基于组合特征的动态垃圾博客过滤算法

任永功 尹明飞 杨荣杰

(辽宁师范大学计算机与信息技术学院 大连 116029)

摘 要 近几年,垃圾博客过滤成为国际上新的热点研究领域。现有的过滤算法大多基于词频特征分类,特征冗余并缺乏关联性。为了解决此问题,提出一种基于组合特征的动态垃圾博客过滤算法(CFDSD),该算法采用作者属性和自相似特征来解决特征冗余和关联性低的问题,并应用贝叶斯分类算法优化词频特征分类。实验表明,该算法能适应博客随时间变化而动态更新的特点,同时提高了过滤效率。

关键词 垃圾博客过滤,词频特征,自相似特征,组合特征,贝叶斯分类

中图法分类号 TP391 **文献标识码** A

Dynamic Splog Filtering Algorithm Based on Combined Features

REN Yong-gong YIN Ming-fei YANG Rong-jie

(School of Computer and Information Technology, Liaoning Normal University, Dalian 116029, China)

Abstract Splog filtering has become a new hot area in the international in recent years. Most of the traditional filtering algorithms are based on word frequency feature classification, which is quite redundancy and lack of relevance. According to this problem, a dynamic filtering algorithm based on the combination of features for splog(CFDSD) was proposed to solve the problem of low relevance and redundancy. The CFDSD algorithm uses self-similarity feathers and the attributes of author, at the same time adopts the Bayesian classification algorithm to optimize word frequency feature classification. Experiments show that the algorithm is adaptable to dynamical updated features of the blog with time changes, and improves filtering efficiency, while reducing the time to filter splog.

Keywords Splog filtering, Term frequency features, Self-similarity features, Combined features, Bayesian classification

1 引言

博客是一种以网络作为载体,由个人管理、不定期张贴新文章的网站。博客研究对市场调查、教育、社会福利、科学研究等具有重大价值。随着博客世界的发展,出现了大量的垃圾博客(spam blog or splog),它产生的方式主要是窃取别人的内容或机器生成,以提高目标网站在搜索引擎中的排名或链接盈利广告。垃圾博客造成的问题包括:1)严重降低信息的检索质量;2)明显浪费网络和存储资源。因此研究垃圾博客的过滤具有重大意义。传统方法主要研究垃圾博客文章常用词比例、重复串长度以及平均词长等内容特征,而忽略了绝大部分的垃圾博客是出自少数专业作者手中这一重要的属性特征[2]。本文以博客文章内容特征作为垃圾博客分类的基础特征,应用垃圾博客动态自相似特征结合作者属性来提高分类算法的准确率和召回率,弥补了传统算法特征冗余且不具有动态性的缺陷。

2 相关定义

设 D 为博客文章,$(W_1,W_2,\cdots,W_i,\cdots)$ 为 D 中 i 个关键词,$P(W_i)$ 指关键词 W_i 在 D 中出现的频率,$S(D)$ 表示 D 属于 splog,同理 $B(D)$ 属于 blog,$V(D)$ 指博客更新速度。

定义 1 基本特征

①博客文章的长度 用 $L(D)$ 表示,N 为关键词总数:

$$L(D)=\sum_{1}^{n}N(w_i) \tag{1}$$

$L(D)\in[100,500]$ 为约束条件,$L(D)$ 满足约束表示 D 属正常博客,$L(D)$ 不满足约束表示 D 属非正常博客。

②内容重复率 用独立 n-gram 概率分布来计算:

$$IndependentLH=-\frac{1}{k}\sum_{i=0}^{k-1}\log P(W_i,\cdots,W_{i+n}) \tag{2}$$

式中,$n=1,2,3$ 时分别表示 unigram, bigram, tri-gram 的独立元似然概率。k 表示 n-gram 的种类个数。P 表示该 n-gram 在文章所有短语中出现的频率。独立 n-gram 概率偏大或偏小都使垃圾博客的概率增加。

定义 2 独有特征

①D 具有的动态性。即博客更新速度 $V(D)\neq 0$。一个垃圾博客站点需要不断更新内容来吸引读者。

②博客中的链接是非认同链接。在 Web 中网站的链接代表了对另一个网站的认同,而在垃圾博客中不是认同而是

到稿日期:2011-06-21 返修日期:2011-08-22 本文受国家自然科学基金项目(60603047),教育部留学回国人员科研启动基金资助项目,辽宁省科技计划项目(2008216014),辽宁省教育厅高等学校科研基金(L2010229),大连市优秀青年科技人才基金(2008J23JH026)资助。

