文章编号：1003-0077(2018)05-0097-08

面向情感聚类的文本相似度计算方法研究

李 欣[1]，李 旸[2]，王素格[2,3]

（1. 山西职工医学院 信息中心，山西 晋中 030619；
2. 山西大学 计算机与信息技术学院，山西 太原 030006；
3. 山西大学 计算智能与中文信息处理教育部重点实验室，山西 太原 030006）

摘 要：在文本情感分析时，使用无监督的聚类方法，可以有效节省人力和数据资源，但同时也面临聚类精度不高的问题。相似性是文本聚类的主要依据，该文从文本相似度计算的角度，针对情感聚类中文本—特征向量的高维和稀疏问题，以及对评论文本潜在情感因素的表示问题，提出一种基于子空间的文本语义相似度计算方法(RESS)。实验结果表明，基于 RESS 的文本相似度计算方法，有效解决了文本向量的高维问题，更好地表达了文本间情感相似性，并获得较好的聚类结果。

关键词：文本情感聚类；文本相似度计算；文本语义子空间

中图分类号：TP391 **文献标识码**：A

Text Similarity Calculation for Text Sentiment Clustering

LI Xin[1], LI Yang[2], WANG Suge[2,3]

(1. Information Center, Shanxi Medical College for Continuing Education, Jinzhong, Shanxi 030619, China;
2. School of Computer and Information Technology, Shanxi University, Taiyuan, Shanxi 030006, China;
3. Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education Shanxi University, Taiyuan, Shanxi 030006, China)

Abstract: In text sentiment analysis, unsupervised clustering method is challenged by low precision. To improve the text similarity measure lying as key to clustering, this paper proposes a semantic subspace (RESS) method to deal with the high dimension and sparseness of sentiment text representation issue. It also helps to caputure the implicit expression of sentiment. The experimental results show that RESS can effectively reduce the feature of data set and generat better results.

Key words: sentiment-based text clustering; text similarity calculation; text semantic subspace

0 引言

随着新兴电子商务平台，微博和微信等社交媒体的广泛使用，人们在享受互联网技术带来便利的同时，也用文字记载了自己的心情、状态、评价和观点。通过挖掘海量微博和评论文本等社会媒体数据，可以获得用户对产品的情感倾向(褒扬或者贬斥)，从而指导企业的决策以及个人的消费行为[1-2]。有监督的机器学习方法需要大量的带标签的文本数据，而无监督的文本聚类方法可以克服这一不足[3]。

目前，聚类方法在文本数据挖掘中发挥了重要作用，情感聚类的相关研究也备受关注[4]。情感聚类常面临三个困难：首先，由于聚类算法的无指导性，使聚类结果总是沿着文本最显著的特点聚簇。而文本一般是按照一定的主题进行组织，因此，情感聚类结果的准确率并不高；其次，由于用户表达的感受和观点等情感蕴含在评论中，其特征表现并不明显。从大量的特征中难以实现情感特征的有效分

收稿日期：2017-03-03 **定稿日期**：2017-05-12

基金项目：国家自然科学基金(61573231，61632011，61672331，61432011)；山西省科技基础条件平台计划项目(2015091001-0102)

离;再次,海量的文本数据也存在文本表示的高维和稀疏问题。为了解决这些问题,一些研究者试图对传统的聚类算法进行改进。希望获得较高的聚类精度,另一些则从特征选择和文本间距离度量等方面展开研究,然而,这些方法过多地依赖人工参与及反馈,且不能有效解决文本向量表示的高维和稀疏问题。

为了充分利用文本的上下文信息,获取文本的语义特征,同时降低文本表示的维度和稀疏性,本文从文本相似度计算角度出发,提出一种基于子空间的文本语义相似度计算方法(RESS)。该方法分别构建文本相关度子空间(RSS)和文本情感子空间(ESS),在此基础上计算文本集的语义相似度,构建相似度矩阵,最后采用聚类算法实现情感聚类。

1 相关研究

近几年,情感分析技术为各行各业及政府提供了重要的信息,体现着不可估量的价值,这些应用推动了情感分析技术的发展。Pang 等[5]人首次使用监督学习方法在电影评论领域做了情感分类研究,2001年,Sanjiv 等[6]人设计出在经济领域进行实时情感挖掘和分析系统,采用分类算法获取股民对股票投资的观点倾向,并分析股票走势对金融市场和股民情绪的影响。随着网络评论文本的指数级增长,采用无监督的聚类方法进行情感分析也备受关注。研究者主要从聚类算法的层面进行改进,如重构文本聚簇[7]、添加约束条件[8]、构建新特征空间[9]、引入反馈机制[10]等。而这些方法过多地依赖人工参与和人类反馈,甚至需要人工阅读大量的评论文本,既耗时又费力。随着研究的深入,人们渐渐发现,对文本情感聚类的研究不能停留在算法层面,情感因素的表示、文本特征选择直接影响着聚类结果,在特征选择、文本距离度量等方面开展了大量的研究。

