# 基于 WEB 使用挖掘的网站个性化服务系统的设计

## Design of Website with Personal Information Service System Based on Web Usage Mining

范利星 张水平 张凤琴 朱 涛 (空军工程大学 电讯工程学院 西安 710077)

**摘要**:在得到用户浏览模式的基础上,设计了一种基于多维关联规则的分类方法,根据不同的浏览模式对历史用户分类,并对不同类中的用户信息进行分析,得到各个类中的用户模式。

**关键词**:Web 使用挖掘 关联规则 分类分析 个性化信息服务

## 1 引言

WWW 自诞生以来,已经发展成为拥有亿万用户和上百万站点的巨大分布式信息空间。Web 上的数据的最大特点是无序性和单结构化。尽管用户可以利用 Web 形式发布信息、下载资料、查询数据,但效果不是很好,用户常常得不到自己需要的数据,有时甚至是一些垃圾数据。如何在浩如烟海的 Web 上找到需要的信息,比传统的数据库领域更加复杂和困难。Web 站点服务器每天产生大量的日志,其中蕴涵了有关用户在网站上的行为的丰富数据。分析这些数据能够发现有意义的访问模式和规则,对分析和改善站点的使用情况及资源配置具有重要的意义。如何对这些数据进行整理和分析,充分了解用户的兴趣爱好和使用模式,设计满足不同用户群体需要的个性化网站变得势在必行。

## 2 Web 使用挖掘与个性化服务

个性化服务是能够根据当前用户的使用模式自动调整站点结构与内容,根据用户的行为特征为其提供个性化服务,如推荐用户可能感兴趣的内容,或者在电子商务中为用户推荐可能会购买的商品等。从而尽可能使得每个用户在浏览该网站时都有自己就是该网站的唯一用户的感觉,网站尽可能地迎合每个用户的浏览兴趣并且不断调整以适应用户浏览兴趣的变化,使站点看起来有用并值得再次访问。

Web 数据挖掘技术就具有很多优点:不需要用户提供主观的评价信息,可以处理大规模的数据量,用户访问模式动态获取,不会过时,使用方便等。在三种 Web 数据挖掘技术中,Web 使用挖掘主要是挖掘用户的使用模式和偏好,所以更适合于实现个性化服务。因此,利用 Web 使用挖掘技术可以发现用户的使用模式和偏好,并对其提供个性化的信息服务。

## 3 用户浏览模式挖掘模块设计

挖掘用户的浏览模式分以下三步:

(1) 设计最大前向路径(Maximal Forward Path)算法,得到所有用户的最大前向引用。

(2) 在得到用户的最大前向引用基础上,基于 Apriori 算法设计适合挖掘用户频繁遍历路径的算法,得到用户访问的最大前向频繁遍历路径。

(3) 对所有最大前向频繁遍历路径进行关联规则分析,找出满足最小支持度和最小可信度的规则,从而发现用户访问的浏览模式。

### 3.1 最大前向路径(MFP)算法

MFP 算法是在一个用户会话中寻找最大前向路径。以一个实际的用户会话{A—B—C—D—C—B—E—F—G—F—H—A—I—J—I—K}为例说明算法的执行过程。首先对该用户会话构造出有向树,如图 1 所示。

MFP 算法对图 1 的用户会话进行 MFP 识别,其中设遍历方向向前时,Flag 为 1;遍历方向向后时,Flag 为

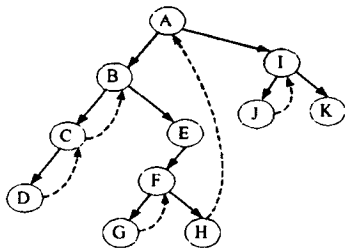


图 1　一个实际的用户会话

表 1　用户的 MFP 表

| 事务编号 | 最大前向路径(MFP) |
|---|---|
| 1 | {A, B, C, D} |
| 2 | {A , B , E , F , G} |
| 3 | {A , B , E , F , H} |
| 4 | { A , I , J } |
| 5 | {A , I , K} |

