

《知识图谱：概念与技术》

百科图谱构建



本章大纲

- 百科图谱概述
- 基于单源的百科图谱构建
 - 数据获取
 - 属性抽取
 - 关系构建
 - 概念层级体系构建
 - 实体分类
- 基于多源的百科图谱融合
 - 基于多个知识图谱的融合
 - 基于多源异构数据的融合

百科图谱概述

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

非百科网页

百科图谱构建

百科网页

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 education, family, and other achievements.

This screenshot shows the Chinese Baidu Baike page for Donald Trump. It features a large portrait of him, a brief summary, and detailed sections on his political career, including his time as the 45th President of the United States and his role as the host of 'The Apprentice'.

• 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

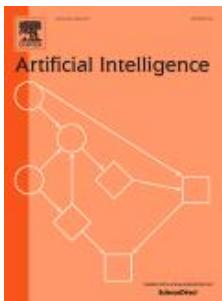
This screenshot shows the IMDB page for the 2017 film 'Beauty and the Beast'. It displays the movie's title, rating (7.9), and a summary. Below the summary is a thumbnail image of the movie poster and a play button.

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 a summary. The page also includes reviews and a 'Buy Now' button.

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



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.ijcar.org/index.php/aij-awards-list-of-previous-winners>

基于单源的百科图谱构建

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

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/>{{File:Euclid.jpg|right|thumb|200px|[歐幾里得]]，西元前三世紀的希臘數學家，現在被認為是幾何之父，此畫為[[拉斐爾|聖齊奧|拉斐爾]]的作品[[拉斐爾|聖齊奧|拉斐爾]]的傑作。}}<br/>「数学」( Mathematics )是利用符号语言研究[[數量]] &lt;ref name="OED"&gt;{{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[[印度历史|古印度
今日，數學使用在不同的領域中，包括[[科学|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}} )<br/>「數學」一詞的大約產生於宋朝末[[元朝元]]時期，多指參數之學，但有時也含有今天上的數學意義，例如，[[秦九韶]]的《{{數學九章}}》 ( {{水經大典}} |author = 鄭重春 )
```

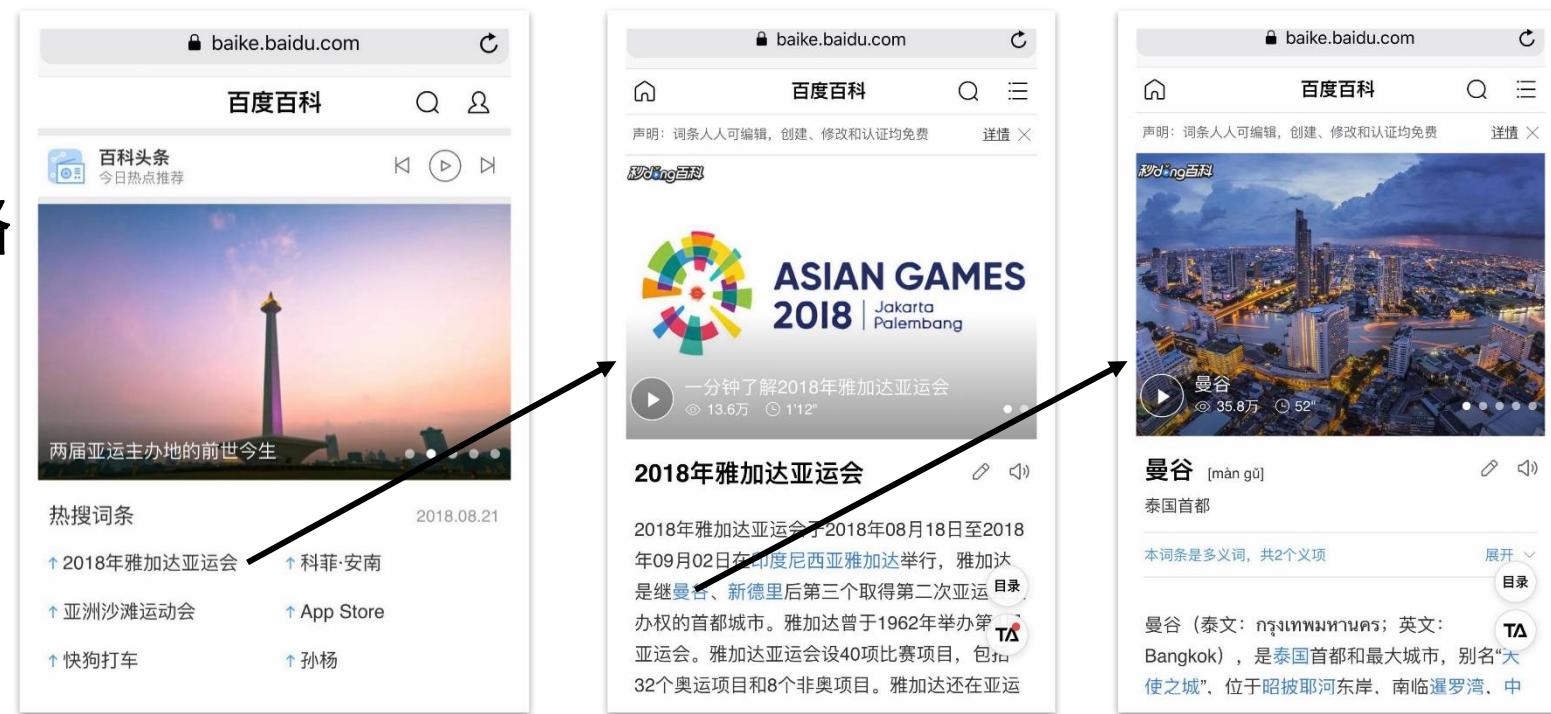
Wikipedia Dump <https://dumps.wikimedia.org/>

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

适用场景：百科网站的页面
URL具有可枚举性

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/周杰伦
http://baike.baidu.com/view/2.htm	http://baike.baidu.com/item/复旦大学
http://baike.baidu.com/view/3.htm	http://baike.baidu.com/item/一出好戏
http://baike.baidu.com/view/4.htm	http://baike.baidu.com/item/黄渤

页面识别

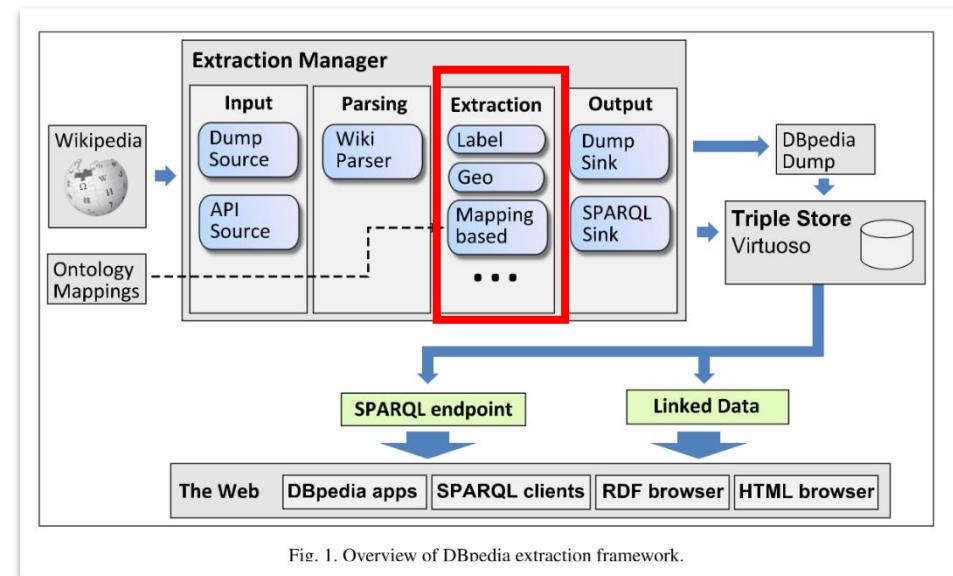
- 目标
 - 筛选出所有介绍实体的网页
- 百科页面的特殊性
 - 每个页面均围绕一个词条进行全方面的介绍
- 方法
 - 每个词条名作为知识图谱中的一个实体
 - 实体发现过程等价于词条页面发现
 - 百科网站的词条页面URL具有一定的规律性
 - `http://baike.baidu.com/view/[ID].htm`
 - `http://baike.baidu.com/item/[NAME]`
 - `https://music.163.com/#/song?id=[ID]`
 - `https://movie.douban.com/subject/[ID]`



属性抽取

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

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^[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。

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)

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》^[1]。1983年主演的武侠剧《神雕侠侣》在香港获得62点的收视纪录^[2-3]。1991年创办天幕电影公司^[4]。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^[5]。1994年担任剧情片《天与地》的制片人^[6]。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^[7]。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^[8]。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’^[9]。2006年获得釜山国际电影节亚洲最有贡献电影人奖^[10]。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^[11]；同年担任第49届台湾电影金马奖评审团主席^[12]。2017年主演警匪动作片《拆弹专家》^[13]。

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



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

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》^[1]。1983年主演的武侠剧《神雕侠侣》在香港获得62点的收视纪录^[2-3]。1991年创办天幕电影公司^[4]。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^[5]。1994年担任剧情片《天与地》的制片人^[6]。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^[7]。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^[8]。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^[9]。2006年获得釜山国际电影节亚洲最有贡献电影人奖^[10]。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^[11]；同年担任第49届台湾电影金马奖评审团主席^[12]。2017年主演警匪动作片《拆弹专家》^[13]。

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

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

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 Encyclopedia page for Fudan University. On the left, there is a table of basic information:

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

On the right, the browser's developer tools are open, specifically the Elements tab. It shows the extracted HTML code for the basicInfo-item value class, which corresponds to the data in the table.

```
<dd class="basicInfo-item value">博士后流动站</dd>
<dd class="basicInfo-item value">35个</dd>
<dd class="basicInfo-item value">校训</dd>
<dd class="basicInfo-item value">博学而笃志，切问而近思</dd>
<dd class="basicInfo-item value">校歌</dd>
<dd class="basicInfo-item value">《复旦大学校歌》</dd>
<dd class="basicInfo-item value">中国科学院院士 21人</dd>
<dd class="basicInfo-item value">中国工程院院士 5人</dd>
<dd class="basicInfo-item value">主要院系</dd>
<dd class="basicInfo-item value">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</dd>
<dd class="basicInfo-item value">国家重点学科</dd>
<dd class="basicInfo-item value">一级学科 11个，二级学科 19个</dd>
<dd class="basicInfo-item value">学校地址</dd>
<dd class="basicInfo-item value">上海市杨浦区邯郸路220号</dd>
<dd class="basicInfo-item value">学校代码</dd>
<dd class="basicInfo-item value">10246</dd>
<dd class="basicInfo-item value">主要奖项</dd>
<dd class="basicInfo-item value">全国优秀博士论文55篇(截至2012年)</dd>
<dd class="basicInfo-item value">校庆日</dd>
```

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	中华人民共和国教育部
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

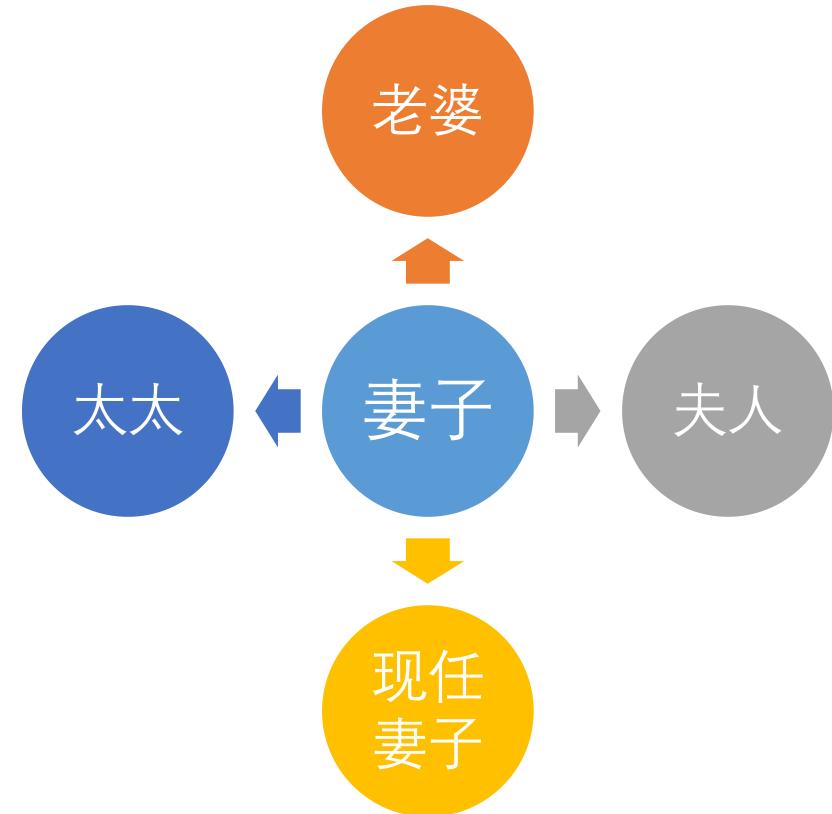
① 生成：找到候选等价属性对

- 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



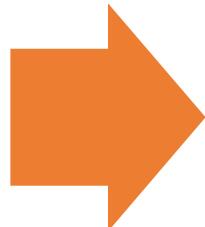
质量问题二：数值属性值格式不统一

- 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})$')↓
```

The screenshot shows a user interface for unit conversion. At the top, there is a navigation bar with tabs for Length, Area, Volume, Mass, Temperature, Pressure, Power, and Energy/Heat. Below the tabs, there is a conversion input field with the value '1' and units '千米(km)'. To the right of this field is a conversion button with a double-headed arrow icon. Next to the button is another input field with the units '米(m)'. Below the input fields, the text '1千米(km)=1000米(m)' is displayed. At the bottom of the interface, it says 'International Unit: 米(m)'.

