文章编号：2095-2163(2021)05-0134-09 中图分类号：TP391.1 文献标志码：A

考察文献活跃度特性的个性化引文推荐研究

崔志慧，彭兰一香，熊 曦，王名扬
（东北林业大学 信息与计算机工程学院，哈尔滨 150040）

摘 要：本文通过将引文推荐问题转化为文献是否被引用的二元分类问题，发掘出影响文献被引用的关键特征，并依据这些特征提升了引文推荐的性能。在提取研究者的个性化引用偏好和常用的文献计量学特征的基础上，加入表征文献活跃度的特征指标，构建用以进行二元分类的特征库。利用 Relief-F、RFE 和 LR3 种方法从特征库中筛选出对文献是否被引具有重要价值的特征；利用朴素贝叶斯、SVM 和 Bagging3 种分类器验证基于这些重要特征的引文推荐效果。实验结果表明，文献的活跃度特征、研究者的个性化引用偏好特征和文献对之间的主题相似性特征是提升引文推荐性能的关键因素。相对于基线方法，在这些关键特征上实施引文推荐，其准确率、召回率和 $F1$ 值分别提升了 6%、29%和 26%。

关键词：引文推荐；文献活跃度；引用偏好；主题相似性

The investigation of personalized citation recommendation based on the characteristics of activity

CUI Zhihui, PENG Lanyixiang, XIONG Xi, WANG Mingyang
(College of Information and Computer Engineering, Northeast Forestry University, Harbin 150040, China)

【Abstract】 This paper explores the key features that affect the citation of an article by transforming the citation recommendation problem into a binary classification problem of whether an article is cited or not, and improves the performance of citation recommendation based on these features. The characteristics of article activity in the scientific community is an important indicator of whether the article can receive attention from researchers. A feature database for binary classification is constructed by the feature representing article activity in combination with extracting previous personalized citation preferences and common bibliometrics features. Three feature selection methods, Relief-F, RFE, and LR methods, are used to select the features which are important for citation from the feature database. Naive Bayes, SVM, and Bagging classifiers are used to verify the effect of citation recommendation based on these important features. The experimental results show that these features, the article activity characteristics, the personalized citation preference characteristics and the topic similarity between the article pairs, are the key factors to improve the performance of citation recommendation. Compared with the baseline method, this paper implements citation recommendation based on these key features, and its accuracy, recall rate and $F1$ index improved by 6%, 29% and 26%, respectively.

【Key words】 citation recommendation; article activity; references preferences; topic similarity

0 引 言

随着互联网的高速发展，每年发表的科技文献总数呈指数增长。据统计，仅 2018 年中国科研人员发表在国际、国内的优秀文献总量就达到 31.59 万篇[1]。对研究者来说，从如此海量的数据中定位满足科研需要的文献是非常困难的。引文推荐能针对某一具体的研究主题和学术文献，自动地为研究者推荐合适的相关成果和文献。借助于引文推荐，研究者可快速获取到与其研究相关的文献资料，从而在一定程度上提高撰写学术文献的效率。

2001 年，Basu C 等首次提出文献推荐的概念，给出文献推荐的过程是如何找到与用户兴趣相匹配的文献的过程，其核心问题在于如何表达用户兴趣和目标文献[2]。在之后的研究中，研究者们也将关注点更多放在对用户兴趣的建模和文献间相似度的计算上。2007 年，Strohma 等提出引文推荐的概念，并结合文本相似性和图模型方法对引文推荐问题进行了初步探索[3]。Bethard S 等结合用户的历史引用信息和引用偏好来为目标文献推荐参考文献列表[4]；2010 年，He Q 等人将词频信息和文献的主题分布作为主要特征，实施引文推荐[5]；Pohl 等基于

基金项目：国家自然科学基金(71473034)。

作者简介：崔志慧(1998-)，女，本科生，主要研究方向：自然语言处理；彭兰一香(1999-)，女，本科生，主要研究方向：自然语言处理；熊 曦(1999-)，男，本科生，主要研究方向：自然语言处理；王名扬(1980-)，女，博士，教授，主要研究方向：自然语言处理、数据挖掘。

通讯作者：王名扬 Email：wangmingyang@ nefu.edu.cn

收稿日期：2021-02-10

用户下载文献的行为记录进行引文推荐[6]。2013年，刘盛博等以全文数据为基础，构建基于引用内容的引文检索与推荐系统[7]；Liu Yaning 实现了基于翻译模型和用户过滤算法的混合推荐模型[8]；2014年，蔡阿妮等结合文献的内容信息和引用关系来对引文进行推荐[9]；王萌星等基于主题社区和双层引用网络的学术推荐方案，向用户推荐作者和论文[10]；刘亚宁等在考察用户的兴趣和其知识水平的基础上实施引文推荐[11]；Guo LT 等运用深度学习技术获取用户的兴趣模型，并改进个性化重排序算法实施推荐[12]；Ali Z 等从 6 个角度对基于深度学习的引文推荐模型进行综述[13]；刘洋利用文献间的语义关联度和作者间的关系构造网络模型实施推荐[14]；Wang J 等将作者信息和引文关系整合到用分布式矢量表示的引文上下文和论文中，提出了基于端到端记忆网络的上下文感知引文推荐模型[15]。

综上，为了实现更精准快捷的推荐，研究者从用户和文献两个角度对引文推荐问题进行了深入的研究，但是这些已有成果的推荐效果仍然差强人意。在这些研究中，学者们均未讨论文献的活跃度特征在引文推荐中的作用。实际上，文献的活跃度体现了文献在科学社区的可见度，活跃度较高的文献将具有更高的被研究者关注的机会，而这种机会将在一定程度上促使文献被研究者引用，成为研究者文献中的参考文献。

