◆生产场景◆

基于 WAE 主题模型和社会网络的农产品在线交互研究——以“沙糖桔吧”为例*

岳　芳,高子雅,樊茂瑞,张发明,肖吉军**
(桂林电子科技大学商学院,广西桂林　541004)

摘要:为了更好地利用农业信息资源,本研究将百度贴吧“沙糖桔吧”中的数据作为研究对象,对无序化的农产品在线交互信息进行数据挖掘,引入大数据技术促进农产品信息传播。首先,构建基于 Wasserstein 的自动编码器(Wasserstein Auto-Encoder,WAE)和高斯先验的主题模型(WAE-G),将其应用于交互信息的主题识别和提取;然后,利用社会网络分析方法对交互用户进行社区聚类,探究聚类主题及网络中的关键用户。实证分析结果验证了该主题模型的有效性,所提取的主题信息能较好地反映用户的信息需求、市场供需现状以及该产业发展的瓶颈。同时,社会网络分析结果发现了不同社区用户交流的主题差异,以及关键用户在信息传播中的作用。

关键词:农产品;交互行为;Wasserstein 自动编码器;主题分析;社会网络分析
中图分类号:G206,TP391　　**文献标识码**:A　　**文章编号**:1005-9164(2024)05-0976-12
DOI:10.13656/j.cnki.gxkx.20240002.001

党的二十大报告中提出“加快发展数字经济,促进数字经济和实体经济深度融合”,这为全面推动数字乡村建设、助力乡村振兴指明了方向。发展数字农业契合我国乡村振兴发展战略要求,也是我国农业产业发展的现实要求[1]。农业农村部相继出台《关于推进农业农村大数据发展的实施意见》和《“十四五”全国农业农村信息化发展规划》等相关政策文件,以数字化推动农业经营增收和农产品质量提升[2]。借助大数据技术来开发利用农业信息资源,对数字农业的建设和发展具有重要意义[3]。

随着互联网的快速发展,农产品在线交互信息已逐渐成为农业信息资源的主体内容之一。然而,我国对这类农业信息资源开发利用不足,缺乏对其高效整合与共享[4]。目前有关农产品在线交互信息的研究主要集中在销售平台中的评价数据和社交网络中的舆论数据,而忽视了社交平台中有关农产品的在线交

收稿日期:2023-07-07　　修回日期:2023-11-10
*国家自然科学基金青年科学基金项目“时变交互网络下知识协同中的观点演化机制研究:基于动力学视角”(72001054),国家自然科学基金项目“复杂信息下的动态激励评价模型与方法研究”(72161006)和桂林电子科技大学研究生教育创新计划项目“开放式知识平台用户知识协同中的群体观点演化模型研究”(2023YCXS086)资助。

【第一作者简介】
岳　芳(1982—),女,讲师,主要从事知识管理、知识协同等研究。
【**通信作者简介】
肖吉军(1975—),男,教授,主要从事工业工程研究,E-mail:84770771@qq.com。

【引用本文】
岳芳,高子雅,樊茂瑞,等.基于 WAE 主题模型和社会网络的农产品在线交互研究:以“沙糖桔吧”为例[J].广西科学,2024,31(5):976-987.
YUE F,GAO Z Y,FAN M R,et al. Online Interaction of Agricultural Products Based on WAE Topic Model and Social Network:A Case Study of “Sugar Orange Bar” [J]. Guangxi Sciences,2024,31(5):976-987.

互数据。这些研究大多是从消费者和企业的角度出发,较少关注农业生产者的需求。例如,人们利用百度贴吧交流农产品种植与出售方面的信息,用户之间的发帖和回帖行为产生的信息数据包含了市场信息、种植经验以及用户情感等内容,但这些海量数据存在无序化的问题,目前尚未得到有效开发和利用。此外,由于用户知识水平参差不齐,发帖和回帖数据不规范、长短不一,在主题识别时会导致主题连贯性和一致性效果较差,难以从中挖掘有效的信息。

为了更好地利用农产品在线交互信息,本研究以百度贴吧中的实际交互数据为研究对象,分析用户间的交流信息和交互行为,同时,为了提升数据挖掘的效果,针对数据特点构建一种新的主题模型。通过主题模型与社会网络分析方法,分析相关用户的信息需求和用户聚类间的主题差异,以及关键用户对信息交流与传播的作用。本研究结果对在线交互平台的可持续发展有着重要意义,有助于平台更好地满足用户的信息需求,促进农产品信息交流与共享。

1 相关工作

1.1 农产品在线交互数据研究

开发和利用交互平台中的农产品信息资源,有助于促进数字农业的快速发展。关于农产品的在线交互数据在各类农业网站、论坛、社交网络以及销售平台大量存在。目前对农产品交互数据的研究主要集中在两个方面:一是销售平台中的在线评论数据,分析购买评论对商家口碑、消费者购买行为以及农产品网络销量的影响[5-7];二是社交媒体中的在线评论数据,分析农产品质量安全谣言的传播和公众负面情绪对农产品价格的影响[8]。上述研究大多是从消费者和企业的角度出发,而较少有学者从农产品生产者的角度出发进行研究。这反映出从生产到消费的整个流通过程中,人们往往更多关注农产品最终的销售环节,而忽视了最初的生产环节。农产品生产者难以及时获取足够的消费者需求信息,导致种植和销售都处于一种盲目状态。此外,由于沙糖橘的品质与地理环境、气候变化、种植水平等直接相关,生产者还可能面临一定的种植风险,收益难以保障。因此,研究在线交互数据对农产品生产者具有十分重要的意义。通过深入分析在线交互平台中农产品的主题信息以及用户交互行为,可以更好地了解用户的信息需求以及关键用户在农产品信息传播网络中的作用机制,有助于提高农业信息资源的传播与共享的效率,为平台的长远发展提供针对性的建议。

1.2 主题分析

农产品在线交互数据中包含大量的用户评论数据,借助主题模型方法,可从无序化的文本数据中发现用户的信息需求。目前旨在挖掘文本中潜在语义关系和隐藏主题信息的主题模型被广泛用于识别网络社区中的主题、信息需求等。例如,吴江等[9]通过 BERT 方法对网易云音乐社区中的短文本评论进行主题挖掘,获得用户自我呈现的主题信息;张瑾等[10]构建主题模型提取教师在线学习社区中的会话主题,为学习社区管理者预测和干预教师学习状态提供参考;周国韬等[11]通过潜在狄利克雷分配(Latent Dirichlet Allocation,LDA)[12]主题模型识别知乎中养生问答数据的主题信息,分析用户养生健康信息需求的变化情况。以上研究表明,借助主题模型可从海量的无规则文本数据中快速识别主题信息,把握用户的信息需求,进而有针对性地采取管理措施。

