

基于支持向量机的中文农业文本分类技术研究^{*}

魏芳芳¹ 段青玲^{1,2} 肖晓琰¹ 张磊¹

(1. 中国农业大学信息与电气工程学院, 北京 100083; 2. 北京市农业物联网工程技术研究中心, 北京 100097)

摘要: 高效地组织、分类信息,是提供个性化农业信息推荐服务的基础。根据农业文本信息特点,提出了一种基于线性支持向量机(Support vector machine,SVM)的中文农业文本分类模型,首先构建农业行业分类关键词库,通过特征词选择和权重计算,构建分类器模型,实现信息的自动分类。实验选取了1 071个测试文档,并按照种植业、林业、畜牧业、渔业进行分类。结果表明,分类准确率为96.5%,召回率为96.4%。实验结果高于贝叶斯、决策树、KNN、SMO等分类算法,将该模型应用于农业物联网行业信息综合服务平台,运行结果表明,该方法能够实现中文农业文本信息的自动分类,响应时间满足系统要求。

关键词: 文本分类 支持向量机 中文农业信息 信息整合

中图分类号: TP391 **文献标识码:** A **文章编号:** 1000-1298(2015)S0-0174-06

Classification Technique of Chinese Agricultural Text Information Based on SVM

Wei Fangfang¹ Duan Qingling^{1,2} Xiao Xiaoyan¹ Zhang Lei¹

(1. College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China

2. Beijing Engineering Research Center of Agricultural Internet of Things, Beijing 100097, China)

Abstract: In order to provide personalized services for agricultural information recommendation, it was needed to organize and classify information efficiently. According to the characteristics of agricultural texts, a Chinese agricultural text classification model was proposed based on linear support vector machine (SVM). Firstly, an agriculture-domain-based dictionary was built. Secondly, a feature vector was extracted and the weight for each feature in a vector was selected. Lastly, a text classification model was established. The model was tested on 1 071 documents which were belonged to four classes: planting, forestry, animal husbandry and fisheries. The results showed that the accuracy was 96.5% and the recall rate was 96.4%. Both of their performances were higher than those of the ones using other classification methods, such as the Bayesian, decision tree, KNN, SMO algorithm and neural network. The model was applied to the platform for agricultural internet of things (IOT) industry integrated information service. The performance showed that the method can automatically classify Chinese agricultural text information and the response time met the system requirements.

Key words: Text classification Support vector machine Chinese agricultural information Information integration

引言

近年来,随着互联网技术的飞速发展,人们已经

从信息较为缺乏的时代过渡到信息丰富的数字化时代。在农业信息服务方面,农业信息资源增长迅速,截至到2012年初,我国已有涉农的网站达3万多

收稿日期: 2015-10-28 修回日期: 2015-11-16
^{*} 国家高技术研究发展计划(863计划)资助项目(2013AA102306)、山东省自主创新资助项目(2014XGA13054)和中央高校基本科研业务费专项资金资助项目(2015XD001)
作者简介: 魏芳芳,硕士生,主要从事农业信息整合研究,E-mail: 13598605510@126.com
通讯作者: 段青玲,教授,博士生导师,主要从事智能信息处理研究,E-mail: dqling@cau.edu.cn

家^[1],80% 以上的数据以非结构化的形式存在^[2]。但是农户要从海量的农业资源中获取有用的信息还较困难,通常需要查询大量的网站,而且信息质量参差不齐。因此,实现农业互联网信息整合,将分散于各个网站的农业资源进行融合,最终形成一个有序、高效的信息资源体系,并利用统一的共享平台,为用户提供个性化的服务,是亟待解决的问题。

目前农业网站中,其信息主要包括:资讯类、新技术、价格行情和供求信息等,其中资讯类和新技术通常没有详细分类。而农业信息服务的对象,如农户或企业只关心某一类的农业信息,例如养殖户更需要养殖业的信息。为了给用户提供更加精准的信息,本文重点研究农业文本自动分类技术,根据农业文本信息特征,把分散在各网络中的信息进行整合并按照农业细类进行分类,从而为用户提供个性化信息推送服务。

