一种基于页面价值和跳转偏爱度挖掘频繁访问路径的模型①

李爱飞, 冀振燕, 王经纬

(北京交通大学 软件学院, 北京 100044)

摘 要: 设计实现了一种从 Web 日志挖掘用户频繁访问路径的模型. 提出网页聚类分析的一个重要基础理论, 以及页面价值和跳转偏爱度的概念, 并建立页面价值模型. 该模型从页面价值-用户矩阵计算出页面价值间的加权欧氏距离, 并由距离大小获得等价值页面集. 再根据跳转偏爱度把等价值页面集转化为 2-项频繁访问子路径集, 并经过自适应的合并算法得到最终的频繁访问路径集. 实验证明该页面价值模型能高效获得更精准的频繁访问路径.

关键词: 页面价值; 跳转偏爱度; 用户频繁浏览路径; 网站日志

A Novel Model for Mining Frequent Paths Based on Page Value and Jump Preference Degree

LI Ai-Fei, JI Zhen-Yan, WANG Jing-Wei

(School of Software, Beijing Jiaotong University, Beijing 100044, China)

Abstract: A model is designed and implemented for mining user frequent paths from Web log. An important basis for webpage clustering analysis is proposed, as well as the concepts of page value and jump preference degree. Then we build the page value model. The distances of page value are calculated out from the Page Value-User Matrix, then the value-equal page set is obtained according to the distances. After that, the set is transformed to binomial frequent path set. Finally the user frequent path set can be obtained by applying an adaptive merging algorithm. Experimental results show the model has better accuracy with high efficiency.

Key words: page value; jump preference; user frequent path; Web log

1 引言

随着 Web 技术的迅速发展, 已经有越来越多的用户在网上进行购物, 学习, 股票交易等活动. 然而, 由于 Internet 信息呈指数增长, 信息过载和资源迷向已经成为制约人们高效使用信息网络的瓶颈, 使得合理的网页架构成为网站的重要需求. 因此, 从大量的 Web 日志数据[1]中挖掘用户频繁访问路径, 在改善网站网页架构, 让用户获得更好体验方面起到了重要推进作用.

目前, 基于 Web 日志的挖掘用户频繁访问路径技术逐渐成熟, 但仍需要进一步研究以获得更合理更实用的模型. 文献[2]采用树形结构来分析, 但未能兼顾到用户“返回上一页”等回溯问题. 文献[3]利用浏览次数、时间和会话数据量, 对页面进行了聚类分析, 但没有区别不同用户的数据作用大小, 且放大了网页数据量这个影响因子.

本文首先对网页聚类分析的前提进行了深入研究, 提出: 用户浏览网页的目的是工作、学习、娱乐等, 表明该网页对用户有一定的价值, 本文称为页面价值(用户访问的次数、浏览时间均与页面价值呈正相关性); 若有用户频繁访问路径 W(W 是页面标号组成的有序序列), 则 W 上所有页面的页面价值对较多用户是相同的. 以上观点, 是对网页聚类分析的根本出发

① 收稿时间:2012-09-05;收到修改稿时间:2012-10-29

点. 以下是模型图解:

