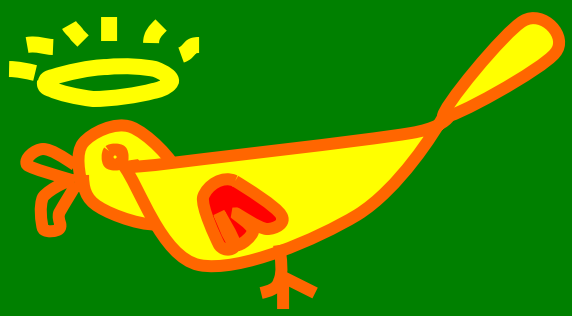
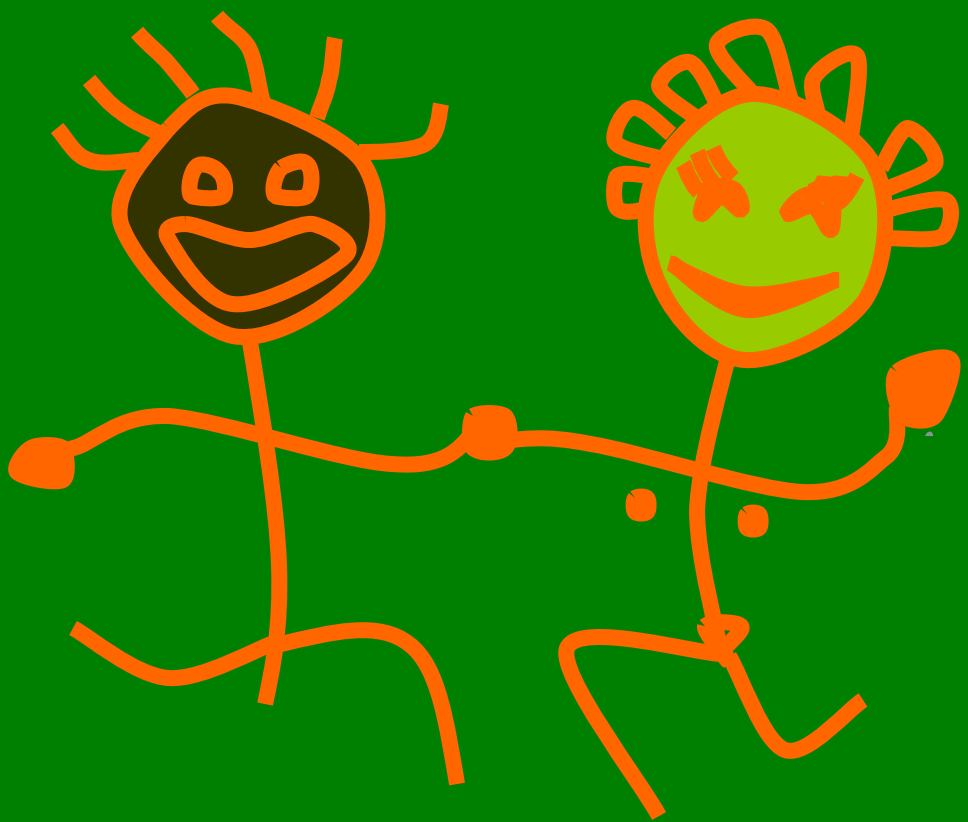
ttnn 观点

200906 vol.5 No.6 总第四十四期





主编

刘庆 (happyscry@gmail.com)

新闻编辑

金鑫

案例编辑

简荣

论坛编辑

傅如南

过刊下载

http://ttnn.appspot.com/mag/index

ttnn 天下会

http://ttnn.appspot.com

ttnn 讨论组

http://groups.google.com/group/ttnn

42

那些死去的人们
并未死去
他们在我记忆里

自从我坠落
被那咒语诅咒
我也同样留在那里

人生短暂
而我确定
必定还有更多

你以为自己也许是一个鬼魂
你没有到达天堂但让它接近
你以为自己也许是一个鬼魂
你没有到达天堂但让它接近

那些死去的人们
并未死去
他们在我记忆里

--Coldplay

目　录

BI 新闻简报

（一）

”六一“儿童节，祝普天下小屁孩和具备小屁孩之心的人们，节日快乐。

【热点】

IBM 发布了一款新的分析软件，“流式计算”的 S 系统，他越来越注重软件市场。有观察说，从他最近几年的活动看，逐步趋软，也是，按照人的年龄来算，这个快一百岁的家伙还能硬吗？SAP 也在拥抱变革，要走向更敏捷、更有响应、更以客户为中心的文化（那是不是意味着他们现在这方面做得不咋地？）最近他们的曝光度频频，如解决方案已经用在跟踪美国金融刺激计划的花费，BO 的副总畅谈他们的在线 BI 服务。

【动态】

Teradata 老总表示自己对数据仓库市场的信息， Oracle 那样的根本形不成威胁。能够这样说，当然是跟他们的产品在 dw 专业市场遥遥领先有关系。不论是在电信，还是在金融，他们走了一种很低调的路线，闷声发大财。

数据仓库专用设备，DWA，最近连续不断有新产品、新版本发布，显示了一派火热市场，有市场，就有对比。Netezza 说自己的盈利高于预期；Infobright，一个开源的 dwa 产品，被 Mavenir 公司选用了企业版作为数据仓库，支撑一种移动通信的业务。又跟 pentaho 合作，宣称发布一款集成的 BI 虚拟机。

在绩效管理方面，Tagetik 发布其 3.0 版本产品，Phophix 发布了金融控制器，Clarity 跟伦敦证券交易所在新的 BPM 服务上进行合作。HSBC 选用 SAS 的欺诈管理方案。还有跟天气相关的 BI 产品， Gold Eagle 是一家机油和添加剂销售商，他们的很多产品跟天气有关，因此，选用了 Planalytics 的业务天气智能服务。一切皆可 BI。

Datawatch 发布了 Monarch V10；XFormity 在很多食品行业，包括 KFC、必胜客等，都部署了其 BI 解决方案，以后我们去吃开封菜的时候，我们的用餐记录就进入 XFormity 系统里面去了。

最后可以看看 SourceMedia 推荐的年度 10 大技术型公司。其中包括 Oracle、SAS。

【技术评论】

知识不够智慧，君请看知识管理和商务智能的关系；

实时决策的技术，需要商务智能；

从社会媒体信息中可以淘金；

Kimball 提出了维度建模的 10 个重要原则；

最后请看 BI、数据挖掘在金融、民航、能源、运输方面的案例；

（二）

【动态】

IBM 2009 IOD（Information OnDemand）信息随需应变大会上，宣布了一项新策略，也是事关 BI 方向的举措，成立了 BAO（宝…包…）部门，业务分析和优化，此举目的在于让其客户跟上这个信息爆炸的时代。BAO，提供面向行业的垂直服务，当然其底层会借助其软硬件技术，来支撑这种服务。而具体的服务将是指导行业，如何管理数据，如何分析数据，如何利用从数据中得到的知识，当然，所有的这一切，都需要记住十八的工具。业务分析（BA）相对与 BI，如果说后者更偏向水平方向的平台，那么前者就是垂直的，面向行业的。Oracle 最近也在其 BI 套件中加入业务分析的模块，比如项目分析、客户忠诚度分析。

