基于机器学习的域名信用评价方法 *

陈 威[1,2,3], 王利明[1,2], 耿光刚[1,2], 毛 伟[1,2], 李晓东[1,2]
(1. 中国互联网络信息中心, 北京 100080; 2. 中国科学院计算机网络信息中心, 北京 100080; 3. 中国科学院研究生院, 北京 100080)

摘 要: 针对域名自身的特点和应用特点,建立一种基于机器学习的域名信用评价自动化方法并进行实验分析。实验结果表明,该方法具有较好的正确率,符合人们的一般认识,其评价结果可以作为域名诚信管理体系的参考依据。

关键词: 不良应用; 失信; 域名; 信用评价; 机器学习
中图分类号: TP391 **文献标志码**: A **文章编号**: 1001-3695(2012)02-0690-03
doi:10.3969/j.issn.1001-3695.2012.02.077

Evaluation method of domain name credit based on machine learning

CHEN Wei[1,2,3], WANG Li-ming[1,2], GENG Guang-gang[1,2], MAO Wei[1,2], LI Xiao-dong[1,2]
(1. *China Internet Network Information Center*, *Beijing* 100080, *China*; 2. *Computer Network Information Center*, *Chinese Academy of Sciences*, *Beijing* 100080, *China*; 3. *Graduate School of Chinese Academy of Sciences*, *Beijing* 100080, *China*)

Abstract: This paper created an automatic method, based on machine learning, to evaluate credit of domain names of their own characteristics and application features, and then made experimental analysis. As the experiment shows, the method can reach a good precision and its' results can be used as a reference of domain integrity and credit management system.
Key words: bad applications; lack of credit; domain name; credit evaluation; machine learning

0 引言

域名服务不仅用于解决地址对应问题,也是各种应用网站建设的基础。当今,域名已成为互联网文化的重要组成部分,也被誉为企业的网上商标。对域名进行信用评价能有效地体现域名的使用情况以及企业的信誉情况,促进企业信息化发展。而对互联网用户而言,了解域名的信用情况能够使自己加深对企业网站的诚信度认识,并获知域名的可靠性相关信息,在网络信息过滤、不良应用防范上获得又一可靠参考指标,为使用互联网提供导向。

国内外目前尚无专门针对域名信用评价的系统方法所进行的研究工作。本文对域名信用评价进行深入研究,通过采用有监督机器学习的方法,将直接的评价问题转换为间接的模式分类问题,从而设计出一种自动化的域名信用评价系统方法。

1 域名信用评价

信用评价[1]一般是指根据规范的指标体系和科学的评估方法,以客观公正的立场,对各类市场参与者(企业、金融机构和社会组织)履行其各种经济承诺的能力及可信任程度进行综合评价,并以一定的标志表示其信用等级的活动。广为熟悉的信用评价工作有个人信贷信用评价、网店信用评价以及企业信用评价等。

对于什么是域名信用评价,目前并没有一个明确的定义,结合一般信用评价的定义和域名本身特点,本文将域名信用理解为域名的可信任程度和域名的应用服务质量,并使用一个具体的数值来标志域名信用,为域名信用评价工作奠定基本方向。

2 基于机器学习的信用评价模型

目前图像质量评价、Web 内容质量评价等方面的评价工作普遍采用了机器学习的方法[2~4],在个人或企业信用评估中机器学习的方法也起到了很好的效果[5]。采用机器学习的方法来进行评价工作,能够将评价问题转换为分类问题,可以避免直接处理一些难以量化的因素,同时基于大量训练样本的学习,能够得到很理想的效果。鉴于机器学习方法在评价工作中表现出的显著效果,本文采用有监督机器学习的方法建立域名信用评价原型系统。操作流程如图 1 所示。

本文将域名信用值作为分类的类别来建立有监督学习的评价原型系统。将域名信用值限定为可数的整数值,如 0,1,…,MAX,MAX 可根据实际情况调整,每个信用值都是一个类别,那么总共有 MAX + 1 个类别 。

