基于多维度特征的不良网站检测①

田双柱[1,2,3], 陈 勇[3], 延志伟[3], 李晓东[3]

[1](中国科学院大学, 北京 100049)
[2](中国科学院计算机网络信息中心, 北京 100190)
[3](中国互联网络信息中心 互联网络域名管理技术国家工程实验室, 北京 100190)

摘 要: 目前主要是通过基于 URL(Uniform Resource Locator)、关键词、图片等网页内容为特征的机器学习方法进行不良网站检测. 但是, 不良网站制作者也会通过更换 URL, 避免常见不良关键词的使用, 对搜索爬虫隐藏图片等做法来规避检测, 这使得基于内容的检测方法会有漏检的情况. 为了更准确的检测出此类网站, 本文提出了注册、解析方面的相关特征, 并通过最主流的机器学习方法构建了检测模型. 用模型预测新数据集, 结果证明, 基于解析和注册特征的检测方法可以有效的在网站集合中检测出前文提到的不良网站,并且对于一般不良也依然能够准确识别. 本次研究为不良网站的检测研究提供了又一思路.

关键词: 解析; 注册; 不良网站; 检测

Illegitimate Website Detection Based on Multi-Dimensional Features

TIAN Shuang-Zhu[1,2,3], CHEN Yong[3], YAN Zhi-Wei[3], LI Xiao-Dong[3]

[1](University of Chinese Academy of Sciences, Beijing 100049, China)
[2](Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China)
[3](National Engineering Laboratory of Internet Domain Name Management Technology, China Internet Network Information Center, Beijing 100190, China)

Abstract: The Web Information Extraction and Knowledge Presentation System is proposed to extract information from data intensive web pages. It downloads dynamic web pages, based on a knowledge database, changes them to XML documents after preprocessing, finds repeated patterns from them, by using a PAT-array based pattern discovery algorithm, recognizes their data display structure models, automatically based on the repeated patterns and an ontology-based keyword library, and then extracts the data and stores them in the knowledge database with the object-relational mapping technology of XML. Through these steps, web data is extracted automatically, and the knowledge database is also expanded automatically. Experiments on the traffic information auto-extraction and mixed traffic travel schemes auto-creation system showed that the system has high precision and is adaptive to web pages in different domains with different structures.

Key words: analysis; registration; illegitimate website; detection

随着网络融入大众生活, 网络用户呈低龄化趋势, 网络环境的净化变得更加重要. 目前对青少年产生严重不良影响的主要是一些涉黄、暴力、赌博网站. 相对于暴力、赌博, 黄色网站有更广大的受众, 传播更为广泛, 尤其对青春期青少年产生的不良影响更为严重, 调查称近半数青少年接触过黄色网站. 由于不良网站不断更新[1], 数量巨大, 靠人工筛选不良网站远远无法达到目的, 所以不良网站的自动检测是关键. 目前常用的检测算法大都是基于网站内容的[2]. 经过长期对不良网站的长期研究, 发现除了网页内容特征外, 这些不良网站在注册、解析层面都存在非常显著的特征. 因此, 本文从这些层面入手, 提取多方面特征, 进

① 收稿时间:2016-05-17;收到修改稿时间:2016-06-27 [doi: 10.15888/j.cnki.csa.005597]

行不良网站的检测研究. 本次研究的主要贡献如下:

1) 分析了不良网站在解析层和注册等方面的数据特征.

2) 提取了解析层和注册方面的特征, 并分析了多种模型的性能.

3) 用模型预测新数据, 证明了特征模型的有效性.

