第16卷 第8期
2006年8月

计算机技术与发展
COMPUTER TECHNOLOGY AND DEVELOPMENT

Vol.16 No.8
Aug. 2006

Web挖掘及其在电子商务中的应用研究

林瑞娟,侯德文
(山东师范大学 信息科学与工程学院,山东 济南 250014)

摘 要:作为一种崭新的信息处理技术,Web挖掘受到了人们极大的关注。电子商务是一种新型的现代商务模式,如何将Web挖掘应用于电子商务,来有效地处理信息,成为企业共同关注的问题。文中介绍了Web挖掘的概念和分类,阐述了Web挖掘在电子商务中的挖掘方法和过程,讨论了Web挖掘在电子商务中的典型应用,并就一个事例进行了具体的分析。

关键词:数据挖掘;Web挖掘;电子商务
中图分类号:TP393 **文献标识码**:A **文章编号**:1673-629X(2006)08-0186-03

Web Mining and Its Applications in Electronic Commerce

LIN Rui-juan, HOU De-wen
(School of Information Science and Engineering, Shandong Normal University, Jinan 250014, China)

Abstract: As a brand new technology to transact information, Web mining is being payed maximum attention. Electronic commerce is a new business mode. Applying data mining to electronic commerce has become a hot issue. This paper introduces Web mining, explains its mining method and process, and sets forth its typical application in electronic commerce. In the end, an electronic commerce example is given in detail.

Key words: data mining; Web mining; electronic commerce

0 前 言

近年来,随着Internet的飞速发展和广泛普及,使得Web上的信息以惊人的速度增长。面对Web上的海量、动态、异质的信息,传统的工具很难满足人们的需要。为了解决这个问题,人们将传统的数据挖掘技术与Web结合起来,产生了一种新的挖掘技术——Web挖掘。

电子商务是指个人或企业通过计算机网络,采用数字化电子方式进行商务数据交换和开展商务业务活动的现代化商业模式。目前国内已有网上商情广告、电子票据交换、网上订购、网上银行、网络结算等多种类型的电子商务形式,电子商务正以其低廉、方便、快捷、安全、可靠、不受时间和空间的限制等突出点而逐步在全球流行[1]。为了更好地利用这一现代商业手段,人们把数据挖掘应用于电子商务系统,极大地推动了电子商务的发展。

1 Web挖掘概述

1.1 Web挖掘定义

Web挖掘就是从Web文档和Web活动中发现、抽取感兴趣的潜在的有用模式和隐藏的信息。它以从Web上挖掘有用信息为目标,以数据挖掘、文档挖掘、多媒体挖掘为基础,并综合运用计算机网络、数据库与数据仓库、人工智能、信息检索、可视化、自然语言理解等技术,将传统的数据挖掘与Web结合起来[2]。

1.2 Web挖掘分类

一般的,Web挖掘可分为3类: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)。图1给出了Web挖掘的分类图。

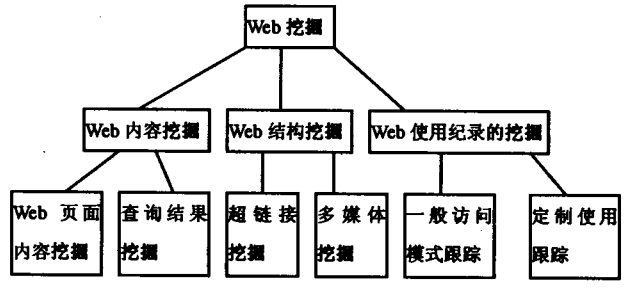

图1 Web挖掘的分类

1)Web内容挖掘:又可以分为Web页面内容挖掘和查询结果挖掘。页面内容挖掘指的就是对Web页面上的数据进行挖掘,而搜索结果挖掘则指的是以某一搜索引擎为基础,对已搜索结果的挖掘,以得到更精确有用的信息。

