作者重名辨识研究进展*

□ 袁军鹏 俞征鹿 苏成 马峥 杨志清 / 中国科学技术信息研究所 北京 100038
宿洁 / 中央财经大学管理科学与工程学院 北京 100081

摘要：作者重名现象将降低文献检索和网络检索的准确性，影响文献数据搜集质量，增加基于作者个人层面分析评价的障碍。目前国内外学者提出了人工辨识、数据库字段修正、基于机器学习的重名辨识等多种方法来解决作者重名问题。文章总结作者重名辨识面临的问题，分析当前各辨识方法的特点以及不足之处，指明作者重名辨识特别是中国作者重名辨识的发展方向。

关键词：作者重名，机器学习，社会网络，聚类
DOI：10.3772/j.issn.1673—2286.2011.10.004

1 引言

作者重名现象将降低文献检索和网络检索的准确性、影响文献数据搜集质量、增加基于作者个人层面分析评价的障碍。Shiffrin和Borner认为作者重名辨识工作是情报学、知识管理、文献计量学与科学计量学等工作的基础[1]。有效的作者重名辨识可以在科技评价、学术研究以及科研管理中广泛应用，如科研项目管理者在立项、评审、管理等过程中寻找评审专家；研究人员寻找某领域的学者信息；期刊编辑寻找审阅论文和办刊、选题策划的专家；学术会议组织者寻找主题发言学者；科研人员引进及招聘过程中核实应聘对象所有的论文时都需要准确的作者信息。另外，人名辨识结果还能建立引文网络（catation network）、合作网络（collaboration）与作者个人名片档案（author profiles）等增值服务。

传统上，处理作者重名问题都是交给图书馆进行人工的权威控制，但在网络普及、数字图书馆充斥的今日，这套方法已经无法有效解决海量数据增长与人工辨识效率偏低的矛盾。所以很有必要对现有各算法来源数据的性质和特点以及特征计算的方法和特征合并方式进行全面而透彻的分析，总结现有研究的特点与不足，为提高现有作者重名辨识算法的效率和辨识结果的准确性提供支持。

2 作者重名辨识面临的问题

对于单一语言体系作者重名辨识问题来说，主要面临以下问题：

（1）多个作者的名字完全相同。现实世界中，多个人物共享一个人名是很普遍的现象，例如中国共有290607人叫“张伟”；排在第二的是王伟，共有281568人[2]。

（2）同一作者的论文在被检索到时，作者姓名可能会有不同的表现方式。外文的多种拼写方式（姓氏和名字的位置、全称和缩写），个人书写、印刷或者数据库加工时的错误，用笔名发表论文等都会导致一个作者的名字有多种形式。书写、印刷错误或数据加工等机器容易识别的问题比较容易被检测出来[3]，因此研究重点在于多种外文拼写或者缩写方式辨识问题。

（3）作者重名辨识所需的元数据不完全或者缺失。理想状态下，我们有足够信息就可以准确识别每一位作者，例如我们在日常生活中遇到重名现象时，采取增加出生年月、性别、父母姓名、住址等信息即可以进行区分[4]。但是，在大规模的文献索引数据库中，由于记录信息的限制，我们无法为每篇论文的作

*基金项目：国家自然科学基金（编号：70973118）；中国科学技术信息研究所预研项目（YY—2010027）。

者找到对应真实作者的元数据。

（4）海量信息导致重名辨识的困难。处理的论文是海量信息，同时，姓名是一个开放的、动态的数据，不但数量十分庞大，难以完全列举，而且随着时间的推移，不断有新的命名实体产生。这些会导致已经辨识的作者信息无法完全用于新产生作者的重名辨识中。

在中国，我们进行文献检索和文献计量研究时，不但要检索、分析一个作者的中文论文，还要检索、研究该作者的外文论文，特别是SCI、EI、CPCI（原ISTP）三大检索论文。在进行中国作者中英文重名辨识时，不但面临以上4种问题，还将遇到更加困难的问题。

