



# 百科知识图谱构建

徐波

复旦大学知识工场实验室

xubo@fudan.edu.cn

2017-08-26



# CN-DBpedia

- 目前**最大的中文开放通用百科知识图谱之一**
- 涵盖**数千万**实体和**数亿**的关系
  - 百科实体数 1600万+
  - 百科关系数 2.1亿+
- 相关知识服务API累计调用量已达**3.5亿**次

<http://kw.fudan.edu.cn/cndbpedia>



# CN-DBpedia数据开放：DUMP数据



## • 版本号

- 2015.07

## • 规模

- 百科实体数 900万+
- 百科关系数 6700万+
  - mention2entity: 110万+
  - triples: 6600万+
    - 摘要：400万+
    - 标签：1980万+
    - infobox：4100万+

## • 下载地址

- <http://openkg.cn/dataset/cndbpedia>

The screenshot shows the OpenKG.CN website interface. At the top, there is a blue header with the OpenKG.CN logo and the text '中文开放知识图谱'. Below the header, the breadcrumb navigation shows '机构 / 复旦大学 / 中文通用百科知识图谱 (CN-DBpedia)'. The main content area features a large red circular logo of Fudan University on the left, with the text '复旦大学' and '没有关于此机构的描述' below it. To the right of the logo, there are tabs for '数据集', '分类', and '活动流', and a '管理' button. The title '中文通用百科知识图谱 (CN-DBpedia)' is prominently displayed, followed by an '介绍' section. The introduction text states that CN-DBpedia is a large-scale structured encyclopedia developed and maintained by Fudan University's Knowledge Factory Laboratory, and that it is derived from Chinese encyclopedia websites. Below the introduction, there is a section titled '数据与资源' (Data and Resources) which lists four data resources, each with a '浏览' (View) button: 1. '数据门户' (Data Portal) with a '浏览' button; 2. '数据流API' (Data Stream API) with a '浏览' button; 3. 'CN-DBpedia Dump数据 (2015.07)' (CN-DBpedia Dump Data (2015.07)) with a '浏览' button and a description: '包含900万+的百科实体以及6600万+的三元组关系。其中摘要信息400万+, 标签信息1980万+, info...'; 4. 'CN-DBpedia Mention2Entity数据 (2015.07)' (CN-DBpedia Mention2Entity Data (2015.07)) with a '浏览' button and a description: '包含110万+的mention2entity数据'.



# CN-DBpedia数据开放：API接口

- API/mention2entity

api/mention2entity

输入实体名称(mention), 返回CN-DBpedia的对应实体(entity)的列表,json格式。

URL  
[http://knowledgeworks.cn:30001/?p=\\*\\*](http://knowledgeworks.cn:30001/?p=**)

Demo  
<http://knowledgeworks.cn:30001/>

Example  
<http://knowledgeworks.cn:30001/?p=南京>

- API/entityAVP

api/entityAVP

输入实体名, 返回实体全部的知识

URL  
[http://knowledgeworks.cn:20313/cndbpedia/api/entityAVP?entity=\\*\\*](http://knowledgeworks.cn:20313/cndbpedia/api/entityAVP?entity=**)

Example  
<http://knowledgeworks.cn:20313/cndbpedia/api/entityAVP?entity=周杰伦>

# CN-DBpedia应用一：语义搜索

e.g., 复旦大学、周杰伦

Query String: fudan

点击更新页面

Named-Entity Disambiguation: 复旦大学

<http://kw.fudan.edu.cn/cndbpedia>

## Information

复旦大学（Fudan University），简称“复旦”，位于上海市，由中华人民共和国教育部直属，中央直管副部级建制，位列“211工程”、“985工程”，入选“珠峰计划”、“111计划”、“2011计划”、“卓越医生教育培养计划”，为“九校联盟”成员、中国大学校长联谊会成员、东亚研究型大学协会成员、环太平洋大学协会成员、21世纪大学协会成员，是一所综合性研究型的全国重点大学。复旦大学创建于1905年，原名复旦公学，是中国人自主创办的第一所高等院校，创始人为中国近代知名教育家马相伯，首任校董为国父孙中山。校名“复旦”二字选自《尚书大传·虞夏传》名句“日月光华，旦复旦兮”，意在自强不息，寄托当时中国知识分子自主办学、教育强国的希望。1917年复旦公学改名为私立复旦大学；1937年抗战爆发后，学校内迁重庆北碚，并于1941年改为“国立”；1946年迁回上海江湾原址；1952年全国高等学校院系调整后，复旦大学成为以文理科为基础的综合性大学；1959年成为全国重点大学。2000年，原复旦大学与原上海医科大学合并成新的复旦大学。复旦师生谨记“博学而笃志，切问而近思”的校训，严守“文明、健康、团结、奋发”的校风，力行“刻苦、严谨、求实、创新”的学风，发扬“爱国奉献、学术独立、海纳百川、追求卓越”的复旦精神，以服务国家为己任，以培养人才为根本，以改革开放为动力，为实现中国梦作出新贡献。

## Infobox

主管部门	中华人民共和国教育部	<input type="button" value="♥"/> <input type="button" value="🗑"/>
学校代码	10246	<input type="button" value="♥"/> <input type="button" value="🗑"/>
学校地址	上海市杨浦区邯郸路220号	<input type="button" value="♥"/> <input type="button" value="🗑"/>
学校类型	综合	<input type="button" value="♥"/> <input type="button" value="🗑"/>
属性	111计划（2006年）	<input type="button" value="♥"/> <input type="button" value="🗑"/>

## Tag

标签	211高校
标签	985高校
标签	上海高校
标签	专科高校

## Type

rdf.type	< <a href="http://dbpedia.org/ontology/Organisation">http://dbpedia.org/ontology/Organisation</a> >
rdf.type	< <a href="http://dbpedia.org/ontology/EducationalInstitution">http://dbpedia.org/ontology/EducationalInstitution</a> >
rdf.type	< <a href="http://dbpedia.org/ontology/University">http://dbpedia.org/ontology/University</a> >

# CN-DBpedia应用二：小Cui问答



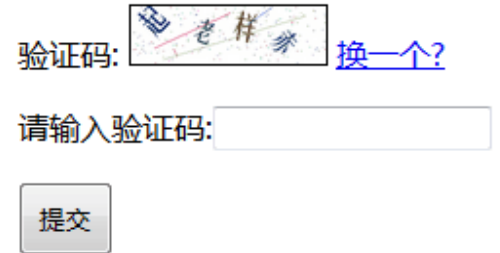
<http://kw.fudan.edu.cn/ddemos/qa/>

增加错误反馈机制，输入1



# CN-DBpedia应用三：超级验证码

- 验证码的作用
  - 区分人和机器
- 一个好的验证码
  - 人很容易识别而机器很难识别
- 传统验证码的困境
  - 随着深度学习在图像领域的快速发展，传统主流图像验证码已经不再安全







# CN-DBpedia应用三：超级验证码

- 人与机器相比，优势在于**阅读理解**
  - 目前机器的语言认知能力还比较弱
- 我们提出了基于知识图谱的验证码系统——超级验证码
- 让用户做“阅读理解”

请通过验证

请点击下文中该问题答案的任意部分：

艾尔伯格迪利安佐酒店的酒店星级是多少？

太难了，换一个

艾尔伯格迪利安佐酒店位于罗马，是家1星级酒店。艾尔伯格迪利安佐酒店让您在罗马这个陌生又熟悉的城市，感受到一丝清浅但又实在的温暖。您一定不能错过。酒店位置较好，距离罗马斗兽场步行22分钟，或打车8分钟，车程约3.6公里。

登录！

Demo地址：<http://kw.fudan.edu.cn/ddemos/vcode/>

API地址：<http://kw.fudan.edu.cn/apis/supervcode/>



# 百科知识图谱背景介绍



# 百科知识图谱



- 是一类专门从**百科类网站**中抽取知识构建而成的知识图谱



WIKIPEDIA  
The Free Encyclopedia

Main page  
Contents  
Featured content  
Current events  
Random article  
Donate to Wikipedia  
Wikipedia store

Interaction  
Help  
About Wikipedia  
Community portal  
Recent changes  
Contact page

Tools  
What links here  
Related changes  
Upload file  
Special pages  
Permanent link  
Page information  
Wikidata item  
Cite this page

Print/export  
Create a book  
Download as PDF  
Printable version

In other projects  
Wikimedia Commons  
Wikinews  
Wikiquote  
Wikisource

Not logged in | Talk | Contributions | Create account | Log in

Article | Talk | Read | View source | View history | Search Wikipedia

## Donald Trump

From Wikipedia, the free encyclopedia  
(Redirected from Trump (president))

*This article is about the incumbent President of the United States. For other uses, see Donald Trump (disambiguation).*

**Donald John Trump** (born June 14, 1946) is an American businessman, television personality, politician, and the 45th President of the United States.

Trump was born and raised in Queens, New York City, and earned an economics degree from the Wharton School of the University of Pennsylvania. He then took charge of The Trump Organization, the real estate and construction firm founded by his paternal grandmother, which he ran for four and a half decades until 2017. During his business career, Trump built, renovated, and managed numerous office towers, hotels, casinos, and golf courses. He has lent the use of his name for the branding of various products and properties. He owned the Miss USA and Miss Universe pageants from 1996 to 2015, and he hosted *The Apprentice*, a reality television series on NBC, from 2004 to 2015. As of 2017, *Forbes* listed him as the 544th wealthiest person in the world (201st in the United States) with a net worth of \$3.5 billion.

Trump first publicly expressed interest in running for political office in 1987. He won two Reform Party presidential primaries in 2000, but withdrew his candidacy early on. In June 2015, he launched his campaign for the 2016 presidential election, and quickly emerged as the front-runner among

Donald Trump



45th President of the United States

Incumbent

**Assumed office**  
January 20, 2017

**Vice President** Mike Pence

**Preceded by** Barack Obama

**Personal details**

**Born**  
Donald John Trump  
June 14, 1946 (age 70)  
New York City

**Political party** Republican (1987–1999, 2009–2011, 2012–present)

**Other political** Reform (1999–2001)



新闻 网页 贴吧 知道 音乐 图片 视频 地图 百科 文库

唐纳德·特朗普

进入词条

搜索词条

帮助

## 唐纳德·特朗普

编辑

收藏 3952 | 747

唐纳德·特朗普（Donald Trump），1946年6月14日生于纽约，美国共和党籍政治家、企业家、商人，第45任美国总统。

1968年从宾夕法尼亚大学沃顿商学院毕业后，进入其父的房地产公司工作，并在1971年开始掌管公司运营，正式进军商界。在随后几十年间，特朗普开始建立自己的房地产王国，人称“地产之王”。除房地产外，特朗普将投资范围延伸到其他行业，包括开设赌场、高尔夫球场等。他还涉足娱乐圈，是美国真人秀《名人学徒》等电视节目的主持人，并担任“环球小姐”选美大赛主席。美国杂志《福布斯》曾评估特朗普资产净值约为45亿美元，特朗普则称超过100亿美元。

特朗普在过去20年间分别支持过共和党和民主党各主要总统竞选者。2015年6月，特朗普以共和党竞选者身份正式参加2016年美国总统选举。此前，特朗普没有担任过公共职务。特朗普结过3次婚，育有5个子女。<sup>[1]</sup>

2016年11月9日，唐纳德·特朗普已获得了276张选举人票，超过270张选举人票的获胜标准，当选美国第45任总统。<sup>[2]</sup>

美国当地时间2017年1月20日中午特朗普在美国首都华盛顿宣誓就职，正式成为美国第45任总统。

人物关系

纠错



中文名	唐纳德·特朗普	毕业院校	宾夕法尼亚大学
外文名	Donald Trump	信仰	基督教
别名	Donald John Trump、川普、川普老爹	主要成就	美国第45任总统
国籍	美国	代表作品	《做生意的艺术》
出生地	美国纽约	性别	男
出生日期	1946年6月14日	星座	双子座
职业	政治家、商人、作家、主持人	现任妻子	梅兰娜·特朗普
		政党	共和党



唐纳德·特朗普图册

词条统计

浏览次数: 15199831次  
编辑次数: 140次历史版本  
最近更新: 昨天  
创建者: jdsbj7sh



# 百科类网站特点

## 一个实体一个页面

- 每个页面均围绕一个实体进行全方面的介绍
- e.g.,
  - [https://en.wikipedia.org/wiki/Donald\\_Trump](https://en.wikipedia.org/wiki/Donald_Trump)
  - <https://baike.baidu.com/item/唐纳德·特朗普>

获取容易

## 包含格式统一的半结构化文本

- 页面格式统一，包含了许多半结构化的数据，方便抽取

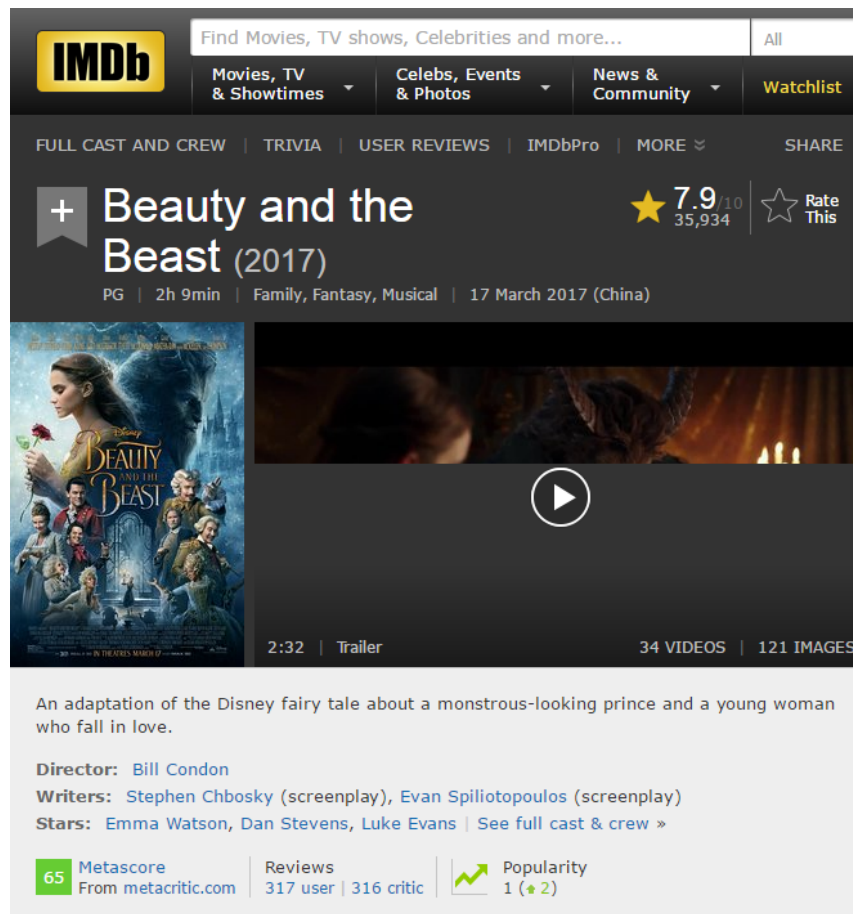
抽取简单

## 内容质量高

- 内容由众包/专业人员编辑，质量相对较高

质量高

百科知识图谱是许多知识图谱构建人士的首选



IMDb Find Movies, TV shows, Celebrities and more... All

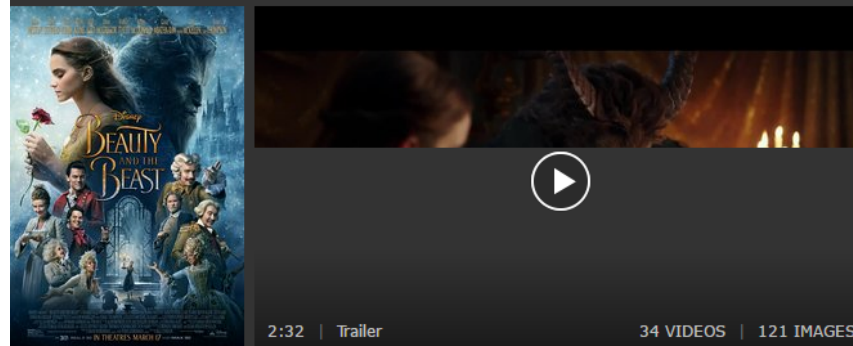
Movies, TV & Showtimes | Celebs, Events & Photos | News & Community | Watchlist

FULL CAST AND CREW | TRIVIA | USER REVIEWS | IMDbPro | MORE | SHARE

## + Beauty and the Beast (2017)

★ 7.9<sup>10</sup>  
35,934 Rate This

PG | 2h 9min | Family, Fantasy, Musical | 17 March 2017 (China)



2:32 | Trailer | 34 VIDEOS | 121 IMAGES

An adaptation of the Disney fairy tale about a monstrous-looking prince and a young woman who fall in love.

