Two-Step Spam Message Filtering Method Based on Optimal Segmentation Strategy

WAN Difei[1], CHEN Jieshu[2]

1. China Mobile Group Design Institute Co., Ltd. Chongqing Branch, Chongqing, China

2. China Mobile Group Design Institute Co., Ltd. Chongqing Branch, Chongqing, China

e-mail: wandifei@ cmdi.chinamobile.com, chenjieshu@ cmdi.chinamobile.com

Abstract: As the rapid development of the Internet and the telecommunications industry, a large number of junk information appears in people's lives, such as garbage message and spam. In this paper, a two-step spam text message filtering method based on optimal segmentation strategy is proposed, which introduces optimal segmentation strategy into two-step classification methods and make use of genetic algorithm to search optimal partition line within the range of non-reliable text near the separate lines in two-step classification method. The first step of this method is characterized by the word BI-GRAM, which avoids complicated word segmentation process and improves the efficiency, while the second step is characterized by the words BI-GRAM to make the classifier to achieve the best performance. The experiments on the Chinese text corpus data sets consisting of 12,600 texts show that the method has higher classification performance and efficiency and its precision, recall and F1 value respectively reached 98.55%, 97.23% and 97.89%, which means it can effectively filter junk information.

Keywords: two-step classification; naive bayesian; genetic algorithm; optimal division strategy

基于最优分割策略的两步垃圾信息过滤方法

万狄飞[1]，陈杰姝[2]

1. 中国移动通信集团设计院有限公司重庆分院，重庆，中国，400042

2. 中国移动通信集团设计院有限公司重庆分院，重庆，中国，400042

e-mail: wandifei@cmdi.chinamobile.com, chenjieshu@cmdi.chinamobile.com

【摘　要】随着互联网、通信行业的快速发展，大量的垃圾信息出现在人们生活中，如垃圾短信、垃圾邮件等。本文提出了一种基于最优分割策略的两步垃圾文本信息过滤方法，该方法将最优分割策略引入到两步分类方法之中，利用遗传算法在两步分类方法中分割线临近的文本不可靠区间内搜索最优文本分割线，第一步以字 BI-GRAM 为特征，免除繁琐的分词过程，提高效率，第二步以词语 BI-GRAM 为特征，从而使分类器达到最佳性能。在由 12600 篇文本构成的中文语料数据集上的实验表明，该方法具有较高的分类性能和效率，精确率、召回率和 F1 值分别达到 98.55％，97.23％和 97.89％，可以有效过滤垃圾信息。

【关键词】两步分类；朴素贝叶斯；遗传算法；最优分割策略

1 引言

随着互联网技术、通信技术的飞速发展和手机普及率的迅速提高，爆炸般增长的垃圾信息侵扰了人们的生活，危害着国家安定团结，例如垃圾短信、垃圾邮件等不良信息。文本分类技术可以对这些垃圾文本信息进行监控和过滤，文本分类就是在给定的分类体系下根据文本的内容确定与其相关联的类别标签，是基于内容的自动信息管理的核心技术。国内外采用统计方法和机器学习技术对文本分类及其相关领域进行了较为深入的研究[1]，典型的文本分类方法有朴素贝叶斯分类器[2]、基于向量空间模型的分类器[3]、基于实例的分类器[4]和用支持向量机建立的分类器[5,6]等。文本过滤（即两类文本分类）是文本分类的基本类型，广泛用于不良信息过滤等现实任务，例如网络、邮件及手机短信中的色情信息和垃圾信息过滤。很显然，不良信息必须得到有效控制；同时使用现有方法的过滤性能往往不够理想。原因在于在这类任务的文档集

资助信息：国家自然基金项目《基于特征联想的中文短文本分类方法研究》（编号：60703010）

978-1-935068-06-8 © 2009 SciRes.

