

# Web信息处理与应用

## 第九节 实体识别

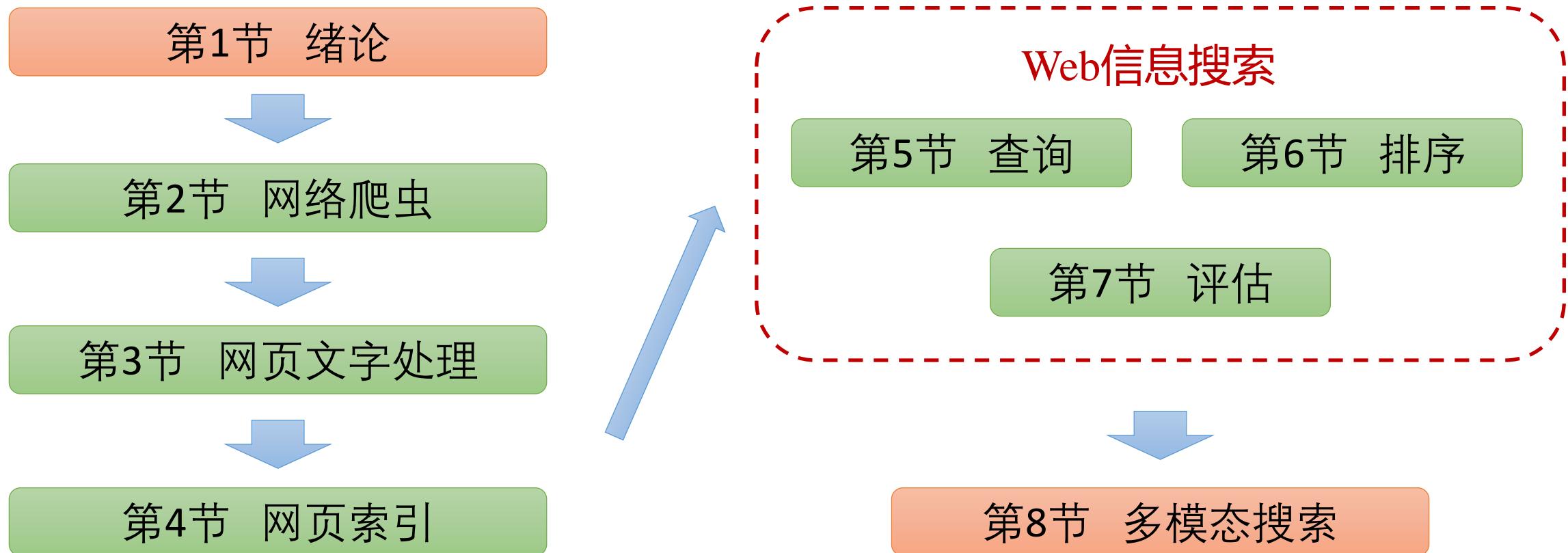
徐童 2021.11.8

- **传统的信息检索是如何实现的**

- 信息检索模型是用来描述文档与查询的表示形式与相关性的框架
  - 信息检索的实质是对文档基于相关性进行排序
  - 好的信息检索模型，可以在理解用户的基础之上，产生近似用户决策的结果，从而在顶部返回最相关的信息
- 信息检索模型的形式化表述：  $[D, Q, F, R(D_i, q)]$ 
  - D：文档表达（可视作索引词项的集合）
  - Q：查询表达
  - F：查询与文档间的匹配框架
  - R：查询与文档间的相关性度量函数（ $D_i$ 与 $q$ 分别表示特定文档与查询）

- 围绕“**检索**”、“**抽取**”与“**挖掘**”三条主线

## 第一部分：Web信息处理与检索



- 然而，传统信息检索返回的是“文档的集合”，而非信息

Baidu 百度

EDG

百度一下

网页 资讯 贴贴吧 图片 视频 知道 文库 采购 地图 更多

百度为您找到相关结果约91,100,000个

搜索工具

EDG获得S11冠军 骑士归来，新王加冕！

S11赛程 战队排名 话题讨论 贴吧

TOP 10 英雄联盟S11

我说edg！你说\_\_!  
EDG最后的排面!  
EDG夺冠  
我们的EDG

- 如今，用户对“信息”的需求更为迫切



人们已不再满足于单纯呈现原始的文档,  
而需要更加精炼的知识表达与更加直观的需求解决。

## • 响应用户需求，搜索引擎的结果日益丰富

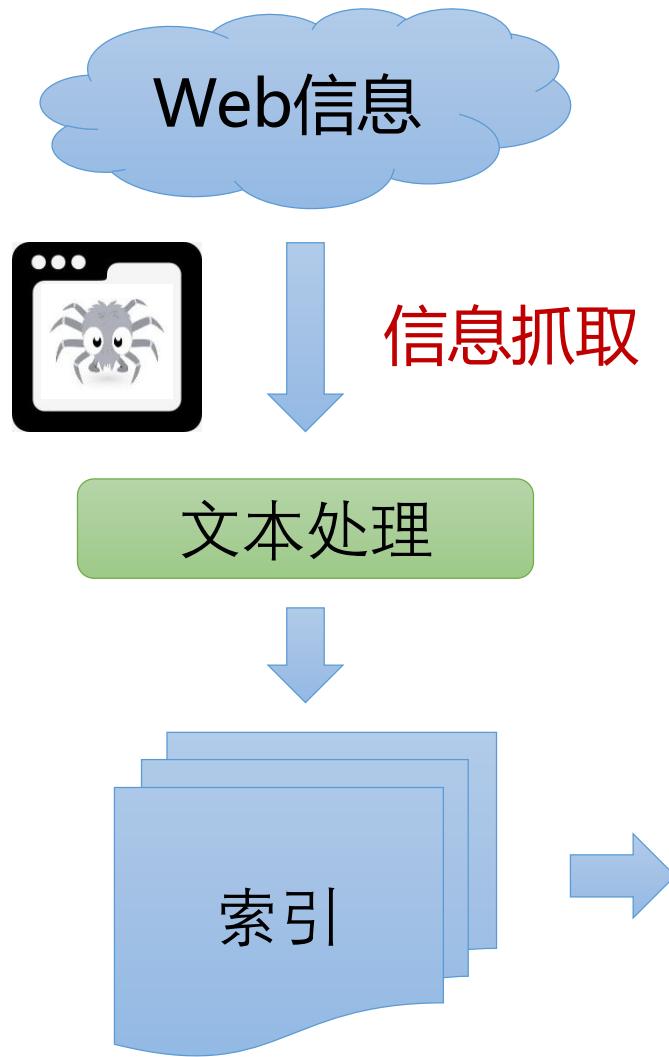
Baidu 百度 鸡腿的热量

百度一下 百度首页 消息 设置 ▾ 展开 ▾

<div style="border: 1px solid red; border-radius: 50%; width: 20px; height: 20px; position: absolute; bottom:

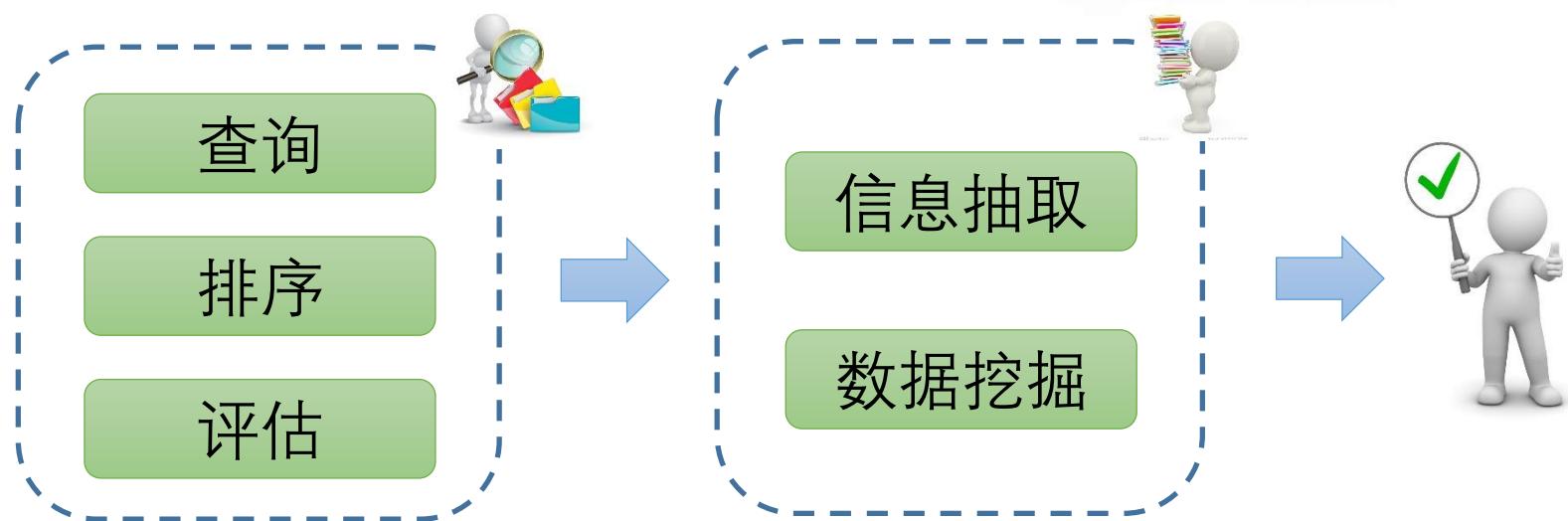
## 本节主题

- 本课程所要解决的问题



# 第八个问题：

## 如何从文档中提取信息和知识？



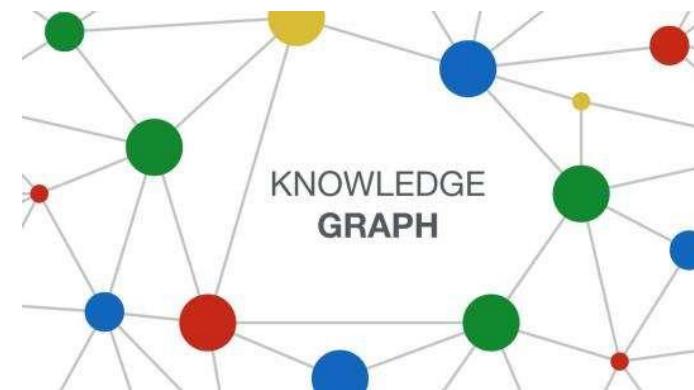
- 围绕“检索”、“抽取”与“挖掘”三条主线

## 第二部分：Web信息抽取与知识图谱

第9节 实体识别



第10节 关系抽取



- 信息抽取概述
- 知识图谱概述
- 命名实体识别

## • 信息抽取的含义

- 从语料中抽取指定的事件、事实等信息，形成结构化的数据
  - 被抽取的信息以预先定义的、结构化的形式描述。
  - 为后续的情报分析、自动文摘、问答系统等一系列应用提供服务。
- 从整体文档中抽取细粒度、结构化信息！