**Director:** Bill Condon  
**Writers:** Stephen Chbosky (screenplay), Evan Spiliotopoulos (screenplay)  
**Stars:** Emma Watson, Dan Stevens, Luke Evans | See full cast & crew »

65 Metascore From metacritic.com | 1 Reviews 317 user | 316 critic | Popularity 1 (★2)

<http://www.imdb.com/title/tt2771200>

## 豆瓣电影

电影、影人、影院、电视剧

影讯&购票 | 选电影 | 电视剧 | 排行榜 | 分类 | 影评 | 2016年度榜单 | 2016观影报告

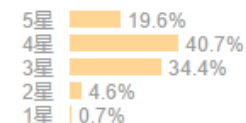
## 美女与野兽 Beauty and the Beast (2017)



导演: 比尔·康顿  
编剧: 斯蒂芬·切波斯基 / 埃文·斯彼里奥托普洛斯 / 琳达·伍尔芙顿 / 珍妮·玛丽·勒普兰斯-德博蒙  
主演: 艾玛·沃森 / 丹·史蒂文斯 / 卢克·伊万斯 / 凯文·克莱恩 / 乔什·加德 / 更多...  
类型: 爱情 / 歌舞 / 奇幻  
制片国家/地区: 美国  
语言: 英语  
上映日期: 2017-03-17(中国大陆/美国)  
片长: 130分钟  
IMDb链接: tt2771200

豆瓣评分

7.5 ★★★★★  
64139人评价



好于 67% 爱情片  
好于 67% 奇幻片

<https://movie.douban.com/subject/25900945/>

# 购物网站



知識工場



解忧杂货店

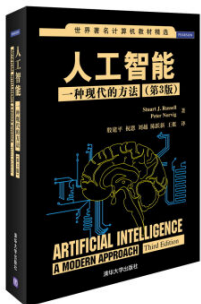
搜索

我的购物车

图书轮播 图书满减 三生三世 通缉令的人 红岩 游戏人间 必经之路 世界虚客

全部商品分类 首页 服装城 美妆馆 京东超市 生鲜 全球购 闪购 拍卖 金融

图书 > 大中专教材教辅 > 大学教材 > 清华大学 > 世界著名计算机教材精选·人工智能：一种现代的方法（第3版）



世界著名计算机教材精选·人工智能：一种现代的方法（第3版） [Artificial Intelligence: a Modern Approach, Third Edition]  
看AlphaGo（阿尔法狗）是如何成长的，美国伯克利大学与Google人工智能科学家合作编写，全世界100多个国家1200多所大学使用。A Must Read for AI

[美] 罗素 (Stuart J. Russell), [美] 诺维格 (Peter Norvig) 著; 殷建平, 祝恩, 刘越 等译



清华大学出版社 京东自营

进店逛逛 关注店铺

服务支持:

④ 京准达 ⑤ 夜间配

⑥ 自提

文轩网旗舰店 ¥88.70

思语华教图书专营店 ¥90.10

润知天下图书专营店 ¥92.16

查看全部商家

京东价: **¥105.70** [8.3折] [定价: ¥128.00] (降价通知) 累计评价 1100+

排名: 自营 大中专教材教辅销量榜 第 37 位

配送至: 上海松江区泗泾镇 | 有货, 支持 99元免基础运费 | 货到付款

服务: 由 京东 发货, 并提供售后服务, 23:00前完成下单, 预计明天(03月23日)送达

选择系列: Wolfram 语言基础入门 Mathematica基础培训教程 **人工智能(第3版)**

白条分期: 30天免息 ¥35.76×3期 ¥18.15×6期 ¥9.34×12期 ¥4.93×24期

加入购物车

温馨提示: 1. 支持7天无理由退货

31个商家在售 ¥88.70 起

相关分类

- 大学教材 研究生教材
- 高职高专教材 中职中专教材
- 成人教育教材 竞赛
- 职业培训教材

商品介绍 商品评价(1100+)

出版社: 清华大学出版社 ISBN: 9787302331094 版次: 3 商品编码: 11343660  
品牌: 清华大学 包装: 平装 丛书名: 世界著名计算机教材精选 外文名称: Artificial Intelligence: a...  
开本: 16开 出版时间: 2013-11-01 用纸: 胶版纸 页数: 918  
字数: 1462000 正文语种: 中文

https://item.jd.com/11343660.html



图书

浏览 全部商品分类 我的亚马逊 Z秒杀 礼品卡 我要开店 海外购 帮助 In English

图书 高级搜索 所有分类 新品排行榜 销售排行榜 新书店 教材教辅 少儿 文学 小说 历史 经营 励志 人文社科 生活 科普  
您可以在Kindle设备上阅读及

图书, 小说, 推理小说

在线阅读



解忧杂货店 精装 - 2014年5月1日

东野圭吾 (作者), 李盈春 (译者)

★★★★☆ 9,690 条商品评论 | 分享

显示所有 2 格式和版本

Kindle电子书 ¥11.85 精装 ¥27.30

使用我们的 免费Kindle阅读软件

促销信息: 满减 中文图书全场满200元减50元 共3个促销

配送至: 上海黄浦区 现在有货

送达日期: 明天(3月23日), 请在1小时17分钟内下单并选择“快速送货上门”。

(精确送达时间请于详情页查询)

目前上海市地区超1100家好德、可的便利店提供自提服务。详情

销售配送: 由亚马逊直接销售和发货。

新品12 售价从 ¥27.30起

退换承诺: 此商品支持30天免费退换 详情



查看全部 6 张图片

基本信息

出版社: 南海出版公司; 第1版 (2014年5月1日)

外文书名: ナミヤ雑貨店の奇蹟

精装: 291页

语种: 简体中文

开本: 32

ISBN: 7544270874, 9787544270878

条形码: 9787544270878

商品尺寸: 21 x 15 x 2 cm

商品重量: 499 g

品牌: 新经典文化

ASIN: B00JZ96Z18

用户评分: ★★★★★ (9,690 条商品评论)

亚马逊热销商品排名: 图书商品里排名第2名 (查看全部商品销售排行榜)

第1位 - 图书 > 在线阅读

第3位 - 图书 > 小说 > 外国现当代小说

第2位 - 图书 > 小说 > 推理小说

https://www.amazon.cn/%E5%9B%BE%E4%B9%A6/dp/B00JZ96Z18

# 工商信息网站

https://www.tianyancha.com/company/68521782



请输入企业名称、人名、品牌等关键词



康成投资（中国）有限公司

企业信用报告

电话: 02126100749 邮箱: fn18@mail.rt-mart.com.cn

网址: www.rt-mart.com.cn

地址: 上海市共和新路3318号

分享

立即更新

天眼查数据更新时间  
— 2017.03.18 —

我要投诉

企业背景	企业发展	司法风险	经营风险	经营状况	知识产权			
基本信息	变更记录 7	融资历史 0	法律诉讼 16	经营异常 0	欠税公告 0	招投标 0	抽查检查 0	商标信息 99+
企业关系	企业年报 3	核心团队 0	法院公告 2	行政处罚 0	债券信息 0	产品信息 0	资质证书 0	专利 1
主要人员 7	分支机构 0	企业业务 1	失信人 0	严重违法 0	购地信息 0	资质证书 0		著作权 1
股东信息 2	投资事件 0	被执行人 2	股权出质 0	招聘 87				网站备案 1
对外投资 99+	竞品信息 10	动产抵押 0		税务评级 2				

## 企业背景

基本信息:

法定代表人  
黄明端 [他的所有公司](#)

注册资本  
24788.6996万美元

注册时间  
2005-03-23

状态  
存续(在营、开业、...

工商注册号: 310000400523532

组织机构代码: 717854767

统一社会信用代码: 91310000717854767L

企业类型: 有限责任公司(台港澳与境内合资)

行业: 商务服务业

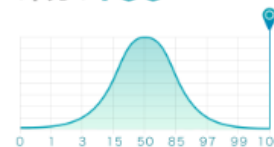
营业期限: 2005-03-23至2055-03-22

核准日期: 2005-03-23

登记机关: 上海市工商局

注册地址: 上海市共和新路3318号

评分 100



[ 以上评分结果仅供参考 ]

经营范围: 一、在国家允许外商投资的领域依法进行投资;二、受其所投资企业的书面委托(经董事会一致通过)向其提供下列服务1、协助或代理其所投资企业从国内外采购该企业自用的机器设备、办公设备和生产所需的原材料、元器件、零部件和在国内外销售其所投资企业生产的产品,并提供售后服务;2、在外汇管理部门的同意和监督下,在其所投资企业之间平衡外汇;3、为其所投资企... [详细](#)



# 法律网站

https://www.itslaw.com/detail?judgementId=d4944de5-1e0a-4ffa-be9b-b679ff6e38fa

陈渭庭与北京市人民政府其他二审行政判决书

北京市高级人民法院 | 二审 | (2016)京行终5748号

【关键词】 拆迁

【文书来源】 北京法院审判信息网

引用法规 \* 本处法规摘自法院观点

## 二审

- 《中华人民共和国行政诉讼法》第八十九条第一款第(一)项 (85000)

检索相关案例

## 文书正文

### 当事人信息

上诉人(一审原告)陈渭庭,男,1928年10月12日出生。

委托代理人陈燕华(陈渭庭之女),女,1963年1月18日出生,住北京市朝阳区。

委托代理人周彤,北京来硕律师事务所律师。

被上诉人(一审被告)北京市人民政府,住所地北京市东城区正义路2号。

法定代表人蔡奇,市长。

委托代理人王仰东,北京市人民政府法制办公室干部。

委托代理人任佳慧,北京高文律师事务所律师。

### 审理经过

上诉人陈渭庭因行政复议不予受理决定一案,不服北京市第二中级人民法院(2016)京02行初78号行政判决,向本院提起上诉。本院受理后依法组成合议庭审理了本案。本案现已审理终结。

收藏 分享 下载

### 快速目录

- 文书正文
- 当事人信息
- 审理经过
- 一审法院认为
- 本院查明
- 本院认为
- 二审裁判结果
- 审判人员
- 裁判日期
- 书记员

### 基本信息

审理法院 北京市高级人民法院

案号 (2016)京行终5748号

案件类型 行政

案由 行政裁决

裁判日期 2017-03-16

合议庭 哈胜男 孙建 姜宇红

审理程序 二审

上诉人 陈渭庭

被上诉人 北京市人民政府

上诉人代理律师 周彤 北京来硕律师事务所

被上诉人代理律师 任佳慧 北京高文律师事务所



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

# 领域百科知识图谱

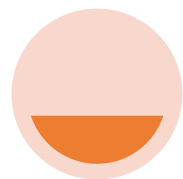


## 资源分类

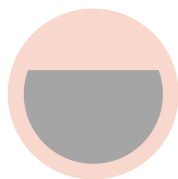


中文开放知识图谱 <http://openkg.cn>

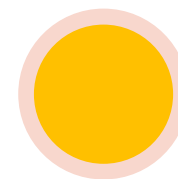
# 单百科数据源的百科知识图谱构建



知识获取



知识填充



知识更新



# 第一部分：知识获取



# 知识获取



知识抽取



数据清洗



# 知识抽取

## 结构化数据的知识抽取

- e.g., 数据库

## 半结构化数据的知识抽取

- e.g., 表格

## 非结构化数据的知识抽取

- e.g., 文本





# 知识抽取

## 结构化数据的知识抽取 ( T5 知识图谱虚拟化 )

- e.g., 数据库

## 半结构化数据的知识抽取 ( T2 百科知识图谱构建 )

- e.g., 表格

## 非结构化数据的知识抽取 ( T3 知识获取方法 )

- e.g., 文本



# 百科知识从哪来 (一)

多义词

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、词作人
- 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
- 原民航局空中交通管理局局长助理
- 四川省广安经济技术开发区国家税务局副局长
- 湖北监利籍烈士
- 清华大学教授
- 新疆青少年出版社出版的著作
- 通川区学生资助中心主任

标题

## 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、词作人）

同义词

刘德华（Andy Lau），1961年9月27日出生于中国香港，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>。1985年因拒签五年合约而被TVB雪藏<sup>[4]</sup>。1988年将事业重心转向电影<sup>[5]</sup>。1991年创办天幕电影公司<sup>[6]</sup>。1994年担任剧情片《天与地》的制片人<sup>[7]</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>[8]</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>[9]</sup>。2005年获得香港UA院线颁发的“1985-2005年全港最高累积票房香港男演员”奖<sup>[10]</sup>。2006年获得釜山国际电影节亚洲最有贡献电影人奖<sup>[11]</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>[12]</sup>；同年担任第49届台湾电影金马奖评审团主席<sup>[13]</sup>。2017年主演警匪动作片《拆弹专家》<sup>[14]</sup>。

摘要

1985年发行首张个人专辑《只知道此刻爱你》<sup>[15]</sup>。1990年凭借专辑《可不可以》在歌坛获得关注<sup>[16]</sup>。1994年获得十大劲歌金曲最受欢迎男歌星奖<sup>[17]</sup>。1995年在央视春晚上演唱歌曲《忘情水》<sup>[18]</sup>。1997年与那英合唱《东方之珠》<sup>[19]</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>[20]</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>[21]</sup>。

1994年创立刘德华慈善基金会<sup>[22]</sup>。2000年被评为世界十大杰出青年<sup>[23]</sup>。2005年发起亚洲新星计划<sup>[24]</sup>。2008年被委任为香港非官守太平绅士<sup>[25]</sup>。2016年连任中国残疾人福利基金会副理事长。<sup>[26]</sup>



图集



# 百科知识从哪来 (二)

是一组 (属性, 属性值) 对  
是对实体的结构化总结

## 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 中国电影百年形象大使 釜山电影节亚洲最有贡献电影人奖
国籍	中国		
民族	汉族		
星座	天秤座		
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

Infobox

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



# 百科知识从哪来（三）

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，与叶德嫻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》中饰演骁勇善战的杨六郎 [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]。

