使用去噪和相异度的电子商务网站用户访问聚类算法①

肖 强 钱晓东 (兰州交通大学经济管理学院 甘肃 兰州 730070)

摘 要: 电子商务网站包含相当大的用户访问信息，对用户信息的数据挖掘，可以加强网站对用户访问信息的准确了解，提高电子商务网站的点击率。为此将提取电子商务网站日志中记录的用户访问链接数据，利用去噪技术对用户访问链接日志记录数据进行过滤分析，将过滤后的用户访问数据利用相异度二元关系组成二元数组，通过对二元数组的相异度分析计算，可实现电商务网站用户的聚类，为网站页面的优化及访问用户的兴趣、爱好的掌握提供参考。

关键词: 聚类; 电子商务; 网站日志; 去噪; 相异度

User Access Clustering Algorithm in Electronic Commerce Website Using Denoising and Dissimilarity

XIAO Qiang, QIAN Xiao-Dong

(Economics and Management School, Lanzhou Jiaotong University, Lanzhou 730070, China)

Abstrat: E-commerce web site contains a large user access to information., by data mining, can enhance the site to an accurate understanding of user access to information to enhance e-commerce web site click-through rate. To do this will extract the e-commerce website logs the user access to link data, the use of denoising techniques logging user access to link data filtering analysis, and the data was became binary array by dissimilarity binary relation, so this data will be computed by differentiation analysis, and the realization electricity commerce website user' s cluster, will provide the reference for the website page optimization and user's interest and habbies.

Keywords: clustering; electronic commerce; web log; denoising; dissimilarity

随着社会信息化进程的加快，网络化技术得到了进一步的提升，正是在这样一种背景下，电子商务网站得到了飞速的发展，然而电子商务网站带来大量用户访问的同时，也带来的大量的用户访问链接数据信息，但这些信息却没有得到有效地分析[1]，致使许多电子网站的客户信息数据白白浪费和闲置。

为了更好地利用这些电子商务网站的用户访问资源，为电子商务网站的改善其本身的服务性能，提供更好、更合适的服务，从而提高电子商务网站访问率和购买率，需采取合适的方法来进行用户访问链接数据的处理。

通过对电子商务网站用户访问链接数据的去噪预处理，以及访问用户间的相异度分析计算，为电子商务网站的用户聚类分析提供依据，为提高电子商务网站的页面优化提供基础。

1 用户访问链接数据的去噪处理

用户在访问网站的过程中，会在网站上形成一系列的点击记录，这些点击会以日志的形式记录在 web 服务器中[2,3]，为此进行电子商务网站访问用户的聚类

① 基金项目:国家社科基金项目(08XTQ010)
收稿时间:2010-03-04;收到修改稿时间:2010-03-29

需要首先获取 web 服务器中的日志记录，并进行去噪预处理。

1.1 web 服务器日志

Web 日志服务器记录用户访问站点时的每个页面的请求信息，常用的 web 日志文件主要内容包括：data(访问的日期)、time(访问的时间)、ip(客户机的 IP 地址)、method(用户请求的方法)、url(请求的链接地址)、status(返回的状态码) size(请求返回的结果大小)、agent(用户代理)、 referee(用户浏览的上一页)[4]，如下图 1 所示：

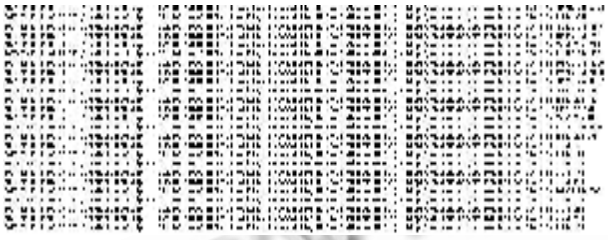

图 1　web 日志服务器记录的用户访问链接记录截图

1.2 用户访问数据的去噪

由于 Web 服务器日志记录的是用户访问站的记录，但并没有进行用户分析，因此记录中含有大量的无用信息即噪声，为了给下一步用的聚类提供可靠的数据资源，需对用户访问数据进行去噪处理。

