基于主题模型的产品在线论坛主题演化分析

蒋翠清, 吕孝忠, 段 锐
(合肥工业大学管理学院, 安徽 合肥 230009)

摘要: 产品论坛主题演化分析对企业的市场营销和产品改进决策具有重要价值. 针对产品论坛的特点构建了一个基于潜在狄利克雷分布((latent Dirichlet allocation, LDA)模型的产品在线评论主题演化分析模型, 从主题标签、主题热度和主题词热度三个层面挖掘海量在线产品评论的主题演化. 实验表明, 该方法能够挖掘产品在线论坛的主题演化规律. 发现不同论坛上同一产品的消费者关注点存在共性和差异性, 关注点热度变化存在随机性, 关注中心存在稳定性, 以及高评论丰富度的论坛更容易形成主题演化关系等规律.

关键词: 主题演化; 产品在线论坛; 潜在狄利克雷分布模型; 主题热度
中图分类号: TP273 **文献标识码**: A **文章编号**: 1000−5781(2019)05−0598−12

doi: 10.13383/j.cnki.jse.2019.05.003

Analyzing topic evolution of online product forum based on topic model

Jiang Cuiqing, Lü Xiaozhong, Duan Rui
(School of Managament, Hefei University of Technology, Hefei 230009, China)

Abstract: Topic analysis of online product forums has an important significance to enterprise marketing and product improvement. According to the features of product forums, a topic evolutionary analysis model is constructed for online product forums based on latent Dirichlet allocation (LDA) model. The model is aimed at mining the topic evolution law of massive online product reviews from three levels: topic label, topic heat, and topic word heat. The experiment results show that the proposed method can mine the topic evolution law of online product forums. Further, consumer concerns of the same product in different forums have both commonness and differences. The change in the heat of consumer concerns are random, while the centres of consumer concerns are steady. Forums with rich reviews are easy to form an evolution relationship of topics.

Key words: topic evolution; online product forum; latent Dirichlet allocation model; heat topic

1 引 言

产品在线论坛已成为人们分享产品信息、传递产品价值和表达产品情感的重要场所, 论坛中蕴含着大量对企业市场决策、售后服务和产品改进等有价值的知识. 企业不仅可以通过这些知识掌握在线评论的主题演化从而把握消费者的诉求, 掌握消费者对产品的总体态度及其变化趋势, 同时还可以分析不同网站的消费者对相同产品关注点的区别, 从而为企业营销决策提供支持; 潜在购买者也可以从相关结果中把握消费者对产品的总体评价情况, 为其购买决策提供更全面的支持[1]. 例如某品牌汽车论坛出现的关于汽车缺

收稿日期: 2016−12−16; 修订日期: 2017−06−22.
基金项目: 国家自然科学基金重点资助项目(71731005); 国家自然科学基金资助项目(71571059); 教育部人文与社会科学项目研究计划资助项目(15YJA630010).

陷的主题以及某品牌手机论坛出现的关于手机电池爆炸的主题. 如果企业能够实时地掌握这类主题及其演化趋势, 制定应对政策, 可以避免消费者信任危机的出现, 提高企业效益. 但是这些知识是以时序文本的形式发布的, 要想全面地获取这些知识就必须对这类典型的非结构化海量信息进行挖掘, 而传统的人工标注和词频统计等方法已经不能有效地予以挖掘. 因此, 本文使用主题模型对蕴含相关知识的评论进行主题演化分析, 探究产品在线评论主题演化影响因素, 以期更好地了解主题演化的形成机理和演化规律.

近年来主题演化的应用分析多集中在新闻、科学文献、电子邮件和社会性在线论坛评论等领域. 这些领域论坛参与人员活跃, 且评论内容丰富度高, 易于进行主题演化分析. 但对产品在线论坛的主题演化分析较少. 究其原因主要是以往网上销售的产品多为快速消费品, 其价值低且受关注度不高, 其在线论坛的用户稀疏且评论较少, 从而导致难以对其在线论坛进行主题演化分析. 近年来, 随着人们消费水平的提高, 一些高价值的耐用品(例如手机、汽车等)已经成为受年轻人青睐的大众产品. 同时相较于社会性在线论坛的主题发散、变化快、粒度粗且突发主题多, 产品的在线论坛主题相对集中、规律性强, 且讨论焦点汇集度高, 易于发现细粒度的主题, 从而适合通过主题挖掘方法进行分析研究. 另外对于同一产品, 不同的在线论坛其主题演化关系也存在着差异. 因此, 本文研究基于主题模型的同一个产品的主题演化规律, 并对不同论坛的评论进行主题演化分析, 重点分析其中蕴含的同一产品消费者关注之差异性.

虽然在线评论内容的挖掘已经成为一个研究热点[2−4], 但大部分方法都需要专业人员的人工标注, 个人的差异必然会导致实验结果出现偏差[5]. 基于潜在狄利克雷分布(latent Dirichlet allocation, LDA)的主题模型可以解决人工标注的偏差问题. 潜在狄利克雷分布模型(LAD 模型)是一种概率生成模型[6], 通过模拟每个文本数据的生成过程, 可以抽取出文本信息中包含的重要主题. 虽然已有基于 LDA 主题模型的在线评论主题演化研究, 但其都只是针对主题的热度变化趋势和主题词语的热度变化趋势两方面进行分析, 忽略了主题标签的演化关系[7−9]和主题演化的影响因素的分析. 另一方面, 在线产品论坛中每一个评论下会有很多回复, 回复数多的评论讨论热度也就更高, 所包含的主题的热度也应该相对更高. 但是, 现有研究在对主题进行抽取时, 把每个评论看成是同等重要的, 忽略了评论的权重.

为此, 本文提出了一种改进的主题演化模型对产品在线论坛进行主题演化分析, 并以汽车在线论坛为实证对象, 结果显示该模型可以有效地挖掘出产品在线论坛中所蕴含的相关知识. 为了便于实验结果的分析, 本文将主题演化关系分为同一演化关系和普通演化关系. 针对现有研究的不足和产品在线评论的特点, 增加了主题过滤和评论权重, 结合夹角余弦相似度、在线评论丰富度和 LDA 模型, 构建了改进的主题演化分析模型. 使抽取的主题更接近于客观事实, 并从三个层次考察主题演化的规律, 一是主题标签在时间上的演化规律，该规律反映消费者对产品关注点的变迁; 二是具有同一演化关系主题的热度随时间的波动规律, 该规律反映消费者对产品某个方面的关注强度变化; 三是具有同一演化关系的主题内部词语热度变化规律, 该规律反映消费者对产品某一方面的关注侧重点的变化. 从这些结果中, 发现了一些对企业决策有用的演化规律和同一产品在不同在线论坛消费者关注点的差异.

