一种基于SVM和领域综合特征的Email自动分类方法*)

耿焕同　蔡庆生
(中国科学技术大学计算机科学技术系　合肥 230027)

摘　要　Email自动分类已成为半结构化文本信息自动处理的研究热点。本文在对已有Email自动分类方法深入研究的基础上,提出了一种基于SVM和领域综合特征的Email自动分类方法。主要包括:一是将SVM引入到Email自动分类研究中,并对SVM学习算法中的核函数和参数选择进行了探讨;二是鉴于词频的特征表示方法难以准确表示Email主要内容,因此将领域知识引入Email特征表示中,并在此基础上提出了一种综合领域知识和词频的特征表示方法,用于Email分类。该方法是在词频特征的基础上加入人工总结出的领域特征,从而更能准确地表示Email的主要内容,以提高Email分类的平均F-score。通过实验,验证了基于SVM和领域综合特征的Email自动分类方法能有效地提高Email自动分类处理的准确性。

关键词　Email自动分类,支持向量机,领域综合特征

A Novel Automatic Email Classification Method Based on Support Vector Machines and Knowledge-based Hybrid Features

GENG Huan-Tong　CAI Qing-Sheng
(Department of Computer Science & Technology, University of Science & Technology of China, Hefei 230027)

Abstract　The process of analyzing and organizing Email messages is a challenging application of Web and Text mining techniques. A novel automatic Email classification method based on support vector machines and knowledge-based hybrid features is put forward on the basis of the research of existing email classification methods in this paper. We firstly apply SVM learning algorithms to Email classification, also investigate the effects of various kernel function and feature selection. Whereas Email feature representation based on word frequency cannot represent the topic of an Email precisely, this paper presents a hybrid feature representation method for Email classification. It adds knowledge-based features in bag-of-word features to improve F-score in Email classification. Experimental results show that this method can effectively improve Email classification accurateness.

Keywords　Automatic Email classification, SVM, Knowledge-based hybrid features

1　引言

随着互联网的日益普及和快速发展,电子邮件(Email)已成为一种既快捷又经济的主要通讯方式。随着Email数量的不断膨胀及Email类型的不断增加,Email分类已成为信息处理领域的研究热点之一。Email自动分类目标包括垃圾邮件的识别和对有用邮件按用户预定义的邮件类别进行自动的分发和归类。因此,Email自动分类在垃圾邮件过滤、Email自动分发等领域具有广阔的应用前景[1,2]。

国内外已有不少学者对Email的智能处理做了不少的研究,取得一定的研究成果,主要集中在:1)垃圾邮件的过滤[3,4]。主要采用基于规则方法和基于黑名单方法,虽具有速度快和简单的特点,但存在着需用户不断地更新垃圾邮件的过滤规则和维护黑名单邮件列表的缺点,且不易扩展和非智能性。2)有用邮件的自动分类[5,6]。主要采用的方法包括基于神经网络和生物免疫机制等分类方法,虽具有自学习的特点,但需要大量的样本对网络进行训练,且训练的时间复杂度高,很难达到实用程度。

准确表示Email是影响Email分类综合性能的主要因素[7]。研究者通常使用基于词频的特征表示方法,如G. Manco等人[8]将Email的标题与内容中过滤掉停用词后的所有词语作为文本特征,并使用TFIDF作为特征权值;Y. Diao等人[9]将Email的标题与内容中不包含数字与符号的词语作为文本特征,并使用词频作为特征权值。然而,由于Email的标题与内容中的文字具有较大的分散性,因此仅使用词及词频难以准确反映Email的主要内容,从而导致Email分类的效果较差。

我们针对已有Email分类方法和特征表示存在的不足,在研究文本自动分类基础上,提出了一种基于支持向量机[10](Support Vector Machines, SVM)和领域综合特征的Email自动分类方法。主要的工作包括:一是将SVM引入到Email自动分类研究中,具有训练样本小、学习速度快、易于扩展等特点,并对SVM学习算法中的核函数和参数选择进行了探讨;二是鉴于词频的特征表示方法难以准确表示Email主要内容,因此将领域知识引入Email特征表示中,并在此基础上提出了一种综合领域知识和词频的特征表示方法,用于Email分类。该方法是在词频特征的基础上加入人工总结出的领域特征,从而更能准确地表示Email的主要内容,以提高

