用于邮件过滤的标准样本生成系统研究

徐选,丁伟

（东南大学 计算机科学与工程系 江苏南京 210096）

摘要：由于缺乏标准的中文邮件样本集，无法评测不同垃圾邮件过滤系统的性能。通过分析邮件样本收集过程中存在的各种问题，对生成标准样本涉及的问题进行了深入的研究。同时设计了一个基于真实环境的标准邮件样本生成系统结构，并希望以此推出一个用于衡量邮件过滤系统性能的标准的邮件样本集，最终作为邮件过滤技术研究的基本语料。

关键词：标准样本生成系统;垃圾邮件;邮件过滤;模糊综合评判

中图分类号：TP393　　文献标识码：A

Research on standard sample generation system for email filtering

XU Xuan and DING Wei

（Department of Computer Science, Southeast University, Nanjing ,210096, Jiangsu, China）

Abstract: Lack of standard Chinese mail dataset, the performance of various Spam-filter systems can't be evaluated. The further research on the issue concerning the standard sample generation are made, through analysis the problems on the collection of email samples. Meanwhile, the design of a standard sample generation system applied in real environment is given. A standard email dataset for evaluating the email filter system is provided, and will be finally developed to be a base corpus of email filtering technique.

Keywords: standard sample generation system; Spam; email filtering; fuzzy comprehensive judgment

0 引言

根据中国互联网协会反垃圾邮件中心[1]最新发布的《2005 年第三次中国反垃圾邮件状况调查报告》显示，2005 年 8 月至 2005 年 10 月期间,中国网民平均每周收到垃圾邮件数量为 17.25 封。垃圾邮件占用大量网络资源和浪费用户的大量时间，并常常成为网络病毒传播工具,威胁互联网信息安全,侵害电子邮件用户的合法权益。与此同时，垃圾邮件的过滤技术也在不断地发展，如规则过滤，基于内容过滤等方法。基于英文邮件过滤技术的邮件样本库 Enron[2,4]几乎是唯一公开的基于“真实邮件”的邮件样本集，为英文邮件过滤系统的性能评测提供了评测基础。与已经成熟的面向英文的垃圾邮件过滤系统相比，针对中文垃圾邮件过滤的各种新技术也在不断出现，但邮件过滤系统的过滤性能，如误报率，查全率等等这些指标均是基于自身的实验样本，公布的实验结果很难令人信服。

从第三方的角度，对垃圾邮件过滤系统进行全方位的评价是一件有广泛的实际需求的事情，无论是这类系统的用户，还是开发商，都有这方面的要求。类似的问题在 IR、在 IDS 测试等许多领域都普遍存在。如何使对垃圾邮件过滤系统的评判公正是一个更加复杂的问题，因为它涉及更多的方面。但就像对 IR 系统的评价一样，优质的评判数据是整个评判的基础。在这里，这个评判数据体现为一个普通的邮件的集合，但其中的每封邮件都有明确的性质标识，优质则体现为数据的真实性和标识的正确性。这样的集合也可称为样本。

邮件样本不同于一般的 web 文档可以任意收集，电子邮件作为日常生活中的一种通讯方式，其中包含了很多邮件用户的个人隐私，收集前需要征得用户的同意；并且邮件样本收集存在着垃圾邮件定义争议性，正常邮件收集困难和垃圾邮件时效性等问题，收集存在一定的难度。为了促进垃圾邮件过滤技术的发展，解决目前没有标准的中文垃圾邮件样本集，本文以 CERNET 一个地区中心的自身邮件服务器环境为依托，对邮件样本收集过程中存在的问题进行了深入的研究，并设计了一个能够在邮件服务器的支持下完成收集邮件并标识垃圾邮件

收稿日期：2006-05

作者简介：徐选（1979-），女，硕士研究生，研究方向：邮件过滤

的全过程的系统。

1 邮件样本收集过程中存在的问题

1.1 收集邮件样本过程中存在的非技术问题

1.1.1 法律问题

关于邮件收集主要体现隐私和利益 2 个方面，明确《中国互联网协会公共电子邮件服务(试行)规范》中电子邮件服务商对于客户的电子邮件地址、邮件内容、个人资料负有保密的义务，未经允许不得以任何形式将客户信息提供给第三方的一些相关规定。