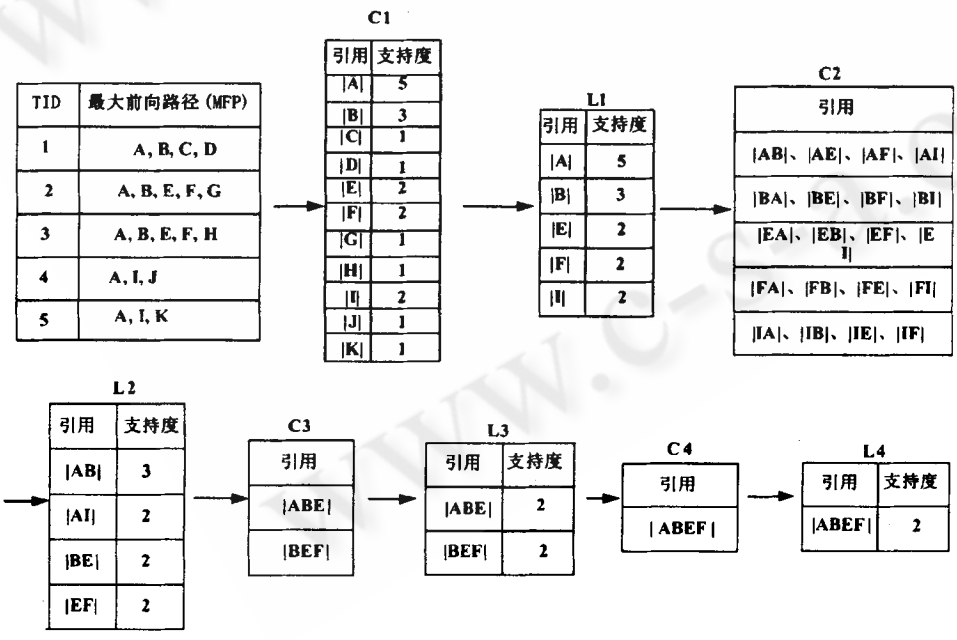


图 2　频繁遍历路径挖掘算法的执行过程

0。最后求得的所有 MFP 都输出到数据库的用户 MFP 表中，见表 1。该表中的数据就可以适用于挖掘频找出了所有 MFP 之后，挖掘频繁遍历路径的问题转化为在所有用户会话的 MFP 中发现频繁出现的连续子序列的问题。

### 3.2　挖掘频繁遍历路径算法

挖掘频繁遍历路径的过程是寻找事务表中的频繁项集的过程，挖掘用户频繁遍历路径的算法基于 Apriori 算法，如表 1 所示，有 5 个事务记录(MFP)。算法的执行过程如图 2 所示。

一个模式 P = {a1,a2,……,a3}是不同属性(1≤k≤n)的属性值集。如果一个样本具有该模式中给出的所有属性的值就说它符合这个模式 P。对于规则 R: P→C ，符合模式 P 且类标识是 C 的数据样本的个数叫做对规则 R 的支持度，用 Sup(R)表示。符合模式 P 且类标识是 C 的数据样本数和类标识是 C 的样本总数的比率叫做 R 的可信度，用 Conf(R)表示。这里设定最小支持度 Minsup = 2，最小可信度 Minconf =60%

首先统计每个页面的支持度，产生候选 1 引用集 C1 = {A, B, C, D, E, F, G, H, I, J, K} ，将满足最小支持度 1 项集的加入到 L1 中，即 L1 = { A, B, E, F, I}；下一个循环根据 L1 生成 C2 = { AB, AE, AF, AI, BA, BE, BF, BI, EA, EB, EF, EI, FA, FB, FE, FI, IA, IB, IE, IF}，将满足最小支持度的 2 项集加入到 L2 中，生成 L2 = {AB, AI, BE, EF} ；下一个循环根据 L2 生成 C3 = {ABE, BEF}，将满足最小支持度的 3 项集加入到 L3 中，生成 L3 = {ABE, BEF}；下一个循环根据 L3 生成 C4 = {ABEF}，将满足最小支持度的 4 项集加入到 L4 中，生成 L4 = {ABEF}；直到发现 L5 为空，循环结束。