整个操作流程分为三个阶段:

a)人工标注,建立带类标记的域名样本集。选取一定数目的域名作为样本集,由人工按照一定的评价准则对样本集中的域名进行信用值打分,每个域名得到(域名,信用值)形式的标注结果,每个样本域名的信用值便是其类标记。

b)训练学习。对带类标记的域名样本集进行特征提取形

收稿日期: 2011-07-11; **修回日期**: 2011-08-20 **基金项目**: 国家自然科学基金资助项目(61005029)

作者简介:陈威(1987-),男,四川人,硕士研究生,主要研究方向为下一代互联网、可信网络(chenwei@cnnic.cn);王利明(1978-),男,博士,主要研究方向为下一代互联网、可信网络;耿光刚(1980-),男,博士,主要研究方向为模式识别、互联网信息检索;毛伟(1968-),男,研究员,博导,博士,主要研究方向为下一代互联网、网络寻址与定位;李晓东(1976-),男,硕导,博士,主要研究方向为互联网基础资源、网络信息安全.

成训练集,使用此训练集对分类器进行训练,并将训练后的分类器作为评价模块。

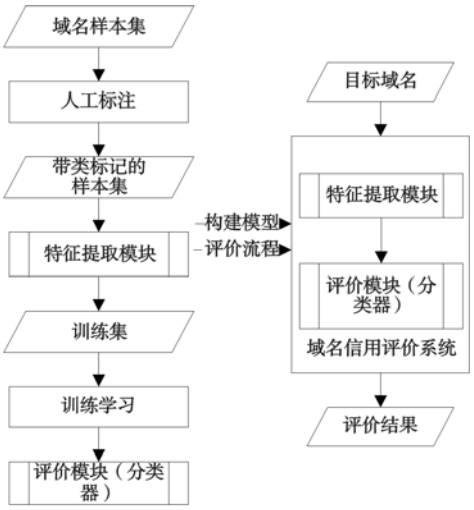

图1　域名评价系统建立流程

c)建立原型。将 b)得到的特征提取模块和评价模块串联结合,形成域名信用评价原型系统。

评价模块实质上是经过训练学习后的分类器,所以评价过程实际上是分类过程:新来的域名经过特征提取后,由评价模块对其分类,域名被分入某个类别后,其类标记便是其信用评价值。

2.1　建立带类标记的样本集

采集一定数目的域名作为样本集,保证样本集的全面性和均衡性。然后,按照一定的准则对该样本集中的域名进行信用评价标注,形成带类标记的样本集。这里的准则具体到本文的研究工作即是域名信用评价准则,2.1.1 小节将详细介绍该准则。

2.1.1　域名信用评价准则

本小节首先分析域名信用的评价指标,进而建立域名信用评价准则。

针对域名自身特点,本文从域名的注册信息和应用服务两个方面来分析域名信用的评价指标。其中,域名的应用当前多为 Web 站点应用,域名应用服务方面的信用可以通过研究其对应的 Web 站点的内容质量来得到。Web 站点的内容质量评价也是近年来研究的热点之一[3,4,6],ECML/PKDD 会议就针对 Web 内容质量在 2010 年举行了挑战赛 ECML/PKDD 2010 Discovery Challenge。

本文最终确立了 7 个具体评价指标。域名注册信息方面,为注册信息可靠度 1 项指标;域名应用服务方面,针对其站点内容,为题材分类、可靠性、事实度、是否存在偏见、是否涉及不良应用、是否是行业内顶尖著名站点 6 项指标,其中题材分类、可靠性、事实度、是否存在偏见 4 项指标借鉴于文献[7]并结合本文的研究作了适当的修改。

