doi:10.3969/j.issn.1000-436x.2013.z2.024

高校误判垃圾邮件自动召回系统的研究与实现

林海卓，王继龙，吴建平，杨家海，徐聪
（清华大学 网络科学与网络空间研究院，北京 100084）

摘 要：垃圾邮件的误判问题一直是反垃圾邮件领域中未能得到根本解决的难点。基于清华大学邮箱系统及反垃圾邮件网关系统进行了一整年的部署和实验（2011 年 9 月至 2012 年 10 月），通过用户对可疑垃圾邮件点击召回的历史行为进行分析，并采用对其感兴趣的垃圾邮件进行文本相似度计算以及关键参数预测的方法来智能化预测用户对当前某一封垃圾邮件的感兴趣程度，即基于用户主观的选择和体验来帮助用户自动召回其可能感兴趣、然而却被反垃圾邮件网关误判的垃圾邮件，解决了传统过滤方法无法杜绝误判的问题。
关键词：隐性索引模型；垃圾邮件；自动召回；向量空间模型；神经网络集成；KNN；贝叶斯
中图分类号：TP393 **文献标识码**：A **文章编号**：1000-436X(2013)Z2-0121-12

Effect of cold-rolling cladding on microstructure and properties of composite aluminum alloy foil

LIN Hai-zhuo, WANG Ji-long, WU Jian-ping, YANG Jia-hai, XU Cong
(Academy of Network Science and Cyber Space, Tsinghua University, Beijing 100084, China)

Abstract: The misjudgement of spam has always been the difficulty in the anti-spam area. Experiments based on Tsinghua university E-mail systems and anti-spam gateway system(Sep,2011–Oct,2012) were deployed and made, analyzing the history of recalling spam behavior of mail users, and using spam text similarity calculation and intelligent key parameters prediction method to predict the user's interest in the current pending spam, which can help users automatically recall their potential interested spams which were misjudged, based on the users' subjective choices and experience, solving the problem that cannot be eliminated by traditional filtering methods.
Key words: LSI; spam; automatic recall; VSM; neural network ensemble; KNN; Bayesian

1 引言

自 1996 年诞生了世界上第一个反垃圾邮件组织 MAPS 以来[1]，通过十多年各国网络安全研究人员的共同努力积累了广泛而成熟的反垃圾邮件技术，包括白名单/黑名单、发送方认证、基于统计学与机器学习的各类智能化过滤方法等。然而，迄今为止，没有一种过滤方法能够彻底杜绝垃圾邮件的误判问题。事实上，在不同的时间，不同的用户对于同一封垃圾邮件的认知和体验差别很大。一部分人认可网关过滤规则的判断，认为是垃圾邮件，而一部分人可能对于这封垃圾邮件的内容和主题非常感兴趣，并不认为这封邮件是垃圾邮件，并希望看到这封邮件。这种现象在高校教师和学生团体中体现得尤为明显，这主要体现在高校学生对求职招聘、网购团购以及电子报刊等方向的广告营销邮件非常感兴趣。然而这部分邮件通常会被各类安全厂商的反垃圾邮件网关拦截并标注为垃圾邮件或可疑邮件。目前，对于这类垃圾邮件误判的问题，工业界诸如网易、Gmail、Hotmail 等国际国内主流邮件运营商尚未有特定的技术手段加以解决。实际上，在“大数据”时代，简单被动的接受用户对“垃圾邮箱”中感兴趣的邮件选择点击找回的方式已经严重过时，传统的方式将不断地被更为智能化的方

收稿日期：2013-09-04
基金项目：国家重点基础研究发展计划（“973”计划）基金资助项目(2009CB320505)
Foundation Item: The National Basic Research Program of China (973 Program) (2009CB320505)

法所替代。本文依托于清华大学邮箱系统及反垃圾邮件网关系统，通过对误判垃圾邮件自动召回系统的研究与实现，基于用户对可疑垃圾邮件点击召回的历史行为进行分析，通过对其感兴趣的垃圾邮件进行文本相似度的计算来预测用户对当前某一封垃圾邮件的感兴趣程度，从而决定是否为用户自动召回这封垃圾邮件, 解决了传统过滤方法无法杜绝误判的问题。

2 垃圾邮件相似度计算框架

根据 MONOSTORI K，ZASLAVSKY A 基于联合文本相似度算法的研究结论，对特定专业背景的文本相似度进行计算，采用多种相似度计算联合求解的方法可以有效地提高计算的精确度[2]。本文将对于垃圾邮件文本相似度的计算划分为基于垃圾邮件主题句的相似度计算以及基于垃圾邮件全文内容的相似度计算。考虑到中文信息处理领域的 3 个研究方向——空间距离计算、语义计算以及隐性语义计算各自在计算方法、理论基础以及在词汇、句子以及段落全文相似度计算中的特点。本文采用了联合向量空间模型/隐性语义索引模型/HowNet 的相似度计算方法，利用各类研究方法的特点，从不同的角度分析垃圾邮件的相似度，并根据用户选择的反馈结果实时地调整参数，从而能够为用户召回基于用户喜好特点的垃圾邮件。本节采用基于向量空间模型以及隐性语义索引模型计算垃圾邮件全文内容的相似度，采用基于 HowNet 的句子相似度计算算法计算垃圾邮件主题句的相似度，并将基于全文的相似度计算结果与基于主题句的相似度计算结果进行加权求和。

$$Spam\_Sim=\alpha_{\mathrm{VSM}}Similarity_{\mathrm{VSM}}+\alpha_{\mathrm{LSI}}Similarity_{\mathrm{LSI}}+\alpha_{\mathrm{HowNet}}Similarity_{\mathrm{HowNet}} \tag{1}$$

$Spam\_Sim$ 为待比较 2 封垃圾邮件的联合相似度，$Similarity_{\mathrm{VSM}}$ 为采用向量空间模型得到的待比较垃圾邮件相似度，α_{VSM} 为向量空间模型相似度计算的权重比例，$Similarity_{\mathrm{LSI}}$ 为采用隐性语义索引模型得到的待比较垃圾邮件相似度，α_{LSI} 为隐性语义索引模型相似度计算的权重比例，$Similarity_{\mathrm{HowNet}}$ 为采用 HowNet 语义相似度计算得到的待比较垃圾邮件相似度，α_{HowNet} 为 HowNet 相似度计算的权重比例。

2.1 基于向量空间模型的邮件文本相似度计算

对于待比较的 2 封垃圾邮件，对二者的全文内容进行相似度的比较。首先要将 2 个全文本进行分词处理，得到相互独立的特征项（K_1，K_2,…，K_n）。待比较垃圾邮件文本集合为（S_1，S_2），对应的垃圾邮件文本特征向量权重采用 TF-IDF 算法。

完整的向量空间模型相似度计算公式为

$$Similarity_{\mathrm{VSM}}=Simila(S_1,S_2)=\frac{\sum_{i=1}^{n}\omega_i(k,s_1)\times\omega_i(k,s_2)}{\sum_{i=1}^{n}\omega_i(k,s_1)^2+\sum_{i=1}^{n}\omega_i(k,s_2)^2-\sum_{i=1}^{n}\omega_i(k,s_1)\times\omega_i(k,s_2)} \tag{2}$$

其中，$\omega_i(k,s_1)$ 为 TF-IDF 的权重比例，表示特征单元 k 在垃圾邮件文本 S_1 中的权重比例。

2.2 基于隐性语义索引模型的邮件相似度计算

美国数学家 DUMAIS S T、FOLTZ P W 以及 PAPADIMITRIOU C H 等人完整地论述了通过隐性语义索引模型对词汇—文本信息矩阵进行奇异值分解，将高维特征向量—文本向量矩阵转换为低维的隐性语义索引空间矩阵，可以极大地保留文本信息中的语义信息含量，并提升文本的检索匹配的效率[3~5]。

邮件文本矩阵奇异值分解定理：设 $\boldsymbol{A}\in R^{p\times q}$ 是 $p\times q$ 垃圾邮件文本实矩阵，取 z 为矩阵 $\boldsymbol{A}$ 的秩，则存在 p 阶左奇异正交矩阵 $\boldsymbol{M}$ 以及 q 阶正交右奇异矩阵 $\boldsymbol{N}$ 使

$$\boldsymbol{M}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{N}=\begin{bmatrix}\Sigma & 0\\ 0 & 0\end{bmatrix} \tag{3}$$

$\boldsymbol{A}=\boldsymbol{M}\begin{bmatrix}\Sigma & 0\\ 0 & 0\end{bmatrix}\boldsymbol{N}^{\mathrm{T}}$ 为矩阵 $\boldsymbol{A}$ 的奇异值分解等式。

隐性语义索引矩阵 $\boldsymbol{A}_i$ 是按照 F—范数评价中，与邮件文本矩阵 $\boldsymbol{A}$ 最为接近的矩阵。隐性语义索引矩阵 $\boldsymbol{A}_i$ 的计算公式为

$$\boldsymbol{A}\xrightarrow{\mathrm{SVD}}\boldsymbol{A}_i=\boldsymbol{X}_1\boldsymbol{H}_1\boldsymbol{Y}_1^{\mathrm{T}} \tag{4}$$

降维系数 i 的确立如下.

