

《知识图谱：概念与技术》

概念图谱构建

本章大纲

- 概念图谱概述
- isA关系抽取
- isA关系补全
- isA关系纠错

概念图谱概述

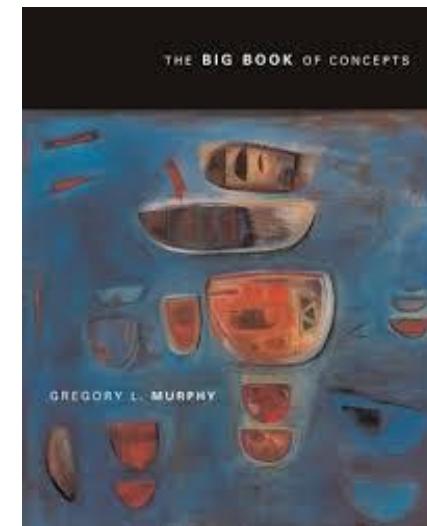
概念认知

- 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
.....

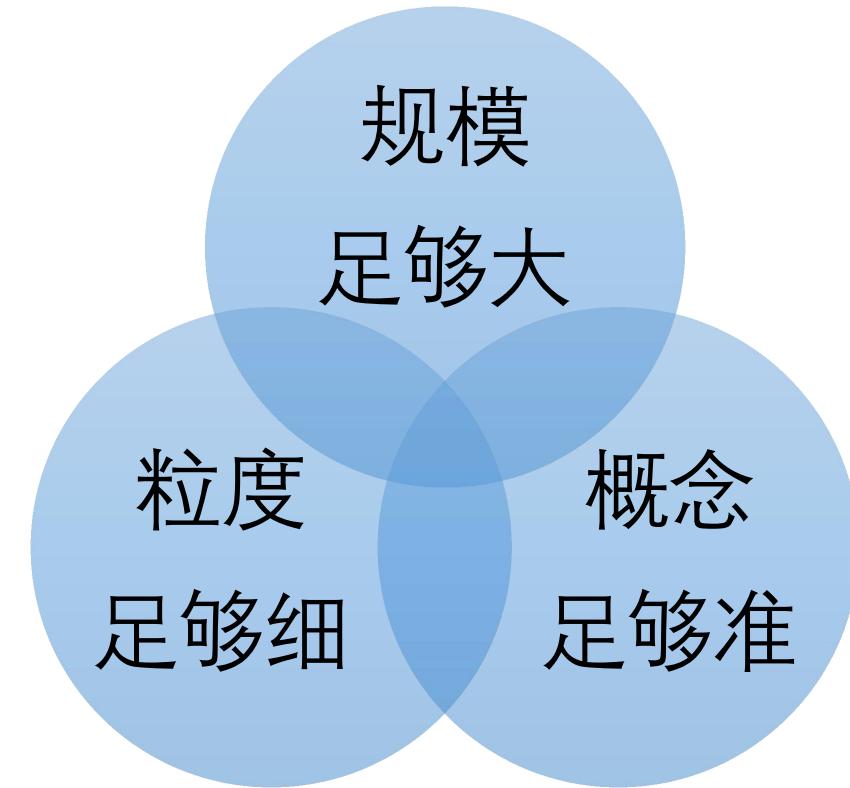
符号形式的概念



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

概念认知的重要

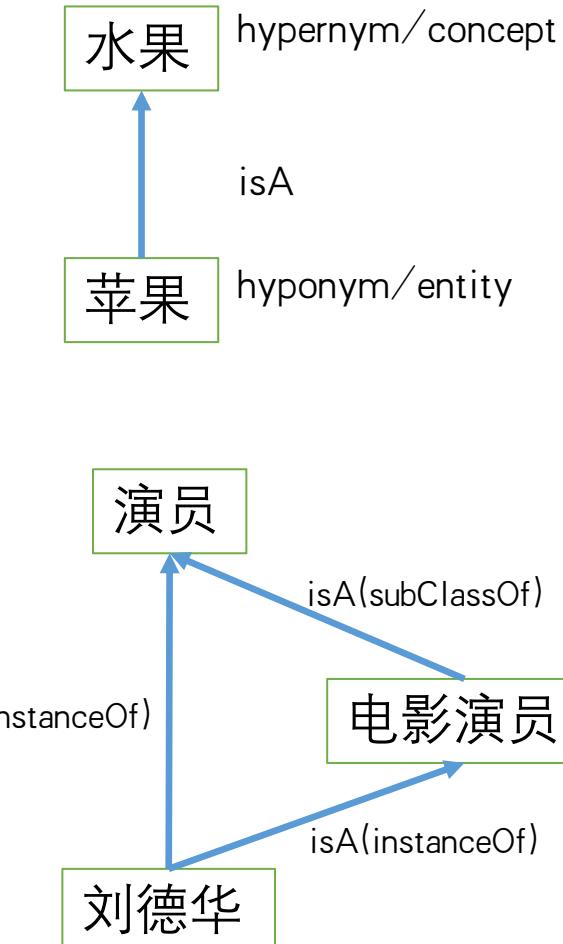
- 人类能“理解”事物的重要体现之一就是**产生概念**
 - 柏拉图（实体） → 哲学家（概念）
- 人类借助概念认知**同类实体**
 - 比如，汽车这一概念使得我们能够认知各种不同类型的汽车，而无需纠缠于各种细节的不同
- 概念是**联想**的重要隐含因素
 - 鸡 → 鸭（家禽）
 - 豆浆 → 油条（早餐）
- 概念是**归纳与推理**的基础
 - 哲学家有自己的哲学观点，柏拉图是哲学家→柏拉图有自己的哲学观点



**大规模概念图谱使得机器
认知实体的概念成为可能**

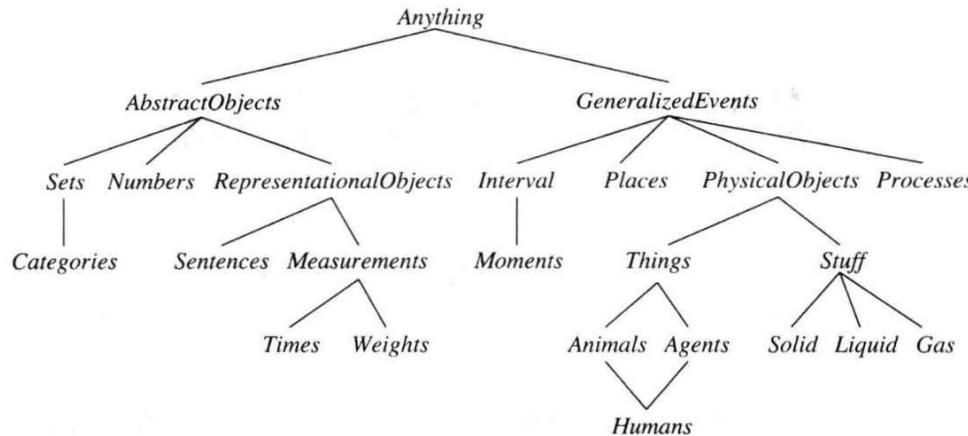
概念图谱的定义

- 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

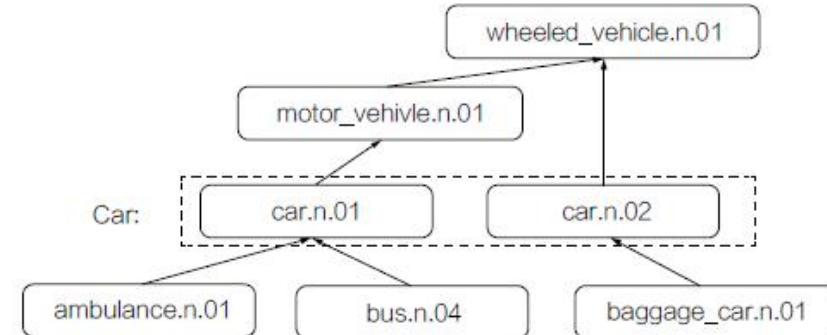
常见的概念图谱

- 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)

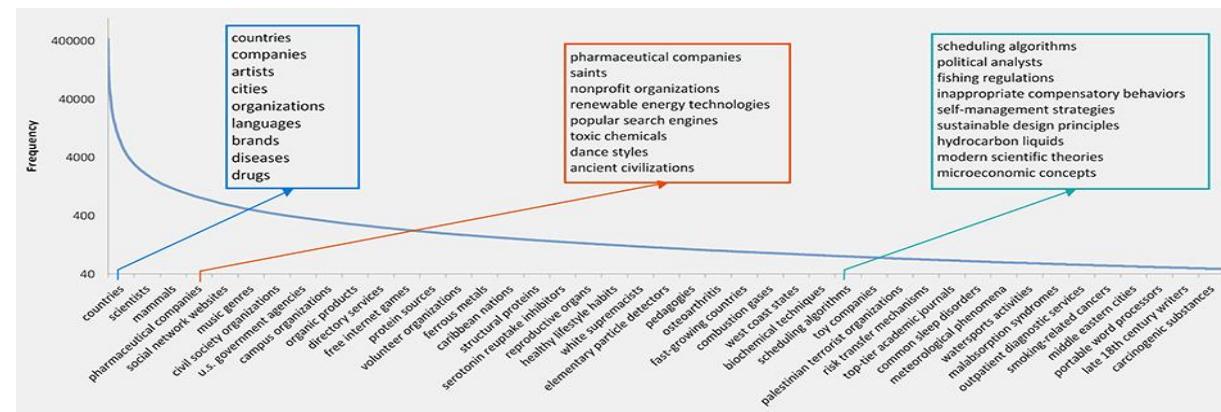
概念图谱	作 者	实 体	概 念	isA 关系数	准确度	权 重
WordNet (英文)	普林斯顿认知科学实验室	—	117 659	84 428	100%	无



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。
- **Probbase**: 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
Probbase (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

