改进的中心向量算法在农业信息分类中的研究

赵新苗,冯向萍,李永可
(新疆农业大学 计算机与信息工程学院,新疆 乌鲁木齐 830052)

摘 要:自 21 世纪以来,农业信息网站开始迅速增加。为了给广大农民朋友和农业科研人员提供方便,需要对农业信息进行分类。将农业信息进行分类有利于农业信息的获取和管理,农业分类的方法有很多种,其中中心法分类相对简单且卓有成效。中心向量计算方法是中心法分类的核心,文中实验目的在于找出效率较高的中心向量计算方法来提高分类的准确率。目前文本类的中心向量计算多数是由该类别文本特征向量的简单算术平均得到的,这样计算得出的中心向量往往会有模型偏差,以至于不能得到很好的分类效果。为解决这个问题,使用总和法、均值法和归一化法计算中心向量,并进行对比实验,结果表明归一化法在查准率、查全率和 F_1 测度都有较好的表现。

关键词:农业信息;分类;中心法;中心向量;文本特征向量
中图分类号:TP301 **文献标识码**:A **文章编号**:1673-629X(2016)08-0146-06
doi:10.3969/j.issn.1673-629X.2016.08.031

Research on Improved Center Vector Algorithm in Agricultural Information Classification

ZHAO Xin-miao,FENG Xiang-ping,LI Yong-ke
(College of Computer and Information Engineering,Xinjiang Agricultural University,Urumqi 830052,China)

Abstract:Since twenty-first century,the site of agricultural information has increased rapidly. In order to provide the convenience for farmers and agricultural researchers,it is need to classify agricultural information. The classification of agricultural information is favor of acquisition and management of the agricultural information. There are several ways to classify agricultural information,in which the centroid-based classification is simple and effective. In this paper,it uses centroid-based classification to find the more efficient one to improve the accuracy of agricultural information. At present,most of the methods for calculating the center vector of the text are the average value of the text feature vector. This method can't get a good classification results due to the model deviation for center vector obtained. In order to solve this problem,the sum method,means method and normalization method is used to calculate the center vector and the result of three methods are compared. The results show that the normalization method has better performance in Precision,Recall and F_1 measure.

Key words:agricultural information;classification;centroid-based method;center vector;text feature vector

0 引 言

自 21 世纪以来,在信息技术迅猛发展的强劲推动下,农业信息化进程明显加快。计算机科学在农业信息领域中发挥着重大作用,农业信息网站开始迅速增加,为广大的农民朋友和农业科研人员提供了极大的方便,但是在众多农业信息中要寻找到自己所需要的信息却面临着极大的挑战。因此如何有效地对农业信息进行分类管理,方便信息的查找成为农业信息化亟待研究的重要领域。

目前使用的网页分类算法主要有 KNN 算法、朴素贝叶斯算法、支持向量机算法和中心法。KNN(K -最邻近算法)最初由 Cover 和 Hart 于 1968 年提出,金一宁等使用该方法对中文网页进行分类,最终分类的准确率达到 80% 以上[1];江小平等采用分布式编程的朴素贝叶斯算法对中文网页文本进行分类,其识别率达到 86%[2];李琼等使用改进的支持向量机算法对文本进行分类,在一定程度上提高了识别的准确率[3]。

中心法是一种相对简单并且高效的文本分类算法,但是其需要满足一个条件,即待分类向量与它所属文本类别的中心向量相似度要大于其他的类别。因此

收稿日期:2015-11-19 **修回日期**:2016-03-04 **网络出版时间**:2016-08-01
基金项目:新疆维吾尔自治区高技术研究发展计划项目(2015X0103)
作者简介:赵新苗(1990-),女,硕士研究生,研究方向为数据库技术;冯向萍,副教授,研究生导师,通讯作者,研究方向为数据库技术及应用。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20160801.0842.020.html