# 信息抽取

## • 信息抽取是整合与分析的基础



查公司 查老板 查关系  
北京百度网讯科技有限公司 天眼一下

VIP会员 ▾



↑ 企业关联

↓ 竞争分析

天眼查 国家中小企业发展基金旗下  
官方备案企业征信机构

查公司 查老板 查关系  
北京百度网讯科技有限公司 天眼一下

VIP会员 ▾

序号	产品名称	当前融资轮次	估值	成立日期	产品标签	所属地	简介
1	魅闪科技	A轮	-	2019-04-29	人工智能	广东	网红打造系统研发商
2	元戎启行	Pre-A轮	-	2019-02-18	人工智能	广东	自动驾驶运营服务提供商
3	觉非科技	A轮	-	2019-01-30	人工智能	北京	自动驾驶技术研发商
4	蝌蚪有读	-	-	2019-01-29	企业服务	上海	IT服务商
5	深蓝爱	-	-	2019-01-04	人工智能	北京	智能办公系统研发商

## • 信息抽取与信息检索

- 两者密切相关，却又存在鲜明差异
  - 功能不同
    - 检索：从文档集合中找文档子集
    - 抽取：从文本中获取用户感兴趣的事实信息
  - 处理技术不同
    - 检索：通常利用统计与关键词等技术
    - 抽取：借助于自然语言处理技术
  - 使用领域不同
    - 检索：通常领域无关
    - 抽取：通常领域相关（借助领域知识辅助抽取）



- **信息抽取的内容**

- 核心的8字方针： “**抽取实体，确定关系**”

- **实体**：即命名实体，指文本中的基本构成块，如人、机构等
- **属性**：实体的特征，如人的年龄、机构的类型等
- **关系**：实体之间存在的联系，也称事实，如公司和地址之间的位置关系、公司与人之间的雇佣关系
- **事性**：实体的行为或实体参与的活动

## • 信息抽取的基本任务

- MUC (Message Understanding Conference) 会议
  - 美国国防部研究计划署 (DARPA) 资助
    - 是否还记得这家神奇的机构? 1969年, 互联网雏形ARPANet由此诞生
  - 该会议主要测评信息抽取系统, 自87年共举行7次 (MUC-1...MUC-7)
- 在MUC-7上, 定义了5类基本的信息抽取任务
  - 命名实体NE、模板元素TE、共指关系CR、模板关系TR、背景模板ST

## • 信息抽取的基本任务

- 命名实体NE (实体抽取)
- 命名实体抽取是信息抽取最重要的任务
- 命名实体是文本中基本的信息元素，是正确理解文本的基础
  - 狹义：指现实世界中具体或抽象的实体，如人、组织、地点等
    - 如：英国混元太极拳协会/Org, 掌门 马保国/Person
  - 广义：还可以包含日期和时间、数量表达式等

- **信息抽取的基本任务**

- **模板元素TE** (属性抽取)

- 模板元素又称为实体的属性，目的在于更加清楚、完整地描述命名实体

- 通过槽 (Slots) 描述了命名实体的基本信息

- 槽：名称、类别、种类等

- 例如：马保国掌门指出，这两个年轻人不讲武德。

- ◆ TE：两个年轻人是不讲武德的 (属性)

- **信息抽取的基本任务**

- **共指关系CR**

- 如果不同的命名实体表达了相同的含义，即为共指关系，也称为等价概念
- 共指关系的抽取任务在于抽取关于共指表达的信息
  - 包括那些已在命名实体和模板元素任务中作了标记的，对于某个命名实体的所有表述
    - 例如：昨天，有两个年轻人……我说小伙子你不讲武德，他说马老师对不起我不懂规矩
  - ◆ CR：他和小伙子均代指“两个年轻人”（中的某一个）

- **信息抽取的基本任务**

- 模板关系TR (关系抽取)
  - 实体之间的各种关系，又称为事实
  - 通过关系抽取，将实体关联起来，并为推理奠定基础
    - 例如，职务 (Post\_of) 、雇佣关系 (Employee\_of) 、生产关系 (Product\_of) 等
      - 如：
        - Post\_of(掌门, 马保国)
        - Employee\_of(英国混元太极拳协会, 马保国)

## • 信息抽取的基本任务

### • 场景模板ST (事件抽取)

#### • 又称事件，是指实体发生的事件

- 例如：会议 (Time<...>, Spot<...>, Convener<...>, Topic<...>)

#### • 常见的新闻事件描述模板 5W1H

➤ Who 、 When 、 Where 、 What 、 Why 、 How

➤ 例如：昨天 (When) , 有两个年轻人 (Who) 因为不讲武德 (Why) , 用左正蹬、右鞭腿和左刺拳 (How) 偷袭 (What) 了马老师

- **信息抽取的基本任务**
- 一个简单的信息抽取实例：人民日报1998-01-07

**19980107-06-016-001**意大利总理普罗迪 4 日说，欧洲国家将采取行动，共同对付库尔德难民涌入问题。普罗迪 4 日晚召开了由意外长、内政和国防部长参加的紧急会议，商讨应付库尔德难民问题的对策。会前，普罗迪说，“在经过最初的混乱后，欧洲国家的行动已经大大加强”，今后几天内将在此问题上进行系统合作。

- **信息抽取的基本任务**

- NE实体抽取结果示例

```
<NamedEntities>
```

```
  <PersonList>
```

库尔德 (occurrence: 1/1/15; 1/2/19;) (类似于倒排表的形式，但表示在文中的位置)

普罗迪 (occurrence: 1/1/3; 1/2/0; 1/3/2;)

```
  </PersonList>
```

```
  <OrgList></OrgList>
```

```
</NamedEntities>
```

- TR关系抽取结果示例

```
<EntityRelations>
```

post\_of( 意大利总理, 普罗迪)

```
</EntityRelations>
```

- **信息抽取的基本任务**

- ST事件抽取结果示例

<EventTemplateInstances>

  <ConferenceInfo>

    <Time> 4 日晚 ( 1998-01 )</Time>

    <Spot> 意大利</Spot>

    <Converner> 普罗迪</Converner>

    <Title>由意外长、内政和国防部长  
    参加的紧急会议

  </Title>

  </ConferenceInfo>

</EventTemplateInstances>

会议时间 Time	4 日晚 (1998-01)	
会议地点 Spot	意大利	
召集人 Convener	姓名/团体名称 Name	普罗迪
	机构、职位 Org/Post	意大利总理
会议名 / 标题 Conf-Title	由意外长、内政和国防部长参加的紧急会议	

