

• 当你看见下面这一串文本你会联想到什么?

Ronaldo Luís Nazório de Lima



•估计绝大多数中国人不明白上面的文本代表什么意思。没关系,我们看看它对应的中文:

罗纳尔多•路易斯•纳萨里奥•德•利马

• 这下大部分人都知道这是一个人的名字了,当然,不出什么意外,还是个外国人。但还是有一部分人不知道这个人具体是谁。下面

是关于他的某张图片:



从这张图片我们又得到了额外信息,他是一位足球运动员。对足球不熟悉的可能还是对他没有什么印象。那么再看看下面这张图片:



• 之所以举这样一个例子,是因为,计算机一直面临着这样的困境——无法获取网络文本的语义信息。尽管近些年人工智能得到了长足的发展,在某些任务上取得超越人类的成绩,但离一台机器拥有一个两三岁小孩的智力这样一个目标还有一段距离。这距离的背后很大一部分原因是机器缺少知识。如同上面的例子,机器看到文本的反应和我们看到罗纳尔多葡萄牙语原名的反应别无二致。为了让机器能够理解文本背后的含义,我们需要对可描述的事物(实体)进行建模,填充它的属性,拓展它和其他事物的联系,即,构建机器的先验知识。

• 围绕罗纳尔多这个实体进行相应的扩展,就可以得到下面这张知识图。 机器拥有了这样的先验知识,当它再次看到Ronaldo Luís Nazário de Lima,它就会"想": "这是一个名字叫Ronaldo Luís Nazário de Lima的巴西足球运动员。"这和人类在看到熟悉的事物,会做一些联想和推理是很类似的

运动项目 Ronaldo Luís Nazário de Lima 里奥·达利马 1976年 180cm

MEMEX - 记忆机器

人的记忆偏重关联

"Wholly new forms of encyclopedias will appear, ready made with a mesh of associative trails running through them, ready to be dropped into the memex and there amplified"

As We May Think, The Atlantic, 1945

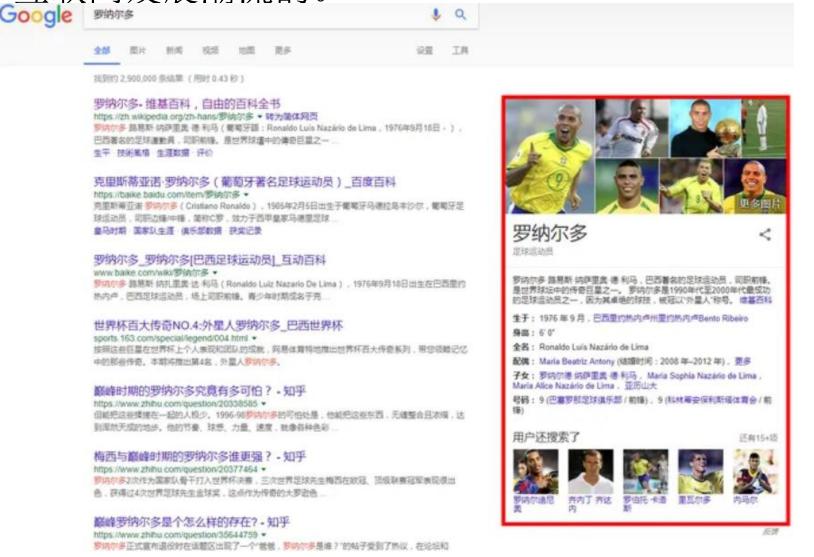


Vannevar Bush

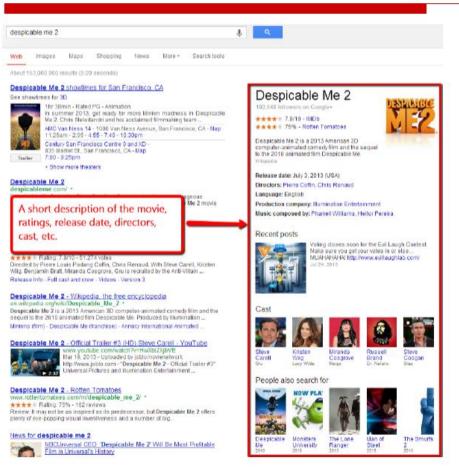
National Science Foundation Manhattan Project First Presidential Science Advisor

- Google为了提升搜索引擎返回的答案质量和用户查询的效率,于 2012年5月16日发布了知识图谱(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。
- 想知道罗纳尔多的相关信息(很多情况下,用户的搜索意图可能也是模糊的),在之前的版本,只能得到包含这个字符串的相关网页作为返回结果;现在,除了相关网页,搜索引擎还会返回一个"知识卡片",包含了查询对象的基本信息和其相关的其他对象。在最短的时间内,我们获取了最为简洁,最为准确的信息。

• 这只是知识图谱在搜索引擎上的一部分应用场景,它的诞生是符合计算机科学、互联网发展潮流的。

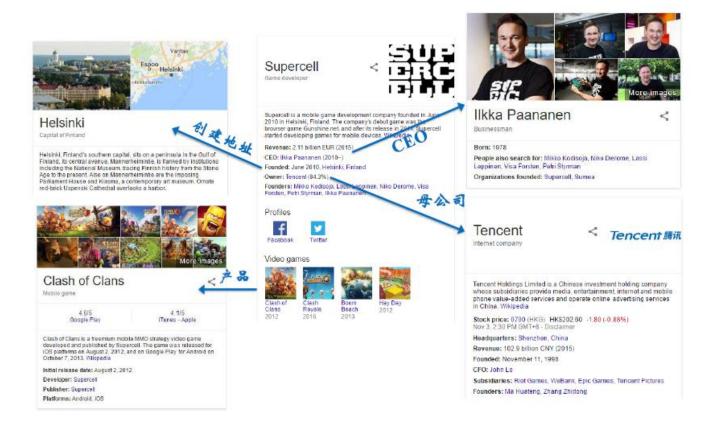


谷歌知识图谱: Things not strings





搜索结果中体现关联



Facebook兴趣图谱



Zhongyuan Xu (徐重远)

Software Engineer at Facebook

Studied at Stony Brook University mutual friend: Ledell Wu

Likes Harvard University, Basketball and 364 others

Structured search within the graph



- •知识图谱的本质是为了表示知识。其实知识图谱的概念并不新,它背后的思想可以追溯到上个世纪五六十年代所提出的一种知识表示形式——语义网络(Semantic Network)。语义网络由相互连接的节点和边组成,节点表示概念或者对象,边表示他们之间的关系(is-a关系,比如:猫是一种哺乳动物;part-of关系,比如:脊椎是哺乳动物的一部分),如下图。
- 在表现形式上,语义网络和知识图谱相似,但语义网络更侧重于描述概念与概念之间的关系,(有点像生物的层次分类体系——界门纲目科属种),而知识图谱则更偏重于描述实体之间的关联。

