一个简单 WEB 使用模式算法的电子商务应用

Application Research of a Simple Web Usage Mining Algorithm in Electronic Commerce

王 昌 （山西财经大学 信息与管理学院 山西太原市 030006）

摘要：在电子商务网站中，Web 使用挖掘可以通过分析 WEB 日志等数据源来获取与用户访问模式相关的信息。本文形式化描述和分析了一个 WEB 使用模式算法，并给出了应用的设计思路。该算法简单、有效、易于实现，适合于构建低成本 B2C 网站。

关键词：WEB 使用挖掘　数据分析　电子商务　B2C

1 引言

Web 挖掘是数据挖掘技术与 WEB 相结合的产物。电子商务是 Web 挖掘应用的领域之一。通过 Web 挖掘，可以为电子商务网站的管理提供深入、准确、详细的分析数据和有价值、易理解的分析知识，进而为电子商务网站的优化提供必需的信息，提高用户对网站的满意程度。本文介绍的就是一种使用统计分析方法实现的挖掘算法原理，并给出电子商务网站应用该算法的设计思路。

2 Web 使用挖掘的内涵

一般地，Web 挖掘可以分为三类：Web 内容挖掘、Web 结构挖掘和 Web 使用模式挖掘。Web 使用挖掘是指通过分析 web 日志等数据源，提取有关用户如何运用浏览器和使用网页间链接的信息，用以获取用户访问 web 页面的模式。其主要目标是从 Web 的访问记录中抽取感兴趣的模式，发现用户访问规律，进而应用到个性化推荐、系统性能改进以及商业智能等方面。

Web 使用挖掘的过程大致包括：数据采集、数据预处理、模式发现和模式分析。在电子商务中，数据源主要是基于点击流的相关记录，如服务器日志和 cookies。对这些数据进行预处理后，通过特定的挖掘方法和算法，并与应用层系统之反映业务过程的数据库信息相结合，可以得到一些有意义的和有价值的知识，如发现潜在客户的访问页面及路线、页面停留时间等偏好，或者发现畅销及滞销商品展示页面等。

3 算法描述

常用的数据挖掘方法有统计分析、关联规则、聚类和分类、序列模式分析等技术，可以根据具体业务的分析需求进行选择。电子商务网站中要挖掘客户访问模式，可以使用的方法之一就是统计分析，算法原理描述如下：

3.1 收集并预处理用户信息

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。这些数据通常都是结构化的，可直接或间接地转化为数据库信息。从数据源中抽取最关键的度量指标作为 Web 信息库的 1 层（最底层原始数据的高一级抽象层）用于挖掘，可形式化地表示为 L = < Userid, Clientip, URL, Time… >，其中，Userid 表示浏览者 id 号，Clientip 表示用户客户端的 ip 地址，URL 表示成功提交的 URL 请求，Time 表示用户请求页面的时间戳。

对该层数据库信息做计数分析，得到用户访问某特定网页的次数。

3.2 构建 Web 站点拓扑结构

把该 Web 站点的拓扑结构抽象为一个有向图 G = (V, R)，如图 1 所示。其中，V = {URL1, URL2, … URLm}，是页面的有穷非空集合；R = { < URL1, URL2 >, < URL1, URL3 >, … }，是页面之间的有序超链接集合。页面抽象化为图的顶点，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他链接页面。

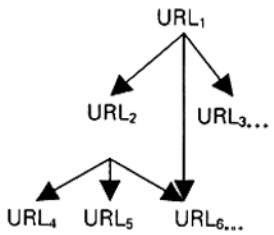

图 1　站点拓扑结构

在图1中，通过 G 可得到站点的所有 URL 链接。按 URLi（i=1…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。

在图2中，Aij 表示 j 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 Si 表示所有用户对 URLi 的访问次数，列向量之和 Sj 表示用户 Userj 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。

3.4 具体应用

该算法能够精确记录每个访问者对站点各页面的访问次数。继而可以应用 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。

