基于访问兴趣的 Web 用户聚类方法①

费洪晓　覃思明　李文兴　李钦秀　董　馨 (中南大学 信息科学与工程学院 湖南 长沙 410083)

摘　要： 基于 Web 日志的信息挖掘具有重要的意义，比如识别兴趣相似的客户群体有利于实现推荐和个性化服务。采用了多元线性回归分析用户浏览行为，直接对兴趣相似矩阵进行 λ 截聚类，最后通过计算项与类的连接强度来调整聚类结果。实验结果证明了该算法具有较高的准确率和良好的扩展性。

关键词： Web 日志挖掘；多元线性回归模型；兴趣相似矩阵；用户兴趣聚类；连接强度

Web User Clustering Based on Interest

FEI Hong-Xiao , QIN Si-Ming , LI Wen-Xing , LI Qin-Xiu, DONG Xin

(School of Information Science and Engineering, Central South University, Changsha 410083, China)

Abstract: Data mining based on Web logs is of great significance. For instance, it can discover groups of people with similar interests and facilitate recommendation and personal service. A new clustering method based on Web users' interests regressively analyzes users' behaviors, partitions the interesting matrix with a threshold λ, and finally relocates some elements of clusters based on the joint strength between an element and a cluster. The favorable precision and scalability of the algorithm are studied through the experiments.

Keywords: Web log mining; multiple linear regression; interesting matrix; clustering based on user interest; joint strength

Web 的方方面面正在飞速地发展着。早期的 Web 主要应用于信息共享，而当今 Web 的应用已经向电子商务、网络游戏、远程登录等领域延伸同时对 Web 站点的设计和功能提出了更高的要求：能够改善网站组织结构以方便用户快捷准确地找到所需要的信息；能够为用户推荐可能感兴趣的页面以提供个性化服务；能够发现潜在的访客群体，为不同访客群体做出准确的市场定位。因此，一种将传统数据挖掘应用于 Web 领域的技术—Web 挖掘应运而生。由于 Web 的信息普遍具有无结构化、缺乏完整性约束和分布松散等特点，直接对 Web 信息进行挖掘具有相当的难度。Web 日志具有完美的结构，其包含的可以揭示用户浏览行为的丰富信息为 Web 挖掘提供了良好的前提条件[1]。因此 Web 日志分析是 Web 挖掘的重要手段。

很多 Web 用户的浏览行为直接反映了用户的兴趣，而 Web 用户浏览行为由许多行为因素构成，因此用户兴趣与多个行为因素之间的关系问题可以认为是回归问题。

本文考虑了以上 Web 用户浏览行为的特点，引入了多元线性回归模型来描述用户兴趣与页面浏览行为的关系从而量化用户对网页的兴趣，在此基础上直接对相似矩阵进行 λ 截聚类，最后通过计算项与类的连接强度来求精从而得到最终的聚类结果。最后通过实验对算法的准确性和性能进行了验证。

1　数据准备

Web 日志包含极其丰富的用户访问信息，但是 Web 日志的数据形式与所需要的相似矩阵数据格式相差甚远，因此数据准备就是将原始的用户访问数据信息转化为结构化的用户相似性数据信息的过程。该过程包括：数据清洗、用户识别、会话识别、过滤信息量不足的数据、用户兴趣识别和构造相似矩阵。

① 基金项目:湖南省科技计划基金(2006JT1040)

收稿时间:2009-07-12;收到修改稿时间:2009-08-30

1.1 数据清洗

Web 日志的数据量异常庞大，若直接对 Web 日志进行挖掘无疑会造成算法效率低下。另外，Web 日志中存在大量的噪音数据和无效数据，这些噪音数据和无效数据很可能影响聚类结果的准确性[2]。因此，对 Web 日志进行分析前需要对其进行清洗。从 Web 日志的统计分析中可以确认，以某些扩展名(如.ico、.gif、.jpg、.css、.wmv、.swf 等)为后缀的URL与挖掘无关，同时Web日志中传输状态为“ 404 ”、“ 301 ”和“ 500 ”的记录同样视为无关记录。将这些无关记录清除，无疑大大减少了数据量，提高了挖掘的精度和质量。

