第 35 卷 第 6 期 计 算 机 工 程 2009 年 3 月
Vol.35 No.6 Computer Engineering March 2009
·软件技术与数据库· 文章编号：1000—3428(2009)06—0088—03 文献标识码：A 中图分类号：TP391

# 一种适应短文本的相关测度及其应用

何海江
(长沙学院计算机中心，长沙 410003)

**摘 要：**针对博客社区和 BBS 论坛充斥 Web 垃圾信息的问题，提出相关度向量空间模型 cVSM，并以此作为评论的特征，采用支持向量机分类算法自动识别垃圾评论。cVSM 包括一种适合短文本的相关测度，用于衡量评论和文章的语义相关程度。在中文博客测试集和中文 BBS 测试集上的实验结果表明，相比纯粹使用评论文本特征的方法，应用该模型时 F1 至少提高 6%。

**关键词：**博客；垃圾评论；支持向量机；文本挖掘；相关测度

# Relevancy Coefficient and Its Application Adapted to Short Texts

HE Hai-jiang
(Computer Center, Changsha University, Changsha 410003)

**【Abstract】** A relevancy coefficient vectort space model named cVSM is proposed to aim at Web spams which flood in blogosphere and forums. The cVSM whose components are employed as features of comments and the support vector machine classification algorithms are used to automatically identify comment spams. The relevancy coefficient included in the cVSM is presented, which is used to evaluate relevancy grade of posts and comments. Chinese blog dataset and Chinese BBS dataset are tested. Experimental results show that compared with traditional method the F1 has been improved at least 6% by this way.

**【Key words】** blog; comment spam; support vector machine; text mining; relevancy coefficient

## 1 概述

博客和 BBS 论坛都是互联网上的代表性应用。博客作者在博客上发表文章，记录个人的日常生活及事务，抒发感情、释放情绪，发表对新闻事件及人物的看法和意见等。一般情况下，博客作者都允许来访的读者对其文章发表评论，实现有效交流。BBS 上热门帖子(文章)也常常吸引众多跟帖(评论)。然而许多人却并非出于交流目的，而是发表一些与文章无关的言论，或者与文章相关但内容不良，这些评论统称为垃圾评论。综合来看，垃圾评论主要有这样几类：涉及产品推销、网站推介、公司宣传等广告信息；在评论中插入超链接，提高其链接指向目标 Web 地址在商业搜索引擎的评分；不顾别人的感受，散布谩骂、下流、人身攻击的言论。

垃圾评论不仅降低网络服务质量，还影响读者的情绪。不清除垃圾评论，博客社区和 BBS 将是不健康的。因此，提出一种文章的相关测度，作为评论文档的一个特征，再结合支持向量机分类算法来识别垃圾评论。

## 2 相关研究工作

依据向量空间模型，文档被看作以词或词权组成的向量，两文档的相关度，即语义关联程度，等于 2 个向量的相似系数。基于 TF×IDF 的内积、余弦值[1-2]等传统相关测度在搜索引擎和 Web 信息检索系统得到广泛运用。然而，这些相关度并不适合评论。

垃圾评论的识别可看作二值(合法评论、垃圾评论)文本分类问题。文本分类技术在垃圾邮件、垃圾博客[3]、Web 文档分类等领域受到广泛研究。其主要内容是：先收集一些带标记的 Web 对象作为训练样本，再找出显著特征，最后运用贝叶斯、KNN、支持向量机等算法分类。

博客和 BBS 的研究大多聚焦于文章，研究评论的报道较少，已有的讨论只限于识别带超链接的垃圾评论[4]。本文的算法能够识别各种类型的垃圾评论。

## 3 评论和文章的相关测度

在垃圾评论识别中，搜索对象为长文本(文章)，待检对象为短文本(评论)。本文的相关测度也基于 TF×IDF 模型，与内积、余弦值等相比，更适合待检对象为短文本的情形。表 1 是相关度模型使用的符号。

