【工业设计】

面向产品设计迭代的缺陷信息挖掘方法研究

梁若愚，张凌浩
（江南大学，无锡 214122）

摘要：目的 基于主题模型与关联算法，研究中文环境下服务于产品设计迭代的缺陷挖掘方法。**方法** 以网络产品社区作为研究对象，利用数据挖掘技术抓取社区用户的贡献内容，归纳面向产品缺陷分析的语法结构与候选词集，使用主题模型分析语料库中所包含的主题（产品属性）数量以及各个主题下的关键词分布，利用关联算法与紧凑规则找出每个主题下的强关联规则，解析后获得产品缺陷信息。**结果** 通过对小米 4 型智能手机用户贡献内容的实证分析，识别出了该产品用户声量最高的十四个关键问题。**结论** 两组对比实验的结果表明，所提出的方法能较好地识别、定位用户贡献内容中所包含的产品缺陷/不足信息，拥有较高的准确率与召回率。该方法能够为企业开展产品设计创新活动提供必要的支持。

关键词：产品设计创新；用户贡献内容；产品缺陷挖掘；主题模型；关联算法
中图分类号：TB472 **文献标识码：**A **文章编号：**1001-3563(2019)24-0150-08
DOI：10.19554/j.cnki.1001-3563.2019.24.024

Defect Information Mining Method for Product Design and Improvement

LIANG Ruo-yu, ZHANG Ling-hao
(Jiangnan University, Wuxi 214122, China)

ABSTRACT: This work aims to explore the defect mining method for product design and improvement in the Chinese context based on topic model and Apriori analysis. Firstly, the online product community was selected as the research object and the user-generated contents were collected with the assist of data mining technique. Then, a set of trivial lexical-Part-of-Speech patterns were proposed to prepare candidate corpus, and a topic model was adopted to find the optimal number of topics and get the words distributions in each topic. Finally, combined Apriori analysis and compactness rules, the expected strong rules in each topic were found out. Product defect information was obtained after analysis. Through the empirical analysis of the content contributed by Mi 4 smartphone users, the 14 key issues with the highest user volume of the product were identified. The results of the comparison experiment reveal that the proposed method can extract the product problems effectively. It has high accuracy and recall rate, can provide support for enterprises to initiate innovation activities.

KEY WORDS: product design innovation; user-generated content; product problem mining; topic model; Apriori analysis

互联网技术的迅速发展使网络社交逐渐成为人们日常生活中不可缺少的部分，出于信息交流、功能比较、兴趣爱好等方面的原因，越来越多产品用户开始在网络社区中通过发表主题帖、评论等方式，分享他们对于产品的使用经验与感受[1]。用户创新理论的提出者 E Hippel 认为，用户所发表的对于产品的见解（如期望、抱怨、需求、评价等），往往反映了他们最真实的感受[2]。这些观点与看法常涉及产品的多个

收稿日期：2019-09-21
基金项目：教育部人文社会科学研究青年基金项目（19YJCZH098）；江苏省社会科学基金重点项目（17YSA001）；江苏省“六大人才高峰”项目（JY-002）；江苏省第五期“333 工程”项目（2016III-2517）
作者简介：梁若愚（1988—），男，山东人，博士，江南大学讲师，主要从事设计方法与网络社区方面的研究。
通信作者：张凌浩（1974—），男，江苏人，博士，江南大学教授，主要从事中国产品设计战略与方法、用户体验设计方面的研究。

方面，有效识别、解析此类内容，可以帮助企业及时发现产品的缺陷与不足，支持决策者提出具有针对性的研发、改进策略，从而提高用户满意度，巩固企业的市场地位。

1　互联网开放创新社区的兴起

网络用户所贡献的内容对于推动企业开展产品设计研发、改进迭代等工作具有重要的意义。在这样一种背景下，众多知名消费品生产企业纷纷建立网络产品社区，鼓励用户通过在线方式贡献内容、参与创新，例如海尔构建了虚拟粉丝俱乐部，吸引消费者分享使用体验、发表诉求等以支持产品创新；戴尔通过在线社区收集用户痛点及市场需求点，以促进产品迭代；小米利用在线产品社区获取用户偏好、需求、对现有产品不满等方面的信息，并邀请用户直接参与到产品的开发过程中[3-4]。除了消费品制造商外，工业品制造商也开始尝试通过网络社区获取外部知识，如丰田建立了网络车友会，鼓励用户在社区中交流有关使用体验、产品问题、未来需求等方面的内容，以获取支持产品开发的有效信息；3M 利用线上与线下相结合的方式，邀请领先用户贡献专业知识以促进工业装备的研发；Seeed 通过在线开发者社区，鼓励用户参与到开源开发板 Arduino 的设计活动中，消费者为 Seeed 提供了大量有价值的创意，并帮助该公司发掘潜在的市场需求[5-6]。

