专利发明人英文重名识别判据及效度比较分析*

王道仁，杨冠灿，傅俊英
（中国科学技术信息研究所，北京 100038）

摘要：本文针对英文专利发明人姓名的字符串匹配问题，利用USPTO发明人姓名的数据集，探索现有字符串匹配算法的适用性。对指向同一发明人不能精确匹配的姓名字符串，分别用10种常用的字符串匹配算法进行处理。比较匹配结果发现：Jaro-Winkler算法对同一发明人姓名字符串匹配效果最好，且结果稳定。通过回归分析可知，杰卡德算法对于发明人的识别效果最佳，基于q-gram的算法在发明人姓名消歧中有重要意义；在发明人消歧中，多种字符串匹配算法的组合运用效果更佳。

关键词：发明人姓名；字符串匹配；Jaro-Winkler算法；杰卡德算法
中图分类号：TP18
DOI：10.3772/j.issn.1673-2286.2016.8.001

1 引言

在专利数据的挖掘利用方面，建立大型专利数据集一直是相关研究的热点。德温特世界专利索引数据库（Derwent World Patent Index，DWPI）最早于1963年开始对专利数据进行加工以生成文摘类数据库，并对专利权人编码进行研究应用，为全球约2.1万家公司分别指定一个长度为4个字符的代码[1]。Hall等最早构建了一个关于美国专利商标局（United States Patent and Trademark Office，USPTO）专利数据的综合数据集——NBER专利和引文数据集，其中包含每条专利的基本字段信息和部分基于引文的指标，为后续学者利用USPTO专利数据提供便利，推动专利发明人相关数据分析研究的发展[2]。

对专利发明人数据进行分析应用，必要的前提是解决发明人的歧义问题。歧义是一种身份不确定的现象，指文本中具有相同姓名的字符串指向现实世界中不同实体人物。无论是合作网络分析还是引文分析，如果不解决姓名歧义问题，都会对研究结果造成严重干扰，降低数据的可信度[3]。在消歧过程中，数据匹配个人姓名和组织机构等信息的质量可能较低，一方面，这些数据可能包含录入错误以及印刷变体等问题；另一方面，姓名和地址信息本身也可能随时间而变化，导致许多机构或个人存在姓名变体的问题。Torvik等使用一种权威控制（Authority）方法对Medline数据库中文献作者姓名进行消歧，为将自动化算法、机器学习等技术引入专利数据处理做铺垫[4]；Trajtenberg等利用NBER数据集和Soundex编码系统，对全球发明人进行基于规则的消歧，得到超过180万含有唯一标识的全球发明人数据，并对该数据集经济学方面的应用途径进行探索[5]；PatentsView专利分析平台最新数据表明，1976年至今，USPTO共有授权专利5 915 849件，消歧后的唯一标识发明人共3 287 305人，平均拥有1.8件/人专利[6]。

人名消歧很长时间一直是个具有挑战性的问题，国内鲜有相关研究著作。目前为止，发明人消歧主要采用基于规则的消歧方法，该方法的精准率在70%—80%，仍有待提高。Zobel等对大词典中近似字符串用字符串匹配算法进行处理，结果表明：语音编码算法比字符串匹配算法更适合于拼、读、校正和个人姓名匹配，多种匹配算法融合使用效果更佳[7]；Donner等研究作者姓

* 本研究得到国家科技支撑计划课题“专利信息支撑科研项目管理应用示范”（编号：2013BAH21B05）资助。

名精确匹配中论文自引检测的缺陷，发现姓名模糊匹配可以增强论文自引检测的效果[8]；Fleming等较早地将字符串匹配算法应用于发明人消歧方面，对发明人姓名进行匹配算法处理后，再基于规则判断匹配与否[9]。匹配与否与比较的字段权重、分值以及设定的阈值相关。目前，字符串匹配与名称编码成为标准化处理发明人数据的两种主要途径。发明人消歧可以通过模糊字符串或者其他特征信息（如地理、技术领域、引文信息等）来辅助识别，本文主要关注通过模糊字符串匹配提升专利发明人识别的效果。

