一种基于RPUC的Web文档索引库的更新算法*)

熊海灵 伍 胜 余建桥 李 航
(西南农业大学信息学院 重庆400716)

摘 要 为提高搜索引擎文档索引库有效性验证的效率,本文提出了一种综合考虑网页更新频度、用户兴趣度及其内容重要程度诸因素相结合以确定文档索引库更新队列的算法。算法将用户的检索率、点击率、网页的Page Rank值和更新频度作为一个特征向量,与不同种类的网页的特征权值组成的矩阵相乘,求得网页的类型向量,依据类型向量实现对文档索引库更新队列的优化,算法改进了统一更新策略周期长、单一更新策略可能产生改变频繁而非常重要的网站长期又得不到更新的问题。

关键词 搜索引擎,索引数据库,检索率,Page Rank,更新频度,点击率

A Refreshment Algorithm for Web Indexed Database Based on RPUC

XIONG Hai-Ling WU Sheng YU Jian-Qiao LI Hang
(College of Information, Southwest Agricultural University, Chongqing 400716)

Abstract In order to improve the efficiency of the validity check on the indexed database of search engine, an algorithm for the refreshment of Web indexed database is presented, which is based on the RPUC, i.e. Retrieval ratio, Page Rank, Updated ratio, Click ratio. They constitute the feature vector of a Web page. Cross multiplying the feature vector and matrix, which is consisted of the characteristic values of various Web pages, type vectors of Web pages can be calculated respectively. By means of the type vectors, indexes in refreshing queue can be arranged optimally. Eventually, demerits of the uniform freshness strategy and personal freshness strategy for indexed database are eliminated effectively.

Keywords Search engine, Indexed database, Retrieval ratio, Page Rank, Updated ratio, Click ratio

1 引言

随着Web信息的迅速增加,搜索引擎从1995年开始逐渐发展了起来。查全率和查准率是衡量一个搜索引擎的主要性能指标[7]。搜索引擎文档索引库的内容的获取、组织与更新是提高搜索引擎精度指标和搜索结果的"新鲜性"的关键因素[2,3]。目前对搜索引擎文档索引库的内容的获取、组织都有比较深入的研究,而对文档索引库更新策略的研究还不多见。事实上,搜索引擎收集的网页数量和其文档索引库的更新速度存在着不可调和的矛盾,因此文档索引库更新策略的问题是一个战略性问题,是一个迟早都要面临和解决的问题。

目前,已存在的更新方案大致可归为以下两类[6]:一是统一更新策略,网络蜘蛛以同样的频率访问集合中的所有网页,而不考虑这些网页的改变频率。二是个体更新策略,不同网页其改变频率也不同。直觉上,更多的刷新应该分配给那些更新快的页面,但研究表明,用较高的频率刷新更新快的页面并不一定是明智之举,频繁刷新改变快的网页不能明显提高搜索效率,因为可能产生那些改变频繁的网站长期得不到更新的问题[1,5]。

本文提出了一种Web文档索引库的快速更新算法,该算法综合利用文档索引的检索情况(检索率:R)、文档的页面权值(Page Rank:P)、文档的更新情况(更新频度:U)以及检索出的文档索引的点击情况(点击率:C)(统称RPUC)对文档索引进行分类,并确定其更新周期,按照此更新周期进行信息

*)本论文得到国家自然科学基金(40731061)和重庆市教委科学技术研究项目资助。**熊海灵** 博士生,主要研究方向为信息检索,数据挖掘。**伍 胜** 硕士,主要研究方向为Web技术。**余建桥** 博士,教授,主要研究方向为数据库与人工智能。**李 航** 教授,博导,主要研究方向为分形理论。

找到了一种如何把离散数字空间和连续空间相对应转化的方法。实现了用混沌系统对文本文件的加密。此算法优点十分明显。

1)利用r_{max}控制密文分布和加密时间。引进了参数r_{max},通过此参数可以有效地控制加密时间和密文分布之间的平衡,如果r_{max}取得大密文分布更为广泛,分布更为均匀,但同时会加长加密解密的时间。反之,如果r_{max}取得太小,则密文分布在狭窄区间,分布不均匀,但同时会缩短加密的时间。

2)加密强度提高。每一次的加密中引入了参数$r(r>250)$,r是一个伪随机序列,从而使得每一次加密的轮数都不一样,增强了密文的强度。

3)密文的空间加大。由于一个汉字由两个字节所组成,因此加密后的密文是明文的两倍,所以增加了加密后的传输时间。

参 考 文 献

1 Baptista M S. Cryptography with chaos Phys. Lett. a, 1998, 240: 50
2 Schneier B. Applied Cryptography: Protocols, Algorithms, and Source Code. In: C. Wiley, New York, 1996
3 Wong W-K, Lee L-P, Wong K-W. Comput. Phys. Commun, 2001, 138: 234
4 郝柏林. 从抛物线谈起——空气动力学引论[M]. 上海:上海科技教育出版社,1997

