高校数字图书馆个性化服务关键技术探究

孙士新,李海燕,王甫成
(亳州职业技术学院,安徽 亳州 236800)

【摘 要】各项信息技术的发展为高校数字图书馆个性化服务提供了可能,通过对高校图书馆个性化服务现状的调研,提出了高校数字图书馆个性化服务可行的策略和关键技术。

【关键词】信息技术;策略;关键技术
【中图分类号】G250.76;G252 【文献标识码】A 【文章编号】1673-1891(2010)02-0048-06

随着信息技术的快速发展,高校图书馆拥有了海量的数字图书资源,如何让用户准确、快速、便捷的找到其需要的图书是图书馆工作人员研究的问题。现有图书馆大多数是使用了传统的图书馆管理系统,只能满足读者日常借阅图书的需要,高校图书馆的读者希望花费尽量少的时间,找到最适合自己的书籍。从图书馆个性化服务现状出发,通过调查研究,掌握我国图书馆个性化服务发展的现状,提炼了图书馆个性化服务的关键技术。

一 高校数字图书馆个性化服务发展现状

如何根据现有数据信息发现用户个性化兴趣,结合现有资源把用户感兴趣的信息便捷的推送给用户,是目前图书馆工作人员努力研究的问题。

(一) 国内外图书馆个性化服务的发展

早在1998年Mylirary@NCState在美国北卡罗莱纳州立大学图书馆建立,主要功能包括:个性化链接(Mylinks)、个性化更新(Myupdates)、个性化内容(Mycontents)、个性化目录(Mycatalogs)、个性化文献传(Mydocumentdelivery)五个组成部分[1]。在1999年美国图书馆与信息技术联合会(LITA)的10个著名图书馆专家在一次研讨会上就把个性化服务列为图书馆技术发展的七大趋势之首。

自1999年清华大学的路海明和徐晋晖合著的《基于Agent技术的个性化主动信息服务》和《基于Multiagent生态进化算法实现个性化主动信息服务》两篇论文发表以来,我国便开始了个性化信息服务的研究。2002年,清华大学和清华同方共同主办了CNKI工程,他的主要功能之一就是主动推送服务。

截止到2009年7月,国内"211工程"的100所大学图书馆网站,采用了基于RSS技术定制信息的有16所,使用了MyLibrary系统的有28所,厦门大学同时使用了我的图书馆和基于RSS的定制技术。

(二) 高校图书馆个性化服务系统研究的内容

高校数字图书馆个性化服务系统的研究内容和研究方向主要包括:

1.个人书架。个人书架是个性化服务系统为浏览者建立的私人数据库[2],可以保存浏览者的私人信息,保存其感兴趣的资源、访问记录、信息定制、信息资源推送及搜索关键字等信息。

2.个性化检索研究。依据读者资料,利用读者兴趣资源库,在显示搜索结果时过滤无关信息。以实现不同用户在输入同一搜索关键字时,显示不同结果,使得搜索结果更贴近用户个性化兴趣。

3.信息分类定制与推送服务研究。信息分类定制是指读者可以按照自己的目的和需求,设定所需信息资源的类型、表现形式、系统服务功能等[1]。信息推送服务是运用退送技术来实现的一种个性化主动信息服务方式[3]。

4.信息智能代理技术。是一种可支持高级复杂自动处理的代理软件技术,具有高度智能性和自主学习性,可以根据用户定义的准则,主动的通过智能化代理服务器为用户搜索感兴趣的信息,并把加工过的信息按时推送给用户,并能推测出用户的意图,从而自主的制定调整和执行工作计划[4]。

5.虚拟咨询服务。指在网络环境下,图书馆以计算机网络为通信手段,以现有数字图书资源为基础,通过电子邮件、QQ、留言板、飞信等形式,向用户提供超时空的咨询服务。

6.Web数据挖掘。所谓数据挖掘(Data Mining)就是从大量的、不完全的、有噪声的、模糊的、随机的原始数据中提取隐含在其中的、事先未知的、但又是潜在有用的信息和知识的过程。数据挖掘模型建立完成后,进行验证和评价非常必要[5]。Web数据挖掘一般可分为Web内容挖掘、Web结构挖掘和Web使用挖掘三类[6]。

