电子商务网站的 Web 数据挖掘方案设计

程 苗
（四川大学工商管理学院 成都 610215）

摘 要 没有有效的数据管理和分析工具，Web 上日益增长的海量数据将变成“数据坟墓”。本文运用数据挖掘技术，从 Web 数据库中提取所感兴趣的信息，从不同角度分析它们，从而有效地利用数据库中的大量数据，将“数据坟墓”转换成“知识金块”。Web 数据挖掘的关键在于如何收集有意义的原始数据，本文将重点阐述如何进行 Web 数据挖掘过程中的数据准备工作。

关键词 数据挖掘，Web 网站，智能查询

The Design of Web Data Mining Based on E-business Website

CHENG Miao
(Management Business Administration Department of Sichuan University, Chengdu 610215)

Abstract It is the purpose of this paper to use the method of data mining and design a scheme of Web data mining which is based on E-business website. The key of Web data mining is to collect more useful data, so this paper is concerned with the question of how to collect and pre-process data.

Keywords Data mining, E-business, Website, Intelligent query

1 引言

随着互联网技术的发展，许多企业都建立了自己的电子商务网站。网上业务的竞争比传统业务的竞争更加激烈，因为顾客只需单击几次便可跳转到竞争对手的网站。网站的内容、结构及服务等任何一处都有可能成为吸引或失去顾客的因素。要想在竞争中生存、获胜，就要比竞争对手更了解顾客。网上顾客决策的程序大致有五个阶段，如图 1 所示。

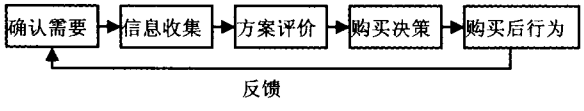


图 1

了解网上顾客的需求和购买过程是制定有效的电子商务战略的基础。通过了解网上顾客如何经历确认需求、收集信息、评价解决方案、购买后行为阶段，企业就能获得怎样满足网上顾客需要的许多线索。换句话说，就是要了解每个阶段网上顾客的行为及其影响因素，这就可以使电子商务人员为目标市场制定比实际切实有效的电子商务方案。

同时，我们还可以对顾客查找留下的那些日志文件进行数据挖掘，筛选关于顾客的信息，通过对顾客的查找行为、频率、内容等的分析，得到关于群体顾客行为和方式的普遍信息，用以改进 Web 页面的设计。通过 Web 数据挖掘，及时了解企业的整体运营情况，针对问题快速做出反映；通过收集各种最新市场信息，并及时反馈给高层决策者和研究开发等有关部门，进行高效、准确的市场决策。通过对销售记录、顾客信息的挖掘与分析，掌握最新的信息以获得更多的市场，甚至可以根据顾客的查找兴趣、查找频率、查找时间动态地调整页面结构，改进服务，给客户个性化的界面，开发有锁定性的电子商务，以更好地满足查找者的需要。

2 Web 挖掘定义

Web 数据挖掘就是指根据跟踪顾客在 Web 上的浏览行为（即 Web 服务器日志文件），对其进行模式分析，以筛选出关于查找兴趣、查找频率等相关知识，从而改进 Web 页面的结构和内容，改进服务，提供个性化界面，从而刺激顾客的购买欲。本文是站在企业的角度，为其设计一个可以改善商业效果的方案，因此将 Web 数据挖掘定义为通过采用数据挖掘技术，从 Web 数据库以及 Web 服务器日志文件中，抽取感兴趣的、有用的模式和隐含信息的过程，其结果可以为企业的决策提供参考。其中，“过程”一词非常重要，它不单单是获得解决方案，还包含了数据的收集、预处理、分析、模型的形成以及对模型的评估，且这个过程是不断反复的。它不光是关联规则、聚类分析、人工神经网络等方法的随意应用，而是一个经过精心策划、深思熟虑的，决定什么是最有前景的一个过程。

3 Web 数据挖掘存在的问题

1）Web 上的信息每天都在不断地更新、增加，各个数据源的结构和信息也各不相同，这就使得在 Web 这个巨大复杂的异构数据库环境中获取所需内容，变得非常困难。因此数据准备就成了 Web 数据挖掘技术的关键，甚至可以将整个数据准备过程独立描述为一个数据挖掘方法。

2）数据挖掘的步骤是：问题定义→数据收集→数据预处理→数据挖掘算法执行→结果解释和评估。然而，为了提出一个有意义的问题定义，拥有领域内详尽的知识和经验是必不可少的，不是每个企业决策者都有敏锐的洞察力，能够在如此庞大的数据中准确地发现问题。因此，用上述数据挖掘过程来发现 Web 上有效的资源信息，就有了一定的局限性。

