Statistics and Application 统计学与应用, 2022, 11(2), 262-272
Published Online April 2022 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2022.112027

Hans汉斯

基于文本挖掘的数字化社区建设政策方向分析

刘羽洋[1]，宋丽娜[2*]，潘雯秋[2]，宣欣祎[3]，李　莉[2]

[1]杭州电子科技大学卓越学院，浙江　杭州
[2]杭州电子科技大学经济学院，浙江　杭州
[3]杭州电子科技大学会计学院，浙江　杭州

收稿日期：2022年3月12日；录用日期：2022年3月25日；发布日期：2022年4月11日

摘　要

数字化改革的浪潮促进了数字化社区的发展，从宏观上把握数字化社区建设的方向可以为其建设及管理提供指导与帮助，对于加快建设数字化社会具有重要意义。本文从各政府网站上获取我国2016~2021年期间数字化社区建设的相关政策文本，利用python3.8对政策典型文本数据进行分词及词频统计；从中抽取关键词，建立共词网络，通过k-means聚类算法对关键词进行聚类，根据结果将数字化社区建设相关政策分为社区基础建设、社区治理、社区服务、社区管理等四类方向，并通过关键词对比验证了其合理性；最后利用Gephi软件将分类结果绘制成词云图，对四个方向的建设提出了不同的建议。

关键词

数字化社区，共词网络，k-means聚类算法，建设方向

Analysis on the Policy Direction of Digital Community Construction Based on Text Mining

Yuyang Liu[1], Lina Song[2*], Wenqiu Pan[2], Xinyi Xuan[3], Li Li[2]

[1]College of Zhuoyue Honors, Hangzhou Dianzi University, Hangzhou Zhejiang
[2]College of Economics, Hangzhou Dianzi University, Hangzhou Zhejiang
[3]School of Accounting, Hangzhou Dianzi University, Hangzhou Zhejiang

Received: Mar. 12th, 2022; accepted: Mar. 25th, 2022; published: Apr. 11th, 2022

Abstract

The wave of digital reform has promoted the development of digital communities. Grasping the di-

*通讯作者。

文章引用: 刘羽洋, 宋丽娜, 潘雯秋, 宣欣祎, 李莉. 基于文本挖掘的数字化社区建设政策方向分析[J]. 统计学与应用, 2022, 11(2): 262-272. DOI: 10.12677/sa.2022.112027

rection of digital community construction from a macro perspective can provide guidance and help for its construction and management, which is of great significance for accelerating the construction of a digital society. This paper firstly obtains the relevant policies of my country's digital community construction from 2016 to 2021 on various government websites, uses python3.8 software to perform word segmentation and word frequency statistics on typical policy text data, and then extracts keywords from it to establish a co-word network, and through the k-means clustering algorithm clusters the keywords, and according to the results, the policies related to digital community construction are divided into four categories, namely community construction, community governance, community service, and community management, and verify its rationality through keyword comparison. Finally, Gephi software is used to draw the classification results into word cloud map, and put forward different suggestions for the construction of the four directions.

Keywords

Digital Communities, Co-Word Network, k-means Clustering Algorithm, Construction Direction

1. 引言

伴随着新一轮科技革命和产业变革持续推进，数字技术正以新理念、新业态、新模式全面融入人类经济、政治、文化、社会、生态文明建设各领域和全过程，给人类生产生活带来广泛而深刻的影响，同时现代科学信息技术的迅猛发展与世界网络化进程的不断加快，使得信息化社会逐渐兴起，社区的数字化建设就成为社会信息化的重要组成部分，成为城市信息化的“细胞工程”，它使得信息技术在城市发展中得到普及和运用[1]，因此数字化社区应运而生，但如今其建设还未普及开来，如何更好地建设数字化社区就成为了一个非常值得关注的问题。