二 高校数字图书存在的问题及改进方法

高校图书馆是师生获取知识的重要源泉,是高校办学实力的象征。

收稿日期:2010-05-11
作者简介:孙士新(1980-),男,安徽亳州人,助教,硕士,主要研究方向:职业教育、网络技术。

(一) 存在的问题

随着信息技术的发展和数字图书馆建设步伐的加大,高校数字图书资源日愈丰富,可供读者查阅的电子图书也越来越多。然而,在实际使用中读者若想精确、快速定位到自己想查找图书的位置非常困难。产生这种现象的原因主要有两点:一是图书资源的急剧增多,图书信息数量级越来越高,查找一条信息所需时间越来越长;二是用户与网站系统缺乏沟通,面对海量的图书信息,用户很难精确表达自己的需求,即便输入同一关键字,不同用户对信息的需求也不尽相同,网站系统很难充分理解用户需求。用户希望进入网站后看到的书籍全是自己感兴趣的,不同用户通过同一搜索关键字所搜索的结果应该是不同的,每个用户进入网站后能够享受到更贴近自己的服务,从而可以使用户方便、快捷的找到自己希望的资源,这就需要高校数字图书馆服务的个性化。

随着图书馆服务理念从"以图书为中心"向"以用户为中心"的转变,图书管理人员更希望把图书馆内的书籍展现给最需要的读者,希望随时可以掌握书籍的使用情况,把访问量低的书籍显示出来,以实现图书资源的有效整合,即图书资源个性化。

个性化服务指的是以用户为中心,基于用户的信息使用行为、习惯、偏好、特点及用户特定的需要,向用户提供满足其个性化需求信息内容和系统功能的一种服务[7]。当前,高校数字图书馆个性化服务在一定程度上有所发展,但还存在许多不足之处,主要表现在个性化服务模式单一、在推送上个性化服务缺乏动态性和实效性、推送模式单调等问题。

(二) 改进途径

RSS技术是Web2.0中的一种Syndication技术,该技术致力于建立标准的频道描述框架和内容收集机制,主要采用XML的标准[8]。结构化是采用RSS标准格式信息的主要特征,处理和分析该标准的信息非常便利。基于RSS推送技术的动态性、即时性、便捷性的特点可以恰好的填补现有高校数字图书馆个性化服务系统推送方法上的缺陷。

基于现有高校数字图书馆个性化服务系统的不足和RSS技术的优越性,将高校图书馆个性化服务和RSS技术有机结合,在一定程度上实现了用户需求与网站信息的对接,并及时的把最新图书资源主动向用户推送,真正的实现了高校数字图书馆个性化服务。

三　高校数字图书馆个性化服务的关键技术探究

高校数字图书馆个性化服务系统的关键技术主要包括用户个性化兴趣、图书使用模型、信息推送三个部分。

(一) 个性化服务技术基础

支持个性化服务系统实现的主要技术包括数据挖掘技术、RSS技术等。

1. Web数据挖掘

(1) Web挖掘的概念和内容

所谓数据挖掘(Data Mining)就是从大量的、不完全的、有噪声的、模糊的、随机的原始数据中提取隐含在其中的、事先未知的、但又是潜在有用的信息和知识的过程[9]。依据挖掘对象不同,Web挖掘分为Web内容挖掘(Web Content Mining)、Web结构挖掘(Web Structure Mining)和Web使用记录挖掘(Web Usage Mining)[10]。如图1所示。

