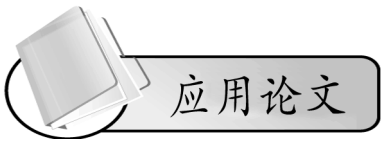

科技文献数据库中机构名称匹配策略研究*

孙海霞[1,2] 王 蕾[2] 吴英杰[2] 华薇娜[1] 李军莲[2]
[1](南京大学信息管理学院 南京 210093)
[2](中国医学科学院医学信息研究所 北京 100020)

摘要: 【**目的**】规范科技文献数据库中机构名称存储与管理, 设计并实现机构名称匹配策略。【**方法**】引入地区、类别和命名特征, 构建 3 类 7 组匹配判定规则, 设计 4 组规则与编辑距离混合的匹配策略, 基于中文生物医学文献数据库 2006 年–2011 年“作者单位”数据进行实现与评估。【**结果**】在 600 余万条“作者单位”数据集上, 对高等院校、医院与科研院所三类机构进行匹配实现, 结果表明综合考虑机构地区和命名特征规则的混合匹配策略表现最佳, 准确率均在 80%以上, 召回率达 64.82%, F 值达 71.66%。【**局限**】辅助词典和规则构建主要依赖人工经验, 覆盖面不全; 机构名称识别存在错误, 对匹配结果产生影响; 提出的匹配策略无法有效解决机构名称形态差异较大的规范问题。【**结论**】本研究提出一种基于规则和编辑距离的机构名称匹配策略, 能够提高科研文献数据库建设的规范性。

关键词: 信息检索 机构名称规范 相似度计算 混合策略 文献数据库
分类号: TP393
DOI: 10.11925/infotech.2096-3467.2018.0178

1 引 言

基于机构名称的信息检索已成为数据库和学术搜索引擎的重要检索行为之一。然而, 现实中由于不同信息载体对机构名称著录要求不一致, 机构更名、合并、拆分引起的机构变名, 作者书写习惯不同, 数据录入和数字化转换错误等客观存在, 使得各类数据库和搜索引擎很难保证数据的查准率和查全率[1-2], 从而影响以机构为中心的各类统计分析和评价结果的可靠性, 如基于科技文献的领域核心机构分析与排名、目标机构科研动态追踪、科研投入产出评估、潜在合作机构选择等[3-5]。构建机构名称规范文档, 进行机构名称规范化, 是目前诸多数据库、搜索引擎和项目提高数据可靠性的主要手段[6-9]。

以基于科技文献的机构名称规范为例, 机构名称规范化任务一般包括两个核心环节。一是从“作者单位”著录项中识别机构名称。在作者提交的论文和科技文献数据库中, 科研机构名称一般连同所在城市和邮编出现在“作者单位”著录项中。二是在同一机构实体不同名称表现形式间建立映射, 实现同一机构实体不同表现形式的同义汇聚。鉴于科技文献数据库“作者单位”项具有一定的结构性和规范性, 机构名识别任务难度相对简单, 学者主要聚焦于后者。众多研究中, 字符串匹配已成为常用方法之一[6,10-11],也构成其他规范方法的基础技术[12-14]。

然而, 仅靠字符串匹配无法解决很多机构名称字面相似度很高, 但并非指向同一个机构实体的问题。对此, 有学者引入机构所在地区、邮编等特征, 用以过

通讯作者: 李军莲, ORCID: 0000-0001-8955-6969, E-mail: li.junlian@imicams.ac.cn。

*本文系中央级公益性科研院所基本科研业务费专项“基于共现分析的著者机构名称规范机制研究”(项目编号: 2016RC330006)和国家科技图书文献中心“下一代国家科技创新开放知识服务系统”先期研发任务“STKOS 自动构建与维护关键技术研究”(项目编号: XQYF0102)的研究成果之一。

滤错误的匹配结果，且效果较佳[10,15]。本文则在此基础上，以经典 Levenshtein 字符串匹配算法[16]为基础，研究加入机构类别和机构名称字符串语言学特征是否能够进一步提升匹配效果。

2　机构名称匹配技术研究现状

机构名称匹配是机构名称规范的核心环节。面向不同应用需要(如信息检索、作者消歧)和不同语料基础，如科技文献、社交网络信息、百度百科、Web 网页等，学者们采用的整体策略也不同。就采用的基础方法来看，基本可以分为 4 类：基于字符串相似度计算的方法、基于统计的方法、基于规则的方法和混合的方法。

(1) 基于字符串相似度方法的基本思路是将机构名称字符串看作是字符序列，字符序列间相同的字符越多，表明这两个字符串越相似。当相似度大于一定阈值，即可认为这两个名称字符串指向同一个机构实体。如 French 等[6,14]分别在 1997 年和 2000 年采用 Hall-Dowling 编辑距离算法和 Jaccard 系数法进行天体物理数据系统作者机构名称规范文档半自动化构建研究；Jacob 等[17]基于 Levenshtein 法对求职简历中的学术机构名称进行匹配、规范。

(2) 基于统计的方法主要是将 Web 作为语料，以机构名称字符串作为检索词，根据搜索引擎返回的前 n 个 Web 页面中是否含有相同 URL 来判断两个机构名称字符串是否匹配，并认为相同的 URL 越多，匹配的可能性越大[18-19]。开始是简单统计相同的 URL 数，后有学者引入权重信息，并考虑相同的 URL 在检索结果中出现的排序信息，如 Aumüller 等[19]在利用 Web 判断两个学术机构名称是否匹配时，参考了 TF-IDF 模型，对相同 URL 在检索结果中的排序位置进行加权。

