文 献 与 数 据 学报
第 1 卷　第 4 期　　2019 年 12 月

基于分段回归的学者学术影响力预测研究 *

池雪花　章成志
（南京理工大学经济管理学院信息管理系，南京　210094）

摘　要：[**目的 / 意义**] 本文对学者未来一段时间内的被引总频次进行预测，以期预测学者的学术影响力，较早发现有发展潜力的学者。[**方法 / 过程**] 考虑到学者的影响力大小不同，本文在传统影响力回归方法上，提出分段回归预测模型。首先，将学者发表的学术论文作为影响力来源，从学术论文中抽取统计类特征、文本内容特征以及网络特征构建特征工程；然后，采用分段回归进行预测，先使用分类方法判断学者论文总被引次数是否为“0”，再使用回归方法预测前一步被引次数为“非 0”的学者的具体被引次数。[**结果 / 结论**] 实验结果表明：融合多种类别特征可以取得更好的分类和回归效果；利用分段回归方法可以取得更好的学者学术影响力预测效果。

关键词：学术影响力预测　被引次数　机器学习　分段回归
分类号：G35
DOI：10.31193/SSAP.J.ISSN.2096-6695.2019.04.04

0　引　言

随着科学技术的不断发展，学术信息不断增多，其中包括会议论文、期刊论文、专利、书籍、会议演示文稿、实验数据等信息资源。相关研究指出，在网上至少包含 1.14 亿个学术文档或记录供用户查询使用，且每天以数以万计的速度增长，标志着学术界也进入了学术大数据时代[1]。丰富的学术数据为研究学者相关信息提供了重要资源，如衡量学者的学术影响力，为学者职称晋升、项目资助等提供依据等。当前，学术界已提出一些衡量学者学术影响力的方法，如：论文被引量、期刊影响因子、作者 h 指数等。其中，论文被引量是一个重要而直观的指标。

由于论文发表和引用具有滞后性，学者的影响力不能及时得到显现。如果能够预测一个学者的学术影响力，就可以及早发现有潜力的学者，这对于专家推荐、评审人寻找等工作具有重要的

* 本文系江苏省“333 工程”科研资助项目“科技文本大数据的知识挖掘与发现技术研究”（项目编号：BRA2018352）研究成果之一。
[作者简介] 池雪花（ORCID：0000-0003-2726-2951），女，硕士研究生，研究方向为文本挖掘与科学计量，E-mail：joyce.chi@qq.com；章成志（ORCID：0000-0001-9522-2914，通讯作者），男，博士，教授，博士生导师，研究方向为信息组织、信息检索、数据挖掘及自然语言处理，E-mail：zhangcz@njust.edu.cn。

意义。

本文以学者发表的学术论文作为影响力来源，对学者的总被引频次进行预测。考虑到学者的影响力大小不同，很多论文的“零被引”造成学者总被引频次为“0”，本文采用分段方法对学者的被引次数进行预测。首先判断学者论文总被引次数是否为“0”，再预测剩余“非 0”学者的论文总被引次数，由此完成学者的被引次数预测。

1　相关研究概述

学术影响力研究是图书情报与科学学领域的关注焦点。本节从学术影响力评价方法与影响力预测两个方面对相关研究进行概述。

1.1　学术影响力评价方法

学者学术影响力评价主要有三种方法。首先是同行评议，这种定性的方法具有一定的权威性，但是花费时间长，具有主观性。其次是通过学者的学术成果来做定量的评价，将成果被引次数作为评价学术影响力的主要依据。这个方法最早于 1970 年由 Garfield 提出，用学者所发表成果的总被引频次来表示学者的影响力大小，代表同行对学者及其学术成果的肯定[2]。由于总被引频次仅考虑了论文数量，忽略了论文质量，Hirsch 提出 h 指数来解决这个问题[3]。其后，出现了一系列基于 h 指数的影响力计算指数，如 g 指数[4]、R 指数和 AR 指数[5]，还有一系列 h 指数的衍生指数，如 hP[6][7]、hm[8][9]、h-maj[10][11]，适用于团队或者学术合作中的学者影响力评价。此外，随着科学交流的网络化，学界引入替代计量指标，用 Twitter、博客、评论、标签、维基百科的引用等指标来评价学者的影响力，这被许多学者认为是补充或者替代同行评议的重要方法。

综上，尽管对学者影响力的评价具有多种方法，但是基于论文总被引频次的评价方法简单且易于执行，众多指数皆是在被引频次的基础上进行拓展和改进，在学术界得到了广泛应用。因此，本文直接选取被引频次来衡量学者的影响力。

