

大规模通用知识图谱构建及应用

百度-知识图谱-张扬

目录

- 知识图谱应用案例
- 知识图谱构建技术
 - 事实获取
 - Ontology构建
 - 实体归一
 - 主动收录
- 知识图谱应用技术
 - 人机平台
 - 问答系统
 - 智能写作
- 知识图谱开放
- 知识图谱展望

知识图谱应用案例 – 搜索



- 屏幕越来越小
- 交互场景越来越广泛



知识
图谱



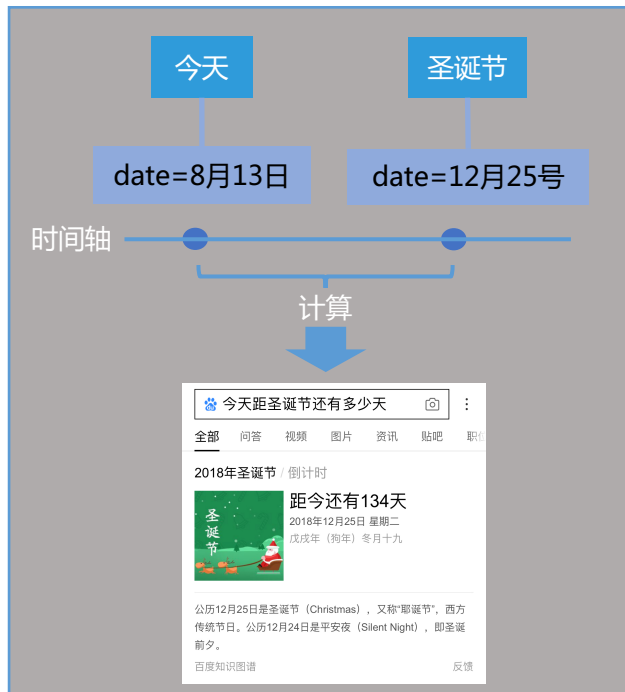
- 更加高效的结果满足
- 更加自然的交互方式
- 每天十几亿次用户请求，覆盖超过85%搜索PV

知识图谱应用案例 – 搜索

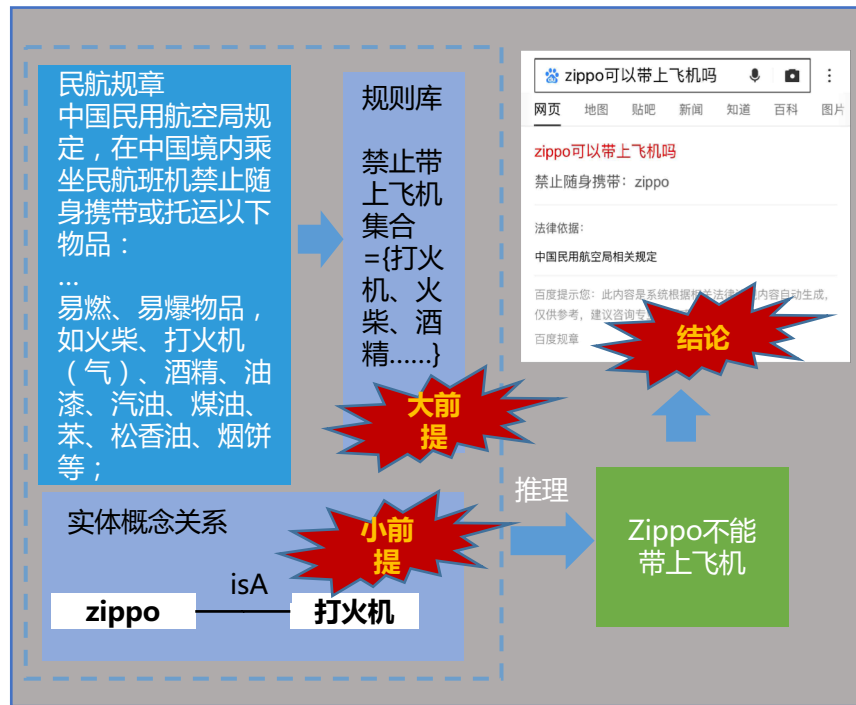
王菲和李亚鹏的... 全部 图片 问答 视频 资讯 贴吧 李亚鹏 王菲 / 女儿 李嫣 年龄：11岁  李嫣，内地演员李亚鹏与香港歌手王菲之女，也是王菲的第二个女儿。2006年5月27日中午12点左右生于北京协和医院。体重8斤，出生后母女平安。由于小字嫣... 百度百科 报错 王菲和李亚鹏的女儿 - 图片  百度图片 查看更多图片 >	北京市的面积 全部 问答 资讯 视频 图片 贴吧 北京 / 面积  1.641万平方千米 人口：2170.5万人（2015年） 北京，简称“京”，中华人民共和国首都、直辖市、国家中心城市、超大城市，全国政治中心、文化中心、国际交往中心、科技创新中心，是中国共产党中央委员会、中华人民共和国中央人民政府和全国人民代表大会的办公所在地，中国中部战区司令部驻地。北京位于华北... 邮政编码：100000 区号：010 简称：京 别名：燕京 蓟城 涿郡 幽州 北平 行政区划代码：110000 景点：天安门广场 故宫 颐和园 八达岭长城 明十三陵 百度百科 报错	梁思成和林徽因... 全部 视频 问答 资讯 图片 文库 夫妻 孩子：梁从诫 梁再冰  梁思成 百度知识图谱  林徽因 徐志摩,林徽因,梁思成是什么关系?_百度知道 4个回答 回答时间：2017年9月1日 [最佳答案] 在林徽因的感情世界里有三个男人,一个是建筑大师梁思成,一个是诗人徐志摩,一个是学界泰斗,为... 林徽因,梁思成和徐志摩到底什么关系啊 林徽因和冰心关系 ...张幼仪,陆小曼,他们到底都是什么关系? 百度知道 更多同站结果 > 梁思成、林徽因故居_百度百科  简介：梁思成、林徽因故居，位于北	高山流水典故中... 全部 资讯 文库 问答 视频 贴吧 伯牙 “自此始有高山流水遇知音，伯牙摔琴谢知音的典故，后有称颂其事，在此筑馆纪念，称为琴台，现琴台东对龟山，西临月湖，成为武汉著名古迹胜地。” 关于钟子期与伯牙关于《高山流水》_百度知道 zhidao.baidu.com 关于这条结果 反馈 《高山流水》的典故中,弹琴者是谁?听琴者是谁?_作业帮 《高山流水》的典故中,弹琴者是谁?听琴者是谁?语文作业帮用户2017-09-18扫二维码下载作业帮 3亿+用户... https://www.zybang... 2017年9月18日 【搜】高山流水故事中善于弹琴的人叫什么?..._王朝网络 10.高山流水典故中,弹琴者是谁?《高山流水》典故中,弹琴者是伯牙以下供参考:高山流水觅知音 伯牙是春秋...	黄日华版天龙八... 全部 视频 资讯 问答 图片 贴吧 难念的经(黄日华版天龙八部主题曲)在线听QQ音乐  歌手：周华健 专辑：生·活 发行时间：1997-01-24 在线试听 QQ音乐 虾米音乐 天龙八部 黄日华版 主题曲是什么,谁唱的_百度知道  3个回答 回答时间：2017年9月3日 [最佳答案] 周华健《难念的经》 附歌词:春风吻雨拜落日未曾伤泣欺山 枉海踏雪径也未绝望 拈花把酒偏折煞世人情狂凭这两眼与百臂或千手... 黄日华版的天龙八部里有很多经典的背景音乐...	一尺九腰围是多... 全部 问答 图片 贴吧 视频 资讯 一尺九腰围是多少码 26码 [尺码换算] • 一尺九(市尺)=26码(英寸)=64(厘米) • 市尺数*10+7=英寸数; 1市尺=33.3厘米 百度知道 报错 1尺9的腰围穿几码的裤子?_百度知道  6个回答 回答时间：2017年10月... [最佳答案] 1、1尺9的腰围穿26码的裤子。2、计算方法 1.9*10=19,19+7=26码 3、尺码对照表 26码=1.9尺 腰27码=2尺腰 28码=2.1尺腰29码... 一尺九的腰围是多少厘米 一尺九的腰是穿几码的裤子? 裤子腰围一尺九是s码还是m码还是l码? 百度知道 更多同站结果 >
---	--	--	--	--	--

直接给出答案：知识图谱精准问答，基于实体的计算，直接给出答案，使搜索引擎更加智能

知识图谱应用案例 – 搜索



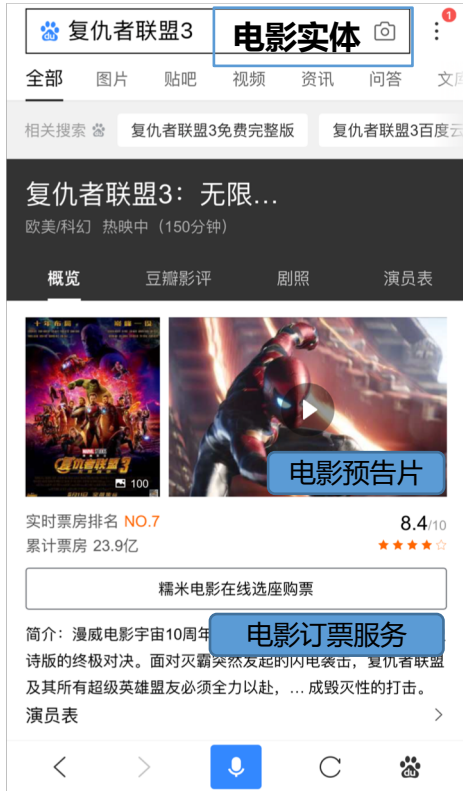
基于知识的动态函数推理计算



基于知识的规则推理计算

知识图谱应用案例 – 搜索

实体信息/服务聚合:以实体为粒度理解用户意图，聚合全网优质资源且展现更全面、更有价值的搜索结果，并提供畅游式浏览体验，让人和搜索引擎的交互更加自然。



知识图谱应用案例 – 搜索

特色垂类能力与服务

基于全网异构中文语言资源，融合多层次细粒度中文语言知识数据，构建了全面的中文语言知识图谱垂类，并研发了中文语言智能化解析、推理、计算引擎，每天满足数千万次汉语搜索请求。



