中国管理科学
Chinese Journal of Management Science
ISSN 1003-207X,CN 11-2835/G3

《中国管理科学》网络首发论文

题目：针对论坛数据特点的汽车质量问题挖掘
作者：王余行，党延忠，徐照光
DOI：10.16381/j.cnki.issn1003-207x.2019.0233
网络首发日期：2020-07-06
引用格式：王余行，党延忠，徐照光．针对论坛数据特点的汽车质量问题挖掘．中国管理科学. https://doi.org/10.16381/j.cnki.issn1003-207x.2019.0233

DOI：10.16381/j.cnki.issn1003-207x.2019.0233

针对论坛数据特点的汽车质量问题挖掘

王余行，党延忠，徐照光
（大连理工大学系统工程研究所，辽宁大连，116024）

摘　要：汽车质量是汽车生产商在市场中立足的根本和发展的保证，了解并掌握用户反馈的汽车质量问题是其维护品牌声誉、提升市场竞争力且最贴近用户的重要途径。本文基于网络论坛数据，对用户的用车体验中隐含的汽车质量问题进行挖掘。针对论坛数据和用户体验的特点，首先选取文本特征，识别出用户体验中涉及汽车质量问题的文本；然后依据质量问题对应汽车部件与问题类型间的关系，提出了一种汽车质量问题的提取方法；最后，利用某实际论坛数据验证了本文方法的可行性与有效性。本文提出的针对论坛数据特点的汽车质量问题挖掘方法，可以帮助汽车生产商及时获取可能存在的汽车质量问题，在质量管理过程中具有重要意义。

关键词： 论坛数据；用户生成内容；汽车质量；质量问题挖掘

1 引言

汽车质量作为汽车生产商核心竞争力的体现，其质量问题深受企业的关注。所谓汽车质量问题是指汽车从生产至使用过程中发生的与客户及社会预期不符的质量状况[1-2]。近年来，由质量问题引起的汽车重大事故、投诉、召回等事件的频繁发生，给企业声誉与经济利益都带来了严重影响。因此，为减少企业损失，企业亟需及时获取并深入了解其汽车质量问题。

企业获取汽车质量问题的信息渠道包括出厂测试、第三方机构(如 JD Power)调研[3]及用户反馈[4]等。其中，用户反馈作为获取用户需求的最直接的方式，被称为提高产品质量最具价值的信息渠道[5]。用户反馈的方法多样，如企业通常直接通过 4S 店获取用户反馈的质量问题，但该方法极大地抑制了用户的自主性，且不能收集维修系统中未记录的质量问题，造成有价值信息的大量流失；问卷调查、车友会等实践社群(COP)[6]、客服热线等形式虽能在一定程度上提高用户的参与度，但这些方法普遍存在样本少、成本高等缺点，而且重要的是用户无法充分且自由地表达用车体验。随着互联网的快速发展，社交媒体等网络平台逐渐成为消费者自由表达用户体验的新渠道[7]。有研究显示，消费者习惯于在社交媒体上了解和反馈汽车的质量状况[8]，专业性的汽车论坛成为用户表达用车体验、评价汽车质量的绝佳载体[9]。用户在论坛中既可以无约束地表达自己的述求和真实体验，又可以与其他用户进行广泛地交流和讨论。这种便利性与自由性使得专业性的汽车论坛中隐含着大量、丰富的有关汽车质量的体验性信息。这些信息直接或间接的影响着消费者的购买决策[10-11]。因此，如果能把汽车论坛中隐含的质量问题挖掘出来，对于汽车生产商而言具有重要的现实意义。

研究论坛用户反馈的质量问题不仅便于企业及早发现并解决问题，而且能够借助质量问题的变化趋势辅助企业进行管理决策，为企业创造更多的商业收益。为此，国内外学者分别从情感和多文本特征两个角度切入，针对论坛用户反馈中隐含的质量问题的挖掘进行了大量研究。Zhang[12]、Abbasi[13]等学者通过情感分析的方法，根据产品各属性的情感极性及综合情感值进行意见分类，利用负面极性消息发现产品缺陷。但他们都假定情感极性与产品质量优劣密切相关，Loughran 指出负面评论不一定针对产品本身[14]。之后，学者们常从多文本特征分析的角度识别产品的质量状况，并对汽车进行了重点研究。如 Abrahams 基于网络媒体提出一个面向汽车缺陷识别的多文本分析框架，完成缺陷的识别[15]；Jiang 额外考虑论坛数据不平衡的特点，基于语言、社会等文本特征构建 HQRM 模型，并结合 OVA 二叉树分离出汽车质量问题相关的内容[16]。Abrahams 在前人基础上，结合词汇等七方面文本特征，总结了一个基于论坛的 SMART 产品缺陷识别框架，并使用逻辑回归的多元解释模型识别汽车等产品的缺陷[17]。蒋等结合中文环境下的文本特征，利用半监督学习算法对汽车产品缺陷内容进行了分类[18]。Liu 指出以往学者的研究大多是对质量缺陷相关评论的发现，并未对质量缺陷本身进行研究，故提出使用 LDA 聚类的方法获取汽车缺陷的主题分布[19]。

综上所述，学者虽已认识到从论坛用户反馈中获取汽车质量问题的可行性，但却很少针对论坛数据特点深入挖掘质量问题对应的汽车部件与问题类型，提取质量问题的模型及方法相对匮乏。针对上述不足，本文在已有研究的基础上，从论坛数据特点出发，对原有获取汽车质量问题的方法进行改进，并继续深入探索论坛中隐含的汽车质量问题对应的汽车部件及问题类型，构建较为完整的自动提取汽车质量问题的方法体系。

2 研究框架

网络论坛是 Web2.0 环境下一种新兴的信息

基金项目：国家自然科学基金资助项目（71871041）
通讯作者简介：徐照光（1989—），男（汉族），江西景德镇人，大连理工大学经济管理学院博士后，研究方向：知识管理、数据挖掘等，E-mail：zhgxu@dlut.edu.cn

传播渠道，拥有开放性、强交互性、快速反响等特点。论坛文本大多表现为用户的生活见闻及真实感受等，这些大量的用户反馈信息为挖掘汽车质量问题提供便利的同时，也增加了挖掘的难度。一方面，用户频繁使用网络用语，语言描述模糊不清，使得论坛数据更加口语化、简洁化，且多表现为短文本，不符合标准的句法结构。另一方面，用户在汽车领域的知识的局限性使其难以用规范化的专业术语表达汽车质量状况。这些特点都使得论坛数据不同于传统数据，大量真实且杂乱的论坛数据造成传统的数据挖掘方法并不完全适用于本文研究问题。因此，本文针对论坛数据特点，提出了本文的研究框架，如图 1 所示。

首先，对论坛数据进行采集与清洗，包括文本规范化、分词等操作。其次，基于论坛数据特点选取文本特征，并结合合适的分类算法识别出论坛中涉及汽车质量问题的文本。之后，利用关联规则统计频繁项的方法从涉及汽车质量问题的文本中自动提取出现质量问题的汽车部件，通过语义聚类获取其对应的问题类型，并依据汽车部件与问题类型的关系，反向提取问题类型对应的汽车部件，重复该提取过程，对二者遗漏的部分进行动态补充。最终组合汽车部件及相应的问题类型，得到可能存在的汽车质量问题，实现针对论坛数据特点的汽车质量问题的挖掘。

