基于页面内容和站点结构的页面聚类挖掘算法*

杨怡玲, 管旭东, 尤晋元

(上海交通大学 计算机科学与工程系 分布计算技术中心,上海 200030)
E-mail: yang-yl@cs.sjtu.edu.cn; guan-xd@cs.sjtu.edu.cn; you-jy@cs.sjtu.edu.cn.ac.cn
http://dctc.sjtu.edu.cn

摘要: 提出了结合站点拓扑结构和Web页面内容的页面聚类改进算法,改进算法引入Web页面的内容链接比和页组的组内链接度,并修改了频繁访问页组支持度的计算公式,以此来提高挖掘结果的兴趣性.通过实验数据的比较,改进算法较一般算法的收敛性好,发现的频繁访问页组的兴趣性高.

关 键 词: Web日志挖掘;日志分析;页面聚类;频繁访问页组
中图法分类号: TP311 文献标识码: A

Web日志挖掘(Web usage mining)[1~4]是指将数据挖掘技术应用于Web服务器日志文件,以发现隐藏在其中的用户访问模式.它主要包括数据预处理和挖掘算法实施两个主要阶段.实施挖掘算法之前要对Web日志文件进行预处理,将其转化为用户会话集.关于数据预处理的具体内容可参见文献[1],此处不再赘述.本文着重讨论Web日志挖掘中的页面聚类技术,即通过分析日志文件,发现经常被用户一起访问的页面,即寻找用户频繁访问页组.

目前已有的日志挖掘系统,如WEBMINER[3]和SpeedTracer[4]等,它们对日志进行深层的分析,包括关联规则、序列模式和页面聚类等.但是,这些挖掘系统的结果并不令人满意,其原因在于,它们仅分析了日志数据,而没有结合用户请求的Web页面的内容和站点的拓扑结构.

通常利用兴趣性评价Web日志挖掘的结果[2],在本文中认为一个兴趣性高的用户频繁访问页组满足3点:(1) 被大量用户访问;(2) 包含尽可能多的内容页;(3) 页组内页面之间包含尽可能少的超链接.本文的工作就是围绕如何提高用户频繁访问页组的兴趣性进行的.

1 页面聚类算法的改进

1.1 范化内容链接比(NCLR)

定义 **1.** 内容链接比(content-link ratio,简称CLR)是指一个Web页面的大小与页面中的链接数之比.

如果一个页面的大小为4KB,其中包含3个链接,那么该页面的CLR为4/3(假定页面的大小以KB为单位).页面的CLR值是大于0小于无穷的数.为了便于设定阈值,将CLR映射为0到1之间的数,该值称为范化内容链接比.

定义 **2.** 范化内容链接比(normalized content-link ratio,简称NCLR)是利用计算公式 $NCLR=(1-e^{-CLR})$ 将CLR映射为0到1之间的浮点数.

如果一个页面的内容少但包含的链接数多,则其NCLR相对较小,反之亦然.对于页面中没有指向其他页面

* 收稿日期: 2000-03-14; 修改日期: 2000-08-14
基金项目: 上海市科技发展基金资助项目(985115035)
作者简介: 杨怡玲(1973－),女,山西太原人,博士生,主要研究领域为数据挖掘,分布移动计算;管旭东(1976－),男,江苏常熟人,博士生,主要研究领域为分布移动计算,数据挖掘;尤晋元(1939－),男,江苏常州人,教授,博士生导师,主要研究领域为操作系统,分布移动计算,构件与协调技术.

的链接的特殊情况,上面的两个定义是无效的,不失一般性,我们将该页的 NCLR 置为 1.

1.2 组内链接度(GILD)

一个页组 G 中页面的超链接关系是一个有向图 $Graph(G)$,有向图的节点是页组中的页面,有向图的边对应页面之间的链接.组内链接度用于刻画组内页面间的链接紧密程度.

定义 **3.** 组内链接度(group inter-link degree,简称 GILD)定义为

$$GILD(G)=\begin{cases}|Graph(G)|/(|G|*(|G|-1)) & |G|>1\\ 0 & |G|=1\end{cases},$$

其中$|Graph(G)|$为有向图 $Graph(G)$中的边数,$|G|$为页组 G 中的页面数.

当页组内的任意两个页面之间都没有链接时,则其 GILD 为 0;反之,页组内的任意两个页面都是相互链接的,则其 GILD 为 1,按照本文对兴趣性的定义,这样的页组就没有必要出现在挖掘结果中.

1.3 改进算法

在传统的页面聚类算法中[4],支持度指包含页组中所有页面的用户会话的个数.在改进算法中,我们将支持度的计算进行了扩展,一个页组 G 的支持度为

$$Support(G)=O(G)*H_{\mathrm{NCLR}}(G)*(1-GILD(G)).$$

其中 $O(G)$为包含 G 中所有页面的用户会话的数目(即传统算法中 support 的定义),$H_{NCLR}(G)$为 G 中所有页面 NCLR 的调和平均值,$GILD(G)$为 G 的页组链接度.

假设 FG_k 是包含 k 个页面的频繁访问页组的集合,其中每个页组的支持度都大于预先设定的阈值 T.发现频繁访问页组是一个递归的过程,首先将 FG_1 初始化为支持度大于 T 的页面,FG_2 是在 FG_1 的基础上产生,FG_3 又是在 FG_2 的基础上产生,依此类推.

