人工智能：知识表示和推理 III

饶洋辉
计算机学院，
中山大学
raoyangh@mail.sysu.edu.cn
http://cse.sysu.edu.cn/node/2471

课件来源：中山大学王甲海教授；浙江大学吴飞教授；海军工程大学贲可荣教授

知识表示和推理

1 谓词逻辑

2 归结推理

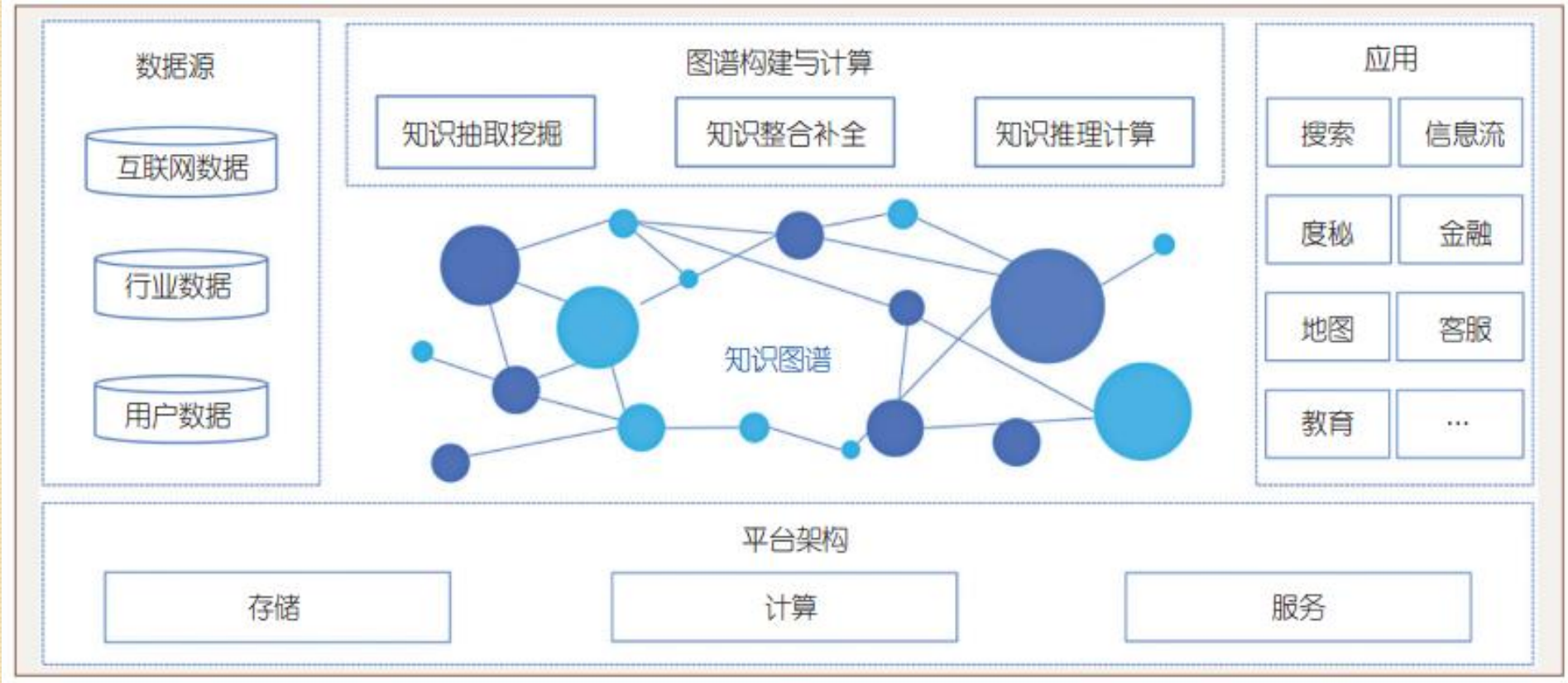
3 知识图谱

知识图谱技术

- 知识图谱技术是指在建立知识图谱中使用的技术，是融合认知计算、知识表示与推理、信息检索与抽取、自然语言处理与语义Web、数据挖掘与机器学习的交叉研究。
  - **知识表示**研究客观世界知识的建模，以便于机器识别和理解，既要考虑知识的表示与存储，又要考虑知识的使用和计算。
  - **知识图谱构建**解决如何建立计算机算法，从客观世界或者互联网的各种数据资源中获取客观世界的知识，主要研究使用什么样的数据和什么样的方法抽取什么样的知识。
  - **知识图谱应用**主要研究如何利用知识图谱，建立基于知识的系统并提供智能的知识服务，更好地解决实际应用问题。

李涓子，侯磊，知识图谱研究综述，中国人工智能学会通讯，2016，6(12): 38-43

百度知识图谱技术视图


数据源
互联网数据
行业数据
用户数据
图谱构建与计算
知识抽取挖掘
知识整合补全
知识推理计算
知识图谱
应用
搜索
信息流
度秘
金融
地图
客服
教育
…
平台架构
存储
计算
服务