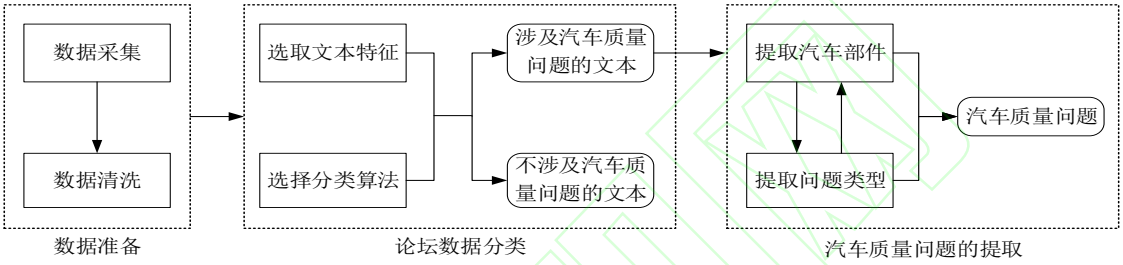

图 1 汽车质量问题的挖掘框架

3 数据准备与数据分类

3.1 数据准备

论坛数据包括非结构化数据与结构化数据两类。其中，非结构化数据主要体现为文本，由于论坛中的汽车质量问题一般会在帖子标题中高度概括，或在帖子内容中展开详细描述，故将帖子的标题与内容作为本研究的文本语料。而对结构化数据而言，现阶段论坛平台提供的很多结构化数据虽有巨大价值但未被充分利用，如与文本相关的数值描述、时间等，故在数据获取阶段，本文尽量完整地收集与汽车质量问题有可能相关的结构化数据，从而更有效地挖掘汽车质量问题。

考虑到论坛数据口语化、网络化的特点，本文采取人工构建规范化词典的方法对文本语料进行规范化处理，统一用户的不同表达。此外，为了减少文本分词错误对后续研究的影响，本文将汽车领域常用词、专有名词等词语添加至用户分词词典中，以提高分词精度。

3.2 文本特征的选取

为获取汽车质量问题，应首先将涉及汽车质量问题的文本从用户反馈内容中识别出来，这一过程一般通过文本分类来实现。如 Liu 按用户评论是否涉及产品缺陷，将其分为质量缺陷相关评论和质量缺陷无关评论两类[19]。但现有研究大多利用传统文本处理方法处理论坛数据，并未考虑论坛数据的特点，文本分类过程中选取的特征并不完全适用于本文的研究问题，因此本文结合论坛数据特点，针对性地选取文本特征。

以往学者在选取文本特征时指出，用户表达文本时所用的词汇、文体、句法、情感对衡量内容有用性有显著作用，可用于判断文本中是否存在产品质量问题[20]。其中，词汇特征反映了词语及短语的分布，常用论坛数据的词语词频表示，但鉴于论坛数据短文本的特点，直接利用论坛文本的词频会产生高维稀疏矩阵，给后续计算带来严重负担，故本文利用信息增益的方法降低特征维度，并结合实际只统计信息增益最高的前 φ 个词的词频；文体特征反映的是文本形式上的特征，常用文本中的词数、句子数、句子长度等衡量；句法特征则刻画了用户在描述汽车质量问题时采用的方式，包括各种词性、标点、句子类型的分布状况等；情感特征则反映了用户对产品的态度，常用情感极性及强度表示。因此，本文构建了词汇特征（F_1）、文体特征（F_2）、句法特征（F_3）、情感特征（F_4），用于识别涉及汽车质量问题的文本。

网络论坛的强交互性、快速反响等特点为挖掘汽车质量问题提供了便利，学者们认为论坛本身的特性可以从侧面反映出汽车质量问题。如蒋指出涉及质量问题的文本通常会引起用户的快速反应及高度关注[18]，用户对帖子的点击量、回复量等关注度越高，则帖子涉及汽车质量问题的可能性越大；Min 发现具有真实产品体验的用户，发表内容的真实性会相对较高[21]，用户若具有真

实驾驶经验或其发帖讨论的车型与其拥有车型一致，则该贴的可信度会相对较高。因此本文基于用户对帖子的关注度及帖子的可信度构建社交特征(F_5)、信息质量特征(F_6)，以便更准确地识别涉及质量问题的文本。

论坛用户非专业词汇的使用造成其表述不清，给识别涉及汽车质量问题的文本带来了困难。Abrahams 指出可从用户发表内容中自动抽取词语构建“Automotive Smoke Words”(简称“烟词”)词表，利用用户表达的体验类词汇检测车辆缺陷是否存在[15]。相比于不涉及质量问题的文本，“烟词”在涉及质量问题的文本中出现地更加频繁，因此本文通过对比文本语料中词语的分布，从用户发表内容中自动抽取词语形成本文的“烟词”词表，并结合汽车领域的术语，构建领域特征(F_7)，以识别涉及质量问题的文本。

本文综合上述分析，构造了词汇、文体、句法、情感、社交、信息质量、领域七方面特征，并按照表 1 标准完成特征选取操作。

表 1 特征选取标准

特征类型	选取标准
词汇特征	文本语料中信息增益最高的前 φ 个词语对应的词频；
文体特征	帖子标题、内容分别对应总字数、汉字数、数字数、英文字符数；帖子内容的句数、平均每句字数；
句法特征	帖子标题、内容分别包含各种词性及标点符号的个数（以《汉语词性标记集》为准）；
情感特征	帖子标题、内容分别对应的综合情感极性与强度（按 Hownet 情感词典计算）；
社交特征	帖子内容的点击量、回复量；帖子发布一天内的回复量及参与讨论人数；
信息质量特征	发帖人是否经过实名认证，讨论车型是否与认证车型一致；
领域特征	本文的“烟词”词表中词语的词频；百度百科汽车术语类目下词语的词频；

3.3 文本分类模型

文本语料经由文本特征选取及特征处理后得到相应的文本数据特征，进而结合数据标签与分类算法，组建为文本分类模型。为提高模型的泛化能力，文本语料通常被划分为训练集、测试集及未标注集，训练集用以训练并优化分类模型，测试集用以衡量模型效果，未标注集则用于预测。其中，训练集与测试集合称为标注集。

本文采用监督学习的方法构建文本分类模型。首先，依照上述特征选取的标准，分别建立文本语料每一类特征 $F_i(i=1,2\cdots I)$，并组建特征矩阵 X，作为文本数据特征；其次，采用人工标注的方法，标注训练集与测试集的数据标签，即按是否涉及质量问题，将标注集的文本语料帖子人为划分到“汽车质量问题相关文本”和“汽车质量问题无关文本”两类中；然后，选取分类效果最优的分类算法进行文本分类，得到文本分类模型。其中，文本数据特征构建方法如下：

$$X=[F_1,\cdots,F_i,\cdots,F_I] \tag{1}$$

由于特征矩阵不同维度的特征量级不同，且文本特征的维度较高，经由特征选取后的特征矩阵 X 难以直接使用，因此本文对 X 进行特征处理。先采用 Min_Max 归一化将各维度特征的值都转化到[0，1]之间，后采用 PCA 降维的方法去除特征矩阵中冗余的维度，如下所示：

$$Y=Min\_Max(X) \tag{2}$$

$$Z=PCA(Y) \tag{3}$$

综上所述，文本语料经过文本特征提取与数据归一化、PCA 降维等特征处理操作后，得到最终的特征矩阵 Z，结合数据标签及分类算法构成文本分类模型，将文本语料划分为“汽车质量问题相关文本”和“汽车质量问题无关文本”两类，为后续利用“汽车质量问题相关文本”提取具体的汽车质量问题提供了数据基础。

