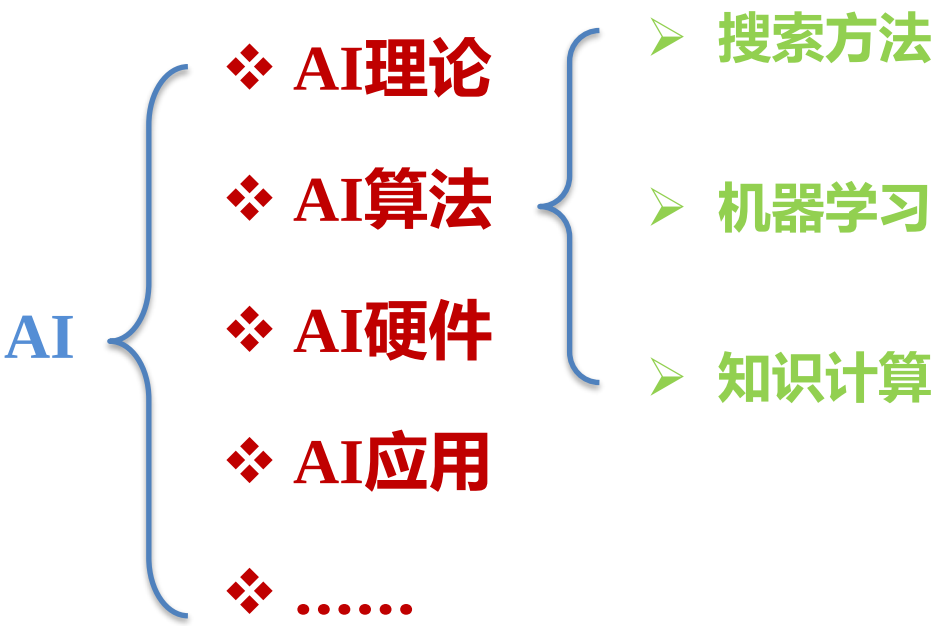
华东师范大学计算机科学与技术学院
School of Computer Science and Technology

2020年春季学期

# 《人工智能》

## 第十三讲：知识图谱（一）

# 课程大纲



# 机器学习大纲

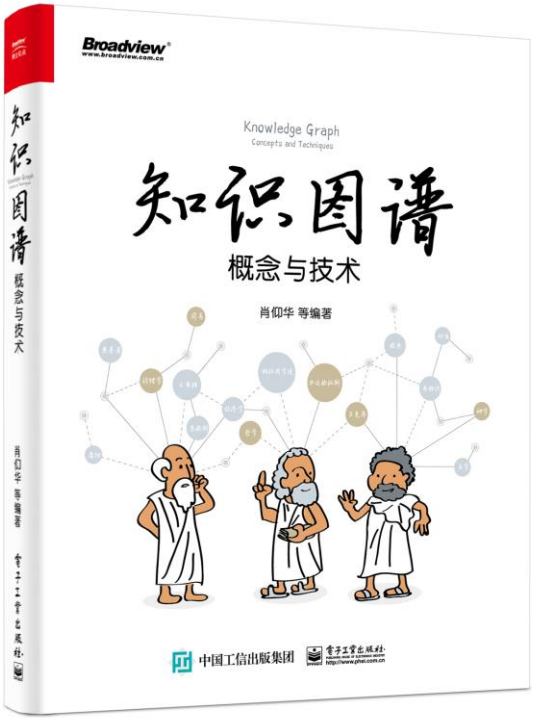
- AI
  - ❖ AI理论
  - ❖ AI算法
    - ➢ 搜索方法
    - ➢ 机器学习
    - ➢ 知识计算
      - ○ 知识工程概况
      - ○ 知识表示
      - ○ 知识图谱
      - ○ 归结推理
      - ○ 贝叶斯推理
  - ❖ AI硬件
  - ❖ AI应用
  - ❖ ……

# 参考教材

- 知识计算阶段

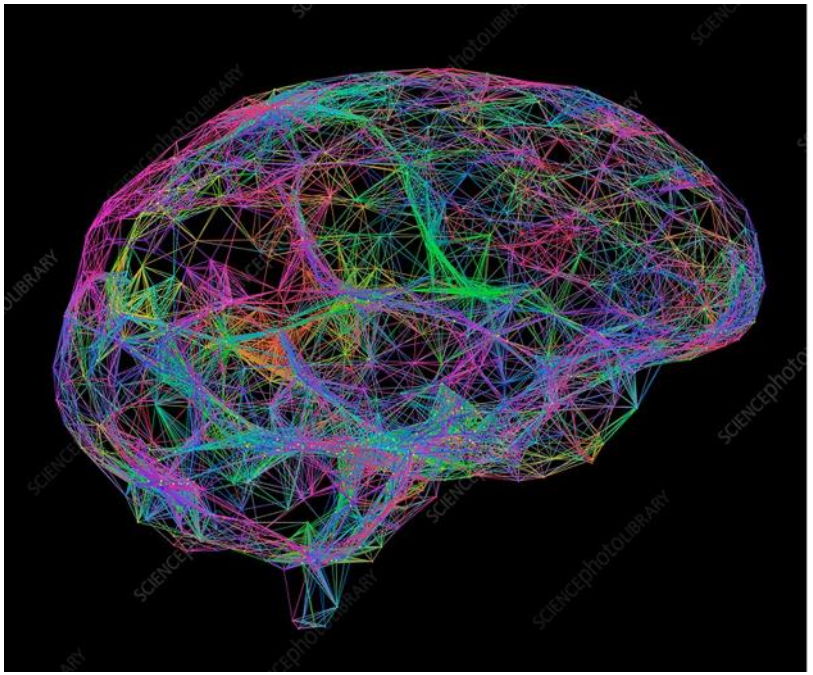
  知识工程概述与知识表示

  肖仰华《知识图谱：概念与技术》，电子工业出版社，第三章——第六章



# 第十三讲：知识图谱（一）

- **章节概要**
  - 13.1 知识图谱概况
  - 13.2 词汇与实体挖掘



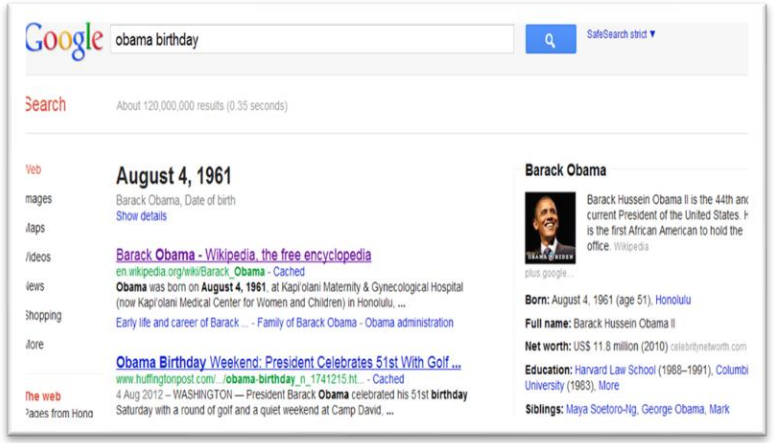
# 13.1 知识图谱概况

# 诞生标志

- **2012年5月，Google收购Metaweb公司，并正式发布知识图谱**

- **搜索核心需求： 让搜索通往答案**
  - ➢无法理解搜索关键词
  - ➢无法精准回答

- **根本问题**
  - ➢缺乏大规模背景知识
  - ➢传统知识表示难以满足需求





https://www.fastcompany.com/1671024/google-buys-metaweb-one-company-could-revolutionize-google-search