在评价引文推荐效果时，往往将被推荐文献是否真正成为目标文献参考文献中的一员来作为评价的依据。这实际上已经将引文的推荐问题转化成了文献是否被引用的二元分类问题。为此，本文将引文推荐问题转换为文献是否被引的二元分类问题，提取表征文献活跃度的特征，结合研究者的个性化引用偏好和常用的文献计量学特征，构建二元分类问题的特征库。利用机器学习方法从特征库中提取有利于文献被引用的关键特征，并基于这些特征实现引文推荐。

1　相关研究

2010 年，He Q 等利用引文上下文的差异性将引文推荐任务细分为局部引文推荐和全局引文推荐[5]。局部引文推荐，是指为目标文献的局部上下文推荐合适的引文列表；而全局引文推荐，是根据目标文献的标题内容和摘要内容为其从整体上推荐引文列表。本文主要对全局引文推荐问题进行研究，仅对全局引文推荐相关的工作进行分析。由于推荐技术主要用于实现用户兴趣与待推荐对象之间的匹配，因此推荐算法是推荐问题的核心，引文推荐问题也不例外。在全局引文推荐领域，研究者主要用到的推荐算法主要包括协同过滤推荐和基于图的引文推荐。

协同过滤推荐根据作者的引用偏好和文献间的相关性来预测作者与文献间的引用关系。McNee 等将作者视为用户，文献视为商品，利用文献之间的引用关系建立评分矩阵，从而将引文推荐问题转化为普通的商品推荐问题[16]；Pohl 等把用户下载文献的行为作为用户的活动记录，并将访问量较高的文献推荐给用户[6]；Tang 等综合引用关系和文献文本内容间的相关性来实施推荐[17]；Choochaiwattana 提出一种基于标签的引文推荐机制，通过用户创建的标签来为用户推荐引文[18]；倪卫杰构建用户兴趣模型和文献兴趣模型，为特定用户推荐引文[19]。Wang 等根据用户的历史行为构建用户偏好模型来实施推荐[20]；Gipp 等在引文推荐中使用了基于内容的协同过滤方法[21]；陈将引文推荐问题视为分类问题，使用文献的内容信息预测文献可能的参考文献列表[22]。Pan 等用标签对用户进行个人配置，计算文献标签向量与个人配置向量间的相似度来实施文献推荐[23]；Khadka 等结合引文位置和引文上下文特征，使用主题建模来实现引文推荐[24]；Zhang 等引入结构上下文的概念来提升引文推荐的效果[25]。

由于异种类型对象和其之间的关系可以简单的用一个图来表示，所以基于图的方法可以很容易地被应用到包含多种类型数据的数据集上来实施推荐。Gori 等构建文献间的同构网络，使用 PageRank 算法计算权重来实施推荐[26]；Meng 等构建四层多元图，利用重启随机游走的方法计算目标文献与候选文献间的相似性来实施推荐[27]；Jardine 等在引文网络图中加入主题分布信息，来改进 PageRank 算法的转移概率以实施推荐[28]；Cai 等构造三层图模型，包括作者层、文献层和出版商层，在此基础上进行推荐[29]；Pan 等提出了一种包含多元信息异构图的引文推荐方法[30]；Gupta 等综合文献内容和文献的结构关系来表示文献，在网络图中计算文献间的相似度进行推荐[31]；李飞构建基于作者和引文的异构图，利用 Deepwalk 算法进行推荐[32]；陈洁等将多粒度属性网络表示学习应用于引文推荐工作中来解决在异质网络中的引文推荐问题[33]。

虽然这些工作实施推荐的角度不同，但其核心问题仍然离不开如何对用户兴趣和目标文献建模，

以及如何度量目标文献和待推荐文献的相似性。尽管这些工作已尽可能广泛地提出了解决以上核心问题的思路,但引文推荐的精度仍不太理想,且有些推荐算法过于复杂,并不能很好地进行推广应用。本文致力于在这些已有工作的基础上,发掘尽可能简洁的特征来实施推荐,取得较为可观的推荐效果。

在当前的推荐工作中,还鲜有研究者考察待推荐文献的活跃程度相关的指标。如果一篇文献在近年来获得了较高的被引频次,说明该文献在科学社区具有较高的认可度,同时也具有较高的可见度。这种较高的可见度能带给文献更高的被研究者关注的机会,从而提升其被研究者引用的可能性。基于这种考虑,本文将文献的活跃度指标引入推荐过程,并探讨这种加入是否能显著提升引文推荐的效果。

2 问题定义

本文构建的考察文献活跃度的引文推荐系统的输入和输出信息如下:

(1)输入

①目标文献:需要被推荐引文的文献集合 P;

②待推荐文献:待推荐文献集合 R, 由目标文献 P 的参考文献列表中实际出现的参考文献集合 B, 和未被目标文献 P 引用的文献集合 N 构成。其中,未被目标文献引用的文献集合 N 中的文献来自于与 B 中文献在同一期刊、同一年份发表的其他文献。