4 汽车质量问题的提取

4.1 汽车质量问题的提取思路

为更好地分析并解决汽车质量问题，还需进一步提取汽车质量问题对应的汽车部件及问题类型。其中，问题类型一般由失效模式表示，失效模式即是指一个系统、子系统或零部件没有满足它的设计的目的或功能[22]。由于用户对汽车质量术语了解不足，因此，用户反馈在文本表现上多为比较模糊的主观感受，难以直接转化为规范化的失效模式。基于上述特点，本文提出弊病感知的概念，即用户对质量问题类型的主观感知的描述。利用弊病感知这一概念对问题类型进行模糊化处理，先从用户反馈中获取弊病感知，再从弊病感知中获取失效模式，从而降低直接提取失效模式的难度。其中，汽车部件与弊病感知具有双指向的特点，即同一个汽车部件可对应多种弊病感知，一种弊病感知也可对应多个汽车部件，且用户不同的弊病感知按所属类型又可对应同种失效模式。因此汽车部件、弊病感知与失效模式之间的关系可由下图 2 表示，本文称之为“AU-DP-FM”三元关系。其中 AU 表示汽车部件（Automobile Unit），DP 表示弊病感知（Defect Perception），FM 表示失效模式（Failure Mode）。

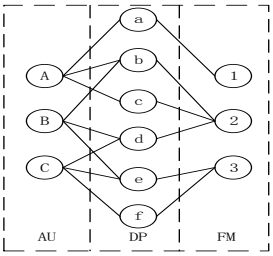

图 2“AU-DP-FM”的三元关系

依照上文，问题类型按视角的不同分为弊病感知与失效模式。将汽车部件分别与弊病感知及失效模式组合，可得到两种视角下的汽车质量问题，本文分别称之为用户及企业层面的汽车质量问题。其中，前者侧重用户主观、精细化的体验描述，能帮助企业细致地分析质量不足，寻找问题发生的原因；后者则侧重企业对质量状况客观、笼统的类别划分，利用质量术语标准化行业内的不同表达，便于企业统计归纳。在提取两种视角下的汽车质量问题时，若忽略汽车部件和弊病感知双指向的特点，对二者独立分析，一旦提取汽车部件时出现遗漏，则其对应的弊病感知及失效模式亦会随之丢失。因此为保证提取的完整度，本文基于“AU-DP-FM”三元关系，提出一种提取汽车质量问题的思路：首先，从“汽车质量问题相关文本”中提取汽车部件；然后，利用汽车部件与弊病感知的关系，从用户对汽车部件的主观体验描述中获取汽车部件对应的弊病感知；然后反向提取弊病感知对应的汽车部件，重复上述过程，对汽车部件及弊病感知遗漏的部分进行螺旋式动态补充，直至二者不再发生变化；之后按弊病感知的所属类别，从中提取失效模式；最终将汽车部件分别与弊病感知、失效模式进行整合，得到用户及企业层面的汽车质量问题，完成汽车质量问题的提取。提取流程如图 3 所示：

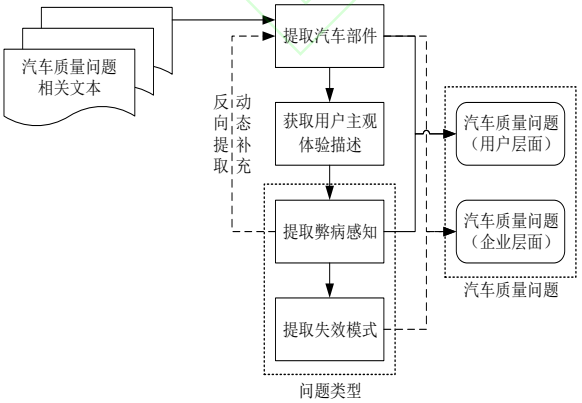

图 3 汽车质量问题提取流程

4.2 汽车部件的提取

Hu 指出产品特征（如产品部件）大多属于名词及名词短语，可采用 Apriori 关联规则统计频繁项的方法，对候选名词进行筛选，提取产品特征[23]。但传统的 Apriori 算法并未考虑句子中项的位置对频繁项集的影响，提取的频繁项集并不完全属于产品特征。李提出可对候选频繁项集进行临近规则剪枝、独立支持度剪枝及非特征项过滤操作，获得最终的产品特征[24]。故本文基于李的思想，通过剪枝与过滤操作，完成汽车部件的提取。

首先，利用语料文本中的名词构建关联事务记录，挖掘候选频繁项集。设 $Cor=\{r_1,\cdots,r_i,\cdots,r_p\}$ 为原始语料文本，$Noun=\{w_1,\cdots,w_j,\cdots,w_q\}$ 为 Cor 中的名词集，$R_i=\{rw_{i1},\cdots,rw_{ij},\cdots,rw_{iq}\}$ 为 $Noun$ 对应的关联事务记录，其中 r_i 表示第 i 条文本语句，w_j 表示 Cor 中包含的第 j 个具有名词特性的词语，rw_{ij} 表示 r_i 中含有 w_j 的个数。从 R_i 中提取支持度高于最小支持度 α 的 δ 项集作为候选频繁项集，记为 $FI=\{fr_1,\cdots,fr_k,\cdots,fr_m\}$。其中 fr_k 为第 k 条项集，fr_k 的支持度为语料 Cor 中包含 fr_k 的语句数占总语句数的比例，而 δ 表示项集 fr_k 中最多含有词语的个数。参考李的研究[24]，取 α 为 1%，δ 为 3。

然后，利用李的方法对候选频繁项集进行剪枝及过滤操作，得到相应的频繁项集 $FI^*=\{fr_1^*,\cdots,fr_i^*,\cdots,fr_n^*\}$。将 FI^* 中的所有词语合并，得到的集合 $FW=\{fw_1,\cdots,fw_g,\cdots,fw_G\}$ 即为汽车部件频繁词集。其中，fw_g 为用户频繁讨论的汽车部件。

基于汽车部件与弊病感知双指向的特点，根据汽车部件频繁词 fw_g 获取相应的弊病感知，进而反向获取弊病感对应的汽车部件，动态补充遗漏的汽车部件词，循环调整直到无法抽取出新的汽车部件及弊病感知为止。若汽车部件非频繁词集 $UW=\{uw_1,\cdots,uw_h,\cdots,uw_H\}\ (uw_h\notin FW)$，则汽车部件词集 AU 为 FW 与 UW 的并集，如下所示：

$$AU=FW\cup UW=\{au_1,\cdots,au_i,\cdots,au_{G+H}\} \quad (4)$$

综上，汽车部件的提取算法可描述为：

步骤 1：对质量问题相关文本 Cor 进行分词及词性标注，获取所有符合产品特征的名词、名词短语、具有名词特性的形容词及动词，组成 $Noun$；

步骤 2：以 r_i 为事务单位，以 w_j 为项，创建 R_i，并从 R_i 中提取高于支持度 α 的 δ 项集，得到 FI；

步骤 3：对 FI 进行剪枝与过滤操作，得到 FI^*，后对 FI^* 中的所有词语进行合并，得到 FW；

步骤 4：提取 FW 中 fw_g 对应的弊病感知，后反向获取弊病感知对应的汽车部件，重复提取过程，直至无法抽取出新的汽车部件及弊病感知，从而得到最终的汽车部件词集 AU。

