

法律声明

□ 本课件包括：演示文稿，示例，代码，题库，视频和声音等，小象学院和讲者共同拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意，我们将保留一切通过法律手段追究违反者的权利。

□ 课程详情请咨询

■ 微信公众号：小象

■ 新浪微博：ChinaHadoop



第十一讲：行业知识图谱构建与应用

目录



Part 1 行业知识图谱简介

Part 2 行业知识图谱应用

Part 3 行业KG应用挑战

Part 4 行业知识图谱生命周期

01

行业KG简介

谷歌知识图谱: Things not strings

despicable me 2

Web Images Maps Shopping News More Search tools

About 163,000,000 results (0.29 seconds)

Despicable Me 2 showtimes for San Francisco, CA
See showtimes for 3D

1hr 38min - Rated PG - Animation
In summer 2013, get ready for more Minion madness in Despicable Me 2. Chris Meledandri and his acclaimed filmmaking team...

AMC Van Ness 14 - 1000 Van Ness Avenue, San Francisco, CA - Map
11:25am - 2:05 - 4:55 - 7:40 - 10:30pm

Century San Francisco Centre 9 and XD -
835 Market St., San Francisco, CA - Map
7:00 - 9:25pm

Trailer

Show more theaters

Despicable Me 2
despicableme.com/

A short description of the movie, ratings, release date, directors, cast, etc.

Despicable Me 2 movie

Despicable Me 2
192,648 followers on Google+

★★★★★ 7.8/10 - IMDb
★★★★★ 75% - Rotten Tomatoes

Despicable Me 2 is a 2013 American 3D computer-animated comedy film and the sequel to the 2010 animated film Despicable Me.

Wikipedia

Release date: July 3, 2013 (USA)
Directors: Pierre Coffin, Chris Renaud
Language: English
Production company: Illumination Entertainment
Music composed by: Pharrell Williams, Heitor Pereira

Recent posts

Voting closes soon for the Evil Laugh Contest. Make sure you get your votes in or else. MUHAHAHAHA! <http://www.evillab.com/>
Jul 24, 2013

Cast

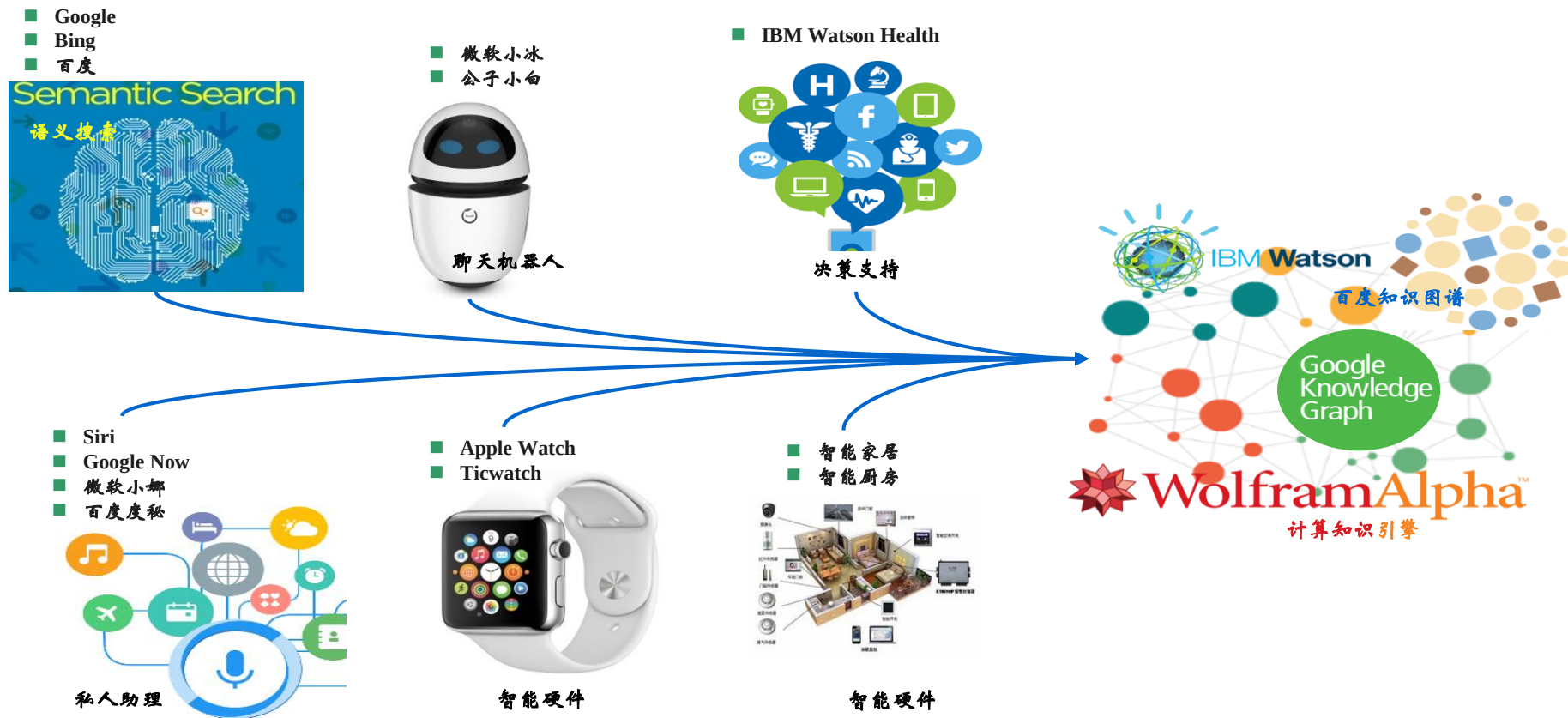
Steve Carell
Kristen Wiig
Miranda Cosgrove
Russell Brand
Steve Coogan

People also search for

Despicable Me 2010
Monsters University 2013
The Lone Ranger 2013
Man of Steel 2013
The Smurfs 2 2013



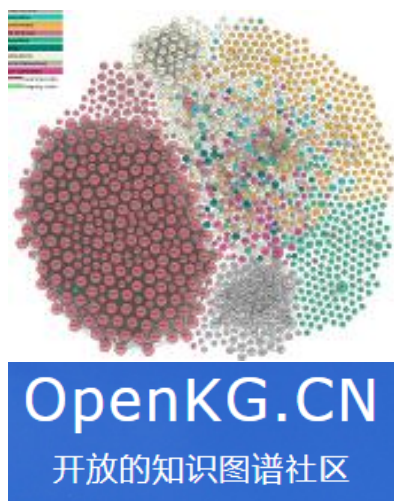
知识图谱助力人工智能应用



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，它强调的是**广度**，因而强调更多的是**实体**，很难生成完整的全局性的本体层的统一管理。

通用知识图谱相关项目



语言学类:

WordNet

MIT - ConceptNet5 的中文部分

汉语开放词网 (Chinese Open WordNet)

百科类:

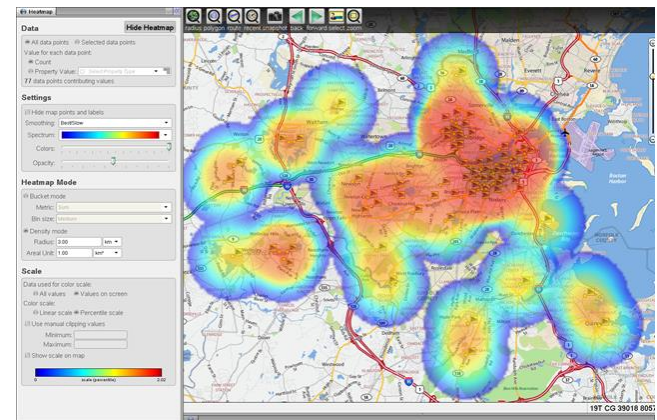
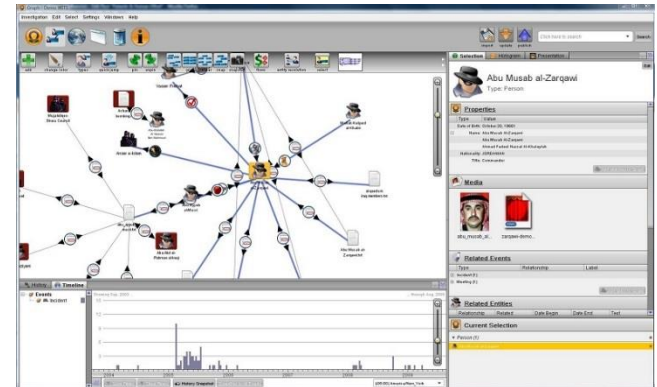
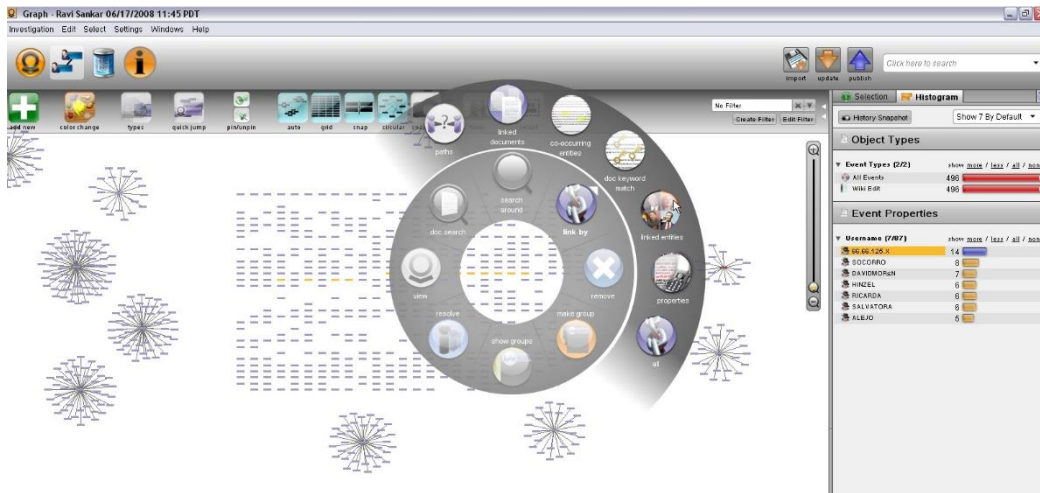
DBpedia

中文通用百科知识图谱 (CN-Dbpedia)

Zhishi.me

PKU-PIE 知识库

行业知识图谱: Palantir



行业知识图谱

- 行业知识图谱指面向**特定领域**的知识图谱。
- 用户目标对象需要考虑行业中各种级别的人员，不同人员对应的操作和业务场景不同，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**严格与丰富的数据模式**，行业知识图谱中的实体通常属性比较多且具有行业意义。

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，行业应用的数据的数量通常都以亿级别计算，存在通常在TB、PB级别甚至更多

行业知识图谱应用一览



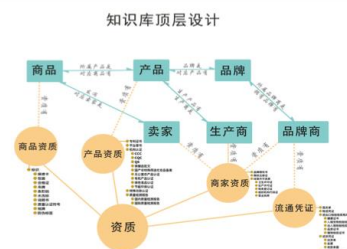
金融证券



生物医药

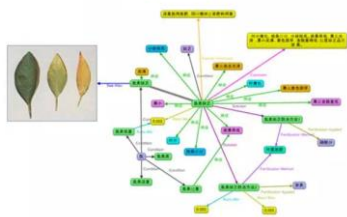


图书情报



电商

Complete Ontology (knowledge) for Lacking Nitrogen



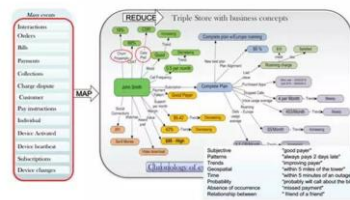
农业

Palantir government

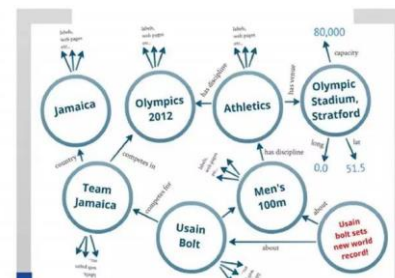


政府

Turn Massive Raw Data Into Business Concepts

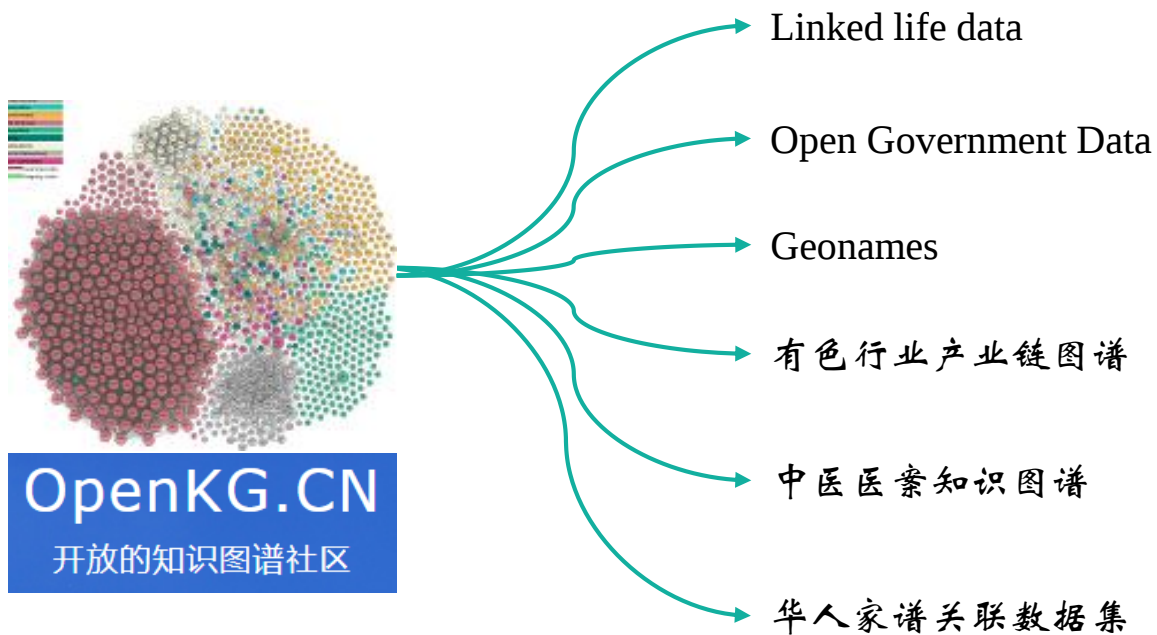


电信



出版

行业知识图谱相关项目



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员

通用知识图谱 + 行业知识图谱

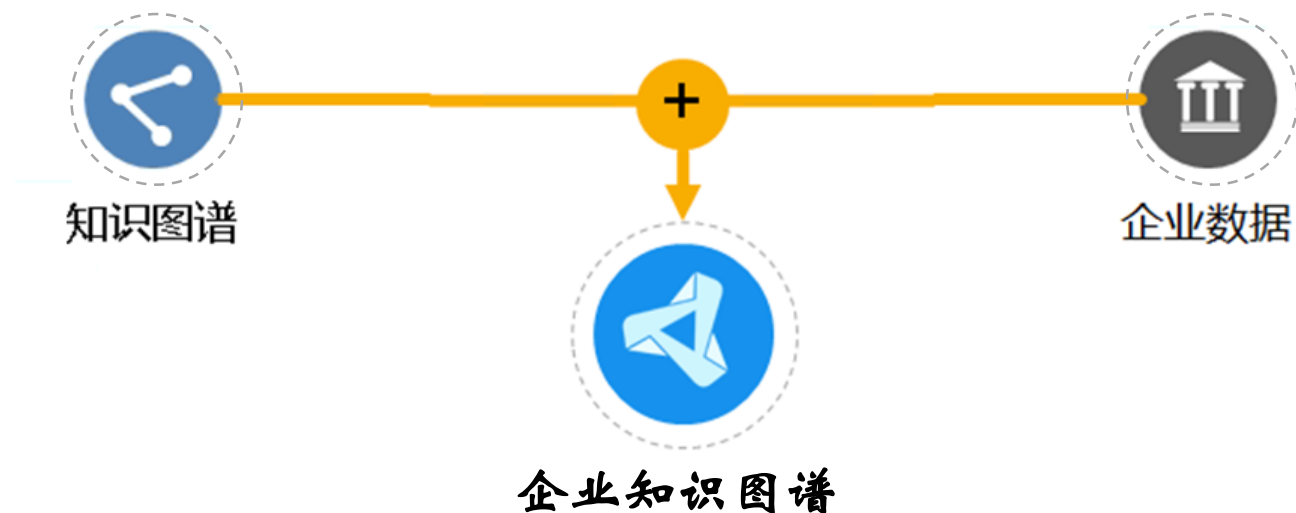
- 通用知识图谱的广度，行业知识图谱的深度，**相互补充**，形成更加完善的知识图谱。
- 通用知识图谱中的知识，可以作为行业知识图谱构建的基础；而构建的行业知识图谱，再融合到通用知识图谱中。



02

行业KG应用

金融证券——企业知识图谱



企业基础数据
投资关系
任职关系

企业专利数据
企业招投标数据
企业招聘数据

企业诉讼数据
企业失信数据
企业新闻数据

企业知识图谱应用——企业风险评估

基于企业的基础信息、投资关系、诉讼、失信等多维度关联数据，利用图计算等方法构建科学、严谨的企业风险评估体系，有效规避潜在的经营风险与资金风险。

用户群体

银行、担保、投行、政府……

应用环节

客户资源分类管理

信贷前期风险评估

采购企业风险审核

招投标企业资质评级

……

综合评估 (2015-05)



综合信用等级

AAA



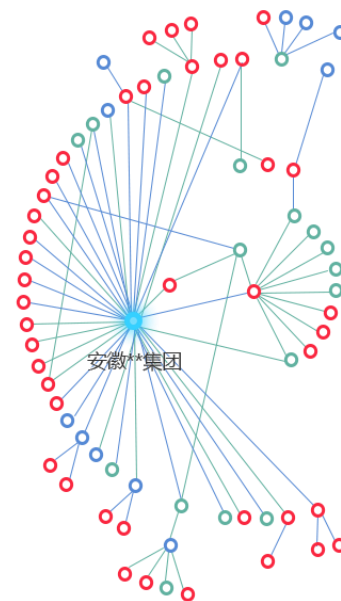
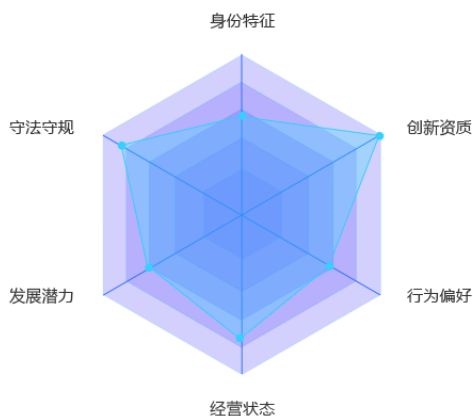
综合信用评价

