计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
48(5): 855-861, 2011

基于赋权粗糙隶属度的文本情感分类方法

王素格[1,2] 李德玉[1,2] 魏英杰[3]
[1](山西大学计算机与信息技术学院 太原 030006)
[2](计算智能与中文信息处理教育部重点实验室(山西大学) 太原 030006)
[3](科学出版社 北京 100717)
(wsg@sxu.edu.cn)

A Method of Text Sentiment Classification Based on Weighted Rough Membership

Wang Suge[1,3], Li Deyu[2,3], and Wei Yingjie[4]
[1] (*School of Computer & Information Technology, Shanxi University, Taiyuan* 030006)
[2] (*Key Laboratory of Computational Intelligence and Chinese Information Processing (Shanxi University), Ministry of Education, Taiyuan* 030006)
[3] (*Science Press, Beijing* 100717)

Abstract Facing with promptly increasing reviews on the Web, it has been great challenge for information science and technology that how people effectively organize and process document data hiding large amounts of information to meet with particular needs. Text sentiment classification aims at developing some new theories and methods to automatically explore the sentiment orientation of a text by mining and analyzing subjective information in texts such as standpoint, view, attitude, mood, and so on. A method of text sentiment classification based on weighted rough membership is proposed in this paper. In the method, the model of text expression is established based on two-tuples attribute (feature, feature orientation intensity), by introducing feature orientation intensity into the method of vector space representation. An attribute discretization method is proposed based on the sentiment orientation sequence for feature selection unifying the discretization processing to depress data dimension. To utilize the feature orientation intensity, a weighted rough membership is defined for classifying new sentiment text. Compared with SVM classifier, on the reality car review corpus, the proposed method based on rough membership for text sentiment classification has the best performance after data being compressed in a certainty extent for text sentiment classification.

Key words text sentiment classification; text expression; sentiment orientation intensity; discretization; rough membership

摘 要 提出了基于赋权粗糙隶属度的文本情感分类方法.该方法将特征倾向强度引入到文本的向量空间表示法中,建立了基于二元组属性(特征,特征倾向强度)的文本表示模型.提出了基于情感倾向强度序的属性离散化方法,将特征选择寓于离散化过程,达到数据降维的目的.利用特征倾向强度,定义了赋权粗糙隶属度,用于新文本的情感分类.在真实汽车评论语料上,与支持向量机分类模型进行比较实验

收稿日期:2010-01-04;修回日期:2010-06-23
基金项目:国家自然科学基金项目(60875040,60970014);高等学校博士学科点专项科研基金项目(200801080006);山西省自然科学基金项目(2007011042,2010011021-1);山西省重点实验室开放基金项目(2007031017);太原市科技局明星专项基金项目(09121001)

表明,基于赋权粗糙隶属度的文本情感分类方法在对数据进行一定程度的压缩后仍表现出较好的分类性能.

关键词 文本情感分类;文本表示;情感倾向强度;离散化;粗糙隶属度

中图法分类号 TP391

Internet 的开放性、虚拟性与共享性使得人们已将其作为表达观点(opinion)、态度(attitude)、感觉(feeling)、情绪(emotion)的公共平台. 这些主观信息表现形式大多以非结构化或半结构化的评论文本形式出现,如产品评论、体育评论、Blog、影视评论、新闻评论、股票评论等. 传统的文本分类主要将文本按照政治、经济、军事、教育、体育、环境、计算机以及汽车等主题进行分类[1]. 与传统的文本信息处理不同,对评论文本的信息处理关注的是文本内容所体现的情感(sentiment). 文本情感分类是指通过挖掘和分析文本中的立场、观点、看法、情绪、好恶等主观信息,对文本的情感倾向作出类别判断. 从当前的研究来看,文本情感类别通常分为两类(正面、反面)或 3 类(正面、反面和中立),且以考虑两种类别者居多[2-9]. 文本情感分类可广泛应用于有害信息过滤、社会舆情分析、产品在线跟踪与质量评价、影视评价、Blogger 声誉评价等方面.