*)本文得到国家自然科学基金项目资助(No. 70171052, No. 90104030)。耿焕同　博士生、副教授,主要研究领域:自然语言处理、数据挖掘;蔡庆生　教授、博导,主要研究领域:人工智能、机器学习、知识发现。

Email 分类的平均 F-score。通过实验，验证了基于 SVM 和领域综合特征的 Email 自动分类方法能有效地提高 Email 自动分类处理的准确性，具有很强的实用性。

本文第 2 节介绍支持向量机理论；第 3 节详细叙述综合领域知识的特征选择和 Email 的表示方法；第 4 节给出 Email 自动分类器的设计；第 5 节是实验设计与分析；最后是结束语。

2 支持向量机理论

V. Vapnik 提出的支持向量机理论因其坚实的理论基础和诸多良好特性在近年获得了广泛的关注，已广泛地应用于模式识别、文本分类和函数逼近等。支持向量机算法是从线性可分情况下的最优分类面提出的。图 1 为一个用某特征空间上的超平面对给定训练数据集做二值分类的问题。

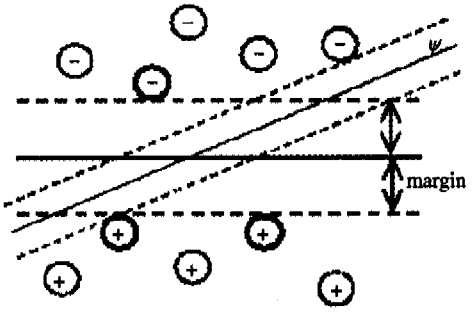

图 1 特征空间中的最优分割平面

一般地，对于给定样本点：

$$(x_1,y_1)\cdots,(x_l,y_l),x_i\in R^n,y_i\in\{-1,+1\}\quad 1\leqslant i\leqslant l \tag{1}$$

其中向量 x_i 可能是从对象样本集抽取某些特征直接构造的向量，也可能是原始向量通过某个核函数映射到核空间中的映射向量。在特征空间中构造分割平面：

$$(w\cdot x)+b=0 \tag{2}$$

使得

$$\begin{cases}(w\cdot x_i)+b\geqslant 1 & y_i=1\\(w\cdot x_i)+b\leqslant -1 & y_i=-1\end{cases}\Leftrightarrow y_i[(w\cdot x_i)+b]\geqslant 1(i=1,2,\cdots,l) \tag{3}$$

可以计算出，训练数据集到给定的分割平面的最小距离为：

$$p(w,b)=\min_{\{x_i|y_i=1\}}\frac{w\cdot x_i+b}{|w|}-\max_{\{x_i|y_i=-1\}}\frac{w\cdot x_i+b}{|w|}=\frac{2}{|w|} \tag{4}$$

根据 SVM 对优化分割平面的定义，可以看出对该平面的求解问题可以简化为：在满足条件式(3)的情况下，计算能最大化 $p(w,b)$ 的分割平面的法向量 w 和偏移量 b。

从 SVM 的理论可知，SVM 的学习能力是独立于特征空间维数的，能很好地应用于高维空间中，同时决定分类面性质的只是训练样本中的支持向量部分，因此用于训练 Email 的特征选择和 Email 的表示直接影响分类的准确性和有效性。

3 特征选择和 Email 的表示

RFC822[11] 规定了电子邮件的格式标准，电子邮件由信头(header)和可省略的信体(body)组成。信头中包含若干个域(field)，有寄信人、收信人、发信时间等信息域，因此我们可以将一封电子邮件看成如下的一个向量：($field_1$, $field_2$, …, $field_n$, body)，其中向量的每一维都是一个文本。由于 Email 中的文本具有较大的分散性，因此采用基于词频的特征表示方法难以准确表示 Email 文本的主要内容，将会导致分类的准确性较差。为了解决这个问题，本文将各类领域知识引入了 Email 的特征表示中，并在此基础上提出了一种综合领域知识和词频的特征表示方法。

3.1 基于领域知识的 Email 特征提取

在 Email 的文本中，由于文本中词语具有较大的分散性，因此一些反映类别特征的词语出现频率并不高，若单从像文本自动分类那样只考虑词语的词频特征来表示文本，反映类别特征的词语将会被基于词频的特征表示方法忽略。为了解决这一问题，我们提出一种基于类别领域知识的特征表示方法。本方法利用归纳出的类别领域关键词作为 Email 的一个特征。类别领域特征的选择方法如下：

