基于文本挖掘的中国跨境电商政策量化研究

施寒潇，毛郁欣

（浙江工商大学 管理工程与电子商务学院，浙江 杭州 310018）

摘　要：跨境电商政策中蕴含着大量引导跨境电商发展的重要信息，这些信息对地方政府和企业规划跨境电商发展具有重要意义。文章采用文本挖掘的方法开展跨境电商政策的量化分析和研究，通过从跨境电商政策文本中提取特征，再进一步使用聚类算法对特征进行聚类，基于聚类结果分析了跨境电商政策关注和聚焦的重点内容，主要包括知识产权、税收、产品等方面。此外，还通过语义网络分析和展示了政策重点内容之间的关联关系。

关键词：跨境电商；政策研究；文本挖掘；特征提取

中图分类号：F49　　**文献标志码：**A　　**文章编号：**1000－2154(2022)11－0005－13

DOI：10.14134/j.cnki.cn33-1336/f.2022.11.001

Quantitative Research on China Cross-border E-commerce Policies Based on Text Mining

SHI Hanxiao, MAO Yuxin

(*School of Management and E-Business, Zhejiang Gongshang University, Hangzhou 310018, China*)

Abstract: Recently, the cross-border e-commerce policies in China contain a lot of important information that guides the development of cross-border e-commerce. The information is of great significance to regional governments and enterprises to plan cross-border e-commerce development. In order to overcome the inadequacy of the existing cross-border e-commerce policy research, we propose to use the text mining approach to carry out quantitative analysis and research on the cross-border e-commerce policy. We try to extract the feature from the policy documents and use some clustering algorithm to cluster the features. Based on the clustering results, we try to analyze and find out the key points of the cross-border e-commerce policy, mainly including intellectual property, taxation, products and so on. At the same time, the relationship between key points of the policies is also investigated and displayed through semantic network analysis.

Key words: cross-border e-commerce; policy research; text mining; feature extraction

一、引　言

跨境电子商务（简称跨境电商）是指分属不同关境的交易主体，通过电子商务平台达成交易、进行电子支付结算，并通过跨境电商物流及异地仓储送达商品，从而完成交易的一种国际商业活动。跨境电商作为一种贸易新业态，正广泛而深刻地影响全球贸易格局。中国跨境电商产业的发展在世界范围内处于领先地位，且已经成为拉动经济增长的新引擎[1-3]。跨境电商目前已经成为我国发展速度最快、潜力最大、带动作用最强的外贸新业态，而跨境电商的快速发展又离不开政策的支持。自2013年以来，我国各级政府和有

收稿日期：2022-06-20

基金项目：浙江省软科学研究计划项目“基于文本挖掘和国内外跨境电商政策量化分析与演进研究”(2021C35128)

作者简介：施寒潇，男，教授，工学博士，主要从事文本挖掘、电子商务研究；毛郁欣，男，教授，工学博士，主要从事社交网络分析、电子商务研究。

关部门密集出台了一系列支持发展跨境电商的政策,政策普遍具备很强的实操性,极大地促进了跨境电商行业的规范发展。跨境电商政策中蕴含着大量指导支持跨境电商发展的重要信息,这些信息对地区和电商企业规划跨境电商发展有重要意义。然而,各级政府部门发布的跨境电商政策较为分散,并不完全统一,再加上跨境电商本身属于新兴行业,相应的政策也会随着行业和时代的发展而迭代,政策内容呈现出较为明显的动态变化的特征。因此,运用科学的方法对不同区域和不同部门出台的跨境电商政策内容进行研究和分析[4],具有十分现实的意义,科学合理的政策有助于推动产业的良性健康发展。然而,从信息学的角度来看,跨境电商政策本身属于自然语言描述的无结构文本,如果单纯依靠人工的方式进行分析,当政策文本的规模较大时,分析的效率必然会下降,而且容易出现疏漏。因此,运用信息技术特别是文本挖掘技术对政策文本进行量化研究和分析是一个比较可行的解决方案。

二、国内外研究现状

目前,国内外已有一些学者开展了与电商或跨境电商政策相关的文本分析研究,和国内的研究成果相比,由于国情和行业差异,国外直接针对跨境电商政策的研究还不多见,或者说,国外学者并未严格区分“一般电商”和“跨境电商”,内容上以电商政策的定性研究为主。而根据研究方法的不同,可以大致将现有研究分为以下三类:

1. 定性或者宏观层面研究。邢光远等(2020)[5]对“一带一路”倡议下中国跨境电商的政策演进与发展态势进行了研究和分析。徐德顺(2021)[6]分析了后疫情时代中国跨境电商面临新的挑战,并给出了政策建议。张晚冰(2021)[7]分析了提出了跨境电商零售出口的体系,并研究了政府政策影响体系的具体路径。Richards 和 Farrokhnia(2016)[8]运用扎根理论对电商政策进行研究,通过具体案例重点研究了世贸组织电商政策面临的困境。Hanna(2016)[9]主要研究了政府电商政策对于企业尤其是中小企业创新的影响。

2. 基于统计方法的定量研究。赵杨等(2018)[10]应用 PMC 指数模型评价方法,通过构建投入产出表计算出单一政策的 PMC 指数得分,对我国跨境电子商务具体政策的实施效果进行重点评价与分析。熊励等(2022)[11]应用多期双重差分模型从国家层面设立综合试验区及地方层面实施政策两个角度评价政策效应,并探讨不同综合试验区的跨境电商政策效应差异。邱国斌等(2022)[12]采用模糊集定性比较分析方法,探究了跨境电商发展各个指标之间相互作用的逻辑关系。Roberta 等(2019)[13]则通过 Agent 模拟的方式研究城市货运政策对电商发展的影响。Lin 等(2011)[14]主要研究了 B2B 电商政策、IT 成熟度和评价实践之间的契合度,及其对电商绩效的影响。

3. 基于文本挖掘的定量研究。和传统的统计方法相比,这类方法更有利于挖掘潜在的规律和特征。李泓烨等(2021)[15]应用政策工具、政策力度、政策主题三维度分析框架,对跨境电商政策文本进行分析,并提出了政策优化建议。钮钦(2016)[16]、侯振兴和闾燕(2017)[17]都采用了内容分析法,前者从政策工具和商业生态系统维度对中国农村电商相关的中央政策文本进行分析,在培育农村电商生态系统方面提出建议;后者从政策工具和农产品生态系统维度对农产品电商发展政策进行分析,发现政策中存在的不足。盛赟等(2019)[18]从文本制定主体、文本类型、文本内容及政策工具角度,对浙江省跨境电商物流相关的政策文本进行分析,为跨境电商物流政策在主体单一和联动性不足等问题上提出改进意见。余传明等(2018)[19]运用主题时间模型,通过计算不同年份下主题的平均强度并提取每个主题下概率高的词汇,分析农村电商扶贫政策内容的演化情况和政策的区域差异性。金珺(2020)等[20]运用文本挖掘工具和社交网络分析工具对省级农村电商政策进行了研究。肖开红等(2019)[21]采用词频分析、共词分析、社会网络分析与文本挖掘等分析方法,对中国涉农电商政策的演进进行研究。