中国作者的英文姓名音译后重名现象更加严重，如据我们统计，在504个百家姓中，翻译到英文时，仅为229个，如俞、庾、于、余、虞、郁、余、禹等8个姓氏都会译为“Yu”，而且我们由英文名字反推中文名字时也无法选择确切的汉字。不仅如此，由于英文姓氏和名字的形式一般与中国相反（英文是名前姓后，中文是姓前名后），在姓氏和名字都是姓名用字时，会导致英文的姓名出现更多重名现象，如两位中文作者分别是苏成和程素，苏成和程素的英文姓名都可以是Su Cheng，也都可以是Cheng Su。另外，由于多数英文文献数据库采用“姓+名的首字母”的检索策略，导致不同作者的姓名相近或相同时，出现重名现象。而且，“三大检索”数据库中，EI仅提供第一作者机构信息，SCI、ISTP仅提供通讯作者机构信息，其余作者的机构信息要么不提供，要么不对应[①]，这也为我们辨识重名增加了障碍。

3 作者重名辨识算法研究现状分析

作者重名辨识已成为当前国内外学者的一个研究热点，2011年3月14日，我们分别以“重名”和“name disambiguation（人名消岐）”在中、英文数据库中检索，共得到76篇期刊论文，其中43篇论文发表在*Journal of the American Society for Information Science and Technology*、*Research and Advanced Technology for Digital Libraries*、*Information Retrieval*、*Scientometrics*等图书情报期刊上，24篇发表在机器学习、数据挖掘等期刊上，其余的9篇发表在专业领域期刊上。人名重名辨识问题的研究主要集中在以下几个领域：网络搜索、文献检索、数字图书馆及文献数据库（电子文档）、文献计量及评价分析、本体论、自然语言处理及信息抽取等。这些领域的研究主要采用以下几种方法进行：

3.1 人工辨识

这种方法主要来源于传统图书馆对馆藏的权威控制（authority control）[5]的思想，如DeRose等撷取特定研究社群的数据库，将数据进行格式统一与人工比对后建立自动辨识平台（DBLife）[6]。

3.2 基于作者互动的数据库字段修正辨识方法

Guha R. V.和Garg A.[7]提出了一种基于用户反馈的半自动化方式进行重名辨识，如请作者提供住址、联系方式（E-mail、电话等）、所在机构等。Xia J. F.[8]提出采用“作者姓名+机构名称+出版日期”综合标示符和由作者添加其姓名的多种变体来进行重名辨识，并提出在数据库中添加一个额外的字段来保存此类信息。更有甚者，Dervos等人建立UAI_Sys系统，采用网页式的登记服务并给予作者辨识码来进行重名辨识[9]，该方法已经在一些文献数据库中得到应用，比如Web of Science的Researcher ID就是通过作者个人注册一个ID来作为在数据库中的唯一标识。

3.3 基于机器学习的重名辨识

这种方法强调利用个体的所有特征，包括作者作品内部信息与外部信息所计算、建立的知识，进一步利用机器学习的方法依据与作者相对应的属性进行自动分类和聚类。根据机器学习的方式可以将其分成两大类：监督式与无监督式。

3.3.1 监督式机器学习

监督式机器学习需要依据训练数据集，由于可以事先调整训练集，所以分类结果一般较好，但是要注意训练集的代表性和完整性，并且要求多次学习，不

① 目前，SCI、ISTP数据库开始对应给出作者机构信息，作者全名，但是以前的数据并未回溯，所以仍然存在以上问题。

断更新辨识模型。在这方面，有许多学者做了开创性的工作，如Han和Giles等提出利用朴素贝叶斯概率模型和支持向量机两种分类算法利用合作者姓名、论文题目和期刊或者会议名称等特征解决引文中的作者重名辨识问题[10]。Torvik和Weeber等提出利用题目、期刊名称、合作者姓名、学科分类、语言、作者机构和名字的属性等特征自动建立训练集进行重名辨识[11]。Yin和Han等结合两种互补性相似关系，并使用支持向量机来标识这些关系之间的权重，自动产生训练集进行重名辨识[12]。