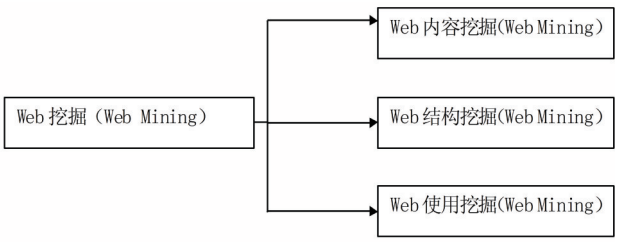

图1　Web挖掘

2.关联规则算法

关联规则(Association Rule)是表示数据库中一组对象之间的某种关联关系的规则[11]。

3. 聚类挖掘

(1)聚类的基本概念

聚类是指将物理的或抽象的对象集合分组成为由相似对象组成的多个类的过程。聚类生成的簇是一组数据对象的集合,同一个簇中的对象是相似的,不同簇中的对象是相异的。

(2)主要的聚类方法

聚类算法大体上可以分为:分层、划分、基于密度、基于网格的方法等。

(3)协同过滤推荐

协同过滤推荐是一种基于聚类技术的推荐算法。常用的协同过滤推荐算法有基于用户的协同过滤推荐、基于模型的协同过滤推荐、基于项目的协同过滤推荐、基于项目评分预测的协同过滤推荐、基于项目分类的协同过滤推荐算法等[12]。

(二) RSS技术

1.XML简介

XML(eXtensible Markup Language)是一种可以对信息进行自我描述的语言。在关系型数据库系

统中的应用研究[13]。

XML可看作一种半结构化的数据模型，可以很容易地将XML的文档描述与关系数据库中的属性一一对应起来，实施精确地查询与模型抽取，以检索出适当的数据[14]。通常一个正规格式的XML文档分三部分：一个可选的序言；文档的主体，通常为层次树状结构形式，由一个或多个元素组成；可选的尾部。所有XML文件都包含实体结构和逻辑结构。实体结构包含文件中所使用的实际数据。逻辑结构则像一个样本，说明该文件中包含哪些元素及元素的顺序。

2.RSS的概念及标准

（1）RSS的概念

RSS是1999年由Netscape发布的，用于发送新闻标题，被称为“推”技术，后来Userland Software发展了其简化版本。与此同时，另外一组开发人员在复兴最初的RDF版本，并最终发布了一个RSS1.0的版本（Rich Site Summary）[15]。RSS是一种用于共享新闻、简讯等Web内容的数据交换规范，是一种基于XML标准的Syndication技术，是广泛应用于互联网上内容封装和投递的协议。

（2）RSS的标准

RSS可以将网站当作一系列频道（channels）的集合，每个频道又包括一系列资源（Items），所以通过对频道和资源的描述可以实现对作为资源集合的网站描述[16]。这个采用RSS元素描述的网站内容汇总文件即为一个RSS feed。

内容的提供者可以为接收者建立一个RSS信息源（RSS Feed），这只需要在其网站中增加些简单的代码。一个RSS文档通常由<rss>、<channel>、<item>三部分构成。

3.RSS信息服务系统解析结构

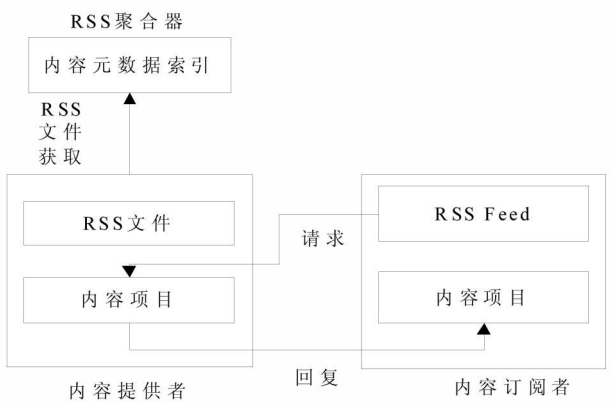

图2 RSS解析体系结构

RSS信息服务方式是内容提供者首先按照内容订阅者的RSS请求把RSS Feed封装，然后把封装后的RSS Feed输出；内容订阅者向RSS聚合器读取RSS摘要且同时获取资讯信息。RSS信息服务系统解析结构包含三个主要部件，如图2所示。

RSS信息服务使互联网信息的发布和共享变得更方便、更高效，用户可以用更少的时间获取更多的信息。但是这种服务模式缺乏用户的参与，个性化特征不明显。

（三）用户个性化兴趣模型

1.用户个性化兴趣模型构架

高校数字图书馆个性化服务系统中用户个性化兴趣模型构架主要包括用户访问信息收集、用户资料收集、用户个性化兴趣提取、用户个性化兴趣更新、用户个性化兴趣表示等。

高校数字图书馆个性化服务系统核心环节是无缝对接用户本身兴趣与系统理解的用户个性化兴趣，以实现系统即时的向用户提供个性化服务。在用户个性化兴趣模型建立的过程中需要考虑以下几个关键问题：