4 模型的建立

针对以上问题，将数据挖掘过程修改为：

数据准备→智能查询→数据挖掘算法执行→结果解释

根据修改的数据挖掘过程，建立以下模型，如图 2。其中数据准备阶段包括数据从 Web 服务器向源数据库的转换，再经清理、变换、集成、装入到数据仓库的建成。智能查询阶段包括用户的查询工具和企业的分析工具。数据挖掘算法执行采用的关联分析和时序分析。而结果解释则为页面维护、更新，制定生产、销售计划，用户个性化界面，新产品的研制和开发等。

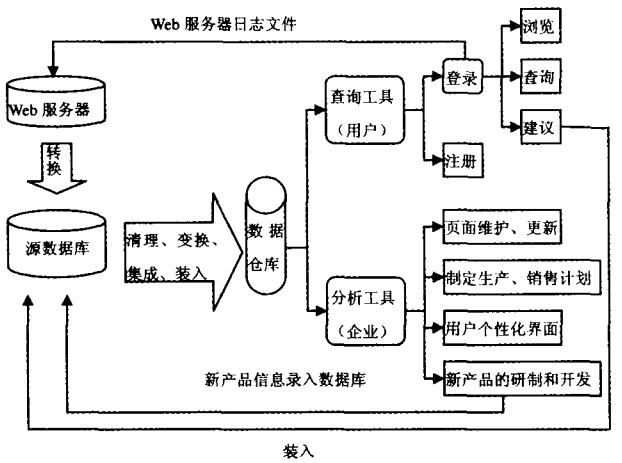


图 2

该模型分为两部分，一是供前台顾客查询的浏览器界面，一是供后台决策者管理的界面。在前台，顾客可以登录企业网站浏览产品相关信息的详细介绍，还可以对产品或服务提出建议而获得个性化功能。在后台，对大量的产品交易信息、顾客注册信息、顾客浏览行为模式、顾客建议等数据进行清理、交换、集成，装入到数据仓库中，通过相应的 Web 数据挖掘技术（例如，关联规则、人工神经网络等），有效地增加市场透明度，为企业提供新的经营生产方式，以增加顾客的数量，提高企业收入。

5 对模型的具体阐述及实现方法

5.1 数据准备

数据准备的中心任务是把数据组织成一种标准形式，使其能被数据挖掘工具和其他基于计算机的工具处理（标准形式即一张关系表）。它包括数据收集和数据预处理两个过程。

(1)数据收集

数据来源主要有两个：一个来自 Web 服务器的日志文件。一个典型的 Web 服务器日志文件包含描述用户浏览网页和沿着连接所产生的点击流序列信息，相关的字段依次有：主机地址、日期/时间，请求，状态，字节，访问页和浏览器类型。这一数据准备过程的工作是从 Web 服务器日志中提取数据，创建适于数据挖掘的文件。另一个来自企业网站数据库的顾客在线注册信息及交易信息，它将直接以一条记录的形式添加进源数据库的相关表中。

(2)数据预处理（数据仓库的创建）

源数据库中保存着大量的原始数据，在我们进行数据挖掘时，并不是每一个数据都是有用的。如：有些感兴趣的属性缺少属性值（有的客户不愿透露自己的收入等）。而更多的时候，源数据库中储存了许多重复的数据。因此，在数据挖掘之前，一定要先对源数据库中的数据进行预处理，经过清理、变换、集成后装入到数据仓库中。数据仓库能够将各种应用系统集成在一起，为统一的历史数据分析提供坚实的平台，对信息处理提供支持。

将数据从源数据库向数据仓库的移动称为数据的预处理。它可以分为几个步骤：提取、净化、集成和汇总。

ⅰ)提取。在这一步中应明确以下几个问题：在源数据库中要访问哪些文件或表？从源数据库中可以提取哪些字段？哪些记录是可用的？哪些记录是可以删除的？

我们进行数据挖掘无外乎是想了解顾客的浏览模式、购买模式。因此，数据提取即是从源数据库中提取相关的数据表：供用户查询的产品表 product；保存客户信息的表 customer；保存客户定单的表 order；由 Web 日志文件转换的相关数据表。