的有效性验证。改进了统一更新策略周期长、单一更新策略可能产生改变频繁而非常重要的网站长期又得不到更新的问题。

2 Web 页的变化规律及其索引更新

2.1 Web 页的变化规律

大量的实验研究表明 Poisson 过程能较好地描述 Web 页面变化的规律[1]。所谓 Poisson 过程，就是一个描述随机事件序列的数学模型，它主要有3个特点：(1)事件随机发生；(2)事件之间相互独立；(3)以一个固定频率反复出现。直观上，互联网上 Web 页面被修改变化的情况也符合 Poisson 过程的3个基本特点。对于一个 Poisson 过程，假设它的变化频率为 λ，用随机变量 $X(t)$ 表示在时间段 $(0,t]$ 内发生变化的次数，那么对任何 $s=0$ 和 $t>0$，随机变量 $X(s+t)-X(s)$ 有以下的 Poisson 概率分布：

$$P\{X(s+t)-X(s)=k\}=(\lambda t^k/k!)e^{-\lambda t} \quad (1)$$

由此，变化频率为 λ 的 Web 页面在时间间隔 $(s,s+t]$ 内不发生变化(取 $k=0$)的可能性为 $e^{-\lambda t}$，进一步，其发生变化的可能性为 $1-e^{-\lambda t}$。

在实际应用中，对网页设定一段监控时间 T，假设在这段时间内网页改变了 X 次，那么网页改变频率的估算值就为 X/T。假定每隔时间 t 访问一次网页 p，并且一共访问了 n 次。用 X_i 表示在第 i 次访问中网页是否变化过，即：

$X_i=1$ (如果在第 i 次访问时网页改变了)；

$X_i=0$ (否则)；

于是，网页总的改变次数为 $X=x_1+x_2+\cdots+x_n$，访问的总时间 $T=nt=n/f$，其中 f 是访问网页的频率，于是网页的改变频率与访问网页的频率之比率为 $u=\lambda/f$。

给定 X_i 和访问的总时间 $T=n/f$，要评测网页的改变频率 λ，由于 $\lambda=fu$，因此可以通过评测 u 来评测 λ。u 的估计值为：

$$\hat{u}=\lambda/f=(1/f)(X/T)=X/n \quad (2)$$

事实上，u 的估计值总是比其真实值小，因为监控所得到的 X 的值总是比实际网页改变的次数要少。为此，Cho[1]提出了另外一种估算 u 的方法：在时间 t 内，网页不改变的概率 $p=e^{-\lambda t}=e^{-u}$，所以有 $u=-\ln p$。如果在时间 T 内，监测到的网页未发生改变的次数为 $n-X$，那么：

$$\hat{u}=-\ln((n-x)/n) \quad (3)$$

根据 Cho 的研究，(3)式比(2)能更好地估计 u。在此基础上 Cho 又修正了 $n-x=0$ 时，$\hat{u}\to 8$ 的情况，得到下面估算 u 的方法：

$$\hat{u}=-\ln((n-x+0.5)/(n+0.5)) \quad (4)$$

(4)式比(3)式更科学。这实际上为我们后边对 Web 页更新频度的估算提供了一种新的思路和方法。

2.2 Web 文档索引的更新过程

索引信息库的更新维护过程通常包含两个方面[5]：验证索引信息库中已有数据的有效性；重新收集从上一次维护结束以来内容发生了变化的文档和新出现的文档信息。有效性验证的根据是文档的最新修改时间，搜索引擎通过发送带有头域 if-modified-since 的 HTTP HEAD 请求来探查文档是否仍然可以访问、文档内容是否发生了变化。如果服务器的响应表明文档已不能被访问，则从索引信息库中删除对应的记录；如果文档的内容发生了变化，则把文档地址加入到一个特定的目标列表中。这个目标列表将用来启动网络蜘蛛进行新一轮文档信息收集过程。

更新网页索引数据库，以反映出网页内容的更新情况，增加新的网页信息，去除死链接，并根据网页内容和链接关系的变化重新排序。这样，网页的具体内容和变化情况就会反映到用户查询的结果中。