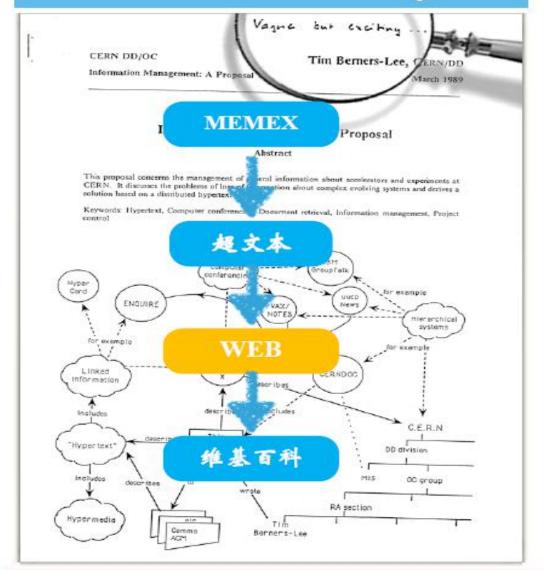
•除了语义网络,人工智能的分支——专家系统,万维网之父Tim Berners Lee于1998年提出的语义网(Semantic Web)和在2006年提出的关联数据(Linked Data)都和知识图谱有着千丝万缕的关系,可以说它们是知识图谱前身。



•即,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SP0三元组(Subject-Predicate-Object)。

Web: 以链接为中心的系统

Linked Information System



...This is why a "web" of notes with links between them is far more useful than a fixed hierarchical system. Circles and arrows leave one free to describe the interrelationships between things in a way that tables, for example, do not. The system we need is like a diagram of circles and arrows, where circles and arrows can stand for anything.

Information Management: A proposal 1989.

以"链接"为中心的系统,在开放的互联网环 境里面更加容易生长和扩展。这一理念逐步被 人们实现,并演化发展成为今天的万维网。



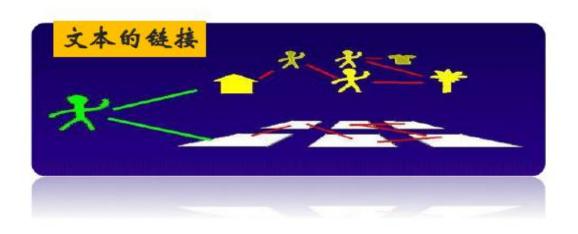
万维网创始人 MIT教授 图灵奖获得者

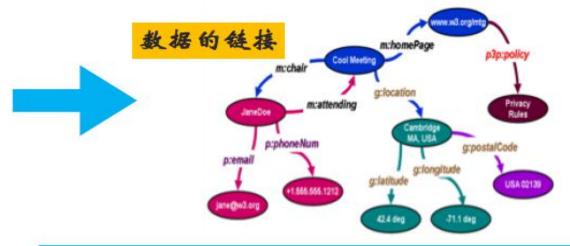
Sir Tim Berners-Lee

语义网:从链接文本到链接数据

"This is a pity, as in fact documents on the web describe real objects and imaginary concepts, and give particular relationships between them but we could not process them at all...

Tim Berners-Lee, Inventor of the Web, @WWW Geneva, 1994

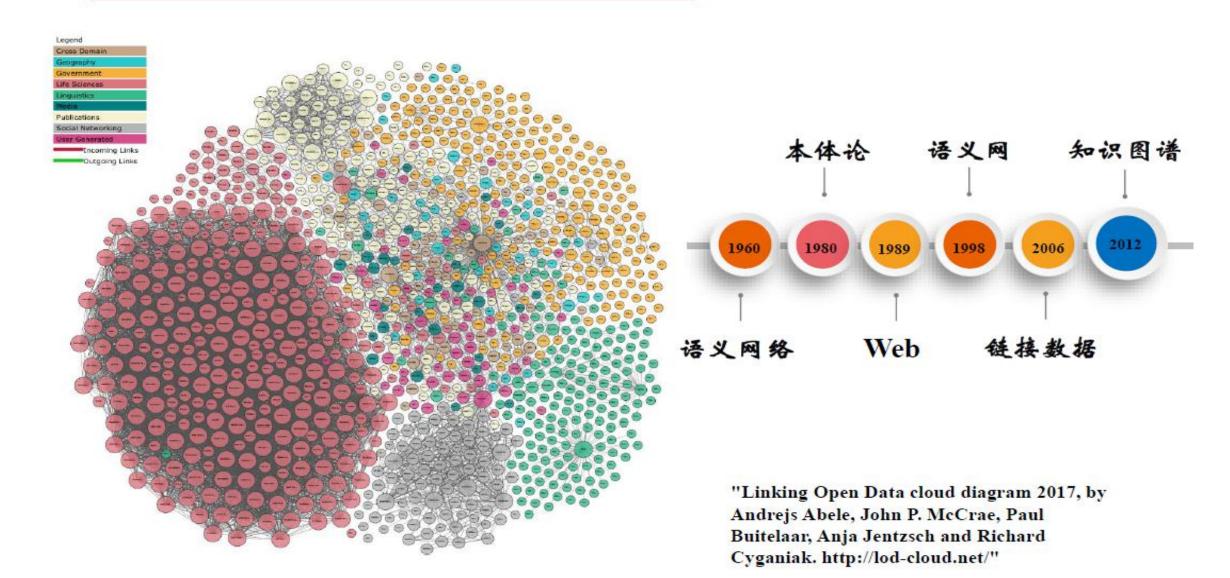




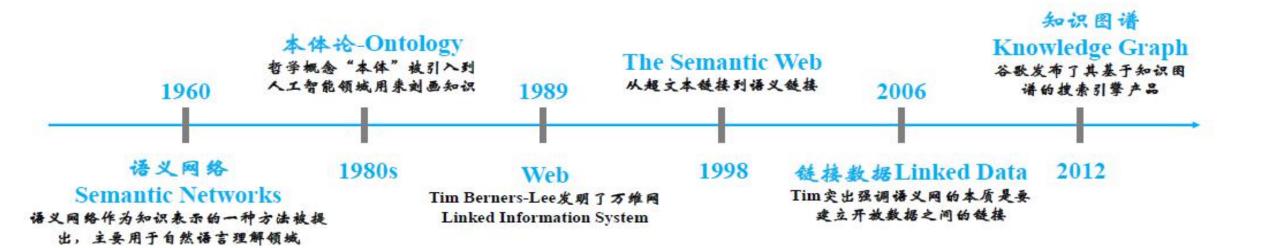
Web of Texts, Web of documents

Web of Objects, Web of Data, Web of Things

知识图谱 (Knowledge Graph, KG)



