一种基于语义距离的Web评论SVM情感分类方法

肖 正 刘 辉 李 兵
(湖南大学信息科学与工程学院 长沙 410012)

摘 要 情感倾向分析本质上可以看作是一个情感极性分类问题。在海量数据处理的大背景下,为了提高文本情感判断的准确率,提出了一种结合潜在语义分析LSA(Latent Semantic Analysis)和支持向量机SVM(Supported Vector Machine)的文本褒贬情感倾向分类方法。从语义的角度利用潜在语义分析方法建立"词-文档"的语义距离向量空间模型,然后使用具有良好分类精度和泛化能力的支持向量机进行情感分类。实验结果表明,该方法在句子简短、情感倾向比较明显的Web评论中的准确率较传统的SVM方法有了一定的提高,在测试集上的分类准确率接近88%。

关键词 文本处理,语义距离,情感极性分类,潜在语义分析

中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2014.09.047

SVM Sentiment Classifier Based on Semantic Distance for Web Comments

XIAO Zheng LIU Hui LI Bing
(Department of Information Science and Engineering, Hunan University, Changsha 410012, China)

Abstract The analysis of sentimental orientation can be regarded as a problem of classification on emotional polarity. Under the background of the mass data processing, we proposed a classification approach in terms of sentimental orientation of texts based on LSA(Latent Semantic Analysis) and SVM(Supported Vector Machine), in order to improve the accuracy of the text emotional judgment. On the concept of semantics, we established a space model of "word-document" semantic distance vectors by the latent semantic analysis, and then on account of the privileges of accuracy and generalization of support vector machine, designed a SVM classifier with semantic distance as the input feature vectors. Experimental results validate that our method effectively improves the classification accuracy compared with the traditional SVM method. The classification accuracy rate rises to near 88% on the test set of Web comments with short sentences and explicit sentimental orientation.

Keywords Text processing, Semantic distance, Sentimental orientation classification, Latent semantic analysis

1 引言

文本情感分析又称意见挖掘,简单而言,是对带有情感色彩的主观性文本进行分析、处理、归纳和推理的过程[1]。文本情感倾向分析任务可大致划分为文本的主客观分析和主观信息的细粒度分析这两个方面,目前的研究主要集中在主观信息的细粒度分析。细粒度分析中的句子级情感倾向分析大都是以情感词为中心,通过情感词褒贬度的线性加权值来最终衡量句子的情感倾向。有代表性的方法是熊德兰等人[2]以HowNet语义词典的词汇语义相似度计算为基础,综合考虑当前句子中词汇间的组合方式对情感词褒贬倾向的加强或减弱作用,最终通过句中各情感词的语义加权值来判定句子的情感倾向。这种方法以语义词典为基础,准确率高,但不足之处是非常依赖词汇级情感分析方法的计算精度,且处理过程复杂,不适合对大量数据进行快速处理。针对大量数据的处理,研究人员引入了机器学习的方法进行情感分析,如K最近邻、朴素贝叶斯、支持向量机以及最大熵模型等方法[3]。机器学习方法的分类准确度虽然没有基于语义词典的语义加权方法高,但优点是能够快速进行大量数据分析,处理过程相对简单。例如徐军等人[4]利用机器学习方法对新闻评论进行情感分类,在最理想的数据集上分类准确率可以达到90%,然而这种方法缺乏语义分析,容易产生向量空间模型数据稀疏问题,对于中文文本处理中普遍存在的一词多义和多词一义问题也不能解决。闻彬等人[5]在情感词识别中引入情感语义概念,基于语义理解来进行文本情感分类,可在一定程度上缓解一词多义和多词一义引起的分类准确率不高的问题,它的不足之处在于只考虑到词语语义层副词的出现规律对词语语义的作用,忽略了整个文本语境对词语语义的影响。有部分研究者在分析过程中将句子的句法结构考虑进来,通过句子中已有的一些规则,实现句子的情感倾向判断[6]。但这种方法处理过程复杂,推广能力差,实用性不高。也有部分研究者尝试借助非标注样本不断训练分类器的方法来提高半监督学习方法的情感分类准确率,实验证明该方法是有效的[7],这对基于机器学习的文本情感分类方法研究会有一定的帮助。

到稿日期:2013-11-19 返修日期:2014-03-24 本文受湖南大学湖南省自然科学基金(13JJ4038),湖南大学"青年教师成长计划"资助。
肖 正(1981-),男,讲师,硕士生导师,主要研究方向为情感计算与文本处理、人工智能等;刘 辉(1987-),男,硕士生,主要研究方向为情感计算与文本处理,E-mail:liuhui_hndx@163.com;李 兵 男,硕士生,主要研究方向为文本处理与电磁计算。