### 3.3　用户模式挖掘模块

用户模式挖掘模块分为三个部分：

(1) 按照不同的浏览模式对历史访问用户分类。

(2) 找出属于某个类中的用户的共同属性，即用户模式。

(3) 基于用户模式对新访问用户分类,预测其访问行为和兴趣偏好。

整个方法实现如下:首先将挖掘出的用户 MFP 输入到多维事务表 MDB 中,见表 2。数据表的结构为(UserID,A,B,C,D,items),其中 UserID 是每一个用户在数据库中的惟一标识。A(性别),B(年龄),C(职称)和 D(学历)是每个用户的不同属性,items 是每个用户的 MFP。每个 m = (UserID,A,B,C,D,items)中所包含的信息可以划分为两个部分:维部分(A,B,C,D)和项集部分(items)。

表 2　多维事务表 MDB

| UserID | A(性别) | B(年龄) | C(职称) | D(学历) | items |
|---|---|---|---|---|---|
| 01 | a1 | b1 | c1 | d1 | {A, B, D} |
| 02 | a1 | b2 | c2 | d3 | {A, B, D, H} |
| 03 | a2 | b3 | c2 | d3 | {A, B, D, E} |
| 04 | a1 | b2 | c3 | d3 | {A, B, D, F, G} |
| 05 | a1 | b2 | c1 | d3 | {A, B, D, G} |
| 06 | a2 | b3 | c1 | d2 | {B, E, G} |
| 07 | a3 | b1 | c2 | d1 | {A, D} |
| 08 | a3 | b2 | c2 | d2 | {C, E, F} |
| 09 | a2 | b1 | c3 | d2 | {C, E, F, G} |
| 10 | a1 | b1 | c1 | d3 | {C, E, F, G, H} |

3.3.1　按已挖掘的浏览模式对历史用户分类

浏览模式就是 MFP 中满足一定支持度的频繁遍历路径,按照浏览模式对用户分类即就是看用户的 MFP 包含哪个频繁遍历路径,那么用户就属于哪一类。在表 2 中有两个满足最小支持度(Minsup = 2)的浏览模式:一个是{A, B, D},其支持度 Sup = 5;另一个是{C, E, F},其支持度 Sup = 3。现将{A, B, D}定为 A 类,{C, E, F}定为 B 类。扫描表 MDB,得到属于 A 类的用户是:01,02,03,04,05 ;属于 B 类的用户是:08,09,10。用户 06,07 不属于任何类,因为他们的浏览行为不具有普遍性,可以忽略。最后得到用户的分类数据库,见表 3。

3.3.2　在用户分类数据库中寻找不同类中的用户模式

由于用户的属性集中各个属性之间没有任何联系,而用户访问路径中各个页面之间有方向和链接关系,所以设计一种基于 FP 树的关联分类算法来生成关联规则,以表 3 的用户分类表为例说明该算法的基本步骤。假设支持度的阈值是 3,可信度的阈值是 60% 。

表 3　用户分类表

| UserID | A(性别) | B(年龄) | C(职称) | D(学历) | 类别 |
|---|---|---|---|---|---|
| 01 | $a_1$ | $b_1$ | $c_1$ | $d_1$ | A |
| 02 | $a_1$ | $b_2$ | $c_2$ | $d_3$ | A |
| 03 | $a_2$ | $b_3$ | $c_2$ | $d_3$ | A |
| 04 | $a_1$ | $b_2$ | $c_3$ | $d_3$ | A |
| 05 | $a_1$ | $b_2$ | $c_1$ | $d_3$ | A |
| 08 | $a_3$ | $b_2$ | $c_2$ | $d_2$ | B |
| 09 | $a_2$ | $b_1$ | $c_3$ | $d_2$ | B |
| 10 | $a_1$ | $b_1$ | $c_1$ | $d_3$ | B |