首先，从训练的 Email 文本中总结出表示各类别特征的领域关键词；

然后，针对每个类别领域关键词 k_i 定义一个类别领域特征 K_i；

最后，确定类别领域特征 K_i 权重 WK_i，本实验中采用词频的计数表示，即如果 Email 中包含类别领域关键词 k_i 的次数为 r 次，则类别领域特征 $WK_i=r$；

3.2 综合领域特征的 Email 表示方法

首先，表示 Email 的信头结构特征，包括 From、To、Subject、Date 和邮件长度等特征；由于 SVM 分类算法对数据的要求，需对上述的特征进行数值化处理；

其次，本方法使用词频特征表示 Email 文本(包括信体和信头的 Subject)，其中词 w 的特征权值 $f(w)$ 由公式(5)确定，但考虑在不同信件位置的词应给不同的权重系数。

$$f(w)=TF(w)*IDF(w) \tag{5}$$

然后，在词频特征统计的基础上，加入代表类别领域知识的类别领域特征。添加过程如下：先从在属于类别 i 的训练 Email 中出现的所有词语中，去掉在 50% 以上的类别中出现的词语；然后在剩余的词语中选择在该类别的 Email 中出现次数较高的前 20 个词语作为类别 i 的领域关键词；最后针对 m 个类别定义 $n(\leqslant 20*m)$ 个领域特征 $K_1\sim K_n$(公式(6)，其中 r 为领域特征权值 WK_i)，并加入权值不为 0 的领域特征。

$$WK_i=r,\text{若 Email 中包含类别领域关键词 } k_i \text{ 的次数为 } r \text{ 次},1\leqslant i\leqslant n \tag{6}$$

最后，将 Email 的信头结构特征、Email 中所有词的词频特征和类别领域特征三者相结合，共同作为 Email 的特征，用于 Email 自动分类。

4 Email 自动分类器的设计

考虑到标准的 SVM 分类器只支持二值分类问题，因此我们需要对每一个类别都设置它的正例和反例的训练样本集合来实现。如果每个类的正反例样本都分别存放，那么当有 m 个类时就要定义 $2m$ 个训练样本集合。为简化系统，我们为所有的类设置一个公共的反例样本集合，只有正例样本集合是各类别专有的，使得训练样本集合的数量减少到 $m+1$ 个。对每个给定的类别，需提供用于 SVM 训练的正例和反例(即公共反例)样本集。训练的过程包括如下步骤(如图 2 所示)。

首先，对提供的正反例训练 Email 样本进行预处理，主要包括 Email 结构化信息的提取、信件内容的英文的词根化、中

文分词、去除停用词等处理。

图 2 基于 SVM 的 Email 分类器的系统模式

然后，利用本文提出的基于领域知识的 Email 综合特征提取方法对每封 Email 进行表示，数据按 Libsvm 软件[12]要求的数据文件格式进行处理：

[label][index1]:[value1][index2]:[value2] …

[indexn]:[valuen]

一行一个样本，如：+1 1:0.708 2: 0.320 … n: 0.105

其中，[label]是训练 Email 样本数据的正反例教师信号，通常用 1 表示正例样本，0 表示反例样本。[index]是从 1 开始的整数，代表样本的特征属性。[value]是一个实数，代表样本对应特征属性的值。

最后，使用 Libsvm 软件对该类的训练样本数据文件进行训练和学习。学习时必须对数据进行归一化处理，以避免大数值的特征属性支配小数值的特征属性；核函数选择 RBF 核函数，因 RBF 核函数能将非线性样本映射到一个高维空间，这样就能处理类别和特征属性间非线性的情况。同时，本文采用交叉验证法寻找 RBF 核函数最好的参数 C 和 γ，使得 SVM 分类器的分类效果最优。

在进行 Email 自动分类的测试中，就可以使用上述得到的 SVM 分类器对 Email 测试样本集进行分类测试和评估。

5 实验设计与分析

为了进一步验证基于 SVM 的 Email 自动分类方法和基于领域知识的 Email 综合特征提取的有效性，本文通过两类实验加以验证：一类是通过对垃圾邮件的过滤（其实就是二值分类）来验证基于 SVM 的 Email 自动分类方法的有效性；另一类是通过对有用邮件按不同 Email 特征提取方法来表示 Email，并应用于基于 SVM 的 Email 自动分类实验中，通过实验结果验证基于领域知识的 Email 综合特征提取的有效性。

