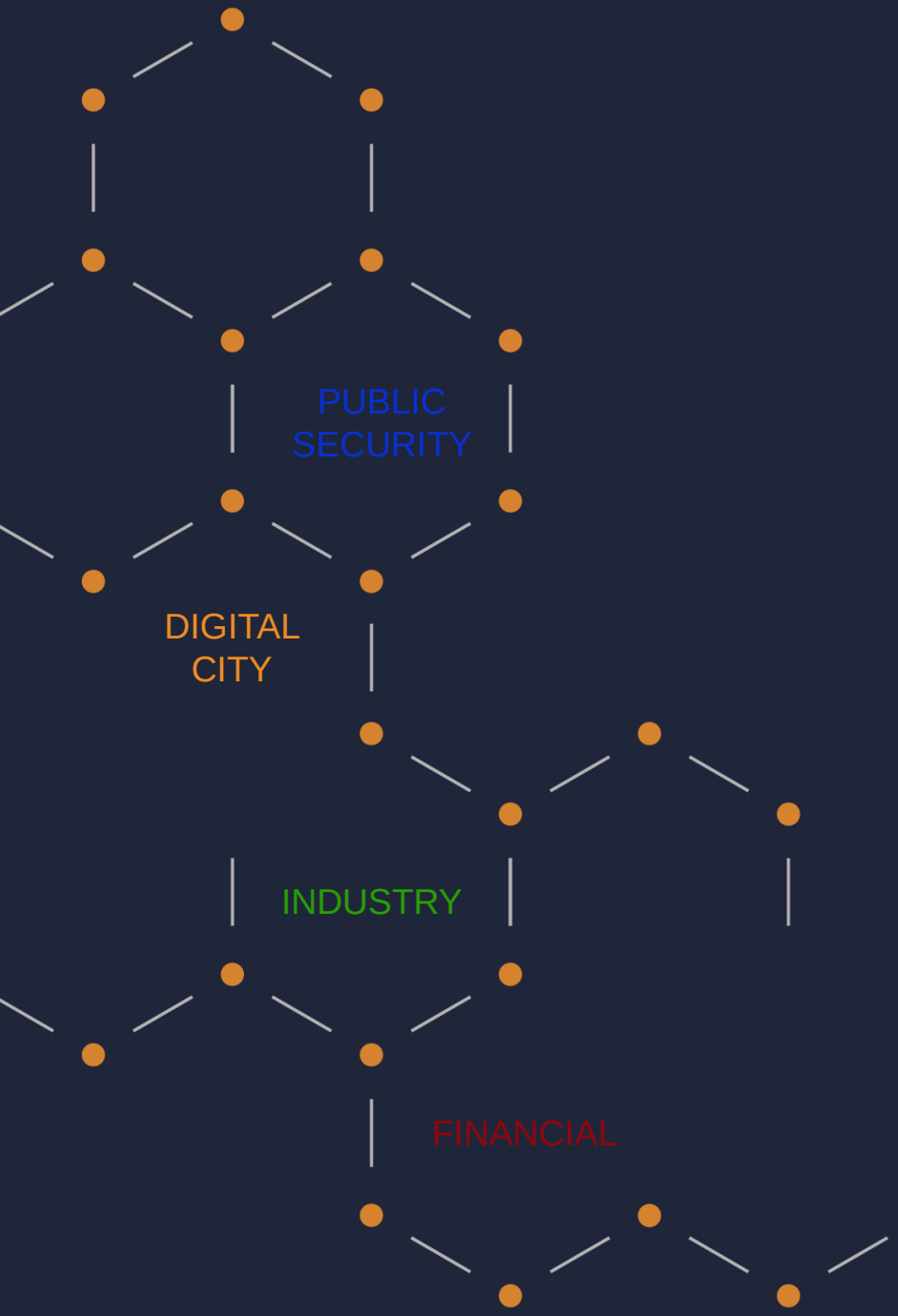


架构迎接未来变化
IAS 2018

MININGLAMP

行业知识图谱构建与应用

于政　博士

MININGLAMP
明　略　数　据

Outline

一、 知识图谱概览

二、 知识图谱关键技术介绍

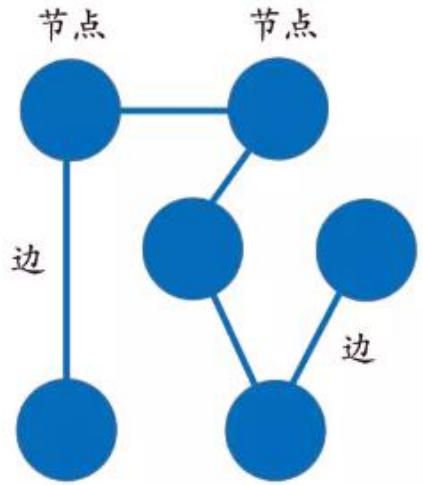
三、 行业知识图谱构建-以汽车行业为例

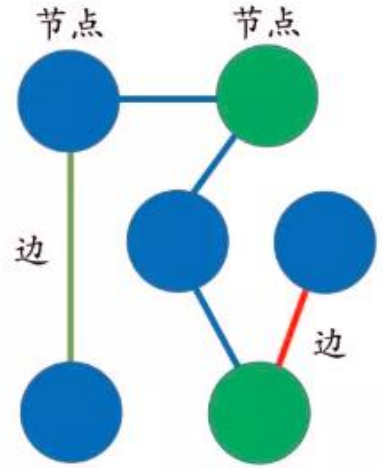
1. 什么是知识图谱

- Knowledge Graph is a **large scale** semantic network
  - Consisting of entities/concepts as well as semantic relationships among them

- 多关系图（Multi-relational Graph）

  - 图（Graph）
    - 节点（Vertex）+ 边（Edge），一种类型的节点和边
  - 多关系图
    - 多种类型的节点和边





2. KG的本质

Web视角	• 像建立文本之间的超链接一样，建立数据之间的语义链接，并支持语义搜索
NLP视角	• 怎样从文本中抽取语义和结构化数据
KR视角	• 怎样利用计算机符号来表示和处理知识
AI视角	• 怎样利用知识库来辅助理解人的语言
DB视角	• 用图的方式去存储知识

做好KG要兼容并蓄，综合利用好KR、NLP、Web、ML、DB等多方面方法和技术

MININGLAMP
明略数据

Outline

一、 知识图谱概览

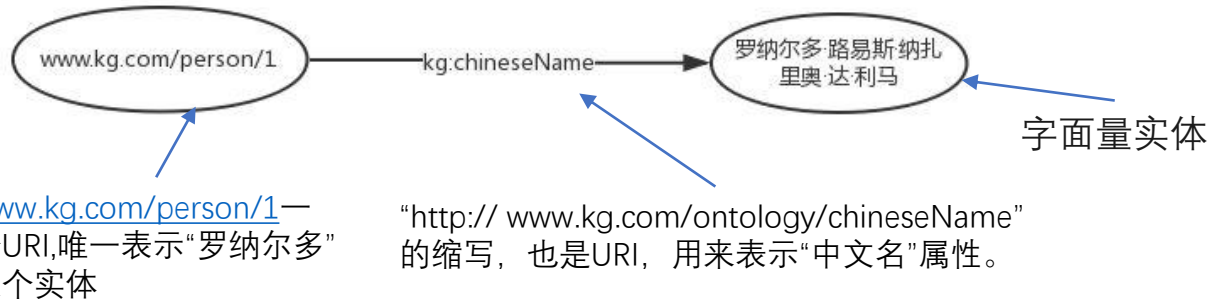
二、 知识图谱关键技术介绍

三、 行业知识图谱构建-以汽车行业为例

1. 知识表示-RDF

- 知识图谱是由一些相互连接的实体和他们的属性构成的。换句话说，知识图谱是由一条条知识组成，每条知识表示为一个SPO三元组(Subject-Predicate-Object)

- 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形式化地表示这种三元关系。RDF(Resource Description Framework)，即资源描述框架，用于描述实体/资源的标准数据模型。