数字化社区建设、社区的数字赋能应用是一些学者的重要研究内容。目前学界主要从数字化社区物业管理建设(金炜玲等，2021)、教育学习地图构建(于莎和刘奉越，2019)、存在的问题及相应对策(许咏华，2019)、通讯管理平台技术(彭蕾，2004)、公众满意度(黄辉庆，2013)等角度探究数字化社区的建设。与已有文献不同，本文对数字化社区建设的政策进行探究，政策是一切实际工作的依据与参考标准，只有对政策进行深度研究，才能更好地将政策付诸于实践。

本文利用文本挖掘和聚类的方法对数字化社区相关政策进行深入研究，通过挖掘国家发布的针对数字化社区的政策文本以及和数字化社区有关的新闻，分析数字化社区的应用方向，以此对我国的数字化社区建设提供一定的建议，推动数字中国的可持续性建设。结果表明，关于数字化社区建设的政策新闻可分为社区建设、社区治理、社区服务、社区管理四个方面。在社区建设方面，社区负责人应紧跟政府节奏，着力建设数字化相关设备平台；在社区治理方面，社区负责人应与相关部门一起形成数字化社区运行体系；在社区服务方面，数字化社区应向智能化、便捷化、现代化发展；在社区管理方面，数字化社区应保证社区安全，形成高效严谨的管理系统。

2. 文献综述

数字化社区自出现以来，一直备受学者关注，如何建设更好的数字化社区成为各学者的研究问题。金炜玲、李熠、李佳(2021) [2]基于政务热线大数据，对数字化社区的物业管理进行创新；于莎和刘奉越

(2019) [3]通过构造数字化社区学习地图，加强社区教育资源建设；许咏华(2015) [4]分析了南宁市社区的硬件设施和服务平台建设的现状，并探究信息化发展的困境及原因，提出了相应的建议；贾炜(2006) [5]以上海市江苏路街道为例，研究了上海市社区信息化的困境和出路；陆伟良(2005) [6]对江苏省数字化社区的建设情况、建设技术、建设问题等进行研究；朱琳和沈夏燕(2014) [7]以上海宝山区若干小区为例，建立在社区信息化影响下社区满意度的测量模型，通过实例调查，分析社区信息化对社区满意度的具体影响并提出相关对策建议。

综上可知，多数学者都是探究数字化社区的建设现状、建设技术、建设问题与解决方案等，少有学者对数字化社区建设的政策进行研究，但政策是一切实际工作的依据与参考指南，是重中之重，只有对政策进行具体的、有针对性的研究，才能更好地推进数字化社区的建设。

目前对政策文本的分析最常用的方法是文本挖掘技术，1995 年，Feldman 等首次提出文本挖掘一词，Justicia 等(2018) [8]对文本挖掘及其应用进行了概述，并对文本挖掘的优势和贡献进行了总结。文本挖掘也称文本数据挖掘，通常是指从非结构化文本文档中提取有趣且非常规模式或知识的过程，可被视为数据挖掘或来自结构化数据库的知识发现的扩展[9]。近年来，文本挖掘被广泛应用于各个领域的文本分析上，华斌等(2022) [10]利用共词分析、LDA 主题建模与相似度计算三种技术对高新技术产业政策文本进行挖掘，得出了高新技术产业政策的变化规律；谢宇等(2022) [11]利用文本挖掘并结合对应分析对当代科学技术伦理研究的态势进行探究；刘家兵等(2021) [12]利用文本挖掘探究局限期小细胞肺癌(LS-SCLC)放疗领域的研究热点；黄晓斌(2009) [13]利用文本挖掘对网络舆情的真实性、关联性和产生原因等进行分析。

由此可见，文本挖掘技术已经较为成熟，多数应用在政治、医疗、社会治理等领域，可用于政策分析、热点研究、演变规律探究等。

对于政策主题的研究，张宝建[14]等利用文本挖掘技术对国家科技创新政策文本内容提取关键词，建立共词网络并聚类，对每一类进行定义，分为了专利类、科技类等八个主题；杨锐[15]等对我国科研诚信政策文本抽取关键词进行聚类，并对各类自行归纳为机制完善与规范等九个主题；唐恒[16]等对中小企业知识产权政策的关键词进行聚类，并根据聚类结果将其分为了九个主题。

