Foundation of Artificial Intelligence

人工智能基础

李翔宇

软件学院

Email: lixiangyu@bjtu.edu.cn

知识表示与知识图谱

- 知识的概述（定义、特性、分类）
- 知识表示的方法 （知识表示方法的分类）
- 产生式表示法
- 状态空间表示法 （定义、表示、发展历史）
- 知识图谱

用产生式表示：如果一种微生物的染色斑是革兰氏阴性，其形状呈杆状，病人是中间宿主，那么该生物是绿**杆菌的可能性有6成。**

IF 本微生物的染色斑是革兰氏阴性∧本微生物的形状呈杆状∧病人是中间宿主, THEN 该生物是绿杆菌(0.6)。

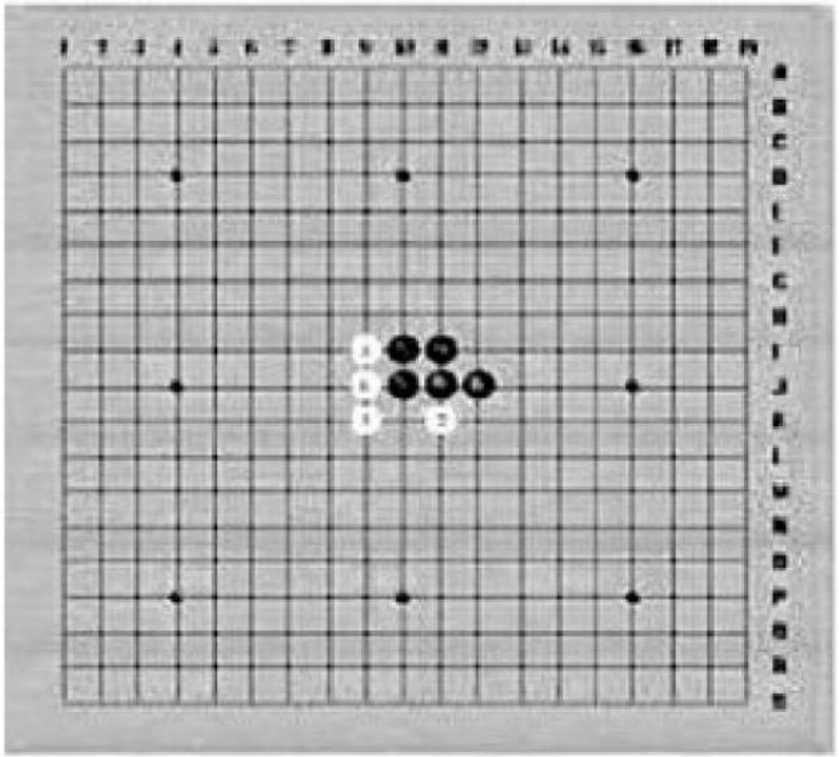
提醒！！！ 第7周 Quiz1，测验内容是1-4章

第13周Quiz2，测验内容是5-7章

2.4 状态空间表示法

- **状态空间**（**state space**） 表示法是人工智能中最基本的形式化方法，是其他形式化方法和问题求解技术的出发点。
- **状态**（state）就是用来描述在问题求解过程中某一个时刻进展情况等**陈述性知识**的一组变量或数组，是某种结构的符号或数据。
- 状态（state）是一组变量q0，q1，q2，…，qn的有序集合，其形式如下：　　Q={ q0，q1，q2，…，qn }

  其中，每个元素qi 称为一个状态变量。
- 状态的表示还可以根据具体应用，采取合适的数据结构，如符号、字符串、多维数组、树和图等。



19× 19的二维数组，0表示无棋子，1表示“黑子”，2表示“白子”

2.4 状态空间表示法

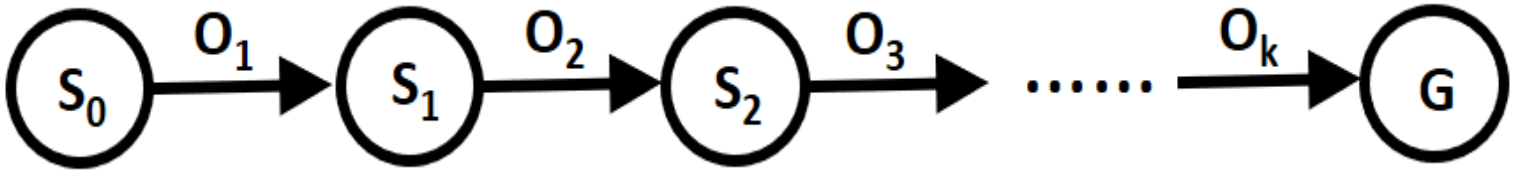
- **操作**也称为**运算**，用来表示引起状态变化的**过程性知识**的一组关系或函数，它会引起状态中的某些分量发生改变，从而使问题由一个具体状态转换到另一个具体状态。
- 操作可以是一个动作（如棋子的移动）、过程、规则、数学算子等，表示状态之间存在的关系。
- 用于表示操作的符号，称为**操作符**（operator）或**操作算子**、**运算符**。
- **状态空间**是采用状态变量和操作符号表示系统或问题的有关知识的符号体系。

2.4 状态空间表示法

- 问题的状态空间是一个表示该问题全部可能状态及其相互关系的集合，常用一个**四元组**（S,O,S0,G）来表示，其中:
  - S 为问题的状态集合；
  - O 为操作符的集合；
  - S_0 是问题的初始状态，是S的一个非空真子集，即$S_0 \subset S$；
  - G 为问题的目标状态，它既可以是若干具体状态，也可以是满足某些性质的路径信息描述，$G \subset S$。

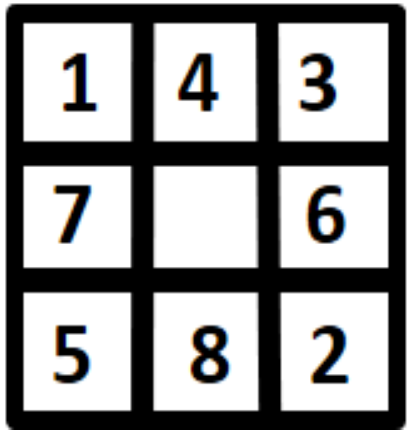
2.4 状态空间表示法

- 状态空间通常用**有向图**来表示，其中，**结点**表示问题的**状态**，结点之间的**有向边**表示引起状态变换的**操作**，有时边上还赋有**权值**，表示变换所需的**代价**。
- 在状态空间中，求解一个问题就是从初始状态出发，不断运用可使用的操作，在满足约束的条件下达到目标状态。
- **问题的解**可能是图中的一个状态，也可能是从初始状态到某个目标状态的一条路径，还可能是达到目标所花费的代价。
- 下图中，**问题的解**便是一条从结点 S_0 到结点 G 的路径，它是一个从初始状态到目标状态的有限的操作算子序列$\{O_1, O_2, \cdots O_k\}$，称为**求解路径**。**问题的解往往并不唯一。**

