基于最大熵的中文短文本情感分析

黄文明[1]，孙艳秋[2]

（1. 桂林电子科技大学 广西可信软件重点实验室，广西 桂林 541004；
2. 桂林电子科技大学 计算机科学与工程学院，广西 桂林 541004）

摘 要：为挖掘用户对互联网产品的使用评价，从用户角度提高产品体验，建立一个基于最大熵的情感模型，对微博、贴吧的中文短文本评论数据进行情感分析。针对中文短文本评论数据的稀疏性，以分类准确率为目标，综合考虑 F1 值，运用有限拟牛顿平滑算法（L-BFGS）对情感分析模型进行平滑优化；针对有效数据特征维度低的特点，引入 3 类话题模型（TopicModel）辅助模型分类；针对小样本产品的冷启动问题，提出情感话题加权模型，实现中文短文本的情感分析，在时效性和冷启动两个方面验证了模型的鲁棒性。通过在以产品名为关键字划分的评论数据集上进行的对比实验，结果验证了该模型的有效性。

关键词：情感分析；最大熵；短文本；话题模型；冷启动
中图法分类号：TP301.6 **文献标识号**：A **文章编号**：1000-7024（2017）01-0138-06
doi：10.16208/j.issn1000-7024.2017.01.026

Chinese short text sentiment analysis based on maximum entropy

HUANG Wen-ming[1], SUN Yan-qiu[2]

(1. Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin 541004, China;
2. School of Computer Science and Engineering, Guilin University of Electronic Technology, Guilin 541004, China)

Abstract: To improve the user experience and provide assistant decision for product improvement by mining useful information from comment data published by users on the net, an emotion analysis model based on maximum entropy was established. In view of the sparse data of Chinese short text obtained from the internet using the crawler, taking classification accuracy as the goal while taking F1-measure into consideration, the model was optimized using limited-memory Quasi-Newton (L-BFGS) algorithm aiming at improving the over-fitting phenomenon during model training caused by the sparse matrix. Considering the characteristics that effect feature in each feature vector is few, a three-dimensional topic model was introduced as important complement of the model to assist sentiment analysis. To reduce the impact of cold start on small sample data set, a weighting model of feature vectors composed of emotional words in comments and topics generated by topic model was proposed, and its validity and robustness were verified through contrast experiments. The emotion analysis model was verified to be solid through experiments based on comment data divided by the names of products.

Key words: emotion analysis; maximum entropy; short text; topic model; cold start

0 引 言

目前微博、贴吧和各大手机软件应用商店已经成为人们发表对互联网和智能终端应用评论的重要途径，这些信息源含有大量用户对产品质量、使用感受和售后服务的评价、观点和情感。通过挖掘分析这些情感信息，有利于发现产品和竞品的特点，优势和劣势，有助于辅助产品改进决策，改善用户体验。

比较公认的文本情感分析研究工作开始于 Pang 等对英文文本进行情感分类的研究[1,2]，英文文本分类和情感分析也取得了很多可观的成果，但是将英文情感分析的成果应用到具体的中文文本情感分类领域，存在局限性[3]。

目前，中文文本情感分析的方法仍大多借助于英文情感分析的理论和研究结果。短文本情感分析和文本分类具

收稿日期：2015-11-24；**修订日期**：2016-01-29
基金项目：广西可信软件重点实验室研究课题基金项目（kx201106）
作者简介：黄文明（1963-），男，广西桂林人，教授，研究方向为网格计算、图形图像处理、软件工程；孙艳秋（1989-），女，吉林长春人，硕士，研究方向为机器学习、情感分析、软件测试。E-mail：sophie_as@163.com

有相同点，文本分类的理论方法可以运用到短文本情感分析中，但同时由于短文本情感分析问题和文本分类问题的不同点，也决定不能将文本分类的方法直接移植到情感分析上。例如“我今天很难过”，在广泛情感分析中的情感极性为负向，而“今天天气真好，我很开心”的情感极性则为相对的正向，而在对互联网产品的评论数据的情感分析问题中，以上两个评论均被认为是情感中性或无关评论。

