基于 SVM 的中文文本分类系统的设计与实现

张昭楠
(陕西职业技术学院 陕西 西安 710000)

摘要：互联网已经成为现代生活中不可或缺的一部分，网络上的信息量也在以数倍的速度快速增长。无论是企事业单位，学校，或者科研院校等等机构中，都积累了非常多的资料，这些资料绝大多数都以文档的形式存在。所以，如何将数以万计且排序混乱的文本信息，按照一定的规则和形式进行统一的管理，以达到方便使用和管理的目的成为了一个不得不去解决的问题。本文就是在 SVM，即支持向量机方法的基础上，设计了一个中文文本分类系统。介绍了系统的需求分析，并对系统进行了详细设计，从概念的初始化设计到之后的详细设计，实现了基于 SVM 的中文文本分类系统的最终目的，达到了设计要求。

关键词：文本分类；支持向量机；文本表示；特征选择
中图分类号：TN99 **文献标识码**：A **文章编号**：1674-6236(2016)16-0139-03

Design and implementation of Chinese text categorization system based on Support Vector Machine

ZHANG Zhao-nan
(*Shaanxi Vocational and Technical College*, *Xi'an* 710000, *China*)

Abstract: The internet has become an indispensable part in modern life, the amount of information on the network also several times at the speed of fast growth. Both the enterprises and institutions, schools, or scientific research in colleges and universities, and so on organization, have accumulated a lot of information, the information is mostly in the form of document. So, in the face of these massive amounts of text document information, how to effectively manage and utilize them becomes a have to solve the problem. This article is in the SVM, namely, on the basis of support vector machine (SVM) method, a Chinese text classification system is designed. Introduces the system requirement analysis, and has carried on the detailed design of system, after the initialization of the concept of design to detailed design, realized the ultimate goal of Chinese text classification system based on SVM, and has reached the design requirements.

Key words: text classification; support vector machine; text presentation; feature selection

随着信息化时代的全面降临，信息资源也已经和能源，物质等常规资源占有同样重要的地位。我国最近大力推行的信息化建设，也正是对这方面越来越重视的充分体现。当今，互联网上出现了各种各样的信息，信息量也以几何倍数的快速增长。而这些信息大部分都是以文本的形式存在的。另外在各个大中型院校，政府机构或者企事业单位等等这些地方，都存放着大量纸质或数字化的文档资料。通常，为了更好的存储和保留，纸质文档都会录成数字文档，存放起来[1-2]。日积月累，数字文档的数量也急剧膨胀。面对海量的文本文档，对它们合理的管理和利用，就显得特别重要。而研究的这些方法就是所谓的文本处理技术。该技术的核心就是本文介绍的文本分类技术。文本分类，以前都是依靠人工操作来进行的，而且不同领域的分类标准和办法是不相同的。但是这样会有一个明显的问题就是，这种方法需要的人力和物力都是惊人的，而且效率非常低，有些情况下只靠人力是无法完成的。正因为如此，设计开发一个方便快捷的文本分类系统，就显得非常重要了。本文，基于 SVM 的中文文本分类系统的设计与实现，通过对系统的需求分析，以及对系统的详细设计，很好的解决了这个问题，大幅度的提高了文本分类的效率和准确性[3-4]。

1 SVM 方法概述

所谓 SVM，全称是支持向量机(Support Vector Machine)。是一种近年来推出的一种全新的分类和统计方法。该方法遵循的原则是结构化风险的最小化，与传统常规的方法相比，其优势也比较明显。比如理论基础特别扎实。根据统计学理论权威专家通过实践得出的结论来看，SVM 方法不仅解决了以前文本分类周期长，准确度低，范围较小等弱点。还可以在极小的样本条件下，仍然可以满足常规的使用方式，且效率不会受到影响。因此该方法逐渐受到人们的重视，并且以广泛的使用到了文本分类，人脸识别，指纹识别等领域。SUV 方

