《知识图谱: 概念与技术》

第 4 讲
概念图谱构建

肖仰华

复旦大学

shawyh@fudan.edu.cn

本章大纲

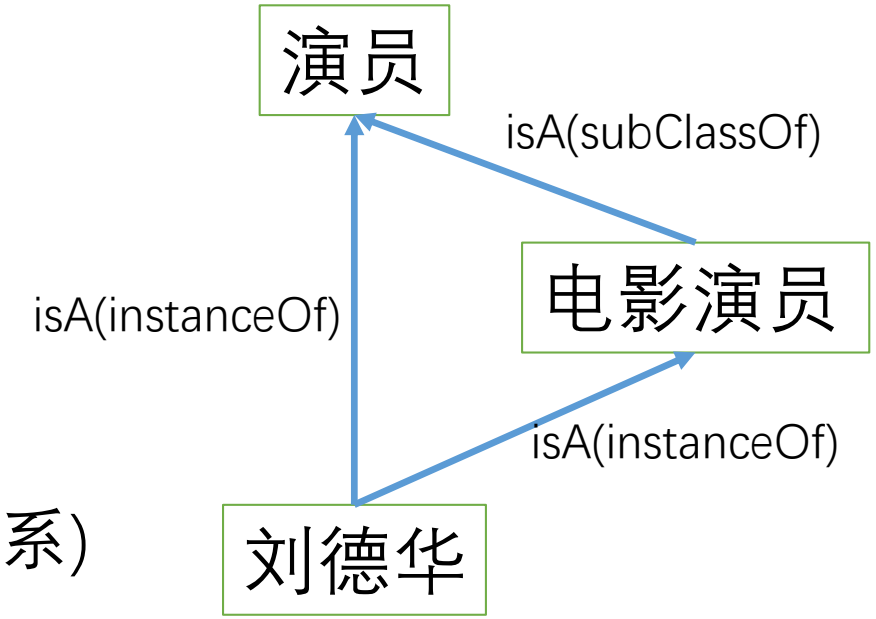
- 概念图谱概述
- isA关系抽取
- isA关系补全
- isA关系纠错

概念图谱概述

概念图谱

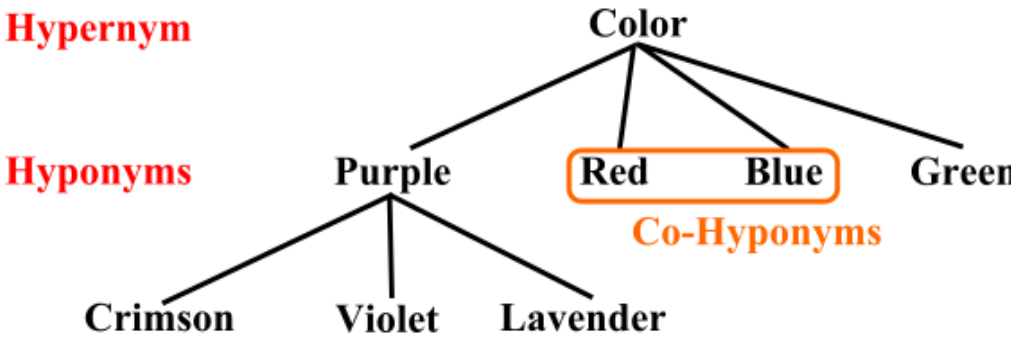
- 概念图谱的组成
  - 节点：实体、概念
  - 关系：实体与概念之间的类属关系(isA)、概念与概念之间的 subclass of 关系组成

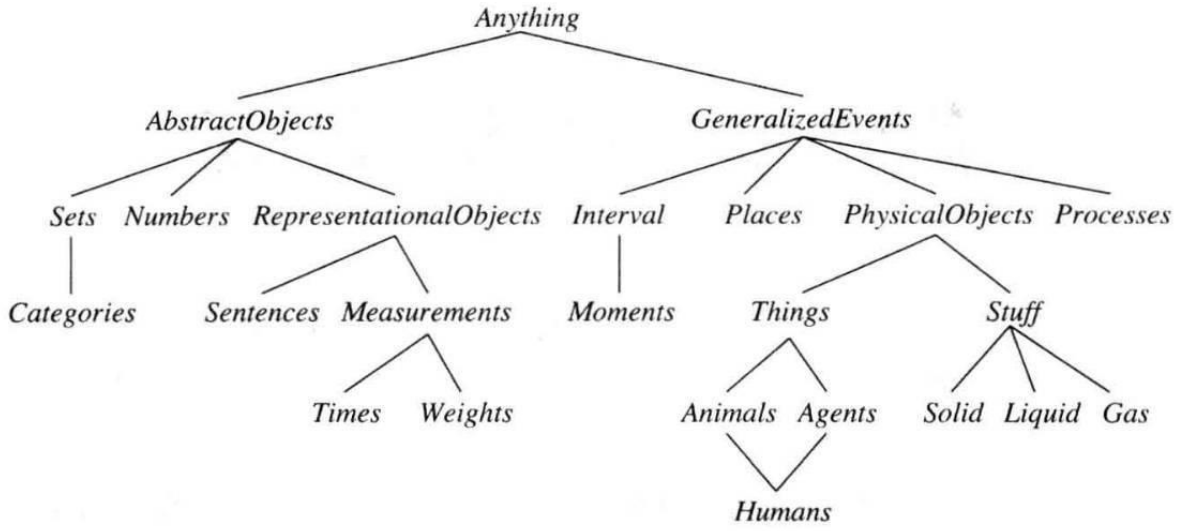
- 实体
  - 比如“刘德华”
- 概念
  - 比如“演员”
- 实体和概念之间的类属关系(isA 关系)
  - 比如“刘德华 isA 演员”
- 概念与概念之间的类属关系(subclassOf 关系)
  - 比如“电影演员 isA 演员”



上下位关系(Hypernym-Hyponym)

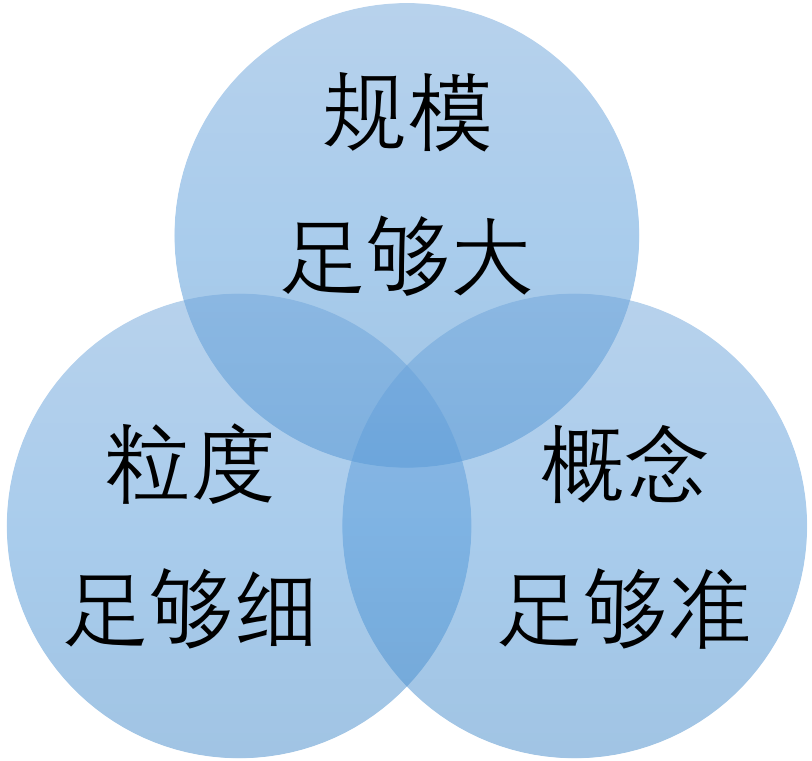
- 实体、概念通常用词汇(term)加以表达
- 实体与概念之间的类属关系以及概念与概念之间的子类关系，对应到语言学角度上下位关系
  - 如果 A isA B，通常称A为B的下位词（hyponym），或者B为A的上位词（hypernym）
- 由概念及其之间的subclass关系构成的有向无环图有时又成为Taxonomy，当实体与概念都用文字描述时，又通常称为lexical taxonomy





概念图谱的重要意义

- 概念是认知的基石
  - 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
- “理解”很多时候体现为产生概念
  - “Trump”-> American President
- 概念是人们解释现象常用的
  - 鲨鱼为何可怕？ 因为它是肉食动物



大规模概念图谱使得机器认知实体的概念成为可能

概念图谱的作用

实例化

列出属于这个概念下的一些典型的实体

Largest company:

- China Mobile
- Google

概念化

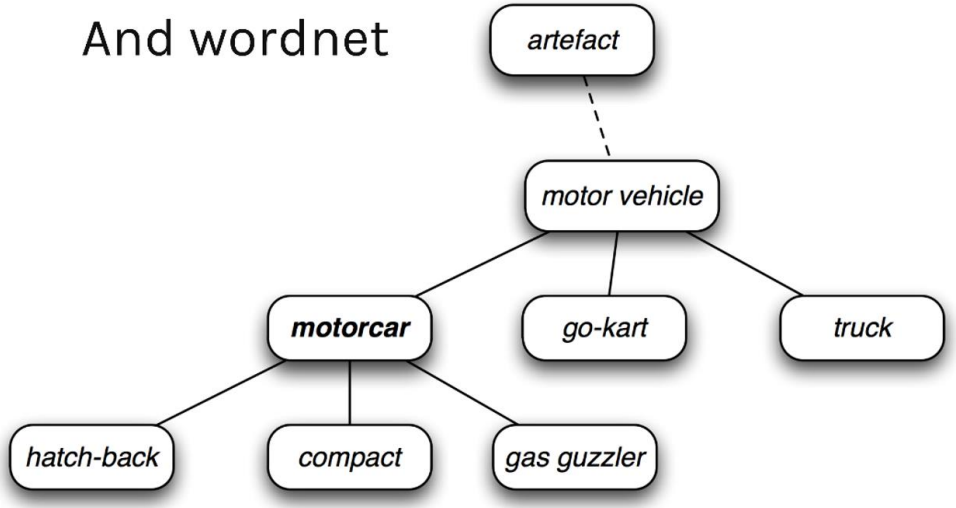
推断一个或一组实体所属的概念

Brazil, India, China:

- BRIC country
- Concept (too vague)

常见的概念图谱

- **WordNet**：普林斯顿认知科学实验室于1995年建立的英文词典
  - 专家构建，准确度极高
  - 实体按sense组织，已经过消歧
  - 规模较小，包含大约155287个单词(117659个词义或同义词集)



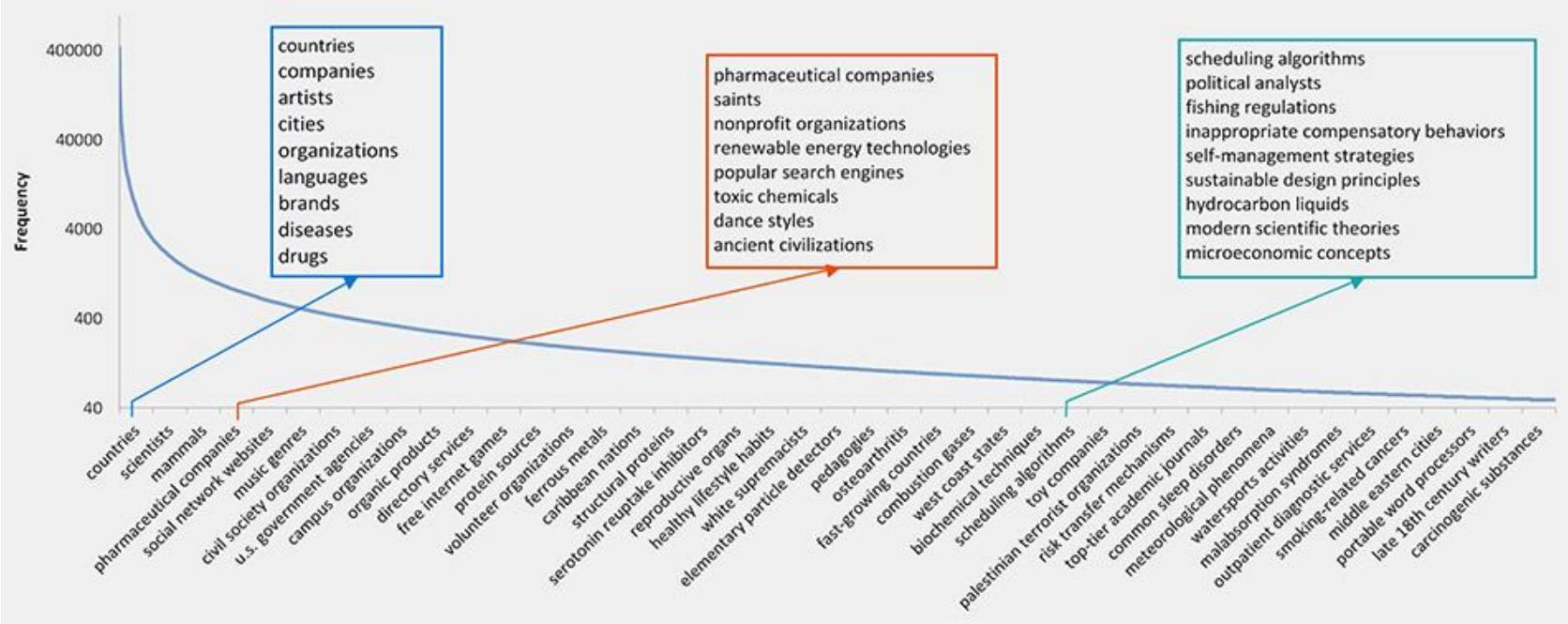
https://sourcedexter.com/find-synonyms-and-hyponyms-using-python-nltk-and-wordnet%E2%80%8B/

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。

常见的概念图谱

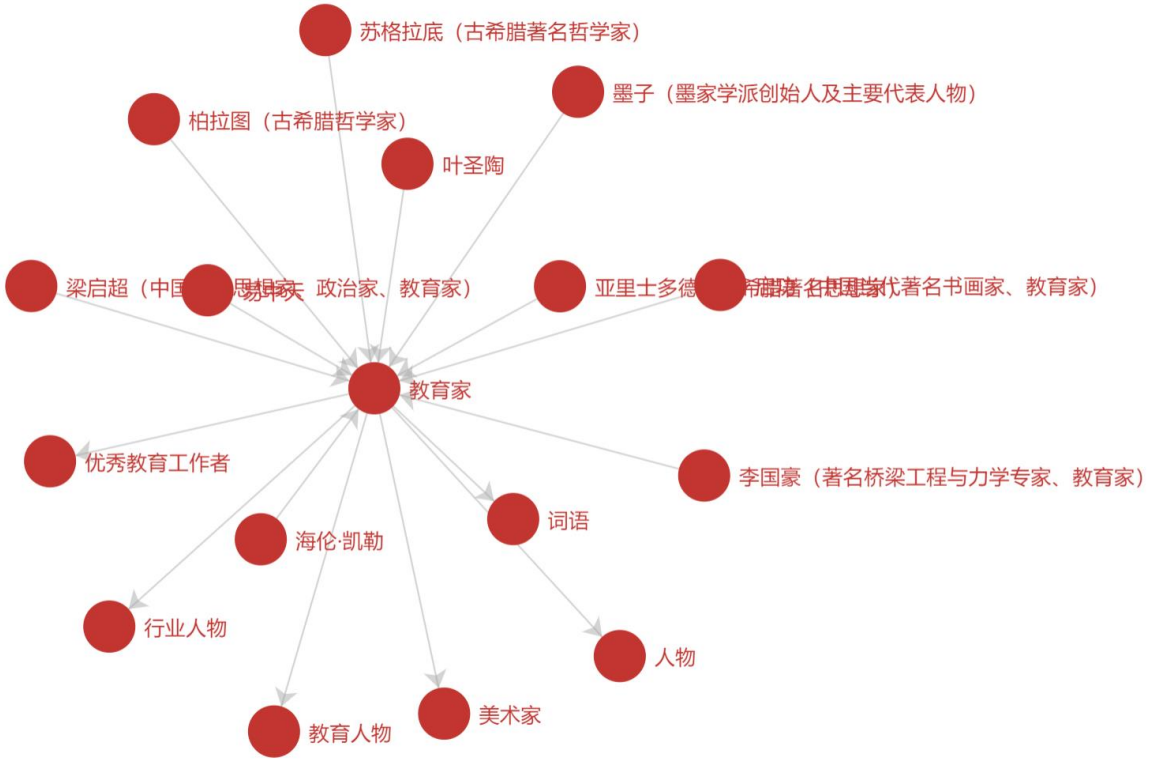
- **Probase：**2012年微软公司提出的研究原型
  - 从网页数据和搜索记录数据构造
  - 包含5,401,933个概念，12,551,613个实例和87,603,947个IsA关系
  - 现已更名为Microsoft Concept Graph



