浙江理工大学学报(社会科学版),第38卷,第5期,2017年10月
Journal of Zhejiang Sci-Tech University (Social Sciences)
Vol. 38, No. 5, Oct. 2017

DOI:10.3969/j.issn.1673-3851.2017.10.015

基于 *k*-means 聚类分析的高校论文统计研究

查香云[1],吕国良[2]

(1.杭州科技职业技术学院信息工程学院,杭州 311402;2.浙江理工大学图书馆,杭州 310018)

摘　要:科研论文的质量能反映学者、学术机构和学术团队的科研水平。文章选取浙江理工大学、浙江工业大学、浙江师范大学等十所高校,提取各高校在2012—2016年所有被WOS核心合集中SCI数据库收录的论文,采用非监督的机器学习*k*-means算法对发文量、去自引被引频次、去自引施引文献及发文量权值等四个特征变量进行统计分析。结果表明:在这十所高校中,浙江工业大学和宁波大学属于第一等级,浙江理工大学、浙江师范大学和杭州师范大学属于第二等级,中国计量大学、杭州电子科技大学、温州大学、浙江农林大学和浙江工商大学属于第三等级。文章研究表明,利用*k*-means方法横向比较高校科研论文质量具有可行性。

关键词:*k*-means;中科院期刊分区;MATLAB;归一化;去自引被引频次

中图分类号:TP311　　**文献标志码**:A　　**文章编号**:1673-3851 (2017) 05-0478-05

科研论文是科研成果的重要表现形式,论文的质量很大程度上反映了个人或机构的科研水平。科研论文的质量评价指标,无疑会对科学研究起到导向性的作用。相比于早期的易受主观因素影响的同行评议,目前人们更认可能体现客观性的论文质量评价指标——论文的被引频次和期刊影响因子。

虽然这两个指标在学术的研究和评价中带来了相当积极的影响,然而,在持续的文献计量的研究中,笔者也发现了其诸多弊端。第一,不同学科领域,受人们的关注的程度不同,会造成论文的被引频次的巨大差异,只用论文被引频次和基于论文被引频次计算的期刊影响因子来表征论文质量的优劣,不够全面。在不同的学科领域,这两个指标完全没有可比性[1]。第二,虽然SCI对期刊做了学科的分门别类,增加这两个指标的可比性,但是由于学科间的渗透融合,期刊的分类标准本身也是研究的一个主题,因此仅从这个层面来考量或研究科研论文的质量和水平,也显得困难重重。第三,基本科学指标(essential science indicators,ESI)作为一种衡量科研水平相对高低的指标,刚一出现,即受到广泛的关注和重视,具有很强的导向性作用。可是,ESI是在学科分类基础上,把论文的被引频次作为唯一指标来衡量论文的质量,这缺少合理性。此外,ESI只显示了达到基线指标的相关的学科、机构和论文等的信息,难以从中了解学术机构的更详实的水平,因此ESI指标在科研管理中也缺乏可操作性。在学术机构作为基本单位进行排序(如ESI的排名)的方法中,鉴于论文的被引频次作为计算指标的不足,本文采用论文的被引频次与其在“中科院期刊分区”(2016年版,以下所称“中科院期刊分区”即指该版本)中的论文权重并举的计算方法;同时,本文也认为,以学术机构作为排序的基本单位,其粒度过细,因此提出了以较粗粒度——学术机构的聚类,作为学术机构(本文以高校作为实证研究对象)排序的基本单位的方法。粗粒度较之于细粒度,其排序结果能够显现出较好的鲁棒性。

k-means是一种非监督机器学习算法,具有易收敛、操作性强的特点。*k*-means在生产实践[2-5]、商务旅游[6-8]、图像处理[9-10]和文本分类等领域都得到了有效应用。朱亮亮[11]把*k*-means应用在数据

收稿日期:2017-06-21　　网络出版日期:2017-09-27

作者简介:查香云(1972-),女,安徽安庆人,讲师,本科,主要从事信息安全、多媒体等方面的研究。

通信作者:吕国良,E-mail:lvgl@zstu.edu.cn

清洗中的人名的消歧,文献[12-14]研究了利用k-means在图书馆服务中实现文献自动推送。但这些研究,均着眼于单一的具体的研究对象内的数据聚类分析,未涉及多个对象的聚类的比较分析。

