



知识图谱原理与 应用概述

彭鹏

湖南大学

hnu16pp@hnu.edu.cn

参考教材

- 知识图谱：概念与技术

作者: 肖仰华 等

出版社: 电子工业出版社

- 考核形式:

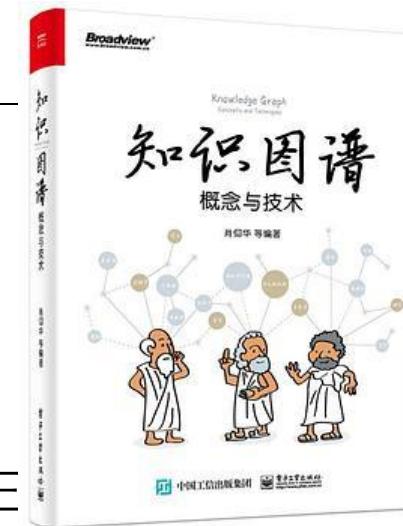
(1) 课程报告 (ISWC、ESWC近三

(2) 课程论文 (占30%)

(3) IBM Skills Academy上的注册与学习 (占40%)

- 课程网站:

<https://bnu05pp.github.io/KGGraduateCourse/2023.html>



目 录

1. 什么是知识图谱?
2. 知识图谱研究的多个维度
3. 从人工智能和大数据的角度看待知识图谱
4. 知识图谱的项目应用

Part 1

什么是知识图谱?

知识图谱 (Knowledge Graph)

2012年5月16日，Google发布“知识图谱”的新一代“智能”搜索功能。

The screenshot shows a Google search results page for the query "瓦特". The top navigation bar includes the Google logo, a search input field containing "瓦特", and a microphone icon. Below the search bar are links for "全部" (All), "图片" (Images), "地图" (Maps), "视频" (Videos), "新闻" (News), and "更多" (More). The search results indicate approximately 15,700,000 results found in 0.63 seconds.

The first result is a link to zh.wikipedia.org, titled "瓦特- 维基百科, 自由的百科全书". Below it is a snippet from baike.baidu.com, also titled "瓦特 (功率单位) _百度百科". The third result is another link to baike.baidu.com, titled "詹姆斯·瓦特_百度百科".

A prominent feature on the right is a "Knowledge Graph" card for "瓦特". It displays the name "瓦特" in large letters, its status as a "功率单位" (unit of power), and a portrait of James Watt. A brief description notes that Watt's definition is 1焦耳/秒 (1 J/s), and it mentions the common use of kilowatts in daily life. It also provides the conversion 1千瓦=1000瓦特 and states that Watt is used for alternating current power. A link to "维基百科" is provided at the bottom of the card.

At the bottom of the card, there is a "反馈" (Feedback) link. Below the card, a box contains the text "查看以下内容的结果:" (Results for the following content:).

传统互联网搜索技术

基于关键词字符串匹配

1. 爬取网页建立倒排索引
2. 基于倒排索引，搜索引擎首先找到了包含关键词的网页
3. 根据打分策略（如PageRank, HITS等），对网页排序
4. 返回用户

Baidu 百度 瓦特 百度一下

网页 资讯 视频 图片 知道 文库 贴吧 采购 地图 更多» 搜索工具

百度为您找到相关结果约30,700,000个

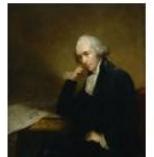
功率 长度 面积 体积 质量 温度 压力 功能/热 <>

1	瓦(W)	全部	
0.0013596	米制马力(ps)	1	焦耳·秒(J/s)
0.000239	千卡/秒(kcal/s)	0.0009478	英热单位/秒(Btu/s)
0.001	千瓦(kW)	0.1019716	公斤·米/秒(kg·m/s)

查看其他单位换算 国际单位：瓦(W)

open.baidu.com/ - ►

[瓦特_百度百科](#)



职业：发明家，学者
生卒：1736年1月19日-1819年8月25日
代表作品：改良蒸汽机
简介：詹姆斯·瓦特 (James Watt, 1736年1月19日 — 1819年8月25日) 英国发明家，第一次工业革命的重要人物。1776年制造...
[人物生平](#) [个人生活](#) [主要影响](#) [人物争议](#) [人物轶事](#) [更多>>](#)
<https://baike.baidu.com/>

计算机并不理解背后的语义

知识图谱的目的

构建知识图谱的目的，就是让机器具备认知能力，理解这个世界。

Baidu 百度 瓦特是哪个学校的校友？ 百度一下

网页 资讯 视频 图片 知道 文库 贴吧 采购 地图 更多»

百度为您找到相关结果约359,000个

▼ 搜索工具

瓦特是哪个大学毕业的_百度知道
2个回答 - 回答时间: 2016年12月1日
最佳答案: 瓦特没受过正规的大学教育。 瓦特的求学经历: 1757年,格拉斯哥大学教授提供给瓦特一个机会,让他在大学里开设了一间小修理店,这帮助瓦特...
[更多关于瓦特是哪个学校的校友?的问题>>](#)
百度知道 - 百度快照

瓦特是以什么身份迈进大学的校门的_百度知道
1个回答 - 回答时间: 2020年1月2日
最佳答案: 像瓦特这样的人才正是当时的格拉斯哥大学所紧缺的,再加上迪克博士和学校教授们的帮助,学校当局终于同意在校园里给瓦特一个工作间,并且授予了他一个“大学...
[更多关于瓦特是哪个学校的校友?的问题>>](#)
百度知道 - 百度快照

瓦特?据说这是最难毕业的学校!_博洛尼亚
2018年12月19日 - 这座城市也因其欧洲著名学术中心的地位闻名于世——天文望远镜的发明者伽利略和以花心浪荡出名的卡萨诺瓦都曾是这所学校的校友。 帕多瓦大学建立于1222年,当年...
搜狐网 - 百度快照

厉害了,进了这个英国名校,你就是亚当斯密的校友了!
2018年3月5日 - 1870年以后,格拉斯哥大学从市中心搬到了西区(West End),学校规模不断扩大,逐渐...瓦特成名后,格拉斯哥大学授予他法医学博士学位,属于“编外”校友。 图片9:格大...
搜狐网 - 百度快照



瓦特是哪个学校的校友?

格拉斯哥大学_百度百科

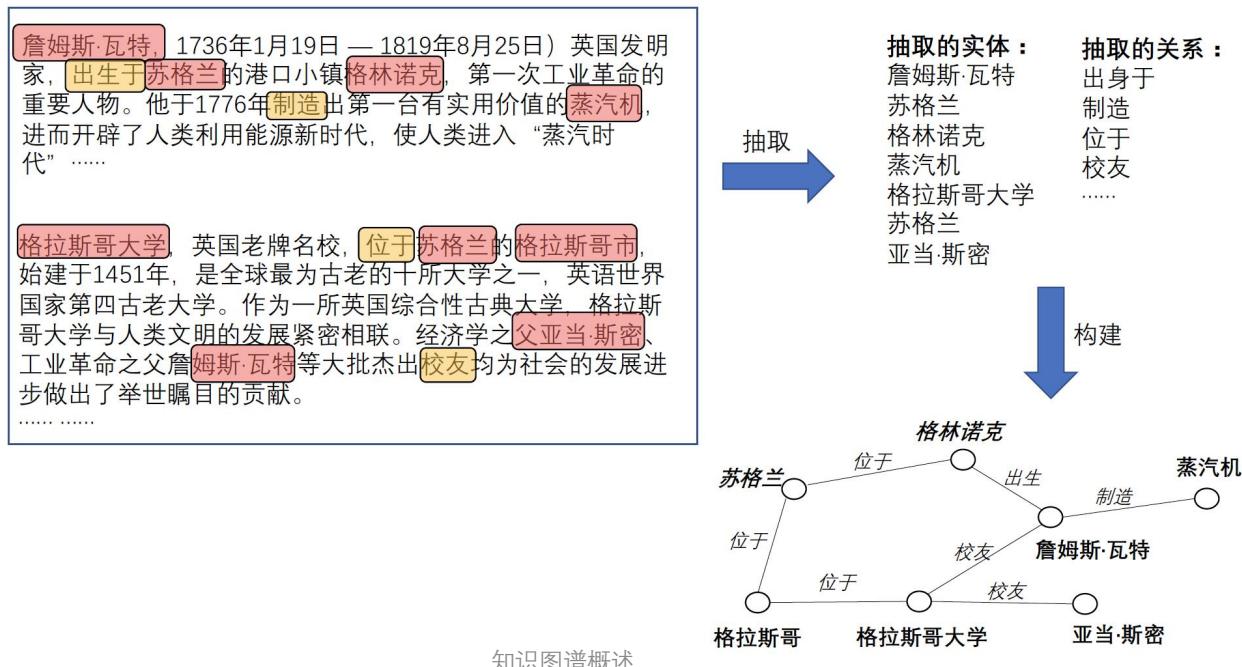


格拉斯哥大学 (University of Glasgow)，简称格大，始建于1451年，位于英国苏格兰最大城市格拉斯哥市，世界百强名校，英国顶尖学府，全球最古老的十所大学之一，英语世界国家第四古老的大学。格大同时也是英国罗素大学集团和Universitas 2...
[历史发展](#) [学校排名](#) [地理交通](#) [学院介绍](#) [学校建筑](#) [更多>>](#)
<https://baike.baidu.com/>

知识图谱 (Knowledge Graph)

什么是知识图谱?

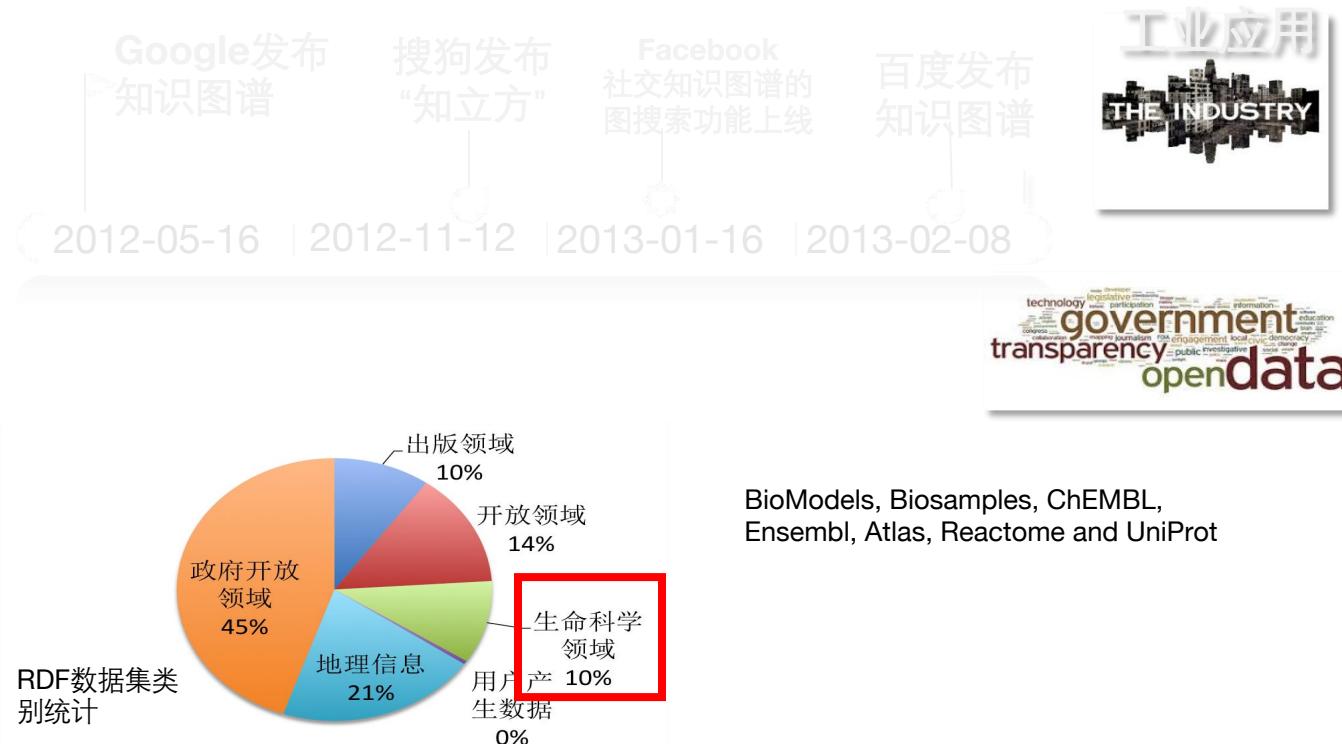
知识图谱本质上是基于图的语义网络，表示实体和实体之间的关系！



知识图谱 (Knowledge Graph)



知识图谱 (Knowledge Graph)



Facebook Social Graph

The screenshot shows the Facebook for Developers website. The top navigation bar includes links for Products, Docs, Tools & Support, News, and Videos, along with a search bar and a 'Register' button. The main content area has a sidebar on the left containing a 'Graph API' section with links for Overview, Using the Graph API, Reference, Common Scenarios, Other APIs, Advanced, and Changelog. The main content area displays the 'Docs / Graph API / Overview / On this page: ▾' path and the title 'The Basics'. It explains that the Graph API represents a 'social graph' composed of nodes, edges, and fields. Below this, it states that the Graph API is HTTP based and can be used directly in a browser.

facebook for developers Products Docs Tools & Support News Videos Search Register

All Docs ▼ Docs / Graph API / Overview / On this page: ▾

Graph API

[Overview](#)
[Using the Graph API](#)
[Reference](#)
[Common Scenarios](#)
[Other APIs](#)
[Advanced](#)
[Changelog](#)

The Basics

The Graph API is named after the idea of a 'social graph' - a representation of the information on Facebook composed of:

- **nodes** - basically "things" such as a User, a Photo, a Page, a Comment
- **edges** - the connections between those "things", such as a Page's Photos, or a Photo's Comments
- **fields** - info about those "things", such as a person's birthday, or the name of a Page

