计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(8):119−125 [doi: 10.15888/j.cnki.csa.006512]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

电子商务网站违法行为监管平台①

高学勤[1], 王　涛[2]

[1](广东工贸职业技术学院, 广州 510510)
[2](华南师范大学, 广州 510631)

摘　要: 针对当前网络交易违法现象人工识别、排查难的监管问题, 本文通过运用互联网搜索引擎、文本检索、数据挖掘以及 bloom 过滤技术设计了电子商务网站违法行为监管平台. 该平台支持对电子商务网络经济主体网站的识别与建库、网络违法经营行为与网络违法广告行为的自动排查、电子标识亮照与防伪检测等功能, 在实践应用中, 该平台充分表现出明显高于人工监管的效率与优势, 提升了网络违法案件的查处效率.

关键词: 电子商务; 电子标识; 行为监管; 搜索引擎

引用格式: 高学勤,王涛.电子商务网站违法行为监管平台.计算机系统应用,2018,27(8):119–125. http://www.c-s-a.org.cn/1003-3254/6512.html

E-Business Website Illegal Behaviors Supervision Platform

GAO Xue-Qin[1], WANG Tao[2]

[1](Guangdong Polytechnic of Industry and Commerce, Guangzhou 510510, China)
[2](South China Normal University, Guangzhou 510631, China)

Abstract: According to the manual recognition and identification difficulty of the E-business illegal behavior, an E-business supervision platform is developed through the use of the internet search engine, text retrieval, data mining, and bloom filtering technologies. This platform supports the identification and establishment of the E-business economic subject database, the illegal E-business behavior and advertising behavior search. It also realizes the electronic identification display and anti-fake detection. This platform shows better efficiency and superiority than manual working in practice.

Key words: E-business; electronic identification; behavior supervision; search engine

随着信息技术的迅速发展, 电子商务已经从一种边缘经营业态发展成为一个新型产业, 并逐渐成为与实体市场并驾齐驱、推动经济社会发展的新动力、新引擎. 尤其是近年来, 我国电子商务发展迅猛, 据中国互联网络信息中心监测[1], 截至 2016 年 12 月, 我国网购用户达 4.67 亿, 同比增长 12.9%, 全年新增网民 4299 万人, 增长率为 6.2%, 我国网民规模已相当于欧洲人口总量. 巨大的电子商务市场空间也给从事电子商务业务的网商提供了更多的发展机会.

但在电商企业蓬勃发展的同时, 在网络购物与消费过程中的消费欺诈、假冒伪劣商品、以次充好、虚假宣传等违法行为也层出不穷. 国家工商总局网站曾经披露和公布了九类典型网络商品交易违法行为[2], 表明目前我国电子商务领域面临着繁重的监管压力, 一定程度上也制约了电子商务行业的健康发展[3–5].

对于电子商务领域的监管, 目前国内还处于探索和起步阶段, 对网站的监管主要依赖于人工力量, 出错率高且费时费力, 而国外由于个人信用体系的完善以及监管体制的差异, 其监管思路也很难有效应用于国内[6]. 针对我国信用体系尚未完全建立的国情, 对电子

① 基金项目: 广东省工商局课题 (GPCGD103117FG257F)
Foundation item: Project of Industrial and Commercial Bureau of Guangdong Province (GPCGD103117FG257F)
收稿时间: 2017-12-25; 修改时间: 2018-01-16; 采用时间: 2018-02-07; csa 在线出版时间: 2018-07-28

商务的监管必须摆脱传统的人工方式, 应当立足于计算机智能化的识别与处理, 监管重点是如何识别与发现从事违法交易的商家和网站, 将各监管结构辖区内的互联网网站作为发掘目标, 充分运用现代数据搜索和处理技术, 建立一个识别率高、信息完备的电子商务经济户口主体数据库, 从而为政府更好履行监管职责提供可靠的保障. 为此本文设计并实现了某省电子商务网站违法行为监管平台.

