摘要

垃圾邮件过滤系统的评估系统的研究与实现

项涛，龚俭　东南大学

垃圾邮件的危害越来越大，已经严重危害了人民的生活和工作。针对垃圾邮件的防范研究是当前的一个研究热点，已经出现了许多优秀的垃圾邮件过滤技术和产品。然而，面对众多的垃圾邮件过滤产品，如何选择一个过滤效果好而又符合用户需求的过滤系统却依然没有一个好的依据。为此，本文以华东（北）地区网络中心为实验环境，以客观、公正为立足点，对影响垃圾邮件过滤系统过滤效果的评估指标、综合评估方法以及标准邮件集进行了研究，并依据研究的结果，设计和实现一个针对垃圾邮件过滤系统过滤能力的评估系统，为用户提供一个可以信赖的选择依据。

论文首先从现有的评估指标出发，参考相关研究领域的成果，总结、归纳和提出了四个基本指标，十二个合成指标和两个基于 ROC 曲线的指标的定义、计算方法及使用范围。在指标的计算中，论文首次将误报代价之比的概念应用到其中，提出“归一化”复合矩阵，从而屏蔽了垃圾邮件与正常邮件在重要性方面存在的巨大差异。

为给用户提供综合的选择依据，论文在第三章介绍了采用模糊综合评估方法、因子分析法和基于 ROC 曲线的 ROCCH 方法来综合评估多个过滤系统的原理和具体实现。前两种方法以当前的评估指标值为依据，判定当前配置下各过滤系统过滤能力的优劣次序。最后一种方法以 ROC 曲线为出发点，判定的是不同阀值下，过滤系统潜在的优劣顺序。

论文在第四章讨论了评测训练方法和标准邮件集对评估结果的影响。实验表明，评测训练方法和标准邮件集中的总邮件数量、垃圾邮件所占比例、训练邮件集所占比例、误报代价比以及邮件的顺序都将对评估的结果产生较大的影响。为此，鉴于现有公开的标准邮件集存在较多的缺陷以及为防止过滤系统刻意的适应静态的邮件数据，论文在第五章介绍了模拟标准邮件集生成系统的设计与实现。该系统能够根据用户的参数配置，动态生成能够用于评估的模拟标准邮件集。

依据上述评估指标和综合评估方法的研究成果，第五章还介绍了评估系统的设计与实现。依据该评估系统，第六章介绍了它采用生成的模拟邮件集和另一标准邮件集对六个垃圾邮件过滤系统的评估结果，并对评估结果进行了分析。

最后，本论文在第七章对论文的主要工作和研究成果进行了总结，并对垃圾邮件过滤系统评估研究的未来发展趋势做出了展望。

【关键字】垃圾邮件过滤系统，评估指标，综合评估方法，误报代价比，标准邮件集

Abstract

Research and Implementation of an Evaluation System Based on Spam Filtering System

XIANG Tao, GONG Jian Southeast University

Nowadays spam becomes more and more harmful, disturbing people's work and life seriously. There has been a keen interest in the research on Anti-Spam thus many excellent Anti-Spam technologies and products have been developed. However, people haven't any criteria to help them choose from numerous spam filtering products one that is both effective and meeting their requirements. Consequently, under the experiment background of CERNET eastern china (north) network center and the research of evaluation index、integrated evaluation methods and standard mail set, this thesis designed and implemented an impartial and fair evaluation system which displays the Anti-Spam ability of Spam Filtering Systems and provides the users with a reliable criterion.

Firstly, based on existing evaluation index and the conclusions of interrelated research fields, the definition and calculating methods and use scope of four basic evaluation index, twelve compositive index and two evaluation index based on ROC curve are concluded and proposed in this thesis. Besides, the rate of misclassification cost is first applied to calculating index and "normalized unit" compound matrix which balances the importance differentia of spam and ham has been proposed in this thesis.

To provide an integrated criterion for users, the third chapter in this thesis introduces the principles of fuzzy comprehensive evaluation, factor analysis and ROCCH method based on ROC curve and how to utilize them to evaluate Spam Filtering Systems. The first two methods are based on the current value of evaluation index and determine the ranking of Spam Filtering Systems in current configuration. The last one is based on ROC curve and determines the potential ranking of Spam Filtering Systems.

The forth chapter in this thesis discusses the effect of test method and standard mail set on the evaluation results. The experiments demonstrate that test method and the total number, spam rate, train rate and mail sequence of standard mail set have important effect on evaluation results. Therefore, due to the deficiencies of public standard mail set and to prevent the Spam Filtering System from adapting to the static mail data to get good results, the fifth chapter in this thesis designs and implements a generated system of simulate standard mail set, which can create a simulated standard mail set dynamicly according the user's configuration.

Based on the former chapters, the fifth chapter has also designed and implemented the evaluation system, which is used to evaluate six Spam Filtering Systems together with a simulated standard mail set and another standard mail set in the sixth chapter.

At last, a summary of this thesis is given in the seventh chapter, and the expectation of the evaluation research of Spam Filtering System is also proposed.

【Key Words】 Spam Filtering System, evaluation index, integrated evaluation methods, Rate of misclassification cost, standard mail set

东南大学学位论文独创性声明

本人声明所呈交的学位论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得东南大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

研究生签名：张涛 日 期：2007.3.22

东南大学学位论文使用授权声明

东南大学、中国科学技术信息研究所、国家图书馆有权保留本人所送交学位论文的复印件和电子文档，可以采用影印、缩印或其他复制手段保存论文。本人电子文档的内容和纸质论文的内容相一致。除在保密期内的保密论文外，允许论文被查阅和借阅，可以公布（包括刊登）论文的全部或部分内容。论文的公布（包括刊登）授权东南大学研究生院办理。

研究生签名：张涛 导师签名：[illegible] 日 期：2007.3.23

第一章 绪论

1.1 引言

随着 Internet 的快速发展，互联网已经得到了大范围的普及。作为互联网中最重要应用之一的电子邮件系统，因其本身具有方便、快捷、成本低等特点,已逐渐成为网民们不可或缺的一种交流方式。根据艾瑞市场咨询（iResearch）今年推出的《2006 年中国电子邮箱研究报告》数据显示，2006 年全球活跃电子邮箱数量（每月登录 1 次以上的电子邮箱）从 2005 年的 12 亿增加到 14 亿，而中国的活跃电子邮箱数量也增长到 2.5 亿左右。与此同时，用户发送的电子邮件也在急剧增加。而据赛迪网发布的消息，2006 年全世界互联网用户每天发送的电子邮件已经达到 600 亿封，电子邮件呈现爆破式发展。

电子邮件迅猛发展的同时，也带来很多的副作用，其中最直接的就是垃圾邮件的大量泛滥。据新浪科技报道，Web 安全和邮件安全解决方案提供商 IronPort 系统公司表示，2006 年 11 月份垃圾邮件数量猛增 35%，该月有两个时期平均每天的邮件数量高达 850 亿之多。据中国互联网协会发布的一个统计数字显示，2006 年 6 月~10 月期间，中国互联网用户收到的垃圾邮件占所收邮件总数的 59.49%，平均每天收到的垃圾邮件数量超过 2.7 封。而在中国互联网协会公布的另一个垃圾邮件调查报告中,垃圾邮件的危害被提升到了一个新的高度:中国网络用户每年收到的垃圾邮件(不算已经被系统过滤掉的) 共计 470 亿封,浪费中国网民的时间达到 15 万亿小时,给中国经济造成的损失达到 48 亿元。

垃圾邮件的大量泛滥给人们的生活、工作带来了巨大的不便和危害。英国的一项调查发现，垃圾邮件成为了最讨厌的工作环境问题之一，仅次于交通阻塞和工作时间太长。垃圾邮件至少带来下面六个方面的危害。

一、 占用大量传输、存储和运算资源，造成邮件服务器拥堵，降低了网络的运行效率，严重影响正常的邮件服务；

二、 我国开始被其他国家视为垃圾邮件的温床，许多 IP 地址有遭受封杀的危险，长期下去可能使我国成为“信息孤岛”；

三、 垃圾邮件以其数量多、反复性、强制性、欺骗性、不健康性和传播速度快等特点，严重干扰用户的正常生活，侵犯收件人的隐私权和信箱空间，并耗费收件人的时间、精力、金钱；

四、 垃圾邮件一旦被黑客利用，危害更大。2000 年 2 月，黑客首先侵入并控制了一些高带宽的网站，集中众多服务器的带宽能力，然后用数以亿计的垃圾邮件发动猛烈攻击，造成部分网站瘫痪；

五、 严重影响电子邮件服务商的形象。收到垃圾邮件的用户可能会因为服务商没有建立完善的垃圾邮件过滤机制，而转向其他服务商；

六、 妖言惑众、骗人钱财、传播色情、反动等内容的垃圾邮件，已经对现实社会造成危害。

当前，越来越多的有识之士和组织机构认识到垃圾邮件危害的严重性和紧迫性。中国互联网协会（ISC）成立的反垃圾邮件中心[1]是专门从事反垃圾邮件咨询服务机构，推动中国的反垃圾邮件事业向良性方向发展。中国教育科研网络中心计算机应急响应小组（CCERT）在 1998 年就成立反垃圾邮件小组（CAST）[2]，专门从事反垃圾邮件技术方面的研究。还有中国反垃圾邮件联盟组织（CASA）[3]，它是由一群致力于中国的反垃圾邮件

事业的朋友成立的一个反垃圾邮件民间组织，它提供一个平台，从技术、法律和人文方面讨论和研究如何反垃圾邮件。

反垃圾邮件的工作任重而道远。目前反垃圾邮件的工作主要有两种途径，一种是建立和完善相应的法律、法规；通过立法的手段来惩治恶意发送垃圾邮件的行为，从而构建和谐、稳定的网络社会。另一种是从技术手段解决垃圾邮件的问题。主要研究从邮件服务器，邮件的客户端对邮件的信封、信头、信体进行过滤。从技术手段解决垃圾邮件问题是目前反垃圾邮件工作的重点。

从技术手段对垃圾邮件进行过滤本质上是一个自动文本分类过程，其过程如图1-1 所示。首先对文本进行预处理,将文本用模型表示,进行特征提取;然后构造并训练分类器;最后用分类器对新文本进行分类。

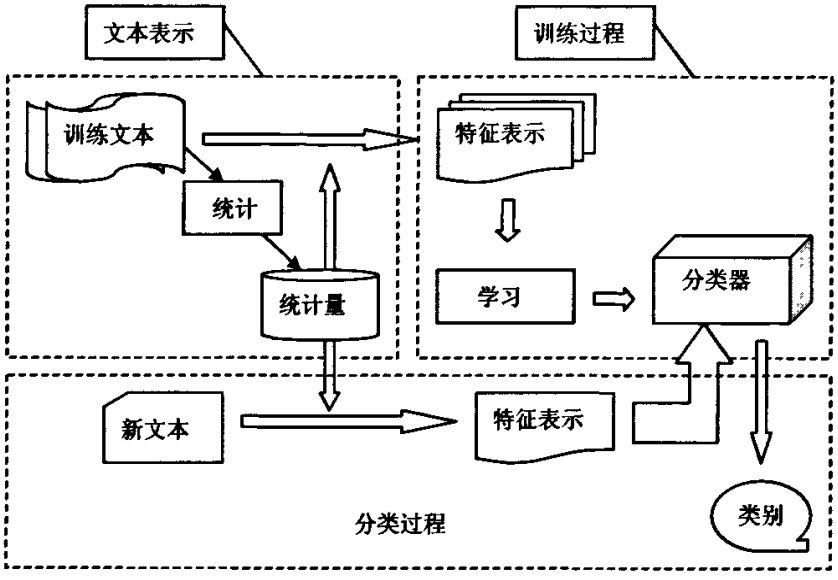

图 1-1 文本分类过程图

对于垃圾邮件过滤，邮件服务器收到的邮件为上述图中描述的文本，而邮件过滤器经过训练后将到来的新邮件分为正常邮件和垃圾邮件两大类。

反垃圾邮件技术的研究是当前的一个研究热点，已经出现了很多的反垃圾邮件产品，它们都是多个层次、多种技术结合的复杂过滤系统。本文所论述的垃圾邮件过滤系统（Spam Filtering System,SFS）是指基于 MDA 的，面向单个邮件服务器内所有用户的垃圾邮件过滤器。常见的开源 SFS 有 CRM114[4]、Dspam[5]、Bogofilter[6]、SpamProbe[7]、SpamAssassin[8]、SpamBayes[9]等等。

1.2 SFS 评估研究的背景

1.2.1 SFS 评估的意义

垃圾邮件的危害越来越大，给网民们的生活和工作带来极大的不便。垃圾邮件的防范的研究是当前的一个研究热点，已出现了很多的垃圾邮件过滤技术，也开发了很多优秀的

垃圾邮件过滤产品。然而，面对众多的垃圾邮件过滤产品，如何选择一个过滤效果好而又符合用户需求的过滤系统却依然没有一个好的依据。SFS 的评估研究是旨在从第三方的角度对 SFS 展开客观的，公正的评估；根据评估的结果，一方面，用户可以从中选择符合自己需求的过滤系统；另一方面，厂商也可以评估他们的各个不同版本的产品。

从相关研究领域来看，当某一领域的研究发展到比较成熟的阶段时，对该技术的评估研究就变得越来越重要了。从医疗诊断评估、文本分类评估到入侵检测系统的评估、网络脆弱性评估都是伴随其依托研究领域的飞速发展而诞生的。SFS 的评估正是针对目前出现的大量垃圾邮件过滤技术和垃圾邮件产品的评估，其意义不言而喻。

1.2.2 SFS 评估的基本原理

垃圾邮件过滤本质上是一个两分类问题。对于新到来的邮件，SFS 依据以往的规则和经验判定该邮件为垃圾邮件或者为正常邮件。它可能正确判定邮件的类别，也可能出现误判。SFS 的评估是将各 SFS 对一组带有标准答案的邮件进行过滤，并将判定的结果和标准答案进行比较，计算各过滤系统正确判定邮件所属类别的能力。在本文中，待评估的 SFS 包括安装程序、分类器、训练器和卸载程序四个部分。分类器用于对新邮件进行分类，而训练器用于对反馈的邮件进行自学习。评估过程中涉及到的概念有标准邮件集、评测邮件集、训练邮件集、评估指标、综合评估方法以及评测训练方法。本节介绍如下：

标准邮件集：大量的带有标准答案的邮件集合。通常由专门的研究机构收集后对外公开发布的。

评测邮件集：从标准邮件集中生成的，专门用于 SFS 评测的邮件集。

训练邮件集：从标准邮件集中生成，用于 SFS 训练的邮件集。有些情况下，评测邮件集即为训练邮件集。

评估指标：用于反映 SFS 过滤效果的参数。依据 SFS 对评测邮件集的评测结果和标准答案的比较而计算得到的。

综合评估方法：反映 SFS 整体过滤能力的方法。

评测训练方法：SFS 如何使用评测邮件集进行训练和评测的方法。

SFS 评估的一般流程为：

1）依据用户的配置，从标准邮件集中生成用户所需的评测邮件集和训练邮件集。

2）各 SFS 使用一定的评测训练方法对步骤 1 中生成训练邮件集及评测邮件集进行训练和评测，并记录评测的结果。

3）依据步骤 2 得到的评测结果，计算各评估指标的值。

4）依据步骤 2 的评测结果和步骤 3 的指标值，按照综合评估方法评估出各 SFS 的优劣顺序。

1.2.3 SFS 评估的特点

SFS 的评估研究有其自身的重要特点，它们是：

1）误报代价的不对称性。在文本分类中，将一类文本误判为另一类的代价被认为是相等的。而在 SFS 的评估中，将正常邮件误报为垃圾邮件的代价远大于将垃圾邮件误报为正常邮件，并且这种代价具有不确定性，不同的用户有不同的要求。

2）垃圾邮件所占比例的不确定性。现实中，由于垃圾邮件的发送具有很强烈的随机性，导致邮件服务器所收到的垃圾邮件所占的比例具有很大的不确定性。除此之外，不同邮件服务器具有各自的特点，所收到的垃圾邮件的比例也不相同。

3）邮件无统一标准答案。在 SFS 的评估中，将 SFS 评测的结果与标准答案进行比较作为评估的基础。但对于同一封邮件，不同的邮件用户有不同的判定标准，可能会有不同的答案而缺乏统一的标准答案。故在 SFS 的评估研究中，邮件标准答案都是通过用户反馈得到的。

4）邮件的隐私性。正常邮件涉及到很多用户的隐私，用户不愿意公开他们自己的正常邮件。标准邮件集中的正常邮件有些是采取替换技术将正常邮件中的隐私替换，有些是用户自己提供的一些正常邮件，还有一些是从其他渠道获得的一些中立的文章片断。

1.3 SFS 评估研究的现状

标准邮件集、评测训练方法、评估指标和综合评估方法是影响 SFS 评估的重要因素，在 SFS 的评估研究中，目前公开的标准邮件集主要有：Ling-Corpus[10]，PU-Corpus[10]，Spamassassin-Corpus[11]和 CCERT 提供的中文邮件标准集[12]。前三个都是英文邮件集，最后一个为中文邮件集。

Ling-Spam：它是由 Ion Androutsopoulos 收集的一组语言学家之间互相交流的 e-mail 和一些已知的垃圾邮件的混合。它包括 481 封垃圾邮件和 2412 封正常邮件，其中邮件中不包括附件，标点符号，HTML 标记等，同一天收到的多封相同邮件也只选取一封。

Spamassassin-Corpus：它包括 6047 封邮件，其中 1897 封垃圾邮件，4150 封正常邮件。正常邮件分为两部分：其中有 250 封属于难以判定的邮件，它们具有很多垃圾邮件中常有的 HTML 标记，加颜色字体等等特点，另 3900 封属于一般的正常邮件。

PU-Corpus: 它由单个用户收集其自身的邮件构成的,属于私人邮件。它包括 481 封正常邮件和 618 封垃圾邮件。正常邮件中的内容(包括单词,数字,标点等)都被替换掉了。可以从站点 http://www.iit.demokritos.gr/~ionandr 获得该标准邮件集。

除此之外，CCERT 于 2005 年对外公开了一组中文的标准邮件集。该邮件集包含两个部分，第一部分包含 25088 封垃圾邮件和 9272 封正常邮件。第二部分包含 20308 封垃圾邮件和 9042 封正常邮件。该标准邮件集中的垃圾邮件是采用 SPAMPOT（即垃圾邮件蜜罐技术）来收集的，并且不删除重复的垃圾邮件和保留邮件的原始信头；正常邮件是通过收集公开论坛所发表的最新帖子用来模仿正常邮件的主题和内容而获得的。收集到的垃圾邮件和正常帖子存放在一个临时数据库中。它们采用人工判别方式，对临时数据库中的内容进行人工分类。来自 SPAMPOT 的邮件，如果被认为是中文垃圾邮件则存放到垃圾邮件数据库；如果被认为是正常邮件则把该邮件的信头信息存放到数据库，删除内容，确保不侵犯个人隐私问题。来自公开论坛的内容，如果符合正常邮件的要求则存放到正常邮件内容数据库。该内容将与以上正常邮件的信头一起构造一个正常邮件。由于垃圾邮件和正常邮件都是实时更新，因此该中文邮件数据库中能够体现出中文邮件最新的特性。

用于SFS评估的评测训练方法有十字交叉验证法、逐一评测训练法和先训练再评测法。

十字交叉验证法的原理是将标准邮件集的邮件平均分为十部分，依次选择其中的九份作为训练邮件集，剩余的一份作为评测邮件集。如此循环十次，每次 SFS 对训练邮件集进行训练，而后对评测邮件集进行评测。将十次评测的结果平均后得到最后的评测结果。文献[14][15][16]采用十字交叉验证法对各自的 SFS 进行了评估。

逐一评测训练法的原理是各 SFS 对评测邮件集中的每一封邮件先评测并记录评测结果，然后依据该邮件的标准答案进行训练。在该方法中，训练邮件集和评测邮件集相同。文献[17]采用了该方法进行评估。

先训练再评测法的原理是首先各 SFS 对训练邮件集进行训练，训练完毕后，各 SFS 对

评测邮件集进行评测，评测的结果作为评估的依据。文献[13][18]采用该方法进行评估。

在评估指标上，为数不多的研究者对此展开了研究。

早在 1998 年，Mehran Sahami[13]等人使用未加权的垃圾邮件查全率、垃圾邮件查对率、正常邮件查全率和正常邮件查对率四个评估指标，采用先训练再评测法和自己收集的 2000 多封邮件对基于 Bayesian 方法的 SFS 进行了简单的评估。但由于他们的评测条件和评测方法与其他文献采用的不一致，他们的评测结果并不具有很强的说服力。

Harris Drucker[14]等人使用错误率、误报率和漏报率等简单指标，采用 AT&T 员工收集的 3000 封邮件和评估了基于 SVM（TF）、SVM（binary）、Rocchio 和 Boosting 算法的四个 SFS，实验结果表明基于 SVM（binary）和 Boosting 算法的两个 SFS 具有最好的过滤效果。

Ion Androutsopoulos[15]等人使用加权正确率，加权错误率以及总代价比率（TCR）等代价敏感指标，采用 Ling-Spam 邮件集和十字交叉验证法评估了基于 Navie Bayes 和基于关键字的 SFS。实验结果表明当权值 λ 较小时（$\lambda=1$或$\lambda=9$），基于 Navie Bayes 的 SFS 具有较好的过滤效果。

Zhang Le[16]等人使用垃圾邮件的查全率、垃圾邮件查对率、错误率和 F1 值四个指标，采用 Spamassassin-Corpus 邮件集和十字交叉验证法评估了基于最大熵模型、改进的最大熵模型和基于 Navie Bayes 的三个 SFS。他们还绘制了三个 SFS 的各个评估指标值随训练邮件集大小的增长而变化的情况。结果表明基于改进的最大熵模型的 SFS 具有最好的过滤效果。

Gordon Cormack[17]使用 ROC 曲线下的面积，以及随邮件数量增长，误报率的大小作为评价指标，以 8 个月的个人邮件作为邮件评测集，评估了 CRM114、DSPAM、Bogofilter、SpamProbe、SpamAssassin 等多个 SFS。

Andrew Tuttle[18]等人使用采用真实的用户反馈的邮件和垃圾邮件的查全率、垃圾邮件的误报率以及过滤的时间评估了基于 Navie Bayes、SVM 和 AdaBoost 算法的三个 SFS。

在国内，李星[19]等人采用CCERT提供中文邮件集，以正确率、错误率及加权错误率与未加权错误率的比率R为评估指标，考虑基于Navie Bayes的SFS在权重 λ 为1、9、999三种情况下，比率R随特征数目m的变化曲线；结果表明，特征数目m以及用于训练的样本个数都存在某个最优值，它们的值逐渐超过最优值时，过滤效果会略微下降并趋于一致。

