Web使用模式挖掘技术

张贵红
(乐山师范学院,四川 乐山 614000)

摘 要:Web挖掘一般可以分为3类:Web内容挖掘、Web结构挖掘和Web使用挖掘。WWW上信息资源的爆炸性增长,Web挖掘已经成为计算机科学的一个重要研究领域。使用模式挖掘是Web挖掘的一个分支,它利用Web服务器的日志中的大量数据及其他相关数据集进行分析挖掘,并从中获得有价值的有关网站访问使用情况的模式知识。对web数据挖掘作了比较详细的介绍,并介绍了Web使用挖掘各阶段的主要工作以及相关技术。

关键词:web挖掘;使用模式;挖掘步骤

1 Web数据挖掘

1.1 Web数据挖掘的概念

数据挖掘是近年来随着数据库技术和人工智能技术的发展而出现的一种全新的信息技术,是指从数据中提取模式的过程。数据挖掘可简单地理解为从数据中挖掘有用的信息。Web数据挖掘,简称Web挖掘,是数据挖掘技术在Web环境下的应用,是集Web技术、数据挖掘、计算机技术、信息科学等多个领域的一项技术。

1.2 Web挖掘研究的内容

1.2.1 个性化特征及推荐系统。解决如何挖掘顾客的个性规律,如对于在线电子商店,了解顾客的个性经验是吸引长期顾客的关键因素。通过Web日志文件中的浏览数据来挖掘顾客的浏览个性,并用这些有价值的数据来提高顾客对网站的浏览效率。

1.2.2 挖掘框架体系及案例研究。集中了来自实际应用数据挖掘技术的厂商在构建系统体系时所要解决的问题,并给出了各自的原型。

1.2.3 用户浏览分析。如何对用户的浏览数据进行挖掘。如得到有价值的电子商务信息,从而了解用户在决定是否购买产品时的细节行为。

1.3 Web数据挖掘的分类

一般地,Web挖掘可以分为Web内容挖掘(Web ContentMining),Web结构挖掘(Web Stricture Mining)和Web使用模式挖掘(Web Usige.schema Mining)三类。

1.3.1 Web内容挖掘是从文档内容或其描述中抽取知识的过程。Web内容挖掘有两种策略:直接挖掘文档的内容和在其它工具搜索的基础上进行改进。根据挖掘处理的数据可以将Web内容挖掘分为文本挖掘和多媒体挖掘两个部分。

1.3.2 Web结构挖掘是从Web组织结构和链接关系中推导知识。挖掘页面的结构和Web结构,可以用来指导对页面进行分类和聚类,找到权威页面、中心页面,从而提高检索的性能。同时还可以用来指导网页采集工作,提高采集效率。Web结构挖掘分为Web文档内部结构挖掘和文档间的超链结构挖掘。

1.3.3 Web使用模式挖掘是从服务器端记录的用户访问日志或从用户的浏览信息中抽取感兴趣的模式,通过分析这些数据可以帮助理解用户的行为,从而改进站点的结构或为用户提供个性化的服务。

2 Web使用模式挖掘

Web使用模式挖掘是在用户访问Web后,对服务器上留下的访问路径进行挖掘,即对用户访问Web站点的存取方式进行挖掘。挖掘的目的是在海量的Web日志数据中自动、快速地发现用户的访问模式,如频繁访问路径、频繁访问页组、用户聚类等。

2.1 Web使用模式挖掘常用技术

Web使用模式挖掘中常用以下一些技术:

2.1.1 关联规则挖掘技术(Associate Mining Technology)。在Web数据挖掘中,关联规则挖掘就是要挖掘出用户在一个访问期间(session)从服务器上访问的页面或文件之间的联系。

2.1.2 序列模式挖掘技术(Sequence Mining Technology)。序列模式挖掘就是要挖掘出交易集之间的有时间序列的模式。在网站服务器日志里,用户的访问是以时间段为单位记录的,经过数据清洗和事务识别以后是一个间断的时间序列。这些序列所反映的用户行为有助于网站确认用户访问网站的兴趣所在。