Query：一丘之貉的貉咋写



VS

Query：衣字中间一个保是什么字



VS

Query：检查的查多音字组词



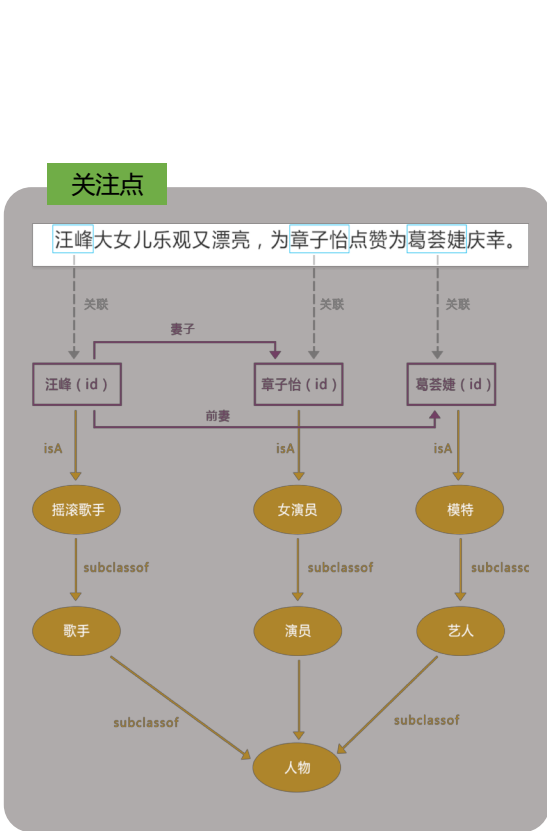
VS

Query：()天()地



VS

知识图谱应用案例 – 信息流



基于多元语义知识图谱的个性化推荐

- 百万关注点，千万关联关系，满足全部用户的兴趣刻画
- 十亿语义与视频的语义标签，精准关联95%的热门实体
- 百万事件，千万关系，事件脉络小时级更新



资源

凭什么是汪峰抱得章子怡归？

【严肃八卦（微信号yansubagua）禁止任何微信号和媒体未经许可转载！要转载联系我谈稿费！其他订阅号随意转载我就举报！网站随意转载我直接投诉总编室！】

虽然章子怡和汪峰一直不承认，但是今天全民星

【章子怡，请汪峰来洛杉矶，这是赴美生子的节奏吗？】

所以，观者千般万般不看好，汪峰和章子怡，已

娱乐 章子怡 汪峰 明星

知识图谱应用案例 – 信息流

- 以知识图谱数据为核心，应用知识加工、计算、聚合等方式产生高时效、高质量的独家内容
- 覆盖娱乐、体育、财经、美食等22个领域，累计产生文章高达80万篇，总阅读量数亿次
- 快讯类写作从事件发生到触达用户仅需要23秒



实体聚合文章

世界杯决赛战报：法国 VS 克罗地亚 全场以比分4:2结束

小度足球早播报
百家号 | 07-16 00:55

已关注

法国 VS 克罗地亚 于7月15日 23:00举行

全场比赛结果：法国 VS 克罗地亚 全场以比分4:2结束

克罗地亚佩里西奇在28分钟、法国7号格列兹曼在38分钟、法国6号博格巴在59分钟、法国10号姆巴佩在65分钟、克罗地亚17号曼朱基奇在69分钟分别踢进一球。

本文由百家号作者上传并发布，百家号仅提供信息发布平台。文章仅代表作者个人观点，不代表百度立场。未经作者许可，不得转载。

举报

体育快讯文章



资讯聚合文章



稀缺类目文章

知识图谱应用案例 – DUEROS

DUEROS

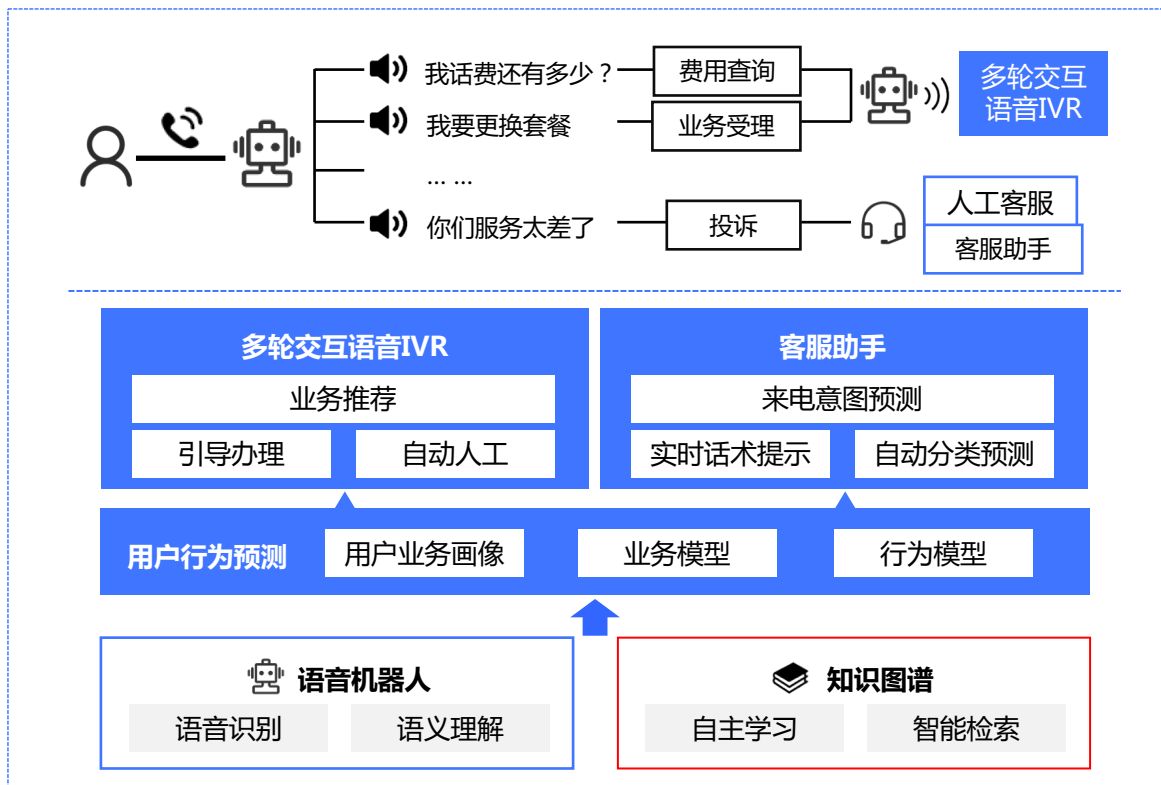
人工智能时代的操作系统，真正能听清、听懂、满足的开放式生态系统



知识图谱助力DuerOS实现10大类目超过100类能力
直接满足 40% 以上的信息需求

知识图谱应用案例 – 行业

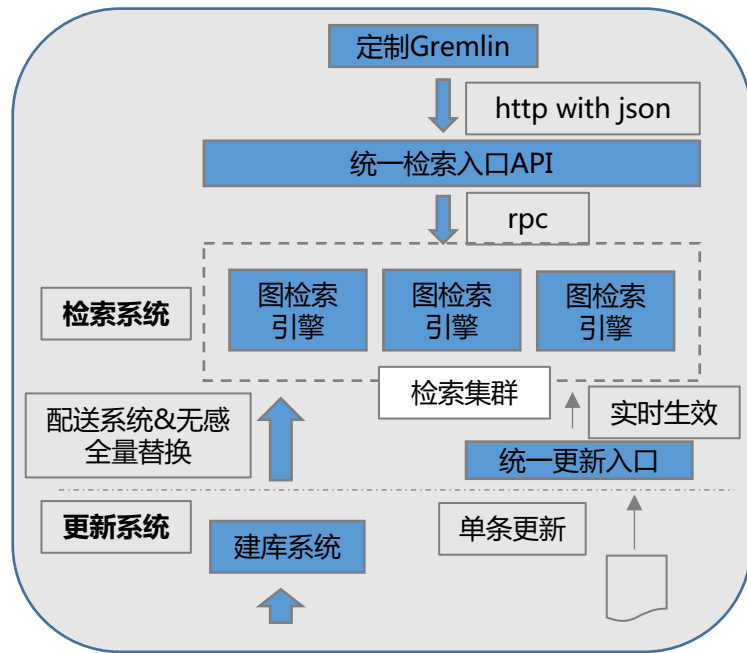
- 引入知识图谱，沉淀积累客服知识和信息资源
- 理解客户问题内容和意图支持客服坐席知识提示
- 面向非投诉类FAQ的智能问答，提升客服工作效率
- 整体提升客服通话效率20%+



知识图谱应用案例 – 行业

BGraph是基于百度知识图谱自主研发并大规模应用的图检索技术，打造的行业图检索产品。

- 支持私有化部署、针对行业应用的丰富的图检索功能
- 高性能：针对图存储和图查询专门优化，搜索场景平均10ms以内响应
- 大容量：能支持十亿级顶点，千亿级边
- 可扩展：分布式可扩展，满足业务的快速增长
- 高稳定：内部业务SLA 99.99%，多年线上大规模应用的检验，每日数十亿调用量
- 开放性：支持业界流行的图数据表达模型和扩展的Gremlin 检索语法接口



知识图谱应用案例 – 行业



百度学术 (学术图谱)

百万中文学者库
亿级学术论文实体



百度教育 (教育图谱)