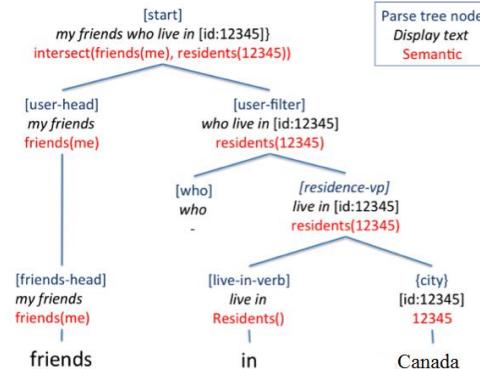
The Graph API is HTTP based, so it works with any language that has an HTTP library, such as cURL, urllib. We'll explain a bit more about what you can do with this in the section below, but it means you can also use the Graph API directly in your browser, for example a Graph API request is equivalent to:

Facebook Social Graph

2013年1月16日 Facebook Graph Search 产品
发布会---Mark Zuckerberg

“My friends who live in Canada”

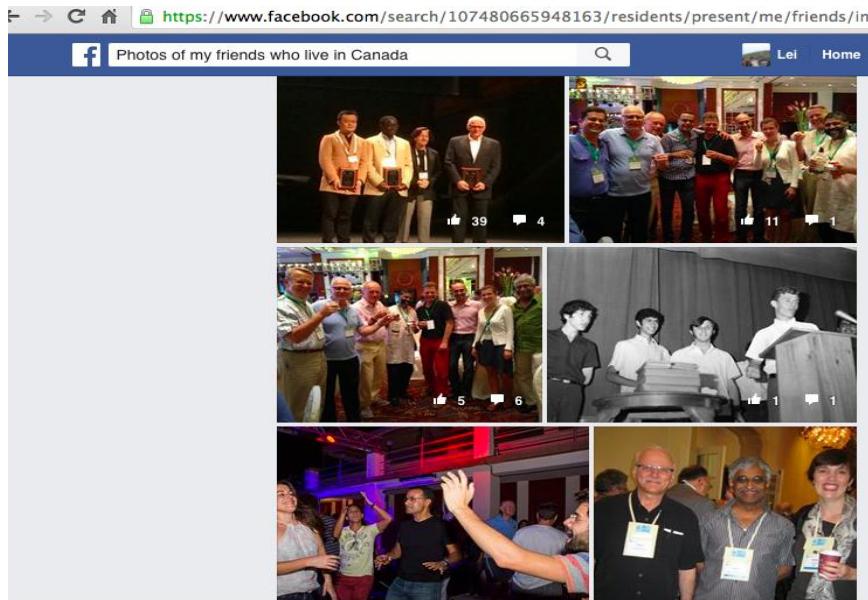
The screenshot shows the Facebook search interface with the query "my friends who live in Canada". The results are filtered by "Top". The first result is "My friends who live in Canada" posted by M. Tamer Ozsu, with 3 new posts. It shows he is a friend since April 2009, lives in Kitchener, Ontario, and works at University of Waterloo. The second result is Bin Zhou, a friend since June 2008, living in Vancouver, British Columbia, working at Microsoft. The third result is Ihab Ilyas, a friend since March 2010, living in Waterloo, Ontario, and a professor at University of Waterloo. The sidebar includes filters for "POSTED BY" (Anyone), "TAGGED LOCATION" (Anywhere, Beijing, China), and "DATE POSTED" (Any time, 2016, 2015, 2014).



The parse tree, semantic and entity ID used in the above example are for illustration only; they do not represent real information used in Graph Search Beta

Facebook Social Graph

“Photos of my friends who live in Canada”



Part 2

知识图谱研究的多个维度

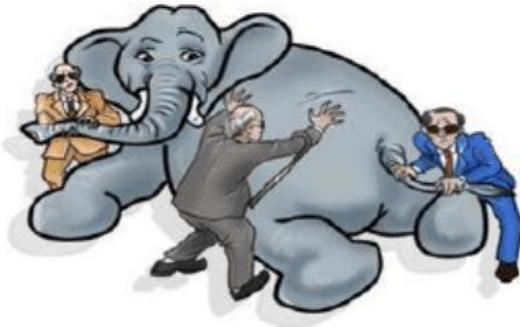
知识图谱的相关领域发展

数据库

RDF数据库系统
数据集成，知识融合

自然语言处理

信息抽取
语义解析



机器学习

知识图谱数据的知识表示
(Graph Embedding)

知识工程

知识库构建
基于规则的推理

知识图谱概述

知识图谱与知识工程

知识图谱是Web 和大数据时代的知识工程新的发展形态。知识工程的核心：[知识库](#)和[推理引擎](#)。

- **领域本体的构建**：面向特定领域的形式化地对于共享概念体系的明确而又详细的说明
- **知识抽取**：从海量的数据中通过信息抽取的方式获取知识
- **知识融合**：通过对多个相关知识图谱的对齐、关联和合并，使其称为一个有机的整体，以提供更全面知识

传统知识工程 Vs. 以知识图谱为代表的新一代知识工程

“Knowledge is the power in AI” --- Edward Albert Feigenbaum

知识本体

本体的定义

- “**本体**”概念来源于哲学领域，指的是对客观存在系统的解释和说明。
- 计算机科学中，“**本体**”用于面向特定领域的形式化地对于共享概念体系的明确而又详细的说明。

它提供了面向特定领域的概念、对象类型、上下位语义关系等以及它们的属性等，是对特定领域之中概念及其相互之间关系的形式化表达，从而方便地进行自动推理等功能

本体语言

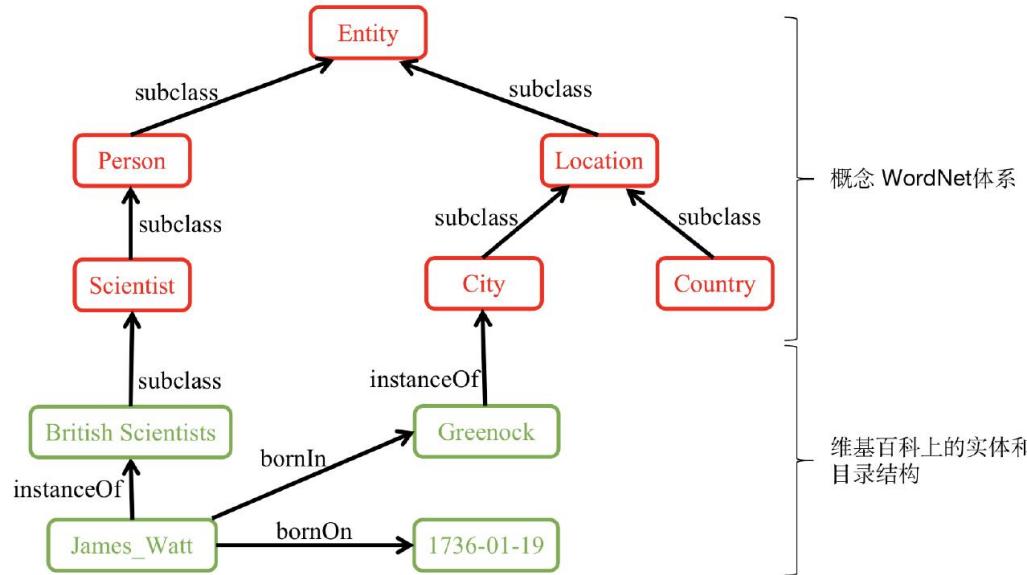
DARPA提出的DAML(DARPA Agent Markup Language)、W3C 提出DAML+OIL3以及目前知识图谱数据集常用W3C 所定义的RDF(S)和OWL 语言等。

本体工具

Protege 和 WebOnto

知识本体

Yago知识库中本体

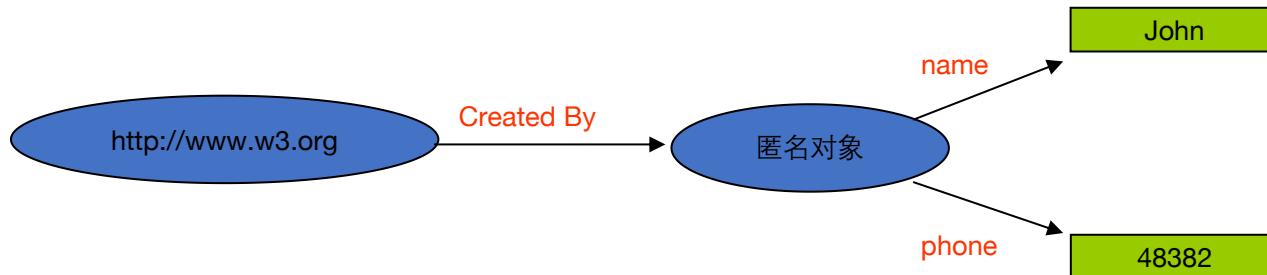


知识图谱数据模型

RDF (Resource Description Framework)

RDF定义了一个简单的模型，用于**描述资源，属性和值之间的关系**。资源是可以用URI标识的所有事物，属性是资源的一个特定的方面或特征，值可以是另一个资源，也可以是字符串。总的来说，一个RDF描述就是一个三元组：<主语、谓词、宾语>。

□ 用有向图表示的RDF示例：



知识图谱数据模型

RDF (Resource Description Framework)

RDF定义了一个简单的模型，用于**描述资源，属性和值之间的关系**。资源是可以用URI标识的所有事物，属性是资源的一个特定的方面或特征，值可以是另一个资源，也可以是字符串。总的来说，一个RDF描述就是一个三元组：<主语、谓词、宾语>。

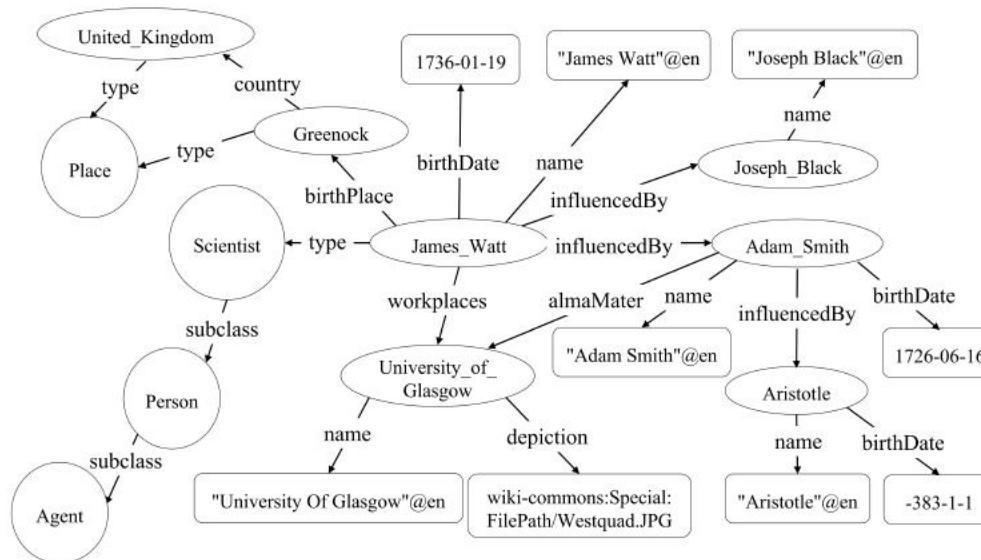
□ 用列表表示的RDF示例：

Subject	Predicate	Object
http://w3.org/	created_by	#anonymous
# anonymous	name	"John"
# anonymous	phone	"477738"

知识图谱数据模型

□ RDF

RDF所表示的瓦特知识图：



知识图谱数据模型

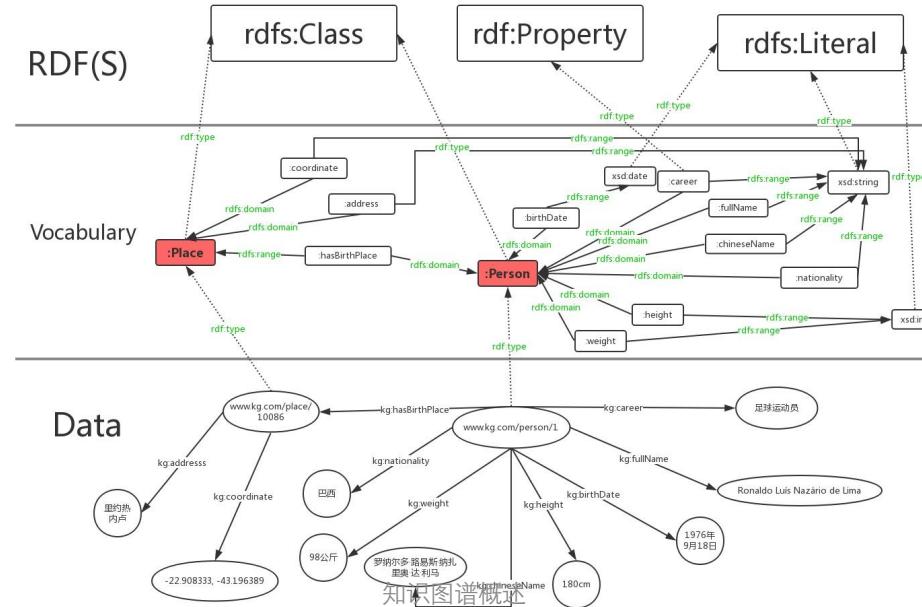
□ RDF

主体	属性	客体
dbr:James_Watt	rdfs:label	"James Watt"@en
dbr:James_Watt	dbo:birthDate	"1736-01-19"^^xsd:date
dbr:James_Watt	dbo:birthPlace	dbr:Greenock
dbr:James_Watt	rdf:type	dbo:Scientist
dbr:James_Watt	dbo:influencedBy	dbr:Joseph_Black
dbr:James_Watt	dbo:influencedBy	dbr:Adam_Smith
dbr:James_Watt	dbp:workplaces	dbr:University_of_Glasgow
dbr:Adam_Smith	rdfs:label	"Adam Smith"@en
dbr:Adam_Smith	dbo:birthDate	"1723-06-16"^^xsd:date
dbr:Adam_Smith	dbo:almaMater	dbr:University_of_Glasgow
dbr:Adam_Smith	dbo:influencedBy	dbr:Aristotle
dbr:Joseph_Black	rdfs:label	"Joseph Black"@en
dbr:University_of_Glasgow	name	"University Of Glasgow"@en
dbr:Aristotle	rdfs:label	"Aristotle"@en
dbr:Aristotle	dbo:birthDate	"-383-1-1"^^xsd:date
dbr:Greenock	dbo:country	dbr:United_Kingdom
dbr:Greenock	dbo:type	dbo:Place
dbr:United_Kingdom	dbo:type	dbo:Place
dbo:Scientist	dbo:subClassOf	dbo:Person
dbo:Person	dbo:subClassOf	dbo:Agent