4.3 问题类型的提取

从论坛用户的用词习惯来看，同一质量问题对应描述词的语义相似度较高，不同问题的描述词的语义相似度较低。故为获取汽车部件对应的弊病感知，应基于语义相似度对汽车部件对应的质量问题描述词进行聚类，即先获取用户对汽车部件的主观描述词（简称为观点词），然后对观点词进行语义聚类。

首先，对于观点词的获取，现有研究大多利用产品特征与观点词的句法依存关系提取特征-观点词对[25]，但此方法对于句法结构不规范的论坛数据却不适用。Jo 指出观点词多为形容词与副词，且常出现于产品特征的临近位置，可定义一个滑动窗口，将与产品特征同在一个窗口内的形容词看作观点词，获取观点词的同时亦得到产品特征与观点词的对应关系[26]。故本文将与汽车部件同在窗口范围内的形容词及副词作为该部件对应的观点词，并结合前人研究，选用词袋模型，以该部件对应的所有观点词中的每一个词为一个维度，将这些观点词表示为空间向量形式，经由 TF-IDF 处理得到弊病感知语料集。

若汽车部件词 fw_g 存在于句子 $s_i(i=1,2,\cdots,M)$ 中，$v_{ij}(j=0,1,\cdots,N)$ 表示 s_i 中 fw_g 的第 j 个观点词。其中 M 为含有 fw_g 的句子的数量，N 为 s_i 中 fw_g 对应观点词的总数。则 fw_g 与 v_{ij} 间的距离应满足下式：

$$\left| loc_{s_i} fw_g - loc_{s_i} v_{ij} \right| \le \omega \tag{5}$$

式中，$loc_{s_i} fw_g$ 表示 fw_g 在 s_i 分词后的句子中的位置，ω 为窗口大小，一般设为 5。

设 $TI_{ih}(h=1,2,\cdots,D)$ 为观点词经由词袋模型及 TF-IDF 处理后对应的权值，D 为 M 条句子中 fw_g 对应的所有观点词的总数，则 fw_g 对应的弊病感知语料集 dc_g 可表示为：

$$dc_g = \begin{bmatrix} TI_{11} & TI_{12} & \cdots & TI_{1D} \\ TI_{21} & TI_{22} & \cdots & TI_{2D} \\ \vdots & \vdots & TI_{ih} & \vdots \\ TI_{M1} & TI_{M2} & \cdots & TI_{MD} \end{bmatrix} \tag{6}$$

之后，基于论坛数据短文本的特点，选取针对短文本也能实现快速收敛的 k-means 算法，对弊病感知语料集进行语义聚类。但传统的 k-means 算法并未考虑文本词语的语义，易产生聚类误差，故本文采用邱提出的基于知网的语义 K-means 算法，完成 dc_g 的语义聚类，以提升聚类效果[27]。其中，词语 W_1 与 W_2 的语义相似度定义如下：

$$Sim(W_1,W_2) = \begin{cases} \max\limits_{i=1\cdots A, j=1\cdots B} Sim(wm_{1i}, wm_{2j}) & W_1, W_2 \in Hownet \\ 0 & \text{其他} \end{cases} \tag{7}$$

$$Sim(wm_{1i}, wm_{2j}) = \frac{\mu}{\mu + min(wd_{ij})} \tag{8}$$

式中，$wm_{1i}(i=1,2,\cdots,A)$ 与 $wm_{2j}(j=1,2,\cdots,B)$ 分别表示 W_1 与 W_2 的含义，$Hownet$ 表示知网词表，wd_{ij} 表示 wm_{1i} 与 wm_{2j} 间的距离，由知网的义原层次树中两含义的路径长度进行获取，而 μ 为调节参数。参照研究[27]，取 μ 为 1.6。

在对 dc_g 进行语义聚类过程中，需事先确定聚类个数，学者指出 k-means 算法的最佳聚类个数可由成本函数获得[28]。随聚类个数的变化过程中，成本函数下降幅度最大处对应的聚类个数即为最佳聚类个数。其中，成本函数为各类簇中心与其内部各成员相似度的加权平方和，如下式所示：

$$CF = \sum_{k=1}^{K} \sum_{C_k} Sim(cw_k, cw_i)^2 \tag{9}$$

式中，C_k 表示第 k 类簇内的数据点，$Sim(cw_k, cw_i)$ 表示第 k 簇的中心词 cw_k 与其簇内成员词 cw_i 的语义相似度。

设 dc_g 经过改进的语义 k-means 算法聚类后，得到的类簇为 $DC_{gk}(k=1,2,\cdots,K)$，DC_{gk} 对应的词集为 $PW_{gk} = \{ow_{gk1},\cdots,ow_{gkl},\cdots,ow_{gkL}\}(l=1,2,\cdots,L)$，其中 ow_{gkl} 表示簇 DC_{gk} 中的第 l 个观点词。将 ow_{gkl} 在原始语料 Cor 中词频最大的词语作为 PW_{gk} 的弊病感知词 AP_{gk}，若 $tf(ow_{gkl})$ 表示 ow_{gkl} 在 Cor 的词频，则：

$$AP_{gk} = arg\,max(tf(ow_{gkl})) \tag{10}$$

由上述分析可知，fw_g 对应的弊病感知 AD_g 即为 AP_{gk} 的合集，最终弊病感知词集 DP 即为汽车部件词 au_i 对应 AD_i 的合集，分别如下式(11)(12)所示，其中 dp_j 为 DP 整合后集合中的弊病感知词。

$$AD_g = AP_{g1} \cup \cdots \cup AP_{gk} \cup \cdots \cup AP_{gK} \tag{11}$$

$$DP = AD_1 \cup \cdots \cup AD_i \cup \cdots \cup AD_{G+H} = \{dp_1,\cdots,dp_j,\cdots,dp_J\} \tag{12}$$

最后，依据弊病感知词的所属类别提取失效模式。由于弊病感知词数较少，且偏口语化，难以直接对词语进行类别划分，故将词语映射为空间向量形式，从而结合聚类算法完成词语的聚类。为更准确地表示论坛数据中的词语，本文以论坛原始文本 Cor 为训练语料，选取 word2vec 算法，通过语料词语的共现关系，得到各弊病感知词的向量表示，并基于弊病感知词数量较少这一特点，采用层次聚类凝聚算法对弊病感知词进行聚类。

设弊病感知词集 DP 中词 dp_j 的向量表示为 $PV_j = [pv_{j1},\cdots,pv_{jt},\cdots,pv_{jT}]$，以词 dp_j 为单独的类，在每次迭代中，将相似度最大的两个类别进行合并，直到满足停止条件，无法继续合并为止。其中，类别相似度常用簇间平均相似度衡量，簇间平均相似度即为不同簇内任意两样本间相似度的平均值，若最邻近的两类簇间平均相似度大于阈值 λ，则停止合并。参考以往研究，λ 常取 0.5。样本间

