文章编号:0253-2778(2016)10-0874-09

基于网格聚类的情感分析研究

缪裕青[1,2],高 韩[1],刘同来[1,2],文益民[1,2]

(1.桂林电子科技大学计算机与信息安全学院,广西桂林 541004;2.广西可信软件重点实验室,广西桂林 541004)

摘要:传统基于语义词典和基于机器学习的中文情感分析方法,其情感分析结果受人的主观因素影响较大,在一定程度上依赖于人工建立的词典,词典的可扩展性不强.本文对于不被包括在知网情感词典中但又含有一定情感倾向的词语,使用点互信息PMI算法、设置参数阈值等方法,进行自动识别、提取和分类,从而达到扩充词典的目的.在此基础上,建立商品评论的特征向量模型,提出情感分类算法SCG,通过网格聚类算法建立分类模型,在网格聚类过程中引入动态衰减因子,周期性地移除稀疏网格,减少计算量.实验结果表明,相比Naive Bayes,SMO(sequential minimal optimization)等分类算法,SCG算法具有更高的准确率和领域适应性.

关键词:情感分析;网格;聚类;点互信息;分类

中图分类号:TP391 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2016.10.012

引用格式:缪裕青,高韩,刘同来,等.基于网格聚类的情感分析研究[J].中国科学技术大学学报,2016,46(10):874-882.

MIAO Yuqing, GAO Han, LIU Tonglai, et al. Sentiment analysis based on grid clustering[J]. Journal of University of Science and Technology of China, 2016,46(10):874-882.

Sentiment analysis based on grid clustering

MIAO Yuqing[1,2], GAO Han[1], LIU Tonglai[1,2], WEN Yimin[1,2]

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

Abstract: To expand a lexicon, the methods of point mutual information (PMI), setting the threshold parameter, etc. were used to automatically identify, extract and classification the words which are not included in the HowNet but have a certain emotional tendency. On that basis, a feature vector model based on commodity comments was established, and the SCG (sentiment classification based on grid clustering) algorithm was presented. Next, the grid-based clustering algorithm was used to build up a classification model. The amount of calculation decreased after the dynamic attenuation factors were introduced and sparse grids were periodically removed in the grid-based clustering process. Experimental results indicate that the classification accuracy and field adaptability of SCG is higher, compared with other algorithms such as Naive Bayes, SMO (sequential minimal optimization).

Key words: sentiment analysis; grid; cluster; point mutual information (PMI); classification

收稿日期:2016-03-09;**修回日期**:2016-09-16

基金项目:广西自然科学基金(2014GXNSFAA118395),国家自然科学基金(61363029),桂林电子科技大学研究生教育创新计划(GDYCSZ201466)资助.

作者简介:缪裕青(通讯作者),女,1966年生,博士/副教授.研究方向:数据挖掘、云计算、并行与分布式计算.
E-mail:miaoyuqing@guet.edu.cn

0 引言

情感分析又称意见挖掘，指对含有一定观点、喜好、情感等主观性的文本进行分析与挖掘，从而获取有用的知识和信息[1]. 它能够对说话者的隐含态度、情绪信息以及意见进行判断或者评估. 情感分析代替人工从大量评论信息中获取有用的褒贬倾向信息，为用户节省了大量时间. 同时，情感分析在互联网舆情监控、个性化推荐、企业营销、用户导购、异常或突发事件检测等领域具有广泛的应用[2].

目前中文情感分析方法主要分为基于语义词典的方法和基于机器学习的方法[3]. 在基于语义词典的方法中，评论信息的情感表达与评论领域息息相关，在不同领域，其情感表达的方式各不相同，甚至截然相反. 有的评论语句需要依据更多的上下文关系才能够了解整个句子或者段落的情感倾向. 基于机器学习的方法需要做大量繁琐的人工标注工作，同时分类算法效果与数据集的领域有关，同一分类算法在不同领域的数据集上产生的分类效果不同.

中文情感分析的本质是利用分类方法将文本的情感分为褒义和贬义两类，或者分为褒义、贬义和中性三类. 分类是一种监督学习，事先知道有哪些类别，对于未分类的数据，根据分类模型，将它们划分到相应的类别. 与分类相比，聚类是一种无监督学习，事先并不知道有哪些类别，通过聚类算法将数据划分为性质相同或相近的子集，即类. 在同一类别中，个体之间的距离较小，而不同类别的个体之间的距离较大. 网格算法是一种聚类算法，该方法首先将数据空间划分为有限个单元的网格结构，所有处理都以单元为单位. 网格数量独立于数据对象的数量，与目标数据集中记录的个数无关，同时对数据的输入顺序不敏感，处理速度独立于数据集的大小，只与数据空间划分的单元个数有关，可伸缩性强、处理速度快.

