一种基于学习向量量化网络的垃圾邮件过滤方法*)

詹 川 卢显良 周 旭 侯孟书
(电子科技大学计算机科学与工程学院 成都610054)

摘 要 伴随着电子邮件的广泛使用,垃圾邮件泛滥成灾,严重影响了人们正常的学习、工作和生活。本文针对目前的垃圾邮件主要是由多种商业或政治性类别的垃圾邮件组成的特点,利用学习向量量化网络能把多个子类合并成一个复杂大类的特性,构建了一个反垃圾邮件的LVQ神经网络模型,我们对该LVQ网络模型进行了与其他算法的对比试验,试验表明它比基于贝叶斯公式算法和基于神经网络BP算法的过滤器有更好的性能。

关键词 学习向量量化,垃圾邮件过滤,互信息,向量空间模型

An Anti-spam E-mail Filter Based on LVQ Network

ZHAN Chuan LU Xian-Liang ZHOU Xu HOU Meng-Shu
(College of Computer Science and Engineering, UESTC of China, Chengdu 610054)

Abstract Along with wide application of E-mail nowadays, a large number of spam E-mails flood into people's life and bring catastrophe to their study and life. This paper presents a novel anti-spam e-mail filter based-LVQ network in terms of spam e-mail's characteristic which are mainly made up of several kinds commercial or political spam emails at present. Our experiment has proved that our filter based-LVQ is superior to anti-spam email filter based-Bayes and that based-BP in total performances apparently.

Keywords LVQ, Anti-spam E-mail filtering, Mutual information, Vector space model

1 引言

伴随着Internet的普及,电子邮件以其快捷、方便、低成本的特点日益得到了广泛的使用,成为互联网上最重要、最普及的应用。但是随之而来的垃圾邮件也越来越猖獗,严重地影响和损害人们的工作、生活和学习。联合国贸易与发展会议(UNCTAD)最近发表的2003年电子商务与发展报告指出,美国是全球最大的垃圾邮件制造者,全球用户收到的垃圾邮件有一半多源自美国。垃圾邮件除了骚扰网络用户外,还给企业带来了巨大的损失。联合国贸发会议援引Message Labs的数据说,垃圾邮件给全球企业带来的损失高达205亿美元。据2004年3月的最新统计[1],中国用户现在平均每人每周发送电子邮件9.8封,收到正常电子邮件12.6封,收到垃圾电子邮件19.3封。收到的垃圾邮件量占收到的总邮件的60.5%,较2003年的垃圾邮件在占所有邮件的比例26.27%上涨了34.23个百分点,上升趋势迅猛,造成了网民的怨声载道,为此公安部、教育部、信息产业部、国务院新闻办在2004年2月专门作出部署,开展互联网垃圾电子邮件专项治理工作,净化互联网环境。造成2004年初垃圾邮件比例大幅度上涨的主要原因是垃圾邮件发送者的技术不断更新和电脑病毒的泛滥。目前国内邮件服务提供商所采用的反垃圾邮件的手段主要是软件自动过滤和人工管理相结合的方式,比如系统全局规则、IP过滤规则、客户过滤规则及内容过滤规则等,但是这些方法相对简单,不能很好地适应垃圾邮件的多样性,因此服务商也只能过滤掉50%左右的垃圾邮件。因此迫切需要更加智能化的垃圾邮件过滤技术来治理日益猖獗的垃圾邮件问题。

在基于内容智能过滤垃圾邮件技术方面,普遍把邮件被当为特殊的文本,对其进行文本识别和分类。Cohen[2]采用TF-IDF表示邮件权重,用RIPPER规则学习算法来对邮件进行分类。Sahami[3]等人利用统计概率学理论和信息检索技术,用向量空间模型表示邮件,假设向量中的各个项是相互独立的,用贝叶斯公式来计算邮件是垃圾邮件的概率,来识别邮件。实验[4]证明,基于贝叶斯公式的方法在垃圾邮件识别效果方面远远好于基于关键字过滤的方法。Xavier Carreras[5]等人应用Boosting算法来进行反垃圾邮件,在公共邮件样本PU1集上试验,显示出这种方法优于基于贝叶斯公式的方法以及运用决策树的方法。Duhong Chen[6]等人对贝叶斯算法、决策树、神经网络和Boosting四种算法对垃圾邮件过滤的效果进行了比较,发现神经网络算法有更高的正确分类率。James Clark[7]等人编写LINGER软件,他们构建了一个3层的BP神经网络,其中输入层的神经元个数为特征项数,隐藏层的神经元数选取在20~40范围内,来进行垃圾邮件过滤,发现BP网络结合用信息增益来选取特征项的方法,有相当好的垃圾邮件识别效果。