目前针对文本数据的主题识别模型研究较多,其中以 LDA 为代表的概率主题模型[13]得到了广泛应用。它们以无监督的方式探索大型文档集合,能够获得语义空间中的隐式层次语义信息。近几年,随着变分自编码器(Variational Auto-Encoder,VAE)的出现,深度神经网络与此类概率主题模型结合。基于神经网络的模型通过重构主题模型的文本生成过程,以生成更具表达力的主题词。例如,Miao 等[14]基于 VAE 提出了神经变分主题模型(Neural Variational Document Model,NVDM);Wang 等[15]基于生成式对抗网络(Generative Adversarial Nets,GANs)提出了对抗神经主题模型(Adversarial-neural Topic Model,ATM)。之后,谷歌提出了 Wasserstein 自动编码器(Wasserstein Auto-Encoder,WAE)[16],将 WAE 迁移到自然语言处理中时,其在长短类型文本的适应性上发挥了显著优势,并且解决了 VAE 后验坍塌的问题。Nan 等[17]基于 WAE 和 LDA 主题模型构建了 W-LDA 神经网络主题模型,在处理大规模数据集时效果显著。随着社交网络的多样化发展,数据文本类型呈现出多样化、规模大和不规范等特点,很多研究学者开始基于 LDA 主题模型的基本框架,从模型假设、主题数量、参数推断方式、监督式模型等角度提出各种改进方法,以解决模型泛化能力弱、主题词可解释性差和分类准确性低等问题[18]。

1.3 社会网络分析

农产品在线交互平台中的用户评论数据承载着

各种信息,用户行为决定着平台的长远发展,因此对用户行为进行研究同样重要。有研究表明,少数关键用户拥有丰富的专业知识,并愿意将经验分享给他人,为平台的发展做出了重要的贡献[19]。针对关键用户的研究已成为学术界的热点话题,众多学者从不同的角度进行了探讨。李长荣等[20]从情感的角度分析了突发公共卫生事件中在线健康社区关键用户的情感特征,有助于监管部门控制舆情的蔓延。张军等[21]从“行为、内容、情绪”3 种用户属性出发,识别了在线健康社区中的关键用户,为社区的健康运营提供决策支持。杨瑞仙等[22]从知识流动的角度识别了学术虚拟社区“丁香园论坛”中的核心用户,揭示了核心用户间的知识流动特征。这些研究表明,对关键用户的识别与分析对于在线交互平台的可持续发展具有重要的意义。

在线交互数据中,用户之间的连接呈现出复杂网络的结构特征,借助社会网络分析工具,可以对这一复杂网络中的用户进行深入分析。社区发现算法在此过程中发挥着关键作用,可根据节点之间的连接紧密程度,对网络中的个体进行有效的社区分类,有助于更好地理解用户行为模式和传播机制。社区发现算法主要包括分裂式算法 Girvan and Newman(G-N)算法[23]、基于模块度优化的 Louvain 算法[24]以及从信息论视角提出的 Infomap 算法[25],它们被广泛应用于用户分析、信息检索及推荐系统等方面[26]。基于模块度优化的 Louvain 算法在大型网络中具有较好的应用效果,常用于社交网络的社区发现。王曰芬等[27]以新浪微博“上海踩踏”事件为研究对象,基于社区发现算法和关键节点识别发现网络舆情;张海涛等[28]通过社区发现算法,将新浪微博“中美贸易摩擦”话题中的整个网络分解为多个话题社群;贾若男等[29]利用 Louvain 算法,对新型冠状病毒疫情期间的“双黄连”辟谣事件的主体关系网络进行社区划分,发现辟谣信息网络中的关键信息主题类型和网络社区结构。综上,通过社区发现算法对用户进行社区划分和分析,可以更好地从大量无序化的用户交互行为中发现有用信息。

以上都是针对主题信息或网络用户的研究,而实际上,主题信息与用户互动是相辅相成的。用户在平台中发表评论,而这些评论包含了丰富的主题信息。不同观点的评论,反过来也会刺激和鼓励用户发表新的内容。为了更全面地分析在线交互信息,本研究以农产品主题百度贴吧中的在线交互数据为研究对象,同时从主题信息和网络用户两个维度分析农产品在线交互信息。首先,构建了一个基于 WAE 的主题模型,并利用该主题模型对评论信息进行主题识别与分析,挖掘农产品信息的主题特征;然后,利用社会网络分析方法对用户交互信息进行社区聚类,并分析不同社区聚类的主题,识别关键用户。通过上述研究,能够更全面了解农产品信息交流的主题特征与网络传播结构,为农产品在线交互平台的长远发展提供科学的建议,更好地满足人们对相关信息的需求,对促进农产品信息的传播与共享具有重要意义。

2 方法

本文在主题模型和社区发现算法的理论框架下,挖掘农产品在线交互数据中的主题信息及社会网络信息,并尝试构建基于 WAE 和高斯先验的主题模型(WAE-G),进一步提高主题识别的多样性和语义连贯性。本文的研究框架如图 1 所示。

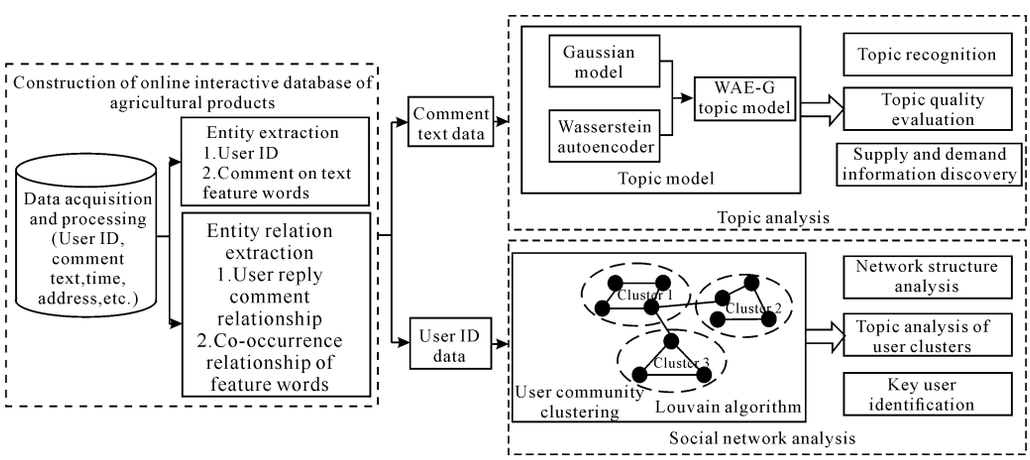

图 1 研究框架

Fig. 1 Research framework

2.1 数据库构建

2.1.1 数据采集

百度贴吧成立于 2003 年,是一种典型的在线交互平台,让对同一个话题感兴趣的人们聚集在一起,方便展开交流和互相帮助[30]。“沙糖桔吧”主要是由沙糖橘果农和果商构成的在线交互平台,用户相互交流和分享种植经验、销售、购买信息等,本质是一个农产品交互平台。截至 2022 年 3 月,“沙糖桔吧”中共有主题数 4 000 多个,会员数 6 000 多人,发布帖子数 40 000 多篇。考虑到百度贴吧是以关键词为主题建立的交互平台,用户比较集中,且其中的数据时间跨度大,相对完整,本文选择“沙糖桔吧”用户的在线交互数据作为研究对象。首先,通过 Python 编写爬虫程序,采集该贴吧从 2006 年 2 月至 2022 年 3 月用户的所有在线交互数据,包括发帖人 ID、发帖内容、发帖时间、回帖人 ID、回帖内容和回帖时间等;然后,对数据进行清洗,包括删除缺失值、空字符串、重复字段、表情符号以及无效值等,共得到 15 152 条数据,涉及 4 243 个用户;最后,按照发帖与回帖的关系对数据进行合并处理作为评论数据,共得到 2 115 个文档。