(3) 基于规则的方法的主要思想是根据机构名称构词特点建立一定的规则库，通过规则首先识别出可能匹配的候选名称字符串，然后利用阈值过滤错误匹配对。如 Huang 等[15]，杨波等[20]在利用 WoS(Web of Science)题录数据研究机构名称规范策略中，综合字面相似度、字长、字顺、子串、地区等特征，构建了识别可能匹配的机构名称对规则，然后利用匹配频率过滤错误匹配结果。

(4) 混合方法主要是采用规则、加权统计弥补单纯基于字符串相似度方法的不足。规则的建立基于地区属性，如 Jonnalagadda 等[10]在开展 PubMed 数据库机构名称规范研究中，引入世界地区、邮编字典，借助“如果是同一机构，则对应的地区或邮编应该一致”规则过滤错误匹配项。加权统计主要考虑了词频和机构名称字符串长度对相似度计算影响，如 Onodera 等[21]在借助作者单位地址信息进行作者消歧的机构名称匹配研究中，首先对所有机构名称中出现的词进行词频统计，然后参考 TF-IDF 模型，对每个机构名称中的词进行加权。

本研究属于基于规则和相似度相结合的混合方法，但在规则构建中，除了考虑地区属性，还重点考虑机构类别、机构名称偏正结构特征和关键区分词。

3　研究方案

3.1　方案设计

整个研究方案以生物医学领域中文学术机构名称为例，如图 1 所示。

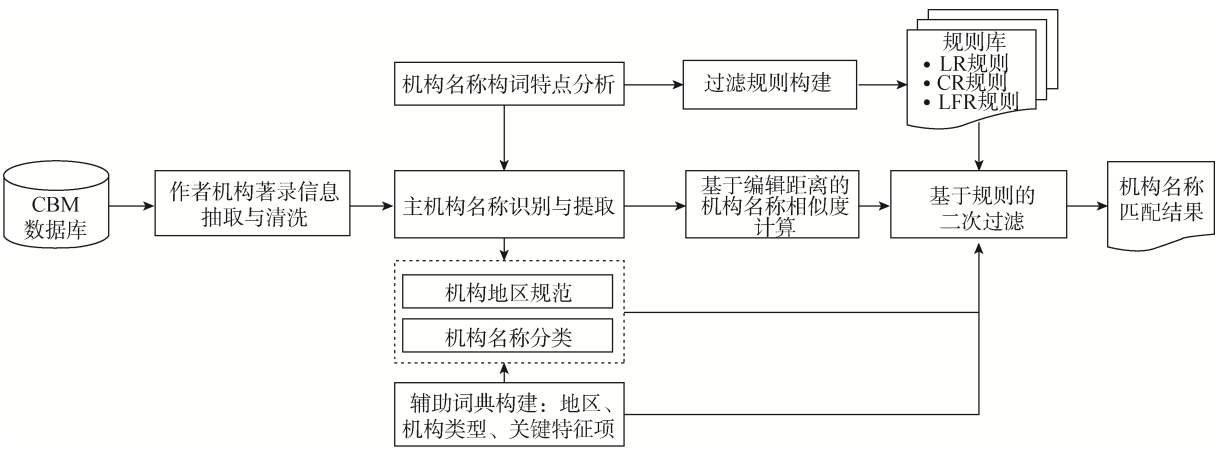

图 1　整体研究思路

(1) 分析中文机构名称字符串构词特征，构建基于机构名称语言学特征的相似度匹配结果过滤规则，并辅助机构名称提取；

(2) 构建地区、机构类别和机构名称关键特征项词典，辅助机构名称地区规范、机构类别分类和过滤规则的应用；

(3) 进行机构名称相似度匹配：即利用 Levenshtein 方法计算源机构名称字符串 Aff_1 和目标机构名称字符串 Aff_2 相似度 $Sim(Aff_1,Aff_2)$，$Sim(Aff_1,Aff_2)$大于既定阈值则认为匹配成功，否则匹配失败；

(4) 利用过滤规则对相似度匹配结果进行二次判断，避免“错误肯定”和“错误否定”。

3.2 学术机构名称构词特点分析与过滤规则构建

(1) 机构名称构词特征分析

从语言学角度来看，学术机构名称是一种偏正式复合名词，由一个或者多个定语加上表示机构称呼的名词性中心语组成[22]。定语可分为前后两部分，前一部分主要为表示地理位置、级别等外部特征信息，如国名、地域名、方位词、人名、序数或基数词、专造词等；后一部分与机构活动内容相关，表达机构所属行业、涉及学科等，如疾病控制、中医药等。中心语则一般用以表明机构性质或类别，如高校、科研院所、医疗机构等，因此在自然语言处理中常被作为类别特征词。通常，不同定语和中心语意味着不同机构实体，这一偏正结构特点构成本研究规则库设计的基础。

由于内在行政隶属、资助、共建、依托等关系，现实中很多机构名称含有多个中心语或类别特征词，如“上海复旦大学附属华山医院”同时含有“大学”和“医院”两个类别特征词，本文称此类机构名称为多类别特征机构名称，相应只含一个中心语的机构名称为单类别特征机构名称。多类别特征机构名称一般具有如下特征：从左到右，不同类别特征项之间存在某种层级关系；机构类别一般由右边出现的类别特征词决定。如“上海复旦大学附属华山医院”的类别为“医院”。这构成从“作者单位”中进行机构名称抽取和分类认知基础。