SVM具有很好的泛化能力和出色的分类性能，因此被广泛应用于分类问题中。传统基于SVM的分类方法直接以情感词在文档中的统计特性（如频度）作为SVM的输入特征向量。但该方法要求情感词在待分类文档中能充分出现，并存在以下3个方面的问题：(1)数据稀疏问题。情感词集不足以覆盖所有文档；(2)一词多义问题。情感词集中情感词可能具有多种语义，导致不同程度的情感倾向；(3)多词一义问题。同样的情感可能用不同的情感词表达。当存在以上3种现象时，传统SVM方法性能下降。实质上，上述3个方面体现了情感词与文档语义关系，基于语义特征进行分类能扩大情感词的分类规模，提高分类的精度。因此，本文引入潜在语义分析技术，将情感词集与待分类文档的在语义上关联起来，以“词-文档”语义距离作为SVM的特征向量。通过融合潜在语义分析和支持向量机，有效解决上述3个问题。实验结果表明，本文提出的方法相比于传统的SVM方法在分类准确度上有所提高。

2 相关背景

2.1 特征向量空间

传统SVM文本分类方法都是以初始“词-文档”向量空间模型作为分类器的输入，利用已标注好的数据集训练分类模型，并将其应用于待分类数据集中，获取分类结果。文本情感分类性能的好坏主要取决于输入数据的形式和分类器性能这两方面。

文本情感分析中，特征词与文档的相互关系大都通过向量空间模型（Vector Space Model，VSM）来表示，将数据向量化，利于问题的后续处理。“词-文档”关系矩阵是文本情感分类中最常见的特征向量空间模型，关系矩阵中的特征值的好坏在一定情况下决定了分类器的性能。特征值一般都是特征词在对应文档中的词频数，常见的特征提取方法有DF（词频）、信息增益（IG）和互信息（MI）等[8]，将“词-文档”之间的关系表示为向量空间。TF-IDF（词频-逆向文档频率）方法是最常见的词频计算方法之一，TF-IDF计算公式如下[9]：

$$Tf\text{-}idf(i,j)_j=\frac{n_{i,j}}{\sum_k n_{k,j}}\times\log\frac{|D|}{1+|\{j:t_i\in d_j\}|} \quad (1)$$

其中，$n_{i,j}$表示第i篇文档中第j个特征情感词，$|D|$表示总文档数。

利用词频计算方法，各文档被映射到所有特征词组成的特征向量空间中，构建初始“词-文档”向量空间模型。

2.2 LSA原理

LSA的出发点是在同一语料库中，文本中的词与词之间存在某种潜在的语义结构，通过这种潜在的语义结构能够获取词语间在上下文环境中的语义，从语义的角度获取词和文档在向量空间中更准确的表示。

LSA是一种用于自动实现知识提取和表达的数学统计技术[10]。LSA的表示形式和VSM相似，不同之处是LSA通过奇异值分解（Singular Value Decomposition，SVD），对初始VSM进行了特征变换。

VSM中假设词语间语义是相互独立的，即在构成的“词-文档”向量空间模型中，每个词语被看作是一个正交基本向量。然而，实际情况中VSM的词语之间在语义上是相互关联的，且向量空间的维度与特征词成正比。LSA将“词-文档”向量空间映射到低维的潜在语义空间中，缩小问题的规模，达到降维的目的”。同时，LSA从语义角度将词语间关联起来，能够消除词语在文档中一词多义和多词一义可能产生的“噪音”问题，从而提高后续处理的效率和精度[11]。

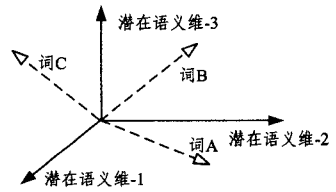

图1 三维潜在语义空间示意图

图1是一个三维潜在语义空间示意图，通过潜在语义维将词向量相互关联起来，词向量之间的相关性可以其在潜在语义空间中的夹角来度量。LSA通过矩阵的SVD对构成的“词-文档”向量空间模型进行变换，将词汇和文档映射到一个低维的潜在语义空间中，建立“词-文档”的语义距离特征向量空间模型。

2.3 LSA在情感分类中的优势

LSA语义分析存在以下优势：

• 数据稀疏性

初始“词-文档”向量空间模型中，很多特征词可能只出现在少数几篇文档中，这样构成的向量空间模型数据非常稀疏。LSA通过潜在语义结构将词语间关联起来，利用概念的传递性，使得向量空间模型数据不再稀疏。

• 一词多义

初始“词-文档”向量空间模型中，特征词都被认为是语义唯一的，特征词的一词多义性增加了整个文档分类的复杂度。LSA从词语间的上下文环境来理解词的语义，能够正确区分特征词在不同文档中的语义，解决特征词的一词多义性，降低整个文档分类的复杂度，进而提高文档分类准确率。

