中文专利发明人重名消解问题研究*

邢晓昭 郑彦宁
（中国科学技术信息研究所，北京 100038）

摘要：专利发明人分析为技术人才评价和科研团队识别提供有力的数据支撑。然而，中文姓名存在大量重名现象，使得基于发明人的研究结果出现偏差。本文提出一种基于规则的中文专利发明人重名消解方法。针对专利申请人因为并购、拆分、重组或战略转型等原因造成的名称不一致情况，采用基于向量空间模型的余弦相似度算法进行识别；针对因门牌号书写不规范而造成的地址不一致情况，采用基于邮编和门牌地址的分级匹配算法进行识别；合作者相似度采用Jaccard系数计算。以中国科学技术信息研究所《电动汽车专题数据库》为例，验证该方法的科学性和有效性。

关键词：重名消解；中文专利；发明人；相似度；向量空间模型
中图分类号：G350；TP391
DOI：10.3772/j.issn.1673-2286.2018.10.001

专利文献是一种具有丰富技术内涵、规范化、可公开获得的技术资料[1]。发明人信息是专利文献信息的重要组成部分，专利发明人是对发明创造的实质特点做出创造性贡献的人[2]。通过发明人专利数量、引文数量和网络中心度等指标，可以对其技术竞争力进行比较和排名，识别核心技术人才；通过对发明人参与发明的时序分析，可了解发明人技术生涯，识别其创新轨迹；通过对发明人合著情况的分析，可以探索技术创新合作模式，发现科研团队。

然而，中文专利发明人姓名中存在大量重名现象。以姓名“李军”为例，在国家知识产权局网站中以“李军”为发明人共检索出专利文献记录21 410条（检索时间：2018年10月8日），分布在生物医药、机械制造、农业、化工、电子电器等多个领域。即使在同一科研机构中重名的情况也非常普遍，仅重庆大学就有7位名为“张伟”的研究人员。发明人姓名重名使得在专利数据库中查询或关联某个发明人的专利时，往往会将所有同名发明人的专利返回，或将某个发明人与其他发明人的专利相链接。如果不解决专利发明人姓名歧义问题，无论是进行技术人才评价、人才成长路径分析，还是开展科研团队识别，都会对结果造成干扰。因此，重名消解研究是建立高质量中文专利数据集，并进行精准团队识别的前提。

1 相关研究

重名消解本质上要解决的是姓名歧义问题。国内外学者针对专利发明人的姓名消歧进行了诸多探索，并积累了许多匹配算法。这些算法主要集中于解决两个问题。一是应该采用何种方式、选择哪些属性计算两组专利数据之间的相似度。字符串匹配与名称编码是标准化发明人姓名的2种主要途径。而专利申请人、分类号、地址信息和合作网络可以作为除姓名外，进行发明人记录对匹配的附加属性。Singh[3]对发明人姓、名全称，以及中间名首字母进行精确匹配，并采用专利分类号进行补充匹配来判定不同的记录是否只指向同一位专利发明人。Miguélez等[4]采用Soundex编码系统对发明人姓名进行重新编码，以降低因匹配不足到造成的“第一类错误”。王道仁等[5]采用汉明距离算法、Jaro-Wrinkler算法和基于q-gram的算法等10种常用字符串匹配算法对发明人的姓名字符串进行模糊匹配，结果显示，Jaro-Winkler算法对于指向同一发明人不同姓名字符串的识

*本研究得到中国科学技术信息研究所创新研究基金青年项目“基于社会网络分析的科研团队识别关键技术研究”（编号：QN2018-01）资助。

别效果最好，而基于q-gram的杰卡德算法对于发明人的整体消歧效果最佳。刘斌等[6]针对中文姓名中可能产生的同音字和形近字歧义，根据汉语拼音和汉字四角码设计了一套专利数据中发明家姓名的消歧算法。二是基于相似度，如何尽可能高效、准确地合并同一发明人，区分不同发明人。常用方法包括基于规则的方法、基于无监督的机器学习方法和基于有监督的机器学习方法。Han等[7]采用无监督的K-way谱聚类算法对引文中的作者姓名进行消歧，并借助Laplacian矩阵对合作者属性、论文标题属性及出版地点属性的相似度进行处理以实现图的划分。在有监督机器学习方面，Li等[8]基于朴素贝叶斯算法，Ventura等[9]使用基于随机森林的条件森林算法进行专利发明人的姓名消歧。朱云霞[10]提出一种基于规则和相似度的重名消解框架，并对两个分解器和一个合并器的规则进行详细描述。

