Web 上的数据挖掘技术和工具设计

谢丹夏
北京大学计算机系　北京 100087)
E-mail:xdx@jbsim.cs.pku.edu.cn

摘　要　电子商务网站的网上业务量巨大,在每天的大量业务中蕴涵了许多用户的隐藏信息。每个顾客在 WEB 上的行为都会产生相关数据,这不只是包括购买的信息,而且还有利用搜索引擎和在站点内浏览的相关数据。所有的交互数据都可以被网站后台的数据库记录下来,这些大量的数据集合包含了对历史记录的市场分析以及数据驱动的市场预测非常有益的潜在信息。利用完善的数据库技术,现在的公司能够比较容易地搜集到大量的客户信息。而通过把数据挖掘在 WEB 上的应用,即 WEB MINING 技术,公司可以利用有效的顾客信息,发掘潜在的市场,提高竞争力。

关键词　WEB 挖掘　数据挖掘　人工智能

Web Mining Technology and Designing of the Tools

Xie Danxia
(Department of Computer,Beijing University,Beijing 100087)

Abstract: More and more commerce-related transactions are becoming digital.The more you know about your customers, the better you can serve them.Every customer action on a Web site generates data,not just high-level interactions such as buying something,but also something as simple as using a search engine or navigating through a site.All these interactions between digital service providers and the consumer can be recorded and stored in digital databases.These large data sets contain information helpful to business marketing strategies,both for retrospective analysis as well as data-driven forecasting.Web mining tools will provide companies with previously unknown statistics and useful insights into the behavior of their online customers via analyzing the data on the web.

Keywords: Web mining,Date mining,AI

1　引言

随着 Internet 的飞速发展，网上的数据资源空前丰富。但是数据资源中 蕴涵的知识却至今未能得到充分的挖掘和利用,"数据丰富而知识贫乏"的问题非常严重。在日益激烈的电子商务买方市场竞争中,任何与消费者行为有关的信息对商家来说都是非常宝贵的。

近年来兴起的数据挖掘 (Data Mining)技术为解决这个问题带来了一线曙光。而通过把数据挖掘在 WEB 上的应用,即 WEB MINING 技术,公司还可以分析和预测顾客的将来行为。通过 WEB MINING 技术,公司利用有效的顾客信息,可以大大降低运营的成本。

2　数据挖掘技术介绍

2.1　数据挖掘 (DataMining)的概念

数据挖掘 (DataMining)是近年来随着数据库和人工智能技术的发展而出现的全新信息技术,同时也是计算机科学与技术,尤其是计算机网络的发展和普遍使用所提出的、迫切需要解决的重要课题。数据挖掘可以描述为 :数据挖掘是指从数据中提取模式的过程。它反复使用多种数据挖掘算法从观测数据中确定模式或合理模型。DataMining 数据挖掘),是一种决策支持过程,它主要基于 AI、机器学习、统计学等技术,高度自动化地分析企业原有的数据,作出归纳性的推理,从中挖掘出潜在的模式,预测客户的行为,帮助企业的决策者调整市场策略,减少风险,作出正确的决策。数据挖掘工具是一种挖掘型的分析工具,重点在于预测。

数据挖掘有两个任务：

()机器的数据库理解 :将数据库变换为在表述上可为计算机理解的更为简洁的模型,然后利用这个模型求解新问题。

()数据库理解 :根据需求简化数据并将其翻译为自然的表示形式 如,数学公式,自然语言与图表等),发现隐含在大量数据中的规律并使之为人理解。数据挖掘可以从实例数据中直接导出规则,用于构造知识库 ;也可在数据库中对已有规则进行验证,因此对知识库的维护和更新也是有用的。