词特征的选择是文本情感分析的关键步骤,Ellen 等[11]人在情感分析和观点挖掘的任务中使用词、n 元语法、短语和词汇语义规则进行文本表示,并使用词语包容关系所构建的层次结构识别复杂特征和约简冗余特征。实验表明这种特征选择方法可以改善情感分析效果。Feng 等[12]人认为博客文本中的情感倾向性在网络中服从一定的分布。它与已有博文聚类方法不同的是,他们认为对于文本特征表示,挖掘博文中潜在的情感因素比抽取其关键词更重要。文中提出了一种概率潜在语义分析方法,首先为隐含的情感因子建模,然后对文本进行聚类。黄永光等[13]人分析了网络中存在的大量不规范的文本数据。这些不仅长度短,而且语言用词极不规范。针对此类问题他们提出的一种"规范文本——拼音串匹配——搜索聚类"处理流程,很好地提高了变异短文本的聚类性能。文本聚类技术有效性的前提是为文本选取合适的特征。但在短文本中,由于特征的稀疏性,单纯使用统计分析方法存在很多弊端。因此,Makrehchi 等[14]人在《同义词词林》的基础上,考虑语义和统计特性,选择最佳特征,使得聚类性能也得到了提升。传统的文本聚类方法,都是在词汇特征的基础上,加入一些简单的语义信息。如利用 WordNet 获得同义或反义关系,而没有利用任何基于短语的语义分析。Zheng 等[15]人从名词短语的角度挖掘更多的语义信息(上位关系、下位关系、整体部分关系),改进了基于 WordNet 的聚类方法,获得了更好的效果。由于特征集过大使得文本表示维度过高,特征集过小,导致文本表示稀疏,信息表现不完整。Jing 等[16]人提出一种新的基于知识的向量空间模型,这种模型考虑了文档之间的非相似性,与传统的只考虑文档之间的相似性方法相比,提高了文本聚类的性能。王素格等[17]人针对文本情感分类中的数据稀疏问题,提出一种新的文本表示模型。该模型利用模糊粗糙理论对文本属性特征进行离散化处理,对包含情感倾向意义的属性加权。计算属性对于情感类别的隶属度,实现属性特征的压缩,提高情感分类效果。夏云庆等[18]人针对歌词情感分析问题,提出了基于情感单元的情感向量空间模型。该模型能够有效地解决文本表示效率、歧义、情感功能、数据稀疏等方面的不足,提高情感分类的效果。针对微博情感分析问题,刘全超等[19]利用微博内容和转发等特征,构建基于短语路径的微博文本情感倾向性判定方法,提高情感分类性能。

2 基于子空间的文本语义相似度计算

传统的文本表示方法将所有的文本构建在共同的特征空间上。文本集的特征个数作为向量的维度,特征数越多,则每篇文本的信息表现得越完整。但同时增加了向量的维度,提高了计算的复杂度。相反,特征数越少,虽然降低了计算的复杂度,但同时减少了文本向量所包含的信息量。针对情感聚类

中文本一特征向量的高维和稀疏问题，以及对评论文本潜在情感因素的表示问题，本文从子空间角度出发，构建文本集的语义相似度矩阵。

在文本情感聚类中，文本相似度既要充分考虑文本在分布上的相关性，又需要计算文本间的情感相关度。因此，本文分别为数据集构建相关性子空间（RSSV）和情感子空间（ESSV），计算基于相关性和情感相融合的文本语义相似度（RESS），在此基础上进行文本情感聚类。其流程图如图 1 所示。

