语义指纹著者姓名消歧理论及应用

韩红旗 ◎ 著

科学技术文献出版社
SCIENTIFIC AND TECHNICAL DOCUMENTATION PRESS

本专著从事的研究受到国家自然科学基金项目“科学合作网络的不连通问题研究”（No.71473237）资助。

本专著是国家新闻出版广电总局首批新闻出版业科技与标准重点实验室“富媒体数字出版内容组织与知识服务重点实验室”的研究成果，受中国科学技术信息研究所“富媒体数字出版内容组织与知识服务关键技术研发与应用示范”项目（ZD2018-07）资助。

本专著部分研究内容是中国工程科技知识中心建设项目“知识组织体系建设”(CKCEST-2017-1-12，CKCEST-2018-1-26)的研究成果。

语义指纹著者姓名消歧理论及应用

Author Name Disambiguation Based on Semantic Fingerprint: Theory and Applications

韩红旗 著

科学技术文献出版社
SCIENTIFIC AND TECHNICAL DOCUMENTATION PRESS

·北京·

图书在版编目（CIP）数据

语义指纹著者姓名消歧理论及应用 / 韩红旗著. —北京：科学技术文献出版社，2018.7
ISBN 978-7-5189-4594-8

Ⅰ.①语… Ⅱ.①韩… Ⅲ.①信息检索 Ⅳ.① G254.9

中国版本图书馆 CIP 数据核字（2018）第 140314 号

语义指纹著者姓名消歧理论及应用

策划编辑：周国臻　　责任编辑：李　晴　　责任校对：张吲哚　　责任出版：张志平

出 版 者　科学技术文献出版社
地　　址　北京市复兴路15号　邮编 100038
编 务 部　(010) 58882938，58882087（传真）
发 行 部　(010) 58882868，58882870（传真）
邮 购 部　(010) 58882873
官方网址　www.stdp.com.cn
发 行 者　科学技术文献出版社发行　全国各地新华书店经销
印 刷 者　北京教图印刷有限公司
版　　次　2018 年 7 月第 1 版　2018 年 7 月第 1 次印刷
开　　本　710×1000　1/16
字　　数　220千
印　　张　14.25
书　　号　ISBN 978-7-5189-4594-8
定　　价　68.00元

自　序

大学期间，曾受几位老师和两位非常有科研天分的学长指点和提携，参加了几个项目的研发工作，做了一点跟科研相关的事情。然而，毕业后并没有走向科研道路，不适合做商业工作的我进入了一个小型的国有企业，成为一名工程师，然后成为一名市场营销人员，将10年最宝贵的青春年华浪费在了绝对不擅长的商业领域。

直到过了而立之年，迷茫之中的我决定读研究生学点管理学知识，以弥补自己欠缺的商业能力。然而毕业后我没有再回到企业，而是阴差阳错地进入了一所地方高校，成为一名光荣的人民教师。这时，环顾四周，我才发现在这个知识分子成堆的地方，自己学到的商业知识不重要，科研能力才是未来赖以生存的能力。

为此，我决定攻读博士学位，转向科研道路，开始了个人转型之路。这期间的酸甜苦辣、压力、困难、挫折难以言表。

在转到科研道路后，一直想将自己的科研成果总结一下，写一本书。这个念头出现之后，曾多次想启动这个工作，但一方面觉得研究尚浅，实在拿不出手，不适合成书；另一方面，编书是一件折磨人的事情，而我不是一个自律性很强的人，如果没有压力和动力，我是很难忍受长时间的枯燥工作的。为此，我在2014年申请的国家自然科学基金项目“科学合作网络的不连通问题研究”中有意设立了一个指标，就是完成一本图书，逼着自己下功夫来完成这个“艰苦卓绝”的工作。

转眼间，3年过去了，指标中的图书还没有踪影。眼看项目周期的最后一年临近，迫不得已，才逼着自己静下心来考虑如何完成这个指标，至于编一本“流芳百世”的书的想法完全烟消云散了。好在这2年我自己在科学合作网络上有了一些研究的积累，我指导的几个硕士生和同研究室的同事也有了一些成果提供了基本的素材让我来编纂这本图书了。

