第 43 卷 第 3 期 计 算 机 工 程 2017 年 3 月
Vol. 43 No. 3 Computer Engineering March 2017

·人工智能及识别技术· 文章编号:1000-3428(2017)03-0225-07 文献标志码:A 中图分类号:TP18

基于中文知识图谱的人物实体识别

李 薇,肖仰华,汪 卫

(复旦大学 计算机科学技术学院,上海 201203)

摘 要:分类是知识图谱构建中的一个重要问题,但是目前多数中文百科都采用人工编辑的方式为词条添加分类,耗费人力并且存在漏标和标错等问题。为此,提出一种自动识别百度百科人物领域下全部实体并添加分类的方法。对百度百科词条已有的分类、属性和副标题进行实体集拓展,使用马尔科夫逻辑网络方法联合推断词条的分类。实验结果表明,与支持向量机和逻辑回归算法相比,该方法在实体识别的精确度和召回率方面性能均有所提升。

关键词:实体分类;实体集拓展;马尔科夫逻辑网络;知识图谱;机器学习;联合推断

中文引用格式:李 薇,肖仰华,汪 卫.基于中文知识图谱的人物实体识别[J].计算机工程,2017,43(3):225-231,240.

英文引用格式:Li Wei,Xiao Yanghua,Wang Wei. People Entity Recognition Based on Chinese Knowledge Graph[J]. Computer Engineering,2017,43(3):225-231,240.

People Entity Recognition Based on Chinese Knowledge Graph

LI Wei,XIAO Yanghua,WANG Wei

(School of Computer Science,Fudan University,Shanghai 201203,China)

【Abstract】 Classification is an important issue in constructing knowledge graph. However,the existing categories are edited by human beings for Chinese Baike websites,leading to increasing manpower cost and lost or error of categories. Aiming at this problem,this paper proposes a method to recognize all entities of Baidu Baike. under people domain automatically and add categories for these entities. Entity Set Expansion(ESE) is used for existing raw data of like categories,attributes and subtitles,and Markov Logic Network(MLN) is used to jointly inference the category of entities. Experimental result shows that,compared with Support Vector Machine(SVM) and Logistic Regression(LR) algorithm,the proposed method performs much better both in precision and recall aspects.

【Key words】 entity classification; Entity Set Expansion(ESE); Markov Logic Network(MLN); knowledge graph; machine learning; joint inference

DOI:10.3969/j.issn.1000-3428.2017.03.038

0 概述

知识图谱最早由 Google 提出,利用知识图谱可以对搜索结果进行知识系统化处理,使关键词获得完整的知识体系。从本质上来看,知识图谱是一种语义网络,其结点代表实体或者概念,边代表实体与概念之间的各种语义关系。知识图谱的直接推动力来自于一系列实际应用,包括语义搜索、机器问答、情报检索、电子阅读、在线学习等[1]。

在知识图谱中,实体的分类信息描述了实体所属的概念域,对实体的解释和用户理解具有重要意义(如苹果属于水果,水果又属于植物)。同一分类下的实体具有领域相关性,如具有相似的属性。百度百科、互动百科等中文百科都已有分类专题,由人工编辑词条的类别。实体分类有利于知识结构化、层次化的展示。分类其实也是为实体标注上层概念的过程,对于构建中文知识图谱中的“is a”关系很有帮助。但由于目前中文百科实体分类仍需要人工编辑,并且主观因素影响较大,漏标和标错现象非常严重,因此迫切需要一种能够自动识别某一类别下实体的方法为中文知识图谱中的实体添加分类。

百度百科约有 550 万词条,其中人物分类下的

基金项目:国家自然科学基金(61472085);国家“973”计划项目(2015CB358800);上海市科技创新行动计划基础研究项目(15JC1400900);上海市青年科技启明星计划项目(13dz2260200)。

作者简介:李 薇(1991—),女,硕士研究生,主研方向为数据挖掘;肖仰华,副教授、博士、博士生导师;汪 卫,教授、博士、博士生导师。

收稿日期:2016-02-18 **修回日期**:2016-05-13 **E-mail**:13210240018@fudan.edu.cn

词条就有 479 018 个,占了总体的近 10% 。百度百科中出现的人物词条都可以认为是名人,有很强的社会影响力,也是搜索中出现的高频人物,挖掘出这些人物信息非常有价值。例如在分词任务中,对于“武汉/市长/江大桥”和“武汉市/长江/大桥”这两种分词歧义,如果有了名人知识图谱,将不再出现这样的歧义[2]。目前百度百科的分类体系还不健全,据统计 30% 的词条在百度百科里是没有分类信息的,对于这部分词条里的人物词条,就需要从其他数据源来推断,例如词条的属性、副标题等;此外,有些人物词条虽然没有人物类别的标注,但是有人物相关的其他分类,例如词条“赵薇”,她有演员、歌手等分类信息,但是没有人物的分类,那么首先要挖掘出人物大类的相关分类,例如演员、歌手等,然后从人物相关的分类、属性特征和副标题等角度推断出这个词条是人物。本文从已有的分类、属性和副标题 3 个部分入手分别拓展人物词条集合,再使用马尔科夫逻辑网络(Markov Logic Network,MLN)对上述3 类拓展后的证据进行联合推断,以优化整体的人物识别效果。

