

# 知识图谱构建



# 本章大纲

---

- 概念图谱构建
- 百科图谱构建

# 本章大纲

---

- 概念图谱构建
  - 概念图谱概述
  - 概念图谱抽取
- 百科图谱构建

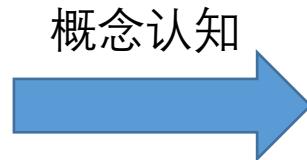
# 本章大纲

---

- 概念图谱构建
  - 概念图谱概述
  - 概念图谱抽取
- 百科图谱构建

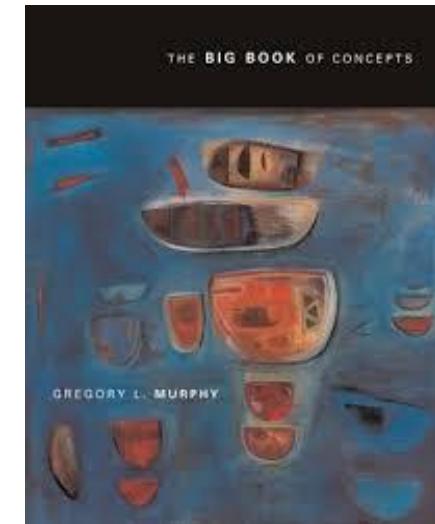
# 概念认知

- 人工智能：“像人一样思考”
- 概念认知是人类思维的基础，是构建人类心灵世界的基石
- 机器的概念认知
  - 是对某个形态的数据输入产生符号化概念输出的过程



猫  
宠物  
动物  
猫科动物  
.....

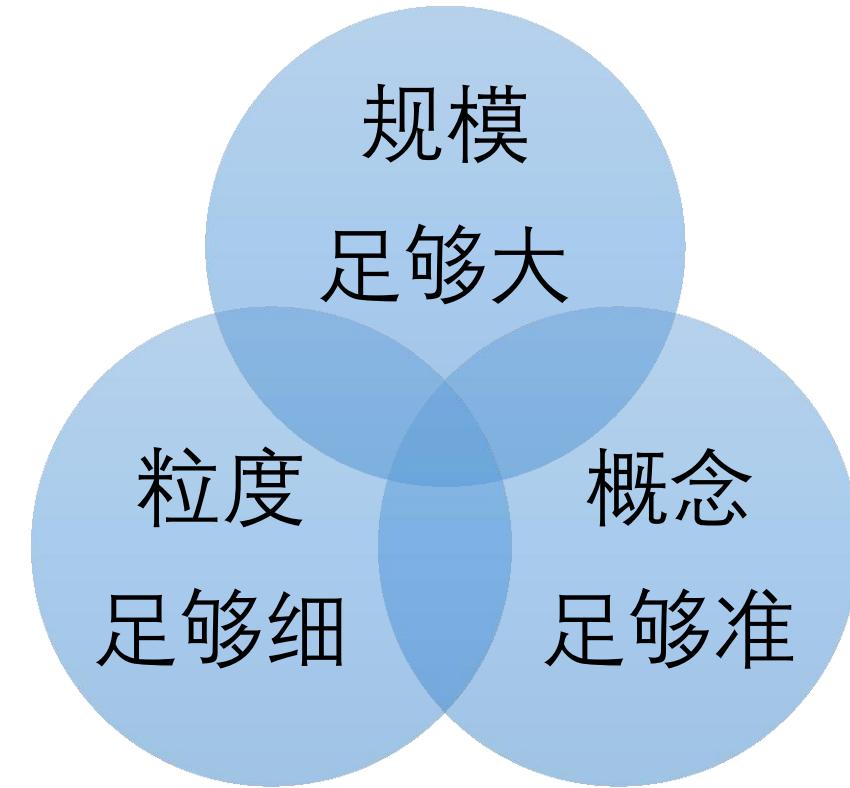
符号形式的概念



“Concepts are the glue that holds our mental world together” , Gregory Murphy , 《The Big Book of Concept》

# 概念认知的重要

- 人类能“理解”事物的重要体现之一就是**产生概念**
  - 柏拉图（实体）→ 哲学家（概念）
- 人类借助概念认知**同类实体**
  - 比如，汽车这一概念使得我们能够认知各种不同类型的汽车，而无需纠缠于各种细节的不同
- 概念是**联想**的重要隐含因素
  - 鸡 → 鸭（家禽）
  - 豆浆 → 油条（早餐）
- 概念是**归纳与推理**的基础
  - 哲学家有自己的哲学观点，柏拉图是哲学家→柏拉图有自己的哲学观点

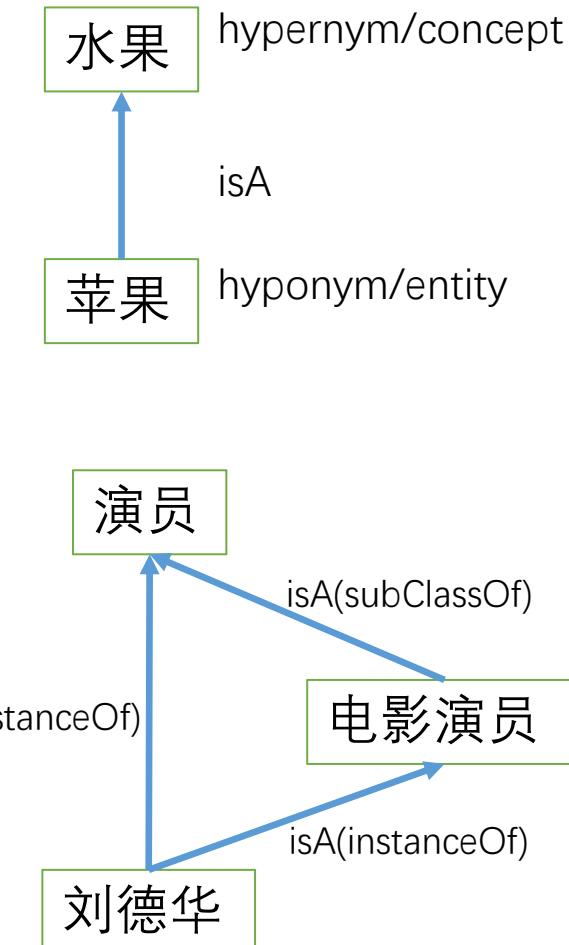


**大规模概念图谱使得机器  
认知实体的概念成为可能**

# 概念图谱的定义

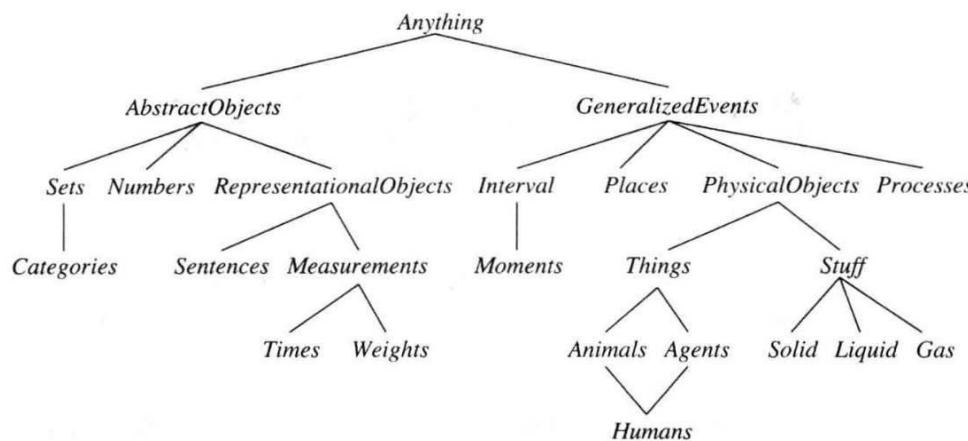
- 概念图谱 (Concept Graph) 是一类专注于实体与概念之间的isA关系的知识图谱。

- 节点：
  - 实体 (如“苹果”，“刘德华”)
  - 概念 (如“水果”，“演员”，“电影演员”)
- 关系：
  - 实体与概念之间的isA关系
    - 如“苹果isA水果”，“刘德华isA演员”
  - 概念与概念之间的subclassOf关系
    - 如“电影演员 subclassOf 演员”



# 概念图谱的分类

- **认知角度**: 概念层级体系 (Taxonomy)
  - 其中的isA关系都是由较具体的实体（或概念）指向较抽象的概念的
  - 有严格的层级结构，形成**有向无环图**
- **语言角度**: 词汇概念层级体系 (Lexical Taxonomy)
  - 基本关系是**词汇**之间的上下位关系
  - 比如，“apple isA fruit”，apple 是fruit 的下位词 (**Hypernym**)，fruit 是apple 的上位词 (**Hyponym**)
  - 可能因为歧义而存在环



概念图谱	图中的节点	边	结 构
概念层级体系 (Taxonomy)，面向认知	概念与实体，如公司、动物	实体与概念之间的instanceOf关系；子概念与父概念之间的subClassOf关系。两类关系统称为isA关系	有严格的层级结构，有向无环图
词汇概念层级体系 (Lexical Taxonomy)，面向语言	自然语言描述的实体与概念，如“苹果”（可能指一种水果，也可能指一家公司）	上下位关系 (Hypernymy-Hyponymy)	有粗略的层级结构，可能由于歧义而存在环

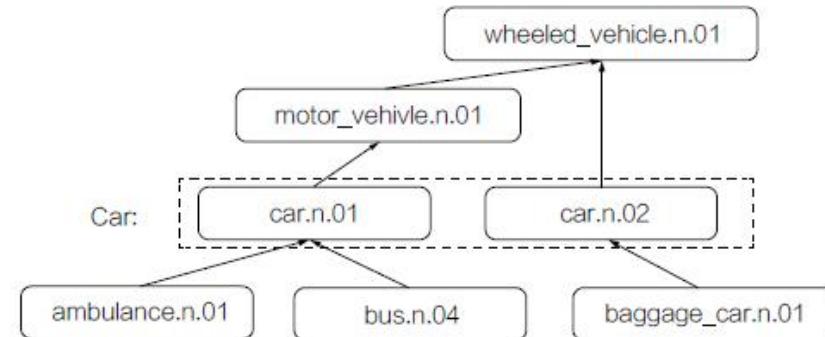
# 常见的概念图谱

## • WordNet

- 普林斯顿认知科学实验室于1985年建立的英文词典

- 专家构建，准确度极高
- 实体按词义 (synset) 组织，已经过消歧
- 包含两种关系：
  - 词汇关系：存在于词形之间
  - 语义关系：存在于词义之间
- 规模较小，包含大约155287个单词(117659个词义或同义词集)

概念图谱	作 者	实 体	概 念	isA 关系数	准确度	权 重
WordNet (英文)	普林斯顿认知科学实验室	—	117 659	84 428	100%	无



WordNet 中部分名词的同义词词集（其中 n 表示词性为名词，n 右边的数字标号是该词的词义序号）

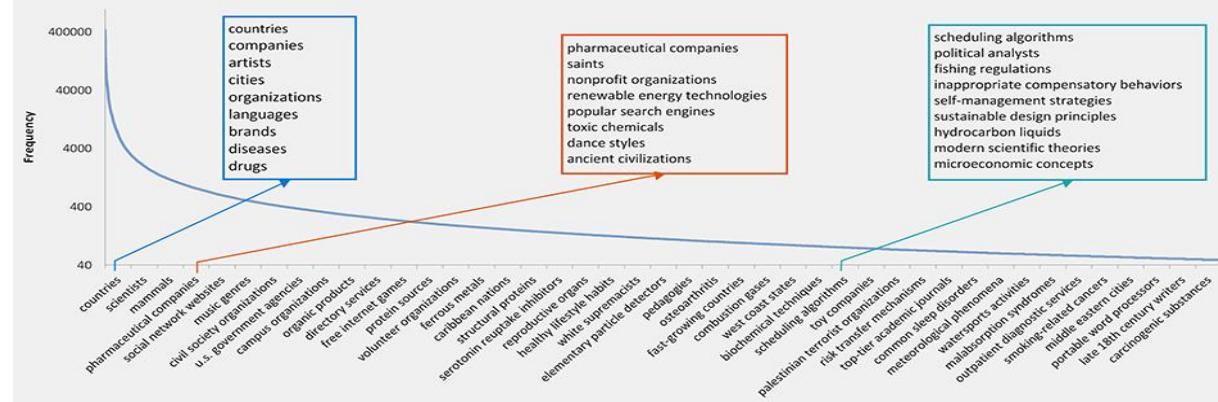
# 常见的概念图谱

- **WikiTaxonomy:** 2008年，Ponzetto和Strube抽取的分类体系

- 数据来源于维基百科数据
- 抽取的isA知识以RDFS形式表示
- 从127,325个类和267,707的链接产生了105,418条IsA关系。

- **Probase:** 2012年微软公司提出的研究原型

- 从网页数据和搜索记录数据构造
- 包含5,401,933个概念，12,551,613个实例和87,603,947个IsA关系
- 现已更名为Microsoft Concept Graph



<https://concept.research.microsoft.com/Home/Introduction>

概念图谱	作 者	实 体	概 念	isA 关系数	准确度	权重
WikiTaxonomy (英文)	欧洲媒体实验室	121 359	76 808	105 418	85%	无
Probase (英文)	微软亚洲研究院	10 390 064	2 653 872	16 285 393	92%	有

# Probbase的频率信息

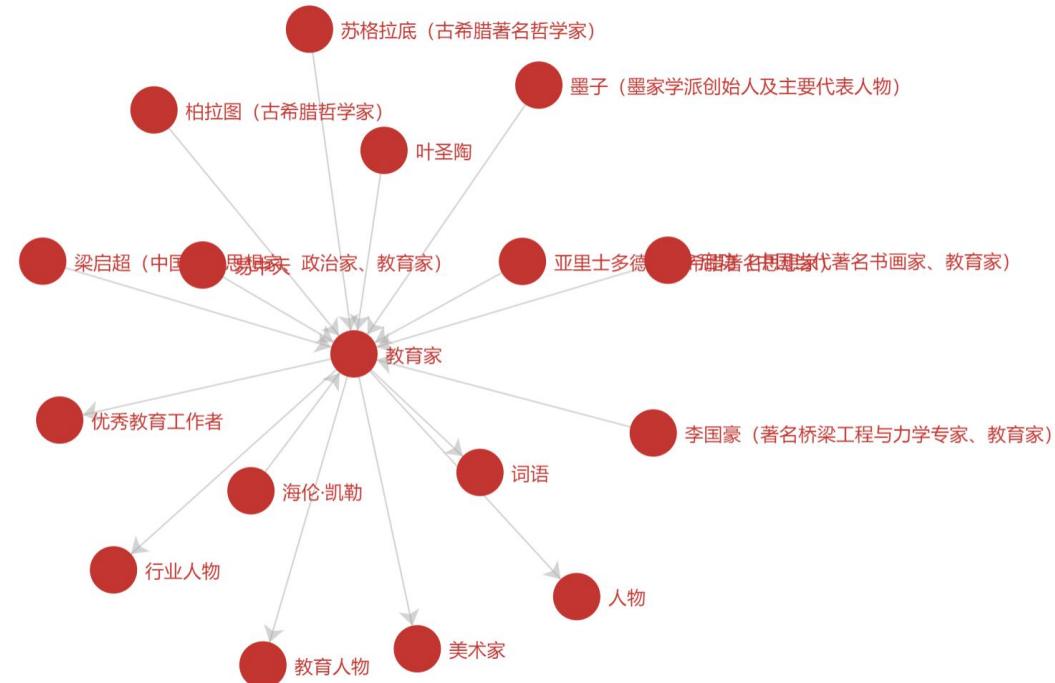
- Probbase中的频数表示该关系在语料中出现的次数
- 对于刻画实体或概念的典型性具有重要意义
  - $P(c|e) = \#(e \text{ isa } c) / \#e$ ;  $P(e|c) = \#(e \text{ isa } c) / \#c$