1 系统概述

本节从系统关键需求与总体功能框架两方面进行阐述.

1.1 系统关键需求

根据调查研究, 对电子商务网站违法行为的监管需要重点解决以下 4 个需求难点.

需求一: 网络经济户口主体识别与建库难. 我国互联网网站的申办目前主要由各省市的通信管理部门负责, 由于部门职能差异, 通信管理部门审核时重点关注网站的域名以及开办者有关联系信息, 通常不会关注网站是否从事商品交易等商务活动行为, 而对电子商务的监管主要落实在工商和商务部门, 但在企业主体登记中往往又不涉及网站开办情况, 更多靠人工等其他手段来搜集和获取网站开办的信息[7]. 这就导致网站开办信息一定程度上存在不一致或缺漏现象, 容易造成监管缺位, 监管对象模糊不清等诸多问题[8,9]. 为此, 在设计过程中重点要考虑如何在庞大的互联网网站中区分出电子商务网站与一般非交易型网站, 并建立一个完备的电子商务网络经济户口主体数据库, 这样才能找准监管对象.

需求二: 应对网站不断更新变化难. 由于互联网经营性网站的生存周期短、更新速度较快, 网站的申办者随时都可能转让到第三方, 甚至直接关闭网站, 如果对监管对象实行静态化管理, 那么可能导致已经搜集信息库的快速陈旧和过时, 在监管需要使用信息时就会因网站信息失效而废弃.

需求三: 识别违法行为难. 识别互联网违法行为需要建立一套标准和规则, 否则识别的潜在违法行为大量需要人工的参与就会降低系统的可用性和有效性. 同时平台还要能依据标准和规则挖掘对象网站中是否存在潜在的违法行为, 将监管尽早的介入到早期预防与指导工作中. 另外违法广告的识别与一般文字型违法判断不同, 通常广告是以图形的方式发布的, 不能直接利用文字判断的方法.

需求四: 为了保护消费者网购的合法权益, 有必要采取措施将合法经营主体网站与非法网站加以分类区分, 从而减少消费者上当受骗的机会. 目前电商网站的分类或信用级别标识主要由行业或企业自发进行分类评级, 在可信度上存在一定的问题.

1.2 系统总体结构分析

针对上述监管需求和难题, 本文进行了一系列创新和改进, 最终设计并形成了如图 1 所示系统总体结构框架: 系统总体上分为面向监管人员的政务内网应用, 面向公众与企业的互联网应用以及基础资源三大功能应用. 政务内网应用中网络经济户口识别建库和网络经济户口更新两个模块用于形成待监管的网站数据库, 是其他模块的数据来源和基础; 网络经营行为监管和网络违法广告监管模块根据网络经济户口库进行违法线索的识别和判定, 是整个系统的核心功能; 电子标识审核发布与检查模块则为网络经济户口建库提供了另外一个数据补充渠道. 互联网应用中包括了网站电子标识申请、相关信息发布以及消费投诉等功能模块, 是平台对外服务的窗口, 同时也与政务内网应用模块之间形成交互, 内外网之间进行逻辑隔离, 防止信息泄漏; 平台引入互联网网站扫描引擎以及工作流引擎作为基础公共资源, 为整个平台各功能模块的运转提供服务.

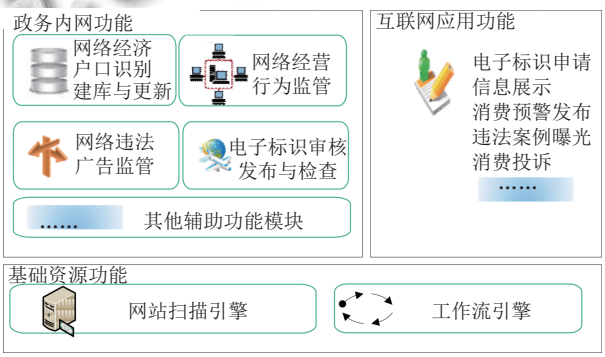

图 1 总体结构框架图

该平台基于.Net 开发技术, 在生成部署中分别采用应用服务器、数据库服务器、采集服务器各 2 台, 实现热备和负载均衡, 网站搜索扫描爬虫服务器 5 台, 实现了对区域内网站全天候的实时监测.

