基于用户访问模式的 WWW 浏览路径优化*

阳小华[1]，周龙骧[2]

[1](中南工学院 计算机系，湖南 衡阳 421001)；

[2](中国科学院 数学研究所 计算机科学研究室，北京 100080)

E-mail: xhyang@ctit.zhnut.edu.cn

摘要：分析了 WWW 用户的浏览活动规律，提出了有关 WWW 浏览路径优化的一些基本概念，设计了一个基于用户访问模式的浏览路径优化算法，并与相关的工作进行了比较.

关键词：万维网；浏览路径；浏览过程；访问模式；路径优化

中图法分类号：TP393 文献标识码：A

浏览和查询是获取 WWW 信息的主要手段. 由于 WWW 信息的提供者常用超链把相关的文档链接在一起，用户通常需要对特定的文档集合进行浏览以获得所需的信息，因此，从某种意义上说，浏览是获取 WWW 信息的基本手段，查询的作用只是为浏览提供适当的起点.

一般地说，WWW 文档之间的超链链接是由信息资源的提供者预先确定的，用户只能在已有的浏览路径中进行选择. 由于不同的用户有不同的目标，同一个用户在不同的时期也会有不同的需求，而在 Web 站点设计时难以预计可能的用户群及其特征，因此预先设计的浏览路径——也就是 WWW 系统结构——常常不能与实际情况很好地吻合，需要根据用户的实际访问规律动态地予以调整、优化.

一个 WWW 站点结构设计的好坏可以简单地以用户为获取所需信息所付出的平均代价作为衡量的标准，而这种代价则可以理解为所经过的超链数目和选择这些超链的困难程度的函数. 我们所说的浏览路径优化就是在不破坏 WWW 系统原有结构——即不删除系统原有文档和超链的前提下，通过增加新的超链或文档来减少用户获取信息所需付出的平均代价.

本文首先分析了 WWW 用户的浏览活动规律，提出了有关 WWW 浏览路径优化的一些基本概念；然后，设计了一个基于用户访问模式的浏览路径优化算法，进行了算法分析；最后，讨论了相关的工作、应用和发展.

1 基本概念

信息获取是 WWW 用户的基本活动，因而一个 Web 用户的活动历史可以用所访问过的文档按时间顺序组成的序列来表示. 由于在 Web 文档之间存在着超链链接，因此，WWW 用户活动的基本特征就是浏览. 用户通过查询或其他手段来获得浏览的起点，然后沿着超链对特定的文档集合进行浏览. 为了描述 Web 用户的这种行为特征，我们引入下面的定义[1].

定义 1. 设 $D=\langle d_1,d_2,\ldots,d_n,\ldots\rangle$ 是一个 Web 用户的活动记录，$S=\langle d_i,\ldots,d_j\rangle$ 是 D 的某

* 收稿日期：1999-10-26；修改日期：2000-03-20

基金项目：国家 863 高科技发展计划资助项目(863-306-02-07-1)；国家部委级重点项目基金(7A.3.1-2)

作者简介：阳小华(1963－)，男，湖南衡阳人，博士，副教授，主要研究领域为 WWW 信息的获取，数据库技术；周龙骧(1938－)，男，浙江温州人，研究员，博士生导师，主要研究领域为数据库技术，电子商务.

个子序列,称 S 为该用户的第 $k(1\leqslant k)$ 次浏览过程,如果

(1) 对 S 中的任一元素 $d_q(i<q\leqslant j)$,如果不存在满足条件 $d_p=d_q$ 的元素 $d_p(i\leqslant p<q)$,那么在 d_{q-1} 中一定有指向 d_q 的超链;

(2) 不存在元素满足条件 $d_p=d_{j-1}$ 的 $d_p(i\leqslant p\leqslant j)$ 并且在 d_j 中不包含指向的 d_{j+1} 超链;

(3) 在 S 之前已经存在 $k-1$ 次浏览过程,并且在第 $k-1$ 次浏览过程中没有与 d_i 相同的元素,在 d_{i-1} 中也不包含指向的 d_i 超链 $(k>1,i>1)$.

浏览过程的第1个元素称为浏览起点.显然,用户第1次浏览过程的起点是文档 d_1.一个非起点文档在一次浏览过程中第1次出现意味着用户点击了前一个文档中指向该文档的超链,而当文档在浏览过程中重复出现时,就不一定对应于超链点击,也可能相应于诸如回退(backward)之类的操作.从定义1可以看出,从浏览起点到浏览过程的其他元素之间存在着由用户点击过的超链所组成的有向路径.

定义2. 设 S 是一个Web用户的某一次浏览过程,我们称以 S 中的元素为结点,以这些元素之间的超链为边构成的有向图为该用户的一个浏览历史区域.

一次浏览过程通常对应于用户一个特定的信息需求,用户在特定的浏览历史区域中漫游的目的通常是为了获取特定类型或主题的信息资源.显而易见,在同一个浏览历史区域中,不同的文档和超链对于用户当前的需求而言,其重要程度通常是不一样的.一般来说,WWW资源的重要性可以用它们被使用的次数或频率来度量——频率越高越重要.但是,由于WWW的网状结构,在浏览的过程中用户有时不得不进行回溯.比如说,为了访问当前文档的兄弟,就常常需要回溯到共同的祖先文档.显然,这种回溯所带来的高频率并不能代表一个文档的重要性,不能表示用户真正的兴趣所在.基于这种认识,我们引入下面的定义.