知识图谱数据模型

□ RDFs

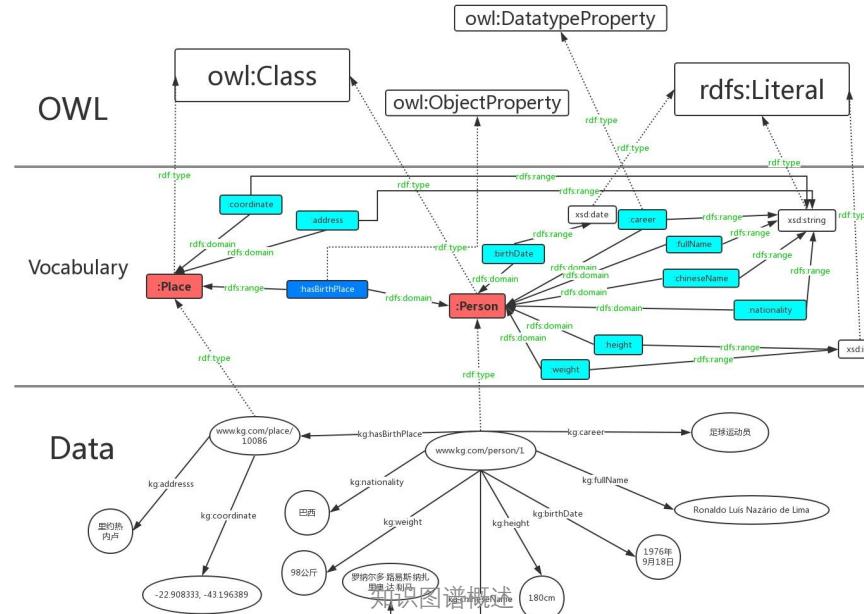
在RDF数据层的基础上引入模式层，定义类、属性、关系、属性的定义域与值域来描述与约束资源，构建**最基本的类层次体系和属性体系**，支持简单的上下位推理。



知识图谱数据模型

口 本体语言 OWL

进一步扩展RDFS词汇，可声明**类间互斥关系、属性的传递性等复杂语义**，支持基于**本体**的自动推理，提供了一组合适web传播的描述逻辑的语法。对机器友好，但认知复杂性限制了工程应用



知识抽取

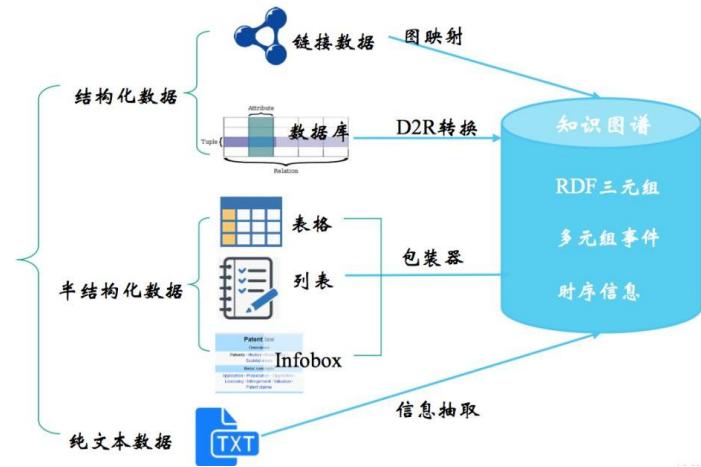
知识抽取

□ 知识获取的目标是从海量的文本数据中通过信息抽取的方式获取知识，其方法根据所处理的数据源的不同而不同。分为：

- 结构化数据
- 半结构化数据
- 非结构化文本数据

□ 文本信息抽取：从非结构化文本数据中进行知识抽取

- 实体识别
- 实体消歧
- 关系抽取
- 事件抽取



知识抽取

Amsterdam	
	
The Keizersgracht at dusk	
Location of Amsterdam	
Coordinates:	52°22'23"N 4°53'32"E
Country	Netherlands
Province	North Holland
Government	
- Type	Municipality
- Mayor	Job Cohen ^[1] (PvdA)
- Aldermen	Lodewijk Asscher Carolin Gehrts Tjeerd Herrema Maarten van Pelgeest Marijke Vos
- Secretary	Erik Gerritsen
Area <small>[2][3]</small>	
- City	219 km ² (84.6 sq mi)
- Land	166 km ² (64.1 sq mi)
- Water	53 km ² (20.5 sq mi)
- Urban	1,003 km ² (387.3 sq mi)
- Metro	1,815 km ² (700.8 sq mi)
Elevation <small>[4]</small>	2 m (7 ft)
Population <small>(1 October 2008)[5][6]</small>	
- City	755,269
- Density	4,459/km ² (11,548.8/sq mi)
- Urban	1,364,422
- Metro	2,158,372
- Demonym	Amsterdamer
Time zone	CET (UTC+1)
- Summer (DST)	CEST (UTC+2)
Postcodes	1011 – 1109
Area code(s)	020
Website: www.amsterdam.nl	

```
@prefix dbpedia <http://dbpedia.org/resource/> .  
@prefix dbterm <http://dbpedia.org/property/> .  
  
dbpedia:Amsterdam  
    dbterm:officialName "Amsterdam" ;  
    dbterm:longd "4" ;  
    dbterm:longm "53" ;  
    dbterm:longs "32" ;  
    dbterm:leaderTitle "Mayor" ;  
    dbterm:leaderName dbpedia:Job_Cohen ;  
    ...  
    dbterm:areaTotalKm "219" ;  
    ...  
  
dbpedia:ABN_AMRO  
    dbterm:location dbpedia:Amsterdam ;  
    ...
```

从Wikipedia
的信息框
(Infor Box)中
进行抽取

知识抽取

Long Tail of Wikipedia

(Intelligence-in-Wikipedia Project) [Wu / Weld: WWW 2008]

YAGO & DBpedia
mappings of
entities onto classes
are valuable assets

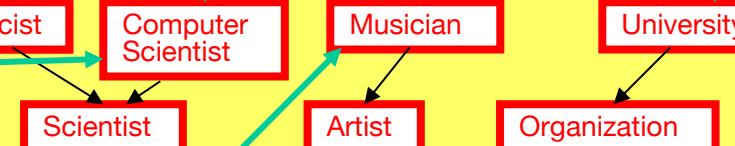
		Born	1944
Born	Apr 1944	Died	lost at sea January 20, 2007
Died	Kiel	Nationality	American
Nationality	Oct 1944	Fields	Computer Science
Fields	Göttingen	Institutions	IBM
Institutions	Phy		Tandem Computers
Alma mater	Univ Univ	Notable awards	DEC Microsoft
	Alma mater		University of California, Berkeley
Alma mater	Kais		Turing Award
	Ludwig-Maximilians-Universität München		

Ricardo Baeza-Yates

From Wikipedia, the free encyclopedia

Ricardo Baeza-Yates (born March 21, 1961) is a Chilean computer scientist and director of the Yahoo! Research labs at Barcelona, Spain and Santiago, Chile. His Ph.D. from the University of Waterloo was entitled *Efficient Text Searching*, supervised by Gaston Gonnet and granted in 1989.

Learning infobox attributes
→ sparse & noisy training data



Frank Zappa

Zappa's interest in composing and arranging proliferated in his last high-school years. By his final year, he was writing, arranging and conducting avant-garde performance pieces for the school orchestra.^[21] He graduated from Antelope Valley High School in 1958, and later acknowledged two of his music teachers on the sleeve of the 1966 album *Freak Out!*^[22] Due to his family's frequent moves, Zappa

从Wikipedia
的文本中进
行抽取

大规模知识抽取

Yago(Yet Another Great Ontology)

2007年，由德国马普研究所发起
融合WordNet和Wikipedia：

- 从Wikipedia的结构中抽取信息
- 利用人工采样评估
- 超过1亿事实和100种关系



Content	Entities of public Interest
Format	TSV,RDF,XML,N3,Web Interface
Sources	Wikipedia, WordNet, Geonames
Main Strength	Focus on Precision, geotemporal annotations,multilingual
Precision	95%
Technique	Extraction from Wikipedia + matching with WordNet & Geonames + consistency checks
Size	Entities: 3 m (+ geonames -> 10m) Facts: 120m (+ geonames -> 460m) Relations: 100, Classes: 200k, Languages: 200
License	Creative Commons BY-SA
URL	http://www.yago.com/
References	[Suchanek, WWW 2007][Hoffart, WWW 2011] [deMelo CIKM 2010]

大规模知识抽取

DBpedia

2007年开放。

目标是构建一个社区，通过社区成员定义和撰写准确的抽取模板，进而从维基百科中抽取结构信息，并将其发布到Web上。

社区通过人工的方式构建分类：

- 280个类别
- 覆盖约50%的维基百科实体



Content	Entities of public Interest
Format	RDF, API, SPARQL
Sources	Wikipedia, WordNet, YAGO
Main Strength	Focus on coverage, interlinking with other data sets
Technique	Extraction from Wikipedia + manual supervision by the community
Size	Entities: 3.5 m (in manual taxonomy: 1.7m) Facts: 670m Attributes: 9k (manually defined: 1k) Manual Classes: 280
License	CC-BY-SA & GNU FDL
URL	http://dbpedia.org
References	[Auer, ISWC 2007], [Bizer09, JWS 2009]]

大规模知识抽取

Freebase

2007年Metaweb公司发布。

2010年被Google收购。

大规模协同构建知识库。

从Wikipedia和其他数据源(如 IMDB、
MusicBrainz)中导入知识

核心思想：

- 在Wikipedia中，人们编辑文章
- 在Freebase中，人们编辑结构化知识

Content	Entities of public Information
Format	RDF, API
Construction	By the community Data import from public sources
Sources	Wikipedia, Libraries, WordNet, MusicBrainz...
Main Strength	Free and large
Size	Facts: several: millions Entities: 20m
License	CC-BY(Creative Commons Attribution)
URL	http://www.freebase.be/



知识融合

知识融合：

- 实体对齐必然涉及到**实体相似度的计算**，假设两个实体的记录 x 和 y ， x 和 y 在第 i 个属性上的值是 x_i, y_i ，那么需要通过两步计算：
 - **属性相似度**：综合单个属性相似度得到属性相似度向量 $[sim(x_1, y_1), sim(x_2, y_2), \dots, sim(x_N, y_N)]$
 - **实体相似度**：根据属性相似度向量得到实体的相似度
- **属性相似度计算方法**：常用的有编辑距离、集合相似度（Jaccard系数、Dice）、向量相似度等
- **实体相似度计算方法**：比如聚合、聚类、表示学习等

知识融合

知识融合方法分类		方法简介	特点
框架/本体 匹配	元素级匹配	<ul style="list-style-type: none">基于字符串：前缀距离、后缀距离、编辑距离基于语言学：利用元素之间的语义关联，如近义词、同根词，利用元素的约束信息，如取值范围，常利用WordNet	只利用元素的直接信息
	结构级匹配	<ul style="list-style-type: none">基于图：转化为发现最大公共子图的问题基于分类体系：类别体系对框架至关重要，只匹配分类关系基于统计分析的匹配：挖掘样本中的规律，对元素进行分组	利用不同元素之间的概念结构关系
实体对齐	成对实体对齐	通过匹配实体属性等特征独立地判断两个实体能否对齐	
	协同实体对齐	不同实体间的对齐相关影响，协调不同对象间的匹配情况	
	基于表示学习	将多个知识库表示在同一语义向量空间中，计算实体相似度	

关系链接举例

实体	类型	实体	类型
格拉斯哥大学	大学	詹姆斯·瓦特	科学家
.....	瓦特 (功率单位)	物理单位
	

校友



格拉斯哥大学

进入词条

基本信息

中文名	格拉斯哥大学	校训	Via Veritas Vita (方法、真理、生命)
外文名	University of Glasgow	地址	University Ave, Glasgow G12 8QQ
简称	格大	主要奖项	2013年英国女王周年纪念奖
创办时间	1451年		2014年泰晤士报年度大学
类别	英国公立大学		2015年先驱报高等教育奖“年度最佳大学”
类型	综合研究型大学		2018年泰晤士报苏格兰年度大学
属性	古典大学	7位诺贝尔奖得主	
罗素大学集团创始成员	Universitas 21创始成员	知名校友	瓦特, 亚当·斯密、开尔文、刘殿爵、蒂姆·库克、杰拉德·巴特勒
校长: Anton Muscatelli		世界排名	151-200 (2019软科世界大学学术排名) [6]
校监: Aamer Anwar		世界排名	99 (2020泰晤士高等教育世界大学排名) [7]

