

斯马里科夫统一方程描述专利信息离散分布的适用性评价^{*}

苑朋彬, 赵蕴华

(中国科学技术信息研究所, 北京 100038)

摘要: 本文主要从图像法的角度探讨多重分类号下专利布氏分布的特点, 并以斯马里科夫统一方程计算值为参考, 通过对比IPC和主IPC区域划分角度下图像的拟合效果和K-S检验结果, 得到在主IPC区域划分角度下方程适用性较高, 多重分类号对专利布氏分布的影响较小。

关键词: 斯马里科夫统一方程; 主IPC; K-S检验; 专利; 离散分布

中图分类号: G350

DOI: 10.3772/j.issn.1673-2286.2016.8.002

科学信息离散分布的特点是全部信息活动的基石, 也是对科学信息活动进行有效管理的基础。因而, 揭示和研究信息离散分布现象是情报学的重要课题^[1]。布拉德福定律(简称布氏定律)首次以位次/频率排序的方法将科技期刊文献信息的离散分布规律予以揭示, 随后该定律被运用到确定核心期刊、制定采购策略、优化馆藏、评价检索系统、比较学科成熟度等多个方面^[2]。

专利信息同科技期刊文献信息一样, 在技术学科间也存在很大交叉性。因此, 近年来诸多学者如张鹏等^[3]、吕义超等^[4]、胡晨希等^[5]、罗爱静等^[6]按照布氏定律的研究思路划分专利技术区域, 从区域法的角度证明专利文献在某一主题领域的分布存在明显的集中、离散现象, 为确定核心专利分类号、核心专利权人等提供一种实用方法。

然而对专利文献涉及的技术类别以IPC分类号进行标引可知: 一篇专利文献可能不仅属于一个技术类别, 还可能属于多个技术类别; 分类号与专利文献的对应关系为n:1, 而期刊与文献的对应关系为1:1。尽管上述学者从区域法的角度证实专利分布遵循布氏分布的比例关系, 但专利形态分布是否遵循布氏分布, 多重分类号对

专利布氏分布是否存在影响, 仍有待探讨。

因此, 本文从图像法的角度探讨专利形态分布的离散特点。图像法常采用一定的数学模型对研究对象的分散规律进行刻画。斯马里科夫统一方程是基于科技期刊论文提出并用来描述布拉德福分散曲线的数学方程之一^[7], 以此方程作为研究基础, 能够在一定程度上揭示专利信息的离散分布状况。通过对比不同技术成熟度主题的方程拟合效果, 确定不同技术成熟度、多重分类号的专利分布特点, 在此基础上, 分别对比IPC分类号和主IPC分类号的方程适用性, 以确定何种划分角度的专利分布更符合布氏形态分布。

1 研究思路

首先, 选取处于能量管理领域不同技术成熟度的两个技术分支进行斯马里科夫统一方程的拟合, 以方程计算值为参考, 观察专利分布形态的特点; 其次, 通过对比IPC和主IPC两种区域划分角度的图像拟合效果和K-S检验结果, 得出不同区域划分角度下多重分类号对专利布氏分布产生影响的大小; 最后, 确定何种区域划

^{*} 本研究得到国家科技支撑计划课题“面向科技创新的专利信息加工与服务关键技术研究与应用示范”(编号: 2015XM56)和国家社会科学基金项目“绿色技术创新驱动我国战略性新兴产业发展的战略研究”(编号: 13CGL007)资助。

分角度的方程适用性更高, 更符合布氏形态分布。

2 研究步骤

2.1 数据来源及预处理

本试验数据选取德温特专利数据库收录的燃料电池汽车能量管理技术领域和纯电动汽车能量管理技术领域的相关专利。选取二者作为研究对象, 能够在一定程度上对比斯马里科夫统一方程在同一技术领域不同技术分支、不同生命周期阶段下的适用性。运用关键词结合德温特手工代码进行检索, 检索日期为2016年1月