表 5-3 Probbase 示例

实 体	isA	概 念	频 数
Google	isA	Company	7816
Basketball	isA	Sport	6423
Apple	isA	Fruit	6315
Microsoft	isA	Company	6189

# 常见的概念图谱

- **CN-Probase:** 复旦大学知识工场实验室研发和维护
  - 目前规模最大的开放领域中文概念图谱和概念分类体系
  - IsA关系的准确率在95%以上
  - 包含约1700万实体、27万概念和3300万isA关系
  - 严格按照实体进行组织，有利于精准理解实体的概念

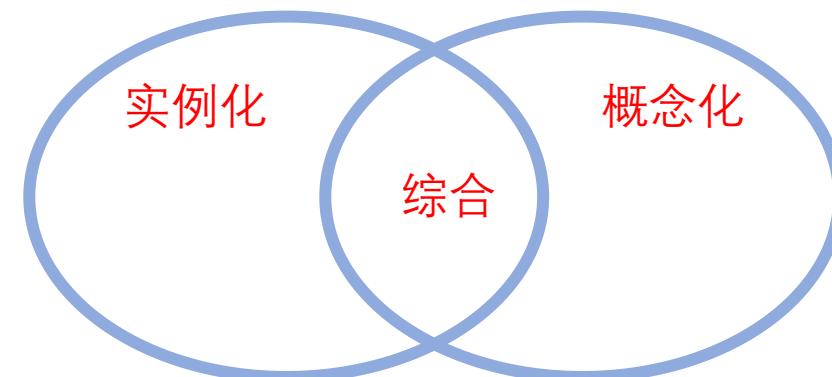


<http://kw.fudan.edu.cn/cnprobase/search/>

概念图谱	作 者	实 体	概 念	isA 关系数	准确度	权重
CN-Probase (中文)	复旦大学	15 066 667	270 025	32 925 306	95%	无

# 概念图谱的应用

- 可以归结为实例化和概念化这两个最基本的功能：
  - 实例化 (Instantiation)
    - 根据给定的概念，列出这个概念下的一些典型实体。
    - 如给出“Largest company”，返回“China Mobile”，“Google”等。
  - 概念化 (Conceptualization)
    - 给出一个或一组实体，推断出这些实体所属的概念。
    - 比如给出“Brazil”，“India”，“China”，返回“BRIC country”（金砖四国）、“Developing country”等概念，后者是更细化的概念
- 在这两个基本功能上，又分化出：
  - 基于实例化的应用
  - 基于概念化的应用
  - 综合使用实例化和概念化的应用



# 基于实例化的应用：实体搜索

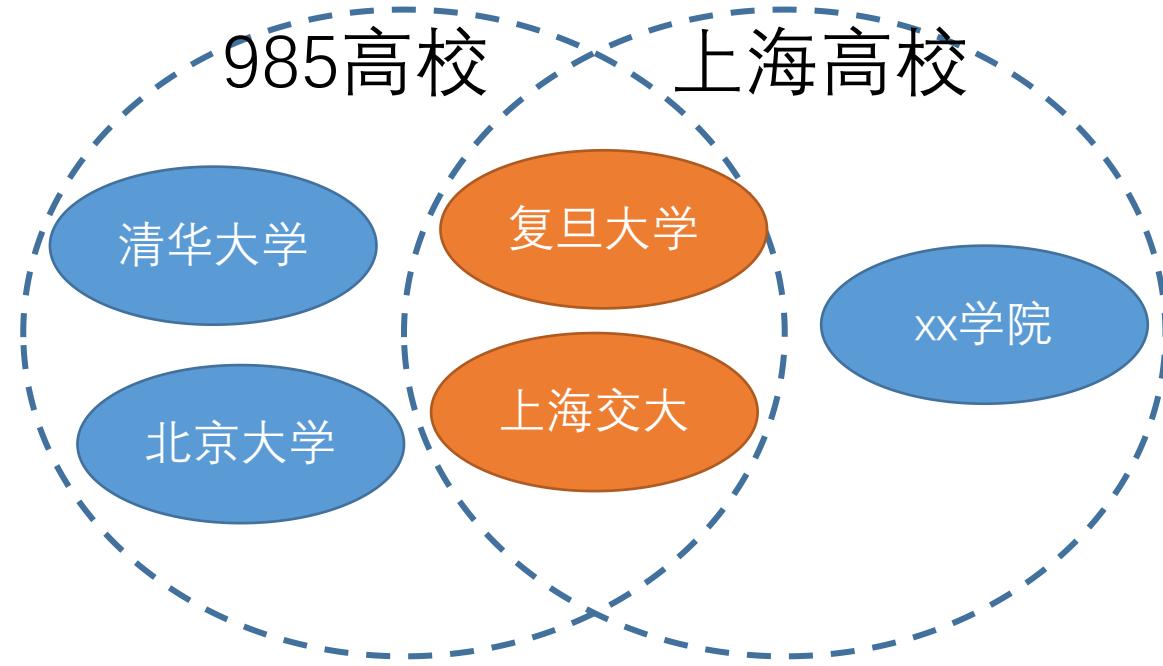
*Query:*水果

苹果
雪梨
西瓜
.....
蛇果
山竹
.....