2 系统设计

各主要功能模块系统设计阐述如下.

2.1 网络经济户口识别建库设计

网络经济户口包括网站经营主体的基本信息 (静态信息) 和监管信息 (动态信息) 两部分. 网络经济户口管理的目标是对网络经营主体的整个生命周期管理. 最终确定了 30 多项建库的数据指标, 并与各个渠道的数据源进行了一对一的匹配, 保证数据指标的一致性. 同时监管中所产生的动态数据 (是否有违法行为发生、何时进行了扫描检查、有无发生过案件处理等) 也要纳入到各个经济户口主体之中, 从而形成静态和动态数据统一刻画下的清晰电商网站监管轮廓图. 这个轮廓图是电子商务网站监管平台的基石, 也是开展网络行为监管、违法广告监管等模块的基础. 网络经济户口在建立过程中分为初始态和监管态两种状态, 初始态户口库的建立过程如下:

针对网站注册与备案集中在通信管理部门的事实, 根据部门之间的协商, 系统取得了定期从通信管理部门提取网站数据的权限, 这样将基本上覆盖全省境内主要的网站数据. 但这里的数据并不全面, 网站也可能加挂在各个第三方交易平台之上, 因此也需要打通与平台运营商的接口, 为此设计了用于导入交易平台中电商数据的数据接口, 实现与平台运营商的对接. 最后还有一部分数据是未进行过登记或备案的网站, 这部分网站往往是最容易产生违法行为的, 因此设计了一个针对省内 IP 地址段进行不间断扫描的网站扫描挖掘搜索引擎, 将疑似网站全部提取到数据库中, 最终三个数据来源的合集经过排重处理形成相对完备的初始态网络经济户口监管数据库 (如图 2 所示).

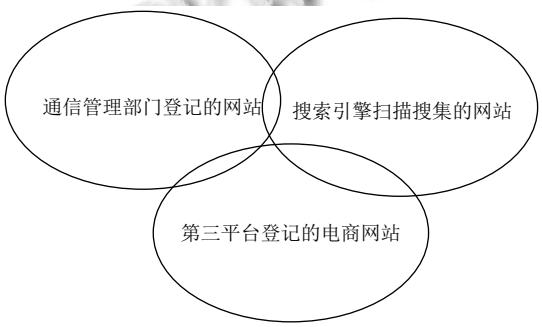

图 2 初始态监管数据库的组成

现在所采集的数据只是包含了省内全部网站的集合, 但这些网站并不一定都从事电子商务活动, 因为其中有相当一部分网站只是用于主题展示、宣传等非交易性行为, 根据有关法律法规, 这类网站暂不列入电子商务交易最终的监管对象. 为此需基于以下策略进一步准确筛出电商网站主体, 并将识别出的初始态对象基于以下判断流程纳入并形成监管态网络经济户口数据库:

(1) 第三方交易平台提供的网站不做排查, 直接转入监管态网络经济户口数据库.

(2) 如果获得的通信管理部门网站数据能够匹配企业登记的网站数据则匹配的初始态网站转入监管态网络经济户口数据库.

(3) 如果搜索引擎挖掘获得的网站能够匹配企业登记网站数据则匹配的网站数据转入监管态网络经济户口数据库.

(4) 如果上述 (2)(3) 两种情况搜集的网站找不到对应的企业登记数据, 则根据特别的主题词 (如包含销售行为的主题词“价格”、“售价”、“促销”等) 进行系统筛选, 将可疑初始态网站筛出并由人工进行最后确认与核对, 确认无误后网站再纳入监管态网络经济户口数据库.