对于对角型矩阵 $\Sigma=\mathrm{diag}(\lambda_1,\lambda_2,\cdots,\lambda_p)$，且 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_z\geqslant\lambda_{z+1}=\cdots=0$，则 i 的确定满足下列语义信息累加不等式

$$\sum_1^i\lambda_i/\sum_1^z\lambda_i\geqslant\theta \tag{5}$$

θ 通常取经验值 40%~60%区间。本文实验部分选取 θ 值为 50%。邮件文本之间的相似度依靠“反向”互倒置矩阵之间的乘积来表示，本节采用的垃圾邮件文本—文本相似度计算为

$$Similarity_{\mathrm{LSI}} = \boldsymbol{B} = \boldsymbol{A}_i^{\mathrm{T}} \times \boldsymbol{A}_i = \boldsymbol{Y}_1 \times \boldsymbol{S}_1 \times \boldsymbol{X}_1^{\mathrm{T}} \times \boldsymbol{X}_1 \times \boldsymbol{S}_1 \times \boldsymbol{Y}_1^{\mathrm{T}} = \boldsymbol{Y}_1 \times \boldsymbol{S}_1^2 \times \boldsymbol{Y}^{\mathrm{T}} \quad (6)$$

2.3 基于 HowNet 的邮件文本相似度计算

HowNet（知网）是学者董振东与董强创建的以中文、英文词汇所表达的概念语义为研究对象进行形式化描述的语言信息结构与知识库[6]。HowNet 进行相似度计算的理论基础是将一个词汇表示成若干概念，而每一个概念由若干的义原进行描述。词汇是通过对句子进行分词得到。基于 HowNet 的相似度计算流程可以简单描述如下：1）对待比较的 2 个句子进行分词处理；2）将分词得到的词汇分解成若干概念及义原；3）通过计算义原之间的相似度从而得到 2 个词汇之间的相似度； 4）通过不同词汇之间相似度的加权求和，从而得到句子之间的相似度。

2.3.1 HowNet 义原相似度计算

假设待比较 2 个义原 o_1、o_2 在 HowNet 义原结构树的距离是 D，结构树中义原相似度为 0.5 的平均语义距离为 S，待比较 2 个义原之间的相似度计算为

$$Similarity(O_1, O_2) = S/(S+D) \quad (7)$$

2.3.2 HowNet 概念相似度计算

基于 HowNet 知识描述语言，将垃圾邮件主题句中的实词表示成若干个概念，而这些邮件主题概念由 4 种类型的义原所组成的特征结构来具体描述(基本义原、其他基本义原、关系义原、关系符号)。

2 个垃圾邮件主题句实词概念之间的相似度计算公式为

$$Similarity(C_{1x}, C_{2y}) = \sum_{i=1}^{4} \partial_i \prod_{j=1}^{i} Similarity_j(O_{1x}, O_{2y}) \quad (8)$$

2.3.3 HowNet 词语相似度计算

每个邮件词汇 W 可以分解为若干个邮件主题概念表示，2 个待比较邮件主题词汇的相似度即为二者对应的概念集中、两两邮件主题概念之间相似度的最大值。2 个垃圾邮件主题句中词语之间的相似度计算公式为

$$Similarity(W_1, W_2) = WMAX_{x=1\to N, y=1\to N}\{Similarity(C_{1x}, C_{2y})\} \quad (9)$$

2.3.4 HowNet 垃圾邮件主题句相似度计算

对于 2 个垃圾邮件主题句分词集合按照代词、量词、名词、动词、形容词、数词及虚词划分。对于 7 个垃圾邮件主题句词集合相似度的计算以及加权求和的计算过程如下。

1) 将 2 个垃圾邮件主题句 T_1、T_2 划分成 6 个实词集合以及一个虚词集合，将 2 个词性相同的集合中的每个词分别进行相似度的比较，得到邮件主题句实词集合相似度比较矩阵 $\boldsymbol{Z}$。

2) 计算矩阵 $\boldsymbol{Z}$ 中相似度最大的实词对组合，记录其相似度的值 $ZMAX_i^{\max\{\gamma,\delta\}}\{Similarity(W_{12i}, W_{22j})\}$，并在矩阵 $\boldsymbol{Z}$ 的行列中删掉这对实词。

3) 在剩下的矩阵 $\boldsymbol{Z}$ 中，重复步骤 2)，直到其中的行或列为零。此时多余的实词设定与空实词组成实词对，其相似度为一较小的常数 k，通常设为 0.05。记录总共的实词对数量 $p = \max\{\gamma, \delta\}$。

4) 对上述得到的所有实词对对应的相似度的值进行加和取平均，其值为 2 个实词集合的相似度。

5) 对 6 个垃圾邮件主题句实词集合对 T_{1x}、T_{2x} 和 1 个主题句虚词集合对 T_1'、T_2' 分别进行步骤 2)、步骤 4）的相似度计算流程，得到句子之间的相似度。

完整的垃圾邮件主题句相似度计算公式描述为

$$\begin{aligned}
&Similarity(T_1, T_2) \\
&= \sum_{i=1}^{6} \lambda_i \{\frac{1}{max\{\gamma,\delta\}} \sum_{j=1}^{\max\{\gamma,\delta\}} \{ZMAX_{j,a\to\gamma,b\to\delta}^{\max\{\gamma,\delta\}} Similarity(W_{1ia}, W_{2ib})\}\} + \lambda_7 Similarity(T_1', T_2') \\
&= \sum_{i=1}^{6} \lambda_i \{\frac{1}{\max\{\gamma,\delta\}} \sum_{j=1}^{\max\{\gamma,\delta\}} \{ZMAX_{j,a\to\gamma,b\to\delta}^{\max\{\gamma,\delta\}} \{WMAX_{x=1\to N, y=1\to N}\{Similarity(C_{1iax}, C_{2iby})\}\}\}\} + \lambda_7 Similarity(T_1', T_2') \\
&= \sum_{i=1}^{6} \lambda_i \{\frac{1}{\max\{\gamma,\delta\}} \sum_{j=1}^{\max\{\gamma,\delta\}} \{ZMAX_{j,a\to\gamma,b\to\delta}^{\max\{\gamma,\delta\}} \{WMAX_{x=1\to N, y=1\to N}\{\sum_{q=1}^{4} \partial_q \prod_{p=1}^{q} Similarity_p(C_{1iax}, C_{2iay})\}\}\}\} + \lambda_7 Similarity(T_1', T_2')
\end{aligned} \quad (10)$$

3 误判垃圾邮件自动召回模型与分析方法

本节将分别实现基于动态以及静态的误判垃圾邮件自动召回的流程与框架，参数的定义、选择，智能化的参数预测以及权值调整算法。同时得到基于用户主观选择的垃圾邮件的分类方法和智能化预测用户感兴趣垃圾邮件的方法。

3.1 基于动态分类的误判垃圾邮件自动召回模型

基于动态分类的误判垃圾邮件自动召回模型，根据不同用户，收到可疑垃圾邮件以及召回垃圾邮件的不同，其召回垃圾邮件所分成的类别也不同。同时根据时间段的变化，类别也在动态的发生变化。本文将采用 6 类算法详细介绍动态分类的自动召回模型实现机制。

本文的具体实验将采集清华大学校园网 EQManager 邮件网关管理系统和亚信 Asiainfo Eventlog 邮箱管理系统的邮件文本和日志进行分析。基于清华大学校园网垃圾邮件网关系统的过滤原理，总共可以归纳为四层过滤规则，介绍如下。

1) 红名单放行规则：在其他的邮件系统提供商中，通常定义为白名单。处于 EQManager-Asiainfo 红名单列表的邮件域名和地址，网关将会无条件放行垃圾邮件。

2) 用户强放行规则：针对特定用户对某一邮件域名和地址加入到其信任列表的行为，网关将会记录对应的用户邮件地址并维护其对应的强放行邮件地址列表。即在该用户信任列表中的邮件源发送给该用户的邮件将会放行。

