情报工程 ISSN 2095-915X
第2卷 第1期 53-58 2016年2月

TECHNOLOGY INTELLIGENCE ENGINEERING
ISSN 2095-915X Vol.2 No.1 53-58 Feb 2016

doi:10.3772/j.issn.2095-915x.2016.01.007

姓名消歧方法研究进展

付媛，朱礼军，韩红旗

（中国科学技术信息研究所 北京 100038）

摘要：为应对日益严重的姓名歧义现象给提高搜索引擎查全率和查准率带来的挑战，同时给姓名消歧方法研究提供参考建议，对研究现状和主要成果进行总结。首先，介绍研究姓名消歧的目的和意义。其次，对国内外现有姓名消歧方法研究进展进行梳理，主要方法包括基于特征的、基于机器学习的、基于社会网络的、基于网络知识资源的姓名消歧等多种方法来解决姓名歧义问题。最后，文章分析各种方法的特征和不足，总结姓名消歧待解决的问题以及未来的研究方向。

关键词：姓名消歧，机器学习，聚类

分类号：G35

A Survey of Name Disambiguation

FU Yuan, ZHU Lijun, HAN Hongqi

(Institute of Scientific and Technical Information of China, Beijing China, 100038)

Abstract: The paper studics the increasingly serious name ambiguity problem to improve the recall ratio and precision ratio of search engine, and provides suggestions for future research. Firstly, the purpose and meaning of name disambiguation are introduced. Secondly, the researches on name disambiguation methods in China and abroad are summarized. Main methods include those based on feature, machine learning, social network and the Internet resources to solve the problem of name ambiguity. Finally, the paper concludes the problem of name disambiguation to be solved and points out the direction of name disambiguation development.

Keywords: Name disambiguation, machine learning, clustering

作者简介：付媛，研究生，研究方向：数据挖掘，知识组织，中国科学技术信息研究所，E-mail：fuyuan2014@istic.ac.cn。朱礼军：男，博士，中国科学技术信息研究所，研究方向：Semantic Web、Web service 知识技术在科技信息服务、电子政务/商务中的应用以及知识组织系统相关研究。韩红旗：男，博士，中国科学技术信息研究所，研究方向：科学社会网络，文本挖掘。
基金项目：国家“十二五”科技支撑计划“面向科技情报分析的信息服务资源与开发与支撑技术研究”（编号：2015BAH25F01）。

1 引言

随着互联网技术迅猛发展，信息爆炸式增长，如何从海量数据中高效快速地找到自己所需要的信息成为信息检索的重要目标，同时用户对信息检索查准率和查全率也提出了更高的要求。查找人名信息是用户在网络上搜索的主要目的之一，是信息查找的关键点，也是信息查询的关键。然而由于姓名歧义现象严重，搜索结果并不能对有歧义的人名有效地组织信息，用户往往需要耗费大量时间来筛选自己感兴趣的人物信息。为此，消除姓名歧义成为了近年来国内外学者的研究热点之一。

姓名歧义性普遍存在于各个国家和各种语言中，指的是不同的实体人物拥有同样的姓名，对于某一姓名的查询结果为具有相同姓名但不同人物实体内容的混合，使得用户对返回结果产生混淆，却难以区分的问题。且当前，网页中的内容在以前所未有的速度增长，人名歧义现象降低了搜索引擎检索的准确性。因此，如何有效的消除人名歧义尤为重要。

姓名消歧是在这样的背景下提出来的，它旨在消除跨文档情况下的人名歧义性，把相同的人名按照现实世界的不同实体进行分类，从而把信息有效地组织和聚类后提供给用户[1]。文献著者姓名消歧也是多文档人名消歧的一种，判断出现在不同文献中的相同作者名是否指向现实生活中的同一个人的处理过程，即将指向现实中的同一个人的文献聚类。有效的文献著者姓名消歧是文献处理、情报工作的基础，也是自然语言处理中姓名识别的一个必要的前提和必不可少的环节，该问题在搜索引擎检索、数据挖掘、人名知识库构建等领域中广泛应用。人名歧义问题是一个亟待解决的问题，需要一个有效的消歧方法。