相似度、类簇间相似度计算公式分别如下：

$$SIM(PV_x,PV_y)=\frac{\sum_{t=1}^{T}(pv_{xt}\cdot pv_{yt})}{\sqrt{\sum_{t=1}^{T}pv_{xt}^2}\cdot\sqrt{\sum_{t=1}^{T}pv_{yt}^2}} \quad (13)$$

$$SIM_{avg}(E_i,E_j)=\frac{1}{|E_i|\times|E_j|}\sum_{x\in E_i}\sum_{y\in E_j}SIM(x,y) \quad (14)$$

设 $DW_i=\{dw_{i1},\cdots,dw_{ij},\cdots,dw_{iQ}\}(i=1,2,\cdots,P,j=1,2,\cdots,Q)$ 为弊病感知词经由向量化、相似的类簇合并后的类簇词集。依上文可知，失效模式为多个相似的弊病感知按所属类别合并后的结果，故 DW_i 表示第 i 类的失效模式。人工总结 DW_i 中 Q 个词语的含义，可得到对应的失效模式词 dw_i^*，P 个 dw_i^* 组成的集合 $FM=\{dw_1^*,\cdots,dw_i^*,\cdots,dw_p^*\}$ 即为最终的失效模式集。

综上，问题类型的提取算法可描述为：

步骤 1：构建汽车部件词 fw_g 对应的 dc_g，利用公式(7)、(8)完成 dc_g 中各词语相似度的计算，并利用公式(9)获得最佳聚类个数，对 dc_g 进行语义 k-means 聚类，得到类簇 DC_{gk} 对应的词集 PW_{gk}；

步骤 2：利用公式(10)获取 PW_{gk} 对应的弊病感知词 AP_{gk}，后反向获取与 AP_{gk} 同在窗口范围内的名词，经由非特征项过滤得到 AP_{gk} 对应的汽车部件词 fw_g'，若 $fw_g'\notin FW$，则将之添加至 UM 中。循环上述过程，直至没有新的汽车部件词及弊病感知词为止，并利用公式(4)得到 AU；

步骤 3：利用公式(11)、(12)将 AU 对应弊病感知词进行合并，得到最终的弊病感知词集 DP；

步骤 4：向量化表示 DP 中的词语，并利用公式(13)(14)分别计算各样本间的相似度以及类簇间的相似度，对相似的类簇进行合并，逐层凝聚直至达到临界条件，得到类簇词集 DW_i；

步骤 5：总结 DW_i 对应的失效模式词 dw_i^*，并由 dw_i^* 组建失效模式集 FM，从而完成问题类型的提取。

5 实例分析

5.1 实验数据

本文以国内某大型汽车论坛的实际数据为基础，探究本文方法的可行性与有效性。为避免销量对论坛数据的影响，本文选取热销且销量稳定的 C 车型作为研究对象，从 C 车型论坛中抓取 2016/1/1-2017/6/30 期间的 91600 条记录，并删除重复记录及空记录等噪声文本，将剩余的 91511 条记录作为原始数据进行后续研究。其中，每条记录包含帖子标题、内容、浏览量、回复量、发帖时间、发帖人等信息。

在进行数据分析之前，需结合本研究目的对原始数据进行预处理，包括停用词表、用户分词词典、规范化词典的构建，数据标签的标注等。首先，为保证分词精度，本文以中科院汉语分词系统 NLPIR 为分词工具，以哈工大停用词词库为停用词表，以汽车领域搜狗细胞词库为用户词典，并在《哈工大信息检索研究室同义词词林扩展版》的基础上人工构建规范化词典。由于篇幅有限，现就规范化词典的部分词表给与显示，如表 2 所示。

表 2　规范化词典

规范前	规范后
无极变速箱,变速箱,变速器,波箱,换挡器,变速箱总成,CVT,MT	变速箱
BAS,EBA,BA,刹车辅助,刹车辅助系统	BAS
4S 店,4s,四儿子,4 儿子	4S 店
…	…
n 档,空挡,N 档,不挂档	N 档

之后，从语料库中随机选取 3000 条记录作为标注集，其余记录则作为未标注集，对标注集进行数据标签的人工标注。标注结果显示，标注数据中共有 489 条涉及汽车质量问题。

5.2 分类结果与评估

依照本文方法，对数据预处理后的文本语料进行文本分类，即可得到涉及汽车质量问题的文本。在实验过程中，结合数据样本，令特征提取过程中的参数 φ 为 10000，提取相应的文本特征，并按 4:1 的比例将标注集随机划分为训练集和测试集，在 SVM、KNN、Naive Bayes、Decision Tree、Random Forest、Xgboost 六种常见的分类算法中选取分类效果最优的算法进行文本分类，完成对“汽车质量问题相关文本”的识别。

为验证本文构建文本分类模型的分类效果，需对本文特征选取方法的效果进行评估。在评估过程中，选取最常用的精确率（Precision）、召回率（Recall）和 F1 值（F1）三个指标，分别衡量分类的查准率、查全率以及综合指数。其中，指标值越接近 1，表示分类效果越好。由于以往研究通常将论坛数据作为传统文本数据，特征选取时只考虑了词汇特征，而本文则针对论坛数据特点选取了词汇、文体等七方面特征，故需将本文的特征选取方法与传统方法进行对比。六种分类算法下，不同特征选取方法对应的分类效果如下图 4 所示。

由图 4 可以看出本文的特征选取效果较传统方法有明显提升，即结合论坛数据特点构建文本特征可提高文本分类的效果，从而证明了本文特征选取方法的有效性。

之后，基于本文特征选取方法选取最优分类算法，各算法分类效果如图 5 所示。

由图 5 可以看出，对于质量问题的识别，SVM 在算法精度上的表现更加出色。即六种分类算法中，SVM 算法更适合区分论坛中涉及汽车质量问题的文本。故本文结合对模型有效性的评估，选取 SVM 算法作为最优分类算法构建最终的文本分类模型。

综上所述，结合本文特征选取方法及 SVM 算法的文本分类模型能有效识别出论坛中涉及汽车质量问题的文本。本文利用该分类模型，预测 88511 条未标注记录的数据标签，最终得到 9360 条涉及汽车质量问题的文本数据，约占未标注集数据总量的 10.6%。