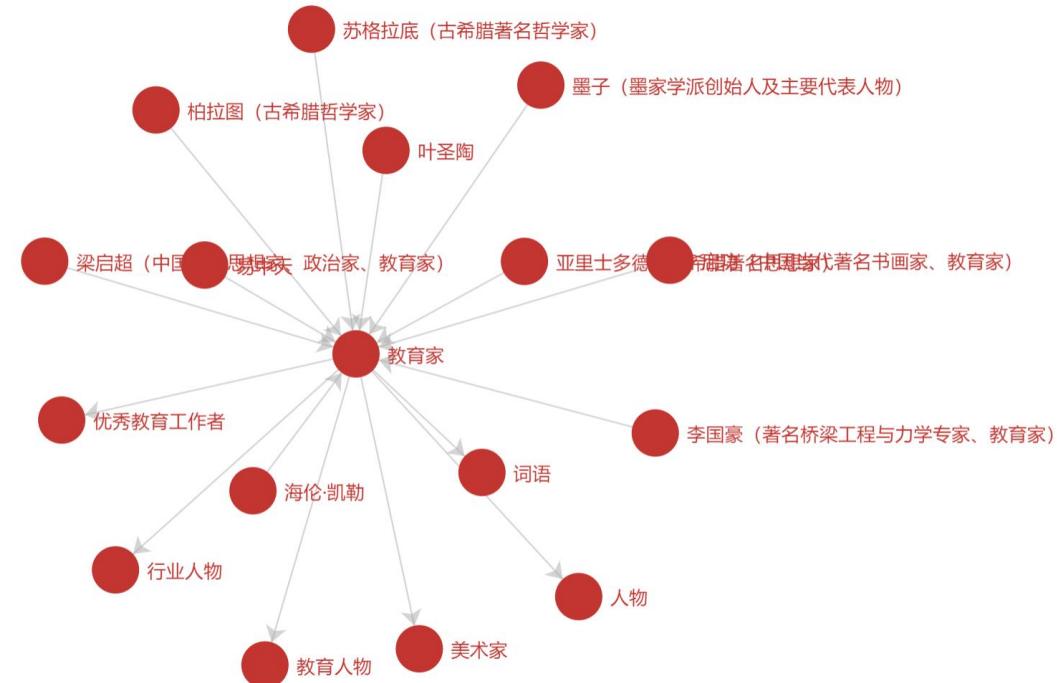
- Probbase中的频数表示该关系在语料中出现的次数
- 对于刻画实体或概念的典型性具有重要意义
 - $P(c|e) = \#(e \text{ isa } c) / \#e$; $P(e|c) = \#(e \text{ isa } c) / \#c$

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

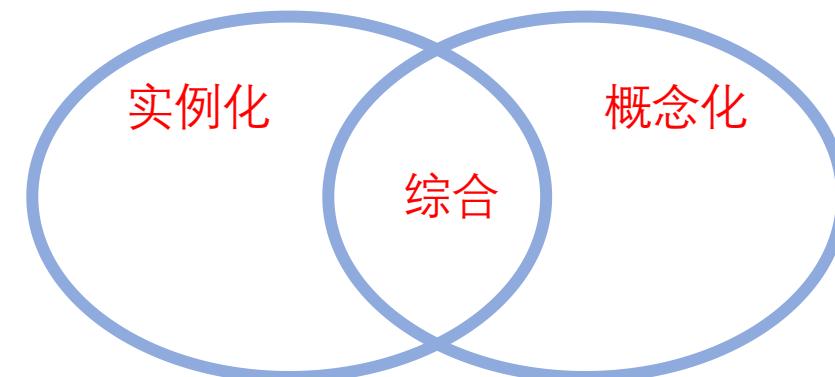


<http://kw.fudan.edu.cn/cnprobase/search/>

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



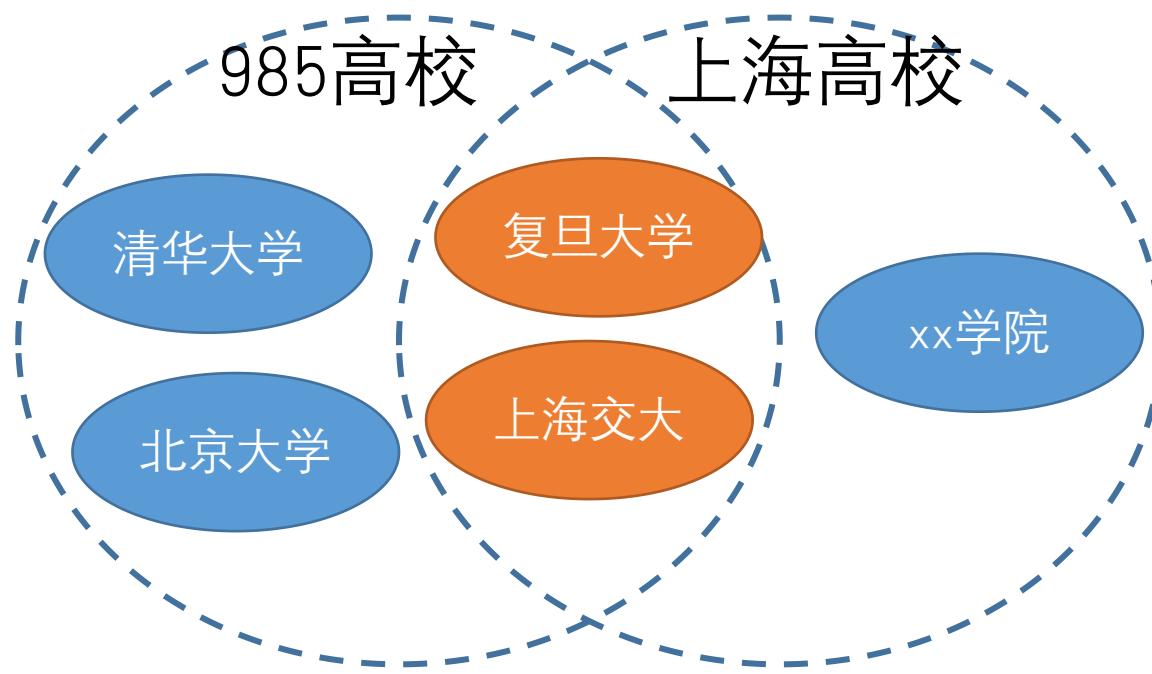
基于实例化的应用：实体搜索

Query: 水果

苹果
雪梨
西瓜
.....
蛇果
山竹
.....

