基于内容特征的垃圾博客过滤*

刘玮[1,2]，廖祥文[1,2]，许洪波[1]

1. 中国科学院 计算技术研究所 信息智能与信息安全研究中心，北京 100190

2. 中国科学院 研究生院，北京 100039

E-mail: liuwei@software.ict.ac.cn, liaoxiangwen@software.ict.ac.cn, hbxu@software.ict.ac.cn.

摘要：本文根据垃圾博客和正常博客在内容特征上的差异，对多种针对博客分类有效的统计特征进行了分析，提出基于博客内容统计特征的过滤方法。在 Blog06 数据集上的实验表明，该方法的过滤准确性达到 97%，比基于词频特征的过滤方法提高了约 7%，在不同规模训练集上的准确性保持在 95%左右，具有更好的泛化能力。

关键词：内容分析，垃圾博客过滤，统计特征，词频特征，泛化能力

Splog Filtering Based On Content Analysis

Wei Liu[1,2], Xiangwen Liao[1,2], Hongbo Xu[1]

1. Research Center of Information Intelligence and Information Security, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190

2. Graduate University of Chinese Academy of Science, Beijing 100039

E-mail: liuwei@software.ict.ac.cn, liaoxiangwen@software.ict.ac.cn, hbxu@software.ict.ac.cn.

Abstract: In this paper, we analyze many effective statistical features for splog filtering by investigating the differences between splogs and normal blogs. Then we present a splog filtering approach based on statistical characteristics of blog content. The experimental results on Blog06 data set show that, our approach can reach an accuracy of 97%, which improves the accuracy by 7% compared with term frequency based method. And with the test sample size increasing, its accuracy keeps around at 95%, indicating a better generalization ability.

Key words: content analyze, splog filtering, statistical features, term frequency features, generalization ability

1 引言

博客（web blog）通过为作者和读者提供交流平台而构建出交互式和动态更新的社会网络，已成为一种重要的信息传播媒介。基于博客的研究分析可以用于市场分析，公共关系，政治，心理学等各个研究领域，所以博客丰富的信息源和巨大的信息量具有重要价值。

博客世界也充斥着大量的垃圾博客，垃圾博客（spam blog or splog）是指内容由机器生成或者从其他网页复制而成的，以提高目标网站排名的博客[1]，垃圾博客的生成非常容易，传播十分迅速。根据垃圾博客过滤问题的特殊性，Finin 等人已于 2006 年将垃圾博客过滤作为 TREC 比赛新增的任务[2]。垃圾博客过滤是一项具有挑战性的任务，第一，检测技术必须是自动的快速的；第二，过滤不能带有主观性；第三，尽量降低误判；第四，检测应该越早越好。

本文正是根据这些问题出发，用基于博客内容的统计特征来作为垃圾博客的区分特征，这些特征来自于博客本身，各特征值的计算相互独立，不需要获取全局信息。正文内容中出现的词特征只是作为一个符号，不考虑其语义含义，所以不会导致过滤带有主观性，同时也避免了常规的

* 本文承国家 973 “大规模文本内容计算” 课题（2004CB318109）的资助。

基于词频特征的分类方法因训练集或关键词有限而带来的局限性问题。多种统计特征相互结合，比单一的词频特征具有更好的可靠性，减少了因偶尔含有垃圾词汇而导致误判的机会。部分统计特征可以在收录博客之前进行计算，从而在第一时间去除垃圾博客。

接下来的内容首先介绍该领域的研究现状和存在的问题，第三部分介绍了本文所用的数据集合，第四部分详细分析多种基于内容的统计特征的分类效果，第五部分实验证明了文中所采用的博客统计特征能有效的用于过滤垃圾博客，第六部分是文章的总结和可以深入研究的问题。

2 相关工作

Kolari 等人根据生成形式和表现形式将垃圾博客分为多类：关键字填充，内容拼接，内容剽窃，偷换链接，页面重定向等[3]，其中内容垃圾最为常见，本文采用基于博客内容的特征对内容垃圾博客进行过滤。