本文针对范畴宽泛的垃圾邮件主要由多种商业性质或政治目的类别的垃圾邮件组成的特点,利用学习向量量化网络能把多个子类合并成一个大类的特性,构建了一个反垃圾邮件的LVQ神经网络模型,对该LVQ网络模型进行了与其他算法的对比试验,试验表明比基于贝叶斯公式算法和神经网络BP算法过滤器有更好的性能。

2 邮件样本及数据预处理

2.1 邮件的表示

为了能让计算机自动地对邮件分类,我们需要把邮件表

*)信息产业部电子工业生产发展基金资助项目,编号:[2002]11006。詹 川 博士研究生,主要研究方向:计算机网络,信息安全,人工智能。卢显良 教授,博士生导师,主要研究方向:计算机网络,分布式系统,信息安全。周 旭 博士研究生。侯孟书 博士研究生。

示为计算机能够处理的形式，向量空间模型(VSM)[8]是一种在信息检索技术中常用而且效果较好的文本表示方法，在该模型中，邮件文本被看作是由一组正交向量组成的向量空间。若该空间的维数为 n，则邮件 d 可被表示为特征向量 $V(d)=(x_1,x_2,\cdots,x_n)$，$V$ 的每个分量表示对应特征在该篇邮件中的权值。

计算邮件的特征权值 x，我们采取的是词频逆文本频率(TFIDF)的方法[9]。这种方法认为词条在文献中的频率正比于其在文献中出现的频率，反比于文本内出现该词条的文本数，词条 t_i 在文档 d 中的 TFIDF 值由式(1)定义：

$$TFIDF_i=TF_i\times log(N/DF_i) \tag{1}$$

其中 TF_i 是词条 t_i 在文档 d 中出现的频数，N 表示全部训练文档的总数，DF_i 表示包含词条的文档频数。

2.2 特征项的提取

特征向量的特征项可以选取邮件中的单词(如：price、adult、shop)，短语(如：on sale、be over 21)以及非文本属性(如：是否带有附件，是否为 html 格式)。我们收集到的邮件样本已经过处理，都不带附件和 html 标志。为了把我们注意力放在我们的算法上，为让测试程序简单，我们本次试验中全部选取单词作为邮件的特征项。

如果我们把邮件里的所有单词作为特征项，会造成向量维数太大，计算量巨大，不切实际。我们采取计算每个单词的各类中的互信息量方法[10]来提取特征项，降低特征空间的维数。互信息(Mutual Information)在文本分类中被广泛采用。词条 t 与类别 s 的互信息可由式(2)计算：

$$MI(t,s)\approx\log\frac{A\times N}{(A+B)\times(A+C)} \tag{2}$$

其中 A 表示包含词条 t 且属于类别 s 的邮件频数，B 为包含 t 但是不属于 s 的邮件频数，C 表示属于 s 但是不包含 t 的邮件频数，N 表示训练集中邮件总数。如果 t 和 s 无关(即 $P(ts)=P(t)\times P(s)$)，则 $MI(t,s)$ 值为零。对于多类别问题，我们采取分别计算该词条对每个类别的互信息量，用式(3)选取最大的互信息量值作为该词条的互信息量。

