情报学报　2020 年 6 月　第 39 卷　第 6 期
Journal of the China Society for Scientific and Technical Information, Jun. 2020, 39(6): 565-578

DOI: 10.3772/j.issn.1000-0135.2020.06.001

基于机构-作者向量的科研机构名称演化识别方法研究

吕冬晴[1]，陆红如[1]，成　颖[1,2]，孙海霞[1,3]

（1. 南京大学信息管理学院，南京　210023；2. 山东师范大学文学院，济南　250014；
3. 中国医学科学院医学信息研究所，北京　100020）

摘　要　机构变迁是引起科研机构名称演化的重要原因。消解科研机构名称的异质性可以提高信息检索的查全率以及科学计量的信度，为此，本文提出了基于科研机构中人员在短期内相对稳定特征的名称演化识别方法。本文构建了机构-作者向量与机构-年度向量，通过综合机构-作者向量的相似度、作者绝对共现量以及1:1、n:1、1:n以及n:m名称映射关系对更名、合并、拆分与重组关系进行了识别；借鉴主成分分析法中的因子识别方法并结合前述4种演化关系，提出了动态相似度阈值设定方法。实验数据采集自CSSCI数据库1999—2015年的论文，实验环节考虑了人员流动以及重名风险对结果的可能影响。结果表明，本研究提出的科研机构名称演化识别方法在准确率与召回率上均有优异的表现。

关键词　机构名称；规范化；更名；作者相似度

A Method for Institution Name Normalization Based on Institution-Author Vectors

Lyu Dongqing[1], Lu Hongru[1], Cheng Ying[1,2] and Sun Haixia[1,3]

(1. *School of Information Management, Nanjing University, Nanjing*　210023;
2. *School of Chinese Language and Literature, Shandong Normal University, Jinan*　250014;
3. *Institution of Medical Information, Chinese Academy of Medical Sciences, Beijing*　100020)

Abstract: Institution transition is one reason behind variety in institution names. Normalization of institution names benefits both information retrieval recall and the reliability of bibliometric research results. Thus, this paper proposes a method for institution name normalization based on the stable feature of personnel in an academic institution in the short term. Specifically, institution-author and institution-annual vectors are constructed for each academic institution, and the similarity of the integrated institution-author vectors, the number of co-authors, and mapping rules are used to identify transition relationships between two institutions, including renaming, merger, split, and reorganization. The method was tested using data from the CSSCI database between 1999 and 2015. After controlling for the impact of personnel turnover and homonymous authors, the proposed method demonstrated excellent performance in both accuracy and recall.

Key words: institution name; disambiguation; change of institution name; author similarity

收稿日期：2019-06-06；**修回日期**：2019-12-10

基金项目：国家社会科学基金项目“施引者引用意向与文献计量视角的学术论文被引影响因素研究”（17BTQ014）。

作者简介：吕冬晴，女，1994年生，博士研究生，主要研究方向为信息检索、信息计量；陆红如，女，1993年生，博士研究生，主要研究方向为信息行为；成颖，男，1971年生，博士，教授，博士生导师，主要研究方向为用户信息行为、信息检索；孙海霞，女，1984年生，博士研究生，助理研究员，主要研究方向为知识组织、信息行为，E-mail：sun.haixia@imicams.ac.cn。

1 引 言

在信息检索、文献计量以及科学评价等领域，机构特征具有独特的价值。规范化的科研机构名称可以提高信息检索的查全率与查准率，增强科学计量以及科研评价结果的可靠性。De Bruin 等[1]发现科学引文索引（SCI）等数据库中存在作者机构信息不一致的现象，并指出该现象弱化了数据的可用性与有用性。学术数据库中机构信息的异质性主要源于：①作者写作习惯的不同，如使用机构简称、别称或者混用一级二级机构等；②机构更名、合并、拆分以及重组等引起的名称变化；③偶发的拼写或输入错误；④尚无约定俗成的译名等。这些因素增加了检索的复杂性[2]，损害了计量或评价结果的信度[3]。对此，学界开展了机构规范文档（authority file）研究[4]。

机构规范文档要求将同一机构实体的所有名称及其他属性有序集中[4]，以达到唯一性与稳定性[5]的要求。研究要解决的关键问题是消除各种原因引起的机构名称变体间的歧义，并将其与机构实体建立精准映射[6]，从而在信息检索等应用中实现该实体所有信息的快速定位。现有的机构规范化研究集中于作者写作习惯、译名不一致以及拼写错误，部分涉及机构变迁中的更名现象，尚未见围绕机构合并以及拆分等引起的名称使用不规范现象的系统研究。研究中多数学者直接从机构名称入手，通过计算相似度或者建立匹配规则等方法力图实现名称归一化；贾君枝等[7]从机构间作者集合的共现视角提取了机构的更名关系。

对于一定规模的科研机构来说，短期内其人员除了少量的退休、调出以及新引进之外，会大体保持稳定。科研机构人员的相对稳定性这一特征可应用于更名、合并和拆分等机构变迁引起的名称变体归一化研究。本文拟基于该特征，综合考虑不同变迁形式特有的名称映射关系全面探测科研机构名称的演化：通过构建科研机构-作者向量，以该向量的相似度作为提取机构变迁关系的观测指标；科研机构的不同变迁类型具有特定的时间分布，据此构建各机构的机构-年度向量用于归并邻近年份的作者向量以减少人员流动对作者相似度计算的影响；结合相似度与映射规则识别机构间的更名、合并与拆分等关系；借鉴主成分分析法中的因子识别方法设置相似度阈值；同时，考虑到机构规模是判断重名风险效应量的关键因素，辅之以作者绝对共现量进一步精准识别结果集。

2 文献综述

王星等[4]将机构规范文档定义为“涉及机构名称的规范化描述、识别、聚类、应用等多方面的综合性记录文档”，其规范化过程主要包含两个环节：①在科学论文的“作者单位”著录项（包括作者名、单位名以及所在省市与邮编）中完成机构实体识别[8-9]；②实现同一机构实体不同名称间的同义汇聚[10]，即科研机构“归一化”。对于前者，主要采用人工构建规则、机器学习以及深度学习等命名实体识别（named entity recognition，NER）技术展开研究，现已取得了阶段性的成果，F1 值徘徊于 80%~90%[11]。对于后者，曾建勋等[12]将其细化为“同一”、“隶属”、“相继”以及“相关”四种关系。现有研究主要致力于解决由偶发错误、机构译名、作者写作习惯以及机构变迁等产生的机构名称多元问题。

2.1 错误名称规范化

该类问题主要涉及拼写与编辑过程中出现的偶发错误、翻译不规范引起的机构名称变体，在外文语料中尤为突出。消解该类异质性最常用的方法是以编辑距离为代表的机构名称字符串的字面相似度匹配策略，即当机构名称变体之间的相似度大于既定阈值，则认为其表征同一机构实体。比如，French 等[2]使用天体物理学数据库收录的书目信息为数据源，提出了 3 种基于编辑距离的聚类方法：①对机构的缩略词进行扩展，计算相似度后采用固定和可变阈值进行聚类；②为解决字符顺序与重复字符造成的识别误差，将机构名称中唯一出现的单词按照字典顺序重新排列后再聚类；③第②步无法解决由于近义词的首字母不同而导致的编辑距离的增加，为此提出了一种基于“词”的聚类方法。Jonnalagadda 等[6]基于 PubMed 构建了 NEMO（Normalization Engine for Matching Organizations）系统，该系统利用全局与局部两类序列对齐算法对机构名称中的字符进行平移，并使用字（word）层面的编辑距离计算相似度以解决由于翻译、拼写或者缩写造成的“同义”问题。