百度百科的原始数据由 3 个部分构成:目前词条下由人工编辑的分类信息,词条属性的结构化信息,重名多义词的副标题。本文将这 3 个方面作为数据源来判断词条是否在人物分类下,首先对 3 个部分的数据各自进行推断,然后使用 MLN 联合推断方法对其进行融合推断,以提升准确率和召回率。

1 相关工作

1.1 马尔科夫逻辑网络

面对海量的网络信息,如何处理信息的复杂性和不确定性问题是人工智能和数据集成的难点之一。为解决这类问题,统计关系学习(Statistical Relational Learning,SRL)方法和概率图模型(Probabilistic Graphical Model,PGM)被相继提出。SRL 通过逻辑表示、概率推理、机器学习和数据挖掘等方法获取关系数据中的似然模型,PGM 则将概率统计信息与数据的结构信息相结合,主要涵盖了贝叶斯网络(Bayesian Network,BN)、隐马尔科夫模型(Hidden Markov Model,HMM)、神经网络(Neural Network,NN)等方法[3]。人们迫切需要一种模型将逻辑表示方法和概率方法结合起来。

MLN 在 2004 年被美国华盛顿大学的 Domingos 和 Richardson 首次提出。该网络可以将统计关系学习和概率图进行较好的结合,比纯逻辑方法和纯概率方法能更好地解决上述问题。MLN 是一种统计关系学习框架,具有强大的描述能力、逻辑推理能力和处理不确定性的能力。从处理不确定性问题看,MLN 为一阶谓词附加权值,可容忍知识库中存在不完整和互相矛盾的知识,具有较好的处理不确定性问题的能力;从概率统计方面看,MLN 为描述 Markov 网络提供了简洁有效的方法。目前,MLN 已成为人工智能、数据集成和机器学习等领域的研究热点,具有广阔的应用前景[4]。

1.2 实体集拓展

实体集拓展(Entity Set Expansion,ESE)的目标是实现自动化地从文本或 Web 页面获取一个特定目标分类下的所有实体。例如给定首都的种子集合{罗马,北京,巴黎},一个实体集拓展系统应该从 Web 页面里抽取出所有的其他首都,例如莫斯科和伦敦。目前实体集拓展系统已经应用在很多领域中,例如字典构建、词义消歧和提问建议[5]。

近年来,实体集拓展系统在学术界和工业界得到了很大的关注(例如 Google Sets)。多数实体集拓展系统都采用 BootStrapping 的方法(例如 DIPRE,Snowball 等)。传统 BootStrapping 方法的主要缺点是拓展边界和语义偏移的问题[6]。目前有2 种策略来解决语义偏移的问题:1)基于排序的方法。该方法基于一个假设:排名高的实体更可能是目标分类下的实体,其通过一个排序算法来挑选出可信度较高的模式和实体;2)基于互斥限制的方法。该方法同时拓展多个分类,并且由预先给定的分类之间的互斥属性来决定拓展界限[7]。

鉴于百度百科数据的特征,本文借鉴上述 2 种策略并且引入马尔科夫逻辑网络的方法进行联合推断。

本文采用 MLN 联合推断方法来进行实体识别。为处理数据缺失的问题,首先从开放分类、属性和副标题 3 个方面分别拓展数据源,扩充 Predicate 数据源;然后通过马尔科夫逻辑网络的方法将三方面的数据源联合起来进行联合推断,从而提升推断的效率。如图 1 所示,通过分类方面的拓展将得到 Category 和 PeopleCategory 2 种 Predicate 数据,而通过属性的拓展将得到 Attribute 和 PeopleAttribute 数据。同时多义词的副标题数据源拓展将会提供 Related 信息,即 2 个词条是相互关联的。通过上述 Predicate 数据可以进行联合推断,判断出哪些词条是 People。

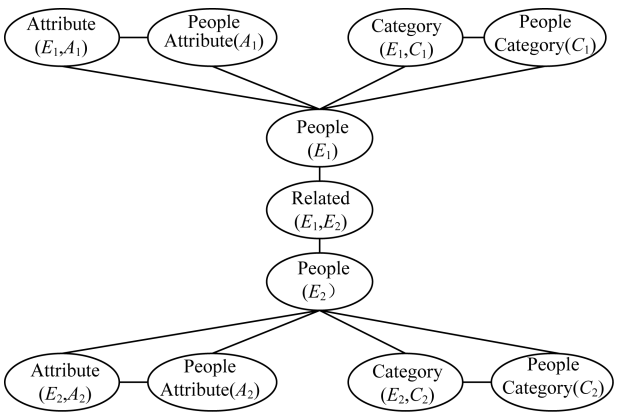

图 1　人物分类马尔科夫逻辑网络无向图模型

2　开放分类拓展

百度百科现有的人物分类下的开放分类共有 112 个,分为自然科学、人文领域、社会科学、文学领域、军事领域、艺术领域、体育领域、虚拟人物 8 个领域,其中各领域还有各自分类,例如自然科学领域又分为科学家、医学家、数学家、物理学家、化学家、农学家、农业学家、古生物学家和医师,如图 2 所示。

