

# 中文知识图谱CN-DBpedia 构建的关键技术

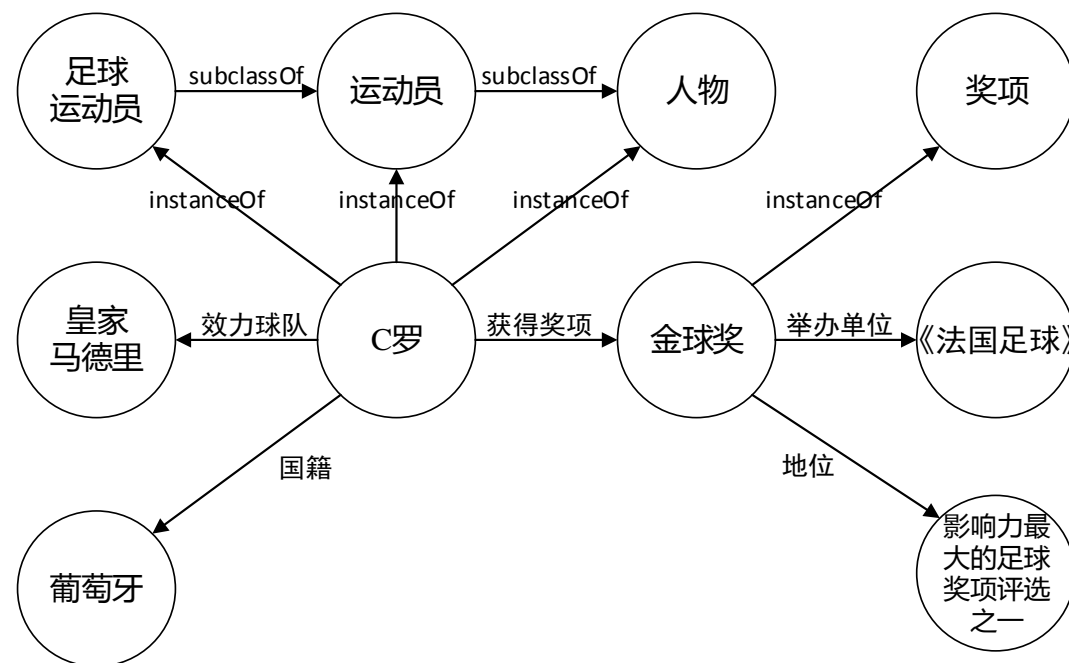
徐 波

复旦大学知识工场实验室

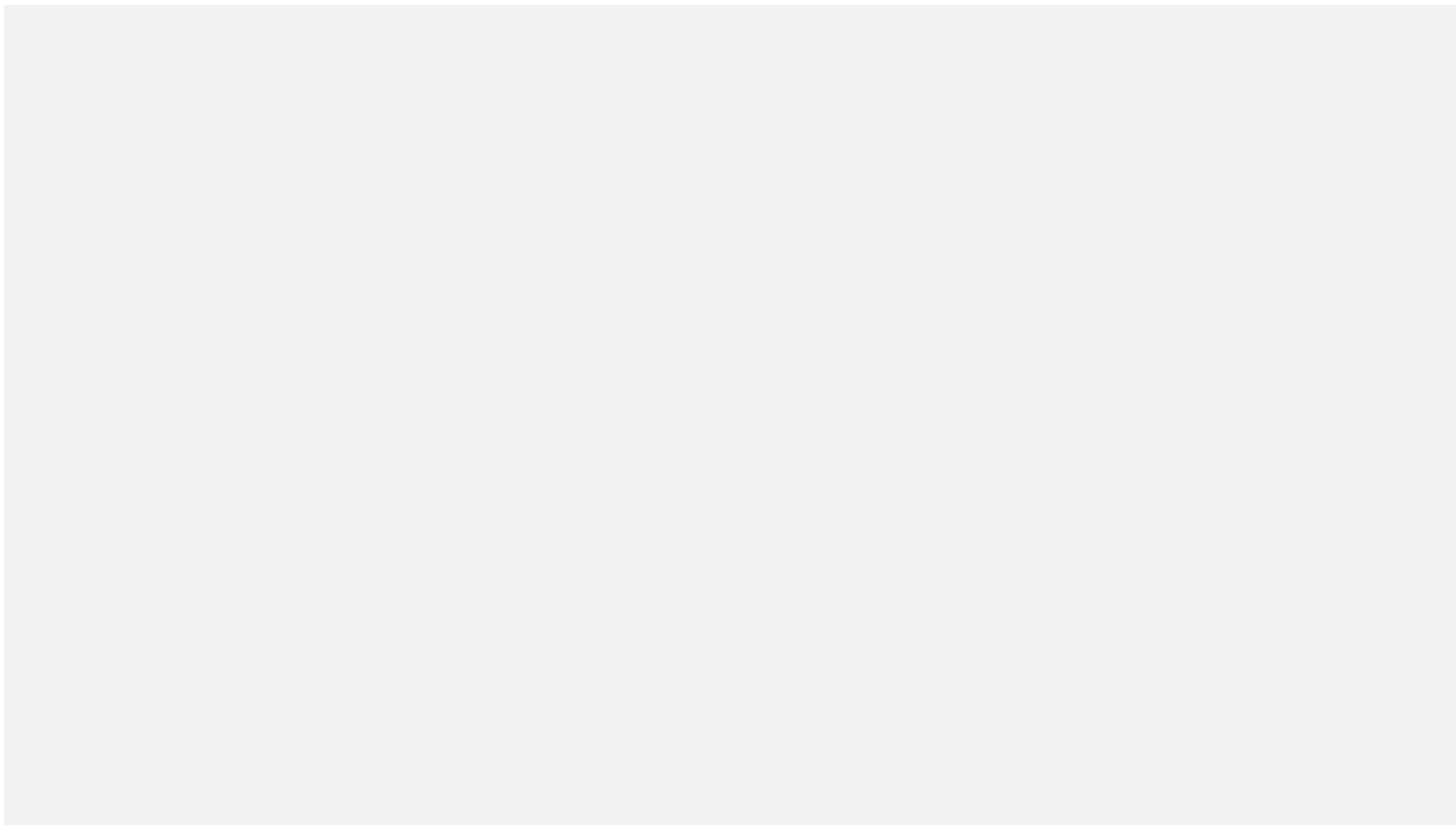
xubo@fudan.edu.cn

# 什么是知识图谱？

- 知识图谱本质上是一种语义网络
  - 结点
    - 实体
    - 概念
  - 边
    - 实体与实体
    - 实体与概念
    - 概念与概念
  - 目标
    - 描述真实世界中存在的各种实体或概念



# 中文开放百科知识图谱CN-DBpedia



# 中文开放百科知识图谱CN-DBpedia

- 是目前**最大规模**的开放百科中文知识图谱**之一**
- 涵盖**数千万**实体和**数亿**的关系
  - 百科实体数 16,537,283
  - 百科关系数 213,506,696
- 相关知识服务API累计调用量已达**2.6亿**次

# CN-DBpedia应用一： 语义搜索

http://kw.fudan.edu.cn/cndbpedi

entity

Search

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

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	<div><div></div><div></div></div>
学校代码	10246	<div><div></div><div></div></div>
学校地址	上海市杨浦区邯郸路220号	<div><div></div><div></div></div>
学校类型	综合	<div><div></div><div></div></div>
属性	111计划（2006年）	<div><div></div><div></div></div>

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	<http://dbpedia.org/ontology/Organisation>
rdf:type	<http://dbpedia.org/ontology/EducationalInstitution>
rdf:type	<http://dbpedia.org/ontology/University>

