# Internet 上支持高质量 E-Services 的个性化技术的研究*)

Personalization Techniques for High Quality E-Services on Internet

于 戈 王大玲 鲍玉斌 王 丹 杨晓春 宋宝燕 王国仁

(东北大学计算机科学与工程系 沈阳110006)

**Abstract** To support high quality E-Services on Internet, this paper proposes a zero-input personalization CMR approach integrating the techniques of data mining, rule resolution, and information integrating, concerning personal data collecting technique, Web data warehousing technique, personalization-oriented Web mining technique, personalization rule resolution technique, and personalization service recommending technique. The paper also gives the design of a CMR based personalization middleware system SmartWeb.

**Keywords** Personalization, Web mining, Data integrating

## 一、前言

随着电子商务、远程程教育等 Internet 上 E-Service 应用的日益普及，大量网站不断涌现，对网上业务量的争夺变得更加激烈。网站业务成败的关键之一在于网站提高服务的质量，即如何为恰当的用户，在恰当的时间里，方便地提供恰当的信息。而另一方面，对 Web 用户的调查表明，70%的 Web 用户认为现有网站难以提供有效帮助[1]。传统的网站采用的是"一对多"的批发方式，即对所有的顾客提供统一的界面、同样的内容。而新的竞争要求采用"一对一"的零售方式，即针对不同的顾客提供他所要求的内容和服务。例如，一个网上书店的顾客，依据不同职业、不同年龄、不同喜好等，不同的人所关心的内容是完全不同的。如同一个老练的店员会对不同的顾客采用不同的策略来推销产品，一个成功的商务网站也应尽可能具备这种自适应性智能。个性化技术正是针对这一问题应运而生的[2]。

个性化技术研究已成为当前学术界和产业界研究开发的热点，各大计算机公司和著名网站纷纷推出个性化服务和个性化开发工具。例如，雅虎公司推出的 My Yahoo!(my.yahoo.com)网站，允许用户选择所希望查看的主题以及主题中特定的栏目，并且可指定展示顺序，为自己构造出专用的网页，此后系统能够进行内容自动更新维护[3]。IBM Almaden 研究中心开发的中间件系统 WBI，通过提供可编程的 HTTP 代理服务器，支持 Web 服务器和浏览器之间的信息流转换，实现 Web 上个性化功能的开发[4]。美国南加州大学研究的 AI 侦察技术，利用人工智能技术，通过探查 Internet 上的应用情况，以改善 Internet 的可使用性[5]。

本文主要研究支持 E-Services 个性化的关键技术。讨论现有 Internet 个性化技术，提出一种新的个性化方法——CMR 方法及个性化支撑软件 SmartWeb 的设计，讨论 SmartWeb 实现中需要解决的关键技术和实现方法，最后，总结全文。

## 二、几种典型的个性化技术

个性化支持可分为初级和高级两种方式。初级方式是由系统在网页上提供选项(如 Check Box)，由用户对网页的形式和内容进行定制。高级形式是系统具有主动学习功能，通过概括和分析用户的个性文件(Profile)和行为，自动地实现某种程度的个性化。个性化支持除了在服务器端实现外，也可在浏览器端实现，如帮助用户整理感兴趣的站点、网页和链接。当前，支持自动个性化的技术可分为三类[6]：第一类方法是手工决策规则系统，它是由 Web 站点管理员，根据用户统计数、静态个性文件或会话(session)历史，制定若干规则，系统根据这些规则为特定的用户提供特定的内容及网页结构，例如，意大利米兰理工大学开发的 Torri 系统[7]及其 Firefly 等著名系统，这种系统容易实现；第二类方法是基于内容的过滤系统(Content based filtering)，它是通过分析用户历史上访问的内容，得出用户关心的内容和形式，向该用户推荐新的类似内容。第三类方法是协作过滤系统(Collaborative Filtering)，它不是通过访问内容的相似性，而是通过用户群的相似性进行内容推荐。

