博弈论在邮件特征选择中的应用

孙晶涛[1,2]，张秋余[1]，袁占亭[1]，董建设[1]

(1. 兰州理工大学计算机与通信学院 兰州 730050; 2. 中国石油化工股份有限公司甘肃石油分公司 兰州 730030)

【摘要】特征选择在垃圾邮件过滤中起着十分重要的作用，本文分析讨论了现有邮件特征选择方法所存在的不足，并在此基础上，提出一种基于博弈论的邮件特征选择模型。该模型将博弈论应用于邮件特征选择中，以达到约减信息规模，提高垃圾邮件过滤效率的目的。在设计特征选择模型时，考虑到邮件样本自身的模糊隶属性对特征选择所产生的影响，在特征点对邮件类别的区分度定义中，通过引入由相融性度量定义的样本模糊隶属度函数，提高博弈邮件特征选择模型对实际问题的处理能力。在CDSCE语料库上的实验表明，该邮件特征选择模型的性能优于同类其他特征选择方法，验证了该邮件特征选择模型的有效性。

关 键 词 中文垃圾邮件; 特征选择; 模糊聚类分析; 博弈论; 隶属度函数

中图分类号 TP393.098 **文献标识码** A doi:10.3969/j.issn.1001-0548.2011.01.018

Application of Game Theory for Email Feature Selection

SUN Jing-tao[1,2], ZHANG Qiu-yu[1], YUAN Zhan-ting[1], and DONG Jian-she[1]

(1. College of Computer and Communication, Lanzhou University of Technology Lanzhou 730050;

2. Gansu Oil Products Company, China Petroleum & Chemical Corporation Lanzhou 730030)

Abstract The shortages existed in the e-mail feature selection method is first analyzed. A new spam filtering feature selection model based on game theory is then proposed. The game theory is applied to feature selection of mail in order to reduce the scale of information and improve the efficiency of spam filtering. When designing the feature selection model, the impact acted by fuzzy membership of mail samples on feature selection is considered. The feature selection model's handling capacity for practical problems is enhanced by using a blending sample measure of fuzzy membership function in the definition of feature points to mail category discrimination. The experiments performed on CDSCE Corpus show that the mail feature selection is better than other feature selection methods.

Key words Chinese spam filtering; feature selection; fuzzy clustering; game theory; membership function

针对垃圾邮件愈演愈烈的现状，人们亟待寻找一种高效、准确的邮件过滤方法[1]。但是目前基于内容的垃圾邮件过滤方法在实际应用中存在如下两方面的问题：1) 人们为了完整地表达每封邮件样本的内容，避免重要特征的遗漏，往往会尽可能多得选取其包含的特征词作为采样特征点，造成特征点冗余度增加，以及邮件样本集特征空间维数地增大，在很大程度上影响了垃圾邮件过滤的效率和精度[2]；2) 在邮件特征选择中，传统特征选择函数[3-4]往往忽视实际邮件样本自身在邮件二分问题中所具有的模糊性、不确定性特点[5]，以及这些特点对邮件特征选择所造成的影响，从而照成部分噪声、冗余特征点的二次引入，降低了垃圾邮件过滤方法在实际问题中的处理性能。

因此，有必要引入新的思想来改进原有邮件特征选择所存在的不足。本文正是研究如何在邮件特征空间中选择出对邮件分类最佳的特征采样点，从而减小邮件过滤方法在问题处理中的空间复杂度。在针对邮件特征选择的研究中，通过利用邮件样本自身在二分问题中的隶属度与特征点在邮件样本集中的权重，定义特征采样点对邮件类别的区分程度，从而达到消除噪声特征点的目的。采用博弈论建立邮件特征选择模型，选择样本集中最佳特征子集，从而减少特征采样点的数量，提高垃圾邮件过滤方法的识别效率。通过在CCERT data sets of Chinese emails(CDSCE)语料库[6]上的实验表明，采用本文方法能够使邮件过滤性能得到显著提高。

收稿日期： 2009 – 07 – 22；修回日期：2010 – 01 – 13

基金项目：十一五国家科技支撑计划资助项目(2006BAF01A21)；甘肃省教育厅科研基金(0703-07)；甘肃省自然科学基金项目(0803RJZA024)

作者简介：孙晶涛(1981 –)，男，博士，主要从事中文信息处理、文本分类、网络与信息安全及网格计算方面的研究.