1.2　学术影响力预测研究

现有研究中关于影响力预测的方法有很多，其中常用的是回归分析法，将影响引用的因素作为自变量，将被引频次作为因变量，然后进行统计分析并预测。此外，还有较多使用机器学习模型、随机模型或者其他数学模型来进行影响力预测。如：Sohrabi 将页数、题目长度、作者数量、期刊年龄、关键词数量、关键词在摘要中的重复率等作为影响特征，使用逻辑回归和最小二乘线性回归方法预测教育学期刊的被引次数[12]；Acuna 等人用 h 指数表示科学家的影响力，然后采用弹性网正则化的方法预测其 h 指数[13]；Lawrence 通过逻辑回归、支持向量机、决策树三种分类器，构建被引量预测模型[14]；Wang 用随机过程理论提出了引用过程的随机模型预测期刊长期被引[15]；Borsuk 和 Budden 等学者使用广义线性模型（GLM）估算了第一作者的性别、作者数量和论文语言这些因素对被引量的影响，发现作者数量因素有着显著的影响[16]。总的来看，预测影响力分为传统回归统计分析方法和机器学习方法，具体使用哪种方法可根据学科差异和影响因素来决定。

2　研究方法

2.1　基本思路

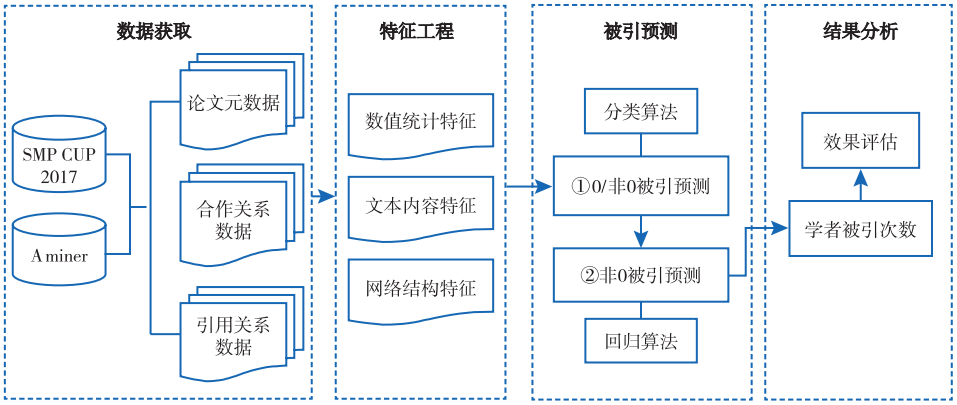

图 1　学者学术影响力预测研究思路

本文以学者总被引频次所代表的学术影响力为研究对象，以学者发表的学术论文作为影响力来源，并利用论文集、论文引用关系和论文合作关系来预测学者的学术影响力。本文基于学术成果的学者学术影响力预测研究思路如图 1 所示。首先，从 SMP CUP 2017 和 Aminer 获得每一位学者的论文元数据、合作关系和引用关系数据。关于 SMP CUP 2017 和 Aminer 数据的详细说明可参见 2.2 小节。然后，对数据进行预处理，主要是构建特征工程，本文充分考虑论文的特征，具体包括数值统计特征、文本内容特征、网络结构特征。

特征工程构建完成后，进行模型的选择、学习训练和预测。本文将被引预测分为两个阶段：第一阶段，进行学者的被引次数是否为“0”的二分类预测，如果学者的被引次数判断为“0”，则预测结束，若学者的被引次数不等于“0”，则进入第二阶段；第二阶段，采用回归算法预测第一阶段被引次数判断为“非 0”的学者的具体被引次数。

在结果分析阶段，进行多种分类模型和回归模型的性能优劣对比分析，选出训练效果最优的分类和回归模型，并对实验结果进行效果评估。

2.2　实验数据说明

本文使用的数据来自 SMP CUP 2017 竞赛数据集（来源网址为：https://biendata.com/competition/scholar/），竞赛旨在构建精准的学术画像，探究前沿的画像技术。论文数据包括题目、作者、发表时间、发表期刊或会议名称和引文信息。为了提取文本特征，本文还通过 Aminer 补充了论文的摘要内容和关键词信息，Aminer 是一个研究者学术搜索网站，已集成了 2 亿多条文献数据（来源网址为：https://www.aminer.cn/citation）。经过筛选，共收集到论文 3 081 997 篇，进一步扩充摘要的论文 1 673 380 篇，扩充关键词的论文 1 710 068 篇，每篇论文数据样例如表 1 所示。