2.1.2 数据处理

抽取评论数据实体:利用集搜客进行分词,包括添加自定义词表(如“盖膜”“炭疽病”“秋梢”等)、创建归并词表和启动过滤词表进行分词,筛选合并后得到 478 个特征词实体。

对于用户 ID,首先抽取实体。基于 ROST CM6 进行分词,由于用户 ID 的特殊性,将全部的用户 ID 添加到自定义词表,得到 4 243 个用户 ID 实体。然后抽取实体关系,依照发帖、回帖之间的关系建立联系,关系数为 10 850 个。

2.2 基于 WAE 和高斯先验的主题模型(WAE-G)构建

2018 年谷歌提出的 WAE 在进行自然语言处理的任务时,能够较好地适应不同长短类型文本。W-LDA 是最早基于 WAE 构建的主题模型,与基于 VAE 的模型不同。VAE 在训练过程中会强制将所有样本的后验分布与先验分布匹配,导致隐变量分布之间出现重叠问题,进而影响重构的效果;而 WAE 则是将先验分布与聚合后验相匹配,因此不同样本的隐变量可以被分隔开,增强了输入的重构效果。此外,相比于 VAE 使用的 KL 散度(Kullback-Leibler Divergence),WAE 使用 Wasserstein 距离来计算重构损失,即使两个概率分布间差异很大也可以衡量它们之间的距离。另外,本文使用的数据集主要包含沙糖橘的相关信息,主题相对较少。为了更准确地识别评论数据集中的主题,本研究使用高斯分布代替 W-LDA 中的 LDA 作为文档主题分布的先验分布,并通过分布匹配对文档主题分布施加先验分布的约束。

2.2.1 WAE

WAE 是一种用于构建数据分布生成模型的算法,结合了 VAE 与 GANs 的优点与结构。与 VAE 结构相似,基于 WAE 的损失函数由重构损失项和正则项组成。

WAE 的重构损失项使用 Wasserstein 距离来衡量原输入数据 X 和重构数据 $\hat{X}$ 之间的重构损失。正则项使用最大均值差异(Maximum Mean Discrepancy,MMD)衡量数据特征 Z 的先验 $p(z)$ 和后验 $q(z \mid x)$ 之间的距离。WAE 的损失函数可以表示为

$$D_{\mathrm{WAE}}[P_X, P_G] = inf_{Q(\theta|X)\in Q} E_{P_X} E_{Q(Z|X)}[c(X, G(Z))] + \beta \cdot D_Z(Q_Z, P_Z), \quad (1)$$

式中,第一项 $E_{P_X} E_{Q(Z|X)}[c(X,G(Z))]$ 是基于 Wasserstein 距离的重构损失项,其中,P_X 是 X 的先验,表示数据的真实分布,P_G 是由 G 生成的分布项,$G(\cdot)$ 可以将 Z 映射到 X 的解码器($G:Z \to X$)。$c(X,G(Z))$ 是原始数据 X 和重构数据 $\hat{X}$ 之间的损失函数,$E_{P_X} E_{Q(Z|X)}$ 是样本的聚合后验分布。

第二项 $\beta \cdot D_Z(Q_Z, P_Z)$ 是用基于 MMD 构成的正则化项,衡量了聚合后验分布 Q_Z 与先验分布 P_Z 之间的差异,其中,$Q_Z := E_{P_X} E_{Q(Z|X)}$,表示编码器经过训练后的样本分布,是数据特征 Z 的聚合后验分布,P_Z 是数据特征 Z 的先验分布,$D_Z(\cdot,\cdot)$ 是 Q_Z 与 P_Z 之间的任意散度测量函数,β 是 D_Z 的权值参数且大于 0。

2.2.2 模型的生成过程

首先是文本的表示,对于给定的文档集合 D,$D = \{d_1, d_2, \cdots, d_N\}$,其中 N 表示文档的数目。X 表示文档 D 的词袋,$X = \{x_1, x_2, \cdots, x_N\}$,其中 $x_i \in R^{|V|}$ 表示文档集中第 i 篇文档的词袋,V 表示整个数据集的词汇表。其次,WAE 由编码器和解码器组成,编码器由一个多层感知器(Multi-Layer Perceptron,MLP)组成,能够将输入数据 X 映射为隐变量 Z,Z 经过 softmax 处理获得文档主题向量 θ。而高斯模型将接收隐变量 Z 作为输入,生成模型的高

斯先验 $p(\theta)$。解码器是由多层感知器组成的生成网络,对于给定的文档主题向量 θ 进行解码,解码过程表示为 $P(\hat{X}\mid\theta)$。通过 softmax 将 θ 映射到 v 维的语义空间,获得文档词汇分布 $\hat{X}$。最后,将文档主题分布重构为文档,并输出主题词汇分布。

2.3 社区发现算法

社区发现中的 Louvain 算法是一种基于模块度的社区发现算法。其基本思想是网络中节点尝试遍历所有邻居的社区标签,并选择最大化模块度增量的社区标签。之后,每个社区作为一个新的节点,重复直到模块度不再增大。通过 Louvain 算法对用户进行聚类,可以使相关性较强的用户聚为一类,同时使聚类之间的异质性达到最大化。相比于其他算法,其在大型网络中具有较好的适应性。在“沙糖桔吧”这类大型社交网络中,用户交流的时间不同,关注的主题内容也有差异,自发形成了以一个或多个关键用户节点为中心的网络社区,这些社区通常围绕某一个主题进行讨论。因此为了更好地发现用户之间的联系以及交流主题,利用 Louvain 算法将用户进行社区聚类。

在此将用户之间发帖、回帖交互的次数作为节点之间边的权重。模块度的计算公式如下:

$$Q=\frac{1}{2m}\left[\sum_{i,j}A_{ij}-\frac{\sum_i k_i\sum_j k_j}{2m}\right]\delta(c_i,c_j)=\frac{1}{2m}\sum_c\left[\sum_{in}-\frac{\left(\sum_{tot}\right)^2}{2m}\right],\tag{2}$$

其中,A_{ij} 表示节点 v_i 和节点 v_j 之间边的权重;$k_i=\sum_i A_{ij}$ 表示所有与节点 v_i 相连的边的权重之和;$m=\frac{1}{2}\sum_{ij}A_{ij}$ 表示所有边的权重之和;c_i 表示节点 v_i 所在的社区,$\delta(c_i,c_j)$ 为判断节点 v_i 和节点 v_j 是否在同一社区的函数,是则值为 1,否则值为 0;$\sum_{in}$ 表示社区 C 内的边的权重之和;$\sum_{tot}$ 表示与社区 C 内的节点相连的所有边的权重之和。