# CN-DBpedia应用二：小Cui问答



# CN-DBpedia应用三：超级验证码

- 传统验证方法已不再安全
- 基于知识图谱的验证码系统
- 以自然语言理解和问答为呈现形式



用户名

test

密码

\*\*\*\*\*

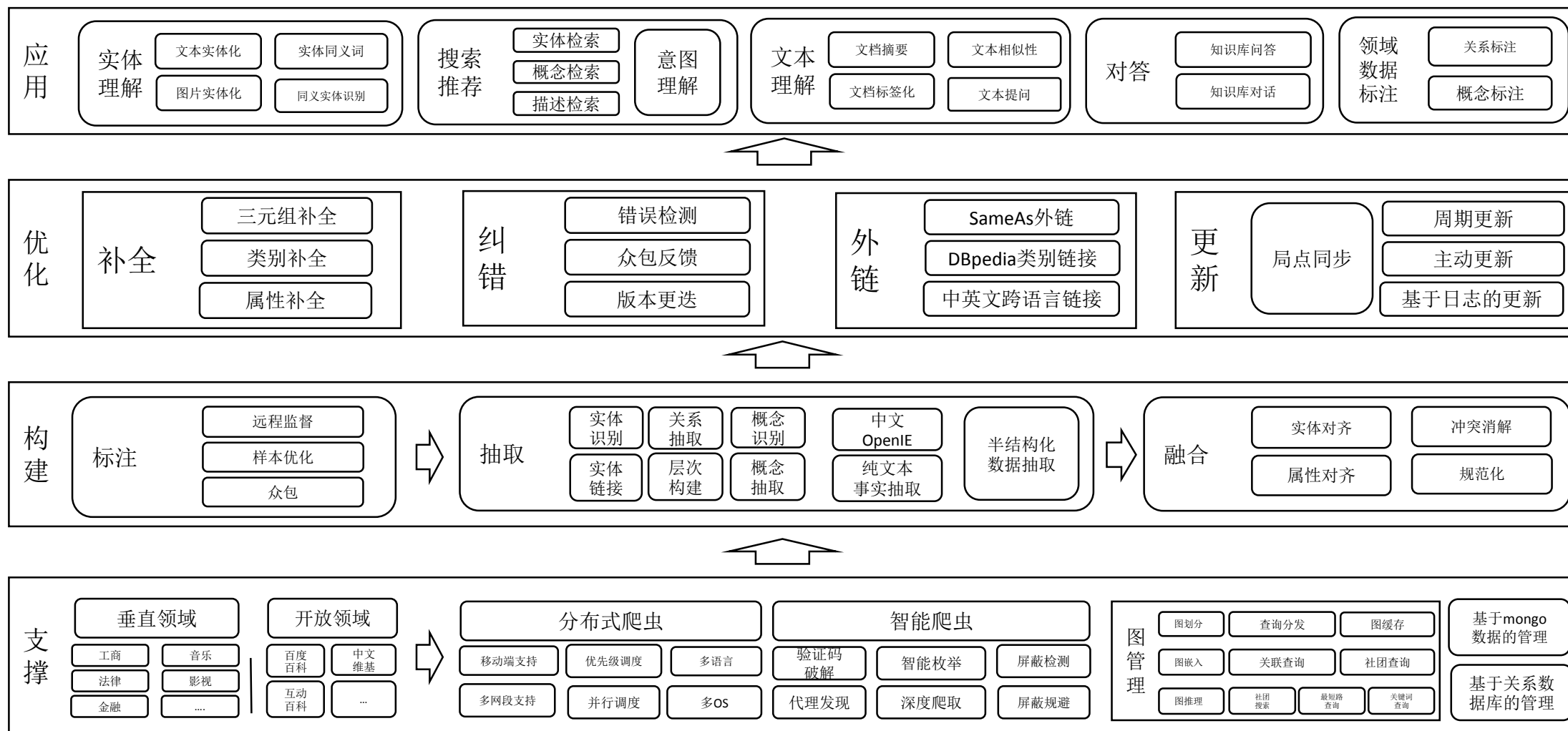
请通过验证

请点击下文该问题答案的任意部分: 岑钊雄的职业是什么? [太难了, 换一个](#)

岑钊雄 (Cen Zhaoxiong), 中国著名企业家、慈善家、社会活动家。1970年11月出生, 广东南海人

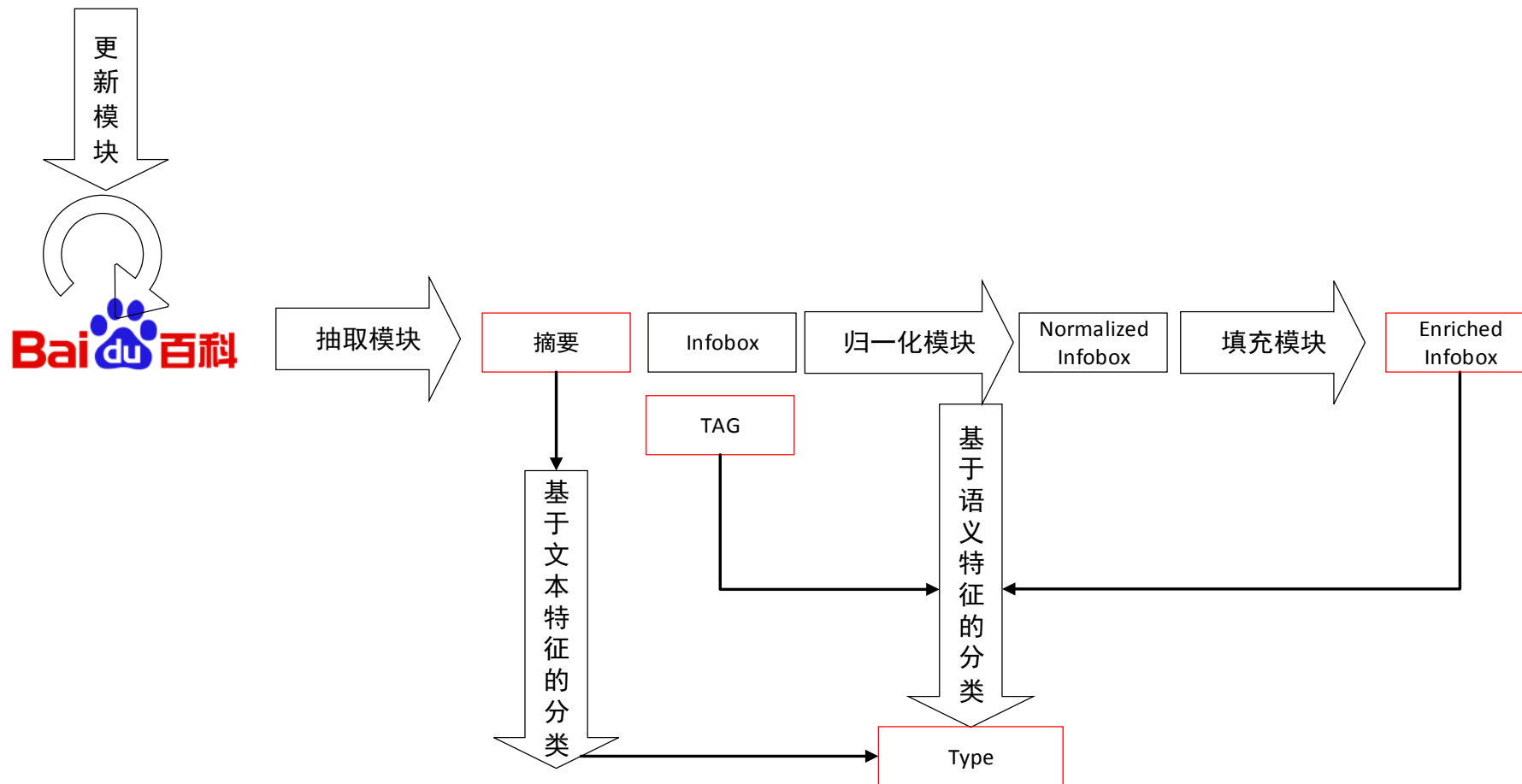
登录!

# CN-DBpedia系统框架

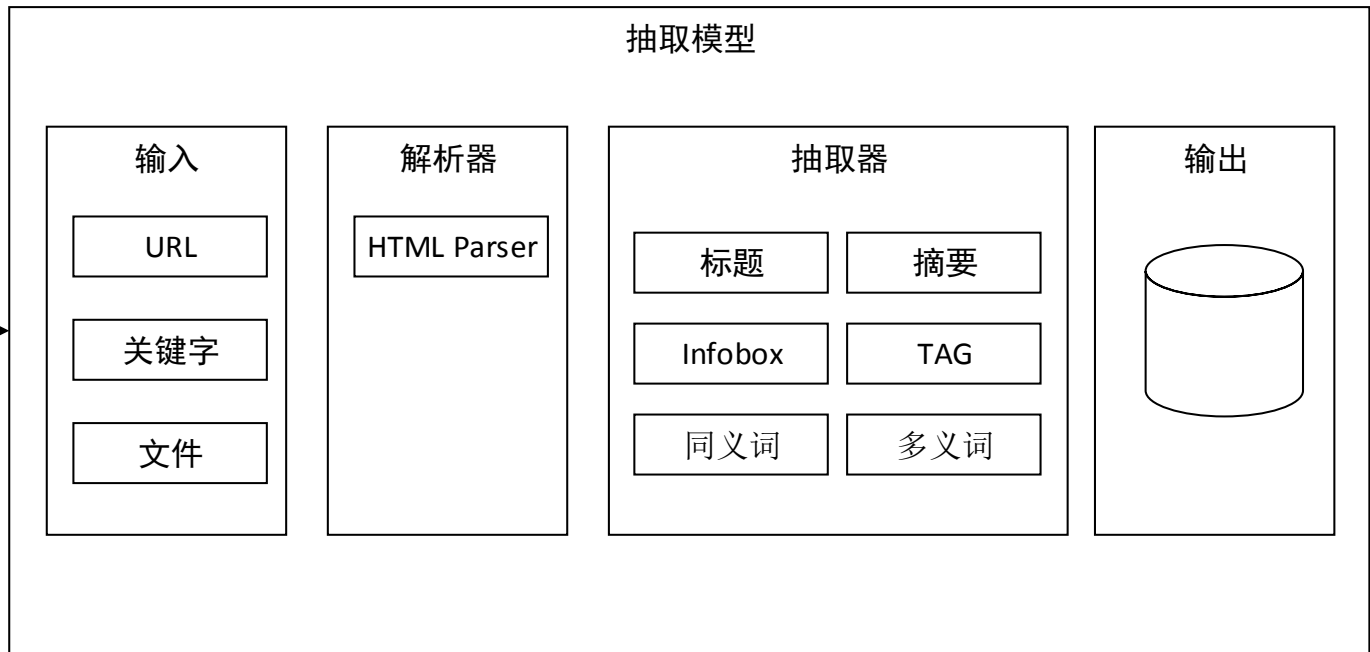




# 本次报告内容



# 抽取模块



输入一：URL

<http://baike.baidu.com/view/1.htm>

输入二：关键字

<http://baike.baidu.com/item/北京理工大学>

输入三：HTML文件

```
<!DOCTYPE html>
<!--STATUS OK-->
<html>
  <#shadow-root (open)
  <head>...</head>
  <body class="wiki-lemma feature small-feature collegeSmall">
    <div id="BAIDU_DUP_fp_wrapper" style="position: absolute; left: -1px; bottom: -1px; z-index: 0;
    <div class="header-wrapper pc-header-new">...</div>
    <div class="navbar-wrapper">...</div>
    <div class="body-wrapper feature feature_small collegeSmall">...</div>
    <div class="wgt-footer-main">...</div>
    <div class="lemmaWgt-searchHeader">...</div>
    <div class="new-bdsharebuttonbox new-side-share" id="side-share">...</div>
    <div class="qrcode-wrapper" id="layer" style="display: none">...</div>
  </div>
</div>
```