3) 一级垃圾邮件过滤规则：即黑名单过滤规则。列入黑名单的邮件域名和地址将会完全被网关过滤。在本节 7 个算法的详细描述中，已判垃圾邮件表示一级垃圾邮件过滤规则中判定为垃圾邮件而被反垃圾邮件网关过滤掉的邮件。

4) 二级垃圾邮件过滤规则：基于规则列表与过滤策略的二级过滤规则。EQManager-Asiainfo 采用了多种过滤规则并组成规则列表，同时对不同过滤规则设置权值分数，当网关接收到邮件时，将会按照规则列表中的过滤规则逐项打分，如果总得分超过一定阈值 M，则判定为垃圾邮件，即已判垃圾邮件，放入过滤垃圾邮件队列中，此时垃圾邮件不会发送到收件人的任何邮箱中。如果总得分介于阈值 N 与 M 之间（$N<M$），则为可疑垃圾邮件，放入可疑垃圾邮件队列中，此时系统会生成可疑垃圾邮件说明文档并发送到收件人的垃圾邮箱中，说明文档中将会提示收件人该邮件是可疑垃圾邮件，并告知收件人该邮件将会保存 15 天，如果收件人希望收到该邮件，可以点击相关链接，系统则会从网关数据库中将邮件发送到收件人的收件箱中。在本节 7 个算法的详细描述中，已召回垃圾邮件为用户垃圾邮箱中用户主动召回的可疑垃圾邮件。网关待处理可疑垃圾邮件即为网关判定为可疑垃圾邮件，但尚未递交到用户的垃圾邮箱，需要根据用户的历史召回行为进行预测，以判断是否需要自动帮助用户召回该邮件。如果总得分低于阈值 N，则为正常邮件，放入放行邮件队列中，同时将这封邮件发送到收件人的收件箱中。其中阈值 M 与 N 以及逐项过滤规则的权值分数可以采用默认值也可以由系统管理员进行修改、添加和删除，基于用户主动召回垃圾邮件 KNN 分类算法如图 1 所示。

算法 1 基于相似度模型的垃圾邮件 KNN 分类算法

Step1 while $Spam_i$ coming
If i=1 go to step2
If i>=2 go to step3

Step2 $j \leftarrow 1$
创建用户 A 的第 j 个邮件文本集 C_j，将第 i 封邮件放入邮件文本集 j，记为：$Collection(Spam_i) \leftarrow j$
$j \leftarrow j+1$

Step3 Compare($Spam_i$, $Spam_{i-1:i-1}$) 与前 $i-1$ 封召回垃圾邮件依次进行相似度计算，取其中最大的相似度值记为 M_i
If M_i>=k go to step4
Else go to step5

Step4 将第 i 封邮件放入与其邮件相似度最大的一封已召回邮件（第 t 封邮件）所属的邮件文本集中，记为
Collection($Spam_i$) ←Collection($Spam_t$)
Go to step2

Step5 创建用户 A 的第 j 个文本集，并将该邮件放入新建的 C_j 邮件文本集中，记为
$Collection(Spam_i) \leftarrow j$
$j \leftarrow j+1$
NNTrain() go to step2

Step6 Collection_num←$j-1$，记为该用户在时间周期 T 内基于相似度模型得到的已召回垃圾邮件分类数
Collection_c_k 该用户已召回垃圾邮件第 c 个分类集对应的相似度阈值
Collection_c_t 该用户已召回垃圾邮件第 c 个分类集中的召回邮件数量

图 1 基于用户主动召回垃圾邮件 KNN 分类算法

3.1.1 基于相似度模型的垃圾邮件 KNN 分类算法

根据 3 个月的数据,对可疑垃圾邮件总数最高的 100 名用户分别对应的已召回垃圾邮件文本集中的邮件文本两两之间进行相似度计算，并根据动态分类方法，分别将这 100 名用户对应的已召回垃圾邮件文本集进行划分，统计对应于每个用户的已召回垃圾邮件文本的分类数。在相似度计算中，分别采用向量空间模型、隐性索引模型以及 HowNet 相似度计算模型，并根据结果进行求和平均作为最终的垃圾邮件文本相似度，相似度阈值 K 设为 0.5，如图 2 所示。

算法 2　网关待处理可疑垃圾邮件的 KNN 分类判别算法

1　*Compare*($Spam_i$,$Spam_{i-1:i-1}$) 与前 i−1 封召回垃圾邮件依次进行相似度的计算，取其中最大的相似度值即为 M_i

　与其邮件相似度最大的一封已召回邮件（第 t 封邮件）所属的邮件文本集记为 C

2　If M_i< Collection_c_k

End

Else

Recall this email to the user

End

图 2　网关待处理可疑垃圾邮件 KNN 分类判别算法

与采集周期内对用户主动召回垃圾邮件的分类判别相类似，均需要将第 i 封可疑垃圾邮件与前 i−1 封用户主动召回邮件、用户查看的系统自动召回可疑垃圾邮件进行相似度的计算，并取其中最大的相似度值即为 M_i，并记与第 i 封垃圾邮件相似度最大的第 t 封垃圾邮件所属的邮件文本集为 c，将 M_i 与该邮件文本集的召回阈值进行比较，若低于阈值，则不召回，反之将该邮件自动召回。

3.1.2　用户正常反馈的参数调节算法

基于用户反馈的参数调节算法如图 3 所示。

算法 3　用户正常反馈的参数调节算法

Step1　If the user if checking this automatic recalled $Spam_i$ then go to step2 Else, go to step3

Step2　a_{VSM}'← (a_{VSM}+ Sim_{VSM})/ (1 + Sim_{VSM}+ Sim_{LSI}+ Sim_{HowNet})

a_{LSI}' ←(a_{LSI} + Sim_{LSI})/ (1 + Sim_{VSM}+ Sim_{LSI}+ Sim_{HowNet})

a_{HowNet}' ←(a_{HowNet}+ Sim_{HowNet})/ (1 + Sim_{VSM}+ Sim_{LSI}+ Sim_{HowNet})

正反馈参数调节，进一步做参数归一化处理

(a_{VSM}，a_{LSI}，a_{HowNet})←Normal(a_{VSM}'，a_{LSI}'，a_{HowNet}')

Go to step5

Step3　η_{VSM} ←Frsort(Sim_{VSM})

η_{LSI} ←Frsort(Sim_{LSI})

η_{HowNet} ←Frsort(Sim_{HowNet})

计算参数调节变化量

Step4　$\tilde{\alpha}_{VSM}' \leftarrow (\alpha_{VSM}+\eta_{VSM})/(1+\eta_{VSM}+\eta_{LSI}+\eta_{HowNet})$

$\tilde{\alpha}_{LSI}' \leftarrow (\alpha_{LSI}+\eta_{LSI})/(1+\eta_{VSM}+\eta_{LSI}+\eta_{HowNet})$

$\tilde{\alpha}_{HowNet}' \leftarrow (\alpha_{HowNet}+\eta_{HowNet})/(1+\eta_{VSM}+\eta_{LSI}+\eta_{HowNet})$ 负反馈参数调节，进一步做参数归一化处理

$(\alpha_{VSM},\alpha_{LSI},\alpha_{HowNet}) \leftarrow Normal(\tilde{\alpha}_{VSM}',\tilde{\alpha}_{LSI}',\tilde{\alpha}_{HowNet}')$

Go to step6

Step5　c←*Collection*($Spam_i$)

Collection_c_k←(*Collection_c_k***Collection_c_t*+0.5)/ (1+ *Collection_c_t*)

Collection_c_t←1+ *Collection_c_t*

相似度阈值奖惩调节，降低或维持该分类集中邮件的相似度召回阈值，更新分类集中邮件数量

End

Step6　c←*Collection*($Spam_i$)

MIN←MIN{ Sim_{VSM}，Sim_{LSI}，Sim_{HowNet} }

If *Collection_c_k*>=*MIN*

Collection_c_t←1+*Collection_c_t*

Nntrain()

End

Else

Collection_c_k←(*Collection_c_k***Collection_c_t*+*MIN*)/ (1+ *Collection_c_t*)

Collection_c_t←1+ *Collection_c_t*

相似度阈值奖惩调节，降低或维持该分类集中邮件的相似度召回阈值，更新分类集中邮件数量

Nntrain()