**表 1 相关度模型符号**

| 符号 | 含义 |
|---|---|
| $q$ | 搜索文档，对应博客文章 |
| $d$ | 待检文档，对应评论 |
| $N$ | 文档集包含的文档篇数 |
| $f_w$ | 词的文档频率，文档集包含词 $w$ 的文档篇数 |
| $f_d$ | 文档 $d$ 包含的词个数 |
| $f_{d,w}$ | 文档 $d$ 包含词 $w$ 的个数 |

内积 *InnPr* 是常见的相关测度，下式内的词频表达文档 $d$ 中词的重要性，*IDF* 为词的反文档频率。为避免词频等于 1 的词乘积为 0，对数内加上 1。

$$InnPr(q,d)=\sum_{w\in q\cap d}\ln(1+f_{d,w})\times IDF_w \quad (1)$$

文本越长，$q\cap d$ 和 $f_{d,w}$ 越大。因此，以 *InnPr* 值为分类阈值，文本长的评论将更倾向于被判为合法评论；而 Web 社区中合法评论往往较短，垃圾评论往往较长。显然，将 *InnPr* 除以 $\sqrt{f_d}$ 后得到的 *InnPrVar* 更合适，较短的评论比长文本评论更可能被判为合法评论。

**基金项目：**长沙学院科研基金资助项目(CDJJ-07010110)
**作者简介：**何海江(1970－)，男，副教授，主研方向：Web 挖掘，数据仓库
**收稿日期：**2008-06-23 **E-mail：**haijianghe@sohu.com

下式的 *Cosine* 是被广泛采用的相关测度[1-2]，2 个文档越相似，向量的夹角越小，*Cosine* 值越大。

$$Cosine(q,d)=\frac{\sum_{w\in q\cap d}\ln(1+f_{d,w})\times IDF_w\times\ln(1+f_{q,w})\times IDF_w}{\sqrt{\sum_{w\in q}\ln^2(1+f_{q,w})\times IDF_w}\times\sqrt{\sum_{w\in d}\ln^2(1+f_{d,w})\times IDF_w}} \quad (2)$$

*Hoad* 和 *Zobel* 为判断文档之间是否存在抄袭和复制[5]，提出了一种如下式的相关测度 *HoadZ*，适合于搜索文档和待检文档文本长度差不多的情形。

$$HoadZ(q,d)=\frac{1}{1+\ln(1+\left|f_d-f_q\right|)}\sum_{w\in q\cap d}\frac{N}{f_w}\times\frac{1}{1+\left|f_{d,w}-f_{q,w}\right|} \quad (3)$$

*Jaccard* 是一种常见的相似性测度，稍作变化，成为下式的 *Sim* 也可用于度量评论和文章的相关程度。

$$Sim(q,d)=\frac{Number\ of\ q\cap d}{Number\ of\ d} \quad (4)$$

垃圾评论往往要发布与文章无关的内容，内容越多，也就是评论的信息熵($IDF_w$ 总和)越大，越能引起读者的注意。相对于垃圾评论来说，合法评论与文章相同的词语则更多。因此，相关度 *CorrPC* 定义为下式，分母表示评论的信息熵，分子表示两者相关内容的信息熵。去除停用词后，*d* 可能为空集，为避免除数为 0，分母加上一个平滑系数 *alpha*，显然有 $0\leq CorrPC<1$。

$$CorrPC(q,d)=\frac{\sum_{w\in q\cap d}\ln(1+f_{d,w})\times IDF_w}{alpha+\sum_{w\in d}\ln(1+f_{d,w})\times IDF_w} \quad (5)$$

相关度反映文章和评论的语义相关程度。而在向量空间模型，只有相同词语才认为相关，语义相近，甚至是同义词都认为无关，误差无法避免。当评论与文章没有相同词时，为了使文本短的评论更倾向于被判为合法评论，在式(5)的分子中也加上 *alpha*，成为 *CorrPCVar*，有 $0<CorrPCVar\leq 1$。

