用于文本分类的特征项权重算法改进

龚 静,胡平霞,胡 灿
(湖南环境生物职业技术学院 信息技术系,湖南 衡阳 421005)

摘 要:TF-IDF 算法是文本分类中一种常用的权重计算方法,但是 TF-IDF 仅仅考虑了特征项在文本中出现的次数以及该特征项在训练集中的出现频率,没有考虑特征项在各个类间的分布情况及特征项的语义信息。因此针对 TF-IDF 的不足提出了一种改进的 TF-IDF 算法,此算法既考虑了特征项在类内的分布情况又考虑了特征项的位置及长度等语义因素,能更好地反映特征项的重要性。用朴素贝叶斯分类器验证其有效性,实验结果表明该算法优于 TF-IDF 算法,能较好地提高文本分类的准确率。

关键词:文本分类;特征项;权重;改进
中图分类号:TP301 **文献标识码**:A **文章编号**:1673-629X(2014)09-0128-05
doi:10.3969/j.issn.1673-629X.2014.09.029

Improvement of Algorithm for Weight of Characteristic Item in Text Classification

GONG Jing, HU Ping-xia, HU Can
(Department of Information Technology, Hunan Environment and Biological Polytechnic, Hengyang 421005, China)

Abstract: TF-IDF algorithm is a commonly used method of calculating weight in text classification, but TF-IDF considers only occurrence of feature in the text, as well as the frequency of characteristic appearing in the training set, and does not take into the distribution of characteristics in each class and the semantic information of characteristics account. In order to solve this problem, the improved TF-IDF algorithm has been proposed which considers not only the distribution condition of feature in class, but also the semantic factors such as the position of the feature, length of the feature. This algorithm can better reflect the importance of feature item, and its validity is verified by Naïve Bayes classifier. The experiment results show that the proposed algorithm outperforms the TF-IDF algorithm, and the algorithm can improve the accuracy of text classification well.

Key words: text classification; feature item; weights; improvement

1 概 述

随着信息技术和互联网技术的飞速发展,人们可以访问的文本信息呈几何级数增长,怎样快速地从海量的文本信息中得到人们所需要的知识成为研究人员关注的热点。近几年来,文本分类取得了较大的发展,它能够有效地组织与管理文本,提高查询质量,方便用户快速地得到所需的文本[1],因此被广泛地应用在信息检索、信息过滤、邮件分类、数字图书馆等领域。

文本分类是指按照预先定义的主题类别,通过分析文本的内容将文本集合中的每个文本分配到预先定义的类别中[2]。文本分类的过程如图 1 所示。

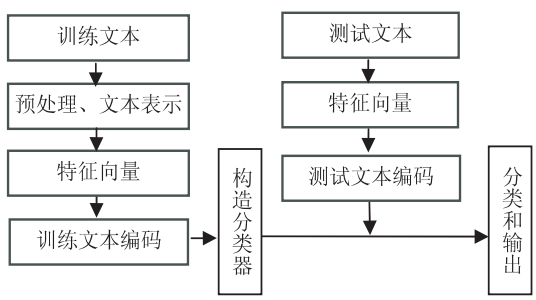

图 1 文本分类结构图

由于中文文本具有有限的结构,甚至没有结构,计算机不能直接理解其语义,因此,在分类之前应先对文本进行预处理和文本表示[3]。预处理是指抽取文本中

收稿日期:2013-11-12 **修回日期**:2014-02-18 **网络出版时间**:2014-07-17
基金项目:湖南省教育科技计划项目(07D036);湖南省教育厅、财政厅联合资助项目(12C1056)
作者简介:龚 静(1972-),女,湖南岳阳人,副教授,硕士,研究方向为自然语言处理。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20140717.1228.015.html

