

基于中文维基百科构建的 知识库的谓词归一

韩喆

iampkuhz@gmail.com

2015 年 4 月 3 日

- 背景
 - 基于维基百科的知识库
- Motivation
- 实验步骤
- 特征选取及分析
- 实验分析和改进

知识库背景

Background

基于中文维基百科 (网页) 的知识库

张家辉



2010年8月24日参加电影《线人》江展官授牌。

男演员

罗马拼音 Cheung Ka Fai

英文名 Nick Cheung

国籍  中国 (香港)

籍贯 广东番禺

出生 1967年12月2日 (47岁)

出生地  英属香港

语言 粤语、英语、普通话

配偶 关咏荷 (2003年至今)

儿女 张童 (Brittany Cheung) - 2006年01月24日 (9岁)

活跃年代 1987年至今

经纪公司 钟珍^[1]

奖项 最佳男演员 - 上海国际电影节
2013年《线人》- 影评
最佳男演员 - 香港电影评论学会

Subject	Predicate	Object
张家辉	罗马拼音	Cheung Ka Fai
张家辉	英文名	Nick Cheung
张家辉	国籍	中国(香港)
张家辉	籍贯	广东番禺
张家辉	出生	1967年12月2日 (46岁)
张家辉	出生地	英属香港
张家辉	语言	粤语
张家辉	语言	英语
张家辉	语言	普通话
张家辉	配偶	关咏荷(2003年至今)
张家辉	儿女	张童(Brittany Cheung)
张家辉	活跃年代	1989年至今
张家辉	经纪公司	钟珍

- 330w 三元组, 1.6w 个谓词

Motivation

Motivation

- ① 知识库的谓词数量多
 - 1.59w, 手工排查后变成 1.4w
- ② 谓词冗余

含有“邮政”的谓词 (17)

INSEE/邮政编码、ISO 3166-2 邮政简写、美国邮政编号、美国邮政编码、邮政、邮政代码、邮政信箱、邮政分区、邮政区号、邮政号码、邮政简称、邮政编号、邮政编号字母、邮政编码、邮政编码 FSA、邮政编码首字母、邮政缩写

- ③ 衍生查询/知识库合并
 - 推荐相似的谓词给查询者
 - 将不同知识库合并时提供谓词归一的规则
- ④ 所以要进行谓词归一(没有搜到相关文章)

实验步骤

- 假设/前提

- 我们在 1.4w 个候选谓词内部进行实验
- **初始问题**: 请提供任意两个谓词的相似度, 进而判断意义是否相同
- 假设所有字符相同的谓词都是同一谓词, 所有字符不同的谓词都非同谓词
 - × 姚明: 出生: 上海 vs 刘翔: 出生地: 上海
 - ✓ 姚明: 出生: 上海 vs 刘翔: 出生: 1983 年 7 月 13 日
- **转换问题**为二分类: 给定任意两个谓词对, 判断其是否是相同谓词
 - 聚类转分类
 - 训练数据格式: $[true/false, PredicateId_1, PredicateId_2]$
 - 测试数据格式: $[PredicateId_1, PredicateId_2]$

- 实验环境/数据
 - 自己手工标注了 1700 多个谓词对
 - 谓词对本身根据规则（有一定拼音、字符串等相似性）抽取
非随机抽取
 - 785 个相同谓词对（47.3%），873 个不同谓词对
 - 测试集 1000 个单词对，训练集 500 个单词对
 - 全部分类为 true: 52.9% correct

● 实验步骤

- ① 对于每个谓词 (1/14000), 统计其信息 (不同类别的特征)

出生 : pinyin={chusheng}, Content={出生}, SubjectCategory={(篮球运动员,10),(足球运动员,100),(政治人物,50)}...

- ② 对于任意两个谓词, 比较其每类特征的相似性, 转化为数值, 生成特征向量

出生, 出生地 : pinyinSim=0.67, ContentSim=0.67, SubjectCategorySim=0.38,...

- ③ 对于训练数据, 提取特征向量, 训练模型
- ④ 对于测试数据, 提取特征向量, 根据模型预测是否为同一谓词

特征选取及分析

- 已选特征
 - 文本相似度
 - 拼音相似度
 - 词频相似度
 - wikitext 相似度
 - 主体的类别相似度

- 文本相似度

- ① 相同单词个数/总长度 (2 维)
- ② $\min(\text{编辑距离}/\text{总长度}, 1)$ (2 维)
- ③ 61.8% correct on SVM

- 拼音相似度

- ① 同文本相似度计算方式, 比较字符相同时改用拼音判段是否相同
- ② 53.3% correct on SVM

- 词频相似度














- ① 初衷是希望出现频率差别越大的谓词越应当合并 (判重), 实际基本没有效果

Method

● wikitext 相似度: 期望的重点

張家輝	
男演員	
罗马拼音	Cheung Ka Fai
英文名	Nick Cheung
国籍	 中国（香港）
籍贯	广东番禺
出生	1967年12月2日（47岁） <div> 英属香港</div>
语言	粤语、英语、普通话
配偶	关咏荷（2003年至今）
儿女	张童（Brittany Cheung） <div>- 2006年01月24日（9岁）</div>
活跃年代	1987年至今
经纪公司	锺珍 ^[1]