本文从文献的四个特征出发,摈弃单纯的使用引文数量的分析方法,结合使用"中科院期刊分区"中论文质量的划分标准,实证研究了 k-means 算法在多个研究对象——十所高校中的聚类分析中的应用。

一、数据的获取和处理

本文分析的文献数据,来自 WOS(web of science)的核心合集中的 SCI 数据库。时间区间:2012—2016 年;检索数据库:web of science "core collection"(SCI);检索字段:organization-enhanced;文献类型:article,review。具体情况如表 1 所示。

表 1　浙江省十所高校论文 SCI 论文特征值(2012—2016)

高校名称	中国计量大学	杭州电子科技大学	杭州师范大学	宁波大学	温州大学	浙江农林大学	浙江工商大学	浙江师范大学	浙江理工大学	浙江工业大学
发文量/篇	2093	2125	3131	4335	1772	1557	1458	2953	2453	4903
去自引被引频次/次	10228	9736	22905	20903	11250	7700	6560	21299	16059	25217
去自引施引文献/篇	8772	7753	18796	17373	9423	6361	5753	17377	13367	21141

注:获取数据的时间为 2017 年 5 月 25 日。

文献计量分析中,论文属性的选择和处理是十分关键的一环,本文选择能反映论文价值的特征作为文献计量的属性,通过归一化等处理方式统一属性间的量度标准。本文所述论文抽取的特征属性如表 2 所示,具体表述如下:

a)属性 a:"发文量",即论文数量,是高校的科研规模和科研产出能力的表征;

b)属性 b:"去自引被引频次",显示了论文所承载的学术成果被他人所认可的程度;

c)属性 c:"去自引施引文献",一定程度上反映了学术成果的影响面;

d)属性 d:"发文量权值",依据中科院期刊分区原则赋予论文权重计算所得的论文的权值。

表 2　文献属性一览

属性 a	属性 b	属性 c	属性 d
发文量/篇	去自引被引频次/次	去自引施引文献/篇	发文量权值

依据"中科院期刊分区",本文把 SCI 期刊分为一区、二区、三区、四区及未收录五个类别。

"中科院期刊分区"的分区原则:将 SCI 期刊分为 13 个大类学科,在每个大类学科内,所有期刊按照学术影响力(3 年平均 IF)由高到低降序排列;依据该期刊排序,将期刊划分为一区、二区、三区和四区四个等级。"中科院期刊分区"的一区到四区的期刊数量不等,呈金字塔状分布。在大类学科中,取前 5%(含 5%)为一区、5%～20%(含 20%)为二区、20%～50%(含 50%)为三区,50%～100%(含 100%)为四区[15-16]。不论领域,只要论文发表的刊物在同一个分区,就可以认为这些论文的质量是相当的[17]。依据这一原则,赋予每个分区的期刊论文的权重。"中科院期刊分区"论文的权重分配见表 3。特别说明,表 3 中的"未收录",是指未被"中科院期刊分区"收录的 SCI 期刊。由于这些期刊毕竟也属于 WOS 中的 SCI 期刊,所以本文赋予了较小的权重"0.5"。

表 3　期刊分区论文的权重分配表

期刊分区	一区	二区	三区	四区	未收录
入选比例/%	5	5～20	20～50	50～100	0
权重	20.0	5.0	2.0	1.0	0.5

发文量权值计算公式:

$$\Delta = \sum \mu_k * \lambda_k \qquad (1)$$

其中:Δ 表示发文量权值;k 表示"中科院期刊分区"的五个类别,即一区、二区、三区、四区和未收录;μ_k 表示分区类别 k 的文献量;λ_k 表示分区类别 k 的权重。

根据"中科院期刊分区"统计各高校论文在期刊分区上的分布,并依据式(1)计算发文量权值,结果见表 4。

表 4　各高校各分区论文的数量及权值

序号	高校名称	一区/篇	二区/篇	三区/篇	四区/篇	未收录/篇	总计/篇	发文量权值
1	中国计量大学	46	393	681	913	60	2093	5190.00
2	杭州电子科技大学	99	491	657	811	67	2125	6593.50
3	杭州师范大学	338	832	892	995	74	3131	13736.00
4	宁波大学	275	926	1365	1708	61	4335	14598.50