1.2 用户识别

用户识别主要目标是区分和标识访问的用户。一种最好的用户识别方法是实现用户的注册访问，但是这种方法无疑限制了站点的游客式访问，一定程度上减少了站点的访问量。另外一种较好的方法是分析 Web 日志记录中访客 IP 和访客状态这两个字段值。若两条记录的访客 IP 和访客状态字段值都不完全相同，可以认为这是两个不同用户的访问记录；否则再根据站点的拓扑结构判断两条记录请求的URL是否存在链接，如果不存在链接可以认为是同一台机器上存在两个不同用户，否则说明两条记录对应于同一用户。

1.3 会话识别

会话是用户对网站的一次连续有效的访问，其表现形式是用户的访问路径序列。不同的用户对站点的访问应该认为是不同的会话。如果同一个用户先后请求的两个页面间隔在规定的时间(即会话有效期)以内，则可以认为这两个页面属于同一个会话；否则，这两个页面属于同一个用户的不同会话。不同站点的会话有效期不同，可以根据实际情况进行设定。

1.4 过滤信息量不足的数据

文献[3]指出可能存在这样的访客，其访问记录过于稀少而不能构成日志文件记录的主体 ;对于构成日志记录主要请求活动的访客，其访问记录中同样也存在点击总次数过少的请求页面。这些数据明显缺乏足够的信息量而不应该参与挖掘过程。因此，过滤日志中信息量不足的数据可以进一步缩小数据规模，缩减空间维数，对提高聚类算法的效率和聚类质量有重要意义。

1.5 用户兴趣识别

Web 用户的浏览行为直接体现了用户兴趣，因为用户访问 Web 站点一般都是按照兴趣进行的。分析用户访问站点的过程，不难发现 Web 用户的浏览行为一般涉及三个行为因素[4]：页面停留时间，页面访问的次数和访问路径次序。由于构造的关联矩阵只关注用户对单个页面的访问情况而无法考察访问路径次序对兴趣度大小的影响，所以我们通过引进多元线性回归模型来描述页面停留时间和页面访问次数对兴趣度的影响。回归线性方程可以表示为：$P_{ji}=aX_{ji}+bY_{ji}+c$，其中 P_{ji} 表示第 j 个用户对第 i 个网页的兴趣度；X_{ji} 表示第 j 个用户浏览第 i 个网页的停留时间；Y_{ji} 表示在某段时间内第 j 个用户重复点击第 i 个网页的次数；a、b、c 均为站点的经验值。

1.6 构造相似矩阵

以 UserID 为行、URL 为列和用户对页面的兴趣度为值构造 UserID – URL 关联矩阵。经过“ 过滤信息量不足的数据 ”这一步骤之后，关联矩阵的维数大大缩小。由关联矩阵的特点可以知道，对行向量进行相似性分析可以得到兴趣相似的客户群体。为了构造用户兴趣相似矩阵，我们定义：$S_{ij}=1-cD(R_i,R_j)$，其中 S_{ij} 为第 i 个用户与第 j 个用户的兴趣相似度；C 为将行相似距离格式化至[0,1]的适当系数；$D(R_i,R_j)$ 表示关联矩阵中第 i 行向量与第 j 行向量的相似距离。该相似距离可以用海明距离来表示为：$D(R_i,R_j)=\sum_{k=1}^{m}|P_{ik}-P_{jk}|$，其中 m 为关联矩阵的 URL 维数；$P_{ik}$ 表示第 i 个用户对第 k 个 URL 的兴趣度。根据行向量相似距离构造的 UserID – UserID 相似矩阵具有自反、对称特性但不具备传递性。

2 Web用户聚类

基于相似矩阵的聚类算法可以通过求相似矩阵的传递闭包来进行 λ 截聚类，但当相似矩阵的维数比较大的时候，求传递闭包的运算量相当大，可能导致算法效率低。理论上已经证明，直接聚类法与传递闭包法具有等价性[5]。因此，可以直接从相似矩阵出发，通过设定截集 λ 对相似矩阵进行截分。λ 截分之后类与类之间可能存在包含与被包含关系，所以需要将存在包含与被包含关系的类进行合并。另外，由于算法本身的原因，不同的类中可能存在相同的项。为了解决相同项的隶属问题，我们定义项与类的连接强度如下。