本文主要研究来源于微博、贴吧、应用商店的针对特定互联网产品的评论数据的情感分析问题。中文评论数据的情感分析存在以下问题：目前尚未有一个通用的情感词库和网络用语词库[4]；评论数据短小，可用的有效词汇特征数少[5]；数据稀疏，会导致模型性能变差；部分新产品数据样本量较少，训练出的模型表现差，无法满足要求。针对以上问题，做了以下工作：在情感词方面，以广泛情感分析的词库为基础进行了扩充；针对评论数据文本短小、情感信息集中的特点，引入了3类话题模型，将话题结果作为模型的辅助特征；针对数据稀疏问题，运用L-BFGS平滑算法对模型进行优化；针对小样本数据和冷启动问题，提出情感话题加权模型，进行情感分析。最后，通过支持向量机和最大熵两种方法进行对比，验证了模型的有效性，利用时间推移数据验证了模型的鲁棒性。

1 相关工作

1.1 支持向量机

统计学习理论的实际风险包括经验风险和置信范围两部分，注重训练样本的经验风险误差，而没有注重最小置信区间值，并不具有很理想的泛化能力。支持向量机（support vector machine，SVM）不同于传统的统计学习理论，它的分类准则是结构风险最小化，在满足训练误差约束的前提下将置信范围最小化作为优化的目标。SVM是二分类分类器，在二分类问题中具有明显的优势，在处理具有小样本、非线性等特点的问题中具有特有的优势，并具有非常优秀的推广能力[6,7]。

SVM算法是一个二分类分类器，在二分类问题中的线性判别函数的一般表达式为

$$g(x)=w^t+b \tag{1}$$

方程 $g(x)=0$ 定义的是一个分类超平面 H，超平面 H 两侧分别为具有不同类别标签的两类样本，超平面 H 将其分隔开来。设有 N 个样本 (x_1, y_1)，(x_2, y_2)，…，(x_n, y_n)，其中 $x_i \in R^n$，$y_i \in \{+1, -1\}$。则有分类规则

$$\begin{aligned} w^t x+b \geqslant 0, y_i &= +1 \\ w^t x+b < 0, y_i &= -1 \end{aligned} \tag{2}$$

由于映射到高维空间后训练样本可分，改写全向量的模，分类规则可写作

$$\begin{aligned} w^t x+b \geqslant \delta, y_i &= +1 \\ w^t x+b < -\delta, y_i &= -1 \end{aligned} \tag{3}$$

归一化处理后

$$\begin{aligned} w^t x+b \geqslant 1, y_i &= +1 \\ w^t x+b < -1, y_i &= -1 \end{aligned} \tag{4}$$

本文的短文本情感分析是一个三分类问题，因此构建两个SVM二分类模型，如图1所示。

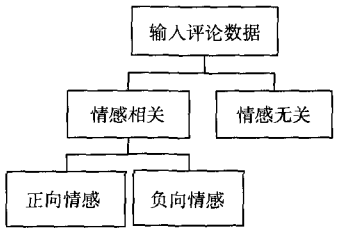

图1 支持向量机三分类情感分析流程

1.2 最大熵

熵最开始是热力学领域的名词，后被推广到计算科学领域，其表达式为[8]

$$\Delta S \approx \frac{Q}{T} \tag{5}$$

熵用来度量随机变量的不确定性，随着随机变量的不确定性增大，熵值增大；反之当随机变量的不确定性减小，熵值减小，当随机变量为定值时，其熵为0。最大熵在满足已有事实（知识）的基础上，对未知事物的特征独立性不做任何假设[9]。

最大熵方法在根本上是一个条件约束问题，最大熵的含义则是在所有满足约束条件的模型中选择熵最大的模型。如果一个随机变量 x 的可能取值为 $x=\{x_1, x_2, \cdots, x_k\}$，其概率分布为 $p(x=x_i)=p_i$，其中 i 取值 $[1, n]$，随机变量 x 的熵定义为

$$H(x)=-\sum_{k=1}^{n} p_k \log(p_k) \tag{6}$$

最大熵原理则是在满足所有约束的情况下，求取随机变量的分布，使熵最大。

最大熵模型的一般表达式

$$\max_{p \in P} H(Y \mid X)=\sum_{(x,y)} p(x,y) \log \frac{1}{p(y \mid x)} \tag{7}$$