已有研究的相似度匹配主要针对姓名字段展开。然而，中文姓名相对简短，且形式固定，出现错误的概率较小。而与发明人相关的申请人及其地址字段相对较长、形式自由，所以变体较多。因此本文针对发明人姓名字段采用精确匹配算法，针对申请人及其地址字段采用模糊匹配算法。此外，用于姓名消歧的相似度算法，无论是基于字符串匹配还是名称编码，都将字符串看作独立字符的集合，没有考虑字符之间的语义联系。考虑到申请人字段自然语言特点明显，且体系庞杂，对此采用基于向量空间模型（Vector Space Model，VSM）的余弦相似度进行匹配。而申请人地址字段中包含邮政编码信息，对申请人所处的实际地理位置进行天然区分，对此采用基于地址和邮编的分级匹配。本文在整体流程上选用基于规则的判别方法。

2 研究方法

2.1 相似度算法

中文姓名只有单一的表述方式（即姓在前，名在后，且没有缩写形式），因而在中文专利发明人重名消解问题中异名同指的情况并不多见。本文首先将姓名完全相同的发明人归为一组，然后在组内进行相似度比较。相似度计算包括专利申请人相似度、地址相似度和合作者相似度3个部分。根据应用目标对于重名消解结果要求的严格或宽松，设置相应的阈值，并选择三者之间的与或逻辑运算。

2.1.1 专利申请人相似度计算

专利申请人是提交专利申请的自然人或法人，一般以机构或者企业居多。与自然人姓名相比，机构名称容易发生变更，如“东风汽车公司”在2017年更名为“东风汽车集团有限公司”；在包含特定关键词的机构名称之间很可能具有隶属、同源、衍变等隐含关系，如“中国第一汽车股份有限公司”是“中国第一汽车集团公司”的控股子公司，“西安交通大学苏州研究院”是由“西安交通大学”与苏州市人民政府共建的事业单位。识别这些关系对于准确判断专利申请人相似度十分关键。如果采用精确匹配的方法，在同一机构或相关机构的名称表述不完全一致的情况下会低估相似度；如果采用模糊匹配方法，仅考虑字符串在形式上的相似性，不考虑字符间语义关联及重要性，则会产生一定误差。如“哈尔滨工业大学”和“哈尔滨工程大学”，基于编辑距离的相似度为85.7%，基于最大公共子串的相似度为57.1%，而二者实际并没有关联。机构名称的另一个特点是长度偏长，而且具有一定的自然语言特点。机构名称一般包含地点、品牌、领域、机构性质等信息要素中的几项或所有项。如果能根据自然语言的特点将这些信息要素分割，并且按照其标识度赋予权重再进行比对，能够减少以上误差。

2.1.1.1 基于VSM的文本表示

VSM由Salton[11]于1975年提出，其基本思想是将文本表示成一个向量，向量的每一个维度表示文本的一个特征，该特征通常是一个字或词。使用VSM进行中文文本表示，需要经过分词、停用词处理及权重计算等步骤[12]。然后，文本集D中的任一文本d_j都可以表示成形如（W_{1j}, W_{2j}, …, W_{ij}, …, W_{nj}）的向量，其中，W_{ij}表示文本d_j中第i个特征（词）的权重。权重计算的方法主要有TFIDF函数、布尔函数、频度函数等。其中使用较多的是由Salton等[13]提出的TFIDF函数。

