Web挖掘技术在高校数字图书馆个性化服务中的应用*

张沛露[1] 王建军[2]

(1:吉林建筑工程学院计算机科学与工程学院,长春 130021;
2:中国科学院长春光学精密机械与物理研究所,长春 130033)

摘要:随着互联网信息的迅猛增加,高校数字图书馆信息资源极大丰富.尽管人们可以借助互联网上功能强大的搜索引擎和快捷的传送手段,使信息资源的搜集和获取变得十分容易,然而在使用中,用户快速、准确地查找自己所需的信息却越来越困难.笔者介绍了Web挖掘技术,并分析了如何将其应用到高校数字图书馆建设中,为用户提供个性化信息的主动推送服务.

关键词:Web挖掘;高校数字图书馆;个性化服务

中图分类号: TP391 **文献标志码**:A **文章编号**:1009-0185(2010)03-0067-03

Web Mining in the Application of Personalized Service for University Digital Library

ZHANG Pei-lu[1],WANG Jian-jun[2]

(1:*School of Computer Science and Engineering, Jilin Institute of Architecture and Civil Engineering, Changchun, China* 130021;
2:*Changchun Institute of Optics Fine Mechanics and Physics of Chinese Academy of Sciences, Changchun, China* 130033)

Abstract: With the rapid increase in Internet information, the information of university digital library was greatly enriched. Although people can make use of the Internet which has powerful search engine and fast means of transmission, the collection of information resources and access to become very easy, but in use, the user quickly and accurately find the information they need has become more and more difficult. This article describes the Web mining techniques and analysis of how to apply it to university digital library construction, to provide users with personalized push services.

Keywords: web Mining;university digital library;personalized service

随着互联网信息的迅猛增加,难以从浩瀚的网络中真正找到自己所需要的信息,因此,人们在寻求一种将用户感兴趣的信息主动推荐给用户的服务方式,即个性化的信息服务.对于数字图书馆,每一个用户所追求的信息目标不同,关注的信息子空间也就不同.现在,大多数图书馆的网站服务模式并没有考虑用户之间的差异性,使每一个用户必须面对相同的用户空间.用户迫切需要一种能够依据自身特点自动组织调整信息的服务模式,因此,将个性化服务应用到数字图书馆已成为一种必然的趋势.数字图书馆的个性化服务,允许用户对自己感兴趣的内容信息进行定制,服务器可以自动对用户的各种浏览信息进行收集,全方位依据用户的需求和喜好提供服务计划,最终形成一个相对完整的信息资料集,反馈给用户.

收稿日期:2010-01-10.
作者简介:张沛露(1977~),女,吉林省长春市人,讲师,硕士.
*基金项目:吉林建筑工程学院青年科技发展基金项目(J20091056).

1 高校数字图书馆个性化服务现状

个性化定制服务技术[1]在国外图书馆推出的时间还很短，但已经普遍受到用户的关注.目前开发出的有MyLibrary[2]，MyGataway等个性化定制服务系统，其中，MyLibrary是美国康纳尔大学图书馆的一个创举.MyLibrary由两部分组成：MyLinks和MyUpdates.MyLinks是为用户个人搜集和组织数字化资源的一种工具；MyUpdates是将图书馆新到资源及时通知用户的一种工具.著名的美国哈佛大学图书馆网站(http：//lib.harvard.edu/)为用户提供了一个搜索工具条，以此提供不同的个性化服务.主要包括E-research和HarvardLibX两种服务方式，其中E-research根据用户对查询内容(期刊、百科全书、词典、书籍等)的需求，为用户提供不同的搜索方式；Harvard LibX提供浏览器工具栏，主要收集每个用户的需求，提供用户最有可能访问的页面链接，同时为用户提供快速查询图书馆网站内部资源的服务.

当前，国内著名高校图书馆网站都在努力尝试为用户提供个性化信息的主动推送服务.北京大学图书馆网站对不同的学科资源进行分类，以此来满足不同专业学生的需要，最近浙江大学图书馆网站开通的个性化服务，是指当用户无法很好地利用关键字进行相关搜索时，可填写需求表单发送给专业老师，由专业老师提供在线的帮助.

2 Web挖掘技术

Web挖掘，就是将数据挖掘技术应用在Web上，从大量类型丰富的Web数据中挖掘隐含知识的过程[3-4].Web上的数据类型丰富，主要包括：HTML文档中的文本数据、多媒体数据、超链数据，以及Web服务器日志文件中登录用户的访问行为数据等.在数据挖掘领域，如果面对的数据类型不同就会采用不同的挖掘算法.因此，根据所挖掘的Web数据的类型，可以将Web挖掘分为以下3类：Web内容挖掘、Web结构挖掘和Web访问日志的挖掘.

2.1 Web内容挖掘

Web内容挖掘，就是从Web文档内容中抽取知识的过程，主要分为文本信息的挖掘和多媒体信息的挖掘.由于文本仍是信息传递的主要方式，而且文本处理技术相对比较成熟，因此文本数据的挖掘，在研究和应用上都比较普遍.文本挖掘主要分为：文本的总结、分类、聚类、关联分析，以及利用Web文档进行趋势推荐等，其中最常见的是文本的分类和聚类.

2.2 Web结构挖掘

Web结构挖掘，是指从Web文档的链接中推导知识的过程.其中比较有代表性的工作是PageRank和CLEVER.