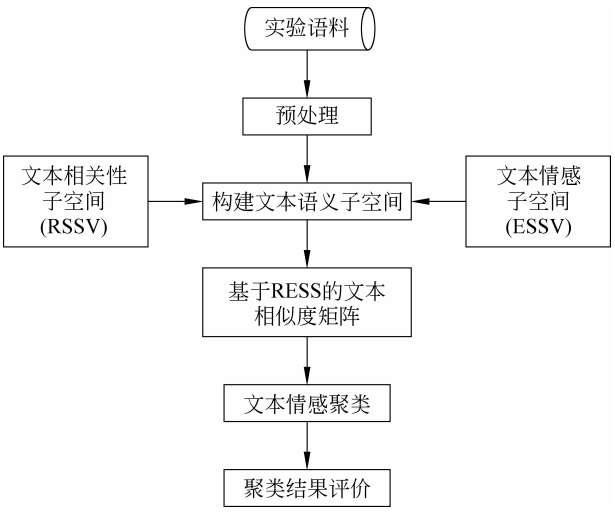

图 1　基于 RESS 的文本情感聚类流程

2.1　词语相关度

文本是由词组成的。在聚类过程中，词语之间的相关度直接影响到文本相似性计算和文本聚类效果。对本文所使用的语料集进行统计和分析发现，当文本数据为 2 000 篇时，词数可达近 20 000 个，这些词语以其复杂的关系构成了不同的表达。因此，词语相关度度量是文本表示和文本聚类研究的前提和基础。

词语相关度反映了词语在语法、语义及语用方面的关联程度。常用的词语相关度计算方法有两种：一种是基于本体知识库（如 HowNet、WordNet、同义词词林、情感词典[20]等）的方法。这种方法把词语作为本体知识网中的节点，通过计算节点之间的距离获得词语之间的相关性。该方法简单、直观，但由于对外部资源的依赖性强，因此在解决多领域和跨领域问题时会表现出一定的局限性；另一种是基于大规模语料库的统计方法。该方法建立在满足以下假设的前提上：凡是语义相近的词，其上下文也相似。对于大规模语料资源，这一假设是成立的。统计两个词语在特定窗口中同时出现的频率，频率越大，其相关性越大。基于统计的方法表面上计算孤立的两个词语之间的关联关系，实际上也利用了词语所在的上下文信息。本文采用基于语料库统计的方法计算词语相关度。

假设 t_i,t_j 是文本集中两个特征词，$rel(t_i,t_j)$ 表示 t_i 与 t_j 的相关度，采用加权对数似然比（WLLR）方法，其公式如式(1)所示。

$$rel(t_i,t_j)=P(t_i\mid t_j)\cdot\lg\frac{P(t_i\mid t_j)}{P(t_i\mid \overline{t_j})} \tag{1}$$

其中，$p(t_i\mid t_j)$ 表示 t_i 与 t_j 共现的频率，$p(t_i\mid \overline{t_j})$ 表示 t_i 与 t_j 不共现的频率。$rel(t_i,t_j)$ 越大，表示特征词 t_i 与 t_j 的相关性越大，也越相似。若 $rel(t_i,t_j)>0$，t_i 与 t_j 呈正相关。否则，成负相关，即语义相关但极性相反。若 $rel(t_i,t_j)=0$，则特征词 t_i 与 t_j 不相关，相似性为 0。

2.2　文本相关性子空间（RSSV）

由于自然语言表达的丰富性，在文本相关性度量中，为了降低向量维度，减少数据的稀疏性，同时利用文本的特征信息，本文将文本向量建立在以任意两篇文本及其特征所构成的子空间上。子空间的维度最大不超过两篇文本中词的个数。在以两篇文本构成的子空间上构建特征向量，不但降低了向量维度，而且能够充分利用文本的词信息。

设评论文本数据集 $X=\{x_1,x_2,\cdots,x_N\}$，x_i 表示第 i 篇评论文本，N 表示评论文本的总数，$T=\{t_1,t_2,\cdots,t_n\}$ 是 X 的原始特征集，n 表示特征个数，假定 $\Theta(X,T)$ 表示文本集的原始特征空间。

对于 $x_i,x_j\in X$，$T_1=\{t_1,t_2,\cdots,t_p\}\subset T$ 表示 x_i,x_j 中所有非停用词构成的特征词集。其中，p 表示特征个数。采用 T_1 为文本 x_i,x_j 构建基于相关度的特征子空间 $\Theta_1((x_i,x_j),T_1)$，$\Theta_1\subset\Theta$。用 $v_i=\{w_{i1},w_{i2},\cdots,w_{ip}\}$ 表示文本 x_i 的特征向量表示。其中，$w_{ik}(k=1,\cdots,p)$ 表示文本 x_i 在特征 t_k 下的权重，其计算的规则如下：

（1）若文本 x_i 中包含特征词 t_k，则 $w_{ik}=1$；

（2）若文本 x_i 中不包含特征词 t_k，$w_{ik}=max_{t\in x_i}rel(t_k,t)$。其中，$rel(t_k,t)$ 表示 t_k 与 x_i 中词 t 的相关度，采用式(1)计算。