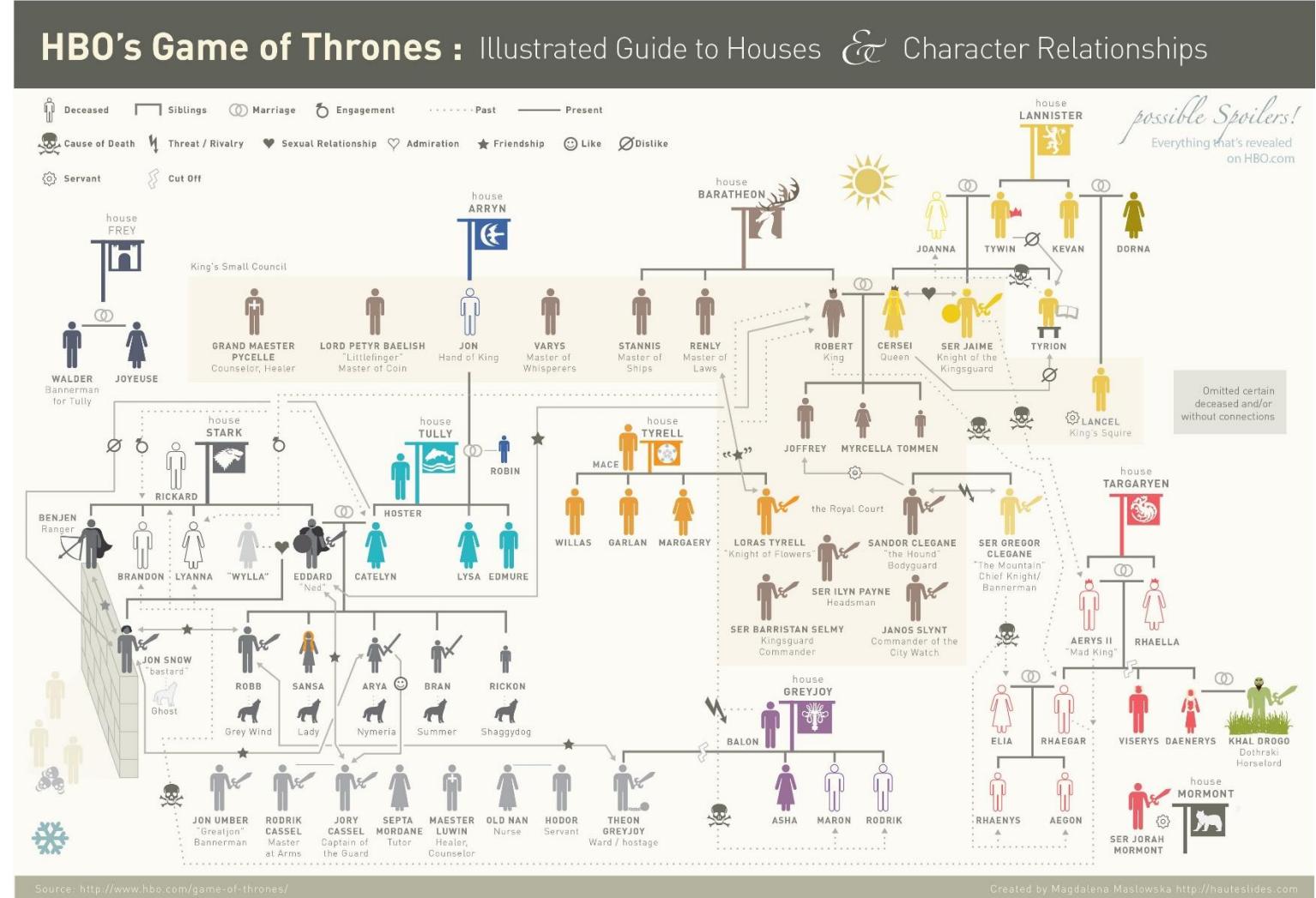
- 基于信息抽取的推理

- 通过对于实体、属性及其关系的描述，使得推理成为可能
  - 例如，如下的例子涉及实体（人物、组织）、属性/事件（在任时间）、关系（任职、前后任）等多种信息。

中国共产党职务		
前任: <a href="#">赵紫阳</a>	<a href="#">中国共产党中央委员会总书记</a> 1989年 - 2002年	继任: <a href="#">胡锦涛</a>
前任: <a href="#">邓小平</a>	<a href="#">中国共产党中央军事委员会主席</a> 1989年 - 2004年	
前任: <a href="#">芮杏文</a>	<a href="#">中国共产党上海市委员会书记</a> 1987年 - 1989年	继任: <a href="#">朱镕基</a>

- 信息抽取概述
- 知识图谱概述
- 命名实体识别

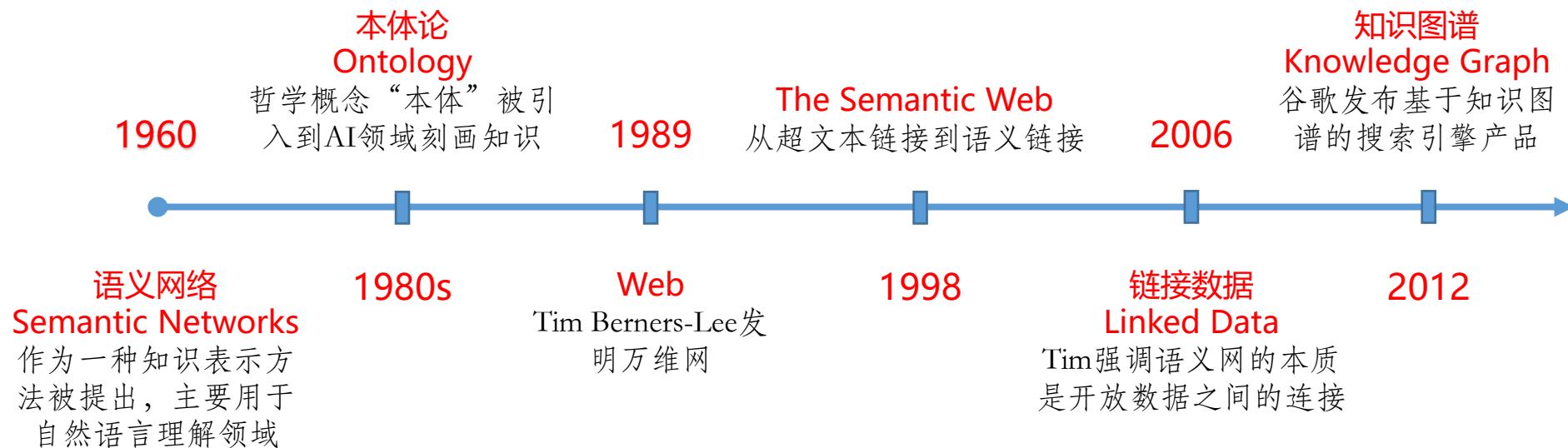
## • 知识的关联与知识本身同样重要



## • 从知识到知识关联

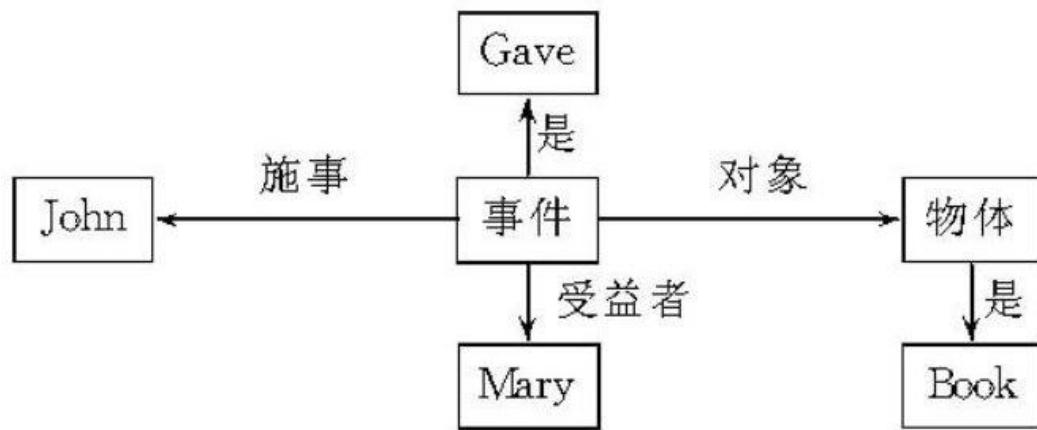
### • 知识图谱的发展历程：从语义网络与本体论衍生而来

- 语义网络的部分内容在第三节有所提及，例如：同义词/相关词，WordNet



## • 雏形理念：语义网络（Semantic Networks）

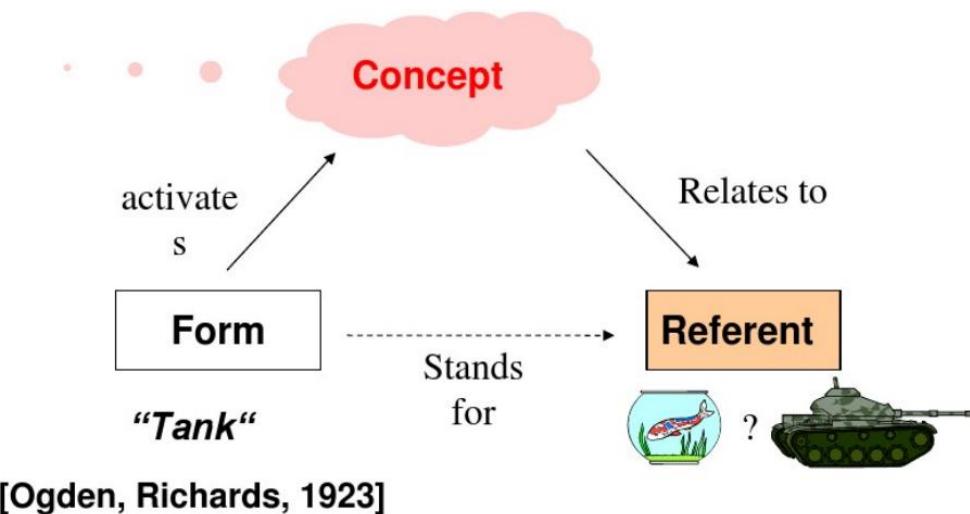
- 一种以有向图结构表达人类知识构造的形式
  - 部分内容在第三节课（“网页文字处理”）有所提及
    - 例如：同义词/相关词，上下位关系，WordNet等



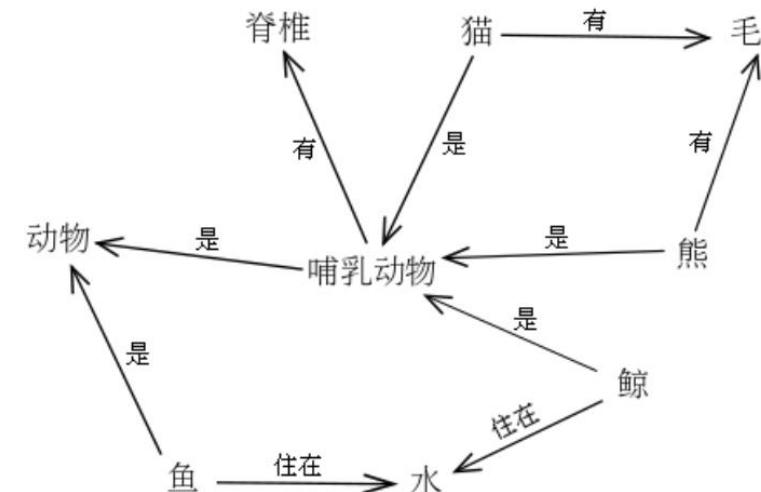
```
dog, domestic dog, Canis familiaris  
=> canine, canid  
=> carnivore  
=> placental, placental mammal, eutherian, eutherian mammal  
=> mammal  
=> vertebrate, craniate  
=> chordate  
=> animal, animate being, beast, brute, creature, fauna  
=> ...
```