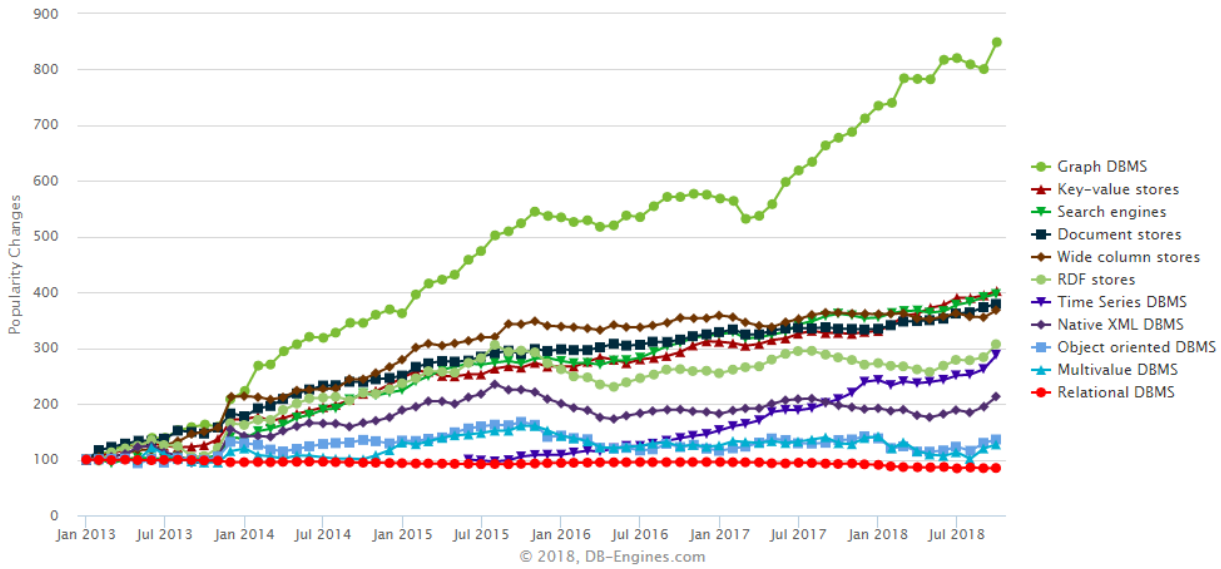
2. 知识存储

- 图结构存储有两种通用的存储方案：**RDF存储** 和 **图数据库(Graph Database)**

	RDF	图数据库
区别	存储三元组（Triple）	节点和关系可以带有属性
	标准的推理引擎	没有标准的推理引擎
	W3C标准	图的遍历效率高
	易于发布数据	事物管理
	多数为学术界场景	基本为工业界场景

Rank Oct 2018	Rank Oct 2017	DBMS	Database Model
22	21	**Neo4j**	**图**
37	34	**MarkLogic**	**XML, RDF**
135	220	**JanusGraph(原Titan)**	**图**
51	45	**OrientDB**	**图，文档**
88	89	**Jena**	**RDF**

Complete trend, starting with January 2013



Rank Oct 2018	Rank Sep 2018	Rank Oct 2017	DBMS	Database Model
1.	1.	1.	Neo4j	Graph DBMS
2.	2.	2.	Microsoft Azure Cosmos DB	Multi-model
3.	3.		Datastax Enterprise	Multi-model
4.	4.	↓ 3.	OrientDB	Multi-model
5.	5.	5.	ArangoDB	Multi-model
6.	6.	6.	Virtuoso	Multi-model
7.	↑ 8.	7.	Giraph	Graph DBMS
8.	↓ 7.		Amazon Neptune	Multi-model
9.	9.	↑ 15.	JanusGraph	Graph DBMS
10.	↑ 11.	↓ 9.	AllegroGraph	Multi-model
11.	↓ 10.	↓ 8.	GraphDB	Multi-model
12.	12.	↓ 10.	Stardog	Multi-model
13.	13.	↑ 16.	Dgraph	Graph DBMS
14.	↑ 16.	14.	Graph Engine	Multi-model
15.	↓ 14.	↓ 13.	Blazegraph	Multi-model

2. 知识存储

数据库选型

- 如果数据量比较小，且图谱节点之间关系较少且不太需要多跳查询，mysql就能搞定。
- 如果数据量比较小，且会进行多跳查询，比如“姚明老婆的女儿叫什么？”，这种用图数据neo4j比较合适（它的开源版本是单机版，如果要支持分布式需要用收费版）
- 如果数据量比较大，且会进行多跳查询，可以考虑用titan等分布式图数据库。
- RDF存储，比如Jena，偏学术研究，工业界应用不多。
- PS：有时候单一的存储无法满足所有需求，可能需要几个系统搭配使用。比如使用图数据库来查多跳查询，使用elasticsearch来进行模糊搜索和相关度排序。

3. 知识抽取

I. 实体识别、链接

II. 关系抽取

III. 事件知识学习

- **实体识别**：命名实体识别的目的是识别文本中指定类别的实体，主要包括人名、地名、机构名、专有名词等的任务。

- **实体链接**：将实体提及与知识库中对应实体进行链接。如“在旧金山的发布会上，**苹果**为开发者推出新编程语言 Swift”，**苹果 -> 苹果公司**。

- **关系抽取**：自动从文本中检测和识别出实体之间具有的某种语义关系，输出结果通常是一个三元组(实体 1，关系，实体 2)。如句子“北京是中国的首都、政治中心和文化中心”中表述的关系可以表示为（中国，首都，北京）。

- **事件学习**：
  - **事件识别和抽取**：研究如何从描述事件信息的文本中识别并抽取出事件信息并以结构化的形式呈现出来，包括其发生的时间、地点、参与角色以及与之相关的动作或者状态的改变
  - **事件检测与追踪**：将文本新闻流按照其报道的事件进行组织，以便让用户了解新闻及其发展。

3.1 关系提取

➢ 关系抽取是从文本中抽取出两个或多个实体之间的语义关系。

- 基于模板的方法
  - 触发词的Pattern,
  - 依存句法分析的Pattern
- 基于监督学习的方法
  - 机器学习方法
  - 深度学习
- 弱监督学习的方法
  - 远程监督
  - Bootstrapping