5日, 检索时间为1963年1月1日至今的所有专利数据, 去重后得到燃料电池汽车能量管理技术相关专利829条, 纯电动汽车能量管理技术相关专利4 956条。

采用主IPC区域划分角度分析时, 需要对德温特专利数据库中PD (Patent Detail) 字段进行逐条提取, 删除重复IPC, 保留同族专利申请国的主分类号; 然后, 将处理好的数据按照位次/频率排序, 形成布氏分布表 (见表1和表2)。其中n代表按照专利数量降序排列的累计序列号, R(n)代表累计专利数, 由于专利与专利分类号间是1:n的对应关系, 所以R(n)量大于实际专利总数。

表 1 燃料电池汽车能量管理技术领域IPC、主IPC布氏排序表

n	专利量	IPC	R(n)	n	专利量	主IPC	R(n)
1	276	B60L-011/18	276	1	128	B60L-011/18	128
2	136	H02P-027/06	412	2	72	H01M-008/04	200
3	131	B60L-003/00	543	3	58	H02P-027/06	258
4	130	H01M-008/04	673	4	45	B60L-003/00	303
...	...	...	...	...	...	...	...
788	1	H05K-005/06	4 060	392	1	H05K-007/02	1 390

表 2 纯电动汽车能量管理技术领域IPC、主IPC布氏排序表

n	专利量	IPC	R(n)	n	专利量	主IPC	R(n)
1	1 644	H02J-007/00	1 644	1	941	H02J-007/00	941
2	1 565	B60L-011/18	3 209	2	713	B60L-011/18	1 654
3	639	H01M-010/48	3 848	3	303	G01R-031/36	1 957
4	607	B60L-003/00	4 455	4	229	B60L-003/00	2 186
...	...	...	...	...	...	...	...
2 039	1	H10M-008/00	17 298	1 037	1	H05K-009/00	6 669

2.2 斯马里科夫统一方程拟合专利分布

斯马里科夫统一方程能有效模拟无格鲁斯下垂时的布拉德福分布。目前, 关于布氏定律图像描述的公式较多, 已产生10多个经验型数学公式^[2], 较为著名的有布鲁克斯模型、莱姆库勒模型、斯马里科夫统一方程等, 许多学者针对各种公式进行择优评价, 发现斯马里科夫统一方程拟合效果更好。例如, 王崇德等运用斯马里科夫统一方程和Egghe公式对应用物理学和润滑工程两个领域

的文献数据分布进行拟合^[8], 得出斯马里科夫统一方程对图像的描述要优于Egghe公式; 周爱民根据图像是否存在格鲁斯下垂, 提出当图像不存在格鲁斯下垂时, 斯马里科夫统一方程的拟合效果比其他模型略好^[9]。

斯马里科夫统一方程中K、P、q、b为参数, n代表相关的专利累计号, 见公式(1)。运用到专利系统中, 以专利量降序排序对应的序列号对数ln(n)为横坐标, 以专利累积量R(n)为纵坐标, 绘制各点形成布拉德福离散分布曲线 (见图1)^[10]。若令x=ln(n), 其方程则变成关

于 x 的方程,如公式(2)所示,此方程可用一元非线性回归来拟合专利布氏分布的4个参数。

$$R(n)=K\lg(n+Pe^{-qn})+b \quad (1)$$

$$R(x)=K\lg(e^x+Pe^{-qe^x})+b \quad (2)$$

2.3 K-S假设检验方程拟合效果

方程最终拟合效果主要通过K-S法进行检验。K-S检验(Kolmogorov-Smirnov)又称D检验,是一种拟合优度检验方法,常被用于检验一组样本数据的实际分布是否符合某一指定的理论分布。其基本原理是通过实际值频数和理论值频数的对比,找出最大的差异点,然后参照抽样分布,确定此差异是否出于偶然。运用到专利系统中,基于IPC和主IPC两种划分角度,作出两组独立假设。

假设1(IPC划分角度):

H_0 : 专利的IPC分布总体符合斯马里科夫统一方程;

H_1 : 专利的IPC分布总体不符合斯马里科夫统一方程。

假设2(主IPC划分角度):

H_0 : 专利的主IPC分布总体符合斯马里科夫统一方程;

