LSA和MD5算法在垃圾邮件过滤系统的应用研究

张秋余[1]，孙晶涛[1]，闫晓文[2]，黄文汉[3]

(1. 兰州理工大学计算机与通信学院 兰州 730050; 2. 陕西西禹高速公路有限公司 陕西 韩城 715400;
3. 陕西理工学院计算机系 陕西 汉中 723003)

【摘要】随着对垃圾邮件问题的普遍关注，针对目前邮件过滤方法中存在着的语义缺失现象和处理群发型垃圾邮件低效问题，提出一种基于潜在语义分析(LSA)和信息–摘要算法5(MD5)的垃圾邮件过滤模型。利用潜在语义分析标注垃圾邮件中潜在特征词，从而在过滤技术中引入语义分析；利用MD5在LSA分析基础上，对群发型垃圾邮件生成“邮件指纹”，解决过滤技术在处理群发型垃圾邮件中低效的问题。结合该模型设计了一个垃圾邮件过滤系统。采用自选数据集对文中设计的系统进行测试评估，经与Naïve Bayes算法过滤器进行比较，证明该方法在垃圾邮件过滤上优于Naïve Bayes方法，实验结果达到了预期的效果，验证了该方法的可行性、优越性。

关 键 词 邮件指纹; 特征提取; 潜在语义分析; MD5算法; 滑动窗口; 垃圾邮件过滤
中图分类号 TP393.098 文献标识码 A

Research of Spam Filtering System Based on Latent Semantic Analysis and MD5

ZHANG Qiu-yu[1]，SUN Jing-tao[1]，YAN Xiao-wen[2]，HUANG Wen-han[3]

(1. School of Computer and Communication, Lanzhou University of Technology Lanzhou 730050;
2. Shaanxi Xiyu Highway Corporation Ltd Hancheng Shaanxi 715400;
3. Department of Computer Science and Technology, Shaanxi University of Technology Hanzhong Shaanxi 723003)

Abstract Along with the widespread concern of spam problem, at present, there are spam filtering system about the problem of semantic imperfection and spam filter low effect in the multi-send spam. This paper proposes a model of spam filtering which based on Latent Semantic Analysis (LSA) and Message-Digest algorithm 5 (MD5). By making use of the LSA marks the latent feature phrase in the spam, a semantic analysis is introduced into the spam filtering technique; the "e-mail fingerprint" of multi-send spam is born with MD5 on the LSA analytical foundation, the problem of filtering technique's low effect in the multi-send spam is resolved with this kind of method. We design a spam filtering system based on this model. This system is evaluated with an optional dataset. The results obtained are compared with Naïve Bayes algorithm filter experiment results. The experiments show the expected results, and the feasibility and advantage of the new spam filtering method is validated.

Key words e-mail fingerprint; feature selection; latent semantic analysis; message-digest algorithm 5; slipping windows; spam filtering

随着Internet的迅速普及，电子邮件在为人们工作、学习和生活提供便利通信手段的同时，也为病毒、黑客程序、色情、反动、暴力、迷信等不良信息的传播提供了重要载体，邮件安全与垃圾邮件问题已被全球普遍关注[1]。垃圾邮件又被称为“不请自来的商业邮件”，给生产或商务活动带来了巨大的损失[2]。虽然陆续推出了几款邮件过滤软件，但在对比几种邮件过滤软件的原理后发现，目前的邮件过滤方法或多或少地存在着语义缺失的问题，当垃圾邮件发展到一定程度时，邮件过滤算法或过滤系统都将难以应付。而且现阶段多数垃圾邮件主体或发信人地址常动态改变，而其正文及附件内容却基本一致，并且对于拥有数万用户的大型局域网而言，垃圾邮件普遍以群发的方式向网内传播。针对这些特点，有必要引入新的邮件过滤思想来改进原有的解决方案。

收稿日期：2007 – 09 – 14
基金项目：“十一五”国家科技支撑计划(2006BAF01A21)
作者简介：张秋余(1966 –)，男，副研究员，主要从事数据仓库和数据挖掘、图像处理与模式识别、多媒体信息处理、信息安全、软件工程方面的研究.