收稿日期：2016-03-22 **稿件编号**：201603297
作者简介：张昭楠(1986—)，女，陕西渭南人，硕士研究生，助教。研究方向：中国古代文学，语言学及应用语言学。

法不是无限制自动调控的系统,其自身也有着安全的监控算法和分类算法。在进行文本分类处理时,一般需要经过两个步骤,分别为训练和分类[5]。训练过程,与之字面意思不同,这里的训练实际就是对词语重新定义的一个过程。由于重新规划所面临的问题很多,计算量又是十分之大。随着技术的优化,如今引入了多维空间理论,不同纬度空间实现了交叉映射,从而避开了线性与非线性的问题,支持向量机也因此成为了一种常用的分类方法[6-7]。

2 系统需求及可行性分析

通过对 SVM 的介绍,其属于智能分类算法。所以在进行文本分类工作之前要进行小部分的人工分类,为分类器提供比对和校准,也就是所谓的训练功能。训练完成后系统的记忆功能将会自动生效,以后同种类型的文本就无需重复的操作。在操作完成后,系统会自动对结果进行统计和分析,以到达最好的效果[8-9]。下面介绍一下 SVM 的训练分类过程,所谓训练,就是将普通词典的文本及文件输入到计算机中去。在计算机分词系统的存储和处理后,以一种以比较特殊的此类表格的数据结构存储在系统的数据库内存中,为中文文本分类程序的使用做好准备。下面对 SVM 进行分类器的使用的过程进行说明,首先就是将需要处理分类的文档打开,对其先进行预处理操作,在对其特征,权重等因素进行计算,最后使用构建成功的文本分类器系统进行自动分类。在分类完成后,系统会将分类前文本的各种参数以及分类后文本的各种参数一并展示在使用者面前,方便使用者查看。一般在设计系统时。都会对其系统的可行性作出分析和说明,这里的重点就是在系统无论进行哪一步操作之前,首先必须完成文本的预先处理工作。预处理是将文本中的中文分词进行简单的,系统可以识别的标识操作[10]。通过对文献的分析以及对类似产品的比对, 发现此方法已经广泛运用到各个分类系统中,所以该方法无论从技术上还是使用经验上都以十分完善和成熟,在使用时直接进行操作和调用即可。

3 系统概要设计

前文已经提到,在系统进行文本分类操作之前,需要将文本中一小部分词语提取出来进行分类器的模拟比对和训练工作,即将文本的一小部分提出进行试验,这一小部分文档就作为分类成功的模板被记录下来[11-12]。本系统的功能示意图如图 1 所示。系统概要设计中一个比较重要的部分就是文本的预处理过程, 该过程是在系统的预处理模块中进行的。该过程其实就分为标识处理和存储处理,这是因为汉语不会像拉丁语一样可以用空格符号来隔开,所以在系统分类前要对词语进行标识处理,而此步骤是后续分类系统顺利进行工作的前提。所以说本模块在整个中文分词系统中的地位是举足轻重的。前文提到的中科院的 ICTCLAS 中文分词系统,正是因为其在预处理过程中性能稳定,处理准确,且提供免费的使用,受到了很多科研机构以及需要此功能人员的青睐。并且该系统还自带了记忆功能,对于经过操作的分词文本再次输入时无需进行重复的操作。中文文本分类系统的权重模块也是设计的重点之一,本文在普通文本分类系统单一算法的基础上,通过对多种算法的比较和实际运用,综合了传统的比对权重算法以及细化权重算法等常用算法,在选择其优势的基础上对出现的问题进行了改进,提出了本文基于 BG*IG 的全新算法。本系统的另一大优势就是文本表示模块的使用,与传统表示模块不同,本系统采用的表示模型是数学上的向量空间的思想。文本经过训练模块和预处理模块的处理后,分本由统一的格式分化成为具有各种不同属性的分词,但是 SVM 只支持向量格式的数据,这时文本表示模块就发挥了作用,其会自动将传递过来的文本信息转化为三维空间向量的形式。所以本文采用该方法进行文本的表示[13]。