• 多词一义

在初始“词-文档”向量空间模型中，特征词之间是互不相关的，特征词在文档中的多词一义问题也无法解决。LSA从词语间的上下文环境来理解词的语义，能够将多个同义词归纳为具有相同特征的词，解决了特征词的多词一义性存在的高维问题。

3 基于LSA的情感倾向分类方法

LSA和SVM的理论原理都非常成熟，应用于文本分类具有很多优势。LSA主要用于文本中特征的提取和表达，它可以降低数据的维度，能够处理文本分析中常见的一词多义和多词一义引起的“噪声”问题，同时也解决了原始“词-文档”向量空间模型中存在的数据稀疏问题。

SVM是主要的机器学习方法之一，可以应用于对大量数据的处理，它建立在统计学习理论的VC维理论和结构风险最小原理基础上，使得它的分类能力和泛化能力都非常不错[12]，作为典型的二元分类器，非常适合应用于文档的褒贬二元情感分类。因此，本文以VSM作为连接点，将两者的优势完美地结合起来，并将其应用到文本的褒贬情感分类中。

图 2 给出了 LSA-SVM 情感分类方法的框架。

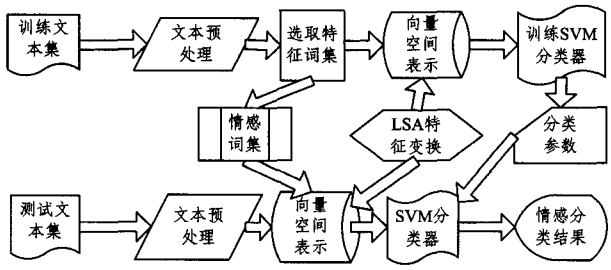

图 2 LSA-SVM 分类方法框架

1)文本预处理:对待分析数据集进行中文分词、去停用词等预处理操作,目前已有一些集成这些操作的中文分词系统;

2)选取特征词集:特征词集是用于表示所有文档的特征词,其选取合理与否对分类有一定的影响,目前已存在许多特征词集选取方法,具体将在 3.1 节中阐述。

3)向量空间表示:对数据进行向量化,利于后续的分类操作。最常见的表示形式是"词-文档"关系矩阵,利用 TF-IDF(文档-逆向文档频率)方法将各文档映射到所有特征词组成的向量空间中,构建初始"词-文档"向量空间模型。

4)LSA 特征变换:将初始"词-文档"向量空间模型进行特征转换,建立"词-文档"的语义距离向量空间模型,在 3.2 节中将进行详细阐述。

5)训练 SVM 分类器:根据已标注好的训练数据集对 SVM 进行训练,获取相应的分类器参数,为文档自动分类提供最佳的分类器。

3.1 情感特征词集的选取

文本情感分析中的特征词大都选择具有情感倾向的形容词、副词和名词作为候选特征词,通过设置一定的阈值,将特征值大于该阈值对应的词选为特征词。其中,特征值往往与特征向量空间中特征值的定义相同。该方法有以下两点不足:1)难以确定合理的阈值;2)某些阈值高的值对应的词可能会产生"噪声"。例如在基于 DF 情感特征词选取方法中,有些词几乎在所有的文档中出现,虽然该词的 DF 值很高,但是该词已经不具备作为整个文档分类的特征之一,不利于整个文档的分类。

本文采取的情感特征词选取原则为:选取那些情感倾向比较强烈且在文本中覆盖范围广的形容词、副词和名词。前者能够更好地表征文档的情感倾向,后者可以有效地缓解数据稀疏问题。具体步骤如下:

1)将分词后所有标注为形容词、副词和名词的词语构成候选特征词集。

2)结合知网情感分析用词语集词典,对候选特征词集进行粗分类,若候选特征词不属于情感分析用词语集词典中的词语,则将其从候选特征词集中滤除。

3)对过滤后的候选特征词集,用 DF 特征提取方法进行进一步的过滤,获取特征情感词集。词的覆盖范围能力可以用包含该词的文档总数即词的 DF 值来衡量。计算公式为:

$$C(t_i)=\{j:t_i\in d_j\} \tag{2}$$

式中,t_i 是候选情感词,d_j 是指包含 t_i 的文档。

4)利用式(2)计算所有候选特征词的 DF 值,选取 DF 值大于某一定值的词组成特征词集。

3.2 基于 LSA 的特征向量空间提取

LSA 的基本原理已经在第 2 节中有过详细的介绍,这里主要介绍基于 LSA 的特征向量空间提取的实现过程。

1)给定构造好的初始"词-文档"向量空间模型。

2)对初始"词-文档"向量空间模型进行奇异值分解。奇异值分解定理:任何一个矩阵均可分解成两个酉矩阵与一个对角矩阵的乘积[13]。

$$X=WSD^{\mathrm{T}} \tag{3}$$

式中,矩阵 W,D^{T} 分别是矩阵 X 的左奇异向量矩阵和右奇异向量矩阵,$S=\{\partial_1,\partial_2,\cdots,\partial_r\}$ 是矩阵 X 的奇异值矩阵,且满足 $\partial_1\geqslant\partial_2\geqslant\cdots\geqslant\partial_r\geqslant 0$。选取前 K 个奇异值,得到原始矩阵秩为 K 的近似矩阵,如下式所示:

