

基于双端惩罚传播算法的搜索引擎排序优化技术

赵迅冕, 牛小飞, 王泽贤, 王耀辉, 宋泓远

(山东建筑大学 计算机科学与技术学院, 山东 济南 250101)

摘要: 一些网页通过欺诈手段欺骗搜索引擎排序算法,使其获得不应有的高排序值,恶化了搜索引擎返回的结果,影响了用户体验。因此,在搜索结果列表中提升信任网页、降级作弊网页可以优化搜索引擎排序结果。然而,已有的传播算法有的只传播信任或非信任值,同时传播信任和非信任值的算法只在传播的源端或目标端进行惩罚,提升信任网页、降级作弊网页的效果有限。为解决此问题,提出了双端惩罚算法,在沿着出链正向传播信任值、沿着入链反向传播非信任值的同时,在传播的源端和目标端对信任与非信任值进行惩罚操作,可以更好地使信任网页排名靠前,作弊网页排名靠后。利用公开的WEBSPAM-UK2006数据集验证了双端惩罚算法的效果,实验结果表明,在排名前4 000的网页中,双端惩罚算法信任网页的比例相比只传播信任值的TrustRank算法高12.9%,相比在源端进行惩罚的GBR算法高3.8%,相比在目标端进行惩罚的TDR算法高2.8%,同时还降级了作弊网页。

关键词: 网页作弊检测;信任传播;非信任传播;双端惩罚算法;传播算法

DOI: 10.11907/rjdk.241818

中图分类号: TP391

文献标识码: A

文章编号: 1672-7800(2025)012-0168-08

扫描二维码阅读全文:



Search Engine Ranking Optimization Technology Based on the Double-Ended Penalty Propagation Algorithm

ZHAO Xunmian, NIU Xiaofei, WANG Zexian, WANG Yaohui, SONG Hongyuan

(College of Computer Science and Technology, Shandong Jianzhu University, Ji'nan 250101, China)

Abstract: Some Web pages cheat the search engine sorting algorithm by fraudulent means, so that they can obtain undeserved high ranking values, which deteriorates the results returned by the search engine and affects the user experience. It can optimize the search engine sorting results to enhance trust pages and degrade spam pages in the search results list. Some of the existing propagation algorithms only propagate trust or distrust values, and the algorithms that propagate both trust and distrust values only punish the source or target end of the propagation, so the performance of enhancing trust web pages and degrading spam Web pages is limited. The proposed double-end penalty algorithm can propagate trust values forward along the outgoing link and distrust values backward along the incoming link, punish trust and distrust values at both the source and target end of the propagation. The new algorithm can better rank the trust Web pages in front and the spam Web pages at the back. The public WEBSPAM-UK2006 data set was used to verify the performance of the double-end penalty algorithm. The experimental results show that among the top 4 000 Web pages, the proportion of trust Web pages of the double-end penalty algorithm is 12.9% higher than that of TrustRank algorithm which only propagates trust values, 3.8% higher than that of GBR algorithm which punishes at the source end, and 2.8% higher than that of TDR algorithm that punishes at the target end, while the new algorithm degrades spam pages.

Key Words: web spam detection; trust propagation; distrust propagation; double-ended penalty algorithm; propagation algorithm

收稿日期: 2024-09-24

基金项目: 山东省自然科学基金项目(ZR2021MF006)

作者简介: 赵迅冕(2000-),男,山东建筑大学计算机科学与技术学院硕士研究生,研究方向为数据挖掘、网页作弊检测;牛小飞(1979-),女,博士,CCF会员,山东建筑大学计算机科学与技术学院教授、硕士生导师,研究方向为数据挖掘、机器学习、网页作弊检测、心理健康辅助诊断;王泽贤(2000-),男,山东建筑大学计算机科学与技术学院硕士研究生,研究方向为数据挖掘、心理健康辅助诊断;王耀辉(2000-),男,山东建筑大学计算机科学与技术学院硕士研究生,研究方向为数据挖掘、心理健康辅助诊断;宋泓远(2001-),男,山东建筑大学计算机科学与技术学院硕士研究生,研究方向为数据挖掘、心理健康辅助诊断。本文通讯作者:牛小飞。

0 引言

日常生活中,搜索引擎可以向用户提供需要查询的相关网页信息。对于一个确定的查询,通常情况下搜索引擎能返回成千上万个结果,但是只有排在前面的一个或几个结果会对用户有所帮助。因此,一些人使用各种手段来欺骗搜索引擎排序算法,使一些网页获得比本身更高的排名。所有这些试图提高网页排名的欺骗行为通常被称为网页作弊(Web Spam)^[1]。网页作弊恶化了搜索引擎排序的结果,降低了用户体验,带有病毒或恶意软件的作弊网页可能会导致用户的隐私泄露、经济损失等。因此,作弊网页检测技术应运而生^[2]。

目前的网页作弊检测技术主要分为分类和传播技术两种。利用分类技术检测作弊网页需要大量已标注样本进行模型训练,但是人工标记网页费时费力,而较少的标注样本会影响分类检测算法的检测效果。利用网页之间链接关系进行传播的算法分为信任值传播和非信任值传播两类。其中,信任值传播算法为所有网页计算一个信任分数,并按照该分数进行排名,使更加可信的网页排名靠前,以此提升信任网页,降级作弊网页,如PageRank算法^[3]、TrustRank算法^[4]。但在实际情况下,一些作弊网页会通过一些作弊手段在可信网页中加入指向自己的链接,从而提高自身的可信分数排名,导致作弊网页排名靠前。基于非信任值传播的算法为所有网页计算一个非信任分数,依此识别出非信任值较高的作弊网页,如Anti-TrustRank算法^[5]。

