doi: 10.12068/j.issn.1005-3026.2020.08.005

基于网页结构与语言特征的垃圾网页链接检测方法

杨 望,江咏涵,张三峰
(东南大学 网络空间安全学院,江苏 南京 211189)

摘 要: 现有的垃圾网站检测方法主要针对自建的垃圾网站,对于通过入侵正常网站注入垃圾网络链接的检测效率不高.本文提出一种基于网页结构与文本多维特征的检测框架,该框架将网页进行分块处理.通过计算优势率的方法提取内容特征,根据标签数、属性键和属性值利用独热率的方法提取结构特征.使用机器学习算法进行训练并得到检测模型,进而有效地检测垃圾网站链接.同时,将本文的检测方法与基于内容分析的检测算法和黑名单匹配算法进行对比,本文提出的方法检测准确率最高有13%的提高.

关 键 词: 垃圾网站检测;黑色SEO;独热率;机器学习;链接注入
中图分类号: TP 393.08 **文献标志码:** A **文章编号:** 1005-3026(2020)08-1091-06

A Web Spam Link Detection Method Based on Web Page Structure and Text Features

YANG Wang, JIANG Yong-han, ZHANG San-feng
(School of Cyber Science, Southeast University, Nanjing 211189, China. Corresponding author: YANG Wang, E-mail: wyang@njnet.edu.cn)

Abstract: The existing spam website detection methods are mainly aimed at self-built spam websites, and not suitable for injected spam websites because of the low efficiency of link detection. This paper proposes a new detection method, in which a detection framework is based on multi-dimensional features of webpage structure and text. The framework divides the webpage into blocks. Then content features are extracted by calculating odd ratio and structural features based on tags, attribute keys and attribute values are extracted by using the one-hot rate. The detection model is generated by proper machine learning and used to detect spam links. The detection accuracy of this framework is increased by up to 13%, compared with the algorithms based on content detection and on blacklist matching.

Key words: web spam link detection; black SEO; one-hot rate; machine learning; URL injection

随着互联网的发展,搜索引擎成为用户使用互联网获取信息的重要工具.根据2018年第四十二次中国互联网发展状况统计报告[1]指出,截至2018年6月30日,我国搜索引擎使用率为81.9%.Jansen等[2]在报告中指出,查找互联网资源时,用户通常只浏览前几页的搜索结果,因此,对于商业网站来说,在用户的搜索结果中获得一个靠前的位置是非常有意义的[3].为了获得更靠前的排名,搜索引擎优化(search engine optimization, SEO)技术因此产生.正常的SEO技术包括优化网站结构、网页代码和内容,以提高搜索排名,从而提高网站的访问量[4].但是也存在一些网站,特别是和色情、赌博、私服、药品等灰色经济相关联的网站,采用一系列方法来欺骗搜索引擎的排名算法.这种欺骗方法一般被称为黑帽SEO(Black Hat SEO)[5],这些使用欺骗方法的网站,被称为垃圾网站(web spam).

为了打击使用黑帽SEO技术的垃圾网站,搜

收稿日期: 2019-09-28
基金项目: 国家重点研发计划项目(2017YFB0801703);国家自然科学基金资助项目(61602114).
作者简介: 杨 望(1979-),男,安徽宣城人,东南大学讲师,博士.

索引擎采用了诸如谷歌公司的 Panda 算法[6]、百度公司的绿箩算法[7],在搜索结果中对垃圾网站进行过滤;因此很多垃圾网站的运行者除了自己搭建垃圾网站外,还尝试入侵教育、政府、正规企业的网站,并在被入侵的网站中插入指向垃圾网站的链接.由于搜索引擎无法判定指向垃圾网站的链接是正常的链接还是被恶意入侵植入的链接,因此这种方法可以有效规避搜索引擎的过滤算法.这种垃圾网站的行为一般还会同时伴有严重的网站入侵和数据泄露事件,相比于传统的垃圾网站技术,给互联网带来了更大的安全威胁,因此本文针对此类威胁提出了一种基于网页结构和内容的多维分类算法来进行检测.