Teradata 建立了一个在线数据仓库开发者社区；HP 和 SAP 联合发布了一款用于直邮营销工具的方案，HP 的 Exstream 提供文档的自动化管理，SAP BO 的 Postalsoft 提供邮件地址的数据质量管理。Informatica 收购了 AddressDoctor，后者从名称上看就知道，这是一个数据质量，特别是地址数据数量方面公司。

Illuminate 发布了 iLluminate 4.0，这是一种相关数据库产品，跟传统关系数据库或者列式数据都不大一样。HighPoint，一个大型 IT 集成商，跟 greenplum 结盟，一个开源数据仓库专用设备。BeyeNetwork 刚刚发布了开源 BI 方案的应用情况，发现财务、销售和市场部分用的还是多些（这个结论好奇怪，主要就是这些部门啊），可以浏览全报告。

SunGard 发布了其 Fame 新版本，10，为财务公司提供细节交易数据的分析。Quantrix 发布了 DataNAV，一款集成数据探索、分析和可视化的工具。NASA 将语义分析技术运用到其星座计划中，用的是 TopQuadrant 的 TopBraid。英国电信跟 Kognitio 合作，更新了数据即服务（DaaS）的合同。Ventyx 和 Visual Mining 都发布其绩效管理产品。

在六月份的第一周，Dataflux，一个数据质量专业厂商突然发力。首先是宣布了支持 SAP NetWeaver 的连接产品。又发布了一串很强大的白皮书，包括数据质量整治，数据监管七式，从创收角度看数据监管的 ROI，从角色理解数据监管 ROI，数据监管的成熟度模型，用相关性指标

来制定数据质量计分卡，衡量数据质量服务水平，数据迁移的三个层次，如何用数据质量来改善分析精度…精品。

【观察和评论】
一些案例：
孤立点分析技术助审计防舞弊；
电信数据仓库项目实施方法论；
电信如何进行精准营销；
学校利用 BI 提高学生成绩；
可口可乐的 RFID 方案重新定义 BI；
从财政部的压力测试看 BI 优点；
跟踪刺激计划，BI 让政务透明；

（三）

【动态】

SAS 发布数据移植工具包，包括数据整合服务器、数据访问引擎以及来自 dataflux 的数据质量方案。同样，是租用制付费的，适用于那些短期项目，比如几个月的项目。全国房地产经纪人协会使用 SAS 的可视化数据发现方案，不知道这个工具是给谁用的，如果放到中国，肯定是不会给消费者，房地产就靠信息不对称来赚钱了，将信息都告诉你，让你看个透气不是太亏。SAP 公布了其 SaaS 的策略，实现一年前的设想。不仅是为了中小企业提供服务，还面向大型企业，野心不小。

一家新西兰的公司 WhereScape 十年磨一剑，发布一款数据仓库的集成开发环境（RED），可以让数据仓库开发更迅速，更容易维护。这对目前还像是手工作坊式的 DW 项目来说，还真可能是一种突破。

印度最大的零售商 Future Group 采用 greenplum 来作为他们散步在全国各地零售网点的数据云。所谓企业数据云（Enterprise Data Cloud），是 greenplum 刚刚倡导的一个新概念。企图重新定义数据仓库和分析市场。他们也刚刚发布 v3.3 版本的数据库，提高了一些并行处理能力并可绑定了更多的硬件平台。

QlikTech 总是卖给小客户，很不爽，他们想要卖大客户。刚刚发布了 QlikView9，号称让 BI 更简单、更快速更具可用性。这是一家快速增长的企业，前景一片光明。MSTR 发布免费的报表套件，面向中小企业。PivotLink 和 Host Analytic，两家宣扬 SaaS 的公司，前者是偏向展现，后

者主攻绩效管理，俩人合作进攻。Information Builder 双喜临门，又是发布 WebFOCUS 新版本，又是发布 iWay B2B 套件，这是个想是企业数据网关的东西，协调企业跟合作伙伴的数据交互。

【观察】

在经济艰难时期，BI 市场有 22%的增长，给 BI 人带来信心了吧。新技术给 BI 带来新的机会，请看 SaaS 带来的改变，或者还有一个概念“按需分析”。

如果说人们对海量数据的管理需求依次是快速、方便、准确、安全的话，现在的阶段即将转移到准确之上，请看数据监管工作的进化史。对于安全却已经很少谈及，但这也不是一个小事。

::ttnn::

Web 挖掘在电子商务货源搜索中的应用

发布：2009-06-19
来源：CIO 时代

1 概述

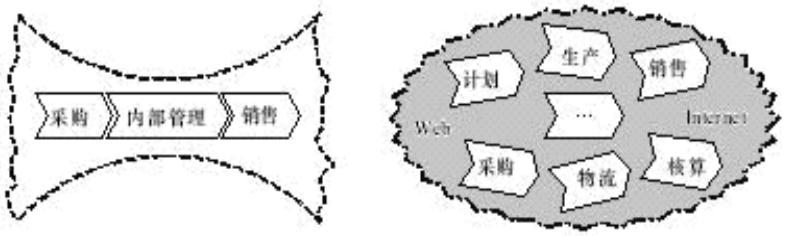
随着计算机网络技术及互联网的发展，电子商务(ElectronicCommerce, EC)在企业经营业务中的应用越来越普遍。它是传统企业的经营业务在互联网环境下电子化的结果，这些经营业务包括产品或服务的交易以及为实现交易而发生的各种相应业务活动。随着电子商务在企业应用不断深入、对外联系增多，企业对潜在客户、供应商和产品等方面的货源信息要求越来越高。电子商务环境中的企业必须超越以往相对狭隘的经营环境，有效地收集、利用货源信息。

Web 挖掘是提高电子商务企业效率的有效工具，它从 Web 内容、结构、使用等方面提炼对电子商务运作有价值的信息。利用 Web 挖掘技术在信息方面支持高效电子商务的研究包括:将数据挖掘(data mining) 技术应用扩展到 Web 挖掘的应用中;采用信息检索技术对 Web 信息进行分类、筛选;支持电子商务运作的信息收集等。