H_1 : 专利的主IPC分布总体不符合斯马里科夫统一方程。

$D=\max|(F_n(x)-F_0(x))|$,其中 $F_n(x)$ 为随机样本的累计分布函数, $F_0(x)$ 表示理论分布函数。当 $D>D(n,\alpha)$ [$D(n,\alpha)$ 是显著性水平为 α ,样本量为 n 的拒绝临界值],拒绝原假设 H_0 ,反之则拒绝原假设 H_1 。取 $\alpha=0.05$,表示该模型在95%的置信区间是可信的。

3 实证对比研究

3.1 拟合效果对比研究

采用斯马里科夫统一方程进行参数估计和曲线拟合,将实际专利分散曲线和理论分散曲线呈现在同一图上。图1分别展示了在IPC和主IPC区域划分角度下,燃料电池汽车能量管理技术领域和纯电动汽车能量管理技术领域的拟合效果,其中横坐标为递减排列的专利对应的序号对数值 $\ln(n)$,纵坐标为专利累计数量 $R(n)$, R^2 代表相关系数,取值范围在0—1, R^2 越大,自变量对因变量的解释程度越高。

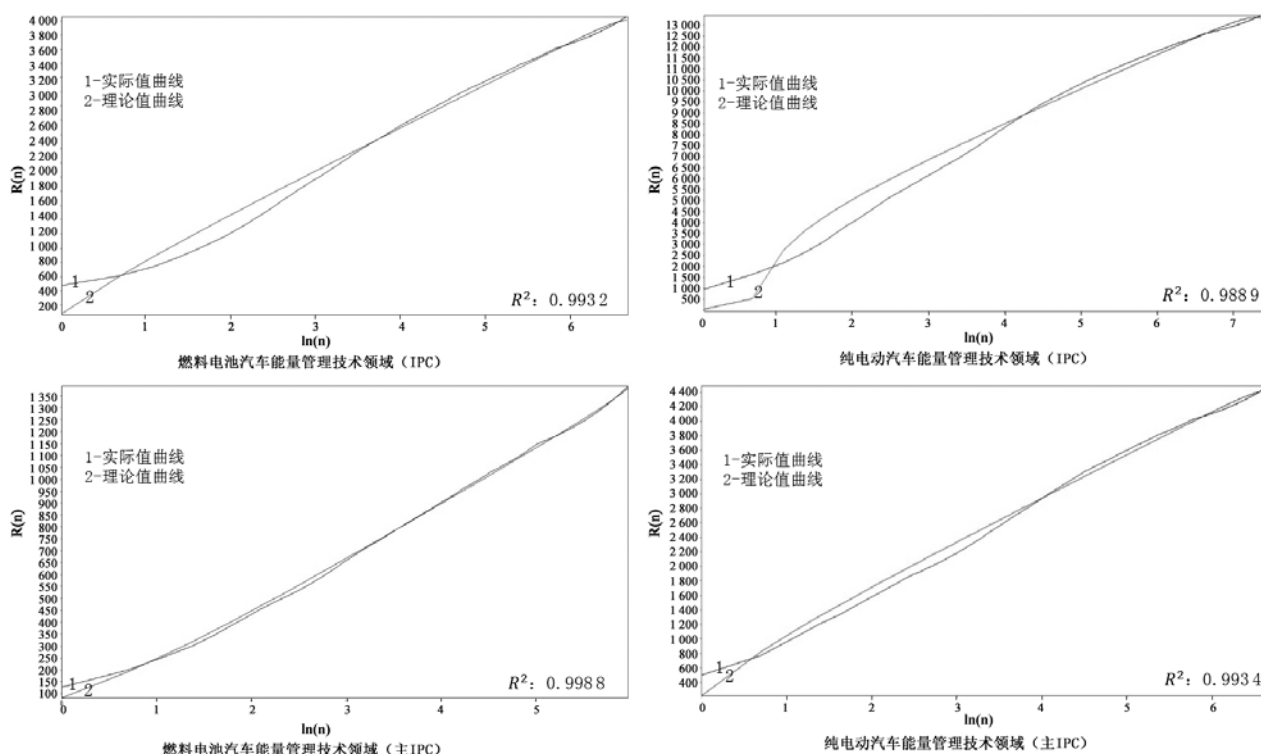


图1 IPC、主IPC划分角度下的方程拟合效果图对比

利用斯马里克夫统一方程对燃料电池汽车能量管理技术领域专利分散曲线的拟合, 分别得到IPC和主IPC划分角度下的基本拟合公式(3)和公式(4)。

$$R(x)=601.763\,65\lg(e^x-0.302\,88e^{-0.007\,53e^x})+88.974\,30 \quad (3)$$

$$R(x)=233.673\,51\lg(e^x+0.703\,33e^{-0.010\,89e^x})-42.772\,34 \quad (4)$$

对纯电动汽车能量管理技术领域专利分散曲线的拟合, 分别得到IPC和主IPC划分角度下的基本拟合公式(5)和公式(6)。

$$R(x)=1\,827.492\,15\lg(e^x-0.805\,58e^{-0.003\,18e^x})+3\,688.858\,42 \quad (5)$$

$$R(x)=743.389\,53\lg(e^x-0.674\,88e^{-0.005\,06e^x})+1\,541.106\,90 \quad (6)$$

标号1的曲线代表专利实际值分散曲线, 观察图1, 两种模式下的专利分散曲线均由一段下凹的曲线和上升的直线组成, 目前尚未出现格鲁斯下垂现象^[11], 符合经典的布氏分散曲线描述。根据周爱民的研究, 当图像不存在格鲁斯下垂时, 斯马里克夫统一方程的拟合效果比其他模型略好^[9], 因此, 该方程对于图形的描述具有一定准确性和适用性。