目前的垃圾网站检测算法主要针对垃圾网站运行者自建的内容式垃圾网站和链接式垃圾网站,对于这种入侵攻击式的垃圾网络链接检测则更多地采用链接的隐藏特征来进行检测[8].由于新的隐藏特征可以不断被生成,因此这种方法虽然对已发现特征的隐藏链接检测效果比较好,却无法检测到采用新的隐藏链接特征的垃圾网络链接.

本文的主要贡献可总结如下:①针对入侵式垃圾网站链接的检测问题,提出了一种基于网页结构与文本多维特征的检测框架,该框架利用独热率和机器学习方法有效地检测垃圾网站链接威胁;②使用该方法针对 CERNET 内含有垃圾网站链接的服务器进行了一次比较全面的检测,通过实际使用证明该方法具有较好的检测效果.

1 研究背景和相关工作

1.1 垃圾网站类型

Gyöngyi 等[9]对垃圾网站技术进行了详细的描述.垃圾网站主要有三种类型(见表1),前两类通过修改内容或者链接干扰搜索引擎的排序结果;第三类则是各种隐藏垃圾内容的技术,比如隐藏文本内容、伪装文本内容和重定向文本内容等.

表1 垃圾网站的三种类型

Table 1 Three types of web spam

类型	原理	方法
内容 spam(content spam)	通过篡改网页内容信息欺骗搜索引擎进而改变排名	关键词堆砌,元标签填充等
链接 spam(link spam)	为网页构造复杂的链接结构进而改变网页的排名	链接农场,链接交换,蜜罐诱饵,维基百科 spam, Splogs
隐藏 spam(hiding spam)	把篡改的内容或者链接隐藏,使其变为搜索引擎爬虫可见但用户不可见	文本内容隐藏,Cloaking,重定向

1.2 已有检测技术

根据垃圾网站的分类,当前主要有三种检测方法,分别是根据内容进行检测,根据链接进行检测,以及对隐藏的垃圾信息进行检测[3],而这些垃圾网站检测方法主要是通过规则检测或机器学习检测.规则检测是指使用固定的规则进行检测,如关键词匹配.这种检测方法误报率比较高,如"金花"、"赌"这两个词都常用于赌博网站的检测,可如果这些词出现在其他地方,比如叫"金花"的人名或者有关于"赌博"的新闻,就会出现误报.随着机器学习的发展,现在许多垃圾网站链接检测都使用了机器学习方法,这些方法一般准确率高,误报率低.

1.2.1 基于内容的 spam 网页检测

根据内容进行检测主要是通过计算网页上的一些测度,对是否是 spam 网页作出判断. Ntoulas 等[10]提出的网页标题词数、网页总词数、锚文本比例、流行词占比等,将检测 spam 网页作为分类问题,运用决策树的技术来判断网页是否是 spam. Fetterly 等[11]分析了 spam 的特点,如重复的网页信息、网页内容的进化信息、域名解析、URL 属性等,以此作为判断 spam 的统计特征.

1.2.2 基于链接的 spam 网页检测

根据链接进行检测已有一些著名的算法,如 Gyöngyi 等[12]提出的 TrustRank 算法,即通过人工选择信任度高的网页种子,再沿着网络图中的链接传播,获取途中各个网页的信任度的数值,从而能够判断网页是否为 spam.

但是由于 TrustRank 算法更适用于规模大的网络社区,同时种子集的主题选择不可能全面,所以 Gyöngyi 等[13]又提出了基于 TrustRank 和 PageRank 的差异来检测 spam 网页.

1.2.3 隐藏 spam 网页检测

Wu 等[14]提供了对于使用重定向技术的隐藏 spam 的检测方法.针对每一个网页,先后爬取三次得到三个版本,其中一次的爬取是遵循浏览器

模式,然后通过使用的术语的差异和提供的链接的差异来计算三个版本之间的差别. 如果短时间内浏览器版和爬虫版本之间的差异大于两个爬虫版本之间的差异,那么可以认定网页是一个 spam 网页. 但是该方法仍具有消耗大、爬虫行为异常等问题. 周文怡等[8]则利用 html 标签中用于隐藏页面内容显示的标签作为特征来训练检测程序,对使用已知标签的垃圾网站链接具有较好的检测效果.