2.2 网络经济户口更新设计

针对已入库网络经济户口数据可能频繁更新或者随时失效的互联网应用现状, 系统拟采取自动化判别与人工审核相结合的策略, 以不断提高监管态网络经济户口数据信息的准确性, 保持监管态数据库的实时有效性, 其策略如下:

(1) 系统与通信管理部门和第三方交易平台最新的数据进行横向比较, 如果有信息的更新, 则及时同步到监管态数据库.

(2) 纵向上通过监管态网站的域名数字证书自动识别网站域名的生存周期, 当域名证书失效时自动提请人工进行关注处理, 表明网站可能关闭或转让. 获取网站的域名证书前期主要通过企业的申报, 后期则计划采取与 CNNIC 等大型域名服务商进行合作, 完成域名证书信息有效性的自动匹配.

(3) 如果存在 CNNIC 无法覆盖的国际注册域名, 在系统实现时考虑以半年为周期, 对国际域名的网站进行一次定期人工判别处理.

2.3 网络经营行为监管设计

网络经营行为监管作为监管平台的一个重要模块,

它承载了整个系统的日常监测与处理工作. 在建立了完备且实时的网络经济户口主体数据之后, 采取结合网络爬虫与基于特征库的自定义搜索策略的数据搜索技术能有效挖掘出可能存在的违法行为. 其中搜索特征数据库的建设是该模块的重点, 也关系到搜索的质量. 通常搜索特征库由监管人员进行搜集和汇总, 在经过一系列验证后导入到监管系统中, 其设计策略如图 3 所示.

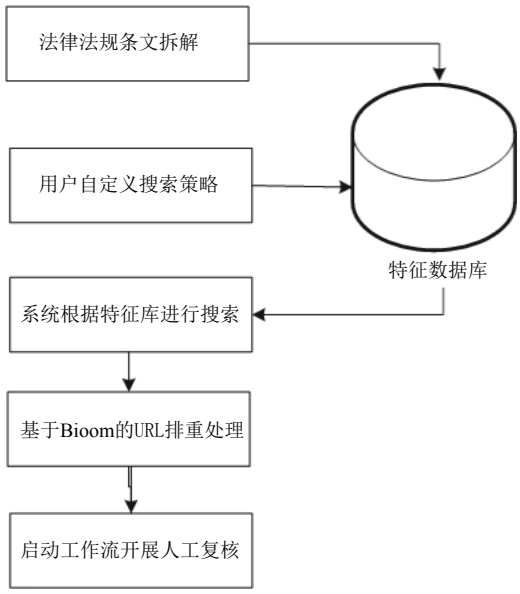

图 3 违法行为发掘与识别流程

(1) 将需要监管的违法行为法律法条拆解映射并转换到特征库中, 将其作为网络可疑线索的来源依据, 总体上可分解为销售违法商品、虚假宣传行为、傍名牌行为、以及其他违法行为等, 每一类行为下面再拆解为具体的违法表现. 比如在虚假宣传行为中将“个体工商户”或“合伙企业”向社会宣传为“公司”或“集团”, 将小规模的“经营场地”进行夸张宣传误导消费者等行为.

(2) 根据用户的自定义策略生成特征库. 用户的自定义策略通常采取违法关键词的录入方式, 由业务部门进行发起并执行, 根据执行效果, 特征数据库需要由监管人员进行不断的优化和改进, 才能提高搜索的效率和准确率. 为了方便监管人员理解搜索关键词的组合含义, 系统简化自定义搜索的关键词录入分为三部分: 必须同时包含的词、或者包含的词以及排除的关键词, 再由系统进行完备的词义转换. 例如图 4(图中数据仅用于案例讲解, 不涉及任何公司或个人的违法行为) 录入的搜索关键词经过翻译后将转换为图 5 所示的搜索关键词组合.