中存在许多具有高度模糊的文本，即在两类文本的边界处存在一个模糊区域。在此区域内，某一类的文本与另一类的文本由于使用了一些相同的词语而具有很大的相似度。例如，黄色信息过滤中， love story 和 sex 往往同时出现在正例和负例中。

樊等观察到了这类文本错误分类的特点，提出了基于朴素贝叶斯分类器的两步分类方法[7, 8, 9]：第一步，以词性为动词、名词、形容词和副词的词语为特征，以朴素贝叶斯分类器分类文本，并直接根据分类器输出确定两类间的模糊区域；第二步，以词性为动词、名词的词语二元串为特征，以类似于前一步使用的朴素贝叶斯分类器对模糊区域内的文本进行二次分类。两步分类方法[7, 8, 9]存在一个不足：分类效率不高，这严重制约了该方法应用于大规模真实文本地实时在线处理。原因在于该方法需要对中文文本进行分词以抽取所需的特征，而目前中文分词系统的速度普遍不高。此外，两步分类方法是直接利用分类器的输出构造一个二维空间，将分类器变换为在此空间中的一条分割线，根据错误分类文本的分布特点（即大多数错误分类的文本聚集在分割直线附近的两侧）来确定模糊区域。本文对错误文本的分布进行了重新观察，受支持向量机中最优分割面思想的启发，发现由贝叶斯分类器直接变换来的分割直线不是一条最优的分割线。

本文提出了一种基于最优分割策略的两步垃圾文本信息过滤方法，在原有两步框架内第一步使用字的二元串作为特征，以提高分类效率。利用遗传算法在构造的二维文本空间中获得一条新的具有较高分类能力的分割直线，构造最优分割分类模型作为两步分类框架中的分类器。通过实验对所提出的方法进行了较为充分的评估。

2 基于最优分割策略的两步垃圾文本信息过滤方法

通过改写两类朴素贝叶斯分类模型来构造一个二维空间，用于观察错误文本分类的分布，并据此固定两步分类所需的模糊区域。

2.1 基于朴素贝叶斯模型的二维空间构造

给定二值文本向量 d=(W1 ,W2 , …,WM)，Wi＝0 或者 1，如果第 i 个特征出现在文本中，Wi＝1，否则 Wi＝0。令 pki=P{Wk=1|ci}，Pr{·}表示求事件{·}发生的概率。

两类朴素贝叶斯分类器的判别函数可表示为：

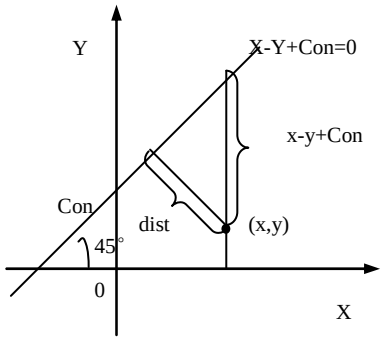

Figure 1. Text-point distance calculation

图 1. 文本点距离计算

$$f(d)=\log\frac{P\{c_1|d\}}{P\{c_2|d\}}=\log\frac{P\{c_1\}}{P\{c_2\}}+\sum_{k=1}^{M}\frac{\log(1-p_{k1})}{\log(1-p_{k2})}+\sum_{k=1}^{M}W_k\log\frac{p_{k1}}{1-p_{k1}}-\sum_{k=1}^{M}W_k\log\frac{p_{k2}}{1-p_{k2}} \quad (1)$$

当 f(d)≥0 时，文本 d 属于类型 c1，否则属于类型 c2。令

$$Con=\log\frac{P\{c_1\}}{P\{c_2\}}+\sum_{k=1}^{M}\frac{\log(1-p_{k1})}{\log(1-p_{k2})} \quad (2)$$

$$X=\sum_{k=1}^{M}W_k\log\frac{p_{k1}}{1-p_{k1}} \quad (3)$$

$$Y=\sum_{k=1}^{M}W_k\log\frac{p_{k2}}{1-p_{k2}} \quad (4)$$

Con 只与所采用的训练样本集有关，不随文本 d 的变化而变化，为常数；X 表示根据特征估本 d 属于类型 c1 的测度；Y 表示根据特征估算出来的文本 d 属于类型 c2 的测度，则公式(1)可改写为：

$$f(d)=X-Y+Con \quad (5)$$

公式(5)表示两类朴素贝叶斯分类器可看作是在由 X 和 Y 构成的二维空间中寻求一条分割直线 f(d)=0。这样，利用式(3)和(4)，可将文本表示为二维空间中的一个点(x,y)，该点到分割直线 f(d)=0 的距离 dist 为：

$$dist=\frac{1}{\sqrt{2}}(X-Y+Con) \quad (6)$$

如图 1 所示。当 dist≥0 时，表示文本 d 属于类型

c1；当 dist<0 时，表示文本 d 属于类型 c2。

我们将公式(1)改写为公式(5)，再演变为公式(6)的目的是：1) 利用公式(6)可以在由 X 和 Y 构成的二维空间中方便地考察、分析文本分类错误，探讨在给定分类方法和文本特征集的条件下，距离 dist 与分类错误的关系；2) 利用公式(6)可以根据距离 dist 的大小方便地评估分类的可靠程度。