文本分类技术主要包括文本预处理和分类器构建,文本预处理包括中文分词、特征选择、特征项权值计算;分类器构建是在预处理的基础上选择分类算法,根据特征向量,实现信息分类。文本自动分类的方法有决策树、朴素贝叶斯、KNN (K-nearest neighbor)、神经网络等^[3-6]。国外学者 Kenekayoro 等^[7]使用决策树算法实现了文本的自动分类;Milt sakaki 等^[8]基于贝叶斯分类方法,开发了一个实时的分类系统,能够自动进行文本分类和难度解析;黄娟娟^[9]采用 Apriori 算法针对不同类别的训练样本提取每个类别的频繁特征集及其关联的文本,为未知类别的文本确定适当的近邻数 k ,根据近邻的类别确定未知文本的类别,准确率为 85%;Shen 等^[10]采用神经网络机器学习的方法进行特征选择;Revathi 等^[11]采用遗传算法进行特征选择,采用神经网络的方法进行文本分类;Ertekin 等^[12]则采用 SVM 实现了文本分类。在农业信息分类方面,文献^[13]提出了基于向量空间模型中文农业网页分类技术,准确率为 92.63%,召回率为 91.5%;朱学芳等^[14]抽取了水产行业网站的 450 个网页,基于 KNN 算法实现了文本分类,准确率为 88.4%。可见,中文文本分类技术已有一定的研究,但是现有的分类效果还不够理想,并且在农业信息分类上的应用较少。

本文根据农业文本信息特点,提出一种基于线性 SVM 的中文农业文本分类模型,首先构建农业行业分类关键词库,通过特征词选择和权重计算,构建分类器模型,实现信息的自动分类。

1 研究方法

1.1 基于 SVM 的文本分类模型

农业 Web 文本分类流程主要分为 2 个阶段:训练阶段和测试阶段。训练阶段主要包括训练样本的分词、特征选择、特征项权值计算等处理,最后采用 SVM 算法获得农业文本分类器。测试阶段的测试样本经过同样的处理,再根据训练阶段得到的关键词库进行特征过滤,最后通过训练好的分类器进行分类。农业文本分类处理流程如图 1 所示。

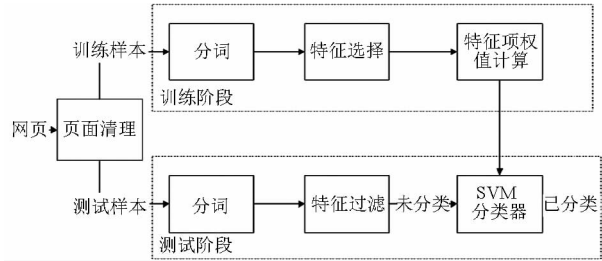


图 1 农业文本分类过程

Fig. 1 Agricultural text classification process

在文本分类中,一般语料库的测试文档中不包含标题,但标题中往往含有文本特征比较明显的词,这些词可以作为特征词,因此获取农业测试文档时都保留了标题信息,更有利于特征选择。获取的文本信息主要来源于中国农业新闻网、河南水产网、农林网等网站,这些信息以 HTML 形式存在,必须去除 HTML 标签、script 脚本、CSS 样式、广告、导航信息,保留标题和正文,然后对训练文本集进行分词、特征选择等后续处理。

1.2 分词和分类关键词库构建

中文文本不同于英文文本,中文文本词与词之间没有明显的分隔,只有句子之间以标点符号来分隔,分词的过程就是按照一定的原则把连续的字序列分成词序列的过程。分词结果需要经过分词、去除停用词、词性标注等处理才能得到。选用基于多层隐马尔可夫模型的汉语词法分析系统 ICTCLAS 对农业信息进行分词,该系统支持分词、词性标注、用户词典和新词识别,分词效果如图 2 和图 3 所示。

根据《国民经济行业分类与代码》^[15],构建如表 1 所示的农业行业分类关键词库,根据该词库调整特征词的文档频率,例如,特征词“牛肉”、“鲫鱼”含有“牛”、“鱼”,调整该特征词的文档频率。

1.3 特征选择和权重计算

特征选择的目的是从原始词序列中找出最具代表性的特征词,构成文本分类的一个特征空间,降低

全国小麦收获过半 夏种工作全面展开
(来源：中国农业新闻网-农民日报)

本报讯当前，华北、黄淮海各小麦主产省相继开镰，全国小麦跨区机收工作由南到北梯次推进，“三夏”生产进入高潮。截至6月9日17时，全国已收获冬小麦19546万亩，超过播种面积50%，夏收小麦进度过半。当日投入联合收割机47.3万台，日机收小麦2398万亩，机收比例达89%。今年大规模小麦跨区机收自5月27日启动以来，总体进展顺利。6月7日-9日，全国小麦日机收面积连续3天超过2300万亩，单日收获机具投放量和日机收小麦面积均创新高，进度明显加快。