## • 雏形理念：本体论（Ontology）

- 本体是指一种形式化的，对于共享概念体系的明确而又详细的说明
  - 指特定领域之中存在的对象类型或概念，及其属性和相互关系



符号三角形：符号、概念、事物三者关系



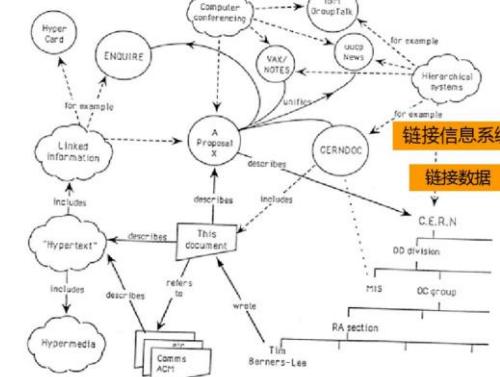
“哺乳动物”概念与相关事物形成的语义网络

- 雏形理念：本体论（Ontology）

- 本体论中的五元组表示法： $O = \{C, R, F, A, I\}$ 
  - C – Concept，概念集合，通常以术语形式组织，如“中科大学生”
  - R – Relations，描述概念或实例间语义关系，如“同班同学”
  - F – Functions，一组特殊关系，其中第N个元素由其他n-1个元素决定
    - 如，二手车的价格由品牌、车况、里程数等因素决定
  - A – Axioms，公理，例如A是B的子女，B是C的子女，得到A与C的关系
  - I – Instances，具体的个体，如“不是助教”是“中科大学生”的实例

- 雉形理念：语义网（Semantic Web）

- 1998年，由Tim Berners-Lee（又是他！）提出
    - 核心思想：通过给万维网上的文档（如HTML、XML文档）添加能够被计算机所理解的语义“元数据”（Meta data），从而使整个互联网成为一个通用的信息交换媒介。
    - 目前，万维网在组织信息资源时主要面向“人”的信息发布和获取，侧重于信息的显示格式和样式（如HTML）。
    - 而语义网必须侧重于信息的语义内容，并考虑计算机对文本内容的“理解”以及它们之间的相互交流和沟通。



## • 知识图谱的诞生

- 2012年5月16日，Google知识图谱（Knowledge Graph, KG）正式发布
- 除了显示网页文档的连接列表外，还提供结构化的、详细的有关主题的信息。
- 其目标是，用户可以使用此功能直接解决查询的问题，而不必导航到其他网站并自行汇总信息。



## • 知识图谱的优点

- 知识图谱至少可以从以下三个层面提升搜索的效果：
  - 找到最想要的信息：不再需要用户自行浏览、阅读和总结，而将信息直接呈现
  - 提供最全面的摘要：对搜索对象进行总结，使得用户获得更完整的信息和关联
  - 让搜索更有深度和广度：构建完整知识体系，使用户获得意想不到的新发现



## • 日益丰富的知识图谱



250概念  
4M实例  
6000属性  
500M三元组  
在线更新



350K概念  
10M实例  
100属性  
120M三元组



NELL



OpenIE  
(Reverb, OLLIE)

850K概念  
8M实例  
70K属性



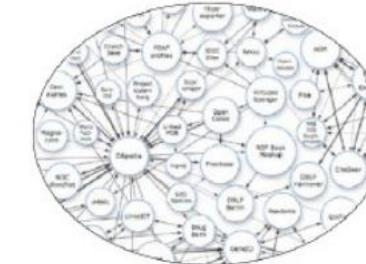
Google KG



15K概念  
40M实例  
4000属性  
1B三元组  
Google KB核心



50M义项  
50+种语言  
262M三元组

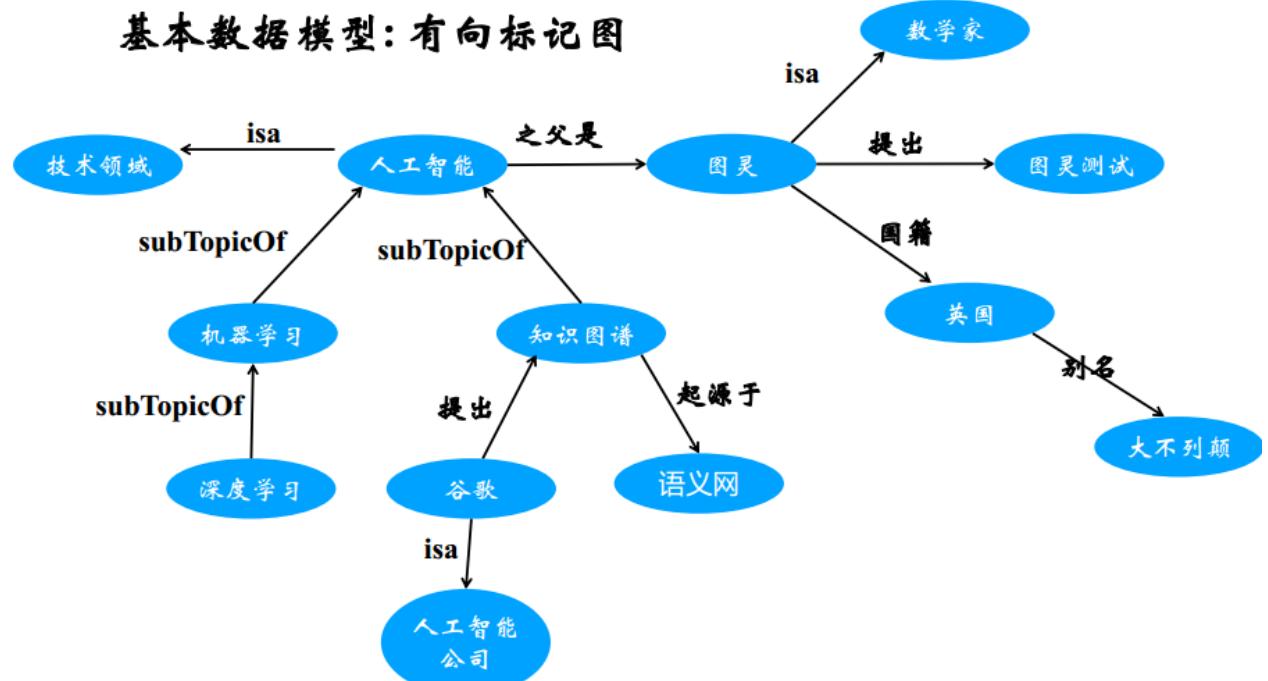


WordNet  
7种欧洲语言  
跨语言链接



## • 知识图谱的基本形式

- 由前述各种网络衍生而来，知识图谱呈现出类似的基本形式
- 由结点和结点之间的边组成，  
结点表示概念（或实体），  
边表示关系（或属性）。
- 在数学上，知识图谱表现为  
一个有向图。



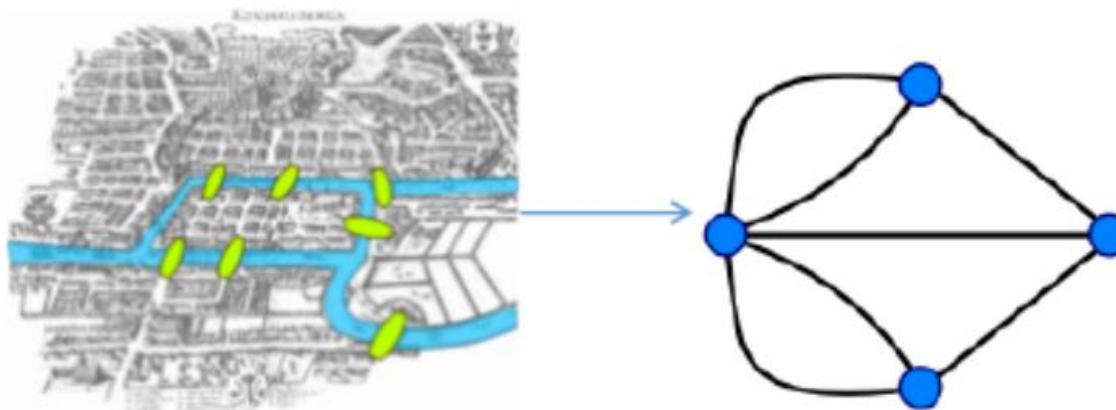
## • 知识图谱的基本元素：节点

- 一般而言，知识图谱中的节点用来表示概念（Concept）和实体。
- 实体（Entity / Object / Instance）
  - 能够独立存在的，作为一切属性的基础和万物本原的东西 —— 黑格尔



## • 知识图谱的基本元素：边

- 一般而言，知识图谱中的节点用来表示关系（Relation）和属性（Attribute）。
  - 关系：侧重实体（Entity）之间的关联，例如“高王”：姚明高王小四
  - 属性：用于描述实体的特征，例如尺寸，颜色、组成等等
- 点和边组成知识图谱的基本单位：三元组（实体-关系-实体）



## • 知识图谱相关应用（1）：语义搜索

- 通过建立事物之间的联系，实现更准确、更直接、更完整的搜索



Freebase

社区协同构建

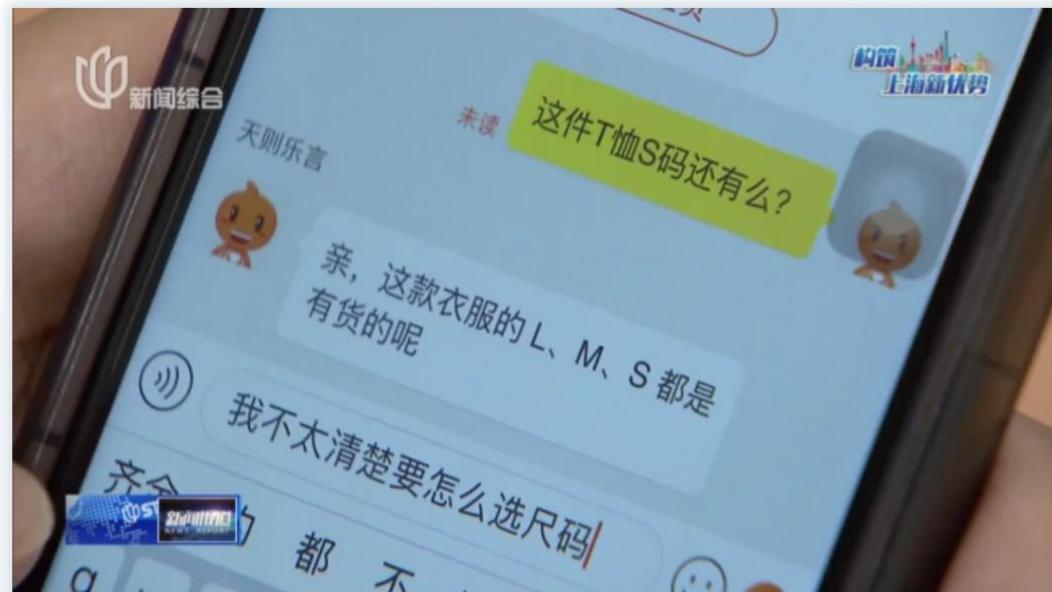


维基众包

schema.org  
....the new SEO?

