doi: 10.3969/j.issn.0490-6756.2019.02.010

基于机器学习的论文作者名消歧方法研究

邓可君, 华 凯, 邓昌明, 姜 宁, 袁 玲, 彭一明, 张治坤
(北京大学计算中心, 北京 100871)

摘 要: 本文提出了一种基于规则匹配和机器学习的论文作者名自动化消歧方法: 首先基于人工构建的人名匹配规则确定候选作者, 对于存在多个候选人的情况, 基于论文的属性信息(例如合作者、标题、摘要、关键词和出版物名称等)提取特征, 然后选取合适的机器学习算法进行消歧. 实验效果表明 K 近邻和 Softmax 分类器较适合于论文作者名消歧任务; 此外, 将作者信息与论文的其他信息分开提取特征能够有效提高作者名消歧的准确性.

关键词: 作者名消歧; 机器学习; 文本特征提取
中图分类号: TP391.1 **文献标识码**: A **文章编号**: 0490-6756(2019)02-0241-05

Research on author name disambiguation method based on machine learning

DENG Ke-Jun, *HUA Kai*, *DENG Chang-Ming*, *JIANG Ning*,
YUAN Ling, *PENG Yi-Ming*, *ZHANG Zhi-Kun*
(Computer Center, Peking University, Beijing 100871, China)

Abstract: This paper proposes an automatic article author name disambiguation method based on rule matching and machine learning. For each article, the candidate authors are determined based on artificial constructed name matching rules firstly. For the cases of multiple candidates, features are extracted from the attribute information of the article, such as collaborators, title, abstract, key words and publication name, and then selected machine learning models are applied to author name disambiguating. The experimental results show that the K-nearest neighbor and Softmax classifier are more suitable for the author name disambiguation task than other models. In addition, extracting features of the authors information separatelycan from other information effectively improve the accuracy of the author namedisambiguation.

Keywords: Author name disambiguation; Machine learning; Text feature extraction

1 引 言

高校和科研机构都需要统计其单位作者的论文信息, 并对该单位的论文进行归档整理, 从而建立本单位的文献数据库. 但目前对于本单位职工的论文整理并不完善, 普遍只记录了论文的标题和署名作者, 没有按作者个体归档. 这样的情况下, 较难评估该单位科研工作者的科研成果和水平, 也难以向外界提供针对特定学者论文的搜索支持.

在论文的自动化归档工作中, 作者名消歧是一个棘手的问题. 一方面, 论文作者名在同一机构中可能会存在重名现象; 另一方面, 国人作者在英文论文中的署名可能存在多种形式. 目前, 自动化识别论文归属作者的方法仍处于探索阶段[1], 很多机构都是采用人工方法进行识别. 然而人工方法费时费力, 且不能保证准确率.

收稿日期: 2018-06-28
作者简介: 邓可君(1986-), 女, 湖南长沙人, 博士生, 工程师, 研究方向为信息处理. E-mail: kejund@pku.edu.cn
通讯作者: 张治坤. E-mail: zhangzhikun@pku.edu.cn

本文提出了一种基于规则匹配和机器学习的论文作者名自动化消歧方法:首先基于人工构建的人名匹配规则确定候选作者,对于存在多个候选人的情况,基于论文的属性信息(例如合作者、标题、摘要、关键词和出版物名称等)提取特征,然后选取合适的机器学习算法进行消歧. 本文选择了北京大学 2004 到 2015 年 7790 条 SCI 论文数据进行了实验,验证了该方法的有效性.

2 相关工作

相比于传统的人名消歧,论文作者名消歧有其特殊性. 一方面,带作者标注的论文数据集较难获取;另一方面,论文信息一般包括作者、标题、摘要、关键词和出版物名称等内容,所包含的信息量较为有限.

论文作者名的自动化消歧可以归为机器学习中的聚类或分类问题,根据所用样本的标注情况可以分为基于监督学习的消歧方法、基于非监督学习的消歧方法和基于半监督学习的消歧方法[2].

(1) 基于监督学习的消歧方法需要利用标注好的训练数据集来学习分类模型,例如朴素贝叶斯(Naïve Bayes)、支持向量机(Support Vector Machine,SVM)和逻辑回归等模型. 学者 Treeratpituk,Han 采用了这些模型进行论文作者名消歧,取得了较好的消歧效果[3,4],但这类方法需要标注好的大量样本,这在论文作者名消歧领域往往是稀缺的.

