

OpenKG.CN

OpenKG: 产学研合作的必经之路

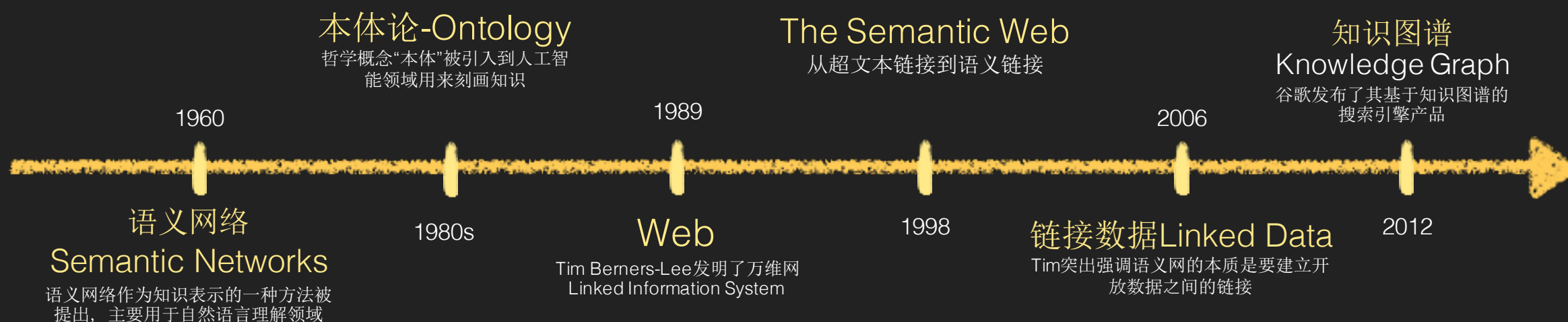
目录

- ▶ 关于KG
- ▶ 关于OPEN
- ▶ 关于OpenKG



关于KG

相关概念的演变



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。

什么是KG?