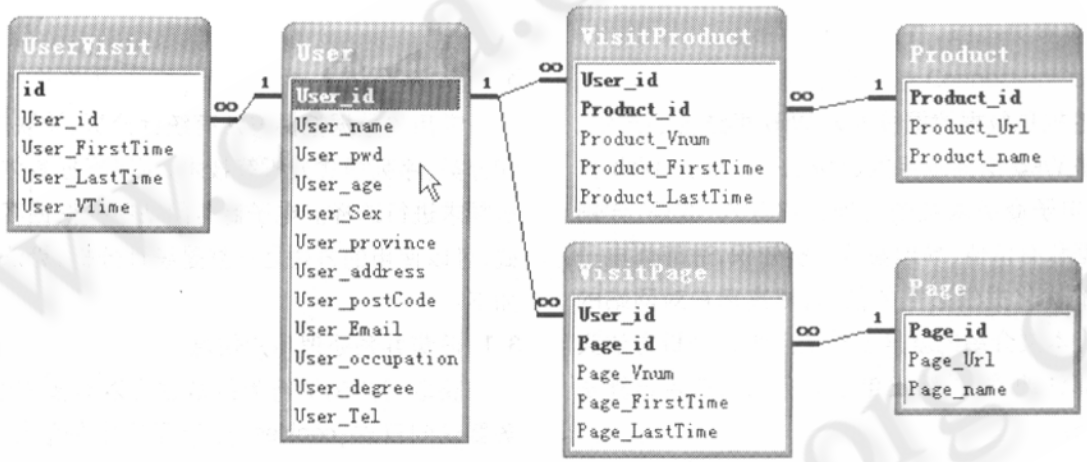

图 3　数据库表结构及关系

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，如图2所示。

$$Fm\times n=\begin{Bmatrix}URL_1\\URL_2\\\cdots\\URL_m\end{Bmatrix}(User_1 User_2\cdots User_n)$$

User_id

$$=\begin{Bmatrix}A_{11}A_{12}\cdots A_{1n}\\A_{21}A_{22}\cdots A_{2n}\\\cdots\cdots\\A_{m1}A_{m2}\cdots A_{mn}\end{Bmatrix}URL\_id$$

图 2　URL_id —User_id 关联矩阵

4　算法分析与实现

假设从服务器日志、cookies 和用户注册信息三种数据源中抽取数据，并建立较完整的用户身份记录和访问信息数据库。数据库可由 User、Page、Product、VisitPage、VisitProduct、Uservisit、Logfile、Order 等表组成，User、Page 和 Product 表分别记录有关用户、网页和商品的静态描述信息，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、所用时间等访问行为，Order 表则记录客户交易的订单信息。

然后，通过 User_id 和 Page_id 相关联，建立 VisitPage 表记录用户访问网页的时间、次数，同理可建立

VisitProduct 表记录用户访问商品页面的行为特征，建立 Uservisit 表记录用户的其他访问行为特征。各表结构及其关系如图 3 所示。对相关信息重新组织后，就可以应用该算法实现特定的业务需求。

4.1 发现重要商品专区页面

网站以商品为线索组织页面，同类商品构成专区，然后采用下列步骤：

① 根据公式 $Si = \sum_{j=1}^{n} Aij$ 计算出所有用户对该网站各商品专区页面的访问次数，构成集合 S = {S1 ，S2，…Sm}，其中，i =1…m

② 在集合 S 中取 S1，S2，…Sm 的值依次按降序排序，排在前面的就是最重要的商品专区页面。

设计实现的主要代码如下：

```
'创建重要商品专区查询(Important_product)——按页面分组统计被访问次数并降序排列
select Product. Product_Url, sum (VisitProduct. Product_Vnum) as TotalPerVnum
from Product inner join VisitProduct on Product. Product _id = Visit Product. Product_id
group by Product. Product_Url
order by 2 desc;
```

4.2 发现重要潜在客户

将潜在客户定义为那些浏览网页却没有在网站存在消费记录的用户，然后采用下列步骤：

① 根据公式 $Si = \sum_{j=1}^{m} Aij$ 计算出各用户对该网站所有页面的访问次数，构成集合 S = {S1 ，S2，…Sn}，其中，j =1…n

② 在集合 S 中取 S1，S2，…Sn 值依次按降序排序，排在前面的就是在站点中访问页面最多的用户。

③ 在集合 S 中减去已经在本网站中购买过商品的客户，就构成潜在客户群体的序列。

设计实现的主要代码如下：

```
'创建重要用户查询(Important_customer)，按用户、页面分组统计访问次数，并依次按访问总数、用户 id 降序排列 select sum(VisitPage. Page_Vnum) as total, User. User_id, User. User_name, Page. Page_Url, UserVisit. User_VTime
from ([User] inner join UserVisit on User. User_id = UserVisit. User_id) inner join (Page inner join VisitPage on Page. Page_id = VisitPage. Page_id) on User. User_id = VisitPage. User_id
group by User. User_id, User. User_name, Page. Page_Url, UserVisit. User_VTime
order by 1 desc , 2 desc;
'判断是否潜在客户
select * from Important_customer where not exists (select User_id from Order inner join Important_customer on Order. User_id = Important_customer. User_id);
```