与传统的向量空间表示不同，基于 RSSV 的文本表示为数据集中任意两篇文本 x_i,x_j 构建向量子空间，其向量模型见表 1 所示。

表 1 基于 RSSV 的文本向量模型

	t_1	t_2	…	t_k	…	t_p
x_i	w_{i1}	w_{i2}	…	w_{ik}	…	w_{ip}
x_j	w_{j1}	w_{j2}	…	w_{jk}	…	w_{jp}

2.3 文本情感子空间(ESSV)

评论者发表其观点时,常常隐含着其情感倾向和情绪表达。因此,包含倾向的观点词可以用情感向量表示。由于文本是词的集合,对观点词向量进行叠加,可以获得文本的情感向量表示,构建文本情感子空间。

2.3.1 情感特征集

对于产品评论和微博,评论者通常采用情绪词表达个人的观点和情感倾向。例如"*这款三星用着真心不爽!!!* "其中,"*不爽*"是一个表示情绪的贬义词,评论者以此表达对"*这款三星手机*"的差评。由此可见,评论者的心情能够反映其对产品的态度。在语料库中,很多评论文本所持有的观点是通过评论者"喜"、"怒"、"哀"、"乐"的情绪表达的,因此,在情感子空间中,情感特征词应同时考虑到情绪词和观点词。

对于第 k 个文本特征词 $t_k \in T$,用情感特征集 $M=\{M_1, M_2, \cdots, M_{12}\}$ 构建特征词的情感向量。对于不同的数据集,情感特征集 M 的选择也是不同的。

对于英文数据集,采用 Mitral[21] 等人提出的情绪类别划分方式[22-23],使用"anger"、"disgust"、"fear"、"guilt"、"sadness"、"shame"、"interest"、"joy"、"surprise"、"desire"、"love"、"courage"共 12 个基本情绪构成情感特征集 M。

对于中文数据集,在情感类别的划分方面至今还没有统一标准。本文采用林鸿飞[24]的分类方法,在七个基本情感类别("恐惧"、"愤怒"、"厌恶"、"悲伤"、"惊讶"、"高兴"、"喜好")的基础上,参考英文情感类的划分,并对数据集进行统计和分析。在中文数据集的每个领域均增加五个与领域相关的观点词,分别是:保险领域:"烦人"、"可恶"、"缺德"、"失望"、"不错";翡翠领域:"漂亮"、"温润"、"精致"、"圆润"、"均匀";手机领域:"失望"、"郁闷"、"伤心"、"不错"、"爽",将情感类别扩充到 12 类,分别作为每个数据集的情感特征集 M。

2.3.2 文本情感子空间(ESSV)

对于 $x_i \in X$, $T_2=\{t_1, t_2, \cdots, t_q\} \subset T$ 表示 x_i 中所有非停用词构成的特征词集,其中,q 表示特征个数。使用情感特征集 M 为特征词 $t_k \in T_2$ 构建情感向量 $I_t^k=\{w_{k1}, w_{k2}, \cdots, w_{k12}\}$,其中,$w_{kj}$ ($j=1,2,\cdots,12$)表示特征词 t_k 在情感向量中的权重,即 t_k 对于第 M_j 的情感强度值,采用 2.1 节中式(1)计算。

由于文本是特征词的集合,因此文本的情感向量 I_x^i 为特征词的情感向量的叠加,其计算见式(2)。其中,I_t^k 表示文本 x_i 中第 k 个特征词的情感向量,n 是 x_i 的特征总数。由此可以构建 x_i 的情感子空间 $\Theta_2(x_i, T_2)$, $\Theta_2 \subset \Theta$。

$$I_x^i = \sum_{k=1}^{q} I_t^k \tag{2}$$

2.4 文本语义相似度计算

对于文本情感聚类,常常面临聚类的方向和结果不是情感相关的。为了解决这一问题,我们提出一种基于 RSSV 和 ESSV 融合的文本语义相似度计算方法(RESS),在文本相关性子空间 Θ_1 和文本情感子空间 Θ_2 结合的基础上,构建文本语义空间 $\Theta_1 \cup \Theta_2 \in \Theta$。

