基于关联规则挖掘的个性化智能推荐服务*)

Intelligent Services of Personalized Recommendation based on Association Rules Mining

汪晓岩[1] 胡庆生[2] 庄镇泉[2]

(国家电力公司电力自动化研究院 南京210003)[1] (中国科技大学电子科学与技术系 合肥230026)[2]

Abstract This paper proposes an intelligent service method on personalized recommendation, which is based on association rules mining for Internet. To alleviate the phenomena of "information overload" and "information bewilderment" in Internet environment, the overall process can be divided into two components: offline part and online part. In offline, Web mining tasks can execute in the logs of Web service resulting in a user transaction file, and the frequent user transaction patterns are extracted by filtering with thresholds of support again, afterwards, constructing aggregating tree of user sessions. In online, the candidate URLs for recommendation can be determined by matching association rules in the aggregating tree with the current active session for the intelligent services of personalization recommendation. The experiments demonstrate that our approach is applicable and effective.

Keywords User transaction, Association rules mining, Intelligent services, Personalization recommendation

1 概述

随着 WWW 上的信息的爆炸性增长，用户的"信息过载"和"资源迷向"问题越来越突出。为了解决用户的信息过载和资源迷向问题，人们发展了许多智能推荐服务系统以及相关技术，帮助用户在 WWW 上快速定位、检索感兴趣的信息。其中，WebWatcher 系统[1,2]采用跟踪用户浏览 Web 站点的行为或者访问路径方法，学习用户的访问模式，将用户可能感兴趣的 Web 页在线推荐给用户。SiteHelper 系统[3]采用分析每一个用户已经访问的 Web 页，学习用户的兴趣模式，从用户感兴趣的 Web 页(浏览时间超过规定门限或者访问频次超过门限的 Web 页)提取关键词形成用户关键词表，然后，提供给用户，系统基于用户相关反馈技术为用户推荐其它的相关 Web 页。Footprints 系统[4]利用可视化技术，为用户提供 Web 站点的被频繁访问的路径。AVANTI 系统[5]利用自适应规则为每一组相同的用户访问模式实现定制化。

近年来，由于电子商务的发展，基于用户个性化模式的智能信息推荐服务日益受到人们的普遍关注。在电子商务中，系统通过跟踪用户的访问操作行为，了解供应商和消费者之间的关联关系，实现有针对性的个性化服务。基于 Web 挖掘一般分为基于信息内容的方法和基于用户访问行为的方法。目前这方面的研究成果已经出现，如，Schechter 等人[6]根据用户的访问路径模式预测用户未来可能的 HTTP 请求，用于代理服务器执行预取相关 Web 页放入其 Cache 中，以加快访问速度。Cooley 等人[7]和 Buchner 等人[8]利用数据挖掘技术从访问 log 文件中提取用户的访问模式，用于市场决策和智能推荐服务。Nasraoui 等人[9]采用聚类用户访问模式方法，预测用户未来的访问行为。

为了解决用户的"信息过载"和"资源迷向"问题，我们开展了面向 Internet 的个性化智能信息检索系统研究工作[10]，本文提出的基于 Web 挖掘的个性化智能推荐服务是我们工作的继续。其基本思想是，在离线方式下，从访问 log 文件中挖掘出用户事务模式，然后，采用多种 Web 挖掘技术和方法，如基于关联规则技术、聚类用户事务模式，实现在线方式的个性化智能推荐服务。

本文首先给出个性化智能推荐服务的系统结构，分别简要地介绍每个部分的作用。然后，按照个性化智能推荐系统结构划分为离线部分和在线部分，分别详细地讨论了离线部分和在线部分的关键步骤。本文重点讨论基于关联规则挖掘的在线个性化智能推荐服务，并且给出了实例，说明在线个性化智能推荐服务的有效性和可行性。最后，总结全文。

2 基于 Web 访问挖掘的个性化智能推荐服务结构

基于 Web 访问挖掘的个性化智能推荐服务过程分成两个部分实现：离线部分和在线部分，如图1所示。离线部分由数据准备和特定的访问挖掘任务组成，数据准备将 Web 服务器的访问 log 文件以及站点的相关文件生成用户文件和事务文件；特定的访问挖掘任务包括关联规则发现和 URL 聚类生成。在线部分利用离线部分生成的频繁项或者 URL 聚类，再根据用户的当前访问操作行为，动态地为用户推荐下一步访问操作。在线部分由个性化智能推荐服务 Agent 和 Web 服务器组成，Web 服务器通过各种方法，如重写 URL、暂存 Web 服务器的访问 log 文件，跟踪用户的访问操作；个性化智能推荐服务 Agent 通过分析用户当前访问操作，发现相关联的访问模式或者所属的 URL 聚类类别，计算生成推荐的 URL 集合作为用户下一步访问操作的候选集合。

Web 个性化服务主要包括：收集 Web 对象(如 Web 页)；识别操作对象的主体(如用户)；对象和主体的分类；对象之间、主体之间以及对象与用户之间的匹配；确定个性化服务的候选集。基于 Web 访问挖掘的个性化智能推荐服务，就是根据用户或者一组用户以往的访问行为规则，预测用户未来可能的行为，为用户提供个性化的智能推荐服务。下面，我们着

