搜索引擎中信息动态采集策略的研究

高 凯
(河北科技大学信息科学与工程学院计算机科学与技术系,河北石家庄 050054)

摘 要: 为了能及时采集到有关网页信息,搜索引擎应根据相应网站及其更新速度,动态调整其信息采集的频度.本文就模型化网页更新过程以及根据相关性动态调整搜索引擎的信息采集频度进行了探讨.一方面使用泊松过程来描述网页更新并分析了搜索引擎如何有效完成信息采集;另一方面采用基于网页从属关系和内容分析的相关性来调节该过程,使得在进行信息采集与数据更新时的针对性更强.实验表明了该方法的有效性.

关键词: 搜索引擎; 数据下载器; 网页更新; 泊松过程; 相关性
中图分类号: TP393 文献标识码: A 文章编号: 0372-2112 (2007) 10-1984-05

Dynamic Refresh Strategy for Crawler in Search Engine

GAO Kai
(*Department of Computer Science and Technology, School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang, Hebei 050054, China*)

Abstract: As for a search engine, keeping up with the evolving Web is necessary. We concern about modeling on an effective Web page collecting policy and propose an adaptive refresh strategy based on the relevance, which is used to adjust the process. On one hand, we think the refresh behavior follows the properties of the Poisson process and analyze the strategy on how to crawl the Web effectively. Further, the relevance is on the basis of the affiliation detecting and the contents analysis. It is used to adjust the process. This makes the process more targeted. The experimental results validate the feasibility of the approach.

Key words: search engine; crawler; refresh; Poisson process; relevance

1 引言

随着计算机网络的迅速普及和应用, Internet 已成为人类的信息宝库,如何有效利用这个信息宝库正日益受到人们的重视.在此应用背景下搜索引擎应运而生.据中国互联网络信息中心 2003 年 7 月至 2006 年 7 月发布的七次统计报告显示[2],在用户经常使用的网络功能中选择搜索引擎的比例分别是 70%、61.6%、64.4%、65%、64.5%、65.7%、66.3%;在用户得知新网站的主要途径中选择搜索引擎的比例分别是 85%、83.4%、86.9%、86.6%、84.5% (注: 2006 年后的调查报告中无此项统计数据),可见搜索引擎正在日益发挥着重要的作用.但同时用户对搜索引擎性能感到非常满意的却只有 23.4%、27.4%、26.9%、28.4% (注: 2005 年 7 月后的调查报告中无此项统计数据).英国 MORI 调查公司的调查统计结果也表明只有 18% 的用户对搜索引擎的返回结果表示满意,而高达 68% 的用户表示很失望,可见搜索引擎仍有许多需改进之处.据中国互联网络信息中心 2005 年 7 月发布的统计报告显示,用户在回答"检索信息时遇到的最大问题"时选择"信息更新慢"选项的占 27.5%,排名第 2 位.因此本文拟针对信息采集中的数据更新问题进行研究.本文认为追求绝对的快不如有的放矢地进行更新,这样既能有效节省网络资源又可保持对相对重要及相关内容的及时下载与更新.

目前网页数量增长十分迅速而且内容更新频繁.虽然人们并不奢望一小时前发生的新闻事件能够马上出现在报纸上,但却希望通过搜索引擎在 Internet 上找到相关新闻.另一方面,由于网络资源的动态变化,搜索引擎链接到的页面有时会变得不可访问.据统计目前搜索引擎链接的失效页面数量大约占全部链接页面数量的 2%~9%[1],可见如不及时有效地进行信息更新势必会影响到搜索引擎的整体性能.但由于不同网站间的更新频度差异很大[2],且更新大多存在着随机性,随时跟踪并完全做到对网页的实时更新几乎也是不现实的.因此一些搜索引擎系统往往根据实际情况采取不同的定时或不定时更新策略.

在相关研究中,文献[6, 10, 12~14]介绍了多种不同的更新方法,而网页相关性问题也在文献[3~5]中被

收稿日期: 2005-08-04; 修回日期: 2007-07-10

提及, 但未涉及如何针对不同网页采取不同的信息采集策略. 本文认为, 如果页面发生了更新, 数据采集器应能及时探查到并更新本地数据, 同时还应对那些相关性较大的网站予以特别关注. 本文通过基于泊松过程的分析来模型化网页更新事件并据此进行信息采集频度的调整. 在基于本研究工作的新闻与教育资源搜索引擎系统中, 由于一些网站对于用户来说具有较大的相关性, 应予特别关注(这在网络资源不足时尤为必要), 因此本文提出基于网页相关性来调整信息采集的方法. 采用上述策略后, 系统会根据当前网站更新情况及其相关性来动态调整针对相应网站的信息采集频度. 实验分析与实用均表明该方法的有效性与合理性, 信息采集与更新效果良好.