模块增益度(ΔQ)是评价本次迭代效果好坏的数值化指标,计算公式如下:

$$\Delta Q=\left[\frac{\sum_{in}+k_{i,in}}{2m}-\left(\frac{\sum_{tot}+k_i}{2m}\right)^2\right]-\left[\frac{\sum_{in}}{2m}-\left(\frac{\sum_{tot}}{2m}\right)^2-\left(\frac{k_i}{2m}\right)^2\right]=\frac{1}{2m}\left(k_{i,in}-\frac{\sum_{tot}k_i}{m}\right),\tag{3}$$

其中,$k_{i,in}$ 表示节点 v_i 与该社区 C 内的节点相连的权重总和。

3 实验与结果分析

3.1 主题分析

3.1.1 基于 WAE-G 的主题识别

基于 WAE-G 对评论数据进行主题识别,其中迭代次数设置为 2 000 次,并过滤频率高于 0.023 的文档特征词。

①评价指标

本文通过主题多样性(Topic Diversity,TD)和主题连贯性(Topic Coherence,TC)两个评价指标来判断主题识别的质量[31,32]。TD 由 K 个主题排名前 N 个单词中唯一词语出现的百分比来确定,取值范围[0,1],接近 0 表示主题冗余,接近 1 表示更加多样的主题。TD 的计算公式:

$$\mathrm{TD}=\frac{\mathrm{unique}(w_{\mathrm{top}N}^{(1)},w_{\mathrm{top}N}^{(2)},\cdots,w_{\mathrm{top}N}^{(K)})}{K\times N}。\tag{4}$$

对于 TC,采用 Röder 等[33]提出的自动度量指标 C_{V} Coherence 来验证。C_{V} Coherence 基于滑动窗口,对主题词进行 one-set 分割,并使用归一化逐点互信息和余弦相似度间接获得连贯度,其取值范围为[0,1],该指标越高,即代表主题的连贯性越好,主题模型效果越好。C_{V} Coherence 的计算公式:

$$C_{\mathrm{V}}=\mu(\{S_{\cos}(\vec{u},\vec{w})\mid\vec{u},\vec{w}\in W\})。\tag{5}$$

②主题个数的确定

为了确定最佳主题个数,本文定义了兼顾 TD、TC 的整体质量(OQ):

$$\mathrm{OQ}=2\times\frac{\mathrm{TD}\times\mathrm{TC}}{\mathrm{TD}+\mathrm{TC}},\tag{6}$$

OQ 越大,意味着提取的主题质量越好。对于这一数据集设置最小主题数为 2,最大主题数为 15,统计在 WAE-G 下迭代 2 000 次的不同主题个数的 OQ,结果如图 2 所示。当主题数为 2 时 OQ 达到最大值,但是考虑 2 个主题数太少,主题划分不够具体,故选择第一个极大值点,即主题数为 4。

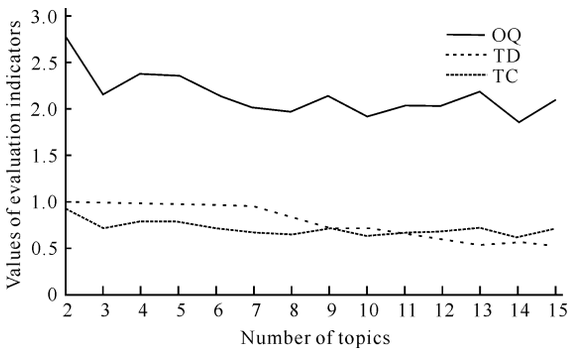

图 2　不同主题数下的 OQ

Fig. 2　OQ with different number of topics

为了验证该模型的主题识别效果,将该模型与现有的基准模型进行比较。一是经典的 LDA 主题模型,该模型被广泛应用于多个领域的主题研究;二是基于 LDA 的狄利克雷先验主题模型 W-LDA,本研究提出的主题模型是在该模型的基础上将高斯先验代替狄利克雷先验;三是基于 WAE 和高斯混合先验主题模型 WAE-GMM,以验证高斯先验优于高斯混合先验。结果如表 1 所示,4 种主题模型都可以在主题多样性上发挥较好的作用,但是在主题连贯性上 WAE-G 明显取得较好的效果。同时,为了增强该模型的说服力,增加百度贴吧“脐橙吧”中的短文本数据进一步验证该模型的有效性。因此,在保证所有参数相同的情况下,验证 4 种主题模型在“沙糖桔吧”和“脐橙吧”2 个数据集上的评价指标 TD 和 TC。根据表 1 可以看出,相较于其他 3 个主题模型,WAE-G 在 2 个数据集中 TD 和 TC 评价指标均取得较好的效果。

表 1　主题识别对比结果

Table 1　Comparison results of topic recognition

主题模型 Topic model	“沙糖桔吧”数据集 “Sugar Orange Bar” dataset		“脐橙吧”数据集 “Navel Orange Bar” dataset	
	TD	TC	TD	TC
LDA	**1.000**	0.677	**1.000**	0.502
W-LDA	**1.000**	0.442	0.983	0.558
WAE-GMM	0.983	0.614	0.983	0.435
WAE-G	**1.000**	**0.789**	**1.000**	**0.680**

Note: bold data in the table indicate the highest scores of the methods in comparison.

根据前面的结果将分词后的评论数据集通过 WAE-G 进行主题识别,设置主题数为 4,每个主题设置 14 个可以体现主题语义的特征词语,结果如表 2 所示。

表 2　“沙糖桔吧”评论数据主题识别结果

Table 2　Topic recognition results of “Sugar orange Bar” comment data

主题编号 Topic ID	主题描述 Topic description	主题特征词 Topic feature words
Topic 1	Sale information	永福、象州、金秀、贺州、来宾、柳州、同步、象州县、实地考察、韶关、实力、议价、鹿寨、西平
Topic 2	Promotion information of purchasing agents	组织过秤、收取、热情周到、手续费、方便快捷、食宿、佛冈县、新鲜、满意、云安、京珠高速公路、国道、批发商、无核
Topic 3	Varieties of fruit tree seedlings	嫁接、果苗、便宜、皇帝柑、枳壳、湖南、适合、金秋、盆栽、病毒、树苗、桔树、苗木、枝条
Topic 4	Planting technique	陈二哥、元素、面肥、施肥、效果、复合肥、肥料、水肥、杀菌、原因、有机肥、红蜘蛛、打药、方法

根据主题识别的结果可以发现,在“沙糖桔吧”中用户讨论的话题主要有 4 个。Topic 1 是关于沙糖橘出售,主题特征词中大部分都是沙糖橘的产地,这也体现出果农为了出售沙糖橘,会在网上发布很多关于出售的地址信息。Topic 2 是关于代办中间商,果农大多是种植散户,他们很难直接联系外地收购商,所以存在很多专门收购沙糖橘的代办,以及集中收购外销的档口,主题特征词体现了他们的招商信息和服务内容。Topic 3 是关于果苗,主题特征词包含了果苗的种类、培养方法以及种植条件等信息,表明人们十分重视果苗的选择,选择优质的果苗是种植沙糖橘的重要环节。Topic 4 是关于种植技术,从主题特征词中可以看出,“陈二哥”是有关种植技术的核心人物,种植方面的问题主要集中在如何施肥和解决病虫害上。总的来看,沙糖橘种植面临的困境主要集中在解决种植技术和出售两个问题上。