## 4 文档表示与文本分类算法

评论不单具有普通文本的特点，它还与相应文章语义关联。同样一篇评论，对文章 A 来说是合法评论，评论文章 B 时则可能变为垃圾。基于向量空间模型(VSM)，提出一种扩展 VSM 的相关度向量空间模型(cVSM)。每一篇评论文档定义为$(\theta,\omega_1,\omega_2,\cdots\omega_n)$，$\theta$ 是评论与文章的相关度，$\omega_i$ 是第 *i* 个词的权重。将评论 *d* 包含第 *i* 个词 *w* 的个数 $f_{d,w}$ 简记为 $df_i$，词权可定义为 $df_i\times IDF_i$，考虑到文本长度不一，采用归一化的词权。

*c* 支持向量机是一种非常有效的分类算法，使用结构风险最小化构造决策超平面，具有很好的泛化能力，在文本分类中得到广泛应用[6]。对两类问题，给定一系列样本 $T=\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2),\cdots,(\boldsymbol{x}_d, y_d)\}$。其中，$\boldsymbol{x}_i\in R^n$ 是输入向量，$y_i\in\{+1,-1\}$ 是类标签，$i=1,2,\cdots,d$；学习问题的目标为

$$\min_{w,b,\xi}\frac{1}{2}w^{\mathrm{T}}w+C_+\sum_{y_i=1}\xi_i+C_-\sum_{y_i=-1}\xi_i$$

使得 $y_i(w^{\mathrm{T}}\sigma(x_i)+b)\geq 1-\xi_i$ 和 $\xi_i\geq 0, i=1,2,\cdots,d$。其中，$C_+$和 $C_-$分别为正类和负类的惩罚因子。本文讨论的焦点是使用 cVSM 比 VSM 作为评论的文档模型时，分类性能显著提高，参数和核函数的选择并不影响结论，实验过程取 $C_+=C_-=1.5$，核函数用线性函数。实际上，改变参数 $C_+$和 $C_-$，或者将线性核替换为高斯核，所取得的结果相类似。采用 LIBSVM 提供的算法完成实验。

召回率 *Recall*、准确率 *Precision*、*F*1 是常用的分类器有效性评价指标，令被正确判为垃圾的评论数为 $N_{stos}$、数据集实际存在的垃圾评论数为 $N_{spam}$、被分类器判为垃圾的评论数为 $N_{mods}$，有

$$Recall=\frac{N_{stos}}{N_{spam}},\ Precision=\frac{N_{stos}}{N_{mods}},\ F1=\frac{2\times Precision\times Recall}{Precision+Recall}$$

## 5 实验结果及分析

与报纸、杂志等媒体不同，Web 社区的文本语法并不严谨，并且有许多的网络用语，因此，计算博客社区和 BBS 的 IDF 很有必要。笔者编写了一个爬虫程序，从博客网站和 BBS 论坛随机下载了大量文章(以博客文章为主)，涵盖教育、娱乐、科技、财经、体育等各个方面。每篇文章不少于 150 个字，筛选后得到 $N$=141 423 篇。使用中科院计算所汉语分词软件 ICTCLAS，按 $idf_w=\ln(N/f_w)$计算每个词的反文档频率。将文档词频小于 10 的词 $f_w$ 一律设为 3($f_w<10$ 的词的词频中位值)，只保留 85 601 个词及其 *idf*。这样做，不仅提高计算速度，也可部分消除由于文章采样不均衡引入的噪声。不再保留词集里的词，其 *idf* 统一为最大值 $\max(idf)=\ln(N/3)$。IDF 为 *idf* 的归一化形式：$IDF_w=idf_w/\max(idf)$。将 $f_w$ 超过 $2/3\times N$ 的词归类为停用词，如“的”、“是”等。后文不再赘述，所有相关测度的计算过程中，都先消除文章和评论的停用词。