$$MI_{max}(t)=\max_{i=1}^{m}MI(t,s_i) \tag{3}$$

其中 m 为类别数，将互信息低于设定阈值的词条从原始特征空间中移除，降低特征空间的维数，保留高于阈值的词条。

3 反垃圾邮件的 LVQ 模型

3.1 垃圾邮件的定义和分类

关于垃圾邮件的定义，目前国内有明确的定义，以下情况属于垃圾邮件范畴：

1. 收件人事先没有提出要求或者同意接受的广告、电子刊物、各种形式的宣传品等宣传性的电子邮件；

2. 收件人无法拒收的电子邮件；

3. 隐藏发件人身份、地址、标题等信息的电子邮件；

4. 含有虚假的信息源、发件人、路由等信息的电子邮件；

5. 含有病毒、恶意代码、色情、反动等不良信息或有害信息的邮件。

含有病毒、恶意代码的邮件其实是属于病毒邮件的范畴，不属于传统垃圾邮件的范围，它将不在我们考虑过滤的范围内。

据统计[1]用户收到的垃圾邮件主要由网上购物、IT 产品推销、网上赚钱、情趣用品、订房/订票/旅游、政治相关、销售商业数据、色情暴力相关及其他类别组成，如图1所示，其中网上购物、IT 产品推销和网上赚钱占前三位。可以看出垃圾邮件的范畴是相当宽泛的，当前垃圾邮件主要是由多种商业性质或政治性质类别的垃圾邮件组成。不同类别的垃圾邮件的特征各不相同，而且有些相差很大；同时在提取特征词方面，如果笼统提取整个垃圾邮件的特征词，提取出的特征词比较离散，特征不明显。上面两个原因造成难于在全局上确定正常邮件和垃圾邮件的界限，不利于准确识别垃圾邮件。如果把垃圾邮件细分成上面几类，把每个类的特征词提取出来，相应特征就比较鲜明，易于智能识别。于是我们把正常邮件与垃圾邮件的分类转换成对邮件具体类别的分类，然后再根据具体分类的情况来判断是否是垃圾邮件。一般情况下正常邮件跟上面各具体类别的垃圾邮件内容上有很大的差别，如果某封邮件被识别为上述的某种类别，我们就认为这封邮件为垃圾邮件。

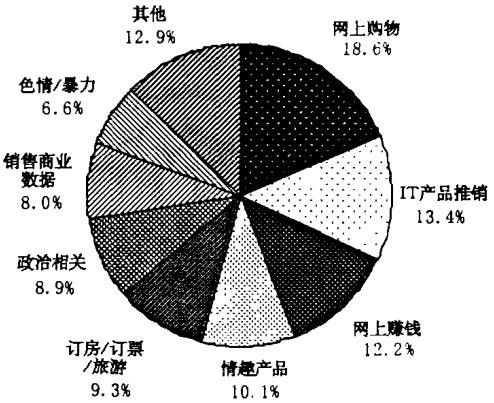

图1 垃圾邮件的种类统计

3.2 学习矢量量化(LVQ)神经网络模型

学习矢量量化神经网络是一种混合网络，通过有监督及无监督的学习来形成分类。该模型分为两层，第一层是竞争层，第二层是将竞争层的分类的结果传递到用户定义的目标分类上。在竞争层，网络将自动学习对输入向量进行分类，竞争获胜的神经元代表的是一个子类而不是一个类，一个类可能由多个不同的神经元(子类)组成。第二层的线性层在把这几个子类组合成一类，组合过程通过 W^2 矩阵来实现的，矩阵的列代表子类，行代表类，每列仅有一个1，其他元素都设置为0，每列中1所在的行表明该子类属于所在行代表的类。

$(w^2_{ki}=1)\rightarrow$子类 i 是类 k 的一部分

通过将多个子类组合成类的过程使得 LVQ 网络可以产生相当复杂的类边界。因此 LVQ 网络适合对存在多种类别的垃圾邮件的识别。

3.3 识别垃圾邮件的 LVQ 算法

1. 初始化权值向量(聚类中心)，$W=\{w_1,w_2,\cdots,w_n\}$，初始化学习率 $\alpha\in[0,1]$。

2. 从训练邮件样本集中抽出一个样本，计算它到各个聚类中心(权值向量)的距离，这里的距离我们采用表示两个文本相似度的余弦(cosine)距离代替传统的欧式距离，计算相似度的余弦距离如式(4)所示：

$$Sim(U,V)=\frac{\sum_{k=1}^{n}w_{uk}\cdot w_{vk}}{\sqrt{\sum_{k=1}^{n}w_{uk}^2}\sqrt{\sum_{k=1}^{n}w_{vk}^2}} \tag{4}$$

然后比较样本与各聚类中心相似度大小式(5)，则与输入样本有最大相似度的那个神经元竞争获胜，其输出为1，其他隐藏

层的神经元输出为0。

$$a^1=\max(Sim(x,x')) \tag{5}$$

3. 修正权值，已知输入样本的类别为 r，而在竞争学习中获胜的神经元 c 的类别为 s，则根据式(6)来修正权值向量。

$$\begin{cases} w_c(t+1)=w_c(t)+u(t)[x(t)-w_c(t)]; & r=s \\ w_c(t+1)=w_c(t)-u(t)[x(t)-w_c(t)]; & r\neq s \\ w_i(t+1)=w_i(t); & i\neq c \end{cases} \tag{6}$$