KG的概念演化



知识表示与知识库- Knowledge Representation / Knowledge Base

人工智能研究者陆续提出了大量知识表示的方法,如框架系统、产生式规则、描述逻辑等。

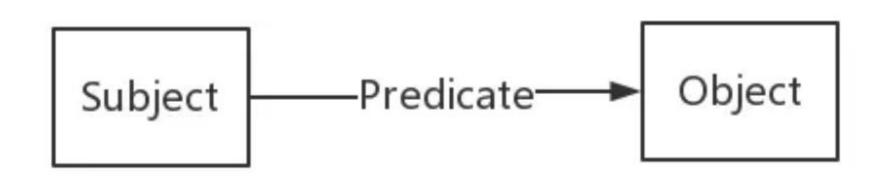
知识图谱得益于Web的发展(更多的是数据层面),有着来源于KR、NLP、 Web、AI多个方面的基因。

• 在知识图谱中,用RDF形式化地表示这种三元关系。RDF(Resource Description Framework),即资源描述框架,是W3C制定的,用于描述实体/资源的标准数据模型。RDF图中一共有三种类型,International Resource Identifiers(IRIs), blank nodes 和literals。

- IRI可以看做是URI或者URL的泛化和推广,它在整个网络或者图中唯一定义了一个实体/资源,和身份证号类似。
- literal是字面量,可以把它看做是带有数据类型的纯文本,比如罗纳尔多原名可以表示为"Ronaldo Luís Nazário de Lima"^xsd:string。
- blank node简单来说就是没有IRI和literal的资源,或者说匿名资源。

• 下面是SPO每个部分的类型约束:

- Subject可以是IRI或blank node。
- Predicate是IRI。
- Object三种类型都可以。

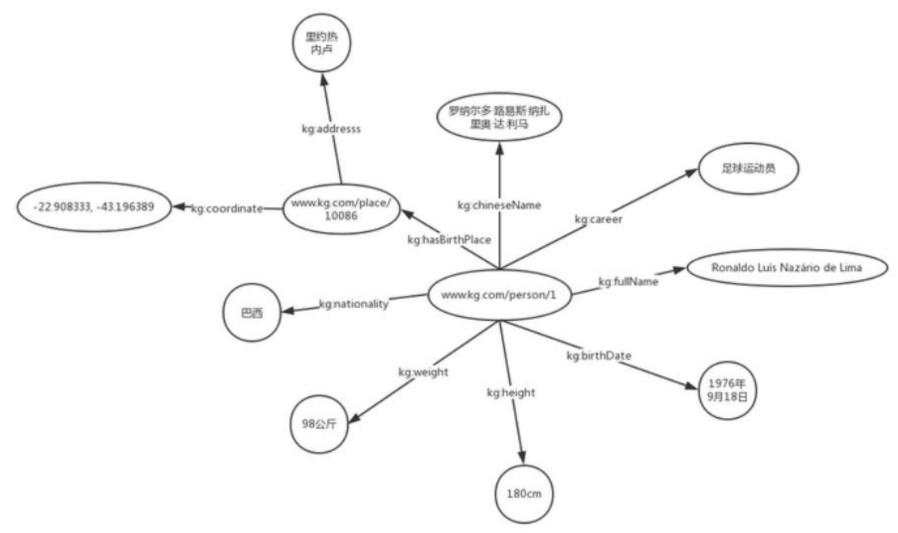


• 那么"罗纳尔多的中文名是罗纳尔多•路易斯•纳扎里奥•达•利马"这样一个三元组用RDF形式来表示就是:



- "www.kg.com/person/1"是一个IRI,用来唯一的表示"罗纳尔多"这个实体。"kg:chineseName"也是一个IRI,用来表示"中文名"这样一个属性。"kg:"是RDF文件中所定义的prefix,如下所示。
- @prefix kg: http://www.kg.com/ontology/
- •即,kg:chineseName其实就是 "http://www.kg.com/ontology/chineseName"的缩写。

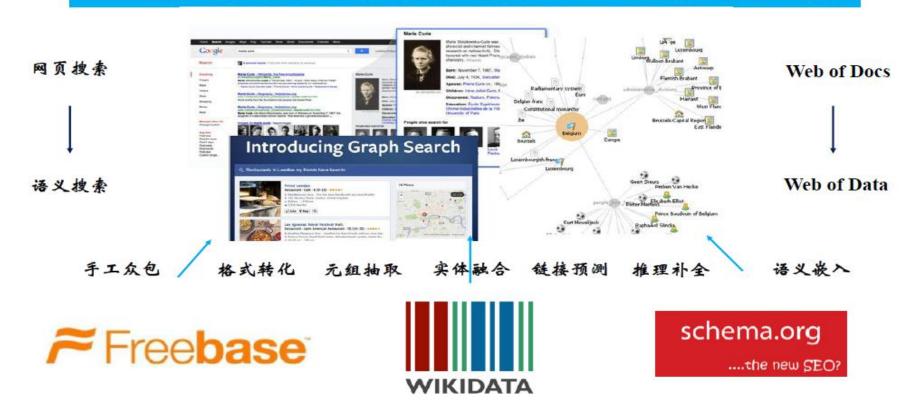
• 将上面的知识图用更正式的形式画出来:



KG辅助搜索

社区协同构建

Web的理想是万物的链接,搜索的理想是事物的搜索



维基众包

网页嵌入语义数据

KG辅助问答

机器人及IoT设备的智能化:给万物都挂接一个背景知识库

对话式的信息获取更加需要精准度和可靠度,知识图谱对于提升用户体验更加不可少。



DUER®S

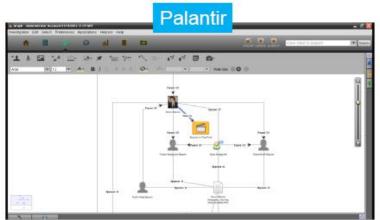
Conversational AI OS for everyone and everywhere





二、知识图谱的前世今生 KG辅助决策







KG辅助AI: 常识推理

当一个人听到或看到一句话的时候,他使用自己所有的知识和智能去理解。这不仅包括语法,也包括他的词汇知识、上下文知识,更重要的,是对相关事务的理解——Terry Winograd (自然语言理解系统SHRDLU作者)

Winograd Schema Challenge

I. The trophy would not fit in the brown suitcase because it was too big (small). What was too big (small)?

Answer 0: the trophy Answer 1: the suitcase

II. The town councilors refused to give the demonstrators a permit because they feared (advocated) violence. Who feared (advocated) violence?