相关实体

标签

词条标签： 音乐人物， 演员， 歌手， 娱乐人物， 制作人， 人物

Ontology : 严格Is-A关系  
Taxonomy : 非严格Is-A关系  
Folksonomy : 开放分类

From CCKS T1 : 知识图谱导论



# 百科知识从哪来（四）

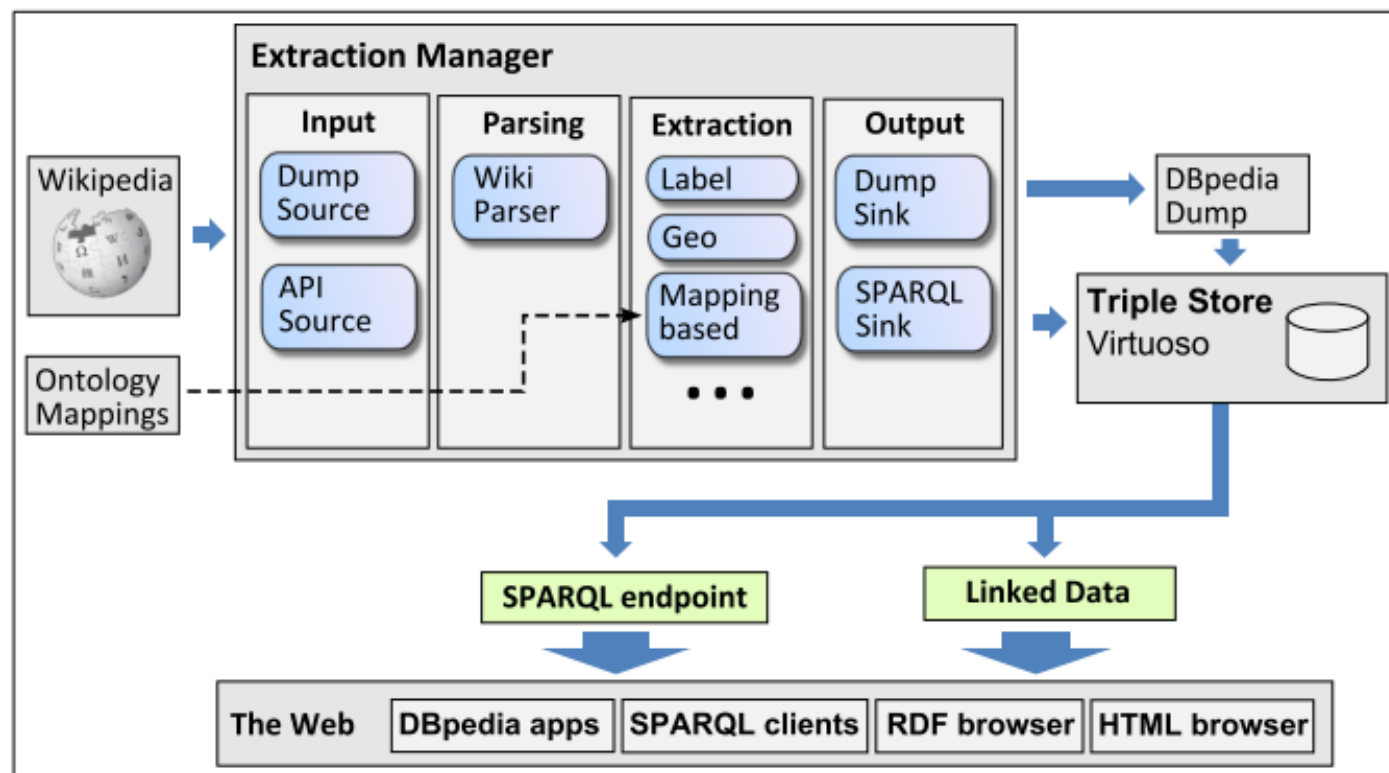
## • 不同百科网站能得到的知识也不尽相同

Name	Description	Example
abstract	Extracts the first lines of the Wikipedia article.	<code>dbr:Berlin dbo:abstract "Berlin is the capital city of (...)".</code>
article categories	Extracts the categorization of the article.	<code>dbr:Oliver_Twist dc:subject dbr:Category:English_novels.</code>
category label	Extracts labels for categories.	<code>dbr:Category:English_novels rdfs:label "English novels".</code>
category hierarchy	Extracts information about which concept is a category and how categories are related using the SKOS Vocabulary.	<code>dbr:Category:World_War_II skos:broader dbr:Category:Modern_history.</code>
disambiguation	Extracts disambiguation links.	<code>dbr:Alien dbo:wikiPageDisambiguates dbr:Alien_(film).</code>
external links	Extracts links to external web pages related to the concept.	<code>dbr:Animal_Farm dbo:wikiPageExternalLink &lt;http://books.google.com/?id=RBGmrDnBs8UC&gt;.</code>
geo coordinates	Extracts geo-coordinates.	<code>dbr:Berlin georss:point "52.5006 13.3989".</code>
grammatical gender	Extracts grammatical genders for persons.	<code>dbr:Abraham_Lincoln foaf:gender "male".</code>
homepage	Extracts links to the official homepage of an instance.	<code>dbr:Alabama foaf:homepage &lt;http://alabama.gov/&gt;.</code>
image	Extracts the first image of a Wikipedia page.	<code>dbr:Berlin foaf:depiction &lt;http://.../Overview_Berlin.jpg&gt;.</code>
infobox	Extracts all properties from all infoboxes.	<code>dbr:Animal_Farm dbo:date "March 2010".</code>
interlanguage	Extracts interwiki links.	<code>dbr:Albedo dbo:wikiPageInterLanguageLink dbr-de:Albedo.</code>
label	Extracts the article title as label.	<code>dbr:Berlin rdfs:label "Berlin".</code>

# 知识抽取框架



- 为每类关系建立一个抽取器



[Jens Lehmann et. al., 2015]

Fig. 1. Overview of DBpedia extraction framework.

# 数据清洗



## 基本信息

中文名	复旦大学	主管部门	中华人民共和国教育部
英文名	Fudan University	硕士点	243个
简称	复旦 FUDAN	博士点	154个
创办时间	1905年(乙巳年)9月14日	博士后流动站	35个
类别	公立大学	校训	博学而笃志，切问而近思
学校类型	综合	校歌	《复旦大学校歌》
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)	专职院士	中国科学院院士 21人 中国工程院院士 5人
所在地区	中国 上海	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
现任校长	许宁生	展开 国家重点学科	一级学科 11个，二级学科 19个
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	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)

Q 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

属性不一致

数值属性值格式不统一

多个对象属性值未分割





# 单数据源属性融合

找到候选属性对

- 属性名称相似性
  - e.g., 英文名, 英文名称
- 外部知识库
  - e.g., 妻子, 老婆
- 人工录入

删除错误属性对

- 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
  - ... ..
- 人工删除



# 数值属性值归一化

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]*(\d){1,2}[\^d]*\.[\^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 | <>

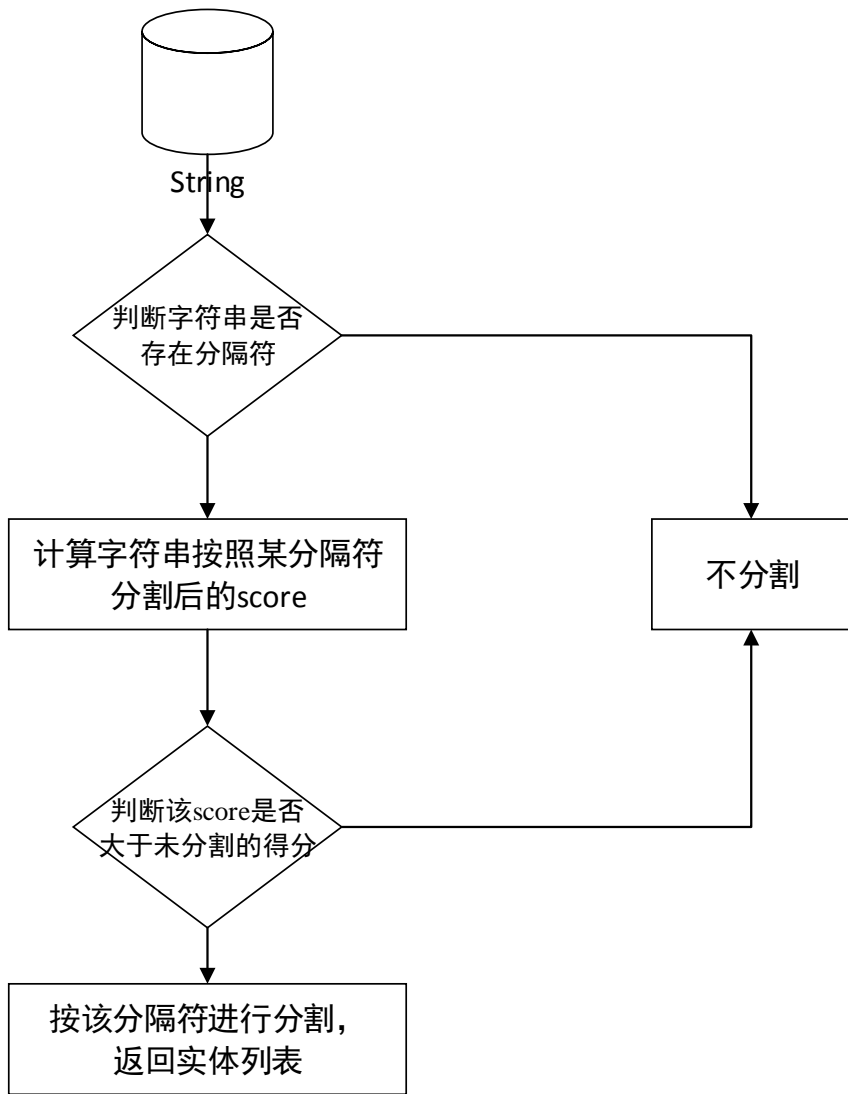
1 千米(km) ⇄ 米(m)

1千米(km)=1000米(m)

国际单位：米(m)

# 对象属性值分割

## • 流程



## • 分割效果打分函数Score

- 不分割的score设为1.1
- 评估每种分隔符分割后的结果
  - 初始值设为0
  - 假设根据分隔符可以将值分为K个部分，对于任意一个子部分：
    - 如果该子部分是一个实体，score+1
    - 反之，score-1

## • 取score最大的方案作为最终的分割方案

# 举例



知識  
工場

## • 属性融合

- Old
  - ( 复旦大学, 英文名, Fudan University )
- New
  - ( 复旦大学, 英文名称, Fudan University )

## • 数值属性值归一化

- Old
  - ( 复旦大学, 创办时间, 1905年(乙巳年)9月14日 )
- New
  - ( 复旦大学, 创办时间, 1905年09月14日 )

## • 对象属性值分割

- Old
  - ( 复旦大学, 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 )
- New
  - ( 复旦大学, 创办时间, 李岚清 )
  - ( 复旦大学, 创办时间, 朱民 )
  - ( 复旦大学, 创办时间, 李源潮 )
  - ( 复旦大学, 创办时间, 竺可桢 )
  - ( 复旦大学, 创办时间, 于右任 )
  - ( 复旦大学, 创办时间, 邵力子 )
  - ( 复旦大学, 创办时间, 王沪宁 )

# 第二部分：知识填充





# 百科知识图谱遇到的挑战

## 知识缺失

- 单个实体的知识缺失
- 实体分类知识缺失
- 概念表示知识缺失

# 举例一：单个实体的知识缺失



中文名	周杰伦	职业	歌手、音乐人、制作人、导演、商人
外文名	Jay Chou	毕业院校	淡江中学
别名	周董	经纪公司	杰威尔音乐有限公司
国籍	中国	代表作品	龙卷风、简单爱、七里香、青花瓷、稻香、告白气球、头文字D、不能说的秘密、逆战、青蜂侠、天台
民族	汉族	主要成就	获得十五座金曲奖
星座	摩羯座		两届台湾金曲奖最佳国语男歌手
血型	O型		四届世界音乐大奖最畅销中华区艺人
身高	175cm		入选美国CNN亚洲25位最具影响力人物 <sup>[23]</sup>
出生地	台湾省新北市		《Fast Company》全球百大创意人物 ~ 展开
出生日期	1979年1月18日		

<https://baike.baidu.com/item/周杰伦>

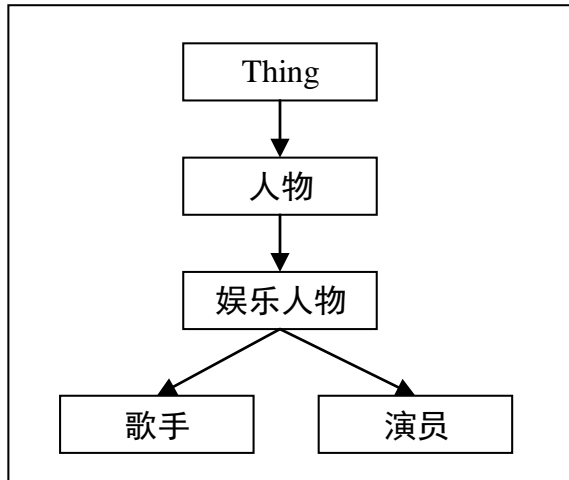
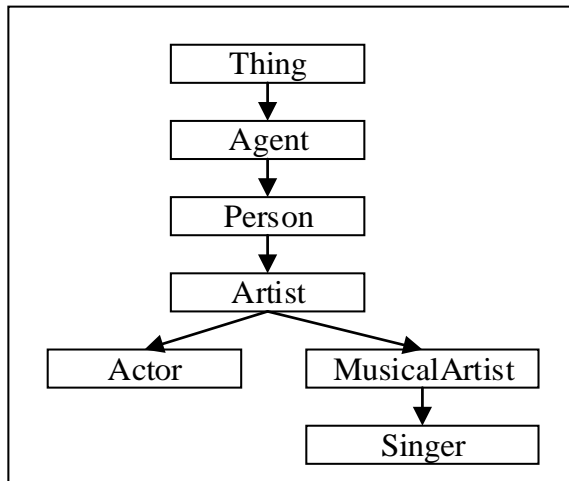
中文名	刘德华	经纪公司	东亚唱片、映艺娱乐
外文名	Andy Lau, Lau Tak Wah	代表作品	暗战、无间道、天若有情、旺角卡门、桃姐、来生缘、忘情水、谢谢你的爱、冰雨、今天、爱你一万年
别名	华仔, 华Dee, 华哥等	主要成就	三届香港电影金像奖最佳男主角 两届台湾电影金马奖最佳男主角 1985-2005年全港最高累积票房香港男演员奖 中国电影百年形象大使 釜山电影节亚洲最有贡献电影人奖 ~ 展开
国籍	中国		
民族	汉族		
星座	天秤座		
血型	AB型		
身高	174cm		
体重	63kg	妻子	朱丽倩
出生地	香港新界大埔镇泰亨村	女儿	刘向慈
出生日期	1961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

<https://baike.baidu.com/item/刘德华>

# 举例二：实体分类知识缺失



From Folksonomy to Ontology







# 举例三：概念表示知识缺失

- 概念的关系较为有限
  - 上下位关系SubclassOf
  - 类属关系Type
- 概念知识缺失
  - 概念是怎么形成的？

Pages in category "English-language films"

The following 200 pages are in this category, out of approximately 54,988 total. This list may not reflect recent changes ([learn more](#)).

([previous page](#)) ([next page](#))

<p>+</p> <ul style="list-style-type: none"> <li>• +1 (film)</li> </ul> <p><b>0-9</b></p> <ul style="list-style-type: none"> <li>• 1 (2013 film)</li> <li>• 1 a Minute</li> <li>• 1 Day</li> <li>• 1 Night in China</li> <li>• 1 Night in Paris</li> <li>• The 1 Second Film</li> <li>• 1-2-3 Go</li> <li>• 1:42.08</li> <li>• 2:22 (2008 film)</li> <li>• 2 Cool 2 Be 4gotten</li> <li>• 2 Days in New York</li> <li>• 2 Days in Paris</li> <li>• 2 Days in the Valley</li> </ul>	<ul style="list-style-type: none"> <li>• 4 Days in May</li> <li>• 4 Devils</li> <li>• 4 for Texas</li> <li>• 4 Little Girls</li> <li>• 4 Minute Mile</li> <li>• 4 Play (film)</li> <li>• 4:30</li> <li>• 4.3.2.1.</li> <li>• 4Chosen: The Documentary</li> <li>• 4D Man</li> <li>• The 4th Dimension (film)</li> <li>• The 4th Floor (1999 film)</li> <li>• 4th Man Out</li> <li>• The 4th Tenor</li> <li>• 5 Against the House</li> <li>• 5 Card Stud</li> <li>• 5 Card Stud (2002 film)</li> <li>• 5 Days of War</li> </ul>
---	---

[https://en.wikipedia.org/wiki/Category:English-language\\_films](https://en.wikipedia.org/wiki/Category:English-language_films)

# 解决方案



## 单个实体的知识缺失

- 实体属性-值关系填充 ( infobox completion )

