

**本 科 生 毕 业 设 计**

**开题报告**



**学生姓名: \_夏梦丽\_\_\_\_\_ \_**

**学生学号: 310102592**

**指导教师: 吴飞**

**年级与专业: 2010级数字媒体技术**

**所在学院: 计算机学院**

**目录**

**本科毕业设计选题报告**…………………………………………………………1

1. 选题依据………………………………………………………………………1
2. 目前产业状况…………………………………………………………………1
3. 可行性分析……………………………………………………………………2
   1. 技术可行性分析…………………………………………………………2
   2. 美工要求…………………………………………………………………2
   3. 工作量估计………………………………………………………………3

**本科毕业设计外文翻译**…………………………………………………………4

**1. 选题依据**

**所选题目：**学术人物关系网络分析及可视化

**毕设内容：**研究学术论文中作者人物关系并实现对此类关系的可视化。具体而言，在充分调研现有论文数据、作者人物关系消歧算法、可视化方法的基础上。尝试利用现有算法与技术手段，基于DBLP数据，实现学术人物关系网络分析及可视化。

引用书目内容中作者人名的歧义是影响数字图书馆或者类似系统服务质量和内容的一个难题。数字书库中往往有数以百万计的引用书目记录（这里可以理解为一组书目的属性，如作者，合著者的人名，作品，和特定期刊发表会议的标题），并已成为学术界信息的一个重要来源，因为这使得搜索和发现中有关出版物以集中的方式来呈现。基于数字书库内容的研究也可以产生有趣的结果，如课题覆盖面，研究趋势，特定的委员会或个人发表物的质量和影响，在社交网络下的协作模式等。这些类型的分析和信息，可以用于如辅助相应机构拨发基金给个人或团队的决策，提高质量。

在课题本身之外，在完成的过程中可以较为深入的对机器学习及信息可视化的课题进行研究。机器学习和可视化研究都是在大数据时代对海量数据分析和操作不可缺少的研究。在海量数据的挖掘中我们可以发现很多平常不被发现的有趣现象。作为数字媒体技术专业的学生，信息可视化是一个与专业较为契合的研究方向。当然我相信可视化和信息的分析之间是相辅相成的，即需要算法对信息的分析整合，也需要结果更直观的表达，从而促进更好的分析。该课题作为自己的兴趣相同点，也会对自己思维方式产生影响。

**2. 目前产业状况**

现有的人名消歧算法主要包括作者分组法（Author Grouping Methods）及作者分配法（Author Assignment Methods）。其中，作者分组方法通过利用引用作者（或一组引用）属性的相似度函数，并用聚类算法来决定是否并入相应分组。相似度函数可以预先定义（基于现有的和根据属性的类型），使用有监督的机器学习技术，或从作者和合作者之间的关系中提取，通常用一些图来表示。这种相似性函数，连同一些聚类技术应该对同一组作者引用，试图分别最大化聚类簇之间的相似性，最小化聚类簇内部发相似性。而作者分配法直接通过构建某一作者的模型（例如，一个作者发表与其他（合作）作者的文章，在一个给定的场景和使用特定的标题的列表，使用一个监督分类方法或基于模型的聚类方法。其中，前者很大程度上依赖于相似性函数的选择及聚类的方法，而后者也很大程度上依赖于选择的特征概率模型的构建。这两种方式都需要对作者特征信息的提取，目前主要的研究方向包括也直接的引用条目信息，如作者/合著者姓名，作品标题，发表地点名称，年份等，web信息，如从互联网中检索出来的与被引用作者相关的发表物文件的其他信息，及隐含的特征如引证的潜主题。还面临着诸如引用条目错误，效率，作者研究领域变化，增量歧义，新研究者的加入等开放性问题的探究。

**3. 可行性分析**

**3.1技术可行性分析**

在数据的采集上，由于有较完整的现成数据，将主要基于DBLP数据从而将精力主要集中在算法及可视化效果的优化上。

在数据的分析上，主要通过解析xml文件节点，利用python语言进行筛选处理，并通过对已有文献引用人名消歧方面的发表论文中算法的总结与改进自动生成得到人物关系图（主要通过人名表示节点，边表示互相关系得到的关系网络）。在这些算法中，