3.1 关系提取

➢ 基于模板的方法

基于模板的方法在小规模数据集上容易实现且构建简单，缺点为难以维护、可移植性差、模板有可能需要专家构建。

□ 基于触发词的Pattern

预先定义一套种子模板，

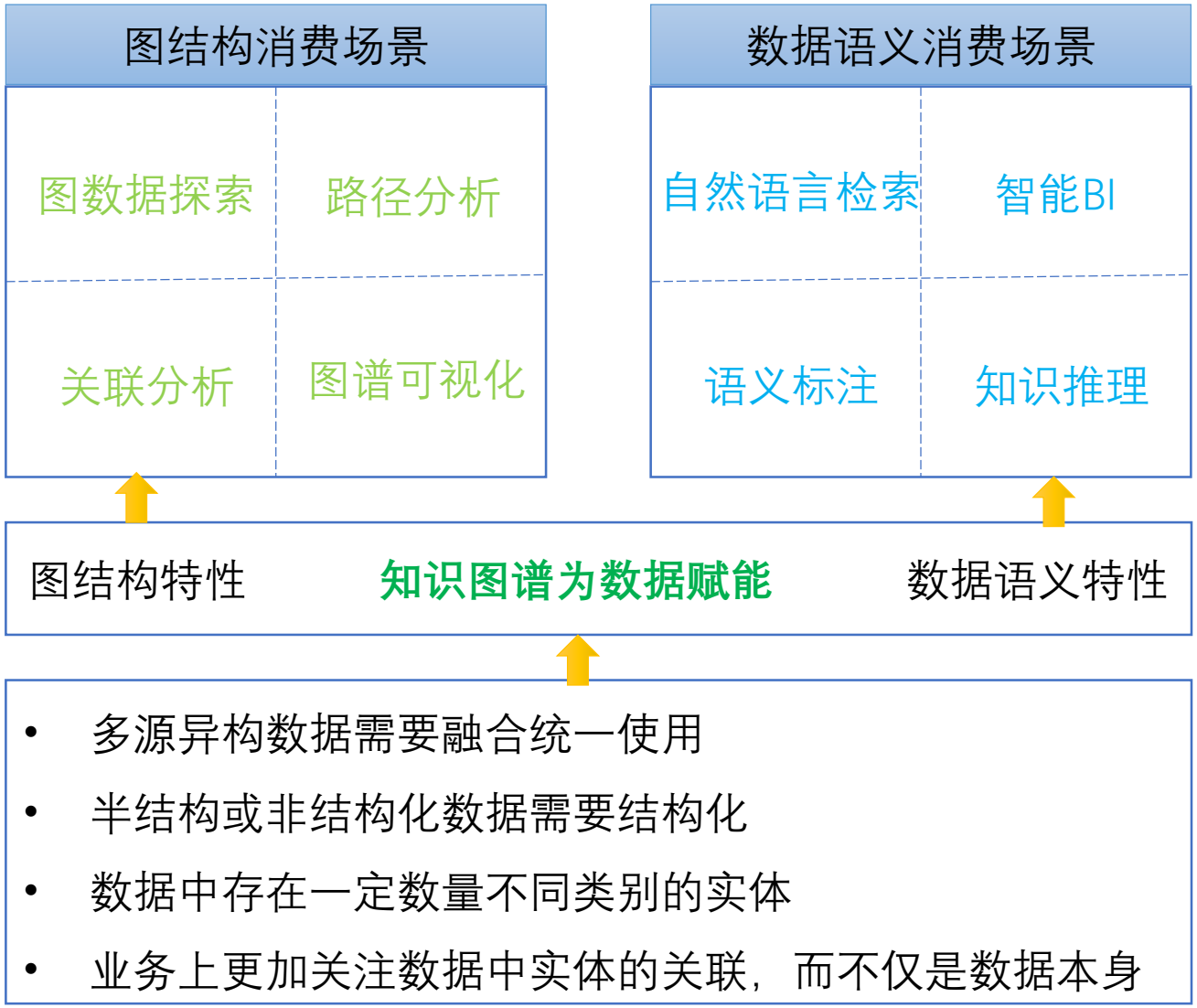
比如

姚明 老婆 叶莉， 徐峥 老婆 陶虹

-> X 老婆 Y

Hearst pattern	Example occurrences
X and other Y	...temples, treasuries, and other important civic buildings.
X or other Y	bruises, wounds, broken bones or other injuries...
Y such as X	The bow lute, such as the Bambara ndang...
such Y as X	...such authors as Herrick, Goldsmith, and Shakespeare.
Y including X	...common-law countries, including Canada and England...
Y, especially X	European countries, especially France, England, and Spain...

4. 知识图谱应用场景



- 不适用的数据场景：

① 通常的二进制数据

② 日志数据

③ 流式数据

- 不适用的消费场景：

① 数据统计

② 数据计算

MININGLAMP 明略数据

4. 知识图谱应用场景

用更简单的方式	选择知识图谱
对可视化需求不高	有强烈的可视化需求
很少涉及到关系的深度搜索	经常涉及到关系的搜索
关系查询效率要求不高	对关系查询效率有实时性要求
数据缺乏多样性	数据多样化、解决数据孤岛问题
暂时没有人力或者成本不够	有能力、有成本搭建系统