KG辅助搜索

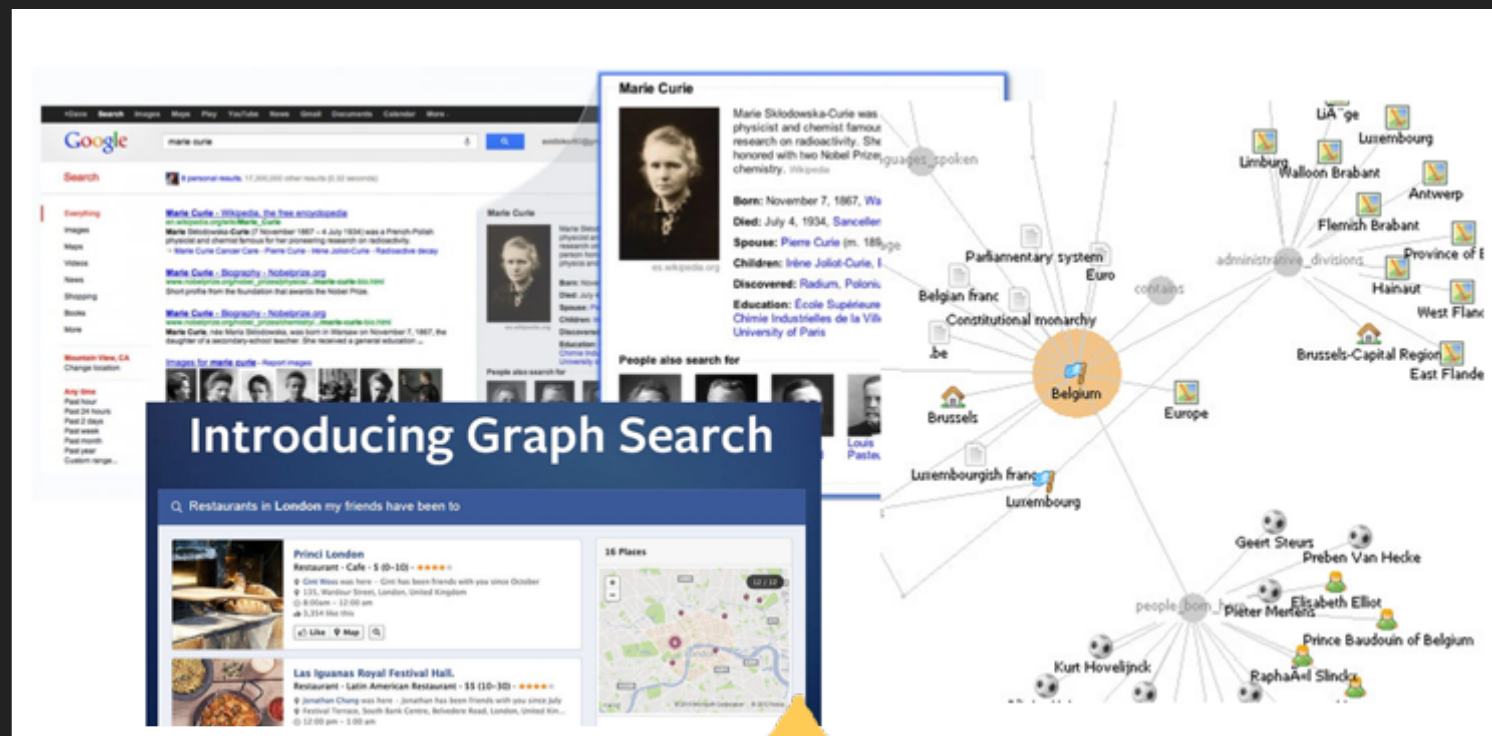
WEB的理想是万物的链接，搜索的理想是事物的搜索

网页搜索

Web of Docs

语义搜索

Web of Data



手工众包

格式转化

元组抽取

实体融合

链接预测

推理补全

语义嵌入



社区协同构建



维基众包



网页嵌入语义数据

什么是KG？

KG辅助问答交互

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。



DBpedia Yago
Wordnet



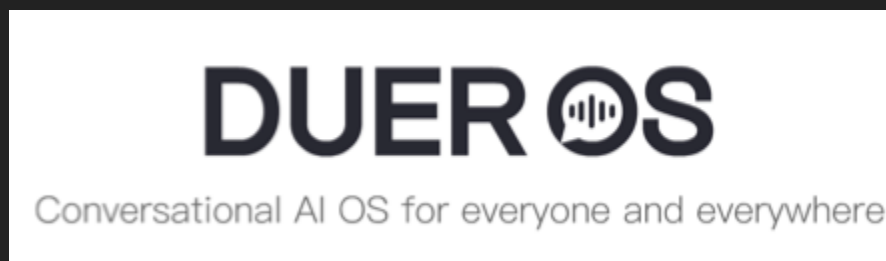
True Knowledge/Evi



WolframAlpha
DBpedia

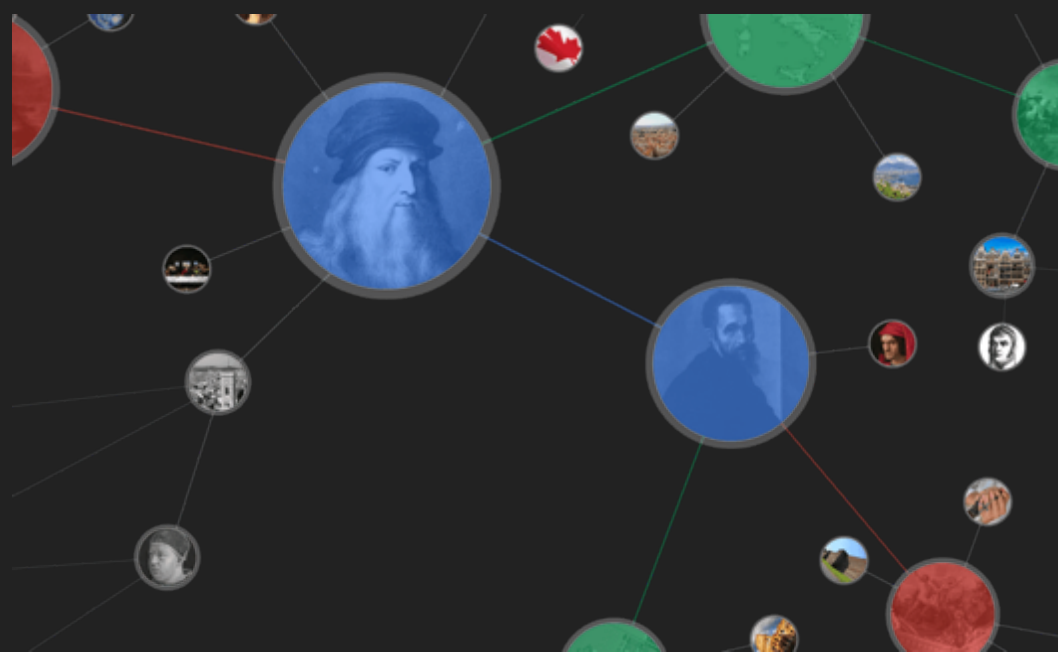


zhishi.me



KG辅助决策

More Machine Understandable



建立数据链接
预先抽取语义

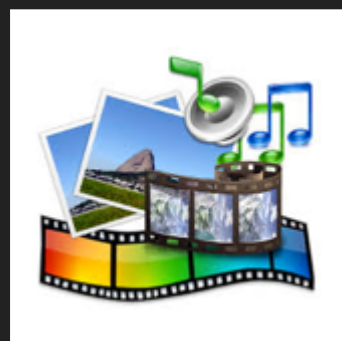
更加规范的数据表达

可计算的数据

粗糙数据



文本



多媒体



传感器

大数据与可计算大数据

...我们现在触手可及的文本型健康医疗数据很多，能够获取的数据规模也都很大，但这是否就意味着这些数据可以供研究机构或商业机构做大数据分析？

.....要把自然语言的描述转换为计算机可以计算的数据，需要依赖自然语言处理中的很多基础技术，比如句子的分词，实体识别，实体的归一化和链接等

.....可以把符号推理与统计学习结合起来，让碎片化知识图谱的知识推理和深度学习的决策模型结合起来，我们称之为**Local Knowledge Powered Global Learning...**

“文本型医疗大数据，拿来就可用？”微软亚洲研究院闫峻博士

KG辅助AI: Commonsense Knowledge

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

纯NLP : 50%



NLP+KB : 超过60%



及格线: 90%

KG辅助AI：未知的难题

► 人的符号记忆是连续的吗？

- Hinton：“...智能的精华是怎样实现推理，推理的关键是形式化逻辑...大脑中并没有符号和图像的存储，而是一些Big Activity Vectors之间相互发生作用。”