具体的指标处理如下:

a)注册信息可靠度。综合域名各方面的注册信息形成注册信息可靠度指标,将该指标分成四个等级 L4、L3、L2、L1,等级越高,可靠度越高。如注册时间越久远、离到期时间越长、注册主体的权威性更高、ns 个数越多,那么注册信息可靠度指标可标注为更高的等级。

b)题材分类。将站点分为新闻或教育类、论坛讨论类、其他类三个类别,分别对应类别 C3、C2、C1。

c)可靠性。该指标界定站点内容来源的权威可靠性,分为三个等级 L3、L2、L1,分别为高可靠性、一般可靠性、不可靠。

d)事实度。该指标界定站点内容事实观点性,也分为三个等级 L3、L2、L1,分别为事实、事实混合观点、观点。

e)是否存在偏见。布尔指标,标注为 Y 或 N,Y 为是,N 为否。

f)是否涉及不良应用。布尔指标,标注为 Y 或 N,其中不良应用涉及到钓鱼、色情、垃圾、作弊、病毒等应用。

g)是否是行业内顶尖著名站点。布尔指标,标注为 Y 或 N,如果该域名对应的站点在其所在行业内属于顶尖著名站点,可标注为 Y,如 amazon.cn、google.com 等可以认为是顶尖著名站点。

评价指标及其处理标记总结如表 1 所示。

表 1　域名信用评价指标及其处理标记

类型	域名信用评价指标	指标标记
域名注册信息	注册信息可靠度	L4 或 L3 或 L2 或 L1
	题材分类	C3 或 C2 或 C1
	可靠性	L3 或 L2 或 L1
	事实度	L3 或 L2 或 L1
域名应用	是否存在偏见	Y 或 N
站点内容	是否涉及不良应用	Y 或 N
	是否是行业内顶尖著名站点	Y 或 N

在确定了具体的评价指标及其处理形式之后,建立如下评价准则。假定域名信用值用 value 来表示,取值范围为 0 到 MAX 间的整数值,对域名按照以下五条准则来进行信用评价:

a)如果“是否是行业内顶尖著名站点”指标标记为 Y,value 赋值为 MAX,评价结束;如果为 N,继续。

b)如果“是否涉及不良应用”指标标记为 Y,value 赋值为 0,评价结束;如果为 N,继续。

c)根据“题材分类”指标为 value 赋值基础分,C3(新闻教育类)、C2(论坛讨论类)、C1(其他类)分别对应 V_3、V_2、V_1 的基础分赋值。

d)对于“注册信息可靠度”“可靠性”和“事实度”三项指标,根据其等级对 value 进行加分,L1、L2、L3、L4 分别对应 K_1、K_2、K_3 和 K_4 的加分。

e)如果“是否存在偏见”指标标记为 Y,对 value 进行 M 分的减分。

实验中,参数具体取值如下:V_3 为 5,V_2 为 4,V_1 为 3,K_1 为 0,K_2 为 1,K_3 为 2,K_4 为 3,M 为 2,从而得到 MAX 为 12,域名信用值的范围为 0~12 的整数值。实验中,每个域名都由多人进行标注评价,然后将结果取平均值,从而减小个人喜好、偏见等主观因素带来的影响。表 2 为实验过程中部分标注实例,credit 为域名信用值。

表 2　实验过程中部分标注实例

domain	credit
ziwojieshao.org.cn	8
3gp.cn	7
wuhunews.cn	9
ukgrimsby.cn	6
95z.cn	3
yanshan.gov.cn	11

2.2 特征提取

选取合理的特征来建立训练样本对训练学习的效果起着重要的作用。本文从以下三个方面进行特征选取分析：

a)域名字符串本身的特征。应用于同一类型网站的域名字符串往往具有相似的特征，如政府网站的域名中一般会有“gov”子串；很多不良应用站点会使用冗长且无字面意义的字符串作为域名，又或者通过模仿著名站点的域名来形成自己的域名字符串。

b)域名注册信息特征。域名注册信息中的各个项可以直接反应域名的特征。

c)域名站点的各种统计特征。如 Google 公司的 PageRank 值[8]、Alexa 的网站排名等，这些第三方机构的统计特征反映了站点页面质量及站点访问情况等的一些特征。