图 2　百度百科人物开放分类

2.1　基于 BootStrapping 的模式学习

BootStrapping 方法是一种弱(半)监督学习方法,其形式化描述如下:

给定标注数据集 L 和未标注数据集 U,不断重复如下过程:

1)使用标注数据集 L 训练分类器 h。

2)用分类器 h 对为标注数据集 U 进行标注分类。

3)在分类结果中选择可信度比较大的标注数据子集,进行如下操作:$L + U' \rightarrow L$,$U - U' \rightarrow U$。

4)如果满足终止条件,则退出 BootStrapping 过程;否则返回步骤 1)[8]。

人物分类拓展的 Bootstrapping 方法步骤如下:

1)选用人物分类,使用该分类下的112 个开放分类(如图 2 所示)以及所有包含“人物”字符串的分类,这些所有分类名称在做并集操作以后被作为种子集合 S。

2)根据 S,在表 1 所示的百度分类中查找出所有分类字段含有 S 中任一元素的词条。表 1 记录了百度百科中人工编辑的分类信息,其中实体由词条的标题和副标题共同表示(如果副标题存在的话),分类就是当前百度百科人工给该实体编辑的分类。对百度百科爬取的页面解析后共得到 12 324 262 条数据,这些词条就被加入到“人物”实体集 P。例如词条“约翰·洛克菲勒”的分类是“人物”,那么词条“约翰·洛克菲勒”就被加入到了人物实体集 P 中。又例如词条“王强_(中国大陆歌手)”的分类是“歌手”,“歌手”是种子集合 S 里的一个元素,因此,“王强_(中国大陆歌手)”也被加入到了“人物”实体集 P 中。

表 1　百度分类

实体(标题 + 副标题)	分类
约翰·洛克菲勒	人物
王强_(中国大陆歌手)	歌手
王强_(中国大陆歌手)	明星

3)对于人物实体集 P 中的每一个元素,将其与表 1 中的的实体字段作匹配,提取出该条数据的分类字段,若分类字段不在 S 中,则加入到候选分类集合 W 中。例如“王强_(中国大陆歌手)”词条分类为“明星”,而“明星”不在集合 S 中,那么“明星”就被加入到候选集 W 里。

4)对于 S 中任一元素 s 和 W 中的任一元素 w,若存在 P 中的一元素 p,(p,s) 和 (p,w) 均存在于表 1的分类字段中,则可计算出相似度 $Similarity(s,w)$(相似性计算见 2.2 节)。

5)如果 $Similarity(s,w) > \theta$(θ 为设定的一个阈值),那么认为 w 也是描述人物的分类,将 w 加入到集合 Y 中。计算 $|Y|$,如果 $|Y| = 0$,则满足终止条件,退出 BootStrapping 过程,否则 $S + Y \rightarrow S$,$P = \phi$,$W = \phi$,$Y = \phi$,返回步骤 1)。

2.2　Category 之间的 Jaccard 相似度

在 2.1 节中提到计算 $Similarity(s,w)$,即计算2 个 Category 之间的相似性。本文使用 Jaccard 相似度来衡量 2 个 Category 的相似度,定义如下:对于2 个Category 数据 s 和 w,在表 1 中分类字段为 s 的数据对应的实体字段的集合为 SP,同理得到 WP,那么:

$$Similarity(s,w) = \frac{SP \cap WP}{SP \cup WP}$$

但是,在实际计算 Jaccard 相似度的过程中存在以下问题:首先,计算次数太多,需要算相似度的 Pair 达到了十万级,每对 Pair 都要进行一个交和并运算,计算量非常大。另外,由于百度百科有 550 万

词条,计算相似度时维度可能会非常大,并且 Boot-Strapping 还需要迭代多次,因此,十分需要一种降维、提高计算效率的算法。

2.3 基于 MinHash 的 Jaccard 相似度计算

2 个集合经随机排列转换之后得到的 2 个最小哈希值相等的概率等于这 2 个集合的 Jaccard 相似度,但对大规模特征矩阵进行显式排列转换是不可行的,因为其计算量过大,耗时较多。然而可以通过一个随机哈希函数来模拟随机排列转换的效果,该函数将行号映射到与行数大致相等的桶中。因此,可以不对行选择 n 个随机排列转换,取而代之的是随机选择 n 个哈希函数 $h_1,h_2,\cdots,h_n$ 作用于行。在上述处理的基础上,就可以根据每行在哈希之后的位置来构建签名矩阵。首先令 $SIG(i,c)$ 都初始化为 ∞。然后对行 r 进行如下处理:计算 $h_1(r),h_2(r),\cdots,h_n(r)$。对每列 c 进行如下操作:如果 c 在第 r 行为 0,则什么都不做;否则,如果 c 在第 r 行为 1,那么对于每个 $i=1,2,\cdots,n$,将 $SIG(i,c)$ 置为原来 $SIG(i,c)$ 和 $h_i(r)$ 之中的较小值。最后,2 个集合的 Jaccard 相似度就是 $h(S_1)=h(S_2)$ 的概率。