5.1 实验的评价指标[13]

实验中，使用 *F-score* 作为测试 Email 分类综合性能的评价指标，在此基础上，使用平均 *F-score* 作为所有类别 Email 分类的平均综合性能评价指标。

$$F\text{-}score=\frac{2\times 查准率\times 查全率}{查准率+查全率} \tag{7}$$

其中，查准率 P=（分类正确的 Email 个数/分为该类别的所有 Email 个数）×100%；查全率 R=（分类正确的 Email 个数/属于该类别所有 Email 个数）×100%。

$$平均\ F\text{-}score=\frac{\sum_{i=1}^{N}第\ i\ 类\ F\text{-}score}{N} \tag{8}$$

5.2 垃圾邮件过滤实验

实验数据来自文[14]中采用的 PU1 英文 Email 样本集[15]，它是提供者一段时间收集到真实邮件。基于预处理方式的不同，PU1 提供了 4 种方式的 Email 样本集，分别为：bare、lemm、stop 和 lemm_stop 。每种方式的 Email 样本集均由 1099 封邮件组成，其中垃圾邮件 481 封、合法邮件 618 封。为减少其他无用信息的干扰，本实验中采用 lemm_stop 形式的 Email 样本集，它由 10 部分组成，每部分约 110 封邮件，每封邮件由主题和内容两部分组成。为保护隐私，PU1 的 Email 样本集对邮件内容进行了“加密”，所有的单词都用相应的数字编码表示。

实验时，先对用于训练和测试 Email 样本集进行特征抽取和 Email 的表示。由于邮件内容加密，无法基于领域知识的 Email 进行综合特征抽取方法对 Email 进行表示，因此只能采用一般的 TFIDF 统计方法对 Email 进行特征抽取和表示，但考虑在不同信件位置应给不同的权重系数。本实验中，若出现在信头 Subject 中，权重系数为 1.5；若出现在信体 Body 中，权重系数为 1.0。然后使用 Libsvm 软件对训练 Email 样本集，按选取不同特征数，分别进行训练垃圾邮件过滤器。其后用训练得到垃圾邮件过滤器对测试 Email 样本集进行测试，共重复交叉实验 10 次，分别以 9 个为训练样本集，另外一个为测试样本集，并计算平均分类的效果，最后跟 Naïve Bayes 方法进行比较。实验的结果如表 1。

表 1 基于 SVM 和 Naïve Bayes 的垃圾邮件过滤实验结果

分类器/特征数	查全率 R(%)	查准率 P(%)	F-score(%)
SVM/300	65.02	74.43	69.41
SVM/600	92.84	93.62	93.23
SVM/1200	93.28	95.06	94.16
SVM/3000	95.57	98.46	96.99
Nave Bayes	89.58	98.29	93.73

5.3 不同 Email 特征表示在基于 SVM 的 Email 自动分类中的比较实验

考虑到邮件保密性问题，同时考虑基于领域知识的 Email 综合特征抽取对 Email 的内容的要求，我们采用的实验数据取自 DELL 技术服务论坛的原帖和回帖[16]。将原帖和回帖的主要信息标题、作者和内容分别对应为 Email 中的 Subject、From 和 Body。共选取了来自 10 不同论坛的 2000 封邮件，论坛对应邮件类别，每个邮件类别大约 200 封邮件，分别随机取每类的 50%作为该类的训练正例集，另外的 50%作为该类的测试集。另外从各类中随机取与训练正例集不同的 30 封，共 300 封作为公共的训练反例集，共重复 10 次取平均值。分别采用不同的 Email 特征表示和不同 Email 文本即 Email 全文（All）和 Email 标题（Subject）进行对比实验。Email 特征表示一是采用完全基于词频统计的 TFIDF 方法来抽取特征并计算权值，另外采用本文提出的基于领域知识的 Email 综合特征抽取方法进行特征抽取和权重的计算，同时考虑在不同信件位置应给不同的权重系数（设置同上）。

实验时，先对用于训练和测试 Email 样本集进行两种不同方式的特征抽取和 Email 的表示，然后使用 Libsvm 软件对训练 Email 样本集进行训练 Email 自动分类器，最后用训练得到的 Email 自动分类器对测试 Email 样本集进行测试，并计算分类效果的平均 *F-score* 值。实验的结果如表 2。