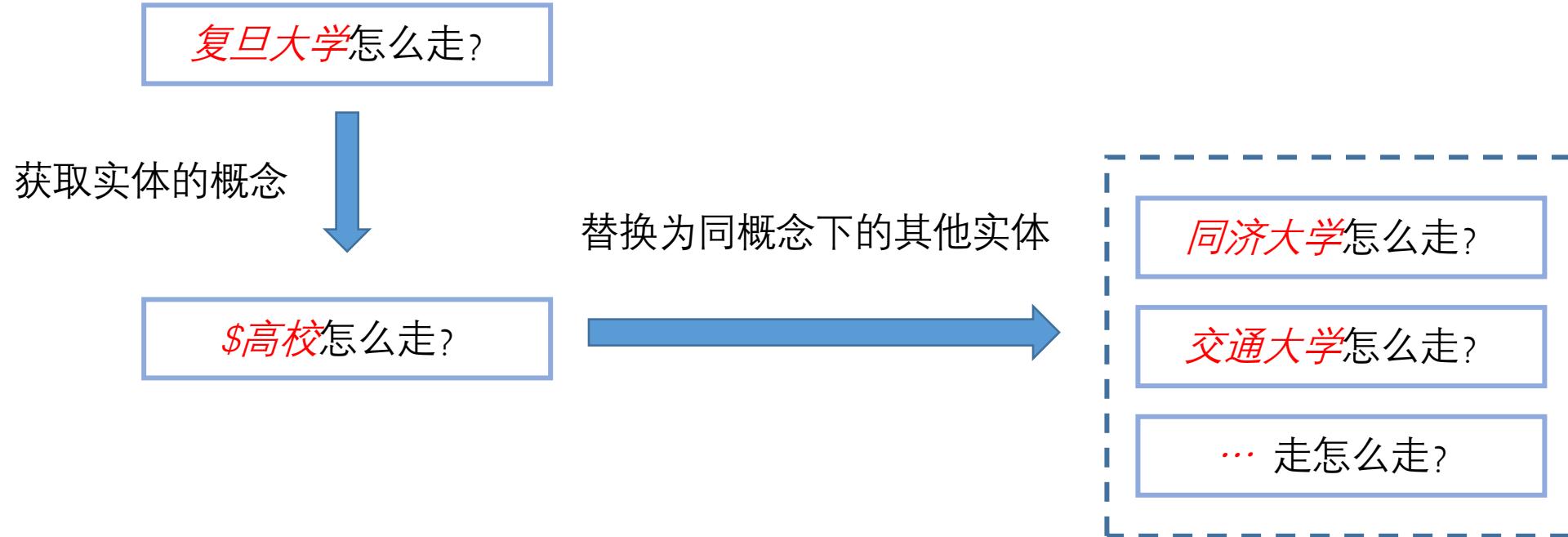
} 更常见
} 更稀有

Query: 上海的985高校

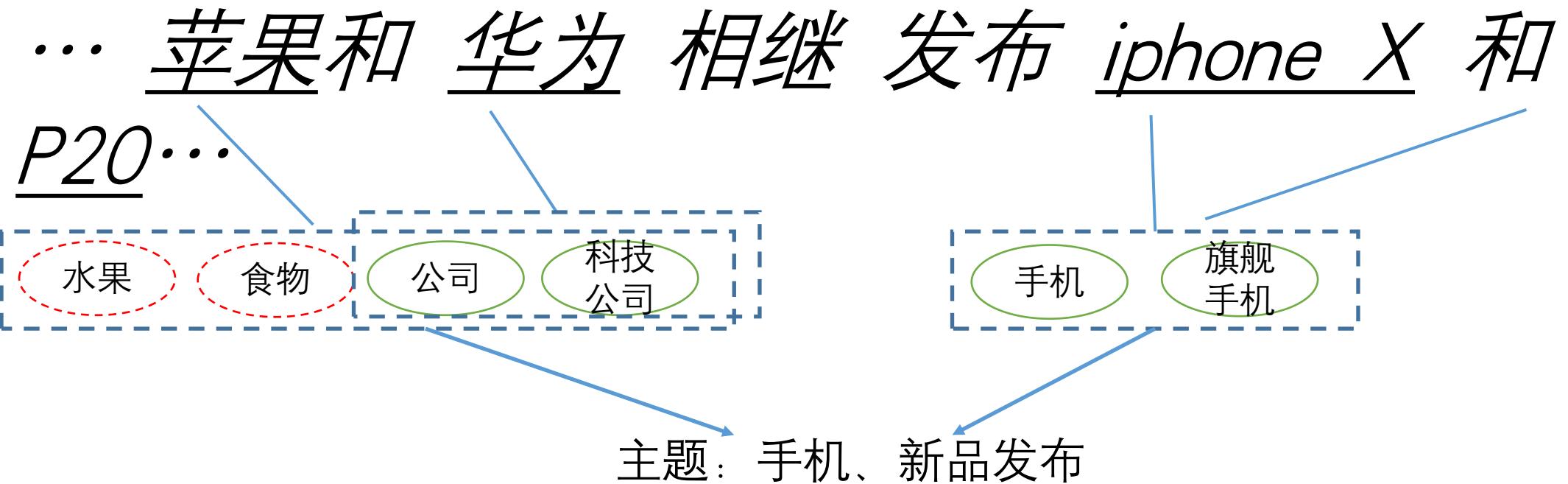


基于实例化的应用：样本增强

- 定义：
 - 利用概念下的实例，增强样本。

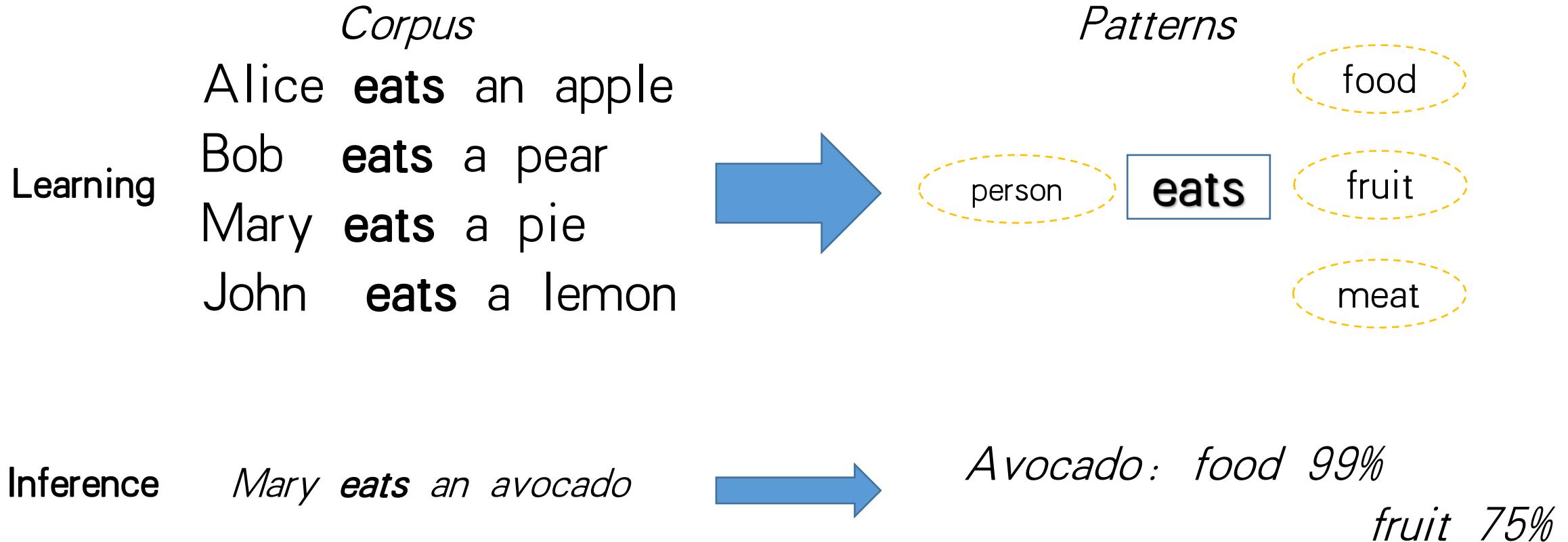


基于概念化的应用：主题理解



基于概念化的应用：语言概念模板

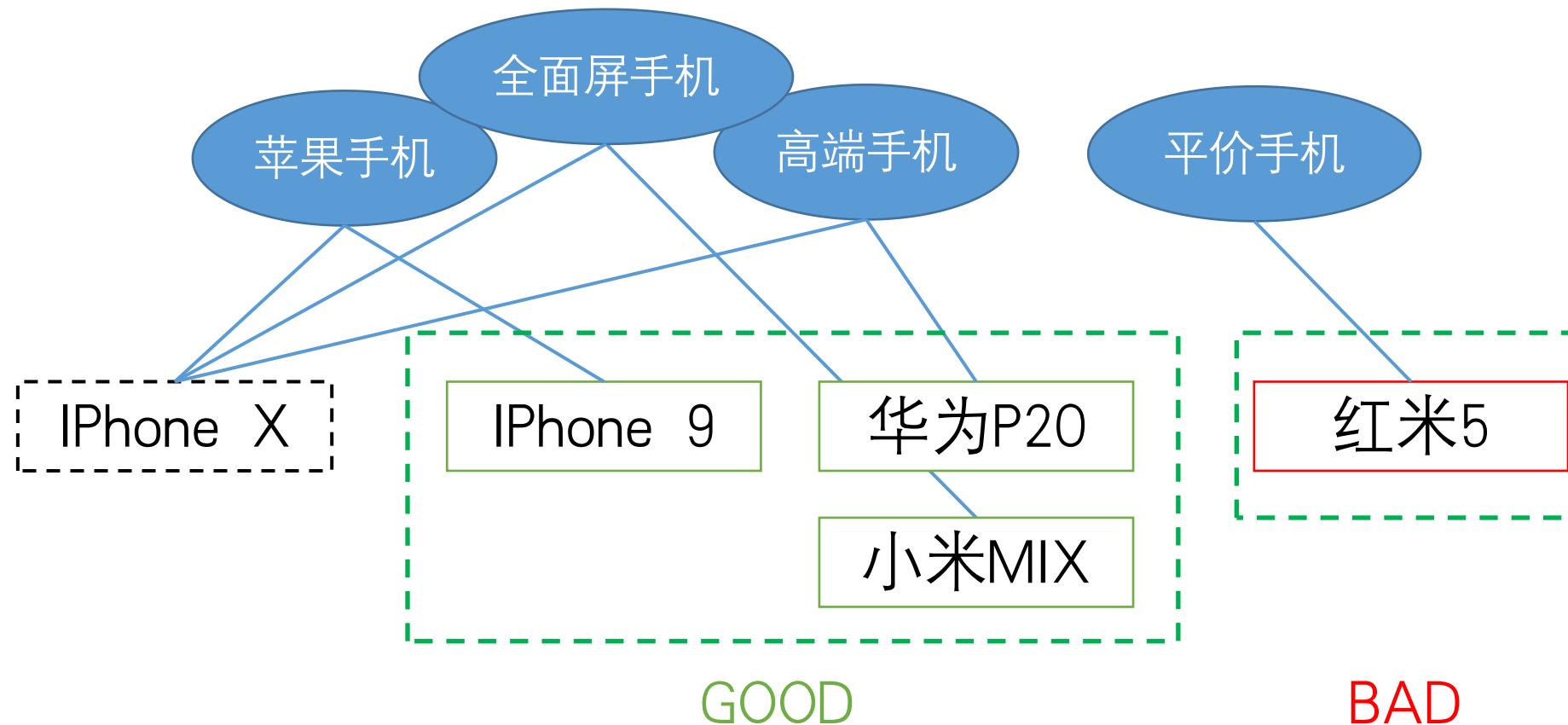
• 语言概念模板



基于概念化的应用

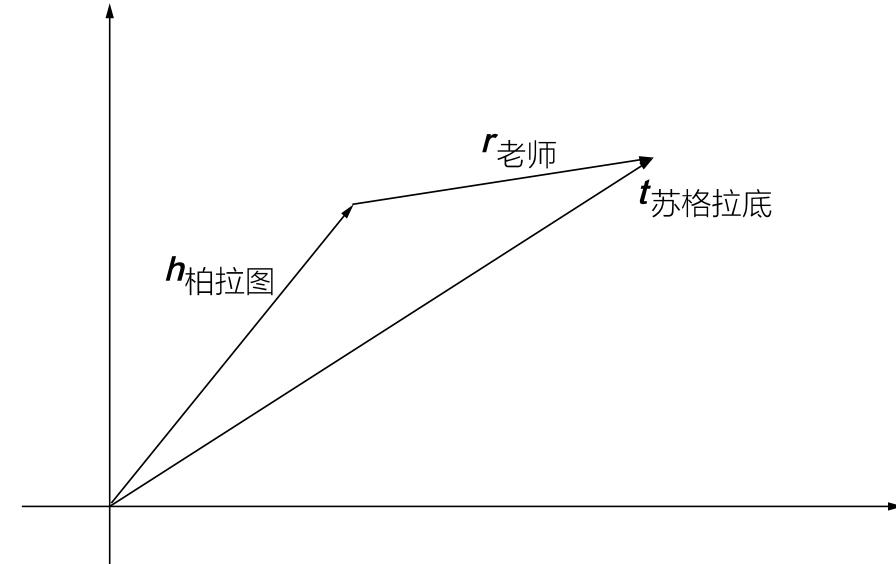
文本分类	根据文本中实体的概念，将文本分为不同类别	包含“足球”“篮球”的文本应当与包含“宝马”“奔驰”的文本属于不同类别
主题分析	给定文本，分析文本属于什么主题	包含“足球”“篮球”的文本应当属于体育类主题
给文档打标签	给定文本，给文本标上若干个概念作为标签	对文本“iPhone进水了怎么办”可以打上“手机，维修”等标签
用户画像	给定用户信息，为用户生成显式的概念	根据用户描述“精通Java、Android开发”，可以为其打上“面向对象编程”“手机App开发”的标签
基于概念的解释	根据概念信息，为事件提供解释	特斯拉Model S的加速性能很好，因为它是“电动汽车”，电动汽车通常具有较好的加速性能
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。
 - 该方法通过人类可以理解的符号概念，使语义表示具有可控、可解释等优点。