代表文本内容的元数据,即特征项。特征项可以为字、词、短语或者语义单元,但是多半是以词作为特征项,因为词是最小的具有语义特性的独立单位。然后用结构化的形式表示特征项,即文本表示。最常见的文本表示有向量空间模型(Vector Space Model,VSM)、概率模型(Probabilistic Model,PM)和布尔模型(BooLean Model,BLM)[4]。布尔模型是基于集合论和布尔代数的一种简单检索模型,其值只有两个,如果特征项属于文本,那么特征项的值是 1;否则,特征项的值为 0。概率模型是按照特征项在相关文档中出现的概率和无关文档中出现的概率来计算该特征项的权重。VSM 是使用最多的文本表示模型,它最基本的思想是用词袋法(Bag -of-Words)表示文本。它将每一个不同的词条全部看作是特征空间中单独的一维,这样每一个文本就是特征空间中的一个向量[5],即 $v(d)=(t_1,w_1;\cdots,t_i,w_i;\cdots,t_m,w_m)$,简记为 $d=(w_1,w_2,\cdots,w_m)$。其中 t_i 为特征项,w_i 为 t_i 在文本 d 中的权重。文本集合用矩阵 $\boldsymbol{A}(n,m)$ 表示,n 代表文本数,m 代表特征数。这样,行是文本,列是文本中出现的特征项,从而将文本相似度计算转化为向量的相似度计算。

特征项的权值计算是向量相似度计算中的关键因素,直接影响计算结果的准确性。研究人员一般采用 TF-IDF 计算方法,其中 TF(Term Frequency)反映特征项在文本内部的分布情况;IDF(Inverse Document Frequency)反映特征项在整个文本集合的分布情况,能够在一定程度上体现特征项的区分能力[6]。但是仅仅依据特征项的频率统计,忽略特征项的语义因素及类内分布信息,并不能准确反映特征项在文本中的重要程度。

因此,文中在 TF-IDF 权值计算的基础上,根据特征项的长度、位置信息及特征项在类内的分布规律对 TF-IDF 权重进行加权,从而对 TF-IDF 计算公式进行修正,以便更加准确地反映特征项在文本中的重要程度,进而提高文本相似度计算的准确性。

2　特征项权重的概念及 TF-IDF 的计算

2.1　特征项权重的概念

对于每一个特征项,它对类别的区分能力是不相同的,而且这种能力影响文本分类的结果,这种能力的大小称之为特征项权重(Term Weight,TW)[7]。影响 TW 大小的因素包括以下几个方面:

(1)特征项的重要性。虽然特征选择评价函数不能完全地评估特征项的重要性,但是相对来说,还是能够体现特征项的重要性。如果存在一个特征项,它的特征选择评价函数值越大,说明该特征项在一定程度上越重要,因此,它对文本分类的影响越大。

(2)特征项的特征性。如果存在一个特征项,该特征项非常多地出现在某个类别的文本集中,而很少出现在其他类别的文本集中,那么该特征项具有较强的类别区分能力,对分类的贡献也就越大。

(3)特征项的代表性[8]。如果存在一个特征项,它均匀地分布在某个类别的文本集中,也就是说该类别中出现该特征项的文本数越多,那么该特征项代表其所在类别的能力越强,对文本分类贡献的也就越大。

2.2　传统的 TF-IDF 公式

特征项频率分为绝对词频与相对词频[9]。绝对词频是指特征项在文本中出现的次数;相对词频是指归一化的词频,它的计算方法主要是采用 TF-IDF 公式[10]。TF-IDF 计算方法强调下面三个因素:

(1)特征项频率(TF):是指特征项在指定文本中出现的次数,是一个与文本相关的统计量。在不同类别的文本中,特征项的出现频率存在很大的差异,所以,TF 是文本分类的一个重要的参考因素,在文本分类的最初阶段就是采用 TF 进行计算的。

(2)特征项的倒排文本频率(IDF):是对特征项在文本集合中的分布情况进行量化,一般采用的计算方法是 $\log(\frac{N}{n_i})+0.01$。其中,N 是指全部训练集的文本数;n_i 是指训练文本集中出现该特征项的文本数。$\log(\frac{N}{n_i})$ 越小,说明该特征项越普遍,如果该特征项出现在训练文本集合的所有文本中,则 $\log(\frac{N}{n_i})$ 值为零。这完全符合人们的经验:出现在所有文本中的特征项对于区分文本的贡献也就越小。

(3)归一化因子(Normalization Factor):TF 与 IDF 的计算没有考虑文本的长度,以免长度对文本分类产生影响,所以对分量进行标准化,将权值规范在[0,1]之间。

依据上述三个因素,得到了 TF-IDF 的权重计算公式,如公式(1)所示:

$$w_{ik}=\frac{\mathrm{tf}_{ik}*\log(\frac{N}{n_i}+0.01)}{\sqrt{\sum_1^n(\mathrm{tf}_{ik})^2*\log^2(\frac{N}{n_i}+0.01)}} \tag{1}$$

