Web数据挖掘在网络教育中的应用

范莉莎,刘 刚,刘志镜
(西安电子科技大学,陕西 西安 710071)

摘 要:随着互联网的发展,网络教育也有了长足的进展。Web数据挖掘以其独特的优点,在网络教育中有很多的用途。介绍了Web数据挖掘的概念的分类、网络教育中的数据资源以及网络教育中的Web数据挖掘的主要过程,并着重介绍了Web数据挖掘在网络教育中的具体应用。

关键词:Web数据挖掘;网络教育;个性化学习
中图分类号:TP393;G434 **文献标识码**:A **文章编号**:1673-629X(2006)06-0068-03

Application of Web Data Mining in Web-Based Education

FAN Li-sha, LIU Gang, LIU Zhi-jing
(Xidian University, Xi'an 710071, China)

Abstract: Under the development of Internet, Web-based education has made a quite great progress. Due to its unique advantages, Web data mining has been in a wide use in Web-based education. This paper introduced the classified concepts of Web data mining, data sources and the main process of Web data mining in Web-based education. Especially, the paper presents the actual applications of Web data mining in Web-based education.

Key words: Web data mining; Web-based education; personalized learning

0 引 言

随着互联网技术的应用和发展,基于Internet技术的网络教育逐渐成为有效利用社会优势教育资源的一种途径。这种教育网络化的趋势不仅为学生提供了便利的学习方式和广泛的选择,也为学校提供了更加深入了解学生需求信息和学生行为特征的可能性。由于受教育对象个体之间存在着极大的差异性,例如:个人学习目标不同、学习能力不同、认知风格不同,所以网络教育应该为不同的受教育者提供个别化的教育,网络教学也必须是一种适应个别化学习需求的个性化教学[1]。

这种个性化教学的提供,是通过将传统的数据挖掘(Data Mining)同Web结合起来,进行Web数据挖掘,即从Web文档和Web活动中抽取学生感兴趣的潜在的有用模式和隐藏的信息,作为对学生提供个性化教学服务的依据,协助管理者优化站点结构,提高站点效率,更好地为网络教育服务。

1 Web数据挖掘的基本概念及分类

Web数据挖掘(Web Data Mining),简称Web挖掘,是从数据挖掘发展过来的集Web技术、数据挖掘、计算机技术、信息科学等多个领域为一体的一项技术[2,3]。Web挖掘是指从大量的Web文档集合中发现蕴涵的、未知的、有潜在应用价值的、非平凡的模式。Web挖掘可分为3类[4]:Web内容挖掘(Web Content Mining)、Web结构挖掘(Web Structure Mining)和Web使用的挖掘(Web Usage Mining),如图1所示。Web所处理的对象包括静态网页、Web数据库、Web结构、用户使用记录等信息。

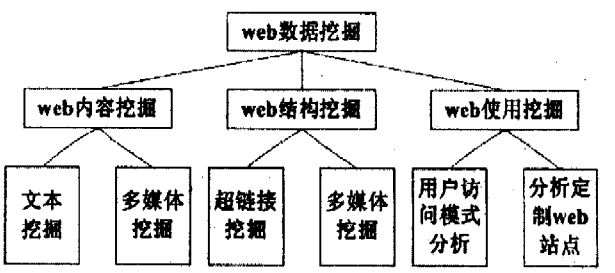

图1 Web挖掘分类

2 网络教育中的数据资源

2.1 服务器数据

学生访问服务器时会在服务器上产生相应的服务器数据,这些数据可以分为两种,即日志文件和查询数据。

(1)日志文件。

日志文件分为server logs,error logs和cookie logs。最常用的格式是标准公用日志文件格式和标准组合日志文

收稿日期:2005-09-01
作者简介:范莉莎(1978-),女,陕西人,硕士研究生,研究方向为远程教育、数字化校园;刘 刚,教授,研究方向为教育信息技术、现代远程教育;刘志镜,教授,博士生导师,研究方向为多媒体技术、远程教育、移动电子商务。

件格式。标准公用日志文件的格式存储关于学生连接的物理信息,如果能够对该文件中存储的一些项进行挖掘分析,就能发现学生的行为。

(2)查询数据。

查询数据是网络教育站点在服务器上产生的一种典型数据,它是在线学生在查询所需信息时生成的,这些查询信息通过Cookie或是登记信息连接到服务器的访问日志上。通常将查询数据和Cookie存入单独的日志中。

2.2 学生登记信息

它是指学生在Web页中输入并提交给服务器的信息,包括注册信息、登录信息、答疑信息、考试成绩、作业情况、交流信息和学习进度等。通过对学生登记信息和日志信息的综合,能够更好地了解学生的行为,并针对不同的学生制定不同的策略。