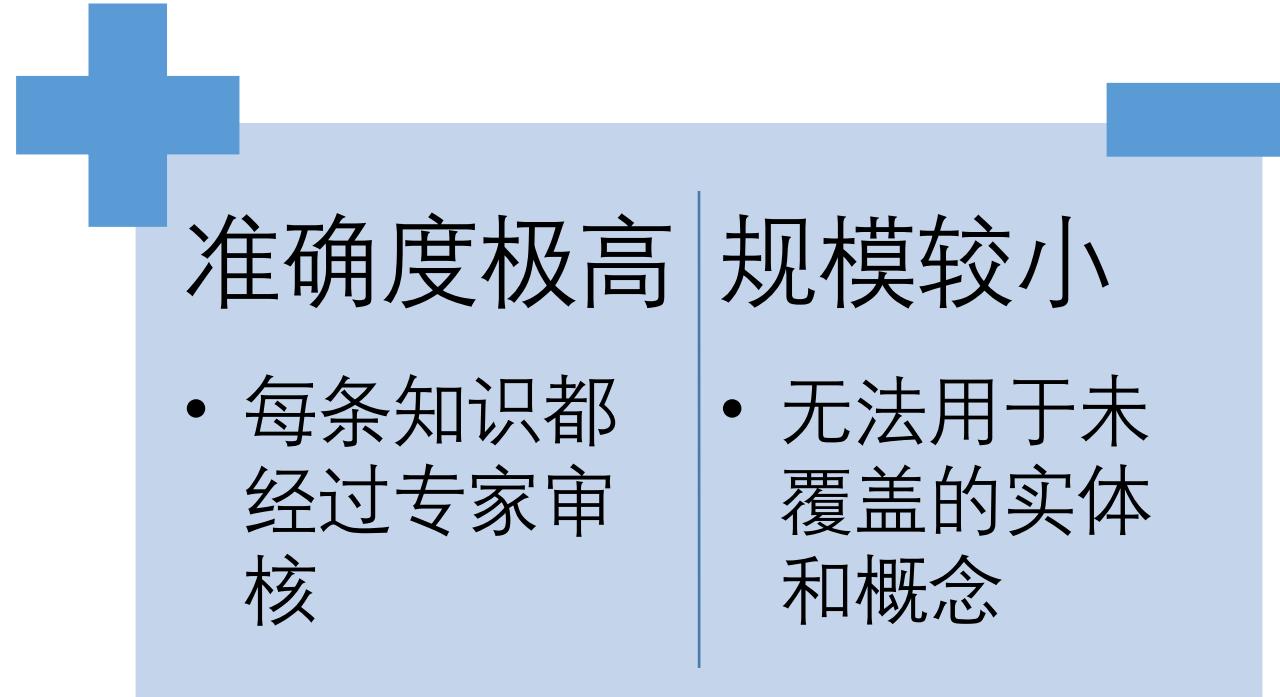


柏拉图：<哲学家、希腊人、男人、古代人>

isA关系抽取

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
- Probbase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-Probbase 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