2　网络社区用户贡献内容挖掘

近年来，针对网络社区用户贡献内容当中的特定信息识别与提取方法研究，逐渐成为了工程与信息领域的关注热点。Gupta 等人[7]认为，每类问题的描述性语段都会包含若干种候选的名词短语，他们训练了一种极大熵分类器对 Twitter 平台中关于 AT&T 公司的产品与服务问题的英文信息进行提取，效果良好。Tutubalina 与 Ivanov[8]利用问题语法结构以及从属关系，对特定领域的产品问题相关语句进行识别，他们以电子类与汽车类社区的在线用户评论作为研究对象，进行了实例分析，证明了所提出方法的有效性。Hagege 与 Brun[9]定义了一系列通用语法格式来识别英文信息中的产品改进/创新建议。与之类似，Ramanand 等人[10]也尝试通过归纳用户常用的问题表述语法格式，如“动词+助动词+积极评价词汇”，来获取用户建议以及他们的购买意愿信息，这种方法在识别特定类别信息时表现出了较高的效率。Goldberg 等人[11]建立了一种支持向量机分类器来发掘用户意愿，它可以利用部分手动输入的二进制语法模板（例如“我希望”），来辅助实现对用户意愿的挖掘。然而，上述研究所提出的信息识别方法，大多依赖精准定义的语法结构，在实践过程中，需要耗费较多时间来分析语段格式，效率和精度相对有限。另外，目前围绕产品相关信息挖掘方法的研究，大多是西方学者在英文语言环境下开展的，针对中文的同类研究相对匮乏。由于中英文在语法结构方面存在巨大差异，计算机往往无法直接识别中文文本[12]，因此针对中文语言环境的信息挖掘较英文环境更为复杂，需要开发更为适宜的方法。

围绕这些问题，基于主题模型（LDA）与关联算法（Apriori）提出一种面向产品缺陷/不足的中文文本挖掘方法。基本思想是：利用 LDA 模型提取用户评论数据库（用户在产品社区中的发帖/评论内容）中所包含的与产品属性相关的所有主题，然后通过 Apriori 算法将与这些属性频繁搭配的词汇/短语（频繁项）挖掘出来，经过分析频繁项中的情感/评价词汇与产品属性词汇之间的关系，找出同产品缺陷相关的内容，定位产品缺陷信息。

3　互联网环境下面向产品设计及改进的缺陷信息挖掘方法研究

3.1　主题模型构建

3.1.1　数据准备

数据准备阶段通过对产品社区用户发表的主题帖/评论进行采集及预处理，构建用户评论数据库。

3.1.2　候选词集挖掘

为了准确识别用户评论数据库中的有效内容，对中文用户语言结构进行了分析，并归纳出一部分用户描述产品问题/缺陷时常用的语法结构，产品缺陷挖掘格式见表 1。

表 1　产品缺陷挖掘格式

Tab.1　Extracted patterns to product problems

格式	示例
PA + *NEGv* or *NEGv* + *PA*	“软件（*PA*）+ 无法下载（*NEGv*）”
PA + *NEGa* or *NEGa* + *PA*	“手机信号（*PA*）+ 不佳（*NEGa*）”
PA + *NEGv* + *n*	“数据线（*PA*）+ 不能连接（*NEGv*）+ 电脑（*n*）”
v + *n* + *NEGv* + *n*/*PA*	“接听（*v*）+ 电话（*n*）+ 没有（*NEGv*）+ 声音（*n*/*PA*）”
an	“死机（*an*）”

注：*n* 代表名词（noun）；*v* 代表动词（verb）；*a* 代表形容词（adjective）

其中：*PA* 是产品属性的简写，多数情况下由名词性词汇或短语构成；*NEGv* 与 *NEGa* 分别表示了含有消极情绪倾向的动词和形容词，由于用户在阐述产品缺陷时很少会用到表达积极情感的词汇，因此这里

对含有积极意味的动词与形容词不予关注；*an* 表示了那些拥有名词词性，但可以发挥形容词作用的词汇。从词性角度，将产品缺陷挖掘的候选词集设定为：

$$CandidateWords = \{noun, verb, adjective\} \quad (1)$$

通过对用户评论数据库的检索，可以获得所有包含候选词集的语段，经过必要的处理即可得到用于主题模型分析的语料库。

3.1.3 主题模型分析

主题模型是一种被广泛应用于各类机器学习与自然语言处理的技术，该技术最早由 Hofmann 在 1999 年提出[13]，它是 LDA 模型的构建基础。LDA 模型是一种贝叶斯概率模型，属于监督学习的降维技术，由 Blei 在 2003 年提出[14]。该模型的基本观点是：在不同潜在主题分布下，词汇往往会显示出不同的概率分布特征；另外，在同一主题下，代表相同主题倾向的词汇/词组出现的概率往往比较接近。基于这些特性，采用 LDA 模型来识别语料库中所包含的主题，以及它们与产品缺陷相关的词汇分布。

LDA 模型是一种典型的有向概率图模型，LDA 的图模型表示见图 1，α 是每个文档中主题分布的狄利克雷先验的参数，它反映了语料库当中隐含主体间的相对强度；β 是每个主题中词汇分布的狄利克雷先验的参数，反映了所有隐含主题自身的概率分布。LDA 模型有两个基本假设：每个文档都是一系列主题的集合；每个词汇都属于文档当中的一个主题。