中心法的准确率在很大程度上依赖于中心向量的计算方法,目前类别的中心向量是由该类别文本特征向量的简单算术平均得到的,但是对于各个类别的文本往往是很分散的,空间上也有和其他类别重叠的区域,这样计算出来的中心向量往往会有模型偏差,以至于不能得到很好的分类效果。

针对上述偏差,文中提出一种改进的中心向量计算方法,并进行了实验对比。

1　农业信息网站的现状与研究

1.1　农业信息网站的现状与分类标准

目前国内农业信息资源的建设在总体上存在缺乏专业特色、信息资源缺乏多样性、信息共享和开放程度较低,以及信息的时效性较差等问题。并且在农业信息资源建设方面缺乏科学权威的农业信息搜索引擎,现代先进的信息科学技术还未得到广泛的应用。现代信息科学技术在农业信息资源建设中的作用需要得到重新认识和足够重视,同时还需要加强对关键技术的研究,如信息自动采集与发布技术、农业信息分类检索技术、网络数据库技术等[4]。

虽然先后在农业领域颁布了 1 064 个国家标准,却没有全面的农业信息分类国家标准。根据农业信息分类的原则,将农业信息分为四级,其中包括一级分类 8 项,二级分类 42 项,三级分类 192 项,四级分类 1 136 项[5]。文中使用一级分类标准,具体为林业、种植业及制品、渔业、园林、畜牧业、农业生产资料、农业机械、植物病理共八类。

1.2　农业信息网页分类流程

文中从各大农业网站上获取与农业信息相关的网页样本进行人工标注,并且将标注样本分为训练集和测试集。先进行文本预处理去除文档标签及非中文字符;然后对网页内容进行中文分词,去除停用词;再进行特征词提取,并且构建特征向量;最后构建文本类的中心向量分类器,使用中心法对测试集网页进行测试,进行性能评估。中文农业信息网页分类流程如图 1 所示。

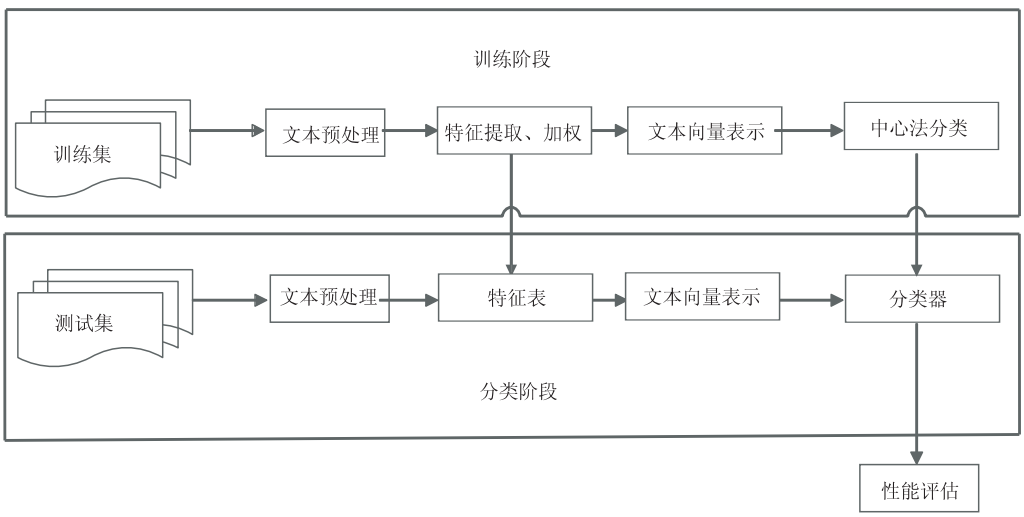

图 1　农业信息网页分类流程

2　主要技术简介

文中使用到的技术主要是分词方法、特征提取、特征加权、中心法。

(1)分词。

对于中文网页分类,分词是很重要的基础。使用高效率的分词方法能够在很大程度上提高网页分类结果的准确性[6]。文献[7-8]对多种分词方法进行了比较,结果表明庖丁解牛分词器从分词效果、性能、准确率方面都表现良好。文中选择庖丁解牛分词器。