2)Web结构挖掘是从WWW的组织结构和链接关系中推导知识[3]。Web结构挖掘的目的:通过聚类和分析网页的链接,发现网页的结构和有用的模式,找出权威页

收稿日期:2005-11-30
基金项目:山东省中青年科学家奖励基金(03BS009)
作者简介:林瑞娟(1981-),女,山东青岛人,硕士研究生,研究方向为数据挖掘、数字图像处理;侯德文,副教授,硕士生导师,研究方向为数字图像处理。

面。

3)Web使用记录挖掘是对用户访问服务器时留下的信息进行挖掘的过程。在WWW中保存了访问日志,分析这些数据可以帮助理解用户的行为,从而改善站点结构或为用户提供个性化的服务。

2 电子商务中的Web挖掘

2.1 电子商务中的Web挖掘数据资源

在Internet电子商务中,客户访问服务器或代理服务器,就会产生相应的服务器数据,包括日志文件和查询数据,并保存在日志文件中,具体数据源有以下几种形式。

2.1.1 服务器端的网页数据以及日志文件

Web结构挖掘是从WWW的组织结构、Web文档结构和链接关系中挖掘有用信息的过程。Web结构挖掘的目的是发现页面的结构,在此基础上对页面进行分类和聚类从而找到权威页面。WWW上每一个提供信息资源的服务器上都有Web访问日志,用于记录和积累关于用户交互作用的数据。

2.1.2 代理服务器端数据

代理服务器[4]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。对此类信息的分析也有助于对客户浏览习惯和目标的归纳和推测。

2.1.3 客户登记信息

客户登记信息[4]是指客户通过Web网页在屏幕上输入并提交给服务器的相关信息。若新客户到商城购物,首先要注册登录,进入购物页面,然后才能购买商品,支付费用。想要更好地了解客户,必须将客户的登记信息和访问日志结合起来分析,才能得出更准确的判断,进而提供更完善的服务。

2.2 电子商务中的Web挖掘数据方法

电子商务中的Web挖掘数据方法[5]分为以下5种:

(1)路径分析:它可以被用于判定在一个Web站点中最频繁访问的路径。还有一些其它的有关路径的信息通过路径分析也可以得出。通过路径分析,可以改进页面及网站结构的设计。

(2)关联规则的发现:关联规则就是为了发现事物之间的意义的联系和规则。进行Web上的数据挖掘,构建关联模型,可以更好地组织站点,减少用户过滤信息的负担。

(3)序列模型的发现:序列模式分析的侧重点在于分析数据间的前后或因果关系。就是在时间有序的事务集中,找到那些"一些项跟随另一些项"的内部事务模式。发现序列模式能够便于电子商务的组织者预测客户的访问模式,对客户提供个性化的服务。

(4)分类规则的发现:数据分类是基于数据的某些属性的值进行的。数据分类方法很多,最为典型的是基于决策树的分类方法。它是从实际数据中构造决策树,是一种有指导的学习方法。得到分类后,就可以针对这一类客户的特点展开商务活动,提供有针对性的个性化的信息服务。

(5)聚类分析的发现:聚类分析法输入集是一组未标定的记录。其目的是根据一定的规则,合理地划分记录集合,并用显式或隐式的方法描述不同的类别。在电子商务中通过聚类具有相似浏览行为的客户,使管理员更多地了解客户,提供更适合、使客户更满意的服务。

2.3 电子商务中的数据挖掘的过程

电子商务中的数据挖掘的过程[1]一般由以下几个主要的阶段组成:数据准备、挖掘操作、结果表达和解释。

(1)数据准备:这个阶段又可进一步分成几个子步骤:数据集成、数据清理、数据预处理。数据集成将多文件或多数据库运行环境中的数据进行合并处理,解决语义模糊数据准备。数据清理目的是去除不相关的记录,找出需要分析的数据集合,缩小处理范围,提高数据挖掘的质量。预处理是为了克服目前数据挖掘工具的局限性。

(2)数据挖掘:这个阶段进行实际的挖掘操作,包括的要点有:决定如何产生假设;选择合适的工具;发掘知识的操作;证实发现的知识。

(3)分析结果:根据最终用户的决策目的对提取的信息进行分析,把最有价值的信息区分开来,并且通过决策支持工具提交给决策者。因此,这一步骤的任务不仅是把结果表达出来,还要对信息进行过滤处理,如果不能令决策者满意,需要重复上述过程。

