文章编号:1001 - 9081(2005)12 - 2940 - 03

基于产品 3D 模型点击流的客户行为分析

谢 凌,陈新度,陈 新
(广东工业大学 信息工程学院,广东 广州 510009)
(lilyly_x@tom.com)

摘 要:将 Web 服务日志进行扩展,以基于 Web 的产品 3D 模型浏览的操作日志和客户基本信息作为源数据,建立了客户操作模型,采用 FP-growth 关联规则分析方法,挖掘客户对产品及产品系列的兴趣度,为企业制定相应的产品开发策略和市场营销策略提供决策支持。

关键词:点击流;日志扩展;FP-growth;兴趣度
中图分类号: TP391 **文献标识码**:A

Analysis of the customer's behavior based on clickstream of product's 3D model

XIE Ling, CHEN Xin-du, CHEN Xin
(*College of Information Engineering, Guangdong University of Technology, Guangzhou Guangdong* 510009, *China*)

Abstract: The Web Server log files were extended and an operation model based on user's operation log files and customer's information was established, the interestingness of customer in product and product series was mined using the FP-growth method. The results shows that the models and methods are useful for dicision-making of product development and marketing strategies in enterprises.

Key words: clickstream; extended log files; FP-growth; interestingness

1 Web 点击分析的研究现状

根据侧重点的不同,网络挖掘可分为三类[1]:1)Web 内容挖掘,是对 Web 页面内容进行挖掘,从 Web 信息中发现用户所需的特定化信息;2)Web 结构挖掘,其目标趋向于 Web 文档的链接结构,揭示出蕴含于文档结构中的个性化信息;3)Web 访问信息挖掘,是用挖掘 Web 服务器日志等数据获取的信息预测用户浏览行为的技术。

点击流分析法是 Web 访问信息挖掘的方法之一[2]。点击流分析以 Web 上的点击流数据为基础,利用 OLAP、数据挖掘等技术对数据进行分析来达到不同的目的。它通过分析采集到的用户在站点上运动情况,跟踪记录访问过的链结点,包括用户的来源地点、浏览站点的路线和最终到达的目标,链接分析包括对点击过的链接的观察、它们在屏幕上的相关位置、用户在网页上停留的时间以及点击过的链接间的关系和最终结果。通过对这些数据的有效分析,不但能够对网站的建设起到指导作用,而且也能够反映出企业在市场、销售、服务等各个方面的状况。点击流分析已经成为企业了解经营状况、了解客户行为的有效工具。

本文采用了一个提供了 3D 模型展示平台的 Web 服务日志,结合用户对 3D 模型的操作记录,进行整合,作为数据仓库的数据源。该展示平台将五金产品分类展示并提供 3D 模型演示功能。

2 基于扩展日志的客户操作模型

尽管 Web 服务日志文件包含很多有用的信息,但是这些日志文件很少包含一个完整地数据流分析所需要的全部数据。为 Web 站点和系统管理者提供用于 Web 站点管理的统计数据是 Web 服务日志设计原始目的,所以它不能作为点击流数据仓库的完整的数据源,但 Web 服务日志中的数据组成了点击流数据仓库的核心信息。

Web 服务器每次收到一个用户请求的时候,网站都会自动在日志文件记下一个事务项。有三种公开的标准日志文件格式:NCSA 的普通日志格式(Common Log Format, CLF),NCSA 的扩展日志格式(Extended Log Format, ECLF)和 W3C 的扩展日志文件(Extended Log File Format, ExLF)。本文采取对 ECLF 格式进行域扩展来获得客户点击行为的数据。ECLF 格式结构如下:

HostId rfc931 authuser【date offset】"method URL protocol" status bytes "Referrer" "Agent"

域含义如表 1 所示。

表 1 ECLF 域描述

域	含义
HostId	远程机器名或 IP 地址
Rfc931	用户登录名称
AuthUser	服务器授权用户名
Date	请求日期和时间
Offset	当地时间与格林威治时间偏移值
Method	请求方法(Get, Post, Head 等)
URL	请求页面完整地址
Protocol	客户端使用的 http 通信协议
Status	http 服务器处理结果状态
Bytes	传输字节数
Referrer	指向当前请求的链接 URL
Agent	客户使用的操作系统和浏览器

收稿日期:2005 - 06 - 17;**修订日期**:2005 - 08 - 30
基金项目:国家 863 计划资助项目(2003AA414023);广东省科技攻关资助项目(2004A10405001)
作者简介:谢凌(1981 -),女,江西赣州人,硕士研究生,主要研究方向:CIMS 及网络化制造; 陈新度(1967 -),男,湖南岳阳人,副教授,主要研究方向:CIMS 及网络化制造; 陈新(1960 -),男,湖南澧县人,教授,博士生导师,主要研究方向:CIMS 及网络化制造.