$$\hat{X}_k=W_kS_kD_k^{\mathrm{T}} \tag{4}$$

具体过程如图 3 所示。

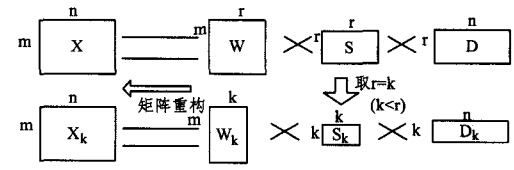

图 3 SVD 过程

SVD 通过一种简单的矩阵分解方法,选取前 $K(K\ll r$,其中 r 是矩阵 X 的秩)个奇异值进行矩阵分解反运算,便能得到原始矩阵维度大大减少的近似矩阵。LSA 在 SVD 的基础上保留了最主要的 K 个奇异值,其中 K 值就是潜在语义空间的维度。已经证明,重构后的矩阵 X_k 是原始矩阵 X 在秩为 K 的前提下最小二乘意义上的最优近似[14]。

3)根据经验选取合适的 K 值,然后对截取后的 3 个矩阵进行重构,获得语义转换后的"词-文档"向量空间模型。

LSA 特征提取通过 SVD 实现,它保留了整个文档中最主要的语义关系,并通过潜在语义空间构建了"词-文档"的语义距离向量空间模型。LSA 中,潜在语义维度 K 值选取不能过大也不能过小,K 值过大,特征词之间被认为是语义上相互独立,忽略了实际中存在的语义相关性;K 值过小,特征词之间都被认为是语义相关的,使得词语间的语义无法区分,两者都会引起分类效果变差。关于如何选取潜在语义 K 值的问题,盖杰、宁健等人[15,16]也进行了相应研究,但大都是根据经验来获取局部最佳 K 值的。如何获取最佳 K 值也是本文后续将要研究的课题之一。

3.3 LSA-SVM 具体算法

图 2 中描述了 LSA-SVM 方法的流程,大致分为两部分,上半部分是对 SVM 模型进行训练;下半部分则利用建立的模型进行分类。具体算法如下:

算法 1 LSA-SVM 训练算法

输入:已标注好的训练文档集 $D=\{d_1,d_2,\cdots,d_n\}$

输出:参数模型 $P=\{\alpha,\beta,\chi\}$

1. 由 D 获取特征情感词集 $C=\{c_1,c_2,\cdots,c_m\}$
2. for i=1 to m,j=1 to n do
3. $\quad tf_{i,j}=Tf\text{-}idf(i,j)$
4. $\quad X=[tf_{i,j}]_{m\times n}$
5. end for
6. $\quad X=[w]_{m\times r}\times[s]_{r\times r}\times[d]_{r\times n}\xrightarrow{SVD}X_k=[w]_{m\times k}\times[s]_{k\times k}\times[d]_{k\times n}$
7. $X_k\xrightarrow{SVM}$参数模型 P

算法 1 中,K 是选取的潜在语义维度。$\{\alpha,\beta,\chi\}$ 是 SVM

分类器训练过程中的3个重要参数，分别是交差验证参数、惩罚因子和核函数参数。

步骤1中，特征情感词集利用3.1节中提到的方法获取。步骤3是TF-IDF词频计算公式，通过该方法构建初始“词-文档”向量空间模型X，如步骤4所示。步骤6是LSA特征向量空间提取过程，获取特征转换后的“词-文档”向量空间模型X_k。步骤7是训练SVM分类参数过程。

算法2 LSA-SVM分类算法

输入：待分类文档集 $T=\{d_1,d_2,\cdots,d_t\}$

输出：$Tab(d_t\in T)$值

1. for i=1 to m,j=1 to n do
2. $tf_{i,j}=Tf\text{-}idf(i,j)$
3. $X_t=[tf_{i,j}]_{m\times t}$
4. end for
5. $X_t=[w]_{m\times r}\times[s]_{r\times r}\times[d]_{r\times t}\xrightarrow{SVD}X_{t,k}=[w]_{m\times k}\times[s]_{k\times k}\times[d]_{k\times t}$
6. $X_{t,k}\xrightarrow{SVM+\{\alpha,\beta,\chi\}}Tab(d_t\in T)$

