

# 基于学术论文全文内容的算法 使用行为及其影响力研究

章成志<sup>1,2,3</sup>, 丁睿祎<sup>1,3</sup>, 王玉琢<sup>1,3</sup>

(1. 南京理工大学经济管理学院信息管理系, 南京 210094; 2. 江苏省数据工程与知识服务重点实验室  
(南京大学), 南京 210093; 3. 江苏省社会公共安全科技协同创新中心, 南京 210094)

**摘要** 数据挖掘算法已被广泛应用于科学研究与实践中。考察数据挖掘算法在学术论文中的使用情况、进而评估其影响力, 能辅助研究者全面了解其所在领域的常用算法, 并根据研究任务类型选择相应算法。本文利用学术论文全文内容, 对算法的使用行为进行分析, 从而考察算法的影响力。具体来说, 本文以自然语言处理领域为例, 收集整理全国计算语言学会议(CCL)1993—2016年收录的学术论文全文数据, 从使用频次、使用位置、使用年代以及使用动机等四个方面全面考察十大经典数据挖掘算法在该领域的使用情况, 并在此基础上对算法的影响力进行评估。实验结果显示, 十大算法的使用行为存在明显差异, 且SVM算法影响力最高, CART与Apriori算法影响力较低。本文研究可为基于数据驱动的相关研究者, 尤其是为初学者在算法选择时提供参考。

**关键词** 算法影响力评估; 使用行为; 全文内容分析

## Using Behavior and Influence Assessment of Algorithms Based on Full-text Academic Articles

Zhang Chengzhi<sup>1,2,3</sup>, Ding Ruiyi<sup>1,3</sup> and Wang Yuzhuo<sup>1,3</sup>

(1. Department of Information Management, Nanjing University of Science & Technology, Nanjing 210094;  
2. Jiangsu Key Laboratory of Data Engineering and Knowledge Service (Nanjing University), Nanjing 210093;  
3. Jiangsu Collaborative Innovation Center of Social Safety Science and Technology, Nanjing 210094)

**Abstract:** Data mining algorithms have been widely used in scientific research and practice. Investigating the mentions of data mining algorithms in academic papers and assessing their influence can help researchers comprehensively understand algorithms used in their field and select those that are appropriate based on a given research task. We used full-text of academic articles to conduct an analysis using the behavior of algorithms and evaluating their influence. This paper considers the field of natural language processing and collects the full-text proceedings accumulated by the *China National Conference on Computing Linguistics* (CCL) from 1993 to 2016 to conduct a comprehensive examination of the top 10 data mining algorithms based on four aspects: frequency of mention, location of mention, motivation for mention, and age distribution. The impacts of the algorithms are evaluated according to these four aspects. The experimental results show that obvious differences exist among the 10 algorithms; the SVM algorithm has the highest influence, while the CART and Apriori algorithms have low influence. This investigation can

收稿日期: 2018-07-04; 修回日期: 2018-10-15

基金项目: 国家社会科学基金重大项目“情报学学科建设与情报工作未来发展路径研究”(17ZDA291)。

作者简介: 章成志, 男, 1977年生, 博士, 教授, 博士生导师, 主要研究领域包括信息组织、信息检索、数据挖掘及自然语言处理, E-mail: zhangcz@njust.edu.cn; 丁睿祎, 女, 1995年生, 硕士研究生, 主要研究方向为信息检索与数据挖掘; 王玉琢, 女, 1995年生, 博士研究生, 主要研究方向为文本挖掘与科学计量。

provide recommendations for researchers, especially novices whose works are related to data-driven research or applications, as well as introduce new ideas toward the assessment of algorithm influence.

**Key words:** influence assessment of algorithms; using behavior; full-text context analysis

## 1 引言

数据挖掘算法被广泛用于科学研究与实践。尤其在当前的大数据环境下,数据驱动型的研究与应用更离不开数据挖掘算法。相关学科领域的科研工作者,尤其是初学者希望快速了解其所在学科领域的常用数据挖掘算法,并根据研究任务有效选择合适的算法<sup>[1]</sup>,从而提高科研效率。研究数据挖掘算法在特定专业领域的使用行为,考察算法在特定领域的实际应用情况,从而评估算法的影响力,对于研究与实践者评估或选择算法等都具有重要的参考意义。

现有的算法影响力评估主要考察算法的使用情况,即:针对特定的研究目标,选择相关数据集,使用不同算法进行实验,依据实验结果判定算法的性能,最终评估算法的影响力<sup>[2-3]</sup>。这种基于使用的算法影响力评估方法具有直接、准确的优点,但有一定的局限性,首先是算法评估依据特定的数据集,然而,现实中并非每个研究问题都有可用的数据集;其次是该方法要求评估者具有较高的专业知识水平,初学者若通过该方法了解算法的性能,实际难度相对较大。因此,我们需要用更具普适性、相对简单的方法来评估算法。

当前,绝大部分的学者在进行数据驱动研究时都会使用相关的数据挖掘算法,并在学术论文中描述这些算法的使用情况。依据学术论文,考察算法在论文全文中的使用频次、使用位置以及使用动机,能反映算法在相关研究领域的整体使用情况,并以此为基础进行算法的影响力评估。已有学者利用这种方法对开源软件或数据集进行评估,但绝大多数研究仅考虑开源软件或数据集在学术论文中的使用频次而忽略使用位置、使用动机,无法深入和全面地了解算法的使用行为,进而无法综合评估算法影响力。随着全文数据库的开放程度越来越高、学术论文全文数据越来越容易获取,以学术论文全文内容为基础的算法评估成为可能,并且普适性更高。因此,本文考虑从学术论文全文内容的角度对算法影响力进行评估。

在被人们广泛使用的数据挖掘算法当中,公认的十大数据挖掘算法于2006年12月在IEEE国际数据挖掘会议(IEEE International Conference on Data

Mining, ICDM)上经由专家投票评选得出<sup>[4]</sup>。这十大算法被广泛应用于数据驱动型的学术研究与实践。本研究以自然语言处理领域为例,从全文内容分析的角度,考察十大数据挖掘算法在该领域学术论文中的使用情况,并以此为基础评价十大算法的影响力。本研究丰富了学术影响力评价的内容,扩充了全文内容方法的应用领域,所用研究方法可为微观实体的科学评价相关研究提供新的思路;所得研究结果可为基于数据驱动的相关研究者,尤其是为初学者在算法选择时提供参考。

本文第2节为相关研究工作概述,第3节详细描述本文所用的研究方法,第4节对所得结果进行具体分析,第5节为对部分结果的讨论与本研究的创新性和局限性,最后给出本文的研究结论与未来工作展望。

## 2 相关工作概述

算法是一种微观实体<sup>[5]</sup>。本文从全文内容分析的角度,研究算法的使用行为,并在此基础上评估算法的影响力。与本文相关的研究工作包括微观实体评估、引文内容分析以及引用动机分类研究,下面对这三个方面的相关工作进行概述。