亿级专业文档
数十万教学视频



权威信息



需求细化

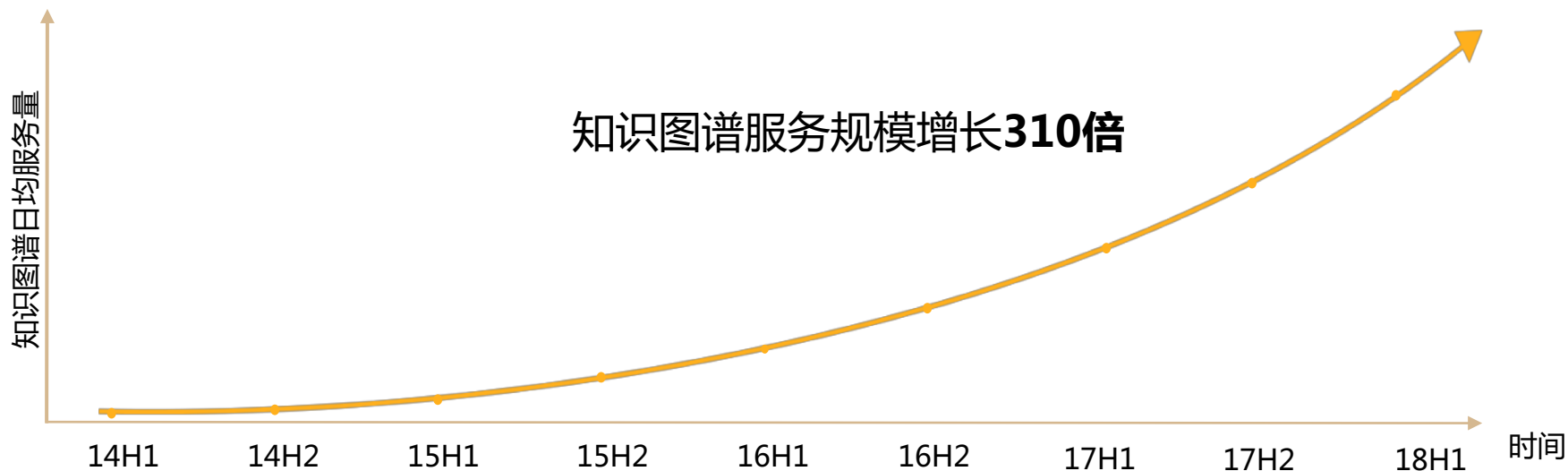


详情引导

百度地图 (地理图谱)

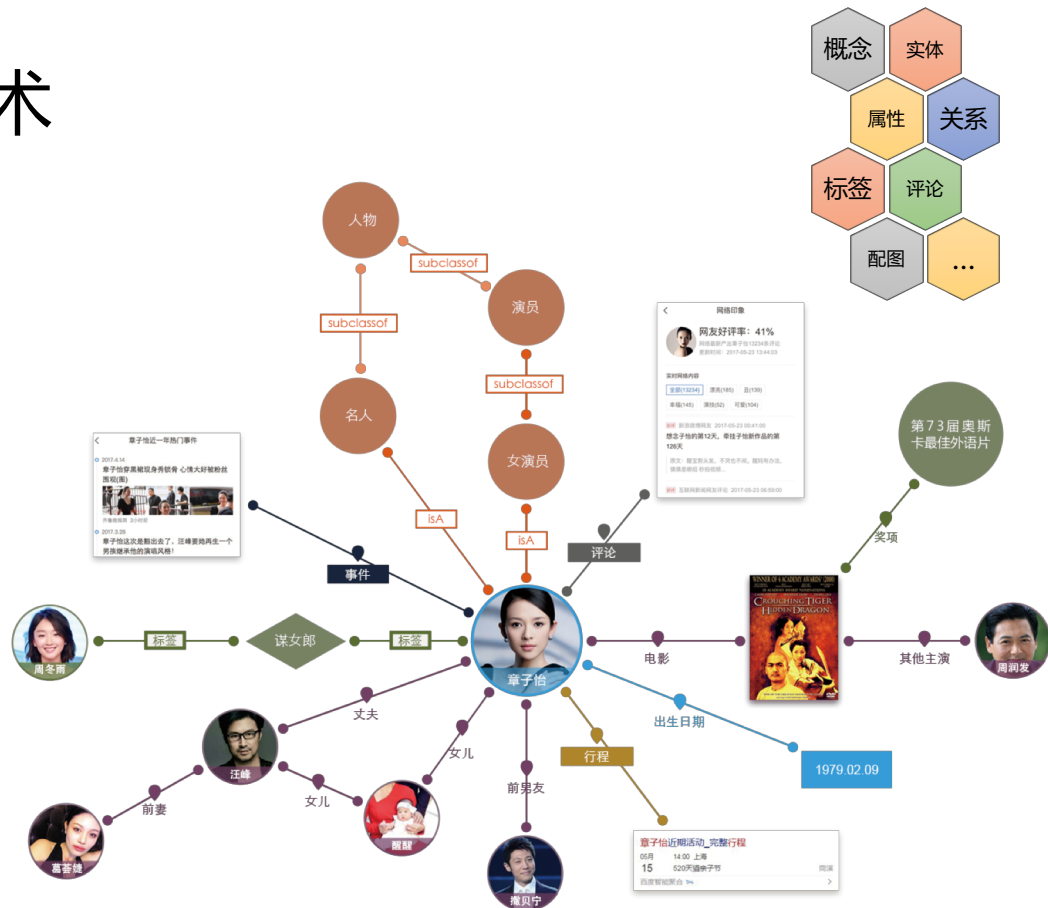
全球POI总数超亿级

知识图谱应用案例

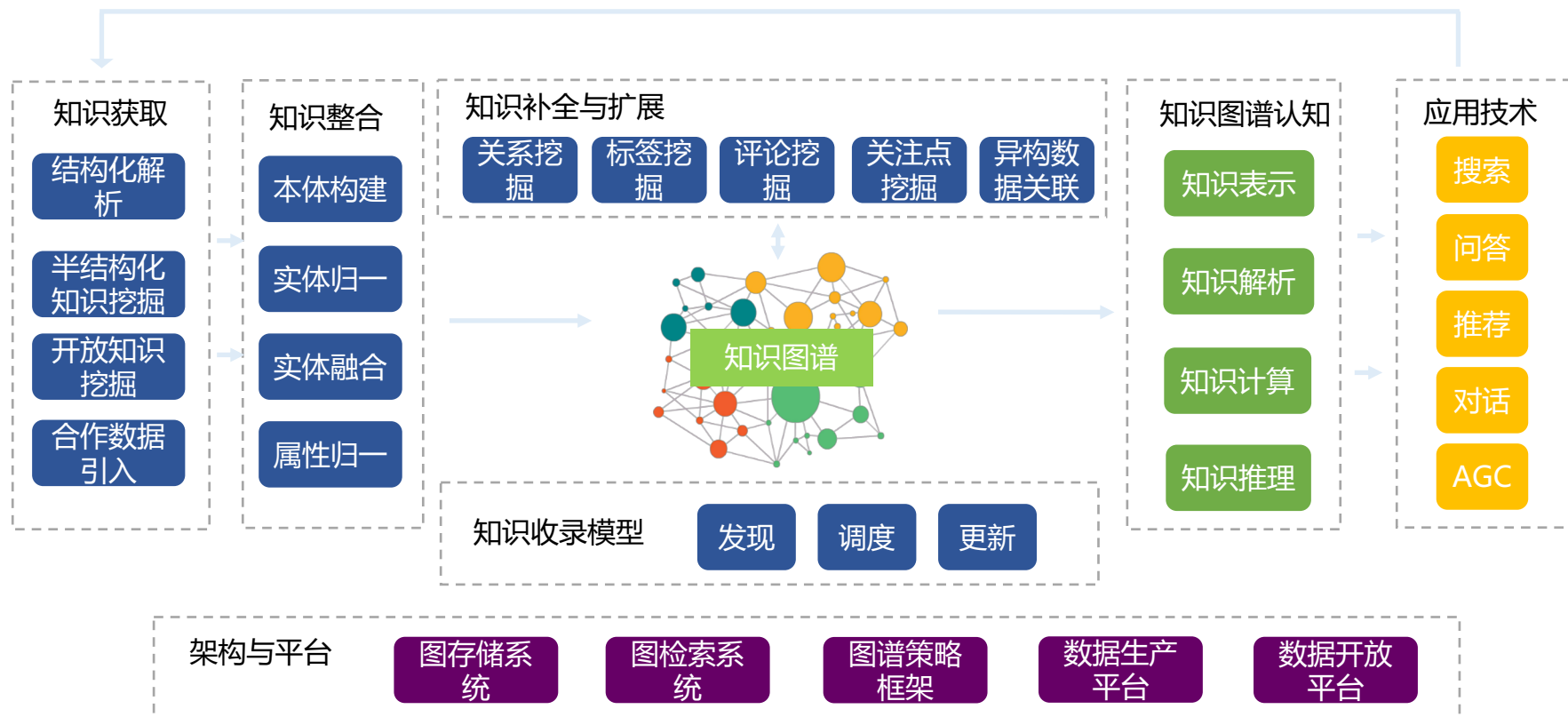


知识图谱构建与应用技术

- 全球最大的知识图谱：亿级实体，千亿事实
- 覆盖人物、影视、音乐、文学、商品、餐饮、旅游、出行、金融、教育、房产、医疗等数十个垂类
- 完备的基础设施建设，高时效性数据秒级更新



知识图谱构建与应用技术



知识图谱构建技术 - 事实获取

结构化抽取：每个xpath节点的语义相同

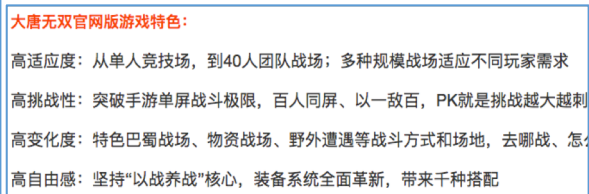
- 抽取方式：人工配置xpath模版
- 优点：
 - 优质站点数据质量较好，性价比高
 - 垂类实体覆盖较好
 - 部分天然的边关系数据
- 难点：
 - 如何高效的进行模版配置
 - 如何预防网站改版
 - 如何保证网页资源收录的实效性，覆盖率