} 更常见

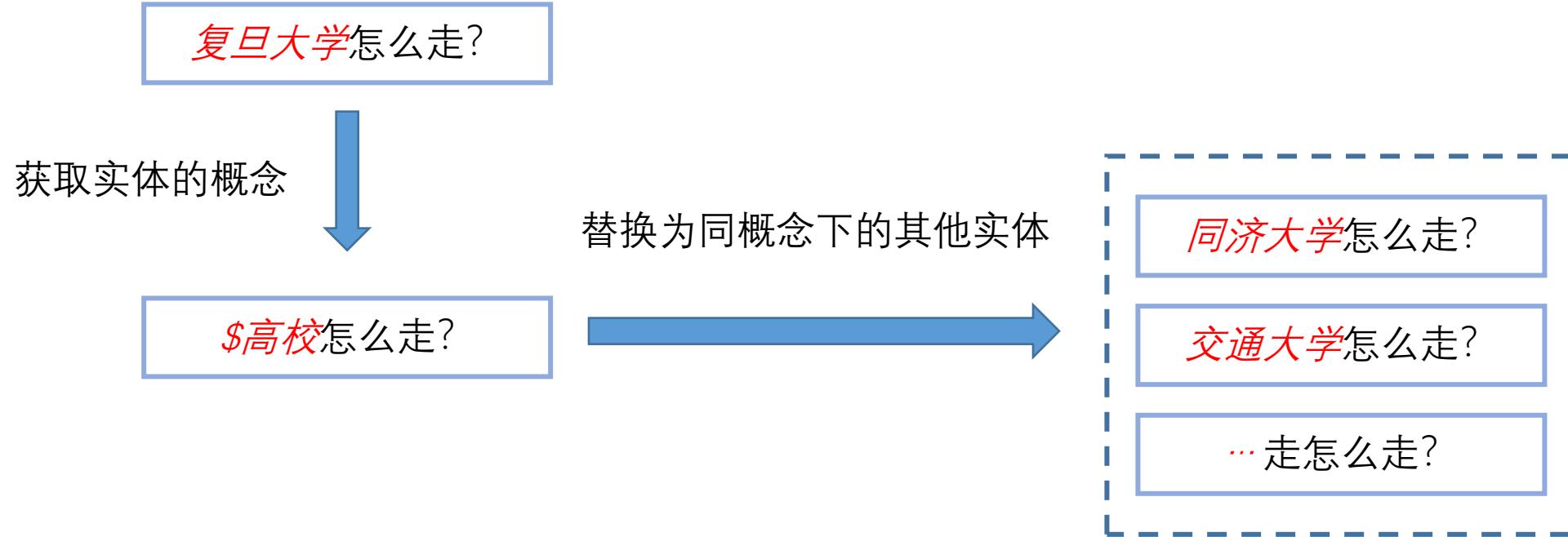
} 更稀有

*Query:*上海的985高校

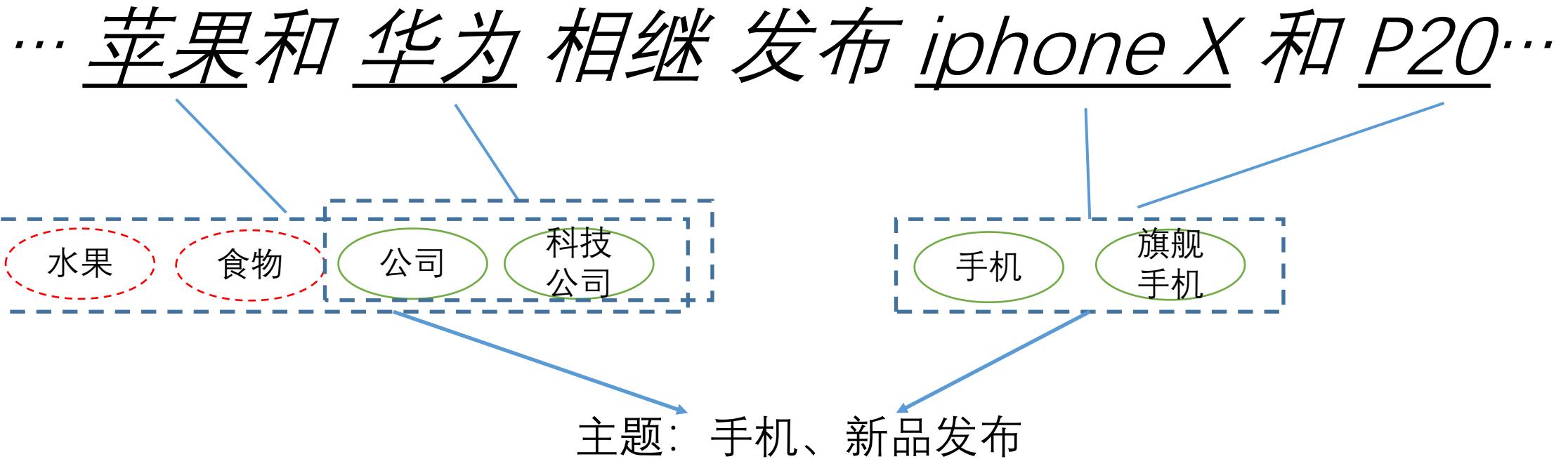


# 基于实例化的应用：样本增强

- 定义：
  - 利用概念下的实例，增强样本。

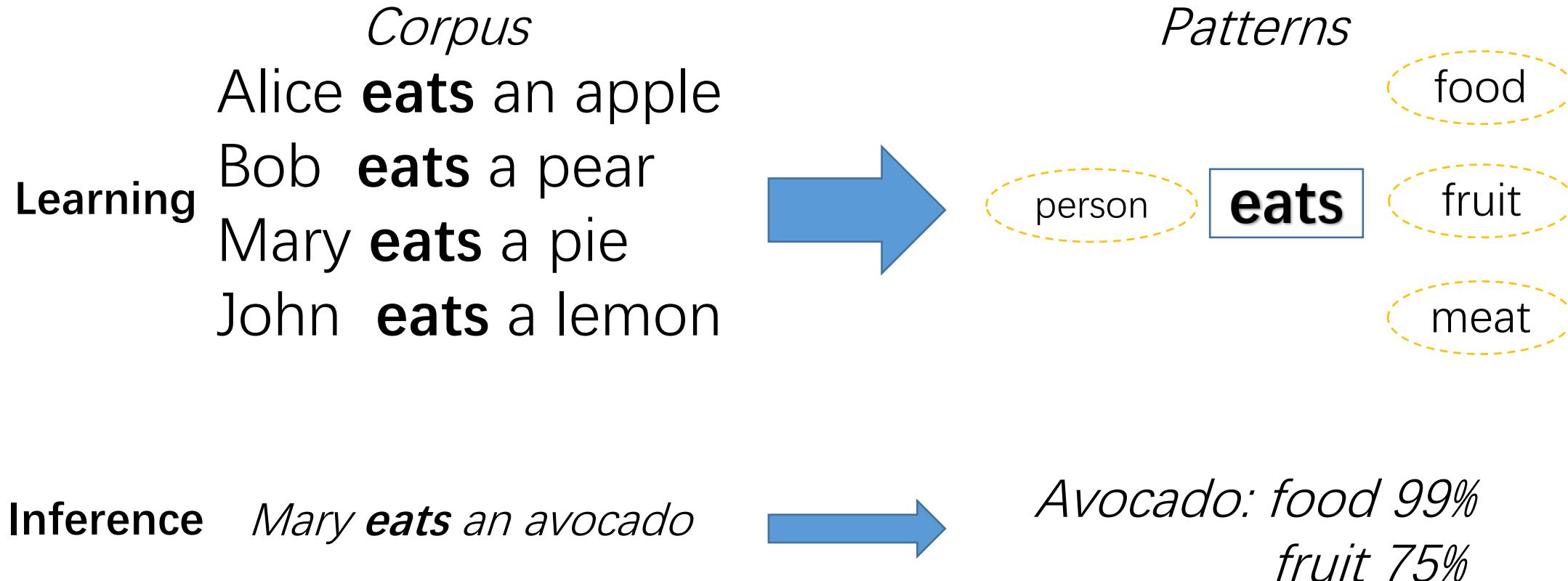


# 基于概念化的应用：主题理解



# 基于概念化的应用：语言概念模板

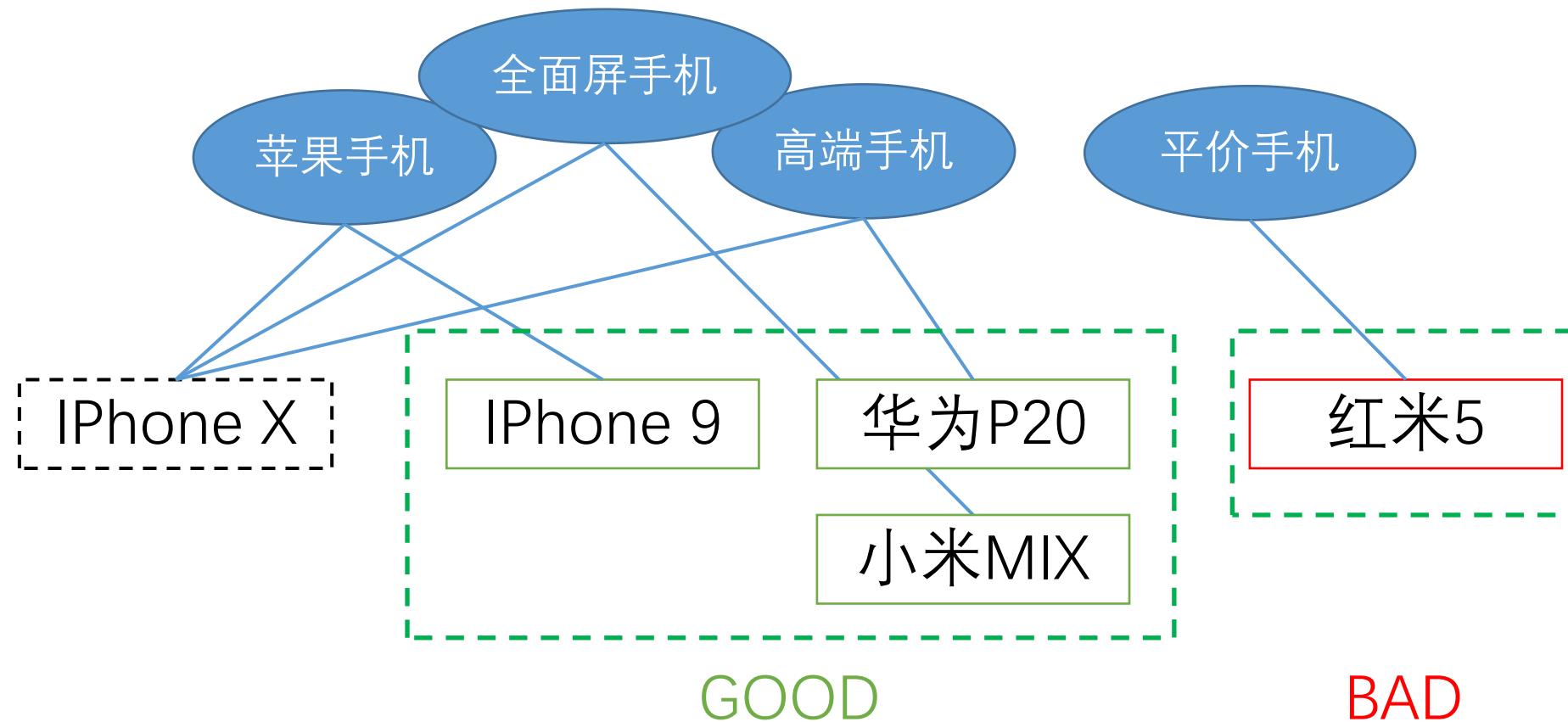
## • 语言概念模板



# 基于概念化的应用

文本分类	根据文本中实体的概念，将文本分为不同类别	包含“足球”“篮球”的文本应当与包含“宝马”“奔驰”的文本属于不同类别
主题分析	给定文本，分析文本属于什么主题	包含“足球”“篮球”的文本应当属于体育类主题
给文档打标签	给定文本，给文本标上若干个概念作为标签	对文本“iPhone进水了怎么办”可以打上“手机，维修”等标签
用户画像	给定用户信息，为用户生成显式的概念	根据用户描述“精通Java、Android开发”，可以为其打上“面向对象编程”“手机App开发”的标签
基于概念的解释	根据概念信息，为事件提供解释	特斯拉Model S的加速性能很好，因为它是“电动汽车”，电动汽车通常具有较好的加速性能
概念归纳	从实体集合或者词袋归纳概念标签	给定“清华大学”“复旦大学”“北京大学”，可以归纳出“中国高校”“985学校”等概念
语义表示	利用概念集合表达实体、词汇的语义	“iPhone X”的语义可表达为其概念集合{“全面屏手机”、“智能手机”、“旗舰手机”}

# 综合使用实例化和概念化的应用：实体推荐



# 隐式表示与显式表示

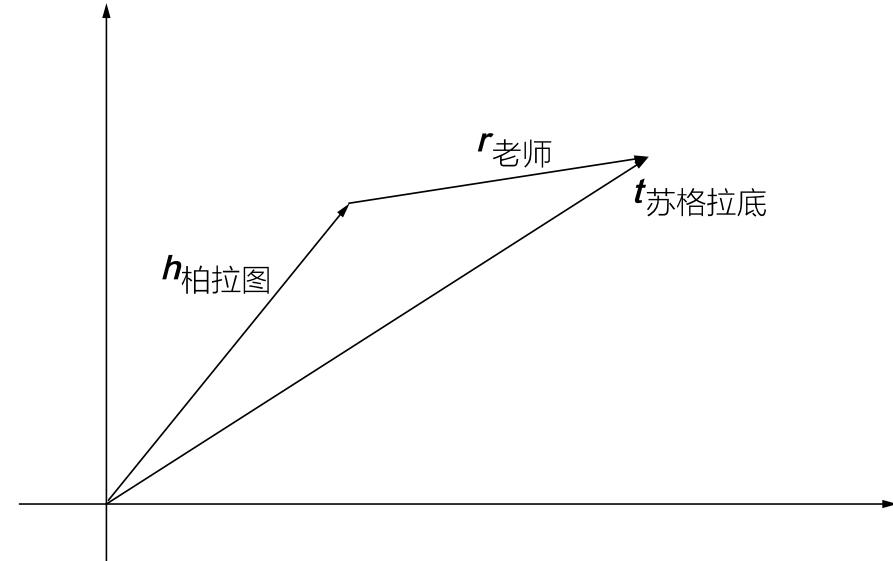
- 隐式语义表示：

- 目前深度学习中基于向量的隐式表示方式
- 向量之间的相似度量，在一定程度上反映向量对应的语义对象之间的关系

- 显式语义表示：

- 上述利用概念作为实体、词汇或者其他对象的语义表示属于显示的语义表示方法。

- 柏拉图：<哲学家、希腊人、男人、古代人>  
汉语入衣小六月丁工、丁肝件守几示。



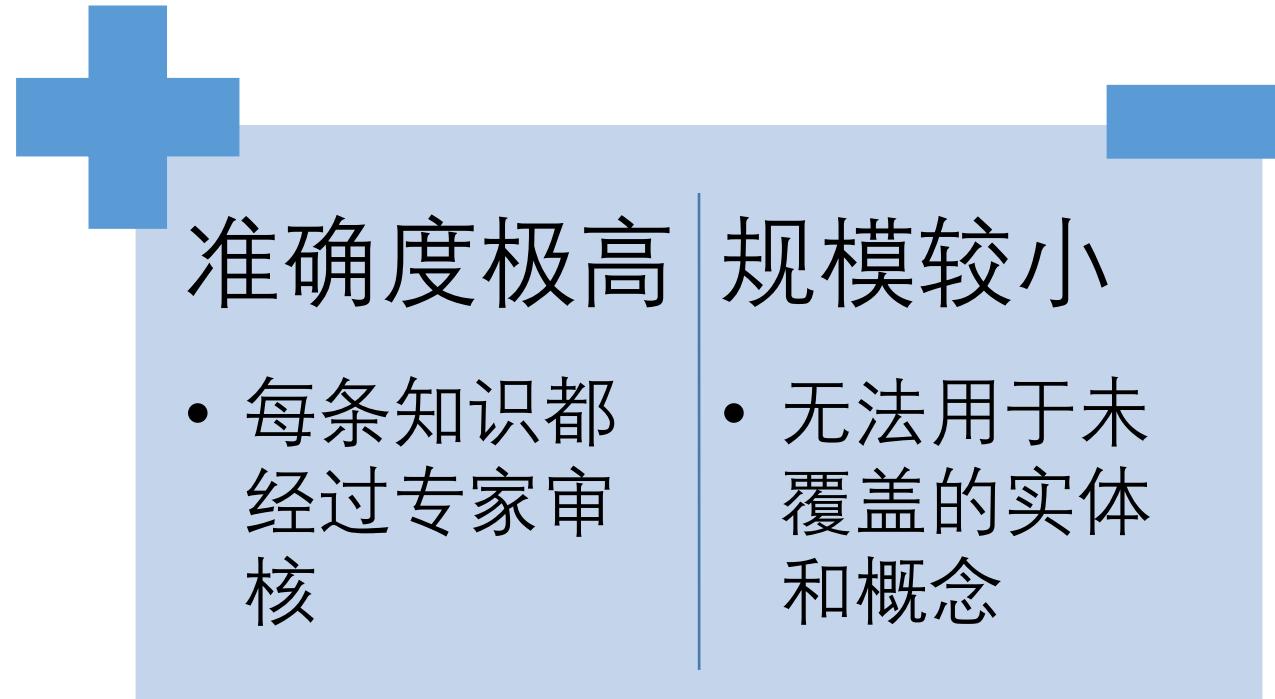
# 本章大纲

---

- 概念图谱构建
  - 概念图谱概述
  - 概念图谱抽取
- 百科图谱构建

# 为什么要自动抽取IsA关系

- 人工构建的概念图谱如WordNet等需要耗费大量人力



- 需要自动抽取IsA关系的方法

# IsA关系抽取：基本方法

## 基于Pattern的方法

- 具有高覆盖率的优点
- Probase包含千万级别的实体和概念，是目前最成功的英文分类体系。

## 基于Wikipedia的方法

- 具有高精度的特点
- 英文的YAGO 和中文的CN-Probase的准确率都在95%以上

## 基于Embedding的方法

- 基于Embedding的方法准确率较低(80%左右)
- 并没有被广泛用于概念图谱构建。

# IsA关系抽取：YAGO

- YAGO概念图谱是一个典型的基于 Wikipedia构建的英文概念图谱
  - 基于维基百科的类别系统构建
  - 包含36万isA关系，准确率在95%左右
- 构建方法
  - 概念型标签识别
  - 概念层级体系构建



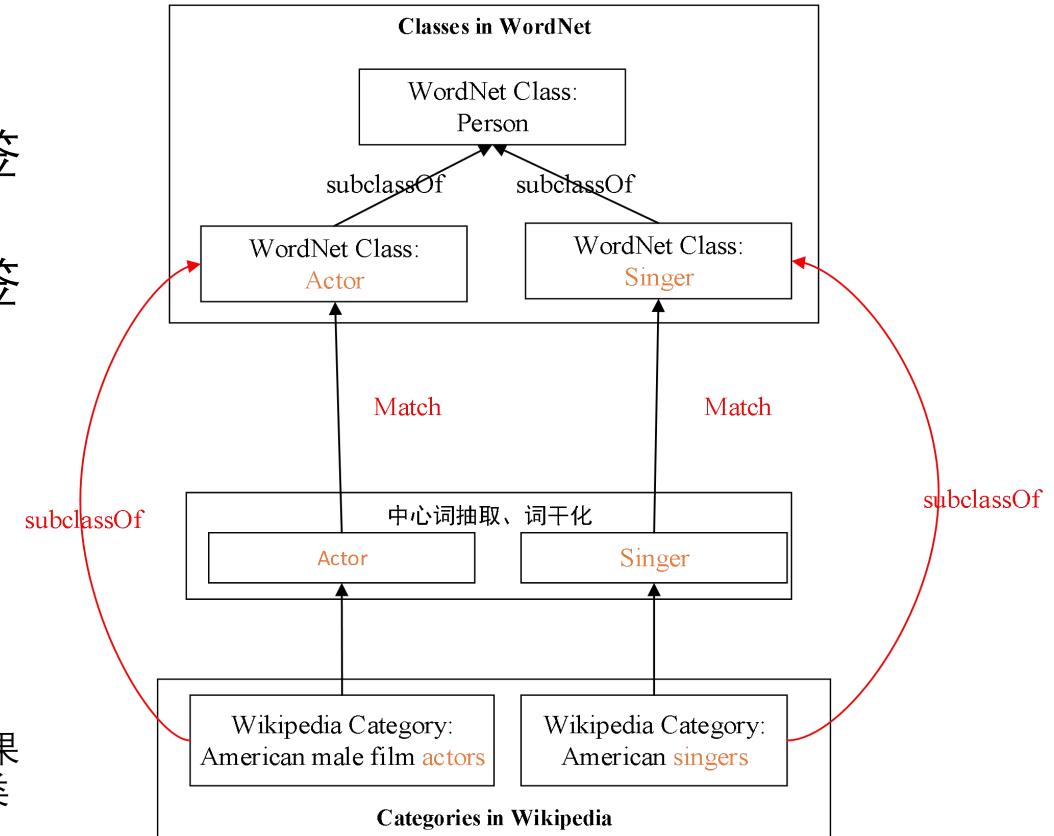
# IsA关系抽取：YAGO

- 概念型标签识别

- 维基百科中包括概念型标签、主题型标签、属性型标签以及管理型标签
- 通过人工或设定简单规则来剔除属性型标签以及管理型标签
- 使用单复数来区分概念型标签、主题型标签

- 概念层级体系构建

- 以WordNet作为基本Taxonomy
- 将更多来自Wikipedia的category加入Taxonomy中
  - 以subclassOf的关系加入，具体方法为：
    - 对Wikipedia的category提取其中心词，并词干化
    - 将处理后的结果与WordNet中结点进行匹配，如果匹配，则认为该category为WordNet中结点的子类



# IsA关系抽取：Hearst Patterns

- **Hearst Patterns:** 有一些固定的句型可以用于抽取IsA关系
  - 左图中列出了Hearst patterns的一部分，这里NP表示名词短语
  - 右图为一些符合Hearst pattern的例子

ID	Pattern
1	$NP \text{ such as } \{NP,\}^* \{(or   and)\} NP$
2	$\text{such } NP \text{ as } \{NP,\}^* \{(or   and)\} NP$
3	$NP\{,\} \text{ including } \{NP,\}^* \{(or   and)\} NP$
4	$NP\{NP\}^* \{,\} \text{ and other } NP$
5	$NP\{NP\}^* \{,\} \text{ or other } NP$
6	$NP\{,\} \text{ especially } \{NP,\}^* \{(or   and)\} NP$

- … animals **other than** dogs **such as** cats …
- … classic movies **such as** Gone with the Wind …
- … companies **such as** IBM, Nokia, Proctor and Gamble …
- … representatives in North America, Europe, the Middle East, Australia, Mexico, Brazil, Japan, China, **and other** countries …



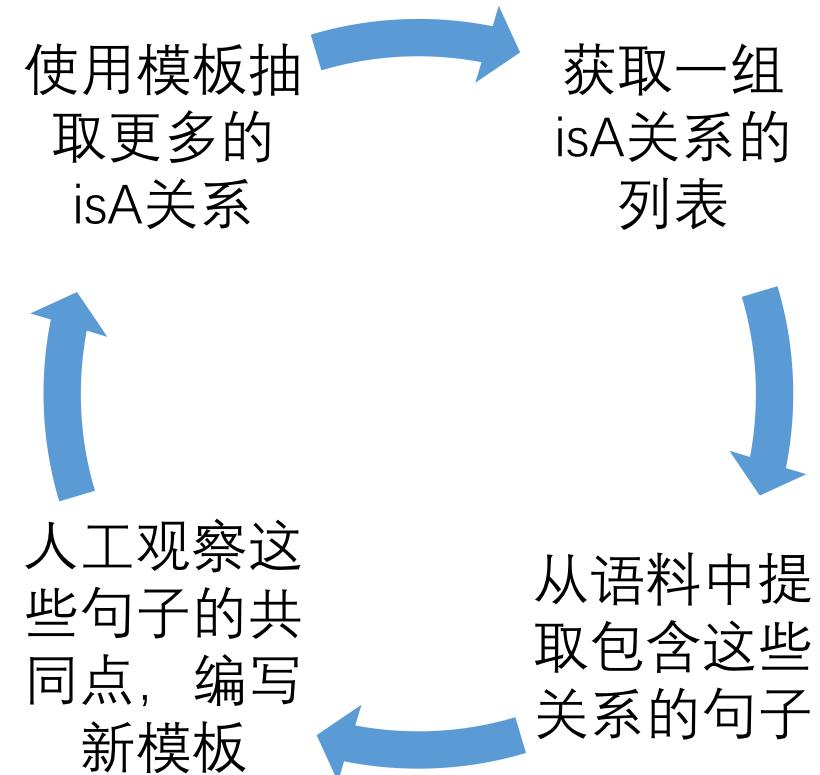
*cat isA animal*

*cat isA dog*

*Gone with the Wind isA classic movie*

# IsA关系抽取：Hearst Patterns

- Hearst Patterns中前3个由专家人手工编写
  - 其余的Hearst Pattern由一个半自动的Bootstrapping方法产生



# IsA关系抽取：Probase

- Probase是基于Pattern从大量英文语料中抽取的概念图谱
  - Step 1 使用Hearst Pattern抽取isA关系
  - Step 2 isA关系清洗

*… animals other than dogs **such as cats** …*

候选概念集合 $X=\{\text{animals}, \text{dogs}\}$ , 候选实体集合 $Y=\{\text{cats}\}$

只选择1个候选概念  
 $p(\text{animals}|\text{cats}) >> p(\text{dogs}|\text{cats})$

$\left\{ \begin{array}{ll} \text{cats isA animals?} & \text{GOOD} \\ \text{cats isA dogs?} & \text{BAD} \end{array} \right.$