表 2 使用不同特征表示方法实现 Email 自动分类测试结果（平均 F-score 值）

特征表示方法	All(%)	Subject(%)
词频方法	75.44	68.05
本文方法	87.32	80.33

5.4 结果分析与比较

垃圾邮件过滤实验表明，使用 SVM 分类方法比 Naïve Bayes 分类方法有更高的分类准确率。当 Email 特征数取

（下转第 57 页）

事务数增加，从而更长的事务跨度增加了与服务器端更新事务数据冲突的可能性。从图中可以看到，NBCC 对移动事务长度更敏感，当移动事务包含多个读操作时，性能急剧下降。而 QBCC-TO 的超截止期率小得多。

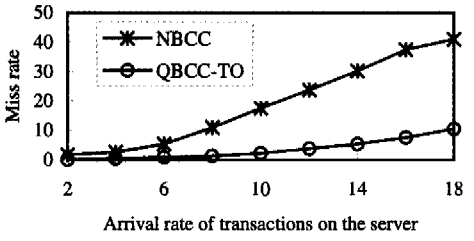

图 2　服务器事务到达率变化时的性能比较

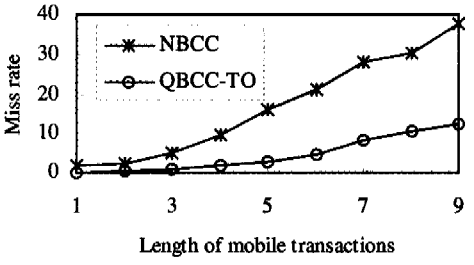

图 3　移动事务长度变化时的性能比较

结论　本文研究了数据广播环境中的一致性问题。提出使用 Q 可串行化，给出了相应的并发控制协议 QBCC-TO。实验结果表明，QBCC-TO 协议能有效改进移动事务的平均响应时间，更好地满足数据广播环境中高性能事务处理的要求。数据广播涉及许多挑战性问题，我们拟进一步研究移动客户缓存、频繁断接等的处理。

参 考 文 献

1　Madria S K, Mohania M, Bhowmick S S, Bhargava B. Mobile data and transaction management. Information Sciences, 2002, 14(1): 279～309
2　Barbara D. Mobile computing and databases—A survey. IEEE Transactions on Knowledge and Data Engineering, 1999, 11(1): 108～117
3　Pitoura E, Bhargava B. Maintaining consistency of data in mobile systems. In: Proc. of the 15th Int'l Conf. on D istributed Computing System s. 1999. 404～ 413
4　Pitoura E. Supporting read only transactions in w ireless broadcasting. In: Proc. of the DEXA '99Workshop on Mobility in Databases and D istributed Systems. 2001. 111～118
5　Acharya S, Franklin M, Zdonik S. Dissem inating updates on broadcast disks. In: The VLDB Conf. India, 1998
6　Dang Depeng, Liu Yunsheng. Concurrency control in real-time broadcast environments. The Journal of System and Software, 2003, 68(2): 137～144
7　Lindstrom J, Raatikajnen K. Dynamic Adjustment of Serialization Order Using Timestamp Intervals in Real-Time Databases. In: IEEE Sixth Intl. Conf. on Real-Time Computing Systems and Applications, Dec. 1999. 13～20
8　Kung H T, Robinson J T. On Optimistic Methods for Concurrency Control. ACM Transaction on Database Systems, 1981, 6(2): 213～226
9　Ramamritham K, Calton P. A Formal Characterization of Epsilon Serializability. IEEE Transactions on Knowledge and Data Engineering, 1995, 7(6): 997～1007
10　Chen G C, Lee S Y. An analytic model for performance analysis of concurrency control strategies in mobile environments. The Computer Journal, 1999, 42(6): 184～196

(上接第 54 页)

3000 时，SVM 方法比 Naïve Bayes 方法在 *F-score* 指标上高出 3.26%。同时，也可以看到使用 SVM 方法进行 Email 分类时，当特征数大于 600 时，分类效果的提高不再明显。这就说明了在使用 SVM 分类时只要能找出主要的关键特征，就能获得非常好的垃圾邮件过滤能力，从而能大大地提高分类的效率。因此，实验验证了基于 SVM 的 Email 自动分类方法是一种有效的分类方法。