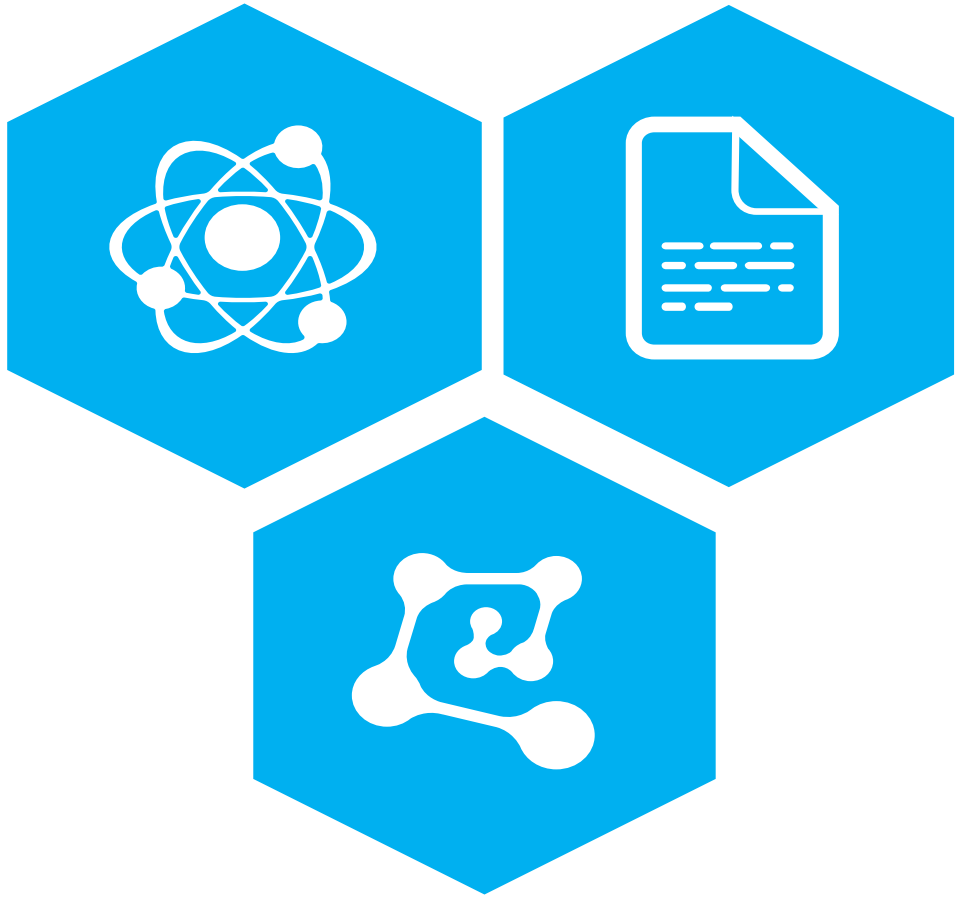
Outline

一、 知识图谱概览

二、 知识图谱关键技术介绍

三、 **行业知识图谱构建-以汽车行业为例**

1. 案例背景



4S店发送维修疑难案例寻求厂商售后技术支持

海量维修案例数据分析

技术支持员工工作负荷大

2. 遇到的问题及挑战

- 4S店寻求厂家支持的问题量大
- 疑难杂症种类繁多

- 维修案例语言表达方式多样
- 无法全面提取维修案例中所有细节

- 无法快速准确定位到问题
- 对员工的专业性要求极高

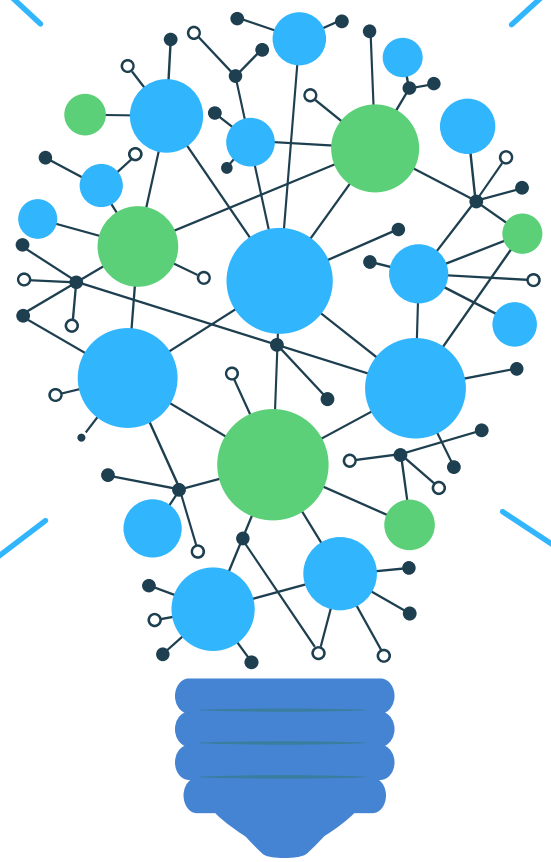
3. 解决方案

技术实现

通过大数据、知识图谱、自然语言处理、机器学习等技术从售后维修案例数据中成功挖掘知识，归纳总结案例问题、维修经验等

基于知识图谱构建知识应用系统，并在检索、问题分类、语义理解等方面展开应用

实现对零部件、故障、工况、维修方法等实体、关系构成的图谱进行图挖掘计算，发现故障图模式及关联关系，并进行预测



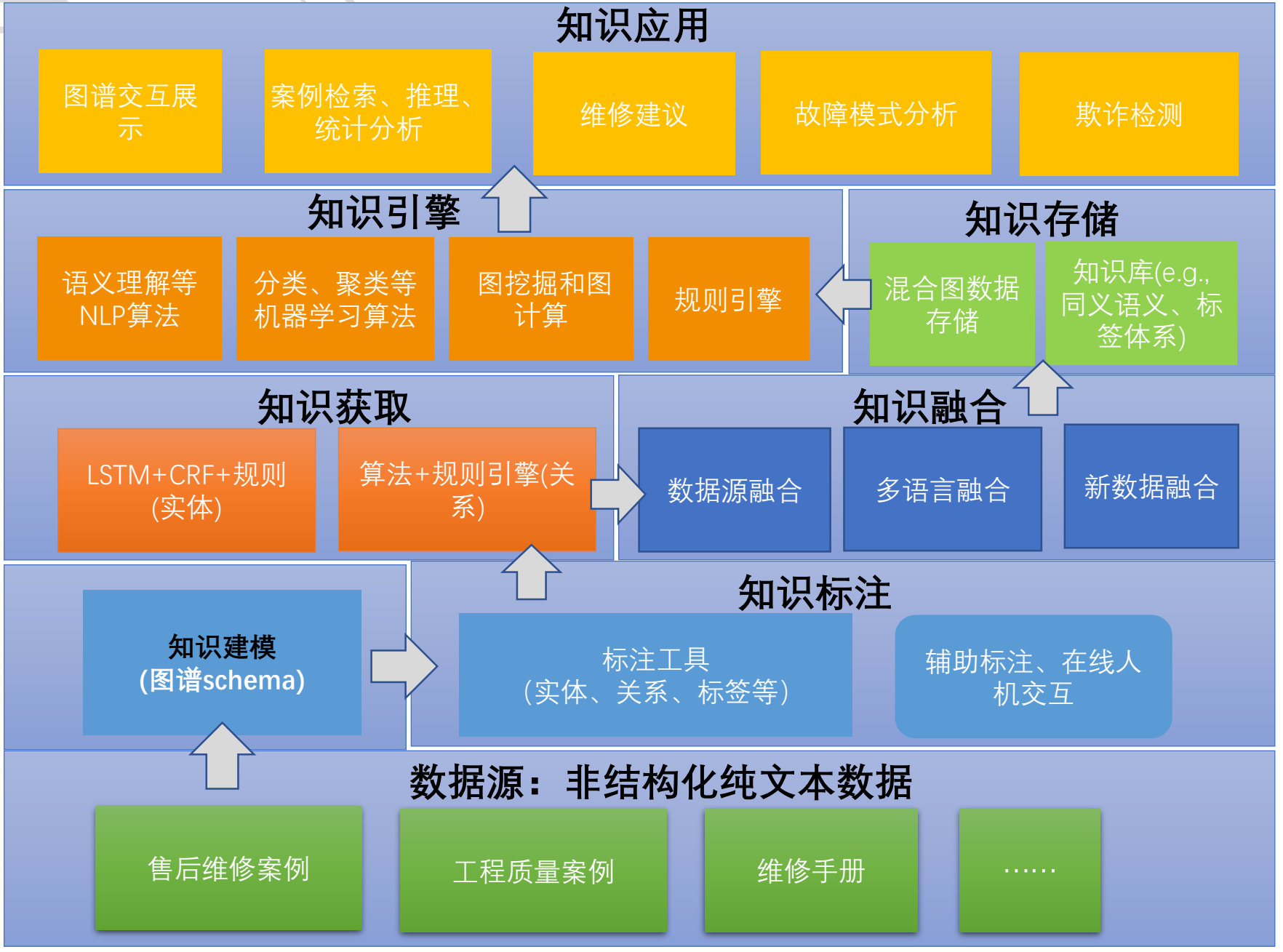
达到效果

将散落的专家知识及系统自动识别的维修方法固化到系统中，减少对人的依赖

给4S店提供一套专业的维修问答系统，用人工智能代替部分人的工作

专家的精力主要用于处理棘手问题，提升品牌美誉度并为设计及制造部门提供持续的质量改善建议

4. 汽车行业知识图谱应用架构



6. 知识建模-图谱Schema定义

- Schema定义:

知识图谱schema相当于领域内的数据模型，定义了实体类型、属性、关系等，具体应用场景、产品需求及数据决定了图谱schema的构建方式。

- 作用:

高质量、标准化的 schema 能有效降低领域数据之间对接的成本，便于图谱扩展及应用开发;

针对该维修案例数据和业务场景，设计图谱schema共包含20+种实体，10+种关系

MININGLAMP 明略数据

6. 知识建模-图谱Schema定义

实体示例：

- Case
  Case编号
- 故障
- 检修方法
- 维修结果
- 无维修方案
- 零部件
- 失效模式
- 正常模式
- 故障灯
- 故障码
- 故障码含义

关系属性示例：

- hasFailure
  主体：Case编号
  受体：故障
  属性1：程度
  属性2：方位
- hasMethod
  主体：Case编号
  受体：检修方法
  属性1：结果成功与否
- hasMode
  主体：Case编号
  受体：工况
- hasPart
  主体：故障(统)
  受体：零部件
- hasLight
  主体：Case编号
  受体：故障灯

MININGLAMP 明略数据

7. 数据标注

➢ 目标及意义：

基于图谱schema定义，使用标注工具对非结构化文本数据进行实体、关系等知识标注，后续将标注样本作为训练数据集用以训练知识抽取模型。

- **标注规范**
  - ✓ 包含20+种实体，10+种关系
  - ✓ 经过7个版本迭代，从粗略到细化再简化，逐步明确、清晰

- **辅助标注**
  - ✓ 使用模型预测结果用于辅助标注，形成闭环
  - ✓ 每条案例的标注时间从4分钟降到1分半

7.1 数据标注-标注样本

人工标注

1 失效模式
-11款1.8L ABC 换档冲击很大

2 失效模式 hasTarget 零部件 检修方法 零部件 工况 hasFailure 工况 失效模式 零部件 程度 工况 失效模式 程度
用户反应车辆挂档发冲，用户反应车辆5月份时在西安4S店修过变速箱更换了变速箱模块，试车挂D档和倒档变速箱冲击很大，在行驶中3档换4档换档冲击很大

3 故障码 hasNorm hasNorm hasTarget 检修方法 零部件 正常模式 正常模式 hasResult hasTarget hasTarget 零部件 hasResult 检修方法 维修结果 hasResult 检修方法 维修结果 hasTarget 检修方法 零部件 hasFailure 失效模式 正常模式 hasTarget 检修方法 零部件 hasResult hasTarget 检修方法 零部件 维修结果
1.检查车辆GDS诊断无相关故障代码，检查变速箱油位正常，无铁屑，对变速箱模块进行编程显示最新程序，配置学习后试车故障未排除 2.解体变速箱检查3-5-倒档离合器片烧蚀，无其它异常，更换3-5-倒档离合器片，清洗控制阀体，试车故障未排除

4 hasResult 维修结果 hasTarget 检修方法 零部件 检修方法 hasTarget 检修方法 零部件
3.与同款车辆对换变速箱模块试车故障排除，更换变速箱模块

5 hasTarget 检修方法 零部件
更换变速箱模块

模型标注

1 失效模式
15年ABC挂档不走车

2 工况 失效模式 工况 失效模式
挂倒挡不走车前进挡加速无力

3 hasTarget 检修方法 正常模式 零部件 方位 检修方法 inPosition 方位 失效模式 hasFailure 零部件 失效模式
检查外部无异常，拆解变速器发现内部进冷却液导致零件损坏

4 hasTarget 检修方法 零部件
更换损坏的内部零件

8. 知识抽取

- **目标及意义：**

结合机器学习模型与规则方法，基于标注样本数据，训练模型和构建规则引擎，使其具备从新数据中抽取相应实体和关系的能力，从而代替人工标注过程，并具备不断迭代优化模型效果的能力。