低风险



综合信用评分

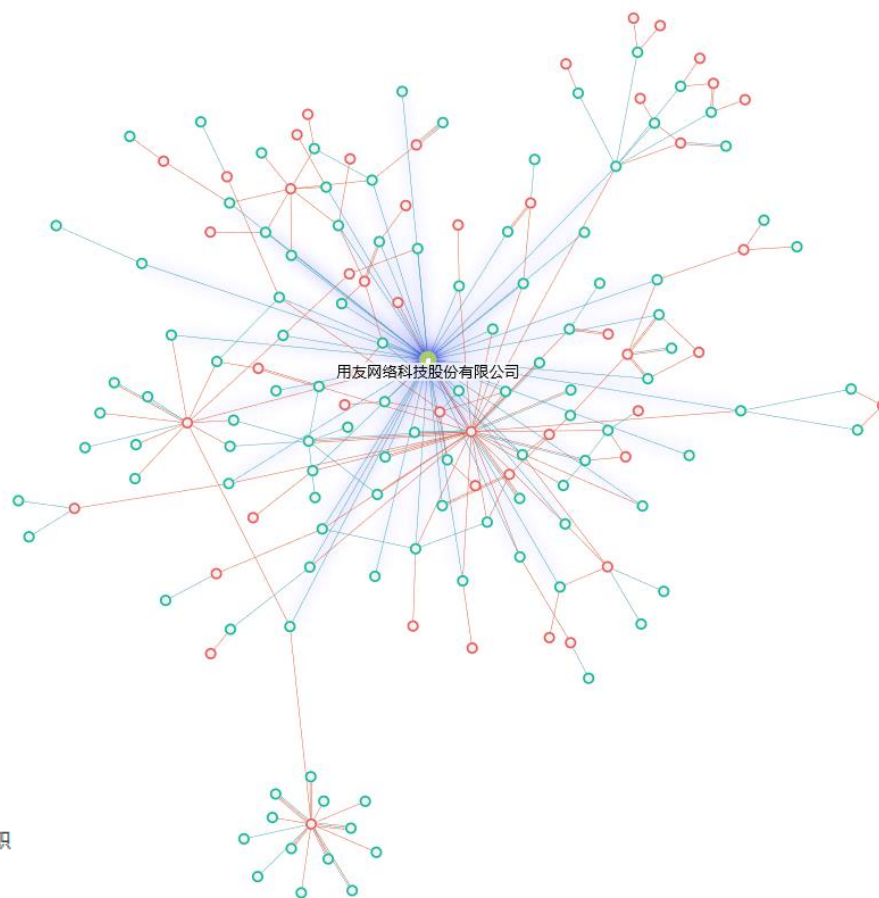
757



● 低风险 ● 中等风险 ● 高风险

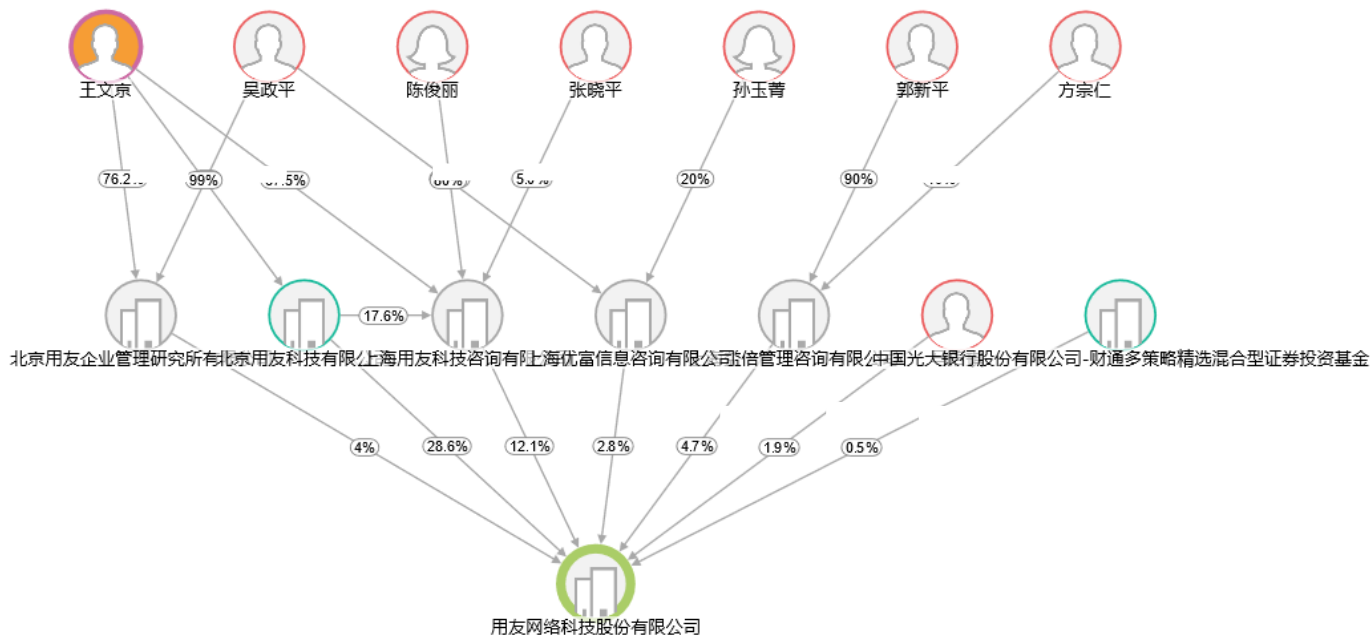
企业知识图谱应用——企业社交图谱查询

□ 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。



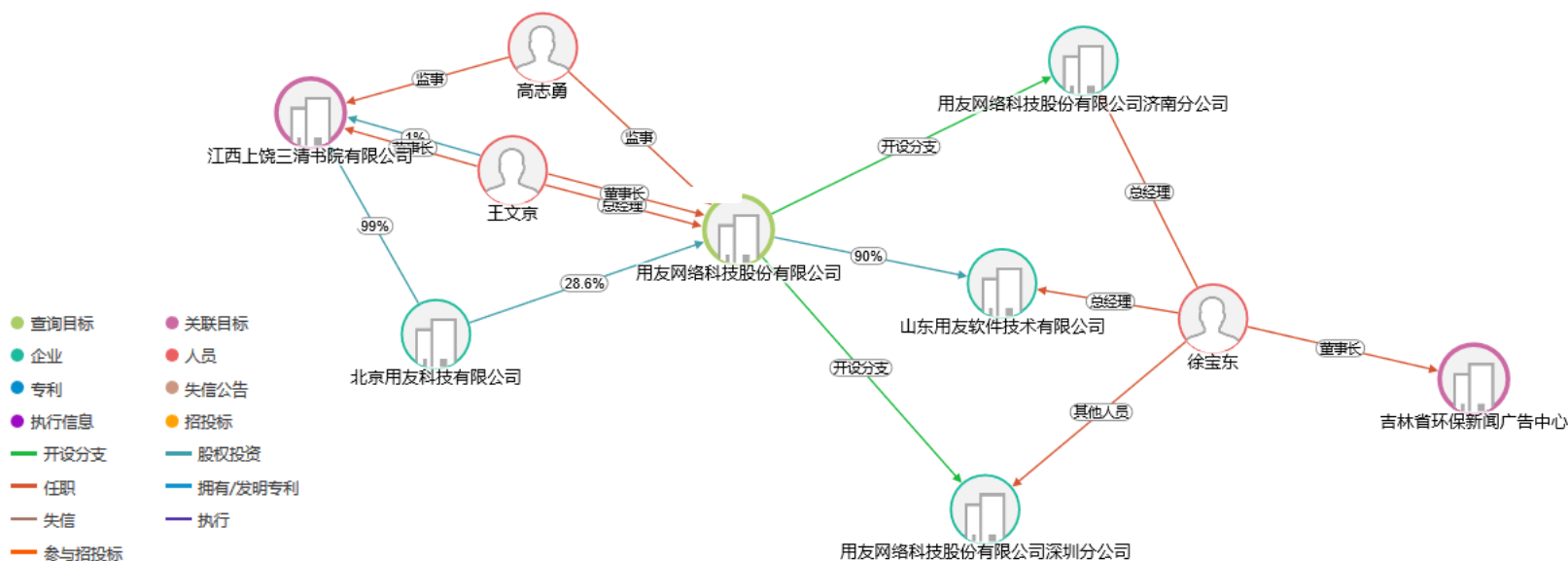
企业知识图谱应用——企业最终控制人查询

□ 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門。



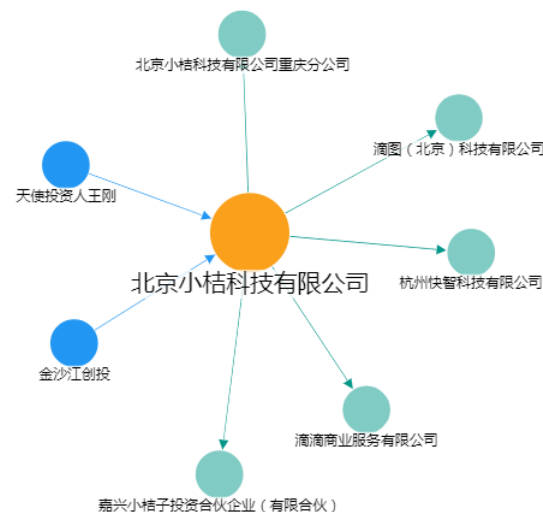
企业知识图谱应用——企业之间路径发现

■ 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。



企业知识图谱应用——初创企业融资发展历程

■ 基于企业知识图谱中的投融资事件发生的时间顺序，记录企业的融资发展历程。



企业知识图谱应用——上市企业智能问答

股票问问

股票问问

搜索

股票查询

- 中信证券上市时间
- 中信证券发行价格
- 注册资本最大的公司
- 保荐机构是中信建投的股票

板块查询

- 迪士尼板块的股票
- 上海食品加工板块的股票
- 上海物联网板块股票负面
- 北京 石油化工板块 股票研究报告

人物查询

- 中信证券法定代表人
- 万达董事长
- 万达高管
- 王健林

关系发现

- 万达信息 上海复高
- 万达信息 宝信软件
- 万达信息 华为
- 万达院线 王健林

事件追踪

- 王健林负面
- 万达信息并购
- 中信证券负面
- 中信证券热点

您的历史搜索

- 中信证券上市时间
- 中信证券上市时间
- 中信证券上市时间

智能搜索答案

1 中信证券的上市时间

2003-01-06

2 注册资本最大的公司

丰田汽车公司

3 地域板块是北京且行业板块是石油化工的

京威股份 福田汽车 七星电子 中国石化

4 万达信息的总经理

史一兵

5 万达信息的董事会秘书

6 万达院线的总经理

7 万达院线的董事会秘书

新闻资讯 负面新闻 热点事件

1 券商“吸金”时代：一边是牌照放开，一边则扎堆“出海”，赴美产业格局重塑 旅游

2 “高田气囊门”：昨日，日本高田公司的场，根据中国国家质检

3 中石化一季度净利十年最低：跌87.5%

4月29日晚间，中石化(微博)发布今年一季度财报，87.5%。这是近十年，中石化...

天信投资：市场变幻 逢低买入勿追高

热点栏目资金流向千股千评个股诊断最新评级反弹。权重板块上，运输设备、航空、有色形

新闻资讯 负面新闻 热点事件 并购事件 公司公告 研究报告

69家公司首推员工持股计划 普惠政策成股价新引擎 2015-04-19 搜狐证券 正面

4月13日，招商银行(600036)抛出一份60亿元的员工持股计划，所有对该行整体业绩和中长期发展具有重要作用的人。此前，招商银行在停牌期间曾一...

4 今年69家公司首推员工持股 普惠政策成股价新引擎 2015-04-18 东方财富网 财经 正面

金融证券——金融交易知识图谱



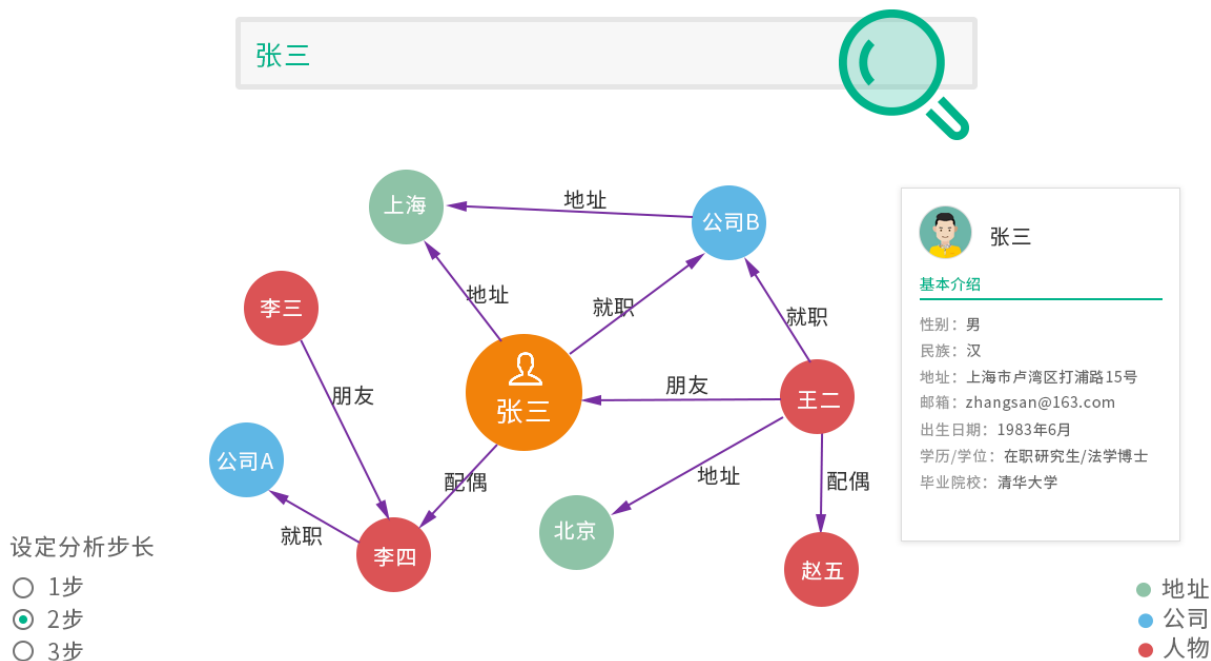
企业知识图谱

交易客户数据
客户之间的关系

交易行为数据

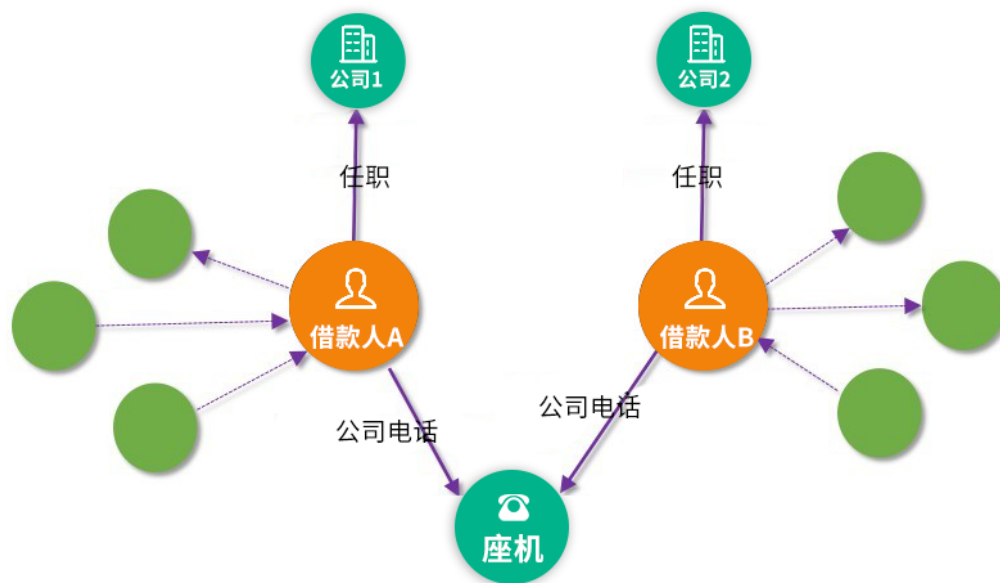
金融交易知识图谱应用——辅助信贷审核

- 基于知识图谱数据的统一查询，全面掌握客户信息；避免由于系统、数据等孤立造成的信息不一致造成信用重复使用、信息不完整等问题。



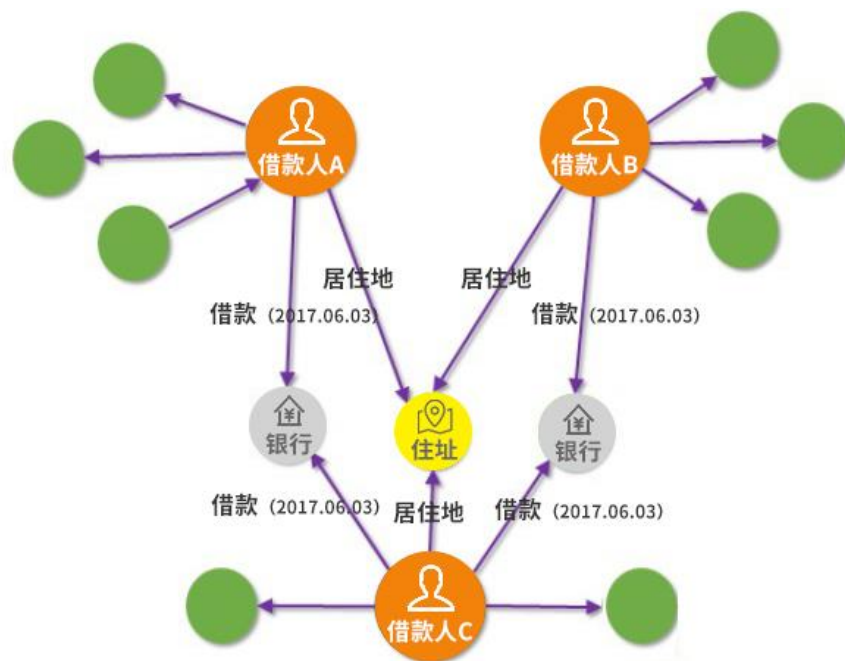
金融交易知识图谱应用——反欺诈 (1)

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。



金融交易知识图谱应用——反欺诈 (2)

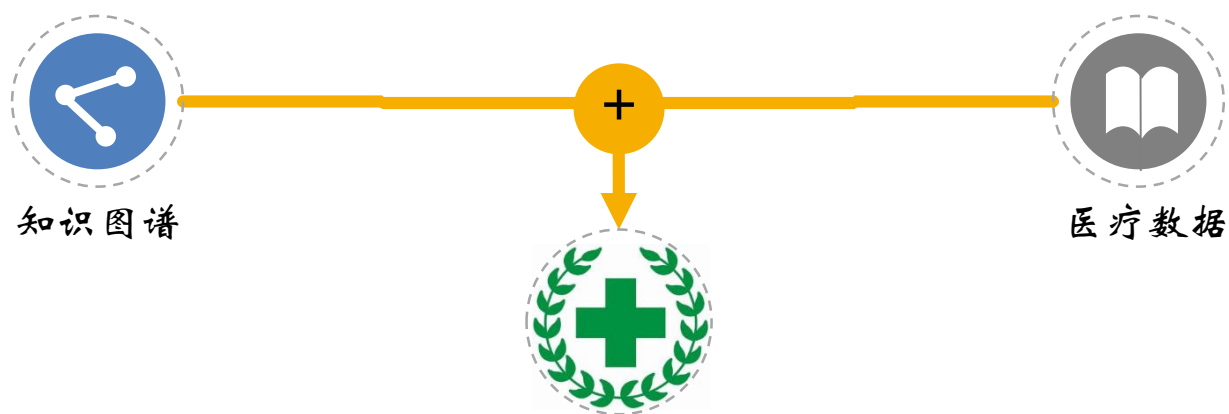
组团进行欺诈的成员会用虚假的身份去申请贷款，但部分信息是共享的。如下图可以看出贷款人A、B和C之间没有直接的关系，但通过知识图谱可以很容易的看出这三者之间都共享着某一部分信息，存在一定的组团骗贷风险。



金融证券——其它应用场景

- 异常分析（异常交易、异常客户）
- 失联客户管理
- 精准营销
- 智能投研
- 智能公告
-

生物医疗——医疗知识图谱



医疗知识图谱

医疗专业知识
医疗文献
医疗常识

电子病历大数据
医案

现有医疗资源
疾病库
指南与规范

医疗知识图谱应用——中医药知识平台

- 针对中医药知识体系系统梳理、建模和展示
- 以图形可视化方式展示核心概念之间的关系
- 辅助中医专家厘清学术发展脉络，浏览中医知识，发现知识点之间的联系。
- 与阅读文献等手段相比，可大幅度节约知识检索获取时间。

中医药知识服务平台

搜索.....