1.2.1 用户访问数据收集

Web 服务器日志记录可用记事本形式浏览，同时根据用户访问日志记录形式编写如下算法，实现对日志记录中用户访问数据的收集。

(1) 打开服务器日志文件；

(2) 判断是否到文件底端，没有执行(3)步，否则执行(5)步；

(3) 分别将服务器日志记录中的<data time ip url>[5,6]取出存于数组 C 中；

(4) 重复(2)步；

(5) 到文件底部，完成服务器日志文件数据的初步收集。

1.2.2 去噪处理

将存于数组中的用户访问的主要信息资源进行去噪处理，去噪过程分为三步：

(1) 读取初步收集数组中的数据，进行 IP 去噪，去噪原理为，若用户访问的记录为单个而无其它链接，则认为该用户为无效访问用户或认为是有噪信息。为此对数组中的 IP 地址进行比较、累加，并依据下述公式去噪。

$$\begin{cases}\sum C_{i\cdot ip}>1, & D_j=C_i \quad i=1\cdots n;\ j=1\cdots m\\ \sum C_{i\cdot ip}=1, & C_i\text{去掉}\end{cases} \tag{1}$$

其中 $\sum C_{i\cdot ip}$ 表示相同 IP 地址用户进行链接访问的次数；

C_i 表示为服务器中日志记录的数据<datai timei Ipi urli >

D_i 表示去噪后新组成的日志记录数据数组。

(2) 将得到的新数组 D 按数组中每行数据中 IP 地址是否一致的将数组结合成下述模式：

U=<IP ,(data1 ,time1,url1),⋯(datan ,timen,urln)>(2)

并进行数据中的时间去噪。去噪原理：凡用户浏览时间为某个门限范围内则认为该访问数据有效，否则认为该用户访问的链接地址无效或认为是有噪信息，为此对数据进行下述处理

$$\begin{cases}ul_j=url_i & t_{\min}<time_{i+1}-time_i<t_{\max}\\ 0 & \text{其它}\end{cases} \tag{3}$$

(3) 将二次去噪的数组 U 根据访问链接 ulj 进行重新组合，得如下形式的数组：

N=<IP,ul1,ul2⋯uln>　　(4)

2　数据去噪后的相异度聚类算法

2.1 相异度

相异度是表征两对象之间的近似程度，通常用 Jaccard 进行评价[7,8]，如下式(5):

$$d(i,j)=\frac{r+s}{q+r+s+t} \tag{5}$$

在本算法中将是否访问链接的状态用 0 或 1 表示，式中(i,j)表示两个不同的 IP 地址的用户，r 是对于用户 i 访问链接时值为 1，用户 j 没有访问链接时值为 0 的变量数量；q 表示用户 i 和 j 都进行访问链接时值为 1 的变量数量； s 是对于用户 i 没有访问链接时值为 0，用户 j 访问链接地址时值为 1 的变量数量；t 表示用户 i 和 j 都没有进行访问链接时值为 0 的变量数量。

2.2 Web 页面数据的相异度聚类算法

通过上述的去噪处理和相异度概念的解释，可将 N 数组中不同 IP 中的 ul 访问链接进行对比，将访问

链接不同的 ul 列在下表 1 中：

表 1　相异度二元关系表

Ip	ul1	ul2	……	uln
N1(172.16.17.77)	1	1	……	1
N2(172.16.100.33)	1	1	……	0
……	……	……	……	1
Nn(172.16.9.45)	1	0	……	0

在上述表中，横向表示某用户所访问的链接，纵向表示某用户是否访问了该链接，其中“1”表示访问，“0”表示没有访问。

设定判断两用户相异度的门限值为 S_{yz}，若满足下式

$$\begin{cases} d(i.j) >= S_{yz}, (i \neq j) & N_i \subseteq N_j \\ d(i.j) < S_{yz}, (i \neq j) & N_i \not\subset N_j \end{cases} \quad (6)$$

则可认为两用户具有相似的访问链接，且这两用户对某电子商务网站具有相似的浏览特性。反之，则认为这两用户对某电子商务网站不具有相似的浏览特性。

3　实验验证

为验证本算法的有效性和实用性，截取某电子商务网站 Web 服务器中的 15 个用户访问该网站图片链接的日志作为验证数据，通过去噪处理和对用户访问链接的相异度二元关系表的分析得到如下数组 A:

$$A = \begin{bmatrix} 1010111100 & 100010100 \\ 0101011100 & 100010010 \\ 1001000110 & 110001001 \\ 0001000111 & 100011000 \\ 1000111110 & 001000100 \\ 0001111100 & 111000110 \\ 0011100011 & 100111111 \\ 0010010010 & 000110001 \\ 1100101000 & 001000010 \\ 0000010001 & 000010000 \\ 1000000010 & 000000000 \\ 1100000100 & 001110000 \\ 1111110001 & 000100010 \\ 1011100000 & 000111111 \\ 0000011110 & 010001000 \end{bmatrix}$$