### 2.1 微观实体评估研究概述

微观实体主要包括数据集、方法、命名实体等<sup>[5-6]</sup>。对微观实体进行评估,能够了解其在特定领域的分布或使用情况,为相关研究人员提供参考。目前,微观实体的评估研究主要针对数据集、软件、算法等研究对象。

对数据集进行评价,能促进数据集的共享与利用。丁楠等<sup>[7]</sup>依据数据发布量、数据被引量、数据平均被引频次以及h指数等指标,构建基于引用的数据评价体系,并以Web of Science的DCI数据库中的人口调查数据为例进行实证研究。Belter<sup>[8]</sup>以海洋学领域的三个数据集为研究对象,依据被引次数研究数据集引用行为,并在此基础上进行数据集的评估。王雪等<sup>[9]</sup>以生物信息学领域论文中数据集被引频次、下载量等为指标,研究科学数据的引用行为及其影响力。

近年来,越来越多的学者开始关注开源软件的

评价。Pan 等<sup>[10]</sup>考察学术论文中 CiteSpace、HistCite 以及 VOSviewer 等三种软件的提及与引用情况，依据软件在文章、期刊、学科等层面上的扩散深度与速度，评估软件的影响力。赵蓉英等<sup>[11]</sup>借助 Python 社区中软件的下载量、文献被引次数以及软件复用次数三个指标，对开源软件的学术影响力进行评估。杨波等<sup>[12]</sup>对生物信息学领域论文中的科学软件利用行为进行分析，并在此基础上提出相关指标来度量软件的质量与影响力。

此外，算法评价问题也是微观实体评价中的一个重要问题。常规的评估方法是在给定数据集基础上选择不同算法进行实验分析与测评，如 Wilbanks 等<sup>[2]</sup>采用“先运用再对比”的方法对用于生物学领域处理 ChIP-seq 峰值的 11 种算法的敏感性、可用性以及准确性进行评估；Nesma 等<sup>[3]</sup>用 KEEL 软件分别测试 10 种最常见的算法，在分类任务上进行比较。王玉琢等<sup>[1,13]</sup>以国际计算语言学会（ACL）论文为研究对象，依据提及频次、提及位置等指标对算法的影响力进行评估。

## 2.2 引文内容分析研究概述

引文内容分析是全文内容分析的一部分，本文的研究借鉴引文内容分析中的相关方法。因此，本节对引文内容分析的相关研究进行概述。

引文内容分析是依据学术全文内容的引文分析<sup>[14]</sup>。早期由于学术论文全文难以获取，引文内容分析相关的研究很难大规模开展。随着全文数据库越来越多地向用户免费开放，以大规模学术论文全文数据为基础的引文内容分析研究成为当前的研究热点之一。Ding 等<sup>[15]</sup>认为引文内容分析是引文分析的下一代发展方向，引文内容分析应包含语法与语义两个层次。胡志刚<sup>[16]</sup>将引文内容分析分为引用位置分析、引用强度分析和引用语境分析三个研究角度。王文娟等<sup>[17]</sup>提出全文引文文本可归类为基于引用功能的引文、基于情感倾向的引文、基于引文影响力的引文等。

目前，引用内容分析相关应用研究包括信息检索、知识图谱构建、科学计量与学术评价等。刘盛博<sup>[18]</sup>利用 PubMed Central 数据库中的所有全文数据，构建一个基于引文内容的引文检索系统，并将其与 Google Scholar、PubMed 做比较，发现采用引文内容中的主题词进行引文检索更具优势。李婷婷等<sup>[19]</sup>利用引文内容来分析信息学期刊互引。An 等<sup>[20]</sup>利用引文位置与引文内容构建作者网络，研究高被引作者的学术特征，发现高被引作者的学术特征与

其文献的被引位置有关。Hassan 等<sup>[21]</sup>使用机器学习方法，借助引文上下文特征来区分重要引用和非重要引用。

## 2.3 引用动机研究概述

目前还鲜有专门针对算法或开源软件等微观实体使用动机的分类体系。算法等微观实体的使用动机与一般文献的引用动机在某些方面存在类似之处。

引用动机即施加引用的目的。学界对引用动机的研究开展的较早，且大部分研究的是文献的引用动机。Garfield<sup>[22]</sup>通过人工归纳，将引用动机较为系统地划分为 15 类，包括向先驱者表示敬意、提供阅读背景资料、证实自己的论点、否认他人的作品或观点等。Weinstock<sup>[23]</sup>指出科研人员的引用动机可分为向前人表示敬意、批评他人的著作、通报未来的工作等 15 种。Brooks<sup>[24]</sup>采访了 26 位来自爱荷华大学的老师和 1 位服务部门教员，将作者的引用动机分为 7 种，即：说服引用、肯定引用、时效性引用、提示引用、概念引用、共识引用和否定引用。Erikson 等<sup>[25]</sup>整合以往基于访谈法的引用动机研究成果，将引用动机分为论证动机、社会联系动机、利益结盟、数据功能这四大类。Moravcsik 等<sup>[26]</sup>将引用动机归纳为有机引用与敷衍引用、概念引用与实际应用、进化引用与并列引用、肯定引用与否定引用四个类别。Teufel 等<sup>[27]</sup>提出一种四大类的引用动机分类系统，包括指出当前工作的不足、自己与他人研究的对比、带有积极情感的引用、完全中立态度的引用或其他引用。Jurgens 等<sup>[28]</sup>参考前人工作，将引用动机分为介绍背景、作为数据或方法等的使用依据、使用数据或方法、扩展数据或方法、比较相似性或差异以及展望未来工作六大类。邱均平等<sup>[29]</sup>提出针对科研人员引用行为的影响因素模型，将引用动机归并为内在引用和外在引用两大类。

综上所述，为提高研究结果的准确性与全面性，本研究采用基于全文内容的方法，从使用频次、使用位置、使用动机以及年份四个方面对十大算法在自然语言处理领域的使用情况进行考察，并在此基础上对其影响力进行评估。与本文工作较为相关的是王玉琢等<sup>[1,13]</sup>的工作。王玉琢等依据提及论文数、提及总次数、提及位置三个方面对十大数据挖掘算法的影响力进行评估，同时分析使用算法所要解决的具体任务<sup>[13]</sup>，但未研究算法的使用动机与算法的使用随着年代的变化情况，且未考虑“被提及算法”和“被使用算法”的影响力差异，故评估结果存在一定

的片面性。与该工作不同的是,本文在使用频次与使用位置的基础上,进一步考察算法的使用动机以及年代分布,详细分析文章为什么使用算法以及算法随时间的变化规律,更加深入地对算法使用行为进行研究。此外,本文将论文的章节类型划分为 7 类,并重点就其中的“method”与“evaluation”2 个章节类型中的算法使用做详细考察,结果更为全面。

### 3 研究方法

#### 3.1 基本思路

如图 1 所示,本研究以十大数据挖掘算法为研究对象,以自然语言处理(NLP)领域为例,采集 NLP 领域学术论文全文,结合所构建的十大数据挖掘算法词典,识别并抽取出算法句(即论文中提及算法的句子)。在此基础上,考察十大数据挖掘算法在 NLP 领域学术全文中的使用情况,具体而言,本文从使用频次、使用位置、使用年代以及使用动机