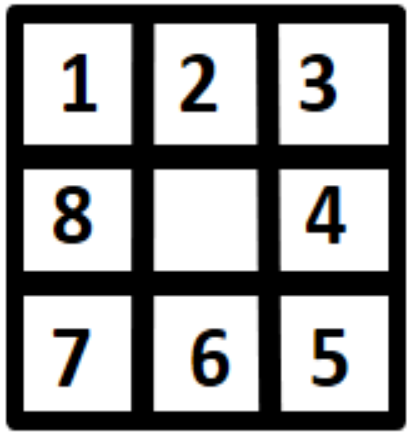


例2.1 八数码问题

又称为重排九宫问题。**首先，需要定义八数码问题的状态集合。**



（a）初始状态



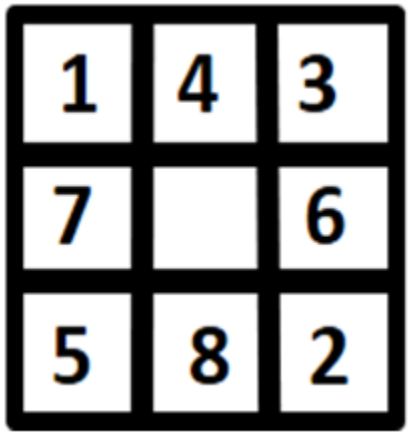
（b）目标状态

- 八个数码的任何一种摆法就是一个**状态**。
- 八数码的所有摆法构成了状态集合S，它们构成了一个**状态空间**.
- 这个状态空间中可以有 9！个状态。

例2.1 八数码问题

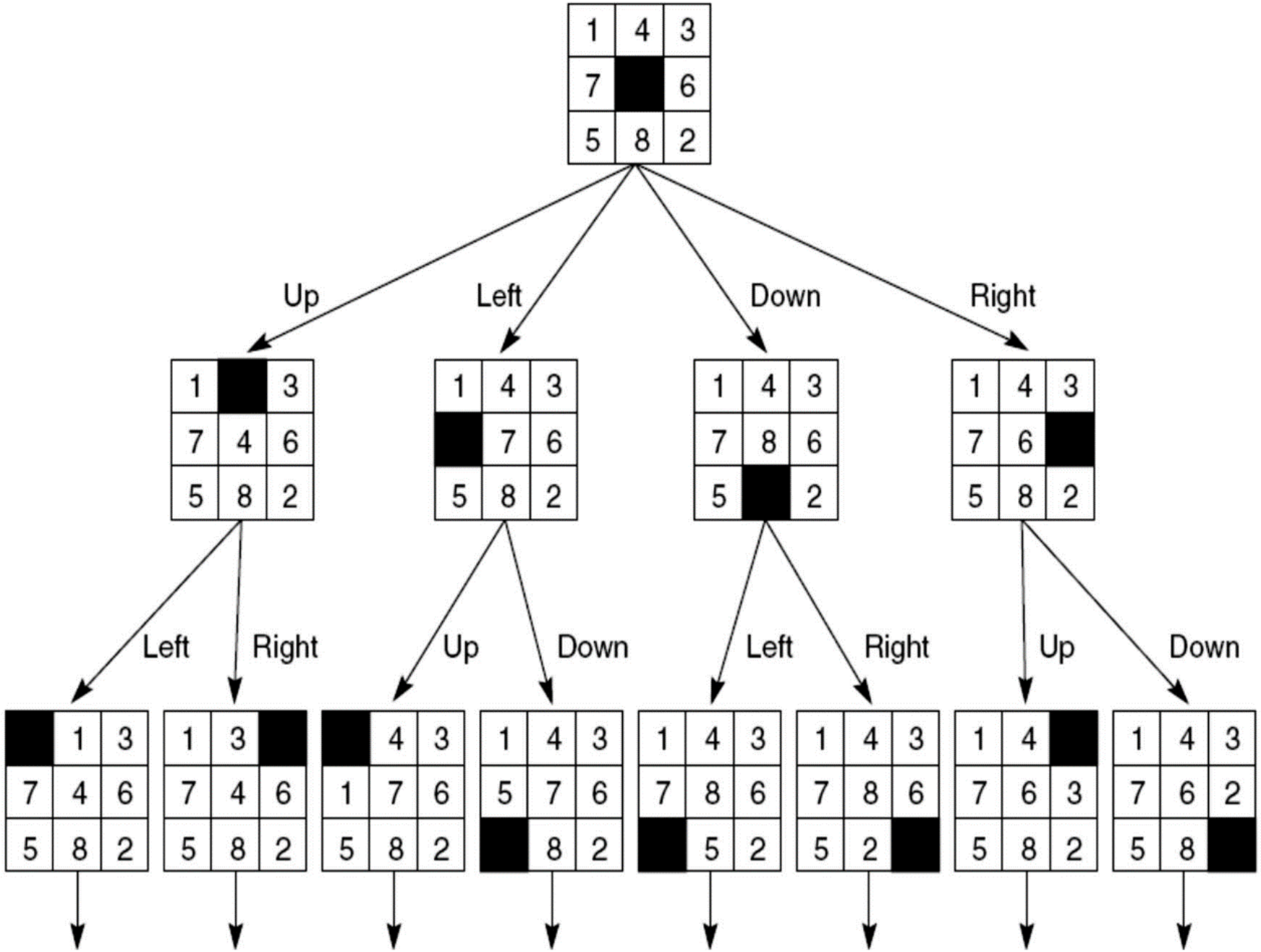
然后，设计操作集合：

将移动空格作为操作，即在方格盘上移动数码等价于移动空格。



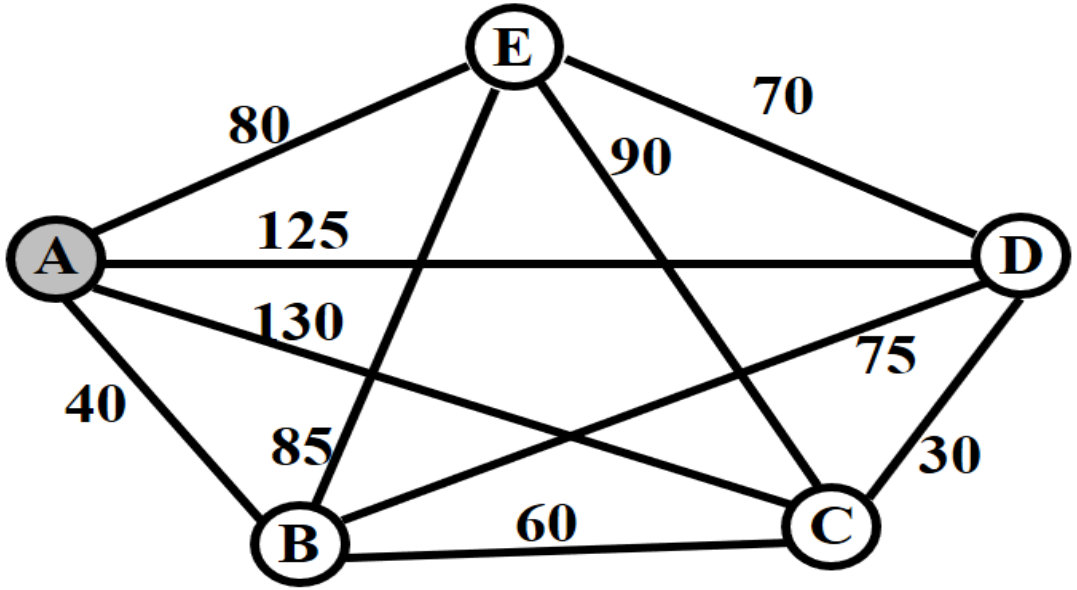
Up： 将空格向上移，if 空格不在最上一行
Down：将空格向下移，if 空格不在最下一行
Left： 将空格向左移，if 空格不在最左一列
Right： 将空格向右移，if 空格不在最右一列

- 八数码问题的**解**就是一个使棋盘从初始状态变化到目标状态的数码牌移动序列。
- 显然，八数码问题的**解并不是唯一的**；
- 可以附加一些**约束条件**，例如要求找到一个移动数码牌次数最少的解。


1 4 3
7 6
5 8 2
Up
Left
Down
Right
1 3
7 4 6
5 8 2
1 4 3
7 6
5 8 2
1 4 3
7 8 6
5 2
1 4 3
7 6
5 8 2
Left
Right
Up
Down
Left
Right
Up
Down
1 3
7 4 6
5 8 2
1 3
7 4 6
5 8 2
4 3
1 7 6
5 8 2
1 4 3
5 7 6
8 2
1 4 3
7 8 6
5 2
1 4 3
7 8 6
5 2
1 4
7 6 3
5 8 2
1 4 3
7 6 2
5 8

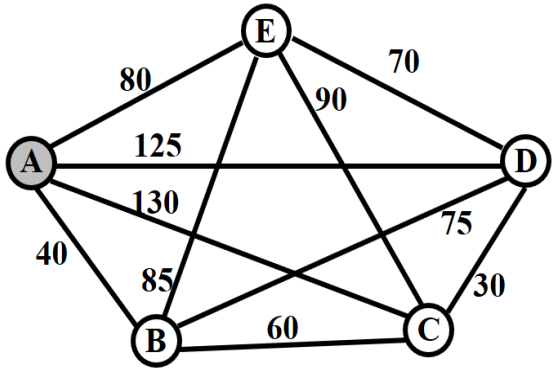
例2. 2 旅行商问题

- 从A城出发回到A城， 每个城市必须且只走一次，问该问题的可行解(我们经常遇见的旅行商问题是求最短路径)

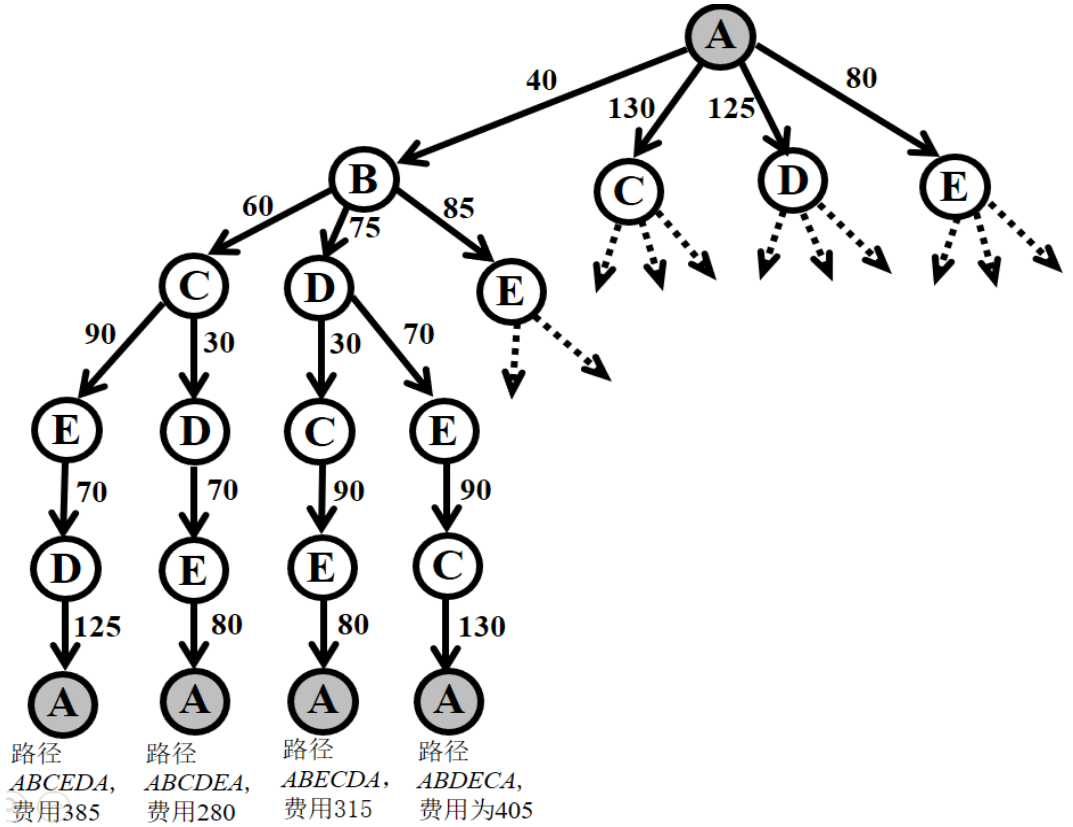


- 旅行推销员问题。一个推销员要到N个城市去推销产品，已知每对城市之间的距离，他从一个城市出发，访问所有城市后，回到出发地。除了出发地，要求每个城市仅经过一次。所要求解的问题是：应该如何设计一条行进路线，才能使得推销员访问每座城市所经过的路径最短或者费用最少?
- 旅行商问题实质是：在一个带有权重的、含有N个结点的完全无向图中，找一个权值最小的哈密尔顿（Hamilton）回路。