2.2　数据挖掘的主要方法

数据挖掘的技术基础是人工智能 (AI)。但它仅仅利用了人工智能中一些已经成熟的算法和技术,例如 :人工神经网络 (Artificial Neural Networks)、遗传算法 (Genetic Algorithms)、决策树 (Decision Trees)、 邻近搜索方法 (Nearest Neighbor Method)、规则推理 (Rule Induction)、模糊逻辑 (Fuzzy Logic)等,其问题的复杂度和难度比人工智能降低了许多。

数据挖掘系统利用的技术越多，得出的结果精确性就越高。这主要取决于问题的类型以及数据的类型和规模,无论采用哪几种技术来完成任务,从功能上可以将 DM 的分析方法划分为以下四种：

()基于关联度的分析 :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。

()基于序列分析 :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。

(3)分类分析:分类分析法的输入集是一组记录集合和几种标记,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。

(4)聚类分析:聚类分析法的输入集是一组未标定的记录,也就是说此时输入的记录还没有被进行任何分类。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。而所依据的这些规则是由聚类分析工具定义的。

在一个实际的 DM 系统中经常是综合地利用这四种方法的。

3 WEB 上的数据挖掘

3.1 WEB 上数据挖掘的用途

到一个站点的所有访问者将会留下浏览的踪迹,这些信息自动存贮在 WEB 服务器的日志文件内。WEB 分析工具通过分析和处理 WEB 服务器的日志文件来生成有意义的信息。例如有多少人访问了该页面,他们从哪儿来,那些页面最受欢迎等。当前经济模式的变化,从传统实体的商店到 INTERNET 上的电子交易,同时也改变了销售商和顾客的关系。现在网上顾客的流动性很大,他们关注的主要因素是商品的价值,而不象以前注意品牌和地理因素。因此,电子销售商一个主要的挑战,是需要了解到顾客尽可能多的爱好,价值取向,以保证在电子商务时代的竞争力。数据挖掘是用来发现不明显的,有潜在价值的数据。WEB 上数据挖掘的潜力在于应用存在的和最新的数据挖掘算法,分析 INTERNET 服务器上的日志以及顾客,销售和产品的外部数据。

综合来说,WEB 数据挖掘有以下三个方面的益处:

理解顾客行为:

(1)通过理解访问者的动态行为来优化电子商务网站的经营模式。

(2)电子销售商可以获知访问者的个人爱好。

(3)决定网站上访问者到购买者的转化率。

(4)决定顾客的回头率(顾客第二次购买同一品牌的概率)。

(5)发现顾客的购买模式和访问者的浏览模式。

(6)发现什么样的顾客群在网站上购买什么商品。

(7)发现电子商务网站上顾客之间的联系。

判断 WEB 站点的效率:

(1)发现站点上的高购买率部分和低购买率部分。

(2)WEB 设计者不在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。

(3)电子销售商可以针对不同顾客提供个性化的服务。

评估电子商务模式的成功与否:

(1)容易将用户按照模式分类。

(2)容易评测广告的投资回报率。

(3)容易得到可靠的市场反馈信息。

3.2 WEB 数据挖掘的分类

WEB 数据挖掘可以分为:

(1)WEB 内容的挖掘

WEB 内容的挖掘是挖掘 INTERNET 的页面和后台交易数据库。

(2)web 结构的挖掘

web 结构的挖掘是运用数据挖掘技术来重建 WEB 站点的结构。

(3)WEB 使用的挖掘

WEB 使用的挖掘是通过挖掘相应站点的日志文件和相关数据来发现该站点上的浏览者和顾客的行为模式。

4 WEB 上的数据挖掘的实现和工具

4.1 WEB 数据挖掘工具的架构和工作步骤

web 数据挖掘器将从 WEB 数据库中提取并集成数据,它需要 WEB 站点的后台数据库支持(包括用户访问日志文件,注册用户的活动信息)以及 WEB 数据仓库(主要是面向电子商务网站的注册用户)。并且要解决数据语义的二义性问题,以及消除脏数据等等,这需要一个过滤器和综合器来完成。由于 HTTP 协议是无状态的,所以无法区分和跟踪一个访问者在网站上的所有行为,这样单纯的依靠日志文件来进行分析所得到的用户信息微乎其微。要吸引访问访问者成为注册用户,以便得到更多的用户信息,并且通过注册给用户加上 COOKIE 访问头,作为标识用户的唯一 ID。