2.3 Web 文档索引更新策略的确定

传统 Web 采集器根据自己的需要采集足量的信息后停止更新，当一段时间后这些数据过时了，它会重新采集一遍来代替原来的采集信息，这叫周期性更新。而另外一种方法，对待旧的页面采取增量式更新，也就是说，采集器在需要的时候采集新产生的或者已经发生变化了的页面，而对于没有变化的页面不进行采集[8]。和周期性信息更新相比，增量式信息更新能极大地减少数据的采集量进而极大地减少采集的时空开销，因此它成为实际系统的首选和研究热点。Google、Mercator 和 Internet Archive 都是增量式信息更新系统。但增量式信息更新在减少时空开销的同时，却增加了算法的复杂性和难度，比如如何判断某个页面的变化频度等。

根据已有的研究和实际需要，搜索引擎系统应该首先更新那些本身变化比较快、内容比较重要、用户比较感兴趣的 Web 页的索引，而且应将这些因素加以综合。基于以上的考虑，本文根据文档索引的检索情况(检索率：R)、文档的页面权值(Page Rank：P)、文档的更新情况(更新频度：U)以及检索出的文档索引的点击情况(点击率：C)对文档索引进行分类，从而为增量式更新过程中需要更新文档的确定提供依据。

3 基于 RPUC 的索引库的更新方法

为描述方便，先定义系统监测周期为搜索引擎连续两次的索引信息有效性验证之间的时间间隔。

检索率(Retrieval ratio：R_i)单位系统监测周期内对该索引文档 i 检索的次数，即自该 Web 页的索引文档加入数据库以来，对该索引文档 i 检索的次数 r 与该 Web 页的索引文档加入数据库以来系统进行有效性验证次数 m 的比值。

$$R_i=r/m \quad (5)$$

页面权值(Page Rank：P_i)最近一次系统有效性验证后计算出的 Page Rank 值[4]。

$$P_i=(1-d)+d(P_{i1}/N_{i1}+\cdots+P_{ij}/N_{ij}) \quad (6)$$

假设 j 个有链接向网页 i 的网页，N_{ij} 表示链接向网页 i 的所有网页 j 的总数。参数 d 可以在0到1之间取值，通常取为0.85。

更新频度(Updated ratio：U_i)理论上为单位系统监测周期内 Web 页面改变的次数，即自该 Web 页的索引文档加入数据库以来，该 Web 页面改变的次数 X 与该 Web 页面的索引文档加入数据库以来系统进行有效性验证次数 n 的比值。根据 Poisson 过程的特点和 Cho 的研究，更新频度可以用下式求出[1]：

$$U_i=-\ln((n-x+0.5)/(n+0.5)) \quad (7)$$

点击率(Click ratio：C_i)简单地讲就是单位系统监测周期内用户对检索出的索引 i 的点击次数 c 与该索引文档 i 检索的次数 h 之比。

$$C_i=c/h \quad (8)$$

检索率和点击率反映了搜索引擎用户对该 Web 页的兴趣度，页面权值反映了 Web 页本身的重要程度，更新频度则反映了 Web 页更新的快慢。显然搜索引擎系统在进行索引数据库的更新时应该首先考虑用户兴趣度高、Page Rank 值大和本身更新快的页面。

为综合考虑每个 Web 页的索引的检索情况(检索率：R)、页面权值(Page Rank：P)、更新情况(更新频度：U)以及

(下转第200页)

由7错成3，或0错成4。同样在译码器初始化时也是非常关键的，可以根据实际情况作适当的修改。对初始化的修改只需对查表逻辑进行修改。

结论 采用部分并行译码结构以及系统码的并行编码方式，从软件仿真和硬件调试来看，设计出的LDPC的性能如下：工作时钟可达28M，译码输出比特率可达3.5Mbits，迭代次数为39次。减少迭代次数到19，译码输出比特率可达到7Mbits，还可以通过提高迭代处理时钟频率来提高最大迭代次数。可分别对20个突发错误和20个随机错误进行纠正。整个电路占用的逻辑单元7510LEs，占用15744ESB位。

参考文献

1 Gallager R G. Low-Density Parity-Check Codes. MLTPress, 1963. Available at:http://justice.mit.edu/people/gallager.html

2 孙韶辉，慕建君，王新梅. 低密度校验码研究及其新进展. 西安电子科技大学学报，2001，28(3)：393～397

3 Oenning T，Moon J. A Low-Density Generator Matrix Interpretation of Parallel Concat-enated Single Bit Parity Codes. IEEE Trans. on Magnetics，2001，37(2)：737～741

4 G. gen：LDPC codes for G. dmt. bis and G. lite. bis. Temporary Document CF-060，STUDY GROUP 4/15，Clearwater，Florida，Jan. 2001

5 Kim S，et al. Parallel VLSI architectures for a class of LDPC codes. Circuits and Systems，2002. In：ISCAS 2002. IEEE Intl. Symposium on，Volume 2，2002. 93～96