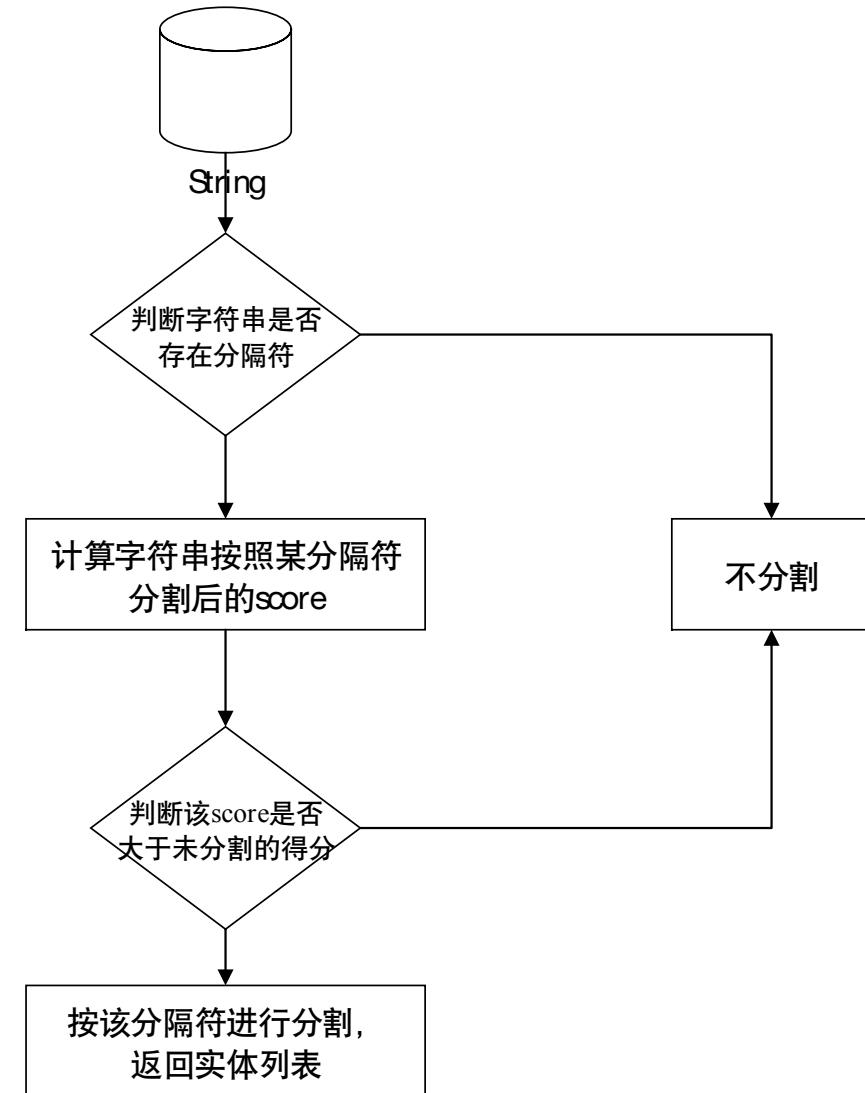
对象属性值分割

- 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

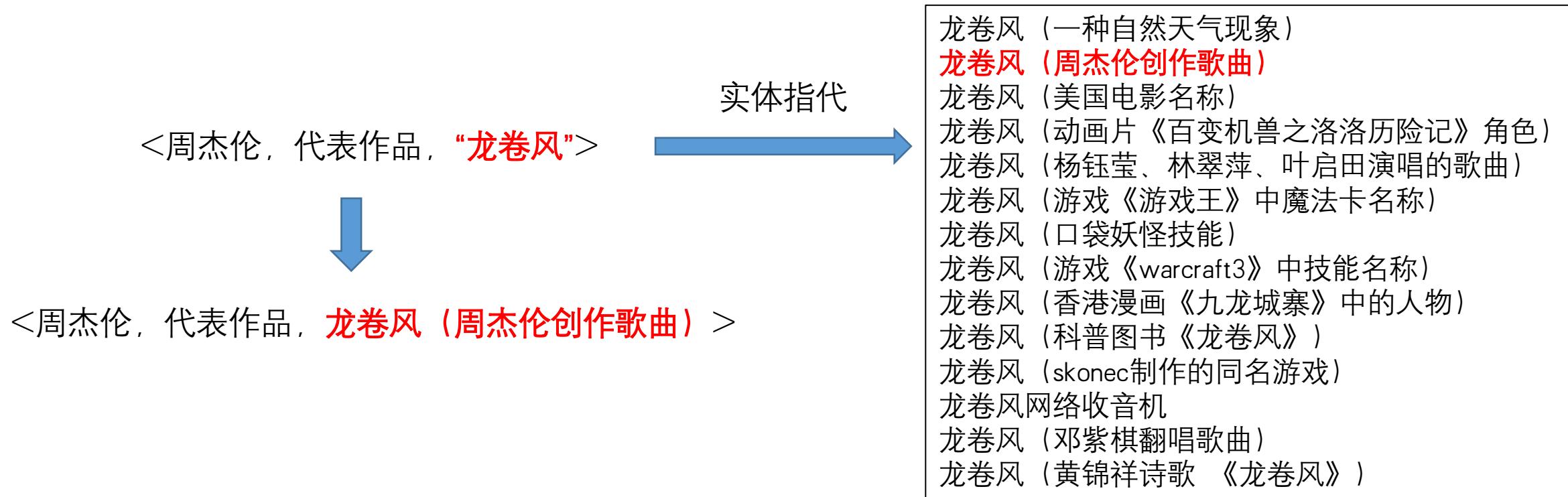
- 空格、中文逗号、英文逗号、中文顿号、英文斜杠、中文分号、英文分号、英文竖号



关系构建

属性抽取步骤的遗留问题

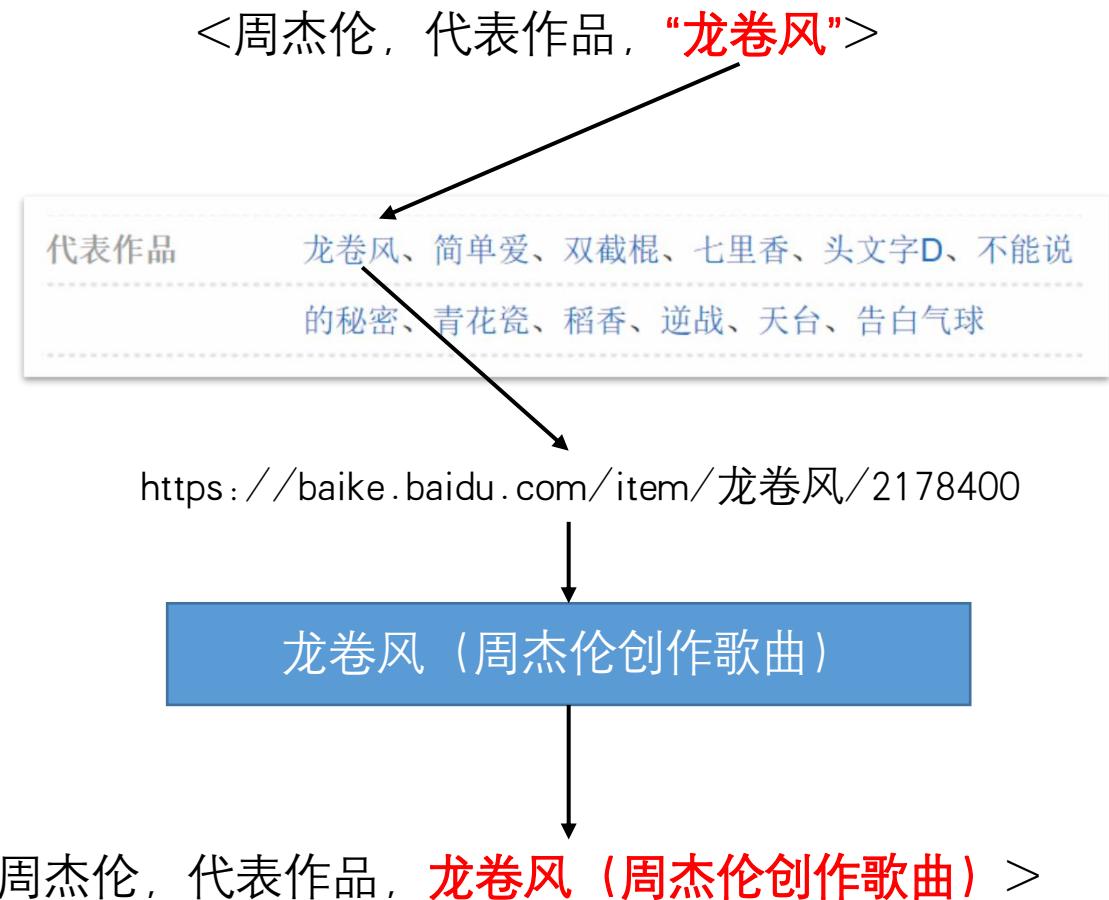
- 未建立实体与实体之间的关系
- 缺少这一步，知识图谱中的实体无法建立关联



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

- 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

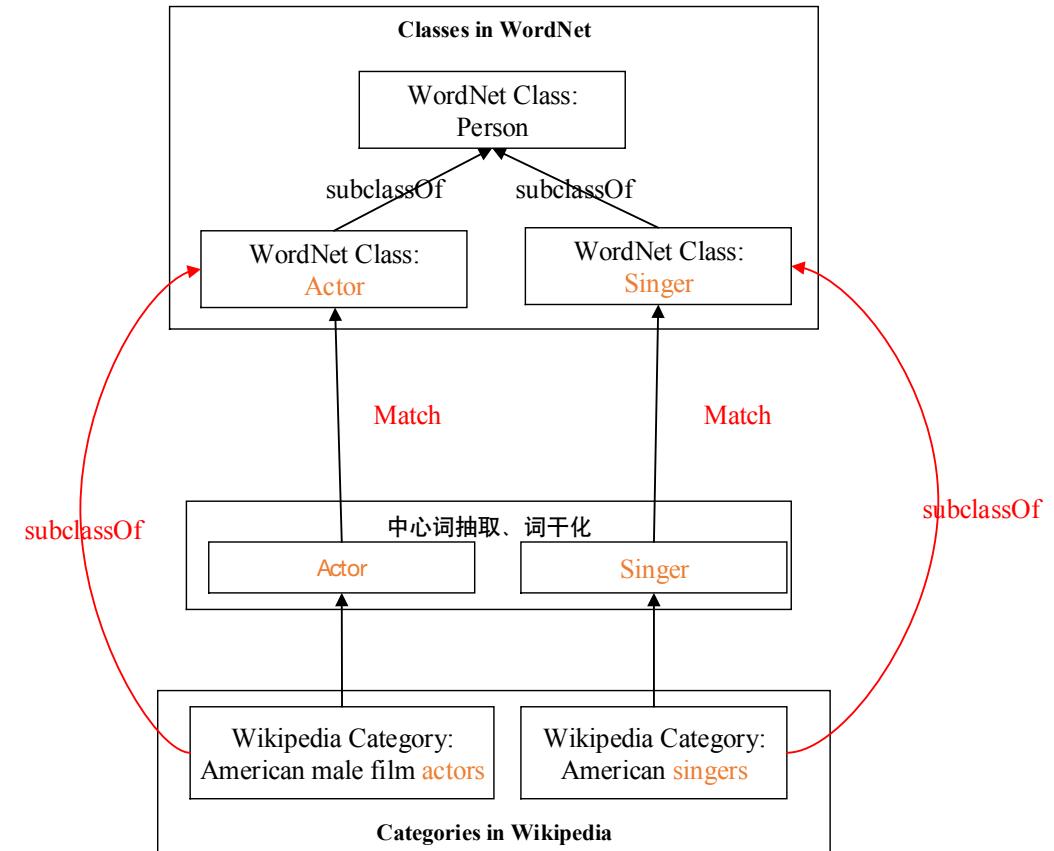
<http://mappings.dbpedia.org/server/ontology/classes/>

