搜索引擎页面刷新策略研究综述①

Survey of Page-Refreshing Strategies of Search Engine

陈丽君　(浙江大学 计算机学院 浙江 杭州 310027; 浙江越秀外国语学院 浙江 绍兴 312000)

林怀忠　(浙江大学 计算机学院 浙江 杭州 310027)

摘　要： 根据判断信息来源的不同，对现有页面刷新策略进行了分类，系统地分析了它们各自的特点，指出了各种策略各自存在的优点与不足。对未来的研究方向进行了展望，给出了若干值得研究的问题。

关键词： 搜索引擎 网络爬虫 页面刷新 更新

Web 页面的快速增长，使搜索引擎成为人们获取信息的主要工具。搜索引擎依靠网络爬虫从 Internet 上采集页面，经处理后存于索引库中供用户查询。一方面，由于 Web 信息资源的动态独立更新(update)，索引库中存储的内容随着时间不断“老化”，需要网络爬虫重新采集，刷新(refresh)现有页面，提高索引库的时新性。另一方面，由于“无限”的页面数量和有限的带宽资源，网络爬虫难以及时刷新所有更新页面，降低了搜索引擎的服务质量。为此，研究人员提出了各种搜索引擎页面刷新策略，优先刷新有更新的、重要的页面，保证索引库的“时新性”和“重要性”。

页面刷新策略研究的主要任务是：如何快速总结出网页更新规律，准确预测页面更新，及时刷新发生变化的页面，使索引库中的内容与各站点的实际内容尽可能保持一致。从分析的数据来源看，主要有以下几种思路：一是由站点提供页面更新信息；二是通过分析历史信息来预测页面的更新；三是通过采样样本页进行分析预测。本文先介绍这三种页面刷新策略的基本思想和主要步骤，再分析它们各自的特点与优缺点，最后指出各种策略适用的场合，并展望了未来的研究方向。

1　基于站点信息的策略

刷新页面最有效、最简单的方法是直接从 web 站点获取网页更新信息。由于 Web 服务器可以记录页面最后修改时间等更新相关信息，Brandman 等人[1]提出由站点搜集更新元数据的刷新方法，即让 Web 服务器周期性地扫描各个网页的更新时间，连同网页地址一起导出到元数据文件，并将这些文件按一定规则组织存放到服务器指定位置。搜索引擎根据实际需求读取元数据文件，判断网页的更新日期，仅刷新那些最后更新时间大于最后刷新时间的页面，从而避免因下载未更新页面引起的带宽浪费。实现此方法只需对服务器做简单扩充，而且可以通过服务器提供的参数填充表单，达到访问动态页面的目的。但其最大的缺点是需要能支持各种元数据格式的网络爬虫，这实际上很难做到。Xu 等人[2]提出采用与用户(内容供应商)合作的方式，可以解决以上弊端。其基本思想是：根据用户提供的更新信息，以网页访问次数为权重，以未能及时刷新有更新页面带来的负面影响计算优先值，根据该值高低确定刷新先后次序。

具体做法可分为以下三步：

1) 确定用户类型。根据用户提供信息的不同分为三类：未提供任何页面更新信息的普通用户(Normal user)，提供页面历史更新信息的高级用户(Smart user)，与既提供页面历史更新信息又提供将来更新概率(P(update))的超级用户(Super user)。

2) 计算优先值。以自最后一次刷新以来，访问到过期页的次数与过期页持续时间的乘积计算。用 t(LC)表示最后刷新时间，t(now)表示当前时间，t(m)表示

① 基金项目:浙江省科技计划(2007C23086)
收稿时间:2008-08-06

自 t(LC)以来第 m 次更新，N(t)表示累计到 t 时刻客户访问到过期页的次数。普通用户优先值的计算公式为 S1=(t(now)-t(LC))*(N(t(now))+c)/2，高级用户的值为 $S2=\sum_{m=0}^{n}(t(m)-t(m-1))*N(t(m))$，而对于超级用户，还要根据历次访问值 C(d)来估计将来的访问概率 $P(f)=P(update)\times\frac{\alpha*C(p-1)+(1-\alpha)*\sum_{d=p-s}^{p-2}C(d)}{\sum_{d=p-s}^{p-1}C(d)}$，结果得 S3=β*S2+(1-β)*P(f)。

3) 确定刷新次序。按优先值降序排列待刷新页面。对于已经更新的页面，直接选择优先值最高的页面刷新；对于即将更新的页面，在更新时刻到来后的第一时间内刷新。

此类方法算法简单、高效，且与具体的用户访问模型无关，同时还考虑了页面的访问流行度，能为用户提供优质的查询服务。但其最致命的弱点是：有赖于用户提供的信息及信息的准确性，是一种被动的刷新方式，当用户不能够提供这些信息时，它的高效性就得不到体现。而且根据实验经验，超级用户数量的比例宜控制在 1/3 到 2/3 之间，否则会影响索引时新性。