# 知识图谱的狭义概念

- **知识图谱(Knowledge Graph)本质上是一种大规模语义网络（semantic network)**
  - 富含**实体**(entity)、**概念**(concepts)及其之间的各种**语义关系**(semantic relationships)

作为一种语义网络，是大数据时代**知识表示**的重要方式之一

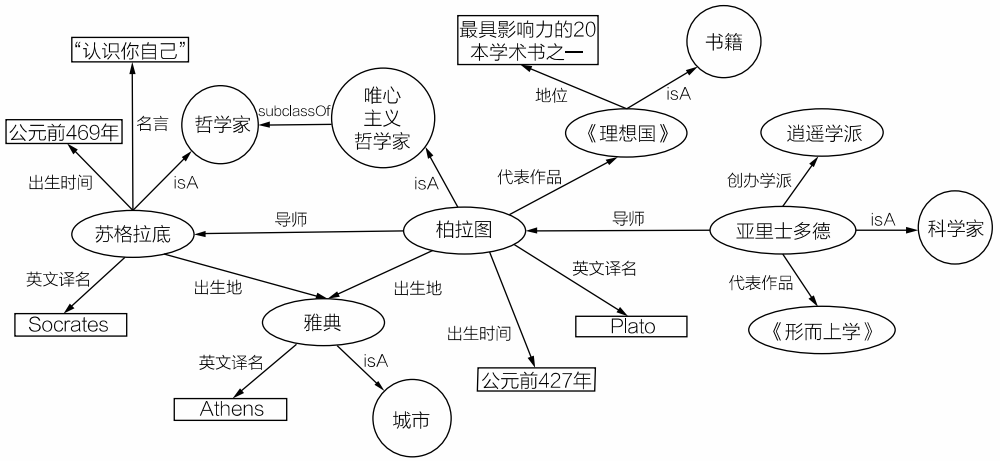
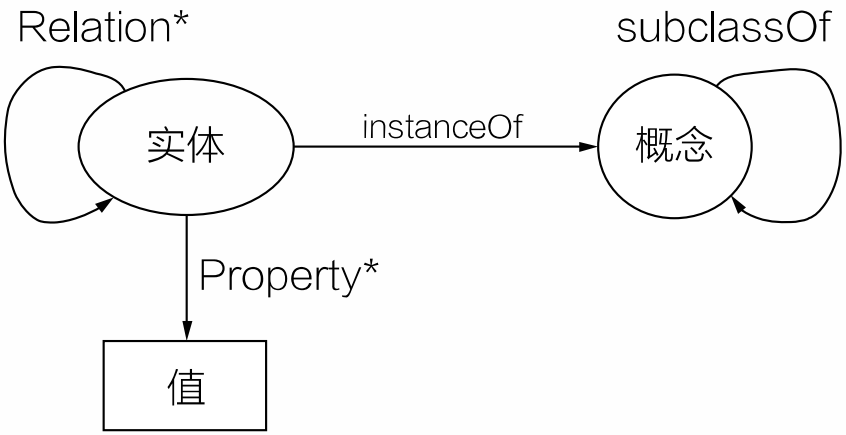



图 1-1　关于古希腊三大哲学家的知识图谱片段

# 知识图谱基本组成

- 知识图谱是一种以图形化的(Graphic)形式通过点和边表达知识的方式[1]，其基本组成元素是点和边



# KG组成- 节点-实体

- **Entity/Objects/Instances**
  - Wikipedia: An **entity** is something that exists as itself, as a subject or as an object, actually or potentially, concretely or abstractly, physically or not.
  - 黑格尔《小逻辑》：能够独立存在的，作为一切属性的基础和万物本原的东西

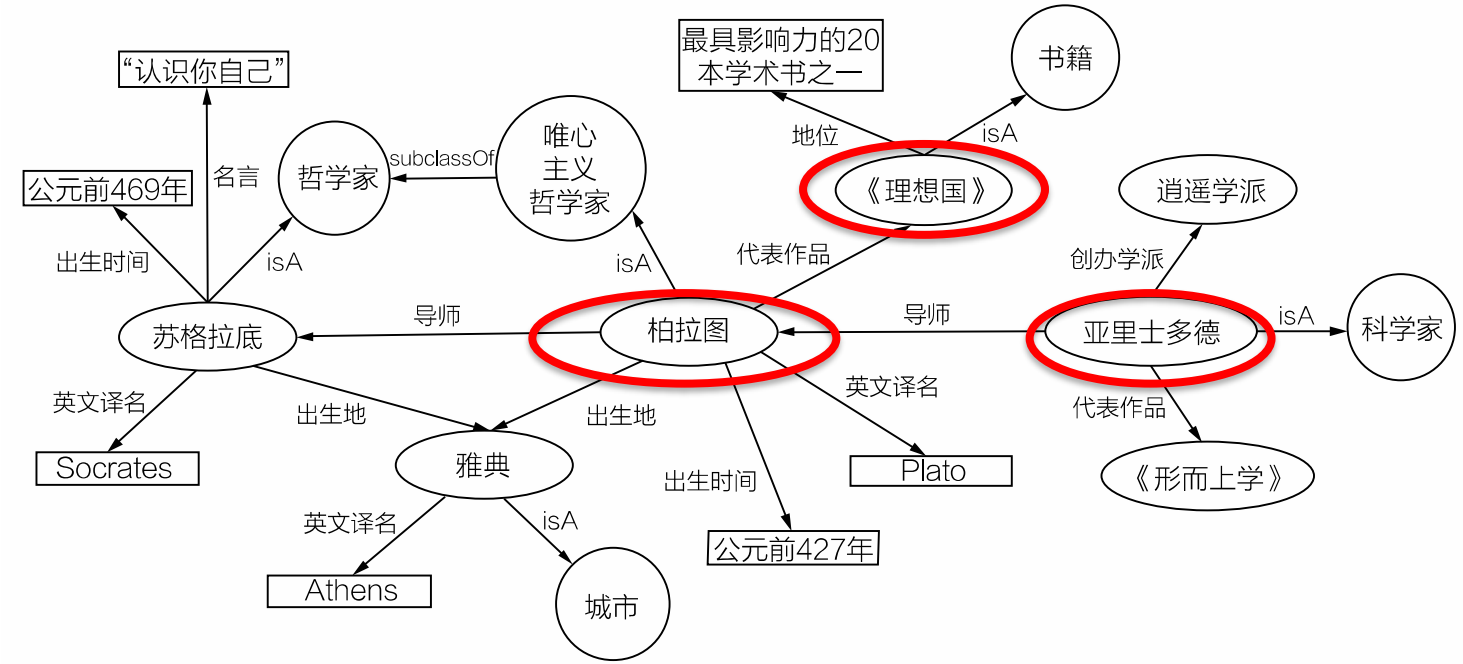



图 1-1　关于古希腊三大哲学家的知识图谱片段

# KG组成- 节点-概念

- **Concept**
  - In metaphysics, and especially ontology, a concept is a fundamental category of existence.
  - (mental) representations of categories
- **Category**
  - Groups of entities which have something in common;