2　检测框架

现有的针对链接进行检测的算法都需要获得网页与外部网页的链接,也就是需要得到检测网页的入度和出度; 而本文研究的算法主要针对组织内部的网页进行检测,这些网页一般不会与外部网站有链接,所以针对链接的检测方法并不实用; 同时,对基于内容分析的检测算法进行了实验,发现 Ntoulas 提出的测度对中文网页的检测也并不准确. 因此,针对现有算法的缺点,本文提出一种基于网页结构和内容多维特征的垃圾网站链接检测算法. 该算法首先对网页进行结构分块处理,然后生成对应的多维结构和内容特征向量集,并使用合适的机器学习算法进行训练,最终得到检测模型,算法流程如图 1 所示.

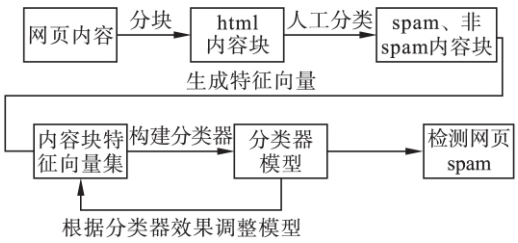


图 1　检测算法流程

Fig. 1　Process of detection algorithm

2.1　网页分块

由于被嵌入链接的网页一般不会全部都是垃圾内容,垃圾网站链接往往只存在于网页的一小部分,如果对网页的全部内容进行建模,非 spam 的词语会对检测质量有很大影响. 所以需要对网页源代码进行分块处理,从而更有针对性地检测出网页中是否存在垃圾网站链接.

根据 html 语言的性质,以及垃圾网站链接内容的特点,一个分块需要包括标签树的每一个节点的标签、叶标签中包含的文本内容,以及叶标签的属性(叶标签为 html 树中没有子标签的标签) .

因为网页源代码一定包括 head 和 body 两部分,所以先将源代码分为这两大块. 在 head 这一大块中,只需要对 meta 和 title 的信息进行提取和处理. 而在 body 这一大块中,就将叶标签作为一块,提取出其标签树、内容和叶标签的属性. 将每一块的信息存入文件系统中,格式见图 2. tag 表示标签树,elem 表示块的内容,attrs 则表示这一块的属性.

```
tag:-a-li-ul-div-body-html-[document]
elem:网络博彩公司
attrs:{u'href':u'/wlbcgs/'}
```

图 2　块存储格式

Fig. 2　Format of a web block

2.2　生成网页块的特征向量

2.2.1　生成网页块标签树的特征向量

因为标签之间的父子关系对 spam 的判断没有贡献,因此不需要单独考虑标签树中标签的父子关系. 又因为 html 语言中标签的数量是有限的,所以首先将每一个标签标号,接着如果块的标签树中有编号则记为 1,没有编号则记为 0,生成块的特征向量.

2.2.2　生成网页块叶标签属性的特征向量

叶标签属性以键值对的方式存入,和值一一对应. 键的数量也是有限的,因此也使用同样的方法,将每一个属性键编号; 如果块中存在某属性键,则为这个属性键生成一维特征向量,否则生成零向量.

属性的值有五种类型: 数值、链接、键值对、自定义字符和 html 内部值. 自定义字符是指用于内容或标题或为当前标签做的记号值,html 内部值是指属性本身具有枚举类型的值. 以上五种类型的值需要进一步处理(见表 2) 再生成特征向量.