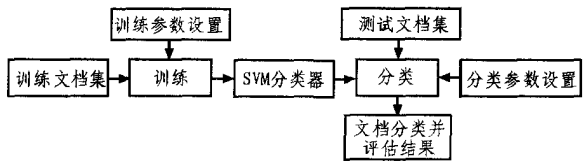

图 1 基于 SVM 的文本分类功能示意图

4 系统详细设计

中文文本分类系统设计的重点首先就是系统的总体界面,系统的总体界面如图 2 所示。其对应的算法格式与之前提到的相同,且是由权重算法自动生成的。作用就是负责对显示系统和响应系统的菜单进行操作和处理,并且所有菜单的操作和处理都是基于此类算法产生的。本系统所有的实现功能首先都是要经过界面上显示的“操作 ”按钮来实现的,换句话说,系统的所有功能在操作界面上都可以体现出来。在对文本进行分类操作时,首先单击“训练 SVM 分类器”菜单选项,这时会跳出一个训练设置界面,在设置完成后点击确认按钮,系统就会自动进行分类训练。这时观察 SVM 文本分类和查看分类结果菜单都是灰色的状态,这就说明分类工作还在进行, 在完成这部分工作后系统才会进行下一步的操作,也就是进行文本的分类。在分类完成后分类结果会出现在显示器上。本系统还有一个优势就是默认的参数往往就是可靠性以及效率最高的值,通常不需要更改,如若需要进行特殊格式的分类,也只需重新单击参数值按钮,重新根据需求设定即可。需要注意的是,若参数更改,则分词器需要进行重新的训练[15]。系统设计的另一个重点就是特征选择方式的设计,本系统进行选择处理的方式分两种,分别是整体选择和分类选择。所谓整体的选择方式,就是将文本中的词按照其根本的词性特点,将其放入统一的数据库中,通过固定的算法,根据其特征词进行筛选,最后按照一定的格式排列起来。而第二种分类选择方式,就是将中文文本中段落细分为类来处理,通过分析,按照权重,词性等因素划分完成后对其分别经行评估。根据评估的的结果,分类放入到数据库中,其最终也是以数据表的形式存在。最后根据实际需求进行调用

和选取。

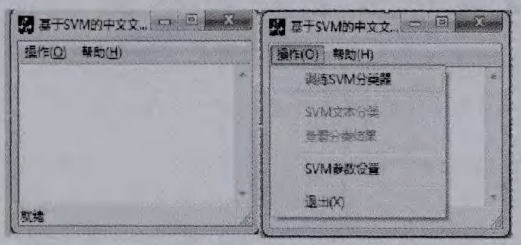

图2　系统界面

5　结　论

文中在充分研究了传统中文文本分类系统的基础上，设计并开发了一个效率高，分类精确的中文文本分类系统，即基于SVM的中文文本分类系统。通过对SVM方法的概述，系统的可行性分析介绍，系统的详细设计，特别是对界面模块以及特征选择模块进行了详细的说明。很好的解决了传统中文文本分类方式所面临的问题，大幅度的提高了文本分类的效率和准确性，达到了设计要求。

参考文献：

[1] 袁彦芹. 基于支持向量机的大规模文本分类研究与设计[D]. 山东：山东师范大学，2007.

[2] 都云琪，肖诗斌.基于支持向量机的中文文本文档自动分类研究[J]. 计算机工程，2002(11)：137-138.

[3] 王永成. 中文信息处理技术及基础[M]. 上海：上海交大出版社，1999.

[4] 成颖，史九林. 自动分类研究现状与展望 [J]. 情报学报，1999，18(2)：20-26.

[5] 王闰强，胡铁军. 中文文本文档自动分类研究进展[J]. 医学情报工作，2002(6)：342-347

[6] 叶新明，徐进鸿. 中文文献自动分类研究 [J]. 情报科学，1992，13(5)：31-34.