（1）使用哪种途径快速精确地获取用户个性化兴趣。

（2）使用哪种用户个性化兴趣表示方法，更能够准确详实地描述用户兴趣。

（3）使用哪种途径更新用户个性化兴趣，使其当用户个性化兴趣发生变更时能精确的重新获得用户当前兴趣。

2.用户个性化兴趣的获取

用户个性化兴趣的提取即根据用户的注册信息（年龄、性别、父母职业、专业、学历、爱好等）、浏览行为信息（检索的关键字、访问内容、停留时间、访问次数等），把用户个性化兴趣组合起来并加以表示的过程。用户个性化兴趣的提取途径有两种：

（1）明获取

明获取方法需要用户的配合，要求主动的向系统发送自己的兴趣，重要方式为：用户提交注册信息、用户对系统推送的个性化信息进行满意度打分、提交自己最近兴趣（用户通过一定方式提交自己最近感兴趣的主题、关键字、文章标题等，用户也可以把自己的兴趣通过文章概述的形式描述出来发送给系统）等。

（2）暗获取

暗获取不需要用户有意识的去发送自己的兴趣描述。系统通过对用户的访问页面的行为和方式等信息进行挖掘以得到用户个性化兴趣。该方式下用户不需要有意识的参与。当前暗获取主要信息来源是Web日志，采用Web日志可以知道用户

访问某页面的次数、在该页面上的停留时间等。通过对Web日志分析系统可以获得页面相关性、用户群兴趣的相似度、访问模式及某一用户所属的兴趣群等信息,图书馆个性化服务系统可以通过这些信息创建、更新用户描述数据库。

3.用户个性化兴趣的表示

表示用户个性化兴趣的方法多样,目前没有形成统一的一个标准,我们经常用以下几种方法表示用户个性化兴趣:

(1)主题表示法

用户个性化模型的主题表示法是指采用用户感兴趣资源的主题来表示用户兴趣模型的方法。如用户对哲学和工学类感兴趣,则用户模型表示为{哲学,工学}。该表示法一般与相应的领域相结合。

(2)关键词表示方法

用户个性化兴趣模型的关键词列表表示法是采用用户感兴趣的资源的关键词表示用户个性化兴趣的方法。例如:用户对计算机感兴趣,则用户个性化模型可以表示为{软件,网络,数据库,硬件,信息管理系统}等。关键词可以是用户有意识的提交给系统,也可以是用户无意识的提交给系统。WebWatcher是典型的利用关键词列表方式表示用户个性化兴趣模型的系统。

(3)加权关键词向量表示法

向量空间模型是20世纪70年代中期由杰拉尔德·索顿提出的检索系统的向量模型,是到目前为止应用最多且效果较好的用户个性化模型表示方法。向量空间模型中的特征向量是由文本中提取的特征项组成的,且以某种形式为其特征项赋权。如文档T可表示成$T(r_1,r_2,\cdots,r_n)$,其中r_x是特征项,$1\leqslant x\leqslant n$。因特征重要程度有异,可以采用附加权值ωx进行量化,文档T则可表示为:$((r_1,w_1),(r_2,w_2),\cdots,(r_n,w_n))$。向量的各维都由一个关键词、权值组成。权值的数值类型为布尔型和实型,依次可以表示用户对某个关键词感兴趣与否,以及感兴趣的强烈程度。

4.用户个性化兴趣的更新

用户个性化兴趣更新是指当系统通过一定的方式获取用户临时兴趣后,怎么把用户临时兴趣与系统中用户原有兴趣合并而得到用户当前兴趣的方法。

时间的推移,用户的兴趣也在不断的变更,当用户个性化兴趣变化后,用户的浏览行为也在发生变化,系统提取的用户个性化兴趣也会不同。这时系统提取的用户个性化兴趣是改变后的用户个性化兴趣,系统把改变后的用户个性化兴趣添加到用户个性化兴趣,结合以前用户个性化兴趣即是用户当前兴趣,用户个性化兴趣的改变在系统中主要体现在用户个性化兴趣的更新。