定义3. 设 $S=\langle d_1,\ldots,d_n\rangle$ 是一个用户浏览历史区域中从浏览起点 (d_1) 开始的有向路径,

(1) 如果 S 中的每一个元素都是不同的文档,则称 S 为一条正向浏览路径;

(2) 正向浏览路径中任意一个连续的子串也是正向浏览路径;

(3) 不被同一用户浏览历史区域中其他正向浏览路径真包含的正向浏览路径称为极大的正向浏览路径.

一条正向浏览路径对应于用户一段信息探求的过程.一般来说,用户之所以会沿着正向浏览路径不断地深入,是因为在路径文档中含有所需的信息.在正向路径上的浏览行为体现了用户真正的需求特征和活动规律,在下面的讨论中,我们将只考虑正向浏览路径.

定义4. 称一条正向浏览路径为常见浏览路径,如果它属于足够多个不同的用户浏览历史区域.

常见浏览路径体现了特定的用户群在特定时期内的活动规律,也就是浏览模式.值得一提的是,在考虑常见浏览路径时,我们忽略了在同一次浏览过程中浏览路径重复出现的问题.因为一次浏览过程对应的只是一个用户在一个特定时期内的行为,而真正起作用的是用户群体在较长时期内稳定的行为模式.

2 路径优化

因为Web站点的路径结构是由站点管理员预先设计的,所以,对于特定的用户浏览过程而言,有些文档并不是由于它们的内容而是由于它们所处的位置而被访问的.有些文档虽然没有包含用户当前所需的信息,但是由于它们位于特定的路径上,为了从浏览起点到达特定的位置不得不经过

它们.因为信息获取的代价与所经过的浏览路径长度成正比,因此,这些位于路径中间就不必要的文档无疑增加了用户获取信息的代价.当然,一个 Web 站点的任务是向广大的用户群提供长期稳定的信息服务,不能保证每个用户的每次浏览过程都能以最小的代价获得所需的资源.但是,在一定时期内,如果大量的用户以及大量的浏览过程都体现出了共同的特征,也都付出了相同的不必要的代价,那么就应该对已有的路径结构进行动态优化,以提高当前站点信息获取的整体效率.

根据定义 4,一条常见浏览路径就是在足够多个不同的用户浏览历史区域重复出现的正向浏览路径.由于位于浏览路径中间的文档有可能不是用户真正的兴趣所在,它们被访问的原因可能只是由于所处的位置而不是所包含的内容,而常见浏览路径则体现了特定用户群在一段时期内的活动特征,因此,如果在常见浏览路径的起点中设计一条直接指向终点的超链,就可以显著地减小特定用户群获取 Web 站点信息的整体代价.这就是我们进行浏览路径优化的基本思想.下面,我们来讨论浏览路径优化的实现算法.

在 WWW 服务器上通常都保存有用户访问的日志,用户的每一次信息请求对应于日志中的一条记录,记录的内容通常包括发出请求的源(IP 地址)、被请求信息资源的 URL 和发出请求的时间.一般来说,来自相同 IP 地址的请求可以认为是由同一个用户发出的.也就是说,每一个 IP 地址对应于一个用户.因此,我们可以根据 IP 地址,从用户访问的日志中得到各个用户的当前活动历史记录.然后,根据定义 1,可以把每个用户的活动历史记录分解为一系列的浏览过程,从而发现常见的浏览路径,实现 Web 站点路径的动态优化.路径优化算法的具体步骤如下:

(1) 根据 IP 地址,从 Web 站点当前的用户访问日志中分解出每个用户的原始活动历史记录,保存在 USER0 中;每个用户对应于一条形式为(IP,UHR0)的记录,其中 UHR0 是由被请求信息资源的 URL 按信息请求发出的时间先后顺序所组成的序列;

(2) 根据定义 1,对 USER0 中各个用户的原始活动历史记录进行分解,得到各个用户由浏览过程构成的活动历史记录,保存在 USER1 中;每个用户对应于一条形式为(IP,UHR1)的记录,其中 UHR1 是由用户的浏览过程按时间顺序构成的序列;

(3) 根据定义 3,对 USER1 中的各个元素进行处理,找出每个浏览过程中的极大正向浏览路径,并统计其出现的次数;处理的结果保存在 DMF 中;每一条极大正向浏览路径对应于一条形式为(MF,NF)的记录,其中 MF 是由 URL 组成的序列,表示极大正向浏览路径,NF 则是 MF 出现的次数;

(4) 对 DMF 中的各个元素进行处理,找出极大正向浏览路径中所有的常见浏览路径,也就是出现次数超过预设阈值的正向浏览路径;处理的结果保存在 DFP 中;每一条常见浏览路径对应于一条形式为(FP,NP)的记录,其中 FP 是由 URL 组成的序列,表示常见浏览路径,NP 则是 FP 出现的次数;

(5) 对 DFP 中的每一个元素 X,进行以下处理:

① 对 DFP 中每一个其他的元素 Y,检查 $X.FP$ 是不是 $Y.FP$ 的子序列;如果是,则 $X.NP=X.NP-Y.NP$;

② 如果 $X.NP\geqslant$预设阈值,则从路径 $X.FP$ 的起点引出一条直接指向路径终点的超链.

