融合多特征 TFIDF 文本分析的汽车造型需求提取方法*

季曹婷,马伟锋,楼　姣,马来宾
(浙江科技学院 信息与电子工程学院,浙江 杭州 310023)

摘　要：针对汽车造型智能设计领域中如何有效提取用户需求的问题,提出一种融合多特征 TFIDF(词频-逆向文件频率)文本分析的汽车造型需求提取方法。首先,通过基于互信息与边界自由度获取大量未登录的专业词汇,优化和修正简单分词后的词汇;然后针对经典 TFIDF 算法的局限性,引入词汇特征因素与情感特征因素,获取用户需求特征候选集;最后根据设定的阈值得到有效的用户需求。实验结果表明,融合多特征 TFIDF 文本分析算法在特征提取方面有一定优势,能有效提取文本中关于汽车造型的用户需求。

关键词：汽车造型;用户需求分析;关键词提取;TFIDF
中图分类号：TN02;TP391　　**文献标识码**：A　　DOI:10.16157/j.issn.0258-7998.200488

中文引用格式：季曹婷,马伟锋,楼姣,等. 融合多特征 TFIDF 文本分析的汽车造型需求提取方法[J].电子技术应用,2021,47(2):16-19,27.
英文引用格式：Ji Caoting,Ma Weifeng,Lou Jiao,et al. An extraction method of car styling requirements by integrating multifeature TFIDF text analysis[J]. Application of Electronic Technique,2021,47(2):16-19,27.

An extraction method of car styling requirements by integrating multi-feature TFIDF text analysis

Ji Caoting,Ma Weifeng,Lou Jiao,Ma Laibin
(School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou 310023,China)

Abstract: In order to extract user requirements effectively in the field of car styling intelligent design, a method for extracting car styling requirements based on multi-feature TFIDF (word frequency-inverse file frequency) text analysis is proposed. Firstly, a large number of unregistered professional vocabularies is obtained through mutual information and boundary degrees of freedom to optimize the vocabulary after simple word segmentation. Next, in order to solve the problem of the limitations of the classic TFIDF algorithm, vocabulary and emotional feature factors are introduced to get user demand feature candidates set. Finally, effective user needs are obtained according to the threshold. The experimental results show that the multi-feature TFIDF text analysis algorithm has certain advantages in feature extraction, and can effectively extract user needs of the car styling in the text.

Key words: car styling;user needs analysis;keyword extraction;TFIDF

0 引言

在智能制造的背景下,个性化生产是未来制造业发展的必然趋势,用户除了对商品基本功能的要求之外,个性化定制的需求正不断地增加[1]。汽车制造业是智能制造的典型应用行业,根据调查,我国超过七成的消费者认为汽车造型是决定购买汽车时的首要考虑因素[2],因此汽车造型能否符合用户需求是个性化汽车造型设计成败的关键[3]。目前,汽车造型的用户需求描述主要以文本数据形式存在[4]。自然语言处理技术是当前文本分析的主流方法,通常采用无监督方法进行自动关键词提取。但是该算法完全基于词频,忽略了词语其他特征对关键词提取影响的问题[5-7]。许多研究人员对此展开研究,赵晓平[8]等人提出文本结构特征与经典的 TFIDF 方法进行融合,应用于科技项目文本的相似度度量计算中;牛永洁[9]等人不仅考虑到词频、词跨度和位置权重特征,还考虑到词性、词长与语义关联度因素,相比经典的 TFIDF 算法有所改进;然而在实际应用中,不仅要考虑到词汇本身的特征信息,而且还需要考虑应用场景的问题。所以余本功[10]等人在解决问答社区关键词提取的问题时融合了词汇特征与社会化问答社区文本的用户关注属性来综合度量词语权重,提升了社区问答关键词提取的效果。

虽然上述研究均取得了一些成果,但是无法有效地对汽车造型的用户需求文本进行提取。本文利用融合多

*基金项目:浙江省基础公益研究计划项目(LGF18F020011)