Answer 0: the town councilors Answer 1: the demonstrators

绝NLP: 50% → NLP+KB: 超过60% → 及格线: 90%

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

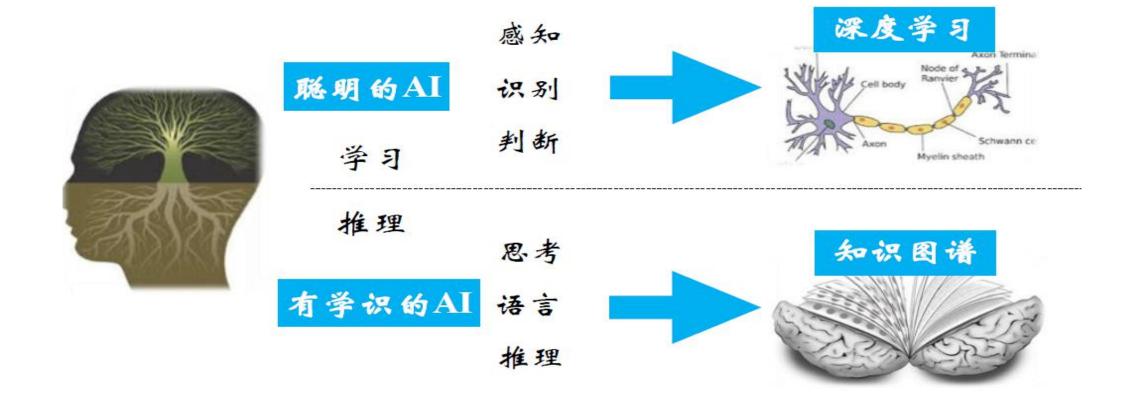
AI视角 怎样利用知识库来辅助理解人的语言

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等多方面的方法和技术

聪明的AIvs有学识的AI

人的大脑依赖所学的知识进行思考、逻辑推理、理解语言



三、典型知识库项目

各种知识图谱项目





































WEBKB

Cyc

- Cyc 是持续时间最久,影响范围较广,争议也较多的知识库项目。Cyc 是在 1984 年由 Douglas Lenat 开始创建。最初的目标是要建立人类最大的常识知识库。典型的常识知识如 "Every tree is a plant", "Plants die eventually"等。
- Cyc 知识库主要由术语 Terms 和断言 Assertions 组成。Terms 包含概念、关系和实体的定义。Assertions 用来建立 Terms 之间的关系,这既包括事实 Fact 描述,也包含规则 Rule 的描述。
- 最新的 Cyc 知识库已经包含有 50 万条 Terms 和 700 万条 Assertions。 Cyc 的主要特点是基于形式化的知识表示方法来刻画知识。形式化的优势是可以支持复杂的推理。但过于形式化也导致知识库的扩展性和应用的灵活性不够。Cyc 提供开放版本 OpenCyc。

WordNet

• WordNet 是最著名的词典知识库,主要用于词义消歧。WordNet 由普林斯顿大学认识科学实验室从1985年开始开发。WordNet 主要定义了名词、动词、形容词和副词之间的语义关系。例如名词之间的上下位关系(如:"猫科动物"是"猫"的上位词),动词之间的蕴含关系(如:"打鼾"蕴含着"睡眠")等。WordNet3.0已经包含超过15万个词和20万个语义关系。

WordNet: a lexical database for English
GA Miller - Communications of the ACM, 1995 - dl.acm.org

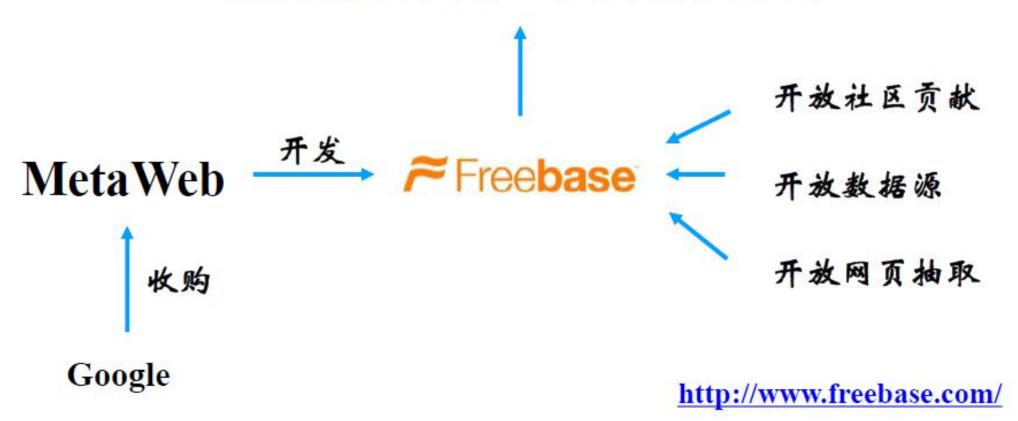
Abstract Because meaningful sentences are composed of meaningful words, any system that hopes to process natural languages as people do must have information about words and their meanings. This information is traditionally provided through dictionaries, and 被引用次数: 9323 相关文章 所有 34 个版本 引用 保存 更多

ConceptNet

- ConceptNet是常识知识库。最早源于MIT媒体实验室的Open Mind Common Sense (OMCS)项目。OMCS项目是由著名人工智能专家Marvin Minsky于1999年建议创立。
- ConceptNet主要依靠互联网众包、专家创建和游戏三种方法来构建。 ConceptNet知识库以三元组形式的关系型知识构成。
- ConceptNet5版本已经包含有2800万关系描述。与Cyc相比, ConceptNet采用了非形式化、更加接近自然语言的描述,而不是像Cyc 那样采用形式化的谓词逻辑。与链接数据和谷歌知识图谱相比, ConceptNet比较侧重于词与词之间的关系。
- 从这个角度看,ConceptNet更加接近于WordNet,但是又比WordNet包含的关系类型多。此外,ConceptNet完全免费开放,并支持多种语言。

Freebase

完全免费并允许商业化的开放许可协议



通过开源免费吸引用户贡献数据,增值的应用及技术服务收费

Wikidata

目标是构建全世界最大的免费知识库,采用CCO完全自由许可协议



维基基金会



开放社区贡献



相对小众、专业、严格的众包社区

提供资金 导入Freebase数据

提供资金 提供资金

谷歌

摩尔基金会

Allen AI Institute

http://www.wikidata.org/

仍然面临知识缺失严重的问题

DBPedia

- DBPedia是早期的语义网项目。DBPedia意指数据库版本的Wikipedia,是从Wikipedia抽取出来的链接数据集。
- DBPedia采用了一个较为严格的本体,包含人、地点、音乐、电影、组织机构、物种、疾病等类定义。
- •此外,DBPedia还与Freebase,OpenCYC、Bio2RDF等多个数据集建立了数据链接。DBPedia采用RDF语义数据模型,总共包含30亿RDF三元组。