和其他产业相比,跨境电商作为新兴的业态,相关的研究成果正在不断涌现,但是直接针对电商或跨境电商政策文本分析的研究还比较少,特别是利用文本挖掘方法进行的研究工作则更加缺乏,因此,目前这一方向上的研究尚不成熟。而从研究方法上来看,现有的政策文本挖掘研究主要集中在特征提取、聚类、分类以及主题提取等方面。针对现有的跨境电商政策研究的不足,本文提出采用文本挖掘的方法开展跨境

电商政策的量化分析和研究。深入分析中国跨境电商政策,有助于我们更好地理解政策的重点,从而把握跨境电商行业的发展趋势。

三、跨境电商政策文本分析框架

本研究主要基于文本挖掘方法进行跨境电商政策文本分析,而文本挖掘分析的重要环节是挖掘方法的选取以及挖掘流程的设计。本研究设计的跨境电商政策文本挖掘的流程如图1所示。主要按照以下步骤进行基于文本挖掘的跨境电商政策文本分析和研究:

(1)首先进行政策文本语料库的构建,以及文本预处理,从而降低后续采用文本挖掘方法分析政策文件提供符合输入要求的数据;

(2)基于 TF-IDF 算法提取初始特征,重点抽取名词和名词短语,从而降低文本挖掘特别是聚类分析的复杂度,提升分析效果;

(3)PMI 算法特征过滤及人工特征筛选,得到特征集合,从而在步骤(2)的基础之上进一步缩小文本处理的范围,提升分析效率;

(4)基于 Word2Vec 训练模型的特征向量化,便于开展基于特征向量的文本聚类;

(5)K-Means 特征聚类形成特征聚类集合,根据聚类结果分析政策文件的关注和聚焦的重点内容;

(6)基于 ROSTCM 进行语义网络分析,从另一个视角分析和展示政策重点内容之间的关联关系,作为步骤(5)结果的补充;

(7)基于聚类分析和语义网络分析的结果,综合形成跨境电商政策文本分析的最终结果,并形成对策和建议。

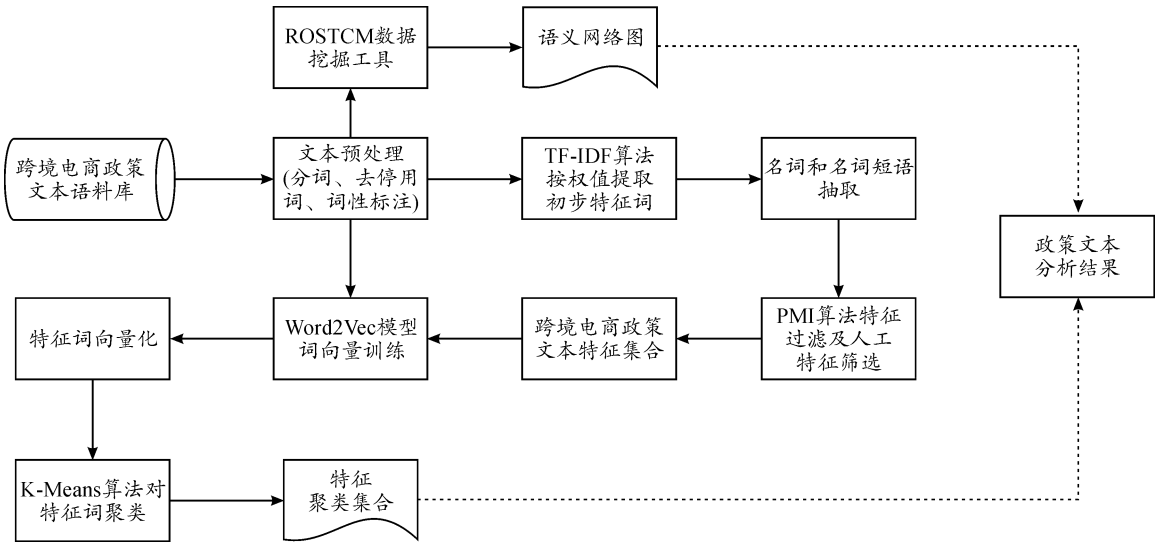

图 1　跨境电商政策文本挖掘的流程

(一) 文本预处理

在建立跨境电商政策文本语料库的前提下,本研究利用开源分词工具来完成文本预处理,主要包括以下三个步骤:

1. 使用分词工具对原始政策文本内容进行分词,并对分词结果中不理想的词进行修正。如“海外仓”是一个词,但自动分词的结果是“海外”“仓”两个词;此外,还有专有名词的修正。如“一带一路”是丝绸之路经济带和21世纪海上丝绸之路的简称,有特定含义,但分词的结果是“一带”“一路”。

2. 加载中文停用词表,对分词后的文本进行停用词过滤。停用词即常出现的词,且词本身不涉及关键

信息,如“一个”“一则”“的”等。

3. 对经过分词和去停用词的文本进行词性标注,将词语标记上词性符号,重点对名词进行标记。

(二) 特征提取

文本预处理之后会产生很多的特征词,如果直接使用预处理后的特征词进行挖掘,不但会造成特征表示上的维度灾难,而且也得不到高质量的聚类结果。因此,特征提取在文本挖掘中十分重要,好的特征提取结果可以给后续的挖掘以及最后的聚类结果带来更好的效果。

1. TF-IDF 算法提取初步特征词。词频-逆文档频率 TF-IDF(Term Frequency-Inverse Document Frequency)被广泛运用于特征词的权重计算[22]。本研究使用 TF-IDF 来计算政策文本中特征词的权值,按权值大小排序,并选择 TF-IDF 值超过特定阈值的特征词作为初始特征。此外,由于同一个特征词在不同的政策文件中会重复出现且权值不同,故同一个特征词取最大的 TF-IDF 值作为权值,并进行去重处理形成初始特征集。

2. PMI 算法特征过滤。点互信息 PMI(Pointwise Mutual Information)是从信息论里的互信息概念中衍生而来的[23]。互信息 MI(Mutual Information)衡量的是两个随机变量之间的相关性,即一个随机变量中包含的关于另一个随机变量的信息量。点互信息 PMI 这个指标常常用来衡量两个事物之间的相关性,比如两个词。本研究使用 PMI 算法将跨境电商政策预处理语料作为输入,先通过频率计算词语的共现概率,然后再计算词语共现的标准化互信息值 NMI(Normalized Mutual Information),最后返回符合 NMI 阈值的特征词列表及 PMI 特征词共现列表。最终通过人工筛选初始特征词和 PMI 算法过滤得到的特征词,形成跨境电商政策文本的特征集,完成特征提取的工作。