当前的 SFS 评估主要为垃圾邮件过滤技术的研究者证明其过滤技术的有效性，而站在中立角度专门针对 SFS 评估的研究还非常少。Ion Androutsopoulos 和 Gordon Cormack 在这方面做了尝试，取得了一定的成绩，但还存在很多缺陷。首先，在评估指标上，他们仅仅采用了少量的几个评估指标，并且没有将这些评估指标的关系，使用的环境进行讨论。其次，他们的研究大都是通过罗列多个评价指标的值来判定 SFS 过滤效果的好坏，然而当各指标值出现不一致的情况时，没有研究综合的评估方法。再次，几乎没有研究标准邮件集和评测训练方法对评估结果的影响。本文将在上述三方面进行研究，旨在建立一个客观的、公正的 SFS 的评估系统，为用户和企业提供依据。

1.4 论文的研究目标和内容

1.4.1 研究目标

本文以 CERNET 华东（北）地区网络中心为实验环境，主要研究 SFS 的评估指标、综合评估方法、标准邮件集等影响 SFS 评估的重要因素；并依据以上研究内容，设计和实现一个尽可能公正的，中立的 SFS 的评估系统 ESFS（Evaluation System of Spam Filtering

System)。使用该评估系统，对多个 SFS 系统的过滤能力进行评估。

1.4.2 研究内容

为客观的、公正的评估 SFS，论文对影响 SFS 评估的主要因素进行研究，它们是 SFS 的评估指标，SFS 的综合评估方法以及标准邮件集。具体内容包括：

- SFS 的评估指标研究

  SFS 评估的目的是尽可能的客观的、中立的反映 SFS 的过滤效果。而过滤效果需要从多个方面来衡量，评估指标的研究主要研究能够反映、评估 SFS 过滤能力的各个参数，以及这些参数的定义、计算方法和使用场合。

- SFS 的综合评估方法研究

  SFS 评估指标的研究是研究从单个指标的角度来反映 SFS 的过滤效果，而对于某些用户或厂商可能更关心的是综合的、整体的过滤能力。SFS 的综合评估方法主要研究如何依据评估指标体系建立综合的评估模型，以便从整体上判定各 SFS 的优劣。

- 模拟邮件集生成系统的研究与实现

  为尽可能公平的反映 SFS，标准邮件集需要动态生成，这样可以防止 SFS 对静态标准邮件集的刻意适应。模拟邮件集生成系统就是按照用户的配置，动态的模拟现实邮件服务器所收到的邮件的集合，该邮件集作为评估被测 SFS 的标准邮件集。

- ESFS 系统的设计与实现

  依据上述研究内容，设计并实现合理的、方便用户使用的 ESFS 系统。

1.5 论文的组织结构

本论文的组织安排如下：

本文的第一章是绪论部分。首先是引言，在这部分论文主要介绍了垃圾邮件的现状和危害。随后介绍了论文的研究背景，它包括 SFS 评估的意义，SFS 评估的原理以及 SFS 评估的特点。本章最后介绍了 SFS 评估的研究现状、论文的研究目标和本文的组织结构。

第二章介绍了 SFS 评估的指标体系研究。首先介绍了复合矩阵和误报代价两个基本的概念，然后依据这两个重要概念，描述了四个基本评估指标、十二个合成性评估指标和两个基于 ROC 曲线的评估指标的定义、计算方法和使用场合。

第三章提出了三种 SFS 的综合评估方法。首先介绍了使用模糊综合评估方法和因子分析法评估 SFS 的原理和流程。随后介绍了使用基于 ROC 曲线的 ROCCH 方法评估 SFS 的原理。最后对这三种综合评估方法的特点和适用场合进行了分析和比较。

第四章介绍了影响 SFS 评估的其他重要因素。首先介绍了评测方法对 SFS 评估可能的影响。随后介绍了标准邮件集的垃圾邮件比例、总邮件数、训练邮件集所占的比例和邮件的顺序对 SFS 的影响。

第五章在前四章的基础上，设计和实现了一个模拟标准邮件集生成系统和 SFS 的评估系统。

第六章采用第五章的评估系统对多个 SFS 进行了评测，并对评估的结果进行了分析。

第七章总结了论文的成果，并对未来的工作提出了展望和建议。

第二章 SFS 的评估指标研究

SFS 的评估指标是衡量 SFS 系统过滤能力的主要参数。在现有的 SFS 评估指标研究中，存在很多的不足。首先，他们采用评估指标较少。大部分文献[13][14][15][16][17][18]使用不到 5 个评估指标评估 SFS 的过滤能力。这使得反映 SFS 的角度不够全面。其次，只有少量研究考虑到 SFS 的评估具有代价敏感性[15][20][21]。再次，对于通过阀值进行分类的 SFS 系统，需要绘制曲线来全面反映 SFS 系统的过滤能力。目前只有文献[17]采用 ROC 曲线评估 SFS 系统，但它没有给出 ROC 曲线评估 SFS 的原理和基于 ROC 曲线的评估指标的计算方法。

本章在现有研究的基础上，借鉴医疗诊断评估、文本分类评估等相关研究领域的成果并结合 SFS 评估的特点，总结出十六个可用于所有的 SFS 系统评估的评估指标。考虑到 SFS 评估具有的代价敏感性特点，本章定义了平均误报代价之比的概念，并认为一封正常邮件的重要性相当于多封垃圾邮件的重要性。依据这个论断，论文给出了各个评估指标的计算方法，这样使得对垃圾邮件和正常邮件的评估指标具有相同的重要性。除以上外，本章还给出两个基于 ROC 曲线的评估指标的含义和计算方法。

依据以上介绍，本章结构如下。首先介绍了 SFS 使用标准邮件集训练评测后得到的复合矩阵和不同邮件的误报代价两个概念；提出“归一化”的复合矩阵，以此为出发点，为 SFS 的评估提出了系统化的指标体系，并给出各个评估指标的定义、计算方法和使用场合；随后针对基于阀值分类的 SFS，介绍了 ROC 曲线和基于该曲线的两个评估指标。

2.1 基本概念

2.1.1 邮件的误报代价

部分文献描述了在 SFS 评估中具有的代价敏感性。在本小节，论文提出了邮件的误报代价、邮件的平均误报代价以及邮件的平均误报代价之比三个概念，并利用这三个概念来描述这种特性。

定义 2.1：邮件的误报代价

邮件的误报代价是指一封新到来的邮件被 SFS 误报为非该邮件真实类别所引起的代价。

邮件的误报代价受很多因素的影响。同一封邮件对不同的用户而言重要性不一样，甚至可能邮件的类别也不一样，有些用户认为是垃圾邮件，而其他用户却认为是正常邮件。邮件的误报代价很大程度上受到用户行为的影响。除此之外，误报代价没有一个统一的衡量标准，没法使用诸如浪费的金钱，消耗的时间等这样的度量值来衡量。

邮件的误报代价可以细分为垃圾邮件的误报代价和正常邮件的误报代价。垃圾邮件的误报代价指 SFS 将垃圾邮件误报为正常邮件而导致的代价。正常邮件的误报代价指 SFS 将正常邮件误报为垃圾邮件所导致的代价。

虽然不同的垃圾邮件间的误报代价也不尽相同，但从整体来说，正常邮件的误报代价要比垃圾邮件的误报代价高得多。邮件用户宁愿多接收一些垃圾邮件也不愿意将一封正常邮件过滤掉。

不同的邮件的误报代价各不相同，但同一类别邮件的误报代价相差较小。为衡量同一类别邮件的误报代价的大小，论文提出邮件的平均误报代价的概念。

定义 2.2：邮件的平均误报代价

邮件的平均误报代价是指所有同类别邮件的误报代价的平均值。

平均误报代价是一个估计值，因为不可能统计所有同类邮件的误报代价来进行平均。平均误报代价分为垃圾邮件的平均误报代价和正常邮件的平均误报代价。垃圾邮件的平均误报代价是所有垃圾邮件的误报代价的平均值，而正常邮件的平均误报代价是所有正常邮件的误报代价的平均值。显然，如以上所描述的那样，正常邮件的平均误报代价要远高于垃圾邮件的误报代价。

为衡量这种不同类别邮件之间的误报代价的巨大差异，引入邮件的平均误报代价之比的概念。

定义 2.3：正常邮件与垃圾邮件的平均误报代价之比（简称误报代价比）

假设正常邮件的平均误报代价为 C(H,S)，垃圾邮件的平均误报代价为 C(S,H)；则正常邮件与垃圾邮件的平均误报代价之比 $\lambda = C(H,S) / C(S,H)$。

正常邮件与垃圾邮件的平均误报代价之比是正常邮件的平均误报代价和垃圾邮件的平均误报代价的比值。通常它反映一封正常邮件被误报的后果相当于 λ 封垃圾邮件被漏报的后果，也说明一封正常邮件的重要性相对于 λ 封垃圾邮件的重要性。SFS 的评估中将使用平均误报代价之比来衡量正常邮件的平均误报代价和垃圾邮件的平均误报代价的差异。

2.1.2 复合矩阵

任何一个 SFS，对任意给定的一个标准邮件集进行训练评测后，都可以得到该 SFS 对邮件集中邮件类别的判定结果，将该结果与相应邮件的"标准答案"进行比较后，可以得到表格 2-1 中的复合矩阵(Compound Matrix)。复合矩阵的行表示 SFS 的判定结果，列表示"标准答案"的结果；复合矩阵中的每一个值表示"标准答案"为对应列所属类别，而 SFS 判定为对应行的类别总邮件数量（Spam 表示垃圾邮件，Ham 表示正常邮件）。文献[17][20]中介绍了这个复合矩阵。

表格2-1 复合矩阵

标准答案 / Filter	Spam	Ham	
Spam	TP	FP	
Ham	FN	TN	

其中：TP 表示"标准答案"为 Spam，且 SFS 也判定为 Spam 的邮件数量；

FP 表示"标准答案"为 Ham，而 SFS 判定为 Spam 的邮件数量；

FN 表示"标准答案"为 Spam，而 SFS 判定为 Ham 的邮件数量；

TN 表示"标准答案"为 Ham，且 SFS 也判定为 Ham 的邮件数量。

令 M = TP + FN、N = FP + TN、T = M + N，则 M、N、T 分别为该标准邮件集中的中

的垃圾邮件数量、正常邮件数量、总邮件数量。

SFS 的评估指标是基于上述复合矩阵中各对应下标元素值的大小及相互间的比例的大小来刻画的，然而对于不同的邮件集，比较 TP、FP、FN、TN 的大小是没有意义的，因为不同的标准邮件集的 M、N、T 可能不完全相同。实际上很多标准邮件集中的垃圾邮件数量、正常邮件数量都相差非常大。除此之外，由于 SFS 评估中存在的代价敏感性，不用标准邮件集下的平均误报代价之比 λ 也不一样。 为能够比较在不同标准邮件集下的 SFS 的评测结果，同时平衡垃圾邮件误报与正常邮件误报之间存在的巨大差异，使他们在相同地位下参与评估，论文对表格 2-1 中的复合矩阵进行了处理，提出“归一化”的复合矩阵。“归一化”后的复合矩阵见表格 2-2。

表格2-2 “归一化” 的复合矩阵

标准答案 / Filter	Spam	Ham	
Spam	tp	fp	
Ham	fn	tn	

TP、FP、FN 和 TN 的定义和获得方式与表格 2-1 一致。M = TP + FN、N = FP + TN 分别表示该标准邮件集中垃圾邮件、正常邮件的数量。

考虑到正常邮件和垃圾邮件在重要性上存在的差异，论文认为一封正常邮件的重要性相当于 λ 封垃圾邮件。基于这个论断，按照公式（2.1）计算得到“归一化”复合矩阵中各参数的值。

$$\begin{cases} T = M + \lambda * N \\ tp = TP / T \\ fp = \lambda * FP/T \\ fn = FN/T \\ tn = \lambda * TN/T \end{cases} \qquad \text{公式（2.1）}$$

“归一化”复合矩阵没有破坏原复合矩阵元素间的比例关系，它将不会影响到依赖于矩阵元素间比例的指标计算。它还引入平均误报代价之比 λ，使得垃圾邮件与正常邮件在同等地位下参与评估。除此之外，“归一化”复合矩阵还方便了记忆。假设垃圾邮件表示为阳性、正常邮件为阴性，那么 tp、fp 可称为为真阳性率、假阳性率；fn、tn 可称为假阴性率和真阴性率。这些概念在医学诊断等相关研究领域被广泛使用，将它们引入 SFS 的评估中来也容易为大家所接受。

2.2 基本评估指标

依据上述的“归一化”复合矩阵，参考相关文献，给出四个基本指标分别定义如下：

定义 2.4： 垃圾邮件查全率（Spam Recall，SR）

$$SR = tp / (tp + fn)$$

它是 SFS 正确判定的垃圾邮件数量与标准邮件集中总垃圾邮件数量的比例。通常使用它来估计 SFS 正确判定新到来的垃圾邮件的概率。

垃圾邮件查全率越高，说明 SFS 判定垃圾邮件的准确率就越高。当它等于 1 时，说明所有的垃圾邮件都能被正确的判定。理想情况下，SR 越大越好，但是由于 SR 的增加可能会导致其他指标的降低（一种极端的情况是：SFS 将所有邮件都判定为垃圾邮件，这种情况下 SR=1，但邮件系统显然已经失效）。SR 是一个常见的评估指标，文献[13][15][16][20][22][23]等都使用了该指标来评估 SFS。

定义 2.5：垃圾邮件查对率（Spam Precision，SP）

$$SP = tp / (tp + fp)$$

它是 SFS 正确判定的垃圾邮件的数量与 SFS 判定的垃圾邮件数量的比例。通常使用 SP 来估计 SFS 判定的垃圾邮件中，被正确判定的概率。

SP 的值越大，说明正常邮件误报为垃圾邮件的概率就越小。当 SP 趋近于 1 时，表示 SFS 判定的垃圾邮件基本上都确实是垃圾邮件，此时系统属于比较"保守"的状态，因为可能漏报了很多垃圾邮件。文献[13][15][16][17] [20][21][22][23]等都使用它来评估 SFS。

SR 和 SP 是两个针对 SFS 对垃圾邮件的过滤效果的评判指标，都是评判 SFS 过滤垃圾邮件的能力。但两者的出发点不一样，SR 是评估 SFS 对已知的一组确认过的垃圾邮件的判定正确的能力，而 SP 反映的是 SFS 判定的一组垃圾邮件中有多少确实是垃圾邮件。SR 是纵向比较，而 SP 则是横向比较。

定义 2.6： 正常邮件查全率（Ham Recall，HR）

$$HR = tn / (tn + fp)$$

类似于 SR，它是 SFS 判定正确的正常邮件数量与邮件集中总正常邮件数量的比例。它可以用来估计 SFS 正确判定新到来的正常邮件的概率。

HR 的值越大，说明 SFS 将越少的正常邮件误报为垃圾邮件，其过滤效果就越好。当 HR 趋近于 1 时，表明 SFS 系统对到来的正常邮件都能正确判定。文献[13][17]使用该指标评估 SFS 系统。

定义 2.7：正常邮件的查对率（Ham Precision，HP）

$$HP = tn / (tn + fn)$$

类似于 SP，HP 是 SFS 正确判定的正常邮件数量与它判定为正常邮件的总数量的比值。通常使用 HP 来估计 SFS 判定的正常邮件中，被正确判定的概率。

HP 值越大，说明 SFS 将垃圾邮件漏报的可能性越小，其过滤的效果也越好。HR 与 HP 是针对 SFS 过滤正常邮件的能力的评估指标，它们之间的关系类似于 SR 与 SP 之间的关系。文献[13][17]使用该指标评估 SFS 系统。

SR、SP、HR 和 HP 是 SFS 评估的四个基本的指标，也是最重要的指标，它们都是"越大越好型"指标。依据以上讨论的"归一化"复合矩阵和四个基本指标，可以得出如下结论：

结论 1 上述的四个指标可以确定唯一的"归一化"复合矩阵。

在四个指标的定义组成的四个方程中，只有 tp、fp、fn、tn 四个变量。根据这四个方

程，很容易求出"归一化"复合矩阵中四个变量的值。

由此，可以得出上述四个指标具有完备性，反映了"归一化"复合矩阵的全部信息。可以使用四个基本评估指标组成的一维向量来取代"归一化"复合矩阵。

另外，在给定标准邮件集和 λ 下，M 和 N 已知，依据公式（2.1）很容易求出表格 2-1 中的复合矩阵。

2.3 合成评估指标

上节叙述了基于"归一化"复合矩阵的基本指标，但有时候用户需要更直观的指标来判断 SFS 在某方面的能力。合成性指标是指可以通过基本指标表示的，反映 SFS 某特定方面能力的指标。

2.3.1 现有的合成评估指标

定义 2.8：正确率（Accuracy Rate，ACC），错误率（Error Rate，ERR）

$$ACC = (tp + tn) = (SR*M + \lambda*HR*N) / (M + \lambda*N)$$

$$ERR = 1-ACC = ((1-SR)*M + \lambda*(1-HR)*N) / (M + \lambda*N)$$

对于给定的标准邮件集，T、M、N 和 λ 是固定的，上述公式中 ACC 是可看作 SR 和 HR 的函数。正确率是 SFS 正确判定的邮件数量与所有的邮件数量的比值。通常使用它来估计 SFS 正确判定新到来的邮件的概率。

正确率越高，说明 SFS 正确判定邮件所属类别的能力就越强。当 ACC = 1 时，说明 SFS 能正确判定所有的邮件。这是最理想的 SFS，但在现实中几乎不可能达到。除此之外，由于 ACC 将 tp、tn 统一计算，它隐藏了 tp，tn 大小的差异。ERR 是 ACC 的相反数，事实上它们是同一指标，任何使用 ACC 指标的地方都可以使用 1 - ERR 来替代。ACC 和 ERR 是最常见的 SFS 评估指标。文献[14][15][16][17][21][22]均使用它们来评估 SFS 系统。

定义 2.9：垃圾邮件的误报率（Spam Misclassification，SM），正常邮件误报率（Ham Misclassification，HM）

$$SM = 1-SR,\ HM = 1-HR$$

SM、HM 分别是 SR、HR 的相反数。SM 反映的是 SFS 将垃圾邮件误报为正常邮件的概率。HM 反映的是 SFS 将正常邮件误报为垃圾邮件的概率。

SM 和 HM 都是"越小越好型"指标，当他们都为 0 时，表示 SFS 没有误报，它正确判定了所有邮件。文献[17]描述了这两个指标的定义。

定义 2.10：总代价比（Total Cost Ratio，TCR）

$$TCR = (tp + fn) / (fp + fn) = M / (M*(1-SR) + \lambda*(1-HR)*N)$$

TCR 表示不使用 SFS 导致的错误率与使用 SFS 导致的错误率的比率。由于对于确定的标准邮件集，不使用 SFS 导致的错误率是恒定的，即将所有的垃圾邮件都误报为正常邮件。

当 TCR 越大，说明由该 SFS 导致的错误率就越小，其过滤效果就越好。当 TCR < 1 时，使用该 SFS 导致的错误率比把垃圾邮件所占比例还要大，这说明再这种情况下，不使

用 SFS，直接将所有邮件判定为正常邮件的过滤效果更好。文献[15][22]介绍了该评估指标的定义。

2.3.2 新提出的合成评估指标

定义 2.11：垃圾邮件的 F_1 值（Spam F_1，SF_1），正常邮件的 F_1 值（Ham F_1，HF_1）

$$SF_1 = 2*SP*SR / (SP + SR),\ HF_1 = 2*HP*HR / (HP + HR)$$

F_1 是被广泛用于文本分类评估的一个指标，它是查全率和查对率的调和平均值。查全率和查对率是在两个不同的角度反映 SFS 的过滤能力（一个"横向"，一个"纵向"），F_1 将这两个指标综合起来，并且认为两者同样重要。

和查全率、查准率一样，F_1 也是"越大越好型"指标。当 F_1 为 1 时，可以推导出查全率和查对率都等于 1，这是最理想的过滤系统。SF_1 和是 HF_1 分属两类邮件下的 F_1 值，它们分别用于评估 SFS 针对本类邮件的过滤能力。文献[16][20][24]中介绍了 F_1 指标，但本文是首次将这个指标分为 SF_1 和 HF_1，并将它们分别用于评估 SFS 对垃圾邮件和正常邮件的过滤能力。

定义 2.12：垃圾邮件概率比 LR，正常邮件概率比 ZR

$$LR = tp / fp = SP / (1-SP),\ ZR = tn/fn = HP / (1-HP)$$

LR 和 ZR 都是从医疗诊断研究中引入的一个评估指标。它们是邮件的查对率与其相反数之比，它们的值越大，表明 SFS 判断准确的概率就越大，过滤系统的过滤效果就越好。文献[25]使用它评估医学影像设备，本文将其引入评估 SFS 系统。

定义 2.13：约登（Youden's）指数 YI，YI = SR-(1-HR)

约登指数是广泛用于医疗诊断评估研究的一个指标，它等于垃圾邮件与正常邮件的查全率之和减 1。它表示 SFS 过滤垃圾邮件和正常邮件的总能力。

和大多数评估指标一样，YI 指数也是"越大越好型"指标，在医疗诊断研究中，YI 指数通常作为诊断实验中 ROC 曲线的最佳临界点的判定指标。文献[25]使用它评估医学影像设备，本文将它引入 SFS 的评估中。

定义 2.14：平均误报代价（Average Misclassification Cost，AMC）

假设垃圾邮件的误报代价为 C(S,H)、正常邮件的误报代价为 C(H,S)，则定义所有邮件的平均误报代价 AMC = (M*(1-SR)*C(S,H) + (N*(1-HR)*C(H,S))) / (M+N)。

邮件的平均误报代价 AMC 是所有邮件导致的总误报代价与邮件总数量的比值，它表示每一封邮件的平均误报代价。如果令 C(S,H)=1，那么平均误报代价可以表示为 AMC = (M*(1-SR) + (N*(1-HR)* λ)) / (M+N)。

AMC 是一个"越小越好型"评估指标，当 AMC = 0 时，表示 SFS 没有误报。当 λ 较大时，平均误报代价不超过 λ。文献[23]使用该指标评估多个基于机器学习算法的分类器，本文结合 SFS 评估的特点，将其应用到 SFS 的评估中来。

定义 2.15：相关系数（Matthew's Correlation Coefficient，MCC）

$$MCC = (tp*tn - fp*fn) / ((tn+fn) * (tn+fp) * (tp+fn) * (tp+fp))^{1/2}$$

MCC 是从用于机器学习性能评估的评估标准，它反映 tp，tn，fp，fn 之间的关联关系。当评测邮件集中的垃圾邮件和正常邮件数量相差很大的时候，MCC 要好于普通的正确

性指标 ACC。因为如果某类邮件占评测邮件集的绝大多数时，对于将所有邮件都判定为这类多数邮件的类别也能获得很高的正确性，而 MCC 可能很好的区分这种情况。MCC 见于评估其它分类研究，，本文首次将该评估指标引入 SFS 的评估中来。