2　主题模型的相关研究

产品在线论坛中的评论是非结构化的文本信息, 具有规模海量、特征空间多维等特点. 挖掘海量、高维度信息一般需要对其进行降维, 将高维词语空间映射到低维语义空间, 以便于对在线论坛评论中的文本信息进行理解与分析. 基于概率图模型的主题模型即是一种有效的文本处理方法[10].

2.1　主题模型研究

作为一种新的统计方法, 主题模型通过概率分布函数分析非结构化文本中的词语分布, 以发现蕴藏于其中的主题, 然后利用获得的主题进行后续数据挖掘与分析(如分类、聚类和演化关系分析等). LDA 模型是一个经典的主题模型[6]. 结合不同的应用场景, 学者们提出了一系列基于 LDA 的改进模型[11−13], 如结合文档以外的作者、题目和链接等信息对模型进行的改进[14]. 这些改进的主题模型不仅在文本信息分析中得到了

广泛应用(如微博的热门博主挖掘[15]和情绪挖掘[16]等), 还在社会网络、图像、源代码和生物信息等领域的分析中得到了应用.

2.2 基于主题模型的主题演化研究

作为一种改进的主题模型, 主题演化模型假设主题随时间变化. 该模型可以从时间序列文本中辨识主题并追踪主题的动态演化. 根据时间引入方式的不同, 目前针对主题演化模型的研究主要分为三种:

1) 后离散方式

这类模型源于 LDA 模型, 即先不考虑时间因素, 在整个语料库上运用 LDA 模型获取所有的主题, 再按照文档的时间信息将文档和主题离散到相应的时间点. 该类模型虽简单易行, 但其假设文本集内的所有文档是可交换的, 从而未能充分利用时间信息, 导致同样建模条件下后离散方式的主题演化模型的复杂度高于其它模型[17], 实验结果也并不能很好地突出一些短时热门主题.

2) 引入时间变量方式

常见的引入时间变量的主题演化模型为时间主题模型(topic over time model, TOTM)和连续时间动态主题模型(continuous time dynamic topic model, CDTM). 前者将时间、文档和词语三者联合起来作为模型参数, 采用 Beta 概率分布模型在给定时间范围内对主题热度的变化进行建模. 但其仅仅展示主题热度的变化, 忽略了主题内容的变化, 不能反映主题的演化关系[18]. 后者运用布朗运动模拟主题分布在时间上的演化, 但其对数据本身有一定的要求, 使得模型的泛化能力差[19].

3) 先离散方式

经典的先离散方式有动态主题模型(dynamic topic model, DTM)[12]、在线 LDA(online LDA, OLDA)模型[20]和序列 LDA(sequential LDA, SLDA)模型[21]. 这三种模型都是先将数据集按时间窗聚合, 各时间窗下的模型参数均依赖于前一个时间窗的状态, 进而进行模型的学习. 但是 DTM 存在粒度选择问题, OLDA 存在主题关联和主题探测问题, 三者都不能反映主题内容的演化变迁.

若采用先离散方式进行 LDA 建模抽取主题, 并选取合理的参数, 使用夹角余弦相似性来建立主题演化关系, 不仅可以得到主题热度的变化, 而且可以反应主题内容的演化变迁. 综合上述分析. 本文采用先离散方式, 构建基于 LDA 和夹角余弦相似性的主题演化模型.

2.3 主题演化模型的应用研究

目前, 国内外对于主题演化的应用分析多集中于新闻[19]、科技文献[22]、电子邮件[18]和英文电影评论[23]等领域. 但针对在线论坛主题演化分析的研究较少[7,9,24−26]. 例如, 文献[9]通过 LDA 模型对 350 篇来自社会性论坛的评论进行分析, 仅度分析了热门主题和冷门主题. 文献[6]通过 DTM 模型挖掘天涯论坛民生版块的主题链, 同样仅从内容和热度两方面研究了论坛主题的演化规律. 这些研究认为主题是单链演化, 没有考虑到单个主题可能朝着多个不同的方向演化, 从而使挖掘出的结果缺乏主题标签的演化分析; 也没有对这些主题出现的原因以及可能带来的结果进行分析. 同时, 这些研究都是针对社会性在线论坛评论的主题演化分析. 再者这些研究基本是使用单一数据集进行实证分析, 实验结果缺乏对比性. 虽然在线产品评论的挖掘分析已成为当前研究的热点[6,27−30], 但相关研究大多没有深入到评论的主题及其演化规律层面, 仅仅是挖掘其语法特征与关键字等. 产品评论主题及其演化规律有待进一步研究.

因此, 本文针对现有研究存在的不足, 研究主题标签的演化规律, 并分析主题演化与产品在线论坛丰富度之间的关联关系. 针对产品在线论坛评论中存在回复和主题可能出现重复的特征, 在主题演化分析模型中增加了评论权重和主题过滤. 同时, 对同一产品的多个在线论坛进行了对比研究, 分析不同论坛中主题及其演化的差异性.

3 主题演化分析框架

为了克服现有研究存在的不足, 本文构建了一个基于主题模型的产品在线评论主题演化分析框架. 如

图 1 所示, 该框架考虑了产品评论热度存在差异性的特点, 使用 LDA 模型对主题进行抽取, 利用夹角余弦相似性构建主题间的演化关系, 使用加权平均算法计算主题热度并结合评论丰富度算法分析产品在线论坛主题及其演化规律.

首先爬取产品在线论坛中评论数据, 选取主评论的文本内容、发布时间和回帖数量作为实验原始数据; 再对数据进行聚合、分词、去停用词和权重赋值等预处理; 然后在不同时间窗内使用 LDA 进行主题抽取; 接着使用夹角余弦相似性进行主题的过滤和主题演化关系的建立; 最后通过主题–词分布和评论–主题分布, 分别进行词语变化分析和主题热度变化分析.