3 实体属性拓展

3.1 相似实体排序模型

如图 3 所示,百度百科中人物与非人物在 infobox 中的属性有着很大差异,为区别描述任务的和描述非人物的属性,将人物、人物下的开放分类、包含人物字符串分类的词条都确定为人物实体,作为正样本 P。百度百科已有的分类体系中已分为了 12 个领域,如图 4所示,负样本则是从除人物外的11 个分类中均匀抽样,构成非人物的集合 N。然后统计这些 P 中元素的属性和 N 中元素的属性,通过比对这两组属性的集合,发现人物与非人物的属性有着非常大的差异。人物识别本质就是一个二元分类问题,是否拥有某条属性则是进行分类的重要特征。

中文名:	王强	出生日期:	1974年5月12日
外文名:	Johnston	职业:	歌手
国籍:	中国	毕业院校:	湖北音乐学院
民族:	汉	代表作品:	《秋天不回来》《不想让你哭》等
出生地:	湖北省宜昌市	主要成就:	05届中国移动彩铃原创歌曲奖

图 3 百度百科信息

分类导航

人物	动漫人物 | 歌手 | 运动员 | 古代嫔妃 | 演员 >>	文化	考古 | 网络用语 | 世界名著 | 诗词 | 神话 >>
技术	土木工程 | 移动通信 | CPU | MP3 | 电脑病毒 >>	历史	侏罗纪 | 清朝 | 明朝 | 洋务运动 | 先秦 >>
艺术	纪念碑 | 戏剧 | 音乐 | 绘画 | 雕塑 | 建筑 >>	生活	影视 | 动漫 | 游戏 | 服饰 | 美容 | 烹饪 >>
地理	大陆架 | 国家 | 地质 | 岛屿 | 山脉 | 河流 >>	社会	外交 | 军事 | 民俗 | 交通 | 法律 | 企业 >>
体育	冰雪运动 | 极限运动 | 电子竞技 | 篮球 | 足球 >>	自然	地震 | 气象 | 天文 | 花卉 | 恐龙 | 细菌 >>
科学	生物 | 遗传学 | 医学 | 化学 | 物理 | 数学 >>	经济	投资 | 保险 | 银行 | 期货 | 基金 | 股票 >>

图 4 百度百科已有分类体系

本文统计出了人物与非人物的属性以及它们出现的频率,其中描述人物的属性共有6 258 个,描述非人物的属性则有 10 670 个。将这些属性都作为特征来对词条分类显然是不可行的:1)经过统计发现许多属性都只在一个词条中出现过,如果使用这些属性作为特征的话将得到一个稀疏矩阵,既浪费了空间,又对分类的帮助不大;2)百度百科有 550 万词条,每个词条用一万多特征来分类的话,计算量过大;3)特征信息存在很大的冗余,如“出生日期”和“星座”,由“出生日期”可以推断出“星座”,同时保留这两个特征就造成了信息的冗余。由此看来,对特征的降维十分必要并且切实可行。那么特征的选择就是下一步要解决的问题。

3.2 特征选择

特征选择[9]过程一般由产生过程、评价函数、停止准则、验证过程 4 个部分组成。本文采用基于关联规则的特征选则 + 最优优化搜索的方法来进行特征选择[10]。产生过程的主要任务是搜索特征子集,提供特征子集给评价函数。本文的产生过程采用最优优化搜索(Best First Search,BFS)算法,这是一种完全搜索方法,算法描述如下:首先选择 N 个得分最高的特征作为特征子集,将这 N 个特征加入到一个长度无限的优先队列中,每次从队列中拿出得分最高的子集,然后穷举向该子集加入一个特征后产生的所有特征集,将这些特征集加入队列[11]。

本文的评价函数采用基于关联规则的特征选则算法(Correlation-based Feature Selection,CFS),这是一种经典的过滤器模式的特征选择[12]方法。特征子集的质量高低是用相关性来衡量的,这种衡量方法的前提假设是:好的特征子集的特点是其所包含的特征与分类的相关度较高,与此同时特征之间的相关度较低(也表示冗余度低)。

本文调用 Weka 中特征选择的相关接口,以 BFS 算法搜索特征子集,用 CFS 来作为评价函数,对人物识别的特征进行降维,最后在近 16 000 个 Feature 中选择 64 个特征来对词条进行分类[13]。

3.3 分类器的选择使用

本文采用了逻辑回归和支持向量机这两种分类器分别对同样的样本进行分类,并对这两种分类器的多个衡量指标进行对比[14]。

本文的特征值都是 0,1,用 0 代表该词条没有此属性,1 代表该词条有此属性,所以,因变量都是离散的,比较适合采用逻辑回归或者支持向量机这两种分类器。本文采用 k-折交叉验证的方法来进行性能评估。