目前已经存在一些垃圾博客过滤方面的研究。Finin 等人[1]提出基于 SVMs 垃圾博客过滤方法，根据博客页面的词频特征进行分类。这种方法不仅需要大量人工标注的训练语料，而且由于训练集有限，分类能力会逐渐下降，泛化能力较差，难以应用于在线的垃圾博客过滤。Lin 等[4]在词频特征的基础上，还考虑了博客的链接，发布时间等自相似特征，但是这些特征需要计算一个博客所有博文的信息。Franco Salvetti[5]等人考虑到博客 url spam，利用 urls 的语言模型进行过滤垃圾博客，具体是将 urls 分割成词后用朴素贝叶斯分类模型计算博客 url 属于垃圾博客的概率。这种方法简单直观但是需要大量训练数据得到先验概率，而且仅考虑了 url 信息。

这些过滤方法大都是基于词频特征的分类，没有考虑博客内容的统计特征。Alexandros 等[6]的基于网页内容的多种过滤技术可以有效检测出垃圾网页，说明基于页面内容可以得到很好的统计特征。本文将针对博客这一类特殊的网页，充分挖掘和利用基于博客内容的统计特征。

3 数据集合

本文用到的数据来自于 Trec06 的 Blog 数据集[7]，该数据集的采集从 2005 年 11 月 6 日开始到 2006 年 2 月 21 日止历时 11 周，包含 100,649 个博客，3,215,171 个 permalinks。为了模拟真实的博客世界，加入了 17,969 个垃圾博客，占所有博客的 17.8%，509,137 个垃圾博客页面，占所有博客页面的 15.8%。其中约 87%的垃圾博客来自 Blogspot.com 域名。我们从 06 年博客检索任务的正确答案中抽取了 5000 个正常英文博客，人工检出 5000 个英文垃圾博客，组成样本数为 10,000 的数据集，本文的实验都在这个数据集中进行。

4 特征提取

博客对于普通网页的特有特征是解决垃圾博客过滤问题的关键。本文的内容特征提取体现出了博客内容所具有的自然语言性质，主要分为三个部分：基于网页标签的特征，基于正文结构的特征和基于正文内容的特征，我们将详细分析这些特征的区分能力。

为了说明某一特征对于区分垃圾博客和正常博客的有效性，接下来的图表由一个柱状图和一个曲线图组成，其中柱状图表示该特征的分布，曲线图表示垃圾博客的概率。横坐标表示该特

征的取值，左边的纵坐标表示所占博客总数的百分比，右边的纵坐标表示垃圾博客所占的比例，可以看到垃圾博客比例随某一统计特征变化而呈现出的变化趋势。

4.1 基于页面 url 和标签内容的特征

4.1.1 用户名长度，用户名非字母符号比例

博客的 url 是一个博文的永久链接地址（permalink），它通常包含博客发布系统的域名和博主（blogger）自己注册的用户名，正常博客的用户名是一个名字或者由有意义的单词组成，不会太长，而垃圾博客为了匹配到更多的检索关键字，使得长度增加。自动生成的用户名还可能包含过多数字符号等非字母字符，用户名中非英文字母符号的比例，也是区分垃圾博客的特征。如图 1 所示，随着 userid 长度的增加，垃圾博客的比例也增大。图 2 所示 userid 中非字母符号比例越大，垃圾博客的比例也越大。

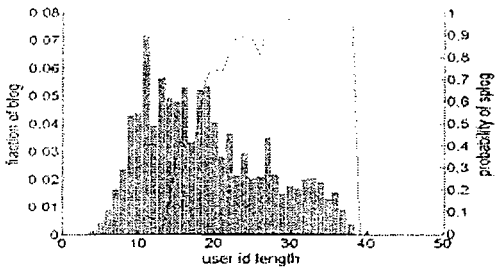


图 1 垃圾博客概率对应 userid 长度的分布

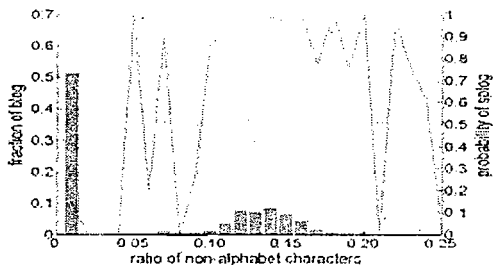


图 2 垃圾博客对应 userid 非字母符号比例的分布

4.1.2 title 标签内容长度

Title 域中的词可以描述博客内容，某些搜索引擎会赋予更高的权值，所以垃圾博客的一种方式是在 title 域中填充大量检索关键字，而使得 title 域变长。从图 3 可以看出，正常博客的 title 长度集中在 5 到 10 之间，过小或者过大的区间里的垃圾博客比例都增大了。