## 实体分类知识缺失

- 实体分类

## 概念表示知识缺失

- 概念符号表示

## 2.1 实体属性-值关系填充 ( infobox completion )



# 方法总结



利用其它知识图谱进行填充

利用百科网站实体标签进行填充

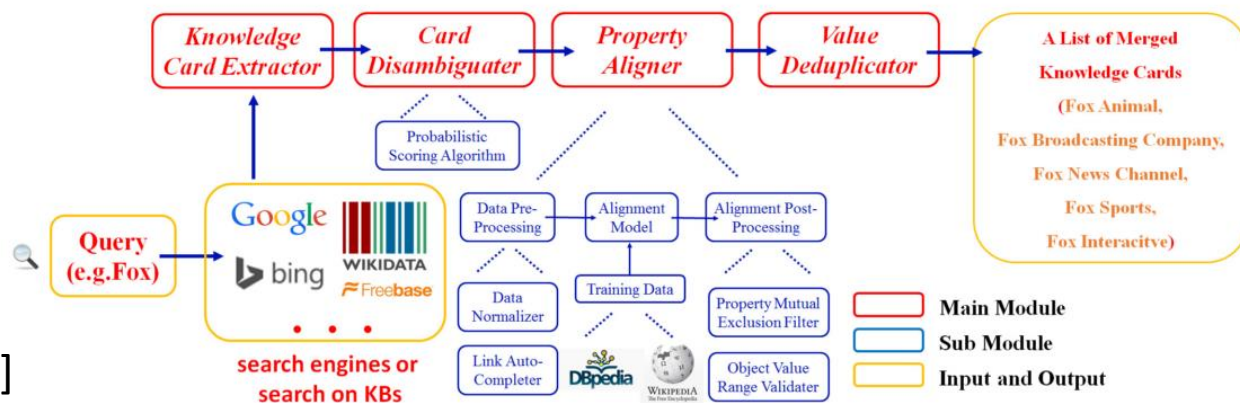
利用百科网站文本进行填充

# 利用其它知识图谱进行填充

- 知识图谱融合
- 基本思路
  - 不同知识图谱的实体之间可以通过等价关系“SameAS”链接在一起
  - 不同知识图谱由于构建方式不同，知识也不尽相同
  - 因此，可利用其它知识图谱来对自身知识图谱Infobox进行填充

## • 难点

- 实体匹配
- 属性匹配
- 属性值融合
- 跨语言 [Bouma, G. et al., 2009]



[Haofen Wang et al. 2015]

# 利用百科网站实体标签进行填充



- 百科网站的标签信息是描述实体的一个重要信息
  - E.g. 如“刘德华”的一个标签信息为“香港男演员”，可以推出
    - ( 刘德华, 出生地, 香港 )
    - ( 刘德华, 性别, 男 )
    - ( 刘德华, 职业, 演员 )
- 目前主流方法倾向于人工建立规则来从标签信息中抽取关系
  - YAGO
  - Catriple
- 基于统计的方法
  - DFs ( 详见概念表示部分 )

- YAGO提出了基于正则表达式从标签信息中抽取关系的方法
- 优点
  - 准确率高
- 缺点
  - 代价大
    - 需要为每个关系定制一套正则表达式

Table 1: Some Category Heuristics

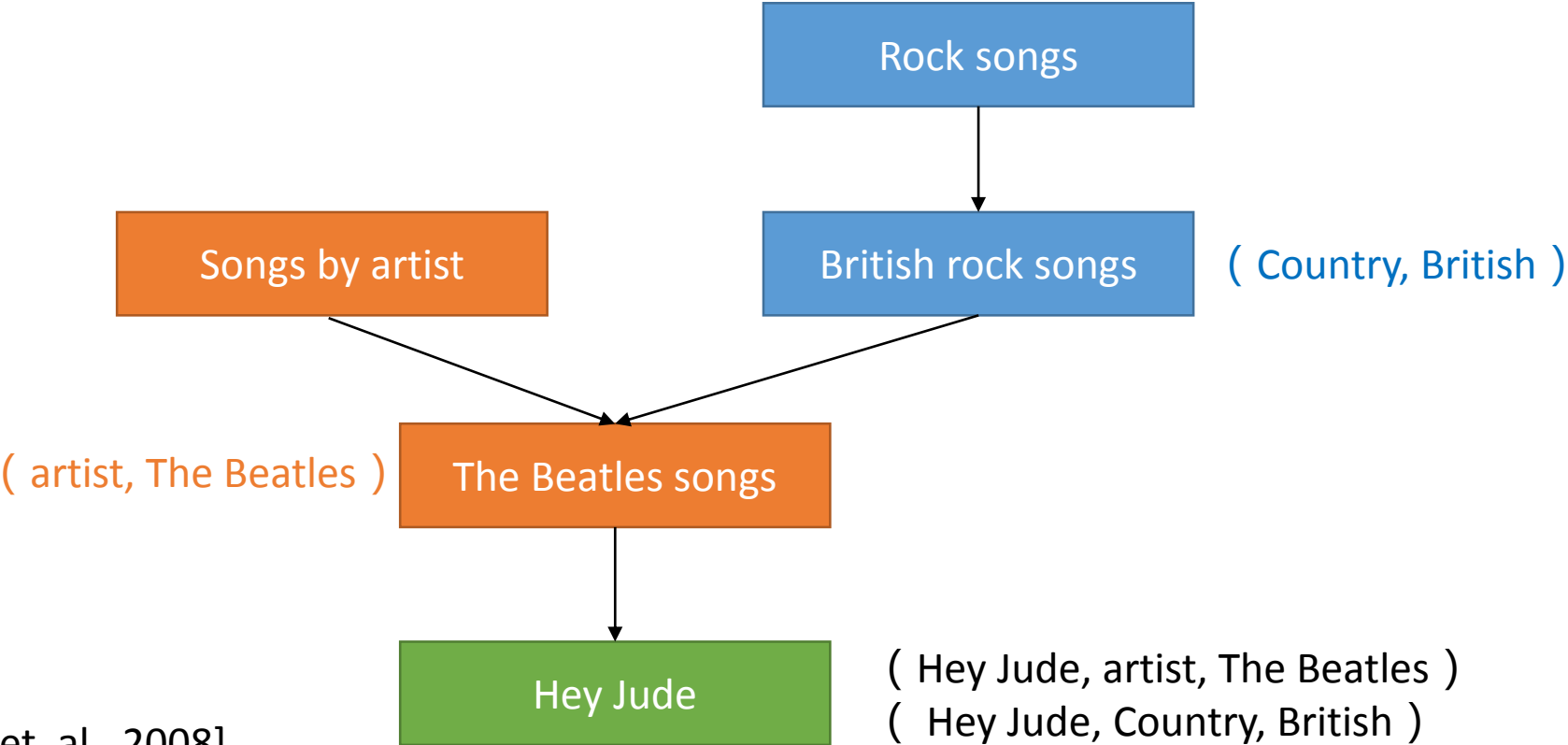
Regular Expression	Relation
<code>([0-9]{3,4}) births</code>	BORNONDATE
<code>([0-9]{3,4}) deaths</code>	DIEDONDATE
<code>([0-9]{3,4}) establishments</code>	ESTABLISHEDONDATE
<code>([0-9]{3,4}) books novels</code>	WRITTENONDATE
<code>Mountains Rivers in (.*)</code>	LOCATEDIN
<code>Presidents Governors of (.*)</code>	POLITICIANOF
<code>(.*) winners</code>	HASWONPRIZE
<code>[A-Za-z]+ (.*) winners</code>	HASWONPRIZE

[Fabian, M. S., et al. 2008]



# Catriple

- 利用Wikipedia的上下位概念来获取实体的infobox信息



[Qiaoling Liu, et. al., 2008]





# 四种有效的上下位概念模式

## Pattern 1: by-prep

- 上位概念：by + 属性
  - e.g., Songs by **theme**
- 下位概念：介词从句且包含属性值
  - e.g., Songs about **divorce**
- 抽取方法
  - 从上位概念抽取属性
  - 从下位概念抽取属性值
  - ( **theme, divorce** )

## Pattern 2: by-noun

- 上位概念：by + 属性
  - e.g., Songs by **artist**
- 下位概念：名词从句且包含属性值
  - e.g., **The Beatles** songs
- 抽取方法
  - 从上位概念抽取属性
  - 从下位概念抽取属性值
  - ( **artist, The Beatles** )



# 四种有效的上下位概念模式

## Pattern 3: \*-prep except by-prep

- 上位概念：不包含属性
  - 上位概念举例：Songs
- 下位概念：介词从句且包含属性值
  - 下位概念举例：Songs from films
- 抽取方法
  - 从下位概念抽取属性值
    - ( ?, films )
  - 通过投票确定属性值对应的属性
    - ( genre, films )

## Pattern 4: \*-noun except by-noun

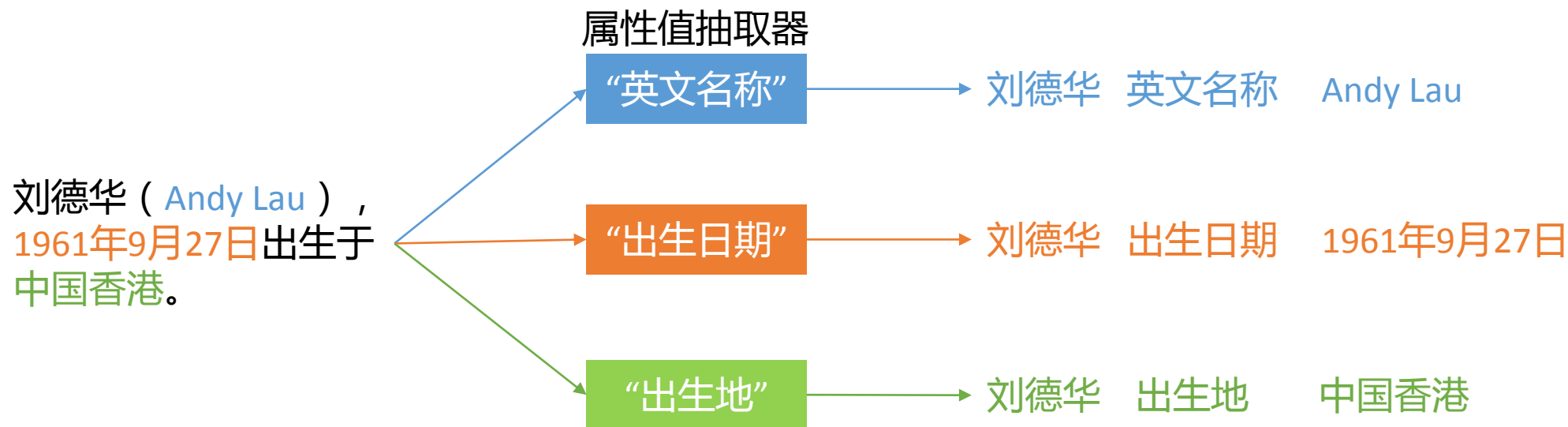
- 上位概念：不包含属性
  - 上位概念：Rock songs
- 下位概念：名词从句且包含属性值
  - 下位概念：British rock songs
- 抽取方法
  - 从下位概念抽取属性值
    - ( ?, British )
  - 通过投票确定属性值对应的属性
    - ( Country, British )



# 利用实体文本内容进行填充

## • 基本思路

- 为每个属性构建一个抽取器（分类器）
- 每个抽取器分别从百科文本（实体名已知）的句子中抽取相应属性的值





# 序列数据标记问题

- 文本属性值抽取被认为是一个序列数据标记问题
  - 将句子当做是一个序列数据
  - 属性值抽取过程即可看作是序列数据标记过程
    - 1表示为属性值
    - 0表示不是属性值

“英文名称”

刘德华 | ( | Andy | Lau | ) | , | 1961年 | 9月 | 27日 | 出生 | 于 | 中国 | 香港 | 。 |  
0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

“出生日期”

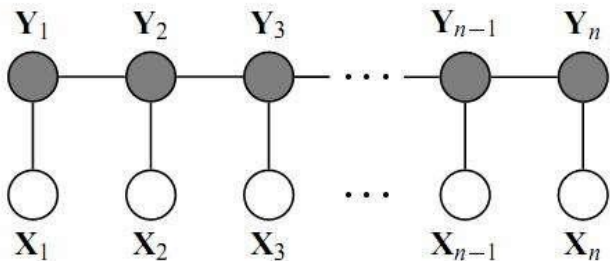
刘德华 | ( | Andy | Lau | ) | , | 1961年 | 9月 | 27日 | 出生 | 于 | 中国 | 香港 | 。 |  
0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |

“出生地”

刘德华 | ( | Andy | Lau | ) | , | 1961年 | 9月 | 27日 | 出生 | 于 | 中国 | 香港 | 。 |  
0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |

# 传统方法

- 条件随机场 ( CRF )
  - 针对序列数据进行分类的模型
  - 每个词组需要人为设定一组特征

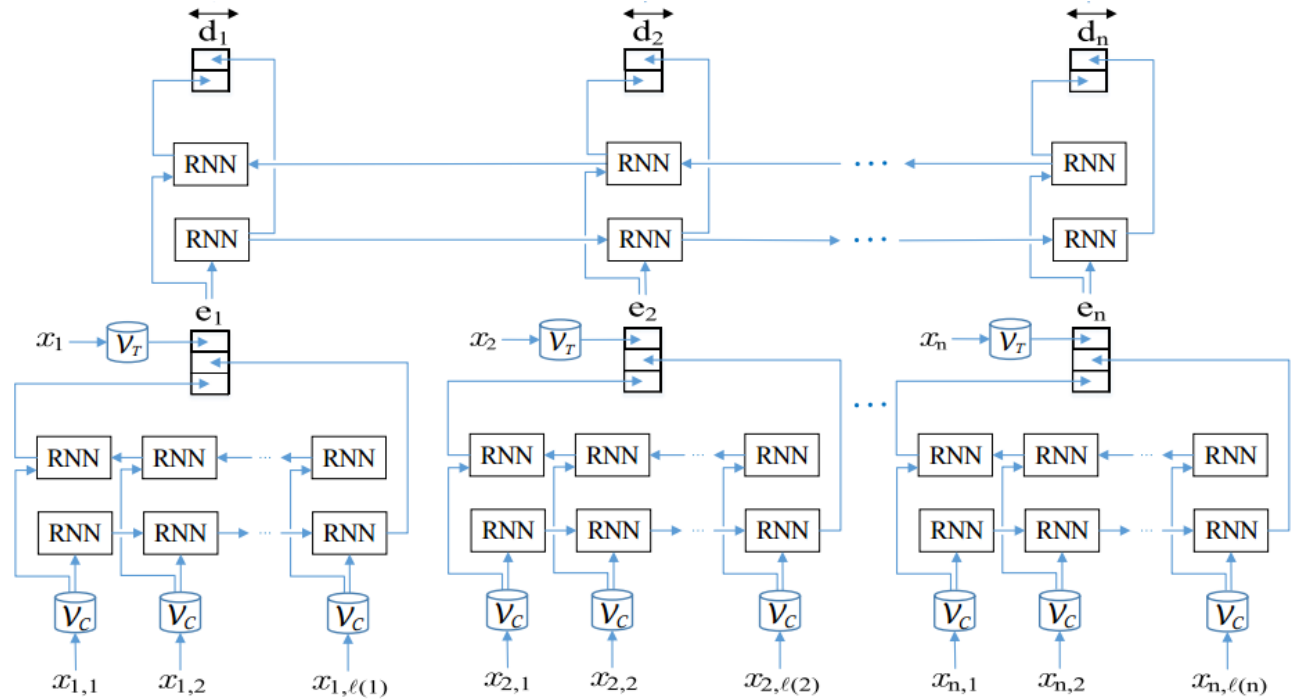


- 缺点
  - 需要专家人为设计特征
  - 不具有通用性

Feature Description	Example
First token of sentence	<i>Hello world</i>
In first half of sentence	<i>Hello world</i>
In second half of sentence	<i>Hello world</i>
Start with capital	Hawaii
Start with capital, end with period	Mr.
Single capital	A
All capital, end with period	CORP.
Contains at least one digit	AB3
Made up of two digits	99
Made up of four digits	1999
Contains a dollar sign	20\$
Contains an underline symbol	km_square
Contains an percentage symbol	20%
Stop word	the; a; of
Purely numeric	1929
Number type	1932; 1,234; 5.6
Part of Speech tag	
Token itself	
NP chunking tag	
String normalization: capital to "A", lowercase to "a", digit to "1", others to "0"	<i>TF - 1</i> $\implies$ <i>AA01</i>
Part of anchor text	<u>Machine Learning</u>
Beginning of anchor text	<u>Machine Learning</u>
Previous tokens (window size 5)	
Following tokens (window size 5)	
Previous token anchored	<u>Machine Learning</u>
Next token anchored	<u>Machine Learning</u>