目前,统计方法是文本情感分类的主流技术. Pang 等人[2]采用了朴素贝叶斯(NB)、最大熵模型(ME)和支持向量机(SVM)3 种机器学习方法对英文电影评论进行了情感分类实验,测试结果表明,基于 SVM 方法显示出最好的分类效果. 徐军等人[3]利用 NB 和 ME 方法进行了新闻及评论语料的情感分类研究. Wang 等人[4]采用混合的特征选择方法和 SVM 对文本进行了情感分类. Tan 等人[5]采用了 MI, IG, CHI, DF 4 种特征选择方法和 KNN, Centroid Classifier, NB, winnow, SVM 5 种机器学习方法,分别进行了中文文本情感分类实验,实验结果表明,IG 和 SVM 仍得到了最好的分类结果. 以上工作思路是将机器学习方法直接用于文本的情感分类,没有考虑情感词汇和短语在判断主观文本情感倾向中的作用,因此,有一些学者将词汇或短语的情感倾向作为研究对象,通过词汇的情感倾向进一步推断文本的情感倾向. Turney 等人[6]通过分析词汇的上下文信息研究了词汇的情感倾向,采用 PMI-IR 和 LSA 两种方法用于度量给定词汇与基准词的关联度,确定词汇的语义倾向,最后将词汇的情感倾向用于判断句子或篇章的情感倾向. 杜伟夫等人[7]提出了一个识别词语语义倾向的通用框架,用于判断文本的情感类别. 这些方法仅利用了情感词汇的简单组合,对于具有复杂的、微妙的情感表达的文本而言,这些方法还不能很好地体现文本的整体类别,胡熠等人[8]提出了一种基于语言建模的文本情感分类方法,该方法从训练数据中分别估计出代表"赞扬"或"批评"两种语言模型,然后通过研究文本自身的语言模型,并用训练好的情感模型和待测文本之间的 Kullback-Leibler 距离对文本进行情感分类,取得了较好的分类性能. 廖祥文等人[9]提出了一个基于概率模型的博客倾向性检索算法,该算法主要将主题相关性评分和倾向性评分统一到概率推理模型中,对给定查询可以有效地识别其在博客空间中的相关观点.

粗糙集理论是一种处理不精确、非完备与非协调知识的数学工具[10]. 该理论以其"不需要数据的任何先验假设"、"可提供非完备、非协调等不确定性知识获取方法"、"所获知识具有较好的直观可理解性"等优势获得广泛关注,并成功地应用于文本分类. Chouchoulas 等人[11]利用粗糙集方法对文本的特征降维,并应用于电子邮件过滤. Bao 等人[12]提出了浅层语义检索与粗糙集方法融合的文本分类方法. Singh 等人[13]提出了一种基于粗糙集的文本过滤方法,将特征选择统一于离散化过程,获得了较高的分类精度.

考虑到特征的情感倾向强度对文本情感倾向的影响,本文提出了基于二元组属性(特征,特征倾向性强度)的文本表示模型. 为了降低文本表示的维数,本文提出了基于情感倾向强度序的属性离散化方法,它体现了倾向强度高的特征对文本情感倾向处于支配地位这一思想. 最后,文本提出了一种赋权粗糙隶属度作为新文本情感类别判定的依据.

1 带情感倾向强度的文本表示

在以往的文本情感分类研究中[2-3],常常将词汇(或短语)作为分类特征来构造文本表示的向量. 然而,直觉上带有情感色彩的词汇比普通词汇对情

感分类会有更大的分类贡献.因此,我们选择情感词汇作为分类特征.进一步,为了体现情感词汇的倾向强度对分类的作用,我们构造一种带有特征情感倾向强度的文本表示模型.形式地,一个文本 Doc_i 被表示为其在一组属性$(F_1,O_1),(F_2,O_2),\cdots,(F_m,O_m)$下的取值所构成的向量$(w_{i1},\cdots,w_{ij},\cdots,w_{im})$,这里,属性$(F_j,O_j)$由特征 F_j 和其情感倾向强度 O_j 组成,w_{ij}表示文本 Doc_i 在特征 F_j 下的权重.在表1中,"C"列表示文本的情感倾向类别,取"正面"或"反面".