因为每个网页有值得信任的一面,也有不值得信任的一面。信任值传播算法在传播过程中不考虑网页的非信任值,只能得到网页的信任值;非信任值传播算法在传播过程中不考虑网页的信任值,只能得到网页的非信任值。同时传播信任值与非信任值的算法在从可信网页种子集沿着网页出链传播信任值的同时,从作弊网页种子集沿着入链反向传播非信任值,以此遍历整个网络链接图,得到每个网页的信任值和非信任值。比较经典的同时传播信任和非信任值的相关算法有Trust-Distrust Rank(TDR)算法^[6]和Good-Bad Rank(GBR)算法^[7]。其中,TDR算法是在信任值/非信任值传播的目标端网页根据其非信任值/信任值进行一个值的衰减操作,即惩罚操作;GBR算法是在信任值/非信任值传播的源端网页根据其非信任值/信任值进行惩罚。本文提出同时传播信任和非信任值的双端惩罚排序算法Double-ended Penalty Rank(DP-Rank),该算法在信任值/非信任值传播的同时,在源端网页和目标端网页同时进行惩罚操作,以此来遍历整个网络链接图得到每个网页的信任值和非信任值,从而提升可信网页、降级作弊网页。

1 相关工作

现有作弊网页检测技术大致可分为两类。基于分类算法检测网页作弊的技术主要通过提取网页的内容、链接等特征,构建分类模型来区分正常网页和作弊网页。例如Chatterjee等^[8]提出一种基于模糊证据的方法来识别网页作弊;王晶晶^[9]提出一种基于网页链接关系和主题相关度的改进算法;陈木生等^[10]提出基于单页内容提取语义特征融合单页统计特征进行分类的方法,以提高网页作弊检测率;苏瑜^[11]提出一种基于源码多维特征的作弊网页检测模型,提取了多个网页的源码特征,并采用深度森林算法构建模型检测作弊网页;Wang等^[12]引入一种智能特征提取方法page2vec和一种新的分类算法RFirF来检测CloT中的作弊网页,通过网页链接结构图计算每个页面的好/坏分数,并利用随机森林进行网页分类,实验结果表明不同评分函数可以产生不同的网页特征。Chakraborty等^[13]构建一个动态作弊信息过滤器,通过使用最常见的朴素贝叶斯算法以更有效地过滤标准消息和作弊信息。Akinyelu等^[14]引入自然启发算法,以进一步提升机器学习算法对作弊网页检测的速度和效果;Aswani等^[15]从评论特征、评论者特征和作弊信息发送者群体特征3方面对信息进行了分类;Waheed等^[16]等利用混合情感分析方法中的Region-based Convolutional Neural Networks(RCNN)算法,可以准确地筛选出假新闻网页;Teja等^[17]针对用于作弊网页检测的机器学习方法,从正确率、精确度、召回率等方面对不同方法进行了比较;Jazzar等^[18]评估了最流行和最有效的用于作弊网页分类与过滤的机器学习算法;Al-Zoubi等^[19]通过考虑不同语言上下文产生的各种重要特征来改进检测方法;Kaddoura等^[20]介绍了社交媒体中作弊信息检测和分类的最新进展,以及现有工作中使用的控制机制和数据集;Batra等^[21]提出一种基于 k 近邻的分类方法,结合几种仿生优化技术对邮件、网页进行分类;Shrivastava等^[22]提出一个集成模型-1,通过集成多层感知器(Multilayer Perceptron, MLP)、Naïve 贝叶斯和随机森林(Random Forest, RF),以获得更高的网页分类精度。Ahsan等^[23]通过主题模型对信息内容进行建模,并将其与面向用户的特性相结合,在预测作弊网页扩散方面具有较高的准确性。

传播算法通过在网页链接结构图上的传播操作为网页计算信任/非信任值,从而提升可信网站的排名或检测作弊网页。例如目前通过可信度对网页进行排序的方法主要是以PageRank为代表的算法,通过手动选择种子集并将信任值沿着种子的出链接进行信任传播的算法主要是TrustRank算法。TDR中的Trust-Rank、GBR中的Good-Rank等算法是从一组挑选好的可信种子集开始,通过同时传播信任值与非信任值的方式计算网站的信任值,以提升可信网站的排名。检测作弊网页的算法如Anti-TrustRank

算法是从一组作弊种子结点开始,在网络中传播非信任分数以识别作弊网页,TDR 中的 Distrust-Rank、GBR 中的 Bad-Rank 等算法从一组挑选好的作弊种子集开始,通过同时传播信任值与非信任值的方式计算网站的非信任值来识别作弊网页。Bevendorff 等^[24]提出以内容和链接农场形式出现的正常网页与作弊网页之间的界限变得越来越模糊,以低质量、大量生产的商业内容形式出现的动态对抗性作弊网页值得更多关注。Zhuang 等^[25]基于网页的链接结构提出基于深度信任网络的网页作弊分类方法。Niu^[26]等提出基于三值主观逻辑模型提升信任网页,降级作弊网页。

对于现实情况下的 Web 网络图来说,当信任从一个可信网站传播到不可信网站的过程是没有意义的。同时,一些作弊网站会恶意制造链接,通过被其它可信网页链接以传播信任值,而作弊网站的信任分数是不值得信任的。因此,从作弊网站传来的分数对于可信网站来说也是没有意义的。本文提出双端惩罚算法,旨在利用信任与非信任同时传播,并且在源端网页和目标端网页同时进行惩罚,尽可能地提高可信网页的排名,同时对作弊网页进行降级。