2.4 Web挖掘在电子商务中的应用

电子商务的产生,改变了企业的经营理念,给社会的各个行业带来了巨大的变化,将成为引导经济发展的新潮流。数据挖掘的应用将极大地提高企业获取信息的能力,使企业信息资源的价值得到充分地体现。它有利于促进企业开拓市场,优化企业资源,提高企业的经营效率和管理水平,为企业资源计划(ERP)、客户关系管理(CRM)、产品数据管理(PDM)和商业信用评估等提供有效的技术途径[6]。

2.4.1 数据挖掘在企业资源计划中的应用

企业资源计划的实施,将促进企业管理的集约化与现代化、企业资源的重组与优化。通过对Web数据挖掘,快速提取商业信息,使企业准确地把握市场动态,极大地提高企业对市场变化的响应能力和创新能力。运用数据挖掘技术,极大地提高公司对过程控制、生产制造和资源管理的能力。使企业最大限度地利用人力资源、物质资源和信息资源,合理协调企业内外部资源的关系,产生最佳的经济效益。

2.4.2 数据挖掘在客户关系管理中的应用

客户关系管理是一种以客户为中心的经营策略,它通过现代信息技术,充分利用客户信息,挖掘有用的商业知

识,指导企业的产品开发、市场营销和管理决策,提高企业的市场竞争能力。基于数据挖掘技术,企业将最大限度地利用客户资源,开展客户行为的分析与预测,对客户进行分类。有助于客户盈利能力分析,寻找潜在的有价值的客户,开展个性化服务,提高客户的满意度和忠诚度,使企业与客户的关系及企业利润得到最优。

2.4.3 *数据挖掘在产品数据管理中的应用*

通过数据挖掘技术来分析产品质量的主要影响因素,选取设计参数,优化产品结构以及成分组合。建立产品质量控制模型,全面提高产品生产和制造的质量,进行产品可靠性分析,对质量超越、零件失效、工艺偏离等情况进行记录和追踪,改善产品的生产工艺流程,减少产品的质量缺陷或质量偏差。通过市场预测对产品的需求能力和需求动向,进行产品性能和产品工艺的创新与改造,从而开发新的产品样式,拓宽产品销售市场。

2.4.4 *数据挖掘在商业信用评估中的应用*

低劣的信用状况是影响商业秩序的突出问题,已经引起世人的广泛关注。由于网上诈骗现象层出不穷,企业财务"造假"现象日益严重,信用危机成为制约电子商务发展的重要因素。利用数据挖掘技术对企业经营进行跟踪,开展企业的资产评估、利润收益分析和发展潜力预测,构建完善的安全保障体系,实施网上全程监控,强化网上交易和在线支付的安全管理。基于数据挖掘的信用评估模型,对交易历史数据进行挖掘,发现客户的交易数据特征,建立客户信誉度级别,有效地防范和化解信用风险,提高企业信用甄别与风险管理的水平和能力。

2.5 电子商务结构组件

由于目前没有一种单一的电子商务网站的标准体系结构,所以电子商务网站的一个普遍需求是需要支持不同的商务活动。不同任务要由不同模块来支持,其中每个模块采用最合适的格式维护它的数据,同时和其它模块共享元数据。电子商务活动包括以下几部分:广告促销,市场交易,报表和分析、数据挖掘、客户交互。这 5 个活动互相重叠,互相影响并依赖于共享信息。这会很自然地有一个连接组件的结构,每一个都有它自己组织共享数据的方式,但每个活动都能和其它活动通信,而且都有访问共享元数据的权限。

2.6 电子商务 IT 体系结构的一个实例

文中采用文献[7]中的一个例子——蓝色马丁尼电子商务平台,它是供包括 Harley Davidion, Gymboree, Sacks Fifth Avenue, Canadian Tire, Levi's 和 Virgin Wine 等网站在内的网络零售商使用的电子商务平台。