实验分为 5 个步骤：数据集构造，相关测度比较，文本特征选择，分类模型训练和分类测试。

### 5.1 数据集构造

笔者构造了 3 个数据集，人工标注所有评论。样本集 BfTrain，下载了 169 篇文章，2 584 条评论，其中垃圾评论 818 条。博客测试集 BlogSet，包含 60 篇文章，1 135 条评论，其中垃圾评论 377 条。BBS 论坛测试集 BBSSet，包含 22 篇文章，235 条评论，其中垃圾评论 70 条。

### 5.2 相关测度的比较

在 BfTrain 上比较相关测度，先剔除一些特殊评论：与文章语义相关，但内容不良，以“垃圾”、“浪货”等侮辱人的词语为主的垃圾评论；与文章语义无关，实为合法评论，以“沙发”、“顶”、“踩”等网络习语组成。剩下 2 324 条评论，其中 673 条垃圾评论。

*Cosine*, *Sim*, *CorrPC*, *InnPrVar* 和 *CorrPCVar* 的值在[0,1]之间，[0,30]和[0,1]能覆盖绝大部分评论的 *HoadZ* 和 *InnPr*，将各区间划分为 100 等份，分别以这些等份值为阈值，相关度超过阈值的为合法评论，否则为垃圾评论。如此分类，取得使 *F*1 最大的相关度阈值 *corr*，并计算召回率和准确率。表 2 是 *F*1 最大时各相关测度的性能比较，*InnPrVar* 和 *CorrPCVar* 的参数 $alpha=1/\max(idf)=0.214$。

**表 2 相关测度比较**

| 参数 | *Recall* | *Precision* | *F*1 | *corr* |
|---|---|---|---|---|
| *InnPr* | 44.2 | 80.2 | 57.0 | 0.13 |
| *Cosine* | 51.4 | 92.3 | 66.1 | 0.03 |
| *HoadZ* | 34.0 | 94.2 | 50.0 | 17.70 |
| *Sim* | 51.2 | 89.2 | 65.0 | 0.25 |
| *CorrPC* | 59.3 | 86.8 | 70.4 | 0.12 |
| *InnPrVar* | 52.5 | 78.6 | 62.9 | 0.15 |
| *CorrPCVar* | 80.9 | 81.0 | 80.9 | 0.13 |

Web 社区的合法评论往往较短，广告式垃圾评论往往较长。文章一般远长于评论，有 $f_q\gg f_d$，短文本的 *HoadZ* 值偏小，更容易被判为垃圾评论，自然 *HoadZ* 性能最差。许多广告式垃圾评论为避免被删除，故意拷贝文章的词句到评论，这样做使得 *InnPr* 增大；但垃圾评论作者始终要突出其广告内容，故而文章相关内容的信息熵小于广告内容的信息熵，

*Sim*, *Cosine* 和 *CorrPC* 都考虑到这一点。文章内很少出现广告内容，广告词的的 IDF 往往很大，*Sim* 却不考虑 IDF，表现也就变差了；*Cosine* 尽管考虑了 IDF，但却引入了文章的信息熵；*CorrPC* 兼顾多个方面，明显比前几个测度合理。*InnPrVar* 和 *CorrPCVar* 倾向于将短文本的评论判为合法评论，自然的，*InnPrVar* 比 *InnPr* 表现好，*CorrPCVar* 比 *CorrPC* 好。综合来看，以 *CorrPCVar* 值为阈值时，性能最优。

表 3 是 *alpha* 对相关测度的影响，*alpha* 从 0.1~1.0，*CorrPC* 的 *F*1 变差，而 *CorrPCVar* 的 *F*1 先增强后变弱，在 0.4 处最优。不少评论与文章没有或很少相同的词语，导致计算 *CorrPC* 时分子为 0 或很小，*alpha* 越大，*CorrPC* 则越小；而 *alpha* 对 *CorrPCVar* 的分子分母都有影响。改变 *alpha*，并不影响文章的结论，计算相关测度时，*alpha*=0.214，后文不再说明。