(三) 基于 Word2Vec 训练的特征词向量化

Word2Vec 是能把词语转化为多维词向量的模型,根据词语的上下文预测词向量。词向量由多维实数表示,虽然不能说明每一维度的实际含义,但它却蕴含了丰富的信息。由于训练时会根据前后就近位置预测词语,考虑了词语间的共现,因此它保持了同义词之间强的相关性。运用 Word2Vec 词向量模型训练跨境电商政策文本语料,可以将其中的跨境电商特征词转化为多维实数向量。与传统的空间向量模型相比,它考虑了词与词之间的共现,同义词所对应的词向量在多维空间中会更加接近,这为后续更准确的挖掘工作做好了铺垫。

Word2Vec 中有两个重要的算法模型:Skip-gram 模型和 CBOW 模型。这两个模型都包含了输入层、投影层和输出层三层。Skip-gram 模型是通过输入特征词来预测特征词上下文的空间向量[24];而 CBOW 模型是通过输入特征词上下文来预测特征词的空间向量。Skip-gram 模型进行预测的次数要多于 CBOW 模型,每个词在作为中心词时,都要使用周围词进行一次预测,相当于比 CBOW 模型的方法多进行了 k 次(假设 k 为窗口大小),所以 Skip-gram 模型训练时间要比 CBOW 模型长。但在 Skip-gram 模型中,每个词都要受到周围词的影响,每个词在作为中心词的时候,都要进行 k 次的预测、调整,这种多次的调整会使得词向量相对更加准确。因此,在政策文本挖掘过程中本研究选择 Skip-gram 模型进行词的向量化训练。

Skip-gram 模型是将一个词语作为输入,来预测它的上下文。假设有一个句子结构为 $w_{n-2}, w_{n-1}, w_n, w_{n+1}, w_{n+2}$,Skip-gram 模型就是通过输入 w_n 来预测 $w_{n-2}, w_{n-1}, w_{n+1}, w_{n+2}$ 的词向量。利用 Skip-gram 模型预测特征词的上下文,对应公式如下:

$$p(Context(w) \mid w) = \prod_{u \in Context(w)} p(u \mid w) \tag{1}$$

其中,w 为当前词,u 为其周围词。

(四) 基于 K-Means 的文本聚类

K-Means 是经典的划分聚类算法,算法的优点是时间复杂度低,聚类效果不错;缺点是初始 k 值比较难选定,且对初始中心敏感,会受离群点的影响。针对 k 值难确定问题可以使用误差平方和的手肘法和轮廓系数来确定具体的 k 值。算法的基本步骤如下:

(1)随机选择 k 个簇类中心点;

(2)遍历所有数据点,把数据点划分到距离最近的一个簇类中;

(3)划分之后就有 k 个簇,计算每个簇类中点的平均值作为新的簇类中心点;

(4)重复步骤(2)和(3),直到聚类中心不再发生变化,或是迭代次数达到设定的值。

对于 K-Means 聚类中 k 值的选择,可以依据基于误差平方和 SSE(Sum of the Squared Errors)的手肘法。SSE 的计算公式如下:

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

其中,C_i 是第 i 个簇,p 是 C_i 中的样本点,m_i 是 C_i 的质心即 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。

手肘法的核心思想是随着簇类数 k 的增大,聚类的划分会更加细致,每个簇的聚合程度会逐渐提高,这使得误差平方和将不断变小。当 k 小于理想聚类数时,由于 k 的增大会使每个簇的聚合程度快速增加,故误差平方和的下降幅度会变得很大。而当 k 达到理想聚类数时,再增加 k 所得到的聚合程度会迅速减小,对应误差平方和的下降幅度会骤减,然后随着 k 值的继续变大,误差平方和的变化会趋于平缓。误差平方和与 k 的关系图是一个类似手肘的形状,而这个肘部对应的 k 值就是数据理想的聚类数。

此外,k 值的选择还可以通过轮廓系数来确定,选择系数较大时所对应的 k 值[25]。轮廓系数的计算公式如下:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{3}$$

其中,$a(i)$是 i 向量到同一个簇内其他点的平均距离;$b(i)$是 i 向量到与它相邻最近的一个簇内所有点的平均距离。轮廓系数的值是在 -1至1之间,越趋近于1代表内聚度和分离度都相对较优,越趋近于 -1 代表内聚度和分离度都相对较差。将所有数据点的轮廓系数求平均,就是聚类结果总的轮廓系数。

四、政策文本挖掘与研究结论

(一) 研究数据采集

本研究的数据采集策略是在法律法规数据库和各级政府网站上用跨境电商相关的关键词(跨境电子商务、跨境电商、跨境 + 电商、跨境贸易等)进行检索,时间跨度设定为2013—2020年。虽然跨境电商模式在我国的发展可以追溯到20世纪90年代,但是目前业界更倾向于将2013年(前后)称为跨境电商的“元年”,其中一个重要原因就是自2013年起,跨境电商出口规模占出口贸易总额的比重不断提升。而随着跨境电商重要性的提升和行业快速发展,也推动了相关政策的制定和发布。2014年7月,海关总署发布的《关于跨境贸易电子商务进出境货物、物品有关监管事宜的公告》和《关于增列海关监管方式代码的公告》,即业内熟知的“56号”和“57号”文件接连出台,使跨境电商获得了政策层面的认可。因此,本研究选取2013年作为政策数据收集的起始年份,对相关的政策文件进行筛选,去除和跨境电商没有直接关系的政策文件,并摘录符合要求的政策文本。主要按照以下两个标准对政策文件进行筛选:

(1)政策文件直接以跨境电商相关关键词命名;

(2)政策文件包含和跨境电商直接相关的内容。

最终采集到的研究数据主要包括近年来我国国务院、商务部、海关总署、税务总局等政府部门以及30个省份(含自治区、直辖市,不含港澳台)发布的276个跨境电商政策文件。以省(含自治区、直辖市)为单位对地方政府部门发布的跨境电商政策进行统计,结果显示发布相关政策最多的是广东,其次是浙江和江苏,都是目前跨境电商最发达的地区。此外,福建、上海、重庆等地也发布了较多的政策。对于非地方性政策文件,按发文机关或部门进行统计,结果显示,除国务院以外,发布相关政策较多的是海关总署、商务部以及原质检总局等部门。

表1 部分跨境电商相关政策示例

政策文件名称	发布时间	类型
浙江省商务厅、中共浙江省委网络安全和信息化委员会办公室关于印发《浙江省数字贸易先行示范区建设方案》的通知	2020年	(2)
国家税务总局浙江省税务局关于发布《浙江省跨境电子商务综合试验区零售出口货物免税管理办法(试行)》的公告	2018年	(1)
海关总署《关于跨境电子商务企业海关注册登记管理有关事宜的公告》	2018年	(1)
浙江省人民政府关于印发《中国(义乌)跨境电子商务综合试验区实施方案》的通知	2018年	(1)
浙江省人民政府关于印发《中国(浙江)自由贸易试验区建设实施方案》的通知	2017年	(2)
浙江省商务厅等8部门关于印发《浙江省跨境电子商务管理暂行办法》的通知	2016年	(1)