3.1.2　供需信息发现

在 Topic 1 出售信息这一主题中,地址特征词尤为突出,充分利用这些地址数据信息,可以发现沙糖橘种植和销售的地理分布情况。因此,将用户作为节点,发帖与回帖作为连接关系,根据地址的经纬度信息确定节点位置,构建关于种植与销售的用户地理分布图谱,其中包括 2 个聚类,果商为绿色节点,果农为红色节点(图 3)。

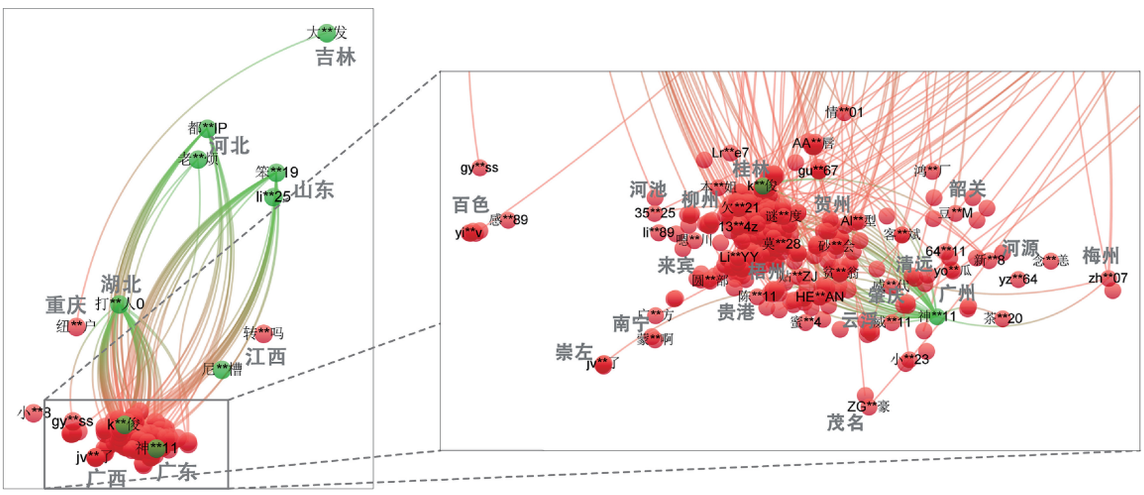

图 3　用户地理分布图谱

Fig. 3　Map of user geographic distribution

根据用户地理分布图谱,可以看出沙糖橘的种植地主要集中在广西和广东两省区,销售地则比较分散,主要是北上广深、华北地区、东北三省地区以及沿海发达地区。将广西和广东两省区的种植分布局部放大分析,其中广西的主要产地有桂林市的阳朔县、荔浦市和永福县等,梧州市的岑溪市、苍梧县和蒙山县等;广东的主要产地有云浮市、清远市和韶关市等(图 3)。

图 3 中果农的数量远远大于果商的数量,这反映出果农的出售信息需求要远大于果商的收购信息需求,也反映出果农对市场信息掌握较少,只能被动等待果商来收购。为了分析沙糖橘的市场供需情况,统计数据中的出售信息和收购信息在时间上的变化情况(图 4)。可以发现,第一,这两类信息的频次在时间分布上具有明显的周期性,在每年的 10 月份到第二年的 3 月份沙糖橘的上市期间达到峰值。第二,出售信息的频次远大于收购信息的频次。

为直观反映市场供需状况,进一步计算出售信息与收购信息的比值,结果如表 3 所示,2017 年和 2018 年出售信息与收购信息的比值最大,根据中国天气网广西站新闻报道,2018 年 1 月上旬中后期受霜冻天气影响,广西地区 2017 年上市的沙糖橘品质受到严重影响,大量滞销,价格大跌。另外 2018 年上市的沙糖橘也出现大量滞销问题,广西梧州市多地人民政府曾出台《关于抓好 2018 年砂糖桔销售工作的紧急通知》[34],要求相关部门引导果农尽快将沙糖橘出售,确保果农增收。可见数据分析结果与实际情况一致,通过对出售与收购信息进行统计分析,可反映和预测沙糖橘的市场供需情况,以指导果农合理安排种植。

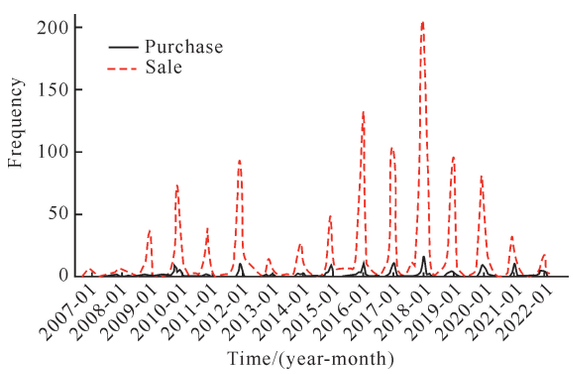

图 4　沙糖橘出售-收购信息时间序列统计图

Fig. 4　Statistical graph of time series of sale-purchase information of sugar orange

表 3　沙糖橘出售-收购信息统计

Table 3　Statistics of sale-purchase information of sugar orange

时间/(年-月) Time/(year-month)	频次 Fequency		比值=出售/收购 Ratio=Sale information/Purchase information
	出售信息 Sale information	收购信息 Purchase information	
2017-10—2018-03	594	32	18.562
2018-10—2019-03	264	17	15.529
2010-10—2011-03	74	5	14.800
2011-10—2012-03	220	16	13.750
2015-10—2016-03	291	23	12.652
2016-10—2017-03	240	22	10.909
2008-10—2009-03	59	6	9.833
2019-10—2020-03	215	27	7.963
2009-10—2010-03	168	22	7.636

续表

Continued table

时间/(年-月) Time/(year-month)	频次 Fequency		比值=出售/收购 Ratio=Sale information/Purchase information
	出售信息 Sale information	收购信息 Purchase information	
2013-10—2014-03	62	9	6.889
2014-10—2015-03	113	18	6.278
2012-10—2013-03	40	7	5.714
2020-10—2021-03	74	16	4.625
2007-10—2008-03	17	5	3.400
2021-10—2022-03	47	17	2.765

3.2 社会网络分析

3.2.1 社区聚类

借助社会网络分析工具,对“沙糖桔吧”在线交互数据中的用户进行社区聚类。利用 Louvain 算法进行聚类,设置分辨率参数为 1,随机重启次数为 1,每次重新启动最大迭代次数为 20,每次迭代中的最大级别数为 20,每一级的最大重复次数为 50,并利用 VOSviewer 绘制用户图谱。图 5 共划分了 26 个聚类,主要对较大的 8 个聚类进行分析。节点的大小由用户的总连接强度决定,节点越大表示该节点与其他节点连接的总次数越多,也表明其越活跃。两节点之间连线的粗细由两实体之间的连接次数决定,连接次数越多,连线越粗,两节点之间的连接越紧密。对用户图谱的网络结构进行分析,用户平均度为 4.365,网络直径为 12,平均聚类系数为 0.143,说明该网络呈现小世界特征,并且部分节点出现聚类或抱团的迹象。网络特征向量中心度系数为 0.002,平均路径长度为 4.539,网络模块度 $Q=0.653$,有研究表明当模块度值在 0.3—0.7 时,说明该网络中出现了较强的社区结构[35]。