1 关键技术概述

20世纪90年代，美国贝尔实验室为了有效提取信息而开发的潜在语义分析(Latent Semantic Analysis，LSA)，在近几年取得了长足的进步，广泛地使用在基于文本的研究中。潜在语义分析是在大量书写文本所形成的语义空间的基础上建立起来的有关知识表征的计算方法[3-4]，它所拥有的揭示词语–文本之间隐含语义关系的能力，在对抗包含隐蔽性信息的垃圾邮件时，提供了一个非常重要且有意义的研究方向。但是针对汉语的一些特点，潜在语义分析尚有多个难点有待进一步解决。信息–摘要算法5 (Message-Digest Algorithm 5，MD5)[5-6]是经MD2、MD3和MD4发展而来，它的作用是让大容量信息在数字签名软件签署私人密匙前被“压缩”成一种保密的格式(把一个任意长度的字节串变换成一定长的大整数)，MD5算法广泛用于加密和解密技术。本文对基于LSA和MD5算法的垃圾邮件过滤系统进行了研究。

2 关键技术分析

2.1 潜在语义分析(LSA)

潜在语义分析的基本观点是把高维的向量空间模型(VSM)表示的文档映射到低维的潜在语义空间中[7-8]。LSA/SVD最早被提出并使用，也是目前普遍使用的典型LSA空间的构造方法。它通过对文本集的词语–文档矩阵的奇异值分解(Singular Value Decomposition，SVD)计算[9-10]，提取K个最大的奇异值及其对应的奇异矢量，以构成新矩阵来近似表示原文本集的词语–文档矩阵。

首先要构造一个可表示为$m\times n$的词语–文档矩阵$\boldsymbol{X}=[x_{ij}]$的文档库，x_{ij}为非负值，表示第i个词在第j个文档中出现的频度。由于词和文档的数量都很大，而单个文档中出现的词又非常有限，$\boldsymbol{X}$一般为稀疏矩阵。通常x_{ij}要考虑来自两方面的贡献，即局部权值$L(i,j)$和全局权值$C(i)$，它们分别表示第i个词在第j个文档和整个文档库中的重要程度：

$$X(i,j)=L(i,j)C(i) \tag{1}$$

在对$\boldsymbol{X}$进行截取-SVD分解(设$m>n$, $\mathrm{rank}(\boldsymbol{X})=r$，存在$K$，$K<r$且$K<<\min(m,n)$)，在2-范数意义下，$\boldsymbol{X}$的秩-$K$近似矩阵$\boldsymbol{X}_k\approx\boldsymbol{X}=\boldsymbol{U}_k\boldsymbol{\mathfrak{R}}_k\boldsymbol{V}_k^{\mathrm{T}}$。其中，$\boldsymbol{U}_k$和$\boldsymbol{V}_k$的列向量均为正交向量；$\boldsymbol{I}_k$为$k$阶单位矩阵，则有：

$$\boldsymbol{U}_k^{\mathrm{T}}\boldsymbol{U}_k=\boldsymbol{V}_k^{\mathrm{T}}\boldsymbol{V}_k=\boldsymbol{I}_k \tag{2}$$

式中 $\boldsymbol{U}_k$和$\boldsymbol{V}_k$的列分别被称为矩阵$\boldsymbol{X}_k$的左右奇异向量；$\boldsymbol{U}_k$和$\boldsymbol{V}_k$的行向量分别作为词向量和文档向量；$\boldsymbol{\mathfrak{R}}_k$是对角矩阵，对角元素被称为矩阵$\boldsymbol{X}_k$的奇异值。

LSA通过奇异值分解和取K-秩近似矩阵，一方面，消减了原词语–文档矩阵中包含的“噪声”因素，更加凸现出词语和文档之间的语义关系；但另一方面，一些对文本影响较大的小语义特征集没有能够得到表现，识别精度也因此被降低。

2.2 MD5算法

MD5算法是20世纪90年代初开发出来的，它仅仅是一种压缩函数，不含任何参数，即无论多长的信息m，通过MD5算法，最后的输出都是128 b的0、1符号串。MD5算法的典型应用是对一段信息(Message)产生信息摘要(Message-Digest)，以防止被篡改。

MD5算法步骤为[11-12]：

(1) 按位补充数据(补位)：在MD5算法中，对信息m进行补位，使得信息m最终的位数对512求余的结果是448。也就是说补充上去的位数，使之作为512的倍数少64 b。具体补位操作是补一个1，然后补0至满足上述要求。

