

第一讲 知识图谱概述

赵 军 (jzhao@nlpr.ia.ac.cn)

中国科学院自动化研究所
模式识别国家重点实验室



中国科学院自动化研究所
Institute of Automation, Chinese Academy of Sciences

课程目的

- 以知识图谱为代表的知识工程的基本问题和基本方法
- 基于知识图谱的语义计算的关键技术
- 使得学生初步具备建立通用或领域知识图谱，以及基于知识图谱进行数据分析、文本分析应用的能力

提纲

- 什么是知识图谱
- 知识图谱发展历程
- 知识图谱的类型和代表性知识图谱
- 知识图谱的生命周期
- 知识图谱和深度学习
- 课程介绍

人工智能

■ 计算智能

- 运算和存储能力

■ 感知智能

- 视觉、听觉、触觉等感知能力
- 人和动物都能够通过智能感知能力与自然界进行交互

■ 认知智能

- 认知是人在感知客观世界以及人与人互动的过程中形成的对世界万物的理性认识——即概念和意义，包括推理、概括、演绎、监控、理解和记忆等一系列心智活动。
- 人类所特有的认知的建立需要思考能力，而思考建立在知识的基础之上

什么是知识

- Plato 认为知识是永恒不变的且适用于世间万物的真理。
- Davenport 认为知识是与经验 (Experience)、背景 (Context)、解释 (Interpretation) 和思考 (reflection) 结合在一起的信息，它是一种可以随时帮助人们决策与行动的高价值信息。
- Harris 认为知识是信息、文化脉络及经验的组合。

什么是知识

- Feigenbaum 认为知识是经过消减、塑造、解释和转换的信息，即知识就是经过加工的信息。
- Bernstein 认为知识是由特定领域的描述、关系和过程组成的。
- Hayes-Roth 认为知识是事实、信念和启发式规则
- Nonaka 认为显式的或者已编码的 (codified) 知识是指一种用正式、系统化的语言传输的知识；另一方面，隐性知识拥有个性化的特征，这使得隐性知识很难进行形式化描述和分享。

知识的分类：陈述性、过程性

- 陈述性知识描述客观事物的性状和关系等静态信息，主要分为事物、概念、命题三个层次。
 - 事物指特定的事或物
 - 概念是对一类事物本质特征的反映
 - 命题是对事物之间关系的陈述
 - 非概括性命题表示特定事物之间关系
 - 概括性命题描述概念之间的普遍关系
- 过程性知识描述问题如何求解等动态信息
 - 规则描述事物的因果关系
 - 控制结构描述问题的求解步骤

什么是知识库

- 在对各种知识进行收集和整理的基础上，进行形式化表示，按照一定方法存储，并提供相应的知识查询手段，从而使知识有序化，是知识共享和应用的基础。
- 知识的编码化和数字化就形成知识库。

知识工程

- 1977年，在第五届国际人工智能会议上，美国斯坦福大学计算机科学家费根鲍姆教授(E. A. Feigenbaum) 发表特约文章 “人工智能的艺术：知识工程课题及实例研究”，系统地阐述了 “专家系统” 的思想，并提出了 “知识工程” 的概念，确定了知识在人工智能中的重要地位。
- 1984年， “ARTIFICIAL INTELLIGENCE: THE SECOND COMPUTER AGE ”

Knowledge is power, and the computer is an amplifier of that power. We are now at the dawn of a new computer revolution... Knowledge itself is to become the new wealth of nations.



互联网知识服务

- 知识蕴含在数据之中。
- 互联网时代，人类在与自然和社会交互中生产了异常庞大的数据，这些数据中包含了大量描述自然界和人类社会客观规律的有用信息，这些信息以图片、声音、文字、视频等各种载体表示和存储。
- 如何让计算机自动阅读、分析、理解这些海量繁杂乃至泛滥的数据，从中挖掘有价值的知识，向用户提供精准知识服务，是构建下一代信息服务，乃至人工智能技术的核心目标之一。

互联网知识服务与知识图谱

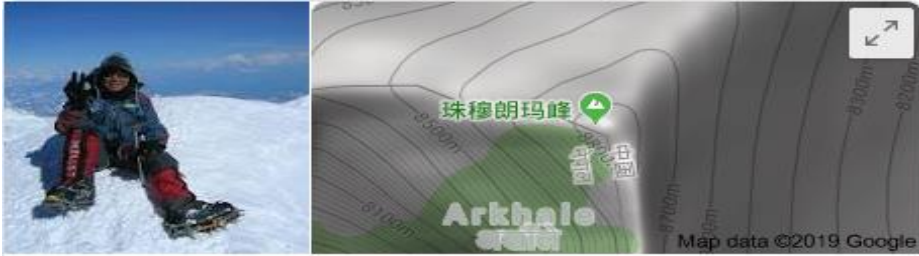
- 互联网中大量的信息以非结构化的形式存储和传播，而计算机擅长处理结构化数据。为了让计算机能够处理这些信息，就需要理解这些非结构化形式数据蕴含的语义，分析其中的语义单元之间的关系，从而转换成结构化形式。
- 图是一种能有效表示数据之间结构的表达形式，因此，人们考虑把数据中蕴含的知识用图结构进行形式化表示。数据的结构化并和已有的结构化数据进行关联，就构成知识图谱。
- 知识图谱对于知识服务有重要的支撑作用，能够将传统基于浅层语义分析的信息服务范式提升到基于深层语义的知识服务。因此，学术界和工业界都给予知识图谱高度关注，将其作为新一代人工智能的基础设施。

什么是知识图谱

- 知识图谱是谷歌公司用来支持从语义角度组织网络数据，从而提供智能搜索服务的知识库。从这个意义上讲，知识图谱是一种比较通用的语义知识的形式化描述框架，它用节点表示语义符号，用边表示符号之间的语义关系。
- 在计算机世界中，节点和边的符号通过“符号具化（symbol grounding）”表征物理世界和认知世界中的对象，并作为不同个体对认知世界中信息和知识进行描述和交换的桥梁。这种使用统一形式描述的知识描述框架便于知识的分享与利用。

Google Knowledge Graph

The Knowledge Graph is a knowledge base used by Google and its services to enhance its search engine's results with information gathered from a variety of sources. The information is presented to users in an infobox next to the search results.



珠穆朗玛峰
4.2 ★★★★★ 9,954条 Google 评价

珠穆朗玛峰，中文简称珠峰，尼泊尔名萨加玛塔峰，英国名埃佛勒斯峰，为地球第一高峰，属于喜马拉雅山脉，位于中国西藏自治区与尼泊尔萨加玛塔专区边界上。 [维基百科](#)

海拔： 29,029'
首次登顶时间： 1953 年 5 月 29 日
突起高度： 29,029'
首位登顶者： [艾德蒙·希拉里](#)，[丹增诺盖](#)
山脉： [喜马拉雅山脉](#)，[马哈朗格·喜玛尔](#)
国家/地区： [尼泊尔](#)，[中华人民共和国](#)

评价
9,954条 Google 评价

[写评价](#) [添加照片](#)

用户还搜索了

 乔戈里峰 28,251'	 干城章嘉峰 28,169'	 冒纳凯阿火山 热点	 洛子峰 27,940'	 乞力马扎罗山 19,341'
--	--	--	--	---

还有15+项

知识组织：一些容易混淆的术语

■ Ontology vs. Knowledge Base vs. Database

- Ontology: 一套对客观世界进行描述的共享概念化体系, 对特定领域中概念 (对象类型) 及其相互关系进行形式化表达。用来进行领域建模或者推理。因此, 重点是对数据的定义进行描述, 而没有描述具体的实例数据。
- Knowledge Base: 服从于ontology 控制的知识实例及其载体
- Database: 计算机科学家为了用电脑表示和存储计算机应用中所需要的数据所设计开发的产品。
- Ontology是蛋糕的模具, Knowledge Base是蛋糕, Database是存放蛋糕的盒子

知识组织：一些容易混淆的术语

■ Ontology

- 树状结构，不同层节点之间是严格的IsA关系
- 优点：可以适用于知识的推理
- 缺点：无法表示概念的二义性（运动员：体育？人物？）

■ Taxonomy

- 树状结构，上下位节点之间非严格的IsA关系
- 优点：可以表示概念的二义性（体育→运动员）
- 缺点：不适用于推理，无法避免概念冗余（餐厅：美食？机构？地点？）

知识图谱

- 人工智能应用中不仅需要具体的知识实例数据，数据的描述和定义也非常关键，
- 知识图谱用统一的形式对知识实例数据的定义和具体知识数据进行描述，即用三元组形式（二元关系）对知识系统进行资源描述和存储。
- 例如：WikiData
 - 知识结构
 - <physicist, subclass of, scientist>
 - <place of birth, value type constraint, geographical object>
 - <place of birth, subproperty of, location>
 - 实例数据
 - <Max Planck, instance of, human>
 - <Max Planck, occupation, physicist>
 - <Max Planck, place of birth, Kiel>

知识图谱

- 各个具体实例数据只有在满足系统约定的“框架”约束下运用才能体现为“知识”，其中框架（Schema，或称“元知识”）就是对知识的描述和定义，知识框架和实例数据共同构成一个完整的知识系统。
- 在约定的框架下，对数据进行结构化，并与已有结构化数据进行关联，就形成了知识图谱。
- 为了将其付诸实现，知识图谱往往需要将自身的框架结构映射到某种数据库系统所支持的框架定义上，必要时可以对数据库进行专门扩展。
- 因此，知识是认知，图谱是载体，数据库是实现，知识图谱就是在数据库系统上利用图谱这种抽象载体表示知识这种认知内容。

知识图谱

- 知识图谱以结构化三元组的形式存储现实世界中的实体及其关系，表示为 $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{S})$ ，三元组通常描述了一个特定领域中的事实，由头实体、尾实体和描述这两个实体之间的关系组成。

- 实体集合: $\mathcal{E} = \{e_1, e_2, \dots, e_{|\mathcal{E}|}\}$

- 关系集合: $\mathcal{R} = \{r_1, r_2, \dots, r_{|\mathcal{R}|}\}$

- 三元组的集合: $\mathcal{S} \subseteq \mathcal{R} \times \mathcal{E} \times \mathcal{E}$

- Freebase

- 三元组 “/people/person/nationality (Jorge Amado, Brazil)” , 它表示 “Jorge Amado” 的国籍是 “Brazil”

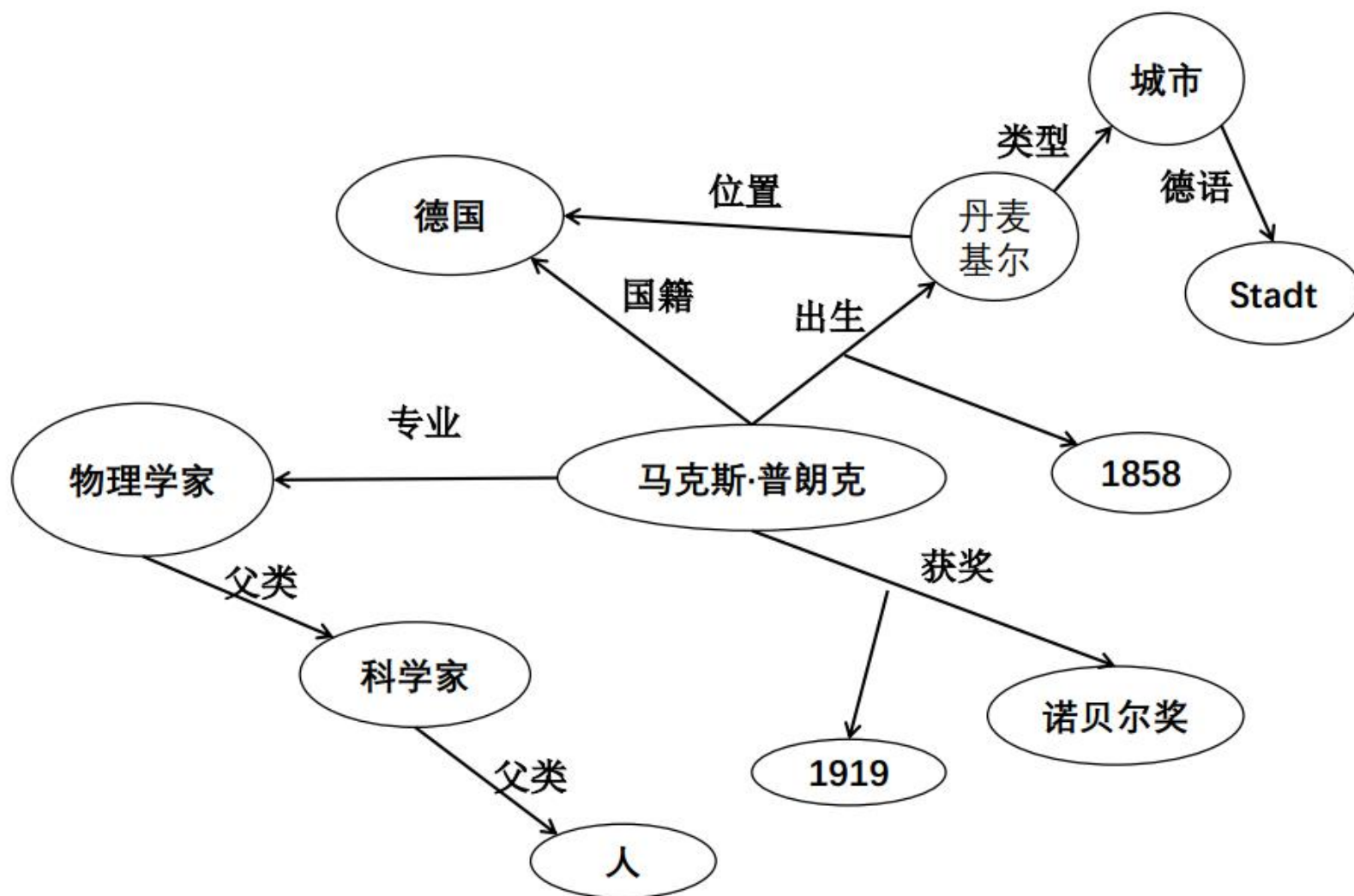
- 头实体: “Jorge Amado”

- 尾实体: “Brazil”

- 关系名称: “/people/person/nationality”

- 关系有时也称为属性，相应地，尾实体被称为属性值。
- 从图结构的角度看，实体是知识图谱中的节点，关系是连接两个节点的有向边。

什么是知识图谱

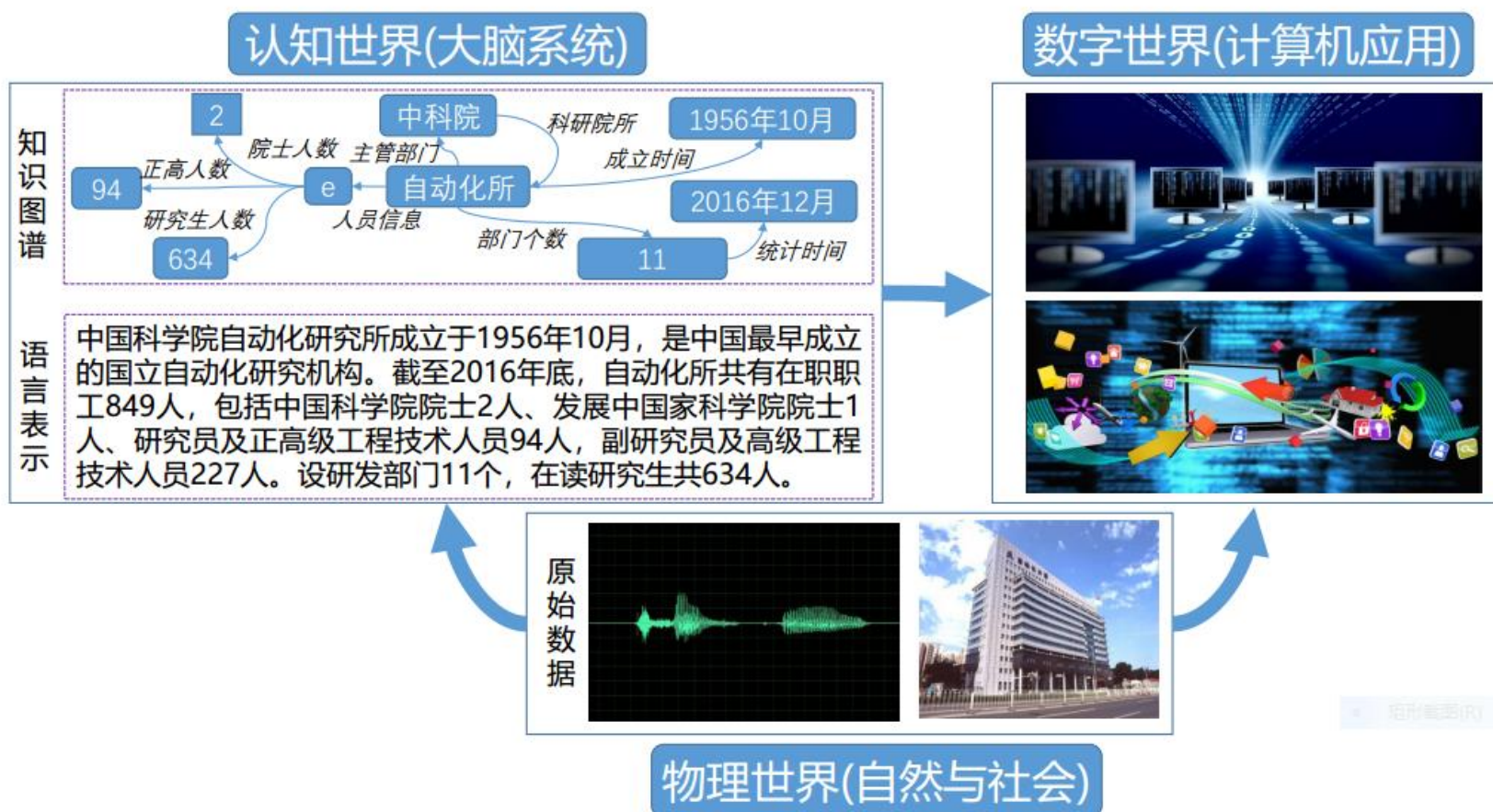


广义的知识图谱

- 从知识工程角度，知识框架一般包含三个层次的知识：
 - 概念知识：给出知识最基本的内容
 - 事实知识：建立概念之间的联系
 - 规则知识：建立事实间的联系
- 对于知识图谱应不应该包含推理规则型的知识，不同领域的人有着完全不同的意见。
 - 知识图谱仅仅是人工智能任务的重要基础设施，不应该包含推理等知识图谱应用部分。他们认为，目前的众多流行知识图谱（如 Freebase、Wikidata、YAGO、DBpedia 等）都在去规则化，仅保留对知识数据的基本描述，因为推理规则确定性更弱，更难管理。
 - 人类文明发展过程中，有意无意总结和传承了很多规则，这些抽象的知识给人们带来了很大的便利，不仅可以节省知识存储和搜索空间，也让人类能够发展出联想、推理、决策等更高层的认知能力。因此，在模拟高层认知能力的人工智能系统中应该要建模这类知识，而知识图谱作为知识表示的基石，应该涵盖推理规则。