1.3 话题模型

人类的语言，可以表达文档或者语句中未显式出现的涵义，随着自然语言处理学科的发展，出现了可以识别潜藏在大规模文档集或语料库中的主题信息[10,11]的隐含狄利克雷分布模型（latent dirichlet allocation，LDA），它是一种非监督学习技术。LDA模型以文本独立性和单词独立性两个假设为基础，它假设文档由一组词构成，词之间没有语法结构的限制和先后顺序，独立可交换，因此每一篇文档可以被看作是一个由词组成的向量，这样文档就可以转换为数字向量，大大简化了建模的复杂性。LDA生成过程

如下：①文档-主题层级：对于每一篇文档，认为是其含有一组主题，在这些主题所构成的分布中随机抽取其中一个主题；②主题-词层级：对于 1 中得到的主题，认为其是由一组词的分布构成的，从其所对应的词所构成的概率分布中抽取出一个单词；③重复①、②中的抽取动作直到文档中的每一个单词都被遍历。本文对每一条评论生成 3 类话题，分别为情感正向、负向和情感无关话题，3 类话题概率和为 1。在每一类话题模型结果中，根据各个话题的概率，选择其中概率最大的一个或多个话题，归一化处理后作为模型训练的补充特征（feature）。

2 本文工作

利用网络爬虫工具，以产品名为关键词从微博、贴吧得到的中文评论数据集，含有以下特点：①含有大量主题不相关数据和无情感类型数据，分类模型需要能够自动将这一类数据标注为不相关数据。②含有错别字、缩写词、网络用语、颜文字、表情符号，针对这一特点，建立了同义词词库、网络用语词库，和表情符号情感极性字典。③文本长度较短，情感较集中，分类模型需要能够充分利用每一个词语特征。④对于不同产品，相同的词汇，可能具有不同的情感极性，针对此问题，分别建立了情感词库和公用的情感词库。

2.1 预处理

（1）情感词典构建

中文情感分析中，尚未有完整词典能涵盖所有中文情感词，并且本文研究的针对特定产品的情感分析与广泛情感分析相比具有特殊性，不能直接借用广泛情感分析的词典。因此，构建一个尽可能丰富的情感词典是本文一项基础并且非常重要的工作。台湾大学自然语言处理实验室发布的情感词库和哈尔滨工业大学发布的停用词库是目前使用较为广泛的基础词库，本文以上述词库和数据堂网络用语词典为基础，加入微博表情符号库和颜文字表情符号库，见表 1。针对不同产品的评论数据集的特点，对基础词库进行删减和补充，形成新的扩充词库，示例见表 2，同时考虑到相同的词汇，在不同产品线分类中可能具有不同的含义，为每个以产品名划分的数据集，分别建立情感词库。对于描述产品功能性的词汇，建立同义词词典，见表 3。基于新扩充的词库，对评论数据进行分词。

（2）数据清洗

本文采用有标注的数据集进行训练，数据的质量会严重影响模型的训练和分类精度，因此首先对数据集中数据相同标签（label）不同的数据进行重新标注，使相同的数据标签统一。对数据中的中英文标点符号进行统一。每一条评论数据，对其分词后所包含的词汇，用其在同义词词典中对应的同义词进行替换。

表 1 表情符号库、颜文字词库

情感极性	示例表情
1	
0	
−1	
1	ヾ(o◕∀◕)ﾉ
0	_(:3」∠)_
−1	(╬▔皿▔)凸

表 2 扩充词库

情感极性	示例扩充词
1	普大喜奔、何厚华
0	城会玩、上交国
−1	坑爹、然并卵、醉不行

表 3 同义词词典

原词	替换后
不可以设置	设置不了
无法设置	设置不了
设置不起	设置不了

2.2 特征选择

特征选择是指在特征候选集合中，选择一个能表达这个随机过程的统计特征的特征集合子集。常用的中文文本分类特征选择方法有文档频率（DF）、信息增益（IG）、互信息（MI）、期望交叉熵（KL 距离）等[12]。