**表 3　*alpha* 对 *CorrPC* 和 *CorrPCVar* 的 *F*1 的影响**

| *alpha* | *CorrPC* | *CorrPCVar* |
|---|---|---|
| 0.100 | 70.9 | 77.6 |
| 0.200 | 70.4 | 80.5 |
| 0.214 | 70.4 | 80.9 |
| 0.300 | 69.7 | 82.6 |
| 0.400 | 69.5 | 83.4 |
| 0.500 | 69.0 | 83.1 |
| 0.600 | 68.7 | 83.0 |
| 0.800 | 67.8 | 82.0 |
| 1.000 | 67.4 | 81.2 |

### 5.3　文本特征选择和分类模型训练

基于支持向量机的文本分类算法中，文档频率 *DF* 和信息增益 *IG* 是常用的特征筛选指标。本文采用 *IG* 和 *DF*+*IG*[6] 2 种方法，后者先将 *DF* 小于阈值的词删除，再计算剩余词的 *IG*，可减少噪声词对分类算法的影响。

在 *BfTrain* 全部评论上计算所有词的 *IG*，共有 12 734 个特征，按照从大到小排列。分别以前 4 000 个~12 734 个作为评论的文本特征，VSM 只包含纯粹文本特征，*Cosine*+*VSM* 则用 *Cosine* 值代入 cVSM 模型的 $\theta$，*CorrPC*+*VSM*, *CorrPCVar*+*VSM* 亦类似，而其他相关测度性能相对要差，则不再参与比较。按照 *DF*+*IG* 选择文本特征时，先将 *DF*<3 的词删除，计算剩余 4 443 个词的 *IG*，从大到小排列，分别以前 1 000 个~4 443 个词作为评论的文本特征。以 *CorrPC*+*VSM*+4 000 为例，计算每条评论的 *CorrPC* 及 4 000 个最大 *IG*(或 *DF*+*IG*)词的 $\omega_i$，$i$=1,2,⋯,4 000，用 *c* 支持向量机学习所有评论的 4 001 个属性，依此生成分类模型。

### 5.4　分类测试

使用 *IG* 选择特征后，再以生成的分类模型在 *BlogSet* 集上测试，表 4 是评价指标 *F*1 的比较。4 种文档表示模型的准确率相差不大，召回率则差距很大，*VSM*<*Cosine*+*VSM*<*CorrPC*+*VSM*<*CorrPCVar*+*VSM*。为节约篇幅，未列出准确率和召回率的比较。

**表 4　*BlogSet* 上 *IG* 特征选择时 *F*1 的比较**

| 特征数量 | *VSM* | *Cosine*+*VSM* | *CorrPC*+*VSM* | *CorrPCVar*+*VSM* |
|---|---|---|---|---|
| 4 000 | 56.88 | 59.68 | 71.88 | 87.48 |
| 5 000 | 50.10 | 55.12 | 69.78 | 87.05 |
| 6 000 | 71.64 | 76.61 | 81.76 | 89.01 |
| 7 000 | 72.58 | 78.78 | 81.62 | 89.93 |
| 8 000 | 72.04 | 82.49 | 83.86 | 89.88 |
| 9 000 | 78.00 | 82.89 | 85.09 | 90.47 |
| 10 000 | 77.91 | 83.53 | 85.01 | 89.52 |
| 11 000 | 80.89 | 84.09 | 86.89 | 90.69 |
| 12 000 | 80.71 | 85.88 | 87.57 | 91.20 |
| 12 734 | 80.35 | 86.00 | 87.75 | 91.23 |

随着相关度特征的加入，召回率和 *F*1 得到大幅提高。而在 3 种相关度中，以 *CorrPCVar* 表现最好，*CorrPC* 次之，这与 5.2 节相关测度的比较结果相吻合。将 *cVSM* 作为文档表示模型，显著提高了垃圾评论识别的能力。当文本特征数从 4 000~5 000 时，4 种文档表示模型的 *F*1 和召回率反而降低了，说明 *IG* 大小处于 4 000~5 000 之间存在一些噪声词，影响了分类模型。在 *BBSSet* 上的测试结果得出同样的结论，表 5 是召回率的比较。由于分类模型的训练集 *BfTrain* 以博客文章为主，*BBSSet* 上 *F*1 的性能相对 *BlogSet* 集略有降低，噪声词的影响也表现不同。