式中,w_{ik} 指特征项 t_i 在文本 d_k 的 TF-IDF 权值;tf_{ik} 是指特征项 t_i 在文本 d_k 中出现的次数;N 是训练文本集的文本数;n_i 是指在训练文本集中出现特征项 t_i 的文本数。

总之,公式(1)的基本思想是:能够区别文本的词语应该是那些出现在少量训练文本中且出现频率足够高,而在训练文本集的其他文本中的出现频率足够低的词语,这符合香农信息学理论。

3 特征项权重算法的改进

3.1 基本思想

通过对 TF-IDF 权重计算方法的分析,可以知道 TF-IDF 仅仅考虑了特征项在文本中出现的次数以及该特征项在训练集中的出现频率,没有考虑该特征项在各个类间的分布情况及特征项的语义信息。因此,文中从特征项在类内的分布情况及它的位置、长度等信息来修正 TF-IDF 算法,使特征项的权重计算方法更合理,以达到提高文本分类的效果。

3.2 类别相关系数

实际上,一个特征项的重要程度不仅与它出现的频率有关,还应该与它对所属类别的贡献程度相关,也就是说特征项的权值应该能反映出其与各类别的相关性。

首先看一个例子,t_i表示文本 d 包含的特征项,n_i 表示训练集中包含特征项 t_i的文本数。TF 表示文本 d 中特征项 t_i的频率,IDF 表示训练集中特征项 t_i的反文本频率,假设 $N=400$,训练集中包含特征项 t_i的文本数 n_i都为 60,t_1,t_2在文本 d 中出现的次数都是 40,如表 1 所示。

表 1 特征项 t_k与类别 C_i的关系

t_i	n_i	TF	IDF	β
t_1	60	40	log(N/60)	0.8
t_2	60	40	log(N/60)	0.05

通过表 1 中的数据,可以知道,特征项 t_1,t_2的 TF-IDF 值是相同的,根据 TF-IDF 加权方法,特征项 t_1,t_2具有相同的类别区分能力。但是假设 t_1只出现在一个类别中,t_2则分散出现在多个类别中,很明显 t_1对文本的代表性肯定比 t_2强,那么 t_1的权重应该比 t_2高才合理,可是两者的权值是相同的,体现不出特征项对类别的贡献程度,也看不出两个特征项的区别。实际上,如果一个特征项主要出现在某一个类别中,而较少出现在其他类中,应该赋予较大的权重。因此,在 TF-IDF 计算方法的基础上,利用特征项在类内的分布信息,提高频繁出现在某一类别中的特征项的权值,降低出现次数少的权值,因此,提出了类别相关系数的概念,以改进 TF-IDF 方法的不足。在表格的后面增加一列作为类别相关系数,定义为 β,β 为类别 C_j中包含特征项 t_i的文本数与训练集中包含特征项 t_i的文本数的比值,比值越大,说明特征项 t_i在类 C_j中频繁出现,分布比较均匀,应该赋予较大的权重;反之就属于稀有词,它不能代表该类的内容,应该具有较低的权重。例如 t_1的 β 值为 0.8,t_2的 β 值为 0.05,t_1比 t_2更能体现文本的类别属性。加入类别相关系数后,权重计算公式如公式(2)所示。

$$w_{ik} = w_{ik} * \beta \tag{2}$$

3.3 位置权重系数

文本分类处理的对象主要是因特网上各种各样的信息,包括各种文章和书籍,对于文章和书籍来说,位于不同位置的词对体现文章主题的作用是不一样的,因此对文本分类的贡献程度也是不一样的。文章或书籍的标题和副标题首先就言简意赅地表现文章或书籍的中心思想,文章的首段往往就点明题目,阐述文章的主题思想,总结性陈述出现在文章的末段[11]。这些规律说明特征项处于不同的位置,它的作用也是不一样的,虽然有的特征项的 TF 不高,但它能很好地体现出文本的内容。因此,在 3.2 节对 TF-IDF 改进的基础上针对不同位置的特征项进行加权处理。根据统计资料及人们的经验,设置特征项的权重系数 p 如表 2 所示。