例2. 2 旅行商问题



旅行商问题的部分状态空间



2.4 状态空间表示法

- 对于大规模的问题，例如旅行商问题中有100个城市，要在有限时间内画出其全部状态空间图，是不可能的。
- 对于**简单问题**，可以采用有向图**直接画出状态空间**。
- 对于大多数**复杂的问题**，是根本**无法完全画出其状态空间**的，此时只需清晰定义状态变换的方式即可，也可以建模。

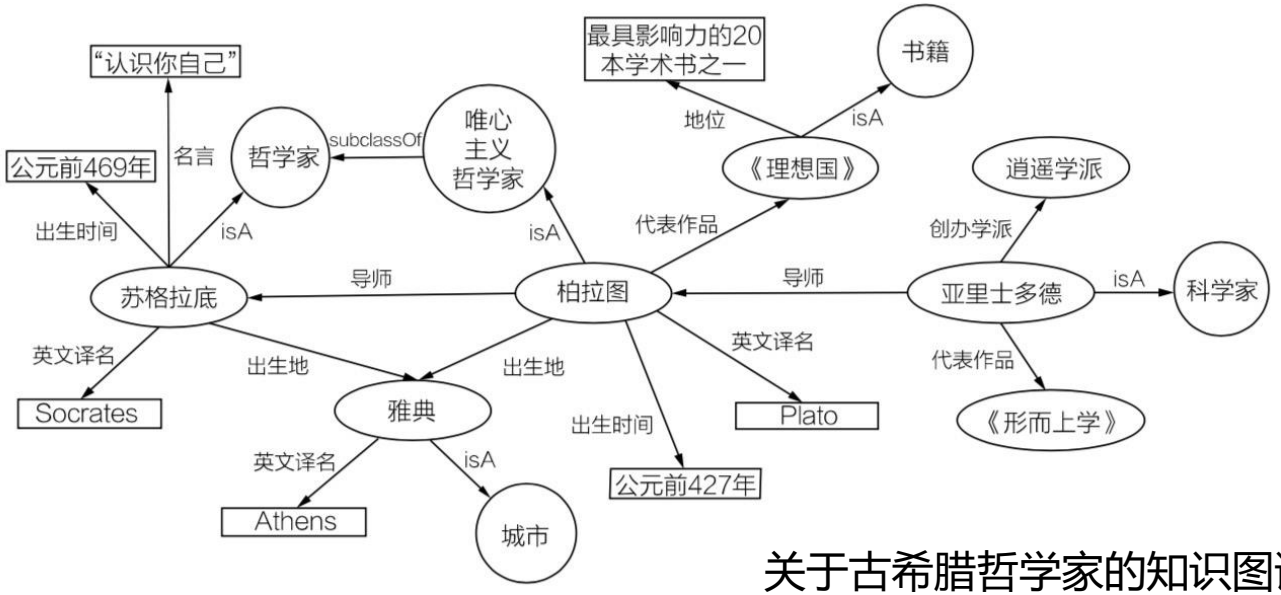
2.5 知识图谱

- **符号主义知识表示方法**经过历代人工智能科研人员的不断完善，**演变为知识图谱**这一符合互联网

2.5.1 知识图谱的定义

- 至今，知识图谱尚未有一个统一的定义。本质上，知识图谱是一种揭示客观世界中存在的**实体**（Entity）、**概念**（concept）及其之间**各种关系**的大规模语义网络，它以图结构表示知识，可理解为是一种描述语义知识的形式化框架， 知识图谱就是这样一类知识表示和应用技术的总称

- 知识图谱是一种**图结构的语义知识库**，组成单位是**实体**、**属性**和**关系**
  - **结点**表示实体（entity）或概念（concept）或属性值(attribute value)
  - 结点之间的**边**（edge）表示属性（attribute）或关系（relationship）
  - **边的方向**表示关系的方向
  - **边上的标记**表示属性名称或关系类型。



关于古希腊哲学家的知识图谱片段

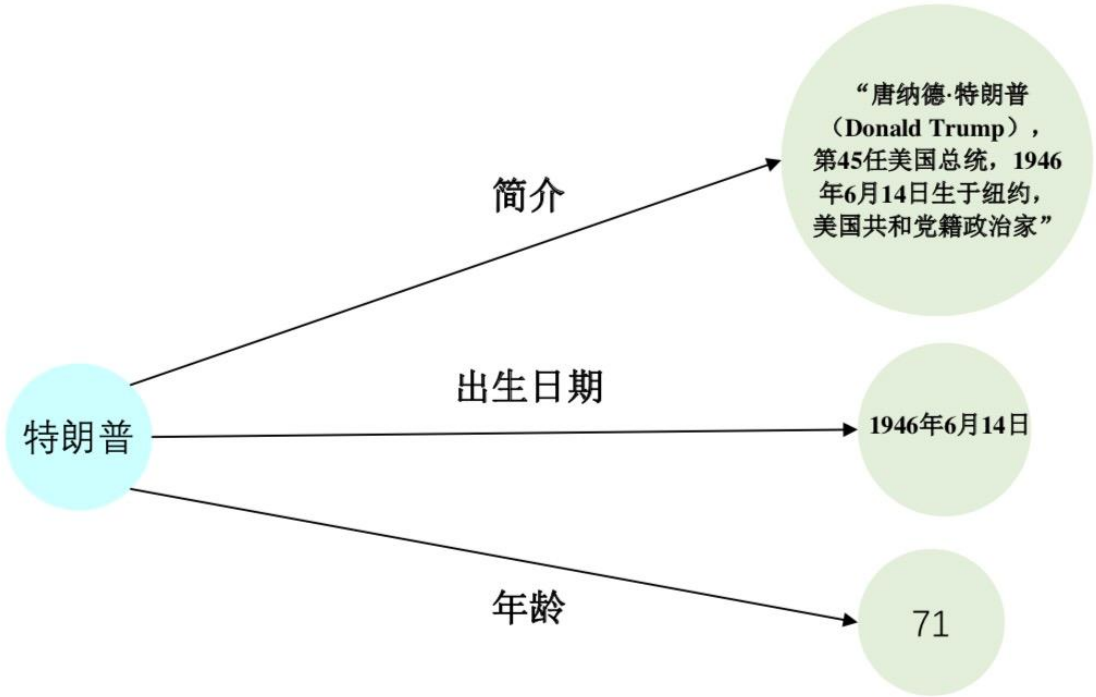
2.5.1 知识图谱的定义

(1) **概念**：也称为**类别**（Type）、**类**（Category或Class），**是某一领域内具有相同性质的对象构成的集合**，如在描述大学领域的知识图谱中，教师、学生和课程是必要的概念，而体育比赛领域中的概念则可能包括运动员、裁判员、教练、奖项等。概念主要用于表示集合、类别、对象类型、事物的种类。

(2) **实体**（entity）：有时也称为**实例**(instance)或**对象**（Object），实体是知识图谱中的最基本元素，是概念中的具体元素，它是**独立存在且可相互区别的客观事物**。例如，“C罗”是“足球运动员”这一概念的一个实例，“金球奖”是“奖项”这一概念的一个实例。