图2 农业文本分词实例

Fig.2 Agricultural text word segmentation example

全国/n 小麦/n 收获/v 过半/v 夏种/vn 工作/vn
全面/ad 展开/v (w 来源/n : /w 中国/ns 农业
/n 新闻网/n -/m 农民/n 日报/n)/w 本报/r 讯
/ng 当前/t , /w 华北/ns /w 黄淮海/j 各/r 小
麦/n 主产省/n 相继/d 开镰/v , /w 全国/n 小麦
/n 跨/v 区/n 机收/j 工作/an 由/d 南/f 到/c
北/f 梯次/n 推进/v , /w “w 三夏/t ” /w 生
产/v 进入/v 高潮/n o/w 截至/v 6月/t 9日/t 17
时/t , /w 全国/n 已/d 收获/v 冬小麦/n 19546
万/m 亩/q , /w 超过/v 播种/v 面积/n 50%/xx ,
/w 夏收/n 小麦/n 进/v 度/v 半/m o/w 当日

图3 农业文本分词结果

Fig.3 Agricultural text word segmentation results

表1 农业行业分类关键词库

Tab.1 Key words of agricultural industry classification

一级词汇	种植业	畜牧业	林业	渔业
	谷物、稻谷、玉米、	牲畜、牛、马、	林木、	水产、
	豆、油料、薯、棉、	猪、羊、骆驼、	育苗、	海水、
	麻、糖、烟草、蔬	家禽、鸡、鸭、	造林、	鱼、虾
二级词汇	菜、食用菌、花卉、	鹅	森林、	
	水果、仁果、葡萄、		木材、	
	柑橘、香蕉、坚果、		竹材	
	香料、茶			

文本表示的维度。目前常用的特征选择方法有基于文档频率(DF)、信息增益(IG)、互信息(MI)、期望交叉熵(ECE)、卡方统计量(CHI)等,本文采用DF和特征词计算规则相结合的方法进行特征选择。

在特征选择时,首先采用DF方法计算每个词的文档频率。

假设文档D是包含所有特征项(词)的集合,D(f₁,f₂,⋯,f_n)表示n维向量空间中的一个向量,其中f_k(k=1,2,⋯,n)表示一个特征项,DF计算公式为

$$DF(f_k)=\frac{a(f_k)}{A}$$

(1)

式中 a(f_k)——具有特征项f_k的文本数

A——训练集文本数

计算出文档中特征词的DF值后,根据特征词是否包含农业行业分类关键词进行调整。假设:农

业行业分类关键词的一级词集合为:T₁= {u_i|i=1,2,⋯,n},农业行业分类关键词的二级词集合为T₂= {v_i|i=1,2,⋯,n},选择后的新特征集合为S={f_k|k=1,2,⋯,m},f_k拆分后的特征子集为:S₁= {f_{ki}|k=1,2,⋯,m;i=1,2,⋯,q},其中f_{ki}为f_k拆分后的特征词,m为特征项的个数,q为拆分后的特征词个数。特征项文档频率阈值为[2,6],文档频率调整规则如下

初始化时,S=φ

For i=1,2,⋯,n

IF ∃f_k∧DF(f_k)<2 THEN f_k∉S

IF ∃u_i∧u_i∈S₁ THEN S={f_k,S},DF(f_k)=6

IF ∃v_i∧v_i∈S₁ THEN S={f_k,S},DF(f_k)=5

通过对文档频率的调整能够比较准确地选择特征词,也有利于特征降维。本文词典如图4所示,第1条记录“农庄”为特征词,5.747 831 257 490 882是文档频率。本文特征选择的结果如图5所示,对每个词给出一个唯一的编号,从1开始,词编号和特征词构成一个特征向量,共选择6 572个特征词。

农庄/n 5.747831257490882+
院长/n 4.4951480280892415+
花/v 3.183264589704584+
吃饭/v 4.831588369835958+
显/v 3.9562312600924665+
花/n 3.6686129692524685+
猪场/n 5.0+
团体/n 5.747831257490882+
杜绝/v 4.831588369835958+
芯片/n 5.747831257490882+
灵敏度/n 5.747831257490882+

图4 农业数据词典

Fig.4 Agricultural data dictionary

Instances: 420

Attributes: 6572

Attributes

All

None

Invert

No.	Name
1	首都/n
2	物理/n
3	犹如/v
4	方便面/n
5	冠心病/n
6	西门/n
7	价值/n
8	财政部/n
9	鲈鱼/n
10	回答/v

