

基于LDA模型的跨境电商研究主题挖掘

娄金沅

江苏大学科技信息研究所, 江苏 镇江

收稿日期: 2025年5月13日; 录用日期: 2025年5月29日; 发布日期: 2025年7月3日

摘要

在全球化与数字技术快速发展的背景下, 我国跨境电商已成为推动外贸增长的重要引擎。本文以2000~2024年CNKI数据库的1516篇中文核心期刊文献为研究对象, 采用LDA主题模型的文本挖掘方法, 系统分析跨境电商研究的主题分布情况。研究发现: (1) 跨境电商研究历程分为萌芽期(2001~2013年)、爆发期(2014~2020年)和成熟期(2021年后)三个阶段, 研究热点与政策发布时间高度相关; (2) 关键词分析显示, “物流” “企业” “政策” “数字化” 等核心研究维度, 物流与供应链管理是跨境电商运营的核心环节; (3) LDA主题聚类识别出五大研究主题, 包括跨境贸易政策与供应链协同创新、制造业进出口绩效评估与数字化转型、国际贸易商品与专业化人才培养、价值链协同与质量安全治理、数字丝路建设与跨境服务争议, 揭示了跨境电子商务发展的演进路径。本研究为跨境电商研究领域提供了系统的文献分析框架, 对政策制定与企业实践具有参考价值。

关键词

跨境电商, LDA模型, 文本挖掘

Research Topic Mining of Cross-Border E-Commerce Based on LDA Model

Jinyuan Lou

Institute of Science and Technology Information, Jiangsu University, Zhenjiang Jiangsu

Received: May 13th, 2025; accepted: May 29th, 2025; published: Jul. 3rd, 2025

Abstract

Against the backdrop of rapid globalization and digital technology development, cross-border e-commerce in our country has become an important engine for promoting foreign trade growth. This article analyzes 1516 core Chinese journal articles from the CNKI database covering the years 2000 to 2024, utilizing the LDA topic model for text mining to systematically examine the topic distribution

in cross-border e-commerce research. The study finds: (1) The research process of cross-border e-commerce can be divided into three stages: the budding period (2001~2013), the explosive period (2014~2020), and the mature period (after 2021), with research hotspots being highly related to the timing of policy releases; (2) Keyword analysis reveals that core research dimensions such as “logistics”, “enterprises”, “policies”, and “digitalization” are pivotal, indicating that logistics and supply chain management are central to the operation of cross-border e-commerce; (3) The LDA topic clustering identifies five major research themes, including cross-border trade policies and supply chain collaborative innovation, assessment of import and export performance in manufacturing and digital transformation, international trade goods and specialized talent cultivation, value chain collaboration and quality safety governance, and the construction of a digital Silk Road with controversies in cross-border services, exposing the evolutionary path of cross-border e-commerce development. This study offers a systematic literature analysis framework for the field of cross-border e-commerce research and holds significant reference value for policy formulation and corporate practices.

Keywords

Cross-Border E-Commerce, LDA Model, Text Mining

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

在全球化浪潮与数字技术蓬勃发展的当下，我国跨境电商呈现爆发式增长，深刻改变着全球商业格局。据海关总署统计，2024 年我国跨境电商进出口规模为 2.63 万亿元，同比增长 10.8% [1]。2025 年 4 月，国际影响导致全球市场剧烈动荡，中国跨境电商平台在海外迅速走红，“中国制造”优势明显，在质量和价格方面都难以替代。但品牌出海导致中国货主支付的物流成本不断提升、政策的变动性增加了运营的不确定性，学术界围绕营销策略[2] [3]、跨境物流运输[4]、数字支付[5] [6]等跨境电商多个关键环节展开研究，但现有研究成果多聚焦宏观层面或单一主体。在研究视角上，“一带一路”与跨境电商存在紧密联系，为跨境电商发展提供了基础设施以支持互联互通[7]；在研究内容上，物流成本、关税政策等话题关注较多；在研究方法上，现有研究主要采用实证分析或定量分析，缺乏对海量文献数据的文本挖掘。为此，为全面剖析跨境电商研究主题，本文基于文本挖掘的方法，运用 LDA 主题模型对 CNKI 中相关文献数据进行系统地分析，从而展示跨境电商研究主题热点与研究趋势。

