[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
G06F 17/30 (2006.01)

[12] 发明专利申请公布说明书

[21] 申请号 200610117787.5

[43] 公开日 2007 年 4 月 11 日

[11] 公开号 CN 1945576A

[22] 申请日 2006.10.31

[21] 申请号 200610117787.5

[71] 申请人 上海态格文化传播有限公司

地址 200433 上海市国定路 101 弄 4 号楼 1302 室

[72] 发明人 邱致中 王少刚

[74] 专利代理机构 上海东亚专利商标代理有限公司

代理人 罗习群

权利要求书 1 页 说明书 5 页 附图 2 页

[54] 发明名称

自适应网页更新时间预测方法

[57] 摘要

本发明为一种自适应网页更新时间预测方法，是改进的邻近法。它能根据网页变化的历史规律预测出其下次的更新时间，在没有网页更新频率先验知识的情况下能快速预测出更新频率的量级，并且能迅速地适应网页更新频率的突变。经 MATLAB 仿真，本方法能较准确地预测网页的更新时间，对比经典的邻近法，本方法能在明显减少系统开销的前提下保证所抓取网页的时新性。该方法适用于网页抓取系统，经在一实际系统应用，性能良好。

1、一种自适应网页更新时间预测方法，其特征在于：该方法通过以下步骤实现：

（1） 首先设一组更新间隔序列,其中元素为最小步长 m 乘以步长缩放因子 f 的指数倍,元素之间应相差一定的数量级；

（2） 设定更新时间间隔初值，若能从网页中解析出网页的真实更新时间，则初值为最近三次更新时间间隔的加权平均，否则取一经验值；

（3） 将上一次的更新间隔与更新间隔序列中的元素匹配，即找出与上次更新间隔最邻近的元素,以确定当前更新时间间隔的数量级；

（4） 判断网页有无更新,若更新,则把下次更新间隔收缩 f 陪；若无更新,则放大 f 倍；

（5） 若网页连续若干次未更新,则更新间隔取间隔序列中的下一个元素,即使得更新间隔增大一个量级；若网页连续若干次更新,则更新间隔取间隔序列中的上一个元素,即使得更新间隔减小一个量级；

（6） 若网页由连续多次未更新而转入更新状态,则更新间隔缩小若干个数量级；若网页由连续多次更新而转入未更新状态，则更新间隔增大若干个数量级。

自适应网页更新时间预测方法

技术领域：

本发明涉及互联网信息处理领域，特别是有关于一种网页更新时间预测方法。

背景技术：

互联网中网页信息量的指数速度增长给诸如搜索引擎之类的网络应用系统的信息搜集带来了巨大的压力，一方面，为了保持信息的时新性，必须以尽可能高的频率来抓取网页，以及时获得更新过的网页；另一方面，受硬件资源的限制，要以尽可能低的频率抓取网页，以减少无效的抓取（即抓取到未更新的网页）。网页更新时间预测是解决上述矛盾的关键，它的目的是准确预测网页的更新时间，使得网页抓取器能够以最小的开销获取时新的网页。但由于网页的纷繁复杂，不同网页的更新频率千差万别，如新闻网站的首页可能过几分钟就会更新一次，而另外一些网页则好几个月才更新一次，甚至可能永远不更新。另外绝大多数的网页并不是以一个特定的频率更新的，网页的更新与否往往是网站维护者的主观意志，故网页的更新频率一般无特定的规律。这要求网页更新时间预测方法对变化莫测的网页具有较强的自适应性。

预测网页更新的经典方法是邻近法。所谓邻近法，该方法即为[Knut Magne Risvik, et al., 2002]文中提到的方法，对新搜集到的网页，系统根据属性设置初始的更新时间，如果网页在该时间内更新，则把更新时间减半；反之，则加倍。这种方法的好处是比较简单，缺点是如果设置的初始更新时间与网页的实际下次更新时间相差较大，则邻近法的收敛速度会比较慢，另外，如果网页的更新频率产生突变，邻近法也很难及时地适

应这种突变。

发明内容

为改进邻近法预测网页更新的缺点，本发明提供一种自适应网页更新时间预测方法，该方法通过下列步骤实现：

（一） 首先设一组更新间隔序列,其中元素为最小步长 m（minStep）乘以步长缩放因子 f (factor)的指数倍,元素之间应相差一定的数量级；

（二） 设定更新时间间隔初值，若能从网页中解析出网页的真实更新时间，则初值为最近三次更新时间间隔的加权平均，否则取一经验值，如 30 分钟；

（三） 将上一次的更新间隔与更新间隔序列中的元素匹配,即找出与上次更新间隔最邻近的元素,以确定当前更新时间间隔的数量级；

（四） 判断网页有无更新,若更新,则把下次更新间隔收缩 f 陪；若无更新,则放大 f 倍；

（五） 若网页连续若干次未更新,则更新间隔取间隔序列中的下一个元素,即使得更新间隔增大一个量级；若网页连续若干次更新,则更新间隔取间隔序列中的上一个元素,即使得更新间隔减小一个量级；