按照上述文本表示模型,一个有监督的文本情感分类的训练数据集可形式化为表1所示的决策表:

Table 1 Decision Table for Text Sentiment Classification

表1 文本情感分类决策表

U	(F_1,O_1)	$\cdots$	(F_m,O_m)	C
D_1	w_{11}	$\cdots$	w_{1m}	c_1
D_2	w_{21}	$\cdots$	w_{2m}	c_2
$\vdots$	$\vdots$		$\vdots$	$\vdots$
D_n	w_{n1}	$\cdots$	w_{nm}	c_m

2 候选特征及其情感倾向强度计算

对文本情感分类而言,我们希望获得区分能力强、且有强烈的主观情感倾向的文本特征.Fisher 准则函数作为一种鉴别特征分类能力的有效度量常被应用于特征选择任务[14],为此,我们设计了基于 Fisher 准则函数与情感词表相结合的候选特征选取方法.候选特征的情感倾向强度计算采用了文献[15]提出的方法.

2.1 Fisher 准则函数[14]

设 P,N 分别表示正、反面文档,t_k 是一个特征项,则 t_k 的 Fisher 准则函数值

$$F(t_k)=\frac{(E(t_k\mid P)-E(t_k\mid N))^2}{D(t_k\mid P)+D(t_k\mid N)} \tag{1}$$

表示特征项 t_k 在正、反两类类间均值差的平方与其在正、反两类类内总方差之比,其值越大表明对分类的贡献越大,反之对分类的贡献越小.

为了计算 Fisher 准则函数值,需要利用训练数据对式(1)中的均值和方差进行估计.设正(反)面文本的第 $i(j)$篇$(i=1,2,\cdots,m;j=1,2,\cdots,n)$的总词次为 $v_{P,i}(v_{N,j})$,特征项 t_k 在第 $i(j)$篇正(反)面文本中出现的次数记为 $w_{P,i}(t_k)(w_{N,j}(t_k))$.则:

$$E(t_k\mid P)=\frac{1}{m}\sum_{i=1}^{m}\frac{w_{P,i}(t_k)}{v_{P,i}};$$

$$E(t_k\mid N)=\frac{1}{n}\sum_{j=1}^{n}\frac{w_{N,j}(t_k)}{v_{N,j}};$$

$$D(t_k\mid P)=\frac{1}{m}\sum_{i=1}^{m}\left(\frac{w_{P,i}(t_k)}{v_{P,i}}-E(t_k\mid P)\right)^2;$$

$$D(t_k\mid N)=\frac{1}{n}\sum_{j-1}^{m}\left(\frac{w_{N,j}(t_k)}{v_{N,j}}-E(t_k\mid N)\right)^2.$$

2.2 基于同义词的词汇情感倾向强度计算[15]

文献[6]提出了基于目标词与褒贬义基准词集间点互信息的词汇情感倾向强度计算方法.考虑到同义词对目标词情感倾向强度的贡献,文献[15]改进了原方法,提出了基于同义词的词汇情感倾向强度计算新方法.

设 w 是一个词,T_w 是词 w 的同义词集合,$Pwords$ 和 $Nwords$ 分别表示褒义基准词集和贬义基准词集,则 w 的情感倾向强度定义为[15]

$$S(w)=\sum\nolimits_{w_i\in T_w\cup\{w\}}S_{Par}(w_i), \tag{2}$$

这里,

$$S_{Par}(w_i)=\sum_{w'\in Pword}PMI(w_i,w')-\sum_{w'\in Nword}PMI(w_i,w'), \tag{3}$$

式(3)中的 $PMI(w_i,w')$为词 w_i 和 w'间的点互信息.

2.3 候选属性获取

候选属性的获取步骤:

Step1. 对训练文本进行分词、词性标注.

Step2. 分别从正反两类文本中抽取词汇.

Step3. 根据式(1),对抽出的词汇计算其 Fisher 准则函数值.

Step4. 给定正整数 N,按照 Fisher 准则函数值由大到小选取前 N 个词汇,并与情感词表 SWT 求交集,得到候选特征.

Step5. 利用式(2)得到每个候选特征的情感倾向强度,进而构成候选属性集.

通过上述5个步骤可以得到带有情感倾向强度的候选属性.这里的情感词表 SWT[16] 主要是借助 GeneralInquirer(GI)词典、《学生褒贬义词典》、知网、《褒义词词典》、《贬义词词典》5种资源构建的中文情感词词表.