人名消歧与词义消歧要求较相似，都是解决自然语言处理问题中出现的歧义现象。但是二者也存在很大的差异，具体表现为对于词义消歧，歧义词语所对应的语义数目是固定且已知的，而歧义姓名对应的人物数目是未知的、不确定的[2]。因此，人名消歧比其他命名实体消歧更为复杂，解决词义消歧的方法不能完全应用于姓名消歧。

2 姓名消歧面临的主要挑战

（1）同一个作者可能会使用多个名字。这种情况的原因可能是外文姓名拼写的变体（名字全称或者缩写）、拼写或者印刷错误、使用笔名以及作者有曾用名等。（2）同一个姓名对应多个作者。在现实中，重名现象非常严重，多个作者对应相同的姓名非常普遍。据一项研究报告显示，有10亿人却仅仅使用了90000个不同的名字[3]。重名现象的普遍性导致了姓名歧义现象严重。由于中文人名用字较为集中，中文姓名具有非常高的歧义性，给中文人名信息查找带来了很大的挑战。（3）大量的文本信息中包含的人物信息不完整，缺乏充足的信息来匹配和识别作者[4]，因此，如何收集更多有效信息也是一个难点。(4) 网页文本格式复杂多样，大多数为非结构化文本，在网页中抽取相关性较高的词语较为困难。

3 姓名消歧方法研究现状

随着Bagga和Baldwin与1998年首次提出的姓名消歧进行研究并逐渐引起人们的关注，越来越多的方法应用到姓名消歧当中。作为自然语言处理中的一个重要部分，姓名消歧的研究也一直在不断地深入，已经有了丰富的研究成果，大致常用方法总结如下：

3.1 基于特征的姓名消歧方法研究现状

基于特征的姓名消歧是一类常用的消歧方法，它研究的重点是，首先尽可能选择一些具有

较大区分度的人名的相关特征；然后进一步优化特征、提取特征，排除掉无关特征，从而选择出最佳特征作为特征子集，同时达到了空间降维的目的；最后，构建姓名消歧模型，选择合适的算法，得到消歧结果。如 2003 年，Mann[5] 等提出了充分利用人物传记信息，提取相关特征，并且使用了一种无监督的层次聚类算法对网页中的人名进行消歧。其核心思想为：先对网页进行预处理，使用一定的规则抽取个人信息，得到人物传记特征，如 Email、生日、职业、机构、电话号码等信息；最后利用所抽取的人物特征进行层次聚类，将具有相同姓名但不同的人物区分开。首次探讨姓名消歧的 Bagga[6] 等，通过抽取待分析实体生成的指代链的有关信息，得到实体摘要，将摘要作为相关特征，并采用 VSM 对文本聚类。Han[7] 等对 DBLP 中的作者进行姓名消歧，首先，抽取文章的作者信息、题目和会议名称等，将其表示为特征向量；然后，计算文章相似度并生成矩阵；最后，采用 QR 分解和特征值分析，并采用“K—Way Spectral”算法对文章进行聚类，最终实现姓名消歧。Bollegala[8] 等丰富了可用特征，考虑到了真实文本环境中人物信息不足的问题，突破了仅有的人物信息的限制，提取了文档中相关实体名等关键词。杨欣欣 [9] 提出一种借助于丰富的网络资源，采用查询扩展方法，使用搜索引擎查询，搜集大量与文本人物相关的特征，克服了本身特征稀疏文本不适用于基于特征的姓名消歧方法的问题，最后对特征两步聚类，先用相关性较强的特征对文本聚类，再用辅助特征聚类，实现较好的姓名消歧结果。王英帅 [10] 基于 CSS 技术的网页内容提取，在浅层语义层次分析特征，探讨了根据网页主题内容相关性和名字上下文噪音小等特性，提出一种基于主题模型 LDA 和上下文摘要聚类相结合的 Web 人名消歧方法。Chen[11] 等首先对文本特征抽取，采用 Soft—TFIDF 对特征量化，计算相似度，并开发了 PolyUHK 系统，最后，用无监督的层次凝聚聚类方法进行聚类，达到姓名消歧的目的。总之，基于特征的姓名消歧分为两种：基于内容的名词短语特征和人物实体特征。名词短语特征也分为：基于全文的名词短语以及基于摘要的名词短语。