6 Zhang T，Parhi K K. VLSI implementation-oriented (3，k)-regular low-density parity-check codes. Signal Processing Systems. In：2001 IEEE Workshop on，Sept. 2001. 25～36

7 Zhang T，Wang Z，Parhi K K. on finite precision implementation of low density parity check cides decier. Circuits and Systems，2001. ISCAS 2001. In：The 2001 IEEE Intl. Symposium on，Volume：4，May 2001. 202～205

8 MacKay D J C. Good error-correcting codes based on very sparse matrices. IEEE Transactions on Information Theory，1999，45：399～431

9 Yeo E. Low Density Parity Check Code (OUTER)Decoder. Available at：http://bwrc.eecs.berkeley.edu/People/Grad_Students/yeo/ee225c/ldpc.htm

10 Parhi K K. VLSI Digital Signal Processing Systems：Design and Implementation. John Wiley & Sons，1999

11 陈宗杰，左孝彪. 纠错编码技术. 人民邮电出版社，1987

（上接第96页）

用户对检索出的文档索引的点击情况（点击率：C），可以将其组织成这一网页的特征向量 $\boldsymbol{V}$：

$$\boldsymbol{V}=(R,P,U,C) \tag{9}$$

不同的网页文档种类对不同的网页特征有不同的权值 W_{ij}，这些权值组成矩阵 $\boldsymbol{W}$，然后计算网页的类型向量[8]：

$$\boldsymbol{T}=\boldsymbol{W} * \boldsymbol{V} \tag{10}$$

最后可以根据 $\boldsymbol{T}$ 来判断网页的类型，需要说明的是：网页的特征权值 W_{ij} 可以经过统计或已有资料确定，一个网页可以同时属于多种类型。

4 基于RPUC的索引库的更新算法描述

根据以上的分析，可以利用以下的算法计算出各网页的基于RPUC的类型向量，然后根据类型向量对网页进行分类，最后结合实际统计的各类网页的变化情况确定各类网页的更新周期。

Step 1 初始化（计算网页的特征向量 v，注意在建立索引数据库时计算出的信息可直接使用，如 P）

Step 2 for(i=0；i<文档索引数；i++)
{计算第 i 个文档索引的类型向量；
switch（第 i 个文档索引的类型向量）
{将索引 i 归入相应的文档索引库的更新队列；}}

Step 3 for(j=0；j<更新队列中的文档索引数；j++)
{更新第 j 个文档索引；}

Step 1中，网页的Page Rank值和更新频度在进行文档索引时可以（或已经）由索引程序直接计算得出，用户的检索率和点击率是随着用户的检索行为动态更新的，所以不会对索引数据库进行有效性验证的时间产生太大的影响。

Step 2中，主要操作是对文档索引类型向量的计算，在此基础上将不同类型向量的索引归入不同的更新队列。

Step 3中，系统针对不同的更新队列，启动不同周期的更新进程。

算法的时间复杂度为：$O(N\log_2 N)$。该算法在充分利用了Web页的索引信息的基础上，结合对用户检索行为的分析，对需要更新的文档索引进行分类，实现索引数据库的智能更新。改进了统一更新策略周期长、单一更新策略可能产生改变频繁而非常重要的网站长期又得不到更新的问题。

结束语 本文提出的搜索引擎中索引数据库的更新算法，综合考虑了用户对文档的兴趣度、文档本身的重要程度以及文档的更新频率，利用文档索引时的相关信息和后天对用户检索时的一些相关信息的统计，确定了文档索引更新的优先队列，既保证了变化的索引信息更新的时效性，也考虑了文档本身的重要性和用户的兴趣度，对索引数据库的更新来说应该是一种有效的方法。

参考文献

1 Cho. Crawling the Web：Discover and Maintenance of Large-Scale Web Data：[PhD. dissertation]. Stanford university，2001

2 Brin S，Page L. The Anatomy of a Large-Scale Hyper textual Web Search Engine. Stanford，CA 94305，USA

3 Kleinberg J. Authoritative sources in a hyperlinked environment. In：Tarjan R E，et al.，eds. Proc. of the 9th ACM-SIAM Symposium on Discrete Algorithms，New Orleans：ACM Press，1997

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

5 陈治平，林亚平. 基于最高响应比的WWW索引库更新方法. 计算机科学，2003，30(5)

6 李盛韬，余智华，程学旗，白硕. Web信息采集研究进展. 计算机科学，2003，30(2)

7 彭洪汇，林作铨. Internet上的搜索引擎和元搜索引擎. 计算机科学，2002，29(9)

8 李剑，金蓓弘. Web链接结构信息研究综述. 计算机科学，2003，30(4)