

用信息-摘要算法提高 Web 信息检索效率的研究

杨文忠, 章 兢

(湖南大学 电气与信息工程学院, 湖南 长沙 410082)

摘 要:针对常用搜索引擎返回给用户的信息中包含大量重复网页的缺陷,提出了一种基于信息-摘要算法的去除重复网页算法。由于算法的成熟,该算法易实现,可移植性强。实验证明该算法能有效地去除常用搜索引擎返回的重复网页,从而为 Internet 用户提高信息检索效率,具有较强的实用价值。

关键词:信息-摘要算法;近似镜像网页;信息检索

中图分类号:TP301.6

文献标识码:A

文章编号:1673-629X(2006)06-0222-02

Using Message-Digest Algorithm for Improving Efficiency of Web Information Searching

YANG Wen-zhong, ZHANG Jing

(College of Electric and Information Engineering of Hunan University, Changsha 410082, China)

Abstract: The returning information of the usual search engines often includes massive repeated pages. Aims at it, an inspecting algorithm of approximate mirror pages is proposed in this paper. Because of the mature of MD5, this algorithm can be implemented easily and is portable. The experiment shows that it can remove the repeated pages from usual search engines effectively and can improve the searching efficiency of Internet users. It has good application foreground.

Key words: message-digest algorithm; approximate mirror pages; information searching

0 引言

目前 Internet 的爆发式发展改变了人们的工作、学习、生活的方式^[1],现有的搜索引擎工具是人们获取网络信息资源的主要手段^[2]。这些搜索引擎有许多优点,但是用户进行 Web 信息检索时,它们往往返回大量的近似镜像网页(可看作重复网页)。针对搜索引擎查询 Web 信息所存在的局限性,很多研究人员进行了近似文本检测算法的研究,通过消除近似镜像网页来提高检索效率。国际上对近似镜像文本的检测算法的研究最初主要是针对大型文件系统的,后来又被拓展应用于数字化图书馆项目和搜索引擎系统。美国 Arizona 大学的研究人员采用计算文档的重叠程度的方法来发现一个大型文件系统中的相似文件。作为 Stanford 大学数字化图书馆项目的一部分, N. Shivakumar 等人研制了 SCAM (Stanford Copy Analysis Mechanism)原型系统^[3],用于发现相似的数字化文档。后来 Narayanan 等人又对 SCAM 原型系统的近似镜像检测算法做了改进^[4],并被应用于 Stanford 大学开发的 google 搜索引擎系统,取得了较好的效果。但这些算法的空间复

杂度和时间复杂度仍然是相当大的,若应用于海量的搜索引擎系统(通常包含上亿个 Web 页面),仍然难以取得理想的效果。

基于此,考虑到基于关键词匹配的搜索引擎系统的特点,结合使用网页的向量空间模型,在此提出一种基于 MD5 算法的近似网页检测算法,用于快速、有效地发现 WWW 上的重复或相似网页。

1 信息-摘要算法描述

信息-摘要算法是在 20 世纪 90 年代初由 MIT Laboratory for Computer Science 和 RSA Data Security Inc 的 Ronald L. Rivest 开发出来,经 MD2, MD3 和 MD4 发展而至 MD5^[5]。它的作用是让大容量信息在用数字签名软件签署私人密钥前被“压缩”成一种保密的格式(就是把一个任意长度的字节串变换成一定长的大整数),MD5 广泛用于加密和解密技术上。

MD5 的典型应用是对一段信息(Message)产生信息摘要(Message-Digest),以防止被篡改。比如,在 UNIX 下有很多软件在下载的时候都有一个文件名相同、文件扩展名为 .md5 的文件,在这个文件中通常只有一行文本,大致结构如下:

MD5 (tanajiya.tar.gz)=0ca175b9c0f726a831d895e269332461

收稿日期:2005-09-13

基金项目:教育部科学与技术研究重点项目(教技司 2001224 号)