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

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

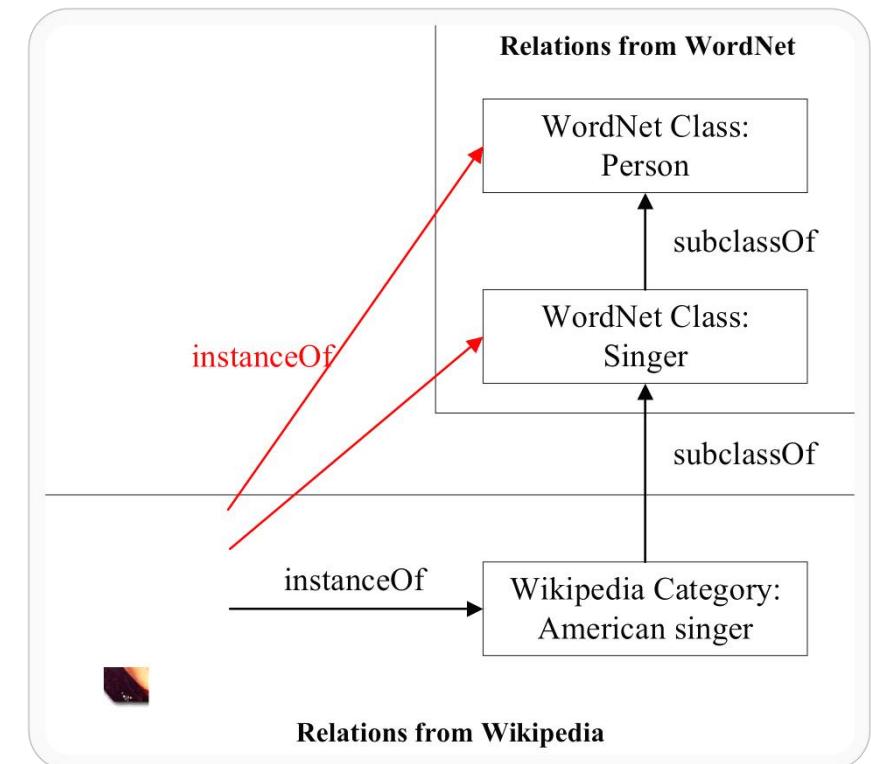
- E.g., 实体名称后缀为“步枪”的很可能属于步枪

- 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

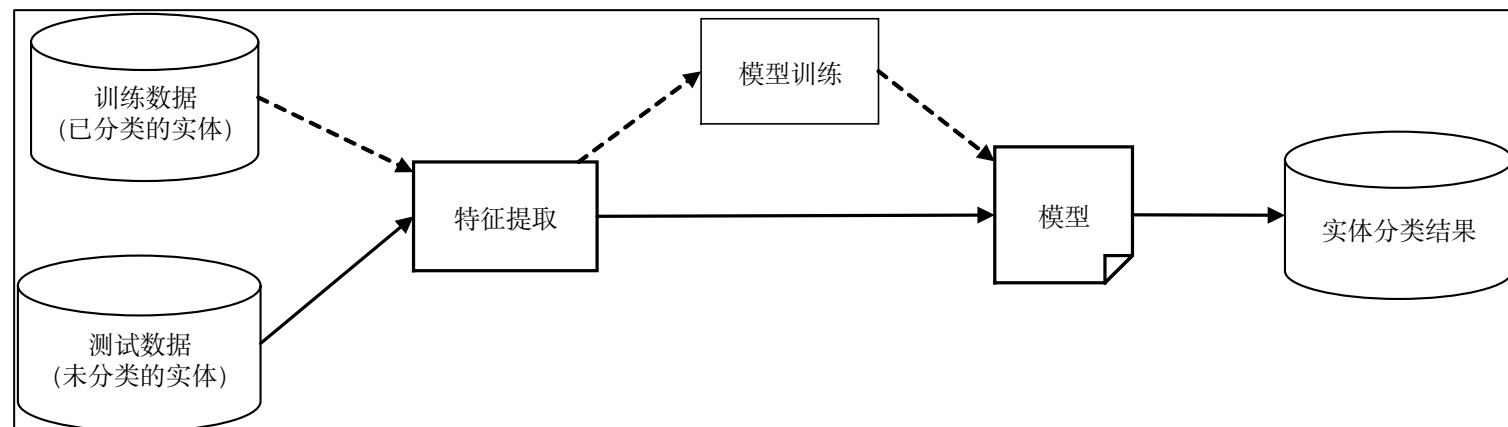
- E.g., 实体包含属性—值对（职业，演员）的属于演员



[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



特征表示

• 单示例特征表示

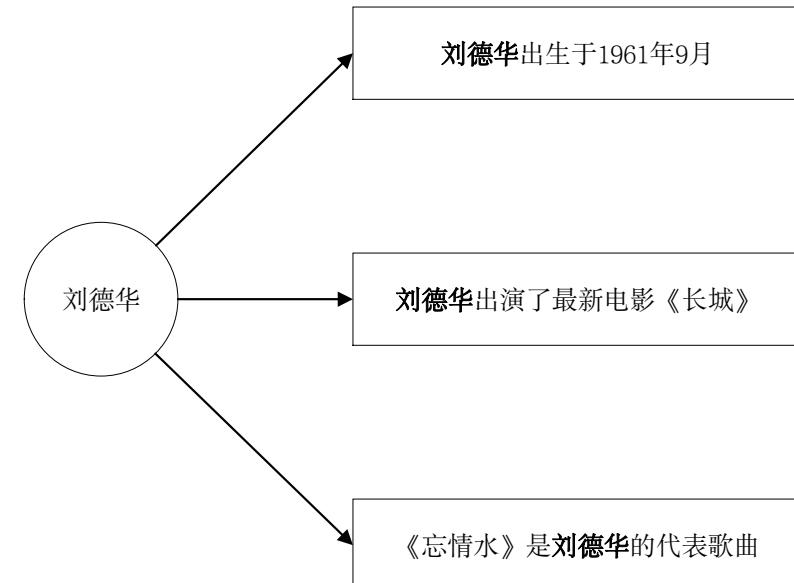
- 一个实体用一组特征集合表示

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

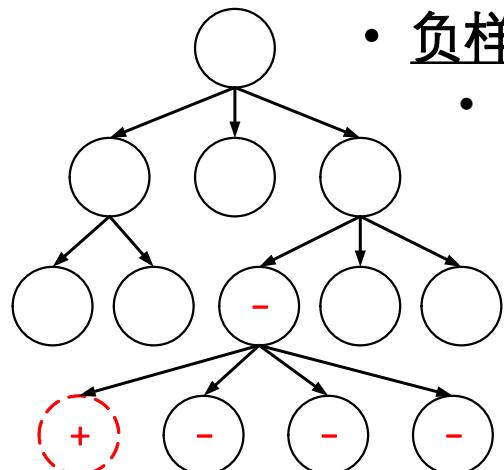
- 输入：实体的特征集合 X
 - $X = [x_1, x_2, \dots, x_i, \dots, x_N]$
 - N 为特征总数
 - $x_i = 1$: 实体包括这一特征
 - $x_i = 0$: 实体不包含这一特征
- 输出：实体的分类结果 Y
 - $Y = [y_1, y_2, \dots, y_i, \dots, y_M]$
 - M 为概念总数
 - $y_i = 1$: 实体属于这个概念
 - $y_i = 0$: 实体不属于这个概念

- 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

单示例实体分类方法：CUTE

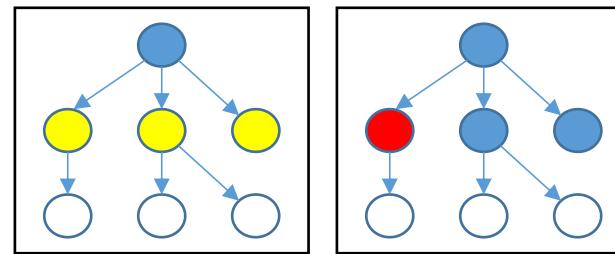
[Bo Xu et al., 2016a]

- 考虑概念之间的层次结构
- 训练过程
 - 为每个概念分别构建一个分类器
 - 为每个分类器定义其正负样本
 - 正样本
 - 所有属于该概念的实体
 - 负样本
 - 所有属于该概念的父概念却不属于该概念的实体

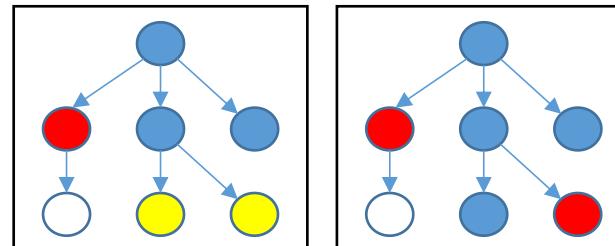


- 预测过程
 - 自顶向下的预测过程

第一轮



第二轮

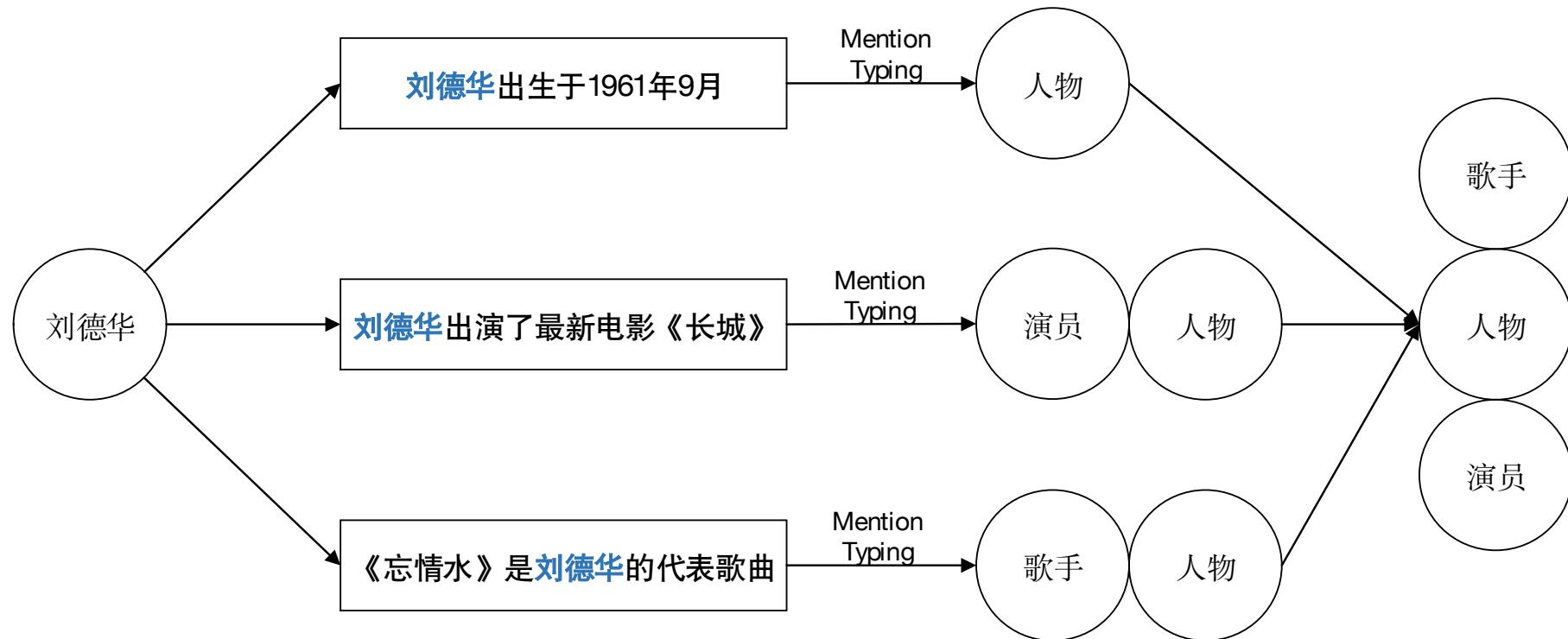


- Classifier Predict 1
- Candidate Classifier
- Classifier Predict 0
- Non-Candidate Classifier