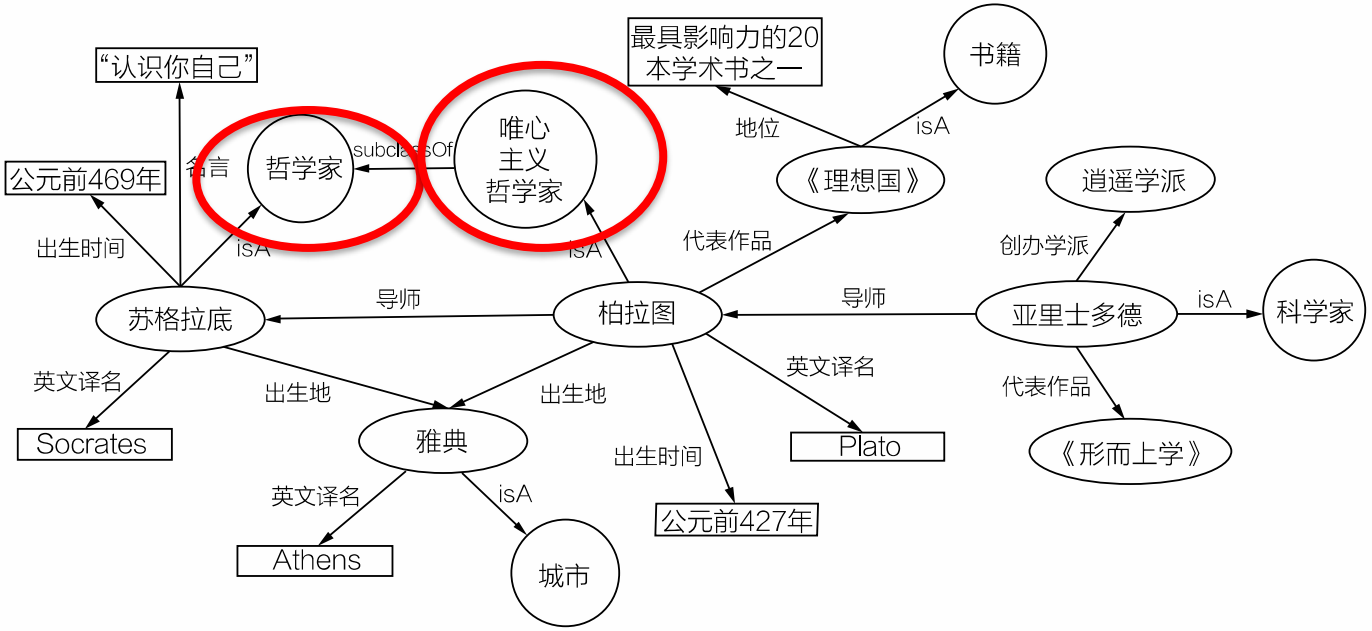



图 1-1　关于古希腊三大哲学家的知识图谱片段

# KG组成- 节点-值

- **Date**
  - 特朗普 出生日期 1946年6月14日
- **String**
  - 特朗普 简介 “唐纳德·特朗普（Donald Trump），第45任美国总统，1946年6月14日生于纽约，美国共和党籍政治家”
- **Numeric**
  - 特朗普 年龄 71

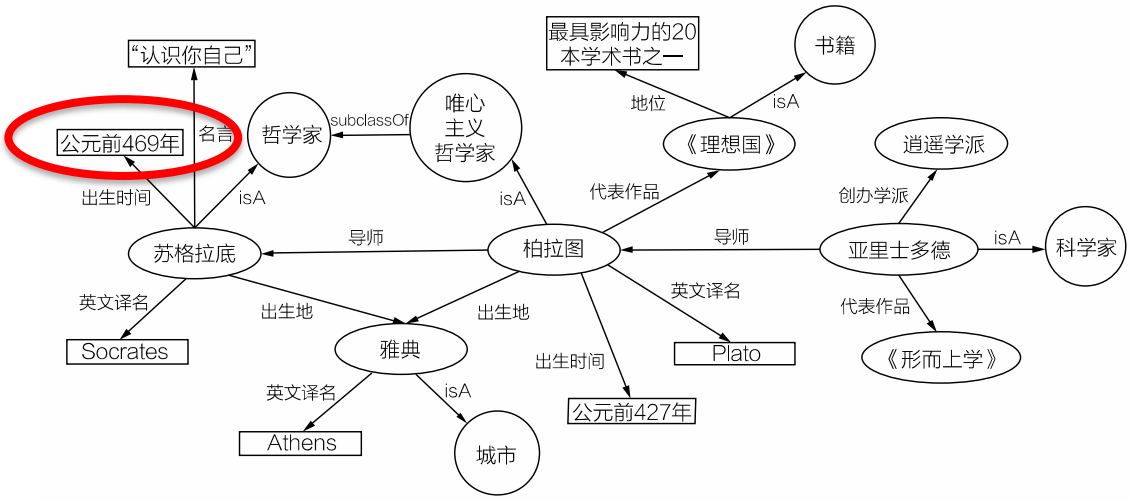



图 1-1　关于古希腊三大哲学家的知识图谱片段

# KG-组成-边

- **Relation**
  - 侧重实体(individual)之间的关系
- **Property/Attribute/Quality**
  - A characteristic/quality that describes an object

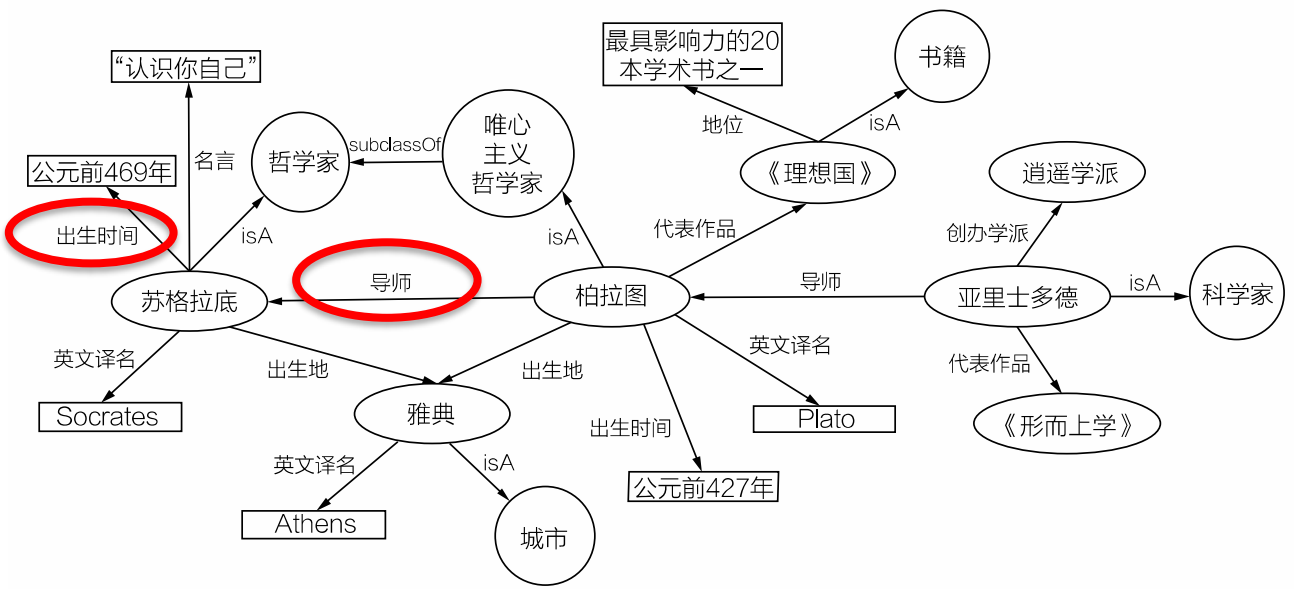
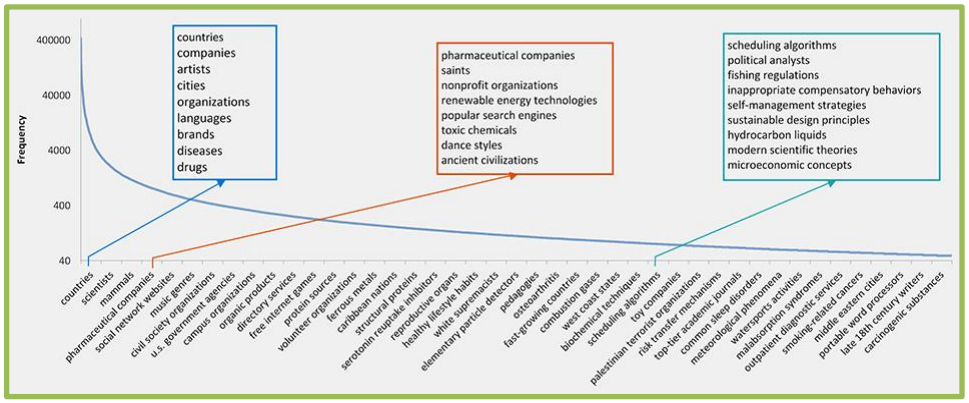