网页嵌入语义数据

- **知识图谱相关应用（2）：问答系统**
- 在理解用户意图的基础上，将知识图谱作为大脑，基于推理能力提升交互体验



## • 知识图谱相关应用（3）：推荐系统

- 利用知识图谱提高推荐系统的推荐多样性和可解释性，提升推荐性能



## • 知识图谱工程技术路线

- 高关联、规范化知识图谱极大提升了数据质量，在各个垂直领域都发挥着重要作用，被各大互联网大厂视为人工智能“新基建”。



## • 公开的知识图谱数据

- 目前，已有多个通用或专业知识图谱开放，典型的中文KG知识图谱有：
  - CN-Dbpedia：由复旦大学知识工场实验室研发并维护的大规模通用领域结构化百科，包含1700万个实体，2亿个三元组
  - Zhishi.me：通过从开放的百科数据中抽取结构化数据，包含1000万个实体，1.2亿个三元组
  - XLORE：由清华大学开放，从异构的跨语言在线百科中抽取结构化信息，包含1600万个实体，44万个属性
- 通过OpenKG (<http://www.openkg.cn>)，可以找到丰富的开放图谱及工具。



- **相关会议与测评**

- 全国知识图谱与语义计算大会 (CCKS)
  - 已成为国内知识图谱、语义技术、链接数据等领域的核心会议

## 2020全国知识图谱与语义计算大会

*China Conference on Knowledge Graph and Semantic Computing – 南昌 – 2020年11月*

- CCKS每年组织有关知识图谱与语义计算相关的测评，2020年的测评包括：
    - 新冠知识图谱构建与问答
    - 面向中文短文本的实体链指
    - 面向中文电子病历的医疗实体及事件抽取
- .....

## • 更丰富的知识图谱类型

### • 事理图谱

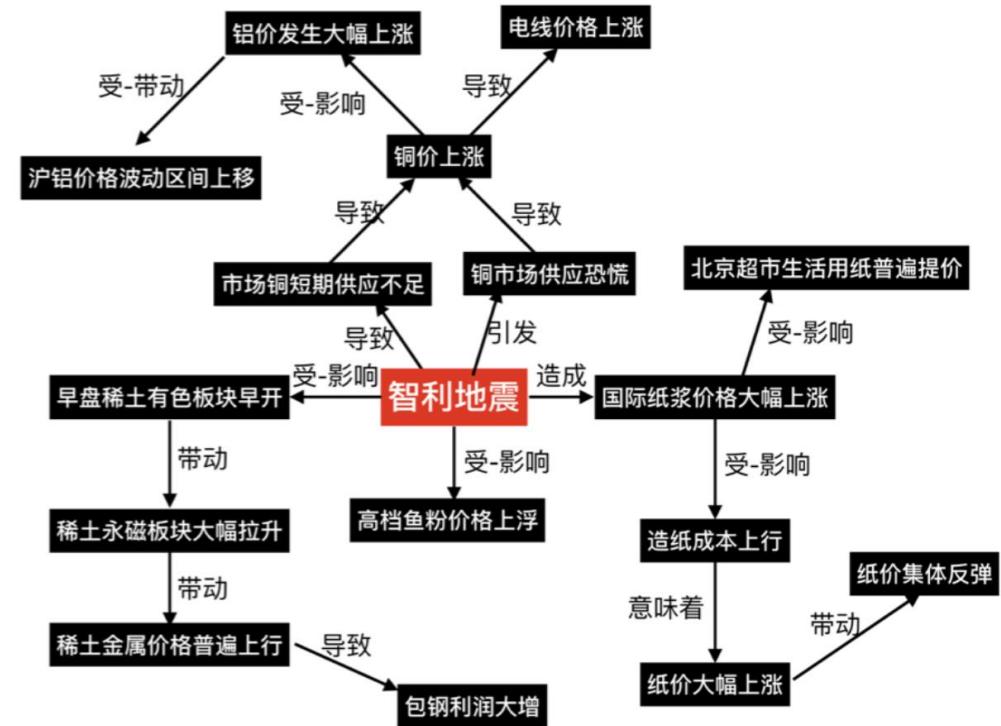
- 人类主要是以“事件”为单位进行记忆和理解现实世界的，传统本体所使用的概念模型难以反映事件这一更高层次和更复杂的语义信息。

	知识图谱	事理图谱
描述知识	万物本体	逻辑社会
研究对象	名词性实体及其属性、关系	谓词性事件及其内外（空间、时间域）联系
构建目标	万物互联	全逻辑库，逻辑演化模型
回答问题	When、Who、What、Where	Why、How
组织形式	有向图	有向图
知识形式	<实体，属性，属性值>,<实体，关系，实体> 三元组	<事件，论元集合，逻辑关系> 多元组
知识确定	事实是确定的	逻辑不确定，有转移概率
知识状态	相对静态，变化缓慢	动态的
知识敏感	精确性要求极高，实时性要求极高	可一定容错，参考逻辑
构建难点	知识本体的搭建、知识抽取与融合	事件的表示、事件的抽取；与知识图谱的融合

## • 更丰富的知识图谱类型

### • 事理图谱

- 事理图谱所要描绘的是一个逻辑社会，研究对象是谓词性事件及其内外联系。
- 事理图谱与知识图谱的组织形式相仿，实体通过头尾相连，形成有向图的组织性质
- 借助图谱中的事理逻辑链接，可以形成对于事件的推理