在语义空间中,依据 Θ_1 有效地解决文本向量的高维问题,实现文本表示的有效降维;依据 Θ_2 将数据集的原始空间映射到情感空间,实现文本表示的情感因素表达。

对于文本向量 x_i 和 x_j,基于相关性的文本相似度 $S_1(x_i, x_j)$ 计算如式(3)所示,其中,v_i, v_j 分别为文本 x_i, x_j 的相关性特征向量表示。

$$S_1(x_i, x_j) = \frac{v_i \cdot v_j}{|v_i| \cdot |v_j|} \tag{3}$$

基于情感子空间的文本相似度 $S_2(x_i, x_j)$ 计算如式(4)所示,其中,I_x^i, I_x^j 分别为文本 x_i, x_j 的情感相关的向量表示。

$$S_2(x_i, x_j) = \frac{I_x^i \cdot I_x^j}{|I_x^i| \cdot |I_x^j|} \tag{4}$$

基于 RESS 的 x_i 和 x_j 文本相似度 $S(x_i, x_j)$ 计算如式(5)所示。

$$S(x_i, x_j) = \alpha S_1(x_i, x_j) + (1-\alpha) S_2(x_i, x_j) \tag{5}$$

其中 α 取值范围为(0,1)。当 $\alpha \to 0$ 时, $S(x_i, x_j) \to S_2(x_i, x_j)$;当 $\alpha \to 1$ 时, $S(x_i, x_j) \to S_1(x_i, x_j)$。

3 实验及结果分析

3.1 实验语料及评价指标

本文所使用的语料包含英文语料和中文语料。英文语料来自亚马逊网站的产品评论数据。含概 Book、DVD、Electronic 和 Kitchen 四个领域,每个领域包含 2 000 篇文本,文本情况统计见表 2;中文语料来自第六

届中文倾向性分析评测(COAE2014)，包含保险、翡翠、手机三个领域的的微博数据，文本情况统计见表 3。

表 2　英文数据集文本情况统计

数据集(文本数)	褒贬分布情况(文本数)	词数	词汇数	最长文本的词汇数	最短文本的词汇数
Book(2 000)	Negative(1 000)	133 052	188 35	1 318	2
	Positive(1 000)				
DVD(2 000)	Negative(1 000)	146 109	19 436	930	3
	Positive(1 000)				
Electronic(2 000)	Negative(1 000)	82 581	9 568	305	2
	Positive(1 000)				
Kitchen(2 000)	Negative(1 000)	66 828	7 808	306	2
	Positive(1 000)				

表 3　中文数据集文本情况统计

数据集(文本数)	褒贬分布情况(文本数)	词数	词汇数	最长文本的词汇数	最短文本的词汇数
保险(2 153)	贬义(1 705)	59 211	7 293	67	2
	褒义(448)				
翡翠(2 220)	贬义(282)	58 953	8 181	65	3
	褒义(1 938)				
手机(2 632)	贬义(1 225)	83 557	8 486	76	3
	褒义(1 407)				

本文的实验主要对产品评论文本进行正面和负面两极情感聚类。为了验证聚类结果的有效性，采用纯度和 F 值两个聚类性能评价指标[25]。所有实验的聚类方法均采用 K-means 聚类方法。

3.2　参数确定

在第 2.4 节中提出，基于 RESS 的文本相似度计算需要确定 α 参数。为了分析基于 RSSV 的文本相关度和基于 ESSV 的文本相关度对于 RESS 的文本相似度的影响，本文对 α 取值为[0，0.1，…，1]，采用图示的形式分别展示中、英文数据集聚类的 F 值，如图 2、图 3 所示。