(二) 特征提取与向量化

在特征词提取过程中将阈值设置为0.1,即特征词的 TF-IDF 值大于0.1才会被提取。表2给出了部分 TF-IDF 值较大的初始特征词。

由于同一个特征词可能会在不同文件中重复出现,因此需要将提取出来的初始特征词进行去重处理,然后根据预处理时的词性标注把非名词的特征去除,再利用 PMI 算法进行过滤(特征词共现的 NMI 值范围为0至1,设置阈值为0.1),提取出部分特征结果如表3所示。

表2 初始特征

初始特征词	TF-IDF 值
经营者	0.56
自贸	0.53
出口	0.46
市场监管	0.38
电子商务	0.36
试验区	0.34
司法	0.34
扶持	0.22

表3 PMI 算法特征过滤结果

特征	NMI 值
海关	0.4211
海外仓	0.3158
质量	0.2467
网购	0.2238
互联网	0.2181
平台	0.2013
消费税	0.1725
进口商品	0.1325

通过 PMI 算法过滤得到446个特征词,由于计算机程序的识别中仍然存在不符合的特征,最后经过人工过滤得到355个特征词,作为最终的特征词集合。

(三) 词向量训练结果

在特征词提取的基础之上,进一步使用 Word2Vec 训练模型得到词向量,设置模型参数如表4所示。

根据训练预处理后的政策文本语料得到语料库词表,词表中每个词对应200维的空间向量。跨境电商特征词则对应语料库词表中的355个200维的词向量。训练出的词向量效果可以用词与词之间的相似度和单个词的相关词列表来查看,相似度计算公式如下:

$$\cos(x,y)=\frac{x\cdot y}{|x|\cdot|y|} \tag{4}$$

以政策文件中的四组词对“检验”和“检疫”、“进口”和“出口”、“税款”和“技术”以及“支付”和“交易”为例,计算每组词对的相似度,结果如表5所示。

另外,以政策文件中的关键词“海关”为例,遍历计算词表中所有词的相似度,得到与其相似度最高的前20个相关词,结果如表6所示。

从对类似表6的结果进行人工分析比对,从词对相似度和相关词相似度来看,处理结果符合认知逻辑,说明通过 Word2Vec 训练模型可以训练得到合理的词向量。因此,可以对政策文本其他特征词实施同样的操作,从而确认训练出的词向量效果。

表4　Word2Vec 的参数描述

参数	说明	设置值
window	上下文划窗长度	5
size	词向量的长度	200
skip-gram	是否采用 skip-gram 模型	1

表5　词对相似度

词对	相似度
检验与检疫	0.9771
支付与交易	0.9409
进口与出口	0.8737
税款与技术	0.3679

表6　相关词相似度

相关词	相似度
归类	0.8675
监管区	0.8499
申报	0.8492
放行	0.8441
网购	0.8438
实行	0.8426
入	0.8356
入区	0.8351
手续	0.8330
检疫	0.8323
进口商品	0.8289
单核	0.8282
税务	0.8257
清单	0.8235
办理	0.8234
运	0.8230
状态	0.8202
退库	0.8176
检验	0.8172
放	0.8150

(四) 基于 K-Means 的聚类结果分析

如前文所述,K-Means 聚类算法中的 k 值选择可以通过基于 SSE 的手肘法来确定。手肘法是根据误差平方和曲线的曲率变化来判定合适的 k 值,曲率越大,越明显的拐点处为越优的 k 值。在得到特征向量后,利用公式(2)计算不同 k 值情况下的误差平方和,结果如图2(a)所示。

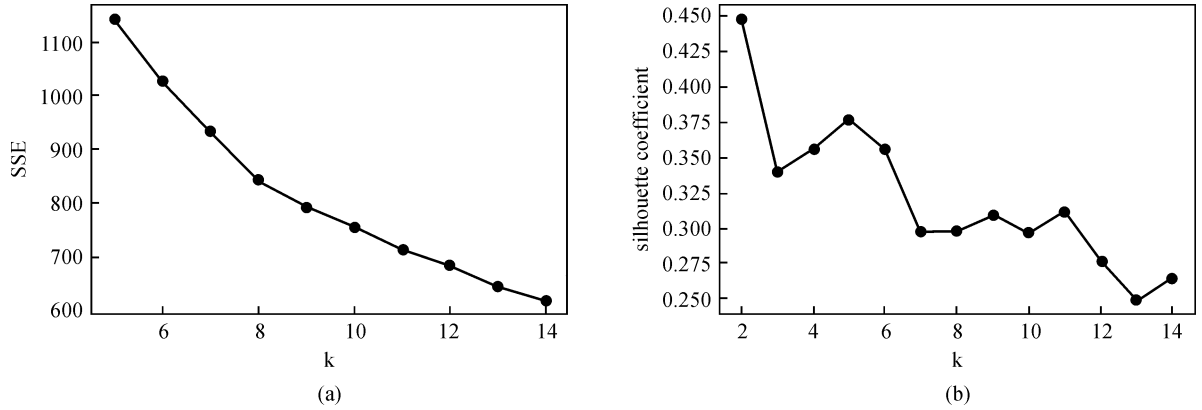

图2　k 值选择结果分析

显然,从图2中的误差平方和曲线来看,其肘部(曲率最高)所对应的 k 值为8,故对于这个数据集的聚类而言,较为合理的聚类数可以设定为8。在得到特征向量后,利用公式(3)计算不同 k 值情况下的轮廓系数,结果如图2(b)所示。选择不同 k 值时轮廓系数的变化情况,如表7所示。

从轮廓系数看随着 k 值的增加轮廓系数逐步降低,在选定的 k 值范围里轮廓系数的变化范围在0.25至0.45之间,而轮廓系数的取值范围在 -1至1之间,轮廓系数越接近1,k 值越优。由于分2个簇与实际情况显然不符,因此,k 值也可以取5。在确定 k 的取值后,使用 K-Means 聚类算法对从政策文本中提取出的特征进行聚类。当 k = 8和 k = 5时,对应的 K-Means 聚类的二维散点图分别如图3(a)和图3(b)所示。

表7 不同 k 值对应的轮廓系数

k	Silhouette Coefficient	k	Silhouette Coefficient
2	0.450	9	0.317
3	0.337	10	0.296
4	0.355	11	0.320
5	0.375	12	0.274
6	0.350	13	0.250
7	0.300	14	0.266
8	0.300		