图 1-1　关于古希腊三大哲学家的知识图谱片段

# KG优势1： large scale

- **Higher coverage over entities and concepts**

| KGs | # of Entities/Concepts | # of Relations |
|---|---|---|
| YAGO | 10 Million | 120 Million |
| DBpedia | 28 Million | 9.5 **Billion** |
| Probase | 2.7 Million | 70 **Billion** |
| BabelNet | 14 Million | **5 Billion** |
| CN-DBpedia | 17 Million | 200 Million |





| Existing Taxonomies | Number of Concepts |
|---|---|
| Freebase [5] | 1,450 |
| WordNet [13] | 25,229 |
| WikiTaxonomy [26] | 111,654 |
| YAGO [35] | 352,297 |
| DBPedia [1] | 259 |
| ResearchCyc [18] | ≈ 120,000 |
| KnowItAll [12] | N/A |
| TextRunner [2] | N/A |
| OMCS [31] | N/A |
| NELL [7] | 123 |
| **Probase** | **2,653,872** |

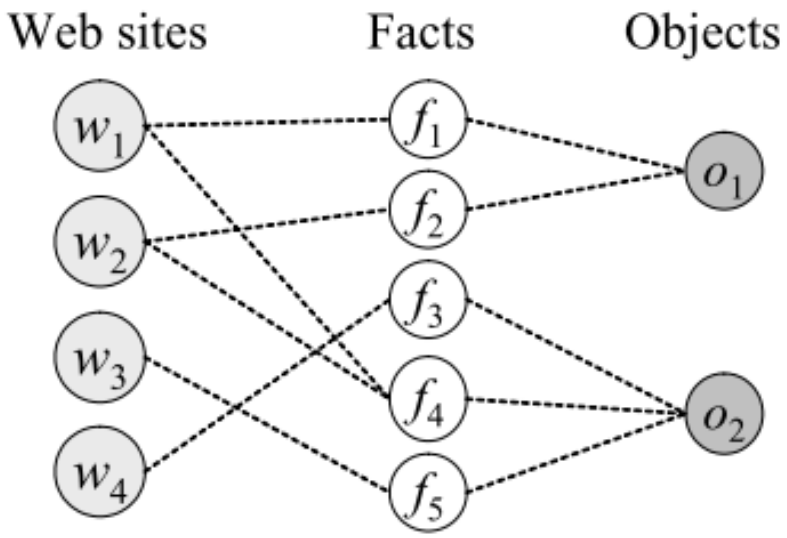
# KG优势2: semantically rich

- **Higher coverage over numerous semantic relationships**

| KGs | # of Relations |
|---|---|
| DBpedia | 1,650 |
| YAGO1 | 14 |
| YAGO3 | 74 |
| CN-DBpedia | 100 Thousands |

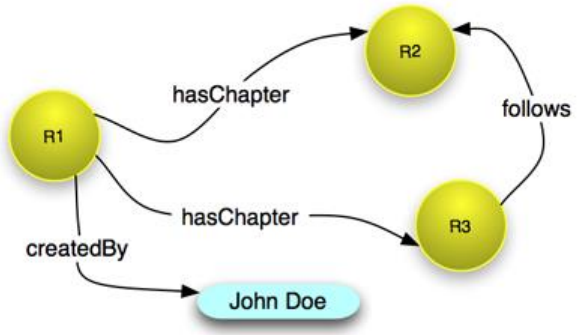
# KG优势3： high quality

- **High quality**
  - Big data: Cross validation by multiple sources
  - Crowd sourcing: quality guarantee

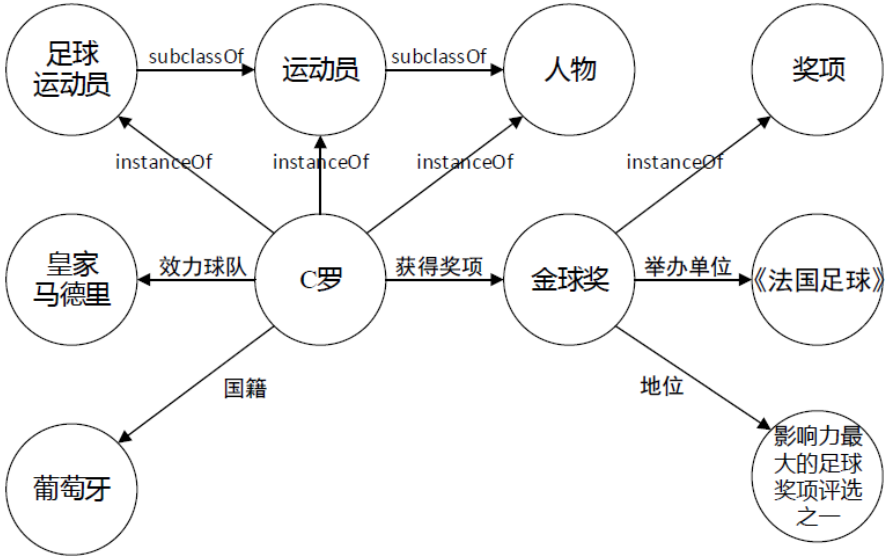




# KG优势4: friendly structure

- **Structured organization**
  - By RDF
  - By graph




| Subject | Predicate | Object |
|---|---|---|
| R1 | hasChapter | R2 |
| R1 | hasChapter | R3 |
| R3 | follows | R2 |
| R1 | createdBy | "John Doe" |



# KG的不足1：高质量模式缺失

- **提升知识图谱的规模往往会付出质量方面的代价**
  - 可以预先定义人的“身高”取值范围为0.5m ~ 2.3m，但可能存在某个人，其身高达到2.31m
  - “妻子”作为一条关系通常只有单一取值，不可以是多值的，但是古代人未必如此，当今世界的某个偏远部落也未必如此
- **知识图谱在设计模式时通常会采取一种“经济、务实”的做法：也就是允许模式（Schema）定义不完善，甚至缺失**