蓝色马丁尼的体系结构是为支持不同市场交易商、推销商和数据挖掘者的需求而设计的,如图 2 所示。对广告推销商来说,这个体系结构支持通过许多的商品层次和工具来控制收集和促销。对市场交易商来说,有些工具可以用来做一些受控试验,跟踪各类广告词和市场规则的有效性。对于数据挖掘者来说,集成了建模软件,无需再靠手工在数十个不同的服务器和应用日志里来收集客户的签名。

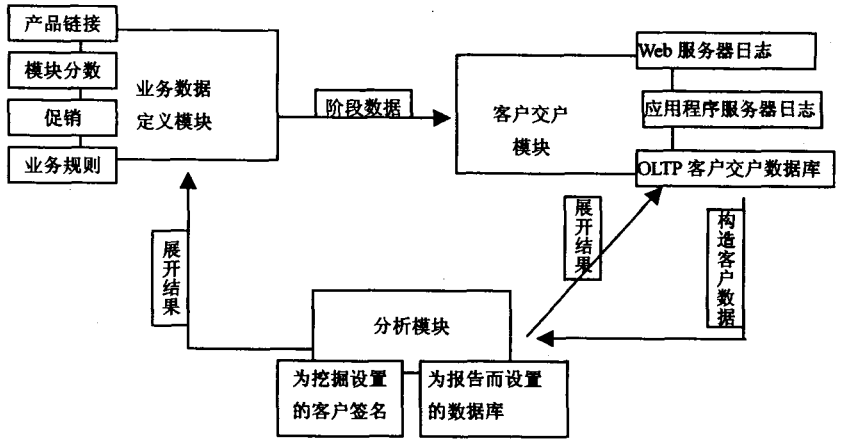

图 2 蓝色马丁尼的电子商务结构

蓝色马丁尼商店里的客户通过网页进行交互。所有客户的交互通过数据中心来管理。报表和挖掘由客户行为数据中心来决定,数据中心包含由客户交户、商品和业务规则数据中心所产生的信息。从交易数据中来的客户签名需要复杂的析取和逻辑变换,这些都是系统的一部分——对于那些曾经对网络日志尝试过数据协调以获得客户信息的人来说,这是一个特别的简化。

系统有 3 个主要的组件,每个组件都有自己的数据中心。这些数据分别跟踪如下事务:业务规则、客户和访问事务、客户行为。

客户行为数据中心,在图 2 中是分析组件的一部分,数据来自于客户交户模块,同时,它给业务数据定义模块和客户交互模块提供规则。业务规则包括问候客户的个性化、商品促销、越区销售等规则。

客户交互模块是系统中和客户直接面对面的一部分,交互通过处理所有客户事务来进行。客户交互模块负责维护客户交易和环境。这个组件实现了实际网络商店,并收集了任何后续分析所需的数据。客户事务数据中心记录客户的选择、支付流程等商务事件的日志。客户交互组件对自己所服务对象有具体认识,可以跟踪许多在网络服务器日志里不能跟踪的事项。客户交互组件通过所收集的数据可以随时跟踪商品和客户。支持客户交互模块的数据库,是一个标准化的支持快速事务处理的关系数据库,输入到分析模块的数据必须要经过析取和变换,以支持建立适合于进行挖掘和报表的结构。

数据挖掘通常需要平面表,每个待研究的客户或商品在表里占有一行。这意味着整平了商品层次的交换,以至于同样的事务可以产生标志。另外的数据来自订单文件、

(下转第 191 页)

$V = \frac{2(|x|-|y|)+\text{const}}{|x|+|y|}$ 小,那么 V 就越小。

Step6: 最终两幅图像的距离定义为:Dist = $D(1+V)$。

4 算法举例

(1) 假设有两个点集 X, Y。

$X = \{x_1, x_2, x_3, x_4, x_5\}$, $Y = \{y_1, y_2, y_3, y_4, y_5\}$, 它们之间的距离矩阵为 $D(i,j)$(见图 3)。

距离 $D(i,j)$	y_1	y_2	y_3	y_4	y_5	该行的最小值
x_1	3	5	5	4	1	1
x_2	2	2	0	2	2	0
x_3	2	4	4	1	0	0
x_4	0	1	1	0	0	0
x_5	1	2	1	3	3	1

