doi:10.3969/j.issn.1673-4785.2010.02.004

一种自反馈垃圾信息综合过滤方法

夏 虎,傅 彦,方育柯,周俊临
(电子科技大学 计算机科学与工程学院, 四川 成都 611731)

摘 要:提出了一种自反馈垃圾信息综合过滤方法.通过构建日志分析模块,在人为参与尽可能少的情况下,根据过滤到的垃圾信息通过自我分析、自我决策、自我优化来实现信息过滤规则的自反馈更新.试验证明该方法克服了传统海量信息过滤中人工参与度高、工作量大、效率和准确率与人的操作高度相关的缺点,大大提高了信息过滤速度和准确率,实现了信息过滤的自动化.

关键词:信息过滤;自反馈更新;日志分析;海量数据处理
中图分类号: TP181 **文献标识码**:A **文章编号**:1673-4785(2010)02-0117-05

A self-feedback synthesis method for spam filtering

XIA Hu, FU Yan, FANG Yu-ke, ZHOU Jun-lin
(School of Computer Science & Engineering, Chengdu 611731, China)

Abstract: A self-feedback based spam filtering method has been developed. In the construction of the log analysis module, the filtering system was implemented in a way that permited self-feedback when updating filtering rules. Self-analysis, self-decision, and self-optimization were all incorporated. In this way minimal human intervention was required. In traditional massive information filtering, human involvement was very high, leaving filtering accuracy and efficiency highly dependent on the skills of the human operator. Experiments proved that this method overcomes these shortcomings, greatly enhancing the speed and accuracy of information filtering and effectively automating information filtering.

Keywords: spam filtration; self-feedback updating; log analysis; massive data processing

垃圾信息的通常定义为:未经请求和允许而收到的、对接收者来说无用的信息.垃圾信息的内容主要包括广告信息、色情信息、假中奖信息、欺诈信息、恶作剧等.垃圾信息的泛滥不仅影响了人们的正常生活,也给正在蓬勃发展的网络和移动行业带来了很大的负面影响,甚至成了一大社会公害.因此,研制垃圾信息过滤系统具有重要的经济价值和社会效益.

在手机短信[1-3]、邮件[4-6]以及Web网页[7-8]中,垃圾信息过滤技术已经得到广泛的应用.目前的垃圾信息的实时过滤方案中主要采用静态方式,即系统运行过程中主要依赖手动设置的关键词或者黑名单等,在系统初始化时加载过滤器来进行信息过滤.这样的静态方式必须经过人工手动的调整参数和知识库来达到优化系统的目的,浪费了大量的人力及物力,同时静态方式也没有充分利用系统运行过程中得到的有价值信息,主要为过滤到的垃圾信息.另外,目前存在的方案中多采用单一方式进行信息过滤,在实时系统中,快速过滤方法如黑名单等可以满足速度要求,但是过滤效果一般不够高,而效率较好的内容过滤如分类等却有致命的速度弱点.

收稿日期:2009-12-04.
基金项目:国家自然科学基金资助项目(60903073);国家"863"计划资助项目(2007AA01Z440);四川省科技支撑计划资助项目(2008GZ0009).
通信作者:夏 虎. E-mail:xiahu@uestc.edu.cn.

1 相关工作

现有的信息过滤主要可以分为规则匹配算法和分类算法.

规则匹配算法即首先定义一系列的垃圾信息规则或正常信息规则,然后通过这些规则来匹配所需过滤的信息. 例如 Jangbok 等人[4]提出针对电子邮件的 url 规则,Deepak[2]和 Ramachandran 等人[9]提出的用户网络规则,Meizhen 等人[1]提出的行为规则,Peizhou 等人[3]提出的群发规则等. 规则匹配算法的速度较快,但是规则的制定需要通过长期的经验总结以及大量的人为参与和筛选,在规则制定不完善的情况下过滤效果不佳.

分类算法即首先根据人工标注的正常信息和垃圾信息进行训练,然后对所需过滤的信息进行分类,得到该信息是正常信息还是垃圾信息. 例如 Bin[6]、Jantima[7]、Deng、Li Qiang 等人提出的各个分类算法等. 分类算法的准确率较高,但是过滤速度较慢.

为了实现海量信息的快速、准确过滤,Pu 与 Calton1、Wu Ningning 和马亮等人提出了将上述方法进行整合,或某些步骤进行自动更新的方法.

针对以上不足,在此基础上提出了一种以信息过滤端、信息存储端、日志分析端三位一体的智能自反馈方法.