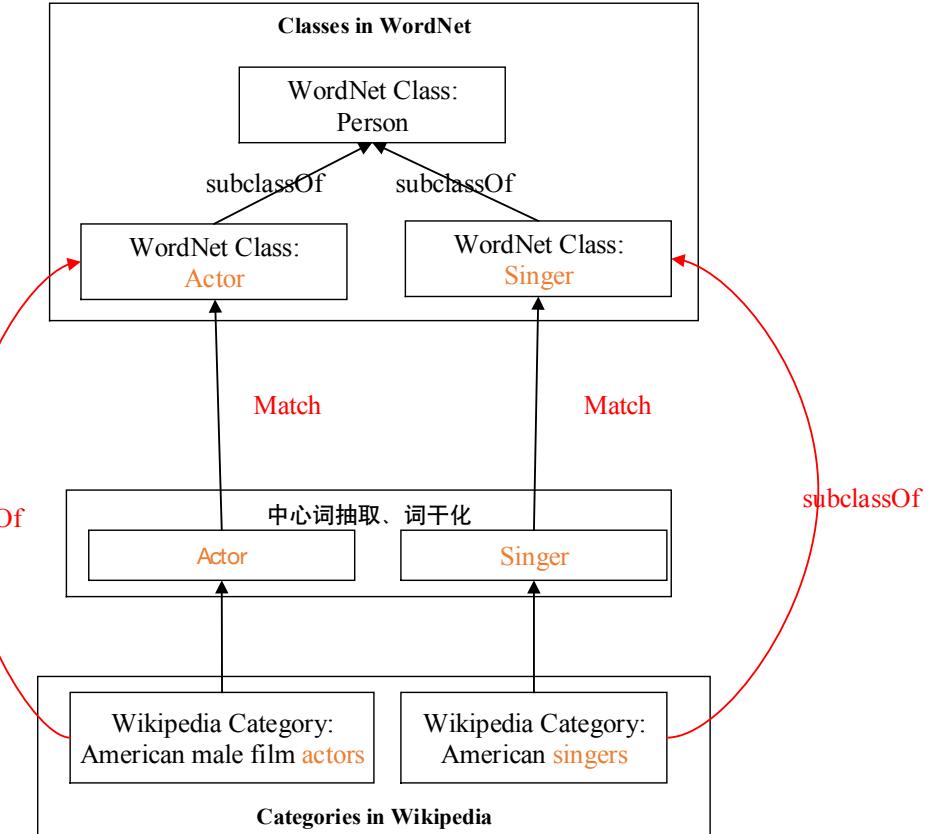
IsA关系抽取：YAGO

- 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 …



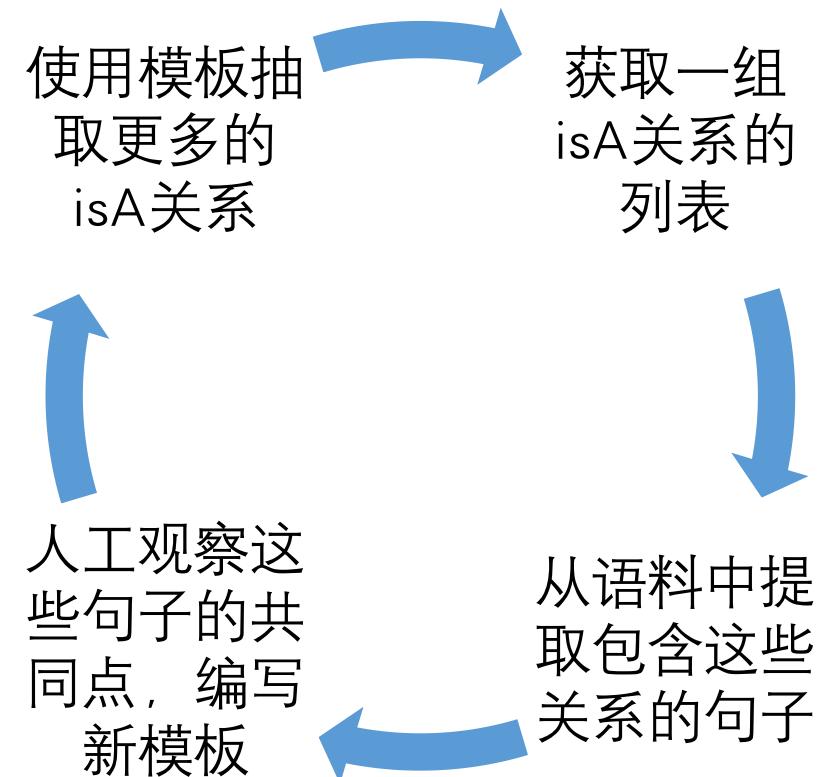
cat isA animal

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

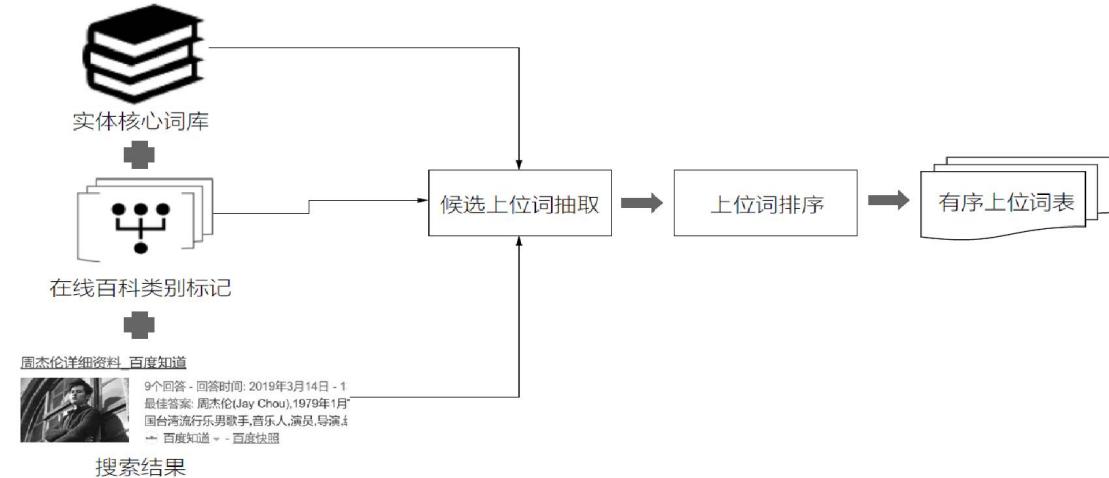
基于在线百科

- 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



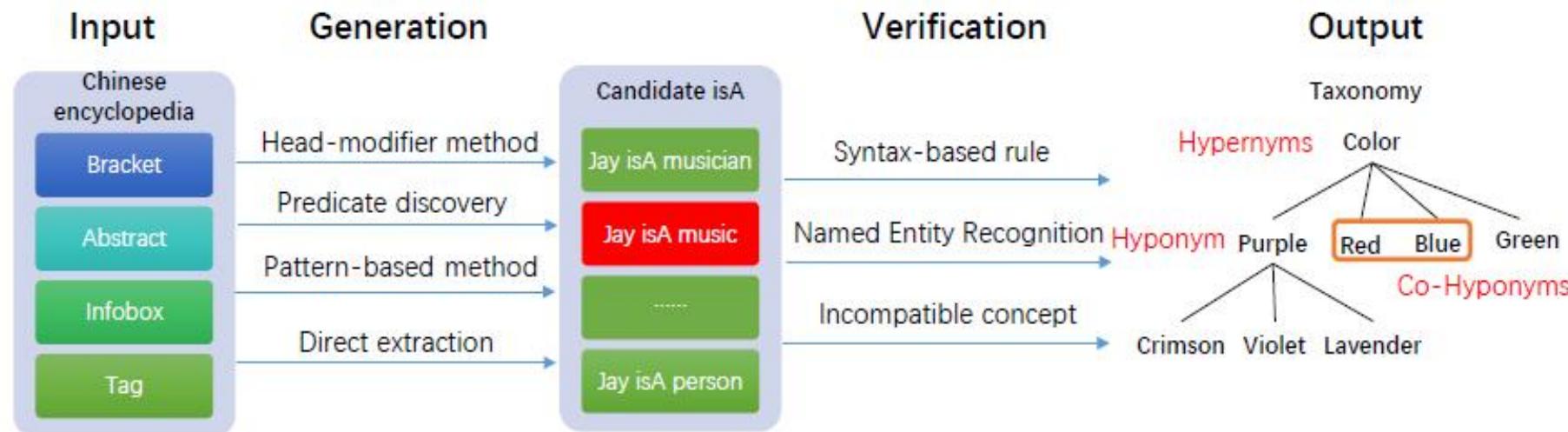
中文isA关系抽取：CN—Probase

- 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—Probbase

- 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

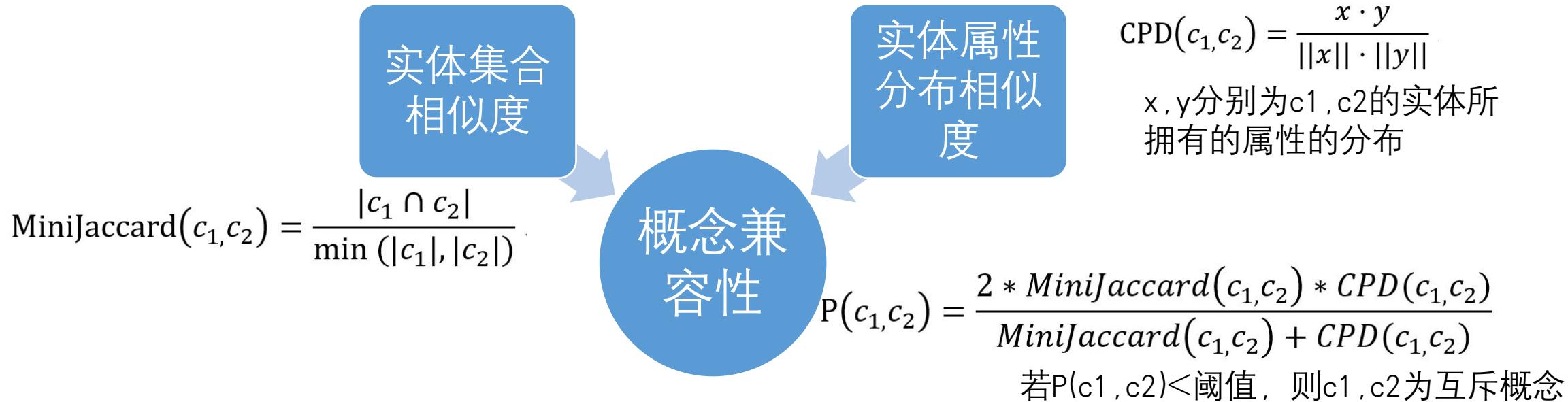
刘德华 (中国香港男演员、歌手、词作人)
Dehua Liu (Hong Kong actor, singer and songwriter)

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。

(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



isA关系补全

缺失成因

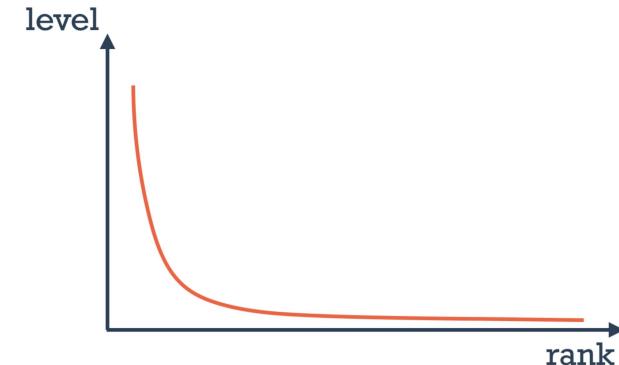
- 某个特定的文本语料只是对知识全集的一个不完整的表达
 - Probase中平均每个实体/概念仅有1.6个上位词

常识相关

- 大量常识性的 isA 关系，在语料中鲜有提及
- 如：“柏拉图”是一个人

低频实体

- 大多数实体在语料中出现的频数非常低（幂律分布）
- 如：“腓特烈二世”仅有“皇帝”一个概念

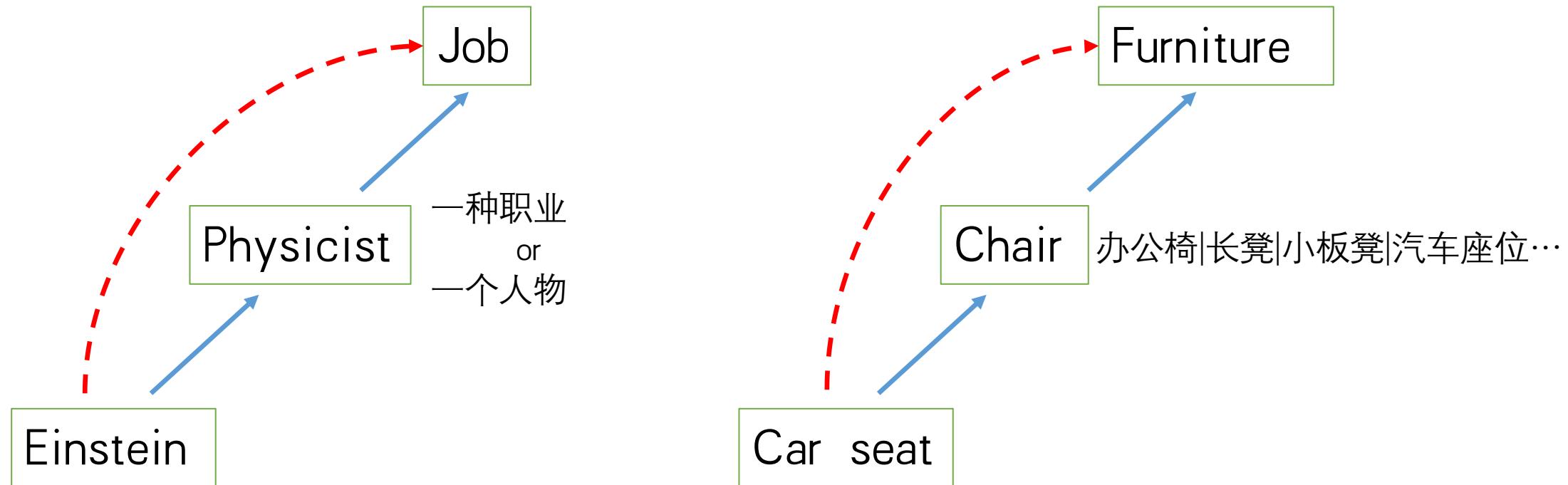


增加语料无法很好解决以上问题！

<https://medium.com/@nicolasterpolilli/the-power-law-of-data-opening-645a35ef03f2>

基于isA关系传递性的概念图谱补全

isA关系在理论上具有传递性。



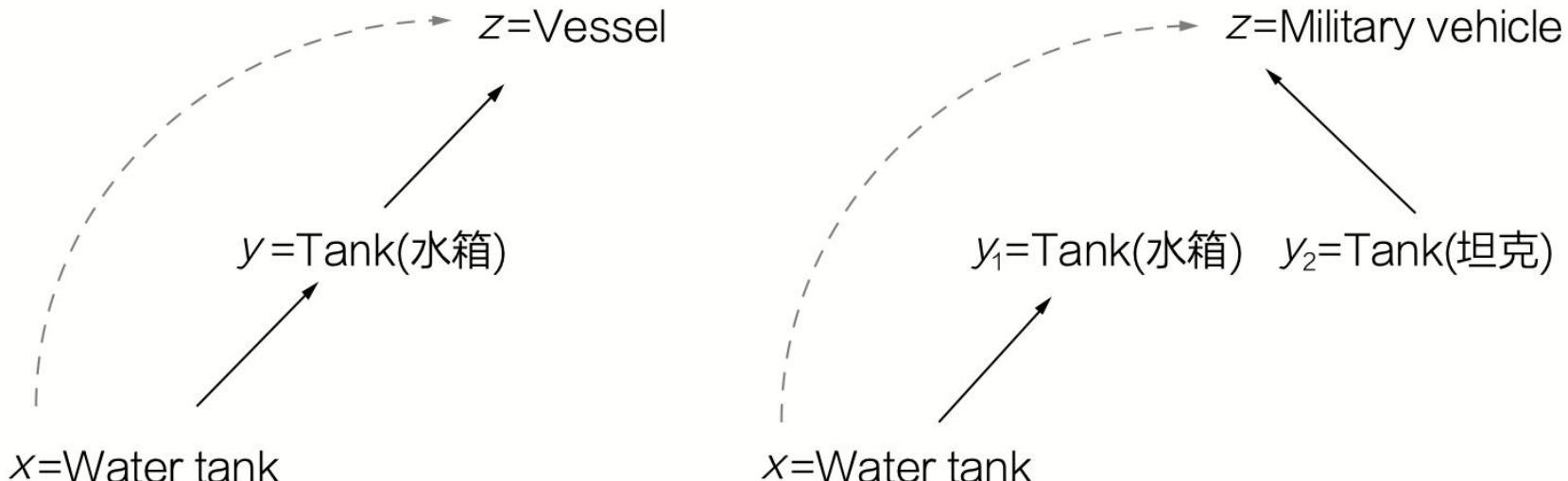
观察：在从大规模文本语料自动抽取构建的大规模词汇概念图谱(比如 Probosc)中，isA 关系的传递性却不一定总是成立的。