(2) 过滤规则设计

假设有两个机构名称字符串 Aff_1 和 Aff_2，如果其相似度 $Sim(Aff_1,Aff_2)$大于既定阈值则认为匹配成功，否则匹配失败。规则主要用于过滤错误匹配(False Positives, FP)和避免错误否定 (False Negatives, FN)。因此，研究中将规则分为 FP 过滤规则和 FN 过滤规则两类。

①FP 过滤规则

规则 1(R1) 如果两个机构名称对应的地区不同，那么当前匹配为 FP 匹配。如“武汉普仁医院”和“北京普仁医院”，前者位于“中国–湖北–武汉”，后者位于“中国–北京”。

规则 2(R2) 如果两个机构名称的类别不同，那么当前匹配为 FP 匹配。如“北京医学院”和“北京医院”，前者属高等院校，后者属医院，故不能匹配。

规则 3(R3) 如果两个机构名称字符串都含有序数，但序数不同，那么当前匹配为 FP 匹配。如“济南市第一人民医院”和“济南市第二人民医院”。

规则 4(R4) 如果只有一个机构名称含有序数，那么当前匹配为不确定匹配。如“济南市第一人民医院”和“济南市人民医院”，因为济南市还有“济南市第二人民医院”和“济南市第三人民医院”。实验中作为 FP 匹配处理。

规则5(R5) 如果关键语义特征词的前 2 个字符不同，那么当前匹配为 FP 匹配。如“东南大学”和“南京大学”，虽然地区和类别相同，但关键语义特征词“大学”前 2 个字符不同，所以匹配失败。

规则6(R6) 如果关键语义特征词的后 2 个字符不同，那么当前匹配为 FP 匹配。如“复旦大学附属华山医院”与“复旦大学附属金山医院”，虽然地区和类别相同，但关键语义特征词“附属”后 2 个字符不同，所以匹配失败。

②FN 过滤规则

规则 7(R7) 如果两个机构名词字符串对应的地区和类别完全相同，且其中一个被另一个完整精确包含，那么该匹配为 FN 匹配。如“第一人民医院”和“济南第一人民医院”，若类别和地区相同，因前者被后者完整精确包含，则认为该匹配为 FN 匹配。

根据规则考虑的特征基础，将上述 7 条规则归纳为三类：基于地区的规则(Location based Rule, LR)，基于类别的规则(Category based Rule, CR)和基于语言学特征的规则(Linguistic Features based Rule, LFR)，如表 1 所示。

表 1 规则分类

	FP	FN
LR	R1	
CR	R2	
LFR	R3-R6	R7

3.3 辅助词典构建

本研究构建了三类词典：行政区划名词典、机构

分类表、关键语义特征词典。

(1) 行政区划名词典参考国家行政区划分表构建，含“国家–省–市–区/县”4 级。

(2) 机构分类表依据《卫生机构(组织)分类与代码》标准(WS218–2002)改编，下设医院、医学科学研究机构(简称科研院所)、医学高等教育机构(简称高校)等 18 大类。机构分类数据预处理中，子类共现词构成了一级类目的关键特征词集。本研究聚焦的医院、高等院校、科研院所三类机构部分类别关键特征词示例，如表 2 所示。

表 2　部分类别关键特征词示例

机构分类	关键特征词示例
医院	医院、临床中心、门诊中心、门诊部、…
医学高等教育机构	学院、大学、学校、学部、…
医学科学研究机构	科学院、研究所、研究院、研究中心、创新中心、…

(3) 关键语义特征词典包括行政区划分特征词(如“国、省、市、区、县、街道、乡、镇、村”)，序文特征词(如“第”)，类别特征词(如医院、大学、学院、系、研究所、研究院、科学院、研究中心、创新中心、科室、研究室等，主要来源于类别词典)和关联特征词(如“附属”)。一般来讲，紧挨着关键语义特征词的前后 n 个字符串对同一类别机构名称的实体指向区分起着关键作用。

3.4　过滤规则在机构名称匹配中的应用

机构名称规范文档可理解为机构名称同义词表，由一系列同义词组组成，每一个同义词组由指向同一个机构实体的不同机构别名组成。聚类和分类是机构名称规范构建的主要技术手段，聚类的基本思路和目标是通过相似度计算实现同义机构名称的自动聚堆；分类的基本思路是实现一个新增机构名称到现有机构名称组中的同义映射。本研究的应用场景主要为分类。

对于新增机构名称字符串 aff 和现有机构名称同义词组集合 Aff(aff_1,aff_2,…aff_i…aff_n), Matching(aff, aff_i)用于判定 aff 与 aff_i 是否匹配成功，如果匹配成功，则将 aff 映射到 aff_i 所在的同义机构名组中，否则 Matching(aff, aff_{i+1})，直至所有机构名称匹配完毕。如图 2 所示，每次匹配中，当 Sim(aff,aff_i)大于等于设定阈值时，依次调用规则 1–规则 6，如通过所有规则，则匹配成功，否则匹配失败；当 Sim(aff,aff_i)小于设定阈值时，调用规则 7，如规则通过，匹配成功，否则匹配失败。