（六） 若网页由连续多次未更新而转入更新状态,则更新间隔缩小若干个数量级；若网页由连续多次更新而转入未更新状态,则更新间隔增大若干个数量级。

本发明的优点在于，它能根据网页变化的历史规律预测出其下次的更新时间，经 MATLAB 仿真，本方法能较准确地预测网页的更新时间，对比经典的邻近法，本方法能在明显减少系统开销的前提下保证所抓取网页的时新性。该方法经一实际的网页抓取系统试验，性能良好。

附图说明

图 1 是本发明的流程图。

图 2 是网页抓取系统的工作流程图。

具体实施方式

本方法可用于各种网页抓取系统，如搜索引擎。网页抓取系统通常由三部分组成：网页下载部件、更新检测部件和更新时间预测部件。参照图 2 系统工作流程如下：

（一） 网页下载部件：根据输入的 url，从网上下载网页，将网页分解为在 html 中作为超链接出现的 url 的列表，以及文本型元素体的列表。

（二） 更新检测部件：将新抓取到的网页与具有相关 url 的本地存储的网页进行比对，以检查网页是否更新，检测部件还可能从网页中提取出网页的真实更新时间。

（三） 更新时间预测部件：根据网页的历史更新情况预测网页的下次更新时间，指导网页下载部件在合适的时间对相同网页进行再次下载。根据图 1,更新时间预测部件的具体流程为：

（1） 将上一次的更新间隔与更新间隔序列中的元素匹配，即找出与上次更新间隔最邻近的元素,以确定当前更新时间间隔的数量级。

（2） 判断网页有无更新,若更新,则把下次更新间隔收缩 f 陪；若无更新,则放大 f 倍。

（3） 检查网页的历史更新情况，若网页连续若干次（这里为 2 次）未更新,则更新间隔取间隔序列中的下一个元素,使得更新间隔增大一个量级；若网页连续若干次更新（这里为 2 次），则更新间隔取间隔序列中的上一个元素,使得更新间隔减小一个量级；若网页由连续多次未更新（这里为 5 次）而转入更新状态,则更新间隔缩小若干个数量级；若网页由连续多次（这里为 6 次）更新而转入未更新状态,则更新间隔增大若干个数量级。

实施例：

举 yahoo 社区的一个网页：

http://cn.bbs.yahoo.com/message/read_talkcar_174080.html 为例，这是个 BBS 页，取其前 60 个更新时间序列（这个序列可从网页上直接读出），以序列第一个值为参考，并将该序列转化为秒，则序列为：

0 935 231883 261484 277037 314594 346493 346601
355709 401795 402343 408114 445925 493502 530610
580559 596884 620318 668050 680267 680267 680270
680282 686234 686533 686609 691639 695092 699361
699813 751811 786379 786384 790780 826472 847222
856377 873258 873687 876733 927321 1014280 1018088
1019502 1027354 1047183 1049073 1086272 1086275 1092288
1103902 1128980 1135175 1135295 1137836 1195896 1214459
1223416 1261189 1304231

网页更新时间预测部件的最小步长设为 minStep=100 秒，步长缩放因子为 factor=1.125,更新间隔序列设为：

假设网页下载部件第一次下载到的这个网页是个新帖（还没有回帖），则检测部件不能提取到网页的真实更新时间，这时更新时间预测部件的初值只能取一经验值，在这里为 4334 秒，网页下载部件经过 4334 秒重新下载该页，更新检测部件发现网页已更新（因为 ），于是更新时间预测部件将下次步长收缩 1.125 倍，变为 3852，下载部件过 3852 秒后重新下载该页，经检测部件后发现页面未更新（因为 ），则更新时间预测部件将下次步长放大 1.125 倍，变为 4334 秒 …… 当更新时间预测部件检测到网页连续两次未更新，便将下次更新间隔增大一个量级，变为 11120 秒，于是再隔 11120 秒后网页下载部件重新下载该页，检测部件检测到该页未更新 …… 根据流程得到的预测序列点为：