ⅱ)净化。数据净化主要是分析提取的表中是否有异常数据，包括缺省值、异常点、噪声等。这些数据的存在会扰乱分析过程，使发现的模式精确性很差。对这些数据的处理可以采用平均或者回归分析、人工神经网络等方法来平滑掉异常数据。但考虑到人工神经网络具有对噪声数据的高度承受能力，以及它对未经训练的数据分类模式的能力等优点，本文采用人工神经网络的方法来发现和分析异常数据。人工神经网络中神经元个数众多以及整个网络存储信息容量的巨大，使得它具有很强的不确定性信息处理能力。即使输入的信息不完全、不准确或模式不清，神经网络仍然能够通过联想思维，展示存在记忆中事物的完整图像。只要输入的模式接近于训练样本，系统就能给出正确的推理结论。

图 3 是一个简单的二层前传人工网络，其中所有神经元状态均取[0,1]中的连续值。图中最下面的一层称为输入层，用以输入已知测量值。中间一层称为处理层或隐单元层，其个数的选取方法可通过试验来决定。最上面一层称为输出层。

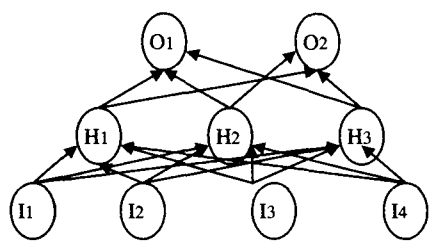


图 3

令 u 表示一个确定的已知样点标号。当将第 u 个样品的原始数据输入网络时，相应的输出单元状态记为 O_i^u，隐藏单元状态为 H_j^u，输入单元取值 I_k^u。i,j,k 依次对应于输出层、中间层及输入层，所以 W_{ij}、W_{jk} 分别表示从中间层到输出层的权和从输入层到中间层的权。如果 W_{ij}、W_{jk} 均已给定，那么对应于任何一组确定的输入 I_k^u，网络中所有单元的取值不难确定。即：对隐单元 j 的输入是 $h_j^u=\sum_j W_{ij} I_k^{-u}$ 相应的输出状态 $H_j^u=1/(1+e^{h_j^u})$。该函数称为挤压函数，因为它将一个较大的输入值域映射到较小的区间 0 到 1。隐单元层 i 所接受到的迭加信号是 $h_i^u=\sum_j W_{ij} H_j^{-u}=\sum_j W_{ij} * 1/(1+e^{-h_j^u})$，网络的最终输出是：$O_i^t=1/(1+e^{-h_i^u})=1/(1+\sum_j W_{ij} * 1/(1+e^{-h_j^u})$。

到此，该问题转换为如何找到一组适当的权值 W_{ij}、W_{jk} 这个问题可以由 BP(Back Propagation，向后传播)方法解决。向后传播(BP)把学习过程分成两个阶段：首先将输入信息从输入层经隐藏层逐层处理，并计算各单元的实际输出值（正向

传播过程)；然后，若在输出层不能得到期望的输出，则逐层计算实际输出与要求输出之差(即误差)，以便据此差值调整权值(反向过程)。通过不断地迭代处理训练样本，修改权 W_{ij}、W_{jk}，使每个样本的网络预测与实际差之间的均方差最小，从而确定最优 W_{ij}、W_{jk}。根据这一模型，可以发现数据库中异常值和估计缺省值。

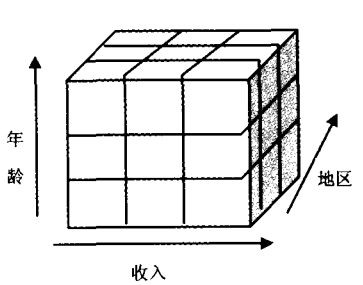
ⅲ)数据集成。数据经过净化后，其异常情况大大减少了。但这时的数据都是"杂而乱"的游离在数据库中的。数据集成就是要将这些"杂而乱"的数据结合起来存放在一个一致的数据仓库中。数据集成涉及的问题有：

(1)实体识别问题。例如计算机如何确定一个数据库的 customer_id 和另一个数据库中的 customer_num 指的是同一实体？通常，数据仓库的元数据(meta data)——关于数据的数据，可以帮助避免模式集成中的错误。元数据库中记录的内容有：数据转换、数据字段的有效日期和范围。记录的源系统或系统、源系统文件和字段映射、提取历史、安全性。