2　相关工作

基于文档内容的更新方法已广泛用于数字图书馆和在线交易系统中. 文献[8]借助数学模型描述了文档内容的变化规律, 但并未考虑网页链接结构的变化; 文献[6]将网页抽象为一颗树型结构, 但该方法不便于处理那些非流行页面的更新. 另一方面, 在基于网页链接的更新处理中, 文献[12]表明不同领域的网页更新频度是不同的; 文献[14]提出 9 种度量网页更新的标准; 文献[7]提出网页生存期的概念, 但未考虑网页本身的相关性和实际更新的频度. 而针对网页相关性问题, 诸多文献亦对此进行了研究: 鱼群算法[3]与鲨鱼算法[4]认为相关网页周围的链接页面也是相关的; PageRank 算法亦有不足; 而在 PageRank 算法基础上衍生出的 Topic-Distillation 算法也只在某些特定情况下效果较好[11]. 和上述工作相比, 本文提出的更新频度调整策略可使更新较接近于实际, 而基于相关性的调整通过网页隶属分析和内容分析, 从权威站点与公众关注热点两个角度入手, 可使得更新时能突出重点.

3　更新行为分析及更新频度调整

虽说不能期望数据下载器总能实时探查到网页的每次更新, 但了解网页更新的统计分布规律可帮助设计人员设计出更加有效和合理的信息采集与数据更新算法. 虽然从理论上说可让数据下载器时刻不停地频繁对某网站进行访问, 但这样做可能会违背网络礼仪, 因此搜索引擎应能根据相应网站更新频度的不同动态调整对其进行信息采集的频度, 如适当缩短或扩大访问时间间隔. 这里仅以缩短时间间隔为例进行分析.

本文认为网页更新事件可用泊松过程来近似描述, 因为它基本符合泊松过程定义的四个条件: 在初始时刻无更新事件发生(满足零初值性); 满足平稳增量性; 在任意多个不相重叠的时间间隔内出现的更新事件相互独立(满足独立增量过程); 在足够短的时间内同时出现两个以上更新的事件为小概率事件. 因此, 网页更新可近似地用泊松过程来描述. 当然, 这只是在特定时间段的一种近似, 例如对白天和夜晚来说, "平稳增量性"就会打折扣, 而按照有关文献[8, 10]的统计, 网页在白天和夜晚的更新频度是有差异的, 这也和我们对国内新闻类门户网站实际的更新情况的统计相接近. 由于在本文所述系统中访问的种子集来自国内新闻网站, 因此本文讨论的更新策略也主要局限在白天.

假设搜索引擎当前的信息采集时间间隔(频度)是常数 T, T_m ($m = 1, 2, \cdots$) 为开始采集时的各时刻点. 如近期对应网站更新较为频繁, 就应缩短这个信息采集时间间隔, 因此就需在 (T_m, T_{m+1}) 之间再插入一个采集时间点 t 并在该时刻追加一次信息采集, 这样就会得到两个新的时间区段 (T_m, t) 和 (t, T_{m+1}). 那么, 变量 t 应取何值时才能使搜索引擎尽量做到对网页的及时采集呢? 显见, 这里应使平均总等待时间为最短. 下面进行分析.

首先, 计算在 (T_m, t) 中出现了更新事件的页面的总的等待被搜索引擎采集到的期望时间. 假定在 (T_m, t) 中总共有 $N(r)$ 个更新事件发生. 设第 1 个更新事件在 W_1 时刻出现, 则其等待时间为 $(W_1 - T_m)$, 而第 i 个 $(i \geqslant 2)$ 更新事件的等待时间为 $(W_i - T_m)$ $(T_m \leqslant W_i \leqslant t)$, 则总的等待被采集到的时间为 $\sum_{i=1}^{N(r)} (W_i - T_m)$, 因而所求的平均总等待时间就可以表示为 $E[\sum_{i=1}^{N(r)} (W_i - T_m)]$. 因 $N(r) = n$, W_i $(i = 1, 2, \cdots, n)$ 的联合分布与 $[T_m, t]$ 上均匀分布的 n 个独立随机变量 Y_i $(i = 1, 2, \cdots, n)$ 导出的一组顺序统计量 Y_i^* $(i = 1, 2, \cdots, n)$ 的联合分布相同, 于是有:

$$
\begin{aligned}
E\Big[\sum_{i=1}^{N(r)} (W_i - T_m)\Big] &= E\Big\{E\Big[\sum_{i=1}^{n} (W_i - T_m) \mid N(r) = n\Big]\Big\} \\
&= E\Big\{E\Big[\sum_{i=1}^{n} W_i \mid N(r) = n\Big] - nT_m\Big\} \\
&= E\Big\{E\Big[\sum_{i=1}^{n} Y_i^*\Big] - nT_m\Big\} \\
&= E\Big\{n \cdot \frac{t - T_m}{2} - nT_m\Big\} \\
&= \frac{\lambda}{2}(t - 3T_m)(t - T_m)
\end{aligned}
$$

故有:

$$
E\Big[\sum_{i=1}^{N(r)} (W_i - T_m)\Big] = \frac{\lambda}{2}(t - 3T_m)(t - T_m)
$$

接下来, 计算在 (t, T_{m+1}) 中出现了更新事件的页面的总的等待被搜索引擎采集到的期望时间. 假定在 (t, T_{m+1}) 中总共有 $N(q)$ 个更新事件发生, 设第 1 个更

新事件在 S_1 时刻出现, 因此其等待时间为$(S_1 - t)$, 而第 i 个$(i \geqslant 2)$更新事件的等待时间为$(S_i - t)$ $(t \leqslant S_i \leqslant T_{m+1})$, 则总的等待被采集到的时间为 $\sum_{i=1}^{N(q)}(S_i - t)$. 和上述分析类似, 有:

$$\begin{aligned} E[\sum_{i=1}^{N(q)}(S_i - t)] &= E\{E[\sum_{i=1}^{n}(S_i - t) \mid N(q) = n]\} \\ &= E\{E[\sum_{i=1}^{n} S_i \mid N(q) = n] - n \cdot t\} \\ &= E\{E[\sum_{i=1}^{n} Y_i^*] - n \cdot t\} \\ &= E\{n \cdot \frac{T_{m+1} - t}{2} - n \cdot t\} \\ &= \frac{\lambda}{2}(T_{m+1} - 3t)(T_{m+1} - t) \\ &= \frac{\lambda}{2}(T_m + T - 3t)(T_m + T - t) \end{aligned}$$

故有:

$$E[\sum_{i=1}^{N(q)}(S_i - t)] = \frac{\lambda}{2}(T_m + T - 3t)(T_m + T - t)$$

因此, 需求出变量 t 使得:$f(t) = \frac{1}{2}(t - 3T_m)(t - T_m) + \frac{1}{2}(T_m + T - 3t)(T_m + T - t)$ 取最小值(即总的等待时间尽可能短). 为此, 令$f'(t) = 4t - 4T_m - 2T = 0$. 易知, 当 t 取唯一驻点$t = T_m + \frac{T}{2}$时,$f(t)$取最小值.

假定搜索引擎在时间点 $T_m(m = 1, 2, \cdots)$进行网页采集与新网页百分比检测. 假设在 T_m 时测得的新网页百分比大于设定的经验阈值v_1(注: 有关新网页百分比的计算参见后续公式(1). 限于项目本身的原因, 经验阈值的确定暂不便列出), 那么就在接下来的时间区间中间追加一次网页采集; 假设在 T_m 时测得的新网页百分比连续两次大于设定的经验阈值v_1, 就缩短这个网页采集时间间隔 T. 在本文所述搜索引擎系统中, 对于绝大部分新闻类门户网站而言, 若页面发生了更新(即发布了新的新闻), 页面中必定会出现新的 URL 来指向该新闻正文. 参照文献[14], 新网页百分比计算方法见公式(1).

$$\text{新网页百分比} = \frac{\text{给定时段内探查到的新出现的 URL 数量}}{\text{给定时段中探查到的所有 URL 的数量}} \quad (1)$$