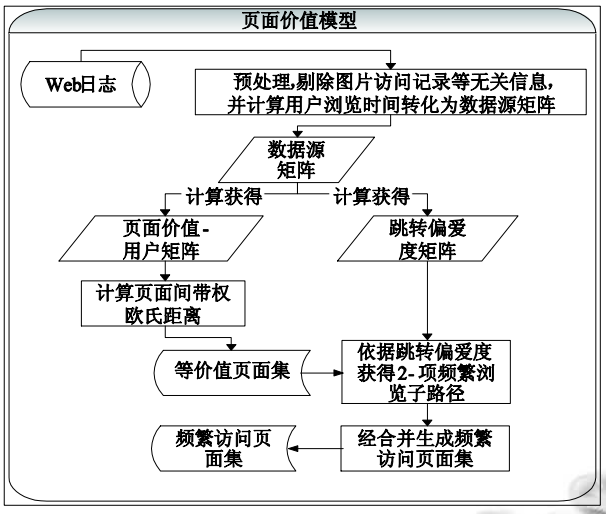

图 1

2 相关定义

定义 1. 页面价值 页面价值-用户矩阵

页面价值: 某页面对单个用户的价值, $V = C \bullet \sum_{i=1}^{n} T_i \cdot$ (C 为某用户访问某页面总次数, T_i 为该用户对该网页第 i 次浏览时间的离散化值[2]).

页面价值-用户矩阵: 以 URL 标号为行标, 以用户标号为列标, 第 i 行第 j 列的元素即为 i 号网页对 j 号用户的页面价值.

定义 2. 跳转偏爱度 跳转偏爱度矩阵

跳转偏爱度: 设 U 是所有页面统一资源定位(URL)的集合, W 是所有频繁访问路径或其子路径的集合, 对 W 中任意一个页面 u 有 u∈U. 取 w∈W, 对于 w 上任意一个页面序列 x, 它的前 m 个页面都相同, 第 m+1 个页面有 n 种可能, 称这 n 种可能的页面中的任意一个页面 k-1 到另一个页面 k 的跳转偏爱度[2]为:

$$p_k = C_k \bullet T_k \Big/ \frac{\sum_{i=1}^{n} C_i \bullet T_i}{n}$$

其中, C_i 表示所有用户从页面 i-1 跳转到页面 i 的次数总和, T_i 是所有用户对页面 i 的离散化浏览时间的总和.

跳转偏爱度矩阵: 以 URL 标号为行号, URL 标号为列号, 建立一个矩阵. 在第一行前加入一行 NULL, 表示“从 NULL 到其它网页的跳转偏爱度 P_{NULL_URL}”. 这里的“其它网页”是用户通过书签、直接输入网址、从其它网站链接接入等方式, 而不是从本网站其它页面链接访问到该页面. 在第一列前加入一列 NULL, 表示“用户从其它网页跳到 NULL 网页的跳转偏爱度 P_{URL_NULL}”, 该网页是用户最后访问本网站的页面, 用户在访问本页面后结束浏览或者跳转到其它网站浏览. 如果网站有 n 个 URL, 那么矩阵是(n+1)方阵.

$JPM_{(n+1)\times(n+1)}$形如:

$$\begin{array}{c} \\ NULL \\ URL_1 \\ \vdots \\ URL_n \end{array} \begin{array}{c} \begin{array}{cccc} NULL & URL_1 & \cdots & URL_n \end{array} \\ \begin{bmatrix} 0 & P_{NULL\_URL_1} & \cdots & P_{NULL\_URL_n} \\ P_{URL_1\_NULL} & 0 & \cdots & P_{URL_1\_NULL_n} \\ \vdots & \vdots & \ddots & \vdots \\ P_{URL_n\_NULL} & P_{URL_n\_URL_1} & \cdots & 0 \end{bmatrix} \end{array}$$

定义 3. 页面价值距离 等价值页面 等价值页面集

页面价值距离: 页面价值-用户矩阵中任意两个行向量的加权欧氏距离. 若行向量用 $\vec{A}$ 和 $\vec{B}$ 表示, 则它们的加权欧式距离 ρ (A, B)为: ρ ($\vec{A}, \vec{B}$) =sqrt[Σ (wi ×(a[i] - b[i])^2)] (i = 1, 2, …, n)

其中: $\vec{A}$ =(a[1], a[2], …, a[n]) , $\vec{B}$ =(b[1], b[2],…, b[n]) . w_i 是页面价值-用户矩阵中第 i 号用户的权值, 表示该用户的网站浏览数据对挖掘频繁访问路径的贡献大小(本文认为对网站的访问次数较多或时间较长的用户数据, 对模型有更大的意义). w_i 计算如下:

$$w_i = \sum_{i=1}^{n} X_{ij} \Big/ \sum_{i=1}^{n}\sum_{j=1}^{n} X_{ij}$$

(X_{ij} 为页面价值-用户矩阵中第 i 行第 j 列的数据)

等价值页面: 页面价值距离在相同范围内的页面.