2.4 曲线性指标

对于通过阀值分类的 SFS，它们的过滤过程是：首先用户设定一个阀值，对于新到来的每一封邮件，过滤系统都将计算该邮件的分值；当分值大于阀值时，将该邮件标记为垃圾邮件，否则标记为正常邮件。阀值的大小直接影响到 SFS 对邮件类别的判定。因此，对于这类 SFS，基本指标和合成指标反映的是某一阀值下该 SFS 的过滤效果，而仅仅使用它们来评估 SFS 的整体过滤能力显然是非常片面的。为全面的评估通过阀值分类的 SFS，需要考虑在不同的阀值下，各评估指标值的情况。

鉴于上述评估指标存在的使用范围的缺陷，在本文中，介绍 ROC 曲线来评估 SFS 系统。本文将详细描述 ROC 曲线评估 SFS 系统的原理、如何使用 ROC 曲线评估 SFS、以及给出基于 ROC 曲线的评估指标的定义和计算方法。

2.4.1 ROC 曲线

ROC（Receiver Operating Characteristic）曲线是对不同的阀值下的真阳性率和假阳性率作图所得的曲线。它源于信号探测理论(signal detect theory)，最早用于描述信号与噪声之间的关系，而后在气象、材料检验、心理物理学等领域都有应用。19 世纪六十年代，ROC 曲线分析开始应用于医学诊断，随后经大量学者的研究与实现，目前 ROC 曲线分析已经成为广泛用于医学诊断和人群筛减试验评价的一种统计方法。它以真阳性率即灵敏度为纵坐标，以假阳性率即（1一特异度）为横坐标，通过构图法反映不同阀值下的灵敏度和特异度之间的关系。通过 ROC 曲线来分析、评价和比较筛检试验的价值等等。

ROC 曲线的特点可以归纳为：

1）横坐标的值通常是越小越好，纵坐标的值通常是越大越好。

2）曲线上点的横坐标与纵坐标相互制约，横坐标的增加会导致纵坐标的增加。

3）相比于纵坐标，通常用户不能接受横坐标值过大的情况。

4）曲线从（0，0）点逐渐变化增大到（1，1）点。

对于垃圾邮件分类，它类似与医疗诊断，本质上也是个二分类问题。SFS 将新收到的邮件分类成垃圾邮件和正常邮件，而分类的准确性是是衡量 SFS 好坏的最重要的因素。

鉴于 SFS 的评估与医学诊断有着非常类似的关系，论文将绘制 SFS 的 ROC 曲线，以此来评估 SFS 的过滤效果。

2.4.2 SFS 的 ROC 曲线

ROC 曲线反映的是真阳性率与假阳性率之间的关系，在 SFS 的评估中，存在这样的指标对有着类似的关系，一方面它们中存在不一致的关系，其中一个指标值越大过滤系统的效果越好，而另一个指标值越小过滤系统的指标值越好；另一方面它们存在协调的关系，其中一个指标值的增大必然会导致另一个指标值的增大。ROC 曲线最擅长处理这类互相牵制而又协调的关系。

SFS 的 ROC 曲线（简称为 ROC 曲线）又可称为 SFS 的 HM-SR 曲线（简称为 HM-SR

曲线)。它是以 HM=1-HR 为 x 轴，SR 为 y 轴的二维曲线，它反映的是垃圾邮件的查全率随正常邮件的误报率变化的情况。当给定一个分类阀值时，分类器对标准邮件集评测后分别计算 HM 和 SR 可得到该 ROC 曲线空间上的一点。通过不断的由小到大的调整分类阀值，得到 ROC 空间上的一系列点，这些点拟合成的曲线就称为该 SFS 的 HM-SR 曲线。

在 HM-SR 曲线空间中，有几个重要的点。左下角的（0，0）点表示 SFS 将所有的邮件都判定为正常邮件，其效果相当于不设置过滤系统。与它相反的另一个极端是，右上角的（1，1）点表示 SFS 将所有到来的邮件都判定为垃圾邮件，此时分类阀值设置为最小。除此之外，左上角的（0，1）点表示所有的邮件都判定正确，它代表最理想的 SFS 的过滤效果，这在现实中很难达到。在 ROC 曲线中，如果一点的在另一点的西北方（即：*SR* 更大或 *HM* 更小或两者皆是），那么该点的过滤效果好于另一点。曲线从（0，0）点开始，沿向上凸的曲线滑到（1，1）点。随着 HM 的增大，SR 也逐渐增大。但由于用户通常不能接受正常邮件的误报率太大的情况，通常选择可接受的垃圾邮件误报率下尽可能大的正常邮件查全率。

ROC 曲线反映的是两评估指标 HM、SR 之间的一种平衡关系，纵坐标 SR 的值越大越好，而横坐标 HM 的值越小越好。

2.4.3 基于 ROC 曲线的评估指标

ROC 曲线是一个二维的曲线图，它反映了 SFS 的过滤效果,但它不够直观。通常采用基于 ROC 曲线的一维评估指标来评价 SFS 的过滤效果。基于 ROC 曲线的评估指标有很多，其中 ROC 曲线下的面积（Area Under Curve，简称 AUC）[26]是最常见也是最重要的一个。它反映过滤系统判断邮件的准确性，理论上该指标在 0.5 到 1 之间。该面积值越大越，.完全无价值的诊断为 0.5，完美的诊断为 1。除此之外，论文还提出了最优分割点指标，该指标给出了一个基于分类阀值的 SFS 可以达到的最优过滤效果和最优效果对应的分类阀值。

- **ROC 曲线下的面积。**ROC 曲线是对点坐标的连接绘制的曲线，它通常是不规则的曲线，所以通常很难通过曲线拟合的方式来计算曲线下的面积。在本论文中，将采取梯形法来计算 ROC 曲线下的面积及误差。梯形法[27]的基本原理是使用 ROC 曲线下的许多微小梯形的面积和来估计 ROC 曲线下的面积。其基本的原理见下图 2.1。

在图 2.1 中，从（0，0）点开始沿曲线选择曲线上的点，分别计算曲线上相邻两点之间构成一个小梯形的面积，如图 2.1 中的梯形 T_{ABED} 和 T_{BCFE} 的面积，并将这些小梯形的面积相加，一直到（1，1）点，最后用所有梯形的面积和来估计整个曲线下的面积。当在曲线上选取足够多的点且相邻点足够近时，用梯形法估计的曲线面积的误差可以忽略不计。

对于给定的一组 ROC 曲线上的点，设为 $(a_1,b_1),(a_2,b_2),\cdots\cdots(a_n,b_n)$，ROC 曲线下的面积可用 *AUC* 来估计，*AUC* 的计算公式为：

$$AUC=\frac{a_1b_1}{2}+\sum_{i=2}^{n}((b_{i-1}*\Delta a_i)+\frac{1}{2}(\Delta b_i*\Delta a_i)) \qquad \text{公式 (2.2)}$$

其中

$$\Delta b_i=b_i-b_{i-1}$$

$$\Delta a_i=a_i-a_{i-1}$$

用 AUC 来估计 ROC 曲线下的面积的误差不超过 $\Delta AUC = \sum_{i=1}^{n} \frac{1}{2}(\Delta b_i * \Delta a_i)$，

即 ROC 曲线的面积范围为 $[AUC, AUC+\Delta AUC]$。

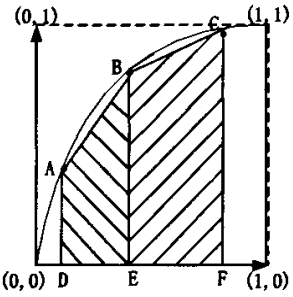

图2.1 梯形法求曲线面积

● **最优分割点。**ROC 曲线下的面积是在整体上对 SFS 的过滤准确性的判定。但有时候用户只关心该 SFS 能够达到的最优的过滤效果，因此本节提出最优分割点[28]指标，用于找到 SFS 能够达到的最优的过滤效果及相应的分类阀值点。最优分割点的选择是基于错判造成的总损失量达到最小的原则，其方法原理如下：

假设在评测邮件集中随机选取一封邮件，它为正常邮件的概率为 P_0，为垃圾邮件的概率是 $P_1=1-P_0$。设 SFS 计算评测邮件的得分为 X，要选择的切入点设为 X_1,并且 $X_0<X_1<X_2$（其中 $X_0=\min$，$X_2=\max$）。若邮件的得分在 $X_0<=X<=X_1$，则判定该邮件为正常邮件，若得分在 $X_1<X<=X_2$，则判定该邮件为垃圾邮件。定义 $F_i(X)$为第 i 类邮件的得分变量 X 的累积分布，$F_i'(X)$为相应的概率密度函数，$F_i(X_0)=0$，$F_i(X_2)=1$（其中 i=0,1）。

以 $L_{2\times2}$ 表示错误判断邮件类别所造成的代价矩阵，该矩阵包含的元素 L_{ij} 表示一个第 i 类的邮件被误报为第 j 类邮件所造成的代价。若 i=j 则 $L_{ij}=0$；否则 $L_{ij}>0$；损失矩阵如下表所示（λ 表示正常邮件与垃圾邮件的平均误报代价之比；i=0 表示正常邮件，i=1 表示垃圾邮件）：

表2-3 损失矩阵 L_{ij}

i/j	0	1	
0	0	λ	
1	1	0	

设 SFS 的由于误报而造成的总的期望损失是 T，则：

$$T(X_1)=\sum_{i=0}^{1}\sum_{j=0}^{1}P_iL_{ij}\{F_i(X_{j+1})-F_i(X_j)\}=P_0\lambda(1-F_0(X_1))+P_1F_1(X_1)$$

若要使得 T 最小，X_1 满足以下条件：

$$\partial T/\partial X_1=0,\qquad \partial^2 T/\partial X_1^{\,2}>0;$$

得到：$\partial(1-F_1(X_1))\Big/\partial(1-F_0(X_1))=\frac{P_0\lambda}{P_1}$。　　公式（2.3）

其中，$1-F_1(X_1)$相当于临界值为X_1时的垃圾邮件查全率，$1-F_0(X_1)$相当于对应的正常邮件误报率，因此 ROC 曲线实际上也就是以$1-F_1(X_1)$纵坐标，$1-F_0(X_1)$为横坐标随X_1变化的曲线。

依据上述公式，在表 2-3 所示的损失矩阵的假设下，ROC 曲线上的最优分割点，也就是 ROC 曲线上斜率等于 $\frac{P_0\lambda}{P_1}$ 的点。它的纵坐标和横坐标的值就是该 SFS 所能达到的最好的过滤效果。

2.6 本章小结

本章首先介绍了用于 SFS 评估的两个重要概念复合矩阵和误报代价，并引申出平均误报代价、平均误报代价之比等概念的定义。依据这些概念，论文将平均误报代价之比的概念融入到复合矩阵中,提出了"归一化"的复合矩阵。在此基础上,论文给出了四个基本指标、现有的五个合成性指标和本文提出的七个合成性指标的定义、计算方法和使用范围。对于通过阀值进行分类的 SFS，为全面反映这类 SFS 的过滤能力，论文还给出了 ROC 曲线，以及基于该曲线的一维评估指标的定义和计算方法。

第三章 SFS 的综合评估方法

第二章中介绍了 SFS 的评估指标体系，但是光有 SFS 的评估指标是不够的。因为各个评估指标都是反映 SFS 某一方面的过滤能力，但对于普通用户来说，由于对 SFS 的内部不太了解，他们需要的是一个对 SFS 的综合的评估结果，用户可以直接比较整体的评估结果而做出系统优劣的判断。因此论文需要建立 SFS 的综合评估方法，用以综合的评估 SFS 的整体过滤能力。

本章将描述应用于 SFS 评估的三种综合评估方法，有对多个评估指标的加权差值进行模糊评估的模糊综合评估方法，有从所有评估指标中提炼出相互独立的少数几个综合指标用于评估的因子分析法，还有考虑邮件的误报代价、邮件到达比例的 ROCCH 评估方法。最后论文对三种方法的特点和使用场合进行了比较，并给出本章小结。

3.1 模糊综合评估方法

模糊综合评估是借助模糊数学的一些原理和方法，将一些边界不清、不易定量的因素定量化，继而进行综合评估的一类方法。其评判的过程是：将评价目标看成是由多种因素组成的模糊集合（称为评估向量 u），再设定这些因素所能选取的评估等级，组成评语的模糊集合（称为评价等级集合 v），分别求出各单一因素对各个评价等级的归属程度（称为模糊矩阵），然后根据各个因素在评价目标中的权重分配，通过计算（称为模糊矩阵合成），求出评价的定量解值。

模糊综合评估法再现实中有着广泛的使用，文献[29]介绍了模糊综合评判的一般原理和方法，并使用它评估了国内各个省、直辖市的综合实力。文献[30]介绍了使用模糊综合评估方法评估入侵响应的效果。文献[31][32]使用该方法评估了入侵检测系统的检测效果。文献[33]采用模糊综合评估法评估了贵阳市、昆明市和成都市在城市品牌方面的先后次序。

针对 SFS 的评估中存在多个评估指标经常出现不一致的情况（一些过滤系统在某几个指标上优于其他系统，而在其他指标上又不如他们），论文认为可以采取模糊综合评估的方法来综合判定 SFS 的过滤效果。使用它评估的基本流程为：

1）确定用于评估 SFS 的各评估指标，组成评估向量。建立各评估指标的评估等级。

2）确定隶属度函数，用于计算隶属度矩阵，它描述了评估指标属于各评估等级的程度。

3）确定各指标权重。

4）综合计算 SFS 属于各评估等级的定量值。

3.1.1 评估指标与评估等级

如上一章所述，可用于评估任何类型的 SFS 的指标有基本指标和合成性指标。这些指标都可用于 SFS 的综合评估，但由于一些指标仅仅是互为相反数的关系，实质上同一指标，故只选取其中的一个。对这些指标进行简单分类可表示如下：

表 3-1 评估指标的分类

指标所属类别	评估指标	取值类型	取值范围

垃圾邮件	垃圾邮件的查全率 SR	实数型	[0, 1]
	垃圾邮件的查对率 SP	实数型	[0,1]
	垃圾邮件的 F_1 值 SF_1	实数型	[0, 1]
	垃圾邮件概率比 LR	实数型	[0,+∞)
正常邮件	正常邮件查全率 HR	实数型	[0, 1]
	正常邮件查对率 HP	实数型	[0, 1]
	正常邮件的 F_1 值 HF_1	实数型	[0, 1]
	正常邮件概率比 ZR	实数型	[0,+∞)
两类邮件	正确率 ACC	实数型	[0, 1]
	约登指数 YI	实数型	[-1,1]
	平均误报代价 AMC	实数型	[0,+∞)
	总代价比 TCR	实数型	[0,+∞)
	相关系数 MCC	实数型	[-1,1]

除平均误报代价 AMC 以外，表 3-1 中所列出的评估指标都是"越大越好型"指标。故在论文中使用 AMC 的倒数作为评估平均误报代价的指标，记 RAMC=1/AMC。SFS 的一维评估向量为 U = {SR，SP，SF_1，LR，HR，HP，HF_1，ZR，ACC，YI，RAMC，TCR，MCC}。

为比较 SFS 两两之间的过滤效果，论文计算 $\Delta U=U_2-U_1$（ΔU 向量中分量的值为 U_2、U_1 中对应分量值的差）。ΔU 向量用来比较两个 SFS 间的差别等级。

表 3-1 所述评估指标、指标的特性在模糊理论中被视为评估因素。SFS 的评估因素呈现树型结构，为更好的描述该集合，论文引入有序模糊评价树（Ordered Fuzzy Evaluation Tree，OFET）[31]来评估 SFS 的过滤效果。OFET 的根节点代表评估的最后结果，中间节点代表类别指标的评估结果，叶节点代表评估指标。OFET 树形成一个高度为 2 的评价树。图 3-1 为 SFS 的 OFET 的示意图，其中 O 为根节点；O_i 为 O 的下一层节点（i=1,2,3），表示根节点的子特性；O_{ij} 为 O_i 的下一层节点，表示各个子特性的评估指标差值。

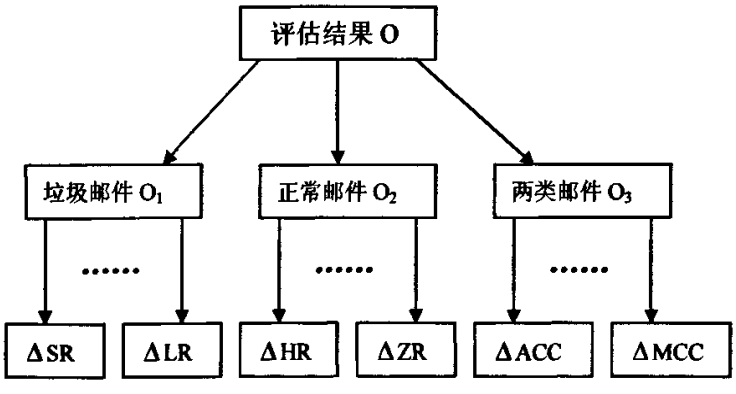

图 3-1 垃圾邮件过滤系统的 OFET 图

两个 SFS 进行相互间的比较时，过滤效果存在优劣等级，评估的目的是确定两两 SFS

之间的优劣等级。在本论文中，将它们分为 9 个等级，记做评估等级集合 $V=\{e_1,e_2,e_3,e_4,e_5,e_6,e_7,e_8,e_9\}$。其中 e_1 ="强烈重要"，e_2 ="明显重要"，e_3 ="一般重要"，e_4 ="稍微重要"，e_5 ="相差无几"，e_6 ="稍微逊色"，e_7 ="一般逊色"，e_8 ="显著更差"，e_9 ="相差甚远"。

评价等级集合 V 是统一的，适用于 OFET 的各个层次。OFET 上的所有节点都使用同一个 V 进行评估。但由于各评估指标表示的含义，取值的范围上存在差异，V 对应的分界点不完全一样。通过对评估指标的分析，论文将指标取值范围相同的评估指标采取相同的分界点。设 $(A_{ij})_{3\times9}$ 表示评估指标的分界点矩阵，当评估指标的取值范围为[0,1]时，分界点向量为 $(A_{11},A_{12},\ldots\ldots A_{19})$；当评估指标的取值范围为[-1,1],分界点向量为 $(A_{21},A_{22},\ldots\ldots A_{29})$；同理，若评估指标取值范围为[0,+∞)，则分界点的向量为 $(A_{31},A_{32},\ldots\ldots A_{39})$。各分界点向量与评估等级集合的关系如下：

取值范围为[0，1]的评估指标有：SR，SP，SF_1，HR，HP，HF_1，ACC。为能够提供较好的区分度和精确性，依据各指标取值的特点和多次实验的结果，论文定义它们的差值分界点与评估等级的关系见表 3-2。

表 3-2 取值范围为[0,1]的评估指标的分界点与评估等级的对应关系

评估等级	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8	e_9
分界点	A_{11} (0.2)	A_{12} (0.12)	A_{13} (0.08)	A_{14} (0.04)	A_{15} (0)	A_{16} (-0.04)	A_{17} (-0.08)	A_{18} (-0.12)	A_{19} (-0.2)

取值范围为[-1，1]的评估指标有：YI，MCC。同取值范围为[0，1]的评估指标，定义它们的差值分界点与评估等级的关系见表 3-3。

表 3-3 取值范围为[-1,1]的评估指标的分界点与评估等级的对应关系

评估等级	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8	e_9
分界点	A_{21} (0.2)	A_{22} (0.15)	A_{23} (0.1)	A_{24} (0.05)	A_{25} (0)	A_{26} (-0.05)	A_{27} (-0.1)	A_{28} (-0.15)	A_{29} (-0.2)

取值范围为[0,+∞)的评估指标有 LR，ZR，RMAC，TCR。由于取值范围大且各指标有不同的特点，本文只能给出较为粗略的分界点值，在后续的分析中尽量减小这些指标对最终评估结果的影响。它们的差值分界点与评估等级的关系见表 3-4。

表 3-4 取值范围为[0,+∞)的评估指标的分界点与评估等级的对应关系

评估等级	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8	e_9
分界点	A_{31} (500)	A_{32} (100)	A_{33} (50)	A_{34} (10)	A_{35} (0)	A_{36} (-10)	A_{37} (-50)	A_{38} (-100)	A_{39} (-500)

3.1.2 隶属度函数

OFET 叶节点上的指标 O_{ij} 在 $e_k \in V$ 上的隶属度函数记为 $f_{e_k}(O_{ij})$。对于非叶节点的隶属度计算下面的叙述中给出。本文采用线性分析法[29]来确定隶属度函数 $f_{e_k}(O_{ij})$。该方法的原理是在评估指标的取值范围内确定一系列的分界点，然后见实际指标值通过线性插值公式进行计算，求出该指标值的隶属度。

对于类型为 i 的指标，假设两 SFS 在该指标值上的差值为 x，其隶属度函数为 $f_i(x)$。该指标的变化属于各评估等级的隶属度为 $f_{i1}(x)$，……，$f_{i9}(x)$，分别计算如下：

$$f_{i1}(x) = \begin{cases} 1 & x \geq A_{i1} \\ \dfrac{x - A_{i2}}{A_{i1} - A_{i2}} & A_{i2} \leq x \leq A_{i1} \\ 0 & x \leq A_{i2} \end{cases} \qquad \text{公式 (3.1)}$$

$$f_{i2}(x) = \begin{cases} 0 & x \geq A_{i1} \\ \dfrac{A_{i1} - x}{A_{i1} - A_{i2}} & A_{i2} \leq x \leq A_{i1} \\ \dfrac{x - A_{i3}}{A_{i2} - A_{i3}} & A_{i3} \leq x \leq A_{i2} \\ 0 & x \leq A_{i3} \end{cases} \qquad \text{公式 (3.2)}$$

…… …… ……

…… …… ……

$$f_{i8}(x) = \begin{cases} 0 & x \geq A_{i7} \\ \dfrac{A_{i7} - x}{A_{i7} - A_{i8}} & A_{i8} \leq x \leq A_{i7} \\ \dfrac{x - A_{i9}}{A_{i8} - A_{i9}} & A_{i9} \leq x \leq A_{i8} \\ 0 & x \leq A_{i9} \end{cases} \qquad \text{公式 (3.3)}$$

$$f_{i9}(x) = \begin{cases} 0 & x \geq A_{i8} \\ \dfrac{A_{i8} - x}{A_{i8} - A_{i9}} & A_{i8} \leq x \leq A_{i9} \\ 1 & x \leq A_{i9} \end{cases} \qquad \text{公式 (3.4)}$$

其中，取值范围为[0,1]的评估指标的类型为 1（即 i=1），取值范围为[-1,1]的评估指标的类型为 2（即 i=2），取值范围为[0,+∞)的评估指标的类型为 3（即 i=3）。

3.1.3 权重系数向量

权重的确定对评估的结果有较大的影响，不同的权重可能会导致评估结果有很大的不同。本文采取常用的层次分析法（AHP）[29]来确定评估指标的权重系数。层次分析法的思想是自上而下，逐层计算各评估指标相对于最高层的权重。在任一层，首先通过 1～9 标度（表 3-5）建立各指标之间的判断矩阵，计算出该矩阵的最大特征值所对应的向量并通过一致性检验，即可得到单层的各指标权重。在此基础上，再与上层指标的权重进行综合，即可得到该层指标对最高目标层次的权重值。

