DOI:10.13833/j.cnki.is.2014.09.049

情报科学
第32卷 第9期 2014年9月

基于访问日志的几类数据挖掘模型体系构建

周文学,王　福

(内蒙古工业大学 图书馆,内蒙古 呼和浩特 010051)

摘　要: 以图书馆各类Web资源平台为基础,通过分析Web日志数据源,获取隐藏在访问日志背后的用户访问模式相关的信息。在此基础上描述和分析了几类基于统计方法的Web模型算法,并探索出在图书馆中应用的设计思路。本模型和算法比较简单适用,且易于实现,适合于图书馆的低成本的构建。

关健词: 图书馆;数据挖掘; 数据分析;算法设计; 模型设计

中图分类号: G250.7　　**文献标识码:** A　　**文章编号:** 1007-7634(2014)09-91-04

Mining Model System Based on the Access Log Several Data

ZHOU Wen-xue,WANG Fu

(*Library of Inner Mongolia University of Technology, Hohhot* 010051, *China*)

Abstract: With the libraries' all types of Web resource platform as the foundation, We can analysis the library Web log data acquisition in the access log, then can found the user access patterns related information which hidden behind. On the basis of the description and analysis We can get a Web model, and explore the applications of Library design. The model and the algorithm is relatively simple to apply, and is easy to realize at low cost, suitable for library construction.

Key words: library; data mining; data analysis; algorithm design; model design

1 引　言

用户访问图书馆资源会在Web页面中留下许多隐含了作者兴趣偏好的访问数据,这些数据具有内容不确定性、信息多样化和分散化的特点。若想从这些数据获取用户兴趣偏好就需要对这些数据进行清洗,先去掉噪声数据,然后利用复杂算法从这些人们事先不知道的、没有规律的、不完全的、模糊的数据中发现用户实际需求,以便更好的为用户提供最佳个性化服务提供依据。通过使用挖掘技术并结合图书馆数字资源Web平台,为图书馆数字资源的采访、订购、评价提供深入、准确、详细的基础数据,提高用户对图书馆数字资源的满意度。为了实现这些功能,提出了一种使用多种挖掘方法实现的模型和算法原理,并给出图书馆数字资源后台挖掘的设计思路,其系统架构如下图1所示[1]。

数据挖掘平台具有元数据处理管理、数据预处理管、数据挖掘管理、模型算法、评价管理和数字资源采访、订购、评价管理等功能。由于对用户行为挖掘的目的不同,所以模型算法和评价管理功需采用多种挖掘的数据模型对给出相应的分析结果来指导挖掘。平台会根据相应的算法来评价分析结果、根据用户兴趣偏好的不同,提出相应的建设性

收稿日期: 2012-08-21

作者简介: 周文学(1976-),男,内蒙古赤峰市人,副研究员,硕士,硕士生导师,主要从事高校图书馆管理研究.

建议，以期使图书馆确定最佳的采访和订购策略，以实现图书馆数字资源采访成本的最小化[2]。

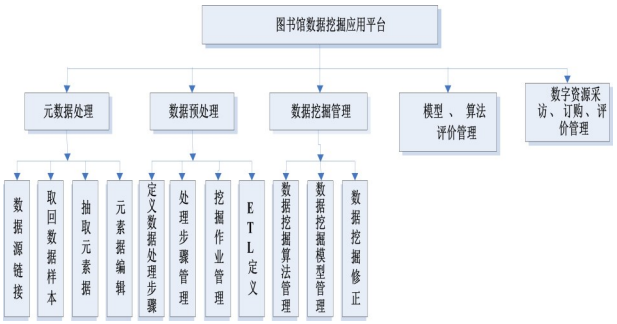

图1 数据挖掘平台架构设计结构图

2 图书馆Web挖掘的内涵

图书馆可以通过多种挖掘方法，如：Web用户操作行为挖掘、Web页面内容挖掘、Web页面结构挖掘等方法发现和感知用户的行为，如：图书馆可以挖掘分析web日志等数据源，提取用户对数字资源的偏好，发现多数用户集中访问资源。也可以从用户操作行为发现用户感兴趣的模式、发现隐藏在用户访问行为后的规律，通过这些挖掘，获取用户对信息的偏好，有针对性的进行个性化服务推送。通过对用户兴趣感知的汇总和分析、发现用户集中需求和分散需求，对图书馆数字资源的采访、订购、评价给出一个合理的分析结果，指导图书馆数字资源的合理采购。