在方法层面还有学者利用语法规则以及概率统计等方法对拼写错误等进行识别与校正。例如，Nguyen 等[13]以越语推特文本为数据集，通过构建字典以识别原始文本中出现的拼写等错误，依据单词结构与音节规则对其进行规范化，并采用 Dice 系数

与修正的 fDice 系数计算校正前后句子的相似度。Cuxac 等[14]致力于解决拼写错误、印刷错误、缩写或者遗漏等名称变体，提出了基于朴素贝叶斯模型的有监督作者所属机构消歧算法；同时，为了减少人力成本，还提出了一种适用于训练数据集受限的半监督学习方法，实验结果表明，两类方法在召回率、精度与 F 值等指标上都有良好表现。

2.2　机构名称归一化

该问题聚焦于机构的别称、简称以及机构变迁过程中因更名、合并与拆分等出现的历史名称的归一化，现有研究大体上可以分为两类。

（1）多个维度或复杂相似性的综合应用。孙海霞等[15]结合编辑距离和基于词集合的语义相似度算法，以中文生物医学文献数据库 CBM（China Biology Medicine）中收录的 1994—2010 年上海市医院类原始机构名称作为实验语料，采用 K-means 算法实现了同一科研机构不同名称的聚类，建立了机构规范名-别名映射表。Onodera 等[16]采用统计加权共现词频的相似度算法识别了机构的多个名称，该算法先统计机构信息中的关键词词频，并进行逆向加权，随之采用机构对之间共现词的总权重计算其名称相似度。Jiang 等[17]则选择归一化压缩距离（normalized compression distance，NCD）计算两个机构名称相似度并进行聚类分析，其以清华大学图书馆收录的 10000 篇论文中的机构信息为数据集，比较了不同的压缩机制在平均精度、F 值、熵以及纯度 4 个评价指标上的表现，发现 NCD 方法相较于 K-means 具有更好的识别效果。

（2）基于规则。即通过人工构建筛选规则，逐步缩小同一机构名称的集合，进而确定指向同一机构实体的名称集合。多项研究的基本思路是将相似度与规则相结合，其中规则大多数依赖于地区、邮政编码、机构类型等附加属性以及机构名称关键词的语言学特征等。例如，Huang 等[18]分析来自同一机构的作者信息，以同名作者为切入点，对机构进行分区，使用机构名称关键词、地区、邮政编码特征建立了 6 组规则，按照区块依次识别组内候选“机构实体对”，最后筛选出频次大于阈值的“机构对”，从而得到指向同一机构实体的名称集合。杨波等[19]在 Huang 等[18]工作的基础上，利用 TF-IDF 加权改进了识别效果。孙海霞等[10]采用 Levenshtein 方法计算机构名称的相似度；将相似度大于或小于阈值的结果分别利用基于机构地区、机构类型以及语言学特征构建的两类规则进行二次判定，实证发现上述三类特征构建的规则可以提高识别的准确率，仅使用机构地区特征的规则在召回率上表现更优。

多数研究者在解释机构实体名称多样化的缘由时都曾提及机构变迁引起的名称演化，但由于更名、合并与拆分往往会造成机构名称在语义、语法或者地区特征等方面的剧烈变化，从而导致基于机构名称相似度的识别方法在该类名称异质性的消歧问题上具有较大的局限性。对此，贾君枝等[7]以《图书情报工作》等 3 本期刊于 2006—2016 年所刊论文的作者单位信息作为数据源，基于各机构作者集合的人员共现率提取两个机构间的更名关系，并以 15% 作为共现率阈值成功识别了“华北工学院”与“中北大学”这一更名关系。

2.3　述　评

学界已充分认识到科研机构名称归一化的价值。现有研究主题多涉及别名、简称以及拼写错误等名称变体的规范化，尚未见对机构变迁中的更名、合并与拆分等名称演化问题的系统研究。研究方法多以基于机构名称字符串相似度的聚类算法为主，辅以地区、邮编以及机构名称的语法和/或语义属性进行优化。不过，基于规则的机构名称匹配方法在一定程度上存在领域受限的不足，而不同类型的组织机构在名称构建上各具特色，故规则通常难以移植，不具有普适性。综上，围绕机构名称特征开展的规范化研究存在较大的局限性，后继研究有必要在此基础上结合更丰富的特征开展。贾君枝等[7]的研究初步证实了科研机构的人员特征对更名关系识别的有效性，为该主题的工作提供了有益的思路。

3　研究方法

3.1　总体思路

机构变迁前后的名称演化关系可归纳为 4 种映射关系：①更名，名称表现为 1:1 映射，如“徐州师范大学”更名为“江苏师范大学”；②拆分，名称表现为 1:n 映射，该类的例子近期鲜见；③合并，名称表现为 n:1 映射，如“中南工业大学”、“湖南医科大学”以及“长沙铁道学院”合并为“中南大学”；④重组，名称表现为 n:m 关系。重组关系所蕴含的机构变迁模式较为复杂，可能同时存在更名、合并与拆分等情况，如武汉水利电力大学拆分后主体部分组建现武汉大学工学部，宜昌校区与湖

北三峡学院合并组建三峡大学。

根据刘进等[20]的调查，中国研究型大学10年间教师的总体流动频率仅为15.9%。该数据表明，一定规模的科研机构，其人员除了少量的离职、退休、调出以及新引进之外，会在相对较短的时期内保持大体稳定。该特征可应用于由机构变迁引起的机构名称演化研究，具体可通过计算机构间的人员相似度实现，如机构更名前后，其人员组成变动较小；在机构拆分中，原机构人员会随之分配到新的机构中，即原机构的部分（如院系）拆分进入新机构，但该部分基本稳定。贾君枝等[7]的工作为该思路的可行性提供了证据。此外，本研究还发现具有特定变迁关系的机构其人员在时间分布上存在特定的规律，因此可以从人员与时间两个维度提取机构的关系。

由此，本研究拟以CSSCI 1999—2015年的论文为数据源，依次构建机构-作者向量与机构-年度向量；依据更名、合并与拆分等特定的机构变迁形式所特有的名称映射关系与时间分布规律筛选潜在的“匹配机构对”；通过归并邻近三年的机构-作者向量以减少人员流动对实验结果的影响，并进行相似度计算；最后依据机构-作者向量相似度、名称映射关系以及作者绝对共现量三个指标识别更名、合并与拆分等变迁关系（图1）。考虑到“经过科研管理部门人工加工和维护机构信息在权威性和准确性方面有一定优势”[19]，得益于期刊的审稿等守门人机制，机构名称中存在拼写错误等发生的概率极低，本文暂不考虑拼写错误对研究的影响。

3.2 机构变迁关系识别

3.2.1 机构-作者向量

对于每个科研机构名称，构建该名称对应的机构-作者向量（以CSSCI 1999—2015年数据为例，其他数据库亦同），具体如下。

（1）识别CSSCI期刊论文中的所有第一作者、机构名称、发表年份，形成<年份, 机构, 作者>三元组。

（2）对三元组进行去重，即若某一记录中，3个属性都相同，则将重复记录删除，仅保留唯一值。该过程中，机构内可能存在重名作者，由于本研究聚焦于总发文量具有一定规模的科研机构，作者基数较大，故个别重名案例对于共现的影响甚微；即使是规模较小的机构，其人员重名的概率也相应降低，因此重名的影响也可忽略不计，后续实验环节会对重名风险进行进一步控制。

（3）将同一年份、同一机构名称的作者进行合并，并统计该年、该机构的作者总人数——“规模”属性（size），形成单位年份的机构-作者向量I_t=(机构名称, 年份, 规模, 作者1, 作者2, …, 作者n)，其中t表示第t年。

3.2.2 机构-年度向量

机构-年度向量$T=(t_{1999}, t_{2000}, \cdots, t_{2014}, t_{2015}, \text{count}, \text{flag})$（图2）构建算法如下。