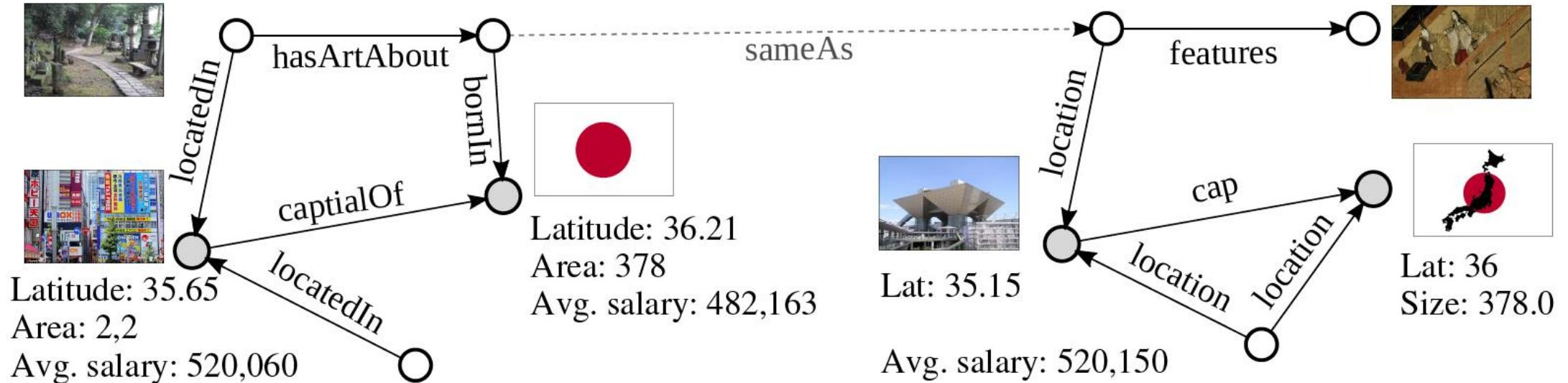


智利地震事件因果图谱（部分）

- 更丰富的知识图谱类型

- 多模态知识图谱

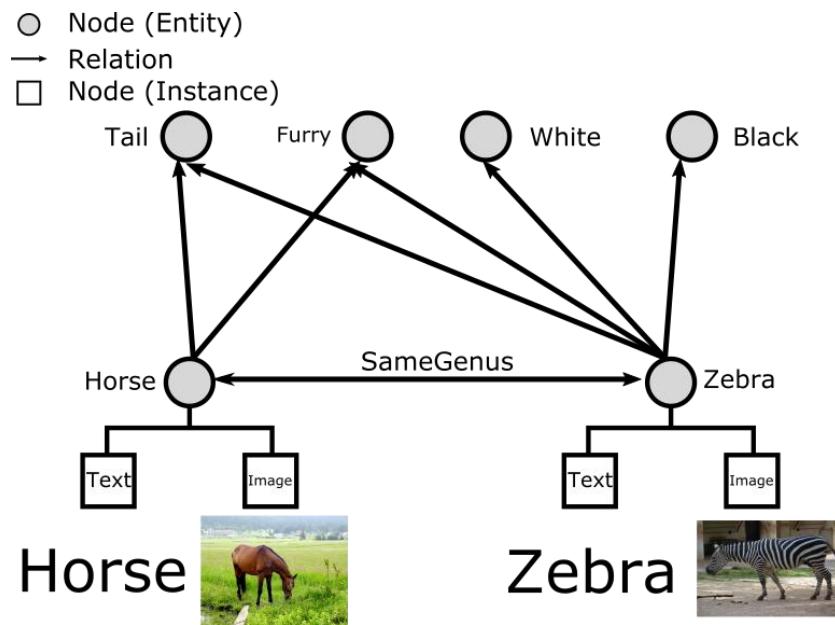
- 现实世界语义模态日益丰富，有效表示与整合多模态知识成为趋势。



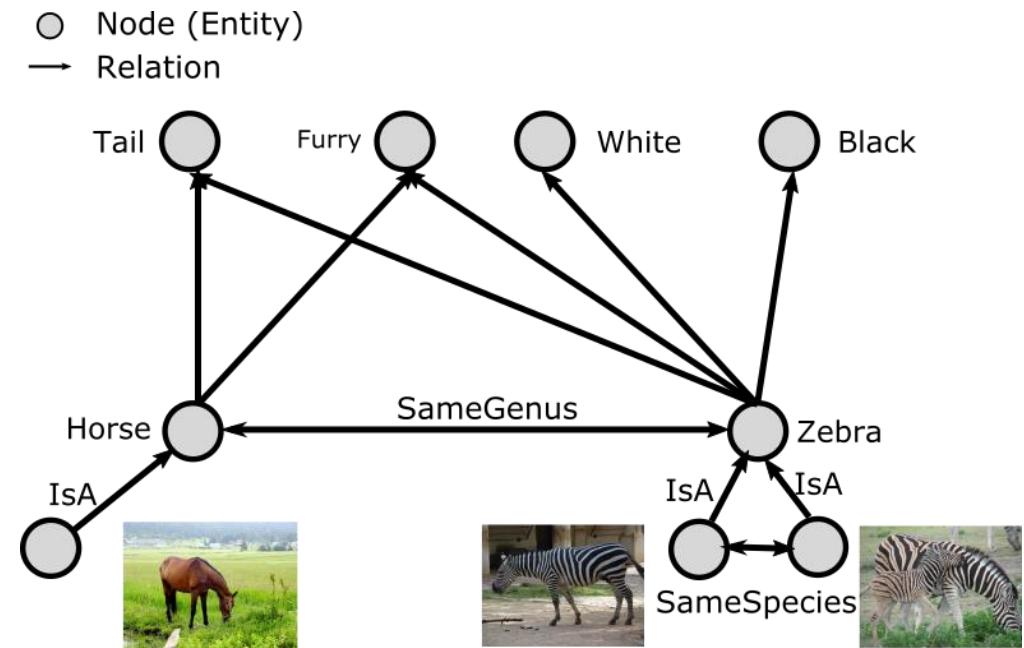
- 更丰富的知识图谱类型

- 多模态知识图谱

- 多模态知识图谱可笼统分为**属性多模态**与**实体多模态**两大类。



属性多模态知识图谱



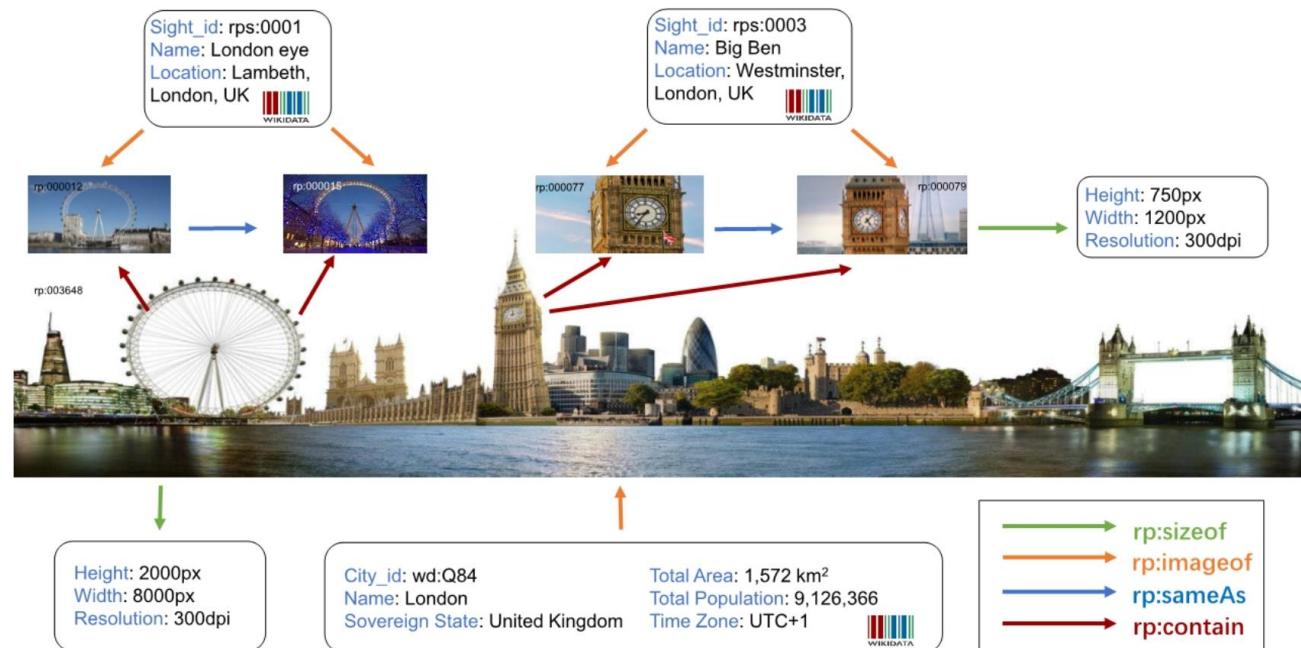
实体多模态知识图谱

- 更丰富的知识图谱类型

- 多模态知识图谱

- 开放多模态知识图谱的典型代表：Richpedia

- 由东南大学认知智能研究所漆桂林老师团队提出，摆脱传统知识图谱中实体局限于文本的束缚

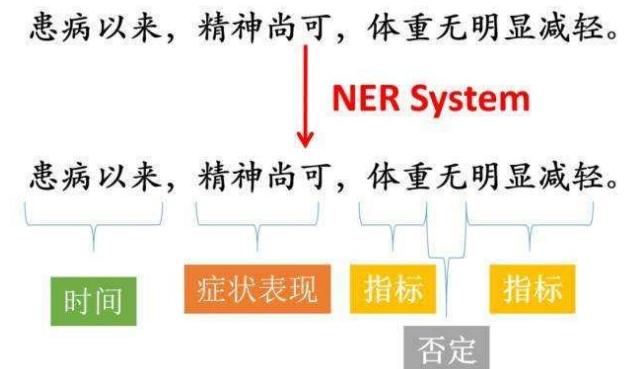


- 信息抽取概述
- 知识图谱概述
- 命名实体识别

- 命名实体识别的基本概念

- 命名实体识别 (Named Entity Recognition, NER)

- 识别出文本中的人名、地名等专有名词，和有意义的时间、日期等数量短语等，并加以归类。
- 命名实体识别是信息抽取中的核心任务，它往往包含两个子任务
  - 判别实体边界
  - 判别实体类型



## • 命名实体识别的内容

- 一般按照MUC-7的定义，分为3大类，7小类

- 实体类：人名、地名、机构名
- 时间类：日期、时间
- 数值类：货币、百分比

- 哪些不是命名实体？（部分例子）

- 重复指代的普通名词：如 飞机、公司 等

- 人的团体名称以及以人命名的法律、奖项等：如 共和国、诺贝尔奖 等

- 非时间、日期、货币、百分比的数字

ACE (Automatic Content Extraction) 定义中的 **NER** 任务：  
人名 (**Person**)、机构名 (**Organization**)、地名 (**Location**)、设备名 (**Facility**)、武器名 (**Weapon**)、交通工具名 (**Vehicle**) 和地理政治实体 (**Geo-Political Entity**)

- 命名实体识别的难点

- 与分词的难点非常相似
  - 不断有新的命名实体涌现，如新的人名、地名、组织名等
  - 命名实体存在严重歧义
    - 如Washington（地名/人名），May（人名/月份）
  - 命名实体构成结构复杂，如别名、缩略词、音译等
    - 如USTC与Univ. Sci. Tech. of China
  - 命名实体类型多样：如John Smith, Mr Smith, John, 实际上是CR关系

- 命名实体识别的性能评价

- 与检索任务大致相同，采用Precision / Recall / F-value加以衡量
- 正确率与召回率的计算方式
  - 方案1：分子为返回的正确答案数量
  - 方案2：分子为返回的正确答案数量 +  $\frac{1}{2}$ 的部分正确答案数量
    - 部分正确的案例：“Severus/Person Snape”（类型正确，边界错误）

- 命名实体识别方法（1）：基于词典

- 基于词典的识别方法（List Lookup）

- 经常作为NER问题的基准算法（baseline）
- 预先构建一个命名实体词典，出现在词典中的词汇即识别为命名实体
- 词典的来源：来自于领域公开数据，例如
  - 人名/组织名，可以来自于黄页、电话簿、公开名单等
  - 地点，可利用一些现有的地理信息列表



- 命名实体识别方法（1）：基于词典

- 基于词典的识别方法（List Lookup）

- 优点：方法简单快速，与具体语境无关，容易部署和更新（只需更新词典）

- 缺点（与基于词典/匹配分词存在类似问题）：

- 大部分情况下很难枚举所有的命名实体名

- 构建和维护词典的代价较大

- 难以有效处理实体歧义



## • 命名实体识别方法（2）：基于规则

- 采用手工构造规则模板，对符合规则的实体进行识别
  - 选用特征包括统计信息、标点符号、关键字、指示词和方向词、位置词（如尾字）、中心词等
  - 以模式和字符串相匹配为主要手段
  - 多数参加MUC-7（1997）会议评测的系统，都采用了此方法
  - 例如：[ORGANIZATION]'s headquarter in [LOCATION]
    - We visited Microsoft /org's headquarter in Seattle /loc.