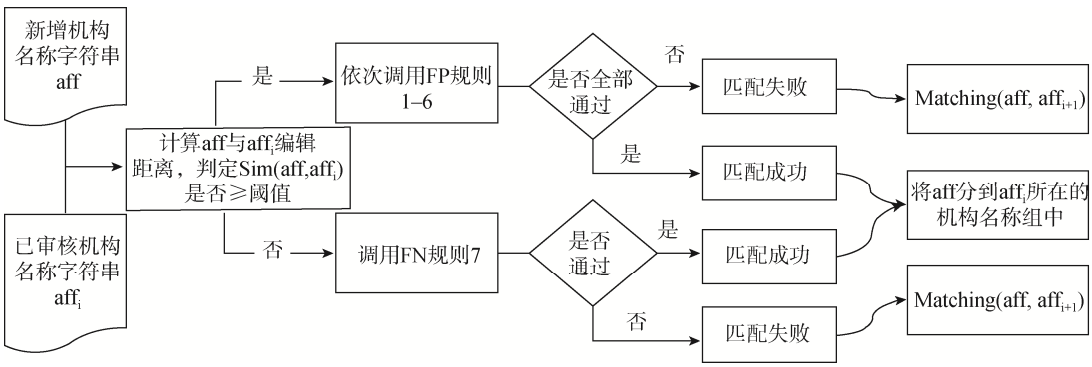

图 2　机构名称匹配算法流程

4　实验与结果分析

4.1　实验方案

(1) 实验目标

地区、邮编已被用于证明能够提升机构名称字符串匹配效果。本实验目标主要有三点：

①验证机构类别和关键特征词是否能够被用以进一步提升机构名称字符串匹配的准确率；

②分析数据集规模大小对算法效果的影响；

③分析不同类别机构名称对方案的敏感性。

(2) 实验数据选择与抽取

以中文生物医学文献数据库 CBM[23]中作者机构著录数据为基础开展研究。CBM 是由中国医学科学院医学信息研究所研制的中国第一个生物医学期刊文献数据库，收录了 1978 年以来 1 800 余种中国生物医学期刊全部摘要信息，是生物医学领域开展查新查引、科技评价等活动的核心数据库之一，于 2013 年开始提供机构检索与分析服务。

不同类别机构实体在医学科学研究活动中功能定

位有所不同, 对科技文献的贡献也不一致。据 CBM 2006 年–2011 年数据统计, 88%文献来源于高等院校、医院与科研院所, 因此本研究重点聚焦于这三类学术机构。

CBM于2011年启动机构名称规范化项目, 目前已完成 2006 年–2011 年期间收录的中文生物医学期刊论文中上述三类中文作者机构名称规范化工作, 形成规范文档, 且所有规范成果均经过人工审核。鉴于此, 本实验抽取 2006 年–2011 年期间 2 842 974 篇中文期刊论文中的“作者单位”项作为本研究数据基础。

(3) 数据预处理

数据预处理主要包括三个环节: 原始机构著录字符串拆分、机构名称字符串抽取与分类、机构名称地区规范。

①原始机构著录字符串拆分

原始机构著录字符串指未经任何处理的作者发文时提交的机构信息字符串。原始机构著录字符串拆分包括两个步骤: 第一步, 多机构拆分。将多机构记录拆成单机构记录。因机构合作客观存在, 很多“作者机构”含有 2 个或 2 个以上机构信息, 简称多机构记录(对应的称为单机构记录), 如“(1)昆山市第一人民医院肿瘤科,江苏昆山 215300; (2)苏州大学医学院分子生物学教研室,江苏苏州 215000”, 不同机构通过分号和序号间隔。第二步, 单机构拆分。将所有原始单机构记录拆分成“机构”、“地址”和“邮编”三个子串, 无则为空。“作者机构”字符串一般由机构名、地区(省市名)、邮编(6 位数字)三项或者某些项组成。6 类常见作者机构字符串组成结构, 如表 3 所示。

表 3 作者机构名称字符串常见组成结构

序号	作者机构字符串常见结构	示例
1	‘机构’+ ‘逗号’+ ‘省份名城市名’+‘邮编’	昆山市第一人民医院肿瘤科, 江苏昆山 215300
2	‘机构’+ ‘逗号’+‘城市名’+‘邮编’	上海复旦大学附属华山医院神外科, 上海 200040
3	‘机构’+‘逗号’+ ‘省份名’+‘邮编’	昆山市第一人民医院, 江苏省 215300
4	‘机构’+‘逗号’+ ‘邮编’	江苏省南通大学附属肿瘤医院, 226361
5	‘机构’+ ‘邮编’	江苏省南通大学附属肿瘤医院 226361
6	‘机构’	安徽医科大学第一附属医院消化内科

②机构名称字符串抽取与分类

不同期刊对作者机构名称著录规范性要求所有不同, 著录深度不一致则是其中一种表现。拆分后的很多“机构”子串含有科室、院系、研究室等二级或者三级机构名。本研究中的机构名称泛指指向一个独立法人机构名。预处理中, 结合机构名长度特点, 首先利用“机构子串长度不能小于 4”进行无意义机构名称字符串过滤, 然后依据类别特征词字典进行机构名抽取和分类。主要步骤如下:

1)机构类别特征词标注。根据类别特征词表对“机构名称信息子串”进行机构类别特征词标注, 并按自左向右出现的顺序进行记录, 如“大学–医院”, “大学–学院”、“学院–医院”、“科学院–研究所”等。并根据类别特征词出现个数, 将所有机构分为单类别机构名称和多类别机构名称。

2)机构名抽取。对于单类别机构子串, 自左向右扫描, 截取类别特征词及其左侧字符串部分作为机构名称; 对于多类别机构子串, 非“大学–学院”类, 自左向右扫描, 截取第二个类别特征词及其左侧字符串部分作为机构名; “大学–学院”类截取“大学”及其左侧的字符串作为机构名。

3)机构名称分类。对机构名进行自右向左扫描, 根据第一个出现的类别特征词, 将机构名称纳入“医院”、“医学高等院校”和“医学科研院所”。

③机构名称地区规范

如表 3 所示, 很多作者机构著录项无地区和邮编信息。研究中, 对于第 2–5 类情况, 根据邮编和行政区划名词典自动获取“省–市”; 对于第 6 种既无地区也无邮编类, 同时应用以下两个条件自动获取: 等同或包含当前机构名称的作者机构著录项; 至少含有 1 个共现作者。当获得多个地区时, 取作者共现数最高者。依然无法获取地区的记录直接剔除。此时, 共获得 6 076 447 条有意义机构记录, 每条记录均含有“中文机构名”、“国家–省–市”和机构类别。

(4) 测试数据集构建

根据CBM收录年份范围, 构建三组测试集, 每组测试集由两部分组成: 基础数据集和新增数据集。基础数据集中所有机构名称均已按人工审核结果规范完毕, 即指向同一个机构实体的不同机构名称字符串已形成一个同义词组。新增数据集合中的机构名称均未在对应基础数据集中出现过, 需根据本研究提出的匹配策略, 分入基础数据集中目标机构名称同义词组中。所有测试集数据均根据“机构名称”和“国家–省–市”进行组内去重。去重后各组测试集统计结果如表 4 所示。

表 4　测试数据集统计

测试数据集分组	基础数据集合				新增数据集合			
	序号	CBM 收录年份范围	机构类别	去重后机构名称串	序号	CBM 收录年份范围	机构类别	去重后机构名称串
第一组 (T1)	TBD1	2006–2008	高等院校	22 685	TID1	2009–2011	高等院校	10 192
			研究所	11 178			研究所	5 182
			医院	93 895			医院	59 937
			合计	127 758			合计	75 311
第二组 (T2)	TBD2	2006–2009	高等院校	26 943	TID 2	2010–2011	高等院校	5 932
			研究所	13 195			研究所	3 165
			医院	113 554			医院	40 281
			合计	153 692			合计	49 378
第三组 (T2)	TBD3	2006–2010	高等院校	31 014	TID3	2011	高等院校	1 862
			研究所	15 051			研究所	1 313
			医院	133 003			医院	20 833
			合计	179 068			合计	24 008

(5) 实验流程

按图 2 所示匹配算法流程, 实现三组测试数据中新增机构名字符串到基础数据集的匹配分类。

相似度阈值的设置是相似度匹配研究另一关键问题, 非本研究重点, 本研究相似度阈值均设为 0.8。

为对比分析不同特征因素对匹配效果的影响, 将三类规则组合成 4 个方案进行实验: LR、LR+CR、LR+LFR 和 LR+CR+LFR。

(6) 评价方法与指标

将实验结果和人工审核判定结果进行对比, 并引入准确率(Precision, P)、召回率 R(Recall, R)和 F 值(F-Measure)三个经典指标。准确率 P 表示正确匹配的机构名称数与全部匹配的机构名称数的比值, 召回率 R 表示正确匹配的机构名称数与应该匹配的机构名称数的比值。

设 A 表示全部匹配的机构名称集合, B 表示正确匹配的机构名称集合, C 表示所有经人工审核后应该匹配上的机构名称集合, 相关计算公式如下。

$$P = \frac{|B|}{|A|} \tag{1}$$

$$R = \frac{|B|}{|C|} \tag{2}$$

$$F = \frac{2 \times P \times R}{P + R} \tag{3}$$

4.2　实验结果与分析

(1) 不同规则组合方案效果对比分析

4 组规则组合方案在三个测试集下的实验评估结果如表 5 所示。其中 C_1 表示“LR”混合; C_2 表示“LR+CR”混合; C_3 表示“LR+LFR”混合; C_4 表示“LR+CR+LFR”混合。

表 5　不同组合方案在 3 个测试集中实验评估结果

方案	T1			T2			T3		
	P	R	F 值	P	R	F 值	P	R	F 值
C_1	71.15%	**62.26%**	**66.41%**	72.68%	**68.66%**	**70.62%**	72.79%	**74.37%**	**73.57%**
C_2	71.23%	60.80%	65.60%	72.23%	66.82%	69.42%	72.45%	72.92%	72.69%
C_3	80.72%	53.29%	64.20%	**80.56%**	59.22%	68.26%	**80.11%**	64.82%	71.66%
C_4	**80.77%**	51.10%	62.59%	80.46%	57.20%	66.86%	80.00%	63.17%	70.59%

从表 5 可以看出, 就准确率而言, 在三组测试数据中, 第一组测试集中 C_4 表现最佳, C_3 紧随其后, 分别高于 C_1 9.62%和 9.57%; 第 2、3 组测试集中 C_3 表现最佳, 与 C_1 相比, 分别提高 7.88%和 7.32%。就召回率和 F 值而言, C_1 在三组数据中均表现最佳, C_4 均表现最差, C_2 和 C_3 居中。