（1）机构-年度向量中包含17个年份属性（t_{1999}~t_{2015}），将所有年份属性初始化为0。

（2）遍历机构-作者向量，若某机构第i年作者向量的“规模”不为零，则将该机构的机构-年度向量中对应t_i值置为1。

（3）统计各机构的机构-年度向量中t_i为“1”的频次，记为count。

（4）根据年度向量中t_i的分布特征，生成“模式”（flag）属性，值域为{0,1,2,3}，不同属性值的含义如下：

模式“0”，即机构-年度向量中的t_i均为“1”，说明该机构未曾经历机构变迁或者发生合并、拆分之后仍沿用该名称。

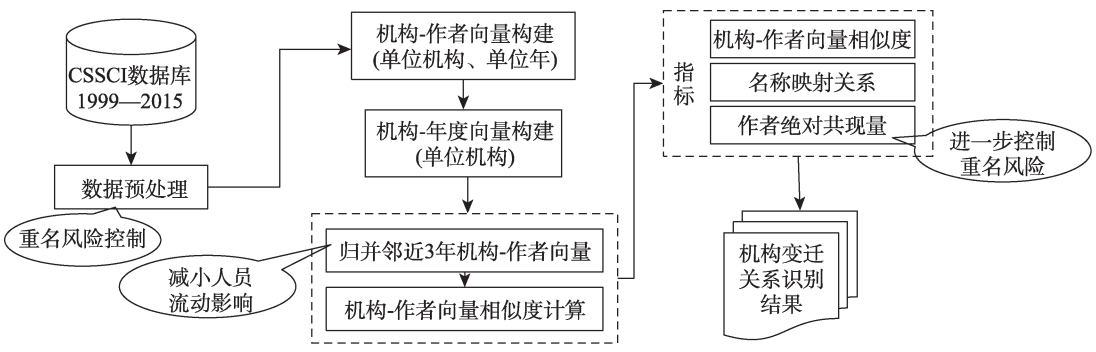

图1 科研机构变迁关系识别总体研究思路

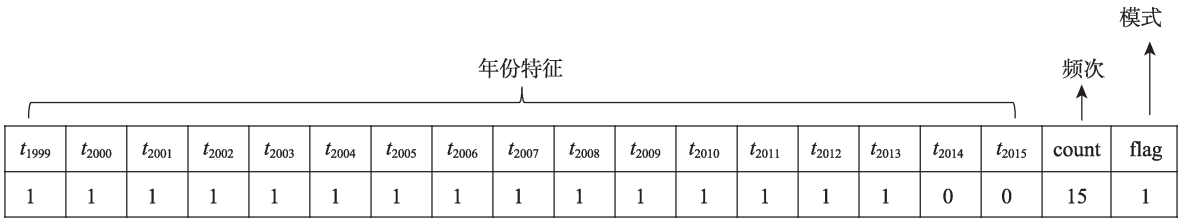

图2　机构-年度向量特征示意图

模式“1”，其年份属性值由连续的“1”变换为连续的“0”，如“徐州师范大学”的年度向量(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,0,15,“1”)，其模式属性的含义是该机构在2013年经历了机构变迁，原机构名称“消失”，2014年起不再有以该名称发表的文献。

模式“2”，与模式“1”相反，即年份属性值由连续的“0”转变为连续的“1”，形如“江苏师范大学”的年度向量(0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,4,“2”)，表示该名称在2012年被首次使用，属于机构变迁后的“新生儿”。

模式“3”，除上述3种特定情形之外的所有形式均属于模式“3”，其属性值“0”、“1”转变次数超过1次以上，如(1,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,1,1,10,“3”)。值得注意的是，机构时间分布特征呈现模式“3”的原因可能有两种：其一是该机构在17年间发生了复杂的机构重组；其二是由于该机构在全年未发文或者系统著录错误等原因造成其年度向量的“混乱”，致使其变迁时间规律变得“模糊”，据此，在识别不同形式的机构变迁关系时，均需要考虑模式“3”机构中可能存在的名称演化。

3.2.3　相似度

1）机构-作者向量归并

识别算法中涉及机构-作者向量的相似度计算，因此需要考虑使用“发文作者表征机构人员组成”的真实性与完整性。根据刘进等[20]的调查，中国研究型大学10年间教师的总体流动频率为15.9%，总体而言，流动比例不高，在实验环节通过压缩年份可以进一步削弱其影响。在高校等科研机构中，除了在职的教师与科研人员之外，硕士研究生、博士研究生已经成为科学论文产出的主力军，随着近些年招生规模的不断扩大，研究生群体已经成为影响机构人员流动的主要因素。国内现行的研究生培养方案中，学制多为3年（部分院校的学术型硕士为2.5年，少量专业型硕士的学制为2年，不过多数院校对于专业型硕士毕业没有发表论文的要求）。综合在职科研人员短时间内流动性较弱以及研究生的学制情况，本研究选择3年作为归并人员的时间窗，从概率角度来说能够最大可能地以发文作者表征机构的人员构成。

据此，本文以机构变迁的年份变量Year作为时间节点，分别归并包括Year在内的前3年数据与不包括Year在内的后3年数据进行相似度计算（图3）；对于无固定变迁时间节点（Year变量）的情形，则采用3年时间窗进行滚动合并的做法，对相邻的3年作者向量依次进行归并，具体做法如图4所示。

2）相似度计算

在对归并后的机构-作者向量进行相似度计算时，主要基于作者名的字面相似度进行匹配，即统计两个机构作者向量中共现作者的比例，

$$S = \frac{n_{co}}{\mathrm{avg}(n_1 + n_2)} \tag{1}$$

式中，S表示相似度；n_1、n_2分别代表两个机构作者向量中的作者人数；n_{co}表示两个机构作者向量中的共现作者数。

3）阈值设置

本研究将机构对之间的机构-作者向量相似度超过阈值作为机构变迁关系识别的必要条件。据此，相似度阈值的设定成为本研究的关键问题之一。现有研究中多通过机构名称的字面编辑距离开展研究，对本文的作者向量相似度计算的参考意义不大；贾君枝等[7]的研究从作者集合的视角探索了更名关系，以15%作为相似度阈值成功识别出一对更名机构，不过文中未阐释该阈值设置的依据。本文借鉴主成分分析中确定因子数的做法，即通过寻找相似度碎石图中的“拐点”作为阈值；该做法具有坚实的数学基础，可以为不同映射关系设定更为合理的阈值，避免了人为设定等不具有普适性的做法。