3 基于情感倾向强度序的属性离散化方法

在文本表示中,一个众所周知的困难是数据稀疏问题,引起这一问题的主要因素之一是文本表示的维数常常很大,降维是解决这一问题的有效方法之一.粗糙集理论的优势之一就在于能对属性进行

有效的约简，删除冗余数据及属性，从而达到数据降维的目的. 本质上，有监督的连续属性离散化就是要在属性空间内寻找一组超平面(数量极小)，使它们能对空间中分布的数据点以极大粒度进行划分而不致产生非协调情况，即同一个数据粒中不会包含不同类别的数据点. 一般地，利用原始决策表，可以将区分两两不同类别的样本点的属性上的所有分割点作为候选分割点. 然后，按照一定的原则和方法，在候选分割点中选取极少(或较少)数量的分割点，使得这个分割点子集同样可以区分任何两个不同类别的样本点，从而完成对连续属性的离散化. 含分割点数最少的离散化称为最优离散化，而寻找决策表的最优离散化是一个 NP-hard 问题[10]，因此人们设计了各种算法以求得次优解，即寻找次优划分. MD-算法是基于 Johson 策略的启发式算法[10]. 对于文本情感分类问题，我们希望得到的分割点不仅区分能力较强，而且情感倾向强度较大. 为此，当两个分割点具有相同的区分能力时，首先选择情感倾向强度绝对值较大的分割点. 因此，我们设计了基于情感倾向强度序的属性离散化算法.

根据情感倾向强度的绝对值由大到小将属性排序，即 $|O_j|>|O_{j+1}|(j=1,2,\cdots,m-1)$，将训练文本表示成如表 1 所示的带有情感倾向强度的决策表 S. 属性离散化算法如下：

算法 1. 属性离散化算法.

输入：决策表 S；

输出：离散化后的决策表 $S^{*'}$.

Step1. 对决策表 S 中的每一列，将属性 (F_j,O_j) 的值由小到大排序，记为 $w_{j_1,j}<w_{j_2,j},\cdots,<w_{j_k,j}$，获得子区间 $[w_{j_1,j},w_{j_2,j})$，$[w_{j_2,j},w_{j_3,j})$，…，$[w_{j_{k-1},j},w_{j_k,j})$，构造候选分割点集 $Cut=\bigcup_{j=1}^{m}Cut_j$，这里 $Cut_j=\{cut_{j_t,j}=((F_j,O_j),\frac{w_{j_{t+1},j}-w_{j_t,j}}{2})\}_{t=1}^{k-1}$.

Step2. 构造二维表 S^*.

将 S 中所有可能的决策值不同的文本对 (Doc_r,Doc_s) 作为 S^* 的对象，将 Cut 中的元素看作二值属性，如果 $cut_{j_t,j}$ 在 $w_{r,j}$ 和 $w_{s,j}$ 之间，则令 (Doc_r,Doc_s) 在 $cut_{j_t,j}$ 下的值为 1，否则为 0.

Step3. 对表 S^*，选取列中含“1”的个数最多的列，若有多个列满足条件，则选取最左面的列. 删除选中的列以及该列中属性值为 1 所对应的所有的行，将形成的新表赋给 S^*，并将该列对应的分割点加入集合 P.

Step4. 如果表 S^* 不空，则转至 Step3.

Step5. 构造决策表 $S^{*'}$.

利用 P 中的关于某个属性 (F_j,O_j) 的所有分割点，将区间 $[w_{j_1,j},w_{j_k,j})$ 分成若干左闭右开的子区间，并给每个子区间赋予一个符号值，这样就定义一个新属性 $(F_j,O_j)^*$. 将所有的新属性作为 $S^{*'}$ 的属性集，$S^{*'}$ 的对象集与 S 的对象集相同. 某个对象 Doc_i，将 $w_{i,j}$ 落在那个新属性 $(F_j,O_j)^*$ 的子区间对应的符号值赋给 Doc_i 作为其新属性 $(F_j,O_j)^*$ 下的取值.

Step6. 算法结束.

决策表 S 经过离散化后，通常属性个数会减少，属性值也转换为符号值，$S^{*'}$ 中保留的属性即为特征选择的最后结果.