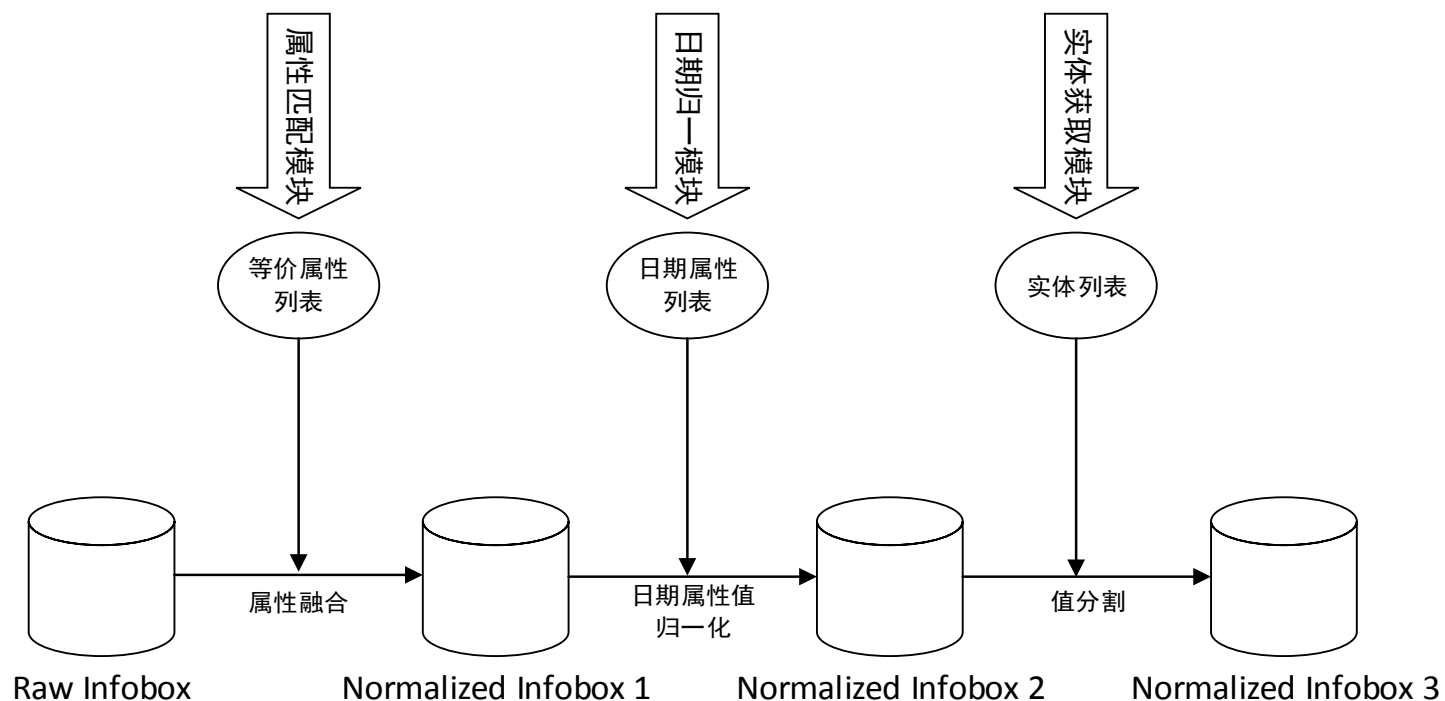
# 归一化模块

## 基本信息

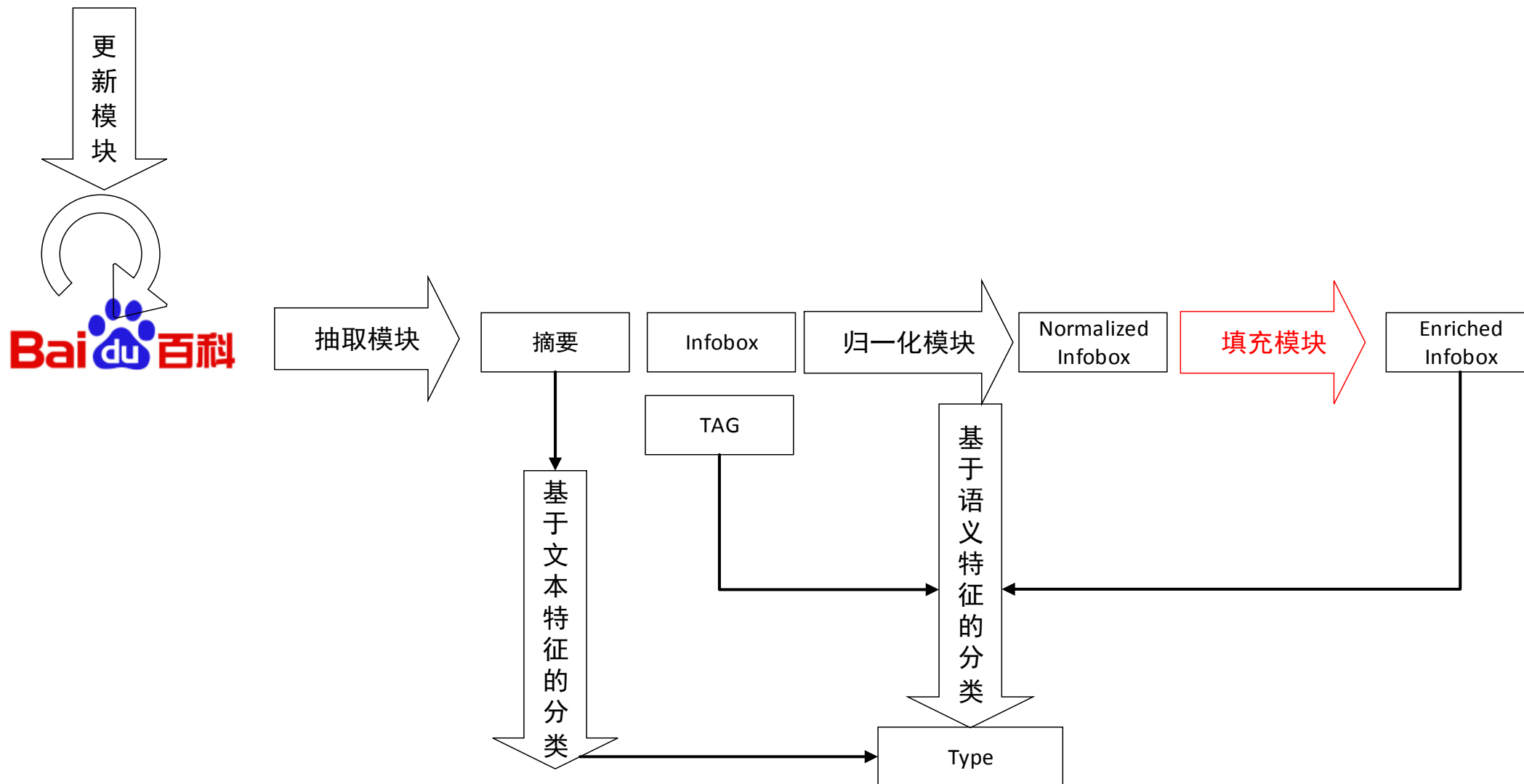
中文名	复旦大学	主管部门	中华人民共和国教育部
英文名	Fudan University	硕士点	243 个
简称	复旦·FUDAN	博士点	154 个
创办时间	1905年（乙巳年）9月14日	博士后流动站	35 个
类别	公立大学	校 训	博学而笃志，切问而近思
学校类型	综合	校 歌	《复旦大学校歌》
属 性	985工程（1999年） 211工程（1994年） 九校联盟（2009年） 珠峰计划（2009年） 111计划（2006年）	专职院士	中国科学院院士 21 人 中国工程院院士 5 人
所属地区	中国 上海	主要院系	中国语言文学系、哲学学院、历史学系、旅游学系 文物和博物馆学系、外国语言文学学院等
现任校长	许宁生	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2013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



# 填充模块（Infobox Completion）



# 填充方法

- 方法一：利用其它知识图谱进行填充
  - e.g. YAGO利用Geonames（一个包含超过1000万地点位置信息的地理知识图谱）来增加YAGO实体的地理位置信息
- 方法二：利用百科网站的其他语种进行填充
  - e.g. Wikipedia
- 方法三：利用百科网站实体标签进行填充
  - e.g. 如“刘德华”的一个分类信息为“香港演员”，可以从中得出（刘德华，出生地，香港）和（刘德华，职业，演员）两组Infobox
- 方法四：利用百科网站实体正文进行填充
  - 百科实体正文内容是对实体最全面的介绍，包含的信息最为丰富

# 利用百科网站实体正文内容进行填充

- 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）的句子中抽取出相应属性的值

## 属性值抽取器

刘德华（Andy Lau），  
1961年9月27日出生于  
中国香港。

“英文名称”

刘德华 英文名称 Andy Lau

“出生日期”

刘德华 出生日期 1961年9月27日

“出生地”

刘德华 出生地 中国香港

# 序列数据标记问题

- 文本属性值抽取本质上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 |

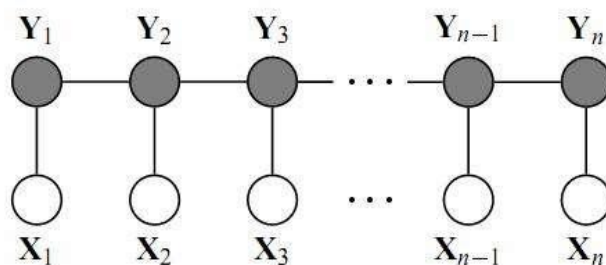
“出生日期”

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 |

# 传统分类方法



- 条件随机场
  - 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"	$TF - 1 \implies AA01$
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>

Wu, F., & Weld, D. S. (2007). Autonomously semantifying wikipedia.