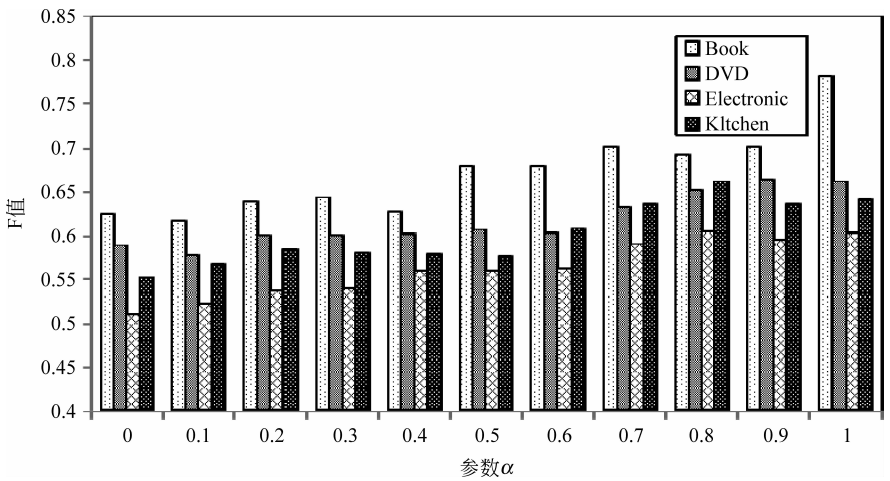

图 2　英文数据集中不同参数 α 下的聚类 F 值

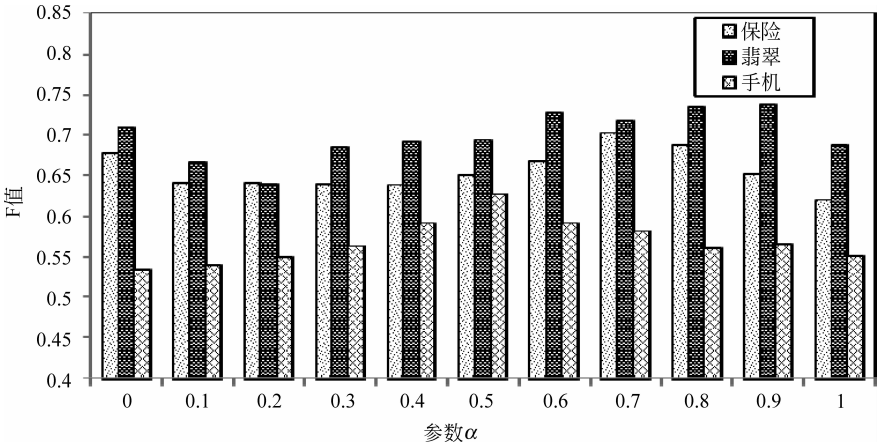

图3 中文数据集不同参数 α 下的聚类F值

从图2和图3中可以看出，对于中、英文数据集，当参数 α 取值为0.6～0.9时，比取值为0.1～0.5所获得的聚类结果更好。这表明文本的上下文(相关性因素)比情感因素在表达文本语义相似度时能够提供更有价值的信息，获得更好的聚类结果。英文数据集的聚类性能峰值保持在参数取值为0.8～0.9之间，说明情感因素在聚类中并未起到很好的作用。而中文数据集的聚类结果峰值在参数 α 取值为0.5－0.9之间，这说明情感因素和语义因素共同作用能够获得更好的聚类结果。尤其对于手机数据集，当参数 α 取0.5时，聚类效果最好。

通过对数据集的分析发现，上述聚类结果与ESSV方法中情感特征集的选取密切相关。对于英文数据集，不同领域选取相同的情感特征词，没有考虑领域相关的情感特征；对于中文数据集，在确定情感类别时采用基于统计的方法，为不同领域选取一定的领域相关的特征词作为补充。情感特征集包含了通用情感词和领域相关的情感词。这种情感特征选择方法能更好地刻画文本中潜在的情感因素，提升情感聚类性能。这也说明，情感集的选定也是影响情感聚类结果的一个重要因素。

3.3 基于语义子空间的情感聚类结果

为了验证语义子空间对文本向量表示的有效降维，本文针对英文数据集和中文数据集，对在原始空间 Θ 和文本语义子空间 $\Theta_1 \cup \Theta_2$ 的文本向量表示进行对比，结果如表4所示。

表4 数据集在 Θ 和 $\Theta_1 \cup \Theta_2$ 的特征数量

		Book	DVD	Electronic	Kitchen	保险	翡翠	手机
文本数		2 000	2 000	2 000	2 000	2 153	2 220	2 632
特征数	Θ	18 835	19 436	9 568	7 808	7 293	8 181	8 486
	$\Theta_1 \cup \Theta_2$	2 636	1 860	610	612	134	130	152

从表4中可以看出，对于中、英文文本数据，采用子空间的文本表示和采用原始特征空间的文本表示在选取的文本特征数量上存在量级的差别。比如，对于Book数据集的2 000篇文本，在 Θ 空间中选取的特征数量为18 835个，在 $\Theta_1 \cup \Theta_2$ 空间特征最多只有2 636个；对于来自微博的中文数据集则更少，保险领域的2 153篇文本在 Θ 空间中选取的特征数为7 293，在 $\Theta_1 \cup \Theta_2$ 子空间，特征数最多为134个。这表明采用语义子空间的文本表示能有效地解决文本向量表示的高维问题。

为了进一步验证语义子空间表示对文本情感聚类的有效性，本文在中英文数据集上进行实验，分别使用基于TF-IDF方法和使用基于概念词典(WordNet、HowNet)的方法构建文本相似度矩阵，聚类的比较结果如表5所示。

从表8中可以看出，本文的方法比采用传统的TF-IDF以及概念词典的文本相似度方法具有更好的F值。采用TF-IDF方法，虽然可以有效地选择对文本聚类具有高区分度的特征词，但没有考虑词语之间的语义关系；基于概念词典的方法只利用词语间的相似度关系，而没有充分考虑词语之间的情感关系；本文方法既考虑了词语之间的相关性，也体现了词语的情感因素，因此能够有效地改进情感聚类的效果。