4　实体副标题的确定

4.1　启发式方法

启发式方法是一种根据经验规则进行发现的方法,其特点是在解决问题时,利用过去的经验,选择已经行之有效的方法,而不是系统地、以确定的步骤去寻找答案[15]。本文通过对人物词条副标题的分析发现,绝大多数副标题都是用人物的职位、职称、职业来概括人物。如图 5 所示,联想到在人物的 infobox 中经常出现职业、职称、职务的属性,并且描述人物的分类也多是关于人物的职业方面的描述,如“歌手”、“演员”等。所以,本文采用启发式的规则来识别人物,用副标题中是否含有人物职称、职务、职业信息来判断此词条是否为人物。首先挖掘人物的职业、职称、职务信息,从百度 infobox 中抽取属性是职务、职业、职称的值,得到集合 M,然后再获取描述人物的分类集合 N,$T=MN$,T 是关于人物职务、职业、职称的描述集合。

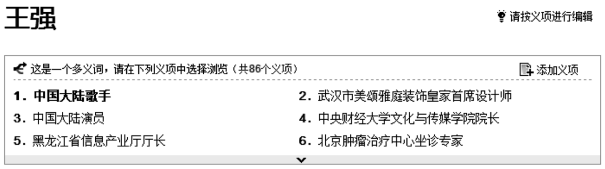

图 5　百度百科副标题示例

4.2　匹配方式

首先利用 IKAnalyzer 分词工具对每个词条的副标题进行分词,然后使用从后向前叠加的方法与职业、职务、职位信息进行匹配。如副标题“中国著名男歌手”,首先进行分词,得到“中国”、“著名”、“男”、“歌手”以及他们的词性。接下来查看“歌手”是否在集合 T 中,若存在则断定该词条是人物,结束匹配。否则查看“男歌手”(“男”+“歌手”)是否在集合 T 中,同理依次匹配“著名男歌手”、“中国著名男歌手”。

中文相较于英文有一个很大的特点是中心词在后面,即尾重原理。百度百科的副标题也符合这个特点,副标题的中心词往往在后面,如“中国著名男歌手”,中心词是“歌手”,前面都是对“歌手”的修饰。所以,在做职称、职务、职业信息匹配时应从后向前匹配,这样可以提高匹配的效率。另一方面,叠加匹配是为了提高人物识别的准确率。如图 6所示,电影“致青春”副标题是“赵薇导演电影”,分词后得到“赵薇”、“导演”、“电影”,从后向前匹配则依次匹配“电影”、“导演电影”、“赵薇导演电影”。这三者均不是在集合 T 中,所以判断该词条不是人物,而如果不是使用叠加匹配的方式,仅仅是对分词后得到的每个词匹配的话,“导演”在集合 T 中,“致青春”将会被判定是人物,产生了错误,降低了人物识别的准确率。

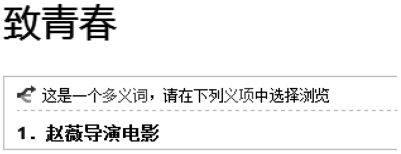

图 6　“致青春”副标题

5　基于马尔科夫逻辑网络的联合推断

上面已经从分类、属性和副标题 3 个方面分别做了实体集拓展,但是 3 种方法是做独立判断的,相互之间并没有关联,但实际上,3 种方法可以相互补充,互为推断条件,因此,本文引入 MLN 方法来进行联合推断,提升实验效果。

5.1　马尔科夫逻辑网络

Markov 逻辑网络 L 是一组二元项(F,w),F 表示一阶逻辑规则,而 w 是一个实数[2]。再和一些常量一起,就定义了 MLN:在网络中的每个谓词的每个实例化作为一个节点,在 MLN 中的每个一阶逻辑规则 F 作为一个特征,并且权重为 w。

一个简单的马尔科夫逻辑网络实例如表 2 所示。

表 2　马尔科夫逻辑网络实例

命题	一阶逻辑规则	权重
酗酒导致高血压	$F_1:\forall x, Dr(x)\Rightarrow Hy(x)$	1.2
如果两人是朋友,他们可能都酗酒,可能都不酗酒	$F_2:\forall x,\forall y, Fr(x,y)\Rightarrow(Dr(x)\Leftrightarrow Dr(y))$	0.8

5.2　人物分类马尔科夫逻辑网络

本文从分类、属性和副标题 3 个方面得到证据谓词,具体解释如表 3 所示。根据已有的证据谓词,本文定义如表 4 所示的一阶逻辑规则。

表 3　证据谓词解释

证据谓词	解释
$att(e,a)$	实体 e 有属性 a
$cat(e,c)$	实体 e 属于分类 c
$related(x,y)$	实体 x 和 y 是同义词
$peopleAttribute(a)$	a 是描述人物的属性
$peopleCategory(c)$	c 是描述人物的分类