$$J(U_i,c_j)=\frac{\sum_{k=1}^{m}sim(U_i,U_k)}{m} \quad (1)$$

其中 $J(U_i,c_j)$ 表示 U_i 与 c_j 类的连接强度，$sim(U_i,U_k)$ 表示

项U_i与C_j类中的U_k项的相似度，m 为C_j类的大小。从定义上看出，项与类的连接强度体现了类的内聚性。

聚类算法描述如下：

输入：截集 λ 的值

输出：Web 用户聚类 C={c_k}

步骤 1：初始化：C=Φ

步骤 2：for 1≤i≤l(l 为相似矩阵维数)do

{

a) 初始化 $c_i = \{U_i\}$

b) for i<j≤l do

{

// sim(U_i,U_j)为相似矩阵第 i 行第 j 列的值

if sim(U_i,U_j) ≥ λ,then

$c_i = c_i \cup \{U_j\}$

}

c) if C 中的各元素与 c_i 不存在包含或被包含关系，then

$C = C \cup \{c_i\}$

else C 中只保留关系的包含者

}

步骤 3：计算各类的相同项与类的连接强度，将连接强度最大的项归入相应的类，消除重复出现的项。

步骤 4：输出聚类结果。

3 实例分析

从原始的 Web 日志出发，按照相似矩阵的构造步骤，可以建立如下的 UserID-UserID 相似矩阵。

$$M=\begin{bmatrix} 1 & 0.749 & 0.020 & 0.259 & 0.755 & 0.999 & 0.996 \\ & 1 & 0.259 & 0.510 & 0.994 & 0.750 & 0.753 \\ & & 1 & 0.748 & 0.265 & 0.020 & 0.017 \\ & & & 1 & 0.504 & 0.260 & 0.263 \\ & & & & 1 & 0.755 & 0.752 \\ & & & & & 1 & 1 \\ & & & & & & 1 \end{bmatrix}$$

当截集 λ=0.36562 时

a) λ 截聚类后的 7 个子类：

i. {U_1,U_2,U_5,U_6,U_7}

ii. {U_2,U_4,U_5,U_6,U_7}

iii. {U_3,U_4}

iv. {U_4,U_5}

v. {U_5,U_6,U_7}

vi. {U_6,U_7}

vii. {U_7}

b) 合并类包含关系得到 3 个子类:

i. {U_1,U_2,U_5,U_6,U_7}

ii. {U_2,U_4,U_5,U_6,U_7}

iii. {U_3,U_4}

c) 处理各子类中相同的项

U_2 与 i 类的连接强度：

$$J(U_2,i)=\frac{0.749+1+0.994+0.750+0.753}{5}$$

U_2 与 ii 类的连接强度：

$$J(U_2,ii)=\frac{1+0.510+0.994+0.750+0.753}{5}$$

由于 J(U_2,i)>J(U_2,ii)，因此 U_2 只属于 i 类。其他相同项可以依照类似的做法进行归类。

d) 输出聚类结果

i. {U_1,U_2,U_5}

ii. {U_6,U_7}

iii. {U_3,U_4}

当截集 λ=0 时，聚类结果为{ U_1,U_2,U_3,U_4,U_5,U_6,U_7}

当截集 λ=1 时，聚类结果为{U_1},{U_2},{U_3},{U_4},{U_5},{U_6,U_7}

当截集 λ= 0.62234 时，聚类结果为{U_1,U_2,U_5,U_6,U_7},{U_3,U_4}

4 实验结果

在 Windows Vista 平台上利用 Java 语言实现了聚类算法的多方面测试，实验数据分为人工合成数据和实际数据两种。测试机器的硬件配置为：处理器 AMD Dual-Core 1.9G，内存 2G。

首先测试的是截集 λ 的大小对簇数目的影响。以站点(http:// 211.66.184.35)2005 年 1 月 10 日 00:00:05 至 00:32:47 的 Web 日志作为实验数据，原始日志记录共有 13509 条，经过数据清洗等步骤之后日志记录为 1005 条，有效记录约占记录总数的 7.44%，共识别出 16 个主要访客和 17 个主要访问路径。不同的 λ 值对簇数目的影响如图 1 所示。可以看到当 λ 取值越大分类精度越高，一个项成为单独簇的机率越大；反之，多个项聚成一类的机率越大。

