兴业银行 福建海峡银行 农业银行	以贷还贷 失信被执行	
	612	福建省南安市立丰石材发展有限公司	福州 泉州	中信银行 光大银行 民生银行 平安银行 泉州银行 招商银行 工商银行	以贷还贷 失信被执行	
	612	本科电器有限公司	福州 泉州	民生银行 泉州银行 兴业银行 福建海峡银行	以贷还贷 失信被执行	
	110	泉州...		中信银行 福建海峡银行	以贷还贷 失信被执行	
	110	福建...		中信银行 泉州银行 招商银行 交通银行	以贷还贷 失信被执行	
	110	龙...		福建海峡银行 农业银行	以贷还贷 失信被执行	
	110	石狮东阳拉链有限公司	福州 泉州	泉州银行 招商银行 建设银行 农业银行	以贷还贷 失信被执行	
	110	南益化工(泉州)有限公司	福州 泉州	泉州银行	以贷还贷 失信被执行	

客户疑点产生的日期

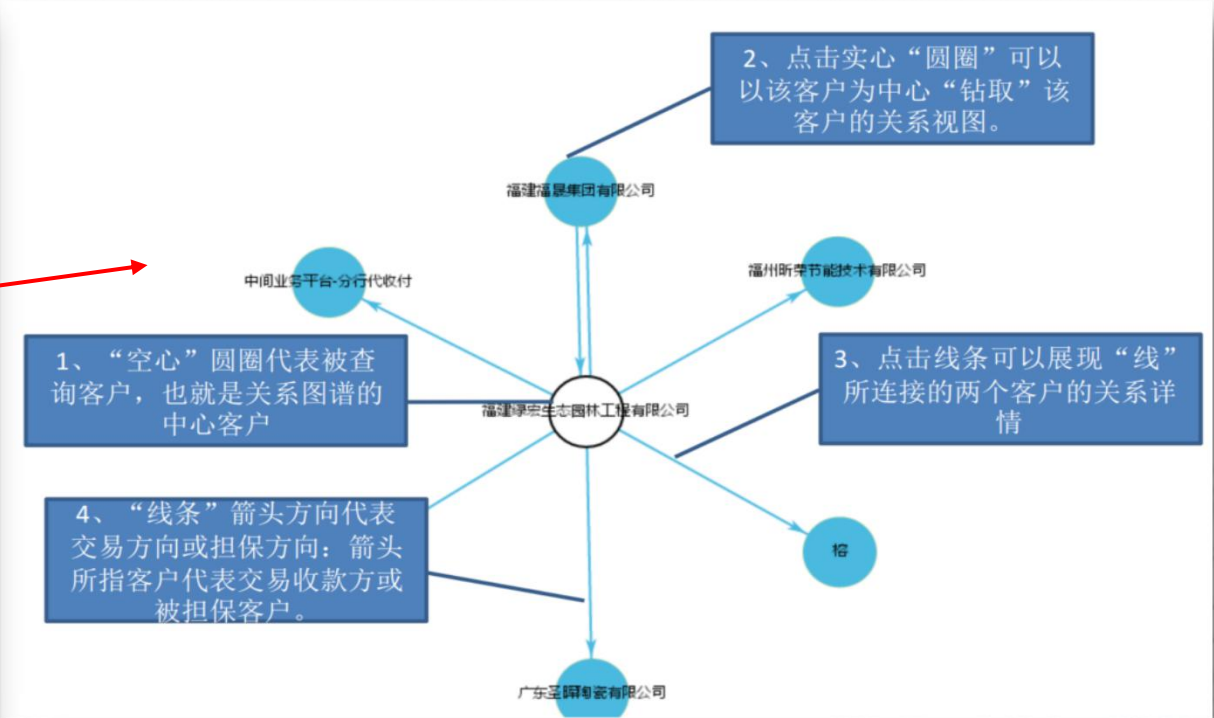
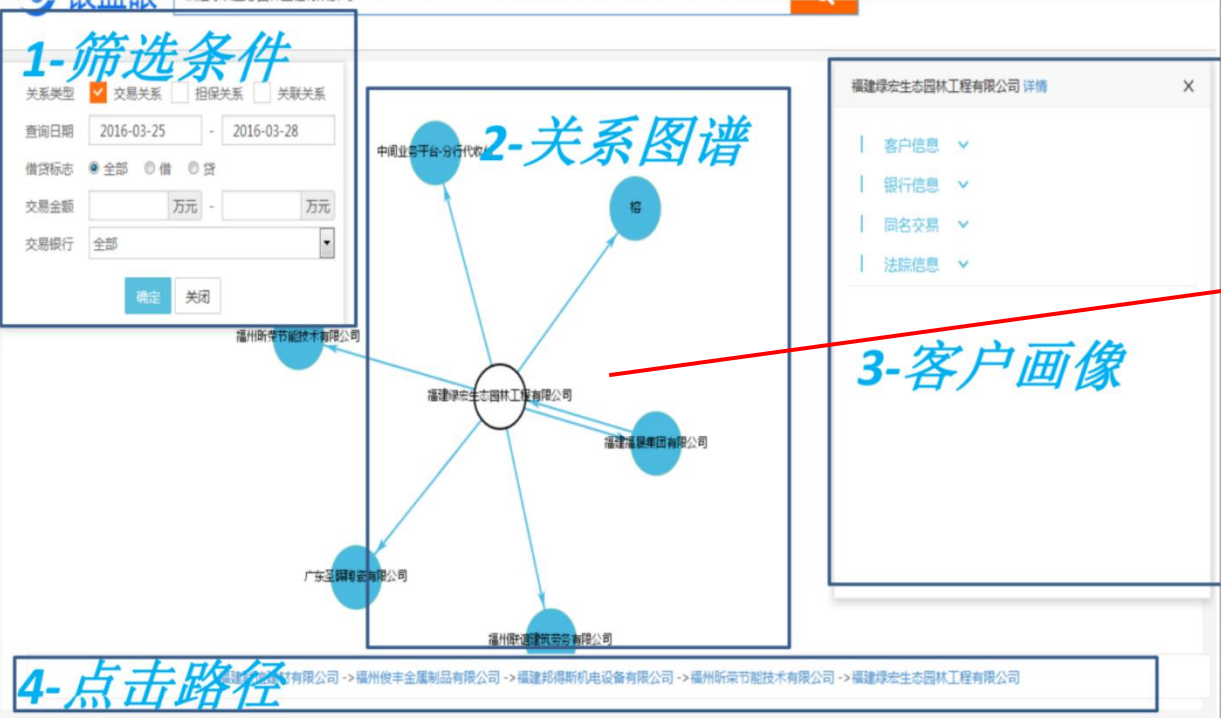
根据疑点的严重程度进行量化，评分规则请查看“疑点说明”