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

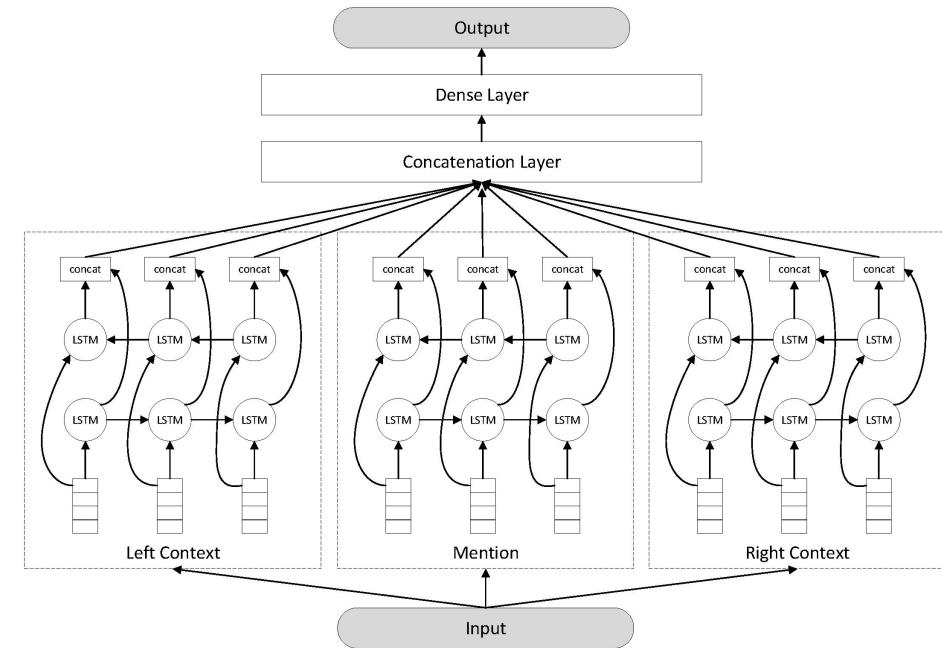
- 解决方案
 - 基于人工特征的方法

- 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)

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

- 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)



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]

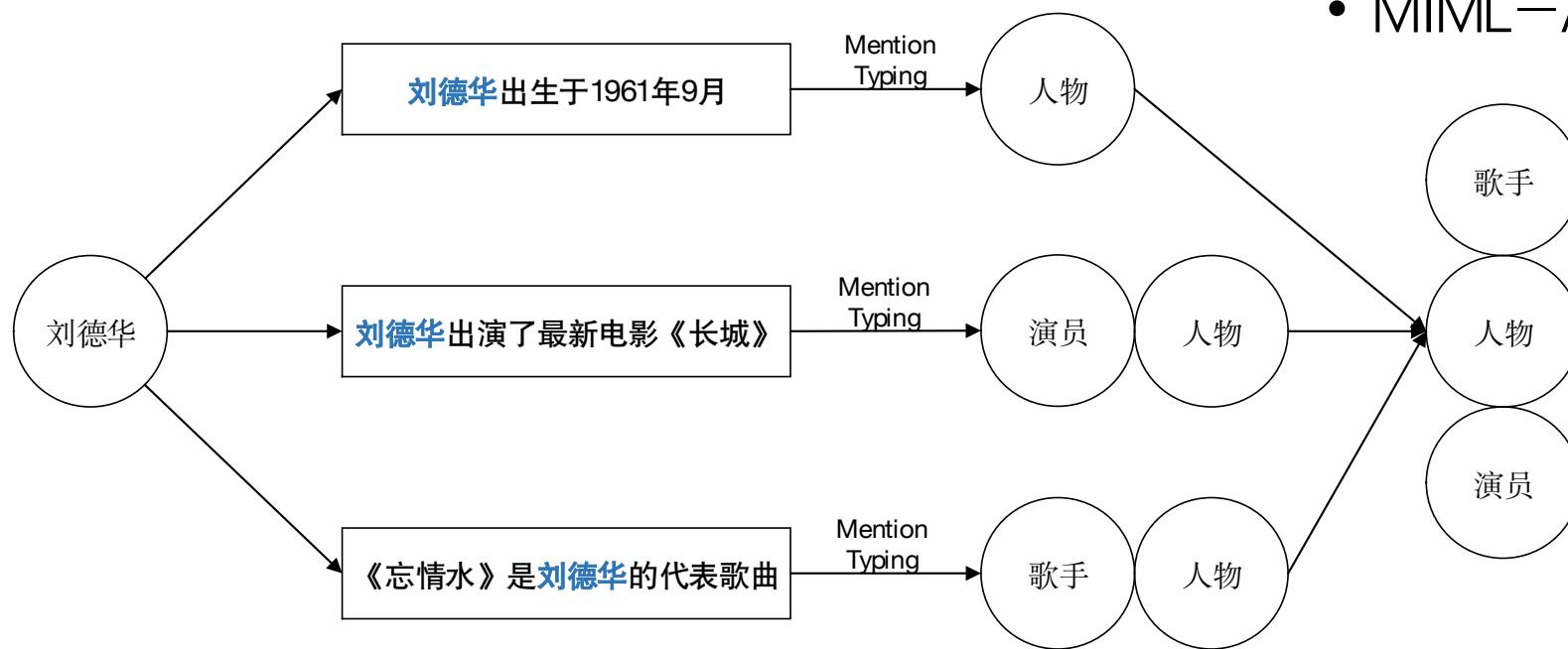
多示例实体分类：多示例学习方法

- 基本思路

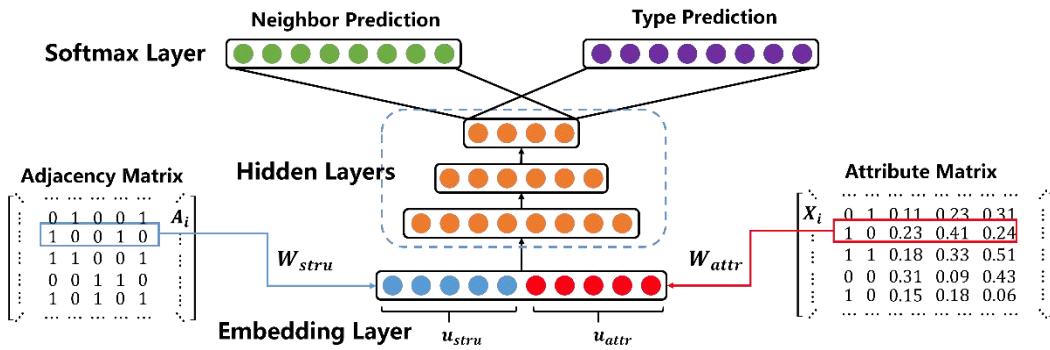
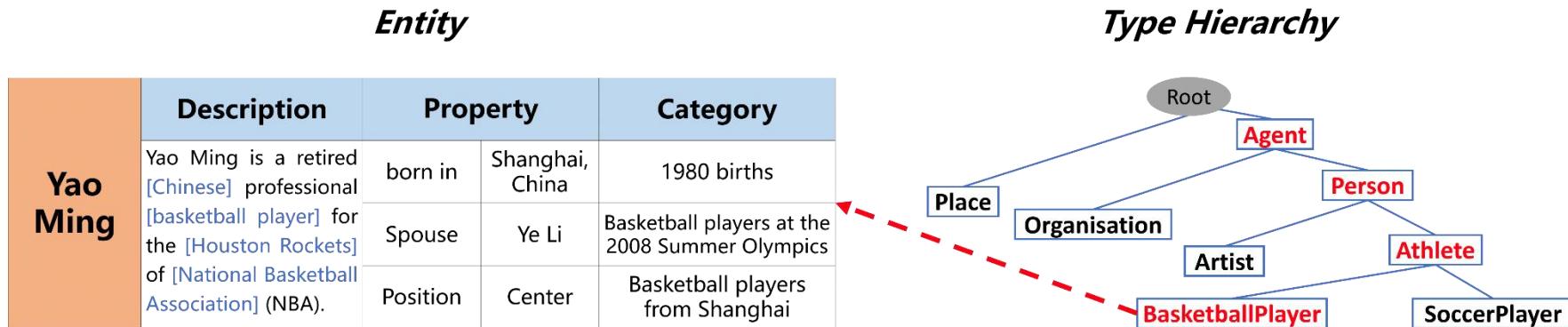
- 输入：一个实体的全部示例
- 输出：一个实体的分类结果

- 方法

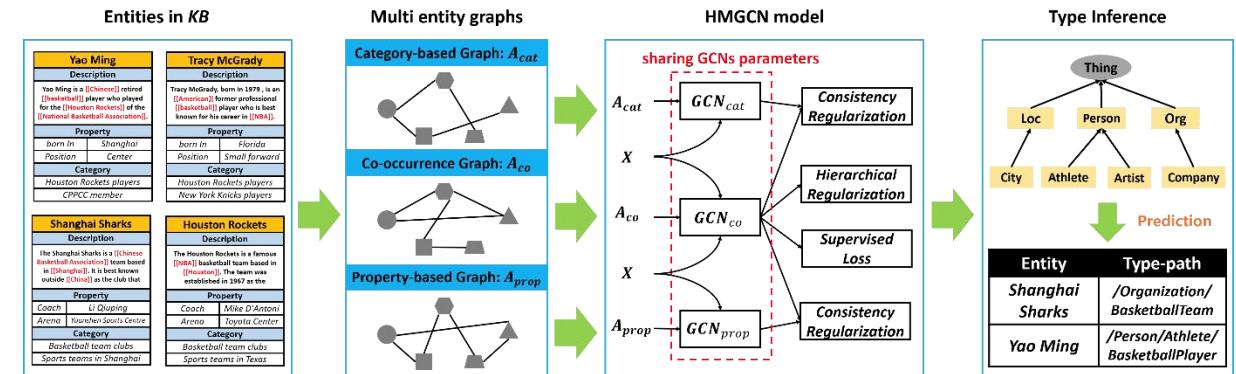
- MIML-MAX
- MIML-AVG
- MIML-MAX-AVG
- MIML-ATT



基于异构特征的实体分类



[COLING2018] Attributed and Predictive Entity Embedding for Fine-Grained Entity Typing in Knowledge Bases



[EMNLP2019] Fine-Grained Entity Typing via Hierarchical Multi Graph Convolutional Networks

基于多源的百科图谱融合

基于多源的百科图谱融合

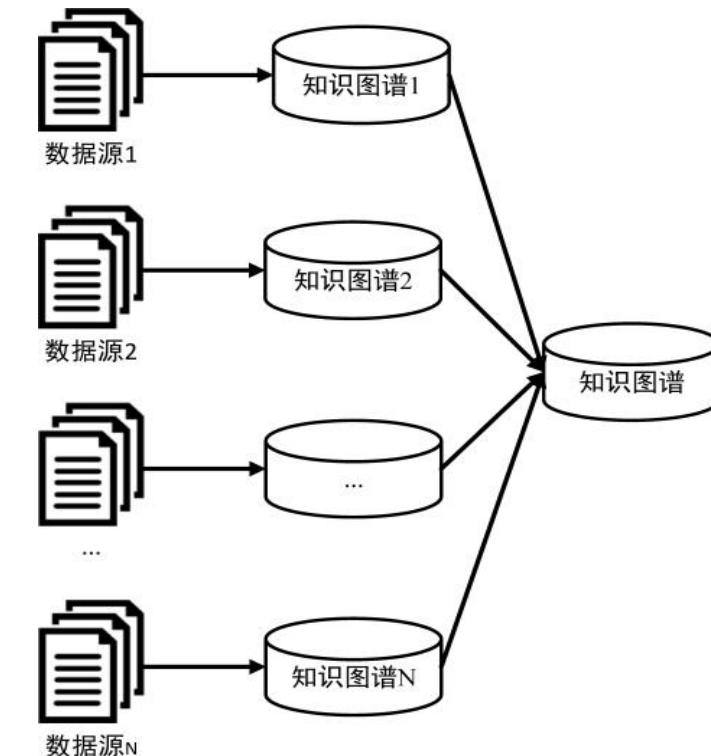
- 问题定义
 - 输入
 - 多源数据
 - 多个知识图谱
 - 多源异构数据
 - 输出
 - 一个融合后的百科知识图谱
 - 分类
 - 基于多个知识图谱的融合方法
 - 基于多源异构数据的融合方法