表 1　论文数据样例

标识	含义	举例
#index	论文编号	*1*
#@	论文作者	*Alberto Villa,JocelynChanussot,*
#*	论文题目	*Unsupervised methods for the classification of hyperspectral images with low spatial resolution.*
#abstract	论文摘要	*The problem of structure detection and unsupervised classification of hyperspectral images with low……*
#keyword	论文关键词	*high spectral resolution\spatial resolution……*
#t	发表时间	*2013*
#c	发表期刊或会议名称	*Pattern Recognition*
#%	引用该文的论文编号	*1775098*

本文将学者数据分为训练集和测试集，划分情况如表 2 所示，训练集的学者数量为 100 万人，测试集数量为 30 万人，总计 130 万人。

表 2　学者数据集

（单位：人）

数据集划分	训练集	测试集	总计
学者数量	1 000 000	300 000	1 300 000

笔者对学者的被引情况进行统计分析后发现，被引次数为“0”的学者占大约 25%，呈现长尾分布。因此，本文提出分段回归的预测方法，将被引次数为“0”的学者排除出预测数据集。

2.3　关键技术描述

2.3.1　数据预处理

对每一篇文章和每一位作者进行编号，进行唯一标识。汇总每一位学者的论文发表、论文引用和合作信息等，为下一步特征体系工程的构建提供基础。处理学者的被引情况时，每一行表示一位学者，数字表示引用的论文编号。对论文内容进行词性标注、停用词过滤以及去除特殊噪音字符等，完成对论文文本的清洗。

2.3.2　特征体系构建

本文的论文特征体系，包含数值统计特征、文本内容特征和网络结构特征，具体如表 3 所示。数值统计特征包括发表论文总数、发表期刊总数、作者发文生涯年限、合作者数量、最大文章被引数、最小文章被引数、平均文章被引数等统计数值，本文对数值统计特征进行了 min-max 标准化处理［17］。论文的文本内容特征包括论文题目、论文关键词、论文摘要，本文采用 TF*IDF［18］方法，对每篇论文进行了文本内容特征提取，维度为 1 000。网络结构特征指论文信

息构成的引文网络和合作网络，本文采用 Line[19] 方法进行了网络结构特征提取，通过网络学习，将合作网络和引文网络中的节点学者转化为低维的网络特征向量，维度各为 500，最大化保留网络结构并将网络中的节点表示成低维、有实值的向量，以直接作为机器模型的输入。

表 3　特征体系类别

特征类别	特征名称
数值统计特征	发表论文总数
	发表期刊总数
	作者发文生涯年限
	合作者数量
	最大文章被引数
	最小文章被引数
	平均文章被引数
文本内容特征	论文题目
	论文摘要
	论文关键词
网络结构特征	引文网络
	合作网络

2.3.3　分类和回归预测模型

根据数据的分布特点，将预测任务分解成两步。首先，利用分类模型判断学者的被引次数是否为“0”，若被引次数为“0”则预测停止，若“非 0”则再利用回归模型预测学者的具体被引次数。

本文尝试多种不同的分类模型和回归模型。采用的分类模型包括朴素贝叶斯模型（Naive Bayes，NB）[20]、支持向量机模型（Support Vector Machines，SVM）[21]、逻辑斯蒂回归模型（Logistic Regression，LR）[22]、决策树模型（Decision Tree，DT）[23]、K 近邻模型（K-Nearest Neighbour，KNN）[24]、梯度提升决策树模型（Gradient Boosting Decision Tree，GBDT）[25]、随机森林模型（Random Forests，RF）[26]。采用的回归模型有线性回归模型（LinearRegression，LR）[27]、决策树回归模型（Decision Tree，DT）[23]、K 近邻回归模型（K-NeighborsRegressor,，KNR）[24]、支持向量机回归模型（Support Vector Regression，SVR）[21]、随机森林模型（RandomForests，RF）[26]、梯度提升回归模型（Gradient Boosting Regression，GBR）[25]。

在进行分类任务时，本文根据训练集中 100 万个学者训练分类模型，采用十折交叉验证的方法训练多种分类模型并对不同类型特征的分类效果表现进行比对分析，最后从中选择出效果最好的特征和分类模型组合。在进行回归任务时，取出训练集中被引次数为“非 0”的学者进行回归模型的训练，也采用十折交叉验证的方法训练多种回归器并对不同类型特征的分类效果表现进行