LDA 模型的分析流程主要由两个步骤组成：文本生成与参数估计。

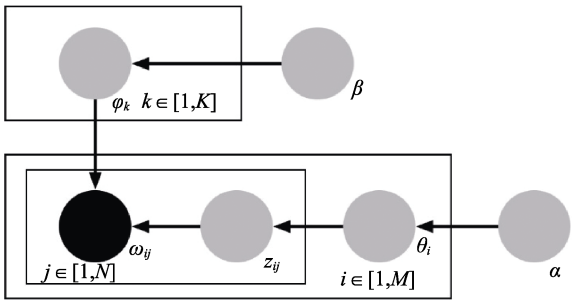

图 1 LDA 的图模型表示

Fig.1 Graphical model representation of LDA

1）文本生成。在文本生成部分，LDA 模型假设语料库 M 当中的一个文档 m 是一个拥有 N 个词汇的向量 $\boldsymbol{w}_i = (w_1, w_2, \cdots, w_N)$，每个 w_j 都来自一个包含 V 维词项的词库。语料库 M 被定义为 $M = \{w_1, w_2, \cdots, w_M\}$，而语料库中的潜在主题则被定义为 $K = \{z_1, z_2, \cdots, z_M\}$。步骤 1 对于主题 z_k：选取 φ_k，φ_k 服从 Dirichlet (β) 分布，即 φ_k ~Dirichlet (β)，φ_k 是主题 z_{kz} 中的词汇分布；选取 θ_i，θ_i 服从 Dirichlet (α) 分布，即 θ_i ~*Dirichlet* (α)，θ_i 是文档 m_i 中的主题分布。步骤 2 对于每个词汇 w_{ij}：选取一个主题 z_{ij}，z_{ij} 服从 Multinomial $(\theta_{i,k})$ 分布，即 z_{ij} ~Multinomial $(\theta_{i,k})$；选取一个词汇 w_{ij}，w_{ij} 服从 Multinomial $(\varphi_{k,t})$ 分布，即 w_{ij} ~Multinomial $(\varphi_{k,t})$。其中：z_{ij} 是文档 m_i 中第 j 个词汇所隶属的主题；w_{ij} 是文档 m_i 中的第 j 个词汇。

2）参数估计。估计的主要目的在于估计每个文档中的主题分布，关联词语的概率分布以及每个文档中特殊主题的混合分布等。通过对 φ 和 θ 的估计，可以获得语料库中所包含主题的相关信息，以及它们在每个文档中所占的比重[15]。采用 Gibbs 抽样[16]对这些参数进行估计。对参数 φ 和 θ 的估计步骤如下。

步骤 1，估计某一文档的概率值：

$$p(w|\alpha,\beta) = \int p(\theta|\alpha)\Big(\prod_{j=1}^{N}\sum_{z_j} p(z_j|\theta)p(w_j|z_j,\beta)\Big)d\theta \quad (2)$$

Collapsed Gibbs 抽样首先估计各个词汇与主题之间的关系，然后通过频率统计来计算这些参数。

步骤 2，计算单词序列下主题序列的条件概率：

$$p(z_i = k|z_{-j}, w) = \frac{p(w,z)}{p(w,z_{-j})} \propto \frac{n_{k,-j}^{t} + \beta_t}{\sum_{t=1}^{V} n_{k,-j}^{t} + \beta_t}(n_{i,-j}^{k} + \alpha_k) \quad (3)$$

其中：z_j 是第 j 个词汇所隶属的主题变量；$-j$ 表示序列当中包含除第 j 项外的所有因子；n_k^{t} 表示主题 z_k 当中词项 t 出现的次数；n_i^{k} 表示文档 m_i 中出现主题 z_k 的次数。

步骤 3，计算 φ 和 θ 的值：

$$\varphi_{k,t} = \frac{n_k^{\mathrm{t}} + \beta_t}{\sum_{t=1}^{V} n_k^{\mathrm{t}} + \beta_t} \quad (4)$$

$$\theta_{i,k} = \frac{n_i^{\mathrm{k}} + \alpha_k}{\sum_{k=1}^{K} n_i^{\mathrm{k}} + \alpha_k} \quad (5)$$

其中：$\varphi_{k,t}$ 表示主题 z_k 中词项 t 的出现概率；$\theta_{i,k}$ 表示文档 m_i 中出现主题 z_k 的概率。通过对概率值的排序，即可获得各个主题与其所属词汇（即每个主题下所包含的词汇）以及各个文档与其所包含主题的关系。在分析过程中，将利用均值复杂度与均值对数似然度，来确定语料库所包含主题的数量。

部分社区用户可能在单一的主题帖中发布涉及多个产品属性的缺陷、问题信息等，且不同的用户在语言习惯方面也存在差异性，他们常常会使用不同的表达方式去描述同一件事情（产品缺陷），因此，在 LDA 模型中，文档往往不会仅仅反映某一个主题，单一文档当中常常会包含多个主题，且主题的分布会因为文档的不同而发生变化。在此，每个主题都聚焦于一个特定的产品属性，LDA 模型可以将与某一属性相关的所有词汇归纳成一个集合，但这些无规律的

词汇并不能直接反映出具体的产品缺陷信息。因此，引入 Apriori 算法，通过关联分析将与属性词频繁搭配的词汇（频繁项）挖掘出来，经过筛选，找出同这些属性缺陷相关的内容，定位产品缺陷信息。

3.2　关联算法分析

关联规则归纳是一种有效的“市场购物篮”分析方法，它常常被用于分析在线商店、超级市场以及邮购公司顾客消费行为中的规律[17]。Apriori 算法是最著名的关联规则归纳技术之一，它主要包含两个分析步骤：（1）找出所有的频繁项集，频繁项集的最小出现次数至少要与预定义的最小支持次数一致；（2）由频繁项集产生强关联规则，强关联规则必须同时满足最小置信度与最小支持度。在此，Apriori 算法将被应用于精确查找产品缺陷与问题。