1 关键技术分析

1.1 博弈论

博弈论主要是研究决策主体的行为发生直接相互作用时的决策，以及决策如何达到均衡的理论[7-8]。

1.2 DCAFEM算法

模糊等价矩阵动态聚类分析算法(dynamic clustering algorithm based on fuzzy equation matrix，DCAFEM)是为了解决分类数不定，事物聚类需根据不同要求进行动态考虑而提出的一种较为实用的模糊聚类分析算法[9-10]。

DCAFEM的主要运算步骤[11]为：

1) 构造模糊聚类分析的特性指标矩阵；

2) 数据规格化；

3) 构造模糊相似矩阵；

4) 构造模糊等价矩阵，进行模糊聚类分析。

2 博弈邮件特征选择模型

2.1 模型的定义

通过采用博弈论中博弈各方策略依存性思想对邮件样本集中的特征点进行选择，其目的是为了获取最佳分类特征点，以提高邮件过滤方法的识别性能。邮件特征选择模型定义如下：

1) 参与人定义

文献[12]指出，在存在冗余和互补关系的特征空间中进行特征选择，可以看作是不同特征之间的博弈问题。在处理邮件特征选择问题中，首先利用模糊聚类算法，对邮件样本集中的特征点进行聚类，将聚类结果中属于同一类的特征点作为一个整体，定义为博弈问题中的一个参与人。对于聚类所得到的 K 个结果，即可定义为该博弈问题中的 K 个参与人。

2) 参与人的行动定义

从邮件特征选择模型对参与人的定义中可以看出，每一个参与人 $i(i=1,2,\cdots,K)$ 都是由第 i 个聚类结果中的 n_i 个同类特征点构成的，因此，将参与人 i 中所包含的 n_i 个同类特征点定义为第 i 个参与人的行动集 $S_i=\{s_1,s_2,\cdots,s_{n_i}\}$。

3) 参与人的战略定义

由于邮件特征选择作为一种信息博弈的过程，其博弈中的参与人都清楚其他方的信息得益函数，而信息博弈又是一个决策的过程，为此可将邮件特征选择的博弈过程看作为完全静态信息博弈过程[7]，因此参与人的战略集与行动集相同[8]。

4) 参与人的支付定义

在博弈问题的分析中，支付表示参与人在博弈中的所得[13]。基于博弈论的邮件特征选择模型中对参与人、参与人的行动(战略)定义可知，模型中 K 个参与人所生成的 $\prod_{i=K} n_i$ 个战略组合所对应的各个支付 $(u_1,u_2,\cdots,u_K)$ 与参与人选取的特征组合相关。例如，在战略组合 $T_j^*=(t_1^*,t_2^*,\cdots,t_K^*)$ 中，$j=\prod_{i=K} n_i$,参与人 i 采取行动 $t_i^*(t_i^*\in S_i)$，其他参与人采取行动 $t_{-i}^*=(t_1^*,t_2^*,\cdots,t_{i-1}^*,t_{i+1}^*,\cdots,t_K^*)$，那么参与人 i 的支付 $u_i=u_i(t_i^*,t_{-i}^*)$。其中 $t_{-i}^*=(t_1^*,t_2^*,\cdots,t_{i-1}^*,t_{i+1}^*,\cdots,t_K^*)$ 表示除参与人 i 以外其他参与人的行动(战略)组合。在邮件特征选择中，参与人选取哪种行动策略要参考其选取的行动对邮件的区分程度，因此特征点与垃圾邮件的相关程度 $M_{t_i^*}$ 就成为支付 u_i 定义中的一个关键要素。而在支付 u_i 的定义中，除了要考虑自身的选择外，还需要考虑到其他参与人的选择。因此参与人在战略组合 T_j^* 下的冲突关系 $R_{T_j^*}$ 也应包含在 u_i 定义中。综合上面的分析，在战略组合 T_j^* 下，支付函数的形式化定义可以表述为：

支付(u_i)=相关程度($M_{t_i^*}$)+冲突关系($R_{T_j^*}$)