4）作者绝对共现量

虽然本文拟聚焦于具有一定规模的科研机构（通过总发文量控制规模），但仍无法避免在单位年份中由于即年发文量低而导致的重名风险，即若某

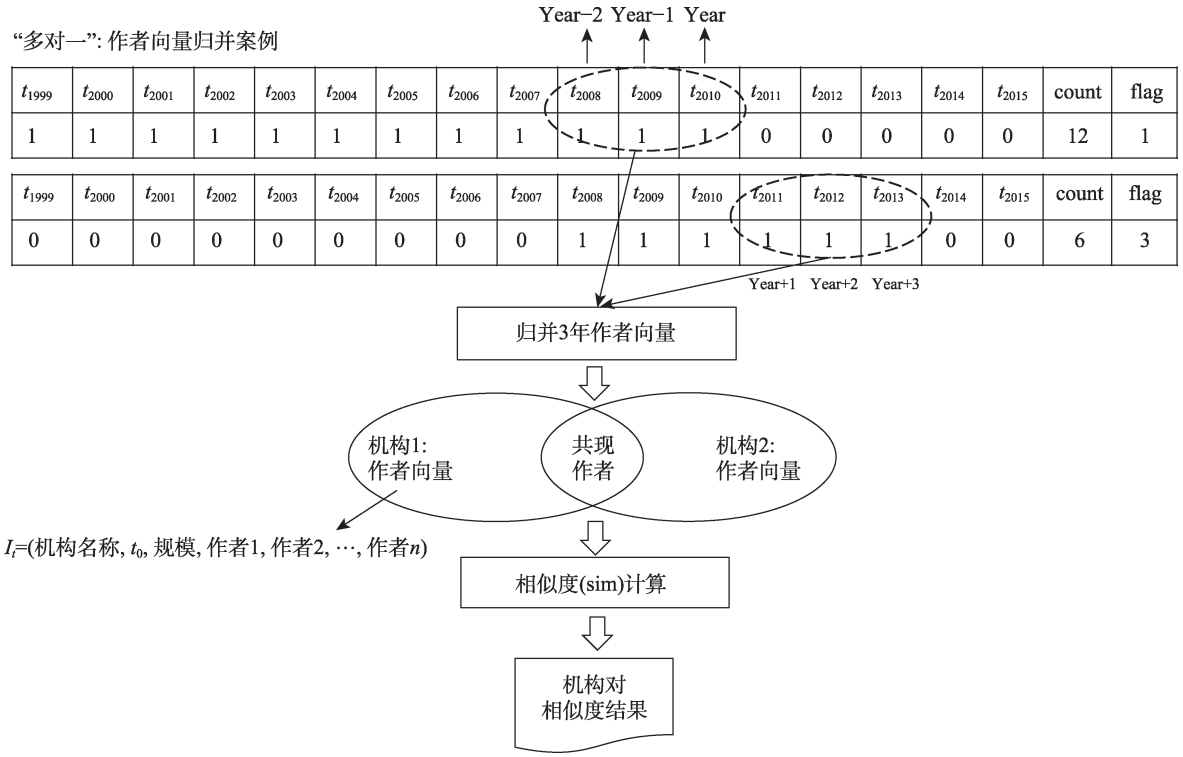

图3 相似度计算示意图

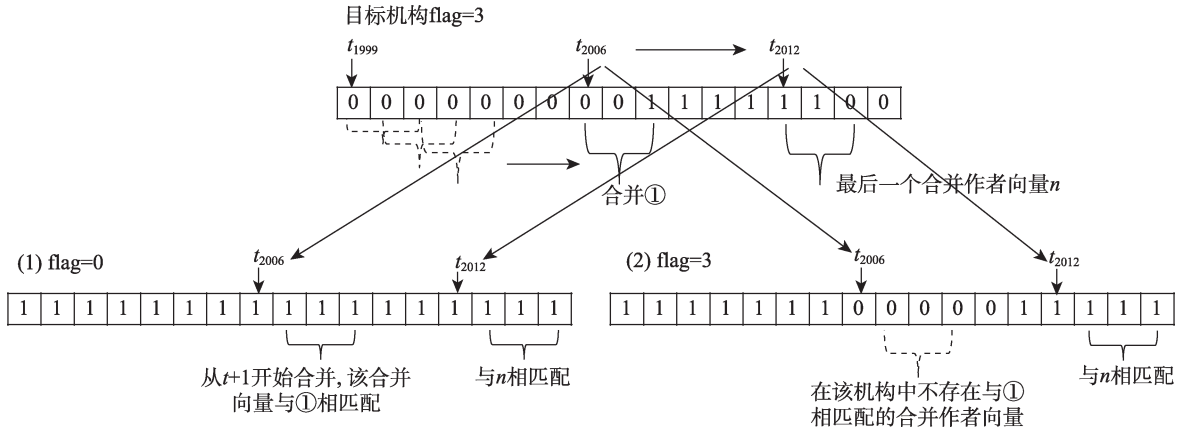

图4 “多对多”作者向量归并示意图

机构在3年内的发文量极低（如小于10），则可能存在因极个别作者的重名而导致相似度超过阈值，因此本文采用机构对之间的作者绝对共现量（C）指标进一步控制重名风险：若作者绝对共现量小于2，则即使机构对满足相似度大于阈值以及映射关系两个条件，仍将其判定为非变迁关系。

3.2.4 识别算法

各机构的机构-年度向量中的“flag”属性是筛选潜在“匹配机构对”（即可能具有变迁关联的机构对）的关键信息，本研究针对4种机构实体变迁关系采用了不同的潜在机构对匹配策略（图5），具体算法如下。

算法1 更名识别算法

输入：机构-作者向量ins_author与机构-年度向量ins_year；

输出：潜在匹配机构对及其相似度；

遍历ins_year{

if (ins_year.flag==1)//潜在更名前的机构，且更名时间点可寻

{

记录机构变迁时间点Year（即由1转变为

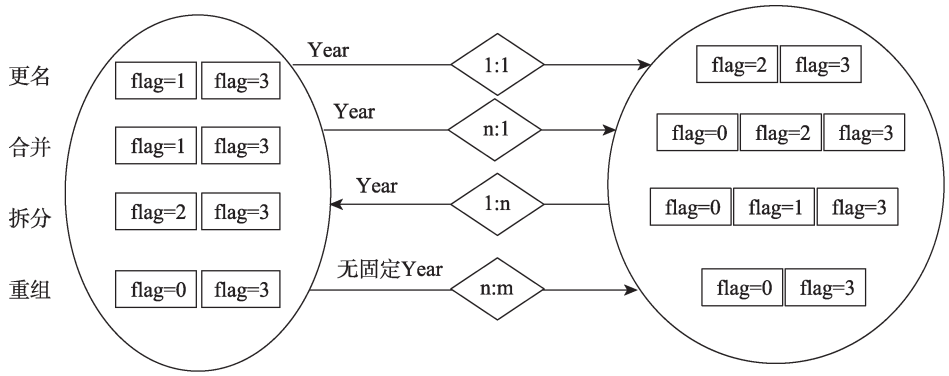

图 5　潜在“匹配机构对”筛选示意图

0 时，“1”的位置）；

归并 ins_author 中该机构 Year-2，Year-1，Year 数据，生成新的机构-作者向量 source；

重新遍历 ins_year{

if (ins_year.flag==2 or 3)//潜在更名后的机构

{

归并 Year+1，Year+2，Year+3 共 3 年数据，生成新的机构-作者向量 destination；

依据公式(1)计算 source 与 destination 的相似度；

}

}

}

}

执行算法 1 之后，根据相似度碎石图中的拐点确定阈值，筛选具有更名关系的潜在机构对；依据“1：1 映射”规则与作者绝对共现量两个条件，进一步过滤，确定机构的更名关系。

算法 2　合并识别算法

输入：机构-作者向量 ins_author 与机构-年度向量 ins_year；

输出：潜在匹配机构对及其相似度；

遍历 ins_year{

if (ins_year.flag==1)//潜在合并前的机构，且合并时间点可寻

{

记录机构变迁时间点 Year（即由 1 转变为 0 时，“1”的位置）；

归并 ins_author 中该机构 Year-2，Year-1，Year 数据，生成新的机构-作者向量 source；

重新遍历 ins_year{

if (ins_year.flag==0 or 2 or 3)//潜在合并后的机构

{

归并 Year+1，Year+2，Year+3 共 3 年数据，生成新的机构-作者向量 destination；

依据公式(1)计算 source 与 destination 的相似度；

}

}

}

}

执行算法 2 之后，根据相似度碎石图中的拐点设定阈值，筛选具有合并关系的潜在机构对；依据“n:1 映射”规则与作者绝对共现量两个条件，进一步过滤，确定机构的合并关系。