2. 数据与方法

2.1. 数据来源

以跨境电子商务研究相关的中文文献为数据基础，选择 CNKI 数据库作为数据来源。结合篇关摘和主题检索两种方式进行数据检索，检索词包括“跨境电商”“跨境电子商务”等，来源类别选择“北大核心”，检索年限设定为 2000 年~2024 年，检索日期为 2025 年 4 月 30 日，最终共得到 1516 篇文献。

2.2. 研究方法

2.2.1. 数据预处理

本研究利用 python3.9 对跨境电子商务文献内容进行 LDA 主题分析，调用中文分词工具 Jieba 进行

分词预处理。在处理前需要构建停用词表，停用词主要将“哈工大停用词词库”“四川大学机器学习智能实验室停用词库”“百度停用词表”汇总去重而得，以减少文本的噪音信息，提高文本分析的效率和准确性，并利用 Jieba 对文献摘要进行分词处理，为后续主题分析做好数据准备。

2.2.2. LDA 主题模型

潜在狄利克雷分配模型(Latent Dirichlet Allocation, LDA)的主题建模分析法是近年来较为新颖的对文本内容的量化分析方法。LDA 主题模型是一种典型的三层贝叶斯词袋模型，即词语层、主题层、文档层。它将词语切分后进行主题概率计算，每个主题可以表现为若干个主题词语的概念分析，再将每个文档中词语对应主题的概率进行计算，得到了文档-主题概率。这种方法提升了对深层次文本内容挖掘和语义分析的精准度与效率[8]，能够帮助研究人员把握学科领域发展趋势，应用广泛。

3. 结果分析

3.1. 时间分布特征

按照发文时间对收集的 1516 篇文献进行了统计分析。如图 1 所示，2001 年以来，我国跨境电商研究发文量整体呈上升状态，其年发展态势随着对外开放、国际贸易形势和政策调整而不断变化。因此，结合我国跨境电子商务的宏观战略部署，本研究将跨境电子商务的研究划分为三个时期。

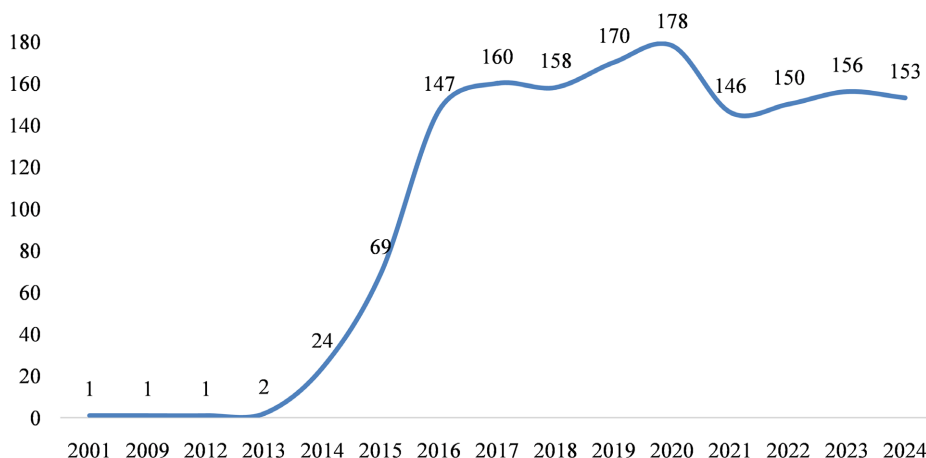


Figure 1. Line chart of the number of published articles from 2001 to 2024

图 1. 2001 年至 2024 年发文量折线图

2001 年至 2013 年是跨境电子商务研究的萌芽期。这一时期的研究多集中于对跨境电子商务运营模式、支付管理等理论内容的探讨。2009 年，深圳被批准创建首个国家电子商务示范城市，2011 年，北京、天津、上海、重庆等 21 个城市创建“国家电子商务示范城市”。以上举措均为了在跨境电子商务交易、支付、物流、通关、退税、结汇等环节通过制度创新、管理创新、服务创新和协同发展，逐步形成一套适应和引领全球跨境电子商务发展的管理制度和规则，为推动中国跨境电子商务健康发展提供经验，为跨境电子商务的后续发展打了坚实的基础。随着全球化程度的不断加深电子商务和跨境电商的发展，从 2014 年起学术界开始大量研究跨境电商，文献数量激增至 24 篇，进入爆发式增长期。这一时期，学者们的研究热点仍集中于对跨境电子商务存在的问题、发展道路、监管模式等方面的讨论，但内容逐渐细化，开始以案例研究、实证分析、问卷调查的方式展开研究，且聚焦群体更加细化、分析视角更加专业。2015 年，国务院同意设立中国(杭州)跨境电子商务综合试验区，为我国发展跨境电商的制度、管理、服务、监