4334 8186 12520 23640 36150 64681 96779

169984 252340 262224 266558 271434 282554 292438

303558 332089 357450 368570 381080 409611 434972

463504 488865 517396 542757 553877 566387 594918

620279 631399 641284 652404 680935 706296 717416

729926 758457 783818 812349 837711 848830 858715

863049 867924 879044 888929 900049 928580 953941

982472 1055677 1120748 1149280 1174641 1203172 1228533

1239653 1252163 1280694

对比两个序列，发现预测序列能较好的拟合实际序列，这说明了算法的有效性。

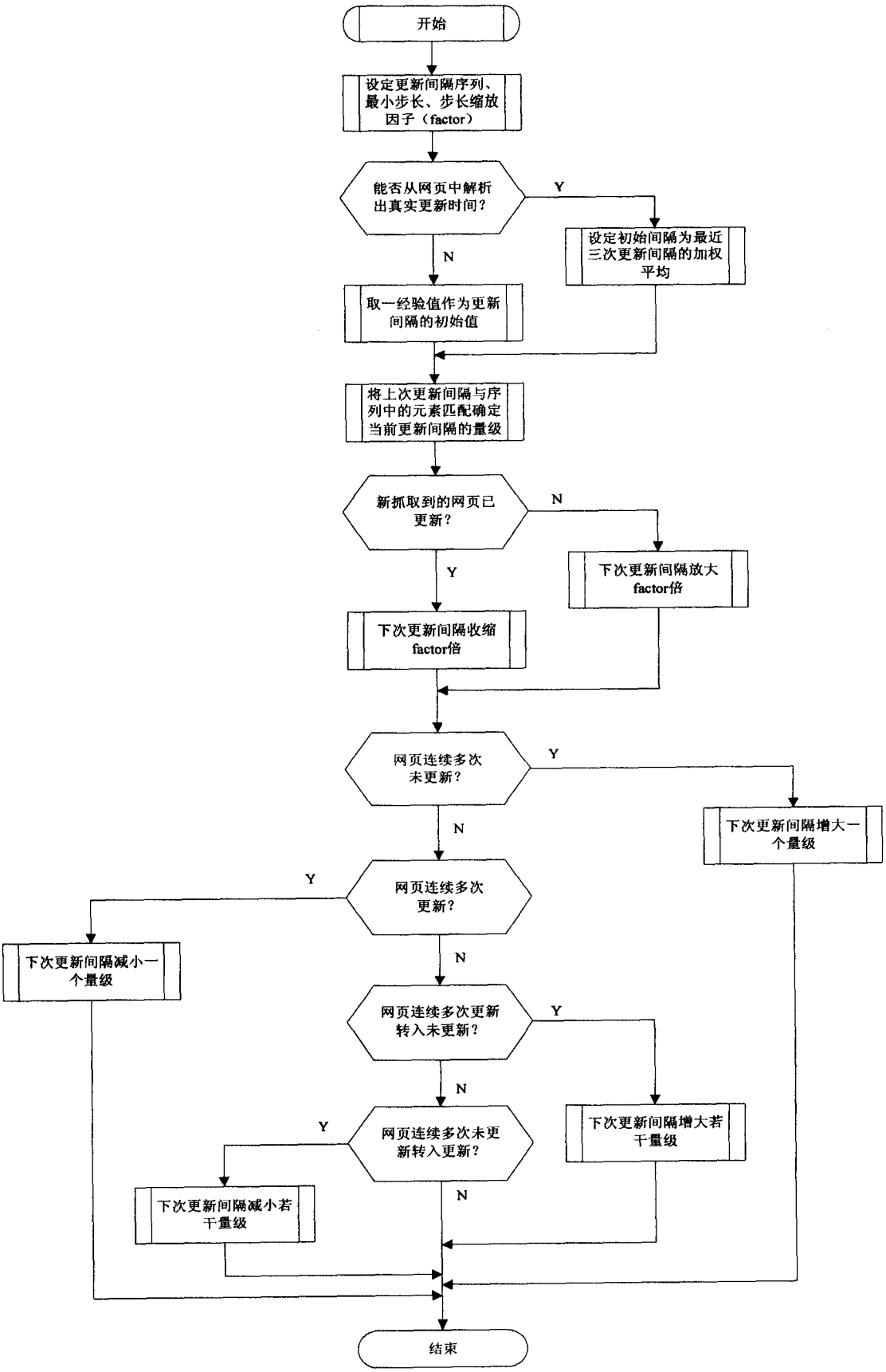
开始
设定更新间隔序列、最小步长、步长缩放因子（factor）
能否从网页中解析出真实更新时间？
Y
设定初始间隔为最近三次更新间隔的加权平均
N
取一经验值作为更新间隔的初始值
将上次更新间隔与序列中的元素匹配确定当前更新间隔的量级
新抓取到的网页已更新？
N
下次更新间隔放大factor倍
Y
下次更新间隔收缩factor倍
网页连续多次未更新？
Y
下次更新间隔增大一个量级
N
网页连续多次更新？
Y
下次更新间隔减小一个量级
N
网页连续多次更新转入未更新？
Y
下次更新间隔增大若干量级
N
网页连续多次未更新转入更新？
Y
下次更新间隔减小若干量级
N
结束

图 1

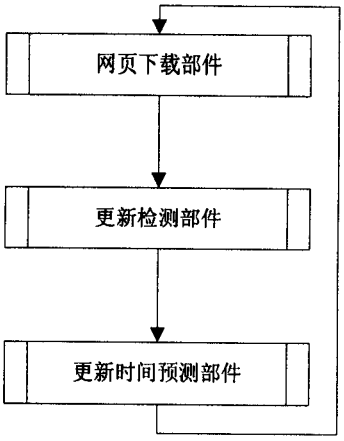

图 2