- 知识的表示是否需要向量化和连续化？

► 人的符号记忆是结构化的吗？

- “大脑分区域管理不同的文字语义。一个区域分管某个语义区域或一组相关的概念。例如：数字的Numerical，空间性的Locational，情感相关的Emotional，社交相关的Social等概念都由不同的区域分管，并且不同人之间具有相似性。” Nature(2016). Natural speech reveals the semantic maps that tile human cerebral cortex.
- 大脑中知识组织的结构是文本化还是结构化？

► 符号记忆的获取和推理过程又是什么？

- 认知科学家道格拉斯-霍夫施塔特：“记忆，是高度重建的。在记忆中进行搜取，需要从数目庞大的事件中挑选出什么是重要的，什么是不重要的，强调重要的东西，忽略不重要的东西。这种选择过程实际上就是感知，认知的核心过程与感知的关系非常非常的密切。”
- 符号推理的过程是符号搜索还是向量计算？

KG的本质

- ▶ **Web**的视角：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**的视角：怎样从文本中抽取语义和结构化数据；
- ▶ **KR**的视角：怎样利用计算机符号来表示和处理知识；
- ▶ **AI**的视角：怎样利用知识库来辅助理解人的语言；
- ▶ **DB**的视角：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

KG的构建与获取: Freebase



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

KG的构建与获取: WikiData

目标是构建全世界最大的免费知识库, 采用CC0完全自由许可协议



维基基金会

导入维基数据



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

谷歌

提供资金

摩尔基金会

提供资金

Allen AI Institute

仍然面临知识缺失严重的问题

KG的构建与获取：SCHEMA.ORG

只针对开放域

Semantic Markup : 在网页、邮件、应用程序中嵌入语义数据

```
<p>Big (1988) is a fantasy comedy movie starting Tom Hanks and directed by Penny Marshall.</p>
```