知识图谱构建技术 - 事实获取

半结构抽取：基于网页特定格式的xpath结构进行抽取

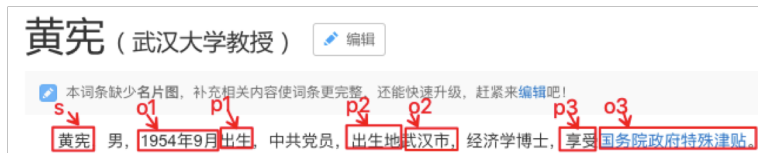
- 抽取方式：
 - KV 识别
 - 主体识别
 - 实体页识别
 - 网页核心mention提取
 - 特定样式xpath识别
- 优点：
 - 数据量庞大，偏长尾信息
 - 策略生成，人工成本小
- 缺点：数据质量略低



知识图谱构建技术 - 事实获取

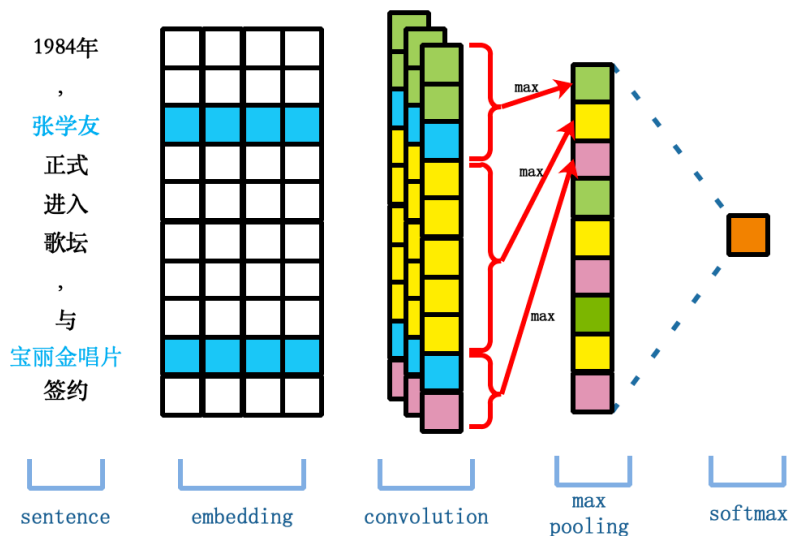
无结构抽取：在无结构的自由文本中，在没有预先定义的schema约束的情况下，通用地抽取开放SPO三元组知识

- **预处理**：获取网页/百科实体，切分句子，添加词法句法信息、实体识别
- **S-O层判别**：识别句中候选S-O pair，基于词法句法规则初步过滤S-O pair，通过PCNN分类模型进行二分类判断是否正确
- **P层标注**：通过bi-LSTM序列标注模型，在句子中标注给定S-O pair对应的P
- **规则过滤**：根据词法句法规则过滤SPO

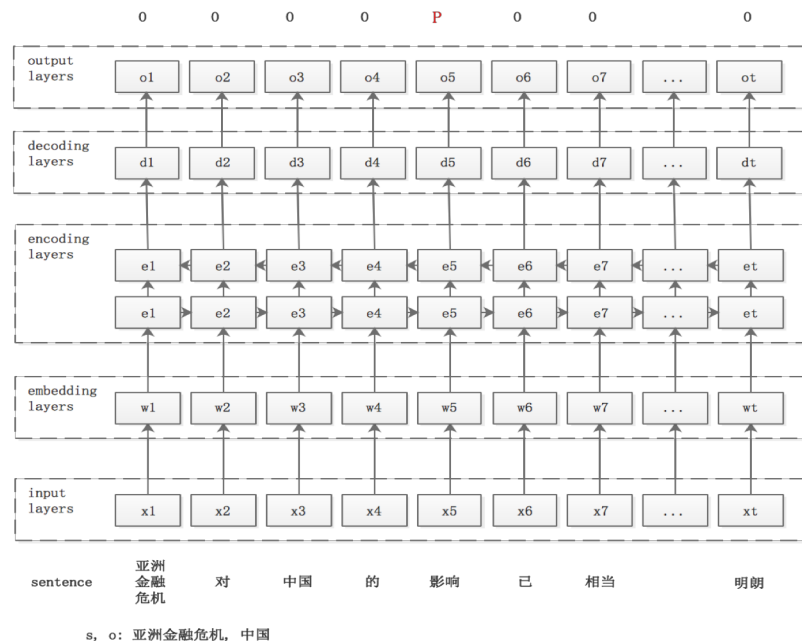


知识图谱构建技术 - 事实获取

无结构抽取：模型示意图



S-O 判别层模型



Predicate 序列标注

知识图谱构建技术 - 事实获取

按抽取知识的类型

- 实体抽取
- 语义标签抽取
- 二元关系抽取
- 多元关系抽取
- 事件抽取
- . . .

网络小说家天蚕土豆在网络小说界大名鼎鼎，其所写的《斗破苍穹》更是人气爆棚

2013年，《南方娱乐周刊》通过网友、媒体评审团、圈中人评审团的力量，最终Angelababy、刘诗诗、杨幂、倪妮被评选为新“四小花旦”

小米上市首日破发 7月23日将正式纳入恒指

7月9日，作为港交所“同股不同权”新规实施后的第一股小米集团（1810.HK）正式在[香港交易所](#)上市。小米董事长雷军亲自敲锣开市。图/视觉中国

7月9日，小米集团在港交所正式挂牌上市，与发行价每股17港元相比，开盘价下跌2.35%，为16.6港元，总市值为3714.4亿港元（约为473.3亿美元）。截至收盘，小米股价固定在16.8港元，相比发行价下跌1.18%，总市值为3759亿港元（约479亿美元）。相比雷军7月8日披露的小米估值543亿美元，小米市值IPO首日蒸发64亿美元。

实体抽取

天蚕土豆 PER
斗破苍穹 NVL

语义标签抽取

天蚕土豆 网络小说界大名鼎鼎
斗破苍穹 人气爆棚

二元关系抽取

天蚕土豆 ISA 网络小说家
斗破苍穹 作者 天蚕土豆

多元关系抽取

新“四小花旦” —Angelababy 刘诗诗 杨幂 倪妮

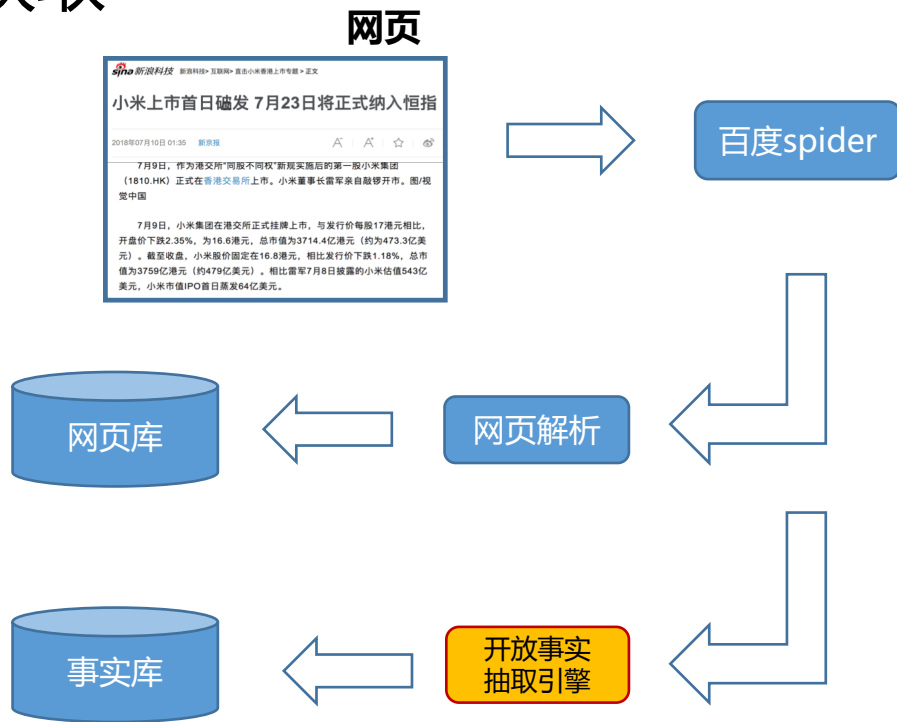
事件抽取

事件：IPO

- 公司：小米集团
- 交易所：香港交易所
- 发行时间：2018年7月9日
- 发行价：17港元

知识图谱构建技术 - 事实获取

- 完全开放语义的深度事实抽取技术，不限定 predicate 类型
- 结构化/半结构/无结构 等多种抽取器并联，保证事实的收录覆盖
- 7*24小时流式不间断计算，分钟级别更新，保证事实抽取的实效性
- 每天完成约十万量级网页的抽取计算，产出百万量级开放 spo 三元组事实数据



知识图谱构建技术 – Ontology构建

- 自顶而下的Ontology构建方式：基于人工（专家）构建
 - 优点：准确性、专业性、完备性较高
 - 缺点：需要专业知识，效率低
 - 适用范围：特定领域图谱构建
- 自下而上的Ontology构建方式：数据驱动
 - 优点：算法生成，高效，不需要领域专家
 - 缺点：准确性、专业性、完备性不足
 - 适用范围：通用图谱构建

kg5.baidu.com/Corporation			
工具 已导入 bd work			
dissolutionDate	Date	DefaultConstraint	dissolutionDate
duns	Text	DefaultConstraint	duns
email	Text	DefaultConstraint	电子邮件
employee	Person	DefaultConstraint	雇员
event	Event	DefaultConstraint	重要事件
faxNumber	Text	DefaultConstraint	传真号码
founder	Person	DefaultConstraint	创始人
foundingDate	Date	DefaultConstraint	成立日期
foundingLocation	Place	DefaultConstraint	foundingLocation
globalLocationNumber	Text	DefaultConstraint	全球位置码
hasPOS	Place	DefaultConstraint	支持POS机
interactionCount	Text	DefaultConstraint	交互数
isicV4	Text	DefaultConstraint	isicV4
legalName	Text	DefaultConstraint	正式名称
location	PostalAddress or Place	DefaultConstraint	地点
logo	URL or ImageObject	DefaultConstraint	logo
makesOffer	Offer	DefaultConstraint	makesOffer
member	Organization or Person	DefaultConstraint	成员