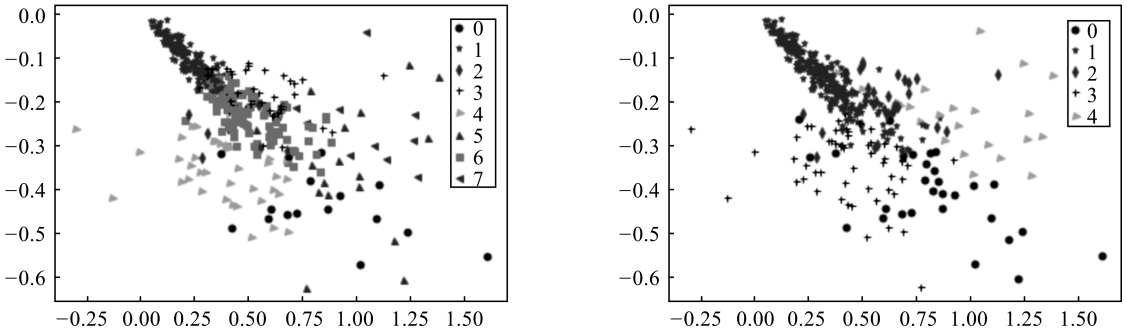

图3 不同 k 值对应的聚类散点图

经过综合比较,最终确定 k 取值为8,然后对跨境电商政策中提取的特征进行聚类,聚类结果如表8所示,表中选择性列出了部分有代表性的聚类结果。

表8 跨境电商政策特征聚类结果

簇类	特征簇	特征数
1	合肥市 安徽省 渤海 常熟市 福州市 兰州 甘肃 沈阳 辽宁省 大连市 浙江省 杭州 上海市 东莞 珠海 深港 广东省 贵阳 湖南省 吉林省 江苏省 山东省 深圳市 重庆市 福建省 闽台 京津冀 天津市 港澳地区 海南省 东北亚 滨海新区 哈尔滨 横琴 长沙市 沈阳市 无锡 南京 南宁 呼和浩特 濮阳市 青岛市 仁川 威海市 日本 韩国 陕西省 苏州市 新乡市 郑州	120
2	综试区 试验区 自贸区 综合 试点 自由贸易 实施方案 一带一路	9
3	科技 知识产权 创新 专利 网络安全 网信 文化 互联网 技术 品牌 侵权 网购 司法 法律 律师	42
4	保税 税款 区域 海关 进口 出口 增值税 税收政策 免税 进口商品 消费税 关税 纳税 出入境	20
5	平台 服务 跨境 支付 交易 融资 投资 电子商务 贸易 物流 业务 代理人	18
6	大宗 油品 食品 快件 农产品 海外仓 服装 保健食品 化妆品 婴幼儿	86
7	检验 监管 消费者 质量 风险 标准 检疫 管理 商品 质检 经营者	17
8	检疫局 商务局 监管局 办公厅 总局 组织 税务局 财政局 人民政府 管理局 人民银行 财政厅 海关总署 商务厅	43

对表8中所列的各个簇类做进一步的解释和归纳如下:簇类1,主要是跨境电商政策实施的热点区域,同时还包括部分跨境贸易的国家和城市;簇类2,主要涉及跨境贸易的城市区域改革与建设;簇类3,主要涉及跨境电商的知识产权、法律法规以及网络安全;簇类4,主要涉及跨境电商的出入境政策和税务;簇类5,主要涉及跨境电商的交易和支付;簇类6,主要和跨境电商的热门产品有关;簇类7,主要和跨境商品的质量安全和检疫检验有关;簇类8,主要是出台相关政策的各种政府部门。

这些特征簇从宏观层面看,涉及跨境电商发展的制度法规、环境建设等;从微观层面看,涉及跨境电商运行的具体环节,如交易、产品、支付、税收、质量管理、知识产权等。由此可见,我国各级政府部门正在努力建立建设健全制度,构建良好的发展平台和环境来推动跨境电商产业的健康发展。从簇类内部的特征数来

看,簇类1和簇类8包含的特征数虽然比较多,但是其特征主要是行政区划和部门,信息比较明确,不需要做过多的解释和分析。而除簇类1和簇类8以外,包含特征数较多的簇类为3、4、6,说明知识产权、税收、产品等是近年来跨境电商政策关注和聚焦的重点内容。簇类3说明,和传统电商相比,跨境电商在行业发展的早期就开始强调知识产权、法律法规以及网络安全问题,这也更有利于保证行业的良性和持续发展。簇类4则强调了跨境电商的出入境政策和税务,和传统电商相比,这些属于跨境电商特有的内容,跨境电商企业在政策层面应重点关注。簇类6指出了跨境电商政策重点关注的商品,一方面说明这些商品属于跨境电商的热销品类,对企业而言是可以重点经营的;另一方面也说明这些商品属于监管重点,企业在生产和销售时,更应重视和确保商品的品质。相比较而言,簇类2、5、7虽然也是跨境电商发展中非常重要的要素,但是和其他簇类相比,包含的特征数较少,单纯从聚类的结果来看并不显著。其中一个可能的原因是,簇类2、5、7涉及的跨境电商区域建设、支付、检疫检验等问题已经相对比较成熟,因此在政策层面不需要通过较多的文本进行阐述,甚至在有些地方性政策中已经不再专门阐述。

此外,还选取了两份典型的省级跨境电商政策,分别是浙江省发布的《浙江省人民政府关于印发中国(湖州)、中国(嘉兴)、中国(衢州)、中国(台州)、中国(丽水)跨境电子商务综合试验区实施方案的通知》(记为政策实例1);江苏省发布的《省政府关于同意中国(常州)、中国(连云港)、中国(淮安)、中国(盐城)、中国(宿迁)跨境电子商务综合试验区实施方案的批复》(记为政策实例2),通过对具体政策的分析来验证聚类结果。将两个政策实例的内容和聚类结果进行对比(见表9),可以看到政策实例和聚类结果的一致性比较高。此外,两个政策实例之间也有较为明显的差异。表8的聚类结果具有一定的典型性,较好地反映了近年来跨境电商政策的关键要素,但是不同地区在制定政策时也会结合和考虑当地的实际情况。

表9　聚类结果实例对比

簇类	特征描述	政策实例1相关度	政策实例2相关度
1	跨境电商政策实施的热点区域	/	/
2	跨境贸易的城市区域改革与建设	高	高
3	跨境电商的知识产权、法律法规以及网络安全	低	中
4	跨境电商的出入境政策和税务	高	高
5	跨境电商的交易和支付	中	高
6	跨境电商的热门产品	高	低
7	跨境商品的质量安全和检疫检验	低	中
8	出台相关政策的各种政府部门	/	/

(五) 语义网络分析

除了聚类分析,本研究还利用 ROSTCM 数据挖掘工具[26]对跨境电商政策文本进行了初步的语义网络分析。本研究主要使用 ROSTCM 中的网络分析功能,对跨境电商政策文本进行语义网络分析。将经过预处理的跨境电商政策文本用 ROSTCM 分析来生成可视化的语义网络(见图4),通过网络图可以进一步分析词与词之间的关联关系,从而更加直观地了解政策文本中重要词语间蕴含的关联关系。