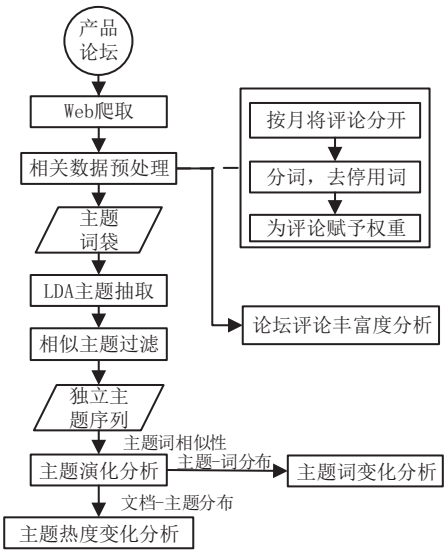


图 1　主题演化分析框架

Fig. 1　The framework of topic evolution analysis

3.1　产品在线论坛主题模型

采用 LDA 模型分析产品在线评论的基本思想是: 每个评论是一系列主题的概率分布, 记为 $\Pr(z)$; 每个主题是一系列词汇的概率分布, 记为 $\Pr(w|z)$. 因此, 评论中每个词的概率分布如下

$$\Pr(w_i) = \sum_{j=1}^{k} \Pr(w_i|z_i = j) \Pr(z_i = j). \tag{1}$$

LDA 概率图模型如图 2 所示.

图 2 中, $\boldsymbol{\alpha}, \beta$ 是 Dirichlet 分布的参数, $\boldsymbol{\alpha}$ 是一个 K 维向量, K 为主题数量; $\boldsymbol{\phi}_k$ 是一个 V 维向量, 表示主题 k 下的词概率分布, 即主题–词分布, V 为词汇表大小; $\boldsymbol{\theta}_d$ 是一个 K 维向量, 表示评论 d 下的主题概率分布, 即评论–主题分布; z 表示评论中分配在词上的主题, w 表示评论中的词, N 表示每个评论中词的数量, M 表示语料库中评论的数量. $\boldsymbol{\phi}_k$ 和 $\boldsymbol{\theta}_d$ 服从狄利克雷分布, z 和 w 服从多项分布.

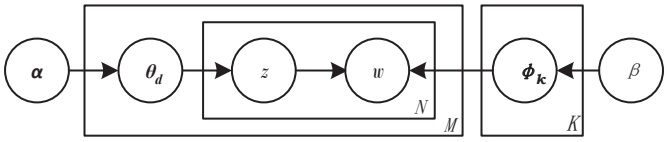


图 2　LDA 概率图模型

Fig. 2　Probabilistic graphical model of LDA

根据每个评论的回复数, 在组织样本时给每个评论赋予权重 $\omega_{d,t}$. 基于 LDA 模型, 一个评论的生成过程如下.

1) 对于某个主题 k, 抽取该主题在词语上的分布 $\boldsymbol{\phi}_k$, 该分布服从 Dirichlet(β);

2) 对于评论 d, 抽取该评论在主题上的分布 $\boldsymbol{\theta}_d$, 该分布服从 Dirichlet($\boldsymbol{\alpha}$);

3) 对评论 d 中的每个词 w_i,

(a) 为其抽取一个主题分配 z_i, 该主题服从 Mult($\boldsymbol{\theta_d}$), $z_i = 1, 2, \ldots, K$,

(b) 抽取这个词 $w_i \sim$, 该词服从 Mult($\boldsymbol{\phi}_z$), $w_i = 1, 2, \ldots, V$;

4) 对于每个评论 d, 根据权重 $\omega_{d,t}$, 重复 $\omega_{d,t}$ 次步骤 2 和步骤 3.

按上述步骤重复抽取生成整个评论, 直到后验参数 $\boldsymbol{\phi}$ 与 $\boldsymbol{\theta}$ 收敛. 本文采用 Gibbs 抽样来学习模型后验参数. 采用评论集的困惑度(记为 P(D))选取最佳主题数 K[31]. 计算公式如下

$$\mathrm{P}(D) = \exp\left(-\ln \Pr(d_i)\left(\sum_{i=1}^{M} N_i\right)^{-1}\right), \tag{2}$$

其中 $\Pr(d_i)$表示模型生成评论 d_i 的概率, N_i 表示评论 d_i 的长度, M 为评论的数量, D 表示评论集. 困惑度越小, 模型学习的结果越能代表数据集. 通过选取不同的 K 值来计算和对比困惑度, 并选择困惑度最小时的 K 值.

3.2 主题演化关系建立

已有研究通过主题间的 KL(Kullback-Leibler)距离和 JS(Jensen-Shannon)距离以及夹角余弦来确立演化关系[8]. KL 距离和 JS 距离在计算时需要将两个主题间的主题词概率相除, 但在本文的实验结果中主题词的概率可以为零, 从而无法得到计算结果. 相比较而言, 夹角余弦更为简洁有效, 完全能满足本文实验要求. 因此, 根据简单有效的原则, 本文选择了主题间的夹角余弦来确立演化关系.

假设两个相邻时间窗 t_i 和 t_{i+1} 的产品评论集经 LDA 建模得到主题 $Z_r^{t_i}$ 和 $Z_s^{t_{i+1}}$, $\boldsymbol{q}$ 是 $Z_r^{t_i}$ 的主题词的概率分布, $\boldsymbol{p}$ 是 $Z_s^{t_{i+1}}$ 的主题词的概率分布, $\boldsymbol{q}$ 和 $\boldsymbol{p}$ 都是 n 维向量. 则主题 $Z_r^{t_i}$ 和 $Z_s^{t_{i+1}}$ 的相似性为

$$\cos(Z_r^{t_i}, Z_s^{t_{i+1}}) = \sum_{i=1}^{n} p_i q_i \left(\sqrt{\sum_{i=1}^{n} {p_i}^2}\sqrt{\sum_{i=1}^{n} {q_i}^2}\right)^{-1}. \tag{3}$$

两个主题的夹角余弦越接近于 1, 其相似度越高, 越有可能存在演化关系. 本文设定一个相似度阈值, 来判定不同时间窗上的两个主题是否存在演化关系. 也就是说, 如果两个主题的相似度大于阈值, 则判定这两个主题具有演化关系.

