文章编号:1006-2475(2016)12-0102-05

基于 AdaBoost 算法的养老信息筛选及应用

程光洋[1],廉　彬[2]

(1.合肥工业大学工业与装备技术研究院,安徽 合肥 230009; 2.安徽省经济信息中心,安徽 合肥 230001)

摘要:面对信息社会中老年人对养老信息的关注与需求,本文使用基于 Python 的网络爬虫技术对民政部网站的新闻和公文进行抓取。针对门户网站的新闻特点,对数据抓取过程以及训练集进行优化,使用 AdaBoost 算法对给定的文本集合进行训练,得到筛选模型。提供一种有效的特征选择方法,采用 χ^2 统计量准则,有效降低了特征维数,然后用该模型对采集的信息进行筛选得到养老信息。最后,对信息筛选结果进行了分析。实验分析结果表明,本文提出的方法可以实现对养老信息的有效筛选,在应用上可以满足老年人对养老信息的获取需求。

关键词:网络爬虫; AdaBoost; 养老信息; 政府新闻; 信息筛选

中图分类号:TP391.1　　**文献标识码**:A　　**doi**: 10.3969/j.issn.1006-2475.2016.12.021

Filtering and Application of Aged Information Based on AdaBoost Algorithm

CHENG Guang-yang[1], LIAN Bin[2]

(1. Institute of Industry & Equipment Technology, Hefei University of Technology, Hefei 230009, China;
2. Anhui Economic Information Center, Hefei 230001, China)

Abstract: Facing attention to the needs of older persons in the information society for aged information, this paper uses Web crawler technology based on Python to crawl the news and official documents from Ministry of Civil Affairs website. Aiming at the characteristics of news on portals, the paper optimizes data fetching process as well as the training set, uses Adaboost algorithm to train a given collection of text and get filtering model. And the paper provides an effective feature selection method which uses the χ^2 statistic principles, effectively reduces the feature dimension, and then uses this model to filter the collection information to get aged information. Finally, the results of information filtering are analyzed. The experimental analysis results show that the proposed method can effectively filter the aged information and meet the elderly demand of aged information acquisition in the practical application.

Key words: Web crawler; AdaBoost; aged information; government press; information filtering

0　引　言

社会的快速老龄化正成为我国的一大特征,随着互联网的快速发展和普及,人们通过网络便可获取海量信息,而面对信息数据的纷繁复杂,从中找到想要的知识需要消耗大量时间。对于老年人来说,如果想在这种信息环境下关注养老有关的新闻和政策越发感到无力。民政部在《“十二五”养老服务体系建设规划(2011-2015 年)》中提出,加强养老服务信息化建设,依托现代技术手段,为老年人提供高效便捷的服务,规范行业管理,不断提高养老服务水平。所以利用现代技术手段整合社会资源,为老年人提供权威有效的养老信息就成为一个值得研究和探讨的课题。

目前国内养老服务体系研究中,大部分的信息技术都是为专门的组织机构设计,或者应用于一些专门的养老机构[1-2]。国内外学者一般围绕如何搞好信息化技术发展而展开[3],对老年人的实际需求大多数仅局限于老年人的健康和安全方面,针对老年人的个性化信息需求存在缺口和不足[4],同时也忽视了老年人对信息的主动获取需求。

本文以民政部网站为例,通过网络爬虫技术对其进行抓取,使用 MySQL 数据库对采集的非结构化数据进行存储,并利用 AdaBoost 分类器筛选出养老相关的文章。在对筛选的结果进行分析后,可得到历年养老文章数量变化分布情况、关键词分布排行以及文章来源分布等一系列具有综合分析意义的信息,从而

收稿日期:2016-05-05

作者简介:程光洋(1991-),男,安徽舒城人,合肥工业大学工业与装备技术研究院硕士研究生,研究方向:智能计算理论与软件; 廉彬(1966-),男,安徽省经济信息中心,硕士,研究方向:软件体系结构。