2.5.1 知识图谱的定义

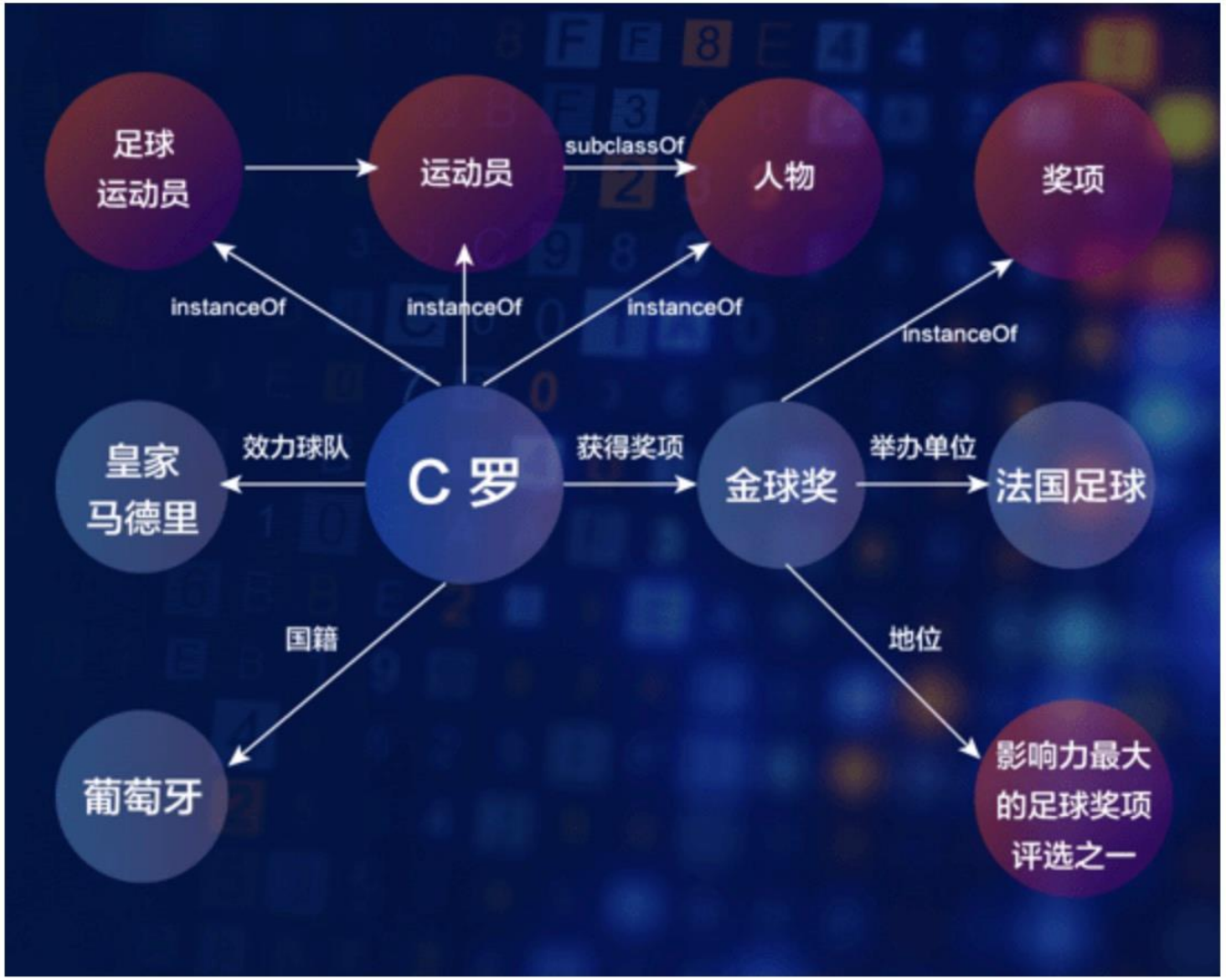
(3) **属性：描述实体或概念的特性或性质**。属性值可能是一个实体、一个字符串或一个数值。例如运动员的属性“国籍”的值是一个具体的国家（实例），属性“性别”的值是一个具体的字符串（male / female），而属性“身高”的值则是一个具体的数值。



2.5.1 知识图谱的定义

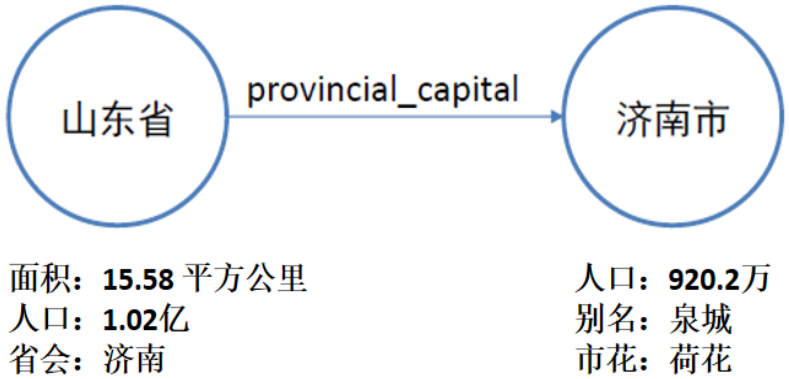
(4) **关系：是指概念之间或实体之间或概念与实例之间的联系**，如“运动员”与“足球运动员”两个**概念之间**存在的父类与子类（subclassOf）的层次关系；“车轮”和“汽车”两个**概念之间**存在的部分与整体（partOf）关系；“中国”与“北京”两个**实体之间**是“首都”关系；“国家”（概念）与“中国”（实体）间是实例化（instanceOf）关系。

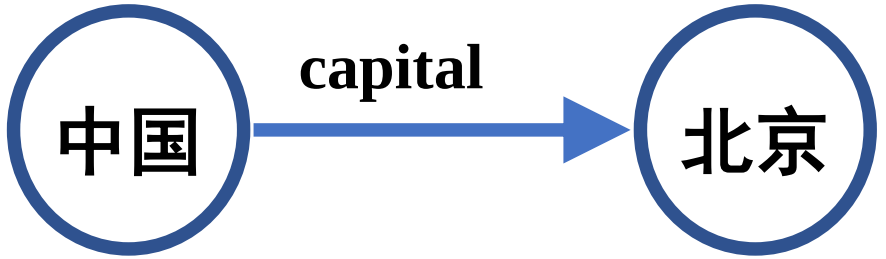
知识图谱示例


足球
运动员
运动员
subclassOf
人物
奖项
instanceOf
instanceOf
instanceOf
instanceOf
皇家
马德里
效力球队
C 罗
获得奖项
金球奖
举办单位
法国足球
国籍
地位
葡萄牙
影响力最大
的足球奖项
评选之一

2.5.2 知识图谱的表示

- 在典型的知识图谱中，每个实体或每个概念用一个全局唯一确定的ID 来标识，称为标识符(identifier)。
- **概念**和**实体**都是通过若干**属性**来刻画其内在特性。
- **概念之间**常见的关系有**父类与子类**（subclassOf）关系、**部分与整体**（partOf）关系
- **实体之间**的关系多种多样，不同实体之间存在不同的关系。例如，
  - “山东省”和“济南市”**两个实体**分别有各自的属性，两者之间存在“provincial_capital”的关系；
  - “中国”和“北京”**两个实体**之间存在“capital”的关系。所有实体和概念相互关联，形成复杂的“图”。





2.5.2 知识图谱的表示

知识图谱的一种通用表示方式是三元组，与事实性知识的产生式表示方法类似，也有两种形式：

（1）属性型联系：用“属性-值”对(Attribute-Value Pair，AVP)来描述一个实体具有某种内在属性，形式为　　（实体，属性，属性值）

例如，“山东省的面积是15.58平方公里”表示为

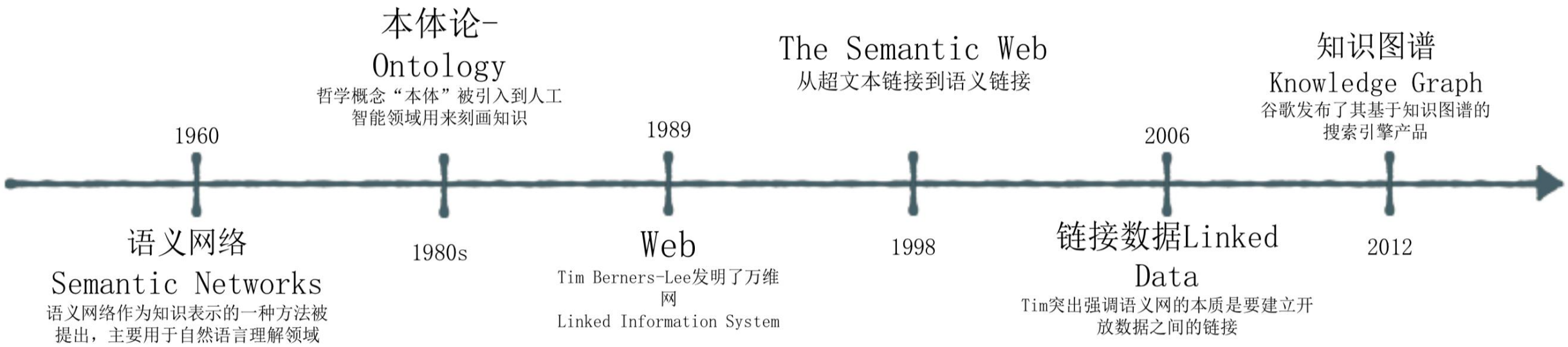
（山东，面积，15.58平方公里）

（2）关系型联系：描述两个实例之间的关系，形式为

（实体1，关系，实体2）

例如，“中国的首都是北京”表示为（中国，首都，北京）

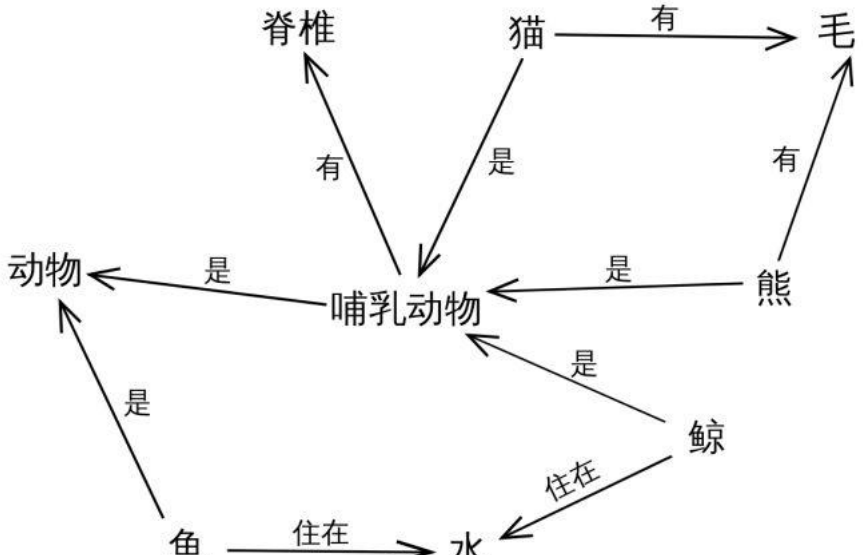
2.5.3 知识图谱的发展历史



1. 语义网络（Semantic Network）（1960s）

- 1968年，认知科学家Allan M. Collins和M. Ross Quillian等人提出了**语义网络**（semantic network，**不是语义网**）的心理学模型：模拟人的联想记忆。
- 随后，Quillian又将它用作人工智能中的一种知识表示方法。
- 语义网络：知识的结构化表示，由结点和弧构成。