知识图谱与自然语言处理

自然语言处理和知识图谱研究是双向互动的关系。



知识图谱与自然语言处理

信息抽取

主要技术：

实体识别与抽取、实体消歧、关系抽取

趋势及挑战：

- 从封闭走向开放
- 大规模信息抽取
- 深层次挖掘信息背后的语义（从抽取到理解）



语义解析

语义解析就是将自然语言映射成机器可以表达的形式。

主要技术：

词义消歧、语义角色标注、指代消解等。

应用：

- 面向知识图谱的自然语言问答
- 聊天机器人等

信息抽取

实体识别：从文本中识别出实体的命名性指称项，并标明其类别

- 三大类：实体类、时间类、数字类
- 七小类：人名、机构名、地名、时间、日期、货币和百分比

□ 实体识别任务的产生：

- 命名实体形式多变：如姚明、小巨人、姚主席、明王都是指同一个人
- 命名实体的语言环境复杂：如彩霞在某些条件下是人名，在另外的条件下可能是自然现象

□ 命名实体识别的方法：

- 基于规则的实体识别方法
- 基于机器学习的实体识别方法

信息抽取

实体识别：

□ 命名实体识别的方法：

➤ 基于规则的实体识别方法

基于命名实体词典的方法：采用字符串完全匹配或部分匹配的方式，从文本中找出与词典最相似的短语完成实体识别

例：中文人名的识别规则示例： $<\text{姓氏}><\text{名字}>$ ，例如：姚明

中文地名的识别规则示例： $<\text{名字部分}><\text{指示部分}>$ ，例如：北京市

优点：规则简单

缺点：需要构建词典和规则；性能受词典规模和质量的影响

信息抽取

实体识别：

□ 命名实体识别的方法：

➤ 基于机器学习的实体识别方法

利用预先标注好的语料训练模型，使模型学习到某个字或词作为命名实体组成部分的概率，进而计算一个候选字段作为命名实体的概率值。若大于某一阈值，则识别为命名实体。

分为：最大熵模型(Maximum Entropy Model) 和条件随机场模型(Conditional Markov Random Field)

格	拉	斯	哥	大	学	位	于	苏	格	兰
B	I	I	I	I	I	O	O	B	I	I

信息抽取

关系抽取：自动识别实体之间具有的某种语义关系

根据抽取文本的范围不同，分为

- 句子级关系抽取
- 语料（篇级）关系抽取

□ 关系抽取任务的难点：

- 同一个关系可以具有多种不同的词汇表示方式
- 同一个短语或词可能表达不同的关系
- 同一对实体之间可能存在不止一种关系
- 需要结合上下文
- 关系有时在文本中找不到任何明确表示，隐含在文本中
- 关系抽取依赖词法、句法分析等基本的自然语言处理工具，但该工具性能并不高

The diagram shows two sentences with entities highlighted in red, green, blue, and orange, and their corresponding extracted relations and arguments.

Top sentence: 巨星刘德华携手巩俐等人气明星打造的都市爱情大片《我知女人心》在博纳悠唐国际影城正式首映。

Bottom sentence: 哈尔滨工业大学校长王树国荣获法国荣誉勋章。

Legend:

- PER (Red)
- LOC (Green)
- ORG (Blue)
- MISC (Orange)

Extracted relations and arguments:

Arg1	Arg2	Relation
刘德华	巩俐	携手
刘德华	《我知女人心》	打造
巩俐	《我知女人心》	打造
《我知女人心》	博纳悠唐国际影城	首映

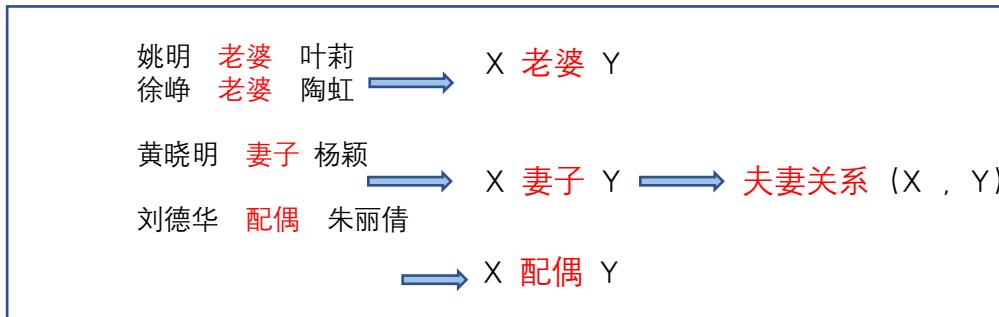
Arg1	Arg2	Relation
王树国	哈尔滨工业大学	校长
王树国	法国荣誉勋章	荣获

信息抽取

关系抽取

关系抽取的方法

➤ 基于模板的方法



基于触发词/字符串

于依存句法

➤ 基于机器学习的方法

董卿现身国家博物馆看展优雅端庄大方

依存分析结果

词顺序	词	词性	依存关系路径	依存关系
0	董卿	人名	1	定语
1	现身	动词	-1 Root 一般是谓语动词	核心词
2	国家博物馆	地名	1	宾语
3	看	动词	1	顺承
4	展	动词	3	补语
5	优雅	形容词	7	定语
6	端庄	形容词	7	定语
7	大方	形容词	4	宾语

规则抽取结果

(董卿, 现身, 国家博物馆) → 位于(董卿, 国家博物馆)

基

信息抽取

事件抽取

- **事件**：发生在某个特定的时间点或时间段、某个特定的地域范围内，由一个或多个角色参与的一个或多个动作组成的事情或者状态的改变。
- **要素**：事件发生的时间、地点、参与事件的角色、与之相关的动作或状态的改变

- **事件抽取**：从描述事件的文本中抽取出用户感兴趣的事件信息并以结构化的形式呈现出来
- **相关概念**：事件指称、事件触发词、事件元素、元素角色、事件类别

信息抽取

事件抽取

事件抽取任务的基础工作：

- 识别事件触发词及事件类型
- 抽取事件元素 (Event Argument)

同时判断其角色 (Argument Role)

- 抽出描述事件的词组或句子

mention trigger

苹果公司将于西部时间9月12日上午10点(北京时间9月13日凌晨1点)举行新品发布会，这一次的发布会地点是全新建造的史蒂夫·乔布斯剧院。根据目前的消息，这次发布会上苹果将会发布iPhone 8(命名不确定，暂且称之为iPhone 8)、Phone 7s、iPhone 7s Plus、Apple Watch 3以及全新Apple TV。

Slot filling

事件类型	发布会
公司	苹果公司
时间	西部时间9月12日上午10点
地点	史蒂夫·乔布斯剧院
产品	iPhone8、iPhone7s、iPhone7s plus、apple watch 3、apple TV

Argument role
Predefined

arguments

语义解析之语义搜索

语义搜索

- 是指搜索引擎的工作不再拘泥于用户所输入请求语句的字面本身，而是透过现象看本质，准确地捕捉到用户所输入语句后面的真正意图，并以此来进行搜索，从而更准确地向用户返回最符合其需求的搜索结果。

语义搜索过程

输入的问句进行解析，找出问句中的实体和关系，理解用户问句的含义。

将用户在知识图谱中匹配查询语句，找出答案

通过一定的形式将结果呈现到用户面前

莱昂纳多是什么时候出生的?

全部 新闻 图片 视频 地图 更多

找到约 3,160,000 条结果 (用时 0.59 秒)

莱昂纳多·迪卡普里奥 / 出生日期

1974 年 11 月 11 日 (44 岁)

用户还搜索了

凯特·温斯莱特 布拉德·皮特 约翰尼·德普

生于：1974 年 11 月 11 日 (44 岁)，加利福尼亚州洛杉矶好莱坞
身高：6'0"

所获奖项：奥斯卡最佳男主角奖，美国演员工会奖最佳男主角，更多
即将上映的电影：好莱坞往事
家长：乔治·迪卡普里奥，艾莫琳·英登比尔肯
提名：奥斯卡最佳影片奖，奥斯卡最佳男配角奖，更多

电影

莱昂纳多·迪卡普里奥_百度百科
<https://baike.baidu.com/item/莱昂纳多·迪卡普里奥>

莱昂纳多·迪卡普里奥 (Leonardo DiCaprio)，1974年11月11日出生于美国加利福尼亚州洛杉矶，美国... 2010年的《禁闭岛》是莱昂纳多与马丁·斯科塞斯的第四次合作。... 莱昂纳多一出生，他的父母便已

还有45+项

语义解析之知识问答

智能问答的方法

1. 基于信息检索的方法

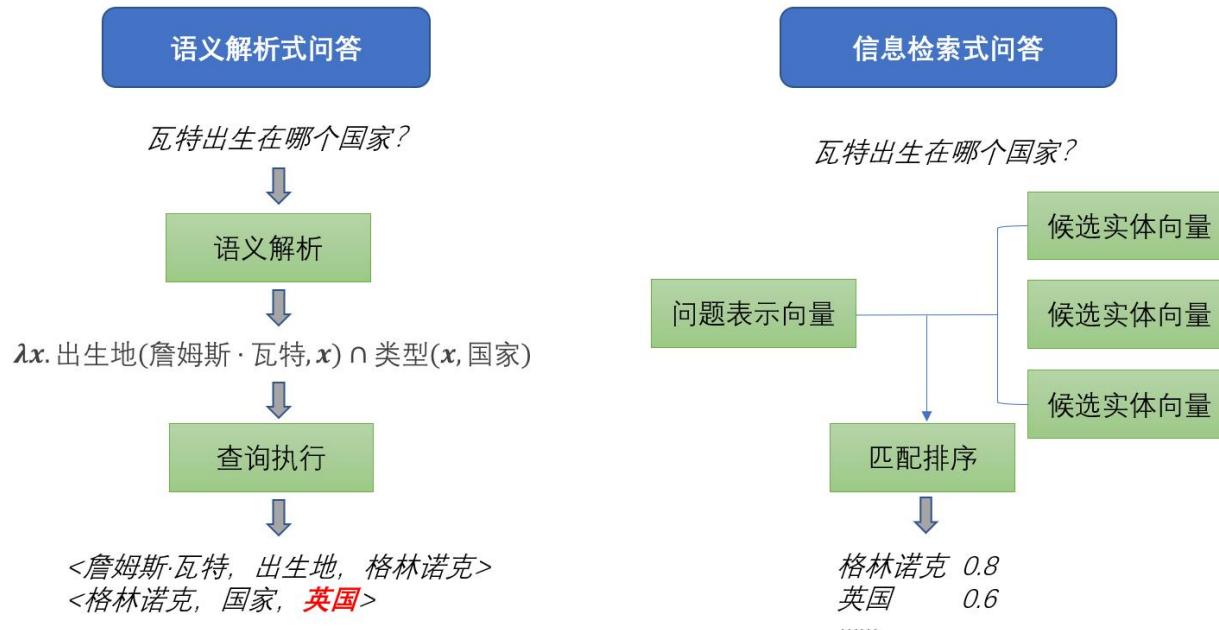
- 首先利用中文分词、命名实体识别等自然语言处理工具找到问句中所涉及到的实体和关键词，然后去知识资源库中去进行检索，并通过打分模型对答案进行排序

2. 基于语义解析的方法

- 将一个自然语言形式的问句，按照特定语言的语法规则，解析成语义表达式，将其转化为某种数据库的查询语言

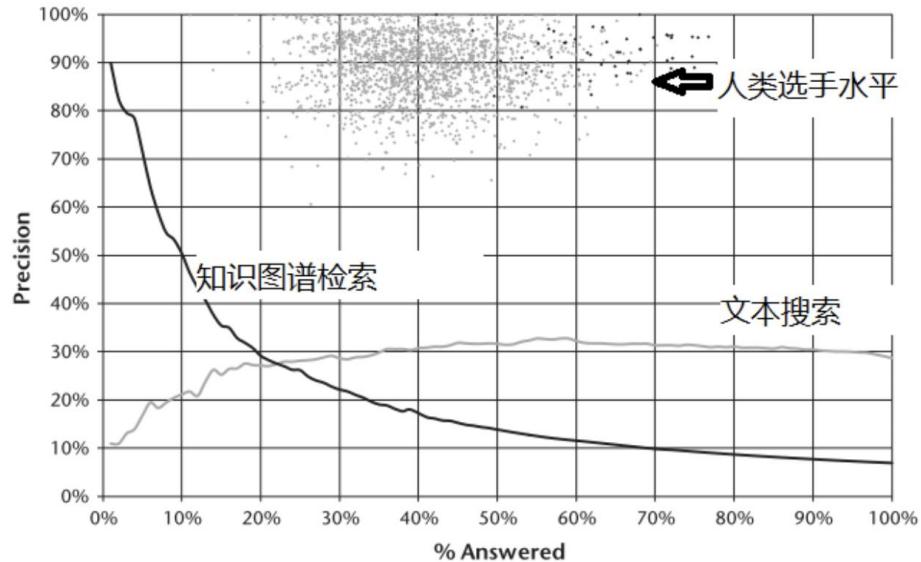
语义解析之知识问答

知识图谱问答的两种主要方法框架对比



IBM Watson系统

Jeopardy “危险边缘”
2011 IBM Watson击败人类冠军



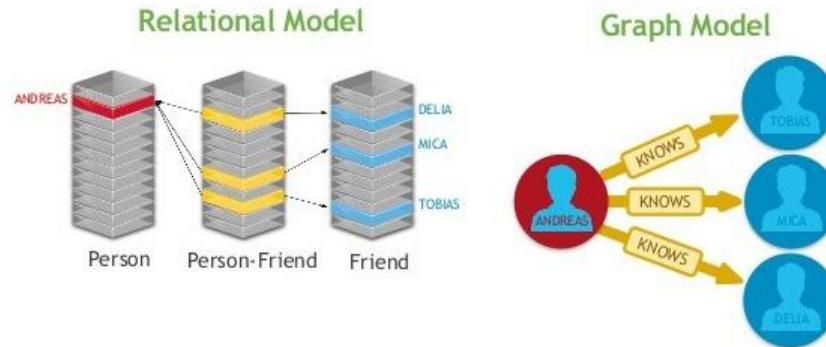
知识图谱与数据管理

知识图谱本质上是多关系图，通常用“**实体**”来表达图里的**结点**、用“**关系**”来表达图里的**边**。

关系型数据库：实体与实体之间的关系通常都是利用外键来实现，对关系的查询需要大量join操作



图数据库：图模型建模实体（结点）和实体之间的关系（边），在对关系的操作上有更高的性能



知识图谱概述

基于关系的知识图谱存储管理

三元组表：主谓宾三列的表

- 优点：简单明了
- 缺点：最大问题在于将知识图谱查询翻译为 SQL 查询后会产生三元组表的大量自连接操作

SPARQL

组成部分	示例
NS	prefix foaf: <http://xmlns.com/foaf/0.1/>
RDF Dataset	FROM NAMED <http://example.org/foaf/bobFoaf>
QF	SELECT ?name ?bd
GP	WHERE { ?p name ?name . ?p birthDate ?bd }
SM	ORDER BY ?X

