基于 Web 使用挖掘的个性化服务技术研究①

Research of personalization technologies based on web usage mining

崔 林 宋瀚涛 龚永罡 （北京理工大学计算机科学与工程系 100081）
陆玉昌 （清华大学计算机科学技术系 100084）

摘要：Internet 的快速增长导致了对个性化服务需求急剧增加，Web 使用挖掘正成为实现个性化系统功能的思想和方法的有价值的源泉。本文讨论了基于 Web 使用挖掘的 Web 个性化技术，并针对个性化系统的功能，介绍了相关数据采集和预处理技术及其在个性化系统中的应用。

关键词：个性化服务 Web 使用挖掘 推荐系统

1 个性化服务系统与功能

1.1 个性化服务系统

个性化服务是指针对不同用户提供不同的服务策略和服务内容的服务模式，其实质就是以用户需求为中心的 Web 服务。个性化服务通过收集和分析用户信息来学习用户的兴趣和行为，进而实现主动推荐服务。因此，通过网络提供的个性化服务不仅可以减轻用户" 信息过载" 的困境，而且可以帮助企业建立友好的客户关系。

创建个性化服务系统的一般步骤为：首先收集用户的各种信息，如注册信息、访问历史等；其次分析用户数据，创建符合用户特性的用户模型；最后结合用户特性，向用户提供符合其特殊需求的个性化服务。当用户对系统提供的服务做出响应或反馈时，系统根据反馈信息调整服务。通过用户与系统之间循环往复的交互，系统最终能够为用户提供个性化服务。

1.2 个性化服务系统的功能

个性化服务系统是通过提供的多种功能实现系统目标的。从目前实现的角度可以将个性化服务系统分为：记忆型、引导型、定制服务型和工作任务辅助支持型。

（1）记忆型。记忆型通过在系统中记录使用者的信息，当使用者再次登录该网站时，系统利用用户过去的历史数据，给用户必要的提示和帮助。具体功能包括：向登录用户致意；为用户建立个性化书签和分配用户个性化的存取权限等。

（2）引导型。引导型是指系统通过提供替代的浏览选项，协助引导使用者更快更容易地获取所寻求的信息。这类个性化服务不但能增加使用者的忠诚度，而且可以减轻用户在大型网站里所面临的" 数据超载" 和" 信息迷航" 问题。具体功能包括：向用户进行超链接的推荐；为用户导航。

（3）定制服务型。这类系统可以按照用户的知识、兴趣和偏好对网页的内容、结构和布局进行个性化设定，达到对数据负荷进行管理，使用户和网站的交互简单化和个性化。具体功能包括：个性化的网站布局设计；个性化的内容定制；个性化的超链接定制；个性化的定价和营销。

（4）工作任务辅助支持型。这类系统能按照用户特点，启动执行特殊的动作程序，给用户的工作辅助帮助和支持。这是最先进的个性化功能，可以在客户端或服务器端实现。具体功能包括：个性化的行动助理；个性化的疑问解答和个性化的谈判助手。

1.3 个性化服务与 WEB 使用挖掘

Web 挖掘技术是实现 Web 个性化服务的核心技术之一。Web 使用挖掘是从用户的网络行为中抽取用户感兴趣的模式。通过对用户浏览网站的使用数据收集、分析和处理，建立起用户行为和兴趣模型，这些模

① 国家 973 基础研究项目“WWW 上的数据集成、数据仓储及知识发现的有效算法与软件系统” 编号：G1998030414）

型可以帮助理解用户行为,改进站点结构以及为用户提供良好的个性化信息服务。由于个性化推荐所面临的关键问题是需要对大量非注册用户的行为模型进行深层理解,传统的协同过滤方法很难处理非注册用户的情况,Web 使用挖掘能较好处理这类问题;同时,借助于 Web 使用挖掘可以从传统的基于使用数据的静态建模转换到基于用户操作行为的动态建模,在系统里帮助改善用户的网络使用经验。因此,基于 Web 使用挖掘建立的个性化系统是实现良好个性化服务的一个有效途径。