(2) 特征集合 X。由用户的个性化引用偏好特征、常用的文献计量学特征和文献的活跃度特征构成。

(3)输出。根据筛选出的特征,取 3 个分类器推荐结果的并集,为每篇目标文献生成一个按照被推荐概率排好序的推荐文献列表 L。

3 考察文献活跃度特性的引文推荐

本文将引文推荐问题看成待推荐文献 R 是否被目标文献 P 引用的二分类问题。为此,需要首先构造用于分类的特征库 X。在已有的推荐工作中,研究者利用不同算法证实了用户的兴趣和文献对间的相似性在引文推荐中的重要作用。本文也将这些特征考虑进来,同时构造表征文献活跃程度的指标,共同生成分类问题的特征库 X。在此特征库基础上,运用 Relief-F、RFE 和 LR3 种特征选择方法,对特征库 X 中的各特征 x 进行重要性排序;利用朴素贝叶斯、SVM 和 Bagging3 种分类器验证特征组合的分类性能,提取对文献是否被引用具有重要影响的特征。依据这些关键特征,生成针对目标文献的待推荐文献列表。本文提出的考察文献活跃度特性的引文推荐算法的示意图如图 1 所示。

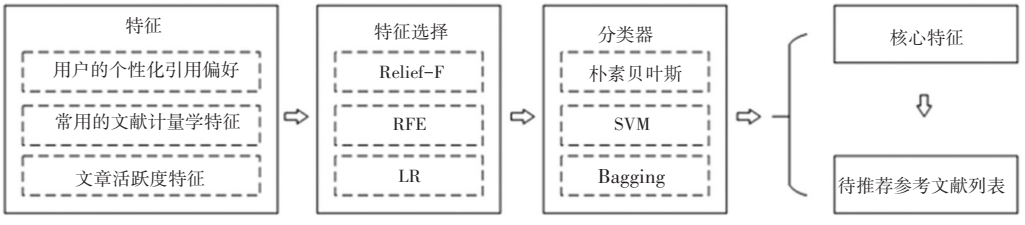

图 1 基于文献活跃度特征的引文推荐实验流程示意图

Fig. 1 Schematic diagram of citation recommendation experiment based on paper activity characteristics

3.1 构建引用分类问题的特征库

3.1.1 用户的个性化引用偏好特征

用户的个性化引用偏好特征主要用来反映用户在撰写科研成果时的引用习惯。本文主要从用户是否偏向于引用其之前发表的文献,是否偏向于引用其之前引用过的文献,是否偏向于引用合作者的文献,以及是否偏向于引用之前引用过的作者所写的文献,4 个角度来表征用户的个性化引用偏好,见表 1。

为获取这些特征,需要为每篇目标文献采集如下信息:

(1)目标文献的所有作者发表的文献构成的集合;

(2)目标文献的所有作者曾经引用过的文献构成的集合;

(3)所有曾经同目标文献的作者合作过的其他作者构成的集合;

(4)目标文献的所有作者曾经引用过的其他作者构成的集合。

表 1 用户的个性化引用偏好特征

Tab. 1 Personalized reference preference characteristics of users

序号	特征
x_1	是否作者之前引用过的文献
x_2	是否为作者之前发表的文献
x_3	是否作者合著者发表的文献
x_4	待推荐文献的作者是否当前作者之前引用过的作者

3.1.2　常用的文献计量学特征

在引文推荐工作中常被研究者用到的文献计量学特征见表 2，符号 p 代指目标文献，符号 r 代指待推荐文献。这些特征涵盖了待推荐文献的作者、所在期刊、基金资助情况，以及待推荐文献与目标文献间的相似度等指标。

表 2　常用的文献计量特征

Tab. 2　Commonly used bibliometric characteristics

序号	特征
x_5	待推荐文献 r 中作者的最高 h 指数
x_6	目标文献 p 和待推荐文献 r 作者关键字的相似度
x_7	待推荐文献 r 发表至今的总被引用频次
x_8	目标文献 p 和待推荐文献 r 标题的相似度
x_9	目标文献 p 和待推荐文献 r 摘要的相似度
x_{10}	目标文献 p 和待推荐文献 r 主题的相似度
x_{11}	目标文献 p 和待推荐文献 r 是否来自于同一个学科
x_{12}	r 所在期刊所发文献在当年的篇均被引频次
x_{13}	待推荐文献 r 所在期刊的 SJR 值
x_{14}	待推荐文献 r 所在期刊的 SNIP 值
x_{15}	待推荐文献 r 的作者是否来源于美国
x_{16}	待推荐文献 r 的作者是否来源于欧洲
x_{17}	目标文献 p 和待推荐文献 r 的发表年代间隔
x_{18}	待推荐文献 r 是否有资金资助

其中，特征 x_6 作者关键字是对文献作者的国籍、所属单位和研究领域等的概括性介绍。本文采用 Jaccard 系数计算目标文献 p 和待推荐文献 r 在作者关键字上的相似度。x_6 的值越大，说明目标文献 p 和待推荐文献 r 的作者越相似。对于目标文献 p 和待推荐文献 r 的作者关键字集合，Jaccard 系数为 p 与 r 交集的大小与 p 与 r 并集的大小的比值，定义如式(1)：

$$J(p,r)=\frac{|p\cap r|}{|p\cup r|}=\frac{|p\cap r|}{|p|+|r|-|p\cap r|}. \quad (1)$$

利用余弦相似度计算特征 x_8、x_9、x_{10} 的值。利用 Python 中的 jieba 算法分别对目标文献 p 和待推荐文献 r 的标题、主题和摘要进行分词，去掉停用词，主题是 Scopus 数据库中对文献研究内容的高度概括。之后结合剩下的词的词频构建标题、主题和摘要的向量，最后利用余弦相似度计算目标文献 p 和待推荐文献 r 在 3 个方面的相似度。余弦相似度的计算公式(2)如下：

$$Sim_i=\frac{p.r}{\sqrt{\sum_{j=1}^{n}p_j^2\sum_{j=1}^{n}t_j^2}}. \quad (2)$$

收集处理完上述特征后，利用线性函数归一化方法将上述特征归一化到[0.01,0.99]范围内，消除不同特征的取值范围对分类的影响。

3.1.3　文献的活跃度特征

本文用待推荐文献在近两年内的引用情况来度量文献的活跃程度。在本实验中，选取的目标文献均发表在 2018 年，则对待推荐文献而言，表征其活跃程度的引用指标均来自于其在 2016 和 2017 年的引用情况。