从算法 1 可以看出，基于情感倾向强度序的属性离散化算法在保持系统分类能力不变(保持系统协调性)的条件下，依据特征情感倾向强度由大到小顺序，以候选分割点的区分能力(能够区分不同类别文本对的数量)为启发信息选择最终分割点，达到离散化目的. 由于特征情感倾向强度序的控制，通常情况下，情感倾向强度较小的特征产生的候选分割点将不会进入最终的离散化结果. 因此，该离散化算法隐含了数据降维功能.

4 赋权粗糙隶属度与文本情感类别判定

一般地，利用粗糙集理论中的约简技术对决策表进行约简旨在获得更为简洁的决策规则，以便对新对象的决策类别进行预测. 由于基于粗糙集理论的离散化技术中隐含对决策表的约简，所以利用它得到的离散化结果自然包含了对决策表中条件属性的约简. 当然利用离散化后得到的决策表可以直接获得表达隐含于数据中的知识，即决策规则，并利用它们对新对象的决策类进行预测. 然而，正如我们在前文中所指出的，中文文本处理的主要困难之一就是文本表示的高维性和数据稀疏问题. 由于决策模型中的属性特征是从大量训练文本中获得的，而通常这些属性特征在某一特定文本中出现是少量的，

这就导致利用获得的规则为一个待分类文本预测类别面临条件属性值匹配困难.为此,本文提出了基于赋权粗糙隶属度的分类方法.其主要思想是利用出现在待分文本中的每个属性及其权重寻找该文本在训练文本集中对应的等价类,并计算其关于某个决策类的置信度,然后在属性特征情感倾向强度意义下进行加权,构造该文本关于此决策类的隶属度.由于本文提出的隶属度是基于出现在待分文本中的每一个属性特征来计算的,所以它可以有效避免数据稀疏问题以及基于决策规则的方法所面临的规则难于匹配问题.

定义 1. 设 $S=(U,AT\cup\{C\})$是一个决策表,其中的属性均为符号值属性,$r\in AT$ 是一个条件属性,$C_v=\{x\in U|C(x)=v\}$是一个决策类.称

$$\mu_{C_v}^r(x)=\frac{\|[x]_r\cap C_v\|}{\|[x]_r\|} \tag{4}$$

为 x 在属性 r 下关于决策类 C_v 的置信度.这里$[x]_r=\{y\in U|r(y)=r(x)\}$称为 x 决定的 r 等价类,$\|\cdot\|$表示集合的基数.

易知,$0\leqslant\mu_{C_v}^r(x)\leqslant 1$,$\mu_{C_v}^r(x)$越大表明 x 隶属于决策类 C_v 的可信度越大.

定义 2. 设 Doc 为一个待分类文本,$\{(F_j,O_j)^*\}_{j=1}^m$为决策表 S^* 中的所有属性,w_j 为文本 Doc 在特征 F_j 下的权重,W_j 为 w_j 被符号化后的取值,C_v 是一个决策类.定义

$$\mu_{C_v}(Doc)=\sum_{j,w_j\neq 0}|O_j^*|\mu_{C_v}^{(F_j,O_j)^*}(Doc) \tag{5}$$

为文本 Doc 关于决策类 C_v 的赋权粗糙隶属度,这里 $O_j^*=\dfrac{O_j}{\sum_{l=1}^m|O_l|}$ 表示特征 F_j 的情感倾向强度归一化后的结果.

易知,$0\leqslant\mu_{C_v}(Doc)\leqslant 1$,$\mu_{C_v}(Doc)$越大,表明文本 Doc 被推断隶属于决策类 C_v 的可信度越大.

设 Doc 为一个待分类文本,C_P,C_N 分别表示正反两个情感类别,构造文本 Doc 的情感分类函数为

$$C(Doc)=\arg\max_{C_v\in\{C_P,C_N\}}\{\mu_{C_v}(Doc)\}. \tag{6}$$

5 实验结果与分析

1) 实验数据:为了验证本文提出的方法,从汽车点评网上收集了 2006 年 1 月至 2007 年 3 月间关于国内外 11 种品牌轿车的评论 500 篇,其中正面文本 250 篇,反面文本 250 篇.该文本库中的评论人群主要是车主和即将购车的人,他们大多是从非专业角度进行评论,另外还有少量媒体评论.将其中 400 篇评论作为训练数据集,另外 100 篇作为测试数据集.所有文本进行了分词、词性标注(采用山西大学的分词软件),人工标注了文本的情感类别.