1 相关研究

针对涉黄、涉赌、涉暴等不良网站, 发现方式主要有人工举报和技术检测两种. 但是前者由于人工操作的限制, 只能发现很少数量的不良网站, 而后者是主要的处理方法. 目前国内外针对不良网站的过滤方法主要包括四种: 基于因特网内容分级平台(PICS)过滤、关键词过滤、数据库过滤以及基于内容理解的过滤[3,4]. 其中, PICS 过滤指的是, 网络评估系统安装色情、暴力、赌博等指标将网站进行分类、分级, 从而进行网站过滤. 但是, 基于实际情况的限制, 一些网站通过各种手段, 贴上与实际内容并不相符的分级标签, 导致这种过滤方式实际并不能起到良好效果. 基于关键词的过滤[5]是指建立不良关键词的词库, 然后根据基于规则的或者机器学习的方式, 检索网站关键词判断网站性质. 这种方式检索速度快, 但是不良网站可以通过将网络关键词改为健康网站关键词的方式来应对搜索引擎, 从而逃避搜索. 数据库过滤则是通过网站的 IP 地址、URL、代理商等信息, 建立黑、白名单的方式进行不良网站过滤. 这种过滤方式准确率高, 但是存在滞后性, 对于伪造 IP 地址, 更换 URL 等手段不能很好的处理. 而基于网站内容的过滤方法[6], 主要是根据网页文本内容、图片内容, 进行数据处理, 然后训练模型, 通过新模型来判断网站性质的方法. 这种方法准确率高, 能够达到较好的效果, 是目前最主流的过滤方法. 本次研究, 结合基于内容的过滤方法, 提出并分析了解析和注册等方面[7]的特征, 进行了相关研究和测试.

2 特征分析

在中国互联网络信息中心的网络监管工作中, 经常要处理大量的不良网站数据. 这些不良网站主要是涉及色情、赌博、暴力等内容, 而且其中部分不良网站也会采取更换 URL、规避使用常规关键词等手段避免检测. 为了对国内网站进行更好的管理, 我们统计了不良网站在注册、解析、运维等多个层面的数据, 并针对各个层面进行了总结、分析. 研究发现, 相对于健康网站, 不良网站在注册、解析等层面的数据都有不同之处. 将这些不同的数据进行提取, 我们得出了此次研究的特征集.

2.1 注册特征

本次训练的数据集采用的解析数据是 2015 年底的权威服务器数据, 所以训练数据集都是在解析日期前已经注册运营的网站. 对于训练集, 我们对于其注册年份, 给出数据统计.

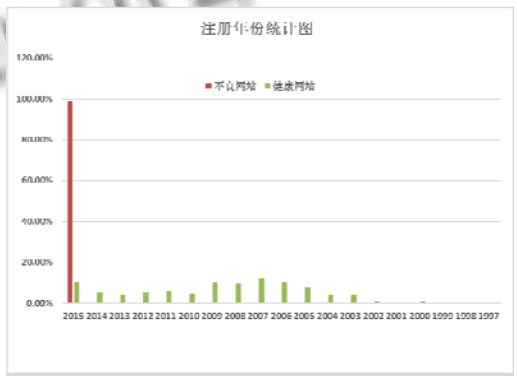

图 1 注册年份统计图

图 1 中横轴为网站的注册年份(按照距离现在的时间排列), 纵轴为对应的注册数量占总量的百分比. 我们可以看到, 健康网站注册年限呈现一个比较均匀的趋势. 相对于健康网站, 90%以上不良网站的注册年份为 2015 年, 部分在 2014 年和 2013 年, 呈现出注册年限普遍比较近的现象. 对这种现象进行分析发现, 普通网站希望将站点做到更好, 对于网站的维护是一个持续的过程. 而不良网站限于国内对网络环境的严格治理, 经常会采用定期更换网络 URL、更改网站关键词等方法来躲避管理. 所以不良网站的目的是快速的吸引网络流量[8], 而并不对网站进行长期的经营管理. 鉴于此, 不良网站的注册时间总是比较新, 而且不会续费太多年份. 基于以上分析, 提出了两个注册方面的特征. 注册年份和网站到续费截止时间存在的年份.

经过长期监管的数据分析, 不良网站的注册商相对于一般健康网站的广泛性, 大部分不良网站的注册商也更加固定, 对数据集网站的注册商进行统计分析.

图 2 中 A~U 代表训练集的主要注册商(具体商家名称不便公布), OTHER 则代表其他一些注册商. 图中

可以看到, 90%以上的不良网站集中在 A、B、C 三家服务商进行注册. 而相对于不良网站的注册集中化, 一般健康网站的注册商明显更为广泛, 分布更加均匀.

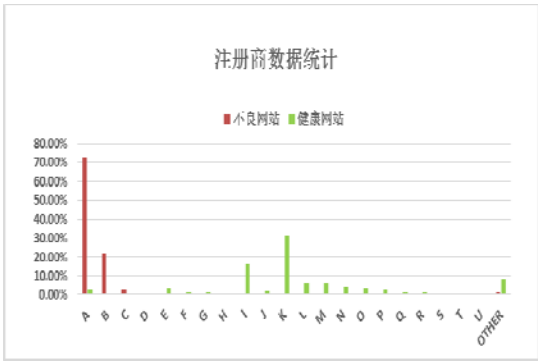

图 2 注册商数据统计图

2.2 解析特征

用户浏览网站时, 需要权威服务器对用户查询进行解析. 统计训练集网站的解析数据, 并对域名解析商进行数据统计分析见图 3.