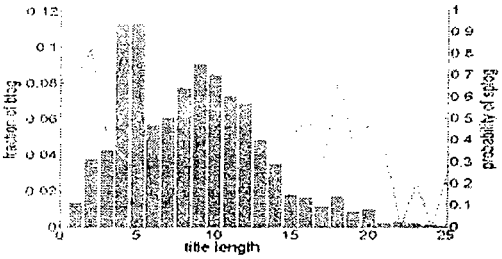


图 3 垃圾博客概率对应 title 长度的分布

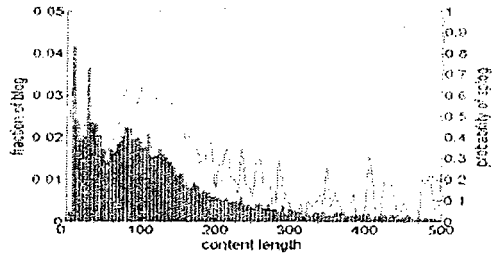


图 4 垃圾博客概率对应正文长度的分布

4.2 基于正文结构的特征

我们从博客中抽取出正文模板，然后将经过人工校对的模板用来抽取博客正文。正文结构包括正文的句长，词长，链接比例等自然语言属性。

4.2.1 正文长度

基于关键字填充的垃圾博客目的是尽可能多的匹配上查询，提高自身被检索到的机会，所以这种页面里会堆砌大量检索关键字，文章长度比较大。通过对图 4 的分析，可以看出博客正文的长度大部分在 100 到 300 个单词之间，而很多垃圾博客长度也集中在正常长度范围内。在长度

约为 50 个单词的地方红色曲线很高，表明还有很多垃圾博客的正文很短，这些博客为了保证频繁更新而采用了仅包含关键字的短文本内容。

4.2.2 平均词长

垃圾博客为了不被关键字过滤技术过滤掉，通常采用的方法就是用单词合成技术，将多个关键字粘连在一起，从而导致平均词长变长，例如“freecreditcard”，我们统计了除去页面结构单词的正文单词的平均长度。如图 5 所示，可以看出平均词长越长属于垃圾博客的概率越大，博客内容是人类自然语言，其中多是常用词，长度比较短，所以正常博客的平均值较小，如图所示平均值在长度为 5 的地方。

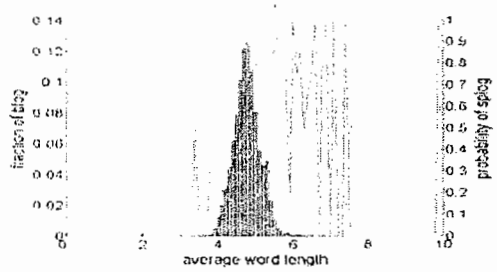


图 5 垃圾博客概率对应正文平均词长的分布

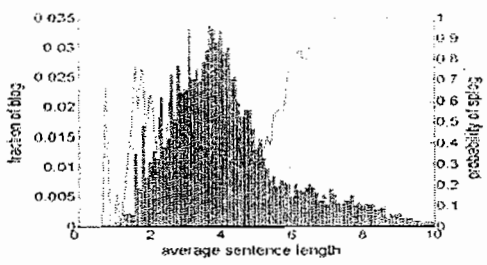


图 6 垃圾博客概率对应平均句长的分布

4.2.3 平均句长

正常博客都是记录生活或者抒发情感的日记形式文章，正文句子的长度不会太长，垃圾博客是复制而来或者机器生成的内容，不是为了供人阅读，而是为了匹配到更多的关键字，从而产生很多长句。如图 6 所示正常博客的句子平均长度集中在 5 到 10 个单词，随着平均句长的增大，垃圾博客的比例也增大。