2) 实验步骤:(1)按照第 2.3 节介绍的方法获取文本表示的候选属性,其中候选特征数 N 采用了文献[4,14]提供的经验数值 1 000;(2)利用得到的候选属性对训练文本进行表示,得到原始文本分类决策表 S,权重采用 TFIDF 度量[1]进行计算;(3)对原始文本分类决策表 S 离散化,获得文本表示的最终属性;(4)用第(3)步中得到的属性对测试数据集中的文本进行表示;(5)利用式(5)计算测试文本的隶属度,利用式(6)为测试文本标注情感类别;(6)用评价指标对分类结果进行分析.

3) 评价指标

(1) 离散化方法的数据压缩能力:即离散化后的数据集规模与原始数据集规模的比例.

(2) 离散化方法的属性压缩能力:即离散化后的属性数与原始属性数的比例.

(3) 系统分类性能的量化指标:精确率(简记为 P)、召回率(简记为 R)、$F1$ 值和正确率(A).正(反)面精确率、召回率和 $F1$ 值分别记为 $PP(PN)$,$RP(RN)$和 $FP(FN)$,$F1$ 值=2(精确率×召回率)/(精确率+召回率).

4) 实验结果

(1) 数据的压缩率

由 400 篇训练文本得到候选特征数为 495,这样得到的文本表示矩阵为 400×495.在未进行数据离散化时,文本的特征为"个性、欺骗、精确、最佳、隐患、潮流、气愤、圆润、脱落、高科技……",情感倾向强度较高的候选特征有"精确、稳定性、隐患"等,部分结果如表 2 所示:

Table 2 Original Expression of Partial Training Texts

表 2 部分训练文本的原始表示

Text	(精确,19.9)	(稳定性,15.1)	(隐患,-12.9)	…	C
Doc_1	1.0	0.3	0.1	…	1
Doc_2	0.8	0.7	0.0	…	1
⋮	⋮	⋮	⋮		⋮
Doc_{400}	0.0	0.0	1.0	…	0

离散化后情感倾向性较强的候选特征被保留下来,并且属性值为符号值,部分结果如表 3 所示.

Table 3 Result of Discretization

表 3 离散化后的结果

Text	(精确,19.9)	(稳定性,15.1)	(隐患,-12.9)	…	C
Doc_1	2	1	0	…	1
Doc_2	2	2	-1	…	1
⋮	⋮	⋮	⋮		⋮
Doc_{400}	-1	-1	2	…	0

属性值的赋值原则为:某个属性(F_j, O_j)的特征 F_j 若在某文本中不出现时,其属性值赋给-1,否则,根据属性值权重落在分割点对应的区间$[w_{j_1,j}, w_{j_k,j})$赋值,赋值的符号根据分割点由小到大所对应的区间$[w_{j_1,j}, w_{j_k,j})$,以此赋给 0,1,2,….

离散化的数据和属性的压缩情况如表 4 所示:

Table 4 Compressed Results of Data and Attributes

表 4 数据和属性的压缩情况

Text/Attribute	Original Data	Data After Discretization	Compression Ratio/%
Text	400	339	84.75
Attribute	495	113	22.85

由表 4 可知,本文提出的离散化方法将 495 个候选属性压缩到 113 个,压缩比达到了 22.85%,表明该离散化方法在数据降维方面效果很好.

(2) 文本分类结果与分析

本节对本文的方法与方法 2[13]中的粗糙隶属度方法进行了比较实验.另外,为了比较基于赋权粗糙隶属度方法与基于支持向量机的分类方法的效果,实验中还用原始的 495 个候选属性中的候选特征对文本进行表示,并利用 SVM 方法构造的分类器进行了对比实验,实验结果如表 5 所示:

Table 5 Comparison of Test Results Respectively Obtained by Diverse Classifiers

表 5 多种分类器测试结果比较 %

Index	Method1 (This Paper)	Method2 (Reference[13])	Method3(SVM)
PP	70.21	67.35	77.27
RP	82.50	82.50	68.00
FP	75.86	74.16	72.37
PN	80.56	79.41	71.43
RN	67.44	62.79	80.00
FN	73.42	70.13	75.47
A	74.70	72.29	74.00

由表 5 可以看出:本文所提方法的测试效果优于方法 2 中的粗糙隶属度方法,正确率提高约 2.5%,表明在文本情感分类中应该考虑特征的情感倾向强度因素.另外,与方法 3(SVM)方法相比,本文方法的正面召回率和 *F*1 值、反面精确率均优于 SVM 分类方法,正确率略有提高,表明压缩后的属性特征可以得到与原特征相近的效率.

