基于聚类的类别模糊邮件过滤方法①

郎加云 胡学钢 (合肥工业大学 计算机与信息学院 安徽 合肥 230009)

摘 要: 目前各种基于规则的分类方法在电子邮件过滤中起到了良好的效果，在邮件过滤器的训练中，训练集中会存在部分邮件具有邮件类别模糊的现象，如何将训练集中的此类类别界限模糊的邮件提取出来将会对邮件的分类效果有明显提高的作用。提出一种基于聚类的过滤方法，根据界限模糊邮件数据之间的共性特征，对邮件训练集进行聚类。实验表明，与单纯的进行基于规则的分类算法相比，这种方法在各项评价指标上具有优越性。

关键词: 聚类；文本分类；垃圾邮件

Clustering-Based Email Filtering Method with Hazy Category

LANG Jia-Yun, HU Xue-Gang

(Department of Computer and Information, Hefei University of Technology, Hefei 230009，China)

Abstract: Presently, a variety of rule-based classification methods in e-mail filtering obtain good results. In the training of e-mail filtering, the training set has the notion that some e-mail messages will be sent to the hazy category. Extracting these e-mails from training set will have a noticeable increase in the results of classification. Therefore, a clustering-based filtering method is proposed in this paper. The common features of the hazy-category email include cluster the training set. Experiments demonstrate that the method has better performance on the appraisal standard than that of a simple rule-based classification algorithm.

Key words: clustering; text categorization; spam

1 引言

随着 Internet 的发展，电子邮件已经成为当前信息交流的重要手段，然而垃圾邮件的泛滥也带来了较大干扰。垃圾邮件，是指未经用户许可，但却被强行塞入用户邮箱的电子邮件。由于垃圾邮件大多是用邮件群发软件散发的，必须借助一定的技术手段进行反垃圾邮件工作，因此自动判别垃圾邮件具有重要意义和应用价值，垃圾邮件过滤也成为重要的课题之一。

传统的反垃圾邮件技术，如实时黑名单过滤、可信白名单、主机反向名验证技术等。尽管垃圾邮件制造者可以通过伪造信头等方式绕过反垃圾邮件技术所设下的其它障碍，但是他们必须传达一定的信息，也就是邮件内容，因此，利用邮件的这些内容信息对邮件进行过滤成为一个有效的方法，内容信息包括文本内容信息，图像内容信息，超链接内容，图像文字混合等方式，其中利用文本内容信息分类算法对垃圾邮件进行识别和过滤逐渐成为反垃圾邮件技术的主流。在传统的文本分类算法中，基于统计的自动分类方法如 KNN，支持向量机，贝叶斯方法等以训练方法简单，采用向量空间模型，得到了广泛的应用。然而，由于邮件过滤采用二值分类，当训练中邮件类别之间的交叉现象比较严重时，邮件过滤的精度会大大下降。针对这一问题，本文将传统邮件二类训练集聚类为四类，分别分离出具有相反类别邮件特征的邮件集，将其作

① 基金项目:安徽省基金课题(090412044)
收稿时间:2009-12-17;收到修改稿时间:2010-01-18

训练集的一个子类，对原训练集进行提纯，去除一些噪音邮件，可以有效增强分类训练样本的可信度。实验表明，采用这种预处理方法比原先算法在垃圾邮件过滤各项评价指标上均取得更优效果，可大大提高过滤器的性能。

2 邮件过滤系统处理流程

2.1 文本的表示

文本分类是有监督的学习任务，任何文本分类算法在学习之前，都要将文本以一种合适的形式表示出来，使其适应于分类算法。本文采用向量空间模型[1]将邮件表示为向量空间中的矢量。每封邮件都可以看作是词(或词组)的序列，所有词构成一个 n 维的向量空间。本实验中，邮件矢量的分量用每个词条的权重 Tfidf 值表示，Tfidf 的计算公式如下：

$$w_i(d)=\frac{tf_i(d)\times\log(N/n_i)}{\sqrt{\sum_j(tf_j(d)\times\log(N/n_j))^2}} \quad (1)$$

其中，$tf_i(d)$ 为词条 t_i 在 d 文档中出现的词频，N 为所有文档的数目，n_i 为出现了词条 t_i 的文档的数目。

2.2 特征选择方法

特征选择是从每一类文档的所有特征中抽取那些能够反映和区分此类文档与其它类文档的特征项。在对大量文本的处理中，考虑到降低向量空间的维度、去除噪音和提高分类的精度的目的，文档在用于分类之前，需要进行特征选择以获得对分类贡献显著的单词[2]。本实验中的特征提取主要分为两步:首先删除一些出现频率很高的词，如连词、量词、语气助词等，进行粗糙的降维，然后根据词频和词熵来进行特征选择。

2.2.1 Tfidf 标准：

实验第一步使用 Tfidf 标准进行降维，用特征的 Tfidf 值(见公式(1))来评估一个特征，默认为出现的次数越少越不重要，因此可以去除一些低频词，引入 idf 逆文档频度则是为了去除一些几乎在所有文档中出现的分类均匀的特征词，其含义为：包含某特征的文档数越多，分布越均匀，则该特征越不重要。