采用**有向图**的结构来表示知识，其中**结点**表示**概念**（事件、事物），**弧**表示概念之间的**语义关系**



- 1960年代，剑桥大学的马斯特曼与其同事们还**将语义网络用于了机器翻译**。
- 1972年，**西蒙**在他的自然语言理解系统中采用了语义网络表示法。

语义网络（Semantic Network）特点

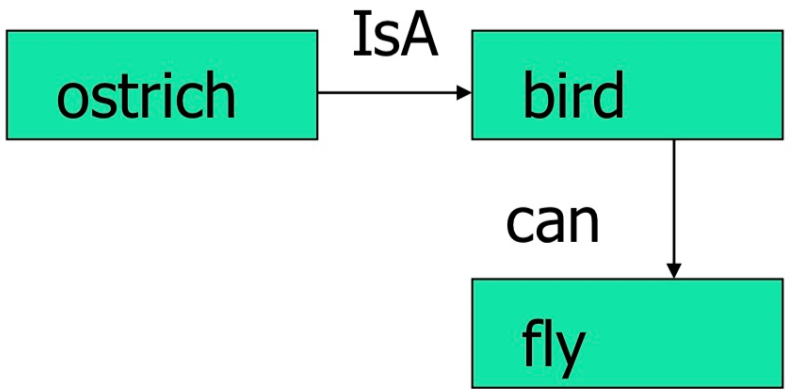
- **优点：**表达形式简单、直观、自然，因此容易理解和展示、相关概念容易聚类。
- **缺点：**

（1）没有定义结点与边的值的标准，完全由用户自己定义；

（2）无法区分**概念**结点和**实体**结点，如哺乳动物是个**抽象概念**，世界上并没有一种动物名字就叫哺乳动物；

（3）无法定义节点和边的标签；

（4）难以融合多源数据，不便于知识的共享。



上述缺点导致**语义网络难以应用于实践**。

2. 本体知识表示（1980s）

- 本体论是研究“存在”的科学，即试图解释存在是什么，世间所有存在的共同特征是什么，本体论的基本元素是概念及概念间的关系。
- 1980 年，“本体”这一哲学概念被引入人工智能领域中用于刻画知识，便产生了基于本体的知识表示方法，这种知识表示是一种“形式化的、对于共享概念体系的明确且详细的说明”。
- 本体显式地定义了领域中的概念、关系和公理（总是为真的陈述）及其之间的联系。
- 人工智能研究人员认为，他们可以创建基于本体的表示模型，从而进行特定类型的自动推理。
- 80年代出现了一批基于本体概念的知识库，例如，CYC和WordNet项目

苹果 **Apple**

3. 语义万维网知识表示（1990--2006）

- 语义万维网（Semantic Web）也称为语义Web或语义网，与语义网络（semantic network）的技术理念完全不同。
- 最主要的区别在于：语义网络知识表示与互联网无关，但语义万维网知识表示却是构建在万维网（world wide web）上的。
- 1963年，泰德·尼尔森（Ted Nelson）创造了“超文本（HyperText）”一词，其含义是用超链接的方法将各种不同空间的文字信息组织在一起的网状文本。
- 1969 年，因特网诞生于美国，它的前身“阿帕网”(ARPAnet)是一个军用研究系统，后来才发展成为覆盖五大洲 150 多个国家的开放型全球计算机网络系统，也称为互联网。

3. 语义万维网知识表示（1990--2006）

- 1989年，英国计算机科学家蒂姆·伯纳斯·李（Lee）创新性地提出了**将超文本用于因特网上**的构想，并于1990年与同事 Robert Cailliau合作发明了**万维网**（world wide web）技术。
- 蒂姆·伯纳斯·李被誉为**万维网之父**，于**2016**年荣获**图灵奖**。
- Web1.0 诞生后，互联网上的网页数量迅速增加，网页之间相互关联形成网络，其中蕴含着大量知识。
- 但这种知识的设计思想是面向人类阅读和理解的，无法被计算机理解和计算。比如我们很容易知道两个网页内容相关，但计算机很难理解网页的内容。
- 因此，蒂姆·伯纳斯·李又于**1998年**提出了“**语义万维网**（语义 Web）”的概念。

3. 语义万维网知识表示（1990--2006）

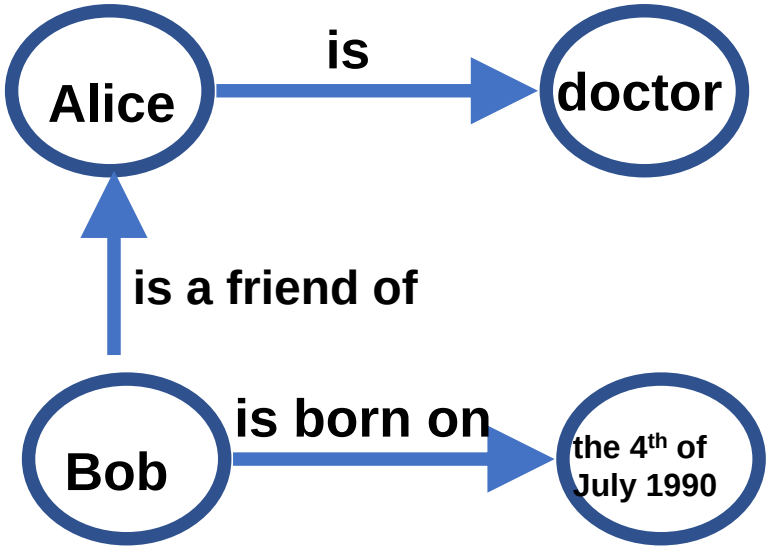
- **语义web**旨在对互联网内容进行**语义化表示**，通过对网页进行语义描述，得到网页的语义信息，从而使计算机能够**理解、推理**互联网信息。
- **语义web**是个庞大的构想，仅靠采用可扩展标记语言(extensible markup language，**XML**)标注web页面的数据内容是远远不够的，而是需要新的知识表示手段和方法。
- 在这样的背景下，科研工作者相继提出了“**资源描述框架**（Resource Description Framework，**RDF**）”和“**网络本体语言**（Web ontology language，**OWL**）”等面向 Web 的知识表示框架。

RDF

- 每个RDF陈述都包含**主语、谓词和宾语**，简称**SPO三元组**，其中主语和宾语分别表示两个资源，谓词表示两个资源间的关系。

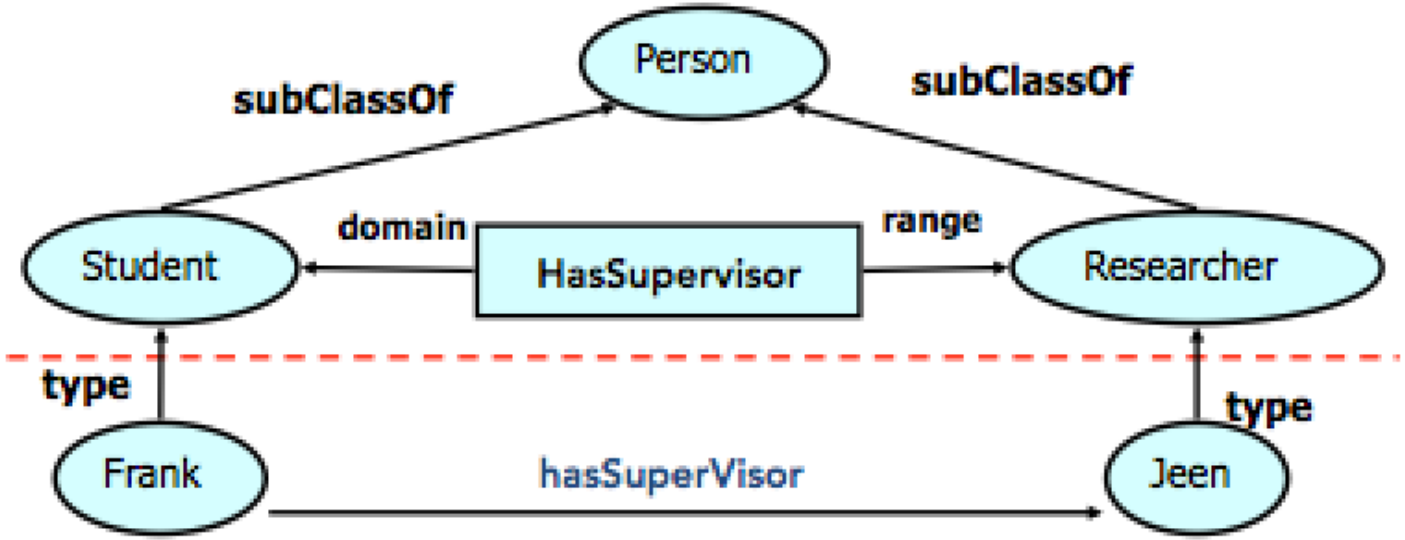
Bob, who is born on the 4th of July 1990 had a good friend, a doctor named Alice.