综上可知，多数学者都是采用词频统计及聚类的方法研究政策主题，并对聚类后的每一类结果进行总结定义，因此本文也利用该方法，将主题拓展至数字化社区政策，对政策主题即社区建设方向进行分析，与已有文献不同的是，本文在对聚类后的每一类结果进行定义后，又对定义的合理性进行了验证。

3. 研究设计

3.1. 研究方法

本文利用文本挖掘和聚类对相关政策进行深入研究，通过挖掘针对数字化社区的政府政策及与相关新闻，用分词及高频词分析的方法提取其中能反应文本特点的关键词，对诸多文本的关键词使用 TF-IDF 方法构建词权重，通过 k-means 聚类分析数字化社区的功能及应用方向，文本聚类流程图如图 1 所示。

3.2. 数据来源

由于数字化社区还未在全国普及，国家社区大多仍处于逐渐转型阶段，为了保证政策的时效性与权威性，本文对政府网站发布的法律法规政策文件以浏览、关键词查找等方式检索了近 5 年国家层面出台的全部与数字化社区有关的政策文件，保证数据资料全面可靠、详尽真实。同时，以全国各地区针对数字化社区的地方性政策细则为依据，以数字化社区相关新闻报道为拓展，通过源头查找和文件追溯补充政策文本遗漏数据。

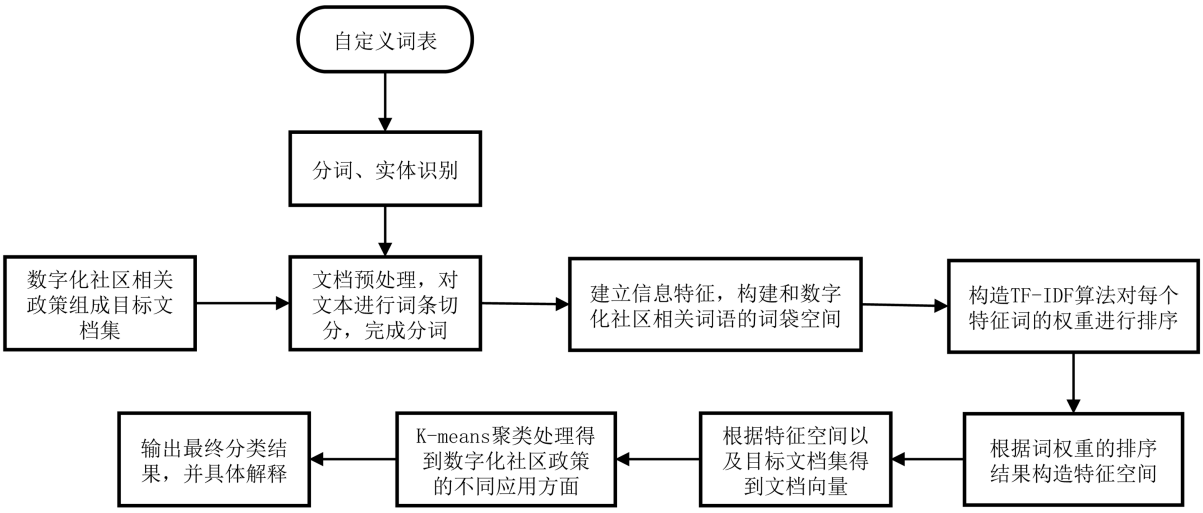

Figure 1. Text clustering flowchart
图 1. 文本聚类流程图

在数据收集过程中，对搜集的政策文本进行了逐条筛选和剔除，对新闻报道的篇幅较短、内容不贴近的文本进行筛选并剔除，保障了所选文本的可靠性与多样性，也使结果更加准确、更有说服力。

3.3. 研究过程

3.3.1. 分词及词频统计

本文首先对文本进行分词处理，使用 Python 中的 jieba 第三方库进行分词和频数统计，在分词过程中本文使用现有的哈工大停用词库去除停用词并构建词袋空间 VSM 来统计词频，并根据词性去除了名词和动词以外的词语，避免了无关词语对分词结果的影响。其中，词汇出现频率的高低与词义在文本中的重要性成正比，统计出的关键词可以有效地反应该文章的中心。