本文采集了待推荐文献在近 2 年内的总被引频次、近 2 年内的引证国家数量、近 2 年内的引证期刊数量、近 2 年内的引证机构数量和近 2 年内的引证学科数量，来构造文献的活跃度特征。这些指标反映了在近 2 年内待推荐文献在科学社区内的影响可见度。对一篇待推荐文献 r 而言，如果在近 2 年内得到了来自更多的国家、机构、期刊和学科的引用，则意味着该文献受到了更多学术同行的认可，在科学社区内产生了较为广泛的影响。而这种影响将推动其继续被学者关注，并持续转化为学者研究成果的参考文献。

为结合以上 5 个引用指标生成综合的文献活跃度特征，本文利用熵权法为每个特征赋权重，求得 5 个特征值的加权和以代表本文的文献活跃度特征。根据待推荐文献在近 2 年内的总被引频次、以及其被不同国家、期刊、机构和学科的引证数量的值构成这 5 项子特征的数据矩阵 A，式(3)，其中 X_{ij} 为第 i 个文献的第 j 个特征的数值。

$$A=\begin{pmatrix}X_{11} & \cdots & X_{15}\\ \vdots & \ddots & \vdots\\ X_{n1} & \cdots & X_{n5}\end{pmatrix}, \quad (3)$$

计算 x_{ij} 占特征 x_j 的比重，式(4)：

$$P_{ij}=\frac{X_{ij}}{\sum_{i=1}^{n}X_{ij}}(j=1,2,3,4,5), \quad (4)$$

计算第 j 个特征的熵值，式(5)：

$$e_j=-k\sum_{i=1}^{n}P_{ij}ln(P_{ij}), \quad (5)$$

根据 e_j 计算第 j 个特征的熵冗余度，式(6)：

$$g_j=1-e_j, \quad (6)$$

根据 g_j 求特征的权数，式(7)：

$$W_j=\frac{g_j}{\sum_{j=1}^{5}g_j}, \quad (7)$$

将求得的每个特征的权重和其值求加权和，得

出本文的文献活跃度特征 x_{19}，式(8)：

$$x_{19} = s_i = \sum_{j=1}^{5} W_j P_{ij} X_{ij}, \tag{8}$$

3.2 特征选择过程

为提取对引文推荐具有重要价值的特征，本文采用 Relief-F、Recursive Feature Eimination(RFE)和 Logistic Regression(LR)3 种特征选择方法对特征进行重要性排序，并结合不同特征组合的分类精度得到影响推荐效果的核心特征子集。

3.2.1 Relief-F

Relief-F 算法通过计算不同特征，区分不同类型样本的能力来为特征赋予权重。其随机从待推荐数据集 R 中选取一个样本 r_i，从 r_i 同类的样本集 C 中找到 k 临近的临近样本 $\{h_j\}$，从与 r_i 不同类的样本集 S 中找到 k 临近的随机样本 $\{m_j\}$，计算特征 x 区分临近样本 $\{h_j\}$ 和随机样本 $\{m_j\}$ 的能力。如果样本 r_i 和临近样本 $\{h_j\}$ 在特征 x 上的距离小于样本 r_i 和随机样本 $\{m_j\}$ 上的距离，则说明该特征对区分同类和不同类的数据是有益的，则增加该特征的权重 W。根据 W 对特征进行排序，获得根据重要程度排序的特征。

求权重 W 的具体算法见公式(9)：

$$W(x) = W(x) - \sum_{j=1}^{k} \frac{diff(x, r_i, h_j)}{mk} + \sum_{C \neq S} \left[\frac{p(C)}{1 - p(S)} \sum_{j=1}^{k} diff(x, r_i, m_j) \right] / (mk), \tag{9}$$

其中，$p(C)$ 为类别 C 在所有类别中所占比例，$p(S)$ 为类别 S 在所有类别中所占比例。diff 定义见公式(10)，其表示样本 r_1 和 r_2 在特征 x 上的差：

$$Diff(x, r_i, r_j) = \begin{cases} \dfrac{|r_i[x] - r_j[x]|}{max(x) - min(x)}, & \text{当 } x \text{ 为连续值时;} \\ 0, & \text{当 } x \text{ 为离散值且} r_i[x] = r_j[x] \text{ 时;} \\ 1, & \text{当 } x \text{ 为离散值且} r_i[x] \neq r_j[x] \text{ 时.} \end{cases} \tag{10}$$

3.2.2 RFE

递归特征消除法是通过递归的方式，不断剔除作用最小的特征，减少特征集的规模来选择需要的特征，RFE 的底层模型很大程度会影响其稳定性。本模型底层采用 SVM，SVM 作为一种基于统计理论的分类方法，将低维线性不可分割的数据在核函数的作用下映射到较高维度而实现线性分割。每个特征对应特定维度，维度的权重由分类器的精度确定，而权重即可视作该特征的重要性。

REF 首先给每个特征赋一初始权重 w_0，然后采用预测模型在这些原始的数据上进行训练，训练后获取特征的最终权重值 w_1，取这些权重值的绝对值，把绝对值最小的特征剔除掉。按照以上步骤，不断循环递归，直至剩余的特征数量达到所需的数量。将剩余特征按照 w_1 排序即得到特征选择的最终结果。

