Web挖掘在个性化远程教育中的应用

石佑红[1],赵 宏[1],乔 敏[2]
(1.北京交通大学 计算机学院,北京 100044;
2.广东工业大学 计算机学院,广东 广州 510090)

摘 要:基于Web挖掘,提出了一种新的个性化远程教育模型。它能充分利用用户Web访问记录,同时结合用户与站点的交互数据进行挖掘,以此来发现学习者的浏览(学习)兴趣,从而改进页面的设计,优化站点结构,更好地满足学习者的个性化需求,提升个性化远程教育的质量。

关键词:Web挖掘;Web使用记录挖掘;远程教育;个性化
中图分类号:TP181;G434 **文献标识码**:A **文章编号**:1673-629X(2006)09-0136-03

Web Mining Used in Personalized Distance-Learning

SHI You-hong[1], ZHAO Hong[1], QIAO Min[2]
(1. School of Computer Science, Beijing Jiaotong University, Beijing 100044, China;
2. School of Computer Science, Guangdong University of Technology, Guangzhou 510090, China)

Abstract: It presents a distance-learning model based on Web mining in this paper, which can take advantage of those students' pages access log information. Moreover, it can mine the useful information between the Web sites and students. This result is not only to find those students' learning interest and improve design of page, but also to optimize the Web sites' structure and meet every student needs. Finally, it also can improve the distance-learning quality.

Key words: Web mining; Web usage mining; distance-learning; personalization

0 引 言

现代远程教育以计算机网络以及卫星数字通讯技术为支撑,具有时空自由、资源共享、系统开放、便于协作等优点。由于接受教育的对象存在着很多个性差异,主要体现在:个人学习目标的不同、学习能力的不同、已有的知识基础的不同、学习习惯的不同等等,这就决定了远程教学必须是一种适应个别化学习需求的个性化教学[1]。目前允许教学资源在课程知识和教学管理水平进行交换的标准还没有很好地认定,妨碍了教学资源的大范围共享与交流。况且,现有的基于Web的远程教育平台并不能解决个别化学习的需求,所以也就无法对学习者实施个性化的远程学习服务。将Web挖掘运用于远程教育系统中,使之量身定做地为每个个体提供个性化的学习方案,是远程教育获得进一步发展的一个重要手段。

一般所说的Web个性化实质上就是一种以用户需求为中心的Web服务。当用户访问一个Web站点时,不同的用户由于其个性化的差异,兴趣喜好通常是不同的,对站点的访问也会带有某种偏好,同时用户的访问路径中蕴藏了用户对站点的兴趣及用户的兴趣转移。因此,可以分析Web访问日志数据,同时充分利用用户与站点的交互数据来发现用户的使用模式,从而向每个用户提供个性化Web站点[2]。当前的个性化远程教育研究大多数只是强调用户的访问日志,或是简单地把交互数据和访问日志合并处理,并提出了一些基于日志挖掘的个性化远程教育模型。鉴于远程教育学习的特殊性,其用户交互数据对个性化教学的实施的好坏有着很大的影响。设计一个能对其使用记录进行挖掘,并充分利用用户的交互数据,通过用户与系统之间循环往复的交互,最终能够为用户提供个性化服务的系统,将能够针对不同的学生,提供不同的学习内容和学习模式,真正做到因材施教。

收稿日期:2006-01-09
作者简介:石佑红(1980-),男,湖北黄冈人,硕士研究生,研究方向为数据挖掘;赵 宏,教授,研究方向为数据库与数据挖掘。

1 Web挖掘综述

1.1 Web挖掘的定义和分类

Web挖掘是一项综合了多领域的技术,涉及到Web、数据挖掘、计算机语言学、信息学等。Web挖掘就是从与WWW相关的资源和行为中抽取感兴趣的、有用的模式和信息[2]。可以对Web挖掘作如下定义[3]:

定义:Web挖掘是指从大量Web文档的集合 C 中发现隐含的模式 P。如果将 C 看作输入,将 P 看作输出的话,那么Web挖掘的过程就是从输入到输出的一个映射 $\xi: C \rightarrow P$。