# 基于深度学习的方法

- 优点
  - 不需要人工设计特征
- 方法
  - LSTM



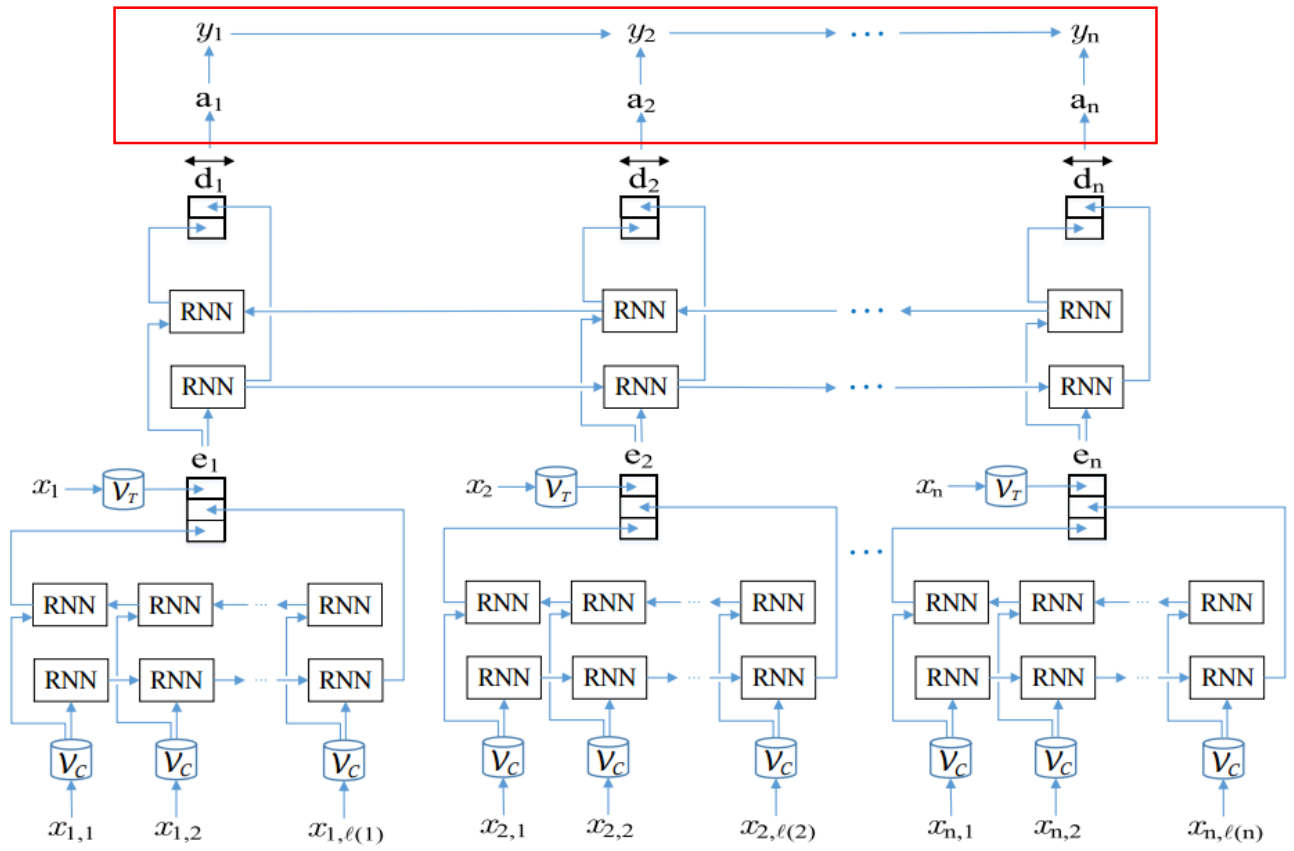
Architecture of the artificial neural network (ANN) model. RNN stands for recurrent neural network. The type of RNN used in this model is Long Short Term Memory (LSTM).  $n$  is the number of tokens, and  $x_i$  is the  $i^{th}$  token.  $V_T$  is the mapping from tokens to token embeddings.  $\ell(i)$  is the number of characters and  $x_{i,j}$  is the  $j^{th}$  character in the  $i^{th}$  token.  $V_C$  is the mapping from characters to character embeddings.  $e_i$  is the character-enhanced token embeddings of the  $i^{th}$  token.  $\vec{d}_i$  is the output of the LSTM of label prediction layer,  $\mathbf{a}_i$  is the probability vector over labels,  $y_i$  is the predicted label of the  $i^{th}$  token.

[Dernoncourt, F. et. al. 2017]

# 基于深度学习的方法

- 优点
  - 不需要人工设计特征
- 方法
  - LSTM
  - LSTM+CRF

$$s(y_{1:n}) = \sum_{i=1}^n \mathbf{a}_i[y_i] + \sum_{i=2}^n T[y_{i-1}, y_i].$$



Architecture of the artificial neural network (ANN) model. RNN stands for recurrent neural network. The type of RNN used in this model is Long Short Term Memory (LSTM).  $n$  is the number of tokens, and  $x_i$  is the  $i^{th}$  token.  $V_T$  is the mapping from tokens to token embeddings.  $\ell(i)$  is the number of characters and  $x_{i,j}$  is the  $j^{th}$  character in the  $i^{th}$  token.  $V_C$  is the mapping from characters to character embeddings.  $e_i$  is the character-enhanced token embeddings of the  $i^{th}$  token.  $d_i$  is the output of the LSTM of label prediction layer,  $a_i$  is the probability vector over labels,  $y_i$  is the predicted label of the  $i^{th}$  token.

[Dernoncourt, F. et. al. 2017]

## 2.2 实体分类





# 实体分类 ( Entity Typing )

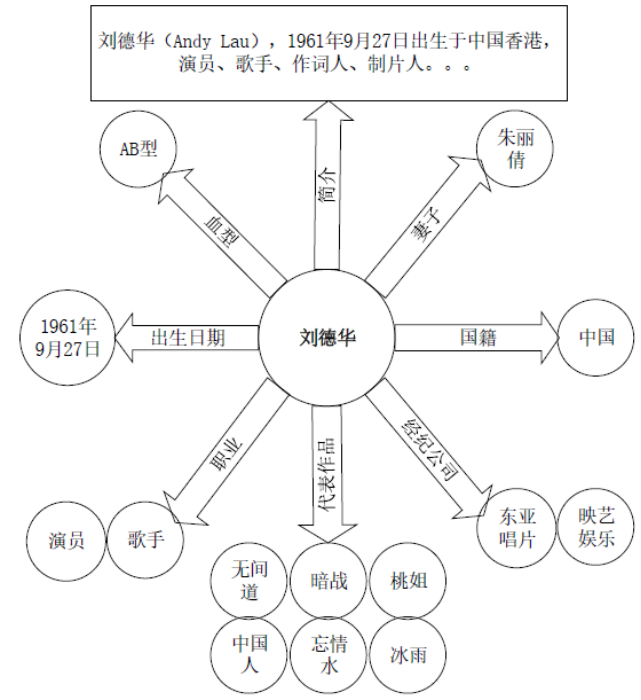
- 定义
  - 将知识图谱中的实体分类到一组预定义的概念集合中
- 和命名实体识别的区别

比较项	命名实体识别	面向知识图谱的实体分类
对象	实体指称项 (entity mention)	实体 (entity)
输入	实体指称项所在的句子	实体在知识图谱中的知识
特征	文本特征	文本特征, 语义关系特征
输出	是现实世界中实体的分类结果的一部分	与真实世界中实体的分类结果理论上一致

刘德华出生于1961年9月

刘德华出演了最新电影《长城》

《忘情水》是刘德华的代表歌曲

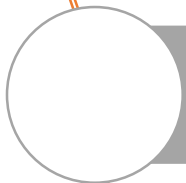




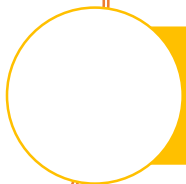
# 方法总结



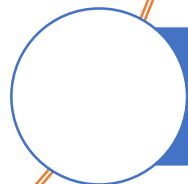
人工方法



自动规则



有监督




远程监督

# 人工方法



**Ford GT40**



**Overview**

**Manufacturer** Ford Advanced Vehicles  
John Wyer Automotive Engineering  
Kar Kraft  
Shelby American

**Production** 1964-1969<sup>[1]</sup>  
105 produced<sup>[2]</sup>

**Assembly** Slough, UK (Mk I, Mk II, and Mk III)  
Wixom, Michigan, United States (Mk IV)

通过人工方法建立infobox模板名称和概念的等价关系

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

## Ontology Classes

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

[Jens Lehmann et al. 2015]



# 自动规则

- 基于等价概念的实体分类

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

- 基于等价实体的实体分类

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

- 基于继承关系的实体分类

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



# 自动规则

## • 概念筛选

- Wikipedia标签体系中包含四类标签
  - **conceptual categories**
    - E.g., **Jay Chou Albums**
  - administrative purposes
  - relational information
    - E.g., **1879 births**
  - thematic vicinity
    - E.g., **Physics**

## • 识别Conceptual Categories

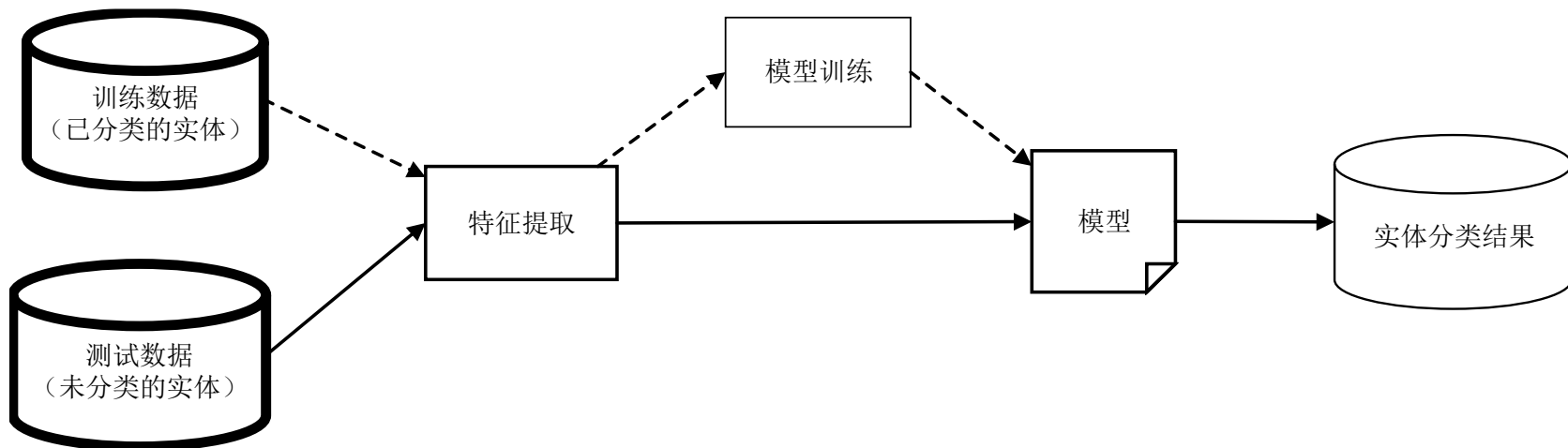
- shallow linguistic parsing
  - if the **head** of the category name is a **plural word**, the category is most likely a **conceptual category**

[Fabian M. S. et al. 2007]

# 有监督实体分类

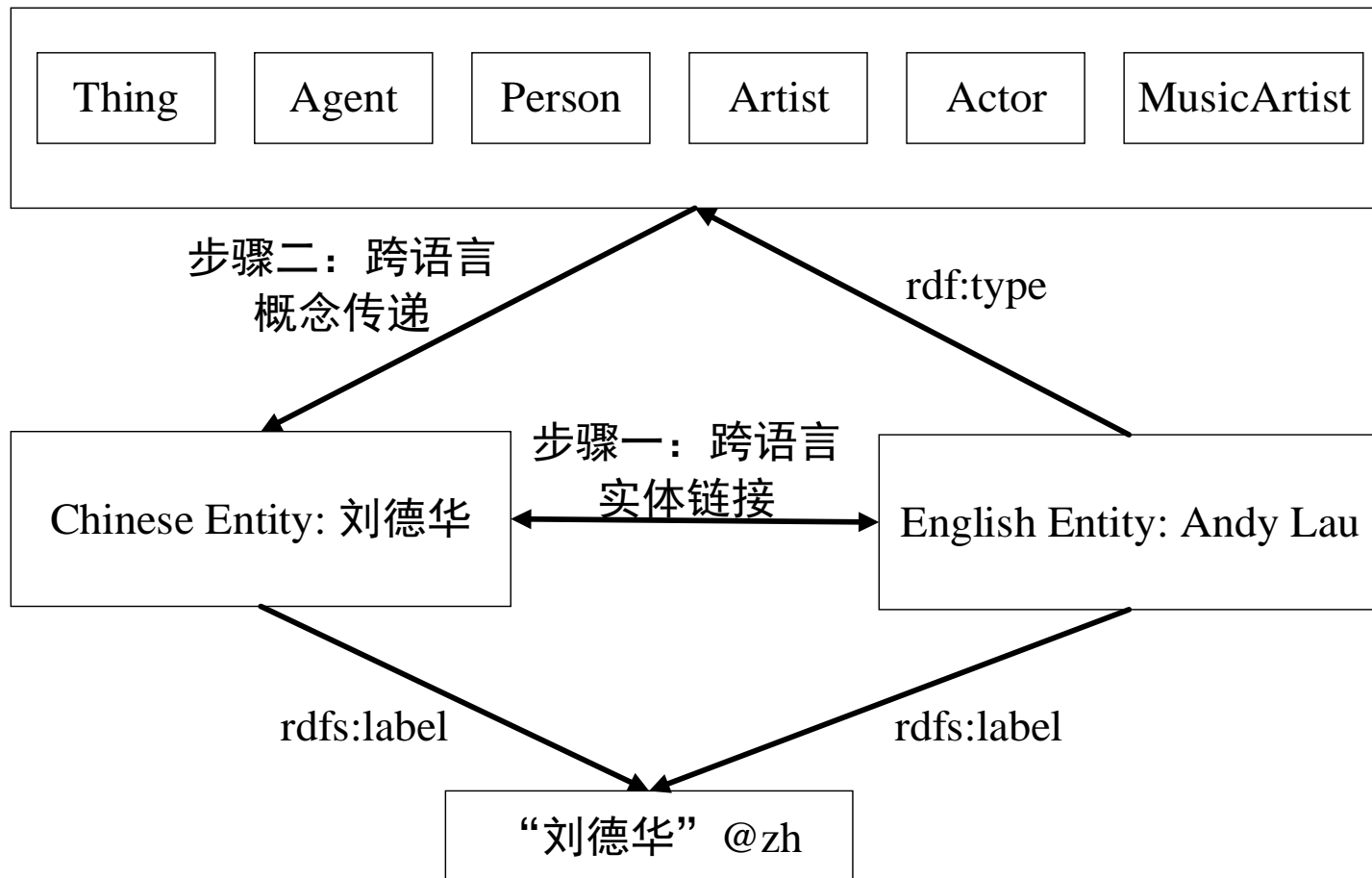
## • 分类模型

- SVM
- 逻辑回归
- 决策树
- ...





