支持向量机在文本分类中的研究与应用

张燕[1]，姚志远[2]，陈文社[1]

（1. 69010部队，乌鲁木齐 830017；2. 69012部队，乌鲁木齐 83001）

摘　要：运用人工智能相关技术实现海量数据文本的自动化分类识别，将有限的人力从海量数据中解放出来，已成为促进工作发展的重要途径。主要运用SVM文本分类技术对数据文本进行自动筛选和智能分类，详细介绍了SVM文本分类方法的系统结构、分词、特征选择、评估方法、模型训练和分类识别的过程，并针对语料库中的大量文本进行分类实验。结果表明，该方法具有较好的分类效果。

关键词：文本分类；机器学习；特征选择；支持向量机

DOI:10.16184/j.cnki.comprg.2018.08.024

1　概述

近年来，随着信息化手段的不断扩充完善，信息获取的能力得到了明显提升，所获数据量更是成几何倍数增长。数据量的增长对信息处理能力提出了更高要求，如何在海量信息资源中获取有效的信息已成为当前急需解决的问题。研究基于SVM的智能文本分类系统，利用SVM文本分类识别技术，在海量的信息素材中，实现数据文件的自动筛选和智能分类，大幅提升价值信息获取的效率。

2　系统设计

文本分类是自然语言处理领域的一个重要应用，文本分类涉及到的学科比较广，包括：数据挖掘、信息学、计算语义学、人工智能等。目前常用的文本分类方法主要包括决策树、支持向量机、K近邻算法（KNN）、贝叶斯算法、粗糙集[1]以及神经网络等方法。其中支持向量机文本分类方法因其实现简单以及具有较高的分类精度，已成为中文文本分类中比较常用的方法。

2.1　文本预处理

文本预处理工作主要分为分词和去停用词两个步骤。

分词是指在中文文本中连续的能够代表语义单元的词，分词的过程是将中文文本的连续字节流转化为离散单词流的过程。拟采用基于特定词典的分词方法进行分词，它是按照特定的分词策略将需要分词的中文字符串与选定的词典中的词条进行逐一的对比匹配，若在词典中找到某个中文字符串，则正确匹配。

停用词是一些完全没有用或者没有意义的词，例如助词（是、的）、语气词等。停用词表是哈工大停用词表，含有767个停用词，过滤掉训练集中的停用词。去掉一些低频词，比如某些单词只在一两个文本中出现过，去掉标记信息，比如标点符号和网页中的记号。

2.2　特征向量选择

由于大多数文本都具有非结构化或者半结构化的特点，计算机很难对它们直接进行运算和处理，因此在文本分类之前，首先需要把文本转化为计算机可以处理的结构化表示形式。拟采取向量空间模型（Vector Space Model）[2]进行文本表示。向量空间模型（Vector Space Model）的核心思想是用向量来表示文本，将文本d看作是向量空间中的一个n维向量，如下式所示：

$$d = [(t_1, w_1), (t_2, w_2), \cdots, (t_n, w_n)]$$

其中t_i表示文本d的第i个特征项；w_i表示文本d的第i个特征项所对应的权重。权重的大小表示该特征在文本d中的重要程度，权重越大该特征越重要，权重越小该特征越不重要。

当前有很多计算特征权重的方法，其中TF-IDF[3]（词频-逆文件频率）是一种比较常用加权技术，在数据挖掘与信息检索领域都具有广泛的应用。拟采用TF-IDF来计算特征权重，其核心思想是在某一特定的文档中，如果某个词出现的频率越高，并且在其他文档中出现的频率较低，则认为此词具有较好的分类区别能力，该词对应的权重自然也就越大。对于在某一特定文档d里的词语t来说权重计算的公式如下：

$$W(t) = TF(t) * IDF(t) = (t_i/c) * \log(N/n+1)$$

其中：t为特定单词t在文档d中出现的次数；c为文档d的总词数，N为全部语料的文本总数；n为包含单

收稿日期：2018-05-19

词 t 的文本总数，为了防止分母为零所以要加 1。通过 TF 和 IDF 计算特征权重，提取权值最大的前 1000 个特征向量作为最终的特征向量。

2.3 SVM 分类算法

支持向量机（SVM）在解决小样本、非线性以及高维模式识别中有许多独特的优势，它是以统计学习理论和结构风险最小化原理为基础，根据样本有限的条件，在模型复杂度和机器学习能力之间寻找最佳折衷，来提高机器学习的泛化能力。