基于多个知识图谱的融合方法

基于多个知识图谱的融合方法

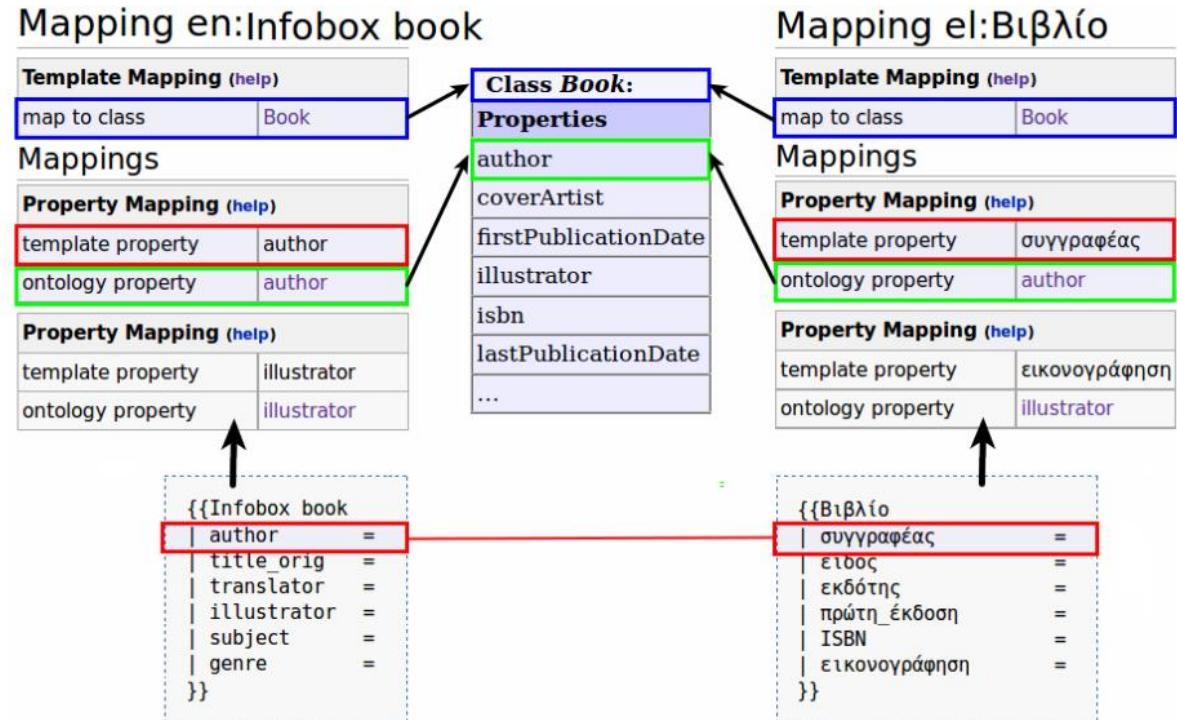
- 步骤

- 概念融合
- 实体对齐
- 属性对齐
- 属性值融合



概念融合

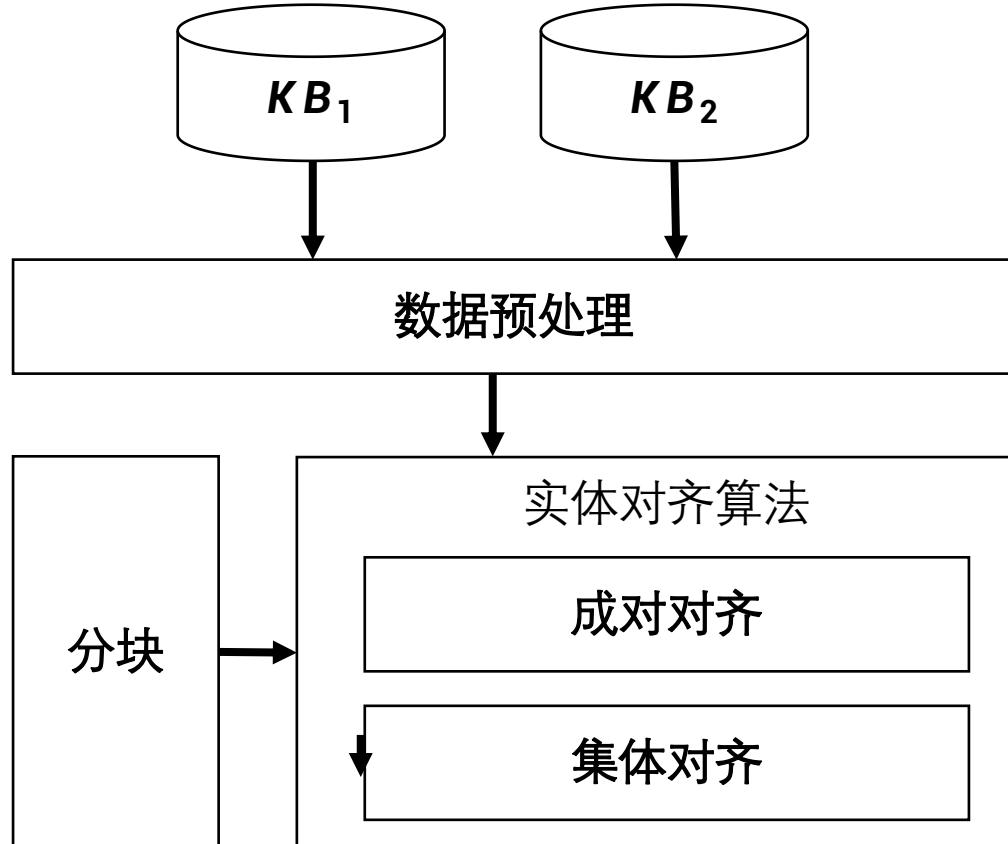
- 不同知识图谱的概念层级体系各不相同，而融合后的知识图谱只能有一个概念层级体系
- 关键技术
 - 找到等价概念
- 融合策略
 - DBpedia
 - 以英文taxonomy为主，删除其他taxonomy上未匹配的概念
 - XLORE
 - 保留所有概念，把等价的概念合并



[Jens Lehmann et al., 2015]

实体对齐

- 实体对齐是知识图谱融合的最关键的步骤，它决定了知识图谱之间是否能够融合
- 实体对齐的主要任务是判断来自两个知识图谱中的实体是否等价
- 流程
 - 预处理
 - 分块索引
 - 成对对齐
 - 集体对齐



预处理和分块索引

- 预处理
 - 目标
 - 处理数据的多源异构性、数据定义的不一致性、数据表达的多样性等
 - 方法
 - 标点去除
 - 同义词扩展
- (1) 肖申克的救赎 = 《肖申克的救赎》
(2) 海尔波普彗星 = 海尔·波普彗星 = 海尔-波普彗星
(3) 奋进号航天飞机 = “奋进号”航天飞机
- Zhishi.me
- 分块索引
 - 目标
 - 通过剪枝过滤掉知识库中不可能相似的实体对，使得相似的实体对尽量分配到一个或几个区块中成为候选对，最终的对齐处理只在这些候选对中进行，从而达到提高匹配效率的目的
 - 索引键值选择的考虑因素
 - 特征的质量高
 - 特征的分布均匀
 - 区块数量和大小适中

[Niu, Xing, et al., 2011]

[Zhuang Yan et al., 2016.]

成对实体对齐方法

- 无监督学习方法
 - 根据现有的知识得到等价关系进行判断
 - 如DBpedia通过维基百科中的实体的多语言版本信息得到了不同语言知识图谱中实体的等价关系
 - 根据实体名称的相似度进行判断
 - 相似度计算的方法包括Jaccard系数、Dice系数和编辑距离等
 - 根据实体的语义相似性进行判断
 - 如同义关系或概念相近等

成对实体对齐方法

- 监督学习方法
 - 二元分类问题（匹配，不匹配）
 - 特征
 - 通过人工方式定义，来自实体的名称、正文、相关实体、标签、Infobox等信息
 - 模型
 - SVM
 - Logistic Regression
 - Decision Tree
 - Factor Graph
 - Heterogeneous Network Embedding

集体实体对齐方法

• 局部集体实体对齐

- 基于简单关系的集体实体对齐
- 在计算实体相似性的时候，将实体的关联实体属性纳入计算，即考虑待匹配实体对的邻居的属性集合，但并不将邻居节点当作平等的实体去计算结构相似度

$$sim(e_1, e_2) = a \times sim_{Attr}(e_1, e_2) + (1 - a) \times sim_{NB}(e_1, e_2)$$

↓ ↓

实体本身属性相似度 实体关联实体相似度

[Zhuang Yan et al., 2016.]

集体实体对齐方法

• 全局集体实体对齐

- 基于实体对齐是相互影响的观察，通过不同匹配决策之间的相互影响调整实体之间的相似度
- 方法
 - 基于相似性传播方法
 - 基本思路
 - 通过初始匹配以bootstrapping方式迭代地产生新的匹配
 - “如果2个作者匹配，则与这2个作者具有“*coauthor*”关系的另外2个相似名字的作者会有较高的相似度，而这个相似度又会对其他作者匹配产生影响”
 - 基于概率模型方法
 - 基本思路
 - 全局概率最大化
 - 方法
 - 贝叶斯网络、LDA模型、条件随机场模型、Markov逻辑网络模型

[Zhuang Yan et al., 2016.]

举例：XLORE跨语言实体对齐

- 目标
 - 将百度百科中的中文实体和维基百科中的英文实体对齐
- 训练集来源
 - 英文维基实体↔中文维基实体
↔中文百度百科实体
- 特征来源
 - 标题
 - 超链接
 - 标签
 - 作者

[Zhichun Wang, et al., 2012]

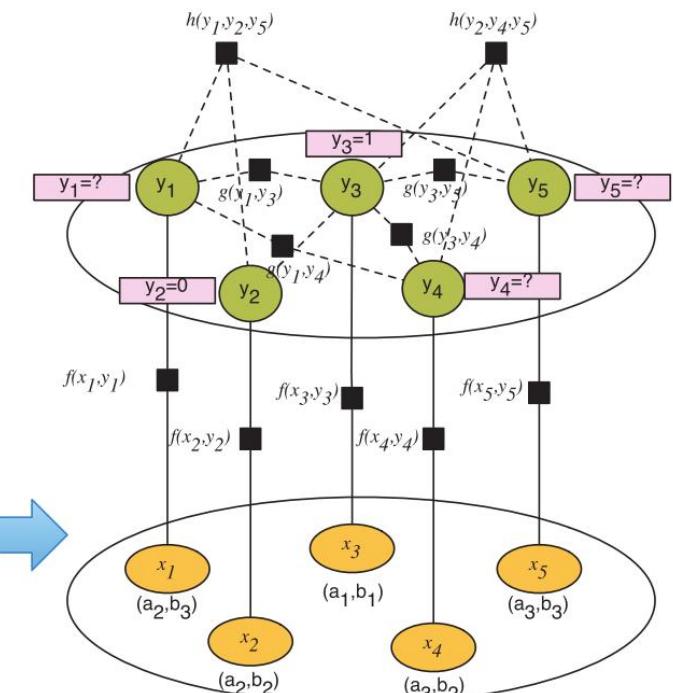
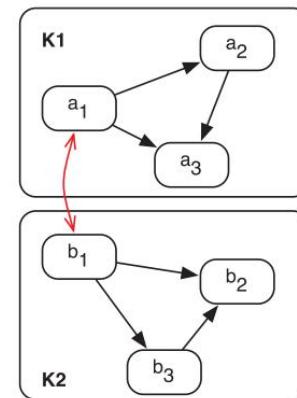


举例：跨语言实体对齐

- 方法
 - 集体实体对齐方法
- 模型
 - Linkage Factor Graph Model

$$p(Y) = \prod_i f(y_i, x_i) g(y_i, G(y_i)) h(y_i, H(y_i))$$

$Y = \{y_i\}_{i=1}^{n \times m}$ 点特征函数 边特征函数 约束特征函数



$$x_i = (a_{i1}, b_{i2})$$

$$y_i = \{0, 1\}$$

[Zhichun Wang, et al., 2012]