算法2中，K是选取的潜在语义维度，$Tab(d)$是文档类别标签函数。步骤2是计算各个特征词在对应文档中的TF-IDF值，步骤3是构建特征词与待分类文档的向量空间模型。步骤5是LSA特征向量空间提取过程，获取特征转换后的“词-文档”向量空间模型$X_{t,k}$。步骤6是通过训练好的分类器进行文档分类，获取最终的文档情感分类结果。

4 一个简单示例

下面将以一些实际的酒店类评论文档为例，详细说明LSA在文本情感倾向分析中所起的作用。

表1 酒店类评论实例

文档编号	文档内容
neg 1	离火车站比较近，隔音差，外边太吵了，一晚上听车子声音听到天亮。
neg 2	酒店离火车站比较远，环境差，服务员素质也差，房间连衣架都没有。
neg 3	房间隔音效果很差，周围比较吵，环境不如锦江干净，不推荐入住。
pos 1	房间比较干净，而且电视离床较远，感觉不错，服务人员态度也很好。
pos 2	距拱北关口近，标准间太小，豪华大床不错，服务好。
pos 3	酒店地理位置好，并且靠近马路也不是很吵，非常舒适。

表2 初始“词-文档”关系矩阵X

情感词	情感词对应的文档中TF-IDF值					
	neg 1	neg 2	neg 3	pos 1	pos 2	pos 3
近	0.8004	0.0	0.0	0.0	0.9782	0.0
差	0.3599	0.7198	0.3599	0.0	0.0	0.0
吵	0.3599	0.0	0.3599	0.0	0.0	0.4398
远	0.0	0.8004	0.0	0.7043	0.0	0.0
干净	0.0	0.0	0.8004	0.7043	0.0	0.0
不错	0.0	0.0	0.0	0.7043	0.9782	0.0
好	0.0	0.0	0.0	0.3167	0.4398	0.4398

表3 SVD后的“词-文档”关系矩阵X_k

情感词	情感词对应文档中的特征值					
	neg 1	neg 2	neg 3	pos 1	pos 2	pos 3
近	0.7423	0.0378	0.0113	−0.033	0.9998	0.1375
差	0.4525	0.6203	0.3715	0.0636	−0.073	0.0775
吵	0.2920	0.2454	0.1358	−0.018	0.0918	0.0479
远	0.0202	0.5312	0.4216	0.6449	−0.009	0.0549
干净	−0.117	0.3664	0.3357	0.7076	0.0631	0.0388
不错	0.0551	−0.091	0.0528	0.7353	0.9306	0.1858
好	0.0608	−0.018	0.0374	0.3393	0.4507	0.4508

在中文概念词典中，概念具有传递性。如果两个词在同一篇文档中多次共现，那么它们之间就可能存在语义相关性[17]。比如词W_1、W_2、W_3是“词-文档”向量空间中的3个特征词，假设词W_1和词W_2在文档D_1中经常共现，词W_2和词W_3在文档D_2中经常共现，那么经过LSA处理后，由于词W_2的传递作用，词W_1和词W_3就具有一定的语义相关性，因此，即使词W_1并没有出现在文档D_2中，但通过词概念的传递作用，使得词W_1有可能出现在文档D_2中，并在D_2表示的向量空间中根据可能性大小赋予一定的权值。因此，表3中，“词-文档”关系矩阵X经SVD后得到的新关系矩阵X_k全部是非零元素，且基本各不相同，解决了表2中的数据稀疏问题。

特征词“远”在原评论文档中的neg2和pos1中出现，权值也差不多，那么这个特征词可能存在一词多义性。经过LSA处理后，“远”这个特征词在所有6篇文档中都有相应的权值，并且该词在neg2和pos1中的权值都下降了，在负面文档neg3中的权值由原来的0增加到0.4216，从整个文档分析，特征词“远”被正确归类为负面情感特征词，解决了特征词中存在的一词多义问题。

特征词“好”和“不错”在原评论文档中都只出现在正面文档中，那么这两个特征词可能属于多词一义现象。经过LSA处理后的表3中，“好”和“不错”在正负文档中对应的权值都相似，也就是说这两个词被自动归类为同一性质的词，解决了多词一义问题。

表3中存在部分特征值为负值的情况，因为这些情感特征词在所有文档中被归类为某一类特征词，然而实际中却出现在另一类文档中。

5 实验分析

5.1 实验环境及评测指标

本文实验部分，情感分析语料库选取了中科院谭松波老师整理的1000篇关于酒店类的论坛评论信息，实验采用Java编程实现。其中，中文分词选用了中科院的ICTC-LAS分词系统，SVM分类器选取了林智仁老师的lib-svm JAVA版本。情感特征词集与领域相关，本文的情感特征词集按照3.1节的方法进行选取，其中$C(t_i)$值选取为2。酒店类论坛评论的情感词集如表4所列，共选取了42个情感特征词。