点击 [注册](#) / [登录](#) , 访问更多资源 ▾

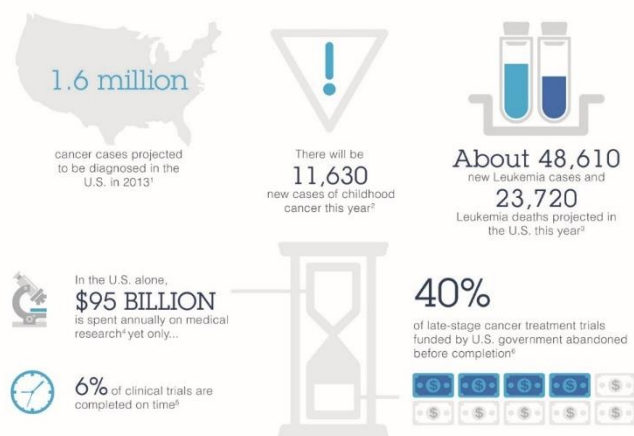
方药 集成包括中药、中成药、西药、院内制剂、药典中药数据在内的各类数据。	名医经验 整合名医医案、名方剂、名医理论、名医介绍等知识资源。	循证 整合系统评价、文献质量评价、临床研究等循证医学数据。	指南与规范 整合国家权威机构发布的指南、编码、术语等数据。
养生 整合中医养生文献，整合各方面中医养生知识。	诊疗技术 整合中医特色疗法、诊疗技术及其评价等临床知识资源。	文献 整合报纸、期刊、古代文献、外文文献等数据资源。	本体 收集证候、临床、养生、中药、中医文献等方面的本体。

<http://www.tcmkb.cn>

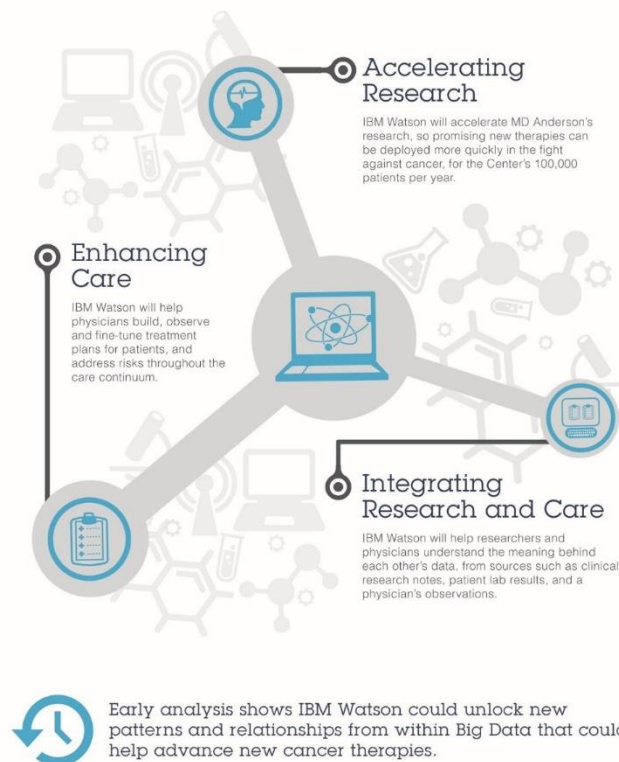
生物医疗——Watson辅助诊断与治疗

安德森癌症中心联合IBM Watson开展终结癌症的任务。

Going Up Against a Deadly Disease



How Watson is Helping the Fight



生物医疗——Open PHACTS 新药物发现



Bringing together **pharmacological data resources** in an integrated, interoperable infrastructure

Explore.

Researchers can use Open PHACTS to access vast amounts of pharmacological data, all from a single, simple interface

Build.

Developers get free access to the Open PHACTS API, to query the pharmacological data resources in our integrated triple store

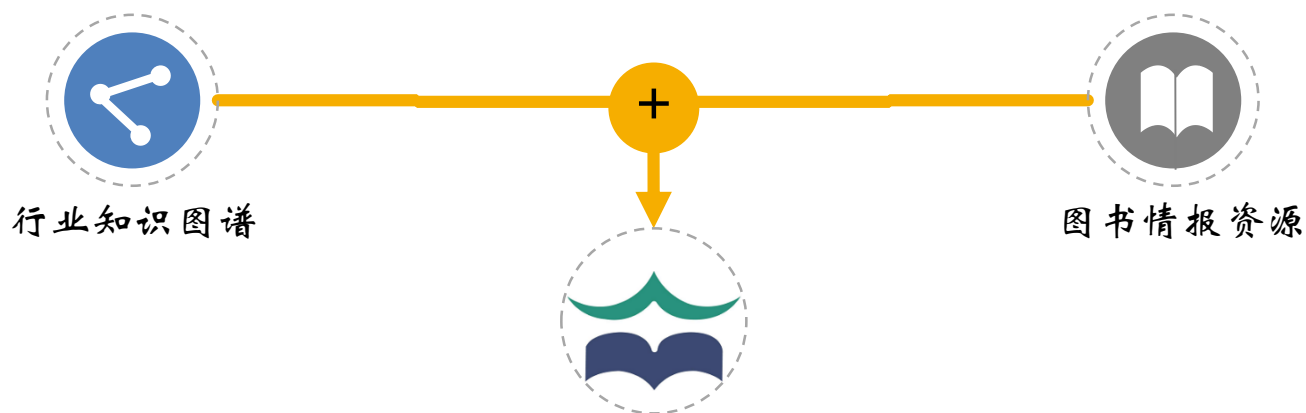
Join.

Members of the Open PHACTS Foundation get prioritised access to data, support and updates, as well as training opportunities

欧盟重大联合攻关项目

面向药物研发的开放数据访问平台开发，其核心技术就是采用语义技术为有关研究人员提供高效的数据访问技术环境的支持。

图书情报——图情资源知识图谱



图情资源知识图谱

图书馆分类学体系
特定方向的知识体系

图书
期刊
论文
专利
报刊

百科数据
行业网站数据

图情资源知识图谱应用——知识导航与资源展示

- 使用知识图谱中的知识体系进行知识导航，引导用户学习知识体系，以及通过实体链接所关联的资源。

响应式系统

主要组成技术

响应式微服务

角色模型

响应云

响应流

快速数据

Kafka

Spark

快数据存储

响应式编程

Kafka (65)

Apache Kafka for Beginners 论文

Gwen Shapira & Jeff

Abstract

When used in the right way and for the right use case, Kafka has unique attributes that make it a highly attractive option for data integration. Apache Kafka is ...

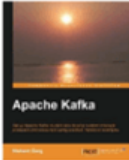
Kafka: a Distributed Messaging System for Log Processing 论文

J Kreps, L Corp

Abstract

Log processing has become a critical component of the data pipeline for consumer internet companies...

Apache Kafka 图书



作者:暂无作者信息
出版社:暂无出版社信息
时间:2013
页数
ISBN:Y9781782167938
资源: 当当网

图情资源知识图谱应用——知识点推荐与搜索

 HUAWEI

Big Data

搜一下

全部 新闻资源 文章资源 图书资源 演讲资源 开源项目资源 论文资源 专利资源

搜索到约 10835 条结果

 **Big Data**
Big data is a term for data sets that are so large or complex that traditional data processing application software is inadequate to deal with them. Challenges include capture, storage, analysis, data curation, search, sharing, transfer, visualization, querying, updating and information privacy. The term "big data" often refers simply to the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from data, and seldom to a particular size of data set.

Safety or no safety in numbers? Governments, big data and public policy formulation.
20150101
Although big data have emerged at the cornerstone of business and management research, past studies have failed to offer explanations and classifications of different levels of capacity and expertise possessed by different countries in utilising big data. The purpose of this paper is to examine the different capacities of governments in utilising big data.

SAFE - Secure and Big Data-Adaptive Framework for Efficient Cross-Domain Communication.
20140101
Today's Cross Domain Communication (CDC) Infrastructure primarily consists of vendor-specific guard products that have little inter-domain coordination at runtime. Unaware of the context and the semantics of the CDC message that is being processed, the guard heavily relies on rudimentary filtering techniques.

Safer@Home Analytics - A Big Data Analytical Solution for Smart Homes.
20130101
Abstract: The vast amounts of data generated from sensors in smart homes, can give valuable insights about social and behavioral patterns on households and their residents. The goal of the project is investigation & implementation of mechanisms to capture/store vast continuous streams of time-series data from optical movement sensors

Big Data
所属： 技术关键词

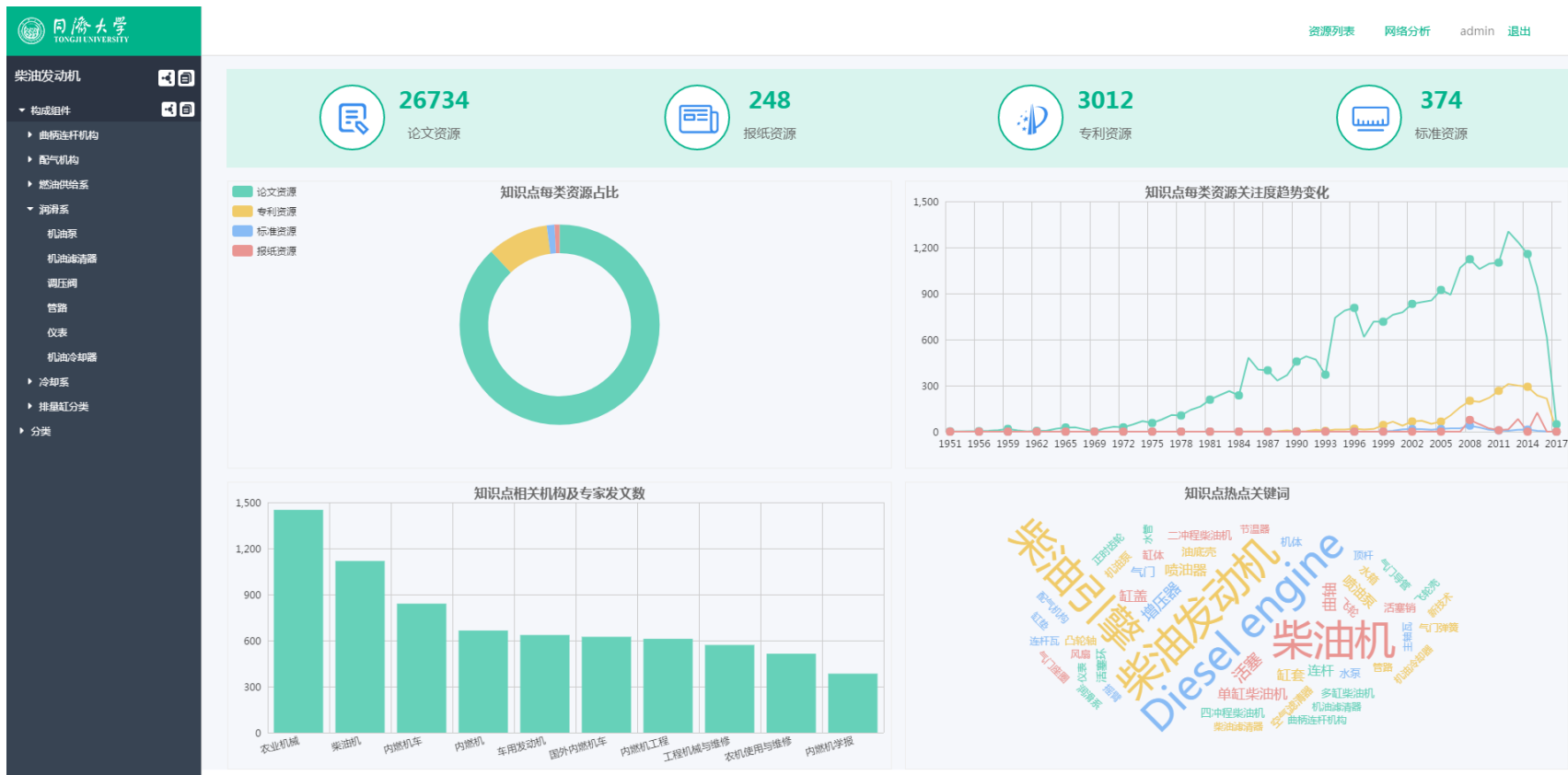
简介
Big data is a term for data sets that are so large or complex that traditional data processing application software is inadequate to deal with them. Challenges include capture, storage, analysis, data [查看全部](#) >>

同义词
-

相关推荐
 EMC
 IBM

相关专家推荐
 Jan L. Bruse
 Arian Bär
 Robin G. Qiu

图情资源知识图谱应用——图情资源统计



知识图谱行业应用——其它行业

□ 农业

- 识别作物危害

□ 政府行业

- 政府大数据管理

□ 客服系统

- 基于知识图谱的智能客服系统

□

03

行业知识图谱 应用挑战

从数据库发展到大数据时代，企业希望融合使用全量数据

DB → BD

融合使用全量数据会遇到
哪些挑战呢？

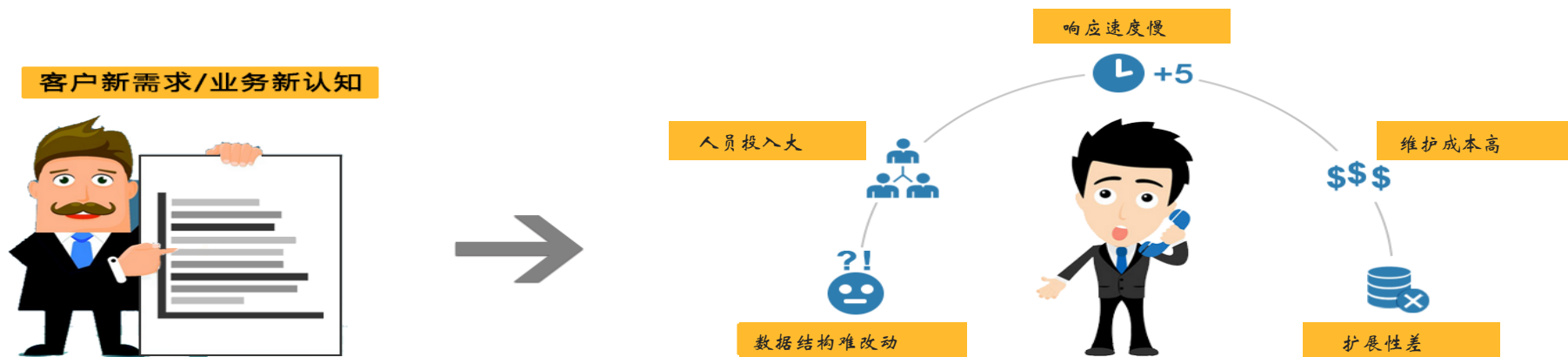
	数据库时代	大数据时代
数据规模	小 MB/GB	大 TB/PB/ZB
数据类型	少 结构化数据为主	多 包含结构化、半结构化、非结构化数据，且后两者越来越多
数据模式	可预先确定 先有数据模式后产生数据； 数据模式相对固定；	无法预先确定 模式在数据出现之后才能确定； 数据模式随数据增长不断演变
处理方法	One Size Fits All	No Size Fits All

企业全量数据应用挑战1: 多源异构数据难以融合



信息聚合、数据融合需求迫切!

企业全量数据应用挑战2: 数据模式动态变迁困难



当前数据模式动态变迁困难,当客户新需求、业务新认知时程序员需痛苦的修改数据结构及业务逻辑,带来扩展性差、对客户响应慢、维护成本高等不良情况。

我们需要: 可自由扩展的数据模式!

企业全量数据应用挑战3: 非结构化数据计算机难以理解

军工科研院所分类改革方案已发至各大军工集团

作者: 陈健健 整理 来源: 中国证券报·中证网 2017-01-11 07:54

“兵工集团非常愿意以资产证券化的方式推动混改, 这样可以解决一股独大的局面。公司化治理是内部运作、外部参与, 操作起来最容易。”李锦说, “在引入社会资本方面, 混改红利会比较快兑现, 空间也比较大。”

Web of Document



人脑



计算机



正确获取文本内信息

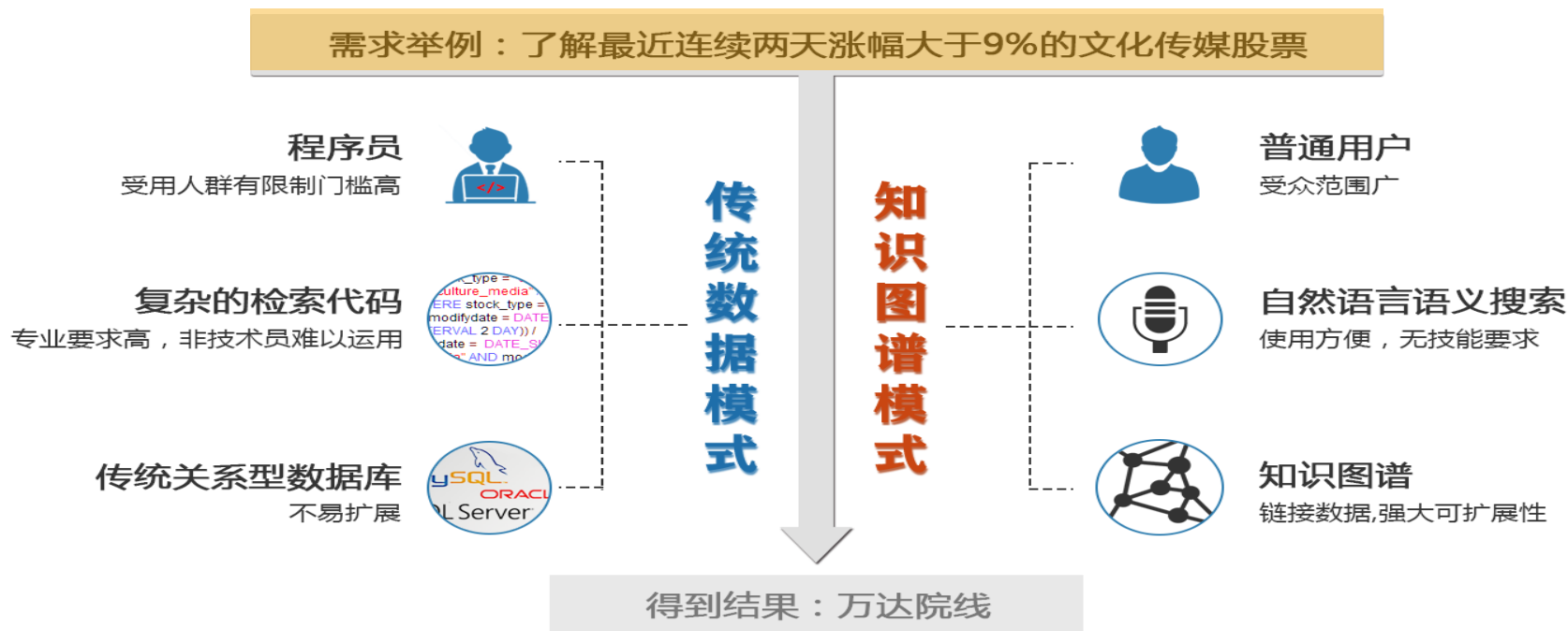


非结构化数据计算机难以正确理解

计算机无法理解非结构化数据的语义

企业迫切需要将非结构化数据结构化

企业大数据应用挑战4: 数据使用专业程度过高



行业智能问答大幅降低数据使用门槛

企业大数据应用挑战5: 分散的数据难以统一消费利用

X 业务系统繁多

X 使用方式各异

X 难以全局把握

✓ 可视化

✓ 统一使用入口

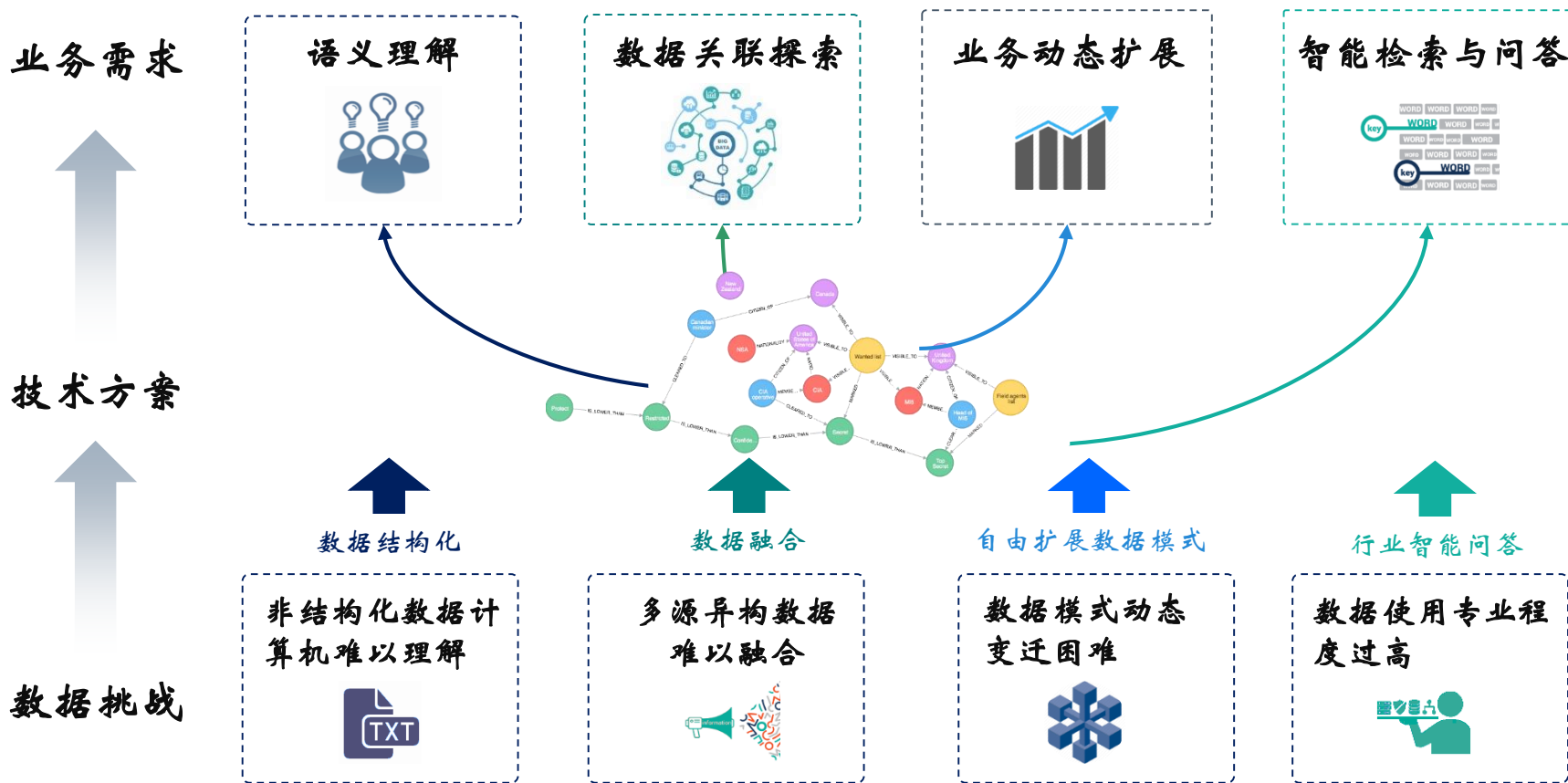
✓ 匹配度高

基于知识图谱数据存储、融合、分析统一平台，为用户提供统一的消费入口，以不同的形态(检索、可视化、分析等)展示给用户。

解决方案：基于行业知识图谱进行数据融合使用

- 挑战1：使用知识图谱（本体）对各种类型的数据进行抽象建模，**基于可动态变化的“概念—实体—属性—关系”数据模型**，实现各类数据的统一建模。
- 挑战2：使用**可支持数据模式动态变化的知识图谱的数据存储**，实现对大数据及数据模式动态变化的支持。
- 挑战3：利用信息抽取技术，**对非结构化数据及半结构化数据进行抽取和转换**，形成知识图谱形式的知识。
- 挑战4、5：在**知识融合**的基础上，基于**语义检索、智能问答、图计算、推理、可视化**等技术，提供统一的数据检索、分析和利用平台。