任永功(1972-),男,博士,教授,主要研究方向为数据挖掘、图像处理技术等,E-mail:renyg@dl.cn;尹明飞(1987-),女,硕士生,主要研究方向为数据挖掘;杨荣杰(1983-),女,硕士生,主要研究方向为数据挖掘。

直接的商业利益的广告链接。

定义 3（作者属性） 博主(Blogger)即博客作者，特指注册了博客空间的人；垃圾博主(Spammer)即垃圾博客的作者。统计实验证明[2]，Blogger 发表的 D 是 blog 的可能很大；Spammer 发表的 D 是 splog 的可能性很大。

定义 4（自相似特征）[5] 即 D 的发表时间特征

$$T=\mathrm{mod}(P_i,\delta_{day}) \tag{3}$$

$$V=|P_i-P_j| \tag{4}$$

$$\Delta T=V/\delta_{day} \tag{5}$$

式中，T 表示博客的横向时间，V 表示纵向时间，P_i 指博客 i 的发表时间，P_j 指博客 j 的发表时间，δ_{day} 等于 24h，即 86400s；ΔT 是博客发表时间的时间间隔率，用来衡量博客发表频度的均衡性。

3 组合特征垃圾博客过滤算法(CFDSD)

3.1 WFD 算法

WFD 算法(Word Feather Detect)即传统的内容特征算法，通过提取大量博客文章的词频，统计出每个词频特征的垃圾博客率，再用这些特征计算待测博文的垃圾博客率。内容特征是一个重要的判别特征，同时算法的可移植性强且减少了主观因素的影响。

在提出对冗余的词频特征进行剪枝时，因仅考虑词频，忽略了特征之间的关联性，造成了剪枝产生较大误差，降低了过滤的准确率。词频数据规模较大，使算法浪费大量存储空间，同时增加了运行时间。

3.2 CFDSD 算法

3.2.1 作者属性

鉴于 Yuuki Sato，Takehito Utsuro 对垃圾博客的统计规律以及对垃圾博客作者的分析[2]，本文采用 UMBC 的 Splog-Blog 数据集。在 Splog-Blog 的训练集上随机摘取一段连续 7 天的博客进行特征分析，本部分博客包括 CEO 博客、公共社区博客以及个人博客等 9 类博客。将其中的垃圾博客统计成一个垃圾博客作者属性表，见表 1。

表 1 垃圾博客作者属性表

作者	垃圾博客数	占垃圾博客率(%)	专业垃圾博客作者数	业余垃圾博客作者数
General	167	35	1	0
Libitina	123	25.9	1	0
caltechgirl	92	19.3	1	0
Andy	26	5.5	0	1
sam	17	3.6	0	1
Derdman	12	2.5	0	1
kcbme	9	1.9	0	1
其他	30	6.3	0	12
总共	476	100	3	16

从表中可看出，476 篇垃圾博客中 382 篇垃圾博客是由专业垃圾博客作者发表，仅有 94 篇是 16 名业余作者创建，由此可知约 80%的垃圾博客作者是出自专业垃圾博客作者之手。垃圾博客本身没有阅读价值就决定了垃圾博客作者必须通过大量地创建垃圾博客来增加博客的点击量。博客的作者属性具有易提取、方便检测的特点，因此可以在过滤垃圾博客时优先检测作者属性来提高过滤效率。

3.2.2 自相似特征的阈值

博客的横向时间指标 T 能体现出垃圾博客自身占据时间段的特点，博客的纵向时间 V 能测量出博客之间的时间关联性。相关概念见定义 4。根据式(3)，计算出的垃圾博客的横向时间自相似不同于正常博客，正常博客的发表时间在早 8:00到晚 22:00 之间，而垃圾博客的发表时间是机器生成的，会出现在午夜 0:00 到早上 6:00 之间。在纵向时间上，正常博客作者发表博文的时间是不均衡的，而垃圾博客的发表时间十分有规律。这是因为垃圾博客的产生方式之一是机器生成，发布时间不受人的情感因素所影响，发表时间具有均衡性；而正常博客的发表时间会根据作息时间的规律，午夜发表博客的数量明显少于正常活动时间，导致午夜时段的垃圾博客率高达 86%。

纵向时间的测量标准是博客发表的时间间隔率 ΔT。随机选取博客的发表时间不具有研究价值，时间间隔率采取时间间隔最近的两篇文章的发表时间之差与一天时间的百分比，其能够客观地挖掘博客作者发表博客时间的整体规律。表 2 是训练集中垃圾博客与正常博客发表博客的时间间隔率。