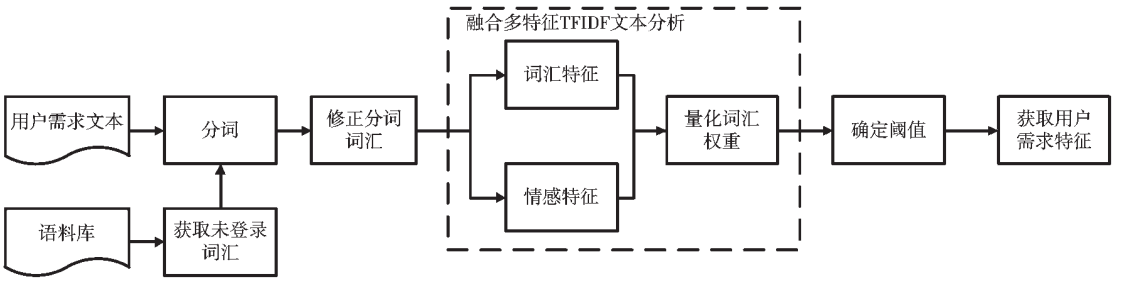

图 1　融合多特征 TFIDF 文本分析的汽车造型需求提取方法流程图

特征 TFIDF 算法对用户需求文本数据进行分析，获取有效的用户需求特征，为汽车造型设计的需求确定提供支撑。

1 方法

本文提出一种融合多特征 TFIDF 文本分析的汽车造型需求提取方法，具体方法流程如图 1 所示。

由图 1 可知，首先基于汽车之家口碑语料库计算得到未登录词汇，结合分词工具从用户需求文本中获取修正后的分词词汇；然后计算词汇特征以及情感特征，并利用改进的 TFIDF 算法量化词汇权重，获取用户需求特征候选集；最后根据实验数据确定阈值，得到有效的用户需求特征。其中，未登录词汇获取方法和融合多特征 TFIDF 算法是有效提取用户需求的关键。

1.1 未登录词汇获取方法

用户需求特征提取首要任务是分词，然而面对口语化的汽车造型风格文本描述，存在着大量未登录词汇，如“腰线很犀利”、“整体车身流线”、“小蛮腰”等出现频率很高但传统分词工具难以区分的词汇。本文基于互信息[11]与边界自由度[12]获取未登录词汇，具体方法流程如图 2 所示。

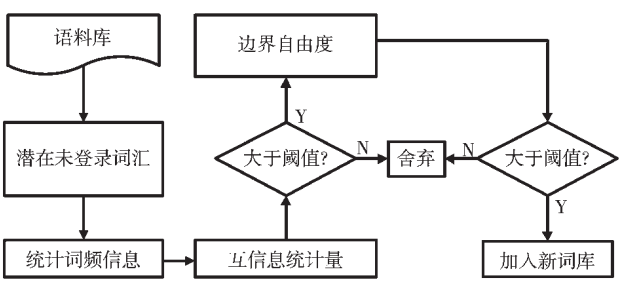

图 2　未登录词汇获取方法流程图

由图 2 可知，首先对语料库进行分词，然后统计分词词汇的频率信息，并根据定义计算边界自由度和互信息，最后根据本文自行确定的阈值确定未登录词汇。

1.2 融合多特征的 TFIDF 算法

1.2.1 词汇特征因素

由于 TFIDF 算法仅考虑了词频信息，没有全面地考虑词汇的本身特性，因此本文从词汇的位置信息、词汇词性、词汇跨度 3 个方面进行考虑，具体内容如表 1 所示。

表 1　词汇特征表

词汇特征	名称	内容	权重
词汇位置	W_{loc}	首句	3
		末句	2
		其他	1
词汇词性	W_{seg}	名词	4
		形容词	3
		动词	2
		其他	1
词汇跨度	W_{span}	词跨度	$\lvert l_i\rvert/\lvert L\rvert$

由表 1 可知，词汇位置信息考虑到首句、末句两个因素，因为文本的首句往往最能体现全文的主题，末句往往是全文的总结性文字描述；词汇词性考虑到名词、形容词和动词 3 个因素，因为在关键词分布中一般以名词或名词性短语、形容词、动词为主。词跨度反映了描述词汇的描述范围，跨段数越多反映该词越重要，全局性越强。$|l_i|$为文档 d_i 中包含词汇的句子总数量，$|L|$为文档 d_i 的分句总数目。

1.2.2 词汇情感特征因素

根据汽车造型设计任务主要是对正向情感文本描述进行用户需求分析的实际要求，提出一种基于语义规则的情感特征计算方法，核心思想是基于汽车造型情感词典，利用词语搭配规则与句型分析规则计算词汇的情感强度，其中情感词典是基于知网词典与 BosonNLP 词典，并结合本文的实际需求，构建了情感词典、否定词典与程度副词词典，详细计算方法与定义如表 2 所示。