3.3.2. 关键词抽取

本文使用关键词权重计算算法 TF-IDF 来对每篇政策文章进行关键词抽取。TF-IDF 是一种基于词袋模型的关键词抽取方法，在文本挖掘中被广泛用于评估一个词语对文本的重要程度，从而提取其中的关键词[17]。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。因此，TF-IDF 倾向于过滤掉常见的词语，保留重要的词语。TF-IDF 计算公式如下：

$$IDF(x)=\log\left(\frac{N+1}{N(x)+1}\right)+1 \tag{1}$$

$$\text{TF-TDF}(x)=TF(x)\times IDF(x) \tag{2}$$

其中，N 代表语料库中文本的总数，$N(x)$ 代表语料库中包含词 x 的文本总数，$TF(x)$ 代表某一个给定的词语在该文件中出现的频率。

本文根据 TF-IDF 算法抽取出每篇政策文章的前十个关键词，并进行整合，对于汇总的关键词进行词频统计，结果如表 1 所示：

由表 1 可得，这些政策和相关新闻中出现频率最多的词是社区和数字化，与本文主题十分贴近，其次是服务、居民、平台、设施、智能等，由此可见，大部分政策都是针对居民旨在为其提供更好的服务，建设数字化设备使得社区更加智能化。

Table 1. Partial vocabulary and word frequency statistics
表 1. 部分词汇及词频统计

序号	词汇	词频	序号	词汇	词频
1	社区	378	11	政府	146
2	数字化	359	12	便捷	107
3	服务	249	13	需求	118
4	居民	213	14	安全	96
5	平台	188	15	设备	192
6	设施	169	16	办事	122
7	建设	187	17	赋能	133

3.3.3. 共词网络

在词频统计分析的基础上，为判断从国家针对数字化社区的相关政策中提取出的关键词之间关联度的强弱，本文以每对词语在同一政策文献中出现次数为根据构建共词矩阵，样例如表 2 所示。通过共词矩阵的构建[18]可以反映出文本主题间的关系，为后续聚类建模提供依据。

Table 2. Partial co-word matrix
表 2. 部分共词矩阵

	社区	数字化	服务	居民
社区	0	29	31	17
数字化	29	0	4	20
服务	31	4	0	29
居民	17	20	29	0
生活	30	30	10	19
设施	23	27	9	10
智能	13	26	6	15
平台	21	28	5	13
管理	26	17	9	2
建设	27	6	35	25
政府	24	35	12	16
便捷	9	11	14	10
需求	18	7	13	6
安全	9	12	5	4
设备	18	33	7	3
办事	27	3	32	15
赋能	15	23	4	2
系统	15	26	13	2
在线	14	10	5	2
创新	12	12	6	3

在共词分析过程中，共词矩阵通常用来表示词语间两两共词的频数，根据共词矩阵可以构建共词网络，通过共词网络体现词与词间的关联[19]，网络节点间位置关系可以反映关键词间的紧密程度。本文利用 Gephi 软件导入共词矩阵形成共词网络如图 2 所示：