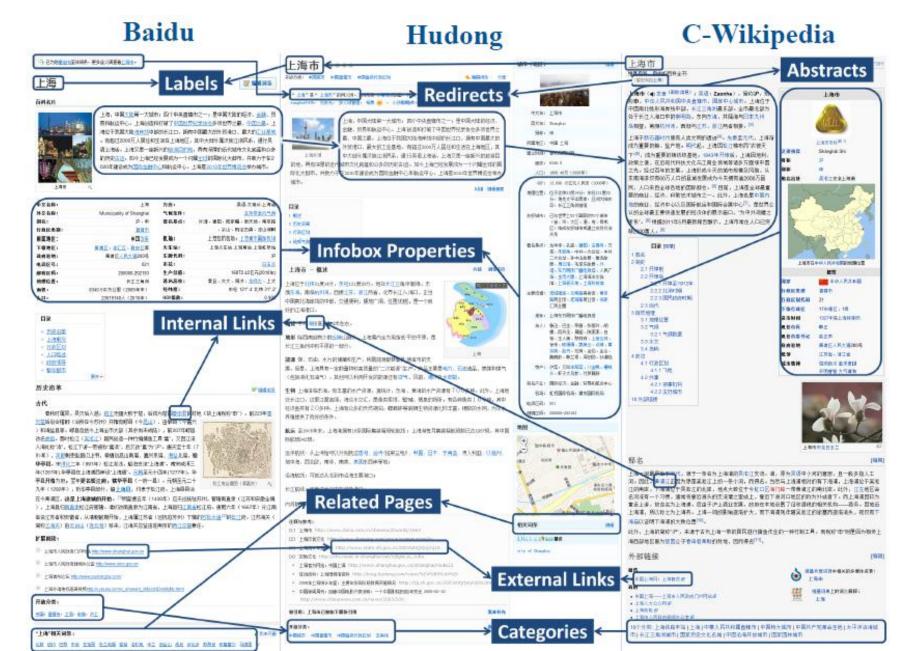
YAGO

- YAGO是由德国马普研究所研制的链接数据库。YAGO主要集成了 Wikipedia、WordNet和GeoNames三个来源的数据。
- YAGO将WordNet的词汇定义与Wikipedia的分类体系进行了融合集成,使得YAGO具有更加丰富的实体分类体系。YAGO还考虑了时间和空间知识,为很多知识条目增加了时间和空间维度的属性描述。
- •目前,YAGO包含1.2亿条三元组知识。YAGO是IBM Watson的后端知识库之一。

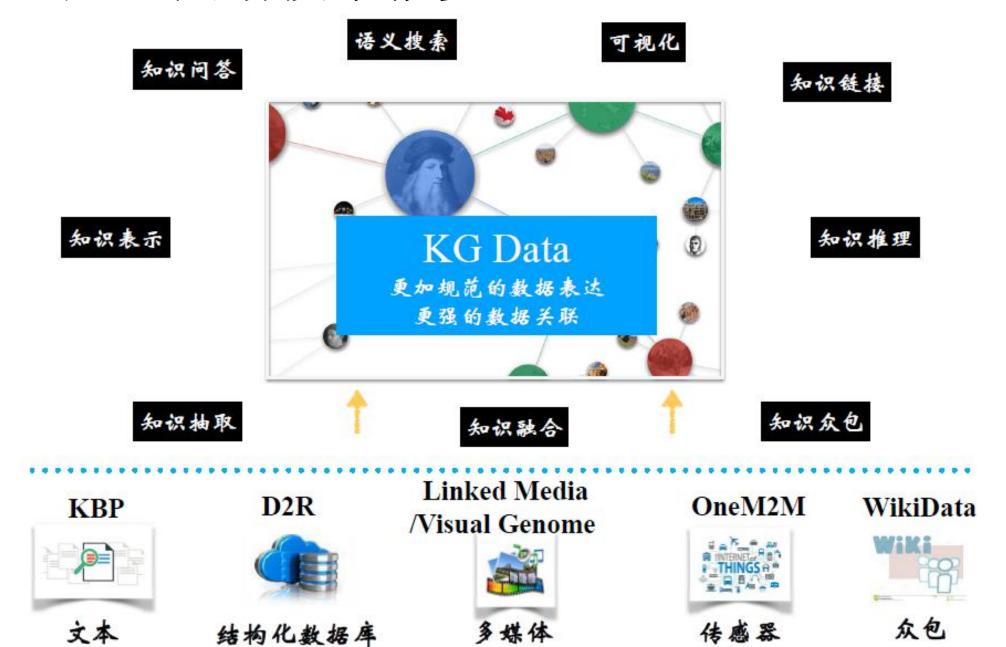
中文知识图谱资源: OpenKG. CN:



Zhishi. me



四、知识图谱技术概览



什么是知识表示

1960s - 1970s

Semantic Nets - Quillian

1970s - 1980s

Production Rules & Expert Systems

1980s - 1990s

Frame Systems & KL-ONE - Marvin Minsky

1990s - 2000s

Description Logic / DAML + OIL - Ian Horrocks

2000s-2010s

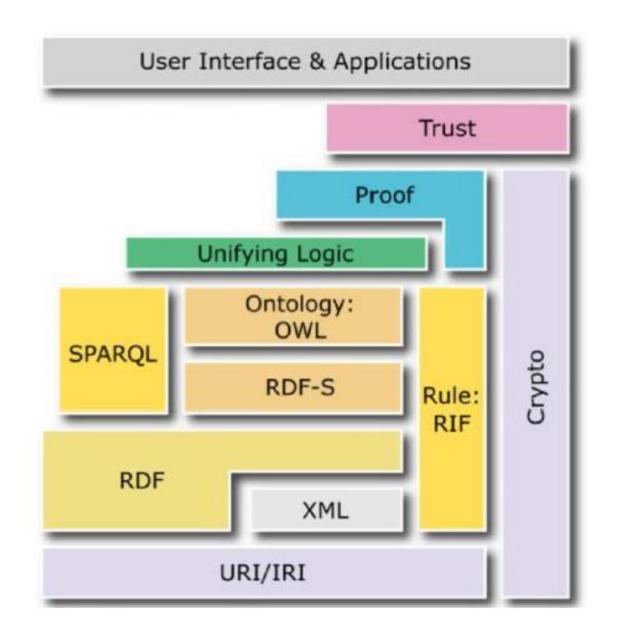
The Semantic Web (RDF & OWL) - Tim Berners Lee