首先扫描表 3,求出超过阈值支持度的属性值的集合 F。在属性 A 中 a1 出现 5 次,超过支持度阈值 3,因此加入集合 F 中。依此方法分别求出属性 B、C、D 中达到阈值的属性值,并加入 F 中。最后集合 F = {a1,b1,b2,c1,c2,d3},叫做频繁项集。其他属性值都没有超过支持度阈值。再对 F 中的属性值按照支持度大小排序,即:F - list = { a1,b2,b1,c1,c2,d3 }。

再一次扫描表 3 以构建 FP 树,如图 3 所示。FP 树是关于 F - list 的前缀树。对于表 3 中的每个记录,出现在 F - list 中的属性值是从记录中抽取并按照 F - list 排序的。对表 3 中第一个记录,抽取(a1,b1,c1)插入树中作为树的最左边的分枝。该记录的类标识和相应的计数器附加到路径中的最后一个节点上。

表 3 中的记录共享前缀。第二个记录支持 F - list 中的属性值(a1,b2,c2,d3)并和第一个记录共享前缀 a1,节点 a1 另外的分枝将被插入具有新节点 b2,c2 和 d3 的树中。第十个记录支持 F - list 中的属性值(a1,b1,c1,d3)并和第一个记录共享前缀(a1,b1,c1),d3 作为新节点被插入到树的末端。计数等于 1 的新类标识 B 也被插入到新的路径的末端。表 3 的最终 FP 树如图 3 a 所示。

分析完所有的记录构建出一个 FP 树后,通过把所有的规则分区成不重叠的子集生成分类规则集。在表 3 中生成的是 6 个子集:包含 d3 值的规则;包含 c1 不包含 d3 的规则;包含 c2 不包含 d3 的规则;包含 b2 但不包含 d3 或 c1 或 c2 的规则;包含 b1 但不包含 d3 或 c1 或 c2 的规则;仅包含 a1 的规则。算法依次地求出

这些子集。

为了求包含 d3 的规则的子集，先遍历所有具有属性值 d3 的节点，向上观察 FP 树搜索 d3 – 投影的记录。FP 树中有 4 个这样的记录，它们是（a1，b1，c1，d3）：B，（a1，b2，c2，d3）：A，（a1，b2，d3）：A，（a1，b2，c1，d3）：A，（c2，d3）：A。求表 3 中所有频繁模式的问题被归结为挖掘 d3 – 投影数据库中的频繁模式。在 d3 – 投影数据库中，由于模式（a1，b2，d3）出现三次，它的支持度等于所要求的阈值 3。同样，基于这个频繁模式的规则，（a1，b2，d3）→A 有 60% 的可信度（等于阈值），这是在所给的数据库投影中生成的唯一规则。

搜索完具有 d3 值的规则后，d3 的所有节点和它们相应的类标识被合并成为它们的 FP 树的双亲节点。FP 树缩小后如图 3b 所示。对于 c1 投影数据库同样重复前面的步骤挖掘剩下的规则集，然后是 c2，最后是 a1 投影数据库。在本例中，（a1，c1）模式虽然在整个表 3 中出现 3 次，但在 A 类中的支持度只有 2，规则（a1，c1）→A 的可信度只有 40% ，所以该模式不满足要求。对于其它模式也可以得出同样的结论，见表 4。因此，表 3 中生成的仅有的关联规则是（a1，b2，d3）→A，支持度等于 3，可信度 60% 。所以，属性集（a1，b2，d3）就是 A 类中的用户模式。当一个新用户同时具有（a1，b2，d3）这三个属性，那么他就属于 A 类，就能按照 A 类的浏览模式对其提供个性化服务。