4.2.4 锚文本比例

链接所带有的简单描述性文字，即为锚文本，搜索引擎会通过锚文本信息来确定所指向网站的内容，基于锚文本的垃圾博客将搜索引擎导向目标网站。图 7 所示的是垃圾博客对应锚文本比例的分布图，可以看出正文中出现大量锚文本是垃圾博客的一个特征。

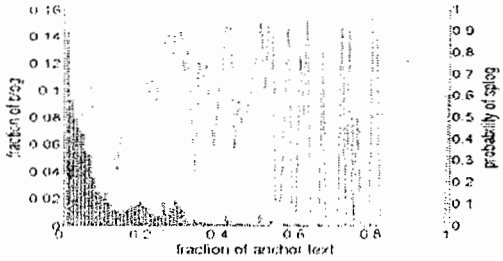


图 7 垃圾博客概率对应锚文本比例的分布

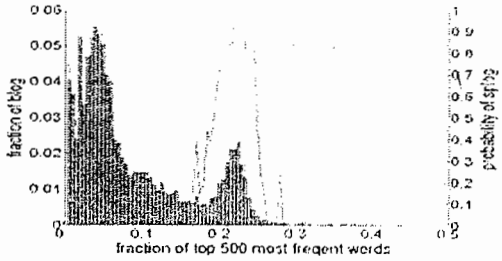


图 8 垃圾博客概率对应英文常用词比例的分布

4.3 基于正文内容的特征

本文考察正文内容是否是人类语言，考察的正文内容特征包括英文常用词所占的比例，重复串的长度和内容重复性。

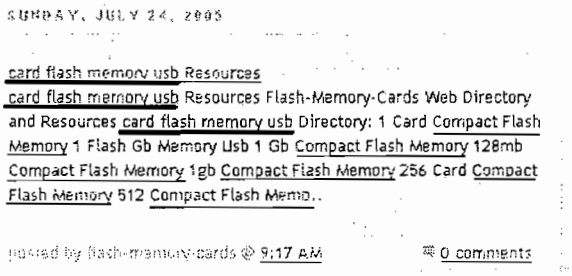
4.3.1 英文常用词比例

因为博客记录的是人类自然语言，所以经常在检索或者传统网页分析中过滤掉的所谓停用词

（比如 a，the，but）在垃圾博客过滤中是一类有用的特征，机器生成的内容通常会忽略人类语言中这种常用词的使用。取 N=500, 那么每包含一个常用词就加 1/500，这样包含停用词种类越多则分值越大，按种类计算而不是按总数计算是为了避免有些垃圾博客采用自动填充特定停用词的方法增加其常用词含有比例。如图 8 所示，正常文章中含有的常用词种类有一定的范围，大多集中在 0.3 以下，停用词比例过少和过多都会使得属于垃圾博客的概率增加。

4.3.2 重复串长度

很多搜索引擎的排名算法都基于关键字的 TFIDF 值，为了提高排名，垃圾博客的正文中会出现大量重复的关键字。于是内容的自我重复特性是垃圾博客的一个显著特征，重复的文本内容越长越不符合人类语言，本文定义最长重复串为博客正文中重复串的最大值，例如图 9 中，如果取出现次数大于 2 的字符串为重复串，重复串“card flash memory usb”的长度是 21，重复串“compact flash memory”的长度是 20，这篇文章的最大重复串长度是 21。从图 10 可以看出，最长重复串的长度越大，是垃圾博客的概率也越大。


SUNDAY, JULY 24, 2005

card flash memory usb Resources

card flash memory usb Resources Flash-Memory-Cards Web Directory and Resources card flash memory usb Directory: 1 Card Compact Flash Memory 1 Flash Gb Memory Usb 1 Gb Compact Flash Memory 128mb Compact Flash Memory 1gb Compact Flash Memory 256 Card Compact Flash Memory 512 Compact Flash Memo..