层次分析法非常适合于本文的多层次模糊综合评估方法中确定权重系数向量。用 AHP 来计算评估指标权重过程如下：

首先考虑图 3-1 中间节点 O_i（i=1,2,3）的权重 $W(O_i)$的计算，设$\vec{W}(O)$={$W(O_1)$,$W(O_2)$, $W(O_3)$}，$W(O_1)+W(O_2)+W(O_3)=1$。依据表 3-5，采用德尔菲法（Delphi）建立上述三个指标间的判定矩阵 A，求得满足 $A\vec{W}(O)=\lambda_{max}\vec{W}(O)$ 的最大的特征根 λ_{max}。此时 λ_{max} 对应的特征向量 W 就为所求的权重向量。计算 $CI=(\lambda_{max}-r)/(r-1)$，r 为判定矩阵的阶数；判定一致性条件 CR=CI/RI<0.1，RI_k 可查表 3-6 得到。如果上述条件满足，则认为判断矩阵的一致性较好；若不满足，则需要重新调整判断矩阵。

类似上述步骤分别计算 $W(O_{ij})$的权重，须注意的是$\vec{W}(O_i)$={$W(O_{i1})$,……,$W(O_{ij})$}，$\sum_{i=1}^{k}W(O_{il})=W(O_i)$。在 $W(O_{ij})$求出以后，需要对总权重做一致性检验 CI_i，并分别计算 $CI=\sum_{i=1}^{3}CI_i * W(O_i)$、$RI=\sum_{i=1}^{3}RI_i * W(O_i)$。若满足 CR＝CI/RI<0.1 时，认为达到了满意的一致性；否则仍需调整判断矩阵。

表 3-5 判断矩阵的取值及含义

判断矩阵中 P_{ij} 取值	定义
1	指标 i 比指标 j 同等重要
3	指标 i 比指标 j 稍微重要
5	指标 i 比指标 j 明显重要
7	指标 i 比指标 j 强烈重要
9	指标 i 比指标 j 极端重要
2，4，6，8	上述情况的中间状态
倒数	$P_{ji}=1/P_{ij}$

表 3-6 RI_k 与阶数 K 的关系

阶数 K	2	3	4	5	6	7	8
RI_k	0.00	0.58	0.90	1.12	1.24	1.32	1.41

在本论文中，首先使用 AHP 方法确定影响评估结果 O 的三要素垃圾邮件 O_1，正常邮件 O_2，两类邮件 O_3 的权重。由于在“归一化”复合矩阵中已经均衡了垃圾邮件与正常邮件的重要性，为此论文认为 SFS 对垃圾邮件的过滤能力的评估指标和对正常邮件的评估指标同等重要。另外，剩余的指标同时从两个方面反映 SFS 的过滤能力，论文认为应减小它对最终评估结果的影响。依据表 3-5，建立判断矩阵 A

$$A=\begin{bmatrix} 1 & 1 & 2 \\ 1 & 1 & 2 \\ 1/2 & 1/2 & 1 \end{bmatrix}$$

求得该矩阵的最大特征根为 $\lambda_{max}=3.000$，满足一致性条件。该特征根对应的特征向量为{0.6667,0.6667,0.3333}。由此可计算出 $W(O_1)=W(O_2)=0.6667/(0.6667+0.6667+0.3333)=0.400$，$W(O_3)=0.200$。即 W(O)={0.400, 0.400, 0.200}。

同理，对于 O_1 下的四个评估指标 SR、SP、SF_1 和 LR，SR 和 SP 是基本指标，具有最大权重；SF_1 反映上述两个指标的平衡能力，重要性仅次于 SR 和 SP；而 LR 是由 SP 与 SP 的相反数之比计算得到的，与 SP 的关联性很大且取值范围为[0,+∞)，论文认为该指标的重要性最小。依据上述判断，构造判断矩阵 A_1

$$A_1=\begin{bmatrix} 1 & 1 & 2 & 4 \\ 1 & 1 & 2 & 4 \\ 1/2 & 1/2 & 1 & 2 \\ 1/4 & 1/4 & 1/2 & 1 \end{bmatrix}$$

A_1 的最大特征根为 $\lambda_{max}=4.000$，满足一致性条件，λ_{max} 对应的特征向量 $\vec{W}$={0.364, 0.364,0.182,0.091}，四个评估指标对 O 的权重为 $\vec{W}(\vec{O_1})=W(O_1)\times\vec{W}$ ={0.1456,0.1456,0.0728,0.0364}。对于 O_2 下的四个评估指标 HR、HP、HF_1 和 ZR，论文认为它们之间的权重与 O_1 下的四个评估指标之间权重相同，即也有 $\vec{W}(\vec{O_2})$={0.1456,0.1456,0.0728,0.0364}。

对于 O_3 下的五个评估指标 ACC、YI、RAMC、TCR 和 MCC，由于 RAMC 和 TCR 取值范围为[0,+∞),它们的分界点确定的较为粗略，论文认为其重要性低于 ACC、YI 和 MCC。构造判断矩阵 A_2 如下。

$$A2=\begin{bmatrix} 1 & 1 & 2 & 2 & 1 \\ 1 & 1 & 2 & 2 & 1 \\ 1/2 & 1/2 & 1 & 1 & 1/2 \\ 1/2 & 1/2 & 1 & 1 & 1/2 \\ 1 & 1 & 2 & 2 & 1 \end{bmatrix}$$

A_2 的最大特征根为 $\lambda_{max}=5.000$，满足一致性条件，对应的特征向量

$\vec{W}$={0.25,0.25,0.125,0.125,0.25}。五个评估指标对 O 的权重$\vec{W}(\vec{O_3})$=W(O_3)×$\vec{W}$={0.05, 0.05,0.025,0.025,0.05}。

3.1.4 模糊矩阵合成

首先对于 O_i 类别下的评估指标 O_{ij}，在评估等级集合V上的隶属度构成了该类指标的隶属度矩阵 R(O_i)。

$$R(O_i)=\begin{bmatrix} f_{L_{i1}1}(O_{i1}) & f_{L_{i1}2}(O_{i1}) & \cdots & f_{L_{i1}9}(O_{i1}) \\ f_{L_{i2}1}(O_{i2}) & f_{L_{i2}2}(O_{i2}) & \cdots & f_{L_{i2}9}(O_{i2}) \\ \cdots & \cdots & \cdots & \cdots \\ f_{L_{ij}1}(O_{ij}) & f_{L_{ij}2}(O_{ij}) & \cdots & f_{L_{ij}9}(O_{ij}) \end{bmatrix} \qquad \text{公式（3.5）}$$

其中 L_{ij} 表示指标 O_{ij} 取值范围对应 3.1.2 节定义的类型。依据上述 O_i 类别下的各评估指标的隶属度矩阵，O_i 在评估等级集合V上的评估结果为

$$B(O_i)=\vec{W}(\vec{O_i})\circ R(O_i) \qquad \text{公式（3.6）}$$

公式（3.6）中 o 算子可以自定义，本文采取乘法运算。据式（3.6）的结果，O 节点下的子类别 O_i 的隶属度矩阵 $R(O)$ 如下所示：

$$R(O)=\begin{bmatrix} B(O_1) \\ B(O_2) \\ B(O_3) \end{bmatrix} \qquad \text{公式（3.7）}$$

再根据公式计算出最后目标所属的评估等级的模糊关系 $B(O)=\vec{W}(\vec{O})\circ R(O)$，依据最大隶属度原则，选择 $B(O)$ 向量中最大的值来判断评估对象所属的类别。

采用模糊评估方法来评估 N 个 SFS 的过滤效果的流程为：

1）N 个 SFS 对相同的标准邮件集进行训练评测后，计算 3.1.1 所需 的各个评估指标的值；
2）选取未比较的两个 SFS 进行两两比较，计算它们各评估指标的差值；
3）根据各评估指标差值和3.1.2 定义的隶属度函数，生成各个层次的隶属度矩阵；
4）根据权重向量和各层次隶属度矩阵，计算出模糊综合评估结果；并依据最大隶属度原则，确定两 SFS 两两比较后的结果；
5）这两个 SFS 的评估结果按表 3-7 的对应关系得到定量分值，并将该分值依据下标记录到 N×N 的比较矩阵中；跳步骤 2）循环上述过程，直至所有 SFS 间都进行过两两比较。
6）将 N×N 矩阵按 $a_{ij}=1/a_{ji}$ 补充完整，该矩阵为一致性判定矩阵，计算该矩阵最大特征根对应的特征向量，依据特征向量各分值的大小对 SFS 进行排序。

表 3-7：评估等级与定量分值的对应关系

评估等级	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8	e_9
定量分值	9	7	5	3	1	1/3	1/5	1/7	1/9

3.2 因子分析法

因子分析法是用少数几个因子来描述许多指标或因素之间的联系，以较少的几个因子反映原有资料中的大部分信息的统计学方法。其特点是：

1）因子的数量远小于原有的指标数量。

2）因子变量不是对原有指标的取舍，而是对原有指标信息的重组，能够反映原有指标的大部分信息。

3）因子变量之间不存在相关性。

4）因子变量具有解释性，是对某一类信息的综合和反映。

文献[34]使用因子分析法评估了成都市及相关竞争城市在旅游、人居、资本聚集、市场等城市品牌方面的先后次序。文献[35]介绍了因子分析法的原理和使用流程。文献[36]介绍了因子分析法的定义和数学模型，并介绍了如何使用 SPSS 统计软件来实现因子分析法。

从本文第二章可以知道，SFS 的合成指标都是通过四个基本指标的四则运算而得到的，它们之间存在着较大的相关性。为此，论文认为可以使用因子分析法来评估 SFS，将 16 个指标降维成较少的互相独立的几个综合指标来评估 SFS。使用因子分析法综合评估 SFS 的数学模型描述如公式（3.8）所述。

$$X = AF + a\varepsilon \qquad \text{公式（3.8）}$$

其中 X 为 16 维的评估指标向量。A 为 16×m 维的因子载荷矩阵，其中 a_{ij} 为因子载荷，表示第 i 个评估指标在第 j 个因子变量上的负荷。F 为 m 维的因子变量，表示 m 个互相独立的综合评估指标，它们由 16 个评估指标线性组合而成。ε 为特殊因子，表示评估指标不能为 m 个因子所表示的部分。m 个因子互相独立，反映 SFS 过滤能力的 m 个方向。

在保证数据信息丢失最少的原则下，将 16 个指标组成的高维评估向量降低为由较少的几个互相独立的因子变量组成的低维评估向量，并使用独立的因子综合评估各 SFS。因子分析法的两个主要问题是构造因子变量和解释因子变量。使用因子分析法评估 SFS 的步骤为：

1）构造因子变量和因子得分

2）利用旋转使得因子变量更具有解释性

3）综合评估 SFS

3.2.1 构造因子变量和因子得分

论文采用主成分分析法[35]来确定因子变量 F。主成分分析的步骤为如下：

1）首先对评估指标值进行标准化处理。标准化处理的过程见公式（3.9）

$$x_{ij}^{*} = \frac{x_{ij} - x_j}{S_j} \qquad \text{公式（3.9）}$$

其中 i=1,2,……，n；n 表示 n 个 SFS。j=1, 2... 16，表示 16 个评估指标。

2）计算数据$[x_{ij}^{*}]_{n\times16}$的协方差矩阵 R。

3）求得 R 的前 m 个特征值：$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$，以及对应的特征向量$\mu_1, \mu_2, \cdots, \mu_m$。根据累计方差贡献率大于 80%的原则，确定 m 的值。即满足公式（3.10）：

$$Q = \frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{16} \lambda_i} \geq 80\% \qquad \text{公式（3.10）}$$

或者通过特征值必须大于 1 的方式确定 m 的值。

4）求得 m 个变量的因子载荷矩阵。计算方法如下：

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ a_{161} & a_{162} & \cdots & a_{16m} \end{bmatrix} = \begin{bmatrix} \mu_{11}\sqrt{\lambda_1} & \mu_{12}\sqrt{\lambda_2} & \cdots & \mu_{1m}\sqrt{\lambda_m} \\ \mu_{21}\sqrt{\lambda_1} & \mu_{22}\sqrt{\lambda_2} & \cdots & \mu_{2m}\sqrt{\lambda_m} \\ \cdots & \cdots & \cdots & \cdots \\ \mu_{161}\sqrt{\lambda_1} & \mu_{162}\sqrt{\lambda_2} & \cdots & \mu_{16m}\sqrt{\lambda_m} \end{bmatrix}$$

5）依据上述因子载荷矩阵 A 和公式 3.8，利用回归法求得因子得分。因子得分将因子变量表示为原 16 个评估指标的线性组合，即：

$$F_j = \beta_{j1}x_1 + \beta_{j2}x_2 + \cdots + \beta_{j16}x_{16} \quad (j=1,2,\ldots\ldots,m) \qquad \text{公式（3.11）}$$

3.2.2 因子变量的解释

载荷矩阵 A 中的某个 a_{ij} 较大，说明评估指标 i 与因子变量 j 有较大的相关关系。载荷矩阵 A 中的某一行可能有多个 a_{ij} 比较大，说明评估评估指标 i 与多个因子变量都有比较大的关联。载荷矩阵 A 中的某一列中也可能存在多个 a_{ij} 比较大，这说明因子变量 j 可能解释多个评估指标的信息。但由于它可能只解释某个评估指标的少部分信息，不能够代表任何一个评估指标，这就使得该因子变量的含义非常模糊。

为此，在实际分析中，为对实际的因子变量的含义获得比较清晰的含义，需要对因子变量做旋转，使得因子载荷矩阵中的元素尽可能的接近 0 或±1，这样得到的因子变量具有比较清晰的实际意义。论文采用极大方差法来随因子变量进行旋转，相关的统计软件中都集成了该方法的实现，其原理请感兴趣的读者参考其他文献。

3.2.3 综合评估 SFS

因子变量反映了 SFS 过滤能力的某一个方面，依据单个因子的得分可以对 SFS 在某一方面的过滤能力进行排名。除此之外，由于各因子变量相互独立，可以使用多个独立因子得分的加权平均和来综合评估各 SFS。综合评估值的计算方法如下：

$$E = w_1 * F_1 + w_2 * F_2 + \cdots + w_m * F_m \qquad \text{公式（3.12）}$$

F_m 表示因子旋转后的得分；$w_i = \lambda_i / \sum_{k=1}^{16} \lambda_k$，称为方差贡献率。

3.3 ROCCH 综合评估方法

本文第二章研究的是静态评估指标，是在确定的平均误报代价之比和给定的标准邮件集下进行评估指标计算。然而现实中邮件服务器收到的垃圾邮件与正常邮件的比例是不断变化的，且不同的服务器收到的垃圾邮件与正常邮件的比例也是不同的[37]。除此之外，在现实中，不同的服务器，不同的时间段，收到的垃圾邮件的误报代价与正常邮件的误报代价之比也是不同的。因此，要客观，公正的评估 SFS，也需要考虑在不同的邮件比例和误报代价比例下进行评估。

ROC 曲线是评估 SFS 的重要手段，目前使用 ROC 曲线评估 SFS 的过滤准确能力的指标主要是 ROC 曲线下的面积和最优分割点，但它反映的是系统的整体过滤效果和最优的过滤效果，不能评估变化邮件比例以及误报代价比例条件下的 SFS 的过滤效果。为此，本节将介绍基于 ROC 曲线的 ROCCH 方法，本并使用它来评估邮件比例和误报代价比例变化情况下的 SFS 的过滤效果，并针对不同的邮件比例和误报代价比例判断出最优的 SFS。

3.3.1 ISO-性能线

假设新到来的邮件是垃圾邮件的概率为 $p(Y)$，是正常邮件的概率是 $p(N)=1-p(Y)$；垃圾邮件误报为正常邮件的代价为 $C(Y,N)$，正常邮件误报为垃圾邮件的代价为 $C(N,Y)$。对于给定的 ROC 空间上的一点 (HM,SR)，定义该"离散"过滤系统的预期误报代价为 $R = p(Y)*(1-SR)*C(Y,N) + p(N)*HM*C(N,Y)$。为使其代价最小，对 R 求导并令导数等于 0。得到：

$$\frac{d_{SR}}{d_{HM}} = \frac{C(N,Y)\cdot p(N)}{C(Y,N)\cdot p(Y)} = m \qquad \text{公式（3.13）}$$

公式（3.11）得到的是 ROC 曲线上某点的斜率，称斜率为 m 的直线为 ISO-性能线。

3.3.2 ROCCH 方法

由于现实中的邮件比例和误报代价比例在不断变化，需要标识出潜在的最优的 SFS 的集合。对于给定的邮件比例和误报代价比例，依据公式（3.13）可以计算出 ISO-性能曲线的斜率，在 ROC 空间中绘制 ISO-性能曲线族。这种条件下最优的 SFS 是与 ISO-性能曲线相切的切点最靠近左上角的那条 ROC 曲线对应的 SFS。

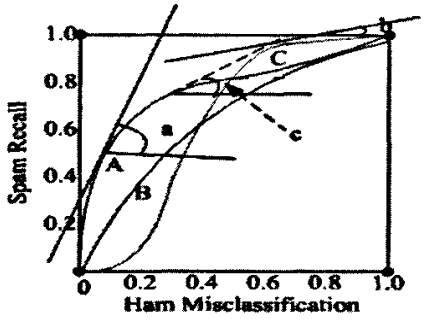

图3-1 标识最优的垃圾邮件过滤系统

如 3-1 图所示，有 A，B，C 三条不同 SFS 的 ROC 曲线，分别记作系统 A，B，C。从图中所示 ROC 曲线可以看出，当 ISO-性能线的斜率 m=a 时，系统 A 是最优；同样，ISO-性能线的斜率 m=b 时，系统 C 最优。系统 A 和系统 C 是潜在的最优的 SFS。斜率 c 是临界点，当 $0 \le m \le c$ 时，系统 C 为最优的 SFS；当 $c \le m \le +\infty$ 时，系统 A 为最优的 SFS。当判定多个 SFS 的优劣顺序也可以采取同样的方法。

3.4 三种方法的比较

模糊综合评估方法以模糊的、多层次的、加权的方式对多个评估指标进行模糊综合，依据最大隶属度原则，判定两 SFS 系统的优劣等级。依此方法建立评估矩阵，求出最后所有 SFS 的优劣等级次序。因子分析法是从各评估指标的相关性出发，从十六个原始评估指标中提取出较少的几个相互独立的因子变量，并依据这几个因子变量独立或综合的评估 SFS。基于 ROCCH 的综合评估方法以 SFS 系统的误报代价最小为目标，依据各 SFS 系统的 ROC 曲线，求出动态条件下潜在的最优的 SFS 系统，以此判定所有 SFS 系统的优劣顺序。

由于三种方法的出发点不完全一样，所以它们之间存在一些差异。模糊综合评估方法和因子分析法都是依据各 SFS 当前给出的指标值，判定各 SFS 当前过滤效果的优劣顺序。而 ROCCH 综合评估方法考虑不同阀值下 SFS 系统的指标值，判定的是动态条件下各 SFS 潜在的过滤能力的优劣顺序。除此之外，由于前两种方法仅考虑当前的评测结果，它们适合于任何的 SFS 系统的评估；而 ROCCH 综合评估方法则是依据 SFS 系统的 ROC 曲线，它仅适合于阀值分类的 SFS 的判定。

另外，模糊综合评估方法需要人工设置各评估等级分界点对应的指标差值和构建各指标间的判定矩阵，而人工设置难免存在误差，而且存在粒度大小的问题，未必适合所有的 SFS 的评估。因子分析法和基于 ROCCH 的综合评估方法不需要任何的用户参数，可以自动的从数据中获得评估的信息，更具有合理性。

3.5 本章小结

本章首先叙述了建立综合评估方法的重要性和必要性。随后叙述模糊综合评估方法、因子分析法和 ROCCH 综合评估方法的原理、特点和在相关研究领域中应用，并给出使用它们对 SFS 进行综合评估的方法和流程。最后，论文对对三种方法的特点和使用范围进行了比较，并给出本章小结。

第四章 影响 SFS 评估的其他重要因素

SFS 是一个复杂的系统，影响它的评估结果的因素有很多。除前两章讨论的 SFS 的评估指标和综合的评估方法之外，还有对 SFS 的评估结果有着重要影响的其他因素。本章将讨论评测训练方法和标准邮件集对 SFS 评估的影响。

4.1 评测训练方法对评估结果的影响

目前被使用的评测训练方法有十字交叉验证法、逐一评测训练法和先训练再评测法。其基本的实现原理如第一章所述。

三种方法各有特点。逐一评测训练法开始的知识比较少，评判的结果误报较多，但随着训练数据的增加，知识越来越丰富，准确性也越来越高。而且由于它不间断的训练，系统能根据邮件内容的变化而丰富自己的知识。采用先训练再评测法，SFS 在开始阶段经过训练获得较多知识，评测的准确性较高，但 SFS 的知识不会再增加，以后的评测都基于开始阶段训练得到的知识。而且不能随着系统的知识不能随邮件内容的变化而变化。十字交叉法本质上是一种特殊的先训练再评测法，它的优点是训练足够充分，准确性较高；缺点是所需时间太长，实现较复杂。

考虑到现实中的邮件服务器，需要尽可能多的获得垃圾邮件的特点，以便取得更好的过滤效果。通常它们都能对用户反馈的垃圾邮件进行自学习。在本文中，结合现实中邮件服务器的使用特点，提出一种新的评测训练方法—反馈训练评测法。该方法吸取了逐一评测训练法和先训练再评测法两者的优点，其基本原理与先训练再评测法非常类似，唯一区别在于当 SFS 对评测邮件集进行评测时，如果被评测邮件的标准答案为垃圾邮件时，该邮件将被 SFS 训练以获得更多的垃圾邮件的特征。这种评测训练法需要用户反馈它们收到的所有垃圾邮件，是理想状态下的过滤效果。

不同的评测训练方法，由于在训练和评测的邮件数量、顺序上存在差异，从而对评估的结果产生一定的影响。为比较四种评测训练方法对评估结果的影响，本文给出 Bogofilter 系统在三种不同的评测训练方法下，对 Spamassassin_Corpus 标准邮件集的评测结果（见表 4.1、表 4.2）。其中在先评测再训练法和反馈训练评测法中，训练邮件集的数量取为 2000 封，占总邮件数量的 33%左右；λ 取为 9。

表 4.1 采用四种评测训练法，Bogofilter 对 Spamassassin_Corpus 的评估结果（一）

指标 / 评测训练法	SR	SP	HR	HP	ACC	ERR	SM	HM
十字交叉法	0.7129	0.9933	0.9998	0.9857	0.9860	0.014	0.2871	0.0002
逐一评测训练法	0.7597	0.9521	0.9981	0.988	0.9866	0.0134	0.2403	0.0019
先训练再评测法	0.5788	1.00	1.00	0.9909	0.9910	0.009	0.4212	0.00
反馈训练评测法	0.7712	0.904	0.9982	0.9950	0.9934	0.0066	0.2288	0.0018