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

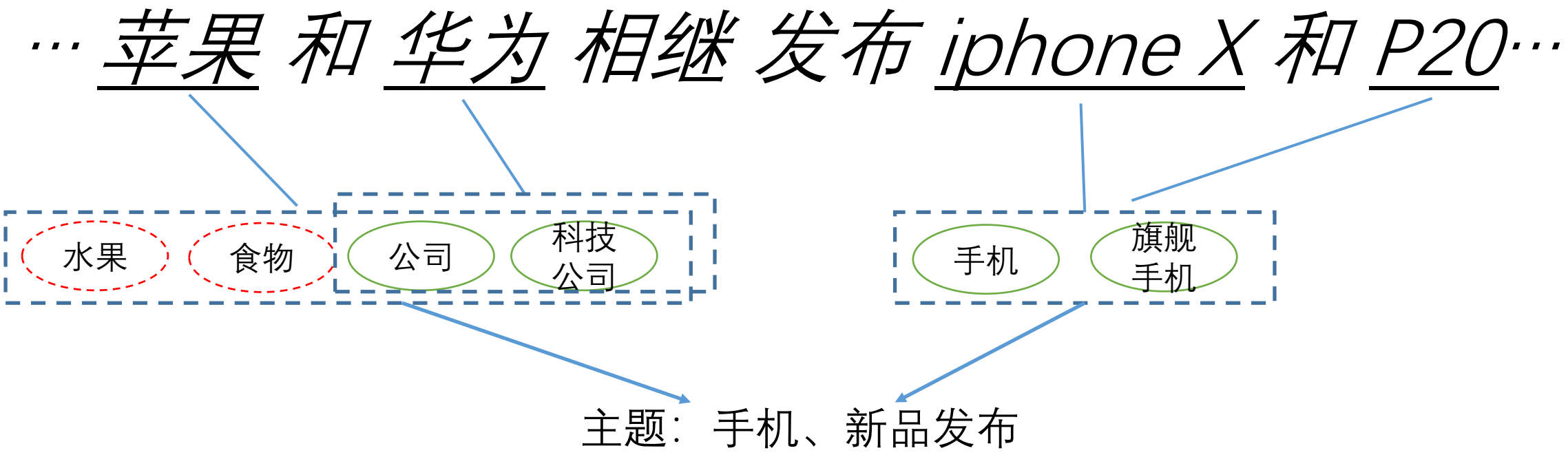
常见的概念图谱

- **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/

概念图谱的应用：主题理解

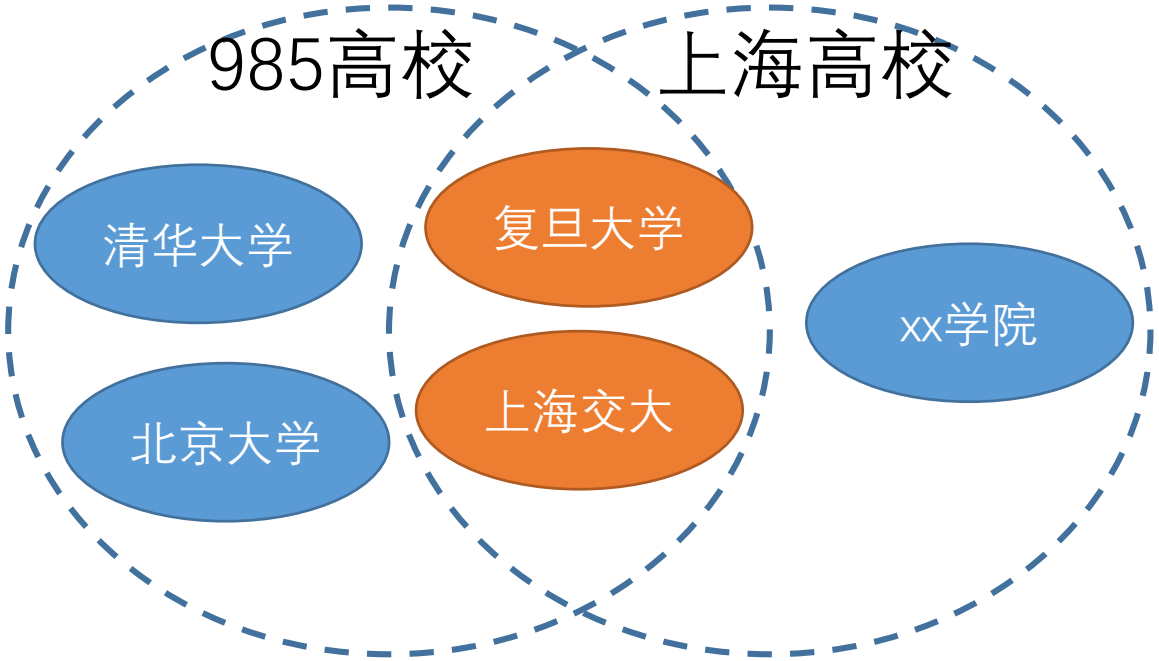


概念图谱的应用：实体搜索

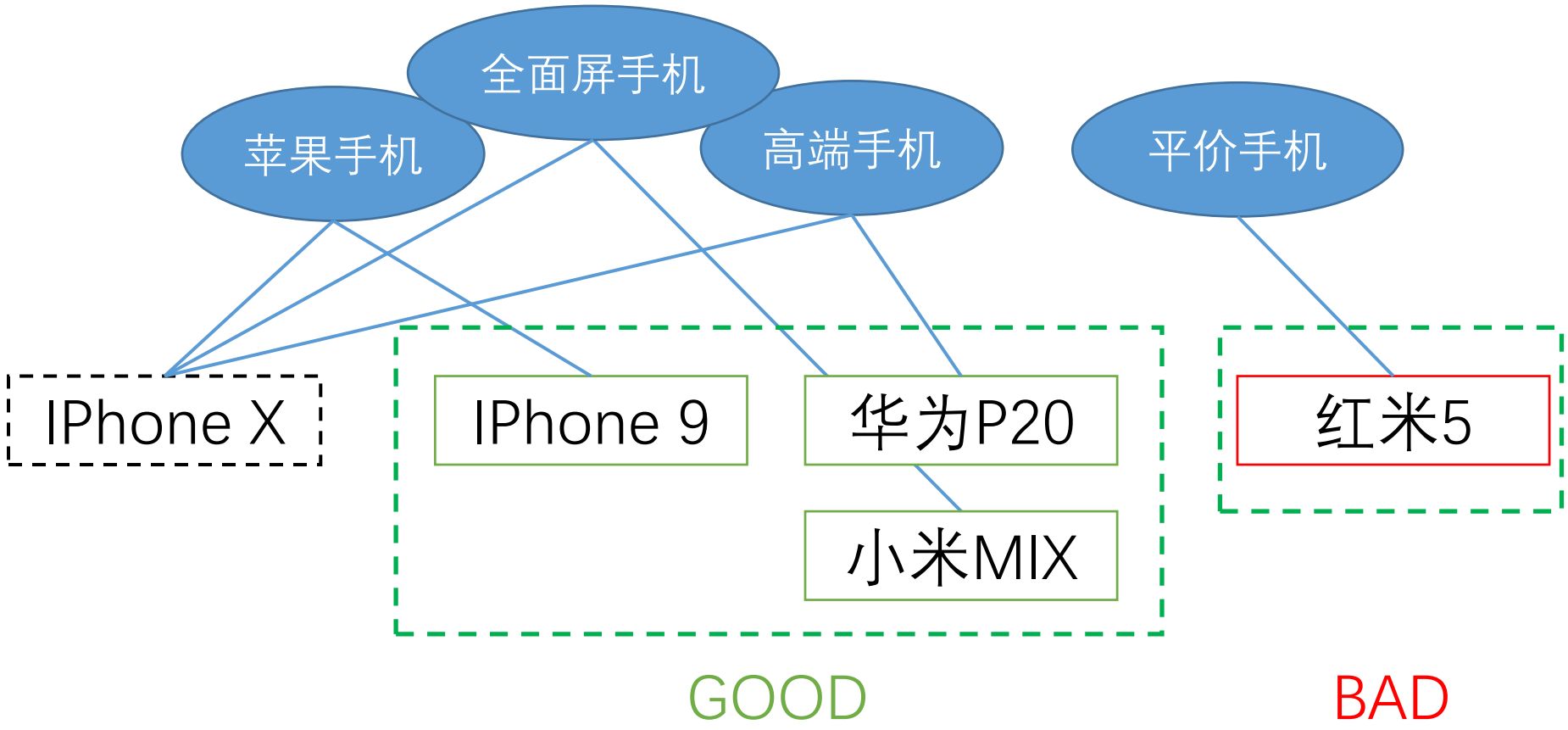
Query:水果

苹果	更常见
雪梨	
西瓜	
……	
蛇果	更稀有
山竹	
……	

Query:上海的985高校



概念图谱的应用：实体推荐



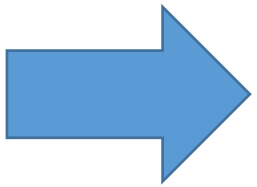
概念图谱的应用：语言概念模板

- **语言概念模板**

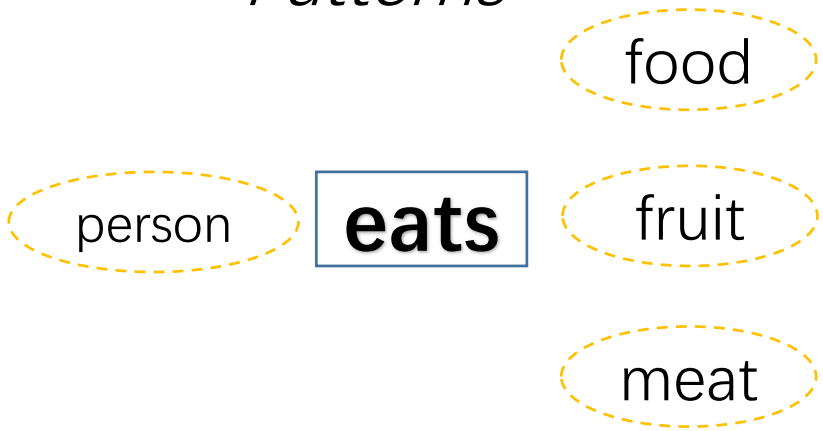
Learning

Corpus

Alice **eats** an apple
Bob **eats** a pear
Mary **eats** a pie
John **eats** a lemon