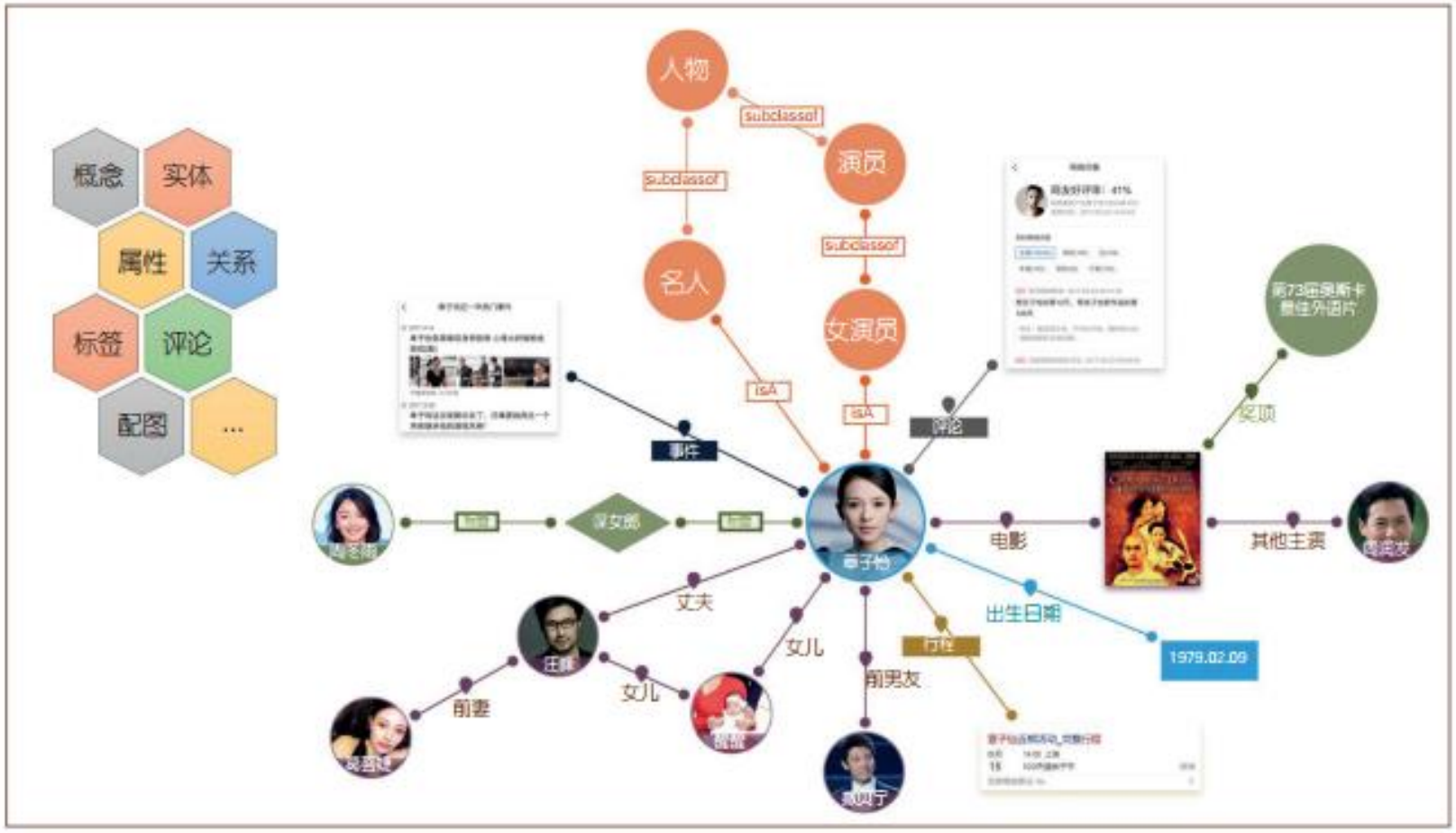
知识图谱及其表示

- 由于互联网内容的大规模、异质多元、组织结构松散的特点，给人们有效获取信息和知识提出了挑战。
- 谷歌于2012年5月16日首先发布了知识图谱（Knowledge Graph）。
- 知识图谱是一种互联网环境下的知识表示方法。
- 知识图谱的目的是为了提高搜索引擎的能力，改善用户的搜索质量以及搜索体验。
- Google、百度和搜狗等搜索引擎公司构建的知识图谱，分别称为知识图谱、知心和知立方。

知识图谱及其表示

- 知识图谱的概念于2013年以后开始在学术界和工业界普及，并在智能问答、医疗、反欺诈等应用中发挥了重要作用。
- 关于知识图谱的2种观点：
  - 知识图谱是一个实体-关系的有向图。
  - 知识图谱包含更抽象的概念之间的关系，例如，谷歌和必应、雅虎一起推出了 Schema.org，用来提供一个覆盖广泛主题（包括人物、地点、事件等）的模式（schema）。