四个方面对十大数据挖掘算法的使用行为以及影响力进行分析。

#### 3.2 语料获取与数据标注

##### 3.2.1 原始语料获取与预处理

十大数据挖掘算法广泛地应用于各个学术领域中。自然语言处理(NLP)是一个以数据和技术为核心的研究领域,大多数学者都需要借助数据挖掘算法或开源工具来完成相关研究任务。本文研究数据为 NLP 领域的学术论文全文,来自于全国计算语言学会(CCL)的历年论文数据。CCL 着重于中国境内各类语言的计算处理,是国内自然语言处理领域权威性最高、规模最大的学术会议,并且公开了历届的会议论文全文数据集<sup>①</sup>。本文获取了 1993—2016 年的所有 CCL 论文(均为 PDF 格式),共 1767 篇,转换为纯文本后,为便于后续处理,对全文的题录与篇章信息进行人工标注,标注所使用的标签集合如表 1 所示。依据表 1 进行人工标注的结果样例如图 2 所示。

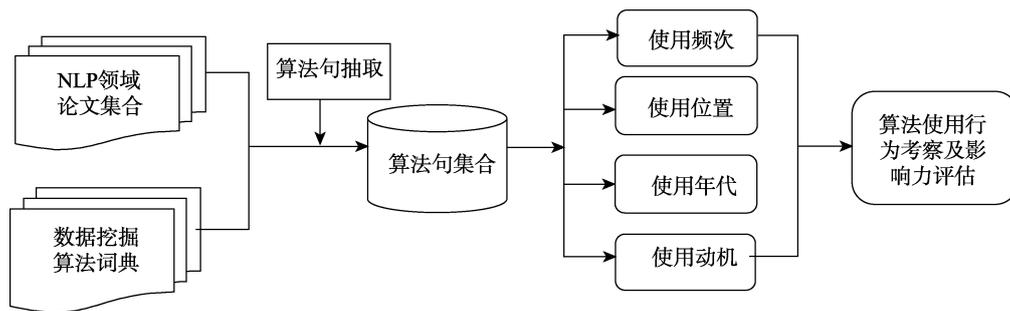


图 1 本文研究思路图

表 1 标注标签集合说明

标签	含义
<title_chinese>...</title_chinese>	中文标题
<title_english>...</title_english>	英文标题
<author_chinese>...</author_chinese>	中文作者名
<author_english>...</author_english>	英文作者名
<address_chinese>...</address_chinese>	中文地址
<address_english>...</address_english>	英文地址
<abstract_chinese>...</abstract_chinese>	中文摘要
<abstract_english>...</abstract_english>	英文摘要
<keyword_chinese>...</keyword_chinese>	中文关键词
<keyword_english>...</keyword_english>	英文关键词
<chapter>...</chapter>	章节
<chapter_title>...</chapter_title>	学术论文的 1 级标题
<chapter_sub_title>...</chapter_sub_title>	学术论文的 2 级标题
<para>...</para>	段落
<reference>...</reference>	参考文献
<section type=<Introduction Related work Method Evaluative Result Conclusion>...</section>	章节类型

① 历年文献下载于中国中文信息学会计算语言学专业委员会官方网站,网址: <http://www.cips-cl.org/webmirror>.

```

<title_chinese>组合中文词义消歧</title_chinese>
<author_chinese>秦颖 王小捷</author_chinese>
<address_chinese>北京邮电大学信息工程学院北京 100876</address_chinese>
<abstract_chinese><section=abstract>.....该组合分类器的性能优于基本 Naive Bayes 分类器和 Ensemble
Naive Bayes 分类器.....</abstract_chinese>
<keyword_chinese>词义消歧, 组合算法, 轨迹</keyword_chinese>
<title_english>Combining Classifier for Chinese Word Sense Disambiguation</title_english>
<author_english>Ying Qin Xiaojie Wang</author_english>
<address_english>School of Information Engineering, Beijing University of Posts and Telecommunications,
Beijing, 100876, China</address_english>
<Abstract_english><section=abstract>The experiment results show the new combining classifier not only
outperforms the Naive Bayesian.....</Abstract_english>
<keyword_english>word sense disambiguation, combining algorithm, trajectory</keyword_english>
<chapter id=1>
<chapter_title><section=introduction>1 引言</chapter_title>
<para> .....词义消歧是一个典型的分类问题, 各种有监督的分类算法如 Bayes 算法、决策列表法、
KNN 法、基于转换学习的方法等都已于用于词义消歧的研究。</para>
.....
</chapter>

```

图 2 标注结果示例<sup>①</sup>

### 3.2.2 数据挖掘算法词典的构建

本研究以 *The Top Ten Algorithms in Data Mining* 一书中的十大经典数据挖掘算法<sup>[4]</sup>为依据, 通过检索, 扩展得到每种算法名称的别名(包括简称、缩略词及中文名), 最后得到十大数据挖掘算法扩展词典, 具体如表 2 所示。

### 3.2.3 算法句的抽取

本文中的“算法句”是指一篇论文中提及某种算法的句子。算法句是考察算法使用情况和评估算法

影响力的重要载体。本文采用基于字典匹配的方法, 从学术论文的全文内容中抽取出具体的算法句。由于题录信息与参考文献中亦存在算法提及, 但对本文研究并无意义, 因此利用论文中所标注的标签将题录信息与参考文献剔除, 仅考虑摘要与正文中的算法提及。首先, 利用数据挖掘算法词典对论文各章节进行算法名匹配, 抽取算法句, 表 3 为抽取出的算法句样例; 然后, 记录算法句所在文章的编号、算法提及次数、算法句所在的章节类型等信息存入数据库; 最终, 得到提及十大数据挖掘算法

表 2 十大数据挖掘算法扩展词典

No.	标准名称	别名	中文名
1	C4.5	—	—
2	K-means	k means	k 均值
3	Support vector machines	support vector machine、svm、svms	支持向量机、支撑向量机
4	Apriori	—	—
5	EM	expectation-maximization、expectation maximization	最大期望算法、期望最大化算法
6	PageRank	PR	—
7	AdaBoost	Adaptive Boosting	—
8	K-nearest neighbor	knn、k-nn、k nearest neighbor、k nearest neighbour、k nearest neighbors、k nearest neighbours、k-nearest neighbors	K 最近邻、k 近邻
9	Naive Bayes	Naïve-bayes、naïve bayes、naïve-bayes、NB、Naive Bayesian	朴素贝叶斯
10	CART	classification and regression trees	—

① 该样本对应的原始文献为: 秦颖, 王小捷. 组合中文词义消歧[C]//全国第八届计算语言学联合学术会议, 南京, 中国, 2005: 127-133.