Patterns



Inference

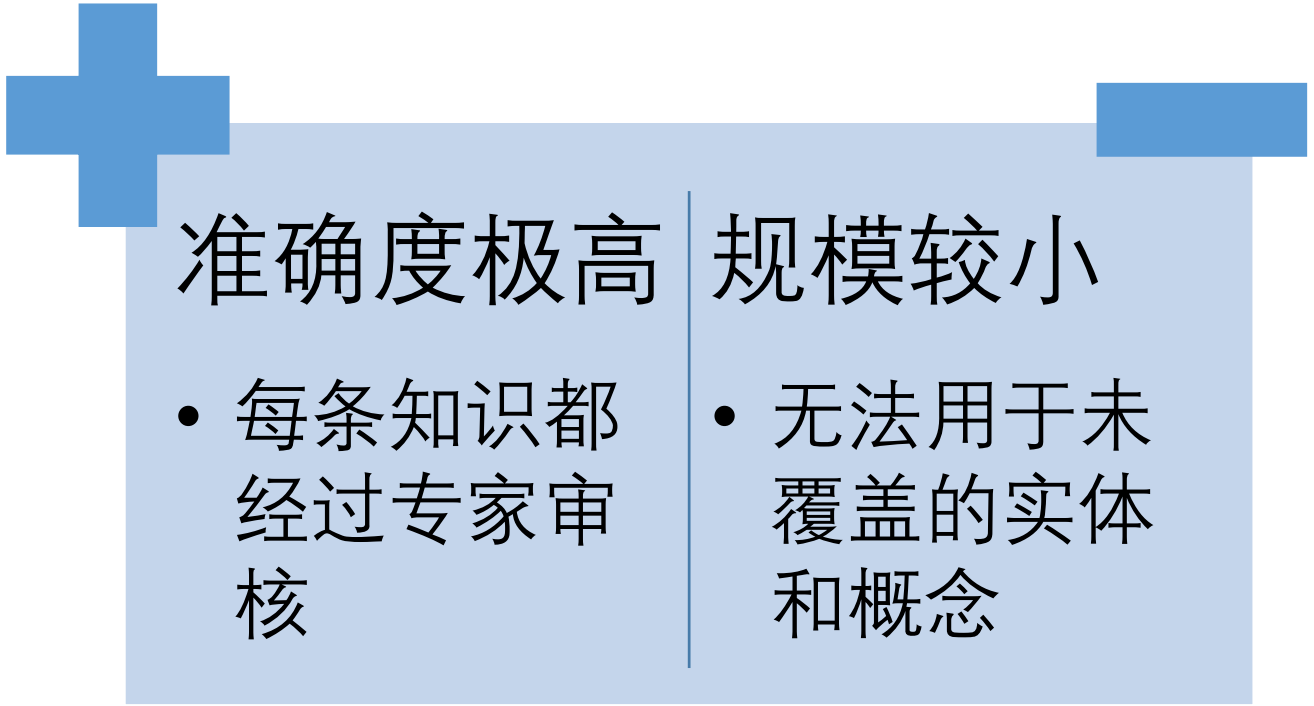
*Mary **eats** an avocado*

Avocado: food 99%
fruit 75%

isA关系抽取

为什么要抽取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

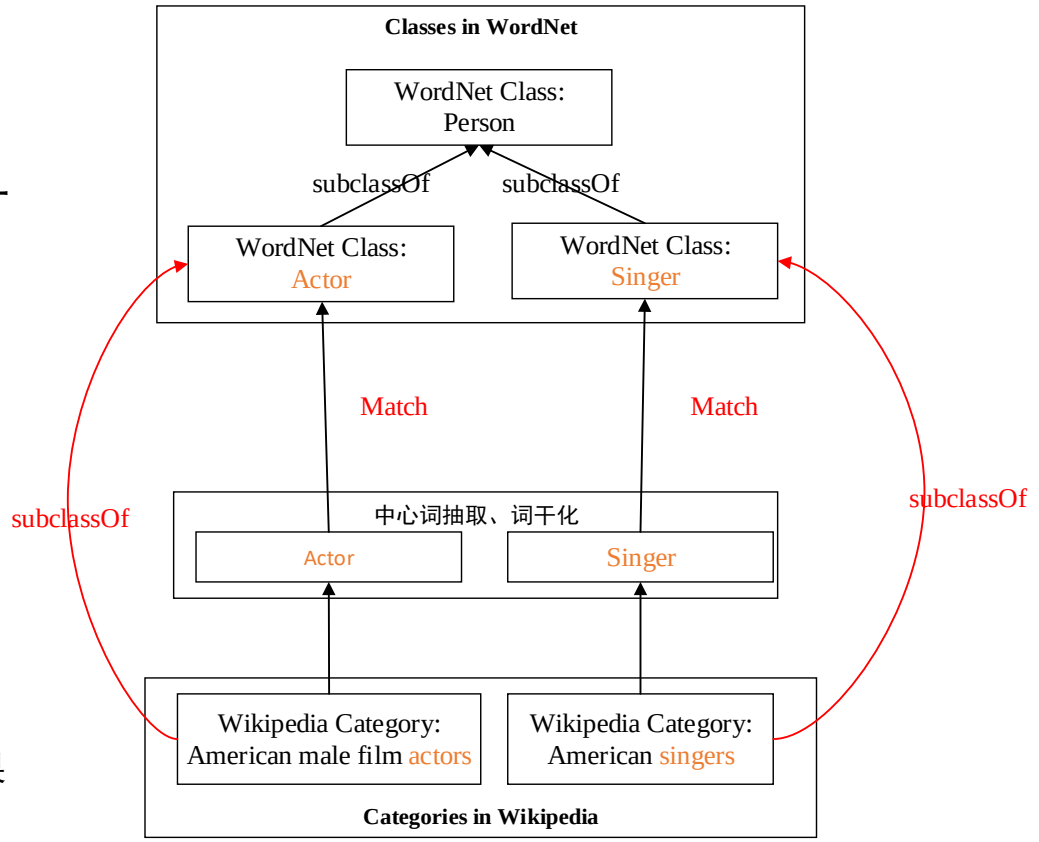
基于Embedding的方法

- 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
  - 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

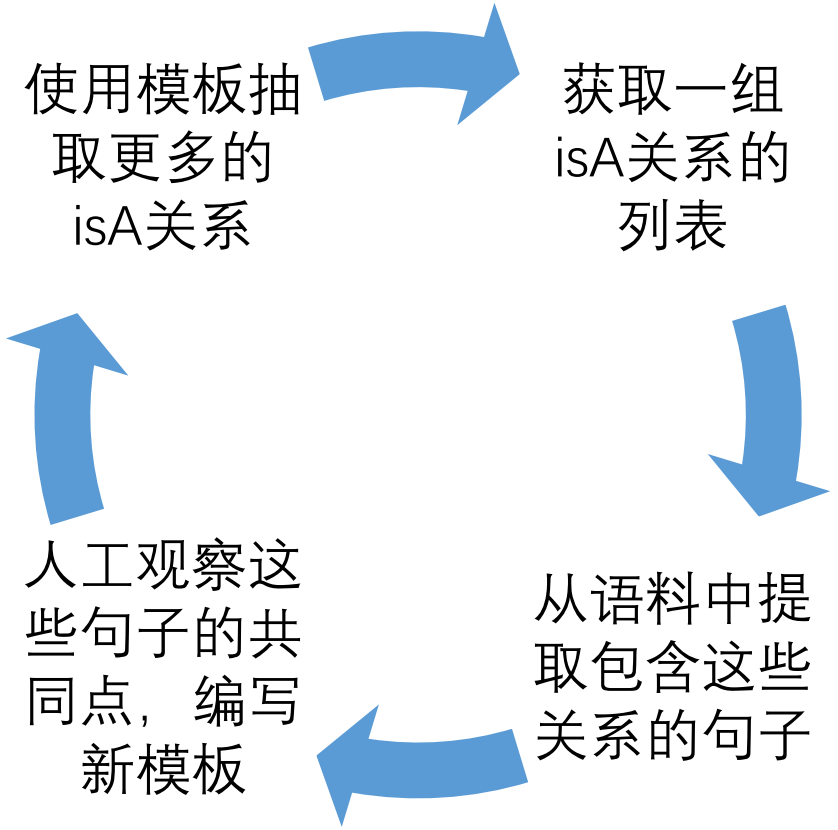
ID	Pattern
1	*NP* such as {*NP*,}*{(or | and)} *NP*
2	such *NP* as {*NP*,}*{(or | and)} *NP*
3	*NP*{,} including {*NP*,}*{(or | and)} *NP*
4	*NP*{,*NP*}*{,} and other *NP*
5	*NP*{,*NP*}*{,} or other *NP*
6	*NP*{,} especially {*NP*,}*{(or | and)} *NP*

- … 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={animals, dogs}，候选实体集合Y={cats}

只选择1个候选概念
p(animals|cats) >> p(dogs|cats)

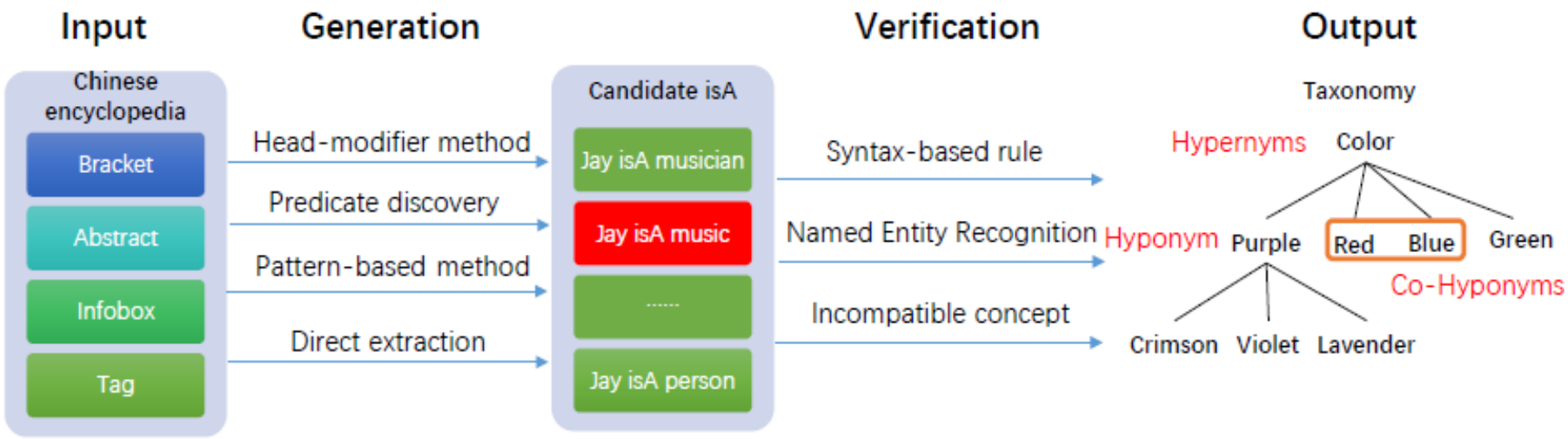
- cats isA animals? GOOD
- cats isA dogs? BAD