- <Alice> <is a> <doctor>，
- <Bob> <is a friend of> <Alice>，
- <Bob> <is born on> <the 4th of July 1990>

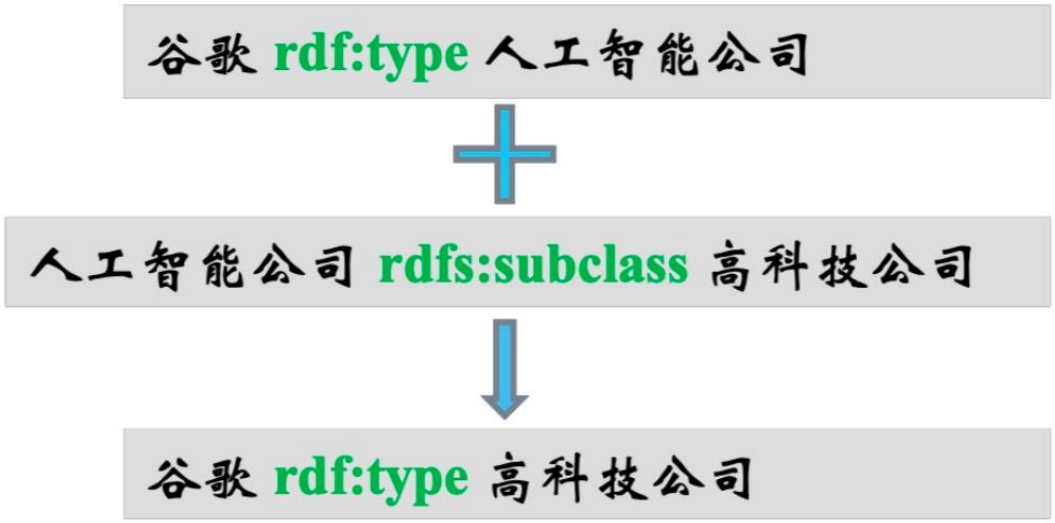


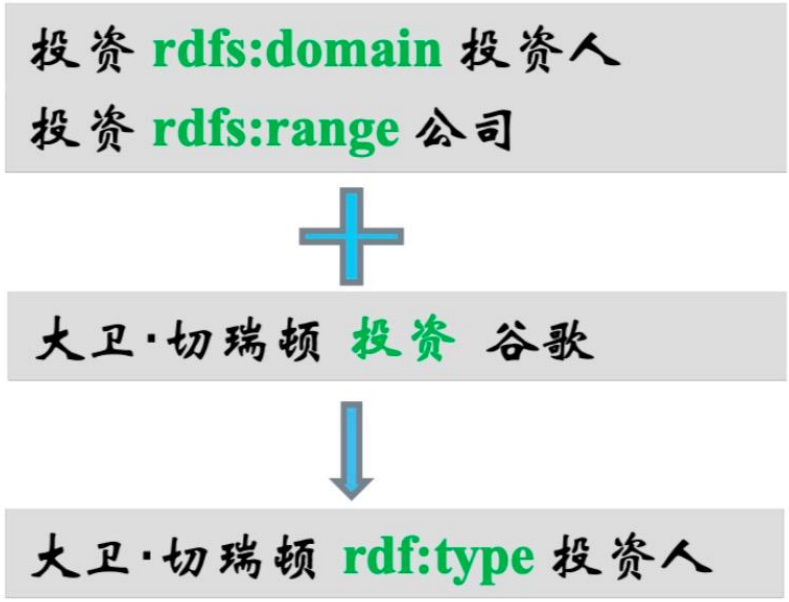
RDF

- **RDF的局限性**：**无法描述**类别和属性的**层级结构、包含关系**。
- W3C又推出了RDF schema（RDFs），在RDF词汇的基础上扩展了一套数据建模词汇（如，class、subClassOf、type、Property、subPropertyOf 等）来描述数据的模式层，可定义类的层次体系和属性体系，如类的继承.



基于RDFs的简单推理





OWL（Web Ontology Language）

- RDFs的表达能力仍不够强大，在**2001年**，W3C又开发了OWL。
- OWL 主要在 RDFs 基础上**扩展了表示类和属性约束的表示能力**，如：复杂类表达（intersection, union 和 complement 等）和属性约束（existential quantification, universal quantification, hasValue 等），使得能构建更为复杂且完备的本体。
- OWL比RDF具有更强的表达能力和推理能力。比如，OWL可以描述“中国所有湖泊”、“美国所有4000米以上的高山”这样的类。
- 但OWL**复杂度非常高**，在逻辑接近完美，但**工程上实现却太过复杂**。

语义网络与语义Web对比

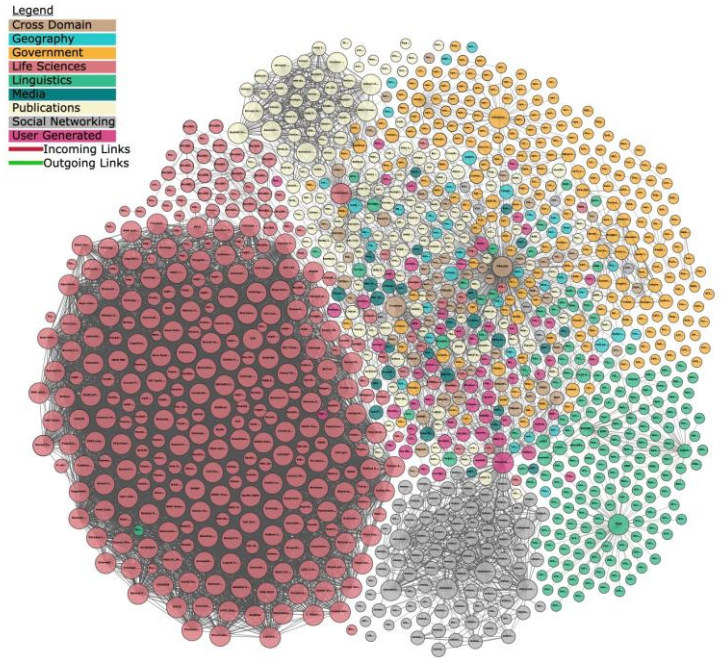
- 相较于语义网络，语义Web更加注重于描述万维网中资源、数据之间的语义关系。
- W3C针对**语义Web**制定的标准**解决了语义网络存在的不足**：
  - RDF为结点和边的取值提供了**统一标准**，为多源数据的融合提供了便利；
  - RDFs/OWL解决了**概念和对象的区分**问题，即定义了Class 和 Object (也称作 Instance, Entity)。
- **这些标准**从三个方面**完善了语义Web**：
  - 一是保证了语义Web的内容有准确的含义；
  - 二是保证了语义Web的内容可以被计算机理解并处理；
  - 三是计算机可从 Web 上整合各种网页中的内容信息。

“弱语义”到“强语义”

- 从2001年到2006年，随着RDF和OWL标准的提出，语义Web技术突飞猛进，各种标准不断升级和复杂化，技术栈层次不断加深，尤其是OWL的复杂程度很高，语义表达能力强大也导致了工程实现的难度大，构建知识库的成本越来越高。
- 期间，语义Web仍然沿袭着**符号主义的核心理念**，尝试建立完美的符号体系来涵盖所有知识。
- 该阶段是**从“弱语义”到“强语义”**的探索。

4. 链接数据（2006年起）

- 2006年，Lee逐渐意识到 **语义web 的发展遇到了瓶颈**，体系结构日益复杂，而工程实现难度越来越大，成本越来越高，各家单位都各自为政开发语义网。
- Lee提出了**链接数据（L**inked Data）的设想，号召各家单位分享自己的知识库，合并起来形成开放的语义网。
- 目前，该设想最大的项目是2007 年5月提出的**LOD**（Linked Open Data），至今，LOD中已经包含了1000多个数据集。https://lod-clou



4. 链接数据（2006年起）

- 自从实践数据链接开始，**在技术层面**，语义web开始**弱化“语义推理”**的功能，而更强调“Web”的作用，即**侧重数据的互联互通**，因此linked data可以看作是语义Web的一个简化集合。
- 在**实现层面**，linked data提倡使用RDF三元组形式描述知识，很少使用理论更完备的OWL系列方法，降低了实现数据链接的技术难度。
- 自此，语义Web开始进入**“弱语义”**的阶段，语义Web的体系结构开始**向知识图谱过渡**发展。
- **“弱语义”是指：**只强调词与词之间存在的语义关系，而不再强调

5. 知识图谱的正式提出（2012年）

- 2012 年 5 月 17 日，Google 正式提出了知识图谱的概念，发布了称之为“知识图谱”的项目，其初衷是为了优化其搜索引擎返回的结果，增强其搜索引擎的信息检索能力，提高用户搜索质量及体验。
- 至此，现代的知识图谱正式登上时代舞台。**谷歌知识图谱进一步弱化了语义，仅保留了RDF三元组的基本形式**，但这种简单的形式非常适合工程应用，以及知识的自动化生成。因此近年来展现出蓬勃的生命力。

2.5.4 典型的知识图谱

从早期人工构建的知识库发展到如今自动构建的知识图谱，期间大致可以划分为“强语义”和“弱语义”阶段，不同发展阶段有各自典型的代表项目。

- **“强语义”阶段的典型知识库**

  是从二十世纪六十年代到2006年，期间，重点研究如何建立语义表示体系，知识库的构建往往依赖于**专家制定、人工添加、合作编辑**的模式。此阶段典型知识库应用有：Cyc、WordNet、HowNet和ConceptNet。