1.1.2 垃圾邮件定义问题

垃圾邮件的定义是处理垃圾邮件问题的一个基本的出发点。只有正确定义垃圾邮件，将它与其它邮件区别开来，才能使邮件样本库中答案标准，成为评测的基础。一直以来关于商业性垃圾邮件的界定标准通常有两种：一种是“Opt-out”式，即选择权在网络邮件发送商的手中，在网络用户未提出要求停止发送垃圾邮件的请求之前，网络邮件发送商可以一直向用户发送垃圾邮件；另一种是“Opt-in”式，即选择权在网络用户手中，在未经过网络用户同意之前，任何向用户发送的邮件都是垃圾邮件。目前，我国采用的是后一种方式，所以一封邮件是否归类于垃圾邮件应该由用户评判，而用户参与邮件评判也属于自愿行为，因此不仅需要用户自愿提交邮件，还需要用户在提交邮件时自愿参与邮件的评判。

非技术问题原则上是要以非技术手段来解决，所有的邮件“贡献者”都必须是自愿的，只有这样才能保障样本的质量。但为了打消“贡献者”的顾虑，对技术支持手段的适当告知是必须的，比如，用户地址和非垃圾邮件在被公布前的保护措施等。

1.2 收集邮件样本过程中存在的技术问题

1.2.1 邮件内容安全的技术支持

用户提交的邮件内容不能进行更改，一旦改动，邮件样本则会不真实，从而会导致得不到标准的邮件样本集，也就失去了收集的意义。但电子邮件特别是用户的正常邮件涉及了众多个人信息，如果直接将这些邮件公布，则会使用户个人信息被泄漏。在正常邮件收集之后公布之前，需要对其中涉及用户个人信息的内容部分进行隐私保护处理，且保护后的邮件不能被更改属性，影响样本集的标准。

1.2.2 邮件样本集中答案标准化问题

样本集的标准性一方面是对收集的邮件包括正常邮件和垃圾邮件均来自于真实的邮件环境中，而不是人工合成的数据；另一方面是邮件样本集中的邮件分类标准，即垃圾邮件和正常邮件的区分不是某个邮件过滤系统的结果，而是遵从垃圾邮件的定义，由用户人工评判邮件的类别。样本收集时还要区分对一封邮件的多重评判和用户的评判是否恶意评判。

1.2.3 与邮件服务器的融合问题

作为收集邮件样本的第三方应该能在不考虑原有邮件服务器的具体类型的情况下融合进整个邮件系统，与原有邮件服务器有相同的安全等级，且能够接受“志愿用户”的邮件反馈，并转发给第三方。同时，不论新增的系统运行与否，原邮件服务器的参数设置和运行方式都不应该受到影响。并需考虑在邮件收集方如何保证所有的邮件信息安全，以及采用何种手段限制用户反馈邮件等问题。

2 标准样本生成过程中技术问题的解决方法

2.1 邮件内容的隐私保护方法

邮件中隐私内容包括了用户姓名、邮件地址等以及和用户的具体相关的信息。尽管用户的姓名和数字很容易抽取，如 Vericept[5]能够从邮件中识别出身份证号码，社会安全号码，信用卡号码和其他一些具体的标识，但很显然的是邮件内容涉及自然语言的处理和具体语义，不是简单匹配所能解决，需要用户的背景知识推理。

本文采用的隐私保护方法是建立一个基于 ontology 的模型[6]，对用户包含的一些隐私信息给出正式描述，然后利用信息抽取（information extraction）技术，根据描述从邮件体中提取信息并进行数据替换。信息抽取是核心部分，传统的方法是通过语义知识对邮件内容进行扫描匹配，但邮件中有大量非正式的语言如 HTML 标记等，计算的代价相当大。这里提出了一种优化的信息提取方法。将邮件分割成邮件头和邮件体两部分，再将邮件体分割成用户签名，邮件回复和其他部分三块。这种分块处理的方法既加快了信息抽取的速度又保证了信息替换的准确性。

隐私保护方法步骤如下：

步骤 1：预处理。将邮件分成邮件头和邮件体，从邮件头中提取发信人，收信人，邮件类型、主题等信息。从邮件主题中提取各种缩写词和分词，将关键信息写入用户的背景知识库用于以后邮件体的内容检测。