图5 特征向量文件

Fig.5 Feature vector file

利用选择好的特征词进行TF-IDF权重计算建立相应的文本表示模型,用于衡量该特征词的重要程度。使用已经编号的类别和特征向量,对每个文档计算TF-IDF值,计算结果如图6所示。例如,第1条中的2表示文本类别,9表示特征词,0.0075 535 433 205 147 58表示该词的特征权重,如

果有 N 个特征向量,每条记录就有 N 维向量。

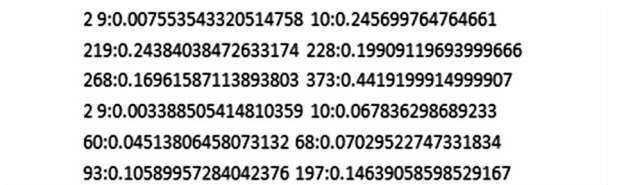


图6 文本特征向量
Fig. 6 Text feature vector

1.4 SVM 分类算法

文本分类为多分类问题,多分类一般转化为二分类问题得到,本文采用的支持向量机(SVM)分类方法的基本思想是:线性可分的情况下,在 n 维空间中找到一个分类超平面,将空间上的 2 类样本分类。SVM 寻求结构风险最小化,从而在统计样本量较少的情况下,获得良好的统计规律。下面给出二分类问题的定义^[16]。假设 SVM 分类训练集 $T = \{(X_i, y_i; i = 1, 2, \cdots, n)\}$,其中样本点 $X_i \in \mathbf{R}^n$,类别 $y_i \in \{-1, 1\}$,搜索 $\mathbf{R}^n$ 上的一个决策函数

$$H(X) = \text{sgn}(g(X)) \tag{2}$$

寻找某样本点 X 对应的 y 值,相当于寻找把 $\mathbf{R}^n$ 上的点分成两部分的规则。如果 $g(X)$ 是线性函数,该分类为线性分类,则采用线性分类模型。在线性可分时,就会存在一个超平面

$$\omega X + b = 0 \tag{3}$$

其中, ω 是常数。

使得训练样本分成两部分,对分类超平面方程归一化,使得 T 满足

$$y_i(\omega X_i + b) \geq 1 \quad (i = 1, 2, \cdots, n) \tag{4}$$

接着将寻找最优超平面转化为求解二次优化问题

$$\min_{w,b} \frac{1}{2} \|\omega\|^2 \tag{5}$$

$$y_i(\omega X_i + b) \geq 1 \quad (i = 1, 2, \cdots, n) \tag{6}$$

然后将二次规划问题转化为对偶问题求解即可。

对于给定未知样本点 X ,计算 $\text{sgn}(\omega X + b)$ 就可以判断 X 所属的类别。多分类问题通过一对一和一对多的方式转化为二分类问题,本文采用 LibSVM 方法分类使用一对一的方式,一对一就是每两类训练一个分类器,对于一个 s 分类问题,会产生 $s(s - 1)/2$ 个分类函数,分类时每个分类器判断类别后,为对应的类别投票,票数最多的为该测试样本的类别。

2 实验与分析

2.1 获取语料库

获取语料库处理流程如下:爬虫程序获取到相

应的网页后,通过 HTML Parser 进行内容解析,建立符合条件的标题 filter 和正文 filter 提取标题和正文,经 HTML Parser 解析后提取的正文可能还含有 script、样式等无用信息,使用 java 的正则表达式将其过滤。最后将解析到的标题和正文文本作为训练语料保存到本地。

测试语料主要来源于中国农业新闻网、河南水产网、农林网,共 1 071 个文档,分为种植业、林业、渔业、畜牧业 4 类,随机抽取训练集 651 个文档,测试集 420 个,训练集与测试集之间的文档互不重叠,如表 2 所示。

表 2 农业文本测试数据
Tab.2 Agricultural text test data

类别	种植业	林业	渔业	畜牧业
训练集	90	50	300	200
测试集	55	32	236	108
总数	145	82	536	308

文本分类模型的评估指标主要包括 5 种,分别是 P_T 、 P_F 、准确率 P 、召回率 R 和 F-measure(F)值。其计算式为

$$P_T = \frac{T_p}{T_p + F_N} \tag{7}$$

$$P_F = \frac{F_p}{F_p + T_N} \tag{8}$$

$$P = \frac{n_t}{n_r} \tag{9}$$

$$R = \frac{n_t}{n_b} \tag{10}$$

$$F = \frac{2PR(\alpha^2 + 1)}{\alpha^2(P + R)} \tag{11}$$

其中,真正类 P_T 是指实例是正类预测也是正类的概率,假正类 P_F 是指实例是负类被预测成正类的概率, T_p 、 T_N 是指被正确分类的正类和负类样本, F_p 、 F_N 是指被错误分类的正类和负类样本, n_t 是分类正确的文本数, n_r 是实际分类的文本数, n_b 是属于该类的文本数, F 是 P 和 R 的加权调和平均。

