网上顾客行为分析系统建设

山东经济学院　李洪磊

信息时代，企业实现上网经营成为大势所趋，这种新型经营方式不但给企业带来可观的经济效益，同时也为企业创造了一种宝贵的信息资源——顾客行为信息。本文从决策支持的角度出发，提出了一种基于数据仓库技术，能够实现网上顾客行为数据采集与分析的系统的建设方案，以飨读者。

引　言

随着Internet的普及和电子商务的兴起，许多企业纷纷建立了网站，不同程度地实现了网上经营。人们在为这一新的企业经营模式所带来的可观经济效益所吸引的同时却忽视了它所带来的另一宝贵资源——顾客行为信息（Customer Behavior Information）。

众所周知，顾客行为信息是市场营销决策的重要因素，在买方市场环境中，掌握顾客行为信息就如同掌握了在市场竞争中获胜的法宝。以往为了获取如此重要的信息，企业往往要进行市场调查，如问卷调查等，这些调查不仅费时费力而且带有浓厚的主观彩色，因而往往最终得出的结论无法反应实际情况，导致经营决策出现失误。在此我们设想一下，如果采用某种措施，将顾客的行为完整地记录下来，然后对这些行为数据进行分析，就可以比较客观地反映实际情况，为企业经营决策提供高质量的信息。但这在传统的企业经营模式下是很难做到的，因为我们不可能做到实时监视每个顾客的活动，而在以信息技术构建的网络经营模式下就可以实现这一设想，因为在该模式下顾客是在由页面构建的“虚拟市场”中活动的，他们的每个活动都是对页面对象的“点击（Click）”操作，如果我们将Click操作记录下来，于是一系列Click操作就是对顾客行为的真实写照。

顾客行为—CLICK的信息结构概述

从业务流程的角度分析，顾客的每一项活动都是由一系列具体操作组成的有序集合，在系统中表现为一组“点击（Click）”操作序列，在此称之为“点击流（ClickStream）”。为了准确反应顾客行为，系统需要记录如下与Click有关的信息：

1. 施动者，即该动作的发起顾客

2. 操作对象：即页面中被操作的对象，这些对象都是“活动对象”。页面由许多页面对象组成，这些对象可分为两类，一类为能够引发具体操作的“活动对象”，如“超链接”、“按钮”等，另一类是起说明或装饰作用的“静态对象”，如文字、图像、声音等。由于顾客操作都是针对“活动对象”进行的，所以在此不考虑“静态对象”。

3. 操作页面：操作对象所在页面。

4. 目标页面：操作对应的页面（也可以理解为操作所产生的页面）

5. 操作发生时间：包括顾客本地时间和系统时间，因为因特网是全球网，当顾客的客户机与网络主机在不同时区时机器时间会不同。

顾客行为数据的存储与采集

上述行为数据的成功采集与存储是实现顾客行为分析的先决条件，要完成对上述顾客行为信息的采集与存储，依靠传统的数据库技术是不可行的，因为对顾客行为数据的处理是一种分析型的处理，而传统的数据库技术是面向业务处理的，在分析型处理面前显得力不从心，为此我们采用一种新型数据处理技术——数据仓库技术。

1. 数据仓库技术简介

数据仓库技术是近几年应数据分析处理需求而发展起来的一项新技术。根据数据仓库之父W.H Inmon的定义，“数据仓库是支持企业经营决策的，面向主题的，集

成的，不可更新的，随时间不断变化的数据集合”。面向主题的数据组织方式是数据仓库的主要特征，是与传统的面向应用的数据组织方式的根本不同。为了直观地说明“面向主题的数据组织方式”在此仅举一例说明：例如，企业经营过程中，与“商品”有关的数据有“商品固有属性（种类、型号、产地、等级、品质等）数据”、“进货方面的数据”、“销售方面的数据”和“库存方面的数据” 等，这些数据在面向应用的数据组织方式下被分散到不同的应用系统中：“进货管理系统”、“销售管理系统”、“库存管理系统”等。而采用面向“商品”主题的数据组织方式则上述数据都被集中放置在一起，形成对商品较为完整的说明，以便为与商品有关的各类分析处理提供丰富的数据。显然“顾客”也是企业数据仓库中的重要主题，其中包含着企业经营决策所需的顾客信息。