知识图谱及其表示



- 知识图谱本质上是语义网络（semantic network）知识库

知识图谱及其表示

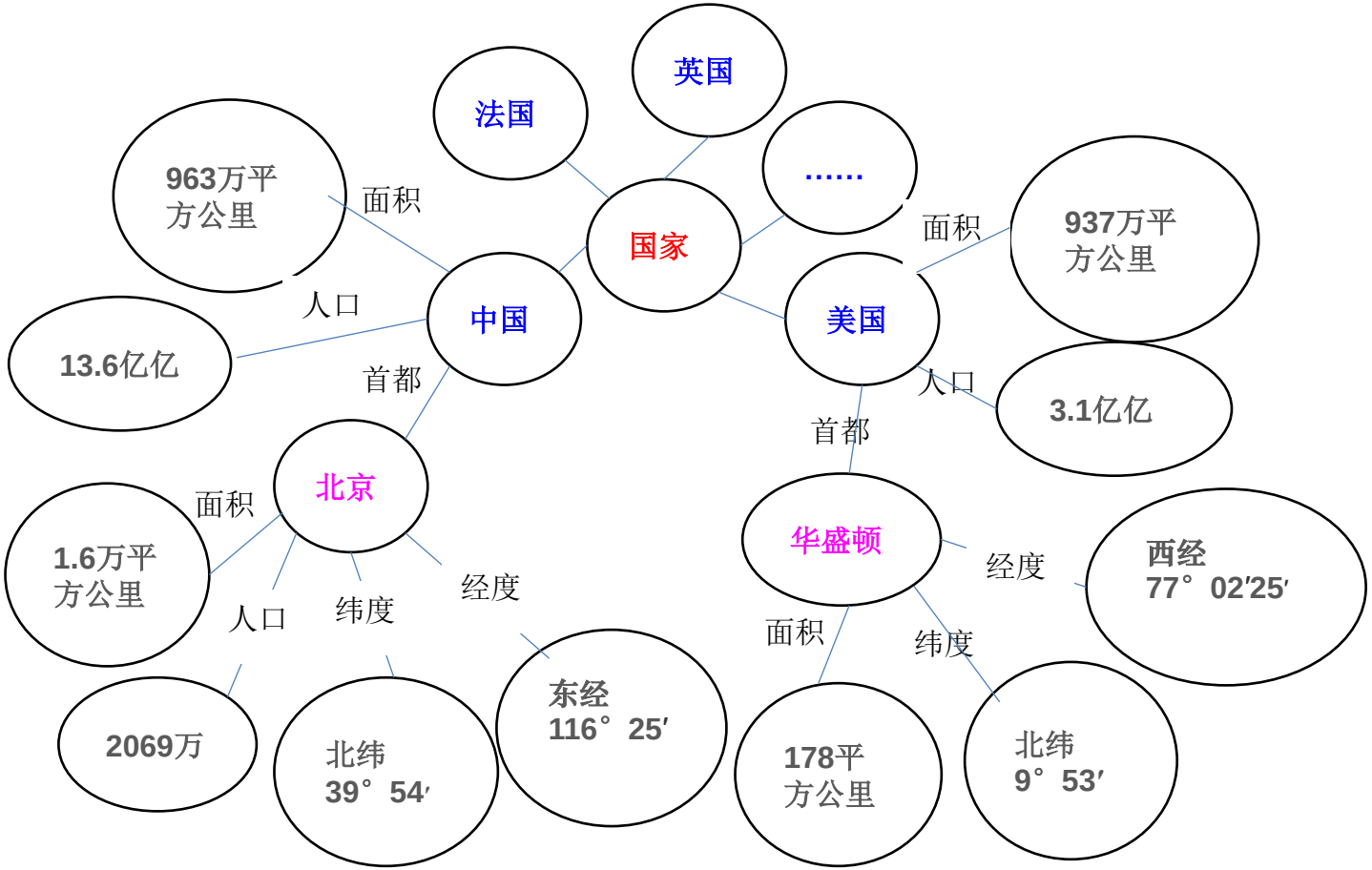
- 知识图谱（Knowledge Graph/Vault），又称科学知识图谱，是用各种图形等可视化技术描述知识资源及其载体，挖掘、分析、构建、绘制和显示知识及它们之间的相互联系。
- 知识图谱是由一些相互连接的实体及其属性构成的。
- 三元组是知识图谱的一种通用表示方式：

  (实体1-关系-实体2)：中国-首都-北京

  (实体-属性-属性值)：北京-人口-2069万

知识图谱及其表示

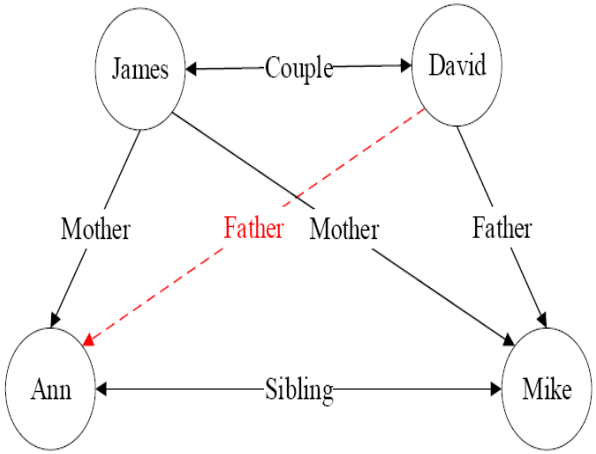
- **知识图谱也可被看作是一张图，图中的节点表示实体或概念，而图中的边则由属性或关系构成。**



知识图谱推理

- **知识图谱推理包括基于符号的推理和基于统计的推理等**
  - **基于符号的推理**一般基于经典逻辑（一阶谓词逻辑或者命题逻辑）或者经典逻辑的变异（比如说缺省逻辑）。基于符号的推理可以从一个已有的知识图谱推理出新的实体间关系；同时可以对知识图谱进行逻辑的冲突检测。
  - **基于统计的推理**一般基于机器学习方法，通过统计规律从知识图谱中学习到新的实体间关系；并且对新学到的关系进行评分，去掉那些可能错误的关系。

基于符号的知识图谱推理



一个简单的家庭关系知识图谱

- 可利用一阶谓词来表达刻画知识图谱中节点之间存在的关系，如图中形如<James,$Couple$,David>的关系可用一阶逻辑的形式来描述，即$Couple$(James, David)。
- $Couple(x,y)$是一阶谓词，$Couple$是图中实体之间具有的关系，x和y是谓词变量
- 从图中已有关系可推知David和Ann具有父女关系，但这一关系在图中初始图(无红线)中并不存在，是需要推理的目标。

$$(\forall x)(\forall y)(\forall z)\big(Mother(z,y) \wedge Couple(x,z) \rightarrow Father(x,y)\big)$$