表 2　不同值的预处理

Table 2　Preprocessing of different values

类型	预处理
数值	无需考虑具体值,直接以标识该键是否存在作为一个特征项
键值对	把内部值与键作为一个特征项
链接	无需考虑具体值,直接以标识该键是否存在作为一个特征项
自定义字符	自然语言处理后生成特征项
html 内部值	因为 html 内部值数量有限,所以给其编号后以有为 1 无为 0 的方式作为特征项

2.2.3　拼接和处理特征向量

因为独热编码方式能够很好地处理非连续性特征,所以使用独热编码将已经生成的特征向量按照标签树、属性键和属性值拼接成一个特征矩

阵. 使用机器学习的方法,要求对于不同的对象,应该有固定的、相同数量的特征,即要求特征矩阵的列数必须是一个定值; 所以对生成的特征矩阵再处理,将缺少的属性用0补全,从而得到每一个块的具有固定列数的特征矩阵,再把每一块的特征矩阵按行拼接,得到整个训练数据的特征矩阵. 特征矩阵的第一列使用1或0标识该块是否为spam,即机器学习的输出结果. 将特征矩阵以csv的文件方式存储,便于后期使用.

2.2.4 提取网页块中文本内容的特征

本文选择优势率方法作为文本内容的特征. 优势率是用于二元分类的一种分类算法,可以更好地突出正向类而弱化负向类,在本算法中也就是更好地突出spam网页,弱化非spam网页. 优势率的二元性质使其特别适合用于web spam词汇的特征提取. 优势率计算方法会给只出现在正向类中而几乎不出现在负面类中的词条打高分. 而且优势率计算方法只考虑了词条在所有文档的出现频率,并没有使用词条在一个文档中的出现频率,因此优势率计算方法不倾向于选择高频词作为文本的特征,这符合spam词汇的特征,因为web spam词汇往往只出现在spam网页中,其总体频率较低. 优势率计算如下:

$$\mathrm{OR}=\frac{A(D-B)}{B(C-A)}. \tag{1}$$

式中: OR为优势率; A 为词条m在spam的网页块中出现的个数; B 是词条m在非spam网页块出现的个数; C 是所有spam网页块的个数; D 是所有非spam网页块的个数.

算法首先将网页头部的文本提取出来,分词并计算平均优势率作为头部的代表优势率; 然后提取每一个网页块中的文本,分词并计算平均优势率作为该块的代表优势率. 算法将比较网页块优势率、可能性阈值和绝对阈值来判断该网页块是否为spam,如果平均优势率在两个阈值中间,则结合网页结构模型检测,逻辑如图3所示.

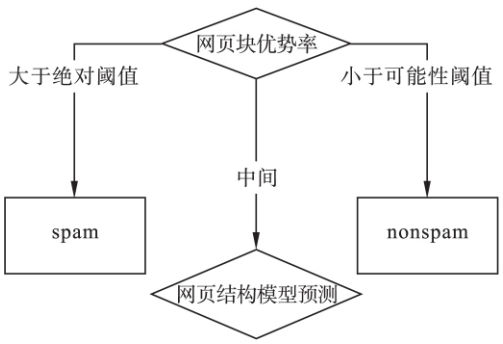


图3 垃圾网页的判断逻辑

Fig. 3 Logic for determining web spam

2.3 模型学习

因为模型的实现使用了增量学习的方法,所以模型和算法对新数据需要有学习的功能,分为纠错性学习和不存在性学习.

纠错性学习针对网页结构模型,当根据结构特征判断的结果和根据内容特征判断的结果不一致时,需要采用纠错性学习. 即某一块计算出的优势率大于绝对阈值,结构模型却预测该网页块不是spam; 或者,某一块计算出的优势率小于可能性阈值,结构模型却预测该网页块是spam. 这两种情况下,结构模型需要被纠正,即采用sklearn函数向旧模型输入新数据得到更新的模型,并用此模型进行下一次检测.

不存在性学习针对自然语言模型,即优势率字典. 检测的样本中一定存在优势率字典中没有的词语,此时便需要使用不存在性学习,根据判定结果将这种词语和对应的优势率存入优势率字典,进行自然语言模型的更新. 针对已经存在的词语,只需要更新优势率字典中该词语与spam样本一起出现的次数和与非spam样本一起出现的次数,并更新其新的优势率.