End

图 3　用户正常反馈的参数调节算法

数据初始化部分，设定用户对应各个垃圾邮件集的召回垃圾邮件相似度阈值为 0.5，3 个相似度模型的权值比例 α_{VSM}、α_{LSI}、α_{HowNet} 均为 1/3。用户反馈的参数调节可以分为正反馈和负反馈 2 个部分分别进行不同的参数调节流程。当用户对系统自动为其召回的垃圾邮件选择查看时，为用户正反馈，对应的调节方式为相似度阈值奖励调节，即允许 3 个相似度权值的变化可以朝着使总相似度增大的方向进行调节，同时对于该封邮件所属的垃圾邮件文本集对应的邮件召回相似度阈值进行维持或者降低，即与最低召回阈值 0.5 做加权平均。

若用户选择将该邮件删除或者阈值周期内未能查看邮件时，则为用户负反馈。对应的调节方式为相似度阈值惩罚调节，即 3 个相似度权值的变化朝着使总相似度减小的方向进行调节，即对当前垃圾邮件采用 3 种相似度计算模型得到的值进行排序，若某模型得到的相似度值为 3 者中的最大值，则其对应的权值变化调节量为 3 者中最小的相似度值。

3.1.3　基于关键参数预测的邮件相似度计算模型

在可疑垃圾邮件 $Spam_i$ 到来时，需要对当前 3 个相似度模型的比例权重进行计算并根据相似度的值进行加权求和得到综合相似度的值，并根据这一相似度的值来进行判断是否召回该可疑垃圾邮件。只有当用户对当前系统为其自动召回垃圾邮件进行反馈时，才能够调整得到对应 $Spam_i$ 邮件的当前 3 个相似度模型的比例权重 α_{VSM_i}、α_{LSI_i}、α_{HowNet_i}，这种逻辑上的前后矛盾使如何对于当前 $Spam_i$ 的 3 个相似度模型的比例权重进行预测成为决策是否召回可疑垃圾邮件的关键。本节将研究对于当前 3 个相似度模型之间权重的预测策略。同时为了能够进一步提升神经网络集成模型预测的精度，将 3 个权重的历史数据对应的时间参数也作为神经网络集成模型预测过程中的训练输入参量。参数定义如图 4 所示。

当第 i 个可疑垃圾邮件 $Spam_i$ 到来时，首先对最新相邻反馈时间间隔进行预测，根据输入序列 θ_{VSM_i-2}，θ_{VSM_i-3}，θ_{VSM_i-4})得到用户对第 i-1 封自动召回邮件的反馈时间与对第 i 封自动召回邮件反馈时间的时间间隔 θ_{VSM_i-1}'，由于计算的即时性，$\theta_{VSM\_i-1}' = \theta_{LSI\_i-1}' = \theta_{HowNet\_i-1}'$。在预测过程中，通

过最近 3 次的反馈时间间隔来对不同相似度模型对应的最近 3 次相似度模型权值进行权值比例的二次限定，即用户对第 i–1 封邮件反馈与第 i 封邮件反馈的时间间隔，对第 i–2 封邮件反馈与第 i 封邮件反馈的时间间隔以及对第 i–3 封邮件反馈与第 i 封邮件反馈的时间间隔进行归一化的权值处理，使参数满足 $\phi_{\mathrm{VSM}\_i-1}'+\phi_{\mathrm{VSM}\_i-2}'+\phi_{\mathrm{VSM}\_i-3}'=1$，其中 ϕ_{VSM_i-1}'、ϕ_{VSM_i-2}'、ϕ_{VSM_i-3}' 在神经网络的预测中与 $\alpha_{\mathrm{VSM}\_i-1}$、$\alpha_{\mathrm{VSM}\_i-2}$、$\alpha_{\mathrm{VSM}\_i-3}$ 共同作为输入序列，并分别代表 $\alpha_{\mathrm{VSM}\_i-1}$、$\alpha_{\mathrm{VSM}\_i-2}$、$\alpha_{\mathrm{VSM}\_i-3}$ 在预测 α_{VSM_i} 中的时间重要性权重，其含义为：与当前时间点越近的用户主观反馈，其对应的反馈参数信息重要性权重越大。因此 ϕ_{VSM_i-1}'、ϕ_{VSM_i-2}'、ϕ_{VSM_i-3}' 之间的大小关系为：$\phi_{\mathrm{VSM}\_i-1}'>\phi_{\mathrm{VSM}\_i-2}'>\phi_{\mathrm{VSM}\_i-3}'$。

算法 4 关键参数预测的邮件相似度计算模型

Step1 $\theta_{\mathrm{VSM}\_i-1}'\leftarrow NN\_Predict(\theta_{\mathrm{VSM}\_i-2},\theta_{\mathrm{VSM}\_i-3},\theta_{\mathrm{VSM}\_i-4})$

$\theta_{\mathrm{LSI}\_i-1}'\leftarrow NN\_Predict(\theta_{\mathrm{LSI}\_i-2},\theta_{\mathrm{LSI}\_i-3},\theta_{\mathrm{LSI}\_i-4})$

$\theta_{\mathrm{HowNet}\_i-1}'\leftarrow NN\_Predict(\theta_{\mathrm{HowNet}\_i-2},\theta_{\mathrm{HowNet}\_i-3},\theta_{\mathrm{HowNet}\_i-4})$

Step2 $\bar{\theta}\leftarrow 1/\theta_{\mathrm{VSM}\_i-1}'+1/(\theta_{\mathrm{VSM}\_i-1}'+\theta_{\mathrm{VSM}\_i-2})+1/(\theta_{\mathrm{VSM}\_i-1}'+\theta_{\mathrm{VSM}\_i-2}+\theta_{\mathrm{VSM}\_i-3})$

$\phi_{\mathrm{VSM}\_i-1}'\leftarrow 1/(\theta_{\mathrm{VSM}\_i-1}\times\bar{\theta})$

$\phi_{\mathrm{VSM}\_i-2}'\leftarrow 1/((\theta_{\mathrm{VSM}\_i-1}'+\theta_{\mathrm{VSM}\_i-2})\times\bar{\theta}))$

$\phi_{\mathrm{VSM}\_i-3}'\leftarrow 1/((\theta_{\mathrm{VSM}\_i-1}'+\theta_{\mathrm{VSM}\_i-2}+\theta_{\mathrm{VSM}\_i-3})\times\bar{\theta}))$

同理可得

ϕ_{LSI_i-1}'、ϕ_{LSI_i-2}'、ϕ_{LSI_i-3}'、$\phi_{\mathrm{HowNet}\_i-1}'$、$\phi_{\mathrm{HowNet}\_i-2}'$、$\phi_{\mathrm{HowNet}\_i-3}'$

Step3

$\alpha_{\mathrm{VSM}\_i}'\leftarrow NN\_Predict(\alpha_{\mathrm{VSM}\_i-1},\alpha_{\mathrm{VSM}\_i-2},\alpha_{\mathrm{VSM}\_i-3},\phi_{\mathrm{VSM}\_i-1}',\phi_{\mathrm{VSM}\_i-2}',\phi_{\mathrm{VSM}\_i-3}')$

$\alpha_{\mathrm{LSI}\_i}'\leftarrow NN\_Predict(\alpha_{\mathrm{LSI}\_i-1},\alpha_{\mathrm{LSI}\_i-2},\alpha_{\mathrm{LSI}\_i-3},\phi_{\mathrm{LSI}\_i-1}',\phi_{\mathrm{LSI}\_i-2}',\phi_{\mathrm{LSI}\_i-3}')$

$\alpha_{\mathrm{VSM}\_i}'\leftarrow NN\_Predict(\alpha_{\mathrm{HowNet}\_i-1},\alpha_{\mathrm{HowNet}\_i-2},\alpha_{\mathrm{HowNet}\_i-3},\phi_{\mathrm{HowNet}\_i-1}',\phi_{\mathrm{HowNet}\_i-2}',\phi_{\mathrm{HowNet}\_i-3}')$

Step4 通过 $\alpha_{\mathrm{VSM}\_i}'\times Sim_{\mathrm{VSM}\_i}+\alpha_{\mathrm{LSI}\_i}'\times Sim_{\mathrm{LSI}\_i}+\alpha_{\mathrm{HowNet}\_i}'\times Sim_{\mathrm{HowNet}\_i}$ 计算当前待处理垃圾邮件的相似度，并通过 UClassify 函数进行进一步计算