前提约束谓词
（学习得到）

目标谓词
（已知）

基于统计的知识图谱推理

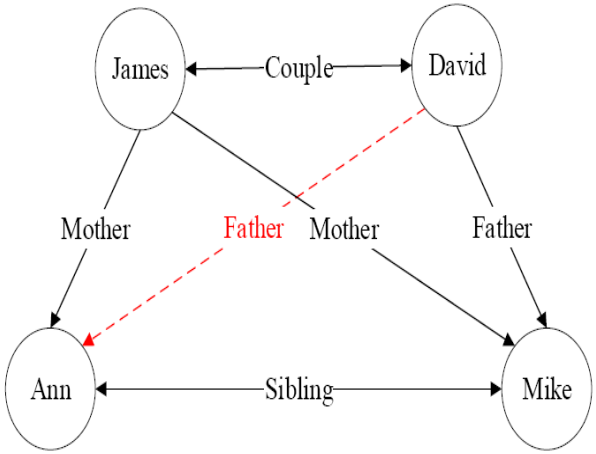
路径排序推理算法：将实体之间的关联路径作为特征，来学习目标关系的分类器。路径排序算法的工作流程主要分为三步：

(1) 特征抽取：生成并选择路径特征集合。生成路径的方式有随机游走（random walk）、宽度优先搜索、深度优先搜索等。

(2) 特征计算：计算每个训练样例的特征值$P(s \to t; \pi_j)$。该特征值可以表示从实体节点s出发，通过关系路径π_j到达实体节点t的概率；也可以表示为布尔值，表示实体s到实体t之间是否存在路径π_j；还可以是实体s和实体t之间路径出现频次、频率等。

(3) 分类器训练：根据训练样例的特征值，为目标关系训练分类器。当训练好分类器后，即可将该分类器用于推理两个实体之间是否存在目标关系。

基于统计的知识图谱推理



一个简单的家庭关系知识图谱

(1) 目标关系：$Father$

(2) 对于目标关系，生成四组训练样例，一个为正例、三个为负例：

正例：(David, Mike)

负例：(David, James)，(James, Ann)，(James, Mike)

(3) 采样得到路径，每一路径链接上述每个训练样例中两个实体：

(David, Mike)对应路径：$Couple \rightarrow Mother$

(David, James)对应路径：$Father \rightarrow Mother^{-1}$（$Mother$的反关系）

(James, Ann)对应路径：$Mother \rightarrow Sibling$

(James, Mike)对应路径：$Couple \rightarrow Father$

(4) 对于每一个正例/负例，判断上述四条路径可否链接其包含的两个实体，将可链接（记为1）和不可链接（记为0）作为特征，于是每一个正例/负例得到一个四维特征向量：

(David, Mike)：$\{[1,0,0,0],1\}$

(David, James)：$\{[0,1,0,0],-1\}$

(James, Ann)：$\{[0,0,1,0],-1\}$

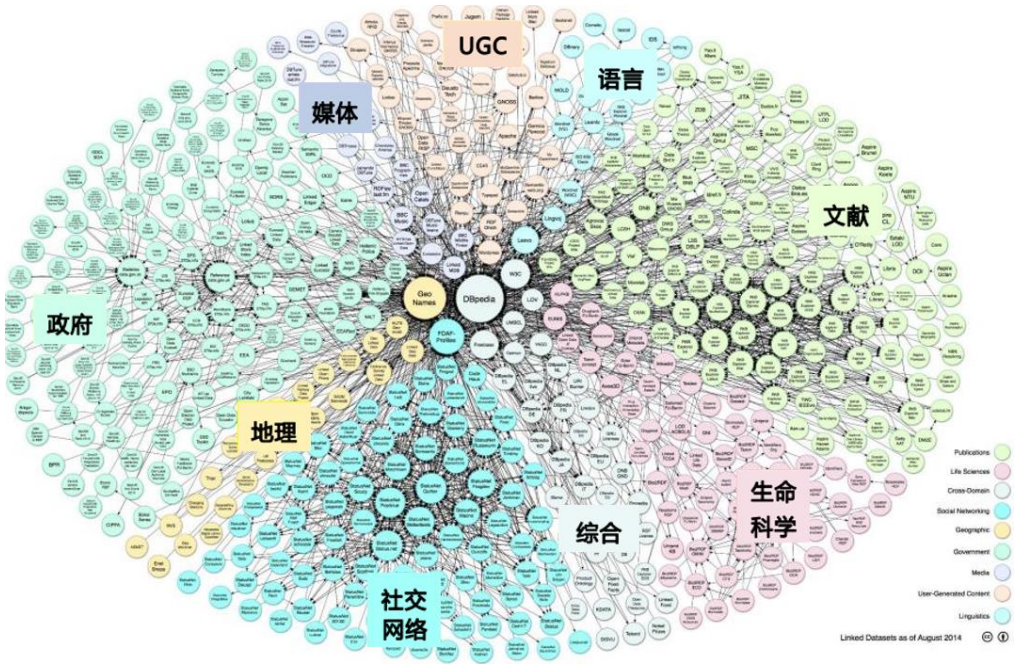
(James, Mike)：$\{[0,0,1,1],-1\}$

(5) 依据(4)中的训练样本，训练分类器M。

(6) 预测。对于形如(David, Ann)的样例，得到其特征值为$[1,0,0,0]$，将特征向量输入到分类器M中，如果分类器M给出分类结果为1，则$Father$(David, Ann)成立。