posted by flash-memory-cards @ 9:17 AM 0 comments

图 9 通过重复关键字提升被检索到的机会

4.3.3 内容重复率

最长重复串是考察文本单个重复串的长度，而内容重复率，本文将考虑整个文本的自我重复性，要计算全部文本的重复性，需要用自然语言处理技术来计算其句法和语义相似性，这样需要耗费大量时间，太过浩大的计算量也不符合快速过滤的原则，所以本文采取一种简单但是有效的计算方法，计算独立 n-gram 概率分布。

$$IndependentLH = -\frac{1}{k}\sum_{i=0}^{k-1}\log P\left(word_{i+1},...word_{i+n}\right) \tag{1}$$

n=1，2，3 时分别表示 unigram，bigram，trigram 的独立元似然概率。k 表示 ngram 的种类个数。P 表示该 ngram 在文章所有短语中出现的频率。如图 11 所示，独立 ngram 概率偏大或偏小都使得成为垃圾博客的可能性增加。

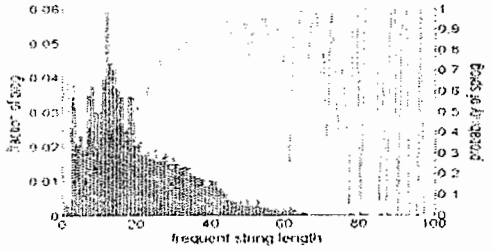


图 10 垃圾博客对应最大重复串长度的分布

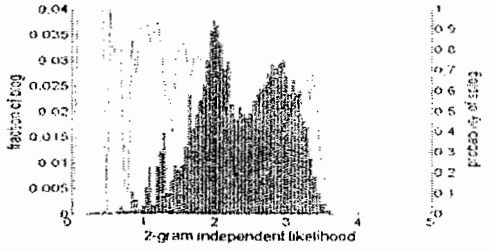


图 11 垃圾博客概率对应 ngram 概率的分布

5 实验结果与分析

实验分为两部分，第一部分是各内容统计特征的分类效果，第二部分采用 SVM 分类器考察博客内容统计特征和词频特征的分类效果。采用的评价指标是正确率（Precision），召回率（Recall）和 F-值，以下所指的准确率都是综合考虑了正确性和召回率的 F 值。

5.1 内容统计特征的分类效果

在第三部分的分析图中可以看出各个统计特征的分类能力有差异，接下来本文分析这些特征是否能有效区分垃圾博客，并找出最好的特征组合进行分类。本文采用决策树分类器 C4.5 作为本试验的分类器，在整个数据集上进行试验，采用 5 折交叉验证的方法计算平均准确率。图 12 是各统计特征的分类准确性，其中用户名长度和其中非英语字母字符比例是很好的区分特征，平均句长，自我重复率等特征也具有较好的分类能力。

由于特征之间可能存在冗余和相互依赖性，所以本文考察了部分特征组合的分类效果，其中常用词比例，平均句长，自我重复率和正文长度组合起来分类准确率达到 96.6%，用户名长度和其中的非英文字符比例组合起来准确率达到 95.1%。减少的特征只损失了少量的准确率，所以在实际应用中，为了进一步减少计算量，可以仅用这些特征组合进行过滤。

5.2 内容统计特征与词频特征的对比试验

SVM 分类器已经用于垃圾博客过滤任务中[1]，本文将这个方法作为对比试验。本文将数据集划分为四个不相交的子数据集 Data-1000， Data-2000，Data-3000 和 Data-4000，其中样本数依次为 1000，2000，3000 和 4000，垃圾博客和正常博客的比例都是 1：1。

首先在 Data-1000 子数据集上采取 5 折交叉验证的方法检验统计特征的分类效果。将 10 个内容统计特征输入 SVM，得到 97.4%的准确率。作为对比，本文取信息增益最大的 500 个正文词作为词频特征输入 SVM，得到 90.5%的准确率，如表 1 所示，用内容统计特征进行分类，大幅减少了特征数目的同时将准确率提高了近 7%。

特征类型（特征个数）	召回率	准确率	F-值
统计特征（10）	0.973	0.973	0.974
词频特征（500）	0.896	0.915	0.905

表 1 SVM 分别采用统计特征和词频特征的分类效果

为了证明由内容统计特征训练得到的分类模型具有更好的泛化能力，本文在 Data-1000 子数据集上训练 SVM 分类模型，然后分别在 Data-2000，Data-3000 和 Data-4000 上进行测试。实验结果如图 13 所示，随着测试集的扩大，基于词频的 SVM 分类器的分类准确性逐渐下降，当数据集扩大到 4000 时，基于词频分类准确性下降到 73.5%，而基于内容统计特征的分类器不仅准确性高，且一直保持在 95%左右。这与预测相符，因为由多种统计特征得到的分类模型具有更普遍的适用性，真正刻画了垃圾博客与正常博客的区别，所以分类效果具有更好的稳定性。