- **实体**
  - 思路

    BI-LSTM-CRF + 规则引擎

- **关系**
  - 思路

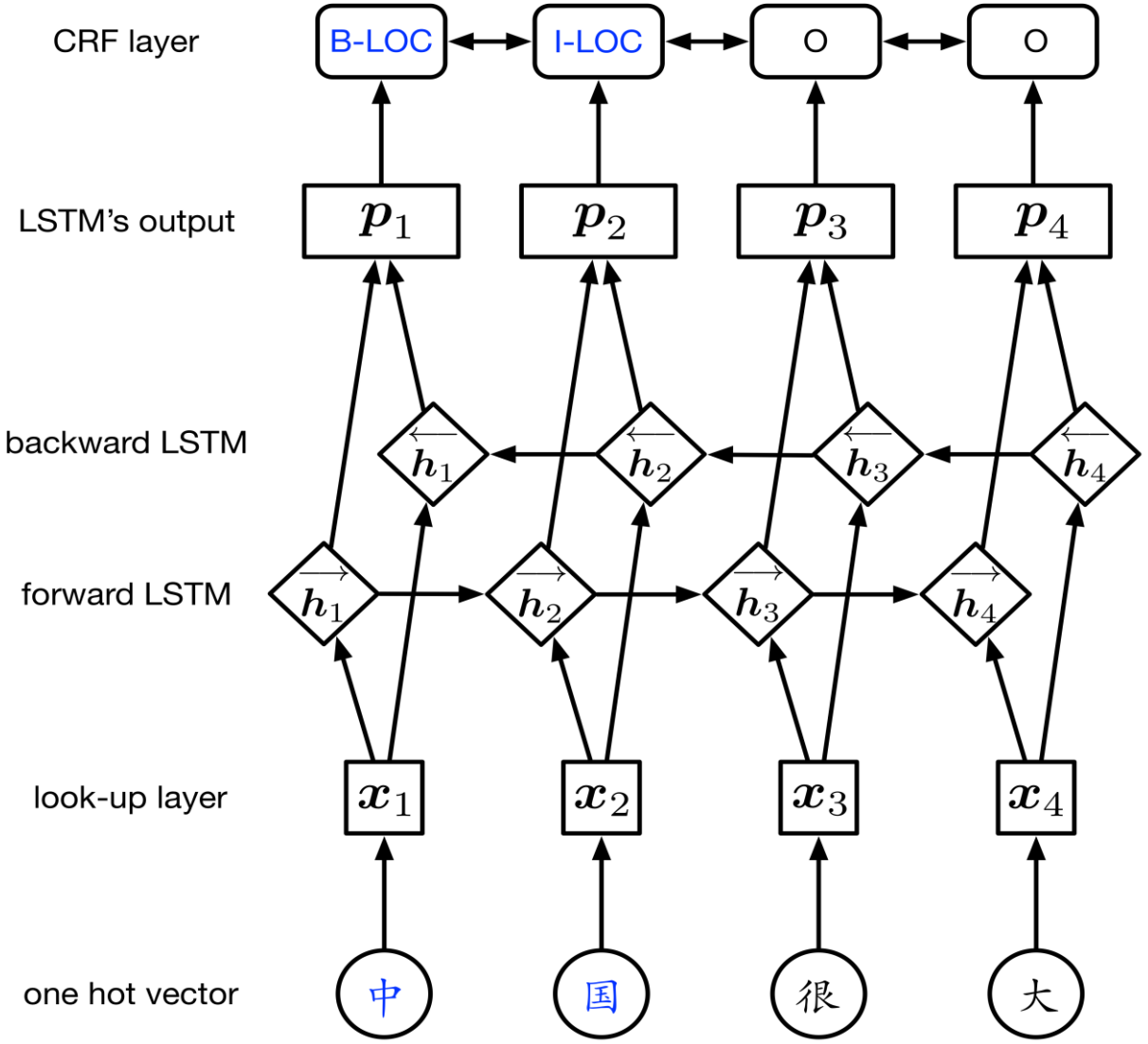
    通过对数据的观察，发现关系比较规律，适合用规则提取

    规则引擎（语义关系+行业用语特点）

8.1 知识抽取模型

架构迎接未来变化
IAS 2018

实体抽取模型架构



MININGLAMP
明略数据

8.2 知识抽取结果

- **效果**

分类			精确率	召回率	F1
(训练2377/测试480)	实体（22）	零部件	0.83	0.86	0.84
		失效模式	0.75	0.78	0.77
		整体	**0.75**	0.79	0.77
	关系（16）	基于预测实体	0.72	0.48	0.58
		基于标注实体	0.95	0.84	0.89

8.2 知识抽取结果

1 15年ABC 挂档不走车

失效模式

2 挂倒挡不走车前进挡加速无力

工况 失效模式 工况 失效模式

3 检查外部无异常，拆解变速器发现内部进冷却液导致零件损坏

hasTarget 检修方法 零部件 正常模式 方位

InPosition 方位 检修方法

hasFailure 零部件 失效模式 失效模式

4 更换损坏的内部零件

hasTarget 检修方法 零部件

8.2 知识抽取结果

- **失效模式同义词**

➢ 方案

聚类辅助人工标注，分类模型（F1: 0.95）

➢ 示例

失效模式	说法	失效模式	说法
加速无力	不给油	异响	变速器声音大
	不能升档		噌噌的响声
	车辆跑不动		咕咚一声
	车速受限制		嗒嗒异响
	发动机转速空转		嗡嗡声
	加不起速度		卡啦卡啦异响
	……		……