Apriori 分析的基本思想是：通过对在线评论数据库的循环检索，计算所有项集的支持度，进而发现所有频繁项集并在它们的基础上产生关联规则。项的集合被称为项集；一个项集包含 k 个项则被称为 k-项集。Apriori 算法使用了一种被称为逐层检索的迭代方法，其中(k-1)项集被用于探索 k 项集。在分析的过程中，通过对评论数据库的扫描，统计每个项的出现次数，并将满足最小支持度的项收集起来，从而找出频繁 1-项集的集合，并将该集合记为 L_1。进而利用 L_1 检索出频繁 2-项集的集合 L_2，然后使用 L_2 找出集合 L_3，以此类推。在第 k 次检索前，频繁(k-1)-项集的集合被用于产生候选集合 C_k。第 k 次检索被用来统计集合 C_k 中每个项的出现频次，满足最小支持度的项将被放入到频繁 k-项集的集合 L_k。当候选集合 C_k 为空时，则迭代结束。每找出一个 L_k，都需要对数据库进行一次完整的扫描[18]。

在用户评论当中，部分关键词汇由于距离较远而无法形成直接的语义关系（即不符合语法搭配）[19]，因此，频繁项集当中有可能包含一些与产品缺陷不相关的集合。为了排除这些无效频繁项集的干扰，利用李实等人[19]提出的紧凑规则，对频繁项集进行剪枝，该紧凑规则定义如下：在中文评论当中，R 被设定为一个拥有 G 个词汇的频繁项集。假设评论当中的一句话包含 R，且 R 中所出现的词汇顺序为 $d_1, d_2, \cdots, d_G$。当频繁项集中两个频繁项 d_g 与 d_{g+1} 之间的距离小于三个词汇，则认为该频繁项集满足紧凑规则，且该频繁项集可被认为有效，紧凑规则示例见表 2。经过该步骤，可以获得全部有效的频繁项集。

表 2　紧凑规则示例

Tab.2　Example of compact rule

项集	有效频繁项	无效频繁项
{手机 电池 消耗 快 不耐用 效果 差}	电池 消耗 快	不耐用 效果 差
{接通 声音 消失 嘈杂 大 没 明显 效果}	接通 消失 嘈杂 大	没 明显 效果
{夜间 屏幕 亮 很刺眼 疲劳 疼痛 闪光}	夜间 屏幕 很刺眼 疲劳 疼痛	闪光
{扬声器 不好用 容易 产生 杂音 电流声 大 不好 接受}	电流声 大 不好	扬声器 不好用 容易 产生 杂音
{续航力 差 充电 时间 久}	差 充电 时间	久

注：灰底部分为项 d_g

被检索出的有效频繁项集可以被用于生成关联规则。置信度与支持度是关联规则相关性的两种度量值，它们分别反映了所发现规则的确定性与有用性，如公式(6—7)。当规则同时满足最小置信度与最小支持度时，则被称为强关联规则。提高度是对规则相关性度量的一种补充值，它反映了规则的价值，如公式(8)所示。

$$support(A \Rightarrow B) = P(A \cup B) \tag{6}$$

$$confidence(A \Rightarrow B) = P(B \mid A) = \frac{support(A \cup B)}{support(A)} = \frac{support\_count(A \cup B)}{support\_count(A)} \tag{7}$$

$$lift(A \Rightarrow B) = \frac{confidence(A \Rightarrow B)}{support(B)} \tag{8}$$

其中：A 与 B 分别是项的集合（项集）；$A \Rightarrow B$ 表示一个关联规则，在数据库中具有支持度 s，s 表示同时包含 A 与 B 的主题帖占所有主题帖的比例，它被表示为 $P(A \cup B)$，如果 s 较小，说明 A 与 B 的关系不大，如果 s 较大，则说明 A 与 B 总是相关。置信度 c 表示在出现项集 A 的主题帖中项集 B 也同时出现的概率，即同时包含 A 和 B 的主题帖占包含 A 主题帖的比例，它被表示为 $P(B \mid A)$，如果 c 较小，说明 A 的出现与 B 是否出现关系不大。支持度计数反映了包含项集的主题帖数量。最后，分析强关联规则当中产品属性词汇与情感/评价词汇之间的关系，即可获得产品缺陷信息。

3.3　实验分析

以在线产品社区小米手机官方论坛为研究对象，选择其中的小米 4 手机的讨论与求助版块作为信息源，采集了 2 013 条主题帖（评论总数为 36 957 条），在经过去重、去空处理之后，获得了 33 210 条有效评论。

3.3.1　LDA 模型分析

1）参数设置。通过分析均值复杂度与均值对数

似然度来确定主题数量 K，α 值被设定为 50/K，β 值被设定为 0.01。对于 Gibbs 抽样，对数似然度的计算方法为：

$$\log(p(w\mid z))=k\log\left(\frac{\Gamma(V\beta)}{\Gamma(\beta)^{\mathrm{V}}}\right)+\sum_{k=1}^{K}\left\{\left[\sum_{t=1}^{V}\log(\Gamma(n_k^{\mathrm{t}}+\beta))\right]-\log(\Gamma(\sum_{t=1}^{V}n_k^{\mathrm{t}}+V\beta))\right\} \quad (9)$$