知识图谱：人工智能的重要基础设施

知识图谱以丰富的语义表示能力和灵活的结构构建了在计算机世界中表示认知世界和物理世界中信息和知识的有效载体。



提纲

- 什么是知识图谱
- 知识图谱发展历程
- 知识图谱的类型和代表性知识图谱
- 知识图谱的生命周期
- 知识图谱和深度学习
- 课程介绍

知识图谱的发展历程

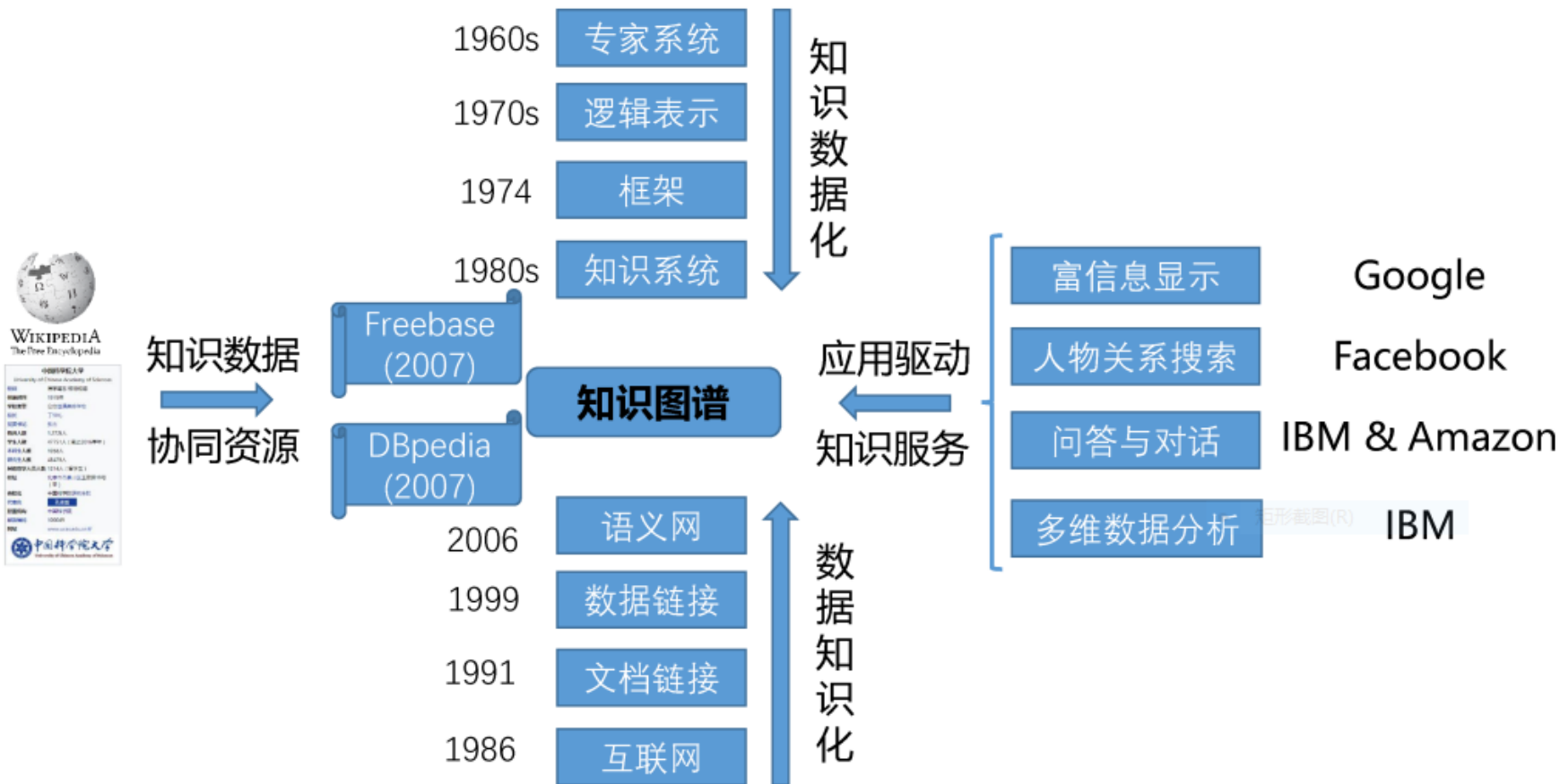
□ 人工智能：知识的数据化——让计算机表示、组织和存储人类的知识

- 人工智能的研究目标是使计算机更智能，让计算机能够完成一些人脑才能完成的任务，例如推理、分析、预测、思考等高级思维活动。人们希望借助知识库完成该目标，即把人类的知识用计算机进行表示和组织，并设计相应算法完成推理、预测等任务。

□ 语义网：数据的知识化——让数据支持推理等智能任务

- 随着互联网的发展，虽然存储和检索海量数据的能力日益提高，但是没有推理、预测等完成复杂任务的功能，不能有效支持商业智能等决策任务。人们希望通过引入知识，使得原始数据能够支撑推理、问题求解等复杂任务（如多维数据分析），这个目标就是语义网。

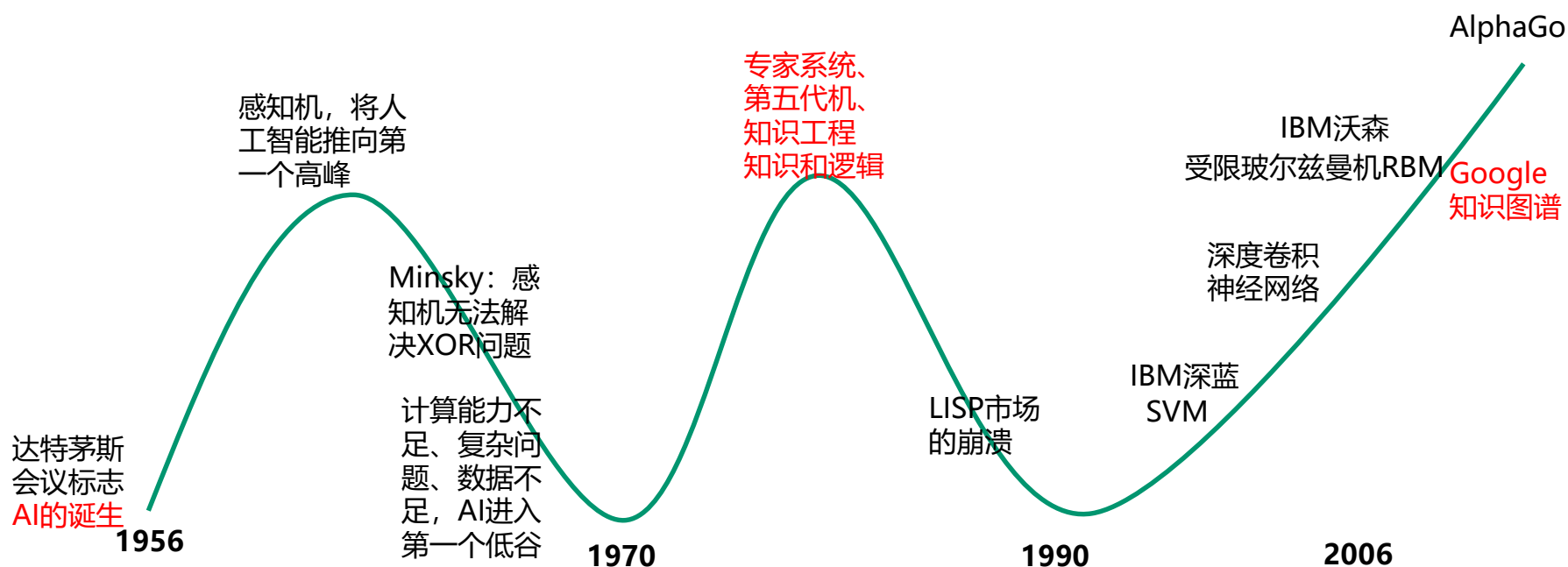
知识图谱的发展历程



提纲

- 什么是知识图谱
- 知识图谱发展历程
 - 知识图谱和人工智能
 - 知识图谱和万维网
- 知识图谱的类型和代表性知识图谱
- 知识图谱的生命周期
- 知识图谱和深度学习
- 课程介绍

人工智能的发展历史



知识图谱与人工智能

- 人工智能发展的第一个十年，研究者们看重的是如何构建一个推理模型进行问题的求解和推理，然而却忽视了对于数据中所蕴含的知识的加工和利用。
- 1977年，图灵奖获得者、斯坦福大学教授费根鲍姆最早提出了知识工程的概念，他认为实现智能行为的主要手段在于知识，在多数实际情况下是特定领域的知识。这对于人工智能技术以及相关研究的发展产生了巨大的影响。研究界提出了一系列各具特点的知识表示理论和方法，人们也试图基于这些知识表示方法构建知识库，知识工程已经成为人工智能的重要组成部分。
 - 上世纪70年代，以一阶谓词逻辑为代表的逻辑表示方法 (Logic)
 - 1974年，闵斯基提出框架表示法 (Frames)
 - 1978年，心理学家西尔文·汤姆金斯提出脚本理论 (Script theory)
 - 1987年：奎利恩提出语义网络 (Semantic Network)
- 上世纪九十年代开始，以统计机器学习为核心的人工智能技术逐步占据主流，“如何自动学习和利用知识进行推断和决策？”成为人工智能的研究热点。

提纲

- 什么是知识图谱
- 知识图谱发展历程
 - 知识图谱和人工智能
 - 知识图谱和万维网
- 知识图谱的类型和代表性知识图谱
- 知识图谱的生命周期
- 知识图谱和深度学习
- 课程介绍

WWW的发展

- World Wide Web (Web 1.0)
- Social Web (Web 2.0)
- Semantic Web (Web 3.0)
- Web Science

万维网技术的局限性（一）

- 万维网现在采用的超文本标记语言HTML，是面向人的存储和共享信息的媒介，Web上的内容是提供给人而不是机器来理解和浏览的。
 - 由于Web内容没有采用形式化的表示方式，并且缺乏明确的语义信息，故而计算机“看到的”Web内容只是普通的二进制数据，对其内容无法进行识别。
 - 如果机器不能充分理解网页内容的含义，就无法实现Web内容的自动处理。

万维网技术的局限性（二）

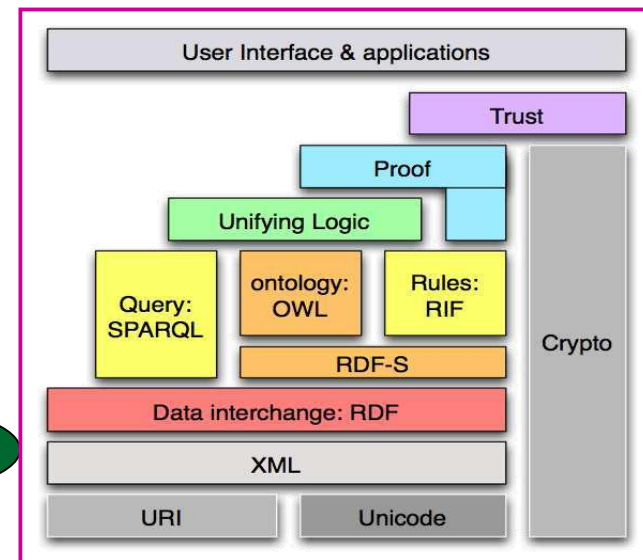
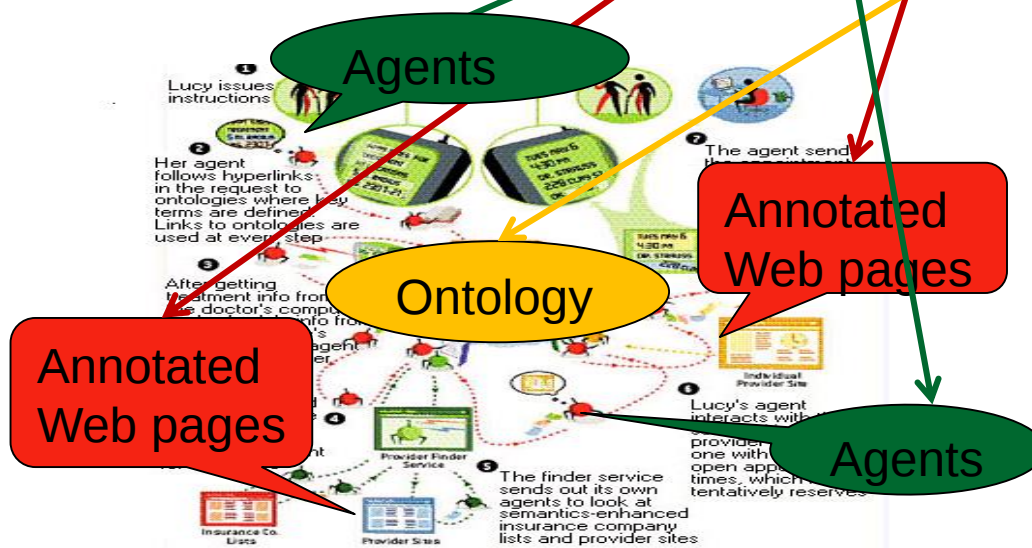
- 万维网是按“网页的地址”，而非“**内容的语义**”来定位信息资源的（缺少语义联接），网上所有信息都是由不同的网站发布的，相同主题的信息分散在全球众多不同的服务器上，又缺少有效工具能将不同来源的相关信息综合起来，因此形成了一个信息孤岛(其中有大量重复)，查找自己所需的信息非常困难。

语义网

■ 定义

The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in co-operation.

– Tim Berners-Lee, James Hendler, Ora Lassila, The Semantic Web, Scientific American, May 2001



语义网：Semantic Web

- 对未来网络的一个设想，是一种智能网络。
- 以Web数据的**内容**（即数据的语义）**为核心**，用机器能够理解和处理的方式链接起来的海量分布式数据库。
- 对现有Web增加了语义支持，它是现有万维网的延伸与变革，其目标是帮助机器在一定程度上理解Web信息的含义，使得高效的信息共享和机器智能协同成为可能。
- 语义网将会为用户提供动态、主动的服务，从而更便于机器和机器、人和机器之间的对话及协同工作。

语义网：困难和挑战

■ Ontology

- 支撑网络应用的大规模知识本体和知识库

■ 数据的语义标注和链接

- 为了更方便地处理 Web 数据，需要标注语义信息。但是，Web 数据量巨大，范围广泛，要求标记方法具备很强的扩展性。

■ 具有语义分析能力的搜索引擎

- 要让计算机能够处理Web 上人类生成的内容，就需要机器具有思考和推理的能力。这就需把数据集转变为知识系统，并赋予描述逻辑的能力。目前，在这个问题上语义网还存在相当的不足，在很多应用场景下需要研究表达能力更强的表示形式。

语义网刚推出的十年间，并没有在实际中大规模应用，但是这方面的成果大力推动了使用本体模型和形式化手段表达数据语义的方法的研究，为后续的知识图谱研究热潮奠定了基础。

知识图谱发展的推动力

- 以维基百科为核心的协同知识资源对于知识图谱的发展起到了决定性的作用。
 - 两个大规模通用领域知识图谱 Freebase 和 DBpedia 都是以 Wikipedia 的 Infobox 数据为基础构建而成的。
 - Freebase 更偏向于知识工程的技术路线：自顶向下
 - DBpedia 更偏向于语义网的技术路线：自底向上
- 真实应用需求是知识图谱发展的另一主要推动力。
 - 具有语义分析能力的信息搜索引擎能够理解人类的自然语言，并且具有一定的推理和判断能力，从而提供符合用户需求的信息服务。

知识图谱的发展

- 谷歌知识图谱：试图通过知识图谱对于网页中的文本内容，特别是客观类事实性文本内容中的语义信息进行刻画，从非结构化的文本内容中提取出实体及其关系，将文本内容转化为相互连接的图谱结构，从而让计算机真正做到内容理解。
- 百度知心：通过对搜索结果的筛选，将相近的信息组织在一起，以知识图谱的形式呈现出来，提高搜索的准确性。
- 搜狗知立方：通过搜集、挖掘互联网上的海量碎片信息并汇总成图谱形式，以更好地支撑语义搜索、智能问答的服务。

知识图谱的特点（一）

■ 知识图谱是人工智能应用不可或缺的基础资源

- 在语义搜索、问答系统、智能客服、个性化推荐等互联网应用中占有核心地位，在金融智能、商业智能、智慧医疗、智慧司法等领域具有广阔的应用前景。

■ 语义表达能力丰富，能够胜任当前知识服务

- 知识图谱源于语义网络，是一阶谓词逻辑的简化形式，并在实际应用中通过定义大量的概念和关系类型丰富了语义网络的内涵。
- 一方面，它能够描述概念、事实、规则等各个层次的认知知识；另一方面，它也能够有效组织和描述人类在自然环境和社会活动中形成的海量数据，从而为各类人工智能应用系统奠定知识基础。

知识图谱的特点（二）

■ 描述形式统一，便于不同类型知识的集成与融合

- 本体（Ontology）和分类系统（Taxonomy）是典型的知识描述载体，数据库是典型的实例数据载体，它们的描述形式截然不同。知识图谱以语义网的资源描述框架（Resource Description Framework，RDF）规范形式对知识描述和实例数据进行统一表示，并通过对齐、匹配等操作对异构知识进行集成和融合，从而支撑更丰富、更灵活的知识服务。

■ 表示方法对人类友好，给众包等方式编辑和构建知识提供了便利

- 传统知识表示方法和描述语言需要知识工程师具备一定的专业知识，普通人群难以操作。知识图谱以实体和实体关系为基础的简洁表示形式，无论是专家还是普通民众都容易接受，这给以众包等方式编辑和构建知识提供了便利，为大众参与大规模知识构建提供了低认知成本的保证。

知识图谱的特点（三）

■ 二元关系为基础的描述形式，便于知识的自动获取

- 知识图谱对各种类型知识采取统一的二元关系进行定义和描述，为基于自然语言处理和机器学习方法进行知识的自动获取提供便利，为大规模、跨领域、高覆盖的知识采集提供技术保障。

■ 表示方法对计算机友好，支持高效推理

- 推理是知识表示的重要目标，传统方法在进行知识推理时复杂度很高，难以快速有效地处理。知识图谱的表示形式以图结构为基础，结合图论相关算法的前沿技术，利用对节点和路径的遍历搜索，可以有效提高推理效率，极大降低计算机处理成本。

■ 基于图结构的数据格式，便于计算机系统的存储与检索

- 知识图谱以三元组为基础，使得在数据的标准化方面更容易推广，相应的工具更便于统一。结合图数据库技术以及语义网描述体系、标准和工具，为计算机系统对大规模知识系统的存储与检索提供技术保障。

狭义和广义知识图谱

- 从狭义上讲，知识图谱是一种知识描述语言，它是语义网络的思想在语义网上的具体实施，使得互联网知识元素以及知识元素之间的关系具有明确的语义和规范的描述，从而易于互联网知识资源的规模化建设，并互相链接，促进知识资源的共享，从而通过知识推理与知识计算赋能各种互联网服务。
- 从广义上讲，知识图谱是一个包括知识表示、知识构建、知识维护以及知识应用的完整生态系统，它不仅包含了特定领域中的知识定义和实例数据，还包含了支撑描述、构建、存储、管理和应用知识所需的配套标准、技术和工具。

提纲

- 什么是知识图谱
- 知识图谱发展历程
- 知识图谱的类型和代表性知识图谱
- 知识图谱的生命周期
- 知识图谱和深度学习
- 课程介绍

知识的分类

- 陈述性知识和过程性知识
- 事实性（或者客观性）知识和主观性知识
- 静态知识和动态知识
- 百科知识、领域知识、场景知识、语言知识、常识知识等
 - 百科知识：涵盖各个行业、领域的通用知识，例如：人物、机构、地点等。
 - 领域知识：某个领域内特有的知识，例如：法律知识、金融知识等。
 - 场景知识：某个特定场景下或者需要完成某项任务时所需要的知识，例如：在订机票过程中需要提供的信息；盖房子的步骤等。
 - 语言知识：那些语言层面的知识，例如：Microsoft 的缩写是 MS；乔丹和佐敦具有同指关系；减肥和瘦身是同义词等。
 - 常识知识：那些大家都认可的知识，例如：狗有四条腿、鸟会飞等。

知识图谱的类型

- 语言知识图谱：主要是存储人类语言方面的知识，其中比较典型是英文词汇知识图谱 WordNet，它由同义词集和描述同义词集之间的关系构成。
- 常识知识图谱：主要有 Cyc 和 ConceptNet 等。其中 Cyc 由大量实体和关系以及支持推理的常识规则构成；ConceptNet 由大量概念以及描述它们之间关系的常识构成。
- 语言认知知识图谱：中文知网词库 HowNet 是一种典型的语言认知知识图谱（语言认知知识与常识知识区别不大，因为语言是人类表达和交换信息的主要载体），HowNet 致力于描述认知世界中人们对词语概念的理解，基于词语义原，揭示词语的更小语义单元的含义。
- 领域知识图谱：针对特定领域构建的知识图谱，专门为特定的领域服务，例如：医学知识图谱 SIDER (Side Effect Resource)，电影知识图谱 IMDB (Internet Movie Database)，音乐知识图谱 MusicBrainz 等，这些知识图谱在各自的领域都有着广泛的应用。
- 百科知识图谱：主要以 Linked Open Data (LOD) 项目支持的开放知识图谱为核心，主要有 Freebase、DBpedia、YAGO 和 Wikidata 等，它们在信息检索、问答系统等任务中有着重要应用。

代表性的知识图谱

名称	开始时间	依赖资源	规模 #(实体 / 概念 / 关系 / 事实)
Cyc ⁵	1984	专家知识	239,261 / 116,822 / 18,014 / 2,093,000
WordNet ⁶	1985	专家知识	155,287 / 117,659 / 18 / -
ConceptNet ⁷	1999	群体智能 (多语言)	- / 8,000,000 / 36 / 21,000,000
YAGO ⁸	2007	WordNet + Wikipedia	4,595,906 / 488,469 / 77 / $\approx 40m$
DBpedia ⁹	2007	Wikipedia + 专家知识	17,315,785 / 754 / 2843 / 79,030,098
Freebase ¹⁰	2008	Wikipedia + 领域知识 + 群体智能	58,726,427 / 2,209 / 39,151 / 3,197,653,841
NELL ¹¹	2010	机器学习	- / 287 / 327 / 2,309,095
BabelNet ¹²	2012	WordNet + Wikipedia(多语言)	9,671,518 / 6,117,108 / 1,307,706,673 / -
WikiData ¹³	2012	Freebase + 群体智能	45,766,755 / - / - / -
Google Knowledge Graph ¹⁴	2012	基于 Freebase	570M / 1500 / 35000 / 18000M
Knowledge Vault ¹⁵	2014	机器学习	45M / 1100 / 4469 / 271M