一般地,Web挖掘可以分成为3类[4]:Web内容挖掘(Web content mining)、Web结构挖掘(Web structure mining)、Web使用记录的挖掘(Web usage mining)。

1.2 Web内容挖掘

Web内容挖掘是从文档内容或其描述中抽取知识的过程。一般可以从两个不同的观点来进行研究:从资源查找(IR)的观点来看,Web内容挖掘的任务是从用户的角度出发,怎样提高信息质量和帮助用户过滤信息;而从DB的角度讲,Web内容挖掘的任务主要是试图对Web上的数据进行集成、建模,以支持对Web数据的复杂查询。

1.3 Web结构挖掘

Web结构挖掘是从WWW的组织结构和链接关系中推导知识。由于文档之间的互连,WWW能够提供除文档内容之外的有用信息,利用这些信息,可以对页面进行排序,发现重要的页面。

1.4 Web使用记录挖掘

Web使用记录挖掘通过对Web日志记录的挖掘,发现用户访问Web页面的模式。其主要目标是从Web的访问记录中抽取感兴趣的模式[5]。Web服务器自动记录了用户的访问日志,这些日志记录了用户的访问信息。利用这些信息,可以更有效地管理Web站点,调整Web站点结构,以及自动改进网站内容安排和帮助理解用户的行为,为用户提供更好的个性化服务。文中主要用到的是Web使用记录的挖掘,图1显示了其应用的领域[6]。

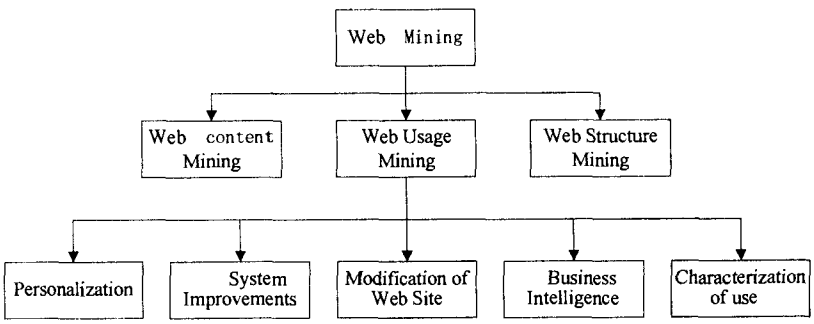

图1　Web挖掘的分类及使用记录挖掘的主要应用领域

2 基于Web挖掘的个性化远程教育模型

2.1 问题描述

实现个性化的远程学习系统的关键就是在学习的各个阶段,对个体进行差异化的分析和处理。首先需要对参加学习的个体情况进行分类,针对不同类别来安排相应的教学内容和学习进程;其次,在学习的过程中,知识表示的内容需要根据对学习者的个性要求的不同而具有不同的形式;最后就是要对学习者每一个阶段的学习进行相应的评估和反馈。在现今远程教育模式中,Web课件、教学录像和其他的一些学习资源都保存在服务器端,整个远程教育系统都以网页的形式在浏览器上运行,包括课程的学习、答疑、讨论、知识的检索、考试等,而每个学生仅仅需要通过一个Web浏览器来访问远程教育站点就可以进行在线学习,这就需要从以网站为中心到以用户为中心的转变。

对于用户使用记录,绝大多数Web服务器都采用通用日志格式(Common Log Format)来记录。日志的每一项由用户IP、用户名、访问时间、访问页面URL、传输量等特征标定。而交互数据库中的信息彼此之间也有不同,例如:作业(练习、测试)信息可以这样标定:<学习者标识,知识点标识,起始时间,完成时间,总用时间,有效时间,正确信息,错误信息,难度信息,……>。提问信息则可以这样标定:<学习者标识,知识点标识,提问时间,提问内容,回答时间,回答人,回答内容,满意程度,……>。正是由于这些不同,为了充分利用所有这些有用的信息,可以把日志记录和交互数据分别进行挖掘,而不用把它们规范化以至丢失了某些有价值的信息。