(2) 基于非监督学习的消歧方法不需要标注,仅凭样本数据的特性对样本聚类,可采用 K 均值算法(K-means)、基于密度的聚类算法(DBSCAN)和凝聚层次聚类等方法将同属于一个作者的论文聚为一类,但这类方法的准确率往往较低. 国内学者如赵铁军提出了多阶段的聚类策略,一定程度上提高了聚类的准确率[5].

(3) 基于半监督学习的消歧方法结合了上述两种方法,国外学者 Levin 提出了一种将聚类和分类结合起来的消歧方法,初始阶段基于规则聚类,得到部分标记样本后训练分类器,最后通过相似度度量再聚类[6],但该方法在初始阶段还需要手工制定规则,无法应用于大规模的数据集.

本文考虑了高校论文的数据特点,首先利用人工构建的匹配规则对给定的论文作者名进行匹配,根据匹配得到的候选作者的论文数据集分开提取特征并训练分类器,预测给定论文的所属作者,从而改进论文消歧效果.

3 基于规则匹配和机器学习的论文作者名消歧方法

3.1 整体框架

如引言所述,论文作者名存在混淆的原因一方面是作者存在重名现象,另一方面是在英文论文中,单个中文名可能存在多个对应的英文名. 高校的人员数量有限,中文重名现象较少,作者名混淆的情况大部分来源于后者. 由于可能采用了不同的姓名顺序和缩写规则,一个作者的中文名可能会对应多种形式的英文名,再加上多音字的现象,会出现大量作者名混淆的情况.

针对这一现象,本文根据“中国人名汉语拼音字母拼写规则”和常见的中文名到英文名的转换形式,制定了一个中文名到英文名的转换规则,并且基于该规则对高校职工的所有中文名进行预处理,即基于转换规则产生所有可能的英文名,保存在人员别名表中. 转换规则和实例如表 1 所示,在后期匹配过程中会统一去掉大小写和特殊符号(例如逗号、分号等).

表 1 中文人名到英文名的转换规则

Tab. 1 Name conversion rules from chinese to english

输入人名	采用规则		对应结果
季羡林	姓名正序	全拼	Ji Xian Lin
		名合并	Ji Xianlin
		缩写	Ji XL
	姓名反序	全拼,名合并	Xianlin Ji
		名缩写	XL Ji

完成中文人名的预处理后,就可以对待处理的论文进行作者名消歧. 基于规则匹配和机器学习的论文作者名消歧方法的整体框架如图 1 所示.

3.2 论文信息预处理

本文使用的论文信息包括标题、作者、出版物名称、摘要和关键词. 由于文本信息中存在噪音数据,而且没有进行分词,所以首先需要进行预处理. 预处理过程依次对论文信息进行去噪处理,包括:去掉特殊字符串、去掉标点符号及特殊符号、去掉多余空格和换行符、去掉长度小于 3 的词、去掉停用词和字符小写化等;然后采用自然语言处理工具(NLTK)对文本进行分词、词性标记和词性还原.

3.3 基于规则匹配确定候选作者

在人名预处理阶段,已经构建了转换规则,并且基于规则对高校所有中文人名构建了对应的别名表. 在规则匹配阶段,将论文中的作者名与人名

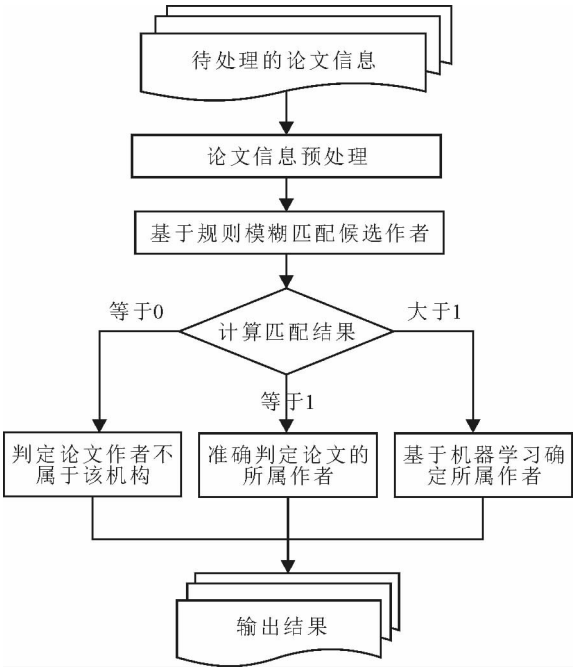

图 1　作者名消歧方法的框架图

Fig. 1　The architecture of authors disambiguation method

系统中的英文别名进行匹配，初步得到候选作者集合.