4 基于网页相关性的更新调整

基于网页相关性的更新调整是建立在网页隶属分析和网页内容分析两个方面, 并通过相关概率来调节信息采集过程的. 这里以教育资讯搜索引擎为例进行说明. 在网页隶属关系分析中, 由于用户常会对那些来自于诸如教育行政主管部门等"权威"站点给予较多的关注, 因此通过调查确定了一部分"权威"教育资讯站点及相应的主页面权值(略), 主页面权值被赋予那些和相应主页面有隶属关系的从属页面; 在网页内容分析中, 如果网页内容包含用户近期常用的流行查询词(可通过对用户查询日志的分析得到), 则包含这些"热点话题"的网页被认为是相对重要的, 如果它们频繁出现在网页的标题、锚文本等处, 则对应网页的相关性一般较高, 它们将会获得较高的信息采集频度. 公式(2)为网页 U_i 的相关性的一种可能的经验计算方法, 设它有 j 个 Tag 标记(标题、锚文本等), 如果 U_i 和权威站点间有从属关系或热点查询词出现在上述 Tag 标记处, 则其相关性较高. 信息采集频度的调整是通过相关概率来实现的, 即通过将相关性转换为对应的相关概率来对相应网站进行更新(限于篇幅, 略). 本文认为引入相关概率后更新仍服从泊松过程(限于篇幅, 证明略), 因此前述的信息采集频度调整仍是可行和有效的. 需要说明的是, 这里要设定更新时间上下限阈值. 上限阈值的设置是为了防止对有些更新较慢的网页长期得不到更新的情况出现, 下限阈值的设定是由于在短时间内对一个网站频繁采集网页可能会引起的一些网络礼仪问题.

$$\text{相关性}(U_i) = \{\sum_{j=1}^{n}[\text{权值}(tag_j) \times tag_j \text{ 上热点查询词数目}]\} \times \text{主页面权值}(U_i) \quad (2)$$

5 试验结果与分析

在上海交通大学纳讯高新技术应用研究所有限公司的协助下, 本文使用前述两个搜索引擎系统来下载相关的教育资讯与热点新闻作为数据集. 图 1 是一个时期不同类别网站更新情况. 在此统计期间, 热点新闻网站每天至少要更新 60% 的内容, 而在教育中却不超过 5%; 热点新闻网站一周后内容要全部更新, 而大学和教育网站中则分别是 60% 和 20%. 因此采用有针对性的动态信息采集策略很有必要.

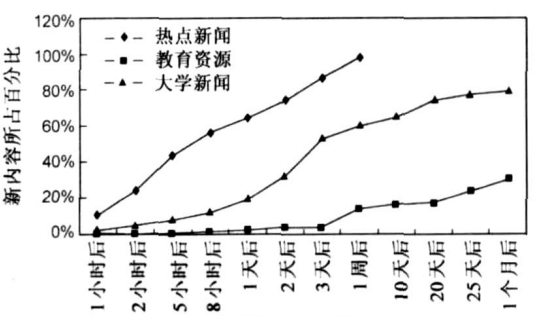

图 1 不同类型的网页更新速度比较

本文以数据集中的几个新闻网站为例进行了实验. 作者抽取某时段这些新闻门户网站中各类别新闻中所有链出页面的 URL 并监控其变化情况, 如本次检

测到有 90% 的 URL 和上一次的发生了冲突(限于篇幅, 此处不介绍该算法的设计), 说明该主页面有 10% 的内容发生了更新, 换句话说新网页百分比为 10%. 通过对一个时期内网页更新情况的监测, 得到这些新闻类站点每 10 分钟的平均新网页百分比数据. 按照在区间$[s, t](s<t)$上的增量过程近似服从参数为 $\lambda(t-s)$ 的泊松分布的假设, 可以计算得到各类新闻的更新概率分布. 图 2 是新闻网站更新情况, 从中可看出各类别新闻更新间隔在 20 分钟左右的居多, 而实际的网站更新情况统计与上述计算得到的结果是基本相符的, 针对这几个新闻网站中各类新闻平均的实际更新情况的统计见图 3. 可见, 按照泊松分布计算得出的更新和实际统计情况是基本接近的. 需要指出的是, 对于在观测期间内上述领域的网页, 用泊松过程建模是比较好的, 但其他领域是否适用, 尚有待进一步研究.