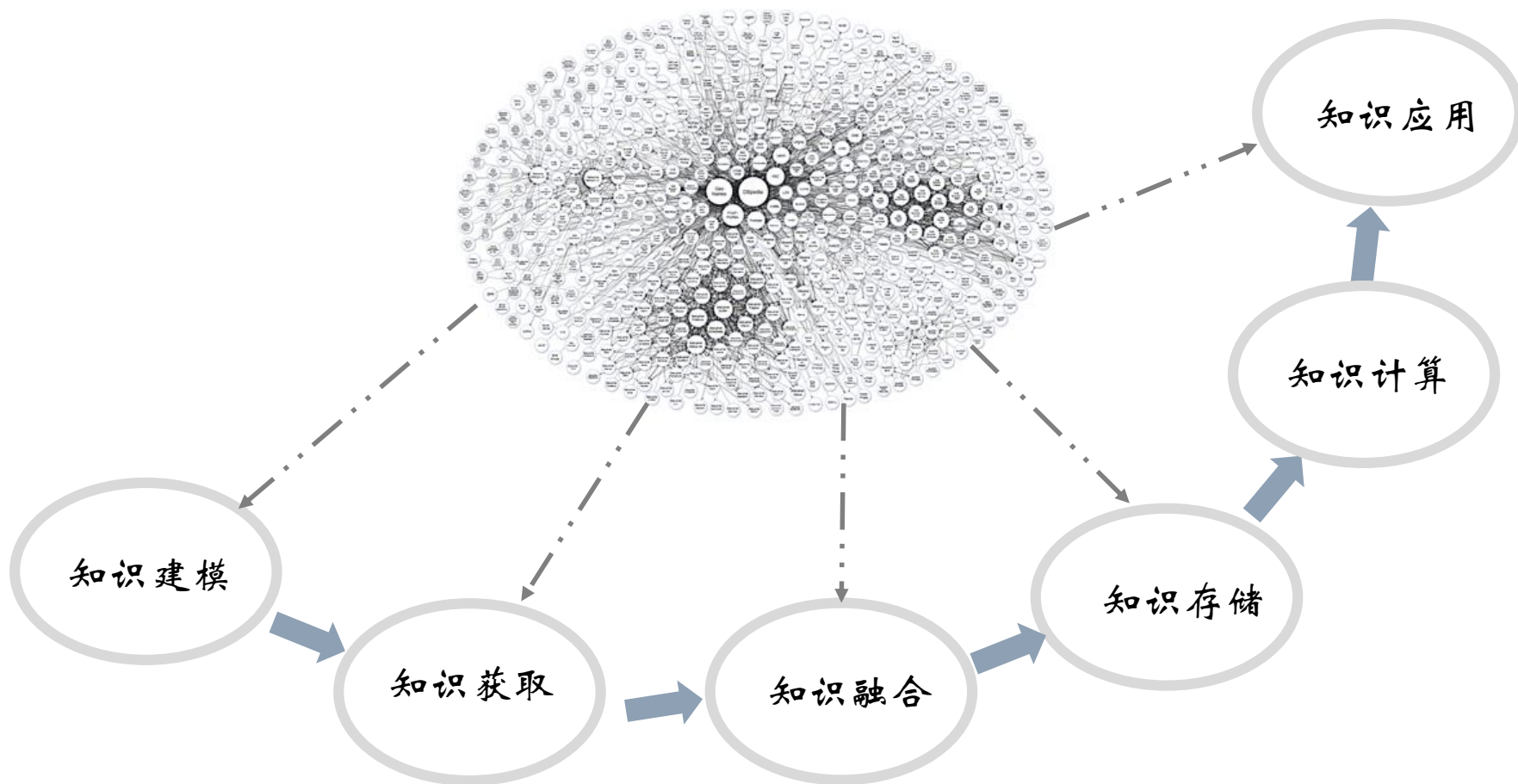
知识图谱助力企业商业智能



04

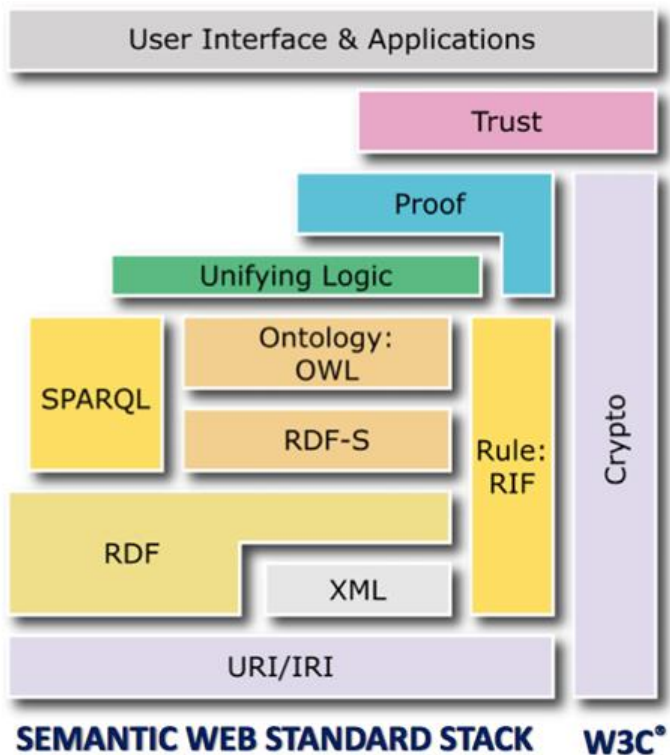
行业知识图谱 生命周期

行业知识图谱生命周期

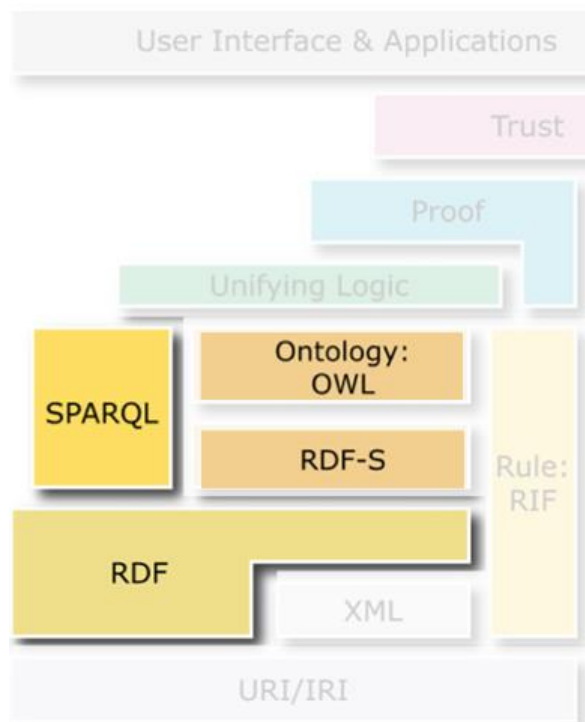


知识图谱基础技术规范

W3C推荐的语义网标准栈

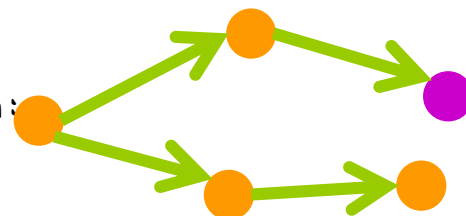
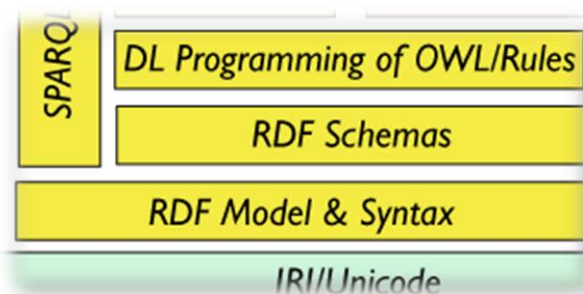


行业知识图谱主要技术标准

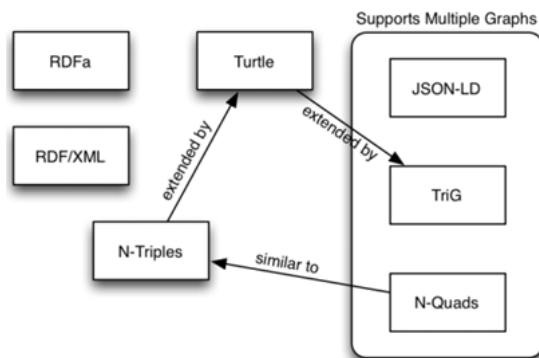


RDF

- ❑ RDF是语义网标准中的第一层
- ❑ RDF 代表
 - **R**esource: 页面、图片、视频等任何具有URI标识符
 - **D**escription: 属性、特征和资源之间的关系
 - **F**ramework: 模型、语言和这些描述的语法
- ❑ RDF是一个三元组 (triple) 模型，即每一份知识可以被分解为如下形式：
(subject (主), predicate (谓), object (宾))
- ❑ RDF 是一个链接资源描述的图模型，其三元组可看作图中的弧
- ❑ RDF其它语法：Turtle、TriG、N-Triples、N-Quads、JSON、RDFa



(顶点, 边, 顶点)



OWL: RDF Schema的扩展

❑ 复杂类：交、并、补

复杂类

```
:Mother owl:equivalentClass [  
  rdf:type owl:Class ;  
  owl:intersectionOf ( :Woman :Parent )  
] .
```

❑ 属性约束：存在量化、全称

量化

❑ 基数约束：最大基数约束、

属性约束

```
:Parent owl:equivalentClass [  
  rdf:type owl:Restriction ;  
  owl:onProperty :hasChild ;  
  owl:someValuesFrom :Person  
] .
```

最小基数约束

❑ 属性特征：反、对称、非对

称、不相交、自反

对称属性

```
:hasSpouse rdf:type owl:SymmetricProperty .
```

❑ 属性链

传递类

```
:hasAncestor rdf:type owl:TransitiveProperty .
```

属性链

```
:hasGrandparent owl:propertyChainAxiom ( :hasParent :hasParent ) .
```

SPARQL 简介

- ❑ RDF 的查询语言：基于 RDF 数据模型
- ❑ 可以对不同的数据集撰写复杂的连接 (joins)
- ❑ 由所有主流图数据库支持
- ❑ SPARQL Protocol and RDF Query Language

A SPARQL query comprises, in order:

- *Prefix declarations*, for abbreviating URIs
- *Dataset definition*, stating what RDF graph(s) are being queried
- A *result clause*, identifying what information to return from the query
- The *query pattern*, specifying what to query for in the underlying dataset
- *Query modifiers*, slicing, ordering, and otherwise rearranging query results

```
# prefix declarations
PREFIX foo: <http://example.com/resources/>
...
# dataset definition
FROM ...
# result clause
SELECT ...
# query pattern
WHERE {
    ...
}
# query modifiers
ORDER BY ...
```

A quick introduction to SPARQL, the de-facto query language for RDF. Those familiar with SQL will have an easy time understanding SPARQL... at the beginning...

```
PREFIX dc: <http://purl.org/dc/elements/1.1/>
SELECT ?title
WHERE { <http://example.org/book/book1> dc:title ?title }
```

A Simple select query

A Join

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT ?name ?mbox
WHERE { ?x foaf:name ?name . ?x foaf:mbox ?mbox . }
```

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT ?name ?mbox
WHERE { ?x foaf:name ?name . OPTIONAL { ?x foaf:mbox ?mbox } }
```

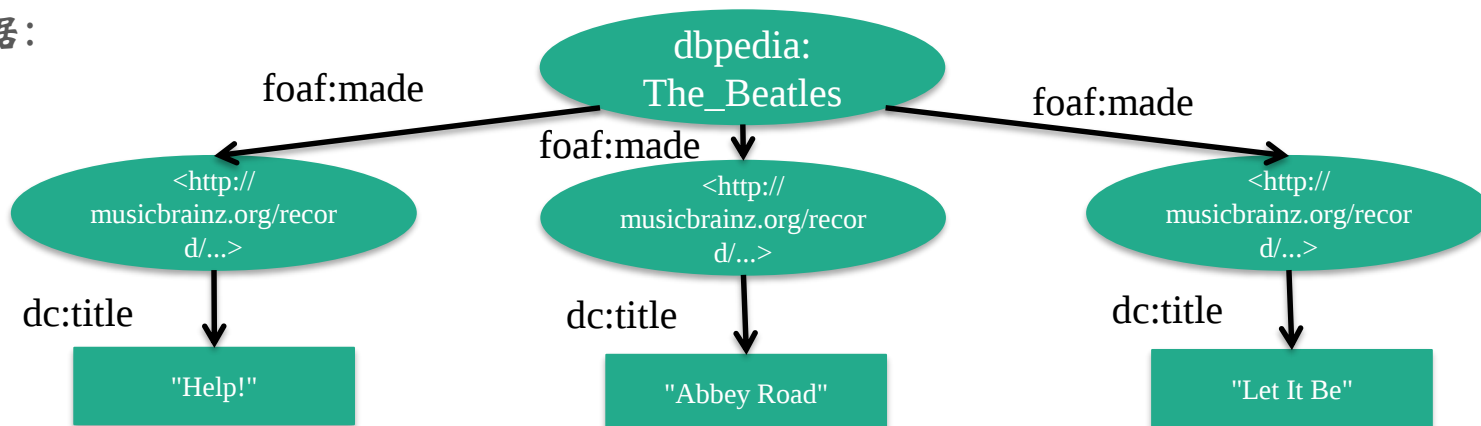
An Outer Join

Like “Like”

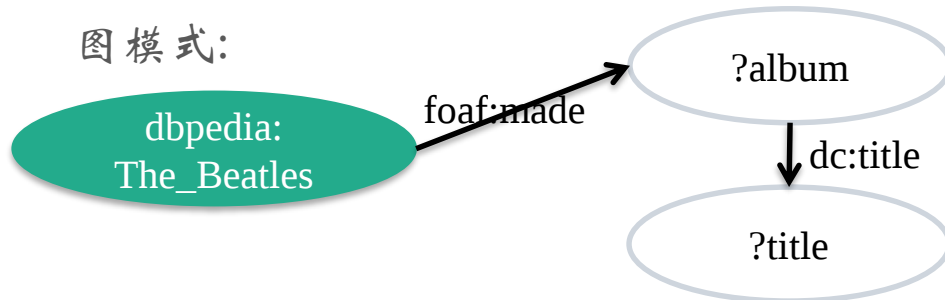
```
PREFIX dc: <http://purl.org/dc/elements/1.1/>
SELECT ?title
WHERE { ?x dc:title ?title . FILTER regex(?title, "SPARQL") }
```

SPARQL查询图可视化

数据:



图模式:



结果:

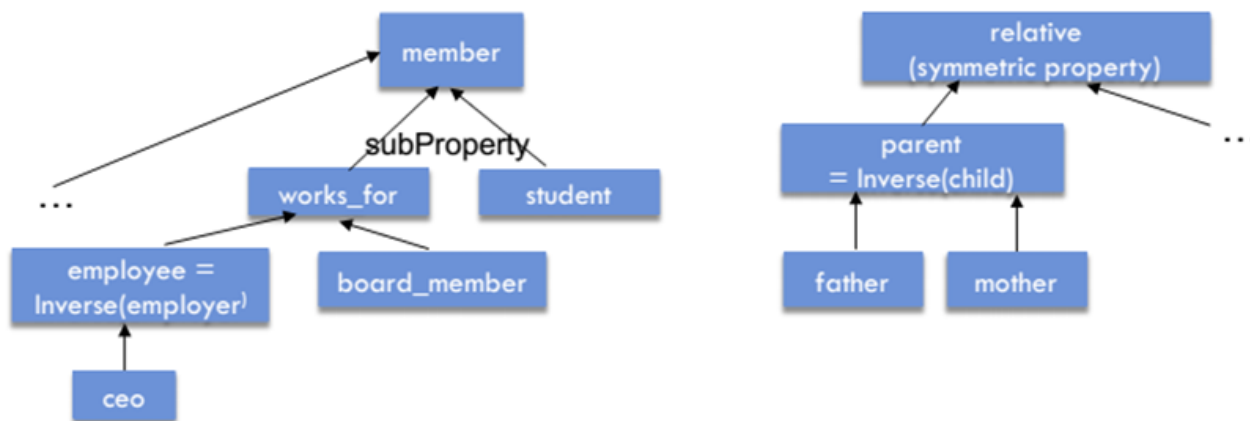
?album	?title
<http://...>	"Help!"
<http://...>	"Abbey Road"
<http://...>	"Let It Be"

本体 (ontology) 可以填充知识与查询之间的语义间隙

Let's revisit our simple conceptual query

"Find people who are members of an organization founded by a relative."