情感分析研究开始于 1997 年[4]. 基于语义词典的情感分析方法利用已标注情感倾向的词典来判断语义相同或相似词语的情感倾向. 在这类研究中 Turney 等[5] 提出点互信息（point mutual information，PMI）算法，通过计算词语分别与“excellent”和“poor”之间的相互信息来判断该词语是否是极性词语. 在后期研究中有不少针对 PMI 算法的改进. 朱嫣岚等[6]提出了基准词表构建方法，选取词频最高的前 40 对褒贬基准词建立词典. 杨昱昊等[7]针对朱嫣岚论文中褒贬词语倾向性计算正确率不平衡问题，对点互信息算法公式进行了改进，加入 α，β 调节参数，平衡褒义词和贬义词判断的差异，并更改褒贬基准词构建方法，在保留原有词典的前 30 对褒贬基准词情况下，将语料中词频最高的前 10 对褒贬词加入词典中，实验结果显示准确率达到 98.94%.

基于机器学习的方法可以跨领域应用而且稳定性较好，对于数据量较小、标注完整的数据集具有较好的效果. 这类研究中主要有朴素贝叶斯方法（Native Bayes，NB）、最大熵方法（maximum entropy，ME）和支持向量机（support vector machines，SVM）等方法. 徐军等[8]分别利用朴素贝叶斯方法和最大熵方法进行新闻分类研究，将新闻文本分为正面情感类和负面情感类，并采用词频和二值作为特征项权重，实验结果显示准确率可达 90%以上. 李寿山等[9]综合使用朴素贝叶斯方法、最大熵方法、支持向量机方法以及随机梯度下降现行分类方法进行文本情感分类，结果显示组合后的分类结果优于单个方法的分类结果. 王磊等[10]在隐含狄利克雷模型基础上提出一个基于主题的情感向量空间模型，将文本的潜在主题特征融入情感模型中，结合情感词典，利用多标签分类算法，对文本中句子的情感极性进行分析与研究. Pang 等在文献[11-12]中采用不同的特征选择方法，综合使用朴素贝叶斯、最大熵、支持向量机等方法对评论进行分类，实验取得较好效果.

本文针对传统情感分析方法的局限性，建立了一个情感词语自动提取分类模型，对知网情感词典以外含有一定情感倾向的词语自动识别、提取和分类，实现情感词典的自动扩充. 同时，建立评论语句的特征向量模型，提出评论语句特征提取方法. 在网格聚类理论基础上提出情感分类算法 SCG（sentiment classification based on grid clustering），通过网络聚类建立分类模型. 在网格聚类过程中引入动态衰减因子，周期性地移除稀疏网格，减少计算量，从而最终提高了分类的准确率.

1 词典自动扩充

随着网络信息交互平台的不断发展，网络中汉语词语的更新速度越来越快，大量的新词不断地涌现出来，如“给力”、“神马”、“高富帅”、“白富美”等. 对于基于语义词典的情感分析方法，由于网络词语

的随意性，词典的更新跟不上网络词语改变的节奏，使得词典中缺少大量网络情感词语.

为了解决以上问题，本文在基于语句语法分析和词性标注的方法基础上实现了知网情感词典以外含有一定情感倾向词语的自动识别、提取和分类，使得不管网络词语如何快速变化，都能够保证情感词典的实时更新.

（Ⅰ）情感词语提取

首先通过 IKAnalyzer2012 中文分词工具包对评论语句进行分词，同时使用隐马尔科夫模型进行词性标注. 然后，从评论语句中提取出知网情感词典以外可能含有一定情感倾向的且词性已经标注的词语，如形容词、状态词、语气词、形名词、叹词等.

（Ⅱ）情感倾向判别模型

为进一步确定提取出的词语是否含有情感以及褒贬倾向信息，使用点互信息 PMI 算法，确定词语的情感倾向.

$$\mathrm{PMI}(\mathrm{word}_1,\mathrm{word}_2)=\log_2\left(\frac{\mathrm{P}(\mathrm{word}_1,\mathrm{word}_2)}{\mathrm{P}(\mathrm{word}_1)\mathrm{P}(\mathrm{word}_2)}\right) \tag{1}$$

式中，$\mathrm{P}(\mathrm{word}_1)$和 $\mathrm{P}(\mathrm{word}_2)$分别表示词 word_1 和词 word_2 在语料库中独立出现的概率，$\mathrm{P}(\mathrm{word}_1,\mathrm{word}_2)$表示词 word_1 和词 word_2 在语料库中同时出现的概率. 所得到的值 $\mathrm{PMI}(\mathrm{word}_1,\mathrm{word}_2)$ 代表词 word_1 和词 word_2 之间的相似程度，$|\mathrm{PMI}(\mathrm{word}_1,\mathrm{word}_2)|$ 越小，代表词 word_1 和词 word_2 之间的相似程度越大.

假设共有 n 对基准词，褒义基准词集 $\mathrm{Wordset_g}=\{\mathrm{C}_1,\mathrm{C}_2,\mathrm{C}_3,\cdots,\mathrm{C}_n\}$，贬义基准词集 $\mathrm{Wordset_n}=\{\mathrm{D}_1,\mathrm{D}_2,\mathrm{D}_3,\cdots,\mathrm{D}_n\}$，则对于待求词 W，基于点互信息 PMI 算法的语义倾向 SO_PMI(W)为

$$\mathrm{SO\_PMI(W)}=\sum_{i=1}^{n}\mathrm{PMI}(\mathrm{W},\mathrm{C}_i)-\sum_{i=1}^{n}\mathrm{PMI}(\mathrm{W},\mathrm{D}_i) \tag{2}$$

若 SO_PMI(W)＞0，则表示待求词 W 为褒义词；若 SO_PMI(W)＜0，则表示待求词 W 为贬义词.