# 基于深度学习的方法

- Embedding Layer

- Word-vec

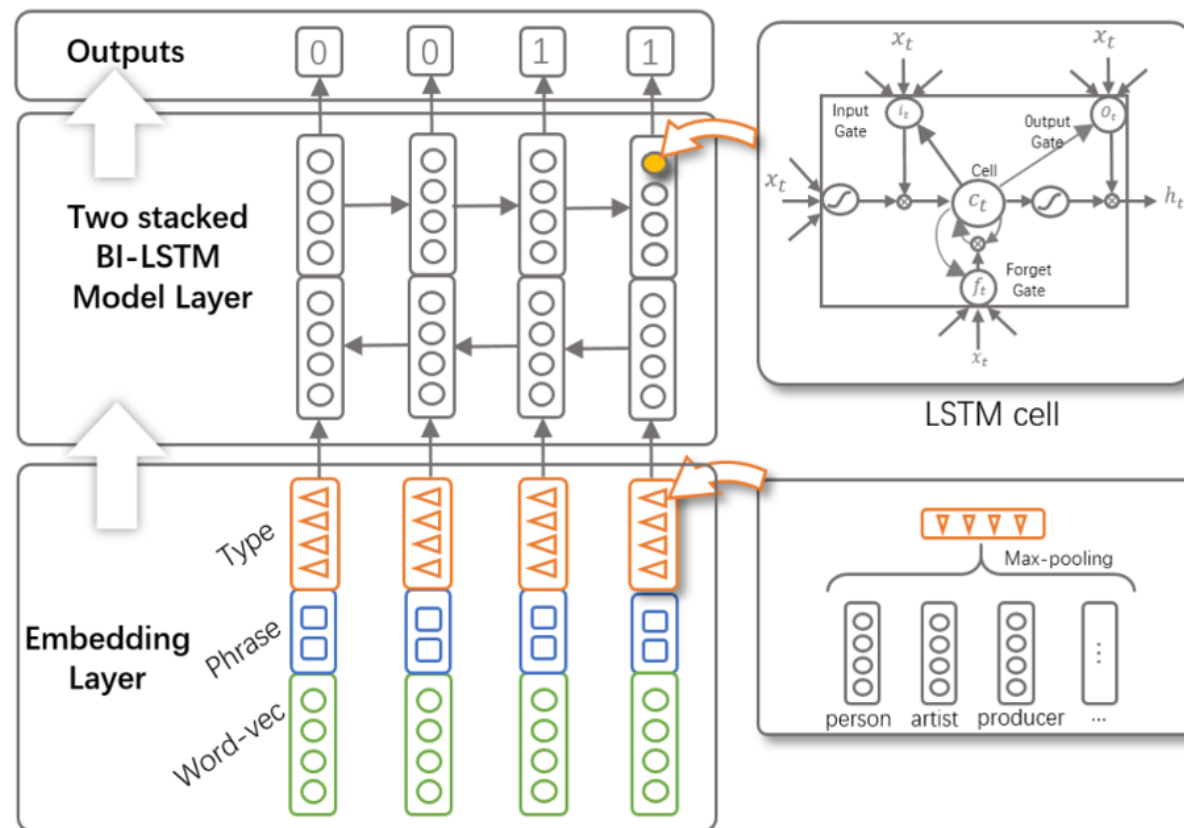
- 刘
    - 德
    - 华

- Phrase

- 刘德华

- Type

- Person
    - Artist
    - ...



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

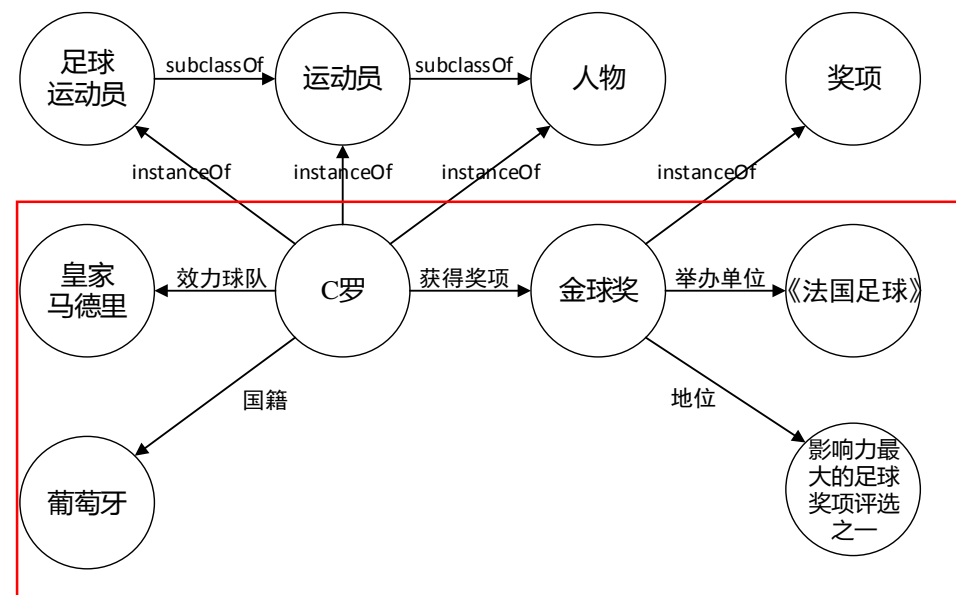
# 实体分类模块

# 知识图谱中的边

- 知识图谱中的边
  - 实体与实体（百科网站抽取）
  - 概念与概念（Taxonomy Construction）
  - 实体与概念（实体分类）

Taxonomy构建需要耗费巨大的人工，代价巨大

如何才能获得一个质量优良、又不需要太多人工的Taxonomy呢？

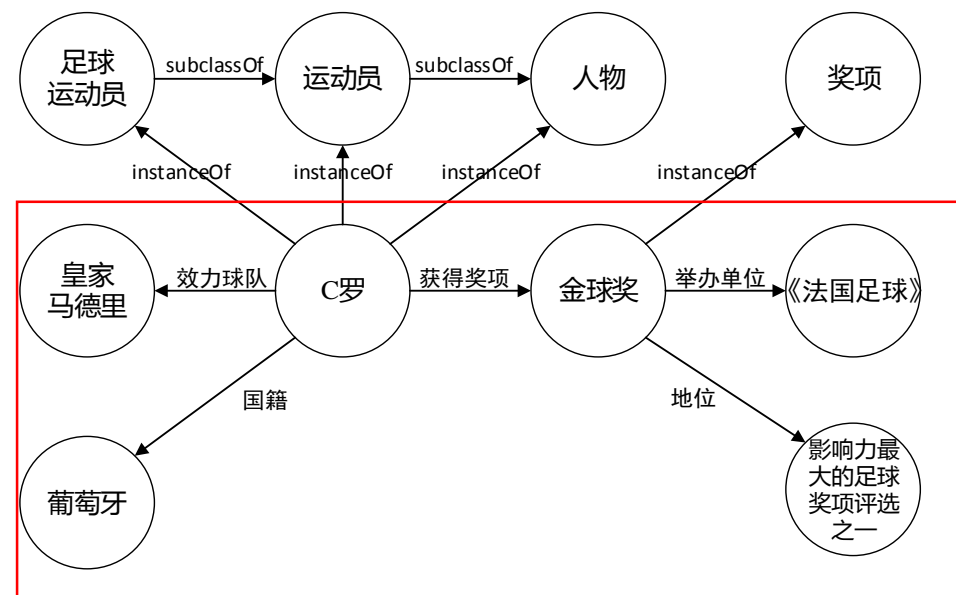


# 知识图谱中的边

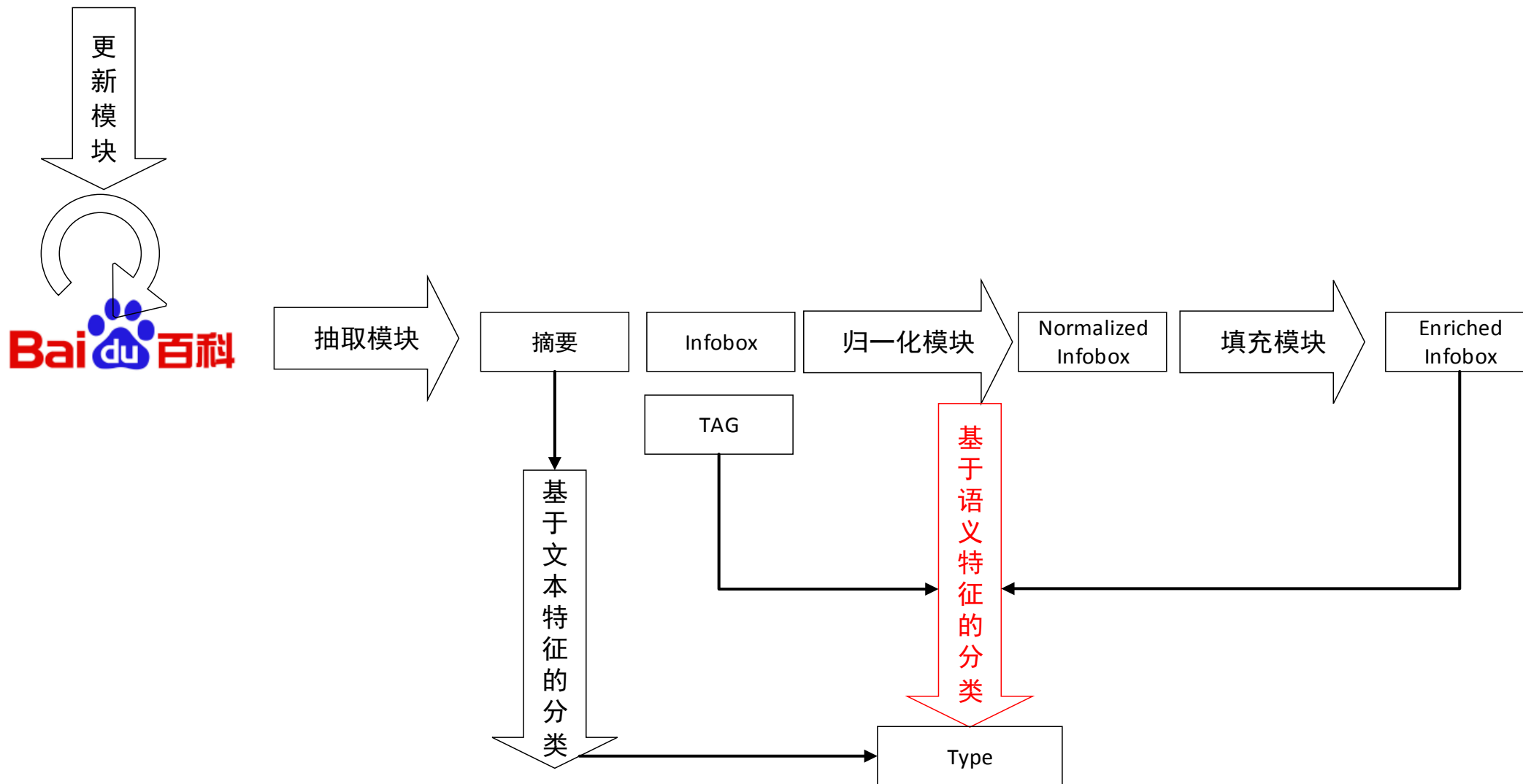
- 知识图谱中的边
  - 实体与实体（百科网站抽取）
  - 概念与概念（Taxonomy Construction）
  - 实体与概念（实体分类）