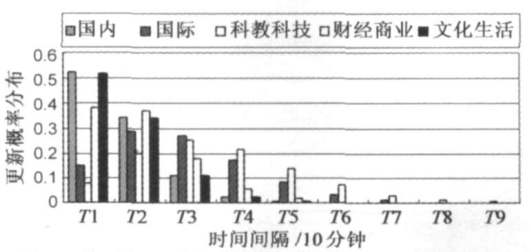

图 2　按泊松分布计算的各类别新闻的更新分布

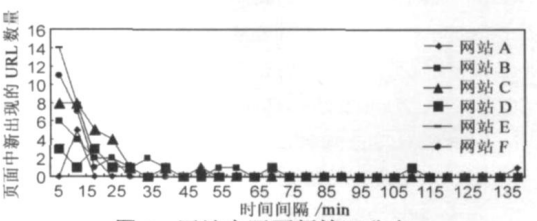

图 3　网站实际更新情况分布

下面对系统资源使用情况进行分析. 未使用动态采集算法前, 系统中以并发方式工作的各数据下载器对网页的采集是在固定时刻集中开始的. 姑且不论该方式是否能够做到对数据的及时更新, 它也易对系统资源造成短时间的集中冲击, 不利于对资源的合理使用. 而动态采集是根据网页自身更新情况选用适当的更新初始值, 并在系统运行中根据具体情况动态调整之, 这样可尽量减少对系统资源的集中冲击. 实际使用

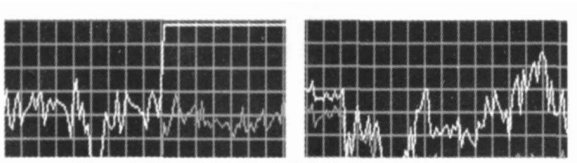

图 4　不同更新方法对系统资源的使用情况

也验证了这点. 作者通过对服务器 CPU 及相关资源使用情况的实际监测, 发现这种错峰使用资源的方法会尽量减少对系统资源的集中冲击, 能更加有效地合理使用资源. 图 4 是采用两种不同采集策略时, 单服务器 CPU 资源使用情况的对比. 从图中可见, 采用固定更新时(左图) 易对系统资源造成冲击, 而采用动态采集策略时(右图) 情况就要好些. 需要说明的是, 采用上述措施后, 也不是说能完全避免图 4 左图中的情况, 只不过这种情况较少发生而已.

最后, 对更新效果进行分析. 这里以新闻搜索引擎为例进行分析说明. 原有搜索引擎每隔 M 分钟(注: 工程经验值, 限于项目本身原因, 此处暂隐去其具体值) 开始对相应网站进行信息采集. 采用上述算法后, 系统会根据当前网站变化情况及其相关性来动态调整针对相应网站的信息采集频度. 为了测试信息采集与更新效果, 作者在搜索引擎上选取当日更新较快且相关性较大的网站进行了跟踪测试, 并选择了部分采用固定更新算法的网站与采用动态采集算法的网站进行了网页采集与数据更新时间的比较. 数据量方面, 数据下载器对国内 30 个主要新闻门户网站中的 9 个类别中的网页进行信息采集(见表 1). 统计时, 采用固定更新周期的方法采集一批网页, 计算它们的平均滞后时间(即采集到该新闻的时间与其发布时间的时间差), 并将其和采用动态采集方法采集到的网页的平均滞后时间进行了对比. 统计表明, 对于更新较快的新闻, 采用动态采集算法后, 采集到相应新闻的时间平均比原先要快(针对该测试期内采集的网页来说, 大致快 10% 以上), 提高了时效性. 但需指出的是, 如果前一段时间该网站变化较慢, 则未来针对它的采集周期可能就要相对长一点, 因此不能说动态采集方法肯定比采用固定周期方法能更快地下载到所需网页, 它只是能根据实际情况动态调节信息采集的频度.

表 1　数据集

数据类别	种子集	数据类别	种子集
国内新闻	35	科技新闻	18
国际新闻	22	体育新闻	24
财经新闻	124	社会新闻	23
军事新闻	8	娱乐新闻	19
文教新闻	13		

不可否认的是, 在多数情况下系统并没有在源信息发布的同时就实时地采集到相关网页, 这主要是受如下因素的影响: 首先, 本文所述搜索引擎不仅需要进行数据下载, 还要进行网页信息标引、自动摘要、相似新闻聚类等数据处理工作, 为了兼顾各方工作, 这里采用的机制是通过定时器定时切换启动网页下载与数据处理工作, 而不是一直在进行网页下载; 第二, 采集器在采集网页时要遍历诸多页面的 URL 序列流, 并在经过网页去重判断后完成信息采集, 这些都需要一定的时间, 而本文所介绍的动态调整方法只是调整采集器

针对该网站信息采集的开始时刻; 最后, 算法中还设定了更新时间的下限,这些因素都会带来上述影响.