SQL

```
SELECT T1.object T2.object  
FROM T as T1, T as T2  
WHERE T1.subject=T2.subject  
and T1.predicate=name and  
T2.predicate=birthdate
```

主体	属性	客体
dbr:James_Watt	rdfs:label	"James Watt"@en
dbr:James_Watt	dbo:birthDate	"1736-01-19"^^xsd:date
dbr:James_Watt	dbo:birthPlace	dbr:Greenock
dbr:James_Watt	rdf:type	dbo:Scientist
dbr:James_Watt	dbo:influencedBy	dbr:Joseph_Black
dbr:James_Watt	dbo:influencedBy	dbr:Adam_Smith
dbr:James_Watt	dbp:workplaces	dbr:University_of_Glasgow
dbr:Adam_Smith	rdfs:label	"Adam Smith"@en
dbr:Adam_Smith	dbo:birthDate	"1723-06-16"^^xsd:date
dbr:Adam_Smith	dbo:almaMater	dbr:University_of_Glasgow
dbr:Adam_Smith	dbo:influencedBy	dbr:Aristotle
dbr:Joseph_Black	rdfs:label	"Joseph Black"@en
dbr:University_of_Glasgow	name	"University Of Glasgow"@en
dbr:Aristotle	rdfs:label	"Aristotle"@en
dbr:Aristotle	dbo:birthDate	"-383-1-1"^^xsd:date
dbr:Greenock	dbo:country	dbr:United_Kingdom
dbr:Greenock	dbo:type	dbo:Place
dbr:United_Kingdom	dbo:type	dbo:Place
dbo:Scientist	dbo:subClassOf	dbo:Person
dbo:Person	dbo:subClassOf	dbo:Agent

基于关系的知识图谱存储管理

属性表：属性相似的聚为一张表

- 优点：克服三元组自连接的问题
- 缺点：一对多联系或多值属性存储问题、RDF的灵活性等
- 代表：采用属性表存储方案的代表系统是 RDF 三元组库 **Jena**

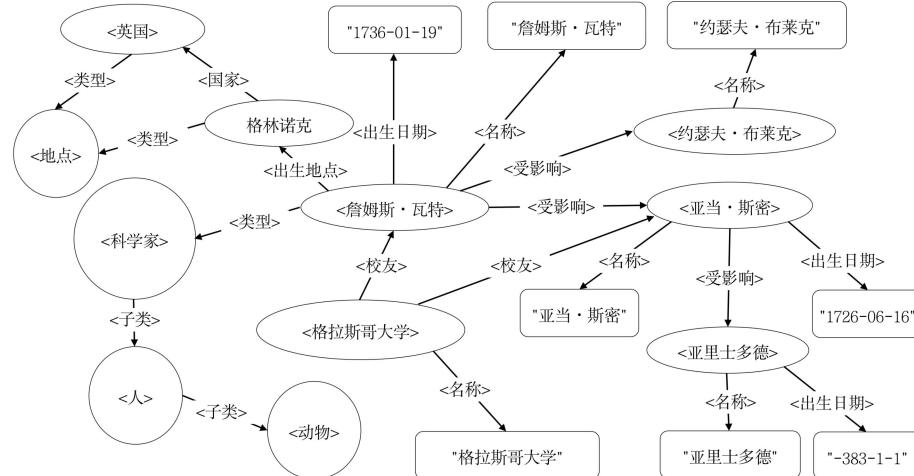
垂直划分：以谓语划分三元组表

- 优点：克服属性表的空值多值问题
- 缺点：大量属性表、删除代价大
- 代表：采用垂直划分存储方案的代表数据库是 **SW-Store**

原生知识图谱存储管理--RDF

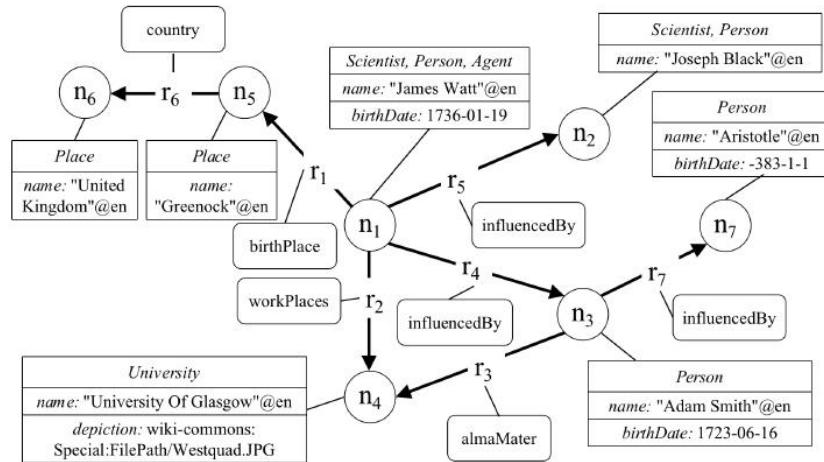
- RDF数据以及其上的结构化查询都可以视为图

回答RDF数据上
SPARQL查询 == 子图
匹配



- 代表性系统：利用子图匹配回答面向RDF知识图谱的
SPARQL查询，gStore系统

原生知识图谱存储管理--属性图



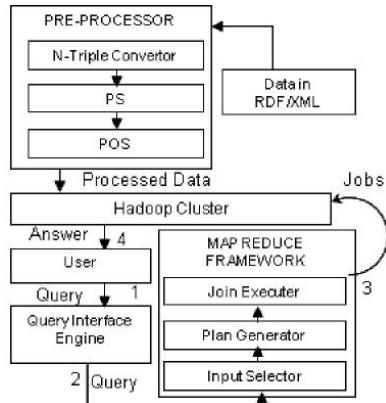
- t₁** **MATCH** (r:Person)
- t₂** **OPTIONAL MATCH** (r) - [:birthPlace] -> (pl:Place)
- t₃** **WITH** r, pl AS birthPlace
- t₄** **MATCH** (r) - [:influncedBy*] -> (p:Person)
- t₅** **RETURN** r.name, birthPlace.name, count(p) AS influenceNum

分布式知识图谱存储管理

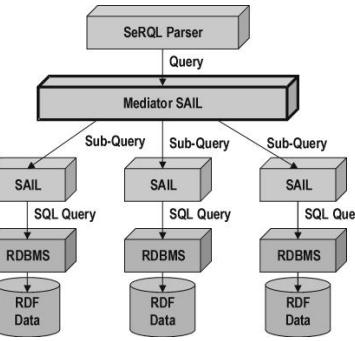
方法	简介	优点	缺点
基于云计算平台的分布式RDF数据管理方法	利用已有云平台进行RDF数据的管理	<ul style="list-style-type: none">有很好的扩展性与容错性	<ul style="list-style-type: none">多是面向离线数据分析查询处理的效率仍是挑战
基于数据划分的分布式RDF数据管理方法	首先将RDF数据图划分成若干子图,然后将这些子图分配到不同计算节点上	<ul style="list-style-type: none">按照自身的算法设计进行RDF数据的划分与分配可减少查询处理阶段的通信代价	<ul style="list-style-type: none">很多实际应用中,RDF数据不能由系统任意划分对于这些不能指定数据划分的应用,有局限性
联邦式系统	对全局RDF数据进行关联.这种联邦式的RDF数据管理系统并不侵犯数据拥有方对于数据的管理权	<ul style="list-style-type: none">局部匹配能保证计算过程中涉及到的边和点是最少的	<ul style="list-style-type: none">将所有局部匹配收集到一台机器上连接得到最终匹配对查询效率以及连接操作是一个挑战

分布式知识图谱存储管理框架

➤ 基于云计算平台的系统框架图

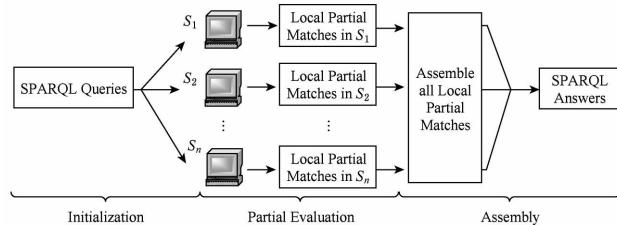


➤ 基于数据划分的系统框架图



- 依赖于现有的云计算系统,例如Hadoop等
- 需要将RDF数据归结到云平台支持的存储的文件格式

➤ 联邦式系统框架图



知识图谱概述

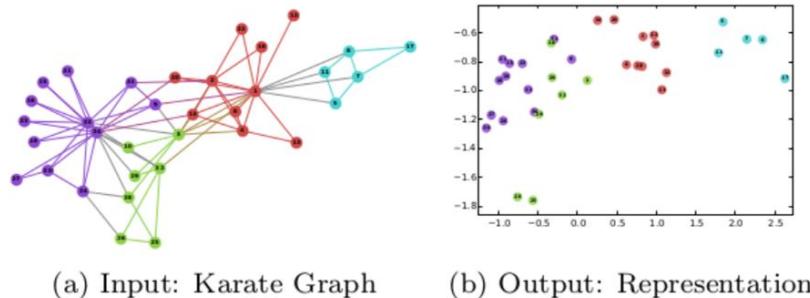
- 主要讨论数据如何划分
- 划分以后的数据分别物理地存储到各自的机器上

- 整体的RDF数据由多个独立节点上的局部数据集成得到
- 联邦系统可以回答针对整体RDF数据集的查询.

知识图谱与机器学习

知识表示学习：

- **背景**：基于网络形式的知识表示存在数据稀疏问题和计算效率问题。
- **知识表示学习 (representation learning)**：主要是面向知识图谱中的实体和关系进行表示学习，使用建模方法将实体和向量表示在低维稠密向量空间中，然后进行计算和推理。
- **优点**：显著提升计算效率，有效缓解数据稀疏，实现异质信息融合。
- **应用**：知识图谱补全、相似度计算、关系抽取、自动问答、实体链指

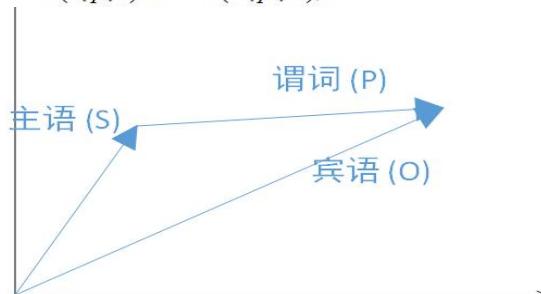


知识图谱与机器学习

知识表示学习：

- **举例：** 知识表示代表模型：TransE [Bordes et al., NIPS 13]。
- 对每个事实(Subject, Predicate, Object)，将其中的predicate作为从subject到object的翻译操作。
- 每个Subject/Predicate/Object，都映射成一个多维向量
- **优化目标：** $S+P=O$

$$\Gamma = \sum_{(s,p,o) \in s} \sum_{(s',p',o') \notin s} [r + d(s+p,o) - d(s'+p',o')]_+$$

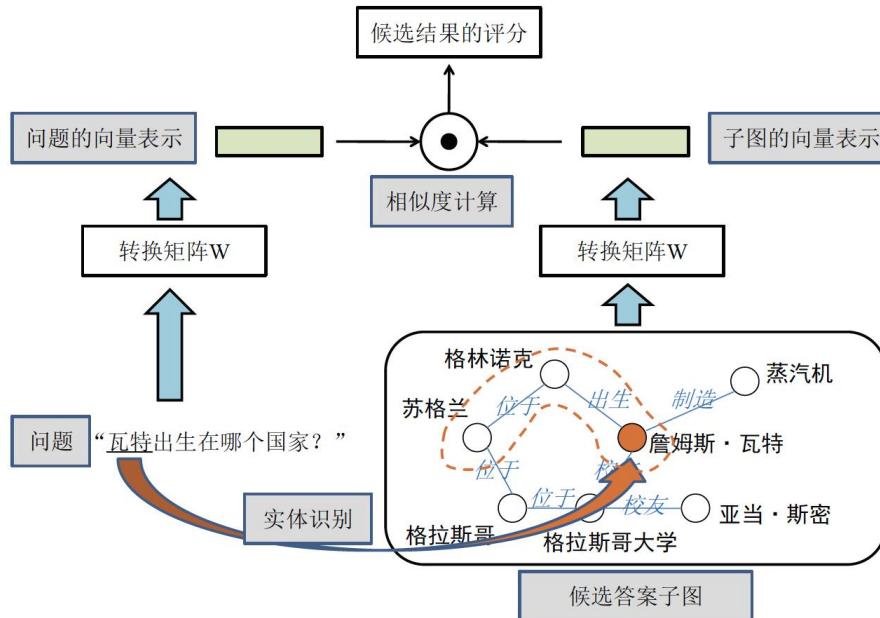


$$\begin{aligned} \text{Beijing} - \text{China} \\ \approx \\ \text{Ottawa} - \text{Canada} \end{aligned} \quad = \text{Capital}$$

S	P	O
China	Capital	Beijing
Canada	Capital	Ottawa
.....