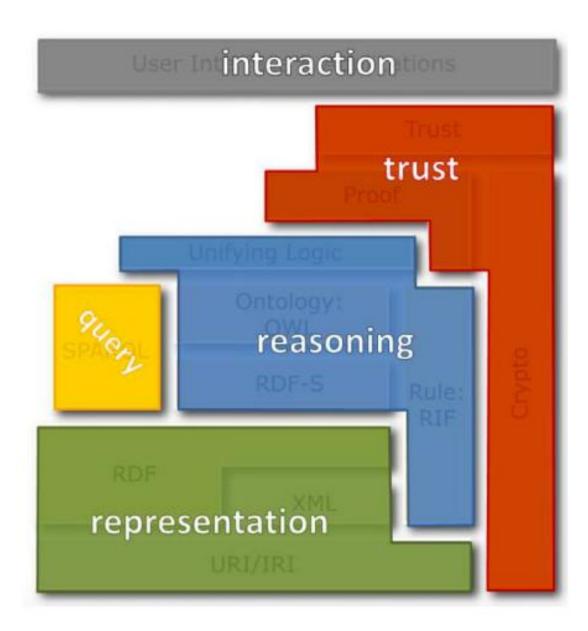
知识表示研究怎样用计算机符号来表示人脑 中的知识,以及怎样通过符号之间的运算来 模拟人脑的推理过程。

基于数理逻辑的知识表示

基于向量空间学习的分布式知识表示

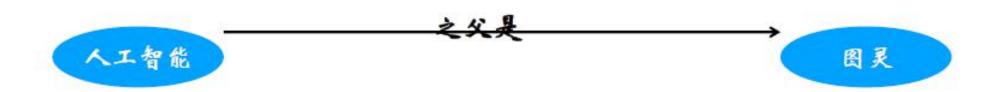
语义网知识表示框架





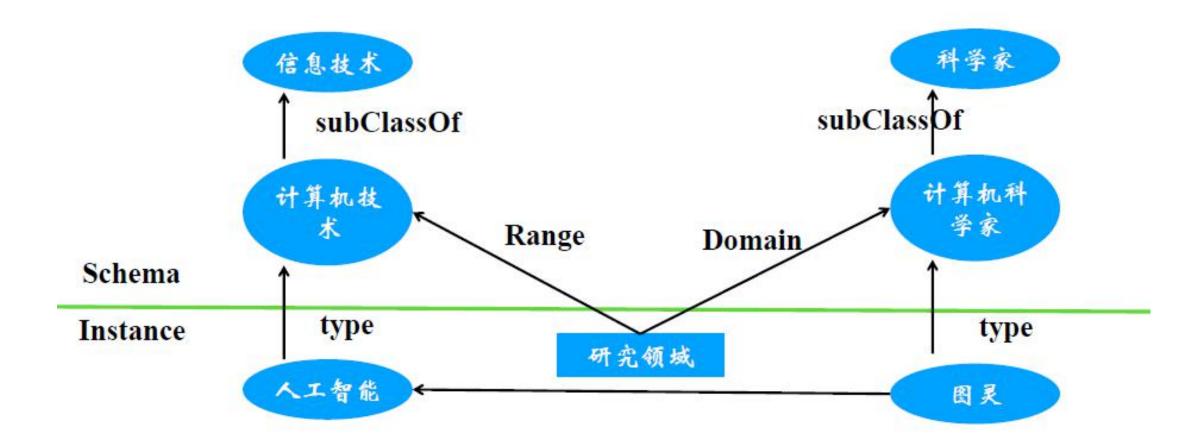
RDF





RDFS

Defines small vocabulary for RDF: Class, subClassOf, type, Property, subPropertyOf, Domain, Range



OWL

Class, subclass, Property, subProperty.....



Complex Classes: intersection, union and complement

Property Restrictions: existential quantification, universal quantification, has Value

Cardinality Restrictions: maxQualifiedCardinality, minQualifiedCardinality, qualifiedCardinality

Property Characteristics: inverseOf, SymmetricProperty, AsymmetricProperty, propertyDisjointWith, ReflexiveProperty, FunctionalProperty

OWL示例

复杂类

属性约束

对称属性

:hasSpouse rdf:type owl:SymmetricProperty .



:hasAncestor rdf:type owl:TransitiveProperty .



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

SPARQL

• SPARQL即SPARQL Protocol and RDF Query Language的 递归缩写,专门用 于访问和操作RDF数 据,是语义网的核 心技术之一。

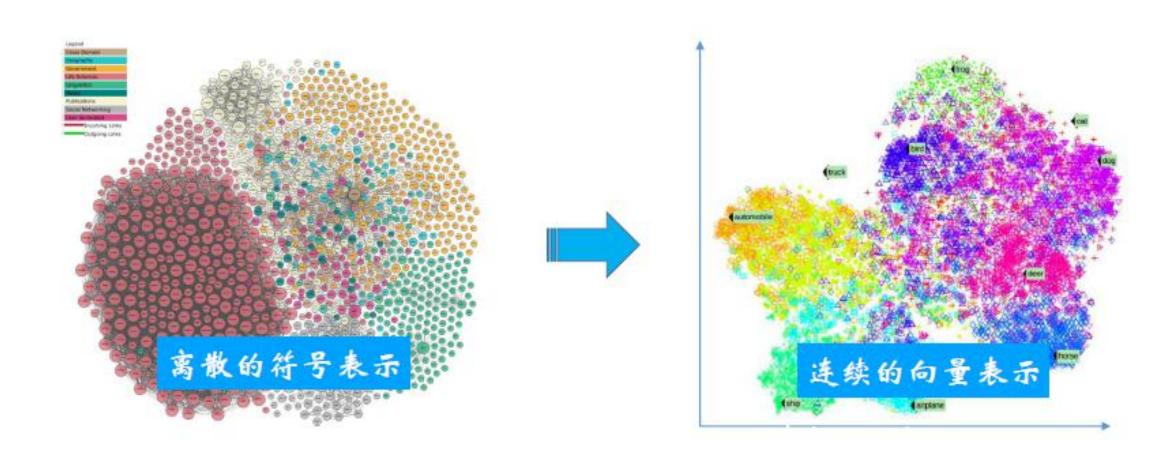
A SPARQL query comprises, in order:

- 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

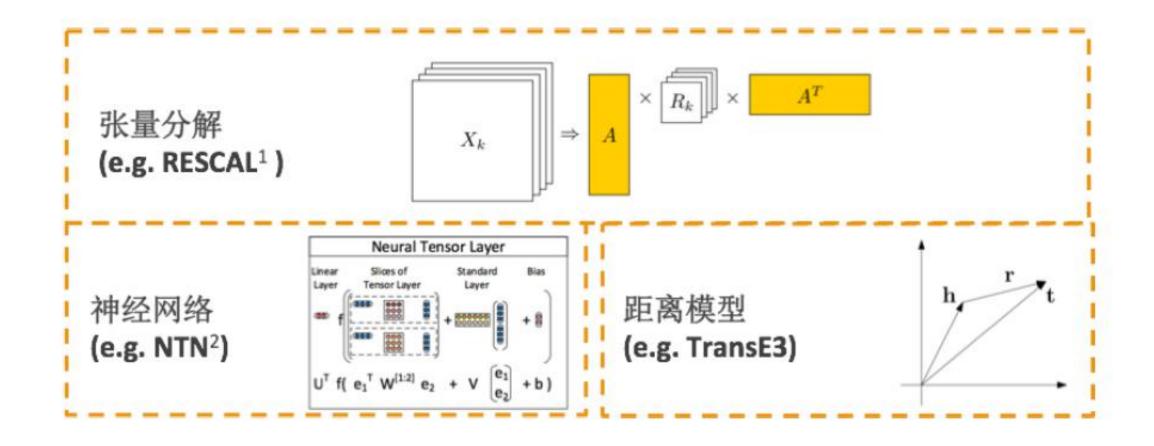
详见: http://www.w3.org/TR/rdf-sparql-query/

知识图谱的分布式表示

在保留语义的同时,将知识图谱中的实体和关系映射到连续的稠密的低维向量空间



知识图谱的分布式表示



M. Nickel, V. Tresp, and H. Kriegel, A three-way model for collective learning on multi- relational data. ICML2011
Socher, R.; Chen, D.; Manning, C. D.; and Ng, A. . Reasoning with neural tensor networks for knowledge base completion. NIPS2013
A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, Translating em- beddings for modeling multi-relational data. NIPS2013

知识抽取: NLP+KR

KR: 三元组、多元关系、模态知识......

关系抽取 事件抽取

NER命名实体识别

Entity Linking 实体链接

分词 词性标注 语法解析 依存分析

文本预处理

各种非结构化文本数据







TAC 的KBP Track

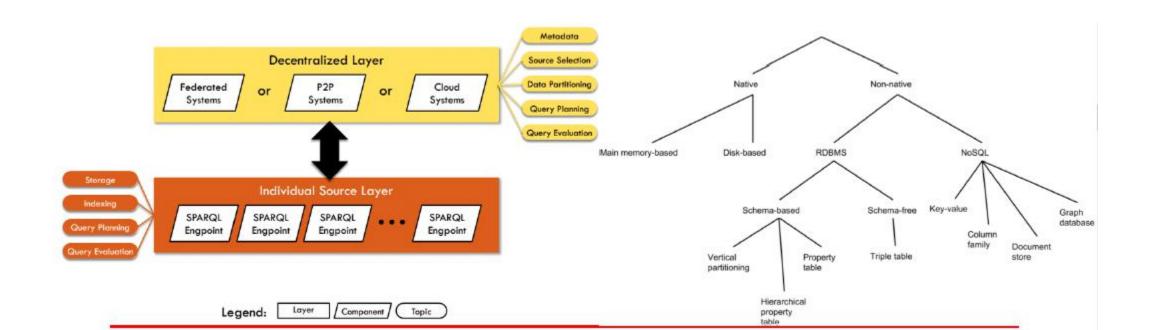




MUC-7

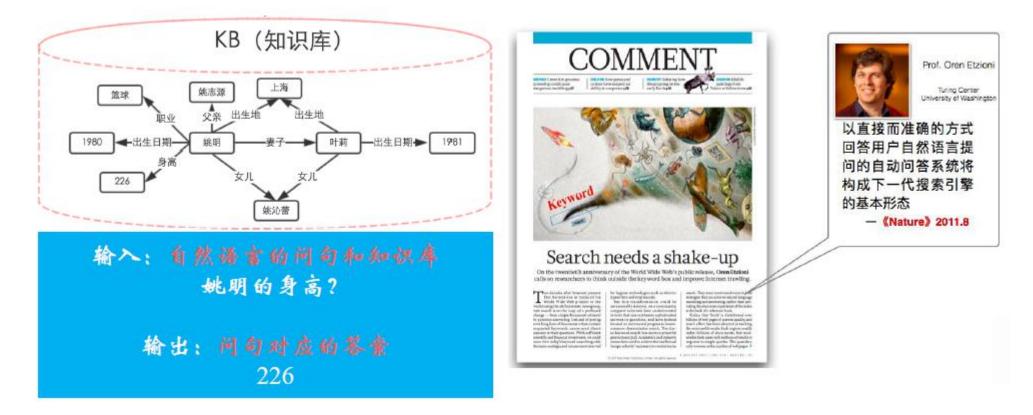
知识存储

- •知识比数据的结构更加复杂,知识的存储需要综合考虑图的特点、复杂的知识结构存储、索引和查询(支持推理)的优化等问题。
- 典型引擎分为基于关系数据库的存储和基于原生图的存储。