3.3.2 无监督式机器学习

按照分类特征不同，我们可以把无监督式的机器学习分为基于论文和基于论文以外信息两大类：

（1）基于论文

主要是采用论文内所提供的信息，包括期刊名称、论文题目、合作者、作者单位、全文、E-mail、引文等特征进行无监督式机器学习。如宾夕法尼亚州立大学的C. Lee Giles创立了CiteSeer全文数据库，Giles的研究团队分别提出了利用合作者、论文题目和期刊名称等特征解决引文中的作者重名辨识问题的K-way光谱自动聚类法[13]和DBSCAN算法[14]。北京大学的王厚峰和梅铮提出将作者姓名、论文题目、职业利用布尔向量进行表示，特征向量用真实值表示，然后集成基于布尔向量的启发式算法和基于特征向量的凝聚聚类算法进行人名的重名辨识[15]。Culotta和McCallum提出了M-H抽样器（Metropolis-Hastings sampler）提取论文题目、作者E-mail、作者单位、期刊名称等特征进行作者重名辨识[16]。McRae-Spencer和Shadbolt提出利用自引、合作者、论文来源等特征在引文网络中识别作者重名[17]。Song、Huang和Giles提出基于两种分层贝叶斯文本模型——概率潜在语义分析（PLSA）和集合概率算法LDA（Latent Dirichlet allocation）的主题模型，利用全文的第一页内容产生作者的主题，以此进行凝聚的层次聚类来进行作者重名辨识[18]。Bhattacharya和Getoor提出了一种集体实体解析方法（collective entity resolution），一组论文的辨识结果会帮助另一组辨识，如A名与B名同时出现在两篇论文，若确定两个A不是同一个人，则会类推两个B是同个人的机率也不高[19]。Soler提出利用作者姓名、E-mail、地址信息、论文题名、关键词、研究领域、期刊名称和发表时间等进行论文之间距离的度量，然后基于此进行聚类分析，识别作者重名[20]。北京邮电大学吴斌等提出先属性匹配，然后基于文献合作网络的结构解析的策略的适用于中文文献索引数据的实体解析方法[21]。北京邮电大学索利军和吴斌提出可以通过作者的科研合作网络辨识作者重名现象[22]。清华大学王建勇团队提出基于图的人名识别框架——GHOST[23]，该框架已经在中国人民大学孟小峰教授创立的以作者为中心的面向计算机领域的中文文献集成系统C-DBLP 系统中用于重名辨识[24]。Kang和Na等提出合作者信息是最容易获得的在重名辨识方面有影响力的指标，因此，提出了一种基于网络辅助合作者信息的重名辨识算法[25]。华中科技大学金海教授等提出一种新的基于语义关联的重名辨识方法——SAND，主要思想是利用名称实体的语义关联进行聚类分析[26]。Zhu、Zhou和Fung提出基于术语驱动的聚类方式进行作者重名辨识，首先建立模仿专家的术语分类法，将之转换为图，然后用基于图的相似度模型和基于图的随机游动模型进行聚类[27]。Ferreira A. A.和Goncalves M. A.等指出作者重名辨识是当前数字图书馆面临的最困难问题之一，利用合著者姓名、论文题目、期刊名称等信息构建了作者重名辨识综合发生器（SyGAR）来进行作者重名辨识[28]。Tang L.和Walsh J. P等利用心理学的认知地图和网络分析的近似等价结构技术，发展了基于同质性知识打分的作者重名辨识算法[29]。Cota R. G.、Ferreira A. A.和Nascimento C.等利用合著者姓名、论文题目、期刊名称等信息提出了一个启发式的层次聚类方法来进行作者重名辨识[30]。

（2）基于论文以外信息

Mann和Yarowsky提出利用作者传记中的特征的自动聚类技术进行重名辨识[31]。Aswani和Bontcheva等提出综合利用摘要、姓名、论文题目、合作者等信息和网络数据挖掘来为重名辨识提供更多的证据[32]。清华大学唐杰等提出采用分类器从网络中抽取相关文档，利用条件随机场（CRF）算法处理作者的相关文档，构建其社会网络，然后利用基于制约的概率模型进行重名辨识[33]。中国台湾中央研究院的Yang和Peng等提出基于研究主题特征和网页共现特征的聚类算法辨识