知识图谱与机器学习

图表示学习用于“自然语言问答”：



Part 3

从人工智能、大数据的角度看待
“知识图谱”

1950–1970：人工智能诞生

1956年达特茅斯会议，提出“人工智能(Artificial Intelligence, AI)”概念。

“用机器来模仿人类学习以及其他方面的智能”

“上古”流派：

- 符号主义 (Symbolism)
- 连接主义 (Connectionism)



“人工智能来了，再过十年机器就要超越人类了！”

1956, 达特茅斯学院
知识图谱概述

符号主义

符号主义(symbolicism)，又称为逻辑主义(logicism)、心理学派(psychologism)或计算机学派(computerism)，其主要原理为认知过程就是在符号表示上的一种运算。

知识图谱起源于符号主义

- 代表人物：



Allen Newell



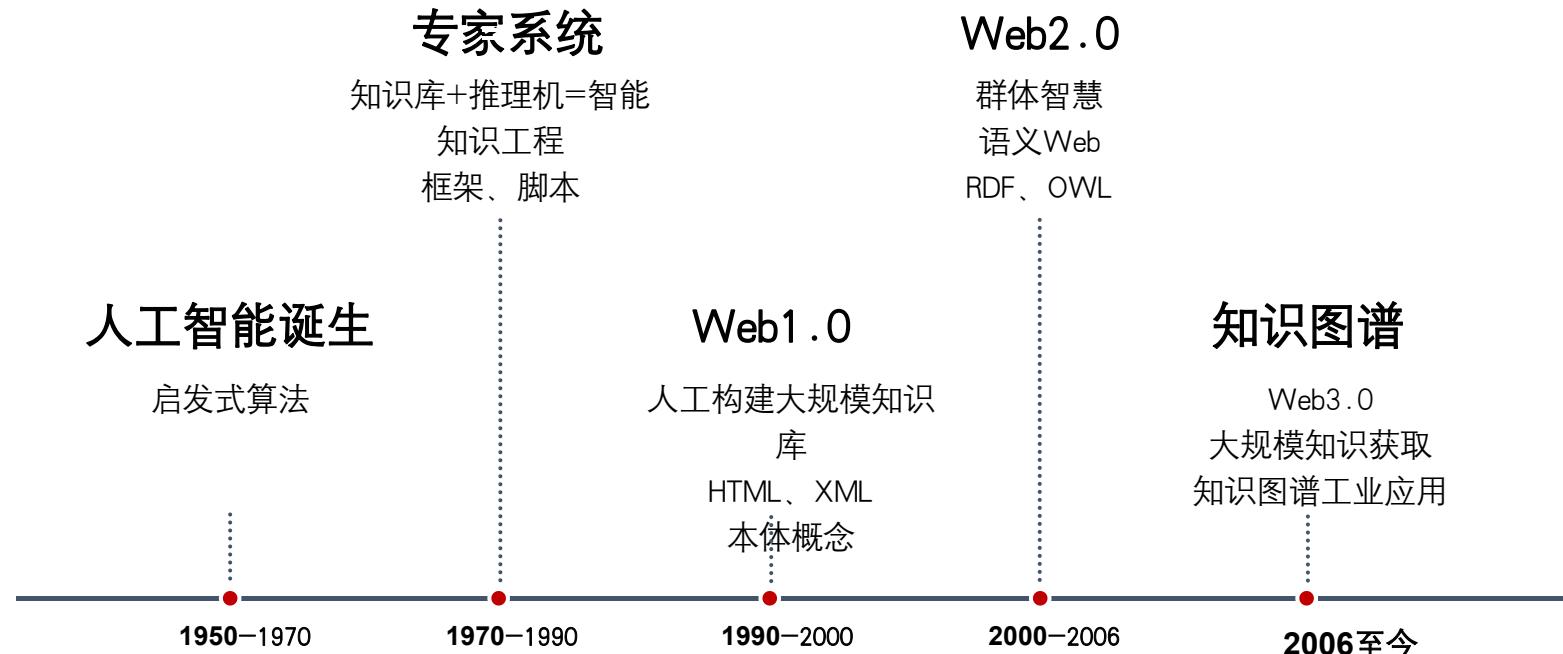
Herbert A. Simon

联合创造“The General Problem Solver”(通用问题求解程序)

小明认识自行车O：
 $O(a,b,c,d,e)$



符号主义发展历史



连接主义

连接主义(connectionism)，又称为仿生学派(bionicsism)或生理学派(physiologism)，其主要原理为智能活动是由大量简单的单元通过复杂的相互连接后并行运行的结果。

当前典型研究：深度学习、深度神经网络

- 代表人物：



Frank Rosenblatt, 提出感知器 (1957)



John Hopfield, 提出Hopfield神经网络 (1982)

小明学骑自行车：
经过长时间练习，小明终于学会了！
却说不清楚“到底该怎样”骑？



1950–1970：人工智能诞生

- 计算机有限的内存和处理速度
- 计算难度指数级增长
- 常识与推理（莫拉维克悖论）

70年代后期

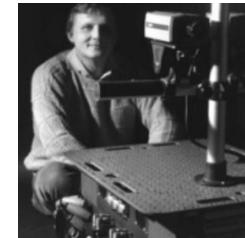


人工智能发展遭遇瓶颈，出现了第一次低谷。

莫拉维克悖论：

实现人类独有的高阶智慧只需要非常少的计算能力，但是实现无意识的技能和感知却需要极大的运算能力。——困难的问题易解，简单的问题难解

“要让电脑如成人般地下棋是相对容易的，但是要让电脑有如一岁小孩般的感知和行动能力却是相当困难，甚至是不可能的。”



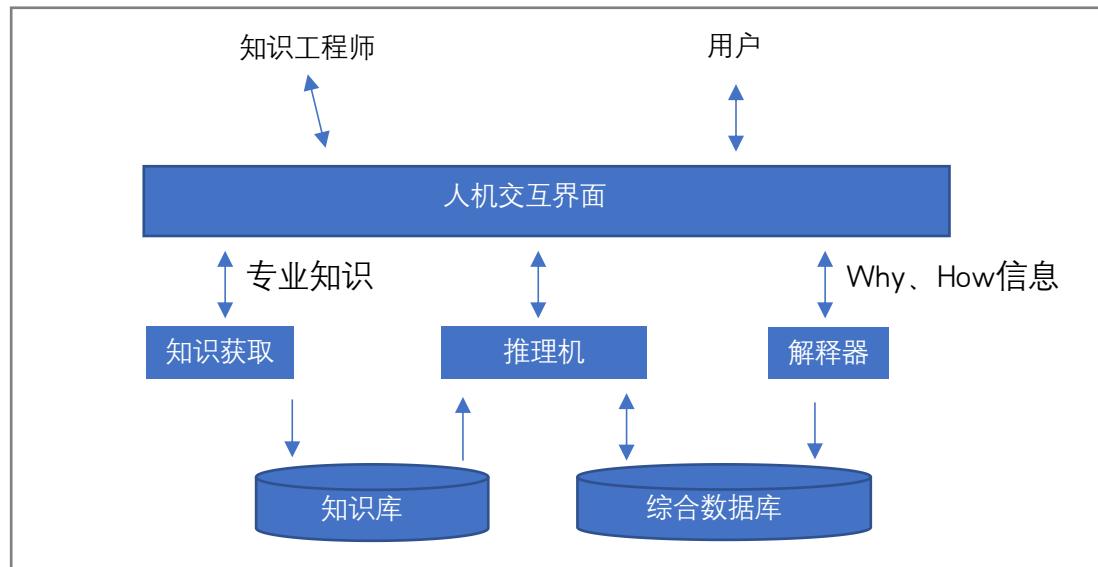
Hans Moravik

1970–1990：专家系统

人工智能开始转向建立基于知识的系统，**通用领域**—>**限定领域**

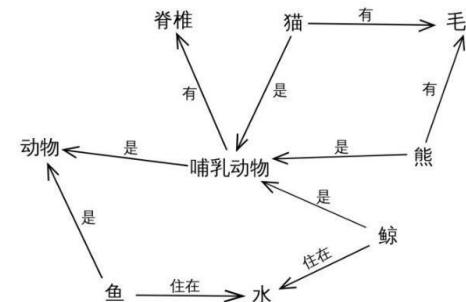
通过**知识库+推理机**实现智能。

专家系统模型：



1970–1990：专家系统

- 语义网络 (Semantic Network): 1970年, Herbert A. Simon正式提出, 通过有向图来表示知识, 作为知识表示的一种通用手段。



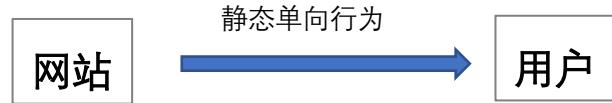
- 知识工程 (Knowledge Engineering): 1977年美国斯坦福大学计算机科学家Edward Albert Feigenbaum教授在第五届国际人工智能会议上提出。确立了知识工程在人工智能中的核心地位。



Edward Albert Feigenbaum
1994年图灵奖得主

1990–2000：万维网 Web 1.0

Web 1.0：文档互联



1989年，英国科学家Tim Berners-Lee发明了万维网(World Wide Web)。

1994年，万维网联盟 (World Wide Web Consortium, W3C) 创建，
是Web技术领域最具权威和影响力的国际中立性技术标准机构。

发布互联网内容标记语言：HTML (1997) , XML (1998)

为互联网环境下大规模知识表示和共享奠定了基础。

1998年，PageRank搜索引擎技术被发明，谷歌 (Google) 成立。



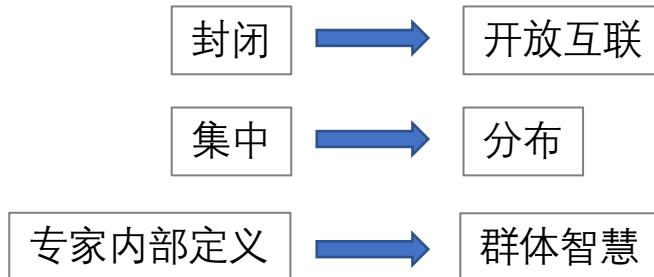
谷歌创始人：
Lawrence Edward Page (左)
Sergey Brin (右)

2000–2006：群体智慧Web2.0

Web2.0：数据互联

强调用户生成内容，易用性，参与文化和终端用户互操作性。

互联网知识：



Wikipedia

开放的在线多语言百科全书

- 2001年开始
- 以众包（crowdsourcing）的方式构建
- 主要特点
 - 数据源质量高
 - 500万概念
 - 富含丰富语义结构的文档：
 - Infobox
 - Table
 - List
 - category



WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file

Article Talk

Not logged in Talk Contributions Create account Log in

Read Edit View history

Search Wikipedia



Ontology (information science)

From Wikipedia, the free encyclopedia

"Knowledge graph" redirects here. For the Google knowledge base, see Knowledge Graph. For other uses, see Knowledge engine (disambiguation).

This article is about ontology in information science. For the study of the nature of being, see Ontology.

In computer science and information science, an **ontology** encompasses a representation, formal naming and definition of the categories, properties and **relations** between the concepts, data and entities that substantiate one, many or all domains.

Every **field** creates ontologies to limit complexity and organize information into **data** and **knowledge**. As new ontologies are made, their use hopefully improves **problem solving** within that domain. Translating **research papers** within every field is a problem made easier when **experts** from different countries maintain a **controlled vocabulary** of **jargon** between each of their languages.^[1]

Since **Google** started an initiative called **Knowledge Graph**, a substantial amount of research has used the phrase **knowledge graph** as a generalized term. Although there is no clear definition for the term **knowledge graph**, it is sometimes used as synonym for **ontology**.^[2] One common interpretation is that a **knowledge graph** represents a collection of interlinked descriptions of entities – real-world objects, events, situations or abstract concepts.^[3] Unlike **ontologies**,

Information science

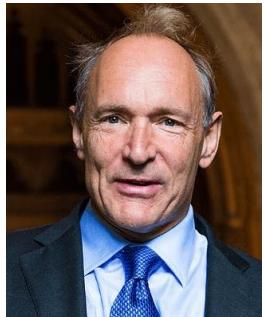
General aspects
Information access •
Information architecture
Information management
Information retrieval
Information seeking • Information society
Knowledge organization • **Ontology** •
Taxonomy
Philosophy of information
Science and technology studies

Related fields and sub-fields

Bibliometrics • Categorization
Censorship • Classification
Computer data storage • Cultural studies

成为大规模构建知识图谱的重要数据基础。

语义Web



Tim Berners-Lee
2016年图灵奖得主
万维网、语义网之父

Tim Berners-Lee于2000年提出语义Web：

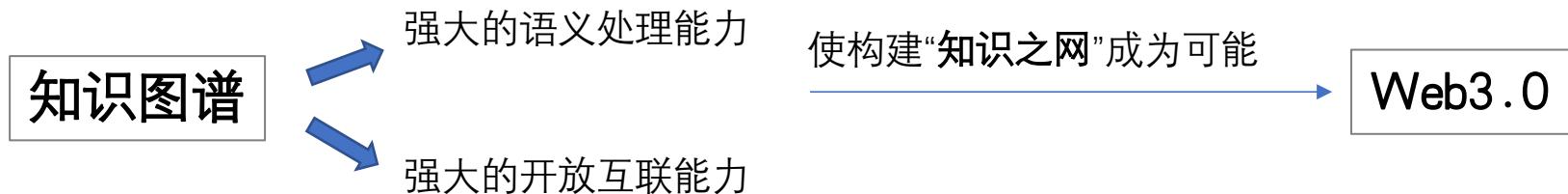
"I have a dream for the Web [in which computers] become capable of analyzing all the data on the Web – the content, links, and transactions between people and computers. A "Semantic Web", which makes this possible, has yet to emerge, but when it does, the day-to-day mechanisms of trade, bureaucracy and our daily lives will be handled by machines talking to machines. The "intelligent agents" people have touted for ages will finally materialize."