算法 3　拆分识别算法

与更名、合并不同，拆分算法是从拆分后的机构“新生儿”视角切入，因新出现的机构在时间分布上具有鲜明特征。

输入：机构-作者向量 ins_author 与机构-年度向量 ins_year；

输出：潜在匹配机构对及其相似度；

遍历 ins_year{

if (ins_year.flag==2)//潜在拆分后的机构，且拆分时间点可寻

{

记录机构变迁时间点 Year（即由 0 转变为 1 时，“1”的位置）；

归并 ins_author 中该机构 Year，Year+1，Year+2 数据，生成新的机构-作者向量 source；

重新遍历 ins_year{

if (ins_year.flag==0 or 1 or 3)//潜在拆分前的机构

{

归并 Year-1，Year-2，Year-3 共 3 年数据，生成新的机构-作者向量 destination；

依据公式(1)计算 source 与 destination 的相似度；

}

}

```
        }
    }
```

执行算法 3 之后，根据相似度碎石图中的拐点设定阈值，筛选具有拆分关系的潜在机构对；依据“1:n 映射”规则与作者绝对共现量两个条件，进一步过滤，确定机构的拆分关系。

算法 4 重组识别算法

在机构重组中，对应机构发文的时间分布规律无明显特征，因此在归并机构-作者向量时采用 3 年滚动时间窗的方式。

输入：机构-作者向量 ins_author 与机构-年度向量 ins_year；

输出：潜在匹配机构对及其相似度；

```
遍历 ins_year{
    if (ins_year.flag==0 or 3)//潜在重组机构，且机构变迁时间点不可寻
    {
        采用 3 年时间窗，滚动归并 ins_author 中相邻 3 年数据，生成新的机构-作者向量 source（若相邻 3 年均无发文作者，则继续向后滚动 1 年）；
        重新遍历 ins_year {
            if (ins_year. flag==0 or 3)//潜在匹配机构
            {
                采用 3 年时间窗，滚动归并相邻 3 年数据，生成新的机构-作者向量 destination（若相邻 3 年均无发文作者，则继续向后滚动 1 年）；
                依据公式(1)计算 source 与 destination 的相似度；
            }
        }
    }
}
```

执行算法 4 之后，根据相似度碎石图中的拐点设定阈值，筛选具有重组关系的潜在机构对；依据“n:m 映射”规则与作者绝对共现量两个条件，进一步过滤，确定机构的重组关系。

需要指出的是，如第 3.2.2 节所述，由于存在部分机构全年不发文或者数据库系统著录错误等特殊情况，会造成机构-年度向量的时间分布规律混乱（即 flag=3），因此除了复杂的机构重组之外，在更名、合并与拆分中，也需考虑 ins_year.flag=3 的机构中存在的变迁关系。由于机构的年度向量无明显变迁时间点可寻，因此采用 3 年滚动时间窗进行机构-作者向量归并，其他做法同算法 1~算法 3。

3.2.5 评价指标

本研究主要采用准确率（公式(2)）与查全率（公式(3)）两个指标来评价本文提出的机构变迁关系识别方法的有效性，

$$P = \frac{n}{N} \tag{2}$$

式中，P 表示识别准确率；n 表示具有机构变迁关系的机构对数量；N 表示识别出的所有机构对数量。

$$R = \frac{m}{M} \tag{3}$$

式中，R 表示识别结果的查全率；m 表示正确识别出的具有相应机构变迁关系的机构对数量；M 表示实验数据集中分别具有更名、合并、拆分或重组关系的机构对总数。

4 结 果

4.1 实验数据

公开发表的科学论文是科研机构的重要研究成果，且科技文献中包含本次研究所需的机构、作者等信息，因此学术论文是研究的理想数据；同时，机构变迁存在一定的时间跨度，据此，本文采集了中文社会科学引文索引（CSSCI）数据库收录的发表于 1999—2015 年文献题录计 1162700 条，第一作者机构数量达 50194。

本文的机构-作者向量的基本思想是“使用发文作者表征机构人员构成”，对于作者发文较少且有一定规模的机构，采用本方法存在发文作者难以代表机构人员的问题。因此，本文将各机构在 17 年间的总发文量降序排列，过滤发文量低于 100 的机构以避免前述问题，最终纳入 453 个独立科研机构。这些机构于 1999—2015 年形成了 6984 个机构-作者向量，其中存在 13 组更名关系以及 8 组合并关系的机构对，不存在拆分与重组关系的机构。本节将依次报告更名与合并关系的识别结果，并采用 CNKI 数据对本文提出的方法在识别拆分与重组关系上的适用性进行检验。

4.2 更 名

4.2.1 相似度阈值确定

1）flag=1 与 flag=2

flag=1 与 flag=2 的机构匹配结果（共 1833 组机构对）以相似度大于 0（310 组案例）的结果作为

碎石图数据源（图 6）。更名关系最可能出现于模式 1 与模式 2 的匹配结果中（即更名后，不再以单位的曾用名发文），由图 6 可见，拐点出现在 0.21333 附近，故本文选取 20% 作为阈值。

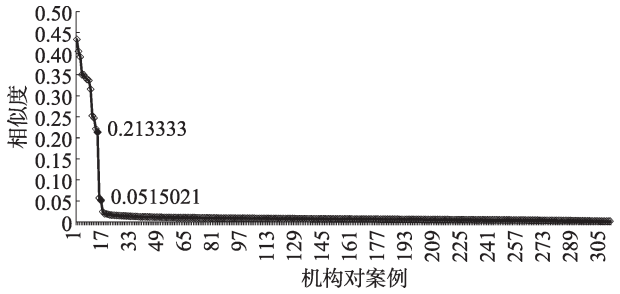

图 6　flag=1 与 flag=2 机构匹配相似度分布

2）flag=1 与 flag=3

flag=1 与 flag=3 的机构匹配结果（共 1006 组机构对）同样以相似度大于 0（48 组案例）的结果为碎石图数据源，由图 7 可见，相似度由 38.5% 骤降至 4.4%，且大于 4.4% 的案例仅 1 例。更名关系中，模式 1 转变为模式 3 多见于机构署名不规范或者全年无发文等情形，考虑到更名前后机构人员的构成理论上应具有较大的相似度，因此选择 30% 作为阈值。

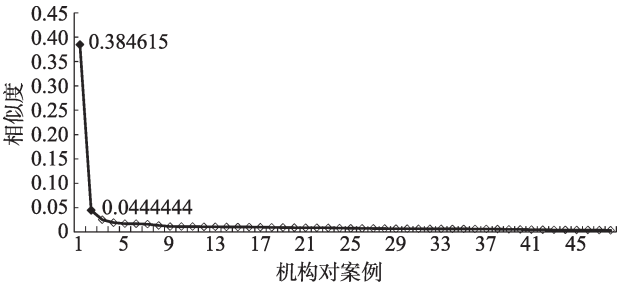

图 7　flag=1 与 flag=3 机构匹配相似度分布

3）flag=3 与 flag=2

flag=3 与 flag=2 的机构匹配结果（共 5103 组机构对）同样以相似度大于 0（96 组案例）的结果为碎石图数据源，如图 8 所示，此处以 20% 作为相似

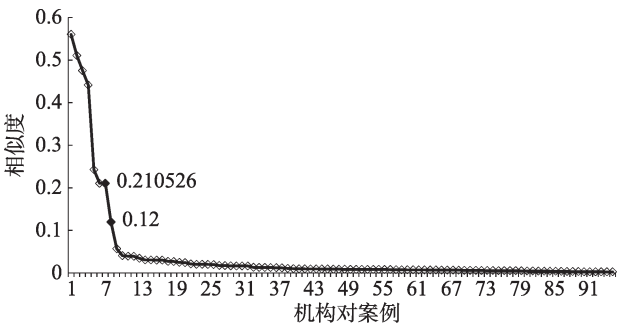

图 8　flag=3 与 flag=2 机构匹配相似度分布

度阈值。

4）flag=3 与 flag=3

flag=3 与 flag=3 机构对意味着两类情形：一是同一机构（名称）自身不同年份的作者向量之间的比较，二是两个不同机构间不同时间段作者向量的比较。考虑到前者与机构变迁无关，故本研究在确定阈值之前，先将其剔除。相似度计算采用类似于 3 元语法的 3 年时间窗进行作者向量归并，因此两个机构间存在多组相似度计算结果。鉴于第 3.2.4 节展示的作者向量归并方式已经有效避免了长时间跨度对人员流动的影响，据此，本研究采取均值作为两机构间的最终相似度（后续在无固定变迁时间点的机构组匹配中均采取该做法）。以相似度均值大于 0（181 组案例）的结果为碎石图数据源，如图 9 所示，选择 18% 作为阈值。