3.4　基于机器学习确定所属作者

经过规则匹配后，对于存在多个候选作者的论文需采用机器学习中的分类模型进行分类，从而进一步确定所属作者. 本文采用基于监督学习的机器学习方法，将论文作者消歧划归为文本分类问题. 整个消歧过程包含三个部分：特征提取、训练分类器和预测所属作者.

(1) 特征提取：本文采用向量空间模型和 TF-IDF (TermFrequency-InverseDocumentFrequency)从论文信息中提取特征[7,12]. TF-IDF 是信息检索领域较为常用且十分有效的特征提取方法[8,10,11]. 该方法用以评估一个字或词对于所在文档的重要程度，该字或词的重要性与它在该文档中出现的频数成正相关，但与它在文档集中出现的频数负相关[8]. 换言之，如果一个字或词在一篇论文的信息中出现的次数越多，且在所有论文信息中出现的次数越少，则其作为该论文的特征的区分能力越强. 本文采用了主流的 TF-IDF 计算公式，如下所示.

$$tfidf_{it}=tf_{it}*idf_t=tf_{it}*(1+\log(\frac{N}{df_t+1})) \tag{1}$$

本文采用 L2-Norm 对 TF-IDF 计算得到的向量进行归一化，归一化可以进一步提升文档查询和本文分类的准确度[8,13]. 此外，在论文信息中包含了论文的合作者信息，科研工作者在一段时间内往往有固定的合作者，因此合作者关系在论文作者名消歧问题中相较于其他信息更为重要. 本文将论文作者信息与其他文本信息(论文标题、期刊名称、摘要和关键词)分开进行特征提取. 分开提取特征具有两个优势：1) 作者信息的 TF-IDF 值会更大一些，从而放大作者信息对于作者名消歧的作用；2) 其他文本信息的向量空间不同于作者信息的向量空间，使得提取出的特征的可解释性更好而且预测能力更强. 最终，分开提取出的特征会再拼接起来形成样本的特征.

(2) 训练分类器：本文采用基于监督学习的分类方法来实现论文消歧. 高校论文信息数量有限，并不适于采用深度神经网络等模型. 因此，本文采用了传统机器学习中的主流分类模型：决策树、随机森林、Softmax、支持向量机、朴素贝叶斯、K 近邻算法和 XGBoost. 本文在北京大学的论文数据集上做了实验，实验表明 K 近邻算法和 Softmax 消歧效果较好.

(3) 预测所属作者：利用带标注的训练数据集训练好分类器后，即可对需要预测的论文数据进行预测，返回所属作者.

4　论文作者名消歧实验评估

4.1　数据集

我们使用了北京大学的论文数据，对本文提出的论文作者名消歧方法进行了实验与评估. 可靠的数据是算法评估和优化的前提，由于本文采用了基于监督学习的作者名消歧算法，因此我们首先要构建带标注的论文数据集. 我们利用 2004～2015 年北京大学职工的 SCI 论文奖励数据来对获取的论文数据进行自动化标注，SCI 论文奖励数据记录了 SCI 论文的标题以及受奖励的第一作者信息. 通过脚本比对论文标题，共获取了 7790 条带标注的论文数据集，一共涉及北京大学的 1457 名职工. 带标注的数据集都为英文论文数据，其中每条记录都包含了完整的论文信息以及所属作者的姓名和职工号，并将唯一的职工号作为样本的标签(label).

为评估消歧效果，我们构建了多个子消歧数据集：遍历每一条样本，经过规则匹配，若候选作者的个数大于 1，则将候选作者的论文集放入新的子消歧数据集. 这样，每一个子消歧数据集中的任意两

篇论文都存在作者名混淆现象. 我们统计了所有子消歧数据集的样本数量情况,整体分布如图 2 所示,可以看到 44.5%的子消歧数据集的样本数在 11～20 之间. 我们选取了样本数量大于 5 的子消歧数据集共 993 个作为实验数据. 在每个子消歧数据集上进行了随机划分,抽取 80%的标记样本作为训练数据集,剩下的 20%作为测试数据集,训练并评估分类模型.