2 预备知识

网页的超链接图可建模为有向图 $G = (V, E, W)$, 其中顶点 $i \in V$ 表示网页,边 $e(i, j) \in E$ 表示从网页 i 到 j 的超链接,并称网页 j 为 i 的出邻居,网页 i 为 j 的入邻居。权重 $\omega(i, j) \in W$ 表示边 $e(i, j)$ 的权值,通常被建模为一个实数。网页入度为指向该网页的网页数量,网页出度为其所指向的网页数量。网页的信任值沿出链正向传播,非信任值沿入链反向传播。

2.1 PageRank 算法

PageRank 算法是用于评估网页重要性的算法。该算法基于图论的思想,通过分析网页之间的链接关系来确定网页排名。PageRank 算法将每个网页赋予一个初始的 PageRank 值,之后通过迭代计算,不断更新每个网页的 PageRank 值,直到达到收敛。

在每次迭代中,PageRank 算法根据链接关系计算每个网页的 PageRank 值。该值由两部分组成:第一部分是来自其它网页传递的 PageRank 值,即该网页的入邻居传递给该网页的值;第二部分是来自随机跳转的 PageRank 值,即用户以一定概率随机访问该网页的概率。PageRank 值计算公式如式(1)所示。

$$PR(p) = \alpha \cdot \frac{1}{N} + (1 - \alpha) \cdot \sum_{q \in IN(p)} \frac{PR(q)}{OUT(q)} \quad (1)$$

其中, α 为衰减因子,通常设置为 0.85, N 为所有网页的数量, OUT 为网页的出度。

通过多次迭代计算,PageRank 算法不断更新每个网页的 PageRank 值,直到收敛为止。最终,网页的 PageRank 值

越高,其在搜索结果中的排名越靠前。

2.2 TrustRank 算法

TrustRank 算法是一种用于评估网页可信度的算法,可以有效提升信任网页,降级作弊网页。TrustRank 算法从一个可信种子集合开始,通过链接分析的方法将信任值逐步扩展到其它网页。计算过程如式(2)所示。

$$r = \alpha \cdot T \cdot r + (1 - \alpha) \cdot d \quad (2)$$

式中, α 是衰减因子, T 是转移矩阵。种子网页在向量 d 中对应位置的值为 $1/m$,其中 m 为种子数量。当网页不是种子时,对应位置为 0。

TrustRank 算法通过可信种子集合沿着网页的出链进行信任传播,以得到所有网页的可信分数,并得到可信网页排名,以达到降级作弊网页的目的。

2.3 TDR 算法

TDR 算法为每个网页分配了一个 T-Rank 分数来表示页面的可信度,同时分配了一个 D-Rank 分数来表示页面的不可信度。因此,可信网页会被分配高 T-Rank 分数和低 D-Rank 分数,作弊网页会被分配低 T-Rank 分数和高 D-Rank 分数。

在 TDR 算法计算过程中,首先选定可信网页和作弊网页种子集,从可信网页种子集出发,沿着出链传播信任值,从作弊网页种子集出发,沿着入链反向传播非信任值。TDR 算法计算公式如式(3)、式(4)所示。

$$t(p) = (1 - \alpha) \cdot s(p) + \alpha \cdot \sum_{q \in IN(p)} \frac{t(q)}{OUT(q)} \cdot \frac{\beta t(p)}{\beta t(p) + (1 - \beta) d(p)} \quad (3)$$

其中,网页的链接方向为 $q \rightarrow p$, α 为衰减因子, β 为信任值所占权重, t 为信任值, d 为非信任值, $s(p)$ 为信任向量。

$$d(p) = (1 - \alpha') \cdot s'(p) + \alpha' \cdot \sum_{q \in OUT(p)} \frac{d(q)}{IN(q)} \cdot \frac{\beta' d(p)}{\beta' d(p) + (1 - \beta') t(p)} \quad (4)$$

其中,网页的链接方向为 $p \rightarrow q$, α' 为衰减因子, β' 为非信任值所占权重, t 为信任值, d 为非信任值, $s'(p)$ 为非信任向量。

通过上式可以看出,在 TDR 算法计算过程中,信任值的计算会被目标端网页的非信任值所影响,非信任值的计算也会被目标端网页当前的信任值所影响。该传播过程实现了在目标端进行惩罚操作,从而有效提升了可信网页的排名,并对作弊网页进行降级。

2.4 GBR 算法

GBR 算法首先会选择出可信网页和不可信网页的集合,并给每个网页一个 G-Rank 值作为可信分数,同时给每个网页一个 B-Rank 值作为不可信分数。在计算目标端网页的 G-Rank 值时,会根据源端网页的 B-Rank 值进行惩罚,同时在计算目标端网页的 B-Rank 值时,会根据源端网页的 G-Rank 值进行惩罚。GBR 算法计算过程如式(5)、式

(6)所示。

$$g(p) = \alpha \cdot \sum_{q \in IN(p)} \frac{g(q)}{OUT(q)} \cdot \frac{g(q)}{g(q) + b(q)} + (1 - \alpha) \cdot s(p) \quad (5)$$

其中,网页的链接方向为 $q \rightarrow p$, α 为衰减因子, g 为信任值, b 为非信任值, $s(p)$ 为信任向量。

$$b(p) = \alpha' \cdot \sum_{q \in OUT(p)} \frac{b(q)}{IN(q)} \cdot \frac{b(q)}{g(q) + b(q)} + (1 - \alpha') \cdot s'(p) \quad (6)$$