Assume some mechanism to specify relationships between predicates as follows:



Simpler precise query against the RDF data enriched with the above semantic knowledge (ontology)

```
SELECT* WHERE {  
  ?x member ?y .  
  ?z founder ?y .  
  ?z relative ?x  
}
```

1. 知识建模

□ 即建立知识图谱的数据模式。行业知识图谱的数据模式对整个知识图谱的结构进行定义，因此需要保证可靠性。

□ 常用方法

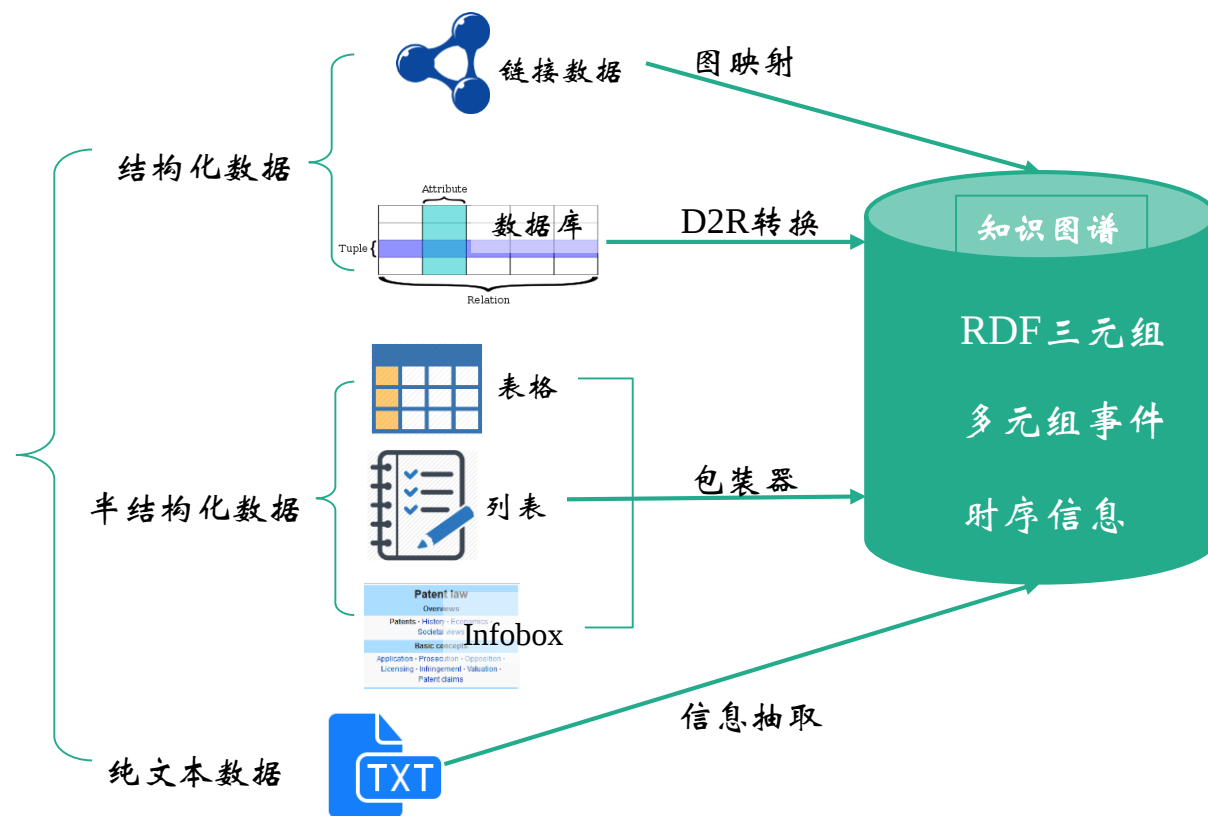
- 自顶向下的方法：专家手工编辑形成数据模式
- 自底向上的方法：
 - 基于行业现有的标准进行转换
 - 从现有的高质量行业数据源（如业务系统数据库表）中进行映射

知识建模关键技术与难点

- ❑ 多人在线协同编辑，并且实时更新
- ❑ 能够导入集成使用现有的（结构化）知识
- ❑ 支持大数据量
- ❑ 能够支撑事件、时序等复杂知识表达
- ❑ 可以与自动算法进行结合，避免全人工操作

2. 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。



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

□ 从链接数据中获取知识：图映射

- 难点：数据对齐

□ 从半结构化（网站）数据中获取知识：使用包装器

- 难点：方便的包装器定义方法，包装器自动生成、更新与维护

□ 从文本中获取知识：信息抽取

- 难点：结果的准确率与覆盖率

3. 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

- 行业知识图谱的数据模式通常采用自顶向下和自底向上结合的方式，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。

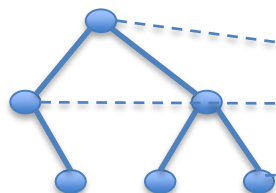
知识融合：跨语言融合



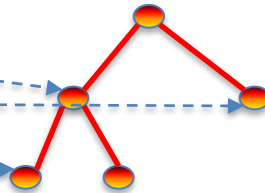
Unified Medical
Language System®

- Abdominal distension
- Abdominal mass
- Abdominal pain
- Abnormal dermatoglyphics
- Abnormal heart sounds
- Abnormal reflexes
- Abnormal splitting of heart sounds
- Abnormal sputum
- Absence of uvula
- Acneiform lesions
- Acute confusional state
- Adrenal cortex hyperplasia
- Akathisia
- Alopecia
- Amaurosis fugax

中文体系结构



英文体系结构



咳嗽 症状

[\[症状起因\]](#) [\[诊断详述\]](#) [\[检查鉴别\]](#) [\[问医生\]](#) [\[找医院\]](#)

咳嗽是呼吸系统疾病最常见的症状之一，它是一种保护性神经反射，通过咳嗽产生呼气性冲击动作，能将呼吸道内的异物或分泌排出体外。[\[详细\]](#)

相关疾病

夏季感冒

病毒性支气管炎

病毒性喉炎

肾咳

痰饮

阴茎短 症状

[\[症状起因\]](#) [\[诊断详述\]](#) [\[检查鉴别\]](#) [\[问医生\]](#) [\[找医院\]](#)

小阴茎(micropenis)是指阴茎外观正常长度与直径比值正常，但阴茎体的长度小于正常阴茎长度平均值2.5个标准差以上。阴茎的长度是指用手提阴茎头尽量拉直即相当于阴茎充分勃起时从阴茎顶到耻骨联合的距...[\[详细\]](#)

相关疾病

男性生殖器畸形

小阴茎

类固醇5α-还原酶2缺乏综合征

先天性睾丸发育不全

阴道出血 症状

[\[症状起因\]](#) [\[诊断详述\]](#) [\[检查鉴别\]](#) [\[问医生\]](#) [\[找医院\]](#)

阴道出血是女性生殖器官疾病常见的症状。出血可来自外阴、阴道、子宫颈和子宫内膜，但来自子宫者为最多。阴道出血量固然可以危及生命，但如良性疾病所致者，预后良好；而出血量少的，也可能是恶性肿瘤...[\[详细\]](#)

相关疾病

重度宫颈糜烂

胎热

血热崩漏

妊娠热病

原发性纤维蛋白溶解症

关节疼痛 症状 (别名：关节痛)

[\[症状起因\]](#) [\[诊断详述\]](#) [\[检查鉴别\]](#) [\[问医生\]](#) [\[找医院\]](#)

关节炎或关节病牵涉范围非常广泛，种类繁多，病因各异，然而，普遍的临床症状均可表现关节疼痛，因此关节疼痛的鉴别诊断至关重要，但是有的病因已知，有的病因尚未明了，总之，明确诊断才是合理治疗...[\[详细\]](#)

R30 Pain associated with micturition
Excludes: psychogenic pain (F45.3)
小便疼痛
不包含：心因性小便疼痛 (F45.3)

R30.0 Dysuria
Strangury
小便困难
尿绞痛

R30.1 Vesical tenesmus
膀胱里急後重

R30.9 Painful micturition, unspecified

ICD编码

DBpedia Mapping

Mapping en:Infobox book

Template Mapping (help)	
map to class	Book

Mappings

Property Mapping (help)	
template property	author
ontology property	author

Property Mapping (help)	
template property	illustrator
ontology property	illustrator

Class Book :	
Properties	
author	
coverArtist	
firstPublicationDate	
illustrator	
isbn	
lastPublicationDate	
...	

{{Infobox book		
author	=	
title_orig	=	
translator	=	
illustrator	=	
subject	=	
genre	=	
}}		

Mapping el:Βιβλίο

Template Mapping (help)	
map to class	Book

Mappings

Property Mapping (help)	
template property	συγγραφέας
ontology property	author

Property Mapping (help)	
template property	εικονογράφηση
ontology property	illustrator

{{Βιβλίο		
συγγραφέας	=	
είδος	=	
εκδότης	=	
πρώτη_έκδοση	=	
ISBN	=	
εικονογράφηση	=	
}}		

<http://mappings.dbpedia.org/>

识别 sameAs 关系



标签

- zhwiki:白癜风
- hudong:白癜风
- baidu:白癜风

Index

- zhishi:abstract
- infobox
- dcterms:subject
- zhishi:relatedImage

owl:sameAs

- ☒ zhwiki:白癜风 (this)
- ☒ baidu:白癜风 (this)
- ☒ hudong:白癜风 (this)
- ☐ dbpedia:Vitiligo

MERGE PAGE

zhishi:abstract

摘要

白癜风 (vitiligo) 一词来源于拉丁语 vitium, 意“缺损”, 或 vi-tellus, 指“小牛白色斑片”。公元一世纪, 罗马医生 Celsus 所写 De Medicina 中, 可见 vitiligo 一词。Vitiligo 在许多古籍中常常出现。印度文献中, kilas (kil 指白色, as 意抛弃) 和 palita (pal 含灰色、年老、老年之意) 可追溯到公元前 1500~1000 年, 指的是皮肤上的白色斑片。佛教圣书 Vinay Pitak (624~544 B.C.) 中记载, 患 kilas 的人不能做牧师。印度 Manusmriti (200 B.C.) 中记载, 患 svitra (白斑) 不受尊重。《古兰经》中 baras 一词意指白色皮肤, 常用来描述耶稣治愈后的状况。世界各地均有发生, 印度发病率最高, 中国约有 1200 万人发病, 本病可以累及所有种族, 男女发病无显著差别。近年白癜风发病率逐年上升, 引起人们的普遍关注。它是一种常见多发的色素性皮肤病。该病以局部或泛发性色素脱失形成白斑为特征, 是一种获得性局限性或泛发性皮肤色素脱失症, 是一影响美容的常见皮肤病, 易诊断, 治疗难。

信息框

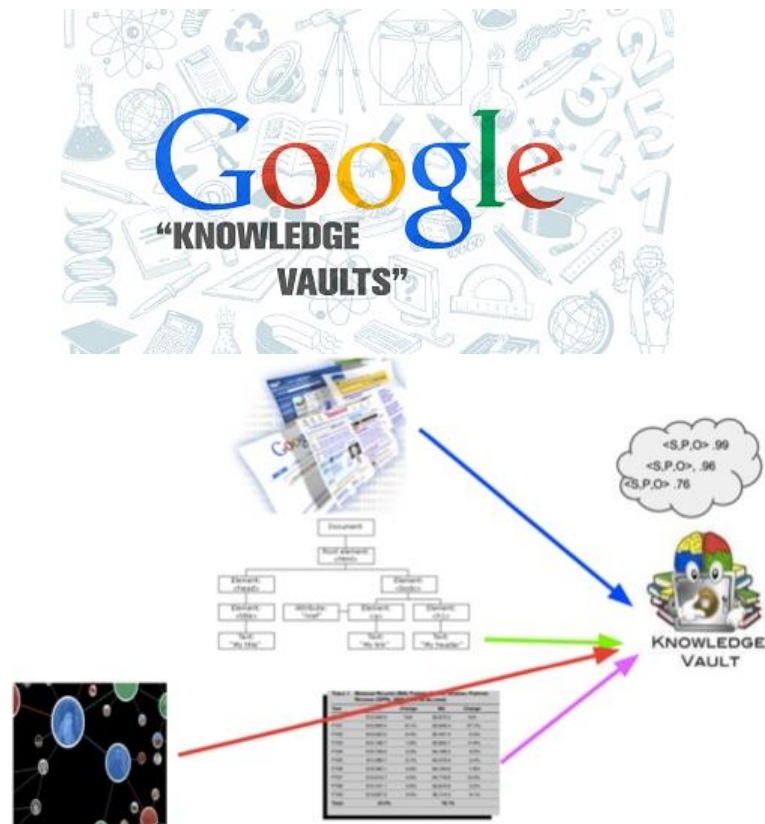
dcterms:subject

主题

- 皮肤和皮下组织疾病
- 应用科学
- 白斑
- 科学
- 疾病大全
- 疾病
- 医学名词
- 皮肤病
- 疾病
- 常见疾病
- 白癜风预防

Knowledge Vault: Google 将建全球最大知识库

- ❑ Knowledge Vault是一种以互联网信息为基础的数据库，即“知识库”。
- ❑ 知识来源：Gmail、Google+、Youtube
- ❑ 通过特定算法自动搜集整编互联网信息，再将其存入数据库中。
- ❑ Knowledge Vault的入库信息已达16亿条（至2014年），其中2.7亿条内容为“事实”（真实性在90%以上）。
- ❑ Knowledge Vault 能够建立历史和社会的模型。



知识融合关键技术难点

□ 实现不同来源、不同形态数据的融合

□ 海量数据的高效融合

□ 新增知识的实时融合

□ 多语言的融合



4. 知识存储

□ 知识图谱数据存储需要完成的基本数据存储：

- 三元组知识的存储
- 事件信息的存储
- 时态信息的存储
- 使用知识图谱组织的数据的存储

□ 知识图谱上层应用需要支持：

- 知识推理
- 知识快速查询
- 图实时计算

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

5. 知识计算

- ❑ 图挖掘计算：基于图论的相关算法，实现对图谱的探索和挖掘。
- ❑ 本体推理：使用本体推理进行新知识发现或冲突检测。
- ❑ 基于规则的推理：使用规则引擎，编写相应的业务规则，通过推理辅助业务决策。

关键技术与难点

□ 图挖掘计算

- 大规模图算法的效率

□ 本体推理与规则推理

- 大数据量下的快速推理
- 对于增量知识和规则的快速加载

6. 知识应用

- 语义搜索：基于知识图谱中的知识，解决传统搜索中遇到的关键字语义多样性及语义消歧的难题；通过实体链接实现知识与文档的混合检索。
- 智能问答：针对用户输入的自然语言进行理解，从知识图谱中或目标数据中给出用户问题的答案。
- 可视化决策支持：通过提供统一的图形接口，结合可视化、推理、检索等，为用户提供信息获取的入口。

关键技术与难点

□ 语义检索

- 自然语言的表达多样性问题
- 自然语言的歧义问题

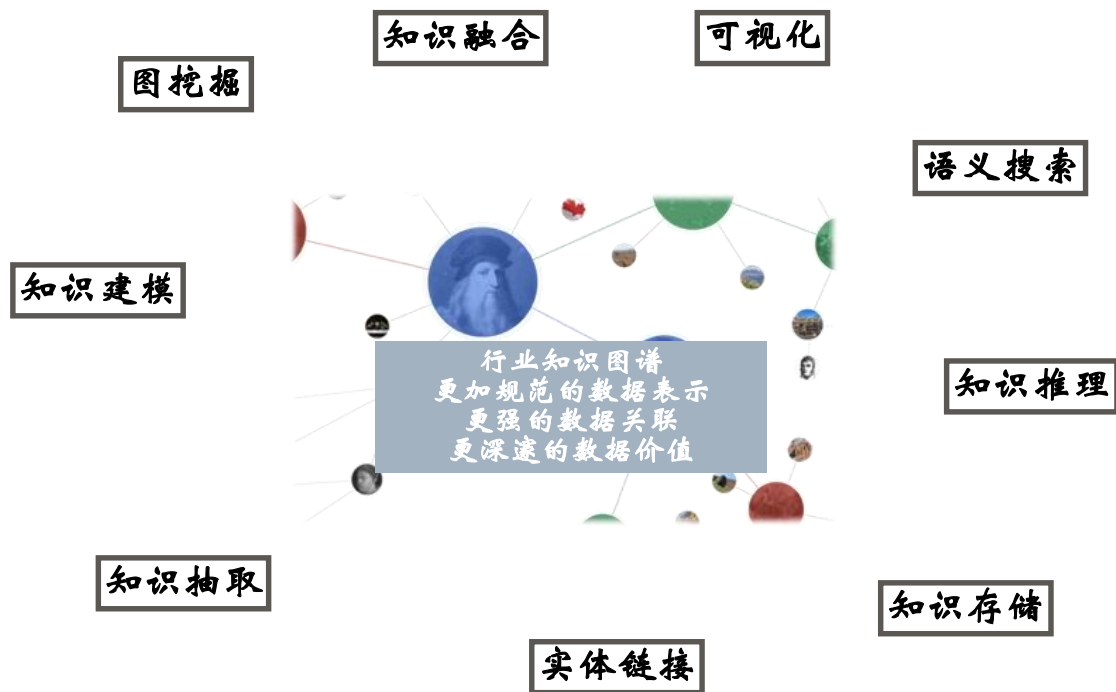
□ 智能问答

- 准确的语义解析
- 正确理解用户的真实意图
- 答案确定与排序

□ 可视化决策支持

- 通过可视化方式辅助用户模式快速发现
- 高效地缩放和导航
- 大图环境下底层算法(图挖掘算法)的效率

行业知识图谱关键技术



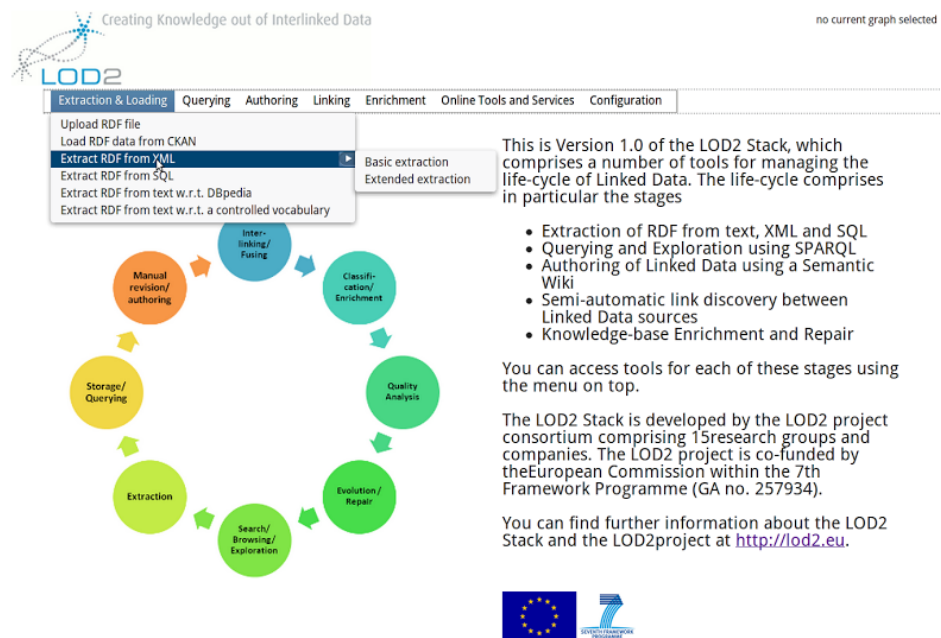
LOD2

□ LOD2 项目的主要目标是构建结构化链接数据的企业级管理工具和方法学，提供一个搜索、浏览和生成链接数据的平台。

□ LOD2 侧重于链接数据的生命周期管理，其它类型的数据需要首先转换成链接数据。

□ LOD2 没有对中文处理的支持

○

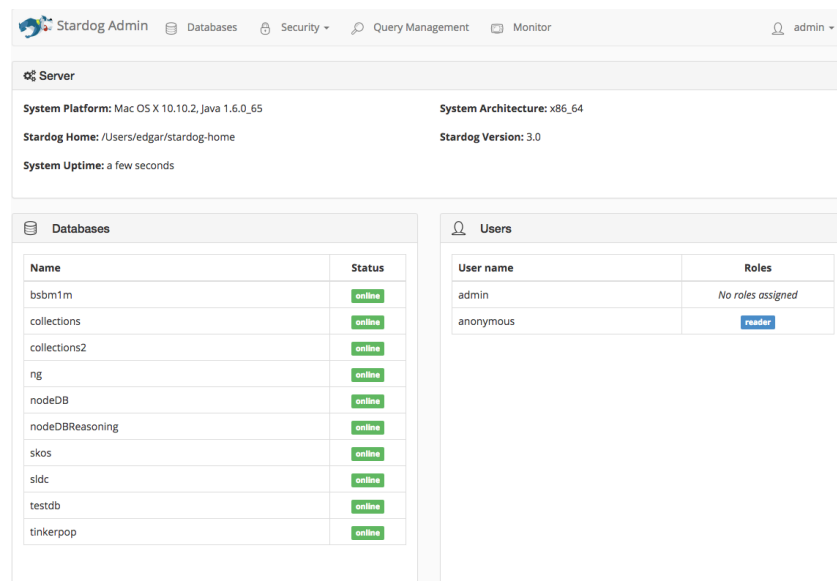


Stardog

❑ Stardog 是一个企业级知识图谱平台，通过把数据转换成知识，使用知识图谱进行组织，对外提供查询、检索、分析服务。主要特点：

- 把关系数据库映射成虚拟图
- 支持OWL2的推理
- 支持Gremlin

❑ 但 Stardog 仅包含对结构化数据 (RDBMS、Excel等) 的处理，没有针对非结构化数据的知识抽取，也没有包含知识融合功能。



The screenshot displays the Stardog Admin web interface. The top navigation bar includes links for Databases, Security, Query Management, and Monitor, along with a user profile icon for 'admin'. The main content area is divided into three sections:

- Server:** Displays system information including the platform (Mac OS X 10.10.2, Java 1.6.0_65), architecture (x86_64), home directory (/Users/edgar/stardog-home), version (3.0), and uptime (a few seconds).
- Databases:** A table listing various databases and their status.
- Users:** A table listing users and their assigned roles.