图 3　两个点集的距离矩阵

(2) 求出每一行的三元组。

第一行$(x_1, y_5, 1)$; 第二行$(x_2, y_3, 0)$; 第三行$(x_3, y_5, 0)$; 第四行$(x_4, y_1, 0)$, $(x_4, y_4, 0)$, $(x_4, y_5, 0)$; 第五行:$(x_5, y_1, 1)$, $(x_5, y_3, 1)$。

(3) 计算距离。

Dxy 为所有三元组的平均距离, $Dxy = 0.375$; Dyx 为利用上述方法求得的 $D(i,j)$ 逆矩阵的平均距离 $Dyx = 0.167$。$D = \max(Dxy, Dyx) = 0.375$。

$$V = \frac{2(5-5)+1}{5+5} = 0.100$$

Dist = $D(1+V) = 0.4125$

Dist 的值越小,代表两幅图像越相似,最终检索出来的图像根据 Dist 的值从小到大排列就得到相似度从大到小的图像序列。

试验结果表明,结果受角检测算法的影响较大,角检测越精确形状匹配就越好。

5 结束语

利用角去表示图像内容的形状,提出利用图论中完美对集的方法和改进的 Hausdorff 距离去测量两幅图像内容形状的相似性,取得了良好的试验结果。但是点集的匹配问题是非常复杂的问题,对旋转、缩放等问题还有待于做进一步的研究。

参考文献:

[1] He X C, Yung N H C. Curvature Scale Space Corner Detector with Adaptive Threshold and Dynamic Region of Support[A]. 17th International Conference on Pattern Recognition[C]. Cambridge, UK:[s.n.],2004.791-794.

[2] Mokhtarian F, Suomela R. Robust image Corner Detection Through Curvature Scale Space[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(12): 1376-1381.

[3] Lowe D G. Object Recognition from Local Scale-Invariant Features[J]. ICCV,1999(15):1150-1157.

[4] Schmid C, Mohr R. Local Gray Value Invariants for Image Retrieval[J]. IEEE PAMI,1997,19:530-534.

[5] 邦迪 J A, 默蒂 U S R. 图论及其应用[M]. 北京:科学出版社,1984.

(上接第 188 页)

会计文件和交易纪录,而且一个客户可能有多个交易事务。通过这些方式将获得的典型数据有分组总开销、平均订单数、个别客户平均订单与总平均订单的差异,以及客户上次前来购买的日期。

报表由支持在各种级别下进行回溯查询的多维数据库完成。数据挖掘和 OLAP 两者都是分析模块的组成部分,但它们解决不同的问题。OLAP 查询可以解决:商品是否畅销;客户最感兴趣的商品;销售量最高的网站等。而数据挖绝解决更复杂的问题:高额消费者的特征;客户和表是否匹配;客户一个月内再回来购物的可能性。

3 结束语

电子商务是现代信息技术迅速发展的必然产物,也是未来企业模式的必然选择。将数据挖掘引入电子商务,增强企业的商务智能,使能向客户提供个性化的服务,将是使电子商务取得更多成就的必然方向。如何更有效地利用数据挖掘解决电子商务中的问题,是电子商务急需解决的重要方面。

参考文献:

[1] 郝先臣,张德干,尹国成,等. 用于电子商务中的数据挖掘技术研究[J]. 小型微型计算机系统,2001,22(7):785-788.

[2] 涂乘胜,鲁明羽,陆玉昌. Web 挖掘研究综述[J]. 计算机工程与应用,2003,39(10):90-93.

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

[4] 万　军,耿东辉. 浅说电子商务中的数据挖掘技术[J]. 东北大学学报(自然科学版),2004,25(增刊 2):194-196.

[5] 蒋良孝,蔡之华. 电子商务中的数据挖掘及其应用[J]. 计算机工程与设计,2003,24(6):74-77.

[6] 黄解军,万幼川. 基于数据挖掘的电子商务策略[J]. 计算机应用与软件,2004,21(7):12-13.

[7] Linoff G S, Berry M J A. Web 数据挖掘:将客户数据转化为客户价值[M]. 沈钧毅,等译. 北京:电子工业出版社,2004.