图 4 关键参数预测的邮件相似度计算模型参数定义

3.1.4 用户延迟反馈的参数调节算法

相似度 *Sim* /相似度模型阈值 α /阈值时间间隔 θ 到达时间甘特图描述如图 5 所示。

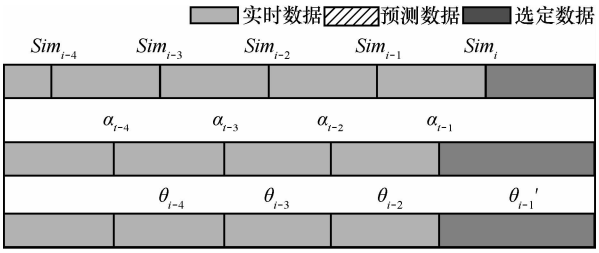


图 5 相似度 *Sim* /相似度模型阈值 α /阈值时间间隔 θ 到达时间甘特图_1

在本例中，可以看到相似度 *Sim* 、相似度模型阈值 α 、阈值时间间隔 θ 之间时间上的对应关系。当第 i 封可疑垃圾邮件 $Spam_i$ 到来时，由于用户尚未有对其进行反馈，因此 α_i 未知，对应的第 i–1 封邮件反馈与第 i 封邮件反馈的时间间隔 θ_{i-1} 未知，在算法 3 中需要首先对其进行预测，进而量化得到 ϕ_{VSM_i-1}'、ϕ_{VSM_i-2}'、ϕ_{VSM_i-3}'，并与 α_{i-1}、α_{i-2}、α_{i-3} 共同作为神经网络的输入参量，预测得到 α_i 的值。下面根据 2 个特例图 6、图 7 总结用户延迟反馈的参数调节策略，如图 8 所示。

在本例中，第 i–2 封和第 i–1 封可疑垃圾邮件到来时，用户尚未对第 i–3 封和第 i–2 封系统召回的可疑邮件进行反馈，第 i–3 封邮件与第 i–2 封邮件到达时间的间距小于阈值 $\tilde{T}_{sim}$，在对 α_{i-2} 进行预测时，采用默认用户正反馈调节 α_{i-3}' 作为第 i–3 封可疑邮件对应的相似度模型权值。当用户对第 i–3 封可疑垃圾邮件的反馈 α_{i-3} 到来时，计算用户对第 i–4 封可疑垃圾邮件与第 i–3 封可疑垃圾邮件到来的时间间隔 θ_{i-4}，如果小于阈值 $\tilde{T}_{\alpha}$，则在后续用到 θ_{i-4}、α_{i-3} 的预测过程中，采用最新反馈的 θ_{i-4}、α_{i-3} 代替 θ_{i-4}'、α_{i-3}' 进行预测。

在本例中，第 i–2 封和第 i–1 封可疑垃圾邮件到来时，用户尚未对第 i–3 封和第 i–2 封系统召回的可疑邮件进行反馈，第 i–3 封邮件与第 i–2 封邮件到达时间的间距大于阈值 $\tilde{T}_{sim}$，在对 α_{i-2} 进行预测时，采用默认用户负反馈调节 $\tilde{\alpha}_{i-3}'$ 作为第 i–3 封可疑邮件对应的相似度模型权值。当用户对第 i–3 封可疑垃圾邮件的反馈 α_{i-3} 到来时，计算用户对第 i–4 封可疑垃圾邮件与第 i–3 封可疑垃圾邮件到来的时间间隔 θ_{i-4}，如果大于阈值 $\tilde{T}_{\alpha}$，则在后续用到 θ_{i-4}、α_{i-3} 的预测过程中，仍采用 θ_{i-4}'、α_{i-3}' 进行预测。

用户延迟反馈参数调节策略的一般描述为

$$\tilde{T}_{sim}=\frac{1}{i-1}\sum_{i=1}^{i-1}T_{sim}\ ,\quad \tilde{T}_{\alpha}=\frac{1}{k-1}\sum_{k=1}^{k-1}T_{\alpha}$$

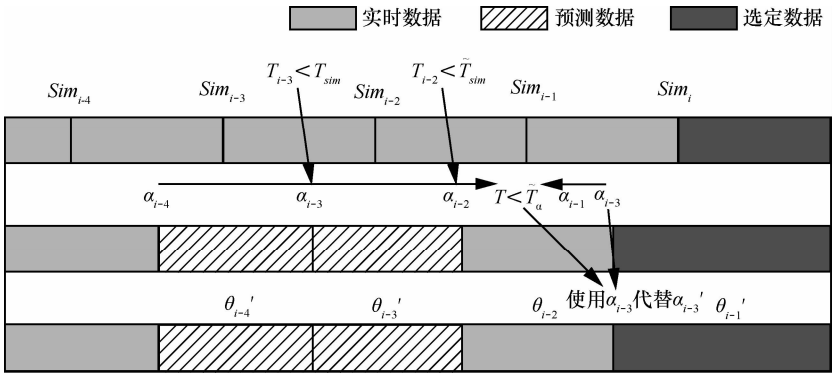


图 6　相似度 *Sim* /相似度模型阈值 α /阈值时间间隔 θ 到达时间甘特图_2

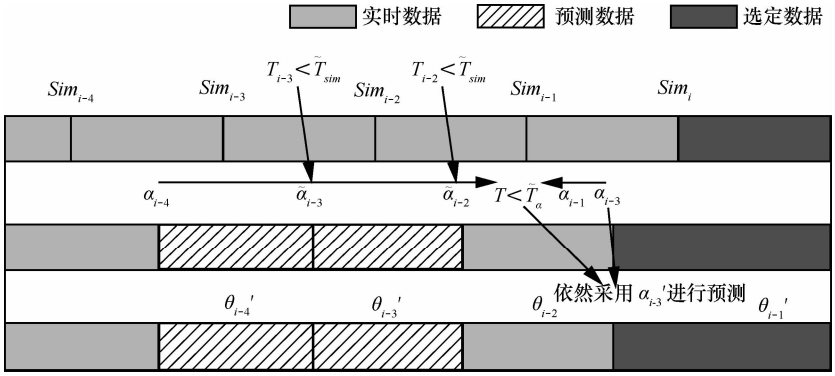


图 7　相似度 *Sim* /相似度模型阈值 α /阈值时间间隔 θ 到达时间甘特图_3

算法 5　用户延迟反馈的参数调节算法

IF 当第 *i* 封自动召回垃圾邮件到来时，用户尚未对第 *i*−1 封自动召回邮件进行反馈

　　IF 第 *i*−1 封自动召回垃圾邮件与第 *i* 封自动找回垃圾邮件到来的时间间隔不超过阈值 $\tilde{T}_{sim}$

　　THEN 在对 α_i 预测时，默认用户对第 *i*−1 封自动召回邮件进行了正反馈，其结果为 α_{i-1}'，同时用户对第 *i*−2 封自动召回垃圾邮件反馈的时间与第 *i*−1 封垃圾邮件反馈的时间间隔由预测值 θ_{i-2}' 替代

　　ELSE

　　THEN 在对 α_i 预测时，默认用户对第 *i*−1 封自动召回邮件进行了负反馈，其结果为 $\tilde{\alpha}_{i-1}'$，同时用户对第 *i*−2 封自动召回垃圾邮件反馈的时间与第 *i*−1 封垃圾邮件反馈的时间间隔由预测值 θ_{i-2}' 替代

END

IF 当用户对第 *k* 封(*i*−4<*k*<*i*)自动找回邮件的反馈到来时，第 *i* 封自动召回垃圾邮件已经到来

　　IF 用户对第 *k*−1 封自动召回邮件反馈的时间与对第 *k* 封自动召回邮件反馈的时间间隔不超过阈值 $\tilde{T}_\alpha$

　　THEN 其后需要用到 α_k 、θ_{k-1} 的预测中，使用用户更新的反馈结果和 θ_{k-1} 值将原值 $\alpha_k'/\tilde{\alpha}_k'$ 、 θ_{k-1}' 替换

　　ELSE

　　THEN 采用 α 默认值 $\alpha_k'/\tilde{\alpha}_k'$， θ 的预测值 θ_{k-1}' 进行下一阶段的预测

END

图 8　用户延迟反馈的参数调节策略

算法 6　基于采集窗口滑动的过期邮件删除算法

FExceed()

Step1　IF *Oldest_Spam_Circle*>*T*

　　Collection_c_order ←*Collection_c_order*−1

　　c ←*Oldest_Spam_c*

　　Collection_c_t ←*Collection_c_t*−1

Step2　IF *Collection_c_t* =0

THEN While(*i*←(*c*+1): (*Collection_c_t* +1))

Collection_c←*Collection_c*+1

　　Collection_c_num ←*Collection_c_num*−1

　　Collection_c_t←*Collection_c_t*−1

Step3　*Oldest_Spam*←*Oldest*()