表 3 算法句抽取结果示例

文章编号	章节类型	算法名	提及次数	算法句
CCL2003-84	method	knn	1	“此外,常用的文本分类算法 KNN 算法,应用在话题跟踪上也有比较好的效果。”
CCL2005-20	result	knn	1	“Daelemans et al 指出 <sup>[21]</sup> ,在 KNN 算法中任何对于数据的编辑都是有害的。”
CCL2005-20	evaluation	naive bayes	1	“我们将轨迹法与 Pederson 的 Ensemble Naive Bayes 分类器 <sup>[9]</sup> 作了比较。”
CCL2005-22	introduction	svm	1	“SVM 的分类准确率最高达到约 80%,为几种方法中分类效果最好的。”

的算法句共 2328 条(涉及论文 352 篇)。

### 3.3 算法使用行为与影响力分析维度

本文通过使用频次、使用位置、使用年代以及使用动机等相关分析维度,具体考察十大算法在学术论文中的使用情况并评估其影响力。

#### 3.3.1 使用频次

使用频次即算法被提及的次数。本文将频次划分为两个指标:提及论文数、平均提及次数。提及论文数参考传统的 Count One 方法,指的是在一篇论文中只要提及某种算法,不论该算法被提及多少次,只记为 1 次;但很多情况下,一种算法在一篇论文中被提及不止一次,只考虑提及论文数过于片面。因此,本研究将平均提及次数列入考察指标。平均提及次数的计算如公式(1)所示。公式中的总提及次数参考 Ding 等<sup>[30]</sup>提出的 Count X 方法,指的是在一篇论文中,某种算法被提及多少次就记为多少次。

$$\text{平均提及次数} = \text{总提及次数} / \text{提及论文数} \quad (1)$$

#### 3.3.2 使用位置

使用位置即算法句所在的章节类型。在一篇学术论文中,算法句分布在不同的章节里。对于一篇学术论文而言,不同的章节重要性不同<sup>[31]</sup>。故而在不同章节被使用的算法影响力也不同。实证型研究论文通常由引言(introduction)、数据与方法(data and methods)、结果(results)、讨论(discussion)和结论(conclusions)这五个部分构成,即 IMRDC 结构<sup>[32]</sup>。结合 An 等<sup>[20]</sup>的研究并考虑到自然语言处理领域里中文论文的特点,本文将学术论文的章节划分为如表 4 所示的 7 种类型。考虑到并非所有论文都存在上述 7 种章节类型,故对于每一篇论文,其存在多少种章节类型,我们就标注多少种。此外,由于不同作者的行文风格各有差异,有些算法提及可能只出现在摘要中,为防止遗漏,本文将“abstract”也作为一种章节类型来进行研究。对已标注的论文数据进行处理,统计算法在不同位置的提及论文数,并基于此分析不同位置算法的使用情况。

表 4 章节类型分类

类型	abstract	introduction	related work	method	evaluation	result	conclusion
功能	摘要	引言	相关工作概述	研究方法描述	实验,通常包括评价方法、具体实验过程	实验结果与结果的深入讨论	结论与未来工作说明

#### 3.3.3 使用年代

本研究中,算法的使用年代即文章的发表年代。对使用年代进行分析,考察算法使用情况演变趋势,能够得出十大算法在自然语言处理领域的使用规律。在本文所使用的数据集中,每篇论文都有包含年份信息唯一的编号(如 CCL2003-73)。在算法句的抽取过程中,记录下文章编号作为每种算法使用的年代信息,在此基础上统计每一年各算法的提及论文数、平均提及论文数进行年代分布分析。

#### 3.3.4 使用动机

使用动机即提及算法的原因。目前还鲜有专门针对算法等实体使用动机的分类体系。算法的使用

与论文引用有类似之处,除使用频次、使用位置外,也有对应的使用动机。本文参考现有的有关引用动机的研究成果,结合算法的特性,以提及某种算法的施引文章为标注单位,提出针对数据挖掘算法的使用动机分类体系,如表 5 所示。

根据表 5 所示的分类体系,对提及算法的论文进行使用动机的人工标注。在一篇文章中,根据文中提及算法所在的算法句以及全文,判断文章提及该算法的动机。本文参考崔明等<sup>[33]</sup>对标注结果的检验方法,从 354 篇提及十大算法的论文中随机抽取 50 篇,分别由一名博士研究生与一名硕士研究生独立标注,经检验二者标注结果的 kappa 系数为 0.703,为高度一致,故标注结果可靠。

表 5 算法使用动机分类

动机类型	二级分类	说明	具体实例
背景提及	—	介绍算法的背景信息、解释算法原理、他人工作中使用	“支持向量机是性能良好的二类分类模型，适用于处理文本分类问题。”
实际应用	单一使用	在实验中使用了算法，且没有与其他算法进行对比	“这里，我们首次采用 SVM 进行中文分词的研究。”
	使用并比较	在实验中使用了算法，且与其他算法进行了比较	“采用 SVM 分类以实现检索目的的方法，取得了比直接用查询检索或者用朴素贝叶斯 (NB) 进行分类更好的效果。”
	改进使用	在实验中改进并使用了算法，且没有与其他算法进行对比	“本文提出一个通用框架，该框架通过半强制解码和变分贝叶斯 EM 对 SMT 模型进行剪枝和优化。”
	改进使用并比较	在实验中改进并使用了算法，且与其他算法进行对比	“实验表明，对日语依存关系训练集进行两次修剪后得到的分类器(MV-LSVM)相比于单纯使用 SVM 在解析精度和解析速度上都表现出了一定的优越性。”

## 4 结果分析

针对前文对数据的处理结果，本节将分别从使用频次、使用位置、使用年代、使用动机四个方面对十大数据挖掘算法的使用情况进行全面的考察，并在使用频次、使用位置以及使用年代的基础上对其影响力进行整体评估。

### 4.1 基于使用频次的算法使用行为及影响力分析

提及论文数是本研究设定的频次评估指标之一。本研究认为算法的提及论文数越高，表明该算法被更多数的学者所使用，故影响力越大。十大数据挖掘算法的提及论文数统计结果如表 6 所示。从表 6 中可以看出，提及论文数最高的为 SVM 算法，占提及算法论文总数的 70% 以上，且高于排名第二位的算法提及论文数的 3 倍。在 *The Top Ten Algorithms in Data Mining* 一书中也有说明，SVM 算法理论基础坚实，是所有已知著名算法中最稳定且最精确的算法之一<sup>[4]</sup>，故大量学者选择 SVM 算法来进行研究。排名垫底的是 CART 算法，仅仅只有 2 篇论文提及；Apriori 算法也只有 4 篇论文提及。由此可知，就提及论文数而言，SVM 是影响力最大的算法，而 CART 算法、Apriori 算法影响力较小。

此外，本文还在提及论文数的基础上统计算法的平均提及次数。对平均提及次数的考察，能够了解算法在单独一篇论文中的提及情况。具体统计结果如表 7 所示。比较表 6 和表 7 可以看出，十大算法的排名并不一致。这说明，算法的提及论文数越高，平均提及次数不一定越高。就平均提及次数而言，主要变化是：adaboost 算法从第八名跃至第一名，在提及论文数中位列第二名的 Naïve Bayes 算法跌至第七名。笔者认为 adaboost 的排名靠前原因可能是