*)本文属于973国家重点基础研究发展规划项目(G1998030413)资助课题。汪晓岩 博士，高级工程师，从事信号处理，人工智能，数据挖掘和通信方面的研究工作。庄镇泉 博士生导师，从事信号处理，人工智能和数据挖掘方面的研究工作。胡庆生 博士后，副教授，从事信号处理、通讯系统的研究工作。

重讨论其中的基于关联规则挖掘的个性化智能推荐服务的实现方法。

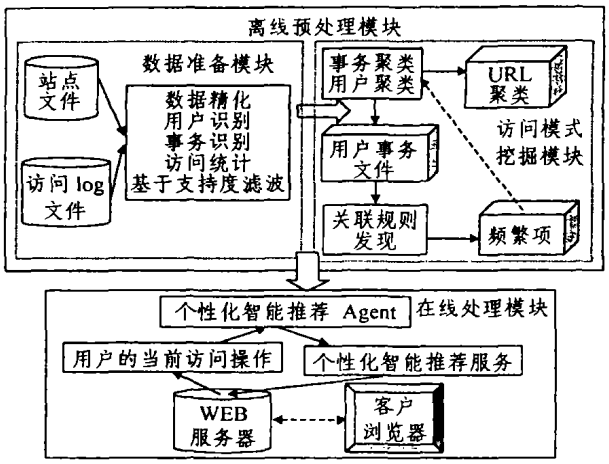

图1 基于 Web 访问挖掘的个性化智能推荐服务流程图

3 离线预处理

3.1 数据准备

由于访问 log 文件存在大量的"噪声"数据，必须对访问 log 文件进行预处理才能进行下一步的关联规则挖掘。预处理主要包括：数据精化、过滤和事务识别。这一步是所有个性化智能推荐服务都必须进行的步骤，我们按照 Cooley 等人[7]提出的方法，对访问 log 文件进行预处理，生成用户事务文件。

典型的访问 log 文件记录格式的每一条请求 log 文件记录包含：(1)客户主机的 IP；(2)时间戳；(3)请求的方法(如：GET、POST 等)；(4)请求文档的 URL；(5)HTTP 版本号；(6)返回码(即请求的状态：成功或错误码)；(7)传输的比特数；(8)引用 Web 页的 URL，Re_URL；(9)代理服务器(如 proxy 或客户端浏览器)Agent_ID。

过滤访问 log 文件涉及二个方面的内容：过滤无关项或冗余项；分析丢失的访问记录。过滤无关项或冗余项就是将对 Web 访问挖掘分析不产生影响的访问记录从访问项集中删除。例如，当访问文件名后缀为 gif、jpg、GIF、JPEG、jpg、JPG 的文件时，由于这些图片文件一般都嵌入在 Web 文档中，并不扩展访问路径，对 Web 访问挖掘分析不产生影响，可以将其删除。分析访问记录丢失是最困难的问题。由于客户端和服务器端使用了 cache 技术，Web 服务器的访问 log 文件无法记录访问储存在 cache 中的 Web 页的访问操作。通常 log 文件所记录的一次访问记录对应多个用户的多次访问。目前解决这一难题的主要方法有：cookies 技术、cache busting 技术和用户登记。cookies 是一种用户标识技术，当用户第一次访问 Web 站点时，Web 服务器或者浏览器动态地分配一个用户 ID 号，表示该次访问操作中代表该用户，在该用户随后的请求中，用户浏览器在发送用户请求的同时也附带该用户的 ID 号表示用户的类别。cache busting 技术实际上是阻止隐含访问存储在局部缓存的 Web 页，迫使浏览器或代理服务器每当接受到访问请求时，都重新从 Web 服务器处加载被请求的 Web 页。用户登记采用的是自愿原则，用户通常因为隐私权问题，提供的登记信息都是虚假信息。处理 cache 问题的方法包括利用站点的拓扑结构或引用 log 记录，再结合访问的时间信息，推理出丢失的访问记录。

利用上面讨论的方法，对访问 log 文件进行过滤、分析处理，生成适合关联规则挖掘的用户事务集合，对于具体的预处理步骤，这里就不详细讨论，可参考文[7]。下面我们着重讨论用户事务的关联规则挖掘方法。

3.2 相关定义

本文提出的基于关联规则的个性化智能推荐服务，主要针对特定的站点组织结构，因此采用最大前向访问路径辅助-内容事务是比较合适的选择。为了方便讨论，我们首先给出相关定义。

设 L 为用户访问操作集合，每一条访问记录 $l \in L$ 包括：用户主机 IP 地址，$l.ip$，用户 ID 号，$l.uid$，被访问 Web 页的 URL，$l.url$ 和访问时间戳，$l.time$。一般情况下，若没有用户的 ID 号，可以任意分配一个标识符给每个用户作为 ID。用户访问操作文件记录还有其它的域，如请求方法，POST 或 GET，传递文件的大小等。然而，这些域在事务模型中都没有使用。一般事务模型 t 定义为三元组，形式化表示为：

定义1(一般事务模型定义)