借助python语言的sklearn库将专利申请人文本转换为向量空间表示。sklearn是python的重要机器学习库，其中封装了大量的机器学习算法，如分类、回归、降维及聚类；还包含监督学习、非监督学习、数据变换三大模块。sklearn拥有完善的文档，使得它具有容易上手的优势。本文选择数据变换模块中的特征抽取模型sklearn.feature_selection，将每一个专利申请人转化为

对应n维特征词空间的权重向量，所有专利申请人的集合就构成一个权重矩阵。具体操作步骤如下。

（1）分词。调用结巴中文分词组件的程序接口实现专利权人分词。现有的中文分词方法可分为三大类，即基于词典的方法、基于理解的方法和基于统计的方法。本文选择基于统计的方法，即根据大量已经分词的文本，利用统计机器学习模型进行训练，从而实现对未知文本的切分。基于人民日报语料进行训练，形成2万多词条。将词条用前缀树表示，以提高查找效率。通过扫描前缀树，将待分词的专利权人字串表示为所有可能成词情况的有向无环图，采用动态规划查找最大概率路径的方式选择切词组合。此外，对于未登陆词，结巴中文分词组件提供基于Viterbi算法的新词发现HMM模型，本文选用此模型。

（2）停用词处理。本文中停用词表由两部分组成。第一部分来自传统停用词表，对专利权人字段特征的考察发现，该字段绝大部分由实词构成，除此以外只包含标点符号和少量的连词。如“台达电子工业股份有限公司；台达电子企业管理（上海）有限公司”中的分号、左括号和右括号，“罗伯特·博世有限公司”中的间隔号，以及“中国科学院上海微系统与信息技术研究所”中的连词“与”。而传统中文停用词表中包含有些停用词在专利权人字段中可能是具有实际意义的专有名词，如“江苏高和机电制造有限公司”中的“和”，以及“广东易事特电源股份有限公司”中的“特”，这部分包括传统停用词表中的所有符号和未同时作为品牌标识的连词。

除此之外，有些实词会在很多专利权人名称中出现。如“有限公司”“公司”“股份”等，这些词区分度不强，反而会降低申请人之间相关关系的识别效果。因此，停用词表的第二部分由词频排在前n位（如n=10），且未包含地点、品牌、领域信息的实词组成。

（3）权重计算。TF-IDF（Term Frequency-Inverse Document Frequency）是一种用于信息检索与数据挖掘的常用加权技术，其主要思想：如果某个词或短语在当前文档中出现的频率高，并且在其他文档中很少出现，则认为此词或者短语对于当前文档具有很好的标识度，适合作为当前文档的特征词[14]。本文采用TF-IDF方法计算专利权人字段中各词语的权重（见表1）。

表1 TF-IDF算法描述

输入：	分词后的文档集	
算法：	（1）获得特征词列表*termlist*	
	（2）计算*TF*	$tf(t_i, d_j)$ 表示在专利申请人文本d_j中词t_i出现的频率
	（3）计算*IDF*	$idf(t_i)=\log\frac{1+n_d}{1+df(t_i)}+1$ 表示词t_i的逆文档频率，n_d表示专利申请人文本总数，$df(t_i)$表示包含词t_i的专利权人文本数
	（4）计算*TF-IDF*	$tf\text{-}idf(t_i, d_j)=tf(t_i, d_j)\times idf(t_i)$
	（5）标准化*TF-IDF*	$v_{norm}(t_i, d_j)=\frac{v}{\sqrt{v_1^2+v_2^2+\cdots v_n^2}}$ $v_{norm}(t_i, d_j)$ 表示词t_i在专利申请人文本d_j中的权重，通过对原*TF-IDF*值的欧几里得范数归一化得到
输出：	特征词列表*termlist*，标准化后的权重矩阵*weight_array*	

（4）特征向量选择。通过权重计算可知，对每条专利申请人文本来说，其特征词之间具有轻重之分。特征向量提取后，包含很多冗余的特征，会对相似度计算结果构成干扰。因此，需要根据特征词的权重设置阈值，选择对于文本表达和相似度比较来说，最有用和最重要的特征向量。