该算法在自然语言处理领域使用较少，因此当有研究使用该算法时，需要花费大量的笔墨去描述解释其原理，从而在文章中反复提及；而 Naïve Bayes 算法排名下降的原因可能是该算法较其他算法而言，原理较为简单，在文章中不需要大篇幅的去解释。其他算法的平均提及次数与提及论文数的结果差距不大。综合提及论文数与平均提及次数结果可知，SVM 影响力最大，CART 算法、Apriori 算法影响力较小。

表 6 提及论文数结果

排名	算法名	提及文章数
1	SVM	247
2	Naïve Bayes	78
3	knn	51
4	PageRank	38
5	k-means	28
6	EM	14
7	C4.5	10
8	adaboost	10
9	apriori	4
10	CART	2

表 7 平均提及次数结果

排名	算法名	平均提及次数
1	adaboost	6.500
2	PageRank	6.105
3	SVM	5.842
4	EM	4.143
5	k-means	3.821
6	knn	3.549
7	Naïve Bayes	2.833
8	C4.5	1.400
9	apriori	1.250
10	CART	1.000

## 4.2 基于使用位置的算法使用行为及影响力分析

在使用频次的基础上,本研究还对算法在全文中出现的使用位置信息进行考察。由于所用数据集中包含中文和英文论文,考虑到重复统计的问题,对于有中英文摘要的中文论文,仅考虑中文摘要中的算法提及。首先,以十大算法在不同位置的提及论文数为指标,考察算法在一篇学术论文中的整体位置分布情况,结果如图3所示。

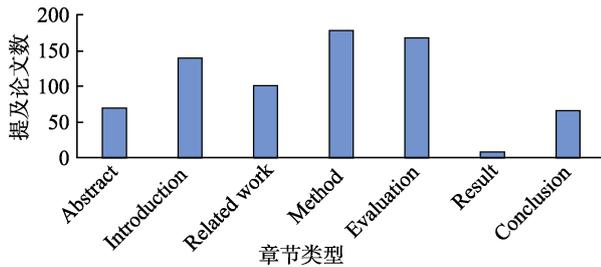


图3 全部十大算法在各章节分布情况

由图3可以看出,算法在不同位置中的提及论

文数最多的是“method”部分,最少的则为“result”部分。在一篇学术论文中,由于“abstract”部分的行文要求,必然会有较低程度的算法提及;同时,学者们需要对其研究中所要使用的算法做简单的背景介绍,因此在“introduction”和“related work”部分会有一定程度的提及;随后,算法作为处理数据时使用的具体方法,在“method”和“evaluation”提及次数显然会有增加;在结果分析部分,主要分析对象是实验所得结果,不需要再对算法的相关内容赘述,因此在“result”部分的算法提及有大幅度下降;最后在总结全文时,需要对研究过程进行总结,包括实验的大致流程与结果等,因此在“conclusion”部分算法提及的次数略有回升。在不同位置使用的算法,其影响力也有差异。本研究统计各算法在学术论文各个章节的提及论文数,结合整体分布情况,且考虑到自然语言处理领域学术论文的特点,以“method”和“evaluation”这两种类型的章节为重点来分析各算法在不同位置的使用情况。各算法在论文中的位置统计情况如表8所示。

表8 各个算法在各章节中的分布情况

章节类型	算法名									
	SVM	NB	KNN	PageRank	k-means	EM	adaboost	C4.5	Apriori	CART
abstract	47	8	6	4	3	0	2	1	2	1
introduction	99	27	15	12	7	2	6	1	1	0
related work	74	28	10	12	3	6	3	1	0	1
method	101	28	21	20	20	8	6	4	2	0
evaluation	113	26	20	14	11	4	4	5	0	0
result	4	3	3	0	1	0	0	0	0	0
conclusion	50	7	5	3	1	2	1	0	0	0

从表8中可以看出,在文章的不同章节中,算法出现的概率各不相同,各个算法在“method”与“evaluation”部分的提及次数普遍高于其他章节,其次是“introduction”与“related work”部分,在“abstract”、“result”与“conclusion”部分提及较少。这说明就每个算法而言,在自然语言处理领域的研究中主要还是作为具体的实验方法被使用。

此外,从“method”章节的统计结果来看,在自然语言处理领域,使用最多的算法仍然是SVM算法,其提及论文数远远高于其他算法;而CART和Apriori算法的使用很少,CART算法的提及论文数为0。从“evaluation”章节的统计结果来看,使用最多的算法仍是SVM算法,而Apriori和CART算法甚至没有提及。因此,就使用位置而言,SVM算法的

影响力稳居十大算法的首位,Apriori和CART算法的影响力仍然垫底。这也与前文所得结果一致。

## 4.3 基于使用年代的算法使用行为及影响力分析

根据论文的收录信息,可得到每篇论文的发表年份。在本研究中,将文章所发表的年份记为某种算法对应的使用年份。对每一种算法在不同年代的提及论文数、平均提及次数进行统计,得到结果如图4和图5所示。可以看出,在本文所考察的自然语言处理论文中,十大算法从2001年才开始出现使用。

从图4中可以看出,就提及论文数而言,随着年代的推进,每个算法均呈现在波折中上升的变化趋势;且在2016年,几乎每个算法都有明显的下降,笔者认为原因可能是随着技术的发展,在自然语言