$t=(ip_t, \text{udi}_t, URL_t)$；

$URL_t=\{(l_1^t.url, l_1^t.time), \cdots, (l_m^t.url, l_m^t.time)\}$

其中，$1 \leqslant k \leqslant m, l_k^t \in L, l_k^t.ip=ip_t, l_k^t.uid=uid_t$ (1)

假设用户浏览 Web 页的平均引用时长(Reference Length)为 T_o，用户访问 Web 页的引用时长大于或者等于 T_o 的 Web 页定义为内容 Web 页，表示用户感兴趣的 Web 页；小于则定义为辅助 Web 页，表示该 Web 页起导航作用。因此，形式化的辅助-内容事务定义如下：

定义2(辅助-内容事务)

$t=(ip_t, \text{uid}_t, URL\_AC_t)$；

$URL\_AC_t=\{(l_1^t.url, l_1^t.time, l_1^t.len), \cdots, (l_m^t.url, l_m^t.time, l_m^t.len)\}$ (2)

其中，$1 \leqslant k \leqslant (m-1): l_k^t.len \leqslant T_o\ AND, k=m: l_k^t.len > T_o, l_k^t \in L, l_k^t.ip=ip_t, l_k^t.uid=uid_t$

执行在线推荐服务，必须进行当前用户访问操作序列与用户访问模式的匹配计算，这里的用户事务模式表示最大前向访问路径辅助-内容事务，每个用户事务为用户一次访问操作中相链接的 Web 页序列，从第一个 Web 页开始到返回 Web 页结束。前向访问 Web 页定义为当前请求的 Web 页不在已访问过的 Web 页集中。与此相对应，后向 Web 页定义为当前请求的 Web 页为已经访问过的 Web 页。每一个新的前向访问链接开始标志着一个新的事务开始。识别最大前向访问路径事务方法不是考虑每个 Web 页的持续时间长度，而是考虑当前 Web 页是否已经访问过，遇到已经访问的 Web 页，表示该事务结束。因此，最大的前向访问 Web 页为内容页，到达内容页所经过的 Web 页为辅助项。下面是最大前向访问路径(Maximal Reference)事务的形式化定义。

定义3(最大前向访问操作路径)

$t=(ip_t, \text{uid}_t, URL\_MRL_t)$；

$URL\_MRL_t=\{(l_1^t.url, l_1^t.time, l_1^t.len), \cdots, (l_m^t.url, l_m^t.time, l_m^t.len)\}$ (3)

其中，$l \leqslant k \leqslant i \leqslant m, l_k^t, l_i^t, l_i^t \in L, l_k^t.ip=ip_t, l_k^t.uid=uid_t, l_k^t.time < l_i^t.time, l_k^t \neq l_j^t, \forall\ j \neq k, 1 \leqslant j \leqslant m$

结合定义2和定义3，我们可以得出最大前向访问操作路径辅助-内容事务的定义，这里，我们就不具体给出。

为了方便讨论关联规则挖掘问题，我们给出 URL 集合定义和用户事务模式集合定义。用户事务模式集合所包含的

所有 URL，称为 URL 集合。用户访问操作集合所包含的所有最大前向访问操作路径辅助-内容事务，称为用户事务模式集合。由 URL 集合和用户事务模式集合构成的空间，我们称为 Web 事务空间。假设用户事务模式集合中包含 m 个用户事务模式，由 n 个不同的 URL 组成。我们给出的 Web 事务空间定义如下：

定义4(Web 事务空间) T 表示包含 m 个用户事务模式的集合，U 表示包含 n 个不同的 URL 组成的集合，分别表示为：$T=\{t_1,\cdots,t_m\}$，$t_i=\langle url_{i1}\cdots,url_{ik}\rangle\subseteq T$，$k\leqslant n$，$i\in[1,m]$和 $U=\{url_1,\cdots,url_n\}$；其中，$t_i\subseteq U$，$\forall\ i$，即 t_i 为 U 的非空子集。T 和 U 构成的空间称为 Web 事务空间 WS。

设 URL_i 为 U 的非空子集，URL_i 关于用户事务模式集合 T 的支持度表示 URL_i 中的 URL 出现在用户事务模式集合 T 中频度。形式化定义如下：

定义5(支持度定义) 设 $URL_i\subseteq U$，$U\subseteq WS$，URL_i 关于 T 的支持度表示为：

$\rho(URL_i)=|\{t|URL_i\subseteq U\}|/|T|$。其中，$|T|$为集合 T 的基数。

定义6(频繁项集) 设 $t\subseteq T$ 为用户事务模式子集，对于每一个 $URL_i\subseteq t$，URL_i 关于 T 的支持度 $\rho(URL_i)=|\{t|URL_i\subseteq U\}|/|T|>\rho_{min}$，其中，$\rho_{min}$为最小支持度，我们称之为频繁用户事务子集，记为 t^F。

定义7(大项集) 设 $t_i\subseteq t^F$ $i\in[1,m]$为 T 中一个用户事务模式，且 $t_i^k\subseteq t_i$ 为 t_i 的访问长度 k 的子用户事务模式，由所有 t_i^k，$i\in[1,m]$构成的集合称为频繁项集 t^F 的 k 大项集。