本文所用短文本来源于用户对特定产品的评论数据，字数大多在 140 字以内，数据去停用词后的词汇量在 1 到 30 之间。特征选择会滤掉部分特征，这种方法可能存在的问题是，会忽略掉一些有用的信息。由于关键词个数较少，为了避免特征选择可能过滤掉任何有用信息，本文在利用重新修正的停用词库去停用词之后，采用全量关键词作为特征。同时，加入评论数据中的产品词个数，竞品词个数，关键词个数，情感词个数和 3 类话题模型作为分类的特征。

2.3 平滑优化

中文短文本评论数据转化而来的特征向量具有稀疏性，在对于本质上是一种指数形式的最大似然模型的最大熵来说，当特征向量稀疏时，模型性质会变差，因此需要对其进行平滑优化，以减少或克服训练过程过适应的影响，而在可选择的多种平滑方法中，高斯先验分布的表现较为出色。

最大熵模型上的高斯先验分布表达式为

$$p(\lambda)=\frac{1}{\sqrt{2\pi\delta^2}}\prod_i \exp\left(-\frac{\lambda_i^2}{2\delta^2}\right) \tag{8}$$

式中：δ^2 是高斯变量。式（8）对模型施加了一个平滑约束后可以放宽最大熵模型的约束，从而希望改善模型训练过适应（over-fitting）的问题。

2.4 情感话题加权模型

对于通过产品划分的数据集，分别以最大熵算法为基础训练产生了各自的情感分类模型。但是对于部分新接入的产品线或小样本产品，可用建模的数据量只有几百条，所训练产生的模型表现差，模型预测准确率极低。考虑到不同产品线的评论数据具有相同的数据源、数据格式和表达方式，通过计算小样本产品和模型表现优秀的产品之间由情感词和话题构成的特征向量的相似度，选择表现优秀的分类模型的加权结果作为预测结果，来对小样本数据进行情感分析，以希望解决小样本产品分类效果差的问题。

现有 M 个表现优秀的分类模型，待预测数据集具有 N 个变量（通常 $N\leqslant 500$），则变量 x 的预测流程为：

（1）从 M 个模型中随机抽取 N 个由情感词和话题组成的特征向量，构成矩阵 A_1，A_2，…，A_M，与待预测数据集的特征向量构成的矩阵 B 对应，计算矩阵乘积和 $\sum_{i=1}^{M} A_i^T B$，本文将此矩阵和称为数据相似度。则待预测模型与 M 个模型的相似度可记为 ∂_i。

（2）对 M 个模型，模型 M_i 的预测结果记为 p_i。根据数据相似度加权公式，可得加权结果 $\sum_{i=1}^{m} \partial_i p_i$。

（3）加权结果为正，则判定预测结果为+1，反之加权结果为负，判定预测结果为−1。对于 M 个预测结果中出现多个分类结果具有相同概率且在所有预测结果中概率最大的情况，本着准确率最大为准则，将其分类结果置为0。

3 实验验证与分析

利用爬虫工具，以产品名为关键词在微博和贴吧上爬取数据，采用K折交叉验证（K=10）的方式，在其中3组数据上进行对比实验，以证明平滑后的最大熵模型有效。对另外两组小样本数据，分别通过3个其它产品表现良好的模型进行预测，并将预测结果与加权模型结果进行对比，以证明加权模型的对抗冷启动问题的能力。两组实验均通过准确率、召回率和F1值3个测试指标进行考量。