2.1.1.2 基于余弦相似度的文本比较

余弦相似度用向量空间中两个向量夹角的余弦值作为衡量两个体间差异的大小。余弦值越接近1，就表明夹角越接近0°，也就是两个向量越相似，即余弦相似性。本文通过计算两个专利权人特征向量的夹角余弦来获得其相似度，计算方法如公式（1）所示。

$$osim(d_i, d_j)=\cos(V_i, V_j)=\frac{\langle V_i, V_j\rangle}{||V_i||\cdot||V_j||} \quad (1)$$

V_j表示专利申请人文本d_j的特征向量，二者的余弦相似度，通过两个向量的内积除以两个向量的模的乘积得到。

2.1.2 地址相似度计算

申请人地址信息由邮政编码和实际街道门牌地址构成。我国邮政编码采用四级六位数编码结构。前两位数字表示省（直辖市、自治区），前三位数字表示邮区，前四位数字表示县（市），最后两位数字表示投递局（所）。①如果发明人记录对的邮编相同，可以二者对应的申请人地址十分接近。②与街道门牌地址相比，邮编更简短规范，可以减少由于语言表述不一致产生的相似度比较误差。③地址信息中包含有效邮编的概率超过98%。所以本文借助邮编对专利申请人地址进行分级匹配。遵循以下分级赋值规则：当记录对（d_i，d_j）的街道门牌地址相同，其地址相似度$asim$（d_i，d_j）=a1（a1=1）；当记录对（d_i，d_j）的街道门牌地址不相同，但邮编相同时，其地址相似度$asim$（d_i，d_j）=a2（a2=0.6）；当记录对（d_i，d_j）的街道门牌地址不相同、邮编也不相同，但邮编的前四位数字相同时，其地址相似度$asim$（d_i，d_j）=a3（a3=0.4）。

2.1.3 合作者相似度计算

参照Jaccard系数计算公式[15]，对发明人的合作者相似度定义见公式（2）。

$$wsim(i,j)=\frac{|A_i\cap B_j|}{|A_i\cup B_j|} \qquad (2)$$

式中，A_i表示发明人i的合作者集合，B_j表示发明人j的合作者集合。$wsim$（i，j）表示发明人i和发明人j的合作者相似度，它等于两个集合中相同合作者数占合作者总数的比例。

2.2 重名消解流程设计

本文采用一种“姓名池投放”的重名消解策略。姓名池是由具有同一姓名字串的所有发明人实体所组成的列表。为所有姓名字串构建姓名池，对于姓名相同的不同实体建立子池，各子池之间通过标签进行区分，子池中存放本发明人实体的所有相关属性。下面以发明人“赵峰”为例，阐释本研究的消解流程（见图1）。首先，将赵峰的第x条记录X与上一条记录X-1进行比较。在重名消解处理开始之前，可以先根据发明人姓名、专利申请人和地址等信息对记录集进行快速排序，这样如果存在与记录X完全相同或相似度很高的记录或记录集，那么X-1有很大可能性在其中。本操作可以降低算法的复杂度，节约时空成本。如果记录X与记录X-1的相似度大于阈值，则认为这两条记录指向同一发明人，那么此时将X-1的发明人标签L_{X-1}赋予X，同时将X中包含的属性并入子池“赵峰L_{X-1}”；如果相似度小于阈值，将X与姓名池中每个实体属性进行逐一比对，直到找到一个实体i，使得X与实体i的相似度大于阈值，则停止循环，将标签i赋予X，将其属性并入子池“赵峰i”。如果姓名池中的实体都无法与X匹配，则为X所指发明人建立新子池“赵峰n+1”，并将其插入姓名池。