# IsA关系抽取：中文概念图谱构建

## 基于模式

- 大部分中文模式比相应的英文模式准确率低

## 基于图谱翻译

- 译法存在歧义
- 不同语种倾向于表达不同的知识

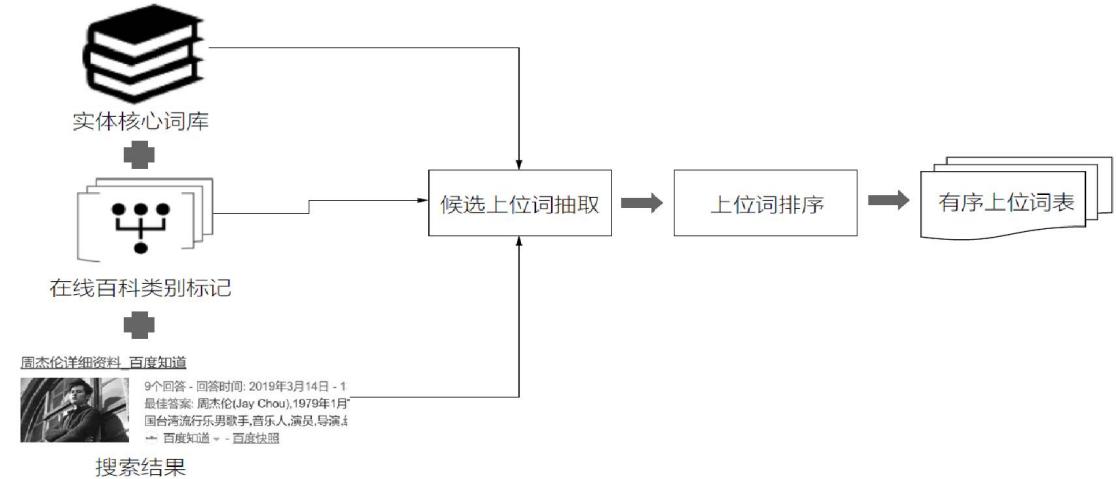
## 基于在线百科

- 覆盖率不高

英 文 模 式	准 确 率	中 文 模 式	准 确 率
NP is a NP	97.2%	NP 是一(个 种 ... 类) NP	95.7%
NP such as {NP,}*{(or   and)} NP	95.7%	例如 NP{、NP}*{	75.3%
such NP as {NP,}*{or   and} NP	96.6%	—	—
NP{,NP}*{,} and other NP	89.3%	NP {、NP}*等 NP	93.0%
NP{,NP}*{,} or other NP	90.7%	—	—
NP{,} including {NP,}*{(or   and)} NP	81.7%	NP 包括{NP、}*NP	80.4%
—	—	NP 是 NP	80.6%

# 中文概念图谱构建：大词林

- 大词林是一个基于**抽取+排序框架**构建的中文概念图谱
  - 输入：实体
  - 输出：有序上位词表
- 候选上位词抽取：
  - 通过搜索引擎搜索实体
  - 从搜索结果、在线百科类别标记和实体核心词库等三类来源获取候选上位词
- 上位词排序：
  - 使用大量命名实体及其候选上位词的标注语料训练排序模型
  - 解决Web得到的候选词召回率高，而准确率低的问题



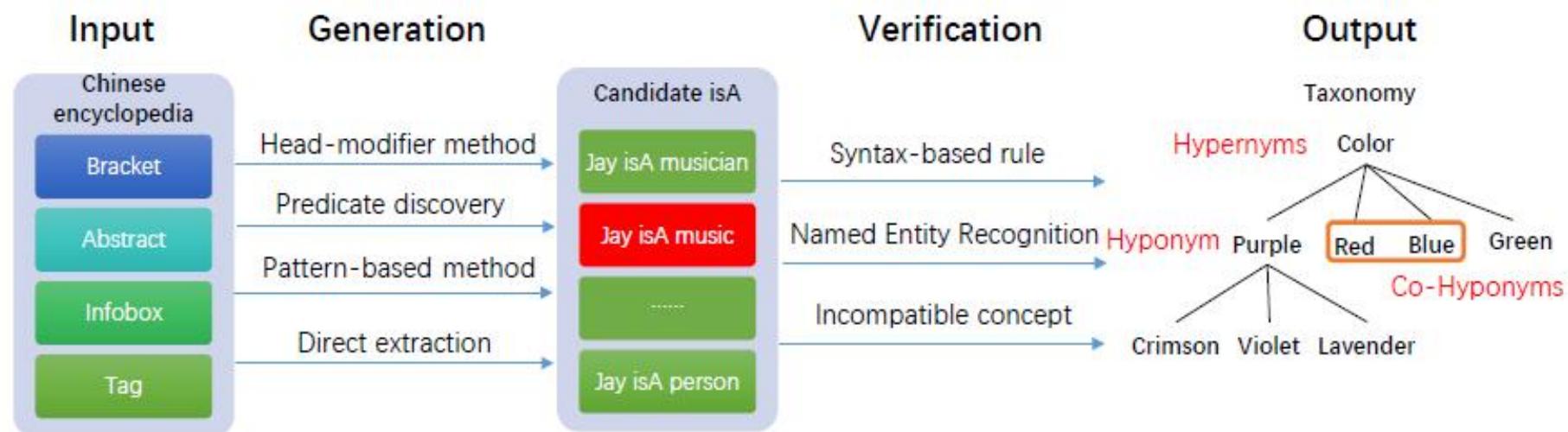
# 中文isA关系抽取：CN-Probase

- Hearst patterns在中文中效果不好
  - “NP such as {NP,}”， 英文：92%准确率， 中文：75%准确率



# 中文isA关系抽取：CN-Probase

- 生成和验证框架
  - 从多个数据源中抽取isA关系，确保覆盖率
  - 验证清洗抽取的结果，确保准确率



# 中文isA关系抽取：CN-Probase

实体括号

- 刘德华isA 歌手

摘要

- 刘德华 isA 制片人

Infobox

- 刘德华 isA 演员

标签

- 刘德华 isA 娱乐人物

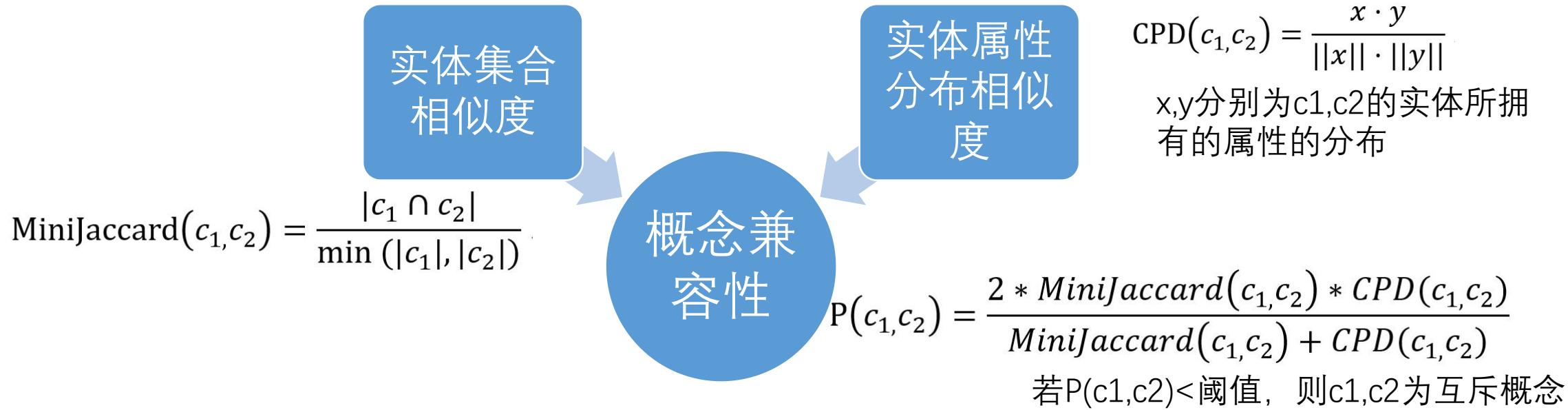
刘德华 (中国香港男演员、歌手、词作人)  (a)Entity with bracket  
Dehua Liu (Hong Kong actor, singer and songwriter)

刘德华 (Andy Lau), 1961 年 9 月 27 日出生于中国香港，男演员、歌手、作词人、制片人。1981 年出演电影处女作《彩云曲》。1983 年主演的武侠剧《神雕侠侣》在香港获得 62 点的收视纪录。1991 年创办天幕电影公司。1992 年，凭借传记片《五亿探长雷洛传》获得第 11 届香港电影金像奖最佳男主角提名。1994 年担任剧情片《天与地》的制片人。2000 年凭借警匪片《暗战》获得第 19 届香港电影金像奖最佳男主角奖。  (b)Abstract

(c)Infobox	中文名	Chinese name	刘德华 Dehua Liu
	职业	Occupation	演员 Actor
	代表作品	Representative works	忘情水 Forget Love Potion
	体重	Weight	63KG 63KG
(d)Tag	标签 Tag		人物 Person
	标签 Tag		演员 Actor
	标签 Tag		娱乐人物 Entertainer
	标签 Tag		音乐 Music

# 中文isA关系验证：CN-Probase

- 互斥的概念不能共存
  - 若发现实体同时存在互斥的概念
  - 只保留其中一个概念（属性分布之间的KL距离较小的一个）
- 互斥概念对发现



# 本章大纲

---

- 概念图谱构建
  - 概念图谱概述
  - 概念图谱抽取
- 百科图谱构建
  - 百科图谱概述
  - 百科图谱抽取

# 本章大纲

---

- 概念图谱构建
  - 概念图谱概述
  - 概念图谱抽取
- 百科图谱构建
  - 百科图谱概述
  - 百科图谱抽取

# 基本概念

- 百科

- “概要介绍人类一切门类知识或某一门类知识的工具书”

——摘自《中国大百科全书·新闻出版》

- 典型百科网站



# 百科网站的特点



# 百科图谱

- 定义

- 是一类以百科类网站作为数据源构建而成的知识图谱

- 区别

- 和纯文本页面不同，百科网站的页面中包含丰富的结构化的知识

新浪娱乐讯 8月18日下午17时，黄渤执导处女作《一出好戏》内地票房正式突破10亿元，成为2018年第12部票房“破十”的影片。该片由黄渤、王宝强、舒淇、张艺兴、于和伟、王迅等出演，自8月10日上映以来，票房口碑双丰收，目前在微博大V推荐度87%（87人评），大众评分8.6。

《一出好戏》于10日上映，首日虽被《爱情公寓》力压，屈居单日票房亚军位置，但隔日便火速实现逆袭，并蝉联日冠至今。8月17日，随着《欧洲攻略》《新乌龙院之笑闹江湖》《精灵旅社3：疯狂假期》《快把我哥带走》等新片的冲击，《一出好戏》仍以单日票房1.07亿的成绩守住冠军位，迫使有梁朝伟、吴亦凡、唐嫣、杜鹃等大牌加盟的《欧洲攻略》成为“老二”。上映次周票房仍然一路飘红，足见《一出好戏》良好口碑的加持有多么重要。（新娱/文）

## 一出好戏



中国 | 134分钟 | 2018年8月10日 (中国)



47

《一出好戏》是由上海瀚纳影视文化传媒有限公司制作的喜剧片，由黄渤执导，黄渤、王宝强、舒淇、张艺兴、于和伟、王迅联袂主演。该片于2018年8月10日在中国内地上映 [1] 。

该片讲述了公司员工团建出游遭遇海难，众人流落在荒岛之上，为了生存他们共同生活，并面对一系列人性问题的寓言故事。 [2]

中文名 一出好戏

主要演员



外文名 The Island

黄渤

舒淇

王宝强

张艺兴

其它译名 大富翁、狂想曲

出品公司 上海瀚纳影视文化传媒有限公司

# 百科图谱的意义

---

- 1. 支撑领域知识图谱的构建
- 2. 为机器语言理解提供通用知识
- 3. 支撑语料自动标注

# 百科图谱分类

根据百科的定义，百科图谱可分为通用百科图谱和领域百科图谱

## • 通用百科图谱

- 来自于通用百科网站：概要介绍人类一切门类知识
- E.g.,
  - 维基百科，百度百科

This screenshot shows the English Wikipedia page for Donald Trump. It includes a large portrait of him, a summary of his life, and sections on his political career as the 45th President of the United States and his business ventures. The page also lists his family, education, and other personal details.

This screenshot shows the Baidu Baike page for Donald Trump. It features a large portrait of him, a brief summary, and sections on his political career as the 45th President of the United States. The page also includes links to related topics like the 2016 US election and his business ventures.

## • 领域百科图谱

- 来自于领域百科网站：概要介绍人类某一门类知识
- E.g.,
  - 电影网站，购物网站

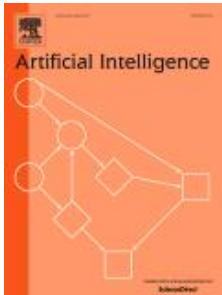
This screenshot shows the IMDB page for the movie "Beauty and the Beast" (2017). It displays the movie's title, rating (7.9), and release date (March 17, 2017). Below the title are images of the cast and crew, and a trailer video.

This screenshot shows the JD.com product page for the book "Artificial Intelligence: A Modern Approach" by Stuart Russell and Peter Norvig. It displays the book's cover, price (\$105.70), and various purchasing options and reviews.