本文定义具有演化关系的两个主题的主题标签相同时, 两者为同一演化关系; 否则为普通演化关系.

3.3 主题演化分析方法

1) 主题标签演化趋势

(a) 主题标签演化表示. 使用圆表示主题, 箭头表示演化关系, 箭头的粗细表示演化关系的强弱, 圆内的词语表示主题标签, 采用人工标记法从主题词中为每个主题分配主题标签.

(b) 主题过滤. 通过式(3)计算主题间的相似性, 在同一时间窗内主题间相似性达到一定阈值时, 过滤掉和下一时间窗内主题演化关系弱的主题, 得到独立的主题集合序列.

(c) 不同时间窗演化关系建立. 当相邻时间窗内主题间相似性达到一定阈值, 认定为上一时间窗内的主题演化成下一时间窗内的主题. 通过以上步骤绘制得到主题演化图. 从而通过主题标签演化分析发现消费者关注点的变化.

2) 主题热度变化趋势

根据后验参数 $\boldsymbol{\theta}_{d,t}$ 的分布, 可以得到每个评论中包含的主题分布. 为从整体上把握各个主题热度在时间轴上的变化, 需构建一个主题热度的计算方法定量描述主题的演化趋势.

对于主题热度的计算, Griffiths 等[22]和 Hall 等[32]在后离散方式(基于 LDA 的模型)中分别提出以平均热度和主题在文档集中出现的次数作为主题热度两种方法.

从 LDA 模型实验结果 $\boldsymbol{\theta}_{d,t}$ 来看, 主题热度基本不为 0, 并且主题在文档中出现的次数是固定的, 所以按

照 Hall 等[32]通过统计主题出现次数的热度计算方法没有意义. 因此, 本文将文献[22]中的平均热度计算方法应用至结果分析中. 由于在组织样本时给每个评论赋予权重 $\omega_{d,t}$, 因此本文对每个时间窗内的主题热度计算方法作如下改进:

记 $\theta_{d,k,t}$ 为评论 d 在时刻 t 下的主题 k 的热度, $\boldsymbol{\theta}_{k,t}$ 的平均值(记为$\overline{\boldsymbol{\theta}}_{k,t}$)表示主题 k 在时间 t 下的热度, $\overline{\boldsymbol{\theta}}_{k,t}$的波动反映了消费者对主题关注度的变化, $\omega_{d,t}$ 为评论 d 在时间 t 的权重. 在时间序列 $t=1,2,\ldots,T$ 下, $\overline{\boldsymbol{\theta}}_{k,t}$ 的具体计算过程如下:

步骤 1 统计时间点 t 下的评论个数 M_t;

步骤 2 在 t 时刻, 第 k 个主题的热度值为

$$\overline{\boldsymbol{\theta}}_{k,t}=\frac{1}{M_t}\sum_{d}\omega_{d,t}\theta_{d,k,t}. \tag{4}$$

本文统计具有同一演化关系主题的热度, 构建主题热度变化图.

评论丰富度反映每个时间窗内评论内容的丰富程度, 如每个时间窗有多少个评论, 这些评论中有多少个词汇. 假设评论丰富度高的产品在线论坛形成的主题演化关系越多. 为了验证这个假设, 本文采用式(5)从多个角度对评论丰富度进行度量.

$$y_{i,t}=\lg(nr_{i,t})+\lg(nw_{i,t})+\lg(nv_{i,t}), \tag{5}$$

其中 $y_{i,t}$ 为网站 i 在 t 时间窗内的评论丰富度, $nr_{i,t}$ 为评论数量, $w_{i,t}$ 为评论词典数量, $nv_{i,t}$ 为评论词汇数量, 对各变量进行对数处理, 以增加结果的平稳性. 将评论丰富度构建成一个绝对值, 方便分析每个论坛自身的评论丰富度变化与主题演化形成的关系.

3) 主题词语变化趋势

通常情况下, 产品在线论坛会讨论多个不同的主题, 即便是针对同一主题, 所包含的主题词语的概率分布随着时间的推移也会发生变化. 因此, 具有同一演化关系主题的词语分布也会发生变化. 词语分布随时间的变化反映了每个主题的讨论中心的变化. 由后验参数 $\boldsymbol{\phi}_{t,k}$ 的结果可以最终确定在每个时间窗中每个主题的词语分布.

通过统计具有同一演化关系的主题间的主题–词分布, 构建主题词语变化趋势图进行分析.

4　主题演化实证分析

4.1　实证对象和数据源选择

本文选择福克斯论坛作为实证研究对象. 根据中国汽车工业协会的统计, 2014 年福克斯以 39.18 万辆的销量成为销售冠军, 故而购买福克斯汽车的消费者就会相对较多, 使得福克斯论坛活跃度较高, 因此福克斯论坛是具有代表性产品在线论坛.

本文选择汽车之家网、太平洋汽车网、易车网和新浪汽车网上的四个福克斯论坛进行实验研究, 验证本文提出的主题演化分析模型对产品在线评论分析的泛化能力, 并发现了一些影响企业和潜在消费者决策的主题演化规律.

4.2　数据统计与清理

原始数据集为上述四个论坛的评论, 时间从 2013–01～2015–09. 数据预处理过程如下.

1) 将评论按月为时间窗聚合;

2) 对数据进行清理, 包括:

(a) 去除评论中的链接信息,

(b) 用 stanford 分词包对评论进行分词处理,

(c) 使用停用词词典去掉标点符号和停用词, 最后将每个评论转化为词袋;

3) 评论赋权. 通常情况下, 产品在线论坛的每个评论的热度是不一样的, 这些评论包含的主题热度也应该不同, 一般认为回复数多的评论热度高. 因此需要预先赋予评论权重, 本文使用评论回复数作为权重.

预处理过程为 LDA 主题模型的学习提供格式化数据. 论坛数据的描述性统计如表 1 所示. 每个论坛单月评论数据都非常的丰富, 适合用于主题演化的挖掘分析. 汽车之家论坛评论数量最多, 新浪汽车论坛评论最少.