模式定义不完善或缺失对知识图谱中的数据语义理解以及数据质量控制提出了挑战

# KG不足2：封闭世界假设不再成立

- **传统数据库与知识库的应用通常建立在封闭世界假设（CWA）基础之上。 CWA 是假定数据库或知识库中不存在（或未观察到）的事实即为不成立的事实**
- **大多数开放性应用不遵守这一假设。也就是说，在这些应用中缺失的事实或知识未必为假**
  - 很难保证知识图谱中关于柏拉图的信息完整，很可能会缺失柏拉图父母的信息。但常识告诉我们柏拉图一定有父母。

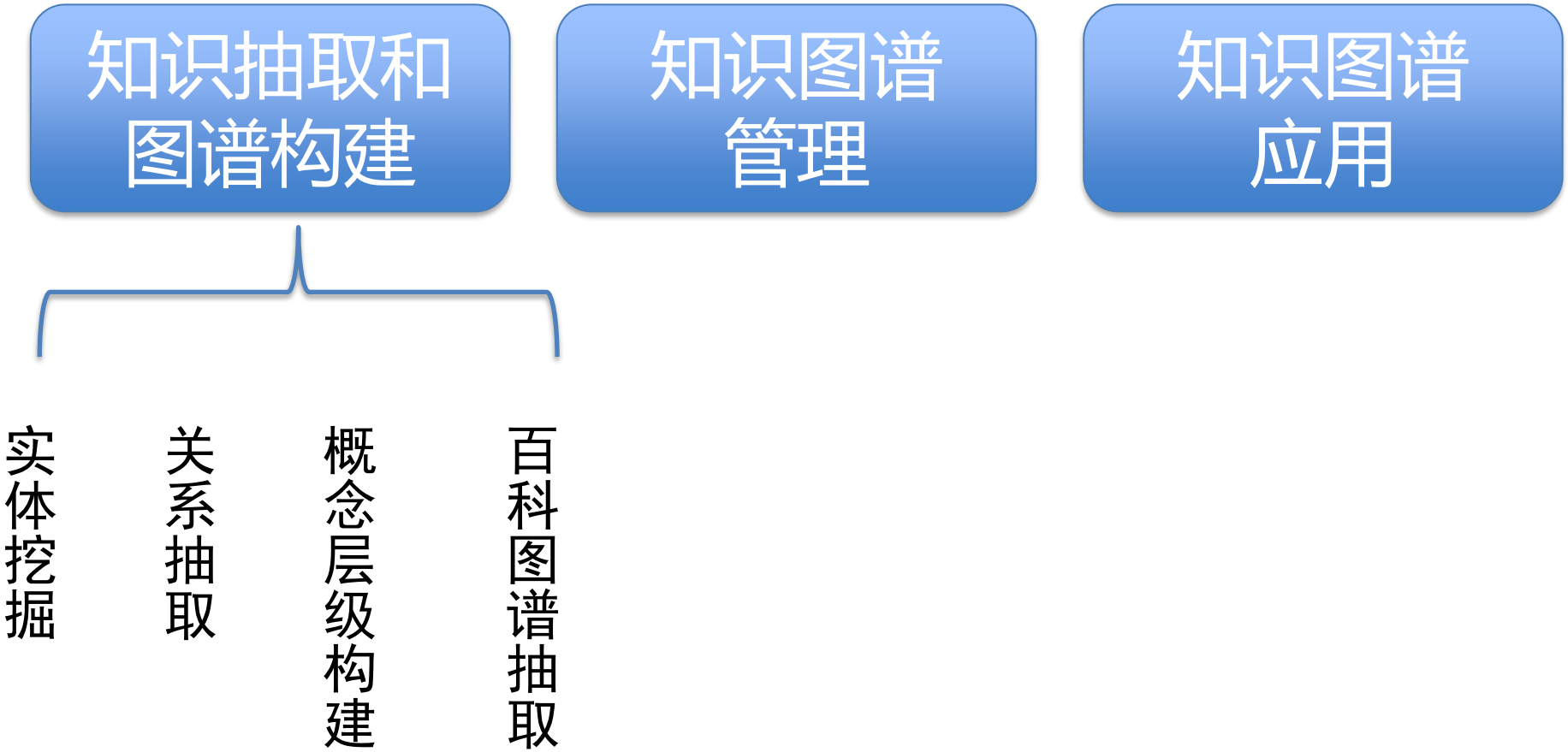
不遵守CWA 给知识图谱上的应用带来了巨大的挑战

# KG不足3：大规模自动化知识获取成为前提

- **传统知识工程依赖专家完成知识获取，这一方式难以实现大规模知识获取，难以满足知识图谱的规模要求**
- **大规模自动化知识获取是知识图谱与传统语义网络的根本区别**