(3) 系统根据特征数据库定期对违法行为基于语义表达式策略进行网络爬虫式的检索和匹配. 传统的基于最大词汇相似度的方法[10]无法准确反映关键词相似的程度, 为此针对项目自身特点采取优化的基于权重的相似语义算法实现对搜索目标的关键字组合进行匹配, 达到根据特征数据库准确查找发现涉嫌违反法律法规规定的网络违法经营信息. 系统设计了 5 台搜索引擎服务器并发通过自动排重、搜索层次、搜索范围等设定进行不间断的扫描.

图 4 自定义关键词录入

(4) 由于搜索的 URL 链接数量较大, 系统设计通过多哈希函数映射的 Bloom 过滤器[11]实现对 URL 链接的快速排重处理, 并减少排重日志的占用空间, 多哈希映射过程原理如图 6 所示.

广东 且 减肥 且 绿瘦 且一天 且 不包含 新闻 且 讲义
或者 广东 且 减肥 且 瘦女郎 且一天 且 不包含 新闻 且 讲义
或者 珠海 且 减肥 且 绿瘦 且一天 且 不包含 新闻 且 讲义
或者 珠海 且 减肥 且 瘦女郎 且一天 且 不包含 新闻 且 讲义

图 5 翻译转换后的搜索关键词

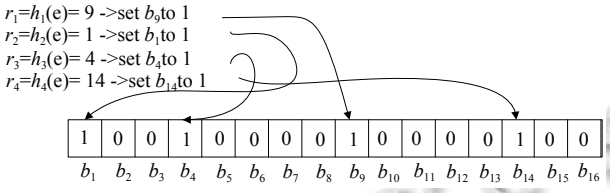

图 6 Bloom 过滤器哈希映射原理

(5) 搜索引擎将找到的可疑线索以工作流机制自动分发到全省各地管辖机关进行人工确认与排查, 并反馈处理结果. 人工判定核实的违法案件则转入案件系统进行后续处理.

2.4 网络违法广告监管设计

网络违法广告监管实际是违法行为监管的一个特例, 但广告的搜索具有自身特殊的规律和需求, 需要采用特殊的搜索技巧. 该模块主要从主体网站和交易平台上监测那些发布有夸大宣传、虚假宣传的广告信息, 并将相关的广告信息形成线索下发到相关的监管人员.

模块提供了多维度的关键字组合匹配与语义识别, 可以识别保健食品广告夸大治疗作用、把保健食品混同为药品、使用与药品相混淆的用语, 超出核准的保健功能范围以及宣传未经核准的功能等违法广告行为. 通过建立常用的广告识别词典, 识别广告中是否含有药品说明书以外的学术理论、观点等内容及使用他人名义保证或者以暗示方法使人误解其效用的宣传, 识别诸如增高、减肥等使用他人名义作保证和使用绝对化用语的广告.

网络违法广告监管与行为监管不同的地方是其通过对网站的链接和网页广告信息分析, 需识别出链接广告、图片广告、文字广告与软文广告, 除了通过特征库匹配关键词技术, 还需要配置广告图文识别规则以准确地识别广告单元. 此外对于轻微的违法行为和违法广告, 系统支持通过邮件、短信等形式将行政指导文书发送给电商主体, 起到警告预防的作用.

2.5 电子标识审核发布与检查设计

为了区分真假网站, 提高消费者的防骗意识, 凡是纳入监管态网络经济户口的主体网站均可以通过登录系统的互联网应用模块申请政府机构统一颁发的监管平台电子标识. 申请电子标识时, 需要网站主体提供相应的网站资质信息、网站域名证书等信息. 申请的信息经由属地监管人员审核. 由监管人员将登记信息与申请信息进行核对, 核查通过后系统生成一串加密的动态电子标识代码, 并通过邮件或短信的形式发送给网站, 网站技术人员将该标识代码按照相关的说明放置到网站的网页代码中, 并将标识加亮显示 (如图 7). 为了防止恶意伪造电子标识, 系统设计了只有加挂的域名和申请信息一致时平台才可以为网站点亮电子标识, 否则系统将提醒网站与申请不匹配, 网站伪造的标识或者域名不匹配标识也无法正常显示 (如图 8).