输入及输出

- 输入：连接图(PCG)
- 输出： $Y = \{y_i\}_{i=1}^{n \times m}$ ，每个 y_i 的取值为0或者1，表示PCG的每个节点中的两个实体是否等价

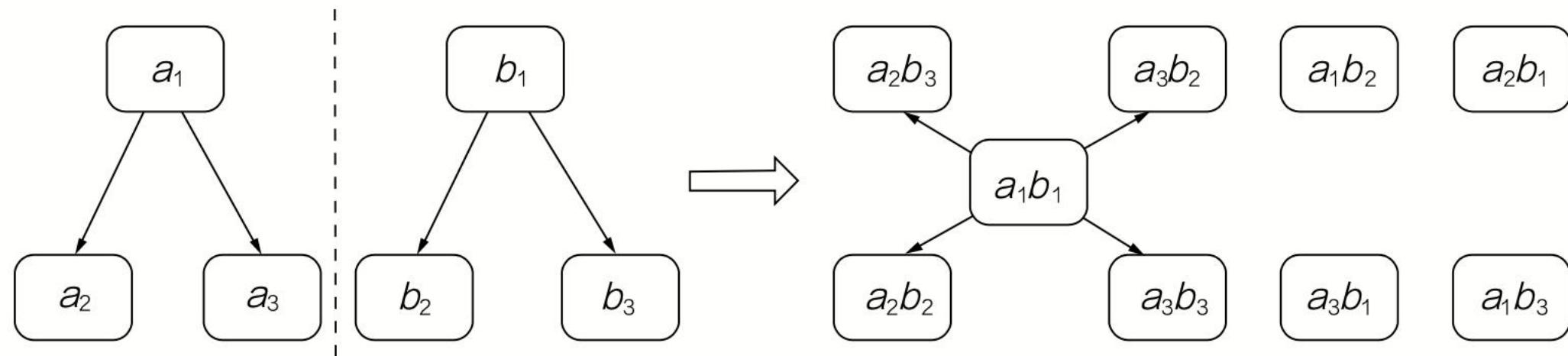


图 6-10 文献[26]提出的将两个知识图谱的引用图转化为 PCG

点特征

- 第一类为PCG中节点的特征函数 $f(y_i, x_i)$ ，如公式(6-6)所示，用来表示PCG中每个节点 x_i 对应的实体对是否等价的先验概率。

$$\cdot f(y_i, x_i) = \frac{1}{Z_a} \exp \{a^T f(y_i, x_i)\} \quad (6-6)$$

- 其中， a 是要学习的特征权重， $f(y_i, x_i)$ 是一个概率向量，由各类似度指标构成（比如文献[26]采用了两个实体在百科中链入（出）实体集的相似度、标签相似度和编写者的相似度等）

边特征

- 第二类为PCG中边的特征函数 $g(y_i, G(y_i))$, 如公式(6-7)所示, 用来表示节点之间的相关性。
 - $$g(y_i, G(y_i)) = \frac{1}{Z_\beta} \exp\left\{\sum_{y_j \in G(y_i)} \beta g(y_i, y_j)\right\} \quad (6-7)$$
- 其中, β 是要学习的特征权重, $G(y_i)$ 是与 y_i 在PCG中的邻居节点集合, 函数 $g(y_i, y_j)$ 是一个指示函数, 表示在PCG中是否存在节点 x_i 到节点 x_j 的边。

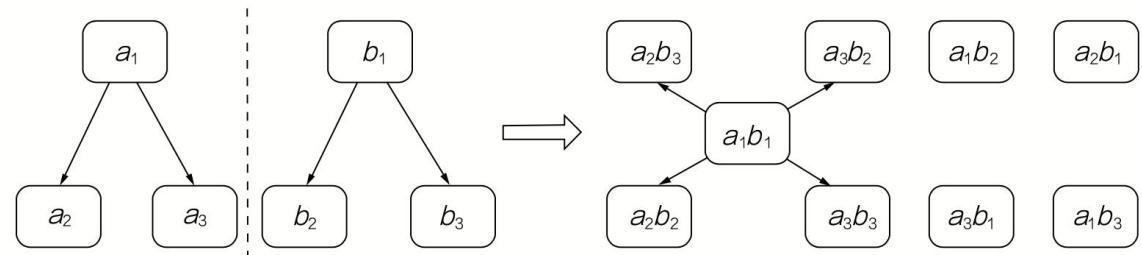


图 6-10 文献[26]提出的将两个知识图谱的引用图转化为 PCG

约束特征

- 第三类为约束特征函数 $h(y_i, H(y_i))$ ，如公式 (6-8) 所示，用来表示PCG中节点之间的约束。这里的约束指的是——对应约束，即一个知识图谱中的实体至多只能和另一个知识图谱中的一个实体等价。

$$\cdot h(y_i, H(y_i)) = \frac{1}{Z_\gamma} \exp \left\{ \sum_{y_j \in H(y_i)} \gamma h(y_i, y_j) \right\} \quad (6-8)$$

- 当 $y_i = 1$ 并且 $y_j = 1$ 时，也就是说两个节点 x_i 和 x_j 中的实体对都是等价的，因此违反了——对应约束，此时 $h(y_i, y_j) = 0$ 。在其他情况下， $h(y_i, y_j) = 1$

属性对齐

- 通用方法

- 属性名称相似性
 - e.g., (英文名, 英文名称)
 - 方法
 - 编辑距离、Jaccard系数、Dice系数
- 同义词相似性
 - e.g., (妻子, 老婆)
 - 利用外部同义词库

- 特有方法

- 数据驱动的属性对齐方法

- 知识图谱中的三元组 (S, P, O)
- 每个属性Property包含多个 Subject-Object pairs
 - 属性值O类型相似度
 - 属性的S-O pairs的overlap程度

KB1

S	P	O
AAA	出生日期	1988-09-01
BBB	出生日期	1966-06-11
CCC	出生日期	1978-11-22

KB2

S	P	O
AAA	生日	1988-09-01
BBB	生日	1966-06-11
CCC	生日	1978-11-22

属性值融合

- 在对齐属性后，需要对来自不同知识图谱的同一实体的同一属性的属性值进行合并
- 属性值融合的任务包括删除重复知识和去除错误知识
- 删除重复知识
 - 属性值的规范化
 - 数值类型的属性值使用同一个标准来表示
 - 单位统一、日期统一等
 - 如果属性值对应一个实体且该实体存在多个名称，则使用统一的实体名称表示

去除错误知识

- 单值属性的属性值融合

- 一个实体的单值属性的值是唯一的
- 举例
 - 出生日期
 - 性别
 - 父亲

- 多值属性的属性值融合

- 一个实体的多值属性的值可能存在多个
- 举例
 - 职业
 - 代表作品
 - 别名

单值属性的属性值融合

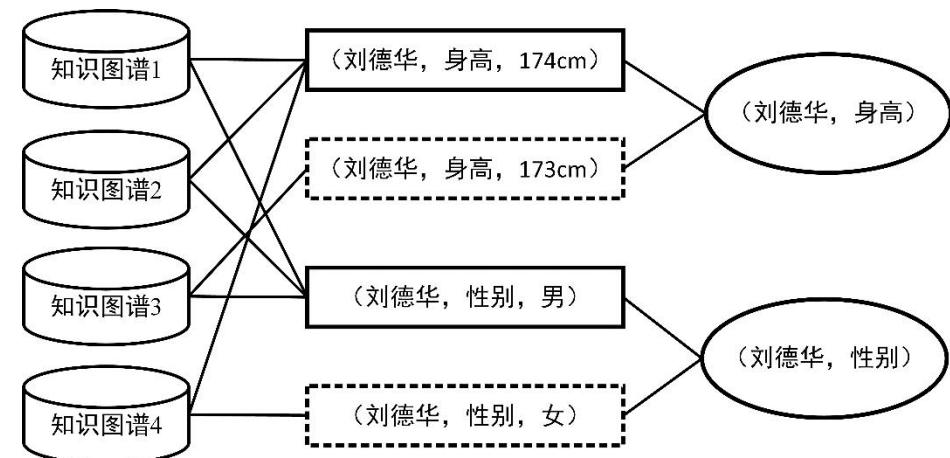
- 基于异构信息网络的真值发现方法

- 三类节点

- 知识图谱
 - 三元组事实
 - 三元组中的实体-属性对（对象）

- 基本思路

- 每个知识图谱都存在大量的（实体，属性，属性值）三元组，可以为多个对象提供属性值
 - 这些三元组的平均准确率决定了知识图谱的质量，而知识图谱的质量也可以用来估计三元组的准确率
 - 当一个对象存在多个属性值时，如果大多数高质量的知识图谱都支持其中某一个属性值，那么这个属性值很有可能就是这个对象的真值



基于异构信息网络的真值发现方法

- 知识图谱质量 VS 事实三元组质量

$$Q(k) = 1 - \frac{\sum_{t \in T(k)} P(t)}{|T(k)|} \quad P(t) = \frac{1}{1 + e^{-\gamma \cdot \sigma^*(t)}}$$

其中， $T(k)$ 为第 k 个知识图谱中的三元组集合， $P(t)$ 为三元组 t 的准确率

三元组 t 的准确率 $P(t)$ 由 $\sigma^*(t)$ 决定 ($P(t)$ 是 $\sigma^*(t)$ 的归一化形式)

$\sigma^*(t)$ 是对三元组 t 的质量的直接度量，其值越大，准确率越高

- $\sigma^*(t)$ 的计算公式如下

$$\sigma^*(t) = \sigma(t) - \rho \cdot \sum_{o(t') = o(t)} \sigma(t')$$

$$\sigma(t) = \sum_{k \in K(t)} Q(k)$$

$\sigma^*(t)$ 由两部分因子组成：

一是来自知识图谱本身的质量（也就是 $\sigma(t)$ ）。当包含 t 的知识图谱可信时， t 也就更可信

二是来自描述实体同一属性但有着不同取值的三元组（所有满足等式 $o(t') = o(t)$ 的三元组 t' ），当实体的某个属性存在不同事实时，这些不同的候选事实之间是此消彼长的关系