# 百科知识图谱构建分类

对单百科数据源深入挖掘

- DBpedia
- YAGO
- CN-DBpedia



2024-3-7

对多百科数据源进行融合

- BabelNet
- Zhishi.me
- XLORE

AIJ 2017 PROMINENT PAPER AWARD

YAGO2 [Johannes Hoffart et. al., 2013]

BabelNet [Roberto Navigli et. al., 2012]

<http://aij.ijcai.org/index.php/aij-awards-list-of-previous-winners> 42

# 本章大纲

---

- 概念图谱构建
  - 概念图谱概述
  - 概念图谱抽取
- 百科图谱构建
  - 百科图谱概述
  - **百科图谱抽取**

# 基于单源的百科图谱构建

- 问题定义

- 输入：一个百科网站的所有页面
- 输出：一个百科知识图谱



- 步骤



CN-DBpedia



# 数据获取

---

- 目标
  - 找到一个百科类网站所有实体的介绍页面
- 步骤
  - 页面获取
  - 页面识别

# 页面获取

- 目标
  - 获取一个百科数据源中全部网页
- 策略
  - 基于Dump数据的下载
    - Wikipedia Dump

适用场景：网站全部数据都以Dump的形式提供下载

现实挑战：大多数网站不提供Dump下载

- 2018-08-11 08:30:12 [enwiki](#): Dump complete
- 2018-08-08 21:24:08 [dewiki](#): Dump complete
- 2018-08-07 17:23:44 [frwiki](#): Dump complete
- 2018-08-07 15:19:36 [ruwiki](#): Dump complete
- 2018-08-07 11:52:12 [zhwiki](#): Dump complete
- 2018-08-07 09:54:55 [commonswiki](#): Dump complete
- 2018-08-07 06:50:40 [ptwiki](#): Dump complete
- 2018-08-06 17:36:31 [plwiki](#): Dump complete
- 2018-08-06 15:16:09 [nlwiki](#): Dump complete
- 2018-08-06 13:10:34 [itwiki](#): Dump complete
- 2018-08-06 04:37:18 [ukwiki](#): Dump complete
- 2018-08-06 02:02:03 [arwiki](#): Dump complete

```
<page>
<title>数学</title>
<ns>0</ns>
<id>13</id>
<revision>
<id>35788162</id>
<parentid>35786218</parentid>
<timestamp>2015-05-21T05:27:52Z</timestamp>
<contributors>
<username>老陳</username>
<id>331249</id>
</contributors>
<model>wikitext</model>
<format>text/x-wiki</format>
<text xml:space="preserve">{{NoteTA|G1=Science}}<br><img alt="Euclid's Elements, Book 1, Proposition 1, showing a geometric proof." data-bbox="100px 100px 200px 200px"/><br>「数学」( Mathematics )是利用符号语言研究[數量] &lt;ref name="OED">{{cite web url="http://oed.com/view/Entry/114974 |title=i
從數學的知識與運用總是個人與團體生活中不可或缺的一環。對數學基本概念的完善，早在[[古埃及]]、[[美索不達米亞|美索不達米亞]]及[[印度]]古印度
今日，數學使用在不同的領域中，包括[[科學|科學]]、[[工程学|工程]]、[[医学|醫學]]和[[经济学|經濟學]]等。數學對這些領域的應用通常被稱為[[应用数学|應用
:= 詞源 ==<br>印歐語系[西方語言]中“數學” ( {{lang-en|mathematics}} ) 一詞源自於[[古希臘語]]的{{lang-el|μαθηματικά}} ( {{lang-la|máthēma}} )
「數學」一詞的大約產生於宋朝末年元朝元時，多指參數之學，但有時也含有今天的數學意義，例如，秦九韶的《數學九章》 ( 《水經大典》
author = 鄭繼春 )
```

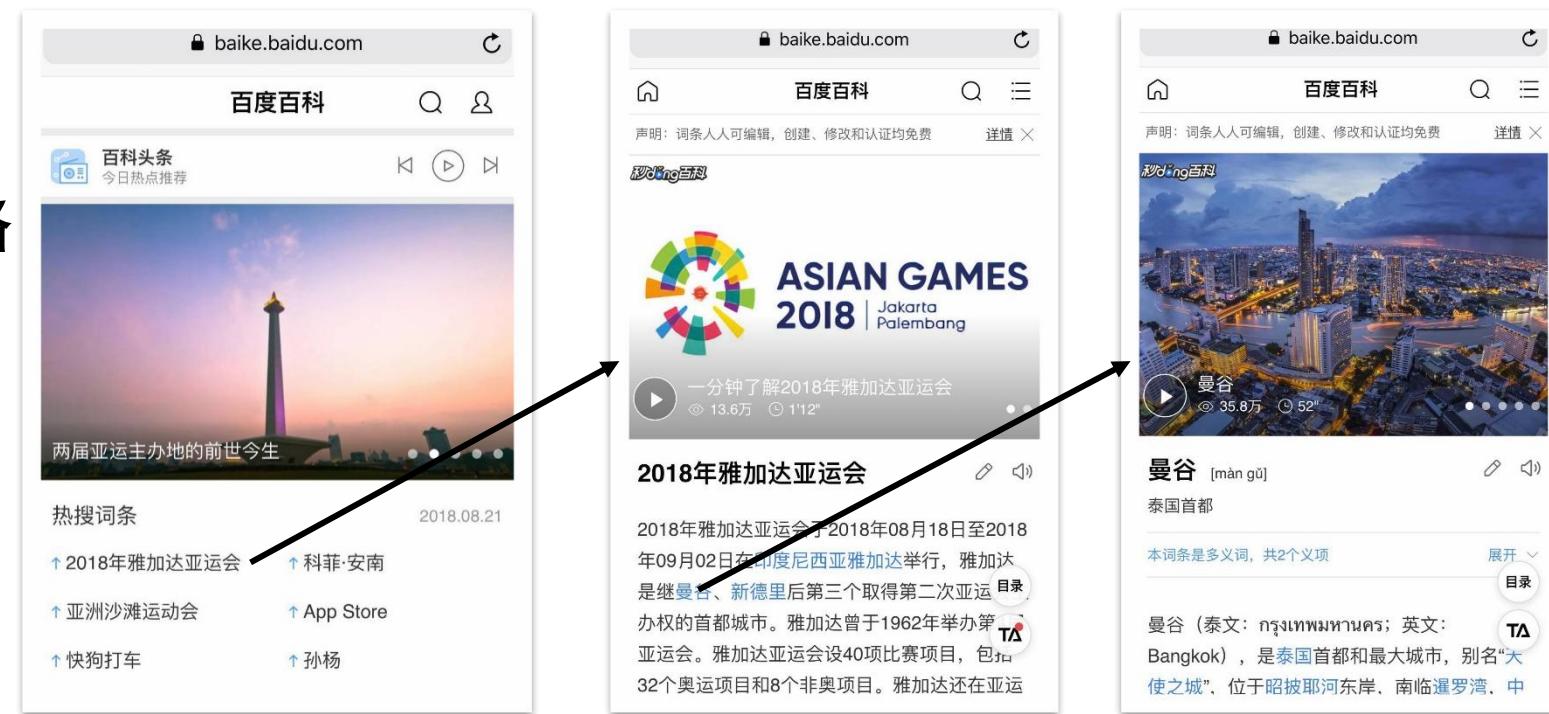
Wikipedia Dump <https://dumps.wikimedia.org/>

# 页面获取

- 目标
  - 获取一个百科数据源中所有网页
- 策略
  - 基于Dump数据的下载
    - Wikipedia Dump
  - 基于超链接的遍历策略
    - BFS / DFS

适用场景：百科数据源中所有网页都通过超链接联接

现实挑战：部分百科页面未被其他页面链接，导致无法获取



# 页面获取

- 目标
  - 获取一个百科数据源中所有网页
- 策略
  - 基于Dump数据的下载
    - Wikipedia Dump
  - 基于超链接的遍历策略
    - BFS / DFS
  - **基于枚举的遍历策略**
    - ID / 名称 / 哈希

ID	NAME
http://baike.baidu.com/view/[ID].htm	http://baike.baidu.com/item/[NAME]
http://baike.baidu.com/view/1.htm	http://baike.baidu.com/item/ <b>周杰伦</b>
http://baike.baidu.com/view/2.htm	http://baike.baidu.com/item/ <b>复旦大学</b>
http://baike.baidu.com/view/3.htm	http://baike.baidu.com/item/ <b>一出好戏</b>
http://baike.baidu.com/view/4.htm	http://baike.baidu.com/item/ <b>黄渤</b>

适用场景：百科网站的页面  
URL具有可枚举性

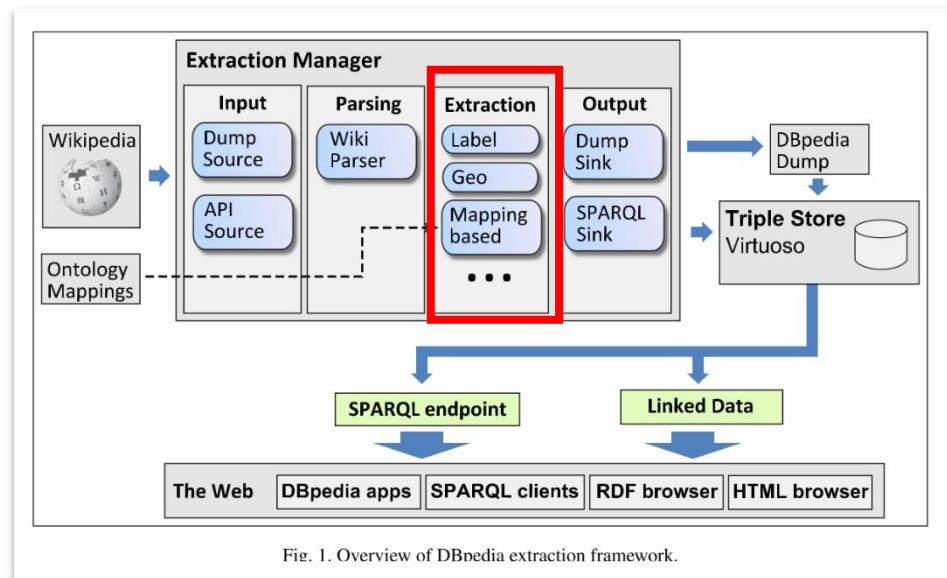
# 页面识别

- 目标
  - 筛选出所有介绍实体的网页
- 百科页面的特殊性
  - 每个页面均围绕一个词条进行全方面的介绍
- 方法
  - 每个词条名作为知识图谱中的一个实体
  - 实体发现过程等价于词条页面发现
  - 百科网站的词条页面URL具有一定的规律性
    - [http://baike.baidu.com/view/\[ID\].htm](http://baike.baidu.com/view/[ID].htm)
    - [http://baike.baidu.com/item/\[NAME\]](http://baike.baidu.com/item/[NAME])
    - [https://music.163.com/#/song?id=\[ID\]](https://music.163.com/#/song?id=[ID])
    - [https://movie.douban.com/subject/\[ID\]](https://movie.douban.com/subject/[ID])



# 属性抽取

- 针对百科页面的知识抽取
  - 本质上是针对其中的半结构化数据进行抽取
  - 对于每个实体页面，使用不同抽取器来抽取不同类型关系



A 刘德华是一个多义词，请在下列义项上选择浏览（共10个义项） 收起 ^ 添加义项 +  
· 中国香港男演员、歌手、制片人、填词人 · 原民航局空中交通管理局局长助理 · 清华大学教授  
· 江西弋阳籍烈士 · 国家税务总局广安开发区税务局副局长 · 新疆青少年出版社出版的著作

B 刘德华 编辑 讨论 99+  
1961年9月27日 | 香港新界大埔镇泰亨村 | 中国  
同义词 华仔一般指刘德华（中国香港男演员、歌手、制片人、填词人）

C 刘德华 (Andy Lau), 1961年9月27日出生于中国香港，籍贯广东新会 [1]，中国香港男演员、歌手、作词人、制片人。

D 1994年创立刘德华慈善基金会 [21]。2000年被评为世界十大杰出青年 [22]。2005年发起亚洲新星导计划 [23]。2008年被委任为香港非官守太平绅士 [24]。2016年连任中国残疾人福利基金会副理事长。 [22]

E 目录 1 早年经历 4 主要作品 5 社会活动 6 获奖记录  
2 演艺经历 6 公益事业  
3 个人生活 7 奥运活动  
8 人物评价  
9 人物事件

F 基本信息  
中文名 刘德华 代表作品 无间道、天若有情、旺角卡门、桃姐、天下无贼、忘情水、谢谢你的爱、爱你一万年、冰雨、今天  
外文名 Andy Lau, Lau Tak Wah  
别 名 华仔, 华Dee, 华哥等 妻 子 朱丽倩  
出生地 香港新界大埔镇泰亨村 女 儿 刘向蕙  
出生日期 1961年9月27日 信 仰 佛教  
职 业 演员, 歌手, 填词人, 制片人 生 肖 牛

G 演艺经历 编辑  
港剧时代  
1983年，主演金庸武侠剧《神雕侠侣》，在剧中饰演外貌俊俏、倜傥不羁的杨过 [33]；该剧在香港播出后取得62点的收视纪录；同年，与黄日华、梁朝伟、苗侨伟、汤镇业组成“无线五虎将” [34]。

H 词条标签：音乐人物，演员，歌手，娱乐人物，制作人，人物

# 名称属性抽取

## • 不存在多义词

- 《实体名》 = 《页面标题》



一出好戏 编辑

中国 | 134分钟 | 2018年8月10日 (中国)

《一出好戏》是由上海瀚纳影视文化传媒有限公司制作的喜剧片，由黄渤执导，黄渤、王宝强、舒淇、张艺兴、于和伟、王迅联袂主演。该片于2018年8月10日在内地上映<sup>[1]</sup>。

该片讲述了公司员工团建出游遭遇海难，众人流落在荒岛之上，为了生存他们共同生活，并面对一系列人性问题的寓言故事。

中文名	一出好戏
外文名	The Island
其它译名	大富翁、狂想曲
出品公司	上海瀚纳影视文化传媒有限公司

主要演员

黄渤 舒淇 王宝强 张艺兴

## • 存在多义词

- 《实体名》 = 《页面标题》 + 《歧义项》



刘德华 (中国香港男演员、歌手、制片人、填词人) 编辑

刘德华是一个多义词，请在下列义项上选择浏览 (共10个义项) 收起

- 中国香港男演员、歌手、制片人、填词人
- 原民航局空中交通管理局局长助理
- 清华大学教授
- 江西弋阳籍烈士
- 四川省广安经济技术开发区国家税务局副局长
- 新疆青少年出版社出版的著作
- 山东钢铁集团有限公司财务总监
- 湖北监利籍烈士
- 通川区学生资助中心主任
- 湖北籍烈士

刘德华 编辑

1961年9月27日 | 香港新界大埔镇泰亨村 | 中国

刘德华 (Andy Lau)，1961年9月27日出生于中国香港，中国香港男演员、歌手、作词人、制片人。

1981年出演电影处女作《彩云曲》<sup>[1]</sup>。1983年主演的武侠剧《神雕侠侣》在香港获得62点的收视纪录<sup>[2-3]</sup>。1991年创办天幕电影公司<sup>[4]</sup>。1992年，凭借传记片《五亿探长雷洛传》获得第11届香港电影金像奖最佳男主角提名<sup>[5]</sup>。1994年担任剧情片《天与地》的制片人<sup>[6]</sup>。2000年凭借警匪片《暗战》获得第19届香港电影金像奖最佳男主角奖<sup>[7]</sup>。2004年凭借警匪片《无间道3：终极无间》获得第41届台湾金马奖最佳男主角奖<sup>[8]</sup>。2005年获得香港UUA学院颁发的全港最高累积票房香港男演员‘奖’<sup>[9]</sup>。2006年获得釜山国际电影节亚洲最有贡献电影人奖<sup>[10]</sup>。2011年主演剧情片《桃姐》，并凭借该片先后获得台湾金马奖最佳男主角奖、香港电影金像奖最佳男主角奖<sup>[11]</sup>；同年担任第49届台湾电影金马奖评审团主席<sup>[12]</sup>。2017年主演警匪动作片《拆弹专家》<sup>[13]</sup>。

1985年发行首张个人专辑《只知道此刻爱你》<sup>[14]</sup>。1990年凭借专辑《可不可以》在歌坛获得关注<sup>[15]</sup>。1994年获得十大劲歌金曲最受欢迎男歌星奖<sup>[16]</sup>。1995年在央视春晚上演唱歌曲《忘情水》<sup>[17]</sup>。2000年被《吉尼斯世界纪录大全》评为“获奖最多的香港男歌手”<sup>[18]</sup>。2004年第六次获得十大劲歌金曲最受欢迎男歌星奖。2016年参与填词的歌曲《原谅我》正式发行<sup>[19]</sup>

# 实体指代抽取

- Mention: 文本中出现的一个命名实体
- Entity: 知识图谱中的一个实体
- Mention2Entity关系
  - 将文本中的一个命名实体和知识图谱中的一个实体对应起来
- 应用
  - 实体链接