2 基于更新历史的策略

网页的更新是随机的，通常被视为是泊松过程，用其为网页建立更新模型，可以估计网页的更新周期与下次更新时间。以这种“根据历史推测将来”的思想为指导，研究者提出了各种基于历史信息的刷新策略。

2.1 用户访问

考虑到搜索引擎主要是为用户提供查询服务，用户对查询结果的满意程度是搜索引擎追求的目标，因此，有学者从查询用户的角度来分析索引质量。比如，虽然索引中存在着过期的页面，但只要它未被用户点击访问，则仍然认为该页面是时新的。

文献[7]中刷新策略的指导思想是：在满足有限带宽资源条件下，降低索引平均过期程度，减少用户“尴尬”(用户打开某个页面，结果却发现该页面并不符合查询要求)的次数。为此，Wolf 等人将网页被用户从索引中访问的概率作为网页重要性权值，同时引入标值点过程，为网页建立更新模型。标值点过程的优势在于它能够用标值空间记录重要的更新信息，如页面更新的概率、更新给搜索引擎带来的影响等。求解过程包括寻找最优频率和调度两个阶段。第一阶段，用概率理论定义网页的平均过期程度，将网页的刷新看作资源分配问题，求出理论上最优的刷新次数和刷新时机。第二阶段，以第一阶段的输出为输入，基于网络流理论，将调度问题映射为运输问题，寻找最优的调度策略。此方法直接运用了资源分配算法上的研究成果，而且算式不仅仅局限于泊松过程，有着更广泛的应用场合。但此方法①输入过于庞大，求解代价较高，不适合在线计算，②只考虑发生“尴尬”的情况，忽略了高质量、查询相关的页面。

Pandey and Olston[8]提出了以用户为中心(user-centric_的刷新策略。其基本思想是：由过去刷新一个页面对索引质量(主要是排序位置)带来影响的平均值，以及最近一次刷新的时间，来确定刷新的优先次序，目标是既要减少用户访问到过期页的次数，又要保证提供优质的页面。

具体实现方法如下：

1) 将查询日志中与页面 p 有关的查询集合 S 中的多词条查询转换成多个单词条查询，得新的查询集合 S'；

2) 对每个查询 q∈S'，根据更新索引时记录的 p 刷新前后的得分 s1 和 s2，计算 p 对于查询 q 的排序位置 r1 和 r2；

3) 计算其他由于 p 的变化而引起的位置调整的页面的得分值；

4) 用 2)和 3)的结果估计索引库质量的变化 △Q(p)，其主要体现为，页面 p 刷新前后，在相关查询 q 返回结果中排列位置的改变，从而引起用户对这些页面(包括 p 和其他页面)访问概率的变化；

5) 将 p 的各个历史△Q(p)值求平均，得 δQ(p)；

6) 计算刷新优先值 P(p,t)=δQ(p)*(t-LR(p,t))，其中 LR(p,t)表示最近一次刷新。

以用户为中心的刷新策略结合了用户访问习惯与

查询历史，最先刷新与查询相关度高的页面，保证了用户的查询质量，体现了“用户至上”的服务思想。同时结合最近刷新时间，防止页面长时间得不到刷新。但是，此方法没有考虑到用户的查询频度，也即那些用户最近关心的热点话题，没有给予特别重视，也没有考虑新增页面。

2.2 更新频率

在页面刷新策略研究中，用得最多的方法是根据网页更新频率计算刷新频率，通过分析两者之间的关系，达到合理分配资源的目的。

文献[3]根据更新频率将网页分组，同一个组中网页的更新频率相近。将整个搜集过程分成一个个搜集周期，每个周期可以分配不同的权重，在每个周期的最后估计每组过期网页数量。目标是在满足总带宽限制的前提下，使已经过期但未刷新的网页，和已发现但未搜集的网页数最少。此方法用不同的桶存放不同更新频率的页面，因此不需要任何关于页面变化率的理论模型。但该方法存在以下局限：①复杂度随时间增长，因此需要周期性地复位才得以继续[4]；②由于太多的系数和约束，使得这个非线性约束条件下的非线性目标优化问题非常难解[12]；③没有考虑新闻等快速变化的页面。