选用python语言一方面是出于已有科学计算库完善的考虑，以及个人对语言的熟悉程度。Python在科学计算领域，有两个重要的扩展模块：Numpy和Scipy。可以很方便的进行高维度数据的分析和媲美matlab的数据变换操作。并且拥有scikit-learn机器学习相关的函数库。scikit-learn 是一个基于SciPy和Numpy的开源机器学习模块，包括分类、回归、聚类系列算法，有SVM、逻辑回归、朴素贝叶斯、Kmeans、DBSCAN等，可以将精力更多放在算法的优化上。

在数据可视化的呈现上也将选用D3.js的javascript库进行快速的可视化实现。D3.是一个开源的基于数据操作文档JavaScript库。为可视化操作提供良好适应性。如CSS3、HTML5和SVG网页标准，减少开销，D3速度非常快。支持大数据集和动态交互和动画效果。并且拥有很多成功的案例可以开拓思路，简化在可视化过程中对dom节点或svg的操作。选用d3可以将更多的精力放在可视化效果的设计上。

**3.2美工要求**

鉴于使用了D3.js的javascript库，美工上的要求主要集中在可视化效果的设计上。可视化是在针对某一研究问题的情况下，将数据通过尽可能直观的展现出来，使信息更好的被获取和接收。当然，在这个前提下能做到视觉上的美观也是十分重要的，主要体现在色彩，大小，动画，交互等方面。相对来说对美工本身的要求较小。

**3.3工作量估计**

总体来说，设计主要集中在算法的学习及可视化效果的设计上。文献翻译的过程中对消歧算法分类有了一定的了解。在算法的分类上，可以分为对引文根据作者姓名进行分组的方法，及对作者特征建模后直接分配引文到相应作者的方法。其中会涉及到的一些机器学习的方法和概率模型等算法的研究会是一个较大的工作量。另外，在可视化的研究上，会主要对图（graph）或者网络（network）的可视化效果进行较为深入的研究，在原有一些可视化算法的基础上进行一些设计。鉴于自身对python及javascript语言的掌握程度，以及对函数库的使用，在设计之后的实现上可以较快的完成。

**外文文献翻译：**A Brief Survey of Automatic Methods for Author Name Disambiguation

**一种自动消除作者人名歧义方法的简述**

**摘要：**

引用书目内容中作者人名的歧义是影响数字图书馆或者类似系统服务质量和内容的一个难题。处理作者人名模糊性的挑战，导致了许多消歧方法。一般来说，比较通常的方法是试图通过对同一作者引用记录的纪录，发现其中的一些相似性或直接将其分配到各自的作者。这两种方法不是用到监督就是非监督的聚类方法。在这篇文章中，我们将提出一种分类方法表征文献中描述的当前作者人名的消歧方法，目前最有代表性的一个简短的综述，并讨论几个开放的挑战。

**1. 引言**

当前世界上包括DBLP1，CiteSeer2 , PubMed3和BDBComp4在内的几个数字书库都提供了相应的功能和服务，以促进文学的研究和发现，以及其他类型的功能。这样的系统可以列出数以百万计的引用书目记录（这里可以理解为一组书目的属性，如作者，合著者的人名，作品，和特定期刊发表会议的标题），并已成为学术界信息的一个重要来源，因为这使得搜索和发现中有关出版物以集中的方式来呈现。基于数字书库内容的研究也可以产生有趣的结果，如课题覆盖面，研究趋势，特定的委员会或个人发表物的质量和影响，在社交网络下的协作模式等。这些类型的分析和信息，可以用于如辅助相应机构拨发基金给个人或团队的决策，提高质量。