(2) 扩展长度：在完成补位工作后，将一个表示原信息m的长度64 b数补在最后，得到的结果数据就被填补成长度为512位的倍数。

(3) 初始化变量：本文用到四个变量，分别为A、B、C、D，均为32 b长。初始化为：A = 0X01234567、B =0X89abcdef、C = 0Xfedcba98 、D = 0X76543210。

(4) 处理信息：首先定义四个辅助函数：$F(X,Y,Z)=(X\&Y)|((\sim X)\&Z)$、$G(X,Y,Z)=(X\&Z)|(Y\&(\sim Z))$、$H(X,Y,Z)=X\wedge Y\wedge Z$、$I(X,Y,Z)=Y\wedge(X|(\sim Z))$。函数中的$X$、$Y$、$Z$均为32 b。如果$X$、$Y$和$Z$的对应位是独立和均匀的，那么结果的每一位也应是独立和均匀的。

(5) 处理数据：每一轮的输入为512 b和128 b的变量($ABCD$)，并更改变量内容。每一轮用由sine的四分之一函数构造的表$T[1,2,\cdots,64]$来表示，即：

$$T[i]=\lfloor 2^{32}|\sin(i)\rfloor \qquad i=1,2,\cdots,64$$

式中 i是弧度；$T[i]$是32 b，用作随机数。

(6) 输出：经上述步骤得到的$ABCD$为输出结果，A、B、C、D连续存放，共16 B，128位。A为低位，D为高位，按十六进制依次输出这个16 B。

3 垃圾邮件过滤系统模型

基于LSA和MD5算法的垃圾邮件过滤系统目标是达到对垃圾邮件高效、准确地过滤，在过滤技术中引入语义分析和“邮件指纹”生成等手段，使系统具有足够的灵活性和较强的适应性。垃圾邮件过滤系统模型如图1所示。

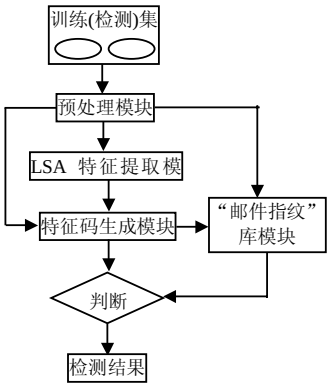

图1　垃圾邮件过滤系统模型

首先系统通过对适当邮件集的训练，让LSA特征提取模块从已知合法邮件和垃圾邮件中，提取出合法邮件和垃圾邮件的特征，并将这些特征自动保存在一个向量空间内。再将经过预处理模块处理的检测邮件交由LSA特征提取模块进行信息提取，得到的信息完全反应了LSA方法对信息的提取能力；将提取得到的信息作为检测邮件的“锚”值，包含“锚”值的检测邮件送交特征码生成模块，利用滑动窗口和MD5算法得到一定长度的特征码。这种方式克服了单纯依靠独立特征词生成的特征码对文本表示不准确的问题。因此，对垃圾邮件的识别就通过生成的特征码(“邮件指纹”)与“邮件指纹”库中的信息进行比对得到。

4　垃圾邮件过滤系统各部分设计

4.1　预处理模块

电子邮件不同于传统数据库中的结构化数据，其头部信息有一定结构，而内容就没有结构。若能对电子邮件这种半结构的数据施加信息处理，必须对其进行预处理，其主要步骤包括：文本特征格式分析、中文分词处理、无用词条过滤、词频加权等一系列的工作。目前这方面的研究很多，成果也很显著。本文使用基于多层隐马尔可夫模型的汉语词法分析系统(Institute of Computing Technology, Chinese Lexical Analysis System，ICTCLAS)[13]，该系统的功能有中文分词、词性标注、未登录词识别等。利用该系统结合禁用词过滤、去除特高频率词和特低频率词等预处理手法，去除意义不大的词条，以达到对邮件进行后续处理的要求。