2.2.2 MI 标准：

由于 Tfidf 标准仅考虑了特征与文档之间的关系，特征与类的关系并没有反映出来，因此实验第二步使用了可以反应类与特征之间关系的 MI 方法。MI 的公式为：

$$MI(W_i,C_j)=\log(\frac{p(W_i\mid C_j)}{p(W_i)}) \quad (2)$$

其中 W_i 为特征项，C_j 为类别，$p(W_i\mid C_j)$ 为特征项 W_i 在类别为 C_j 的文本中出现的概率，$p(W_i)$ 为特征项 W_i 在所有文本中出现的概率。

针对互信息的不足：没有考虑关键词在文档集合中出现的频率，可能会造成不同频率词互信息量的相近，如对于特征 a 和 b，a 和 b 特征在所有文档中出现的频率比为 n(n>1)倍，其在类别中出现的概率分别比也为 n 倍，那么根据公式，特征 a、b 对 C_j 的 MI 值是一样的，但是实际上，特征 a 对 C_j 明显更重要一些，因此对互信息量的计算作如下改变：

$$MI(W_i,C_j)=p(W_i\mid C_j)\log(\frac{p(W_i\mid C_j)}{p(W_i)}) \quad (3)$$

2.3 基于聚类邮件过滤方法

2.3.1 贝叶斯分类

目前基于统计的邮件分类中，贝叶斯分类算法的应用最为广泛，其中朴素贝叶斯假设一个属性虽给定类的影响独立于其他属性，即特征独立性假设，当假设成立时，与其他分类算法相比，朴素贝叶斯分类器是最精确的[3]。

朴素贝叶斯公式：

$$P(C_j\mid W_i)=\frac{P(W_i\mid C_j)P(C_j)}{\sum_{k=1}^{n}P(W_k\mid C_j)P(C_j)} \quad (4)$$

其中 $P(C_j\mid W_i)$ 表示特征项为 W_i 时文件类别为 C_j 的概率，$P(W_i\mid C_j)$ 为特征项 W_i 在类别为 C_j 的文本中出现的概率，$P(C_j)$ 为文本类别是 C_j 的概率。

2.3.2 k-means 聚类方法

数据聚类是发现事物自然分类的一种方法，也是机器学习和模式识别的一个重要研究领域。k-means 算法是最常用的方法之一。其主要思想为：

1) 从 n 个数据对象中选个 k 个对象作为初始聚类中心。

2) 根据聚类中每个对象的均值，计算样本集中每个对象与这些中心对象的距离，并根据最小距离重新对相应对象进行划分。

3) 重新计算聚类中心。

4) 循环进行 2) 3)直至聚类趋于稳定。

在 k-means 中选择不同的初始聚类中心对其聚类结果影响很大，因此如何找到与数据分类分布相一致的初始聚类中心是个关键问题[4]。

2.3.3 基于聚类思想的邮件过滤方法

目前的邮件分类系统都是将邮件分为两类，非垃圾邮件和垃圾邮件。由于邮件分类类别少且邮件维数多，当训练中两种邮件类别之间的交叉现象比较严重(两类之间的特征重复较多)或类别界限模糊时，属于合法邮件类别的邮件往往会具有垃圾邮件类别邮件的特征，此时根据传统的基于规则的二值分类器进行分类，邮件过滤的精度会大大下降。针对这一现象，可以将判断错误的邮件从原有种类中分离出来成为独立的类别，当邮件符合此类邮件特征的邮件时，将其判定为相反类别的邮件种类。

算法描述：

1) 使用贝叶斯分类方法对邮件训练集进行初步分类，由此可以产生以下四种情况，将原两类邮件标记为四类：

原为合法邮件，比较后为合法邮件，判定为合法邮件；

原为合法邮件，比较后为垃圾邮件，判定为伪合法邮件；

原为垃圾邮件，比较后为合法邮件，判定为伪垃圾邮件；

原为垃圾邮件，比较后为垃圾邮件，判定为垃圾邮件；

2) 根据步骤 1 所获得的四个类别将其作为一个四值分类来获得分类类别，计算四个类别的类中心点作为 k-means 聚类的初始聚类中心。

3) 通过 k-means 聚类方法对获得的四个小类进行聚类训练，直至聚类中心的变化幅度小于阈值 t，即每封邮件的种类趋于稳定。

4) 获得了分布稳定的四个小分类之后，将测试数据集中的数据进行贝叶斯类别判定，属于合法邮件和伪垃圾邮件的邮件将被判定为合法邮件，而属于垃圾邮件和伪合法邮件的邮件则被判定为垃圾邮件。