1.2.3 算法步骤

TFIDF 算法是基于统计的自动关键字提取最具代表性的方法之一，其核心思想是提取某一文档内容的关键

表 2　基于语义规则的计算方法

规则	组合	计算方式
词语搭配规则	程度副词+情感词	$M_i=w\times e$（e 为情感权重，w 为程度副词权重）
	否定词+情感词	$M_i=w_{no}\times e$（w_{no} 为否定词权重）
	否定词+程度副词+情感词	$M_i=w_{no}\times e+w\times e$
句型分析规则	含有疑问句标志	$M_i=w_{ques}\times M_i$（w_{ques} 为疑问句权重）
	含有感叹句标志词	$M_i=w_{sign}\times M_i$（w_{ques} 为感叹句权重）

字候选集以及对应的权重[13]。如果某关键词出现在某一文档的频率越高，同时出现在其他文档的频率越少，表明该词具备本文档与其他文档区别的能力。TF为某个词出现在一篇文档的次数，IDF是该词区别于其他文档的能力。TF与IDF具体计算方法如式(1)所示，融合多特征的TFIDF方法具体定义如式(2)所示。

$$\begin{cases} \mathrm{TF}_{ij}=\dfrac{n_{ij}}{\sum\limits_{k} n_{kj}} \\ \mathrm{IDF}_{i}=\log\dfrac{|D|}{|D_i|+1} \end{cases} \tag{1}$$

其中，n_{ij}为关键词t_i在文档d_j中出现的次数，$\sum\limits_{k} n_{kj}$表示所有文档中关键字出现的次数之和；$|D|$为语料库中的文档总数，$|D_i|$为包含关键词t_i的文档总数目。

$$W_{ij}=(\mathrm{TF}_{ij}\times\mathrm{IDF}_{ij}+W_{\mathrm{span}_{ij}})\times W_{\mathrm{loc}_{ij}}\times W_{\mathrm{speech}_{ij}}\times M_{ij} \tag{2}$$

权重W_{ij}反映了关键字t_i在文档d_j占比，数值越大，反映了关键词所占比重越大。其中，W_{span}为词汇跨权重，W_{loc}为词汇位置权重，W_{seepch}为词性权重，M_{ij}反映了关键词t_i在文档d_j中的情感权重。具体算法步骤描述如下：

(1)对用户需求文本描述进行文本预处理，将文本d_j划分为n个句子。并载入人工构建的词典、未登录词汇和停用词去除重复词汇和停用词，对分句s进行分词，形成相应的词汇集C。

(2)记录每个分词C_i的词汇信息与在句中的位置I_{index}，并以字典形式存储。

(3)若C_i为情感词汇，在情感词表中寻找情感词，以每个情感词为基准，向前依次寻找程度副词、否定词，并作相应分值计算。

(4)判断该句是否为感叹句，是否为反问句，并作相应分值计算。获得该词汇所在分句的情感强度，即词汇C_i的情感特征权重M_{ij}。

(5)计算词汇C_i的位置特征权重W_{loc}、词性特征权重W_{seepch}与词跨度权重W_{span}，并根据式(2)量化词汇权重W_{ij}，利用改进的TFIDF算法分别得出用户需求特征的关键词候选集k及其权重w。

2 实验与结果分析

2.1 数据集

为了验证本文方法的有效性，选取来自汽车之家网站的用户口碑语料库进行实验对比与分析，并选取2 952篇口碑汽车造型评价数据作为验证集，人工标注合计9 351个关键词标签。关键词标签数据主要描述了用户属性(如用户性别、年龄阶段、用途和工作性质)和汽车风格属性(如时尚、霸气、硬朗等)，实验命名这个数据集为PUBLIC-PRAISE。