表 4.2 采用四种评测训练法，Bogofilter 对 Spamassassin_Corpus 的评估结果（二）

指标 / 评测训练法	SF_1	HF_1	LR	ZR	YI	AMC	TCR	MCC
十字交叉法	0.830	0.9927	149	69.133	0.7127	0.091	3.426	0.8353
逐一评测训练法	0.8451	0.9930	179	9.1413	0.7578	0.0870	3.5904	0.8239
先训练再评测法	0.7332	0.9954	10000	109.23	0.5788	0.0689	2.3741	0.7573
反馈训练评测法	0.8323	0.9966	9.4259	200.74	0.7694	0.0508	3.2195	0.8318

从表 4.1、表 4.2 可以看出，在训练了 1/3 的邮件后，采用先训练再评测法仅能过滤掉不到 60%的垃圾邮件；而其他三种方法都能过滤掉 70%以上的垃圾邮件。按照 3.1 小节介绍的模糊综合评估方法，在 SR 指标上，先训练再评测法与其他三种方法相差 3 个等级以上（$\Delta SR > 0.12$）。而在 SP 指标上，采用不同评测训练方法得到的指标值的差异同样很大。十字交叉法和先训练再评测法获得的 SP 值接近于 1，比逐一评测训练法和反馈训练评测法分别高 1、2 个等级。在其他的评估指标上，四种不同的评测训练方法给出的评估结果同样也存在较大的差异。

由此可见，评测训练法对 SFS 的评估结果有较大的影响。为客观、中立的评估 SFS，在对不同的 SFS 评估时，就必须采用相同的评测训练方法。依据前面的分析，逐一评测训练法需要用户反馈所有的正常邮件和垃圾邮件，在现实中比较难实现。十字交叉法所需的时间太长，而且训练邮件所占比例固定。在本文中，将采用先训练再评测法或反馈训练评测法来评估 SFS。

4.2 标准邮件集对评估结果的影响

标准邮件集模拟着现实中邮件服务器到来的邮件，使用不同的标准邮件集评估 SFS 将会有不同的结果，标准邮件集也将对评估的结果有着重要的影响。本节将从标准邮件集的总邮件数量、垃圾邮件所占比例比例、训练邮件所占比例、误报代价比到邮件的到达顺序等方面讨论标准邮件集对 SFS 评估的影响。为保证评测训练法不对 SFS 评估的产生影响，本节所有的实验均采用先训练再评测法。

4.2.1 邮件总数量对评估结果的影响

标准邮件集的总邮件数量是影响 SFS 评估的重要因素。用于评估的总邮件数量太少，则不能充分反映邮件服务器的各方面的过滤能力，且偶然性也大。而总邮件数量太多，又导致 SFS 训练和评测花费的时间太长；而且当邮件数量达到一定程度后，对评估结果的影响将减少到很小。

为反映邮件数量对评估结果可能造成的影响，本文使用 Bogofilter 系统对 CCERT 提供的邮件集进行训练和评测。从 CCERT 提供的邮件集中分别选出 1000、3000、5000、8000、10000、12000、15000、18000、20000 封邮件作为标准邮件集。所有的标准邮件集中垃圾邮件所占比例均为 6/10，训练集所占比例为 2/10，误报代价比 $\lambda = 9$。对上述标准邮件集进行评测后，部分评估指标随邮件数量变化的曲线如图 4-1 所示：

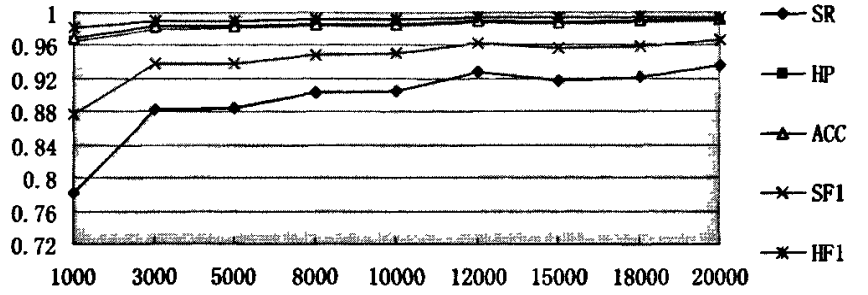

图4-1 Bogofilter下，部分评估指标随总邮件数量变化曲线图

从图 4.1 可以看出，SR、HP、ACC、SF_1、HF_1五个指标随总邮件数量的增长而增大。当总邮件数量较小时（<12000），上述五个评估指标随总邮件数量的增长速度较快。当总邮件数量的足够大时，各评估指标值变化较小，趋于平稳。从单个指标来看，SR 和 SF_1 两指标的增长幅度较大，SR 从 0.78 增长到 0.93，增长幅度超过 19%；SF_1的增长幅度也达到 10%以上。剩余的评估指标增长均不超过 3%。另外，由于 Bogofilter 过滤系统具有对正常邮件超强过滤能力的特点，HR、SP 两评估指标的值在上述评测中恒为 1，其他的评估指标与上述指标有相同的变化趋势。

从上述分析可知，SFS 的过滤能力随总邮件数量的增大而增强。这是因为邮件数量较少时，SFS 获得的知识较少，容易导致误报；随着邮件数量的增加，SFS 获得的知识也在增加，其过滤准确度也逐渐提高了；而 SFS 获得了足够的知识，邮件数量对过滤的结果的影响就变得很小了。依前面所述，鉴于邮件数量对评估有较大的影响，为客观、中立的评估各 SFS，标准邮件集的邮件数量应该足够大。

4.2.2 训练邮件集所占比例对评估结果的影响

训练邮件集所占比例对评估结果的影响是显然的。一个经过充分训练的 SFS 的过滤能力肯定比经过较少训练的 SFS 的过滤能力要强些。为全面的反映单个 SFS 的过滤能力，就必须要考虑在不同的训练集比例下 SFS 的过滤能力。同样，在评估多个 SFS 时，也必须让它们在相同的比例下进行训练和评测。

为反映 SFS 随训练集和评测集邮件数量比例的变化情况，本文使用 CRM114 系统对 CCERT 提供的邮件集进行训练和评测。从 CCERT 提供的邮件集中选择 10000 封邮件作为标准邮件集，其垃圾邮件数量所占的比例为 6/10。将训练集邮件的数量所占的比例依次取为 0.1、0.2、……、0.9，误报代价比取为 $\lambda=9$。部分评估指标随训练邮件集所占比例的变化情况如图 4-2 所示。

从图 4-2 可以看出，当训练集所占比例较小时（<3/10），CRM114 系统各评估指标的值随训练集所占比例的增加而迅速增加；增长最大的 SP 指标从 0.864 增长到 0.987，增长幅度超过 12%；增长最小的 HF_1 幅度在 1.2%左右。这是由于训练集所占比例越大，SFS 从训练集中获得知识就越多，其过滤的准确性也就越高。当训练集所占的比例足够大时（>3/10），各指标值都趋于平稳，SFS 的过滤准确度也基本保持不变。这说明 SFS 获得了足够的知识，训练集所占比例对 SFS 的过滤结果影响不大。从上图中还可以看到，当训练集所占的比例非常大时（>9/10），SP 和 SF_1 的指标值出现了小幅的降低，本文认为这是由于评测邮件集所占比例太小而导致的。在上述实验中，SR 和 HP 指标的值接近 1，变化非常小；而其他的一些评估指标具有与上述指标相同的变化趋势。

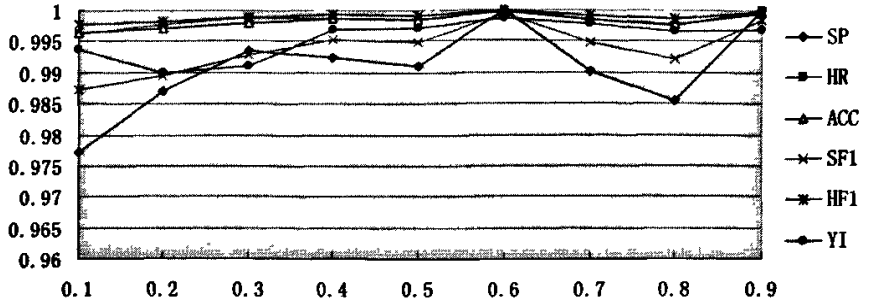

图4-2 CRM114系统下，部分指标随训练邮件集所占比例的变化图

鉴于训练集所占的比例对评估结果的影响较大，为全面评估单个 SFS，需要在不同的训练集比例下的进行评估；而为公平的评估多个 SFS，则需要让它们在相同的训练集比例下进行评估。

4.2.3 垃圾邮件数量所占比例对评估结果的影响

对于给定数量的标准邮件集，垃圾邮件数量所占比例也对SFS的评估有着重要的影响。比如：假设两个有 10000 封邮件的标准邮件集，第一个包含 1 封垃圾邮件和 9999 封正常邮件；另一个包含 9999 封垃圾邮件和 1 封正常邮件。对于同一个 SFS，使用这两个标准邮件集进行评测就非常不公平。对于第一个标准邮件集，SFS 不做任何过滤，将所有邮件都判定为正常邮件就能够得到 99.99%的正确率，而对于第二个标准邮件集，SFS 很难获得这么高的正确率。垃圾邮件所占比例的巨大差异导致这两个评估结果的不可比性。

本节主要描述由于垃圾邮件所占比例的变化，对评估结果的影响。为保证评估的结果尽量不受其他因素的影响，本文使用 Bogofilter 系统对从 CCERT 邮件集中选择 10000 封邮件组成标准邮件集进行训练和评测。垃圾邮件所占的比例依次取 1/10、2/10、3/10、4/10、5/10、6/10、7/10、8/10、9/10，训练集所占的比例固定为 2/10，误报代价比为 $\lambda=9$。四个基本指标的随两类邮件比例变化情况见图 4-3。

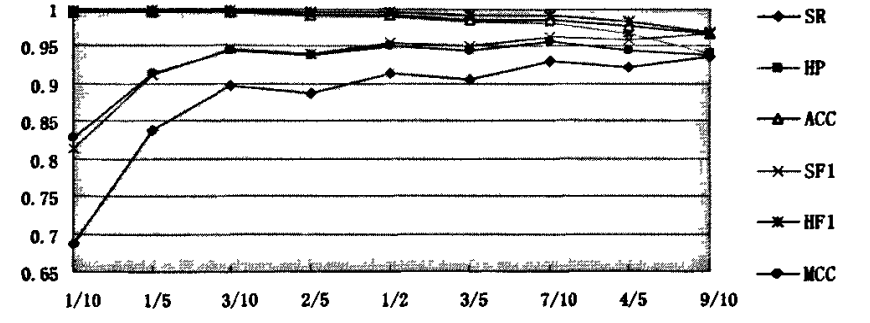

图4-3 bogofilter系统下，部分指标随垃圾邮件所占比例的变化图

从图 4.3 可以看出，SR、SF1 和 MCC 三个指标随垃圾邮件所占比例的增加而增大，平均增长幅度超过 22.7%；而 HP、HF_1 和 ACC 指标随垃圾邮件所占比例的增加而减小，平均下降幅度超过 3.7%。除此之外，当垃圾邮件所占比例大于 3/10 时，SR 和 SF_1 的指标

值增长变缓，趋近于固定值。而 HP、HF_1 和 ACC 的值在垃圾邮件所占比例较小时，变化很小；直到它大于 7/10 时，三个指标的值减小的速度变快。

上述分析可见标准邮件集中，垃圾邮件所占的比例对评估的结果有非常大的影响。文献[1]叙述了当前中国网民收到的垃圾邮件占所有邮件的比例大约为 6/10，为此，在本文中使用的标准邮件集中的垃圾邮件所占的比例应尽可能与这个数值相当。

4.2.4 误报代价比对评估结果的影响

误报代价比衡量的是正常邮件比垃圾邮件重要的程度。对于不同的邮件用户来说，误报代价比是不一样的。误报代价比的大小直接影响到了一些指标值的计算结果，从而对评估的结果产生影响。文献[17]讨论了误报代价比 $\lambda=1$、9、999 的情况下，SP 随 SR 变化的情况。

在本文中，为描述这种影响，本节对 Bogofilter 系统对 Spamassassin_corpus 标准邮件集进行训练和评测的结果采用 $\lambda=1$、9、999 计算各指标的值。评估的结果见表 4-3。

表 4-3 不同误报代价比下，Bogofilter 系统对 Spamassassin_Corpus 评估结果（一）

指标 / λ 的值	SR	SP	HR	HP	ACC	ERR	SM	HM
$\lambda=1$	0.5788	1.0	1.0	0.9239	0.9311	0.0689	0.4212	0.0
$\lambda=9$	0.5788	1.0	1.0	0.9909	0.9910	0.0089	0.4212	0.0
$\lambda=999$	0.5788	1.0	1.0	0.9999	0.9999	0.0001	0.4212	0.0

表 4-3 不同误报代价比下，Bogofilter 系统对 Spamassassin_Corpus 评估结果（二）

指标 / λ 的值	SF_1	HF_1	LR	ZR	YI	AMC	TCR	MCC
$\lambda=1$	0.7332	0.9604	10000	12.137	0.5788	0.0689	2.3741	0.7313
$\lambda=9$	0.7332	0.9954	10000	12.137	0.5788	0.0689	2.3741	0.7573
$\lambda=999$	0.733	0.9999	10000	12.137	0.5788	0.0689	2.3741	0.7608

从表 4-3 可以看出，误报代价比对部分评估指标的值有非常大的影响。$\lambda=1$ 时的 HP 值比 $\lambda=999$ 时少了 0.05 以上，而 MCC 也少了 0.03 左右。而对于 SR、HR、SM、HM 等指标值没有任何变化。这是由误报代价比的含义所导致的，它定义了正常邮件相当于垃圾邮件的数量。误报代价比是一个很难定量的值，不同的用户、不同的邮件都有不同的误报代价比。在 SFS 评估中，为保证公平性，需要采用相同的，符合大多数现实情况的误报代价比。本文将使用 $\lambda=9$ 作为系统默认的值。

4.2.5 邮件的顺序对评估结果的影响

邮件的顺序是指标准邮件集中，邮件用于训练和评测的先后顺序。如果具有某种特征的邮件被训练过了，那么由于SFS已经获得该特征的相关知识，将会影响到后面具有相似特征的邮件的判定。举个极端的例子，对于相同的标准邮件集，将所有的正常邮件和垃圾邮件独立开来与将它们相互交错对评估结果的影响应该不一样。正是基于这样的判断，本文认为，邮件的顺序对评估的结果是有一定影响的。

为从实验中证实这种影响，本文使用Bogofilter系统对从CCERT中选择的10000封邮件的标准邮件集进行训练和评测。该集合先后采取了所有的垃圾邮件在前面、垃圾邮件与正常邮件按照它们的数量比例随机的交错在一起、所有的正常邮件在前面三种不同的邮件顺序排列方式。为保证尽量不受其他因素的影响，实验采用$\lambda=9$，训练集所占比例为2/10，垃圾邮件所占比例为6/10。实验数据结果如表4-4所示。

表4-4 不同邮件顺序下，Bogofilter系统对CCERT标准邮件集的评估结果（一）

指标 / 邮件的顺序	SR	SP	HR	HP	ACC	ERR	SM	HM
垃圾邮件在前面	0	0	1	0.9	0.9	0.1	0	0
两类邮件交错	0.7712	0.9041	0.998	0.9950	0.9934	0.0066	0.2288	0.0018
正常邮件在前面	0.9112	1	1	0.9712	0.9778	0.0222	0.0888	0

表4-4 不同邮件顺序下，Bogofilter系统对CCERT标准邮件集的评估结果（二）

指标 / 邮件的顺序	SF_1	HF_1	LR	ZR	YI	AMC	TCR	MCC
垃圾邮件在前面	0.0	0.9474	0	9	0	0.5	1	1
两类邮件交错	0.8324	0.9966	9.4259	200.74	0.7694	0.0508	3.2195	0.8318
正常邮件在前面	0.9535	0.9854	0.5994	10000	3.7523	0.9112	11.257	0.8850

从表4-4可以看出，在不同的三种顺序下，评估指标值有非常大的差别，邮件的顺序对SFS的评估结果也有着重要的影响。但由于垃圾邮件和正常邮件的到达顺序还没有人研究过，并且不同的邮件服务器收到的邮件的顺序的分布也可能不一样。在本文中，对于给出的邮件顺序的标准邮件集，按照给定邮件顺序对SFS进行评估；对于没有给定顺序的标准邮件集，将从标准邮件集中按照一定概率随机选取。

4.3 本章小结

本章首先分析了评测训练方法对 SFS 评估的影响，并根据现实邮件服务器的情况，提出一种新的评测训练法，然后给出 Bogofilter 系统采用四种不同的评测训练方法对 Spamassassin_Corpus 标准集的评测结果。随后，本章以实验的方式从邮件的总数量、垃圾邮件占的比例、训练邮件占的比例、误报代价比到邮件的到达顺序等方面讨论了标准邮件集对评估结果的影响，并给出本文选择用于评估 SFS 的各参数的值。

第五章 SFS 评估系统的设计与实现

由第四章可以看出，标准邮件集是影响 SFS 评估的重要因素，目前公开的标准邮件集都是由研究机构收集的静态标准邮件集。一方面，为防止有些 SFS 刻意适应这些静态的标准邮件集，另一方面由于静态的标准邮件集自身存在的不足；本章首先设计和实现了一个模拟标准邮件集的生成系统，该系统可以接收用户的配置动态生成模拟标准邮件集。随后本章介绍了 SFS 评估系统的总体结构，以及它的各子系统的实现机制和功能结构。

5.1 模拟标准邮件集生成系统的研究与实现

当前已经有一些研究机构提供了公用的标准邮件集，这些邮件集被其他的研究者所使用，具有较好的使用价值，它们具有以下优点：

1）这些标准邮件集都是从邮件服务器上收集的真实邮件。

2）各标准邮件集中邮件通常按时间顺序排列，给定了邮件的到达顺序。

但作为标准邮件集，它们也存在如下不足：

1）总邮件数量参次不齐，很多标准邮件的邮件数量太少。

2）垃圾邮件的比例变化较大，有的太大而有的又太小。

3）从单个邮件服务器收集的邮件，邮件的内容单一。

基于上述原因，为尽可能客观的、公正的评估 SFS，本文设计和实现了模拟标准邮件集的生成系统，它可以从邮件库中按照用户的配置动态生成一个模拟的标准邮件集。这样一方面它可以防止 SFS 刻意适应静态的标准邮件集；另一方面可以让用户根据不同邮件服务器的实际情况，动态生成合理的标准邮件集。

模拟标准邮件集生成系统包括用户参数配置，配置文件解析，垃圾邮件和正常邮件集以及邮件集生成等模块，其结构如图 5.1 所示：

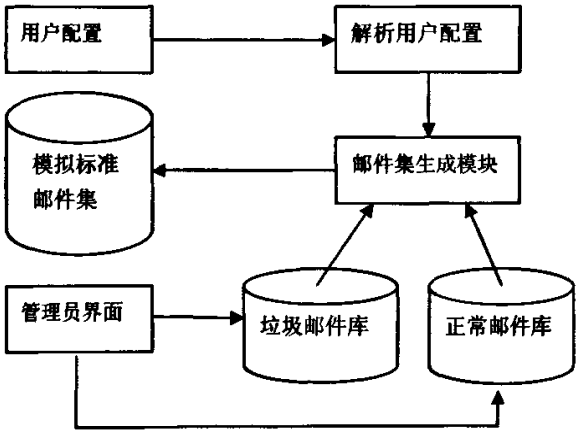

图 5-1 模拟标准邮件集生成系统结构图

模拟标准邮件集生成系统的实现流程为：

1）用户配置系统所需的各个参数，系统解析后传给邮件集生成模块。

2）邮件集生成模块依照一定的生成策略和用户配置参数，生成模拟标准邮件集并存放在指定路径。
3）管理员可通过管理员界面添加垃圾邮件和正常邮件。

5.1.1 用户的配置

用户配置界面的参数见表 5-1。

表 5-1 用户配置参数列表

参数描述			参数名	参数类型	取值范围
用户邮箱数			mail_num	整型	(0, M)　M 为限定的最大邮箱数
总邮件数			total_mail	整型	(0，N)　N 为邮件库最大邮件数
邮件类型			mail_type	整型	{0,1,2} 0：中文邮件；1：英文邮件； 2：两者混合
邮件类型	2	英文邮件的数量	enmail_num	整型	（0，total_num）
		英文邮件中垃圾邮件的数量	enspam_num	整型	（0，enmail_num）
		中文邮件中垃圾邮件的数量	cnspam_num	整型	（0，total_num－enmail_num）
	0/1	垃圾邮件的数量	spam_num	整型	（0，total_num）
邮箱邮件数量的分布(设置可选分布和缺省分布)			distribution	字符串	可选的分布
模拟邮件集的名称			libname	字符串	任何可被计算机识别的字符串
邮件服务器域名			do-name	字符串	任何可被计算机识别的字符串

5.1.2 邮件集生成模块

邮件集生成模块的实现流程为：
1）获取解析的用户配置参数，并依据生成策略从垃圾邮件集、正常邮件集生成邮件；
2）复制生成的邮件到生成邮件集中，并依用户设定的分布函数改写邮件的部分字段，发布该邮件到指定的用户邮箱中，最后将该邮件的路径添加到邮件索引中。

正常邮件集和垃圾邮件集中的邮件都是以数字命名的，编号从 1 开始直到邮件集中总邮件数量。由于在同一邮件集中的各个邮件的地位是一样的，故邮件的生成采用随机生成策略。系统自动生成 1 到邮件最大编号间的随机数，并将该随机数命名的邮件作为生成的邮件。当所需生成的邮件数量与邮件集中的总邮件数量非常接近时，由于生成的相同的随机数的概率非常大，从而使得许多邮件被重复添加到标准邮件集中。为此，本文采用来线性探测法[38]来避免生成重复的邮件。

依据 4.2.4 小节可知，邮件的到达顺序是影响 SFS 评估一个重要影响因素。在现实生活中邮件的到达往往是不可预料的，目前还没有对垃圾邮件与正常邮件的到达顺序的研究。在本文中，假设邮件到达的是随机且相互独立的，新邮件的到达过程如图 5.2 所示。

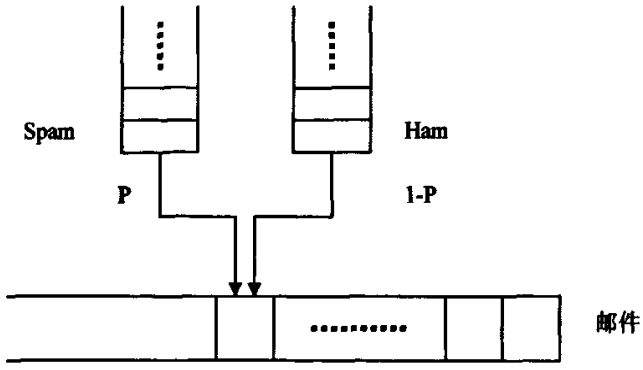

图 5-2 新邮件到达过程图

根据假设，邮件的到达满足两点分布。假设垃圾邮件所占的比重为 P，则新到来的邮件 x 的分布：

表 5-2 两点分布

x	0（正常邮件）	1（垃圾邮件）
P(x)	1 - P	P

依据上述两点分布的定义，邮件的生成策略如下：随机生成[0, 1]之间的数 x，若 x 属于[0, P]则生成垃圾邮件，否则生成正常邮件。同样，假设中文邮件和英文邮件的到达也是互相独立并随机的，它们的生成类似于上述过程。