2.1.3 分类与聚类技术(Classification & Clustering)。分类规则可以挖掘Web日志中某些共同的特性,利用该特性对新添到数据库里的数据项进行分类,根据访问模式得出访问某一服务器文件的用户特征。聚类分析用于将有相似特性的用户、数据项集合到一起。聚类的目标是将大量的数据项聚集成类,使得类与类之间的相似度尽量小,而类内的相似度尽量大。

2.1.4 路径分析技术(Route Analysis Technology)。在Web使用模式挖掘过程中,通过路径分析技术可以确定网站的频繁访问路径,可以对频繁访问的路径进行优化,并可以在频繁访问的路径上放置重要的信息,如导航信息等,以方便用户使用。通过路径分析技术得出的网站结构图在模式挖掘中非常有用。

2.2 Web使用模式挖掘流程

Web使用模式挖掘主要是存在于服务器日志中的用户访问信息,它将数据挖掘技术应用到Web中,形成了自己的挖掘方式。一般对Web使用模式挖掘流程的划分可分为三步和四步两种不同的看法。三步法认为应分为数据准备阶段、模式发现阶段和模式分析阶段。四步法是将流程分为源数据收集、数据预处理、模式挖掘和模式分析四个阶段。因为源数据收集和数据预处理可以归并为数据准备,所以本文采用三步划分法。其流程如图1所示。

Web使用模式挖掘是从用户浏览网站的数据中抽取感兴趣的模式,理解用户的浏览兴趣行为,以便进一步改善网站结构或为用户提供个性化的服务。

Web使用模式挖掘必须解决以下两个基本问题:

2.2.1 如何准确收集用户身份、访问行为、访问频度、访问内容等浏览信息。

2.2.2 如何正确度量和表达用户的浏览兴趣。

2.3 基于用户浏览行为的挖掘步骤

2.3.1 数据准备:采集用户的浏览信息,并将用户信息记录到用户浏览行为库,数据进行清洗,滤掉脏数据,识别用户,提取关键字。数据清洗是指删除采集来的Web日志中与挖掘算法无关的内容,包括图片、框架等非用户请求单位、robot浏览日志记录以及一些噪声、错误数据等。

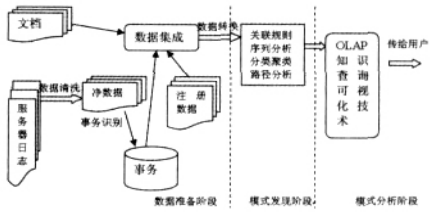

图1 web使用模式挖掘流程

2.3.2 用户兴趣度计算:计算用户浏览过的网页的兴趣度,并生成网页兴趣度矩阵,以便于为进一步的计算做准备。根据已经计算出的网页兴趣度矩阵计算主题兴趣度,根据已经计算出的网页兴趣度矩阵计算关键词兴趣度,并将计算结果按兴趣度大小依次排列。

2.3.3 推荐:根据挖掘结果实施推荐,将用户感兴趣的内容的链接添加到用户正在浏览的网页。

2.3.4 结果修正:根据用户浏览行为的反馈对推荐进行修正,如果用户接受推荐,则进行巩固;否则,根据用户的反馈进行重新推荐。

结束语

Web使用模式挖掘是一个对服务器日志的挖掘,它旨在得出日志中有用的用户访问信息,以使网站有针对性地完善自身,能更好地服务用户并取得较好的经济效益。本文对Web数据挖掘作了比较详细的介绍,并对Web使用模式的挖掘作了较深入的探讨。

参考文献

[1]何波,李建国.基于XML的WEB数据挖掘系统框架的设计与实现[J].西南师范大学学报:自然科学版.
[2]葛昕,黄永慧,陈锐.WEB使用模式挖掘系统的设计与实现[J].柳州师专学报.
[3]乔智勇,刘志镜.WEB数据挖掘系统的设计及实现研究[J].计算机工程与设计.
[4]许建潮,王颖楠,胥桂仙.WEB文本信息抽取与挖掘方法[J].长春工业大学学报.

作者简介:张贵红(1973~),女,四川乐山市人,乐山师范学院计算机系,讲师,硕士学位。

责任编辑:周宝军