2 自反馈过滤模型

结合已有的信息过滤方法,将整个信息过滤分解为 3 个部分:信息过滤模块、信息存储模块和日志分析模块. 其中,信息过滤模块根据信息存储模块提供的过滤规则进行自动信息过滤,并将过滤结果存入信息存储模块的信息日志库中,日志分析模块则定时读取信息存储模块中的信息日志库,并进行自动分析,提取出有效过滤规则并更新信息存储模块中各个过滤规则,从而实现了信息过滤规则的自反馈更新. 自反馈过滤模型框架如图 1 所示.

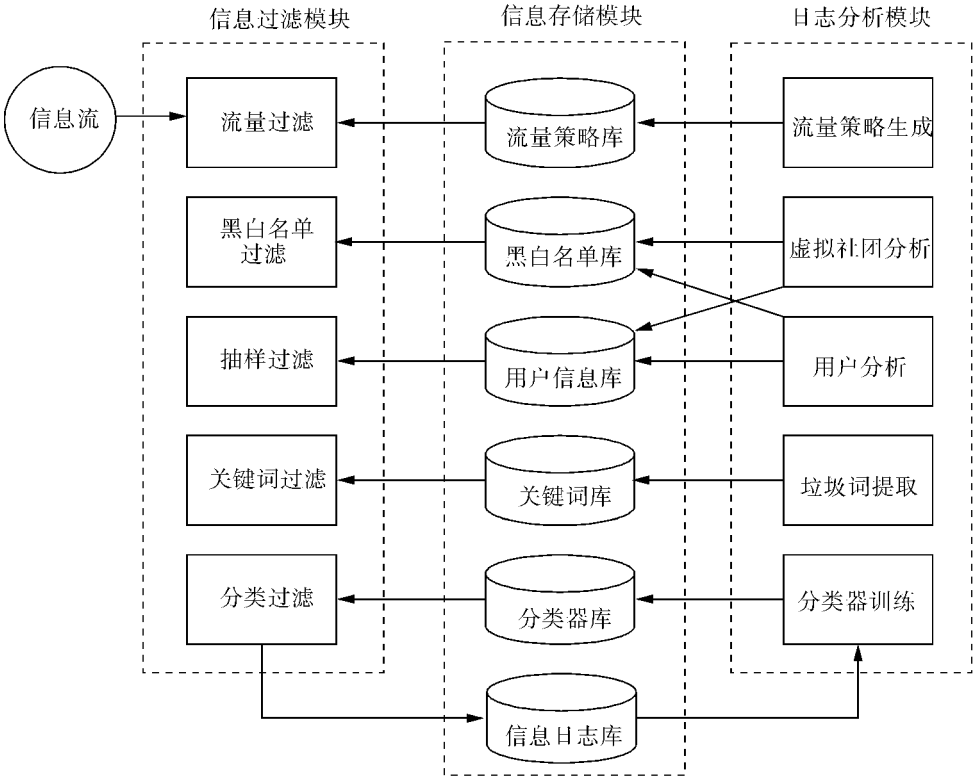


图 1 自反馈垃圾信息综合过滤模型框架

Fig. 1 A self-feedback spam filtering integrated framework

根据如图 1 的自反馈垃圾信息综合过滤模型框架,本方法的具体算法步骤如下:

1)建立信息过滤规则库,各种信息过滤规则存入其中;

2)信息过滤器从过滤规则库中得到过滤规则,对信息流进行过滤,并将过滤结果存入信息日志库中;

3)构建日志分析模块,从信息日志库中读取垃圾信息,分析后更新信息过滤规则库中的过滤规则.

以下分别介绍自反馈过滤算法中的自反馈过滤算法和自反馈更新算法. 其中自反馈过滤算法如图 2 所示.

```
1 Algorithm feedbackFilter(DB,cache,interval)
2 Begin
3       Initialize(DB, rules);
4       while(true)
5       do begin
6       d = cache(first);
7       if(d=null)
8             then begin
9             wait(interval);
10            continue;
11      end
12      else begin
13            cache.remove(first);
14            IntegratedFilter(d, rules);
15            if(d is spam) then
16                  Save(d,DB);
17      end
18    end
19 end
```