2.2 参与人的设计

上节已经对文中博弈问题的参与人进行了定义，通过采用DCAFEM对邮件样本集中的特征点进行聚类，以聚类结果设计邮件特征选择博弈问题中的参与人。

2.3 参与人的支付函数设计

在博弈邮件特征选择模型的定义中，本文对战略组合 T_j^* 情况下参与人 i 的支付函数进行了形式化描述。本文在利用常用统计方法对相关程度求取的基础上，引入了邮件样本自身在二分问题中的隶属度函数；而对于支付函数 u_i 中冲突关系 $R_{T_j^*}$ 的度量，采用参与人 i 的选择与其他参与人的选择所构成的行动组合与邮件类别的相关程度作为他们之间的冲突量。由 $M_{t_i^*}=\frac{1}{G}\sum_{j=1}^{G} tf_{t_i^*,j}P(a_j)idf_{t_i^*}$ 和 $R_{T_j^*}=\frac{1}{\sum_{i=1}^{K}\frac{1}{M_{t_i^*}}}$ 可得：

$$u_i=\frac{1}{G}\sum_{j=1}^{G} tf_{t_i^*,j}P(a_j)idf_{t_i^*}+\frac{1}{\sum_{i=1}^{K}\frac{1}{M_{t_i^*}}} \tag{1}$$

式中，$tf_{t_i^*,j}$ 称为行动 $t_i^*(i=1,2,\cdots,K)$ 的局部权重；

$P(a_j)$ 即为样本 a_j 在邮件二分问题中的隶属度；$idf_{t_i^*}=\lg\left(\dfrac{G}{df_{t_i^*}}\right)$ 表示所选特征点在整个邮件样本集 A 中的全局权重；$df_{t_i^*}$ 指包含所选特征点的样本数量。

在支付的定义中，$P(a_j)$ 的求取是文中研究的一个关键点，具体计算过程如下。为了表述的方便和符号的简明，计算过程中用 x 指代邮件样本 a_j。

1) 样本 x 的局部最大离散度 $V(x)$ 定义

假设 $k\in N+$，且 $k>1$，样本 x 有 k 个近邻样本，其中属于第1类的有 p 个样本，则属于第2类的有 $k-p$ 个样本，分别记为 $x_1^*,x_2^*,\cdots,x_p^*$ 和 $x_1^{\Delta},x_2^{\Delta},\cdots,x_{k-p}^{\Delta}$。

样本 x 关于第1类的局部离散度 $V_1(x)$ 定义[14]为：

$$V_1(x)=\frac{1}{p}\sum_{i=1}^{p}(x_i^*-x)^{\mathrm{T}}(x_i^*-x) \tag{2}$$

样本 x 关于第2类的局部离散度 $V_2(x)$ 定义[15]为：

$$V_2(x)=\frac{1}{k-p}\sum_{i=1}^{k-p}(x_i^{\Delta}-x)^{\mathrm{T}}(x_i^{\Delta}-x) \tag{3}$$

样本 x 的局部最大离散度 $V(x)$ 定义为：

$$V(x)=\mathrm{Max}\{V_1(x),V_2(x)\} \tag{4}$$

2) 样本 x 的相融性 $\mathrm{Co}(x)$ 定义

对于二分类问题，由最大局部离散度可知，样本 x 对于第 q 类的局部离散度最大，$q\in\{1,2\}$，那么对于样本 x 的第 q 类的相融性 $\mathrm{Co}(x)$ 定义为：

$$\mathrm{Co}(x)=\begin{cases}\dfrac{\left(\sum_{i=1}^{p}V_1(x_i^*)/p\right)}{V(x)} & q=1\\[2ex] \dfrac{\left(\sum_{i=1}^{k-p}V_2(x_i^{\Delta})/(k-p)\right)}{V(x)} & q=2\end{cases} \tag{5}$$

3) 基于相融性度量的隶属度函数 $P(x)$ 定义

通过上述分析，样本 x 的隶属度函数定义式为：

$$P(x)=D(x)\times\mathrm{Co}(x) \tag{6}$$

3　简单数值算例

本文通过一个简单测试集说明博弈邮件特征选择模型的实际处理过程。提取6封测试邮件样本中的12个关键词，构建词-文档矩阵 X，如表1所示。

表1　词-文档原始矩阵

词	Mail1	Mail2	Mail3	s-mail1	s-mail2	s-mail3
网上	2	1	1	0	0	0
闲聊	2	0	0	0	0	0
游戏	1	1	0	0	0	0
娱乐	1	2	0	0	0	0
赚钱	1	3	2	0	0	0
上网	0	1	1	0	0	0
收入	0	1	2	0	0	0
美女	2	0	1	5	4	3
照片	0	0	0	1	3	1
写真	1	0	0	1	3	1
视频	0	0	0	1	1	1
漂亮	0	0	0	1	1	1