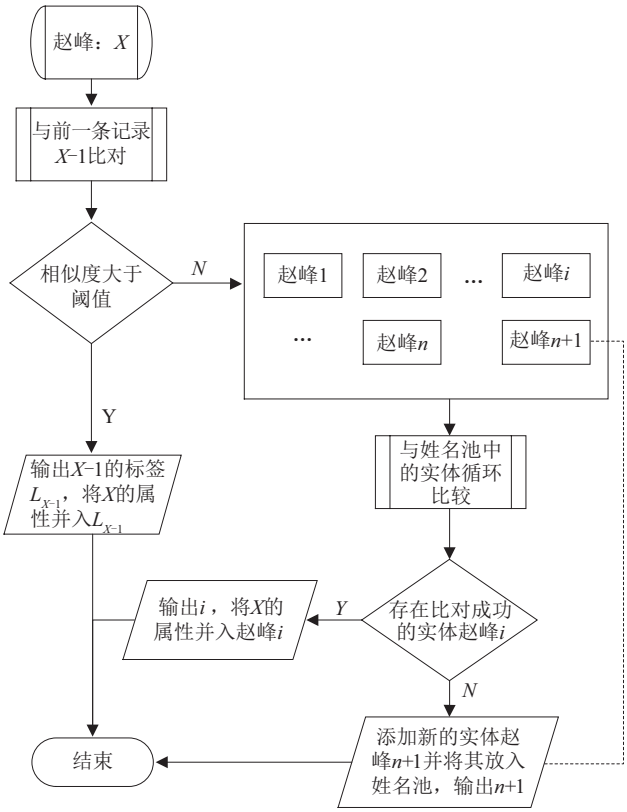

图1 算法流程

3 实验验证

3.1 测试集与阈值设定

本文基于中国科学技术信息研究所《电动汽车专题数据库》开展实证研究，从库中抽取中国专利申请人申请的中国国家专利3 899条，拆分专利发明人后，形成“发明人-专利”记录13 758条，其中需要进行重名消解的记录10 511条。在相似度匹配方面，如果两条记录满足以下4个条件之一，则判定两条记录中的发明人为同一位发明人。

条件1：（name1=name2）AND（*osim*>0.5）
/*发明人姓名相同且机构相似度大于0.5*/
条件2：（name1=name2）AND（*asim*>0.5）
/*发明人姓名相同且地址相似度大于0.5*/
条件3：（name1=name2）AND（（*osim*+*asim*）>0.7）
/*发明人姓名相同且机构相似度与地址相似度之和大于0.7*/
条件4：（name1=name2）AND（*wsim*>0）
/*发明人姓名相同且具有相同合作者*/

3.2 验证方法与评价指标

采用抽样人工确认的方式对实验结果进行验证，抽样方法选择分层抽样。借鉴布拉德福定律中对于核心区、相关区和外围区的定义，本文将专利发明人按其参与的发明数量（*P*）倒序排列，并将其划分为3个类别，控制每个类别中的专利总数（指总人次）相等。其中，①核心发明人：$P\geqslant 10$；②相关发明人：$4\leqslant P<10$；③外围发明人：$P<4$。从每一类别中随机抽取10个姓名，对其所指向的实体进行人工确认。采用正确率指标$A=\frac{TC}{N}$来评价算法优度，其中，*N*表示某一姓名所包含的所有记录条目，*TC*表示正确分类的记录条目；即当某一发明人被正确分配到其所在子池时，我们认为判定结果正确。在进行结果确认的过程中，发现有4条关于“胡明晖”的记录及9条关于“胡明辉”的记录，经询问本人得知2个姓名指向同一发明人“胡明辉”，因此将两部分记录合并，将“胡明晖”列为核心发明人，再从相关发明人姓名集中随机抽取1个姓名样本。因此，最终样本数据包含姓名31个、记录341条，其中核心发明人姓名11个。