- **大规模自动化知识获取的方式是多样的**
  - 从文本中自动抽取
  - 基于大规模众包平台的知识标注
  - 多种方式混合

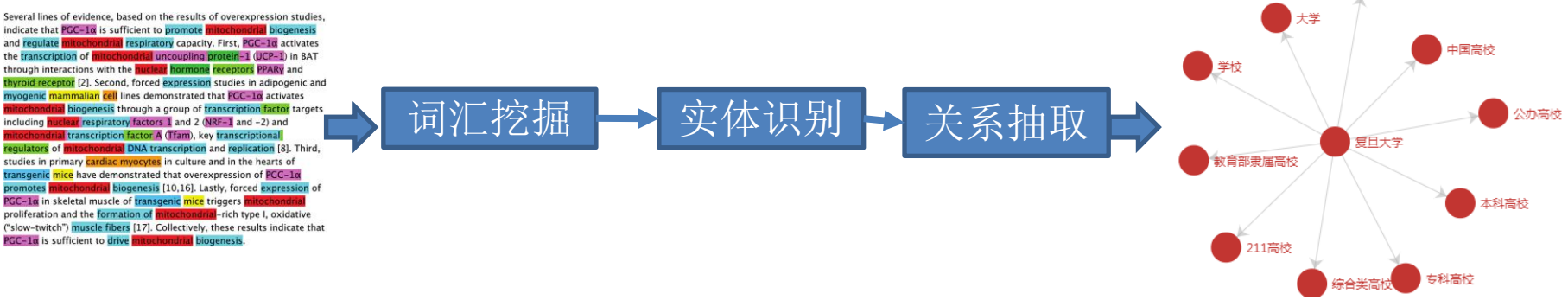
# KG的主要问题



# 13.2 词汇和实体挖掘

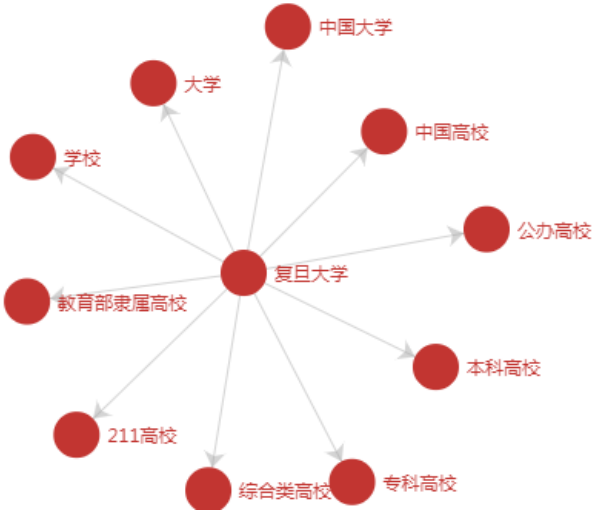
# 知识图谱构建的第一步

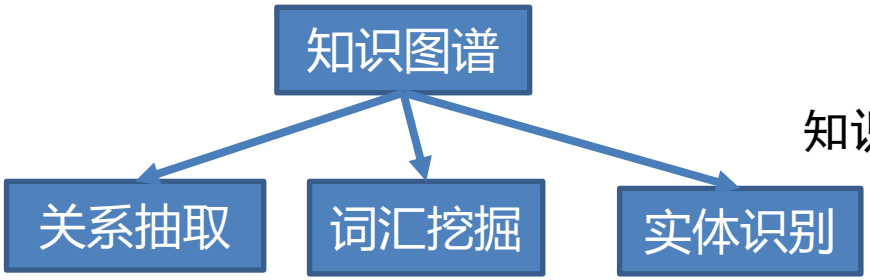
- **知识图谱中的点**
  - 词汇，eg，“知识图谱”
  - 实体，eg，刘德华



# 词汇挖掘必要性

- **理解一个领域往往是从理解领域词汇开始的**
- **与图书情报领域的叙词表（主题词表）构建相关**
- **词汇知识是理解用户意图的关键知识**
- **广泛应用**
  - 比如，猎头如果要寻找知识图谱领域的专家或学者，只需要判断候选人的简历或者论文题目中是否包含知识图谱的领域词汇

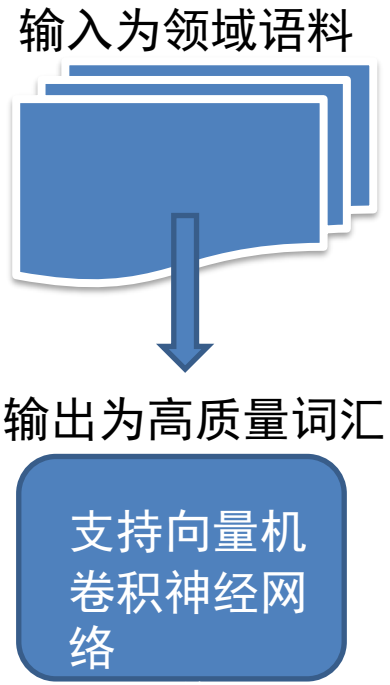



知识图谱中的实体、概念是通过词汇表达的

# 高质量短语挖掘

- **领域词汇挖掘指的是从给定的领域语料中自动挖掘属于该领域的高质量词汇的过程。**
- **高质量短语**
  - 高频率：一个 N-Gram 在给定的文档集合中要出现得足够频繁才能被视作高质量短语
  - 一致性：N-Gram 中不同单词的搭配是否合理或者是否常见
  - 信息量:一个高质量短语应该传达一定的信息，表达一定的主题或者概念
    - 比如，“机器学习”与“这篇论文”
  - 完整性:一个高质量短语还必须在特定的上下文中是一个完整的语 义单元。
    - 比如，“vector machine” vs“support vector machine”



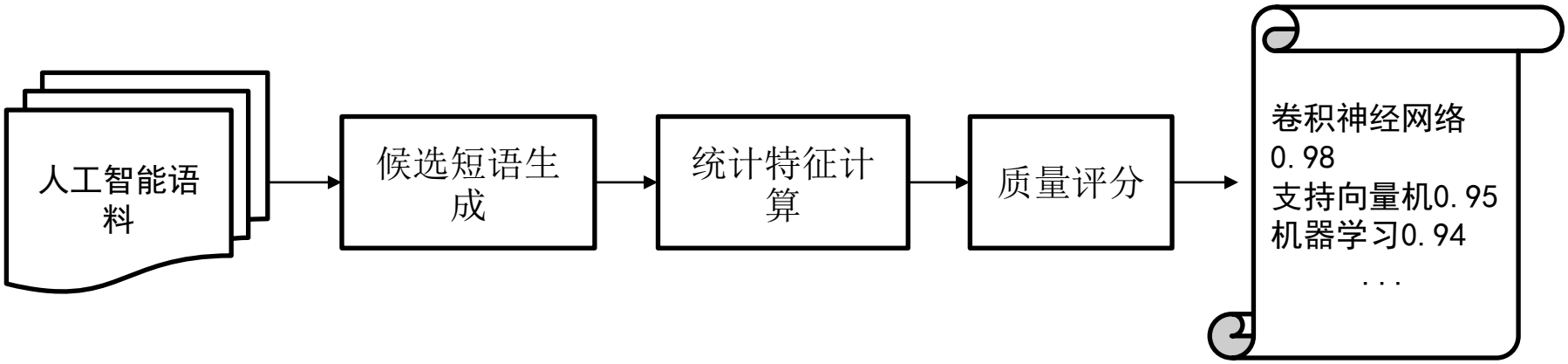
# 领域短语挖掘

- **基于规则的**
- **无监督**
  - 通过预定义的词性标签（POS Tag）规则来识别文档中的高质量名词短语。
  - 缺陷：规则一般是针对特定领域手工设计的，难以适用于其他领域。人工定义规则代价高昂，难以穷举所有的规则，因此召回率存在一定的局限性。
- **有监督**
  - 利用标注好词性的语料来自动学习规则
  - 缺陷：依赖于领域的语言规则以及昂贵的人工标记，不适用于新兴的大型语料。另外词性标注不能做到百分百的准确，这会在一定程度上影响后续规则学习的准确率。

- **基于统计学习**
- 无监督
  - 通过计算候选短语的统计指标特征从而给词汇打分、排序来进行领域词汇挖掘。
- 有监督
  - 根据人工或自动标注的高质量短语，建立高质量短语分类模型。
  - 使用wiki中存在的词条做自动标注

# 基于无监督学习的领域短语挖掘

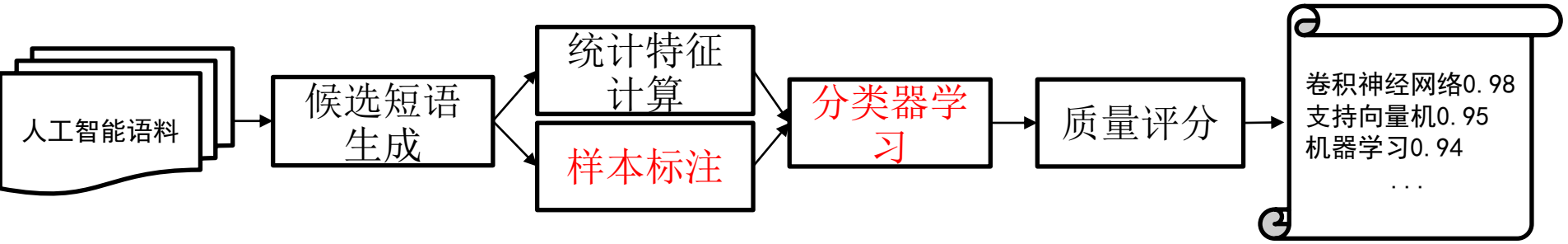
1. 候选短语生成：N-Gram得到高频候选短语。
2. 统计特征计算：如计算TF-IDF和PMI等。
3. 质量评分：融合这些特征的值（如加权求和等）得到短语的最终分数。
4. 排序输出：取topK或根据阈值筛选词汇输出。