2.3 代理服务器数据

代理服务器相当于在客户浏览器和Web服务器之间提供了缓存功能的中介服务器,它的缓存功能减少了Web服务器的网络流量,加快了网页的运行速度,同时将大量的用户访问信息通过代理日志的形式保存起来。

3 网络教育中Web数据挖掘的过程

Web数据挖掘过程分为3个阶段:数据预处理、模式发现和模式分析[5]。

3.1 数据预处理

实际系统中的数据一般都具有不完全性、冗余性和模糊性,要使挖掘内核更有效地挖掘出知识,就必须为它提供干净、准确、简洁的数据。预处理主要对用户访问日志进行数据清洗、用户惟一性识别、用户会话识别、完善访问路径和事务识别等处理。

1)数据清洗。其目的是从服务器日志文件中消除不相关的项,缩小被挖掘数据对象的范围。实现方法可以通过查找URL地址名称的后缀,如:以.gif,.GIF,.JPEG,.JPG等为后缀的文件就可以移去。

2)用户惟一性识别。可以通过分析用户方Cookies文件和采用catch busting技术,并且借助其他一些信息来实现。如:对具有同一IP地址的用户,可以参考代理方参考日志文件中的信息来判断,若其中显示的用户使用的浏览器软件及操作系统是不同的,那么即使同IP地址的用户也被作为不同的用户而考虑。另外还可以参考网络站点的拓扑结构信息。

3)用户会话识别。目的是将每个用户的访问信息划分成若干个独立的会话进程,最简单的方法是采用超时估计的办法,即当对页面之间的请求时间间隔超出所给定值时,即可以认为用户已经开始了一次新的会话。

4)完善访问路径。由于存在客户端缓存,当用户使用浏览器的后退功能时会产生路径信息不完整的描述。解决这一类问题的方法类似于用户识别,如果一个页面请求信息与该用户上次请求的页面没有直接的链接关系,可以查看参考日志文件来决定这个页面来自于哪个页面的链接。

5)事务识别。事务识别建立在对用户会话识别的基础上,目的是依据数据挖掘任务的需求将事务做分割或合并处理,使其适合于数据挖掘需求的分析。它要求以一组事务列表或一些参数作为输入条件,输出为一组与输入同格式的事务列表。可以采用3种分割方法来实现事务识别:参考时长法、最大前向参考和时间窗法。

3.2 模式发现

模式发现阶段就是利用挖掘算法挖掘出有效的、新颖的、潜在的、有用的及最终可以理解的信息和知识。可用于Web数据挖掘的技术有路径分析、关联规则、序列模式、分类聚类技术和依赖性建模,其中路径分析技术是Web使用挖掘所特有的。

1)访问路径分析。访问路径是用户在网络上浏览时,从一个网页到另一个网页的路径。访问路径分析就是通过对Web服务器的日志文件中学生访问站点的访问次数和路径进行分析,从图中确定最频繁的路径访问模式或大的参引访问序列。图最直接的来源是网站结构图,其他图也都是建立在页面和页面之间的联系,或者是一定数量的用户浏览页面顺序基础之上的。它可以被用于判定一个Web站点中最频繁访问的路径,例如通过路径分析可以得出:85%的学生存取这个站点是从/course/network开始的;65%的学生在浏览3个或更少的页面后就离开了。

2)关联规则的发现。就是要找到客户对网站上各种文件之间访问的相互联系。可以用Apriori算法,从事务数据库中挖掘出最大的频繁访问项集,这个项集就是关联规则挖掘出来的用户访问模式。关联分析的目的就是为了挖掘出隐藏在数据间的相互关系,在网络教育中关联规则的发现就是找到学生对网站上各种知识之间访问的相互联系。如:40%的学生访问Web页面/course/network时,也访问了/course/java。即学生对某一知识点感兴趣,那么他也可能对相关的知识点也会留意。这有利于学校更好地组织站点,为学生减少过滤信息的负担。

3)序列模式的发现。就是在时间戳有序的事务集中,找到那些"一些项跟随另一个项"的内部事务模式。如:在/course/web programming上进行在线浏览学习的学生中,有60%的人在过去15天内也在/course/network处浏览过。在网站服务器日志里,学生的访问是以一段时间为单位记载的,经过数据清洗和事务处理后是一个间断的时间序列,这些序列所反映的学生行为有助于帮助学校印证其课程所处的学习周期阶段。

4)分类和聚类技术。分类技术可以从个人信息或共同的访问模式中得出访问某一服务器文件的用户特征。分类分析法的输入集是一组记录集合的几种标记。首先为每一个记录赋予一个标记,即按标记分类记录,然后检查这些标定的记录,描述出这些记录的特征。在网络教育中通过分类技术对学生进行细分,得到分类后,就可以针