设 A 是词语m和spam样本一起出现的次数, B 是词语m和非spam样本一起出现的次数, $\mathrm{OR}_{\mathrm{new}}$ 指新的优势率,则对于优势率字典中已经存在的词语m,如果检测样本中,该词语所在的网页块被判定成spam网页块,则

$$B_{\mathrm{new}}=B_{\mathrm{old}}, \tag{2}$$

$$A_{\mathrm{new}}=A_{\mathrm{old}}+1, \tag{3}$$

$$\mathrm{OR}_{\mathrm{new}}=\left(1+\frac{1}{A_{\mathrm{old}}}\right)\times\mathrm{OR}_{\mathrm{old}}; \tag{4}$$

如果检测样本中,该词语所在的网页块被判定为非spam网页块, 则

$$A_{\mathrm{new}}=A_{\mathrm{old}}, \tag{5}$$

$$B_{\mathrm{new}}=B_{\mathrm{old}}+1, \tag{6}$$

$$\mathrm{OR}_{\mathrm{new}}=\left(\frac{1}{1+\frac{1}{B_{\mathrm{old}}}}\right)\times\mathrm{OR}_{\mathrm{old}}. \tag{7}$$

使用式(2) ~式(7),可以更新优势率词典中已存在的词语的优势率.

纠错性学习和不存在性学习使网页结构模型和自然语言模型在检测数据时也能得到更新. 使用增量学习方式适应新的数据,可以适应不断更新的spam种类,减少了重新训练的时间和工作量.

3　实　　验

3.1　实验数据集

本文使用的实验数据集,是通过对校园网出口处的流量镜像进行解析并爬取,从而获取的教育类网站的页面,共有 840 个网页; 通过人工检查和标记,将其分为 spam 网页和非 spam 网页. 同时,本文研究的算法需要根据 spam 和非 spam 网页块的特征提取自然语言模型和网页结构模型,所以在算法对训练集网页分块后,对数据集网页的每一块内容进行了人工标记,划分为 spam 块和非 spam 块,在这 623 个训练集网页中,共有 14 406 个 spam 块和 473 629 个非 spam 块. 算法通过分析这些网页块来提取模型并用于检测.

一般情况下,机器学习的训练集和测试集的比例应该为 7∶3,所以实验的训练集包含了 623 个网页,其中包含 504 个非 spam 网页,119 个 spam 网页. 测试集包含 217 个网页,其中包含 169 个非 spam 网页,48 个 spam 网页. 测试集和训练集都是随机从实验数据集中抽取的.

为了真实反馈不同算法之间性能的差别,需要保证除了算法不同,其他条件都相同. 因此,将相同的训练集和测试集分别用于本文算法及 Alexandros Ntoulas 算法,进行训练和测试. 由于黑名单匹配算法不需要训练,所以直接以测试集作为其数据集,进行对比实验.

3.2　测试标准

实验通过正确率、错误率、误报率和漏报率对算法进行评估. 当算法检测的结果与该网页人工标记的结果一致时,表明该网页被准确检测; 不一致则表明对该网页的检测错误. 算法将 spam 网页检测成非 spam 网页的情况属于漏报,将非 spam 网页检测为 spam 网页属于误报. 统计以下四种情况,并记为: A,spam 网页被分类为 spam 网页的数量; B,非 spam 网页被分类为非 spam 网页的数量; C,spam 网页被分类为非 spam 网页的数量; D,非 spam 网页被分类为 spam 网页的数量. 然后便可以通过式(8) ~式(11) 计算对应的正确率 R,错误率 W,误报率 FR,漏报率 MR,以此对算法进行评估,并以此为标准将几个算法进行对比.

$$R = \frac{A + B}{A + B + C + D}, \tag{8}$$

$$W = \frac{C + D}{A + B + C + D}, \tag{9}$$

$$\mathrm{FR} = \frac{D}{B + D}, \tag{10}$$

$$\mathrm{MR} = \frac{C}{A + C}. \tag{11}$$

3.3　各种算法的测试结果

3.3.1　本文检测方法测试结果

根据统计,本文算法(算法 1) 针对这个测试数据集,检测出 54 个 spam 网页,163 个非 spam 网页,其中正确分类的 spam 网页 A 有 45 个,正确分类的非 spam 网页 B 有 160 个,漏报的网页 C 有 3 个,误报的网页 D 有 9 个. 由此可以计算出,本文的算法正确率达到 0.944 7,错误率为 0.055 3,漏报率为 0.062 5,误报率为 0.053 2; 正确率较高,漏报率和误报率都比较低.