3.1 数据分析

由于本文研究问题与传统的情感分析具有一定差异性，对数据的处理也有所不同，以下对数据进行简要分析展示。

示例：

原始数据：更新后总是出现闪退问题，希望解决一下这个问题！

分词结果：‘更新’后‘总是’出现‘闪退’问题，‘希望’解决一下这个‘问题’！

分析：

1　2　3　4　5　6　7　8　9　10　11　12　13

‘更新’后‘总是’出现‘闪退’问题，‘希望’解决‘一下’这个‘问题’！

该条数据中共有单词13个。其中单词3、8、9在情感分析中的分别为停用词、正向极性、无极性词，在本文问题中，均为情感极性为−1的情感词。单词7、11为停用词。单词13为叹号，本文问题中作为特殊符号被认为是情感特征之一。

3.2 模型有效性

（1）改进模型的有效性

3组数据，分别采用SVM算法、最大熵算法和改进后的最大熵算法进行模型训练和评测。实验结果见表4～表6。

表4 算法在产品A上的预测结果

产品A	准确率	召回率	F1值
最大熵	0.681	0.616	0.647
最大熵+词库	0.727	0.568	0.637
SVM+词库	0.705	0.761	0.732
改进最大熵	0.791	0.698	0.742

表5 算法在产品B上的预测结果

产品B	准确率	召回率	F1值
最大熵	0.61	0.565	0.587
最大熵+词库	0.746	0.517	0.611
SVM+词库	0.668	0.652	0.66
改进最大熵	0.799	0.735	0.765

表6 算法在产品C上的预测结果

产品C	准确率	召回率	F1值
最大熵	0.522	0.242	0.33
最大熵+词库	0.62	0.31	0.413
SVM+词库	0.834	0.152	0.257
改进最大熵	0.662	0.439	0.528

产品A数据集，正向评论数据16 439条，中性评论数据36 449条，负向评论数据28 858条，样本较均衡。产品B数据集，正向评论数据17 842条，中性评论数据44 961条，负向评论数据12 263条，样本较均衡。产品C数据集，正向评论数据681条，中性评论数据22 999条，负向评论数据1296条。中性评论数据超过正向数据和负向数据和的十倍，训练和测试样本严重不平衡，因此导致最大熵模型分类结果整体偏低。尽管如此，从表4～表6中数据仍可以看出，在3个产品线中，扩充词库后的分类效果均优于扩充词库前。产品A、B中，加入了3类话题模型的最大熵模型在准确率、召回率和F值上均比最大熵、支持向量机算法表现优秀。在产品C中，加入了3类话题模型的最

大熵算法在准确率上低于支持向量机（17%），但是在召回率和 F 值均大幅优于支持向量机（分别为 28%和 27%）。

（2）情感话题模型有效性

对于两组小样本数据，分别采用其自身数据训练的模型，其它产品线中准确率最高的模型和情感话题模型进行预测，对比结果见表 7、表 8。

表 7 数据集 A 的预测结果

数据集 A	准确率	召回率	F1 值
模型 A	0.395	0.539	0.456
模型 B	0.444	0.459	0.451
模型 C	0.460	0.457	0.458
加权模型	0.510	0.489	0.499

表 8 数据集 B 的预测结果

数据集 B	准确率	召回率	F1 值
模型 A	0.480	0.537	0.507
模型 B	0.503	0.476	0.489
模型 C	0.497	0.437	0.465
加权模型	0.534	0.481	0.506

加权模型主要用于在新产品接入情感分析时，可用数据少的情况。在这种情况下，相比于召回率，准确率具有更高的优先级。因此在处理模型加权结果时，当出现对数据的情感预测出现多种分类概率相当的情况时，策略性的将数据归为情感 0 分类。可以预见，在这种策略干预下，召回率不会有大幅提高，甚至会有相当程度的下降。而表 4、表 5 的两组数据也验证了这一点。表 4 中 3 个模型加权准确率较 3 个模型准确率均值（43.3%）提高了 10%，召回率与召回率均值（48.5%）相当。表 5 中，模型加权准确率较准确率均值（49.3%）上升了 5%，召回率与召回率均值（48.3%）相当。