2.2 实验结果分析

实验采用准确率[14](precision)、召回率[15](recall)和F1值[16](F1-Measure)来评价关键词提取的效果。

2.2.1 融合不同特征的TFIDF效果对比

为了验证获取未登录词汇方法与融合多特征TFIDF方法的有效性，在PUBLIC-PRAISE数据集合上进行不同组合的实验效果对比，具体实验结果数据如表3所示。

表3 不同特征组合的TFIDF效果对比

方法	precision	recall	F1-Measure
TFIDF	0.352	0.401	0.375
TFIDF+未登录词汇	0.379	0.427	0.401
TFIDF+未登录词汇+词汇特征	0.519	0.593	0.554
TFIDF+未登录词汇+词汇特征+情感特征	0.527	0.600	0.561

对表3分析可知，相比于经典的TFIDF算法而言，本文方法在关键词提取效果上有明显提升，原因在于：(1)引入未登录词汇方法解决了用户需求文本描述中出现传统分词工具不能识别的词汇，一定程度上提升了传统分词工具的分词能力；(2)引入词汇特征解决了经典的TFIDF方法仅考虑词频信息的问题，从词性、词位置与词跨度角度考虑能够提升关键词提取能力；(3)由于包含负面情绪的文本数量较少，因此引入情感特征准确率稍有提升，也说明引入情感特征符合本实验的实际需求，能够去除文本中负面情绪的相关词汇。总体上，本文的方法相比于经典的TFIDF方法在关键词提取效果上有所提升，不仅解决了仅考虑词频信息的问题，而且考虑到了正向情感的用户需求分析的实际问题。

为了提升本文方法的关键词提取的性能，分别设置不同关键词提取个数进行探索，实验结果如图3所示。

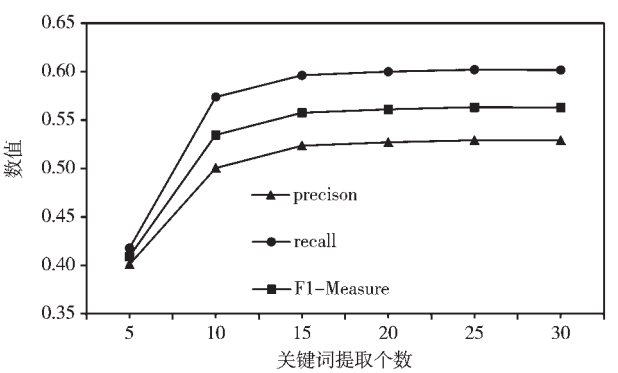

图3 不同关键词提取个数效果对比

对图3分析可知，当关键词个数$K\leqslant 25$时，随着关键词个数的增加，提取效果呈现不断上升的趋势；当$K>25$时，提取效果呈现趋于平稳的趋势。所以，选取$K=25$作为关键词提取个数。

2.2.2 与两种改进的TFIDF方法对比

根据文献[10]提出的基于多属性线性加权的TFIDF与文献[9]提出的融合多因素的TFIDF两种关键词提取方法，基于PUBLIC-PRAISE数据集合，引入未登录词汇，

并统一关键词提取个数 $K=25$，将两种改进的 TFIDF 方法与本文改进的关键词提取方法进行实验对比，具体实验结果如表 4 所示。

表 4 本文方法与改进的 TFIDF 方法对比

方法	precision	recall	F1-Measure
文献[10]	0.448	0.534	0.487
文献[9]	0.498	0.587	0.539
本文	0.529	0.602	0.563

对表 4 分析可知，本文方法相比于两种改进的算法，在准确率、召回率与综合评价指标的 F1 值上提取效果有了明显的提升。原因在于：(1)本文基于文献[10]的思想，引入词频、词性特征以及用户评论数、赞同数和浏览数用户关注属性特征。根据实验结果分析可知，引入用户关注属性对关键词提取意义不大。(2)文献[9]仅考虑了词汇本身的特征，如词频、词性等特征，而本文需要提取出正向情感的用户需求特征，因此该方法不适用于本文研究的实际情况。

为了对比 3 种方法应用于不同文本数量的效果，分别随机选取 500、1 000、1 500、2 000、2 500 条文本集，引入未登录词汇，并统一关键词提取个数 $K=25$，进行关键词提取，得到的实验结果如图 4 所示。