图书馆采用的Web挖掘顺序是：先进行元数据的采集：就是采集访问电子资源的数据的数据，通过对这些数据进行清洗、集成、转换和消减，将来自多个数据源的数据按照一定的格式集中形成比较完整的数据集，从而为数据挖掘提供基础数据。在被挖掘的数据中，有些是对结果影响较大的，在不影响最终挖掘结果的情况下，对重点数据进行挖掘，可以大幅度缩小挖掘数据的规模，缩短数据挖掘的用时，提高挖掘的效率[3]。

可以对数据挖掘进行管理。数据挖掘的算法非常多，涉及到机器学习、统计学等诸多学科。目前常用的有：统计方法、判别分析、聚类分析、探索性分析、决策树分析、神经网络法、以及模糊集、粗糙集、支持向量集等。这些方法可以单独使用也可以多种方法配合使用，形成一个可视化的、完整的数据挖掘流程。也可以将每个模型都包装成节点，实现对生成模型进行增加、删除、修改、查询等编辑维护工作。

3 数据挖掘的模型及算法描述

由于要对图书馆数字资源用户行为进行挖掘，所以选择的数据挖掘方法有：统计分析、关联规则方法 、决策数分析方法、序列模式分析等方法的单独或结合，不同的挖掘目的可以采用不同的挖掘方法或多个方法的结合使用，以实现最优挖掘。

3.1 收集并预处理用户信息

如果把用户访问Web页面的操作行为定义为一个数据库系统，那么用户（User）、用户会话（User Session）、页面文件（Page File）、页面视图（PageView）、服务器会话（Server Session）等可以看作是数据库的字段，用户的访问行为衍生许多包含行为信息的数据可以看做是记录，这些记录在Web环境中交互。获取用户浏览行为的数据可以有多种途径，如服务器访问日志、cookies应用程序中的注册信息和相关记录等多种查询语言来获得Web的信息来完成信息的抽取[4]。

3.2 构建Web站点拓扑结构

图书馆的Web站点涉及到不同的层次关系和逻辑关系，各页面间有相互链接关系。为此，把图书馆Web各页面拓扑关系可以抽象为一个有向图 $G=(V,R)$ 其中 $V=\{URL_1, URL_2, \cdots, URL_m\}$ 是图书馆各类资源Web页面用户访问的的有穷非空集合；$R=\{<URL_1, URL_2>, <URL_1, URL_3>, ...\}$ 是页面之间的有序超链接集合。如图2所示，主页面抽象化为图的顶点，分支页面之间的超链接抽象化为图的有向边，顶点的入边表示其它页面对该页面的链接，出边表示该页面所指向的其它链接页面。当然在实际的处理中我们会发现网页的超链接指向的不仅仅是网页还有许多是指向图片、CSS样式文件和一些特殊类型文件。由于挖掘要获取的网站拓扑结构只反映不同网页间的关系而且上述提到的图片和样式文件等均为构成某一网页的元素，所以数据应忽略此类超链。参照有向图的邻接表存储形式在数据库中设计了两个表节点表和边表来存储网站拓扑结构，为此如下算法来获得网站的拓扑结构[5]。算法GetClearLog对数据进行日志清洗得到干净的分析数据；算法GetStructure根据前一算法得到的数据生成Node节点表和Association边表，从而得到网站拓扑结构。

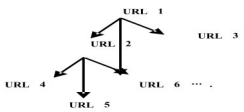

图2 图书馆WEB站点拓扑结构示意

在图2中，通过G可得到站点的所有URL链接。按URLi(i =l…m)进行分组计数，可以获得用户访问每个页面及相应的访问次数。

3.3 构造用户访问页面关联矩阵

建立Web站点的以URL id为行、User id为列、元素值为用户访问次数的关联矩阵。

$$\text{Fmxn}=\begin{Bmatrix}Url_1\\Url_2\\\cdots\\Url_m\end{Bmatrix}\{User_1 \quad User_2 \quad \cdots \quad User_n\}$$

$$=\begin{Bmatrix}A_{11} & A_{12} & \cdots & A_{1n}\\A_{21} & A_{22} & \cdots & A_{2n}\\\cdots & \cdots & \cdots & \cdots\\A_{m1} & A_{m2} & \cdots & A_mn\end{Bmatrix}$$