管等探索经验。此后七年(2014~2020 年)文献量持续攀升,年均增长率达 35.6%。这一阶段的快速增长与政策试点使得跨境电子商务的研究热度达到历史高点。2021 年后文献数量呈现高位震荡特征,增速趋缓并出现小幅波动,表明该研究领域已进入成熟期,学者们开始从宏观模式研究转向细分领域深化。

3.2. 关键词分析

对文献进行分词处理,通过词频统计,绘制了如下词云图(如图 2)。从整体上看,“跨境电商”“企业”与“物流”是词云图最突出的词汇,表明在跨境电商的研究中,企业是运营的主体,企业的管理模式、政策对企业的各种措施都会影响跨境电商行业的发展;此外,物流是跨境电商运营中最关键的环节,联系也最为紧密,其发展水平的高低直接影响消费者的购物体验、服务质量和跨境电商的长期发展。具体来看,词云图主要描绘了以下四部分内容:



Figure 2. Keyword word cloud diagrams

图 2. 关键词词云图

(1) 物流与供应链。“物流”“供应链”“仓储”“通关”“运输”“货物”等词汇频次较高,说明物流和供应链管理是跨境电商的重要组成部分,涉及到货物的运输、仓储、清关等环节,是跨境电商运营实现的核心环节。

(2) 国际贸易与市场。“国际贸易”“进出口”“对外贸易”“出口贸易”“国际市场”等词汇强调了跨境电商在全球贸易中的作用,以及对国际市场的开拓和参与。

(3) 数字化与技术。“信息化”“信息技术”“互联网实践”“数字化”“电商平台”等词汇显示了技术在跨境电商中的重要性,技术的发展正在深刻地改变着跨境电商的运营模式,包括跨境电商平台的开发、跨境支付、信息系统的建设以及数字化转型等。

(4) 政策与法规。“政策”“法律法规”“监管”“合规”等词汇指出政策环境和法律框架对跨境电商的影响，企业需要遵守相关法规并应对政策变化。

此外，相关研究也聚焦于跨境电商的未来发展。“创新”“转型升级”“高质量发展”“新模式”“新业态”等词汇反映了跨境电商行业正在经历的变革和发展趋势，强调创新和高质量发展的重要性，“困境”“贸易壁垒”“纠纷”揭示了跨境电商面临的各种挑战和困难，包括市场竞争、贸易壁垒等。

3.3. 主题聚类分析

采用 LDA 主题模型对跨境电子商务的研究进行主题聚类分析。由于 LDA 主题抽取的效果与主题数的确定有直接关系, 本文结合一致性评价指标来确定最优主题数目。在 LDA 模型参数设定方面, 本研究

采用对称先验分布 $\alpha = 50/K$ (K 为主题数) 和 $\beta = 0.01$ 的默认参数配置, 运行程序后得到文档 - 主题分布和主题 - 词语分布。跨境电子商务主题一致性变化趋势见图 3, pyLDAvis 可视化展示的结果见图 4。