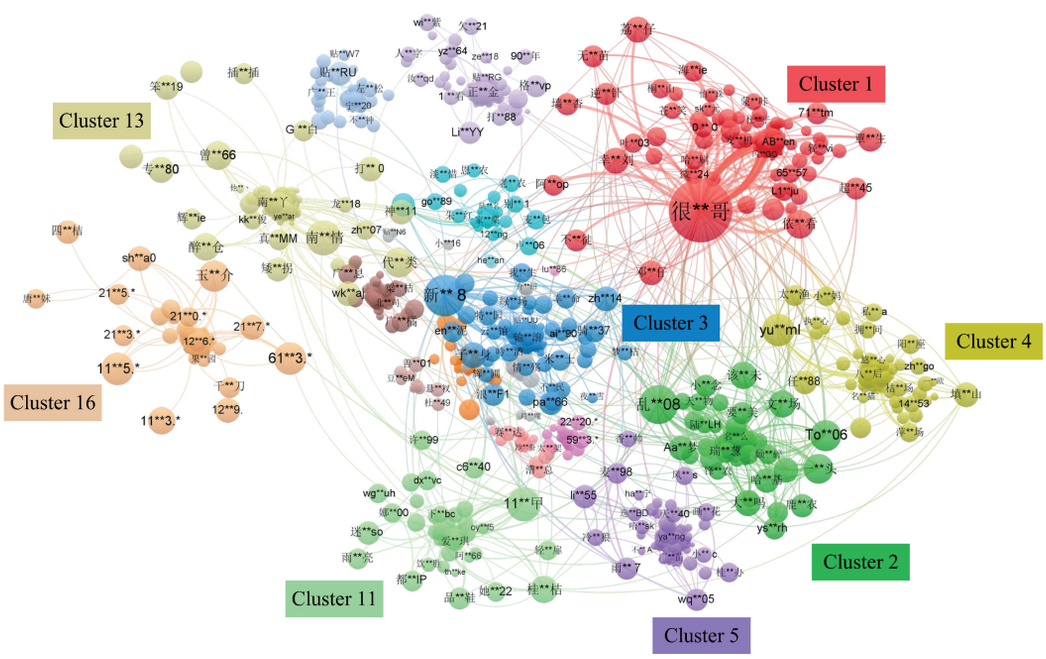

图 5 用户社区聚类图谱

Fig. 5 Cluster map of user community

为了分析这 8 个聚类中用户交流的主题,借助词云图对各个聚类中用户评论的高频特征词进行可视化呈现(图 6)。聚类 1、聚类 2 和聚类 5 主要讨论的是种植技术信息;聚类 3 主要讨论的是种植技术和交易信息;聚类 4 主要讨论的是果苗的交易信息;聚类 11、聚类 13 和聚类 16 主要讨论的是沙糖橘的收购与出售信息。由于时间的跨度长,讨论相同主题信息的用户也被划分成不同的聚类。可见,将社区发现算法用于用户社区聚类具有较好的效果。

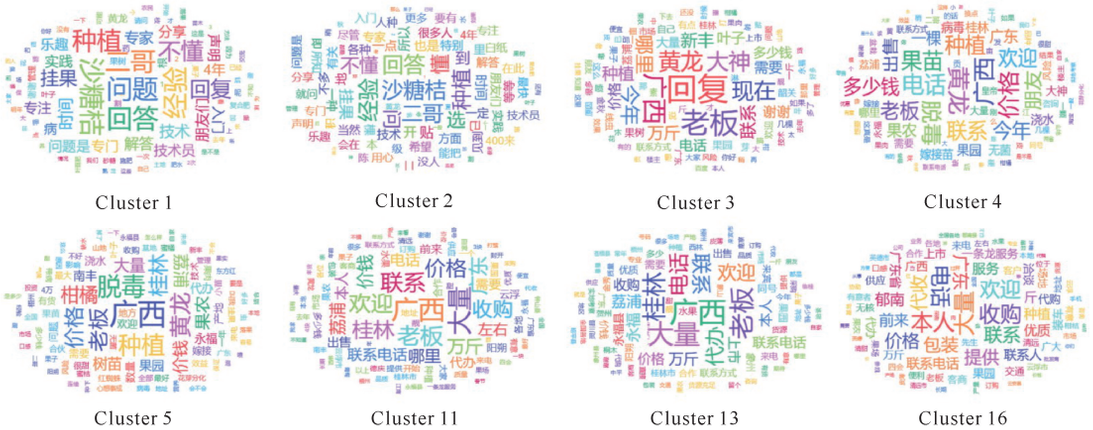

图 6　用户社区聚类主题词云

Fig. 6　Topic word cloud of user community clusters

以上用户聚类的交流主题可归纳为 4 种,分别是种植技术信息、种植技术与沙糖橘出售信息、果苗出售与购买信息、沙糖橘出售与购买信息。结合时间信息,绘制各个聚类在主题和时间两个维度上的分布(图 7)。在 2015 年之前存在较少大的聚类,大多是关于沙糖橘出售与收购方面的信息,种植技术的讨论也是出现在 2015 年之后。分析其原因,是由于用户图谱中节点很大的“很 * * 哥”在 2015 年开始分享并解答关于种植技术方面的问题,引起很多果农的交流。这也体现了网络中的关键用户对于促进信息共享和交流具有重要作用,因此对于关键用户的识别也具有重要意义。

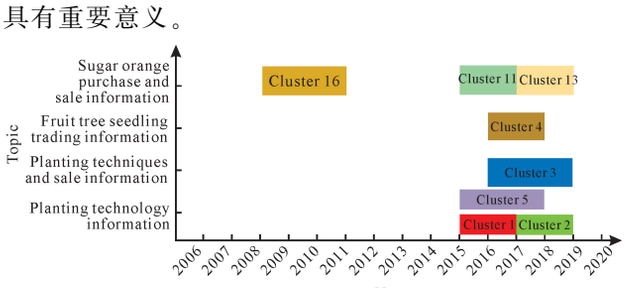

图 7　用户社区聚类主题分布

Fig. 7　Topic distribution of user community clusters

3.2.2　关键用户识别

沙糖橘的种植,从选苗到挂果每一个环节都会影响到果实的品相、口感和产量,而果农的种植技术参差不齐。目前针对果农开展的农业信息服务仍以电话直接咨询、开展种植技术讲座、集中技术培训和专家现场指导为主,熟悉农业生产活动、掌握信息资源的从业人员和专家的数量远不能满足需求[36],而且时效性和精准性较差。“沙糖桔吧”中一些种植经验丰富的用户愿意在线分享种植经验,使得有技术需求的果农能够得到更为及时和有针对性的帮助。但是,图 5 中一些较大的节点,如“很 * * 哥”、“新 * * 8”、“南 * * 情”、“乱 * * 08”等用户,仅反映出他们与其他用户连接较多,发言积极活跃,难以识别他们在网络中的影响机制。因此为有效促进在线种植技术相关知识的传播,科学识别关键用户就显得尤为重要。