- **“弱语义”阶段的典型知识图谱**

  自2006年起进入互联网时代后，随着知识库规模的不断增大，搜索引擎成为获取信息的主要手段，人们更多关注的是“是否存在某种知识，且能否找到某种知识”，而不是“是否可以理解、推理某种知识”。显然，这种需求使得知识库越来越**倾向于“弱语义、大规模”，不再强调逻辑复杂的语义及其推理，**而是**强调**如何利用互联网知识**自动构建大规模知识图谱**。

2.5.4 典型的知识图谱

- OpenCyc:23.9万个实体，1.5万个关系属性，209.3万个事实三元组
- WordNet:155, 327个单词，同义词集117,597个，同义词集之间由22种关系连接
- ConceptNet(5.0): 21种关系，包含约2800万三元组关系描述
- HowNet: 2000多个义原标注了约 10 万个中文/英文词或短语
- Freebase:4000多万实体，上万个属性关系，24多亿个事实三元组
- DBpedia:400多万实体，48,293种属性关系，10亿个事实三元组
- YAGO2 :980万实体，超过100个属性关系， 1亿多个事实三元组
- BabelNet: 284种语言、610多万个概念、960多万个实体、1500多万同义词组、13亿多个词汇和语义关系
- 百度百科:词条数1000万个
- 互动百科:800万词条，5万个分类，68亿文字
- Kinships:描述人物之间的亲属关系，104个实体，26种关系, 10,800个三元组
- UMLS:医学领域，描述医学概念之间的联系，135个实体，49种关系，6,800个三元组
- Cora:2,497个实体，7种关系，39,255个三元组
- NELL: 519万实体，306种关系， 5亿候选三元组
- Knowledge Vault: 4500万实体，4469种关系，2.7亿三元组

2.5.4 典型的知识图谱

类别	名称	
词法知识图谱	WordNet	wordnet.princeton.edu
	HowNet	
常识知识图谱	Cyc	https://www.cyc.com/
	ConceptNet	https://conceptnet.io
世界知识图谱	WikiData	https://www.wikidata.org/
	Freebase	
	DBpedia	dbpedia.org
	YAGO	yago-knowledge.org
中文知识图谱	百度知心	www.baidu.com
	搜狗知立方	www.sogou.com
领域知识图谱	AMiner	aminer.org
	FOFA	www.foaf-project.org/
	阿里商品	www.alibaba.com

2.5.5 知识图谱的应用

- 知识图谱为互联网上海量、异构、动态的大数据提供了一种更有效的表示、组织、管理及利用的方式，提高了网络应用的智能化水平，更加接近于人类的认知思维。
- 知识图谱已成为知识驱动的智能应用的基础设施，且已在许多领域中有了较为成功的应用。例如：语义搜索、知识问答。
- 此外，知识图谱还应用于大数据分析与决策、推荐系统以及一些垂直行业（教育、医疗、金融风险控制等）中，成为支撑这些应用发展的动力源泉。
- 目前，基于知识图谱的服务与应用已成为当前的研究热点，知识图谱与大数据、深度学习相结合，也已成为推动互联网和人工智能发展的核心驱动力之一。

知识图谱的应用：语义搜索

- 现代知识图谱最初提出的目的是增强搜索引擎的搜索结果，改善用户搜索体验。这就是“**语义搜索**”，是目前知识图谱最典型的应用方式。
- 在信息搜索方面，**传统的方法**是**基于关键词的搜索**。这种方式往往无法理解用户的意图，而是直接根据关键词给出若干网页。用户需要自己再次甄选，获取信息。
- 知识图谱引入搜索引擎之后，利用其推理技术，可以发现用户检索词的深层含义，从而以更精确的方式给出搜索结果。

知识图谱的应用：Google语义搜索的例子

知识图谱的应用：百度语义搜索的例子

知识图谱的应用：知识问答

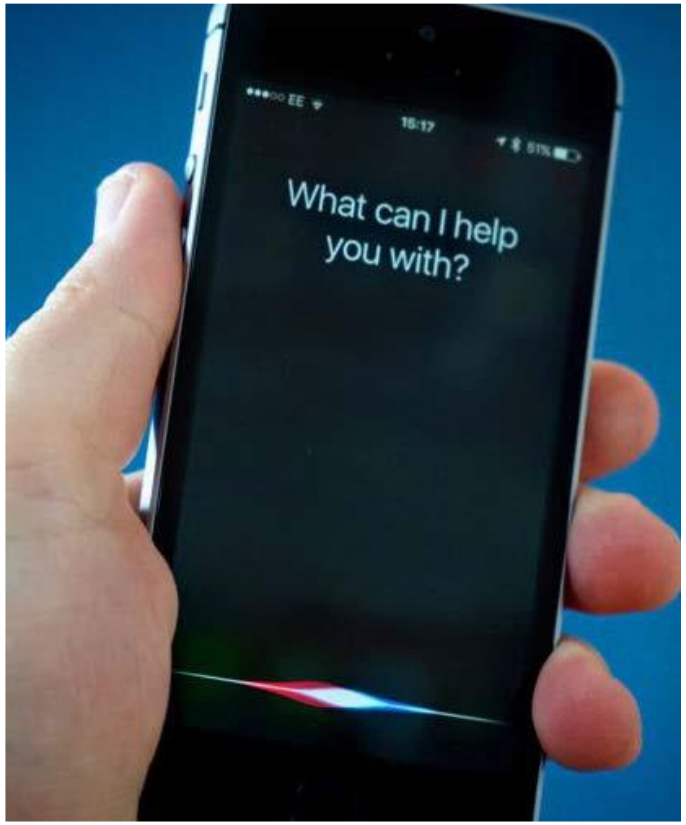
- 问答系统（Question Answering, QA）是指让计算机自动回答用户所提出的问题，是信息服务的一种高级形式。不同于现有的搜索引擎，问答系统返回用户的不再是若干相关文档，而是精准的、单一的语言形式的答案。

- 2011年，**IBM Watson** 在智力竞赛节目**“Jeopardy!”**中战胜人类选手，引起了巨大轰动。在后续几年，各大IT巨头相继推出以问答系统为核心技术的产品和服务，如 **Siri**、**Cortana**、**百度小度**等。

知识图谱的应用：知识问答

- **尽管 IBM Watson 系统**在战胜了人类选手，但是在2011年，其核心技术仍然是“检索”模式，即在大规模数据库中直接搜索答案。
- 近几年，随着知识图谱规模扩大和技术成熟，研究者逐步开始利用知识图谱回答问题，也就是我们要介绍的“知识问答”。
- 知识问答实现过程**分为两步**：
  - **提问分析**：将用户提问语言中的语义、意图提取出来，形成可供三元组推理使用的“查询”
  - **答案推理**：将该“查询”与知识图谱中的三元组进行检索、匹配或推理，获取正确答案

知识图谱的应用：知识问答



苹果公司siri

微软 cortana

百度小度

2.5.5 知识图谱的应用

- 知识图谱为互联网上海量、异构、动态的大数据提供了一种更有效的表示、组织、管理及利用的方式，提高了网络应用的智能化水平，更加接近于人类的认知思维。
- 知识图谱已成为知识驱动的智能应用的基础设施，且已在许多领域中有了较为成功的应用。例如：语义搜索、知识问答。
- 此外，知识图谱还应用于大数据分析与决策、推荐系统以及一些垂直行业（教育、医疗、金融风险控制等）中，成为支撑这些应用发展的动力源泉。
- 目前，基于知识图谱的服务与应用已成为当前的研究热点，知识图谱与大数据、深度学习相结合，也已成为推动互联网和人工智能发展的核心驱动力之一。

- 典型的知识图谱阅读材料（2.5.4）

1984，CYC 知识库

- 第一个例子，叫做CYC，是早期知识库项目的代表。也是目前持续时间最长的知识库项目。CYC 最早由 Douglas Lenat 在1984年创建，并延续至今。
- CYC 最初的目标是要建设人类最大的常识知识库，它认为，常识可以通过“实体”和“断言”来描述。类似于“每棵树都是植物”、“植物最终都会死亡”。
- CYC知识库的知识是以一阶谓词逻辑的形式存储的。
- CYC 设想，当用户提出“树是否会死亡”的问题时，CYC推理引擎可以通过自动推理得到正确的结论。

1984，CYC 知识库

- CYC 项目的知识事实主要通过手工添加到知识库中，类似定理库。这使得CYC的推理效率很高，可以支持复杂推理。但缺点同样突出：构建成本太高，知识更新慢，推理死板，适应性差。
- 近几年，CYC也开始通过机器学习来自动获取知识。截至目前，该知识库仍在运行，目前已经包含了700万条人类定义的断言，涉及50万个实体，15000个谓词。
- 目前在其官网上还提供了免费的版本 openCYC。有兴趣的同学可以关注一下。
- CYC 官网（https://www.cyc.com/）

1985，WordNet

- 第二个知识库是WordNet，也是目前**知名度最高**的词典知识库，它最早于**1985年**，由普林斯顿大学的认知科学实验室主持构建，最开始的目的是针对多义词的词义消歧。
- Wordnet 认为，每个词（word）可能有多个不同的语义（sense）根据词去组织词典，则会忽略同义词信息。
- 同样，每个语义（sense）也可能对应多个词。如果按照sense组织词典，**把语义近似相同的词打包放在一起**，是否可以解决多义词问题?
- 据此，WordNet设计了**同义词集合(Synset)**，作为基本单位来组织词典。
- WordNet 朴实的官网（ https://wordnet.princeton.edu/）