WEB 使用挖掘一般包括:(1)数据收集;(2)数据预处理;(3)模式发现和评价运用几个阶段。

2 数据采集的个性化技术

2.1 数据采集技术

数据可信性是影响 Web 个性化服务质量的重要因素,准确的用户使用数据对识别用户、发现用户的兴趣很重要。数据采集阶段就是要根据系统要求,确定从何处采集用户的使用数据,识别出它们的内容和结构。Web 服务器,客户端以及代理服务器是目前三个主要数据来源。

2.1.1 服务器端数据

服务器端的数据主要包括:服务器日志文件、Cookies、用户显示数据输入和外部统计数据。

(1) 服务器日志文件。Web 服务器日志文件记载的是多用户访问单服务器数据,这些日志记录了用户对 Web 页面的存取情况,主要有两种文件格式:常规日志文件格式 CLF 和扩展日志文件格式 ECLF。由于网络高速缓存以及中间代理服务器存在造成 IP 地址的动态变化,从而使得准确提取用户数据并不容易,目前主要采用各种启发式方法帮助解决。此外,日志文件也可能造成对个人隐私的威胁(Broder,2000)。

(2) Cookies。Cookies 用来追踪用户浏览过的页面,通过 Cookies 在客户机器上储存一个信息,当用户下次访问该网站时,这个信息会送回到服务器,从而识别出用户。Cookies 也能储存其他类型的数据,如页面是否访问过,是否购买了产品等等。但当用户从不同的机器连接上网或一台机器几个用户上网时,利用 Cookies 判断就不准确。此外,用户也可能出于对隐私和安全考虑而拒绝接受 Cookies,不定期删除 Cookies,以及机器内对 Cookies 的容量限制等,都会对 Cookies 分析的准确性带来影响。

(3) 用户显示输入数据。用户提交的各种数据能较好反映用户的偏爱兴趣,对个性化服务的实现也非常有用。对用户输入的查询和检索词的分析要设法将检索词和网络的结构,内容以及关键词语义,领域背景知识相结合,才能较好发现用户的偏好。但仅依靠用户输入获取数据会增加用户负担,还需要借助其他方法隐式获取用户的使用信息。

(4) 外部统计数据获取。最后,可以从数据库第三方购买获得一些用户统计信息,但对于数据的隐私保护可能限制随意数据转移。

2.1.2 客户端数据

客户端数据记录了单用户访问多服务器的模式,客户端数据需要有专门的程序收集。早期是采用在客户端修改浏览器 获取客户端数据。目前客户端的数据主要依靠远程 agent 获得,这可以通过 Java 或 JavaScript (Shahabi et al.,1997 ,2001)实现。客户端的数据比服务器端的数据要更可靠,它们避免了高速缓存与 IP 地址误解问题,但获取客户端数据必须用户给予合作。

2.1.3 代理服务器和包侦测

代理服务器端记载的是多用户访问多服务器的访问模式。代理服务器也使用类似 Web 服务器的日志格式,记录 Web 页面请求和服务器的响应,使用这些日志可以了解用户在代理服务器后面的动作行为。然而,前面提到的高速缓存和 IP 地址误解问题在代理服务器数据中依然存在。包侦测是采用软件或硬件装置监视网络通信情况,如从 TCP/ IP 包中提取数据。包侦测的优点是能实时采集和分析这些在日志文件里难以获取的数据,非常有用。但出于安全考虑,电子商务数据会以加密格式传输,这给获取有用数据带来困难。

2.2 数据采集在个性化系统的应用

用户使用数据收集是迈向 Web 个性化服务的第一步,上述的所有方法都能用于各种不同的个性化功能中。

(1) 记忆型。不同的数据收集方法适合不同的记忆功能要求。实现向用户致意需要用户显示注册输入用户名字,将名字可存在本地数据库或 Cookie 文件里,其他的注册数据在实现个性化的存取策略上也非

常有用。另外，通过收集用户已经访问的网页可帮助实现书签功能，这可通过对日志文件分析和客户端 Agent 数据收集来支持这项功能。

（2）引导型。引导型功能需要了解用户当前的行为和知识水平，以便进行指导。获得这类数据需要将客户端和服务器端的数据汇集起来进行综合判断分析，可能还需要中间数据帮助。一般来自客户端的数据比较准确，特别适合进行指导性的个性化服务功能的实现。