为此，我们提出了Taxonomy复用的方法

也就是将现有的、成熟的Taxonomy（如DBpedia、Yago、Freebase等）作为CN-DBpedia的Taxonomy

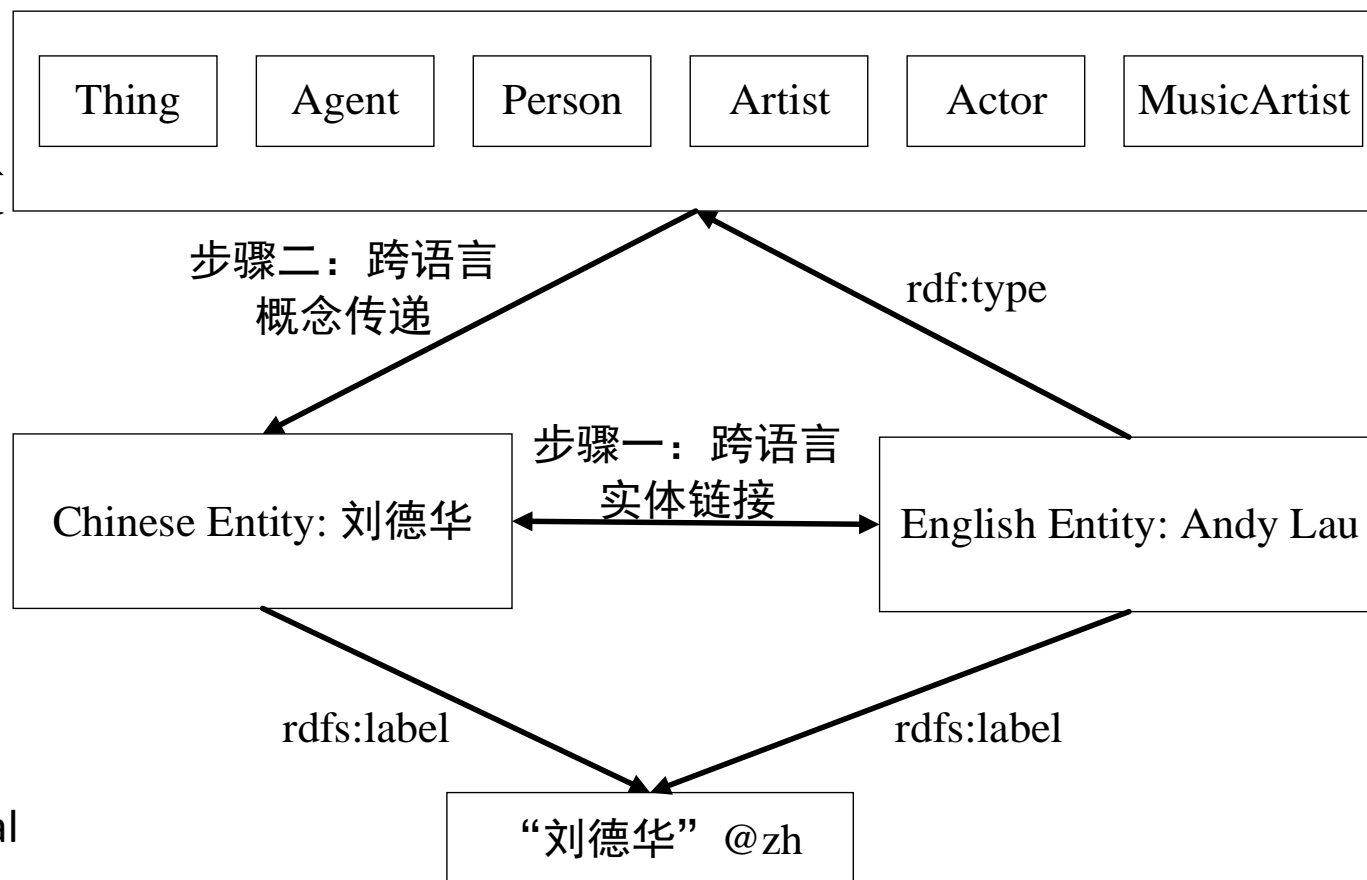


# 基于Taxonomy复用的实体分类



# 基于Taxonomy复用的实体分类

- 难点1：训练集构建
  - 中文实体无法直接分类到英文Taxonomy上
- 解决方案
  - 跨语言实体链接
  - 跨语言概念传递



Wang, Z., Li, J., Wang, Z., & Tang, J. (2012). Cross-lingual Knowledge Linking Across Wiki Knowledge Bases.

Wang, Z., Li, J., & Tang, J. (2013). Boosting cross-lingual knowledge linking via concept annotation.

# 基于Taxonomy复用的实体分类

- 难点2：训练集存在噪声
  - CASE 1：DBpedia 中的实体本身存在分类错误，这将导致对应的中文实体也分类错误
  - CASE 2：由于跨语言实体链接错误，导致中文实体分类错误
  - CASE 3：由于中文实体语义特征缺失，导致无法推断部分来自其对应英文实体的概念
- 解决方案
  - 对训练集中实体的分类结果进行多分类器投票过滤
  - 将训练集分为N份，其中每N-1份作为训练集，用来过滤另一份的结果
  - 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认为的噪声数据
  - 通过过滤策略对结果进行过滤