等价值页面集: 以若干对等价值页面为单位, 构成的不重复的等价值页面集合.

3 主要算法描述

算法 1. 等价值页面集生成算法

目的: 由页面价值-用户矩阵, 通过比较页面间加权欧氏距离得出等价值页面矩阵, 该矩阵为后来的 2-项频繁访问路径计算提供数据.

输入: 页面价值-用户矩阵 PVUM, 加权欧氏距离阈值 threadhold, 存储等价值页面的矩阵 TM, 页面跳转偏爱度矩阵 PSM.

输出: 以 tm[ ][ ]数组为依据, 用于存放等价值页面的路径集.

算法: for: int i=0,tm_line=0{//i 为循环遍历 PVUM 中的行值, 大于//page_num 为跳出循环的条件

for: int j=i+1;// j 为列标, 大于 user_num 为跳出循环的条件 count=0; //每次循环前置 0, 表示页面 i 和页面 j 之间加权欧式距离 for: k=0 {// k 为列标, 大于

Al_const_user_num 为跳出循环条件 count=count+wi×Math.pow(PVUM[i][k]-PVUM[j][k],2);}

If(count<threadhold){//如果两个页面之间的加权欧式距离小于阈值则把这两个页面存入到 TM 三元组中,该页面标志拷贝到 TM}}

算法 2. 频繁访问路径合并算法

目的: 根据处理前的矩阵 pm_before_combine 中子路径上页面数目 path_item_num 把路径合并到矩阵 combined_pm 中.

输入: 处理前的矩阵 pm_before_combine .

输出: 处理后的矩阵 combined_pm; 算法返回值: 新路径的个数.

算法: for:int line_i,comcount=0;

{// line_i 是矩阵的行号, comcount 表示本次合并成功数目

{for: line_j =0; {　//line_j 是矩阵的行号

if(line_i 和 line_j 可以合并, 并且合并后的路径不出现重复网页, 合并后不与其他路径重合) {//输出到 combined_pm 相应位置

for: q=1{//q 为 pm_before_combine 矩阵的列标, 把要合并的两路径 WA 和 WB 的 WA 数据拷贝到 combined_pm}

for:int r=1; {//r 表示源数据 pm_before_combine 的列标

把要合并的两路径 WA 和 WB 的 WB 数据拷贝到 combined_pm}

comcount++; Al_path_count++;}}}//为 pm_before_combine 的行标

return 本次合并产生的新的路径个数 path_num_temp;

4 模型实现

4.1 对 Web 日志的预处理

因为相关技术非常成熟, 这里不再赘述. 其中用户对页面 A 浏览时间定义为某用户最近两次打开网页(先后打开 A 和 B)的时间差值. 预处理[5]后得到的存储矩阵[6]为:

$$DM_{376\times4}=\begin{vmatrix}4 & 2 & 2.5 & 1\\ 2 & 5 & 13.5 & 1\\ 3 & 6 & 1 & 2\\ \vdots & \vdots & \vdots & \vdots\\ 2 & 6 & 6.2 & 4\end{vmatrix}$$

其中 376 是日志记录个数. 存储矩阵每一行对应一次会话. 第一列是会话对应的页面 URL 的标号, 第二列是会话对应页面的 reURL(即用户由 preURL 页面跳转到当前 URL 页面), 第三列是会话对应的离散化浏览时间, 第四列是会话对应的用户标号.