其中,网页的链接方向为 $p \rightarrow q$, α' 为衰减因子, g 为信任值, b 为非信任值, $s'(p)$ 为非信任向量。

通过上式可以得出结论,GBR算法在源端进行信任/非信任的惩罚,可以一定程度上降低当源端网页是作弊网页时对信任传播的影响。TDR和GBR算法同时实现了信任值与非信任值在Web链接图中的传播,并且分别在目标端和源端进行了惩罚操作,从而为本文提出的双端惩罚算法提供了研究思路。

3 双端惩罚算法

TDR和GBR算法可分别在目标端与源端进行信任/非信任的惩罚操作,但有两种情况这两类算法无法同时解决。第一种情况是当传递信任的网页本身是作弊网页,其传播出的信任值是没有意义的,GBR算法可以有效解决此种情况。第二种情况是当信任值通过可信网页的出链接传播到一个不可信网页时,该信任的传播过程是没有意义的,而TDR算法可在一定程度上避免该情况的发生。因此,为了能更精确地得到可信网页的排名并降级作弊网页,本文综合考虑TDR和GBR算法的优缺点提出了双端惩罚算法。利用网络的有向链接图,在沿着网页之间的出链传播信任值时,将非信任值按入链的方向进行反方向传播,并在计算网页信任值/非信任值时,同时考虑源端网页和目标端网页的信任值与非信任值。然后利用传播和组合的操作,在整个网页链接图上进行迭代计算,最终对所有网页链接图中的网页进行有效排名。

3.1 双端惩罚:在源端和目标端同时惩罚

双端惩罚算法会选择出值得信任的种子集和作弊网页种子集,设值得信任的种子集共有 m 个网页,对其中每个网页分配 $1/m$ 作为 $t1$ 和 $t2$ 的可信分数初始值,设作弊网页种子集共有 n 个网页,对其中每个网页分配 $1/n$ 作为 $d1$ 和 $d2$ 的不可信分数初始值。在每一次迭代计算过程中,首先利用给定的初始信任值 $t2$ 和非信任值 $d2$ 在源端进行惩罚计算得到 $t1$ 分数与 $d1$ 分数,计算过程如式(7)、式(8)所示。然后将得到的 $t1$ 分数作为信任值, $d1$ 分数作为非信任值,在目标端进行惩罚计算得到 $t2$ 和 $d2$ 分数,计算过程如式(9)、式(10)所示。经过计算后,信任值与非信任值得到更新。之后通过迭代计算至算法收敛,最终确定所有网页的信任值和非信任值。

$$t1(p) = \alpha \cdot \sum_{q \in IN(p)} \frac{t2(q)}{OUT(q)} \cdot \frac{t2(q)}{t2(q) + d2(q)} \quad (7)$$

$$d1(p) = \alpha' \cdot \sum_{q \in OUT(p)} \frac{d2(q)}{IN(q)} \cdot \frac{d2(q)}{t2(q) + d2(q)} \quad (8)$$

$$t2(p) = \alpha \cdot \sum_{q \in IN(p)} \frac{t1(q)}{OUT(q)} \cdot \frac{\beta t1(p)}{\beta t2(p) + (1 - \beta)d2(p)} + (1 - \alpha) \cdot s(p) \quad (9)$$

$$d2(p) = \alpha' \cdot \sum_{q \in OUT(p)} \frac{d1(q)}{IN(q)} \cdot \frac{\beta' d1(p)}{(1 - \beta')t1(p) + \beta' d1(p)} + (1 - \alpha') \cdot s'(p) \quad (10)$$

其中, $t1$ 、 $t2$ 为可信分数, $d1$ 、 $d2$ 为不可信分数, α 、 α' 为衰减因子, β 、 β' 为信任值和非信任值所占权重, $s(p)$ 为信任向量, $s'(p)$ 为非信任向量。

双端惩罚算法结合TDR算法和GBR算法的思想,在给定种子集网页初始值后,首先利用GBR算法的思想在源端进行惩罚计算,得到每个网页第一次的信任值 $t1$ 和非信任值 $d1$ 。之后将TDR算法公式中出现的信任值和非信任值替换为刚计算得到的信任值 $t1$ 与非信任值 $d1$,以此实现在网页目标端的惩罚计算,得到每个网站确定的信任值 $t2$ 和非信任值 $d2$,并以此迭代计算至算法收敛。DP-Rank算法同时进行信任值与非信任值的传播和组合,并在源端和目标端同时进行惩罚计算,可得到更准确的信任值与非信任值,从而有效提升可信网页的排名,并降级作弊网页。

3.2 双端惩罚算法实例

对于传统的信任传播算法如TrustRank算法,以及非信任传播算法如Anti-TrustRank算法,都是单向地通过网页之间的链接进行传播。这两种方式会分别用到种子集中的可信网页或作弊网页,而种子集通常都是由专家手动进行标注的,若只使用一种标注样本则会丢失另一种标注样本的信息。因此,本文考虑对网页赋予信任值和非信任值。