为了得到更多的用户操作记录,把标准日志文件 ECLF 域进行扩展,加入 OperateObject、Field、LoadTime、TimeTaken、Operation。域含义如表 2 所示,这些记录将为本文在下面的对客户的兴趣度的挖掘起到很大的作用。

表 2　ECLF 扩展域描述

域	含义
OperateObject	操作对象
Field	操作对象所属的域
LoadTime	加载时间
TimeTaken	在对象上停留时间
Operation	对对象进行的操作

操作对象可以是指一个网页,也可以具体到一个文件。进行的操作对于一些提供功能组件的网站来说,这部分内容尤为重要。对于一个 3D 模型,可进行的操作可以有放大、缩小、旋转和漫游。对一些组合件甚至提供拆分及重组功能。

在表 1,2 的日志文件中,将部分域抽取出来存储在一张表中,为下面的数据挖掘做准备,表结构如表 3 所示。利用表 3 的数据,将 TimeTaken 减去 LoadTime,就得到实际操作时间(ActualTime)。而进行数据挖掘真正用到的就是实际操作时间。通过用户对操作对象进行的操作以及在该对象上的实际操作时间来进行感兴趣度的分析,这里的感兴趣度是针对某一操作对象的;也可以通过分析操作对象所属的域,而对用户对整个域的感兴趣度做一个界定。

表 3　数据结构表

域	含义
HostId	远程机器名或 IP 地址
Rfc931	用户登录名称
OperateObject	操作对象
Field	操作对象所属的域
LoadTime	加载时间
TimeTaken	在对象上停留时间
Operation	对对象进行的操作

以上讨论了一个点击流数据仓库的主要数据源,在对源数据进行接收、分析、抽取、净化、汇总、交换、存储等之后,再从源数据库中分析抽取面向主题的集成数据,以该主题数据作为分析型应用的数据,就得到了数据仓库的数据。本文采用数据挖掘方法来分析客户行为,为企业进行市场预测,提供决策依据。

3　客户行为分析

在数据挖掘的模式中,关联规则模式是比较重要的一种。最基本的关联规则发现算法是 Apriori 算法,但使用 Apriori 算法可能会产生大量的频繁集,并且需要重复地扫描数据库,产生大量地频繁模式,而且这些模式中有些是用户不感兴趣的模式。FP-growth 的方法采用了分而治之的策略:在经过了第一次的扫描之后,把数据库中的频集压缩进一棵频繁模式树(FP-tree),同时依然保留其中的关联信息。随后再将 FP-tree 分化成一些条件库,对这些条件库分别进行挖掘,大大提高了挖掘效率[3],因此本文采用 FP-growth 方法。

在数据挖掘的开始,必须先对数据进行预处理,本文一个重要的工作就是将数据库中的数量属性按照一定的规则离散化,得到真正需要的数据之后,再利用 FP-growth 算法进行关联规则发现。

3.1　数据预处理

需要用到表 3 所列出的数据项:操作对象和操作对象所属的域,以及根据表 3 的数据得到的数据实际操作时间。文献[4]提出了时间窗口模型,认为有意义的事务具有与之相关的整体平均长度,可以简单地以一个预定的参数作为时间段来分割用户访问日志。但是对于实际操作时间是一些具体的时间,在数据库中称为数量属性,因此必须根据一定的规则将这些数值属性离散化,将数值属性的取值映射到一个个区间上,每个区间对应着一个离散的符号,如表 4 所示。

表 4　数据分组规则

时间段	特征	特征标识(Identify)
0 ~ 20s 或 > 600s	不感兴趣,或只是打开模型,没有看,忘记关闭	N
20 ~ 60s	一般感兴趣	G
60 ~ 120s	感兴趣	L
120 ~ 300s	很感兴趣	M
300 ~ 600s	非常感兴趣	H