RDFa, JSON-LD, HTML5 Microdata

```
<div itemscope itemtype="http://schema.org/Movie"><p>  
  <span itemprop="name">Big </span> (1988) is a  
  <span itemprop="genre">fantasy comedy</span> movie starting  
  <span itemprop="actor" itemscope itemtype="http://schema.org/Person">  
    <span itemprop="name">Tom Hanks</span></span> and directed by  
  <span itemprop="director" itemscope itemtype="http://schema.org/person">  
    <span itemprop="name">Penny Marshall</span></span>.</p></div>
```

搜索引擎优化

已经包含642个类，992个关系



schema.org

Home Schemas Documentation

截止目前，全球约有1.2亿的网站，超过30%的网页已经嵌入有Schema.org的语义数据。谷歌2015年12月推出的Cutomizable Knowledg Graph，允许各网站基于Schema.org，以RDFa，JSON-LD，HTML Microdata等方式在网页、邮件等数据源中嵌入语义化数据，支持个人和企业定制自己的知识图谱信息。

KG的构建与获取：ConceptNet

起源于MIT Marvin Minsky倡导的Open Mind Common Sense项目

采用CC BY SA开放许可协议



早期版本主要来源于专家录入、众包和游戏。

新版本导入大量开放的结构化数据，如 DBPedia、Wikinary, Wordnet等



关于OPEN

关于KG的几个事实

现有的大规模知识图谱都大量的依赖开放域的数据抽取
和开放社区协作的方式构建

Freebase Yago DBPedia Wikidata ConceptNet ...

Zhishi.me CN-DBPedia X-Lore PKU-PIE

关于KG的几个事实

不能指望NLP和机器学习能自动的去构建你真正需要的知识图谱。
但KG的确需要大量的使用NLP/ML的各种技术手段，来帮助冷启动、后期补全和质量提升。

NLP只是获取KG手段之一，文本也只是KG数据的一种来源

关于KG的几个事实

构建KG的成本非常高，单一机构无法构建覆盖全面的知识图谱，即便是金融这样的垂直领域也是如此。每个从事KG研发的机构都会也应该去寻求开放社区和外部数据的帮助。

所有KG的共同问题：知识缺失严重！

纯手工众包

内部结构化或文本数据

开放域的结构化/半结构化数据

开放域的文本抽取

关于OPEN的解读

OpenKG $\neq$ Open Data

开放域的数据

结构化并链接已经开放的数据



开源的数据

类似于开源的软件



开放的数据



提高可利用性：语义表达规范的数据更加容易被利用和重用



提高可发现性：开放是为了让更多人的发现你，并给你带来收益



提高可公开性：开放你所拥有的数据

开源的数据：可商业化的开放数据许可协议

- ▶ 做KG生意绕不过开放的问题，并面临数据版权问题
 - ▶ Creative Common
 - ▶ CC0：完全自由，例如WikiData
 - ▶ CC-BY：署名许可，适应性最强的许可，鼓励商业化，如Yago
 - ▶ CC-BY-SA：署名且要求采用相同的许可条款，鼓励商业化，例如：ConceptNet
 - ▶ Open Knowledge Foundation
 - ▶ PDDL\ODC-By\ODbL

可能的KG商业模式（大佬企业除外）

KG开放或部分开放 + 增值技术服务 / 优先使用数据 / 高质量KG服务等

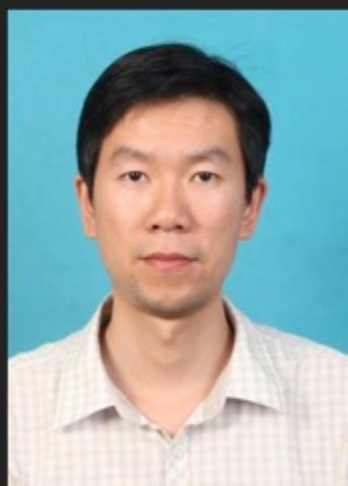


关于OpenKG

OpenKG的使命

- ▶ 传播开放之于KG的价值及理念
- ▶ 培育开放的知识图谱社区
- ▶ 构建开放域的链接知识图谱
- ▶ 共同探索开放知识图谱的商业模式与商业价值

主要发起人



- ▶ 陈华钧，浙江大学教授，中国中文信息学会理事、语言与知识计算专委会副主任，浙江省大数据智能计算重点实验室副主任，国际大数据期刊Elsevier Journal of Big Data Research副主编。曾获国际语义网大会ISWC最佳论文奖。CCKS2016全国知识图谱与语义计算大会程序委员会主席。