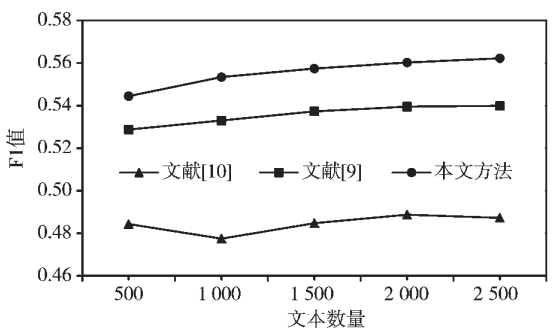

图 4 不同文本数量下 3 种方法对比

对图 4 分析可知，文献[10]随着本文数量的增加提取关键词的能力变弱，文献[9]的方法随着本文数量的增加提取关键词的能力趋于平稳，而本文方法的综合指标 F1 值不仅明显大于其他两种方法，而且呈现增长的趋势，反映了本文方法具备良好的性能。

总体而言，本文方法相比于现有基于 TFIDF 改进的方法效果有所提升，并取得了一定的实验效果。

2.2.3 用户需求特征提取

以 3 位用户的汽车造型风格评价文本描述为例，利用本文方法进行用户需求特征提取，获取文本描述的关键词以及对应的权重，具体文本描述和关键词提取结果如表 5 所示。

由表 5 可知，用户 1 仅包含正向情感的用户需求文本描述，而用户 2 和用户 3 不仅包含正向情感的用户需求文本描述，而且存在负向情感的文本描述。所以设置

表 5 用户需求文本特征提取结果

用户	文本描述	用户需求特征提取	结果
1	喜欢它霸气的马丁脸，特别适合年轻人上下班代步，比较有线条感且外形时尚。	上下班:13.622，年轻人:12.167，霸气:11.003，线条感: 7.089，时尚:7.070，代步:6.221，马丁脸:2.235	上下班，年轻人，霸气，线条感，时尚，代步，马丁脸
2	喜欢外观的成熟车型，适合我这样快步入中年的三十多岁的男性，不适合年轻人开。我不太喜欢个性太张扬的车，已经不适合我这年纪了，用来接送小孩和商务接待不错。	成熟:7.755，商务接待: 2.938，接送小孩: 2.580，中年: 0.059，男性: 0.008，年轻人: −1.492，个性: −14.211	成熟，商务接待，接送小孩，中年，男性
3	热衷于低调，动力足，外观大气，内饰豪华。男性开很不错，其它品牌的小车太张扬霸气了我不喜欢，适合城市公路代步。	男性: 4.989，低调: 3.461，代步: 2.230，大气: 1.399，豪华: 0.347，小车:−5.474，霸气: −6.366	男性，低调，代步，大气，豪华

阈值 $P=0$，筛选出大于阈值的用户需求特征，根据用户需求特征提取结果可知，用户 3 中去除了无效的负向情感词汇，得到了有效的用户需求特征。

3 结论

本文基于统计思想的关键词提取方法，综合考虑词汇特征与情感特征，提出适用于汽车造型设计领域的用户需求文本特征提取方法，相比于经典的无监督提取方法和现阶段研究的无监督关键词提取方法性能有所提升。结果表明，该方法能够有效获取用户需求特征，且辅助汽车造型设计师完成用户需求分析的任务。当然，该方法还存在不足之处：仍需要人工构造词汇集和人工筛选未登录词汇的手段，确保关键词提取的有效性，且该方法采用的词汇特征和情感特征不能完全反映文本的的语义信息，所以该方法的关键词提取性能仍需进一步提升。

参考文献

[1] 陶金泽亚，吴凤羽.工业 4.0 背景下的个性化定制探讨[J].改革与开放，2015(21)：17-18.

[2] 国务院发展研究中心产业经济研究部，中国汽车工程学会，大众汽车集团(中国).中国汽车产业发展报告(2013)[M].北京：社会科学文献出版社，2013.

[3] GOEL A，VATTAM S，WILTGEN B，et al.Cognitive，collaborative，conceptual and creative：four chara-cteristics of the next generation of knowledge-based CAD systems：a study in biologically inspired design[J].Computer-Aided Design，2012，44(10)：879-900.

