

基于权重均值的不良网页过滤算法研究

唐坚刚, 魏然

(上海理工大学计算机工程学院, 上海 200093)

摘要: 传统的网页权重过滤算法中的权重大都根据词频统计方法来确定, 该方法不能很好地表达关键词对主题的表达程度, 且易被某些网站利用反关键字过滤策略逃避检测。在传统方法的基础上, 设置加权的关键词矩阵词典, 从关联规则出发, 应用汉语语料库里的同义词定义, 提出基于同类词权重均值的关联过滤算法。试验结果表明, 该算法过滤更为高效, 并且能够很好地应对色情网站的反关键字过滤策略, 尤其在色情与医学网页的分离上有明显的效果。

关键词: 网页过滤; 关键字; 矩阵词典; 关联规则; 权重均值

中图分类号: TP309 **文献标识码:** A **文章编号:** 1000-7024(2008)05-1088-02

Study and realization of method to webpage filtrating based on weight equal value

TANG Jian-gang, WEI Ran

(Institute of Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: The weights of traditional keywords webpage filtering are mainly determined by the frequency of statistical methods. This method can't expression degree of keywords characterization of the theme very well, and some websites are easy to use anti-keyword filtering strategy to evade detection. Based on the way of the traditional keywords webpage filtering, intercalate a keyword matrix dictionary with weight value, setting out from the connection rule, make use of the same kind word definition in the Chinese glossary database, creatively brought forward a connection rule filtering algorithm base on the weight equal value of the same kind word, which makes filtering more effective and cope with the strategy to the anti-keyword filtering of eroticism website. Especially in the separation of the eroticism webpage and the medical science webpage have the obvious effect.

Key words: webpage filtering; keywords; matrix dictionary; connection rule; weight equal value

0 引言

互联网技术飞速发展,信息的共享和发布跨越了时空的限制,网络成为人们获取信息的重要源泉,但网络上各种不良文本(色情、暴力、反动)也随之泛滥。如何控制这些信息的传播,进行有效信息过滤,已成为网络内容安全研究的主要内容之一。

网页是信息的载体,信息的过滤便是网页的过滤。当今关于网页过滤的方法有很多,而对于色情网页过滤方法研究上大致有这么一个现象:在研究的试验结果处往往是3个网页库的对比:一个库是纯色情网页,一个库是关于医学健康网页,还有一个库就是一般的新闻正规信息网页。大部分的模型与算法,可以在正确率上将前两者和最后一个拉开距离,比如第1个库的过滤误差率是13%,第2个库的过滤误差率为12%,第3个库的过滤误差率是6%。虽然在过滤一般网页上的确有了很显著的效果,但是在和医学健康方面网页没有拉开距离。所以,本文将在这里入手,从关键字的角度出发,在基于一个色情文本语料库的基础上,结合同类词的权重,进行

二次分离过滤研究,将色情网页从正规的医学性知识等健康网站里分离出来。

1 研究方法介绍

基于关键词的网页过滤系统中,其权重往往是通过词频等统计方法确定的,这不能很好的表达关键词对主题的表达程度,而且现在很多色情网站都有采取反关键字过滤方法,故基于这种词频取值的关键词过滤系统的过滤准确率到达一定值后其精度就很难提升。本文采用的方法不同于传统文本内容或者关键字过滤的方法,它并不基于某些特殊的关键字出现的次数和频率来划分。在汉语语料库里,每个词语都是有其意思相关或者意义相近的词汇。将这一点放入关键词的网页过滤系统中,我们人工设置好一个敏感关键词的语料库,如果在给这些关键词之间设上链接模型,使得他们两两互连,每个词语都给予一定的权重。接着基于这个库开始扫描一篇文章。按照词汇的权重均值进行判定,可得出是色情网页还是医学健康网页。

收稿日期: 2007-03-19 E-mail: william212@163.com

基金项目: 上海市高等学校青年科学基金项目 (03SQ05)。

作者简介: 唐坚刚 (1962—), 男, 上海人, 博士, 副教授, 研究方向为网络信息安全; 魏然 (1984—), 男, 江西南昌人, 硕士研究生, 研究方向为网络信息安全。