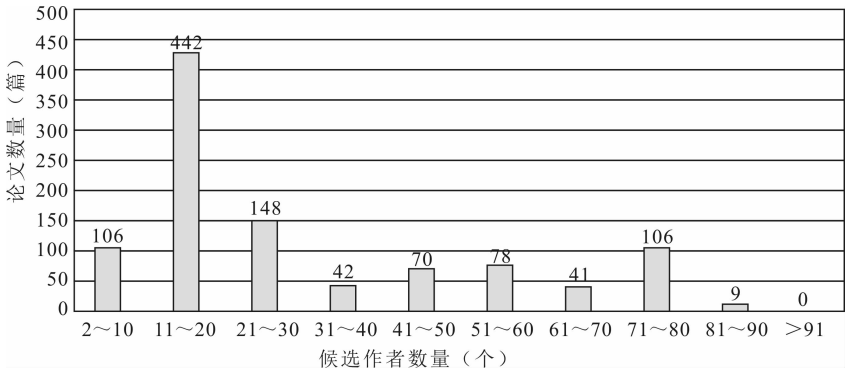

图 2 子消歧数据集的论文数量分布

Fig. 2 Number distribution of articles in disambiguation sub-datasets

4.2 评价指标

对于分类问题常用的评价指标是精确率(precision)、召回率(recall) 和 F_1 值,但这些指标仅适用于二分类的问题[9]. 需将这些指标进一步拓展,以适用于多分类的情况. 二分类问题的精确度和召回率的计算公式如下.

$$precision_i = \frac{TP_i}{TP_i + FP_i} \tag{2}$$

$$recall_i = \frac{TP_i}{TP_i + FN_i} \tag{3}$$

通过 macro 方法扩展二分类的 precision 和 recall,仅仅是做了算术平均,没有考虑到样本类别不平衡的问题,无法有效地评价模型的性能. 本文所用的消歧数据集可能会出现一个类别的样本数比其他类别的样本数多一个量级的情况. 因此,本文采用了加权平均法,将每一类别的样本数占总样本数的比例作为其权重,加权的分类指标可计算如下.

$$precision_{\text{weighted}} = \sum_{i=1}^{m} \frac{N_i}{N} * \text{precision}_i \tag{4}$$

$$\text{recall}_{weighted} = \sum_{i=1}^{m} \frac{N_i}{N} * recall_i \tag{5}$$

其中,N_i 表示类别 i 的样本数;N 表示总样本数;m 表示总类别数.

由以上两式可得 F_1 值计算公式

$$F_{1,\text{weighted}} = 2\frac{recall_{\text{weighted}} \times precision_{\text{weighted}}}{recall_{\text{weighted}} + precision_{\text{weighted}}} \tag{6}$$

4.3 不同分类模型的预测结果与分析

如 3.4 节所述,在上述子消歧数据集上,我们尝试了传统机器学习中主流的分类模型,分别是决策树、随机森林、Softmax、支持向量机、朴素贝叶斯、K 近邻算法和 XGBoost. 这些模型在多个子消歧数据集上的平均分类结果如表 2 所示.

表 2 不同分类模型的预测结果

Tab. 2 Prediction results of different classification models

	朴素贝叶斯	K 邻近算法	Softmax	随机森林	决策树	支持向量机	XGBoost
precision	95.08%	91.70%	92.13%	83.34%	89.40%	90.07%	85.13%
recall	89.68%	93.23%	92.91%	85.28%	89.86%	91.46%	87.42%
$f1$-score	0.9117	0.9246	0.9252	0.8274	0.8872	0.9005	0.8513

从表 2 中可以看到,朴素贝叶斯模型的精确度最高,K 近邻算法的召回率最好,而 Softmax 的 F_1 值表现最为突出. 在小样本分类问题中,往往是越简单的模型可以获得越好的分类效果,如表 2 所示,K 近邻算法和 Softmax 的各方面的分类指标都较好,显著优于其他模型. 在树型算法中,从模型的复杂度来讲,决策树小于随机森林,而随机森林又小于 XGBoost;在实验结果中,决策树的分类效果优于 XGBoost,而后者又优于随机森林. 总的来说,K 近邻算法和 Softmax 模型更适用于高校论文作者名消歧问题.