得到离散区间后,将原始数据的实际操作时间用特征标识表示,如表 5 所示(由于数据太多,本文只列出部分有关数据)。操作对象列,取值有{1,2,3,4,5,6,7,8},操作对象所属的域列,取值有{a,b,c},其中 OperateObject{1,2,3}属于 Field{a},OperateObject{4,5}属于 Field{b},OperateObject{6,7,8}属于 Field{c}。

3.2　基于 FP-growth 算法的关联规则发现

表 5　部分日志数据

OperateObject	Field	Identify
1	a	L
7	c	H
3	a	N
6	c	H
8	c	G
5	b	M
4	b	N
2	a	H
6	c	M
7	c	M
…		

数据预处理之后,就可以利用 FP-growth 算法进行数据挖掘工作。FP-growth 算法过程可以总结为:①对数据库进行扫描,把数据库中的频繁集压缩到一棵频繁模式树(FP-tree)上;②对 FP-tree 进行数据挖掘,从中挖掘出关联规则。下面详细描述该算法的实现过程。

3.2.1　构造 FP-tree

FP-tree 的构造步骤如下[5]:(1)扫描一遍数据库得到项目的支持技术,根据最小支持度阈值筛选出频繁项目,并对项目按其频度降序排列,形成项头表;(2)对每条事务中的频繁项目按项目表中项目的次序组成频繁模式并插入到 FP-tree。

FP-growth 算法的第一部分就是构造一棵 FP-tree,并定下最小支持度,找出频繁集。本文设定最小支持度 $minsup$ = 10%,完成 FP-tree,如图 1 所示。

3.2.2　利用 FP-tree 进行规则挖掘及规则总结

在 FP-tree 上进行规则挖掘时,设最小支持度 $minsup$ = 3%,最小置信度 $Minconf$ = 40%,得到所有的频繁集并找出关联规则。进行数据挖掘必须明确挖掘的任务或目标是什么,本文的目的是根据这些数据得到操作对象的感兴趣度,或者是对对象所属域的感兴趣度。考虑产品复杂度影响到操作时间的区间离散化,因此在 3D 模型展示平台的 Web 服务日志中选取某洁具公司复杂度接近的三种产品系列(对应为操作对象所属的域 field):延时龙头、恒温龙头、感应龙头以及各系列中多种产品(对应为操作对象)进行分析。

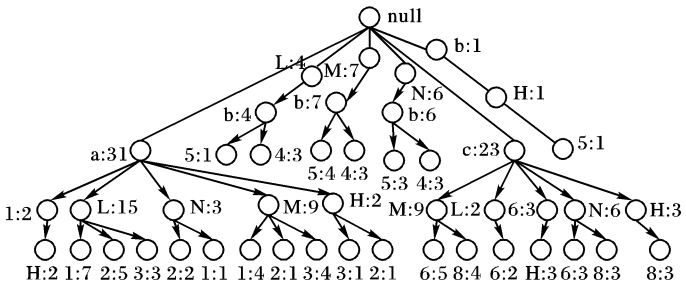

图 1 FP-tree

本文基于以下两种模式进行数据挖掘,找出有价值的关联规则,并按照两种模式进行分类,分类结果如下:

1)市场对某产品的反应

Confidence(1 => L) = 46.7%

Confidence(2 => L) = 41.7%

Confidence(3 => M) = 40%

Confidence (a => 1) = 46.4%

2)市场对某产品系列的反应

Confidence(L => a) = 68.1%

Confidence(H => c) = 69.2%

Confidence(b => N) = 42.9%

从这些关联规则中,总结出以下知识:

1)延时龙头产品系列中,产品 1 的查阅频率最高,产品 1 和产品 2 的感兴趣度水平是 L,产品 3 的感兴趣度水平为 M。

2)在感兴趣度水平为 L 的所有产品中,延时龙头产品占了很大部分;而在感兴趣度为水平为 H 的所有事务中,感应龙头产品以为主。

3)对恒温龙头产品系列,感兴趣度水平仅为 N。

4 结语

本文采用一个提供了 3D 模型展示平台的 Web 服务器日志,将其进行扩展作为数据仓库的数据源,并在此基础上用 FP-growth 方法进行关联分析的规则发现。此外,在数值属性离散化时,还可以进一步考虑操作对象的复杂度,增加权重的思想,使数据分组更加具有合理性。

参考文献:

[1] 张锋, 常会友. Web 使用挖掘系统研制中的主要问题和应对策略[J]. 计算机科学, 2003,30:(6).