刘德华



1961年9月27日 | 香港新界大埔镇泰亨村 | 中国

同义词 华仔一般指刘德华（中国香港男演员、歌手、制片人、填词人）

刘德华是一个多义词，请在下列义项上选择浏览（共10个义项）

- 中国香港男演员、歌手、制片人、填词人
- 江西弋阳籍烈士
- 山东钢铁集团有限公司财务总监
- 湖北郧西籍烈士

别名

华仔，华Dee，华哥等

# 摘要属性抽取

- 摘要
  - 一段概括实体的文本
- 应用
  - 实体展示
  - 相似度计算
  - Embedding

刘德华 编辑

1961年9月27日 | 香港新界大埔镇泰亨村 | 中国

同义词 华仔一般指刘德华（中国香港男演员、歌手、制片人、填词人）

刘德华 (Andy Lau) , 1961年9月27日出生于中国香港，中国香港男演员、歌手、作词人、制片人。1981年出演电影处女作《彩云曲》<sup>[1]</sup>。1983年主演的武侠剧《神雕侠侣》在香港获得62点的收视纪录<sup>[2-3]</sup>。1991年创办天幕电影公司<sup>[4]</sup>。1992年，凭借传记片《五亿探长雷洛传》获得第11届香港电影金像奖最佳男主角提名<sup>[5]</sup>。1994年担任剧情片《天与地》的制片人<sup>[6]</sup>。2000年凭借警匪片《暗战》获得第19届香港电影金像奖最佳男主角奖<sup>[7]</sup>。2004年凭借警匪片《无间道3：终极无间》获得第41届台湾金马奖最佳男主角奖<sup>[8]</sup>。2005年获得香港UA院线颁发的全港最高累积票房香港男演员”奖<sup>[9]</sup>。2006年获得釜山国际电影节亚洲最有贡献电影人奖<sup>[10]</sup>。2011年主演剧情片《桃姐》，并凭借该片先后获得台湾金马奖最佳男主角奖、香港电影金像奖最佳男主角奖<sup>[11]</sup>；同年担任第49届台湾电影金马奖评审团主席<sup>[12]</sup>。2017年主演警匪动作片《拆弹专家》<sup>[13]</sup>。

1985年发行首张个人专辑《只知道此刻爱你》<sup>[14]</sup>。1990年凭借专辑《可不可以》在歌坛获得关注<sup>[15]</sup>。1994年获得十大劲歌金曲最受欢迎男歌星奖<sup>[16]</sup>。1995年在央视春晚上演唱歌曲《忘情水》<sup>[17]</sup>。2000年被《吉尼斯世界纪录大全》评为“获奖最多的香港男歌手”<sup>[18]</sup>。2004年第六次获得十大劲歌金曲最受欢迎男歌星奖。2016年参与填词的歌曲《原谅我》正式发行<sup>[19]</sup>。

1994年创立刘德华慈善基金会<sup>[20]</sup>。2000年被评为世界十大杰出青年<sup>[21]</sup>。2005年发起亚洲新星导计划<sup>[22]</sup>。2008年被委任为香港非官守太平绅士<sup>[23]</sup>。2016年连任中国残疾人福利基金会副理事长。<sup>[24]</sup>

# 基本属性抽取

- 基本属性/Infobox
  - 对实体的结构化总结
  - 以表格的形式展示
    - 第一列表示属性
    - 第二列表示属性值

基本信息			
中文名	刘德华	经纪公司	东亚唱片、映艺娱乐
外文名	Andy Lau, Lau Tak Wah	代表作品	暗战、无间道、天若有情、旺角卡门、桃姐、来生缘、忘情水、谢谢你的爱、冰雨、今天、爱你一万年
别 名	华仔，华Dee，华哥等	主要成就	三届香港电影金像奖最佳男主角 两届台湾电影金马奖最佳男主角 1985-2005年全港最高累积票房香港男演员奖
国 籍	中国		中国电影百年形象大使
民 族	汉族		釜山电影节亚洲最有贡献电影人奖
星 座	天秤座		~ 展开
血 型	AB型		
身 高	174cm		
体 重	63kg	妻 子	朱丽倩
出生地	香港新界大埔镇泰亨村	女 儿	刘向蕙
出生日期	1961年9月27日	全球粉丝会	华仔天地
职 业	演员，歌手，填词人，制片人	信 仰	佛教
毕业院校	可立中学；第十期无线艺员训练班	生 肖	牛

是百科知识图谱**最重要的**知识来源之一  
从数量上来说，它是能提供**最多知识**的一类关系

# 相关关系/分类关系抽取

- 相关关系

- 以超链接的形式展示与实体相关的其他实体

- 分类关系

- 对实体进行分类
- 标签来自于用户众包

1982年，刘德华以甲级成绩从艺员训练班毕业后正式签约TVB [34]；同年在喜剧《花艇小英雄》中饰演败家仔钱日添；12月，与叶德娴搭档主演时装警匪剧《猎鹰》，凭借卧底警察江大伟一角获得关注 [35]。

1983年，主演金庸武侠剧《神雕侠侣》，在剧中饰演外貌俊俏、倜傥不羁的杨过 [36]；该剧在香港播出后取得62点的收视纪录；同年，与黄日华、梁朝伟、苗侨伟、汤镇业组成“无线五虎将” [37]。

1984年，与赵雅芝合作主演古装武侠剧《魔域桃源》，在剧中饰演资质出众、武功高强的傅青云 [38]；同年，与梁朝伟共同主演金庸武侠剧《鹿鼎记》，在剧中饰演英明果断的康熙 [39]。

1985年，在古装武侠剧《杨家将》中饰演饶勇善战的杨六郎 [40]；同年，TVB向刘德华提出加签五年的合约，刘德华因拒绝而被TVB雪藏400天 [41-42]。1986年，在邵逸夫的调解下，刘德华与TVB和解并签下合约；同年，主演古装剧《真命天子》。1988年，在出演了武侠剧《天狼劫》后，刘德华将演艺事业的重心转向影坛 [42]。

词条标签： 音乐人物， 演员， 歌手， 娱乐人物， 制作人， 人物

# 基于正则表达式的抽取

e.g., 基本属性抽取器

**属性抽取正则表达式** : <dd class="basicInfo-item name">(.\*)</dd>

**属性值抽取正则表达式**: <dd class="basicInfo-item value">(.\*)</dd>

The screenshot shows a Baidu百科 page for Fudan University. The left side displays basic information in a table format:

中文名	复旦大学
英文名	Fudan University
简 称	复旦·FUDAN
创办时间	1905年(乙巳年)9月14日
类 别	公立大学
学校类型	综合类
属 性	985工程(1999年) 211工程(1994年) 九校联盟(2009年) 珠峰计划(2009年) 111计划(2006年)
所属地区	中国·上海
现任校长	许宁生
知名校友	李岚清、朱民、李源潮、竺可桢、于右任、邵力子、王沪宁等

The right side shows the developer tools Elements tab with the DOM tree expanded to show the structure of the page. The 'basicInfo-item.value' class is highlighted in several `<dd>` elements, such as those for '博士后流动站' (Postdoctoral流动站), '校训' (Motto), and '国家重点学科' (Key Disciplines).

# 数据清洗

- 数据质量问题

属性表述不一致

数值属性值格式不统一

多个对象属性值未分割

- 清洗后结果

基本信息	
中文名	复旦大学
英文名	Fudan University
简称	复旦 FUDAN
创办时间	1905年（乙巳年）9月14日
类别	公立大学
学校类型	综合
属性	985工程（1999年） 211工程（1994年） 九校联盟（2009年） 珠峰计划（2009年） 111计划（2006年）
所属地区	中国 上海
现任校长	许宁生
知名校友	李岚清、朱民、李源潮、竺可桢、于右任、邵力子、王沪宁等
主管部门	
硕士点	243 个
博士点	154 个
博士后流动站	35 个
校训	博学而笃志，切问而近思
校歌	《复旦大学校歌》
专职院士	中国科学院院士 21 人 中国工程院院士 5 人
主要院系	中国语言文学系、哲学学院、历史学系、旅游学系、文物和博物馆学系、外国语言文学学院等
国家重点学科	一级学科 11 个，二级学科 19 个
学校地址	上海市杨浦区邯郸路 220 号
学校代码	10246
主要奖项	全国优秀博士论文 55 篇（截至 2012 年）
校庆日	5 月 27 日（上海解放纪念日）

InfoBox

中文名	复旦大学
创办时间	1905年09月14日
知名校友	于右任
知名校友	朱民
知名校友	李岚清
知名校友	李源潮
知名校友	王沪宁
知名校友	竺可桢
知名校友	邵力子
英文名称	Fudan University

# 属性对齐：生成+过滤+验证

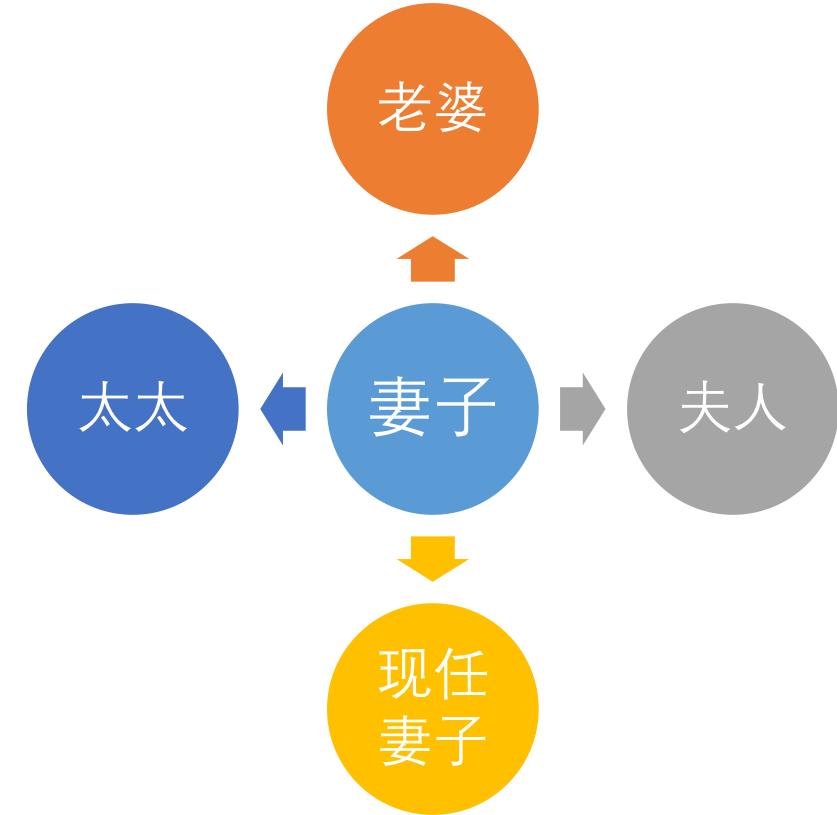
## ① 生成：找到候选等价属性对

- 属性名称相似性
  - Jaccard, Dice, 编辑距离
- 外部同义词知识库
  - 妻子, 老婆
- 属性取值相似度

## ② 过滤：删除错误候选属性对

- 启发式规则
  - 等价属性不同时出现在一个实体中
  - 等价属性domain和range相同

## ③ 验证：人工验证



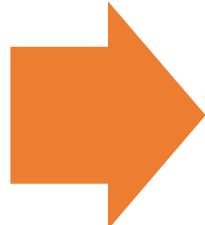
# 质量问题二：数值属性值格式不统一

- 日期表述不一致
  - 2020年02月02日
  - 20200202
  - 2020/2/2
- 单位表述不一致
  - 1.76米
  - 176cm
  - 176厘米
  - 1.76

# 数值属性值归一化

- 将所有的数值属性值统一表示

数值抽取



单位统一

```
#Patterns抽取年、月、日↓  
ymd_re1_=re.compile(r'(\d{3,4})[^\d]*-[^\d]*(\d{1,2})[^\d]*-[^\d]*(\d{1,2})')↓  
ymd_re2_=re.compile(r'(\d{3,4})[^\d]*\.(\d{1,2})[^\d]*\.(\d{1,2})')↓  
ymd_re3_=re.compile(u'(\d{3,4})[^\d]*年[^\d]*(\d{1,2})[^\d]*月[^\d]*(\d{1,2})[^\d]*')↓  
ymd_re4_=re.compile(r'^(\d{3,4})/(\d{1,2})/(\d{1,2})$')↓
```

长度 面积 体积 质量 温度 压力 功率 功能/热 <>

1 千米(km) 米(m)

1千米(km)=1000米(m)

国际单位：米(m)