为合理的将生成的邮件将被分发到各个用户邮箱中，论文假设各用户邮箱中邮件的数量满足一定的函数分布。在本文中，实现了"离散型"均匀分布，"离散型"指数分布和"离散型"正态分布三种可选的函数分布，其中缺省的是正态分布。用户可以选择其中的一个作为用户邮箱中邮件数量的分布函数。

假设用户邮箱数为 n，各用户邮箱名分别为 U_1，U_2，……，U_n，新邮件到达各用户邮箱的概率分别为 P_1，P_2，……，P_n。用户邮件数量满足一定的函数分布可以反映到其对应的概率满足相应的分布。各函数分布的实现如下：

"离散型"均匀分布的各变量值的概率都等于 1/n。其概率分布函数如表 5-3 所示。

表 5-3 "离散型" 均匀分布

X	U_1	U_2	……	U_n
P(x)	1/n	1/n	……	1/n

指数分布 $e(\lambda)$ 的分布函数如公式 5.1 所示。

$$F(x)=\begin{cases}1-e^{-\lambda x} & x>0\\ 0 & x\le 0\end{cases} \qquad \text{公式（5.1）}$$

指数分布的自变量的空间为实数空间。为使得"离散型"随机变量能够满足指数分布的特点，本文将指数分布的自变量划分为 n 个等长的区间，将每个区间当作一个用户的邮箱，落入各个邮箱中的邮件数量因此也指数分布的特征。"离散型"指数分布的概率分布如图 5.4 所示。当 $\lambda=1$ 时，F(10)>0.99995；故本文对[0, 10]区间进行划分。

表 5-4 “离散型”指数分布（$\lambda=1$）

X	U_1	U_2	.....	U_n
P(x)	F(10/n)-F(0)	F(2*10/n)-F(10/n)	.....	1-F((n-1)*10/n)

标准正态分布 $N(0,1)$ 的分布函数如公式 5.2 所示。

$$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{x^2}{2}}dx \qquad \text{公式（5.2）}$$

标准正态分布关于 Y 轴对称。类似于指数分布，为使“离散型”随机变量满足正态分布，需要对自变量的取值范围进行等长划分。由于 $\Phi(-4)\le 0.0001$，故本文将对区间[−4，4]进行等长划分。表 5.5 给出了新邮件进入各邮箱的概率。文献[39]给出了 $\Phi(x),(x\in[-4,4])$ 的取值表。

表 5-5 “离散型”正态分布

X	U_1	U_2	.....	U_n
Y	y1 = -4+8/n	y2 = -4 +2*8/n	.....	yn = 4
P(x)	$\Phi(y1)$	$\Phi(y2)-\Phi(y1)$	.....	$1-\Phi(yn-1)$

采用类似实现两点分布的方式实现上述分布，将[0，1]区间按上述概率划分成 N 个区间，系统随机产生一个[0，1]区间数，该随机数落入的区间号就是对应的用户邮箱号。

邮件生成模块的流程如图 5-3 所示。

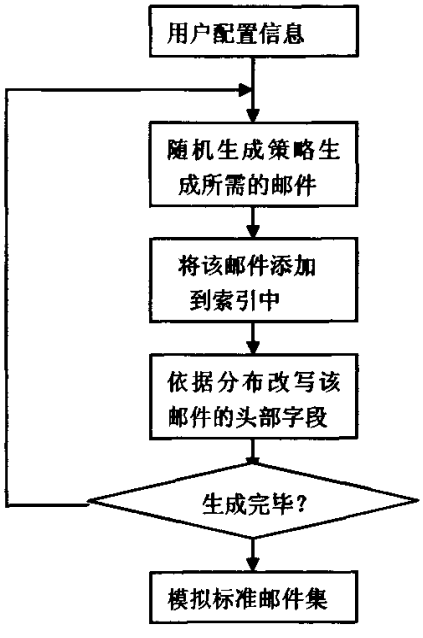

图 5-3 邮件生成模块流程图

5.1.3 管理员界面

管理员界面提供添加正常邮件和垃圾邮件到正常邮件库和垃圾邮件库中的用户接口。通常正常英文邮件的产生是汇集现有的标准邮件集，中文的可以 bbs 或网站上截取一些中立的文章片断作为正常邮件的内容。在本文中，作者将自己的邮箱定制了 google 的论坛的多个讨论组中，论坛用户之间的讨论都将发到笔者的邮箱，这样笔者每天可以获得超过 100 封的正常邮件。英文的垃圾邮件从现有邮件评测集获取的，中文的垃圾邮件通过从众多朋友那收集。作者注册的一个 gmail 邮箱，曾留在 google-talk 刚发布的讨论区中，现在每天能收到 30 封左右的垃圾邮件，经过几个月的积累，已经收集到几千封垃圾邮件。

作者将公开的中英标准邮件集和个人收集的正常邮件和垃圾邮件一起添加到正常邮件和垃圾邮件库中，目前邮件库中包含的正常的中文邮件近 20000 封、中文垃圾邮件近 50000 封、英文的正常邮件和垃圾邮件也都在 50000 封左右。

5.2 SFS 评估系统的总体设计

5.2.1 系统总体结构图

SFS 评估系统主要由评估结果计算子系统和训练评测子系统三个子系统以及用户配置和结果显示模块组成。各子系统之间既相互独立，完成各自独立的功能；又相互依赖，彼此间的数据共享。其系统总体结构图如下：

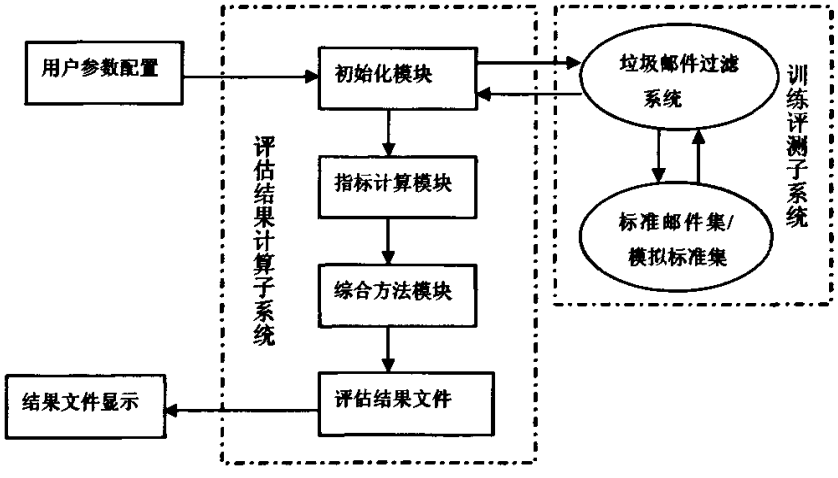

图 5-4 SFS 评估系统的总结构图

用户交互子系统包括用户参数配置和结果文件显示。用户参数配置包括 SFS 名，标准邮件集/模拟标准邮件名等。结果界面将显示 SFS 的各个指标值，多个 SFS 的好坏评判结果。

评估结果计算子系统包括初始化，指标计算，综合评估，评估结果等子模块。初始化模块解析用户的配置文件，启动过滤模块并获得 SFS 的过滤结果。指标计算模块依据初始化子模块的过滤结果计算出文章第二章描述的各个评估指标的值。综合方法模块采用第三章描述的多个综合评估方法对多个 SFS 进行评估。评估结果文件将存放 SFS 的各评估指标值和各 SFS 的综合评估结果。

训练与评测子系统包括 SFS 采用用户配置好的评测训练方法对标准邮件集/模拟标准邮件集进行训练和评测。标准邮件集将由一些研究机构提供，而模拟标准邮件集由模拟标准邮件集生成系统动态生成。

该评估系统的工作流程如下：

1）用户对评估系统进行参数配置，包括（SFS 名，标准邮件名集，垃圾邮件所占的比例、平均误报代价之比等等）。
2）初始化模块解析用户的配置，并驱动 SFS 对标准邮件集进行训练和评测。
3）SFS 对标准邮件集进行训练和评测，将结果反馈给初始化模块。
4）初始化模块将过滤结果传递给指标计算模块和综合方法模块。
5）指标计算模块依据评估指标的计算方法，计算各评估指标的值，并存放在评估结果文件中。
6）综合方法模块依据 5）和 6）的计算结果，判定各 SFS 过滤效果的好坏顺序，并存放在评估结果文件中。
7）结果显示模块将评估结果文件中的评估结果显示给用户。

5.2.2 参数配置文件

参数配置文件定义了系统所需的各个参数的值。评估系统的配置参数见表 5-6。

表 5-6 用户配置参数列表

参数描述			参数名	参数类型	取值范围
配置文件编号			Confid	整型	所有正整数
SFS 系统名			SystemName	字符串	可用的 SFS，用“|”隔开
标准邮件集名			SetName	字符串	可用的标准邮件集，用“|”隔开
误报代价比			Weight	实数型	大于 0 的实数
是否变化训练集所占比例			LoadMultriTrain	布尔型	TURE | FALSE
LoadMultriTrain	TURE	增长的比率	IncreaseRate	实数型	（0，1）
	FALSE	训练集所占比率	TrainRate	实数型	（0，1）
评测训练方法			TrainMethod	字符串	可用的评测训练方法

5.3 训练评测子系统

5.3.1 训练邮件集所占比例变化方案

由第四章可知，训练邮件集所占的比例不同，评估的结果将有很大的差异。为反映 SFS 的过滤能力，有时还需要考虑在不同训练邮件集所占比例下，SFS 的评估结果的情况。

在本文中，将根据用户配置文件中的 LoadMutriTrain 参数来判定是否需要在不同的训练邮件集所占比例下对 SFS 进行评估。当 LoadMutriTrain 为 TURE 时，说明用户需要在不

同的比例下进行评估，这时用户还需要配置比例的增长率；当 LoadMutriTrain 为 FALSE 时，说明用户不需要变化的训练集所占比例，这时用户仅需配置用于当前训练的邮件集的比例。

举例说明，当 LoadMutriTrain::=TRUE 时，IncreasedRate::=0.1 表明训练邮件集所占的比例分别以 0.1、0.2、……、0.9 来对 SFS 进行评估。当 LoadMutriTrain::=FALSE 时，TrainRate::=0.2 表明在当前评估过程中，训练邮件集所占的比例为 20%。也即前 20%的邮件用于训练，后 80%的邮件用于评测。

5.3.2 训练评测方法

配置好了标准邮件集和训练集所占的比例后，训练邮件集和评测邮件集也就给定了。SFS 将采用一定的方法对训练邮件集和评测邮件集进行训练和评测，评测的结果将提交给评估结果计算子系统。由第四章可知，评测训练法对评估结果也存在一定的影响。在本系统中，用户可以自行配置提供前面所述的四种评测训练法中的一种。表 5-7 给出配置名和各方法的对应关系。

表 5-7 配置名与评测训练法的对应关系

配置名	评测训练方法
TrainCla	训练再评测法
ClaTrain	逐一评测训练法
TenCross	十字交叉训练法
FbTrainCla	反馈训练评测法

考虑到现实情况中 SFS 的特点，本文建议用户采用训练评测法或反馈训练评测法对 SFS 进行评估。反馈训练评测法的实现流程如下：

1）按顺序读取训练邮件集中每一封邮件和标准类别，依照标准答案进行训练；

2）对评测邮件集中的每一封邮件，SFS 对其进行分类，记录该邮件的分类结果和得分；

3）将该邮件名，邮件标准类别，邮件分类结果，邮件评分值写入结果文件；

4）如果该评测邮件标准答案为垃圾邮件，则 SFS 对该邮件依照标准答案进行训练；

5）当评测邮件集中的所有邮件都被评测后，将过滤结果提交给评估结果计算子系统。

训练评测法的实现流程与上述类似，唯一不同的是它没有步骤 4）。不管采用哪种评测训练法，SFS 对评测邮件集的过滤结果的格式均为如下所示。

file=邮件名 judge=标准类别 class=分类结果 score=分值

以下是 bogo 过滤系统对某邮件集的部分过滤结果：

file=chinese/spam/31366 judge=spam class=ham score=0.520000

file=chinese/ham/12359 judge=ham class=ham score=0.520000

file=chinese/ham/9385 judge=ham class=ham score=0.504421

file=chinese/spam/26791 judge=spam class=ham score=0.498310

file=chinese/spam/8061 judge=spam class=ham score=0.501722

file=chinese/spam/3276 judge=spam class=ham score=0.520000

file=chinese/ham/13789 judge=ham class=ham score=0.520000

file=chinese/ham/382 judge=ham class=ham score=0.07821

5.4 评估结果计算子系统

5.4.1 指标计算模块

依据第二章所讲述的原理，统计评测训练子系统的过滤结果中的 judge、class 的值得到“归一化”复合矩阵。依据该矩阵和各评估指标的定义，可以很容易计算出各基本指标和合成性指标的值。

在第二章中，论文已经介绍了 SFS 的 ROC 曲线的基本特征。依据 SFS 的特点，在绘制 ROC 曲线中，论文采用从大到小的 score 值作为 SFS 的分类阀值，以各阀值下的坐标点绘制曲线。其算法如下：

SFS 的 ROC 曲线的点坐标生成算法如下（算法 5.1）

输入：L 表示标准邮件评测集；i 表示某一封邮件； f(i)表示 SFS 对邮件 i 计算的分值；min 和 max 表示垃圾邮件对邮件所评判的最小和最大值；increment 表示两个相邻评分值之间的最小的间隔。TP 和 FP 分别表示 SFS 正确和错误判定垃圾邮件的数量。N 和 P 分别表示标准邮件集中的正常邮件和垃圾邮件的数量。

1） 分类阀值增加 increment，开始时分类阀值为 min。

2） 将 FP，TP 都赋 0。

3） 从 L 中选择邮件 i；如果 f(i)的值小于分类阀值，则跳回 3）循环。

4） 否则，如果 i 是垃圾邮件，则 TP++；如果不是，则 FP++。

5） 跳回 3），直至 L 中所有邮件都被选择过。

6） 计算点坐标（FP/N，TP/P），并将它添加到点坐标的文件中。

7） 跳回 1），循环至分类阀值 max。

当点坐标生成之后，论文采用开源的 Gnuplot[40]画图软件来绘制 SFS 的 ROC 曲线。

- ROC 曲线下的面积。在本论文中，根据第二章讲述的梯形法原理，,给出计算 ROC 曲线下的面积的算法如下：

梯形法计算 ROC 曲线下的面积的算法（算法 5.2）：

输入：L 表示评测邮件集；i 表示邮件编号； f(i)表示 SFS 对邮件 i 计算的分值；N 和 P 分别表示标准邮件集中的正常邮件和垃圾邮件的数量。TP 和 FP 分别表示当前阀值下，SFS 正确和错误判定垃圾邮件的数量。TP_{prev} 和 FP_{prev} 分别表示阀值变化前，SFS 正确和错误判定垃圾邮件的数量。

输出：A 表示 ROC 曲线下的面积。

1） 将 L 依据 f 的值降序排序得到 L_{sort} 邮件集记录；

2） 将 TP、FP、TP_{prev}、FP_{prev}、A 都赋 0，i 赋 1，f_{prev} 赋 $-\infty$；

3） 如果 f(i) != f_{prev}, 则 TRAPEZOID_AREA(FP, FP_{prev}, TP, TP_{prev})+A 赋给 A；

4） 如果邮件 i 是垃圾邮件，则 TP++，否则 FP++；

5） 当 i<=|L_{sort}|，跳回 3）循环；

6） 将 TRAPEZOID_AREA(N, FP_{prev}, P, TP_{prev})+A 赋给 A；

7） 将 A/(P*N)赋给 A；

函数 TRAPEZOID_AREA(X1, X2, Y1, Y2)

功能：计算以(X1,Y1),(X2,Y2),(X1, 0),(X2,0)四点构成的梯形的面积。

1） return |X1-X2|×(Y1+Y2)/2;

● 最优分割点。假设评测邮件集中正常邮件占总邮件数量的比例为 P_0，正常邮件的评均误报代价是垃圾邮件的平均误报代价的 λ 倍时，那么 ROC 曲线的最优分割点是 ROC 曲线上斜率等于 $\lambda * P_0/(1- P_0)$的点。计算的方法类似与求曲线面积的算法。计算 ROC 曲线上每一点得斜率，选择斜率与 $\lambda * P_0/(1- P_0)$最接近的点为最优分割点。

5.4.2 综合评估方法模块

综合评估方法模块包括模糊综合评估方法和基于 ROCCH 方法的综合评估方法的实现。模糊综合评估的实现流程见图 5-5。

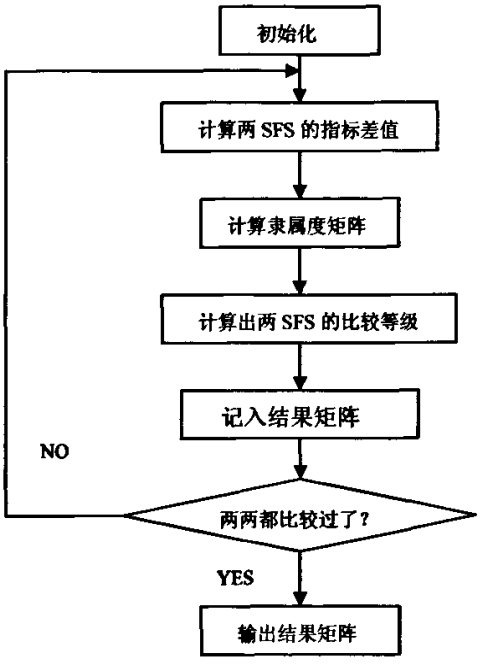

图 5-5 模糊综合评估方法流程图

通过图 5-5 得到的是多个 SFS 的结果矩阵。为得到各 SFS 之间的优劣次序，依据 3.1 节的原理，需要求得该矩阵的最大特征根对应的特征向量。在本文中，需要用户使用 Matlab[41]来计算对应的特征向量，并依照特征向量各下标值的大小来排列 SFS 的优劣次序。

依据 3.2 节的原理，基于 ROCCH 方法综合评估的实现流程如图 5-6 所示。

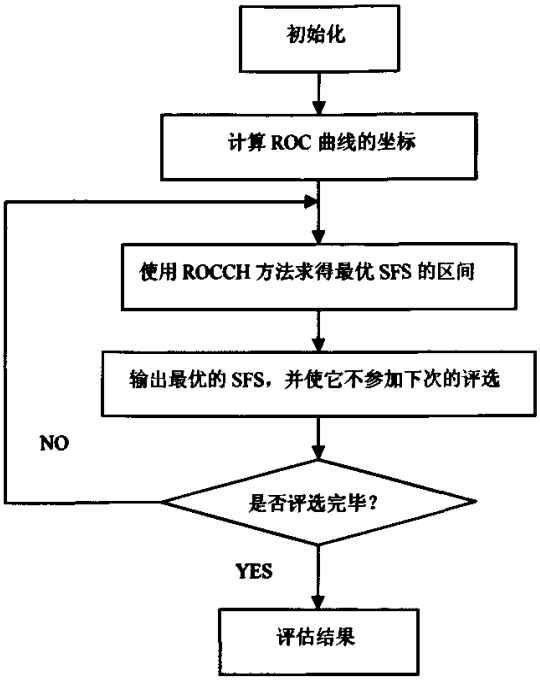

图 5.6 基于 ROCCH 方法的综合评估流程图

在图 5-6 中，依据算法 5.1 求得在标准邮件集下 ROC 曲线的点坐标。ROCCH 方法将采用 Tom Fawcett 在 HP 实验室开发的 ROCCH 软件包来实现。

5.5 其他模块

系统的评估结果，将存放在以配置文件编号命名的目录下。在该目录中，包括各 SFS 在标准邮件集下计算的所有指标值，ROC 曲线的点坐标和模糊综合评判及基于 ROCCH 方法的综合评估结果。

其中各 SFS 系统的评估指标值存放在以 SFS 名命名的目录中，该结果文件名为 SFS 名＋标准邮件集名＋训练邮件数量＋.index。例如 bogo 系统对 spamassassin_corpus 邮件集评测（训练邮件数为 2000）后的各指标存放在 bogo 目录下 bogospamassassin_corpus2000.index 文件中。对应的 ROC 曲线点坐标存放在 bogo 目录下的 bogospamassassin_corpus2000.SRHM 文件中。

模糊综合评估的结果存放在以 confid+编号+.fy 命名的文件中，它包含了多个 SFS 之间的比较矩阵。基于 ROCCH 方法的综合评估结果存放在 average_roc 命名的目录下，该目录包括多个 SFS 的 ROC 曲线的点坐标。

除此之外，对于需要变换训练邮件集所占比例的 SFS 评估中，将在结果目录中生成 MutriTra.index 文件，用来存放各不同训练集比例下的各评估指标值。

5.6 本章小结

本章首先介绍了模拟标准邮件集的生成系统的设计与实现。在该生成系统的设计与实现中，论文定义了它所需的各个用户参数配置，提出邮件的生成采用随机生成策略，假设邮件的到达是独立的并给出该假设下邮件的生成顺序的实现，定义和实现了用户邮箱中邮件数量的分布函数，以及管理员模块的设计与实现。

论文随后介绍了 SFS 的评估系统的设计与实现。它包括系统的总体结构图，用户配置参数的定义，评测训练方法，指标的计算，综合评估方法的计算等多个子模块的设计与实现。

第六章 评估结果及分析

6.1 实验背景介绍

本实验使用的标准邮件集是生成的模拟标准邮件集 alice（简称邮件集 A）。该邮件集由模拟标准邮件集生成系统生成的，有 20 个邮件用户，共计 20000 封邮件，中英文邮件数量的比例为 4：1，其中英文垃圾邮件 2400 封、占总数的 12%，英文正常邮件 1600、占总数的 8%，中文垃圾邮件 9600、占总数的 48%，中文正常邮件 6400、占总数的 32%，各用户的邮件数量的分布为正态分布，邮件服务器的域名为 njnet.edu.cn。该邮件集 A 的配置文件如表 6-1 所示。

表 6-1 生成邮件集 A 的配置文件

User::=20
Total::=20000
Mailtype::=both
Enmail::=4000,Enspam::=2400,Cnspam::=9600
Distribute::=default
Libname::=alice
Domain-name::=njnet.edu.cn

本实验使用的另一个标准邮件集 dataset 是徐选[42]等人发布的。该标准邮件集依托华东（北）地区网络中心的邮件服务器所收到的邮件，其收集采取用户反馈的方式，邮件集中的正常邮件都进行了匿名化。该标准邮件集共计 25944 封邮件，绝大多数都是中文邮件，其中正常邮件 3183 封、占 12.27%，垃圾邮件 22760 封、占 87.73%。称该标准邮件集为邮件集 B。