[4] 卢兆麟，王波，石清吟.面向汽车造型设计模糊前端的思

(下转第 27 页)

参考文献
[1] 赵彦富,随力,李月如.基于脑电图的脑疲劳检测研究进展[J].中国医学物理学杂志,2019,36(11):1312-1316.
[2] 王军,万憬,汪东军,等.脑电反馈训练在高性能战机飞行员中的应用与分析[J].空军医学杂志,2016,32(4):231-233.
[3] CHEN S C,HUANG C K,SU S B.The relationship between attention assessment and EEG control[C].Asia-Pacific Chemical,Biological & Environmental Engineering Society (APCBEES).Proceedings of 2012 2nd International Conference on Biomedical Engineering and Technology.Asia-Pacific Chemical, Biological & Environmental Engineering Society (APCBEES),2012.
[4] Yin Liyong,Zhang Chao,Cui Zhijie.Experimental research on real-time acquisition and monitoring of wearable EEG based on TGAM module[J].Computer Communications,2020,151:76-85.
[5] 鲁在清.临床脑电图学概论[M].南京:东南大学出版社,2018.
[6] nRF24LE1 2.4GHz RF system-on-chip with flash[EB/OL].(2016-11-01)[2020-06-28].http://www.nordicsemi.com/eng/Products/2.4GHz-RF/nRF24LE1.
[7] 凌双明.基于 atmega128 清扫机器人的控制系统设计与研究[D].长沙:湖南大学,2019.
[8] 胡唯唯,王宜怀,张永.基于 K64 的 USB 驱动构件化设计[J].电子技术应用,2017,43(7):55-58.
[9] (美)阿克塞尔森.USB 开发大全(第 4 版)[M].李鸿鹏,郑瑞霞,陈香凝,等,译.北京:人民邮电出版社,2011.
[10] 陈翔宇.基于 TeeChart 的无人机实时参数显示及告警系统[J].遥测遥控,2018,39(3):38-40.

(收稿日期:2020-06-28)

作者简介:
丛林(1989-),通信作者,男,硕士研究生,研究实习员,主要研究方向:航空航天医学、通信与信息系统,E-mail:conglin8383@126.com。
马进(1978-),男,博士研究生,副研究员,主要研究方向:航空航天医学。
胡文东(1964-),男,硕士研究生,研究员,主要研究方向:航空航天功效学、应用心理学。

(上接第 19 页)

维机制研究[J].汽车工程,2017,39(5):869-875.
[5] 罗燕,赵书良,李晓超,等.基于词频统计的文本关键词提取方法[J].计算机应用,2016,36(3):718-725.
[6] 陈伟鹤,刘云.基于词或词组长度和频数的短中文文本关键词提取算法[J].计算机科学,2016,43(12):50-57.
[7] 张建娥.基于多特征融合的中文文本关键词提取方法[J].情报理论与实践,2013,36(10):150-158.
[8] 赵晓平,马文,刘雪萍,等.一种面向科技项目文本的相似度度量方法[J].电子技术应用,2020,46(5):31-34.
[9] 牛永洁,田成龙.融合多因素的 TFIDF 关键词提取算法研究[J].计算机技术与发展,2019,29(7):80-83.
[10] 余本功,李婷,杨颖.基于多属性加权的社会化问答社区关键词提取方法[J].图书情报工作,2018,62(5):132-139.
[11] 赵秦怡,王丽珍.一种基于互信息的串扫描中文文本分词方法[J].情报杂志,2010,29(7):161-162.
[12] 李文坤,张仰森,陈若愚.基于词内部结合度和边界自由度的新词发现[J].计算机应用研究,2015,32(8):2302-2304.
[13] 赵京胜,朱巧明,周国栋,等.自动关键词提取研究综述[J].软件学报,2017,28(9):2431-2449.
[14] SALTON G,BUCKLEY C.Term-weighting approaches in automatic text retrieval[J].Information Processing & Management,1988,13(24):512-523.
[15] MIHALCEA R,TARAU P.TextRank:bringing order into text[C].Proceedings of the Conference on Empirical Methods in Natural Language Processing.Barcelona:Association for Computational Linguistics,2004:404-411.
[16] BLEI D M,NG A Y,JORDAN M I.Latent dirichlet allocation[J].Journal of Machine Learning Research,2003(3):3993-3997.

(收稿日期:2020-06-11)

作者简介:
季曹婷(1996-),女,硕士研究生,主要研究方向:车联网技术。
马伟锋(1979-),通信作者,男,副教授,硕士生导师,主要研究方向:软件架构、大数据与 AI 应用、智能物联系统等,E-mail:mawf@zust.edu.cn。
楼姣(1995-),女,硕士研究生,主要研究方向:车联网技术、自然语言处理。

版权声明