挖掘频繁访问页组的改进算法如下:

```
1. count the NCLR of all distinct pages appeared;
2. initialize FG_1 as the top requested single page groups with Support>=T;
3. for (i=2;i<=k;i++) {
4.     Sort the pages of groups in FG_{i-1} in lexicographical order;
5.     for each group {x_1,…,x_{i-1}} in FG_{i-1} {
6.         for each group {y_1,…,y_{i-1}} in FG_{i-1} {
7.             if (x_2=y_1 and …and x_{i-1}=y_{i-2}) {
8.                 construct a new group G={x_1,…,x_{i-1},y_{i-1}};
9.                 if (G not already in CG_i) {
10.                    test all other combinations of subgroups of G with size (i-1);
11.                    if (all such subgroups are in FG_{i-1})
12.                        if (Support(G)>=T ) add G into FG_i;
13. } } } } }
```

由于引入内容链接比和组内链接度,改进的页面聚类算法的计算量要比一般算法大,但是对于内容链接比以及页面间的链接关系的计算可以在预处理阶段完成,且可保留直到下一次站点内容的更新,在一定程度上减少了程序 1 中的计算量.另一方面,改进算法可尽早地过滤掉兴趣性低的页组,使得后面迭代过程中的数据量提前减小,快速收敛(参见第 2 节).

2 实验结果

改进的页面聚类算法在上海交通大学 Web 服务器(http://www.sjtu.edu.cn)的日志上进行了验证,试验环境是一台 Pentium Ⅲ 450MHz 处理器和 128M 内存的机器,运行平台为 Windows 98.实验数据为 9MB 的日志文件,其中包含 10 万条记录.日志数据中有 417 个不同的 HTML 页面,从中识别出 1 902 个用户会话.将一般的页

面聚类算法($Support(G)=O(G)$)和改进后的算法对实验数据进行挖掘,得到的 FG_i 的数目见表 1.

Table 1 Comparison of the experimental results of the normal and the enhanced algorithms

表 **1** 一般的页面聚类算法和改进的算法实验数据比较

Algorithm used①	T	$\|FG_1\|$	$\|FG_2\|$	$\|FG_3\|$	$\|FG_4\|$	$\|FG_5\|$	$\|FG_6\|$	$\|FG_7\|$
Normal②	50	29	87	126	97	36	5^{*}	0
Enhanced③	30	29	16	1^{+}	0	0	0	0
	15	49	71	24	1^{++}	0	0	0
	9	68	128	109	36	2^{+++}	0	0

①所使用的算法,②一般算法,③改进算法.

表 1 中$|FG_i|$表示 FG_i 包含页组的个数.一般算法结果中标识为*的 5 个页组包括:(1)人们不感兴趣的 3 个页面——首页("/")、中文版的首页("/chinese/index.htm")和一个导航页("/chinese/Navigate.htm");(2) 人们较少感兴趣的其他 3 个页面.而改进算法结果中标识为+,++,+++的页组中均包含了兴趣较高的页面.同时,通过分析迭代过程中$|FG_i|$数目的变化可以发现,改进算法可以使迭代过程很快收敛,提高了计算过程的效率.

3 结 论

本文提出一个兼顾站点拓扑结构和页面内容的改进算法,使得我们在挖掘过程中更注重内容页之间的关系.在改进算法中,页组的支持度不仅与用户会话有关,而且与 Web 页面的内容链接比和页组的组内链接度有关.通过实验证明,利用改进的页面聚类算法得到的挖掘结果有显著改善.

References:

[1] Cooley, R., Srivastava, J. Data preparation for mining World Wide Web browsing patterns. Journal of Knowledge and Information Systems, 1999,1(1):5～32.

[2] Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P. The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM, 1996,39(11):27～34.

[3] Mobasher, B., Jain, N., Han, E.H., *et al*. Web mining: pattern discovery and from World Wide Web transactions. Technical Report, 96-050, University of Minnesota, 1996.

[4] Wu, K.L., Yu, P.S., Ballman, A. SpeedTracer: a web usage mining and analysis tool. IBM System Journal, 1998,37(1):89～105.

Mining the Page Clustering Based on the Content of Web Pages and the Site Topology*

YANG Yi-ling, GUAN Xu-dong, YOU Jin-yuan

(*Distributed Computing Technology Center, Department of Computer Science and Engineering, Shanghai Jiaotong University, Shanghai* 200030, *China*)

E-mail: yang-yl@cs.sjtu.edu.cn; guan-xd@cs.sjtu.edu.cn; you-jy@cs.sjtu.edu.cn.ac.cn

http://dctc.sjtu.edu.cn

Abstract: In this paper, an enhanced algorithm is proposed for page clustering, which considers both the content of web pages and the site topology. By introducing the content-link ratio and the group inter-link degree and modifying the computation of the support of frequently visited page group, the algorithm can increase the interestingness of the mining result. The experimental results show that the algorithm converges more rapidly and could find out more interesting page groups than the normal algorithm.

Key words: Web log mining; log analysis; page clustering; frequently visited page group

* Received March 14, 2000; accepted August 14, 2000

Supported by the Science and Technology Development Foundation of Shanghai of China under Grant No.985115035