发明人英文姓名的匹配研究是发明人识别（归一化）的基础，可以提高消歧的效果。常用的英文字符串匹配算法对于具体的发明人姓名数据是否适用，以及各种匹配算法的效果都有待进一步验证。首先，对于各种匹配算法对发明人消歧效果的具体影响，并没有统一的认知；其次，不同算法间适用的场景有所差异，解决的问题不同，以及如何综合运用多种字符串匹配算法，提升发明人识别效果是本文重点关注的问题。

2 研究方法

2.1 数据来源

本文采用PatentsView研讨会提供的td数据集进行字符串相关匹配算法的研究[6]，是由Ge等于2014年公布的有关工程师和科学家的USPTO专利数据集组成[10]。该数据集已经过消歧处理，并赋予发明人唯一标识，可据此评价字符串的匹配效果。

从表1可见，常见的11种英文姓名形式及处理方式。最基础的英文姓名是由姓氏和名字两部分组成，其他还有前缀、后缀和中名，中名又有1个首字母缩写和2个首字母缩写的区分。姓氏和名字会有顺序变换的情况，因此需要识别每一种具体情况，将一个整体的姓名字符串，拆分为名字（first name）、中名（middle name）、姓氏（last name）和后缀（suffix）。USPTO专利发明人姓名信息更加复杂，处理时需要总结每一种情形，进行调整，不断反馈、改进。

表 1 英文姓名形式及其处理方式[11]

英文姓名	特 点	名 字	中 名	姓 氏	后 缀
Jeff Smith	没有中名	Jeff	-	Smith	-
Eric S. Kurjan	有中名，1个首字母缩写	Eric	S.	Kurjan	-
Janaina B. G. Bueno	有中名，2个首字母缩写	Janaina	B. G.	Bueno	-
Kahn, Wendy Beth	姓氏和名字顺序变换	Wendy	Beth	Kahn	-
Mary Kay D. Andersen	名字有2个字符串	Mary Kay	D.	Andersen	-
Paula Barreto de Mattos	姓氏有3个字符串	Paula	-	Barreto de Mattos	-
James van Eaton	姓氏有2个字符串	James	-	van Eaton	-
Bacon Jr., Dan K.	姓氏和名字顺序变换，后缀	Dan	K.	Bacon	Jr.
Gary Altman III	后缀	Gary	-	Altman	III
Mr. Ryan Ihrig	前缀	Ryan	-	Ihrig	-
Julie Taft‐Rider	姓氏是合成的	Julie	-	Taft‐Rider	-

2.2 精确匹配和模糊匹配

字符串匹配对比较的方法有多种，根据字段特征进行划分，可分为长字段和短字段两种类型。在专利发明人相关字段中，长字段类型主要包括发明人姓氏、名字等；短字段类型主要包括发明人的国别、州名、专利分类号等。短字段类型出现错误的概率较小，可直接使用精确匹配；长字段类型极易出现拼写错误或者印刷变体，需使用匹配算法返回一个代表相似性程度的数，现有研究一般将这种不能精确匹配的相似字符串间的匹配称为模糊匹配。精确匹配和模糊匹配的效果对比，可以明确匹配算法的效用。

统计可知，td数据集中被标识的记录共19 303条，指向6 715个含有唯一标识的发明人。对这些数据进行

数据清洗、数据预处理等基础工作（包括去除后缀、提取中名等）；另外，定义一个新字段full name，由预处理后的first name、middle name和last name合并组成；最终可利用的字段包含first name、last name、middle name和full name。

对含有唯一标识的同一发明人姓名数据（实际匹配）的匹配对，利用字符串精确匹配进行识别。发现其中first name有384对匹配对不能精确匹配，last name有66对，full name有550对。统计同一发明人的姓名产生变体的情况（字符串不相同但实际上指向同一发明人），可具体归纳为6种。