从图4的语义网络可以看出,电子商务、平台、发展、创新、建设、试验、改革、制度、政策、模式等关键词在网络中属于关键节点,从一定程度上反映出我国近年来跨境电商政策的聚焦重点。另外,根据关键节点在网络中的连通关系,通过组合关键词可以进一步得到平台建设、改革创新、制度改革、制度创新、模式创新、政府政策、海关出口等关联短语。

此外,还基于 PMI 方法计算获得了高频词共现矩阵,并通过 Gephi 工具进行可视化展示,包括平均加权度、模块化、平均聚类系数等参数的计算。通过合理地布局,得到如图5所示的聚类图像展示。

在此基础上,过滤出聚类系数为0.6以上的数据,并重新处理和展示,最终得到如图6所示的结果。从图6中可以观察到,管理/规划、运营/服务、信用/规则、检疫/检验、卖家/供货商、商品质量/消费者、税务、司法、证书/营业执照等特征词属于网络的关键节点。

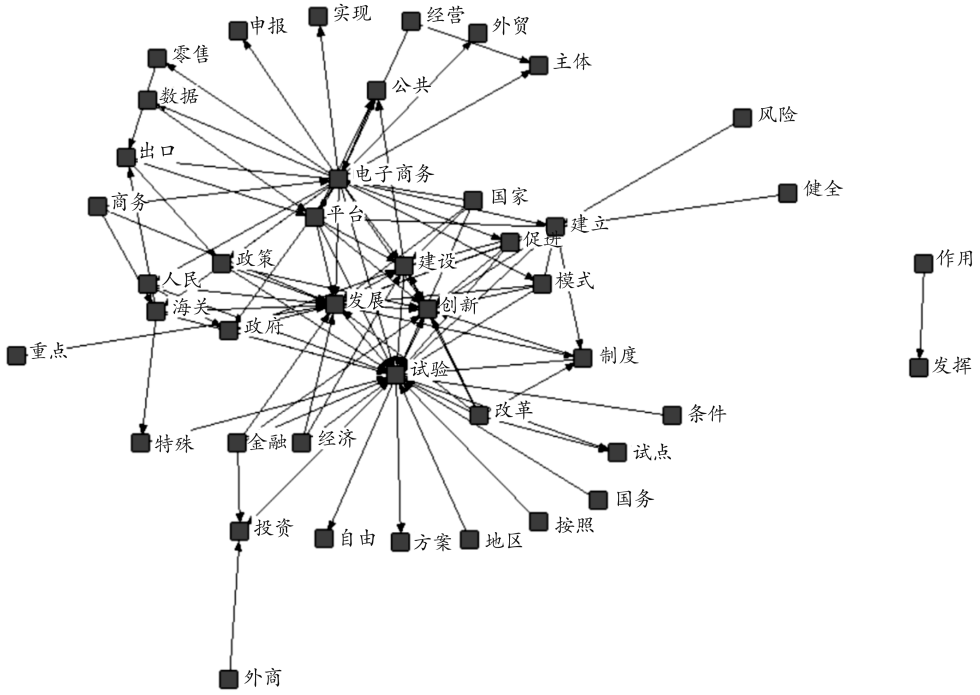

图4 跨境电商政策语义网络

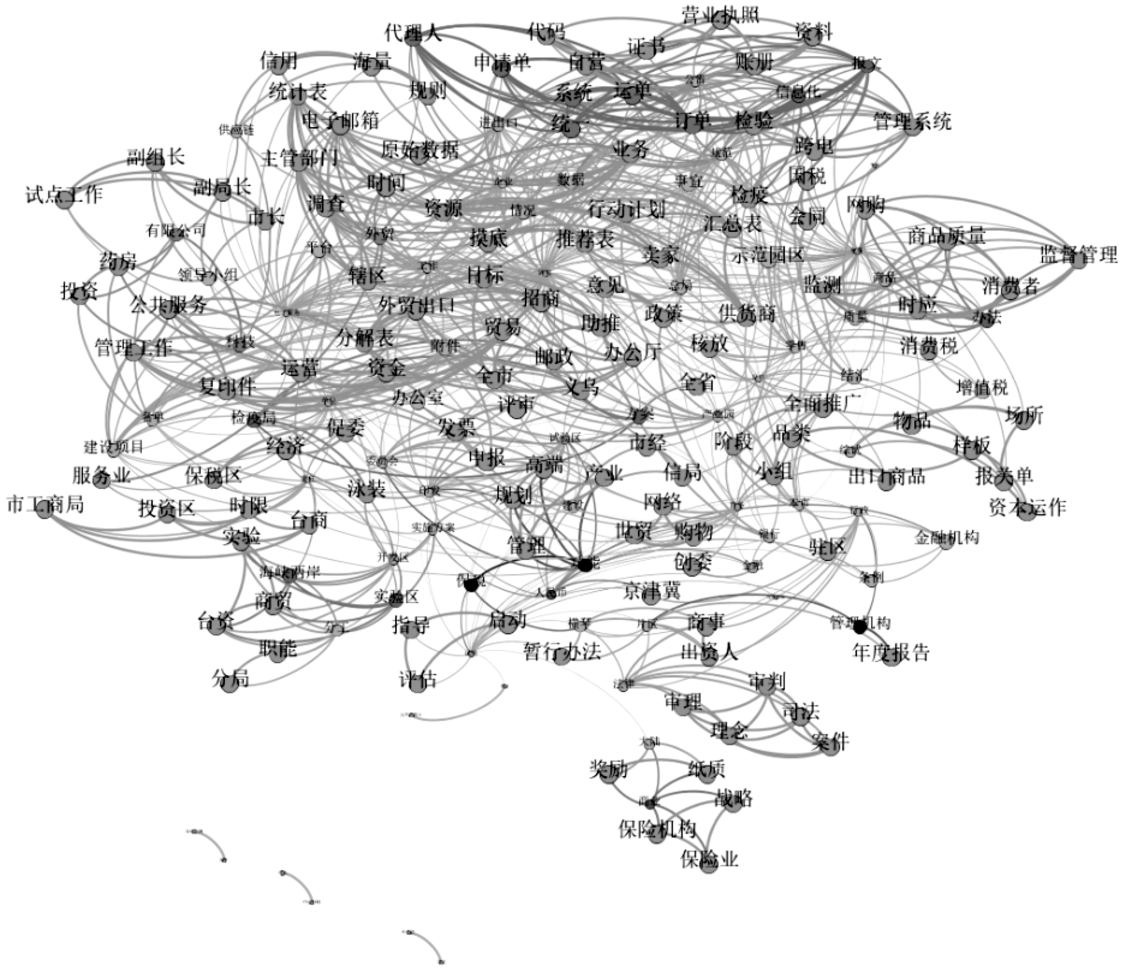

图5 高频词共现矩阵的聚类图像展示

通过上述语义网络分析,可以初步得出我国近年来跨境电商政策一方面聚焦创新、试验、建设、发展等宏观问题,另一方面也重点关注税务、商品质量、检疫检验等具体问题。语义网络分析的结果和聚类分析的结果具有一定的重合性和关联性,从另一个维度体现了跨境电商政策的要点和重点。因此,语义网络分析可以作为聚类分析的补充。