根据D.Lee,等人的研究，高品质的内容所面临的挑战来自于数据输入的错误，不同的引文格式，缺乏（执法）标准，不完善的引用收集软件，暧昧的作者人名和出版地点头衔的缩写。在这些挑战中，由于其固有的困难，作者人名模糊性需要大量的对数字书库研究界的关注。具体来说，人名的歧义是当一组引用记录包含模棱两可的作者人名，也就是说，同一个作者可能会出现在不同的名称（同义词） 下，或不同的作者可能具有相似的名称（多义性）情况下出现的问题。这个问题可能是多种原因引起的，包括缺乏标准和惯例，以及分散的内容引起的（即通过自动收录方法） 。

为了说明这个问题，表1示出了一组三个引文的集合{C1，C2，C3}，以便每个引用的作者人名由rj来区分，其中1≤j≤16。对于每个引用，每个名字rj是一个参考的作者和拥有与之相关的属性，如合著者人名的列表（即引用同一引用其他作者的列表），作品标题，期刊发表会议的标题，发表年份等。查看表1，我们可以发现同义词和多义词的例子，如前所述，是人名歧义问题的子问题。作者人名R3和R15是人名多义的一个例子，其中R3是指IBM印度研究院的“Ajay Gupta”，而R15是指美国NEC实验室的“Aarti Gupta”。作者人名R3和R7是人名同义的例子。两者都是指IBM印度研究院的“Ajay Gupta”。

更严格的，人名消歧的问题可以表示如下：

设集合C={C1，C2，...，CK}是一组引用记录，每一个引用记录ci具有相应的属性列表，其中包括至少作者人名，作品名称和发表期刊标题。一个引用的每个属性都与特定的值相关，可以具有多个组成部分。对于作者人名这个属性，一个分量对应一个唯一的作者人名，并用Rj表示一个真正的作者。对于其他属性，一个分量对应于一个单词/术语。消歧的目的是生产用于分区一个作者集合{r1, . . . , rm}为n个子集合{a1 , . . . , an }，使每个子集包含ai（尽可能全部）对应到同一个作者。

为了区分一个数字书库的书目资料，首先我们可以将作者集合中有歧义的作者人名进行分组。这些被称为歧义组（即，有作者人名的值具有相似名称的属性引用的分组）。歧义组可以通过使用阻断方法 [ 37 ]来获得，这避免了在所有的作者人名变量中，需要进行比较的地址可扩展性问题。

处理作者人名模糊性的挑战，导致了许多消歧方法的研究[3, 4, 7, 9, 15, 16, 20, 21, 22, 24, 26, 27, 33, 35, 38, 40, 41, 42, 43, 44, 46, 49]。其中一个这样的挑战是，通常情况下，只有一个最小属性集（在大多数情况下，只有作者人名，发表物和地点职称）可以处理。在任何情况下，现有的消歧方法通常只是使用某种比较它们之间的相似性的方法对同意作者的作品进行分组，或者直接分配引文记录到各自的作者。

早期的一些初步的消歧方法的研究可以在[ 28 ]中找到。在那篇文章中，Klass 将已有的方法划分为监督或非人监督的，并介绍了直到2006年发表的一些方法。然而，这一研究领域在去年变的非常多产，随着最近许多方法被提出。在这篇文章中，，我们将提出一种分类方法表征文献中描述的当前作者人名的消歧方法并将展示一些最有代表性的研究综述。

本文的结构如下。第2节介绍我们的分类表征作者人名消歧方法。第3节介绍具有代表性的作者人名消歧方法的概述。方法特性的概要总结在第四节。第5节讨论的作者人名消歧任务的一些开放的挑战。最后，第6节给出我们的结论。

**2.作者人名消歧方法的一种分类**

本节以层次分类的方式对文献中发现的最有代表性的自动作者姓名消歧方法进行分组。建议的分类参考图2。该方法可根据不同开发方法的主要类型可分为：作者分组[ 4 ， 7 ， 9 ， 15 ， 16 ， 22 ， 24 ， 26 ， 27 ， 36 ， 38 ， 41 ， 42 ， 44 ， 45 ， 46 ， 35 ， 49 ] ，它试图根据某种类型的相似性之间的计算将同一作者使用的引用进行分组，或作者分配[ 3 ， 16 ， 20 ， 21 ， 43] ，其目的是直接分配到引用各自的作者。另外，该方法可根据在消歧任务探索的证据进行分组为：引文属性（单一），Web信息，或可从现有资料中提取隐含的数据。

