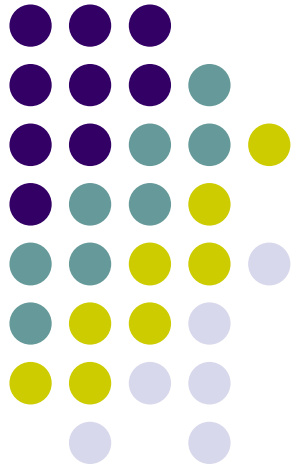
垃圾邮件过滤系统



垃圾邮件过滤

1、什么是垃圾邮件？

2、什么是过滤？如何过滤？

垃圾邮件的定义

垃圾邮件是指向未主动请求的用户发送的电子邮件如广告、刊物或其他资料;或没有明确的退信方法、发信人、回信地址等的邮件;或者利用网络从事违反网络服务供应商的安全策略或服务条款的行为和其他预计会导致投诉的邮件。

垃圾邮件的防范

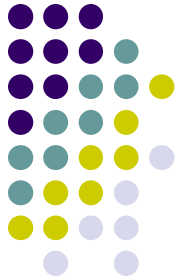
现在，采用的反垃圾邮件技术主要从三个方面来防范垃圾邮件:邮件发送方、邮件传输过程、邮件接收方。采用的主要技术有：

1、邮件服务系统的安全加固:主要措施有增强邮件服务器的安全性、提供邮件服务安全身份认证、添加反垃圾邮件的专用设备或插件等。

2、邮件过滤技术。主要技术有基于规则(如IP地址、域名、邮件地址等)和基于统计的过滤方式(基于邮件内容过滤)。

3、提高发送垃圾邮件成本，从源头上阻止垃圾邮件的产生。主要技术有电子邮票、Challenge-Response, SPE (sender policy framework)等。

Graham使用Naive Bayesian过滤垃圾邮件的理论



Paul Graham于2002年8月发表了一篇文章：A Plan for Spam，在文章中Graham提议建立垃圾邮件和非垃圾邮件单词的贝叶斯概率模型。其大体思想是，在已知的垃圾邮件中，一些单词出现的频率较高。运用一些众所周知的数学知识，对于每个特征，可以生成一个“垃圾邮件指示性概率”(spamminess probability)。根据邮件中所包含的一组词，可以用另一个简单的数学公式来确定文本邮件的“整体垃圾邮件概率”(combined probability)，也称邮件的联合概率。

算法说明

我们之所以选择贝叶斯算法，原因是由于该算法的优点在于：

1、实现简单；

2、贝叶斯模型能够自我纠正。

特征概率的算法

对于训练集中的一个特征w：

b(w)＝含有特征w的垃圾邮件数量/总的垃圾邮件 的数量；

g(w)＝含有特征w的合法邮件数量/总的合法邮件的数量；

$$p(w) = \frac{b(w)}{b(w) + g(w)}$$

p(w)是Graham方法对特征概率的估计。

特征w概率f(w)的计算：

$$f(w) = \frac{(s * x) + (n * p(w))}{s + n}$$

上式中：

n：含特征w的邮件数量；

s：一个常数参量，通常为1；

x：当n＝0时，我们需要假设的常量，也是特征w的概率，通常设为0.5；

邮件特征联合概率的算法

在过滤过程中，对于进入过滤的邮件，我们要根据训练的结果和该邮件的特征表示，给该邮件一个综合的判定值，即联合概率。然后根据设定的阈值，判定此邮件是垃圾邮件还是合法邮件。