其中，Aij表示用户在一段时间内访问URLi的次数，行向量F[*,J]表示所有用户对URL "*"的访问情况，列向量F[i, *]表示用户“*”对该站点中所有URL的访问情况。行向量之和S_i表示所有用户对URL_i的访问次数，列向量之和S_j表示用户$User_j$对该站点中所有的URL的访问次数，反映用户访问的个性化网页。该模型及算法能够精准地记录用户对图书馆Web页面访问次数，从而利用SQL语句或多维OLAP构造查询：找出前N个页面，反映网页受欢迎的程度；找出前N个用户，用于发现访问次数最多的用户；定位某个用户的访问特征，如访问时间等[2]。

4 平台数据挖掘算法分析与实现

图书馆数据挖掘需要从服务器日志、cookies和用户注册信息等多种元数据采集数据，并获得用户访问行为数据库。数据库中由下面各表组成：User、Page、Resource、VisitPage、VisitResourcre、Uservisit 、Logfile、Ponder。其中，User、Page和Resource表各自保存有关用户、网页和资源的静态描述信息，Logfile表记录用户访问某特定网页的日期、时间、引用网页、客户端IP地址、操作行为、访问时间等，Ponder表则用户下载各类资源信息记录。通过Use-id和Page_ id相关联，建立Visit-Page表记录用户访问网页的时间、次数，同理可建立VisitResource表记录用户访问数字资源行为特征，建立Uservisit表记录用户的其他访问行为特征[2]。

4.1 发现重要数字资源集合页面

图书馆网站以各类电子资源为线索组织页面，同类电子资源成一个集合，按照下面步骤完成数据挖掘：①根据公式$S_i=\sum_{i=1}^{n}A_{ij}$计算出所有用户对该数字资源集合页面的访问次数，构成集合S={S_l, S_2, …,S_m}，其中，i=l,…,m；②在集合S中取S_1,S_2, …S_m的值依次按降序排序，排在前面的就是最重要的数字资源集合页面。具体在ASP页面设计查询，并按页面分组统计被访问次数并降序排列，同时编写SQL语句在数据库中检索，返回检索结果。

4.2 发现数字资源潜在用户

图书馆数字资源潜在用户是指那些浏览了Web页面却没有使用数字资源的用户。根据公式S_j各A_{ij}计算出所有用户对该网站所有页面的访问次数，构成集合S={S_l ,S_2, …S_n}，其中，j=1, …n；在集合S中取S_1,S_2, …S_n值依次按降序排序，排在前面的即为访问页面最多的用户；在集合S中除去已经在图书馆网站中利用数字资源的用户，这样就构成了潜在用户群体的序列[2]，用SQL语句进行相关查询，并把结果返回到查询结果数据库中。

4.3 发现个性用户偏好等特征

可以在发现重要用户的基础上，获得与该用户有关的信息并加以处理和分析。首先根据公式S_j各A_{ij}计算出各用户对该网站所有页面的访问次数，构成集合S={S_l ,S_2, …S_n}，其中，j=1, …,n。再根据公式$\text{Rate}_{i,j}=A_{ij}\sum_{j=1}^{n}S_j$计算出某用户对网站中各页面的访问率，构成二维数组。

$$\text{RATE}_{mxn}=\begin{Bmatrix}rate_{11} & rate_{12} & \cdots & rate_{1n}\\rate_{21} & rate_{22} & \cdots & rate_{2n}\\\cdots & \cdots & \cdots & \cdots\\rate_{m1} & rate_{m2} & \cdots & rate_{mn}\end{Bmatrix}$$