2.2 模型的提出

基于Web挖掘的个性化远程教育模型主要分为3个阶段:数据的获取阶段、使用挖掘阶段、挖掘结果的解释阶段,结构如图2所示。

该模型的特点是有两个基本平行的挖掘过程分别对访问日志和交互数据进行挖掘,使得每一个挖掘过程相对简单,又能各自独立地处理不同的数据源。这种对数据源的区分主要是基于其二者不同的特点:首先,二者来源不同,访问日志记录了关于用户访问的信息;交互数据则记录了用户和远程教育系统之间的交互信息。再者,对二者的兴趣点不同,日志记录中我们主要关心的是用户访问的行为方式,如:兴趣的转移、页面滞留时间、访问次数等;交互数据中我们主要感兴趣的是用户访问记录的具体内容,例如:作业、考试完成情况和答案,答疑时给老师提的问题,与同学交流的主题等。因此,用两个相对独立的挖掘模块对进行挖掘,更能充分利用这些信息来个性化用户群。

此外,该模型还体现了远程教育的以用户为中心的思想,从用户出发,最终反馈到用户,变被动为主动,能够尽可能地迎合每个学习者的浏览(学习)兴趣,并且不断调整自己来适应学习者浏览(学习)兴趣的变化,进行个性化教学服务。在模型中,对两种不同数据库一般会采用不同的挖掘算法,以期望使个性化得到最好的体现,我们的目标不是个性化到每个相似的用户群体,而是个性化到每个不同的用户,使每个用户感觉到整个教育过程就是专门为他(她)量身定做的一样。

2.3 模型的实现

模型的实施是建立在Web挖掘技术之上的。因此,数据的获取和挖掘依赖于Web挖掘技术的研究进展。目前,对Web使用记录挖掘的研究有很多,其过程的总体描述如图3所示[7]。从整个过程可以看到,首先要解决的问题就是数据的预处理,它主要包括如下两个部分:数

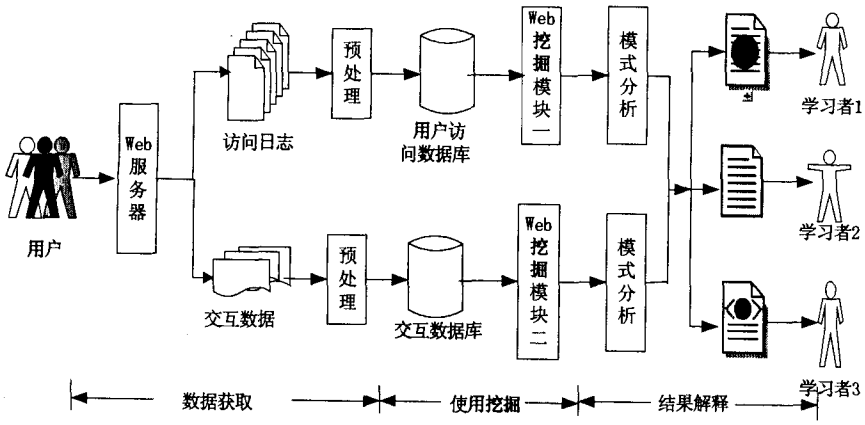

图 2 基于 Web 挖掘的个性化远程教育模型

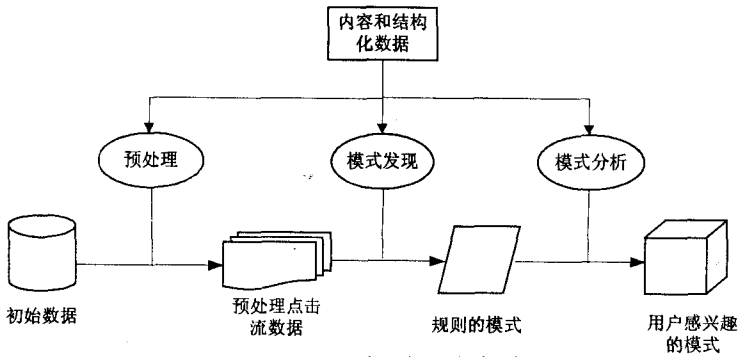

图 3 Web 使用记录挖掘过程

据清洗,包括无关记录的剔除、判断是否有重要的访问没有被记录、用户的识别等问题;事务识别,就是与要挖掘什么样的知识有关,将用户会话针对挖掘活动的特定需要进行事件定义。其次要解决的问题就是选择合适的挖掘算法,Web 使用记录挖掘中用到的 Web 日志分析及用户行为模式的挖掘方法,主要用了统计分析、关联规则、分类、聚类、序列模式等技术,这些挖掘技术相对来说比较成熟,可以用在我们的模型中。