拼写错误（a类）：包括字符的增加、缺失、替换以及字符顺序的调换；

缩写（b类）：姓名只取前几位字符来代表姓名，或者以几个字母缩写代替姓名的情况；

空格（c类）：录入错误等因素造成的多余空格；

缺失值（d类）：只有first name没有last name，或者与之相反；

乱码值（e类）：姓名的某一字符或者全部字符以其他形式的编码出现；

预处理余留的问题（f类）：前缀、后缀没有处理完善，中名或者姓氏由两部分组成。

2.3 匹配算法

字符串匹配是指寻找对应真实世界中同一实体的所有字符串，具体可分为同一数据库中字符串匹配和不同数据库中字符串匹配。目前研究者已提出许多度量方法用于字符串匹配，相似度度量可以将一对字符串(x,y)映射为[0,1]的一个数值，数值越高，意味着x和y的相似度越高。

从表2可见，10种相似度度量方法可以分为3类。

（1）基于序列的相似度度量算法，指将字符串看作字符的序列，然后计算将一个字符串转化为另一个字符串的代价。汉明距离算法只能识别相同长度的字符串，应用范围较小，同时运算量也较小；Levenshtein距离算法和Damerau-Levenshtein编辑距离算法增加编辑操作，对于一般的字符串匹配都可以运用；最长公共子串距离算法可以最大程度减少多个单点字符串变化带来的影响，对不连续的单个字符替换或者缺失等处理效果较好；Jaro-Winker算法增加了姓名前几个字符的权重，对字符串后半部分错误较多的情况可以较好

表 2 10种常用字符串匹配算法

序 号	字符串匹配算法	算法特点
1	汉明距离算法	只对相同长度的字符串进行比较，且编辑操作只有替换一种
2	Levenshtein距离算法	将一个字符转化为另一个字符所需要的单一字符插入，删除和替换的步骤数
3	Damerau-Levenshtein编辑距离算法	Levenshtein编辑距离的直接拓展，编辑操作增加相邻字符的顺序调换
4	加权Damerau-Levenshtein编辑距离算法	Damerau-Levenshtein编辑距离拓展，包括赋予不同编辑操作不同的权重
5	最长公共子串距离算法	用一种迭代的方式，找到并移除2个字符串相同的最长公共子串，子串指连续的字符组
6	Jaro以及Jaro-Winker算法	$\mathrm{sim}_{\mathrm{jaro}}(s_1,s_2)=\frac{1}{3}\left(\frac{c}{\lvert s_1\rvert}+\frac{c}{\lvert s_2\rvert}+\frac{c-t}{c}\right)$， $\mathrm{sim}_{\mathrm{winkler}}(s_1,s_2)=\mathrm{sim}_{\mathrm{jaro}}(s_1,s_2)+\frac{p}{10}\left(1.0-\mathrm{sim}_{\mathrm{jaro}}(s_1,s_2)\right)$， c是匹配的字符数，t是换位的数目，p是两个字符串开头相同的首字母个数
7	基于q-gram的算法	q-gram是长度为q的字符串，其位置用首字符在模式串或文本串中的偏移量表示
8	基于q-gram的余弦算法	$\mathrm{sim}_{\mathrm{cosine}}(A,B)=\cos(\theta)=\frac{A\cdot B}{\lVert A\rVert\times\lVert B\rVert}=\frac{\sum_{i=1}^{n}A_iB_i}{\sqrt{\sum_{i=1}^{n}A_i^2}\sqrt{\sum_{i=1}^{n}AB_i^2}}$， A和B分别代表被比较的字符串
9	基于q-gram的杰卡德算法	$sim_{jaccard}(A,B)\lvert=\frac{\lvert A\cap B\rvert}{\lvert A\cup B\rvert}=\frac{\lvert A\cap B\rvert}{\lvert A\rvert+\lvert B\rvert-\lvert A\cup B\rvert}$， A和B中的元素为从字符串中抽取的q-gram
10	Soundex算法	返回由4个字符组成的代码，以评估两个字符串的相似性