4.2 计算页面价值-用户矩阵和跳转偏爱度矩阵

对矩阵 DM376×4 处理, 把相同的用户浏览次数和时间相加, 获得页面价值, 并根据页面价值-用户矩阵定义可得:

$$PVUM_{6\times6}=\begin{vmatrix}32 & 3 & 81 & 45 & 0 & 7\\ 59 & 27 & 42 & 36 & 33 & 38\\ 171 & 53 & 74 & 75 & 83 & 44\\ 58 & 80 & 56 & 136 & 53 & 76\\ 53 & 6 & 35 & 99 & 57 & 6\\ 42 & 15 & 41 & 62 & 83 & 38\end{vmatrix}$$

由存储矩阵得到从 URLa 跳转到 URLb 的跳转偏爱度, 根据跳转偏爱度矩阵定义得到:

$$JPM_{7\times7}=\begin{vmatrix}0 & 0 & 0 & 0.1757 & 0 & 0 & 0.1444\\ 0.0915 & 0 & 0 & 0.0584 & 0.1373 & 0.017 & 0.356\\ 0.0886 & 0.3487 & 0 & 0 & 0 & 0 & 0.5431\\ 0.0888 & 0.2289 & 0.7346 & 0 & 0 & 0.0615 & 0.3075\\ 0.1746 & 0.2194 & 0.3268 & 0.5283 & 0 & 0 & 0.2686\\ 0 & 0.0839 & 0.1198 & 0.4315 & 0.6953 & 0 & 0.0599\\ 0.1248 & 0 & 0.9362 & 0.0221 & 1.9895 & 1.4590 & 0\end{vmatrix}$$

4.3 计算等价值页面集

算法一已经实现该步骤. 首先计算页面价值间加权欧式距离, PVUM 的第一行 PVUM1 和第二行 PVUM2 的距离为:

$distance_{12}=0.2185\times(32-59)^2+0.0968\times(3-57)^2+0.1732\times(81-42)^2+0.2385\times(45-36)^2+0.1627\times(0-33)^2+0.11\times(7-38)^2$

以此类推得到其它 distanceij. 根据页面间加权欧式距离阈值 12000, 得到等价值 页面集{{1,2}, {1,6}, {2,6}, {5,6}}.

4.4 从等价值页面集获得 2-项频繁访问子路径

等价值页面集中: {1,2}对应路径 1→2 和 2→1. 在跳转偏爱度矩阵中找到 : 路径 1→2 对应的第一行第二列的数据 0<0.2(0.2 是切换偏爱度的阈值), 路径 2→1 对应的第二行第一列的数据 0.3487>0.2. 所以得到一个 2-项频繁访问子路径 2→1, 将它存储到矩阵中.

{2,6}对应路径 2→5 和 5→2. 以此类推. 获得 2-项频繁访问子路径集为: {2→1, 1→6, 2→6, 6→2, 6→5}

4.5 循环合并获得 频繁访问路径集

对 2-项频繁访问子路径集进行一次 合并(即算法二完成的功能), 如果合并产生了新的路径 w, 则继续把 w 加入 2-项频繁访问子路径集所在集合. 并对该集合继续使用算法二. 直到不再产生新的路径为止, 最后得到频繁访问路径集为: {2→1, 1→6, 2→6, 6→2, 6→5, 2→1→6, 1→6→2, 1→6→5, 2→6→5, 6→2→1, 2→1→6→5}

5 结果分析

我们在 WindowsNT 平台上, 用 java 和 mysql 实现了本文的模型和文献 3 和文献 4 的算法. 并对结果进行了对比. 以下是效率比较图(图中横轴为数据条数(单位兆 M), 纵轴为 cpu 处理需要的时间(单位秒 S). 试验环境:cpu2.53G(i3 双核), 可用内存 2.99G, WindowsNT)