知识图谱的类型

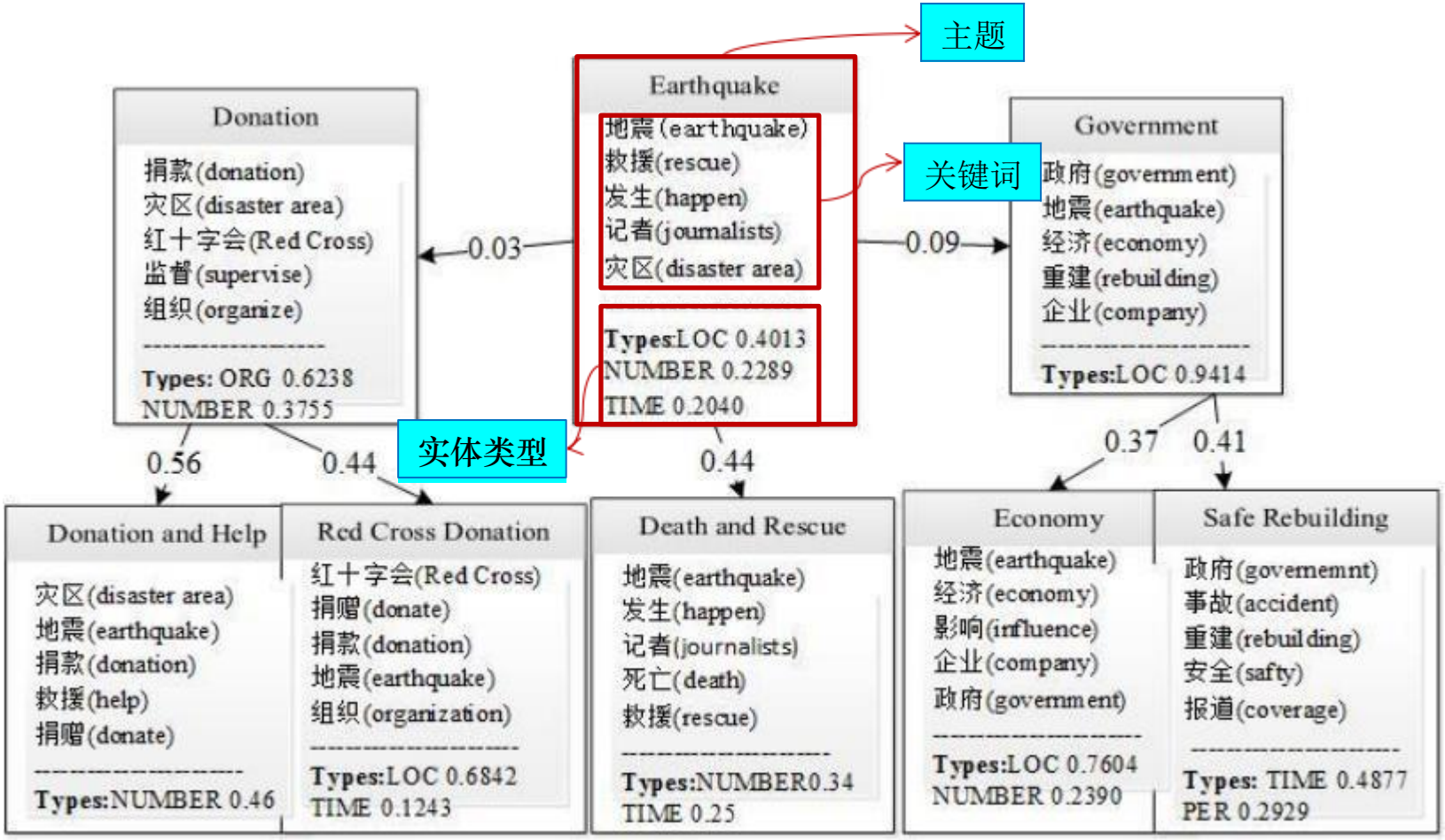
- 知识图谱是对处理数据的结构化结果表示。
- 知识图谱可以表达：
  - 实体及其关系知识
  - 事件知识
  - 计算知识
  - 限定领域知识
  - 面向特定任务的知识
  - 跨语言知识
  - ……