多值属性的属性值融合

- 多策略融合
 - 直接合并策略
 - 投票策略
 - 大多数投票
 - 一致性投票
 - 加权投票
 - 自定义融合策略



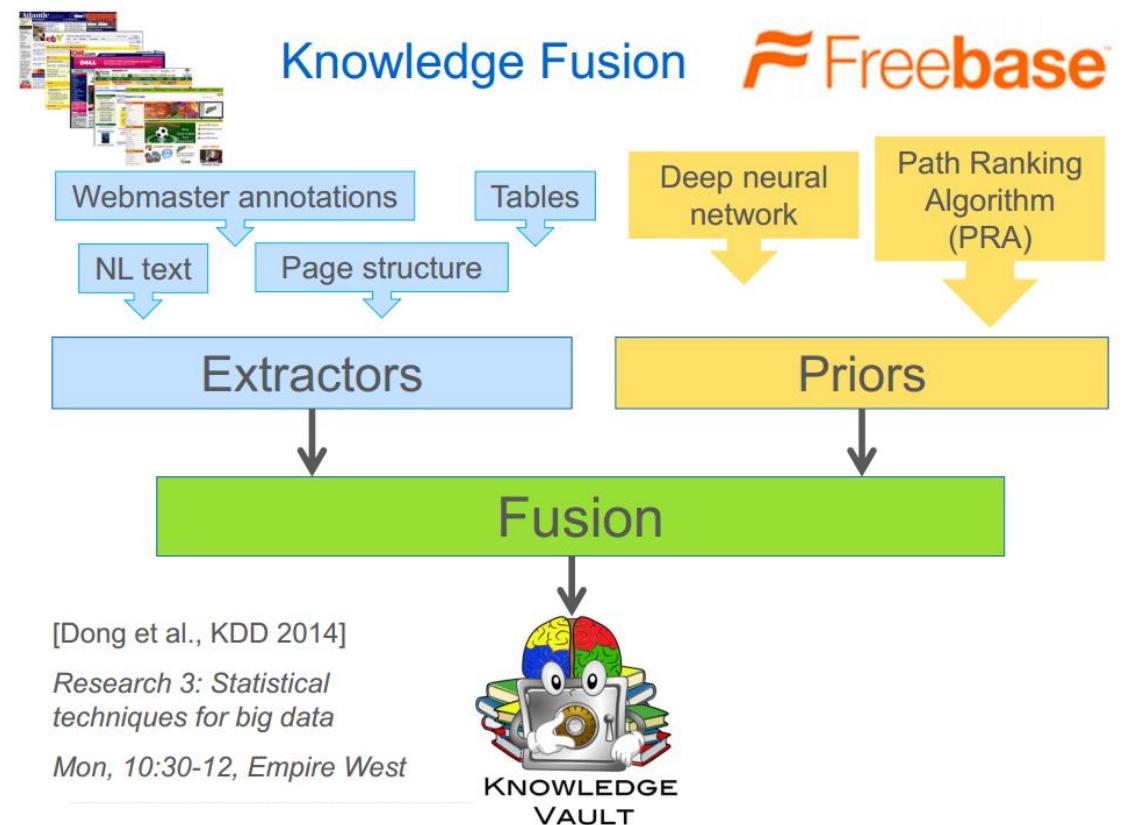
基于多源异构数据的融合方法

基于多源异构数据的融合方法

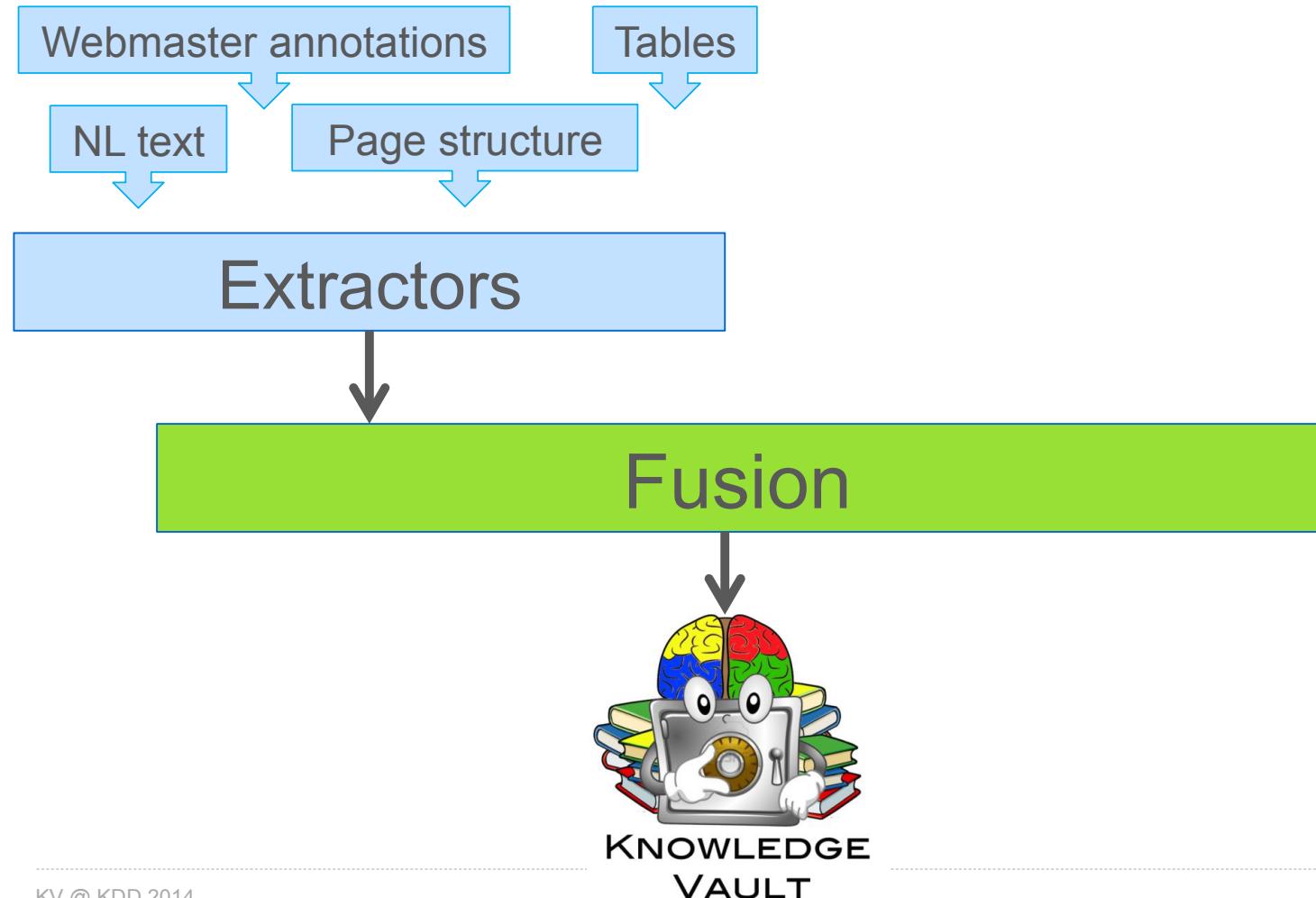
- 输入
 - 多源异构的数据源，包括互联网页面和知识图谱等
- 输出
 - 一个融合后的百科图谱
- 典型代表
 - Knowledge Vault

Knowledge Vault

- 三大组件
 - 知识抽取器
 - 以互联网页面作为数据源进行知识抽取
 - 为知识图谱中的每条关系训练一个抽取器，再利用这些抽取器从多源异构的互联网页面中抽取出更多的知识
 - 知识推理器
 - 从知识图谱自身推理出新知识
 - 知识融合器
 - 从知识抽取器和知识推理器中得到每条知识的最终可信度



Fact extraction from the web



Fact extraction from text (TXT)

- First identify named entities (entity linkage).
- Then classify verb phrase as one of 2000 relations

Patrick Newport ,who has been working at IHS Global Insight, noted...



The result is a probabilistic triple:

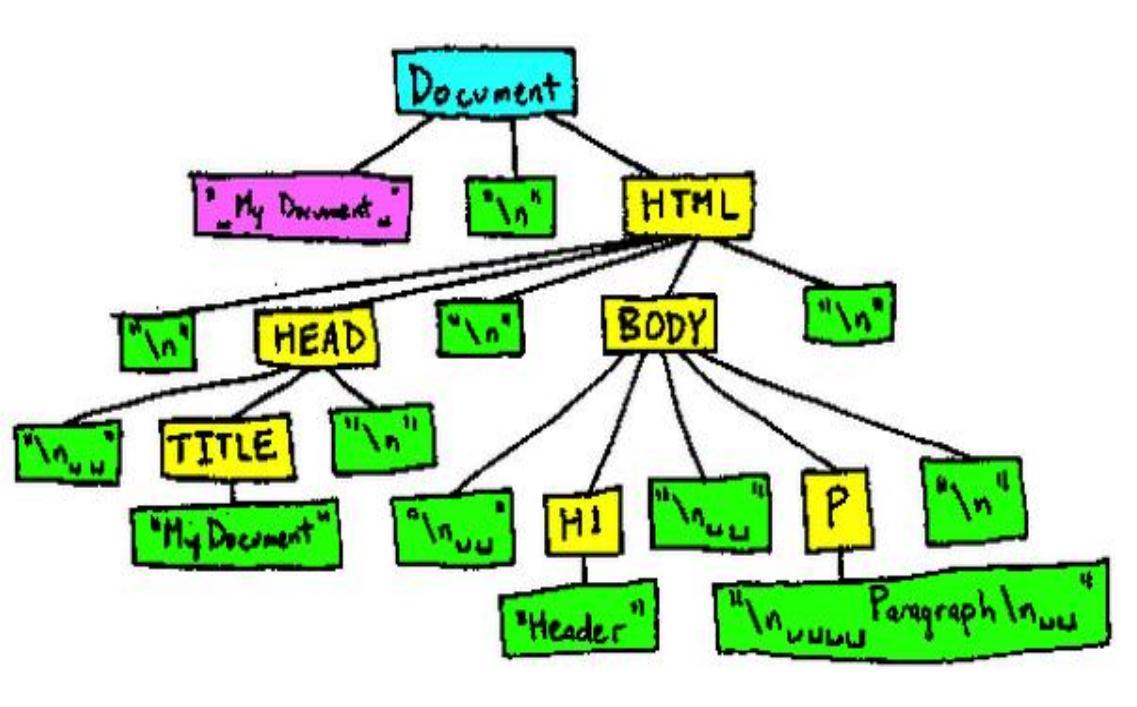
$$\Pr(\langle \text{subject}, \text{reln}, \text{object} \rangle = 1 \mid \text{text})$$

Classifier trained using distant supervision.*

Details: see eg tutorial by Ralph Grishman (NYU):
“Information Extraction: Capabilities and Challenges”,
2012

Fact extraction from DOM trees*

- First identify named entities on page
- Then classify X-path connecting each entity pair as one of 2000 relations

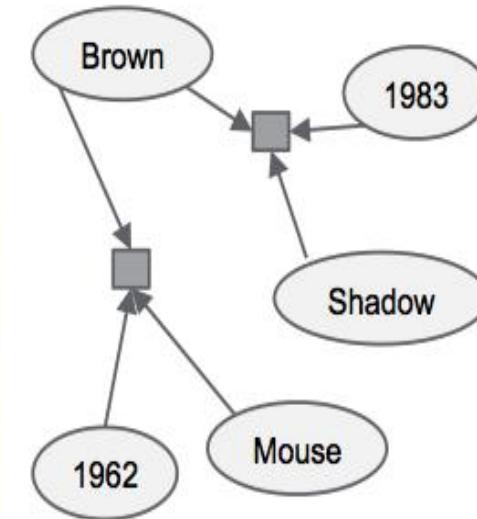


Fact extraction from tables (TBL)*

Caldecott Medal

From Wikipedia, the free encyclopedia

	Illustrator	Title
1999	Mary Azarian	<i>Snowflake Bentley</i>
1954	Ludwig Bemelmans	<i>Madeline's Rescue</i>
1983	Marcia Brown	<i>Shadow</i>
1962	Marcia Brown	<i>Once a Mouse</i>
1955	Marcia Brown	<i>Cinderella, or the Little Glass Slipper</i>
1943	Virginia Lee Burton	<i>The Little House</i>
1980	Barbara Cooney	<i>Ox-Cart Man</i>
1959	Barbara Cooney	<i>Chanticleer and the Fox</i>



Squares are CVT nodes

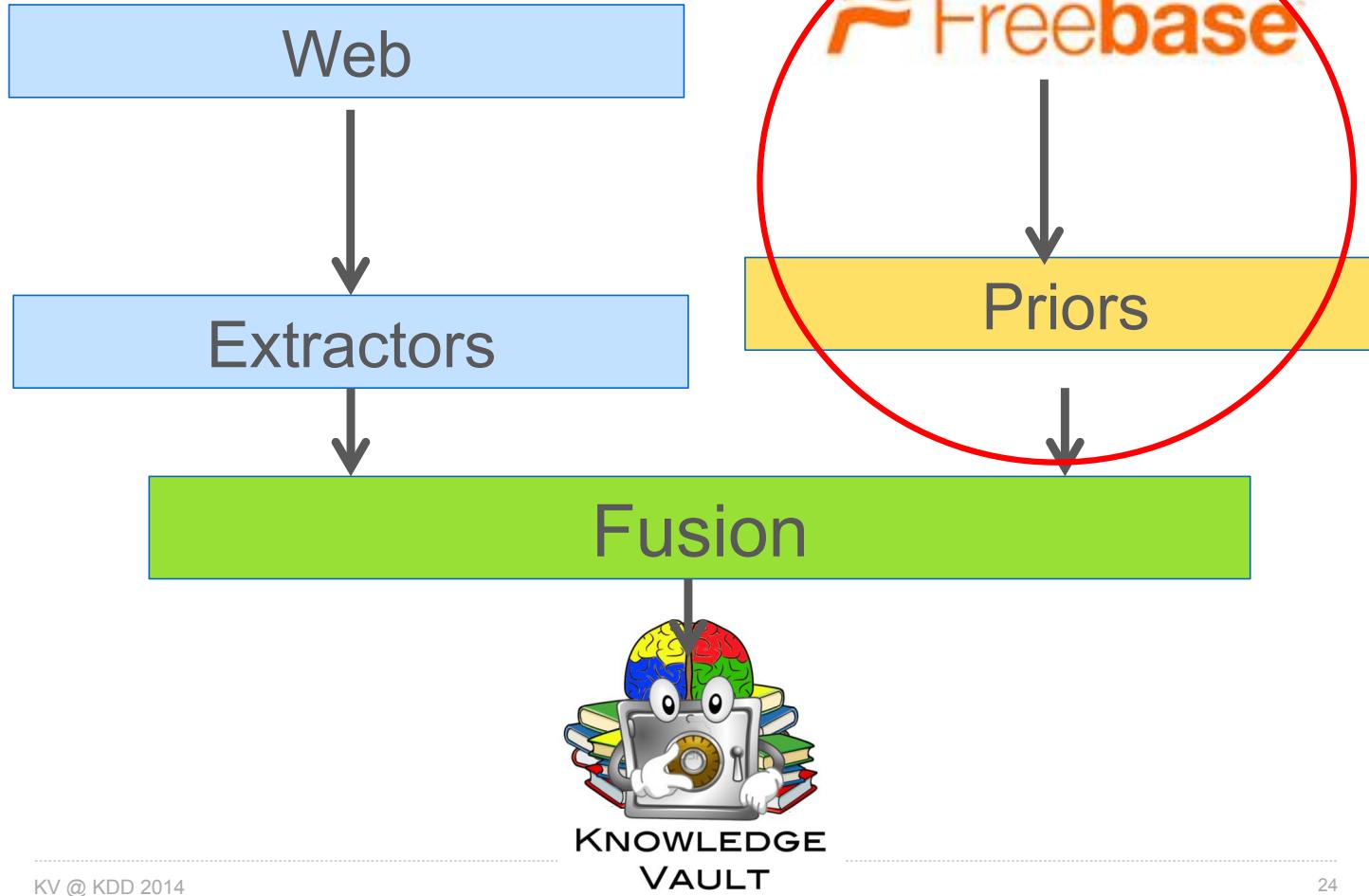
Fact extraction from schema.org annotation (ANO)



```
<script type="application/ld+json">
{"@context": "http://www.schema.org",
 "@type": "Event",
 "startDate": "2014-07-26",
 ...}
</script>
```