表 2 特征项权重系数表

特征项在文本中的位置	权重系数(p)
标题	1.0
首段	0.8
末段	0.5
其他位置	0.3

设 l_{t_i} 为特征项 t_i出现在相应位置的次数,p_{t_i} 为特征项 t_i的位置权重系数,引入了位置加权系数的权重计算方法如公式(3)所示:

$$w_{ik} = w_{ik} * \frac{\sum l_{t_i} * p_{t_i}}{l_{t_i}} \tag{3}$$

3.4 特征长度函数

词语的长度是影响特征项权值大小的一个非常重要的因素。在中文文本中,词语的长度越长,在文本中出现这个词语的概率就越小,但是长的词语的信息内涵肯定比短的词语更丰富,重要性更高。相对来说,长词具有较低的频率,是面向内容的,而短词具有较高的频率和较多的含义。适当增加长词的权值,有利于对词汇进行分割,这样可以更好地反映特征项对于文本的重要性。例如,“污染”“尾气污染”“机动车尾气污染”三个词语专指性逐渐增强,但概括性逐渐减弱。所以,给长词赋予较高的权重,引入词长系数 λ 后,公式(3)修正为公式(4)。

$$w_{ik} = w_{ik} * \frac{\lambda}{\lambda + 1} \tag{4}$$

其中,λ 表示词语的长度,如“机动车尾气污染”中 $\lambda = 7$。

3.5 改进的算法描述

(1)对训练文本集进行预处理和文本表示,具体为分词、建立停用词表将在文本中出现较多的而又没

有实在意义的虚词和功能词删除;

(2)计算特征项 t_i 在文本 d_k 中出现的频率 tf_{ik}、t_i 在 C_j 中出现的文本数 n_j、t_i 在训练文本集中的出现次数 n_j,同时,统计特征项 t_i 的词长 λ 及位置信息;

(3)利用步骤(2)计算得到的 tf_{ik} 与 n_j 计算出 tf 与 idf,从而得到 t_i 的 TF-IDF 计算公式,即 $w_{ik}=tf*idf$,依次计算出向量空间中的每个特征项的权重;

(4)利用步骤(2)统计到的 n_j、n_i,求出 β 值,利用公式 $w_{ik}=w_{ik}*\beta$ 计算出引入类别相关系数后的特征项权值;

(5)利用步骤(2)得到的位置信息,然后根据公式 $w_{ik}=w_{ik}*\frac{\sum l_{t_i}*p_{t_i}}{l_{t_i}}$,计算出引入位置信息后的特征项权值;

(6)利用步骤(2)中得到的词语长度值 λ,然后根据公式 $w_{ik}=w_{ik}*\frac{\lambda}{\lambda+1}$ 修正特征项权值 w_{ik};

(7) 从而计算出每个文本的特征项的最终权值,每个文本表示为向量 $(t_1,w_1;\cdots,t_i,w_i;\cdots,t_m,w_m)$,$t_i$ 为特征项,w_i 为对应的权值,这样每个文本用向量空间中一个点表示。

4　实验与结果

4.1　分类算法

中文文本分类算法有很多,一般的分类算法都是趋向于二分问题,也就是说一个文本要么与预先确定的主题相关,要么不相关[12]。使用最为广泛的算法有:K-最近距离方法、朴素贝叶斯分类算法、Rocchio 算法、支持向量机以及基于语义网络的概念推理网分类算法等,实验中采用朴素贝叶斯分类算法。

朴素贝叶斯分类算法[13]是采用贝叶斯公式通过对词的分布和类别的先验概率来计算未知文本属于某一类别的概率,计算见公式(5)。

$$P(C_j\mid T)=\frac{P(C_j)P(T\mid C_j)}{P(T)} \tag{5}$$

其中,$P(T\mid C_j)$ 表示类别 C_j 中含有文本 T 的概率;$P(C_j\mid T)$ 表示文本 T 属于类别 C_j 的概率;在所有的 $P(C_j\mid T)$ 中,如果 $P(C_k\mid T)$ 的值最大,则文本 T 属于 C_k 类。因为 $P(T)$ 是常数,所以只需要求解 $P(C_j)P(T\mid C_j)$。