标号2的曲线代表斯马里克夫统一方程拟合的理论值分散曲线。分别对比燃料电池能量管理技术领域和纯电动汽车能量管理技术领域的曲线拟合效果可知: 在主IPC区域划分角度下的拟合效果图较IPC划分角度拟合效果图存在明显的差距收缩。一方面, 主要体现在专利分散曲线头部拟合效果的提升以及实际值和理论值间差距的减小; 另一方面, 体现在相关系数 R^2 在主IPC划分角度下均有所增大, 相关性有所提高。

从图1可见, 无论是在IPC还是主IPC区域划分角度下, 分布曲线图像均符合布氏定律, 即由一段曲线和上升的直线组成。专利分布在一定程度上服从布氏分布的规律特点, 但集中度更高, 主要体现在曲线头部的实际值要大于理论值; 从图像拟合效果看, 两种区域划分

角度下专利分散曲线均可用斯马里克夫统一方程拟合, 且主IPC划分角度下, 相关系数 R^2 增大; 斯马里克夫统一方程对两个领域的拟合效果均好于IPC区域划分方式, 在一定程度上说明主IPC区域划分角度更适合方程的拟合。

3.2 K-S检验对比研究

尽管上述相关系数 R^2 都在0.9以上, 相关程度较高, 但方程最终能否拟合实际分布状况要通过K-S检验来验证。分别计算实际累计频率和理论累计频率, 然后根据 $D=\text{Max}|(F_n(x)-F_0(x))|$ 找出最大临界值MaxD。表3和表4分别展示了在IPC、主IPC区域划分角度下, 燃料电池汽车能量管理技术领域和纯电动汽车能量管理技术领域的K-S检验情况。

燃料电池汽车能量管理技术领域经计算得到不同区域划分角度下D的拒绝临界值分别为 $D_{\text{IPC}}(788, 0.05)=0.048\,4$, $D_{\text{主IPC}}(392, 0.05)=0.068\,7$ 。

纯电动汽车能量管理技术领域经计算得到不同区域划分角度下D的拒绝临界值分别为 $D_{\text{IPC}}(2\,040, 0.05)=0.030\,1$, $D_{\text{主IPC}}(1\,037, 0.05)=0.042\,2$ 。

根据表3和表4, 将IPC、主IPC划分角度下各技术领域的最大临界值MaxD和D检验值整理得到K-S检验结果表(见表5)。按照K-S检验的思路, $\text{MaxD}>D(n, 0.05)$, 则拒绝原假设 H_0 , 认为不能通过K-S检验, 反之接受假设 H_0 。可以看出, IPC区域划分角度下的燃料电池汽车能量管理技术领域和纯电动汽车能量管理技术领域的MaxD均大于D检验值, 即不能通过检验, 因此, 可认为专利IPC总体分布并不符合斯马里克夫统一方程的分布; 而主IPC区域划分角度下的燃料电池汽车