表4续

序号	高校名称	一区/篇	二区/篇	三区/篇	四区/篇	未收录/篇	总计/篇	发文量权值
5	温州大学	125	446	557	581	63	1772	6456.50
6	浙江农林大学	90	331	457	641	38	1557	5029.00
7	浙江工商大学	124	384	347	567	36	1458	5679.00
8	浙江师范大学	236	756	871	974	116	2953	11274.00
9	浙江理工大学	179	573	781	875	45	2453	8904.50
10	浙江工业大学	342	1167	1363	1913	118	4903	17373.00

对表1和表4作了汇总,结果见表5。

表5 四个特征属性值汇总

序号	高校名称	属性a	属性b	属性c	属性d
1	中国计量大学	2093	10228	8772	5190.00
2	杭州电子科技大学	2125	9736	7753	6593.50
3	杭州师范大学	3131	22905	18796	13736.00
4	宁波大学	4335	20903	17373	14598.50
5	温州大学	1772	11250	9423	6456.50
6	浙江农林大学	1557	7700	6361	5029.00
7	浙江工商大学	1458	6560	5753	5679.00
8	浙江师范大学	2953	21299	17377	11274.00
9	浙江理工大学	2453	16059	13367	8904.50
10	浙江工业大学	4903	25217	21141	17373.00

为避免特征属性的不同量纲对k-means的结果的影响,对四个特征属性的值进行线性归一化处理。归一化映射公式如下:

$$a_i'=\frac{a_i-\min(a_i)}{\max(a_i)-\min(a_i)} \tag{2}$$

其中:a_i为一组数据集中的第i个值;$\min(a_i)$为该数据集的最小值,$\max(a_i)$为该数据集的最大值;a_i'为第i个数据的归一化处理后的值。

二、相似度和聚类的计算

本文采用欧氏距离来计算元素之间的相似度。两个元素的欧氏距离值越小,两者相似度越高。距离值越大,则相异度越高。

设有两元素X和Y,其都具有n个属性,则X、Y之间的欧氏距离$D(X,Y)$表示为:

$$D(X,Y)=\sqrt{((X_1-Y_1)^2+(X_2-Y_2)^2+(X_3-Y_3)^3+\cdots+(X_n-Y_n)^2)} \tag{3}$$

k-means算法,是指含有n个元素的集合D,$D=\{X_1,X_2,X_3,\cdots,X_n\}$,每个有可观察属性有$m$个,即$X_1$有属性$\{X_{11},X_{12},X_{13},\cdots,X_{1m}\}$,$X_2$有属性$\{X_{21},X_{22},X_{23},\cdots,X_{2m}\}$,$\cdots$,$X_n$有属性$\{X_{n1},X_{n2},X_{n3},\cdots,X_{mm}\}$。假定要把这$n$个对象分成$k$个子集,即$k$个簇($k<n$),则具体计算步骤为:第一步,在$n$个元素中取$k$个元素作为第一次聚类的初始中心点,分别计算这$n$个元素与这$k$个初始中心点的距离(本文取欧氏距离),取与这$k$个中心点的距离最近的对象为簇,即有$k$个簇,第一次迭代结束;第二步,计算上一步骤的$k$个簇中各自构成的元素的平均值,作为其中心点,再一次计算所有n个元素与这新的中心点的欧氏距离,取距离最近的对象为簇,得新的k个中心点,第二次迭代结束;如此反复迭代,直至中心点不再变化。围绕着这k个中心点的元素,也不再更新。k-means聚类收敛,迭代结束。