本文研究电子商务环境中企业如何有效利用互联网收集、挖掘业务信息的问题，分析了 Web 挖掘在电子商务中的作用，利用 Agent 和 Web 挖掘技术设计了以元搜索引擎为核心的货源搜索机器人。元搜索引擎利用通用搜索引擎扩大信息搜索范围，搜索有关货源信息，采用 Web 挖掘方法对货源信息进行过滤分析，从中筛选对企业有潜在价值的客户、供应商和产品信息，为企业电子商务中的业务处理和决策提供依据。

2 货源搜索

电子商务在给企业带来巨大发展机遇的同时，也使企业面临超出传统经营模式的挑战。如图 1(a) 所示，在传统经营环境下，企业的市场范围物理上受到地区或国家的限制，与客户、供应商的业务范围主要集中在采购、销售等外部业务环节，企业竞争压力小。如图 1(b)所示，在电子商务环境下，市场和业务范围延伸扩展，企业之间合作加强，并向企业内部渗透，企业与伙伴之间合作紧密，竞争对手增多，竞争压力变大。为了适应电子商务这种大范围、高强度的竞争环境，企业必须充分利用电子商务的有利条件，大力挖掘潜在客户、供应商、产品和竞争对手的相关信息。本文将这些信息统称为货源信息。



(a)传统环境　　(b)电子商务环境

图 1　传统企业经营环境与现代电子商务环境

采用 Agent 与 Web 挖掘技术自动进行货源信息搜索、分检的计算机软件称为货源搜索机器人(Business Search Robot)。货源搜索机器人的主要功能包括:

(1)根据用户需要，设置搜索线索条件信息;

(2)在互联网上收集满足搜索条件的页面信息;

(3)分检所得页面信息，按特指领域知识进行页面过滤、分类、索引;

(4)存储搜索结果于数据库中;

(5)以用户所需形式提供相关货源信息。

实现这些功能的关键在于(2)和(3)的实现。本文从搜索引擎和货源信息分检方面介绍货源搜索机器人的设计与实现。

3 货源搜索引擎

搜索引擎(Search Engine，SE)是以互联网上 Web 站点提供的页面为信息源，为方便信息使用者检索所需信息而设计开发的计算机软件。目前商业化的通用搜索引擎有很多，如:百度，搜狐，Yahoo!, Google, Excite,A1ta Vista 等。尽管通用搜索引擎正不断提高计算和网络访问能力，但仍存在以下几点不足:

(1)每个通用搜索引擎相对于整个互联网的覆盖范围是有限的;

(2)在搜索结果中有相当一部分是和搜索内容无关的;

(3)搜索结果的链接有些是无效链接。

为避免以上问题，本文提出的货源搜索机器人采用元搜索模式设计搜索引擎。元搜索引擎利用多个通用搜索引擎来实现完成自身的搜索，搜索范围要大于单个通用搜索引擎。通过筛选、过滤搜索结果，得到与搜索目标内容尽可能接近的结果。该元搜索引擎的设计结构如图 2 所示。

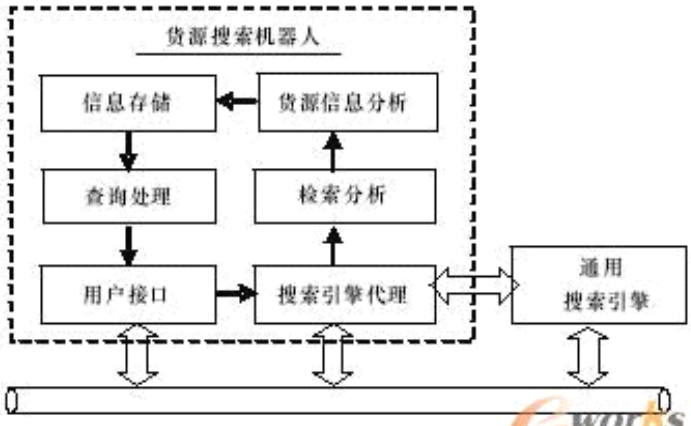


图 2　基于元搜索引擎的货源搜索机器人

货源搜索机器人分为 6 个功能模块:

(1)搜索引擎代理模块按预定的搜索线索制定通用搜索引擎使用的搜索条件，并提交给通用搜索引擎，通用搜索引擎再把搜索结果返回给搜索引擎代理。

(2)检索分析模块对从通用搜索引擎得到的搜索结果进行解析，检验网络链接的有效性。

(3)货源信息分析模块对检索分析结果进行整理、归纳和分类，得到与领域相关的货源信息数据。

(4)信息存储模块负责把分检得到的货源信息存储在数据库中。

(5)查询处理模块根据服务请求在货源信息数据库中检索，并把检索结果反馈给用户。

(6)用户接口模块负责接收用户的检索服务请求，设置系统的参数。

4 货源信息分检

在运用元搜索引擎收集到货源相关信息(raw informarion)后，下一步是对这些信息进行货源信息分检。货源信息分检分为 2 个方面来实现:系统分检和用户分检，如图 3 所示。系统分检的处理对象是从搜索引擎获得的搜索结果，输出是特指领域相关的货源信息。用户分检的处理对象是系统分检的结果，输出是用户需求相关的货源信息。

图 3　货源信息分检

4.1 系统分检

系统分检对收集的货源信息进行分析整理。由搜索引擎获得的搜索结果虽然经过搜索词的过滤，但为了尽可能扩大搜索范围，搜索结果页面往往很多，其中有很多是与特指领域不相关的。系统分检相当于对搜索结果的预处理，筛选出利用价值更高的货源信息。系统分检的实现步骤如下:

(1)校验搜索结果页面的有效性;
(2)从搜索结果摘要中抽取描述词汇;
(3)分析描述词汇与领域知识叙词的相关性;
(4)根据叙词相关性分检搜索结果;
(5)排除相关度低于预设 17 值占.的搜索结果;
(6)解析搜索结果页面;
(7)将系统分检结果存入货源信息数据库待查。

其中，确定搜索结果与特指领域知识叙词的相关性可以根据需要采用不同判断模型。本文以向量模型为例加以说明。假设搜索引擎搜索到 S 个页面，搜索特指领域知识有 N 个叙词。系统分检中还可以采用其他方法或模型对搜索结果领域相关性进行确定，例如基于概率、模糊集合、隐含语义等的判断模型。在特指领域知识表示上，可以采用多层面、多角度的方法选择叙词，设置相应权重。具体实现可借鉴信息检索中全局或局部聚类方法。

4.2 用户分检

用户分检是按用户需求进行的。如果把系统分检看作一次分检，那么用户分检就相当于对货源信息的二次分检。用户需求表示为 DNF(Disjuncrive Normal Form)范式的形式，用户分检的实现步骤如下:

(1)用户输入需求，设置货源查询词;
(2)在货源数据库中检索满足用户需求的信息;
(3)分析检索结果与用户需求的相关性;
(4)保留相关度高于预设 17 值民的检索结果;
(5)以用户所需形式输出检索结果。

综上所述，在货源信息分检过程中，利用系统分检对所关注的领域相关信息进行大范围搜索和初步筛选过滤，再在用户的参与下利用用户分检对货源信息进行小范围的搜索，就可以找到用户需求满意度较高的货源信息。

5 实验结果及分析

本节通过实例计算说明了货源搜索机器人的搜索效果。通过 2 次对货源信息的分检，使搜索结果的查准率和查全率均得到一定的提高。实例采用网络新闻组文献(选自 USE-NETnewsgroups)作为实验数据，其中，包括汽车类、摩托车类等其他类文献共 2，000 篇。

5.1 系统分检结果

搜索目标领域是与汽车类相关的货源信息，汽车类文献共 600 篇。系统分检结果如表 1 所示。从表 1 可以看出，系统分检可以有效地从通用搜索引擎的返回结果中提取出与搜索领域相关的货源信息，为之后的用户分检做好充分准备。

5.2 用户分检结果

假设用户搜索目标是满足表达式,并与汽车类相关的货源信息。用户分检结果说明，如果直接在通用搜索引擎返回的结果中进行搜索(不经过系统分检)，则用户分检的查准率平均值在 50%左右，经过系统分检后，查准率平均值能达到 75%，并且在返回文献数相同的情况下，经过系统分检后的查准率比不经过系统分检的查准率平均提高 22.1%，查全率平均提高 15.9%。

用户分检的比较结果表明，在查全率相同时，经过系统分检后的用户分检的查准率明显高于不经过系统分检的查准率。因此，货源搜索机器人通过系统和用户的 2 次分检搜索领域相关的货源信息是非常有效的。

6 结束语

本文针对电子商务环境下的货源信息搜索问题，采用 Web 挖掘和信息检索技术，提出一种货源搜索机器人设计与实现的方法。这种基于元搜索引擎的搜索方法扩大了货源搜索范围，通过对系统和用户的 2 次分检发现更有价值的货源相关信息。文中提出的搜索机器人的设计方法对其他领域知识相关的大范围信息搜索也有很好的应用价值。

::ttnn::

建立汽车维修行业管理数据仓库满足分析需求

来源： http://www.itxinwen.com/ View/new/html /2009-06/2009-06-19-566651.html

发布：2009/6/19

0 引 言

随着汽车工业的飞速发展和人民生活水平的逐步提高，汽车维修行业和汽车配件行业得以迅猛发展。以宁波为例：截至 2004 年底，全市现有各类汽车 (摩托车)维修业户 3 747 家，其中一类维修企业 85 家，二类维修企业 424 家，三类专项修理企业 1 246 家，摩托车维修业户 1 001 家，

机动车配件经营户 956 家，车辆综合监测站 9 个，客运车辆安全门检站 26 个。

全市汽车维修企业(户)从业人员共计 14 922 人，其中生产工人 11 639 人，占从业人员总数 78%，生产工人持证率(上岗证)达到 99.3%。生产工人持证等级：初级工 5 684 人，占生产工人总数 48.8%；中级工 5 249 人，占生产工人总数 45．1%；高级工 628 人，占生产工人总数 5．4%；技术人员(包括管理和双证书人员)421 人，占生产工人总数 3．6%。

2002 年与 1997 年全市维修企业维修车辆同比增长情况如下：汽车大修增加 46．4%，二级维护增加 58.8%，摩托车维修增加 77.6%，维修企业共维修车辆增加 23.8%，维修营业额增加 39.3%。

在发展如此迅速的现状下，如何推动汽车维修企业开展良性竞争，引导行业的正确发展方向成为亟待研究和解决的问题，同时也是需要行业管理部门予以关注和指导的问题。汽车维修行业管理部门在规划和指导行业发展时，需要大量长期积累的关于维修企业及汽车工业发展的历史数据作为数据支持和依据。目前，我国汽车维修企业在日常管理及经营活动中已开始推广使用基于联机事务处理(OLTP)的汽车维修企业信息系统，积累了大量汽车维修企业的细节数据，但是 OLTP 本身的特点决定了它所采集的数据缺少综合性，无法为管理者的决策提供综合有效的数据支持。如何将各企业积累的数据进行综合管理并用于维修行业管理和决策已成为当前需要解决的主要问题。

综上所述，本研究考虑将数据仓库用于管理大量的汽车维修企业积累的数据，让这些数据为汽车维修行业的管理发挥作用。

1 设计与建立汽车维修行业管理数据仓库

1.1 总体框架构建

汽车维修及配件行业管理数据仓库系统的结构可以划分为如图 1 所示的 3 个部分：数据采集、多维数据结构构建和管理以及 OLAP 应用系统的开发。

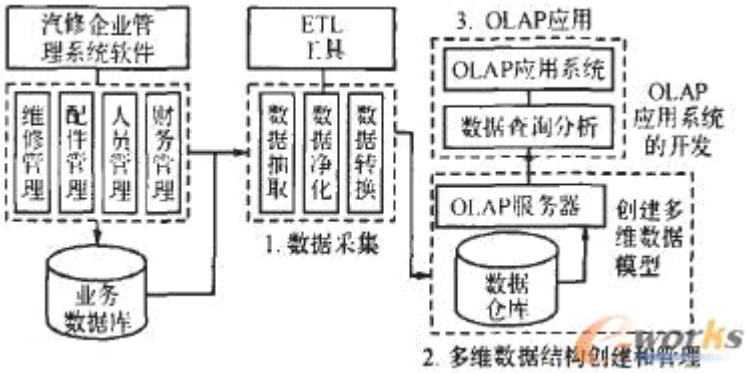


图 1 汽车维修及配件行业管理数据仓库系统结构

在数据采集阶段，根据数据仓库的业务需求分析和主题域分析，从宁波市公路管理处(管理当地汽车检测维修和配件销售企业的政府主管部门，以下简称公管处)下属各企业的上报数据中抽取数据，进行清理、转换后载人数据仓库中。

在多维数据结构创建和管理阶段，根据公管处管理人员的分析建立数据模型，将数据仓库中的数据按照一定层次进行聚合和汇总，构成信息分析的多维视图，最后选择一定的存储模式，将这些多维视图存储在 OLAP 服务器上。

构建数据仓库的目的是满足管理人员决策分析的需求，因此用户界面友好、软件易用也是十分重要的，另外还要考虑用户所支付的开发和使用成本，所以在 OLAP 应用系统中考虑使用目前成熟的 BI 产品作为数据仓库的前端工具，以减少开发成本。