（3）定制服务型。这类功能主要需要获取关于使用者的兴趣和偏爱的情况信息。这些数据要通过分析用户的浏览历史获得，客户端和服务器端的是这类数据的主要来源，特别是来自服务器端的日志数据。

（4）工作任务辅助支持型。实现对用户的工作辅助支持的功能往往需要融合多种数据收集方法。因此，来自服务器日志和客户端 agent 收集的数据能够反映用户的浏览行为，再结合这些数据去推断用户的意图。然而，要更准确获得用户的意图往往需要进一步分析用户的注册数据和用户当前一段时间输入的查询数据。

3 数据预处理中的个性化应用

收集到的数据要进行必要的处理，通过数据预处理可以使数据更为精细详尽，适合分析处理。此外，数据预处理在很大程度上是和领域相关的，数据预处理的好坏和数据本身的类型与质量也有很大关系，对数据的处理往往还要结合领域背景知识，在数据预处理的粗细程度上加以认真权衡，过粗和过细都会影响知识发现。这一阶段主要包括数据过滤，用户识别和事务识别。

3.1 数据预处理技术

（1）数据过滤。数据过滤主要是检查采集来的数据，将不恰当的或冗余的数据项从数据集中清去。这里主要涉及到对 Web 服务器端和代理服务器端收集的数据，由于这些数据记录了多用户与系统交互，需要提取和清理。从客户端采集来的数据因为没有太多用户干涉，数据相对要比较干净。此外，还要对用户给出的数据（如注册数据）进行确认、校正和形式化，以便于模式发现。最后，还要去掉日志文件里冗余的请求，比如图片，代理请求和网络蜘蛛的访问等等。

（2）用户识别。由于高速缓存、防火墙和代理服务器等的存在，准确识别出每个用户很困难，除非在客户端跟踪用户的行踪。但是，在客户端跟踪用户的访问行为涉及到用户隐私，必须用户配合才行。表 1 归纳了目前常用的用户识别方法及其优缺点。

（3）用户事务识别。用户事务对于分析用户的浏览行为很重要，用户事务是用户在对网站访问期间，在一定时间内访问的一组页面集合，是具有一定语义和目的的有序动作。目前多采用各种不同的启发式方法来识别用户使用事务。这些方法可分为基于时间的和基于上下文内容的两类（Spiliopoulou 1999）。例如，基于时间的启发式方法可以是对在网页上花费的时间限定上限或是对一个事务上面的总时间限定上限。而对特定的类型页存取或完成一定意义的工作可认为是基于上下文内容的。

然而，基于网页花费时间的方法由于用户行为变化很大，并不是非常可靠。此外，这个时间界限很大程度上取决于网站本身的内容。在时间方面，主要问题是高速缓存的使用带来的错误判断，这个问题可通过引入特定的 http 头解决（Shahabi 2001），但是一些外部因素，像网络畅通情况、用户的浏览器类型等会影响这一方法的使用效果。Yan 等人提出了在页面加入识别标记来帮助识别用户事务，然而，高速缓存还是会影响这个方式的准确性。

基于上下文方面，Cooley 等人在假定事务与一个用户行为有很强联系的基础上，将网站的页面分为导航页面，内容页面和混合页面帮助确定事务，但这种基于上下文的分类还取决于用户的浏览期望目标，一个用户的导航页可能是另一个使用者的内容页。而最大前向引用路径方法是 Chen 等人提出的，依照这个方式，每个最大前向引用路径为一个事务，这种方法与网站的内容无关，但高速缓存会影响日志记录。最后，Rdissono 和 Torasso（2000）提出了通过定义“当前关注中心”概念，将没有改变关注中心的行为构成浏览的“本地历史”，用来分析用户的行为。