对于如图1所示的Web链接图,可信网页显示为白色,作弊网页显示为灰色。将 $\{1, 2\}$ 作为好种子,将 $\{7, 8\}$ 作为作弊种子。本文使用TrustRank算法,设 $\alpha=0.85$,迭代计算至收敛,可以得到网页的分数向量分别为 $t = [0.075; 0.25; 0.185; 0.079; 0.107; 0.033; 0.157; 0.079; 0.033]$ 。可以看到结果并不理想,因为7、8网页都获得了很高的TrustRank分数。因此,本文选择在信任值按照网络链接图节点沿着出链进行传播的同时,将非信任值沿着入链进行反向传播,并且将非信任值考虑到信任值的计算中。例如之前已有的信任与非信任同时传播的GBR算法,按照如图1所示的例子,使用相同的参数包括迭代次数、衰减因子等,使用GBR算法将其运行至收敛,得到信任值的分数向量为 $g = [0.132; 0.341; 0.202; 0.078; 0.144; 0.001; 0.044; 0.057; 0.001]$ 。从结果中可以看出,所有好页面的分数都高于坏页面,也证明了同时传播信任值与非信任值的方法是可行的。这些算法虽然实现了信任值与非信任

值的同时传播,但也随之产生了一些问题。比如当信任从一个可信网站传播到不可信网站时,该信任的传播过程是没有意义的。为了解决上述问题,本文使用双端惩罚算法进行网页筛选,图1说明了双端惩罚算法的计算过程。

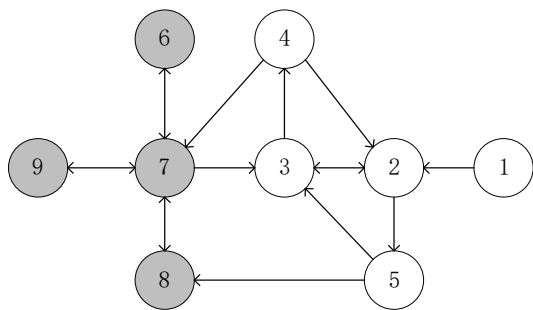


Fig. 1 Web link graph example

图1 网络链接图示例

假设节点1、2被标注为可信网页,节点7、8被标注为作弊网页,则信任值从1、2开始沿着网页的出链进行传播,而非信任值从7、8开始沿着入链进行反向传播,在计算信任值时要考虑到当前网页的非信任值和信任来源网页的信任值。采用本文提出的双端惩罚算法计算图1中各节点的信任值,得到信任值分数向量 $t=[0.108; 0.343; 0.165; 0.086; 0.192; 0.001; 0.055; 0.048; 0.001]$ 。通过与TrustRank算法、TDR算法和GBR算法的结果进行比较,可以发现,所有好页面的信任值都比坏页面更高,而且好页面的分数与坏页面分数的区分度也更大。这是因为双端惩罚算法实现了信任与非信任在Web图中的同时传播,并且双端惩罚算法在源端和目标端都进行惩罚的特性有效避免了源端网页为作弊网页时信任传播的无效性,减少了作弊网页在传播时对排名结果的干扰,同时减少了一些恶意操作,如作弊网页通过恶意手段制作从可信网页到作弊网页的链接等情况对算法结果的影响。

3.3 双端惩罚算法的收敛性

双端惩罚算法从可信网页种子集和作弊网页种子集出发,通过Web网页有向图进行信任与非信任的传播和组合操作,以此来迭代更新所有其它网页的信任值和非信任值,最终得到一个可信分数的排名。计算过程类似于TDR算法和GBR算法,只是在计算每一代的信任值/非信任值时,先计算得到 $t1/d1$ 的分数,再将 $t1$ 的分数作为信任值,将 $d-d1$ 的分数作为非信任值,对新得到的信任值/非信任值进行 $t2$ 分数和 $d-d2$ 分数计算,之后迭代计算至设定好的迭代次数。由于TDR算法和GBR算法都已被证明为收敛的,所以对于双端惩罚算法,随着信任值/非信任值在Web网络内的传播,第 $k+1$ 次的迭代结果与第 k 次的迭代结果之间的差异逐渐减小,即双端惩罚算法逐渐收敛。

4 实验与结果分析

为了验证双端惩罚算法的有效性,本文分析了迭代次

数和不同种子个数对双端惩罚算法的影响,并与PageRank、TrustRank、GBR和TDR算法进行了比较。

4.1 实验数据集

WEBSpAM-UK2006和WEBSpAM-UK2007是两个用于研究可信或作弊网站检测的公共数据集,其数据是由网络爬虫收集的。WEBSpAM-UK2006和WEBSpAM-UK2007数据集标注样本数如表1所示。

Table 1 Number of annotated samples in the dataset

表1 数据集标注样本数

数据集	网站总数	标注正常网站	标注作弊网站
WEBSpAM-UK2006	11 402	5 549	1 924
WEBSpAM-UK2007	114 529	5 709	344

其中,WEBSpAM-UK2006含有比WEBSpAM-UK2007更多的已标记网页,更适合评估算法效果。因此,本文实验采用WEBSpAM-UK2006数据集来评估所提出的双端惩罚算法。

4.2 参数设置

在执行算法前,首先使用PageRank算法与InverseRank算法挑选出200、400、800个正常网页和作弊网页,分别作为正常网页种子集和作弊网页种子集。在研究双端惩罚算法的收敛速度时,设置迭代次数 e 分别为2、3、5、10、15、20、25,衰减因子 α 为0.85。为了解种子数量对结果的影响,使用不同数量的信任和作弊网站执行双端惩罚算法。实验具体参数设置如表2所示。

Table 2 Parameter settings

表2 参数设置

参数	数值
作弊网站个数	200, 400, 800
信任网站个数	200, 400, 800
迭代次数 e	2, 3, 5, 10, 15, 20, 25

4.3 种子选择

本文使用经典的网页排名算法PageRank以迭代次数20和衰减因子0.85在数据集WEBSpAM-UK2006上进行实验,将排名靠前且标注为normal的网页作为可信网页的种子集挑选出来,同时使用相同的参数运行InversePageRank算法,并将排名靠前且标注为spam的样本作为作弊网页的种子集。表3显示了利用PageRank算法和InversePageRank算法所挑选网页中前200个网页的平均入度和出度。入度和出度越多的网页更有利于信任/非信任的传播,使用PageRank算法和InversePageRank算法可以有效地帮助挑选实验的种子集。