后两种系统是由系统自动生成动态的用户个性文

*)该课题得到教育部跨世纪优秀人才基金、骨干教师人才基金及科研教学奖励计划资助。

件，实现高级的个性化支持，典型的系统有 WebWatcher[8]。这些系统的特点通过对用户访问历史的分析，获得该用户的访问模式，再将该模式解释成内容需求，将其与 URL 结合，形成用户的个性化服务。但是，这两种方法仍需要用户一定程度的参与，必须反馈明确的要求，方可为其提供个性化服务。为了彻底减轻用户的负担，特别是广大初级用户的负担，提高 E-Services 服务质量的有效途径之一是实现零负担个性化，即在用户正常浏览时，不增加额外操作负担的前提下，实现面向用户个性化要求的网上信息发现与推荐。爱尔兰 College Dublin 大学 K. Nicholas 等将其定义为"零输入个性化(zero-input personalization)"，并给出了支持这一技术的基于信息提供者网页的推荐系统 RBPR[9]。如何根据用户的访问历史获取该用户的需求，是零输入个性化要解决的一个主要问题。RBPR 采用传统的机器学习方法，在一定程度上解决了这一问题，但它获得的访问规则形式单一。事实上，从用户访问历史数据中获取关于用户的访问信息也是一个知识发现的过程，因此，支持知识发现的技术之一——数据挖掘技术被应用，将会使知识发现的过程基于客观的数据，获得的用户访问知识更加全面、真实。另一方面，大多数现有系统仅简单提供与用户访问历史相关的 URL 地址列表，并不做进一步深入的处理。实际上，当我们浏览某个 URL 网页时，常遇到这样的问题，即网页中的部分信息是我们所关心的，而其它信息是与我们的查询主题不相关的。当系统提供的每个 URL 指向的网页都保存有大量与查询不相关的其它信息时，将对用户产生误导作用，这不但没有达到对个性化支持的目的，反而为用户的访问设置了障碍。

关于数据挖掘技术在 Web 中的应用，即 Web 挖掘的研究是目前数据挖掘技术研究的一个热点，但是，针对 Web 个性化支持方面，即如何通过 Web 挖掘技术的引入，获得用户个性化的需求，并提供给用户个性化的信息，则是一个新的研究课题。特别是零输入个性化技术，国内外仍处于起步阶段[10,11]。而且，国内外研究大多集中在规则挖掘和推荐引擎方面[6,12~14]，而对于两者的接口，即对生成的规则如何解释，并提供给推荐引擎的问题则比较模糊。零输入个性化问题在源数据的收集和预处理、Web 数据仓库的建立、基于该数据仓库的规则挖掘、规则解释以及信息推荐等问题上将更为复杂，单独采用上述报道中的某项技术尚不能很好解决。

综上所述，我们提出了采用 Web 挖掘技术获取用户需求，采用信息集成技术提供用户所需的信息，采用规则解析技术作为两者的接口，将这三者有机地集于一体的 CMR 方法与 SmartWeb 系统，以解决 Web 个性化信息服务问题。CMR 方法与 SmartWeb 系统的研究涉及到了如下关键技术：

1)实现零输入个性化的 Web 仓库数据模型和领域知识模型；

2)Web 挖掘所需信息源的收集和预处理；

3)自动获取用户的个性化需求的 Web 挖掘技术；

4)支持信息推荐的规则解释机制和解析算法；

5)面向相关 URL 列表的集成信息生成机制；

6)个性化信息的推荐模式。

## 三、CMR 方法与 SmartWeb 系统

本文提出的 Internet 环境上实现零输入个性化支持的关键技术，包括个性数据采集技术、Web 数据仓库技术、面向个性化的 Web 数据挖掘技术、个性化规则解析技术以及个性化服务推荐技术，并在此基础上开发一个个性化支持中间件软件系统 SmartWeb。SmartWeb 系统的体系结构如图1所示，它由 C(Collecting)、M(Mining)和 R(Recommending)三部分组成。C 完成用户个性数据采集和预处理，M 完成个性化规则的挖掘，R 实现个性化信息的集成和推荐。因此，我们将该综合型个性化技术命名为 CMR 技术。