3.2.3 LR

LR 是统计学中一种经典的分类算法，对回归或分类问题建立代价函数并迭代优化，求解出最优参数，该参数即特征的权值。具体步骤如下：

将线性回归函数带入 Sigmoid 函数，得到的 h 函数，若 $h_\theta(x) > 0.5$，则 $Y \in A$；若 $h_\theta(x) < 0.5$，则 $Y \in B$。

线性回归函数，式(11)：

$$z = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \cdots + \theta_n X_n = \theta^T X. \tag{11}$$

Sigmoid 函数，式(12)：

$$g(z) = \frac{1}{1 + e^{-z}}. \tag{12}$$

h 函数，式(13)：

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T X}}. \tag{13}$$

然后构造代价函数 $C(\theta)$，$C(\theta)$ 能够描述模型预测值 $h(\theta)$ 和真实值 y 之间的差异。若有多个样本，则取所有代价函数的均值，计作 $J(\theta)$。该均值 $J(\theta)$ 可用于评价该模型的好坏。$J(\theta)$ 越小，则当前模型的参数与训练样本越相符。于是基于最大似然估计可得 $J(\theta)$，式(14)：

$$J(\theta) = -\frac{1}{m} \left[\sum_{i=1}^{m} \left(y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log(1 - h_0(x^{(i)})) \right) \right], \tag{14}$$

基础梯度下降法求 $J(\theta)$ 最小值，更新参数，得到最符合当前数据的模型，式(15)：

$$\theta_j = \theta_j - \alpha \left(\frac{\partial y}{\partial \theta_j} \right) J(\theta) = \alpha \left(\frac{1}{m} \right) \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right) x_j^{(i)}. \tag{15}$$

特征对应的系数 θ_j 越大代表对期望的贡献越大，该系数也就是特征的权值。将系数 θ_j 从大到小排序，获得根据重要程度排序的特征。

3.3 筛选关键特征

在通过特征选择方法获得特征排序结果的基础上，本文利用朴素贝叶斯，SVM 和基于决策树的 Bagging 3 种分类器来检验不同特征组合的分类效

果，得到影响文献是否被引用的关键特征。

朴素贝叶斯是一种基于概率的分类器算法，其假设每个输入变量是独立的，根据训练集中每个特征的取值是否被引的先验概率，推算出测试集中特征给定时被引的后验概率，决定该元组是否被引。本实验中使用的是高斯朴素贝叶斯模型，假定数据符合高斯分布。

SVM 是一种二分类算法，可以支持线性和非线性的分类，其把划分数据的决策平面统称为超平面。离这个超平面最近的点叫支持向量，点到平面的距离叫间隔，通过在特征空间中寻找最佳的分离超平面，从而使训练集中正样本和负样本的间隔最大。利用该最优超平面，将文献集输入模型后即可得到合适的引文集并推荐给目标文献。本实验使用线性核函数的 SVM 并进行概率估计。

Bagging 是一种基于决策树的分类器，它是一种并行的集成学习方法，使用多棵树进行训练和预测，并结合训练结果输出预测值。本实验中使用决策树分类器，考虑到该分类问题为二分问题，构建 9 棵决策树进行投票，在避免过拟合的情况下尽可能收缩，使最终结果趋于均值。

4　实验过程和结果分析

4.1　数据集

本实验的原始数据均来自 Scopus 数据库。Scopus 收录了来自于全球 4 000 家出版社的19 000 种来源期刊，是全球最大的文摘和引文数据库，为科研人员提供一站式获取科技文献的平台。本文以科学计量学领域下的国际顶级期刊 Scientometrics 为文献样本来源，来获取目标文献集合。

数据的获取为利用爬虫算法在 Python3.7 环境下，爬取 Scopus 数据库中期刊 Scientometrics 中发表时间为 2018 年且被引频次排名前 100 的文献作为目标文献集合 P。收集 100 篇目标文献 P 的参考文献共 4 250 篇，将标题、作者、摘要和 DOI 为空的文献删除，剩余的 3 555 篇文献作为被引文献 B。按照 1:4 的比例收取与被引文献 B 在同一期刊、同一年份发表的且未被目标文献 P 引用的文献 N。被引文献 B 和未被引文献 N 共同构成待推荐的文献集 R。

数据的处理分为对目标文献集 P 的处理，以及对待推荐的文献集 R 的处理，处理步骤如下：

(1) 目标文献

①从 Scopus 数据库上直接导出文献的标题、作者、作者 ID、摘要、来源出版物、发表时间、施引文献数量、作者关键字以及在 Scopus 上的链接、文献的 EID 号和 DOI 号；

②在 Scopus 数据库上手工收集每篇目标文献 p 的每个作者之前写过的所有文献、每个作者的之前的合著者、每个作者引用过的文献以及每个作者引用过的作者；

③利用爬虫爬取每篇目标文献 p 的主题、学科和国家。

(2) 推荐的文献

①从 Scopus 数据库上直接导出文献的标题、作者、作者 ID、摘要、来源出版物、发表时间、施引文献数量、作者关键字、出资详情以及在 Scopus 上的链接、文献的 EID 号和 DOI 号；