3.2 基于机器学习的姓名消歧方法现状研究

机器学习分为有监督和无监督的，即为分类方法和聚类方法。聚类分析是一种无监督的学习方法，不需要预先对文档进行手工标注，不依赖于训练数据，具有较高的灵活性。聚类方法多种多样，如何选择有效的聚类方法是姓名消歧的研究方向之一。朱亮亮 [12] 对 k-means 算法进行研究改进，提出了新的 k-means 算法实现文献著者姓名消歧，新的方法根据最大最小原则选择初始聚点，突破了传统 k-means 算法对初始聚类中心敏感，不同的初始中心对应不同的聚类结果的问题，使 k-means 聚类算法得到优化，从而提高了消歧效果。任景华 [13] 提出了改进的 DBSCAN 算法，优化了 DBSCAN 算法初始参数选择，使参数的决定更具客观性，实现了在大容量数据集中完成姓名消歧。丁海波 [14] 等通过多次使用聚类方法，得到更优的消歧结果。作者提出了多阶段聚类，采用较确定的人物传记属性进行第一步聚类，在第一次聚类结果基础上，结合上下文特征进行第二步聚类，克服了第一阶段人物传记属性数据稀疏的问题，从而召回更多的正确结果。Wang[15] 等改进了常规的基于聚类的姓名消歧方法，提出了基于两步策略的自适应共振理论来解决姓名消歧的问题，该方法在第一步进行对待消歧姓名字符串进行聚类，第二步合并相似簇。Xu[16] 提出了先使用层次聚类方法，然后通过基于中心的方法找出聚类结果簇中的偏离点，再利用基于短语的字符串，采用支持向量机方法，使更相似的文本聚为一簇。

基于分类的姓名消歧方法是一种监督学习方法，对每个排歧目标进行训练和学习来建立相应的模型，再利用模型实现分类的目的，对于任何一个预测性模型，足够的训练数据是很关键的，且训练数据应该代表全部数据而不是只代表一部分数据特征。Huang[17] 等利用 Online—SVM 分类学习算法计算文献之间相似度，再用 DBSCAN 聚类算法实现作者消歧。Han[18] 等首先人工构建训练集，之后利用朴素贝叶斯概率模型和支持向量机分类算法解决引文中的作者姓名消歧问题。该类方法需要人工构建训练集，面对海量数据进行人工标注非常困难，从而限制了该方法在姓名消歧中的应用。

3.3 基于社会网络的姓名消歧方法研究现状

基于成员在社会网络中的关系来进行姓名消歧是目前一个新的研究趋势，这类方法首先要建立待消歧人名的相关社会网络图，转化为了图论的问题，根据人物关系进行姓名消歧。如何正确有效地利用社会网络还有待进一步研究。如 Malin[19] 首先挖掘和构建待消歧人物的社会网络，通过局部社会网络计算相似度，并采用层次聚类进行划分；第二种方法基于全局社会网络度量相似度，采用随机游走的方法实现类别划分，实现姓名消歧。郎君 [20] 等采用基于社会网络的姓名消歧方法，认为不同人物一定具有各自的不同的社会网络，结合社会网络信息进行姓名消歧。首先要发现并探索待消歧人物的潜在社会网络，转化为图论问题，结合图的分割算法进行社会网络聚类，从而实现人名检索结果的姓名消歧。Tang[21] 等同样是构建待消歧人名的社会网络，并将其表示成二部图，提出了一种基于二部图的社会网络相似度计算方法，根据相似度进行自底向上聚类，该方法在 WePS-2 人名消歧任务的测试集上取得了高于最好评测结果的好成绩。