徐激[43]依据网络中心的研究人员和工作人员的实际需求，设计和实现了一个基于网络中心邮件服务器的 SFS。为获得更高的过滤效果，相关研究人员对该系统进行了更改。未修改的过滤系统称为 NC-1，而经过修改后的过滤系统称为 NC-2。

本实验依据第五章设计的评估系统，使用上述两个标准邮件集评估 Bogofilter、CRM114、Spamassassin 和 SpamProbe 四个开源的 SFS 以及网络中心的 NC-1 和 NC-2 两个 SFS 的过滤能力。

6.2 SFS 的评估指标结果

依据第五章的评估系统，采用邮件集 A 和邮件集 B，对 Bogofilter、CRM114、Spamassassin、SpamProbe、NC-1 和 NC-2 六个开源的 SFS 的过滤能力进行评测，评测训练法采用先训练再评测法，训练邮件的比例为 20%，误报代价比 9，不变化训练邮件所占比例，该评估实验的参数配置如表 6-2 所示。

表 6-2 评估实验的参数配置

Confid::=15
SystemName::=bogo|crm114|dbacl|spamassassin|spambayes|spamprobe
SetName::=alice|dataset

TotalMailNum::=20000|25944
Weight::=9
LoadMultriTrain::=FALSE
TrainRate::=0.2
TrainMethod::=TrainCla

6.2.1 评估指标

上述实验得到的各 SFS 在邮件集 A 下的十六个评估指标的值见表 6-3。

表 6-3 六个 SFS 在邮件集 A 下的评估指标值（一）

指标 / 过滤系统	SR	SP	HR	HP	ACC	ERR	SM	HM
Bogofilter	0.8628	0.9978	0.9997	0.9779	0.9804	0.0196	0.1372	0.0003
CRM114	0.9642	0.9199	0.9862	0.9941	0.9831	0.0169	0.0358	0.0138
Spamassassin	0.9621	0.9780	0.9964	0.9938	0.9916	0.0084	0.0379	0.0036
SpamProbe	0.9313	0.9890	0.9983	0.9888	0.9888	0.0112	0.0687	0.0017
NC-1	0.9738	0.9663	0.9944	0.9957	0.9915	0.0085	0.0262	0.0056
NC-2	0.9769	0.9628	0.9938	0.9962	0.9914	0.0086	0.0230	0.0062

表 6-3 六个 SFS 在邮件集 A 下的评估指标值（二）

指标 / 过滤系统	SF_1	HF_1	LR	ZR	YI	AMC	TCR	MCC
Bogofilter	0.9253	0.9887	457.5	44.33	0.8624	0.083	7.1875	0.9174
CRM114	0.9415	0.9901	11.489	167.53	0.9504	0.0714	8.3508	0.9320
Spamassassin	0.9700	0.9951	44.362	159.91	0.9585	0.0356	16.775	0.9651
SpamProbe	0.9593	0.9935	89.788	88.409	0.9296	0.0472	12.642	0.9534
NC-1	0.970	0.9951	28.657	230.796	0.9682	0.0359	16.612	0.9651
NC-2	0.9698	0.9950	25.875	262.432	0.9707	0.0363	16.44	0.9648

从表 6-3 的评估结果中可以看出，Bogofilter 在 SP、HR、HM、LR 四个指标值上有最优值，而在 SR、 HP、ACC、ERR、SF_1、HF_1、ZR、YI、AMC 和 TCR 等指标上最差；这说明该系统对正常邮件的过滤能力极强，但对垃圾邮件的过滤能力比较弱。与之相反的是 CRM114 在 SR、HP、SM、ZR 指标上有较好的表现，说明它对垃圾邮件的过滤能力很强，但对正常邮件的额过滤能力却最差。Spamassassin 和 Spamprobe 两个 SFS 的过滤能力比较平衡，对两类邮件的过滤能力都较强。NC-1 和 NC-2 对垃圾邮件的过滤能力最强，但对正常邮件的过滤能力只能算中等。

同理，在邮件集 B 下，各 SFS 的评估结果如表 6-4 所示。

表 6-4 六个 SFS 在邮件集 B 下的评估指标值（一）

指标 / 过滤系统	SR	SP	HR	HP	ACC	ERR	SM	HM
Bogofilter	0.6534	0.9762	0.9881	0.7930	0.8453	0.1549	0.3466	0.0119
CRM114	0.9949	0.5548	0.4060	0.9907	0.6572	0.3428	0.0052	0.5940
Spamassassin	0.7616	0.8551	0.9040	0.8360	0.8432	0.1568	0.2383	0.0960
SpamProbe	0.7268	0.9486	0.9707	0.8269	0.8667	0.1333	0.2732	0.0293
NC-1	0.8270	0.9007	0.9321	0.8786	0.8872	0.1127	0.1730	0.0679
NC-2	0.8546	0.8942	0.9247	0.8953	0.8948	0.1052	0.1454	0.0753

表 6-4 六个 SFS 在邮件集 B 下的评估指标值（二）

指标 / 过滤系统	SF_1	HF_1	LR	ZR	YI	AMC	TCR	MCC
Bogofilter	0.7828	0.8799	40.969	3.8314	0.6415	0.3155	2.7579	0.7024
CRM114	0.7123	0.5760	1.2461	105.97	0.4009	0.6991	1.2445	0.4676
Spamassassin	0.8056	0.8686	5.9005	5.0969	0.6656	0.3197	2.7214	0.6782
SpamProbe	0.8231	0.8930	18.461	4.7764	0.6975	0.2719	3.1997	0.7355
NC-1	0.8622	0.9046	9.067	7.2403	0.7591	0.2299	3.784	0.7691
NC-2	0.8739	0.9098	8.4477	8.5508	0.7794	0.2145	4.0564	0.7844

从表 6-4 可以看出，类似于表 6-3，Bogfilter 对正常邮件的过滤能力超强，而对垃圾邮件的过滤能力却非常弱；CRM114 正好和它相反。其他 SFS 的过滤结果也与 6-3 类似。对垃圾邮件过滤强的，对正常邮件过滤就弱些，反之也亦然。

另外，可以发现，表 6-3 中各 SFS 对应的评估指标值要比表 6-4 的要高得多。这也说明标准邮件集对评估得结果有较大影响。本文认为是邮件的内容和两类邮件的比例的较大差别导致了上述情况。

6.2.2 曲线指标评估结果

上述六个 SFS 中，NC-1 和 NC-2 都是对多个开源的 SFS 修改后，再叠加的方式组成，没有对邮件进行评分。为此，论文只能计算剩余四个 SFS 的曲线性指标。

图 6-1 为四个 SFS 在邮件集 A 下的 ROC 曲线。四条曲线非常接近，且非常靠近（0，1）点，说明这四个 SFS 之间相差不多，且都具有较好的过滤能力。表 6-5 为邮件集 A 下，各 SFS 的 AUC。

表 6-5 在邮件集 A 下，各 SFS 的 AUC

SFS	Bogofilter	CRM114	Spamassassin	SpamProbe
AUC	0.999514	0.994187	0.996029	0.996694

表 6-5 可以看出，各 SFS 的 AUC 都大于 0.99，介于 0.994～0.9995 之间，差别非常小。

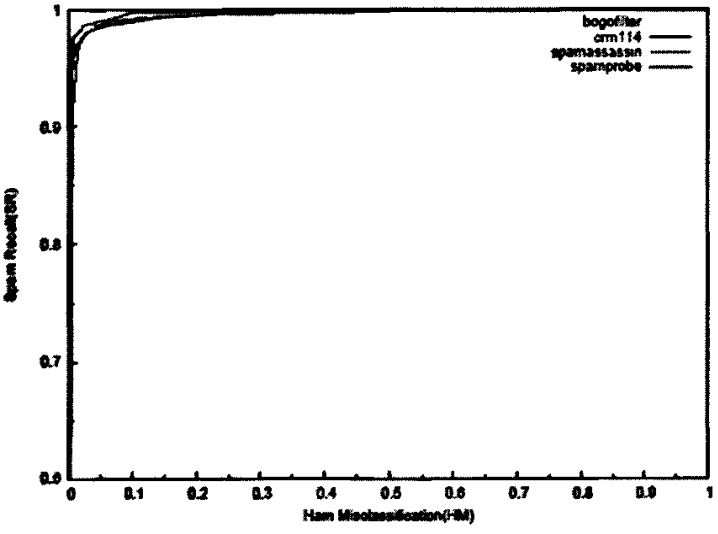

图 6-1 SFS 在邮件集 A 下的 ROC 曲线

- SFS 的最优分割点。各 SFS 的最优分割点的分类阀值见表 6-6。

表 6-6 SFS 最优分割点对应的阀值

SFS	Bogofilter	CRM114	Spamassassin	SpamProbe
Best Threshold	0.999978	-2.1171	4.6	0.077198

在最优分类阀值下，各 SFS 对应的各评估指标的值如表 6-7 所示。

表 6-7 最优分割点对应的各评估指标值（一）

指标 / 过滤系统	SR	SP	HR	HP	ACC	ERR	SM	HM
Bogofilter	0.8628	0.9978	0.9997	0.9779	0.9804	0.0196	0.1372	0.0003
CRM114	0.9723	0.9044	0.9831	0.9954	0.9816	0.0184	0.0277	0.0169
Spamassassin	0.9648	0.9743	0.9958	0.9942	0.9914	0.0086	0.0352	0.0042
SpamProbe	0.9605	0.9559	0.9927	0.9935	0.9882	0.0118	0.039	0.0073

表 6-7 最优分割点对应的各评估指标值（二）

指标 / 过滤系统	SF_1	HF_1	LR	ZR	YI	AMC	TCR	MCC

Bogofilter	0.9253	0.9887	457.5	44.33	0.8624	0.083	7.1875	0.9174
CRM114	0.9371	0.9892	9.4608	216.34	0.9555	0.0778	7.6667	0.9272
Spamassassin	0.9695	0.9950	37.897	172.18	0.9606	0.0362	16.485	0.9646
SpamProbe	0.9582	0.9931	21.674	152.98	0.9532	0.0500	11.931	0.9513

将表 6-7 与表 6-3 对比可以发现，SpamProbe 在最优分割点的 SR、HP 等指标上更优，而在 HR、SP 等指标上更差，其他的 SFS 与 SpamProbe 类似。由于用户希望尽可能的少丢失正常邮件，所以 SFS 在缺省状态下，阀值都设置的较高，也就导致了通常缺省阀值下，SFS 对正常邮件的过滤能力非常强，而对垃圾邮件的过滤能力相对较弱。在最优分割点，是在给定误报代价后，基于总误报代价最小的原则，而对垃圾邮件和正常邮件过滤能力达成的"折中"。

同样，四个 SFS 在邮件集 B 下的 ROC 曲线如图 6-2 所示。从图中可以看出，Bogofilter 的曲线最靠近左上角；Spamassassin 和 Spamprobe 相互交叉，各有所长；CRM114 最差。表 6-8 为各 SFS 在邮件集 B 下的 AUC。

表 6-8 邮件集 B 下，各 SFS 的 AUC

SFS	Bogofilter	CRM114	Spamassassin	SpamProbe
AUC	0.975411	0.818564	0.896945	0.899517

表 6-8 可以发现，Bogofilter 的 AUC 最大，CRM114 的最小，其余两个相差不大。

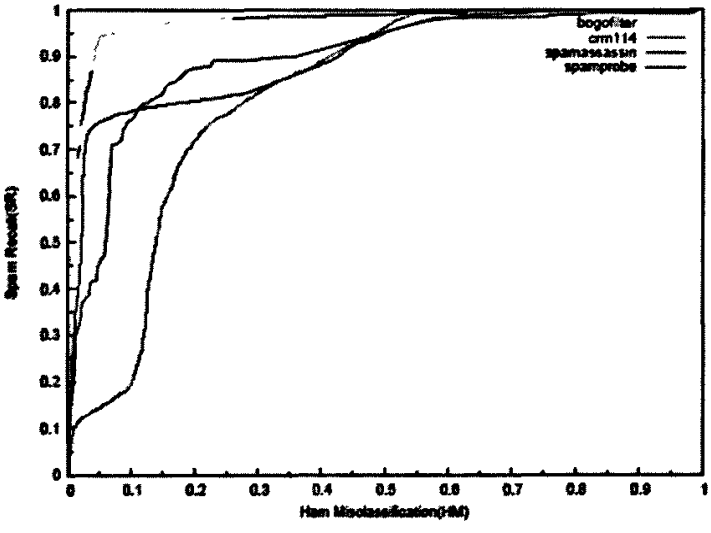

图 6-2 SFS 在邮件集 B 下的 ROC 曲线

- SFS 的最优分割点。在邮件集 B 下，各 SFS 的最优分割点对应阀值如表 6-9 所示。

表 6-9 邮件集 B 下，SFS 最优分割点对应的阀值

SFS	Bogofilter	CRM114	Spamassassin	SpamProbe
Best Threshold	0.500214	3.2031	4.5	0.008852

在最优分类阀值下，各 SFS 对应的各评估指标的值如表 6-10 所示。

表 6-10 最优分割点对应的各评估指标值（一）

指标 / 过滤系统	SR	SP	HR	HP	ACC	ERR	SM	HM
Bogofilter	0.9079	0.9416	0.9581	0.9333	0.9367	0.0633	0.0921	0.0419
CRM114	0.8841	0.7466	0.7768	0.9001	0.8226	0.1774	0.1159	0.2232
Spamassassin	0.7673	0.8466	0.8965	0.8381	0.8414	0.1586	0.2327	0.1034
SpamProbe	0.7734	0.8848	0.9251	0.8458	0.8604	0.1396	0.2266	0.0749

表 6-10 最优分割点对应的各评估指标值（二）

指标 / 过滤系统	SF_1	HF_1	LR	ZR	YI	AMC	TCR	MCC
Bogofilter	0.9244	0.9455	16.122	13.984	0.8660	0.1291	6.738	0.8704
CRM114	0.8096	0.8339	2.9468	9.0086	0.6609	0.3619	2.4043	0.6538
Spamassassin	0.8049	0.8663	5.5181	5.1777	0.6638	0.3235	2.6897	0.6742
SpamProbe	0.8253	0.8837	7.682	5.486	0.6984	0.2848	3.0551	0.7144

6.3 SFS 的综合评估结果

对于上述评估指标值，分别采用模糊综合评估方法、因子分析法和 ROCCH 方法进行评估，评估的结果如下。

6.3.1 模糊综合评估法

首先计算邮件集 A 下的模糊综合评估结果。

计算两两 SFS 之间的比较等级，Bogofilter 与 CRM114 两系统的比较等级计算如下：

1）将两个 SFS 的各指标值相减，得到$\Delta U=\{\Delta O_1, \Delta O_2, \Delta O_3\}$其中$\Delta O_1=\{-0.1014, 0.0779, 0.0162, 446.011\}$，$\Delta O_2=\{0.0135, 0.0162, 0.0014, 123.2\}$，$\Delta O_3=\{0.0027, -0.088, 1.9574, -1.1633, 0.0146\}$。

2) 依据上述指标差值和 3.1.2 小节定义的隶属度函数,计算出各个隶属度矩阵.

$$R(O_1)=\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0.465 & 0.535 & 0 \\ 0 & 0.9475 & 0.0525 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.595 & 0.405 & 0 & 0 & 0 \\ 0.8650 & 0.1350 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$R(O_2)=\begin{bmatrix}0&0&0&0.3375&0.6625&0&0&0&0\\0&0&0&0&0.595&0.405&0&0&0\\0&0&0&0&0.965&0.035&0&0&0\\0&0&0&0&0&0&0&0.942&0.058\end{bmatrix}$$

$$R(O_3)=\begin{bmatrix}0&0&0&0&0.9325&0.0675&0&0&0\\0&0&0&0&0&0.240&0.760&0&0\\0&0&0&0&0.8043&0.1957&0&0&0\\0&0&0&0&0.8837&0.1163&0&0&0\\0&0&0&0&0.7080&0.2920&0&0&0\end{bmatrix}$$

3) 依据 3.1.3 小节求得的权重向量，上述隶属度矩阵以及公式(2.9),求出评估结果 O 的隶属度矩阵如下。

$$R(O)=\begin{bmatrix}0.0787&0.0123&0.3449&0.0191&0.1083&0.0737&0.1693&0.1947&0\\0&0&0&0.1229&0.6334&0.1538&0&0.0857&0.0053\\0.5844&0&0&0&0.5844&0.1691&0.2455&0&0\end{bmatrix}$$

依据公式，得 $B(O)=\{0.032\ \ 0.005\ \ 0.138\ \ 0.057\ \ 0.414\ \ 0.125\ \ 0.117\ \ 0.112\ \ 0.002\}$。

4) 依据 $B(O)$ 的值和最大隶属度原则,评估 Bogofilter 和 CRM114 两个过滤系统的结果为 e_4。通过表 7 可以查得,对应结果矩阵 $A=(a_{ij})_{6\times6}$ 中 $a_{12}=1$。

5）采取同样的方法，六个 SFS 按照上述方法两两比较后得到的结果矩阵如下：

$$A=\begin{bmatrix}1&1&1&1&1/2&1/2\\1&1&1&1&1/2&1/2\\1&1&1&1&1&1\\1&1&1&1&1&1\\2&2&1&1&1&1\\2&2&1&1&1&1\end{bmatrix}$$

对上述矩阵求得最大特征根 $\lambda=6.1072$，满足一致性条件，该特征根对应得特征向量为 V={0.3129，0.3129，0.3942，0.3942，0.4967，0.4967}。依据特征向量值的大小，可以将六个 SFS 按照优劣分为{NC-1,NC-2}，{Spamassassin,SpamProbe}和{Bogofilter,CRM114}三组。由于模糊综合评估方法评估等级的粒度原因，未能给出更详细的排名。

同理，类似与上述流程，计算邮件集 B 下，模糊综合评判得到的最后结果矩阵为如下：

$$A=\begin{bmatrix}1&9&1/2&1/2&1/5&1/6\\1/9&1&1/9&1/9&1/9&1/9\\2&9&1&1/2&1/3&1/4\\2&9&2&1&1/3&1/3\\5&9&3&3&1&1\\6&9&4&3&1&1\end{bmatrix}$$

对上述矩阵求得最大特征根 λ =6.4142，计算 CI=(λ −6)/5=0.08284<0.1*RI_6，满足一致性条件。该结果矩阵对应的特征向量 V={0.1575, 0.0403, 0.2210, 0.2866, 0.6189, 0.6779}。为此，模糊综合法判定在邮件集 B 下，各 SFS 的好坏顺序为 NC-2，NC-1，SpamProbe，Spamassassin，Bogofilter，CRM114。这与邮件集 A 的模糊综合评估结果相同。

6.3.2 因子分析法

本节将采用 3.2 节介绍得因子分析法和 6.2.1 节的各个指标值来评估六个开源的 SFS。论文采用 SPSS 软件对邮件集 A 下的评估指标评估采用因子分析法如下。

1）构造因子变量和计算因子得分

采用主成分分析法处理后，从原来的十六个指标中提取出了 2 个因子变量。这两个因子的特征值、特征值占方差的百分数、特征值占方差百分数的累加值如表 6-11 所示。2 个因子所解释的方差占整个方差的 98%以上，能够比较全面的反映原指标的信息。

依据 3.2 节介绍的原理，计算因子载荷矩阵、Varimax 旋转后后的因子载荷矩阵以及因子得分矩阵，分别见于表 6-12、表 6-13。

表 6-11 方差贡献表

因子变量	特征值	贡献率	累计贡献率
1	11.489	71.809	71.809
2	4.219	26.367	98.176

表6-12 因子载荷矩阵

评估指标	因子载荷矩阵 A_0		旋转后的因子载荷矩阵 A	
	第一因子F1	第二因子F2	第一因子F1	第一因子F2
SR	0.953	-0.302	0.726	0.687
SP	-0.355	0.935	0.087	-0.996
HR	-0.341	0.940	0.103	-0.994
HP	0.954	-0.301	0.727	0.686
ACC	0.912	0.410	0.999	0.028
ERR	-0.912	-0.410	-0.999	-0.028
SM	-0.953	0.302	-0.726	-0.687
HM	0.341	-0.940	-0.103	0.994
SF_1	0.947	0.317	0.991	0.128
HF_1	0.905	0.425	1.000	0.012
ZR	-0.911	0.313	-0.684	-0.679
LR	0.868	-0.235	0.679	0.590
YI	0.977	-0.212	0.787	0.617
AMC	-0.912	-0.408	-0.999	-0.030

TCR	0.886	0.440	0.989	-0.010
MCC	0.932	0.361	0.996	0.082

表6-13 因子得分矩阵

评估指标	因子得分矩阵	
	第一因子F1	第一因子F2
SR	0.043	0.101
SP	0.069	-0.213
HR	0.070	-0.213
HP	0.044	0.100
ACC	0.114	-0.053
ERR	-0.114	0.053
SM	-0.043	-0.101
HM	-0.070	0.213
SF1	0.107	-0.032
HF1	0.115	-0.056
ZR	-0.039	-0.101
LR	0.044	0.083
YI	0.055	0.082
AMC	-0.114	0.053
TCR	0.115	-0.060
MCC	0.110	-0.042

2）对因子变量的解释

从表 6-12 可以看出，在旋转后的因子载荷矩阵 A 中，ACC、SF_1、HF_1、TCR 和 MCC 指标在第一主因子变量 F1 都超过 0.99 的负荷，表明第一因子变量与它们有较大关系，论文认为第一因子变量反映的是 SFS 对两类邮件整体的过滤能力。指标 SR、HP、HM、LR 和 YI 在第二因子变量上有较大负荷，而 SP、HR、SM 和 ZR 在第二因子变量上有很大的负负荷，为此论文认为第二因子变量反映 SFS 对垃圾邮件的过滤能力。

另外，表 6-11 说明了第一因子变量的贡献是第二因子变量贡献的 3 倍左右，表明在 SFS 的评估中，SFS 整体的过滤能力比对垃圾邮件的过滤能力更加重要得多。

3）综合评估 SFS

依据因子得分矩阵表6-13、因子的方差贡献率表6-11和公式3.12，计算出各SFS的评估结果见表6-14。

表6-14 因子分析法的评估结果

SFS	第一因子变量 F_1	排名	第二因子变量 F_2	排名	总得分 E	综合排名
Bogofilter	-1.46712	6	-1.18578	6	-1.36618	6
CRM114	-1.02295	5	1.70167	1	-0.28589	5
Spamassassin	0.77828	1	-0.29328	4	0.48155	3

SpamProbe	0.18890	4	-0.69845	5	-0.04851	4
NC-1	0.76873	2	0.16311	3	0.59503	2
NC-2	0.75418	3	0.31272	2	0.62402	1

表 6-14 给出了在邮件集 A 下，使用因子分析法评估六个开源的 SFS 的结果。从中可以看出，综合排名的顺序依次为 NC-2、NC-1、Spamassassin、SpamProbe、CRM114、Bogofilter。这个结果与邮件集 A 下，采用模糊综合评估的结果相吻合。