文献[4]中，Cho and Garcia-Molina 用泊松过程为网页建立更新模型，以提高平均时新性或降低平均年龄为目标。先在假设固定更新频率、固定资源分配的前提下，比较研究固定、随机和纯随机三种刷新次序，结果表明，用固定的刷新次序能获得最大的时新性；接着在假设不同更新频率、固定刷新次序的前提下，比较研究固定资源分配和按比例资源分配策略，结果表明，固定资源分配策略能取得更大的时新性；最后用拉格朗日多项式得到了更新频率与刷新频率之间的最优关系图，即更新频率与刷新频率之间存在着一定的关系，但这种关系并不是线性的。文献[5]将网页权重设置为与更新频率成正比，将页面的刷新看作轮询过程，从理论角度分析如何使过期页面的比例最少。文献[6]研究在信息不完整的情况下，如何更好地估计网页的更新频率。

以网页更新频率为依据的刷新策略，方法比较直观，算法容易理解。其存在的主要问题：一是页面的更新频率①在初始状态的估计值并不可靠；②很难在短期内准确估计；③本身可能也是在不断变化；二是只用 0 和 1 表示页面更新与否，没能反映出页面变化程度的高低。

2.3 信息效用

考虑到 Web 页面往往包含着各种不同的信息块，如导航栏、广告条、相关链接、正文等，各块的生存周期不尽相同，广告条等与主题不太相关的块往往变化得比较快，可能没来得及刷新就又发生了变化，而正文等相对比较稳定。Olston and Pandey[9]尝试结合页面变化程度与速度，计算刷新一个页面的效用(utility)来决定该页面是否需要刷新。即在给定刷新代价的前提下，仅刷新那些效用达到指定阀值 T 的页面，最终使页面的平均差异程度最小。

具体来说，包括以下过程：

1) 根据实际资源情况，确定效用阀值 T，即在刷新页面与下载新页面之间合理分配资源；

2) 将页面文档看作连续的词序列，用 shingling 技术将文档分段，使其成为信息块的集合；

3) 每次刷新时，用变化文档(change profile)记录页面的刷新时间及相对变化程度(t,D(P(tB),P(t)))，这样的文档至多保留 h 个；

4) 结合各个变化文档中的历史记录，计算下一次刷新时间 φp，有两种方法可以选择：

① 曲线拟合法：将 h 个变化文档中的各个相对变化程度值，按相对于各自基准时间 tB 的差值排列，相对时间相同的取变化程度的平均值；根据各时间点的变化程度值拟合出一条连续的曲线，观测其变化方式是更接近于 churn(短期的替换更新)，还是更接近于 scroll(周期地添加更新)，最后根据其变化方式预算出效用刚好为 T 时的时刻 t，即 U(t)=T，作为下次刷新时间 φp。其中 $U(t)=t*D(t)-\int_0^t D(x)dx$

② 基于上下边界法：首先确定每个变化文档变化程度曲线的上下边界，再分别计算出 h 个文档上边界和下边界的平均值，根据这两个平均值分别求出效用的最大和最小值，即 Umax 和 Umin，最后根据 Umax 和 Umin

与阀值 T 的关系就可以确定下一个更新时刻 ϕp。

与以往的方法相比，根据信息效用的刷新策略：① 将页面内容分块，按块衡量变化程度，而不是把页面看作一个整体，直接忽略所有变化快的页面；② 更倾向于关注长期性内容的刷新，节约了资源；③ 只需根据实际资源调节效用阀值 T，而不依赖于全局优化，适用于大规模的并行爬行；④ 使用变化文档记录页面信息，占用的空间少；⑤ 考虑到对新增页面的刷新。但此策略需要累积到一定数量的文档时才能对更新作出比较准确的判断。

3 基于采样预测的策略

3.1 站点采样

基于历史信息策略的特点是：1) 需要保存大量的历史轨迹，存储需求量大；2) 为得到有效的信息，需要长时间的搜集，这在实际的应用中是不允许的。有研究表明，Web 信息资源的变化呈现出一定的规律性，如商业类网站的变化速度比较快，教育、政府类网站要慢得多。因此，Cho and Ntoulas[10]提出了基于站点采样的策略，它通过比较各站点采样页面的变化率，将资源分配给变化率高的站点。该基于贪婪算法的策略主要包括以下过程：

1) 根据可用资源情况确定采样粒度 k 和采样置信度 α；

2) 从各个网站采样 k 个网页，来估计各站点的变化率 ρi，再结合 α 计算出最高与最低变化率 hi 与 li；

3) 根据变化率 ρi 的分布，计算变化率阀值 ρt；

4) 将各站点的 hi 与 li 和 ρt 比较，如果 $hi < \rho t$，则放弃对该站点的刷新；如果 $li > \rho t$，则刷新该站点的所有网页；

5) 重复 2)~4)直到所有资源耗尽。

此策略方法简单快捷、容易实现，不需要保存历史信息。实验数据表明，与基于更新频率的策略相比，此策略判断的准确率接近于前者的两倍，尤其是在开始一段时间内。其关键在于采样，如果采样不够均匀，将导致错误的资源分配，从而大大降低刷新效率。