基于isA关系传递性的概念图谱补全

- isA关系的传递性是否成立?
 - 只有在isA传递性成立的情况下，才能利用isA传递性来进行补全
 - 三元组 $\langle x, y, z \rangle$, isA传递性成立, $x \text{ isA } y, y \text{ isA } z$, 则补全 $x \text{ isA } z$
- 建模为机器学习二分类问题:
 - 样本标注
 - 特征

样本标注

- 标注数据：利用专家构建的高质量概念图谱
 - WordNet：经过消歧的、专家构建的、isA自然传递性的概念图谱
- 方案：
 - 在WordNet中，若 x isA y 且 y isA z ，则 x isA z 成立，因此 $\langle x, y, z \rangle$ 是一个正例。
 - 在WordNet中，若 x isA y_1 且 y_2 isA z ，且 y_1 和 y_2 是同一个词 y 的不同词义，则 x isA z 不成立，因此 $\langle x, y, z \rangle$ 是一个负例。



Positive: water tank - tank - vessel Negative: water tank - tank - military vehicle

特征

- 特征1：基本特征。

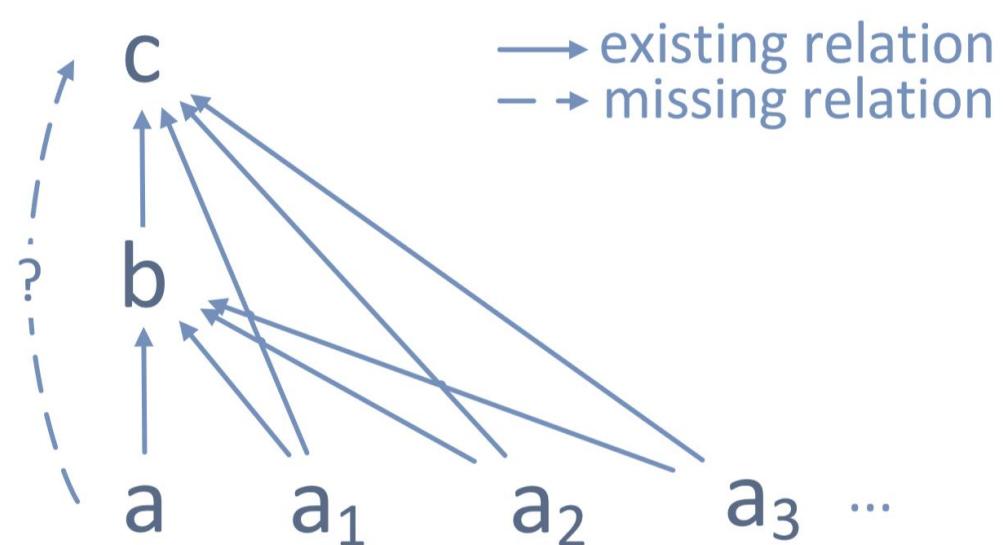
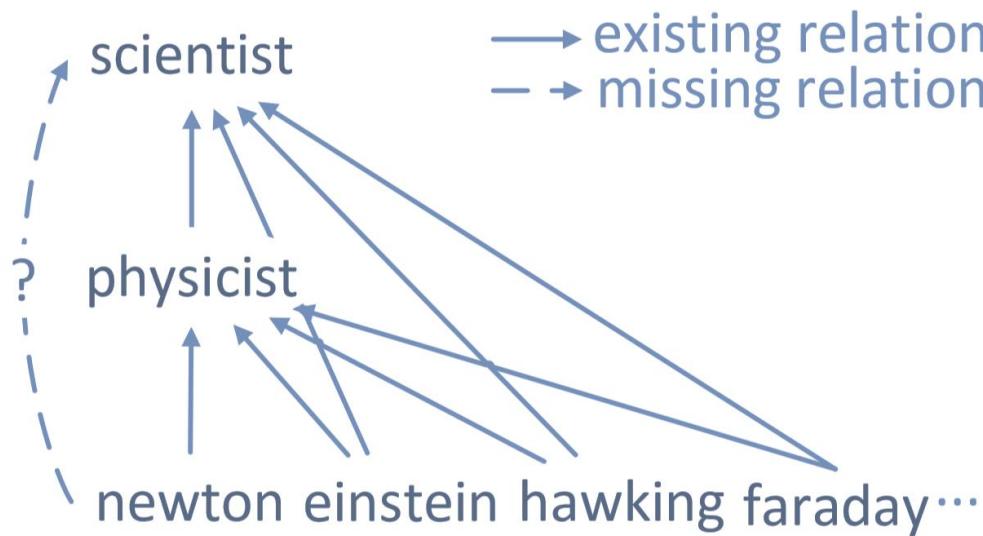
- 从 $x \text{ isA } y$ 和 $y \text{ isA } z$ 这两个已知的 isA 关系中提取一系列有效特征来判定传递性是否成立。
- 基本特征包括各实体/概念 x 、 y 、 z 的一些**统计信息**。
- 例如：其上(下)位词的数量、在语料中出现的频次等。

同样，对于 $x \text{ isA } y$ 和 $y \text{ isA } z$ 这两条已知的 isA 关系，也可以提取其边缘(即此关系在语料中出现的频次)、关系两端实体/概念的点互信息量(Pointwise Mutual Information)等特征。

特征

- 特征2：来自于同类实体的信息。

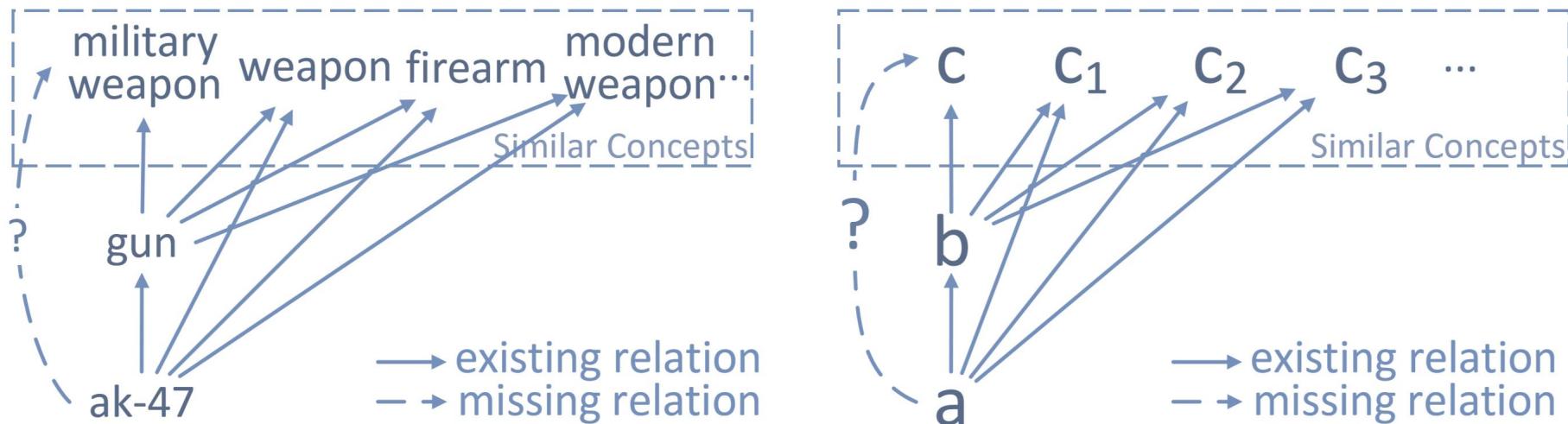
- 为了判断<Einstein, Physicist, Scientist>的传递性是否成立，可以观察与 Einstein 类似的实体，如 Newton、Faraday 相应的传递性。
- 由于 Newton、Faraday 等实体都同时具有 Physicist 和 Scientist 等上位词，因此有理由相信 Einstein 的 isA 关系也可以通过 Physicist 传递到 Scientist。



特征

- 特征3：来自于相似概念的信息。

- 对于三元组 $\langle x, y, z \rangle$ ，考虑与之相关的正例三元组 $\langle x, y, z' \rangle$ ，概念 z' 和 z 的相似度可以用来推断 $\langle x, y, z \rangle$ 的传递性。
- 例：为了判定三元组 $\langle ak47, gun, military weapon \rangle$ 的传递性是否成立，可以观察与 $military weapon$ 相似的概念 $weapon$ 、 $firearm$ 等，这些概念与 $ak47$ 、 gun 构成的三元组传递性均成立，因此可判定三元组传递性成立。