表 3 燃料电池汽车能量管理技术领域的K-S检验数据表

IPC区域划分角度					主IPC区域划分角度				
n	R(n)	实际累计频率	理论累计频率	D检验值	n	R(n)	实际累计频率	理论累计频率	D检验值
1	276	0.068 0	-0.032 0	0.100 1*	1	128	0.092 1	0.059 5	0.032 6*
2	412	0.101 5	0.099 9	0.001 6	2	200	0.143 9	0.137 4	0.006 5
3	543	0.133 7	0.168 6	0.034 8	3	258	0.185 6	0.190 4	0.004 8
4	673	0.165 8	0.215 3	0.049 6	4	303	0.218 0	0.230 6	0.012 6
5	785	0.193 3	0.250 8	0.057 5	5	347	0.249 6	0.263 1	0.013 4
...	...	...	...	...	...	...	...	...	...
788	4 060	1.000 0	0.987 2	0.012 8	392	1 390	1.000 0	0.993 4	0.006 6

注: “*”表示最大临界值。

表 4 纯电动汽车能量管理技术领域的K-S检验数据表

IPC区域划分角度					主IPC区域划分角度				
n	R(n)	实际累计频率	理论累计频率	D检验值	n	R(n)	实际累计频率	理论累计频率	D检验值
1	1 644	0.095 0	0.038 8	0.056 2	1	941	0.141 1	0.104 7	0.036 4
2	3 209	0.185 5	0.231 6	0.046 1	2	713	0.248 0	0.261 9	0.013 9
3	3 848	0.222 5	0.295 9	0.073 5	3	303	0.293 4	0.324 6	0.031 2
4	4 455	0.257 5	0.335 6	0.078 1*	...	...	...	...	...
5	5 036	0.291 1	0.364 4	0.073 3	11	2 927	0.438 9	0.479 5	0.040 6*
...	...	...	...	...	...	...	...	...	...
2 040	17 298	1.000 0	0.986 6	0.013 4	1 037	6 669	1.000 0	0.990 4	0.009 6

注：“*”表示最大临界值。

表 5 K-S检验结果表

技术领域	IPC区域划分角度			主IPC区域划分角度		
	MaxD	D检验值	K-S检验	MaxD	D检验值	K-S检验
燃料电池汽车能量管理技术领域	0.100 1	0.048 4	未通过	0.032 6	0.068 7	通过
纯电动汽车能量管理技术领域	0.078 1	0.030 1	未通过	0.040 6	0.042 2	通过

能量管理技术领域和纯电动汽车能量管理技术领域的MaxD均小于D检验值，即通过检验，所以可以用斯马利科夫统一方程来描述其分布。从统计学角度进一步说明，多重分类号下的专利分布在整体上服从布氏形态分布，且主IPC角度下采用斯马利科夫统一方程描述专利信息离散分布的适用性要好于IPC角度。

4 结论

采用布氏定律研究方法揭示专利信息离散分布的特点与规律，对于确定核心专利技术、核心专利权人、制定专利检索策略、评价技术发展成熟度有重要的现实意义。通过图像拟合效果和K-S检验结果对多重分类号下的专利分布进行探讨，结果表明：专利信息同样存在集中与分散趋势，而且其分布曲线图像符合经典布拉德福分散曲线走势。因为斯马利科夫统一方程是基于布拉德福定律提出的反映科学信息离散分布的方程，所以通过斯马利科夫统一方程论证专利信息离散分布是否符合布氏定律是可行的。通过对IPC和主IPC不同区域划分角度方程适用性的对比研究，发现主IPC划分角度的图像拟合效果更好，且能通过K-S检验，因此，该分类角度下的专利分布更符合布氏形态分布，且多重分类号对专利布氏形态分布并未造成明显影响。

本研究从图像法的角度探讨专利布氏形态分布特

点，将描述布氏定律的斯马利科夫统一方程运用到专利主IPC的分散曲线描述中，并采用数学模型和假设检验的方法进行验证，进一步揭示将布氏定律研究思路运用到专利信息系统的可行性。同时本研究存在一定的局限性，仅选取同一技术领域的两个技术分支进行专利文献离散分布探讨，而未深入探讨技术发展成熟度对专利布氏分布的影响。未来研究工作将利用更多的技术领域实践验证该方程在专利分布中的适用性，还将探索核心专利权人的布氏分布特点。

参考文献

[1] 马费成,陈锐,袁红.科学信息离散分布规律的研究:从文献单元到内容单元的实证分析(I):总体研究框架[J].情报学报,1999,18(1):79-84.