PageRank的核心思想在于发现权威性页面.权威性页面可由Web页面间的超链来反映：当一个Web页面的作者建立指向另一个页面的超链时，可以看成该作者对另一Web页面的引用；如果一个页面被引用的次数越多，而且引用该页面的重要性越高，该页面也就越权威.PageRank的具体描述是这样的：假设有n个页面T_1，T_2，…，T_n指向(引用)页面A.参数d可以在0到1之间取值，通常取为0.85.C(A)为A中指向外边的链接数量.那么A的PageRank定义如下：

$$PR(A)=(1-d)+d(PR(T_1)/C(T_1)+\cdots+PR(T_n)/C(T_n))$$

PageRank在网页上形成一个概率颁布，所有网页的PageRank的和是1.

CLEVER系统主要采用了HITS算法，该算法的主要思想在于，如何识别hub/authority页面.著名的搜索引擎Google中就采用了该算法，比较于其它基于词类索引检索的搜索引擎，可以得到明显优化的查询结果.

2.3 Web访问日志的挖掘

Web访问日志，是登录某个Web站点的用户经过一系列的站点浏览后，系统自动记录的用户浏览行为

数据，诸如用户的 IP、用户的访问时间、浏览过页面的 URL、请求方法、请求的字节数、客户端的操作系统和浏览器版本号等．通过对 Web 站点上用户访问日志文件中的数据的挖掘，可以了解登录 Web 站点的大多数用户经常采用的浏览模式、浏览路径，从而改进站点的设计．对于访问模式相似的用户，进行分类或聚类；针对类型不同的用户，提供不同的个性化服务方案．Web 日志挖掘系统框架见图 1．

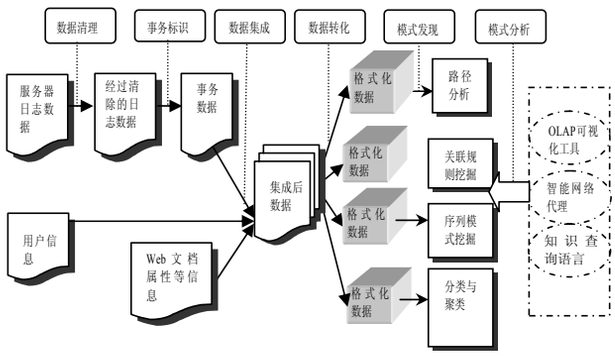

图 1 Web 日志挖掘系统

3　Web 挖掘技术在高校数字图书馆个性化服务的应用

高校数字图书馆为师生主要提供以下功能：查找图书、期刊论文、会议文献等数字资源；图书借阅、归还等服务；发布图书信息、管理制度；导航到图书光盘、视频资源等数据库系统．师生时常登录到网站中查找其需要的信息，根据师生所学专业、研究方向不同，关注目标也不同．通常这类师生会到常用的图书馆网站上，查找自己所需要的特定领域的资源；浏览一下有哪些内容发生变化，是否有新知识增加，而且所有改变常常是用户所关注的内容；另外，当目标网页所在的位置有所改变或这个网站的组织结构、层次关系有所变动时，所有这些问题只要稍加改动，容易使用户难以找到所需内容．

本课题采用 Web 挖掘技术与搜索技术相结合．首先允许用户对感兴趣的内容进行定制，构造数据挖掘的先验知识，然后通过构造浏览器插件，捕获用户在浏览器上的行为数据，采用 Web 数据挖掘的方法，深入分析用户的浏览行为数据，获得用户的信息资料集，最终为用户提供不同的个性化服务页面，并提供用户对站内信息进行搜索功能，同时可以满足师生对于图书馆资源进行查找访问的需求，实现高校图书馆网站资源真正意义上的个性化服务．具体流程如下：① 数据采集．包括用户定制信息的收集和用户浏览行为信息的收集；② 数据预处理．把网页中的文本、图片及其他文件转换为数据挖掘技术可用的形式；③ 用户模式识别．选择数据挖掘方法从 Web 中挖掘知识；④ 模式分析，分析挖掘结果，得到用户模式；⑤ 动态生成 HTML 界面，应用 PUSH 技术，将挖掘得到的用户模式主动推送给数字图书馆的用户；⑥ 利用数据挖掘的结果进行自学习，完善挖掘知识．

4　结语

笔者首先分析了当前高校数字图书馆实现个性化信息服务的必要性．在对 Web 挖掘技术研究的基础上，总结了 Web 挖掘技术在高校数字图书馆个性化服务中的应用方法．笔者所讨论的个性化服务能够实现对用户需要的信息及时、准确的主动推送，同时可以满足用户寻找自己关心内容的需求．

参　考　文　献

[1] 常勇生．国内外数字图书馆个性化信息服务现状与建设趋势[J]．科技情报开发与经济，2007，17(28)：92-93．

[2] 周　慧．数字图书馆 My Library 个性化服务探析[J]．情报探索，2006(5)：75-76．

[3] JiaweiHan，Michclinc Kamber．数据挖掘概念与技术[M]．北京：机械工业出版社，2001：290-295．

[4] 丁　一，孙玉霞．基于 Web 挖掘的用户个性化服务研究[J]．湖北师范学院学报，2005，25(3)：23-27．