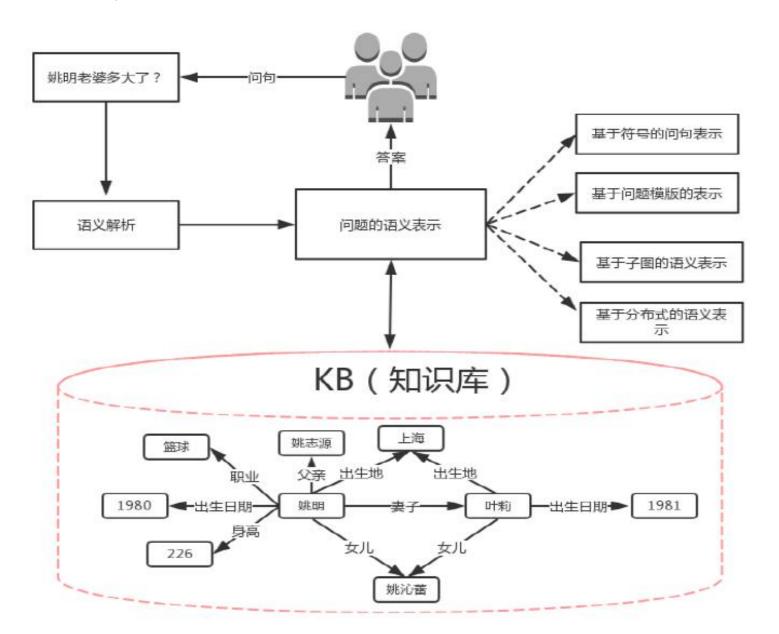


知识问答KBQA

• 以直接而准确的方式回答用户自然语言提问的自动问答系统将构成下一代搜索引擎的基本形态。



KBQA基本实现流程



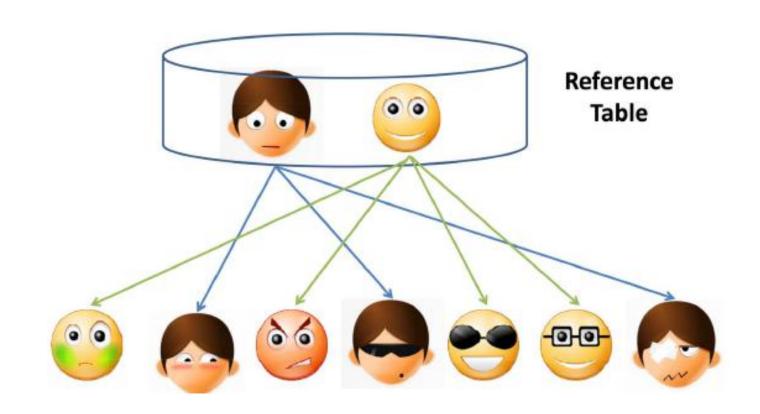
知识推理

- 推理是指基于已知事实推出未知的事实的计算过程。
- 大部分搜索引擎目前都已经支持一定的推理能力。



知识融合

• 指在不同数据集中找出同一个实体的描述记录,主要目的是对不同数据源中的实体信息进行整合,形成更加全面的实体信息。



五、典型案例



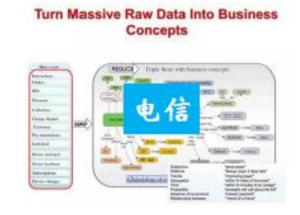














中医药知识平台



主要目标

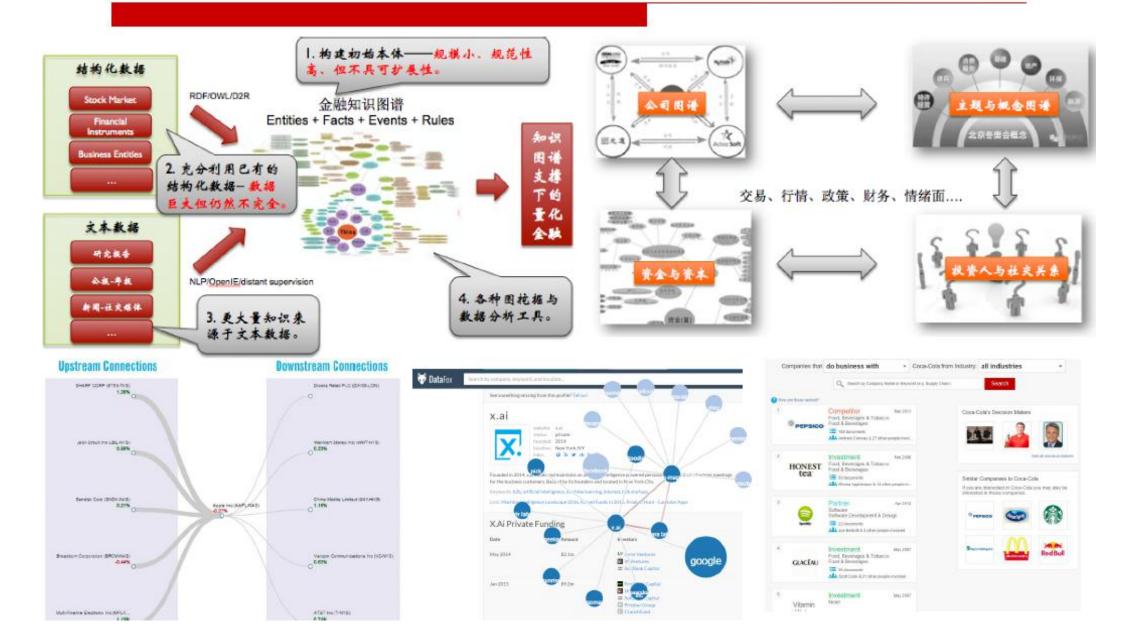
- •对中医药知识体系进行系统梳理、建模和展示
- •知识可视化以图形方式凸显核心概念之间的关系
- ·辅助中医专家厘清学术发展脉络,浏览中医知识,发现知识点之间的联系。与阅读文献等手段相比,可节约知识检索获取时间

电商知识图谱



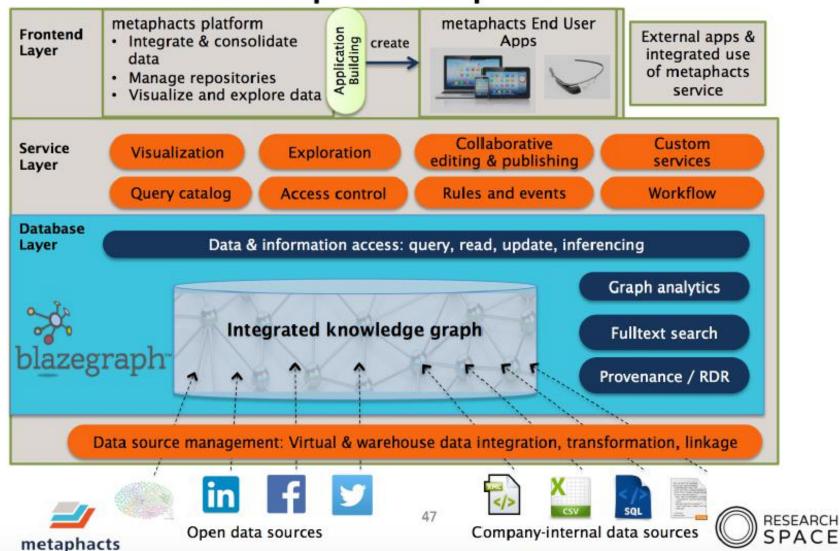
引自:张伟(览图) — "阿里巴巴知识图谱"

金融: 链接金融数据



大英博物馆: 后端平台

metaphacts platform



BBC: 打通BBC全领域

BBC ONTOLOGIES

- Programmes Ontology
- Wildlife Ontology
- Sport Ontology
- Curriculum Ontology
- Corenews Ontology
- Storyline Ontology