公式假设特征项之间是条件独立的,但是这个假设在实际情况中是不成立的,为了在朴素贝叶斯分类算法中能够取得良好的结果,设计 $P(C_j\mid T)$ 计算公式如下:

$$P(C_j\mid T)=P(C_j)\prod_{i=1}^{M}P(t_i\mid C_j) \tag{6}$$

其中,$P(C_j)=\frac{C_j\text{中文本个数}}{\text{总文本个数}}$;$P(t_i\mid C_j)=\frac{t_i\text{在类别}C_j\text{中出现的次数}}{C_j\text{中所有词的个数}}$。

朴素贝叶斯分类算法是一种统计方法,与其他分类算法相比,主要有下面几个优点:速度快而稳定,利于创建线性模型,分类效果好,便于维护。

4.2　实验数据

实验数据从搜狗中文文本分类语料库完整版中选取 8 个类别,分别是财经、体育、IT、教育、汽车、旅游、健康、军事,每个类别 700 篇,训练样本与测试样本各一半。首先使用中国科学院的中文分词系统 ICTCLAS 进行分词,去停用词和功能词,然后使用 CHI 特征选择方法删除无关和冗余特征。采用朴素贝叶斯分类算法,在进行特征项权值计算时使用传统的 TF-IDF 公式和改进的权值算法来进行对比实验,实验平台是 Win7,采用 C++编程语言。

4.3　评价标准

文中分类的评价标准采用宏平均查准率(Macro-averaging Precision,MP)、宏平均查全率(Macro-averaging Recall,MR)和宏平均 F_1(Macro-averaging F_1,MF_1)测度值,各评价参数定义如下[14]:

(1)平均准确率 P。

分类的准确率 P=分类正确文本数/实际分类的文本数

$$MP=\frac{1}{n}\sum_{j=1}^{n}P_j$$

其中,n 为实际参与分类的文本数;P_j 为第 j 类的准确率。

(2)平均召回率 R。

分类的召回率 R = 分类正确文本数 / 分类应有的文本数

$$MR=\frac{1}{n}\sum_{j=1}^{n}R_j$$

其中,n 为实际参与分类的文本数;R_j 为第 j 类的召回率。

(3)宏平均 F_1 值。

$$MF_1=\frac{MP\times MR\times 2}{MP+MR}$$

4.4　实验结果

表 3 给出了实验中采用 TF-IDF 方法和改进权值算法计算特征项权值时,每个类别的性能指标与 8 个类别的宏平均指标。

表 3 的实验结果说明了在计算特征项权值时采用不同的计算方法,对分类器带来的性能变化。由表 3 中相应指标可知,除了个别类,如旅游的性能有一些降低外,其余部分的分类性能指标都提高了。

表 3 TF-IDF 方法与改进算法的性能评估指标 %

类别	R (TF-IDF)	P (TF-IDF)	F_1 (TF-IDF)	R(改进算法)	P(改进算法)	F_1(改进算法)
财经	82.12	83.27	82.69	85.43	87.27	86.34
体育	79.23	76.43	77.80	83.11	79.56	81.30
IT	69.21	72.34	70.74	73.65	76.09	74.85
教育	85.01	80.21	82.54	90.02	89.34	89.68
汽车	79.03	74.56	76.73	83.21	80.86	82.02
旅游	76.19	73.98	75.07	72.45	68.58	70.46
健康	80.89	84.32	82.57	83.95	89.45	86.61
军事	80.15	85.76	82.86	85.34	89.4	87.32
MP		78.86			82.57	
MR		78.98			82.15	
MF_1		78.92			82.36	

为避免实验中的偶然性,在同样的实验条件下重复进行了 10 次实验,取实验结果的算术平均值作为最终的分类性能指标,如表 4 所示。其中,TF-IDF(R)、TF-IDF(P)和 TF-IDF(F_1)分别指利用 TF-IDF 方法计算特征项权值时整个分类器的平均查全率、平均查准率及宏平均 F_1 值。改进算法(R)、改进算法(P)和改进算法(F_1)分别指利用改进的权重算法计算特征项权值时整个分类器的平均查全率、平均查准率及宏平均 F_1 值。从表中的数据可以知道,TF-IDF 方法与改进的权重算法相比,平均查全率从 78.48% 提高到 82.35%,提高了 3.87%,也就是说改进的算法比 TF-IDF 方法多训练出正确的文本 100 多篇;平均查准率从 78.55% 提高到 82.09%,提高了 3.54%;宏平均 F_1 值从 78.51% 提高到 82.22%,提高了 3.71%。实验结果表明,改进的权值计算方法更能体现特征项在文本中的重要程度,能有效地提高中文文本分类的性能。