根据文中博弈邮件特征选择模型的定义，利用模糊聚类算法，对选取出的12个关键词进行模糊聚类，以构建该博弈问题中的参与人。具体过程如下。

首先采用夹角余弦公式构造该12个特征词的相似系数矩阵 $\boldsymbol{R}$，再通过对 $\boldsymbol{R}$ 进行关系合成运算，得到 $\boldsymbol{R}^*=\boldsymbol{R}^4\circ\boldsymbol{R}^4=\boldsymbol{R}^4$，最后选取合适的 α 值，得到相应的分类结果。文中选取 $\alpha=0.3$，得到截集 $(\boldsymbol{R}^*)_{0.3}$。由此得到{网上、闲聊、游戏、娱乐、赚钱、上网、收入}、{美女、观看、照片、视频、漂亮}这两个聚类结果，即可以定义为该博弈问题中的参与人。

$$\boldsymbol{R}^4=\begin{bmatrix}
1.000\,0 & 0.816\,5 & 0.866\,0 & 0.866\,0 & 0.836\,7 & 0.836\,7 & 0.836\,7 & 0.288\,7 & 0.288\,7 & 0.288\,7 & 0.288\,7 & 0.288\,7\\
0.816\,5 & 1.000\,0 & 0.816\,5 & 0.816\,5 & 0.816\,5 & 0.763\,8 & 0.763\,8 & 0.288\,7 & 0.288\,7 & 0.288\,7 & 0.288\,7 & 0.288\,7\\
0.866\,0 & 0.816\,5 & 1.000\,0 & 0.948\,7 & 0.836\,7 & 0.836\,7 & 0.836\,7 & 0.288\,7 & 0.288\,7 & 0.288\,7 & 0.288\,7 & 0.288\,7\\
0.866\,0 & 0.816\,5 & 0.948\,7 & 1.000\,0 & 0.836\,7 & 0.836\,7 & 0.836\,7 & 0.288\,7 & 0.288\,7 & 0.288\,7 & 0.288\,7 & 0.288\,7\\
0.836\,7 & 0.816\,5 & 0.836\,7 & 0.836\,7 & 1.000\,0 & 0.944\,9 & 0.944\,9 & 0.288\,7 & 0.288\,7 & 0.288\,7 & 0.288\,7 & 0.288\,7\\
0.836\,7 & 0.763\,8 & 0.836\,7 & 0.836\,7 & 0.944\,9 & 1.000\,0 & 0.948\,7 & 0.288\,7 & 0.288\,7 & 0.288\,7 & 0.288\,7 & 0.288\,7\\
0.836\,7 & 0.763\,8 & 0.836\,7 & 0.836\,7 & 0.944\,9 & 0.948\,7 & 1.000\,0 & 0.288\,7 & 0.288\,7 & 0.288\,7 & 0.288\,7 & 0.288\,7\\
0.288\,7 & 0.288\,7 & 0.288\,7 & 0.288\,7 & 0.288\,7 & 0.288\,7 & 0.288\,7 & 1.000\,0 & 0.870\,4 & 0.870\,4 & 0.934\,2 & 0.934\,2\\
0.288\,7 & 0.288\,7 & 0.288\,7 & 0.288\,7 & 0.288\,7 & 0.288\,7 & 0.288\,7 & 0.870\,4 & 1.000\,0 & 0.957\,4 & 0.870\,4 & 0.870\,4\\
0.288\,7 & 0.288\,7 & 0.288\,7 & 0.288\,7 & 0.288\,7 & 0.288\,7 & 0.288\,7 & 0.870\,4 & 0.957\,4 & 1.000\,0 & 0.870\,4 & 0.870\,4\\
0.288\,7 & 0.288\,7 & 0.288\,7 & 0.288\,7 & 0.288\,7 & 0.288\,7 & 0.288\,7 & 0.934\,2 & 0.870\,4 & 0.870\,4 & 1.000\,0 & 1.000\,0\\
0.288\,7 & 0.288\,7 & 0.288\,7 & 0.288\,7 & 0.288\,7 & 0.288\,7 & 0.288\,7 & 0.934\,2 & 0.870\,4 & 0.870\,4 & 1.000\,0 & 1.000\,0
\end{bmatrix}$$

通过博弈邮件特征选择模型中对支付函数的定义，计算得到参与人的支付矩阵，并根据重复剔除劣战略方法对支付矩阵进行筛选，最后得到战略组合{赚钱，美女}为参与人的最佳特征选择策略，也就是行动(战略)集的最佳特征子集。