4.2　LSA特征提取模块

在垃圾邮件过滤系统中，LSA特征提取模块是系统的核心部件之一。模块采用LSA技术，通过奇异值分解和取K-秩近似矩阵，凸显出词语和文档之间的语义关系。当每个词都有其矢量表示时，在进行潜在特征词(词组)识别过程中，只需依据已获取的识别结果(历史)去预测潜在特征词(词组)，将组成识别历史的所有词矢量予以加权求均值，产生历史矢量，此过程是为了产生的历史矢量融入长距离语义信息。$\{X_1,X_2,\cdots,X_{i-1}\}$和$P_{i-1}$分别表示$i$–1时刻之前所获得的词语矢量和相应时刻的历史矢量；扩展i时刻识别的结果，增加一个新词W_i，$\{X_1,X_2,\cdots,X_{i-1},X_i\}$和$P_i$分别表示$i$时刻所获词语的矢量和相应时刻的历史矢量，$\omega_j$表示$j(j=1,2,\cdots,i)$时刻所获得的词$W_j$相应于训练语料的熵。根据：

$$P_{i-1}=\frac{1}{i-1}\sum_{j=1}^{i-1}X_j\left[1-\omega_j\right],\ P_{\mathrm{i}}=\frac{1}{i}\sum_{j=1}^{i}X_j\left[1-\omega_j\right] \quad (4)$$

可得：

$$P_i=\frac{1}{i-1}P_{i-1}+\frac{1}{i}X_i\left[1-\omega_i\right] \quad (5)$$

式(4)作为潜在特征词(词组)识别时历史矢量的更新公式。更新识别历史矢量时，利用熵进行加权以区分每个词对识别历史的贡献。

4.3　特征码生成模块

垃圾邮件过滤系统中的特征码生成模块，它主要功能是生成垃圾邮件的“邮件指纹”。其过程实现如图2所示。

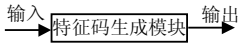

图2　“邮件指纹”实现过程

(1) 输入：为了使检测邮件经由 MD5 算法快速、高效、准确地得到“邮件指纹”，需要在邮件文本中设立采样点。经上层模块处理得到包含“锚”值的邮件文本，以这些“锚”值作为邮件的采样点。为了更有效反映文本信息，也为了不破坏词汇之间的联系，引入滑动窗口的特征提取算法对“锚”值周围区域的词汇重组，进一步扩大特征的选择范围，以使得提取出的特征词能够更准确地反映文本的特征，并以这些特征词作为输入。

(2) 输出：经由滑动窗口的特征提取算法得到的特征值通过安全哈希 MD5 算法输出一定长的特征码，并将输出的特征码存储到后台的“邮件指纹”库中。

4.4　“邮件指纹”库模块

本模块数据库采用MySQL数据库，它是一个快速、多线程、多用户和强壮的SQL数据库服务器。与目前现有的数据库系统相比，具有反应速度快的特点，可跨平台使用[14]。数据库主要存储tab-files表，它包括files字段及fingerprint字段。files字段保存文本信息，fingerprint字段保存文本生成的数字指纹信

息，主要是为了避免出现文件重复的情况。tab-files表中导入数据的过程如下：

(1) 经处理得到邮件文本 M，计算 M 的“邮件指纹”FP_M。(2) 查询数据库中是否有与 FP_M 相同的“邮件指纹”。(3) 若存在，跳过此文本，转到过程(1)，处理下一邮件，直到不存在未导入数据库的邮件。(4) 若不存在，将此邮件文本及对应的“邮件指纹”保存在表 tab-files 中，转到过程(1)，继续处理下一篇邮件，直到不存在未导入数据库的邮件。

5 系统测试与分析

进行系统测试时，语料库的选择是非常重要的，在国外已经有了一些标准权威的语料库，如：PU1语料库[15-16]。而在中文垃圾邮件分类方面，目前还没有一个公认权威的中文语料库。因此，本文选用了广泛收集的1 500篇各类垃圾邮件，构成近15 MB的训练集，实验平台为PM2.1 GB，2 GB内存。