申请的自然科学基金项目是科学合作网络领域的。科学合作网络是一个跟人物有关的网络，准确地说，是科研工作者由于合作而产生的人物社会网络。我想做跟人物相关的研究有两个原因。一个原因源自于我的博士导师北京理工大学朱东华教授给我的一个教诲，他说做科技监测要“三盯”，即“盯人、盯机构、盯项目”。“三盯”让我们知道一个科研人员、一个科研机构、一个国家在做什么、发展什么，我想把“盯人”这个先做了，因为人是科研的第一要素，没有“人”就没有科研、就没有创新，而且机构的研究是靠科研人员做的，项目也是科研人员申请的。另一个原因源自于我到中国科学技术信息研究所工作以后承担的一个国家科技支撑计划课题，这个课题的名字是“基于多源信息的电动汽车数据挖掘关键技术研究”。我在这个课题中承担了一个任务，是关于学术社会网络的相关技术研究的。学术社会网络是由一个个科研人员构成的，通过社会网络“盯人”是一个好的方法，而这正是我愿意做的一个方向。

在研究中我发现构建科学合作网络的第一个问题是姓名消歧。只有将重名人员“各归其位”，构建的网络才更准确，分析的结论才更可靠。我在进入中国科学技术信息研究所工作之前一直在做文本挖掘相关的研究，对社会网络知之甚少，对姓名消歧的难度

也没有深刻认识。所以，当一个学术造诣深厚的同事跟我说“姓名消歧是世界难题”时，我不以为然，曾“豪情万丈”地说这个问题将被我们解决。然而，随着对姓名消歧研究的逐渐深入，我才真正认识到这个问题的复杂性和难度，暗暗为当时的吹牛脸红。姓名消歧问题的复杂性和难度不仅在于现实中有些姓名重名的人很多、科研人员的流动性很强、有些重名人员可能只有1篇或2篇论文等客观条件，更在于其技术上的难度。首先，聚类的数量没有先验知识，歧义程度不服从正态分布，网页是噪声源，属性和其他索引很难抽取；其次，姓名消歧涉及文本聚类、信息抽取、词义区分等复杂的自然语言处理和信息检索技术，当前的姓名唯一标识系统也难以从根本上解决这个问题，而跨学科研究重名人员、姓名的变体、跨语言姓名消歧等更增加了这个问题解决的难度。

我们在前人研究的基础上，提出了一套以语义指纹技术为基础的著者姓名消歧方法，达到了较好的姓名消歧效果，较同类研究在时间成本、运算代价和存储空间需求上均有明显的优势，具有较强的实用性。现将研究成果编纂成书，一方面希望能为后来者研究这个问题提供一点资料和想法，算是抛砖引玉吧；另一方面我们出版此书也希望更多的学者来继续关注这一问题，让我们在互联网的海洋中不迷失自己，能知道我是谁、他是谁。

尽管我们在编纂此书过程中力求严谨，但受个人能力、时间所限，本书难免存在不足甚至是错误之处，恳请读者批评指正。任何咨询、批评、建议可以发邮件给我，邮箱地址为 bithhq@163.com。

感谢那些为本书提供了大量素材的研究室同事和研究生们。

如果本书有些亮点，那一定是你们的功劳；如果本书有问题，那一定是我的粗心和马虎。感谢国家自然科学基金委员会给予我们项目的支持，才让我们有条件完成这本书的编纂。感谢在这一课题研究过程中给予我们指导和批评的专家和学者，正是你们的真知灼见让我们的研究得以完善。感谢我所在工作单位的领导和同事，正是你们给了我支持，才让我完成了这项对我来说很艰难的工作。感谢我挚爱的妻子和儿子，由于你们对家庭的默默付出，以及对我工作无底线的支持，我才有了大量时间来完成这一工作。

韩红旗

2018 年 2 月于北京

前　言

狭义上讲，科学合作网络（Scientific Collaboration Network）是指以论文或专利数据中的合著关系为基础构建的社会网络。论文的作者、专利的发明人在本书统一称为著者或作者。科学合作网络中，节点是著者，边是著者之间的合著关系（Co-authorship）。虽然也有研究国家或地区间合作的科学合作网络，但本研究限定其为著者合著关系构建的网络。科学合作网络中的合著关系是一种强社会关系，合著者一般认识，往往是同一个机构、同一个项目、同一个科研工作中的“同事”，或者通过其他合著者间接认识，所以它在揭示科研工作者的关系、发现科研合作社区、提升学术信息检索质量、评价科研人员的能力、提供学术推荐和科研合作建议、服务科研论文和项目评审等方面有着重要的应用，从而受到了不少研究人员的关注。