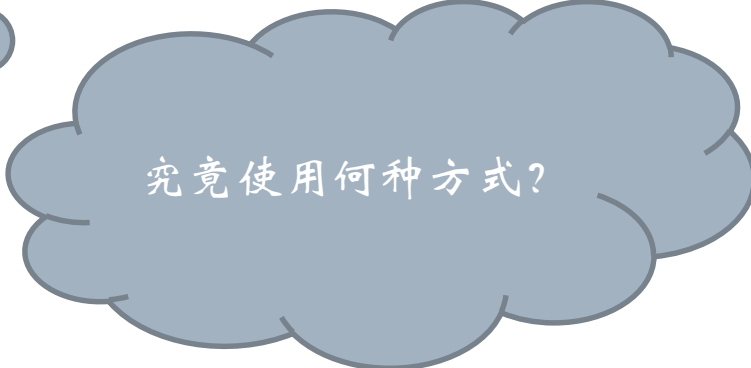
Name	Status
bsbm1m	online
collections	online
collections2	online
ng	online
nodeDB	online
nodeDBReasoning	online
skos	online
slidc	online
testdb	online
tinkerpop	online

User name	Roles
admin	No roles assigned
anonymous	reader

知识图谱生命周期总结

□ 在行业应用中使用知识图谱，大致有如下几种方式：

- 使用现有的套装工具 (如 LOD2、Stardog)
- 在现有套装工具的基础上进行扩充：
- 使用各生命周期过程的相应工具并进行组合使用
- 针对性开发或扩展生命周期中特定工具
- 完全从零开始构建



究竟使用何种方式？



Part 1 知识建模

Part 2 知识获取

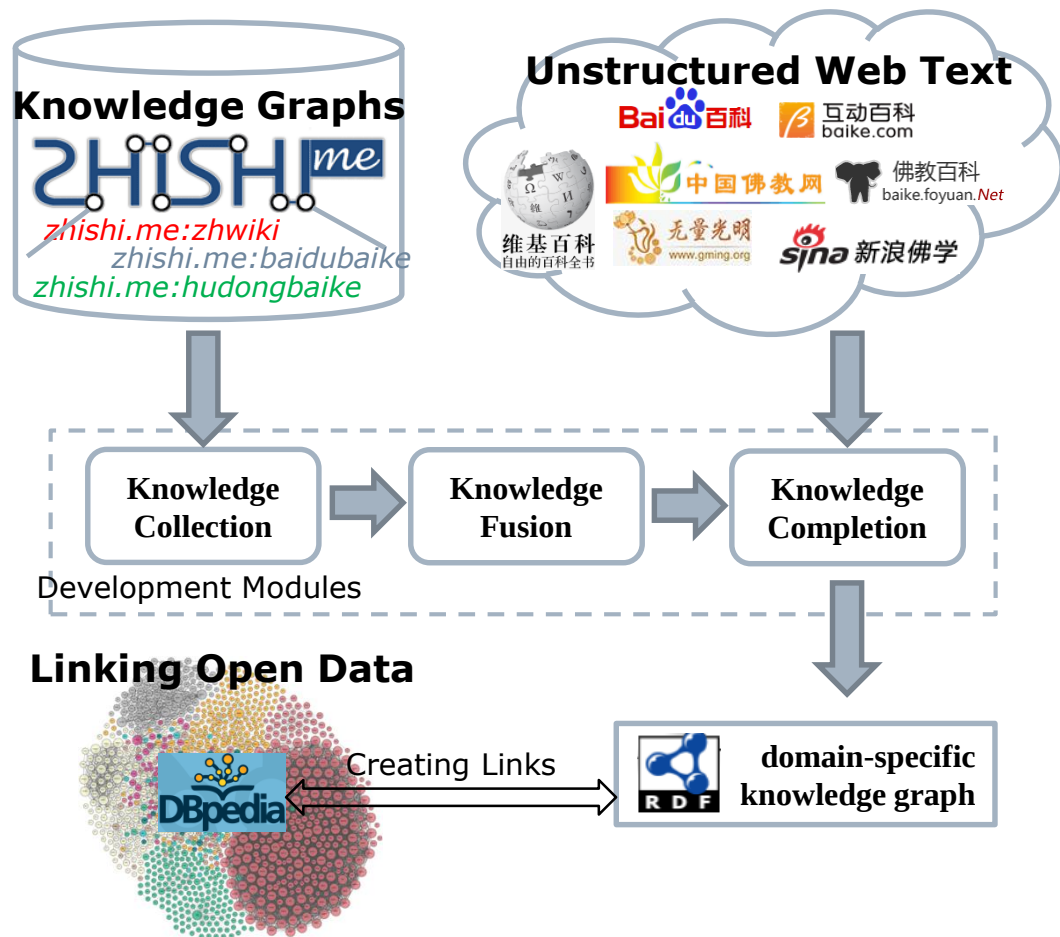
Part 3 知识融合

Part 4 知识存储

Part 5 知识计算

Part 6 知识应用

领域知识图谱构建框架



知识收集（佛教）

- Category方法

- 人工观察百科中与佛教人物相关的分类 (于佛教的子类中观察)
- 抽取佛教人物分类下所有文章对应的实体

- 命名规则方法

例： “.+菩萨” “.+禅师”

- 维基百科“佛教头衔”分类下的所有实体
- 已抽取出的实体名中高频的公共字符串

佛

- 佛教之最 (1个页面)
- 佛教修行法門 (29个页面)
- 佛教历史 (8个分类, 90个页面)
- 佛教咒語 (15个页面)
- 佛教哲學 (3个分类, 6个页面)
- 佛教圣者 (3个分类)
- 佛教宗派 (10个分类, 11个页面)
- 佛教建筑 (1个分类, 1个页面)
- 佛教戒律 (2个分类, 23个页面)
- 佛教政治 (2个分类, 1个页面)
- 佛教文化 (12个分类, 15个页面)
- 佛教文学 (8个页面)

- 佛教文献 (6个分类, 2个页面)
- 佛教相關人物 (5个分类, 1个页面)
- 佛教研究 (2个分类, 6个页面)
- 佛教神祇 (8个分类, 37个页面)
- 佛教神話 (14个页面)
- 佛教組織 (5个分类, 8个页面)
- 佛教與其他宗教 (2个页面)
- 佛教與女性 (1个分类, 1个页面)
- 佛教节日 (3个分类, 5个页面)
- 佛教術語 (7个分类, 449个页面)
- 佛教題材作品 (5个分类, 12个页面)

反

- 反佛教 (1个分类, 13个页面)

知识融合（佛教人物）

- 主语融合（佛学人物图谱中的应用）
 - 实体的“别名”属性和重定向作为实体的别名集合
 - 不同来源的实体存在一个完全匹配的别名则认为相同实体
 - 人工检查相同实体数多于三个的映射

例：

Entities: 百度百科: <http://zhishi.me/hudongbaike/resource/额尔德尼·确吉坚赞>
互动百科: <http://zhishi.me/baidubaike/resource/班禅额尔德尼·确吉坚赞>
维基百科: <http://zhishi.me/zhwiki/resource/十世班禅>



Alias Sets: {十世班禅, 班禅额尔德尼·确吉坚赞, 第十世班禅额尔德尼}
{确吉坚赞, 班禅额尔德尼·确吉坚赞, 罗桑赤烈伦珠}
{班禅额尔德尼·确吉坚赞, 额尔德尼·确吉坚赞, 罗桑赤烈伦珠确吉坚赞}



Result: <http://www.kg-Buddhism.com/entity/额尔德尼·确吉坚赞>

知识融合（佛教寺庙）

□ 主语融合（寺庙知识图谱中的应用）

问题

— 同名不同实体

<baidubaike/龙泉寺> : {龙泉寺、北京龙泉寺}

<zhwiki/龙泉寺> : {龙泉寺、南京龙泉寺}

— 同实体不同名

<baidubaike/龙泉寺> : {龙泉寺、北京龙泉寺}

<hudongbaike/龙泉禅寺> : {龙泉禅寺、北京凤凰岭龙泉寺}

<zhwiki/龙泉寺_(海淀区)> : {龙泉寺_(海淀区)、海淀区龙泉寺}

知识融合（佛教寺庙）

□ 主语融合（寺庙知识图谱中的应用）

方案

- 多个相同别名判断实体是否相同
- 实体的“地址”、“建造时间”属性是否冲突

例：

Entities: 百度百科: <http://zhishi.me/baidubaike/resource/三祖寺>
互动百科: <http://zhishi.me/hudongbaike/resource/乾元禅寺>
维基百科: <http://zhishi.me/zhwiki/resource/三祖寺>



Alias Sets: {三祖寺, 山谷寺}
{三祖寺, 山谷寺, 乾元禅寺}
{三祖寺, 三祖山谷乾元禅寺, 山谷寺, 乾元禅寺}

Address: 安徽省潜山县城西北9公里处的谷口凤形山上
安徽省潜山县天柱山风景区
安徽省安庆市天柱山南

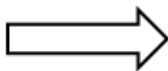


Result: <http://www.kg-Buddhism.com/entity/三祖寺>

知识融合（佛教寺庙）



精确性原则



□ 宾语融合

■ 单值属性

精确性原则：日期、地点等类型的属性值出现冲突时选择最精确的一个。

大多数原则：不同来源的属性值出现冲突时，选择出现次数最多的值。

■ 多值属性

直接合并去重

知识融合（佛教寺庙）

```
triple1:  
<http://www.agriculture.kg/entity/中国农业大学>  
<http://www.agriculture.kg/property/主要院系>  
"农学与生物技术学院、生物学院、动物医学院、资源与环境学院"@zh
```

```
triple2:  
<http://www.agriculture.kg/entity/中国农业大学>  
<http://www.agriculture.kg/property/主要院系>  
"农学与生物技术学院、工学院、理学院、信息与电气工程学院"@zh
```

```
triple3:  
<http://www.agriculture.kg/entity/中国农业大学>  
<http://www.agriculture.kg/property/主要院系>  
"农学与生物技术学院、生物学院、动物医学院"@zh
```

↓
多值属性合并去重

```
<http://www.agriculture.kg/entity/中国农业大学>  
<http://www.agriculture.kg/property/主要院系>  
"农学与生物技术学院、生物学院、动物医学院、资源与环境学院、工学院、理学院、信息与电气工程学院"@zh
```

□ 宾语融合

■ 单值属性

精确性原则：日期、地点等类型的属性值出现冲突时选择最精确的一个。

大多数原则：不同来源的属性值出现冲突时，选择出现次数最多的值。

■ 多值属性

直接合并去重

知识补全（佛教人物）

□ 对infobox属性进行补全

- 人工编写规则从非结构化文本中抽取属性值
- 依照知识融合方法将属性值对转换为三元组

例：

Text:

title：李叔同

李叔同，谱名文涛，幼名成蹊，学名广侯，字息霜，别号漱筒；出家后法名演音，号弘一，晚号晚晴老人。

Pattern:

.*(法名|法号)(为|曰|称|叫|即){0,1}([\S]+?)(字|祖籍|，|。){0,1}.*

Property-value pair

法名：（李叔同，演音）

Triple:

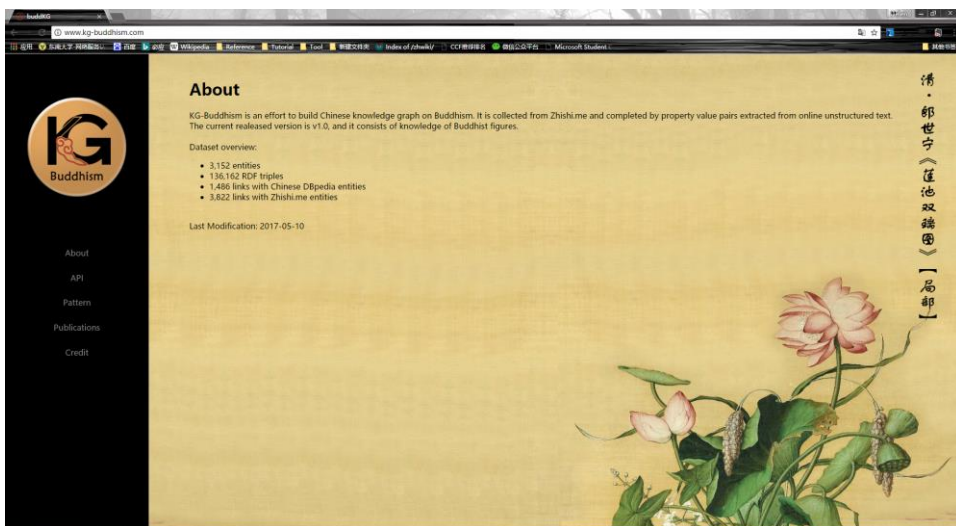
<http://www.kg-buddhism.com/entity/李叔同>
<http://www.kg-buddhism.com/property/法名>
"演音"@zh

知识问答（佛教人物）

□ 在线接口

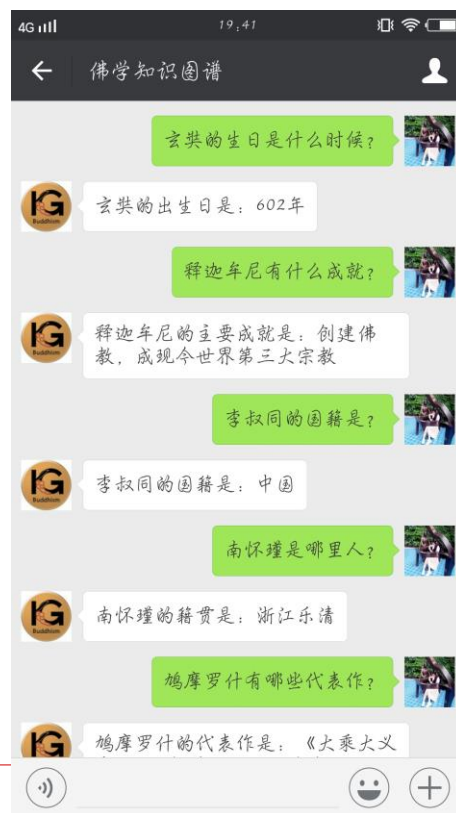
■ 数据访问地址：

www.kg-Buddhism.com



• 应用

— 问答微信公众号



佛学考试机器人

考试机器人：Todai Robot Project (2014, Japan)
the Project Aristo (2015, America)

863课题：高考地理 考试机器人

佛教入学考试：一个更具挑战性的任务。

1. 僧伽是①涅槃义②和合众③杀贼义。
2. 「诸行无常、诸法无我、涅槃寂静」称为①三种无常②三法印③三乘道。
3. 人生最大的错误是①杀生②妄语③邪见。



佛学考试机器人

典型问题（内容/答案类型划分）

1. 人物类 佛陀在世时，印度最忠诚拥护佛法的国王是()?
2. 地名类 中国佛教四大名山中，观音菩萨的道场是()?
3. 佛经名类 善财童子五十三参出自()?
4. 术语类 十善业又分身、口、意三类，意善业有三，是指()?

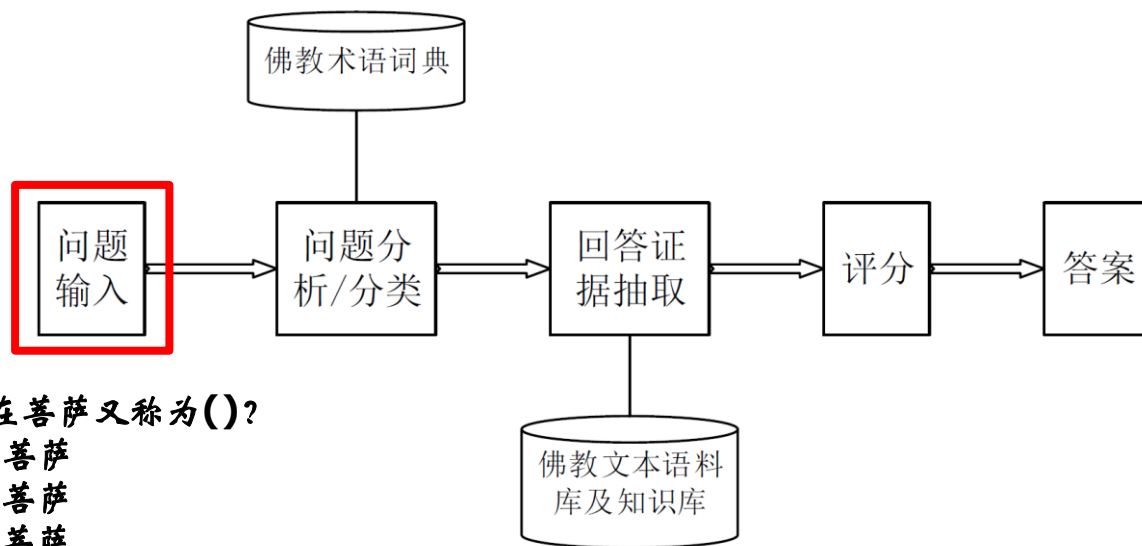
佛学考试机器人

典型问题（提问方式划分）

1. 直答类 佛陀在()为五比丘说法?
2. 解释类 佛陀十号之一的两足尊，其含义是指()?
3. 列举类 一般佛教徒经常诵读的经典是()?

以及，选择题中特有的问题句式，如：以下何部经典不是谈般若空性？

框架



《心经》中的观自在菩萨又称为()?

- A 弥勒菩萨
- B 地藏菩萨
- C 观音菩萨
- D 文殊师利菩萨

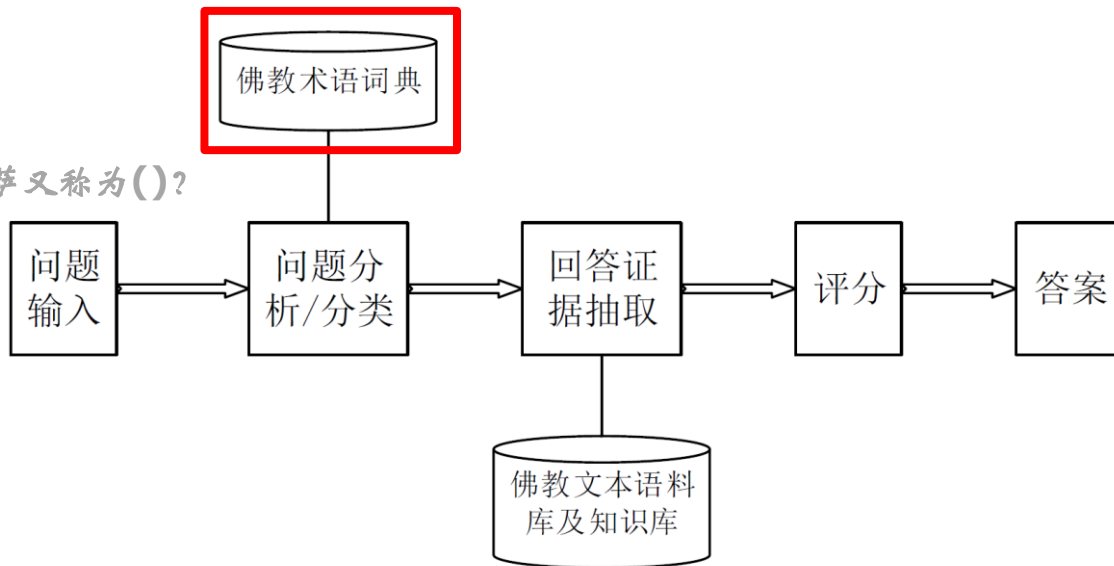
框架

佛学人物词典/书籍词典等

佛教术语词典

《心经》中观自在菩萨又称为()?

- A 弥勒菩萨
- B 地藏菩萨
- C 观音菩萨
- D 文殊师利菩萨

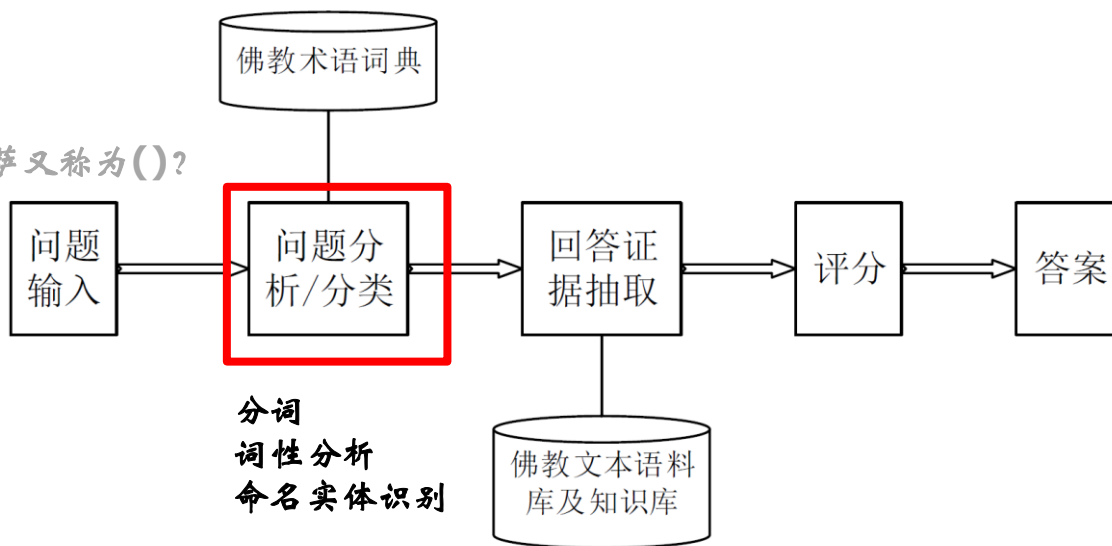


框架

佛学人物词典/书籍词典

《心经》中观自在菩萨又称为()?