表 4　挖掘出的用户模式

| 模式 | 支持度 | 可信度 | 类别 |
|---|---|---|---|
| a1，b2，d3 | 3 | 60% | A |
| a1，c1 | 2 | 40% | A |
| a1，c1 | 1 | 33.3% | B |
| b2，c2 | 1 | 20% | A |
| b2，c2 | 1 | 33.3% | B |
| a1，b1 | 1 | 20% | A |
| a1，b1 | 1 | 33.3% | B |

### 3.3.3 对新的访问用户分类

当有新用户访问时，分类算法从整个规则集中寻找符合该用户的模式，即判断该用户是否具有某个模式中属性值集合的所有属性。如果有，则属于这个类；没有，则不属于这个类。如果用户恰好属于哪个类，分类算法就把某个类标识分配给该用户。如果用户不属于其中的任何一个类，则将用户的属性值和所有用户模式的属性值集合进行比较，找出与其相似度最高的某个模式，然后将这个模式的类标识分配给该

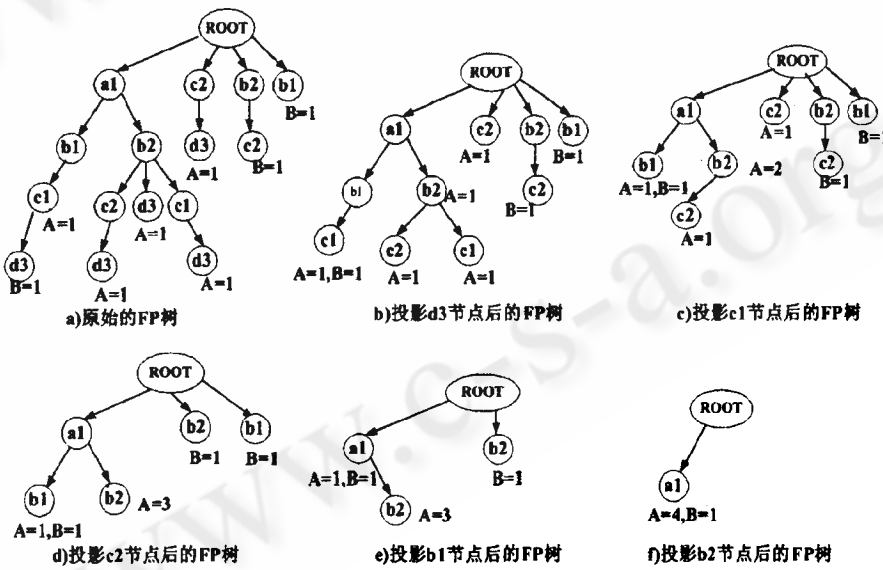


图 3　用户分类的 FP 树

用户。

要度量用户和模式的相似度，这里采用多维特征空间的欧氏距离作为度量标准。首先给出以下定义：

(1) 设每一个规则 $\{r_i:(y_{i1},y_{i2},\cdots,y_{im})\Rightarrow A, r_i\in Y, i=1,\cdots,n\}$ 都用向量 $r_i=\{y_{i1},y_{i2},\cdots,y_{im}\}$ 来表示。m 的值是一个规则的维数，n 是挖掘出的所有规则集 Y 中的规则个数。

(2) 设每一个新用户 $x_i\in X, i=1,\cdots,n$ 都用向量 $x_i=\{x_{i1},x_{i2},\cdots,x_{im}\}$ 来表示。m 的值是一个用户的维数(特征)，n 是所有新用户集合 X 中的用户个数。

(3) 某个规则的单分量 $y_{ij}$ 和某个用户的单分量 $x_{ij}$ 是一个特征或属性值。$y_{ij}$ 和 $x_{ij}$ $(j=1,\cdots,m)$ 是一个域 $P_j$。其中，$P_j$ 可以是不同类型的数据，可以是定量的或是定性的。具体分类如下：