表 5　不同文本表示的相似度计算方法的情感聚类 F 值

	Book	DVD	Electronic	Kitchen	保险	翡翠	手机
TF-IDF	0.641 2	0.607 9	0.520 6	0.551 9	0.611 7	0.688 2	0.533 5
概念词典	0.657 4	0.621 1	0.601 8	0.630 0	0.659 5	0.707 0	0.550 5
本文方法	**0.700 2**	**0.663 7**	**0.613 3**	**0.661 4**	**0.686 8**	**0.737 7**	**0.627 3**

对比分析在文本相关性子空间 Θ_1、文本情感子空间 Θ_2、文本语义子空间 $\Theta_1 \cup \Theta_2$ 和原始特征空间 Θ 中的情感聚类结果，实验结果见表 6 和表 7 所示。

表 6　数据集在不同表示空间中的聚类纯度

	Book	DVD	Electronic	Kitchen	保险	翡翠	手机
Θ_1	**0.815 3**	0.699 3	0.621 2	0.640 8	0.663 1	0.702 5	0.563 4
Θ_2	0.623 4	0.587 9	0.509 1	0.551 0	0.677 0	0.707 8	0.533 5
$\Theta_1 \cup \Theta_2$	0.756 6	**0.707 1**	**0.626 1**	**0.662 8**	**0.750 2**	**0.794 1**	**0.643 9**
Θ	0.645 0	0.509 1	0.552 9	0.555 3	0.607 3	0.677 5	0.514 4

表 7　数据集在不同表示空间中的聚类 F 值

	Book	DVD	Electronic	Kitchen	保险	翡翠	手机
Θ_1	**0.781 5**	0.661 8	0.601 7	0.640 4	0.619 7	0.687 7	0.550 5
Θ_2	0.623 4	0.587 9	0.509 1	0.551 0	0.677 0	0.707 8	0.533 5
$\Theta_1 \cup \Theta_2$	0.700 2	**0.663 7**	**0.603 3**	**0.661 4**	**0.686 8**	**0.737 7**	**0.627 3**
Θ	0.635 0	0.507 1	0.520 0	0.528 1	0.551 7	0.642 0	0.515 5

从表 6 和表 7 可以看出：

(1) 对于中、英文领域的七个数据集，在 Θ_1、Θ_2、$\Theta_1 \cup \Theta_2$ 三种子空间上的聚类结果均优于原始空间 Θ，对于 DVD、Electronic、Kitchen、保险、翡翠、手机数据集，在空间 $\Theta_1 \cup \Theta_2$ 上获得最好的聚类纯度和 F 值。这说明本文提出的基于语义子空间的文本表示和相似度计算方法在情感聚类中是有效的。

(2) 在英文数据集中，Book 数据集在 Θ_1 上获得的情感聚类效果最好，并且优于 $\Theta_1 \cup \Theta_2$（文本相关性子空间），这与 Book 数据集本身的特点有关。Book 数据集的评论相对其他数据集较长，文本中不但包含了阅读者对某一本书的整体评价和感受，而且也包含大量对书中故事情节和人物情感的客观描述，因此，在 Θ_2 上构建的情感向量是不准确的，会影响评论文本的情感极性。

(3) 在中文数据集中，虽然保险和翡翠数据集是非平衡的，但其聚类的纯度和 F 值均比手机数据集高。通过对数据分类结果的分析发现，保险、翡翠领域的评论文本的语言风格、评价对象和评价词相对固定、单一。这种语言现象有助于提高非平衡数据集多数类的聚类效果，从而改善了数据集整体的聚类结果。这说明本文提出的基于 RESS 的情感聚类方法同样适用于不平衡数据集，这在大数据中有更广泛的应用价值。

4　结论与展望

本文针对情感聚类中文本-特征向量的高维和稀疏问题，以及对评论文本潜在情感因素的表示问题，提出基于子空间的文本语义相似度计算方法(RESS)，通过构建文本相关度子空间(RSS)和文本情感子空间(ESS)，计算文本集语义相似度矩阵，实现情感聚类。在中、英文七个领域的数据集上分别进行实验，结果表明：基于 RESS 的文本语义相似度计算从文本相关性和情感角度实现文本的语义表示，有效地解决文本向量的高维问题，并获得较好的聚类结果。同时，该方法也适用于非平衡数据集。

本文的情感聚类结果将文本分为正面和负面两类，但是五级情感标签可以更细地刻画情感的强度。因此，今后将在五级情感聚类方面开展研究。

参考文献

[1] 孟小峰，慈祥. 大数据管理：概念、技术与挑战[J]. 计算机研究与发展，2013，50(01)：146-169.