(2)特征提取。

在文本分类领域,目前比较常用的特征提取方法有:文档频率(Document Frequency,DF)、互信息(Mutual Information,MI)、期望交叉熵(Expected Cross Entropy,ECE)、信息增益(Information Gain,IG)和 X^2 统计方法(CHI-square,CHI)等。

特征提取的质量将在很大程度上决定分类效果的效率与优劣,因而寻找有效的特征提取方法,不仅能降低文本特征向量的维数,而且可以抑制干扰词语对分类的影响,从而提高分类精度[9]。文献[10-11]对这几种方法都进行了不同程度的介绍,并针对不同的中文语料集通过实验进行分析比较。实验结果表明,CHI 分类效果较好,且在样本数据分布不平衡或是样本数据差异较大等情况下,同样表现稳定。文中选择 CHI 作为特征提取的方法。

(3)特征加权。

特征集合中不同的特征词对样本文档的重要程度和区分度不同,所以需要对特征集合中的所有特征词

进行赋权重处理。常用的加权算法有:布尔权重(BooLean,BL)、词频权重(Term Frequency,TF)、倒文档权重(Inverse Document Frequency,IDF)和 TFIDF 权重(Term Frequency, Inverse Document Frequency)等等[12]。

TFIDF 权重规避了 TF 权重与 IDF 权重的缺点,将两种权重算法结合起来,寻求一种折中[13]。即将以下两种思想综合考虑:特征词 t 在样本集中的文档频率 $\mathrm{df}(t)$ 越高,词语 t 越不重要;特征词 t 在文档 d_i 中出现频率越高,词语 t 越重要。因此文中选择了 TFIDF 方法进行特征权重的计算。

(4)中心法。

中心法文本分类的思想是:对于每个已知的类别 i 都存在一个类别的中心向量 $\boldsymbol{C}_i$,这个中心向量为该类别的代表向量。当需要对一个未知类别文本分类时,把文本向量 $\boldsymbol{d}$ 和每个类别的中心向量 $\boldsymbol{C}_i$ 进行对比,通过最大相似度确定该文本类别。但是,中心法分类需要满足一个条件,即待分类向量与它所属于的向量相似度要大于其他的类别,但是数据分布是有偏差的,这些偏差会导致判断失误,如图 2 所示。

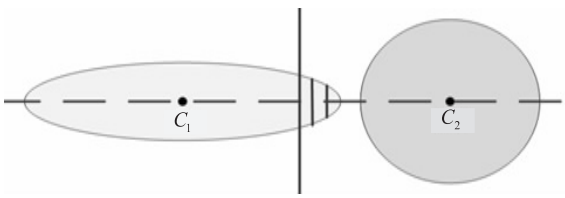

图 2 中心法偏差图

如图 2 所示,C_1 和 C_2 分别为两个类别,但是如果属于 C_1 类别的待分类文本位于两个类别中心右侧的位置,那么模型会将其判断为 C_2 类别,因为这样待分类文本距离 C_2 距离较近,若是如此将会出现偏差。

为了解决这一问题,进行两点改进:一是修改中心向量的计算方法,使中心向量更具有显著性和通用性;二是在中心法分类中尽量选择更加合适的相似度计算方法。

通过以上分析,文中将尝试使用不同的中心向量计算方法:均值法、总和法和归一化法。分别对这些方法得到的结果进行评估,以确保中心法有更好的表现。

3 实验设计

3.1 样本的选取及处理

(1)样本的选取。

由于目前国内尚无关于农业网页的开放语料库,文中的农业信息网页样本主要来源于国内农业相关网站。下面将简单介绍样本的选取和处理。

首先,从各大农业网站上抓取林业、种植业及制品、渔业、园林、畜牧业、农业生产资料、农业机械、植物病理农业相关的网页,选取 18 000 张网页。然后,组织 30 名学生每组 3 人共 10 组,每组标记 1 800 张网页,每组三人对相同的 1 800 张网页分别进行标记,标定结果写入到结果表中。对每 1 张网页 3 人标记相同时,才判定该网页类别,不同则重新标记,标记结果差异数据统计见表 1。