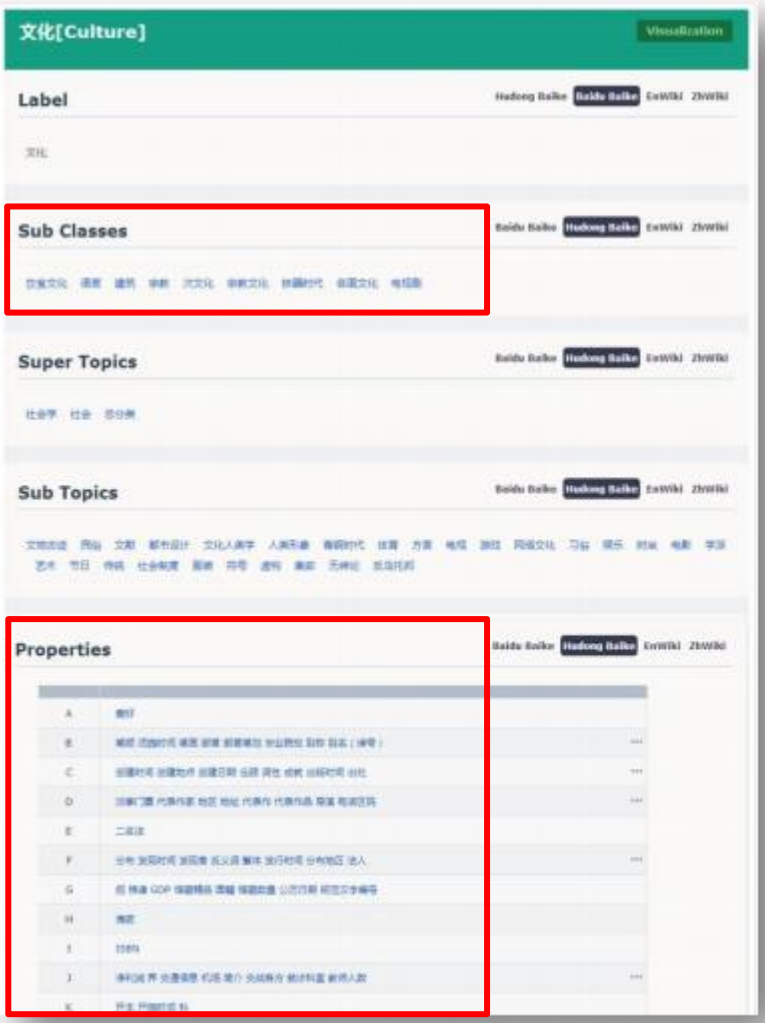
- 知识图谱是实现语义互操作的基础。

事件知识图谱

- 事件学习：从多个相似事件实例中学习层次主题模式



跨语言知识图谱



http://xlore.org/

知识图谱的构建

- （1）自顶向下指的是先为知识图谱定义好本体与数据模式，再将实体加入到知识库。

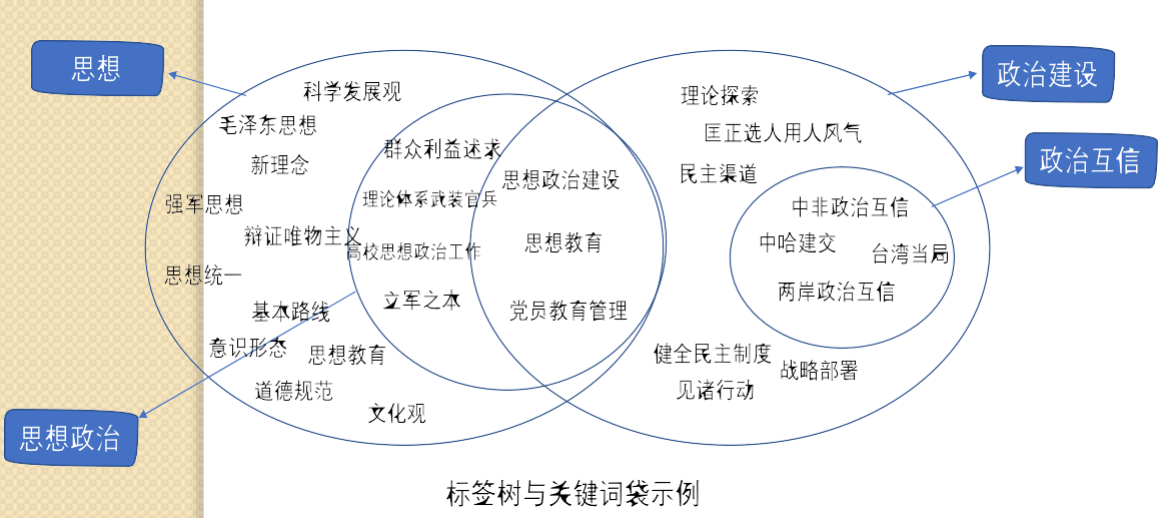
- （2）自底向上指的是从一些开放链接数据中提取出实体，选择其中置信度较高的加入到知识库，再构建顶层的本体模式。

知识图谱的应用

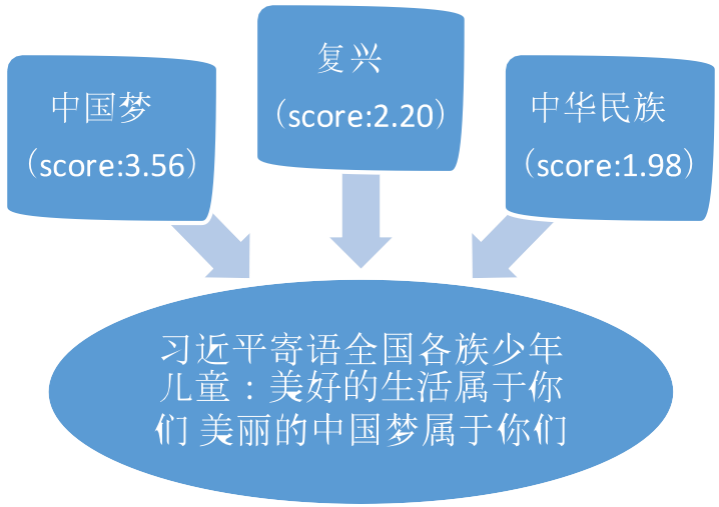
- 认知智能应用需求广泛多样，需要对传统信息化手段的全面而彻底的革新
- 认知智能：人类脑力解放，机器生产力显著提高
- 认知智能包括：
  - 精准分析
  - 智慧搜索
  - 智能推荐
  - 智能解释
  - 自然人机交互
  - 深层关系推理

精准分析

- 基于行业知识图谱，形成行业数据理解能力。
- 实现数据中的实体、概念、主题认知，实现可视化洞察。



标签树与关键词袋示例



总书记文章的标签体系自动构建

总书记文章自动标签化

智慧搜索

- 精准搜索意图理解
  - 精准分类、语义理解、个性化
- 复杂多元对象搜索
  - 表格、文本、图片、视频
  - 文案、素材、代码、专家
- 多粒度搜索
  - 篇章级、段落级、语句级
- 跨媒体搜索
  - 不同媒体数据联合完成搜索任务

