文章编号:1001 - 9081(2004)10 - 0158 - 03

基于 Web 结构挖掘的搜索引擎作弊检测方法

冉 丽[1],何毅舟[2],许龙飞[1]
(1. 暨南大学 计算机科学系,广东 广州 510632; 2. 暨南大学 网络中心,广东 广州 510632)
(ranli @jnu. edu. cn)

摘 要:搜索引擎作弊行为从搜索引擎优化中演变而来,却对网络发展带来负面影响。通过构造站内站外精简模型用于判断几类作弊行为,得出 PageRank 改进算法中惩罚因子的公式和其中三个函数的特征,展望了搜索引擎作弊检测方法的发展前景。

关键词:Web 结构挖掘;搜索引擎作弊;精简模型;PageRank
中图分类号: TP393. 07 **文献标识码**:A

Detection method for search engine spam based on Web structure mining

RAN Li[1], HE Yi-zhou[2], XU Long-fei[1]
(1. *Department of Computer Science, Jinan University, Guangzhou Guangdong* 510632, *China*;
2. *Network Center, Jinan University, Guangzhou Guangdong* 510632, *China*)

Abstract: Search engine spam is an offset outcome of search engine spam, imposing negative effect on Web development however. With reduced model of Web site inside and out to judge some kinds of search engine spam, a punish multiplier formula and characteristics of three functions in proved algorithm on pagerank were described. At last, a future look at the detection method for search engine spam was presented.

Key words: Web structure mining; search engine spam; reduced model; pagerank

0 引言

网络将零星的资源汇集到一起,连接成让全世界范围内网络用户共享的财富。一方面,各开放资源的管理者希望更多的用户能共享他提供的信息和服务,以获得更大的商机;另一方面,用户希望能省时省力的寻找到自己真正所需要的资源,于是搜索引擎应运而生。如今搜索引擎已成为继电子邮件之后被用户最为广泛使用的网络工具。在搜索引擎中的排名对各网站意味着商机和实力对比,对网络用户意味着获取资源的准确度和效率,因此搜索引擎排名对各公开网站、特别是电子商务类网站显得尤为重要。

为提高排名而衍生出的搜索引擎优化是近年来的热门技术,其不仅包括计算机技术,还涉及网络营销、网站推广等商业因素。搜索引擎优化使各网站更加注重自身网页的质量和结构的优化,由此带来了网络整体质量的提高,与此同时,假借"优化"之名而出现的各类作弊手段却影响了搜索引擎排名的秩序,给网络用户带来不便。基于 Web 结构挖掘的搜索引擎作弊检测方法正是针对以上问题提出的解决方案之一。

1 搜索引擎作弊检测方法概述

1.1 搜索引擎作弊

搜索引擎作弊是指采用一些特殊的、有悖常规的网页设计手法,以期提高网站排名的行为。虽然"作弊"和"优化"的目的都是为了取得最佳排名,但区别在于:优化使用与内容相关的密度适当的关键词,并对网站结构、页面因素和外部链接进行优化,方便了用户的使用;而作弊使用的是不合常规的方法,其结果使网页质量下降、Web 结构恶化,妨碍了用户的使用。目前,搜索引擎 Google 把属于作弊性质的优化技术分为九种,其搜索策略和排名算法仍在不断研究和变化之中。

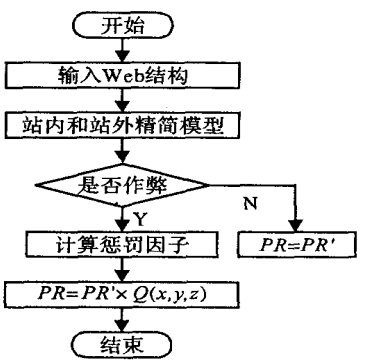

图 1 检测方法步骤

1.2 Web 结构挖掘

Web 结构挖掘是指从 URL 字符串中的目录路径结构、网页内部的树形结构、网页和网站之间的超链接结构中推导信息、知识,其目的在于研究网络的超链接结构、网页归类和概要信息(如网页间的相似性和联系等)。其常见应用有 Google 排名中的 PageRank 算法,用于主题提取的 HITS 算法和在网络社区中的算法。在进行 Web 结构挖掘时,忽略网页内容而着

收稿日期:2004 - 04 - 06;**修订日期**:2004 - 06 - 03
基金项目:广东省科技计划项目(2003C101037);广东省自然科学基金重点项目(010421)
作者简介:冉丽(1978 -),女,四川绵阳人,硕士研究生,主要研究方向:数据挖掘和知识工程; 何毅舟(1978 -),男,河北石家庄人,硕士研究生,主要研究方向:计算机网络; 许龙飞(1946 -),男,广东广州人,教授,主要研究方向:数据挖掘与知识工程.