知识图谱构建技术 – Ontology构建

开放分类体系构建

- IC关系挖掘

国家电网有限公司成立于2002年12月29日，是根据《公司法》规定设立的中央直接管理的国有独资公司

IsA 抽取



IsA

国有独资公司

百度从不满足于自身取得的成绩，也从未停止发展的步伐，自2005年在纳斯达克上市以来，截至2015年，百度的市值已达 800亿美元。如今，百度已经发展成一家国际性企业，在日

百度-上市时间-2005年

规则式生成



IsA

上市公司



百科义项分类



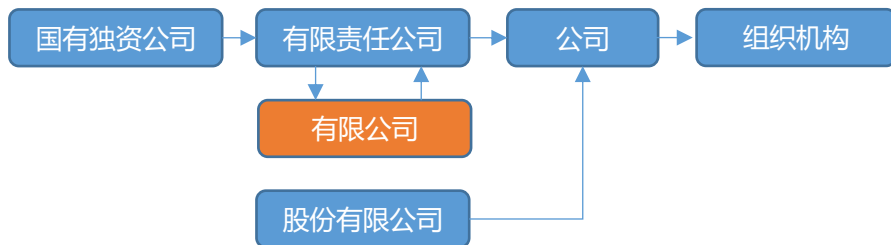
IsA

互联网公司

知识图谱构建技术 – Ontology构建

开放分类体系构建

- CC 关系构建：
 - 上下位挖掘
 - 同位判断



国有独资公司 编辑

国有独资公司，是指国家单独出资、由国务院或者地方人民政府授权本级人民政府国有资产监督管理机构履行出资人职责的有限责任公司。国有独资公司的公司章程由国有资产监督管理机构制定，或者由董事会制定报国有资产监督管理机构批准。联我国有独资公司不设股东会，由国有资产监督管理机构行使股东会职权。国有资产监督管理机构可以授权公司董事会行使股东会的

有限责任公司 编辑

有限责任公司，简称有限公司。中国的有限责任公司是指根据《中华人民共和国公司登记管理条例》规定登记注册，由五十个以下的股东出资设立，每个股东以其所认缴的出资额为限对公司承担有限责任，公司法人以其全部资产对公司债务承担全部责任的**经济组织**。有限责任公司包括**国有独资公司**以及其他有限责任公司。^[1]

我国法定公司有两种形式：有限责任公司(co.,ltd或ltd.) 和股份有限公司 (co.,ltd.)

Baidu 百科 公司 进入词条

公司

范主说：这次说的不是你呀~（商务范出品，转载请注明出处）近日，麦当劳的中国公司正式改名为“金拱门”，这一消息发布后...

而今，公司这个词我们太熟悉了，谁的生活都离不开它。但你说公司制为什么是市场经济的主旋律呢？我想它一定有什么...

1976年9月10日，有“发明大王”之称的“爱迪生”在纽约逝世。

2017-10-27 0 837 2017-11-30 0 139 2017-09-14

参考资料

1. 赵旭东等，《公司法》，高等教育出版社，2012：29

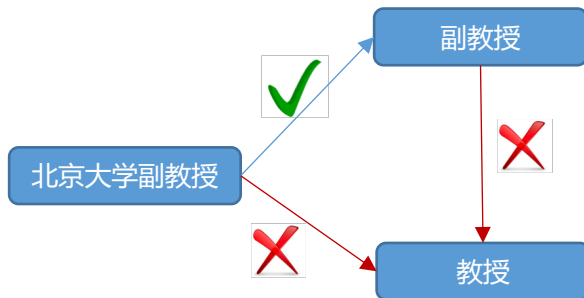
2. 赵旭东，《公司法》，高等教育出版社，2012：34

词条标签： 组织机构、法律术语

知识图谱构建技术 – Ontology构建

开放分类体系构建

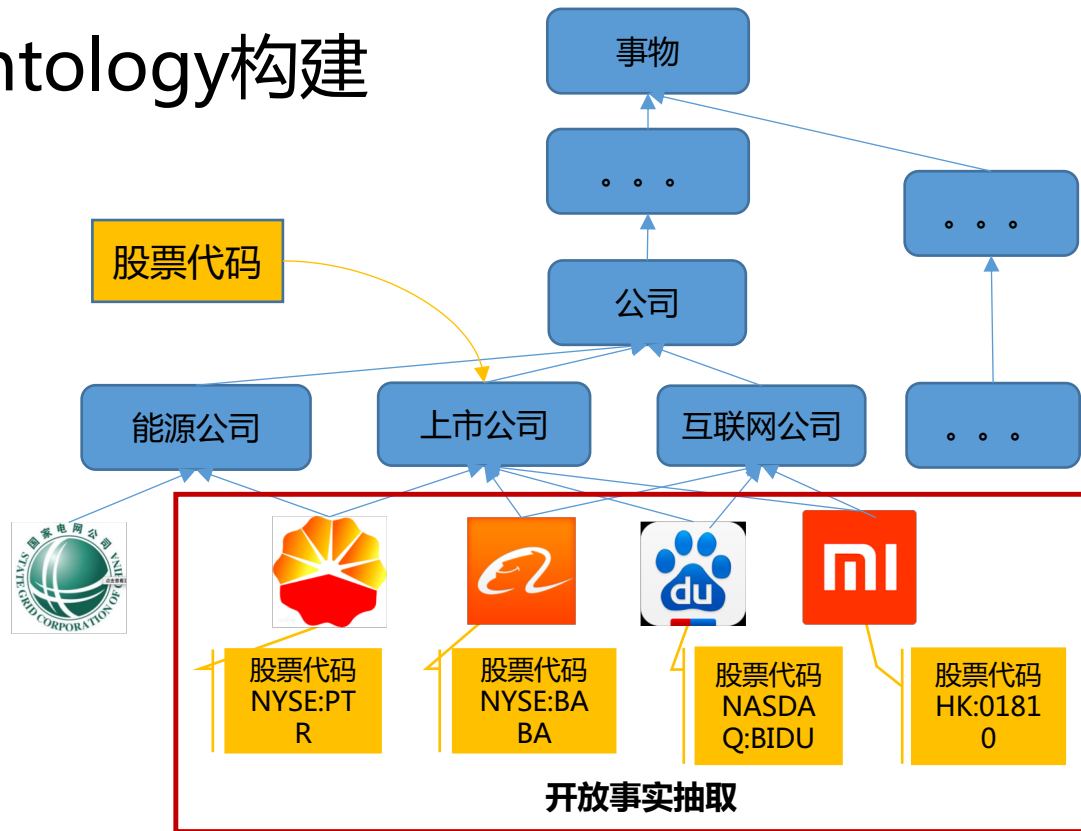
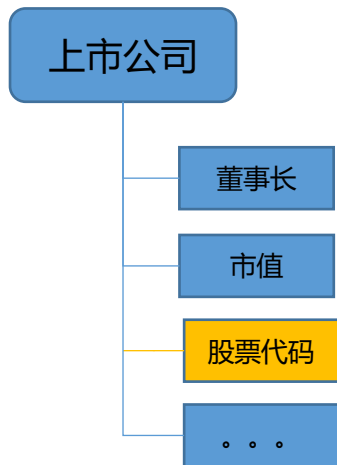
- CC 关系构建：
 - 人工维护
 - 基于词法关系的语义主干提取