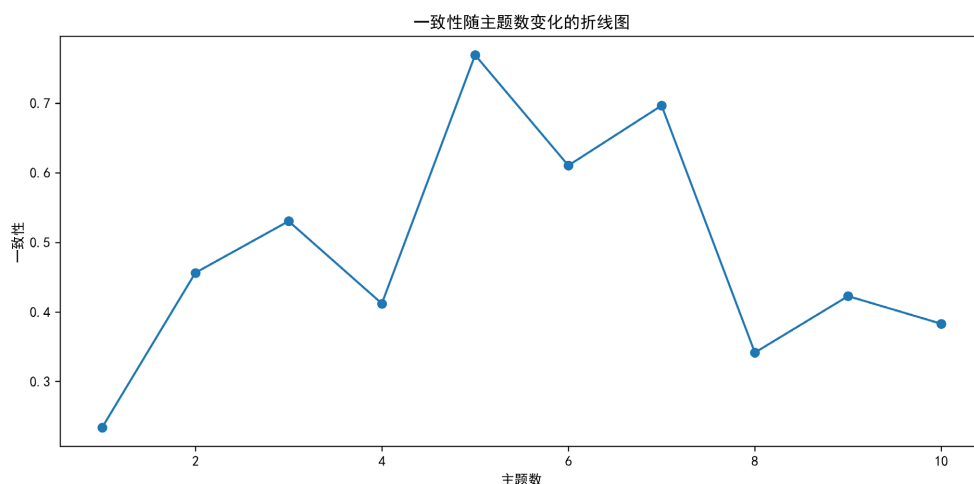


Figure 3. The changing trend of theme consistency
图 3. 主题一致性变化趋势

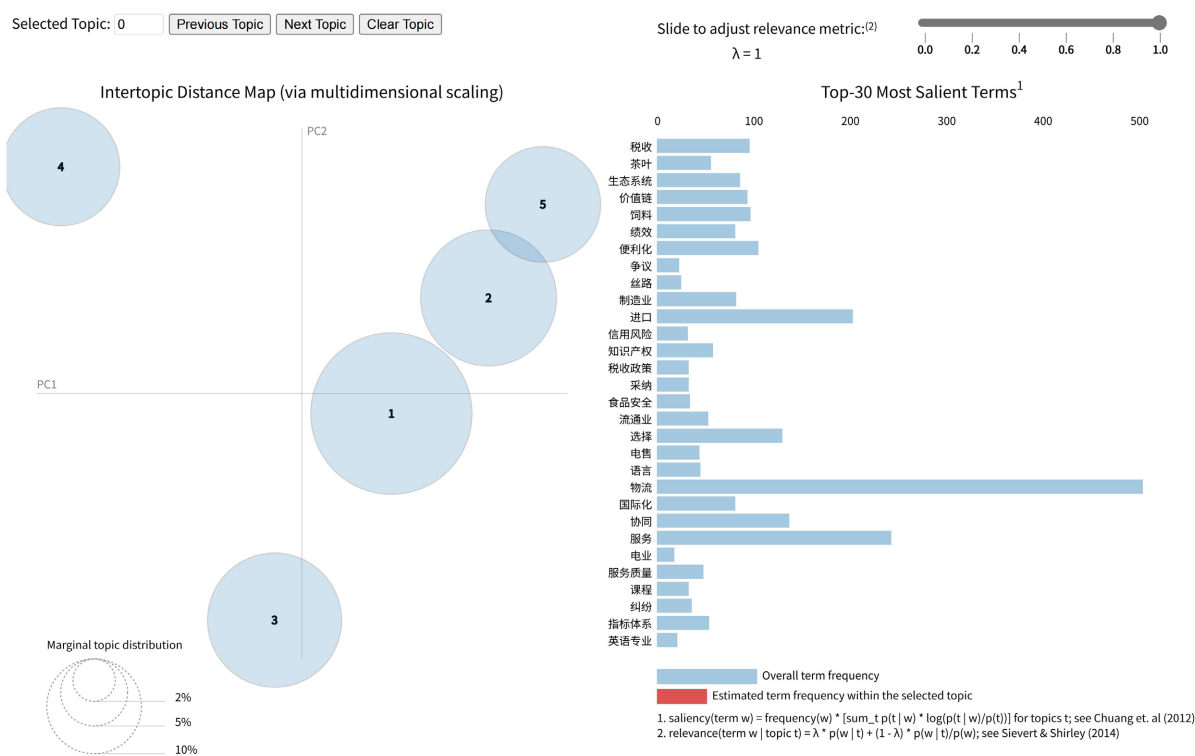


Figure 4. Topic clustering pyLDAvis visualization results
图 4. 主题聚类 pyLDAvis 可视化结果

根据主题一致性变化和可视化聚类分析, 可以看出跨境电子商务研究领域在主题数为 5 时, 各个主题之间重合区域较少, 因此本研究将最优主题数定为 5, 各个主题的命名采用关键词概括的形式, 如表 1 所示。