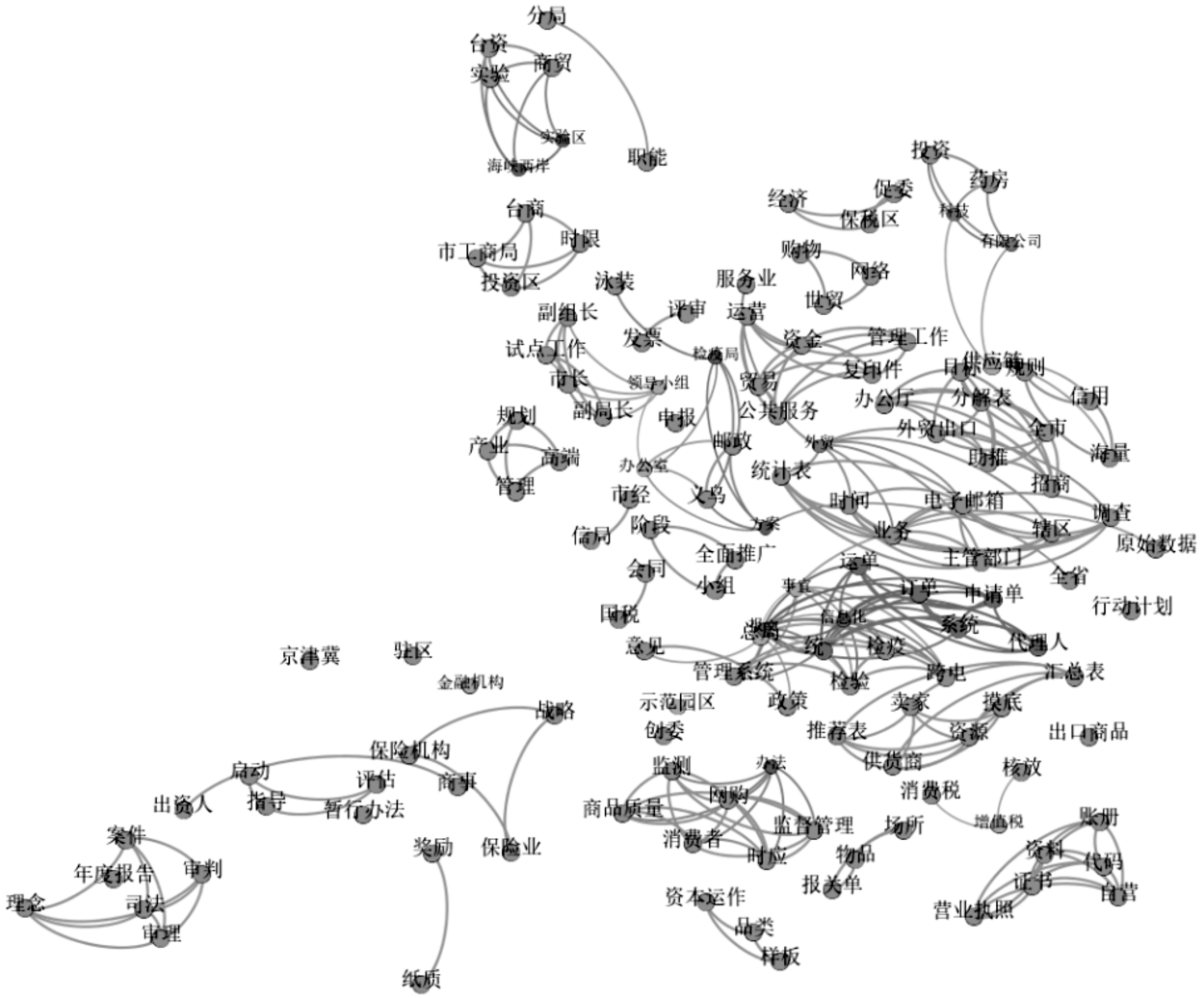

图6　聚类系数过滤后的图像展示

五、研究结论与建议

(一) 研究结论

本研究基于文本挖掘的跨境电商政策量化分析方法,通过从跨境电商政策文本中提取特征,再进一步使用聚类方法和语义网络方法分析了跨境电商政策的重点内容。该方法能够从大量的政策文本中挖掘出特征信息,并为跨境电商政策分析和行业研究提供决策支持,得到如下结论。

第一,跨境电商领域已经形成较为充分的政策数据集。近年来,我国各级政府部门发布了一系列跨境电商政策文件,通过分类整理,可以获得一个较为完整的跨境电商政策数据集,包含了我国30个省份(含自治区、直辖市,不含港澳台),以及国务院、商务部、海关总署、税务总局等部门发布的跨境电商政策文件。这个数据集能够被用于研究和分析跨境电商领域的政策特征和演化规律,并为跨境电商行业发展提供

决策支持。

第二,跨境电商政策文本呈现出较为显著的簇类特征。从聚类分析的结果来看,我国跨境电商政策的内容呈现出较为显著的特征簇。这些特征簇从宏观层面看,涉及跨境电商发展的制度法规、环境建设等;从微观层面看,涉及跨境电商运行的具体环节,如交易、产品、支付、税收、质量管理、知识产权等。说明我国各级政府部门正在努力建立健全制度,构建良好的发展平台和环境来推动跨境电商产业的健康发展。同时也说明,和传统电商相比,跨境电商在发展早期就有比较完善的制度来规范行业发展。

第三,语义网络分析能进一步发现跨境电商的政策重点。通过对跨境电商政策文本进行语义网络分析,挖掘网络中的关键节点,能从不同维度进一步反映我国近年来跨境电商政策的聚焦重点。从语义网络分析的结果来看,可以初步得出我国近年来跨境电商政策主要聚焦在创新、试验、建设、发展等方面。语义网络分析的结果和聚类分析的部分结果有一定的重合度和关联性,可以作为聚类分析的补充。

第四,文本挖掘为跨境电商政策研究提供了新的视角。利用文本挖掘的方法,对跨境电商政策文本进行挖掘和分析,有助于分析我国跨境电商政策中的关键内容。本研究融合多种文本挖掘和分析方法,形成以“文本预处理、特征提取、特征过滤、特征向量化、文本特征聚类、语义网络分析”为流程的研究模式,为跨境电商政策研究提供了新的思路,同时也可以扩展到其他领域的政策文本研究中去。

(二) 对策建议

跨境电商目前已经成为我国发展速度最快、潜力最大、带动作用最强的外贸新业态,而跨境电商的快速发展又离不开政策的支持。科学合理的政策有助于推动产业的良性健康发展。结合上述研究结论,本研究提出如下建议。