- Cyc 是一个通用的常识知识库，始建于 1984 年，其目的是将上百万条知识编码为机器可处理的形式，并在此基础上实现知识推理等智能信息处理任务。
- Cyc 中不仅包含了大量实体和关系，还包含用于推理的常识规则，并提供多种推理引擎，支持演绎推理和归纳推理，同时也提供扩展推理机制的模块。
- Cyc 项目试图用表达能力受限的一阶谓词对自然语言表达的知识进行描述，由于自然语言本身的复杂性和歧义性，难以完整地自然语言中蕴含的包括隐含关系在内的各种复杂知识进行表示。

Cyc



- Cyc基本上通过人工构建，项目负责人 Lenat 估计整个Cyc 需要 350 人年的时间才能够完成。
- 目前 Cyc 包含 50 万实体，接近 3 万个关系以及 5 百万事实。
- OpenCyc是 Cyc 项目开放出来免费供大众使用的一个子集，目前 OpenCyc 包含 24 万实体，2 百万事实。
- 近年来，Cyc 也开始使用一些自动构建的方法从非结构化文本中自动抽取知识。
- 从 2008 年开始，Cyc 开始将其资源与Wikipedia、DBpedia 等资源进行关联，建立它们之间的链接。

The Cyc Ontology and Content

Tens of thousands of Predicates

between, owns, isa, beliefs

Hundreds of thousands of Concepts



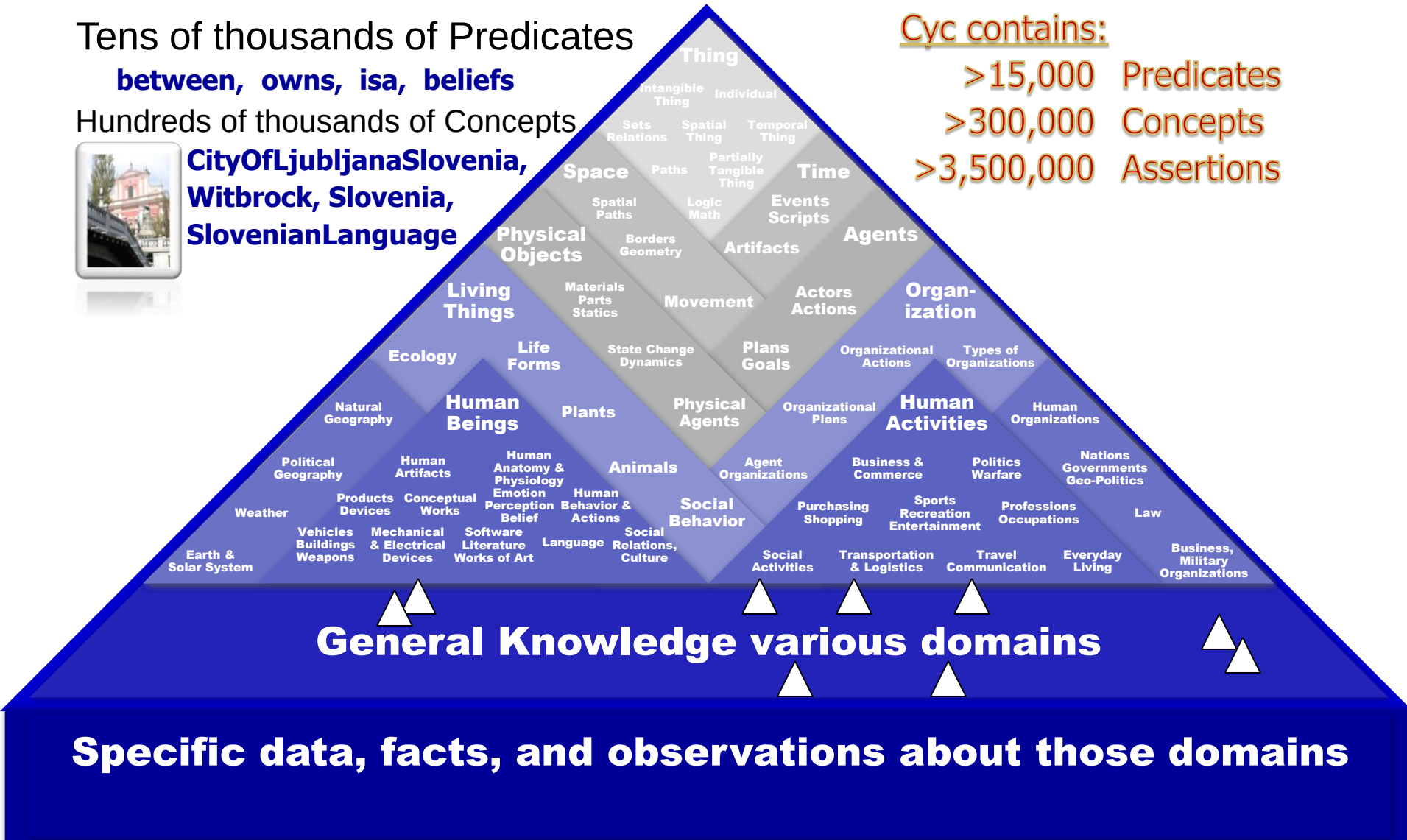
**CityOfLjubljanaSlovenia,
Witbrock, Slovenia,
SlovenianLanguage**

Cyc contains:

>15,000 Predicates

>300,000 Concepts

>3,500,000 Assertions

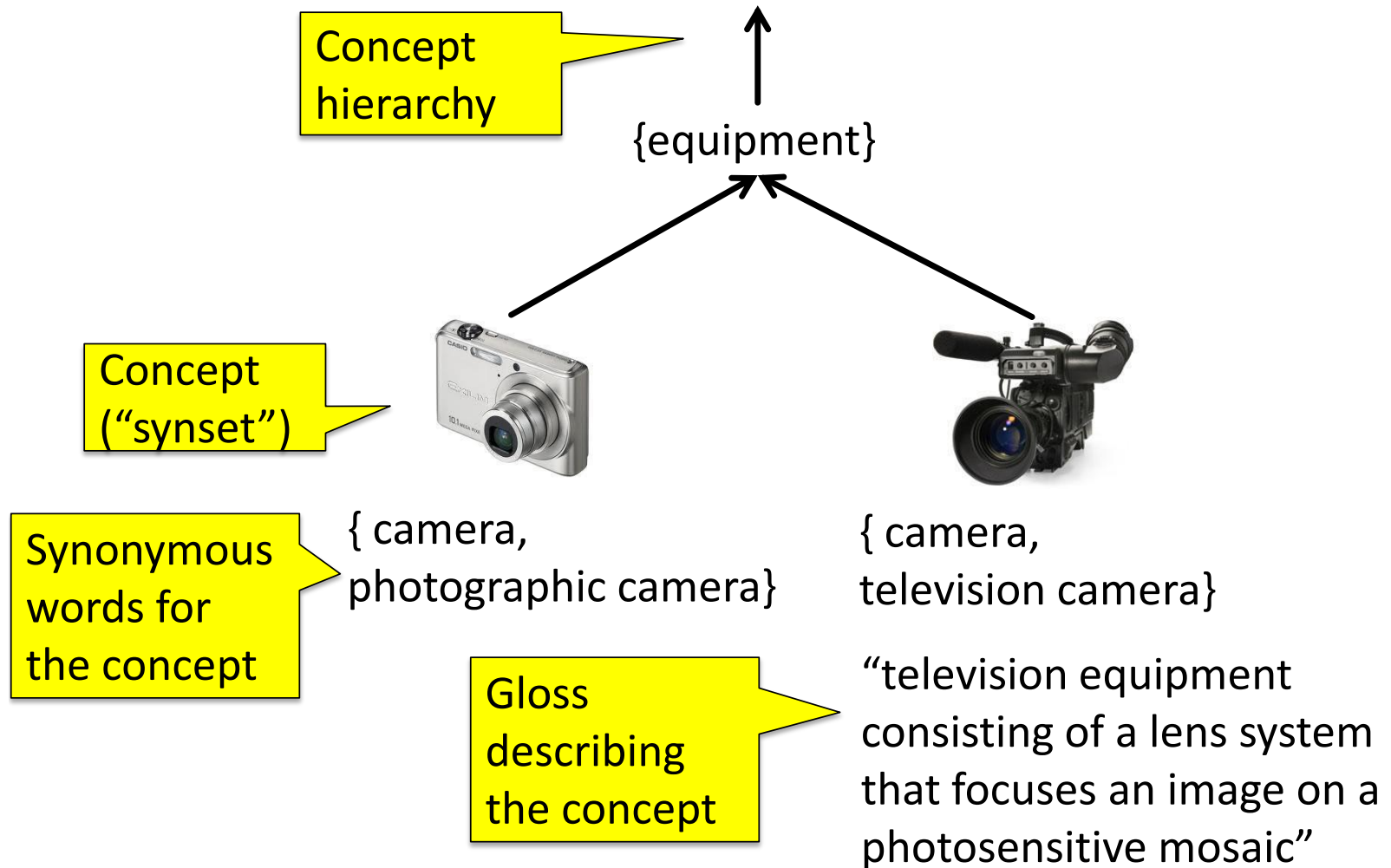


WordNet

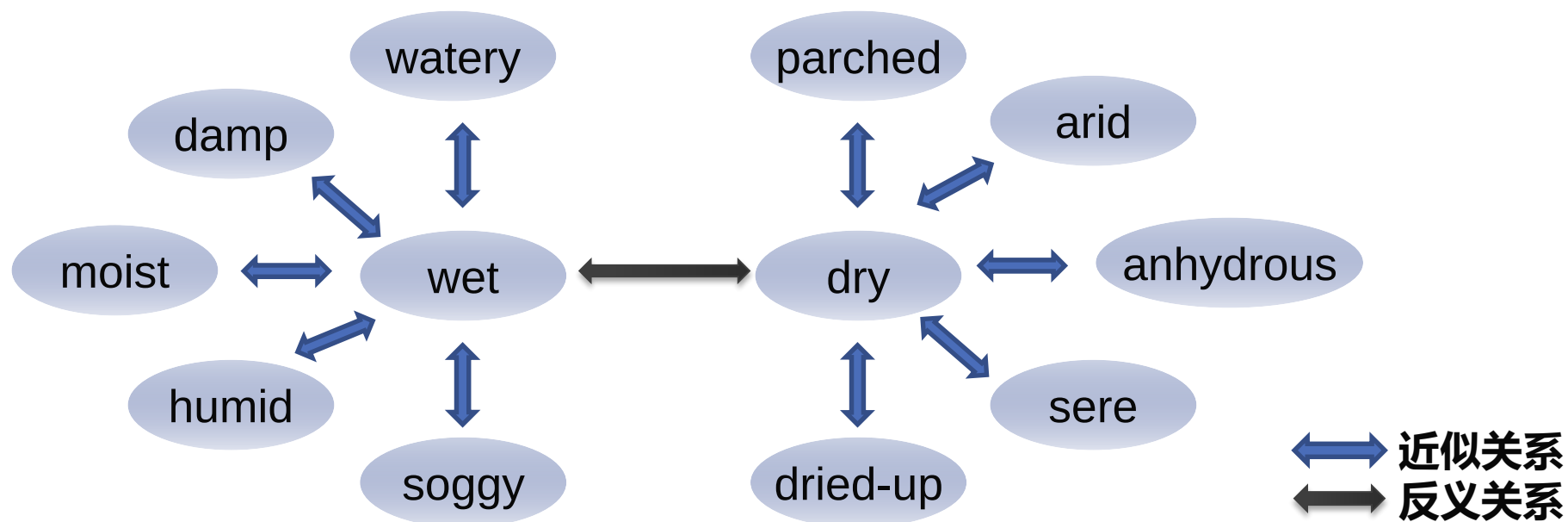
WordNet
A lexical database for English

- WordNet 是 1985 年由 Princeton 大学公布的一个英文电子词典和本体。采用人工标注的方法，将英文单词按照单词的语义组成一个大的概念网络。
- 词语被聚类成同义词集 (Synset)，每个同义词集表示一个基本的词汇语义概念。
- 词集之间的语义关系包括同义关系、反义关系、上位关系、下位关系、整体关系、部分关系、蕴含关系、因果关系、近似关系等。
- 1991 年，WordNet1.0 版本正式公布，目前 WordNet 包含 146,350 个单词, 111,223 个同义词集。

WordNet组织示例：上下位、同义



WordNet组织示例：近义、反义



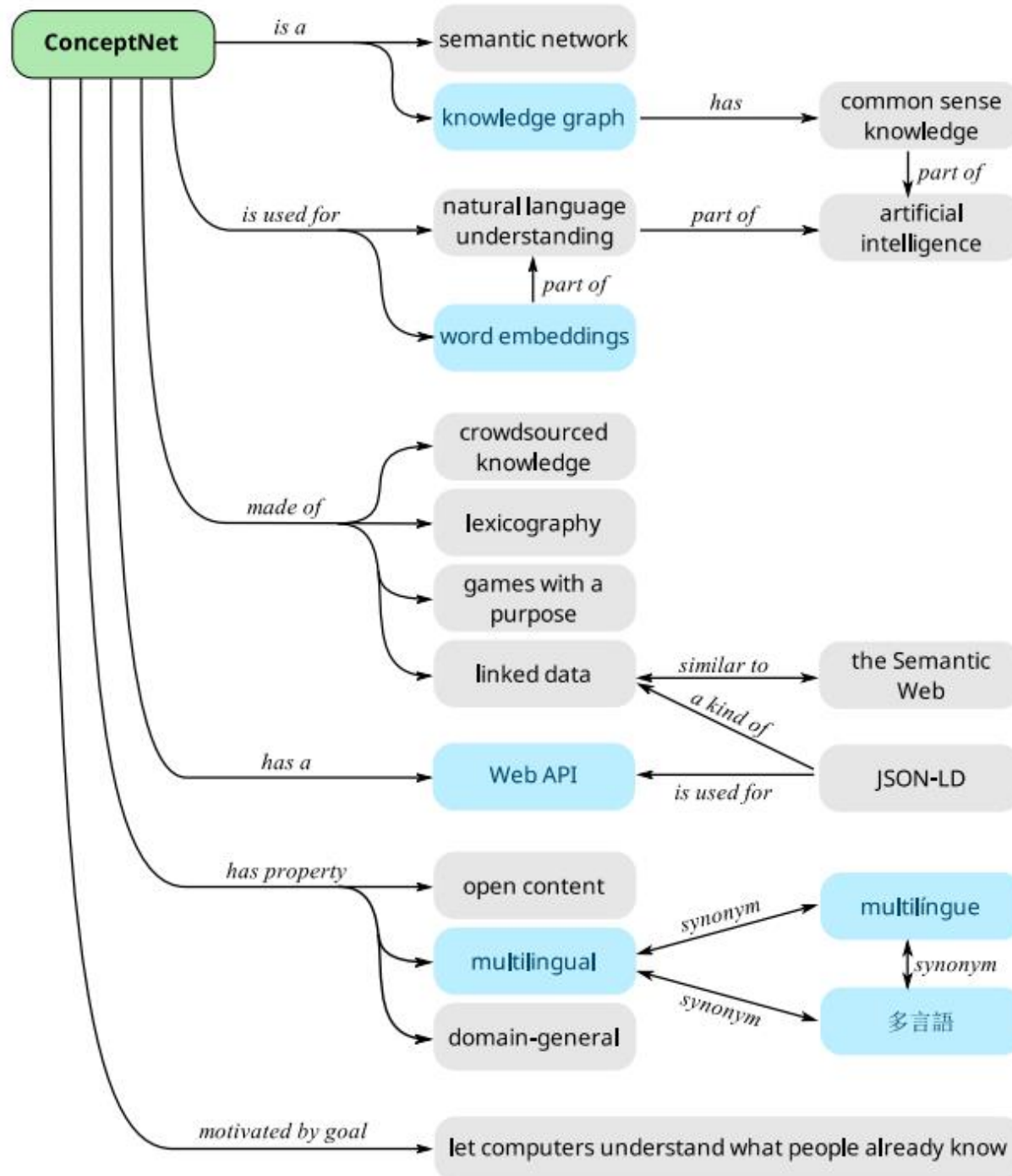
基于反义、近义组织的形容词synset

知网：HowNet

- 知网（英文名称为 HowNet）是由董振东教授主持开发的一个语言认知知识库/常识知识库。
- 以概念为中心，基于义原描述了概念与概念之间以及概念所具有的属性之间的关系，每一个概念可以由多种语言的词汇进行描述（主要是中文和英语）。
- 目前知网总共包含了 800多个义原，11,000 个词语。

ConceptNet

- ConceptNet 是一个开放的、多语言的知识图谱。最早起源于 MITMedia Lab 的一个众包项目：Open Mind Common Sense，致力于帮助计算机理解人们日常使用的单词的意义。
- ConceptNet 主要是由大量概念以及描述它们之间关系的常识构成，支持多种语言，可以用于自然语言处理等多种人工智能应用。



Wikipedia

■ 免费的在线百科全书

- 2001年开始
- crowdsourcing的方式构建
- 目标：构建全世界最大的百科全书



高质量数据源
500万概念
多语言
富含丰富语义结构的文档：
Infobox, table, list, category...

基于Wikipedia的知识库

Wikipedia: 文档结构

标题=概念

唐太宗

唐太宗李世民（598年1月28日－649年7月10日^{[1][2][3]}），[中国唐朝](#)第二任皇帝。[祖籍陇西郡成纪县](#)（今[甘肃省天水市秦安县](#)北），生于陕西[武功县](#)，626年至649年在位。父亲是[唐高祖](#)李渊，母亲是窦皇后

概念文本描述

Infobox:以(属性, 值)对形式呈现的信息表格

每个页面有多个类别，类别组成Taxonomy

•分类: [598年出生](#), [649年逝世](#) [唐朝皇帝](#)

唐太宗



概要

姓名 [李世民](#)

庙号 [太宗](#)

谥号 [文皇帝](#)（649年初谥）

[文武圣皇帝](#)（674年加谥）

[文武大圣皇帝](#)（749年加谥）

[文武大圣大广孝皇帝](#)（754年加谥）

陵墓 [昭陵](#)

政权 [唐朝](#)

在世 [598年1月28日](#)－[649年7月10日](#)（52岁）

在位 [626年9月4日](#)－[649年7月10日](#)

年号 [贞观](#)：627年－649年

基于Wikipedia的知识库

■ 基于Wikipedia的知识库都基于几乎相同的思路：

- 从Wikipedia丰富的半结构化信息中挖掘知识

- 包括：Infobox, Category, 超链接, Table, List...