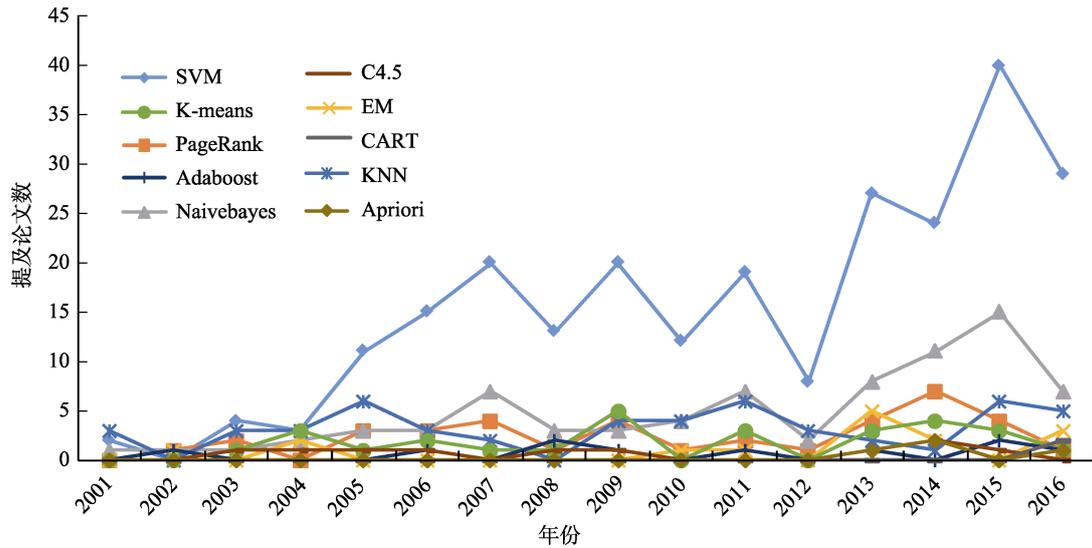


图 4 不同算法提及论文数演变图

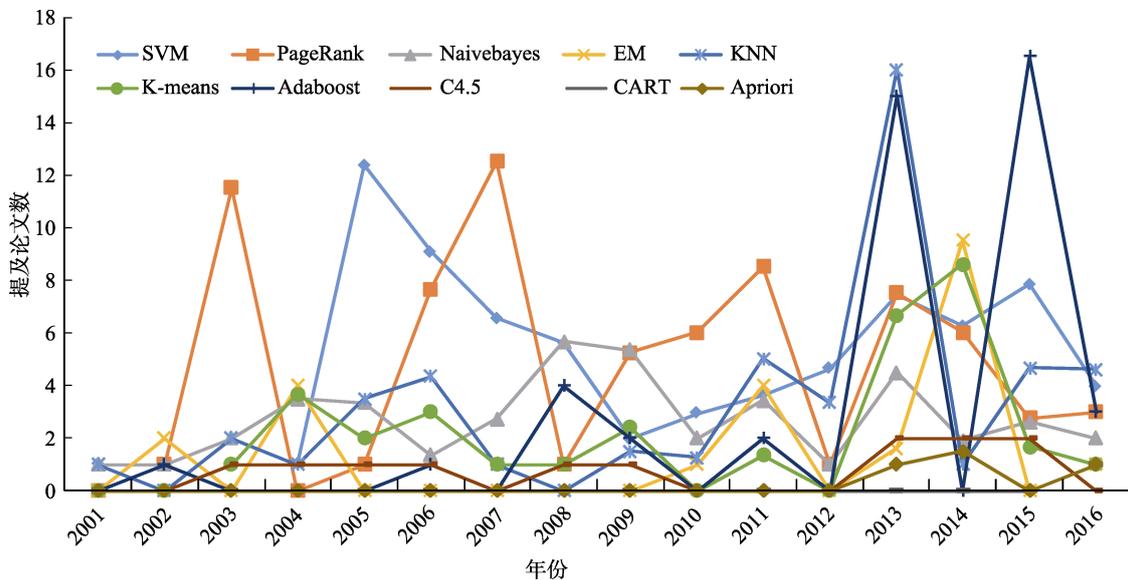


图 5 不同算法平均提及次数演变图

处理领域产生了新的算法（例如神经网络算法等），这些新的算法对处理不同的问题有着更好的性能，使学者们有了更大的选择空间。但从图 5 可知，就每一个算法而言，平均提及次数与年代没有什么关联。笔者认为，不管这些算法是诞生初期还是发展成熟期，当学者们要在研究中使用某个算法时，依然会在文章中花费笔墨来解释算法的背景、原理等，并不会因为算法的发展足够成熟，而仅仅只是在文章中一笔带过。就各个算法而言，每一年里 SVM 算法的提及论文数和平均提及次数都远高于其他算法，影响力最高；而 CART、Apriori 算法的提及论文数和平均提及次数较少，影响力较低。

#### 4.4 基于使用动机的算法使用行为分析

对算法使用动机的考察，能够提高算法使用行为研究的全面性。本研究依据第 3.5 节中提出的算法使用动机分类进行人工标注，统计得到所有算法的使用动机整体分布如图 6 所示。由前文结果可知，CART 算法与 Apriori 算法的提及论文数过少，为避免偶然性，在本节我们主要考察其他八种算法的使用情况。

由图 6 可以看出，就所有算法的使用动机整体分布情况而言，实际应用的占比略高于背景提及，这说明在自然语言处理领域，学者们使用算法的目的偏向于将其当做具体的数据处理方法来应用，这与对使用位置考察所得出的结论一致。表 9 是各算

法的背景提及与实际应用分布情况。

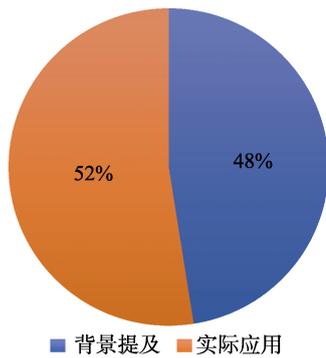


图6 十大算法的整体使用动机分布

表9 各算法背景提及与实际应用分布

算法名	使用动机	
	背景提及	实际应用
SVM	46.96%	53.04%
PageRank	55.26%	44.74%
NB	51.28%	48.72%
EM	57.14%	42.86%
KNN	50.98%	49.02%
K-means	32.14%	67.86%
Adaboost	40.00%	60.00%
C4.5	20.00%	80.00%

就各算法而言, PageRank、NaiveBayes、EM以及KNN算法的背景提及均高于实际应用。笔者推测原因可能为在一项研究中, 若要使用这四种算法, 则需要对算法的原理进行一定程度的阐述以及对他人使用对应算法的工作进行总结, 故背景提及必不可少; 同时, 在处理相应的问题时, 还有众多非本文考察对象的针对同种问题、性能更优的算法, 研究者在选用其他算法的同时, 仍然需要在相关工作部分整合他人使用这四种算法进行的研究, 导致这几种算法背景提及高于实际应用。在实际应用的基础上, 本文细致地统计了四种实际应用的使用动机分布情况, 结果如表10所示。

表10 各算法的使用动机分布

算法名	实际应用			
	单一使用	使用并比较	改进使用	改进并比较
SVM	45.04%	46.56%	0	8.40%
PageRank	35.29%	47.06%	0	17.65%
NB	28.95%	63.16%	0	7.89%
EM	66.67%	0	16.67%	16.67%
KNN	36.00%	48.00%	0	16.00%
K-means	73.68%	21.05%	0	5.26%
Adaboost	33.33%	16.67%	0	50.00%
C4.5	50.00%	50.00%	0	0