2.2 二维空间中错误分类的文本观察及模糊区间的固定

以第 3 节实验中所使用的语料为样本，以 X 为横坐标，以 Y 为纵坐标，文本点在二维空间中的分布情况如图 2 所示。从图中可以看出，在二维空间中两类文本以条带形状分布在分割直线的两边，被错误分类的文本（即位于分割直线下方的整例文本，和位于分割直线上方的负例文本）到分割直线的距离都很近。根据观察，可将由 X 和 Y 构成的二维平面分成可靠和不可靠两个区域。根据式(7)进行分类判别。

$$\begin{cases} Dist2 \le dist \le Dist1, & \text{对文本d的任何分类决策都是不可靠的} \\ dist > Dist1, & \text{文本d属于类型}c_1\text{且分类结果可靠} \\ dist < Dist2, & \text{文本d属于类型}c_2\text{且分类结果可靠} \end{cases} \quad (7)$$

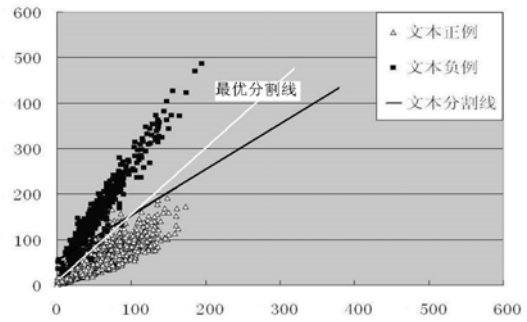

Figure 2. The distribution of the text points

图 2. 文本点的分布情况

由图 2 中可以看出，原有的分割线（黑）不是一条最优分割线，如果在不可靠区域内将其调整为如图所示的最优分割线（白），则分类器的分类能力将提高。

2.3 最优分割分类模型(O-Naive)

最优分割线可以通过在二维文本空间中不可靠区域内对原有分割直线的平移和旋转来实现. 即在改写的判定函数公式(5)中，加入了 β 和 α 两个参数，将其改写为公式(8)。

$$f(d)=\alpha*X-Y+\beta*Con=0 \quad (8)$$

公式(8)中，β 用于对原有分割直线的平移，而 α 用于对原有分割直线的旋转。利用公式(8)，我们可以通过遗传算法在文本分布的不可靠区域对参数 β 和 α 的搜索来确定最优分割直线。

对由公式(3)和(4)确定的二维空间文本点(x, y)，其到由公式(8)确定的最优分割直线的距离 Dist 为：

$$dist=\frac{1}{\sqrt{2}}(\alpha*x-y+\beta*Con) \quad (9)$$

利用公式(7)可固定最优分割分类模型的模糊区域。

2.4 利用遗传算法寻求最优分割直线

2.4.1 遗传算法基本原理

遗传算法(GA)是一种基于自然选择和遗传变异等生物进化机制的全局性概率搜索算法。与基于导数的解析方法和其他启发式搜索方法(如爬山方法，模拟退火方法，Monte Carlo 方法)一样，进化算法在形式上也是一种迭代方法。它从选定的初始解出发，通过不断迭代逐步改进当前解，直到最后搜索到最优解或满意解；在进化计算中，迭代计算过程采用了模拟生物体的进化机制，从一组解(群体)出发，采用类似于自然选择和有性繁殖的方式，在继承原有优良基因的基础上，生成具有更好性能指标的下一代解的群体。由于 GA 在问题空间搜索最优值所表现的优良特性，我们考虑将 GA 引入到基于最优朴素贝叶斯分类模型中以确定阈值 β 和 α。

2.4.2 染色体表示

利用遗传算法在二维文本空间不可靠的区域对阈值 β 和 α 进行选择。阈值 β 和 α 的取值范围与二维文本空间中不可靠的区域的范围有关，

$$\text{若 } Con>0, \text{则} \beta \in \left(1-\sqrt{2}*\frac{|Dist2|}{Con}, 1+\sqrt{2}*\frac{|Dist1|}{Con}\right) \quad (10)$$