6 结束语

针对文本情感分类,本文推广了文本表示的向量空间模型,提出了带情感倾向强度的文本向量表示模型,在此模型表示下,提出了基于情感倾向强度序的属性离散化方法,用于对文本表示降维;借助于粗糙集理论中的隶属度思想,构造了赋权粗糙隶属度,用于文本类别的判定,推广了文献[13]中的方法.与同类方法及基于支持向量机方法的比较实验表明,本文提出的文本情感分类方法在数据压缩、各项分类评价指标,以及分类结果的可解释性等方面均优于其他方法.尽管本文的方法是以词汇层次语言粒度为文本表示特征的分析,但方法本身也适用于以搭配、关联对等为特征的文本情感倾向性分析任务.

致谢 感谢哈尔滨工业大学信息检索研究室提供的"语言技术平台 LTP"中的《同义词词林扩展版》.感谢董振东先生提供的 HowNet 的情感词汇和评价词汇!

参 考 文 献

[1] Yang Y, Pedersen J O. A comparative study on feature selection in text categorization [C] //Proc of the 14th Int Conf on Machine Learning. San Francisco, CA: Morgan Kaufmann, 1997: 412-420

[2] Pang B, Lee L, Vaithyanathan S. Thumbs up? sentiment classification using machine learning techniques [C] //Proc of the Conf on Empirical Methods in Natural Language Processing (EMNLP). Philadelphia, PA: Association for Computaional Linguistics, 2002: 79-86

[3] 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 (in Chinese)
(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)

[4] Wang Suge, Wei Yingjie, Li Deyu, et al. A hybrid method of feature selection for Chinese text sentiment classification [C] //Proc of the 4th Int Conf on Fuzzy Systems and Knowledge Discovery. Los Alamitos, CA: IEEE Computer Society, 2007: 435-439

[5] Tan Songbo, Zhang Jin. An empirical study of sentiment analysis for Chinese documents [J]. Expert Systems with Application, 2008, 34(4): 2622-2629

[6] Turney P D, Littman M L. Measuring praise and criticism: inference of semantic orientation from association [J]. ACM Trans on Information Systems, 2003, 21(4): 315-346

[7] Du Weifu, Tan Songbo, Yun Xiaochun, et al. A new method to compute semantic orientation [J]. Journal of Computer Research and Development, 2009, 46(10): 1713-1720 (in Chinese)
(杜伟夫，谭松波，云晓春，等. 一种新的情感词汇语义倾向计算方法[J]. 计算机研究与发展，2009，46(10)：1713-1720)

[8] Hu Yi, Lu Ruzhan, Li Xuening, et al. Research on language modeling based on sentiment classification of text [J]. Journal of Computer Research and Development, 2007, 44(9): 1469-1475 (in Chinese)
(胡熠，陆汝占，李学宁，等. 基于语言建模的文本情感分类研究[J]. 计算机研究与发展，2007，44(9)：1469-1475)

[9] Liao Xiangwen, Cao Donglin, Fang Binxing, et al. Research on blog opinion retrieval based on probabilistic inference model [J]. Journal of Computer Research and Development, 2009, 46(9): 1530-1536 (in Chinese)
(廖祥文，曹冬林，方滨兴，等. 基于概率推理模型的博客倾向性检索研究[J]. 计算机研究与发展，2009，46(9)：1530-1536)

[10] Komorowski J, Pawlak Z, Polkowski L, et al. Rough Sets: A Tutorial [M]. Rough Fuzzy Hybridization: A New Trend in Decision Making. Berlin: Springer, 1999: 3-98