首先对这些邮件正文进行提取和处理，经过中文分词、禁用词过滤、去除特高频率词和特低频率词等预处理，生成5 672×1 500的词语–文本矩阵，然后进行SVD分解生成潜在语义空间。在这个过程中，降维因子K值的选取直接关系到语义空间模型的效率，K值过小会丢失一些有用的信息，K值过大则会使运算量增加。因此，根据不同的文本集和处理要求，选取最佳的K值十分必要。本文利用贡献率不等式，即 $\mathfrak{R}=\mathrm{diag}(a_1,a_2,\cdots,a_n)$，且 $a_1\geqslant a_2\geqslant\cdots\geqslant a_t=\cdots=a_n=0$，$K$满足贡献不等式：

$$\sum_{i=1}^{k}a_i\Big/\sum_{i=1}^{t}a_i\geqslant\varphi \tag{5}$$

式中 φ 为包含原始信息的阈值，取70%。贡献率不等式是参考因子分析的相应概念提出的用以衡量K维子空间对于整个空间的表示程度[17]。图3的K值分析图表明，K值提高到400时，表示程度达到最高，随后继续加大K值，表示程度产生异变，甚至下滑。经分析，当K值提高到一定阶段后，代表特征已经基本表示，接着表示的都是噪音空间。当K=300时表示程度与K=400时基本相同，但K=300，时间花费较少，因此，本文选取K=300。

在特征码(“邮件指纹”)生成阶段，滑动窗口大小的设置，也会影响到整个过滤器的效能。如图4所示，滑动窗口越大，整个系统的过滤效果越好[18]。因为窗口越大，选择的特征越多，也就越能选出代表文本的特征，不论是召回率还是正确率都有所提高。但窗口越大，运行速度会随之降低，运行时间也会增加，当窗口大小为2时，效果和运行速度之间平衡最好。

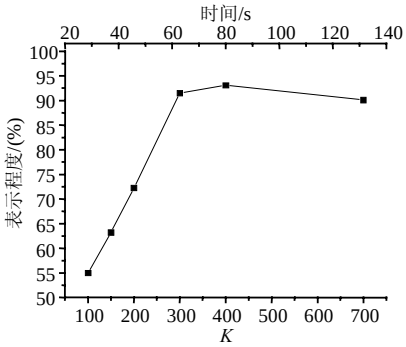

图3 K值分析

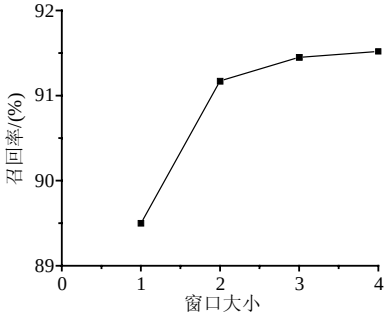

图4 窗口大小分析

为了说明系统在垃圾邮件过滤中的实际效果，将它与Naïve Bayes方法[19]进行了实验比较，实验结果如表1所示。

表1 LSA和MD5算法与Naïve Bayes的系统实验结果

算法	召回率/(%)	正确率/(%)	F_1值/(%)
Naïve Bayes	76.37	94.21	84.36
LSA和MD5	91.17	92.83	91.99