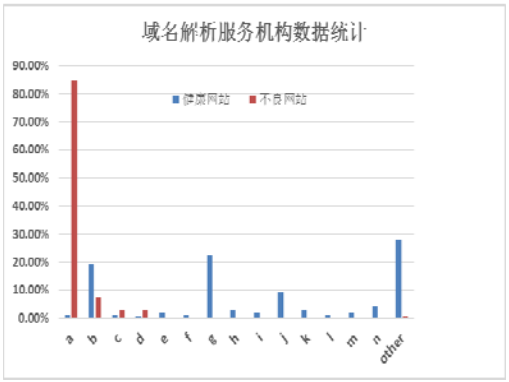

图 3 域名解析服务器数据统计图

图 3 中 a~n 代表解析网站域名的权威服务器名称, 从图中可以看出, 不良网站的解析服务器以 a、b、c、d 为主. 这几大服务器为训练集中 90%以上的不良网站提供解析服务. 而健康网站可以看到虽然也有部分解析商提供大量网站的解析, 但是解析商分布明显更加发散. 此外, 统计数据中还有部分数据没有在图表中显现出来. other 项, 15 个不良网站由其他服务商解析, 而健康网站有将近 500 个, 占总数约 30%由其他服务商解析. 更进一步体现了健康网站域名解析的分散性, 而不良网站会相对聚簇.

此次训练用的解析日志数据为.cn 权威服务器一天的数据, 处理解析数据, 保留 A 类查询信息, 然后对训练数据集进行数据统计. 分别提取了训练集网站的 24 小时的解析量, 以及当天的解析总量. 并按照解析总量对网站数据进行了分类统计.

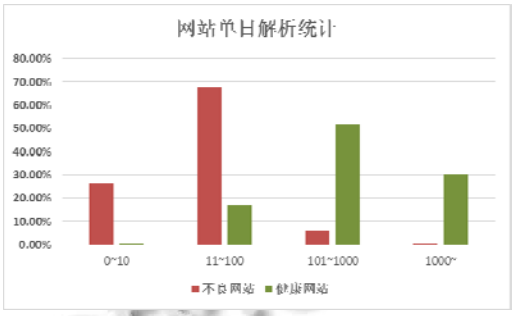

图 4 网站单日解析量统计图

根据提取日期当天的解析数据量, 对训练集数据进行分类统计. 从图 4 可以看出不良网站的日访问量相对健康网站偏低(解析日志为某个周三的数据). 而且, 大部分不良网站的解析次数在单日 100 次以内. 相对于不良网站, 健康网站解析次数更多, 单日 100 次以内的网站数量大概占 20%左右. 由此可见, 虽然不良网站会通过各种不良手段来吸引网络流量, 但是限于国内严格的网络监管环境, 不良网站的访问量还是普遍偏低的.

2.3 其他特征

为了使模型更加准确, 模型的训练在使用了解析数据和注册数据的基础上还添加了网站具体内容——title 特征集来训练模型. Title 特征集的获取步骤如下:

获取网站的 title 信息;

用 jieba 分词工具对 title 信息进行分词;

对分词结果去掉停用词, 取前 2000 的高频词汇;

对高频分词运用信息增益的方法进行特征选择并降维[9], 得到 title 特征词集合.

训练模型时, 对于每个域名的 title 信息, 根据 title 特征词集合, 转换为相应的特征词矩阵. 将特征矩阵用于模型训练.

此外, 不良网站相对于健康网站还有很多其他明显特征. 相对于国内对网络监管比较严格的大环境, 大部分不良网站的实际接入地址一般会选择外国. 针对网站 IP 的物理接入地址, 进行了数据统计.

通过图 5 可以观察到, 超过 90%的不良网站的接入地位为美国(US)或香港(HK), 而少部分(不到 1%)

在、中国大陆(CN)、法国(FR)等地. 而健康网站则大部分在中国大陆接入, 一部分在美国、香港接入, 非常少的一部分在日本等国家接入(图中由于部分数据量非常小, 所以柱形图并不明显). 可以得出, 接入 IP 地址这一特征也有很大价值. 此外, 对于国外接入的网站应该进行更加严格的监管.