复杂度等价于每个词汇似然值的几何平均数：

$$Perplexity(w)=\exp\left\{-\frac{\log(p(w))}{\sum_{i=1}^{N}\sum_{t=1}^{V}n_i^{\mathrm{t}}}\right\} \quad (10)$$

其中：V 是语料库 M 中的词项（维度）数量，n_i^t 表示第 i 个文档里出现词典中第 t 维词项的次数。较低的均值复杂度或较高的均值对数似然度可以反映出较好的泛化能力，运行 LDA 模型，计算当主题数在 5~60 的均值复杂度与均值对数似然度，结果见图 2。可以看出，当主题数为 20（K=20）时，模型具有较好的泛化性能。

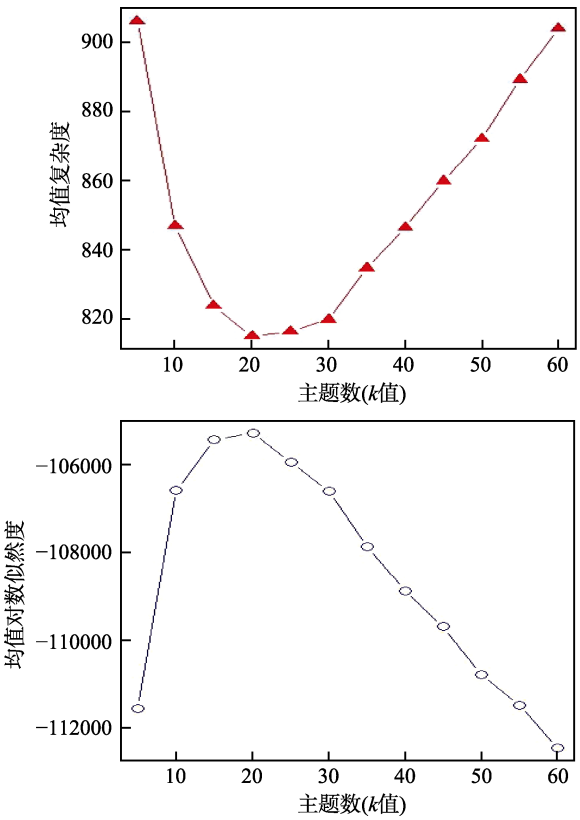

图 2 不同主题数量下的均值复杂度与均值对数似然度

Fig.2 Mean perplexity and mean log-likelihood on different topics

2）分析结果。将 LDA 模型主题数设置为 20，使用 Gibbs 抽样对数据进行 1 000 次迭代分析，对每 100 次迭代的结果进行记录，以找出每个主题当中的关键词分布。通过对 20 个主题内容的分析可以发现，LDA 模型所归纳出的主题并非完全聚焦于产品属性/缺陷等内容，各主题内容总结见表 3，将无意义主题的 17~20 删除后，最终获得 16 个与产品相关的候选主题。

表 3 各主题内容总结

Tab.3 Summary of each topic

主题	主要内容（标签）	主题	主要内容（标签）
1	接听电话	11	售后服务
2	WiFi/无线信号	12	软件安装/卸载
3	软件版本	13	相机/相册
4	购买	14	系统升级
5	电池/充电器	15	质量/感受
6	GPS	16	硬件
7	屏幕/触屏	17	智能
8	物流	18	小米
9	设置	19	摄影大赛
10	RAM	20	骁龙

主题分析可以确定用户评论当中所包含的特定产品属性，并能够获得与这些属性相关的一系列词语，但是这种层次的内容无法直观反映出具体的产品缺陷信息，因此需要使用 Apriori 算法对数据进行处理，以获得更为直接的结果。

3.3.2 Apriori 算法分析

以主题 1“接听电话”作为示例对关联算法分析进行阐释。在进行 Apriori 分析前，需要设定最小支持度与置信度值，参照廖晓等人[20]与安兴茹[21]的研究结果，这两个阈值分别设置为 0.005 与 0.5。主题 1 当中项集的有向图见图 3，直观地反映出了主题 1 当中各个项集之间的关系：“接”、“接听”、“打电话”等描述用户通话行为的动词出现频率较高，且与“电话”、“听筒”等名词频繁搭配，进而又与“不”、

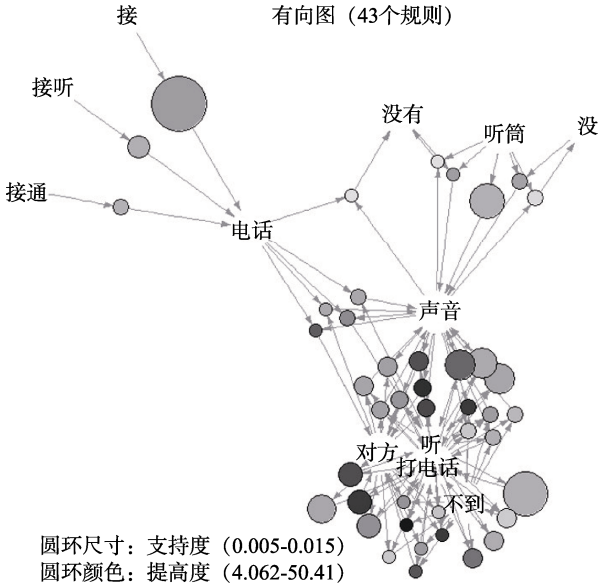

图 3 主题 1 当中项集的有向图

Fig.3 Directed graph of itemsets in Topic 1

“没有”等否定副词发生联系，并指向“声音”等名词，反映出该主题下用户发表评论的内容，大多围绕“接听电话无声音”这一主旨展开。

通过分析主题 1 中的频繁项集，可以得到六个强关联规则，见表 4，这六个规则分别是对同一事件的六种不同描述方式，它们当中有可能包含了用户对产品缺陷/问题的阐述。将这六个强关联规则当中的词汇顺序进行调整，便可以获得易于理解的中文语段。经过分析可知，主题 1 所反映的产品缺陷是“接听电话时无法听到对方的声音”。