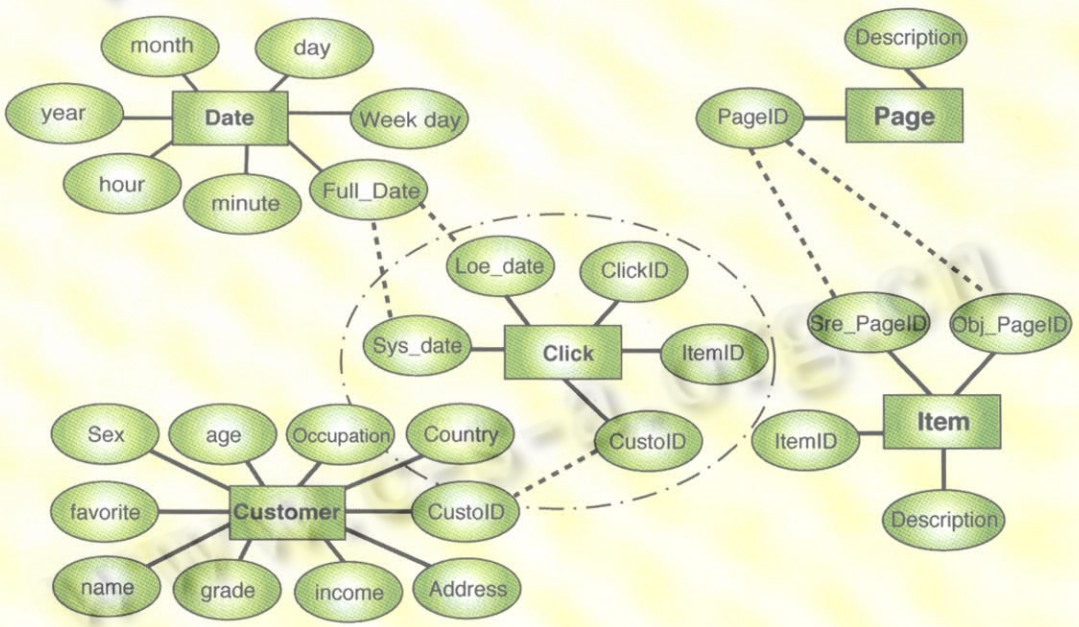
数据仓库技术不但提供了对数据的采集与存储机制，而且还提供了专用的分析方法——联机分析处理（OLAP）和数据挖掘（Data Mining）。联机分析处理用于从多个角度（维）对数据进行分析，从而尽可能体现出数据中所隐含的各种信息；数据挖掘则采用人工智能、统计分析等技术从数据中挖掘规律性的信息（称之为知识）。

2. 数据存储部件——顾客行为数据仓库的设计

由于顾客行为数据量极大，因此需单独建立一个数据仓库——顾客行为数据仓库对它们进行管理。以下是该数据仓库的建立过程：

（1）概念模型设计。以下的E-R数据模型描述了顾客行为信息的概念模型。

（2）行为数据的逻辑模型设计。由于目前数据仓库主要采用关系型数据库技术建立，所以针对上述概念模型



模型说明

① 方框表示实体，椭圆表示属性，虚线表示联系，实线表示属性与实体的所属关系

② Click 实体为模型的核心，其余四个实体是从四个方面（维）对该实体的补充，以支持对顾客行为的多维分析

③ Customer 实体是对 CustID 的描述，其中包含了顾客各方面的信息

④ Date 实体是对时间的描述，它将一个完整的时间分割为 6 个部分，分别为年、月、日、周日、时、分，这种细化可满足针对不同时间划分的分析

⑤ Item 实体是对页面活动对象的描述。ItemID 是对象唯一标识，Description 描述了其职能，如“新客户注册”、“购买产品”、“查阅产品目录”等

⑥ Page 实体是对系统页面的描述。PageID 是页面的唯一标识，Description 是对页面的描述，如“产品说明页面”、“销售页面”、“注册页面”等。

可得出如下逻辑模型（由二维表构成）

Click_table（行为记录表）/*表格式样请参照打印稿或随盘WORD文档*/

Click_table（行为记录表）

属性（或字段）	数据类型	说明
ClickID	Long	行为记录标识
CustID	Varchar(50)	顾客标识
ItemID	Varchar(50)	操作对象标识
Sys_Date	Datetime	系统时间
Loc_Date	Datetime	顾客本地时间