从表 6-14 还可以发现，Spamassassin、NC-1 和 NC-2 的整体过滤能力最强，Bogofilter 和 CRM114 的整体过滤能力最差；CRM114、NC-1 和 NC-2 对垃圾邮件的过滤能力最强，而 Bogofilter 对垃圾邮件的过滤能力最差，这与前面的结论一致。

类似于邮件集 A，使用因子分析法综合评估各 SFS 在邮件集 B 下的结果如下。

1） 构造因子变量和计算因子得分

采用主成分分析法处理后，从原来的十六个指标中提取出了 2 个因子变量。方差贡献表、因子载荷矩阵、Varimax 旋转后后的因子载荷矩阵以及因子得分矩阵，分别见于表 6-15、表 6-16 和表 6-17。

表 6-15 方差贡献表

因子变量	特征值	贡献率	累计贡献率
1	13.205	82.529	82.529
2	2.510	15.684	98.213

表6-16 因子载荷矩阵

评估指标	因子载荷矩阵 A_0		旋转后的因子载荷矩阵 A	
	第一因子 F1	第二因子 F2	第一因子 F1	第一因子 F2
SR	-0.794	0.601	-0.311	-0.946
SP	0.980	-0.179	0.705	0.705
HR	0.995	-0.094	0.764	0.644
HP	-0.839	0.521	-0.394	-0.905
ACC	0.978	0.207	0.922	0.386
ERR	-0.978	-0.208	-0.923	-0.385
SM	0.794	-0.601	0.310	0.946
HM	-0.995	0.094	-0.764	-0.644
SF_1	0.827	0.560	0.998	0.010
HF_1	0.995	0.089	0.869	0.493
ZR	0.491	-0.765	-0.031	0.909
LR	-0.989	0.037	-0.793	-0.593
YI	0.929	0.368	0.974	0.226
AMC	-0.978	-0.208	-0.923	-0.386

TCR	0.868	0.472	0.982	0.106
MCC	0.960	0.264	0.939	0.329

表6-17 因子得分矩阵

评估指标	因子得分矩阵	
	第一因子F1	第一因子F2
SR	0.087	-0.231
SP	0.021	0.101
HR	0.041	0.074
HP	0.066	-0.207
ACC	0.108	-0.026
ERR	-0.108	0.026
SM	-0.087	0.231
HM	-0.041	-0.074
SF1	0.178	-0.148
HF1	0.082	0.014
ZR	-0.143	0.272
LR	-0.053	-0.055
YI	0.141	-0.081
AMC	-0.108	0.026
TCR	0.161	-0.117
MCC	0.120	-0.045

2）对因子变量的解释

类似于对表 6-12 中第一因子变量的解释，表 6-16 中的第一因子变量反映的是 SFS 整体的过滤能力。与先前解释不同的是，表 6-16 中第二因子变量与 SP、HR、SM、ZR 等指标具有较高的相关度，论文认为它反映了 SFS 对正常邮件的过滤能力。

3）综合评估 SFS

依据因子得分矩阵表6-17、因子的方差贡献率表6-15和公式3.12，计算出各SFS的评估结果见表6-18。

表6-18 因子分析法的评估结果

SFS	第一因子变量 F_1	排名	第二因子变量 F_2	排名	总得分 E	综合排名
Bogofilter	-0.51639	5	1.60810	1	-0.17396	5
CRM114	-1.66753	6	-1.14535	6	-1.55583	6
Spamassassin	0.03222	4	0.19840	3	0.05771	4
SpamProbe	0.20898	3	0.54635	2	0.25816	3
NC-1	0.86909	2	-0.47088	4	0.64340	2

NC-2	1.07362	1	-0.73661	5	0.770518	1

由表 6-18 可以看出，在邮件集 B 下，采用因子分析法综合评估各 SFS 的优劣顺序为 NC-2、NC-1、SpamProbe、Spamassassin、Bogofilter、CRM114。这与邮件集 B 下，采用模糊综合评估的结果完全一致。

从表 6-18 中还可以发现，NC-2 的整体过滤能力最好，而 Bogofilter 对正常邮件的过滤能力最好，这与前面所得到的结论保持一致。

6.3.3 ROCCH 评估方法

ROCCH方法基于ROC曲线，故在本节只能评估四个开源的SFS的综合过滤能力。

在邮件集A中，垃圾邮件占总邮件数量的比例为60%，误报代价 λ 取为9，则最优性能曲线的斜率 $m=\frac{40\%*9}{60\%*1}=6$。

使用ROCCH方法对四个SFS进行评估后，评估得结果如表6-11所示。

表6-19 邮件集A下，各SFS对应的最优斜率范围

最优斜率 / SFS	第一次评估	第二次评估	第三次评估	第四次评估
Bogofilter	[0,1547.8]	√		
CRM114	无	无	[0.03,0.373]	[0,+∞]
Spamassassin	无	[0.003,15.16]	√	
SpamProbe	[1547.8,∞]	[0,0.003] [15.16,+∞]	[0,0.03] [0.373,+∞]	√

依据表 6-11，ROCCH 方法评估六个 SFS 的顺序为 Bogofilter、Spamassassin、SpamProbe、CRM114。这与因子分析法的评估结果基本一致。

同理，在邮件集 B 中，垃圾邮件数量占总邮件数量的 87.73%，误报代价比为 9，则最优曲线的斜率 $m=\frac{12.27\%*9}{87.73\%*1}=1.259$。使用 ROCCH 方法对四个 SFS 进行评估后，评估结果如表 6-12 所示。

表6-20 邮件集B下，各SFS对应的最优斜率范围

最优斜率 / SFS	第一次评估	第二次评估	第三次评估	第四次评估
Bogofilter	[0.033, 99.601]	√		
CRM114	[0,0.003] [0.018, 0.033]	[0,0.003] [0.018, 0.033]	[0,0.315]	[0,+∞]
Spamassassin	无	[0.315, 0.795]	[0.315,+∞]	√
SpamProbe	[0.003, 0.018] [99.601,+∞]	[0.003, 0.018] [0.795,+∞]	√	

在邮件集 B 下，ROCCH 评估各 SFS 的顺序为 Bogofilter、SpamProbe、Spamassassin 和 CRM114。这与邮件集 A 下的判定结果差不多。

综合上述三种方法的评估结果，论文认为从整体上来说上述六个 SFS 的优劣顺序可以表示为 NC-2>NC-1>{Spamassassin, SpamProbe}>{Bogofilter,CRM114}。{}内部的 SFS 表示它们的过滤效果相差不多，好坏难以确定，需要在实际的情况中作出选择。除此之外，论文还认为由于 Bogofilter 系统在 ROC 曲线及 ROCCH 方法评估上表现最好，它是潜在的最优的 SFS。

6.5 本章小结

本章使用模拟标准邮件集生成系统生成的邮件集 A 和网络中心相关人员收集的邮件集 B，依据第五章设计的评估系统，对六个 SFS 进行了评估。通过计算各 SFS 的十六个评估指标值，部分 SFS 的 ROC 曲线以及采用模糊综合评估方法、因子分析法、ROCCH 方法对它们进行了综合评估，论文给出了它们之间的优劣顺序和各 SFS 自身的特点。

第七章 总结与展望

7.1 论文成果总结

目前在国内外对 SFS 评估的研究都非常少，本论文在紧扣研究目标的基础上，对影响 SFS 评估的几个重要因素展开较为深入的研究，并依据这些研究成果，设计和实现了一个客观的、公正的、全面的反映 SFS 过滤能力的评估系统。论文的主要成果如下所述。

论文首先从已有的参考文献出发，系统化的归纳、总结和提出了用于评估 SFS 系统的指标体系。论文定义了四个基本指标、十二个合成指标和两个基于 ROC 曲线的指标。在指标的计算方法中，论文首次将误报代价之比的概念应用到其中，提出了“归一化”复合矩阵，从而屏蔽了垃圾邮件与正常邮件在重要性上面存在的巨大差异。论文还对各评估指标的使用范围进行了详尽的阐述，明确了基本指标和合成指标可用于评估任何的 SFS 系统，而曲线性指标仅对基于阀值分类的 SFS 系统起评估作用。

为给用户提供综合的评估结果，论文依据建立的评估指标体系，采用模糊综合评估方法、因子分析法和基于 ROC 曲线的 ROCCH 方法来综合评估多个 SFS 的整体优劣顺序。模糊综合评估法以模糊数学理论为理论基础，以基本指标和合成指标为出发点，从 2 个层次上归并评估指标对最终的评估结果的影响。模糊综合评估方法评估的是各 SFS 系统当前配置下的整体过滤效果。因子分析法是从各评估指标的相关性出发，从十六个原始评估指标中提取出较少的几个相互独立的因子变量，并依据这几个因子变量独立或综合的评估 SFS。它对六个开源的 SFS 的综合评估结果与模糊综合评估方法类似。基于 ROCCH 方法的综合评估是评估 SFS 的潜在的最优过滤能力。它以 SFS 的 ROC 曲线为出发点，考虑在不同垃圾邮件与正常邮件比例、不同误报代价比下，SFS 可能的最优的过滤能力。

上述讨论的评估指标和综合评估方法是影响 SFS 评估结果的重要因素。除它们之外，本论文还首次提出标准邮件集和评测训练方法对 SFS 的评估结果有着较大的影响。为验证这些因素的影响，论文在保证其他影响因素一致的情况下，通过实验的方式查看待研究的因素在不同条件下各评估指标的变化情况。实验表明标准邮件集中的总邮件数量，训练邮件集所占的比例，垃圾邮件所占的比例，误报代价比以及邮件的到达顺序等参数都将对 SFS 的评估结果有较大的影响；同样，在不同的评测训练法下进行评估，得到的评估结果也存在较大的差异。为此，本论文认为，为客观、公正的评估各 SFS，就应该使得各 SFS 在相同的条件下进行评估。而为全面反映各个 SFS 的过滤能力，需要实验 SFS 在某一参数的不同大小下的评估结果。

正是考虑到标准邮件集在 SFS 评估中的重要地位，为防止 SFS 对静态标准邮件集的刻意适应，本论文设计和实现了一个模拟的标准邮件集生成系统。该系统可以依据用户的配置参数从大量收集而来的邮件集中动态生成一个模拟的标准邮件集。该模拟标准邮件集将和研究机构公开的标准邮件集一样，可以提供给 SFS 进行评估。

综合上述的研究结果，论文在第五章设计和实现了一个 SFS 的评估系统。该系统接收用户配置的 SFS 系统名、标准邮件集、误报代价比、评测训练方法等参数，并驱动各 SFS 对指定的其他参数进行评估。在第六章，论文给出采用生成的标准邮件集和华东北地区网络中心收集的标准邮件集评估六个 SFS 的评估指标值、综合评估的结果。

7.2 未来工作的展望

依据论文前面所讨论的结果可知，标准邮件集是 SFS 评估研究中的重要组成部分，它对评估结果有着重要的影响。目前的标准邮件集都是由一些研究结构收集和发布的，它们中的垃圾邮件或者是收集某邮件服务器在一段时间内的垃圾邮件、或者是通过 HoneyPot 的方式对外收集用户反馈的垃圾邮件。无论上述哪种方式，所获得的垃圾邮件在邮件内容、邮件的结构上都存在鱼龙混杂、多种多样的特点。这些具有不同邮件内容、邮件结构的垃圾邮件在目前发布的标准邮件集中的地位都是一样。但是在现实中，对大多数用户来说，带有色情、政治、宗教色彩的垃圾邮件要比带有商业目的的垃圾邮件更令人讨厌的多，带有大量图片信息的垃圾邮件也要比仅仅是文字信息的垃圾邮件要浪费用户更多的时间。同样，正常邮件也存在为同样的问题，发给重要客户的邮件显然要比普通客户邮件的重要性要高很多。

为此，若能将标准邮件集按照邮件的内容，邮件的结构进行分类，那么它至少能带来下述两方面的好处。

1) 具有相同类别的邮件由于在邮件内容，邮件结构上比较相近，它们的误报代价也相差无几。为每个小类别设置一个平均误报代价显然要比为仅仅设置垃圾邮件和正常邮件两个平均误报代价要精确的多。这对保证 SFS 评估的公平性有着重要的意义。

2）通过将 SFS 独立的对具有相似内容和结构的邮件集合进行评测，可以发现 SFS 对邮件内容和邮件结构的“偏好”，以此来发现 SFS 自身的特点。

正是由于能够带来上述好处，论文认为需要将标准邮件集按照邮件内容、邮件结构进行进一步的分类，并研究在各个类别中邮件的误报代价；SFS 独立的对各个小类别中的邮件集进行评测。可以综合计算 SFS 对多个小类别的评测结果来反映 SFS 的过滤能力。事实上，在文本分类研究领域，也是将待评测的文本集按照一定的特性分为若干个文本组来进行评测的。在本文中，由于缺乏大量的标准邮件的来源以及作者的精力有限，未能对标准邮件集的分类工作展开研究，故在此做出展望，希望能够给从事相关研究的人员提供一些思路。

除此之外，由于各个邮箱用户之间的习性、爱好和认识的不同，SFS 对各个不同的邮箱用户的过滤结果来进行评估是未来评估研究的一个方向。

参考文献

1 中国互联网协会反垃圾邮件中心 http://www.anti-spam.cn/

2 CERNET 应急响应中心反垃圾邮件小组（CAST） http://www.ccert.edu.cn/spam/index.htm

3 中国反垃圾邮件联盟 http://www.anti-spam.org.cn/

4 CRM114 - the Controllable Regex Mutilator http://crm114.sourceforge.net/

5 Dspam's Home Page http://sourceforge.net/projects/dspam-filter

6 Bogofilter's Home Page http://bogofilter.sourceforge.net/

7 SpamProbe http://sourceforge.net/projects/spamprobe/

8 The spamassassin public mail corpus. http://www.spamassassin.org/publiccorpus

9 SpamBayes http://sourceforge.net/projects/spambayes

10 Corpus download http://iit.demokritos.gr/skel/i-config/downloads/

11 Spamassassin-Corpus http://spamassassin.apache.org/publiccorpus/

12 Quang-Anh.CCERT Data Sets of Chinese Emails (CDSCE) [EB/OL].
http://www.ccert.edu.cn/spam/sa/datasets.htm. 2005.

13 Sahami M, Dumais S, Heckerman D, et al. A Bayesian Approach to Filtering Junk Email[C]. In: Proceedings of AAAI-98 Workshop on Learning for Text Categorization. Madison Wisconsin. May. 1998. 55–62.

14 Harris Drucker, Donghui Wu, Vladimir N Vapnik. Support Vector Machine for Spam Categorization [J]. IEEE TRANSACTIONS ON NEURAL NETWORKS. Sep. 1999. 10(5): 1048-1054

15 Androutsopoulos I, Koutsias J, Chandrinos K, et al. An evaluation of naive Bayesian anti-spam filtering[C].In:G.Potamias, V Moustakis, eds. Proceedings of the workshop on Machine Learning in the New Information Age. Barcelona: 2000.9-17.

16 Zhang Le, Yao Tian-shun. Filtering Junk Mail with A Maximum Entropy Model[c]. In: ICCPOL. ShenYang, China. 446-453

17 Gordon Cormack, Thomas Lynam, A Study of Supervised Spam Detection applied to Eight Months of Personal E-Mail[C]. In: Proceedings of Conference on Email and Anti-Spam (CEAS) 2004, Mountain View, CA, July 30 and 31, 2004.

18 Andrew T, Evangrlos Milios, Nauzer Kalyaniwalla.An Evaluation of Machine Learning

Techniques for Enterprise Spam Filters[R]. http://www.cs.dal.ca/news/def-1156.shtml

19 李星，田莹，段海星。中文垃圾邮件过滤系统的实现和评估[J]。大连理工大学学报，2005，第45卷，增刊，189-195。

20 Dragos D. Margineantu, Thomas G. Dietterich. Bootstrap Methods for the Cost-Sensitive Evaluation of Classifiers

21 Jose Mara, Gomez Hidalgo. Evaluating Cost-Sensitive Unsolicited Bulk Email Categorization

22 Androutsopoulos I, Koutsias J, Chandrinos K, et al. An Experimental Comparison of Naïve Bayes and Keyword-Based Anti-Spam Filtering with Personal E-mail Messages[C].In N.J.Belkin(eds).Proceesing of 23rd Annual Internal ACM SIGIR Conference on Research and Development in Information Retrieval. Greece. July, 2000. P160-167.

23 Foster Provost, Tom Fawcett. Robust Classification for Imprecise Environments. Machine Learning, 42,203-231, 2001

24 宋枫溪，高林.文本分类器性能评估指标.[J]. 计算机工程，第30卷，第13期，2004年7月

25 陈英茂，田嘉禾，耿建华等.ROC 曲线分析及诊断分界点确定程序[J]。中国医学影像技术，2004，20(4)：614-617。

26 Andrew P. Bradley. The Use of the Area under the ROC Curve in the Evaluation of Machine Learning Algorithms [J].Pattern Recognition, 1997, 30:1145-1159

27 Tow Fawcett. ROC Graphs: Notes and Practical Considerations for Researchers [EB/OL] Tech.Report HPL-2003-4, 2003. http://www.purl.org/NET/tfawcett/papers/ROC101.pdf.

28 邹莉玲。ROC 曲线方法及其在医学诊断试验评价中的应用研究[D]：硕士学位论文，南京：东南大学公共卫生学院。

29 郭亚军著. 评估理论与方法[M]. 北京：科学出版社，2002

30 李杰，龚俭.一种基于模糊综合评判的入侵响应效果评估方法[C].见：江苏省计算机学会编，计算机科学技术发展：第二界江苏计算机大会论文集.南京.东南大学出版社.2006.273-278

31 甘早斌，何建国.入侵检测系统的多层次模糊评估研究[J].计算机应用研究.2006.第 4 期，96-99.

32 基于多层模糊综合评判的入侵检测系统报警验证。穆成坡,黄厚宽,田盛丰,北京交通大学计算机与信息技术学院,北京

33 方丽. 城市平派要素研究及实证分析[D]：硕士学位论文,西安，西南交通大学

34 杜青龙 中国城市品牌理论研究与实证分析[D]：硕士学位论文，西安，西南交通大学

35 胡永宏，贺思辉著.综合评估方法[M].北京：科学出版社，2000，53-78

36 余建英，何旭宏著.数据统计分析与 SPSS 应用[M].北京：人民邮电出版社，2003，291-310

37 Tow Fawcett. Evaluating Classifiers for Changing Environment.

http://home.comcast.net/~tom.fawcett/public_html/ROCCH/ROCCH-talk-public.html

38 殷人昆，陶永雷等著.数据结构（用面向对象方法与 C++描述）[M].北京：清华大学出版社，1999.369-372

39 曹振华，赵平，胡跃清著. 概率论与数理统计[M]. 南京：东南大学出版社，2001.

40 gnuplot http://www.gnuplot.info/

41 张智星著。Matlab 程序设计与应用[M]。北京：清华大学出版社，2002

42 徐选,丁伟. 用于邮件过滤的标准样本生成系统研究[J].山东大学学报（理学版）.2006,第 41 卷,第 3 期,85-89.

43 徐激。综合邮件过滤系统的设计与实现[D]：硕士学位论文，南京，东南大学计算机系

致 谢

随着论文的完成，我在学校的日子已接近尾声。回首过去在东南大学的这七年的时光，晃如昨日，感慨万千，历历在目。它是我生命中一段珍贵的，值得回忆的篇章。在过去的七年里，将我从一个单纯、无知的少年培养成一个有一定的独立动手和思考能力的年轻人。我很幸运能结识到这么多的优秀师长和同学，感谢东南大学特别是网络中心对我的教育和培养。它使我积累了丰富的专业知识、思维方式、动手能力和心理素质，同时也使得自己的人生观和价值观有了更深刻的了解和认识。这些都是我宝贵的财富。

本文能够顺利完成，最应该感谢的是我的导师龚俭老师和指导老师丁伟老师。论文的选题、研究到最后的设计、实现、完善，两位老师都给予了悉心的指导。他们严谨的治学态度、一丝不苟的工作作风、广博的知识、丰富的科研经验和正直的做人态度深深影响了我，使我受益匪浅。其次，我要感谢网络中心的曹争老师、程光老师、吴桦老师、吴剑章老师，感谢他们对我的帮助，感谢他们为网络中心营造了美好的科研和生活环境。感谢张黎明、李代强、戈志强等为我提供的帮助，感谢他们为网络中心所做的一切。

此外，我还要感谢在我读研期间所结识的优秀的同学们。感谢网络中心的全体同学，感谢孙美凤、成卫青、彭艳兵、周明中、杨望、陈亮、魏薇、刑苏宵、朱海婷博士，感谢他们在我的学习和工作中热忱的给予了许多指导和帮助。同时还要感谢魏德昊、周渔、徐敏、徐选、戴宣、高亚东、程龙、许春蝶、李杰、吴雄、徐嘉羚、汤晓波、孙毅、薛冠鹏、杨艳、王远等人，他们是我生活上的伙伴，是我学习工作上的良师益友。

同时，我还要感谢我本科的同学们、和我在学生社团 SCDA 一起共事的伙伴们、和我一起参加数模竞赛的战友们以及和我同宿一舍的室友们，和你们的交流开拓了我的眼界，忠心的感谢和祝福你们。

最后，我要感谢我的家人，感谢我的女友张玮，感谢我的父母多年来给我的默默支持，感谢你们对我的抚育、培养和启迪。正是你们多年前的选见才成就了今天的我。感谢我的女友，你一贯的支持是我不竭的动力，感谢你的陪伴、支持和鼓励。

项 涛

2007 年 3 月于东南大学

作者简介

项涛，男，1983 年 10 月出生，籍贯江西分宜

履历：

2000 年 9 月—2004 年 7 月，就读于东南大学数学系信息与计算科学专业，获学士学位。

2004 年 9 月至今于东南大学计算机科学与工程学院计算机系统结构专业攻读硕士学位，从事计算机网络应用方面的研究。研究生就读期间，共完成 18 门课程的学习，总学分 34 分，其中必修学分 20、实践环节 2 学分和选修学分 14。

硕士研究生期间从事的科研工作：

- 2004.7-2004.11 参与国家自然科学基金课题“面向大规模网络的分布式入侵检测和预警”背景下的基于滥用检测的入侵检测和响应系统的开发，负责用户界面的开发。
- 2005.7-2005.8 参与国家教育部 211 项目《CERNET 主干网运行安全基本保障系统》的开发。

发表的论文：

- 项涛，龚俭，丁伟 基于 ROCCH 方法的垃圾邮件过滤系统的评估，计算机科学技术论文展——第二届江苏计算机大会论文集，2006 年 11 月。
- 项涛，龚俭，丁伟 垃圾邮件过滤系统的评估模型研究，计算机工程与设计（已录用）。

硕士期间所获奖励：

- 2005 全国研究生数模竞赛三等奖
- 2006 全国研究生数模竞赛二等奖
- 2005～2006 东南大学优秀硕士二等奖学金