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

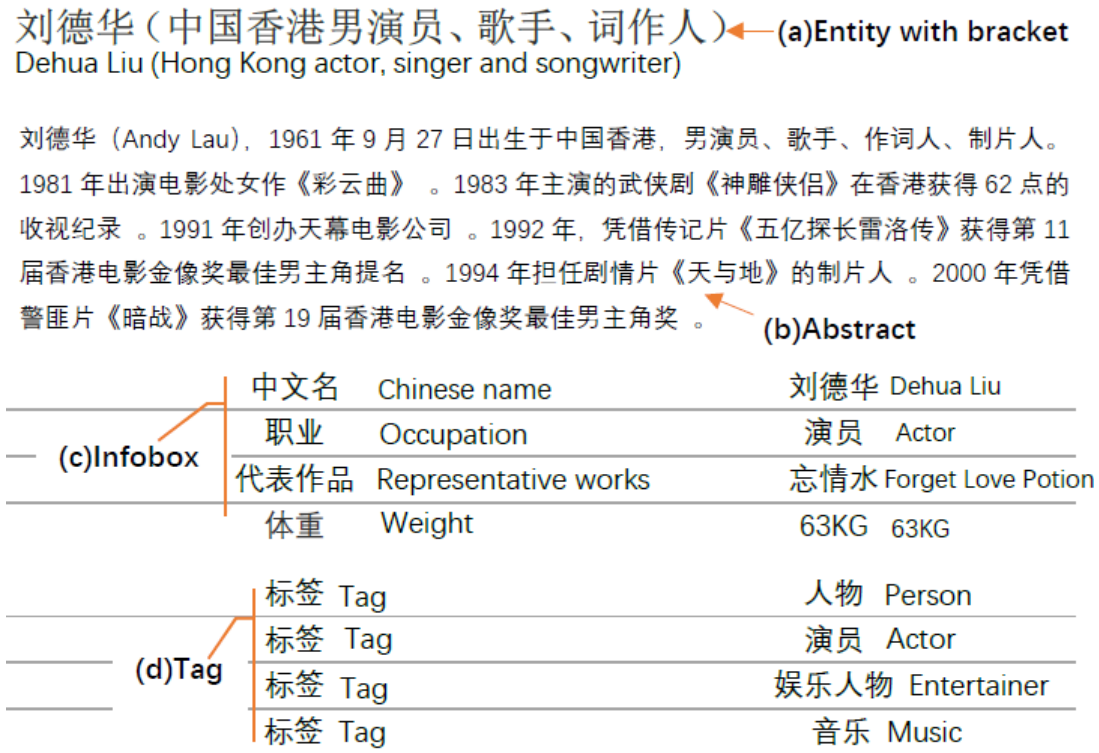
中文isA关系抽取：CN-Probase

- 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

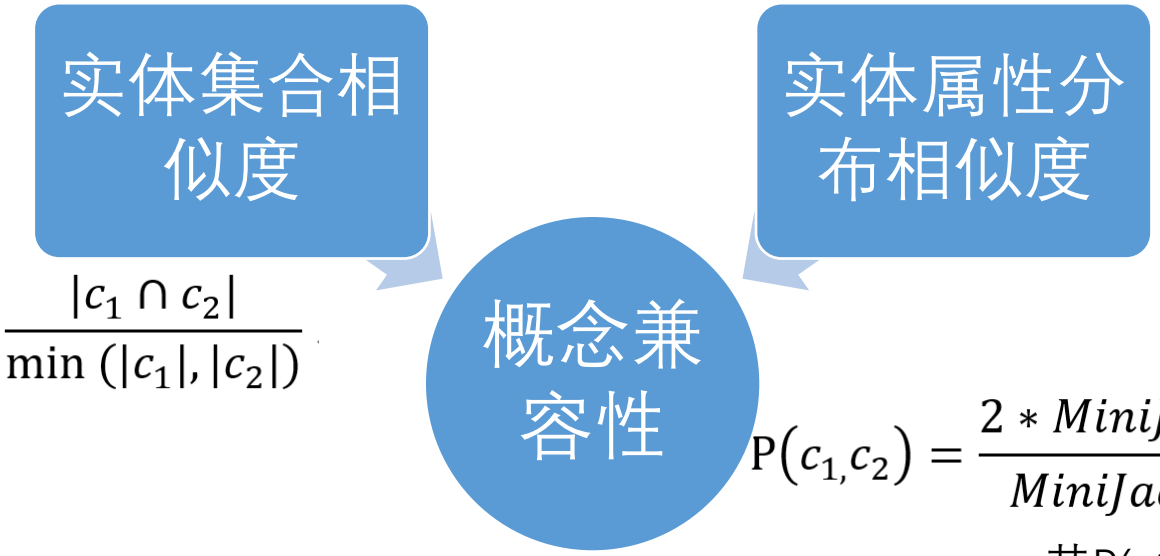
实体括号	• 刘德华isA 歌手
摘要	• 刘德华 isA 制片人
Infobox	• 刘德华 isA 演员
标签	• 刘德华 isA 娱乐人物



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



$$\mathrm{MiniJaccard}(c_1,c_2)=\frac{|c_1\cap c_2|}{\min(|c_1|,|c_2|)}$$

$$\mathrm{CPD}(c_1,c_2)=\frac{x\cdot y}{||x||\cdot||y||}$$

x,y分别为c1,c2的实体所拥有的属性的分布

$$\mathrm{P}(c_1,c_2)=\frac{2*MiniJaccard(c_1,c_2)*CPD(c_1,c_2)}{MiniJaccard(c_1,c_2)+CPD(c_1,c_2)}$$

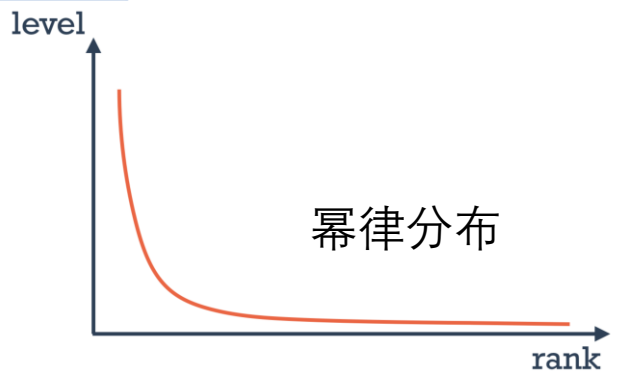
若P(c1,c2)<阈值，则c1,c2为互斥概念

isA关系补全

概念图谱知识缺失的成因

- 有大量的正确isA关系并没有出现在抽取的概念图谱之中
  - Probase中平均每个实体/概念仅有**1.6**个关系

常识相关	低频实体
• 许多常识性的isA关系不会在语料中以书面表达方式出现 • 如：Steve Jobs是一个亿万富翁	• 大多数实体在语料中出现的频数非常低（幂律分布） • 如：Tesco仅有Big UK supermarket一个概念