[11] Chouchoulas A, Shen Q. Rough set-aided keyword reduction for text categorization [J]. Applied Artificial Intelligence, 2001, 15(9): 843-873

[12] Bao Yongguang, Aoyama Satoshi, Yamada Kazutaka, et al. A rough set based hybrid method to text categorization [C] // Proc of WISE'01. Los Alamitos, CA: IEEE Computer Society, 2001: 254-261

[13] Singh S, Dey L. A new customized document categorization scheme using rough memship [J]. Applied Soft Computing, 2005 (5): 373-390

[14] Wang Suge, Li Deyu, Wei Yingjie, et al. A feature selection method based on fisher's discriminant ratio for text sentiment classification [G] //LNCS 5854: Proc of WISM 2009. Berlin: Springer, 2009: 88-97

[15] Wang Suge, Li Deyu, Wei Yingjie, et al. A method for word sentiment orientation discriminating based on synonyms [J]. Journal of Chinese Information Processing, 2009, 23(5): 68-74 (in Chinese)
(王素格，李德玉，魏英杰，等. 基于同义词的词汇情感倾向分类研究[J]. 中文信息学报，2009，23(5)：68-74)

[16] Wang Suge, Yang Anna, Li Deyu. Research on Sentence sentiment classification based on Chinese sentiment word table [J]. Computer Engineering and Application, 2009, 45(24): 153-155, 161 (in Chinese)
(王素格，杨安娜，李德玉. 基于汉语情感词表的句子情感倾向分类研究[J]. 计算机工程与应用，2009，45(24)：153-155，161

Wang Suge, bron in 1964. PhD and associate professor. Her main research interests include intelligence information retrieval, natural language processing and text mining.

Li Deyu, bron in 1965. Professor and PhD supervisor. Senior member of China Computer Federation. His main research interests include rough set theory, granular computing and data mining.

Wei Yingjie, born in 1982. Master. His main research interests include text minging.

基于赋权粗糙隶属度的文本情感分类方法

作者： 王素格， 李德玉， 魏英杰， Wang Suge， Li Deyu， Wei Yingjie

作者单位： 王素格,李德玉,Wang Suge,Li Deyu(山西大学计算机与信息技术学院,太原,030006;计算智能与中文信息处理教育部重点实验室(山西大学),太原,030006)， 魏英杰,Wei Yingjie(科学出版社,北京,100717)

刊名： 计算机研究与发展 ISTIC EI PKU

英文刊名： JOURNAL OF COMPUTER RESEARCH AND DEVELOPMENT

年，卷(期)： 2011,48(5)

参考文献(16条)

1. 王素格;杨安娜;李德玉 基于汉语情感词表的句子情感倾向分类研究 2009(24)

2. 王素格;李德玉;魏英杰 基于同义词的词汇情感倾向分类研究 2009(05)

3. Wang Suge;Li Deyu;Wei Yingjie A feature selection method based on fisher's discriminant ratio for text sentiment classification 2009

4. Singh S;Dey L A new customized document categorization scheme using rough memship 2005(05)

5. Bao Yongguang;Aoyama Satoshi;Yamada Kazutaka A rough set based hybrid method to text categorization 2001

6. Chouchoulas A;Shen Q Rough set-aided keyword reduction for text categorization 2001(09)

7. Komorowski J;Pawlak Z;Polkowski L Rough Sets:A Tutorial 1999

8. 廖祥文;曹冬林;方滨兴 基于概率推理模型的博客倾向性检索研究 2009(09)

9. 胡熠;陆汝占;李学宁 基于语言建模的文本情感分类研究 2007(09)

10. 杜伟夫;谭松波;云晓春 一种新的情感词汇语义倾向计算方法 2009(10)

11. Turney P D;Littman M L Measuring praise and criticism:inference of semantic orientation from association[外文期刊] 2003(04)

12. Tan Songbo;Zhang Jin An empirical study of sentiment analysis for Chinese documents 2008(04)

13. Wang Suge;Wei Yingjie;Li Deyu A hybrid method of feature selection for Chinese text sentiment classification 2007

14. 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 2007(06)

15. Pang B;Lee L;Vaithyanathan S Thumbs up? sentiment classification using machine learning techniques 2002

16. Yang Y;Pedersen J O A comparative study on feature selection in text categorization 1997