- ▶ 漆桂林，东南大学教授，中国中文信息学会语言与知识计算专委会副主任和CCF中文信息技术专委会委员。Journal of Web Semantics的编委。在国际人工智能著名会议IJCAI、AAAI、KR发表多篇论文。CCKS2017全国知识图谱与语义计算大会程序委员会主席。

主要发起人



- ▶ 王昊奋，深圳狗尾草智能科技CTO，著名知识图谱专家，中文知识图谱Zhsihi.me创始人，YOCSEF上海主席，中国计算机学会术语专委会执委，公子小白、琥珀机器人技术负责人。



- ▶ 谢殿侠，上海海知智能科技创始人兼CEO，中国中文信息学会理事，主攻个性化聊天机器人开放技术平台ruyi.ai，负责贤二机器僧的知识库研发。

主要发起人



- ▶ 曾毅，中国科学院自动化研究所研究员，类脑智能研究中心副主任，中国科学院大学岗位教授。Belief Engine中英文知识图谱、最大规模的脑科学知识图谱Linked Brain Data创始人。Cognitive Systems Research, Computational Cognitive Science国际期刊副主编。



- ▶ 鲍捷，文因互联创始人兼CEO，著名金融知识图谱专家，美国Iowa State University博士，三星问答系统S-Voice第二代系统核心设计师。

其他主要发起人

- ▶ 李娟子 清华大学 教授
- ▶ 赵东岩 北京大学 教授
- ▶ 俞思伟 武汉大学 教授
- ▶ 阮 彤 华东理工大学 教授
- ▶ 顾进广 武汉科技大学 教授
- ▶ 黄志生 荷兰阿姆斯特丹自由大学 教授
- ▶ 陈文亮 苏州大学 教授
- ▶ 陈运文 达观数据 CTO
- ▶ 朱其立 上海交通大学 教授
- ▶ 丁 军 上海海翼知科技有限公司 CEO
- ▶ 张 洁 神龙良品科技有限公司 CEO
- ▶ 丁 力 上海海知智能科技 CTO
- ▶ Jeff Pan 英国Aberdeen大学 教授
- ▶ 张小旺 天津大学 副教授
- ▶ 冯岩松 北京大学 副教授
- ▶ 肖仰华 复旦大学 副教授
- ▶ 闫 俊 微软亚洲研究院 研究员
- ▶ 赵 丹 海云数据 CTO
- ▶ 刘升平 云知声科技有限公司
- ▶ 杜剑峰 广东外贸外语大学 教授
- ▶ 胡 伟 南京大学 副教授
- ▶ 刘知远 清华大学
- ▶ 刘译景 百分点科技
- ▶ 徐叶强 广州索达科技有限公司 CTO
- ▶ 邹 磊 北京大学 副教授
- ▶ 朱烨林 OMAHA开放健康医疗联盟
- ▶ 于 彤 中国中医科学院信息研究所

近期工作一： 培育开放的社区

- ▶ KG社区：
 - ▶ 语义网 & 知识图谱微信群
 - ▶ OpenKG公众号
 - ▶ 各个垂直领域的开放社区：医疗、金融
- ▶ 参与组织各种社区活动
 - ▶ KG沙龙：每月一次，全国巡回
 - ▶ CCKS全国知识图谱与语义计算大会