表 1 数据标记差异统计

标记次数	标定相同记录数	标定不同记录数
1	13 898	4 102
2	2 652	1 450
3	867	583
4	156	472

去除无法确定类别的剩余 472 条记录。对分类结果进行数据统计,见表 2。

表 2 分类结果统计

分类	数量
种植业及制品	2 895
渔业	1 976
园林	2 597
畜牧业	1 589
农业生产资料	2 487
农业机械	1 892
植物病理	1 934
林业	2 158

取其中每类 70% 作为训练样本集,其他每类 30% 作为测试样本集。最终得到训练样本集共 12 270 个样本,测试样本集共 5 258 个样本。

(2)样本的处理。

首先,对样本进行分词,去除停用词,提取特征词。提取特征词使用的方法是 CHI,公式如下:

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

其中,N 为训练集中的文档总数;A 为属于类别 C_i 且包含特征词 t 的文档频数;B 为不属于类别 C_i 但包含特征词 t 的文档频数;C 为属于类别 C_i 但不包含特征词 t 的文档频数;D 为不属于类别 C_i 也不包含特征词 t 的文档频数。

其次,计算每个样本的特征向量,最后构成样本的向量集合 $\vec{D}(\vec{d}_1,\vec{d}_2,\cdots,\vec{d}_n)$。文本的特征向量由特征空间的特征权重组成,公式如下:

$$\vec{d}_i=(\vec{w}_1,\vec{w}_2,\cdots,\vec{w}_n) \tag{2}$$

其中,w 为使用 TFIDF 计算各项特征的权重。

$$w_{ij}=\mathrm{tf}_{ij}\times\mathrm{idf}(t_j)=\mathrm{tf}_{ij}\times\log\left(\frac{N}{\mathrm{df}(t_j)}+0.01\right) \tag{3}$$

其中, tf_{ij} 为特征词 t_j 在文档 d_i 中的词频权重; $df(t_j)$ 为特征词 t_j 在样本集中的文档频率。

然后,计算出每类样本的中心向量,三种中心向量计算方法如下所示。

均值法计算公式为:

$$\vec{c_i} = \frac{1}{|c_i|}\sum_{d\in c_i}\vec{d_i} \tag{4}$$

总和法计算公式为:

$$\vec{c_i} = \sum_{d\in c_i}\vec{d_i} \tag{5}$$

归一化法计算公式为:

$$x_i^* = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{6}$$

其中, x_i 、x_i^* 分别表示数据归一化前后的值; $x_{\max}$ 、$x_{\min}$ 分别表示样本数据中的最大和最小值。

最后,使用欧氏距离计算待分类样本与各类中心向量的相似度,待分类样本与某类的相似度最高即属于该类。

欧氏距离公式为:

$$d_{ij} = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})^2} \tag{7}$$

3.2　实验方案

中心法分类中需要两步:第一步是训练阶段,即使用已给定的样本标记集合生成具有分类能力的中心向量分类器;第二步是分类阶段,即使用训练阶段生成的分类器对未知类别样本进行分类,确定其所属的类别。

(1)训练阶段。

①将从农业网站上获取的农业网页进行预处理,去除文档标签。

②使用庖丁解牛分词器对中文文本进行中文分词和去除停用词等操作。

③使用 X^2 统计提取出对分类贡献较高的特征词,将每篇文档用一个特征词集的向量表示。

④使用 TFIDF 对特征集合中的特征词进行赋权重处理,得到由特征权重组成的文档的特征向量空间 $\vec{D}(\vec{d_1},\vec{d_2},\cdots,\vec{d_n})$ 。