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}

# 跨语言实体分类—系统框架

解决英文知识图谱  
实体类别不完整的问题

Chinese KB

DBpedia Type Completion

English  
DBpedia

Entity Type  
Completion

提高训练集质量

Training Data Construction

Cross-Lingual  
Entity Linking

Cross-Lingual  
Type Propagation

Chinese Entities  
with DBpedia Types

Noise  
Filter

Training Data

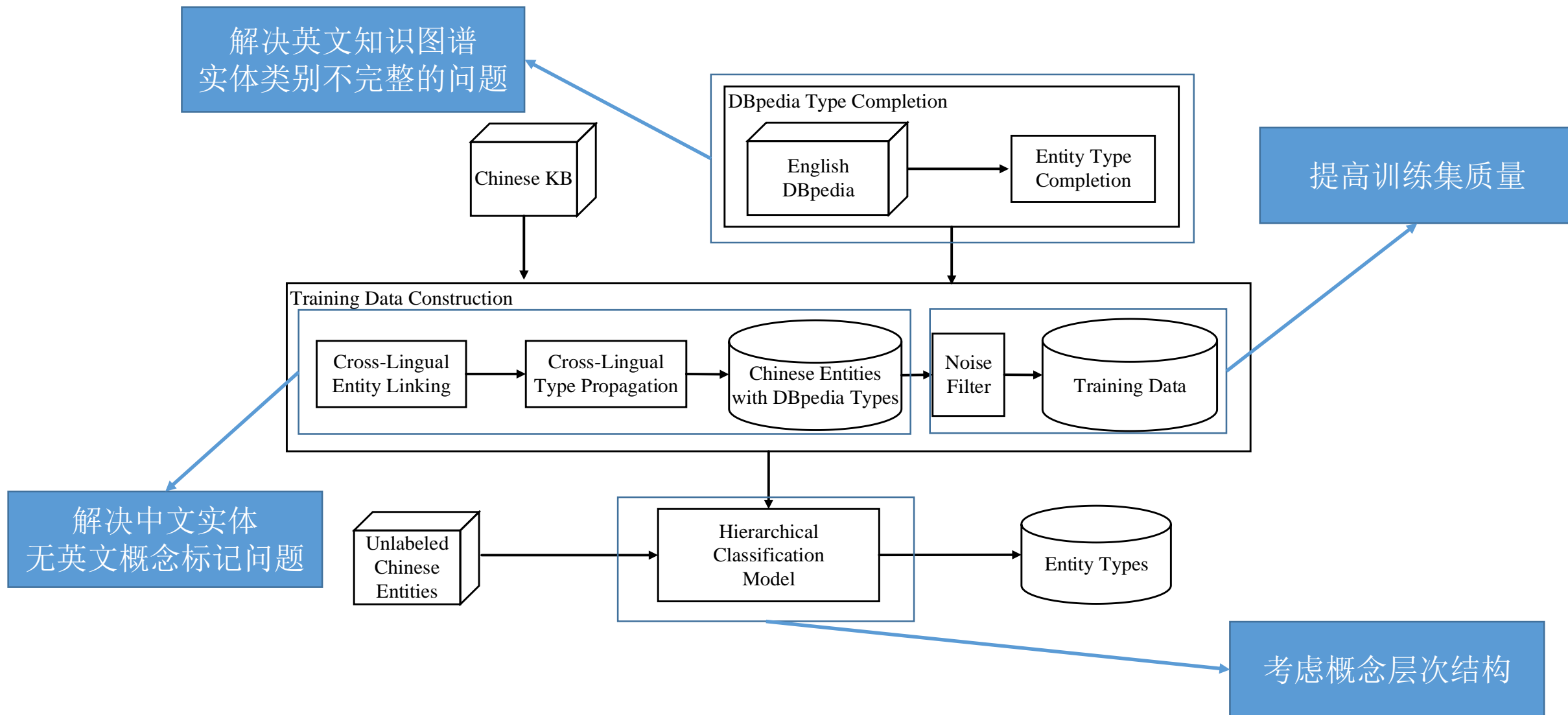
解决中文实体  
无英文概念标记问题

Unlabeled  
Chinese  
Entities

Hierarchical  
Classification  
Model

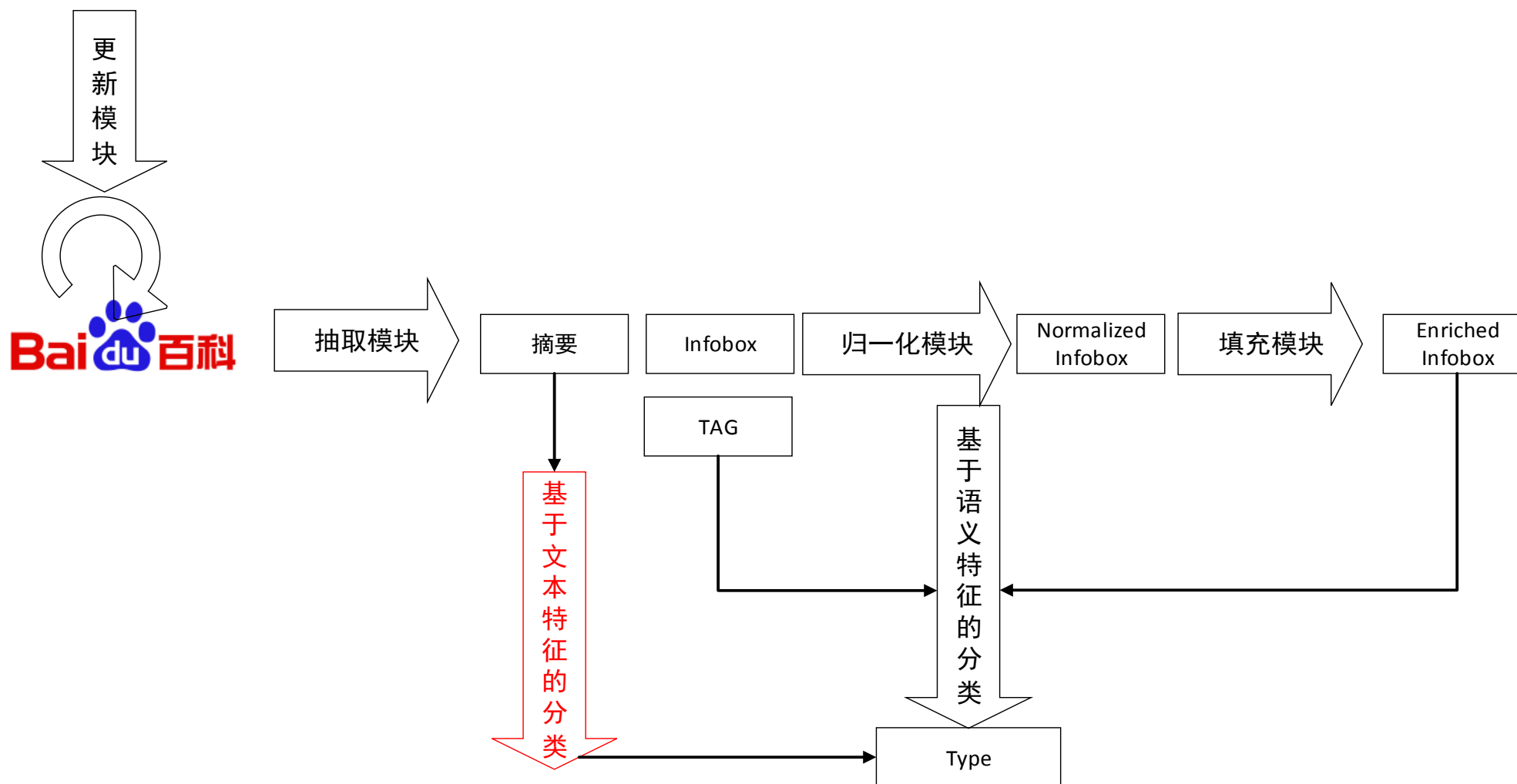
Entity Types

考虑概念层次结构

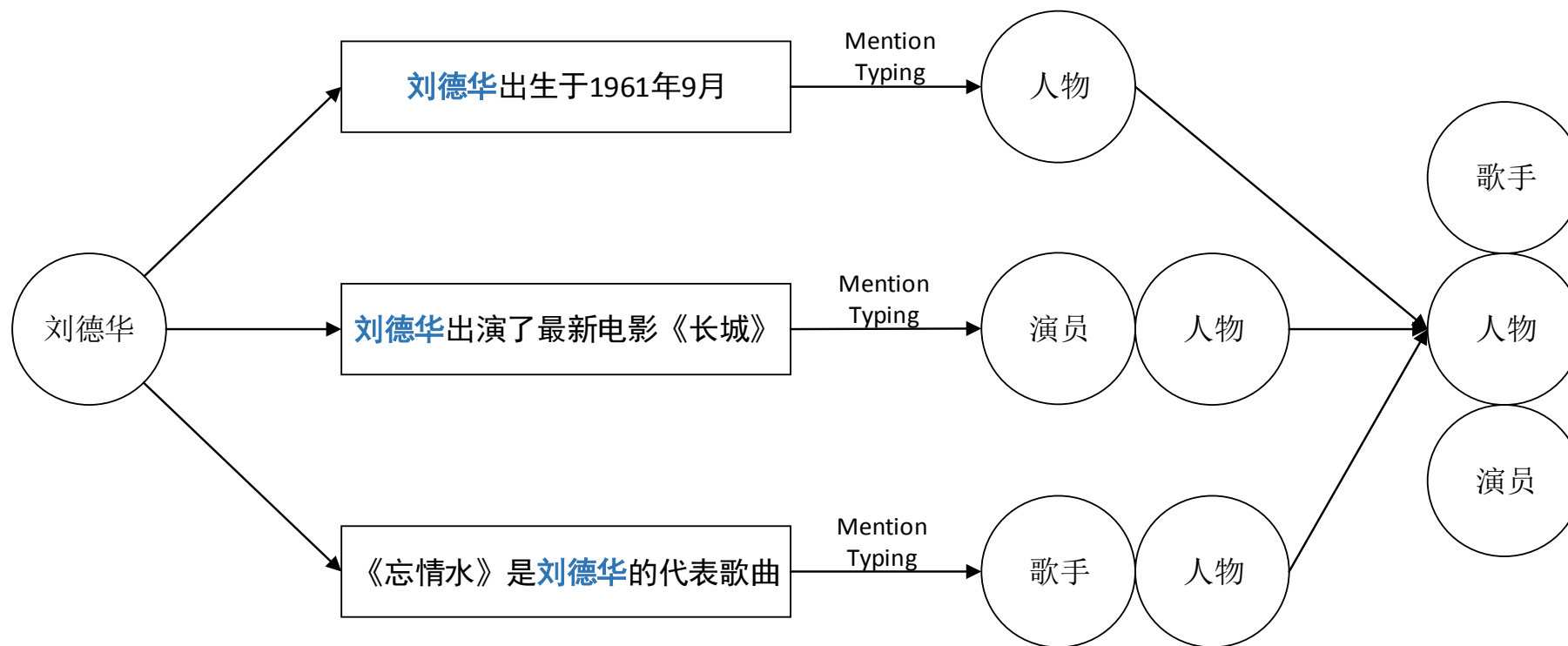




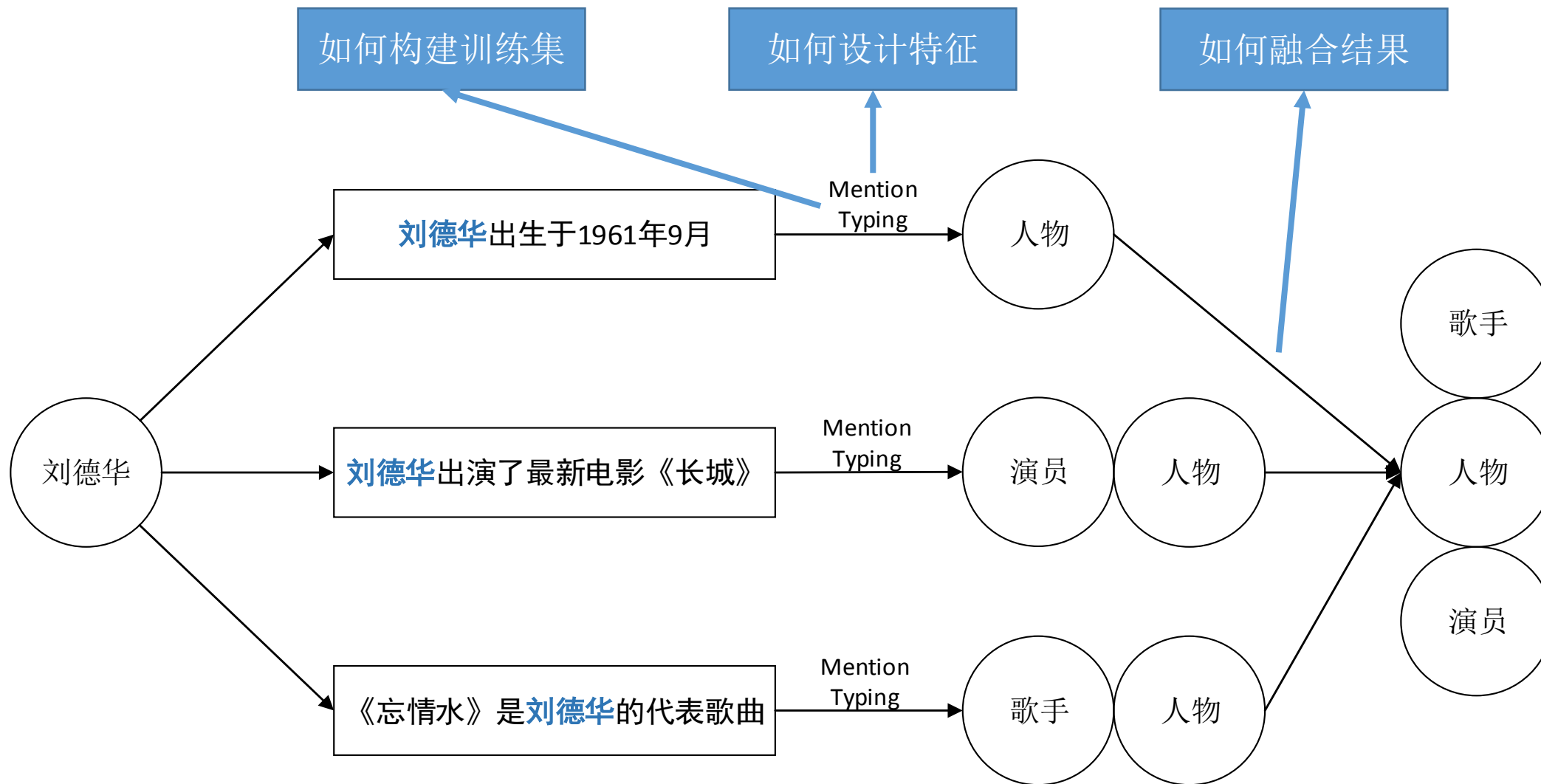
# 基于文本特征的分类



# 基本思路

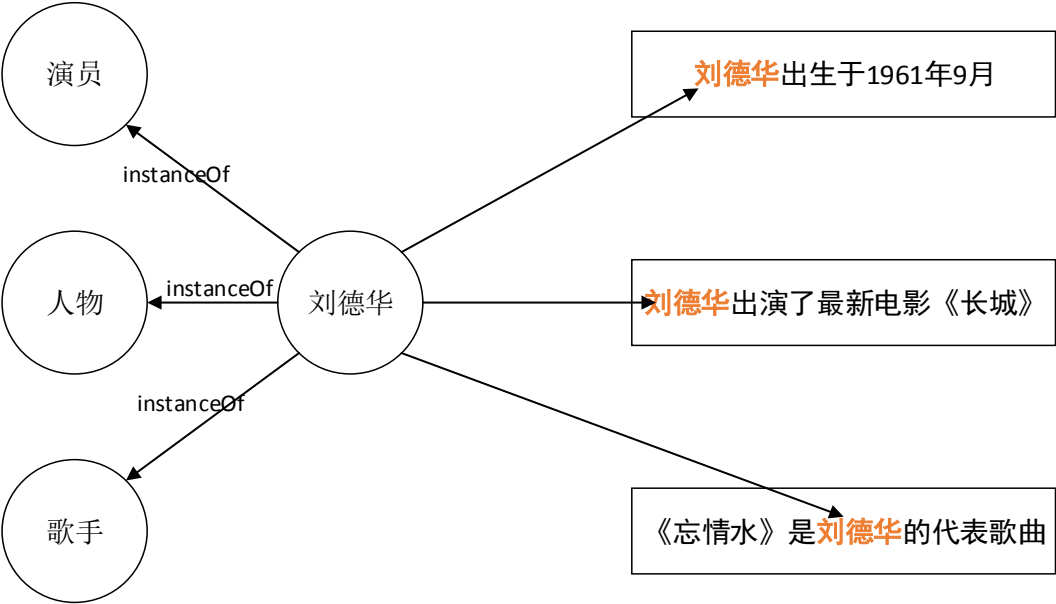