判定isA传递性成立：特征

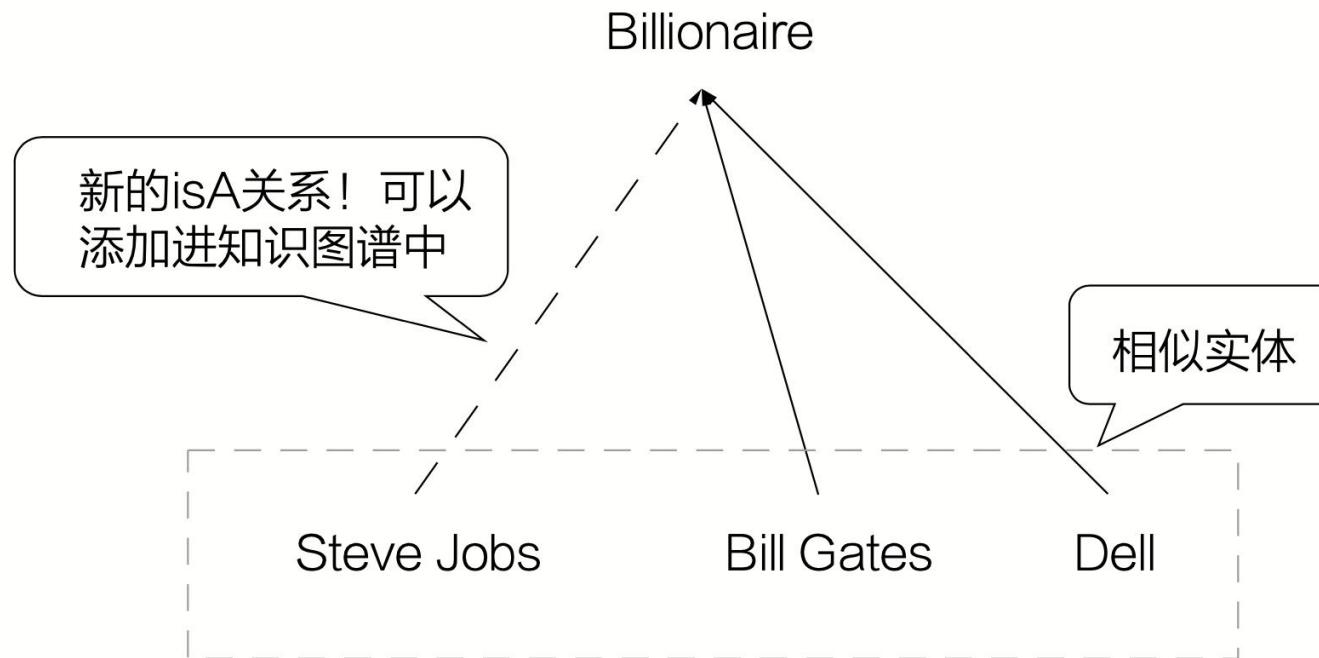
- 特征3：中间词的歧义性
 - 中间词的意思越多，传递性越有可能不成立
 - 使用WordNet获取三元组中间词的意思个数
 - 若该词在WordNet中，直接获取其意思的个数
 - 若该词不在WordNet中，说明它是低频词，一般只有一个意思
 - 另外，排除掉作为某特定实体的歧义
 - 三元组 $\langle a, b, c \rangle$ 的中间词b一定拥有实体a作为其下位词，故b不可能为一个底层实体

$$sc_b(t) = \begin{cases} synsets(b) - \theta(b) & b \in \text{WordNet}; \\ 1 & \text{otherwise.} \end{cases}, t = \langle a, b, c \rangle$$

基于协同过滤思想的概念图谱补全

- 相似实体拥有类似的上位词

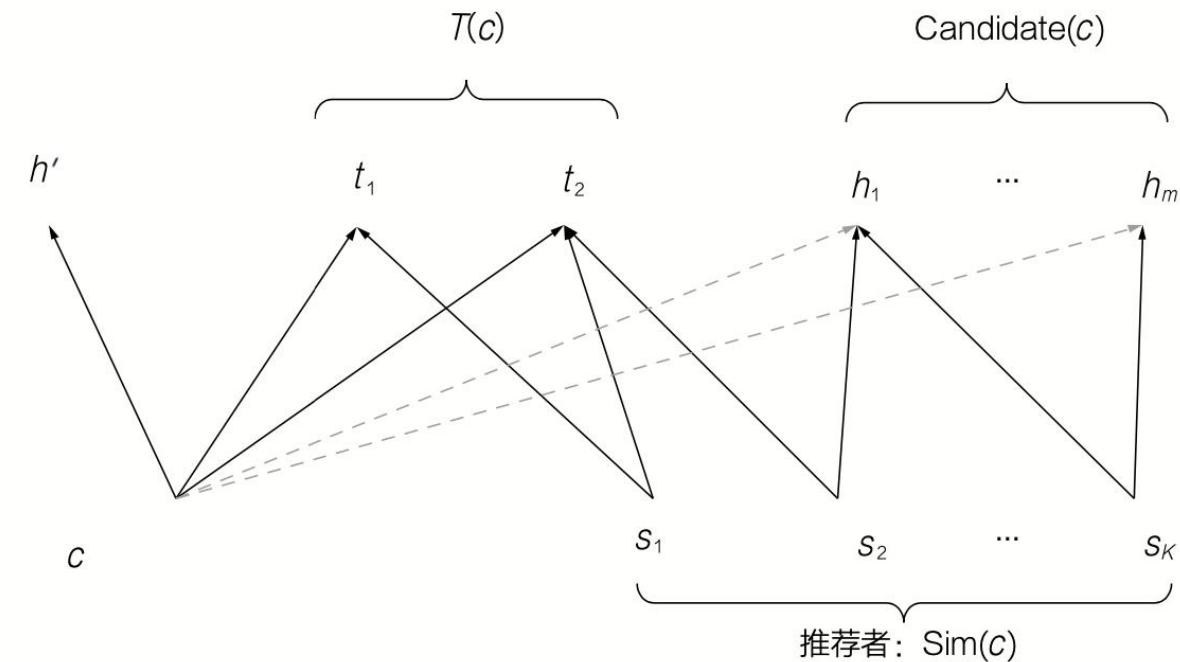
- 在考虑实体“Steve Jobs”时，人们很容易联想到其他相似的人物，如“Bill Gates”等。由于后面这些实体都属于“Billionaire”这一概念，因此可以推测“Steve Jobs”也属于“Billionaire”。



基于传递性进行补全的方法能找到大量的 new isA 关系，但是这种方法只适用于存在一个中间“桥梁”概念的 isA 关系，因而具有一定的局限性。

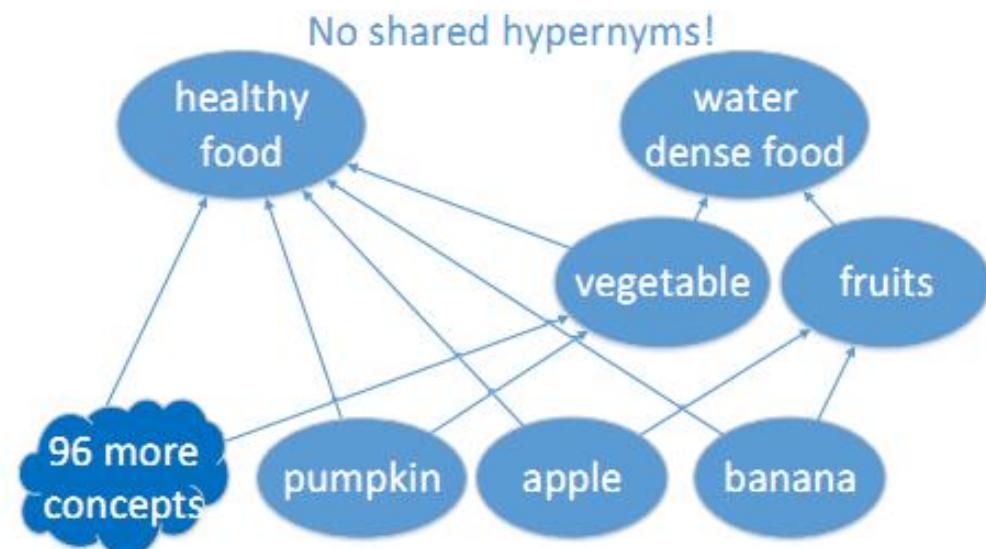
基于相似实体的图谱补全：框架

- 框架：协同过滤
 - 原理：相似的实体很有可能拥有类似的概念
- 目的：每个目标实体 c 寻找其缺失的上位词
 - 步骤1：在概念图谱中寻找 c 的相似概念或实体集合 $\text{Sim}(c) = \{s_1, s_2, \dots\}$ ，这一过程的核心是相似度计算。
 - 步骤2：从 $\text{Sim}(c)$ 的上位概念集合中，寻找 c 的候选缺失上位词。这一过程的核心是计算上位词的推荐分数。



相似度计算

- 首先要为目标实体 c 寻找与其相似的实体/概念集合 $\text{Sim}(c)$
 - 先要定义计算两个实体/概念相似度的函数 $\text{sim}(c_1, c_2)$
 - $\text{sim} = \mathbf{f}$ (Jaccard metric, Random walk metric)
 - Jaccard metric: 高精确度, 直接考虑两个实体间的共同上下位概念/实体
 - RW metric: 高召回率, 挖掘图谱中的远程关系
 - 右图: healthy food 和 water dense food
 - 只有很少的共同上下位概念/实体
 - 但是关联仍然非常紧密
 - 不能只使用简单直接的 Jaccard 相似



相似度计算

- 首先要为目标实体 c 寻找与其相似的实体/概念集合 $\text{Sim}(c)$
 - 先要定义计算两个实体/概念相似度的函数 $\text{sim}(c_1, c_2)$
- 常用方法：Jaccard 相似度
 - 对于集合 A, B , 其 Jaccard 相似度为 $J(A, B) = |A \cap B| / |A \cup B|$, 之后需要对这两个 Jaccard 相似度进行合并

由于概念图谱通常非常稀疏，两个相似性度量结果往往比真实的相似度偏小。因此，通常采用 Noisy-Or 模型合并各个相似性度量结果以增强信号：

$$N(x, y) = 1 - (1 - x)(1 - y).$$

显然，合并的结果不会小于被合并的各个分量。利用 Noisy-Or 模型增强信号是稀疏图谱中的常用做法，有一定的代表性。

随机游走相似度

- 随机游走相似度：
 - 分别计算两个实体为起点的随机游走向量
 - 计算两个向量的Cosine相似度
- 实体/概念c的随机游走向量：
 - 维度为 $2N$ 的向量， N 为图谱中的节点数‘
 - 每 N 维为以c为起点，按上位/下位方向随机游走后落到每一个节点的概率
 - 计算时，模拟走 L 步即可
 - 实验表明 $L=2$ 即满足要求