值得注意的是，在篇综述中，我们只包括了自动消歧的方法。其他类型的方法，如图书馆管理员交互式的分配[39]或协作努力[6]的方法，很大程度上依赖于人的努力，这在大量姓名消歧任务中是很难使用的。出于这个原因，他们没有在本文中提及。也有试图为每个作者建立一个唯一的标识，如使用开放研究员投稿识别[7]的方法（ORCID），但这些都不将被讨论。

由于人名歧义的问题并不局限于文本中，其他几个名字消歧方法，其中利用不同的特征或针对其他应用程序（如在网页搜索结果的人名消歧），在文献[2，12，18，48，50]中描述的也是值得注意的。然而，这些方法的讨论超出了本文的范围之内。

最后，我们想强调的是类别的区分在我们的分类方法中不完全脱节。例如，有使用两个或更多个类型的特征或混合方法的方法。在接下来的小节中，我们将详细我们提出的分类方法。

**2.1方法的类型**

正如前面所说，一种组织已经存在的作者姓名消歧方法的方式是根据他们开发方法的不同。我们将在下面进一步阐述这种区别。

**2.1.1作者分组法**

作者分组方法通过利用引用作者（或一组引用）属性的相似度函数，并用聚类算法来决定是否并入相应分组。相似度函数可以预先定义（基于现有的和根据属性的类型） [ 4 ， 7 ， 22 ， 36 ， 41 ] ，使用有监督的机器学习技术[ 9 ， 24 ， 44 ， 45 ， 46 ]，或从作者和合作者之间的关系中提取，通常用一些图来表示[ 15 ， 33 ， 35 ] 。这种相似性函数，连同一些聚类技术应该对同一组作者引用，试图分别最大化聚类簇之间的相似性，最小化聚类簇内部发相似性。

**定义相似度函数**

在这里，一个相似的功能是负责确定两个引用（或参照组）作者的相似度。我们的目标是获得一个函数，当引用同一作者返回相似度极高值和引用不同作者返回较低的相似值。此外，期望的相似性函数是传递的。更具体地，设C1 ，C2和C3是3个引文，如果c1和c2是非常相似的（根据函数）并且c2和c3也非常相似，那么C1和C3通过我们的函数的计算也应具有较高的相似性。接下来，我们来讨论确定这种相似性函数的方法。

使用预定义的相似度函数。

这类的方法都有一个特定的预定义相似性函数S嵌入到他们的算法来检查是否两个引用的参考文献或分组指的是同一作者。这样的函数S的例子包括[6]：Levenshtein距离，Jaccard系数，余弦​​相似性，soft-TFIDF 和一些其他方法[6]，应用的参考属性的元素。这样的功能的组合也被使用于（如，在[4，41]）。

这些方法不需要任何在训练数据上的监督，但其相似性函数通常需要进行调整来消除歧义的引文记录的特定集合。对于不同的集合，可能需要一个新的调整过程。最后，并非所有这样的函数都可以很自然的过渡。

学习型的相似度函数。

学习一个特定的相似度函数通常会产生更好的结果，因为这些教训函数直接优化手头消歧问题。学习的相似度函数消除歧义的方法是接收一个引用对集合{sij}和这两个相应的引用是否指的是同一作者的特殊变量为输入。这对引用，Ri和Rj∈R（引用的集合）是通常由相似性向量s⃗表示。每个相似性向量s⃗是由一组q个特征的集合F{f1,f2,...,fq}组成。这些向量的每个特征fp都代表了在属性ri，Al 和 rj上的对比。Al 是两个引用ri和rj。

每个特征的值都由其他函数定义，如Levenshtein距离，Jaccard系数，Jaro-Winkler，余弦​​相似性，soft-TFIDF，欧氏距离等，或某些特定的启发式，如术语或共同作者的数量，普通或特殊值如名字随着姓氏出现的初值等。