作者简介:杨文忠(1970-),男,湖南花垣人,讲师,硕士研究生,研究方向为数据挖掘;章 兢,博士生导师,教授,研究方向为数据挖掘。

这就是 tanajiya. tar. gz 文件的数字签名。MD5 将整个文件当作一个大文本信息,通过其不可逆的字符串变换算法,产生了这个惟一的 MD5 信息摘要。如果在以后传播这个文件的过程中,无论文件的内容发生了任何形式的改变(包括人为修改或者下载过程中线路不稳定引起的传输错误等),只要对这个文件重新计算 MD5 时就会发现信息摘要不相同,由此可以确定得到的只是一个不正确的文件。如果再有一个第三方的认证机构,用 MD5 还可以防止文件作者的“抵赖”,这就是所谓的数字签名应用。

对 MD5 算法简要的叙述可以为:MD5 以 512 位分组来处理输入的信息,且每一分组又被划分为 16 个 32 位子分组,经过了一系列的处理后,算法的输出由 4 个 32 位分组组成,将这 4 个 32 位分组合级联后将生成一个 128 位散列值。

在 MD5 算法中,首先需要对信息进行填充,使其字节长度对 512 求余的结果等于 448。因此,信息的字节长度 (Bits Length) 将被扩展至 $N * 512 + 448$, 即 $N * 64 + 56$ 个字节 (Bytes), N 为一个正整数。填充的方法如下:在信息的后面填充一个 1 和无数个 0,直到满足上面的条件时才停止用 0 对信息的填充。然后,在这个结果后面附加一个以 64 位二进制表示的填充前信息长度。经过这两步的处理,现在的信息字节长度 $= N * 512 + 448 + 64 = (N + 1) * 512$, 即长度恰好是 512 的整数倍。这样做的原因是为满足后面处理中对信息长度的要求。

将上面 4 个链接变量复制到另外 4 个变量中: A 到 a , B 到 b , C 到 c , D 到 d 。

主循环有四轮,每轮循环都很相似。第一轮进行 16 次操作。每次操作对 a, b, c 和 d 中的其中 3 个做一次非线性函数运算,然后将所得结果加上第四个变量、文本的一个子分组和一个常数。再将所得结果向右环移一个不定的数,并加上 a, b, c 或 d 中之一。最后用该结果取代 a, b, c 或 d 中之一。

以下是每次操作中用到的 4 个非线性函数(每轮一个)。

$$F(X, Y, Z) = (X \& Y) \mid (\bar{X} \& Z)$$

$$G(X, Y, Z) = (X \& Z) \mid (Y \& \bar{Z})$$

$$H(X, Y, Z) = X \oplus Y \oplus Z$$

$$I(X, Y, Z) = Y \mid (X \oplus \bar{Z})$$

这 4 个函数的说明:如果 X, Y 和 Z 的对应位是独立和均匀的,那么结果的每一位也应是独立和均匀的。

F 是一个逐位运算的函数。即,如果 X , 那么 Y , 否则 Z 。函数 H 是逐位奇偶操作符。

假设 M_j 表示消息的第 j 个子分组(从 0 到 15), $<< FF(a, b, c, d, M_j, s, ti)$ 表示 $a = b + ((a + (F(b, c, d) + M_j + ti)) <<$

$GG(a, b, c, d, M_j, s, ti)$ 表示 $a = b + ((a + (G(b, c, d) + M_j + ti)) <<$

$HH(a, b, c, d, M_j, s, ti)$ 表示 $a = b + ((a + (H(b, c, d)$

$+ M_j + ti)) <<$

$I(a, b, c, d, M_j, s, ti)$ 表示 $a = b + ((a + (I(b, c, d) + M_j + ti)) <<$

MD5 中有 4 个 32 位被称作链接变量 (Chaining Variable) 的整数参数,它们分别为: $A = 0x01234567, B = 0x89abcdef, C = 0xfedcba98, D = 0x76543210$ 。

当设置好这 4 个链接变量后,就开始进入算法的四轮循环运算。循环的次数是信息中 512 位信息分组的数目。

常数 ti 可以如下选择:在第 i 步中, ti 是 $4294967296 * \text{abs}(\sin(i))$ 的整数部分, i 的单位是弧度。(4294967296 等于 2 的 32 次方)。