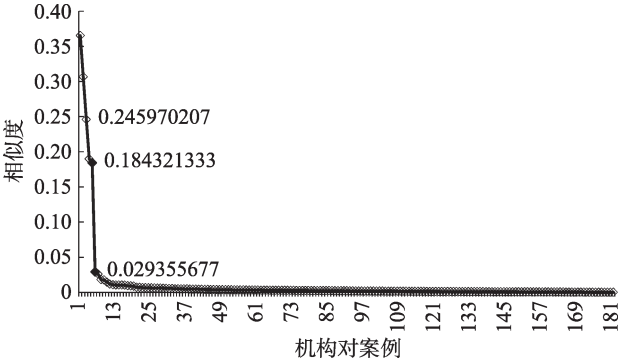

图 9　flag=3 与 flag=3 机构匹配相似度分布

4.2.2　识别结果

筛选出上述 4 类不同模式匹配结果中相似度大于阈值，且满足 1∶1 映射的机构对作为候选集，其次在候选集剔除作者绝对共现量小于 2 的机构对，该结果即为机构之间存在更名关系的识别集，共 16 组（表 1）。结合相似度与一一映射规则正确识别出 13 组具有更名关系的机构（机构变迁形式为“1”），具有更名关系的机构对之间相似度均在 20% 以上，最大值高达 56%，且作者绝对共现量分布于 12~102，表明相似度结果可靠，这些数据充分说明本方法的有效性。

在表 1 所展示的 16 组关系中，包含 1 组合并关系与 2 组错误识别机构对。通过人工审核发现“西南师范大学”于 2005 年与“西南农业大学”合并为“西南大学”，然而在候选集中未出现“西南农业大学”与“西南大学”机构对，因此该结果通过了“一一映射”的检验，故将该更合并关系误判为更

表 1 更名关系识别结果

机构 1			机构 2			相似度	作者绝对共现量	识别结果	真实机构变迁形式
名称	模式	规模	名称	模式	规模				
上海国际问题研究所	3	24	上海国际问题研究院	2	26	56.00%	14	1	1
陕西省考古研究所	3	29	陕西省考古研究院	2	18	51.06%	12	1	1
中国科学技术促进发展研究中心	3	35	中国科学技术发展战略研究院	2	45	47.50%	19	1	1
浙江工商职业技术学院	3	144	浙江工商大学	2	282	44.13%	94	1	0
南京经济学院	1	118	南京财经大学	2	228	43.35%	75	1	1
西北政法学院	1	104	西北政法大学	2	133	40.51%	48	1	1
安徽财贸学院	1	162	安徽财经大学	2	144	39.22%	60	1	1
杭州商学院	1	132	浙江工商职业技术学院	3	154	38.46%	55	1	0
天津财经学院	1	90	天津财经大学	2	116	34.95%	36	1	1
徐州师范大学	1	217	江苏师范大学	2	281	34.94%	87	1	1
杭州商学院	1	132	浙江工商大学	2	223	34.37%	61	1	1
北京广播学院	1	254	中国传媒大学	2	349	33.83%	102	1	1
中央教育科学研究所	1	34	中国教育科学研究院	2	47	24.69%	10	1	1
华东政法学院	3	102	华东政法大学	3	128	24.60%	28	1	1
西南师范大学	3	318	西南大学	2	506	24.27%	100	1	2
云南财贸学院	1	94	云南财经大学	2	87	22.10%	20	1	1
说明			识别结果:0(非更名);1(更名) 机构变迁形式:0(不相关匹配);1(更名);2(合并);3(拆分);4(重组)						
准确率			修正后 100%						
查全率			100%						

名关系，该误差是由于“西南农业大学”是以农学为主的高校，发文主要集中在自然科学，在 CSSCI 中发文较少，所以该错误产生的原因与数据集有关而非本文提出的算法。

对于“浙江工商职业技术学院”与“浙江工商大学”，以及“杭州商学院”和“浙江工商职业技术学院”2 组机构对，其相似度分别达到 44.13% 与 38.46%，但是通过资料审查发现二者并不存在任何机构变迁关系。该现象超乎常识认知，因此对本文的数据处理过程进行校验，以寻找原因。首先检查“浙江工商职业技术学院”的历年作者向量，发现其年均发文量为 10 篇左右，而在 2004 年突然增至 140 多篇论文，通过人工审核 CSSCI 数据库中作者机构为“浙江工商职业技术学院”、“浙江工商大学”以及“杭州商学院”的论文发现，2004 年“浙江工商职业技术学院”的发文多为机构著录错误，实际属于“浙江工商大学”的研究成果。由于产生该识别结果误差的原因是 CSSCI 偶发的著录错误，而非本研究的算法所致，故在评价识别准确率时，可将这两组结果排除在外。

通过修正上述数据，本方法更名识别的准确率与查全率均达到 100%。

4.3 合 并

4.3.1 相似度阈值设定

1）flag=1 与 flag=2，flag=1 与 flag=3，flag=3 与 flag=2

flag=1 与 flag=2、flag=1 与 flag=3、flag=3 与 flag=2 机构匹配相似度结果分别如图 6、图 7 和图 8 所示。相较于更名，合并前后规模的变化可能较大，会引起共现作者占比降低，根据“拐点”的位置，均选择较为“宽松”的阈值，即分别以 5%、4% 和 10% 作为阈值，通过后续“n:1 映射”以及作者绝对共现量过滤非合并关系的机构对。

2）flag=1 与 flag=0

flag=1 与 flag=0 两类机构间的合并属于机构“兼并”现象，即经过合并，先前某个机构仍沿用原名称，共 5620 组机构对，仅展示相似度大于 1% 的 197 组案例，其相似度分布碎石图如图 10 所示，同 1），此处以 4% 作为相似度阈值。

4.3.2 识别结果

筛选出上述 4 组结果中相似度大于阈值，且满足“n:1 映射”的机构对作为候选集，并在候选集

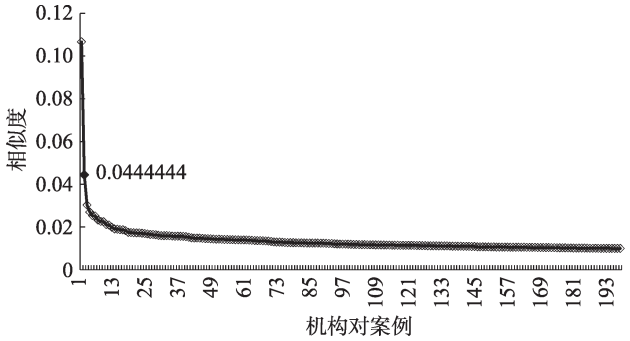

图 10　flag=1 与 flag=0 机构匹配相似度分布

中筛选出作者绝对共现量大于 1 的组别，该结果即为机构之间存在合并关系的识别集（表 2）。本文共正确识别出 7 组具有合并关系的机构对，准确率达 100%。对于“中共中央党校科社部”、“中共中央党校科社教研部”以及“中共中央党校哲学部”与“中共中央党校马克思主义理论教研部”3 组机构对之间的关系，由于“中共中央党校”官网上并未介绍其所属部门的沿革，考虑到前三者均与“中共中央党校马克思主义理论教研部”形成多对一映射，说明后者的概念范围更广；同时，在这 3 组机构对中，其机构的即年发文规模处于 20 上下，分别有 4 位或者 3 位共现作者，同一机构中能有如此规模的重名作者的可能性可以忽略，因此将其判断为合并关系。