文献数据库中的作者重名问题[34]。清华大学王建勇团队提出网络搜索中人名重名辨识算法——GRAPE，利用姓名、机构、E-mail等特征进行基于图的无监督学习的自动聚类并对结果进行评价[35]。由于重名的不同人物所属的社会网络具有区分性，很多学者都提出了构建社会网络来进行作者重名辨识，如Byung-Won提出将作者重名问题进行可视化表达，并用社会网络分析方法（SNA，Social Network Analysis）进行处理[36]。哈尔滨工业大学的郎君等利用检索结果中共现的人名发现并拓展检索人物相关的潜在社会网络，结合图的谱分割算法和模块度指标进行社会网络的自动聚类，在此基础上实现人名检索结果的重名辨识[37]。Levin F. H.和Heuser C. A.研究了将社会网络分析方法集成到传统的作者重名辨识过程，并用真实数据验证了利用社会网络分析可以提高辨识的准确性[38]。

3.4 其他方法

另外，还有些学者提出的研究方法可以提供借鉴，如Culotta等提出利用总体约束（aggregate constraints）方式帮助判别，如任一作者在某特定的一年中不可能发表超过30篇论文、只会有2个以下E-mail与服务机构等[39]。Churches等提出基于隐蔽式马尔科夫链的记录连结模型，选取电话号码、生日、性别、通信地址等唯一性特征处理姓名的变异与地址的标准化，该研究丰富了相似性测量的方法[40]。

4 当前主要研究模式的特点与瓶颈分析

从以上的分析可以看出，当前的作者重名辨识算法主要可以分为人工辨识、数据库字段修正、基于机器学习的辨识以及其他4类。

人工辨识相对可靠，但是效率偏低，已经无法满足数据日益增长的需求。而且，由于人为差异会造成结果充满不确定性，如Smalheiser和Torvik随机选取了MEDLINE中的100名作者和他们的一篇论文，利用可以获取的所有信息用人工辨识，结果发现两位辨识者对1/3的论文有不同意见[41]。另外，在面对大量与自己重名的论文时，作者本人也可能弄错自己的论文。

基于作者互动的数据库字段修正辨识方法是人们在进行重名辨识时首先会想到的方法，技术上也容易实现，但是，这种机制却存在许多问题。首先就是只有数据库厂商才有可能实现；其次依赖于作者的参与，但是这个机制却不太可行，我们无法激励作者们自愿、主动、正确无误、周期性地去更新其信息，Garfield E.指出很难说服每位作者发表论文时提供完整的姓名，如中间姓（middle initial）等[42]，况且是要花费时间维护自己的信息呢。Torvik和Smalheiser指出，MEDLINE中有46%的作者只发表一篇论文[43]，要这些人参与这类型计划是不容易的，因为他们通常不觉得参与其中会得到任何实质的好处。另外，理想上，这种人名ID数据库的服务范围应该是扩及所有国家、任何语言、各个学科以及不同的数据类型。随着越来越多的机构或服务单位建立这些ID后，每个作者拥有多个ID，使得这种ID的人名辨识价值也随之降低[44]。

作者重名辨识问题实质上是将同名作者的论文进行分类或者聚类。所以基于机器学习实现作者重名的自动识别，是当前研究中主要方向。基于机器学习的重名辨识离不开训练数据集、特征、特征选择与合并和结果评估。

监督式学习需要依据训练数据集，训练数据集是产出预测模型的重要因素，可以自动产生，也可以人工编制。由于可以事先调整训练集，所以分类结果一般较好，但是要注意训练集的代表性和完整性，并且要求多次学习，不断更新辨识模型。另外，监督式学习一般会将数据集分成一块一块的区块，这样可以仅处理同名数据集，减少计算与比对的时间。