第一,跨境电商政策在强化“长板”的同时,也应积极补足“短板”。通过文本挖掘能发现当前政策的关注焦点,但同时也从一个侧面反映出当前政策存在一定的“盲区”。因此,各地区、各部门在制定跨境电商政策时,应强化“长板”,补足“短板”,积极推动跨境电商产业的均衡发展。目前,各地在制定和发布跨境电商政策文件时,充分考虑了区域性特征和实际情况,而如何针对自身的不足通过政策引导和推动加以弥补,也是值得深入思考和研究的问题。

第二,跨境电商政策应尽可能覆盖和适应动态变化的行业形势。当前国内外形势正在发生深刻复杂的变化,我国发展仍处于重要战略机遇期。和动态变化的经济社会形势相比,政策文件的制定和实施会有一定的滞后性。为此,应更加科学合理地研究和发布政策,让政策能够尽可能地覆盖和适应动态变化的行业形势(如后疫情、双循环等)。另外,还应适当结合和参考国外电商相关政策文本,不断改进和完善我们的政策。

第三,要充分利用信息技术和大数据思维来研究和分析政策。和人工分析相比,基于文本挖掘方法进行政策分析,能够实现对大量文本数据的有效处理和分析,同时也有助于发现一些隐性的或潜在的规律,并确保政策的一致性和延续性。当然,目前对于政策文本的分析还停留在相对浅层,如何通过加强分析粒度来进行更加深入具体的分析,也是后续需要重点研究和解决的问题。

第四,跨境电商政策对企业有较强的导向作用,企业应通过解读政策来更好地理解行业和洞察商机。从聚类分析结果来看,跨境电商政策本身就指出了重点监管的商品品类,对企业生产经营具有一定的导向作用。同时,聚类分析结果也体现了政府部门对跨境电商知识产权、法律法规、税收等要素的重视,意味着企业在生产经营时,也应更多关注这些问题。

第五,跨境电商政策的内容应适当考虑垂直行业特征。从目前政策文本的内容分析来看,绝大部分政策并没有针对垂直行业或者跨境电商热门品类的内容或描述。而实际上垂直行业或者品类之间的差异对跨境电商运营是有较大影响的,监管的要求也会有所不同。因此,后续可以考虑针对跨境电商的热门品类(如宠物用品、户外用品、消费电子等)制定和出台专门的政策,或者在政策中有专门的内容体现。

参考文献:

[1]马述忠,房超.跨境电商与中国出口新增长——基于信息成本和规模经济的双重视角[J].经济研究,2021(6):159-

176.
[2]吴国英,闫建钢.“趋利”还是“避害”? ——直播电商退换服务对消费者行为的影响研究[J].现代财经(天津财经大学学报),2021(12):65-77.
[3]李宏兵,王爽,赵春明.农村电子商务发展的收入分配效应研究——来自“淘宝村”的经验证据[J].经济经纬,2021(1):37-47.
[4]MA S Z,GUO X Y,ZHANG H S. Policy analysis and development evaluation of digital trade:an international comparison[J]. China & World Economy,2019,27(3):49-75.
[5]邢光远,史金召,路程.“一带一路”倡议下中国跨境电商的政策演进与发展态势[J].西安交通大学学报(社会科学版),2020(5):11-19.
[6]徐德顺.后疫情时代中国跨境电商发展的政策建议[J].对外经贸实务,2021(7):4-7.
[7]张晚冰.影响我国跨境电商零售出口的政策因素分析[J].全国流通经济,2021(17):38-40.
[8]RICHARDS C,FARROKHNIA F. Optimizing grounded theory for policy research:a knowledge-building approach to analyzing WTO e-commerce policies[J]. International Journal of Qualitative Methods,2016,15(1):1609406915621380.
[9]HANNA N K. E-commerce as a techno-managerial innovation ecosystem:policy implications[J]. Journal of Innovation Management,2016,4(1):4-10.
[10]赵杨,陈雨涵,陈亚文.基于 PMC 指数模型的跨境电子商务政策评价研究[J].国际商务(对外经济贸易大学学报),2018(6):114-126.
[11]熊励,郭梦滢,叶凯雯.基于多期双重差分模型的跨境电商政策效应评价研究[J].智库理论与实践,2022(3):41-52.
[12]邱国斌,于梦鑫,胡佳星,等.基于 fsQCA 方法的江西省跨境电商发展政策研究[J].科技和产业,2022(1):81-87.
[13]ROBERTA A,RENATO S L,DAVID C S,et al. Agent-based simulation model for evaluating urban freight policy to e-commerce[J]. Sustainability,2019,11(15):4020.
[14]LIN C,LIN H C K,HUANG Y A,et al. The fit between organizational B2B e-commerce policy,IT maturity and evaluation practices on B2B e-commerce performance in Australian healthcare organizations[J]. African Journal of Business Management,2011,5(5):1983-2005.
[15]李泓烨,王旭,梁颖.政策工具视角下中国跨境电商政策文本量化评价研究[J].图书情报导刊,2021(12):37-44.
[16]钮钦.中国农村电子商务政策文本计量研究——基于政策工具和商业生态系统的内容分析[J].经济体制改革,2016(4):25-31.
[17]侯振兴,阎燕.区域农产品电子商务政策文本计量研究——以甘肃省为例[J].中国流通经济,2017(11):45-53.
[18]盛赟,阮钰涵,任嘉英,等.基于文本分析的浙江省跨境电子商务物流政策研究[J].物流工程与管理,2019(5):26-30.
[19]余传明,郭亚静,龚雨田,等.基于主题时间模型的农村电商扶贫政策演化及地区差异分析[J].数据分析与知识发现,2018(7):34-45.
[20]金珺,李诗婧,陈赞.基于社会网络分析的农村电商政策研究——以浙江、甘肃为对比[J].创新科技,2020(2):38-48.
[21]肖开红,雷兵,钟镇.中国涉农电子商务政策的演进——基于 2001—2018 年国家层面政策文本的计量分析[J].电子政务,2019(11):91-103.
[22]RAJARMAN A,ULLMAN J. Mining of massive datasets[M]. New York:Eambridge University Press,2011:1-17.
[23]VERGARA J R,ESTEVEZ P A. A review of feature selection methods based on mutual information[J]. Neural Computing and Applications,2014,24(1):175-186.
[24]MIKOLOV T,CHEN K,CORRADO G,et al. Efficient estimation of word representations in vector space[EB/OL]. 2013,https://doi.org/10.48550/arXiv.1301.3781.
[25]KAUFMAN L,ROUSSEEUW P J. Finding groups in data:an introduction to cluster analysis[M]. New York:John Wiley & Sons,2009:1-67.
[26]张幸芝,雷润玲,杨超.文本挖掘——基于 ROSTCM 和 NetDraw 的内容分析[J].科技文献信息管理,2017(1):17-21.

(责任编辑　游旭平)