8.2 知识抽取结果

- **工况分类**

➢ 方案

聚类辅助人工标注，分类模型（F1: 0.87）

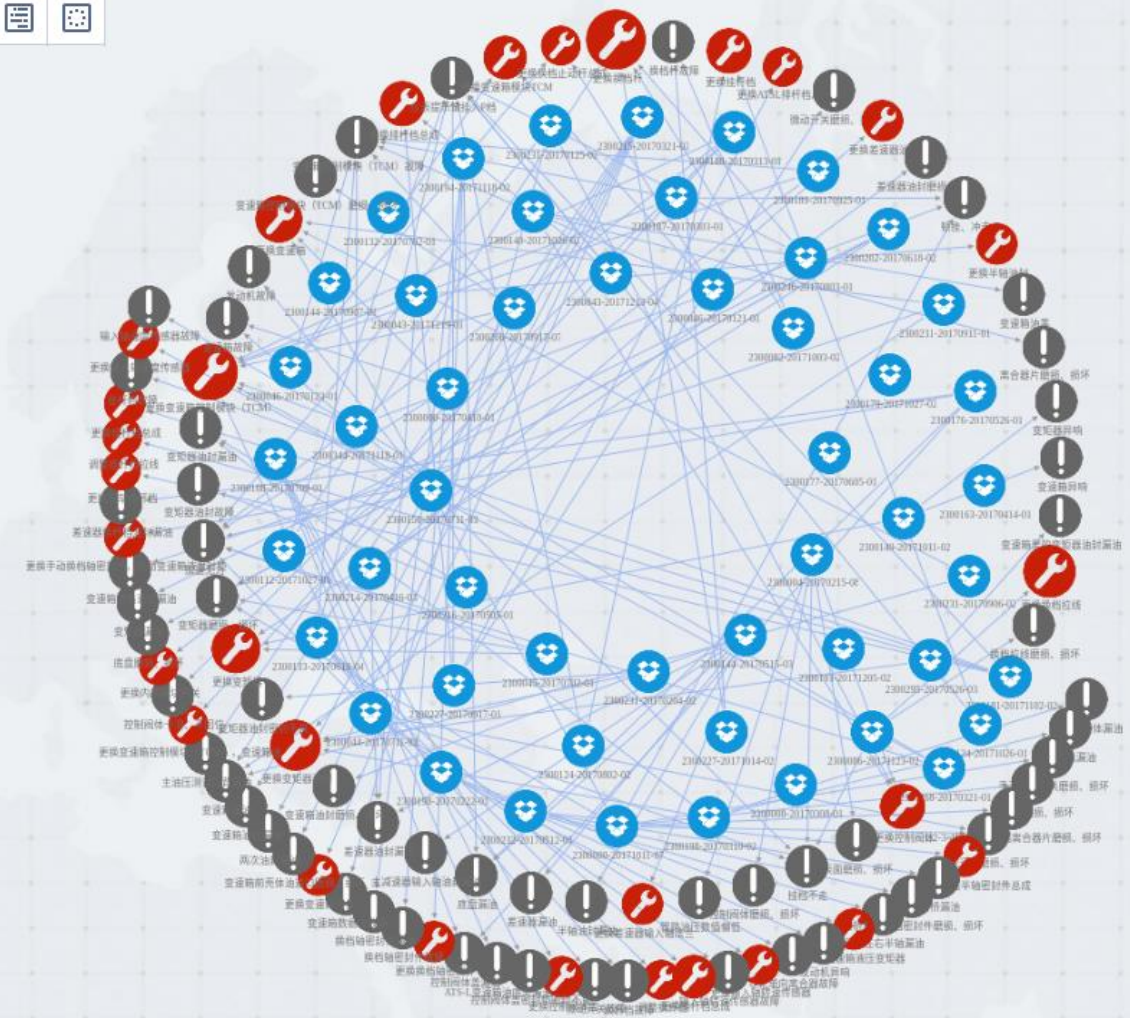
➢ 示例

工况	说法	工况	说法
车速	加速至20码	档位工况	挂D挡
	正常行驶40码时		挂5挡
	低速行车		挂前进档
	低速滑行		挂入倒挡后
	55km/h开始平稳加油门时		3挡行驶时撤油后3挡滑行时
	5-40码		挂杆切换到倒挡时
	……		……

9. 图谱应用--语义检索（同车型所有故障案例）

检索　推理　信息统计

车型: ABC　故障: Ex: 离合器　Ex: 不回位　维修部件: Ex: TCM　工况: Ex: 挂入R档　故障码: Ex: P0700　查 询



- 案例编号
- VIN
- 故障
- 维修方法
- 零部件
- 车型
- 工况
- 故障灯
- 仪表提示
- 故障码
- 故障码含义
- 发动机型号
- 变速箱型号

导航器

ABC　(总计:104 成功:85 失败:19)

维修方法	尝试案例	成功	失败	占比	成功率
更换换档杆	11	11	0	10.58%	100.00%
更换变速箱控制模...	10	9	1	8.65%	90.00%
更换换档拉线	7	7	0	6.73%	100.00%
更换变矩器油封	7	6	1	5.77%	85.71%
更换变矩器	6	5	1	4.81%	83.33%
更换变速箱	4	4	0	3.85%	100.00%
更换挂杆档	3	3	0	2.88%	100.00%
更换排杆档总成	3	3	0	2.88%	100.00%
更换控制阀体	4	3	1	2.88%	75.00%
更换换杆档总成	2	2	0	1.92%	100.00%
更换排杆档	2	2	0	1.92%	100.00%

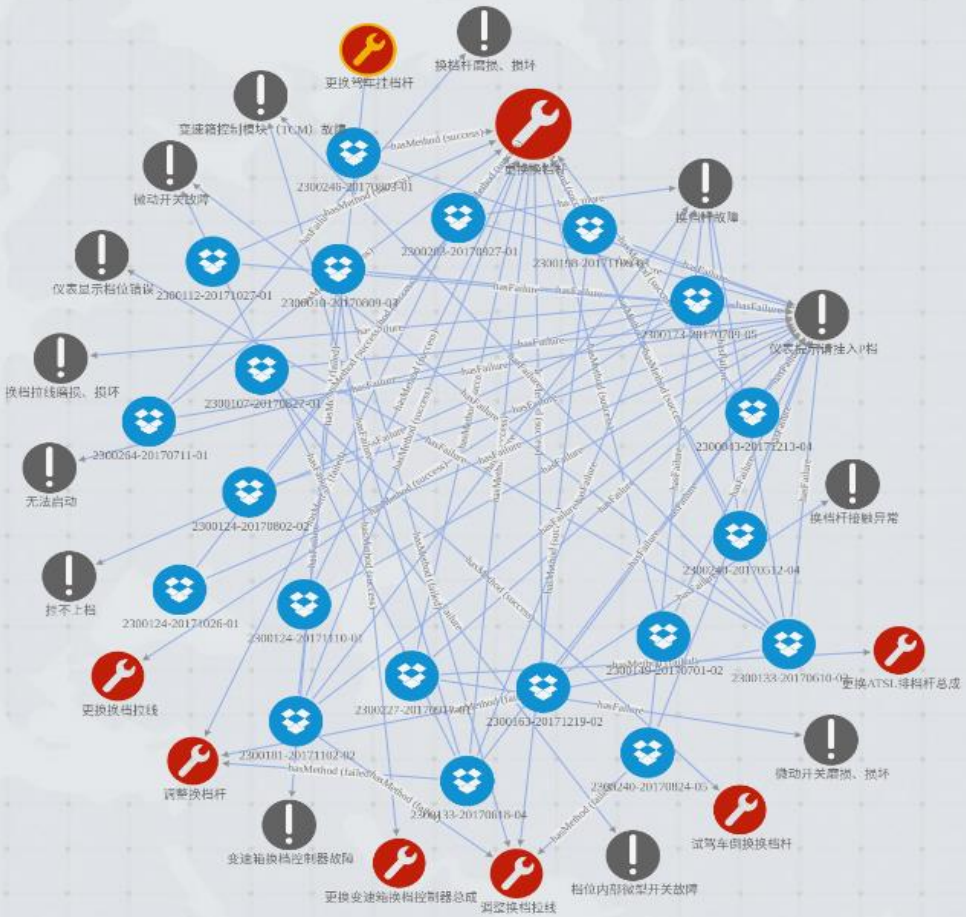
相关案例

2300004-20170215-08
2300144-20170515-03
2300231-20170204-02
2300045-20170702-01
2300216-20170505-01
2300156-20170711-03
2300000-20170410-01
2300206-20170917-07

图谱应用--语义检索（同故障所有维修方法）

检索　推理　信息统计

车型: ABC　故障: Ex: 离合器　仪表提示请挂　维修部件: Ex: TCM　工况: Ex: 挂入R档　故障码: Ex: P0700　查 询



- 案例编号
- VIN
- 故障
- 维修方法
- 零部件
- 车型
- 工况
- 故障灯
- 仪表提示
- 故障码
- 故障码含义
- 发动机型号
- 变速箱型号

导航器

仪表提示请挂入P档　ABC　（总计:20 成功:19 失败:1）　全部

维修方法	尝试案例	成功	失败	占比	成功率
更换换档杆	16	16	0	80.00%	100.00%
更换变速箱换档控...	1	1	0	5.00%	100.00%
更换换档拉线	1	1	0	5.00%	100.00%
试驾车倒换换档杆	1	1	0	5.00%	100.00%
调整换档拉线	4	1	3	5.00%	25.00%
更换ATSL排档杆总成	1	0	1	0.00%	0.00%
更换驾车挂档杆	1	0	1	0.00%	0.00%
调整换档杆	3	0	3	0.00%	0.00%