表 4 一阶逻辑规则

命题	一阶逻辑规则
实体有表述人物的属性,则该实体是人物	$F_1:\forall e,\forall a,att(e,a)\ \hat{}\ peopleAttribute(a)\Rightarrow people(e)$
人物实体的属性是表述人物的属性	$F_2:\forall e,\forall a,People(e)\hat{}att(e,a)\Rightarrow peopleAttribute(a)$
实体属于人物相关的分类,那么该实体是人物	$F_3:\forall e,\forall c,cat(e,c)\ \hat{}\ peopleCategory(c)\Rightarrow people(e)$
人物属于某分类,则该分类是人物相关的分类	$F_4:\forall e,\forall c,people(e)\hat{}cat(e,c)\Rightarrow peopleCategory(c)$
实体 x 和实体 y 是多义词下的两个词条,则一个实体是人物,另一个实体也是人物	$F_5:\forall x,\forall y,related(x,y)\Rightarrow(people(x)\Leftrightarrow people(y))$
多义词实体的反身性	$F_6:\forall x,\forall y,related(x,y)\Leftrightarrow related(y,x)$

本文使用了华盛顿大学提供的 alchemy 工具,在提供了证明谓词和相关一阶逻辑规则后,MLN 学习分为结构学习和参数学习[16]。参数学习是指在 MLN 结构确定的前提下,进一步学习和优化模型的参数。

6 实验结果与分析

6.1 实验数据集

百度百科的 url 基本格式如下:http://baike.baidu.com/view/298381.htm,最后的数字是页面 ID。百度百科页面 ID 是连续的,这为网页爬虫的实现带来了极大的便利,只需要不断递增 URL 的 ID 项即可爬取百度百科中的所有页面。

爬虫仅仅是爬取了所需要的整个页面,还需要一个解析的过程来抽取需要的信息。经过上文对百度百科人物特征的分析,解析出如表 5 所示的内容。该表记录了百度百科中存在 infobox 的词条的属性信息。其中,实体同表 1 中的定义,属性就是 infobox 中出现在":"前的内容,值就是对应属性的值。如歌手王强的出生地是湖北省宜昌市,如表中第2 行所示存储。解析爬取的页面共得到 2 007 399 条这样的数据。

表 5 百度信息表

实体(标题 + 副标题)	属性	值
邓肯·弗格森	专业特点	前锋
王强_(中国大陆歌手)	出生地	湖北省宜昌市
电车	发明日期	1881 年

表 6 记录了百度百科中出现的多义词。其中,第 1 列是词条的标题,如人名王杨,而第 2 列副标题就是多义词里对该词条的进一步的描述,共得到了 291 411 条数据。

表 6 百度副标题

标题	副标题
王杨	王杨_(安徽师范大学教师)
假面	假面_(瑞典 1966 年英格玛·伯格曼执导电影)
假面	假面_(盗墓笔记》广播剧片尾曲《假面)

表 7 记录了百度百科中出现的同义词,示例如图 7 所示,当在百度百科中搜索陈平,会显示"陈平"与"三毛"是同义词。在表中一行代表一组同义词,从百度百科中共解析出 4 453 条类似的数据。

表 7 百度同义词

实体 1	实体 2
月亮_(中国大陆演员)	许十云
爱情_(韩国 1998 年发行电视剧)	我怕恋爱
基德_(怪盗基德)	怪盗基德

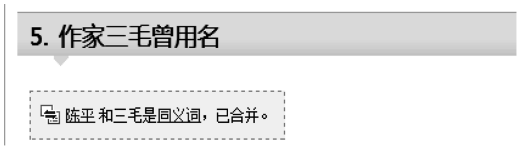

图 7 同义词示例

6.2 实验方法及结果

6.2.1 分类拓展实验

统计人物、人物开放分类、含有人物字段的分类共 12 287 个。根据这些分类在 Baiducategory 中一共找到 639 702 词条,确认是人物的集合 S。再在 Baiducategory 中查找这些词条的 Category,共有 92 385 个。除去原来的 12 287 个 Category,剩下 80 104 个候选的 Category,并将这些 Category 按在已有人物词条 S 中出现的次数排序。

Precision@ k 是排序的前 k 个样本的精确度,即前 k 个样本里是正样本的个数除以 k。本文人工标注了前 500 个 Category,计算 *Precision@ k*。实验结果如图 8 所示。

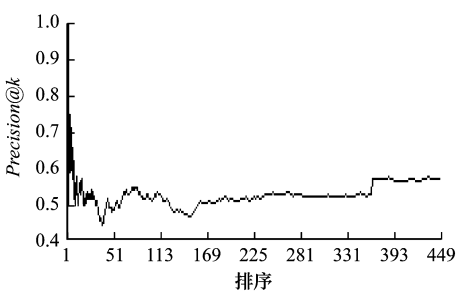

图 8 *Precision@ k* 曲线

可以看出,前 500 位的 *Precision@ k* 并没有明显递减的趋势,由此可知仅采用 Category 在已有人物集合 S 中出现的次数作为评判 Category 是否与人物相关是不可取的。

6.2.2　属性拓展实验

根据已确定人物集合 S,在 Baiduinfobox 表格中共找到 6 258 个属性来描述这些人物,并且统计这些属性在人物词条中出现的频率。再在整个 Baiduinfobox 中统计共有 24 735 个属性,可以看到有近两万的属性只用来描述非人物,由此看来人物与非人物的属性方面存在很大的差异。