表 4 TF-IDF 方法与权重改进算法 10 次实验结果 %

实验编号	TF-IDF (R)	TF-IDF (P)	TF-IDF (F_1)	改进算法(R)	改进算法(P)	改进算法(F_1)
1	78.23	78.67	78.45	83.01	82.09	82.55
2	78.86	78.33	78.59	82.57	82.15	82.36
3	77.97	78.76	78.36	82.87	83.06	82.96
4	78.09	78.45	78.27	81.79	82.13	81.96
5	78.56	78.91	78.73	82.45	82.77	82.61
6	78.88	77.87	78.37	82.84	82.67	82.75
7	79.09	78.55	78.82	81.88	81.97	81.92
8	78.14	78.47	78.30	82.11	80.67	81.38
9	78.34	78.43	78.38	81.69	82.35	82.02
10	78.65	79.02	78.83	82.31	80.99	81.64
平均值	78.48	78.55	78.51	82.35	82.09	82.22

5 结束语

特征项权重算法是文本分类的一个基础环节,采用不同的权重算法对文本分类的效果有很大的不同。TF-IDF 公式仅仅考虑了特征项的出现频率,单纯地认为特征项在文本中出现得越多,在训练集中包含该特征项的文本数越少,越能代表文本,从而权值越大。文中在 TF-IDF 权值计算方法的基础上,用特征项的类内分布情况、位置信息及特征项长度等因素来修正其计算方法,将之应用在文本分类中,取得了比较好的效果。

下一步的工作方向:一是采用多种分类器进行分类实验,尝试将此权值计算方法与分类器相结合以期提高文本分类的整体效果;二是扩大实验范围,将之应用到中文与英文两种文本中,以便实现多语种文本分类。

参考文献:

[1] 台德艺,王 俊. 文本分类特征权重改进算法[J]. 计算机工程,2010,36(9):197-199.

[2] 熊忠阳,黎 刚,陈小莉,等. 文本分类中词语权重计算方法的改进与应用[J]. 计算机工程与应用,2008,44(5):187-189.

[3] 寇莎莎,魏振军. 自动文本分类中权值公式的改进[J]. 计算机工程与设计,2005,26(6):1616-1618.

[4] Salton G,Buckley B. Term-weighting approaches in automatic text retrieval[J]. Information Processing and Management,1988,24(5):513-523.

[5] 谭金波. 文本层次分类中特征项权重算法的比较研究[J]. 情报杂志,2007,26(9):87-88.

[6] 路永和,李焰锋. 改进 TF-IDF 算法的文本特征项权值计算方法[J]. 图书情报工作,2013,57(3):90-95.

[7] Naveenkmar N,Batri K. An empirical study on term weights for text categorization[J]. International Journal of Advanced Information Science and Technology,2012,11:43-46.

[8] 任永功,杨荣杰,尹明飞. 基于特征权重与词间相关性的文本特征选择算法[J]. 计算机应用与软件,2012,29(9):33-36.

[9] Xu Fengyang,Luo Zhengsheng. An improved approach to term weighting in automated text classification[J]. Computer Engineering and Applications,2005,41(1):181-184.

[10] Lan M,Tan C L,Su Jian,et al. Supervised and traditional term weighting methods for automatic text categorization[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2009,31(4):721-735.

[11] 裴颂文,吴百锋. 动态自适应特征权重的多类文本分类算法研究[J]. 计算机应用研究,2011,28(11):4092-4096.

[12] 侯艳钗,沈西挺. 文本分类中基于改进的词语权重算法的研究[J]. 微计算机信息,2011,27(6):211-213.

[13] 李艳姣,蒋同海. 基于改进权重贝叶斯的维文文本分类模型[J]. 计算机工程与设计,2012,33(12):4726-4730.

[14] 宋惟然. 中文文本分类中的特征选择和权重计算方法研究[D]. 北京:北京工业大学,2013.

用于文本分类的特征项权重算法改进

作者： 龚静， 胡平霞， 胡灿， GONG Jing， HU Ping-xia， HU Can
作者单位： 湖南环境生物职业技术学院 信息技术系,湖南 衡阳,421005
刊名： 计算机技术与发展 ISTIC
英文刊名： Computer Technology and Development
年，卷(期)： 2014(9)