依据表5中的四个特征属性作为可观察属性的项,计算欧氏距离。设定把10所高校分为三个层次,则在k-means聚类中,取$k=3$。

k-means算法的终止条件可以是以下中的任何一个:

a)没有数据对象被重新分配到不同的聚类;

b)聚类中心收敛;

c)误差平方和局部最小。

根据式(2),在MATLAB软件下运行程序,结合表5,输出的归一化值如表6所示。

表6 四个特征属性归一化值

序号	高校名称	属性a	属性b	属性c	属性d
1	中国计量大学	0.18433	0.19660	0.19619	0.01304
2	杭州电子科技大学	0.19361	0.17023	0.12997	0.12674
3	杭州师范大学	0.48563	0.87608	0.84761	0.70536
4	宁波大学	0.83512	0.76877	0.75513	0.77523
5	温州大学	0.09115	0.25138	0.23850	0.11564
6	浙江农林大学	0.02874	0.06110	0.03951	0.00000
7	浙江工商大学	0.00000	0.00000	0.00000	0.05266
8	浙江师范大学	0.43396	0.79000	0.75539	0.50591
9	浙江理工大学	0.28882	0.50914	0.49480	0.31396
10	浙江工业大学	1.00000	1.00000	1.00000	1.00000

根据式(3)和表6,在MATLAB运行k-means程序,输出的结果为表征各个簇(即聚类)的代码。同一个簇,其代码数字是相同的。对应表6高校名称的排列顺序,程序运算结果见表7。表7中的簇代码,不表示大小或顺序,数字相同的数据对象位于同一个簇。

表 7　簇代码与高校对应表

簇代码	高校名称
3	中国计量大学
3	杭州电子科技大学
2	杭州师范大学
1	宁波大学
3	温州大学
3	浙江农林大学
3	浙江工商大学
2	浙江师范大学
2	浙江理工大学
1	浙江工业大学

显见,三个簇的对象分别是:簇 1:浙江工业大学和宁波大学;簇 2:浙江理工大学、浙江师范大学和杭州师范大学;簇 3:中国计量大学、杭州电子科技大学、温州大学、浙江农林大学和浙江工商大学。

三、结　语

综合分析表明,所统计的这四个指标属性中,浙江工业大学在 10 所高校中都处于榜首位置,无疑是这 10 所高校的领军者。宁波大学以四项相对比较均衡的指标值显示出其较强的科研能力,*k*-means 算法聚类结果显示,它与浙江工业大学位居同一层次。

a)文献发文量显示了该统计区间(2012—2016 年)高校的科研成果的产出。科学研究是需要投入的,科研投入与产出一般是正相关关系,因此文献的发文量与该高校获得科研经费的能力相关,这也是一种科研能力的体现。本文中,浙江工业大学以总量 4903、占比 18.31%占居首位;其次为宁波大学,总量 4335,占比 16.19%。杭州师范大学、浙江师范大学和浙江理工大学紧随其后。

b)文献数量只是评价科研能力的一个指标,科研能力还体现在文献的质量、学术影响的深度和广度上面。WOS 平台为我们提供了现成的影响力指标——文献被引频次和文献的施引文献。本文选择更具有客观性的"去自引被引频次"和"去自引施引文献"两种指标。去自引被引频次,是文献被他人关注和认可的客观反映。去自引被引频次越高,表明文献所承载的研究成果越被他人所推崇和认可,影响也就越是深远;施引文献是被引文献的知识的发展面,揭示了知识流动的方向,也即原始文献所承载的研究成果的影响广度。表 1 显示,浙江工业大学以去自引被引频次 25217 次,去自引施引文献 21141 篇,独占鳌头,杭州师范大学则分别以 22905 次和 18796 篇位居其二。

c)根据"中科院期刊分区"加权获得的发文量权值反映了论文的整体质量,发文量权值越大,论文的总体质量越高。从表 4 可以看出,浙江工业大学、宁波大学和杭州师范大学位列三甲。

在高校内部的科研管理中,使用该方法统计分析各学科、各学术团队或各学术机构如研究所和学院的科研论文,利用 *k*-means 对他们的学术发展水平做一个统一的聚类分析和评估,简单方便,操作性强。本文中的不足之处在于:a)本文只统计了 SCI 的论文,其聚类排名只限于在理工科方面的学术水平的展现;b)未考虑作者在具体的论文中的排名,而致在合作发表的论文中对各高校的学术贡献程度的揭示不够充分。

参考文献:

[1] 丁佐奇,郑晓南.期刊影响因子、论文被引证次数与学术质量评价的矛盾分析[J].中国科技期刊研究,2009(2):286-288.