■ 不同之处在于

- 如何处理有歧义的属性映射

- 如何构建知识库的Taxonomy

在世	598年1月28日 - 649年7月10日 (52岁)
----	-----------------------------

出生	隋文帝仁寿四年604年
----	-------------



BirthDate (李世民, 598年)

BirthDate (李元吉, 604年)

基于Wikipedia的知识库具有相同的模型

- 一个知识库包含一个集合的实体
 - 猫王、李世民、赵高、唐朝、分封制...
- 实体被划分到不同的类别中
 - 歌手(猫王), 皇帝(李世民), 朝代(唐朝), 制度(分封制)
- 类别通过上下位等关系相互关联
 - SubClassOf(歌手, 人), SubClassOf(皇帝, 人)
- 实体和类别都通过属性和相互之间的关系来描述
 - BirthDate (李世民, 598年), Has(歌手, 歌曲)
- 关系可以通过蕴含关系来进行推理
 - 歌曲 → 作品, 收购 → 持有

DBPedia

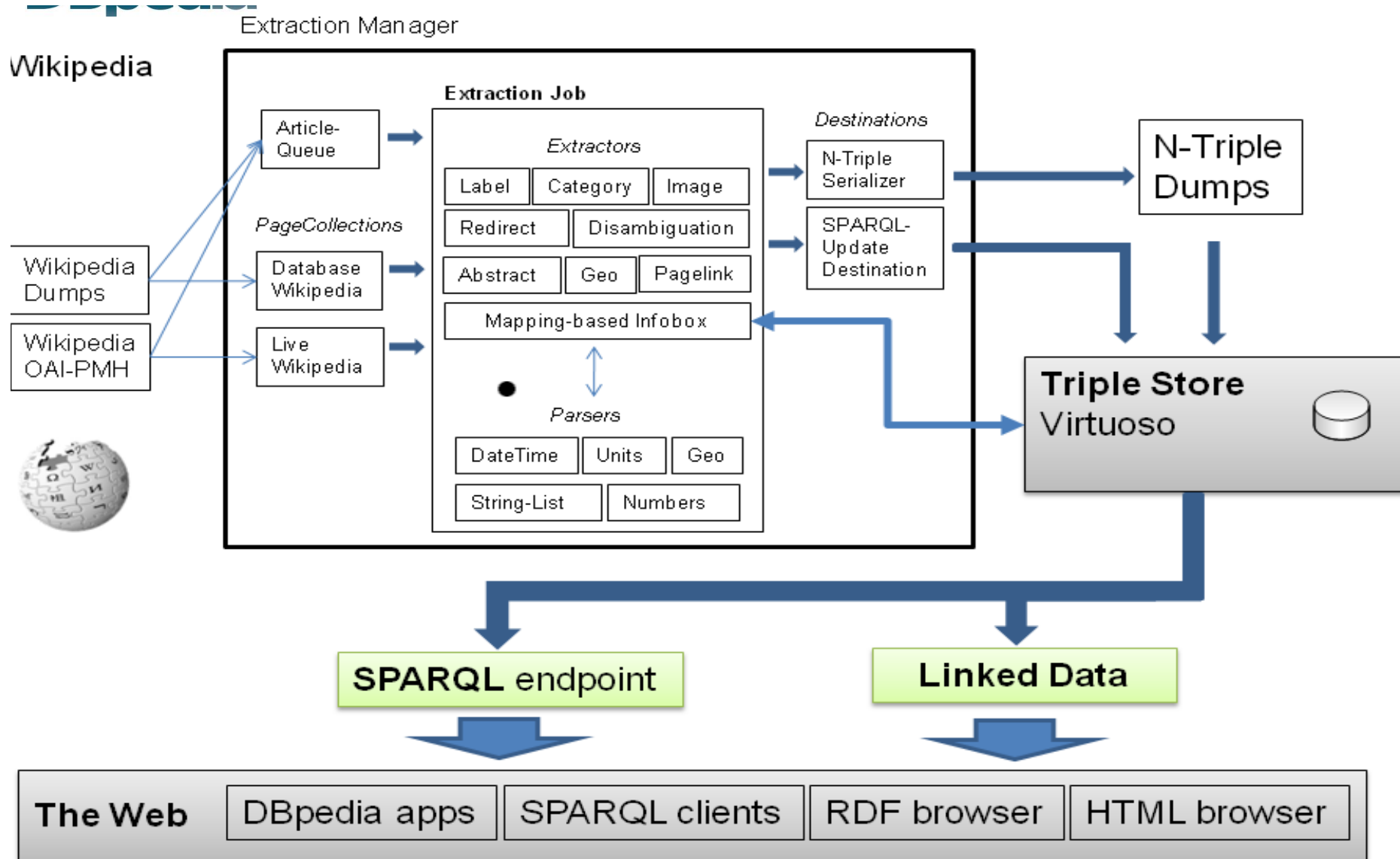
- 2007年开始，其主要目标是构建一个社区，通过社区成员来**定义和撰写准确的抽取模板，从维基百科中抽取结构信息，并将其发布到Web上**
- DBPedia 总共包含 95 亿事实三元组
 - 13 亿数据抽取自英文版维基百科
 - 50 亿数据抽取自其他语言的维基百科
 - 32 亿抽取自 Wikidata 数据



Dbpedia的抽取方法

- 社区通过人工的方式构建了Ontology
 - 280个类别
 - 覆盖约50%的维基百科实体
- DIEF - DBpedia Information Extraction Framework
 - 目标：抽取Wikipedia中的结构化信息
 - 方法：基于属性mapping的Infobox抽取, Raw Infobox Extraction, Feature Extraction, Statistical Extraction
 - 编程语言：Scala & Java
- DBpediaLive：持续保持与Wikipedia的同步
 - 2013年六月，英语维基百科有将近330万次编辑(每分钟越77次)

抽取框架图



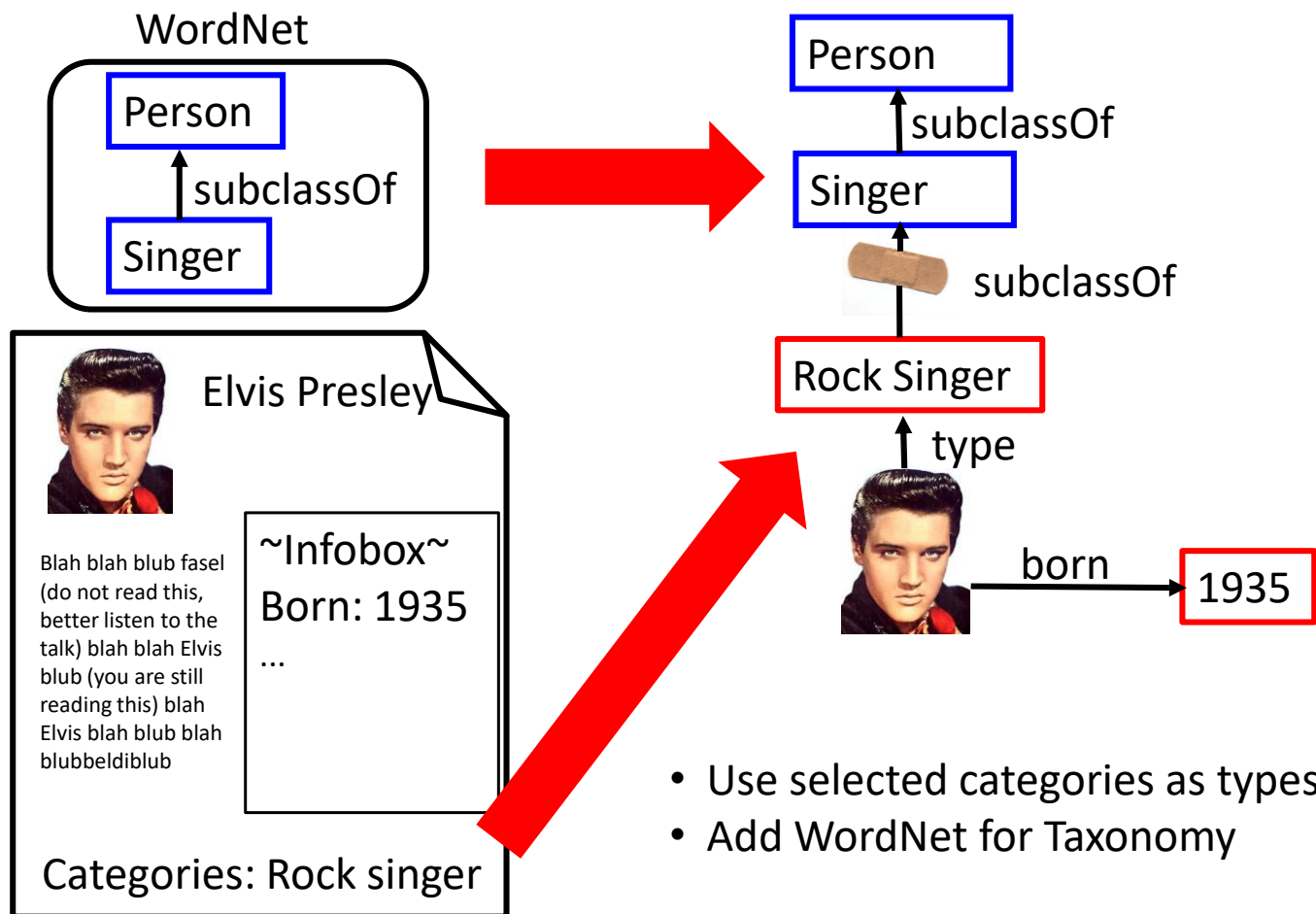
Yago

- 德国马普研究所从2007年开始的一个项目
- Yago基于WordNet的知识体系，将 Wikipedia中的类别与 WordNet 中的类别进行关联，同时将Wikipedia 中的条目挂载到 WordNet 的体系下。
- 通过语言本体和世界知识的融合，一方面扩充了语言知识库，另一方面对海量的世界知识进行了组织和整理。



Yago Ontology构建

- 使用WordNet的Ontology作为基础
- 将Wikipedia中的类别加入到WordNet中



Yago的抽取方法

- Yago的实体关系数据都依赖人工严格定义的规则从 Wikipedia 中自动抽取产生，包括实体之间的上下位 (Is-A) 和实体属性等关系。
- 目前在 YAGO 中，关系实例数据的准确率超过 95%。
- 整个知识库使用 RDFs (RDF Schema) 语言与 OWL (OntologyWeb Language) 语言描述，构成一个具有清晰完整逻辑定义的知识系统。
- YAGO2在 2017 年人工智能国际顶级学术会议 IJCAI2017 上获得由 Artificial Intelligence Journal(AIJ) 颁发的卓越论文奖(Prominent Paper Award) 。

BabelNet

- BabelNet是一个多语言词汇级的语义网络和本体，由罗马萨皮恩扎大学计算机科学系计算语言学实验室所创建。
- 与YAGO类似，BabelNet 的主要特点也是将维基百科链接到 WordNet 上。但是 BabelNet 加入了多语言支持，目前覆盖了 271种语言，包括全部的欧洲语言、大多数亚洲语言及拉丁语。
- 提供了多种语言的概念和实体，以及它们之间丰富的语义关系。BabelNet 包含了大约 1400 万个同义词集和 7.46 亿个词义。
- BabelNet 中所定义的语义关系主要来源于：
 - WordNet 中所定义的语义关系，如上下位关系、部分整体关系、反义、同义关系等，总共大约 36.4 万条关系；
 - 维基百科中非特定的相关关系，例如，国籍、首都等，总共大约 3.8 亿条关系。
- BabelNet 同 YAGO一起，在IJCAI2017上获得卓越论文奖。

Freebase

- Metaweb公司2000年开始构建，2010年被Google收购
- 从Wikipedia和其他数据源（如IMDB、MusicBrainz）中导入知识
- 核心理念：
 - 在Wikipedia中，人们编辑文章
 - **在Freebase中，人们编辑结构化知识**



Freebase



- Freebase 是基于维基百科、使用群体智能方法建立的结构化知识资源。
- Freebase 是第一个尝试利用协同智慧构建完全结构化知识图谱的系统。
- Freebase是公开可获取的规模最大的知识图谱之一
 - 包含 4726 万实体、19亿个实体关系三元组
- 2010年被谷歌收购并纳入到谷歌知识图谱中。
- 2015 年，谷歌关闭了 Freebase，并把数据全部迁移到 Wikidata。

用户是Freebase知识构建的核心

■ 编辑实体

- 创建实体
- 将实体分到类别
- 增加/修改属性/关系
- 上传图片

■ 编辑Schema

- 定义新类别
- 定义类别的属性

■ Review

- 验证知识准确性
- 投票
- 删除错误知识

■ DataGame

- 寻找别名
- 抽取事件日期
- 使用Yahoo图片搜索加入图片

用户在Freebase中的作用

KnowItAll, TextRunner & ReVerb

- 华盛顿大学图灵中心的 KnowItAll项目
- 目标是让机器自动阅读互联网上的文本内容，从大量非结构化的文本中抽取结构化的实体关系三元组信息
- 区别于传统的文本信息抽取系统，这里要抽取的关系不再是预定义的，也就是说知识抽取的范围是开放域文本。
- TextRunner和Reverb系统是 KnowItAll 项目中的两个代表系统
 - 致力于从文本中通过识别句子的谓词抽取所有的二元关系
 - 都利用了网络数据的冗余信息，对初步认定可信的信息进行评估。
 - 可以在系统中直接查询知识（实体关系三元组）

KnowItAll, TextRunner & ReVerb的知识查询



Open Information Extraction

Hosted by



Created at



Example Queries: ⓘ
[What kills bacteria?](#)
[Who built the Pyramids?](#)
[What did Thomas Edison invent?](#)
[What contains antioxidants?](#)

Typed Example Queries: ⓘ
[What countries are located in Africa?](#)
[What actors starred in which films?](#)
[What is the symbol of which country?](#)
[What foods are grown in which countries?](#)
[What drug ingredients has the FDA approved?](#)

Argument 1:

Relation:

Argument 2:

Corpus:

- NELL是卡内基梅隆大学基于Read the Web项目开发的一套“永不停歇的语言学习”系统。
 - Never-Ending Language Learning
- NELL本身是一套语言学习系统，每天不间断地执行两项任务：阅读和学习。
 - 阅读任务从Web文本中获取知识，并添加到内部知识库；
 - 学习任务目标是提升机器学习算法的性能
- NELL 可以抽取大量的事实，并标注所抽取的迭代轮数、时间及系统置信度，供人工进行校验。
- 系统从 2010 年开始学习，经过半年的学习和抽取，总共得到将近 35 万的实体关系三元组，经过人工少量矫正和标注，在抽取更多事实的前提下，知识抽取的准确率可以达到 87%。
- NELL 目前还需要进一步完善学习算法来消除错误信息，使其能够更加有效地完成对知识的学习、总结和归纳。

抽取结果示例

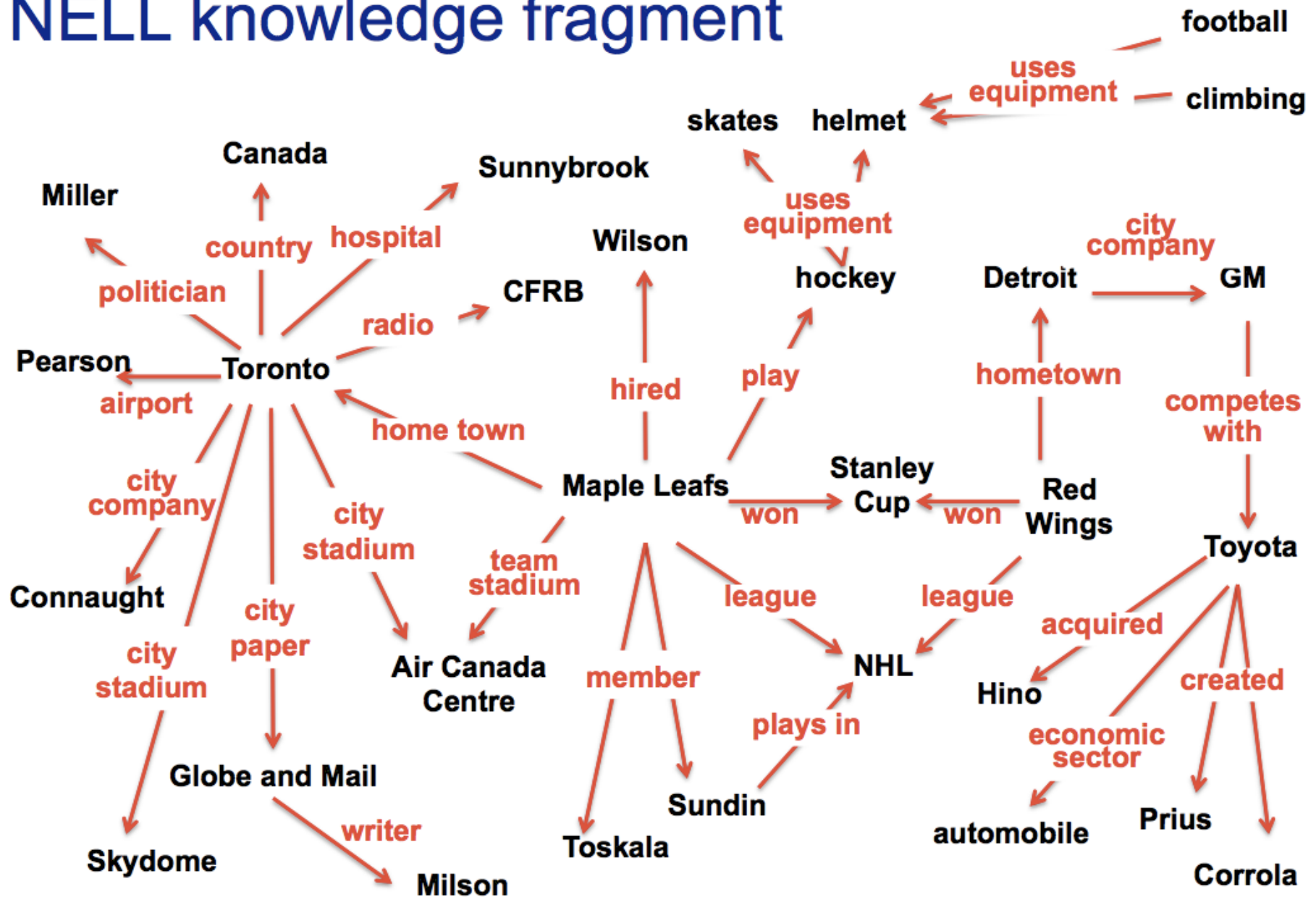
Recently-Learned Facts

[Refresh](#)

instance	iteration	date learned	confidence		
<u>rillito river</u> is a <u>river</u>	1064	26-jun-2017	99.9		
<u>james dexter</u> is a <u>CEO</u>	1064	26-jun-2017	99.7		
<u>richard thompson</u> is a <u>Mexican person</u>	1064	26-jun-2017	100.0		
<u>htc droid eris</u> is a <u>consumer electronic device</u>	1064	26-jun-2017	92.8		
<u>radiant snow</u> is a <u>weather phenomenon</u>	1066	15-jul-2017	99.8		
<u>the london eye001</u> is a tourist attraction <u>in the city london</u>	1069	03-aug-2017	100.0		
<u>state university</u> is a sports team <u>also known as michigan state university</u>	1067	21-jul-2017	100.0		
<u>state university</u> is an organization <u>also known as clemson</u>	1066	15-jul-2017	96.9		
<u>irene kirkaldy belongs to</u> the religion <u>seven day adventist</u>	1069	03-aug-2017	100.0		
<u>kcal</u> is a <u>TV station in</u> the city <u>los banos</u>	1069	03-aug-2017	100.0		

<http://rtw.ml.cmu.edu>

NELL knowledge fragment



Knowledge Vault

- Google 于 2014 年创建的一个大规模知识图谱。
- 相较于 Google 之前基于 Freebase 的知识图谱版本，Knowledge Vault 不再采用众包的方式进行图谱构建，而是试图通过算法自动搜集网上信息，通过机器学习方法对已有的结构化数据进行集成和融合，将其变成可用知识。
 - 集成和融合YAGO、Freebase、网页中的表格数据等
- 目前，Knowledge Vault 已经收集了 16 亿个事实
 - 其中，2.71 亿事实具有高置信度，其准确率大致在 90% 左右。

通用知识图谱 vs. 领域知识图谱



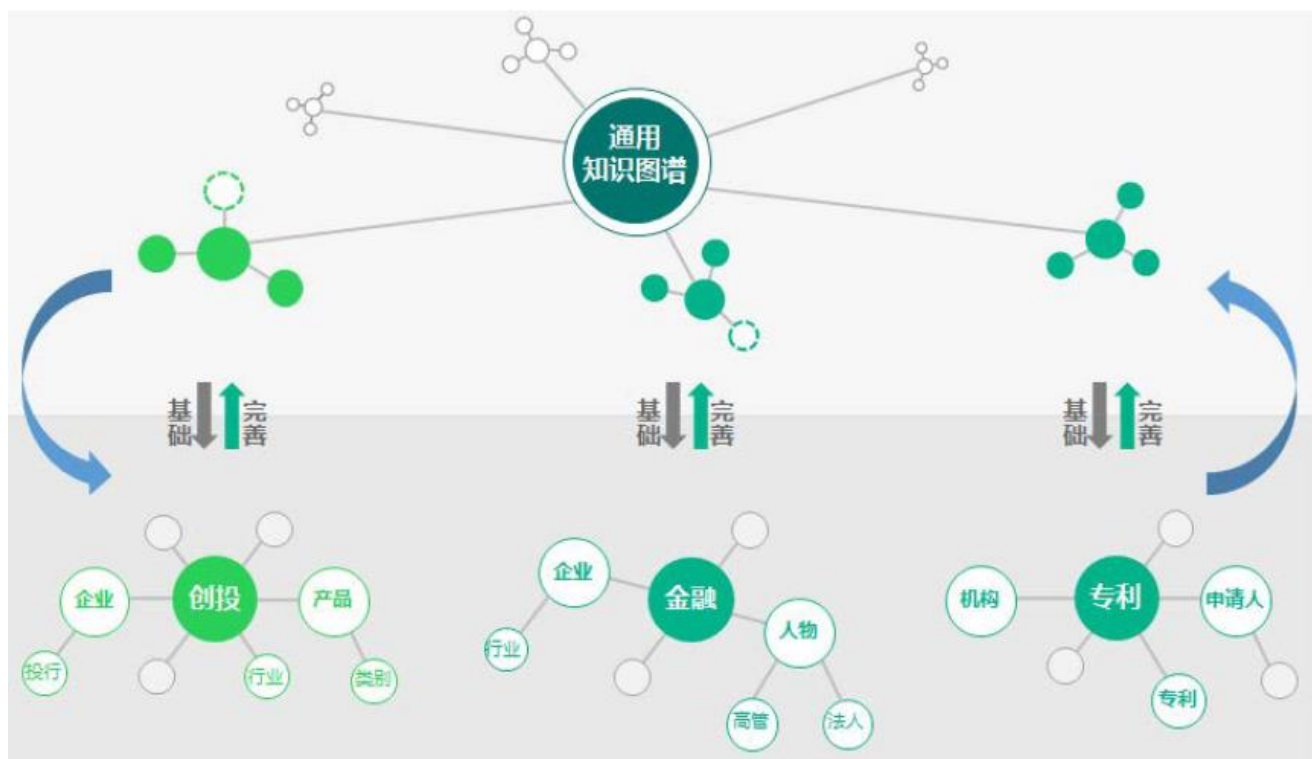
- 面向通用领域
- 以常识性知识为主
- 结构化的百科知识
- 强调知识的广度
- 使用者是普通用户