3.4 基于网络知识资源姓名消歧方法研究现状

这类方法利用网络上现有的公开资源，构建新的规则和类别，使待消歧姓名与现实世界中人物信息中区分度较强且准确的社会属性建立联系，得到更丰富的人物特征，结合社会属性进行分类，从而达到消歧的目的。Han[22] 等从 Freebase 中抽取了待消歧人物的职业目录，利用粒度适中的近两千种职业分类的相关文档构建训练数据，采用 KNN 方法将职业文档分类，进而通过职业的异同确定各个文档中不同的人物实体。Vu[23] 等通过引入在线的网络资源有效地丰富了传统的文档特征，作者提出了一种新的文档相似度比较标准，充分利用了互联网用户认可的文档类别目录，使待消歧文档与各个目录类别的建立联系，以及结合各个文档间的文本特征，在传统 tf − idf 特征的基础上增加了相关目录参数，以达到改善文档相似度计算评价的目的。周晓 [24] 等人利用常识信息人工设定规则，通过互斥信息来判断待消歧人物信息中不可能属于同一人物实体的情况，避免了相似的文档错分到同一类别中，以达到更好的消歧效果。如 Bunescu[25] 等引入了维基百科的相关信息，并生成语义字典，通过将待消歧姓名的文档特征映射到字典的相应条目上，以达到姓名消歧的目的。Han[26] 等同样是利用 Wikipedia 构建了大规模的语义网络，并将共现的人物信息向量化，最后通过凝聚层次算法。

4 结束语

国内外学者提出了许多的姓名消歧方法，随着日趋庞大的数据规模，以及人名的复杂性、数据库格式等要素的变化而影响特定方法的姓名消

歧的效果。每种方法都有其特点和不足，基于特征的姓名消歧方法往往受到信息量的限制以及难以对传记信息准确抽取和定位，且大量的文档中没有足够的人物传记信息，此类方法的改进主要集中在特征的选择和提取；聚类分析是一种无监督的学习方法，不依赖预先设定的训练数据，实用性较高，是当前姓名消歧的核心技术和主流方法；基于社会网络的姓名消歧方法关键在于构建完备的社会网络以及如何正确有效地利用社会网络，这是十分复杂和困难的，还需进一步研究；基于网络知识资源的姓名消歧方法存在的问题是网络上公开资源丰富而零散，且有的网络信息不一定准确，给姓名消歧工作带来了一定的困难，如何最大限度地搜集到网络上的可利用资源来解决姓名消歧的问题，是下一步发展的趋势；现有的姓名消歧方法只考虑了对已有文献进行相似度计算，聚类，实现将同名作者论文区分，然而在信息爆炸式增长的今天，对于随时产生的新数据或者新收录的论文，现有方法需要重新对包含该姓名的所有文献先计算相似度，再聚类，才能得到较为准确的新的区分结果，不能利用已有的分类信息，不仅效率较低，而且不能快速对其进行姓名消歧。近年来，指纹技术的发展使其成为成熟的生物鉴定方式，广泛应用于身份认证，然而是否能够通过对比信息携带的独特的语义指纹特征，来识别信息的作者，达到人名消歧的目的，是今后研究的方向之一。

参考文献

[1] 杨欣欣 . 基于两步聚类和查询扩展的人名消歧 [D]. 苏州 : 苏州大学 ,2015.

[2] 邓龙 . 基于语义的中文词义消歧技术研究 [D]. 哈尔滨 : 哈尔滨理工大学 ,2015.

[3]Ide N. Introduction to the special issue on word sense disambiguation: the state of the art [J]. Computational Linguistics， 1998,24(1):2-40.

[4] 袁军鹏 , 俞征鹿 , 苏成 , 等 . 作者重名辨识研究进展 [J]. 数字图书馆论坛 ,2011(10): 60-65.