表 4　主题 1 中的强关联规则
Tab.4　Strong rules in Topic 1

关联规则	中文语句	支持度	置信度	提高度
{接通，没有，声音} → {电话}	接通电话没有声音	0.005 595	0.647 059	13.678 68
{接听，没有，声音} → {电话}	接听电话没有声音	0.007 121	0.666 667	14.093 19
{接，没有，声音} → {电话}	接电话没有声音	0.014 751	0.878 788	18.577 39
{没，听筒} → {声音}	听筒没声音	0.005 595	0.785 714	16.974 88
{没有，听筒} → {声音}	听筒没有声音	0.005 086	0.833 333	18.003 66
{打电话，对方，不到，声音} → {听}	打电话听不到对方声音	0.006 104	0.923 077	37.807 69

与之类似的，通过 Apriori 分析可以找到每个主题当中的全部强关联规则，通过分析这些规则中的词汇关系，能够解析出用户所反映的产品缺陷/问题。当最小支持度为 0.005 时，可以从关键用户评论中解析出十四个产品缺陷，解析出的产品缺陷/问题见表 5。然而，并非所有主题都能反映出有价值的信息，经过分析，可能是以下三个原因造成了这一现象。

表 5　解析出的产品缺陷/问题（最小支持度 0.005）
Tab.5　Extraction of product problems (min sup=0.005)

主题	产品缺陷/问题	支持度	排序
1	接听电话时无法听到对方的声音	0.044252	1
2	WiFi 频繁断开	0.006104	7
2	手机信号比较弱	0.010173	5
5	手机电池非常不耐用	0.005595	8
5	手机充电器不能正常工作	0.008647	6
6	GPS 定位失败/不准确	0.008647	6
7	屏幕显示会出现意外黑屏	0.022415	2
7	触屏失效/触摸控制失效	0.005595	8
8	物流速度太慢	0.006104	7
10	RAM 太小/不够大	0.016785	4
11	对售后服务不满意	0.019329	3
12	手机无法安装软件	0.006104	7
13	照片不清晰/拍照效果不好	0.008647	6
14	操作系统无法升级/系统升级失败	0.005595	8

注：排序为产品问题出现频率的高低次序

（1）原始语料库所包含的数据规模不够大。部分产品缺陷/问题在用户主题帖当中所出现的频率相对较小，无法达到最小支持度，因此被排除在外。（2）用户评论与候选词集/语法结构不匹配。缺陷/问题涉及多个产品属性、用户描述方式较为含蓄等，都有可能导致评论文本与候选语法结构不匹配而无法被 LDA 模型识别的问题。由于汉语言相近词义较多，单纯的词义挖掘很容易使含义相关词汇的相似程度得不到数学意义上的体现。LDA 模型建立在“词袋模型”（不考虑文法以及词的顺序）的假设基础之上，认为每个词都独立存在，这样的强假设不符合实际，会忽略词语之间的关联性，这是 LDA 模型在处理中文词汇时暴露出的缺点之一。（3）由于部分中文词汇包含多个词性而导致系统词性标注错误。

围绕以上几个问题，提出了一些解决思路。对于第一个问题，接下来的研究应当扩大数据库容量，以消除数据规模对结果造成的不利影响。对于第二个问题，未来的研究有可能需要通过人工手段来分析用户所发表评论内容的语义特点，以解析产品缺陷信息；另外，需要对 LDA 的基本算法进行优化，以提高其对中文环境的适应度。对于第三个问题，应该通过构建拓展词库来解决。本文使用了由第三方开发的词典系统来为词汇标注词性，该系统通过将评论中的词语与内置词库中的词汇（预置词性）进行比对来实现。由于本文的用户评论中包含大量专业词语，它们的部分词义并不存在于该系统的词库当中。不同的词义有可能具有不同的词性，系统有可能因为错误的词义而为词汇标注了不恰当的词性，所以，下一步的研究可以通过构建专业词库来克服这些问题。

3.3.3　有效性验证

为了验证方法与结果的有效性，提供了两组对比实验，将该方法所获得结果设置为实验组。

1）与传统挖掘方法的实证对比。目前，面向中文互联网社区环境下的产品缺陷挖掘，主要是基于大规模词汇检索实现的，其步骤如下：构建产品属性、情感词汇、评价词汇词库，并根据研究对象语法特点，设置被采集文本的候选格式；采集文本，经过预处理后形成原始语料库；结合所构建的词库与候选格式，对语料库进行大规模检索；对检索到的匹配内容进行筛选，归纳出产品缺陷信息。利用这一方法，设置对