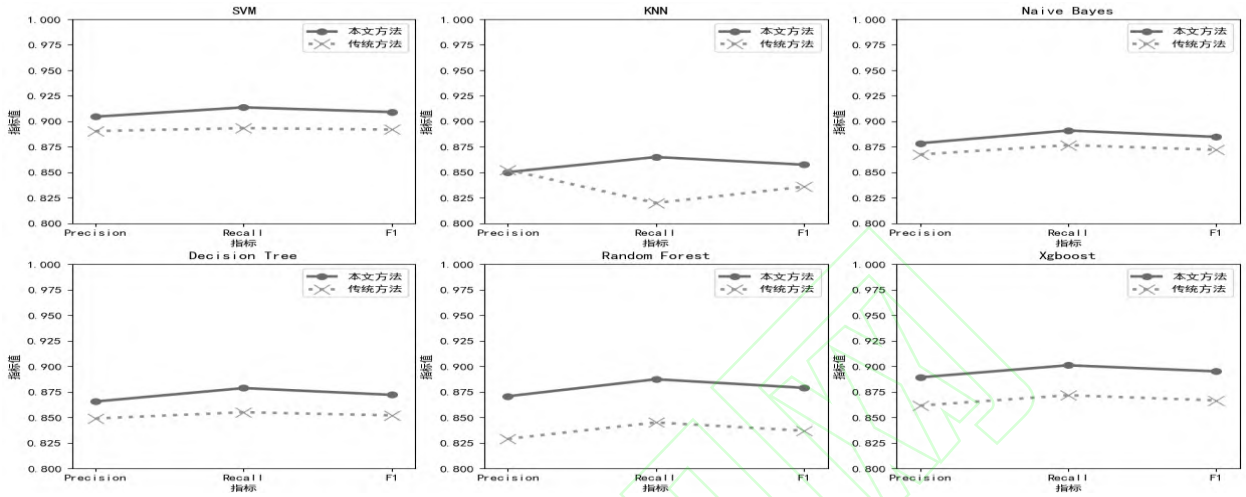

图 4 基于不同特征提取方法的分类效果对比

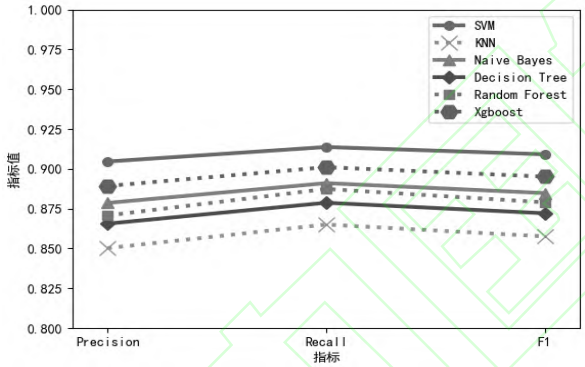

图 5 六种算法的分类效果对比

5.3 提取结果与分析

获取涉及汽车质量问题的文本后，在其基础上提取具体的汽车质量问题。由于论坛数据含有时间标签，能进一步研究汽车质量问题随时间的变化情况，故本文依次以静态、动态两个角度切入，探究汽车质量问题的分布状况及其变化趋势。首先从静态角度分析各时间段对应汽车部件、弊病感知、失效模式以及汽车质量问题的分布状况，帮助企业具体分析汽车质量问题对应的部件及类型；其次从动态角度研究四者随时间的动态变化情况，帮助企业分析趋势变化背后蕴含的机理。

本文将标注集中 489 条及预测集中 9360 条，共计 9849 条涉及汽车质量问题的文本数据作为实验数据，并以季度为周期将其划分为 6 个时间段，具体划分如下表 3 所示。

表 3 时间划分

时间段	时间段简称
2016.1-2016.3	2016 年春
2016.4-2016.6	2016 年夏
2016.7-2016.9	2016 年秋
2016.10-2016.12	2016 年冬
2017.1-2017.3	2017 年春
2017.4-2017.6	2017 年夏

按照本文方法提取各时间段内用户讨论的汽车部件词，篇幅所限以 2017 年夏为例，将汽车部件词按被讨论的频次降序排列，如表 4 所示。

表 4 用户讨论的汽车部件词（2017 年夏）

时间段	频繁词	非频繁词
2017 年夏	发动机、变速箱、油门、方向盘、空调、中控、车门、轮胎、钥匙、手刹、玻璃、车窗、底盘、前轮、车身、涡轮、电池、内饰、后备箱、天窗、机舱、后轮、仪表盘、风扇、车载电脑、后视镜、座椅、发电机、收音机、音响、电瓶、大灯、后排、驾驶室、排气管、喇叭	节气门

经过语义 K-means 聚类及层次聚类，可分别获取各时间段每种部件对应的弊病感知及失效模式，以 2017 年夏被讨论最频繁的汽车部件（发动机）为例，探究其对应的问题类型，如表 5 所示。

表 5 发动机对应问题类型（2017 年夏）

部件	类别	关键词	频次	问题类型	
				弊病感知	失效模式
发动机	0	异响声,发动机,机舱,声音,舱盖,打开	38	异响声	异响问题
发动机	1	顿挫感,发动机,行驶,晃,很大,开车	303	顿挫感	抖动问题
发动机	2	震动,发动机,启停,抖动,刹车,手刹	24	震动	抖动问题
…	…	…	…	…	…
发动机	23	吱吱声,发动机,启动,机舱,启停,很大	46	吱吱声	异响问题

以上述结果为基础，对汽车质量状况进行展开分析。首先，从静态角度出发，用上文的“AU-DP-FM”三元关系描述一时间段内汽车质量状况的整体分布。利用 Gephi 软件得到该时间段的“AU-DP-FM”三元图，用节点表示具体的汽车部件、弊病感知及失效模式，用层间的边连接表示节点间的关系，用节点大小、颜色深浅、边的粗细描述节点被讨论的频繁程度，节点越大、颜色越深、边越粗，则表示用户对其的讨论越频繁。2017 年夏的汽车质量状况分布如下图 6 所示。

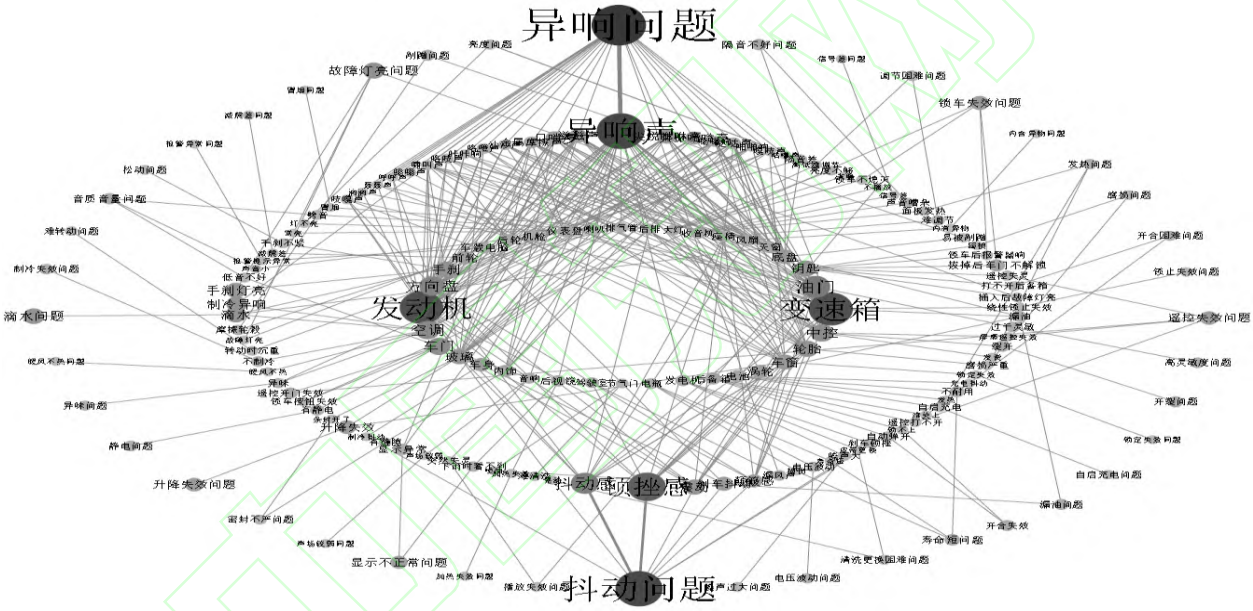

图 6 汽车质量状况分布（2017 年夏）

图 6 中汽车质量状况包含三层，分别对应内层的汽车部件 AU、中间层的弊病感知 DP、外层的失效模式 FM。AU 层与 DP 层间的边连接代表用户层面的汽车质量问题，DP 层与 FM 层间的边连接代表弊病感知的类别划分。观察图 6 可知：从节点来看，AU 层的发动机、变速箱，DP 层的异响声、顿挫感，FM 层的异响问题、抖动问题被用户讨论的比例明显较高。从边连接来看，AU 层与 DP 层间的边连接中，“变速箱顿挫感”被讨论的比例相对较高；DP 层与 FM 层间的边连接中，异响声在异响问题中占的比例最高。