特征可以通过计算字符串间的距离、地理上的距离、对应实体属性之间的距离得到，在计算距离时又可以用不同的权重计算方式得到，所以导致基于机器学习的重名辨识方法复杂多样，精彩纷呈。特征选择是基于机器学习的重名辨识方法的核心，一般来说，特征越多越好，过少的特征会造成低的查全率。不过在绝大多数的研究中，并不会只采单一的特征，因此特征合并就显得格外重要。特征合并主要解决将所有特征分数转为单一权重值（应确保特征间的独立性），以便比较与计算两两纪录之间的相似程度问题。

另外，在选择无监督学习的方法时，要注意一些方法会使用低复杂度的数据集，或只给出事先规划的相应参数的结果，其实际效率不是很高，要注意区分高效率是算法带来的还是其他带来的。

5 今后主要的研究方向

国内外研究者提出了大量的作者重名辨识技术，但是随数据库的规模、作者姓名的复杂性、数据库的格式、论文语言等要素的变化而影响特定技术的识别能力和效果，特别是有些方法还需要一些私密信息。因此目前尚未有一以贯之的典型辨识方法，急需发展一个基于各数据通用信息的作者重名辨识方法。

随着数据量的急速增长，基于机器学习的方法是未来发展的主要方向，目前急需在现有算法基础上发展高效率、高准确率的作者重名辨识方法。

随着信息社会的到来，新的文献不仅有纸版期刊和图书出版，电子版的期刊图书也越来越多，特别是一些开放存取的论文仅在网络上发表。对于电子版论文、图书作者的重名辨识需要发展在线辨识研究，目前该领域尚未见到有关的论文。

另外，从国内外研究来看，当前对作者重名辨识的研究一般都是基于单一数据库、单一语言进行的。在科技全球化时代，随着国际科技合作、海外留学人员归国、国际访问学者交流日益增多，越来越多的中国作者同时发表中文、外文论文，并且在将来这一现象会更加突出。这样导致中国作者论文数据来源形式多样、发表论文的语言多种，数据较为复杂。集成中、英文论文作者重名辨识技术的研究将是中国作者重名辨识发展的方向之一。