表4 酒店类论坛评论的情感词集

破	恶劣	舒适	周到	脏	坏	一般
方便	愉快	热情	差	舒服	乱	近
小	偏僻	失望	大	宽敞	便宜	简陋
不错	合理	不便	烂	美	潮湿	丰富
贴心	暗	干净	旧	不好	豪华	划算
贵	优越	安静	嘈杂	好	陈旧	糟糕

本实验采用准确率、假阴阳比作为衡量方法性能的指标，其中假阴阳比是衡量方法稳定性的重要指标。参数P代表初始正面文档总数，参数N代表初始负面文档总数，参数P_P代表初始正面文档分类后结果仍然是正面的文档数，参数N_N代表初始负面文档分类后的结果仍为负面的文档数，P_N代表初始正面文档分类结果为负面的文档数，N_P代表初始负面文档分类结果为正面的文档数。性能指标计算式如下：

准确率：$C=(P_P+N_N)/(P+N)$；

总判误率：$E=(P_N+N_P)/(P+N)$；

假阴性：$EP=P_N/P$；

假阳性：$EN=N_P/N$；

假阴阳比：EP/EN。

5.2 实验结果分析

本文采用传统的 SVM 情感分析方法和 LSA-SVM 情感分析方法进行对比分析，实验过程中，将 1000 篇文档（正负各为 500 篇）按训练集和测试集 3∶2 的比例随机分成 3 组，最终取 3 组的平均值作为结果。

针对 LSA-SVM 方法流程中，训练过程和测试过程中潜在语义维度的取值问题进行了对比试验，实验结果如表 5 所列。

表 5 训练和测试阶段中参数 K 值的选取对分类结果的影响

测试阶段 K 值 / 训练阶段 K 值	K=10 的分类结果（准确率%）	K=12 的分类结果（准确率%）	K=8 的分类结果（准确率%）
K=10	85.25	85.75	79.75
K=8	80.5	83.25	79.5
K=12	82.5	80.25	80.75

从表 5 的结果可知，训练和测试阶段最优的 K 值是不同的，但当 K 均为 10 时，分类准确率接近于最优 K 值（训练取 10，测试取 12）对应的结果。其原因在于本文分析的数据都是在同一主题下的文档，特征词与文档存在类似的关系，其最佳 K 值比较接近。因此，在接下来的实验中训练阶段和测试阶段的 K 值相同。

根据经验[18]，本实验中潜在语义维 K 值取 10。实验结果如表 6 所列。从表 6 的结果中可知，LSA-SVM 方法的分类准确率比 SVM 方法有了 6.25%的提高，其中 LSA-SVM 方法的 EP/EN 值接近 1，即 LSA-SVM 方法的误判率不会出现向某一方严重倾斜的情况。综合这两项可以得出 LSA-SVM 方法的整体性能较 SVM 方法好。因此，实验验证了本文提出的 LSA-SVM 模型方法的性能比传统的 SVM 方法更优越。

表 6 SVM 和 LSA-SVM 方法的结果对比

编号 \ 结果		准确率 C(%)	总判误 E(%)	假阴性 EP(%)	假阳性 EN(%)	假阴阳比 EP/EN
1	SVM	80.5	19.5	16.13	22.42	0.72
	LSA-SVM	85.0	15.0	15.05	14.95	1.01
2	SVM	77.25	22.75	20.83	24.52	0.85
	LSA-SVM	84.75	15.25	14.58	15.86	0.92
3	SVM	80.75	19.25	19.60	18.88	1.04
	LSA-SVM	87.5	12.5	16.67	13.26	1.26
均值	SVM	79.5	20.5	19.98	21.94	0.87
	LSA-SVM	85.75	14.25	15.43	14.69	1.06

针对情感词集对该分类模型可能产生的影响，本文在表 1 中随机选取了 10 个、20 个、30 个和 40 个情感特征词构成特征词集进行试验，考虑到情感特征词数量大于当前情感特征词集的影响，将情感特征词增加到 50 个、60 个和 70 个，实验结果如图 4 所示。

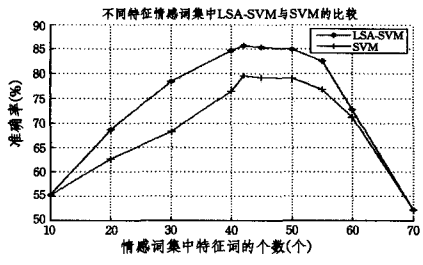

图 4 情感词集对分类模型的影响

图 4 中，情感分类的准确率随着情感特征词的增加而上升，当情感特征词数少于 20 个时，SVM 方法分类准确率只有接近 63%，其原因是构成的原始“词-文档”向量空间表示非常稀疏，造成很多文档的向量表示中包含很多零元素，使得 SVM 分类性能下降。而对于 LSA-SVM，经过 LSA 处理后，向量空间模型变得不再稀疏，分类准确率接近 70%。在情感词集相同的情况下，LSA-SVM 方法的分类准确率比 SVM 方法有明显提高。然而，也不是特征词数越多越好，当特征词数量超过 60 个以后，分类器性能会急剧下降。其原因应该是当分类特征超过一定数量时，数据本身已变得无序，使得分类器出现不可分状态。