2.2 实验结果分析

根据获得的语料库,采用 weka 进行实验,首先选用决策树、贝叶斯、KNN、神经网络几种算法分别测试,结果如表 3 所示。

由表 3 得出,SVM 分类方法比 C4.5 决策树、贝叶斯、SMO 的准确率(P)、召回率(R)、F-measure(F)值高,分别达到了 96.5%、96.4%、96.4%。贝叶斯、SMO、C4.5 决策树、SVM 算法明显比 KNN 和神经网络算法分类效果理想,而且 KNN 还涉及最佳 K 值的确定问题,神经网络需要训练学习样本,运行

时间较长。

表 3 农业文本分类测试结果

Tab.3 Agricultural text classification test results %					
算法	P_T	P_F	P	R	F
KNN	56.0	54.4	67.9	56.0	40.9
神经网络	71.4	24.3	70.3	71.4	69.9
贝叶斯	90.0	7.7	89.9	90.0	89.8
SMO	93.8	3.9	94.0	93.8	93.8
C4.5 决策树	94.5	1.8	94.8	94.5	94.4
SVM	96.4	3.5	96.5	96.4	96.4

其次,针对 SVM 算法的不同核函数进行实验,结果如表 4 所示。

表 4 RBF 核函数与线性核函数的比较

Tab.4 Comparison between RBF and linear kernel function %					
算法	P_T	P_F	P	R	F
RBF-SVM	87.1	12.1	88.0	87.1	86.4
线性 SVM	96.4	3.5	96.5	96.4	96.4

由表 4 得出,SVM 分类器的效果还取决于核函数的选择,结果表明,线性 SVM 的分类准确率和召回率均大于 RBF-SVM,因此,线性 SVM 更适合用于文本分类。

再次,对影响分类的因素数据样本规模也进行了研究,针对线性 SVM 算法,将数据量减少进行测试,训练集减少到 448 个,测试集减少为 94 个,测试结果如表 5 所示。

表 5 不同数据规模的线性 SVM

Tab.5 Linear SVM with different data sizes %					
数据	P_T	P_F	P	R	F
小数据集	87.2	7.9	87.8	87.2	86.4
大数据集	96.4	3.5	96.5	96.4	96.4

由表 5 得出,训练集与测试集的比例和数据规模也会影响分类器分类效果。

最后,绘制 7 种算法的比较实验结果图,如图 7 和图 8 所示。

由图 7 和图 8 得出,线性 SVM、C4.5 决策树、贝叶斯、SMO 算法对种植业、林业、渔业、畜牧业 4 种类别分类的准确率和召回率较高,且分类性能稳定,神经网络、RBF-SVM、KNN 算法的分类性能