表 2 发表博客的时间间隔率

时间间隔率(ΔT)	垃圾博客(%)	正常博客(%)
0～1	47.5	12.5
1～5	35	25
5～10	15	45.8
10 以上	2.5	16.7

由表 2 可知，垃圾博客的时间间隔率在 0～1 之间的情况约占 50%，而正常博客的时间间隔率集中于 5～10 之间。这表明垃圾博客的时间间隔率是很小的，而正常博客的时间间隔率相对较大。这由垃圾博客的自身特点所决定，垃圾博客为了提高网页的点击量进而提高网站在 ALEXA 中的排名，须在短时间内发表大量的博文，同时机器生成垃圾博客文章的速度非常快。另一原因是博客世界具有实时性，文章的更新是高速的，故垃圾博客的作者需要缩小博客发表的时间间隔率。正常博客的作者会因工作繁忙或个人情感的起伏导致发表博客的时间间隔在某一段时间内偏大，致使时间间隔率在 10 以上还占有较大比例。

根据博客的作者属性和自相似特征易提取、方便过滤且特征高效的优点，在过滤垃圾博客时优先过滤该两项关联特征，能够大大提高过滤效率。垃圾列表 tableS 与非垃圾列表 tableB 的创建如算法 1 所描述。

算法 1 创建 tableS、tableB

输入：训练博客文章

输出：关联特征的博客表

Begin

for every splog、blog under train

1. collect(D(author)，D(time))；
   /* 提取训练集垃圾博客文章的作者及自相似属性 */
2. establish(tableS，table B)；
   /* 建立两个空表 tableS，tableB，分别存储垃圾博客和非垃圾博客的作者及发表时间 */
3. compute(T，ΔT)；
   /* 计算博客横向发表时间以及时间间隔率 */
4. while(D(T)$\in$(0:00～6:00) $\|$ (22:00～24:00)&& D(ΔT)$\in$(0～5)) insert(D(author)->tableS)；
   /* 将博客的作者插入垃圾博客列表 */
5. else
   insert(D(author)->tableB))；
6. goto step 3；

```
    until n=0;
  /*n为训练集中所有的文档数*/
End;
```

算法1根据博客的自身特点主动选择高效的关联特征，自适应地过滤掉大量的冗余特征来减轻博客世界的样本量大、样本特征不均衡所带来的影响。将选取作者属性特征以及自相似特征作为CFDSD算法的数据预处理，能准确过滤约80%的博文。剩余的博文没有时间、没有注明作者或列表中没有相对应的信息，需要采用词频分类方法。

3.2.3 贝叶斯词频分类

数据预处理无法检测的博文，彼此之间的关联性很弱，此时博文自身的内容就显得尤为重要。博客的内容特点主要体现在词频上。本文采用贝叶斯分类算法优化了词频特征的提取并应用REP方法[9]对大量的冗余特征进行了有效的剪枝，提高了算法的运行效率。

算法2 贝叶斯优化词频特征算法

```
输入:垃圾博客集和非垃圾博客集
输出:特征词根的垃圾概率hash表
Begin
1. collect(D);
  /*分别采集700篇的垃圾博客和700篇非垃圾博客建立splog集
    和blog集*/
2.   for(i=0;i<700;i++) collect skey,num(skey);
  /*提取垃圾博客的所有特征词根并统计每一个词根的词频*/
3.   for(i=0;i<700;i++) collect bkey;
4. create TreeS(skey),TreeB(skey);
  /*创建splog关键词树及blog关键词树*/
5. pruning(treeS,treeB);
  /* 采用REP方法进行剪枝*/
6. generate(tableS,tableB);
7. compute(p(tableS),p(tableB));
  /*计算两个表中的所有关键词的概率*/
8. match Key(TableS,TableB);
9. compute(P(hash(key)));
  /*根据贝叶斯公式计算*/
10. establish(hashT(key));
End;
```

其中贝叶斯公式如下：

$$P(C_j|D)=\frac{(C_j)\prod_i p(w_i|c_j)}{p(D)} \tag{6}$$

为减小大量词频特征的规模，同时保证算法的召回率，可在博客特征词根的垃圾概率表中随机选取几个连续片段的词频特征。这种策略具有普遍性，同时优化了算法。REP是最简单的事后剪枝方法之一，在数据量较大的情况下经常被应用且计算复杂度仅为$O(n)$。