表 1 数据集的统计结果

Table 1 The statistical result of data set

名称	平均数	均方差	最大值	最小值
汽车之家评论数量	6 030	12 147	68 811	269
太平洋汽车网评论数量	4 492	3 402	12 173	881
易车网评论数量	1 880	2 668	10 878	181
新浪汽车网评论数量	202	117	554	53
汽车之家网词典词数	7 115	4 641	26 115	1 768
太平洋汽车网词典词数	6 899	3 973	13 381	1 639
易车网词典词数	2 592	2 484	10 251	443
新浪汽车网词典词数	1 766	712	3 250	408
汽车之家网语料库	36 081 882	59 839 079	340 983 555	1 562 028
太平洋汽车网语料库	31 692 881	29 119 265	89 518 890	1 342 341
易车网语料库	6 445 283	11 255 490	52 535 936	97 903
新浪汽车网语料库	1 814 176	1 307 789	5 279 625	83 028

4.3 主题数量的确定

由于每天的评论数量较少, 最终选择以月为主题演化的时间窗. 根据式(2)的困惑度计算生成得主题数为 20 个.

4.4 主题演化分析结果

1) 主题内容变迁分析

利用式(3)计算主题间相似性, 确定演化关系. 本文将演化关系超过五个月的长期主题从主题集中抽取出来进行分析, 并通过 3.4 节提出的方法构建四个论坛的主题演化图(示例见图 3).

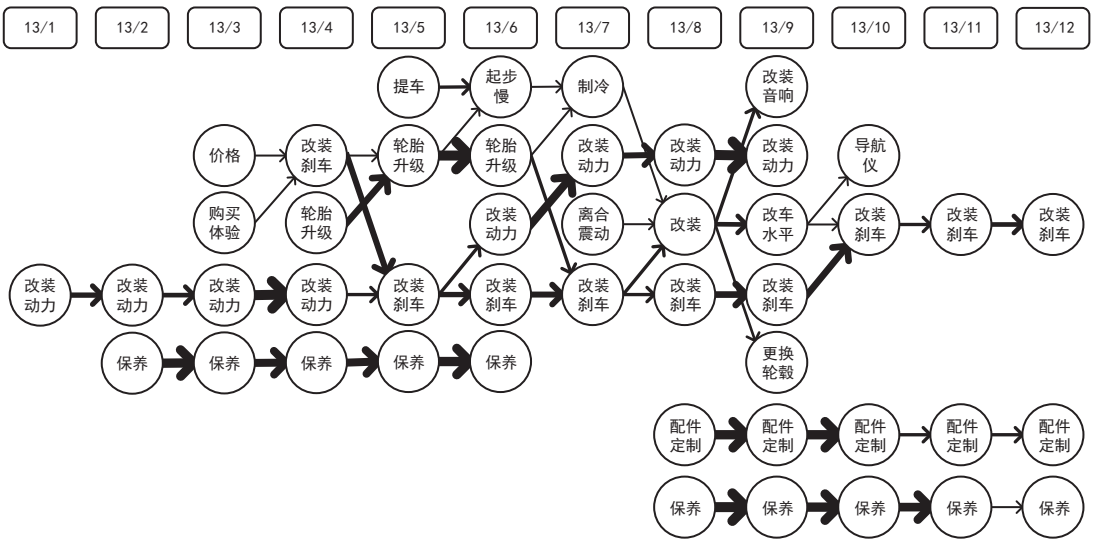


图 3 汽车之家福克斯论坛 2013 年主题演化图

Fig. 3 Focus BBS 2013 theme evolution map

从主题演化图中发现以下规律:

(a) 消费者的关注点主要包括“福克斯汽车相关信息”, “车友线下活动”和“论坛线上活动”三大类. 其中“福克斯汽车相关信息”是消费者的主要关注点.

(b) 同一产品的消费者虽然有共同的关注点, 但是不同网站的消费者群体对产品有着不同的关注偏好. 例如消费者长期共同关注福克斯汽车的“口碑”、“保养”和“故障”等主题, 只有汽车之家网站消费者会长期关注“改装”等技术性较强的主题演化关系, 说明汽车之家福克斯论坛的用户知识专业性较强. 而易车网福克斯论坛消费者主要关注如“提车记”、“车友活动”和“ 论坛精华帖”等主题的车友线下活动和线上活动, 说明易车福克斯论坛的用户互动性较强.

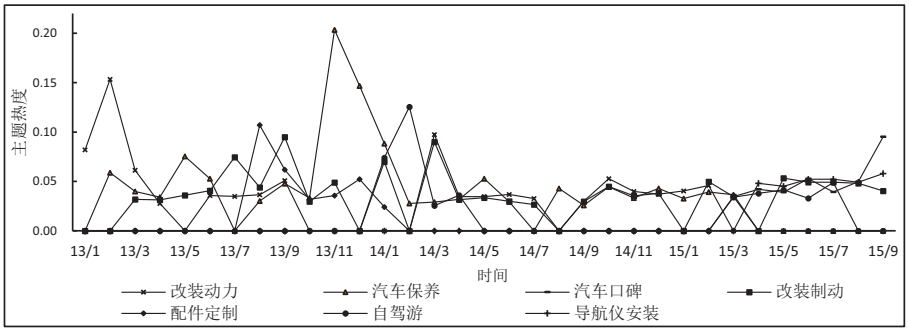
(c) 当某个主题出现时, 随着时间的推移, 消费者的关注重点会发生变化, 该主题就有可能演化成其它主题, 或者分裂成几个主题. 例如汽车之家网福克斯论坛 2013 年 8 月“改装”主题演化为 2013 年 9 月“ 改装音响”、“改车水平”和“更换轮毂”主题. 同样的, 当主题间讨论的内容有实质联系时, 这些主题会演化为一个主题. 例如汽车之家网福克斯论坛 2013 年 7 月“制冷”、“离合震动”和“改装刹车”主题演化为 2013 年 8 月的“改装”主题, 说明此主题主要讲的是针对汽车制冷, 离合和制动的改装.

(d) 消费者购买汽车以后将长时间关注其维护和升级, 它们之间的演化关系也更强烈,对此产品的口碑、故障以及购买产品对生活质量改善的关注时间较短. 例如, 有关“改装”主题大家讨论的时间最长, 其次是有关“保养”的主题, 而“口碑”、“故障”、“自驾游”等主题的关注时间持续相对较短.