WordNet 的缺点

- WordNet的注意力不是在文本和话语水平上来描述词和概念的语义，因此WordNet 中**没有考虑特定语境下的相关概念之间的联系**。
  - 例如，WordNet 中没有将网球拍、网球、球网等词语以联系到一起。这就是著名的“tennis problem” （**网球问题**）。
  - 类似还有医生、医院之间的关系；
  - 教师、学生、学校之间的关系；
  - 大海、沙滩之间的关系等。
- **网球问题**涉及到许多世界知识的描述和关联，也是目前通用人工智能亟待解决的问题之一。

1999，ConceptNet

- 第三个知识库，是 ConceptNet，它最早源于MIT媒体实验室的 OpenMind commonsense 项目，该项目是由明斯基1999年创建的（这个明斯基就是达特茅斯会议的那个，神奇的老头）。
- ConceptNet 最初的目标是构建一个描述人类常识的大型语义web。
- 在1999年，RDF技术已经成熟，因此 ConceptNet 直接采用三元组的形式来构建，而不是谓词逻辑。
- 在构建方法上， ConceptNet并不是完全由专家来制定结构、层级、语义体系，而是通过“众包”方式，结合一定的文本抽取，半自动半人工地构建。

1999，ConceptNet

- 在ConceptNet中，所有的概念都来自于真实文本，概念之间的关系通过文本的统计数据确定。比如，在文本中多次出现“化妆... ... 漂亮”则可以推断“化妆”和“漂亮”之间存在导致关系。
- 这种从文本中发现的关系，并不是由专家事先制定好的。这就意味着，ConceptNet 本身已经**是一个“弱语义”的知识库**，只强调词与词之间存在的关系，而不再强调知识库整体的语义完整性。
- 经过多年发展，目前ConceptNet的主流版本已经升级到5.0，在onceptNet5.0中，一共定义了21种关系，包含约2800万三元组，支持多种语言。

HowNet（知网）

- 以上介绍的知识库都以英文为主。
- 近几年也开始扩展到中文，如 wordnet、conceptnet 都已经加入了中文词汇。
- **HowNet 是纯中文的知识库。**
- **知网**（HowNet）是一个以汉语和英语的词语所代表的概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库
- HowNet 还在持续发展中，并且获得越来越多的关注。目前也已经有公开版本 OpenHowNet 问世。

HowNet 的基本思想

- HowNet 最早的理念可以追溯到**1988年**，知网的作者**董振东**先生曾在他的几篇文章中提出：
  - 自然语言处理系统需要知识库。
  - 知识库应包含概念、概念的属性、以及概念之间、属性之间的关系
  - 应首先建立常识性知识库，描述通用概念
  - 应由知识工程师来设计知识库的框架，并建立知识库的原型。
- 知网就是在这些理念的指导下，历经多年开发得到的中文知识系统。

HowNet 的构建

知网作为一个知识系统，是一个**网状结构**。知网的建设方法的一个重要特点是自下而上的归纳的方法。

- 知网知识体系的最底层，是800多个“**义元**”，是“**最基本的、不易于再分割的意义的最小单位**”。义元由人工专家大量阅读文本，逐步精炼得到，是 HowNet 的精华。
- 有了义元，HowNet 进一步**用义元来标注、解释事件和概念。然后加入概念、属性之间的关系，构成网络**。

知识图谱规模化挑战

- 进入互联网时代后，尤其是**搜索引擎成为人们获取信息的主要手段**以后，工业界对知识库的规模提出了越来越高的要求。以往“小而美”的知识库，已经无法满足智能应用的需求。
- 另一方面，以搜索引擎为例，人们**更多关注**的是“**是否具有并且找到某种知识**”，而不是“**是否可以理解、推理某种知识**”。显然这种需求，使得知识库**越来越倾向于“弱语义、大规模”**。
- 因此，从2010年开始，许多学者开始尝试利用机器学习、信息抽取等技术，自动从互联网获取词汇知识。

自动获取web知识：知识源瓶颈

- 典型的例子包括
  - 华盛顿大学的TextRunner（现改为OpenIE，开放信息抽取系统）
  - 卡内基梅隆大学的NELL (Never-Ending Language Learning)
- 这两个系统，都是完全根据算法，以互联网网页上的文本为知识源，试图自动分析、发现其中的概念以及概念之间的关系。
- 这样做的**好处**是**很容易获得大量知识**。
- **缺点**在于开放互联网上的信息质量差别大，数量虽然庞大，但知识密度非常低，使得**系统准确率**和**知识获取效率都比较低**。

另一条路：Wikipedia

- 可以说：**自动构建**知识库，**前提**是准备好知识密集、格式统一、大规模的知识源。
- 在2010年前后，随着在线百科网站的兴起，这种知识源逐渐成熟，其典型代表就是大名鼎鼎的**维基百科**，wikipedia。

在线百科全书 Wikipedia

- 维基百科是世界上最著名的在线百科全书，它致力于向读者提供免费的百科全书知识。
- 在线百科全书的概念来**自理查德•斯托曼**（同时他也是开源软件的倡导者、精神领袖）
- Wikipedia 始于2001年1月15日，目前发展为全球性的项目
- **Wikipedia 特点**：众包、词条存储、累计有千万级别的百科词条。

基于在线百科的知识图谱

- 在 wikipedia 取得成功之后，大批在线百科网站兴起，在维基百科的“在线百科全书列表”词条中，记录的目前知名的在线百科网站已经达到139个。
- 这些网站以基本相同的结构，存储了大量词条以及描述文本。其中包含了方方面面的知识，为知识图谱自动构建奠定基础。
- 目前，大多数通用知识图谱，也都采用类似的方法，通过对在线百科网站进行自动分析，构建知识图谱。

从 FreeBase 到 Wikidata

- Freebase 是较早期的开放共享知识库。由硅谷创业公司 MetaWeb 在2005年启动。其主要数据来源包括维基百科、世界名人数据库、开放音乐数据库，以及社区用户的贡献等。
- 早期的 FreeBase 以人工转化为主，即，**由社区成员协作，将知识源中的知识提取，构建为 Freebase 格式的三元组**。
- **Freebase是典型的“弱语义”知识库**，它对知识库中的实体和关系不做严格的控制，完全由用户来创建、编辑。
- 2010年，谷歌收购了Freebase 作为其知识图谱数据来源，并于**2012年发布谷歌知识图谱**。
- 2016年，谷歌将Freebase 的数据**迁移至新的Wikidata**，正式关闭了

从 FreeBase 到 Wikidata

- 在关闭前，freebase 大约包含了6800万个实体、约10亿条关系，超过24亿条三元组。
- 作为继任者，**wikidata** 对 freebase 的结构进行了改进，以提高质量，并**与 wikipedia 深度结合**，到2017年底已经具有2500万个词条，**是现代知识图谱的典型代表**。

DBpedia

- DBPedia是早期的**基于维基百科的语义网项目**。DBPedia 的本意就是指**数据库版本的 Wikipedia**，旨在将 wikipedia 的知识系统化、规范化、结构化。
- 与 Freebase 不同，**DBPedia定义了一套较为严格的语义体系**，其中包含人、地点、音乐、电影、组织机构、物种、疾病等类定义。
- 此外，DBPedia 还是LOD (**Linked Open Data**)计划的核心，与 Freebase，OpenCYC、Bio2RDF 等多个数据集建立了数据链接。
- DBPedia**采用RDF三元组模型**，2016年的版本中，已经包括了660万实体，130亿个三元组。

YAGO

- YAGO 是由**德国马普**研究所研制的知识图谱，主要集成了**Wikipedia、WordNet 和 GeoNames**三个来源的数据。
- YAGO的**特点**：将WordNet的词汇定义与Wikipedia的分类体系进行了知识融合，使得YAGO具有更加丰富的实体分类体系。
- YAGO 还考虑了**时间和空间**知识，为很多知识条目增加了时间和空间维度的属性描述。
- 目前，YAGO 包含1.2亿条三元组知识。
- 值得一提的是，**YAGO 是 IBM Watson 的后端知识库之一**。

BabelNet

- BabelNet的功能类似于WordNet，是个词汇知识库，是多语言百科全书式的字典和语义网络。
- BabelNet**特点：**将WordNet词典与Wikipedia多语言百科做知识融合，使得wordnet 支持更多的语言，**解决小语种 wordnet 数据缺乏的问题**。
- BabelNet **核心思想**是：许多Wikipedia 词条都具有多语言版本，因此如果wordnet 中的词条可以与wikipedia中的条目匹配，则相当于获得了**多语言版本的wordnet**。
- 目前，BabelNet 3.7包含了271种语言，包含1400万同义词组，36.4万词语关系，超过19亿的三元组，是目前最大规模的多语言词典知识库。

中文知识图谱的现在

- 中国中文信息学会语言与知识计算专业委员会于2015年发起和倡导了**开放知识图谱社区联盟（OpenKG.cn）项目**，OpenKG是公益性中立项目，旨在推动基于中文的知识图谱数据的开放、互联与众包，以及知识图谱算法、工具和平台的开源开放工作。

中文知识图谱的现在

- 近些年，以中文为主语言的知识图谱主要是基于百度百科和维基百科的结构化信息构建起来的，如哈尔滨工业大学的“大词林”、上海交通大学的“zhishi.me”、清华大学的“XLore”及复旦大学的“CN-pedia”等。

Thank you !