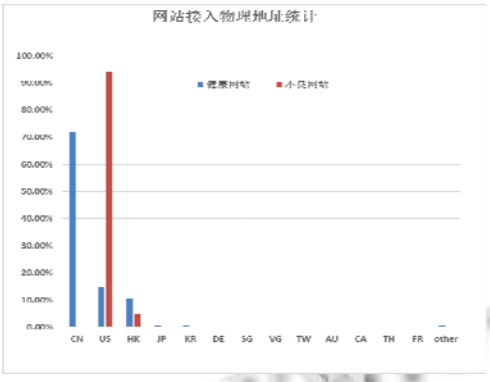

图 5　网站接入物理地址分布统计图

3　不良网站检测

基于上节对相关特征的分析, 本次研究建立了检测模型. 为了得到更准确的模型, 我们首先严格清理了之前采集的数据, 然后用目前最广为流行的 4 种机器学习算法(Naïve Bayes 算法[10]、C4.5 算法[11]、Logistic Regression 算法[12]、Random Forest 算法[13])进行建模, 通过 3 折交叉验证(将数据集随机分为 3 份, 以其中 2 份为训练数据, 另 1 份为测试数据, 3 份数据依次做测试数据, 取结果的平均值)的方法来验证模型准确度[14]. 并用其中的优秀模型对新数据进行预测, 进一步验证了模型准确性和特征的有效性.

3.1　评估参数

我们用 Precision, Recall, F1-Measure 以及 ROC 面积作为模型的评测指标. 下面基本介绍一下各指标及其参数.

$$\Pr ecision = \frac{TP}{TP + FP} \quad (1)$$

$$\mathrm{Re}call = \frac{TP}{TP + FN} \quad (2)$$

TP: 在样本中为正类, 并且被模型预测为正类的样本数量.

FP: 在样本中为负类, 但被模型预测为正类的样本数量.

FN: 在样本中为正类, 但被模型预测为负类的样本数量.

TN: 在样本中为负类, 被模型预测为负类的样本数量.

ROC 面积: 以 FP RATE 为横坐标, TP RATE 为纵坐标, 绘制坐标曲线, 所围成的面积.

$$TP\ Rate = \frac{TP}{TP + FN} \quad (3)$$

$$FP\ Rate = \frac{FP}{TN + FP} \quad (4)$$

$$F1 - Measure = \frac{2 \times \Pr ecision \times \mathrm{Re}call}{\Pr ecision + \mathrm{Re}call} \quad (5)$$

3.2　模型评估

此次实验测试使用数据为 CN 顶级域下的域名数据, 我们获取了.cn 权威服务器的大量日志数据. 数据集分为不良网站数据集和健康网站数据集. 在网站的监管工作中, 通过人工标注的方法, 得到不良网站数据集. 数据集中, 多为一般不良网站, 部分则为采用隐藏图片等手段规避检测的不良网站. 健康网站的数据集则来自于 Dmoz 目录. Dmoz 是一个开放式的分类目录, 由全球各地的志愿者共同维护, 提供比较优良的网站目录. 训练数据集共包含不良网站 1807 个, 健康网站 1778 个.

针对训练数据集, 实验用机器学习常用软件 weka 进行模型构建和测试. 上节的参数为测试指标. 具体评估结果见表 1.

表 1　分类模型交叉验证结果

分类算法	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	网站类别
Naive Bayes	0.991	0.134	0.883	0.991	0.934	0.937	不良网站
	0.866	0.009	0.99	0.866	0.924	0.987	健康网站
C4.5	0.964	0.042	0.959	0.964	0.962	0.975	不良网站
	0.958	0.036	0.963	0.958	0.961	0.975	健康网站
Logistic Regression	0.944	0.047	0.953	0.944	0.948	0.975	不良网站
	0.953	0.056	0.943	0.953	0.948	0.978	健康网站
Random Forest	0.98	0.034	0.967	0.98	0.973	0.997	不良网站
	0.966	0,02	0.979	0.966	0.972	0.997	健康网站

通过表 1 数据可以看出, C4.5 算法和 Random Forest 算法查准率(Precision)、查全率(Recall)高, 其他指标也非常良好. 此外, 训练集中的不良网站, 不仅包含一般的不良网站也包含那些采取各种手段、更加隐蔽的不良网站. 由此可见, 在基于内容的特征基础上, 添加了注册、解析等特征建立的检测模型可以非常精准的检测出各种不良网站, 不仅解决了单纯基于内容的检测容易被规避的问题, 也没有影响模型对于一般不良网站的检测效果.