对于 1 个不含有情感倾向的词语 W^* 来说，可能存在这样的现象：$\mathrm{SO\_PMI}(\mathrm{W}^*)=0$ 或者 $|\mathrm{SO\_PMI}(\mathrm{W}^*)|<\delta$，其中 δ 是一个可调节参数且 $\delta>0$. 对于第 1 种情况，直接可以认为词语 W^* 不含有情感倾向；对于第 2 种情况，$\mathrm{SO\_PMI}(\mathrm{W}^*)\neq0$，且 $\mathrm{SO\_PMI}(\mathrm{W}^*)\in(-\delta,\delta)$，则需要确定参数 δ 的大小，才能认定词语 W^* 是否含有情感倾向.

将能够确定具有情感倾向的词语分别存入褒、贬词典中，没有情感倾向的词语则存入到另一个词典中，以防再次提取出相同的词语，从而节省自动提取的时间.

2 特征向量模型

本文是针对网络商品评论的研究. 为了表示每条评论的特征，创建 1 个 7 维特征向量 $\boldsymbol{\alpha}$：

$$\boldsymbol{\alpha}=(u,u_j,v,v_j,x,y,z) \tag{3}$$

式中，u 表示褒义词数；u_j 表示褒义极性强度；v 表示贬义词数；v_j 表示贬义极性强度；x 表示评论时间；y 表示评论地点；z 表示虚假度.

（Ⅰ）褒义词数和贬义词数

褒义词数和贬义词数分别表示评论语句中褒义词和贬义词的用词量，一定程度上反映出评论者的情绪，同时也反映出评论语句的褒贬特征. 因此，在特征向量中加入了褒义词数和贬义词数因子.

（Ⅱ）褒义极性强度和贬义极性强度

褒义极性强度和贬义极性强度分别表示评论语句中褒义词语和贬义词语的情感强度. 褒义词语的情感极性值为正，贬义词语的情感极性值为负，极性值的绝对值越大，则情感强度越强. 它们反映出整条评论语句的情感强度大小，是辅助评论语句情感倾向判断不可缺少的特征.

（Ⅲ）评论时间

Golder 等[13]考察了来自 84 个国家的 240 万个 Twitter 用户，分析每人的 400 多个帖子，并分析他们使用的词语，积极的如 awesome，消极的如 annoy 等. 如图 1 中所示，不同文化背景的人都有相似的日常情绪节律. 以西海岸（West Coast）数据线为例，早晨起来（05:00），人们的情绪慢慢高涨，在 9 点时达到最高峰，然后情绪渐渐消沉，晚上睡觉前（17:00）又高涨一直到达凌晨（01:00）. 因此，时间对人的情

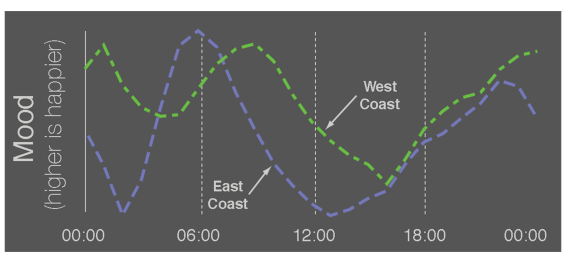

图 1 情绪随时间变化[13]

Fig. 1 Mood changes over time[13]

感是有影响的.

(Ⅳ) 评论地点与虚假度

考虑到评论者的情绪往往会受到环境的影响,所以,在特征向量中加入了评论地点因子. 此外,在商品评论中也掺杂了不少虚假评论,这些评论可能会影响情感倾向判定的结果.

3　基于网格聚类的分类模型

3.1　网格聚类理论

2007 年 Chen[14] 提出一种基于网格的增量式数据流聚类 D-stream 算法,把整个数据空间划分为边长为 Len 的立体子空间. 算法设定 1 个密度阈值,根据网格单元密度的稀疏程度将网格单元分为稠密、稀疏以及过渡 3 种形式. 以稠密网格单元为中心,形成聚类.

定义 3.1　假设输入数据的维度为 d,那么整个数据空间定义为

$$S = S_1 \times S_2 \times \cdots \times S_d \tag{4}$$

式中,S_i 表示第 i 维空间,$i=1,2,\cdots,d$. S_i 又可以按长度 Len 等量划分为 p_i 个部分,于是:

$$S_i = S_{i,1} \cup S_{i,2} \cup \cdots \cup S_{i,p_i} \tag{5}$$

因此,数据空间 S 被划分为 M 个密度网格.

$$M = \prod_{i=1}^{d} p_i \tag{6}$$

于是,1 个密度网格 g 可以表示为 $S_{1,j_1} \times S_{2,j_2} \times \cdots \times S_{d,j_d}$,$j_i=1,\cdots,p_i$,记为

$$g = (j_1, j_2, \cdots, j_d) \tag{7}$$

对于 1 个新到达的数据记录 $x=(x_1,x_2,\cdots,x_d)$,映射到数据空间的网格为 $g(x)$,且

$$g(x) = (j_1, j_2, \cdots, j_d) \tag{8}$$

式中,$j_i=x_i/\text{Len}$.