地进行识别。

（2）基于集合的相似度度量算法，指将字符串看作词项的集合或包，然后使用集合有关的性质计算相似度得分，最常见的词项类型是q-gram。q-gram指字符串长度为q的子串，其基本思路是使用一种滑动窗口的方法将输入的2个字符串分割为短的长度为q的子字符（q-gram），然后统计两个字符串中q-gram的个数。q-gram算法和余弦算法、杰卡德算法对于字符串中任意位置出现错误的情况都能识别，相应的运算空间和运算时间要求比较高。

（3）语音度量（phonetic）最初是为处理人口统计中姓氏匹配问题而提出的。最常见的是基于姓名发音的Soundex算法，可以将姓氏转变成一个由4个字符串组成的编码表示这个单词的发音。Soundex算法比较适合由语音拼写引起的错误（通过电话或者口头输入），增强元音字节的作用。

不同度量算法均是从短文本的某一特征或者几种特征出发，具有各自的特点，在实际消歧过程中，最好能结合多种特征，以提高消歧精度。

目前，对于各种匹配算法运用到发明人消歧中的效用，缺乏综合的认知。首先，本文用10种算法处理实际匹配的USPTO发明人数据（模糊匹配），对相似性分值进行初步统计判断，并对USPTO发明人姓名的数据分布进行初步了解；然后，分别对不使用匹配算法和使用匹配算法的发明人消歧进行简单对比分析；最后，通过回归分析对10种算法于USPTO发明人识别的具体影响进行深入对比分析。

3 字符串算法匹配的结果

3.1 同一发明人Full Name的算法匹配

对模糊匹配的同一发明人的full name，分别用表2的10种匹配算法进行处理，得到的结果（见表3）。由相似性数值可知Jaro-Winkler算法和最长公共子串距离算法是效果最好的2种算法。

表 3 同一发明人Full Name算法匹配的算法结果

分 类	字符串对	hamming	osa	lv	dl	lcs	jw	q-gram	cosine	jaccard	soundex
a类	HarilosMavridis HarilaosMavridis	0	0.94	0.94	0.94	0.97	0.96	0.90	0.91	0.81	1
b类	Kenneth J. Biddle Ken Biddle	0	0.59	0.59	0.59	0.74	0.83	0.64	0.67	0.47	0
c类	James DeVore James De Vore	0	0.67	0.67	0.67	0.77	0.90	0.69	0.70	0.53	0
d类	NA Andrew Felix GTI Andrew	0	0.44	0.44	0.44	0.56	0.60	0.52	0.55	0.35	0
e类	Andres F. Andr<U+00C3>s F.	0	0.89	0.89	0.89	0.91	0.97	0.85	0.85	0.74	1
f类	James R. York James R York	0	0.92	0.92	0.92	0.96	0.98	0.87	0.87	0.77	1

注：“hamming”表示汉明距离算法，“osa”表示加权Damerau-Levenshtein编辑距离算法，“lv”表示Levenshtein距离算法，“dl”表示Damerau-Levenshtein编辑距离算法，“lcs”表示最长公共子串距离算法，“jw”表示Jaro-Winker算法（参数p=0.1），“q-gram”表示q-gram算法（参数q=2），“cosine”表示基于q-gram的余弦算法（参数q=2），“jaccard”表示基于qgram的杰卡德算法（参数q=2），“soundex”表示Soundex算法。

统计可知，除汉明距离算法，其他算法均能将大部分属于同一发明人的full name识别出来。除缺失值外（类别为d），Jaro-Winkler算法基本上对所有模糊匹配的字符串处理效果都比较好，大部分相似值在0.50以

上，且结果稳定。对于姓名缺失的情况，10种算法的识别效果都很差。Jaro-Winkler算法由美国人口统计局根据美国人的姓名特点发展而来，在比较姓名前几个字符匹配的情况下，会被赋予较高的相似性分值。