# 有监督学习

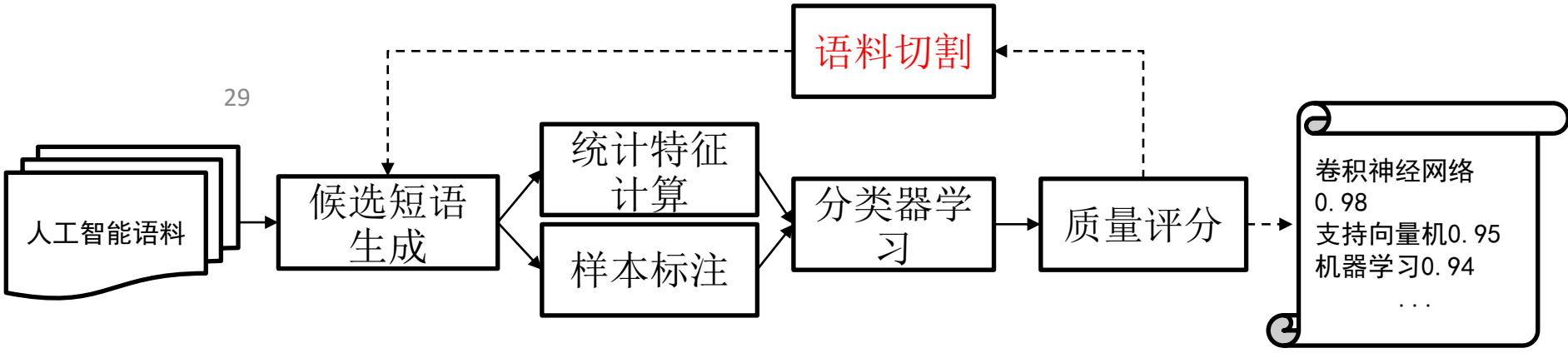
新增两个步骤：

1. 样本标注：人工标注或者远程监督标注样本。
2. 分类器学习：根据正负样本，学习一个二元分类器。分类器模型可以是决策树、随机森林或者支持向量机。对于每个样本，使用 统计指标（TF-IDF、C-value以及 PMI 等）构造相应的特征向量。



# 有监督+频次统计优化

- 基于 N-Gram 的原始频次统计方法需要修正与优化：
  - 因为该统计下子短语的词频一定大于父短语。比如在人工智能语料中“向量机”和“支持向量”的词频一定大于或等于“支持向量机”。但事实上“支持向量机”的质量更高。
- 语料切割：
  - 利用模型来识别高质量短语，再根据已经发现的高质量短语对语料进行切割，在切割的基础上重新统计词频，改进词频统计的精度。



# 统计指标：TF-IDF（词频-逆文档频率）

- Motivation:“的”、“是”和“由于”等词汇过于普遍，不适合用来刻画该领域语料的特征。
- 一个词的重要程度与其tf 正相关，idf反相关

$$\mathrm{tf}(u) = \frac{f(u)}{\sum_{u'} f(u')} \qquad \mathrm{idf}(u) = \log \frac{|D| + \delta}{\left|\{j : u \in d_j\}\right| + \delta}$$

避免出现次数为零所导致的分母为零

- Principle: 如果某个短语在领域语料中频繁出现但是在外部文档中很少出现，则该短语很可能是该领域的高质量短语

# 统计指标： C-value

- 词频与长度决定候选短语质量
- 优化词频
  - 父短语的重复统计会带来频次估计的偏差
  - “支持向量机”是个高质量短语，那么“向量机”和“支持向量”的词频就不应该重复计数

$$\mathrm{C-value(u)} = \begin{cases} \log_2|u| \cdot f(u), & u\text{没有父短语} \\ \log_2|u| \left( f(u) - \frac{1}{|T_u|} \sum_{b \in T_u} f(b) \right), & u\text{有父短语} \end{cases}$$

Principle： 一般而言，在很多专业领域（比如医学领域）越长的短语越有可能是专有名词, 从而极可能是高质量短语。

# 统计指标： NC-value

- **在C-value的基础上利用短语丰富的上下文信息。**
  - 先通过 C-value 值对候选短语进行了初步排序，再选取前 5%候选短语的上下文中所出现的单词作为b。上下文单词 b 的重要性值 为 weight(b)。
  - weight(b)越大，说明 b 越倾向于出现在高质量短语的上下文中，因此越有助于找到高质量短语。

$$\mathrm{NC-value}(u) = 0.8\mathrm{C-value}(u) + 0.2\sum_{b \in C_u} f_u(b)\mathrm{weight}(b)$$

b与u的共现次数

$$\mathrm{weight}(b) = \frac{t(b)}{n}$$

与高质量候选短语的共现次数

# 统计指标： PMI（点互信息）

- PMI 值刻画了短语组成部分之间的一致性（Concordance）程度。
  - 如果“电影”和“院”独立地出现在语料中，那么它们联合出现的概率“电影院”应该等于其分别出现的概率之乘积。
  - 对于候选短语，枚举所有可能的拆分方式，计算相应的PMI，取最小为候选短语的PMI
  - PMI（“电影院”）> PMI(“的电影”)

$$\mathrm{PMI}(u_l, u_r) = \log \frac{p(u)}{p(u_l)p(u_r)}$$

- Principle：如果两部分联合出现的概率远大于两者在独立情况下随机共现的概率，说明这两个部分的共现是一个有意义的搭配，预示着两者应该组成一个有意义的短语而非纯粹偶然共现。

# 统计指标：左（右）邻字熵

- 描述词汇的自由搭配程度，也就是用来衡量一个词的左(右)邻字集合的丰富程度。
  - “亚里士多”这个短语的右邻字比较集中，总是“德”字，所以一般不会把它当作一个完整短语，而应将“亚里士多德”当作一个完整短语。
  - H(“亚里士多德”)> H(“亚里士多”)

$$H(u) = -\sum_{x \in \chi} p(x) \log p(x)$$

- Principle：一个词汇的左（右）邻熵越大，左（右）搭配越丰富，则该词汇越有可能是个好的词汇。

# 统计指标特征

| 常用统计指标特征 | 作用 |
| --- | --- |
| TF-IDF | 挖掘能够有效代表某篇文档特征的短语 |
| C-value | 考虑了短语与其父短语的关系来挖掘高质量短语 |
| NC-value | 在C-value 的基础上进一步考虑了上下文来挖掘高质量短语 |
| PMI | 挖掘组成部分一致性较高（经常一起搭配）的短语 |
| 左（右）邻字熵 | 挖掘左（右）邻丰富的短语 |

在领域词汇挖掘中要融合多种统计指标特征，取长补短，挖掘出高质量的领域短语。只考虑单一特征不足以挖掘高质量短语

# 命名实体识别概述

- **命名实体**
  - 是单词或短语
  - 具有标识和区别作用

- **实体**
  - 认知概念，指代世界上存在的某个特定事物
  - 有不同的表示形式，或者不同的提及方式

- **实体指代**
  - 实体在文本中的表示形式通常被称作实体指代(Mention，或者直接被称为指代)
    - 如：周杰伦这个实体，在文本中有时被称作“周董”，有时被称作“Jay Chou”。“周董”、“Jay Chou”就是实体指代。