2 设置加以权重的关键词词典

首先用人工的方法收集整理常见的色情词汇,并做成一个独立的色情词典(小词典)。为了计算关键词权重,本文借鉴了 WordNet 中采用的词汇的矩阵模型^[1],提出了构造基于语义的矩阵词典,如表 1 所示,简述如下:

矩阵词典中的矩阵的行存放的是同一类型相互关联而不同词形的词,且该行的第一个单词为这些词中最通用的词,称为矩阵首列词。文档中其它地方出现的与该词汇同类的其它词汇均可与首列词相关,以使概念扩大化后的关键词可以增减。

矩阵词典中的矩阵的列存放的是分类后的不同词义的词汇,是在按人工分类排序存放的,可将该列看作概念缩小化后的传统意义上的关键词词典^[2]。

表 1 矩阵词典

首列词	同类关联词				
	W1	W2	...	Wj	...
M1(F1)	T11(F11)	T12(F12)	...	T1j(F1j)	...
M2(F2)	T21(F21)	T22(F22)	...	T2j(F2j)	...
...
Mi(Fi)	Ti1(Fi1)	Ti2(Fi2)	...	Tij(Fij)	...

注: W1, W2, ..., Wj 为同类关联词, M1, ..., Mi 为矩阵首列词, Fij 为每个词汇对应权重, 其中 $1 \leq i \leq m, 1 \leq j \leq n$, Tij 可为空值, Fij 不可为空值。

设置好矩阵词典后,给每个关键词加上一定的权重。关键词的权重并不以统计词频为基础,这也就是本方法与其它关键词过滤算法一个最大的不同。关键词将由人工赋值权重。对于一般生活用语将赋予比较低的权重,而色情文章里频频出现的词汇将会赋予相对较高的权重。在试验中,我们给词典里的关键字的权重范围 F 设置为 $F \in [1, 5]$ ^[3]。

3 基于词典的权重算法模型

得到网页后,对网页进行预处理,仅抽取其标题及正文内容文本^[4]。当扫描到关键词 M_i , 则开始词典 M_i 同类(也就是词典中的同列)关联词 T_{ij} , 并记录其对应权重 F_{ij} 。这时说项 T_{ij} 的权重是 $F_{ij}, 1 \leq i \leq m, 1 \leq j \leq n$ 。

统计权重均值: 取得同类关键词权重后, 获得第 i 个关键词 M_i 的权重均值 $\overline{F(T_i)}$, 即 $\overline{F(T_i)} = \frac{1}{n} \sum_{j=1}^n F_{ij}$; 重复扫描词典里的敏感关键词, 在去除已扫描关键词的基础上, 重复上述步骤。直到没有新得关键词出现为止。全部扫描结束后, 得到 K 个 $\overline{F(T_i)}$, 再取这 K 个 $\overline{F(T_i)}$ 的权重均值 $\overline{F(T_k)}$, 即 $\overline{F(T_k)} = \frac{1}{k} \sum_{i=1}^k \overline{F(T_i)}$; 根据关联规则设置得最小支持度 $MS(\text{minsup})$ ^[5]。将 $\overline{F(T_k)}$ 与这个最小阈值 MS 比较, 如果 $\overline{F(T_k)} < MS$ 则文本输出; 否则, 认为该网页为疑似网页, 需进一步处理。

4 实验与简化的网页过滤模型

本实验以色情网页为主要研究对象, 搜集了 260 篇色情网页和 275 篇医学类健康网页作为实验样本集。另外, 搜集 97 篇色情网页、103 篇性知识医学类网页和 75 篇新闻正规类网页作为测试样本集。实验中, 用人工的方法收集整理了常

见的色情词汇和常见的医学健康类及主要的性相关文学类词汇, 并按照本文提及的矩阵词典理论做成一个独立的小型矩阵词典, 选用了 Java、Microsoft SQL Server 2000 等作为开发工具。为了比较分析, 实验中采用两种过滤算法: 一是通常的词频统计算法^[6], 二是本文所提出的借助矩阵词典的关联词汇权重统计算法。模型如图 1 所示。