# 弱监督/远程监督





# 远程监督的问题

- 远程监督构建的训练集存在噪声问题
  - 目标知识图谱本身存在噪声
  - 实体链接错误
  - 实体特征缺失





# 解决方法

## • 多分类器投票过滤

- 将训练集分为N份，其中每N-1份作为训练集，用来过滤剩下一份的噪声
- 每个分类器分别对实体进行重新预测，与原结果比较，未预测出的结果即视为该分类器发现的噪声数据
- 综合多个分类器的噪声数据，通过过滤策略对训练集进行过滤
  - 大多数投票过滤
  - 一致性过滤

表 3.2: 一个实体在训练集中的概念集合为 {A, B, C, D}, 通过不同分类器识别出不同的噪声集合

分类器	预测概念集合	噪声概念集合
1	{A, B, C}	{D}
2	{A, B}	{C, D}
3	{A, B}	{C, D}

表 3.3: 使用不同的策略对表 3.2中实体的概念集合进行过滤

过滤策略	最终噪声集合	过滤后的概念集合
大多数投票过滤	{C, D}	{A, B}
一致性过滤	{D}	{A, B, C}

分类器的要求：

- (1) 每个分类器的准确率要高于50%
- (2) 每个分类器产生的错误是独立的

# CN-DBpedia中的实体分类



## • 问题

- 将百度百科中的实体分类到DBpedia的概念集合中

## • 特征

### • 语义关系特征

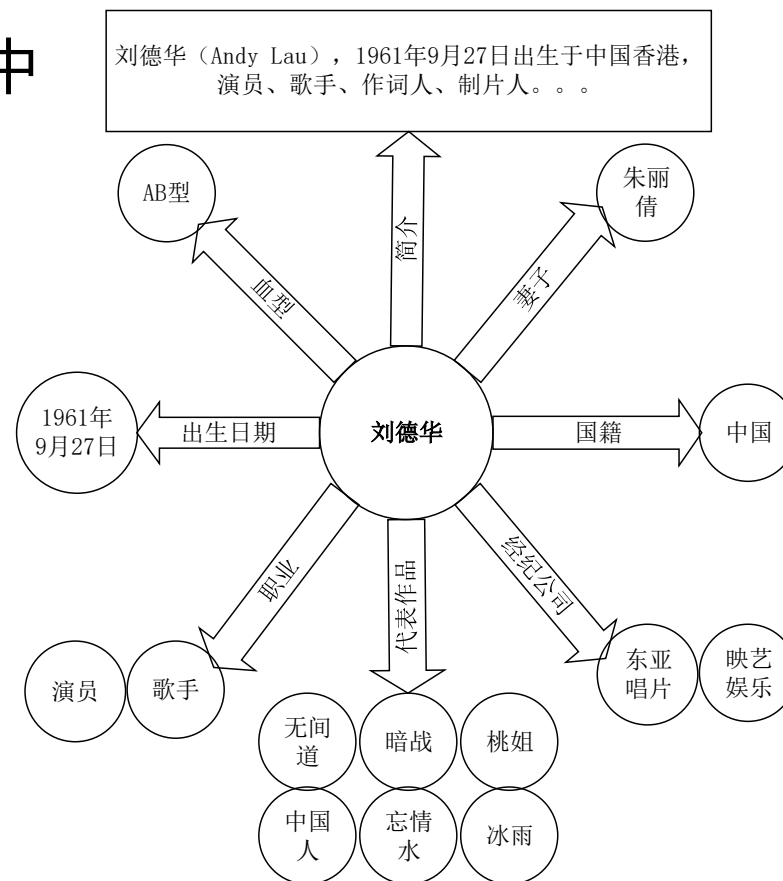
- 属性
- 属性-值
- 标签

### • 文本特征

- 来自于摘要及正文信息

## • 方法

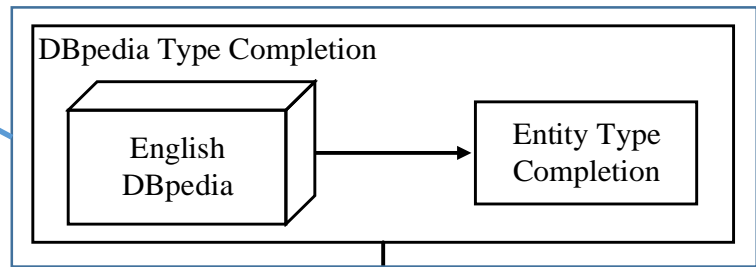
- 利用语义关系特征进行分类
- 利用文本特征进行分类



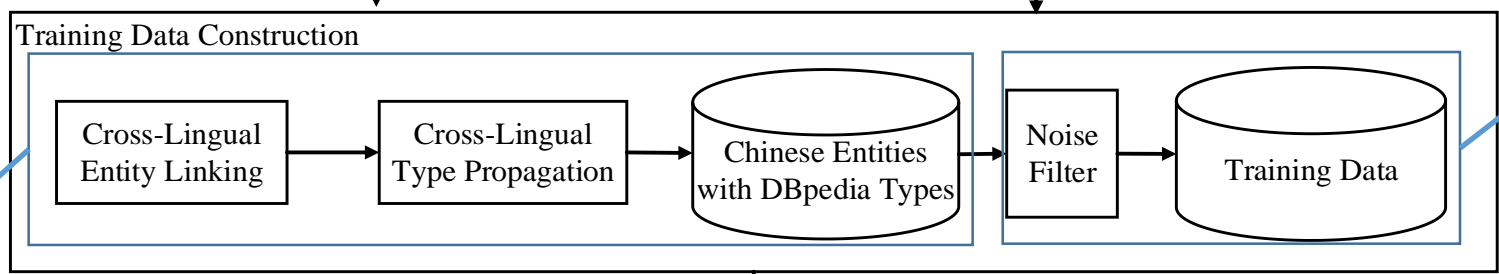
# 利用语义关系特征进行分类



解决英文知识图谱  
实体类别不完整的问题

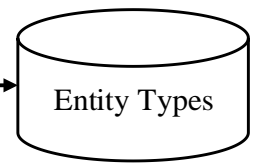
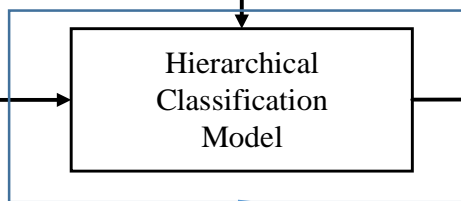
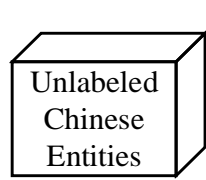


提高训练集质量



[Wang, Z. et al., 2012]  
[Wang, Z. et al., 2013]

解决中文实体  
无英文概念标记问题

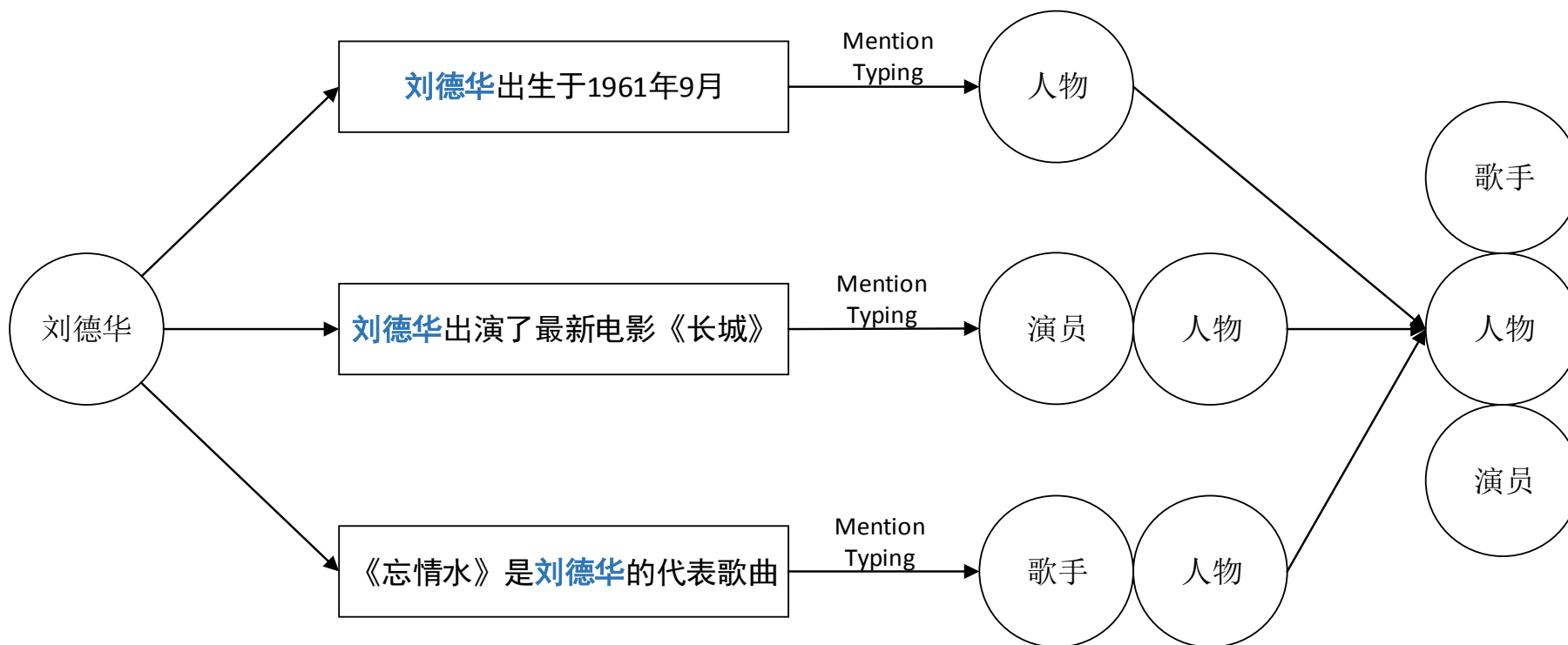


考虑概念层次结构

[Bo Xu et al., 2017]

# 基于文本特征的实体分类

- 基本思路
  - Mention Typing + Type Fusion



# 基于文本的实体分类

- 难点1：训练集构建
  - 人工标记代价大

- 解决方案

- STEP 1：基于远程监督的训练集构建
- STEP 2：训练集噪声过滤
  - 多分类器投票过滤方法

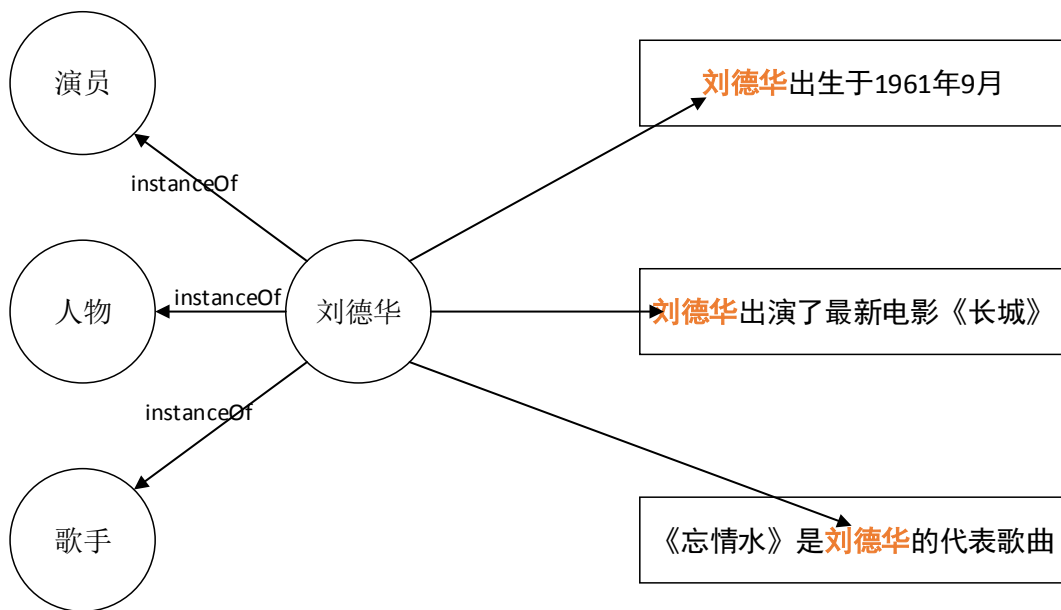


表 4.2: 训练集过滤前后效果。

ID	包含实体指称项的句子	过滤前概念集合	过滤后概念集合
1	刘德华出生于 1961 年 9 月	{人物、演员、歌手}	{人物}
2	刘德华出演了最新电影《长城》	{人物、演员、歌手}	{人物、演员}
3	《忘情水》是刘德华的代表歌曲	{人物、演员、歌手}	{人物、歌手}



# 基于文本的实体分类

- 难点2：特征选择
  - 人工设计代价大

## • 解决方案

- 基于神经网络的实体指称项分类
- 一个句子分为三部分

- Left Context
- Mention [Li Dong et al., 2015]
- Right Context

## • 对句子进行向量化处理

$$[c_{-S}, \dots, c_{-1}] [m_1, \dots, m_n] [c_1, \dots, c_S]$$

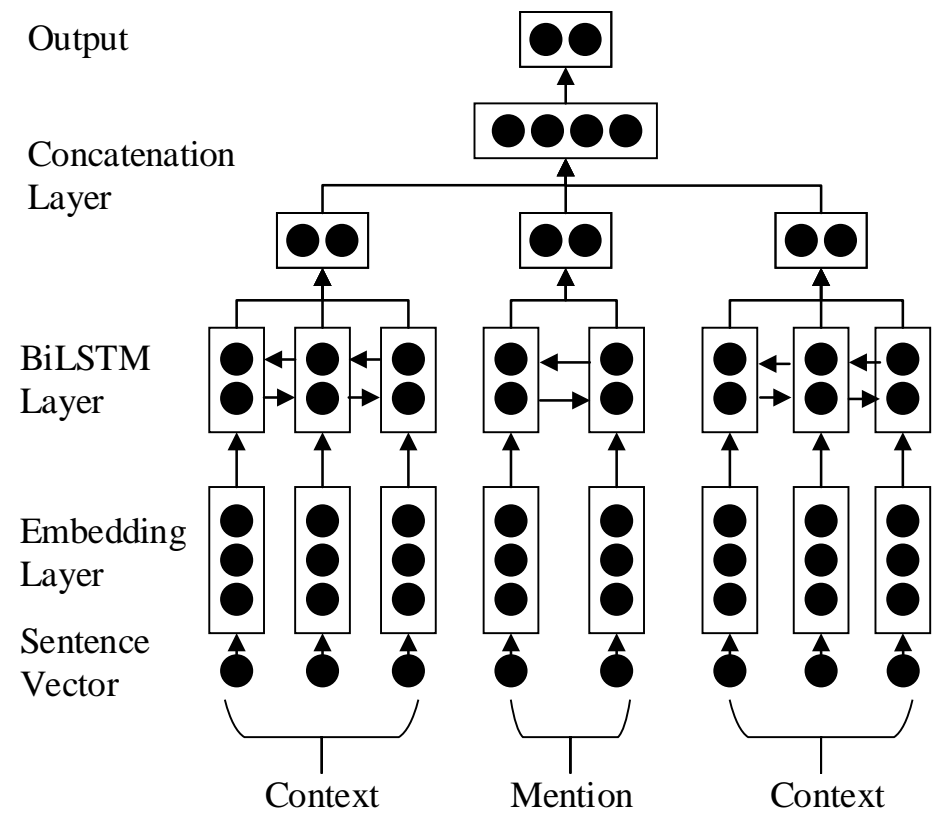


表 4.4: TEX 系统中，中文句子的向量化表示形式

Format	$[c_{-S} \dots c_{-1}] [m_1 \dots m_N] [c_1 \dots c_S]$ , $S = 5$ and $N = 5$
Sentence	皇家马德里的明星克里斯蒂亚诺·罗纳尔多 在星期天和他的家人庆祝他的第 32 个生日
Segmentation	皇家, 马德里, 的, 明星, 克里斯蒂亚诺·罗纳尔多, 在, 星期天, 和, 他, 的, 家人, 庆祝, 他, 的, 第 32, 个, 生日
Partition	[Null, 皇家, 马德里, 的, 明星] [克里斯蒂亚诺·罗纳尔多, Null, Null, Null, Null] [在, 星期天, 和, 他, 的]
Vector	[0 334 346 75545 8456] [2478 0 0 0 0] [678 883 2793 67094 24679]



# 基于文本的实体分类

- 难点3：结果融合
  - 简单的合并算法无法取得良好的效果

表 4.9: 不同融合策略对实体分类效果的影响

Strategy	pE	rE	fE
Consider all possibilities	0.79	0.93	0.85
No Gossiping	0.98	0.46	0.63
Majority Voting	0.98	0.77	0.86
TEX-TF-Disjointness	0.90	0.92	0.91
TEX-TF-Hierarchy	0.81	0.93	0.87
TEX-TF-ALL	0.93	0.92	0.92

[Dong, Xin Luna et al., 2009]

- 解决方案
  - 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

Maximize

$$\sum_{c \in C} \sum_{s \in S} w(c|e, s) \times x_{e,c} \quad x_{e,c} = \begin{cases} 1 & \text{if entity } e \text{ belongs to type } c \\ 0 & \text{else} \end{cases}$$

Subject to

$$\begin{aligned} \forall_{ME(c_1, c_2)} \quad x_{e,c_1} + x_{e,c_2} &\leq 1 \\ \forall_{ISA(c_1, c_2)} \quad x_{e,c_1} - x_{e,c_2} &\leq 0 \end{aligned} \quad w(c|e, s) = \begin{cases} P(c|e, s) & \text{if } P(c|e, s) > \theta \\ 0 & \text{else} \end{cases}$$