训练数据被用来产生一个相似性函数S，从R X R的矩阵映射到{0，1​​}的值，其中1表示该两个引用都指的是同一个作者，0表示不是同一个作者。正如前面提到的，依靠机器学习技术来定义相似性函数的方法在引文不同的集合下是相当有效，但他们通常需要许多例子和足够的特征才能很好的起效，这可能是很难得到的。

利用基于图的相似性函数。利用基于图形的相似性函数的方法对作者姓名消歧通常为模糊的组创建一个合作者关系图G=（V，E）。每个作者姓名及合著者姓名元素的属性是由一个顶点v表示。同一个合作者通常用一个唯一的结点表示。对于每一个合作关系（即一对作者发表的一篇文章）创建一条边⟨Vi，Vj⟩。边的权重和Vi，Vj结点对应作者姓名合作的文章数有关。

基于图的度量（例如，最短路径[33]），可能会与其他相似性函数组合用在引用作者的属性或用作相似度向量新特征。

**聚类方法**

作者分组方法经常在其消歧任务中用到聚类算法。最常用的技术包括： 1 ）分类[23] ，它创建了一个在迭代过程中提到的作者的预定数k ，2）分层合并聚类[23] ，通过层次分级的方式将引用作者分组，3）基于密度的聚类[23] ，其中一个簇对应的引用的作者（根据一些密度条件）所包围的低密度的区域的致密区域 - 具有低密度的区域引用视为噪声，以及4）谱聚类[51] ，其对应于基于图形的技术，计算拉普拉斯算子矩阵的特征值和特征向量，频谱信息，即，在作者姓名消歧任务，表示为一个相似性矩阵加权图。在一般情况下，这些聚类技术依赖于一个好的“相似度函数”才能对引用进行分组。

**2.1.2作者分配法**

作者分配法直接通过构建某一作者的模型（例如，一个作者发表与其他（合作）作者的文章，在一个给定的场景和使用特定的标题的列表，使用一个监督分类方法[ 16 ， 20 ]或基于模型的聚类方法[ 3 ， 21 ]。

分类。在这种将引用直接分配给作者类的方法使用监督机器学习的算法。更具体地讲，它们把一组作者引用作为属性输入，成为训练数据D ，由实例，或更具体地，正确的作者是已知的。每个实例是由m个特征{f1, f2, . . . , fm}的集合F，和一个作者名的特殊变量组成。这个作者变量从一组离散的标签{a1,a2,..., an}中获取，其中每个标签都有唯一标识作者。训练实例通过与关联特征值和正确的作者变量来生成一个消歧函数（如消歧子）。测试集（记为T）的消歧任务由一组已知特征征而正确作者未知的集合组成。消歧子就是一个将特征集合{f1,f2,...,fm} 映射到作者集合{a1,a2,..., an}，从而来预测测试集合中引用和正确作者的一个函数。这种方法下，消歧子将纪录集合T分成n个集合{a1,a2,...,an}，其中ai在理想情况下仅全部的包含第i个作者的所有引用。

当面临每个作者大量的引文的案例中，这些方法通常是非常有效率的。另一个优点是，如果集合已经被消除歧义（手动或自动），该方法可能只需通过运行已经经过学习的模型并将新的引用插入到相应集合中就可以了。虽然这些方法的应用已经有很多成功案例，训练集合的采集，通常需要熟练的人力来手动标记引用。数字书库是非常动态的系统，因而大量的手动贴标是不可行的。此外，消歧任务呈现细微差别，并需要与特定能力的方法。举例来说，因为把所有可能的作者都包含在训练数据的假设是不合理的，而且随着时间的推移，作者们总是会改变自己研究的兴趣领域，新的实例需要被不断地插入到训练数据，方法就要被重新训练来保证其有效性。

聚类。聚类的算法通过优化一组引用的作者和一些用来表示作者的数学模型之间的配合试图直接把引用分配到相应的作者，他们用概率方法来确定一个迭代的方式来拟合模型（或估计概率参数的方法）与作者之间的关系。例如，在这种方法的第一次迭代过程中，每个引用会随机地分配到一个作者ai和一个函数使得从特征集合{f1,f2,...,fm} 到作者集合{a1,a2,..., an}的映射。在第二次迭代时，这个函数就用来预测每个参考文献的作者和一个新的函数被用于下一次迭代。这个过程继续，直到一个停止条件被满足，例如，一些次迭代后。两种常用的适合消歧问题中模型的算法是，期望最大化（ EM ） [11]和Gibbs采样[ 19 ] 。