可以给老年人提供重点和指导信息,也有利于提高政府及相关社会机构的养老服务水平。这对于提升老年人的个性化信息需求服务体验,推动养老信息化进程具有重要意义。

1　网站数据采集

从商业利益、网络信息泛滥等方面考虑,与传统数据获取方式相比,用爬虫程序得到权威网站的数据更为安全和真实,时效性更高。随着养老信息数据量的不断增加,使用爬虫系统实现动态抓取和分析更是降低成本、提高效率的可靠途径。

1.1　数据库表字段框架设计

本文采用 MySQL 数据库对采集的数据和筛选结果进行结构化存储。定义存储的文章对象有:文章标题、正文源码、发文时间、公文号、公文关键字、文章页面链接、网页源码、文章的初始引用来源,其中公文号和公文关键字可为空。UUID 使数据库每条记录都能有唯一的辨识信息,而不需要通过中央控制端来做辨识资讯的指定。设 doc_id 为主键,不需考虑数据库数据写入时的名称重复问题。数据库表字段框架如表 1 所示。

表 1　数据库表字段框架

字段	类型	长度	NULL	存储对象
doc_id	varchar	255	not	(主键)
doc_title	varchar	255	not	文章标题
doc_content	mediumtext	0	not	正文源码
doc_date	date	0	not	发文时间
doc_num	varchar	30		公文号
doc_keyword	varchar	50		关键字
doc_url	varchar	255	not	页面链接
doc_source	mediumtext	0	not	网页源码
doc_domain	varchar	50	not	文章来源

1.2　Python-Scrapy 抓取数据

Scrapy 是一种用 Python 语言实现的爬虫框架,采用 Twisted 异步网络库来处理网络通讯。本文中自定义的爬虫类继承自 CrawlSpider 类,可实现迭代爬取。针对民政部网站的特点,根据所要获取的养老相关信息,对减灾救灾、区域地名、党风建设以及婚姻登记等栏目直接进行过滤。与全网抓取相比,不仅提高了数据采集效率,也减少了分类所需时间。

Scrapy 框架的数据流,先从初始队列中的 URL 地址获取相应页面内容后返回给爬虫(Spiders),爬虫分析出来的结果有 2 种:1)需要进一步抓取的链接,放回链接队列中;2)需要保存的数据,被送到项目管道(Item Pipeline)进行数据清洗。解析网页是数据清洗过程中最为核心的地方,通过自定义 parse 方法来处理 HTTP 响应的 Response 对象,使用 XPath 表达式和正则表达式相结合来提取网页中的对象,再传递给存储数据的 Item 容器,然后通过项目管道(Pipeline)将清洗后的结构化数据保存到 MySQL 数据库对应表中。如果此时链接队列中还有未响应的 URL 则继续进行抓取操作。该爬虫算法的逻辑框图如图 1 所示。