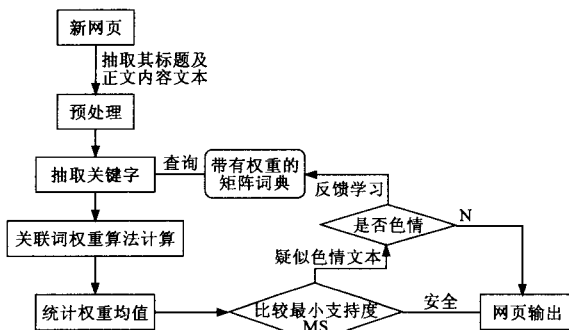


图 1 简化的过滤模型

5 结果与分析

准确率是评估不良网页过滤模型的指标之一^[7-8], 可以定义为: 所有过滤的文本中与人工分类结果吻合的文本所占的比率。其数学公式表示如下

$$\text{准确率} = \frac{\text{过滤的正确文本数}}{\text{实际分类的文本数}}$$

从表 2 的数据可以看出, 在色情与医学网页构成的实验样本集里, 新的算法比传统的过滤算法可以产生更准确的过滤效果。从而顺利实现医学与色情网页的分离。

表 2 算法测试结果

定义类	采用传统关键词过滤权重算法的准确率	采用基于词典关联词汇权重统计算法的准确率	差距
实验样本集	74.23%	85.78%	11.55%
测试样本集	67.67%	73.92%	6.25%
均值	70.79%	79.82%	8.87%

6 结束语

本文提出了带有权重的矩阵词典和关键词关联策略, 通过对关键词同类词汇的扫描得到网页的权重均值。实验证明该算法在医学与色情网页的过滤系统应用中更合理有效, 尤其是可以避免一些色情网站的反关键字过滤策略。但是由于本文矩阵词典中词汇是由人工收集并且其权重是由人工依靠经验所设置, 随意性较大且矩阵词典中收录的词汇量较少, 这都将影响过滤准确率。而在面对网页文本较大的时候关联算法的效率需要改进。

参考文献:

[1] Mohamed Hammami, Youssef Chahir. WebGuard: Web based adult content detection and filtering system[C]. Proceedings of the IEEE/WIC International Conference on Web Intelligence. Computer Society, 2003. (下转第 1107 页)

供商提供的服务无缝结合,不必改变 Intranet 应用。②用户边缘路由器只与服务提供商网络边缘路由器相连,而不与 VPN 中的其它节点直接相连,服务提供商边缘路由器只接收并保持与其直接相连路由器的有关 VPN 的路由信息,所以用户在管理自己的 VPN 时会发现使用 MPLS 模式时路由配置非常简单。他们可以把运营商的骨干网当作他们到所有地点的缺省路由来使用,而不需要与非常复杂的、包括了大量 2 层 PVC 或 3 层路由表的网络打交道。③由于在骨干网络采用 VPN-ID,可以保持全网的惟一性,因此 VPN 用户可以沿用原有的专用 IP 地址而无须 NAT(网络地址翻译)。

3.5 MPLS 技术能为 VPN 提供良好的安全性

MPLS 网络的 LSP 通道机制实现透明报文传输,具有与帧中继和 ATM VCC(virtual channel connection,虚通道连接)类似的高可靠安全性。体现在:

(1)MPLS VPN 实现了 VPN 之间的路由隔离。如前所述,每个 PE 路由器为每个所连接的 VPN 维护一个独立的 VRF,每个 VRF 处理来自同一 VPN 的路由(静态配置或在 PE 和 CE 之间运行路由协议)。因为每个 VPN 都产生一个独立的 VRF,因此不会受到该 PE 路由器上其它 VPN 的影响。在穿越 MPLS 核心到其它 PE 路由器时,这种隔离是通过为 MP-BGP 增加惟一的 VPN 标志符来实现的(这是在 BGP 方式下,虚拟路由的方式与此类似)。MP-BGP 穿越核心网专门交换 VPN 路由,只把路由信息重新分发给其它 PE 路由器,并保存在其它 PE 的特定 VPN 的 VRF 中,而不会把这些 BGP 信息重新分发给核心网络。因此穿越 MPLS 网络的每个 VPN 的路由是相互隔离的。MPLS 完全可以隔离无关用户的通信,使得无关用户的通信不会混杂,从而提高了安全性。

(2)隐藏了 MPLS 核心结构。出于安全考虑,运营商和终端用户通常并不希望把它们的网络拓扑暴露给外界,这可以使攻击变得更加困难。如果知道了 IP 地址,一个潜在的攻击者至少可以对该设备发起 DoS 攻击。但由于使用了“路由隔离”,MPLS 不会将不必要的信息泄露给外界。