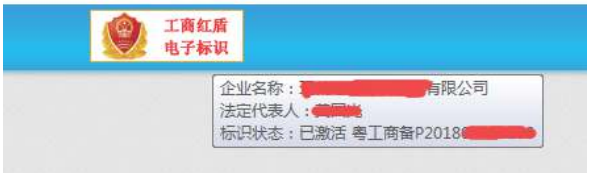

图 7 正常显示的电子标识

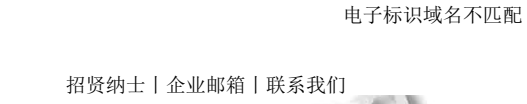

图 8 域名不匹配的电子标识

对于已经领取电子标识的电商主体, 系统依据预先设定的规则, 将定期对其进行自动检查, 获取其在网上标识的信息, 并与登记信息进行比对. 当发现标识信息与实际登记信息不符时, 系统能自动提醒监管人员进行检查, 等待做出进一步处理. 电子标识申报与检查流程如图 9 所示.

2.6 互联网应用功能设计

作为监管机构日常工作需要以及服务大众的需要, 系统也需要建立对外服务的窗口, 系统通过在互联网设立公众服务入口, 除向企业开放电子标识申请业务外, 还向公众与社会展示国家职能部门的法律法规、舆情监测信息、消费预警信息以及违法典型案例, 及时公示企业信用及违法企业黑名单, 并接受消费者的投诉举报.

3 应用效果

平台自上线以来, 已在全省 20 个市分别投入和使

用, 在电子商务监管工作中发挥了积极作用, 目前已经建立了近 10 万网络经济户口库, 为合法电商主体发布 3000 多个电子标识, 出色的完成了上级部门部署的多次区域内违法网站检查专项行动. 表 1 显示出平台应用后在年平均检查网站数量以及年平均查办网络违法案件量上比应用平台前的突出优势, 表明平台的应用大大提高了网站监管的效率以及处理违法案件的能力. 同时平台为电商网站年均发布电子标识超过了 1500 个, 并呈逐年递增趋势, 进一步净化了网络购物和消费环境. 图 10 显示出采用基于权重的相似语义匹配算法较普通的单纯关键词匹配方法在识别可疑违法网站准确率上有较大的提升, 但实践中由于关键词设置的准确性以及监管人员判定违法行为能力的差异, 对实验结果的准确率都会有一定的影响, 待关键词库优化后能进一步提升准确率, 总体上初步实现了平台设计目标.