对这类学生的特点提供个性化的服务。另外,通过学生登记信息、在线调查表也可以得到学生的一些特征。分类可以通过决策树技术、贝叶斯分类法、k－相似相邻分类等技术实现。

聚类分析与分类不同,它的输入集是一组未标定的记录,可以从 Web 访问信息数据中聚集出具有相似特性的学生。聚类分为对学生群体的聚类和 Web 页面的聚类。其中学生群体的聚类在网络教育和用户提供个性化服务的应用中起着很重要的作用。通过分组聚类出具有相似浏览行为的学生,并分析学生的共同特征,更好地帮助网络教育的教师了解自己的学生,向学生提供更合适的教学服务。

5) 依赖性建模。建模的目标是开发出一种能表达出 Web 领域中各种变量之间显著依赖性的模型。例如,在教育网站中一名学生从经常访问的常客到潜在的注册选课学习者,这个行为选择过程,也许会经历几个不同的阶段。有几种概率学习方法可以用来为用户的浏览行为建模,如:隐马尔可夫链模型、贝叶斯信念网络等。Web 使用模式的建模不仅能为分析用户行为提供理论框架,还具有预测 Web 资源消耗的潜力。

3.3 模式分析

模式分析主要是为了从模式发现算法找到的模式集合中筛选出有趣的模式。精确的分析方法通常是由 Web 挖掘的具体应用来控制的。模式分析的形式可以是 SQL 那样的知识查询机制,也可以把 Web 使用数据装入数据仓库,以便执行 OLAP 操作。诸如图形化模式或为不同值赋不同颜色的可视化技术,可以使得数据中的总体模式或趋势变得更加直观。

4 Web 数据挖掘在网络教育中的应用

数据挖掘最早起源于商业的直接需求,把它放在网络教育领域中,同样也有着广泛的应用。下面主要介绍几种应用。

4.1 了解学生

随着"以学生为中心"、"以人为本"教育理念的不断深化,分析学生、了解学生,针对不同学生提供"量身定做"的课程设置,引导学生的需求已成为学校工作的重要课题。通过对网络教育系统收集的数据进行分类分析,可以按各种学生层次对学生分类,然后确定不同类型学生的知识需求,以便提供相应的课程设置,促使网络教育优势的最大化。同时有利于提高学生的满意度,最终达到留住学生的目的。

4.2 个性化学习

根据学生的注册信息和需求纪录,系统可以向学生显示那些可能引起学生特殊兴趣的新知识。当学生注意到下一知识点时,系统会建议一些在学习中会用到的知识点和相关知识。一般的知识点常常简单地按类型对知识进行分组,以简化学生在选择中的步骤。然而对于在线辅导,知识点分组可能是完全不同的,它常常以针对学生的需求知识为基础。针对不同学生的个性化学习,强调信息的个性化,亦即识别、建立、调整学生的喜好,使客户能以自己的方式来访问。不仅考虑学生看到的知识点,而且还考虑学生已掌握的知识,结果就会使服务更加个性化。

4.3 改进网站设计

教育网站的设计者可以不再完全依靠专家的定性指导来设计网站,而是根据访问者的信息来修改和设计网站结构和外观;找出如何优化一个网站组织结构的策略;确定预传哪些页面到客户端,从而提高网站的效率。例如,可以根据学生的访问路径,找出学生访问模式的频繁路径,再根据这个路径来改进网站的结构和网页的链接,有助于节约学生的访问时间,也节约了网站的开支。

4.4 分析需求趋势

分析学生浏览学习的历史资料不仅可以预测学生学习需求趋势,还可以评估需求倾向的改变,有助于提高教育网站的竞争力。通过 Web 数据挖掘得来的资料,及时调整网站的课程设置与专业设置,满足广大学生的需求,留住现有的学生,同时吸引更多的学生来到本网站进行学习。

5 结束语

Web 数据挖掘是在传统的数据挖掘的基础上发展起来的一门综合技术,它主要致力于在网络上海量的异构的信息资源中寻找蕴涵的有价值的知识。近年来,随着网络教育的迅速发展,Web 数据挖掘有了更多的用途,它能够有效地帮助校方扩大影响,吸引学生,同时还可根据学生的实际情况及时调整教学计划,以满足广大同学多层次、多方位的需要。相信 Web 数据挖掘在网络教育中会发挥更大的作用。

参考文献:

[1] 雍全明 . 基于 Web 的远程教育参考模型研究[J]. 鄂州大学学报,2003(10):28－29.

[2] 姜传菊 . 试论 Web 中的数据挖掘[J]. 网络资源与建设,2003(年刊):162－164.

[3] 黄 茜 .Web 日志挖掘在个性化网络教育中的应用[J]. 现代教育技术,2004(5):52－55.

[4] 严华云 .Web 挖掘在网络教育中的应用研究[J]. 湖州师范学院学报,2003(6):72－75.

[5] 张 娥,冯秋红,宣慧玉,等 .Web 使用模式研究中的数据挖掘[J]. 计算机应用研究,2001(3):80－83.