定义 3.2　在每个网格中均使用 1 个特征向量(Position,D,Label)记录该网格的信息,Position 表示网格在数据空间中的位置即 $g(x)$,D 表示网格的密度,Label 表示网格的类标识.

假设网格的划分长度为 Len,数据记录 x 映射到网格 g 中,那么网格 g 的特征向量初始化方法为:

```
Initial_gird(g){
    Position=x/Len;
    D=1;
    Label=NULL;
}
```

定义 3.3　对于网格 g,当 1 个在 t 时刻新到达的数据记录落入到该网格时,其密度更新方式为

$$D(g,t) = D(g,t-1) + 1 \tag{9}$$

定义 3.4　对于每个网格 $g(x)$,它的密度每隔 1 个单位时间窗口衰减 1 次,也就是说,每处理完 N 个数据,所有网格的密度都减少 θ,衰减方式为

$$D = D - \theta \tag{10}$$

定义 3.5　假设 1 个单位时间窗口(时间窗口大小为 N),数据分别落入到 n 个网格中,那么网格的平均密度为

$$\overline{D} = N/n \tag{11}$$

定义 3.6　假设 1 个网格 g 在 t 时刻接收到 1 个新的数据记录,如果有

$$D(g,t) \geqslant \alpha\overline{D} = Dm \tag{12}$$

那么网格 g 是 1 个稠密网格,在此刻形成 1 个新微簇. 式中,$\alpha>1$.

如果有

$$D(g,t) \leqslant \beta\overline{D} = Dl \tag{13}$$

那么网格 g 是 1 个稀疏网格. 式中,$0<\beta<1$.

对于 1 个网格 g,如果有

$$\beta\overline{D} < D(g,t) < \alpha\overline{D} \tag{14}$$

那么网格 g 是 1 个过渡网格.

定义 3.7　对于两个稠密网格 $g(x)=(x_1,x_2,\cdots,x_d)$ 和 $g(y)=(y_1,y_2,\cdots,y_d)$,如果存在 k,$1\leqslant k\leqslant d$,使得:

① $x_i=y_i$,$i=1,\cdots,k-1,k+1,\cdots,d$

② $|x_k-y_k|=1$

则表示 $g(x)$ 和 $g(y)$ 在第 k 维是相互连通的近邻网格,记为 $g(x)\sim g(y)$. 这样,1 个稠密网格代表 1 个微簇,所有近邻微簇组成 1 个聚簇.

3.2　SCG 算法

本文在网格聚类理论基础上提出情感分类算法 SCG. 这里先描述算法的基本思想.

对于网格密度衰减,关键问题之一是确定衰减因子 θ 的大小. θ 不能设置得太大也不能太小. 如果 θ 太大,网格密度衰减会很快,网格就不能很好地形成簇;如果 θ 太小,会导致历史数据积累,占用大量网格,增加空间复杂度. θ 的设置无论太大还是太小都会影响聚类.

θ 值的设置方法为,1 个稀疏网格只要其密度增长速度大于衰减速度,那么它就有机会成为 1 个稠密网格,并且所需的单位时间和平均每个稠密网格形成的单位时间是差不多的.

设平均每个稠密网格形成的时间为 t，当前数据空间中稠密网格的数量为 N_g，已经运行了 N_t 个单位时间，那么有 $t=N_t/N_g$（若 $N_g=0, t=0$），一般认为 $t\geqslant 1$. 于是有

$$Dl+\theta t\leqslant Dm \tag{15}$$

式中，$Dl=\beta\overline{D}$，$Dm=\alpha\overline{D}$，$\overline{D}$ 为网格的平均密度. 所以得到

$$\theta\leqslant(\alpha-\beta)\overline{D}/t \tag{16}$$

α,β 为阈值参数，θ 的变化实际只与 t 有关. 衰减因子 θ 控制着稠密网格的形成速度，当稠密网格形成得太快时，t 减少，θ 随之增大，抑制网格的形成；当网格形成太慢时，t 增大，θ 随之变小，使新的稠密网格出现. 另外，网格平均密度 $\overline{D}$ 是某个单位时间窗口网格的平均密度，它是不断变化的，能够通过上一次的平均密度去指导下一次聚类.

理想中的数据集大部分数据都会映射到少量网格上面，绝大多数的网格都不会接受到任何数据，因此可以给非空网格动态分配内存，而不需要浪费大量的内存保存全部的网格. 然而，现实中的数据集不可能完美，会导致许多网格中含有孤立点，非空网格数量不断增加，并且这些网格中包含的数据量长期不增加而无法形成簇，甚至 1 个孤立数据点就占据 1 个网格. 若数据记录是无限的，孤立点经过长期积累必然占据大量网格，占用愈来愈多的内存资源，同时降低算法的计算效率. 所以，在网格密度周期性衰减前，会删除密度小于 Dl 的稀疏网格.