眼于网络中的链式结构,从而形成数亿个节点和数十亿条边的巨大有向图。Web 结构图的节点不仅可以是网页,也可以是网站,甚至是主题相关的网站类等,这根据研究需要的抽象程度而定。

1.3　两者的结合

基于 Web 结构挖掘的排名欺骗检测技术是根据一个网站内部或外部的 Web 结构,分别建立站内站外精简模型,根据由该模型得出的阈值来判断该网站是否存在搜索引擎作弊行为,然后根据作弊程度对网站中每个网页的 PageRank 值乘上惩罚因子,使企图利用作弊手段来达到提高排名的网页或网站达不到预期效果并且受到惩罚。具体实现步骤如图 1。

2　搜索引擎作弊的 Web 结构分析

2.1　站内链接分析

在搜索引擎作弊中最常使用的手段是增加导入链接数,PageRank 算法决定的排名对网页不仅要求导入链接的数量,更重要的是导入链接的质量。一个拥有 k 页网页的网站在不存在重复链接的情况下,站内最多可以建立 $k\times(k-1)$ 条链接。搜索引擎作弊行为之一就是建立多条不必要的链接指向目的页,以提高该页 PageRank 值,同时又不希望用户看到如此混乱的链接,常采用隐藏链接的手段达到目的。这种隐藏的链接结构旨在逃过用户的肉眼,却难以逃过软件工具的检测,如检索程序 spider 就能一目了然地识别这种结构。这里介绍一种站内精简模型用来衡量网站内存在的冗余链接数目。

建立方法:设一个网站,它的 k 页网页所组成的层次结构形成一棵高度为 h 的网站结构树,该树的根节点是该站主页,且根节点高度为 0。用深度优先的遍历方法遍历该树,将各个节点的节点名称、节点所在高度、节点的子节点数存入二重数组内,用 n_i 表示网页 i,以上属性分别用 $n_i.name$, $n_i.ht$, $n_i.chs$ 表示。$Link_i$ 表示 i 类链接数目,站内精简模型认为在站内只有如下三类链接是必要的:

(1) 每个节点到自己的子节点的链接,站内共有:

$$Link_1 = k - 1$$

(2) 每个节点到自己父节点或祖先节点的链接,站内共有:$Link_2 = \sum_{i=1}^{k} n_i.ht$

(3) 每个节点到自己兄弟节点的链接,站内共有:

$$Link_3 = \sum_{i=1}^{k}((n_i.chs - 1)\times n_i.chs)$$

站内精简模型得出链接数为 $\mathbb{L}ink$,如果站内链接数超过 $\mathbb{L}ink$ 则认为存在冗余链接。

$$\mathbb{L}ink = \sum_{j=1}^{3} Link_j = \sum_{i=1}^{k}(1 + n_i.ht + ((n_i.chs - 1)\times n_i.chs)) - 1 \quad (1)$$

2.2　站外链接分析

站外链接结构比站内链接结构更为复杂,不仅因为其研究的范围过大,以及其对应的图形不是简单的树形而是网状图,而且还关系到主题相关、主题相关类规模大小等问题。在站外的互联中,“链接工厂”被视为搜索引擎作弊行为,“链接工厂”指由大量网页交叉链接而构成的一个网络系统,系统内的站点以“互惠”方式相互提供链接,往往忽略主题相关性原则,以达到提高站点排名的目的。在 PageRank 算法中,如果存在网页 A 到网页 B 的链接,则认为 A 投了 B 一票。在链接工厂中,由于交换链接,以站点为节点、站外链接为边的 Web 结构图中必然存在大量回路,该回路上除顶点与终点不再有相同的站点。回路可以理解为该站点投了自己一票,当一个站点存在多条回路时,将被视为重复投票,而且应该受到适当的惩罚。同时,互联网的目的是资源共享,因而合理使用链接而形成回路链接是必然存在的。如何有效惩罚作弊行为而不影响合理优化的积极性正是判断惩罚因子质量的尺度,首先要建立站外精简模型,确定站外链接形成的合理回路数目。