"我有一个梦想，网络中的所有计算机能够分析网络中的数据，包括内容、链接、人与计算机之间的往来。语义Web会让这一切成为可能，一旦该网络出现，日常的交易机制、事务以及我们的日常生活都会由机器与机器之间的沟通来处理。人们吹嘘多年的“智能代理”将最终实现。”

2006至今：知识图谱

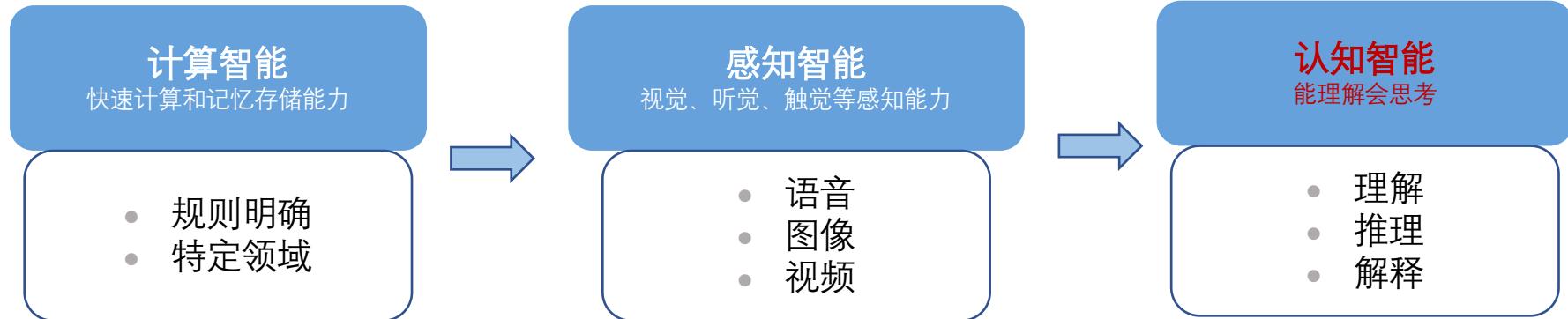
Web3.0：知识互联

构建人与机器都可理解的万维网，使网络更加智能化。



知识图谱与人工智能

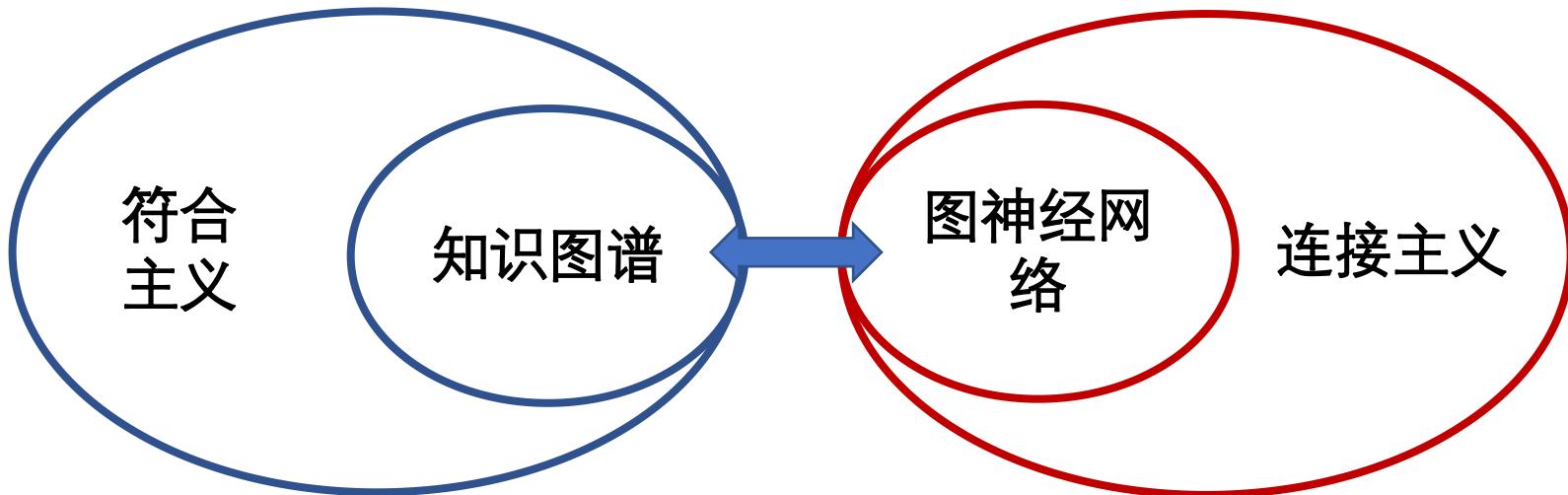
人工智能需要机器智能，特别是认知智能，认知智能依赖知识图谱



- 认知智能的理解、推理、解释任务不仅需要数据，更需要**知识背景**
- 知识图谱是**知识的图谱表示**，这种知识表示方式适合理解、推理、解释
- 知识图谱是实现**认知智能**的关键技术，是实现机器认知智能的使能器

知识图谱与人工智能

知识图谱脱胎于符号主义，但是和连接主义的结合是目前的重要研究方向（例如知识图谱的表示学习等）。

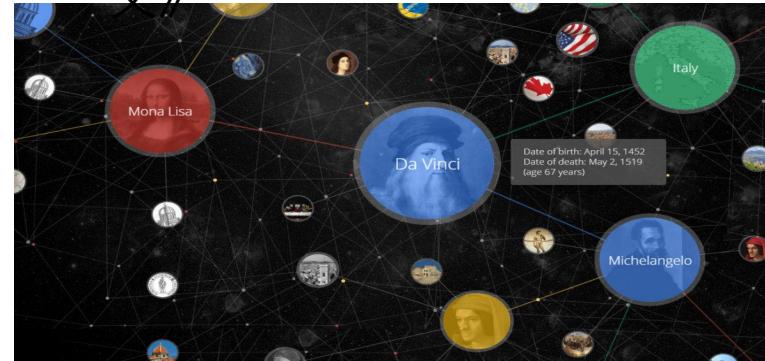


知识图谱与大数据

大数据的特点：

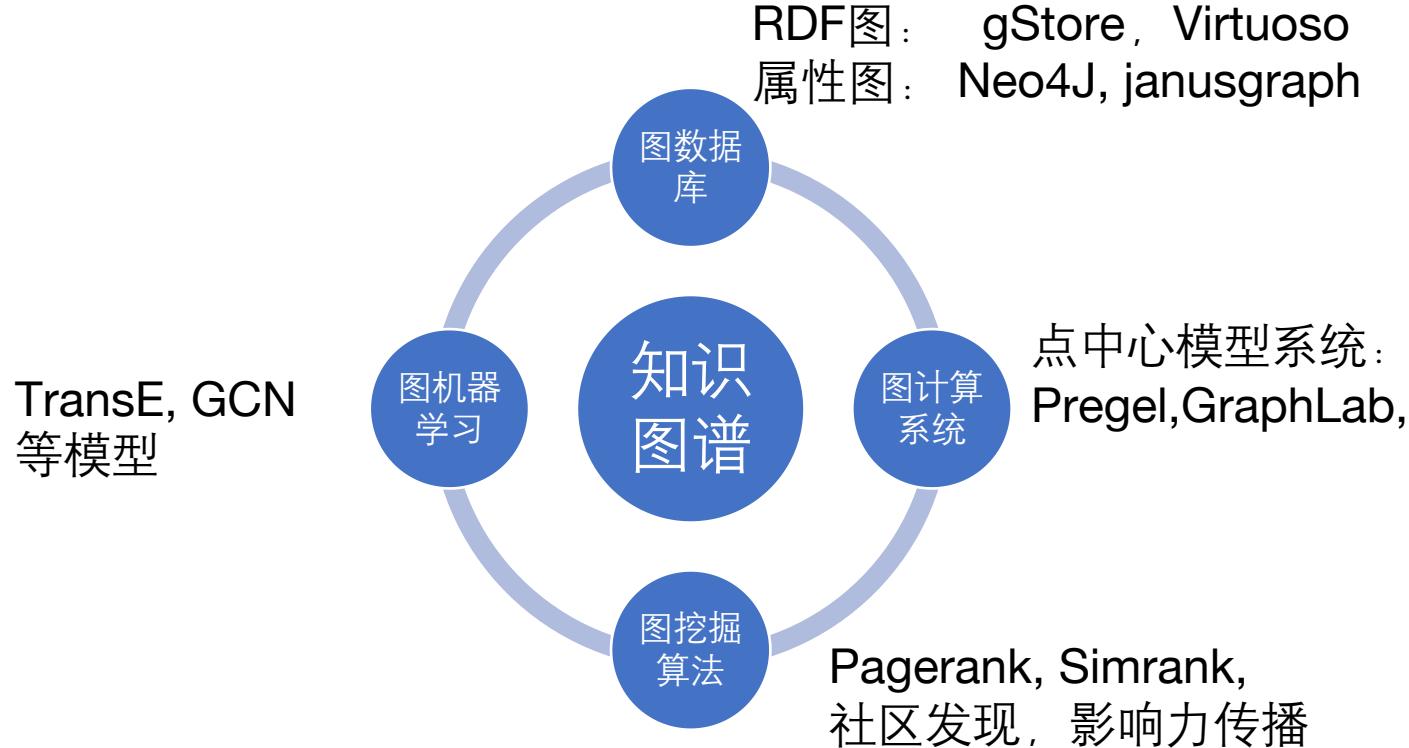
多样化 (variety)
规模大 (volume)
速度化 (velocity)

“世界是普遍联系的整体，任何事物之间都是相互联系的”----- 马克思《辩证唯物主义》



“知识图谱”是面向关联分析的大数据模型

知识图谱与大数据



Part 4

知识图谱的项目应用

知识图谱的行业应用



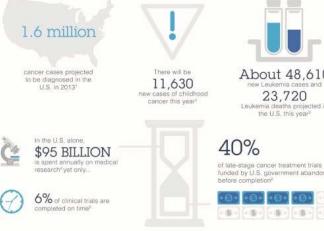
情报分析

股票问问



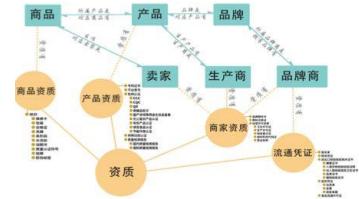
金融证券

Going Up Against a Deadly Disease



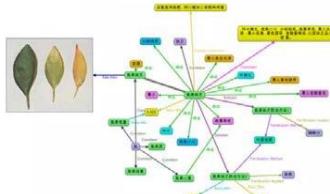
医疗

知识库顶层设计



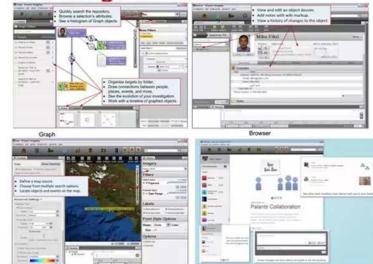
电商

Complete Ontology (knowledge) for Lacking Nitrogen



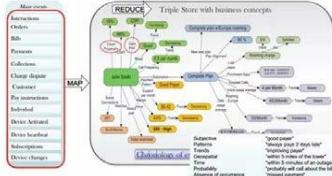
农业

Palantir government

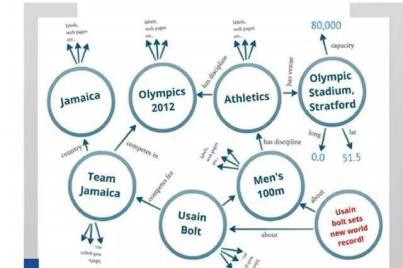


政府

Turn Massive Raw Data Into Business Concepts

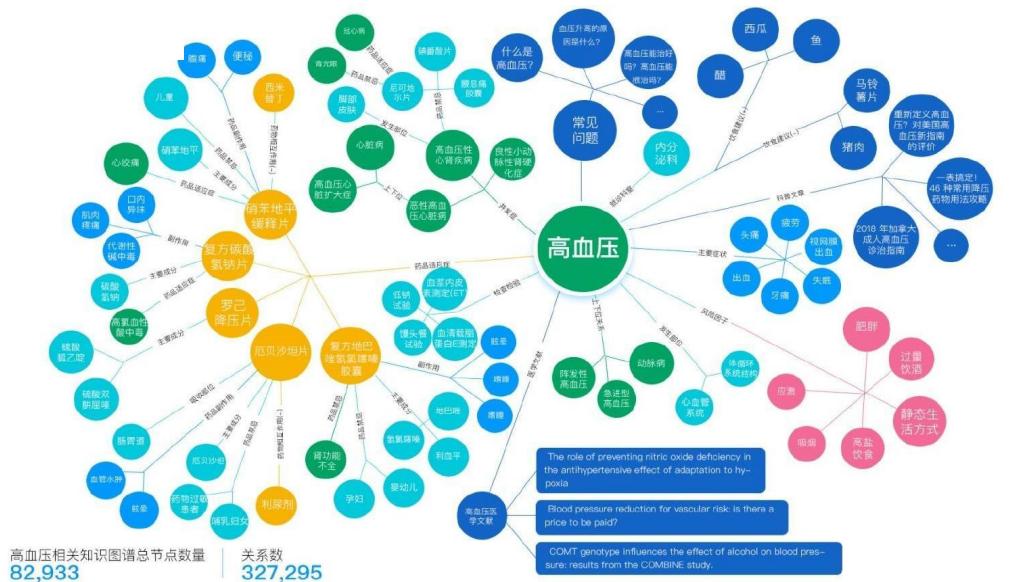


电信



出版

医疗领域知识图谱



高血压知识图谱示例

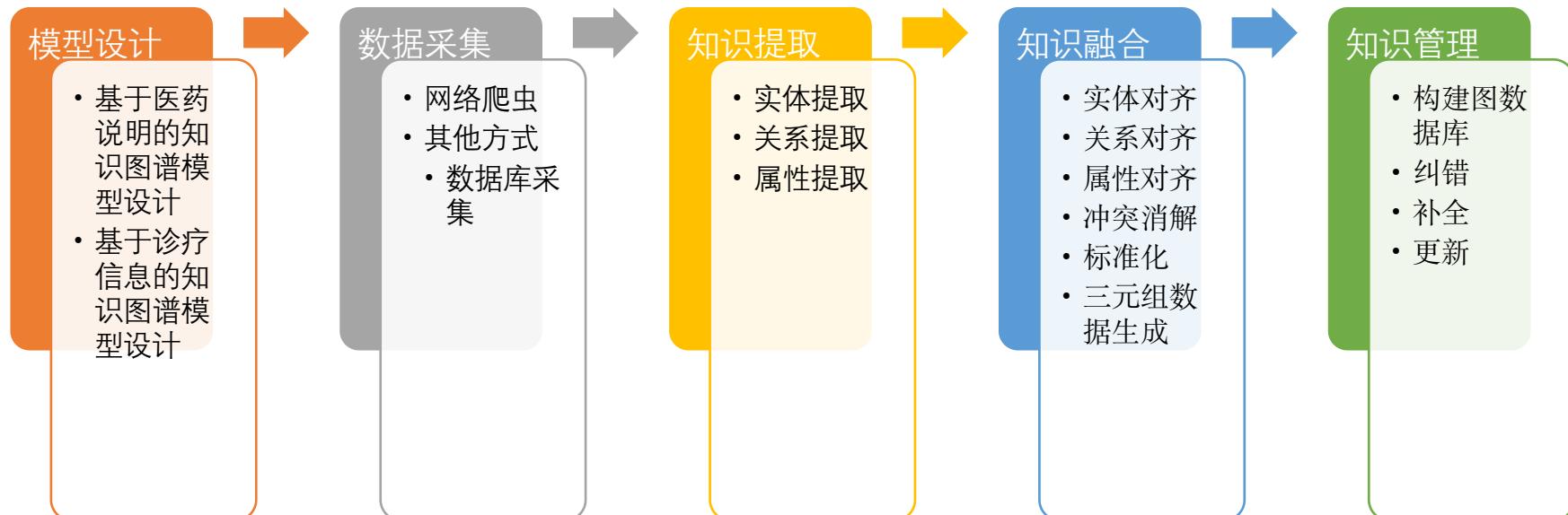
医疗知识图谱的主要应用

- 智能问诊/分诊服务
- 智能影像辅助诊断
- 智能疾病辅助诊疗
- 精准用药推荐
- 智能患者教育/ 随访
- 智能临床科研平台
- ...