[2] 张海燕.关于布拉德福定律及其应用的若干思考[J].情报探索,2013,194(12):19-21.

[3] 张鹏,刘平,唐田田等.布拉德福定律在专利分析系统中的应用[J].现代图书情报技术,2010(Z1):84-87.

[4] 吕义超,刘红光,王君.布拉德福定律在专利文献中应用的可行性研究[J].图书情报研究,2011,4(2):49-52.

[5] 胡晨希,邵蓉.基于布拉德福定律的药品核心专利分析[J].中国药事,2012,26(2):134-136,139.

[6] 罗爱静,尹瑾.基于信息分析的中药领域核心专利技术发展研究[J].情报杂志,2009(S1):37-39.

- [7] 邱均平.信息计量学[M]. 武汉:武汉大学出版社,2007:109-123.
- [8] 王崇德,赵艳.布拉德福分布解析式的择优评鉴[J].情报学报, 1998(4):66-71.
- [9] 周爱民.几种布拉德福分散曲线拟合模型的实证比较[J].情报杂志, 2013,32(1):59-62.
- [10] 何荣利,黄振文.关于布拉德福定律中的两个问题[J].中国科技期刊研究,2009,20(6):1078-1080.
- [11] GROOS O V. Bradford' s Law and the Keenan-Atherton data[J]. American Documentation,2007,18(1):46.

作者简介

苑朋彬, 男, 1990年生, 硕士研究生, 研究方向: 技术竞争情报, E-mail: yuanpb2014@istic.ac.cn。
赵蕴华, 女, 1967年生, 研究馆员, 研究方向: 重点科技领域研究, E-mail: zhaoyh@istic.ac.cn。

The Suitability Evaluation to Use Sri LankaMali Cove Unified Equation for Describing the Patent Information Discrete Distribution

YUAN PengBin, ZHAO YunHua
(Institute of Scientific and Technical Information of China, Beijing 100038, China)

Abstract: Adopting the image method, this paper mainly discusses the characteristics of patent distribution basing on the multi-classification. By comparing the fitting effect and K-S inspection result reference to Sri LankaMali eventually comes to the conclusion that under the division angle of "main IPC", the equation of applicability is higher and the influence of multi-classification to the distribution is smaller.

Keywords: Sri LankaMali Cove Unified Equation; Main IPC; Kolmogorov-Smirnov Inspection; Patent Discrete Distribution

(收稿日期: 2016-07-04)

■ 书 讯 ■

《汉语主题词表》(工程技术卷)

《汉语主题词表》自1980年问世以后,经1991年进行自然科学版修订,在我国图书情报界发挥了应有的作用,曾经获得了国家科学技术进步二等奖。为了适应网络环境下知识组织与数据处理的需要,2009年由中国科学技术信息研究所主持,并联合全国图书情报界相关机构,完成《汉语主题词表(工程技术卷)》的重新编制工作。

全书共收录优选词19.6万条,非优选词16.4万条,等同率0.84。在体系结构、词汇术语、词间关系等方面进行改进创新。为了方便工程技术领域不同专业用户使用,《汉语主题词表》(工程技术卷)按专业分13个分册出版,同时建立《汉语主题词表》服务系统,提供在线概念检索和辅助标引服务,通过可视化技术展示各类概念关系,是图书馆、档案馆、出版社、期刊杂志社、文献信息中心等专业工作者及科研、教育及工程技术领域人员必备的参考书。

《汉语主题词表(工程技术卷)》已于2014年由科学技术文献出版社出版,全书2300余万字,总定价3880元,可分册购买。