对于社会网络中的用户节点的影响力,可以通过度中心性、中间中心性和特征向量中心性等多个指标从不同角度进行衡量。对于这些社会网络分析指标,可依据用户之间的连接关系构建邻接矩阵,利用 UCINET 软件计算得到,结果见表 4。

表 4　用户节点中心性指标

Table 4　Centrality indicators of user nodes

序号 No.	用户 ID User ID	度中心性 Degree centrality	中间中心性 Betweenness centrality	特征向量中心性 Eigenvector centrality
1	很 * * 哥	1 122	1 096 484.625	98.424
2	南 * * 情	261	267 478.531	1.595
3	乱 * * 08	237	483 970.031	23.203
4	11 * * 甲	210	409 564.406	1.463
5	11 * * 5. *	208	0.000	0.381
6	新 * * 8	192	566 111.000	1.578
7	61 * * 3. *	192	0.000	0.051
8	玉 * * 介	176	1 765.160	0.006
9	yu * * ml	141	178 106.891	2.381
10	T * * 06	136	233 900.516	14.663

Note: the table lists top 10 users from a total of 4 243 users based on the degree centrality.

表 4 中有些用户度中心性很大,但中间中心性很小甚至为 0,比如用户“11 * * 5. * ”、“61 * * 3. * ”和“玉 * * 介”,这些用户均为图 5 聚类 11 中的用户,讨论的主题信息主要是“沙糖橘购买与出售信息”。

这些用户身份都是沙糖橘收购商,发布收购信息之后,很多果农对其进行跟帖,但他们很少对其他用户进行回复,这些用户影响力较小,并不能促进用户之间的信息交流和知识传播,因此考虑采用一种可以综合各指标的方法来确定关键用户。

为了综合考虑每项评价指标,引入综合评价 TOPSIS 算法对每个用户的影响力进行评价。首先确定每一评级指标的正理想值和负理想值,然后利用欧式距离公式 $d_i^+ = \sqrt{\sum_j^3 (Z_{ij} - Z_j^{\max})^2}$ 和 $d_i^- = \sqrt{\sum_j^3 (Z_{ij} - Z_j^{\min})^2}$,计算第 i 个用户评分值与正、负理想值之间的距离,最后再根据公式 $S_i = \frac{d_i^-}{d_i^+ + d_i^-}$,综合每个评价指标与最优值的接近程度和最劣值的远离程度,即综合评价指数 S_i,从而评价第 i 个用户在网络中的影响力。

由于表 4 中的评级指标统计标准不同,可先对表 4 中的结果进行标准化处理,然后根据综合评价 TOPSIS 算法,计算每个用户的综合评价指数。由于用户数量较多,选择影响力排名前 100 的用户,并按照影响力由大到小编号,绘制用户综合评价指数折线图(图 8)。

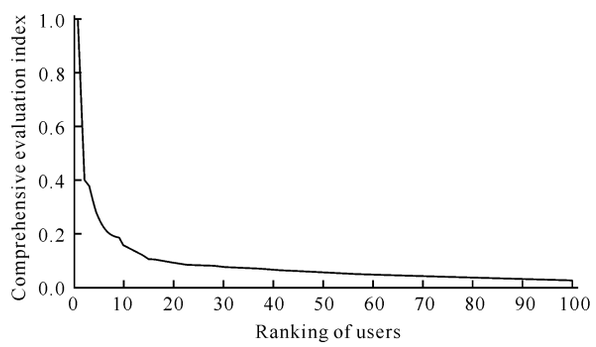

图 8 用户综合评价指数折线图

Fig. 8 Line chart of user comprehensive evaluation index

从图 8 可以看出,排名第一的“很 * * 哥”影响力远远高于其他用户,该用户在百度贴吧中分享相关的种植技术,在网络结构中具有很大的影响力。排名第二至第九的用户具有相对较高的综合评价指数,之后综合评价指数快速下降,因此可以将排名前 9 的用户作为关键用户。对于种植经验丰富且乐意分享的“很 * * 哥”,以及发言较为活跃且在网络中促进信息交流的其他关键用户,应该给予相应的鼓励与引导,激励他们分享种植技术,帮助更多新手解决种植难题。

实证研究表明,从评论数据的主题识别结果来看,“沙糖桔吧”中交流的内容主要集中在沙糖橘出售、种植技术交流、代办与宣传、果苗的栽培与交易 4 个方面。对于果农来说,沙糖橘的种植与出售是他们面临的最大难题。其中沙糖橘出售信息中地址信息尤为突出,通过用户地理分布图谱和出售与收购信息的统计分析结果,可以发现市场的供需分布情况。通过主题识别发现用户的信息需求以及沙糖橘产业存在的痛点,相关部门应及时给予针对性的帮扶,同时也可以根据以往市场信息预测市场需求,避免果农盲目种植带来的经济损失。从用户交互数据分析结果来看,不同用户聚类交流的主题存在一定差异,“沙糖桔吧”初期,用户主要讨论沙糖橘出售与收购方面的信息,到 2015 年随着一些关键用户的出现,用户逐步增加对种植技术和果苗出售等信息的讨论,其中一些关键用户在促进种植技术交流上发挥了重要作用。

4 结束语

本文将文本挖掘技术应用于百度贴吧农产品在线交互数据,从信息主题和用户网络两个维度出发,分析了农产品信息的传播特征。研究发现,农产品在线交互数据中隐含了用户的信息需求、市场供需现状以及产业发展的痛点等信息,并且其中的关键用户在促进话题讨论和知识共享上起重要作用。综上,为促进农产品信息的传播与共享以及百度贴吧等在线交互平台的可持续发展,提出以下建议。

①对农产品信息进行细分。农产品主题吧中的信息具有海量且无序化的特征,平台可以根据用户交流的主题进行细分,这有助于用户快速高效地获取和利用相关信息,对于传播农产品种植知识和促进农产品交易具有重要意义。

②建立奖励机制。鼓励百度贴吧中的用户积极参与信息交流,促进农产品信息的传播。例如,可以通过积分或者等级等对关键用户进行激励,对于优质信息的分享者也可以提供额外的奖励,留住这些用户并发挥他们作用。

③重视对百度贴吧等平台数据的利用。由于平台用户比较活跃,相关部门可以实时获取农产品种植与销售方面的信息,将其作为传统数据的一个有益补充。及时发现农产品生产环节中的问题和瓶颈,积极采取相应措施。例如,评估农产品供需状态,为果农提供种植建议,减小果农因盲目种植造成的经济损失。

实际上,农产品在线交互平台种类繁多,本研究

仅针对百度贴吧“沙糖桔吧”中的在线交互数据进行了分析。为了更科学、全面地开发利用农产品信息资源,后续的研究将扩大数据来源,增加更多平台的数据进行分析,以便更好地了解农产品在线交互信息的特征和传播规律。

参考文献

[1] 夏玉林,唐剑. 乡村振兴背景下数字农业发展困境与战略研究[J]. 农业经济,2023(5):3-6.

[2] 李丽莉,曾亿武,郭红东. 数字乡村建设:底层逻辑、实践误区与优化路径[J]. 中国农村经济,2023(1):77-92.