# 基于文本的实体分类—挑战



# 基于文本的实体分类

- 难点1：训练集构建
  - 人工标记代价大



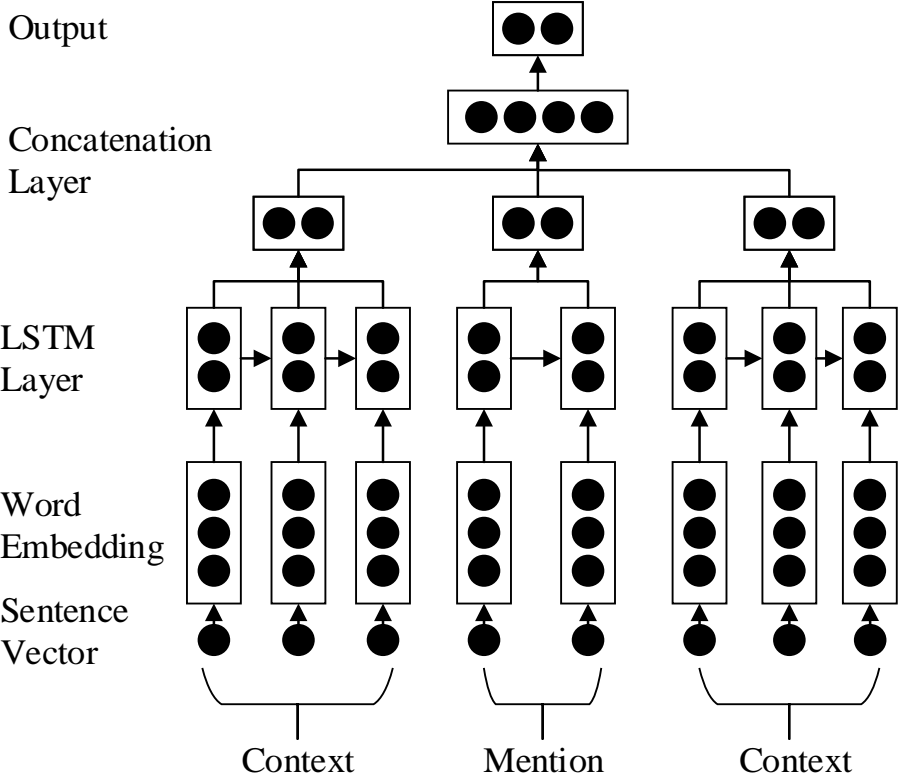
- 解决方案
  - STEP 1：基于远程监督的训练集构建
  - STEP 2：训练集噪声过滤
    - 多分类器投票过滤方法

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
    - 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	$[c_{-s} \cdots c_{-1}] [m_1 \cdots m_N] [c_1 \cdots c_s]$ , $S = 5$ 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

Maximize

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

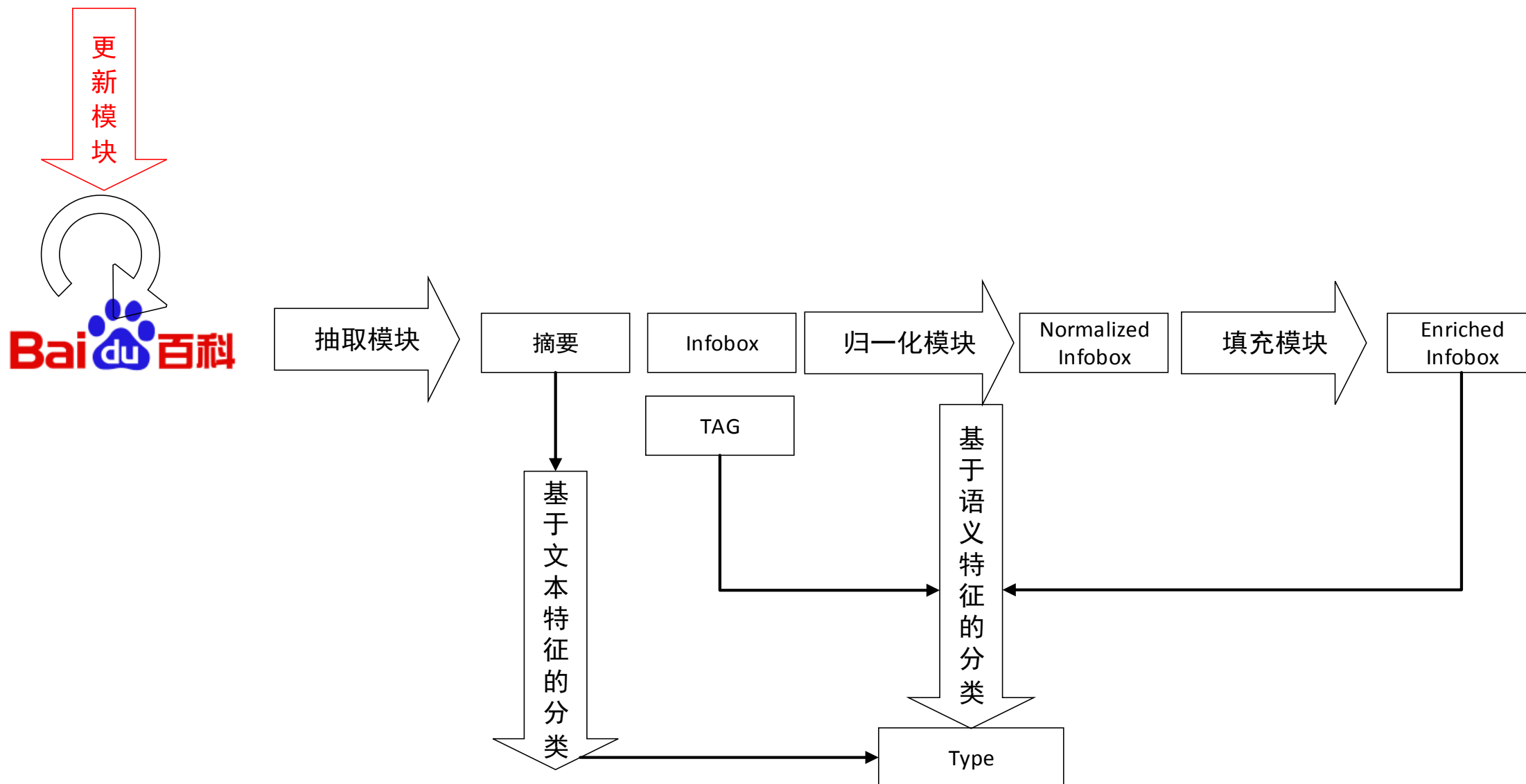
Subject to

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

## • 解决方案

- 将其看作是一个整数线性规划问题
  - 一个带约束的优化问题, 并且模型中的每个参数都要求为非负数
- 模型
  - 将所有mention的分类结果累加
- 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