⑤分别使用均值法、总和法和归一化法计算出每类的中心向量 $\vec{C}(\vec{c_1},\vec{c_2},\cdots,\vec{c_n})$ 。

(2)分类阶段。

①待分类文档需要经过训练阶段的步骤 1 ~4,得到测试文档向量空间 $\vec{t}(\vec{t_1},\vec{t_2},\cdots,\vec{t_n})$ 。

②将待分类 $\vec{t_i}$ 与每个类的中心向量 $\vec{c_i}$ 进行比较,通过欧氏距离得出 $\vec{t_i}$ 与每个类的相似度。

③比较 $\vec{t_i}$ 与各类的中心向量的相似度,哪个类的相似度较高则该文档属于哪类文档。

4　实验结果

4.1　分类器性能评估

评估分类准确程度的依据是通过对网页分类结果与人工分类结果进行比较,结果越相近,分类的准确程度就越高。国际上通用的评价指标有:查准率(Precision, P)、查全率(Recall, R)和 F_1 测度[14]。

假设 A 、B 、C 、D 含义如下: A 表示样本集中原本是正例,被模型判断为正例的样本数; B 表示样本集中原本是正例,却被模型判断为反例的样本数; C 表示样本集中原本是反例,被模型判断为反例的样本数; D 表示样本集中原本是反例,却被模型判断为正例的样本数。

查准率评价指标公式为:

$$\text{Precision} = \frac{A}{A + D} \tag{8}$$

查全率评价指标公式为:

$$\text{Recall} = \frac{A}{A + B} \tag{9}$$

F_1 测度是对查准率和查全率两个指标进行加权和平均后形成的一个综合指标。公式为:

$$F_1 = \frac{\text{Precision} \times \text{Recall} \times 2}{\text{Precision} + \text{Recall}} \tag{10}$$

4.2　结果与分析

实验首先通过人工标记的方式选取出了 17 528 篇网页作为实验数据来源。这些网页来源于各大农业网站,其中包含:种植业及制品(2 895 篇)、渔业(1 976 篇)、园林(2 597 篇)、畜牧业(1 589 篇)、农业生产资料(2 487)、农业机械(1 892 篇)、植物病理(1 934 篇)、林业(2 158 篇)。将这些网页的 70% 作为训练样本,30% 作为测试样本,如表 3 所示。

表 3　训练集和测试集划分表

分类	训练集	测试集
种植业及制品	2 027	868
渔业	1 383	593
园林	1 818	779
畜牧业	1 112	477
农业生产资料	1 741	746
农业机械	1 324	568
植物病理	1 354	580
林业	1 511	647

分别对基于总和法、均值法、归一化法三种不同的中心向量计算方法的中心法进行测试,测试指标包括:查准率(P)、查全率(R)、F_1 测度。测试结果如表 4 ~6 所示。

每一类文章经过预处理、特征提取、特征加权后会变成一个矩阵,每一行代表一篇文章,每个值代表特征词的重要性。总和法是指将同一类的所有文章特征向

表 4 总和法计算中心向量的中心法分类结果

	种植业及制品	渔业	园林	畜牧业	农业生产资料	农业机械	植物病理	林业
种植业及制品	739	6	11	17	29	21	12	33
渔业	11	523	0	7	23	27	2	0
园林	19	0	648	1	18	21	33	39
畜牧业	15	3	0	346	57	51	2	3
农业生产资料	57	27	19	33	519	31	33	27
农业机械	37	27	17	23	21	424	4	15
植物病理	31	4	22	5	25	17	444	32
林业	17	0	35	0	27	15	37	516
P	0.798 1	0.886 4	0.861 7	0.800 9	0.721 8	0.698 5	0.783 1	0.775 9
R	0.851 4	0.882 0	0.831 8	0.725 4	0.695 7	0.746 5	0.765 5	0.797 5
F_1	0.823 9	0.884 2	0.846 5	0.761 3	0.708 5	0.721 7	0.774 2	0.786 6