表1是对300多封邮件进行实验的结果，可以看出，本文的方法使邮件系统召回率上升了14.8%，识别准确率仅降低1.38%，而F_1值提高了7.63%。

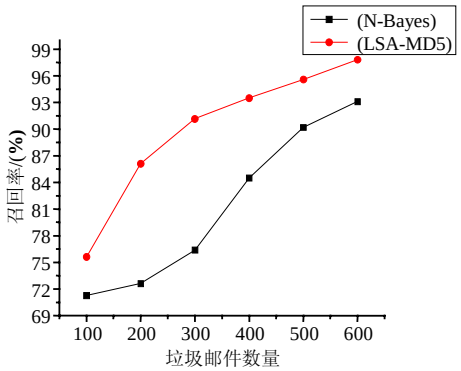

图5 召回率

图5、6是对基于LSA和MD5算法的垃圾邮件过滤系统与Naïve Bayes算法的系统在不同检测集下的

系统性能曲线图。通过对这些数据的分析，可以看出，系统在垃圾邮件过滤性能上优于Naïve Bayes算法过滤器，其设计达到了预期的效果，具有较好的应用前景。

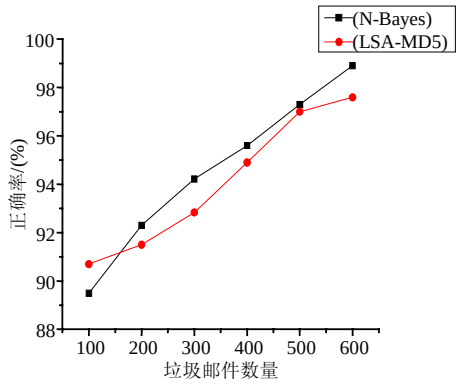

图6　正确率

6 结　论

本文提出并实现基于LSA和MD5算法的垃圾邮件过滤系统，该系统利用了LSA、MD5算法、数据库等技术，在过滤技术中引入文本潜在语义分析，并针对群发型垃圾邮件特征，结合MD5算法生成“邮件指纹”，达到了对群发型垃圾邮件的高效、准确地过滤，从而找出了一种垃圾邮件过滤的新思路。

参　考　文　献

[1] 中国互联网反垃圾邮件联盟[EB/OL]. http://www.anti-spam.org.cn, 2007-03-10.

[2] GUO Y, ZHANG Y, LIU J, et al. Research on the comprehensive anti-spam filter[J]. Industrial Informatics, 2006: (2): 98-100.

[3] 王　怡, 盖　杰, 武港山, 等. 基于潜在语义分析的中文文本层次分类技术[J]. 计算机应用研究, 2004, 21(8): 32-33.

[4] YEH J Y, KE H R, YANG W P, et al. Text summarization using a trainable summarizer and latent semantic analysis[J]. Information Processing & Management, 2005, 41(1): 75-95.

[5] JARVINEN K, TOMMISKA M, SKYTTA J. Hardware implementation analysis of the MD5 hash algorithm[C]// Proceedings of the 38th Hawaii International Conference on System Sciences. Hawaii, USA: IEEE Computers Society Press, 2005: 320-322.

[6] PRENEEL B, VAN O P C. On the security of iterated message authentication codes[J]. Information Theory, 1999, 45(1): 121-123.

[7] Weizhong Zhu C C. Storylines: Visual exploration and analysis in latent semantic spaces[J]. Computers & Graphics, 2007, 31(3): 78-79.

[8] MALETIC J I, MARCUS A. Using latent semantic analysis to identify similarities in source code to support program understanding[C]//Tools with Artificial Intelligence. Vancouver: IEEE, 2000: 321-323.

[9] MARTIN D I, MARTIN J C, BERRY M W, et al. Out-of-core SVD performance for document indexing[J]. Applied Numerical Mathematics, 57(11-12): 1994:224-226.

[10] 盖　杰, 王　怡, 武港山. 潜在语义分析理论及其应用[J]. 计算机应用研究, 2004, 21(3): 161-164.

[11] 卢开澄. 计算机密码学-计算机网络中的数据保密与安全[M]. 第3版. 北京: 清华大学出版社, 2003.

[12] CHAVEZ E, TELLEZ E S. A universal full text index with access control and annotation driven information retrieval[C]//CIC06, AI & Society. [S.l.]: [s.n.], 2006: 1123-1127.

[13] ZHANG Hua-ping, YU Hong-kui. HHMM-based Chinese lexical analyzer ICTCLAS[C]//41st Annual Meeting of the Association for Computational Linguistic. Sappor: [s.n.], 2003: 1431-1435.

[14] 杜波依斯. MySQL权威指南[M]. 北京: 机械工业出版社, 2004.

[15] ANDROU T I, PAL I G M K E. Learning to filter unsolicited commercial E-mail [EB/OL]. http://www. aueb. gr/users/ion/docs/TR2004_updated.pdf, 2007- 01-16.

[16] ANDROU T I, PAL I G M K E. An Evaluation of Naïve Bayesian Anti-Spam Filtering[C]//11 European Conferences on Machine Learning, System Sciences. Barcelona, Spain: [s.n.], 2000: 1165-1168.

[17] 林鸿飞, 姚天顺. 基于潜在语义索引的文本浏览机制[J]. 中文信息学报, 2000, 14(5): 111-114.

[18] 李建中, 张冬冬. 滑动窗口规模的动态调整算法[J]. 软件学报, 2004, 15(12): 13-16.

[19] STORY R E. An explanation of the effectiveness of latent semantic indexing by means of a Bayesian regression model[J]. Information Processing & Management, 1996, 32(3): 329-344.

编　辑　漆　蓉