[2] Berry M W，Castellanos M. Survey of text mining

[M]. New York: Springer, 2004:219-232.
[3] Turney P D. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews [C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 2002: 417-424.
[4] 李欣,王素格,李德玉. 面向文本情感聚类的维度判别方法[J]. 计算机工程与应用, 2015,51(7):124-130.
[5] 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-Volume 10. Association for Computational Linguistics, 2002:79-86.
[6] Das S R, Chen M Y. Yahoo! For amazon: sentiment parsing from small talk on the Web[J]. Management-Science, 2007, 53(9):1375-1388.
[7] 陈笑蓉,刘作国. 文本聚类的重构策略研究[J]. 中文信息学报,2016,30(02):189-195.
[8] Bilenko M, Basu S, Mooney R J. Integrating constraints and metric learning in semi-supervised clustering[C]//Proceedings of the 21st International Conference on Machine Learning. ICML, 2004:81-88.
[9] Bekkerman R, Raghavan H, Allan J, et al. Interactive clustering of text collections according to a user-specified Criterion [C]//Proceedings of the International Joint Conference on Artificial Intelligence. IJCAI, 2007: 684-689.
[10] Dasgupta S, Ng V. Mining clustering dimensions [C]//Proceedings of the 27th International Conference on Machine Learning (ICML-10), 2010: 26270.
[11] Riloff E, Patwardhan S, Wiebe J. Feature subsumption for opinion analysis[C]//Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing,2006: 440-448.
[12] Feng S, Wang D, Yu G, et al. Extracting common emotions from blogs based on fine-grained sentiment clustering[J]. Knowledge and Information Systems, 2011, 27(2): 281-302.
[13] 黄永光,刘挺,车万翔,胡晓光. 面向变异短文本的快速聚类算法[J]. 中文信息学报,2007,21(02): 63-68.
[14] Makrehchi M, Kamel M S. Text classification using small number of features[M]. Machine Learning and Data Mining in Pattern Recognition. Springer Berlin Heidelberg, 2005: 580-589.
[15] Zheng H T, Kang B Y, Kim H G. Exploiting noun phrases and semantic relationships for text document clustering[J]. Information Sciences, 2009, 179(13): 2249-2262.
[16] Jing L, Ng M K, Huang J Z. Knowledge-based vector space model for text clustering[J]. Knowledge and Information Systems, 2010, 25(1): 35-55.
[17] 王素格,李德玉,魏英杰. 基于赋权粗糙隶属度的文本情感分类方法[J]. 计算机研究与发展, 2011, 48(05):855-861.
[18] 夏云庆,杨莹,张鹏洲,刘宇飞. 基于情感向量空间模型的歌词情感分析[J]. 中文信息学报,2010, 24(01): 99-103.
[19] 刘全超,黄河燕,冯冲. 基于多特征微博话题情感倾向性判定算法研究[J]. 中文信息学报,2014, 28(04):123-131.
[20] 郗亚辉.产品评论中领域情感词典的构建[J]. 中文信息学报, 2016,30 (05):136-144.
[21] Mitral M, Hadi A, Man L, et. al. Sense Sentiment Similarity: An Analysis[C]//Proceedings of the 26th Association for the Advancement of Artificial Intelligence, 2012:1706-1712.
[22] Neviarouskaya A, Ishizuka M. SentiFul: Generating a reliable lexicon for sentiment analysis[C]//Proceedings of the 3th International Conference on Affective Computing and Intelligent Interaction and Workshops (ACII), 2009:1-6.
[23] Ortony A, Turner T J. What's basic about basic emotions? [J]. Psychological Review, 1990, 97(3): 315-331.
[24] 徐琳宏,林鸿飞,潘宇. 情感词汇本体的构造[J]. 情报学报, 2008,27(2):180-185.
[25] Dunning T. Accurate methods for the statistics of surprise and Coincidence[J]. Computational Linguistics, 1993, 19(1): 61-74.

李欣(1990—),硕士,助教,主要研究领域为自然语言处理。
E-mail: 694877398@qq.com

王素格(1964—),博士,教授,主要研究领域为自然语言处理、文本情感分析。
E-mail: wsg@sxu.edu.cn

李旸(1988—),博士研究生,主要研究领域为情感分析。
E-mail: 770624917@qq.com