任何侵权内容将会删除 | 百科内容须附有来源，以供查证

A A                   

Method

- wikitext 相似度.
 - 没有固定的对应规则
 - (比方说) 编辑者在“Template: 男艺人”页面写了一个转换说明, 把“出生日期”自动转化为“出生”显示。如果没有定义, 则用模板“Template: 人物”的规则匹配。且说明页面非结构化, 不能自动抽
 - 收集了从 wikitext 抽取的三元组, 利用手写规则与从网页抽取的三元组做对应, 然后做统计

Method

- **wikitext 相似度.**
 - 没有固定的对应规则
 - (比方说) 编辑者在“Template: 男艺人”页面写了一个转换说明, 把“出生日期”自动转化为“出生”显示。如果没有定义, 则用模板“Template: 人物”的规则匹配。且说明页面非结构化, 不能自动抽
 - 收集了从 wikitext 抽取的三元组, 利用手写规则与从网页抽取的三元组做对应, 然后做统计

内核类别

- $\{(\text{kernel type}, 132), (\text{screenshot}, 2), (\text{logo}, 2), (\text{name}, 2), (\text{kernel}, 1)\}$

出生

- $\{(\text{birth place}, 11470), (\text{birth date}, 7598), (\text{出生地点}, 7241), (\text{出生日期}, 6789), (\text{date of birth}, 3775), (\text{place of birth}, 3690), (\text{term start}, 2346), (\text{出生地}, 2076), (\text{term end}, 1511), (\text{birthplace}, 1156) \dots\}$

Method

- wikitext 相似度.
 - 实验效果

Method

- wikitext 相似度.
 - 实验效果
 - SVM 分类失败（全部预测为 1...）

- wikitext 相似度.
 - 实验效果
 - SVM 分类失败（全部预测为 1...）
 - 失败原因
 - 80% 的测试数据的相似值为 0。很多时候有一个谓词没有对应的 wikitext，尤其是出现频率少的谓词
 - 下一步修正
 - 观察没有抽到 wikitext 的谓词信息，修改代码（理论上都是可以对应有 wikitext 的）

- 主体的类别相似度 二级类别分布。
 - 假设前提：意义相同的谓词，其所在的三元组的主体的类型分布应该是一致的。
 - “出生”的主语类别分布 {(人物,10000),(动物,100)}
 - “出生日期”的主语类别分布 {(人物, 2000), (动物,500)}
 - 实验方法
 - 利用中文维基百科的类别，“页面分类”下面的子类 (22-2) 作为类别分布的规约终点

语言, 跨學科領域, 应用科学, 文学, 艺术, 宗教, 休閒, 科技, 心理学, 人物, 地理, 人文學科, 技术, 社会, 历史, 幫助, 資訊, 科学, 總類, 自然科学, 社会科学, 哲学,

- 主体的类别相似度 **二级类别分布**。

- 利用维基百科的类别体系，建立所有类别到这 22 个类别的对应关系

分类:美国篮球运动员

页面分类 > 人物 > 职业 > 各职业美国人 > 美国运动员 > 美国篮球运动员

页面分类 > 人物 > 各国人物 > 各国运动员 > 美国运动员 > 美国篮球运动员

页面分类 > 人物 > 各职业人物 > 运动员 > 篮球运动员 > 美国篮球运动员

... > ... > 球类运动 > 篮球 > 篮球运动员 > 美国篮球运动员

... > ... > 篮球 > 各国篮球 > 美国篮球 > 美国篮球运动员

... > ... > 各国体育 > 美国体育 > 美国运动员 > 美国篮球运动员

... > ... > 各国体育 > 美国体育 > 美国篮球 > 美国篮球运动员

... > ... > 各国体育 > 各国运动员 > 美国运动员 > 美国篮球运动员

... > ... > 各国体育 > 各国篮球 > 美国篮球 > 美国篮球运动员

- 20 个节点**宽度优先**向下搜索

- 深度优先失败, 所有类别都是**语言**的子类

- “**雷·阿伦**: **出生**: 加利福尼亚州”

雷·阿伦属于类别“美国篮球运动员”

出生: {(人物, 100), (科技, 10), ...} -> {(人物, 101), (科技, 10), ...}

总结

- 目前效果

- 69.1% correct on SVM; f1: 0.713
 - 类别信息、wikitext 特征虽然有，但是 bug 太多
觉得应该做到 80% 左右是可以接受的程度
 - 二级类别分布特征还在（bu）改（ren）进（zhi）中（shi）...
- 当时抽特征的时候，没有及时仔细检查，能跑出结果就行...

实验分析和改进

Method (to do)

- 下一步工作

- ① 类别信息特征 bug
- ② 规约类别修正
 - 科学是自然科学的父类，但都是规约重点
 - 类别分布向量直接加入特征向量
原来是做的余弦相似度的值
- ③ 增加含有 wikitext 信息的谓词数量
- ④ 频率特征重利用
 - 待完善思路：频率低的谓词，应当抽取其客体的语义信息
- ⑤ 命名实体类别分布特征
 - freebase.NER 提供了维基百科实体的映射关系
通过 type->People 的类别判断 freebase 实体的类型 {people, Location, organization, other}

any questions || any suggestions ?