合并关系识别结果中查全率仅为 87.5%，同样是由于本研究数据集中未出现“西南农业大学”，从而导致即使“西南师范大学”与“西南大学”的相似度超过阈值，但由于其映射关系不满足“n:1 映射”，不能纳入合并关系结果集。

4.4　CNKI 检验

考虑到：①CSSCI 仅面向人文社会科学领域，未覆盖自然科学；②CSSCI 数据库中收录的文献均为发表于国内核心期刊的论文，具有较高的质量，仅占我国人文社科学术期刊总量的 20%；③CSSCI 收录的数据始于 1998 年，该数据难以覆盖发生于 20 世纪 90 年代初的我国科研机构合并潮；这些原因导致使用 CSSCI 的数据不能充分、完整反映科研机构的名称演化关系，同时，第 4.1 节与第 4.2 节的结果中尚未发现机构的拆分关系。据此，本研究尝试分别使用 3 组具有不同机构变迁关系的机构对（更名、合并与拆分各一组）对本文提出的识别方法进行验证。鉴于 CNKI 的数据覆盖范围更为全面，故拟从中下载相关机构的发文进行验证。

通过查阅资料，本文确定了具有 3 种机构变迁关系的代表性机构：①华东工学院于 1993 年更名为南京理工大学；②成都科技大学与四川大学于 1994 年合并为四川联合大学；③武汉水利电力大学于 2000 年被拆分，其主体部分成为武汉大学工学部，另一部分与湖北三峡学院合并为三峡大学。

同样地，为了减少人员流动对实验结果的影响，分别从 CNKI 中下载以下数据：①华东工学院于 1991—1993 年和南京理工大学于 1994—1996 年发表的论文；②成都科技大学与四川大学于 1992—1994 年和四川联合大学 1995—1997 年发表的论文；③武汉水利电力大学于 1998—2000 年、武汉大学于 2001—2003 年以及三峡大学于 2001—2003 年发表的论文。

参照第 3.2 节的方法进行作者合并，随后计算相似度，结果如表 3 所示。从相似度结果可以看

表 2　合并关系识别结果

机构 1			机构 2			相似度	作者绝对共现量	识别结果	真实机构变迁形式
名称	模式	规模	名称	模式	规模				
河南财经学院	1	112	河南财经政法大学	2	199	31.51%	49	1	2
山东经济学院	1	102	山东财经大学	2	263	25.21%	46	1	2
山东财政学院	1	112	山东财经大学	2	263	21.33%	40	1	2
中共中央党校科社部	3	22	中共中央党校马克思主义理论教研部	2	16	21.05%	4	1	2
中共中央党校科社教研部	3	22	中共中央党校马克思主义理论教研部	2	16	21.05%	4	1	2
中共中央党校哲学部	3	34	中共中央党校马克思主义理论教研部	2	16	12.00%	3	1	2
河南省政法管理干部学院	1	33	河南财经政法大学	2	200	5.15%	6	1	2
说明			识别结果：0（非合并）；1（合并） 机构变迁形式：0（不相关匹配）；1（更名）；2（合并）；3（拆分）；4（重组）						
查准率			100%						
查全率			87.5%						

表3 检验结果

机构1	机构2	Year	规模1	规模2	相似度	作者绝对共现量	机构变迁形式
华东工学院	南京理工大学	1993	239	409	26.54%	86	1
成都科技大学	四川联合大学	1994	238	432	16.72%	56	2
四川大学	四川联合大学	1994	728	432	10.17%	118	2
三峡大学	武汉水利电力大学	2001	140	748	6.98%	31	3
武汉大学	武汉水利电力大学	2001	1391	748	7.39%	79	3
说明	机构变迁形式:1(更名);2(合并);3(拆分)						

出，该方法对更名与合并关系的识别效果良好，与CSSCI的结果一致；对于机构拆分关系而言，其相似度均在5%~10%；3种变迁关系中作者绝对共现量均较大，分布于31~118，表明作者的高相似度是源于真实作者共现，因重名引起的可能性极低。

拆分关系相似度低的可能原因是机构拆分往往伴随着其他形式的组织重构。武汉水利电力大学是一个典型案例，其主体部分拆分后归并至武汉大学工学部，而其宜昌校区则与湖北三峡学院合并组建三峡大学。从武汉水利电力大学的角度看，其与武汉大学以及三峡大学存在1:2映射关系，且武汉水利电力大学消失，可以判定该校存在拆分关系；从武汉大学以及三峡大学角度看，武汉大学由原武汉大学与武汉水利电力大学合并而成，三峡大学由原武汉水利电力大学与湖北三峡学院合并而成；因此，三者间存在拆分以及合并等复杂的重构关系（限于篇幅武汉大学与三峡大学的合并关系不再进一步分析数据）。

5 讨 论

5.1 机构变迁时间节点

本研究在进行机构-作者向量合并时，基于模式1（flag=1）和模式2（flag=2）机构-年度向量中年份属性值的“突变”，探索了其机构变迁时间节点，并在结果中进行了展示。通过人工审核发现，由该方法识别的变迁时间点（Year）与真实机构更名、合并年份（真实t）会产生一年左右的误差（表4）。对于实验识别Year比真实机构变迁时间推迟一年的合理解释有两种：其一是名称变更时间接近年末，而由于期刊论文存在“出版时滞”[21]，尤其是审稿过程造成的延迟，导致由公开出版物中提取机构变迁时间也相应地滞后；其二则由作者的使用习惯引起，即机构变更后的短时间内原机构人员仍惯于使用曾用名作为发文的单位名称。

而对于实验识别Year较真实机构变迁时间早一年的现象，主要是由于科研机构在申请更名、分立或合并时具有“审核时滞”，而相关机构在提出申请之日即开始使用新的机构名称引起的。例如，根据教育部公布的《实施本科及以上教育的高等学校的设立、分立、合并、变更和终止审批服务指南》规定，“申请设立公办本科学校（含更名‘大学’）的，审批时限为3个月加上专家考察和评审时间；申请设立民办本科学校（含更名‘大学’）的，审批时限为6个月”。

表4 实验变迁时间点与真实变迁时间点对比

	flag=1	flag=2	实验Year	真实t	时间差
更名关系	南京经济学院	南京财经大学	2003	2003	0
	西北政法学院	西北政法大学	2006	2006	0
	安徽财贸学院	安徽财经大学	2003	2004	−1
	天津财经学院	天津财经大学	2004	2004	0
	徐州师范大学	江苏师范大学	2012	2011	+1
	杭州商学院	浙江工商大学	2003	2004	−1
	北京广播学院	中国传媒大学	2003	2004	−1
	中央教育科学研究所	中国教育科学研究院	2012	2011	+1
	云南财贸学院	云南财经大学	2005	2006	−1
合并关系	河南财经学院	河南财经政法大学	2011	2010	+1
	山东经济学院	山东财经大学	2012	2011	+1
	山东财政学院	山东财经大学	2012	2011	+1
	河南省政法管理干部学院	河南财经政法大学	2010	2010	0

5.2　作者绝对共现量

本研究除了关注科研机构变迁前后内部的人员构成的相似性的相对值之外（即机构-作者向量相似度），同时考虑了作者共现的绝对值（作者绝对共现量），以避免在即年发文量较小的机构中由于极个别作者重名而导致的机构间人员的高相似度。据此，在作者向量相似度与具体变迁形式的映射规则的基础之上，进一步增加该指标作为判断机构变迁关系的必要条件，实验结果与预期一致：作者绝对共现量指标的应用可以进一步提高机构变迁关系识别的准确率，能够避免“错误肯定”的风险。