所有这些完成之后,将 A, B, C, D 分别加上 a, b, c, d 。然后用下一分组数据继续运行算法,最后的输出是 A, B, C 和 D 的级联。下面是对该 MD5 算法的测试结果:

MD5 (“”) = d41d8cd98f00b204e9800998ecf8427e

MD5 (“a”) = 0cc175b9c0f1b6a831c399e269772661

MD5 (“abc”) = 900150983cd24fb0d6963f7d28e17f72

MD5 (“message digest”) = f96b697d7cb7938d525a2f31aaf161d0

MD5 (“abcdefghijklmnopqrstuvwxyz”) = c3fed3d76192e4007dfb496cca67e13b

MD5 (“ABCDEFGHIJKLMNOPQRSTUVWXYZabcdefghijklmnopqrstuvwxyz0123456789”) = d174ab98d277d9f5a5611c2c9f419d9f

MD5 (“12345678901234567890123456789012345678901234567890123456789012345678901234567890”) = 57edf4a22be3c955ac49da2e2107b67a

2 基于信息-摘要算法的重复网页去除算法

式(1)描述了利用 MD5 散列值的近似镜像网页的检测算法,用 P_i 表示第 i 个网页,其权值最高的前 N 个关键词构成的特征项集合为 $T_i = \{t_1, t_2, \dots, t_m\}$, 其对应的特征向量为 $W_i = \langle W_{i1}, W_{i2}, \dots, W_{im} \rangle$, 前 N 个关键词拼接成的字符串用 $\text{Concatenate}(T_i)$ 表示,用 $\text{MD5}(X)$ 来表示字符串 X 的散列值,用 $\text{Repeat}(P_i, P_j)$ 表示 P_i 和 P_j 互为近似镜像网页,用 $A \Rightarrow B$ 表示“若 A 成立则 B 成立”。

$$(\text{MD5}(\text{Concatenate}(T_i))) =$$

$$\text{MD5}(\text{Concatenate}(T_j)) \Rightarrow \text{Repeat}(P_i, P_j) \quad (1)$$

可以看出,上述算法待检测两个文本前 N 个关键词求 MD5 散列值,当两个网页的权值最高的前 N 个关键词集合相同时就认为二者是互为近似镜像的。它不仅要求两个相似网页的前 N 个关键词相同,其顺序也是一致的(按权值排序),因而这种算法在消除近似镜像网页时不会误消相似网页,不影响查全率而达到减少返回页面的效果。

3 结论

由于 MD5 算法的成熟与易实现性,采用比较对象间的 MD5 散列值的方法来消除近似镜像网页,程序的移植性强,消除重复网页的效果也比较好的,克服了现有算法时间复杂度和空间复杂度的缺点。此外,笔者所做的研究

(下转第 226 页)

$$W(k+1) = W(k), \text{ for } d(k)X^T(k)W(k) \geq \delta$$

$$W(k+1) = W(k) + \frac{\gamma d(k)}{n+1} X(k)[L - d(k)X^T(k)W(k)], \text{ for } d(k)X^T(k)W(k) < \delta$$

对式中各个参数的选取文中给出了具体的经验值。这个方法和感知机学习规则都遵循了 Hebb 学习规则的数学描述,新的权值是旧的权值加上或减去输入:

$$W_{ij}^{\text{new}} = W_{ij}^{\text{old}} + eP_{ij} \quad (e \text{ 可取正负值}) \quad (1)$$

并证明经过一步计算新的权值就可以翻转该神经元的数字输出量。为了使新的权值改动尽量的小,以保证它的改动对前面训练好的样本产生的影响尽量小,将(1)式改写成 $W_{ij}^{\text{new}} = W_{ij}^{\text{old}} + \gamma eP_{ij}$,把 γ 称为学习速率, γ 在(0,1)之间取值, γ 越小新的权值对旧的权值的改动也越小,但这样可能使一次迭代无法满足翻转神经元的数字输出量的要求,因此必须验证迭代后的新权值输出的正确性,一旦新权值满足了翻转要求,就把它作为下次学习的新的权值,这样既保证了翻转又保证了在特定学习速率下权值的改动最小。