978-1-935068-06-8 © 2009 SciRes.

$$若Con<0,则\beta\in\left(1+\sqrt{2}*\frac{|Dist1|}{Con},1-\sqrt{2}*\frac{|Dist2|}{Con}\right) \quad (11)$$

$$若Con=0,则\beta=0 \quad (12)$$

在二维文本空间不可靠的区域内，文本分割线与 X 轴夹角的范围理论上可取 0 度－90 度，这里我们取阈值α经验值范围为：$\alpha=(0.36,2.75)$，即约为 20 度－70 度之间。

阈值β和α是取值在一定范围之内的实数，可以视为遗传算法的表现型形式，从表现型到基因型的映射称为编码。我们采用二进制编码形式，将β和α变量值代表的个体表示为一个{0, 1}二进制串，当然，串长取决于求解的精度。例如：求解的精度精确到 3 位小数, 区间长度为 0. 5，必须将区间分为0.5×10^3等份。因为$256=2^8<0.5\times10^3<2^9=512$，所以编码的二进制串长至少需要 9 位。

2.4.3 适应度函数

文本分类中有三个主要的性能、效率评估指标：精确率、召回率和 F-measure。用遗传算法对阈值β和α的搜索时用 F1 作为适应度函数，F1 值越大证明该分类器的分类性能越好。

精确率（Precision，P）

$$P=\frac{正确分为某类的文本数}{数据集中分为该类型的文本总数}\times100\%$$

准确率是所有输入系统进行分类处理的文本中与专家分类结果完全吻合的文本所占的比率，P 描述了分类结果中的准确程度，即分类结果中有多少是正确的。

召回率（Recall，R）

$$R=\frac{正确分为某类的文本数}{数据集中属于该类型的文本总数}\times100\%$$

召回率是人工分类结果应有文本中分类系统分类正确的文本所占的比率。R 描述了正确分类的能力，即已知的文本中，有多少分类正确。

F-measure

对于一次测试，准确率和召回率一般是成反比的。提高准确率，召回率会下降；提高召回率，准确率会下降。F-measure 综合了 P 和 R 两个指标，可以对分类器进行整体评价。

$$F_{\beta_i}=\frac{(\beta^2+1)\cdot P_i\cdot R_i}{\beta\cdot P_i+R_i}$$

其中：β是调节 P 和 R 相对重要程度的常数，当$\beta=0$时，F-measure 为准确率；当$\beta\to\infty$时，为召回率。通常取$\beta=1$，对 P 和 R 平等看待，这时得到最常用的 F1-measure 指标（简称 F1）。

$$F_1=\frac{2\times P\times R}{P+R}$$

2.4.4 遗传操作设定

图 3 所示，为生成子代种群的方法。首先把当代种群的染色体从优到劣进行排序，然后选择一定比例的下位个体淘汰掉，淘汰比例一般设为 40%，在上位个体中实行均匀交叉，生成的子个体填补到种群中，以保持种群规模不变，最后按照设定的变异概率实行变异操作，生成子代种群。

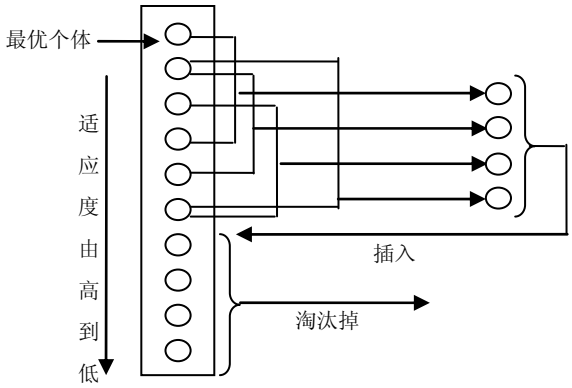

Figure 3. Individual offspring generation
图 3. 子代个体生成

2.5 两步分类方法

方法 1（词语+词 bi-gram +Naive）：第一步，以词语为特征，以朴素贝叶斯分类器（Naive）分类文本，并直接根据分类器输出确定两类间的模糊区域（利用(6)计算 Dist, (7)固定模糊区域）；第二步，以词性为动词、名词的词语二元串为特征，以类似于前一步使用的朴素贝叶斯分类器对模糊区域内的文本进行二次分类。