4 算法应用实例分析

已有许多统计分类和机器学习技术被应用于垃圾邮件过滤，为了检验博弈邮件特征选择在实际应用中的效果，本文选取其中的朴素贝叶斯(Naïve Bayes)方法[16]。选择Naïve Bayes方法是因为它是最有效的启发学习算法之一，分类效果也较好[17]。在语料库的选择方面，本文选取CDSCE语料库，其中包含20 308封垃圾邮件和9 042封正常邮件。为了使实验更能突出各种特征选择算法的优势，对CDSCE语料库进行了一定程度的补充和优化，如补充了部分新类型的邮件，去除了一些过时类型的邮件，均衡了一些小样本类型的邮件数量。在通过上述补充和优化处理过的CDSCE语料库中，随机选取21 447封邮件生成训练集合 A 和5 620封邮件测试集合 B。实验平台选用IBM ThinkPad T43p，CPU为Intel Pentium M Dothan Processor 2.13GHz、内存2G SDRAM的PC机。

图1为DF、IG、MI和博弈邮件特征选择在选取的CDSCE语料集上用Naïve Bayes邮件过滤器所得到过滤效果图。从图中可以明显看出，博弈邮件特征选择是效果最好的，甚至超过了IG，DF稍差，MI的效果最差。

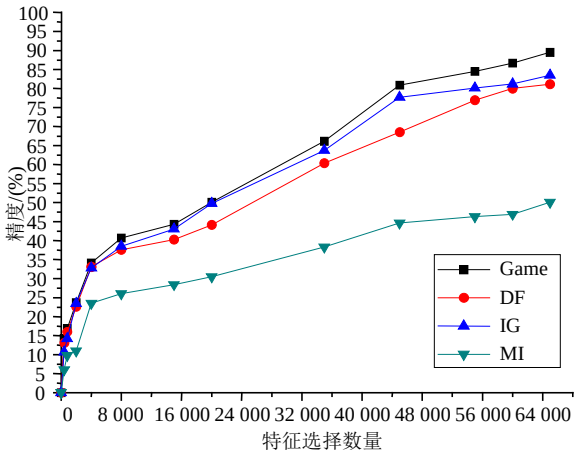

图1 在CDSCE语料集上用Naïve Bayes邮件过滤器

为了进一步验证文中所提算法的有效性，将SVM的邮件过滤算法[18-19]、Naïve Bayes的邮件过滤算法和KNN的邮件过滤算法[20]应用于训练集 A 上进行训练，然后在测试集 B 上进行邮件识别。实验结果如表2所示。

表2 博弈邮件特征选择用于Naïve Bayes方法与SVM、Naïve Bayes、KNN的实验结果比较

算法	召回率/(%)	正确率/(%)	F_1值/(%)	执行时间/s
博弈邮件特征选择应用于Naïve Bayes方法	86.3	93.37	89.7	11.16
Naïve Bayes	56.7	84.53	67.87	9.32
KNN	71.7	77.9	74.67	12.67
SVM	79.6	88.1	83.63	20.36

5 结 论

本文首先讨论了基于内容的垃圾邮件过滤方法在特征选择方面所存在的不足，并在此基础上研究和分析了将博弈论的思想应用于邮件特征选择中，在最大程度保留训练样本集信息的前提下，选择出最佳特征子集，以减少特征数量，提高邮件过滤效率。而在邮件特征选择中，邮件样本自身在二分问题中所具有的隶属度也是博弈邮件特征选择模型设计中的关键，因为其在邮件特征选择中也能做出很大的贡献。文中将相融性度量应用于原有的基于距离的样本隶属度函数定义中，既考虑了样本到所在类中心之间的距离，又考虑了样本与类的相融性，由此构造出的隶属度函数能够更加客观、准确地反映样本所存在的不确定性。通过对中文垃圾邮件的分类实验，运用博弈邮件特征选择的Bayesian邮件过滤算法，在召回率、正确率等方面都取得了较好的表现。

但该模型存在一些有待进一步研究的方面，如模型在特征选择时的收敛性以及时间复杂度等。

参 考 文 献

[1] 王斌, 潘文锋. 基于内容的垃圾邮件过滤技术综述[J]. 中文信息学报, 2005, 19(05): 1-10.
WANG Bin, PAN Wen-feng. A survey of content-based anti-spam entail filtering[J]. Journal of Chinese Information Processing, 2005, 19(05): 1-10.

[2] ZHANG Q Y, SUN J T. Technology of spam filtering based on latent semantic analysis[J]. The Systemics and Informatics World Network, 2007(04): 1265-1270.

[3] Al-MUBAID H, UMAIR S A. A new text categorization technique using distributional clustering and learning logic[J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(9): 1156-1165.