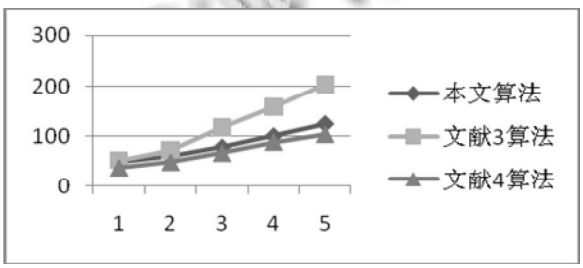

图 2 算法效率比较图

由上图可知, 本文算法效率接近文献 4 的算法, 但因为考虑了时间因素, 比文献 4 的算法更有说服力. 同时本文算法比文献 3 更高效, 同时也保证了准确性.

另外: 文献 3 求出的频繁访问路径集为: {2→1, 1→6, 2→6, 5→2, 6→2, 6→5, 2→1→6, 1→6→2, 1→6→5, 5→2→6, 2→6→5, 6→2→1, 2→1→6→5}与本文所得结果仅在 5→2, 5→2→6 处有区别. 经过分析 5→2 的切换偏爱度为 0.0839, 显然跳转偏爱度很低, 所以 5→2, 5→2→6 不适合做频繁访问路径集的成员. 显然本文算法更准确.

因此本文的算法更加全面, 因为去掉了会话数据大小这个影响因子, 比文献 3 有更好的稳定性和更强的说服力. 同时本文在获得 2-项频繁访问路径时只需对跳转偏爱度矩阵进行访问, 算法效率极大提高.

6 结语

本文对当前基于 Web 日志页面的聚类分析提供了重要理论支持, 并提出页面价值理论和跳转偏爱度理论. 以页面价值模型为基础, 对某网站后台管理系统的 Web 日志进行了分析, 证明了本模型比其它模型在准确性和有效性上有很大提高, 可以更好的挖掘到频繁访问路径. 同时, 本模型还有可以提高的地方: 一、我们在研究过程中发现, 当网页的用户数据量超大时, 可对用户先进行聚类分析然后再获得等价值页面集. 这使得模型算法的效率大幅提升. 二 、在离散化浏览时间的区间划分界限方面, 经过进一步的研究可以得到更准确的取值.

参考文献

1 黄磊,黄汉永.XML 技术在 Web 挖掘中的应用.信息技术, 2003,27(5):6–13.
2 邢东山,沈钧毅.一个可以准确反映 Web 浏览兴趣的度量值——偏爱度.控制与决策,2004,19(3):307–310.
3 任永功,付玉,张亮.一种改进的用户浏览偏爱路径挖掘方法.计算机工程,2009,35(8):47–49.
4 杜家强,韩其睿,王科,杜家兴.web 日志中用户频繁路径快速挖掘算法.计算机工程与应用,2005,41(22):164–167.
5 彭曙蓉,王耀南,杨文忠.基于马尔可夫链的 Web 访问序列挖掘算法.计算机工程与设计,2006,27(2):332–334.
6 冯晨,张旭翔.数据挖掘技术及算法综述.电脑知识与技术, 2009,5(13):3331–3332.

(上接第 80 页)

8 何芝霞,黄昶,何云东.基于 CompactRIO 的数据采集系统.仪器仪表用户,2009,16(1):37–39.
9 张万峰,景永刚,等.基于NI-CompactRIO平台的水声信号参数估计实现.声学技术,2010,6–13,29–3.
10 齐晶晶,黄彩霞.基于 FPGA 和 LABVIEW 的信号源设计.电脑知识与技术,2008,11.
11 包敬民,齐新社,马刚.基于 Labview8.2 的虚拟频谱分析的设计.现代电子技术,2007,22(261):200–202.
12 周求湛.虚拟仪器与 Labview 7 Express 程序设计.北京:北京航天航空大学出版社,2004.1–3.
13 孙俊卿,罗云林,等.基于 Labview 的虚拟信号分析仪设计与实现.微计算机信息,2010,26:9–12.