SCG 算法：

```
1. Procedure
2. 初始化一个特征向量为(Position,D,Label)的网格链表
   grid_list,初始化网格划分长度 Len;
3. While(有数据到达)
4.   For(对于每 N 个数据)
5.     读取数据记录 x=(x1,x2,…,xd);
6.     将 x 映射到相应密度网格 g 中;
7.     If (网格 g 不在 grid_list 中)
8.       将网格 g 插入到 grid_list 当中,初始化 g 的特征
         向量;
9.     Else
10.      更新网格 g 的密度 D(g),D(g)+=1;
11.      If (网格 g 是稠密网格 &&. g 未分类)
12.        使用邻接网格判别方法,分配类标
             Identify(g){
               For(对 grid_list 中的每个网格 h){
                 If (g~h&&Label(g)==NULL)
                   则 g 为旧类别,Label(g)=Label(h);
                 If (g~h&&Label(g)! =NULL)
                   Label(h) = Label(g);}
               If( Label(g) ==NULL)
                 g 为新类别,Label(g)=new Cluster_ID;
             }
13.    End For
14.    For(对 grid_list 中的每个网格 grid)
15.      If (网格 grid 是稀疏网格)
16.        将 grid 从 grid_list 中删除;
17.      Else
18.        动态衰减过渡网格 grid 的密度 D(grid);
19.    End For
20. End While
21. For(对网格中的每个聚簇 c)
      If (Count(褒)>Count(贬))
        Label(c)=褒;
      Else
        Label(c)=贬;
    End For
22. End Procedure
```

4 实验结果与分析

4.1 实验数据

（Ⅰ）实验基础语料

① 基础情感词典为知网情感词典[15]，褒义词 4 566 个、贬义词 4 370 个.

② 情感词极性表为清华大学于 2012 年公布的词表[16]，其中包含 23 419 个汉语情感词的情感极性值.

③ 词性标注语料库使用 1998 年 1 月份《人民日报》语料[17]，共计 200 万字，46 个标记集.

④ 实验语料使用真实网络衣服购买评论，2 万条褒义评论和 2 万条贬义评论作为训练集，1 万条褒义评论和 1 万条贬义评论作为测试集.

⑤ 新增否定词表，用来进行逻辑判断. 当出现否定词后，将与之最近的情感词的情感倾向反转，若为褒义则反转为贬义，若为贬义则反转为褒义. 否定词表如表 1 所示.

表 1 否定词表

Tab. 1 Privative word list

不	不要	不让	没	没有	木有
没什么	不怎么	不是	还没	还未	没啥
极不	不能	买不到	无	几乎没有	

（Ⅱ）PMI 算法种子词集

词典扩充过程中，计算语义倾向度必须选择 1 组褒贬基准词对，分析待测词语与这组基准词的语义关系紧密程度，计算语义倾向值，并以此来判断待测词语的情感倾向. 因此，所选的基准词必须是具有代表性的含有强烈褒贬倾向的词语. 基于以上原因，建立 PMI 算法种子词集 $Wordset_g$ 和 $Wordset_n$，$Wordset_g$是由词性标注语料库中词频最高的前 30 个褒义词和待评语料中词频最高的前 10 个褒义词组成. $Wordset_n$是由词性标注语料库中词频最高的前 30 个贬义词和待评语料中词频最高的前 10 个贬义词组成. 种子词集如表 2 所示.

表 2　种子词集

Tab. 2　Seed word set

褒义种子词									
健康	超级	英雄	稳定	最好	高手	漂亮	开通	优质	不错
安全	保险	精选	优秀	最佳	文明	完美	真实	欢乐	出色
良	卫生	快乐	高级	幸福	积极	简单	先进	美好	成熟
很好	舒服	满意	喜欢	合适	好看	好评	合身	值得	舒适
贬义种子词									
伤心	事故	麻烦	色情	落后	自负	野蛮	无情	虚假	劣质
黑客	非法	不良	暴力	漏洞	不安	陷阱	失误	残酷	愚人
莠	失败	病人	黄色	有害	魔鬼	不当	淫秽	变态	恶劣
差	垃圾	不好	一般	不值	掉色	太大	骗子	失望	色差

4.2　评论语句特征提取方法

（Ⅰ）褒义词数和贬义词数

褒义词数和贬义词数特征提取使用基于语义词典的中文情感分析方法，分别提取出评论语句中的褒义词个数和贬义词个数，以作为该条评论语句的两维特征.

（Ⅱ）褒义极性强度和贬义极性强度

在情感词极性词表的基础上，分别提取出评论语句中褒义词和贬义词的最大褒义极性值和最小贬义极性值，以此作为该条评论语句的褒义极性强度和贬义极性强度特征.

（Ⅲ）评论时间

根据 Golder 的研究结果，以西海岸（West Coast）的日常情绪变化趋势为参考依据，将商品评论发表时间分为两类：03:00～06:00 和 11:00～19:00；06:00～11:00 和 19:00～03:00. 处在该两类时段的评论时间分别标记为 1 和 2，前者表示评论者含有的贬义情绪大于褒义情绪，后者表示评论者含有的褒义情绪大于贬义情绪.

（Ⅳ）评论地点与虚假度

本文在评论语句特征向量中提出包含评论地点和虚假度两维特征，但在实验中未提取，因此，将每条评论的评论地点和虚假度特征均设置为 1.