3.3 模型鲁棒性

（1）训练数据的时效性

分别用 2014-11-01 到 2015-06-01 之间以月为单位的数据作为训练数据集，以 2015-07-06 至 2015-07-19 时间跨度两周的数据作为测试数据，对模型进行评测，验证结果见表 9。

表 9 训练数据时效性验证结果

时间	准确率/%	召回率/%
2014/11/1	88.24	16.32
2014/12/1	92.82	16.11
2015/1/1	93.06	15.36
2015/2/1	94.30	69.73
2015/3/1	96.50	58.15
2015/4/1	94.49	77.27
2015/5/1	93.42	79.50
2015/6/1	94.80	85.48

从表 9 可以看出：①准确率基本保持不变，都在可以接受的范围内。②召回率随测试数据和训练数据在时间上相隔越远而越低。③2015-04-01 之后的模型都是可以接受的，可以认为该模型的保质期有 3 个月。

（2）模型随时间衰减

用 2014-07-01 至 2015-02-01 作为训练数据训练模型，2015-02-01 之后的数据按周切割作为测试数据，验证结果如图 2 所示。

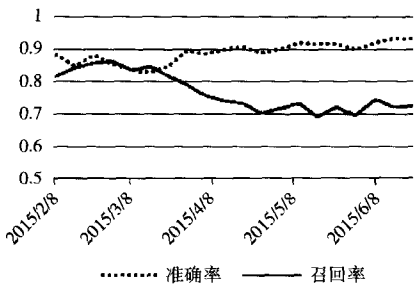

图 2 时效性验证结果

从图 2 中可以看出，准确率基本不会随着时间的推移而改变。召回率会随着时间的推移，慢慢降低。因此可以认为模型再学习的时间长度在较大程度上取决于对召回率的要求。

（3）冷启动

使用 2015-07-06 至 2015-07-19 之间时长两周的数据作为测试数据，从 2015-05-01 起，分别选取 1 天～30 天的数据作为训练数据（该产品线数据增长级数为 1 千条/天），来考验情感分析模型在冷启动产品上的表现。结果见表 10。

表 10 冷启动验证结果

准确率	召回率	F1 值
0.906	0.736	0.620
0.903	0.868	0.835
0.928	0.872	0.822
0.944	0.872	0.730
0.936	0.852	0.782
0.943	0.847	0.768
0.926	0.879	0.837
0.927	0.876	0.830
0.914	0.885	0.830

从表 10 可以看出：准确率和召回率均会随着数据量的增大而提升，而此产品线中，2 日的数据量基本可以完成要求。

4 结束语

在情感分析中，相同词汇在不同产品中，可以表达不同或者完全相反的情感。同时，文本短小，情感集中，不同的句子可以表达相同或类似的情感类型。本文针对以上

两个问题进行分析和研究，对不同产品分别构建了情感词库，同时基于最大熵，增加 3 类话题模型。首先该方法适用于二分类、三分类问题。其次，该方法具有鲁棒性，可以应对冷启动和时间衰减。最后，通过相关实验，结果表明了该方法的有效性、鲁棒性，并且具有一定的对抗时间衰减和冷启动的能力。

参考文献：

[1] LIU Zhiming, LIU Lu. Empirical study of sentiment classification for Chinese microblog based on machine learning [J]. Computer Engineering and Applications, 2012, 48 (1): 1-4 (in Chinese). [刘志明，刘鲁. 基于机器学习的中文微博情感分类实证研究 [J]. 计算机工程与应用，2012，48 (1)：1-4.]

[2] WANG Zhengzhong, ZHANG Hongyuan. Research of sentiment analysis on English blog text [J]. Computer Technology and Development, 2011, 21 (8): 153-156 (in Chinese). [汪正中，张洪渊. 基于英文博客文本的情感分析 [J]. 计算机技术与发展，2011，21 (8)：153-156.]

[3] ZHOU Shengchen, QU Wenting, SHI Yingzi, et al. Overview on sentiment analysis of Chinese microblogging [J]. Computer Applications and Software, 2013, 30 (3): 161-164 (in Chinese). [周胜臣，瞿文婷，石英子，等. 中文微博情感分析研究综述 [J]. 计算机应用与软件，2013，30 (3)：161-164.]