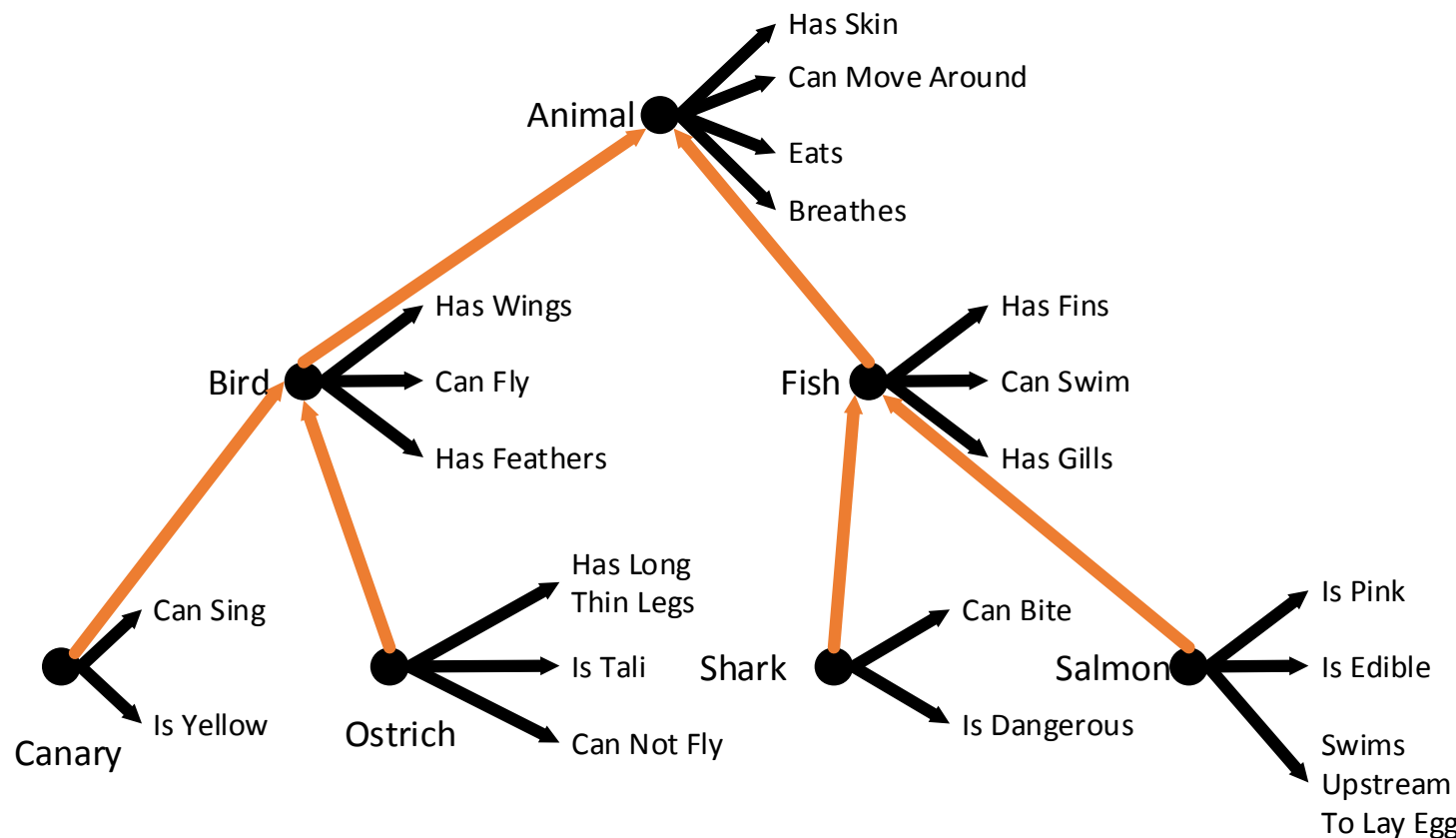
## 2.3 概念表示





# 概念的形成

- 经典范畴理论认为，概念是由一组包含相同特征集合的实体组成，这组特征集合称为**概念的固有特征集合**





# 概念的固有特征集合性质

## • 性质一

- 如果一个实体包括某个概念的固有特征集合，那么它一定属于这个概念

## • 应用

- Entity Typing

$$\vec{f}_c \subseteq \vec{f}_e \Rightarrow e \in c$$

## • 性质二：

- 如果一个实体属于某个概念，那么它也一定包含这个概念的固有特征集合

## • 应用

- Infobox Completion

$$e \in c \Rightarrow \vec{f}_c \subseteq \vec{f}_e$$

# 研究问题



- 人工方法无法解决规模巨大的概念集合
  - 在很早之前，心理学家通过人工的方法为少量的粗粒度的概念定义了它们的固有特征集合，如鸟类、动物、汽车等
  - 然而，这种方法仅适合概念数量较少的分类当概念数量上万时，人工方法已经不再适用了
- 因此，我们研究如何通过**自动**的方法，利用知识图谱中的**已有知识**来表示概念的固有特征集合

# 问题一：哪些概念适合用知识图谱中的关系来表示？



- 粗粒度概念 or 细粒度概念？
  - 粗粒度概念代表
    - WordNet , DBpedia Ontology
    - E.g.,
      - “花”, “鸟”, “鱼”, “虫”
  - 细粒度概念代表
    - Wikipedia标签系统, 百度百科标签系统
    - E.g.,
      - “周杰伦歌曲”, “香港男演员”

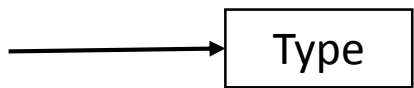
细粒度概念☑

# 问题二：知识图谱中的哪些知识能够用来表示概念？



## • Bird

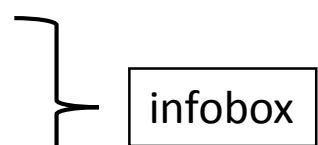
• Animal



• Has Wings

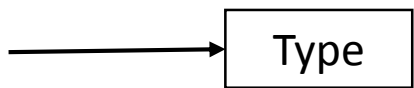
• Can Fly

• Has Feathers



## • Fish

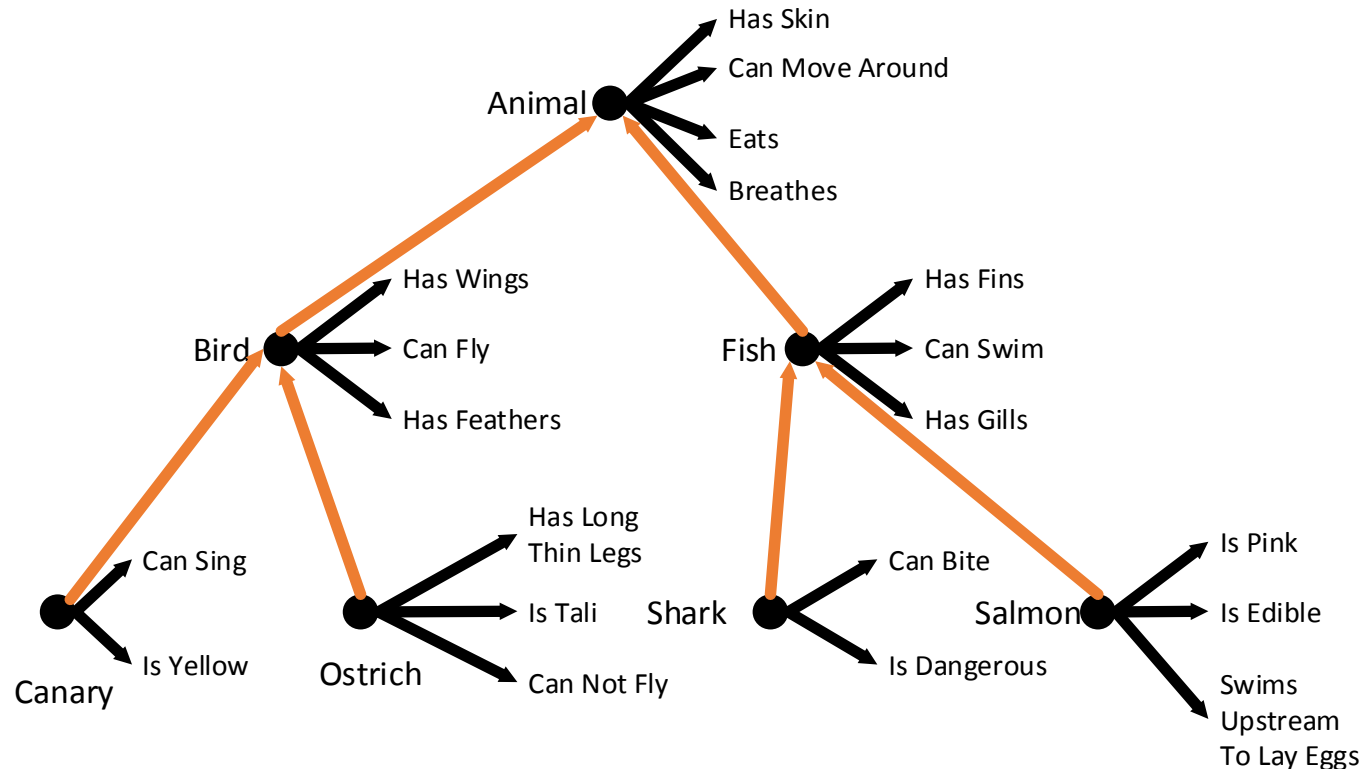
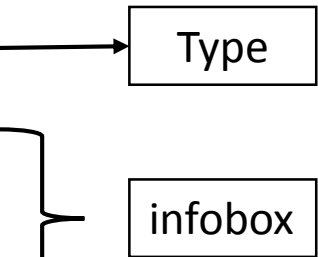
• Animal



• Has Fins

• Can Swim

• Has Gills



利用知识图谱中实体的infobox和Type信息来表示实体的细粒度概念（标签）

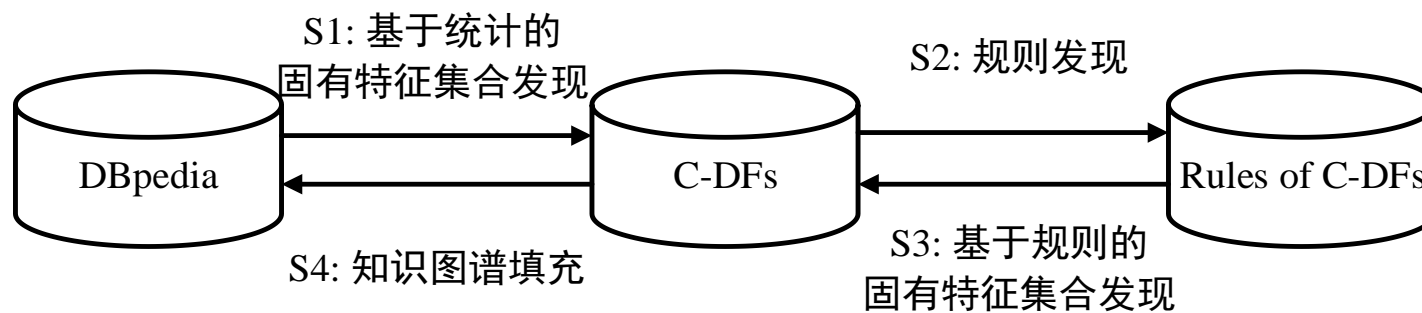
# 举例：知识图谱中细粒度概念表示



细粒度概念	Infobox	Type
Films directed by Christopher Nolan	(director, Christopher Nolan)	(Type, Film)
Jay Chou albums	(artist, Jay Chou)	(Type, Album)
American screenwriters	(birthPlace, United States) (occupation, Screenwriter)	(Type, Person)
American crime films	(country, United States) (genre, crime)	(Type, Film)

概念的固有特征集合是唯一的，集合中的特征不能多一个，也不能少一个

# 框架





# 基于统计的固有特征集合发现

## • 基本思路

- 每个属于概念的实体都包含该概念的固有特征集合

$$e \in c \iff \vec{f}_c \subseteq \vec{f}_e$$

- 然而，一个实体不仅包含概念的固有特征集合，也包含更多的非固有特征集合的特征

Category: Films directed by Christopher Nolan	
Entity: Inception (盗梦空间)	
Infobox	Type
(name, Inception)	(Type, Thing)
(director, Christopher Nolan)	(Type, Work)
(producer, Emma Thomas)	(Type, Film)
(producer, Christopher Nolan)	
(writer, Christopher Nolan)	
(starring, Leonardo DiCaprio)	
(starring, Ken Watanabe)	
(starring, Joseph Gordon-Levitt)	
(starring, Ellen Page)	
(starring, Tom Hardy)	
(music, Hans Zimmer)	
(cinematography, Wally Pfister)	
... ..	





# Naïve方法

- 穷举概念 $c$ 的所有可能的特征集合 $\vec{f}$ ，并计算每个集合的得分
- 打分函数
  - 衡量一组特征集合 $\vec{f} = (f_1, f_2, \dots, f_k)$ 是某个概念 $c$ 的固有特征集合的概率

$$score(c, \vec{f}) = P(\vec{f}|c) \times P(c|\vec{f})$$

$$P(\vec{f}|c) = \frac{\# \text{ of entities in } c \text{ contain } \vec{f}}{\# \text{ of entities in } c}$$

$$P(c|\vec{f}) = \frac{\# \text{ of entities in } c \text{ contain } \vec{f}}{\# \text{ of entities contain } \vec{f}}$$

$$e \in c \Rightarrow \vec{f}_c \subseteq \vec{f}_e$$

$$\vec{f}_c \subseteq \vec{f}_e \Rightarrow e \in c$$

- 得分为**1**，表示特征集合 $f$ 为概念 $c$ 的**固有**特征集合

Jay Chou albums	(artist, Jay Chou)	(Type, Album)
-----------------	--------------------	---------------



# Naïve 方法的缺陷

- 缺陷一
  - 由于知识图谱的不完整性，很多概念的固有特征集合的得分都小于1
- 缺陷二
  - 穷举所有特征集合的代价太大



# 解决方案

- 针对缺陷一，我们降低了  $score(c, \vec{f})$  的取值

$$\vec{f}_c = \arg \max_{\vec{f} \subseteq F(c)} score(c, \vec{f}) \quad score(c, \vec{f}) > \alpha$$

- 针对缺陷二，我们利用频繁项集挖掘方法进行剪枝

- 全部项items
  - 所有特征
- 事务Transaction
  - 每个实体及其特征
- 项集itemset
  - 每个实体的特征集合

实体	特征	Naïve 个数
$e_1$	$\{f_1, f_2, f_3, f_4\}$	11
$e_2$	$\{f_1, f_2, f_5\}$	7
$e_3$	$\{f_1, f_2, f_6, f_7\}$	11
$e_4$	$\{f_1, f_2, f_8, f_9\}$	11
$e_5$	$\{f_3, f_{10}\}$	3
频繁项集 ( $\beta > 0.4$ ) : $\{f_1\}, \{f_2\}, \{f_3\}, \{f_1, f_2\}$		

当  $\alpha = \beta$  时，能保证非频繁特征集合一定不是概念的固有特征集合



# 基于统计方法的缺陷

- 由于知识图谱的不完整性，很多概念的固有特征集合可能是不频繁的，导致其无法通过基于统计的方法得到结果
- 因此，我们又提出了基于规则的方法

# 正则表达式？



C	DFs
Films directed by <b>Christopher Nolan</b>	{{(Type, Film), (director, <b>Christopher Nolan</b> )}}
Films directed by <b>James Cameron</b>	{{(Type, Film), (director, <b>James Cameron</b> )}}
Films directed by <b>Steven Spielberg</b>	{{(Type, Film), (director, <b>Steven Spielberg</b> )}}
Films directed by <b>David Fincher</b>	{{(Type, Film), (director, <b>David Fincher</b> )}}
Films directed by <b>Ben Affleck</b>	{{(Type, Film), (director, <b>Ben Affleck</b> )}}

C Pattern	DFs Pattern
Films directed by <b>(.*)</b>	{{(Type, Film), (director, <b>(.*)</b> )}}