数组 A 中横向表示 15 个用户中每个用户所访问的电子商务网站上的 19 个产品图片链接，“1”表示链接，“0”表示没有链接。

$$B = [N1, N2, N3, N4, N5, N6, N7, N7, N8, N9, N10, N11, N12, N13, N14, N15]$$

数组 B 表示 15 个用户。

其主要实现程序如下

```
Dim S as single ’表示判断两用户的相异度阈值
  For i=1 to 15 ‘所有用户相比较相异度
      For x=1 to 15
       r=0:s=0:q=0:t=0
         For j=1 to 19’对链接的地址进行比较
            If  A(i,j)=1 and A( x,j)=1 then
             Elseif A(i,j)=1 and   A(x,j)=0 then
                 q=q+1
              Elseif A(i,j)=0 and   A(x,j)=0 then
                  r=r+1
                Elseif A(i,j)=0 and   A(x,j)=0 then
                     s=s+1
                else
                     t=t+1
        end if
       next
       d = (r + s) / (q + r + s + t)
       If    d >=S Then      '如果相异度大于阈值
     Print c(z); Spc(2); '则将两用户放在一个数组中
                  e(i) = e(i) + c(z)
        End if
    Next
Next
```

经 VB 程序实验其聚类类结果如下表所示：

聚类阈值 S 设定　聚类结果

表 2　用户访问链接聚类表

聚类阈值 S 设定	聚类结果
S=0.2	1.N1,N2,N3,N4,N5,N6,N7,N8,N9,N10,N11,N12,N13,N14,N15
S=0.3	1.N1,N2,N3,N4,N5,N6,N7,N8,N9,N10,N11,N12,N13,N14,N15
S=0.5	1.N1,N3,N14;2.N6,N7,N9,N10,N11,N14,N15;3.N2,N4,N15; 4.N8,N11,N12 5.N9,N10,N13　6.N5,N9,N13; 7.N1,N6,N7,N10,N14;　8.N2,N7,N10,N11,N15
S=0.6	1.N1;2.N2;3.N3; 4.N4;　5.N5;　6.N7,N10,N11,N15; 7.N8;　8.N9,N13　9. N6,N7,N10,N14;　10.N12
S=0.7	1.N1; 2.N2; 3.N3; 4.N4; 5.N5; 6.N6; 7.N7; 8.N8; 9.N9;10.N10;11.N11;12.N12;13.N13;14.N14;15.N15
S=0.8	1.N1;2.N2;3.N3;4.N4;5.N5;6.N6;7.N7;8.N8;9.N9;10.N10; 11.N11;12.N12;13.N13;14.N14;15.N15

从表中我们可以看出，阈值 S 越小，聚类分组就越少，电子商务网站用户的聚类效果就不很明显，但如果阈值 S 越大，聚类分组就越多，电子商务网站用户聚类的效果就越明显，但如果过大，聚类效果反而不好，因为聚类分组过多，影响了对同类型用户的判断，不利于电子商务网站对用户的分析。

4 总结

通过对电子商务网站 Web 服务器用户访问日志数据的去噪，以及通过相异度计算，对用户进行了聚类，从中可以发现有相同浏览特性的用户，为电子商务网站的进一步优化提供依据。

参考文献

1 王泽彬,金飞,李夏等. Web 数据挖掘技术及实现.哈尔滨工业大学学报, 2005,37(10):1403－1405.
2 王玉珍. Web 使用模式挖掘研究.计算机应用, 2003,23(7):86－88.
3 周晓梅,王潜平,苏琳.基于 XML 的 Web 数据挖掘模型的设计.计算机工程与设计, 2008,34(13):54－56.
4 肖剑,姜良华,章彪.WEB 浏览行为的客户端追踪的研究, 2007,23(11):270－272.
5 刘茂福,何炎祥,彭敏. Web 模糊聚类方法及其应用.计算机应用, 2009,32(1):155－158.
6 李净,袁小华,沈晓晶. Web 权威信息自动提取技术的研究及应用.计算机工程, 2008,34(13):54－56.
7 万仁霞,王立新,刘振文.基于等价相异度矩阵的聚类.计算机工程与应用, 2008,44(25):149－151.
8 赵明清,蒋昌俊,陶树平.基于等价相异度矩阵的聚类.计算机科学, 2004,31(7):183－184.