这些方法不需要训练数据，但它们通常需要优先的正确作者个数或正确作者分组的个数的信息，可能需要一些时间来估计其参数（例如，由于多次迭代）。此外，这些方法可能要到最后一次迭代才能直接把分配到他们的引用当中。

**2.2 特征的探测**

在本节中，我们将对消除歧义的方法最常用的特征种类进行探讨。

引文信息。引用信息从引文中直接提取出来的属性，如作者/合著者姓名，作品标题，发表地点名称，年份等。这些属性是那些在所有引文中常见，但通常都不足以完全消除作者所有引用的歧义。有些方法还会利用一些附加信息，如电子邮件，地址，纸头，但这并不总是可用的，或容易获得的信息，但如果存在，他们通常会对消歧的过程中有帮助（很多！）。

Web信息。 Web信息表示从互联网中检索出来的与被引用作者相关的发表物文件的其他信息。这种信息通常是根据通过在搜索引擎中，根据引文属性的值提交查询条件来获取，利用返回的网页页面作为一个新的特征来计算引用之间的相似度。新的特征通常帮助了消歧任务。但其中的一个问题是从Web文档中提取所需的全部信息的额外成本。

隐含的特征。隐含的证据是从属性的可见元素推断。几种找到隐含的证据方法已经被应用，如引证的潜主题。一个例子是隐含狄利克雷分布（LDA）[5]，它可以估计主题分布引文（即，LDA估计给定的引​​文每个主题的概率）。估价分布化作为新的证据（属性）来计算引用作者之间的相似性。

**2.3 评估指标**

虽然不是我们的分类法的一部分，但是要了解下面的讨论中很重要的一点是，在实验评估使用的每一个提出的方法的评价指标。最常用的指标是：精度，这基本上是所有的预测中正确结果的比例;传统的准确率的评估和F1，通常用于在信息检索和分类问题8; 常和F1一起出现的，F1的另一种形式，是认为对引文组正确地分配给同一作者（或不）; 聚类F1，计算正确的簇的精度（即包含所有与只有所有引用作者分组）;K指标[7]，纯度（一个包含的只有一个作者的纯粹的分组）和分裂（发生一个作者被分成一个或多个分组的情况，分裂）的组合；B -Cubed模型[1]，将计算出的精度和召回每个引用的作者;和MUC [1]。在这最后的指标中，召回由正确元素簇包含这些元素和数目减去实验获得簇的数目（用相关算法获得的）之和除以元素的总量减去这个数理论集群来获得。精度计算也是相似的。

**4.特征总结**

在本节中，我们给出了在所描述的作者姓名消歧的方法中找到的特征，总结在表2和表3的概述中。其中用于评估方法的集合，我们有：（1）CiteSeer，DBLP，BDBComp，ArnetMiner和REXA含计算机科学期刊版本；（2）arXiv包含从高能物理出版物引用，（3）BIOBASE，含从生物出版物引用，（4）MEDLINE和生物医学与生物医学刊物的数据;（5）ISI-Thomson和几个知识领域的出版物;（6）科拉，这是计算机科学的重复引文和人名从提取的构成网络，以及（7）IMDB约演员的数据。

大部分所描述的方法[4，7，9，15，22，24，26，27，33，35，38，40，41，42，45，46，49]尝试使用引用作者相似度函数来表示两个引用是否来自同一个作者来消除歧义，而不是直接分配相应的作者，如[3，16，20，21，43，47]。

有些这类方法需要一个正确的作者数量作为输入([15, 21, 22])，或者这个数量是由训练数据获得的[20]。 其他的一些方法，比如那些在[3], [43] 和 [16]中提及的则是通过估计获得这个数量。