消费者对产品关注的偏好、演化变迁和持续性都是营销人员决策所要关注的基础信息. 主题内容演化分析可以快速而客观地为其制定营销方案提供信息支持. 同时潜在消费者可以清楚地观察到此产品消费者对其总体舆论内容的变化情况, 从而为其购买决策提供支持.

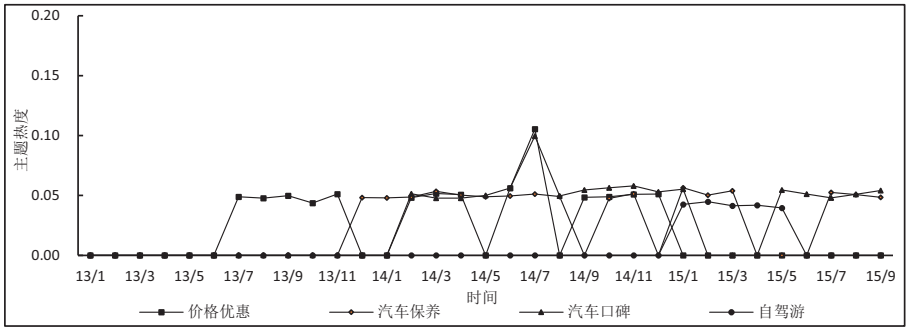
2) 主题热度分析

对具有同一演化关系的主题利用式(4)进行主题热度计算, 得出如图 4 所示的主题热度变化图.



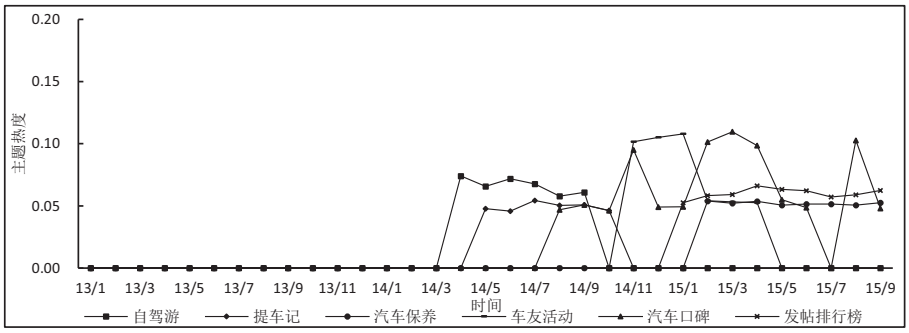
(a)　汽车之家网福克斯论坛主题热度

(a) The home of the automobile network fox BBS theme heat



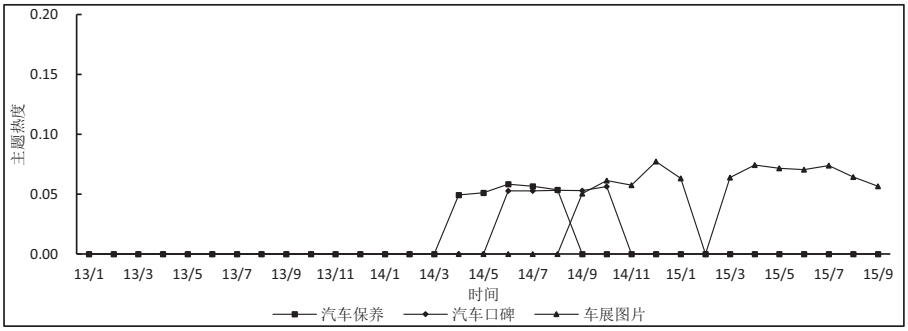
(b)　太平洋汽车网福克斯论坛主题热度

(b) Pacific automotive network fox BBS theme heat



(c) 易车网福克斯论坛主题热度

(c) Easy car network fox BBS theme heat



(d) 新浪汽车网福克斯论坛主题热度

(d) Sina car network fox BBS theme heat

图 4 福克斯论坛主题热度变化

Fig. 4 The topic heat's change of Focus forum

图 4 中, 并没有发现这些长期出现的主题的热度变化在时间轴上的变化规律, 说明消费者对产品关注点的热度变化是很难预测的. 产品在线论坛并没有像社会在线论坛那样, 随着节日或者特殊日期的到来, 其中的主题热度会发生急剧变化. 另外, 通过主题热度变化分析可以观察消费者对产品关注点的热度变化, 让营销人员了解到在消费者长期关注的主题中哪些主题比较重要, 从而在制定营销策略时有所侧重.

为探究是否评论的内容和数量越丰富, 越容易形成主题演化关系. 利用式(5)计算得出各网站福克斯论坛每个月的评论丰富度, 如图 5 所示.

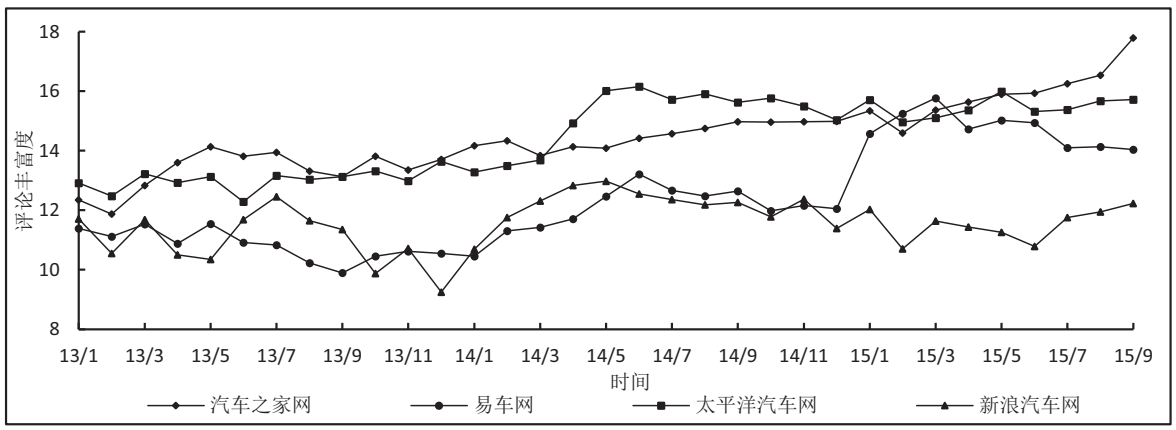


图 5 福克斯论坛评论丰富度

Fig. 5 Review richness of Focus forum

对比图 4 和图 5 发现以下两个规律:

(a) 像“发帖排行”这样讨论个人论坛荣誉的主题出现时, 可以刺激消费者更多地发表产品评论, 从而使产品论坛评论丰富度有一个短期地提升, 但随着时间的推移这种提升作用不断减弱. 例如易车网福克斯论坛在 2014 年 12 月以后出现了一个新的主题演化关系, 内容主要是讨论发帖排行榜和精华帖排行榜. 当“发帖排行榜”主题出现并且保持一定的热度时, 同一时期的易车网福克斯论坛评论丰富度有了一个瞬间的提高, 但随着时间的推移论坛的评论丰富度开始不断下降. 所以要想使论坛吸引更多地用户, 并使其积极参与对产品的评论, 还需要论坛管理者积极管理, 建立长效的机制.

(b) 产品论坛评论丰富度越高, 主题演化关系形成的就越多, 其热度也越高. 例如从图 4 和图 5 中发现评论丰富度高的汽车之家福克斯论坛和太平洋汽车福克斯论坛形成的主题演化关系较多, 且热度也较高.

3) 主题词语变化分析

从汽车之家福克斯论坛中选取“保养”主题进行主题词语热度变化分析. 如图 6 所示, 并发现以下规律:

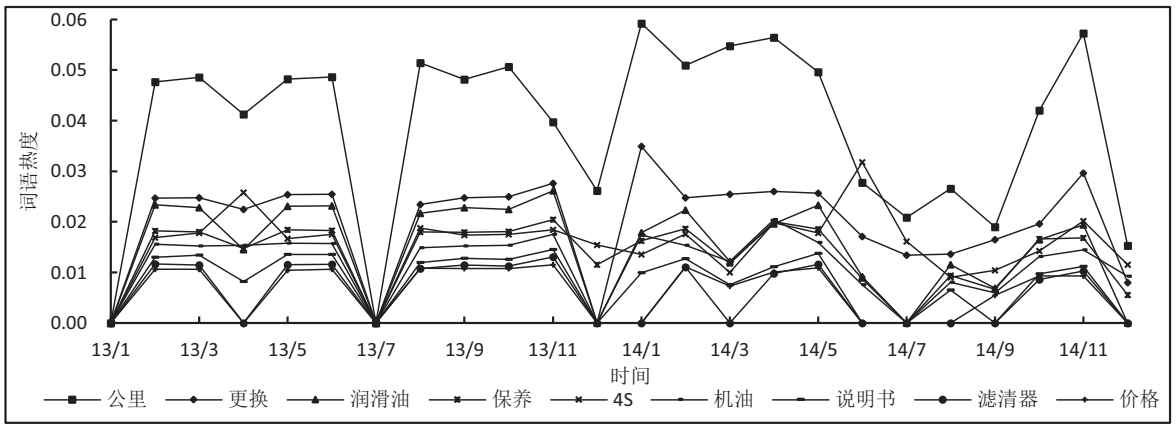


图 6　“保养”主题下词语热度变化

Fig. 6　Word heat change of ‘upkeep’ topic

(a) 同一演化主题的单个词语热度在时间上波动并无规律.

(b) 主题词的热度顺序在时间上基本不会发生变化, 热度成比例波动. 例如图 6 中 2014 年 6 月主题词的热度比例有了一个很小的波动, 其余时间点主题词的热度顺序基本是不会改变的.

以往学者研究社会性在线论坛中主题词的热度顺序在时间上是会发生改变的. 因为在社会在线论坛中讨论的主题一般是社会事件, 随着时间的推移, 一些新证据的曝光使得网民对事件的认识不断改变, 从而关注的事件中心发生改变. 即使对同一个主题, 主题词间的热度顺序也会发生改变. 然而产品在线论坛的主题一般比较客观, 因此对于同一个主题内容消费者关注的侧重点不会发生变化. 所以, 这些主题的词语热度排序随着时间的推移基本不会发生变化, 即产品在线论坛中具有同一演化关系的主题中心在时间轴上一般不会发生变化.

5　结束语

针对产品在线论坛评论热度存在差异性的特点, 构建了基于 LDA 模型的主题演化分析框架, 从主题标签、主题热度和主题词热度三个层次挖掘产品在线评论主题演化规律. 实验结果表明, 该方法能够客观地抽取主题的演化规律; 同时发现了同一个产品的不同在线论坛间消费者关注点演化规律的共性与区别, 例如福克斯论坛消费者共同关注“口碑”、“保养”和“故障”三大类主题, 汽车之家福克斯论坛侧重关注技术性主题, 易车网福克斯论坛侧重于互动性主题; 另外, 发现了消费者关注点热度变化的时间随机性, 以及消费者关注点演化关系形成与在线评论丰富度的关联; 对于同一个消费者关注点, 其关注中心是基本不会改变的.

对于网络营销人员来说, 尤其当一个新产品出现时, 该方法能够快速地为其呈现消费者关注点的演化和不同网站间消费者关注点的差异, 从而选择相应的营销侧重点; 研究结论揭示了产品网络论坛主题演化关系形成的一些影响因素, 从而为相关企业正确引导消费者舆论走向提供帮助; 同时为产品论坛的编辑提供了新的途径, 如利用本文分析模型向产品网络论坛用户展示消费者关注点演化信息等. 但本文研究侧重分析具有长期演化关系的主题, 短期主题的分析有待进一步研究.

参考文献:

[1] Archak N, Ghose A, Ipeirotis P G. Deriving the pricing power of product features by mining consumer reviews. Management Science, 2011, 57(8): 1485–1509.

[2] Gopinath S, Thomas J S, Krishnamurthi L. Investigating the relationship between the content of online word of mouth, advertising, and brand performance. Marketing Science, 2014, 33(2): 241–258.

[3] Chau M, Xu J. Business intelligence in blogs: Understanding consumer interactions and communities. MIS Quarterly, 2012, 36(4): 1189–1216.