可单选或复选筛选不同的疑点客户类型

点击客户名称进入“关系图谱”

点击可以展现详情，目前只展示“失信被执行人”的疑点详情数据。

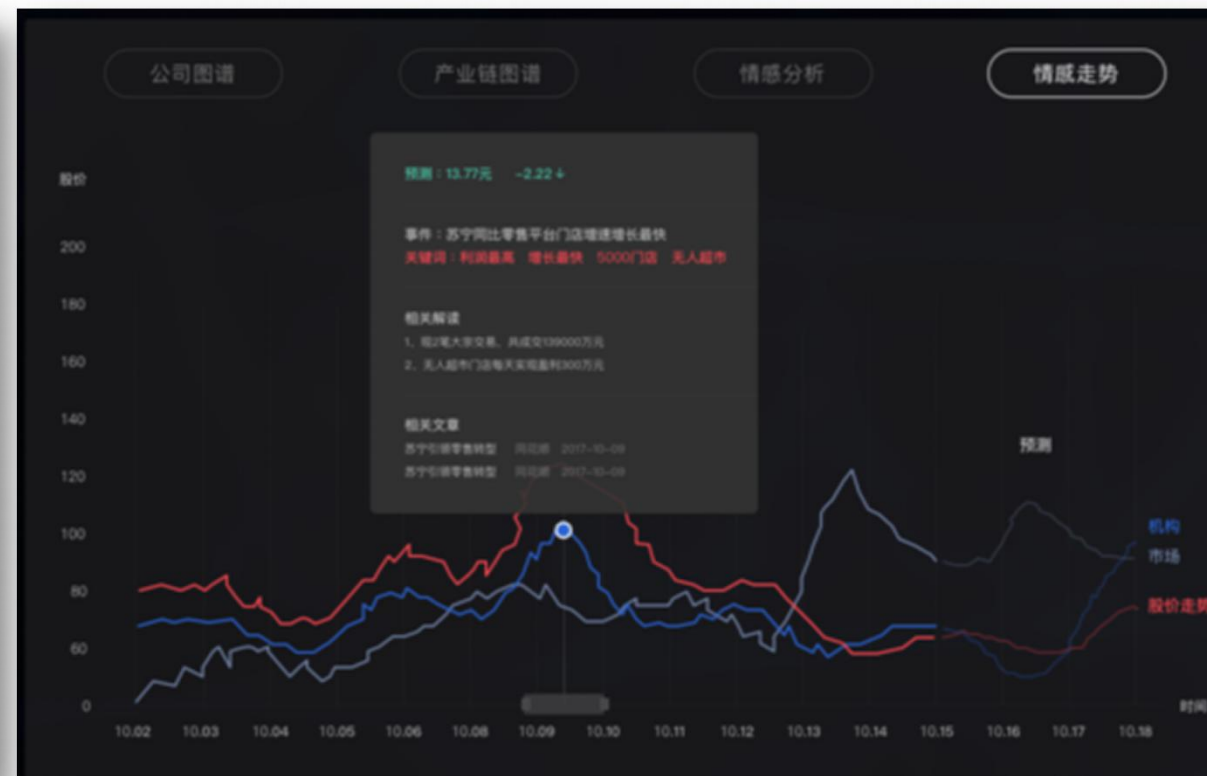
案例：银监会 — 企业监管图谱







智能投研 — 舆情分析



引用

- Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion, Xin Luna Dong, Evgeniy Gabrilovich, Jeremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, Wei Zhang
- 深度解析机器学习五大流派中主算法精髓, Pedro Domingos, 华盛顿大学计算机科学与工程学教授
- The Master Algorithm, Pedro Domingos, 华盛顿大学计算机科学与工程系教授
- 知识图谱与大数据, 陈华钧, 浙江大学计算机科学与技术学院教授



THANKS