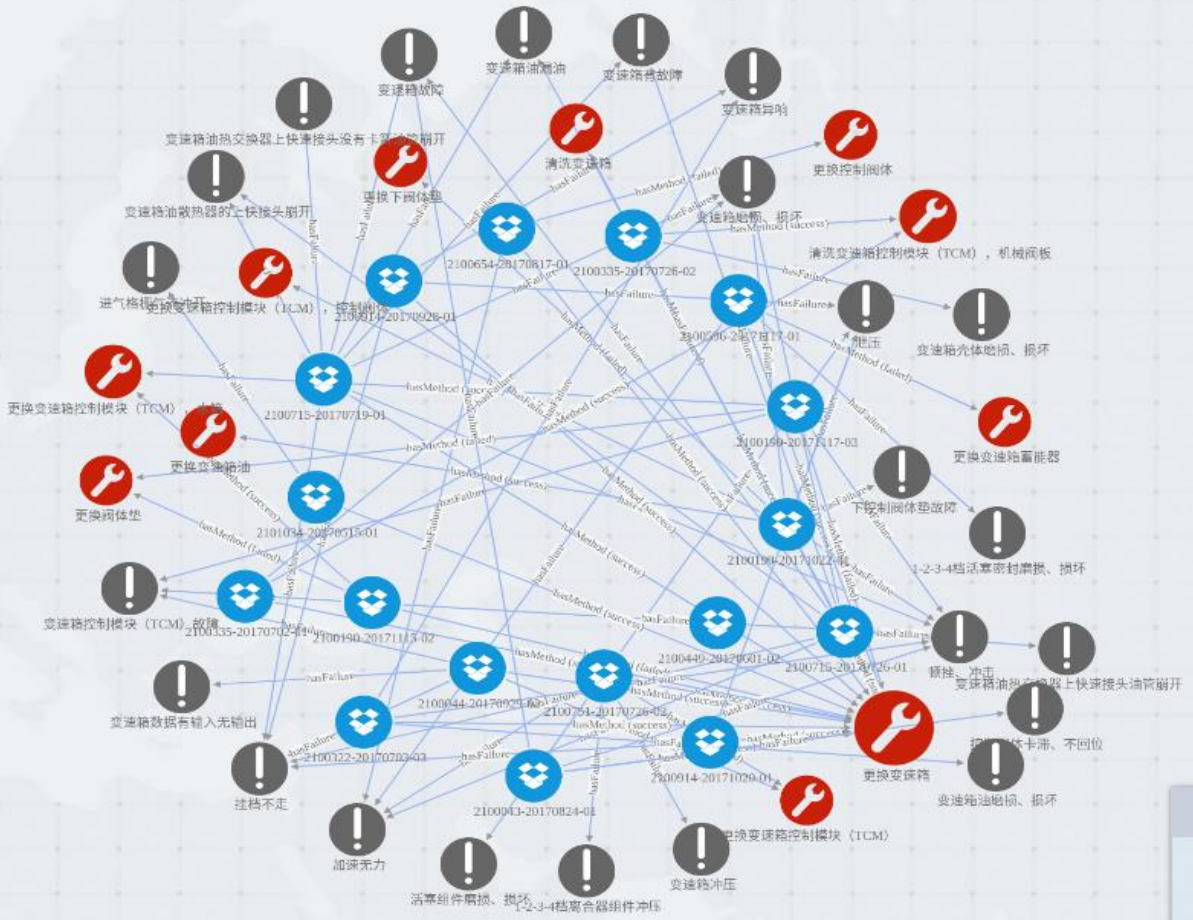
更换驾车挂档杆 （1）　全部

2300181-20171102-02

图谱应用--语义检索（同部件故障不同的维修方法）

检索　推理　信息统计

车型: ABC　故障: Ex: 离合器　Ex: 不回位　维修部件: 变速箱　工况: Ex: 挂入R档　故障码: Ex: P0700　查 询



- 案例编号
- VIN
- 故障
- 维修方法
- 零部件
- 车型
- 工况
- 故障灯
- 仪表提示
- 故障码
- 故障码含义
- 发动机型号
- 变速箱型号

导航器

变速箱　ABC　（总计:17 成功:16 失败:1）

维修方法	尝试案例	成功	失败	占比	成功率
更换变速箱	15	13	2	76.47%	86.67%
清洗变速箱控制模...	2	2	0	11.76%	100.00%
更换变速箱控制模...	2	2	0	11.76%	100.00%
更换变速箱油	1	1	0	5.88%	100.00%
更换变速箱控制模...	1	0	1	0.00%	0.00%
更换变速箱控制模...	2	0	2	0.00%	0.00%
更换变速箱蓄能器	1	0	1	0.00%	0.00%
更换控制阀体	1	0	1	0.00%	0.00%
更换阀体垫	2	0	2	0.00%	0.00%
清洗变速箱	1	0	1	0.00%	0.00%
更换下阀体垫	1	0	1	0.00%	0.00%

相关案例

2100190-20171022-01
2100449-20170601-02
2100751-20170726-02
2100044-20170929-02
2100190-20171115-02
2101034-20170515-01
2100715-20170719-01
2100914-20170928-01