比对分析，选取出效果最好的回归模型和特征组合模式。训练后，使用最优组合模型来预测测试集中学者的被引次数是否为“0”。

3 结果分析

3.1 分类和回归模型的测试与选择

本文通过准确率（P）、召回率（R）、F1 值[28]来评估模型的训练效果（P 反应分类预测模型的准确性；R 反应预测模型的完善性；F1 值则综合准确率和召回率，确定最优的类型特征和分类模型组合），并进行不同类型特征、不同分类模型的对比分析。

对数值统计特征、文本内容特征、网络结构特征和融合三类特征的总特征分别在不同分类模型下进行对比分析，实验结果见表 4。可以发现，在分类过程中利用总特征进行训练的效果是最好的，所有分类模型下的 F1 值都在 0.8 以上，其他三类特征的效果无明显差异。因此，进一步利用总特征对测试集数据进行不同分类模型的测试。

利用总特征对测试集数据进行不同分类模型的测试效果见表 5。可以看出，朴素贝叶斯模型的效果是最差的，F1 值为 0.85，其他分类模型的效果都在 0.90 以上，分类效果较好的是支持向量机、决策树、梯度提升决策树模型，F1 值皆在 0.91 ~ 0.94 之间，效果最好的是随机森林模型，高达 0.97。从模型和数据角度分析原因，相对其他分类模型，随机森林有很大的优势，模型泛化能力强，且很好地平衡了实验数据存在的类别不均衡问题，在分类实验过程中表现较好。因此，在判断被引次数是否为“0”的分类阶段，本文采用随机森林模型进行分类。

表 4 训练集不同特征分类模型效果

特征指标 \ 分类模型		NB	SVM	LR	DT	KNN	GBDT	RF
统计特征	P	0.79	0.70	0.72	0.78	0.76	0.72	0.78
	R	0.76	0.70	0.71	0.78	0.75	0.72	0.76
	F_1	0.77	0.70	0.71	0.78	0.75	0.72	0.77
文本特征	P	0.83	0.76	0.77	0.80	0.63	0.77	0.83
	R	0.57	0.72	0.82	0.79	0.60	0.82	0.84
	F_1	0.67	0.73	0.79	0.79	0.61	0.79	0.83
网络特征	P	0.72	0.85	0.87	0.74	0.88	0.87	0.85
	R	0.85	0.79	0.83	0.58	0.85	0.83	0.81
	F_1	0.78	0.81	0.84	0.65	0.86	0.84	0.83
总特征	P	0.88	0.94	0.94	0.99	0.93	0.94	0.98
	R	0.77	0.94	0.94	0.97	0.91	0.94	0.98
	F_1	0.82	0.94	0.94	0.98	0.92	0.94	0.98

表 5　测试集不同分类模型效果

指标 \ 分类模型	NB	SVM	LR	DT	KNN	GBDT	RF
P	0.91	0.93	0.93	0.96	0.93	0.93	0.98
R	0.79	0.93	0.90	0.94	0.92	0.94	0.97
F_1	0.85	0.93	0.91	0.94	0.92	0.93	0.97

3.2　“非 0”被引次数回归预测与结果

接下来利用回归方法对被引次数为“非 0”的学者的被引次数进行预测，同样，使用不同类型特征进行回归，且采用不同回归模型进行对比分析。回归模型的评估方法采用相关指数 R^2，R^2 反映的是在因变量的变差中回归方程所能解释的比例，用来揭示回归方程的拟合程度。R^2 越接近 1，说明回归模型越好，R^2 出现负值则说明模型较差。

从表 6 可以看出，对三类特征进行回归训练的效果各异，只利用数值统计特征进行回归训练的整体效果相比于文本内容特征和网络结构特征较好，在不同回归方法下的 R^2 值皆在 0.5 以上，其中梯度提升回归模型的效果最好。此外，利用总特征进行不同回归模型下的训练，得到的模型效果远好于单类特征，同样是梯度提升回归模型的效果最好。因此，结合训练集模型效果，本文将利用总特征进行不同回归模型的预测。

表 7 为对测试集数据基于总特征的不同回归模型的预测效果。从其中 R^2 值可以看出，GBR 模型的回归效果最好，其次是 RF，KNR，DT，然后是 LR 和 SVR 模型。因此，在“非 0”被引次数的回归阶段，本文将利用总特征以及选用 GBR 模型进行学者被引次数的回归预测。