知识图谱构建技术 – Ontology构建

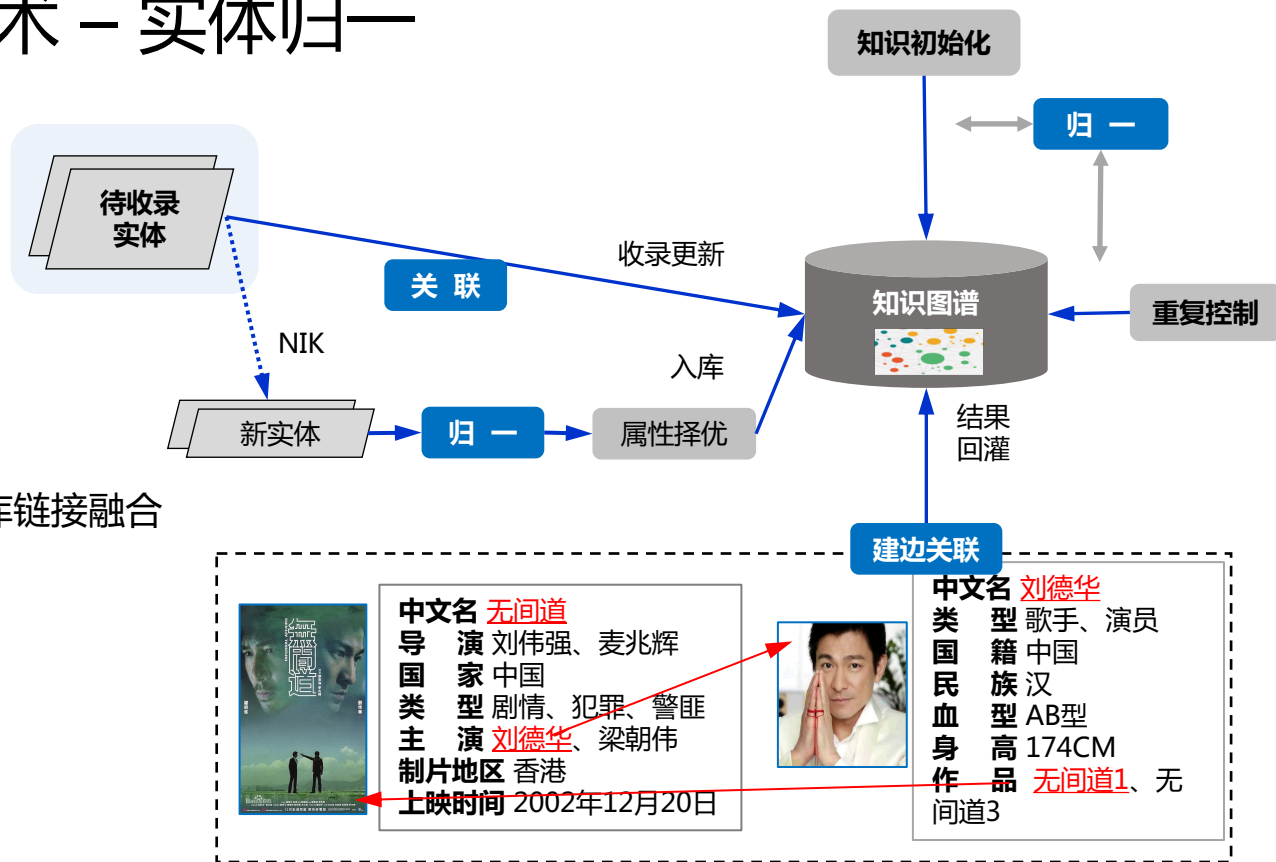
开放Schema构建

- 开放事实抽取
- 开放分类体系构建
- Schema自动生成
- Predicate归一 & O值约束挖掘

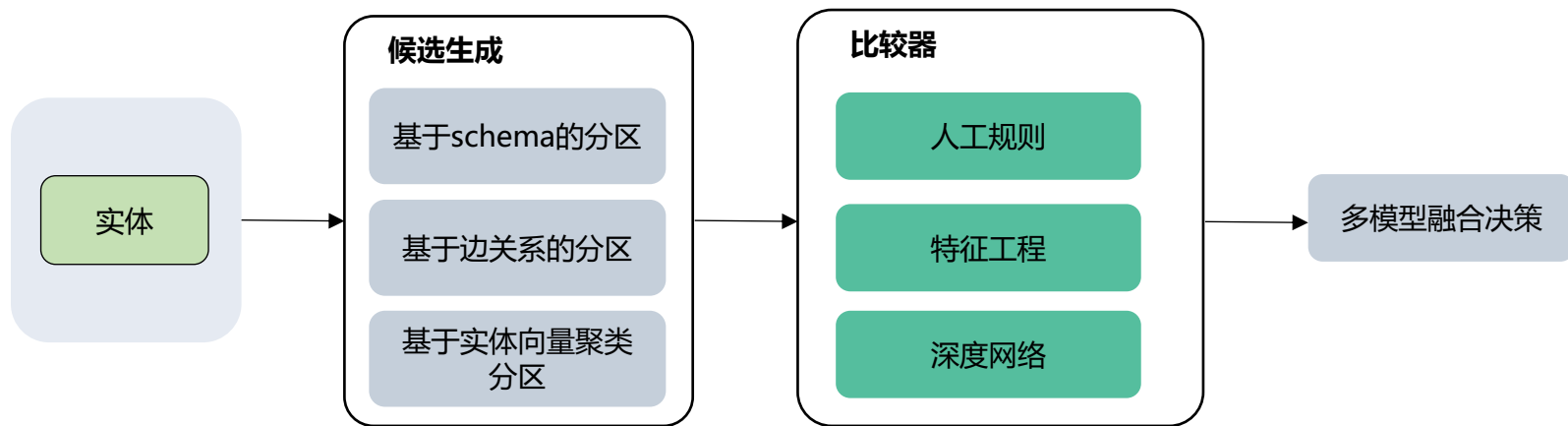


问题场景

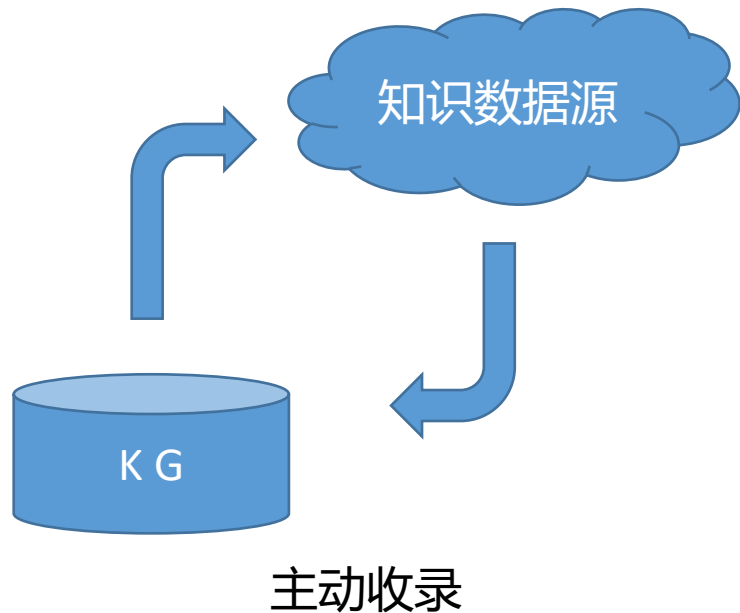
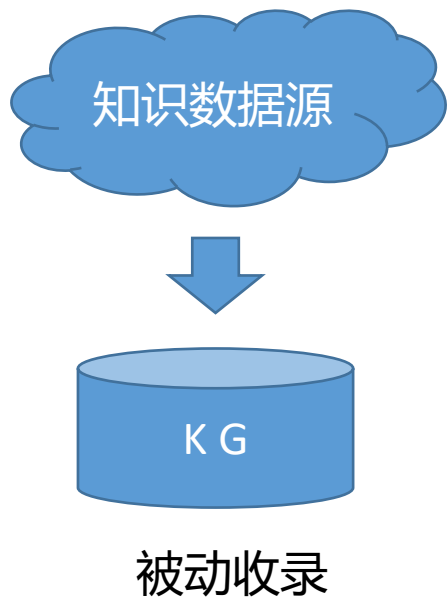
- 实体的同一性判断
- 知识图谱重复控制
- 图谱连通（O值ID化）
- 单源、多源（异构）知识库链接融合



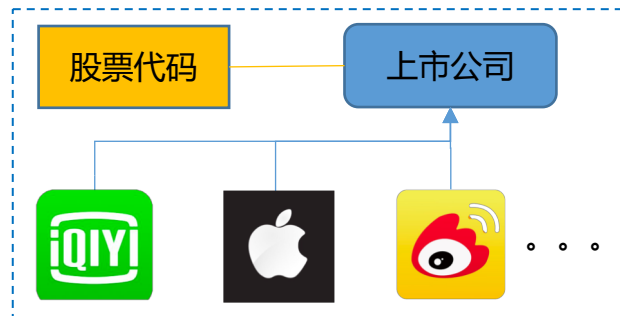
知识图谱构建技术 - 实体归一



知识图谱构建技术 – 主动收录



知识图谱构建技术 – 主动收录



Query 生成



Query Ranking

- 爱奇艺的股票代码？
- 苹果的股票代码？
- 谷歌的股票代码？
- 搜狗的股票代码？
- ...

任务分发

抽取队列

苹果的股票代码？
爱奇艺的股票代码？
谷歌的股票代码？
... ..

推理队列

条件推理
引擎

人工队列

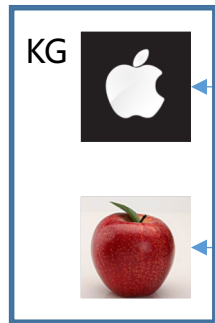


知识图谱构建技术 – 主动收录

Slot filling (抽取队列)



- Para 离线处理
 - 全网高质量para获取
 - 命名实体识别
 - 实体链接
- Para 检索 & ranking
 - 高性能para索引库
 - Para 相关性 ranking 技术
- 答案抽取
 - 基于 LSTM 的序列标注模型



苹果公司1980年在美国纳斯达克证券交易所上市。苹果公司的股票代码是：AAPL；截止到2016年11月29日，其收盘价为111.27美元/股，总市值为5946.68亿美元。

苹果 - 股票代码 - AAPL

苹果作为鲜食水果，等级价差就大，烟台苹果、洛川苹果都属于世界的知名品牌。尤其近些年来，洛川苹果风头盖过烟台苹果。这是得益于当地政府的大力支持扶持，果农的辛勤付出，果商的大量宣传。未来的苹果，在资本市场的引领下，等级划分将更加清晰：

知识图谱应用技术 - 人机平台

刘德华的女儿

百度一下

网页 图片 新闻 视频 知道 贴吧 文库 音乐 地图 更多»

百度为您找到相关结果约3,060,000个

刘德华女儿:
刘向蕙
刘向蕙 (小龙女, Hanna), 2012年5月9日出生于香港, 少数看过刘向蕙面貌的圈内人士称, 女儿的颜值不输妈妈朱丽倩, 耳朵则像天王爸爸刘德华, 长得眉清目秀...