参考文献

[1] SHIFFRIN R M, BÖRNER K. Mapping knowledge domains [C]// Proceedings of the National Academy of Sciences of the United States of America,2004:5183-5185.
[2] 陈佳宜. 中国重名最多50姓名公布,近30万人叫张伟[EB/OL]. [2011-03-18]. http://newssinacomcn/s/2007-07-25/144113525318shtml.
[3] BILENKO M, MOONEY R, COHEN W, et al. Adaptive Name Matching in Information Integration [J]. IEEE Intelligent Systems,2003,18(5):16-23.
[4] 郭绍武. 候选人重名怎么办[J]. 乡镇论坛,2002(21).
[5] MAXWELL R L. Maxwell's guide to authority work [C]// Chicago: ALA, 2002.
[6] DEROSE P, SHEN W, CHEN F, et al. DBLife: A Community Information Management Platform for the Database Research Community (Demo) [C]// CIDR 2007, Third Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, 2007:169-172.
[7] GUHA R V, GARG A. Disambiguating People in Search [C]// Proceedings of the 13th World Wide Web Conference (WWW 2004), ACM Press, 2004.
[8] XIA J F. Personal name identification in the practice of digital repositories [J]. Program: Electronic Library & Information Systems,2006,40(3):256-267.
[9] DERVOS D A, SAMARAS N, EVANGELIDIS G, et al. The Universal Author Identifier System (UAI_Sys) [C/OL]// Proceedings 1st International Scientific Conference, eRA: The Contribution of Information Technology in Science, Economy, Society and Education; 2007 [2011-03-09]. http://dlist.sir.arizona.edu/1716/ 2007.
[10] HAN H, GILES L, ZHA H, et al. Two supervised learning approaches for name disambiguation in author citations [C]// Proceedings of the Fourth ACM/IEEE Joint Conference on Digital Libraries (IEEE Cat No04TH8766), 2004: 296-305|xiv+429.
[11] 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.
[12] YIN X X, HAN J W, YU P S. Object Distinction: Distinguishing Objects with Identical Names [C]// Data Engineering, 2007 ICDE 2007 IEEE 23rd International Conference, 2007: 1242-1246.
[13] HAN H, ZHA H Y, GILES C. L. Name disambiguation spectral in author citations using a K-way clustering method [C]// Proceedings of the 5th Acm/Ieee Joint Conference on Digital Libraries, Proceedings, 2005: 334-343.
[14] HUANG J, ERTEKIN S, GILES C L. Efficient name disambiguation for large-scale databases [C]// Knowledge Discovery in Databases: PKDD 2006 10th European Conference on Principle and Practice of Knowledge Discovery in Databases Proceedings (Lecture Notes in Artificial Intelligence Vol 4213), 2006: 536-544|xxii+660.
[15] WANG H F, MEI Z. Chinese multi-document personal name disambiguation [J]. High Technology Letters (English Language Edition), 2005, 11(3): 280-283.
[16] CULOTTA A, MCCALLUM A. Tractable learning and inference with high-order representations[C]// ICML Workshop on Open Problems in Statistical Relational Learning, 2006.
[17] MCRAE-SPENCER D M, SHADBOLT N R. A citation graph approach to name disambiguation [C]// 2006 IEEE/ACM 6th Joint Conference on Digital Libraries, 2006: 2 pp.|CD-ROM.
[18] SONG Y, HUANG J, COUNCILL I G, et al. Efficient Topic-based Unsupervised Name Disambiguation [C]// Proceedings of the 7th Acm/Iee Joint Conference on Digital Libraries, 2007: 342-351.
[19] BHATTACHARYA I, GETOOR L. Collective Entity Resolution In Relational Data [C]// ACM Transactions on Knowledge Discovery from Data, 2007, 1(1): 1-36.
[20] JOSÉ S. Separating the articles of authors with the same name [J]. Scientometrics, 2007, 72(2): 281-290.
[21] 吴斌,徐超群,王文彬,等. 基于链接的作者重名处理方法研究与应用[J]. 计算机科学,2008,35(3):197-199.
[22] 索利军,吴斌. 生命科学领域科研合作网的分析[J]. 数字图书馆论坛,2008(6):2-6.
[23] FAN X M, WANG J Y, LV B, et al. GHOST: an effective graph-based framework for name distinction [C]// CIKM 2008, Napa Valley, California, USA, 2008: 1449-1450.
[24] 陈威,,王仲远. C-DBLP: 中文文献信息集成系统[EB/OL]. [2011-03-19]. idkeruceducn/reports/report2008/Systems/C-DBLP.pdf.
[25] KANG I S, NA S H, LEE S, et al. On co-authorship for author disambiguation [J]. Information Processing & Management, 2009, 45(1): 84-97.
[26] JIN H, HUANG L, YUAN P P. Name disambiguation using semantic association clustering [C]// 2009 IEEE International Conference on e-Business Engineering ICEBE 2009, Macau, China, 2009: 42-48.