3.3 结果分析

3.3.1 错误案例分析

实验结果的总体正确率为96.48%，每个姓名的具体正确率见表2。出现的错误主要为“第一类”错误，即2条记录实际指向同一发明人，却将其判别为2位不同发明人。如图2所示，根据实验结果可知：①“秦大同”指向2个实体，一个在重庆市，另一个在江苏省，二者的申请人、合作者和地址信息相似度都为0。经核实发现，秦大同为重庆大学的教授，同时受聘于苏州安远新能源动力有限公司担任技术副总监。曹秉刚、孙力也属于这种“在外兼职”的情况。②“孙辉”指向2个实体，一个属于哈尔滨工业大学，另一个属于徐工集团工程机械股份有限公司，从实体属性来看没有关联。经核实发现，孙辉曾在哈尔滨工业大学攻读博士学位，2009年毕业后就职于徐工集团，二者实际为同一人。陈慧勇同属这种“单位变更”的情况。③“胡明晖”与“胡明辉”的实体属性非常相似，但是由于二者姓名不同，将其判定为不同实体。经核实发现，关于“胡明晖”和“胡明辉”的专利都属于发明人胡明辉，“胡明晖”是因为誊抄失误而造成的姓名误差。

3.3.2 算法比较

为对比不同算法的重名消解效果，本文分别单独使用申请人相似度、地址相似度和发明人相似度对实验数据进行处理（见表3）。基于VSM的申请人余弦相似度可以准确辨别76.83%的姓名歧义，比精确匹配

表2 重名消解算法实验结果

核心发明人			相关发明人			外围发明人		
序 号	姓 名	正确率/%	序 号	姓 名	正确率/%	序 号	姓 名	正确率/%
#1	任勇	100.00	#1	沈晞	100.00	#1	包寿红	100.00
#2	秦大同	94.29	#2	陈立建	100.00	#2	罗念宁	66.67
#3	赵川林	100.00	#3	顾伟	100.00	#3	孙力	66.67
#4	孙辉	96.67	#4	李涛	100.00	#4	王磊	100.00
#5	曾小华	100.00	#5	于远彬	100.00	#5	侯哲	100.00
#6	杨亚联	100.00	#6	曹秉刚	66.67	#6	李建朋	100.00
#7	陈慧勇	94.12	#7	曹正策	100.00	#7	石万凯	100.00
#8	张永生	100.00	#8	余涛	100.00	#8	王军平	100.00
#9	朱光海	100.00	#9	顾红娟	100.00	#9	王林	100.00
#10	胡明辉	69.00	#10	张德平	100.00	#10	杨洋	100.00
#11	高卫民	100.00						
总体		96.89	总体		96.88	总体		91.67