表 5 均值法计算中心向量的中心法分类结果

	种植业制品	渔业	园林	畜牧业	农业生产资料	农业机械	植物病理	林业
种植业及制品	761	6	10	17	27	18	2	27
渔业	8	530	0	7	21	25	2	0
园林	15	0	655	1	17	19	33	39
畜牧业	13	3	0	360	51	45	2	3
农业生产资料	57	27	19	33	519	31	33	27
农业机械	37	27	17	23	21	424	4	15
植物病理	26	4	18	5	21	11	466	29
林业	17	0	33	0	25	13	35	524
P	0.814 8	0.887 8	0.871 0	0.807 2	0.739 3	0.723 5	0.807 6	0.789 2
R	0.876 7	0.893 8	0.840 8	0.754 7	0.695 7	0.746 5	0.803 4	0.809 9
F_1	0.844 6	0.890 8	0.855 6	0.780 1	0.716 9	0.734 8	0.805 5	0.799 4

表 6 归一化法计算中心向量的中心法分类结果

	种植业及制品	渔业	园林	畜牧业	农业生产资料	农业机械	植物病理	林业
种植业及制品	794	6	10	17	9	5	2	25
渔业	6	546	1	6	15	17	2	0
园林	11	0	690	2	6	14	24	32
畜牧业	9	3	0	386	31	43	2	3
农业生产资料	56	23	14	26	543	25	32	27
农业机械	37	25	13	23	17	437	4	12
植物病理	22	2	14	3	15	9	491	24
林业	12	0	23	0	21	7	31	553
P	0.838 4	0.902 5	0.902 0	0.833 7	0.826 5	0.784 6	0.835 0	0.818 0
R	0.914 7	0.951 1	0.885 8	0.809 2	0.727 9	0.769 4	0.846 6	0.854 7
F_1	0.874 9	0.926 2	0.893 8	0.821 3	0.774 1	0.776 9	0.840 8	0.836 0

量进行求和运算,最后得到的和向量即为该类的中心向量;均值法是在总和法的基础上取得和向量的平均值,由实验结果得出,均值法在查全率、查准率和 F_1 测度上比总和法高出了 5% 左右;归一化法是为了消除指标之间的量纲影响,由实验结果得出,归一化法在查全率、查准率和 F_1 测度上比总和法高出了 10% 左右,比均值法高出了 5% 左右。

通过测试结果可以发现,农业生产资料和农业机

械分类结果在查全率、查准率和 F_1 测度上表现的都比较差,原因可能是因为这两类与其他类具有一定的关联性,所以导致分类结果与其他相比较低。

实验结果表明,改进后的使用归一化方法计算中心向量的中心法的测试结果优于总和法和均值法,几乎各项指标都达到了 80%。由此可以推断出改进中心向量的计算方法对中心法的改良起到了一定的作用,得到的分类结果更好,准确率更高。

5　结束语

文中采用中心法对农业信息网页的分类进行了研究,实验结果表明归一化法计算中心向量对于农业信息网页的分类判别有着较高的准确率。下一步将继续调整中心向量的算法以及相似度算法,从而进一步提高模型在农业信息网页分类中的准确率。与此同时探索其他文本分类方法,并设计对比实验,以获得更佳的农业信息网页分类模型。

参考文献:

[1] 金一宁,王华兵,王德峰. 基于 KNN 及相关链接的中文网页分类研究[J]. 哈尔滨商业大学学报:自然科学版,2011,27(2):203-207.

[2] 江小平,李成华,向　文,等. 云计算环境下朴素贝叶斯文本分类算法的实现[J]. 计算机应用,2011,31(9):2551-2554.

[3] 李　琼,陈　利. 一种改进的支持向量机文本分类方法[J]. 计算机技术与发展,2015,25(5):78-82.