接下来探究不同种子数量对双端惩罚算法的影响,设置种子数量分别为200、400和800,结果如图2所示。

Table 3 The average outdegree and indegree of the selected top 200 seeds

表3 所选前200个种子的平均入度和出度

算法	平均出度	平均入度
High PageRank	242.460	571.765
Inverse PageRank	643.235	276.385

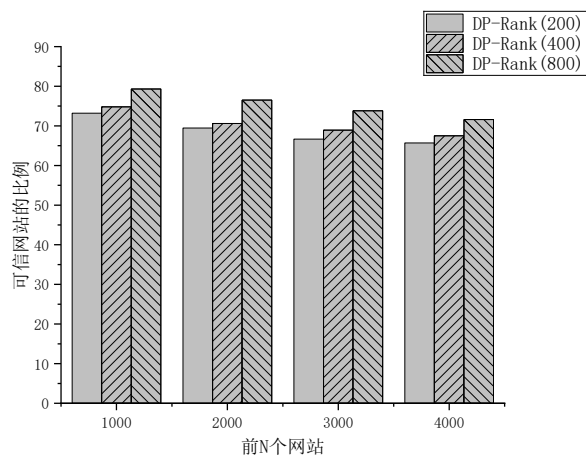


Fig. 2 The effect of different number of seeds on double-end penalty algorithm

图2 不同种子数量对双端惩罚算法的影响

通过对比不同种子数量对双端惩罚算法的影响,发现随着种子数量的增加,双端惩罚算法的效果呈上升趋势。在前1 000个网页中,可信网页的占比从73.2%提升到了79.3%。这也说明当选择的可信种子集和作弊种子集越多时,双端惩罚检测算法效果越好。

4.4 迭代次数对算法的影响

为了探究迭代次数对双端惩罚算法的影响,本文首先在种子集中选出200个可信网页和作弊网页作为种子集进行实验。实验尝试了不同的迭代次数,分别为2、3、5、10、15、20、25,通过计算每一次迭代后每个网页的信任值变化,将其差值平方求和后再开方。计算过程如式(11)所示。

$$c - score = \sqrt{(k_1^{t+1} - k_1^t)^2 + (k_2^{t+1} - k_2^t)^2 + \dots + (k_N^{t+1} - k_N^t)^2} \quad (11)$$

其中,c-score表示每次迭代后与上一次迭代时网页信任值的变化, k_i^t 表示第*i*个网页在第*t*次迭代后所得的信任值。

不同迭代次数下c-score值变化趋势如图3所示,发现双端惩罚算法在迭代计算次数达到20次后信任值变化得非常缓慢,即双端惩罚算法在迭代次数达到20次时基本收敛。因此,本文选择后续实验的迭代次数为20。

4.5 算法效果对比

选择经典的基于链接的网页作弊检测算法TrustRank算法、TDR算法、GBR算法和本文提出的双端惩罚算法,对其针对可信网页的排名效果进行比较。实验统一设置衰减因子为0.85,迭代次数为20,正常网页和作弊网页的种子集数量都为200。图4为可信网页在排名前1 000、2 000、3 000和4 000网页中的比例,图5为作弊网页在排名前4 000、6 000和8 000网页中的比例。网页排名中靠前的信任网页越多、作弊网页越少说明排序结果越好。

从图4可以看出,双端惩罚算法在可信网页排名方面

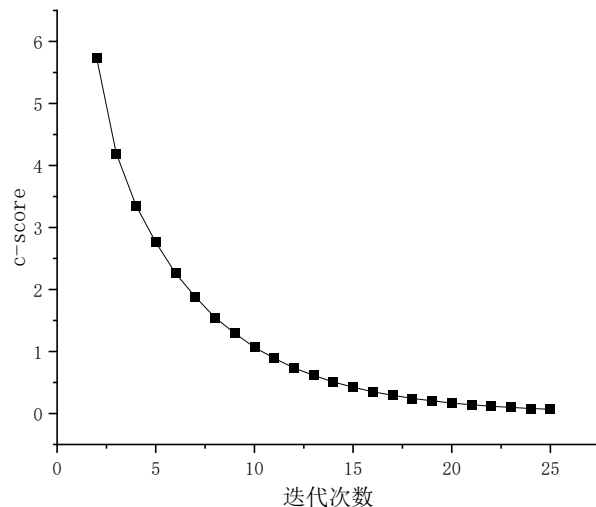


Fig. 3 The change trend of c-score under different iterations

图3 不同迭代次数下c-score值变化趋势

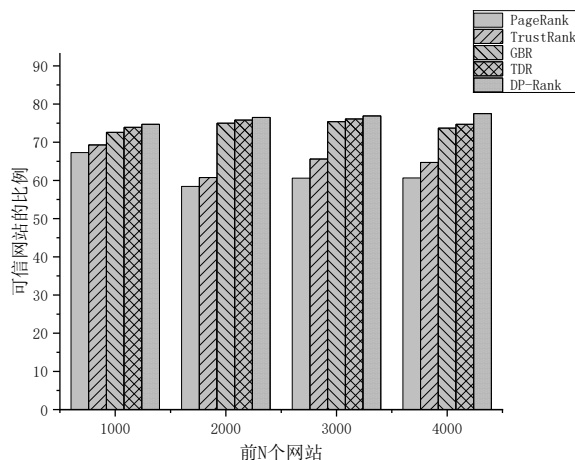


Fig. 4 The percentage of trust pages in the first *N* pages of different algorithm results