1.2 汽车维修及配件行业管理数据仓库建模

数据仓库建模是构建数据仓库的重要组成部分，是数据仓库构造开始的第一步。一般来说，在数据仓库开发过程中，数据模型具有 3 个不同的层次：概念模型、逻辑模型和物理模型。每一个层次实质上是前一种数据模型的精炼或更加详细的表达。在设计期间，通过多层次细化，建立与用户需求更加一致的面向主题的数据仓库。

1.2.1 概念模型的建立

笔者多次到公管处调研，通过了解他们的工作流程，确定以下几个内容是决策者希望执行的数据分析：全市各地区汽车维修企业的数量、情况及变化趋势；全市各地区汽车维修企业设备总价值和厂房面积变化趋势；全市各地区汽车维修量和维修营业额的变化趋势；维修人员技术水平(工种等级)、教育背景和培训时间的变化趋势；汽车维修合格率和返修率；客户投诉情况趋势等。

汽车维修及配件行业管理数据仓库在建设过程中将综合考虑公管处的业务需求，同时仔细研究汽修行业目前积累的数据，通过业务驱动和数据驱动相结合来构建数据仓库。根据各方面分析结果，确定主题域描述如表 1 所示。

表 1 主题域描述

主题名	公共码键	属性组
企业基本情况	企业_ID	企业信息：企业_ID，企业编号，区域，建立日期，注册资金，地址，电话，结束日期，是否营业……
		企业资金产值信息：企业_ID，原始资金总数，资产净值，流动资产，总产值，……
		企业维修设备场地信息：企业_ID，通用设备总数，通用设备总金额，维修主厂房面积，企业占地面积……
维修	维修编号	维修信息：维修编号，汽车_ID，维修种类_ID，维修费用……
职工	职工_ID	企业人员工资教育表：职工_ID，工种，工种等级，月培训时间，月工资……
		企业职员固有信息：职工_ID，企业编号，职工姓名，职工性别，所属部门……
投诉	投诉编号	投诉情况表：投诉编号，投诉日期，企业编号，投诉件数，解决投诉数量……

1.2.2 逻辑模型的建立

1.2.2.1 分析主题域

数据仓库的设计是一个逐步求精的过程，要对概念模型设计步骤中确定的基本主题域进行分析，上述概念模型中的企业基本情况主题联结各主题域。所以先考虑用这个主题来实施数据仓库的构建；维修主题中的维修是维修企业基本业务，又是进行决策分析的主要领域，第2 步确定“维修”主题并实施数据仓库的构建；最后再完成职工和投诉主题。

1.2.2.2 粒度层次划分

由于在汽车维修行业信息管理系统中，对数据的查询和分析是多层次的，为了提高查询和分析质量，数据仓库按多重粒度来组织数据。

分析数据时，在时间维上要对每月、每年企业的设备、人员和维修情况进行分析。所以在本项目中企业的基本情况数据和职员的情况数据是按照月份来组织的；维修事实表则是按照 El、月、年 3 种粒度组织数据；由于投诉事件比较少，为了方便管理人员进行处理，投诉主题是按照投诉日期组织数据的。在本项目的数据仓库中采用多重粒度组织数据。

1.2.2.3 关系模式定义

数据仓库的每个主题可以由多个表来实现，笔者对选定的当前事实的主题进行模式划分，形成多个表，并确定各表的关系模式。

以企业基本情况主题为例：在本主题中有企业资金产值表和企业设备场地表等多个事实表。由于篇幅所限，这里只截取企业资金产值信息事实表，如图 2 所示。

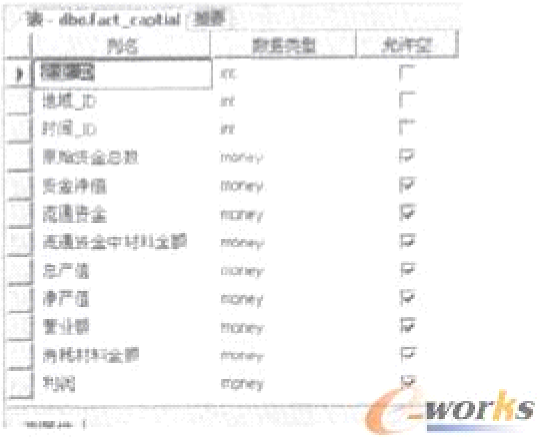


图 2 企业资金产值信息事实表截图

常见的基于关系表的存储方式有两种：星型模型和雪花型模型。星型聚合快，效率高；雪花型结构明确，便于与 OLTP 系统交互，占用空间小，但是模式较复杂，浏览困难，并且额外的连接将使查询性能下降。在本项目中，数据量较小，数据结构简单，所以本项目选用星型架构，本项目数据仓库中维度模式的星型架构模式图如图 3 所示。

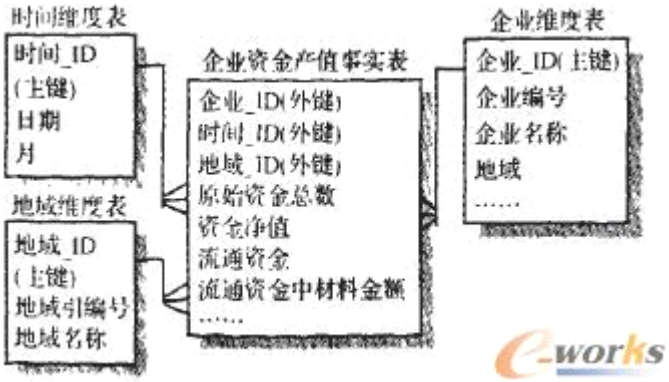


图 3 企业基本情况主题的企业资金维度星型架构模式图

在数据仓库构建过程中采用维度建模，经过对主题的分析和数据的研究，数据仓库的每个主题可以由多个表来实现，对选定的当前事实的主题进行模式划分，形成多个表。本项目中关系模式均采用星型架构，数据仓库中目前共建立了 5 个事实表、9 个维度表。

1.2.3 物理模型的建立

在考虑存储结构时应考虑 3 个方面的因素：存取时间、存取空间和维护代价。为了提高分析质量和响应速度，采用了一定的数据冗余，这些数据冗余对数据库查询速度的影响不大，可以忽略不计。

在逻辑设计时使用 Microsoft 公司的 SQL Server 2005。本项目中索引建立策略为 :首先为各表(包括事实表和维表)的主键建立聚集索引；在事实表的外键上建立非聚集索引；然后根据实际的运行情况，通过 RDBMS(关系型数据库管理系统)提供的数据库监控工具，建立一些合适的非聚集索引，从而获得最高查询性能。数据存放位置涉及存储设备的存取速度。由于目前高速存储设备较为便宜，性价比较高，因此主要以硬盘为存储媒介。