回答集程序 (Answer Set Programming) 求解器groc

追从者

0

➕ 追随

🏢 机构

WESTERN SYDNEY
UNIVERSITY



Western Sydney
University

没有关于此机构的描述

🌐 社会

📱 Google+

🐦 Twitter

📘 Facebook

🛠 工具

🗂 分类

🕒 活动流

🔧 管理

回答集程序 (Answer Set Programming) 求解器 groc

回答集程序 (Answer Set Programming)，简称ASP，是一个描述性编程语言 (Declarative Programming Language)，是人工智能知识表示领域中最重要几个语言之一。很多知识并不适合用经典一阶逻辑 (First-Order Logic) 来表示，比如：“如果一个人有VIP卡，而且不知道其是否黑名单上的话，可以给折扣”。这种情况下，可以把它表示成规则如下：Discount(x) :- VIP(x), not BlackList(x) 其中，Discount(x)，VIP(x)以及BlackList(x)分别表示可以给x折扣，x是VIP以及x在黑名单上。“:-”是用来表示规则的符号，可以读成“如果”，含义是如果该符号右侧所有事实成立，那么该符号左侧事实也成立。“not”符号是表示缺省符号“不知道”。回答集程序就是用来计算如果有一堆如上所述规则，那么能推出来的结果应该是什么。回答集程序是很多经典表示和求解语言的扩充，包括命题可满足求解 (SAT)，约束满足求解 (CSP)，非单调推理 (NMR)，演绎数据库 (Datalog) 等等。简而言之，ASP=SAT+Datalog+NMR+CSP