②利用爬虫爬取待推荐文献 r 的常用的科学计量学特征和文献活跃度特征，利用程序判断待推荐文献 r 和对应目标文献 p 的关系，获取作者偏好特征。

4.2　实验过程

首先，利用 Relief-F、RFE、LR 3 个特征选择算法分别对实验收集的 19 个特征进行特征排序；其次，选取每个方法排名前 10 的特征完成进一步实验。对于某一种特征组合 $\{x_i\}$，$(i = 1,2,\cdots,10)$，取一篇种子文献 p_i，p_i 作为目标文献，p_i 的待推荐文献集 R_A 作为测试集，其余 99 篇种子文献的待推荐文献集 R_B 作为训练集。将训练集 R_B 放入分类器进行训练后，输入测试集 R_A，通过比较分类器对测试集 R_A 的预测结果和目标文献 p_i 实际引用情况的吻合程度，衡量分类效果。求取 3 个分类器分别输出的 F1 的均值作为该特征组合 $\{x_i\}$ 对该篇目标文献 p_i 的分类效果值。对 100 篇种子文献都重复以上步骤后，将获得的 100 个 F1 值求取均值，来代表该特征组合 $\{x_i\}$ 对本实验数据集的分类效果值。

按照上述实验思路，逐个去掉每个特征选择中得分最低的特征，输入到 3 个不同的分类器模型中，得出新的子特征组合对应的 F1 均值。提取 F1 均值最高时对应的特征子集为最终的约简子集。将在 3 种分类器下得到的约简子集取交集运算，即可得最终筛选出的特征。

4.3　评价指标

为评价本文提出的算法在引文推荐任务中的有效性，本文采用准确率 P，召回率 R 和 $F1$ 值来衡量推荐列表的质量。公式中符号的具体说明见表 3。

表 3 评价指标

Tab. 3 Evaluation indexes

	实际被引	实际未被引
预测引用	TP	FP
预测不引用	FN	TN

准确率是指分类正确的文献在文献总数中的占比,是对推荐系统查准率的衡量。在本文中,即被正确分类的待推荐文献与总的待推荐文献的比值,式(16):

$$P = \frac{TP + TN}{\mathrm{TP} + FP + FN + TN}. \tag{16}$$

召回率指的是正确推荐给目标文献 p 的引文与其实际引用的比率,是对推荐系统查全率的衡量,式(17):

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{17}$$

由于准确率与召回率有时候会出现相矛盾的情况,故引入衡量指标 $F1$ 值对二者进行综合考虑,式(18):

$$F1 = \frac{2PR}{P + R}. \tag{18}$$

4.4 推荐结果及分析

4.4.1 重要特征的选择

表 4 列出了 3 种特征选择算法下选出的前 10 个特征,可以看出由近期引用状况特征所确定的文献的活跃度特征,在 3 种方法中的排名均比较靠前,说明文献活跃度的特征有助于提升推荐效果;在常用的科学计量学特征中,主题和标题的相似度具有更大的优势;作者偏好特征中,大部分的特征排名都靠前,说明作者的兴趣对推荐具有较大的影响。

表 4 特征选择的结果

Tab. 4 Result of feature selection

特征选择方法	特征选择前 10 结果
RELIEF-F	$\{x_1, x_9, x_8, x_{10}, x_{18}, x_{19}, x_2, x_4, x_{16}, x_{17}\}$
RFE	$\{x_{19}, x_{10}, x_1, x_8, x_6, x_2, x_4, x_5, x_{17}, x_3\}$
LR	$\{x_{19}, x_{10}, x_8, x_4, x_6, x_1, x_2, x_5, x_{17}, x_3\}$

为了得到对文献是否被引具有重要影响的特征,在由每个特征选择方法得到的特征排序结果中,本文依次去掉权重得分最低的特征,将剩下的特征集合放入分类器中,记录分类的精度,循环进行,直到分类器的精度下降,取此时在特征集中剩余的特征为选出的特征约简子集。在分类器 Relief-F、RFE 和 LR 下分类精度的变化趋势如图 2~4 所示,按照此过程选出的特征约简子集的结果见表 5。

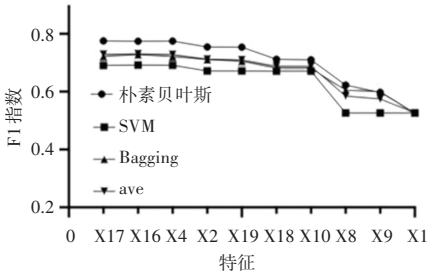

图 2 Relief-F 方法下 F1 值变化趋势图

Fig. 2 Change trend diagram of F1 index under Relief-F method

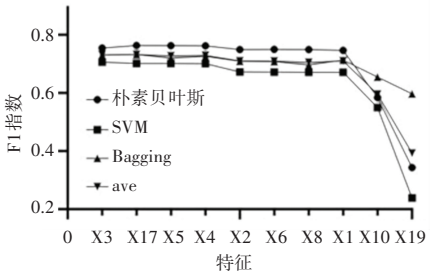

图 3 RFE 方法下 F1 值化趋势图

Fig. 3 Change trend diagram of F1 index under RFE method

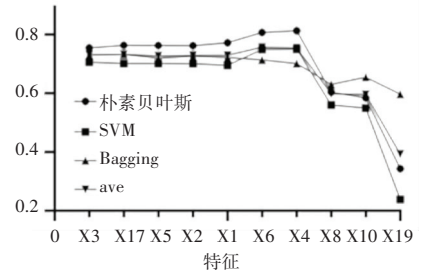

图 4 LR 方法下 F1 值变化趋势图

Fig. 4 Change trend diagram of F1 index under LR method

表 5 特征选择结果

Tab. 5 Results of feature selection

特征选择方法	精度下降前的特征选择结果
Relief-F	$\{x_1, x_9, x_8, x_{10}, x_{18}, x_{19}, x_2, x_4, x_{16}\}$
RFE	$\{x_{19}, x_{10}, x_1, x_8, x_6, x_2, x_4, x_5, x_{17}\}$
LR	$\{x_{19}, x_{10}, x_8, x_4, x_6\}$