2 联机分析处理的设计与实现

利用 SQL Server 2005 中 Analysis Services 功能建立 OLAP 多维数据库，利用上文中构建的数据仓库作为数据源。建立多维数据集时，根据需求分析和数据仓库的建模分析，选择合适的事实表、度量值和维表，利用 SQL 生成维度。本项目共生成 5 个多维数据集，并在此过程中生成 9 个共事维度，如图 4 所示。

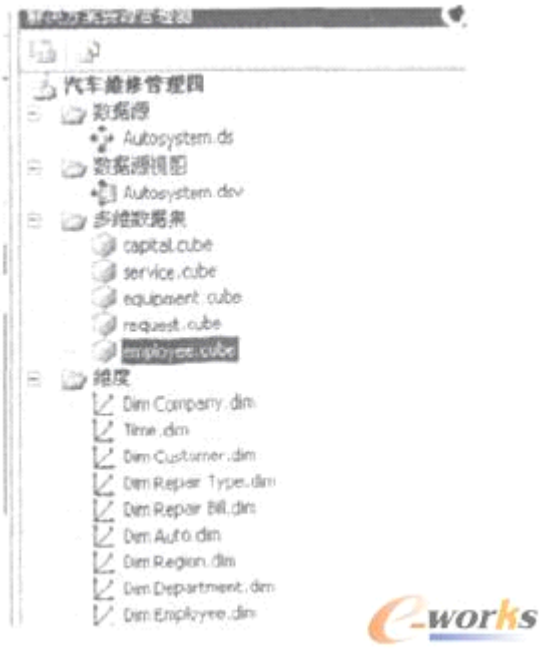


图 4 多维数据集

本研究可以采用 Microsoft 的 EXCEL 软件或普科(ProClarity) 公司的 ProClarity6．0 作为数据显示分析的前端工具。前者具有成本低廉的优点；后者功能强大且能和 SQL Server 2005 实现无缝联结，使用方便。使用 ProClarity6．0 风险报警功能分析 OLAP 立方数据，其界面如图 5 所示，图中数据为企业的利润率，粗体数字表示该企业的利润率低于警戒值，这时管理部门应该密切注意有报警标志的企业，尽量避免由于企业经营不善引发的恶性连锁事件。

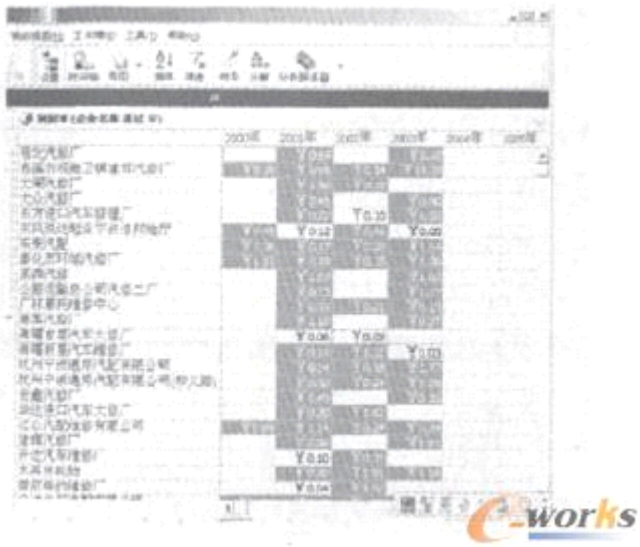


图 5 使用 ProClarity6．0 风险报警功能分析 OLAP 立方数据

3 结束语

在全面调研汽车维修行业管理现状的基础上，笔者研究了数据仓库系统的设计与实现方法，提出了将数据仓库应用于汽车维修及配件行业管理的设想，构建了汽车维修行业管理数据仓库体系结构，并研究其设计方法、构建技术及联机分析等技术。

在此过程中，数据仓库的建模是整个开发过程中的关键技术，本研究采用维度建模，首先考虑如何满足和适应汽车维修行业管理的需求，通过业务驱动和数据驱动相结合的方法实现了该目标；其次是如何提高查询效率，本研究采用星型架构和适度数据冗余来提高查询效率；最后要考虑数据仓库的未来可扩展性，汽车维修行业管理的业务除了在文中列出的企业情况、维修、职工和投诉外，还有汽车检测和汽车配件出售业务，这些业务的数据可通过添加相关主题域并构建事实表和维度表，方便地加入到目前的数据仓库中。

::ttnn::

孤立点分析技术助审计防舞弊

来源：http://finance.ifeng.com/stock/roll/20090605/750101.shtml
发布：2009/6/5

最有效的治理财务报告舞弊的方法就是提高各方的识别能力，而海量的电子数据使审计面临前所未有的困难，不仅在识别财务报告舞弊方面有困难，连无意识造成的虚假财务报告也难以识别，从而大大增加了审计风险。因此，关键是借助各种审计技术，提高审计人员的识别能力。在这里专门介绍一种孤立点计算机识别法。

孤立点分析的 5 种方法
孤立点是数据集中与一般数据模型不相符的数据。一般情况下，在数据导入数据仓库之前，应经过数据清理，消除其不一致的现象。

但在实际应用中，经常会出现一些客观存在的，非操作人员的人为因素而导致的异常数据。对于这些异常数据，既无法按照一般可行的分类规则对其划分，也无法通过聚类方法将其与其他数据建立有效的联系，应用孤立点分析技术对数据进行分析，却能有效识别这些异常数据，从而把虚假财务报告从中识别出来。一般情况下，应用于虚假财务报告识别的孤立点分析主要有如下方法：

1．基于统计的方法采用某种概率分布拟合数据集，根据该分布对数据集中的每个数据对象进行“不一致性测试”，如果与分布不符合，就认为它是一个孤立点。基于分布的方法易于理解，对数据分布满足某种概率分布的数值型单变量数据较为有效。但对大多数挖掘应用来说，数据分布形式事先并不知道，需要多次的实验才能得到合适的数据分布形式。而且有的数据分布并不满足任何概率分布，该方法难以适应。

2．基于距离的方法基于距离的孤立点的概念认为，如果一个数据对象与数据集中大多数对象之间的距离（相异度）都大于某个阈值，就是一个孤立点。基于距离的孤立点定义体现孤立点的本质，避免了基于分布方法中的数据分布适应性问题。

3．基于偏差的方法基于偏差的方法不采用基于分布或基于距离的度量值来确定孤立点。相反，它通过检查一组对象的主要特征来确定孤立点。与给出的描述“偏离”的对象被认为是孤立点。

4．基于密度的方法基于密度的孤立点的定义是在基于距离的基础上建立起来的。

这种方法将数据对象之间的距离和某一对象在其指定的邻近范围内包含的对象个数这两个参数结合起来，得到“密度”的概念。根据密度来判断一个对象是否是孤立点。该方法能有效的发现局部的孤立点。