3.3.2　黑名单匹配算法测试结果

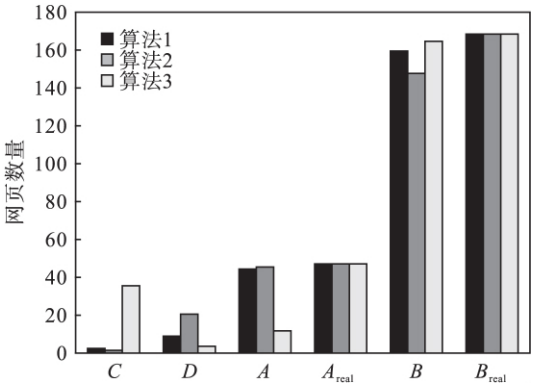
根据统计,黑名单匹配算法(算法 2) 针对该数据测试集,检测结果 A 为 46 个,B 为 148 个,C 为 2 个,D 为 21 个. 由此可以计算出,该算法检测的正确率为 0.894 0,错误率为 0.106 0,误报率为 0.124 3,漏报率为 0.041 7.

3.3.3　基于内容分析检测算法测试结果

根据统计,基于内容分析检测算法(算法 3) 针对该数据测试集,检测结果 A 为 12 个,B 为 165 个,C 为 36 个,D 为 4 个. 由此可以计算出,该算法检测的正确率为 0.815 7,错误率为 0.184 3,误报率为 0.023 7,漏报率为 0.75.

3.4　算法对比与分析

将以上三种算法的结果进行对比,如图 4 和表 3 所示.



A_{real} 和 B_{real} 分别是实际 spam 网页和非 spam 网页数量

图 4　不同算法测试结果对比

Fig. 4　Comparison of test results from different algorithms

表 3　测试结果对比

Table 3　Comparison of test results　%

算法	正确率	误报率	漏报率
算法 1	94.47	5.32	6.25
算法 2	89.4	12.43	4.17
算法 3	81.57	2.37	75

从结果可以看出,本文算法在正确率、误报率和漏报率上表现都比较好,能够较准确地检测出垃圾网页.

因为黑名单匹配算法对 spam 网页的检测比较严格,只要在网页中匹配到了黑名单中的关键词,就会将其归类为 spam 网页,所以虽然它能够准确检测到 spam 网页,漏报较少,但是其误报的网页也非常多,误报率比较高.

基于内容分析的检测算法效果最差. 文献[10]中,是对外文网页进行分析,并提取出相应的内容特征,投入分类器中进行检测. 这些内容特征,例如页面词数、标题词数,包括独立、条件 n - gram 可能性的计算,都是在分出的单词基础上进行提取的. 中文与外文有很大的区别,外文每一个单词就是一个词语,可以简单地通过空格和标点对外语进行分词,从而获取准确的词语,进而得到各个特征值. 然而,中文一个句子中,所有的词语都是连在一起的,不可能用某个标志进行分词. 算法使用的是 jieba 分词[15],jieba 分词精确度仍不能达到理想的情况;而且,由于训练集中,spam 网页的占比仅为训练集总数的 19%,会出现不平衡分类问题,即分类算法会偏向样本个数多的类别,导致正类样本的准确率较高,而负类样本的预测准确率则不理想. 这也解释了漏报率比较高,而整体准确率没有很低的原因. 同时,由于 Ntoulas 等的论文[10]发表较早,随着技术的更新换代,spam 网页和非 spam 网页在这些特征上已经相差无几,所以该方法检测出的效果是比较差的.

4 结 语

检测 spam 网页仍是一个具有挑战性的研究领域. 经过实验与对比,本文研究的算法已经可以基本解决 spam 检测的问题,准确率较高,又不会出现过高的误报率和漏报率. 同时,如果用户对误报率和漏报率有所要求时,也可以通过修正优势率的阈值来实现.