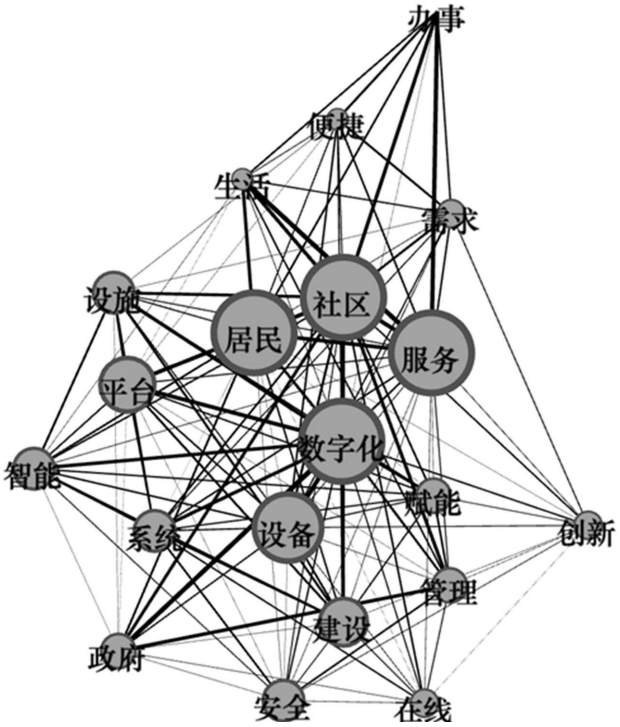

Figure 2. Keyword co-word network diagram
图 2. 关键词共词网络图

3.3.4. 聚类

关键词聚类能够凸显各关键词间的类别联系，聚类分析根据关键词间共词强度将关键词划分为不同类别[15]。本文使用 k-means 聚类算法对共词矩阵中的关键词进行聚类分析。为了确定聚类簇数 k，本文通过测算政策文本与提取关键词之间的相互距离和平均距离寻找度中心性最大的节点作为初始聚类中心，以优化初始聚类中心选择，然后采用依次剔除度中心性最高的节点来找出其他的中心点，再进行多次迭代计算直至找出第 k 个中心点，在本文中采用“手肘法”来确定聚类簇数 k，“手肘法”的核心指标是误差平方和，其公式为

$$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2 \tag{3}$$

其中，C_i 是第 i 个簇，p 是 C_i 中的样本点，m_i 是 C_i 中所有样本的均值，SSE 是所有样本的聚类误差，代表了聚类效果的好坏。基于政府政策数据，根据误差平方和公式可以算出 $k = 4$ 时曲线有明显转折，因此本文取 k 值为 9，即对提取的关键词通过聚类的方法分成了 4 类。

4. 结论与建议

4.1. 主要结论

从整体来看，这些政策和新闻均与数字化社区紧密相连，其中出现频率最多的词是社区和数字化，

与本文主题十分贴近，其次是服务、居民、平台、设施、智能等，由此可见，大部分政策都是针对居民旨在为其提供更好的服务，建设数字化设备使得社区更加智能化。

针对每一类结果，本文根据各类中的关键词对主题进行定义，如在第一类中，“建设”、“设备”、“系统”等反映社区基础建设的关键词出现频率较高，因此将该类定义为社区基础建设方向；同理，将其他三类定义为了“社区治理”、“社区服务”、“社区管理”。

四类结果绘制词云图如图 3~6 所示：

Figure 3. The first category—community building word cloud figure
图 3. 第一类——社区基础建设词云图

Figure 4. The second category—community governance word cloud figure
图 4. 第二类——社区治理词云图

Figure 5. The third category—community service word cloud figure
图 5. 第三类——社区服务词云图

Figure 6. The fourth category—community management word cloud figure
图 6. 第四类——社区管理词云图

这四类的范畴和要义如表 3 所示。

4.2. 结论验证

由于对四个建设方向的定义存在主观性，因此本文通过查找有关“社区基础建设”、“社区治理”、“社区服务”、“社区管理”的具体政策，从中提取关键词，与以上四类结果进行比对，数据来源于全国各地政府网站。特定的四类相关政策分析出的关键词词频统计结果如表 4 所示。

Table 3. Thematic analysis of digital community construction policy
表 3. 数字化社区建设政策的主题分析

类别	范畴	要义
社区基础建设	基础设施、数字化平台搭建、信息共享等	社区建设类的政策以社区基础设施、平台搭建等硬件设备为主体，要求完善社区人文环境
社区治理	社区制度规范、社区秩序、数字运行等	社区治理类的政策为政府、社区组织、居民及辖区等单位解决社区存在的问题，完成特定的、具体的经济社会发展任务进行指导
社区服务	生活便民服务、政府便民服务、精神文化建设等	社区服务类的政策明确社区负责人和物业等组织应为社区居民提供的各类便民服务，以及完善精神文化建设
社区管理	小区安全、家庭安全、物业管理等	社区管理类的政策明确了社区内部机构、组织等如何进行自我管理或行政管理活动，以维护社区正常运行秩序