从四种实际应用动机的具体统计结果中可以看出, 算法的使用主要集中在“单一使用”与“使用并比较”两个方面, 这说明在多数情况下, 学者们倾向于直接使用十大算法来处理问题, 并试图通过结果的比较找出最佳解决办法, 体现出十大算法的性能优越性。在本文所划分的四种具体实际应用动机中, “改进使用”的情况最少; 结合“改进并比较使用”的情况来看, 不难推测, 若有学者在原有算法的基础上进行改进, 他们会将自己改进的算法与其他标准算法的处理结果进行对比, 从而证明自己算法的优越性。值得注意的是, Naïve Bayes算法的“使用并比较”远远高于“单一使用”, 前者所占的比重是后者的两倍之多。笔者认为原因可能是 Naïve Bayes算法有着坚实的数学原理, 且简单、易实现, 分类效果稳定, 在处理分类问题时, 非常适合作为基准方法 (baseline) 来与学者所采用的其他方法进行性能的比较。同时, 同样作为分类算法的 SVM、KNN以及C4.5算法的“使用并比较”所占比重均最高, 这说明在自然语言处理领域中, 分类算法多种多样, 学者们倾向于同时使用多种分类算法来处理数据, 并通过比较结果的优劣来判断性能最优的算法。相对的, K-means算法和EM算法的“单一使用”远远高于“使用并比较”。在十大算法中, 这两种算法的应用针对性较强。K-means算法是一种经典聚类算法, 原理简单、运算速度快, 笔者推测在自然语言处理领域中, 由于性能的优越性, K-means算法成为学者们在处理聚类问题时的首选算法, 故K-means算法的“单一使用”频率远远大于“比较使用”。同理, EM算法是一种针对统计学习的算法, 有着与K-means算法类似的结果。此外, Adaboost算法的改进并比较使用最为频繁。Adaboost算法是一种精度很高并且不会过度拟合的分类算法, 但其分类效果依赖于弱分类器的选择。对于这种性能极好的算法, 对其不足之处加以改进就能获得更好的结果, 笔者推测这正是导致Adaboost算法的“改进并比较”使用频率最高的原因。

## 5 讨论

由上述研究结果可以看出, 就各个算法而言, 从使用频次和使用位置的角度来看, SVM算法均展现出较高的影响力, 而CART、Apriori在每一项评估指标中影响力均较低。王玉琢等<sup>[1,13]</sup>以ACL会议论文为研究对象, 从提及频次与提及位置的角度对十大算法进行评估, 结果显示影响力最大的是SVM算法, 影响力明显低于其他算法的是Apriori算法。

考虑到本研究所用论文集与王玉琢等所用论文集分别来源于国内外 NLP 领域的顶级会议，能够分别代表国内外 NLP 领域的前沿研究，因此将本文结果与王玉琢等的研究结果进行比较，能够在一定程度上帮助学者掌握研究发展的最新动态及存在的差异，从而全面了解领域内的研究趋势。就十大算法的使用频次与使用位置而言，本文与其结果十分相似，表明在 NLP 领域，国内与国外在这两方面的差异较小。因此，我们提出猜想：在 NLP 领域，国内与国

外对于十大算法的使用差别是否会体现在使用时间上？从本文关于算法随时间演变的结果中可以看出，在 2001 年之后，NLP 领域的中文会议论文中才出现十大算法的使用。而其在国外的相关论文中是否会更早出现？但目前还没有针对国外 NLP 领域论文中十大算法随时间演变规律的研究。故本文以 1993—2016 年历届 ACL 论文为数据集，考察每年十大算法的总提及论文数，并与本文对使用年代的分析结果进行比较。所得结果如图 7 所示。

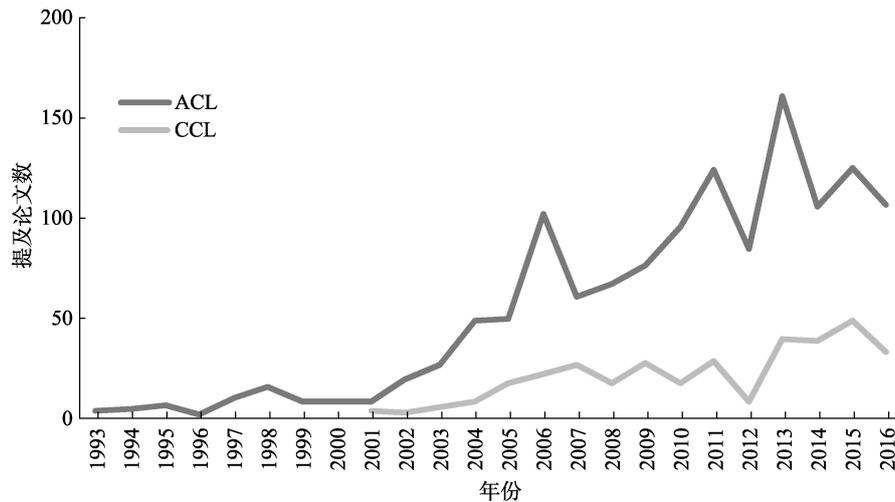


图 7 国内外十大算法总提及论文数演变对比图

从图 7 可以看出，不同于 CCL，在 ACL 论文集中，十大算法的使用从 1993 年开始出现。由此可知，与国外 NLP 领域相比，就十大算法的使用而言，国内的起步比国外滞后。但从整体演变趋势可以看出，国内与国外 NLP 领域对于十大算法提及论文数的波动规律较为相似，如在 2005—2006 年均有大量增长；在 2012 年、2016 年均有明显下滑，可以推断：就十大算法的使用而言，国内 NLP 领域虽然起步较晚，但总体发展过程与国外接近，说明国内 NLP 领域与国际前沿较为同步。此外，对于两者随年份变化的具体提及论文数量差异显著的情况，笔者认为原因在于会议的规模差距，因此不做具体数量的比较。对于每种算法是否存在时间上的使用差异，是我们下一步要研究的内容。

此外，本文重点考察算法的使用动机。通过构建算法的使用动机分类体系，统计出各算法在学术论文中的使用动机分布情况。王玉琢等<sup>[13]</sup>以 ACL 会议论文为研究对象，对论文中提及的算法与需要解决的实验任务进行分析。但在论文中提及算法可能只是介绍研究背景或相关工作，并未真正应用算法

来解决实验任务，因此仅考虑算法的提及具有一定的片面性。而本研究明确区分算法的背景提及与实际应用，并将实际应用进一步细分为单一使用、使用并比较、改进使用以及改进并比较 4 类，对于算法在文章中的角色定位更加明确。因此，未来可考虑将算法的使用动机与需要解决的实验任务相结合，从而更为细致地考察算法与任务之间的不同关系。

## 6 结论与展望

对不同领域算法影响力的评估有利于科研人员快速了解所研究领域的算法使用情况。本文以数据挖掘十大经典算法为研究对象，以自然语言处理领域为例，通过研究算法在相关学术论文中的使用频次、使用位置、使用年代以及使用动机，考察每类算法在学术论文中的使用行为，并在此基础上对各类算法在该领域的影响力进行评估。在专家票选结果中，第一名是 C4.5 算法，最后一名是 CART 算法<sup>[4]</sup>；而从本研究的影响力评估结果上看，总体影响力最大的是 SVM 算法，C4.5 算法影响力明显较低。两者有明显差异，说明在特定的学科领域中，对算法进

行量化分析的客观评估结果有别于人为的主观评估结果。此外,本文在前人研究成果的基础上,提出算法的使用动机分类体系,并以CCL论文集为例进行实证研究。研究表明,该动机分类体系能够适用于数据集中所有的算法提及,可为算法等微观实体的使用动机研究提供参考。另一方面,有别于普遍认知,对于算法这种工具型微观实体,在学术论文中不仅仅是作为“实际应用”的工具被使用,同时也存在多数作为“背景介绍”被提及的情况;若要对微观实体的使用进行研究,应当重视并区分这两种不同类型的使用,从而得到更正确的研究成果。本研究丰富了学术影响力评价的内容,扩充了全文内容方法的应用领域;所用方法可为微观实体的科学评价相关研究提供新的思路,所得结果可帮助初涉自然语言处理领域的科研工作者了解经典数据挖掘算法的使用情况以及影响力分布,为其在数据挖掘算法选择上提供一定的意见,解决他们对相关信息的需求。