建立方法:以研究的网站为中心划出主题高度相关类。主题高度相关类划分办法是利用 2.1 节中的公式(5),建议取与该站相关度最高的 5 ±2 个网站进入该类。主题高度相关类是站外精简模型的基础,它是以该站为中心的主题相关类的真子集。设该类包括 m 个站点,如果任意两个站点之间都存在相互链接,利用排列组合的方法可得到从某个站点出发的回路链接数目为 $OLink$,如果站外回路链接超过 $OLink$ 的回路则认为是冗余的。

$$OLink = \sum_{k=0}^{m-2}\frac{(m-1)!}{k!} \quad (2)$$

2.3　重复链接分析

重复链接不论存在于站内还是站外,都被认为是作弊行为,通常的表现形式为隐藏链接,在论坛上发布重复链接等。

3　作弊行为对 PageRank 算法的影响

3.1　PageRank 改进算法原理

实现方法是把网站作为一个整体看待,如果站内存在冗余链接,或者该站加入了链接工厂,或者存在重复链接,那么站内的所有页面都应受到惩罚,其 PageRank 值都会乘上一个惩罚因子,以削弱虚假链接对 PR 值的正面影响并进行相应惩罚。$PR'(i)$ 是网页 i 原来的 PageRank 值,改进算法形如:

$$PR(i) = PR'(i)\times Q(x,y,z) \quad (3)$$

3.2　惩罚因子

以下是对惩罚因子 $Q(x,y,z)$ 的讨论:

(1) 对站内冗余链接

设 x 表示所研究网站的实际链接数目,$\mathbb{L}ink$ 由公式(1)求得,表示该网站的站内精简模型得出的链接数目,k 表示该网站拥有的网页数目,$f(x)$ 是函数,对于变量 x, $Q(x,0,0) = 1/F(x)$,函数 $F(x)$ 定义如下:

$$\leq x \leq F(x) = \begin{cases} 1 & 0 \leq x \leq \mathbb{L}ink \\ f(x - \mathbb{L}ink) & \\ f(k\times(k-1) - \mathbb{L}ink) & x > k\times(k-1) \end{cases} \quad (4)$$

(2) 对站外冗余链接

计算站外冗余链接的惩罚因子,首先考虑互相链接的网站间的相关度问题。将所有网站进行概念分层形成网站分类树,树中每个节点表示一个主题类,越靠近根节点的节点所代表的主题类范围越大,然后层层细分,直到分成一定规模的主题类为止,在互联网中的每个网站都可以在这棵网站分类树中找到所属类别的节点,而且该节点的子节点中再也找不到适合这个网站的主题相关类。网站间相关度的计算方法可以借鉴“知网”中词语相似度的计算方法[5]。需要强调的是这里

是基于网站分类树,而不是"知网"中的语义分类树。

设 $DIS(S_i, S_j)$ 是 S_i, S_j 在网站分类树上的最短距离,α 是可调节参数,$REL(S_i, S_j)$ 表示网站 i 和网站 j 的相关性,以下是转换关系式的形式之一:

$$REL(S_i, S_j) = \alpha / (DIS(S_i, S_j) + \alpha) \quad (5)$$

另外,$REL(S_i, \Delta_j)$ 表示由网站 S_i 出发经过集合 Δ_j 中 m 个网站形成的回路链接的相关度,其中 $S_k \in \Delta$,S_k 和 S_{k+1} 是回路 j 中的相邻网站,计算公式为:

$$REL(S_i, \Delta_j) = \frac{REL(S_i, S_1) + \sum_{k=1}^{m-1} REL(S_k, S_{k+1}) + REL(S_m, S_i)}{m + 1} \quad (6)$$

y 表示该网站在外部链接中形成的回路数,$OLink$ 由公式(2)求得,表示该网站的站外精简模型得出的链接数目,$g(y)$ 是函数,由于互联网是无边无际的,所有 y 取值为 $[0, +\infty)$,即上限趋于无穷,对于变量 y,$Q(0, y, 0) = 1/G(y)$,函数 $G(y)$ 定义如下:

$$G(y) = \begin{cases} 1 & 0 \le y \le OLink \\ (g(y - OLink) \times y) / \sum_{j=1}^{y} REL(S_i, \Delta_j) & \end{cases} \quad (7)$$

(3) 重复链接

z 表示该网站在站内外存在的重复链接数,$h(z)$ 是函数,对于变量 z,$Q(0, 0, z) = 1/H(z)$,函数 $H(z)$ 定义如下:

$$H(z) = \begin{cases} 1 & z = 0 \\ h(z) & \end{cases} \quad (8)$$

(4) 惩罚因子合成公式

所以在只考虑站内冗余链接、站外回路链接、重复链接三类作弊行为时,"惩罚因子"的数学表达式如下:

a) 当 $0 \le x \le k \times (k - 1)$,$y \ge 0$,$z \ge 0$:

$$Q(x, y, z) = 1/(F(x) \times G(y) \times H(z)) \quad (9)$$

b) 当 $x > k \times (k - 1)$,$y \ge 0$,$z \ge 0$:

$$Q(x, y, z) = \frac{1}{F(k \times (k - 1) - ILink)} \times \frac{1}{G(y) \times H(z + x - k \times (k - 1))} \quad (10)$$