# 正则表达式 ? NO!

C	DFs
Jay Chou albums	(Type, Album) (artist, Jay Chou)
Justin Bieber albums	(Type, Album) (artist, Justin Bieber)
Lady Gaga albums	(Type, Album) (artist, Lady Gaga)
Sony Music Taiwan albums	(Type, Album) (recordLabel, Sony Music Taiwan)
Cherrytree Records albums	(Type, Album) (recordLabel, Cherrytree Records)
Def Jam Recordings albums	(Type, Album) (recordLabel, Def Jam Recordings)

C Pattern	DFs Pattern
(.*) albums	(Type, Album) (artist, (.*))
(.*) albums	(Type, Album) (recordLabel, (.*))

# 规则



C Pattern	DFs Pattern
(.*) albums	(Type, Album) (artist, (.*))
(.*) albums	(Type, Album) (recordLabel, (.*))

任意字符串 (.\* ) 替换为  
特定类型的实体<Agent>,<RecordLabel>

属性名	Domain	Range
artist	MusicalWork	Agent
recordLabel	Thing	RecordLabel

C Pattern	DFs Pattern
<Agent> albums	(Type, Album) (artist, <Agent>)
<RecordLabel> albums	(Type, Album) (recordLabel, <RecordLabel>)

# 规则评估



- 好的规则

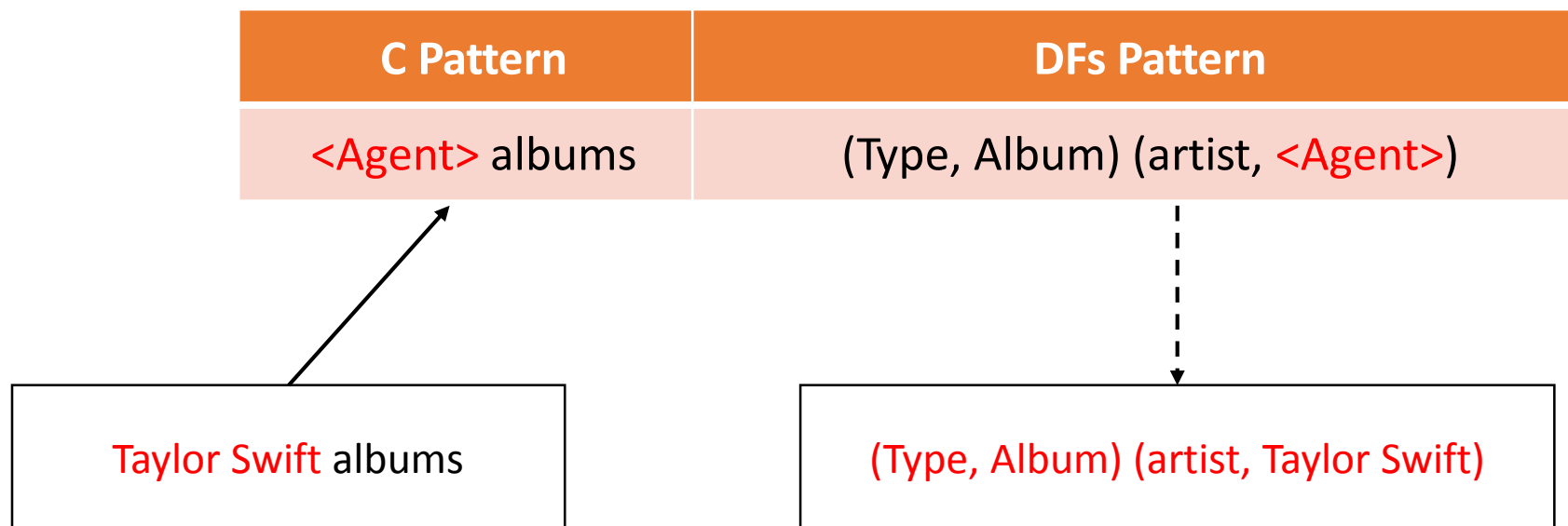
- $supportCount(r) > \gamma$
- $confidence(r) > \delta$

$SupportCount(r) = \# \text{ of CDFs matching } r_l \text{ and } r_r$

$$Confidence(r) = \frac{SupportCount(r)}{\# \text{ of CDFs matching } r_l}$$



# 基于规则的固有特征集合发现

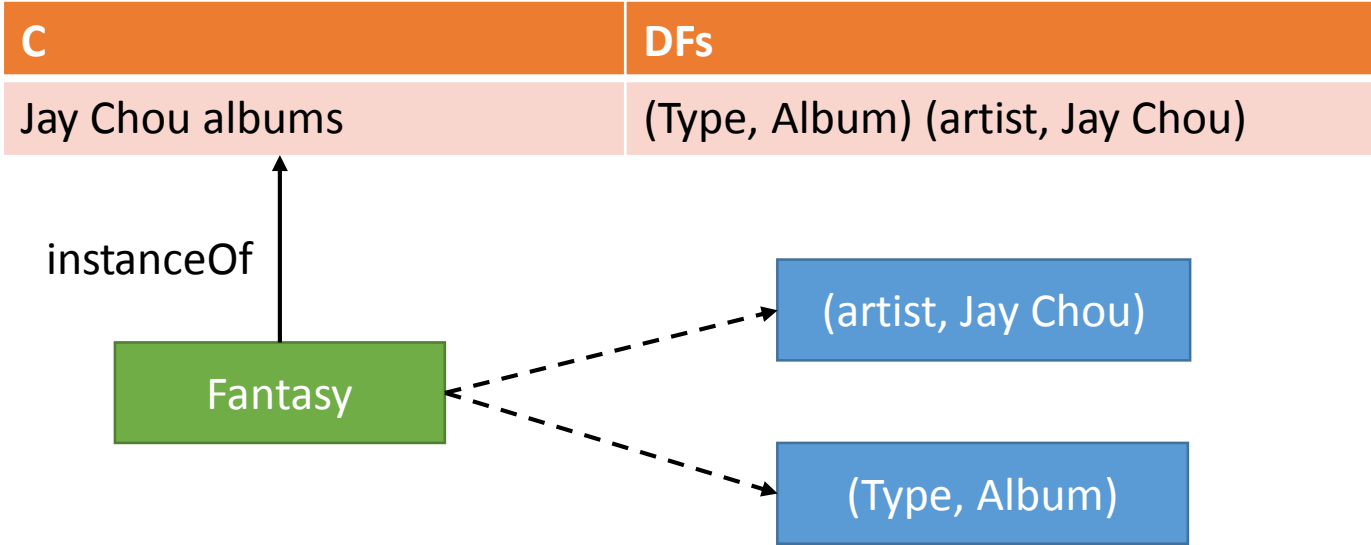




# 知识图谱填充

- 实体infobox填充
- 实体type填充

$$e \in c \Rightarrow \vec{f}_c \subseteq \vec{f}_e$$

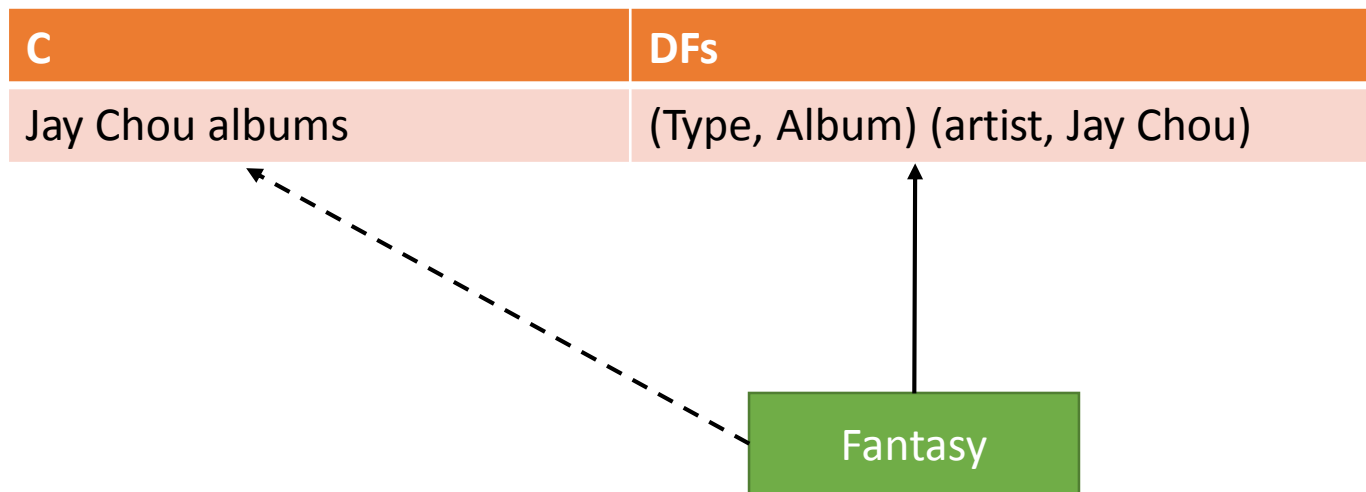




# 知识图谱填充

- 实体细粒度概念填充

$$\vec{f}_c \subseteq \vec{f}_e \Rightarrow e \in c$$



# 第三部分：知识更新





# 传统更新方法

- 基于更新日志的更新
- Wikipedia有这个功能，但百度百科没有
- 周期性更新
  - E.g., 每半年重新爬取一遍数据并进行解析
- 依然无法保证周期内数据的时效性



# 基于语义搜索引擎的更新

- 基于用户反馈的更新
  - 用户点击更新按钮，进行更新

---

e.g., 复旦大学、周杰伦

Query String: 复旦大学

点击更新页面

Named-Entity Disambiguation: 复旦大学

---

- 基于搜索日志的新词发现
  - 用户搜索一个词时，未在知识库中找到，即认为是一个新词

---

e.g., 复旦大学、周杰伦

Query String: 顺丰菜鸟大战

Not Found in CN-DBpedia

---



# 主动更新方法

## • 基本思路

### • 监控互联网上的**热词**

#### • 热词分为两种情况

- 新词
- 旧词，但信息发生了变化

### • 更新热词以及与之**相关**的词条

## 热搜新闻词 HOT WORDS

从1到7“数”读 习近平扶贫方略		习近平将对哈萨克斯 坦进行国事访问		离岸人民币本周 涨484点	军报批动漫迷扮 军人下跪
陈刚任雄安新区 临时党委书记	高考期间全国大 部气温适宜	菜鸟顺丰恢复数 据传输	普京自曝如何 躲过5次暗杀	第三代社保卡 年内试点发放	北上广深二手房 价和租金齐跌

热搜词条 今天 | 昨天

<p><b>顺丰菜鸟大战</b> ↑</p> <p>国家邮政局宣布，菜鸟与顺丰同意从6月3日12时起，全面恢复业务合作和数据传输。</p>	<p><b>菲律宾恐怖袭击</b> ↑</p> <p>6月2日凌晨，一名蒙面者手持长枪闯进菲律宾首都马尼拉一酒店的赌场并开枪射击，已造成至少34人死亡。</p>
<p><b>李晨</b> ↑</p> <p>李晨在节目中自曝父母已离婚，自己还有个相差18岁的妹妹。</p>	<p><b>福特号航空母舰</b> ↓</p> <p>美国首艘“福特”级航母交付美国海军。</p>
<p><b>西班牙大厦</b> ↑</p>	<p><b>孙怀山</b> ↑</p>
<p><b>住房公积金</b> ↑</p>	<p><b>星耀五洲</b> ↑</p>

热点要闻

个性推荐

进入推荐版

- 国际社会高度关注习近平哈萨克斯坦之行  
引领上合发展 共建一带一路 砥砺奋进的五年
- 李克强出席第十九次中欧领导人会晤 专题  
访德4大成果 张德江 俞正声 张高丽
- 上海等10省份今日举行事业单位招聘考试  
总招录人数超4.5万 多地强调要严肃考试纪律
- 安理会通过决议：强烈谴责朝鲜核导活动 扩大制裁
- 媒体：美国退出气候协定，中国的机遇期来了？
- COSER穿中国军装向日式少女下跪 军媒：丢人且违法
- 内蒙古阿拉善盟阿拉善左旗附近发生5.0级左右地震
- 中国机动车近3亿辆 系PM2.5污染重要原因
- 误机掌握工作人员女博士：我知错了 能少关几天吗
- 境外消费超千元要“汇报” 微信支付宝不在范围内

# 为什么要做实体扩展更新？

- 原因：“牵一发而动全身”
- 例如：王宝强离婚事件
  - 热词：**王宝强**
    - 知识库中的婚姻关系进行了更新
  - 扩展实体：**马蓉**
    - 同样更新其婚姻关系







# 更新步骤

- 步骤一：从互联网上发现热词作为种子结点
- 步骤二：更新这些热词（从百科网站中获取新词或更新旧词）
- 步骤三：从这些更新的热词的页面中的超链接中获取更多的待更新实体，并为每个待更新实体设置更新优先级
  - 这是由于扩展会得到非常多的实体，超过每日的更新限制k
- 步骤四：按照优先级顺序更新扩展实体



# 优先级如何设置？

## • 原则

- 如果是一个新词，那么优先级设置为最高
- 如果是一个旧词，估计其上一次更新结束到当前时间内可能更新的次数，该次数作为优先级指标  $E[u(x)]$

$$E[u(x)] = P(x) \times (t_{now} - t_s(x))$$

- $P(x)$ ：为实体x预期的更新频率，通过预测器得到

## • 模型：回归

- 随机森林回归

## • 特征

#	Feature	$\chi^2$	IG( $10^{-3}$ )
1	#Weeks of existence	41.8	19.1
2	#Total updates	<b>481.1</b>	<b>55.9</b>
3	#Times viewed by users	203.5	46.2
4	#All hyperlinks	460.9	35.8
5	#Hyperlinks to entities	444.9	32.1
6	Page length	131.9	32.9
7	Main content length	202.1	19.1
8	Historical update frequency	287.6	54.7

# 总结



- 知识抽取
- 数据清洗

知识获取

- 属性-值填充
- 实体分类
- 概念表示

知识填充

- 反馈更新
- 主动更新

知识更新

Thank YOU !



- Our LAB: Knowledge Works at Fudan University
  - <http://kw.fudan.edu.cn>

# 参考文献



- [Jens Lehmann et al., 2015] DBpedia: A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia.
- [Haofen Wang et al., 2015] Effective Online Knowledge Graph Fusion
- [F. Deroncourt et al., 2017] De-identification of Patient Notes with Recurrent Neural Networks
- [Fabian, M. S. et al. 2007] Yago: A core of semantic knowledge unifying wordnet and wikipedia
- [Fabian, M. S. et al. 2008] YAGO: A Large Ontology from Wikipedia and WordNet
- [Bo Xu et al., 2016] Cross-lingual type inference
- [Brodley Carla E. et al. 1999] Identifying mislabeled training data
- [Allan M. Collins et al., 1969] Retrieval time from semantic memory
- [Jiaqing Liang et al. 2017] How to Keep a Knowledge Base Synchronized with Its Encyclopedia Source
- [Qiaoling Liu et al., 2008] Catriple: Extracting Triples from Wikipedia Categories

# 参考文献



- [Fei Wu et al., 2007] Autonomously semantifying wikipedia
- [Omer Levy and Yoav Goldberg, 2014] Dependency-Based Word Embeddings
- [Mikolov, Tomas, et al., 2013] Distributed representations of words and phrases and their compositionality
- [Pennington Jeffrey et al., 2014] Glove: Global vectors for word representation
- [Komninos and Manandhar, 2016] Dependency Based Embeddings for Sentence Classification Tasks
- [Bo Xu et al., 2017] CN-DBpedia: A Never-Ending Chinese Knowledge Extraction System
- [Dong, Xin Luna et al., 2009] Data fusion: resolving data conflicts for integration
- [Bouma, G. et al., 2009] Cross-lingual alignment and completion of Wikipedia templates
- [Wang, Z. et al., 2012] Cross-lingual Knowledge Linking Across Wiki Knowledge Bases.
- [Wang, Z. et al., 2013] Boosting cross-lingual knowledge linking via concept annotation
- [Fabian M. S. et al., 2011] Paris: Probabilistic alignment of relations, instances, and schema

# 参考文献



- [Li Dong et al., 2015] A hybrid neural model for type classification of entity mentions
- [Johannes Hoffart et. al., 2013] YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia
- [Roberto Navigli et. al., 2012] BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network
- [Bo Xu et al., 2016b] Learning Defining Features for Categories