- 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 + 领域知识图谱

- 通用知识图谱的广度，领域知识图谱的深度，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。



领域知识图谱

■ 影视领域知识图谱 IMDB

- Internet Movie Database
- 关于电影、电视节目、演员和电影制作的知识图谱
- 截止到 2012 年 2 月, IMDB 共收录了 2,132,383 部作品资料以及 4,530,159 名人物资料。

■ 音乐领域知识图谱 MusicBrainZ

- 致力于收藏音乐数据, 有许多著名的音乐服务网站的数据均来自 MusicBrainZ
- 用户注册后即可使用并且向其中增加数据

■ 医疗卫生领域知识图谱SIDER

- Side Effect Resource
- 主要包含了已经上市的药物及其记录在册的副作用, 这些数据可以为相关疾病的治疗提供依据。

企业知识图谱： 阿里电商知识图谱

- 以商品、产品、品牌、条码、分类为核心
- 整合关联了百科、国家标准等9大类一级本体，包含了**百亿级别**的三元组，形成了巨大的知识图谱

女装男装

潮流女装	羽绒服	毛呢大衣	毛衣	冬季外套
	新品	裤子	连衣裙	腔调
羽绒服	轻薄款	长款	短款	毛领
	加厚	裤子	麂绒	新品
呢外套	廓形	双面呢	羊绒	中长款
	短款	毛领	设计师款	系带
毛衣	马海毛	貂绒	羊绒	羊手
	开衫	中长款	短款	
外套上衣	外套	套装	风衣	
	真皮皮衣	马甲	小西装	

时尚男装

时尚男装 秋冬新品 淘特莱斯 淘先生 拾货
秋冬外套 时尚套装 潮牌 爸爸装

秋外：

秋外
衬衫

1688

秋
衫
套
装

里十裤子

男十裤子 休闲裤 工装裤 运动裤 长裤

性感內衣

性感内衣 春新品 性感诱惑 甜美清新 简约优雅
奢华高贵 运动风 塑身 基础内衣

文胸

文胸 无钢圈 无痕文胸 蕾丝内衣 运动文胸
聚拢文胸 大码文胸 抹胸式 隐形

家居服

家居服 睡衣套装 睡裙 睡袍浴袍 外穿家居
女士睡衣 男士睡衣 情侣睡衣 亲子睡衣

内裤

内裤 女士内裤 男士内裤 三角裤 平角裤
丁字裤 阿罗裤 早期裤 低腰

練

袜 船袜 男人袜 连裤袜 隐形袜



母婴用品

宝宝奶粉	英国牛栏	英国爱他	美赞臣	雅培
	澳洲爱他	可瑞康	惠氏	贝因美
辅食营养	米粉	肉松	磨牙棒	果泥
	益生菌	清火开胃	钙铁锌	维生素
童装	T恤	连衣裙	泳装	套装
	衬衫	防晒服	半身裙	短裤
玩具	沙滩戏水	早教启蒙	拼搭益智	遥控模型
	运动户外	学习爱好	卡通公仔	亲子互动

嬰童

理发器
安全座椅

纸尿

纸尿裤 花王 moon 大王 帮宝适
雀氏 好奇 妈咪宝贝 安儿乐

童鞋

童鞋	凉鞋	沙滩鞋	洞洞鞋	网鞋
	学步鞋	拖鞋	帆布鞋	宝宝鞋

童车

童车	电动车	自行车	学步车	手推车
	三轮车	滑板车	扭扭车	儿童轮滑

孕产必备

孕产必备	内衣	内裤	喂奶枕	收腹带
	妈咪包	待产包	防辐射服	储奶袋

海外直邮

海外直邮 海淘奶粉 海淘辅食 海淘营养 直邮花王
海淘洗护 海淘奶瓶 海淘餐具 海淘孕产

亲子鞋服

亲子鞋服 母女裙 父子装 亲子T恤 亲子衬衫
亲子套装 母女鞋 父子鞋 家庭鞋

早教启蒙

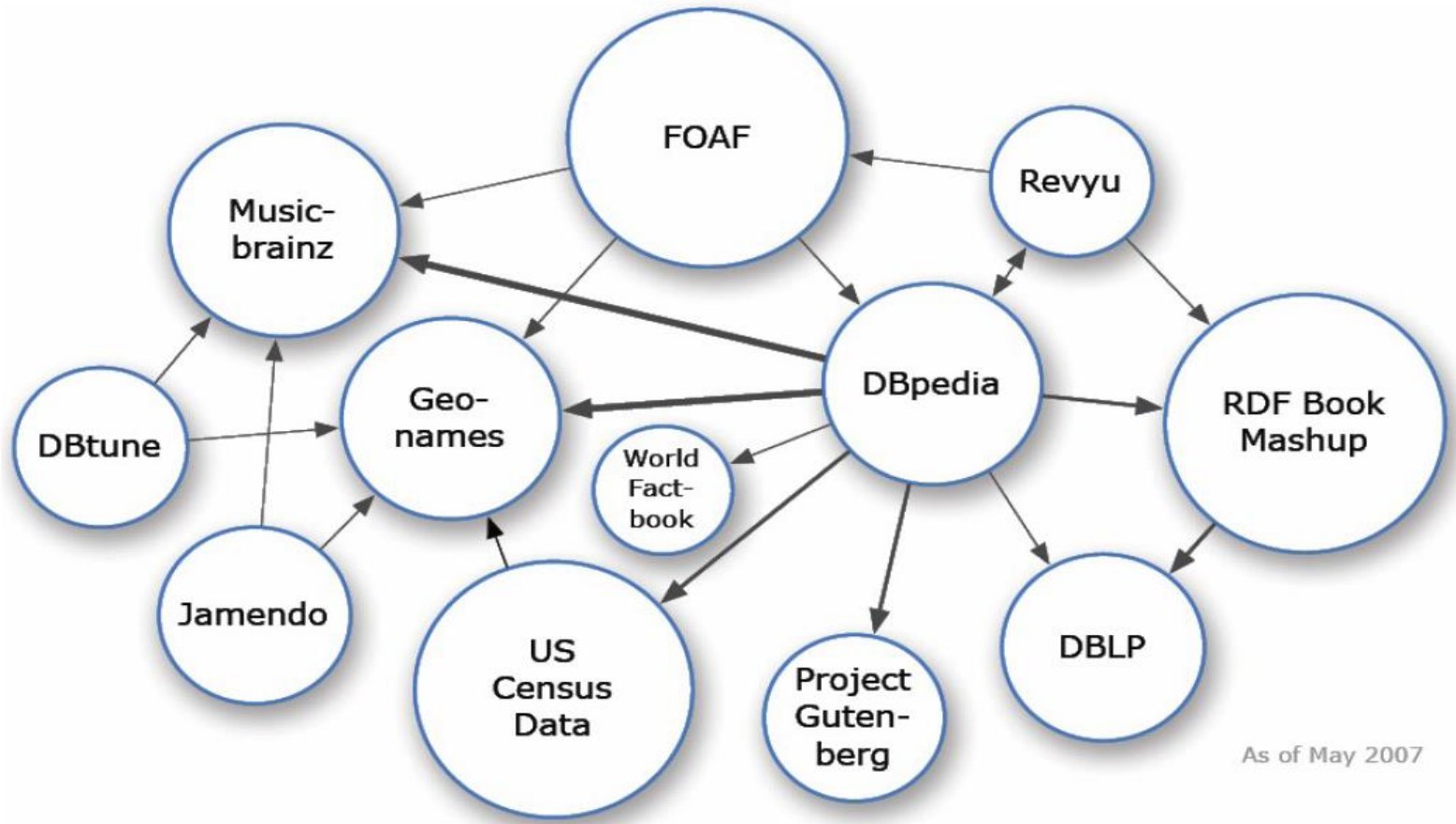
早教启蒙 早教机 点读机 健身架 布书
串/绕珠 床/摇铃 爬行垫 木质拼图

- 应用于搜索、前端导购、平台治理、智能问答、品牌商运营等核心、创新业务

Linked Open Data(LOD)

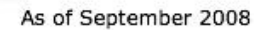
- 国际万维网组织 W3C 于 2007 年发起的开放互联数据项目
- 目的是将由互联文档组成的网络 (Web of Documents) 扩展成为由互联数据组成的全球数据及知识共享平台 (Web of Data)
- LOD的目标是：以 RDF 语义形式在 Web 上发布各种开放数据集，并在不同来源的数据项之间建立语义链接，从而增强 Web 上的数据共享，最终实现语义 Web 知识资源的开放共享。
- LOD 项目发布了 570 余个数据集，包含上千亿 RDF 三元组，并且随着链接数据的推广和 LOD 项目的发展，将有越来越多的数据以链接数据的形式发布在万维网上。
- LOD 为实现网络环境下的知识发布、互联、共享和服务提供了创新技术，为智能搜索、知识问答和语义集成提供了创新源动力。

Linked data in May 2007

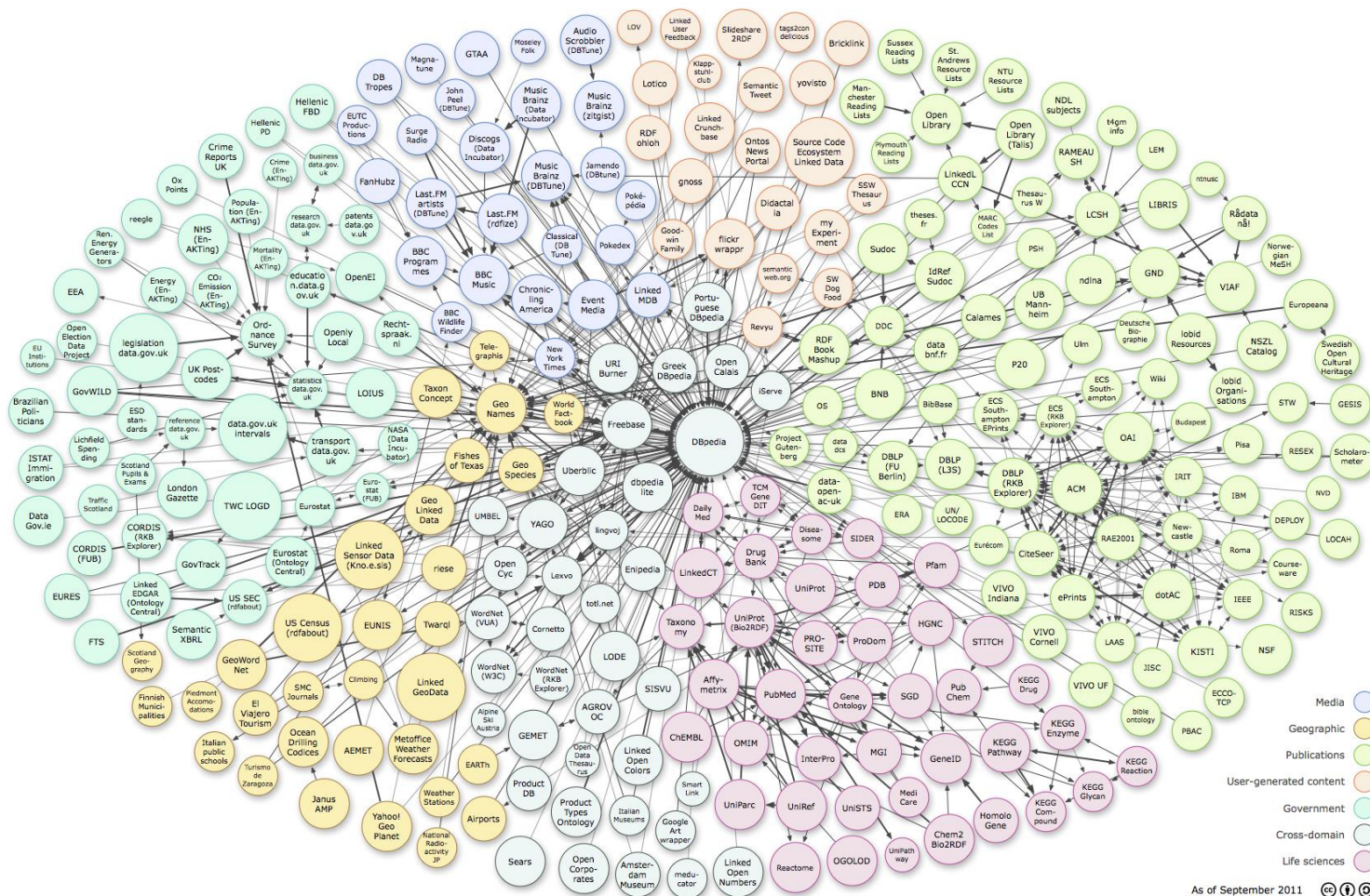


Over 500M RDF triples

Around 120K RDF links between data sources



Linked data in Sep. 2014



提纲

- 什么是知识图谱
- 知识图谱发展历程
- 知识图谱的类型和代表性知识图谱
- 知识图谱的生命周期
- 知识图谱和深度学习
- 课程介绍

知识图谱的生命周期

■ 知识体系构建

- 也叫知识建模
- 建模领域知识结构

■ 知识获取和验证

- 获取领域内的事实知识
- 估计知识的可信度

■ 知识融合

- 也叫知识集成
- 将碎片知识组装成知识网络

■ 知识存储和查询

- 采用何种方式对知识图谱进行存储

■ 知识推理

- 发现已有知识中隐含的知识

■ 知识应用

- 提供高性能知识服务

生命周期：知识体系构建

■ 知识建模

■ 指采用什么样的方式表达知识，其核心是构建一个本体对目标知识进行描述。

- 在这个本体中需要定义出知识的类别体系
- 每个类别下所属的概念和实体
- 某类概念和实体所具有的属性以及概念之间、实体之间的语义关系
- 同时也包括定义在这个本体上的一些推理规则

■ Freebase的知识体系

- 定义了超过 1.5 万个概念类型和 4,000 个属性
- 对每个类型定义了若干关系，并制定关系的值域约束其取值。

知识体系构建

■ 输入:

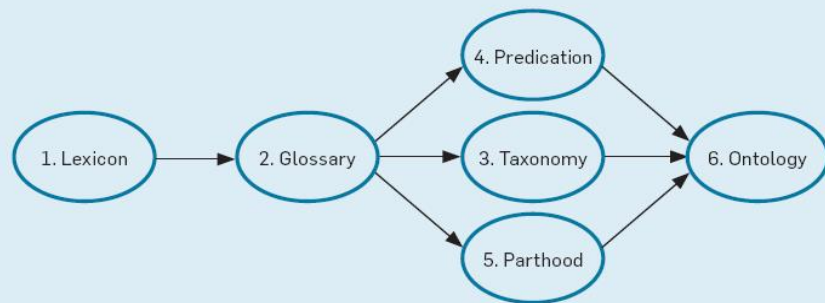
- 领域 (医疗、金融...)
- 应用场景

■ 输出: 领域知识本体

- 领域实体类别体系
- 实体属性
- 领域语义关系
- 语义关系之间的关系

■ 关键技术:

- Ontology Engineering



	A	B	C	D
	Term	Synonyms	Kind	Description [source]
1	Delivery address	Shipping address	Complex property	Location to which goods are to be sent [1].
2	Invoice	Bill	Object	Itemized list of goods shipped, usually specifying the price and the terms of sale [2].
3	Postal address	Address	Complex property	Information that locates and identifies a specific address, as defined by the postal service [3].
4	Purchasing conditions	Purchase terms and conditions	Object	Conditions related to the transaction and the trade [4].
5	Purchase order	PO	Object	Commercial document issued by a buyer to a seller, indicating types, quantities, and agreed prices for products or services the seller will provide to the buyer [5].
6	Customer	Client	Actor	One who purchases a commodity or service [2].
7	Invoicing	Issuing invoice	Process	Making or issuing an invoice for goods or services [6].
8	Purchasing	Buying	Process	Acquisition of something for payment [6].

知识体系构建

- 作为语义网的应用，知识图谱的知识建模采用语义网的知识建模方式，分为概念、关系、概念关系三元组三个层次，并利用“资源描述框架 (RDF)” 进行描述。
- RDF 的基本数据模型包括了三个对象类型：
 - 资源 (Resource)
 - 能够使用 RDF 表示的对象称之为资源，包括互联网上的实体、事件和概念等。
 - 谓词 (Predicate)
 - 主要描述资源本身的特征和资源之间的关系。每一个谓词可以定义元知识，例如，谓词的头尾部数据值的类型（如定义域和值域）、谓词与其他谓词的关系（如逆关系）。
 - 陈述 (Statements)
 - 一条陈述包含三个部分，通常称之为 RDF 三元组 < 主体 (subject), 谓词 (predicate), 宾语 (object)>。其中主体是被描述的资源，谓词可以表示主体的属性，也可以表示主体和宾语之间关系。当表示属性时，宾语就是属性值；当表示关系时，宾语也是一个资源。

生命周期：知识获取

- 知识获取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，其方法根据所处理数据源的不同而不同。
- 知识图谱中数据的主要来源有
 - 结构化数据
 - 半结构化数据
 - 非结构化文本数据
 - 纯文本

知识获取

■ 输入:

- 领域知识本体
- 海量数据: 文本、垂直站点、百科

■ 输出: 领域知识

- 实体集合
- 实体关系/属性

■ 主要技术:

- 信息抽取
- 文本挖掘



网络文本信息结构

■ 结构化数据 (Infobox)

- 置信度高
- 规模小
- 缺乏个性化的属性信息

中文名:	姚明	专业特点:	20英尺外精确跳投
外文名:	Yao Ming	主要奖项:	NBA全明星赛(7次)
别名:	小巨人 移动长城		ESPN全球最有潜力运动员奖(2000)
国籍:	中国		劳伦斯世界最佳新秀奖(2003)
民族:	汉族		中国篮球杰出贡献奖
出生地:	上海	重要事件:	专题影片《姚明年》发行
出生日期:	1980年9月12日	祖籍:	江苏苏州吴江震泽
身高:	2.23米 (7.32英尺)	位置:	中锋
体重:	140.6kg	鞋码:	18码
运动项目:	篮球	自传:	《我的世界我的梦》
所属运动队:	NBA火箭队	生涯最高分:	41分

■ 半结构化数据

- 置信度较高
- 规模较大
- 个性化的信息
- 形式多样
- 含有噪声

1984:《一个和八个》摄影	1984:《黄土地》摄影	1986:《大阅兵》摄影	1987:《老井》主演	1987:《红高粱》导演	1989:《古今大战秦俑情》	1989:《代号美洲豹》	1990:《菊豆》导演	1991:《大红灯笼高高挂》
栖息环境					沙狐栖息在广袤荒原及半沙漠中,行动快而敏捷。狐臭不显。			饮食习惯
					食物包括各种鼠类、鼠兔、			生长繁殖
					常在每年1-3月发情,妊娠期常每胎3~6仔。寿命约6年。			个人简介
					姓名: 姚明			祖籍: 江苏苏州吴江震泽
					出生地: 上海市徐汇区			出生医院: 上海交通大学附属第六人民医院
					小学: 上海市高安路第一小学			初中: 上海市第二中学
					大学: 上海交通大学			星座: 处女座
					曾效力球队: CBA上海东方大鲨鱼、NBA休斯敦火箭			前任主教练: 里克·阿德尔曼
					休斯敦火箭队球衣号码: 11			国家队球衣号码: 13
					2008奥运会号码: 13			手掌: 19厘米
					净身高: 2.26米			臂展: 7英尺4.75英寸 (2.25米)
					站立摸高: 9英尺7英寸 (2.91米)			原地弹跳: 19英寸 (约48厘米)