3.2 数据预处理技术在个性化系统中的应用

除去噪音和不恰当的数据是实现网络个性化的第一步。此外，用户识别是网络个性化系统最关键因素之一，因此，大部份个性化功能都需要进行数据过滤和用户识别。

(1) 记忆型。用户识别是记忆功能需要的唯一数据预处理步骤,特别是使用者招呼和个性化存取都需要准确的用户识别,可以采用用户注册登记形式。前面讨论的启发式方法对于这类功能不太合适。

(2) 引导型。用户事务识别是实现良好的指导型功能的最基本要求,需要基于上下文的方法,这种方法能更详细地了解用户知识。而超链接推荐可以采用基于上下文或基于时间的用户事务识别方法。

(3) 定制服务型。用户事务识别也是实现这类个性化功能的基础。像内容定制需要有关用户访问网页之间的关联信息,基于上下文的方法更合适。

(4) 工作任务辅助支持型。用户事务识别对于识别用户运行某工作的意图很重要,特别是基于上下文的方法对于实现谈判助理功能比较合适,因为能用来发现使用者对某一个商议主题是否有兴趣。

4 模式发现中个性化技术应用

4.1 模式发现技术

这个阶段主要是发现用户的网络行为偏好和兴趣知识,帮助自动建立用户行为模型。

有各种机器学习方法用于 Web 使用挖掘的模式发现,如聚类、分类、关联规则发现和序列模式发现。表 2 为几个典型个性化系统采用的技术。

4.2 模式发现技术在个性化系统的应用

从 Web 数据中抽取使用模式是有效构造用户模型的基础。除了记忆型的最简单功能外,用户模型对个性化的所有功能都非常有用。

(1) 记忆型。向用户致意和书签功能由于要使用显示提供数据,一般不需要模式发现。然而,个性化存取权利要求按照存取策略将用户分类,因此,分类方法对个性化存取权功能的实现是必要的。

(2) 引导型。引导型功能主要需要关联规则和序列模式发现方法,以便标识相关的页面或各自的浏览模式,这样以后可以对用户进行新页面的推荐或对用户进行导航。

(3) 定制服务型。定制型需要按照用户的知识、兴趣和偏好对网站用户以及给网站页面分类。因此,分类技术可用于有预先分类定义的场合;而聚类技术用于需要从使用数据发现分类的场合。允许交叠的聚类算法有时更为适合,它可以将用户分在不同的类别,也更为灵活。

(4) 工作任务辅助支持型。考虑到了解用户需求主要是发现用户典型的浏览路径,以便决定何时执行辅助动作。序列模式发现可以为这类任务提供帮助,此外关联规则也可用于对网站用户的行为进行分析。

5 结束语

基于 Web 使用挖掘的个性化技术由于可以隐式发现用户的行为特征,成为当前个性化系统中的一项重要技术。然而,这些技术也是处于不断发展和完善之中,在具体运用这些技术的同时,必须了解这些技术的特点和局限,要结合应用系统的开发目标,选择合适的数据采集、数据预处理方法和模式发现技术,特别要善于将相应的技术融合,优势互补。本文介绍了这些相应的技术特点以及在个性化系统中的运用,随着对 Web 使用挖掘以及在个性化服务系统中的应用与研究的不断深入,我们相信面向用户的个性化服务技术将会逐渐不断完善和成熟。

参考文献

1 Mobasher B., Cooley R., Srivastava, J. Automatic personalization based on Web usage mining[J]. Communication of the ACM, 2000, 43(8): 142 – 151.

2 Pierrakos D., Paliouras G., Papatheodorou C. Web usage mining as a tool for personalization: A survey [J]. User Modeling and User – Adapted Interaction, 2003, 13: 311 – 372.

3 曾春、邢春晓、周立柱,个性化服务技术综述[J],软件学报, 2002, 13(10): 1952 – 1961。

4 Yew – Kwong Woon, Wee – Keong Ng, Xiang Li. Efficient Web log mining for product development [A]. In Proceedings 2003 International Conference on Cyberworlds[C], 2003: 294 – 301.

5 Srivastava J., Cooley R., Deshpande M. Web usage mining: discovery and applications of usage patterns from web data[A]. In Proceedings of the ACM SIGKDD Explorations[C], ACM Press, 2000, 1(2): 12 – 23.