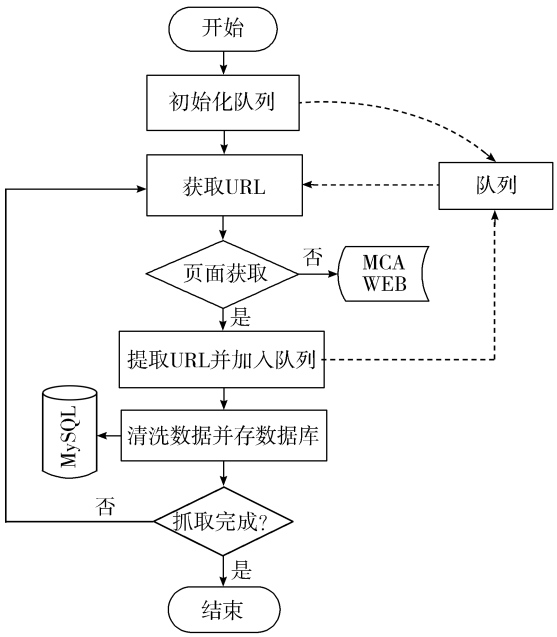

图 1　爬虫算法的逻辑框图

2　数据库信息筛选

从采集的民政部网站数据中筛选出的养老相关信息实际上是两类文本分类问题[5]。所谓文本分类,是根据某分类算法和预定义的类别标号将待分类文本归类。数据库文本分类过程如图 2 所示。其中训练样本集的建立[6]充分考虑了民政部网站新闻特点,对正例集(即养老信息集)进行基于 TF-IDF 算法的关键词抽取,校验相关性文本的选取质量;反例集除了来源于各类新闻源外,还将网站中与养老信息不相关却又特征不明确的文章加入到反例集,这是在不断优化过程中得到的经验样本。

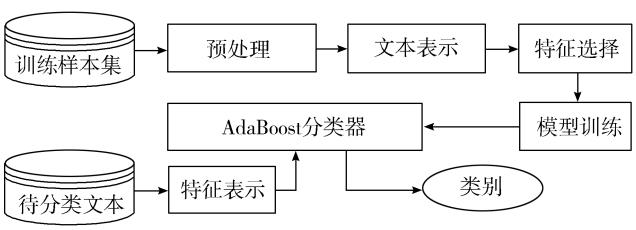

图 2　数据库文本分类过程

在预处理阶段,对存储的信息是否对称进行检查,纠正个别文章标题空置,以及剔除正文源码部分无内容的记录。对于每条记录的文本内容还需要进

行中文分词处理。

2.1 中文分词

机器学习算法只能作用在数值数据上,算法期望使用定长的数值特征而不是不定长的原始文本文件,本文先将文本数据集转换成数值数据集。对中文文本先通过中文分词技术从文本内容得到字符串列表,再用向量空间模型(Vector Space Model,VSM)表示文本。其中在预处理阶段便完成了中文分词的过程。

本文采用基于 Python 的中文分词组件结巴(Jieba)分词工具,对文本进行中文分词和词性标注。结巴分词的算法是基于前缀词典实现高效的词图扫描,生成句子中汉字所有可能成词情况所构成的有向无环图(DAG);采用动态规划查找最大概率路径,找出基于词频的最大切分组合;对于未登录词,采用基于汉字成词能力的 HMM 模型。主要功能包括分词、添加自定义词典、关键词提取、词性标注、并行分词以及命令行分词等。与目前比较流行的中科院 NLPIR 分词工具相比,结巴分词不仅有优秀的中文分词结果,还提供了方便的使用接口,能使代码在设计上做到简洁。

2.2 文本表示

计算机并不具有人类的智慧,不能读懂文字,所以必须先将文本转化成计算机能够理解的形式,即进行文本表示。目前主要是用 Gerard Salton 和 McGill 于 1969 年提出的向量空间模型(Vector Space Model,VSM)来表示文本,一个文本对应一个向量,其基本思想是把文档简化为以特征项的权重为分量的向量表示[7-8]:$(w_1, w_2, \cdots, w_m)$,其中 w_i 为第 i 个特征项的权重。本文选取词作为特征项,用词频表示权重,用特征向量来表示文本,分别计算文本中每个特征出现的次数作为特征向量空间中每维特征的值[9]。

2.3 特征选择

数据库中存储了大量信息,并且每段信息的文本词汇也不相同,这就使得表示文本样本集的特征向量空间的维数过大。很多像标点符号、虚词、副词以及助词等,还有长度为 1 的词,这些词对文本分类是不需要的,除了耗费计算资源,也会引起“过拟合问题”而影响分类效果,因此有必要对分词词汇进行特征筛选来减少向量空间的维数[10-11]。

目前比较常用的特征筛选方法有信息增益(Information Gain)、互信息(Mutual Information)、词频(Document Frequency)以及 χ^2 统计量等[12]。