[5]Mann G S, Yarowsky D. Unsupervised personal name disambiguation[C]// Proceedings of Computational Natural Language Learning,2003:33-40.

[6]BAGGA A, BALDWIN B. Entity-based cross-document conferencing using the vector space model [C]// Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics,1998:79-85.

[7]HAN H, ZHANG C L. Giles. Name disambiguation in author citations using a K-way spectral clustering method [C] // ACM IEEE/CS Joint Conference on Digital Library, 2005:334-343.

[8]BOLLEGALA D, MATSUO Y, ISHIZUKA M. Disambiguating personal names on the web using automatically extracted key phrases[C] // European Conference on Artificial Intelligence,2006:553 -557.

[9] 杨欣欣 . 基于查询扩展的人名消歧 [J]. 计算机应用 ,2012,32(9):2488—2490.

[10] 王英帅 . Web 人名消歧方法的研究与实现 [D]. 苏州 : 大学 ,2010.

[11]CHEN YING, MARTIN J. Towards robust unsupervised personal name disambiguation[C] // Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Leaning(EMNLP-CoNLL), 2007:190-198.

[12] 朱亮亮 . 利用改进的 K-means 算法实现文献著者人名消歧 [J]. 软件导刊 ,2013(05):63-66.

[13] 任景华 . 利用优化的 DBSCAN 算法进行文献著者人名消歧 [J]. 图书馆理论与实践 ,2014(12):61-65.

[14] 丁海波 , 肖桐 , 朱靖波 . 基于多阶段的中文人名消歧聚类技术的研究 [C]// 第六届全国信息检索学术会 (CCIR2010),2010:316-324.

[15]Wang X, Liu Y, Wang X, et al. Adaptive resonance theory based two-stage Chinese name disambiguation[J]. International Journal,2012(2):83 -88.

[16]Xu J, Lu Q, Liu Z Z. Combining classification with clustering for web person disambiguation[C]// Proceedings of the 21st International Conference Companion on World Wide Web,2012:637-638.

[17]Huang J, Ertekin S, Giles C L, et a1. Efficient name disambiguation for large--scale databases[C]// Knowledge Discovery in Databases(PKDD),2006:536-544.

[18]HAN H, GILES C L, ZHA H, et al. Two supervised learning approaches for name disambiguation in author citations[C]// ACM IEEE/CS Jt. Conf. on Digital Library, 2004:296-305.

[19]MALIN B. Unsupervised name disambiguation via social network similarity[C]// Workshop on Link Analysis, Counterterrorism, and Security in conjunction with the SIAM Intemational Conference on Data Mining, 2005:93-102.

[20] 郎君 , 秦兵 , 宋魏 , 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): 1365-1373.

[21]Tang J,Fong A C M,Wang B,et al. A unified probabilistic framework for name disambiguation in digital library [J]. IEEE Transactions on Knowledge and Data Engineering,2012,24(6):975-987.

[22]Han X,Zhao J. Web personal name disambiguation based on reference entity tables mined from the web[C]// Proceeding of the Eleventh International Workshop on Web Information and Data Management,2009:75-82.

[23]Vu Q M,Takasu A, Adachi J. Improving the performance of personal name disambiguation using web directories [J]. Information Processing and Management,2008,44(4):1546 -1561.

[24] 周晓 , 李超 , 胡明涵 . 基于人物互斥属性的中文人名消歧 [C]// 第六届全国信息检索学术会议 (CCIR2010), 2010:333-340.

[25]Bunescu R C,Pasca M. Using Encyclopedic Knowledge for Named Entity Disambiguation[C]// Proceedings of the 11st Conference of the European Chapter of the Association for Computational Linguistics,2006:9 -16.

[26]Han X, Zhao J. Named entity disambiguation by leveraging Wikipedia semantic knowledge[C] // Proceedings of the 18th ACM Conference on Information and Knowledge Management. New York: ACM,2009:215 -224.