(3)抗攻击性。因为进行了路由隔离,因此不可能从一个 VPN 攻击另外一个 VPN 或核心网络。

(4)标记欺骗。在 MPLS 网络中,包的转发不是基于 IP 目的地址,而是基于由 PE 路由器预先添加的标记。与 IP 欺骗攻击时攻击者替代包的 IP 源地址和目的地址相似,理论上有可能出现 MPLS 包的标记欺骗。任何 CE 路由器和它的对等 PE

路由器之间的接口主要是 IP 接口(也就是说没有标记)。CE 路由器不知道 MPLS 核心的存在,所有的“标记”工作都应该是由 PE 完成的。因此出于安全考虑,PE 路由器应该不接受来自 CE 路由器的任何标记包。当然,发送到 MPLS 网络中的包仍然存在 IP 地址欺骗的可能性,但这可以通过地址隔离来实现,使得属于某个 VPN 的用户只可能攻击他自己的网络,而无法攻击别人的网络。

4 结束语

MPLS 技术在兼容原来的第三层 IP 技术的同时,以更新换代的气魄改进其下层支撑网络的结构,既保留 IP 网络的灵活性优点,又能达到网络换代的自然过渡。种种迹象表明,MPLS 有望成为继 IP 技术后的新一代主流网络技术。MPLS VPN 能够充分利用 MPLS 技术各方面的优势,在对企业单位原有网络配置影响不大的基础上,极大地提高用户网络运营和管理的灵活性,同时能够满足用户对信息传输安全性、实时性、宽频带和方便性以及 QoS 等各方面的需要。因此,与现有传统 VPN 相比,MPLS VPN 将提供令用户更为满意的服务。

参考文献:

- [1] 石晶林,丁炜.MPLS 宽带网络互联技术[M].北京:人民邮电出版社,2001.
- [2] 华为-3com 公司技术研发小组. MPLS 技术白皮书 [Z]. 华为-3com 公司,1999.
- [3] Rekhter B D Y. 多协议标记交换技术与应用[M]. 罗志祥,朱志时,译.北京:机械工业出版社,2001.
- [4] Luca Martini, Steve Vogelsang, Daniel Tappan, et al, draft-martini-l2circuit-trans-mpls-0x.txt, Internet Draft[Z]. 2001.
- [5] Rosen E, Rekhter Y. RFC2547, BGP/MPLS VPNs[S]. Cisco Systems Inc,1999.
- [6] Gleeson B, Lin A, Heinanen J, et al. RFC2764, A framework for IP based virtual private networks[S].
- [7] Muthukrishnan K, Malis A. RFC2917, a core MPLS IP VPN architecture[S]. 2000.
- [8] 城域网 MPLS VPN 的几种实现方法[J].华为技术,2002,7(149):135-147.
- [9] 韩海雯,林生. 新一代网络技术—MPLS 工作机制剖析[J].微机发展,2005,15(12):129-134.

(上接第 1089 页)

- [2] 张文修,吴志伟,梁吉业,等. 粗糙集理论与方法[M].北京:科学出版社,2001.
- [3] Pawlak Z. Rough, theoretical aspects of reasoning about data[C]. Proceedings of Computer and Information Sciences, 1982: 341-356.
- [4] 杨红菊,梁吉业. 布尔加权关联规则的几种开采算法及比较[J]. 电脑开发与应用,2004,17(4):12-13.
- [5] Cook D, Holder L B. Graph-based data mining[J]. IEEE Intelligent Systems, 2000,15(2):32-41.
- [6] 陆建江. 加权关联规则挖掘算法的研究[J]. 计算机研究与发展,2002,39(10):1281-1286.
- [7] Woon Y K, Ng W K, Lim E P. A support-ordered trie for fast frequent itemset discovery[J]. IEEE Transactions on Knowledge and Data Engineering, 2004,16(7):875-879.
- [8] Mohamed Hammami, Youssef Chahir. WebGuard: Web based adult content detection and filtering system[C]. Proceedings of the IEEE/WIC International Conference on Web Intelligence. Computer Society, 2003,14(5):55-59.