这表明，与只引入地区特征匹配策略相比：加入机构类别特征，并未能够提升机构名称字符串的匹配效果，准确率、召回率和 F 值均下降；加入机构名称构词特征，能够显著提高机构名称字符串的匹配准确率，但召回率和 F 值出现下降。

(2) 基础数据集规模对算法效果影响分析

图 3 是方案 C1-C4 在三个测试集中的准确率、召回率和 F 值及其变化情况。可以看出，随着基础规范数据集增大，准确率有升有降，但幅度较小，只有一个变幅为 1.53%，其他均在 0.5%以内。这说明 4 组算法的准确率均具有较好的稳定性；召回率和 F 值均持续提升，且变化提升幅度较大。这在一定程度上表明，低召回率可以通过扩大基础数据集规模加以改善。

(3) 不同类别机构名称字符串适应性分析

重点分析高等院校、科研院所、医院三类学术机构名称对不同规则组合敏感性：哪种组合方案准确率最佳；随着基础数据集规模增大，准确率最佳方案对应的召回率和 F 值是否呈上升趋势。

表 6 是不同组合方案在不同类别机构名称测试子集中准确率实验结果：高等院校类，第一组测试子集 C_4 表现最佳，C_3 紧随其后，仅相差 0.01%；第二、三组测试子集 C_3 表现最佳，C_4 紧随其后，优于 C_1 和 C_2。科研院所类，三组测试子集 C_3 均表现最佳，分别为 84.91%、85.41%、84.83%，显著优于其他方案。医院类，第一、二组测试子集 C_3 表现最佳，分别为 81.75%、81.45%；第三组测试子集 C_4 表现最佳，80.94%，C_3 紧随其后，80.79%，均显著优于 C_1 和 C_2。

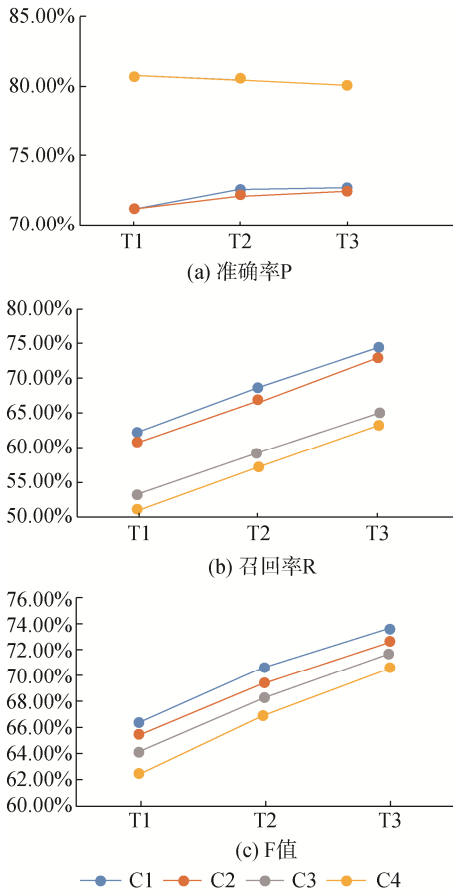

图 3　不同方案在三个测试集下效果变化趋势

表 6　不同组合方案在不同类别子集中准确率评估结果

方案	高等院校			科研院所			医院		
	T1	T2	T3	T1	T2	T3	T1	T2	T3
C_1	72.55%	72.37%	68.84%	79.51%	79.64%	77.94%	71.05%	72.33%	72.85
C_2	72.35%	71.86%	67.70%	77.33%	78.79%	77.08%	71.05%	72.25%	72.77%
C_3	74.43%	**74.24%**	**70.41%**	**84.91%**	**85.41%**	**84.83%**	**81.75%**	**81.45%**	80.79%
C_4	**74.44%**	74.00%	69.51%	77.46%	79.57%	77.27%	81.61%	81.25%	**80.94%**

由表 6 可知，就准确率而言，对于高等院校、科研院所、医院三类机构名称中，综合考虑地区特征和机构名称构词特征的混合策略表现均优于只考虑地区特征的混合组合方案，其中科研院所和医院类更为显著。

图 4 是 C_3 在高等院校、科研院所、医院不同测试子集中的召回率及其变化趋势。可以看出：医院类召回率最高，高等院校次之；随着基础数据集规模的增大，三类机构的召回率均呈现上升趋势。

图 5 是 C_3 在高等院校、科研院所、医院不同测试子集中的 F 值及其变化趋势。可以看出：医院类 F 值最高，高等院校次之；随着基础数据集规模的增大，科研院所和医院类均呈上升趋势，高校类机构先升后降，幅度均较小。