一切皆可搜索，搜索必达。

汉语语言知识图谱

- 汉语语言博大精深，百度有专门针对汉语语言的知识图谱。比如，搜索“凹的笔顺”，知识图谱可以直接把笔顺反馈给用户。针对现在大多数用户使用拼音或语音输入时，一些字不会念，无法输入拼音，百度会把汉字拆解，用语言描述它。比如，如果不知道“怼”字怎么念，就可以这样提问：“上面是‘对’下面是‘心’怎么念”。搜索结果页除了满足用户的主需求外，还有很多与这个汉字有关的扩展内容，如拼音、释义、组词、近反义词、成语、成语故事等。

知识图谱应用：新一代搜索引擎

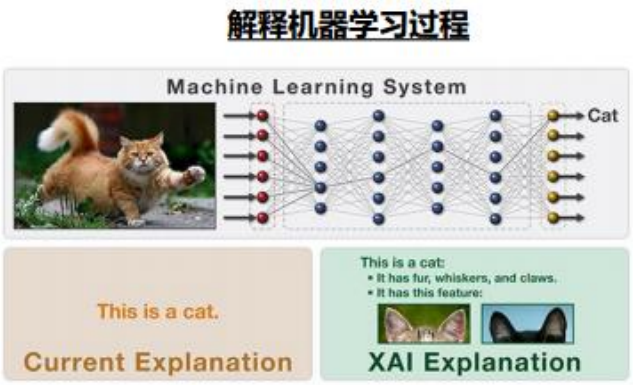
- **Baidu & Google：**对搜索引擎提出问题，它们会通过知识图谱给出很多知识性问题的答案。

在智能问答领域，借助知识图谱问答技术，图谱问答搜索产品能够直接回答搜索查询(query)中有客观答案的问题，如刘德华的年龄、上善若水的拼音、薰衣草的花期等。答案是一个有共同特征的集合问题，如一部电视剧的演员表、形容春天的成语、清朝皇帝列表、李白的诗等。

智能推荐

- 用户在发起一次搜索时，满足当前的需求后，仍有可能存在尚未满足的潜在需求。利用知识图谱中实体之间的丰富联系，我们可以给出优质的推荐，激发用户潜在的需求。根据查询的相关性、用户点击推荐内容的可能性等多方面，对推荐内容综合动态排序，向用户展示。
- 比如，当用户搜索“杨幂”的时候，除了直接给出和杨幂有关的各个维度的人物信息外，还能够给出可解释的推荐理由。
- 知识图谱还有生成摘要功能，即借助知识图谱技术提炼页面内容，在搜索结果中展示最关键的内容与服务，优化搜索体验。