结束语 本文从词语语义的角度对 Web 评论信息进行情感倾向分析，利用 LSA 方法进行词语间语义计算，将情感词集与待分类文档的在语义上关联起来，以“词-文档”语义距离作为 SVM 的特征向量，解决了机器学习中广泛存在的数据稀疏问题，同时也有效地降低了一词多义和多词一义问题对情感分类的影响。结合具有较好分类性能和泛化能力的 SVM 分类器，提出了文本的正负二元情感分类的 LSA-SVM 方法。实验结果表明，该方法较传统的 SVM 方法在准确率上有了一定的提高，但同时也发现情感特征词的选取对分类性能有很大的影响。另外，LSA 中潜在语义维度的选取对分类效果可能也会有一些影响，这些都将是我们今后研究的方向。

参 考 文 献

[1] Subasic P, Huettner A. Affect analysis of text using fuzzy semantic typing[J]. IEEE Transactions on Fuzzy Systems, 2001, 9(4): 483-496

[2] 熊德兰，程菊明，田胜利. 基于 HowNet 的句子褒贬倾向性研究[J]. 计算机工程与应用，2008，44(22)：143-145

[3] 唐慧丰，谭松波，程学旗. 基于监督学习的中文情感分类技术比较研究[J]. 中文信息学报，2007，21(6)：88-94

[4] 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

[5] 闻彬，何婷婷，罗乐，等. 基于语义理解的文本情感分类方法研究[J]. 计算机科学，2010，37(6)：261-264

[6] 段建勇，谢宇超，张梅. 基于句法语义的网络舆论情感倾向性评价技术研究[J]. 情报杂志，2012，31(1)：147-150

[7] 高伟，王中卿，李寿山. 基于集成学习的半监督情感分类方法研究[J]. 中文信息学报，2013，27(3)：120-126

[8] 代六玲，黄河燕，陈肇雄. 中文文本分类中特征抽取方法的比较研究[J]. 中文信息学报，2004，18(1)：26-32

[9] Ramos J. Using tf-idf to determine word relevance in document queries[C]// Proceedings of the First Instructional Conference on Machine Learning. 2003

[10] Dennis S, Landauer T, Kintsch W, et al. Introduction to latent semantic analysis[C]// Slides from the tutorial given at the 25th Annual Meeting of the Cognitive Science Society. Boston, 2003

[11] Landauer T K. Latent semantic analysis [M]// Encyclopedia of Cognitive Science. Nature Pub Group, 2006

[12] Chang C C, Lin C J. LIBSVM: a library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2011, 2(3): 27

（下转第 284 页）

引方案较适合对大型文本库的处理。

结束语 本文提出了一种能够支持多种子近似串匹配的q-gram索引结构，通过该索引能够快速获取任意长度连续种子的地址集合。文中首先详细介绍了所使用的q-gram索引结构，然后详细地阐述了获取任意长度种子的地址集合的相关理论，最后给出了在本文索引中快速提取任意连续种子地址集合的相关算法。相关实验数据表明，相比文献[19]中的多索引方案，本文的索引方案在牺牲很小的速度的基础上就成倍地减少了索引空间的消耗。因此，本文索引方案较适合作为大型文本库多种子近似串匹配的索引结构。

虽然本文索引方案减少了多种子近似匹配时索引空间的消耗，但索引速度却有所下降，本文将进一步研究索引的优化方法，以提高索引的性能。另外，近似串匹配中留空种子的研究已成为热点问题，作者将研究支持多留空种子的索引方法。

参 考 文 献

[1] Dorneles C F, Goncalves R, Mello R D. Approximate data instance matching: a survey[J]. Knowledge and Information Systems, 2011, 27(1): 1-21

[2] Navarro G. A guided tour to approximate string matching[J]. ACM Computing Surveys, 2001, 33(1): 31-88

[3] Lu C W, Lu C L, Lee R C T. A new filtration method and a hybrid strategy for approximate string matching[J]. Springer Berlin Heidelberg, 2013, 20: 143-155

[4] Levenshtein V. Binary codes capable of correcting deletions, insertions, and reversals[J]. Soviet Physics Doklady, 1966, 10(8): 707-710

[5] Burkhardt S. Filter algorithms for approximate string matching [D]. Saarland: Department of Computer Science, Saarland University, 2002