4.3　情感词语自动提取分类

在情感倾向判别模型中，为了确定参数 δ 的大小，现对 3 万条贬义评论进行如下实验：

已知经过词性标注后，提取出来的词语中含有褒义词 96 个、贬义词 566 个、非情感词 8 996 个. 然后通过设定不同参数 δ 的大小，重复实验，分别得到

表 3　不同参数 δ 下的结果

Tab. 3　Results based on different values of δ

δ	褒义词识别			贬义词识别			非情感词识别		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
1	4.69%	52.08%	8.60%	6.99%	68.90%	12.68%	94.68%	31.66%	47.45%
2	5.36%	51.04%	9.70%	7.02%	66.96%	12.71%	94.41%	35.14%	51.22%
3	6.24%	47.92%	11.04%	7.12%	64.31%	12.82%	94.30%	39.92%	56.09%
4	7.24%	43.75%	12.43%	7.29%	60.07%	13.00%	94.17%	46.19%	61.98%
5	7.98%	39.58%	13.29%	7.65%	54.59%	13.42%	94.19%	53.84%	68.51%
6	8.73%	34.38%	13.92%	8.02%	51.24%	13.88%	94.18%	59.32%	72.79%
7	9.63%	32.29%	14.83%	8.26%	46.64%	14.04%	94.10%	64.23%	76.35%
8	10.82%	30.21%	15.93%	8.77%	44.35%	14.64%	94.13%	68.31%	79.17%
9	12.72%	30.21%	17.90%	9.25%	42.23%	15.17%	94.21%	71.69%	81.42%
10	**14.89%**	**29.17%**	**19.72%**	**9.53%**	**39.22%**	**15.34%**	**94.16%**	**74.74%**	**83.34%**
11	16.00%	25.00%	19.51%	9.34%	34.28%	14.68%	93.96%	77.61%	85.01%
12	16.54%	22.92%	19.21%	9.11%	30.39%	14.02%	93.81%	79.64%	86.14%

表 4 不同新增词语个数下的特征提取时间

Tab. 4 Feature extraction time based on different numbers of new words

贬义训练数据	0	0.5 万	1 万	1.5 万	2 万	2.5 万	3 万
新增褒义词数	0	28	67	85	85	91	96
新增贬义词数	0	240	383	440	482	519	566
新增非情感词数	0	2 237	4 270	5 826	6 973	8 081	8 996
3 万条褒义评论语句特征提取时间/ms	8 767 751	5 708 110	4 026 504	3 638 625	3 239 519	2 972 756	2 838 570

准确率 P、召回率 R 和 F 值，结果如表 3 所示，其中，

P=正确识别词语个数/识别词语总数，

R=正确识别词语个数/存在词语总数，

$F=2*P*R/(P+R)$.

从表 3 可以看出，当 $\delta=10$ 时，褒义词识别和贬义词识别均达到最大的 F 值. 本次提出的情感词语自动提取分类模型是一个循环渐进的过程. 在初次识别时，会有大量的非情感词语对实验结果造成干扰，导致表 3 的准确率、召回率和 F 值均很低. 但是此后的识别是以之前的识别结果为基础，所以在时间消耗方面，是一个越来越好的过程. 为了验证该结论，设计如下实验：

将不同数量级的贬义评论语句作为训练数据，分别得到新增的褒义词语、贬义词语和非情感词语个数，然后使用 3 万条褒义评论语句作为测试数据，其特征提取时间如表 4 所示.

从表 4 可以看出，将贬义训练数据中得到的新增褒义词、新增贬义词和新增非情感词作为 3 个词典均加入到 3 万条褒义评论语句的测试中，得到其进行特征提取的时间. 随着训练数据的增加，新增的褒义词、贬义词和非情感词个数也随之增加，但 3 万条褒义评论语句的特征提取时间随之减少. 因此，该实验很好地验证了本文所提出的情感词语自动提取分类模型是一个越来越优化的过程，能够缩减后续评论语句特征提取时间，提高效率.

4.4 SCG 算法动态衰减因子

衰减因子主要控制着稠密网格的形成速度. SCG 算法使用动态衰减因子对网格密度进行衰减. 将动态衰减因子与静态衰减因子、无衰减因子等衰减方法进行比较，以网格聚类中形成的聚簇数作为衡量不同种类衰减因子好坏的标准.

使用不同数量级的褒义评论语句和贬义评论语句特征作为测试数据，更改 SCG 算法中的衰减因子性质，得到其聚簇数分别如图 2 和图 3 所示.

从图 2 和图 3 可以看出，使用动态衰减因子形成的聚簇数均小于使用静态衰减因子和无衰减因子形成的聚簇数. 说明动态衰减因子很好地控制了聚簇的形成速度，使聚类更加集中，而使用其他的衰减因子的聚类结果比较稀疏松散. 使用动态衰减因子，聚簇节点保存的个数比使用其他的衰减因子聚簇节点保存的个数少，降低了内存的占用.