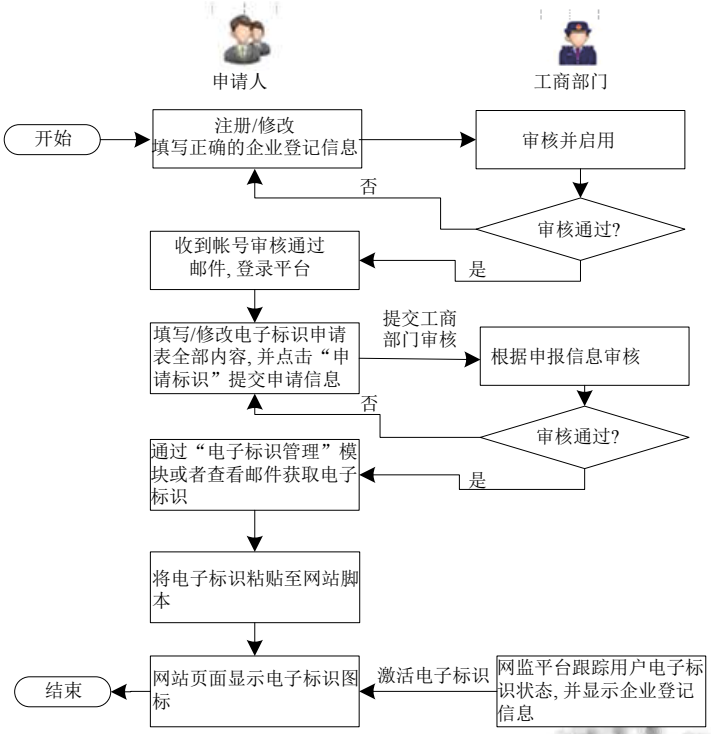

图 9 电子标识申报流程

表 1 平台应用前后效果对比

平台应用前后情况	年平均检查网站数量 (个)	年平均查办网络违法案件量 (件)	平均发布电子标识量 (个/年)
未应用监管平台前	8000	<300	-
应用监管平台后	>50 000	>1100	>1500

4 结语

电子商务网站违法行为监管平台融合了互联网智能搜索引擎、文本检索与分析等先进技术, 实现了对网络经济户口信息的采集、分析、汇总, 并通过对采集数据的分析、整理, 实现对电商主体经营行为的监测和处理. 平台的所有数据统一存储在省级部门, 并通过政务内网在市县区三级部署使用, 提升了对电商网站违法行为监管的效率和能力. 同时该平台很好的解决了电子商务经营主体的虚拟性、跨地域性、违法隐蔽性等突出问题, 破解了把监管视野和范围从传统经济向互联网络延伸和拓展的难题, 具有一定的借鉴和参考价值.

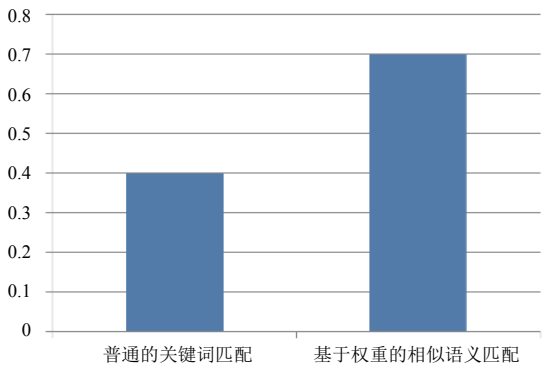

图 10 识别可疑违法网站准确率

参考文献

1 中国互联网络信息中心. 中国互联网络发展状况统计报告. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/ 201701/P020170123364672657408.pdf. [2017-01].

2 国家工商行政管理总局. 工商总局公布九类典型网络商品交易违法行为. 中国工商报, 2013–07–10(1)

3 张新民. 织好“全国一张网”推进市场监管新发展. 中国市场监管研究, 2016, (5): 40–41.

4 王健, 张原天. 强化市场监管促进电子商务健康发展. 工商行政管理, 2016, (21): 23–24.

5 王海军. 电子商务安全监管模式研究. 计算机安全, 2013, (9): 25–28.

6 国家工商行政管理总局市场规范管理司, 中国工商行政管理学会, 德国国际合作机构. 中德网络商品交易监管比较研究. 北京: 中国工商出版社, 2011. 78–100.

7 邓晖珞. 广州市工商行政管理局网络商品交易监管的案例研究[硕士学位论文]. 成都: 电子科技大学, 2013. 28–31.

8 吴琼衡. 我国网络商品交易监管制度比较研究. 商场现代化, 2016, (10): 10–13. [doi: 10.3969/j.issn.1006-3102.2016.10.006]

9 罗苏晨, 丁维, 廖思敏. 电子商务环境下的消费者权益保护-基于大学生网购维权调查分析. 技术与市场, 2013, 20(12): 272–273. [doi: 10.3969/j.issn.1006-8554.2013.12.173]

10 朱征宇, 孙俊华. 改进的基于《知网》的词汇语义相似度计算. 计算机应用, 2013, 33(8): 2276–2279, 2228.

11 田小梅. 多布鲁姆过滤器查询算法及其应用研究[博士学位论文]. 长沙: 湖南大学, 2013. 8–22.