[2] 边振兴,杨子娇,钱凤魁,等.基于 LESA 体系的高标准基本农田建设时序研究[J].自然资源学报,2016(3):436-446.

[3] 刘艳秋,武佩,张丽娜,等.母羊产前行为特征分析与识别:基于可穿戴检测装置构架[J].农机化研究,2017(9):163-168.

[4] 常亮,郭垚嘉,贾炯,等.利用聚类算法分析河北省地震分布状况[J].高原地震,2017(2):12-16.

[5] 刘仕兵,葛俊祥.基于 K-means 聚类法的牵引供电隔离开关故障状态监测[J].华东交通大学学报,2017(3):109-117.

[6] 陈钢华,黄远水.旅游者重游决策的影响因素实证研究:基于网络调查[J].旅游学刊,2008(11):69-74.

[7] 陈晓艳,黄震方,胡小海,等.事件旅游城市居民分类及影响因素研究:以常州花博会为例[J].南京师大学报(自然科学版),2016(1):108-116.

[8] 丛丽,吴必虎,张玉钧,等.野生动物旅游场所涉入实证分析:以澳大利亚班布里海豚探索中心为例[J].北京大学学报(自然科学版),2017(4):1-6.

[9] 蔡志华.基于 K 均值聚类的彩色图像快速分割方法[J].计算机与数字工程,2013(8):1328-1330.

[10] 李文博,强少卫.基于 BMP 位图的簇绒机花型图像处理技术初探[J].纺织科技进展,2017(6):9-11.

[11] 朱亮亮.利用改进的 K-means 算法实现文献著者人名消歧[J].软件导刊,2013(5):63-66.

[12] 常盛.*k*-means 聚类算法在提高图书馆数字文献服务效能中的应用[J].电子技术与软件工程,2016(23):163-164.

[13] 吉雍慧.数字图书馆中的检索结果聚类和关联推荐研

究[J].现代图书情报技术,2008(2):69-75.
[14] 张宏,王新玲,张丽.基于读者文献推送需求分析的医院图书馆精准服务实践[J].中华医学图书情报杂志,2016(4):74-77.
[15] 中科院网站.JCR期刊分区数据在线平台[EB/OL].(2016－10－15)[2017－06－15]. http://www.fenqubiao.com.
[16] 刘芳,朱沙.学术期刊主要评价体系差异性研究[J].高等教育研究学报,2015(1):33-38.
[17] 李秋实,刘红玉.基于文献计量的期刊分区与论文学术评价量化实证研究[J].图书馆工作与研究,2015(4):60-66.

Statistical Research of University Papers Based on *K*-means Cluster Analysis

ZHA Xiangyun[1], *LÜ Guoliang*[2]
(1. School of Info Engineering, Hangzhou Polytechnic, Hangzhou 311402, China;
2. Library, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: The quality of research papers can reflect scientific research level of scholars, academic institutions and academic team. Ten universities such as Zhejiang Sci-Tech University, Zhejiang University of Technology and Zhejiang Normal University were chosen, and their articles' data downloaded from SCI database of "core collection" in WOS (Web of Science, WOS) during five years (2012—2016) were extracted in this paper. *K*-means, an unsupervised algorithm, was employed for statistical analysis of four characteristic variables including quantity of publications, citation frequency without self-citation, citing articles without self-citation and weight of publications. The results showed that among these ten universities, Zhejiang University of Technology and Ningbo University are clustered to the first level; Zhejiang Sci-Tech University, Zhejiang Normal University and Hangzhou Normal University fall into the 2nd level and the other five universities (China Jiliang University, Hangzhou Dianzi University, Wenzhou University, Zhejiang A & F University and Zhejiang Gongshang University) belong to the 3rd level. The study showed that it is feasible to apply *k*-means for horizontal comparison of the quality of universality papers.

Key words: *k*-means; CAS Journal Section; MATLAB; normalization; citation frequency without self-citation

(责任编辑:任中峰)