对550对属于同一发明人模糊匹配的full name进行算法匹配，共5 279条匹配数据，得到的相似性分值分布情况，如图1所示。其中，Jaro-Winkler算法匹配结果中有3 978个匹配对的相似性分值在[0.9，1]，远超以1 674对排在第2位的最长公共子串距离算法；另外，Jaro-Winkler算法匹配结果的相似性分值大部分都在0.7以上，表明大部分属于同一发明人，不能精确匹配的姓名字符串被算法正确识别。

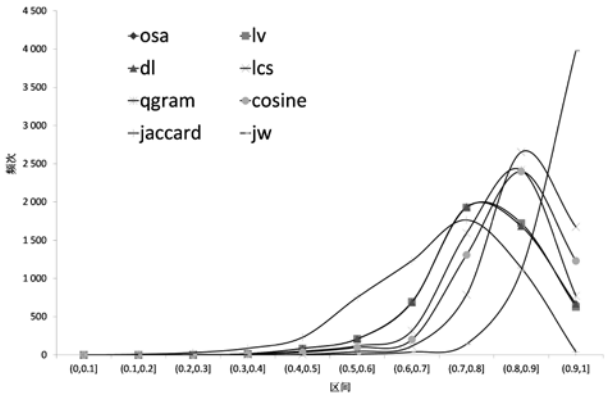

图 1 同一发明人Full Name模糊匹配相似值分布

3.2 精确匹配和发明人消歧

为进一步验证各种字符串匹配算法对发明人数据最终消歧效果的影响，本文以消歧中使用的字符串匹配算法为因变量，探索其对USPTO发明人识别的影响。发明人消歧中可利用的数据按发明人相关的属性划分，主要包括发明人姓名的拆分、组合，以及专利其他相关属性（见表4）。

发明人姓名是发明人消歧中最重要的判别因素，将发明人的英文姓名看作字符串，则发明人消歧最终转化为字符串匹配问题。发明人姓名的处理方法对消歧效果有直接影响，在合理范围内所有可能的姓名组合方式，是对发明人姓名数据的最有效利用。由发明人链接到其相关专利，由此再链接到更多的相关数据信息。比如，在发明人姓名模糊匹配的情况下，专利权人和专利分类号是否一致可以提高判断的精准性。发明人消歧算法，具体可以分为基于监督学习、基于无监督学习、基于半监督学习以及基于规则的算法。本文使用基于规则的消歧算法，设置阈值为0.6，即分值大于0.6的数据对视为指向同一发明人。

如果将所有数据进行两两比较生成消歧数据对，会占用大量的运算空间和时间。限制比较的数据在合理范围内的过程称为区块化（Blocking），本文使用city和

表 4 发明人及其专利相关属性

序 号	名 称	英文表示	内 容
1	名字	first name	发明人的名字
2	中名	middle name	发明人的中名
3	姓氏	last name	发明人的姓氏
4	全名	full name	将名字、中名和姓氏联合起来
5	城市	city	发明人记录里的城市名称
6	州	state	发明人记录里的州名简称
7	国家	country	发明人记录里的国家简称
8	专利权人	organization	清洗后的对应专利发明人的专利权人
9	USPC分类号	subclass	对应专利的技术分类号

first name、last name的首字母进行限定，即city和first name、last name的首字母相同的发明人数据才进行匹配。从表5可见，第1条记录分别和第36条、第1 002条记录匹配的结果，0代表不同，1代表相同。

删除数据中有空值的记录，共有158 148条记录生成4 231 091个比较数据对。精确匹配时的消歧结果如表6所示，使用Levenshtein距离算法消歧结果的准确性比不使用匹配算法的准确性提高10.1%，Jaro-Winkler算法为7.3%。

表 5 精确匹配的数据对

id1	id2	city	state	country	subclass	organization	first name	middle name	last name	full name
1	36	1	1	1	1	1	1	0	1	1
1	1002	1	1	1	0	1	1	0	1	1