3.3 新数据预测

将新标记的不良网站 534 个, Dmoz 中未在训练集中出现的健康网站 1373 个作为测试集. 对测试集的相应特征进行提取, 整理成测试数据矩阵. 由于随机森林算法在之前的交叉验证中各项指标都非常优秀, 使用随机森林算法建立好的模型对测试集进行测试, 依然使用前面使用的评估指标, 得到表 2 的混淆矩阵和表 3 的测试结果.

表 2 预测结果的混淆矩阵

实际分类 \ 预测分类	健康网站	不良网站
健康网站	1353	20
不良网站	17	517

表 3 随机森林模型预测新数据结果

TP Rate	FP Rate	Precision	Recall	F1-Measure	ROC Area	class
0.968	0.015	0.963	0.968	0.965	0.996	不良网站
0.985	0.032	0.988	0.985	0.987	0.996	健康网站
0.981	0.027	0.981	0.981	0.981	0.996	平均值

通过测试结果可以看到, 使用随机森林建立的模型预测新数据, 表现良好, 充分证明了所提出特征的有效性. 进一步分析可以发现, 不良网站在解析商、注册商、IP 接入地址等特征均体现出聚集性, 注册年限一般也都在 3 年以内, 而健康网站在各特征上的分布均比较均匀. 与这种现象相对应, 随机森林算法算法由于其本身特点, 可以有效解决不平衡问题, 且抗噪强, 不易出现过拟合. 综上分析, 随机森林算法尤其适合于基于所提出特征的数据集分类, 所以此次建模和预测都非常成功.

4 总结

本文分析了不良网站在注册、解析层面与健康网站的区别, 并基于此, 提出了新的分类特征. 然后分别使用朴素贝叶斯、决策树、逻辑回归、随机森林算法对数据集进行建模和交叉验证. 最后用随机森林模型预测新数据, 得到了非常好的预测结果. 实验证明, 新模型可以精准的检测出各类不良网站, 其中既包括隐蔽的不良网站, 也包括一般的不良网站. 这也证明了新特征的可用性. 下一步的工作是将这些特征更有效的和网页特征相结合, 并实际应用于不良网站的日常检测当中.

参考文献

1 关超.网络敏感信息过滤技术研究与实现[硕士学位论文].郑州:解放军信息工程大学,2009.

2 尹显东,唐丹,邓君,李在铭.基于内容的特定特像过滤方法.计算机测量与控制,2004,12(3):283–284.

3 王子强,张文阁,王洪艳.基于内容的网络异常信息过滤.硅谷,2012,9(18):9–10.

4 Theodoridis S, Koutroumbas K. 李晶皎等译.模式识别.北京:电子工业出版社,2006.

5 何苗,全宇.基于关键词的文本内容过滤算法的改进.微计算机应用,2007,(8).

6 邹飞.内容过滤关键技术研究.科技信息(科学教研).2008,(6).

7 巫锡洪,刘宝旭,杨沛安.基于域名的僵尸网络行为分析.信息网络安全,2013,(9):10–13.

8 Liu YQ, Chen F, Kong WZ, et al. Identifying web spam with the wisdom of the crowds. ACM Trans. on the Web (TWEB), 2012, 6(1): 13–15.

9 Theodoridis S, Koutroumbas K. Pattern Recognition. Academic Press, Inc., 2009.

10 Elkan C. Naive Bayes learning [Technical Report]. Department of Computer Science and Engineering, University of California. 1998.

11 Quinlan JR. C4.5 Programs for Machine Learning. San Mateo CA: Morgan Kaufmann, 1993.

12 Han JW, Kamber M, Pei J.范明,孟小峰译.数据挖掘概念与技术.北京:机械工业出版社,2012.

13 Breiman L. Random forests. Machine Learning, 2001, 45(1): 5–32.

14 Olivier Dubrule. Cross validation of kriging in a unique neighborhood. Journal of the International Association for Mathematical Geology, 1983, (6): 687–699.