[4] ZHOU Yongmei, YANG Aimin, YANG Jianeng. Construction method of sentiment lexicon for news reviews [J]. Computer Science, 2014, 41 (8): 67-69 (in Chinese). [周咏梅，阳爱民，杨佳能. 一种新闻评论请词典的构建方法 [J]. 计算机科学，2014，41 (8)：67-69.]

[5] ZHANG Lin, QIAN Guanqun, FAN Weiguo, et al. Sentiment analysis based on light reviews [J]. Journal of Software, 2014, 25 (12): 2790-2807 (in Chinese). [张林，钱冠群，樊卫国，等. 轻型评论的情感分析研究 [J]. 软件学报，2014，25 (12)：2790-2807.]

[6] Chang CC, Lin CJ. LIBSVM: A library for support vector machines [J]. ACM Transactions on Intelligent Systems & Technology, 2011, 2 (3): 27-65.

[7] Kim KI, Jung K, Park SH, et al. Support vector machines for texture classification [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 24 (11): 1512-1550.

[8] Wicentowski Richard, Sydes Matthew R. Emotion detection in suicide notes using maximum entropy classification [J]. Biomedical Informatics Insight, 2012, 5 (1): 51-60.

[9] FAN Na, CAI Wandong, ZHAO Yu. Extraction of subjective relation in opinion sentences based on maximum entropy model [J]. Computer Engineering, 2010, 36 (2): 4-6 (in Chinese). [樊娜，蔡皖东，赵煜. 基于最大熵模型的观点句主观关系提取 [J]. 计算机工程，2010，36 (2)：4-6.]

[10] REN Yuan. Sentiment analysis of Chinese microblog using topic self-adaptation [J]. Computer Science, 2013, 40 (1): 231-235 (in Chinese). [任远. 基于话题自适应的中文微博情感分析 [J]. 计算机科学，2013，40 (1)：231-235.]

[11] Galligan MC, Saldoca R, Campbell MP, et al. Greedy feature selection for glycan chromatography data with the generalized dirichlet distribution [J]. BMC Bioinformatics, 2014, 14 (1): 155.

[12] SONG Weiran. Research on feature selection and weighting method for Chinese text classification [D]. Beijing: Beijing University of Technology, 2013 (in Chinese). [宋惟然. 中文文本分类中的特征选择和权重计算方法研究 [D]. 北京：北京工业大学，2013.]

(上接第 126 页)

[14] ZHU Dehai. Point cloud library learning tutorial [M]. Beijing: Beijing University of Aeronautics and Astronautics Press, 2012 (in Chinese). [朱德海. 点云库 PCL 学习教程 [M]. 北京：北京航空航天大学出版社，2012.]

[15] WEI Yingzi, LIU Xiaoli. Robust point cloud plane fitting based on the random sample consensus [J]. Journal of Beijing University of Technology, 2014, 40 (3): 400-403 (in Chinese). [魏英姿，刘晓莉. 基于随机抽取一致性的稳健点云平面拟合 [J]. 北京工业大学学报，2014，40 (3)：400-403.]

[16] WANG Li, LI Guangyun, ZHANG Qifu, et al. Plane fitting and transformation in laser scanning [J]. Journal of Geomatics Science and Technology, 2012, 29 (2): 101-108 (in Chinese). [王力，李广云，张启福，等. 激光扫描中平面拟合及坐标转换模型构建 [J]. 测绘科学技术学报，2012，29 (2)：101-108.]

[17] YE Minlv, HUA Xianghong, CHEN Xijiang, et al. Research on method for the denoising of point cloud based on orthogonal TLS fitting [J]. Bulletin of Surveying and Mapping, 2013 (11): 37-39 (in Chinese). [叶珉昌，花向红，陈西江，等. 基于整体最小二乘平面拟合的点云数据去噪方法研究 [J]. 测绘通报，2013 (11)：37-39.]