有了支持度的定义，就可以给出 Web 事务空间的关联规则定义。

定义8(关联规则定义) 对于 $X,Y\subseteq WS$，关联规则表示为 $X^{\rho_{min}}\Rightarrow^{\sigma_{min}}Y$，其中，$\rho_{min}$和 σ_{min}分别为最小支持度和最小置信度，且关联规则的支持度定义为 $\rho(X^{\rho_{min}}\Rightarrow^{\sigma_{min}}Y)=\rho(X\cup Y)/|T|$，置信度定义为 $\sigma(X^{\rho_{min}}\Rightarrow^{\sigma_{min}}Y=\rho(X\cup Y)/\rho(X)$，其中，$|T|$为用户事务模式集 T 的事务支持度的总和。

定义9(聚集树) 聚集树为一加权有向树 $G=(V,W(V),E)$，其中，V 为顶点集，对应 URL 集，E 为有向图的有向边，对应两个 Web 页之间的超链。$W(V)$为对应图中顶点的权值，表示对应顶点的 Web 页被访问的次数。

3.3 聚集树的生成算法

为了实现基于关联规则的个性化智能推荐服务，首先必须生成关联规则。生成关联规则的算法在事务数据库中主要采用 Apriori 算法[11,12]，该算法由两个主要步骤组成，第一步是发现所有的频繁项作为候选项，第二步从频繁项中生成关联规则。由于该算法在生成候选项时并不考虑候选项之间的顺序或时间顺序关系，因此生成的候选项存在很多无关项，即，实际上在用户事务模式集合中不存在子用户事务模式。本文采用基于聚集树生成关联规则有效地克服了这一缺点，提高了算法的效益。

基于聚集树的关联规则生成算法由两个主要步骤组成，即：(1)聚集树生成；(2)从聚集树上生成关联规则。

聚集树生成算法的输入为频繁的最大前向访问路径集，输出为聚集树。首先对访问 log 文件进行预处理，生成所有的用户事务，然后，采用基于支持度滤波算法过滤掉小于最小支持度的用户事务，最后，将获得的频繁的用户事务输入到聚集树生成算法。聚集树生成算法的代码如下：

算法1 聚集树生成算法
输入：频繁的用户事务模式集 Max_T^F；
输出：聚集树 Tree_T^F。
(1) $Tree\_T^F\leftarrow\phi$;
(2) *while* $Max\_T^F\neq\phi$ do
(3) begin
(4) $t_i\leftarrow Max\_T^F$;
(5) while $v_i^k\in t_i$ do
(6) begin
(7) if $v_i^k\notin Tree\_T^F$ then $Tree\_T^F\leftarrow v_i^k$; $Tree\_T^F\leftarrow(v_{i-1}^k,v_i^k)$; //添加新的路径；
(8) else 将 v_i^k 与 $Tree\_T^F$ 合并，v_i^k 的支持度与 $Tree\_T^F$ 中相应的节点的支持度相加；
(9) k=k+1;
(10) end;
(11) end; //while

为了讨论上述算法的特点，我们用实例说明如下。需要说明的是，在表1中，用频繁度代表了支持度，目的是更容易理解算法的执行过程。

表1 满足最小支持度的最大前向访问路径集的一个实例

事务序号	事务序列	支持度
1	A—B—C—D—E—F	11
2	A—G—C—D—H—F	10
3	^—A—B—C—D—E	9
4	^—A—B—C—D—F	9
5	A—I—C—D—J—F	7
6	^—A—I—C—D—J	7

在表1的实例中，用户事务模式集为最大前向访问路径辅助-内容事务集，最小支持度设为7。算法1对于每一用户事务模式 t_i，判别每一个被访问的 Web 页 $v_i^k\in t_i$ 是否已经在聚集树上，如果存在，则合并该节点，将两者的支持度相加作为该节点在聚集树中的新支持度；如果不存在，则在聚集树中创建新的节点增加新的分支。最后生成的聚集树如图2所示。

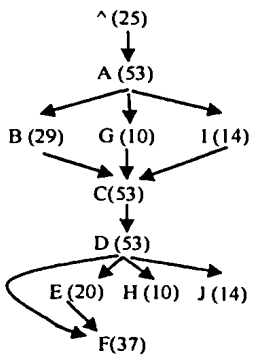

图2 表1实例的聚集树

4 在线个性化智能推荐服务 Agent

个性化智能推荐 Agent 是一个基于访问挖掘的 Web 个性化智能服务系统的在线模块，其任务就是根据当前用户访问操作，计算生成推荐集。推荐集是由与当前用户访问操作窗相匹配的访问操作模式组成，每一个访问操作模式都是根据用户当前访问站点的方式，分析发现的潜在有用的、相链接的 Web 页。这些被推荐的链接 Web 页被添加到用户当前已经访问过的 Web 页的后面。