1. 用户个性数据收集器。Server PDC(Personal Data Collector)为服务器端的收集器代理，主要从 Web 服务器的 Log 文件中获得访问路径等信息。Client PDC 为客户器端的收集器代理，从客户端获得关于用户查询的关键字等与内容相关的信息。由于该信息受用户隐私权和安全性保护的限制，我们采用 W3C 国际组织建立的 P3P(Platform for Privacy Preferences)协议[16]。

2. 用户个性数据的预处理器。根据领域知识(Domain Knowledge)，将收集器收集到的个性数据按用户分类，将对 Web 页的访问序列组成逻辑单元，以表征事务或用户会话，对于页面内容信息，进行关键字的分解及相关路径的匹配处理。最后，根据个性挖掘的要求，将各类数据统一以 XML 标准结构存储到数据仓库中。上述处理结果作为 Web 个性挖掘的数据源。关于数据源描述的元数据采用 RDF 格式存储。

3. 个性化信息挖掘器。参照领域知识，通过关联规则、序列模式、聚类分析、分类分析等挖掘算法，挖掘出用户需求的个性化信息，得到关键字与 Web 站点的关联规则、用户的分类规则、单个用户的访问模式，以及用户与站点之间的关联规则和分类规则等知识。主要方法有：

(1)通过对各用户访问的 URL 的关联规则挖掘，获得每个用户频繁访问的 URL 集合，得到 URL 的关联序列。在此基础上，将数据集中各项定义成多个 URL 站点序列，进行关联规则的挖掘，以便获得每个用户的频繁访问序列集合，进行路径分析；

(2)通过页面内容分类、概念层次树的建立及概念

提升，生成满足不同用户要求的个性化查询索引；

(3)通过用户聚类分析，将用户按浏览兴趣分成不同的簇，以便为各簇进行满足该簇要求的个性化的信息模式匹配；

(4)通过页面内容、站点分类及用户聚类，得到形如：

IF 用户=XXX&{站点$_i$=XXX$_i$}$^+$THEN {页面$_i$=XXX$_i$}$^+$

的规则，

另外，一部分挖掘结果，如用户聚类结果、频繁访问的 URL 集合等，又可作为另一种挖掘算法的输入信息。

4. 规则解析器。通过对各种规则进行分析，将规则解析成形如

(用户名，站点序列，站点与关键字匹配序列，站点与页面匹配序列)

的结构化形式并存储在用户的个性文件中，并建立各种索引，如按用户名的索引，以便向集成器提供每个用户访问行为频度及内容等。

5. 信息集成器。对数据挖掘后产生的结果进行分类，对于"URL 序列"或"站点与页面匹配序列"，可以直接"短路"到相应站点，获取信息；对于"URL 与关键字匹配序列"借助搜索引擎获得与关键字匹配的信息，并过滤掉与站点不匹配的信息，最后采用基于 XML 的信息集成技术，生成集成信息。对于信息获取机制的实现，采用移动代理技术。