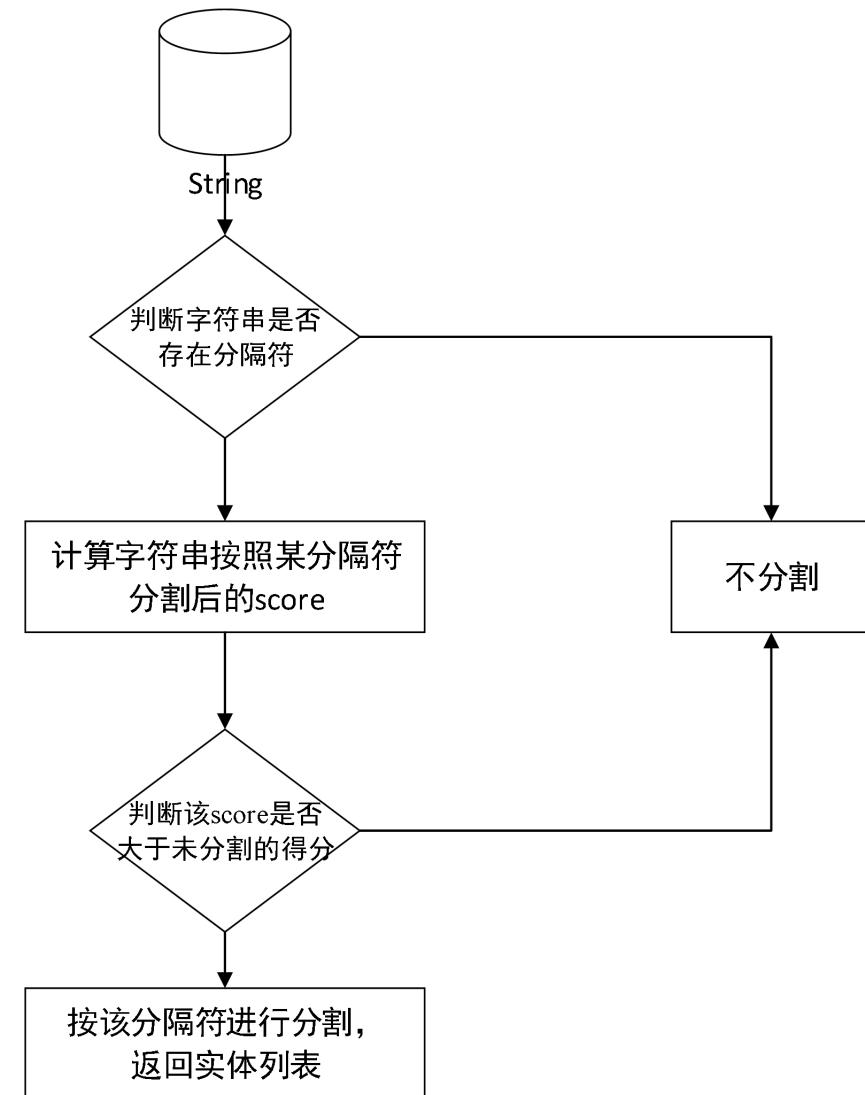
# 对象属性值分割

- 分割思路

- 对于任意一个属性，如果分割后的属性值集合中的大部分属性值都指代特定实体，那么这个属性很可能是多值的对象属性，对相应属性值进行分割是合理的尝试

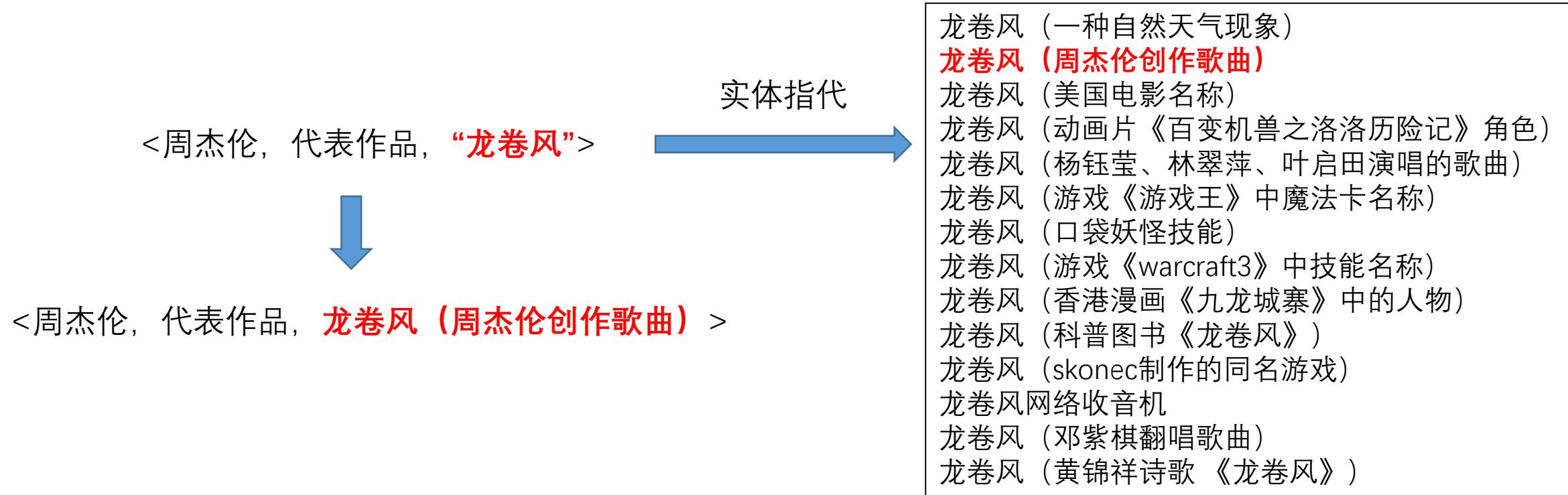
- 分隔符

- 空格、中文逗号、英文逗号、中文顿号、英文斜杠、中文分号、英文分号、英文竖号



# 属性抽取步骤的遗留问题

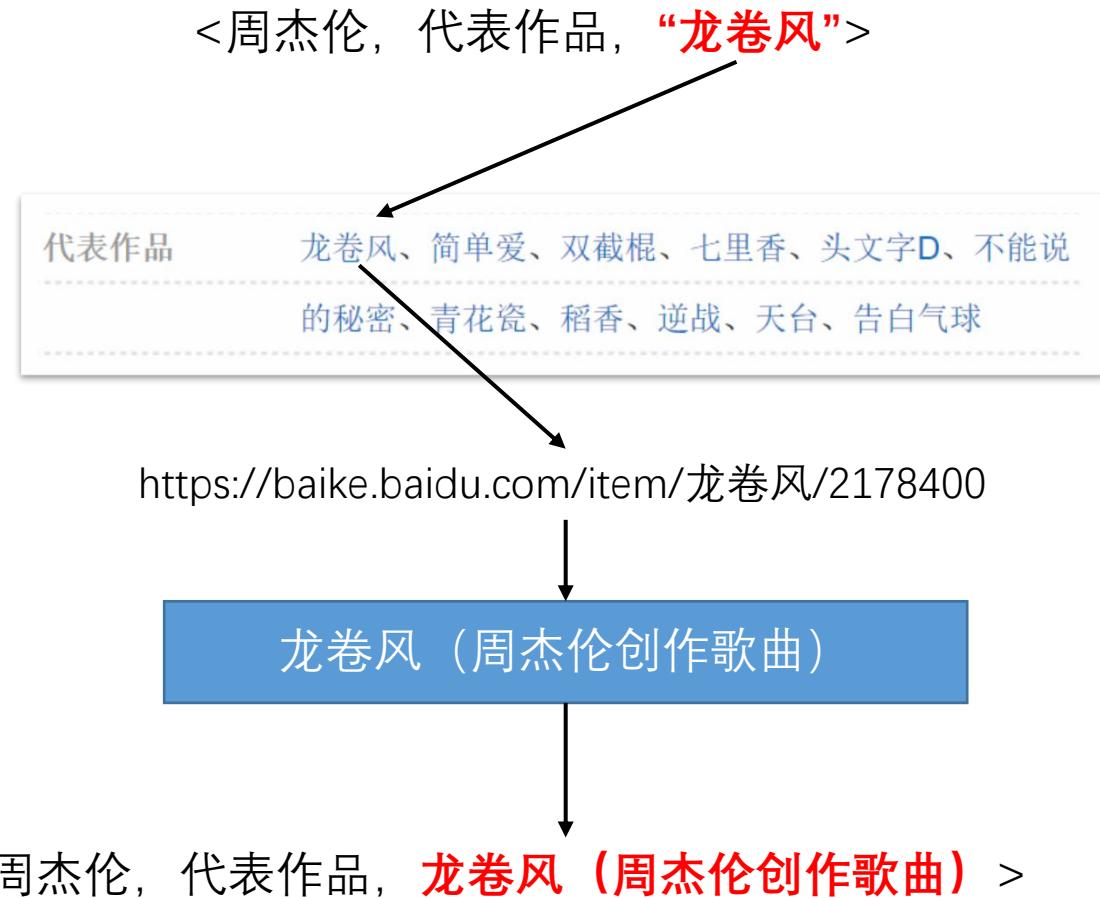
- 未建立实体与实体之间的关系
- 缺少这一步，知识图谱中的实体无法建立关联



# 对象属性值实体链接

- 场景一：当属性值存在超链接时

- 解析超链接对应的URL



# 对象属性值实体链接

## • 场景二：当属性值不存在超链接时

- 建模为分类问题

- 输入：

- 一个（实体，属性，属性值）三元组
    - 属性值对应的所有候选实体列表

- 输出
    - 0个或1个正确的实体

- 模型

$$s(m, e) = \sum_{i=1}^7 w_i \times f_i(m, e)$$

[Mengling Xu etc., 2013]

<周杰伦，代表作品，“**龙卷风**”>

龙卷风（一种自然天气现象）  
**龙卷风（周杰伦创作歌曲）**  
龙卷风（美国电影名称）  
龙卷风（动画片《百变机兽之洛洛历险记》角色）  
龙卷风（杨钰莹、林翠萍、叶启田演唱的歌曲）  
龙卷风（游戏《游戏王》中魔法卡名称）  
龙卷风（口袋妖怪技能）  
龙卷风（游戏《warcraft3》中技能名称）  
龙卷风（香港漫画《九龙城寨》中的人物）  
龙卷风（科普图书《龙卷风》）  
龙卷风（skonec制作的同名游戏）  
龙卷风网络收音机  
龙卷风（邓紫棋翻唱歌曲）  
龙卷风（黄锦祥诗歌《龙卷风》）

Feature 1: Entity Occurrence  
Feature 2: Link Probability  
Feature 3: Infobox Context Relatedness  
Feature 4: Article Context Relatedness  
Feature 5: Abstract Context Relatedness  
Feature 6: Attribute Range Context Relatedness  
Feature 7: Attribute Domain Context Relatedness

# 概念层级体系构建

---

- 概念层级体系构建便于对知识图谱中的实体进行组织和管理
- 目前主要包括人工构建和半自动构建两种方式
- 百科图谱对概念层级体系的质量要求较高，一般不采用全自动的方式构建

# 人工构建方式

- 典型代表：DBpedia

- 通过众包的方式将来自维基百科数据源的所有实体用几百个概念进行有效的组织

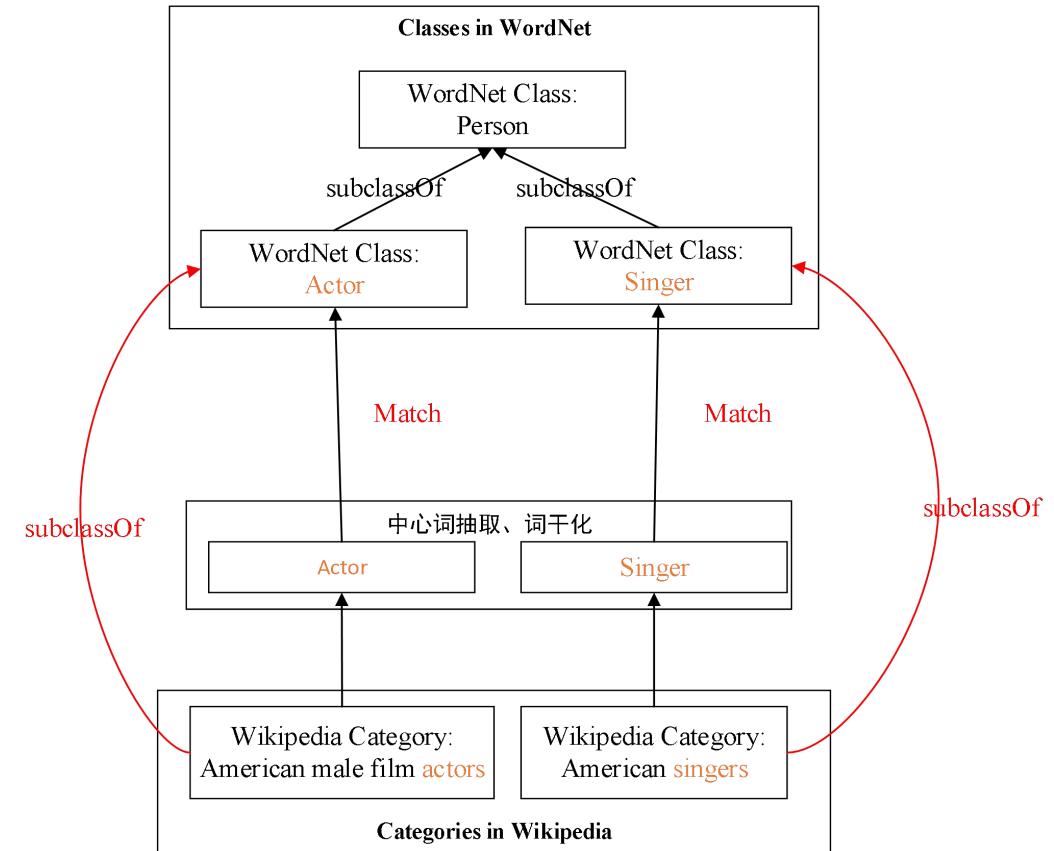
<http://mappings.dbpedia.org/server/ontology/classes/>

## Ontology Classes

- owl:Thing
  - Activity (edit)
  - Game (edit)
    - BoardGame (edit)
    - CardGame (edit)
  - Sales (edit)
  - Sport (edit)
    - Athletics (edit)
    - TeamSport (edit)
  - Agent (edit)
    - Deity (edit)
    - Employer (edit)
    - Family (edit)
      - NobleFamily (edit)
    - FictionalCharacter (edit)
      - ComicsCharacter (edit)
      - AnimangaCharacter (edit)
      - DisneyCharacter (edit)
      - MythologicalFigure (edit)
      - NarutoCharacter (edit)
      - SoapCharacter (edit)
    - Organisation (edit)
      - Broadcaster (edit)
        - BroadcastNetwork (edit)
        - RadioStation (edit)
        - TelevisionStation (edit)
      - Company (edit)
        - Bank (edit)
        - Brewery (edit)
        - Caterer (edit)
        - LawFirm (edit)
        - PublicTransitSystem (edit)
          - Airline (edit)
          - BusCompany (edit)
        - Publisher (edit)
        - RecordLabel (edit)
        - Winery (edit)
      - EducationalInstitution (edit)

# 半自动构建方式

- 典型代表：YAGO
  - 将WordNet作为上层本体
  - 建立Wikipedia conceptual categories与WordNet概念之间的subclassof关系



# 实体分类

---

- 定义
  - 将知识图谱中的实体分类到一组预定义的概念集合中，这组预定义的概念集合来自概念层级体系
- 分类方法
  - 人工方法
  - 基于规则的方法
  - 基于机器学习的方法

# 人工方法

- 利用人工来对知识图谱中的实体进行分类，参与人员包括领域专家和广大志愿者



通过人工方法建立infobox模板名称和概念的等价关系

## Ontology Classes

```
{|Infobox automobile
| name = Ford GT40
| production = 1964-1969
| engine = 4181 cc
| ...
|}
```

- owl:Thing
  - MeanOfTransportation (edit)
    - Aircraft (edit)
    - MilitaryAircraft (edit)
  - Automobile (edit)
    - Locomotive (edit)
    - MilitaryVehicle (edit)
    - Motorcycle (edit)

[Jens Lehmann et al. 2015]

# 基于规则的方法

- 使用一组IF-THEN规则来对实体进行分类

- 通用推理规则

- 基于等价实体关系的规则推理

$$(e_1 \in c) \wedge (e_1 = e_2) \Rightarrow e_2 \in c$$

- 基于概念子类关系的规则推理

$$(e \in c_1) \wedge (c_1 \subset c_2) \Rightarrow e \in c_2$$

- 启发式推理规则

- 基于标题的规则推理

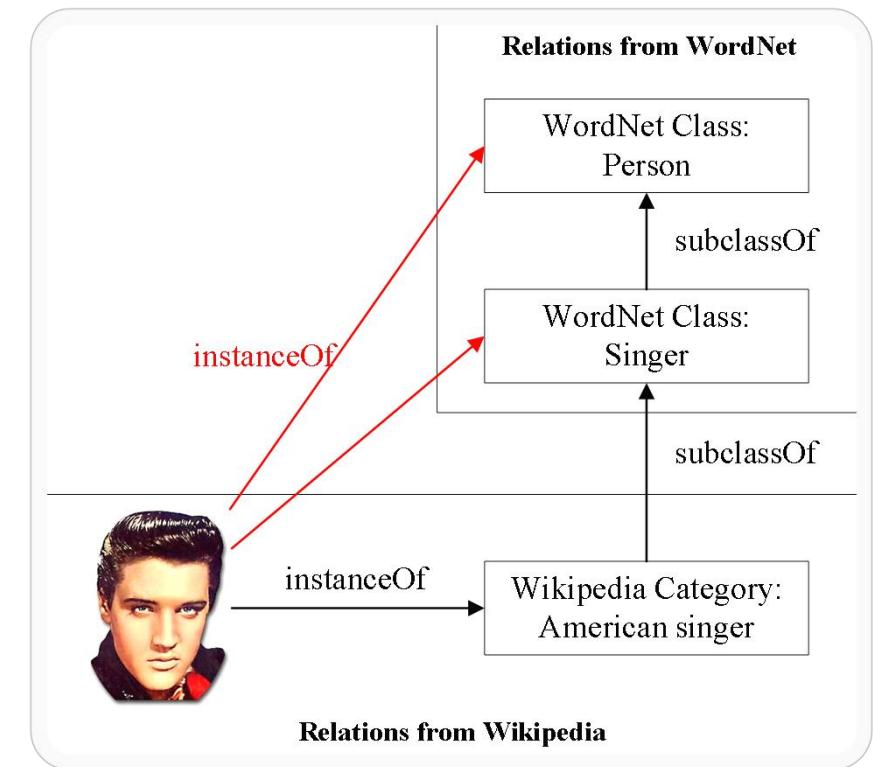
- E.g., 实体名称后缀为“步枪”的很可能属于步枪

- 基于属性的规则推理

- E.g., 实体包含属性“毕业院校”的属于人物

- 基于属性-值的规则推理

- E.g., 实体包含属性-值对（职业，演员）的属于演员



[Fabian, M. S. et al. 2007]

# 基于机器学习的方法

- 问题建模
    - 多标记分类问题 (Multi-label Classification)
    - 每个标记类代表知识图谱中的一个概念，一个实体可以属于知识图谱中的多个概念
  - 监督学习通用框架
    - 训练集构建
    - 特征抽取
    - 模型训练
    - 结果预测
- 
- ```
graph LR; TD[训练数据<br>(已分类的实体)] --> FE[特征提取]; TD --> MT[模型训练]; TD -.-> MT; TD -.-> M[模型]; FE --> M; M --> ER[实体分类结果]; TE[测试数据<br>(未分类的实体)] --> FE;
```