Cust_table（顾客信息表）

属性（或字段）	数据类型	说明
CustID	Varchar(50)	顾客标识
Name	Varchar(50)	姓名
Sex	Char(1)	性别
Birthday	Date	出生日期
Grade	Varchar(5)	文化水平
Country	Varchar(10)	国家
Address	Varchar(40)	住址
Occupation	Varchar(20)	职业
Income	Varchar(20)	收入
Favorite	Varchar(50)	爱好

Date_table（时间因素表）

属性（或字段）	数据类型	说明
Full_Date	Datetime	时间
Year	Numeric(4)	年
Month	Numeric(2)	月
Day	Numeric(2)	日
Week_day	Numeric(1)	周日
Hour	Numeric(2)	时
Minute	Numeric(2)	分

Item_table（操作对象字典）

属性（或字段）	数据类型	说明
ItemID	Varchar(40)	操作对象标识
Src_PageID	Varchar(40)	操作页面标识
Obj_PageID	Varchar(40)	目标页面标识
Description	Varchar(40)	Numeric(2)

Page_table（页面字典）

属性（或字段）	数据类型	说明
PageID	Varchar(40)	页面标识
Description	Varchar(40)	页面说明

完成上述逻辑设计后即可建立顾客行为数据仓库，即建立一个含有以上五个二维表的数据库。（因篇幅所限，有关“粒度”、“分割”、“存储策略”等方面的设计在此不作描述）

3. 顾客行为数据的采集

实现对顾客行为（Click）数据的采集是成功进行顾客行为分析的关键。在对Click数据进行采集之前需明确它的来源，通过对Click数据的分析，它有两个来源：客户端和服务器端。客户端的数据源包含：顾客标识和本地时间；服务器端的数据源包含：页面信息、操作对象信息、系统时间。把顾客标识划分在客户端是因为虽然顾客信息存储在服务器端的顾客行为数据仓库中，但究竟是哪个顾客在活动需要从客户端确定。明确了数据来源，下一步工作就是如何实现数据采集。笔者设想了三种实现模式：

(1)修改WWW服务器程序使其具备记录顾客行为数据的功能。

(2)增设与WWW服务内核进行通信的外挂模块，监视WWW服务器的操作，实现对顾客行为数据的记录。

(3)在每个页面中添加采集顾客行为数据的小程序，此程序可用JavaScript或ASPScript编写。

笔者建议使用方式(3)，因为前两种方法涉及与服务器内核的通信或改动，既复杂又破坏了服务器功能的独立性与完整性，而方式(3)仅对页面进行改进，从编程到实现均不复杂。方式(3)的工作原理：

①在主页中加入登录、注册控制，使老顾客能进行登录而新顾客可以进行注册。注册时要求顾客输入一些个人信息，其中必需的信息有姓名、住址等，因为这涉及购买商品或服务的配送等问题。

②每个页面中的操作对象被激发时均要传递两个参数：CustID和Loc_datetime，以此确保每个传送至服务器的操作均可识别其施动者和发生时间。

③在每个操作对象的目标页面(首部)中加入如下程序代码（伪代码，仅说明算法）：

CustID=getmessage(CustID)//接收顾客编号

Loc-datetime=getmessage(Loc_datetime) //接收顾客当地时间

ItemID=ITEMID//页面活动对象标识，为常量，已事先命名

Sys_datetime=now()//获取系统当前时间

ClickID++//确定记录序号

Click-Table.Insert(ClickID, CustID, ItemID,Loc-datetime, Sys-datetime)//写入行为数据仓库的Click表中