使用各种数据挖掘方法分析数据库中的数据,为了更方便的加入和替换挖掘方法,把方法做成调用库的形式,可以使用选项来选择挖掘方法。

可以得到以下的信息:

(1)用户的生活模式,爱好,购买频率,所属的用户群。

(2)不同用户群的共同特征。

(3)页面的访问情况。

(4)广告的点击情况。

对挖掘出的规律和模式进行评价和验证,可以通过可视化的工具来进行评价。也可以通过相应的 WEB 数据仓库工具去验证得出的结论。

进而可以对证实的结论和模式进行应用,主要在以下几个方面。

(1)信息反馈和广告发送。

(2)网站设计的相应修正。

(3)对用户定制个性化的页面。

(4)对广告设置的修改。

4.2 WEB 数据挖掘工具的系统框图

如图 1 所示:

4.3 WEB 数据挖掘工具的设计

该工具是辅助电子商务网站开发的分析工具。它运行在电子商务网站的用户数据库和数据仓库之上。包括以下功能模块:

(1)过滤器

用来从 WEB SERVER 数据库中抽取相关数据,进行二义性分析,消除不一致性。

(2)挖掘综合器

是一个挖掘驱动引擎。

它根据挖掘要求,或者根据挖掘方法的知识库到 Web 数据挖掘算法库中去选择合适的挖掘方法,并且使用该方法去执行挖掘任务。

(3)Web 数据挖掘算法库

是一个数据挖掘分析方法的算法库。它以插板的方式来组织各种挖掘算法,使各种方法可以方便的插入,实现了可扩展性和易选择性。

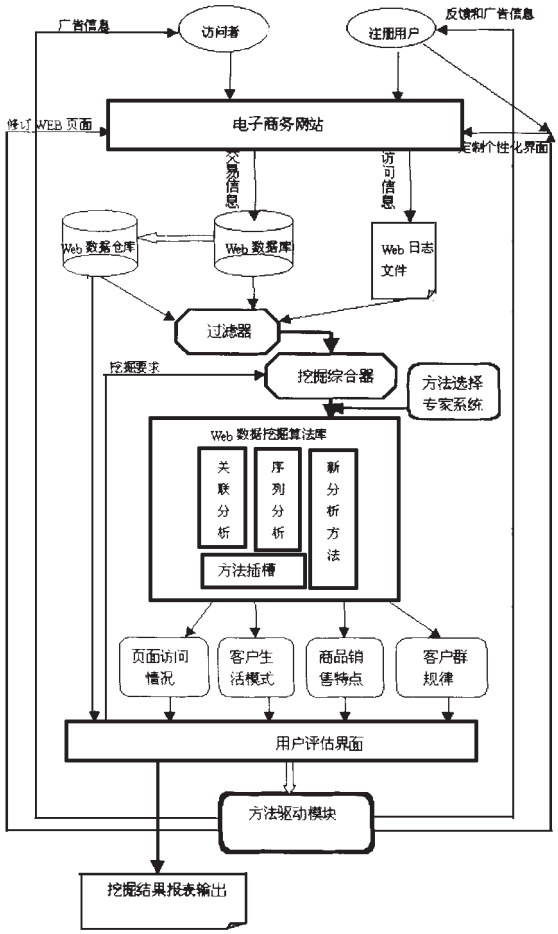

图 1

(4)用户评估界面

以一种直观的方式来表现数据挖掘的结果，提供一个和分析人员交互的友好界面。如果本次的挖掘结果不能满足分析人员的需要或者还有进一步的猜想，就可以从这里输入挖掘需求。

(5)方法驱动模块

它利用挖掘出来的信息，去进行相应的工作。

其中页面访问情况用来指导网站页面的重新设计和修改。

分析出的客户生活模式可以作为反馈信息，以电子邮件的形式把相应的商品广告等发送给客户。

根据客户的爱好等来定制针对他的个性化 WEB 界面。把他所喜欢的栏目排放在前面，把客户所常用的商品尽量前排。

5 结论

WEB 上的电子商务是交互式的，它的发展方向是在这里顾客可以定制和指定产品和服务，交换信息。WEB MINING 工具可以用来对 WEB 上的商业模式建模，预测，了解影响销售的各种因素，以便迅速调整他们的市场，价格，存货等。WEB MINING 工具还可以发现顾客和访问者的爱好，生活模式等，并且可以充分利用这些信息来发展新的客户，发掘新的商业机会。

在网络社会中，顾客成为真正的“上帝”，网上销售商必须很好地考虑到顾客的需要和利益。在竞争如此强烈的网络经济中，作为电子商务成功的一个重要因素，WEB MINING 将成为一个关键技术。(收稿日期：2000 年 9 月)

参考文献

1.Jesus Mena.Data Mining Your Website.Digital Press，1999.7 ISBN 1-55558-222

2.Wayne W Eckerson.Marrying E-Commerce and Customer Intelli gence

3.Usama M Fayyad.Advances in knowledge discovery and data mining.1996