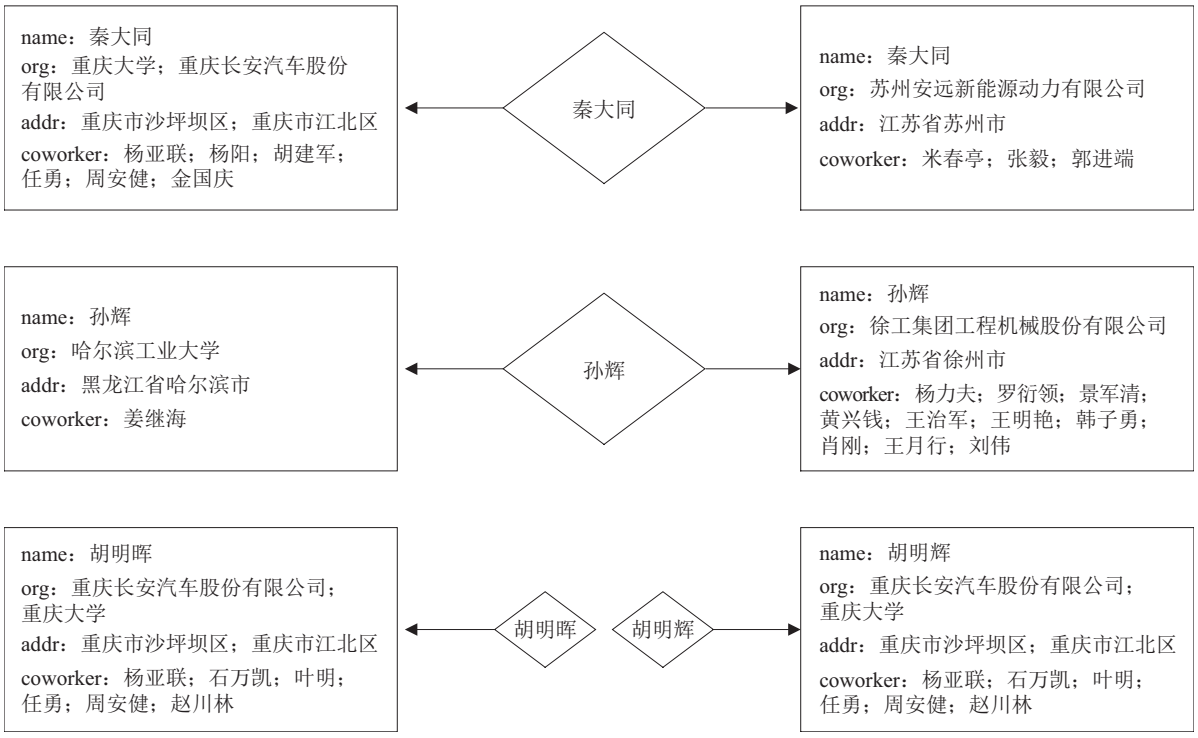

图2 错误案例详细属性

算法提高30%。基于分级匹配算法的地址相似度可以准确辨别72.43%的姓名歧义，比精确匹配算法提高20%。单独使用合作者相似度可以准确辨别91.20%的姓名歧义。由此可知，本研究选用的方法可以有效提高重名消解的准确率。此外，发明人合作网络对于实体链接的辅助判断作用，相对于专利申请人和地址属性来说更为重要。

表3 不同相似度算法正确率比较

申请人相似度		地址相似度		合作者相似度
精确匹配	基于VSM的余弦相似度	精确匹配	分级匹配	91.20%
58.94%	76.83%	60.11%	72.43%	

4 结论与展望

专利发明人的重名消解是一项重要工作，海量专利数据的诞生使得依靠人工标注的消解方式不能满足需求。本文针对中文专利题录数据的特征，提出一种适应大规模专利数据环境的自动化重名消解方法，旨在提高发明人重名消解的效率和准确率。本文的主要贡献在于，首先，将发明人相似度的计算拆分为申请人相似度、地址相似度和合作者相似度3个部分。其中申请人相似度计算采用基于向量空间的文本表示模型和基于余弦相似度的相似度计算模型，以识别专利申请人之间的隶属、更名等潜在关系。地址相似度采用基于邮编和街道门牌地址的分级匹配方法，以解决因为街道门牌地址书写不规范而造成的匹配不足问题。合作者相似度通过Jaccard系数计算。其次，在流程设计上采用基于“姓名池投放”的重名消解策略，将已经确认实体指向的记录中所包含的属性并入相应子池，新记录在与优先匹配记录匹配不成功时，只需要与各子池进行逐一比较，直至匹配成功，而记录之间不必两两比较，降低了算法的时空成本。

本文所提出的专利发明人重名消解方法也存在一些局限性。首先，由于同音或形近字造成的姓名誊写错误不能很好地识别；其次，专利题录数据中包含的发明人背景信息较少，不足以完全识别出所有的相同发明人。在未来研究中将考虑汉字的特殊性，探索针对中文姓名的模糊匹配方法。另外，拓展用于重名消解的数据

基础，除专利题录数据外，将诸如发明人履历信息等纳入参考范畴。

参考文献

[1] 邢晓昭. 专利信息生命长度测度及规律研究——以运算处理和计算领域为例[D]. 北京：中国科学技术信息研究所，2013.

[2] 中华人民共和国国务院. 中华人民共和国专利法实施细则[EB/OL]. [2018-10-01]. http://www.sipo.gov.cn/zhfwpt/zlsqzn/zlfssxzsczn/201508/t20150824_1164885.html.