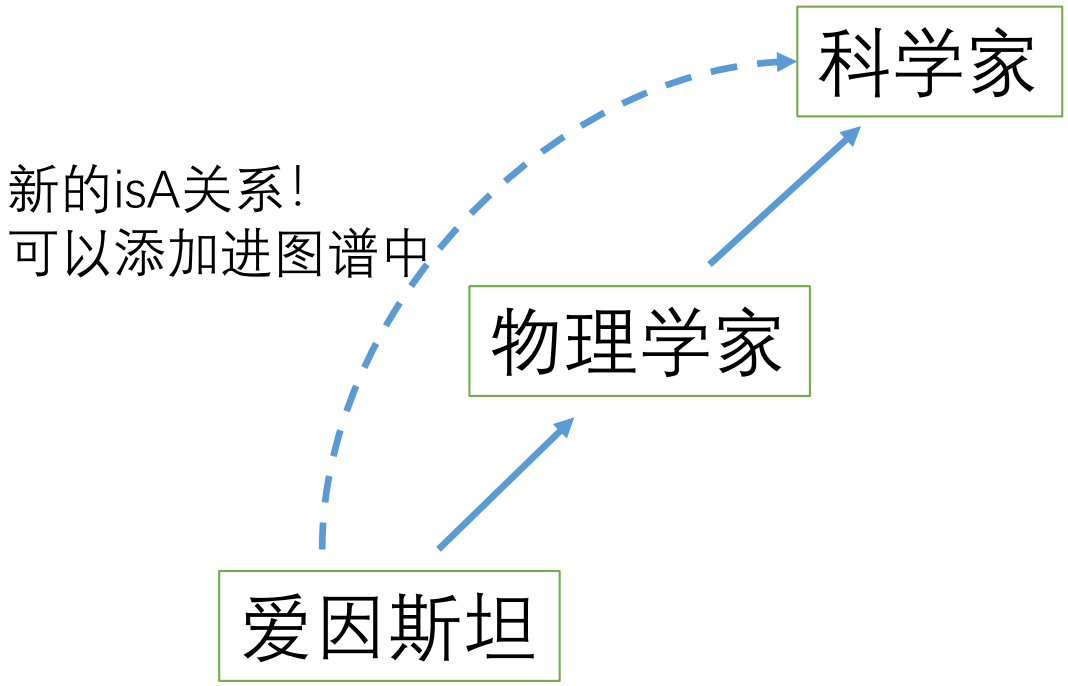


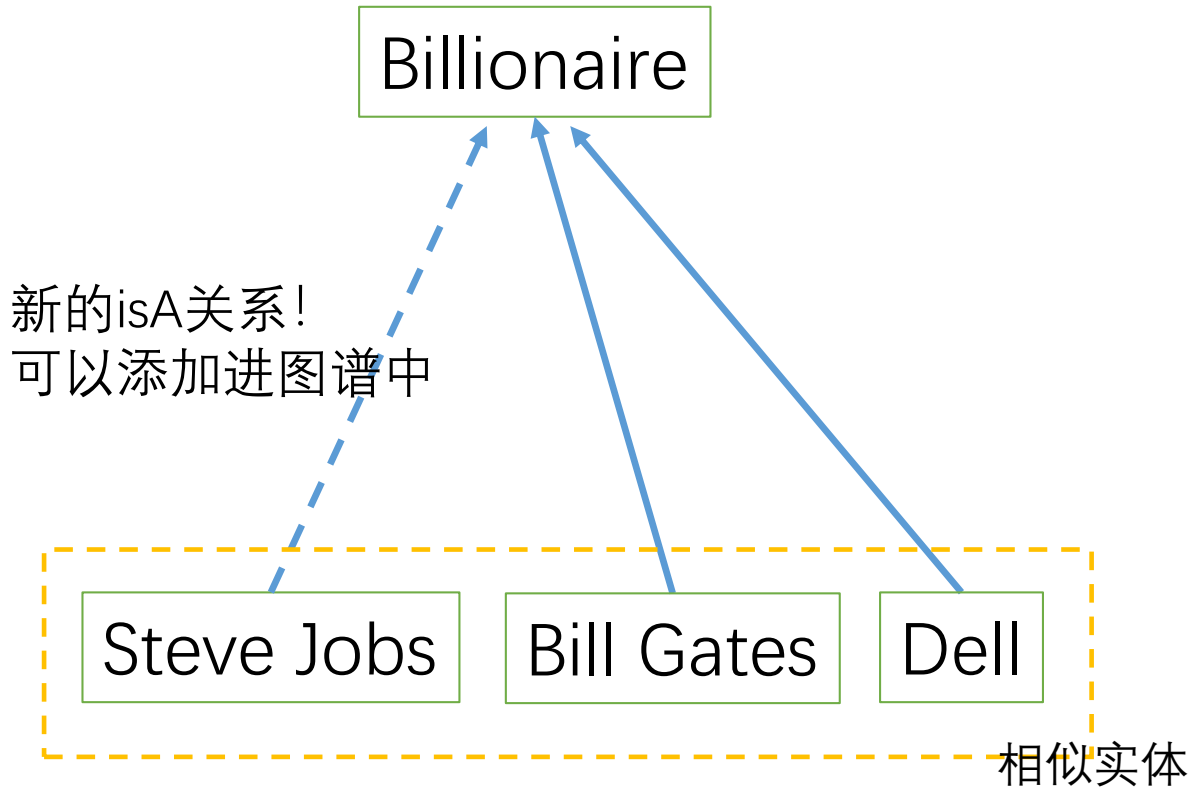
增加语料无法很好解决以上问题！

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

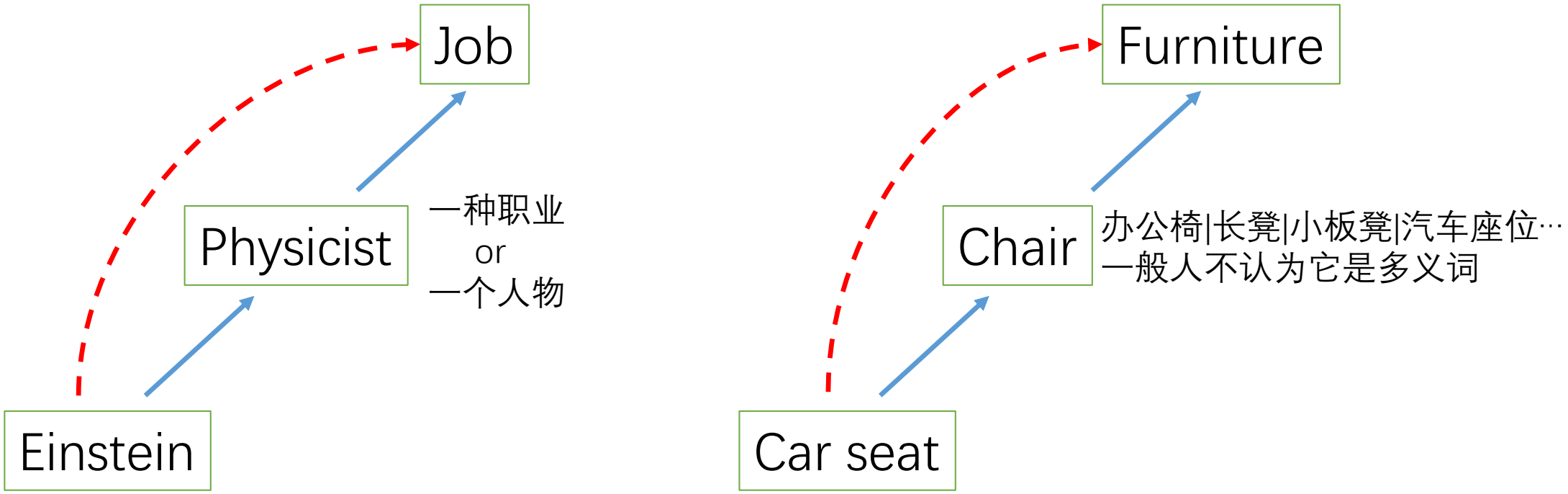
概念图谱补全：方案

- 基于isA关系的传递性推理
- 基于相似实体的信息推理





传递性并不总是成立



无法以消除歧义的方法来让Probase这样的大规模概念图谱变得和WordNet一样规整!

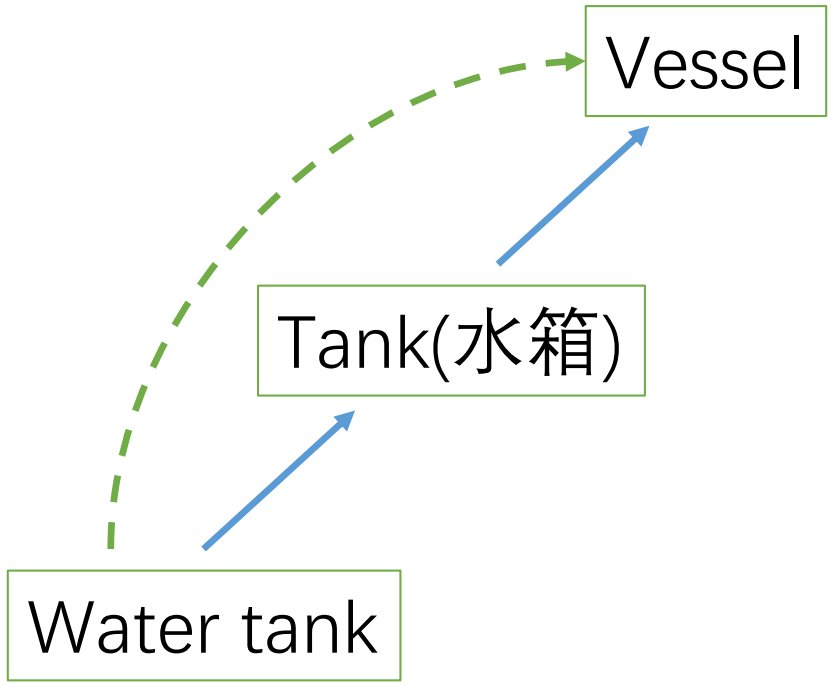
利用isA传递性进行图谱补全

- 问题：**isA传递性在什么情况下成立？**
  - 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<x,y,z>，isA传递性成立，x isA y，y isA z，则补全 x isA z

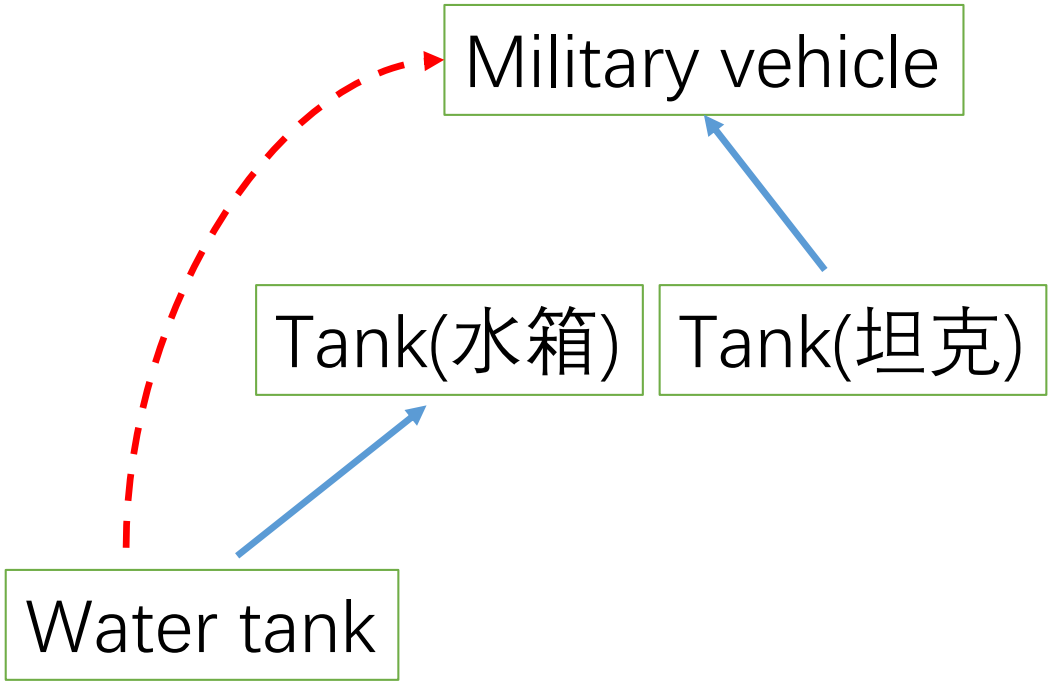
- 机器学习二分类问题：isA传递性成立(positive)与不成立(negative)
  - 标注数据
  - 特征
  - 模型：Random Forest