3.2 分组采样

考虑到基于站点采样的粒度比较粗糙，影响判断的准确性。也就是说，同一站点页面的更新频率可能相差很大，而相近更新频率的页面又会分散于不同的站点。因此 Tan 等人[11]提出了分组采样的策略，其基本思想的依据是：①网页的变化频率与某些网页特征密切相关；②特征相似的网页，在变化规律上也表现出相似性。

其具体方法描述如下：

1) 抽取与更新模型相关的静态与动态特征，包括网页内容、URL 长度、链接 PR 值、图像数等，并将它们作为网页的特征向量来量化网页；

2) 将寻找具有相同更新模型网页问题看作聚类问题，用 RBC(Repeated Bisection Clustering)算法将网页分组；

3) 从各组中选取靠近组质心的典型页面作为样本采样，用泊松过程建立网页更新模型，以它们的历史平均更新频率来预测整个组下次的更新概率 ϕ；

4) 各组按 ϕ 值降序排列，优先下载 ϕ 值大的页面，直到资源耗尽。

分组采样策略将更新频率相近的页面组合到一起，巧妙地结合了历史更新频率与采样各自的优点，提高了刷新效率和判断的准确性。同时对于新发现的页面可以根据其特征很方便地加入到对应的组，追随该组的刷新规律。关键是特征的选取，它将直接决定着刷新的效率。

4 结束语

不同的页面刷新策略有不同的特点，适用的场合也不尽相同。本文第一种策略的前提是需要有内容供应商的合作，比较适合于大型网站；第二种策略在能够正确估计网页更新规律并收集到一定历史信息时，能取得很好的刷新效果，比较适合用在后期的刷新过程和变化较有规律的网站；第三种策略的关键在于采样的样本是否具有代表性，比较适合于变化分布均匀的网站。

本质上搜索引擎的页面刷新就是根据网页的某些属性，分析预测其将来的变化趋势。因此，如何提高预测的准确性仍然是未来重点研究的问题，如在文献[11]中考虑结合多种策略的思想，实现优势互

补。网页的变化错综复杂，变化规律本身也不固定，单种策略往往难于适应，如何适时地根据实情动态调整刷新策略，提高自适应性，也是个有价值的研究方向。目前大多数研究都把页面的变化近似为泊松过程，但它并不适用于新闻类等快速变化的页面，而这样的页面实际占有相当的比例，因此有必要增进非泊松过程模型的研究。另外，由于页面数量的快速增长，算法的效率，以及对新增页面的刷新都有待进一步研究。

随着人们对搜索引擎依赖程度的日渐增长，搜索引擎的服务质量、索引时新性备受人们关注，如何准确、高效、灵活地刷新页面仍然是个值得研究的课题。本文对现有刷新策略进行分类，系统分析、比较它们的优缺点，指出它们各自适用的场合，希望能为今后的工作提供基础性的支持。

参考文献

1 Brandman O, Cho J, Garcia-Molina H, Shivakumar N. Crawler-Friendly web servers. In PAWS'00, 2000.

2 Xu J, Li QL,Qu HM, Labrinidis A.Towards a Content-Provider-Friendly Web page crawler. Proc. of the 10th International Workshop on Web and Databases, 2007.

3 Edwards J, McCurley K, Tomlin J. An adaptive model for optimizing performance of an incremental Web crawler. Proc. of the 10th International Conf. On World Wide Web, 2001.

4 Cho J, Garcia-Molina H. Effective page refresh policies for Web crawlers. ACM Trans. on Database Systems, 2003.

5 Jr Coffman EG, Liu Z, Weber RR. Optimal robot scheduling for Web search engines. Journal of Scheduling, 1998.

6 Cho J, Garcia-Molina H. Estimating frequency of change. ACM Trans. on Internet Technology, 2003.

7 Wolf JL, Squillante MS, Yu PS, Sethuraman J, Ozsen L. Optimal crawling strategies for Web search engines. Proc. of the 11th World Wide Web Conf., 2002.

8 Pandey S, Olston C. User-Centric Web crawling. Proc. of the 14th International Conf. on World Wide Web, 2005.

9 Olston C, Pandey S. Recrawl scheduling based on information longevity. Proc. of the 17th International Conf. on World Wide Web,2008.

10 Cho J, Ntoulas A. Effective change detection using sampling. Proc. of the 28th International Conf. on Very Large Database, 2002.

11 Tan QZ, Zhuang ZM, Mitra P, Giles CL. A Clustering-Based sampling approach for refreshing search engine's database. Proc. of the 10th International Workshop on Web and Databases, 2007.

12 孟涛,王继民,闫宏飞. 网页变化与增量收集技术. 软件学报, 2006,17(5):1051－1067.