按照上面的思路，实验中选择如表 3 所示的各项特征进行训练学习，特征总的维度为 18。

表 3 系统选取的学习特征

特征类型	具体特征
域名字符串特征	域名的长度、域名的二级域、包含数字的个数、包含“-”的个数、包含字母的个数、是否包含年份等
注册信息特征	注册时间、过期时间、域名状态、域名 ns 个数、域名注册组织(个人或单位)、域名注册商等
域名站点第三方统计特征	Pagerank 值、Google 索引数、Yahoo 索引数、Alexa 链接个数、Alexa 排名、DMOZ 索引数等

2.3 分类算法策略

实验中，采用 C4.5 决策树算法[9]结合 Bagging 算法[10]来构建分类器进行训练学习，这是当前广泛使用且效果突出的一种策略。

C4.5 作为一种改进的决策树算法，具有分类精度高、构成模式简单、对噪声数据有很好的健壮性等优点。同时，本文实验中所选取的学习特征针对性强、维度不高，也比较适合采用 C4.5 算法。

Bagging 算法的主要思想是给定一弱学习算法和一原始训练集，经过多轮训练将该弱学习算法提升为强学习算法[11,12]。让该弱学习算法训练 t 个轮次，每轮的训练子集是从原始训练集随机抽取 n 个训练例构成的。t 轮训练之后，可以得到预测函数序列 $h_1,\cdots,h_t$，最终的预测函数 H 对分类问题由预测函数序列采用投票的方式完成或者将预测函数序列的分类结果加权求和。

实验中，经过 t 轮训练，得到分类器序列(预测函数序列) $h_1,\cdots,h_t$，最终的信用评价结果 credit(d) 为各个分类器的分类结果加权求和(参见式(1))，将式(1)作为系统的评价模块。

$$\text{credit}(d)=\sum_{i=0}^{t}p_i\times h_i(d) \tag{1}$$

其中：d 为被评价的域名；credit(d) 为域名的信用评价结果；p_i 为第 i 个分类器的权值，$\sum_{i=0}^{t}p_i=1$；$h_i(d)$ 为第 i 个分类器对域名 d 作出的分类评价结果。

经过式(1)的计算，被评价域名的信用值从类标记形式的整数值演变为实数值。经过归一化处理，最终可以将所有的信用评价值归一化到特定的范围内，如[0,10]。

2.4 举例介绍

假设原始样本集里有 example1.com 和 example2.com 两个域名。

首先人工按照域名信用评价准则来对样本集中的域名进行信用值标注。其中，假设 example1.com 和 example2.com 的标注结果为(example1.com, 10)(example2.com, 8)。

将标注后的带类标记的样本集交由特征提取模块处理，提取域名的各项特征形成训练集，这些特征是 2.2 节表 3 中所描述的各项特征。

按照 Bagging 算法思想，假设能将训练集随机划分成 5 个子训练集，进行 5 轮训练后，得到分类器序列 h_1、h_2、h_3、h_4、h_5。

现有目标域名 target.com 需要进行信用评价。在进行同样的特征提取后，由分类器序列对其进行分类。假设 5 个分类器对 target.com 的分类结果范围都只在 example1.com 和 example2.com 所在的两类中，它们对 target.com 给出的分类结果分别是 10、8、8、10、8，并假设 5 个分类器的权重一样，那么最终 target.com 的信用评价值将会是 (10+8+8+10+8)/5=8.80。

2.5 实验结果

将所有评价结果经过归一化处理后，信用评价值范围为[0,10]的实数值，保留两位小数，部分信用评价实验结果如表 4 所示。从表 4 中可以看出 sohu.com、ifeng.com 等著名域名得到比较高的信用评价值；radio5.cn 提供了比较丰富的广播资源，得到了中等的信用评价值；2se.cn 为在线电影网站，但是其提供的资源很多都无法播放，所以其信用评价值不高。

表 4 部分实验结果示例

domain	credit
sohu.com	9.39
ifeng.com	9.04
tianya.cn	8.54
radio5.cn	5.06
2se.cn	3.90

2.6 实验系统性能评估分析

实验系统采用 NDCG(normalized discounted cumulative gain)[13]作为性能评估指标，并采用 5 交叉验证(5 cross-validation)[14]方式进行。NDCG 的值越接近 1.0，说明机器评价的结果越接近样本标注的结果，性能评估结果如表 5 所示。