交互数据的来源有很多:用户注册,作业、考试的完成过程与结果,答疑,学生在论坛中的讨论内容等等。对于这些用户数据,由于关注的是其内容信息,在对这些数据进行预处理时,要特别注意保持其语义特征的完整。因此,预处理主要是指抽取代表信息文本特征的元数据(特征项),对元数据进行标记、语形学分析、词性标注、短语边界辨认等。挖掘算法的选取,则可以借鉴 Web 内容挖掘中所采用的比较成熟的算法,如层次聚类[8,9]等。

挖掘结果的解释就是确定挖掘出的信息哪些是有价值的、将采取什么样的方式表示出来、怎样把它们反馈给用户等。当前流行的做法有两种:一种是对不同的用户,根据挖掘的结果提供不同的学习内容、进度安排等;再一种就是,在同一页面上根据不同的挖掘结果进行个性化的页面推荐[10,11]。在系统中,可以把这二者结合起来,让用户根据自己的学习情况可以有更多的选择。

3 结束语

目前,国内外对 Web 挖掘研究方兴未艾,还没有形成成熟的理论和统一的体系结构。因此,基于 Web 挖掘的个性化远程教育还处在发展阶段[12]。在本模型中,如何把对用户使用记录和对交互数据进行挖掘、模式分析后的信息特征统一起来,如何对同一用户不同解释的取舍将是下一步要做的工作。随着网络教育的发展,那种只是将传统课堂教育简单移植到远程教育上、教育系统模式单一、以系统自身为中心、学生只能被动地接受完全相同的学习内容、并没有真正体现出个性化教育优势的学习方式必将被逐步取代。新的远程教育系统会尽可能满足学习者的个性化需求,这种个性化服务势必会促进远程教育的进一步发展。

参考文献:

[1] 黄名选,冯 平.基于 web 挖掘的个性化远程教学模型研究[J].广西工学院学报,2005,6(3):68-72.

[2] 高 鹏,高 岭.基于 Web 挖掘的个性化算法及其在网络教学平台的应用[J].计算机应用,2005,25(5):1012-1015.

[3] Gravano L, Garcia-Molina H, Iomasic A. The effectiveness of gloss for the next database discovery problem[A]. SIGMOD'94[C]. Minneapolis, MN:[s.n.], 1994.126-137.

[4] 韩家炜,孟小峰,王 静,等.Web 挖掘研究[J].计算机研究与发展,2001,38(4):405-414.

[5] Srivastava J. Web usage mining: Discovery and application of usage patterns from Web data[A]. SIGKDD Explorations[C]. New York: ACM Press, 2000.43-56.

[6] Arayaa S, Silvab M, Weber R. A methodology for web usage mining and its application to target group identification[J]. Fuzzy Sets and Systems, 2004, 148:139-152.

[7] Chen M S, Park J S, Yu P S. Data Mining for Path Traversal Patterns in a Web Environment[A]. Proceedings of the 16th International Conference on Distributed Computing Systems[C]. Anaheim, California: ACTA Press, 1996.27-30.

[8] Willet P. Recent Trends in Hierarchical Document Clustering: A critical Review[J]. Information Processing and Management, 1988(24):5772-5971.

[9] 闫 莺,王大玲.支持个性化推荐的 Web 页面关联规则挖掘算法[J].计算机工程,2005,31(1):79-82.

[10] 石 建,孔祥成.论个性化信息提取中的 Web 挖掘技术[J].情报技术,2003(2):10-14.

[11] 石 晶,龚震宇,裘杭萍.基于 Web 使用挖掘的个性化服务系统[J].电子科技大学学报,2002,31(4):399-403.

[12] 梁开健.Web 挖掘在现代化远程教育中的应用[J].微机发展,2005,15(8):101-104.