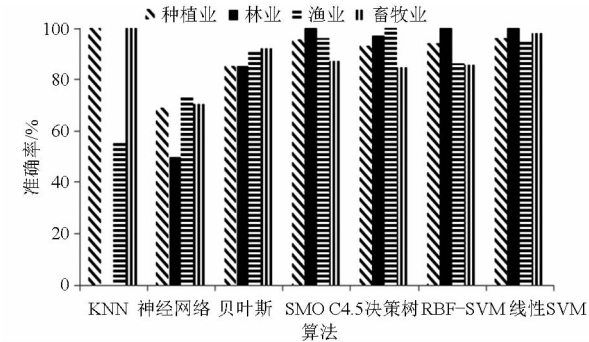


图 7 7 种分类算法的准确率

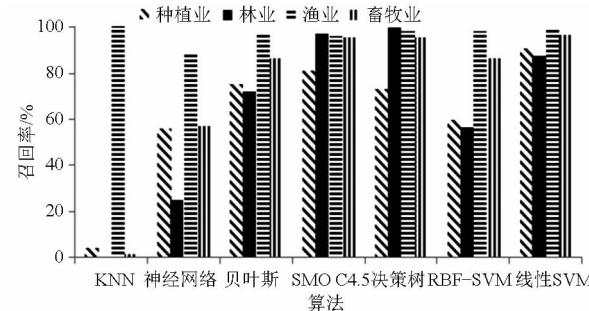


图 8 7 种分类算法的召回率

Fig.8 Recall rates of seven classification algorithms

不稳定且准确率、召回率比较低。

3 结论

(1)构建了一种基于 SVM 的中文农业文本分类模型,根据建立的农业行业分类关键词库和设置的规则调整 DF 值,能够获取到相关度较高的特征词项,构建的 SVM 分类器能够实现文本分类,实验结果证明其准确率为 96.5%,召回率为 96.4%。

(2)对比了 KNN、神经网络、贝叶斯、SMO、C4.5 决策树、RBF-SVM 分类和线性 SVM 算法,结果表明线性 SVM 的分类准确率、召回率、 F -measure 值优于其他分类算法。就分类的稳定性来说,线性 SVM、C4.5 决策树、贝叶斯、SMO 算法优于 KNN、神经网络、RBF-SVM 算法。分类的效果不仅取决于分类算法的选择还要受到语料库构建和样本倾斜的影响。

(3)SVM 分类的训练和预测比较耗时,增加了时间复杂度,今后可以考虑增量学习的方法减少运行时间,提高运行速度。同时,SVM 分类涉及参数选择,可以考虑参数寻优方面的改进。

参 考 文 献

1 郭作玉. 信息化与现代农业发展[J]. 中国信息界,2011(9):5-8.
2 董萍. 基于知网语义关系的中文事件信息抽取技术研究[D]. 西安: 西安电子科技大学,2010.
3 Quoc Le, Tomas Mikolov. Distributed representations of sentences and documents[C] // Proceedings of the 31st International Conference on Machine Learning, 2014.
4 Nemanja Djuric, Vladan Radosavljevic, Mihajlo Grbovic. Hierarchical neural language models for joint representation of streaming

- documents and their content[C]//International World Wide Web Conference Committee (IW3C2),2015;248 – 255.
- 5 Stephen Robertson. Understanding inverse document frequency: on theoretical arguments for IDF[J]. Journal of Documentation, 2004,60(5):503 – 520.
- 6 Keyur J Patel, Ketan J Sarvakar. Web page classification using data mining[J]. International Journal of Advanced Research in Computer and Communication Engineering, 2013(7):2513 – 2519.
- 7 Patrick Kenekayoro, Kevan Buckley. Automatic classification of academic web page types[J]. Scientometrics,2014,101(2):1015 – 1026.
- 8 Eleni Miltsakaki, Audrey Truett. Real-time web text classification and analysis of reading difficulty[C]//Proceedings of the Third ACL Workshop on Innovative Use of NLP for Building Educational Applications,2008: 89 – 97.
- 9 黄娟娟. 基于 KNN 的分类特征选择与分类算法的研究与改进[D]. 厦门:厦门大学,2014.
- 10 Shen Feng, Luo Xiong, Chen Yi. Text classification dimension reduction algorithm for Chinese web page based on deep learning [C]//International Conference on Cyberspace Technology(CCT 2013),2013:451 – 456.
- 11 Revathi N, Anjana Peter, Jagadeesh Kumar. Web text classification using genetic algorithm and a dynamic neural network model [J]. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET),2013,2(2):436 – 442.
- 12 Seyda Ertekin, C Lee Giles. A comparative study on representation of web pages in automatic text categorization[R]. University Park, PA:The Pennsylvania State University,2010.
- 13 段园磊,张太红. 基于向量空间模型中文农业网页分类技术研究[J]. 新疆农业大学学报,2012,35(2):164 – 167.
Duan Yuanlei, Zhang Taihong. Research of on technology Chinese agricultural web page classification based on vector space model[J]. Journal of Xinjiang Agricultural University, 2012,35(2):164 – 167. (in Chinese)
- 14 朱学芳,冯曦曦. 基于文本内容的农业网页信息抽取和分类研究[J]. 情报科学,2012(7):1012 – 1015.
Zhu Xuefang, Feng Xixi. Text oriented information extraction and classification technology for agricultural webs[J]. Information Science, 2012(7):1012 – 1015. (in Chinese)
- 15 GB/T 4754—2011 国民经济行业分类与代码[S]. 2011.
- 16 刘佳. 基于 SVM 的 WEB 中文文本分类系统研究与实现[D]. 西安:西安电子科技大学,2014.