表 5 展示了一组由于作者绝对共现量过小而被判定为非变迁关系的机构对。由于“中国电视艺术委员会”与“中国电视杂志社”的机构模式均为“3”，故在提取更名关系的算法中进行了多组相似度计算（滚动 3 年时间窗）。在前期检验中，该组机构对满足相似度均值阈值大于 18% 条件，且符合“1:1 映射”关系，故进一步观测其作者绝对共现量指标以最终确认其是否具有更名关系。由表 5 可知，“中国电视艺术委员会”与“中国电视杂志社”属于即年发文量较低的小规模机构，其历年相似度均达到 10% 以上，但作者绝对共现量表明，其个别年份的高相似度的有效性欠佳。

通过人工审查发现，“中国电视艺术委员会”与“中国电视杂志社”确实不存在更名关系。事实上，中国电视艺术委员会是《中国电视》杂志的主办单位，中国电视杂志社与其存在一定的关联，但由于其非机构变迁关系，且其高相似度的可靠性较低，故本研究结果未将其判断为更名关系具有合理性。

6　结　论

与现有围绕机构名称本身相似度开展机构规范化的多数研究相比，本文基于机构-人员向量相似度的视角，结合机构变迁特有的时间特征与名称映射关系解决了由机构变迁引起的机构名称演变问题，为具有变迁关系的机构对建立了“关联”，以服务于“机构信息”在信息检索、文献计量等领域的实践应用。

本文提出的机构名称演化识别方法的核心在于构建机构-作者向量与机构-年度向量，其中机构-年度向量用于探索不同机构变迁方式（更名、合并、拆分与重组）中潜在的“匹配机构对”以及其发生机构变迁的时间节点。同时，为降低因数据集时间跨度大而导致的人员流动现象对计算作者向量相似度时造成的结果偏差，首先依据实验识别的变迁时间点进行邻近 3 年作者向量归并，再对其进行相似度计算，判定相似度超过阈值且满足映射关系的机构对之间存在相应的变迁关系；对于该类工作的难点之一是相似度阈值的设定，本研究采用 PCA（principal component analysis）中具有坚实数学基础并已成熟应用的碎石图“拐点”法确定阈值，并根据不同的名称演化关系选择严格和宽泛的具体值；另外，为了进一步降低由于即年发文量低造成的重

表 5　“中国电视艺术委员会”与“中国电视杂志社”相似度计算结果

flag=3	flag=3	归并时间点	规模 1	规模 2	相似度	作者绝对共现量	相似度均值
中国电视艺术委员会	中国电视杂志社	2013	9	4	61.54%	4	18.99%
		2011	10	8	44.44%	4	
		2012	12	6	44.44%	4	
		2014	6	3	44.44%	2	
		2010	8	9	23.53%	2	
		2008	7	8	13.33%	1	
		2008	8	10	11.11%	1	
		2009	8	11	10.53%	1	
中国电视杂志社	中国电视艺术委员会	2010	11	10	38.10%	4	18.43%
		2012	8	9	35.29%	3	
		2009	10	8	33.33%	3	
		2013	6	6	33.33%	2	
		2014	4	3	28.57%	1	
		2008	7	8	13.33%	1	
		2008	8	8	12.50%	1	

名风险，辅以作者绝对共现量对结果进行二次过滤。通过采集 CSSCI 1999—2015 年的数据进行实证，结果表明，该方法在机构变迁关系的识别上具有优异的表现，修正后的准确率达 100%，查全率则达到 87.5%~100%，从 CNKI 下载的验证数据的测试结果也体现了该方法的合理性与有效性。

在实验过程中，部分机构中上下级混杂的情况影响了机构变迁关系的识别，例如，在第 4.2.2 节结果中报告的有关“中共中央党校”下级机构间的相关关系（简称、别称），此类机构对的发现主要是由于事先未对机构名称进行等级划分引起的，这也是本研究的可改进之处。后续研究可以在严格区分一级、二级机构的基础上再次对本文提出的方法进行检验，从而获得更精确的机构变迁关系。本研究中“西南农业大学”由于 17 年间在 CSSCI 中的发文量较少从而被拒于数据集之外，导致了机构合并关系识别中查全率的不足，故应在后续研究中选取覆盖面更广的数据集（如 CNKI）进行系统实验。本研究所提出的识别方法可应用于 CNKI 以及万方等收录文献较为全面的数据库，以提高检索的查全率和准确率。

参 考 文 献

[1] De Bruin R E, Mode H F. Delimitation of scientific subfields using cognitive words from corporate addresses in scientific publications[J]. Scientometrics, 1993, 26(1): 65-80.

[2] French J C, Powell A L, Schulman E. Using clustering strategies for creating authority files[J]. Journal of the American Society for Information Science and Technology, 2000, 51(8): 774-786.

[3] Galvez C, Moya-Anegón F. The unification of institutional addresses applying parametrized finite-state graphs (P-FSG)[J]. Scientometrics, 2006, 69(2): 323-345.

[4] 王星, 曾建勋, 苏静, 等. 机构规范文档构建方式研究[J]. 数字图书馆论坛, 2015(7): 2-8.

[5] 黄俊贵. 规范控制概说[J]. 高校图书馆工作, 1999, 19(3): 1-8.

[6] Jonnalagadda S, Topham P. NEMO: Extraction and normalization of organization names from PubMed affiliation strings[J]. Journal of Biomedical Discovery and Collaboration, 2010, 5: 50-57.

[7] 贾君枝, 曾建勋, 李捷佳, 等. 科研机构名称归一化实现[J]. 图书情报工作, 2018, 62(13): 103-110.

[8] Christen P, Belacic D. Automated probabilistic address standardization and verification[C]// Proceedings of the 4th Australasian Data Mining Conference, 2005.

[9] Guo H L, Zhu H J, Guo Z L, et al. Address standardization with latent semantic association[C]// Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York:ACM Press, 2009: 1155-1163.

[10] 孙海霞, 王蕾, 吴英杰, 等. 科技文献数据库中机构名称匹配策略研究[J]. 数据分析与知识发现, 2018, 2(8): 88-97.

[11] 刘浏, 王东波. 命名实体识别研究综述[J]. 情报学报, 2018, 37(3): 329-340.

[12] 曾建勋, 贾君枝. 机构名称规范数据的语义模型构建[J]. 大学图书馆学报, 2019, 37(1): 42-47.

[13] Nguyen V H, Nguyen H T, Snasel V. Text normalization for named entity recognition in Vietnamese tweets[J]. Computational Social Networks, 2016, 3: 10.

[14] Cuxac P, Lamirel J C, Bonvallot V. Efficient supervised and semi-supervised approaches for affiliations disambiguation[J]. Scientometrics, 2013, 97(1): 47-58.

[15] 孙海霞, 李军莲, 吴英杰. 基于 K-means 的机构归一化研究[J]. 医学信息学杂志, 2013, 34(7): 41-44, 71.

[16] 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.

[17] Jiang Y, Zheng H T, Wang X M, 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.

[18] Huang S Q, Yang B, Yan S L, et al. Institution name disambiguation for research assessment[J]. Scientometrics, 2014, 99(3): 823-838.

[19] 杨波, 杨军威, 阎素兰. 基于规则的机构名规范化研究[J]. 现代图书情报技术, 2015(6): 57-63.

[20] 刘进, 沈红. 中国研究型大学教师流动: 频率、路径与类型[J]. 复旦教育论坛, 2014, 12(1): 42-48, 92.

[21] 李江, 伍军红. 论文发表时滞与优先数字出版[J]. 编辑学报, 2011, 23(4): 357-359.

（责任编辑 魏瑞斌）