图 9　基于采集窗口滑动的邮件筛选算法

3.1.5　基于采集窗口滑动的邮件筛选算法

对于当前用户召回垃圾邮件集中最早召回的邮件，当其生存周期大于采集窗口周期长度时，将其删除，同时将对应的垃圾邮件的序号、所属垃圾邮件类别集中的邮件按时间顺序重新进行排列，若被删除的垃圾邮件是其所属邮件类别中唯一的垃圾邮件，则删除该封邮件所在的邮件类别，如图 9 所示。

3.1.6 动态分类邮件自动召回模型综合计算框架

在前期采集与神经网络训练阶段，根据用户主动召回垃圾邮件分类算法 *Classify*()对不同的用户建立其对应的垃圾邮件类别集以及不同的垃圾邮件类别集中各自的垃圾邮件文本，根据周期 T 内用户的相似度模型权值 α 以及其对应的时间参数权重 θ 的历史数据按照算法 4 的预测模型进行训练，如图 10 所示。

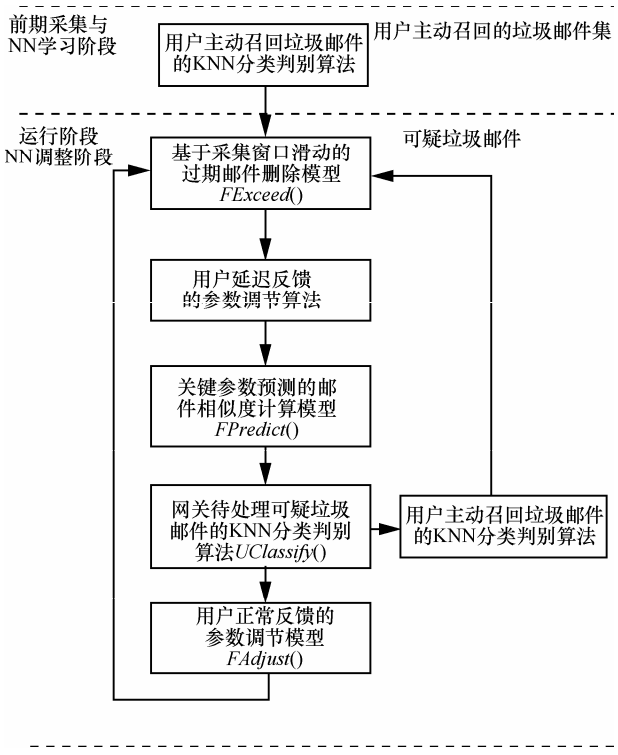


图 10 动态分类邮件自动召回模型综合计算框架

在实时运行与神经网络调整阶段，首先基于采集窗口滑动的过期邮件删除模型 *FExceed*()将超过采集时间窗口的过期数据删除，接着在进行实时参数预测之前，需要对部分用户延迟反馈的特殊情况进行参数调整处理，即采用用户延迟反馈的参数调整算法。然后通过关键参数预测的邮件相似度计算模型 *FPredict*()，通过神经网络输入参量 ϕ_{i-1}'、ϕ_{i-2}'、ϕ_{i-3}'、α_{i-1}、α_{i-2}、α_{i-3} 共同作为输入序列，预测当前待反馈邮件对应的 α_i 值，并得到邮件相似度的计算模型。其后通过网关待处理可疑垃圾邮件分类判别算法 *UClassify*()判断是否将该邮件召回。对于自动为用户召回的可疑垃圾邮件，通过用户对该邮件的主观反馈进一步调整相应的相似度模型权重，即基于用户反馈的参数调整模型 *FAdjust*()。用户反馈的参数调节可以分为正反馈和负反馈 2 个部分，分别进行不同的参数调节流程。当用户对系统自动为其召回的垃圾邮件选择查看时，为用户正反馈，对应的调节方式为相似度阈值奖励调节；若用户选择将该邮件删除或者阈值周期内未能查看邮件时，则为用户负反馈，对应的调节方式为相似度阈值惩罚调节。同时，对于未自动召回的可疑垃圾邮件，若用户主观选择召回，则依然按照用户主动召回垃圾邮件分类算法 *Classify*)处理。

3.2 基于静态分类的误判垃圾邮件自动召回模型

与动态分类的误判垃圾邮件自动召回模型相比，对于不同用户对应不同的可疑垃圾邮件数量和内容，均会将其召回垃圾邮件划分成固定的 H 个类别，同时 H 个类别也不会随时间段的变化而变化。

贝叶斯分类是目前模式识别、数据挖掘领域中应用最为广泛的技术，在包括搜索引擎、电子商务、IM、电子邮件以及密码学等领域有着广泛的应用,是统计学领域中，以内容为中心进行智能化分类的数学方法。本节进一步将清华大学校园网垃圾邮件划分为 5 个主题：广告营销类、宗教政治类、学术论文类、求职招聘类以及其他类。选取 2011 年 9 月至 2011 年 11 月 3 个月的垃圾邮件 750 封进行贝叶斯分类训练，获得已知 5 类邮件主题集所对应的各个文本关键词出现的先验概率。750 封垃圾邮件中广告营销类、宗教政治类、学术论文类、求职招聘类、其他类邮件各 150 封。由于贝叶斯分类在理论上要求训练集无限大并且保证所有特征项之间相互独立，这与实际问题研究背景不符，从而影响最后的分类结果。在下一节中，会进一步通过实验比较静态分类和动态分类误判垃圾邮件自动召回技术的实验结果。

4 基于神经网络集成系统的关键参数预测模型

HANSEN 和 SALAMON 在 1990 年率先提出了神经网络集成的研究方法，并从权值初始化、算法复杂性和神经元灵敏度的角度，在理论上证明了通过对于缺乏先验信息的问题采用相互独立的、能改善不同的系统性能指标的多种神经网络进行训练并将结果进行合成，较之单一使用其某一个神经网络进行预测，在相同训练周期的设定下可以显著地提高神经网络逼近任何复杂函数的精度。增加神经网络系统稳定性同时提升神经网络系统的泛化能力[7~9]。

本节将基于神经网络系统的 5 类主要的性能指标：泛化能力、稳定性、顽健性、容错性、收敛能

力为基础设计 6 类不同的神经网络——BP 神经网络、遗传神经网络、PID 神经网络、小波神经网络、混沌神经网络以及灰色神经网络，并根据待预测的误判垃圾邮件自动召回模型的关键参数相似度模型权值 α 与用户反馈时间间隔 θ 建立完整的神经网络集成系统。同时横向比较动态分类与静态分类，采用单一神经网络与神经网络集成，考虑反馈时间间隔 θ 与不考虑反馈时间间隔 θ ，采用对 α_t 进行预测与直接使用 α_{t-1} 进行相似度计算的垃圾邮件自动召回模型最后的实验结果。

多神经网络系统组成的神经网络集成结构描述如图 11 所示。

4.1　基于动态分类邮件相似度分布与用户反馈结果

在动态分类的神经网络集成系统预测的训练过程中，原始参照文本集选取 2011 年 9 月至 2011 年 11 月 EQManager 邮件网关管理系统与亚信 Asiainfo Eventlog 邮箱管理系统中用户主动召回的垃圾邮件文本。训练过程中的垃圾邮件数据选取清华大学校园网 2011 年 12 月至 2012 年 2 月的可疑垃圾邮件文本数据，依据 5.2.2 节训练集生成办法生成垃圾邮件文本训练集，并根据 3 个月数据对神经网络集成系统的训练对未来 3 个月 2012 年 3 月至 2012 年 5 月对应的 90 位用户进行参数预测，为用户自动召回其潜在需求的可疑垃圾邮件。

本节选取为用户自动召回最多可疑垃圾邮件的前 12 位用户，计算为每位用户自动召回垃圾邮件的相似度值及垃圾邮件类别划分如图 12 所示。其中雷达图所平分的扇区代表为该用户自动召回垃圾邮件共划分的类别，半径代表该垃圾邮件的相似度值。黑色圆点代表用户正反馈的已召回可疑邮件，灰色圆点代表用户负反馈的已召回可疑邮件。

4.2　动态与静态分类误判垃圾邮件召回模型实验对比

统计基于动态与静态分类的误判垃圾邮件召回模型的实验结果，分别按照 To24 用户正反馈率、Top80 用户正反馈率、Top24 负反馈相似度值域、Top80 负反馈相似度值域统计如表 1 所示。