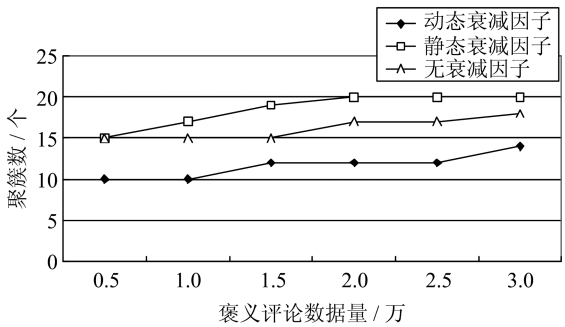

图 2 褒义评论语句聚簇数

Fig. 2 Clustered numbers of commendatory comment statements

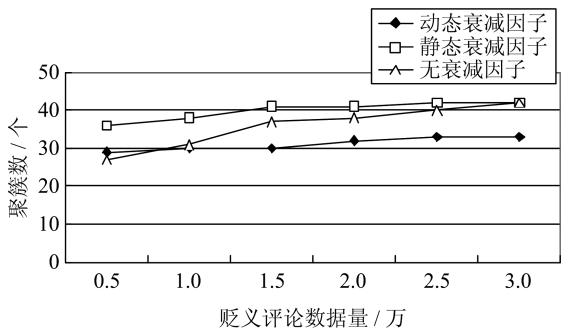

图 3 贬义评论语句聚簇数

Fig. 3 Clustered numbers of derogatory comment statements

4.5 评论语句特征对比实验

为了研究褒义词数、褒义极性强度、贬义词数、贬义极性强度、评论时间等 5 维特征在实验中的影响，去除其中的一个或多个特征，使用 SCG 算法进行准确率判断，实验结果如表 5 所示.

从表 5 可以看出，测试语料缺少 1 维或多维特征的准确率均低于具有 5 维特征时的准确率，并且维数越多，SCG 算法的准确率越高. 该实验充分说

明了本文提出的评论语句特征向量模型的有效性，在多个特征维度的基础上能够提高 SCG 算法的准确率.

表 5　特征对比实验结果

Tab. 5　Results of feature contrast experiment

第 1 维	第 2 维	第 3 维	第 4 维	第 5 维	准确率
√	√	√	√	√	94.34%
√	√	√	√		93.39%
√		√		√	92.81%
	√		√	√	88.25%
	√		√		51.91%

4.6　情感分类实验

为了验证 SCG 算法的性能，将 2 万条褒义评论和 2 万条贬义评论作为训练数据，导入网格进行聚类，再将 1 万条褒义评论和 1 万条贬义评论作为测试数据，导入网格进行分类. 同时选取 Naive Bayes，SMO（sequential minimal optimization），J48 和 Random Forest 等机器学习算法与其进行准确率比较，结果如表 6 所示. 其中，对 Naive Bayes，SMO，J48 和 Random Forest 的实验均使用 weka 数据挖掘平台进行.

表 6　不同算法的准确率(1)

Tab. 6　Accuracy of different algorithms (1)

算法	准确率
SCG	94.340%
Naive Bayes	90.915%
SMO	91.325%
J48	91.010%
Random Forest	91.515%

从表 6 可以看出，SCG 算法的准确率明显超过 Naive Bayes，SMO，J48 和 Random Forest 的准确率，均高出 3 个百分点左右. SCG 算法能够获得更高准确率的原因是：① 在现有情感词典的基础上实现了词典的自动扩充，提取评论语句的褒义强度和贬义强度特征更加准确，更能体现出评论语句的真实特征；② SCG 算法使用网格聚类方法，将数据空间划分为有限个网格单元，研究网格单元中的数据对象，使特征相同或相似的数据落入同一细分的网格，并且依靠邻接网格来形成聚簇，识别任意形状的聚簇，比传统的聚类方法好.

为了进一步研究 SCG 算法在其他领域的作用，现选用 1 个包含 2 000 条褒义评论和 2 000 条贬义评论的计算机语料（缺少评论时间特征），将其中 1 500 条褒义评论和 1 500 条贬义评论作为训练集，余下的 500 条褒义评论和 500 条贬义评论作为测试集. 对比实验结果如表 7 所示.

表 7　不同算法的准确率(2)

Tab. 7　Accuracy of different algorithms (2)

算法	准确率
SCG	93.3%
Naive Bayes	78.8%
SMO	81.5%
J48	75.2%
Random Forest	76.3%

从表 6 和表 7 可以看出，更改测试语料后，SCG 算法依然具有较高的准确率. 测试语料的变动使 Naive Bayes，SMO，J48，Random Forest 4 类分类算法的准确率变动较大. 说明 SCG 算法相比 Naive Bayes，SMO，J48，Random Forest 算法具有更好的领域适应性.

5　结论

对商品评论信息进行褒贬分类能够为商家和消费者提供决策支持. 本文针对传统中文情感分析方法结果依赖人工建立的词典、词典可扩展性不强等局限性，对不被包括在知网情感词典中但又含有一定情感倾向的词语，使用点互信息 PMI 算法、设置参数阈值等方法，进行自动识别、提取和分类，从而达到了扩充词典的目的. 同时，提出了构建商品评论的 7 维特征向量模型，包括褒义词数、褒义极性强度、贬义词数、贬义极性强度、评论时间、评论地点和虚假度. 并在网格聚类理论的基础上提出了情感分类算法 SCG，通过网络聚类建立分类模型，在网格聚类过程中引入动态衰减因子，周期性地移除稀疏网格，从而减少内存占用量和计算量. 实验结果表明，SCG 算法具有更高的分类准确率和领域适应性.