1)人物属性和整个百度百科属性分布特征的比较。

首先根据属性在人物中出现的频率统计在人物词条中出现 N 次的属性的个数(N 从 1 到属性出现的最大次数)。然后计算出现 N 次的属性个数在所有描述人物属性所占的比例。最后计算从出现 1 次到 N 次的属性所占比例的累积和。同理,在整个百度百科的属性中计算累积和。

属性分布特征如图 9 所示,其中,横坐标是属性出现的频率,纵坐标是出现了 N 次的属性在所有属性中所占比例的累积和。实线是百度百科所有属性的分布特征,虚线是人物的属性的分布特征。从图中发现百度百科的属性和人物属性的曲线几乎重叠,说明两者符合一致的分布特征,也说明仅通过属性的频率分布特征难以得到人物与非人物分类。通过此图同时可以看出大部分属性只出现了少于 20 次,只有少数的属性出现的次数较多,这充分地说明了特征的选择非常重要,如果以每个属性的有无作为一个特征,那么将得到一个十分稀疏的矩阵,既会降低效率,又对分类模型的建立没有什么帮助。

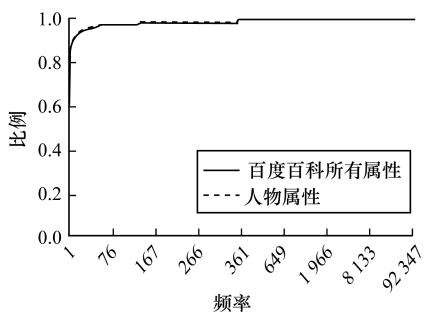

图 9　属性分布特征

2)逻辑回归和支持向量机分类器的建立。

在上文所述的 S 集合中,查找在 Baiduinfobox 中出现过的词条作为正样本。在百度百科除人物外的 11 个类别中均匀取样,将在 Baiduinfobox 中出现过的词条作为负样本,共得到 134 714 个负样本、86 497 个正样本,共 221 211 个样本。

根据这 221 211 个样本,在 Baiduinfobox 中共找到 12 180 个属性。以某词条是否有此属性则可以得到 12 180 个取值为 0 或者 1 的特征。在上文已经分析过,此处需要降维,进行特征选择。

本文采用 CFS + BFS 方法进行特征选择,最终得到了 64 个特征。下面分别采用逻辑回归和支持向量机 2 种分类器建立分类模型,并用十折交叉检验的方法来评估分类器的性能,结果如表 8 所示,可以看出,SVM 的效果明显优于逻辑回归方法,无论准确率还是召回率都比逻辑回归要高。

表 8　3 种方法的效果比较

方法	准确率	召回率	F-score 值
逻辑回归	0.847	0.847	0.847
支持向量机	0.901	0.885	0.892
MLN	0.915	0.893	0.903

6.2.3　马尔科夫逻辑网络实验

本文使用了华盛顿大学提供的马尔科夫逻辑网络工具 alchemy,数据集与 6.2.2 节使用的相同,在与传统的逻辑回归和支持向量机分类器比较后,发现马尔科夫逻辑网络无论是准确率还是召回率都比传统方法效果有所提升。

7　结束语

分类是构建中文知识图谱中的一个重要环节,实体的分类信息描述了实体所属的概念域,对实体的解释和用户理解具有重要意义。而当前各种中文百科均是通过人工编辑的方式来为实体分类,漏标和标错现象非常严重,这给中文知识图谱的构建带来了问题。因此,本文提出一种领域识别的方法为词条自动添加分类。以人物领域为例,识别出该分类下的实体,并为这些实体添加人物分类。这种自动添加分类的方法也可以推广到其他的分类中去。本文使用了联合推断的方法——马尔科夫逻辑网络(MLN),在前期进行了大量的数据拓展,给出合理的一阶逻辑规则。实验结果表明,与逻辑回归和支持向量机方法相比,本文方法可有效提升分类效果。

参考文献

[1] 郭云峰,韩　龙,皮立华,等. 知识图谱在大数据中的应用[J]. 电信技术,2015(6):25-29.

[2] 孙茂松,左正平,黄昌宁. 汉语自动分词词典机制的实验研究[J]. 中文信息学报,2000,14(1):1-6.

[3] 徐从富,郝春亮,苏保君,等. 马尔可夫逻辑网络研究[J]. 软件学报,2011,22(8):1699-1713.

[4] 楼俊杰,徐从富,郝春亮. 基于马尔科夫逻辑网络的实体解析改进算法[J]. 计算机科学,2010,37(8):243-247.