**表 5　*BBSSet* 上 *IG* 特征选择时 *Recall* 的比较**

| 特征数量 | *VSM* | *Cosine*+*VSM* | *CorrPC*+*VSM* | *CorrPCVar*+*VSM* |
|---|---|---|---|---|
| 4 000 | 35.29 | 34.29 | 52.86 | 80.00 |
| 5 000 | 27.54 | 35.71 | 47.14 | 77.14 |
| 6 000 | 37.14 | 48.57 | 60.00 | 81.43 |
| 7 000 | 41.43 | 47.14 | 60.00 | 78.57 |
| 8 000 | 45.71 | 52.86 | 60.00 | 78.57 |
| 9 000 | 48.57 | 54.29 | 58.57 | 81.43 |
| 10 000 | 51.43 | 54.29 | 60.00 | 77.14 |
| 11 000 | 52.86 | 57.14 | 62.86 | 80.00 |
| 12 000 | 54.29 | 55.71 | 65.71 | 82.86 |
| 12 734 | 51.43 | 55.71 | 65.71 | 81.43 |

使用 *DF*+*IG* 选择特征后所生成的支持向量机分类模型，在 *BlogSet* 和 *BBSSet* 上的测试结果与 *IG* 大致相同。图 1 是 4 种文档表示模型在 *BBSSet* 的 *F*1 比较，*Cosine*+*VSM* 简记为 *Cosine*+，*CorrPC*+和 *CorrPCVar*+亦如此。特征数少于 3 000 时，*VSM* 和 *Cosine*+都不足以分类垃圾评论，但 *CorrPCVar*+仍然表现良好，说明该相关度很好地反映了评论和文章的相关程度。

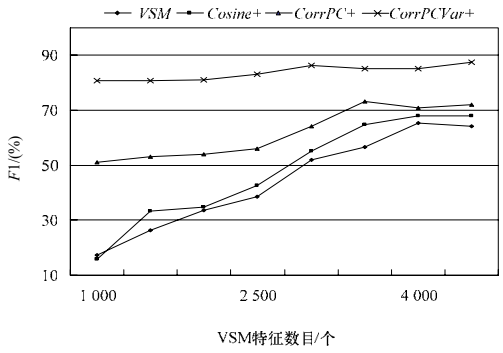


**图 1　*BBSSet* 上 *DF*+*IG* 特征选择时 *F*1 的比较**

当文本特征数从 4 000 到 4 443 时，*VSM* 和 *Cosine*+的 *F*1 下降，仍然是噪声词的影响，当然 *IG* 的前 4 000 个词与 *DF*+*IG* 的前 4 000 个词不同。文本特征非常稀疏，要建立一个噪声词少的分类模型，至少要 GB 级的训练集，人工标注将非常困难。此时，*CorrPC*+和 *CorrPCVar*+的 *F*1 却没有下降，并且当特征数从 3 500 到 4 000 时 *CorrPC*+的 *F*1 下降，当特征数从 3 000 到 3 500 再到 4 000 时，*CorrPCVar*+的 *F*1 也下降，这些都是因为相关度和 *VSM* 之间存在冗余，两者都基于 *TF*×*IDF*。相关度和 *VSM* 之间到底存在何种冗余，暂时还难以作出理论分析。

## 6　结束语

博客和 BBS 可随意发表评论，由此产生许多不良信息。Web 社区的繁荣，也吸引了垃圾制造者的目光，将评论变成他们传播信息的平台，提高搜索引擎排名的工具。本文提出了适应评论和短文本的相关测度 *CorrPC* 和 *CorrPCVar*，明显

(下转第 96 页)