尽管本文对不被包括在知网情感词典中但又含有一定情感倾向的词语，设置了参数阈值来进行自动识别、提取和分类，但还是提取出了大量的非情感词语，这个过程无疑增加了系统的计算量和运行时间. 同时，在评论信息特征向量构建过程中没有充分利用评论地点和虚假度两维特征. 本次实验过程中使用的语料只分为褒义和贬义两类，并未考虑中性评论在 SCG 算法中的影响. 实验过程除了从词语级别进行情感分析研究外，也未从句子及篇章整体角度进行情感分析探讨. 这些内容将在下一步工作中研究.

参考文献(References)

[1] 魏韡，向阳，陈千. 中文文本情感分析综述[J]. 计算机应用，2011，31(12)：3 321-3 323.
WEI Wei，XIANG Yang，CHEN Qian. Survey on Chinese text sentiment analysis [J]. Journal of Computer Applications，2011，31(12)：3 321-3 323.

[2] 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.
ZHOU Shengchen，QU Wenting，SHI Yingzi，et al. Overview on sentiment analysis of Chinese microblogging [J]. Computer Application and Software，2013，30(3)：161-164.

[3] 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)：2 790-2 807.
ZHANG Lin，QIAN Guanqun，FAN Weiguo，et al. Sentiment analysis based on light reviews[J]. Journal of Software，2014，25(12)：2 790-2 807.

[4] HATZIVASSILOGLOU V，MCKEOWN K R. Predicting the semantic orientation of adjectives [C]// Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the ACL. Morristown，NJ，USA：ACL，1997：174-181.

[5] TURNEY P D. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews [C]//Proceedings of 40th Meeting of the Association for Computational Linguistics. Stroudsburg，PA，USA：Association for Computational Linguistics，2002：417-424.

[6] 朱嫣岚，闵锦，周雅倩，等. 基于HowNet的词汇语义倾向计算[J]. 中文信息学报，2006，20(1)：14-20.
ZHU Yanlan，MIN Jin，ZHOU Yaqian，et al. Semantic orientation computing based on HowNet[J]. Journal of Chinese Information Processing，2006，20(1)：14-20.

[7] 杨昱昺，吴贤伟. 改进的基于知网词汇语义褒贬倾向性计算[J]. 计算机工程与应用，2009，45(21)：91-93.
YANG Yubing，WU Xianwei. Improved lexical semantic tendentiousness recognition computing [J]. Computer Engineering and Applications，2009，45(21)：91-93.

[8] 徐军，丁宇新，王晓龙. 使用机器学习方法进行新闻的情感自动分类[J]. 中文信息学报，2007，21(6)：95-100.
XU Jun，DING Yuxin，WANG Xiaolong. Sentiment classification for Chinese news using machine learning methods [J]. Journal of Chinese Information Processing，2007，21(6)：95-100.

[9] 李寿山，李逸薇，黄居仁，等. 基于双语信息和标签传播算法的中文情感词典构建方法[J]. 中文信息学报，2013，27(6)：75-80.
LI Shoushan，LI Yiwei，HUANG Juren，et al. Costruction of Chinese sentiment lexicon using bilingual information and label propagation algorithm [J]. Journal of Chinese Information Processing，2013，27(6)：75-80.

[10] 王磊，苗夺谦，张志飞，等. 基于主题的文本句情感分析[J]. 计算机科学，2014，41(3)：32-35.
WANG Lei，MIAO Duoqian，ZHANG Zhifei，et al. Emotion analysis on text sentences based on topics[J]. Computer Science，2014，41(3)：32-35.

[11] PANG B，LEE L，VAITHYANATHAN S. Thumbs up? Sentiment classification using machine learning techniques [C]// Proceedings of the ACL-02 Conference on Empirical Methods In Natural Language Processing. Stroudsburg，PA，USA：Association for Computational Linguistics，2002：79-86 .

[12] PANG B，LEE L. Seeing stars：Exploiting class relationships for sentiment categorization with respect rating scales [C]// Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Stroudsburg，PA，USA：Association for Computational Linguistics，2005：115-124.

[13] GOLDER S A，MACY M W. Diurnal and seasonal mood vary with work，sleep，and daylength across diverse cultures[J]. Science，2011，333(6051)：1 878-1 881.

[14] CHEN Y，TU L. Density-Based Clustering For Real-Time Stream Data[C]// Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. New York，NY，USA：ACM，2007：133-142.

[15] 董振东，董强. 知网：HowNet Knowledge Database [DB/OL]. [2016-02-10]. http://www.keenage.com.

[16] 汉语情感词极值表[DB/OL]. (2012-07-19)[2016-02-10]. http://www.datatang.com/data/43216.

[17] 北京大学计算语言学研究所. 人民日报切分/标注语料库下载[DB/OL]. [2016-02-10]. http://www.icl.pku.edu.cn/icl_res/.