表 5 NDCG 评估结果

特征	域名字串特征	注册信息特征	第三方网站统计特征	所有特征
NDCG	0.898 8	0.920 2	0.914 9	0.928 3

从表 5 可以看出，单方面地利用域名串特征、注册信息特征或第三方网站统计特征进行学习均能得到较好的效果；利用所有特征则取得了最好的效果，说明各视角特征存在互补性、辅助性。

3 结束语

本文采用有监督机器学习的方法将信用评价问题顺利地转换为模式分类问题，有效地对域名信用进行自动化地评价。下一步的研究工作还将可能包括优化信用评价准则、扩大样本集的规模并深入研究样本集的全面性、扩充特征空间、对比分类算法策略等。

(下转第 697 页)

分能力强的特征项集,再利用特征项集与分类模型对待过滤邮件进行分类。特征选择介于分类器与邮件集合之间,是影响邮件过滤性能的重要因素。其通过降低特征空间维数减少分类器在计算和存储上的开销,同时过滤部分噪声以提高分类的准确性。

本文研究并总结了文本分类中常用的特征选择算法,分析了其优缺点,并重点研究其应用于邮件过滤时的性能差异。本文在朴素贝叶斯、支持向量机两个分类器上分析各特征选择算法对分类器的适应性,得到各算法在两个分类器上的召回率、准确率维持在 80% 以上。其中,OR、CHI、IG、DF 的性能较好,WET、MI 性能较差。并在 SVM 分类器上分析了各算法对数据集分布的依赖性,各算法在邮件样本集分布不平衡时性能均有所下降,OR、DF 下降幅度较小,受不平衡数据分布影响小;在各算法运行时间上,DF 计算最简单,CHI、IG 运行时间较长。综合考虑各因素,OR 在邮件过滤中性能较好。

参考文献:

[1] Symantec Intelligence Report[EB/OL]. (2011-07)[2011-07-22]. http://www. symantec. com/content/en/us/enterprise/other _ resources/b-intelligence_report_07-2011. en-us. pdf.

[2] 中国互联网协会反垃圾邮件中心. 2011 年第一季度中国反垃圾邮件状况调查报告[EB/OL]. (2011-05-24)[2011-07-22]. http://www. anti-spam. cn/ShowArticle. php? id = 11129.

[3] 靖红芳,王斌,杨雅辉,等. 基于类别分布的特征选择框架[J]. 计算机研究与发展, 2009, 46(9):1586-1593.

[4] SEBASTIANI F. Machine learning in auomated text categorization [J]. ACM Computing Surveys,2002,34(1):1-47.

[5] 陆玉昌,鲁明羽,李凡,等. 向量空间法中单词权重函数的分析与构造[J]. 计算机研究与发展, 2002,39(10):1205-1210.

[6] 刘海峰,张学仁,姚泽清,等. 基于类别选择的改进 KNN 文本分类[J]. 计算机科学, 2009,36(11):213-216.

[7] YANG Yi-ming, PEDERSEN J O. A comparative study on feature selection in text categorization[C]//Proc of the 14th International Conference on Machine Learning (ICML-97). San Francisco: Morgan Kaufmann,1997:412-420.

[8] 单松巍,冯是聪,李晓明. 几种典型特征选取方法在中文网页分类上的效果比较[J]. 计算机工程与应用, 2003,29(22):146-148.

[9] 单丽莉,刘秉权,孙承杰. 文本分类中特征选择方法的比较与改进[J]. 哈尔滨工业大学学报, 2011, 43(1):319-324.

[10] MLADENIC D, GROBELNIK M. Feature selection for unbalanced class distribution and Naïve Bayes[C]//Proc of ICML'99. San Francisco: Morgan Kaufmann,1999:258-267.

[11] 周茜,赵明生,扈旻. 中文文本分类中的特征选择研究[J]. 中文信息学报, 2004,18(3):17-23.