一般来讲，确定推荐集需要考虑如下因素：

(1)针对当前用户的访问操作窗，确定与每一个频繁项的

匹配准则；

(2)确定作为候选的 Web 页是否为当前用户已经访问过的 Web 页；

(3)确定当前用户访问操作窗是否为推荐的访问操作序列部分匹配；

(4)与当前用户访问操作窗中 Web 页存在物理链接的 Web 页作为推荐的候选 URL。

利用固定大小的滑窗滑动覆盖当前的用户访问操作序列，是实现在线个性化智能推荐服务的一个有效方法。滑窗内的当前用户服务操作序列随着访问进程的进展，不断地向前更新。例如，假设滑窗大小为3，当前滑窗内的用户访问操作序列为〈A,B,C〉，当用户访问了 D 之后，新的滑窗内的用户访问操作序列变成了〈B,C,D〉。这样处理对于个性化智能推荐服务是非常有意义的，因为采用过长的当前用户服务操作序列，在与频繁项匹配操作时，很难获取非常多的信息，即相匹配的项很少。短序列能够获得非常多的相匹配的项，从推荐服务意义上，这是很有价值的。一般情况下，如果滑窗大小为 n，那么只有最近被访问的 n 个 Web 页对推荐集产生影响。然而，滑窗大小对推荐集的影响程度还没有得到有效的解决，是值得研究的问题。实验证明[13]，平均访问长度作为滑窗大小是一种比较好的选择。

从信息量或者新颖性的角度看，推荐一个物理链接远离当前用户访问操作的 Web 页是优先考虑的对象。为此，我们定义一种链接距离因子表示这种选择策略。

物理链接路径长度由表示站点拓扑结构的有向图确定。有向图中的每一个节点代表了站点中相应的一个 Web 页的 URL。如果 Web 页 X 到 Web 页 Y 存在一条物理链接，则相应的节点 X 到节点 Y 存在一条有向边。两个 URL(即 u_1 和 u_2)之间的物理链接路径距离定义为：在站点有向图中，从 u_1 到 u_2 的最小访问路径长度。

定义10(链接距离因子) 给定聚集树有向图 $G=(V,E)$，其中，$V\subseteq U$。设 s 为当前用户的访问操作序列，$u\notin s$ 为当前推荐的一个 URL；$dist(u,s,G)$ 表示 u 到 s 中的 URL 之间的最小物理链接路径距离，u 关于 s 链接距离因子定义为：

$$idf(u,s)=\log(dist(u,s,G))+1 \qquad (3)$$

如果 u 属于当前用户的访问操作，则定义距离因子为0。我们对距离取对数计算，目的是在下面讨论计算推荐度时，不要过度影响置信度对推荐度因子的贡献。

5 基于关联规则的个性化智能推荐服务算法

计算推荐集的一个有效方法，就是直接利用离线方式下获得的频繁访问模式。在预处理阶段识别出用户事务(即，最大前向访问路径辅助-内容事务)后，利用最小支持度过滤掉不频繁的项，然后，利用生成的频繁用户事务集合产生聚集树。基于关联规则的个性化智能推荐服务算法首先从聚集树发现匹配用户当前访问操作路径的关联规则，然后，再根据推荐度因子的大小确定推荐项，推荐度因子定义为关联规则的置信度乘以距离因子。

为了能够实现在线推荐，有效、实时地跟踪用户的访问操作，采用了滑窗采样获取当前用户的访问操作路径的方法。假设滑窗大小为 W，关联规则集实际上是从 W 大项发现 W+1 大项集，换句话说，就是用滑窗覆盖的长度为 W 的当前用户访问操作路径去匹配聚集树上的子访问路径，获取所有长度为 W+1的频繁子访问路径。

下面给出基于聚集树的关联规则发现算法：

算法2 基于聚集树的关联规则发现算法
输入：当前滑窗覆盖的用户访问操作路径 S_W，聚集树 $Tree\_T^F$；最小支持度 ρ_{min}，最小置信度 σ_{min}；
输出：关联规则集 GL；
(1) GL←ϕ；
(2) 从聚集树 $Tree\_T^F$ 中发现匹配 S_W，且长度为 W+1的访问路径 S_{W+1}的候选大项集；
(3) for 对于第 i 个候选大项 $S'_{W+1}\in S_{W+1}$ do
(4) begin
(5) if $\rho(S'_{W+1})\geqslant\rho_{min}$ then
(6) begin
(7) 计算关联规则 $S_W\Rightarrow S'_{W+1}$的置信度 $\sigma(S_W\Rightarrow S'_{W+1})$；
(8) if $\sigma(S_W\Rightarrow S'_{W+1})\geqslant\sigma_{min}$ then $GL\leftarrow S_W\Rightarrow S'_{W+1}$；
(9) end; //if
(10)end; //for

需要说明的是第五步计算支持度 $\rho(S'_{W+1})$，对于每一个候选大项 $S'_{W+1}\in S_{W+1}, i\in[1,|S_{W+1}|]$，其支持度定义为 $\rho(S'_{W+1})=MIN_k\{\rho(v'_k)\}$。其中，$v'_k\in S'_{W+1}, k\in[1,|S'_{W+1}|]$，$|*|$ 表示集合的基数或序列的长度。

我们以图2为例子，说明算法2的计算过程，设滑窗大小为 W=2，如果当前滑窗所覆盖的用户访问操作路径为 $S_2=\langle C,D\rangle$，则获取的大小为3的候选访问路径有4条，表示为 $S_3=\{\langle C,D,E\rangle,\langle C,D,F\rangle,\langle C,D,H\rangle,\langle C,D,J\rangle\}$，其相应的支持度和关联规则的置信度的计算结果如表2所示。