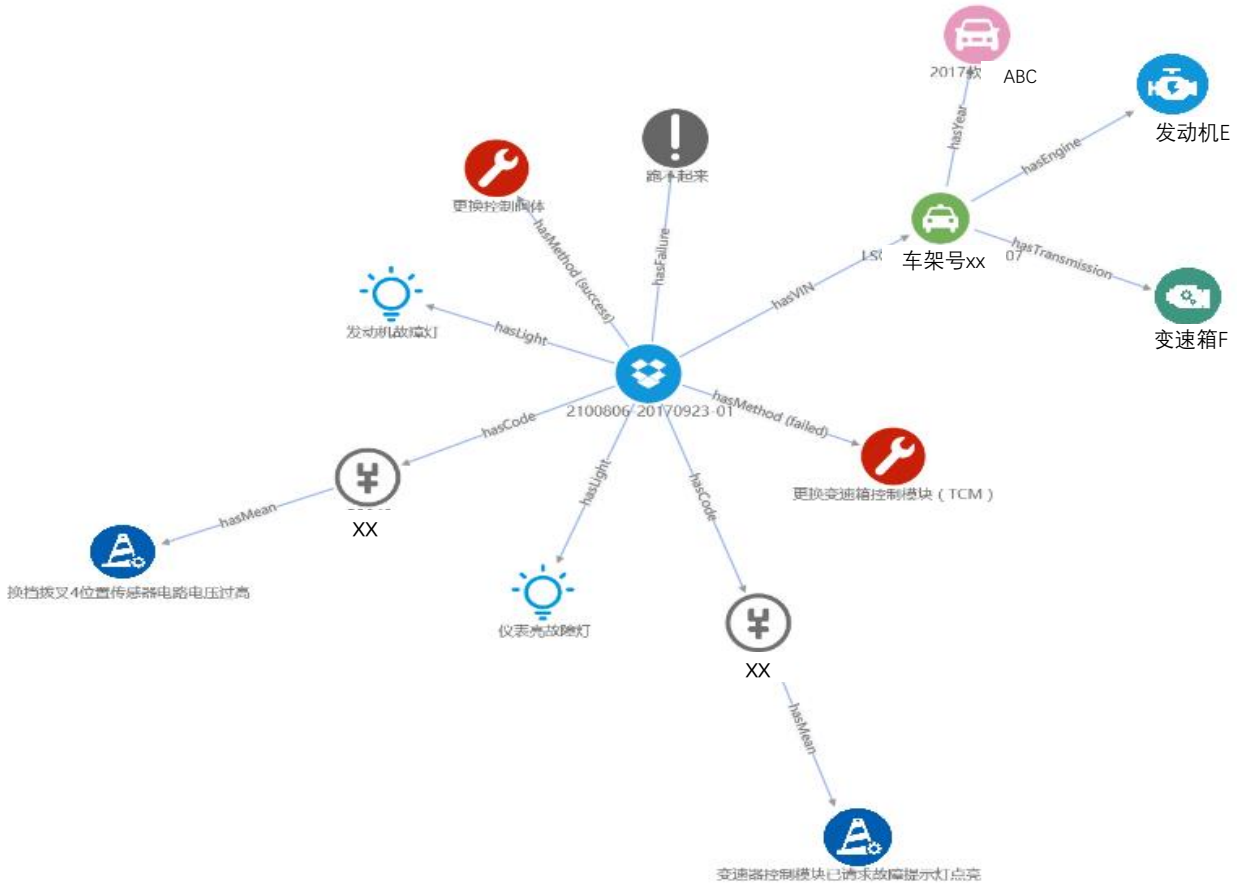
图谱应用--
自动抽取案例中的实体，关系，失效模式，维修方法等

架构面接未来变化
IAS 2018

2100806-20170923-01

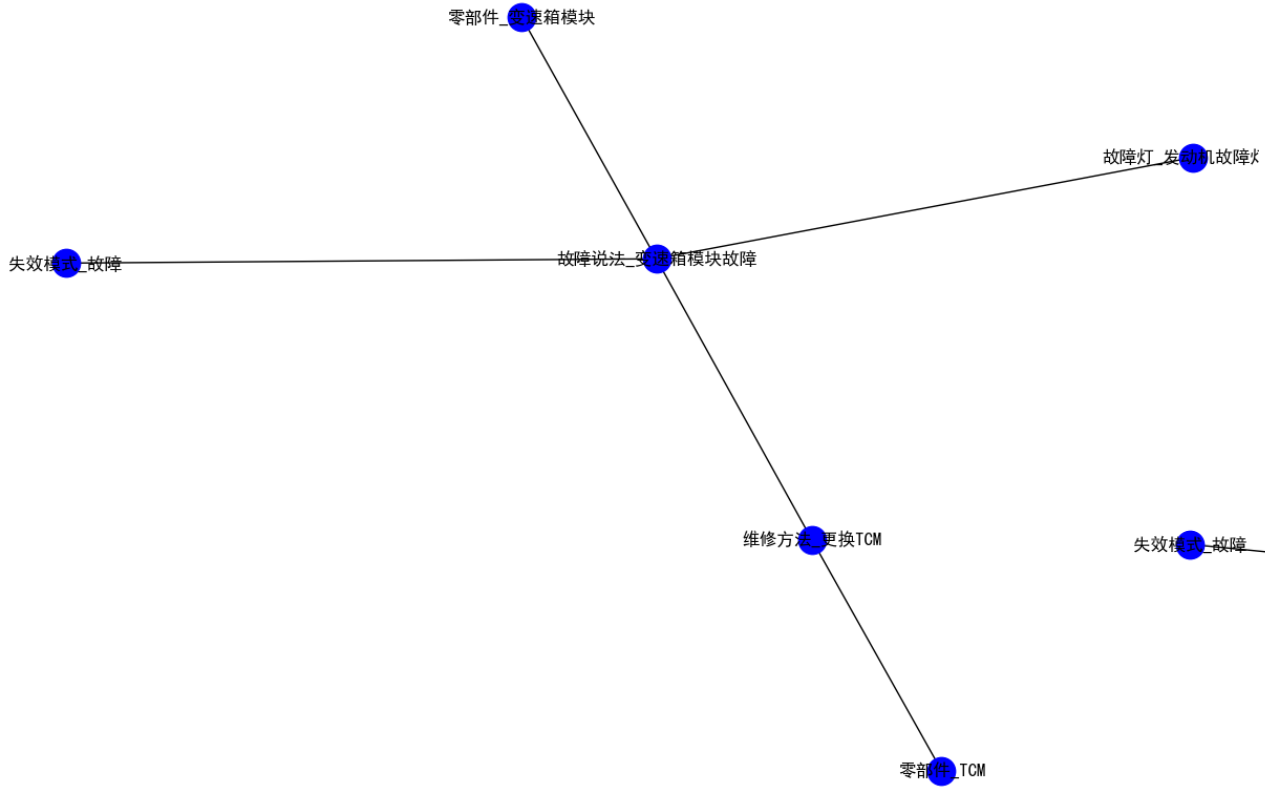
案例详情

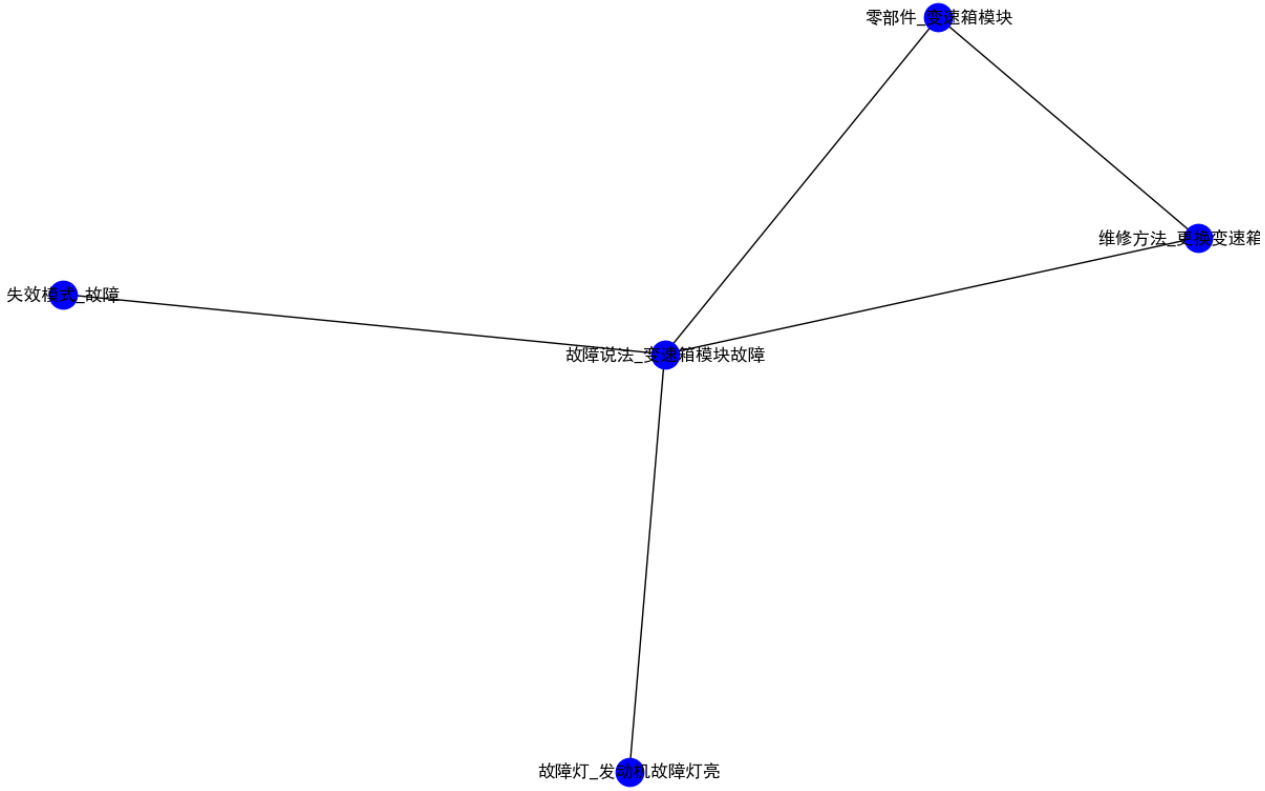
17年 AB 亮发动机故障灯仪表亮故障灯，有时跑不起来，转速很高。诊断维修：电脑检查有故障码 XX 换挡拨叉4位置传感器电路电压过高， XX 变速器控制模块已请求故障指示灯点亮。对故障码清除，变速箱学习后，故障又出现。根据维修手册测量线路都正常，对换变速箱模块后故障依旧，申请更换变速箱阀体总成。申请更换变速箱阀体总成。



MININGLAMP
略 数 据

图谱应用-频繁子图挖掘


零部件_变速箱模块
故障灯_发动机故障灯
失效模式_故障
故障说法_变速箱模块故障
维修方法_更换TCM
零部件_TCM


零部件_变速箱模块
维修方法_更换变速箱
失效模式_故障
故障说法_变速箱模块故障
故障灯_发动机故障灯亮

10. 业务价值

降低运维成本	• 让故障得到有效远程解决，降低故障解决时间，减少差旅成本； • 数据实时在线，打通数据孤岛，提高各部门协作效率，流程更加透明； • 对故障的发展、爆发等做到实时监控、跟踪和预测分析，使资源配置、方案制定等更加合理；
提高工作效率和问题解决质量	• 维修方法、失效模式的积累能够大大降低相似案例、疑难案例的分析成本，并提高售后支持解决方案的有效性； • 更加自动化、智能化的检索方式使用户能够快速、精确查找到期望结果，大大提高工作效率
数据金矿	• 不断完善和迭代知识抽取模型，故障分析模型，维修知识等，同时平台积累的案例数据、索赔数据、人的经验数据等金库被充分挖掘；

联系我们 Contact us

架构迎接未来变化
IAS 2018

关注明略数据官方微信
获取最新行业人工智能动态

个人微信，欢迎相关技术与解决方案交流

Add：北京市海淀区中关村东路1号院1号楼 清华科技园创新大厦A座10层1002
Tel：010-82151987
Web：www.mininglamp.com
E-mail：mkt@mininglamp.com

MININGLAMP
明 略 数 据

THANKS!