5．发掘时序孤立点数据的方法时序孤立点数据对象一般是指那些与时间上相邻的对象相比，幅度变化比较大，且持续时间比较短，将它们从序列中移去，剩下的序列将变的很平滑，可以获得比较简洁的表示。

企业的财务报表数据会随着企业经营业务的变化而变化。实践表明，真实的财务报表中主要项目的数据变动具有一定的规律性，如果其变动表现异常，说明数据中可能存在虚假成分。孤立点分析对虚假财务报告中数据的异常变动识别有着非常重要的应用价值。在实际操作中，首先选择能够显著显示财务欺诈征兆的一些关键财务指标，如应收款项比率，应收款项周转率，资产负债率，速动比率，主营业务税金及附加比率，资产质量，管理费用和销售费用率等，并为其设定一个阀值，通过孤立点分析方法的应用分析，一旦财务报告中的相关财务指标数值超过这个阀值，说明报告有可能具有虚假性。

孤立点分析的规则和算法

经过实证分析，虚假财务报告的形成受一定的环境因素的影响，因此，可以利用数据挖掘技术分析行业环境与企业环境，以识别出有会计造假倾向的公司，在此基础上进一步增加审计程序，提高审计质量，降低审计风险。ＪｏｓｅｐｈＴ．Ｗｅｌｌｓ指出“财务报告舞弊”不是始于管理层的不诚实，而是发端于某种环境—— —这种环境中存在两个特征（激进的财务业绩目标），目标未实现将被视为不可宽恕的氛围。换言之，财务报告舞弊缘于压力。我国的研究者通过大量的统计研究（陈信元，杜滨等 2001），总结出极有可能采取会计造假的公司的特征，通常包括如下几个特征：

1．前两年连续亏损，本年经营业绩没有得到根本改善的公司（为了避免被特殊处理）。

2．前两年平均净资产报酬率达到 10%，本年行业不景气的公司（为了争取配股的资格）。

3．资本运作和关联交易频繁的上市公司。

4．业绩和股价波动幅度较大的上市公司。

5．全行业亏损或行业过度竞争的上市公司。

美国Ｃｏｏｐｅｒｓ＆Ｌｙｂｒａｎｄ会计师事务及知名学者对美国上市公司财务报告欺诈行为进行多年研究后曾经总结出 29 面红旗（即特征）。

一旦出现这些红旗，就需要格外关注公司管理当局是否存在财务报告舞弊的可能，比较典型的情形有：

1．现金短缺，负的现金流量，营运资金及 / 或信用短缺，影响营运周转。

2．融资能力（包括借款及增资）减低，营业扩充的资金来源只能依赖盈余。
3．成本增长超过收入或遭受低价进口品的竞争。
4．发展中或竞争产业对新资金的大量需求。
5．对单一或少数产品，顾客或交易的依赖。
6．夕阳工业或濒临倒闭的产业。
7．因经济或其他情况导致的产能过剩。
8．现有借款合约对流动比率，额外借款及偿还时间的规定缺乏弹性。
9．迫切需要维持有利的盈余记录以维持股价。

在审计领域利用孤立点分析技术识别虚假财务报告的研究目的是确定孤立点分析的方法论，建立相应的规则和算法。具体而言，需要运用孤立点分析技术整合上市公司财务数据、经营管理、证券市场交易及宏观经济环境等多方面的非财务信息，然后在大量数据模拟和试验的基础上，给出识别各种类型的财务造假模式的数据挖掘解决方案、规则、算法等。在应用数据挖掘来识别虚假财务报告时，可以利用上述特征建立相应的规则与算法，同时应用孤立点分析技术来识别出可能采用会计造假的公司，在保持应有的职业谨慎的态度下，进一步增加对这些公司的审计程序，以降低审计风险，提高审计质量。

::ttnn::

数据挖掘技术油田生产决策支持研究

来源：http://cio.ccw.com.cn/research/hangye/htm2009/20090529_632204.shtml
发布：2009/5/29

数据仓库体系结构的分析与设计
文中提出了基于油田生产数据采用数据抽取、转换和加载技术的数据仓库的构建策略以及基于多维数据集的数据挖掘的实施方案，并对包括数据仓库体系结构的设计、数据仓库的构建、多维数据集数据存储模式的优选策略、在线分析处理以及基于分层聚类分析的方法实现数据挖掘等在内的各主要环节进行了系统详细的阐述。最后结合油田生产数据，综合运用数据仓库、联机分析处理和数据挖掘技术构建了一套油田企业生产决策支持系统，并提取和挖掘出了对于油田生产决策支持有用的信息。

随着油气勘探开发工作的不断深入，经常需要处理、使用大量的信息数据，而在这一过程中往往出现以下问题：管理人员的操作日趋复杂、用户分散、相互联系程度低、信息共享程度低；信息加工、处理手段差，无法直接从各级各类业务信息系统采集数据并加以综合利用，业务系统产生的大量数据无法及时提供给决策部门。作为油田管理人员，仍需在查询多个基于各种异构数据源的业务系统和外部系统，并进行大量的数据分析后才能做出决策。工作量大，且容易出现人为差错，从而影响决策的质量。

在油田生产过程中，积累了大量的生产管理历史数据和成果数据，从事务型数据中得到有价值的决策信息越来越困难。因此，通过建立有兴趣的模型，提取和挖掘出大量数据后面的“知识”，探索出油田生产中的规律性，可以预测油藏开发指标、未来的生产情况等，从而更有效地进行生产调整和优化，并为参与市场竞争做出重要的决策。数据挖掘是实现油田的智能化决策的现代化油藏管理的重要技术，因此，在合理构建数据仓库平台的基础上，开展在线分析处理与数据挖掘技术的决策支持系统的研究工作是有意义的，并为决策人员研究油田生产的发展走势提供可靠的技术支持。

1 数据仓库体系结构的分析与设计

数据仓库的数据来源广泛，使用要求多变，查询要求复杂，传统的数据库系统结构无法提供足够的灵活性来满足这种复杂多变的使用要求。因此，从用户角度来分析与设计数据仓库的体系结构，首先应根据数据仓库的使用要求确定分析的主题和各种分析指标，数据在进人数据仓库的存储之前，必须经过数据抽取、清洗和转换等预处理过程。然后，选择合适的存储模型，将它们进行有效的组织，并存储在数据仓库之中，继而从中分析并挖掘出潜在的、隐藏的有用知识，为决策支持提供可靠信息。

一般数据仓库系统的体系结构可设计 3 个独立的数据层次：信息获取层、数据管理层和应用服务层。而考虑到油田生产数据信息的特点，本文提出的油田生产决策支持系统是由源数据层、数据获取层、数据管理层、数据分析层和数据展示层共 5 层构成的系统体系结构。如图 1 所示。