步骤 2 ：邮件体中用户签名和回复行识别。用户的个性签名中包含了用户设置的各种特殊符号、ASCII图形和用户邮件地址，用户联系方式等个人信息。一般采用H. Chen, J. Hu[7]基于语言学和图形学的两维文本结构的方法识别用户签名，但计算量代价很大。本文采用基于Vitor R. Carvalho [8]特征识别的方法。将每封邮件看成若干个顺序行，通过寻找邮件的最后N行是否有如表1的特征或从上下行推出的特征。对于回复行的识别只需识别是否在每行头部以'>'开头，并且通过上下行是否有这种特征来提高识别性能。从实践上来看，这种方法通过应用签名行和回复行的显著特征，较通过文本块的特征识别更容易实现。

步骤 3 ：用户模型建立和信息替换。信息替换时根据每个用户的知识数据库中的内容语义匹配，进行数据替换和对步骤 2 中识别出的用户个性签名和回复行进行删除。

2.2 样本集答案的标准化

取得用户授权后，已经保证邮件来自于真实环境中。邮件分类采用用户评判需考虑两类问题：一是用户恶意评判邮件混淆邮件分类；二是同一封邮件出现多重评判。

为了防止用户为恶意评判混淆邮件的分类，系统一方面服从用户对邮件的判别，另一方面通过对用户提交邮件的行为打分评价用户。,用户的分数决定是否采用该用户提交的邮件。用户个性化信息大致可分为用户统计信息、用户操作历史、用户访问过的资源特征等[9]。Susan Gauch 等人在文献[10]中总结了 50 个个性化系统,讨论了不同的个性化系统中用户个性化信息的表示方法,如布尔模型、加权关键字向量、语义网络、n-grams 等模型。用户的评价考虑多方面因素，如反馈结果与最终判断结果的差异、用户反馈的频率、用户本身是否是垃圾邮件制造者等等。为这些因素定义相应的数据指标，采用模糊综合评判[11]方法，对用户的反馈效果进行评分。具体步骤如下：

表 1　用户签名和回复行的具体特征

Tab. 1 Complete list of features used for line extraction

行特征描述	当前行	上一行	下一行
是否有空白行	×	×	×
有邮件地址格式	×	×	×
是否为最后一行	×		
是否有 URL 格式	×	×	×
是否有电话号码格式	×		
行以一些特定符号结束，例如正则表达式: "\"$"	×		
上一行或下一行和本行用了同一个符号结尾，例如"\"$"	×		
标点符号（如正则表达式"\p{Punct}"）在行中的比例大于 40%	×	×	×
签名行是否符合以下正则表达式："^[\s]*---*[\s]*$"	×	×	×
行中是否有发信人的姓名，（可以从邮件头得到姓名）	×		
一行中有 10 个或者更多的特殊字符，如以下正则表达式："^[\s]*([*]|#|[\+]|[\^]|-|[\~]|[\&]|[///]|[\$]|_|[\!]|[\/]|[\%]|[\:]|[\=]){10 ,}[\s]*$"	×	×	×
回复行是否有明显标识，如 '>'	×	×	×
以标点符号作为当前行的开头	×	×	×
上一行或下一行和本行用了同一个符号开始	×		
是否以 1 或 2 个标点符号作为行的开头，类似典型的信件回复形式，如以下的正则表达式："^\p{Punct}{1,2}\>"	×	×	×
回复行标识的结尾是否如以下的正则表达式：" wrote:$" or " writes:$"	×	×	×
字母和数字（如正则表达式："\[a-zA-Z0-9]"）的比例小于 80%	×	×	×

步骤 1：确定对用户评分需要考虑的因素集 U

该因素集包括的所有因素反映了一个用户提供反馈的效果，因素集表示为 $U=\{u_1,u_2,\mathbf{L},u_n\}$

步骤 2：评价集合 V

评价集合包含的元素表示希望将用户分为几个等级，如将用户分为可信（10 分），不确定(5 分)以及不可信(0 分)三种，则得到如下评价集：$V=\{V_1,V_2,V_3\}$ 则 V_1=10，表示用户可信；V_2=5，表示用户不确定； V_3=1，表示用户不可信。

步骤 3：建立权重集 A

权重集中的元素对应了因素集中的各个元素，表示了每个因素的权重。如对于上面例子中的因素集，设权重集 $A=\{a_1,a_2,\mathbf{L},a_n\}$，权重集满足 $\sum_{i=1}^{m}a_i=1,a_i\geq 0,(i=1,2,\mathbf{L},m)$，权重集的设定可以用确定隶属度的方法求出，也可以根据具体的因素主观地确定。