郭靖

全部 图片 视频 贴吧 问答 资讯 文库

郭靖_百度百科
金庸武侠小说《射雕英雄传》男主角

郭靖, 是金庸武侠小说《射雕英雄传》中的男主角和《神雕侠侣》中的男主角。他出生于蒙古, 是成吉思汗的义子, 也是蒙古国大汗。他自幼在蒙古长大, 以杭州西湖为蓝本, 汲取了杭州西湖的美景, 成为了一名出色的弓箭手。他的一生充满了传奇色彩, 他的事迹是贯穿整个《射雕英雄传》的主线。

出生: 1205年(大漠)
父母: 郭啸天 李萍
子女: 郭芙 郭破虏 郭襄
出处: 《射雕英雄传》
职业: 侠客 武者
成就: 武林的领袖、当世威望最高的大侠、因守护南宋襄阳前线边境数十年而受万民景仰、死后威名历百余年而不衰
百度百科 反馈

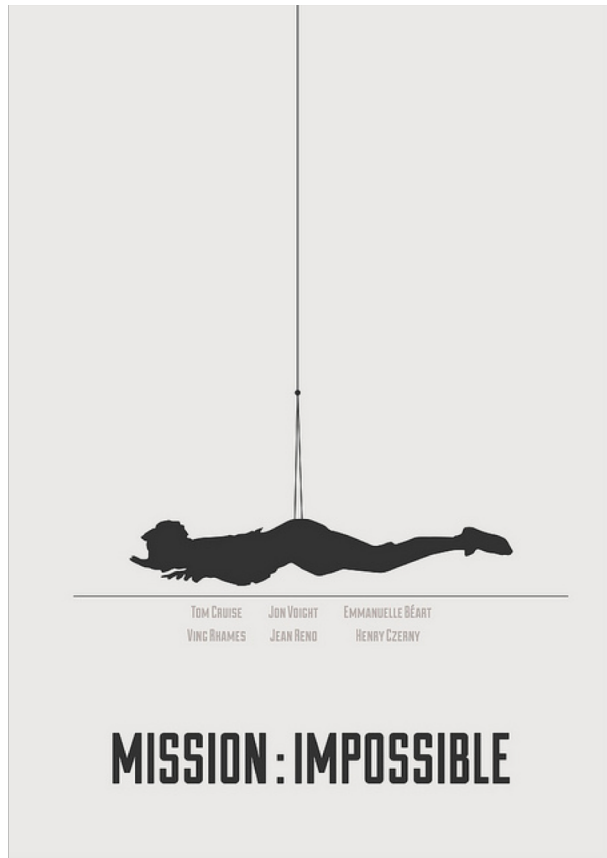
颐和园

颐和园 4.7分 18631人评价 AAAAA 景点

概览 图片 购票 攻略

数据准确率需求 99%

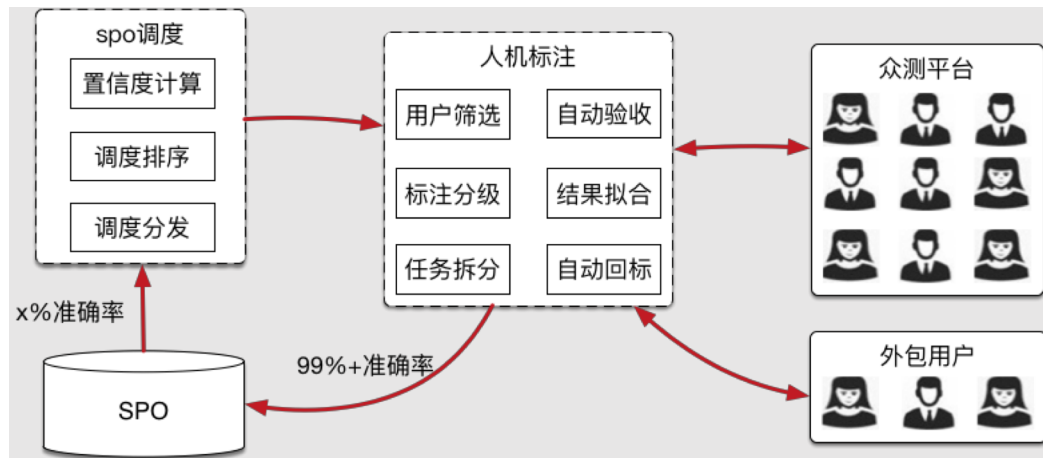
算法?



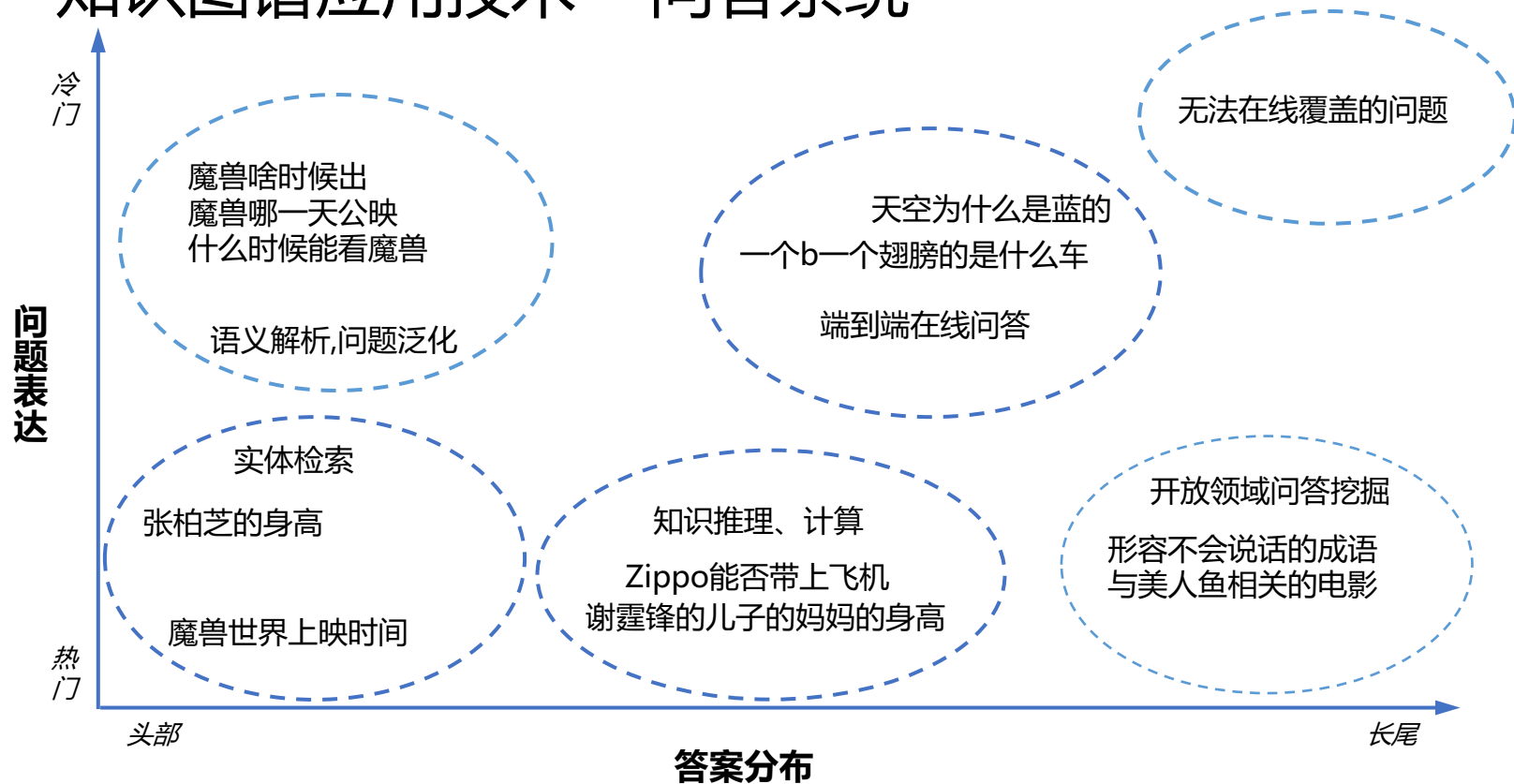
知识图谱应用技术 – 人机平台

人机结合子系统辅助突破算法瓶颈 —— 最后一公里

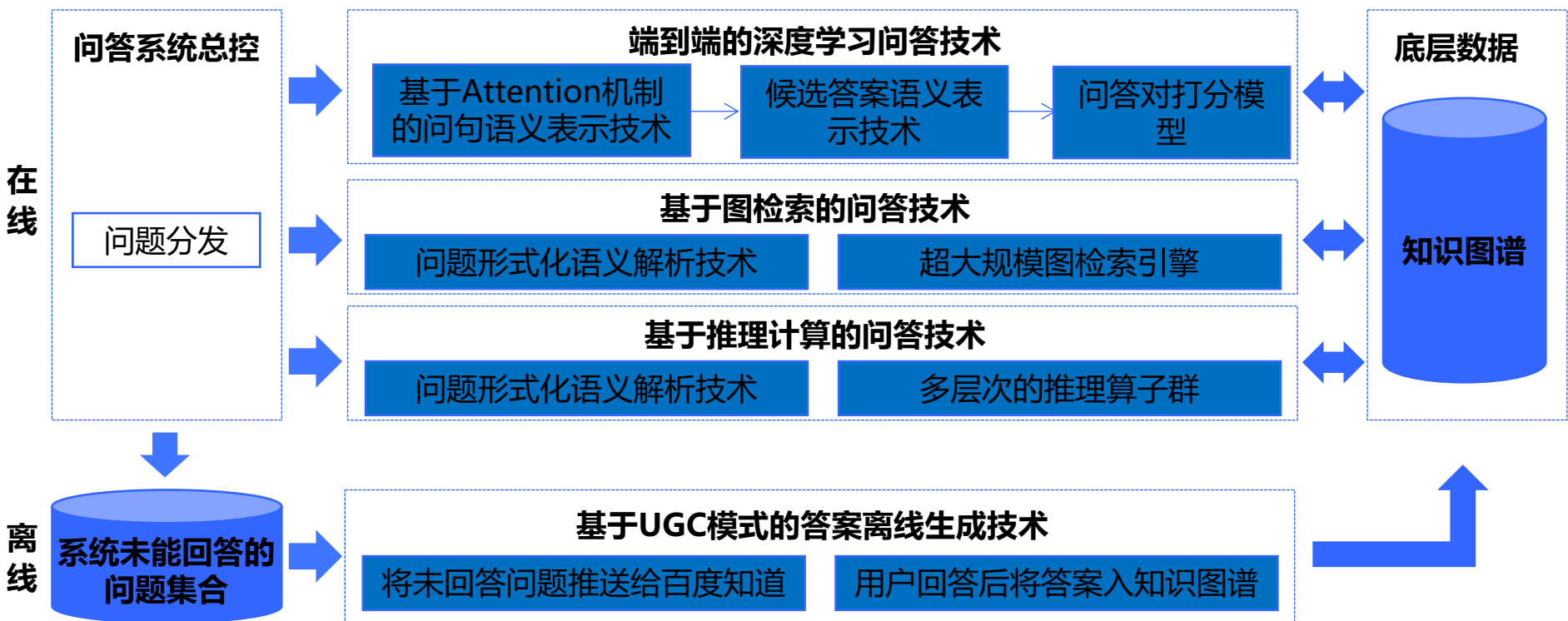
- 成本：自动化对接众测
- 门槛：题型设计 & 辅助工具
- 质量：自动回标 & 自动验收
- 效率：用户筛选 & 任务调度