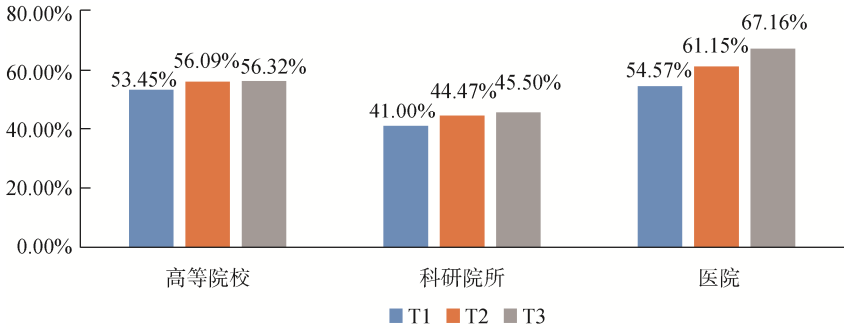

图 4　组合方案 C_3 在三类机构名称不同测试子集中召回率及其变化情况

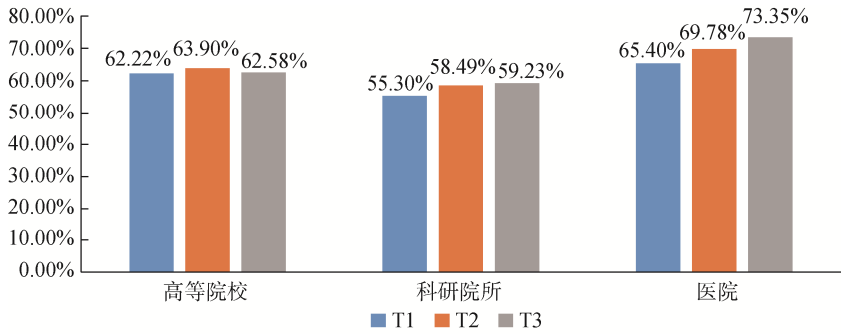

图 5　组合方案 C_3 在三类机构名称不同测试子集中 F 值及其变化情况

(4) 小　结

综上分析可得，无论是整体测评，还是分类测评，就准确率而言，综合考虑机构地区、机构名称构词特征混合策略表现最佳；就召回率而言，只考虑机构地区混合策略表现最佳。

对此，笔者认为，在机构名称规范工程中，高的准确率要比高的召回率更重要，因为低的召回率可以通过扩大基础数据集规模加以改善。上述基础数据集规模对算法效果的影响分析结果(图 3–图 5)也证明，随着基础数据集规模的扩大，综合考虑机构地区、机构名称构词特征的召回率和 F 值呈现显著上升趋势。

5　结　论

本文在现有研究引入地区特征基础上，尝试进一步从机构名称类别和字符串构成的关键语义特征项来提升传统基于编辑距离的机构名称匹配效果，并以中文生物医学领域权威数据库 CBM 中“作者单位”为数据基础，开展实证研究，证明机构名称构词特征能够有效辅助地区特征提高不同类别机构名称的匹配效果，且具有较好的稳定性。

辅助词典和规则的人工构建是本研究得以实现的重要因素，也是研究本身固有的局限性，限制了匹配结果准确率和召回率的进一步提升，后续将探讨机器学习在辅助词典和规则构建中的应用。

参考文献:

[1] Khalid M A, Jijkoun V, De Rijke M. The Impact of Named Entity Normalization on Information Retrieval for Question Answering[C]//Proceeding of the IR Research, 30th European Conference on Advances in Information Retrieval，Glasgow, UK. Berlin, Heidelberg: Springer-Verlag, 2008: 705-710.

[2] 唐金玲. 国际三大检索系统论文作者机构名称问题研究——以高校机构名称为例[J]. 情报探索, 2014(9): 80-84. (Tang Jinling. Study on Issues of Author Affiliations on Papers Included in International Three Key Retrieval Systems: Case Study of Name of University[J]. Information Research, 2014(9): 80-84.)

[3] 苏新宁. 图书馆、情报与文献学学术影响力研究报告(2000–2004)——基于 CSSCI 的分析[J]. 情报学报, 2006, 25(2): 131-153. (Su Xinning. Report on Academic Influence in Library, Information and Documentation Science (2000–2004) [J]. Journal of the China Society for Scientific

and Technical Information, 2006, 25(2): 131-153.)

[4] 曾建勋, 王立学. 面向知识评价的规范文档建设方法[J]. 图书情报工作, 2012, 56(10): 101-106. (Zeng Jianxun, Wang Lixue. Construction of Knowledge Evaluation-oriented Authority Files[J]. Library and Information Service, 2012, 56(10): 101-106.)

[5] Abramo G, D'Angelo C A, Pugini F. The Measurement of Italian Universities' Research Productivity by a Non Parametric-Bibliometric Methodology[J]. Scientometrics, 2008, 76(2): 225-244.

[6] French J C, Powell A L, Schulman E. Automating the Construction of Authority Files in Digital Libraries: A Case Study[C]//Proceedings of International Conference on Theory and Practice of Digital Libraries.Berlin,Heidelberg: Springer, 1997: 55-71.

[7] Liu W L, Doğan R I, Sun K, et al. Author Name Disambiguation for PubMed [J]. Journal of the Association for Information Science and Technology, 2014, 65(4): 765-781.

[8] 孙海霞, 李军莲. 学术论文作者机构规范文档构建[J]. 医学信息学杂志, 2015, 36(11): 42-47. (Sun Haixia, Li Junlian. Construction of Authority File of Author Affiliations[J]. Journal of Medical Informatics, 2015, 36(11): 42-47.)

[9] 陈金星, 祝忠明. 责任者名称规范控制研究及进展[J]. 现代图书情报技术, 2009(12): 12-17. (Chen Jinxing, Zhu Zhongming. Research Progress of the Name Authority Control for the Contributor[J]. New Technology of Library and Information Service, 2009(12): 12-17.)