由此可以得到带学习速率的感知机学习规则在 MR II 算法中进行调权的改进的 Madaline 学习算法,其步骤如下:

(1)初始化网络结构包括网络层数和各层结点数,用一个随机数作为网络中的各个权值的初始值。保存这些权值数据。

(2)随机输入一个样本矢量 X_K 和它的理想输出 T_K ,按层一步一步地计算出每个 Adaline 的实际输出,分别保留它们的模拟量和数字量。

(3)比较理想的输出 T_K 和实际输出的不相同个数,如果出错个数为 0,则到(7);如果出错个数不为 0,则从第一层开始,根据最小干扰原则,即找出第一层中神经元模拟量最接近于 0(即其加权后的绝对值最小)的那个神经元。

(4)翻转该神经元(使其输出从 0 变为 1 或从 1 变为 0),即让它的输出改变符号。逐层计算网络最后的输出,比较它和理想输出之间的误差,如果输出误差个数减少,则接受这个翻转的尝试,记下此时该神经元的输出作为它

的理想输出,用带学习速率的感知机学习规则调整它的权值,每次迭代后新的权值要验证是否已经使该神经元翻转,一旦翻转就保留这个权值到(3);如果输出误差个数没有减少,就不接受它的翻转仍保留原来数据,即上一步改变的输出数字量的符号恢复,到(5)。

(5)转入下一个神经元,即其模拟输出次接近 0 的神经元,仍按(4)规则训练。

(6)当这一层的单个神经元训练结束后,再按两个一组地训练,然后再按三个一组,……,直到 k 个一组(k 为第一层神经元个数)。如仍然不能符合要求,再用同样的方法训练第二层,第三层……直到输出与要求响应之间的理想输出相吻合。保存符合该样本理想输出的数据。

(7)输入另一个新样本,用同样的方法训练。

3 总 结

计算和算法是人类自古以来十分重视的研究领域。在神经网络的发展中,找到适合特定网络模型的最佳算法,有效提高网络的运行能力,是很具有现实意义的工作。Madaline 网络由于具有明显的离散特点,且其学习算法 MR II 也为其改进提供了很好的平台。故恰当地运用这个特点,结合改进的感知机学习规则来有效减少权值的改变量,可达到提高学习能力的目的。同时,MR II 中还有很多细节值得更进一步的研究探讨。

参考文献:

- [1] Widrow B, Winter B, Rodney Gerard Madaline RULE II [M]. CA, USA: Stanford University, 1989.
- [2] 胡守仁. 神经网络导论 [M]. 长沙: 国防科技大学出版社, 1993.
- [3] 张立明. 人工神经网络模型及其应用 [M]. 上海: 复旦大学出版社, 1993.
- [4] Hagan M T, Demuth H B, Beale M H. 神经网络设计 [M]. 戴葵等译. 北京: 机械工业出版社, 2002.
- [5] Winter R G. Madaline Rule II: A new method for training networks of Adalines [D]. CA, USA: Stanford University, 1989.

(上接第 223 页)

是对常用搜索引擎的返回结果进行二次处理,消除重复网页,从而大大提高用户的搜索效率,具有较大的实用价值。

参考文献:

- [1] 李增智,李平均,王广荣. 计算机网络管理系统的若干重要问题[J]. 微机发展, 2000, 10(2): 7-10.
- [2] 葛新红. 数据挖掘软件应用分析[J]. 微计算机应用, 2005, 26(3): 374-377.
- [3] Shivakumar N, Garcia - Molina H. Finding near - replicas of

documents on the Web [A]. In proceedings of the Workshop on Web Databases [C]. [s. l.]: [s. n.], 1998. 204-212.

- [4] Shivakumar N. SCAM: A copy detection mechanism for digital documents [A]. In proceedings of 2nd International Conference in Theory and Practice of Digital Libraries [C]. Austin, Texas: [s. n.], 1995.
- [5] 王贵竹,李津生,洪佩琳. MD5 报文摘要算法与 IPv6 认证 [J]. 小型微型计算机系统, 2001, 22(1): 126-128.