将近一半的方法[3, 4, 7, 15, 16, 20, 21, 22, 33, 35, 40]用最多三个主要的引用属性：作者姓名，论文标题和发表刊物标题作为特征值。这三个属性是引用纪录中最常见的，在大多数情况下都能进行消歧。很少的方法[26, 27, 38, 49]利用了一些附加的特征，比如电子邮件，地质，论文头部等不能总是很容易获得的信息。

表2和3同样总结了每种方法的评估标准和子问题的类型（如同义，多义或都存在的情况）。

**5. 开放性挑战**

有很多的开放性挑战都需要被提出，为了产生更多可以应用于真实数字书库下生产模式中的解决方案。下面我们来讨论其中的一些。

引用中数据非常少。在大多数情况下，我们只能获得关于引用很基本的信息：作者（合作者）的名字，作品，发表期刊的标题，和发表的年份。并且，有些情况下，作者名字只包括了姓，最有一个名，发表期刊的标题还是缩写。新的在一定场景下有前景策略会试图从这些含蓄的信息（如研究课题）或者结合从web获取的附加信息来推导。

非常模糊的情况。有些方法利用启发式的基于合作者的算法, 通过明确的假设：（一）很少有模棱两可的引用而合作者名字也补明确的情况（ii）两位作者用非常相似的名称在同一研究领域的工作是罕见的。这些假设在大多数情况下都成立，但是当他们失败了，它们所产生的错误是很难修复。例如，在亚洲名作者的情况下，第一个假设出现错误的频率比为英文或拉丁文名作者更高。

引文的错误。引用信息的错误有时是无法检测到的。方法需要能够容忍这样的错误。

效率。随着大量的文章发表在不同知识领域的今天，该解决方案必须能够有效地处理这个问题。少数提出的方法有这样明确的关注。

不同的知识领域。正如我们所看到的，最常用的用于评估方法的集合都是计算机科学相关领域的。然而，其他知识领域（如人文，医学）可能有不同的出版模式（例如，具有独立作者或是很多合作者的出版物）对目前这些方法造成额外的困难。

增量歧义。理想的消歧方法应在数字书库中随着新引文的加入能够持续的起作用，因为在整个数字书库在每个新的输入都能被消除歧义的假设通常被认为是不合理的。然而，大多数，不是全部，方法忽略了这个事实。一个有前景的解决方案被提出[10]。

作者简介变化。随着时间的推移作者的研究兴趣会变化是很常见的。因为新的合作，在变化研究小组或机构，研究领域的自然演化等都导致这可能发生，这些变化导致模型中代表的修改造成方法的困难。一种可能的解决方案可能涉及再训练，但确定何时再训练也是一个难题。然而，这个问题已经在很大程度上被所有方法忽略了。

 新的作者。该方法应该能够识别在DL中还没有存在引用的模糊作者的名字。唯一发表的方法之一[ 47]明确解决这一问题。

**6 结论**

本文介绍了作者姓名消歧方法的综述。我们提出了一个分类的方法对已有的方法进行分类，并提供了最具代表性的一个概述。有些模式就变得清晰。大部分被调查的方法，通过使用某种类型的相似性函数的引用比较记录执行消歧。这个函数，它可以被预定义或通过专门的学习用于消歧任务，被直接应用到引用属性，它可以与从Web检索或从自己引用推断附加信息的属性（例如，主题）得到增强。其他一些方法消除歧义通过监督和无监督机器学习技术，或者直接指派引用记录它们的作者。还讨论了一些开放式的问题。

在该领域的一个主要差距是缺乏的是在相同条件下方法的直接比较：例如，同一集合（例如，许多方法中使用不同版本的集合，如DBLP的） ，相同的计算环境，相同的实验设计。这可能是由于缺乏像那些由TREC比赛提供的标准的集合。此外，存在的几个涉及最多三个或四个方法，并在静态情况下被执行。事实上，没有研究如何将这些方法在一个动态并且实时的数字图书馆的比较。连同先前讨论的开放性问题，这些问题都是在我们看来，是在不久的将来应该引导开发新的作者姓名消歧方法的研究工作。