判定isA传递性成立：标注数据

- **标注数据：**
  - WordNet：经过消歧的、专家构建的、isA自然传递性的概念图谱



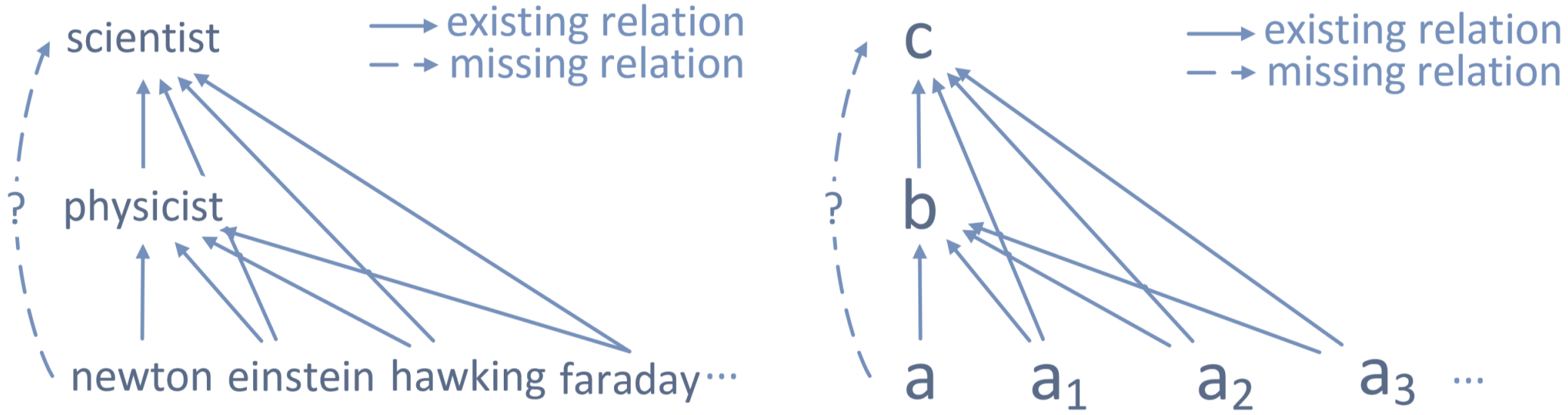
Positive: water tank - tank - vessel



Negative: water tank - tank - military vehicle

判定isA传递性成立：特征

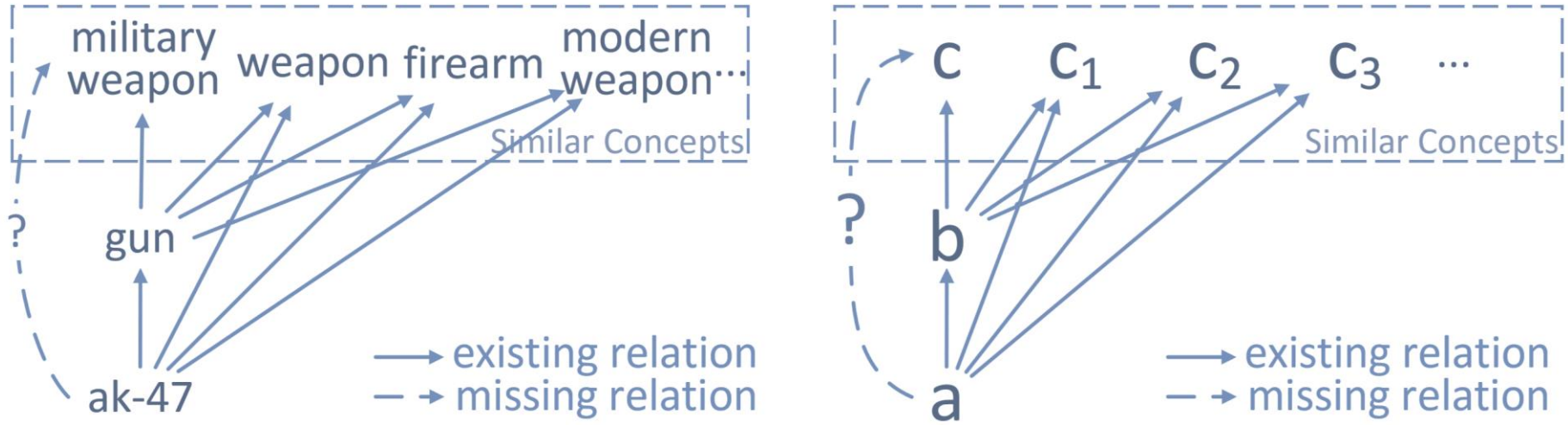
- 特征1：来自于同类实体的信息。
  - Einstein - Physicist - Scientist
  - Einstein ~ Newton，　Einstein ~ Faraday



$$sib_r(t) = \frac{|hypo(\mathtt{b}) \cap hypo(\mathtt{c})|}{|hypo(\mathtt{b})|}, t = \langle \mathtt{a}, \mathtt{b}, \mathtt{c} \rangle$$

判定isA传递性成立：特征

- 特征2：来自于相似概念的信息。
  - Ak47 - gun - military weapon
  - ak47 isA weapon, gun isA weapon: Sim(weapon, military weapon)



$$sim(t) = \frac{\sum_{c_i \in hype(\mathtt{a},\mathtt{b})} sim_c(\mathtt{c}, c_i)}{|hype(\mathtt{a},\mathtt{b})|}, t = \langle \mathtt{a},\mathtt{b},\mathtt{c} \rangle$$

$$sim_c(c_1, c_2) = 1 - (1 - s_\epsilon(c_1, c_2)) \times (1 - s_o(c_1, c_2))$$

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<a,b,c>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

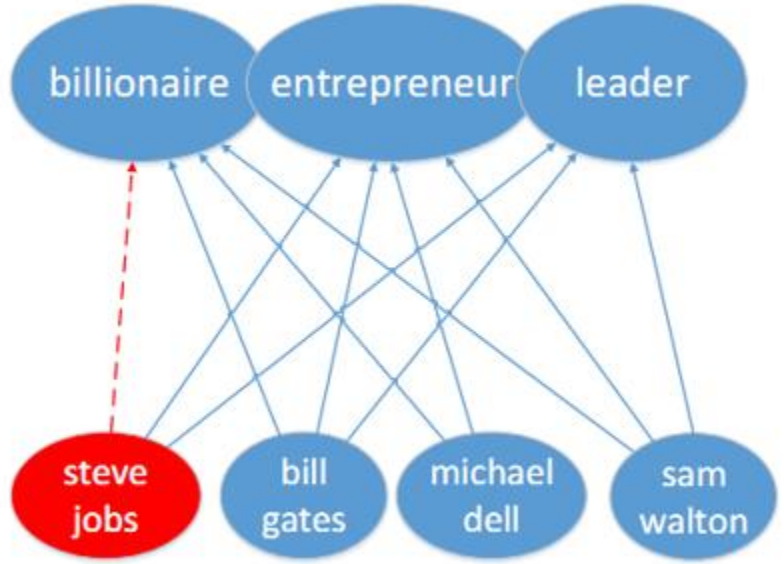
$$sc_b(t) = \begin{cases} synsets(\mathrm{b}) - \theta(\mathrm{b}) & \mathrm{b} \in \text{WordNet;} \\ 1 & \text{otherwise.} \end{cases}, t = \langle \mathrm{a}, \mathrm{b}, \mathrm{c} \rangle$$

基于相似实体进行图谱补全

基于传递性的方法

- 大量的低频实体没有足够的传递性的信息
- 往往会连接到比较高层的抽象概念

考虑实体“Steve Jobs”
很容易找到类似的人物，如“Bill Gates”
这些类似实体都属于“Billionaire”
可以推测，“Steve Jobs”也属于“Billionaire



基于相似实体的方法

- 可以通过其他方法寻找相似实体
- 可以连接到更近的概念

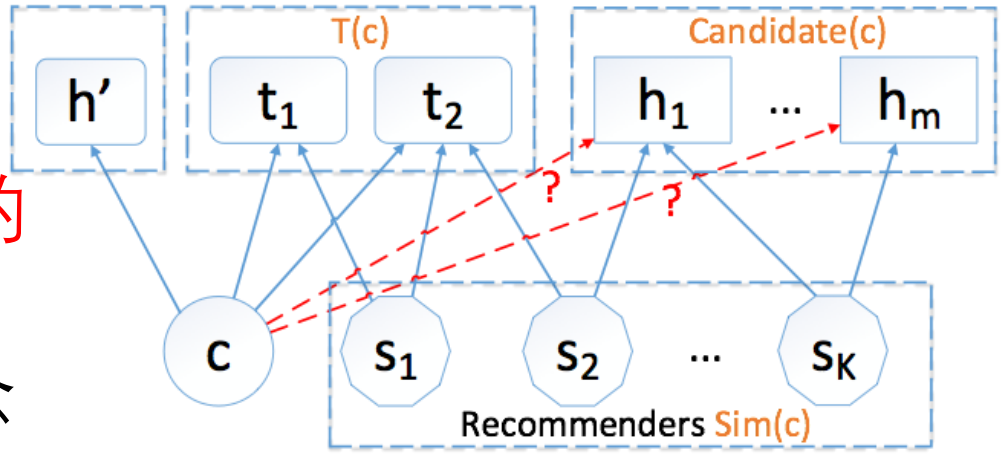
基于相似实体的图谱补全：框架

- 框架：协同过滤
  - **原理：相似的实体很有可能拥有类似的概念**

- 协同过滤的优点
  - 协同过滤和基于相似实体的思路一致
  - 协同过滤非常灵活
    - 相似度和推荐打分都可以灵活地根据实际情况进行调整
  - 协同过滤已有很多缓解“冷启动”问题的优化
    - 正好能用于大量的低频实体

协同过滤

- 基于用户的协同过滤
  - Hypernyms --- 物品
  - Concepts --- 用户
  - Synonyms or Siblings --- 相似用户
- 有相似意思的term很可能拥有相同的上位概念/下位实体
- 为了为实体/概念c寻找新的上位概念
  - Step 1 寻找c的相似实体/概念
  - Step 2 将c的相似实体/概念的上位概念共享给c



Idea:

如果c的大多数相似项都有上位概念h，c也很可能拥有上位概念h

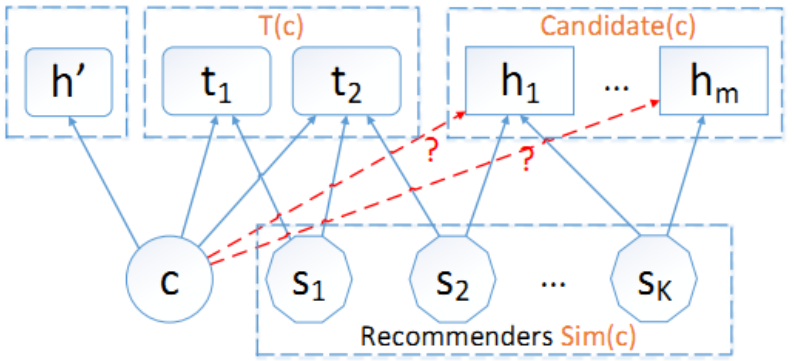
协同过滤框架

- 迭代式框架
  - 对每一个实体c
    - 寻找和c最相似的k个实体
    - 将这k个实体的概念作为待选概念
    - 对这些待选概念进行打分和排序
    - 将所有高分的待选概念推荐给c
  - **T(c)**是c的已知概念和待选概念的交集
    - 这个集合可以用于作为训练数据以确定打分算法的参数和阈值