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，步骤如下：

① 根据公式 $Sj = \sum_{i=1}^{m} Aij$ 计算出各用户对该网站所有页面的访问次数，构成集合 S = {S1 ，S2，…Sn}，其中，j =1…n

② 根据公式 $ratei,j = Aij / \sum_{i=1}^{n} Sj$ 计算出某用户对网站中各页面的访问率，构成二维数组

$$RATEm \times n = \begin{Bmatrix} rate_{11}\, rate_{12} \cdots rate_{1n} \\ rate_{21}\, rate_{22} \cdots rate_{2n} \\ \cdots\cdots\cdots \\ rate_{m1}\, rate_{m2} \cdots rate_{mn} \end{Bmatrix}$$

③ 根据公式 time = Page_LastTime – Page_FirstTime，计算某用户在各网页滞留的时间，构成二维数组

$$TIME_{m \times n} = \begin{Bmatrix} time_{11}\, time_{12} \cdots time_{1n} \\ time_{21}\, time_{22} \cdots time_{2n} \\ \cdots\cdots\cdots \\ time_{m1}\, time_{m2} \cdots time_{mn} \end{Bmatrix}$$

④ 分别从数组 RATE 和 TIME 中取值，依次按 rate 和 time 降序排序，可以得到客户感兴趣的（即访问频率最高和停留时间最长）页面。

设计实现的主要代码如下：

```
'创建个性用户查询(Individual)，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
<% @LANGUAGE = " VBSCRIPT" % >
<%
```

```
……
Dim connStr,conn,sql1, sql2, sql3,rs1,rs2, rs3
Set Conn = Server. CreateObject ( " ADODB. Connection" )
    connStr = " provider = microsoft. jet. oledb. 4. 0; data source = " &server. mappath ( " web. mdb" )
    Conn. open connStr
set rs1 = Server. CreateObject ( " ADODB. recordset" )
set rs2 = Server. CreateObject ( " ADODB. recordset" )
set rs3 = Server. CreateObject ( " ADODB. recordset" )
sql1 = " Select * from Individual where user_id = " &user_id
sql2 = " Select * from Individual_Feature"
sql3 = " Select * from VisitPage where user_id" &user_id
rs1. Open sql1,conn
rs2. Open sql2,conn,3,3,1
rs3. Open sql3,conn
while not rs1. eof and not rs3. eof
  time1 = rs3 ( " page_lasttime" ) - rs3 ( " page_firsttime" )
  rate = rs3 ( " Page_Vnum" ) / rs1 ( " total" )
  rs2. addnew
  rs2 ( " user_id" ) = user_id
  rs2 ( " rate" ) = rate
  rs2 ( " time" ) = time1
  ……
  rs2. update
  rs1. movenext
  rs2. movenext
  rs3. movenext
wend
rs1. close
rs2. close
rs3. close
conn. close
Set rs1 = nothing
Set rs2 = nothing
Set rs3 = nothing
Set Conn = nothing
% >
……
```

创建基于 Individual_Feature 表的查询，依次按用户访问率、时间降序排列，方法基本同上，不再赘述。

5 结束语

客户访问 Web 页面时在日志中留下海量信息，对这些信息和业务数据库信息加以分析和利用，有助于更好地理解客户行为和完善网站设计，对实现提高网站设计可用性、改善客户关系、提高系统性能等需求起到重要作用。本文形式化描述和分析了一个 WEB 使用模式算法，并给出了算法在电子商务网站中的应用实例。该算法简单、有效、易于实现，适合构建低成本 B2C 网站的 web 使用挖掘需求。在此基础上，结合其它方法，还可以进一步挖掘出更多信息，如相关 WEB 页面、相似客户群体、客户频繁访问路径等。

参考文献

1 宋擒豹、沈均毅等，Web 日志的高效多能挖掘算法[J]，计算机研究与发展，2001 年，第 38 卷，第 3 期：P328 - 332.

2 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 年，第 13 卷，第 4 期：P22 - 24.

3 Gordon S. Linoff , Michael J. A. Berry, Mining the Web: Transforming Customer Data into Customer Value[M]. American: John Wiley & Sons, Inc. , 2002:P34 - 42.

4 Jiawei Han , Micheline Kamber, Data Mining: Concepts and Techniques [M]. American: Morgan Kaufmann Publishers, Inc. ,2001:P440 - 443.