6. 个性化信息裁剪器。将上述取得的 URL 路径信息或 XML 内容信息，裁剪制作成满足用户个性化要求的页面。

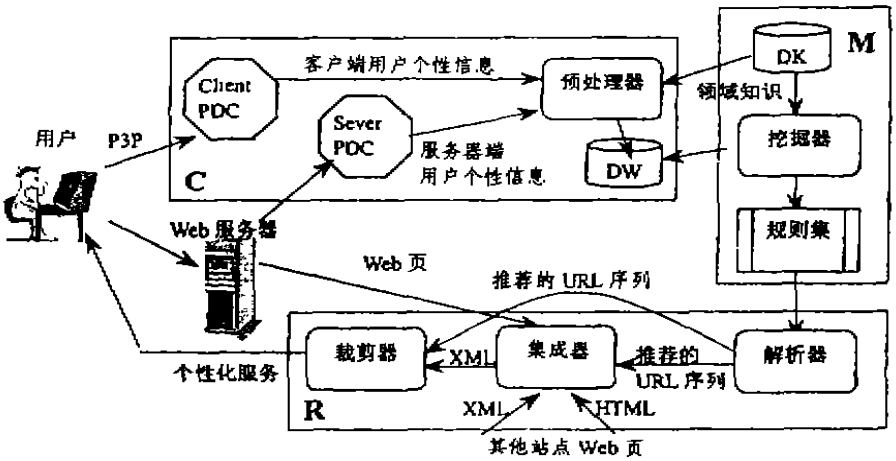


图1 SmartWeb 系统的体系结构

## 四、主要关键技术和实现方法

以下讨论 SmartWeb 系统实现过程中解决的主要问题。

1. Web 个性化信息的收集和预处理相关问题。其目的是为 Web 个性化挖掘提供数据源，涉及如下问题：

(1)目前 Web Log 信息质量较低。①典型的 Log 文件包含以下信息：IP、时间戳、Method、URL、HTTP 版本、返回码、传送的字节数、引用页 URL 和代理(Agent-浏览器和用户操作系统)。由于使用代理服务器和防火墙，用户 ID 通常不可用；由于使用局部(本地)缓存，导致 Log 文件不完整，即重复访问的页不被 Log 文件所记载。②在 Web 服务器和客户浏览器间需要用户注册或使用"cookies"，但用户为了隐私而不愿意注册或匿名注册。这些都给信息的收集和预处理造成很大困难，因此，SmartWeb 的预处理需要将多种渠道(Server PDC 和 Client PDC 等)收集到的信息进行分析和综合。

(2)面向用户的 Web 个性使用分析必须首先确认用户会话(user session)。确认用户会话有以下三种机制：①Web Server 提供 Cookies，②若服务器不提供 Cookies，需要每个浏览器注册 ID，③如果 Server 不提供 Cookies 或 ID，则使用 Host 地址确认用户会话。由此带来的问题是：来自同一个代理或防火墙的访问被认为是同一个会话，显然这是不正确的，SmartWeb 需要采用专门的鉴别机制。

(3)Web 数据仓库的设计和建立，用以描述数据源的元数据的管理，以及多维索引的建立和维护等问题。

2. 数据挖掘算法的研究与设计。Web 挖掘，特别是个性化的 Web 挖掘有其不同于一般数据挖掘的特

殊之处，主要包括：

(1)目前 Web 挖掘研究的目的大多是发现群体信息，而个性化 Web 挖掘更多地要求关于用户个体信息的规则。如何在一个具有多个用户访问信息的数据库中获得每个用户的模式是新的课题。显然，将数据库按用户名分成多个子库再对子库进行挖掘并非可取之策。我们的解决方法是，将现有算法做相应的改进，通过增加定义，在具有多个用户访问信息的数据库中，只需一次挖掘就能获得每个用户的个体信息；

(2)支持个性化的 Web 挖掘，不仅包括单一规则的发现，还包括复合规则(如聚类与关联规则、关联规则与分类规则等)的发现；

(3)目前一般数据挖掘算法大多针对静态数据源，而 Internet 网上用户不断加入和退出，浏览的信息也不断改变，这些都造成 Web Log 数据的频繁更新，数据量急剧增大。因此，信息收集、预处理和数据挖掘算法是针对这种动态数据源进行的，一方面，考虑一种适用于这种动态变化的存储结构，另一方面，设计针对这种存储结构的数据挖掘算法。我们采用 J. Pei 和 J. W. Han 提出的"Web Access Pattern Tree(WAP-Tree)及该结构的关联规则挖掘"方法解决此问题[15]；

(4)不同的数据挖掘方法需要使用不同格式的数据源，如 URL 的关联规则要求一个用户的一次浏览记录作为一个事务，将 WWW 上的每个站点的 URL 作为一个项目，而分类规则需要将 URL、用户名等作为条件属性，将页面内容或关键字作为决策属性。这些都要求在挖掘之前对数据仓库的挖掘对象进行转换。

3. 用户个性化信息的获取和处理，包括 Profile 记录的分解、组合以及所要展示信息的结构、内容等。个性文件包含两部分内容，即事实 Profile 和行为 Profile。前者包括一些统计信息和从事务数据中获得的信息，后者表征用户的行为，来自挖掘到的单一的和复合的规则。规则解析器将事实 Profile 和行为 Profile 的内容结合起来，组织成研究目标中制定的存储形式。