2.4 实验评价指标

假设待测试的邮件集合中共有 N 封邮件，N=A+B+C+D，如表 1 所示：

表 1　变量定义表

	实际垃圾邮件	实际正常邮件
系统判定为垃圾邮件	A	B
系统判定为正常邮件	C	D

定义如下评价指标衡量不同垃圾邮件过滤系统的性能：

(1) 召回率：$R=\frac{A}{A+C}$，即垃圾邮件检出率。反映了过滤系统发现垃圾邮件的能力。

(2) 正确率：$P=\frac{A}{A+B}$，即垃圾邮件检对率。反应了过滤系统“找对”垃圾邮件的能力。

(3) 精确率：$Accur=\frac{A+D}{A+B+C+D}$，即对所有邮件(包括垃圾邮件和合法邮件) 的判对率。

3 实验结果及分析

3.1 实验数据集

实验使用了 pu 语料库中的原始数据集 Pu1，Pu2，Pu3 和 Pua，PU 系列语料由希腊学者 Androutsopoulos 提供，来源于积累的真实邮件，并因个人私密将不同的词汇编码成不同的整数进行加密，可以每次取其中的 9 份作为训练集，另外一份作为测试集进行交叉验证。

3.2 算法性能评估实验

在实验中使用了 Pu1，Pu2，Pu3 和 Pua 进行了四组试验，在对原始经过加密后的邮件文本进行向量化后，首先对得到的高维文本特征向量用 tf-idf 方法去掉一些低频词和分布均匀的词后，通过 MI 方法对其进行特征降维，在实验中通过学习得到：当 MI 阈值在 [0.001，0.02]区间时比较合适。

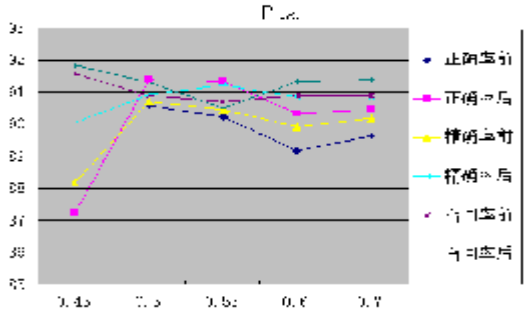

图 1　Pua 各评价指标比较

在特征选择后的数据集上，实验横向比较了通过聚类后再进行朴素贝叶斯过滤和直接进行朴素贝叶斯方法过滤的分类效果，分类结果见下图，分三个指标进行对比(X、Y 轴坐标单位均为 0.01)：

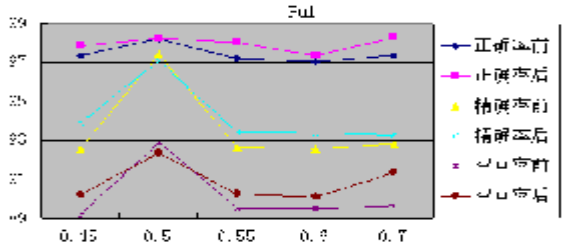

图 2　Pu1 各评价指标比较

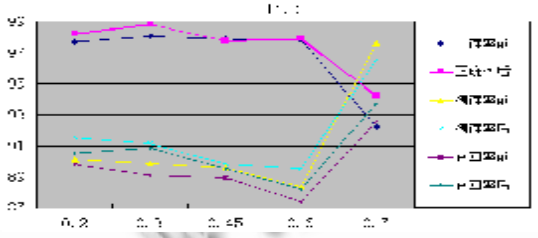

图 3　Pu2 各评价指标比较

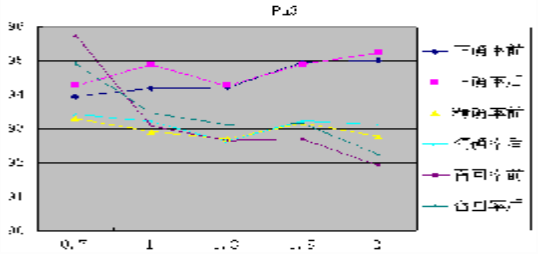

图 4　Pu3 各评价指标比较

通过实验结果比较可以看出，在对数据集进行聚类处理成四类后，系统的召回率，精确率，正确率均有所提高。可见进行此预处理是有效的。

4　总结

本文给出了一种基于聚类的垃圾邮件过滤技术。通过实验，我们可以看出，在对训练集进行聚类后，可以将具有相反类别特征的邮件很好的提取出来，将训练集进行提纯，减少分类的错误率，从前提高了精确率，正确率和召回率，取得较好的分类效果。在实际应用中，我们还可以结合人工个性化，将人对邮件的兴趣度，关注度作为评价因素，以及和为现今涌现出的大量的以图片作为背景的邮件而考虑的图像邮件过滤方法综合使用，提高邮件分类的准确性。后续工作是进一步将该系统应用到实际的电子邮件系统中。

参考文献：

1 黄萱青,吴立德.基于向量空间模型的文档分类系统.模式识别与人工智能, 1998,11(2):147－153.

2 Yang YM, Pedersen JO. A comparative study on feature selection in text categorization. Proc. of ICML-97,14th International Conference on Machine Learning, San Francisco: Morgan Kaufmann , 1997:412－420.

3 张铭峰,李云春,李巍.垃圾邮件过滤的贝叶斯方法综述.计算机应用研究, 2005,8:14－19.

4 Yang XH, Yu K, Deng W. A k-means clustering algorithm based on self-adoptively selecting density radius. International Journal of Computer Science and Network Security, 2006,6(8A):43－47.