Table 1. Cross-border e-commerce topic clustering and its keywords
表 1. 跨境电商主题聚类及其关键词

主题	关键词(部分)
跨境贸易政策与供应链协同创新	税收, 监管, 创新, 供应链, 互联网, 生态系统, 竞争力, 进口, 一带一路, 高质量发展, 服务, 新业态, 流通业, 税收政策, 协同
制造业进出口绩效评估与数字化转型	进口, 绩效, 便利化, 制造业, 基础, 进出口, 出口贸易, 互联网, 一带一路, 系统, 对外贸易, 服务, 海外, 指标体系, 信用风险, 竞争力
国际贸易商品与专业化人才培养	供应链, 企业, 海外, 国际贸易, 服务, 饲料, 人才, 对外贸易, 人才培养, 体系, 出口贸易, 监管, 创新, 茶叶, 互联网, 转型升级, 实践
价值链协同与质量安全治理	企业, 价值链, 对外贸易, 进口, 协同, 一带一路, 创新, 监管, 服务, 体系, 产业链, 流通, 协同发展, 国际贸易, 互联网, 路径, 食品安全
数字丝路建设与跨境服务争议	服务, 企业, 一带一路, 进口, 基础, 生态系统, 监管, 便利化, 进出口, 供应链, 国际化, 通关, 国际贸易, 知识产权, 创新, 竞争力, 丝路

主题 1 可命名为“跨境贸易政策与供应链协同创新”，该主题聚焦于政策调控与供应链创新的相互作用。具体表现为通过税收优惠、监管机制等政策工具降低跨境贸易成本，同时借助物流体系升级、产业链协同提升供应链韧性。尤其是在“一带一路”背景下，通过技术创新构建数字生态系统，形成涵盖跨国流通、通关便利化的新型贸易模式。

主题 2 可命名为“制造业进出口绩效评估与数字化转型”，该主题围绕制造业出口竞争力提升路径，通过构建绩效评估体系、量化便利化措施效果，重点考察数字化转型对出口贸易的促进作用。在基础设施支撑下，企业借助互联网实现服务标准化，应对成员国市场准入差异。

主题 3 可命名为“国际贸易商品与专业化人才培养”，该主题展示了跨境电商交易的商品种类有“饲料”“茶叶”等，同时还强调了国际贸易中对“人才”的需求。通过制定完善的人才培养体系和实践体系培养跨境电商交易中的专业人才，从而有效应对全球贸易发展，提升国际竞争力。

主题 4 可命名为“价值链协同与质量安全治理”，该主题聚焦价值链重构过程中的治理难题，通过协同机制整合跨境服务资源，建立覆盖食品安全、产品认证的全流程监管体系，同时，企业还需应对不同市场复杂度带来的合规成本。

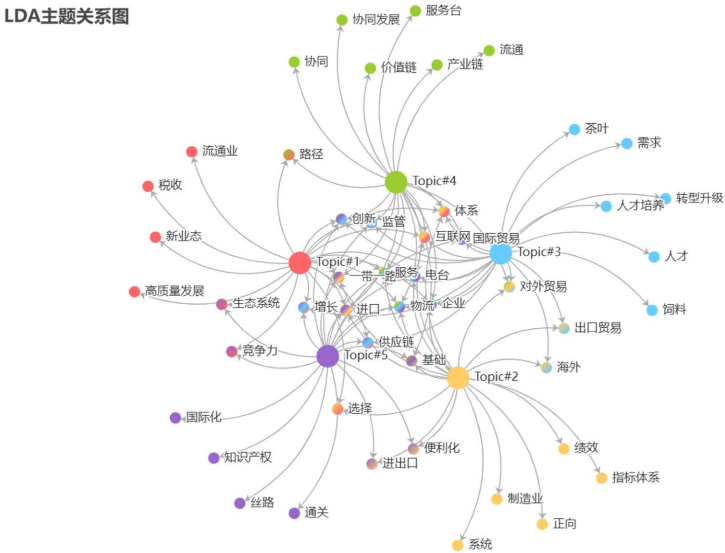


Figure 5. Co-occurrence map of themes and keywords
图 5. 主题 - 关键词共现图

主题 5 可命名为“数字丝路建设与跨境服务争议”。一方面该主题强调了在丝绸之路建设过程中,对跨境电子商务的数字化推动跨境服务效率提升,衍生出知识产权、支付结算等新型争议。在语言、标准等软联通要素支撑下,企业需建立覆盖境外市场的风险预警系统。