4. 个性化规则的解析与结构化存储，即如何将不同类型、不同结构的规则解释并转换成同样的结构。这里涉及到了规则库的设计和建立，个性索引的建立和维护等问题。

5. 基于用户使用记录的全自动模式匹配技术。通常，模式匹配由于涉及到大量的语义信息，必须要求有适当的人工干预，不可能实现完全的自动处理。但在零输入情况下，因不要求用户人工干预，因此我们采用基于用户已有的页面操作(即在一个页面中选择感兴趣的超链的过程)，分析用户历史上访问的引用页，进行自动模式匹配。

6. 支持个性化的信息推荐形式。包括访问模式的推荐和内容的推荐。访问模式推荐指用户感兴趣的 URL 列表，指导用户的上网行为；而内容推荐是较高级的个性化支持方式。根据自动模式匹配的结果，包括对数据挖掘后产生的相关 URL 列表进行分析，对 URL 列表中每个网页中的信息进行剪裁，过滤掉与用户访问主题无关的信息，最终经过对相关信息的集成处理，形成一个集成后的信息推荐给用户。

7. 基于 XML 的信息集成处理技术。推荐的 URL 列表的内容信息可能会存在异构性，主要体现在数据表示上的异构性，例如 XML、HTML 等格式，通过集成器对异构的网页信息进行处理，统一生成以 XML 形式表示的格式。

8. 支持个性化信息采集的移动代理技术，主要是要建立一个支持移动代理的迁移、恢复、运行、通信、终止等的运行环境。在此环境的支持下，移动代理运用挖掘得到的规则自主地进行信息采集和过滤。

SmartWeb 的实现，采用了在信息集成、数据仓库、数据挖掘、XML 数据处理、移动代理以及信息安全等方面的有关技术，具体方法如下：

1)将智能移动代理技术应用于 Web 用户当前访问信息的在线采集和历史访问信息的获取与收集，移动代理在收集信息时还需结合信息安全技术中的访问控制机制以保护用户的隐私权。必要时需采用 IBM 公司的 WBI 代理软件和 W3C 的 P3P 协议[16]，获取用户个性信息如点击流(Clickstream)。

2)将数据仓库技术应用于各种信息源信息的预处理过程中，采用视图维护技术、多维存取技术处理，即时更新的个性数据。

3)将传统的数据挖掘技术、Web 挖掘技术应用于支持零输入个性化的数据挖掘过程中。通过在数据预处理过程中将各种格式的数据转换成 XML 的方法提供个性挖掘所需的数据源，通过改进现有算法的方法实现个性化信息和复合规则等信息的发现。另外，在概念层次树的建立过程中引入领域专家知识。

4)将 XML 数据处理、文档数据库、SQL 技术以及相关的数学函数应用于挖掘结果的分析、整理中，形成具有个性化特征的 URL 推荐序列。

5)将信息集成技术、智能搜索技术、XML 语义分析技术以及移动代理技术应用于信息获取、信息过滤、信息分析以及信息的剪裁和制作过程。

**结束语** 随着电子商务、电子政务、网上教育等 E-Services 在我国的发展，研究 Web 环境下支持高质量 E-Services 的零输入个性化服务的有关技术，具有重要的理论意义和广阔的应用前景。我们提出了将 Web 挖掘技术、信息集成技术及规则解析技术三者有机地集于一体的 CMR 方法与 SmartWeb 系统，用以

解决 Web 个性化信息服务问题，在个性化技术的研究中作了新的探索。由于该领域的研究国内外都处于探索阶段，因此，我们认为，CMR 方法与 SmartWeb 系统在以下几方面有所创新：