智能解释

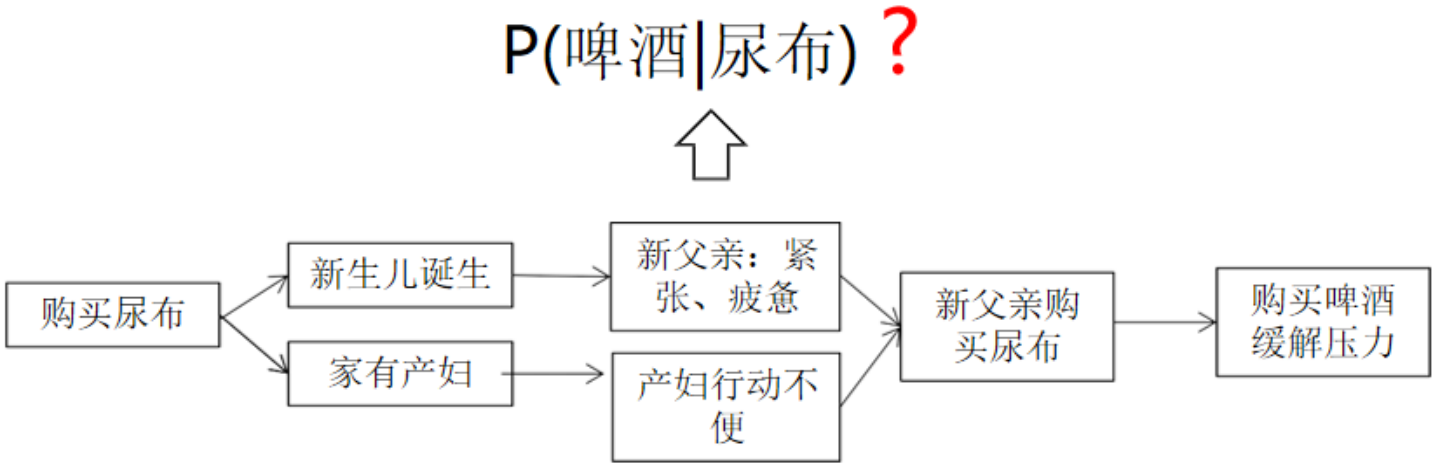


- 事实解释
- 关系解释
- 过程解释
- 结果解释

- 解释是智能的重要体现之一，将是人们对于智能系统的普遍期望。
- 可解释是智能系统决策结果被采信的前提。

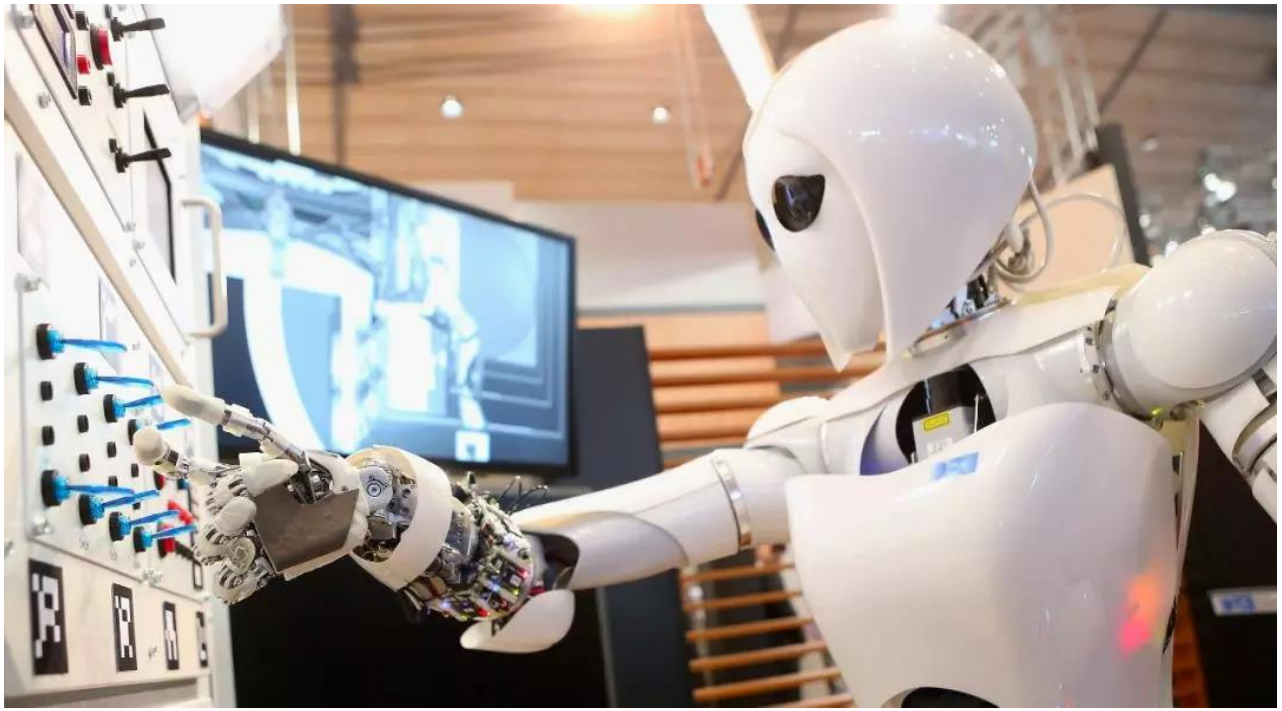
补全缺失的因果链条

- 万事万物都处在一个复杂的因果网络中。
- 当前的大数据多是业务结果数据，缺乏产生这些数据背景因果。
- 知其然，不知其所以然。



自然人机交互

- 人机交互方式将更加自然，对话式交互取代关键词搜索成为主流交互方式。
- 一切皆可问答： 图片问答、新闻问答、百科问答



深层关系发现与推理

Why baoqiang select Qizhun Zhang as his lawyer?
Why A invests B?

- 隐式关系发现、深层关系推理将成为智能的主要体现之一。

现有知识图谱资源

- **依靠人工构建的知识资源**
  - 早期知识资源建立是通过人工添加和群体智能合作编辑得到，如英文WordNet和Cyc项目，以及中文的HowNet。Cyc是一个通用的世界知识库，始建于1984年，其目的是将上百万条知识编码为机器可处理形式，并在此基础上实现知识推理等人工智能相关任务。Cyc包含了50万实体，接近3万个关系以及5百万事实。

- **基于群体智能的知识图谱**
  - 维基百科是至今利用群体智能建立的互联网上最大的知识资源，因此出现了很多使用维基百科构建知识库的项目，如DBpedia、YAGO和Freebase等。DBpedia以构建本体的形式对知识条目进行组织。YAGO融合WordNet良好的概念层次结构和维基百科中的大量实体数据。Freebase是基于维基百科，使用群体智能方法建立的包含6800万实体的结构化数据的知识图谱。清华大学和上海交通大学通过利用互动百科、百度百科，建立大规模知识图谱XLORE和Zhishi.me。XLORE以英文维基百科为桥梁，通过跨语言链接技术，建立融合了四大中英文百科数据的跨语言知识库。

现有知识图谱资源

- **基于互联网上链接数据构建的知识资源**
  - 国际万维网组织W3C于2007年发起的开放互联数据项目（Linked Open Data，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。Sean Bechhofer 等人在科学领域自建了一个近似于Linked Data的语义数据资源，该资源包含更准确的学术用语，并能很好地反映研究者的影响力。

- **基于机器学习和信息抽取构建的知识图谱**
  - 从互联网数据自动获取知识是建立可持续发展知识图谱的发展趋势。这类知识图谱构建的特点是面向互联网的大规模、开放、异构环境，利用机器学习和信息抽取技术自动获取Web上的信息构建知识库。如华盛顿大学图灵中心的KnowItAll和TextRunner项目、卡内基梅隆大学的“永不停歇的语言学习者”（Never-Ending Language Learner，NELL）项目都是这种类型的知识库。

现有知识图谱资源

- **维基百科（Wikipedia）**由维基媒体基金会负责运营的一个自由内容、自由编辑的多语言知识库。
- **DBpedia**由2007年德国柏林自由大学及莱比锡大学的研究者从维基百科里萃取结构化知识的项目开始建立。
- **YAGO**由德国马克斯-普朗克研究所（MPI）构建的大型多语言的语义知识库，是从10个维基百科以不同语言提取事实和事实的组合。
- **XLORE**是清华大学构建的基于中、英文维基和百度百科的开放知识平台，是第一个中英文知识规模较为平衡的大规模中英文知识图谱。