但是,算法仍有一些需要深入研究的问题. 后续工作需要更进一步扩大并完善初始数据,从而优化对优势率字典的构建. 同时,希望更进一步地研究优势率阈值的选取,从而更准确地对网页进行检测. 最后,网页中,块与块之间是有联系的,而本算法简化了这一部分,把块当作独立的个体进行处理. 所以,在后续研究中,可以将块与块之间的相似度,如结构相似度、文本内容相似度等加入特征的选取和计算.

参考文献:

[1] 中国互联网信息中心(CNNIC). 第42次中国互联网络发展状况统计报告[EB/OL]. [2019-08-19]. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201808/P020180820630889299840.pdf.
(China Internet Network Information Center. The 42nd statistical report on China's Internet development [EB/OL]. [2019-08-19]. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201808/P020180820630889299840.pdf.)

[2] Jansen B J, Spink A. An analysis of web documents retrieved and viewed [C]//The 4th International Conference on Internet Computing. Las Vegas, 2003: 65-69.

[3] 杨向军. Web spam 检测系统的设计和实现[D]. 广州: 华南理工大学, 2010.
(Yang Xiang-jun. Design and implementation of web spam detection system [D]. Guangzhou: South China University of Technology, 2010.)

[4] da Costa Carvalho A L, Chirita P A, de Moura E S, et al. Site level noise removal for search engines [C]//Proceedings of the 15th International Conference on World Wide Web. Edinburgh, Scotland, 2006: 73-82.

[5] Malaga R A. Search engine optimization—black and white hat approaches [J]. *Advances in Computers*, 2010, 78: 1-39.

[6] Google Inc. Google Panda [EB/OL]. [2019-07-15]. https://baike.baidu.com/item/%E7%86%8A%E7%8C%AB%E7%AE%97%E6%B3%95.

[7] Baidu Inc. Baidu Luluo algorithm [EB/OL]. [2019-07-18]. https://baike.baidu.com/item/%E7%99%BE%E5%BA%A6%E7%BB%BF%E8%90%9D%E7%AE%97%E6%B3%95/6023432?fromtitle=%E7%BB%BF%E8%90%9D%E7%AE%97%E6%B3%95&fromid=5994878&fr=aladdin.

[8] 周文怡, 顾徐波, 施勇, 等. 基于机器学习的网页暗链检测方法[J]. 计算机工程, 2018, 44(10): 22-27.
(Zhou Wen-yi, Gu Xu-bo, Shi Yong, et al. Detection method for hidden hyperlink based on machine learning [J]. *Computer Engineering*, 2018, 44(10): 22-27.)

[9] Gyöngyi Z, Garcia-Molina H. Web spam taxonomy [C/OL]. Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web [2019-08-05]. http://airweb.cse.lehigh.edu/2005/gyongyi.pdf.

[10] Ntoulas A, Najork M, Manasse M, et al. Detecting spam web pages through content analysis [C]//Proceedings of the 15th International Conference on World Wide Web. Edinburgh, Scotland, 2006: 83-92.

[11] Fetterly D, Manasse M, Najork M. Spam, damn spam, and statistics: using statistical analysis to locate spam web pages [C]//Proceedings of the 7th International Workshop on the Web and Databases. Paris, 2004: 1-6.

[12] Gyöngyi Z, Garcia-Molina H, Pedersen J. Combating web spam with trustrank [C]//Proceedings of the 30th International VLDB Conference. New York: ACM Press, 2004: 576-587.

[13] Gyöngyi Z, Berkhin P, Garcia-Molina H, et al. Link spam detection based on mass estimation [C]//Proceedings of the 32nd International Conference on Very Large Data Bases. [S.l.]: VLDB Endowment, 2006: 439-450.

[14] Wu B, Davison B D. Cloaking and redirection: a preliminary study [J/OL]. [2019-08-16]. https://www.researchgate.net/publication/303137682_Cloaking_and_Redirection_A_Preliminary_Study.

[15] Sun J Y. jieba [EB/OL]. [2019-07-28]. https://pypi.org/project/jieba/.