在对科学合作网络进行研究的过程中，我们发现构建的科学合作网络由很多不连通的子网络构成，其中通常存在一个作者数量很多的“中心网络”和很多作者数量较少的“边缘网络”。文献调研中发现，虽然不少研究者已经注意到构建的科学合作网络是由很多不连通的子网络构成的，但并未就其进行深入研究。科学合作网络的不连通问题会在一些实际应用产生错误的结果，影响其应用效果。为此，本项目拟对不连通问题的原因、不连通问题对实际应用的影响进行分析，探索提高科学合作网络连通性的方法。基于不连通是因为数据不全面的假设，对中英双语言体系

下的跨数据库姓名消歧问题进行研究，在此基础上，对采用关系扩展方法提高科学合作网络连通性的方法进行研究。基于关系扩展法成本高、效率低等不足，对采用链接预测算法提高科学合作网络连通性的方法进行研究。通过对科学合作网络不连通问题的研究，探索提高连通性的方法，希望为科学合作网络的深入研究和应用提供新的理论或方法的指导。

在科学合作网络研究中，著者姓名消歧是普遍认可的一个重要步骤，也是一个关键步骤。来源于传统图书馆权威控制（Authority Control）思想的人工消歧固然是一种较为可靠的方法，但该方法效率较低，人为因素可能造成消歧效果充满不确定性，使其难以满足文献数据量激增和数字图书馆服务及时化的需求。因此，采用自动化的姓名消歧技术是一种更为现实的解决方案，也是当前的研究热点和重点。自动姓名消歧技术是自然语言处理的基本问题之一，最初是作为实体共指现象来研究的，后来在一些会议和评测竞赛的推动下，姓名消歧作为一个单独的研究问题被提出来。提出的众多自动化消歧技术大体上可以落入无监督和有监督的机器学习的范畴。虽然很多算法被提出来以解决这个问题，但迄今为止不少研究者认为它依然是学术界的一大难题。此外，当前的自动化处理技术多研究静态数据环境下的姓名消歧，大多需要较大的运算量，较少考虑真实的、动态的数字图书馆情景下的消歧。

随着数字图书馆、互联网的进一步发展，姓名歧义问题越来越得到重视。对于著者姓名消歧来说，图书情报领域的国内外研究人员和机构提出构建一个注册系统，为每一名科研人员分配唯一的 ID，这样就能从根本上解决著者姓名歧义问题。表面上这个

提议很好，但有些专家指出，该系统不符合现实世界的人类行为，没有考虑到基于 Web 的资源天性是脆弱的，也没有考虑到这样一个注册系统谁愿意来长期资助并维护，因此这样的系统并不能从根本上解决姓名歧义问题。尽管国内外一些机构建立了 ResearcherID、ORCID、ScholarID 等科研人员唯一标识系统，对这种思想进行了具体实践，然而数年过去了，仍未见该提议起到根本的效果。

我们在语义指纹、著者姓名消歧相关理论和应用的研究过程中，以及在本书的编写过程中，参考了大量国内外出版物和网上资料，在此谨向各位作者表示由衷的敬意和感谢。我们尽可能地遵循学术规范要求，将有关观点、文字、图表的引用出处列入参考文献中，如有疏漏，也请谅解。

本书是团队工作的结晶，参加编写的人员主要来自于中国科学技术信息研究所知识组织与知识工程研究室的研究人员和研究生。研究人员主要有韩红旗、姚长青、董诚、刘志辉、张运良、李琳娜、王莉军、李颖、刘泸颖、王力、高影繁、张兆锋、桂婕，研究生主要有付媛、于永胜、李仲、翟晓瑞。其中，姚长青、董诚、刘志辉、张运良对本书的编纂和研究工作提出了具体的意见和指导，李琳娜、王莉军、李颖、刘泸颖、王力、高影繁、张兆锋、桂婕、付媛、于永胜、李仲、翟晓瑞等参与了前期的资料收集和整理及文字内容的检查、修订工作，付出了辛勤的劳动，在此表示最诚挚的谢意。

目　录

图表目录