(四)图书使用模型

1.图书使用模型构架

数字图书馆个性化服务系统中图书使用模型主要包括图书访问记录收集、图书访问记录提取、呆滞图书更新、呆滞图书表示等。

图书模型成功的关键准确提取呆滞图书资源,并形成一个随时间变化而周期性变化的呆滞图书曲线,以合理的安排图书资源,做到图书馆资源整合。呆滞图书是指图书馆中访问量极低的图书。在建立图书模型时需要解决以下问题:

(1)采用何种方法准确快速的提取图书访问记录。

(2)采用何种方法精确方便的更新呆滞图书。

(3)采用何种方法能够把呆滞图书分门别类的表示出来。

2.图书访问记录提取

图书访问记录提取是根据对访问记录的提取,来发现各种书籍的访问情况,进而确定在某段时间内哪些是闲置书籍。访问量提取的方法有三种:

(1)基于日志文件的方式

通过预处理服务器上记录的访问日志文件,提取有效信息,进而获取图书资源的访问次数。服务器上的日志文件主要有ISS、APACHE日志等。

(2)基于监听方式

监听方式主要借助用户请求被服务器过滤特点,用户对任何一个页面的请求都能被服务器检测到,而且服务器可以修改一个request,这样服务器可以借助程序记录用户对所有图书页面的访问次数。

(3)代码嵌入方式

代码嵌入方式就是在图书页面中嵌入统计页面代码,记录、统计读者浏览页面时间和行为,当读者访问那些页面时,图书的页面ID、被访问时间、访问次数等就会被提交到接收页面,接收页面利用内部对象将被访问的页面记录到数据库中。

3.呆滞图书更新

呆滞图书更新是指由于用户每天访问图书资源信息不同,数据库中未被访问的信息也发生改变,采用一定的统计策略,真实统计图书空闲情况的过程。

用户的兴趣发生着改变,他们每天访问书籍类

型也在改变，访问的页面也在发生变化，根据日、周、月为单位对每个图书信息页面的访问情况进行统计，结合现有图书数据库，得到的呆滞图书数据库也在发生改变。如何及时更新呆滞图书？如何准确真实的反映图书的空闲情况？

4.呆滞图书表示

与用户个性化模型一样，我们也要为呆滞图书建立图书模型。个性化服务系统的应用领域决定所处理资源的对象。如：Smart Push、Anatagonomy是报纸；GropLens应用领域是Usenet新闻；Cite Seer处理对象是科技文档等。目前，图书馆个性化服务系统所处理的资源都属于文本范畴。

图书资源的描述与用户的描述密切相关，一般采用相同的机制来表达用户与图书资源，图书资源的描述可以是基于内容的方法和基于分类的方法表示。基于内容的方法。

（五）个性化推荐技术

数字图书馆个性化服务系统建设在很多院校已开始实现，常用的推荐技术有：基于聚类的推荐算法、基于项目评分预测的协同过滤算法、基于关联规则的推荐算法、基于最近邻技术的协同过滤算法等。

1.基于聚类的推荐算法

聚类是数据挖掘技术的一个分支，是用于从数据集中找出相似的数据并组成不同的组[17]，传统上，聚类算法大体上可以分为：划分法、分层法、基于网格的方法、基于密度的方法、基于模型的方法等[18]。

2.基于项目评分预测的协同过滤算法

基于项目评分预测的协同过滤算法原理是依据目标用户评价过的项目与目标项目间的相似性，。该算法经历分项目相似性和推荐产生两个阶段。

3.基于关联规则的推荐算法

所谓关联规则，就是寻找描述数据库中数据项之间存在的关联，利用关联规则的数据挖掘技术，可以找出大量数据之间未知的依赖关系。

4.基于最近邻技术的协同过滤算法

最近邻技术就是使用统计方法查找与目标用户具有相同或相似兴趣的邻居用户。

（六）信息推送

目前高校数字图书馆个性化信息服务系统采用的推送服务方式有邮件式推送、频道式推送、手机短信推送、用户专用网页四种。

1.邮件式推送

邮件式推送，即主动的把需要推送的信息发电子邮件给相关用户，发送的内容一般为对用户的通知、新书列表、学术会议等。该方式需要借助于一个基于Web的电子邮件发送系统。该推送方式打破了时空限制，用户无论在任何地方、任何时间只要打开自己的邮箱就可以看到被推送的信息，减少了网上搜索的过程，将用户感兴趣的信息推送给用户，简化了用户获取信息的过程。