照组 1，其所使用的研究数据与实验组一致，与传统挖掘方法的实证对比结果见表 6。由表 6 可知，所提出的方法可以准确、高效地发掘出用户评论当中所反映的大部分产品缺陷，同时这一对比结果也突出了传统方法所存在的不足：词库构建过程较为繁琐，且容易产生纰漏；需要大量人工处理环节，工作量大，效率低下；受限于词库的内容，无法找到预置词汇以外的信息。另外，对于该方法未能识别出对照组 1 结果中的“系统版本无法降级”这一问题，经过分析，发现是由于该缺陷相关的规则支持度小于 0.005，而导致其未被系统收录，因此，这一问题可以通过降低最小支持度得到解决。

表 6 与传统挖掘方法的实证对比结果
Tab.6 Empirical comparison with traditional mining methods

实验组结果	对照组 1 结果
接听电话时无法听到对方的声音	接听电话时无法听到对方的声音
WiFi 频繁断开	WiFi 频繁断开
手机信号比较弱	/
手机电池非常不耐用	手机电池非常不耐用
手机充电器不能正常工作	/
GPS 定位失败/不准确	GPS 定位失败/不准确
屏幕显示会出现意外黑屏	/
触屏失效/触摸控制失效	触屏失效/触摸控制失效
物流速度太慢	/
RAM 太小/不够大	RAM 太小/不够大
对售后服务不满意	/
手机无法安装软件	手机无法安装软件
照片不清晰/拍照效果不好	照片不清晰/拍照效果不好
操作系统无法升级/系统升级失败	操作系统无法升级/系统升级失败
/	系统版本无法降级

注：“/”为本组未挖掘到该信息

2）挖掘结果有效性验证。为了验证该方法所取得结果的有效性，将中关村在线的用户评测报告（对照组 2）与所取得的结果进行了对比。中关村在线是国内著名的专业 IT 测评网站，其测评内容涵盖手机、计算机、相机等多个方面，通过对本网站小米 4 手机用户评测报告的挖掘，总共找到六个相关产品问题，小米论坛与中关村在线挖掘结果对比见表 7。通过对比可以发现，实验组的产品缺陷信息涵盖了对照组 2 当中出现的大部分内容，从而证明了实验组结果的有效性。然而，实验组与对照组的结果在排序方面却存在巨大差异，经过分析，有可能是以下原因造成了这一结果。（1）对照组 2 当中仅仅包含 390 个用户评测报告，样本规模较小，有可能导致结果出现误差。（2）小米社区用户在发表产品缺陷相关内容时，其原始动机在于希望企业能够发现并解决这些问题，从而改善产品使用体验与产品性能，其产生内容的行为不受外界影响。中关村在线用户则是在网站的激励下产生内容（如该网站为用户提供半结构化的评测模板），有可能受到网站经营者与其他用户的影响，因此，实验组与对照组用户在网站使用动机方面存在差异，有可能导致两者的结果出现不一致。（3）使用 Apriori 算法分析文本内容的频繁度，需要预先对支持度、置信度、提高度等参数进行设置，如果参数值设置过大，会导致满足条件的关键词过少，从而识别出的产品问题也较少；反之，则识别出的问题会较多。适当降低实验组在关联分析中的支持度与置信度参数值，有可能会发现更多产品问题，从而消除与对照组的差异。

表 7 小米论坛与中关村在线挖掘结果对比
Tab.7 Mining results comparison of Mi forum and ZOL

小米 4 手机缺陷	评测报告数量	支持度	对照组 2 排序	实验组排序
手机信号比较弱	158	0.4051	1	5
产品外观令人不满	84	0.2154	2	/
照片不清晰/拍照效果不好	54	0.1385	3	6
手机电池非常不耐用	46	0.1179	4	8
手机价格虚高	26	0.0667	5	/
RAM 太小/不够大	22	0.0564	6	4

注：“/”为本组未挖掘到该信息

4 结语

目前，虽然围绕网络文本特定信息/主旨提取的方法，如 Gupta 等人的极大熵分类器与 Goldberg 等人的支持向量机分类器，能够较为精确地识别出英文文本的主旨信息，但是这些工具需要预先对输入的语法结构进行精准定义，在实践过程中要投入较多精力对语料进行预处理，且分析过程对于专业词库的依赖性较高。另外，这些方法大多难以直接处理中文文本。针对这一现状，提出一种基于 LDA 模型与 Apriori 算法的中文产品缺陷信息挖掘方法，该方法首先利用 LDA 模型对文本主旨进行分类，识别出用户评论信息中所包含的产品属性词；然后利用 Apriori 算法找出与这些属性频繁搭配的评价性词汇，直观地呈现用户对各产品属性的观点和态度，进而定位产品缺陷与不足。相较现有的产品缺陷挖掘手段，该方法无需对文本语法结构进行精准定义，并且降低了挖掘过程对词库的依赖，在操作效率、结果覆盖率、准确率等方面都具有一定优势。

研究结论对企业开展创新活动具有一定的实践意义。首先，利用所提出的方法，消费品/工业品制造商可以通过分析网络社区、测评网站等用户的贡献

内容，来准确、高效地定位产品缺陷与不足，所获得信息能够用来支持产品研发与改进。其次，该方法可以降低产品缺陷挖掘过程中的人力成本，提高分析效率。最后，该方法能够帮助企业实时获取产品缺陷与用户抱怨信息，从而保证经营者及时作出研发与管理决策。

参考文献：

[1] GUO W, LIANG R Y, WANG L, et al. Exploring Sustained Participation in Firm-hosted Communities in China: the Effects of Social Capital and Active Degree[J]. Behaviour and Information Technology, 2017, 36 (3): 223-242.