然后根据公式time = Page_Last-Time−Page_FirstTime，计算某用户在各网页逗留的时间，构成二维数组。

$$\text{TIME}_{mxn}=\begin{Bmatrix}time_{11} & time_{12} & \cdots & time_{1n}\\time_{21} & time_{22} & \cdots & time_{2n}\\\cdots & \cdots & \cdots & \cdots\\time_{m1} & time_{m2} & \cdots & time_{mn}\end{Bmatrix}$$

按照数组RATE和TIME的rate和time值降序排序，可以获得用户访问频率最高和停留时间最长也就是用户感兴趣的页面，实现步骤如下：创建个性用户查询，按用户、页面分组统计访问次数，并按用户id排序，方法基本同上，最后一行改为order by User. User id;计算页面访问率rate,滞留时间time等数据，并存入Individual_ Feature表[2-3]。

4.4 对个体用户进行关联数字资源推荐

相关数字资源推荐是通过分析用户所利用资源的联系，来分析用户的数字资源的利用偏好，以向其推送感兴趣的资源信息。对用户需要进行某类数字资源的推广和宣传，就需要在系统中寻找利用这些数字资源的用户同时还需要利用哪些数字资源，这样就可以把这些数字资源进行聚合，向其推送。首先根据公式 S_j 和各 A_{ij} 计算出各用户对相关网站访问次数，构成集合 $S=\{S_1, S_2, \cdots, S_n\}$，其中 $j=1, 2, \cdots, n$。其次根据公式 $\text{Relation}_{i,j}=\text{Aij}\sum_{j=1}^{n}S_j$ 计算出某用户对网站中相关页面的访问率，构成二维数组。这种推送目标群明确，节约了推送成本，效果要远远好于不分对象的全部推送。

$$\text{Relate}_{m\times n}=\begin{Bmatrix} relate_{11} & relate_{12} & \cdots & relate_{1n} \\ relate_{21} & relate_{22} & \cdots & relate_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ relate_{m1} & relate_{m2} & \cdots & relate_{mn} \end{Bmatrix}$$

最后可以从数组Relate中取值，依次按Relate降序排序，得到某一用户关联的数字资源页面[6]。

5 结 语

图书馆用户访问日志蕴含了丰富信息，如能对这些隐含信息加以分析和利用，可以更好的感知用户兴趣，对数字资源的利用情况进行分析评价。通过数据挖掘可以了解那些潜在用户的需求，对潜在用户的利用信息进行整理，对数字资源的采访、订购进行可行性分析，为馆领导和校领导决策提供依据。该平台提出和设计了数据挖掘模型和相关算法，该模型和算法具有简单、有效、易于实现的特点，如能结合其它方法，还可以进一步挖掘出更多信息。

参考文献

1 王仁武.基于社区Web日志挖掘的用户行为实证研究[J].图书馆论坛,2011,(4):100-102.
2 王 昌.一个简单WEB使用模式算法的电子商务应用[J]. 计算机系统应用,2007,(2):39-42.
3 王玉珍. 基于电子商务的Web挖掘技术研究[J]. 北京电子科技学院学报,2005,(12):22-25.
4 杨柏刚. WEB使用挖掘系统数据预处理子系统的设计[D]:北京:北京邮电大学,2008.
5 刘绍清,黄章树. 数据挖掘商业应用平台的设计与实现[J].计算机系统应用,2007,(7):11-14.
6 向坚持,等. 基于用户行为的Web使用挖掘数据采集技术研究[J].计算机与现代化,2007,(12):59-62.

（责任编辑：毛秀梅）

（上接第90页）

of Educational Technology & Society, 2013, (3): 179-190.
5 Gilakjani, A.P., Leong, L.M., & Ismail, H.N. Teachers ' Use of Technology and Constructivism [J]. International Journal of Modern Education and Computer Science, 2013, (4): 49-63.
6 马冲宇,陈坚林. 虚拟语言学习环境VILL@GE的项目分析及启示 [J]. 中国电化教育,2013, (2): 121-125.
7 蔡龙权,吴维屏. 关于把信息技术作为现代外语教师能力构成的思考 [J]. 外语电化教学,2014, (1): 45-53.

（责任编辑：毛秀梅）