贝叶斯词频分类算法主要针对正文内容，通过分析各种关键词在正文中所占的频率，得到了关键词与垃圾博客之间的统计规律。通过增加关键词的频率特征，可以增加各种特征之间的互补信息，使得CFDSD分类算法能更好地挖掘各种特征之间的关联性，从而提高对垃圾博客分类的准确率。

3.2.4 组合特征的贝叶斯分类算法

本文设计的组合特征分类算法属于二分类算法，其可以检测一个样本是否属于垃圾博客。对待检测样本的检测过程如图1所示。如果一个博客样本被组合特征分类模型的一个节点判断为Y(即该样本的特征属于关联特征表)，则可以用算法1进行分类。若该节点被判断为N，则用算法2进行分类。算法3在算法2的基础上实现了对一篇待测博客文章的分类。

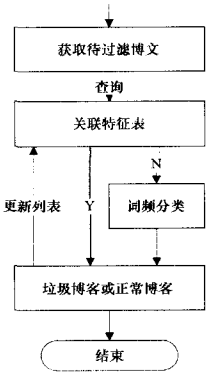

图1 组合算法的流程图

算法3 待测博客分类

```
输入:一篇文档
输出:文档的垃圾博客率以及是不是垃圾博客
Begin
1. Read(D);
2. Extract(key);
  /*提取独立的字符串(key),建立key*/
3. Pretreat(key);
  /*对key集进行预处理,及还原词根*/
4. if(match(key,hash(key))
      key1=key;
  /*根据输入的key和hash表中的key表进行匹配,查找满足输出
    key1*/
5. Check(p(key1));
  /*查找算法2 hash表中对应key1的垃圾概率*/
6. Acording(key1) Compute(P(D));
  /*用贝叶斯公式计算垃圾博客概率*/
7.   if(P(D)>0.5)
     printf("D is splog");
  /*阈值选为0.5*/
8. else
     printf("D is blog");
9. return 0;
End;
```

在算法3中，首先检测待测博文的关联特征，将关联特征与算法1中的表进行匹配。若满足条件，则直接进行分类；否则继续提取独立词频特征，删除不满足算法2中条件的词频特征，再用贝叶斯优化词频算法分类。算法3中的概率阈值由训练集中的实验结果选定。

组合特征分类算法在最坏的情况下是每个表都需要进行匹配，其时间复杂度为$O(2n^2)$。但是出现最坏情况的概率很小，因为组合特征分类算法选取了高效的关联特征。而基于词频特征分类的传统WFD算法的时间复杂度为$O(2n^2)$，因此组合特征分类算法的时间复杂度与传统算法相比并没有增加。同时，对大量冗余特征的剪枝并选取高效的关联特征使CFDSD算法为过滤垃圾博客节省了大量的存储空间。

4 实验结果分析

本文实验所采用的数据来自UMBC ebiquity的Splog-

(下转第212页)

2007,10(2):133-155

[8] Khare R, An Yuan, Song I-Y. Understanding Deep Web Search Interfaces: A Survey[J]. ACM SIGMOD Record, 2010, 39(1): 33-40

[9] Wu W, Doan A, Yu C, et al. Modeling and Extracting Deep-Web Query Interfaces[C]// Advances in Information and Intelligent Systems. Springer Berlin, Heidelberg, 2009: 65-90

[10] Khare R, An Y. An Empirical Study On Using Hidden Markov Model for Search Interface Segmentation[C]//Proc. of the 18th Int'l Conf. on Information and Knowledge Management. Hongkong, China, ACM Press, New York, NY, 2009: 17-26

[11] Dragut E C, Kabisch T, Yu C, et al. Hierarchical Approach to Model Web Query Interfaces for Web Source Integration[C]// Proc. of the 35th Int' Conf. on VLDB (Lyon, France). IEEE Computing Society, Washington DC, 2009: 325-335

[12] Raghavan S, Garcia-Molina H. Crawling the hidden Web[C]// Proc. of the 27th International Conference on very Large Data Bases. 2001: 129-138

[13] UIUC Web integration repository[EB/OL]. http://metaquerier.cs.uiuc.edu/repository/, 2010-10

(上接第 179 页)

Blog 数据集。该数据集采集于 2005 年 8 月 29 日，一共包含 3000 个垃圾博客网页，其中有 700 个 splog 网页、700 个 blog 网页，剩余 1600 个网页作为测试集。博客内容涵盖科学研究、个人情感、公共事业、新闻、图片、国际政治以及博客链接等，其完全可以模拟真实的博客世界。