[3] 谢康,易法敏,古飞婷. 大数据驱动的农业数字化转型与创新[J]. 农业经济问题,2022,43(5):37-48.

[4] 陈诗. 新媒体背景下农业信息化建设的应用研究[J]. 农村经济,2019(9):110-115.

[5] 陈静,王磊,张海彬. 在线评论对农产品扩散的影响研究[J]. 资源开发与市场,2021,37(3):301-305.

[6] 冯建英,吴丹丹,王博,等. 中文在线评论文本分析对生鲜农产品电商影响研究综述[J]. 农业机械学报,2021,52(S1):504-512.

[7] 胡雅淇,林海. 在线评论特征对生鲜电商农产品销量的影响:来自淘宝羊肉大数据的证据[J]. 中国农业大学学报,2021,26(6):206-218.

[8] LIU Y,LIU S,YE D,et al. Dynamic impact of negative public sentiment on agricultural product prices during COVID-19 [J]. Journal of Retailing and Consumer Services,2022,64:102790.

[9] 吴江,刘涛,刘洋. 在线社区用户画像及自我呈现主题挖掘:以网易云音乐社区为例[J]. 数据分析与知识发现,2022,6(7):56-69.

[10] 张瑾,张夷楠,叶海智,等. 教师在线学习社区中会话主题发现及演化分析[J]. 远程教育杂志,2021,39(2):85-94.

[11] 周国韬,龚栩,邓胜利. 基于主题分析的用户养生健康信息需求演化趋势研究[J]. 情报科学,2022,40(4):118-126,137.

[12] BLEI D M,NG A Y,JORDAN M I. Latent dirichlet allocation [J]. The Journal of Machine Learning Research,2003,3:993-1022.

[13] HOFFMAN M D,BLEI D M,BACH F. Online learning for latent dirichlet allocation [C]//Proceedings of the 24th International Conference on Neural Information Processing Systems - Volume 1. New York:ACM,2010:856-864.

[14] MIAO Y S,YU L,BLUNSOM P. Neural variational inference for text processing [C]//Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48. New York:ACM,2016:1727-1736.

[15] WANG R,ZHOU D,HE Y. ATM:adversarial-neural topic model [J]. Information Processing & Management,2019,56(6):102098

[16] TOLSTIKHIN I,BOUSQUET O,GELLY S,et al. Wasserstein auto-encoders [EB/OL]. (2017-11-05)[2023-06-02]. https://arxiv.org/abs/1711.01558.

[17] NAN F,DING R,NALLAPATI R,et al. Topic modeling with Wasserstein autoencoders [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2019:6345-6381.

[18] 黄佳佳,李鹏伟,彭敏,等. 基于深度学习的主题模型研究[J]. 计算机学报,2020,43(5):827-855.

[19] 王闯,王亚民. 基于 K 核分解的网络知识社区关键用户挖掘研究[J]. 情报理论与实践,2019,42(6):149-153.

[20] 李长荣,纪雪梅,郭凤仪. 突发公共卫生事件中在线健康社区关键用户的情感表达特征研究[J]. 现代情报,2021,41(7):85-93,102.

[21] 张军,李新旺,李鹏. 多维属性融合视角下的在线健康社区关键用户识别研究[J]. 情报科学,2022,40(3):82-90.

[22] 杨瑞仙,郭孟含. 学术虚拟社区核心用户知识流动特征研究[J]. 现代情报,2022,42(8):50-60.

[23] GIRVAN M,NEWMAN M E J. Community structure in social and biological networks [J]. Proceedings of the National Academy of Sciences of the United States of America,2002,99(12):7821-7826.

[24] BLONDEL V D,GUILLAUME J L,LAMBIOTTE R,et al. Fast unfolding of communities in large networks [J]. Journal of Statistical Mechanics: Theory and Experiment,2008,2008(10):P10008.

[25] ROSVALL M,BERGSTROM C T. Maps of random walks on complex networks reveal community structure [J]. Proceedings of the National Academy of Sciences of the United States of America,2008,105(4):1118-1123.

[26] 张海涛,周红磊,张鑫蕊,等. 在线社交网络的社区发现研究进展[J]. 图书情报工作,2020,64(9):142-152.

[27] 王曰芬,王一山,杨洁. 基于社区发现和关键节点识别的网络舆情主题发现与实证分析[J]. 图书与情报,2020(5):48-58.

[28] 张海涛,刘雅姝,张枭慧,等. 基于模块度的话题发现及网民情感波动研究:以新浪微博“中美间贸易摩擦”话题为例[J]. 图书情报工作,2019,63(4):6-14.

[29] 贾若男,王晰巍,孙玉姣.社交媒体中突发公共卫生事件网络辟谣信息主体研究[J].图书情报工作,2021,65(19):16-25.
[30] 岳芳,黄鹏,李小锋.基于社会网络分析的百度贴吧“沙糖桔吧”中的交互行为研究[J].中国管理信息化,2019,22(24):152-154.
[31] 王秀红,高敏.基于 BERT-LDA 的关键技术识别方法及其实证研究:以农业机器人为例[J].图书情报工作,2021,65(22):114-125.
[32] 毛存礼,梁昊远,余正涛,等.基于神经自回归分布估计的涉案新闻主题模型构建方法[J].中文信息学报,2021,35(2):89-98.
[33] RÖDER M,BOTH A,HINNEBURG A. Exploring the space of topic coherence measures [C]//Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. New York:ACM,2015:399-408.
[34] 广西苍梧县人民政府办公室.关于抓好 2018 年砂糖桔销售工作的紧急通知[Z].(2018-01-08).[2023-03-16]. https://www.sohu.com/a/288978453_120052760.
[35] TANG L,WANG X,LIU H. Community detection *via* heterogeneous interaction analysis [J]. Data Mining and Knowledge Discovery,2012,25(1):1-33.
[36] 张博凯,李想.基于知识图谱的 Android 端农技智能问答系统研究[J].农业机械学报,2021,52(S1):164-171.

Online Interaction of Agricultural Products Based on WAE Topic Model and Social Network: A Case Study of “Sugar Orange Bar”

YUE Fang,GAO Ziya,FAN Maorui,ZHANG Faming,XIAO Jijun**
(Business School,Guilin University of Electronic Technology,Guilin,Guangxi,541004,China)

Abstract: To make better use of agricultural information resources,this paper takes the data of “Sugar Orange Bar” on Baidu Post Bar as the research object to conduct data mining on the disordered online interactive information of agricultural products,so as to promote the application of big data technology in the information dissemination of agricultural products. Firstly,based on Wasserstein Auto-Encoder (WAE) and Gaussian prior,the topic model (WAE-G) is constructed and applied to the topic recognition and extraction of interactive information. Then,social network analysis is employed to cluster interactive users and explore the topics between clusters and the key users of the network. The results of empirical analysis verify the validity of the model,and the extracted topic information reflects the information demand of users,the current situation of market supply and demand,and the bottleneck in the development of the industry. At the same time,the results of social network analysis reveal the thematic differences of user communication in different communities and the role of key users in the dissemination of information.

Key words: agricultural products; interactive behaviors; Wasserstein Auto-Encoder (WAC); topic analysis; social network analysis

责任编辑:陆　雁