■ 纯文本

- 置信度低
- 复杂多样
- 规模大

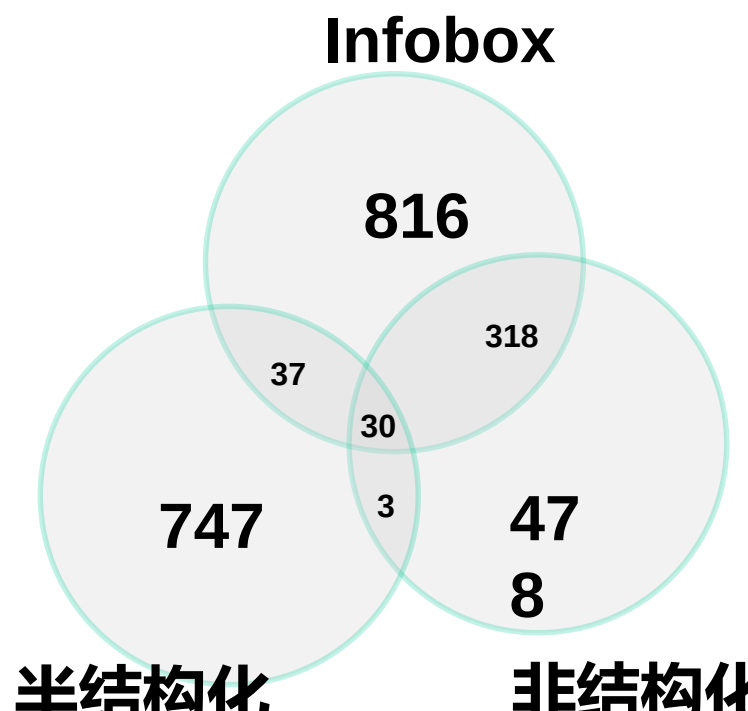
张艺谋, 陕西省西安人, 中国电影导演, 北京奥运会开幕式总导演。在电影学院学的是摄影专业。张艺谋是中国在国际影坛最具影响力的导演, 曾多次荣获国际电影节大奖并成功执导北京奥运会开幕式。其代表作《红高粱》被认为是中国电影走向世界的新开始, 《英雄》则是开启了中国电影的大片时代。

结构化 vs. 半结构化 vs. 非结构化

■ 随机抽取100篇百科文档（共5类）

□ 对于其中三部分都包含的网页进行了统计

	人物	地理	电影	动物	图书
InfoBox	87	182	260	183	104
非结构化	79	147	109	107	36
半结构化	119	96	327	129	76
OverLap Info vs. 非	62	101	87	72	26
OverLap Info vs. 半	11	11	21	7	17
OverLap 非 vs. 半	8	7	4	4	10
三方OverLap	7	7	3	4	9



非结构化文本的实体关系抽取非常重要

知识获取

■ 结构化和半结构化数据源中的知识抽取

- 因为数据噪声少，这类数据源的信息抽取方法相对简单，经过人工过滤后能够得到高质量的结构化三元组。
- 工业界常用的技术手段

■ 非结构化文本中实体的知识抽取

- 涉及到自然语言分析和处理技术，难度较大。
- 但是，因为互联网上更多的信息都是以非结构化文本的形式存在，而非结构化文本的信息抽取能够为知识图谱提供大量高质量的三元组事实，因此它是构建知识图谱的核心技术。
- 学术研究的重点

知识获取：实体识别

- 实体识别任务的目标是从文本中识别实体信息。
 - 例如，对于“姚明 1981 年出生于上海”，需要从中识别出“姚明”、“1981 年”、“上海”三个实体，并确定实体的它们的类别。
- 早期有关实体识别的研究主要是针对命名实体的识别。
 - 命名实体指的是文本中具有特定意义的实体，一般包含三大类（实体类、时间类和数字类）、七小类（人名、地名、机构名、时间、日期、货币和百分比）。
- 在知识图谱领域，从文本中识别实体不仅仅局限于命名实体，还包括其他类别的实体，特别是领域实体。
 - 例如：股票名、汽车品牌、餐馆名等。
- 与实体识别相关的任务是实体抽取
 - 目标是在给定语料的情况下，构建一个实体列表。
 - 例如：构建一个歌曲名列表。

知识获取：实体消歧

■ 目标是消除指定实体的歧义。

- 对于 “李娜于 2011 年获得法国网球公开赛的冠军” 中的 “李娜”，系统需要自动判别出这个实体指称项 “李娜” 指的是打网球的 “李娜”，而不是其他的 “李娜”

■ 实体消歧对于知识图谱构建和应用有着重要的作用，也是建立语言表达和知识图谱联系的关键环节。

■ 从技术路线上划分，实体消歧任务可以分为实体链接和实体聚类两种类型。

- 实体链接是将给定文本中的某一个实体指称项链接到已有知识图谱中的某一个实体上，因为在知识图谱中，每个实体具有唯一的编号，链接的结果就是消除了文本指称项的歧义。
- 实体聚类的假设是已有知识图谱中并没有已经确定的实体，在给定一个语料库的前提下，通过聚类的方法消除语料中所有同一实体指称项的歧义，具有相同所指的实体指称项应该被聚为同一类别。

知识获取：关系抽取

- 目标是获取两个实体之间的语义关系。
- 语义关系可以是一元关系（例如实体的类型），也可以是二元关系（例如实体的属性）甚至是更高阶的关系。
- 在现有知识图谱中，所处理的语义关系通常指的是一元关系和二元关系。
- 在知识图谱中，包含多种类型的语义关系
 - 上下位关系、部分整体关系、属性关系等。
- 关系抽取对于构建知识图谱非常重要，是图谱中确定两个节点之间边上语义信息的关键环节。
- 关系抽取也十分困难，因为对于同一语义关系的自然语言表达多种多样、十分丰富。

知识获取：关系抽取

■ 根据抽取目标的不同，关系抽取任务可以分为：

- 关系分类任务：判别一句话中两个指定实体之间的语义关系。

- “姚明 1981 年出生于上海”，假设已经识别出“姚明”和“上海”两个实体，关系分类的目标是判别出它们之间具有“出生地”关系。

- 属性抽取任务：在给定一个实体以及一个预定义关系的条件下，抽取另外一个实体。

- 例如：“姚明 1981 年出生于上海”，假设已经识别出“姚明”，给定“出生地”这个关系，属性抽取的目标是从这句话中把表达姚明出生地的属性值抽取出来，即“上海”。

- 关系实例抽取任务：给定关系类型，抽取满足该关系的实例数据。

- 例如，给定“出生地”，抽取出满足该关系的“姚明，上海”、“郎平，北京”。

■ 根据标注数据以及运用场景不同，关系抽取方法可以分为：

- 有监督关系抽取、无监督关系抽取、弱监督关系抽取以及开放关系抽取等。

知识获取：事件抽取

- 目标是从描述事件信息的文本中抽取用户感兴趣的事件信息并以结构化的形式呈现出来。
- 事件是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，由一个或者多个角色参与的，一个或者多个动作组成的事情或者状态的改变。
 - “成龙和林凤娇于 1982 年12 月 1 日在洛杉矶举行婚礼。”
 - 事件抽取的目标是识别出：
 - 这个句子描述了一个 “结婚” 事件
 - “结婚的双方” 是 “成龙” 和 “林凤娇”
 - “结婚时间” 是 “1982年 12 月 1 日”
 - “结婚地点” 是 “洛杉矶”

知识获取：事件抽取

- 现有知识图谱大多以实体和实体之间的关系为核心，缺乏事件知识。
- 事件知识能弥补现有以实体和实体关系为核心的知识图谱知识表达能力不足的问题，是构建知识图谱不可或缺的技术。
- 事件结构本身的复杂性以及自然语言表达的歧义性和灵活性，对事件抽取提出了很大的挑战。
- 根据抽取方法的不同，已有的事件抽取方法可以分为
 - 基于模式匹配的事件抽取
 - 基于机器学习的事件抽取

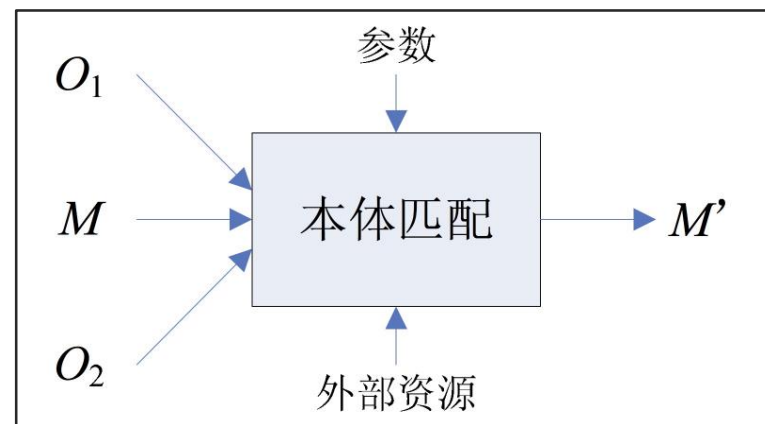
生命周期：知识融合

- 任务：对不同来源、不同语言或不同结构的知识进行融合，从而对于已有知识图谱进行补充、更新和去重。
- YAGO：对专家构建的高质量语言知识图谱 WordNet 和网民协同构建的大规模实体知识图谱 Wikipedia 进行融合而形成的，实现质量和数量的互补；
- BabelNet：融合不同语言的知识图谱，实现跨语言的知识关联和共享。

生命周期—知识融合

■输入：

- 抽取出来的知识
- 现有知识库
- 知识本体



■输出：

- 统一知识库
- 知识置信度

■关键技术：

- Ontology Matching
- Entity Linking



知识融合

■ 从融合的对象看，知识融合分为知识体系的融合和实例的融合。核心问题是计算两个知识图谱中两个节点或边之间的语义映射关系。

- 知识体系的融合：两个或多个异构知识体系进行融合，即对相同的类别、属性、关系进行映射。

- 实例的融合：对于两个不同知识图谱中的实例（实体实例、关系实例）进行融合，包括不同知识体系下的实例、不同语言的实例。

■ 从融合的知识图谱类型看，知识融合分为：垂直方向的融合和水平方向的融合。

- 垂直方向的融合：融合（较）高层通用本体与（较）底层领域本体或实例数据，

 - 融合 Wordnet 和 Wikipedia

- 水平方向的融合：融合同层次的知识图谱，实现实例数据的互补。

 - 融合 Freebase 和 DBpedia

生命周期：知识存储和查询

- 任务：研究采用何种方式将已有知识图谱进行存储。
- 因为目前知识图谱大多是基于图的数据结构，它的存储方式主要有两种形式：
 - RDF 格式存储：以三元组的形式存储数据
 - Google 开放的 Freebase 知识图谱，就是以文本的形式逐行存储三元组 SPO(subject, predicate, object)
 - 这种存储方式使得三元组的搜索效率低下。为了提升三元组的搜索效率，通常采用六重索引的方法。
 - 图数据库 (Graph Database)
 - 比 RDF 数据库更加通用，目前典型的开源图数据库是 Neo4j
 - 优点：具有完善的图查询语言，支持大多数的图挖掘算法
 - 缺点：数据更新慢，大节点的处理开销大。为了解决上述问题，子图筛选、子图同构判定等技术是目前图数据库的研究热点。

生命周期—知识存储和查询

■ 输入:

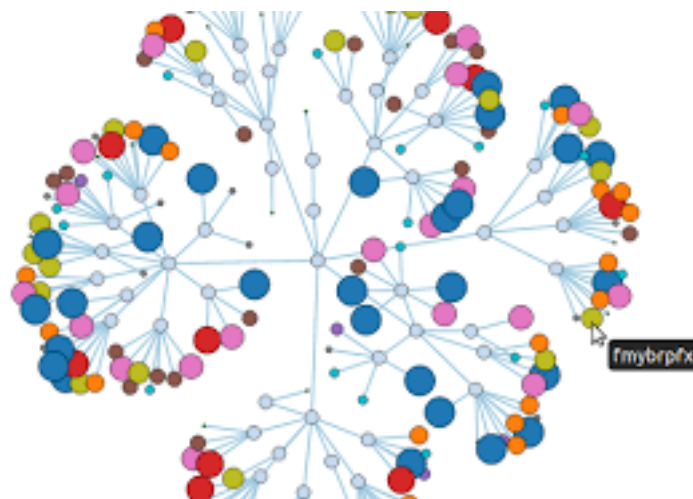
- 大规模知识库知识

■ 输出:

- 知识存储和查询服务

■ 主要技术:

- 知识表示
- 知识查询语言
- 存储/检索引擎



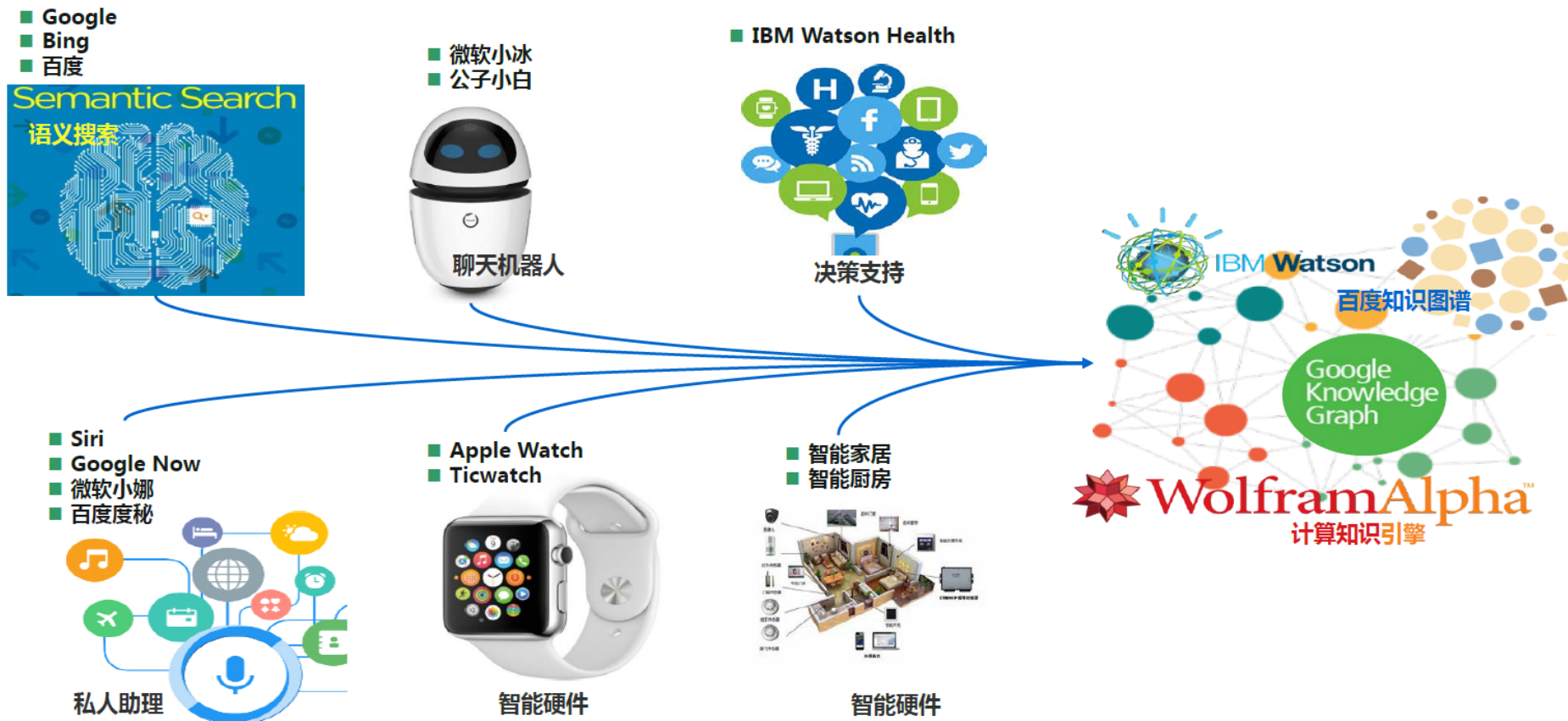
生命周期：知识推理

- 由于处理数据的不完备性，知识图谱中肯定存在知识缺失现象（包括实体缺失、关系缺失）。我们也很难利用抽取或者融合的方法对于缺失的知识进行补齐。因此，需要采用推理的手段发现已有知识中隐含的知识。
- 目前知识推理的研究主要集中在针对知识图谱中缺失关系的补足，即挖掘两个实体之间隐含的语义关系。所采用的方法可以分为两种：
 - 基于传统逻辑规则的方法进行推理：研究热点在于如何自动学习推理规则，以及如何解决推理过程中的规则冲突问题。
 - 基于表示学习的推理：即采用学习的方式，将传统推理过程转化为基于分布式表示的语义向量相似度计算任务。这类方法优点是容错率高、可学习，缺点也显而易见，即不可解释，缺乏语义约束。

知识推理

- 知识推理也直接应用于相关应用任务，例如自动问答系统中往往也需要知识推理。
- 关键在于如何将问题映射到知识图谱所支撑的结构表示中，在此基础上才能利用知识图谱中的上下文语义约束以及已有的推理规则，并结合常识等相关知识，得到正确的答案。

生命周期：知识应用



精准语义搜索



姚明 运动员



Sign in

All

Images

News

Videos

Maps

More

Settings

Tools

About 5,770,000 results (0.53 seconds)

Images for 姚明 运动员



More images for 姚明 运动员

Report images

姚明- 维基百科，自由的百科全书

<https://zh.wikipedia.org/zh-hans/姚明> • Translate this page

姚明（1980年9月12日－），生于中国上海市，祖籍江苏省苏州市吴江区震泽镇，著名篮球运动员，曾为中国国家篮球运动员，曾效力于中国篮球职业联赛（CBA）上海大...

青年时代及CBA生涯·NBA职业生涯·退役后·生涯数据

姚明（中职联公司董事长兼总经理，前NBA球星）- 百度百科

baike.baidu.com/view/2271.htm • Translate this page

姚明（Yao Ming），1980年9月12日出生于上海市徐汇区，祖籍江苏省苏州市吴江区震泽镇，前中国职业篮球运动员，司职中锋，现任中职联公司董事长兼总经理。1998...

总局发言挺姚明掌舵篮协：优秀运动员可当领导_CBA_新浪竞技风暴_...

sports.sina.com.cn • 篮球-CBA • Translate this page

Jan 9, 2017 - 记者陈临报道1月2日，本报独家报道，国家体育总局对中国篮球发展做出了10点重要批示，其中包括陈临和任志强主席，并兼任男篮主教练。

姚明_互动百科

www.baike.com/wiki/姚明 • Translate this page

姚明-姚明（1980年9月12日），中国上海市人，毕业于中国上海交通大学，前中国篮球运动员。姚明17岁入选国家青年队，18岁穿上了中国队服，曾效力于中国篮球职业...

工资最高的华人运动员，不是姚明，是他！_网易订阅

dy.163.com/weimedia/article/detail/8ELN83790512814T.html • Translate this page

姚明当年在NBA打球的时候创造了一项记录：北美五大职业联赛中拿到第一高薪的华人球员。2006年夏天火箭为他开出5年7500万美元续约合同，平均年薪1500万...