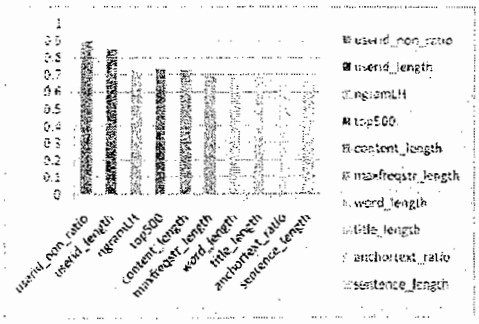


图 12 各种统计特征的分类准确性

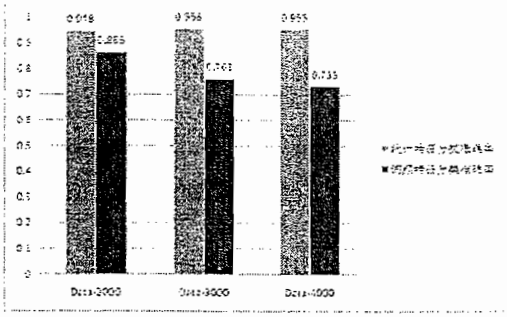


图 13 统计特征分类与词频特征的泛化能力对比

6 结论

本文分析了基于博客内容的统计特征，比较它们的分类能力和用分类器融合后的分类效果。结果表明，用内容统计特征进行垃圾博客过滤，不仅减少了特征数目也提高了过滤准确性，且具有更好的泛化能力。这是因为垃圾博客无论采用什么关键词，要达到被搜索引擎检索的目的就必然会在其内容统计特征上显现出异常，所以内容统计特征是很好的判别特征。

各种特征可以单独使用也可以与其他过滤方法结合使用，在实际应用中，可以按照需要选择部分统计特征，实现快速的垃圾博客过滤。文中对博客内容统计特征的分析研究是挖掘和利用博客信息的第一步，其在垃圾博客过滤任务上的应用是有效的。目前的试验都只是在英文博客中进行，下一步，我们将考察这些内容统计特征的推广能力，即这些特征在中文垃圾博客里是否具有同样的区分能力，其变化规律是否和英文垃圾博客类似。

参考文献

[1] Kolari P. , and Finin T. , Joshi A. . SVMs for the blogosphere: Blog identification and splog detection. In: Proc. of the AAAI Spring Symp. on Computational Approaches to Analyzing Weblogs. California: AAAI Press, 2006. 92 - 99.

[2] Kolari P. , Java A. , Finin T. , Mayfield J. , Joshi A. , Martineau J. . Blog Track Open Task: Spam Blog Classification. TREC 2006 Blog Track Notebook.

[3] Kolari P. , Java A. , Finin T. . Characterizing the splogosphere. In: Proc. of the World Wide Web 2006 Workshop on the Webloggging Ecosystem: Aggregation, Analysis and Dynamics. Edinburgh, 2006.

[4] Yu-Ru Lin, Hari Sundaram, Yun Chi, Junichi Tatemura, Belle L. Tseng. Splog Detection using self-similarity analysis on blog temporal dynamics. In: Proc. of the ACM Workshop on Adversarial information retrieval on the web. 2007, 1 - 8.

[5]Salvetti F. , Nicolov N. . Weblog Classification for Fast Splog Filtering: A URL Language Model Segmentation Approach. In: Proc. of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers, 137 - 140.

[6]Ntoulas A. , Najork M. , Manasse M. , Fetterly D. : Detecting spam web pages through content analysis. In: Proc of the 15th international conference on World Wide Web, Edinburgh, Scotland, 2006, 83 - 92.

[7]Macdonald C. , Ounis I. . The TREC Blog06 Collection: Creating and Analysing a Blog Test Collection. DCS Technical Report TR-2006-224. Department of Computing Science, University of Glasgow. 2006.