本文采用 χ^2 统计量方法进行特征选择[13]。χ^2 统计用来度量两者(特征值和类别)独立性的缺乏程度,χ^2 越大,独立性越小,相关性越大(若 AD < BC,则类和特征项独立,N = A + B + C + D。χ^2 统计量(Chi-Square Statistic,CHI)特征选择方法又被称作开方拟合检验(CHI,2-test),这个概念来自列联表检验(contingency table test),用来衡量特征 t 与类别 c 之间的统计相关性。

χ^2 是数理统计中的一个统计量。设 t 为某个词,c 为某个类别,计算公式如下:

$$\chi^2 = \frac{N(AD - BC)^2}{(A + C)(B + D)(A + B)(C + D)} \tag{1}$$

其中,A 是 t 和 c 共同出现的频数,B 是训练语料中包含 t 的文档,且该文档不属于 c 类的频数,C 是 c 类文档中不包括 t 的文档频数,D 是出现既不包括 t 也不属于 c 类的文档频数,N 为训练语料中的总文档树。当 $\chi^2(t,c) = 0$ 时,t,c 独立。

此外,通过词性标注过滤法以及去停用词清理法(采用哈工大停用词表)对特征词向量空间的维度进行选择性降低,有效提高了分类效率和精度。

2.4 构建分类器

AdaBoost 算法是 Freund 和 Schapire 在 1995 年根据在线分配算法提出的,与 Boosting 算法不同的是,AdaBoost 算法不需要预先知道弱学习算法正确率的下限,并且最后得到的强分类器的分类精度依赖于所有弱分类器的分类精度[14]。基于该算法的分类器分析过程如下:

当给定的训练样本集 S 为:$\{(X_1, Y_1), \cdots, (X_m, Y_m)\}$时,其中:$X_i \in X$ 表示 X 中第 i 个元素,$Y_i \in \{-1, +1\}$表示与 X_i 对应元素的属性值。在训练数据中的每个样本都分配一个权重,这些权重构成了向量 D。本文所选正反样本集数目相同,故对初始化训练样本的权重 D(i)设置为$\frac{1}{m}$,即每个训练样本的权重相同。设 T 为训练最大循环次数,在训练数据上计算该弱分类器在权值为均等 D_t 条件下的错误率 ε_t 为:

$$\varepsilon_t = \sum_{i=1}^{m} D_t(X_i) P Y_i \neq h_t(X_i) P \tag{2}$$

在分类器的第二次训练中,将会调整每个样本的权重,其中第一次分类正确样本的权重将会降低,而第一次分类错误的样本的权重将会提高。为了从所有弱分类器中得到最终的分类结果,AdaBoost 算法为每个分类器都分配了一个权重值 α,这些 α 值是基于每个弱分类器的错误率进行计算的,其计算公式为:

$$\alpha_t = \frac{1}{2}\ln\left(\frac{1 - \varepsilon_t}{\varepsilon_t}\right) \tag{3}$$

计算出 α 值后可以对权重向量进行更新,使得那些正确分类的样本的权重降低,而错分样本的权重升高。更新样本的权值公式如下:

$$D_{i+1}(i) = \begin{cases} \dfrac{D_t(i)e^{-\alpha}}{Sum(D)}, & 若\ Y_i = h_t(x_i) \\ \dfrac{D_t(i)e^{-\alpha}}{Sum(D)}, & 若\ Y_i \neq h_t(x_i) \end{cases} \tag{4}$$

在计算出新样本的权值后,该算法又开始下一轮迭代,直至训练错误率为 0 或弱分类器的数目达到用户指定的值为止。最后,将由各个弱分类器投票得到强分类器[15-16],计算公式如下:

$$H(x) = \text{sign}\left(\sum_{t=1}^{T}\alpha_t h_t(x_i)\right) \quad (5)$$

训练样本集经过文本表示以及特征降维后,将得到的特征向量和类标签作为 AdaBoost 算法训练函数的输入,通过训练后得到筛选器模型,然后可使用此模型对未知文本进行分类。通过自定义设置最大循环次数 T 来决定弱分类器的数目,为了对分类器的性能进行测试,对训练数据集采用十折交叉验证的统计方法,得到的分类精度和朴素贝叶斯分类器比较如图 3 所示。