借助“AU-DP-FM”三元关系图，企业可了解并把握汽车的质量状况：①从整体出发，企业可掌握汽车质量的全局状况。根据“AU-DP-FM”图中三层节点得到各汽车部件、弊病感知与失效模式的频率分布，并通过层间的边连接了解用户层面中的各种汽车质量问题的分布、弊病感知的类别划分，进而依照质量状况的严重性进行优先级排序，如优先解决变速箱的异响与抖动问题等。根据优先级顺序，企业可分析问题发生的原因，并调整工作重心，重点关注用户讨论较多的汽车质量问题，并以此为切入点进行针对性营销。②从局部出发，了解具体某部件的质量状况。由于上图 6 中弊病感知与失效模式节点大小及颜色深浅是所有汽车部件对应问题类型叠加的结果，而企业各部门在实际生产中主要关心本部门生产部件的质量状况，因此企业各部门可根据自身需求从上图 6 中选取相应的汽车部件单独分析，获得选定汽车部件对应的“AU-DP-FM”三元图，从局部分析该部件对应的质量状况，进而把握其生

产质量不足并加以解决。

其次，从动态角度出发，企业可分别分析各汽车部件、弊病感知、失效模式及汽车质量问题随时间变化的情况。篇幅所限，现只对汽车质量问题的时变情况展开分析，并依照上文，将汽车质量问题按视角不同分为用户及企业两个层面。选取表 4 中被讨论最频繁的五个部件（发动机、变速箱、油门、方向盘、空调）及对应的问题类型，分析用户及企业层面的汽车质量问题的时变趋势，即用户对话题的讨论频次随时间的变化情况，如下图 7 所示。

由图 7 可以看出，随着时间的推移，变速箱的异响问题呈明显下降趋势，发动机顿挫等抖动问题却有显著上升趋势，空调制冷时的异响问题与时间呈周期性变化，而油门灵敏度过高及方向盘难转动问题的变化与时间却无明显关系。这些实验结果均与现实情况相吻合，如 2015 年底生产商因 C 车型变速箱异响问题对汽车进行了大范围召回，之后变速箱的质量得到相对改善，用户对其讨论也相对减少；2016 年 8 月 C 车型改款，发动机进气方式由自然吸气更换为涡轮增压，引发了用户对发动机相关问题的大量讨论；空调制冷在夏秋两季使用频繁，相应地，制冷时出现异响等问题被讨论的次数也较春冬两季明显增加。

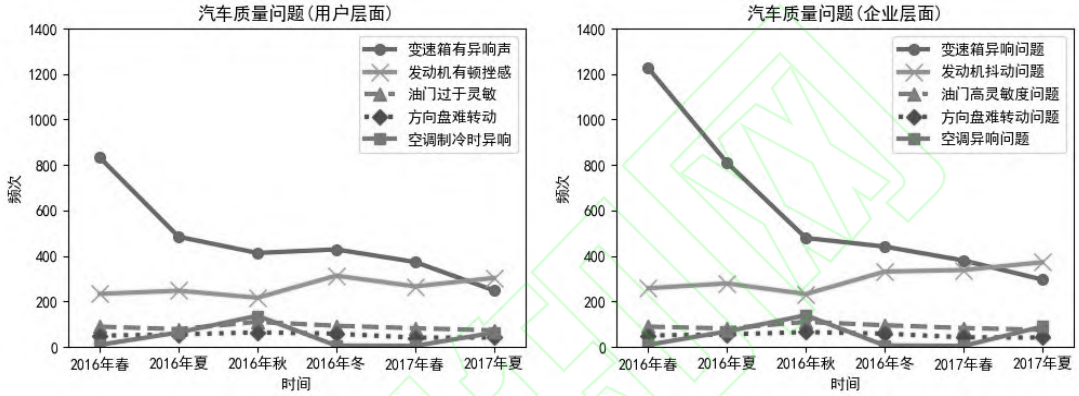

图 7 汽车质量问题时变趋势

借助汽车质量问题被讨论的时变情况，企业可深入分析动态变化趋势背后蕴含的机理：①企业可分析汽车质量问题的热度变化，如突变点、峰值、峰谷及波动状况，进而结合实际情况分析造成曲线变化的深层原因，以便及时调整。如上图 7 中 C 车型发动机改款后，发动机顿挫感等抖动问题持续增长，可以推测此次改款可能是造成发动机抖动的重要因素。②企业可通过对比生产及管理决策对应时间点前后质量问题的变化情况，判断决策成效。如 2015 年底实施汽车异响召回的决策后，汽车异响问题得到明显改善，可以推测此次决策调整成效显著。③企业还可通过对比多条质量问题的时变趋势分析各汽车质量问题间的内部联系，进而系统性的分析问题。如上图 7 油门高灵敏度问题与方向盘难转动问题，二者存在同步变化。企业应深入分析出现“此消彼长”、“同增同减”等现象的原因是用户在体验时的感知误差，还是偶然现象，或是二者确实存在连锁效应，进而深入分析并解决问题。

综合上述分析可知，汽车生产商可从自身与竞争者两个层次展开分析。一方面，企业通过不断获取自身可能存在的各种质量问题及其变化情况，可及时分析并解决质量问题，提高经济收益；另一方面，借助论坛开放性的特点，企业可通过获取并分析竞争者的汽车质量问题，促进产品的差异化与优质化发展，从而在行业中保持竞争优势。针对不同的质量状况，企业可结合不同需求从静态和动态两个角度着手。从静态角度，企业可了解并把握汽车的整体与局部的质量状况；从动态角度，企业则可结合实际深入分析汽车质量状况时变趋势背后蕴含的机理。在对可能出现的汽车质量问题分别进行具体分析时，企业可依照不同视角将其划分为用户及企业层面的汽车质量问题。用户层面的质量问题能帮助企业快速分析问题，寻根溯源进而找到问题发生的原因；而企业层面的质量问题则能标准化行业内的不同表达形式，利于企业进行系统性的统计归纳。汽车生产商若能及时分析并解决各种汽车质量问题，加强对汽车质量的控制，即可减少质量问题的发生，提高品牌声誉及经济收益。

6 结语

论坛的快速发展为用户表达用车体验提供了便利，论坛中大量的用户反馈信息也为汽车生产商全面高效地获取汽车质量问题及其变化趋势创造了条件。以往研究缺少对汽车质量问题的深入挖掘，本文针对论坛数据特点，选取多种文本特征，实现了对涉及汽车质量问题的文本的识别，并在此基础上分析了汽车部件、弊病感知和失效模式的“AU-DP-FM”三元关系，基于三者关系