# 命名实体识别概述

- **命名实体识别（NER）：**
  - 在文本中定位实体并分类为预定义类别
  - Input：token序列$s = \langle w_1, w_2, \cdots, w_N \rangle$
  - Output：$\langle I_s, I_e, t \rangle$
    分别代表开始、结束位置和实体类型


- **NER的类型：**
  - 粗粒度、细粒度
    - 人物，还可以进一步细分为导演、
  - 多标签、单标签
    - 如：“吴京”既是演员又是导演。

# 传统命名实体识别方法

- **基于规则、词典和在线知识库的方法**
  - 基于规则的NER系统：

    LaSIE-II，NetOwl，Facile，

    SAR，FASTUS和LTG
  - 如“White, 33”中的“White”为人名

LTG 系统使用的部分规则

| 规　则 | 标　注 | 举　例 |
|---|---|---|
| Xxxx+, DD+ | 人物 | White, 33， |
| Xxxx+ is? a? JJ* PROF | 人物 | Yuri Gromov, a former director |
| Xxxx+ is? a? JJ* REL | 人物 | John White is beloved brother |
| Xxxx+ himself | 人物 | White himself |
| Xxxx+ area | 地点 | Beribidjan area |
| PROF of/at/with Xxxx+ | 组织机构 | director of Trinity Motors |
| shares in Xxxx+ | 组织机构 | shares in Trinity Motors |

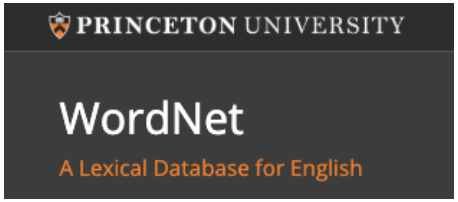
其中，“Xxxx+”代表大写单词序列，“DD”代表数字，“PROF”代表职业，“REL”代表人物关系，“JJ*”代表形容序列

Principle： 基于规则的实 体识别系统往往还需要借助实体词典，对候选实体进行进一步的确认。当词典详尽无遗时，基于规则的系统效果很好。

# 传统命名实体识别方法

**基于WordNet的方法：**

- **基本思想**
  - 计算某个词或实体与 WordNet 中的概念或者实例的语义相似性
  - 将目标词挂载到相应的概念或者实例的上位词下，从而完成实体分类。
- **比如，Mordor 与 WordNet 中的 Country 有着足够强的相似性，因此其应该归类为 Country**




Principle： WordNet 具有丰富的类别体系，因此这一方法可以极大地拓展普 通 NER 模型类别的数量。这一方法无须人为定义模式，也无须标注样本，有时也被归类到无监督学习方法中。

# 监督学习方法

- **建模为多分类或序列标记任务**
- **BIO标注法**
  - 其中 B 表示实体的起 始位置，I 表示实体的中间或结束位置，O 表示相应字符不是实体。

| 亚 | 里 | 士 | 多 | 德 | 出 | 生 | 于 | 斯 | 塔 | 基 | 拉 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| B-PER | I-PER | I-PER | I-PER | I-PER | O | O | O | B-LOC | I-LOC | I-LOC | I-LOC |

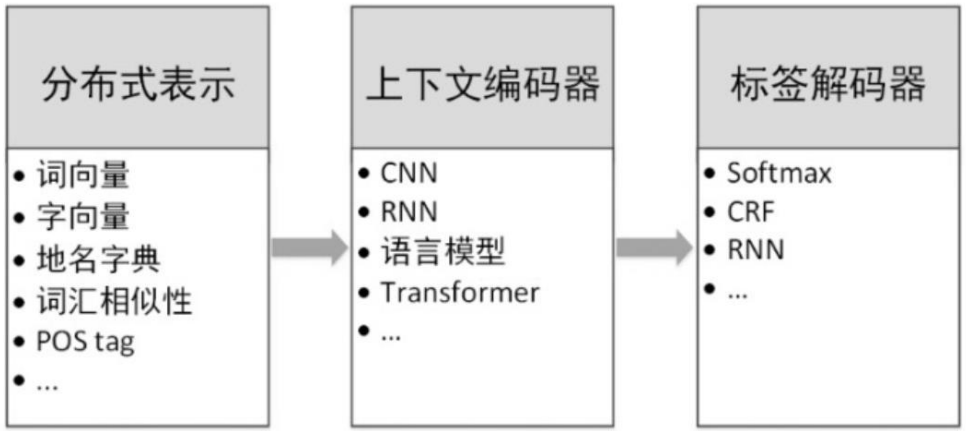
B-PER 表示这个字符是一个人物命名实体的起始位置

I-PER 表 示相应字符为人物实体的中间或结束位置。

类似的，B-LOC 与 I-LOC 代表地点名的起始和中间或结束位置。

# 深度命名实体识别方法

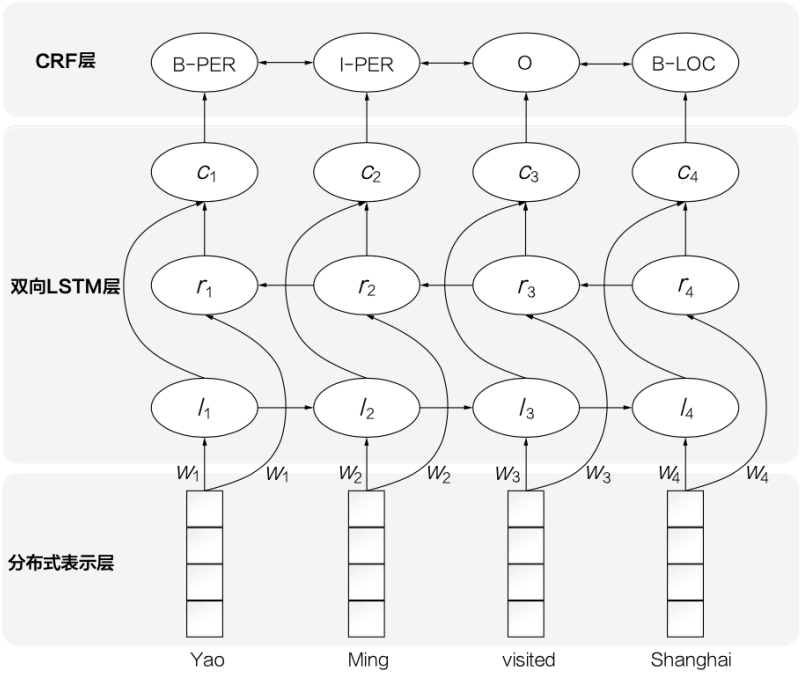
- **基于深度学习的NER框架包含三个模块：**
  - 输入的分布式表示
  (distributed representatio
  - 上下文编码器
  (context encoder)
  - 标签解码器
  (tag decoder)



# 深度命名实体识别方法

- **BiLSTM-CRF**
  - 是基于深度学习的NER最常见的架构
  - 由分布式表示层、双向 LSTM 层，以及 CRF 层构成。

- **相比于传统机器学习模型，深度模型有以下优点：**
  - 并不需要特定的人工制定规则或者繁琐的特征工程
  - 易于从输入提取隐含的语义信息
  - 灵活且便于迁移到新的领域或其他语言




其中$l_i$、 $r_i$分别表示$w_i$左右两边的上下文信息，$c_i$是将$l_i$、 $r_i$拼接起来 后的整体上下文信息 。