支持向量机（SVM）在模式识别中的主要思想是寻找一个超平面作为分类决策面实现对正例和反例的分割，并尽量使两个类别之间的空白间隔最大，以保证分类的高可信度。在线性的情况下，该问题可以归结为一个二次规划问题。在非线性的情况下，可以将输入矢量利用核函数映射到一个高维的特征空间，在高维空间非线性的问题仍然可以转化为线性可分问题。分类原理示意如图 1 所示。

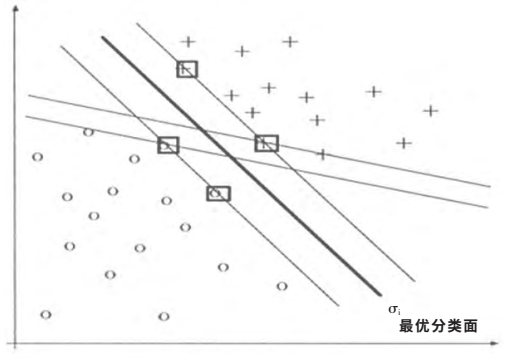

图 1 SVM 分类原理示意图

采用高斯核函数作为支持向量机的核函数：

$$K(x,y)=\exp(-\gamma(x-y)2)$$

按照文本预处理和文本特征向量选择得到训练样本的特征向量，根据特征向量训练得到一个训练模型，最终将利用这个训练模型对待分类的测试样本进行分类识别。

2.4 文本分类评价指标

准确率和召回率[4]是普遍用于统计分类和信息检索领域的两个评价指标，主要对分类的效果进行评价。

准确率（P）是指被正确分类的文本数与被识别为该类别的文本数的比率，代表分类器做出正确判断的概率。

P = 提取出的正确信息条数/提取出的信息条数。

召回率（R）是指被正确分类的文本数占测试文本总数的比率，代表文本被分类器正确识别的概率。

R = 提取出的正确信息条数/样本中的信息条数。

F1 值即为正确率和召回率的调和平均值，F1 值是对分类效果的综合评价，一般 F1 值越大，分类的效果就越好。

F1= 正确率 * 召回率 * 2/(正确率 + 召回率)

3 算法流程及实验结果

3.1 算法流程

本系统实现采用 Java 作为开发工具，算法流程：首先分别对提前准备好的训练样本和测试样本按照词库匹配的方法进行分词，去除停用词、低频词以及标记信息，将文本用向量空间模型表示为特征向量，通过 TF 和 IDF 计算得到特征项权重，将计算到的特征权重从高到低排序，取前 1000 个作为最终的特征向量，将这 1000 维特征向量输入到 SVM 分类器中训练，得到训练模型，最后可以用这个训练模型对测试样本进行分类识别。算法流程图如图 2 所示。

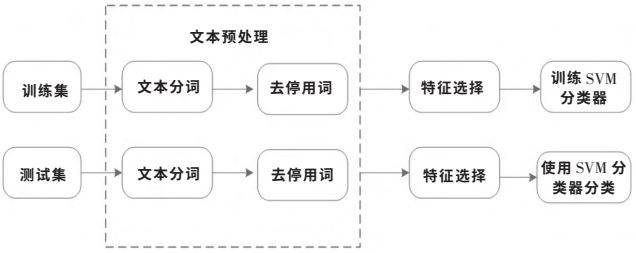

图 2 SVM 文本分类流程图

3.2 实验结果

实验数据来源于网上下载的一个文本语料库，从语料库中随机选取军事、娱乐、健康、农业、体育 5 大类文章各 1000 篇，再从每个类别随机抽取 600 篇文章作为训练样本，剩下的所有文章作为测试样本。实验结果如表 1，图 3 所示。

表 1 SVM 分类结果表

评价指标 / 类别	准确率 P	召回率 R	F1 值
军事	83.70%	85.50%	84.59%
娱乐	82.20%	81.40%	81.80%
健康	84.10%	82.60%	83.34%
农业	80.30%	82.70%	81.48%
体育	81.80%	80.20%	80.99%