图 2　垃圾信息自反馈过滤算法

Fig. 2　Self-feedback algorithm for spam filtering

其中,第 1 行列出需要传入的参数,包括 DB 数据库连接,待过滤信息缓冲池 cache、interval 缓存空时等待时间间隔. 第 3 行初始化数据库并且读取过滤规则. 6 ~ 11 行读取 cache 中第一条待过滤信息 d, 若 cache 中无待过滤信息则等待时间间隔 interval 后继续. 否则行从 cache 中删除信息 d 后,使用规则 rules 对 d 进行过滤,若识别为垃圾信息 spam,则将其保存入数据库的日志库中 12 ~ 17,否则继续.

```
1 Algorithm feedbackRenew(DB, hourcount)
2 Begin
3       Initialize(DB, rules);
4       newrules= ϕ
5       crules= ϕ krules=  ϕurules=   ϕ
6       srules= ϕ; frules=  ϕ
7       while(hourcount >0)
8       begin
9             wait a hour;
10            hourcount = hourcount -1;
11      end;
12      Initialize(log);
13      crules =trainclassifier(log);
14      krules =extractkeyword(log);
15      urules =useranalyse(log);
16      srules =socialnetwok(log);
17      frules =flowanalyse(log);
18      newrules=curles∩krules∩urules∩srules∩frules;
19      newrules=mergerules(newrules, rules);
20      save(newrules,DB);
21 end
```

图 3　垃圾信息自反馈更新算法

Fig. 3　Self-feedback algorithm for rules update

自反馈更新算法如图 3 所示. 其中,第 1 行列出需要传入的参数,包括 DB 数据库连接,自反馈更新的时间间隔 hourcount(例如 24 h). 3 ~ 6 行初始化各参数,包括 DB 数据库连接、原始过滤规则集 rules、新分类规则 crules、新关键词规则 krules、新用户规则 urules、新用户网络规则 srules,新流量过滤规则 frules. 7 ~ 11 行实现定时机制,循环等待过滤开始时间. 第 12 行初始化待分析的日志文件. 13 ~ 18 行分别从日志中分析得到上述所有新规则并并入新规则集 newrules 中. 19 ~ 20 行将新规则集 newrules 与原始规则集 rules 进行整合并将新规则存入数据库中.

3　试验与分析

为了对比以往方法中垃圾信息过滤中的准确率和过滤速度不能兼顾的问题,分别从自反馈过滤的准确率和效率两方面来进行试验和分析.

3.1　自反馈准确率分析

本文的垃圾信息综合过滤方法目的是将信息分为垃圾信息和非垃圾信息,因此采用 2 种在分类算法中常用的准确率和召回率作为评价指标. 由于非垃圾信息识别为垃圾信息与垃圾信息识别为非垃圾信息相比会造成更大的损失,因此同时也采用误判率(FAR)来评价正常短信误判的情况.

在信息过滤中,垃圾信息识别的准确率的计算公式如下:

$$\text{precision} = \frac{N(\text{Correct})}{N(\text{Detected})} \times 100\%.$$

式中: N(Correct)表示正确识别出来的垃圾信息, N(Detected)表示识别出来的垃圾信息.

垃圾信息的召回率的计算公式如下:

$$\text{recall} = \frac{N(\text{Correct})}{N(\text{Correct}) + N(\text{Missed})} \times 100\%.$$

式中: N(Correct) 表示正确识别出来的垃圾信息, N(Missed) 表示被遗漏的垃圾信息.

正常信息的误判率(*false accept rate*)的计算公式如下:

$$\text{FAR} = \frac{N(\text{Normal}) \cap N(\text{Detected})}{N(\text{Normal})} \times 100\%.$$

式中: N(Normal)表示正常信息, N(Detected)表示识别出来的垃圾信息, N(Normal) ∩ N(Detected)即表示被识别为垃圾信息的正常信息.

实验中,通过对采集到的短信数据进行测试,其中,训练数据集包含4 867条正常短信和3 791条垃圾短信,测试数据集分别采用1 579条和40 000条的

大小不同的2个数据集.测试结果如表1所示.

表1 垃圾信息过滤结果

Table 1 Filtering results

试验1						试验2					
短信	总数	识别		准确率/%	召回率/%	误判率/%	短信	总数	识别	准确率/%	召回率/%	误判率/%
		正常	垃圾						正常 垃圾			
正常	1 170	1 169	1	99.7	82.1	<2.1	正常	28 000	27 729 271	97.2	81.0	<3.02
垃圾	409	73	336				垃圾	12 000	2 277 9 723			

由表1可以看出,提出的综合过滤方法针对小数据集和大数据集的准确率均值为98.15%,召回率均值为81.59%,误判率均值为2.56%.

3.2 自反馈效率分析

由于采用自反馈方法,具有高过滤速度的过滤方法(如黑白名单过滤、抽样过滤和关键词过滤等)在得到日志分析的反馈更新后,在过滤的比例上随系统运行而提高.在本实验中,系统运行的人工参与时间定为每天一次,对反馈更新的各新增信息进行人工筛选和确认.实验采用200万大数据量进行测试,得到过滤效率随时间的变化如下图4和图5所示.