groc是一个新的ASP求解器，支持normal ASP rule, aggregate, not等等。

数据与资源



回答集程序求解器(ASP solver)groc

下载后先解压。使用前请先阅读“回答集程序求解器groc”和“使用说明”。

↪ 浏览

Datalog

NP问题求解

描述性编程

约束满足问题求解

逻辑程序设计

其他信息

域

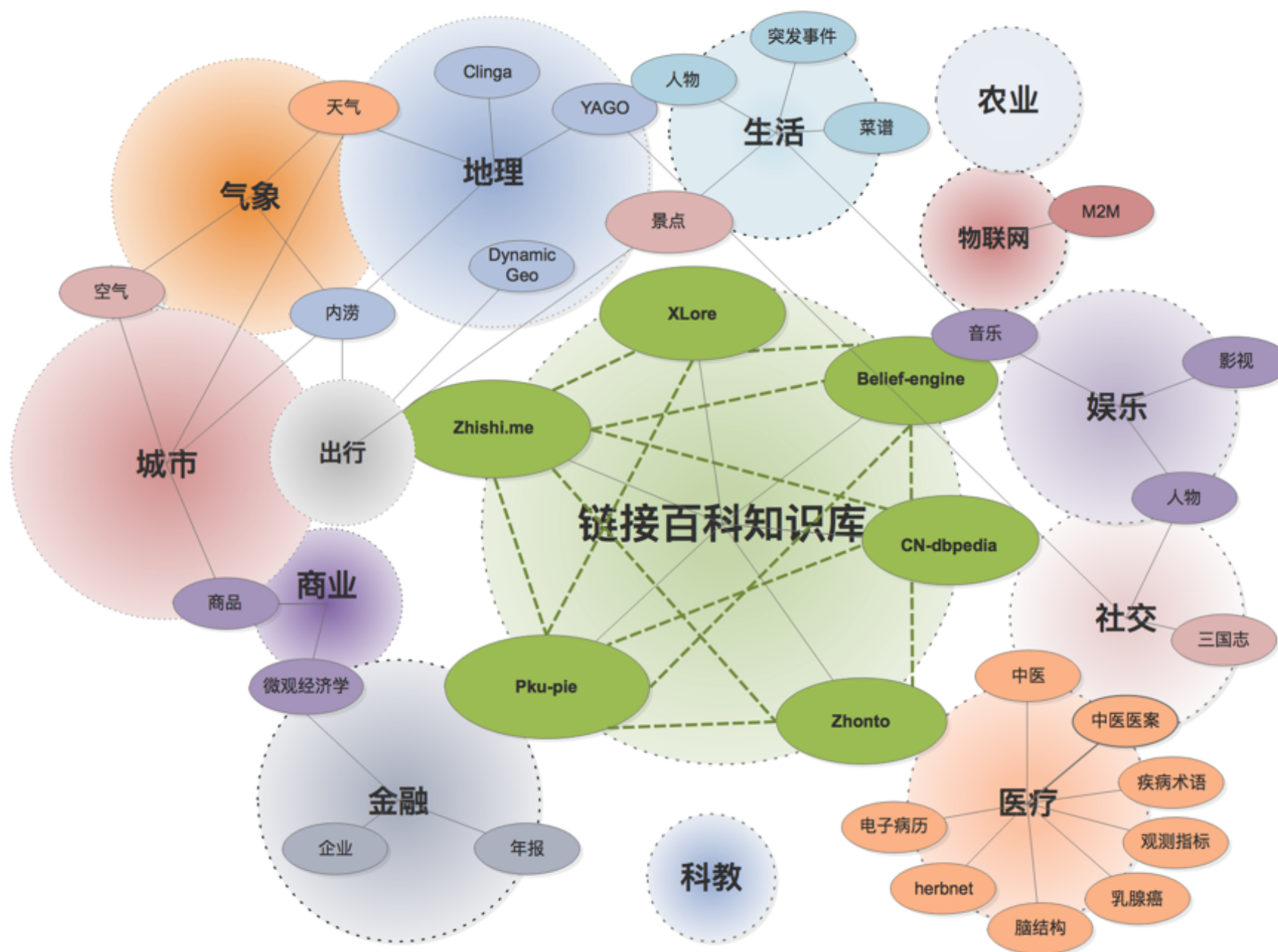
价值

已经聚集的几个典型中文知识库的规模

- ▶ Zhishi.me : 近1000万实体, 10亿三元组
- ▶ CN-DBpedia : 近1000万实体, 8000多万三元组
- ▶ XLORE双语百科知识图谱: 1000万实体
- ▶ Belief Engine : 5000万三元组
- ▶ PKU-PIE知识库: 70万个实体, 900万条三元组
- ▶ ConceptNet中文部分 : 2800万三元组
- ▶ WikiData中文部分 : 60万Items

近期工作三：OpenKG.LINK

构建开放域的链接知识图谱：从链接百科类知识库开始做起



链接开放
内容私有

近期工作四：垂直领域合作

OpenKG与OMAHA开放医疗与健康联盟的合作

KNOWLEDGE
知识服务平台

图谱 工具 关于OMAHA 加入我们



OMAHA & OpenKG

医学

追从者 0 图谱 17

机构

- 中国中医科学院中药信息研究所 (5)
- Vrije University of... (5)
- 中国科学院自动化研究所 (4)
- 开放医疗与健康联盟 (1)
- 华东理工大学 (1)
- 个人 (1)

标签

资源集 活动流 关于

搜索资源集...

找到17个资源集 排序: 相关度

疾病术语集

疾病术语集包了疾病实体及疾病相关同义词，同时建立了疾病实体与UMLS概念的映射，术语集的数据来源于国内权威的临床医学术语网站。

ZIP

观测指标标识符逻辑命名与编码系统

观测指标标识符逻辑命名与编码系统 (Logical Observation Identifiers Names and...

ZIP

虚拟电子病历生成工具

APDG是一个基于知识的虚拟电子病历生成工具，可用于生成语义电子病历或其他格式（如XML）的电子病历

java

乳腺癌临床试验语义标注

乳腺癌临床试验语义标注

ZIP RDF



近期工作五：开放的Bot Schema

BOT SCHEMA和KG应用生态建设

丁力：博士，海知智能CTO。全球首款语义搜索引擎Swoogle作者，美国开放政府数据Data.gov语义技术核心贡献者。近二十年在知识图谱，语义搜索，政府数据公开，社交网络，机器学习与大数据分析等领域研究和实践，发表上百篇论文，引用量过万，google H-index>30。前高通研究院科学家, 斯坦福大学博士后。

加入OpenKG的价值

- ▶ 提高你的数据的可发现性
- ▶ 与其它数据进行链接
- ▶ 帮助你的数据实现商业化
- ▶ 共同构建开放域的知识图谱

总结

开放是KG的源生基因

开放是做好KG的必要手段和途径

OpenKG需要产学研密切的合作

CCKS 2017-全国知识图谱与语义计算大会

China Conference on Knowledge Graph and Semantic Computing – 成都, 2017年8月26- 29日

[首页](#) / [组织委员会](#) / [论文征集](#) / [日程](#) / [注册](#) / [会议地点](#) / [特邀报告](#) / [讲习班](#) / [评测竞赛](#) / [工业界论坛](#) / [赞助](#) / [English](#)



加入我们 请勿问

<http://OpenKG.CN>



开放推动互联 链接创造价值