Yao Ming

Basketball player



姚明

Yao Ming is a Chinese retired professional basketball player who played for the Shanghai Sharks of the Chinese Basketball Association and the Houston Rockets of the National Basketball Association. Wikipedia

Born: September 12, 1980 (age 36 years), Shanghai, China

Height: 7' 6"

Shoe size: 18

Weight: 311 lbs

Spouse: Ye Li (m. 2007)

Parents: Yao Zhiyuan, Fang Fengdi

Profiles



Facebook



Twitter



Instagram

People also search for

View 15+ more



Ye Li
Spouse



Shaquille
O'Neal



Manute Bol



Yi Jianlian



Jeremy Lin

Feedback

精准语义搜索：天眼查



王健林 [下载报告](#)

介绍：王健林，男，1954年10月24日出生于四川省广元市，1989年起担任大连万达集团股份有限公司董事长。1970年入伍，1986年毕业于辽宁大学，同年7月进入大连市西岗区人民政府任办公室主任，1989年进入房地产行业，1993年起担任大连万达集团股份有限公司董事长、总裁。是中共十七大代表、第十一届全国政协常委、第十一届全国工商联副主席，兼任... [详细](#)

- 他的所有商业角色
- 他的所有企业

角色	企业	省份地区	开业日期	注册资本	经营状态
	大连万达集团股份有限公司	辽宁	1992-09-28	100000 万人民币	存续（在营、开业、在册）
	北京中住联华科技发展有限公司	北京	2001-04-25	120 万元	吊销
	北京环球高尔夫俱乐部有限公司	北京	1993-04-10	1120万元美元	吊销，未注销
	大连万达商业开发有限公司	辽宁	1994-12-12	0 万	吊销，未注销
	大连合兴投资有限公司	辽宁	2007-04-27	7860 万人民币	存续（在营、开业、在册）
	大连万达通信有限公司	辽宁	1998-01-23	15000 万	吊销，未注销
	大连万达酒业有限公司	辽宁	1998-03-30	240 万	其他
	大连万达中心大厦有限公司	辽宁	1995-04-29	10000 万人民币	吊销，未注销
	四川万达房地产有限公司（转成都市锦江区）	四川	1999-07-22	2000万人民币	存续（在营、开业、在册）
	大连宏大物业管理有限公司	辽宁	1995-04-11	1000 万人民币	吊销，未注销
	大连天兴大酒店有限公司	辽宁	1997-01-20	100 万	吊销，未注销

关系搜索

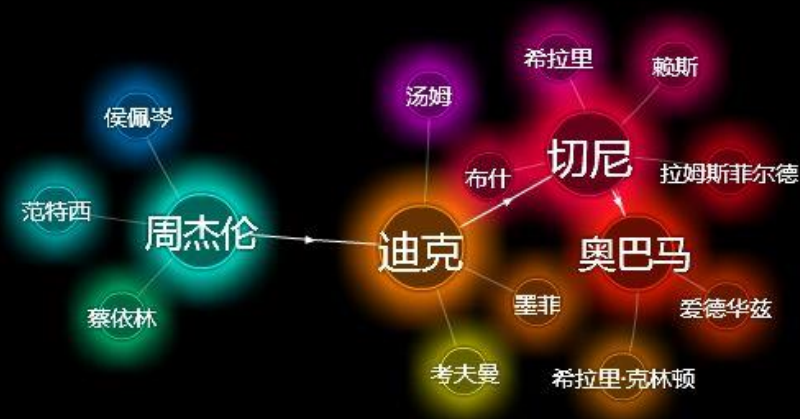


六度秀
分享
帮助

周杰伦 → 奥巴马 六度搜索



点击上方人物，自定义关系链



智能问答



屠呦呦是哪里人



全部

新闻

图片

视频

地图

更多

设置

工具

找到约 1,600,000 条结果 (用时 0.38 秒)

屠呦呦 / 出生地



智能问答：有推理能力

梁启超的儿子的老婆的情人的父亲

搜狗搜索

网页

论坛

知识

新闻

博客

百科

更多

什么是分类搜索

找到约 173,657 条结果

梁启超的儿子的老婆的情人的父亲：



徐申如

推理说明：梁启超的儿子是梁思成。梁思成的妻子是林徽因。林徽因的情人是徐志摩。徐志摩的父亲是徐申如

[梁启超的儿子的老婆的情人的老婆-读书-DoNews.COM](#)

2004年6月15日... 作者 帖子主题：知道是谁不？ 作者 帖子主题：RE：梁启超的儿子的老婆的情人的老婆 【(shengfang) 回复 (cool) 的大作】 陆小慢 作者 帖子主题：RE:....
[donewsIT门户 - home.donews.com/.../477746.html - 2004-6-15 - 快照 - 预览](#)

[梁启超的儿子的老婆的情人的父亲](#) 最佳答案 搜狗知识搜索

[梁启超的儿子的老婆的情人的老婆是谁 - 已解决 搜搜问问 2007-11-25](#)

答：梁启超的儿子呢是中国的著名建筑师梁思成 梁思成的老婆呢叫林徽茵看过《人间四月天》的人应该知道啊 那么林徽茵的情人呢就是大名鼎鼎的徐志摩啦 那么徐志摩的老婆...

[梁启超的儿子的老婆的情人的老婆是谁??? - 已解决 搜搜问问 2011-3-4](#)

[梁启超的儿子的太太的情人的太太分别是谁 - 已解决 搜搜问问 2007-12-9](#)

[梁启超的儿子的妻子的情人的老婆是谁? 百度知道 2006-10-4](#)

梁启超



梁启超(1873.2.23—1929.1.19)，生于广东新会。1894年，梁启超提倡变法，并于上海主撰《时务报》，著《变法通议》，刊布报端，启发国人之革新思想。与谭嗣同...

[相关阅读](#)

出生：1873-02-23 / 广东新会

逝世：1929-01-19

妻子：李蕙仙 (正室) / 王桂荃 (老婆)

人物关系：梁宝瑛 (父亲) / 梁思达 (儿子) / 梁思忠 (儿子) / 梁思懿 (女儿) / 梁思成 (儿子)

个人名言：享受工作的同时享受生活

著作

[更多>>](#)



中国近三百年学...



中国历史研究法



新大陆游记



李鸿章传



清代学术概论

推荐

战狼2



百度一下

网页 新闻 贴吧 知道 音乐 图片 视频 地图 文库 更多»

百度为您找到相关结果约5,830,000个

搜索工具

战狼2_百度百科



关注点 TA说：深度解读功与过

《战狼 II》是吴京执导的动作军事电影，由吴京、弗兰克·格里罗、吴刚、张翰、卢靖姗、淳于珊珊、丁海峰等主演。该片于2017年7月27日在中国内地上映。影片讲述了脱下军装的冷锋被卷入了一场非洲国家的叛乱，本来能够安全撤离的他无法忘记军...

[剧情简介](#) [演职员表](#) [角色介绍](#) [音乐原声](#) [幕后花絮](#) [更多>>](#)

baike.baidu.com/

战狼2的最新相关信息

战狼2依然无敌 其他片即便是“炮灰”也各有特色



上周五，《心理罪》《侠盗联盟》《鲛珠传》集体进入电影市场，对票房强势的《战狼2》形成围剿。但姜还是老的辣！从上周末的市场走向来看，《战狼2》依然...

新浪 17分钟前

最全票房信息都在这里！战狼2票房将超6亿... 人民网重庆站

37分钟前

《战狼2》助力国产片票房逆袭 大批国产片火... 电影网

45分钟前

专访《战狼2》航拍主飞手：辛苦与压力并存 华夏经纬

4小时前

外媒：《战狼2》触动中国观众的民族情结... 观察者网

6小时前

战狼2吧_百度贴吧

关注用户：10万人 累计发帖：52万

精品贴(21)

★【战狼2】??【BD1280】[全]完整出了。在线观...

点击：2402

回复：9

战狼2是我贡献过票房的电影

点击：1334

回复：5

战狼2一出，我就是元老了，哈哈

点击：488

回复：57

[查看更多战狼2吧的内容>>](#)

相关影视作品

收起 ^



叶问4

甄子丹主演
的动作电影



印囡

徐峥编导公
路喜剧片



火蓝刀锋2之
雄鹰展翅

杨志刚主演
当代军旅剧



使徒行者2

苏万聪执导
警匪电视剧



僵尸世界大
战2



大闹东海



复仇者联盟3



敢死队4



环太平洋2



金刚狼3



杀破狼2



极限特工3



孤狼特别突
击队



利刃出击



拆弹专家



第一滴血5

领域知识图谱的应用



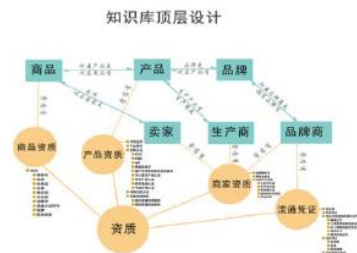
金融证券



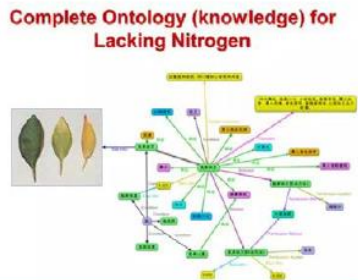
生物医疗



图书情报



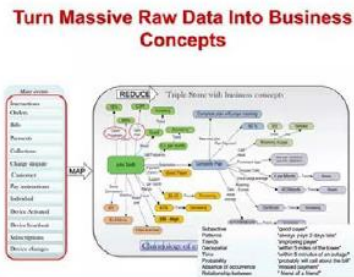
电商



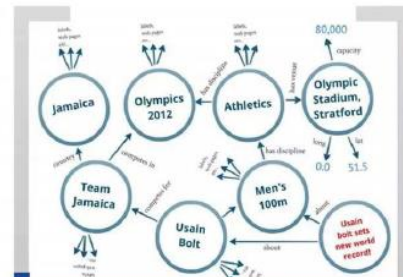
农业



政府



电信



出版

金融证券领域

■ 企业风险评估

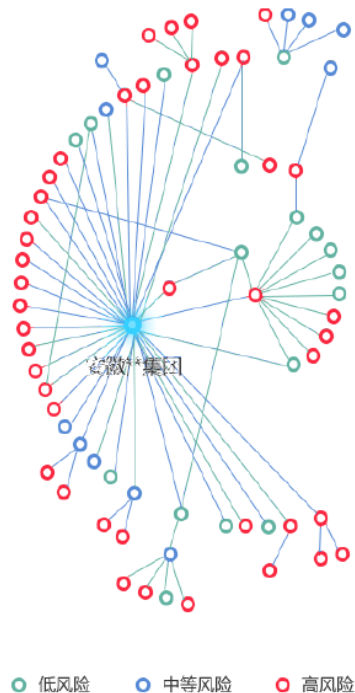
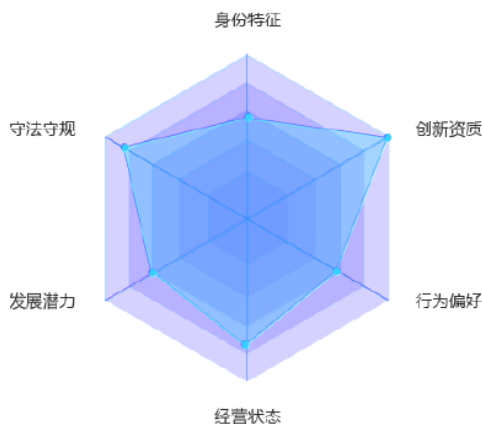
- 基于企业的基础信息、投资关系、诉讼、失信等多维度关联数据，利用图计算等方法构建科学、严谨的企业风险评估体系，有效规避潜在的经营风险与资金风险。

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



■ 用户群体

银行、担保、投行、政府.....

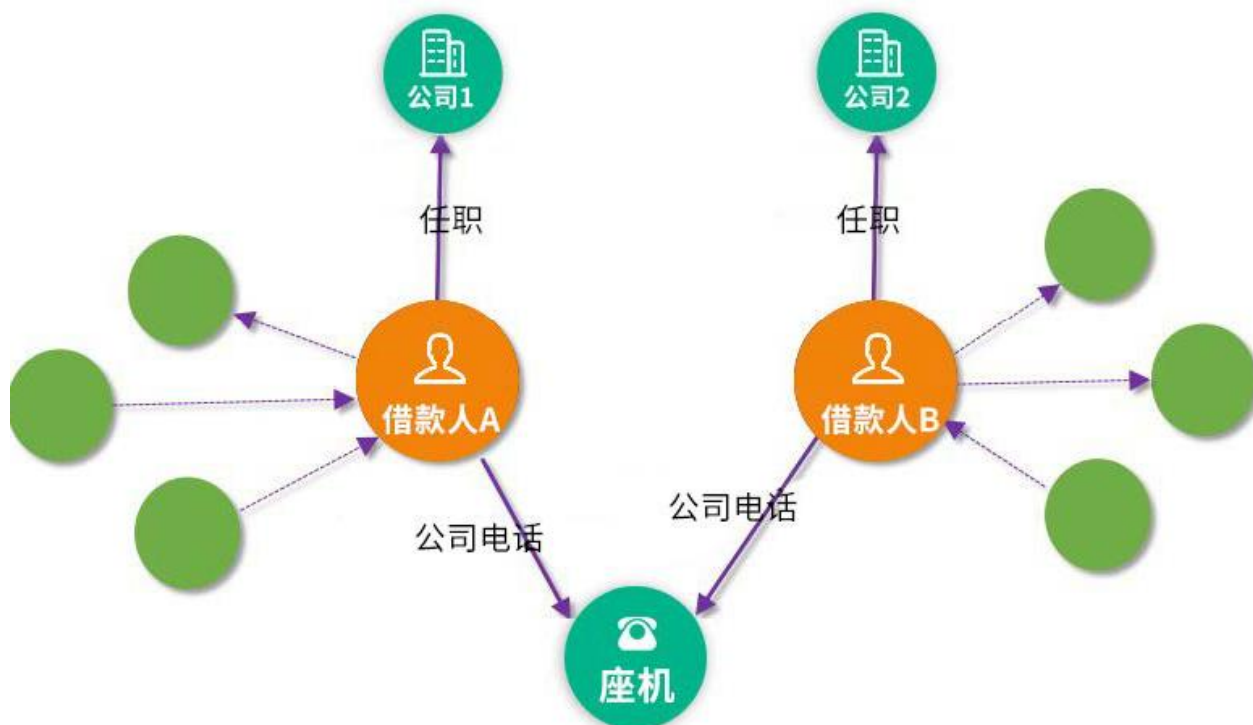
■ 应用环节

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
.....

金融证券领域

■ 反欺诈

- 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。



医疗领域

■ 中医药知识服务平台

- 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。

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

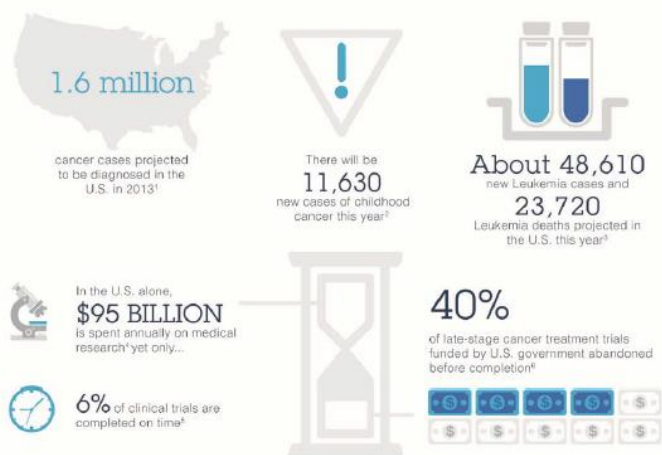
收集证候、临床、养生、中药、中医文献等方面的本体。

医疗领域

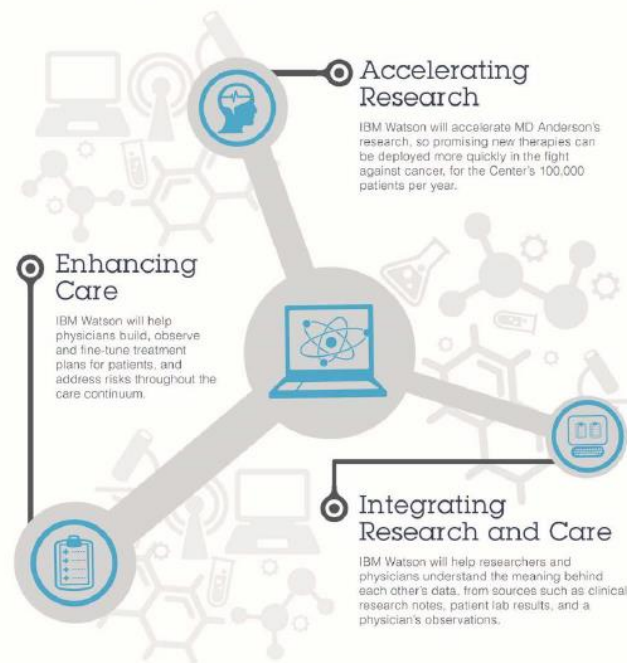
■ Watson辅助诊断与治疗

□ 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



Early analysis shows IBM Watson could unlock new patterns and relationships from within Big Data that could help advance new cancer therapies.

医疗领域

■ Open PHACTS 新药物发现

- 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。



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

提纲

- 什么是知识图谱
- 知识图谱发展历程
- 知识图谱的类型和代表性知识图谱
- 知识图谱的生命周期
- 知识图谱和深度学习
- 课程介绍

知识图谱 vs. 深度学习

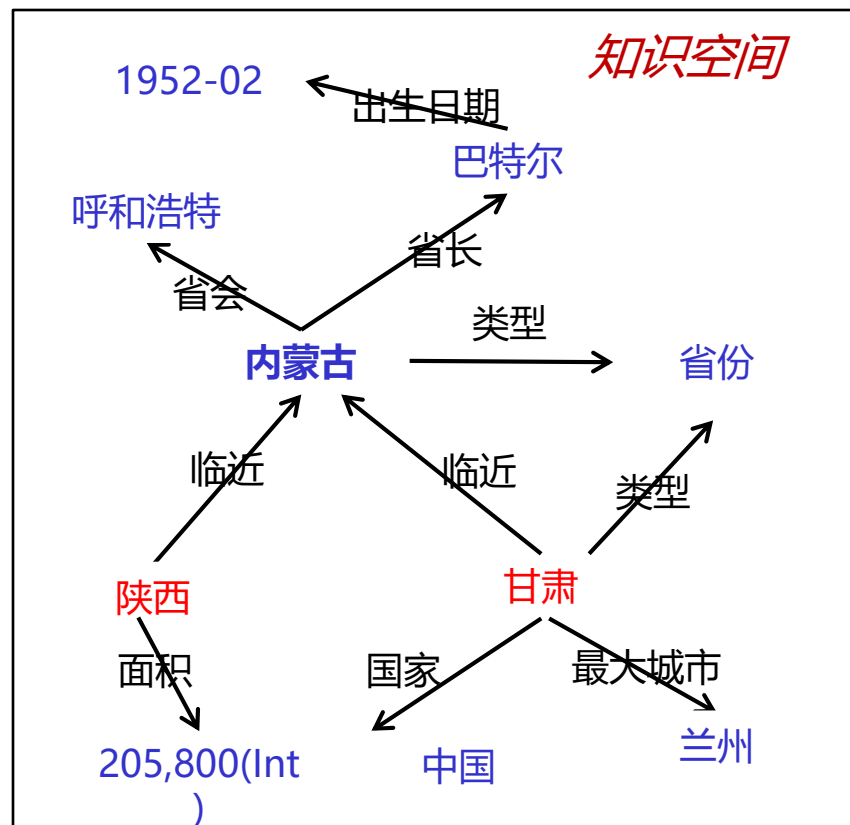
■ 知识图谱：基于符号计算的人工智能的基础

- 语言符号和知识符号
- 符号逻辑系统
- 语言和知识单元及其关系的语义蕴含在形式化符号及其结构之中
- 通过精确匹配和数理演算进行语义计算

■ 深度学习：基于数值计算的人工智能的基础

- 基于分布的表示
- 数值计算系统
- 语言和知识单元及其关系的语义蕴含在数值空间中的点、线、面、子空间等单元之中
- 通过数值演算进行语义计算

符号化的知识表示：结构化知识

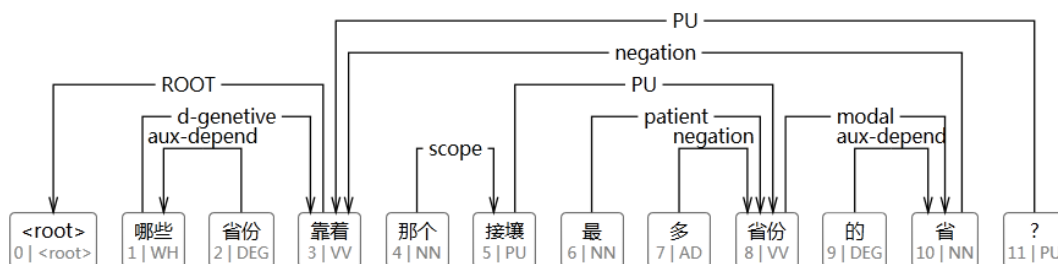


符号化的知识表示：非结构化知识

问句：哪些省份靠着那个接壤最多省份的省？

■ 关键词组合 {靠着、省份、接壤、最多}

■ 语义树(图)结构



■ 逻辑表达式结构

$\lambda x. \text{省}(x) \wedge \text{临近}(x, \text{argmax}(\lambda y. \text{省}(y), \lambda y. \text{count}(\lambda z. \text{省}(z) \wedge \text{临近}(y, z))))$

数值计算的知识表示：分布表示

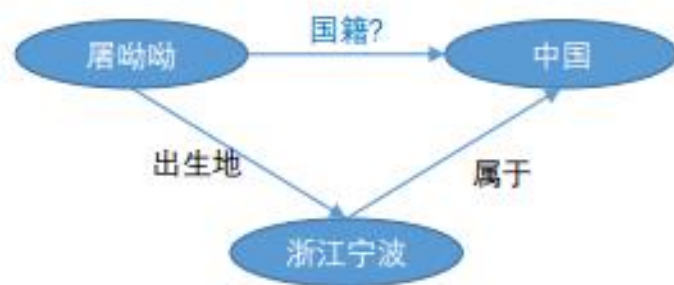
■ 计算机擅长处理什么

- 二进制：进行布尔运算
- 数字：进行数值计算

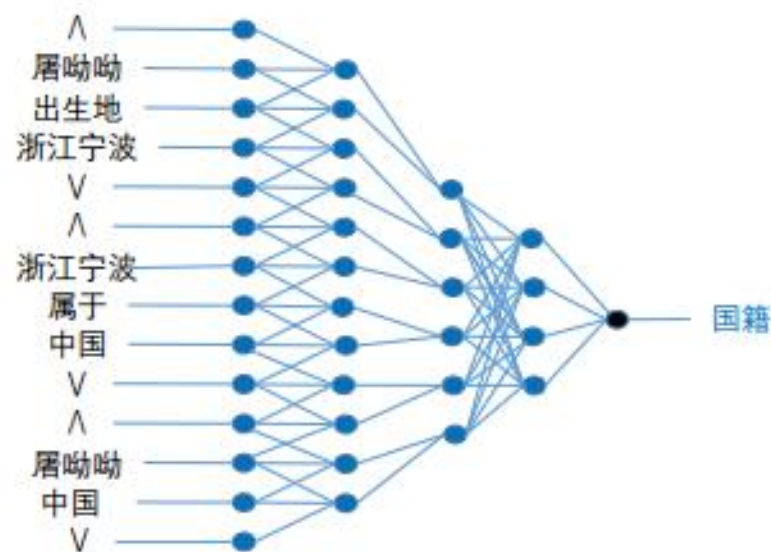
■ 基于神经元的深度学习方法本质

- 通过多层神经网络，自动学习处理对象的抽象表示，把它们表示为数值（标量/向量/矩阵），并通过数值计算（如数值相乘、开方、乘以非线性函数等）完成各种语义计算任务。
- 知识单元及其结构
- 语言单元及其结构