Date-Table.Insert(Loc-datetime)//将时间按规定维数拆开，写入时间因素表中

Date-Table.Insert(Sys_datetime)

基于顾客行为数据的分析

拥有了丰富的顾客行为数据后就可以根据决策的需要进行各类分析了。对于网上经营来说，利用顾客行为数据分析能解决的问题有：

1. 哪些页面访问率最高？

2. 哪些页面是多余的或最不受欢迎的？

3. 哪些页面与实际销售关系最大？

4. 哪些页面是“死页”。所谓“死页”指使用顾客中断继续访问的页面。

5. 确定顾客购买模式

6. 确定顾客在进行商品或服务购买前的预先访问次数等。

针对以上问题的分析结果对企业经营业绩的提高是很有帮助的，如确定网页的访问率有助于了解顾客的消费倾向，以便即时调整经营方向，提高经营业绩；“死页”的确定和排除会增加顾客对网上服务的访问；购买模式的确定有助于企业更好的理解顾客的消费行为，针对不同购买模式的顾客实施不同的营销策略等等。

目前数据分析的方法有多种，但对不同的问题要选择不同的分析方法。一般来说对于趋势或对比类型的分析（如针对问题1、2、4、6进行的分析），应采用多维视图分析工具或统计分析方法；对于规律（或知识）挖掘型的分析（如针对问题3、5进行的分析）则应采用数据挖掘方法。在实际工作中我们可以采用目前的一些商品化分析软件如面向多维数据分析的OLAP软件（如INFORMIX公司的MetaCube，ORACLE公司的Discover和Express）；面向数据挖掘的分析软件（如加拿大Simon Fraser大学研制的DBMiner）和一些统计分析软件（如SPSS）进行数据分析，当然除了使用商品化的分析软件外还可根据需要自行设计分析程序进行数据分析，两者配合使用往往会取得更好的效果。

存在问题及探讨

以上从理论上探讨了网上顾客行为分析系统的设计方法，但在实现过程中仍存在一些需要解决问题，突出表现在以下几个方面：

1. 如何保持顾客使用唯一标识（顾客号）

采用匿名方式访问网站会使顾客行为记录无确定的施动者，因而行为信息将产生一定程度的混乱，但将不注册顾客拒之门外显然是不明智的，企业所要做的是如何设计一种激励机制促使尽量多的顾客采用注册方式访问网站。在此我们可以借鉴“会员制”的作法，即对注册会员给予一定的回报（如价格打折），并且为了使顾客长期使用唯一标识（以实现对顾客行为的长期考察），可以将回报率与唯一标识的使用次数挂钩，以此鼓励顾客采用唯一标识访问网站，确保顾客行为信息的准确性。

2. 如何确定每个顾客活动（ClickStream）

如前所述，顾客在网站上的每项活动（如采购商品）是由一系列的Click操作组成的有序集合，如不确定每个Click操作记录与具体活动的所属关系（即Click属于哪个ClickStream），就会将顾客的所有活动混合在一起，顾客行为因此而失去了实际意义，这将无法解答诸如“顾客行为模式确定”之类的问题。解决办法之一就是在Click模型中加入活动标识（SessionID），具有相同SessionID的Click数据就构成了一个ClickStream，以实现与顾客每项活动的一一对应。但SessionID的确定是一个难题，因为只有顾客自己才知道操作与活动的隶属关系，但在操作过程中他们不会下意识的去区分操作与活动的关系，如果让顾客在进行每项操作时都向服务器声明操作所属的活动显然是不现实的，而且顾客一旦发觉其活动被“监视”很可能产生反感和抵触情绪，对企业经营造成不良影响，因此还应从系统自动识别上下工夫。■

参考文献

(1)《数据仓库技术与联机分析处理》王珊等 科学出版社

(2)《数据库系统概论》萨师煊 王珊 高等教育出版社

(3)《数据仓库—客户／服务器计算指南》(美) Harjinder S.Gill 清华大学出版社

(4)《The ClickingStream Data Mart:Window into Customer Behavior》(美) Ralph Kimball Intellegent Enterprise Vol 2,No.1