3.3 对 $F(x)$,$G(y)$,$H(z)$ 的函数形式特征的讨论

公式(4),(7),(8)给出了 $F(x)$,$G(y)$,$H(z)$ 的函数定义,但具体形式需根据各搜索引擎对站内冗余链接、站外回路链接和重复链接这三类搜索引擎作弊行为的惩罚力度而定,并不存在同一形式。在确定这三个函数时,应按照以下原则:

(1) 三个函数在其取值范围内都是非递减函数;

(2) 三个函数在其取值范围内,函数值都必须大于等于 1;

(3) 在搜索引擎作弊行为中,站内冗余链接、站外回路链接和重复链接三类作弊行为的恶劣程度依次增加,故 $F(x)$,$G(y)$,$H(z)$ 函数的递增速度也应依次增加。

图 2 是这些函数形式特征的图示。

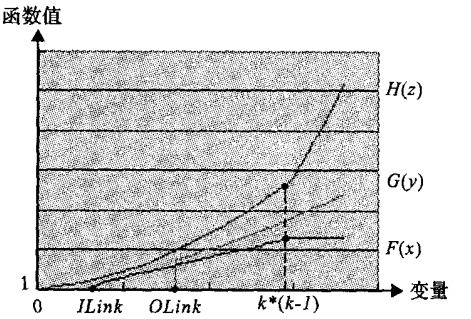

图 2 三个函数比较图形

4 结语

本文所介绍的方法是在分析 Web 结构的基础上来判断搜索引擎作弊行为,但是只能用于判别和计算有限的几类作弊行为。此外还存在重复性关键词、误导性关键词、日志欺骗等多种形式的作弊行为,所以应将 Web 内容挖掘、Web 结构挖掘和 Web 行为挖掘结合起来应用于搜索引擎作弊检测方法中。

参考文献:

[1] 李晓星,李星. 搜索引擎与 Web 挖掘进展[M]. 北京:高等教育出版社,2003.

[2] 杨炳儒,李岩,陈新中,等. Web 结构挖掘[J]. 计算机工程,2003,39(29):28 - 30.

[3] HAN J IAWEI, KAMBER M. 数据挖掘概念与技术[M]. 北京:机械工业出版社,2001.

[4] 王晓宇,周傲英. 万维网的链接结构分析及其应用综述[J]. 软件学报,2003,14(10):1768 - 1780.

[5] 刘群,李素建. 基于《知网》的词汇语义相似度计算[J]. Computational Linguistics and Chinese Language Processing, 2002,7:59 - 76.

(上接第 157 页)

论分析是一致的,CIPM 的回溯时间比方案 J 明显短。

2 结语

我们提出的 CIPM 能够改善方案 J 中的分组丢失问题,回溯时间低于方案 J,回溯时间是影响分组丢失的主要原因。

当前的所有 IP 回溯方案都是通过寻找攻击路径来回溯攻击源。CIPM 没有依循这种思路:它首先确定攻击源所在的网络,然后在该网络中将攻击源识别出来。

参考文献:

[1] BABA T, MATSUDA S. Tracing Network Attacks to Their Sources[J]. IEEE Internet Computing March/April 2002,6(2):20 - 26.

[2] KUZNETSOV V. An evaluation of different IP traceback approaches [EB/OL]. http://www.sm.luth.se/csee/csn/publications/ip-traceback.pdf,2004.

[3] BELOVIN S, LEECH M, TAYLOR T. ICMP Traceback Messages. Internet draft, work in progress, OCT[EB/OL]. http://www,ietf.org/internet-frafts/fraft-ietf-itrace-01.test,2001.

[4] SAVAGE S, WETHERALL D, KARLIN A, *et al*. Practical network support for ip traceback[A]. Proceedings of the 2000 ACM SIGCOMM Conference[C]. 2000.

[5] SNOEREN AC, *et al*. Single-Packet IP Traceback[J]. IEEE/ACM TRANSACTIONS ON NETWORKING, 2002,10(6).

[6] 梁丰,赵新建, Yau D. 通过自适应随机数据包标记实现实时 IP 回溯[J]. 软件学报,2003,14(05):1005 - 1010.