步骤 4：多因素模糊评判

记 $B=$

$$(a_1,a_2,\mathbf{L},a_n)\bullet R=(a_1,a_2,\mathbf{L},a_n)\bullet\begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ & \mathbf{L} & \\ r_{n1} & r_{n2} & r_{n3} \end{bmatrix}$$

其中，r_{ij} 表示因素集 U 中元素 u_i 对评价集 V 中元素 v_j 的隶属度。每个用户都对应了一个 R 矩阵。可按实际因素的意义来确定 R 中的元素值。

$B=(b_1,b_2,b_3)$表示了综合评判集，其元素 bi 表示了评判指标，即按不同权重，综合考虑了所有因素之后，该用户为可信，不确定，不可信的隶属度。

步骤 5：确定最后得分 M

则该用户的得分为 $M=\sum_{j=1}^{n}b_jV_j/\sum_{j=1}^{n}b_j$，$M$ 值就是利用模糊综合评判，为该用户的分数。

最后根据用户得分 M 决定以后是否采用该用户提交的邮件。

如果出现多个用户对同一封邮件的多重评判，那么就可以根据用户的评分和该邮件总的评判个数相结合决定评判结果。

2.3 与邮件服务器融合的解决方法

标准样本生成系统相当于“反向SMTP代理”，监听25端口，接收并检查邮件服务器转发的“志愿用户”的邮件。在邮件服务器增设了该生成系统后，一般需要修改对应邮件服务提供商的DNS MX记录，使得原定交由该服务商处理的邮件经SMTP转发到生成系统。如图1所示。因此系统是借助网络结构的重新规划，它的架设不会引发复杂的邮件用户端设置修改。在系统方为用户提供良好的Web支持，并借助Web远程管理服务使用户反馈邮件，如可以让用户在收取邮件时反馈。并采用数据库方式进行用户登录认证和管理，在生成系统最终完成判断并替换正常邮件前，采用SSL/TLS进行加密处理保护用户邮件安全。

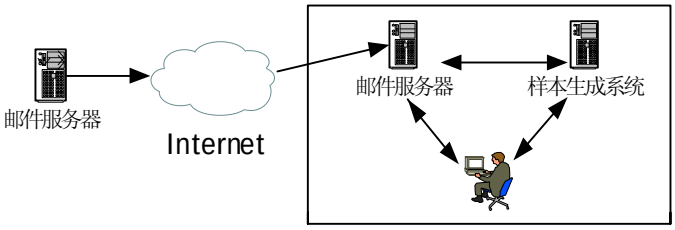

图 1 生成系统与邮件服务器关系图

Fig.1 relation between generation system and mail server

3 邮件样本生成系统的总体结构

针对以上邮件收集过程中各种问题的解决方案和系统的设计方案，我们设计了系统的总体结构。系统可以相对独立地划分成样本添加模块、用户交互模块、用户评分模块、管理员交互模块和邮件隐私保护模块。其总体结构如图 2 所示。

样本添加模块在需要采集邮件的邮件服务器上增设代理，将邮件服务器上的志愿用户的邮件转发到生成系统的样本库中，设置与邮件服务器相同的安全等级；

用户交互模块和管理员交互模块用于系统用户、管理员与系统交互。通过交互界面，用户可以查看、阅读和删除邮件，评判后提交邮件入样本库，查看系统对自己的评分；系统管理员能够查看样本库的状态，邮件总体数量和正常邮件、垃圾邮件数量，用户的评分，用户的行为等信息，并可以通过查看系统日志了解整个生成系统的运行状态和维护生成系统；