表2 例图2的关联规则的支持度、置信度和距离因子的计算结果

候选访问路径	支持度	关联规则	关联规则支持度	关联规则置信度	距离因子
$S_3^1=\langle C,D,E\rangle$	$\rho(S_3^1)=20$	$S_2\Rightarrow S_3^1$	$\rho(S_2\Rightarrow S_3^1)=20/81=0.25$	$\sigma(S_2\Rightarrow S_3^1)=20/53=0.38$	1
$S_3^2=\langle C,D,F\rangle$	$\rho(S_3^2)=37$	$S_2\Rightarrow S_3^2$	$\rho(S_2\Rightarrow S_3^2)=37/81=0.46$	$\rho(S_2\Rightarrow S_3^2)=37/53=0.70$	1
$S_3^3=\langle C,D,H\rangle$	$\rho(S_3^3)=10$	$S_2\Rightarrow S_3^3$	$\rho(S_2\Rightarrow S_3^3)=10/81=0.12$	$\rho(S_2\Rightarrow S_3^3)=10/53=0.19$	1
$S_3^4=\langle C,D,J\rangle$	$\rho(S_3^4)=14$	$S_2\Rightarrow S_3^4$	$\rho(S_2\Rightarrow S_3^4)=14/81=0.17$	$\rho(S_2\Rightarrow S_3^4)=14/53=0.26$	1

如果设 $\rho_{min}=0.20$，$\sigma_{min}=0.20$，则获得的关联规则集为：$GL=\{(S_2\Rightarrow S_3^1),(S_2\Rightarrow S_3^2)\}$。

在获得关联规则集后，下一步根据推荐度因子，计算确定推荐集，具体算法如下：

算法3 基于关联规则的推荐集生成算法
输入：关联规则集合 GL；最小推荐度因子 $SCORE_{min}$；
输出：推荐集 Recommend。
(1)Recommend←ϕ；i=0；
(2)while $S'_{W+1}\leftarrow GL$，AND $GL\neq\phi$ do
(3)begin
(4) for $u'_i\in S'_{W+1}$ do
(5) begin
(6) 计算距离因子 $idf(u'_i,S_w)$；
(7) 计算 $SCOCE(u'_i)=\sigma(S_W\Rightarrow S'_{w+1})*idf(u'_i,S_w)$；
(8) if $SCORE(u'_i)\geqslant SCORE_{min}$ then Recommend←u'_k；
(9) end；
(10) i=i+1；
(11)end；

基于关联规则的个性化智能推荐服务很适合实现用户智能接口的个性化智能信息检索主动服务，也适合于 Web 服务器站点的个性化自适应服务。通过上面的小例子分析，我们已经理解了基于关联规则挖掘的个性化智能推荐服务的核心思想。为了更进一步说明该方法的有效性和可行性，我们以一个

真实的访问 log 文件为例，测试我们提出的算法的性能。

6 实验结果与讨论

实验数据取自 Hyperreal 站点(http://www.Hyperreal.org)的服务器的实际访问 log 文件，总的数据量是二个月的访问 log 记录(1997.9～10)。经过预处理，获得大约1200个用户事务模式，2500个不同的 URL。用这些预处理后，我们进行基于关联规则挖掘的个性化智能推荐服务试验。

首先是滑窗大小的选择问题，统计分析实验样本数据得出平均最大前向访问路径长度为2.7，因此，我们选择滑窗大小2。为了获得充分多的候选项，支持度和置信度门限不宜选择太高。这里我们设定支持度门限为25%，置信度门限设定为0.1。

为了评估实验的推荐服务的信息新颖程度和有效性，实验采用用户初始进入站点的三种不同方式：从站点主页进入；从站点的相关主题的主页进入；从用户浏览器的"Bookmarks"收藏的兴趣页进入。实验结果如表3所示。

表3 三种初始进入方式的实验结果比较

用户开始的 Web 页	第一次推荐		第二次推荐		第三次推荐	
	推荐集数量	最高推荐度值	推荐集数量	最高推荐度值	推荐集数量	最高推荐度值
/(站点主页)	25	0.39	20	0.26	10	0.42
/music(相关主题页)	17	0.41	12	0.50	8	0.28
/categories (收藏 Web 页)	16	0.62	8	0.70	3	0.50

表4 用户采用相关主页初始进入站点的第一次生成的推荐集结果

滑窗用户访问操作	推荐集	推荐度因子
/music 支持度因子 =3.05%	/categories	0.41
	/guide	0.40
	/manufacturers/	0.38
	/manufacturers/Roland	0.33
	/manufacturers/Korg	0.33
	/machines/categories	0.30
	/manufacturers/Paia	0.29
	/manufacturers/Yamaha	0.29
	/categories/manufacturers	0.28
	/categories/drum-machines	0.25
	/manufacturers/Ensoniq	0.25
	/manufacturers/Octave	0.23
	/categories/DIY	0.23
	/manufacturers/Kawai	0.20
	/manufacturers/Moog	0.18
	/categories/images	0.18
	/categories/software	0.17
		0.15
		0.14