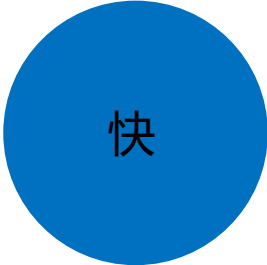
知识图谱应用技术 – 问答系统



知识图谱应用技术 – 问答系统



知识图谱应用技术 – 智能写作

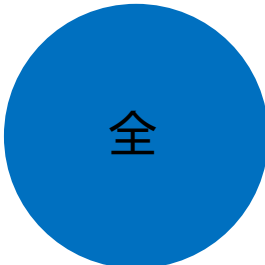


快

超高时效性

知识处理能力 + 实时计算框架

财经播报、体育快讯，快速触达用户

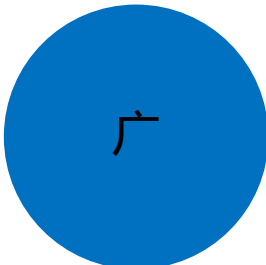


全

覆盖中长尾，补全稀缺资源

海量知识数据 + 知识理解能力

科普类、教育类、轻知识，补充内容短板，体现生态价值



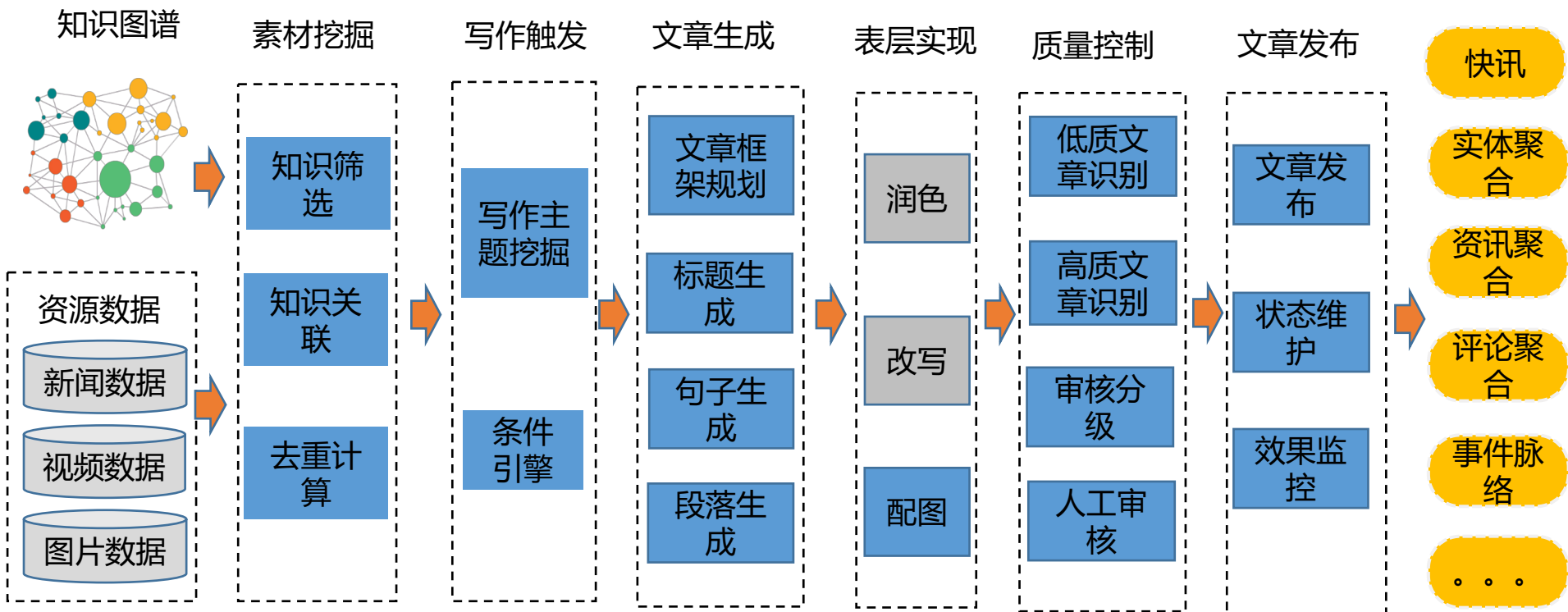
广

多维度聚合，提高获取信息效率

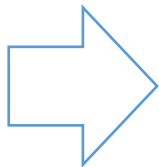
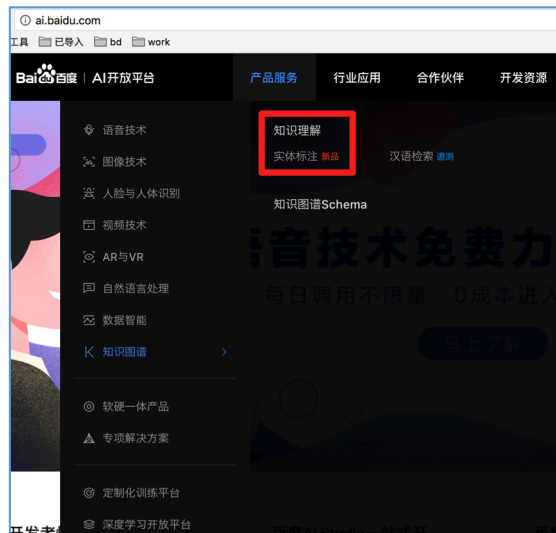
知识理解能力 + 知识计算能力

实体聚合、资讯聚合、评论聚合，不同维度信息整理，满足个性化知识需求

知识图谱应用技术 – 智能写作

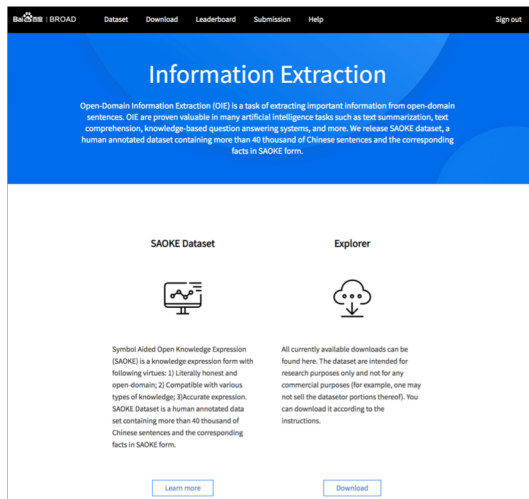


知识图谱开放

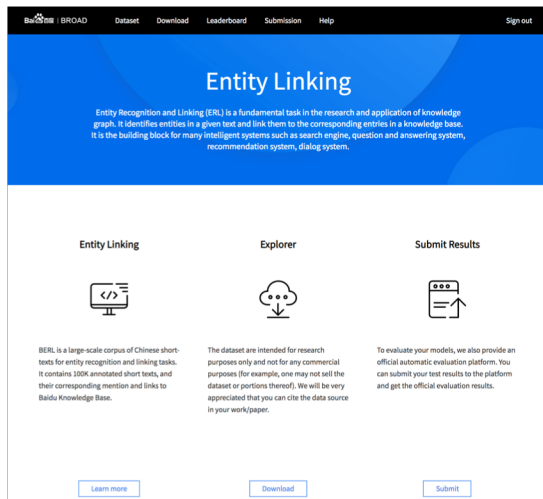


爱奇艺头条实体标注服务产品示例

知识图谱开放



- 在SPO三元组关系抽取/关系分类领域中，规模最大、关系类型最丰富的中文数据集。
- 文本来源多样，关系类型贴近真实应用。
- 包含50条Schema数据，22万句子，及对应的44万SPO三元组数据

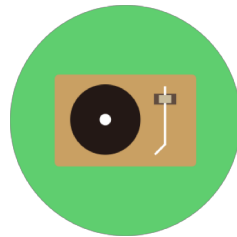


- 大规模的真实场景的中文短文本实体标注开放数据集(共12万)，高准确率（全人工标注，准确率96%+）
- 面向通用实体（通用领域的实例、概念）的标注集

知识图谱展望



文本



多媒体

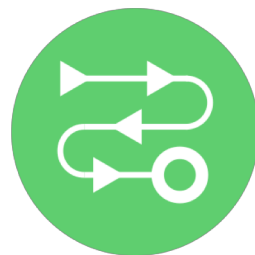
知识图谱展望



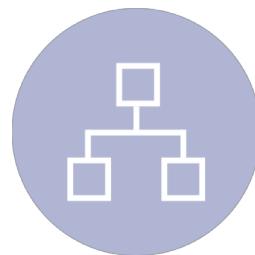
SPO



流程



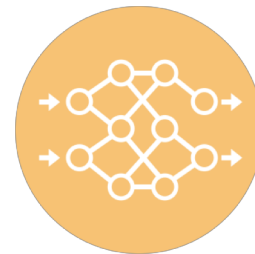
时序



逻辑



函数



模型

知识图谱展望



谢谢