方法 2 （字 bi-gram+词 bi-gram +Naive）：第一步，以字二元串（字 bi-gram）为特征，以朴素贝叶斯分类器（Naive）分类文本，并直接根据分类器输出确定两类间的模糊区域（利用(6)计算 Dist, (7)固定模糊区域）；第二步，以词性为动词、名词的词语二元串为特征，以类似于前一步使用的朴素贝叶斯分类器对模糊区域内的文本进行二次分类。

方法 3（字 bi-gram+词 bi-gram +O-Naive）：该方法与方法 2 类似，其区别在于第一步和第二步中使用的分类器为式(8)确定的优化朴素贝叶斯分类器，第一步中利用(9)计算 Dist。

方法 1 是[7]中所使用的方法，方法 2 是对[7]中两步方法的一次改进，它在第一步中不需要分词，第二步需要分词但需要处理的文本很少，因此能够提高分类的效率。方法 3 是对方法 2 的又一次改进，由于使用了最优分割分类模型，其分类性能将进一步提高。

3 实验

为了评估本文提出的方法，分别在[9]中的中文数据集，对三种两步分类方法进行了对比实验。

3.1 实验数据集

本文用于实验的，中文数据集收集文本共 12600 篇，其中宣扬、传播对国家安全有害内容的文本为 1800 篇，它们构成属于类型 c1 的文本集；揭露、批判这种对国家安全有害内容的文本为 3716 篇，它们构成属于类型 c12 的文本集；内容与那些对国家安全有害的内容完全不同，但它们使用的词语中有相当部分是相同的文本为 828 篇，它们构成属于类型 c22 的文本集；其他文本为 6256 篇，它们构成属于类型 c32 的文本集；文本集 c12 、c22 和 c32 共同构成属于类型 c2 的文本集，共 10800 篇。为了模拟现实环境中两类文本出现的实际情况，属于类型 c1 和属于类型 c2 的文本数目比例为 1:6。将属于类型 c1 和属于类型 c2 的文本集随机地平均分为四份，以其中的一份构成测试集，另外的三份构成训练集, 按四栏进行交叉验证，以四栏实验的平均值作为最终的性能指标。

3.2 分类性能、分类效率评估指标

对文本分类的性能采用如下三种指标进行评估：

精确率(Precision)：

$$P=\frac{\text{正确分为某类的文本数}}{\text{数据集中分为该类型的文本总数}}\times 100\%$$

召回率(Recall)：

$$R=\frac{\text{正确分为某类的文本数}}{\text{数据集中属于该类型的文本总数}}\times 100\%$$

F1 值：

$$F_1=\frac{2\times P\times R}{P+R}$$

对文本分类的效率用分类所耗费时间来进行评估。实验所用 PC 配置如下：

CUP：Intel Pentium4 3.0，内存：DDRII533 1G。

3.3 实验参数配置

特征选择：采用了改进的互信息公式[11] (13)进行特征选择。

$$MI_1(t_k,c_i)=\sum_{i=1}^{n} P\{t_k,c_i\}\log\frac{P\{t_k,c_i\}}{P\{t_k\}P\{c_i\}} \qquad (13)$$

上面等式中，tk 表示第 k 个特征，文中为一个词语或者一个二元词语串；ci 表示文本预定义类型中的第 i 个类型，文中 i 取值为 1 或者 2,类型意义定义同引言中的定义；MI(tk,ci)表示特征 tk 和类型 ci 之间的互信息；P{tk}表示特征 tk 发生的概率；P{ci}表示类型 ci 发生的概率；P{tk,ci}表示特征 tk 和类型 ci 共现的概率。

分词系统：选用清华大学开发的 CsegTag3.0 进行中文分词，以抽取所需特征。

特征集规模确定方法：以所使用的特征数为横坐标，以分类性能（F1）为纵坐标，划出分类型能随特征数变化曲线，选择曲线上第一个拐点附近对应的特征数为特征规模集。

参数 Dist1 和 Dist2 的确定方法：以距离 Dist 为横坐标，以[7]中定义的错误率为纵坐标，划出错误率随特征数变化曲线，参照[7]中的方法确定 Dist1 和 Dist2。遗传算法参数：确定阈值β和α时，遗传算法初始参数设定参考经验值，这些经验在一定程度上具有一定的代表性，如表 1 所示。

Table 1. Genetic algorithm parameters
表 1. 遗传算法参数

种群大小	30
染色体变异概率	0.1
染色体淘汰概率	0.4
最大世代数	240
精度要求	3
种群大小	30
染色体变异概率	0.1
染色体淘汰概率	0.4

3.4 实验结果

由于测试集中，属于类型 c1 和属于类型 c2 的文档比例为 1:6,如果将所有文本都标记为 c2, 类型 c2 的分类精度也能达到 85.71%。因此在只给出了类型 c1 的分类结果，如表 2 所示。

978-1-935068-06-8 © 2009 SciRes.