# 特征表示

## • 单示例特征表示

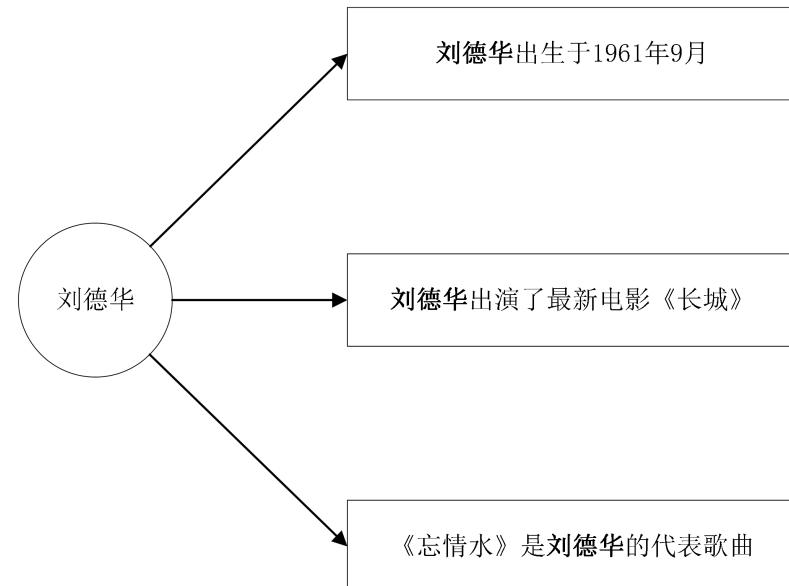
- 一个实体用一组特征集合表示

| Features    | 特征类型 |
|-------------|------|
| 血型          | 属性   |
| 妻子          |      |
| 国籍          |      |
| (职业, 演员)    | 属性-值 |
| (职业, 歌手)    |      |
| (代表作品, 忘情水) |      |
| 香港人         | 标签   |
| 港台男歌手       |      |
| 艺人          |      |

“刘德华”的单示例特征集合

## • 多示例特征表示

- 一个实体用多个示例表示，每次示例为一组特征集合
- 每个示例可能只表示实体部分分类结果



“刘德华”的多示例表示

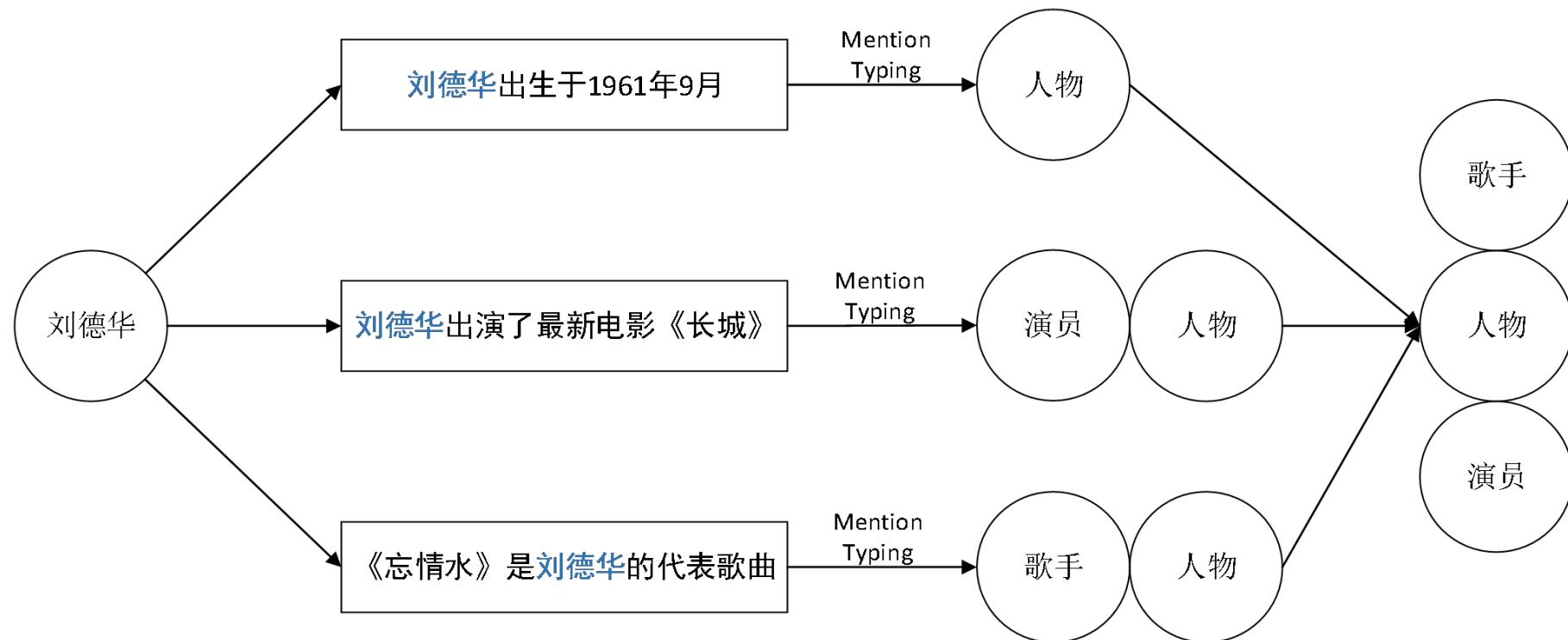
# 单示例实体分类

- 输入：实体的特征集合 $X$ 
  - $X = [x_1, x_2, \dots, x_i, \dots, x_N]$
  - $N$ 为特征总数
  - $x_i = 1$ : 实体包括这一特征
  - $x_i = 0$ : 实体不包含这一特征
- 输出：实体的分类结果 $Y$ 
  - $Y = [y_1, y_2, \dots, y_i, \dots, y_M]$
  - $M$ 为概念总数
  - $y_i = 1$ : 实体属于这个概念
  - $y_i = 0$ : 实体不属于这个概念

- 问题归类
  - 多标记分类 (Multi-label Classification)
    - 一个实体可以属于多个概念
- 分类模型
  - 朴素贝叶斯
  - Logistic回归
  - 支持向量机
  - 决策树

# 多示例实体分类：Pipeline方法

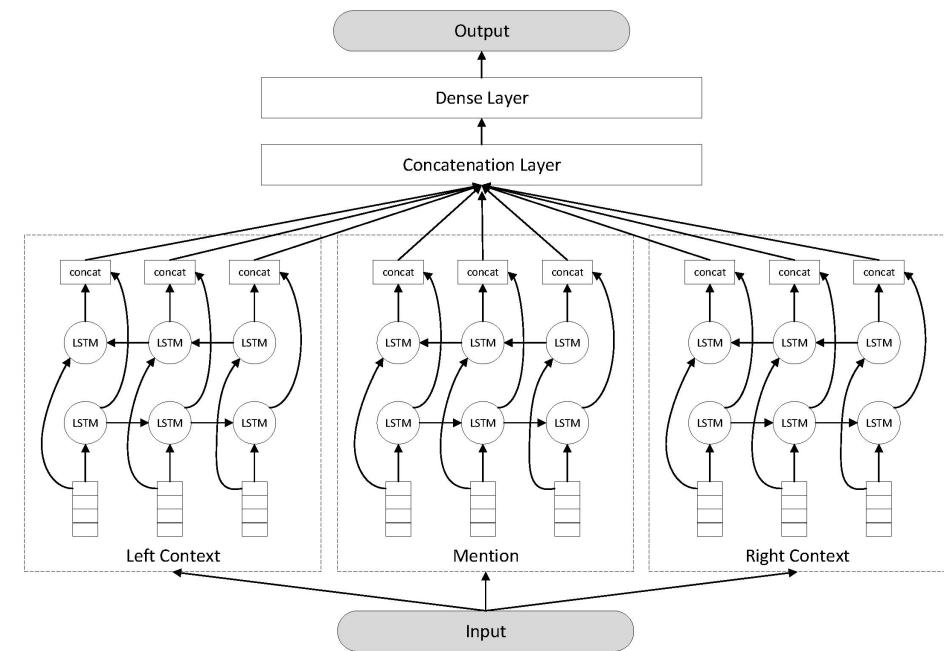
- 基本思路
  - Mention Typing + Type Fusion



[Bo Xu et al., 2018]

# Mention Typing

- 解决方案
  - 基于人工特征的方法
    - PL-SVM (Nguyen and Caruana, 2008)
    - CLPL (Cour et al., 2011)
    - FIGER (Ling and Weld, 2012)
    - FIGER-Min (Gillick et al., 2014)
    - HYENA (Yosef et al., 2012)
    - ClusType (Ren et al., 2015)
    - DeepWalk (Perozzi et al., 2014)
    - LINE (Tang et al., 2015b)
    - PTE (Tang et al., 2015a)
    - WSABIE(Yogatama et al., 2015)
    - AFET (Ren et al., 2016)
- 基于神经网络的自动特征抽取方法
  - 方法
    - HNM (Dong et al., 2015)
    - METIC (Bo Xu et al., 2018)
    - KNET (Ji Xin et al., 2018)



| Feature       | Description                                                                         |
|---------------|-------------------------------------------------------------------------------------|
| Head Token    | Syntactic head token of the mention                                                 |
| POS           | Tokens in the mention                                                               |
| Character     | Part-of-Speech tag of tokens in the mention                                         |
| Word Shape    | All character trigrams in the head of the mention                                   |
| Length        | Word shape of the tokens in the mention                                             |
| Context       | Number of tokens in the mention                                                     |
| Brown Cluster | Unigrams/bigrams before and after the mention                                       |
| Dependency    | Brown cluster ID for the head token (learned using $\mathcal{D}$ )                  |
|               | Stanford syntactic dependency (Manning et al., 2014) associated with the head token |

# Type Fusion

- 融合策略

- 直接合并
- 一致性投票
- 大多数投票
- 带约束合并

Maximize

$$\sum_{t \in \mathcal{T}} (\max_{m \in M_e} P(t|m) - \theta) \times x_{e,t}$$

Subject to

$$\forall_{ME(t_1, t_2)} x_{e,t_1} + x_{e,t_2} \leq 1$$

$$\forall_{IsA(t_1, t_2)} x_{e,t_1} - x_{e,t_2} \leq 0$$

- 带约束合并

- 将其看作是一个整数线性规划问题
- 目标函数
  - 最大化所有mention的分类结果
- 约束
  - 概念互斥约束
    - 一个实体不能同时属于两个语义互斥的概念
  - $PMI(c_1, c_2) = \log \frac{P(c_1, c_2)}{P(c_1) \times P(c_2)}$
  - 概念层次化约束
    - 一个实体如果不属于某个概念，那么也不能属于这个概念的任意子概念

[Bo Xu et al., 2018]

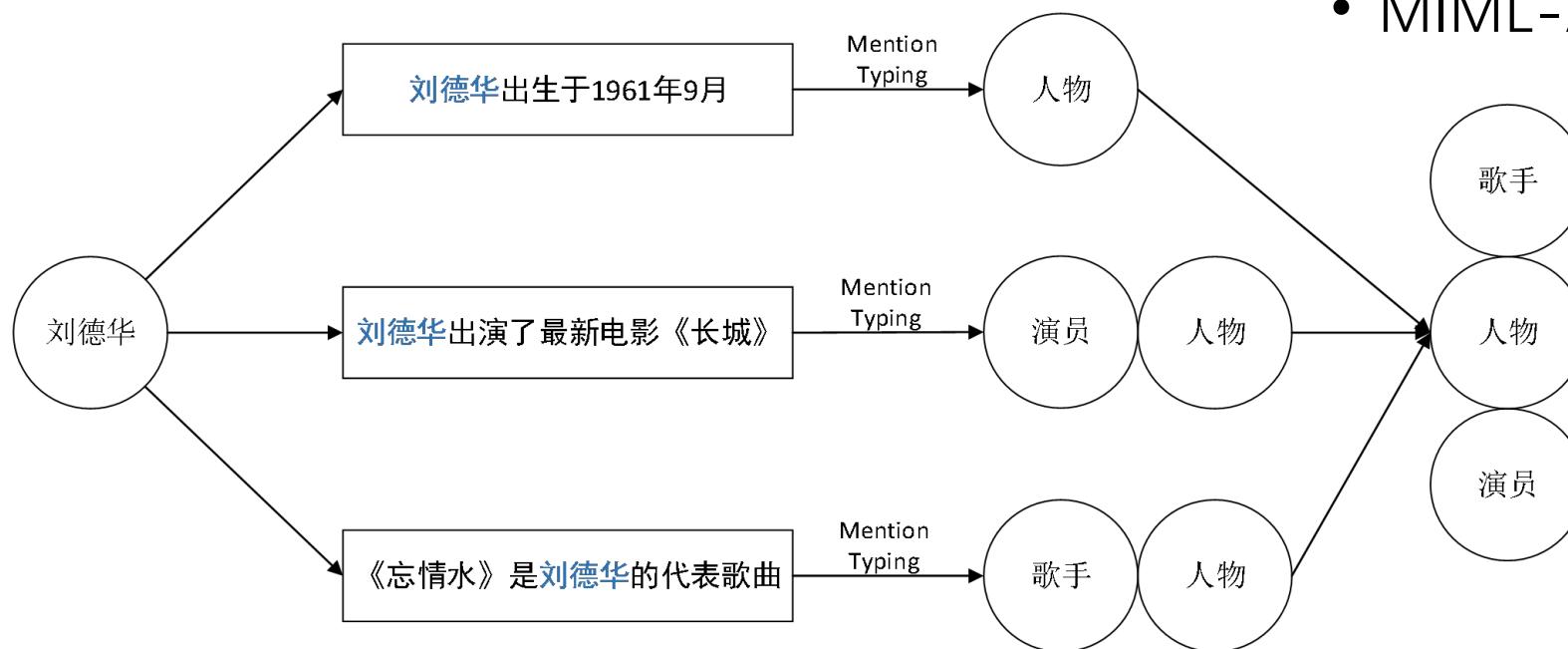
# 多示例实体分类：多示例学习方法

- 基本思路

- 输入：一个实体的全部示例
- 输出：一个实体的分类结果

- 方法

- MIML-MAX
- MIML-AVG
- MIML-MAX-AVG
- MIML-ATT



(Yadollah Yaghoobzadeh et al., 2017)

# References

---

- Miller, George A. "WordNet: a lexical database for English." *Communications of the ACM* 38.11 (1995): 39-41.
- Hearst, Marti A. "Automatic acquisition of hyponyms from large text corpora." *Proceedings of the 14th conference on Computational linguistics- Volume 2*. Association for Computational Linguistics, 1992.
- Wu, Wentao, et al. "Probase: A probabilistic taxonomy for text understanding." *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*. ACM, 2012.
- Liang, Jiaqing, et al. "On the Transitivity of Hypernym-Hyponym Relations in Data-Driven Lexical Taxonomies." *AAAI*. 2017.
- Liang, Jiaqing, et al. "Graph-Based Wrong IsA Relation Detection in a Large-Scale Lexical Taxonomy." *AAAI*. 2017.
- Liang, Jiaqing, et al. "Probase+: Inferring Missing Links in Conceptual Taxonomies." *IEEE Transactions on Knowledge and Data Engineering* 29.6 (2017): 1281-1295.
- Ponzetto, Simone Paolo, and Michael Strube. "WikiTaxonomy: A Large Scale Knowledge Resource." *ECAI*. Vol. 178. 2008.
- Fabian, M. S., K. Gjergji, and W. E. I. K. U. M. Gerhard. "Yago: A core of semantic knowledge unifying wordnet and wikipedia." *16th International World Wide Web Conference, WWW*. 2007.