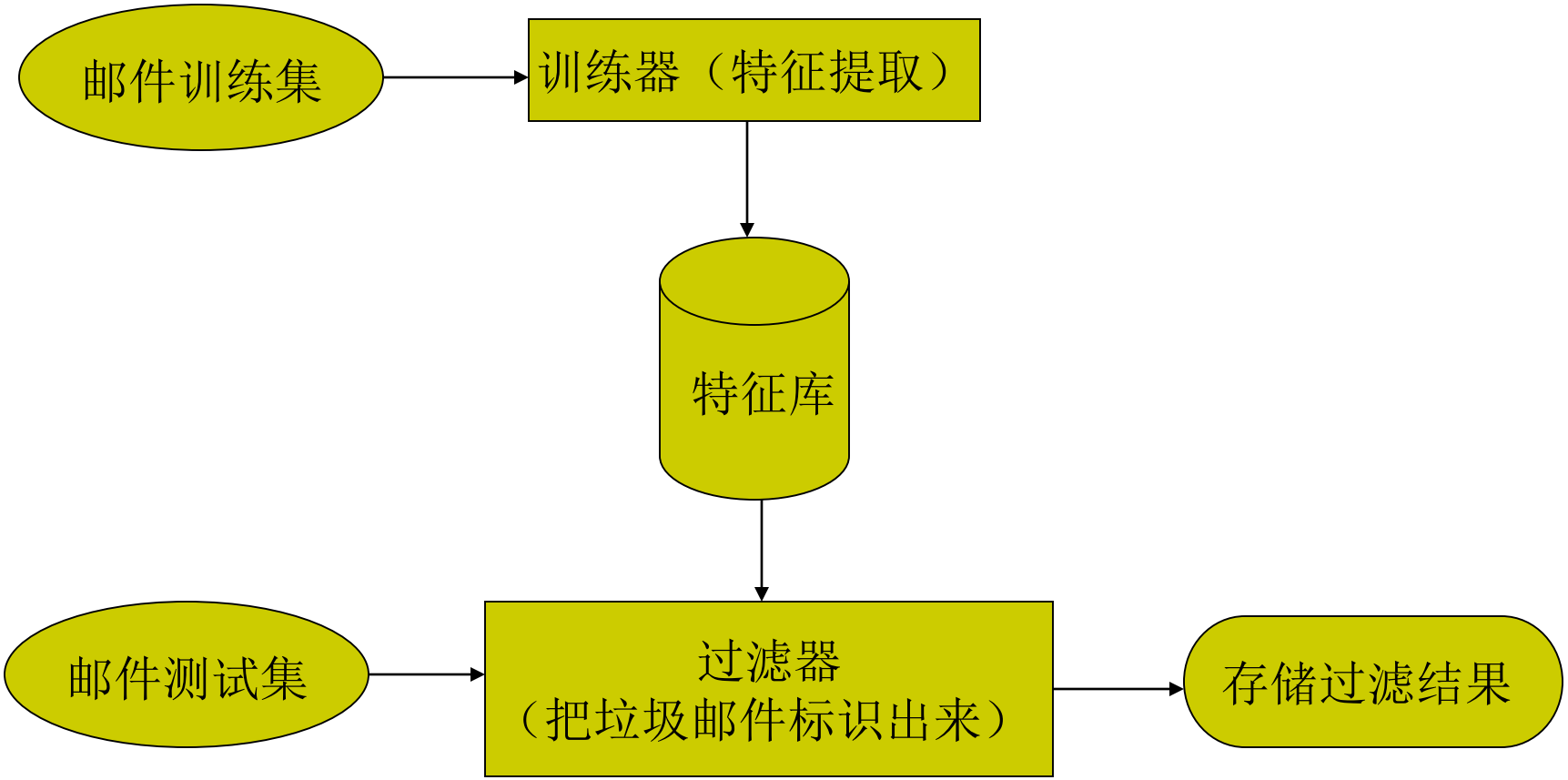
计算方法如下：

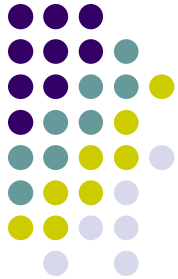
$$\frac{ABC\ldots N}{ABC\ldots N+(1-A)(1-B)(1-C)\ldots(1-N)}$$

A,B,C,...,N代表了各个特征的在哈希表hash-spamminess中的值。当邮件特征中包含以前没有从来没有出现的特征，建议特征概率为0.4。

本垃圾邮件过滤系统的工作说明

垃圾邮件过滤系统的系统流程图：

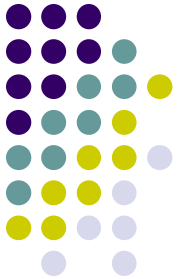




邮件提取：利用对邮件的解码提取邮件信息，包括对测试集邮件的提取和训练集邮件的提取。

特征提取：将由训练集或者测试集中的邮件，进行特征提取。在训练集中，把从邮件提取的特征，按照已定的算法进行计算，再用放入特征库中来进行储存；在测试集中，提取邮件的特征，交给下个环节处理。

在系统中，邮件阶段是通过对邮件的解码获取信息，并将邮件转换为文本。特征提取是提取邮件的主题和邮件体中的字符串，利用token串统计提取出的token串中各个token出现的次数。



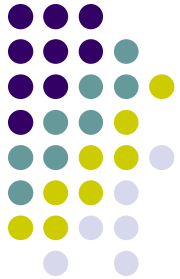
分类：接受特征提取中后的信息，根据规则数据库中的规则，按照某种相似度计算算法计算信息与实际需求的相关性，在达到一定的阀值后，输出过滤的结果。

信息表示：提供对过滤后的邮件的浏览，以及对过滤效果的评价。

分类阶段分为两个阶段：训练和测试。

训练阶段主要是训练规则库，提取spam和ham的特征；主要分三步：

解析邮件和提取特征；



建立三个哈希表：hash-good、hash-bad、 hash-spamminess； hash-good存放合法邮件中提取的特征和各特征在合法邮件中出现的次数；hash-bad存放垃圾邮件中提取的特征和各特征在垃圾邮件中出现的次数；

综合考虑hash-good和hash-bad，建立各特征的垃圾邮件指示概率(spammniess probability)， 存入哈希表hash-spamminess中。

测试阶段是利用已训练完成的规则库，对邮件进行判断。并向用户提交结果。 其过滤过程为：

对于新的邮件，提取邮件的特征，通常是最能代表邮件内容的若干个特征(这里的特征应该是它们的垃圾邮件指示性概率远离0.5的)，通过哈希表hash-spamminess计算这封新邮件的联合概率(combined probability)。如果邮件的联合概率超过某个阀值，就判此邮件为垃圾邮件，其他的为合法邮件。

- 实验报告
  - 实验名称：垃圾邮件过滤系统
  - 实验目的：
    - （1）掌握垃圾邮件过滤系统主要功能模块
    - （2）掌握在WINDOWS下安装和使用垃圾邮件过滤系统
    - （3）掌握文本内容过滤的原理
  - 实验内容
    - （1）分析并调试垃圾邮件过滤系统程序主要功能模块
    - （2）选取实验数据集
    - （3）运行WINDOWS下的垃圾邮件过滤系统
    - （4）用垃圾邮件过滤系统对实验数据集进行过滤实验
  - 系统整体描述和分功能描述
  - 实验步骤、结果及分析
  - 实验中遇到的问题及改正的方法

谢谢！！！