4. 调整学习率 $u(t)$，随着迭代逐渐减小。

5. 检查终止条件，看是否到达所要求的迭代步数。

3.4 反垃圾邮件 LVQ 模型参数选择

3.4.1 *反垃圾邮件的 LVQ 模型结构* 图2为我们构建的反垃圾邮件的 LVQ 网络试验模型，在输入层，我们选择100个特征项作为输入节点，根据我们以前的试验发现，在选择100特征项时有比较好的结果，当继续增加特征项时，发现性能改变甚微，并且大大增加了计算量。在竞争层，神经元的个数我们根据垃圾邮件的分类，大致分为九类，我们认为基于某个人的正常邮件，其内容相对集中在一定的领域，然后我们把正常邮件再归为一类，因此我们在竞争层设置了10个神经元，选用的是竞争函数，在输出层，设置了正常邮件和垃圾邮件两个神经元，选用的线性函数。

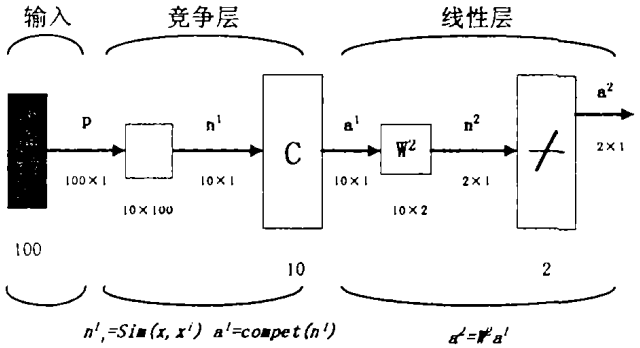

图2 反垃圾邮件的 LVQ 网络

3.4.2 *学习率的选择* 为了加快学习效率，并且能够尽快收敛，我们把学习分为两个阶段，第一阶段为粗学习，在这阶段 $u(t)$ 选取大于0.05的参数，加快学习步伐，当各输入向量有了相对的映射位置，就转入第二阶段，细调整阶段。在这阶段网络学习集中在相对较小范围内的权值进行调整，这阶段我们把 $u(t)$ 设为0.05。

3.4.3 *初始化权值矢量的选定* 为了防止 LVQ 算法中可能出现的在不断学习的过程中，竞争层各神经元权值无法收敛的现象，我们从各类垃圾邮件及正常邮件样本中各抽取一个样本，矢量化后作为初始的权值矢量。

4 实验及结果

4.1 邮件样本的收集

本次试验采用的训练和测试邮件样本来自于 http://www.spamassassin.org/publiccorpus。它是一个用于垃圾邮件过滤研究的公开可用的邮件样本。我们用的版本为20030228样本集，我们按中国目前的垃圾邮件大致比例从中选取了1000封邮件，其中，580封垃圾邮件，420封正常邮件。然后我们把训练集按前面列举的垃圾邮件的种类，人工地把垃圾邮件分成网上购物、IT 产品推销、网上赚钱、情趣用品、订房/订票/旅游、政治相关、销售商业数据、色情暴力相关及其他类别，以用于 LVQ 网络的训练学习。这些邮件全是英文邮件，其中的 html 标记和附件都已被除去。

4.2 评价指标

我们在试验中测试了过滤器的两个评定指标：SP(垃圾邮件识别的准确率)和 SR(垃圾邮件识别的查全率)。

$$SP=\frac{n_{spam\to spam}}{n_{spam\to spam}+n_{legit\to spam}} \tag{7}$$

$n_{spam\to spam}$ 为正确识别出的垃圾邮件数，

$n_{legit\to spam}$ 为正常邮件被误识别为垃圾邮件数。

$$SR=\frac{n_{spam\to spam}}{N_{spam}} \tag{8}$$

N_{spam} 为样本邮件中垃圾邮件的总数。

SP 和 SR 反映的是邮件过滤器质量的两个方面，为了综合考虑邮件过滤器的性能，我们引入一个新的评估指标 $F1$，它综合考虑了准确率和查全率两方面的指标，其数学公式如下：

$$F1=\frac{SP\times SR\times 2}{SP+SR} \tag{9}$$

4.3 实验结果

在试验中，我们先测试了我们建立的 LVQ 神经网络在进行不同次数训练后，其过滤器对垃圾邮件识别的效果，见表1，同时测试了开放集和封闭集两种状态。然后我们测试了基于神经网路 LVQ 网络(ANN－LVQ)反垃圾邮件过滤器与基于贝叶斯算法(Naïve Bayes)以及基于神经网络 BP 网络(ANN－BP)反垃圾邮件过滤器的过滤性能，见表2。