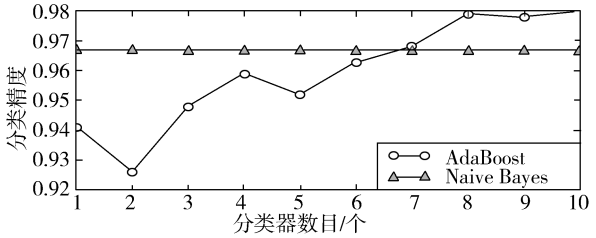

图 3　分类器精度比较

由图 3 可以看出,分类器个数为 2 时,分类器的精度反而降低,这是因为上一分类器的分类精度过高而使这一分类器又很关注那些错误的数据,从而造成整体的分类效果偏低。在第 3 章的验证中,分类精度得到强有力提升,在经过 8 次提升后,分类精度基本稳定在 97.8%,这比朴素贝叶斯分类器分类精度的 96.7% 还要高出 1.1%。随着待分类文本的逐渐增加,这种精度的提升会更可观。

3　实　验

3.1　实验数据

本实验从实际应用考虑,数据全部来源于网络爬虫在民政部网站的抓取结果。数据爬虫环境在 Windows 7 的 64 位系统下进行,其中 Python 版本为 2.7.10rc1,Scrapy 版本为 1.0.1,MySQL 数据库版本为 5.6,Jieba 分词工具版本为 0.38,所有程序由纯 Python 开发实现。

截止到 2016 年 1 月 8 日共采集到 3481 条有效信息,通过已经训练好的分类器对采集的数据库信息进行筛选,最终得到养老相关信息 290 条,其中含文号和关键字的政府公文 58 条,将筛选结果写入数据库新建表中。

在实验中,先抓取后分类的筛选模型比抓取时对每条信息进行筛选的方式效率要高很多。采用前者还可以对数据进行二次分析,发掘更多潜在的规律。运用 AdaBoost 算法对养老信息进行筛选属于文本分类的问题,所以最终评估的主要标准是分类的精度。

1)分类精度:

$$A_m = \frac{N_1}{N_1 + N_2}$$

2)漏分率:

$$B_m = \frac{N_2}{N_3}$$

其中,N_1 表示分类结果中将养老信息归为养老信息的数量,N_2 表示分类结果中将养老信息归为非养老信息的数量,N_3 表示分类为非养老信息的数码。为了更好验证 AdaBoost 算法在养老信息筛选,特别是随着待分类数目增加情况下的优良性能,实验中把该算法分类结果与朴素贝叶斯分类结果进行了比较,比较结果如表 2 所示。

表 2　分类结果比较

	分类耗时/s	漏分率/%	精度/%
AdaBoost	1642	0.38	95.87
Naive Bayes	1487	0.56	93.33

由表 2 的分析可知,AdaBoost 算法在精度和漏分率方面要优于朴素贝叶斯算法,随着测试样本容量的增加,将会影响应用系统的体验。在分类效率方面,朴素贝叶斯算法则更胜一筹。从最终的分类结果来看,基于 AdaBoost 算法的分类结果中还包含少量朴素贝叶斯分类结果中没有的内容,这些记录间接地与养老信息相关。因为前者是根据多个弱分类器的权重来投票决定最终结果,这样不容易错漏与养老信息有关的文章。