(下转第 240 页)

[5] Corne D W, Jerram N R, Knowles J D, et al. PESA-II: Region-based Selection in Evolutionary Multiobjective Optimization [C]//Proceedings of Genetic and Evolutionary Computation Conference. Washington D. C., USA: IEEE Press, 2002: 283-290.
[6] Coello C A C, Lechuga M S. MOPSO: A Proposal for Multiple Objective Particle Swarm Optimization [C]// Proceedings of Congress on Evolutionary Computation. Washington D. C., USA: IEEE Press, 2002: 1051-1056.
[7] Zhang Qingfu, Li Hui. MOEA/D: A Multi-objective Evolutionary Algorithm Based on Decomposition [J]. IEEE Transactions on Evolutionary Computation, 2007, 11(6): 712-731.
[8] Li Hui, Zhang Qingfu. Multiobjective Optimization Problems with Complicated Pareto Sets, MOEA/D and NSGA-II [J]. IEEE Transactions on Evolutionary Computation, 2009, 13(2): 284-302.
[9] Zhang Qingfu, Liu Wudong, Li Hui. The Performance of a New Version of MOEA/D on CEC'09 Unconstrained MOP Test Instances [C]//Proceedings of Congress on Evolutionary Computation. Washington D. C., USA: IEEE Press, 2009: 203-208.
[10] Li Ke, Fialho A, Kwong S, et al. Adaptive Operator Selection with Bandits for Multi-objective Evolutionary Algorithm Based Decomposition [J]. IEEE Transactions on Evolutionary Computation, 2014, 18(1): 114-130.
[11] Coello C A C, Veldhuizen D A V, Lamont G B. Evolutionary Algorithms for Solving Multi-objective Problems [M]. Boston, USA: Kluwer Academic Publishers, 2002.
[12] Chiang T C, Lai Y P. MOEA/D-AMS: Improving MOEA/D by an Adaptive Mating Selection Mechanism [C]// Proceedings of IEEE Congress on Evolutionary Computation. Washington D. C., USA: IEEE Press, 2011: 1473-1480.
[13] Zhao Shizheng, Suganthan P N, Zhang Qingfu. Decomposition Based Multiobjective Evolutionary Algorithm with an Ensemble of Neighborhood Sizes [J]. IEEE Transactions on Evolutionary Computation, 2012, 16(3): 442-446.
[14] Li Ke, Zhang Qingfu, Kwong S, et al. Stable Matching Based Selection in Evolutionary Multiobjective Optimization [J]. IEEE Transactions on Evolutionary Computation, 2013, 18(6): 909-923.
[15] Konstantinidis A, Charalambous C, Zhou Aimin, et al. Multi-objective Mobile Agent-based Sensor [C]// Proceedings of IEEE Congress on Evolutionary Computation. Washington D. C., USA: IEEE Press, 2010: 1-8.
[16] Yazdi J. Decomposition Based Multi Objective Evolutionary Algorithms for Design of Large-scale Water Distribution Networks [J]. Water Resources Management, 2016, 30(8): 2749-2766.
[17] 吕铭晟,沈洪远,李志高,等. 多变异策略差分进化算法的研究与应用[J]. 计算机工程, 2014, 40(12): 146-150.
[18] Ishibuchi H, Akedo N, Nojima Y. Relation Between Neighborhood Size and MOEA/D Performance on Many-objective Problems [C]//Proceedings of the 7th International Conference on Evolutionary Multi-criterion Optimization. Berlin, Germany: Springer, 2013: 459-474.
[19] Ishibuchi H, Narukawa K. An Empirical Study on Similarity-based Mating for Evolutionary Multiobjective Combinatorial Optimization [J]. European Journal of Operational Research, 2008, 188(1): 57-75.

编辑 顾逸斐

(上接第 231 页)

[5] Shami M, Do T, Wright K, et al. Entity Expansion and Grouping: US 8312018 B2[P]. 2012.
[6] 段宇锋,朱雯晶,陈 巧,等. 朴素贝叶斯算法与Bootstrapping方法相结合的中文物种描述文本语义标注研究[J]. 现代图书情报技术, 2014(5): 83-89.
[7] 罗 军,高 琦,王 翊. 基于Bootstrapping的本体标注方法[D]. 重庆: 重庆大学, 2010.
[8] 赵江江,秦 兵. 基于BootStrapping的中文事件元素抽取系统设计与实现[J]. 智能计算机与应用, 2012, 2(1): 16-17.
[9] 邓 琦,苏一丹,曹 波,等. 中文文本体裁分类中特征选择的研究[J]. 计算机工程, 2008, 34(23): 89-91.
[10] 朱 明,王 军,王俊普. Web网页识别中的特征选择问题研究[J]. 计算机工程, 2000, 26(8): 35-37.
[11] 李 宁. 基于流统计特性的应用协议识别技术研究[D]. 南京: 南京邮电大学, 2013.
[12] Hall M A. Correlation-based Feature Selection for Machine Learning[D]. Hamilton, New Zealand: The University of Waikato, 1999.
[13] Bouckaert R R, Frank E, Hall M, et al. WEKA——Experiences with a Java Open-source Project [J]. Journal of Machine Learning Research, 2010, 11(5): 2533-2541.
[14] Koh K, Kim S J, Boyd S. An Interior-point Method for Large -scale ℓ_1-Regularized Logistic Regression [J]. Journal of Machine Learning Research, 2007, 8(3): 1519-1555.
[15] 白思俊. 资源有限网络计划启发式方法的评价(中): 启发式方法的综合比较和评价[J]. 运筹与管理, 1999(3): 51-56.
[16] Sumner M, Domingos P. The Alchemy Tutorial [D]. Washington D. C., USA: University of Washington, 2010.

编辑 金胡考