- A 弥勒菩萨
- B 地藏菩萨
- C 观音菩萨
- D 文殊师利菩萨

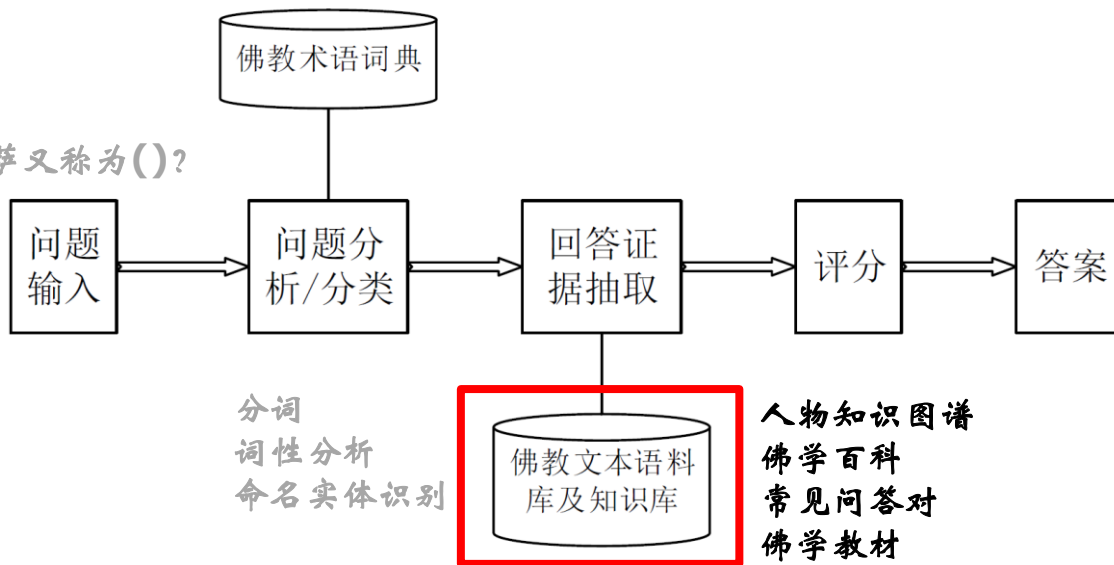


框架

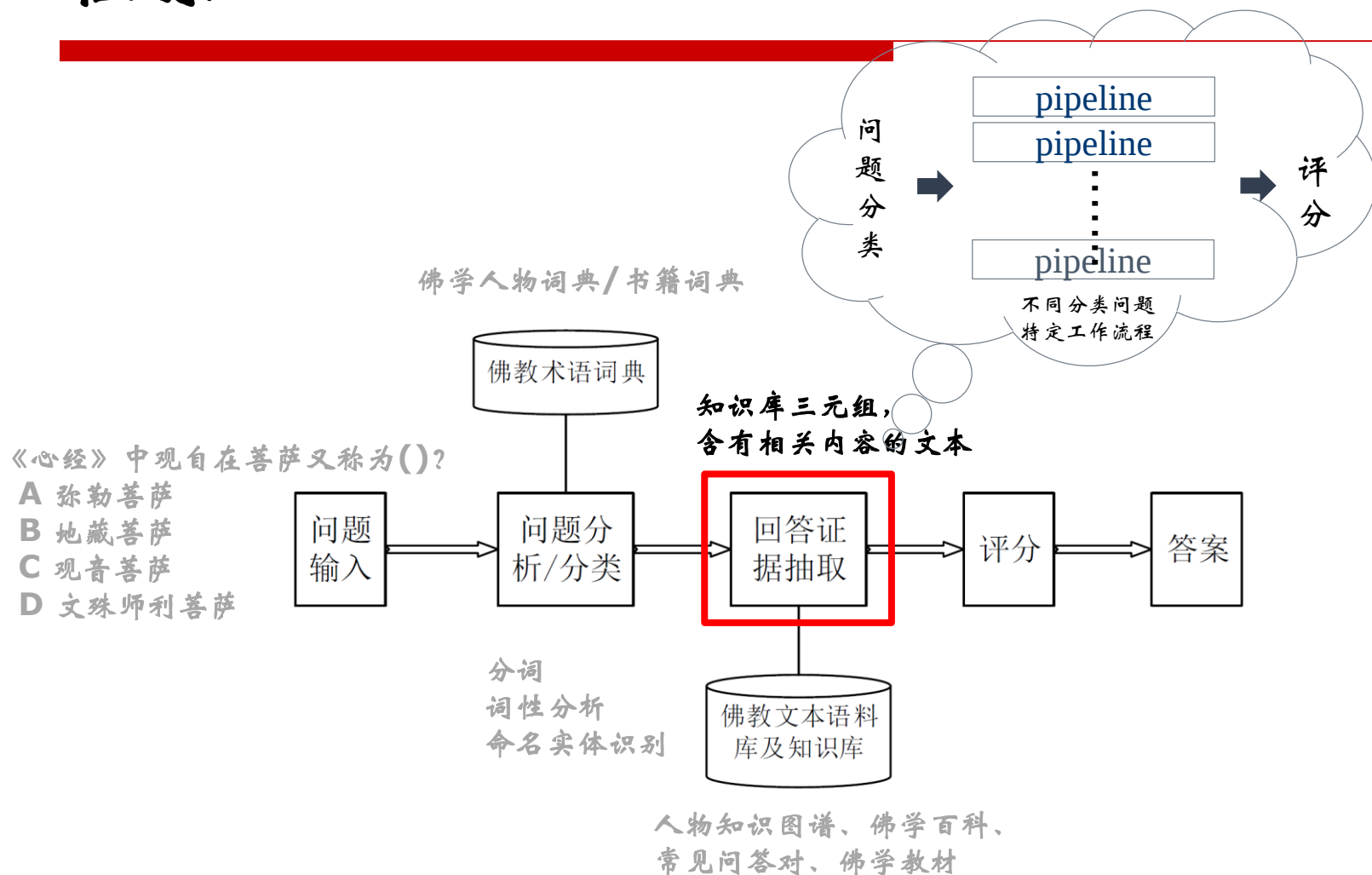
佛学人物词典/书籍词典

《心经》中观自在菩萨又称为()?

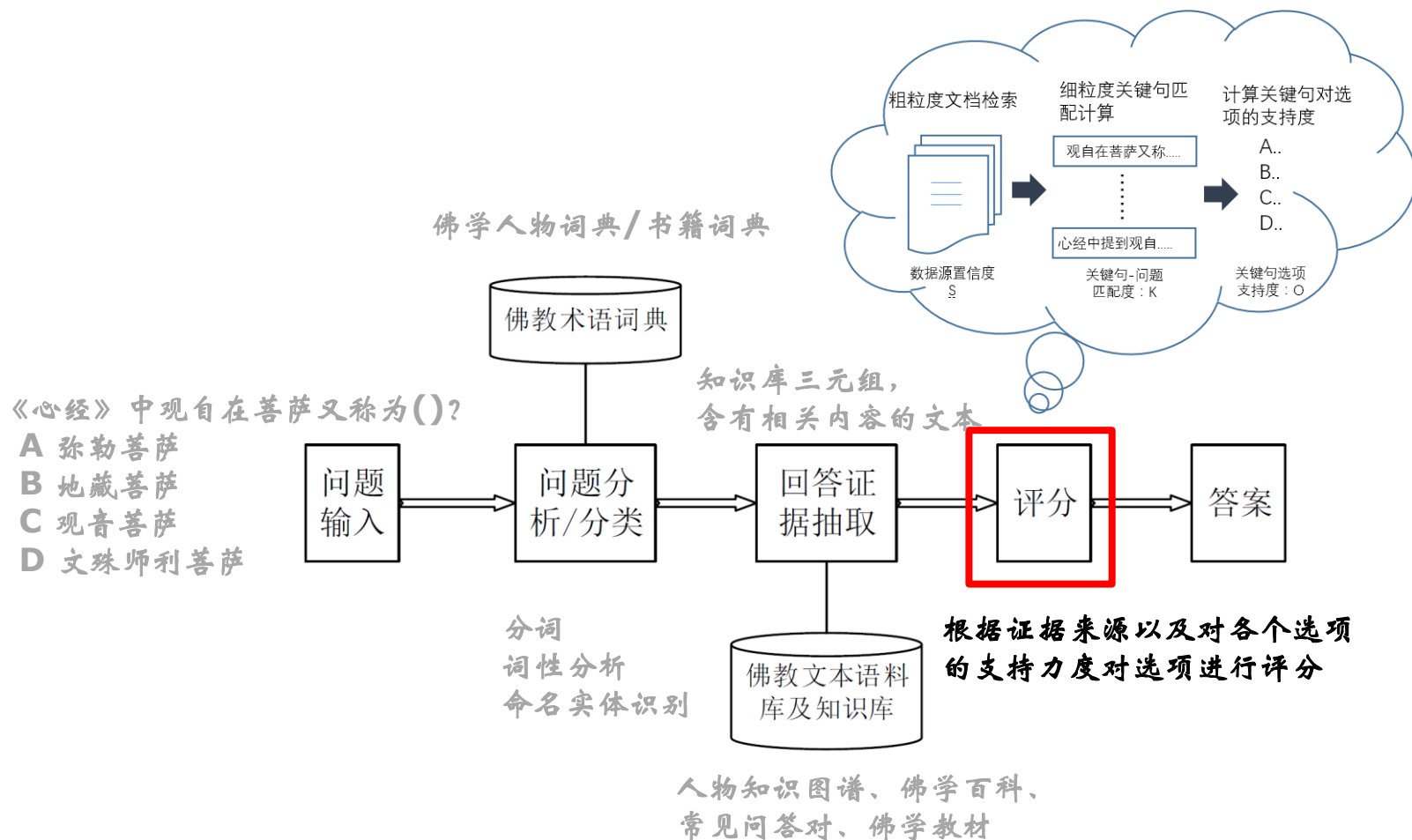
- A 弥勒菩萨
- B 地藏菩萨
- C 观音菩萨
- D 文殊师利菩萨



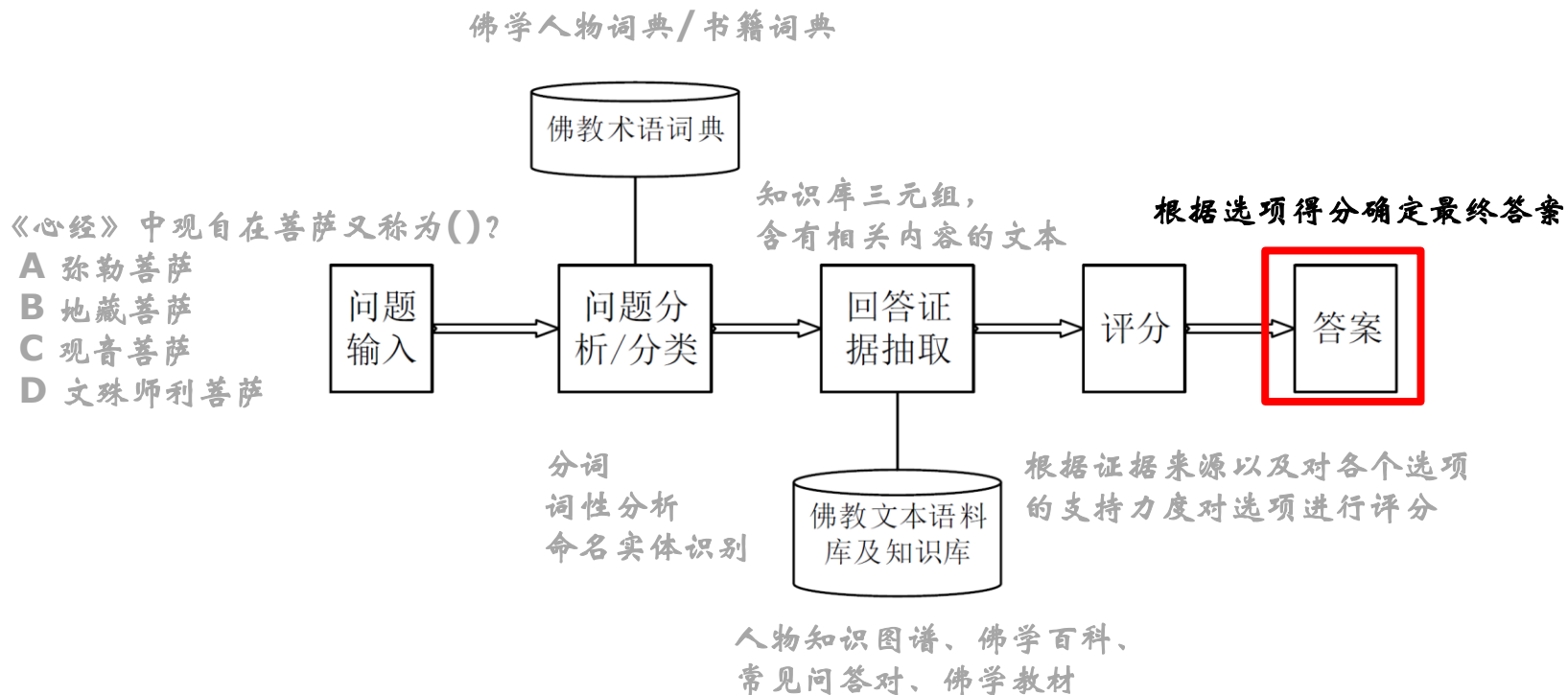
框架



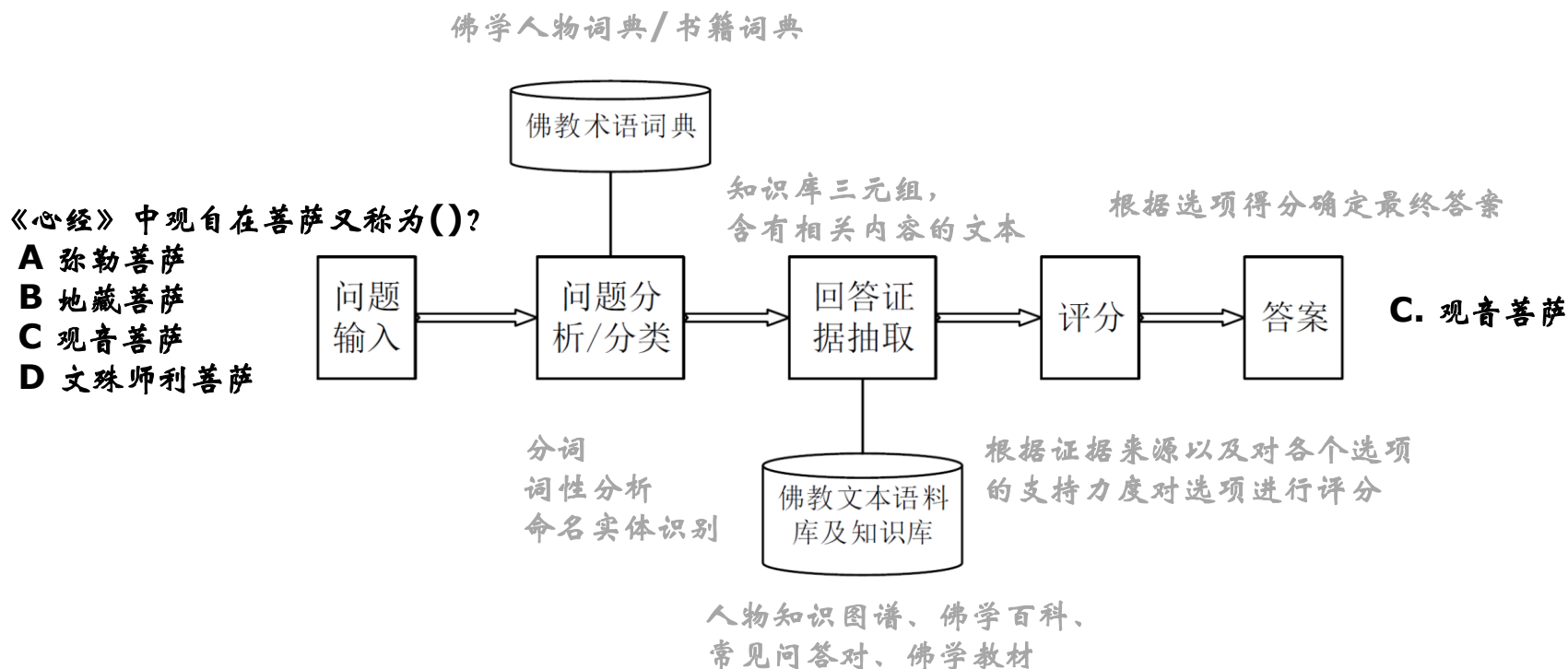
框架



框架



框架



Demo

佛学考试机器人

SEU-KSE

正确答案: 124

问题总数: 175

正确率: 71%

1: 佛陀在()为五比丘说法

正确答案:鹿野苑

A 七叶窟

答案分数:0.30578512396694213

B 王舍城

答案分数:0.2892561983471074

C 鹿野苑

答案分数:0.4049586776859504

输出答案:鹿野苑

2: 佛陀告诉我们人生有()

正确答案:四苦 八苦

A 四苦

答案分数:0.23306772908366533

B 八苦

答案分数:0.23306772908366533

C 四苦 八苦

答案分数:0.5338645418326693

输出答案:四苦 八苦

模式知识构建

■ 模式知识构建：Taxonomy(分类体系) 构建

- 抓取互动百科和中文维基百科分类系统中关于农业的分类分支
(百度百科中没有这样的分类体系，因此不选用百度百科)



互动百科分类体系农业分支



中文维基百科分类体系农业分支

模式知识构建

■ 模式知识构建：Taxonomy(分类体系) 构建

- 抓取互动百科和中文维基百科分类系统中关于农业的分类分支
- 邀请多位志愿者对抓取下来的分类体系中概念之间的subclass关系进行正确与否的标注。如果志愿者中的多数人认为某一个subclass关系错误，就将该关系剔除。
- 对筛选过后的两个百科的分类体系进行合并，得到最终的农业图谱的分类体系。

模式知识构建

■ 模式知识构建：类别属性抽取

- 统计某一个类别中所有实体的属性，若大部分实体都有该属性，则认为该类别有这一属性

— 例：农作物的属性有场地和修剪时间等属性

- 通过实验确定一个最优的阈值，某一类别下的所有实体中，拥有某一相同属性的实体数量或百分比超过该阈值时，该类别概念就具有该属性。

知识融合

□ 主语融合（农业图谱中的应用）

- 实体的“别名”属性和重定向作为实体的别名集合
- 不同来源的实体存在一个完全匹配的别名则认为相同实体
- 人工检查相同实体数多于三个的映射

例：

Entities: 百度百科: <http://zhishi.me/baidubaike/resource/鸡蛋果>
互动百科: <http://zhishi.me/hudongbaike/resource/百香果>
维基百科: <http://zhishi.me/zhwiki/resource/西番莲>



{鸡蛋果、洋石榴、紫果西番莲、百香果、藤石榴}

Alias Sets: {百香果、西番莲果、热情果、西番果}

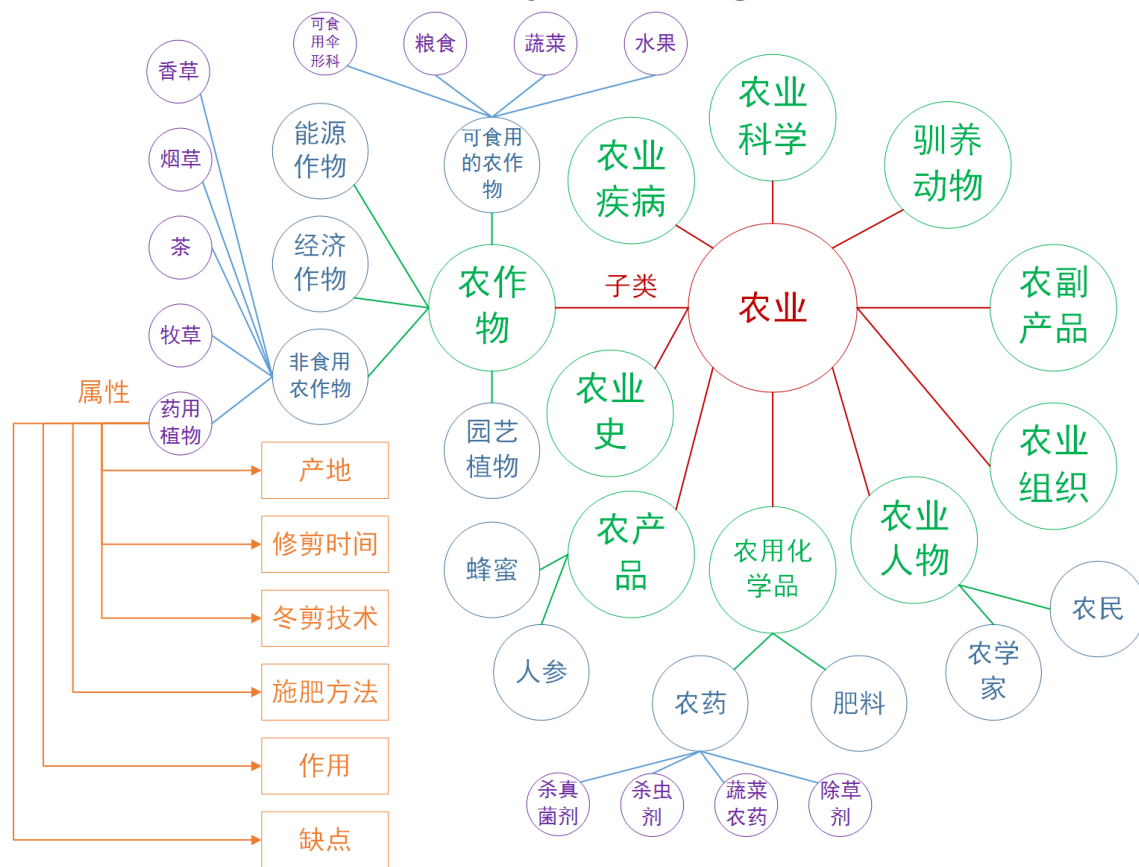
{热情果、百香果、西番果、巴西果}



Result: <http://www.agriculture.kg/entity/百香果>

农业知识图谱示例

- 使用知识图谱构建知识库，用于事实查询，知识图谱来源为结合已有百科知识图谱的相关内容、百科知识、文档等，经过人工审核，建立的图谱。下图为农业知识图谱的示例（Toy knowledge base）。



系统背景

■ 系统开发目标：

- 在互联网中，海量信息分布于不同的信息源，相关性稀疏
- 使用传统搜索引擎，准确、快速地获得有价值的信息变得愈发困难
- 相较而言，问答系统的出现旨在快速得到高质量信息或答案
- 同一问题，在传统搜索引擎与问答系统分别检索，结果对比如下图所示，可见通过问答系统得到的信息更加明确、简练

有哪些预防猪瘟的措施？

[猪瘟的主要临床症状及防治措施有哪些?|猪传染病【搜猪网_中国生猪...](#)

2016年4月27日 - 猪瘟是由猪瘟病毒引起的一种急性、热性、高度接触性传染病。(1)症状目前在临床上,典型的猪瘟病较难看到,但温和型猪瘟(非典型猪瘟)较常见,该病型由...