通过主题-关键词共现情况绘制了 LDA 主题关系图(见图 5)。图的中心主要由五个主题节点所构成,图的内围主要是五个主题中共同出现的高频词,围绕在主题节点周围的是些关键词节点,这些关键词与各主题相关联。“物流”“服务”“企业”在五个主题中同时出现,说明这是跨境商务的重要组成。从主题 3“国际贸易商品与专业化人才培养”到主题 5“数字丝路建设与跨境服务争议”展现了实体商品流动到数字服务扩展,完善的供应链是重要的实现途径。主题 1“跨境贸易政策与供应链协同创新”从单向政策输出到主题 4“价值链协同与质量安全治理”双向标准协同,共同指向“创新”“监管”等词,反映了我国在全球经贸体系中的角色转变。

4. 结论与展望

4.1. 结论

通过对我国跨境电子商务相关研究的分析,得出以下结论:(1) 研究主题具有阶段性特征。2001~2013 年为萌芽期,聚焦基础性议题,以理论探讨为主;2014~2020 年为爆发期,政策驱动下研究主题细化,供应链协同、监管创新与案例实证成为热点;2021 年后为成熟期,研究转向数字化转型等深层问题,文献增速趋缓但内容深化。(2) 关键词分析显示“企业”与“物流”是跨境电商研究的核心议题,涉及仓储、通关、运输等环节的效率优化;此外,物流与供应链、国际贸易与市场、数字化与技术以及政策与法规四组关系也在相关研究中反复出现。(3) 通过 LDA 主题模型共聚类得到跨境电商五大主题,分别是跨境贸易政策协同、制造业数字化转型、专业化人才培养、质量安全治理及数字丝路建设其关联性,各主题间具有一定联系,既展现了我国跨境电商发展的脉络,也反映了我国在国际贸易中的地位与作用。

4.2. 展望

未来研究可从以下几方面深化探索:(1) 深化细分领域研究,当前文献对具体贸易商品关注较多,但对部分新兴市场的案例分析不足,可结合区域特色展开研究,弥补现有文献对新兴市场的分析不足。(2) 创新探索技术驱动的跨境电商新业态,如今生成式 AI、区块链等新技术正重塑跨境电商生态,需加强 AIGC 在跨境交易、区块链技术在跨境支付中的应用,进一步完善研究体系。(3) 本研究数据来源限于中文文献,未来可结合 WOS 等国际数据库,开展中英文研究的比较分析,在方法层面也可融合社会网络分析(SNA)与 LDA 模型,揭示作者合作网络与主题演化的关联性。

参考文献

- [1] 新华社. 图表: 2024 年我国跨境电商进出口同比增长 10.8% [EB/OL]. 2025-01-13. https://www.gov.cn/zhengce/jiedu/tujie/202501/content_6998286.htm, 2025-04-30.
- [2] 李楠. 跨境电商中的文化适应性营销策略研究[J]. 中国集体经济, 2025(11): 101-104.
- [3] 吕欣蓉. 基于价值链的中小微跨境电商企业直播营销管理诊断研究[J]. 商展经济, 2025(7): 125-128.
- [4] 石学刚, 樊浩天. 跨境电商与航空物流的协同发展研究——基于城市层面的实证分析[J]. 价格理论与实践, 2024(9): 152-159.
- [5] 车丽波. 数字金融赋能跨境电商: 理论逻辑、现实困境与发展路径[J]. 西南金融, 2024(8): 81-92.
- [6] 房朔杨. 跨境电商发展对服务贸易出口的影响机制研究——基于跨境支付的中介效应检验[J]. 商业经济研究, 2024(19): 138-141.
- [7] 侯养兵. “一带一路”背景下贸易便利化对我国跨境电商出口贸易的影响[J]. 商业经济研究, 2025(7): 127-130.
- [8] 马海群, 张斌. 基于 LDA 模型的我国开放公共数据政策供给特征分析[J]. 现代情报, 2023, 43(8): 35-44.