2.频道式推送方式

频道式推送方式是当前广泛采用的一种推送模式，它将相关站点定义为浏览器中的频道，用户可以根据自己兴趣，像选择电视频道一样选择自己感兴趣的信息。RSS是基于频道推送的服务方式，被称为真正的简单聚合，基于RSS的频道式推送在Blog和新闻聚合上广泛应用。大型的新闻站点一般都使用RSS制定用户感兴趣的新闻信息。现在RSS还被广泛应用在网络信息资源推荐、图书书目推荐、机构最新信息发布等方面。

3.手机短信推送方式

手机短信推送方式就是借助手机短信平台推送信息的一种方式。随着读者对手机拥有率的提高，发送手机短信成为人们日常传递信息、获取信息的主要方式。如图3。

图3 手机短信推送方式图例

4.用户专用网页

用户专用网页又称为动态网页，系统根据不同的登录用户显示不同的内容，将用户感兴趣的信息分栏目展示给用户。一般采用点对点通信方式，该方式要求用户必须访问网站。目前很多高校数字图书馆仍然采用该推送模式。

注释及参考文献：

[1]周军.基于数据挖掘的数字图书馆个性化服务系统的构建[J].图书馆学研究，2007(3)：15-17.

[2]董倩.利用数据挖掘技术创建数字图书馆个性化服务系统[J].科技情报开发与经济，2007(12)：13-14.

[3]高文，刘峰，黄铁军.数字图书馆原理与技术实现[M].北京：清华大学出版社，2000，10-158.

[4]鲍静,范生万.关联规则挖掘在图书馆中的应用[J].数字图书馆论坛,2008(5):49-52.
[5]M.J.A.Berry and G.Linoff.Data Mining Techniques:For Marketing, Sales,and Customer Relationship Management [M].Wiley Computer Publishing,2nd edition,2004:10-20.
[6]数据挖掘——概念、模型、方法和算法[M].闪四清,陈茵,程雁等译.北京:清华大学出版社,2003:154-156.
[7][美]Mehmed Kantardzic.欧阳烽.Web数据挖掘与高校数字图书馆个性化服务[J].数字图书馆技术论坛,2008(1):103-107.
[8]周志峰.基于RSS的高校图书馆重点学科信息导航系统研究[J].现代情报,2008(11):67—72.
[9]王伊蕾,李涛,王福生,等.一种基于库存理论的图书订购策略[J].情报科学,2008(5):668-700.
[10]黄浩,王建军.WEB使用挖掘研究[J].计算机系统应用,2008(1):124-128.
[11]耿晓中,张军.Web挖掘及其在电子商务中的应用[J].长春工程学院学报(自然科学版),2007(4):79-82.
[12]Xu Rui.Survey of cluster algorithms[J].IEEE Transaction on Neural Network,2005,13:645-678.
[13][美]Vipin Kumar.数据挖掘导论[M].范明,范宏建等译.北京:人民邮电出版社,2006.305-400
[14]Rudi Cilibrasi.Clustering by compression[J].IEEE Transactions on Information Theory,2005(4):1523-1545.
[15]毛海波.RSS技术在数字图书馆中的应用[J].新世纪图书馆,2007(2):77-78.
[16]黄继征.RSS技术在图书馆信息推送服务中的应用[J].大学图书情报学刊,2006(5):35-42.
[17]张丽宁.基于RSS的数字图书馆信息服务应用[J].农业图书情报学刊,2007(6):61-65.
[18]陈建芳.基于RSS的教学资源整合与创新服务[J].情报探索,2008(9):42-43.

The Key Technical Research about the Individual Service in College Digital Library

SUN Shi-xin, LI Hai-yan, WANG Fu-cheng
(Bozhou Vocational and Technical College, Bozhou, Anhui 236800)

Abstract: The improvement of each information technology has provided the possibility of individual service for college digital library. Through the investigation and survey on the present situation of the individual service in the college library, it puts forward the feasible tactics and key technology about the individual services in college digital library.

Key words: Information technology; Tactics; Key technologies