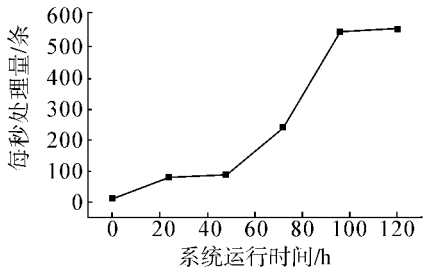


图4 信息过滤随系统运行每秒的处理量

Fig.4 System handling capacity per second

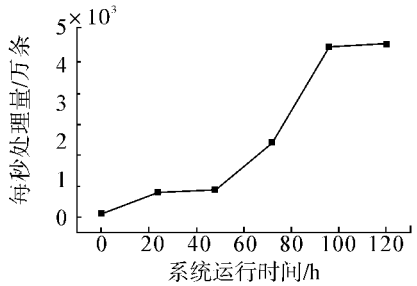


图5 信息过滤随系统运行每天的处理量

Fig.5 System handling capacity per day

由图5可以看出,随着系统的不断运行和过滤规则的自反馈更新,信息过滤速度有了显著的提高.

4 结 论

针对以往垃圾信息过滤方法中过滤速度和过滤准确率不能兼顾,以及过滤规则人工参与度高的缺点,提出了一种自反馈垃圾信息综合过滤算法.通过将快速过滤算法和高准确率算法进行整合,并使用过滤规则的自反馈更新,在系统运行前期使用高准确率算法(如分类)进行过滤,并定时通过日志分析功能提取新规则并更新规则库,使得快速过滤算法(如关键词)的规则库得到自反馈更新后,在后期系统运行中不仅可以占据更大的过滤比例从而使得过滤速度大为提高,还可以应用优化的规则库得到更高的准确率,在目前的海量数据应用领域具有较高的应用价值.

下一步针对海量数据过滤方法的研究,有以下几个重点:

1)针对信息过滤过程中用户信息的隐私保护研究;

2)针对垃圾短信、垃圾邮件等具有多变特性数据的不确定性研究;

3)针对垃圾信息传播的社会网络结构演化的分析研究;

4)针对海量数据的云计算研究.

参考文献:

[1]WANG Meizhen, LI Zhitang. Research on behavior statistic based spam filter[C]//Proceedings of the 1st International Workshop on Education Technology and Computer Science (ETCS 2009). Wuhan, China, 2009:687-691.

[2]DEEPAK P, JYOTHI J. A community based approach for spam filtering[C]//Proceedings 2004 International Conference on Information and Communication Technologies: From Theory to Applications (ICTTA 2004). Damascus, Syria,

2004:611-612.

[3] HE Peizhou,SUN Yong. Filtering short message spam of group sending using CAPTCHA[C]//Proceedings 1st International Workshop on Knowledge Discovery and Data Mining. Washington DC,USA:IEEE Computer Society,2008:558-561.

[4] KIM J,CHUNG K,CHOI K. Spam filtering with dynamically updated URL statistics[J]. IEEE Security and Privacy,2007,5(4):33-39.

[5] WU Yu,LI Zhijun,LUO Ping,WANG Guoyin. A new anti-spam filter based on data mining and analysis of email seuriy[C]//Proceedings of SPIE—The International Society for Optical Engineering. [S.1.],2003:147-154.

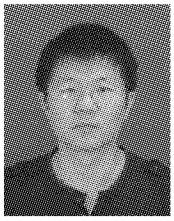
[6] CHEN Bin,DONG Shoubin,FANG Weidong. Email header feature study for improving Bayesian anti-spam filter[J]. Journal of Computational Information Systems,2008,4(3):1205-1212.

[7] JANTIMA P,ANIRUT C,CHUMSAK S,RAPEEPORN C,SOMNUK P. Content-based text classifiers for pornographic Web filtering[C]//Proceedings IEEE International Conference on Systems,Man,and Cybernetics. Taipei,China,2006:1481-1485.

[8] ZHOU Xujuan,LI Yuefeng,BRUZA P,WU Shengtang,XU Yue,RAYMOND Y K. Using information fitering in Web data mining process[C]//Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence. Washington DC,USA:IEEE Computer Society,2007:163-169.

[9] ANIRUDH R,NICK F. Understanding the network-level behavior of spammers [J]. Computer Communication Review,2006,36(4):291-302..

作者简介:



夏　虎,男,1981年生,博士研究生,主要研究方向为数据挖掘、异常检测、隐私保护.发表学术论文多篇.

傅　彦,女,1962年生,教授,博士生导师,电子科技大学计算机科学与工程学院副院长,四川省教学名师.主要从事模式识别、数据挖掘应用、信息安全等.主持和参与科研项目30余项,发表学术论文50余篇,被SCI、EI等检索10余篇.