图4 不同算法结果在前*N*个网页中被标记为可信网页的百分比

相比其他算法表现更为优异,在前1 000、2 000、3 000、4 000个网页中与PageRank相比分别提高了7.4%、18.1%、16.3%、16.9%;与TrustRank相比分别提高了5.4%、15.7%、11.3%、12.9%;与TDR算法相比分别提高了0.8%、0.65%、0.8%、2.8%;在略微增加时间成本的基础上,与GBR算法相比分别提高了2.1%、1.0%、1.5%、3.8%。由实验结果可以得出结论,基于信任值和非信任值同时传播的算法相比只传播信任值的算法在提升可信网页排名方面更有优势。同时通过对比双端惩罚算法与TDR算法和GBR算法,在信任传播过程中,根据目标端和源端的非信任值同时对信任值进行惩罚的操作也能更好地提升可信网页的排名,证明了双端惩罚算法的有效性。

从图5可以看出,双端惩罚算法在降级作弊网页方面也更有优势,在前4 000、6 000、8 000个网页中,与TrustRank、GBR算法和TDR算法相比,作弊网页所占比例更

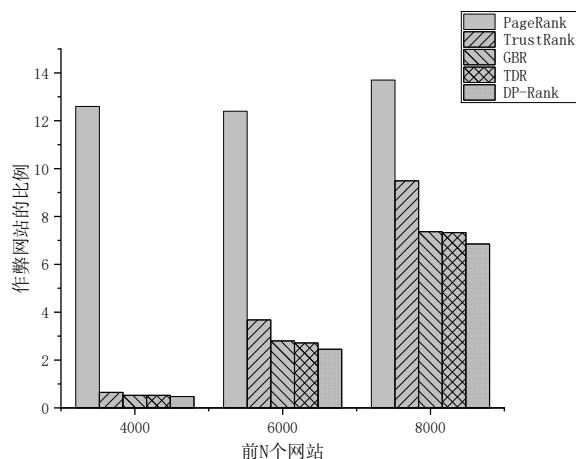


Fig. 5 The percentage of spam pages in the first N pages of different algorithm results

图5 不同算法结果在前 N 个网页中被标记为作弊网页的百分比

小,与PageRank相比分别降低了12.1%、9.9%、6.9%;与TrustRank算法相比分别降低了0.175%、1.23%、2.64%;与GBR算法相比分别降低了0.15%、0.35%、0.51%;与TDR相比分别降低了0.15%、0.26%、0.475%。说明双端惩罚算法在提升信任网页的同时也降级了作弊网页。

双端惩罚算法实现了信任与非信任的同时传播,同时在源端和目标端都对网页进行惩罚计算。不同传播算法在不同迭代次数下的运行时间如表4所示,可以看出将双端惩罚算法应用于公开的网页数据集时,其在进行20次迭代计算使算法达到收敛时,所用时间仅比TDR算法多6.2 s,比GBR算法多7.9 s。说明双端惩罚算法在提升可信网页排名时,并没有牺牲较多的时间成本,体现了双端惩罚算法的实用性。

Table 4 The running time of different propagation algorithms under different iterations

表4 不同传播算法在不同迭代次数下的运行时间 (s)

算法	迭代5次	迭代10次	迭代15次	迭代20次
DP-Rank	5.563	12.347	18.878	24.045
TDR	3.908	8.132	12.640	17.889
GBR	3.693	8.007	11.793	16.159

综上,双端惩罚算法在已有的信任值与非信任值同时传播的基础上进行了改进,在信任的传播和组合操作中,根据源端网页和目标端网页的非信任值同时做出了惩罚操作,实现了双端惩罚的效果。实验结果表明,双端惩罚算法提高了对于Web网络链接图中可信网页的识别率,通过给出更高分数让更多可信网页获得更高的排名,并降低作弊网页的排名。另外,在使用信任值对网页进行排名的同时,也可以使用双端惩罚算法中的非信任值沿着入链传播,根据网页信任值对非信任值进行惩罚,利用非信任值检测出作弊网页,以达到优化搜索引擎的目的。

5 结语

本文结合TDR与GBR算法的优点提出双端惩罚算法DP-Rank,在计算信任值时,从源端和目标端双端进行惩罚,在几乎不增加时间成本的基础上提升了对可信网页的排名效果,在WEBSPPAM-UK2006数据集上的实验验证了DP-Rank算法可获得更好的输出结果。当设定可信种子与作弊种子数量分别为200时,与现有基于链接的PageRank、TrustRank算法及信任与非信任同时传播的TDR、GBR算法相比,排名效果均获得提升。

在未来的工作中,可以将双端惩罚算法与分类算法进行融合,以提高可信网页和作弊网页的识别率;根据网页链接结构的动态变化以及时更新网页的信任值/非信任值,从而更好地适应网络的发展。当网页链接结构发生改变时,通过算法更新网络改变部分链接关系值,而非重新计算整个Web网络中的值,以进一步提升双端惩罚算法对可信网页排名的效果。另外,还可以考虑将双端惩罚算法应用于社交网络中,对社交网络中的信息进行信任评估。

参考文献:

- [1] MOHAMMED M A, IBRAHIM D A, SALMAN A O. Adaptive intelligent learning approach based on visual anti-spam email model for multi-natural language[J]. Journal of Intelligent Systems, 2021, 30(1): 774-792.
- [2] NIU X F, LIU G C, YANG Q. OpinionRank: trustworthy website detection using three valued subjective logic[J]. IEEE Transactions on Big Data, 2022, 8(3): 855-866.
- [3] SARDAR A, PRAMANIK P K D. Estimating authors' research impact using PageRank algorithm[C]// Proceedings of ICDMAI, 2022: 471-483.
- [4] GYONGYI Z, GARCIA-MOLINA H, PEDERSEN J. Combating web spam with trustrank[C]// Proceedings of the 30th International Conference on Very Large Data Bases, 2004: 576-587.
- [5] KRISHNAN V, RAJ R. Web spam detection with anti-trustrank[C]// Proceedings of the 2nd International Workshop Adversarial Information Retrieval Web, 2006: 37-40.
- [6] ZHANG X, WANG Y, MOU N, et al. Propagating both trust and distrust with target differentiation for combating link-based web spam[J]. ACM Transactions on the Web, 2014, 8(3): 1-33.
- [7] LIU X, WANG Y, ZHU S, et al. Combating web spam through trust-distrust propagation with confidence[J]. Pattern Recognition Letters, 2013, 34(13): 1462-1469.
- [8] CHATTERJEE M, NAMIN A S. A fuzzy dempster-shafer classifier for detecting web spams[J]. Journal of Information Security and Applications, 2021, 59: 1-9.
- [9] WANG J J. Research on spam detection algorithm based on link and topic relevance [D]. Wuhan: Huazhong University of Science and Technology, 2021.
- [10] 王晶晶. 基于链接与主题相关度的垃圾网页检测算法研究[D]. 武汉: 华中科技大学, 2021.
- [10] CHEN M S, GAO F, WU J H. Web spam detection based on semantic

- features from current page [J]. *Application of Electronic Technique*, 2023, 49(6): 24–29.
- 陈木生, 高斐, 吴俊华. 基于单页语义特征的垃圾网页检测[J]. *电子技术应用*, 2023, 49(6): 24–29.
- [11] SU Y. Web spam detection technology based on multidimensional features of source code and tex[D]. Chengdu: Sichuan University, 2021.
- 苏瑜. 基于源码及文本多维特征的垃圾网页检测技术研究[D]. 成都: 四川大学, 2021.
- [12] WANG R, ZHUANG X, ZHU X, et al. Intelligent web spam detection in the consumer Internet of Things [J]. *IEEE Transactions on Consumer Electronics*, 2024, 71: 1260–1273.
- [13] CHAKRABORTY A, DAS U K, SIKDER J, et al. Content based email spam classifier as a web application using Naïve Bayes classifier [C]//*Proceedings of International Conference on Intelligent Computing & Optimization*, 2022: 389–398.
- [14] AKINYELU A A. Advances in spam detection for email spam, web spam, social network spam, and review spam: ML-based and nature-inspired-based techniques [J]. *Journal of Computer Security*, 2021, 29(5): 473–529.
- [15] ASWANI R, GHRERA S P, CHANDRA S, et al. A hybrid evolutionary approach for identifying spam websites for search engine marketing [J]. *Evolutionary Intelligence*, 2021, 14: 1803–1815.
- [16] WAHEED A, SALAM A, BANGASH J I, et al. Lexicon and learn-based sentiment analysis for web spam detection [J]. *IEEE-SEM*, 2021, 9(3): 97–107.
- [17] TEJA N P, SHAIS K M. Machine learning for spam detection [J]. *Asian Journal of Advances in Research*, 2023, 6(1): 167–179.
- [18] JAZZAR M, YOUSEF F R, ELEYAN D. Evaluation of machine learning techniques for email spam classification [J]. *International Journal of Education and Management Engineering*, 2021, 11(4): 35–42.
- [19] AL-ZOUBI A M, ALQATAWNA J, FARIS H, et al. Spam profiles detection on social networks using computational intelligence methods: the effect of the lingual context [J]. *Journal of Information Science*, 2021, 47(1): 58–81.
- [20] KADDOURA S, CHANDRASEKARAN G, POPESCU D E, et al. A systematic literature review on spam content detection and classification [J]. *PeerJ Computer Science*, 2022, 8: e830.
- [21] BATRA J, JAIN R, TIKKIWAL V A, et al. A comprehensive study of spam detection in e-mails using bio-inspired optimization techniques [J]. *International Journal of Information Management Data Insights*, 2021(1): 100006.
- [22] SHRIVAS A K, DEWANGAN A K, GHOSH S M, et al. Development of proposed ensemble model for spam e-mail classification [J]. *Information Technology and Control*, 2021, 50(3): 411–423.
- [23] AHSAN M, SHARMA T P. Spams classification and their diffusibility prediction on Twitter through sentiment and topic models [J]. *International Journal of Computers and Applications*, 2022, 44(4): 365–375.
- [24] BEVENDORFF J, WIEGMANN M, POTTHAST M, et al. Is Google getting worse? a longitudinal investigation of SEO spam in search engines [C]//*Proceedings of European Conference on Information Retrieval*, 2024: 56–71.
- [25] ZHUANG X, ZHU Y, PENG Q, et al. Using deep belief network to demote web spam [J]. *Future Generation Computer Systems*, 2021, 118: 94–106.
- [26] NIU X F, LIU G C, YANG Q. Trustworthy website detection based on social hyperlink network analysis [J]. *IEEE Transactions on Network Science and Engineering*, 2020, 7(1): 54–65.

(责任编辑:黄健)