Table 4. Word frequency statistics for specific policies
表 4. 特定政策词频统计

社区基础建设		社区治理		社区服务		社区管理	
建设	196	治理	231	居民	202	管理	186
信息	185	政府	175	服务	174	物业	164
设施	164	机构	165	办事	137	巡查	159
未来	138	活动	154	城乡	128	建立	143
数字	121	工作	135	体系	119	登记	127
发展	90	开展	131	规划	100	管家	99
生活	71	人大	110	群众	75	业主	84
打造	52	城市	68	功能	46	监控	72
目标	42	主题	55	强化	33	党员	56
城市	20	街道	42	党建	19	小区	38

由结果可知，每一类的关键词均与本文根据聚类结果所定义的四类数字化社区建设方向的关键词存在较高的重合度，即可以认为本文所定义的四类方向是合理的。

其中针对有关社区基础建设方向的政策，如“某镇健康社区建设工作计划和实施方案”等通过词频统计得出主要的关键词有建设、信息、设施等，与第一类关键词设施、系统、信息等重合度高，联系紧密，因此将第一类定义为社区基础建设方向是合理的。

针对有关社区治理方向的政策，如“某省民政厅关于加强社会工作专业人才队伍建设加快推进社会工作发展的意见”等通过词频统计得出的关键词主要是治理、政府、机构等，与第二类关键词发展、政府、体系、机构重合度高，因此将第二类定义为社区治理方向是合理的。

针对社区服务方向的政策，如新华社发布的“更好解决人民群众操心事、烦心事、揪心事——聚焦《‘十四五’城乡社区服务体系建设规划》”等经过词频统计得出关键词主要为居民、服务、办事等，与第三类关键词教育、生活、智慧、服务、文明、养老、办事、居民重合度高，因此将第三类定义为社区服务的方向是合理的。

针对社区管理方向的政策，例如“某市以红色管家引领社区管理优化提升”等，提取后主要的关键

词有物业、巡查、管理等，与第四类关键词物业、报警、设施、监控、管理、巡查、登记重合度高，联系紧密，因此将第四类定义为社区管理的方向是合理的。

4.3. 对策建议

通过对词云图的分析可以看出，在社区建设方面，设施、平台、系统、信息、设备等关键词十分突出，设施、设备等关键词可以说明更应该看重社区的数字化基础设施建设，平台、系统、信息等关键词说明应当尽快完善整个数字化平台与系统。只有建设好数字化平台，才能让更多的居民享受到数字化的福利，也能更好地推进国家数字化进程。当下科技发展迅速，数字化已经成为必然趋势，在社区方面，社区负责人要紧跟政府节奏，建设好数字化平台，让社区向数字化社区迈进，进而更大程度地满足居民的需求。

在社区治理方面，发展、政府、体系、机构、治理等关键词尤为突出。从政府、机构等关键词可见，数字化社区的治理离不开政府机构的参与，数字化社区的建设可以让政府高效治理社区；从发展、体系等关键词可见社区的治理是数字化形式且成体系的。加快建设数字化社区可以更有效地帮助政府治理社区，提高治理的效率。以数字化的形式可以更便捷、更全面地解决社区存在的问题。

在社区服务方面，教育、生活、智慧、服务、文明、养老、办事、居民、便捷等关键词很突出。这些关键词大多关乎到社区内的民生，和居民以及服务息息相关。这可以反映出数字化社区的建设极大地便捷了居民生活的方方面面。从儿童的教育到老人的养老再到居民日常办事，数字化平台都可以很好地帮助他们。因此，建设数字化社区对居民生活有着很大改善，不仅可以高效地解决居民生活中遇到的很多问题，还可以让社区更加智能、更加现代化、更加符合当前社会的发展趋势。