(下转第 85 页)

在 DBMS 外层加密比内核加密要简单得多，仅仅只需要技术人员制作外层加密工作就能实现加密效果，成本投入小。

3 结语

网络信息技术的普及与运用，给人们的生活带来了极大的便利，但也面临着许多安全问题，尤其是数据库安全问题受到了人们的普遍关注。为了提高网络数据库的安全稳定性，就必须要使用各种数据库安全技术，挖掘网络资源优势，并充分利用网络资源优势，解决网络中遇到的问题，是广大从业者必须要思考和解决的问题，需要不断地去研究探索，满足新时代网络发展需求。

参考文献

[1] 蒋继洪. 计算机系统、数据库系统和通信网络的安全与保密 [J] . 成都: 电子科技大学出版社, 2014, (16) : 129-135.

[2] 朱鲁华, 陈荣良. 数据库加密系统的设计与实现[J] . 计算机工程技术, 2015, (25) : 235-242.

[3] 李军, 孙玉芳. 计算机的防毒原则及中毒后的修复处理方法 [J] . 河北北方学院学报 (自然科学版) , 2016, (19) : 321-325.

[4] 刘延华. 数据库安全技术的理论探讨 [J] . 福州大学学报, 2015, (20) : 214-218.

[5] 钟勇, 秦小麟. 数据库入侵检测研究综述 [J] . 计算机科学, 2015, (08) : 108-115.

[6] 陈明忠. 入侵检测技术在数据库系统的应用研究 [J] . 计算机工程与科学, 2016, (02) : 321-326.

(上接第 69 页)

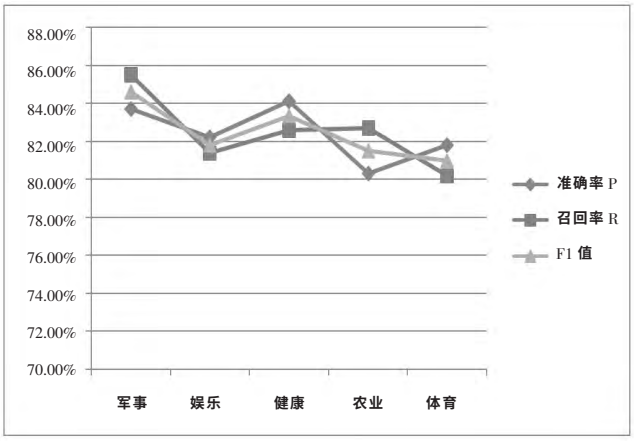

图 3 SVM 五类样本分类结果图

上述结果表明使用该分类模型对文档进行筛选分类，分类结果的 F1 值达到了 80%以上。

4 结语

讨论了 SVM 文本分类方法在数据文本智能分类中的应用，提出了基于 SVM 模型的文本分类系统的设计和实现。详细介绍了该系统的模型训练和分类识别的具体步骤和流程，通过实验效果看，该方法具有较好的准确率和召回率。

参考文献

[1] 林珣, 李志蜀, 周勇. 基于粗糙集理论的文本分类算法研究 [J] . 计算机科学, 2011.

[2] 牛强, 王志晓, 陈岱. 基于 SVM 的中文网页分类方法研究 [J] . 计算机工程与应用, 2007, (8).

[3] 杨凯峰, 张毅坤, 李燕. 基于文档频率的特征选择方法 [J] . 计算机工程, 2010, (9).

[4] 汪光庆. 基于 SVM 的网页分类技术研究 [D] . 中国石油大学硕士论文, 2011.

(上接第 82 页)

部分构成了 JS 的强大能力基础。它的弱类型，简单易用性、解释性及跨平台性，使其在 Web 的应用开发中可以说无处不在。随着 WebAPP 的需求不断增加，JS 的异步执行能力、性能优化、后端 JS 的发展，必将使 JavaScript 的总体发展达到前所未有的高度。

参考文献

[1] 沈昶军. 藏族民居建筑色彩文化研究 [J] . 大众文化, 2016.

[2] 刘莲花, 陈瑛. JavaScript 的面向对象特性分析 [J] . 电脑知识与技术, 2017.

[3] Nicholas C.Zakas [美] . JavaScript 高级程序设计 [M] . 3 版. 人民邮电出版社.