3.2　结果分析

通过对筛选结果进行分析,不仅可以对养老信息筛选的实际效果进行满意度评估,还可以从已存在的养老数据中发掘潜在规律,为老年人提供更多的信息指导和服务。

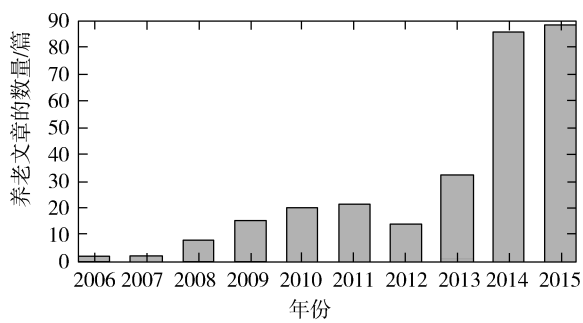

图 4　民政部 2006 ~ 2015 年间养老文章的分布

在对每篇文章的发文时间进行统计后,得到民政部 2006 ~ 2015 年间养老文章的分布情况如图 4 所示。通过观察发现,从 2013 年开始,社会对于养老问题的关注程度已经有了明显提升,而在 2014 年出现了爆发式的增长,并继续保持较高的关注程度。根据

国家统计局网站公布的数据显示:2011 ~ 2014 年全国 65 岁及以上的老年人口占比分别为:9.12%,9.39%,9.67%,10.06%,根据中国国家卫生和计划生育委员会统计显示:目前中国 60 岁以上人口约有 2.12 亿人,占总人口的 15.5%。可见,随着人口老龄化程度的不断加深,社会对养老问题的关注度也在增加,显示出前所未有的重视。中国的老龄化人口没有下降的趋势,可以预测在未来几年养老问题都将成为社会和政府所关注的热点。

为更清楚了解筛选的养老信息是否能满足老年人的信息需求,实际效果的满意度如何,除了关注整体趋势外,还要对数据库所有养老文章进行量化统计。

通过对每条数据库信息的正文源码字段所存储的文本信息进行关键词抽取(设定抽取数量为 λ 个),得到多组包含特定数量关键词的字符列表,再对所有代表文本信息的关键词列表进行词频统计,选出代表整体筛选信息的关键词。因为文本数量关系,会造成统计结果的关键词过多而失去意义,所以需设定一个阈值来对关键词进行调节,将词频小于阈值的关键词剔除,最终获得统计词频靠前的关键词,用这些词表示文本的内容重心。图 5 所示为信息筛选结果的关键词分布情况。

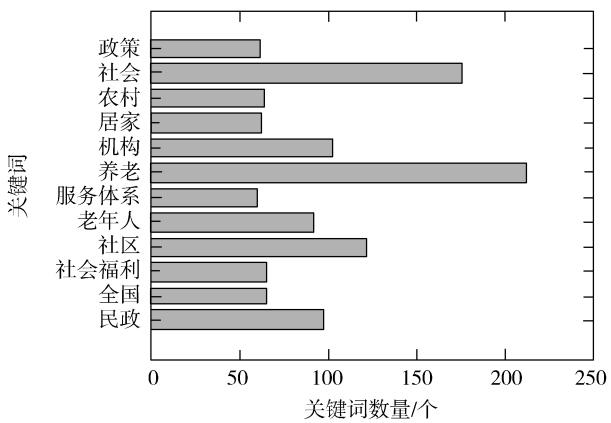

图 5 信息筛选结果的关键词分布

由图 5 可知,养老、老年人以及与养老相关的关键词均出现在靠前位置,而在目前的养老方式中,机构养老、社区养老以及居家养老这些关键词均分布在热点关键词中。由此说明了筛选结果的可靠性,以及在养老信息服务应用上的可行性。除此之外,还出现了一些服务类的关键词,这也是养老信息中对于养老服务建设的反映。

考虑到老年人对养老信息有主动获取的需求,为了能向他们推荐可靠的信息获取途径,通过对文章的来源进行统计,以便更好了解哪些地方在养老方面更为活跃,这就使得老年人能更高效地从推荐的信息通道获取到养老相关信息。图 6 所示是从所有信息来源中,统计出的文章主要来源分布情况。