不可否认的是, 并非所有站点的更新事件都严格满足泊松过程的条件. 对于非平稳泊松过程来说, 仍可采用相应的分析方法近似之, 例如可根据最优化理论采取黄金分割法($T_m + 0.618T$) 等更新策略, 具体策略可参阅相关文献, 限于篇幅在此不再赘述; 而对一些变化很慢或很有规律的网站采用定时更新的策略效果较好, 需针对具体情况进行具体处理. 在实际项目中也是对种子集中的部分网站采用了上述动态更新算法. 另外, 上述方法虽已在具体的工程实践中应用, 并且也已通过了项目验收, 但从算法性能的分析上来说, 尚缺少和其他文献中同类方法的比较. 结合下一个科研课题的开发, 除上述比较外, 计划进一步分析"新网页百分比"和"网页采集间隔"的关系, 以期能进一步优化信息采集过程.

6 结论

搜索引擎中的数据下载器应根据网页实际更新情况动态采集并更新本地数据, 这样既能有效节省网络资源又可保持对相对重要及相关内容的及时下载与更新. 本文提出一种网页更新策略. 如果页面发生了更新, 网页自动采集器应能及时探查到并及时下载网页及更新本地数据, 同时对那些相关性较大的网站予以特别关注. 基于泊松过程分析的更新调整策略可使信息采集更接近于实际, 而基于相关性的调整通过网页隶属和内容分析, 从权威站点与公众关注热点两个角度入手, 使得更新时更能突出重点. 实验分析与用户实用均表明该方法可有效地在尽量少占用资源的情况下保证对相关网页的及时更新.

致谢 向对本文提供帮助与支持的上海交通大学纳讯应用技术研究所有限公司的王永成老师及公司工作人员表示衷心谢意.

参考文献:

[1] Ricardo Baeza-Yates, Berthier Ribeiro-Neto. Modern Information Retrieval[M]. USA: Addison Wesley & ACM Press, 1999. 367- 395.

[2] http://www.cnnic.net.cn/index/0E/00/11/index.htm [OL]. 2007.

[3] P M E De Bra, R D J Post. Searching for arbitrary information in the www: the fish search for mosaic[A]. Proceedings of the 2nd World Wide Web Conference[C]. USA, 1994.

[4] M Hersovici, M Jacobi, Y S Maarek, D Pelleg, M Shtahaim, S Ur. The shark-search algorithm. an application: tailored web site mapping[A]. Proceedings of 7th World Wide Web Conference [C]. Australia, 1998. 317- 326.

[5] Y S Maarek, M Jacobi, M Shtalhaim, S U D Zernik, I Z Ben Shaul. WebCutter: a system for dynamic and tailorable site mapping[A]. Proceedings of 6th World Wide Web Conference [C]. USA, 1997. 713- 722.

[6] S Flesca, E Masciari. Efficient and effective web change detection[J]. Journal of Data & Knowledge Engineering, 2003, 46(2): 203- 224.

[7] Jinghoo Cho. Crawling the Web Discovery and Maintenance of Large-Scale Web Data [D]. Dissertation for the PhD of Stanford University, 2001. 57- 70.

[8] Brian E Brewington, George Cybenko. How dynamic is the web [J]. Journal of Computer Network, 2000, 33(1- 6): 257- 276.

[9] Alexandros Ntoulas, Junghoo Cho, Christopher Olston. What's new on the web the evolution of the web from a search engine perspective[A]. Proceedings of the 13th World Wide Web Conference[C]. USA: ACM Press, 2004. 1- 12.

[10] Brian E Brewington, George Cybenko. Keeping up with the changing web[J]. Computer, 2000, 33(5): 52- 58.

[11] http://www.cs.toronto.edu/~georgem/hilltop[OL]. 2004.

[12] Dennis Fetterly, Mark Manasse, Marc Najork, Janet Wiener. A large-scale study of the evolution of web pages[A]. Proceedings of the 12th World Wide Web Conference[C]. Hungary: ACM Press, 2003. 669- 678.

[13] Jinghoo Cho, Hector Garcia-Molina. Effective page refresh policies for web crawlers[J]. ACM Transactions on Database Systems, 2003, 28(4): 390- 426.

[14] Sung Jin Kim, Sang Ho Lee. An empirical study on the change of web pages[A]. Web Technologies Research and Development APWeb 2005[C]. Heidelberg: Springer Berlin, Volume 3399, 2005. 632- 642.

作者简介:

高 凯 男, 1968 年 11 月出生于天津, 1990 年、2002 年和 2007 年分别在河北科技大学、燕山大学和上海交通大学获工学学士、硕士和博士学位. 现主要从事网络信息智能处理、信息检索、数据挖掘等方面的研究与开发工作.

E-mail: gaokai@sjtu.edu.cn