不同的特征选择算法侧重点各异,单个特征选择方法选出的特征具有局限性,因此,对 3 个特征选择算法所得到的约简子集取交集运算,以得到在不同的特征选择算法下都比较重要的特征。这些特征,将是影响文献是否被引用的最核心的指标,得到的结果见表 6。

表 6　最终选择的特征结果

Tab. 6　The final selection of feature results

序号	特征
x_1	是否作者之前引用过的文献
x_2	是否为作者之前的文献
x_4	文献作者是否作者之前引用过的作者
x_6	作者关键字相似度
x_8	标题相似性
x_{10}	主题相似性
x_{19}	文献的活跃度特征

4.4.2　利用分类器实现推荐

将选出来的 7 个特征放入分类器，验证基于这些特征的引文推荐效果。本文将推荐问题转化为二元分类问题，对每篇目标文献 p，生成一个按照被推荐概率排序的推荐文献列表 l，将推荐结果 l 与每篇目标文献 p 的实际引用进行比较，算出相应的得分。同时与仅考虑文本相似度，利用标题相似度和主题相似度进行推荐的结果作对比见表 7。可以看出，相对于基线方法，利用本文提取出的 7 个核心特征进行是否被引用的识别，其准确率、召回率和 $F1$ 值分别提升了 6%、29% 和 26%，由此证明了这些特征是影响文献是否被引，实际上也是文献是否应该被推荐的关键指标。

表 7　分类器实现推荐的结果

Tab. 7　Results of classifier implementation recommendations

		准确率	召回率	$F1$ 值
本文采用方法	朴素贝叶斯	0.89	0.7	0.76
	SVM	0.87	0.63	0.7
	Bagging	0.85	0.66	0.73
	均值	0.87	0.66	0.73
基于基线方法	朴素贝叶斯	0.81	0.39	0.48
	SVM	0.81	0.36	0.46
	Bagging	0.81	0.37	0.48
	均值	0.81	0.37	0.47

相对于之前的研究工作而言，本文用较少的非常容易获取的特征较好地实现了引文推荐的工作，这对研究者开展实际的引文推荐研究具有重要的价值。在这些特征中，本文引入的文献的活跃度特征在引文推荐过程中起到了非常重要的作用，这实际上反映了引用过程中的“优先链接”的思想，说明那些在近期内得到较高引用的文章将具有更高的被再次引用的可能性。

5　结束语

本文将引文推荐问题转换为文献是否被引的二元分类问题，提取表征文献活跃度的特征，结合研究者的个性化引用偏好和常用的文献计量学特征，构建用以二元分类问题的特征库。利用 Relief-F、RFE 和 LR 特征选择方法从特征库中提取有利于文献被引用的关键特征，并基于这些特征利用朴素贝叶斯、SVM 和 Bagging 分类器实现引文推荐。本文的实验结果表明，文献的近期活跃度特性、作者的个性化引用偏好和文献对间的主题相似性是影响文献是否被推荐的核心因素。本文通过较为精简的特征实现了较好的引文推荐工作，这将对研究者开展实际的引文推荐研究提供有价值的参考。

参考文献

[1] 中国科技论文统计与分析课题组. 2018 年中国科技论文统计与分析简报[J]. 中国科技期刊研究，2020，31(1)：88-98.

[2] BASU C，HIRSH H，COHEN W W，et al. Technical paper recommendation：A study in combining multiple information sources[J]. Journal of Artificial Intelligence Research，2001，14：241-262.

[3] STROHMAN T，CROFT W B，JENSEN D. Recommending citations for academic papers[C]// International Acm Sigir Conference on Research & Development in Information Retrieval. ACM，2007：705-706

[4] BETHARD S，JURAFSKY D，Who should i Cite?：Learning literature search models from citation behavior[J]. International Conference on Information and Knowledge Management，Proceedings，2007，609-617

[5] HE Q，PEI J，KIFER D，et al. Context-aware citation recommendation.[J]. Proceedings of the 19th International Conference on World Wide Web，WWW ′10.2010.421-430.

[6] POHL S，RADLINSKI F，JOACHIMS T. Recommending related papers based on digital library access records[C]//Proceedings of the ACM International Conference on Digital Libraries，2007：417-418

[7] 刘盛博，丁堃，刘则渊. 基于引用内容的引文检索与推荐系统[J]. 情报学报，2013，32(11)：1157-1163.

[8] LIU Ya’ning，YAN R，YAN H. Guess What You Will Cite：Personalized Citation Recommendation Based on Users′ Preference[J]. 2013：428-439.

[9] 蔡阿妮. 基于内容与引用关系的学术论文推荐[D]. 华东师范大学，2014.

[10] 王萌星. 基于社区和引用网络的学术推荐关键技术研究与实现[D]. 北京邮电大学，2014.

[11] 刘亚宁，严睿，闫宏飞. 基于用户偏好与语言模型的个性化引文推荐[J]. 中文信息学报，2016，30(2)：128-135.