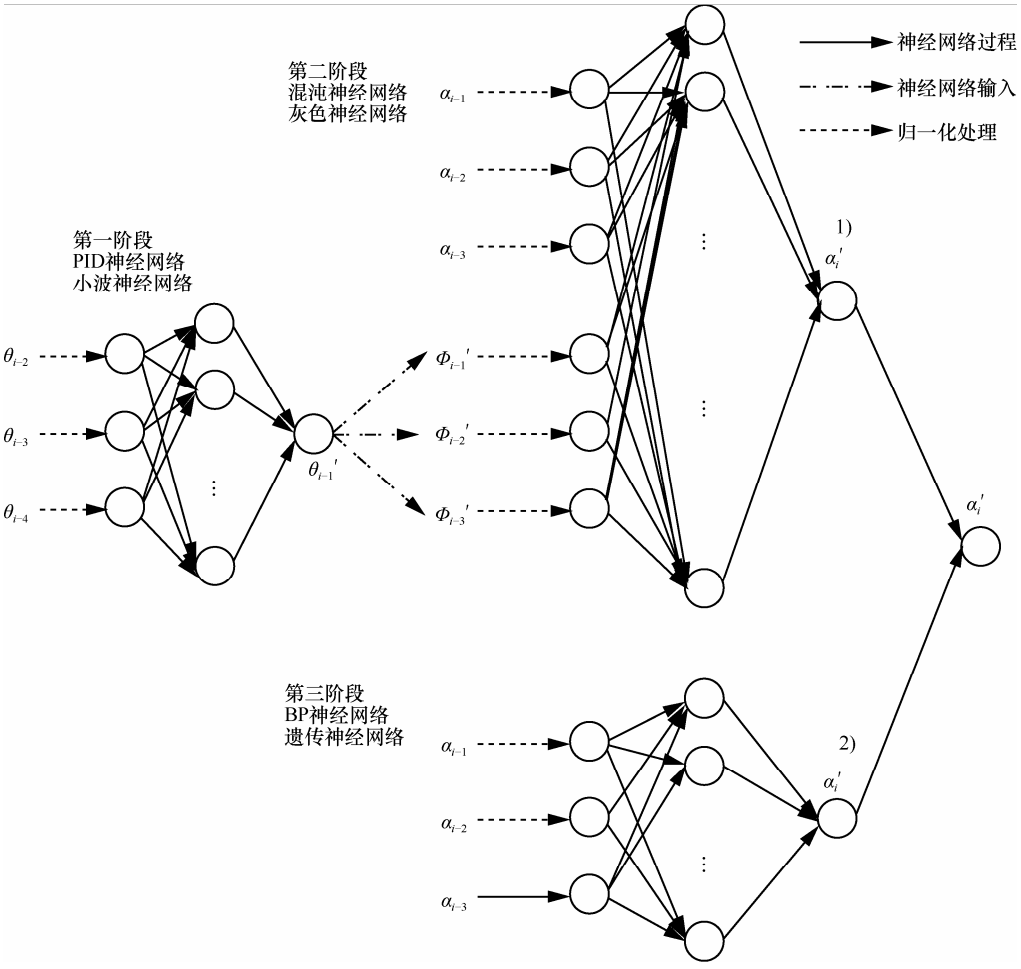


图 11　关键参数预测的神经网络集成系统结构

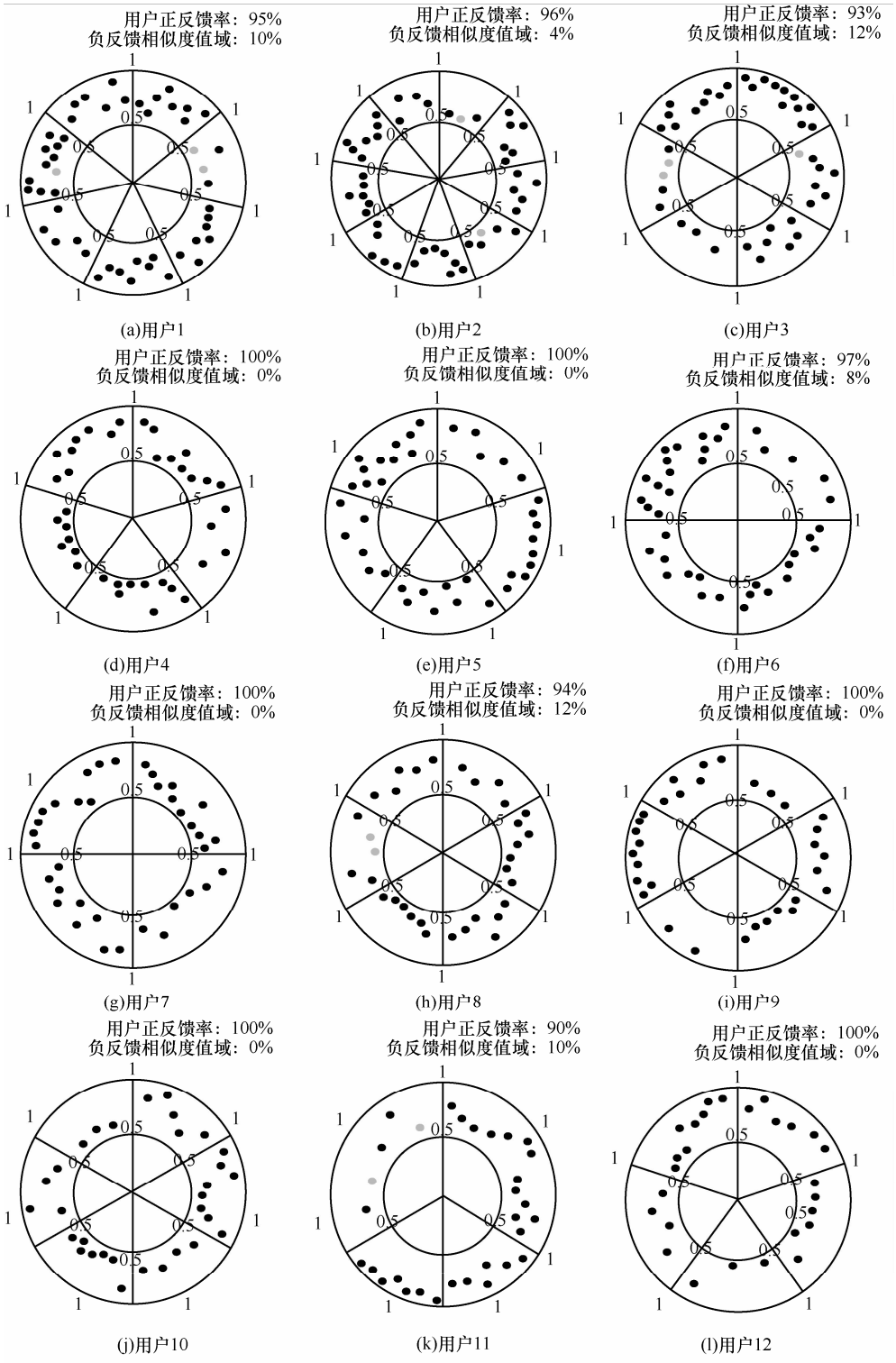


图 12　动态分类-相似度值及垃圾邮件类别划分雷达图

表 1　　动态/静态分类相似度值域比较

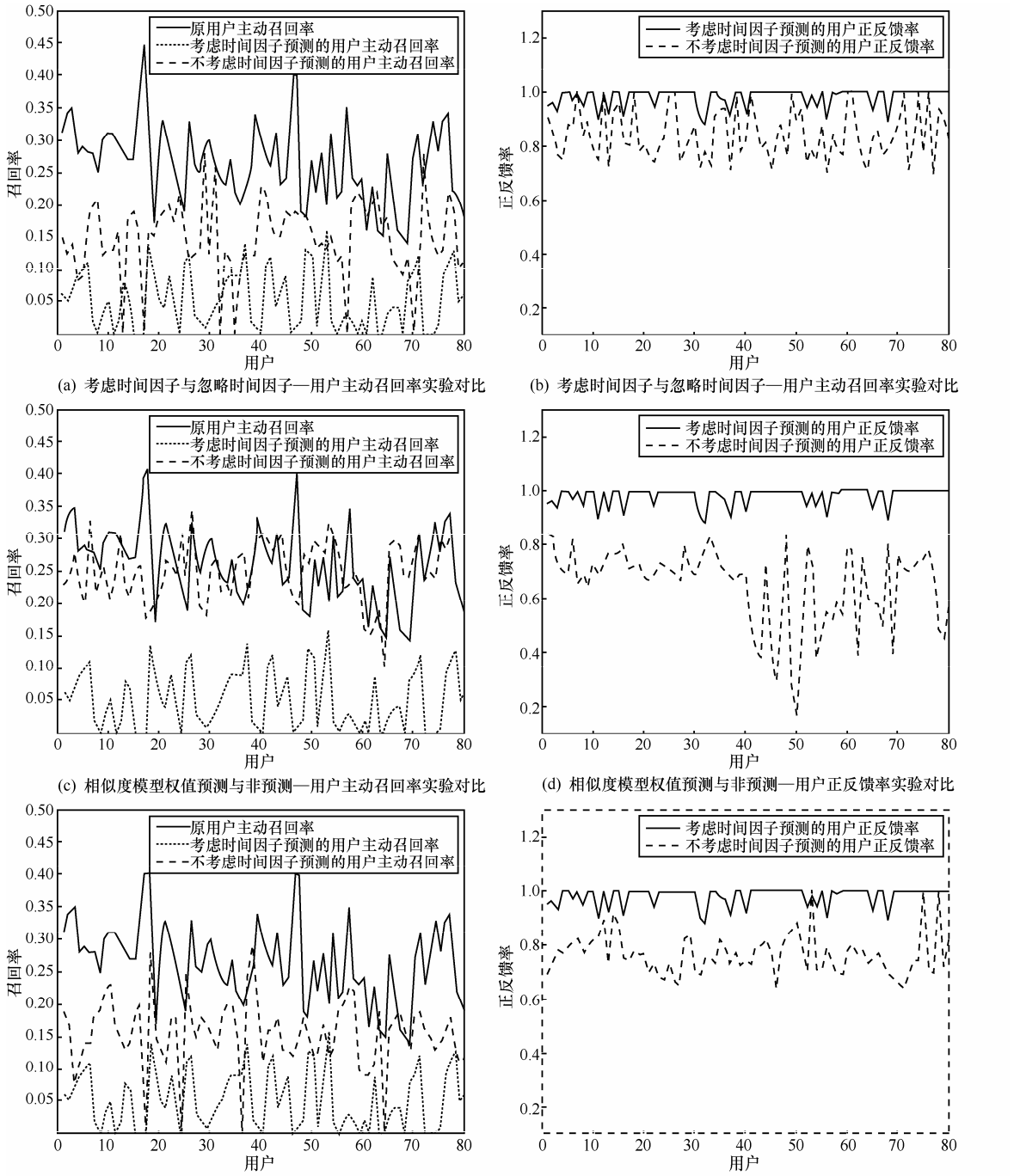
	动态分数	静态分类
Top24 用户平均正反馈率	0.975 8	0.788 8
Top80 用户平均正反馈率	0.982 0	0.786 4
Top24 负反馈相似度值域	0.034 6	0.488 3
Top80 负反馈相似度值域	0.031 2	0.512 8

基于上述数据统计，基于动态分类的用户正反馈率远高于基于静态分类的用户正反馈率，同时基于动态分类的负反馈相似度值域集中在 4%以内（3.12%），与平均负反馈率（1.8%）相比，误差仅有 1.32%，具有统计学上的分类意义。而

基于静态分类的负反馈相似度值域约为 51.28%，与平均负反馈率（21.36%）相比，误差为 29.92%，不具有统计学上的分类意义。同时，基于动态分类的用户主动召回率也远低于基于静态分类的用户主动召回率。

4.3　用户正反馈率/用户主动召回率对比

选取清华大学校园网 2012 年 5 月至 2012 年 7 月 EQManager 邮件网关管理系统与亚信 Asiainfo Eventlog 邮箱管理系统的可疑邮件文本数据，在对本次相似度值 α_i' 的预测中不考虑时间因子 ϕ_{i-1}'、ϕ_{i-2}'、ϕ_{i-3}' 的预测，为用户自动召回 2012 年 8 月其潜在需求的可疑垃圾邮件。在相似度模型权值预测与非预测实验对比中，不采用权值预测的方法，而是采用上一次计算得到的相似度值 α_{i-1} 代替 α_i' 进行相似度的计算，为用户自动召回 2012 年 9 月其潜在需求的可疑垃圾邮件。在神经网络集成系统



(a) 考虑时间因子与忽略时间因子—用户主动召回率实验对比　(b) 考虑时间因子与忽略时间因子—用户主动召回率实验对比

(c) 相似度模型权值预测与非预测—用户主动召回率实验对比　(d) 相似度模型权值预测与非预测—用户正反馈率实验对比

(e) 神经网络集成系统与单一神经网络—用户主动召回率实验对比　(f) 神经网络集成系统与单一神经网络实验对比—用户正反馈率实验对比

图 13　用户正反馈率/用户主动召回率对比

与单一神经网络实验对比中，不采用神经网络集成系统的预测方法，而是采用单一 BP 神经网络对本次相似度值α_i'的预测，为用户自动召回 2012 年 10 月其潜在需求的可疑垃圾邮件。最后计算用户主动召回率以及正反馈率如图 13 所示。

基于数据统计，无论是用户正反馈率还是用户主动召回率，考虑时间因子预测的神经网络集成系统以及考虑相似度模型权值预测的垃圾邮件自动召回反馈结果要显著优于其他分析方法的垃圾邮件自动召回反馈结果。

5　结束语

本文研究并实现了误判垃圾邮件自动召回系统，首先本文采用了联合 VSM/LSI/HowNet 的相似度计算方法，利用各类研究方法的特点，从不同的角度分析垃圾邮件的相似度。其次，本文分别提出了基于动态分类和静态分类的误判垃圾邮件自动召回模型，通过 6 个具体的算法得到误判垃圾邮件分类标准、自动召回的判别标准、用户反馈的参数调节方法以及数据采集周期的确定和滑动调节方法和策略。再次，本文以神经网络泛化能力、稳定性、顽健性、容错性、收敛能力为基础设计了 6 类多神经网络——BP 神经网络、遗传神经网络、PID 神经网络、小波神经网络、混沌神经网络以及灰色神经网络，并根据待预测的误判垃圾邮件自动召回模型的关键参数（相似度模型权值与用户反馈时间间隔）建立完整的神经网络集成系统。最后根据清华大学邮箱管理系统以及反垃圾邮件网关系统的日志与垃圾邮件文本选取 90 位校园网活跃用户进行前后历时一年时间（2011 年 9 月至 2012 年 10 月）实验。实验证明，采用神经网络集成系统对动态分类的误判垃圾邮件自动召回模型进行分析与计算，其结果优于其他类别的分析方法，达到了很高的用户正反馈率，同时显著降低了用户主动召回率。