Table 2. Performance comparison of three classification
表 2. 三种分类方法的性能比较

序号	1	2	3
第一步	词	字 BI-GRAM	字 BI-GRAM
第二步	词 BI-GRAM	词 BI-GRAM	词 BI-GRAM
分类器	Naive	Naive	o-Naive
精确率（%）	97.19	97.65	98.55
召回率（%）	93.94	97.00	97.23
F1 值（%）	95.54	97.31	97.89
特征数（个）	500+3000	800+8500	800+8300
分词耗时	00:49:21	00:19:48	00:18:54

实验结果表明，以 BI-GRAM 为特征的，基于最优分割策略的两步垃圾文本信息过滤方法(方法 3)具有最好的过滤效率和性能，比原有的两步分类方法 F1 值提高了 2.35 个百分点。

4 结论

本文提出一种基于最优分割策略的两步垃圾文本信息过滤方法，该方法采用字 BI-GRAM 为第一步使用特征，词 BI-GRAM 为第二步使用特征，最优分割分类模型为分类器的改进两步分类方法。实验表明，该方法具有高的分类性能和效率，可以对垃圾文本信息进行很好的过滤。随着视频、音频以及彩信等信息承载媒体的大量涌现，如何有效过滤这些垃圾信息将是我们下一步的工作。

致谢

本课题在选题及研究过程中得到王老师、樊老师的悉心指导。樊老师多次询问研究进程，并为我指点迷津，帮助我开拓研究思路，精心点拨、热忱鼓励。王老师一丝不苟的作风，严谨求实的态度，踏踏实实的精神，不仅授我以文，而且教我做人，虽历时三载，却给以终生受益无穷之道。对王老师、樊老师的感激之情是无法用言语表达的。

最后，向我的家人致谢，感谢他们对我的理解与支持。

References (参考文献)

[1] Sebastiani, F. Machine Learning in Automated Text Categorization. ACM Computing Surveys, Vol.34, No.1, March 2002, pp. 1-47.

[2] Lewis, D. D. Naive Bayes at forty: The independence assumption in information retrieval. In Proceedings of ECML-98, 10th European Conference on Machine Learning, Chemnitz, Germany, 4-15. 1998.

[3] Salton, G. Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley，Reading, MA, 1989.

[4] Mitchell，T.M. Machine Learning. McCraw Hill, New York, NY, 1996.

[5] Joachims, T. Text categorization with Support Vector Machines: Learning with many relevant features. In Machine Learning: ECML-98, Tenth European Conference on Machine Learning, pp. 137--142.1998.

[6] Yang, Y. and Liu, X. A re-examination of text categorization methods. In Proceedings of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval, Berkeley, CA, 42-49,1999.

[7] Xinghua Fan, Maosong Sun, Key-sun Choi, and Qin Zhang. Classifying Chinese texts in two steps. IJCNLP2005, LNAI3651, pp.302-313, 2005.

[8] Xinghua Fan. A High Performance Prototype System for Chinese Text Categorization. A. Gelbukh and C.A. Reyes-Garcia (Eds.): MICAI 2006, LNAI 4293, pp. 1017 – 1026, 2006.

[9] Xinghua Fan, Difei Wan, and Guoyin Wang. Combining Bi-gram of Character and Word to Classify Two-Class Chinese Texts in Two Steps. S. Greco et al. (Eds.): RSCTC 2006, LNAI 4259, pp. 597 – 606, 2006.

[10] Sahami, M., Dumais, S., Hecherman, D., Horvitz, E. A Bayesian Approach to Filtering Junk E-Mail. In Learning for Text Categorization: Papers from the AAAI Workshop, pp. 55-62, Madison Wisconsin. AAAI Technical Report WS-98-05, 1998.