# 更新模块



# 如何更新？


- 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
- 反馈更新
  - 用户点击更新按钮，进行更新
- 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

---

entity	Search
--------	--------

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

---

entity	Search
--------	--------

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

热搜词条 今天 | 昨天

顺丰菜鸟大战

↑

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。

菲律宾恐怖袭击

↑

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。

李晨

↑

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。

福特号航空母舰

↓

美国首艘“福特”级航母交付美国海军。

西班牙大厦

↑

孙怀山

↑

住房公积金

↑

星耀五洲

↑

热点要闻

个性推荐

进入推荐版

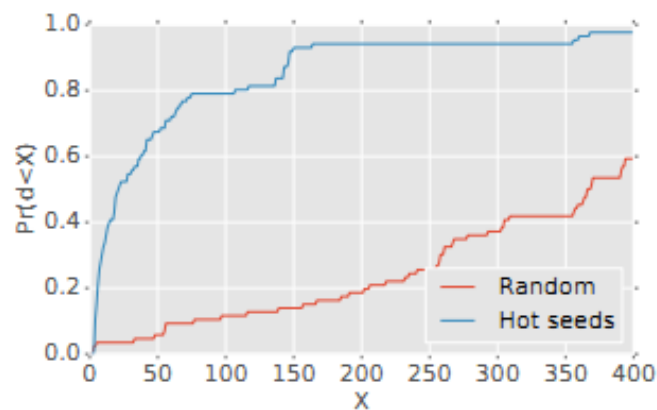
- 国际社会高度关注习近平哈萨克斯坦之行  
引领上合发展 共建一带一路 砥砺奋进的五年
- 李克强出席第十九次中欧领导人会晤 专题  
访德4大成果 张德江 俞正声 张高丽
- 上海等10省份今日举行事业单位招聘考试  
总招录人数超4.5万 多地强调要严肃考试纪律
- 安理会通过决议：强烈谴责朝鲜核导活动 扩大制裁
- 媒体：美国退出气候协定，中国的机遇来了？
- COSER穿中国军装向日式少女下跪 军媒：丢人且违法
- 内蒙古阿拉善盟阿拉善左旗附近发生5.0级左右地震
- 中国机动车近3亿辆 系PM2.5污染重要原因
- 误机掌握工作人员女博士：我知错了 能少关几天吗
- 境外消费超千元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

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

# 为什么将热词作为更新的种子结点？

- 实证分析
  - 实验
    - 统计热词的更新频率和随机选择的实体的更新频率
  - 结果
    - 80%的热词在100天内更新过了
    - 10%的随机选择的实体在100天内更新过了



# 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
- 实证分析
  - 实验
    - 统计80个种子实体扩展出来的687个实体的更新频率和随机选择的实体的更新频率
  - 结果
    - 269/687（大约40%）的扩展实体在一个月进行了更新
    - 而只有3%的随机实体在一个月进行了更新



# 更新框架

- 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
  - 如果是旧词，从知识库中获取
  - 如果是新词，从最新的百科页面中获取
  - 之所以要设置优先级而不是更新所有扩展实体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 $E[u(x)]$

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

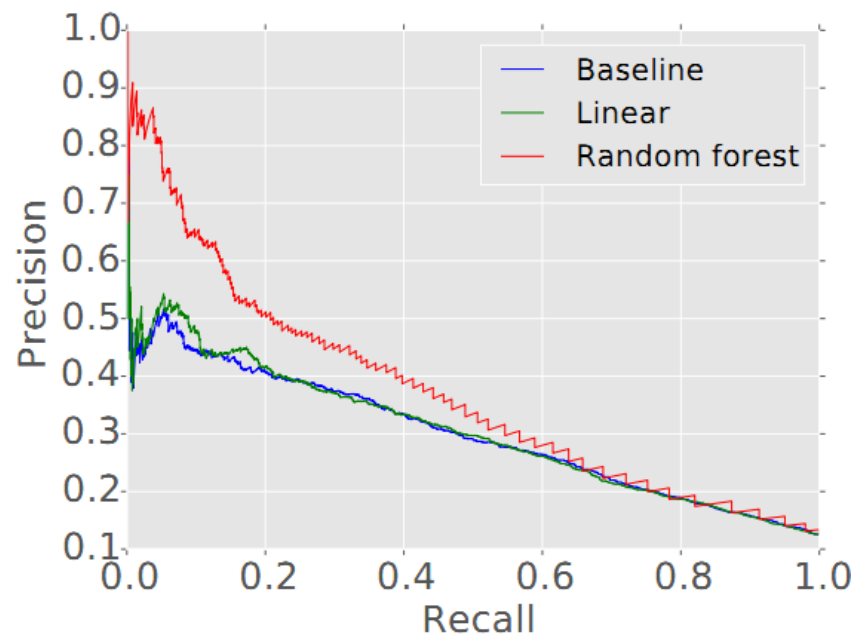
- $P(x)$ : 为实体 $x$ 预期的更新频率，通过预测器得到
- $t_s(x)$ : 为最近一次更新的时间
  - 如果 $x$ 是一个新词， $t_s(x) = -\infty$

# 期望更新频率预测器

- 模型：回归
  - 线性回归
  - 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

- 评估



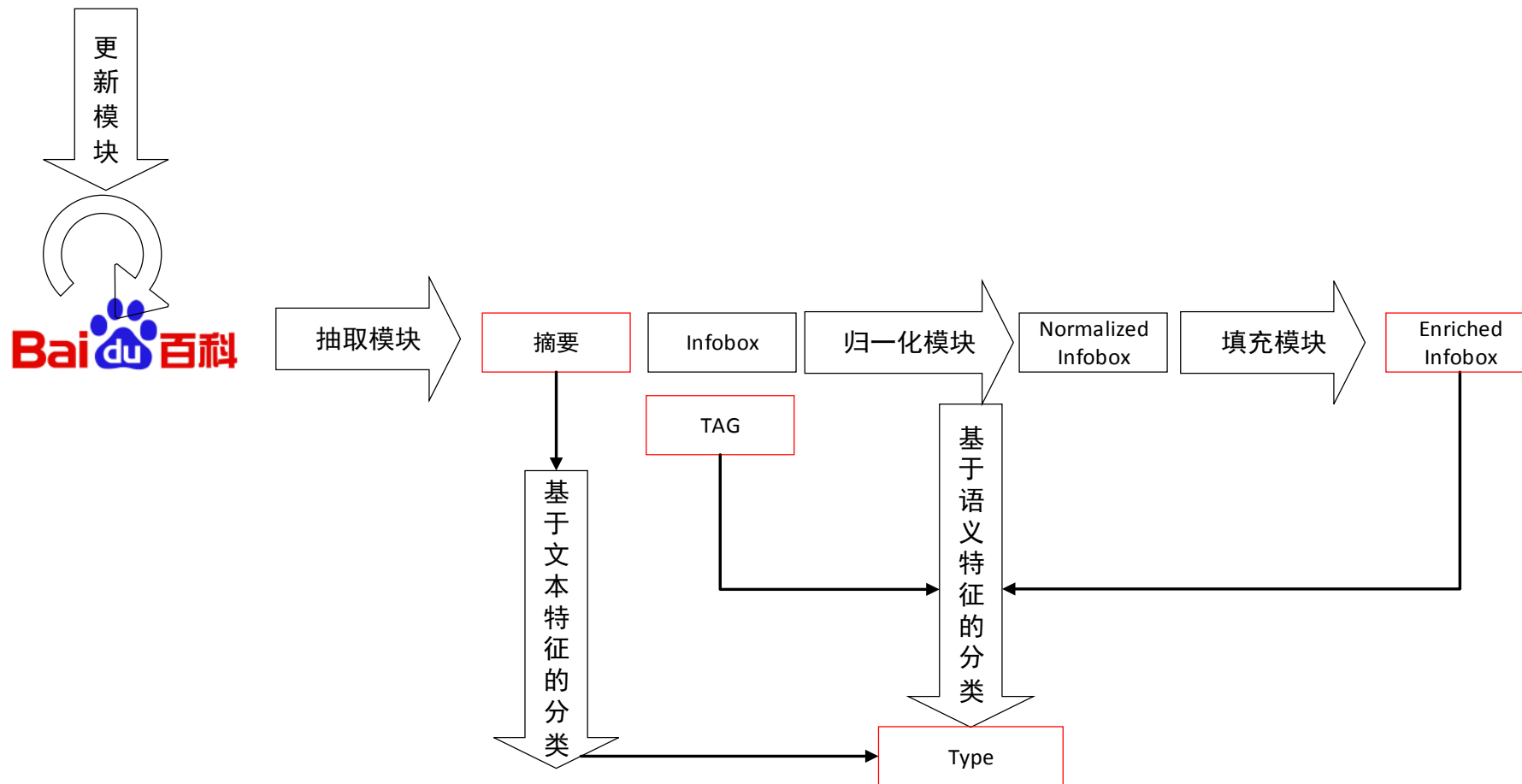
Model	MSE	AUC
Baseline	0.0400	0.2992
Linear	0.0367	0.3021
Random forest	<b>0.0315</b>	<b>0.3692</b>

# 更新系统评估

- 我们将这套更新机制布置到CN-DBpedia中
- 设置K（每日更新实体个数上限）为1000
- 我们系统在一天中爬取了1000个实体，其中68.7%的实体的信息发生了变化

Total visits	Success updates	Success ratio
50	46	92.0%
100	90	90.0%
200	175	87.5%
500	398	79.6%
1000	687	68.7%

# 总结





# 知识工场实验室介绍