提出了一种自动提取汽车质量问题的方法，全面高效地实现了对论坛中隐含的汽车质量问题的挖掘，并形象化地展示了汽车质量问题的分布及其动态变化情况。

本文提出的汽车质量问题挖掘方法为企业提供了一种了解汽车质量状况的新思路。利用用户的用车体验信息获取汽车质量问题及其变化态势，便于企业从用户角度出发分析质量不足，深入地剖析问题，从而利于企业的质量提升及产品改进，在质量管理、管理决策以及竞争情报分析等方面都具有重要意义。在未来的研究中，需对文本分类过程中数据标签的人工标注进一步优化，如使用半监督学习等方法进行改进等。并且，本文只针对质量问题本身进行了研究，在后续研究中应加入对问题发生的原因、解决方案及预防措施的深入挖掘。另外，将这一系列挖掘方法整合为固定流程，存入知识库，进而推广应用至其他领域，也是未来质量管理值得探索的工作。

参考文献

[1] 王宗水, 赵红, 秦绪中. 我国家用汽车顾客感知价值及提升策略研究[J]. 中国管理科学, 2016, 24(2):125-133.

[2] Xu Z, Dang Y, Munro P. Knowledge-driven intelligent quality problem-solving system in the automotive industry[J]. Advanced Engineering Informatics, 2018, 38: 441-457.

[3] Jang S, Prasad A, Ratchford B T. Consumer search of multiple information sources and its impact on consumer price satisfaction[J]. Journal of Interactive Marketing, 2017, 40: 24-40.

[4] 闫强, 孟跃. 在线评论的感知有用性影响因素——基于在线影评的实证研究[J]. 中国管理科学, 2013(s1):126-131.

[5] Li Y M, Chen H M, Liou J H, et al. Creating social intelligence for product portfolio design[J]. Decision Support Systems, 2014, 66(C):123-134.

[6] Lee-Kelley L, Turner N. PMO managers' self-determined participation in a purposeful virtual community-of-practice[J]. International Journal of Project Management, 2017, 35(1): 64-77.

[7] 刘宇, 梁循,杨小平. 基于 Petri 网的微博网络信息传播模型[J]. 中国管理科学, 2018, 26(12):161-170.

[8] Raji R A, Mohd Rashid S, Mohd Ishak S, et al. Do firm-created contents on social media enhance brand equity and consumer response among consumers of automotive brands?[J]. Journal of Promotion Management, 2020, 26(1): 19-49.

[9] Hammond M. Users of the world, unite! The challenges and opportunities of Social Media[J]. Business Horizons, 2010, 53(1):59-68.

[10] 冯娇, 姚忠. 基于社会学习理论的在线评论信息对购买决策的影响研究[J]. 中国管理科学, 2016, 24(9):106-114.

[11] 施晓菁, 梁循,孙晓蕾. 基于在线评级和评论的评价者效用机制研究[J]. 中国管理科学, 2016, 24(5):149-157.

[12] Zhang W, Xu H, Wan W. Weakness Finder: Find product weakness from Chinese reviews by using aspects based sentiment analysis[J]. Expert Systems with Applications, 2012, 39(11):10283-10291.

[13] Abbasi A, Chen H, Salem A. Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums[J]. Acm Transactions on Information Systems, 2008, 26(3):1-34.

[14] Loughran T, McDonald B. When is a liability not a liability? Textual analysis, dictionaries, and 10‐Ks[J]. The Journal of Finance, 2011, 66(1): 35-65.

[15] Abrahams A S, Jiao J, Wang G A, et al. Vehicle defect discovery from social media[J]. Decision Support Systems, 2012, 54(1):87-97.

[16] Jiang C, Liu Y, Ding Y, et al. Capturing helpful reviews from social media for product quality improvement: a multi-class classification approach[J]. International Journal of Production Research, 2017, 55(12):3528-3541.

[17] Abrahams A S, Fan W, Wang G A, et al. An Integrated Text Analytic Framework for Product Defect Discovery[J]. Production & Operations Management, 2015, 24(6):975-990.

[18] 蒋翠清,王齐林,刘士喜,等.中文社会媒体环境下半监督学习的汽车缺陷识别方法[J].中国管理科学, 2014(s1).

[19] Liu Y, Jiang C, Ding Y, et al. Identifying helpful quality-related reviews from social media based on attractive quality theory[J]. Total Quality Management & Business Excellence, 2017(1):1-20.

[20] 郝媛媛, 叶强, 李一军. 基于影评数据的在线评论有用性影响因素研究[J]. 管理科学学报, 2010, 13(8):78-88.

[21] Min H J, Park J C. Identifying helpful reviews based on customer's mentions about experiences[J]. Expert Systems with Applications, 2012, 39(15): 11830-11838.

[22] Xu Z, Dang Y, Munro P, et al. A data-driven approach for constructing the component-failure mode matrix for FMEA[J]. Journal of Intelligent Manufacturing, 2020, 31(1): 249-265.

[23] Hu M, Liu B. Mining and summarizing customer reviews[C]//Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004: 168-177.

[24] 李实, 叶强, 李一军,等. 中文网络客户评论的产品特征挖掘方法研究[J]. 管理科学学报, 2009(2):142-152.

[25] Guo H, Zhu H, Guo Z, et al. Product feature categorization with multilevel latent semantic association[C]//Proceedings of the 18th ACM conference on Information and knowledge management. ACM, 2009: 1087-1096.

[26] Jo Y, Oh A H. Aspect and sentiment unification model for online review analysis[C]// ACM International Conference on Web Search and Data Mining. ACM, 2011:815-824.

[27] 邱云飞, 赵彬, 林明明,等. 结合语义改进的 K-means 短文本聚类算法[J]. 计算机工程与应用, 2016, 52(19):78-83.

[28] Kodinariya T M, Makwana P R. Review on determining number of Cluster in K-Means Clustering[J]. International Journal, 2013, 1(6): 90-95.

Mining Automobile Quality Problems based on the characteristics of forum data

Wang Yuhang, Dang Yanzhong,Xu Zhaoguang
(Institute of Systems Engineering,Dalian University of Technology, Dalian 116024, China)

Abstract: As the embodiment of the core competitiveness of automobile manufacturers, automobile quality is the basis and guarantee for the development of automobile manufacturers in the market. Understanding and mastering the automobile quality problems from user feedback is an important means to maintain brand reputation, enhance market competitiveness, and be close to users.Based on the data of online forums, this paper excavates the car quality problems found by the users when using or driving the cars. According to the characteristics of the forum data and user experience,firstly, the text features are selected to identify the texts related to automobile quality problems in the user experience. Then, according to the relationship between automobile units corresponding to quality problems and the types of problems，this dissertation proposed a method to extract automobile quality problems.Uses the Apriori algorithm to extract the automobile units, and uses the semantic K-means clustering and hierarchical clustering algorithm to extract the corresponding problem types. The combination of automobile units and the types of problems leads to the quality problems of automobiles.Finally, the feasibility and effectiveness of this method are verified by actual forum data. The proposed method to mine automobile quality problems based on the characteristics of forum datacan help automobile manufacturers obtain and analyze potential automobile quality problems in time and assist companiesin making management decisions, which is of great significance in the process of quality management.

Keywords: forum data; user-generated content;automobile quality; quality problemsmining