不同 Email 特征表示在基于 SVM 的 Email 自动分类中的比较实验表明，由于综合方法在领域知识特征中增加了词频特征，削弱了对领域关键词的依赖，同时可以弥补词频特征表示方法容易忽略类别特征词语的缺陷，进而提高了分类的平均 *F-score*。当 Email 全文(All)和 Email 标题(Subject)实现分类时，与词频方法和本文方法相比，本文方法将平均 *F-score* 分别提高了 11.88%和 12.28%，从而达到了 87.32%和 80.33%的平均 *F-score*。基于领域知识的 Email 综合特征表示方法在分别对 Email 全文(All)和 Email 标题(Subject)进行分类时，结果表明 Email 全文(All)比 Email 标题(Subject)约高出 7%，说明了全文比标题含有更多的信息，反过来也说明了标题已具有很强的信息量。

结束语　Email 自动分类方法的选择和特征表示是困扰 Email 分类研究的瓶颈问题之一。针对已有 Email 分类方法和特征表示存在的不足，提出了一种基于 SVM 和领域综合特征的 Email 自动分类方法。在分类方法的选择上，我们选择了 SVM，它具有训练样本小、学习速度快、易于扩展等特点，并对 SVM 学习算法中的核函数和参数选择进行了探讨；同时在 Email 的特征表示上，将领域知识引入 Email 特征表示中，以克服词频特征表示方法难以准确表示 Email 主要内容的缺陷，并在此基础上提出了一种综合领域知识和词频的特征表示方法，用于 Email 分类中。通过实验，验证了基于 SVM 和领域综合特征的 Email 自动分类方法能有效地提高 Email 自动分类处理的准确性，具有很强的实用性。

然而在领域知识的获取上采用了人工总结的方式，在一定程度上影响了它在不同应用领域之间的适应性。因此，在今后工作中将考虑引入基于机器学习的信息抽取方法，自动获取领域知识，从而提高本方法的可适应性。

参 考 文 献

1　Bagga N A. Email classification for contact centers. In: Proc. of the 2003 ACM symposium on applied computing. Melbourne, Florida, 2003. 789～792
2　Sahami M, Dumais S, Heckerman D, et al. A Bayesian approach to filtering junk email. In: Proc. of AAAI Workshop on Learning for Text Categorization. Madison Wisconsin, 1998. 55～62
3　Segal J K B L R, Crawford J. Spamguru: An enterprise anti-spam ltering system. In: Proc. of First Conference on Email and Anti-Spam(CEAS), 2004
4　Sakkis G, Androutsopoulos I, Paliouras G, et al. A Memory-Based Approach to Anti-Spam Filtering for Mailing Lists. Information Retrieval, 2003, 6(1): 49～73
5　Clark J, Koprinska I, Poon J. A Neural Network Based Approach to Automated E-Mail Classification. Web Intelligence, 2003. 702～705
6　Secker A, Freitas A, Timmis J. AISEC: An Artificial Immune System for E-mail Classification. In: Proc. of the Congress on Evolutionary Computation, IEEE. 2003. 131～139
7　Klmit B, Yang Y. The Enron Corpus: A New Dataset for Email Classification Research. ECML04. Pisa, Italy: 2004
8　Manco G, Masciari E, Ruffolo M, et al. Towards an Adaptive Mail Classifier. AIIA, 2002. 2002
9　Diao Y, Lu H, Wu D. A Comparative Study of Classification-based Personal Email Filtering. In: PAKDD'00. Kyoto, Japan, 2000. 408～419
10　Vapnik V N. 统计学习理论的本质. 张学工, 译. 北京: 清华大学出版社, 2000
11　RFC822 http://www. faqs. org/ rfcs/ rfc822. html 2005-5-19
12　http://www. csie. ntu. edu. tw/～cjlin/libsvm 2005-5-19
13　Sebastiani F. Machine Learning in Automated Text Categorization. ACM Computing Surveys, 2002, 34(1): 1～47
14　Androutsopoulos I, Koutsias J, Chandrinos K V, et al. An Experimental Comparison of Naive Bayesian and Keyword-Based Anti-Spam Filtering with Encrypted Personal E-mail Messages. In: Proc. of the 23rd Annual International ACM SIGIR, Athens, Greece, 2000. 160～167
15　http://www. aueb. gr/users/ion/data/PU123ACorpora. tar. gz 2005-5-19
16　http://forums. us. dell. com/supportforums? ～ck=mn 2005-5-19