[27] ZHU J, ZHOU X F, FUNG G P C. A term-based driven clustering approach for name disambiguation [C]// Advances in Data and Web Management Proceedings Joint International Conferences, APWeb/WAIM 2009, 2009: 320-331.
[28] FERREIRA A A, GONCALVES M A, ALMEIDA J M, et al. SyGAR - A Synthetic Data Generator for Evaluating Name Disambiguation Methods [J]. Research and Advanced Technology for Digital Libraries, 2009 (5714): 437-441.
[29] TANG L, WALSH J P. Bibliometric fingerprints: name disambiguation based on approximate structure equivalence of cognitive maps [J]. Scientometrics, 2010, 84(3): 763-784.
[30] COTA R G, FERREIRA A A, NASCIMENTO C, et al. An Unsupervised Heuristic-Based Hierarchical Method for Name Disambiguation in Bibliographic Citations [J]. Journal of the American Society for Information Science and Technology, 2010, 61(9): 1853-1870.
[31] MANN G S, YAROWSKY D. Unsupervised personal name disambiguation [C]// Proceedings of CoNLL-7, 2003: 33-40.
[32] ASWANI N, BONTCHEVA K, CUNNINGHAM H. Mining information for instance unification [C]// The Semantic Web - ISWC 2006 OTM 2006 Workshops 5th International Semantic Web Conference, ISWC 2006 Proceedings (Lecture Notes in Computer Science Vol 4273), 2006: 329-342|xxiv+1001.
[33] TANG J, ZHANG D, YAO L M. Social network extraction of academic researchers[C]// ICDM 2007: Proceedings of the Seventh Ieee International Conference on Data Mining, 2007: 292-301.
[34] YANG K H, PENG H T, JIANG J Y, et al. Author Name Disambiguation for Citations Using Topic and Web Correlation [J]. Research and Advanced Technology for Digital Libraries, 2008, 5173: 185-196.
[35] JIANG L L, WANG J Y, AN N, et al. GRAPE: a graph-based framework for disambiguating people appearances in Web search [C]// Proceedings of the 2009 Ninth IEEE International Conference on Data Mining (ICDM 2009), 2009: 199-208.
[36] ON BYUNG-WON. Social network analysis on name disambiguation and more [C]// Third International Conference on Convergence and Hybrid Information Technology (ICCIT), 2008: 1081-1088.
[37] 郎君,秦兵,宋巍,等. 基于社会网络的人名检索结果重名消解[J]. 计算机学报,2009,32(7):1365-1374.
[38] LEVIN F H, HEUSER C A. Evaluating the Use of Social Networks in Author Name Disambiguation in Digital Libraries [J]. Journal of Information and Data Management. 2010, 1(2): 183-197.
[39] CULOTTA A, KANANI P, HALL R, et al. Author disambiguation using error-driven machine learning with a ranking loss function [C]// Proceedings of the AAAI 6th International Workshop on Information Integration on the Web, 2007: 32-37.
[40] CHURCHES T, CHRISTEN P, LIM K, et al. Preparation of name and address data for record linkage using hidden Markov models [J]. BMC Medical Informatics and Decision Making, 2002 (2): 9.
[41] SMALHEISER N R, TORVIK V I. Author Name Disambiguation [J]. Annual Review of Information Science and Technology, 2009 (43): 287-313.
[42] GARFIELD E. British quest for uniqueness versus American egocentrism[J]. Nature. 1969 (223): 763.
[43] TORVIK V I, SMALHEISER N R. Author name disambiguation in MEDLINE [J]. ACM Trans Knowl Discov Data, 2009, 3(3): 1-29.
[44] MERALI Z, GILES J. Databases in peril [J]. Nature,2005(435):1010-1011.

作者简介

袁军鹏（1973-），男，博士，副研究员，研究方向：情报学，科学计量学，科技政策。E-mail: junpengyuan@gmail.com
宿洁（1975-），女，博士，副教授，主要研究方向：优化理论与算法。
俞征鹿（1980-），女，硕士，助理研究员，主要研究方向：科学计量学、科技评价。
苏成（1973-），男，硕士，副研究员，主要研究方向：科学计量学、科技评价。
马峥（1975-），男，副研究员，主要研究方向：科学计量学、科技评价。
杨志清（1957-），女，副编审，主要研究方向：情报学、科技评价。

A Survey of Author Name Disambiguation

Yuan Junpeng, Yu Zhenglu, Su Cheng, Ma Zheng, Yang Zhiqing / Institute of Scientific and Technical Information of China, Beijing, 100038
Su Jie / School of Management Science and Engineering, Central University of Finance and Economics, Beijing, 100081

Abstract: Because of name variations, an author may have multiple names and multiple authors may share the same name. Name disambiguation affects the performance of document retrieval, web search, database integration, and may cause improper attribution to authors. This paper makes a thorough investigation of the whole problem, analyses the current characteristics of the various disambiguation methods and points out the direction of name disambiguation development.

Keywords: Name disambiguation, Machine learning, Social network, Clustering

（收稿日期：2011-07-14）