用户评分模块则根据系统日志采用综合评判的方法，设定用户的各个因素集和评价集，并计算评价集的权重集，得出用户模糊综合评判值；

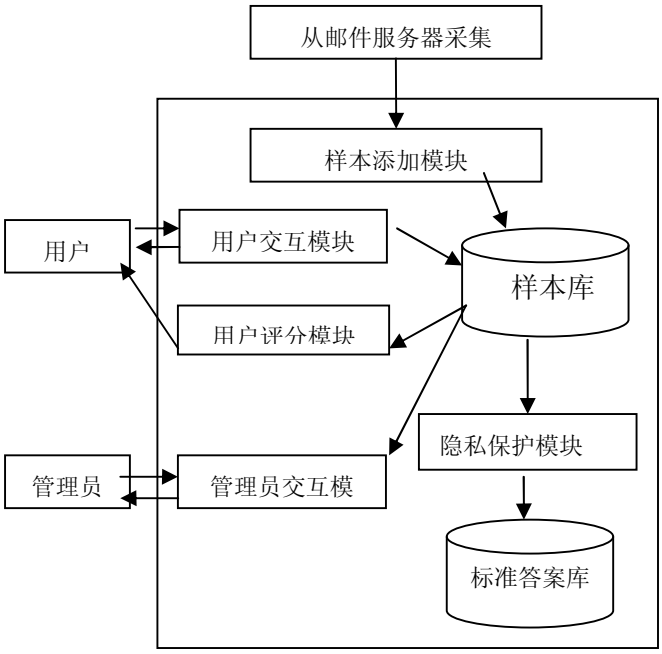

图 2 标准样本生成系统结构

Fig. 2 the Standard sample generation system

邮件隐私保护模块对正常邮件中用户不想公开的邮件内容进行隐私保护，并保证保护后的邮件属性不会改变。隐私保护后的正常邮件和垃圾邮件一起进入系统标准答案库。

4 小结与展望

随着垃圾邮件数量的不断上升，垃圾邮件的过滤技术也在不断地发展。但到目前为止，标准的中文邮件样本集还没有出现，因此也没有针对垃圾邮件过滤系统性能的评测，则急需构建一个标准中文邮件样本集。本文针对收集邮件样本过程中相关的技术和非技术问题，探讨了样本生成过程中涉及问题的解决方法，并设计了标准样本生成系统的总体结构和系统的基本功能模块。系统正在开发过程中，收集了 CERNET 一个地区中心的邮件服务器环境中内部用户提交的邮件生成标准邮件样本集，下一步，我们将会于最近公布首批实验数据。

参考文献：

[1] 李红.中国互联网协会 2006 年第一次反垃圾邮件调查结果发布[EB/OL].http://www.anti-spam.cn,2006-03-22/2006-03-28

[2] Klimt B, Yang Y. Introducing the Enron Corpus [EB/OL]. http://www.ceas.cc/papers-2004/168.pdf,2004-07/2006-03-28

[3] Bekkerman, R., A. McCallum, G Huang, et al. Automatic Categorization of Email into Folders: Benchmark Experiments on Enron and SRI Corpora [EB/OL].http://www.cs.umass.edu/~mccallum/papers/foldering-tr05.pdf, 2004-01/2006-03-28

[4] Jitesh Shetty, Jafat Adibi. The Enron Email Dataset Database Schema and Brief Statistical Report [EB/OL].http://www.isi.edu/~adibi/Enron/Enron_Dataset_Report.pdf, 2004-07/2006-03-28

[5] DENVER, Colo. Vericept Announces General Availability of Version 7.5 of the Vericept 360 Risk Management Platform[EB/OL].http://www.vericept.com,2006-03-15/2006-03-28

[6] Narj`es Boufaden, William Elazmeh, Yimin Ma, et al. PEEP - An Information Extraction based approach for Privacy Protection in Email [EB/OL]. http://www.lib.unb.ca/Texts/PST/2005/pdf/boufaden.pdf,2005-07/2006-03-28

[7] H.Chen, J. Hu, R. Sproat. Integrating geometrical and linguistic analysis for e-mail signature block parsing. [A], ACM Transactions on Information Systems[C]. NY, USA: ACM Press New York, 1999, 17(4):343~366

[8] Vitor R. Carvalho, William W. Cohen. Learning to Extract Signature and Reply Lines from Email [EB/OL]. http://www.cs.cmu.edu/~wcohen/postscript/email-2004.pdf,2004/2006-03-28

[9] Zeng C, Xing CX, Zhou LZ.A survey of personalization technology [J]. Journal of Software, 2002, 13(10):1952~1961

[10] Pretschner A, Gauch S. Personal ontologies for web navigation [M]. NY, USA: ACM Press New York, 2000

[11] 韩立岩,汪培庄. 应用模糊数学[M].北京：首都经济贸易大学出版社，1998