表1 不同训练次数的效果

训练次数	开放集		封闭集	
	SP(%)	SR(%)	SP(%)	SR(%)
500	90.64	87.45	91.06	88.96
1000	95.83	92.36	96.74	95.29
1500	98.97	93.58	99.51	96.86

表2 不同算法的比较

	SP(%)	SR(%)	$F1$(%)
Naïve Bayes	97.63	86.48	91.72
ANN－BP	98.42	91.26	94.70
ANN－LVQ	98.97	93.58	96.20

从表1可看出，我们设计的 LVQ 网络在从训练500次 到1000次后，查准率和查全率都有明显的增高，训练次数在500次时，网络还不收敛，训练还不够充分。当我们继续把训练次数增加到1500次，查准率和查全率有一定的提高，但增加幅度较小，逐渐到达饱和。此时在开放集状态下，查准率 SP 达到98.97%，查全率 SR 达到93.5%，封闭集状态下，准确率 SP 为99.51%，查全率 SR 为96.86%。

在表2中，我们比较了三种基于不同算法过滤器的性能，神经网络实现的垃圾邮件过滤器的效果普遍要好于基于 Naïve Bayes 算法的。基于神经网络的方法比基于 Naïve Bayes 算法的在准确率上有一定的提高，但相差不大，但在查全率方面，提高的幅度较大。对于基于神经网络两种算法，基于 LVQ 网络过滤器相对于基于 BP 网络的来说，在查准率和查全率都有一定的提高。

结论 本文针对范畴宽泛的垃圾邮件特点，对其细化分类，经过分类的各类垃圾邮件，其相应特征更加突出，便于我们进行识别，然后利用神经网络 LVQ 的特点，它可以把多个子类组合成一个复杂的大类，来进行垃圾邮件识别。通过测试

(下转第87页)

根据跟踪信息，我们清楚地发现各服务器的执行过程。代理服务器将在1000，3000，4000，6000的地方进行服务器切换，如图5(b)，可以发现1000，3000，4000，6000条记录附近出现了轻微的峰值，这正是代理在进行服务器切换。例如0到1000期间，服务器1为活动服务器，第1000条请求到来时，活动服务器被切换到服务器2，依次类推。同时，各服务器的归档过程也在响应时刻按时执行。总体来看，这种无缝的服务器切换过程得益于良好的调度策略。

实验结果表明，本文的目录服务模型对目录服务请求的平均响应速度远远小于 Kwok-Yan Lam 模型的平均响应速度。切换服务器的代价远远小于归档带来的影响，这一点可以比较图5(a)、(b)得出。在使用良好的调度策略后，切换服务器所带来的目录响应延迟几乎可以忽略不计。

我们在测试过程中发现，目录的大小和归档线程的间隔是有密切联系的。如果目录较大，归档间隔小，则由于前一次归档未完后继的大部分归档得不到执行。因此，制定恰当的归档策略很重要。同时，该试验结果证明我们的调度策略是非常成功的，根据调度策略制定的初始化归档计数，有效地保证了在某一目录服务环境——目录数据、请求频率、服务器性能等下，交替归档的成功实现。总之，对于相同的目录服务环境，基于代理的目录服务模型不但保证了目录服务的安全，也在一定程度上提高了目录服务的性能，减少了归档过程对目录服务性能的影响。

小结 本文在分析虚拟组织资源管理的动态性、多样性等特点的基础上，结合其资源管理需求，提出了一种基于代理的目录服务模型。该模型保证了目录服务系统本身的安全性，同时在多目录服务器的环境下，利用调度策略有效平衡系统负载，实现了高效稳定的目录服务。在虚拟组织规模较大、资源访问过于频繁的情况下，多台服务器交互保证了目录服务的高响应性能，提高了虚拟组织的资源访问速度。

参 考 文 献

1 Katzy B R. Design and Implementation of Virtual Organization, System Sciences. In: Proc. of the Thirty-First Hawaii Intl. Conf. on, 1998, 4: 142～151