·支持 Internet 环境中零输入个性化技术的、集数据仓库技术、数据挖掘技术、规则解析技术和信息集成技术于一体的体系结构。

·支持个性化需求的数据挖掘、Web 挖掘技术，包括数据源的存储结构及在该存储结构下的挖掘算法。

·Web 挖掘规则解析机制以及结构化存储模式转换机制。

·具有个性化特征的基于 XML 的信息集成和裁剪技术。

## 参 考 文 献


1 http://www.personalizaton.org/
2 Doug R Introduction:Personalized Views of Personalization. ACM Computer,2000,43(8):26～28
3 Manber U,Patel A,Robision J. Experience with Personalization on Yahoo! ACM Computer,2000,43(8):41～43
4 Malio P,Barrett R. Intermediaries Personalize Information Streams ACM Computer,2000,43(8):96～101
5 O'Leary D The Internet,intraners,and rhe AI renaisance lEEE Compuuer,Jan. 1997
6 Aggarwal C,Yu P. Data Mining Techniques for Personalization. lEEE Data Engineering Bulletin,2000,23(1):4～9
7 Ceri S,Fraternali P,Paraboschi S. One-ro-One Personalization of Data-Intensive Web Sites. In:Proc. of VLDB'2000
8 Joachims T,et al. WebWatcher:A Tour Guide for the World Wide Web. In,Proc. of the Int. Joint. Conf in AI(IJCAI97),Aug. 1997
9 Nicholas K,McKee J,Toolan F. Towards Zero-input Personalization:Referrer-Based Page Prediction. In:Proc. of AH 2000. 131～143
10 Mobasher B,Cooley R,Srivastava J. Automatic Personalization based on Web usage mining. ACM Computer,2000,43(8):142～151
11 Yu P. Data Mining and Personalization Technologies. In:Proc. of DASFAA. 1999. 6～13
12 王继成，等. Web 文本挖掘技术研究. 计算机研究与发展，2000,37(5):513～520
13 王实，等. Web 数据挖掘. 计算机科学，2000,27(4),28～31
14 陈宁，等. 数据采掘在 Internet 中的应用 计算机科学，2000,26(7):44～53
15 Pei J,Han J. Mining Access Pattern Efficiently from Web Logs,Conf. of PAKDD'2000
16 http://www.w3.org/


(上接第41页)

```
2) receiveShadow:
 [A message ("MoblieShadow",shadow) has arrived]
 receive(("MoblieShadow",shadow));
  if (shadow.shadowTTL != 0)
    if (shadow.homePlace != place.name())
    {
      shadow.currentPlace = place.name();
      shadow.List.add(shadow);
    }
    else
    {
      //shadow comes back home
      surrogateID = shadow.shadowed;
      suurogate = shadowList.find(surrgateID);
      shadowList.remove(surrogate);
      shadowList.add(shadow);
      shadow currentPlace = Null;
  }
 ;
3) shadowProxyPathTimeOut:
 [The timer triggered a (timer,shadow) message]
 receive((rimer,shadow));
 shadowList.remove(shadow);
4) terminateShadow()
 if (currentPlace != Null)
  //Shadow moved
    send(currenrPlace,("Terminate",shadowID));
  delete(rhis);
5) receiveTerminate.
 [A message ("Terminate",shadowID) has been received]
 shadow = shadowList.find(shadowed);
 if (shadow != Null)
 shadow terminareShadow();
```

**评价** 基本影子协议、层次影子协议、移动影子协议完整地构成了移动 Agent 的影子控制协议。影子协议通过检测影子是否存在，来探测孤儿。

通过路径信息的帮助找到 Agent 的 Proxy。按照 Proxy 指示的信息，给目标场所发送查找消息，然后目标场所检查 Agent 是否在当地，如果不在，继续到下一个场所去找，直到找到 Agent 的场所，即找到路径的终点，实现了 Agent 的定位。

Agent 的 TTL 计时器决定何时终止 Agent 的执行。

总之，影子协议可以方便实现 Agent 的定位、Agent 的终止和孤儿的探测。

## 参 考 文 献


1 Genesereth M, Ketchpel S. Software Agenr. USA,1994
2 Shaham Y. Agent-oriented Programming. USA,1993
3 Baumann V. Conrrol Algorithms for Mobile Agent. University of Sruttgart,1999
4 Jennings N, Wooldridge M. Agent-Oriented Sofrware Engineering. University of London, UK, 1999
5 EURESCOM. Project712: Intelligent and Mobile Agents and Their Applicability ro Service and Network Management. Germany. 1999