知识图谱 vs 深度学习



基于符号演算的知识推理



基于深度学习的语义计算

深度学习的优势和局限性

- 深度学习通过多层神经网络，自动学习处理对象的抽象表示。
- 当前的深度学习模型需要依赖大量的标注数据。
 - Bengio 在其最新的《Deep Learning》著作中称，平均每个类别需要 5 千个标注样本才能够达到理想的性能
- 这种纯粹的数据驱动学习方式不仅结果不可解释而且学习过程也不可调控。
- 端到端的学习过程很难加入先验知识，而大量应用任务中人类的先验知识非常重要。

知识图谱的优势和局限性

- 为了模拟人类的认知过程，进行更好的知识抽象和推理，人们定义了符号逻辑对知识进行表示和推理，其中知识图谱是典型代表之一。
- 符号逻辑难以从历史数据中自动学习，目前大都还是采用人工编撰和校验的方式获得符号逻辑推理规则
- 尽管可以利用概率图模型等方法学习有限的知识推理规则，但是受限于模型复杂度还是难以扩展到大规模数据上。

知识图谱和深度学习的融合

当前知识图谱的限制和不足

■ 领域限制

- 一些知识库侧重于语言: WordNet, BabelNet
- 一些知识库侧重于Schema: ConceptNet, UMBEL
- 一些知识库侧重于Fact: DBPedia, Yago

■ 对时空属性的建模

- 对动态性的实体, 如Event建模不足
- Yago 3在一定程度上考虑时间和地理属性

■ 自动构建

- 自动构建是维护 and 保持知识图谱质量和覆盖的核心技术

■ 与LOD的集成

- 缺乏Schema之间的alignment
- 往往只用到底层的表达能力, OWL的高级功能很少涉及

知识图谱的发展方向

■ 新的知识表示模型

- Ontology engineering 已经被用了超过15年

■ 新类型的知识图谱

- 不再围绕实体和关系的存储
- 如Event-centric KG

■ 知识图谱自动构建技术

- 在Freebase中, 71%的人没有出生日期
- 新技术: Distant Supervision, KG embedding, 知识集成 (如Google的Knowledge Vault)

提纲

- 什么是知识图谱
- 知识图谱发展历程
- 知识图谱的类型和代表性知识图谱
- 知识图谱的生命周期
- 知识图谱和深度学习
- 课程介绍

知识工程

S.L. Kendal and M. Creen

An Introduction to Knowledge Engineering

1	An Introduction to Knowledge Engineering.....	1
	Section 1: Data, Information and Knowledge	2
	Section 2: Skills of a Knowledge Engineer	10
	Section 3: An Introduction to Knowledge-Based Systems	18
2	Types of Knowledge-Based Systems	26
	Section 1: Expert Systems.....	27
	Section 2: Neural Networks	36
	Section 3: Case-Based Reasoning.....	55
	Section 4: Genetic Algorithms.....	66
	Section 5: Intelligent Agents.....	74
	Section 6: Data Mining	83
3	Knowledge Acquisition.....	89
4	Knowledge Representation and Reasoning	108
	Section 1: Using Knowledge.....	109
	Section 2: Logic, Rules and Representation	116
	Section 3: Developing Rule-Based Systems	126
	Section 4: Semantic Networks.....	140
	Section 5: Frames	149
5	Expert System Shells, Environments and Languages.....	159
	Section 1: Expert System Shells.....	160
	Section 2: Expert System Development Environments	165
	Section 3: Use of AI Languages.....	168

知识工程

x

Contents

6	Life Cycles and Methodologies.....	183
	Section 1: The Need for Methodologies.....	185
	Section 2: Blackboard Architectures.....	192
	Section 3: Problem-Solving Methods.....	199
	Section 4: Knowledge Acquisition Design System (KADS).....	209
	Section 5: The Hybrid Methodology (HyM).....	218
	Section 6: Building a Well-Structured Application Using Aion BRE....	232
7	Uncertain Reasoning.....	239
	Section 1: Uncertainty and Expert Systems.....	240
	Section 2: Confidence Factors.....	243
	Section 3: Probabilistic Reasoning.....	248
	Section 4: Fuzzy Logic.....	259
8	Hybrid Knowledge-Based Systems.....	270

课程安排

■ 第一周 课程介绍及知识图谱基础 （赵军）

- 1.1 课程简介
- 1.2 知识图谱概述
- 1.3 知识图谱与深度学习

■ 第二周 机器学习基础 （刘康）

- 2.1 机器学习基础
- 2.2 深度学习基础
- 2.3 典型深度学习模型（一）
- 2.4 典型深度学习模型（二）

■ 第三周 图算法基础 （何世柱）

- 3.1 典型图算法
- 3.2 图匹配算法
- 3.3 面向图结构的随机游走算法
- 3.4 基于图结构的深度学习模型

课程安排

■ 第四周 知识表示方法 (赵军)

- 4.1 经典知识表示理论
- 4.2 语义网中的知识表示
- 4.3 知识图谱中的知识表示
- 4.4 知识图谱的数值化表示

■ 第五周 知识体系构建和知识融合 (刘康)

- 5.1 知识体系构建
- 5.2 知识融合

■ 第六周 实体识别和扩展 (赵军)

- 6.1 实体识别
- 6.2 细粒度实体识别
- 6.3 实体扩展

课程安排

■第七周 实体消歧

(赵军)

- 7.1 实体消歧概述
- 7.2 基于聚类的实体消歧
- 7.3 基于链接的实体消歧

■第八周 关系抽取

(赵军, 陈玉博)

- 8.1 关系抽取概述
- 8.2 限定域关系抽取
- 8.3 开放域关系抽取

■第九周 事件抽取

(赵军, 陈玉博)

- 9.1 事件知识表示
- 9.2 事件知识抽取
- 9.3 基于事件知识的故事概要抽取

考核方式

■ 第十周 存储和检索 (何世柱)

- 10.1 基于关系数据库的知识图谱存储
- 10.2 基于图数据库的知识图谱存储
- 10.3 图数据库
- 10.4 知识图谱检索 (SPARQL/Cypher)

■ 第十一周 知识推理和语义计算 (何世柱)

- 11.1 推理概述
- 11.2 知识推理分类
- 11.3 归纳推理：推理规则获取
- 11.4 演绎推理：推理具体事实

考核方式

■第十二周 知识图谱应用1：问答系统 (刘康)

- 12.1 知识问答概述
- 12.2 基于语义解析的方法
- 12.3 基于搜索排序的方法

■十三周 知识图谱应用2：对话系统 (刘康)

- 13.1 知识对话概述
- 13.2 任务导向型对话系统
- 13.3 通用知识对话

教学方式

- 脉络梳理
- 兼顾学术和应用
- 教学相长

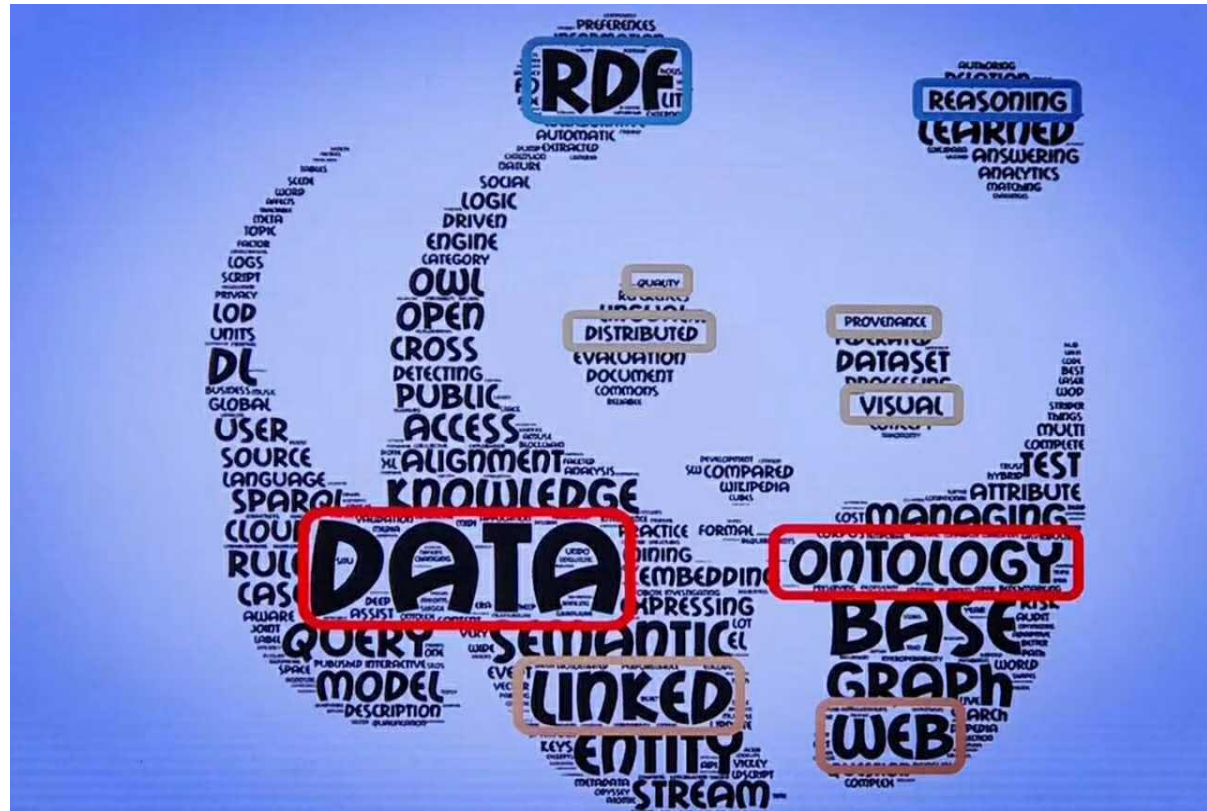
参考书

- 赵军, 刘康, 何世柱, 陈玉博, 《知识图谱》, 高等教育出版社, 2018.
- Stuart J. Russell, Peter Norvig. 人工智能:一种现代方法(第3版)(影印版). 清华大学出版社. 2011年7月出版.
- Brachman, Ronald J., Hector J. Levesque. Knowledge Representation and Reasoning. MIT press, 2004.
- Goodfellow I, Bengio Y, Courville A. Deep learning. MIT press, 2016.
- Blackburn P, Bos J. Representation and inference for natural language[J]. A first course in computational semantics. CSLI, 2005.

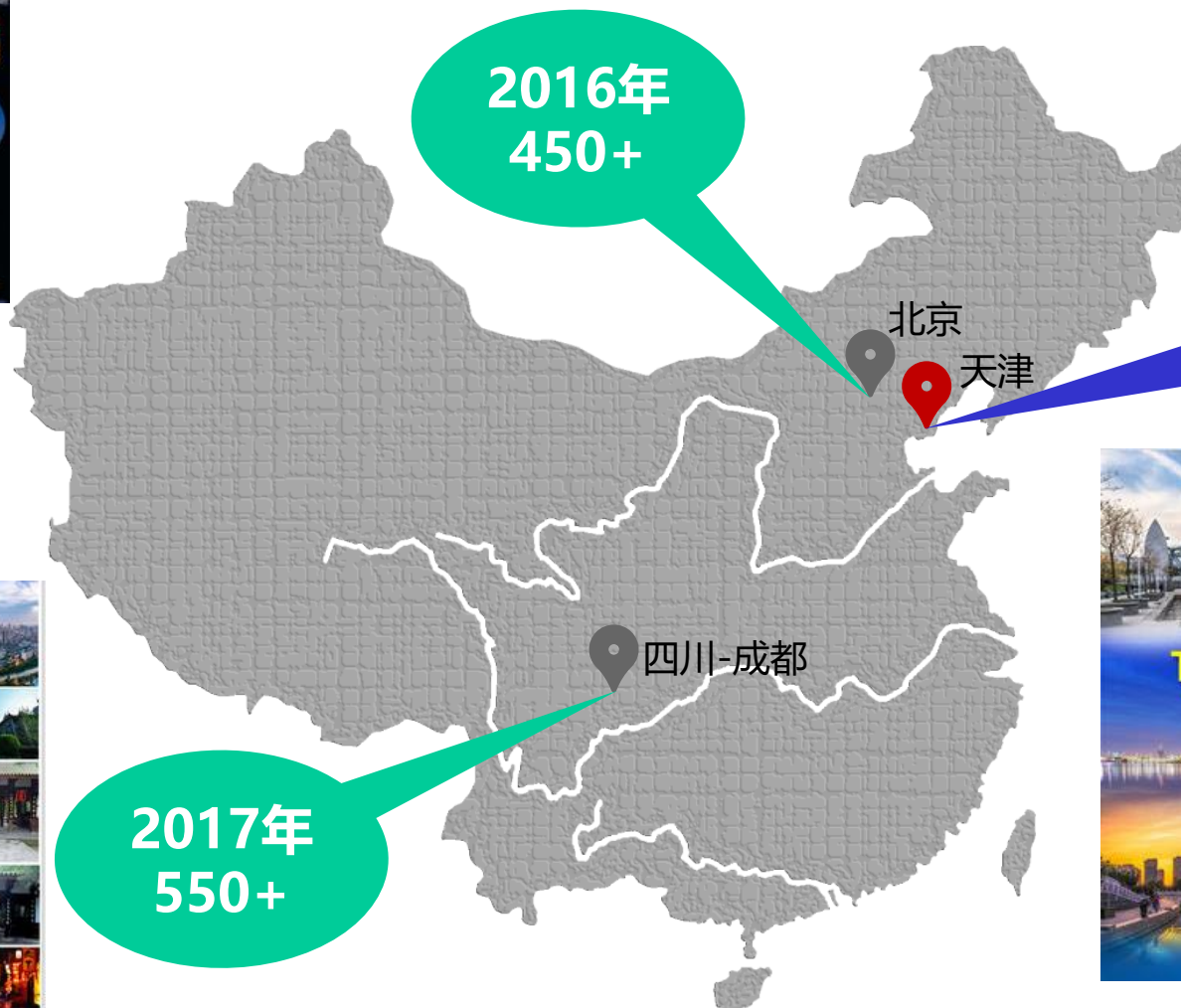
学术界-国际

■ 国际上顶级会议相关的文章和专题：云图

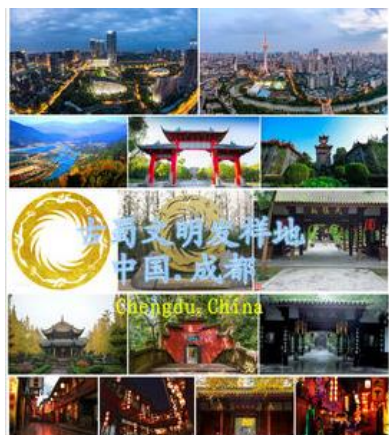
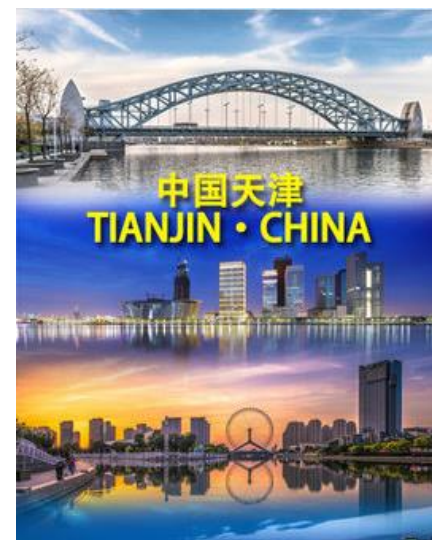
- ACL
- IJCAI
- AAAI
- KDD
- SIGIR
- COLING
- EMNLP
- ISWC



全国知识图谱与语义计算学术会议



2018年
800+



CCKS 2018

■ 主题：知识计算与语言理解

■ 特邀报告：

Prof. James A. Hendler

The **Semantic Web**: Vision,
Reality and Revision

02

Prof. Roberto Navigli

What can you do with
**multilingual knowledge
graphs**? Experiences at
Sapienza and Babelscape

04



01

张钹院士

知识与人工智能



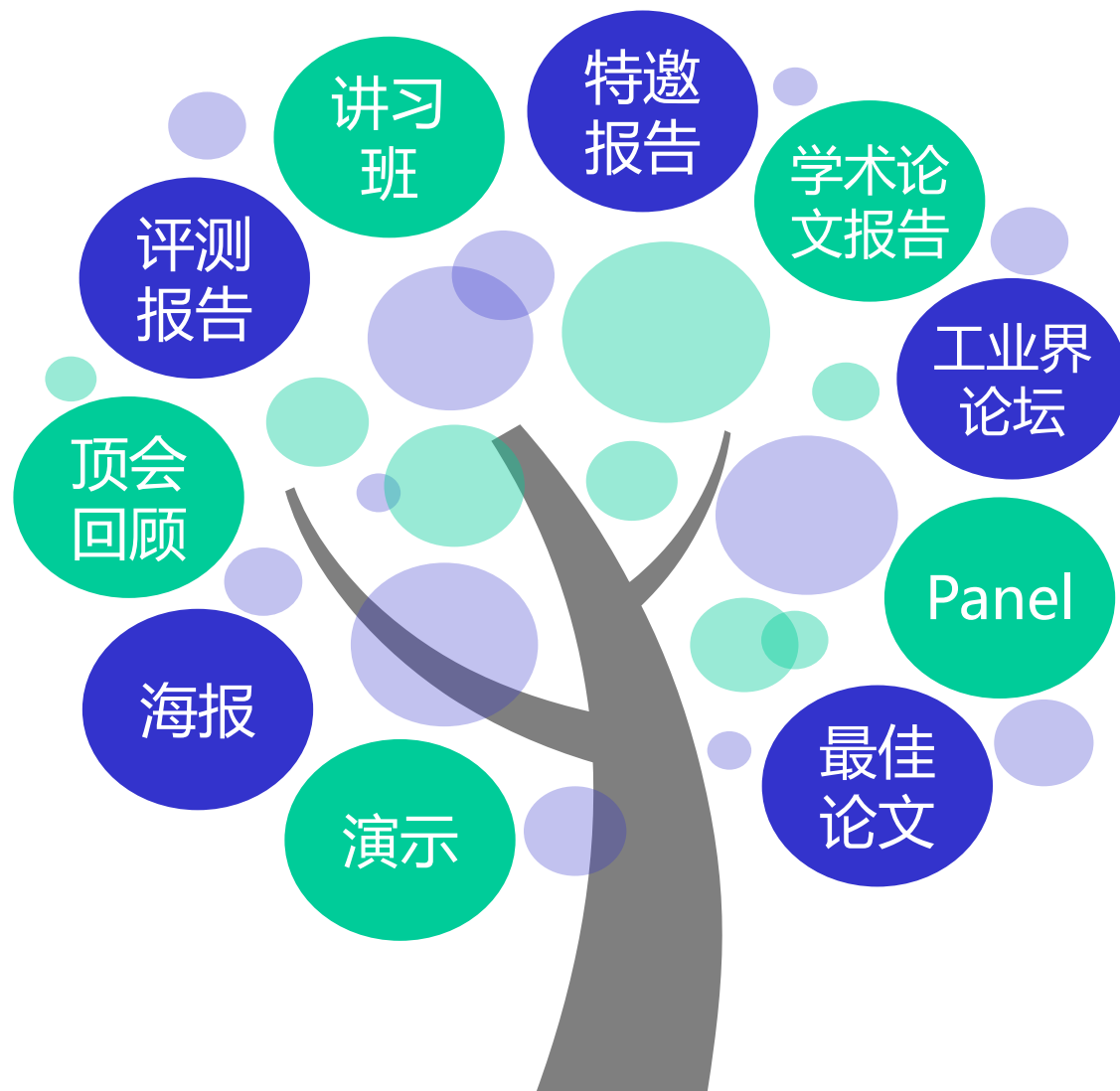
03

玄难副总裁

阿里巴巴生态下
的知识引擎



CCKS 2018



参考

■ 李涓子、韩先培、王昊奋等老师的工作

谢谢

jzhao@nlpr.ia.ac.cn