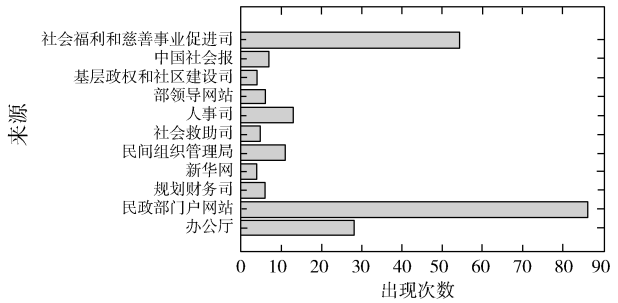

图 6 文章主要来源分布

通过对信息的筛选结果的一系列分析后,可以看出养老本身的关注度随着社会老龄化程度的加深也在不断提高。对结果进行关键词统计表明,筛选结果可以满足老年人对于养老信息的关注需求,其在为老年人信息服务的应用上也是可行的。最后还分析了文章主要来源的分布情况,给老年人提供更多的信息指导。

4 结束语

本文在社会老龄化背景下,针对信息社会中老年人对养老信息的关注需求,通过网络爬虫技术对政府网站新闻进行采集,介绍了一种基于 Python 的爬虫框架 Scrapy,设计了 MySQL 数据库字段和存储方式,运用 AdaBoost 算法对采集的信息进行筛选,得到人们所需要的养老信息,最后对结果进行了有效分析,完成整个数据采集与处理的过程。实验分析结果表明,本方法在养老信息筛选中具有较为满意的效果,可以应用在老年人信息服务上。

信息技术在老年人服务上的应用越来越广泛,养老服务的发展进入了多样化和个性化阶段,但实验中仅以民政部网站为例,抓取的信息量有限。今后的研究将对更多的政府网站新闻进行采集,建立政府养老信息采集系统,研究完整的为老年人服务的信息化平台。同时还需对训练集进行迭代训练以实现更好的分类结果,对程序进行优化以实现更高效的处理和分析。

参考文献:

[1] 李洪心,李巍. 国内外养老模式研究[J]. 经济与管理,2012,26(12):18-22.

[2] 刘晓梅. 我国社会养老服务面临的形势及路径选择[J]. 人口研究, 2012(5):104-112.

[3] 张丽雅,宋晓阳. 信息技术在养老服务业中的应用与对策研究[J]. 科技管理研究, 2015,35(5):170-174.

[4] 李璞. 老年人居家养老管理系统的研究及实现[D]. 兰州:兰州大学, 2015.

[5] 樊兴华,孙茂松. 一种高性能的两类中文文本分类方法[J]. 计算机学报, 2006,29(1):124-131.

(下转第 110 页)

```
private void setData(int count) {
ArrayList < String > xVals = new ArrayList < String > ();
for (int i = 0; i < count; i + + ) { xVals. add(xArray
[i]); }
ArrayList < Entry > yVals = new ArrayList < Entry > ();
for (int i = 1; i < count; i + + ) {
float val = Float. parseFloat(yArray[i - 1]);
yVals. add(new Entry(val, i));
}
// 创建 LineDataSet 并初始化
LineDataSet set1 = new LineDataSet(yVals, "");
set1. setLineWidth(1.5f);
set1. setCircleSize(4f);
// 创建 LineData 变量
LineData data = new LineData(xVals, set1);
mChart. setData(data); // mChart 赋值
mChart. invalidate();// mChart 刷新
}
```

3 结束语

本文从农业种植管理的实际需求出发,对传统农业种植管理系统进行改进,通过移动智能终端技术的应用,为农业种植管理提供了一个简单、快捷的采集与管理的渠道。农业种植管理 APP 充分利用移动智能终端的智能、无线的设备优势,结合二维码技术的应用,实现了农业种植管理工作的合理优化。通过采用 Restful Web Service 架构,实现了 APP 与服务端的数据交互;通过调用终端设备扫描装置,实现了二维码的自动扫描和信息的快速采集;通过 MPAndroidChart 类库的使用,为 APP 的开发提供了良好的人机交互界面。该 APP 为农业种植过程信息的移动采集与管理提供了一个可行的途径。

参考文献:

[1] 卫荣,王秀东. 我国种植业主要农产品生产现状及对策建议[J]. 农业经济, 2015(1):42-44.