## • 命名实体识别方法（2）：基于规则

### • 基于手工规则的识别方法

- 优点：当提取的规则能较精确地反映语言现象时，性能较好

- 缺点：

- 规则往往依赖于具体语言、领域和文本风格

- 不同领域的句法往往差异极大，如学术圈与二次元

- 代价太大，系统建设周期长、移植性差而且需要建立不同领域知识库

## • 命名实体识别方法（3）：基于统计

- 基于统计的命名实体识别方法是当下的主流方法

类型	采用的模型或方法	代表工作
有监督的学习方法	隐马尔可夫模型或语言模型	Liu <i>et al.</i> (2005); Zhang <i>et al.</i> (2003a); Sun <i>et al.</i> (2002); Zhou and Su(2002); Bikel <i>et al.</i> (1997)
	最大熵模型	Tsai <i>et al.</i> (2004); Borthwick(1999); Mikheev <i>et al.</i> (1998)
	支持向量机	Yi <i>et al.</i> (2004); Asahara and Matsumoto (2003)
	条件随机场	Leaman and Gonzalez (2008); Finkel <i>et al.</i> (2005); McCallum and Li (2003)
	决策树	Isozaki(2001); Palouras <i>et al.</i> (2000); Sekine <i>et al.</i> (1998)
半监督的学习方法 (弱监督学习方法)	利用标注的小数据集(种子数据)自举学习	Singh <i>et al.</i> (2010); Nadeau(2007); Niu <i>et al.</i> (2003); Collins (2002b); Collins and Singer (1999)
无监督的学习方法	利用词汇资源(如 WordNet)等进行上下文聚类	Etzioni <i>et al.</i> (2005); Shinyama and Sekine (2004)
混合方法	几种模型相结合或利用统计方法和人工总结的知识库	Liu <i>et al.</i> (2011b); Finkel and Manning(2009); Zhou(2006); Wu <i>et al.</i> (2003, 2005); Jansche and Abney(2002)

- **回顾：分词时的序列标注问题**

- 基于统计模型的分词方法，进一步抽象而言，可以得到一个序列标注问题
  - 四类标注：B（词的开始）、M（词的中间）、E（词的结束）、S（单字词）
  - 例子：中国科学技术大学是中国最好的大学
    - 标注：BMMMMMMME S BE BME BE
    - 分词结果：中国科学技术大学 / 是 / 中国 / 最好的 / 大学
  - 类似的序列标注，在命名实体识别问题中也得到广泛应用。

## • 命名实体识别方法（3）：基于统计

- 分支一：基于分类的命名实体识别方法
- 将NER视作一个多分类问题，通过设计特征训练分类器的方法加以解决。
- 例如： Hideki Isozaki, et al., Efficient Support Vector Classifiers for Named Entity Recognition, COLING 2002
  - 一共33个标签：8种实体，每种对应Begin, Middle, End, Single四种类型，加上Other（即不属于任何一类实体），得到 $8 \times 4 + 1 = 33$ 类标签。
  - 选取15维特征：当前词及前后各两个词（共计5个词）
    - 每个词3维特征：词性、字符类型、单词

## • 命名实体识别方法（3）：基于统计

- 通过以上方式，得到一个高维稀疏的向量，仅15维为1，其余均为0
- 特征实例：对于 “President George Herbert Bush said Clinton is ...” 中 “Bush” 这个词

```
x[1] = 0      // Current word is not 'Alice'  
x[2] = 1      // Current word is 'Bush'  
x[3] = 0      // Current word is not 'Charlie'  
           :  
x[15029] = 1 // Current POS is a proper noun  
x[15030] = 0 // Current POS is not a verb  
           :  
x[39181] = 0 // Previous word is not 'Henry'  
x[39182] = 1 // Previous word is 'Herbert'  
           .
```

- 基于该向量，通过支持向量机（SVM）模型+Sigmoid函数属于何种标签

## • 命名实体识别方法（3）：基于统计

- 事实上，早期基于统计的方法，需要精心设计大量的相关特征
  - 以机构名识别为例，常见的内部特征包括单词特征、核心词特征、词性特征、语义特征等

标注	类型	示例
F	机构特征词	北京搜狐畅游时代网络技术有限公司
R	机构名中的人名	法国马蒂尼埃集团
NR	其它人名	俞昊然创立了“计蒜客”
S	机构名中的地名	北京市文化局相关领导表示
NS	其它地名	在前不久的中国游戏行业年会上
O	常见机构名	中国人民银行
E	机构名中的其它词	侵犯腾讯公司相关游戏著作权一案
L	机构名之间的连接词	中国移动和中国联通慢慢掌控了很多版权
P	职位名	友达董事长李焜耀
Z	其它词	
.....	.....	.....

## • 命名实体识别方法（3）：基于统计

- 事实上，早期基于统计的方法，需要精心设计大量的相关特征
  - 以机构名识别为例，常见的内部特征包括单词特征、核心词特征、词性特征、语义特征等

标注	类型	示例
M	修饰词	国内知名厂商长虹
C	中心词	华为市场份额
W	谓语动词	诺基亚终于发布了其第一款TD产品
N	主谓之间的词	诺基亚终于发布了其第一款TD产品
K	谓宾之间的词	中国联通联合了中国电信
J	介词	在央视广告招标中
B	机构名上文的前一个词	瑞典正是爱立信的总部所在地。
A	机构名下文的后一个词	北京市文化局相关领导表示
.....	.....	.....

## • 命名实体识别方法（3）：基于统计

- 与统计特征相关的一个问题：词性标注问题
  - 词性（part-of-speech）是词汇基本的语法属性，通常也称为词类。
  - 词性标注就是在给定句子中判定每个词的语法范畴，确定其词性并加以标注的过程。词性标注是自然语言处理中一项非常重要的基础性工作。
- 词性标注问题，尤其是中文词性标注问题，也面临一些困难和挑战：
  - 汉语是一种缺乏词形态变化的语言，无法从单词形态上来判别
  - 常用词兼类现象严重，例如：科学技术（名词） / 这不科学（形容词）

## • 命名实体识别方法（3）：基于统计

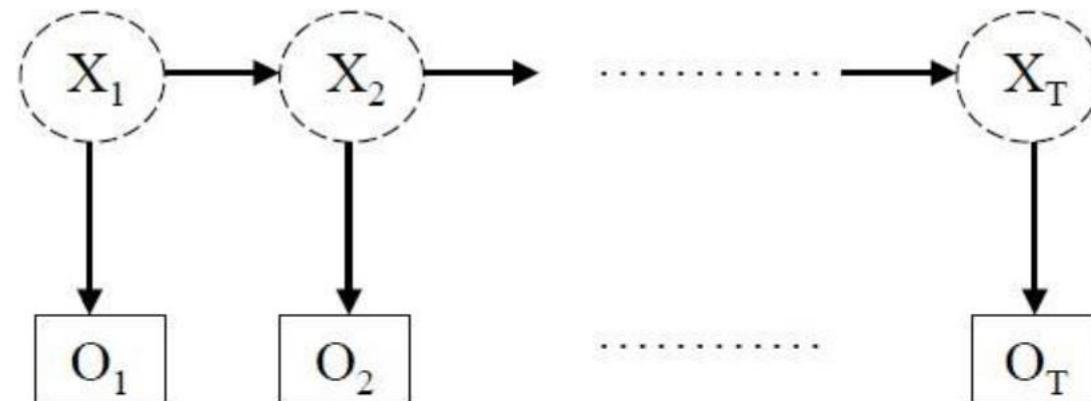
### • 基本的词性标注方法

- 从思路上说，词性标注方法与实体识别，乃至分词，总体思路都是类似的
- 基于规则的方法：人工或通过大规模语料学习规则
  - 核心思想是按兼类词搭配关系和上下文语境建造词类消歧规则
- 基于统计模型的标注方法：HMM等面向词序列的方法
- 统计方法与规则方法相结合的词性标注方法
  - 先基于规则排除明显歧义，再基于统计模型标注，最后人工校验

➤ 可参考统计自然语言处理（第7.5节），宗成庆著，北京大学出版社

## • 命名实体识别方法（3）：基于统计

- 分支二：基于序列模型的命名实体识别方法
- 与分词中的序列标注方法思路类似，区别在于标注的不同
  - 针对命名实体的类别不同，引入了更多、更细致的标签种类
  - 常用模型亦采用HMM、CRF以及各种序列深度学习方法（如LSTM）等



## • 命名实体识别方法（3）：基于统计

- 基于统计的命名实体识别方法具有以下特点：
  - 对特征选取的要求较高，需要从文本中选择对NER有影响的特征来构建特征向量（尤其是早期工作，深度学习技术发展后相对要求降低）
  - 通常做法是对训练语料所包含的语言信息进行统计和分析，从中挖掘出特征
  - 对语料的依赖也较大，目前缺少通用的大规模语料
    - 对深度学习技术影响尤甚，特定专业领域影响最为明显
    - 大部分技术仍需要进行人工标注训练数据

## • 命名实体识别方法（3）：基于统计

- 除了模型的改进之外，研究者们也在尝试引入更多信息和领域知识以提升效果
  - 例如，英文的词根词缀信息，或者汉字的部首信息，都蕴含着丰富语义
  - 以医疗实体识别为例，人们发现，中文五行等特征部首常出现在医疗实体中
    - 不仅是实体开始的提示符，不同类型的部首，还往往对应着不同类型的医疗实体

• 金：人体内微量元素

• 木：中成药物

• 水：人体体液与体液症状（溶、溃…）

.....

• 月：身体部位（腿、胃、肾…）

• 口：头部器官

• 疖：疾病名称

.....