从 Web 站点当前的用户访问日志中分解出每个用户的原始活动历史记录是一项比较简单的工作,只需对日志进行一次扫描即可.但从每个用户的原始活动历史记录中分解出浏览过程则要稍微复杂一些,需要抽取和利用文档中的超链信息.为了得到极大正向浏览路径,首先要构造浏览历史区域,然后只要对浏览历史区域进行一次图遍历即可.为了找出所有的常见浏览路径,需要对

DMF 进行多次扫描,具体的做法可参见文献[2].

值得注意的是,因为常见浏览路径的子路径也一定是常见浏览路径,所以并不是所有的常见浏览路径都应该设置捷径.如果一条浏览路径之所以成为常见浏览路径,完全是因为它是其他常见浏览路径的子路径的缘故,那么我们就不应该设置从起点到终点的捷径.因此,在路径优化算法的第(5)步,我们考虑了由于子路径所带来的影响.只有那些在排除了子路径所带来的影响之后,其访问次数仍然超过阈值的常见浏览路径,才是真正独立的常见浏览路径,我们才为之设置捷径.

3 小 结

浏览路径优化主要是 Web 服务器结构的优化,文献[3]在这方面进行了许多有益的探讨和研究工作.Mike Perkowitz 和 Oren Etzioni 探讨了 Web 站点结构优化的一些问题,并给出了一个通过自动合成 Web 索引页来优化服务器结构的方法,他们的主要做法是:(1) 处理用户访问日志,把它分解为一系列的"访问".每一个"访问"对应于一天的历史记录.(2) 计算 Web 页之间共同出现的频率,构造相应的相似矩阵.(3) 构造相应于相似矩阵的图,利用簇聚技术挖掘出图中的"集团"(cliques),位于同一"集团"中的 Web 页将具有较高的共同出现的频率.(4) 对于每一个"集团",构造一个包含指向"集团"内每一个文档的超链的 Web 索引页.

与我们的方法相比,文献[3]对于用户活动历史的分段仅仅是基于时间,而并不是对应于用户的信息需求,未能体现出 WWW 用户活动的浏览特征.另外,在文献[3]中,自动生成的 Web 索引页应该放在站点的什么位置也是一个未能得到满意解决的问题.

除了 Web 服务器以外,浏览路径的优化也可以在网关上进行.一个特定的网关对应于一个特定的用户群,这些用户只能通过这个网关访问 WWW.由于网关通常也保存有用户的访问日志,而且在网关上通常还有一个保存着用户近期所获资源的缓冲区,因此我们可以在网关上进行浏览路径的优化.

值得一提的是,由于我们依据的是用户当前的活动规律,浏览路径优化是动态的,需要定期或不定期地重复进行.因此,对于那些当前已过时的捷径——也就是从前所设置的超链,需要适当地进行清理.

References:

[1] Yang, Xiao-hua, Zhou, Long-xiang. The view of Web users. Journal of Software, 1999,10(7):690~693 (in Chinese).

[2] Chen, Ming-syan, Jong Soo Park, Yu, P. S., *et al*. Data mining for path traversal patterns in a Web environment. In: Kavanugh, M. E., ed. IEEE Proceedings of the 16th International Conference on Distributed Computing Systems (ICDCS). Hong Kong: IEEE Computer Society Press, 1996. 385~392.

[3] Perkowitz, M., Etzioni, O. Adaptive Web sites: automatically synthesizing Web pages. In: Mostow, J., Charles, R., eds. Proceedings of the 15th National Conference on Artificial Intelligence and the 10th Innovative Applications of Artificial Intelligence Conference. Madison, Wisconsin: AAAI Press, 1998. 727~732.

附中文参考文献:

[1] 阳小华,周龙骧.Web 用户的视图.软件学报,1999,10(7):690~693.

Optimization of WWW Navigation Path Based on User Access Patterns*

YANG Xiao-hua[1], ZHOU Long-xiang[2]

[1](*Department of Computer Science, Central-South Institute of Technology, Hengyang* 421001, *China*);

[2](*Department of Computer Science, Institute of Mathematics, The Chinese Academy of Sciences, Beijing* 100080, *China*)

E-mail: xhyang@ctit.zhnut.edu.cn

Abstract: In this paper, the authors analyze the activities of WWW users and present a series of concepts about optimization of WWW navigating path. Furthermore, an algorithm of optimization of WWW navigating path is given, that is based on user access pattern and other related works are compared with.

Key words: world wide web; navigation path; navigation process; access patterns; path optimization

* Received October 26, 1999; accepted March 20, 2000

Supported by the National High Technology Development Program of China under Grant No. 863-306-02-07-1; the Ministry & Commision-Level Research Foundation of China under Grant No. 7A. 3.1-2