表5 用户采用相关主页初始进入站点的第二次生成的推荐集结果

滑窗用户访问操作	推荐集	推荐度因子
/music /machines/ categories 支持度因子 =5.15%	/features	0.50
	/categories/drum-machines	0.49
	/categories/manufacturers	0.48
	http://bong.com/machines/categories/	0.42
	/manufacturers/Paia	0.33
	/manufacturers/Yamaha	0.32
	/manufacturers/Roland	0.32
	/manufacturers/Ensoniq	0.32
	/manufacturers/Kawai	0.28
	/manufacturers/Moog	0.23
	/categories/images	0.23
	/categories/software	0.17
		0.15
		0.14

表6 用户采用相关主页初始进入站点的第三次生成的推荐集结果

滑窗用户访问操作	推荐集	推荐度因子
/music /machines /categories /manufacturers /Roland 支持度因子 =4.0%	/features/first-synth.html	0.28
	/manufacturers/Kawai	0.26
	/manufacturers/Moog	0.26
	/manufacturers/Roland/MKS/MKS	0.24
	/manufacturers/Roland /MKS/Roland-MKS	0.24
		0.23
	/manufacturers/Roland /TR-606/samples	0.19
		0.18
	/manufacturers/Roland/MKS/ /Roland/TR-808	0.13

从表3结果比较可以看出，用户初始进入站点的方式不同，对生成推荐集的影响是比较大的，随着推荐过程的深入，推荐集越来越小。对于固定大小置信度，最后推荐集有可能为空集。如果出现这种情况，可以调整置信度门限的方法克服。

由此可见，基于关联规则挖掘的个性化推荐服务能够快速帮助用户定位到感兴趣的主题上，但是，从获取新的信息角度看，其性能不是很好。此外，随着推荐服务的深入，当前推荐集有可能丢失部分上一次推荐集中用户感兴趣的项，为了说明此问题，我们以用户用相关主题也初始进入站点为例，具体结果如下：

我们从表4到表6三次推荐过程所形成的推荐集可以看到，用户对分类"Roland"感兴趣，整个推荐过程很快帮助用户定位到感兴趣的 Web 页，减少了许多中间导航 Web 页，起到了所谓的"短接"作用。但是，上一次推荐所形成的推荐集中，许多用户感兴趣的项目可能在随后的推荐集中丢失。例如，用户可能感兴趣的/manufacturers/Paia 和/manufacturers/Yamaha 等项都丢失了。但是，另一方面，却增加了有关分类"Roland"的进一步信息项。从这一点可以看出，基于关联规则挖掘的个性化智能推荐服务显示了快速、准确的特点。

总结 从访问 log 文件中挖掘用户访问站点的用户事务模式，实现个性化智能推荐服务是一个新兴的研究方向。个性化主动服务在电子商务系统中以及在 WWW 上实现个性化信息检索都有很大的实用意义。

本文所讨论的基于 Web 访问挖掘的个性化智能推荐服务是一种新尝试。从站点服务器的访问 log 文件中挖掘相似的用户访问模式是获取用户知识的有效手段，实验结果初步显示，本文提出的基于关联规则挖掘的个性化智能推荐服务方法的尝试是有效的和可行的。

参考文献

1 Joachims T, Freitag D, Mitchell T. WebWatcher: A Tour Guide for the World Wide Web. In: Proc. of the 15th Intl. Joint Conf. on Artificial Intelligence IJCAI-97, 1997. 770～775

2 Armstrong P, et al. Webwatcher: A learning apprentice for the world wide web. In Working Notes of the AAAI Spring Symposium: Information Gathering from Heterogeneous, Distributed Environments, Stanford University, AAAI Press, 1995. 6～12

3 Ngu D S W, Wu X. SiteHelper: A localized agent that helps incremental exploration of the world wide web. In: 6th Intl. exploration of the World Wide Web Conf., Santa Clara, CA, 1997

(下转第86页)

6)}。

对于每一个 $P\subseteq Q$ 确定一个对论域 U 的划分 U/I_P，在相应的等价类中的对象关于属性 P 有相同的描述，如，对于 $P'=\{A_1,A_2\}$，$U|I_{P'}=\{\{1\},\{2,3,4\},\{5\},\{6\}\}$。这样，{1}，{2,3,4}，{5}，{6}是 P'-基础集。

若选属性集为 $P=\{A_1,A_2,A_3\}$，则总体评价为"好"的学生的近似集 $X=\{1,3,6\}$，因为 $U|I_P=\{\{1\},\{2,3\},\{4\},\{5\},\{6\}\}$，于是

$\underline{P}(X)=\{1,6\}$，$\overline{P}(X)=\{1,2,3,6\}$，$B_{n_P}=\{2,3\}$

以上结果表明 P-边界集 $B_{n_P}(X)$ 非空，学生2和学生3属于同一 P-边界集，但学生2的评价为"完全坏"，而学生3的评价为"好"。基于属性 P 的信息，我们可以看出：