(2)冗余问题。该问题可被相关分析检测到。例如，给定两个属性，根据可用的数据，由公式 $r_{AB}=(n\sum A_iB_i-\sum A_i\sum B_i)/\sqrt{(n\sum A_i^2-\sum A_i\sum A_i)(n\sum BA_i^2-\sum B_i\sum B_i)}$ 计算出属性 A 和 B 的相关性。其中 n 是元组个数。若 $0<r_{AB}<1$，表明 A 与 B 之间存在不完全相关关系，且该值越大，一个属性蕴涵另一个的可能性越大。若 $-1<r_{AB}<0$，表明 A 与 B 之间存在不完全相关关系，一个属性会阻止另一个出现。若 $r_{AB}=1$，表明 A 与 B 之间存在完全正相关关系。若 $r_{AB}=-1$，表明 A 与 B 之间存在完全负相关关系。若 $r_{AB}=0$，表明 A 与 B 是独立的。利用这个公式便可以检测到 customer_id 和 customer_num 的相关性。

(3)数据值冲突的检测和处理。例如：不同国家价格的计量单位不同。该问题可以由一个统一的规范标准来解决。例如，遍历 product 表中的属性 price，发现其单位不为 RMB，则由相关公式转换。

ⅳ)数据汇总。将进行了提取、净化、集成后的数据加载到数据仓库进行汇总。这里的汇总也可理解为对数据做聚类分析。聚类与分类是不同的，聚类要划分的类是未知的。将该方法用于 Web 数据挖掘的汇总阶段，能够分析网上顾客浏览模式和购买模式，刻画出不同顾客群的特征，从而生成许多不同的簇(该簇是一组数据对象的集合，这些对象与同一簇中的对象彼此相似，与其它簇中的对象相异)。通过观察每簇的特点，集中对特定的某些簇做进一步分析，使可以挖掘大量有价值的信息，做出正确的决策。例如，进入数据仓库后的数据可以汇总成下列的数据立方体，其中每一个小立方体即为一个"簇"。



5.2 智能查询

数据挖掘即从上述经过收集、净化、集成、汇总而形成的数据仓库中有效地发现有价值但不明显的信息。通过数据挖掘调查，企业可以预测以下几个问题：哪些类型的顾客明年将成为购买最多的顾客？何种附加产品可以最大限度的刺激顾客的购买？明年我们将开发何种新产品以拓展市场？明年我们将增大哪个地区的产品投入？然而在大多情况下，企业可能并不精确地知道要挖掘什么知识或数据库有什么限制，因此不能给出精确的查询。智能查询，通过结合数据挖掘技术，帮助分析企业的目的，用智能的方式很好地返回给企业用于决策的信息。

(1)通过关联分析挖掘数据仓库中的关联规则。关联规则也就是数据集中项之间的有趣联系。例如，通过关联分析发现啤酒与尿不湿之间的关联规则是购买啤酒的顾客同时也会购买尿不湿。这就可以将这两种相关商品摆在一起组合销售或"购买 200 元尿不湿送两罐啤酒"来促进销售。

(2)通过时序分析和序列模式的挖掘来预测企业产品未来的销售情况，以及发掘潜在客户，促进产品销售。例如，根据前三个月企业产品的销售情况来预测下个月的销售情况。当客户在线购买一台个人电脑时，系统会根据挖掘出来的序列模式"购买这种电脑的人在一个月以后很可能再来购买一台打印机"而建议他同时购买一台打印机。

(3)通过数据预处理时得到的数据立方体来回答查询。借助 OLAP 对数据立方体进行切片、切块、下钻、上卷及旋转等操作，能方便地对任何一部分数据或不同抽象级别的数据进行挖掘，提供给企业强大的统计、分析、趋势预测能力。

总结 Internet 技术的发展和网络的普及扩展了数据挖掘的应用的范围，将 Web 技术与数据挖掘结合，已成为数据挖掘发展的一个新方向。如何充分了解顾客的喜好和购买模式，甚至是顾客的一时冲动，从而设计出满足于不同顾客群体需要的个性化网站，是企业在竞争中生存并获胜的关键。本文首先指出 Web 数据挖掘存在的问题，然后针对这些问题修改了数据挖掘过程，提出了基于 Web 数据挖掘的模型，并重点阐述了其中的数据准备阶段。最后还提出了智能查询的三种方法。

总之，Web 数据挖掘的出现让企业掌握顾客背景知识成为可能，也为其带来新的曙光。

参 考 文 献

1 (加)Han Jiawei, Kamber M 著. 数据挖掘：概念与技术
2 (美)Kantardzic M 著. 数据挖掘：概念、模型、方法和算法
3 薛惠锋，等编著. 智能数据挖掘技术
4 (美)Roiger R j, Geatz M W 著. 数据挖掘教程
5 邵峰晶，于忠清编著. 数据挖掘原理与算法
6 彭木根编著. 数据仓库技术与实施
7 (美)Sperley E 著. 企业数据仓库：规划、建立与实现
8 Mallach E G 著. 决策支持与数据仓库系统