[2] 宋艳,程改兰. 基于物联网技术的农业种植环境监控系统设计[J]. 电子设计工程, 2014,22(8):101-103.

[3] 金炜,顾玉琦,陈浩. 基于物联网的农产品种植监控与质量安全溯源[J]. 安徽农业科学, 2014,42(30):10788-10790.

[4] 鲁帆. 移动智能终端发展趋势研究[J]. 现代传播, 2011(11):139-140.

[5] 王跃,肖丽. 移动智能终端技术架构模型研究[J]. 现代电信科技, 2013(6):13-23.

[6] 马金平,陈彦珍,王佳,等. 二维码追溯技术在葡萄栽培及葡萄酒上的研究与应用[J]. 北方园艺, 2015(21):205-207.

[7] 杨彦. 基于 RFID 和二维码技术的农产品溯源商务平台建设的探讨[J]. 浙江农业科学, 2013(9):1218-1222.

[8] 张玉清,王凯,杨欢,等. Android 安全综述[J]. 计算机研究与发展, 2014,51(7):1385-1396.

[9] 倪红军. 基于 Android 平台的消息推送研究与实现[J]. 实验室研究与探索, 2014,33(5):96-100.

[10] 杨林楠,郜鲁涛,林尔升,等. 基于 Android 系统手机的甜玉米病虫害智能诊断系统[J]. 农业工程学报, 2012,28(18):163-168.

[11] 江燕良. 基于 Android 智能终端的远程控制系统[J]. 电子技术应用, 2012,38(8):129-132.

[12] 李光明,孙英爽,党小娟. 基于安卓的远程监控系统的设计与实现[J]. 计算机工程与设计, 2016,37(2):556-561.

[13] 黄晓沛,白恩健,邓美琛,等. 基于 Android 智能终端的环境监测系统设计[J]. 信息通信, 2014(8):40-41.

[14] 程涛,毛林,毛烨. 农产品质量安全追溯智能终端系统的构建与实现[J]. 江苏农业科学, 2013,41(6):273-275,282.

(上接第 106 页)

[6] 张启蕊,张凌,董守斌,等. 训练集类别分布对文本分类的影响[J]. 清华大学学报(自然科学版), 2005,45(S1):1802-1805.

[7] 李惠娟,高峰,管晓宏,等. 基于贝叶斯神经网络的垃圾邮件过滤方法[J]. 微电子学与计算机, 2005,22(4):107-111.

[8] 庞剑锋,卜东波,白硕. 基于向量空间模型的文本自动分类系统的研究与实现[J]. 计算机应用研究, 2001,18(9):23-26.

[9] 伍洋,钟鸣,姜艳,等. 面向审计领域的短文本分类技术研究[J]. 微电子学与计算机, 2015,32(1):5-10.

[10] Guo Qiang. Research and improvement for feature selection on Naive Bayes text classifier[C]// 2010 2nd International Conference on Future Computer and Communication (ICFCC). 2010:V2-156-V2-159.

[11] 姜百宁. 机器学习中的特征选择算法研究[D]. 青岛:中国海洋大学, 2009.

[12] Yang Y, Pedersen J O. A comparative study on feature selection in text categorization[C]// Proceedings of the 14th International Conference on Machine Learning. 1998:412-420.

[13] 裴英博,刘晓霞. 文本分类中改进型 CHI 特征选择方法的研究[J]. 计算机工程与应用, 2011,47(4):128-130.

[14] 董乐红,耿国华,周明全. 基于 Boosting 算法的文本自动分类器设计[J]. 计算机应用, 2007,27(2):384-386.

[15] Winata G I, Khodra M L. Handling imbalanced dataset in multi-label text categorization using Bagging and Adaptive Boosting[C]// 2015 International Conference on Electrical Engineering and Informatics(ICEEI). 2015.

[16] Yoon Y, Lee G G. Text categorization based on boosting association rules[C]// The IEEE International Conference on Semantic Computing. 2008:136-143.