[10] Jonnalagadda S R, Topham P. NEMO: Extraction and Normalization of Organization Names from PubMed Affiliation String[J]. Journal of Biomedical Discovery and Collaboration, 2010, 5(1): 50-75.

[11] Jiang Y, Zheng H T, Wang X, et al. Affiliation Disambiguation for Constructing Semantic Digital Libraries[J]. Journal of the American Society for Information Science and Technology, 2011, 62(6): 1029-1041.

[12] 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.

[13] Cuxac P, Lamirel J C, Bonvallot V. Efficient Supervised and Semi-Supervised Approaches for Affiliations Disambiguation[J]. Scientometrics, 2013, 97(1): 47-58.

[14] French J C, Powell A L, Schulman E. Using Clustering Strategies for Creating Authority Files[J]. Journal of the American Society for Information Science, 2000, 51(8): 774-786.

[15] Huang S, Yang B, Yan S, et al. Institution Name Disambiguation for Research Assessment[J]. Scientometrics, 2014, 99(3): 823-838.

[16] 孙海霞, 成颖. 信息集成中的字符串匹配技术研究[J]. 现代图书情报技术, 2007(7): 22-26. (Sun Haixia, Cheng Ying. Study on String-based Matching of Information Intergration[J]. New Technology of Library and Information Service, 2007(7): 22-26.)

[17] Jacob F, Javed F, Zhao M, et al. sCooL: A System for Academic Institution Name Normalization[C]//Proceeding of 2014 International Conference on Collaboration Technologies & Systems.IEEE, 2014: 86-93.

[18] Bollegala D, Ishizuka M, Matsuo Y. Measuring Ssemantic Similarity Between Words Using Web Search Engines[C]// Proceeding of the 14th International Conference on World Wide Web. 2007: 757-766.

[19] Aumüller D, Rahm E. Web-based Affiliation Matching[C]// Proceeding of International Conference on Information Quality. DBLP, 2009: 246-256.

[20] 杨波, 杨军威, 阎素兰. 基于规则的机构名称规范化研究[J]. 现代图书情报技术, 2015(6): 57-63. (Yang Bo, Yang Junwei, Yan Sulan. Research on Rule-based Normalization of Institution Name[J]. New Technology of Library and Information Service, 2015(6): 57-63.)

[21] Onodera N, Iwasawa M, Midorikawa N, et al. A Method for Eliminating Articles by Homonymous Authors from the Large Number of Articles Retrieved by Author Search[J]. Journal of the American Society for Information Science and Technology, 2011, 62(4): 677-690.

[22] 张小衡, 王玲玲. 中文机构名称的识别与分析[J]. 中文信息学报, 1997, 11(4): 21-32. (Zhang Xiaoheng, Wang Lingling. Identification and Analysis of Chinese Organization and Institution Names[J]. Journal of Chinese Information Processing, 1997, 11(4): 21-32.)

[23] 中国生物医学文献数据库 [EB/OL]. [2017-10-30]. http://www.sinomed.ac.cn/zh.(SinoMed[EB/OL]. [2017-10-30]. http://www.sinomed.ac.cn/zh.)

作者贡献声明:

孙海霞: 提出研究思路, 设计研究方案, 算法设计, 论文起草与最终版本修订;

王蕾: 辅助字典构建, 进行实验;

吴英杰: 辅助字典构建, 采集、清洗和分析数据;

华薇娜: 设计研究方案;

李军莲: 部分算法设计和算法评测。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: sun.haixia@imicams.ac.cn。

[1] 孙海霞，吴英杰．source_data.accdb．CBM 著者机构原始数据．

[2] 孙海霞，吴英杰．source_pre-process_data.accdb．清洗后的 CBM 著者机构数据．

[3] 王蕾. Levenstein.py, Matching_Process.py. 算法实现程序．

[4] 孙海霞，王蕾. result_data.accdb．匹配结果数据．

收稿日期: 2018-02-11
收修改稿日期: 2018-03-26

Matching Strategies for Institution Names in Literature Database

Sun Haixia[1,2]　Wang Lei[2]　Wu Yingjie[2]　Hua Weina[1]　Li Junlian[2]
[1](School of Information Management, Nanjing University, Nanjing 210093, China)
[2](Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100020, China)

Abstract: [Objective] This paper designs and implements matching strategies for institution names in literature database, aiming to regulate their storage and management. **[Methods]** We first established seven name matching rules based on their regions, types and naming characteristics. Then, we designed four hybrid matching strategies combining rules and Levenstein distance. Finally, we evaluated the four hybrid strategies with institution names from the papers indexed by Chinese Biomedical Literature (CBM) database during 2006-2011. **[Results]** More than six million affiliation strings from CBM were matched, which included higher education institutions, hospitals and research institutes. We found that the hybrid matching strategy based on region, naming characteristics and Levenstein distance obtained the highest precision (all above 80%), recall (64.82%), and F-value (71.66%). **[Limitations]** The rules and related dictionary were mainly constructed with human experience and their coverage is limited. There are some errors in the identifying institution names. The proposed strategy cannot address the issues caused by the transformative actions of institutions. **[Conclusions]** The proposed strategies could improve the performance of scientific research literature databases.

Keywords: Information Retrieval　Normalization of Affiliation Strings　Similarity Measure　Hybrid Strategy　Literature Database