[2] VON H E. Democratizing Innovation: the Evolving Phenomenon of User Innovation[J]. Journal Für Betriebswirtschaft, 2005, 55(1): 63-78.

[3] 秦敏, 梁溯. 在线产品创新社区用户识别机制与用户贡献行为研究:基于亲社会行为理论视角[J]. 南开管理评论, 2017, 20(3): 28-39.
QIN Min, LIANG Su. Study on User Recognition Mechanism and Contribution Behavior in Online Innovation Communities: Based on Prosocial Behavior Theory[J]. Nankai Business Review, 2017, 20(3): 28-39.

[4] 杨德清, 张静, 郭伟, 等. 基于在线产品社区的动态用户需求 Kano 模型构建研究[J]. 机械设计, 2018, 35(3): 12-19.
YANG De-qing, ZHANG Jing, GUO Wei, et al. Research on Dynamic Kano Model Construction of Customer Requirements Based on Online Product Community[J]. Journal of Machine Design, 2018, 35(3): 12-19.

[5] LIANG R Y, GUO W, ZHANG L H, et al. Investigating Sustained Participation in Open Design Community in China: the Antecedents of User Loyalty[J]. Sustainability, 2019, 11(8): 2420-2438.

[6] 邱丹逸. 众包模式下产品设计任务推荐[J]. 机械设计, 2017, 34(12): 48-52.
QIU Dan-yi. Product Design Task Recommendation Based on Crowdsourcing Mode[J]. Journal of Machine Design, 2017, 34(12): 48-52.

[7] GUPTA N K. Extracting Phrases Describing Problems with Products and Services from Twitter Messages[J]. Computación y Sistemas, 2013, 17(2): 197-206.

[8] TUTUBALINA E, IVANOV V. Unsupervised Approach to Extracting Problem Phrases from User Reviews of Products[C]. Dublin: Association for Computational Linguistics, 2014.

[9] BRUN C, HAGEGE C. Suggestion Mining: Detecting Suggestions for Improvement in Users' Comments[J]. Research in Computing Science, 2013, 70: 199-209.

[10] RAMANAND J, BHAVSAR K, PEDANEKAR N. Wishful Thinking: Finding Suggestions and Buy Wishes from Product Reviews[C]. Los Angeles: Association for Computational Linguistics, 2010.

[11] GOLDBERG A B, FILLMORE N, ANDRZEJEWSKI D, et al. May All Your Wishes Come True: a Study of Wishes and How to Recognize Them[C]. Boulder: Association for Computational Linguistics, 2009.

[12] 李实, 叶强, 李一军, 等. 挖掘中文网络客户评论的产品特征及情感倾向[J]. 计算机应用研究, 2010, 27(8): 3016-3019.
LI Shi, YE Qiang, LI Yi-jun, et al. Mining Product Features and Sentiment Orientation from Chinese Customer Reviews[J]. Application Research of Computers, 2010, 27(8): 3016-3019.

[13] HOFMANN T. Probabilistic Topic Maps: Navigating through Large Text Collections[C]. Amsterdam: ACM, 1999.

[14] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet Allocation[J]. Journal of Machine Learning Research, 2003, 3(1): 993-1022.

[15] ROSEN-ZVI M, GRIFFITHS T, STEYVERS M, et al. The Author-topic Model for Authors and Documents [C]. Banff: AUAI Press, 2004.

[16] 王振振, 何明, 杜永萍. 基于 LDA 主题模型的文本相似度计算[J]. 计算机科学, 2013, 40(12): 229-232.
WANG Zhen-zhen, HE Ming, DU Yong-ping. Text Similarity Computing Based on Topic Model LDA[J]. Computer Science, 2013, 40(12): 229-232.

[17] BORGELT C, KRUSE R. Induction of Association Rules: Apriori Implementation[C]. Berlin: Center of Applied Statistics and Eco, 2002.

[18] KAMBER M, HAN J, PEI J. Data Mining: Concepts and Techniques[M]. Singapore: Elsevier, 2012.

[19] 李实, 李秋实. 中文评论中产品特征挖掘的剪枝算法研究[J]. 计算机工程, 2011, 37(23): 43-45.
LI Shi, LI Qiu-shi. Research on Pruning Algorithm of Product Feature Mining in Chinese Review[J]. Computer Engineering, 2011, 37(23): 43-45.

[20] 廖晓, 李志宏, 席运江. 基于加权知识网络的企业社区用户创新知识建模及分析方法[J]. 系统工程理论与实践, 2016, 36(1): 94-105.
LIAO Xiao, LI Zhi-hong, XI Yun-jiang. Modeling and Analyzing Methods of User-innovation Knowledge in Enterprise Communities Based on Weighted Knowledge Network[J]. Systems Engineering-Theory and Practice, 2016, 36(1): 94-105.

[21] 安兴茹. 基于正态分布的词频分析法高频词阈值研究[J]. 情报杂志, 2014, 33(10): 129-136.
AN Xing-ru. The Research on the Threshold of High-Frequency Words Based on the Normal Distribution in Word Frequency Analysis[J]. Journal of Intelligence, 2014, 33(10): 129-136.