4.W H Inmon,John A Zachman,Jonathan G Geiger.Data stores,data warehousing,and the Zachman Framework:managing enterprise knowledge.1999

5.Alex Berson,Stephen J Smith.Data warehousing,data mining,and OLAP.1997

(上接 56 页)

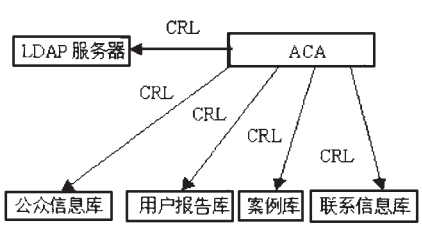

图 4 证书的取消模型

3.3 属性证书的发放、储存和取消

系统管理员设计好属性证书后，证书的签发由 CCERT 的属性证书权威(Attribute Certificate Authority,ACA)来执行。用户向 ACA 申请证书，证书存储于用户本地，备份在 LDAP 服务器中，用户的证书丢失后，可向 LDAP 服务器索要证书的备份。授权签发的证书由授权签发人取消，授权签发的证书的有效期一般很短，所以大都不是被取消而是自动失效。其余的属性证书由系统管理员委托 CCERT 的证书权威取消，证书取消列表(Certificate Revoke List,CRL)发往 LDAP 服务器和相关的资源服务器。证书的取消方式，如图 4。

4 结论

文章采用属性证书，给出了 CCERT 信息服务的访问控制模型及其实现。这个模型既便于集中管理，又具有很好的灵活性，并用公钥算法和随机数技术保证了系统的安全性。这种模型不但适用于 CCERT 分布式信息服务的访问控制，也适用于其他的分布式系统的访问控制。(收稿日期：2000 年 2 月)

参考文献

1.Sandhu R S,Samarati P.Access control:Principles and Practice.IEEE Communications Magazine,1994.9,32(9):40-48

2.James Davis,Dong Jacobson,Stephanie Bridges,et al.An Implementation of MLS on a network workstations using X.500/509.IEEE International Performance,Computing & Communications Conference,Proceedings,Feb 5-7,1997

3.Dave Kosiur.LDAP:The next-generation directory?Sunworld,1996.10,http://www.sunworld.com/swol-10-1996/swol-10-ldap.html

4.R Housley SPYRUS,W Ford VeriSign,W Polk NIST,et al.1999.1,http://www.ietf.org/rfc2459.txt

5.ITU-T recommendation on X.509.1994.2,http://sirius.ac.upc.es/~jbt/learning/x509/ITUx509/97x509final.html