[6] Tian Y, Tata S, Patel J M, et al. Practical methods for constructing suffix trees[J]. VLDB Journal, 2005, 14(3): 281-99

[7] Manber U, Myers G. Suffix arrays: A new method for on-line string searches[J]. SIAM Journal on Computing, 1993, 22(5): 935-948

[8] Navarro G, Baeza-yates R. A practical q-gram index for text retrieval allowing errors[J]. CLEI Electronic Journal, 1998, 1(2): 1-16

[9] Kim M S, Whang K Y, Lee J G. n-Gram/2L- approximation: A two-level n-gram inverted index structure for approximate string matching[J]. Computer Systems Science and Engineering, 2007, 22(6): 365-379

[10] Puglisi S J, Smyth W F, Turpin A, Inverted files versus suffix arrays for locating patterns in primary memory[C]// Proceedings of the 13th Symposium on String Processing and Information Retrieval. Berlin, Germany: Springer Verlag, 2006: 122-133

[11] Altschul S F, Gish W, Miller W, et al. Basic local alignment search tool[J]. Journal of Molecular Biology, 1990, 215(3): 403-410

[12] Altschul S F, Madden T L, Alejandro A S, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs[J]. Nucleic Acids Research, 1997, 25(17): 3389-3402

[13] Wu S, Manber U. Fast text searching allowing errors[J]. Communications of the ACM, 1992, 35(10): 83-91

[14] Chang Y I, Chen J R, Hsu M T. A hash trie filter method for approximate string matching in genomic databases[J]. Applied Intelligence, 2010, 33(1): 21-38

[15] Burkhardt S, Crauser A, Ferragina P, et al. Q-gram based database searching using a suffix array[C]// Proceedings of the Annual International Conference on Computational Molecular Biology, RECOMB 99. New York, USA: ACM, 1999: 77-83

[16] Jokinen P, Ukkonen E. Two algorithms for approximate string matching in static texts[C]// 16th International Symposium Proceedings on Mathematical Foundations of Computer Science. Berlin, Germany: Springer-Verlag, 1991: 240-248

[17] Rasmussen KR, Stoye J, Myers EW. Efficient q-gram filters for finding all epsilon-matches over a given length[J]. Journal of Computational Biology, 2006, 13(2): 296-308

[18] Sutinen E, Tarhio J. Filtration with q-samples in approximate string matching[C]// Proceedings of 7th Annual Symposium on Combinatorial Pattern Matching, CPM 96. Berlin, Germany: Springer-Verlag, 1996: 50-63

[19] Ma B, Tromp J, Li M, et al. PatternHunter: faster and more sensitive homology search[J]. BIOINFORMATICS, 2002, 18(3): 440-445

[20] Egidi L, Manzini G. Better spaced seeds using Quadratic Residues[J]. Journal of Computer and System Sciences, 2013, 79(7): 1144-1155

[21] Baeza-Yates R, Navarro G. Faster approximate string matching [J]. Algorithmica, 1999, 23(2): 127-158

[22] Myers E W. A sublinear algorithm for approximate keyword searching[J]. Algorithmica, 1994, 12(4/5): 345-374

[23] Ilie S. Efficient computation of spaced seeds[J]. BMC Research Notes, 2012, 5(1): 1-7

[24] Kim Y, Park H, Shim K. Efficient processing of substring match queries with inverted variable-length gram indexes[J]. Information Sciences, 2013, 244: 119-141

[25] Yang X, Wang B, Li C. Cost-based variable-length-gram selection for string collections to support approximate queries efficiently[C]// Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. New York, USA: ACM, 2008: 353-364

[26] Karp R M, Rabin M O. Efficient randomized pattern-matching algorithms[J]. IBM Journal of Research and Development, 1987, 31(2): 249-260

[27] NCBI. UniGene Build #223, Homo sapiens[DB/OL]. (2010-1-27)[2010-04-28]. ftp. ncbi. nih. gov//repository/ UniGene/Homo_sapiens/Hs. seq. uniq. gz.

(上接第252页)

[13] Kalman D. A singularly valuable decomposition: the SVD of a matrix[J]. College Math Journal, 1996

[14] Golub G H, Van Loan C F. Matrix computations [M]. Baltimore, MD, USA: Johns Hopkins University Press, 1996: 374-426

[15] 盖杰，王怡，武港山. 潜在语义分析理论及其应用[J]. 计算机应用研究，2004，21(3)：9-12

[16] 宁健，林鸿飞. 基于改进潜在语义分析的跨语言检索 [J]. 中文信息学报，2010，24(3)：105-111

[17] 于江生，俞士汶. 中文概念词典的结构[J]. 中文信息学报，2002，16(4)：12-20

[18] 王卫国，徐炜民. 基于潜在语义分析的个性化查询扩展模型[J]. Computer Engineering，2010，36(21)：43-45