[4] CHIANG D A, KE H C, HUANG H H. The Chinese text categorization system with association rule and category priority[J]. Expert Systems with Applications: An International Journal, 2008, 35(01): 102-110.

[5] SUN J T, ZHANG Q Y, YUAN Z T. Calculation of latent semantic weight based on fuzzy membership[C]//The 5th

International Symposium on Neural Networks. Heidelberg: Springer Verlag, 2008: 91-99.

[6] 陈光英. CCERT 中文垃圾邮件过滤规则集[EB/OL]. [2009-01-10]. http://www.ccert.edu.cn/spam/sa/2005-Jul.tar.gz.

[7] 王刚, 赵海, 魏守智. 基于威胁博弈理论的决策级融合模型[J]. 东北大学学报(自然科学版), 2004, 25(1): 32-35.
WANG Gang, ZHAO Hai, WEI Shou-zhi. Decision-level fusion model based on threat game theory[J]. Journal of Northeastern University (Natural Science), 2004, 25(1): 32-35.

[8] 罗云峰. 博弈论教程[M]. 北京: 清华大学出版社, 2007.
LUO Yun-feng. Game theory[M]. Beijing: Tsinghua University Press, 2007.

[9] 陈东锋, 雷英杰, 田野. 基于直觉模糊等价关系的聚类算法[J]. 空军工程大学学报(自然科学版), 2007, 8(1): 63-65.
CHEN Dong-feng, LEI Ying-jie, TIAN Ye. Clustering algorithm based on intuitionistic fuzzy equivalent relations[J]. Journal of Air Force Engineering University (Natural Science Edition), 2007, 8(1): 63-65.

[10] 杜静, 敖富江, 杨学军, 等. 基于模糊聚类分析的构件并行技术研究[J]. 计算机学报, 2007, 30(11): 1939-1946.
DU Jing, AO Fu-jiang, YANG Xue-jun, et al. Research on component parallel technology based on fuzzy clustering analysis[J]. Chinese Journal of Computers, 2007, 30(11): 1939-1946.

[11] 叶玉玲, 伞冶. 基于遗传算法的粗糙集混合数据属性约简[J]. 哈尔滨工业大学学报, 2008, 40(5): 683-687.
YE Yu-ling, SAN Ye. Rough set reduction for hybrid data based on genetic algorithm[J]. Journal of Harbin Institute of Technology. 2008, 40(5): 683-687.

[12] 王刚. 多源信息冲突环境下的博弈融合问题[D]. 沈阳: 东北大学, 2004.
WANG Gang. Problems on the game fusion under the conflict environment of multiple source information[D]. Shenyang: Northeastern University, 2004.

[13] 王从陆, 尹长林. 基于博弈论的安全决策信息融合[J]. 中国安全科学学报, 2005, 15(4): 74-76.
WANG Cong-lu, YIN Chang-lin. Fusion of safety decision-making information based on game theory[J]. China Safety Science Journal, 2005, 15(4): 74-76.

[14] WILCOX R. A measure of coherence for human information filters[J]. Synthese, 1957, 22(3): 269-274.

[15] SANCHO R A, VERDEGAY J L. Fuzzy coherence measures[J]. International Journal of Intelligent Systems, 2005, 20(01): 1-11.

[16] ZHAN Chuan, LU Xian-liang, ZHOU Xu. An improved bayesian with application to anti-spam email[J]. Journal of Electronic Science and Technology of China, 2005, 3(1): 30-33.

[17] 肖旻, 刘晓璐, 屠立忠. 基于贝叶斯分类的邮件过滤方法及模型研究[J]. 南京师范大学学报(工程技术版), 2006, 6(2): 86-89.
XIAO Min, LIU Xiao-lu, TU Li-zhong. Research in a method and model of spam filtering based on bayesian classifier[J]. Journal of Nanjing Nor University (Eng and Technol), 2006, 6(2): 86-89.

[18] IKONOMAKIS M, KOTSIANTIS S, TAMPAKAS V. Text classification: a recent overview[C]//Proceedings of the 9th WSEAS International Conference on Computers. Wisconsin, USA: Stevens Point, 2005: 1-6.

[19] LISHUANG L, TINGTING M. Extracting location names from Chinese texts based on SVM and KNN[C]// Proceedings of 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering. Wuhan: IEEE Signal Processing Society, 2005: 371-375.

[20] LIU C M, FU S Y. Effective protocols for kNN search on broadcast multi-dimensional index trees[J]. Information Systems, 2008, 33(1): 18-35.

编 辑 蒋 晓