Algorithm 1 CF-based Missing isA Relationship Finding

Input: Taxonomy T, parameters K, θ

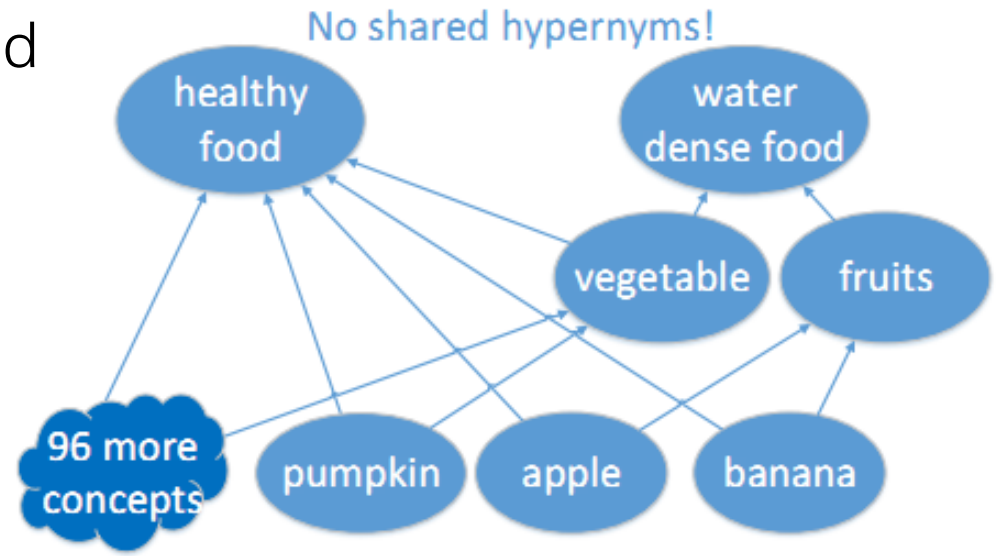
while $iteration < max\_iteration$ **do**
  for term $c \in T$ **do**
    $Sim(c) \leftarrow$ top-K similar terms of c;
    $Candidate(c) \leftarrow [\bigcup_{s \in Sim(c)} hype(s)] - hype(c)$;
    Rank candidates in $Candidate(c)$ by a scoring function f;
    Update T by attaching c to any $x \in Candidate(c)$ s.t. $f(x) \geq \theta$;
  end for
  $iteration \leftarrow iteration + 1$;
end while



协同过滤：相似度计算

- 如何在概念图谱中寻找寻找和c最相似的k个实体？
  - 先要定义一个相似度函数$\mathrm{sim}(c1, c2)$
  - sim = 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相似度



协同过滤：Jaccard相似度

- Jaccard相似度：
  - 直接的考虑，上位概念和下位概念集合重叠越多的实体越相似
  - 分别对上位概念集合和下位概念集合计算：j_e & j_o
  - 使用noisy-or合并这两个相似度
    - 使用noisy-or的原因
      - 由于概念图谱缺失问题，这两个值可能偏小
      - 对底层实体对，j_o经常为0

$$Jacc(U,V) = \frac{|U \cap V|}{|U \cup V|} = \frac{|U \cap V|}{|U| + |V| - |U \cap V|}$$

$$j_e(c_1, c_2) = Jacc(hype(c_1), hype(c_2))$$
$$j_o(c_1, c_2) = Jacc(hypo(c_1), hypo(c_2))$$

$$jacc(c_1, c_2) = 1 - (1 - j_e(c_1, c_2)) \cdot (1 - j_o(c_1, c_2))$$

协同过滤：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

$$\vec{v}_c^{(i)} = \frac{1}{2}\vec{v}_c^{(0)} + \frac{1}{2}M\vec{v}_c^{(i-1)}$$

$$rw(c_1, c_2) = \frac{V_{c_1} \cdot V_{c_2}}{\|V_{c_1}\|\|V_{c_2}\|}.$$

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

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

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

协同过滤：备选概念打分排序

- 备选概念推荐分数：带权和-协同过滤最基本的打分方法
  - 对每个备选概念，将所有产生此备选概念的相似实体的相似分数求和
    - $\mathrm{Score}(h_j) = \sum_{s_i \in Sim(c)} w(s_i\ isA\ h_j)\, sim(s_i, c)$
  - 利用T(c) (c已存在的备选概念)中的分数来计算阈值
    - 即推荐分数超过大部分（80%）c已有的概念的分数才能被补全

isA关系纠错

概念图谱纠错：概述

- 自动抽取技术可以轻易产生千万级别规模的概念图谱
- 由于基数十分巨大，里面的错误信息的数量也非常多

- 有必要对抽取得到的概念图谱进行清洗

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

Exciting city isA Paris
句子“…Paris is such as(an) exciting city”
笔误，其中的an写成了as，符合Hearst模板

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

概念图谱纠错

- 不可能自动化地找到所有错误
- 考虑优先能够处理一些或者某一类比较常见且容易判断的错误

简单的想法：知识的支持度

- 通过每一条知识寻找支持证据，来“证明”每一条知识
  - 若某句子通过抽取得到了一条知识，那么此句子就“支持”此知识
    - 如果有大量的句子都可以抽取到同一条知识，那么它非常有可能是正确无误的
  - 右表：Probase中按不同支持度采样的结果
    - 若支持度足够高，知识的正确率非常高
    - 低支持度的知识有可能是错误的
    - 低支持度的知识太多，全部删除过于浪费

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%

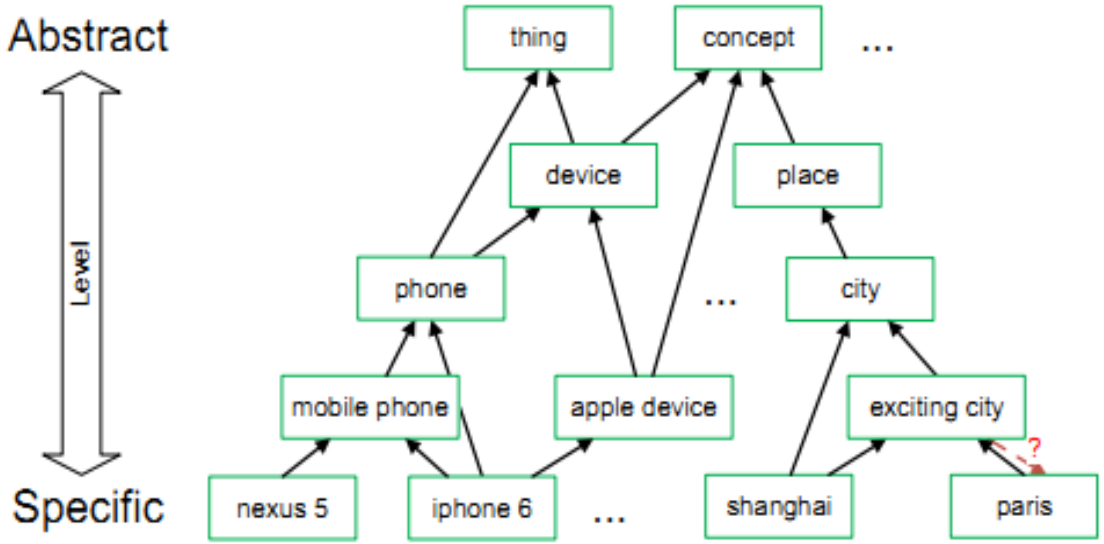
实例分析：Probase中的错误

- 通过Case Study寻找常见的错误
  - 共性：*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
    - *抽象的概念* isA *具体实体*可能导致图谱中产生环

Probase中的部分错误

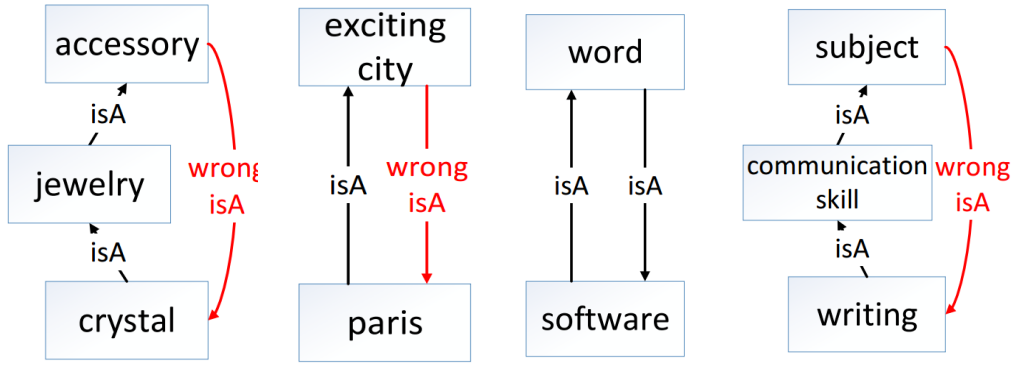
Entity	*isA*	Concept	Entity	*isA*	Concept
exciting city	*isA*	paris	battery	*isA*	fuel cell
automobile	*isA*	lead acid battery	cause	*isA*	tsunami
music video	*isA*	youtube video	sweet	*isA*	glucose
world cup	*isA*	football	grape	*isA*	purple
college	*isA*	basketball	juice	*isA*	tomato

Table 1: Examples of incorrect isA relations in Probase



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
  - -> 通过寻找概念图谱中的环，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

在概念图谱中进行消环

- 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
Minimize $\sum_{e \in E'} w(e)$, 其中w(e) 是e的可信程度

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

输出错误边集 E’ 应当尽可能包含不可信的边

边可信度定义

- 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**

$$P_h(X\ isA\ Y) = \log(1 + \frac{hypo(Y)}{hypo(X)})$$

- 两指标之积作为最终可信度

模型求解

- **给定有向图G(V, E)，可信度函数$w: E \rightarrow$ 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**
  - -> 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。

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.

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