[www.soozhu.com/article...](#) - 百度快照

[猪瘟的预防方法有哪些? - 爱问知识人](#)

2017年8月10日 - 猪瘟的预防方法有哪些?:猪瘟的预防1.平时的预防措施着重提高猪群的免疫水平,防止引入病猪,切断传播途径,广泛持久地开展猪瘟疫苗的预防接种,是预防猪瘟...

[iask.sina.com.cn/b/6i1...](#) - 百度快照

[猪瘟防治措施_百度文库](#)

2016年12月11日 - 猪瘟防治措施_畜牧兽医_农林牧渔_专业资料。猪瘟防治措施 防治措施 猪瘟对养猪业的危害很大,目前还没有有效的治疗方法。临床出现猪瘟症状后可使用一些...

农业问答

有哪些预防猪瘟的措施？

提问

坚持自繁自养，必须引进种猪时严格做好检疫
发病时，严格隔离 疫苗防治

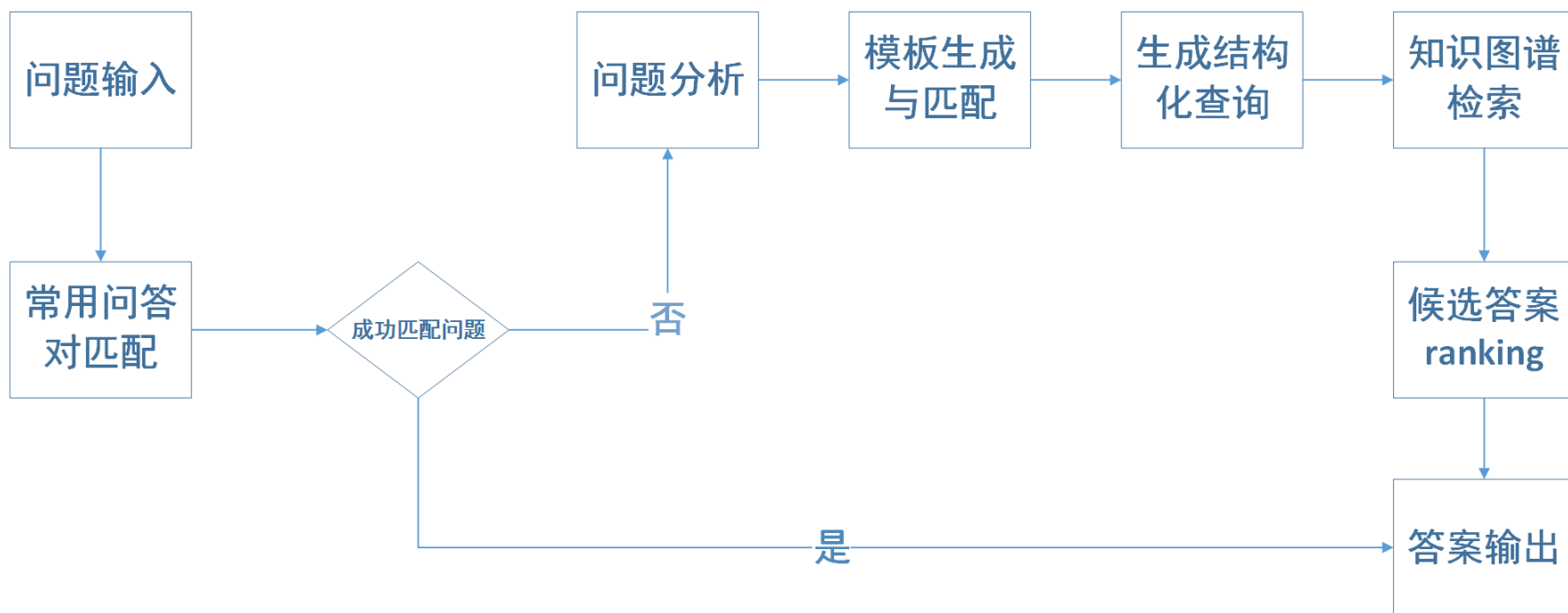
系统展示

系统开放web service API服务，允许用户进行网络访问，远程访问运行成果如下图示：



系统框架

- 使用pipeline+模板的主要方法，进行前期框架设计



问题模板收集——问题来源

- 数据类型均为文本类型：
 - 百科知识：农业百科、百度百科

韭菜的种植时间和种植条件

韭菜的叶、花葶和花均作蔬菜食用，种子等可入药，具有补肾，健胃，提神，止汗固涩等功效，韭菜适应性强，抗寒耐热，全国各地到处都有栽培，下面我们就一起来看一看韭菜种植时...韭菜的种植时间和种植条件

发布时间:2017-07-29|文章主题:茎叶蔬菜|文章来源: www.nongye.com

- 社区问答：知乎、大众养生等

是黄豆有营养？还是豆芽更有营养？

 地主易
都市养花种菜，产品价值加工，绿色低碳生活

1 人赞同了该回答

豆芽除了含有豆类的营养外，幼苗状态的根茎还含有纤维素，因此更有营养。

比如西兰花种子长出来的嫩芽里含有的对预防癌症与防止老化颇具效果的胡萝卜素、维生素C、B族维生素的含量也比成熟西兰花的多。

- 垂直网站：黔农网等

韭菜的种植时间

发布时间：2015-06-20 21:50:01 来源：黔农网 标签：韭菜的种植时间

内容提要：韭菜的叶、花葶和花均作蔬菜食用，种子等可入药，具有补肾，健胃，提神，止汗固涩等功效，韭菜适应性强，抗寒耐热，全国各地到处都有栽培，下面我们就一起来看一看韭菜种植时间和种植条件！

问题模板收集——问题来源

■ 数据类型均为文本类型：

- 专业问答网站：WikiHow、百度知道、爱问知识人、百度经验

如何种植豆类和豌豆

3 方法： ■ 种植豆类的准备工作 ■ 种植豌豆的准备工作 ■ 种植豆类和豌豆

豆类和豌豆这类植物相对容易生长，如果你是一个刚入门的园丁或刚刚获得一片新的园地，种植它们是很好的选择。这些豆类还与固氮细菌有着共生关系，这意味着他们可以改善土壤中的营养成分。遵循以下这些指示种植豆子——然后将它们从藤上摘下来吃掉，你就会发现它们有多美味！

- 专业论文网站：中国知网

☐ 1	食物寒热属性研究进展 优先出版	张娣; 安玉香	亚太传统医药	2017-04-05 15:01	期刊	23	↓	HTML	——
☐ 2	口腔上火别急着吃西瓜	健康时报记者 杨绪军	健康时报	2009-07-13	报纸	14	↓	📖	——
☐ 3	“靳三针”疗法合饮食澄清计划治疗Crohn's病的临床研究	杰弗里	广州中医药大学	2002-04-01	博士	10	↓	📖	——

- 书籍、期刊、文献、APP（如搜狗微信）等：

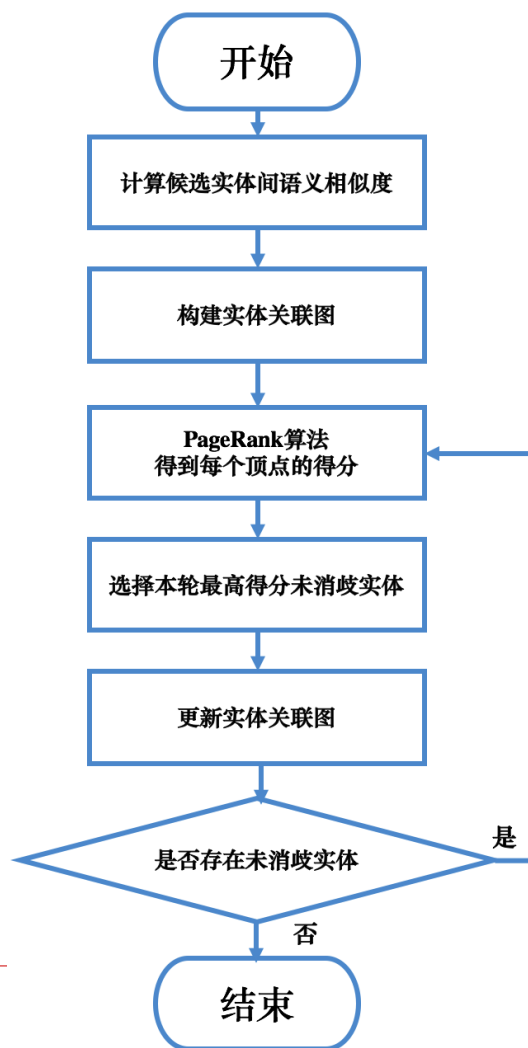
- APP：农业通，农管家，农医生
- 期刊：《广东农业科学》、《贵州农业科学》、《甘肃农业科技》、《现代农业科技》、《中国畜牧兽医》、《福建农业科技》

实体识别

- 从知识图谱中统计得到所有实体集合
 - 将自然语言问句与实体集（辅以知识图谱中的等价实体）进行字符串匹配，得到mention字符串集合
 - 若一个子串包含于另一个子串，则将最长子串视为一个识别到的实体mention
 - 若一个问题中包含多个不同的最长子串，则视为识别到多个实体mention
 - 对每一个mention进行实体链接，并将句子中剩余的部分视为待映射为谓词的句子模板
 - 对于每一个mention-句子模板配对，进行后续分析
 - 例如：阿里山高山茶分布在哪里？
 - 山茶
 - 高山茶
 - 阿里山高山茶
- > 实体：阿里山高山茶，问句模板：_____分布在哪里？

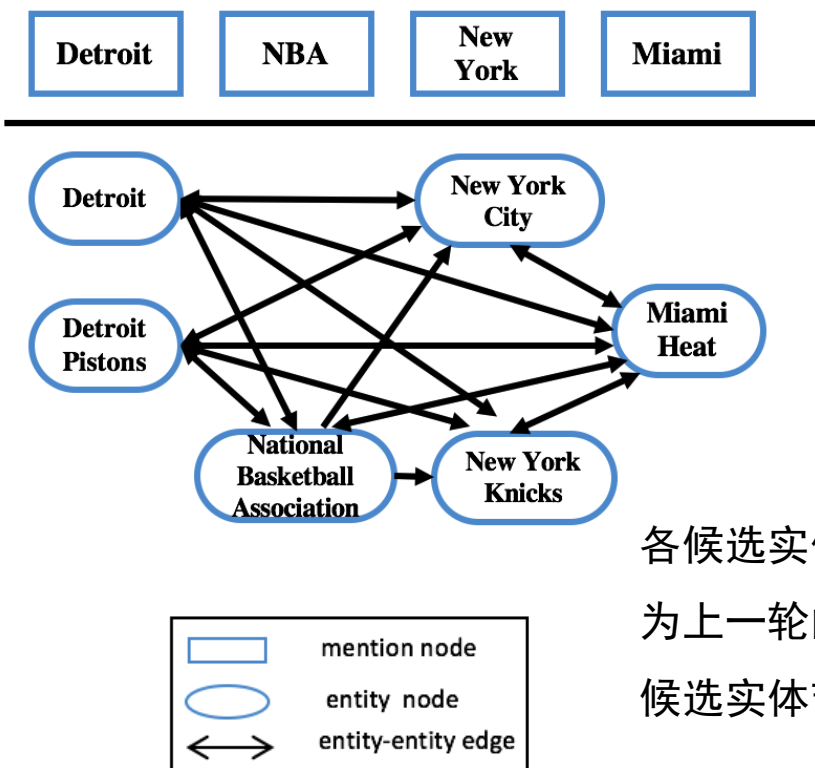
实体链接

■ 实体消歧算法流程图



实体链接

■ 构建实体关联图



实体关联图由四个部分组成：

- (1) 实体指称节点
- (2) 候选实体节点
- (3) 候选实体节点顶点值：代表该候选实体是实体指称的目标实体概率大小
- (4) 候选实体节点边权值：代表两个候选实体间的转化概率大小

各候选实体节点顶点值：初始化为均等，之后每轮更新为上一轮的PageRank得分

候选实体节点边权值计算公式如下：

$$SM(e_a^i, e_b^j) = \cos(v(e_a^i), v(e_b^j)) \longrightarrow$$

两个候选实体间的相似度大小

$$ETP(e_a^i, e_b^j) = \frac{SM(e_a^i, e_b^j)}{\sum_{k \in (V \setminus V_i)} SM(e_a^i, k)} \longrightarrow$$

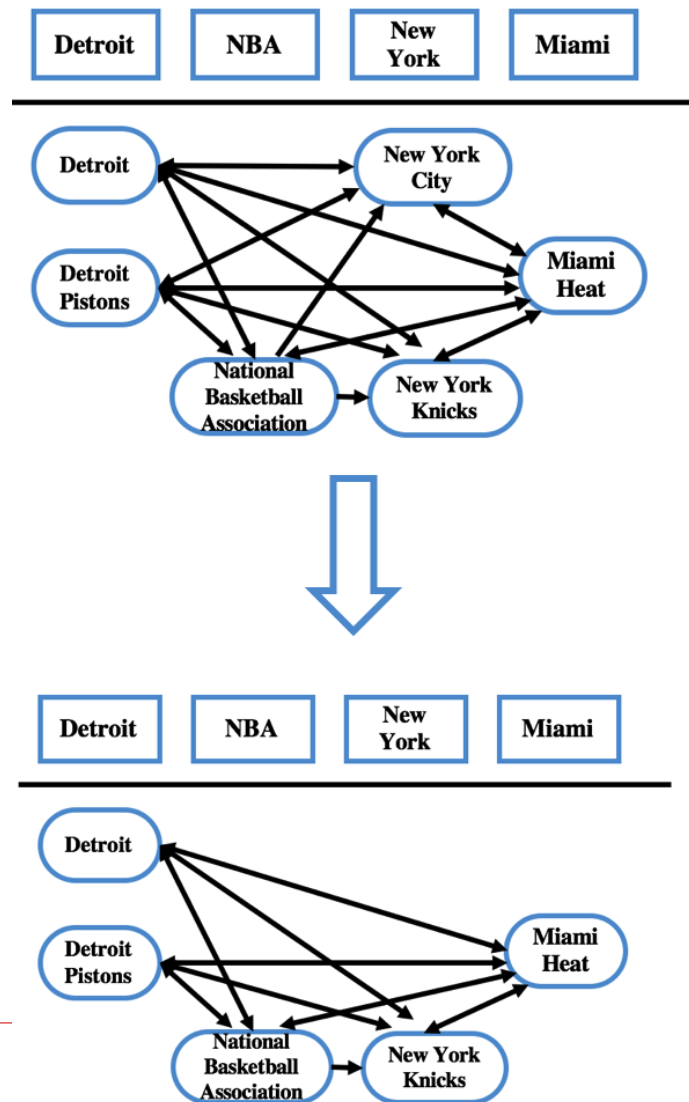
两个候选实体间的转化概率

实体链接

■ 更新实体关联图

节点代表实体	节点 PageRank 得分
Detroit	0.44314869
Detroit Pistons	0.77259475
Nation Basketball Association	0.85422741
New York City	0.36443149
New York Knicks	0.78134111
Miami Heat	0.78425656

选择本轮最高得分的未消歧实体 New York Knicks作为实体指称New York的最佳实体，删除其他候选实体 New York City及相关的边，更新图中的边权值。



问题分析

■ 五种类别的问题：

— 有什么

- 棉花的作用有什么？黄瓜有什么营养？西瓜有效的杀虫剂有哪些？

— 怎么做

- 怎样种好棉花？大白菜怎么种植？如何用南瓜嫁接黄瓜？

— 是什么

- 韭菜是什么样子的？花生是什么科目的？玉米是温性还是凉性的食物？

— 什么时候

- 胡萝卜什么时候种植？花生最佳收获时间是什么时候？生豆芽要几天？

— 以上四种问题回答方法

- 答案为知识图谱实体与文本信息 -> 事实类问答+文档检索 -> 知识图谱+IR
- 前期以知识图谱为知识来源，后期辅以文档检索为补充

— 概念判断

- 黄瓜和豆腐能一起吃吗？豌豆和毛豆有什么区别？土豆吃多了会上火吗？

— 以上这种问题回答方法

- 答案为正确与否的判断-> 基于文档检索抽取证据做判断 -> IR

问题分析

- 问题分析：问法类似（有什么……/怎么做……/是什么……/什么时候……） -> 适合采用模板（template）的方法

— 大白菜怎么种植？

-> <\$蔬菜> 怎么种植 -> <\$蔬菜> <\$p> <?x>

（“怎么种植”映射为谓词<\$p>）

```
SELECT ?x WHERE{  
  ?w type vegetable .  
  ?w p ?x.}
```

SPARQL

— 胡萝卜什么时候种植？

-> <\$蔬菜>什么时候种植 -> <\$蔬菜> <\$p> <?x>

（“什么时候种植”映射为谓词<\$p>）

```
SELECT ?x WHERE{  
  ?w type vegetable .  
  ?w p ?x.  
  ?x type date}
```

SPARQL

候选答案打分

- 从模板得到的候选答案：将实体从自然问句中抽去，并将剩余的问句部分（如：“_____分布在哪里？”）与模板库中的问句模板进行相似度计算（基于编辑距离）
- 从谓词-同义词集合得到的候选答案：将实体从自然问句中抽去，并对剩余的问句部分进行分词，得到问句的tokens（如“_____分布在哪里？”分词为“分布、哪里”），与同义词集合（如“分布、哪里、区域、地方”等）进行相似度计算（基于Jaccard相似度），问句的tokens与同义词集合的交集越大，说明相似度越高
- 从百科词条得到的答案：若自然问句是询问百科词条类问题（如“白菜是什么”），则从问句中识别出实体（“白菜”），并从百科词条库中搜寻相关词条（如“大白菜”、“白菜”、“青菜”等），并将问题中识别的实体与查询返回的相关百科词条，进行编辑距离相似度计算
- 从FQA常用问答对得到的答案：将问题与返回的问答对中的问题进行字符串相似度计算

谢谢大家！

本课程课件由OpenKG提供支持

联系我们

小象学院：互联网新技术在线教育领航者

- 微信公众号：大数据分析挖掘
- 新浪微博：ChinaHadoop