难点：非结构化的电子病历理解成本大

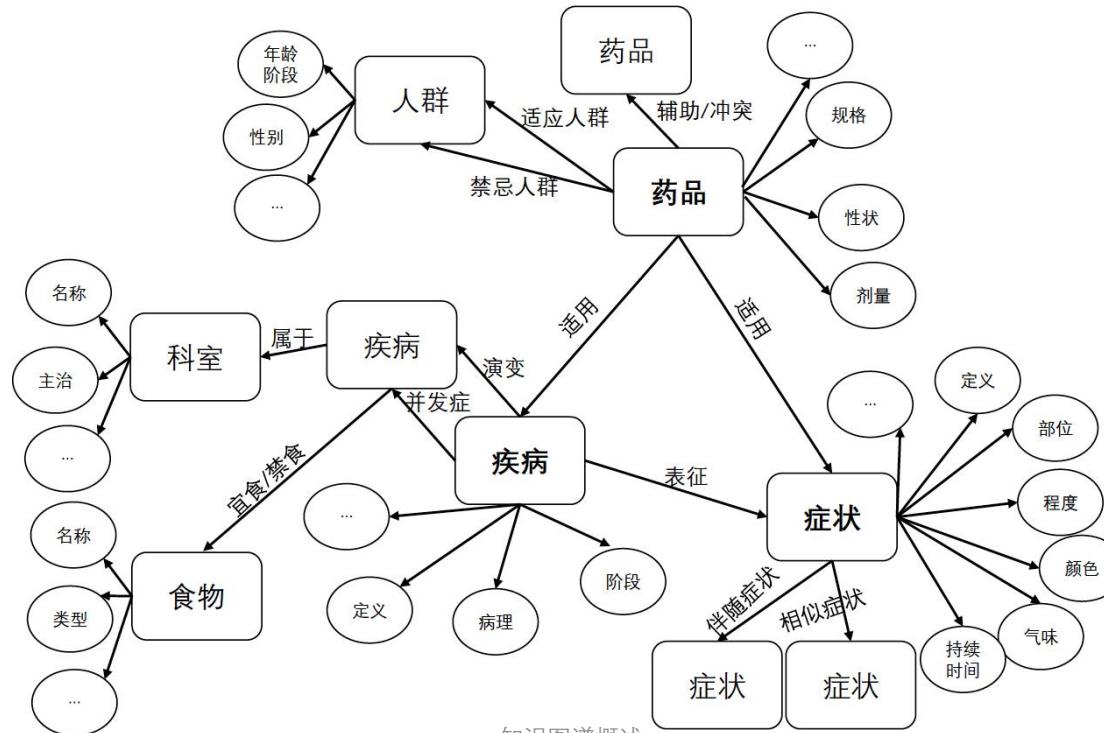
医疗领域知识图谱

❖ 医疗知识图谱构建流程



医疗领域知识图谱

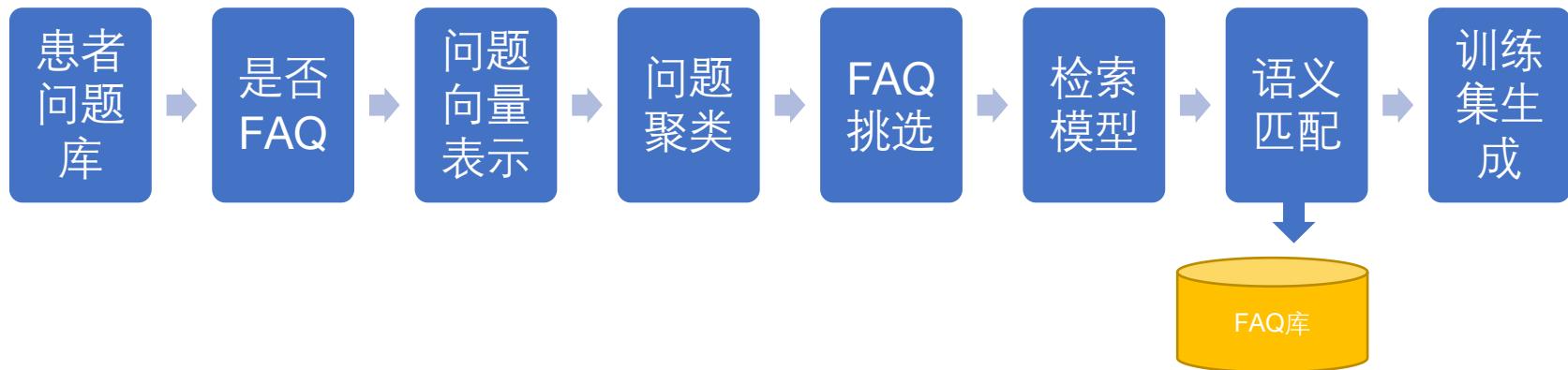
❖ 医疗知识图谱Schema



医疗领域知识图谱

智能问诊服务—疾病健康问答

- 定义：基于患者在慢病管理当中会有各种疑问，大量健康教育等相关问题，从医疗知识库当中找到答案并输出
- 举例：糖尿病患者应该如何饮食？立普妥和络活喜可以一起吃吗？空腹血糖8.6，有问题吗？



金融领域知识图谱



金融知识图谱

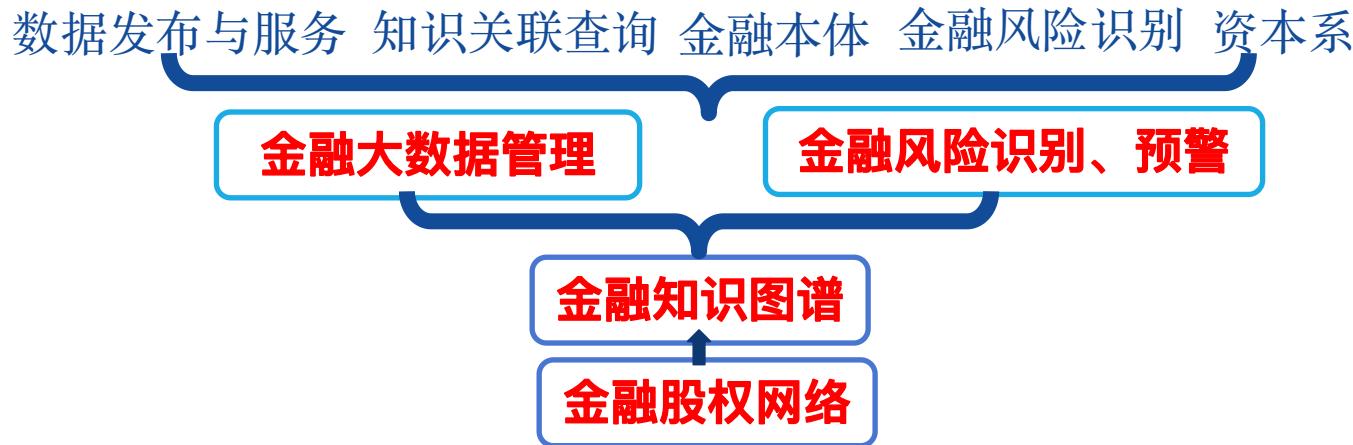
- 丰富的数据源采集
- 强大的图谱构建关系挖掘
- 场景化业务分析与挖掘
- 商业智能解决方案

具体应用场景

- 企业信贷风险评估
- 反欺诈
- 智能投研
- 投资产品推荐
- ...

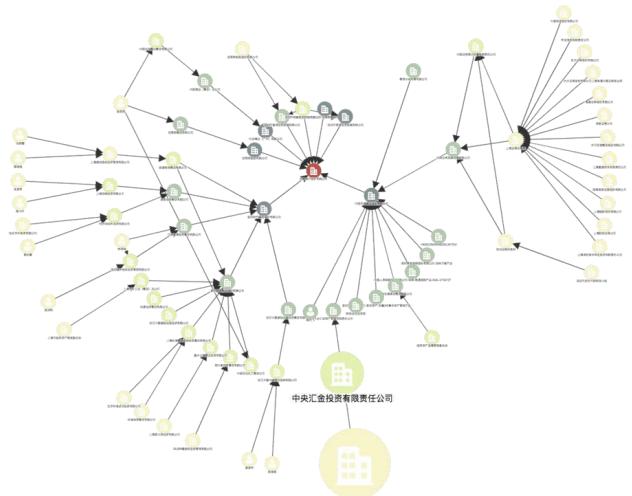
金融领域知识图谱—知融金融知识图谱平台

- 融合商业银行（400+）的股权数据和全量（4000万+）工商企业注册数据，自动构建了亿级实体-关系三元组的金融知识图谱
- 结点：工商企业、银行；边：股权关系
- 结点信息：类型、行业、注册资本、人员、历史变更……
- 边信息：控股、共同股东、一致行动、关联交易、共同人员……



知融金融知识图谱平台

招商银行股份有限公司



第零层

第一层

第二层

第三层

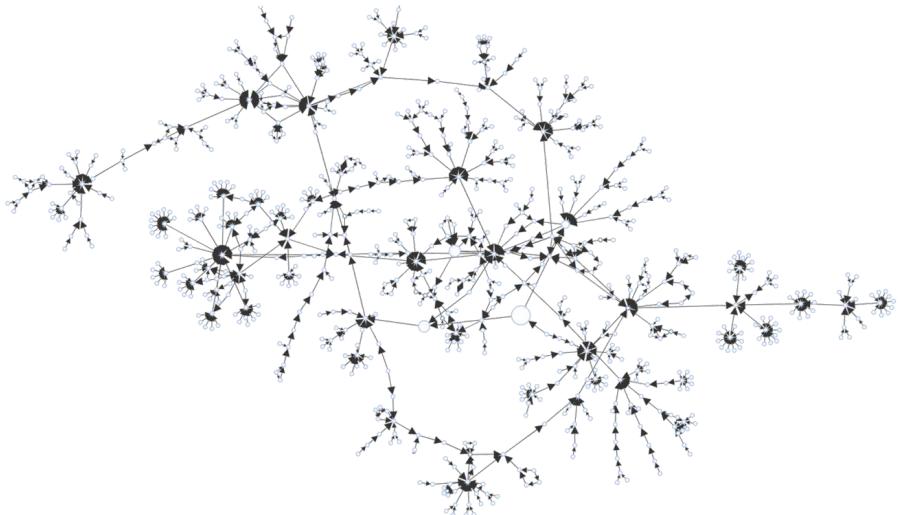
第四层

金融机构

资本系核心

其他

希望系



穿透式多层次股权查询

资本系查询

知融金融知识图谱平台

综合评估 (2015-05)



综合信用等级



综合信用评价



综合信用评分

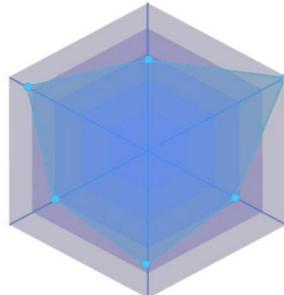
AAA

低风险

757

身份特征

守法合规

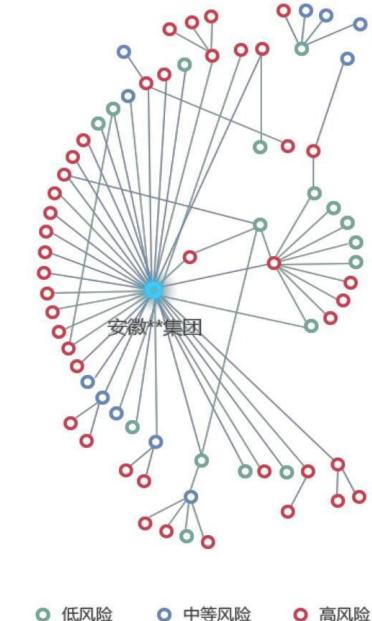


创新资质

发展潜力

经营状态

行为偏好



知识图谱概述

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

➤ 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
-

总结

本讲主要内容：

- 什么是知识图谱？
- 从各学科角度探讨知识图谱研究的问题
- 从人工智能角度看待知识图谱
- 知识图谱应用项目的普遍流程

THANK YOU