2 Chrysanthis P K, Znati T, Banerjee S, Chang Shi-Kuo. Establishing virtual enterprises by means of mobile agents, Research Issues on Data Engineering: Information Technology for Virtual Enterprises. In: RIDE-VE '99. Proc, Ninth Intl. Workshop on, March 1999. 116～123

3 Fitzgerald S, Foster I, Kesselman C, et al. A Directory Service for Configuring High-Performance Distributed Computations, High Performance Distributed Computing. In: Proc. The Sixth IEEE Intl. Symposium on, 1997. 365 ～ 375

4 Foster I, Kesselman C, Tuecke S. The Anatomy of the Grid ---- Enabling Scalable Virtual Organizations, Cluster Computing and the Grid. In: Proc. First IEEE/ACM Intl. Symposium on, 2001. 6 ～ 7

5 Czajkowski K, Fitzgerald S, Foster I, Kesselman C. Grid Information Services for Distributed Resource Sharing, High Performance Distributed Computing. In: Proc. 10th IEEE Intl. Symposium on, 2001. 181 ～ 194

6 Birrell A D, Jones M B, Wobber E P. A Simple and Efficient Implementation for Small Database. In: Proc. of the 11th ACM Symposium on Operating Systems Principles, 1987. 149～154

7 Lam K-Y, Salkield T. Implementing a Highly Available Network Directory Service. Systems Software, 1997, 37: 41～47

8 Cannon S, Chan S, Olson D, Tull C. Using CAS to Manage Role-Based VO Sub-Groups. In: Computing in High Energy and Nuclear Physics Conf. San Diego, 2003

9 Pearlman L, Welch V, Foster I, Kesselman C, Tuecke S. A Community Authorization Service for Group Collaboration. Policies for Distributed Systems and Networks. In: Proc. Third Intl. Workshop on, 2002. 50～59

10 Yang C S, Liu C Y, Chen J H, Sung C Y. Design and Implementation of Secure Web-based LDAP Management System, Information Networking. In: Proc. 15th Intl. Conf. on, 2001. 259～264

(上接第68页)

基于 LVQ 网络反垃圾邮件过滤器发现：

1. 网络的训练次数对于 LVQ 的性能有很大的影响，如果没有充分的训练，性能还达不到要求，在训练1500次左右，性能基本达到稳定。

2. 基于神经网络的方法普遍好于贝叶斯算法。分析大概是因为神经网络更加整体考虑了各个特征单词之间的关系，而在贝叶斯算法中，它是过于简单地假设各个特征单词之间是独立的。

3. 基于 LVQ 算法的过滤器优于基于 BP 算法的，是因为我们细化了垃圾邮件的类型，每个类型的垃圾邮件有更加明确的特征向量范围，更利于计算机的识别。

参 考 文 献

1 上海艾瑞市场咨询公司. 2004年中国反垃圾邮件研究报告, 2004, 3

2 Cohen W W. Learning rules that classify e-mail. In: proc. of the 1996 AAAI Spring symposium in information access, 1996

3 Sahami M, Dumais S, et al. A Bayesian Approach to Filtering Junk E-Mail. Learing for Text Categorization -Papers from the AAAI Workshop, Madison Wisconsin. 1998

4 Androutsopoulos I, Koutsias J, et al. An experimnental comparison of Naive Bayesian and Keyword-based anti-spam filtering wiht encrypted personal messages. In: Proc. of the 23rd annual intl. ACM SIGIR conf. on research and development in information retrieval, Athens, Greece

5 Carreras X, Mrquez L. Boosting trees for anti-spam email filtering. In: Proc. of RANLP-01, Jth Intl. Conf. on Recent Advances in Natural Language Processing, Tzigov Chark, BG, 2001

6 Chen Duhong, Tong jie, et al. Spam Email Filter Using Naive Bayesian, Decision Tree, Neural Network and AdaBoost. http://www.cs.iastate.edu/~tongjie/spamfilter/paper.pdf

7 Clark J, Koprinska I, Poon J. A neural network based approach to automated e-mail classification. In: Proc. of the IEEE/WIC intl. conf. on web intelligence

8 Salton G, Wong A, Yang C S. A vector space model for automatic indexing. Communication s of the ACM, 1975

9 Salton G. Introduction to modern information retrieval. New York. McGraw-Hill Book company, 1983

10 Church K W, Hanks T. Workd association norms, mutual information and lexicography. In: Proc. of ACL27, Wancouver, Canada, 1989

11 Hagan M T, Demuth H B, Beale N H. Neural network design. China Machine Press, 2002, 8