基于随机游走的度量可以充分利用全图信息，也就是说可以利用非直接邻接的远程关系的信息，从而更加有效地利用图中的结构信息放大稀疏图谱中的微弱信号。

协同过滤：合并相似度

- 使用WordSim353–similarity作为开发集
 - $F_{2.36}$ 为最好的合并算法
 - 右图一些例子表明了对数个实体找到的前k个最相似实体

$$\begin{aligned} cs(c_1, c_2) &= F_\beta(jacc(c_1, c_2), rw(c_1, c_2)) \\ &= \frac{(1 + \beta^2) \cdot jacc(c_1, c_2) \cdot rw(c_1, c_2)}{\beta^2 \cdot jacc(c_1, c_2) + rw(c_1, c_2)} \end{aligned}$$

Term	Similar terms
facebook	twitter, linkedin, myspace, flickr, digg
python	perl, ruby, visual basic, tcl, basic
haskell	erlang, standard ml, scheme, ocaml, lisp
iphone	ipod touch, apple iphone, smartphones, psp, smart phone
microsoft windows	mac os, windows xp, windows, windows 95, mac os x
warcraft	starcraft, warcraft iii, company of heroes, age of empires, half life

上位概念推荐

- 对于实体 c ，根据 $\text{Sim}(c)$ 的上位词可以构建待推荐的上位词候选集

$$H = \{h | s_i \text{ isA } h, s_i \in \text{Sim}(c), c \text{ isA } h \text{ not in KB}\}$$

- 集合 H 中分数较高的上位词 将作为 c 的新上位词。

$$\text{Score}(h_j) = \sum_{s_i \in \text{Sim}(c)} w(s_i \text{ isA } h_j) \text{sim}(s_i, c)$$

其中，每个累加项包含相似项 s_i 和 c 的相似程度 $\text{Sim}(s_i, c)$ ，以及相似项 s_i 与上位概念 h_j 的 isA 关系权值 $w(s_i \text{ isA } h_j)$ 。基于 $\text{Score}(h_j)$ ，对每一个 c ，选择有较高得分的新上位概念补充给 c 即可。

在 Probase 中，某条 isA 关系的权值就是该 isA 关系在语料中被观察到的次数。使用协同过滤框架最终能为 Probase 添加接近 500 万条新的 isA 关系，抽样检测表明其正确率超过 85%。

isA关系纠错

概念图谱错误成因

来自语料中的错误

- 不能从字面意思直接理解的修辞(如反话、比喻、抽象等)
- 错误的句子、不当的表达甚至笔误

句子…Paris is such as(an)
exciting city
笔误, an写成了as,
符合Hearst模板, 得到错误的
Exciting city isA Paris

来自抽取方法的错误

- 依赖于大量NLP工具, 错误会累积

来自自动推理的错误

- 自动推理技术本身效果未达100%
- 原来的概念图谱中存在错误, garbage in garbage out
- 存在大量的特例不能通过简单推理/归纳等技术产生

企鹅不会飞
即使是人也很容易错误地推断
企鹅不是鸟类

简单的想法：基于知识的支持度的纠错

- 每一条知识寻找支持它的证据，来“证明”其正确性
 - 出现某条知识的句子的数量作为这条知识的支持度
 - Probase中按不同支持度采样的结果如右表所示
 - 更高支持度的 isA 关系通常更可能是正确的
 - 低支持度的知识多，全部删除过于浪费且会误删
 - 可信度度量

支持度	占比	正确率
1	85.88%	78%
2-10	13.27%	86%
11-100	0.80%	94%
>100	0.05%	100%

可信度定义

- 前面提到的“**支持度**”可以作为很好的边可信度定义
 - 但 86% 的边拥有相同的支持度 1, 不具有区分度
- 额外的**启发式可信度**
 - 一个底层实体不应有下位词
 - 一个更具体的概念应该相比更抽象的概念含有更少的下位词
 - juice(173 hyponyms) isA tomato(69 hyponyms) → unreliable
 - exciting city(29 hyponyms) isA paris (9 hyponyms) → more unreliable
- 两指标之积作为最终可信度

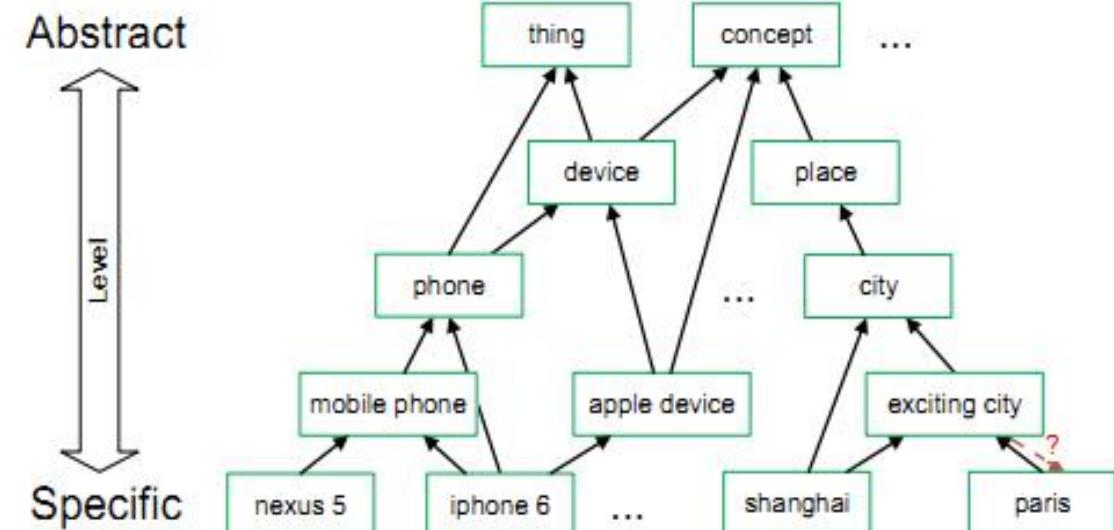
实例分析：Probbase中的错误

- 通过Case Study寻找常见的错误(左表)
 - 常见错误：一个较抽象的概念 isA一个较具体的实体
 - 一般而言，概念图谱应当是底部为具体实体，往上为抽象概念的形式（右图）
 - 抽象的概念isA具体实体可能导致图谱中产生环

Probbase中的部分错误

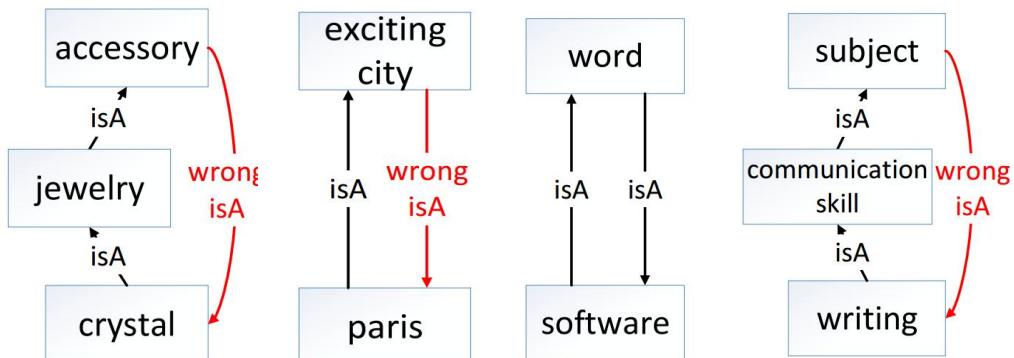
Entity	isA	Concept	Entity	isA	Concept
exciting city	isA	paris	battery	isA	fuel cell
automobile	isA	lead acid battery	cause	isA	tsunami
music video	isA	youtube video	sweet	isA	glucose
world cup	isA	football	grape	isA	purple
college	isA	basketball	juice	isA	tomato

Table 1: Examples of incorrect isA relations in Probbase



Probase中的错误与环

- 猜想：概念图谱中的环很有可能包含错误的边（isA关系）
 - 左图：Probase中的环的例子，除了第3个是由歧义造成，其余都是由于错误isA关系造成
 - 右图：对Probase中大小为2或3的环的采样测试
 - 超过95%的小环中包含错误isA关系
 - 环的存在可以定位其中的错误isA关系



Size	Have error	Null model	z-score	p-value
2	97%	15%	22.96	<0.0001
3	96%	24%	16.86	<0.0001

基于图模型的纠错

- 在概念图谱中消环
- 问题定义
 - Input: 图 $G(V, E)$
 - Output: 包含错误边的集合 E'
 - Constraint:
 - $G(V, E - E')$ 是一个有向无环图 DAG
删除边后的图应当不存在环
与人们对概念图谱的树形层次直觉相符
 - Minimize $\sum_{e \in E'} w(e)$, 其中 $w(e)$ 是 e 的可信程度, 可信度定义参照前面
输出错误边集 E' 应当尽可能包含不可信的边

概念图谱中消环

- 给定有向图 $G(V, E)$, 可信度函数 $w: E \rightarrow \mathbb{R}$ 是定义在边上的实数权重。求边集 E' , 使 $G(V, E-E')$ 为有向无环图, 且 $\sum_{e \in E'} w(e)$ 最小
 - \rightarrow 带权 MFAS 问题 NP-HARD
- 贪心算法
 - Step 1:
 - 随机顺序枚举图中的每个环, 每次找到一个环, 将环中最小权值的边全部删除, 直到图中不存在环为止。
 - Step 2:
 - 将前一步中删除的边按权值从大到小排序, 逐个尝试。
 - 若当前被删除的边加回图中不会产生环, 则将其加回图中。否则删除这条边作为最终输出的一部分。

结论

- 本章主要介绍了概念图谱
 - 一类有着广泛用途的，主要包含isA关系的知识图谱
 - 概念图谱可以用于查询各种实体或概念的从属关系，以支撑概念化、推理、归纳等智能应用。
- 人工构建的概念图谱虽然拥有很高的精度，但是其规模过小，不能覆盖实际情况中的大量实体和概念。
- 从大规模语料中自动构建的概念图谱拥有更大的规模和可接受的准确度。
- 本章介绍了一系列构建大规模概念图谱的方法。
 - 从大规模的互联网语料中抽取isA关系的方法
 - 对初步构建完成的概念图谱进行补全的方法
 - 对初步构建完成的概念图谱进行清洗的方法

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. "Probbase: 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. "Probbase+: 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.