[3] SINGH J. Collaborative networks as determinants of knowledge diffusion patterns [J]. Management Science，2005，51(5)：756-770.

[4] MIGUÉLEZ E，GÓMEZMIGUÉLEZ I. Singling Out Individual Inventors from Patent Data [EB/OL]. (2011-06-03) [2018-04-25]. https://www.researchgate.net/publication/228272261_Singling_Out_Individual_Inventors_from_Patent_Data.

[5] 王道仁，杨冠灿，傅俊英. 专利发明人英文重名识别判据及效度比较分析[J]. 数字图书馆论坛，2016(8)：2-9.

[6] 刘斌，赵升，孙笑明，等.我国专利数据中发明家姓名消歧算法研究[J]. 情报学报，2014，35(4)：405-414.

[7] HAN H，ZHA H Y，GILES C L. Name disambiguation in author citations using a k-way spectral clustering method [C]//2005 Proceedings of the 5th ACM/IEEE Joint Conference on Digital Libraries in Digital. Denver：ACM，2005：334-343.

[8] LI G C，LAI R，D'AMOUR A，et al. Disambiguation and co-authorship networks of the U.S. patent inventor database (1975—2010) [J]. Research Policy，2014，43(6)：941-955.

[9] VENTURA S L，NUGENT R，FUCHS E R H. Methods Matter：Revamping Inventor Disambiguation Algorithms with Classification Models and Labeled Inventor Records [EB/OL]. (2013-03-01) [2018-04-25]. https://www.researchgate.net/publication/256021263_Methods_Matter_Revamping_Inventor_Disambiguation_Algorithms_with_Classification_Models_and_Labeled_Inventor_Records.

[10] 朱云霞. 中文文献题录数据作者重名消解问题研究[J]. 图书情报工作，2016，58(3)：143-148.

[11] SALTON G. A vector space model for automatic indexing [J]. Communication of the Acm，1975，18(11)：613-620.

[12] 吴夙慧，成颖，郑彦宁，等. 文本聚类中文本表示和相似度计算研究综述[J]. 情报科学，2012，30(4)：622-627.

[13] SALTON G，BUCKLEY C. Term-weighting approaches in automatic text retrieval [J]. Information Processing & Management，1987，24(5)：513-523.

[14] 刘晓豫，朱东华，汪雪峰，等. 多专长专家识别方法研究[J]. 图书情报工作，2018，62(3)：55-63.

[15] 周朝阳. 杰卡德相似度在图书推荐中的应用研究[J]. 情报探索，2017(7)：43-46.

作者简介

邢晓昭，女，1988年生，硕士，助理研究员，研究方向：科研团队识别、专利分析，E-mail：xingxz@istic.ac.cn。

郑彦宁，男，1965年生，研究馆员，博士生导师，研究方向：情报学理论方法。

Research on Inventors' Name Disambiguation for Chinese Patent Information

XING XiaoZhao ZHENG YanNing

(Institute of Scientific and Technical Information of China, Beijing 100038, China)

Abstract: The analysis of patent inventors provides powerful data support for the evaluation of technical talents and the identification of scientific research teams. However, there are large number of duplicate names existing in the Chinese names, making the research results based on inventors deviate. In this paper, we propose an algorithm dealing with the duplicate names of the inventors based on rules. Given the inconsistencies in names of the patent applicants caused by reasons such as merger, split, restructuring or strategic transformation, a cosine similarity algorithm based on Vector Space Model is adopted to judge the relevant institutions. In view of the inconsistent addresses due to the incorrect writing of the house number, a hierarchical matching algorithm based on zip code and house number is applied to identify the similar address information. The similarity of the collaborators is calculated by the Jaccard coefficient. Finally, taking the *thematic database on electric vehicles* of ISTIC as an example, an empirical research is carried out to verify the scientificity and effectiveness of the method.

Keywords: Name Disambiguation; Chinese Patent Information; Inventor; Similarity; Vector Space Model

（收稿日期：2018-10-08）