同前人的研究相比,本文的创新点在于:从全文内容的角度出发,不仅考虑使用频次和使用位置,还深入分析算法的使用动机及算法的使用随年代的变化规律,所得结果较为全面。但仍然存在许多不足之处:首先,本研究仅对十大数据挖掘算法进行考察,实际上在NLP领域中常用的算法并不局限于此,未来将考虑扩大待研究算法的种类,并转变算法句识别的方式,结合自然语言处理的技术进行实体识别。其次,本研究所采用的使用动机标注体系是在前人对论文引用动机的研究成果上、结合本文数据集构建得出,针对性不够完全。未来将在此方面进行改进,更深入地细化动机的分类。另外,本研究只是以NLP领域的中文学术论文为例来进行研究,数据集较小,未来将考虑扩充中文语料(如《中文信息学报》等相关期刊论文)。我们将进一步以自然语言处理领域的英文论文作为研究对象,对其中的算法使用行为进行研究,并将本文的研究结果与以英文语料为数据集所得的结果进行比较。最后,在未来研究中,我们还可以结合基于引文内容分析的方法,对算法在学术论文中的引用行为进行深入地考察。

### 参 考 文 献

[1] 王玉琢,章成志.考虑全文本内容的算法学术影响力分析研究[J].图书情报工作,2017,61(23):6-14.  
[2] Wilbanks E G, Facciotti M T. Evaluation of algorithm perform-

ance in ChIP-seq peak detection[J]. PLoS ONE, 2010, 5(7): e11471.  
[3] Nesma S, Mohammed B, Mohammed C. Statistical comparisons of the Top 10 algorithms in data mining for classification task[J]. International Journal of Interactive Multimedia and Artificial Intelligence, 2016, 4(1): 46-51.  
[4] Wu X. Top 10 algorithms in data mining[J]. Knowledge Information System, 2008, 14(1): 1-37.  
[5] Ding Y, Song M, Han J, et al. Entitymetrics: Measuring the impact of entities[J]. PLoS ONE, 2013, 8(8): e71416.  
[6] Kathy M, Hal D, Snigdha C, et al. Predicting the impact of scientific concepts using full-text features[J]. Journal of the Association for Information Science and Technology, 2016, 67(11): 2684-2696.  
[7] 丁楠,黎娇,李文雨泽,等.基于引用的科学数据评价研究[J].图书与情报,2014(5):95-99.  
[8] Belter C W. Measuring the value of research data: A citation analysis of oceanographic data sets[J]. PLoS ONE, 2014, 9(3): e92590.  
[9] 王雪,马胜利,余曾漂,等.科学数据的引用行为及其影响力研究[J].情报学报,2016,35(11):1132-1139.  
[10] Pan X, Yan E, Ming S, et al. Examining the usage, citation, and diffusion patterns of bibliometric mapping software: A comparative study of three tools[J]. Journal of Informetrics, 2018, 12 (2): 481-493.  
[11] 赵蓉英,魏明坤,汪少震.基于Altmetrics的开源软件学术影响力评价研究[J].中国图书馆学报,2017,43(2):80-95.  
[12] 杨波,王雪,余曾漂.生物信息学文献中的科学软件利用行为研究[J].情报学报,2016,35(11):1140-1147.  
[13] Wang Y Z, Zhang C Z. Using full-text of research articles to analyze academic impact of algorithms[C]// Proceedings of International Conference on Information. Springer, 2018: 395-401.  
[14] 赵蓉英,曾宪琴,陈必坤.全文本引文分析—引文分析的新发展[J].图书情报工作,2014,58(9):129-135.  
[15] Ding Y, Zhang G, Chambers T, et al. Content-based citation analysis: The next generation of citation analysis[J]. Journal of the Association for Information Science and Technology, 2014, 65(9): 1820-1833.  
[16] 胡志刚.全文引文分析方法与应用[M].北京:科学出版社,2017.  
[17] 王文娟,马建霞,陈春,等.引文文本分类与实现方法研究综述[J].图书情报工作,2016,60(6):118-127.  
[18] 刘盛博.科学论文的引用内容分析及其应用[D].大连:大连理工大学,2014.  
[19] 李婷婷,李秀霞.基于引文内容的信息学期刊互引分析[J].情报杂志,2016,35(2):110-115.  
[20] An J, Kim N, Kan M Y, et al. Exploring characteristics of highly cited authors according to citation location and content[J]. Journal

- of the Association for Information Science and Technology, 2017, 68(8): 1975-1988.
- [21] Hassan S U, Safder I, Akram A, et al. A novel machine-learning approach to measuring scientific knowledge flows using citation context analysis[J]. *Scientometrics*, 2018, 116(2): 973-996.
- [22] Garfield E. Can citation indexing be automated?[C]// *Proceedings of the Symposium on Statistical Association*, Washington DC, 1964: 84-90.
- [23] Weinstock M. Citation indexes[J]. *Encyclopedia of Library and Information Science*, 1971, 5(1): 16-40.
- [24] Brooks T A. Private acts and public objects: an investigation of citer motivations[J]. *Journal of the American Society for Information Science*, 2010, 36(4): 223-229.
- [25] Erikson M G, Erlandson P. A taxonomy of motives to cite[J]. *Social Studies of Science*, 2014, 44(4): 625-637.
- [26] Moravcsik M J, Murugesan P. Some results on the function and quality of citations[J]. *Social Studies of Science*, 1975, 5(1): 86-92.
- [27] Teufel S, Siddharthan A, Tidhar D. An annotation scheme for citation function[C]// *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*. Stroudsburg: Association for Computational Linguistics, 2006: 80-87.
- [28] Jurgens D, Kumar S, Hoover R, et al. Measuring the evolution of a scientific field through citation frames[J]. *Transactions of the Association for Computational Linguistics*, 2018, 6: 391-406.
- [29] 邱均平, 陈晓宇, 何文静. 科研人员论文引用动机及相互影响关系研究[J]. *图书情报工作*, 2015, 59(9): 36-44.
- [30] Ding Y, Liu X, Guo C, et al. The distribution of references across texts: Some implications for citation analysis[J]. *Journal of Informetrics*, 2013, 7(3): 583-592.
- [31] McCain K, Tuhneh K. Citation context analysis and aging patterns of journal articles in molecular genetics[J]. *Scientometrics*, 1989, 17(1-2): 127-163.
- [32] Lin L, Evans S. Structural patterns in empirical research articles: A cross-disciplinary study[J]. *English for Specific Purposes*, 2012, 31(3): 150-160.
- [33] 崔明, 潘雪莲, 华薇娜. 我国图书情报领域的软件使用和引用研究[J]. *中国图书馆学报*, 2018(3): 66-78.

(责任编辑 魏瑞斌)