4.4 不同特征提取策略的预测结果与分析

在 3.3 节中提到,我们认为相比于其他论文信息,合作者关系所包含的信息量更大,因此将论文的作者信息与其他论文信息分开提取特征. 在实验中,我们对分开提取特征和混合提取特征这两种特征提取方式分别进行了实验,对比了朴素贝叶斯、

K 近邻算法和 Softmax 这三种模型采用不同特征提取策略的分类效果，如表 3 所示.

表 3　不同特征提取策略的预测结果

Tab. 3　Prediction results of different feature extraction strategies

	混合提取特征			分开提取特征		
	朴素贝叶斯	K 近邻算法	Softmax	朴素贝叶斯	K 近邻算法	Softmax
precision	94.30%	90.70%	87.46%	95.08%	91.70%	92.13%
recall	88.48%	92.86%	90.35%	89.68%	93.23%	92.91%
*f*1-score	0.9011	0.9177	0.8828	0.9117	0.9246	0.9252

通过上表可以看到，将作者信息与其他论文信息分开提取，模型的预测效果更好，验证了合作者信息的重要性，因此分开提取特征的策略更有利于论文作者名消歧任务.

5　结　论

本文设计了一种基于规则匹配和机器学习消歧的论文作者名识别框架，实现了英文作者名到中文作者的规则匹配，并设计了合适的特征提取和分类器用于论文作者名的消歧. 本文通过比对北京大学论文数据库信息，进行了 7790 条样本的自动化标注，并构建多个子消歧数据集. 在这些数据集上，相应实验表明，通过 TF-IDF 将文本信息和作者信息分开处理的特征提取方法有较好的特征提取效果；K 邻近和 Softmax 分类器在样本数极少的消歧数据集上有较高的预测精度，这两类分类器较适合于高校论文作者名消歧任务. 在实际工作场景中，论文作者名消歧的样本规模并不大，因此针对分类器的改进对于消歧效果的提升并不显著，而文本特征提取的有效性很大程度上决定了消歧效果. 本文后续研究的重点是更有效的特征提取方法.

参考文献：

[1] Smalheiser N R, Torvik V I. Author name disambiguation [J]. Annu Rev Inf Sci Tec, 2009, 43: 1.

[2] 郭舒. 文献数据库中作者名自动化消歧方法应用研究 [J]. 情报杂志, 2013, 32: 132.

[3] Treeratpituk P, Giles C L. Disambiguating authors in academic publications using random forests [C]// Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries. Austin, TX, USA: ACM, 2009.

[4] Han H, Giles L, Zha H, *et al*. Two supervised learning approaches for name disambiguation in author citations [C]//Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries. Tucson, AZ, USA: ACM, 2004.

[5] Han W, Xu B, Zhao T. Study on Chinese person name disambiguation based on multi-stage strategy [C]//Proceedings of 2011 the 8th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD). Shanghai, China: IEEE, 2011.

[6] Levin M, Krawczyk S, Bethard S, *et al*. Citation-based bootstrapping for large-scale author disambiguation [J]. J Am Soc Inf Sci Tec, 2012, 63: 1030.

[7] Salton G, Wong A, Yang C S. A vector space model for automatic indexing [J]. Commun ACM, 1975, 18: 613.

[8] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval [J]. Inform Process Manag, 1988, 24: 513.

[9] 李航. 统计学习方法 [M]. 北京：清华大学出版社, 2012.

[10] 高云龙，左万利，王英，等. 基于集成神经网络的短文本分类模型 [J]. 吉林大学学报：理学版, 2018, 56: 933.

[11] 陈晨，张璐，伍之昂. 词句协同排序的自动摘要算法 [J]. 江苏大学学报：自然科学版, 2016, 37: 443.

[12] 周顺先，蒋励，林霜巧，等. 基于 Word2vector 的文本特征化表示方法 [J]. 重庆邮电大学学报：自然科学版, 2018, 30: 272.

[13] 黄江平，姬东鸿. 基于句子语义距离的释义识别研究 [J]. 四川大学学报：工程科学版, 2016, 48: 202.

引用本文格式：

中　文：邓可君，华凯，邓昌明，等. 基于机器学习的论文作者名消歧方法研究[J]. 四川大学学报：自然科学版, 2019, 56: 241.

英　文：Deng K J, Hua K, Deng C M, *et al*. Research on author name disambiguation method based on machine learning [J]. J Sichuan Univ: Nat Sci Ed, 2019, 56: 241.