[12] GUO LT，CAI X Y，HAO F，et al. Exploiting Fine-Grained Co-Authorship for Personalized Citation Recommendation[J]. IEEE ACCESS，2017，5：12714-12725

[13] ALI Z，KEFALAS P，MUHAMMAD K，et al. Deep learning in citation recommendation models survey[J]. Expert Systems with

Applications, 2020, 162:113790.
[14]刘洋. 基于属性网络表示学习的引文推荐问题研究[D]. 安徽大学,2020.
[15]WANG J, ZHU L, DAI T, et al. Deep Memory Network with Bi-LSTM for Personalized Context-aware Citation Recommendation [J]. Neurocomputing, 2020, 410:101-113
[16] MCNEE S M, ALBERT I, COSLEY D, et al. On the recommending of citations for research papers[C]//Proceedings of the 2002ACM conference on computer supported cooperative work, ACM, 2002:116-125.
[17] TANG J, ZHANG J. A Discriminative Approach to Topic-Based Citation Recommendation [C]//Advances in Knowledge Discovery and Data Mining, 13th Pacific - Asia Conference, PAKDD 2009, Bangkok, Thailand, April 27 - 30, 2009, Proceedings. Springer-Verlag, 2009:572-579.
[18]CHOOCHAIWATTANA W. Usage of tagging for research paper recommendation [C]//InternationalConference on Advanced Computer Theory & Engineering, 2010:V2439-V2442
[19]倪卫杰. 基于用户兴趣模型的个性化论文推荐系统研究[D]. 天津大学,2010.
[20]WANG Y, LIU J, DONG X L, et al. Personalized Paper [8] Personalized paper recommendation Based on User Historical Behavior[C]//Springer Berlin Heidelberg, 2012:1-12.
[21] BEEL J, GIPP B, LANGER S, et al. Research - paper recommender systems: a literature survey. International Journal on Digital Libraries, 2016,17(4), 305-338. DOI:10.1007/s00799-015-0156-0
[22]陈俊鹏. 基于梯度渐进回归树的引文推荐方法研究[D]. 北京理工大学,2016.
[23]DAI T, ZHU L, CAI X, et al. Explore semantic topics and author communities for citation recommendation in bipartite bibliographic network. Journal of Ambient Intelligence and Humanized Computing,2018, 9(4), 957-975. DOI:10.1007/s12652-017-0497-1
[24]KHADKA A, KNOTH P. Using citation-context to reduce topic drifting on pure citation-based recommendation. In Proceedings of the 12th ACM conference on recommender systems, 2018:362-366.
[25]ZHANG Y, MA Q. Citation recommendations considering content and structural context embedding. arXiv:2001.02344:1-7
[26]GORI M, MAGGINI M, SARTI L. Exact and approximate graph matching using random walks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(7), 1100-1111. DOI:10.1109/TPAMI.2005.138
[27] MENG, Z, SHEN H, HUANG H, et al. Search result diversification on attributed networks via nonnegative matrix factorization. Information Processing and Management, 2018, 54(6), 1277-1291. DOI:10.1016/j.ipm.2018.05.005:1277-1291
[28] JARDINE J, TEUFEL S. Topical PageRank: A model of scientific expertise for bibliographic search. Paper presented at the 14th Conference of the European Chapter of the Association for Computational Linguistics 2014, EACL 2014:501-510
[29] Cai X, Han J, Li W, et al. A Three-Layered Mutually Reinforced Model for Personalized Citation Recommendation. IEEE Transactions on Neural Networks and Learning Systems, 2018,29(12), 6026-6037. DOI:10.1109/TNNLS.2018.2817245
[30] PAN L, DAI X, HUANG S, et al. Academic paper recommendation based on heterogeneous graph. In: Vol. 9427. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),2015: 381-392.
[31]PRADHAN T, GUPTA A, PAL S. HASVRec: A modularized Hierarchical Attention - based Scholarly Venue Recommender system. Knowledge-Based Systems, 204. DOI:10.1016/j.knosys.2020.106181
[32]李飞. 基于文本向量表示学习的引文推荐方法研究[D]. 西北农林科技大学,2018.
[33]陈洁, 刘洋, 赵姝, 张燕平. 利用多粒度属性网络表示学习进行引文推荐 [J/OL]. 计算机科学与探索:1-13[2020-12-17].

(上接第 133 页)

[2] 王荣本,李兵,施树明,等. 世界智能车辆研究概述[J]. 公路交通科技,2001(5):93-97.
[3] 韩俊淑 ,韩佳文,高翔,等. 智能车辆的研究及发展[J]. 世界汽车,2003(9):79-80.
[4] 郭晋昌,徐鹏. 基于 K60 的遥控巡逻小车设计、实现及试验[J]. 陇东学院学报,2019,30(2):16-19.
[5] 于少东,黄丹平,田建平,等. 基于 Kinetis K60 的智能车控制系统设计[J]. 四川理工学院学报(自然科学版),2014,27(5):37-42.
[6] 徐彦钦,石子昊,夏佳宁. 基于 ESP8266 智能空调控制系统的设计[J]. 信息与电脑(理论版),2018(9):82-83.
[7]李征文. 基于路灯改建充电桩系统的设计研究[D]. 大连理工大学,2018.
[8] 赵风财,肖广兵.基于树莓派的四轮独立电驱动监控系统设计[J]. 软件,2020,41(8):78-82.
[9]林家泉,程绪宇,周贤民,等. 一种小型直流电机控制系统硬件设计方案[J]. 自动化与仪表,2014,29(11):73-76.
[10]熊云龙,李志扬. 搭载激光雷达的智能小车系统开发[J]. 科技与创新,2020(5):35-37.
[11]金展. 对于导航机器人关于 SLAM 实现的研究[J]. 科技与创新,2019(19):142-143.
[12]向亚军,严华. 基于激光雷达的移动机器人避障策略研究[J]. 四川大学学报(自然科学版),2017,54(3):529-534.
[13]陈文澄,张辉,张晋滔. ESP8266 Wi-Fi 模块在智能小车控制中的应用[J]. 工业控制计算机,2019,32(7):134-136.