表 6　训练集不同特征回归模型预测效果

指标 \ 回归模型 \ 特征		LR	DT	KNR	SVR	RF	GBR
R^2	数值统计特征	0.529	0.504	0.590	0.523	0.60	0.673
	文本内容特征	0.257	0.151	0.246	0.135	0.253	0.310
	网络结构特征	0.367	0.042	0.301	0.184	0.484	0.503
	总特征	0.607	0.574	0.634	0.573	0.68	0.708

表 7　测试集不同回归模型预测效果

回归模型	LR	DT	KNR	SVR	RF	GBR
R^2	0.567	0.620	0.63	0.553	0.675	0.72

3.3　单一回归预测对比实验结果

单一回归预测指不事先进行被引次数是否为“0”的判断，直接采用不同回归模型对学者被引次数进行预测，与上小节实验进行对比。具体做法为，先利用训练集训练回归模型，训练效果

如表 8 所示。可以发现，利用总特征进行回归训练的效果是最好的，然后利用训练好的回归模型在测试集上进行测试，结果如表 9 所示。

从表 9 可以看出，对测试集进行单一回归预测时，GBR 模型的效果最好，R^2 值为 0.62。而在上文进行的分段预测中，GBR 模型的 R^2 值为 0.72，这说明分段预测方法相比于单一回归方法能取得更好的效果。将表 7 和表 9 进行整体对比可以发现，不同回归模型在单一回归预测时均比分段预测的效果差。由此可见，通过二分类方法排除掉大量被引次数为“0”的学者数据，可以使得被引次数为“非 0”的学者数据更加规整，从而提高回归预测的效果。

表 8　训练集单一回归预测效果

指标特征 \ 回归模型		LR	DT	KNR	SVR	RF	GBR
R^2	数值统计特征	0.503	0.431	0.587	0.510	0.532	0.635
	文本内容特征	0.238	−0.05	−0.513	0.129	0.240	0.288
	网络结构特征	0.345	0.053	0.270	0.160	0.450	0.502
	总特征	0.510	0.434	0.620	0.445	0.622	0.625

表 9　测试集单一回归预测效果

回归模型	LR	DT	KNR	SVR	RF	GBR
R^2	0.513	0.474	0.612	0.450	0.615	0.62

4　结论与展望

预测学者的学术影响力，更早发现有潜力的学者，有利于全面掌握科学研究和科研人才的发展状况。本文将学术论文的被引次数作为评估学者影响力的计量指标，从学术论文中提取多种类型特征，并采用机器学习的方法进行被引次数的预测实验。预测实验过程分为两个阶段，首先判断学者的被引次数是否为“0”，进行二分类划分，然后利用回归方法进行具体被引次数的预测。在分类和回归过程中，本文针对学术论文不同类型的特征进行了多种分类和回归模型测试，并选择效果最好的一种分类和回归模型用于最终的预测。实验发现，基于二分类判断的分段回归方法比直接进行预测的单一回归具有更好的预测效果。

参考文献

［1］XIA F., WANG W., BEKELE T. M., et al. Big scholarly data: A survey［J］. IEEE Transactions on Big Data, 2017, 3(1): 18-35.

［2］GARFIELD E. Citation Indexing for Studying Science［J］. Nature, 1970, 227(5259): 669-671.

［3］HIRSCH J E. An index to quantify an individual's scientific research output［J］. Proceedings of the National Academy of Sciences of the United States of America, 2005, 102(46): 16569-16572.

[4]EGGHE L. Theory and Practice of the g-Index[J]. Scientometrics, 2006, 69(1): 131-152.

[5] 金碧辉 , RONALD R. R 指数、AR 指数：h 指数功能扩展的补充指标 [J]. 科学观察 , 2007, 2(3):1-8.

[6]WAN J K, HUA P H, ROUSSEAU R. The pure h-index: calculating an author's h-index by taking co-authors into account[J]. Collnet Journal of Scientometrics& Information Management, 2007, 1(2):1-5.

[7]CHAI J C, HUA P H, ROUSSEAU R, et al. The adapted pure h-index [C]. In: Proceedings of the Fourth International Conference on Webometrics, Informetrics and Scientometrics & Ninth COLLNET Meeting. Berlin, Germany, 2008:1-6.

[8]SCHREIBER M. A modification of the -index: The -index accounts for multi-authored manuscripts[J]. Journal of Informetrics, 2008, 2(3):211-216.