## • 命名实体识别方法（3）：基于统计

### • 如何利用部首信息进行建模？三点观察

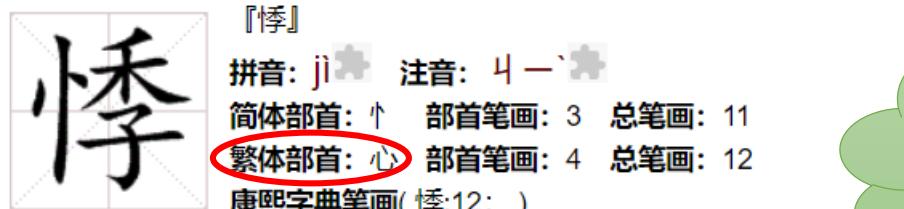
- 简体中文汉字往往都是由繁体中文汉字简化演变而来，因而简体部首可能不具有很强的解释性。
- 存在不同部首表达同一种释义的情况：“阝”与“耳”、“忄”与“心”、“火”与“灬”等。增大了训练代价，也降低了部首信息的表达能力。
- 释义不同的字符有着相同的部首：“朝”（释义为早晨）与“脚（释义为足）”的部首都是“月”。同样影响了部首的表达能力。

惭 患

朝 腿

## • 命名实体识别方法（3）：基于统计

- 基于前述观察，通过采用繁体部首的方式，还原语义信息



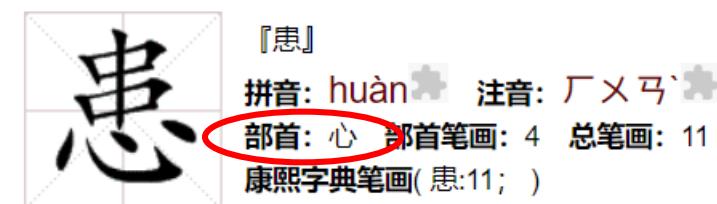
五笔86: NTBG 五笔98: NTBG 仓颉: PHDD  
 四角号码: 92047 UniCode: U+60B8 规范汉字编号: 4988



五笔86: EVEP 五笔98: EVPY 仓颉: BYAV  
 四角号码: 77233 UniCode: U+817F 规范汉字编号: 2976

释义相同，在繁体部首上统一

释义不同，在繁体部首上区分开



五笔86: KKHN 五笔98: KKHN 仓颉: LLP  
 四角号码: 50336 UniCode: U+60A3 规范汉字编号: 2285



五笔86: FJEG 五笔98: FJEG 仓颉: JJB  
 四角号码: 47420 UniCode: U+671D 规范汉字编号: 2559

## • 命名实体识别方法（3）：基于统计

- 相应的，基于部首信息，对模型进行改进
  - 基础模型：采用LSTM+CRF的方式实现医疗命名实体识别
    - 单纯LSTM忽略了标签序列的关联性，CRF将提升标签序列的合理性
  - 部首信息对于模型的改进体现在以下两个方面
    - LSTM部分，在字向量编码中加入部首编码，与字符向量拼接来表示字符
    - CRF部分，加入部首标签矩阵，区分不同部首对应不同类型的不同可能性
- 李丹等，部首感知的中文医疗命名实体识别，中文信息学报，2020

## • 命名实体识别方法（3）：基于统计

- 实验证实，引入部首信息之后，在医疗命名实体识别任务上取得了更好效果
- 案例1：“患者肺部轻度慢性发炎”
  - 实体嵌套，传统方法容易拆成多个实体对待，导致实体支零破碎
  - 基于部首信息，可抽取出以“肺”为开始，“炎”为结尾的实体词
- 实例2：“粘膜上层及粘膜层均见低分化浆液性腺癌浸润”
  - 典型的结构复杂的较长实体，传统方法往往仅能识别“腺癌”
  - 由于其中多个字符具有鲜明部首特征，因此可以识别完整实体

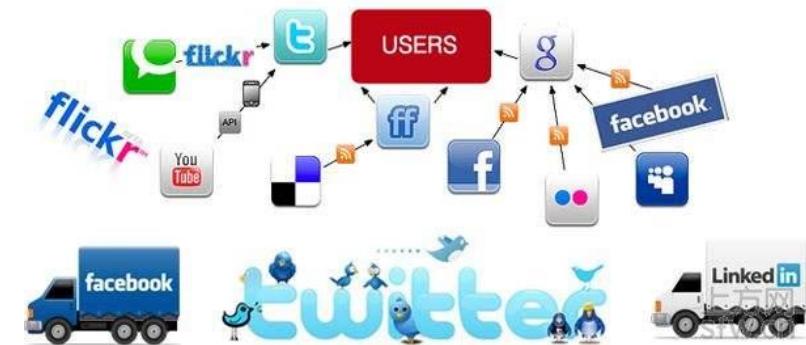
- 与命名实体识别相关的问题：实体对齐

- 实体对齐 (Entity Alignment) , 也称实体匹配 (Entity Matching)
- 指对于异构数据源知识库中的各个实体，找出属于现实世界中的同一实体。
  - 例如，不同药物可能在不同数据库中采用不同的名称
    - E.g., 利君沙 (琥乙红霉素片)
- 一般而言，利用实体的属性信息判定不同源实体是否可对齐



- 与命名实体识别相关的问题：实体对齐

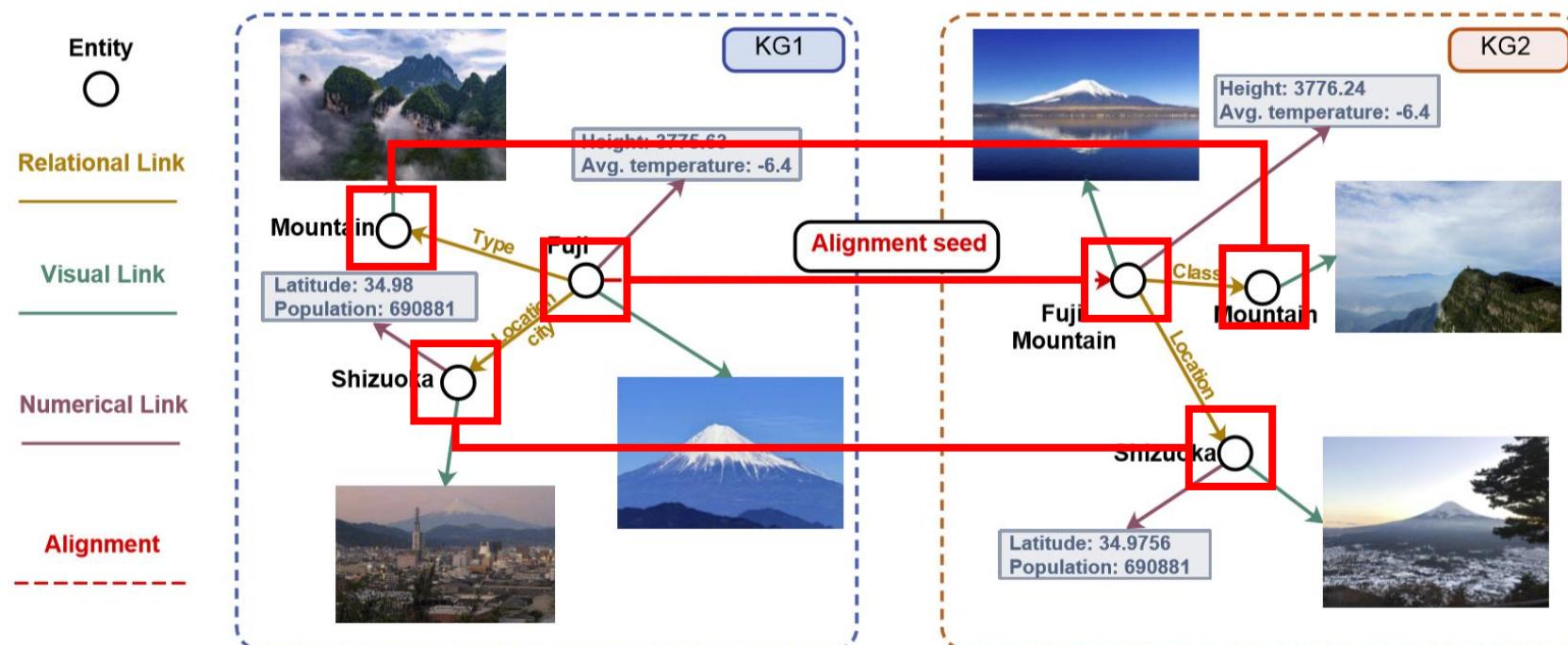
- 近来，针对跨知识图谱（KGs）的实体对齐任务，研究者提出并改进了多种基于表征（Embedding）的模型。
  - 不仅利用实体的属性和语义信息，还利用实体间的关系。
  - 换言之，这些模型更关注于关系三元组（relationship triple）。
  - 一个类似的任务：跨社交网络用户匹配
    - 不仅考虑用户画像，也考虑社交关系相似性



- 与命名实体识别相关的问题：实体对齐

- 进阶任务：多模态知识图谱中的实体对齐

- 不仅需要利用实体的属性、语义和关系，还要有效处理其多模态特性



## • 与命名实体识别相关的问题：实体消歧

- 实体消歧 (Entity Disambiguation)，本质在于一个单词很可能有多个意思
- 这就意味着，在不同的上下文中所表达的含义可能不太一样。
  - 例如，介绍查询意图的歧义问题时提及的“苹果”

id	实体名	实体描述
1001	苹果	美国一家高科技公司，经典的产品有Iphone手机
1002	苹果	水果的一种，一般产自于...
...	...	...

## • 与命名实体识别相关的问题：实体消歧

- 解决实体歧义问题，首先需要获取实体的各种不同含义
  - 首先，对不同的含义抽取其相关内容，如描述文本，并建立关键词表
  - 其次，通过对关键词表的语义分析，从中抽取和归并相应的“概念”
    - 例如，苹果（水果）可能对应“富士”、“烟台”等，而苹果（手机）可能对应“iPhone”、“刘海屏”等。
  - 最终，对关键词进行语义表征，得到不同语义的表征向量。
- 由此，可以通过语义相似性（如余弦相似度）判断究竟属于哪种语义的实体。

- 命名实体识别的常用工具

- 英文命名实体识别的常用工具

- Stanford NER

- 斯坦福大学开发的基于CRF的NER系统，基于CoNLL、MUC-7和ACE等语料训练

- <https://nlp.stanford.edu/software/CRF-NER.shtml>

- MALLET

- 麻省大学开发的统计自然语言处理的开源包，其序列标注工具的应用中能够实现NER。

- <http://mallet.cs.umass.edu/>

- 命名实体识别的常用工具

- 中文命名实体识别的常用工具

- NLPIR-ICTCLAS: <http://ictclas.nlpir.org/nlpir/>

- 介绍中文分词时提到的可视化分词工具，同时可实现词性判别与实体识别

- HanLP: <http://hanlp.linrunsoft.com/>

- 一系列模型与算法组成的NLP工具包，支持命名实体识别

- NLTK: <http://www.nltk.org/>

- 一个高效的Python构建的平台，用来处理人类自然语言数据。

# 本章小结

## 实体识别

- 信息抽取概述：定义、五种基本任务
- 知识图谱概述
  - 发展历史、雏形思想
  - 基本要素：点（实体）、边（关系）、三元组
- 命名实体识别
  - 基于词典或规则的基础方法
  - 基于统计的方法：分类、序列标注

[tongxu@ustc.edu.cn](mailto:tongxu@ustc.edu.cn)