[2] SWEIGER M, MADSEN MR, LANGSTON J, 等. 点击流数据仓库[M]. 北京:电子工业出版社, 2004.

[3] 冯志新, 钟诚. 基于 FP-tree 的最大频繁模式挖掘算法[J]. 计算机工程,2004,30(11).

[4] COOLEY R, MOBASHER B, SRIVASTAVA J. Grouping Web page references into transactions for mining world wide Web browsing pattern[A]. Knowledge and Data Engineering Workshop[C]. Newport Beach, CA; IEEE, 1997. 2 - 9.

[5] (德)巴斯蒂安. Data Warehousing and Data Mining[M]. 北京:冶金工业出版社, 2003.

(上接第 2939 页)

近,传统的 GPS 接收机的水平定位精度约为 12 米(95%)。

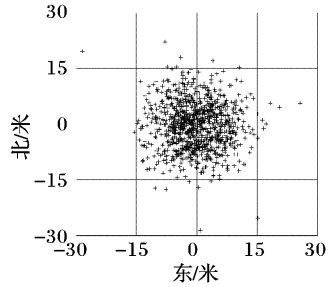

图 3 无遮蔽的操场上的现场测试结果

在 GPS 卫星信号受到遮蔽的街道边上的现场测试结果则如图 4 所示。在这种环境下,传统的 GPS 接收机大部分时间都只能捕获到两颗 GPS 卫星的信号,导致输出的定位结果长期没有变化。而本文的组件式软件 GPS 接收机平均可以检测到六颗 GPS 卫星的信号,定位成功率达到了约 99.5%,水平定位精度也达到了 30 米(95%)。

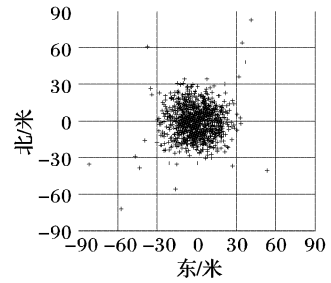

图 4 有遮蔽的街道边上的现场测试结果

从以上现场测试结果可以看出,本文的组件式软件 GPS 接收机与传统的GPS接收机相比,具有更好的稳定性和更高灵敏度;特别是在 GPS 卫星信号受到遮蔽使得传统的 GPS 接收机无法定位的情况下,本文的组件式软件 GPS 接收机仍然能够正常工作,对扩大 GPS 的应用领域有重要意义。

4 结语

本文设计了一个面向嵌入式系统、易于集成的组件式软件 GPS 接收机,通过自主定位方式、快照定位方式和辅助定位方式的有机结合,具有传统 GPS 接收机所没有的优秀的综合性能,即高稳定度、高灵敏度、低功耗和可观的定位精度。同时,它还具有寄主 CPU 时间占用率低的特点,且软硬件移植都很方便,因而是将定位功能集成到各种移动设备中的高效率低成本的方案。

参考文献:

[1] KAPLAN ED . Understanding GPS : Principles and Applications [M]. Boston: Artech House Publishers, 1996.

[2] JAMES BAO-YEN TSUI. Fundamentals of Global Positioning System Receivers: A Software Approach[M]. New York: Wiley Inter-Science, 2000.

[3] MOEGLEIN, KRASNER. An Introduction to Snaptrack Server-Aided GPS[A]. Proceedings of the Institute of Navigation conference, ION-GPS 1998[C]. 1998.

[4] FENG SJ, CHOI LOOK LAW. Assisted GPS and Its Impact on Navigation in Intelligent Transportation Systems[A]. Proceedings of the IEEE 5th International Conference on Intellitent Transportation Systems[C]. 2002. 926 - 931.

[5] DJUKNIC GM, RICHTON RE. Geolocation and Assisted GPS[J]. IEEE Computer, 2001,34(2):123 - 125.