① 从 X 的 P-下近似得知，学生1和学生6一定属于"好学生"的集合；

② 从 X 的 P-上近似得知，学生1、2、3和学生6可能属于"好学生"的集合；

③ 从 X 的 P-边界得知，学生2和学生3是"好学生"的集合中不确定成员。

在属性集 Q 中，如果把条件属性 $C=\{A_1,A_2,A_3\}$ 和决策属性 $D=\{A_4\}$ 加以区别，则数据表就可以视为决策表。下面给出经过约简后的决策规则：

(1)如果 $f(x,A_1)=$好，则 $f(x,A_4)=$好； (1,6)

(2)如果 $f(x,A_1)=$坏，则 $f(x,A_4)=$坏， (4)

(3)如果 $f(x,A_1)=$中等，且 $f(x,A_2)=$好，则 $f(x,A_4)=$坏； (5)

(4)如果 $f(x,A_1)=$中等，且 $f(x,A_2)=$坏，则 $f(x,A_4)=$好或坏。 (2,3)

由上述规则不难看出，规则(1)，(2)，(3)只有一个单义结果，所以它们是确切的规则，但规则(4)的结论有两个结果，这是因为(4)是一个近似规则。

最后，我们把计算出的近似质量、在属性 C 中的属性所有子集的相互影响指标 I_S 和 I_B 及 Möbius 表示列在表2中。在表2中的第三列相应于 $\{A_1,A_2,A_3\}$ 的负值将作为在三个属性联合贡献中信息冗余的度量。在第四列中的前三个值是 Shapley 值，且可以解释为在粗糙近似中相应属性重要程度的度量，其中 A_2 和 A_3 是互补的，而 A_1 和 A_2 与 A_1 和 A_3 是可以互相替代的，这列中相应影响指标的负值说明 A_1,A_2,A_3 间有一个冗余。对于第五列也可类似地做出解释。

表2 近似质量、Möbius 表示、相互影响指标

属性	质量	Möbius 表示	Shapley	Banzhaf
$\{A_1\}$	0.5	0.5	0.44	0.5
$\{A_2\}$	0	0	0.11	0.17
$\{A_3\}$	0	0	0.11	0.17
$\{A_1,A_2\}$	0.67	0.17	−0.17	−0.17
$\{A_1,A_3\}$	0.67	0.17	−0.17	−0.17
$\{A_2,A_3\}$	0.5	0.5	0.17	0.17
$\{A_1,A_2,A_3\}$	0.67	−0.67	−0.67	−0.67

结束语 本文在叙述有关粗糙集的基本概念的基础上，重点介绍用模糊测度来度量粗糙质量的基本方法：将对策论的 N 人合作对策的 Shapley 值及 Banzhaf 值作为度量多属性决策的个体属性之间的相互影响的指标，为多准则决策分析提供了新的有效方法。

参考文献

1 Greco S, Matarazzo B, Slowinski R. Rough sets theory for multicriteria decision analysis. European Journal of Operational Research, 2001, 129(1): 1～47

2 Pawlak Z. Rough sets and Fuzzy sets. Fuzzy Sets and systems, 1985, 17(1): 99～102

3 Yao Y. A comparative study of fuzzy and rough sets. Information Sciences, 1998, 109: 227～242

4 王国胤. Rough 集理论与知识获取[M]. 西安：西安交大出版社，2001

5 曾黄麟. 粗集理论及应用[M]. 重庆：重庆大学出版社，1996

6 刘清. Rough 集及 Rough 推理[M]. 北京：科学出版社，2001

7 张文修，吴伟志，梁吉业，等. 粗糙集理论与方法[M]. 北京：科学出版社，2001

(上接第83页)

4 Wexelblat A, Maes P. Using history to assist information browsing, RIAO'97: Computer-assisted information retrieval on the Internet, Montreal, 1997

5 Fink J, Kobsa A, Nill A. User-oriented adaptivity and adaptability in the AVANTI project. Designing for the Web: Empirical studies, Microsoft Usability group, Redmond, WA

6 Schechter S, Krishnan M, Smith M D. Using path profiles to predict HTTP requests. In: Proc. of 7th Intl. World Wide Web Conf. Brisbane, Australia, 1998

7 Cooley R, Mobasher B, Srivastava J. Data preparation for mining World Wide Web browsing patterns. Journal of Knowledge and Information Systems, 1999(1)

8 Buchner A. Mulvenna M D. Discovering internet marketing intelligence through online analytical Web usage mining. SIGMOD Record, 1999, 27(4)

9 Nasraoui O, et al. Mining Web access logs using relational competitive fuzzy clustering. To appear in the Proceedings of the Eight International Fuzzy Systems Association World Congress, Aug. 1999

10 汪晓岩，胡庆生，李斌，庄镇泉. 面向 Interent 的个性化智能信息检索. 计算机研究与发展，1999. 9

11 Agrawal R, Srikant R. Fast algorithms for mining association rules. In: Proc. of the 20th VLDB Conf. Santiago, Chile, 1994. 487～499

12 Savasere A, Omiecinski E, Navathe S. An efficient algorithm for mining association rules in large databases. In: Proc. of the 21th VLDB Conf. Zurich, Switzerland, 1995. 432～443

13 Yan T, Jacobsen M, Garcia-Molina H, Dayal U. From user access patterns to dynamic hypertext linking. In: Proc. of the 5th Intl. World Wide Web Conf. Paris, France, 1996