在社区管理方面，物业、报警、设施、监控、管理、巡查、登记这些关键词非常突出。报警、设施、监控等关键词体现了社区在数字化上的管理方向。物业、报警等关键词可以反映出社区的智能报警系统、智能监控系统都离不开数字化。有了数字化平台，人们还可以网上登记、网上填写资料等，方便了物业的管理，减少了人力物力的浪费。通过数字化的管理方式，可以让物业管理者和居民更加轻松。做好物业的管理，可以提高物业的声誉，提升小区的价值。当下，越来越多的公司采用数字化管理手段，社区也应该完善数字化管理，落实政府政策，符合社会发展趋势，让居民的生活从点到面都能体现方便、高效的数字化形式。

参考文献

[1] 黄辉庆. “数字社区”建设公众满意度指数模型构建及实证研究[D]: [硕士学位论文]. 湘潭: 湘潭大学, 2013.

[2] 金炜玲, 李熠, 李佳. 数字化社区治理: 应用政务热线大数据创新物业管理[J]. 电子政务, 2021(2): 27-37.

[3] 于莎, 刘奉越. 行动者网络下数字化社区学习地图构成及运行机制[J]. 现代远距离教育, 2019(1): 83-89.

[4] 许咏华. 南宁市社区信息化的现状、问题与对策研究[D]: [硕士学位论文]. 南宁: 广西大学, 2015.

[5] 贾炜. 上海市社区信息化建设的困境与出路研究——以上海市江苏路街道为例[D]: [硕士学位论文]. 上海: 华东师范大学, 2006.

[6] 陆伟良, 彭蕾. 对江苏省数字社区发展对策的探讨[J]. 江苏建筑, 2005(1): 59-62.

[7] 朱琳, 沈夏燕. 社区信息化对居民满意度影响的实证研究——以上海市宝山区若干小区为例[J]. 电子政务, 2014(2): 17-28.

[8] Justicia de la Torre, C., *et al*. (2018) Text Mining: Techniques, Applications, and Challenges. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, **26**, 553-582. https://doi.org/10.1142/S0218488518500265

[9] Tan, A.H. (1999) Text Mining: The State of the Art and the Challenges. *Proceedings of the PAKDD* 1999 *Workshop on Knowledge Discovery from Advanced Databases*, Vol. 8, 65-70.

[10] 华斌, 康月, 范林昊. 中国高新技术产业政策层级性特征与演化研究——基于 1991-2020 年 6043 份政策文本的

分析[J]. 科学学与科学技术管理, 2022, 43(1): 87-106.

[11] 谢宇, 孔燕. 当代科学技术伦理研究的态势分析——以《科学与工程伦理》文本挖掘为例[J]. 自然辩证法通讯, 2022, 44(1): 85-92.

[12] 刘家兵, 林桂, 倪渊, 傅小龙. 基于文献挖掘局限期小细胞肺癌放疗热点的研究[J]. 中华肿瘤防治杂志, 2021, 28(21): 1660-1665.

[13] 黄晓斌, 赵超. 文本挖掘在网络舆情信息分析中的应用[J]. 情报科学, 2009, 27(1): 94-99.

[14] 张宝建, 李鹏利, 陈劲, 郭琦, 吴延瑞. 国家科技创新政策的主题分析与演化过程——基于文本挖掘的视角[J]. 科学学与科学技术管理, 2019, 40(11): 15-31.

[15] 杨锐, 杨亮, 李良强, 张楠, 廖觅燕. 我国科研诚信政策特征及演化逻辑——基于文本挖掘法[J]. 科技进步与对策, 2020, 37(20): 89-98.

[16] 唐恒, 高清, 孙莹琳, 肖寒姿. 基于文本挖掘的中小企业知识产权政策研究——来自中央层面的数据[J]. 科技管理研究, 2022, 42(1): 92-100.

[17] 吴晓秋, 吕娜. 基于关键词共现频率的热点分析方法研究[J]. 情报理论与实践, 2012, 35(8): 115-119.

[18] 伍若梅, 孔悦凡. 共词分析与共引分析方法的比较研究[J]. 情报资料工作, 2010(1): 25-28.

[19] 张燕刚, 成全. 基于共词分析的我国乡村振兴与田园综合体政策研究[J]. 农村经济与科技, 2019, 30(13): 25-29.