[4] 胡金有,张　健,游龙勇. 我国农业信息网站现状分析[J]. 农机化研究,2005(6):38-40.

[5] 王　健,甘国辉. 多维农业信息分类体系[J]. 农业工程学报,2004,20(4):152-156.

[6] Feng Xia, Tang Xianchao. An improved dictionary-based chinese word segmentation approach in Lucene[C]//Proceedings of 2010 international conference on services science, management and engineering. [s. l.]:[s. n.],2010.

[7] 孙殿哲,魏海平,陈　岩. Nutch 中庖丁解牛中文分词的实现与评测[J]. 计算机与现代化,2010(6):187-190.

[8] Zhang Qun, Cheng Yu. Research on Chinese word segmentation algorithm based on special identifiers[C]//Proceedings of the 2011 international conference on computing, information and control. [s. l.]:Intelligent Information Technology Application Association,2011.

[9] 王霜霜,张太红,冯向萍,等. 农业网站导航页面识别模型研究[J]. 新疆农业大学学报,2011,34(5):447-453.

[10] 苏金树,张博锋,徐　昕. 基于机器学习的文本分类技术研究进展[J]. 软件学报,2006,17(9):1848-1859.

[11] Wang Bingkun, Huang Yongfeng, Yang Wanxia, et al. Short text classification based on strong feature thesaurus[J]. Journal of Zhejiang University-Science C (Computers & Electronics),2012(9):649-659.

[12] Zhang Yuntao, Gong Ling, Wang Yongcheng. An improved TF-IDF approach for text classification[J]. Journal of Zhejiang University Science A (Science in Engineering), 2005, 6A(1):49-55.

[13] 张保富,施化吉,马素琴. 基于 TFIDF 文本特征加权方法的改进研究[J]. 计算机应用与软件,2011,28(2):17-20.

[14] 宋枫溪,高　林. 文本分类器性能评估指标[J]. 计算机工程,2004,30(13):107-109.

(上接第 145 页)

参考文献:

[1] 张炯森. 科技项目档案管理中存在的问题及对策[J]. 科技情报开发与经济,2009,19(1):98-99.

[2] 陈文英,陈开魁,徐迟默,等. 热带农业科技档案信息资源共享现状分析及可行性研究[J]. 安徽农业科学,2014,42(25):8846-8848.

[3] 肖琬蓉,张　静. 科技档案全文数字化信息系统开发研究[J]. 计算机应用与软件,2013,30(3):145-147.

[4] 许雯倩,李　娟. 论信息化背景下高校档案推送服务[J]. 兰台世界,2013(8):18-19.

[5] 王兰成,黄永勤. 大数据时代国防科技档案服务的信息技术研究[J]. 浙江档案,2014(11):6-9.

[6] 廖淑莉. 构建科技档案云平台　支撑科技创新驱动——以粤西高校科技档案云平台关键技术研究为例[J]. 档案时空,2016(2):16-18.

[7] 王兰成. 科技档案异构数据整合及其检索的研究[J]. 中国科技资源导刊,2009,41(5):36-41.

[8] 毕春华. "文档一体化"管理的建设及启示[J]. 江汉大学学报:社会科学版,2003,20(1):93-96.

[9] 张艳霞,齐兰英. 数字化档案管理与查询[J]. 黑龙江科技信息,2010(7):161-161.

[10] 张　红,王　敏. 利用 ACCESS 建立基建档案全文数据库管理系统[J]. 机电兵船档案,2007(3):24-26.

[11] 刘　轩. 档案数字化的策略分析及系统构建[J]. 北京档案,2004(12):28-29.

[12] 杨春子,孙小康. 科技档案管理系统的研究与实现[J]. 武汉理工大学学报:信息与管理工程版,2006,28(3):62-65.

[13] 周　丽. 电子档案加密探讨[J]. 国网技术学院学报,2015,18(1):85-88.