表 6 精确匹配的消歧结果

真实状态	不匹配	潜在匹配	匹配
0	165 943	0	0
1	701 729	0	3 363 419

3.3 模糊匹配和发明人消歧

利用线性回归，可进一步验证运用字符串匹配算法的数据是否比精确匹配的数据效果更好。以发明人是否匹配为因变量，可分别对发明人的名字和姓氏等分别设置单因素变化，进行对比分析，本文以full name为例。

从表7可见，全名不精确匹配的发明人数据线性回归中，R^2为5.92%，其中发明人的城市，对应专利的分类号和专利权人与是否匹配是显著相关的。

以发明人的全名字符串匹配为例，对10种字符串算法处理的数据分别进行回归分析。结果如表8所示，杰卡德距离算法（jaccard）效果最佳，R^2提升最多，其次是q-gram算法，余弦距离算法（cosine）紧随其后，三者都是以q-gram为基础的算法。Jaro-Winkler算法对相似发明人姓名赋予的相似性分数最高，同时也会造成一个问题，即对不是同一发明人的相似姓名赋分过高。10种算法的R^2都高于5.92%，对于发明人消歧效果都有提高。

表 7 基本发明人数据的线性回归

参 数	系 数	方 差	*t*值	*p*值
city	0.055 990	0.000 883	63.405	<2e-16 ***
state	0.000 025	0.000 227	0.112	0.91 ***
subclass	0.013 580	0.002 088	6.503	0.000 000 000 1 ***
organization	0.032 410	0.000 872	37.177	<2e-16 ***

注：“***”表示极显著，$p<0.01$；无“***”表示不显著，$p>0.05$。

表 8 Full Name的字符串算法匹配线性回归结果

算 法	系 数	方 差	*t*值	*p*值	R^2
jaccard	0.023 930	0.000 622	38.474	<2e-16 ***	7.30%
q-gram	0.012 060	0.000 426	28.292	<2e-16 ***	6.67%
cosine	0.011 340	0.000 408	27.801	<2e-16 ***	6.65%
lcs	0.011 510	0.000 482	23.877	<2e-16 ***	6.46%
lv	0.010 530	0.000 445	23.671	<2e-16 ***	6.45%
osa	0.010 501	0.000 444	23.629	<2e-16 ***	6.45%
dl	0.010 493	0.000 445	23.610	<2e-16 ***	6.45%
jw	0.007 695	0.000 480	16.041	<2e-16 ***	6.17%
soundex	0.002 404	0.000 247	9.720	<2e-16***	6.01%
hamming	0.002 573	0.000 720	3.572	0.000 354***	5.94%

注：“***”表示极显著，p<0.01。

如图2所示，10种匹配算法模糊匹配的数据相关系数热力图，可以发现10种算法可大致归为6类。基于序列的相似度度量算法可具体分为3类，Levenshtein距离算法（lv）、Damerau-Levenshtein编辑距离算法（dl）和加权Damerau-Levenshtein编辑距离算法（osa）比较接近，Jaro-Winkler算法（jw）、最长公共子串距离算法（lcs）和汉明距离算法（hamming）和其他算法均无明显的关联性。各种编辑距离算法经过不断发展，展现出不同的特色。以q-gram为基础的3种算法差异较小，相互间关联最近；Soundex算法是唯一基于语音相似性的匹配算法，和其他算法关联性最小。早期发展的字符串匹配算法对发明人姓名虽有所识别，但效果较差。