参考文献：

[1] Mail abuse prevention system[EB/OL]. http://ers.trendmicro.com.

[2] MONOSTORI K，ZASLAVSKY A，SCHMIDT H. Document overlap detection system for distributed digital libraries[A]. Proeeedings of the ACM Digital Libraries 2000[C]. San Antonio,USA, 2000.226-227.

[3] MICHAEL W. B, SUSAN T. D, GAVIN W O'B. Using linear algebra for intelligent information retrieval[J]. SIAM Review, 1997, 37(1): 145-149.

[4] LANDAUER T K, FOLTZ P W, LAHAM D. Introduction to latent semantic analysis[J]. Discourse Processes, 1998, 25(5):259-284.

[5] PAPADIMITRIOU C, RAGHAVAN P, TAMAKI H. Latent semantic indexing: a probabilistic analysis[J]. Journal of Computer and System Science 2000, 61(2):82-93.

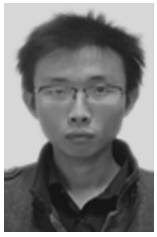
[6] DONG Z D, DONG Q. Bigger context and better understanding-expectation on future MT technology[A]. Proc of the International Conference on Machine Translation & Computer Language Information[C]. 1999.17-25.

[7] HANSEN L K, SALAMON P. Neural network ensembles[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 1990, 12(10): 993-1001.

[8] SOLLICH P, KROGH A. Learning with Ensemble: How Over-Fitting Can be Useful[M]. Cambridge: MIT Press, 1996. 337-353.

[9] PERRONE M P, COOPLER L N. When networks disagree: ensemble method for neural networks[A]. Artificial Neural Networks for Speech and Vision[C]. London: Chapman-Hall, 1993. 126-142.

作者简介：



林海卓（1988-），男，辽宁大连人，清华大学博士生，主要研究方向为网络管理与测量。

王继龙（1973-），男，黑龙江大兴安岭人，清华大学教授、博士生导师，主要研究方向为下一代互联网体系、网络管理与测量等。

吴建平（1953-），男，山西太原人，清华大学教授、博士生导师，主要研究方向为下一代互联网。

杨家海（1966-），男，浙江丽水人，清华大学教授、博士生导师，主要研究方向为计算机网络体系结构、IPv6 与下一代互联网技术。

徐聪（1986-），男，河北保定人，清华大学博士生，主要研究方向为网络管理与测量。