[12] 卢杨竹,张新有,祁玉. 邮件过滤中特征选择算法的研究及改进[J]. 计算机应用, 2009,29(10):2812-2815.

[13] 闫鹏,郑雪峰,李明祥,等. 二值分类中基于 Bayes 推理的特征选择方法[J]. 计算机科学,2008,35(7):173-176.

[14] 计算所汉语词法分析系统 ICTCLAS[EB/OL]. (2002-08-16)[2011-07-01]. http://www. nlp. org. cn/project/project. php? proj_id = 6.

[15] 杨创新. 基于机器学习的高性能中文文本分类研究[D]. 广州:华南理工大学,2009.

[16] 王清翔,广凯,潘金贵. 基于支持向量机的邮件过滤[J]. 计算机科学,2007,34(9):93-94.

[17] 中国互联网协会反垃圾邮件中心. 2010 年第四季度中国反垃圾邮件状况调查报告[EB/OL]. (2011-03-25)[2011-07-22]. http://www. anti - spam. cn/ShowArticle. php? id = 11010.

[18] YANG Y, LIU X. A re-examination of text categorization methods [C]//Proc of the 22nd ACM International Conference on Research and Development in Information Retrieval. Berkeley: ACM Press, 1999:42-49.

(上接第 692 页)

参考文献:

[1] MBAlib. Credit rating[EB/OL]. (2011-02-28)[2011-03-20]. http://wiki. mbalib. com/wiki/Credit_Rating.

[2] 佟雨兵,张其善,祁云山. 基于 PSNR 与 SSIM 联合的图像质量评价模型[J]. 中国图象图形学报,2006,11(12):1758-1763.

[3] GENG Guang-gang, JIN Xiao-bo, ZHANG De-xian. Evaluating Web content quality via multi-scale features[C]//Proc of ECML/PKDD 2010 Discovery Challenge. 2010.

[4] ERDELYI M, GARZO A, BENCZUR A A. Web spam classification: a few features worth more[C]//Proc of the 2011 Joint WICOW/AIRWeb Workshop on Web Quality. [S. l.]:ACM Press, 2011.

[5] DAVIS R H, EDELMAN D B, GAMMERMAN A J. Machine-learning algorithm for credit-card applications[J]. IMA J Management Math,1992,4(1):43-51.

[6] VIEDMA H E, PASI G, PORCEL C, *et al*. Evaluating the information quality of Web sites: a methodology based on fuzzy computing with words[J]. Journal of the American Society for Information Science and Technology,2006,57(4):538-549.

[7] ECML/PKDD. Rules: ECML/PKDD 2010 discovery challenge[EB/OL]. (2010-06-25)[2011-03-28]. http://datamining. sztaki. hu/? q = en/DiscoveryChallenge/rules.

[8] PAGE L, BRIN S, MOTWANI R, *et al*. The PageRank citation ranking: bring order to the Web, Technical Report SIDL-WP-1999-0120[R]. [S. l.]:Stanford InfoLab,Stanford University,1999.

[9] QUINLAN J R. C4.5: programs for machine learning[M]. San Mateo: Morgan Kaufmann Publishers, 1993.

[10] BREIMAN L. Bagging predictors[J]. Machine Learning,1996,24(2):123-140.

[11] KEARNS M, VALIANT L G. Learning boolean formulae or factoring, Technical Report TR-1488[R]. Cambridge, MA:Aiken Computation Laboratory,Harvard University,1988.

[12] KEARNS M, VALIANT L G. Cryptographic limitation on learning boolean formulae and finite automata[J]. Journal of ACM,1994,41(1):433-444.

[13] JAERVELN K, KEKAELAEINEN J. Cumulated gain-based evaluation of IR techniques[J]. ACM Trans on Information Systems, 2002,20(4):422-446.

[14] KOHAVI R. A study of cross-validation and bootstrap for accuracy estimation and model selection[C]//Proc of International Joint Conference on Artificial Intelligence. Montréal, Québec: Morgan Kaufmann Publishers,1995.