[7] 朱兰娟. 中文文献自动分类的理论与实践 [J]. 情报科学，1987，6(6)：433-437.

[8] 肖明，沈英. 自动分类研究进展. 现代图书情报技术[J]. 2000，5(3)：25-28.

[9] 田军. 图书自动分类的数学建模型及实现[J]. 图书情报工作，2001，9(2)：44-47.

[10]李晓黎，刘继敏，史忠植. 概念推理网及其在文本分类中的应用[J]. 计算机研究与发展，2000:37.

[11]邹涛，王继成，黄源等. 中文文档自动分类系统的设计与实现[J]. 中文信息学报，1999，13(3)：124-157.

[12]陶兰，申军霞.文本信息自动分类系统[J]. 中国农业大学学报，1999，4(4)：341-357.

[13]马忠宝. 基于支持向量机的中文文本分类系统研究[D].武汉：武汉理工大学，2006.

[14]朱德熙. 语法讲义[M]. 上海：商务印书馆，1982.

[15]周程远. 中文自动分词系统的研究与实现[D]. 上海：华东师范大学，2008.

(上接第138页)

助解析XML文件、动态添加组件等方法完成拼写单词或释义等操作，该软件也具有随机生成单词列表、自行添加单元、单词、词库管理等功能，设计界面美观，客户端运行稳定，获得出色的用户体验。

参考文献：

[1] 潘香萍. 基于Skype的网络英文交际文本分析与研究[J]. 科技视界，2014，25(19)：170-170，186.

[2] 张爽，朱志良，于瑞云等. 软件工程课程全英文教学模式研究[J]. 计算机教育，2013，35(22)：55-57，61.

[3] 郑深. 基于Flash益智游戏“学字母，打气球”的设计与开发[J]. 软件工程师，2014，17(8)：27-29，20.

[4] 曾利，李自力. 英文"Internet of Things"研究热点与趋势分析[J]. 国防科技，2015，36(5)：100-109.

[5] 康来松，杜晖，沈奕娜等. 中华文化英文BBS系统开发研究[J]. 计算机技术与发展，2013，11(5)：17-21.

[6] 高天寒，郭楠. 以现代密码学与加解密技术为基础的全英文教学模式[J]. 计算机教育，2013，33(24)：50-52.

[7] 刘慧云，曾加劲. 基于统计分析的英文影视词汇习得研究[J]. 教育导刊(上半月)，2014，23(12)：72-75，76.

[8] 戴光荣，宋玉春. 哈希算法与语义映射在语料库对齐中的运用[J]. 福建工程学院学报，2014，16(5)：454-458，463.

[9] 王天剑. 基于语料库的英文软件EULA格式与语言分析[J]. 河北北方学院学报(社会科学版)，2015，31(1)：12-16.

[10]曹琳，汤静芳，程张根等. 基于字幕语料库的英文电影教学模式初探 [J]. 安徽商贸职业技术学院学报 (社会科学版)，2014，13(2)：77-80.

[11]王思鹏，田萍芳，丁胜等. 基于Android的自助式英文学习软件设计与实现[J]. 现代计算机(专业版)，2013，9(2)：69-72.

[12]张巍，汪雪锋，朱东华等. “主题词簇”方法研究——英文科技文献主题词清洗、合并与聚类[J]. 科学学研究，2013，31(11)：1615-1622.

[13]康卉，史子明. 匿名约束网络反馈平台在EFL写作教学中的实证研究[J]. 现代教育技术，2014，24(9)：65-71.

[14]江业峰，姚红岩，王瑞等. Visual Basic程序设计全英文授课的现状与发展思路——以辽宁科技大学为例[J]. 软件工程师，2015，18(9)：9-10.

[15]宋容嘉，杜晖，曹玉玺等. 基于WordPress的中华文化英文博客的设计与实现[J]. 计算机与现代化，2013，15(7)：217-219，223.