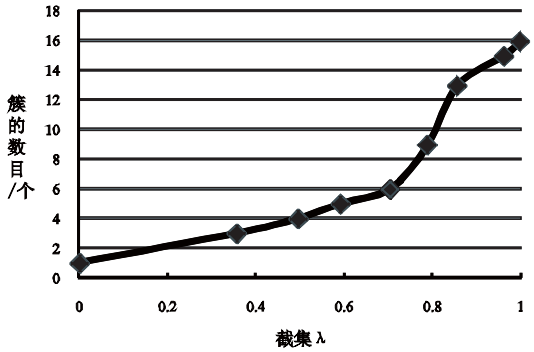

图 1 λ的值对簇数目的影响

接着测试的是算法的准确性。考虑到当用户数较少时，簇的数目较少可能导致聚类算法的准确度很高的极端情况，我们构造了 5 个包含较多用户的测试用例。这 5 个测试用例的用户数在 8 至 15 之间。首先分别对 5 个测试用例中的用户进行手工聚类，然后调整聚类算法的λ值以使簇的数目与手工聚类数目尽可能相近。定义准确度的评价标准为 $P=\frac{\sum_{i=1}^{n}P_i}{n}$，其中 n 为聚类数，代表各类的准确度，即各类与手工聚类相应类的重叠项与该类大小的比值。最后得到 5 个测试用例的准确度如图 2 所示，从图可以求得算法的准确度大概在 80.616% 左右。

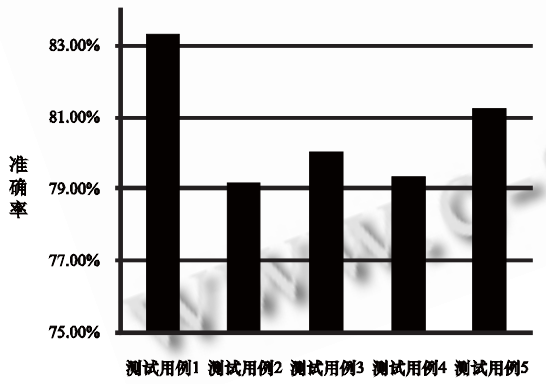

图 2 测试用例的准确度

最后测试的是算法的性能。为了使实验结果更有实际意义，我们选取了 5 个数据量较大的 Web 日志作为测试用例，它们的大小分别为 1203KB、1340KB、1466KB、1668KB 和 1856KB。调整λ值以获取较为合理的聚类数目，实验得到的算法所消耗 CPU 时间与日志数据量的关系如图 3 所示。从图中可以看到，当聚类数目较合理时，随着数据量增大，算法所消耗的 CPU 时间增幅较缓，扩展性良好。

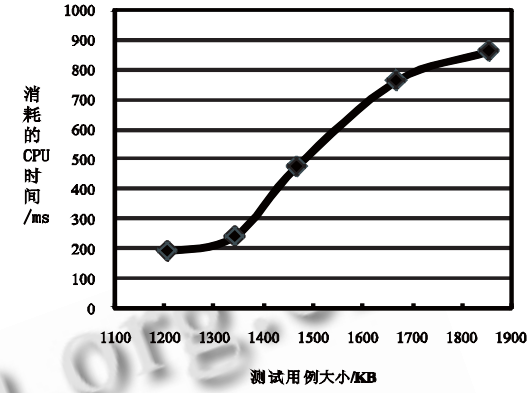

图 3 算法的性能

5 结语

随着 Web 的飞速发展，电子商务已经成为一种不可或缺的新型商业运营模式。电子商务网站存在着大量有商业价值的信息，如何挖掘这些有价值的信息对企业和商家的发展至关重要。从 Web 日志中识别用户兴趣和聚类用户，有利于实现推荐和个性化服务。本文提出了一种以多元线性回归分析用户浏览行为和以项与类的连接强度调整聚类结果的 Web 用户聚类方法。首先建立了多元线性回归模型以很好地将用户页面兴趣和用户浏览行为关联起来，通过行向量的相似性分析将用户访问矩阵转化为用户兴趣相似矩阵，再直接对相似矩阵进行λ截聚类，最后通过合并子类和计算项与类的连接强度来调整聚类结果。实验表明，该算法准确率较高，扩展性较好。

参考文献

1 Agosti M, Nunzio GMD. Gathering and Mining Information from Web Log Files. Berlin: Springer, 2007.

2 张欣,孙强,张蕾.基于兴趣的用户聚类分析在入侵检测中的应用.计算机工程与设计, 2008,29(6):1446－1447.

3 Rangarajan SK, Phoha VV, Balagani KS, Selmic RR, Iyengar SS. Adaptive neural network clustering of Web users. Computer Publication Date, 2004,37(4):34－40.

4 Dong Y, Zhang HY, Jiao LN. Research on Application of User Navigation Pattern Mining Recommendation. The sixth world congress on intelligent control and automation, 2006,2:6106－6110.

5 李鸿吉.模糊数学基础及实用算法.北京:科学出版社, 2005.