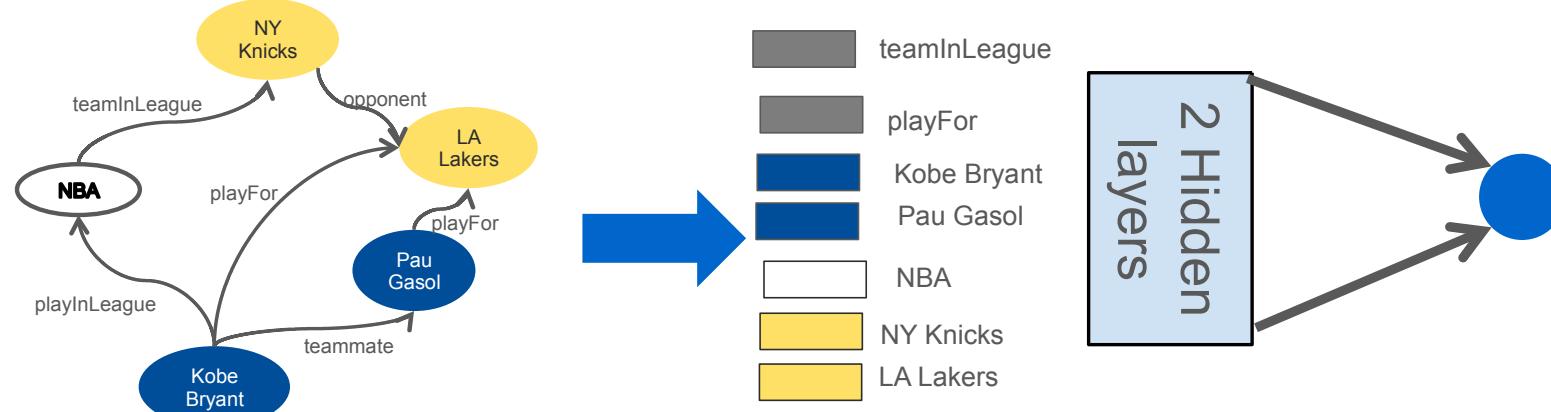
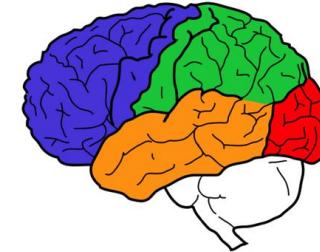
- About 20% of webpages have machine-readable annotations of commercial events, products, etc.
- Automatically map to KG schema.
- We still need to do entity linking.

Mining facts from graphs



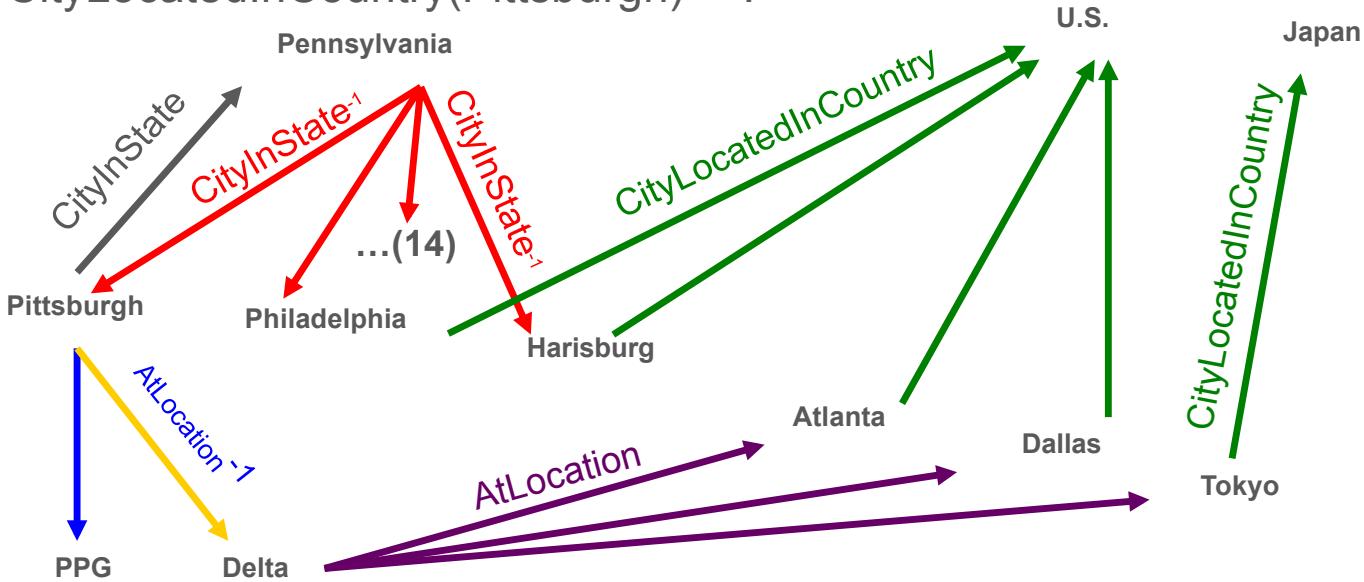
(Deep) neural network for link prediction

- Represent each entity and relation by its own low-dimensional (100D) embedding vector.
- Stack together, feed into neural net.
- Train model to maximize log-likelihood of observed positive and negative triples.
- Outperforms neural tensor model (Socher et al).



Path Ranking Algorithm [Lao et al., EMNLP11]

$\text{CityLocatedInCountry}(\text{Pittsburgh}) = ?$



<u>Feature = Typed Path</u>	<u>Feature Value</u>	<u>Logistic Regression Weight</u>
CityInState, CityInstate^{-1} , $\text{CityLocatedInCountry}$	0.8	0.32
Allocation^{-1} , AtLocation , $\text{CityLocatedInCountry}$	0.6	0.20
...

$\text{CityLocatedInCountry}(\text{Pittsburgh}) = \text{U.S. } p=0.58$

Example of paths / rules learned by PRA

CityLocatedInCountry(*city*, *country*):

7 of the 2985 learned paths

8.04 cityliesonriver, cityliesonriver⁻¹, citylocatedincountry

5.42 hasofficeincity⁻¹, hasofficeincity, citylocatedincountry

4.98 cityalsoknownas, cityalsoknownas, citylocatedincountry

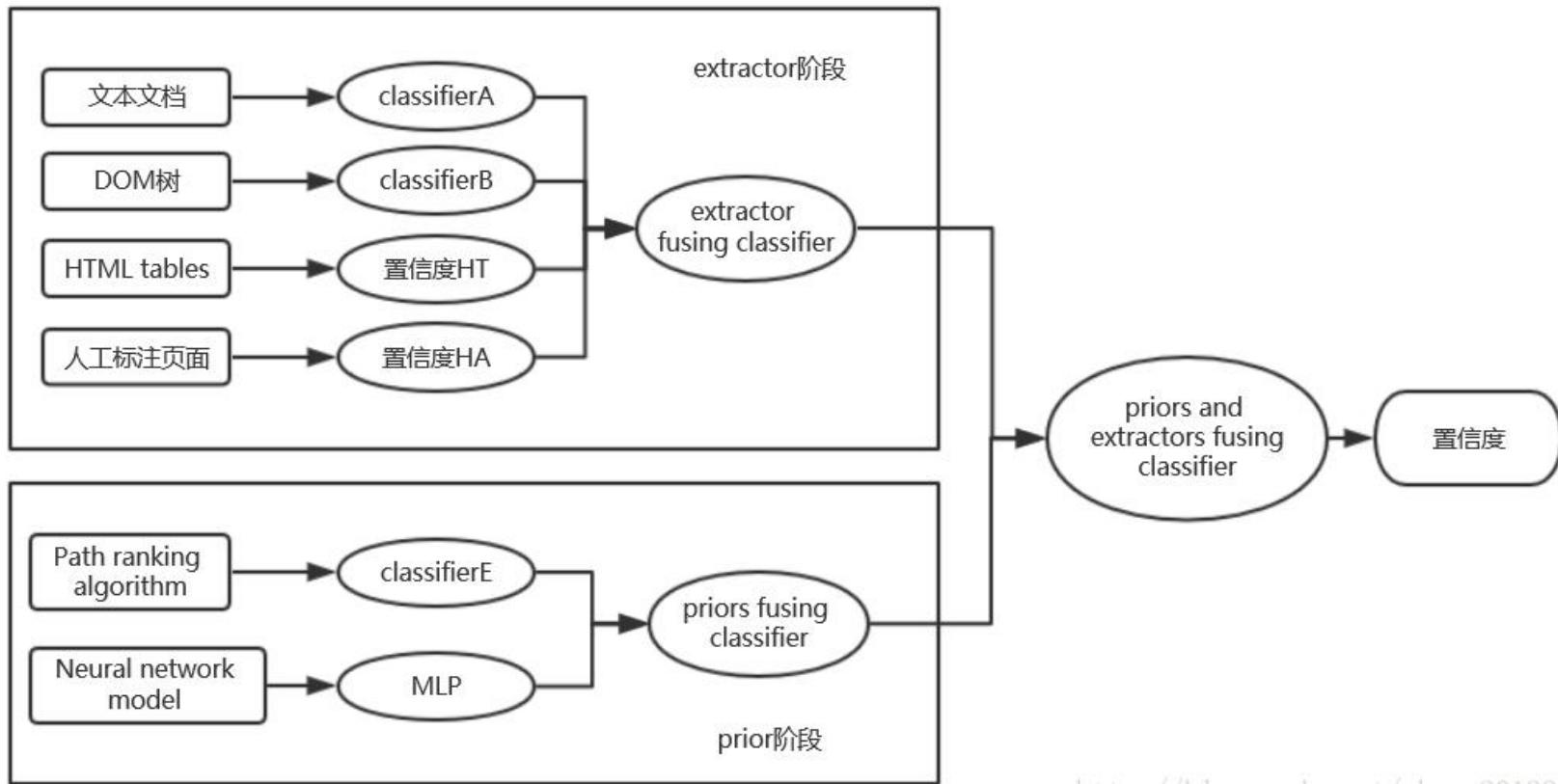
2.85 citycapitalofcountry,citylocatedincountry⁻¹,citylocatedincountry

2.29 agentactsinlocation⁻¹, agentactsinlocation, citylocatedincountry

1.22 statehascapital⁻¹, statelocatedincountry

0.66 citycapitalofcountry

完整流程



<http://blog.csdn.net/zhang201322>

总结

- 基于单源的百科图谱构建
 - 数据获取
 - 属性抽取
 - 关系构建
 - 概念层级体系构建
 - 实体分类
- 基于多源的百科图谱融合
 - 基于多个知识图谱的融合
 - 概念融合
 - 实体对齐
 - 属性对齐
 - 属性值融合
 - 基于多源异构数据的融合
 - 知识抽取器
 - 知识推理器
 - 知识融合器