[9]SCHREIBER Michael. A case study of the modified Hirsch index hm accounting for multiple coauthors[J]. Journal of the American Society for Information Science and Technology, 2009, 60(6): 1274-1282.

[10]HU X, ROUSSEAU R, JIN C. In those fields where multiple authorship is the rule, the h-index should be supplemented by role-based h-indices[J]. Journal of Information Science, 2010, 36(1):73-85.

[11]ROUSSEAU R, HU X, Association K. An Outgrow Index[J]. Science Focus, 2011, 57(3): 287-290.

[12]SOHRABI B, IRAY H. The effect of keyword repetition in abstract and keyword frequency per journal in predicting citation counts[J]. Scientometrics, 2017, 110(1): 1-9.

[13]ACUNA D E,ALLESINA S, KORDING K P. Future impact predicting scientific success[J]. Nature, 2012, 489(7415):201-202.

[14]FU Lawrence D, ALIFERIS C F. Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature[J]. Scientometrics, 2010, 85(1):257-270.

[15]WANG D, SONG C, BARABASI A. Quantifying Long-Term Scientific Impact[J]. Science, 2013, 342(6154):127-132.

[16]BORSUK R M, BUDDEN A E, LEIMU R, et al. The influence of author gender, national language and number of authors on citation rate in ecology[J]. Open Ecology Journal, 2009, 2(1): 25-28.

[17]AL Shalabi L, SHAABAN Z, KASASBEH B. Data mining: A preprocessing engine[J]. Journal of Computer Science, 2006, 2(9): 735-739.

[18]AIZAWA A. An information-theoretic perspective of tf-idf measures[J]. Information Processing & Management, 2003, 39(1): 45-65.

[19]TANG J, QU M, WANG M, et al. Line: Large-scale information network embedding [C]. In: Proceedings of the 24th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee.New York, USA,2015: 1067-1077.

[20]MCCALLUM A, NIGAN K. A comparison of event models for naive bayes text classification [C]. In: Proceedings of the AAAI-98 workshop on learning for text categorization. Madison, USA,1998, 752(1): 41-48.

[21]HEARST M A, DUMAIS S T, OSUNA E, et al. Support vector machines[J]. IEEE Intelligent Systems and their applications, 1998, 13(4): 18-28.

[22]HARRELL Jr, FRANK E. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis[M]. Berlin: Springer, 2015: 311-325.

[23]SAFAVIAN S R, LANDGREBE D. A survey of decision tree classifier methodology[J]. IEEE transactions on systems, man, and cybernetics, 1991, 21(3): 660-674.

[24]SOUCY P, MINEAU G W. A simple KNN algorithm for text categorization [C]. In: Proceedings of the 2001International Conference on Data Mining.Washington, USA,2001: 647-648.

［25］KE G, MENG Q, FINLEY T, et al. Lightgbm: a highly efficient gradient boosting decision tree［C］. In: Proceedings of the 2017 Advances in Neural Information Processing Systems. Long Beach, USA,2017: 3146-3154.

［26］BREIMAN L. Random forests［J］. Machine Learning, 2001, 45(1): 5-32.

［27］SEBER G A F, Lee A J. Linear regression analysis［M］. New York: John Wiley & Sons, 2015.

［28］SALTON G, McGill M J. Introduction to modern information retrieval［M］. McGraw-Hill. 305-306.

Research on Predicting the Academic Influence of Scholars Based on Stepwise Regression

CHI Xuehua　ZHANG Chengzhi

(Department of Information Management, Nanjing University of Science & Technology, Nanjing 210094, China)

Abstract: **［Purpose/significance］** This paper predicts total citation frequency of scholars within a certain period of time in the future, in the hope of predicting scholar's academic influence and discovering potential scholars earlier. **［Method/process］** Considering that the academic influence of scholars can be divided into different levels, this paper proposes a stepwise regression prediction model based on the traditional influence regression method. Firstly, taking the academic papers that scholars published as the sources of influence, construct the feature engineering of academic papers from their statistical features, text content features and network features. Then, whether the total citation is zero or not will be judged by automatic classification method. Furthermore, this paper applies regression technology to predict scholar's citation frequency whose total citation is non-zero. **［Result/conclusion］** The experiment shows that better categorization and regression can be achieved through combining multiple features and better prediction by stepwise regression.

Keywords: Academic influence prediction; Citation frequency; Machine learning; Stepwise regression

（本文责编：王秀玲）