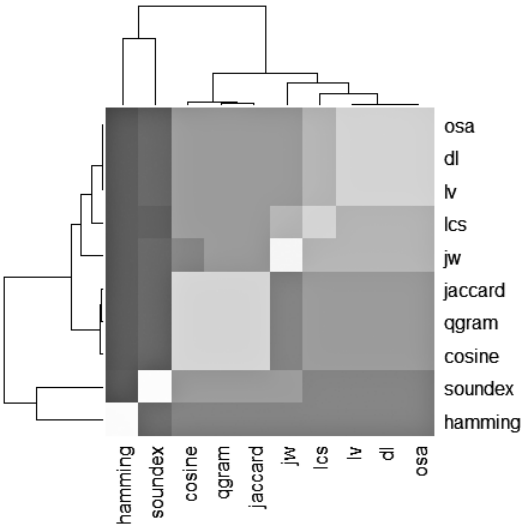

图 2 10种算法的相关系数热力图

汉明距离算法是最早的编辑距离算法[12]，只能识别相同长度的字符串，匹配结果只有0和1。随后的广义Levenshtein距离算法[13]，将编辑操作拓展到插入、删除和替换，但是每一次编辑增加一个相同单位的成本；Jaro-Winkler算法改进了Jaro算法在处理两个具有相同前缀，很有可能匹配的字符串相似性分值较低的情况，即提高了2个字符串开头字母相同时匹配的权重[14-15]；基于q-gram字符串匹配算法可以最大程度地利用2个字符串片段重叠的部分，对于在姓名中间部分出现录入错误或顺序变化等有较好的平滑效果，可以保证其余部分词项（子字符串）的匹配[16]。近年来，相关算法研究的重点是将字符串编辑距离和q-gram集合相结合。Soundex算法的匹配结果也只有0和1，一般适合与其他算法联合使用。作为一种语音度量方法，Soundex算法最初就是用于移民信息统计和人口普查统计，不过其对东亚地区的人名不能很好地处理。如果只是十分模糊的语音匹配处理，可以用Soundex算法[17]。

选取Jaro-Winkler算法、最长公共子串距离算法、Damerau-Levenshtein编辑距离算法、杰卡德算法和Soundex算法，以及发明人城市、专利分类号和专利权人，利用逐步回归中的后向回归，观察各种算法交叉组合和是否匹配的关系。根据精确AIC准则，可以发现，2种算法的交叉中，Jaro-Winkler算法和最长公共子串距离算法效果最佳；3种算法的交叉中，Jaro-Winkler算法、最长公共子串距离算法和杰卡德算法的组合效果最佳；4种算法的交叉中，Jaro-Winkler算法、最长公共子串距离算法、Damerau-Levenshtein编辑距离算法和Soundex算法的组合效果最佳。统计各回归结果，匹配算法组合的R^2均比单一算法要高，且组合越多，效果越好。

4 结语

类似于USPTO数据库，大部分专利数据库并没有赋予专利发明人唯一的、一致的可识别标识。无论是合作网络分析还是引文分析，如果不解决姓名的歧义问题，都会对研究结果造成严重干扰，降低数据的可信度。对于发明人进行计量分析的首要任务是对发明人姓名进行消歧处理，这其中的关键点是判断相似的英文姓名字符串是否指向同一发明人。本文通过USPTO中同一发明人姓名模糊匹配的的结果，发现Jaro-Winkler算法匹配效果最好。通过回归分析，发现杰卡德距离算法是实际的数据匹配中效果最佳的单一算法，基于q-gram的算法对于英文发明人姓名消歧有着广泛的应用前景。最后通过逐步回归，可知各种类型的匹配算法的组合效果会比单一算法更适合发明人的识别。

归一化后的发明人数据，可以对申请国际专利的发明人进行精确的专利计量，有助于人才评价、技术测评等。对于专利发明人进行归一化研究，可以进一步定量地研究发明人的国际流动、发明人的合作网络变化和技术溢出等。本文主要是针对USPTO专利数据，研究对象主要是英文姓名识别，对于中国发明人的英文姓名识别，需要结合英文姓名消歧算法和中国国家知识产权局的专利数据，这也是后续研究方向。

参考文献

[1] Derwent Innovations Index. Derwent patent assignee/company codes [EB/

OL].[2015-09-12].http://images.webofknowledge.com/WOKRS518B4/help/zh_CN/DII/hs_assignee_name.html

[2] HALL B H,JAFFE A B,TRAJTENBERG M.The NBER patent citation data file: lessons, insights and methodological tools[EB/OL]. [2016-08-01]. https://ideas.repec.org/p/nbr/nberwo/8498.html.