[4] 刘 洋, 廖貅武, 刘 莹. 在线评论对应用软件及平台定价策略的影响. 系统工程学报, 2014, 29(4): 560–570.
Liu Y, Liao X W, Liu Y. The impact of online review on software and platform's pricing strategies. Journal of Systems Engineering, 2014, 29(4): 560–570. (in Chinese)

[5] Ghose A, Ipeirotis P G, Li B. Designing ranking systems for hotels on travel search engines by mining user-generated and crowd-sourced content. Marketing Science, 2012, 31(3): 493–520.

[6] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation. Journal of Machine Learning Research, 2003, 3(4/5): 993 – 1022.

[7] 曹丽娜, 唐锡晋. 基于主题模型的BBS话题演化趋势分析. 管理科学学报, 2014, 17(11): 109–121.
Cao L N, Tang X J. Trends of BBS topics based on dynamic topic model. Journal of Management Sciences in China, 2014, 17 (11): 109–121. (in Chinese)

[8] 唐晓波, 王洪艳. 基于潜在狄利克雷分配模型的微博主题演化分析. 情报学报, 2013, 32(3): 281–287.
Tang X B, Wang H Y. Analysis of microblog topic evolution based on latent Dirichlet allocation model. Journal of the China Society for Scientific and Technical Information, 2013, 32 (3): 281–287. (in Chinese)

[9] 石大文, 张 晖. 基于 LDA 模型的 BBS 话题演化. 工业控制计算机, 2012, 25(5): 82–84.
Shi D W, Zhang H. LDA model-based BBS topic evolution. Industrial Control Computer, 2012, 25(5): 82–84. (in Chinese)

[10] Blei D M. Probabilistic topic models. Communications of the ACM, 2012, 55(4): 77–84.

[11] Wallach H M. Topic modeling: Beyond bag-of-words // Proceedings of the 23rd International Conference on Machine Learning. ACM, 2006: 977–984.

[12] Blei D M, Lafferty J D. Dynamic topic models // Proceedings of the 23rd International Conference on Machine Learning. ACM, 2006: 113–120.

[13] Blei D M, Lafferty J D. A correlated topic model of science. The Annals of Applied Statistics, 2007, 1(1): 17–35.

[14] 徐 戈, 王厚峰. 自然语言处理中主题模型的发展. 计算机学报, 2011, 34(8): 1423–1436.
Xu G, Wang H F. The development of topic models in natural language processing. Chinese Journal of Computers, 2011, 34(8): 1423–1436. (in Chinese)

[15] Weng J, Lim E P, Jiang J, et al. Twitterrank: Finding topic-sensitive influential twitterers // Proceedings of the Third ACM International Conference on Web Search and Data Mining. ACM, 2010: 261–270.

[16] Mei Q, Ling X, Wondra M, et al. Topic sentiment mixture: Modeling facets and opinions in weblogs // Proceedings of the 16th International Conference on World Wide Web. ACM, 2007: 171–180.

[17] Iwata T, Yamada T, Sakurai Y, et al. Online multiscale dynamic topic models // Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2010: 663–672.

[18] Wang X, McCallum A. Topics over time: A non-Markov continuous-time model of topical trends // Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2006: 424–433.

[19] Wang C, Blei D, Heckerman D. Continuous time dynamic topic models. https://arxiv.org/ftp/arxiv/papers/1206/1206.32-98.pdf, 2015.

[20] AlSumait L, Barbará D, Domeniconi C. On-line LDA: Adaptive topic models for mining text streams with applications to topic detection and tracking // Eighth IEEE International Conference on Data Mining. IEEE, 2008: 3–12.

[21] Du L, Buntine W, Jin H, et al. Sequential latent Dirichlet allocation. Knowledge and Information Systems, 2012, 31(3): 475–503.

[22] Meng C, Zhang M, Guo W. Evolution of movie topics over time. http://cs229.stanford.edu/proj2012/MengZhangGuoEvolutionofMovieTopicsOverTime.pdf, 2012.

[23] Shi L, Sun B, Kong L, et al. Web forum sentiment analysis based on topics // Ninth IEEE International Conference on Computer and Information Technology. IEEE, 2009, 2: 148–153.

[24] You L, Du Y, Ge J, et al. BBS based hot topic retrieval using back-propagation neural network // International Conference on Natural Language Processing. Berlin, Heidelberg: Springer, 2004: 139–148.

[25] 胡艳丽, 白　亮, 张维明. 网络舆情中一种基于 OLDA 的在线话题演化方法. 国防科技大学学报, 2012, 34(1): 150–154.
Hu Y L, Bai L, Zhang W M. OLDA-based method for online topic evolution in network public opinion analysis. Journal of National University of Defense Technology, 2012, 34(1): 150–154. (in Chinese)

[26] Hu N, Koh N S, Reddy S K. Ratings lead you to the product, reviews help you clinch it: The mediating role of online review sentiments on product sales. Decision Support Systems, 2014, 57: 42–53.

[27] Ghose A, Ipeirotis P. The EconoMining project at NYU: Studying the economic value of user-generated content on the internet. Journal of Revenue and Pricing Management, 2009, 8(2/3): 241–246.

[28] Yu X, Liu Y, Huang X, et al. Mining online reviews for predicting sales performance: A case study in the movie domain. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(4): 720–734.

[29] Yan Q, Wu S, Wang L, et al. E-WOM from e-commerce websites and social media: Which will consumers adopt. Electronic Commerce Research and Applications, 2016, 17: 62–73.

[30] Cao J, Xia T, Li J, et al. A density-based method for adaptive LDA model selection. Neurocomputing, 2009, 72(7): 1775–1781.

[31] Griffiths T L, Steyvers M. Finding scientific topics // Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(1): 5228 – 5235.

[32] Hall D, Jurafsky D, Manning C D. Studying the history of ideas using topic models // Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2008: 363–371.

作者简介:

蒋翠清(1965—), 男, 安徽合肥人, 教授, 博士生导师, 研究方向: 社交媒体 UGC 挖掘, 大数据环境下信用评价, Email: jiangcuiq@163.com;

吕孝忠(1988—), 男, 安徽六安人, 博士生, 研究方向: 社交媒体文本挖掘, Email:xiaozhong.lv@qq.com;

段　锐(1991—), 男, 山东德州人, 博士生, 研究方向:社交媒体协同推荐, Email: duanrui.haoren@qq.com.