定量特征：连续值(例如实数 $P_j\subseteq R$)，离散值(例如二元类型数 $P_j=\{0,1\}$)和区间值

定性特征：名义型或无序型，如一组颜色{R，G，B}；顺序型，如表示职称{教授，讲师，助教}

(4) 使用多维特征空间的欧氏距离计算相似度，$d(x_i,y_j)=(\sum_{k=1}^{m}(x_{ik}-y_{jk})^2)^{\frac{1}{2}}$，$x_i$ 表示一个用户，$y_i$ 表示一个规则。当 d 值越小时，距离越小，相似度越大。

(5) 如果某一个样本 $y_i$ 的属性 $y_{ij}$ 的值域是 $\{Y|y_{ij}\in Y\}$；另一个样本 $x_i$ 的属性 $x_{ij}$ 的值域是 $\{X|x_{ij}\in X\}$。则计算欧氏距离 d 时，若 $X\subset Y$，则 $(x_{ij}-y_{ij})=0$；若 $X\not\subset Y$，则 $(x_{ij}-y_{ij})=1$。

例如，已经挖掘出两个规则，一个是 $\{r_1:(y_{11},y_{12},y_{13},y_{14},y_{15})\Rightarrow A\}$，另一个是 $\{r_2:(y_{21},y_{22},y_{23},y_{24},y_{25})\Rightarrow B\}$。

各属性的含义和值见表 5：

表 5

| | $y_{i1}$(性别) | $y_{i2}$(年龄) | $y_{i3}$(职称) | $y_{i4}$(学历) | $y_{i5}$(工资) | 类别 |
|---|---|---|---|---|---|---|
| $r_1$ | 男 | 20 — 30 | 助教 | 本科 | 1000 — 2000 | A |
| $r_2$ | 男 | 30 — 40 | 讲师 | 硕士 | 2000 — 3000 | B |

现有一新用户 x1，其基本属性见表 6：

表 6

| | $x_{i1}$(性别) | $x_{i2}$(年龄) | $x_{i3}$(职称) | $x_{i4}$(学历) | $x_{i5}$(工资) |
|---|---|---|---|---|---|
| $x_1$ | 男 | 25 | 助教 | 本科 | 1500 |

分别计算欧氏距离得：

$d(x_1,r_1)=\sum_{k=1}^{5}(x_{1k}-y_{1k})^2)^{\frac{1}{2}}=(0+0+0+1+0)^{\frac{1}{2}}=1$ ($x_1$ 与 $r_1$ 的欧氏距离是 1)

$d(x_1,r_2)=\sum_{k=1}^{5}(x_{1k}-y_{2k})^2)^{\frac{1}{2}}=(0+1+1+0+1)^{\frac{1}{2}}=1.732$ ($x_1$ 与 $r_2$ 的欧氏距离是 1.732)

因为 $d(x_1,r_1)<d(x_1,r_2)$，所以用户 $x_1$ 被归于 A 类。

## 4 结束语

基于多维关联规则设计出的分类方法，得到用户的模式挖掘模块。较完美地解决了 Web 上数据的无序性和单结构化的问题：根据用户访问历史，可以了解用户的兴趣爱好，为不同用户群定制个性化的信息服务，从而增强网站的吸引力和亲和力；通过使用挖掘，可以提供网站服务效率全方面的信息，从而有助于平衡服务器的负荷，优化传输，减少阻塞，缩短用户等待时间，提高系统效率和服务质量；通过挖掘用户使用网站的信息，还可以帮助网站设计者确定如何修改网站结构。

## 参考文献


1 陈强、黄国兴，一种适用于关联规则挖掘的优化选样算法[J]，微型电脑与应用，2005(3)：11 - 12。

2 阮幼林、李庆华、杨世达，一种基于事物树的快速频繁集挖掘与更新算法[J]，计算机科学，2005(2)：210 - 212。

3 Mehmed Kantardzic 著，闪四清等译，数据挖掘：概念、模型、方法、和算法[M]，清华大学出版社，2003.8。

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