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

[4] TORVIK V I,WEEBER M,SWANSON D R,et al. A probabilistic similarity metric for medline records:a model for author name disambiguation[J].Journal of the American Society for Information Science and Technology,2005,56 (2):140-158.

[5] TRAJTENBERG,SHIFF,MELAMED.The"NamesGame":harnessing inventors' patent data for economic research[J].National Bureau of Economic Research,2006(8):12479.

[6] PatentsView.PatentsView inventor disambiguation technical workshop participant information[EB/OL].[2015-09-12].http://www.PatentsView.org/workshop/participants.html#how.

[7] ZOBEL J,DART P.Finding approximate matches in large lexicons[J]. Software Practice & Experience,2015,25(3):331-345.

[8] DONNER P. Enhanced self-citation detection by fuzzy author name matching and complementary error estimates[J].Journal of the Association for Information Science & Technology,2016,67(3):662-670.

[9] FLEMING L,KING C,JUDA A I.Small worlds and regional innovation[J]. Organization Science, 2007,18(6):938-954.

[10] GE C,HUANG K W,PNG I P L.Engineer/Scientist careers: patents, online profiles, and misclassification bias[EB/OL].(2014-11-27)[2016-08-01]. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2531477.

[11] Microsoft office support.Split text among columns by using functions[EB/OL].[2016-04-12].https://support.office.com/en-us/article/Split-text-among-columns-by-using-functions-c2930414-9678-49d7-89bc-1bf66e219ea8?CorrelationId=bd77b568-5d86-4fda-8476-a1772bbb3d22&ui=en-US&rs=en-US&ad=US14-9678-49d7-89bc-1bf66e219ea8?CorrelationId=65d0a205-2e7f-4bc8-92a4-66def716a327&ui=en-US&rs=en-US&ad=US.

[12] HAMMING R W. Error detecting and error correcting codes[J].The Bell System Technical Journal,1950(29):147-160.

[13] LEVENSHTEIN V I.Binary codes capable of correcting deletions, insertions,and reversals[J].Soviet Physics Doklady,1996(10):707-710.

[14] MATTHEW J. UNIMATCH:arecord linkage system:user manual[M]. Washington:United States bureau of the census,1978:103-108.

[15] WINKLER W.String comparator metrics and enhanced decision rules in the Fellegi-Sunter model ofrecord linkage[M]//Proceedings of the Section on Survey Research Methods.American Statistical Association,1990:354-359.

[16] CAVNAR W B,TRENKLE J M.N-gram-based text categorization[C]//3rd Annual Symposium on Document Analysis and Information Retrieval. Proceedings of SDAIR-94,1994:161-175.

[17] 马立东.Soundex语音匹配算法综述[J].现代计算机(专业版), 2010(5):17-20.

作者简介

王道仁，男，1990年生，硕士研究生，研究方向：前沿领域分析和专利分析，E-mail：wangdr2014@istic.ac.cn。
杨冠灿，男，1981年生，博士，助理研究员，研究方向：专利数据、技术竞争情报。
傅俊英，女，1972年生，博士，研究员，研究方向：专利分析、生物医药领域情报研究。

A Comparative Analysis of English Name Recognition Criterion and the Validity of the Patent Inventor

WANG DaoRen, YANG GuanCan, FU JunYing
(Institute of Scientific and Technical Information of China, Beijing 100038, China)

Abstract: To solve the string matching problem of the English name for the patent inventor, the paper used USPTO data sets, exploring the applicability of the existing string matching algorithm. The string belonging to the same invention, but not exactly matching, were processed with 10 kinds of common string matching algorithms. By comparing the match results, analysis, Jaro-Winkler algorithm name string best match effect to the same inventor, and the results are stable. Regression analysis shows the Jaccard algorithm (Jaccard) works best for inventor recognition, algorithm based on q-gram in the name of the inventor in the disambiguation is important. Disambiguation of the inventor, multiple strings matching algorithm used is better.

Keywords: Inventor Name; String Matching; Jaro-Winkler Algorithm; Jaccard Algorithm

（收稿日期：2016-07-29）