SVM 分类器已经用于垃圾博客过滤任务中[3]，本文基于这个方法进行对比实验。所有的实验在相同环境下进行。实验环境为：Windows XP 系统，2.66GHz 奔腾四处理器，512MB 内存，VC 环境。采用的评价指标是准确率(Precision)、召回率(Recall)和 F 值，以下所指的准确率都综合考虑了准确率、召回率及 F 值。

所设计的实验方法如下：设定特征值数量从 100 到 400。分别比较组合特征算法和传统分类算法的准确率。实验结果如图 2 所示。

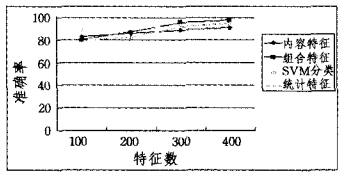

图 2 准确率结果对比

从图 2 来看，当特征数量比较少时，SVM 的准确率比较好；随着特征数量的增加，组合特征的准确率缓慢提高并逐渐显现出组合特征的优势。这是因为 SVM 需要大量人工标注的训练语料，由于训练集有限，而特征规模不断增大，使得 SVM 的分类面会出现较大的误差。特征数量增加，会完善 CFDSD 算法，使之具有高达 97.7%的准确率。

选定 400 个特征，比较分类算法的准确率(Precision)、召回率(Recall)和 F 值，如图 3 所示。

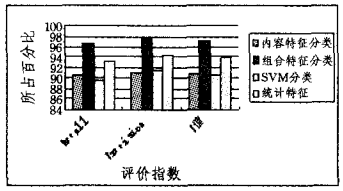

图 3 评价指数对比

当选定特征数量时，组合特征分类的准确率和召回率都高于传统特征分类、统计特征和 SVM。因组合特征是提取垃圾博客的关联特征，且特征之间存在互补性，所以其提高了算法对垃圾博客分类的能力；内容特征分类算法提取独立词频特征而忽略了特征的关联性；统计特征缺乏动态性而 SVM 不容易找到一个最佳的分类面。

结束语 本文分析了博客的关联特征和词频特征，比较了特征融合后分类算法与现有算法的分类效果。结果表明，基于组合特征的动态垃圾博客过滤算法，在同等时间复杂度下大大提高了过滤的准确率，具有更好的泛化能力，适应了博客的内容随着时间的不同而不断更新的特点。下一步工作是改进 CFDSD 算法阈值的提取，使阈值是由算法根据不同训练集而自适应地提取的。同时将 CFDSD 算法扩展到其他语言，如中文，以及将博客的 ping 时间系列特征加入到 CFDSD 算法中，以进一步提高 CFDSD 算法的泛化能力。

参 考 文 献

[1] Nanno T, Fujiki T, Suzuki Y. Automatically collecting, monitoring, and mining Japanese weblogs[C]//Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers & Posters. ACM Press(WWW Alt. '04), 2004: 320-321

[2] Sato Y, Utsuro T, Fukuhara T. Analysing features of Japanese splogs and characteristics of keywords[C]//Proc. 4th AIRWeb. 2008

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

[4] Melville P, Gryc W, Lawrence R D. Sentiment Analysis of Blog by Combining Lexical Knowledge with Text Classification[C]// Proc KDD 09. June 2009

[5] Ru Yu, Sundaram L H, Chi Yun. Splog Detection Using Self-similarity Analysis on Blog Temporal Dynamics[C]// Proc 5th AIRWeb Press. 2007

[6] Katayama T, Utsuro T, Sato Y. An Empirical Study on Selective Sampling in Active Learning for Splog Detection[C]// Proc 4th AIRWeb Press. 2009

[7] Kolari P, Finin T, Joshi A. Svms for the blogosphere: Blog identification and splog detection[C]// AAAI Spring Symposium on Computational Approaches to Analysing Weblogs. Baltimore County: Computer Science and Electrical Engineering. University of Maryland, March 2006

[8] Cormack G V, Smucker M D, Clarke C L A. Efficient and effective spam filtering and re-ranking for large Web datasets[J]. Computing Research Repository, 2010, 14(5): 441-465

[9] 魏红宁. 决策树剪枝方法的比较[J]. 西南交通大学学报, 2005, 40(1): 44-48

[10] 刘玮, 廖祥文, 许洪波. 基于统计特征的垃圾博客过滤[J]. 中文信息学报, 2008, 22(6): 86-91

[11] http://ebiquity.umbc.edu/resource/

[12] http://technorati.com