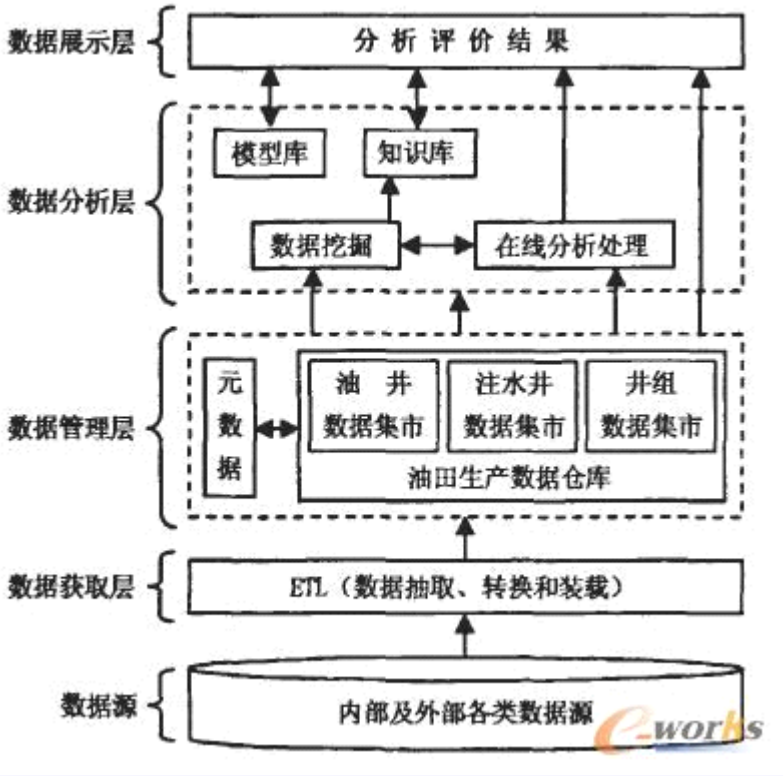


图 1 油田生产决策支持系统数据仓库体系结构

数据源层主要存放着油田生产过程中的大量历史数据和在分析决策时需要用的外部数据。数据获取层从源数据层中抽取分析决策所必须的相关数据，然后将净化和转换后的数据集成到油田生产数据仓库中。通过数据管理层对数据仓库中的数据和数据源进行存储和管理，根据不同的主题建立数据集市来减少数据处理量。针对不同主题的数据集市，数据分析层中进行在线分析处理与数据挖掘，实现数据的多层次的分析和挖掘。然后数据挖掘工具将数据仓库中挖掘的知识放入专家系统的知识库中，通过知识推理达到定性分析辅助决策。而模型库则实现多个模型的综合决策。最后数据展示层将分析结果通过图件或表格的形式提供给相关决策人员，辅助决策。

2 油田生产数据仓库的构建

数据仓库的构建过程中首先需要进行数据建模，确定系统主题域。以井组生产为例确定的系统主题为：不同层位注采工艺和注水量的不同对油井生产的影响。

主题域一经确定，就可以对每个主题的内容进行较明确的描述，通过分析所需使用的数据包括：生产时间、油井属性数据、油井生产数据、注水井属性数据、注水井生产数据及层位属性数据，进而可以确定每个主题的事实和维度，并使用多维数据模型建立数据仓库的概念模型。对于井组

生产主题来说，决策者所关心的事实数据为日产液量、日产气量、气油比、含水和日配注水量等。传统的概念模型注重的是数据的结构，对于分析型应用是不合适的，而多维数据模型注重的是数据的含义，能够清楚地表达分析领域的数据模型，因此，数据仓库的概念模型可采用多维数据模型来建模。如图 2 所示。

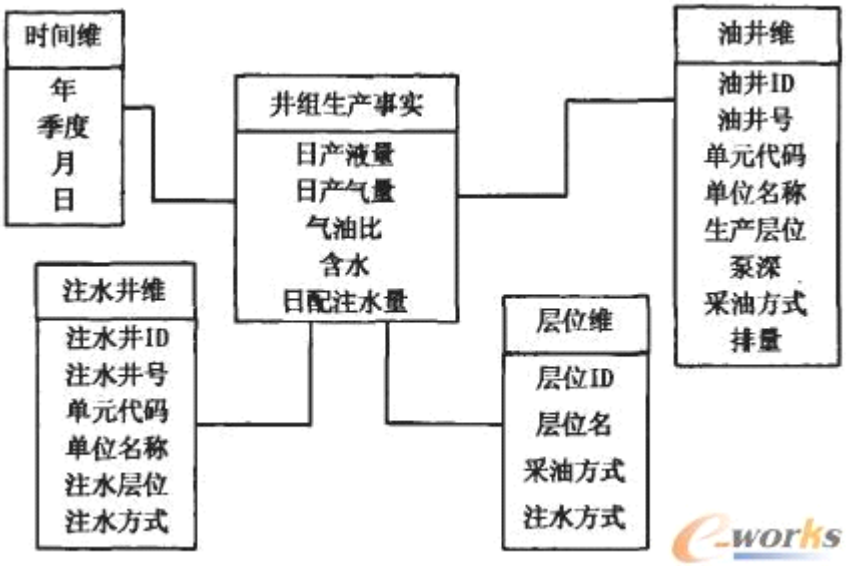


图 2 井组生产的多维数据模型

根据上面的概念模型还不能直接建立数据仓库的物理模型。必须先建立逻辑模型，由逻辑模型来指导数据仓库的物理实施。在数据仓库逻辑模型的设计主要包括粒度层次的划分，关系模式的定义，数据源及数据抽取模型的确定等。而关系模式的确定与粒度层次的划分有关，关于粒度的大小则遵循在充分考虑数据仓库的分析能力的前提下，也要兼顾数据量的大小和查询分析效率。

数据源中的数据在数据的组织方式、数据格式等许多方面与数据仓库对数据的要求有很大差别，因此在进入数据仓库之前，必须进行数据的抽取与清理工作。

数据抽取包括对数据源的说明、数据抽取规则、数据源的列与数据仓库列的对应关系等，并不是所有的数据源中的数据都需要抽取到准备区，抽取的数据必须满足一定的条件。在很多情况下，需抽取的数据可能分散在不同的表中，这时还需要指定表的连接条件。抽取后的数据还不能直接加载到数据仓库中去，还需要对数据进行各种清理工作，包括格式转换、类型转换、统一单位，或将数据按照划分的粒度层次进行汇总、聚集等。经过抽取和清理的数据，才能从数据准备区加载到数据仓库中去。

3 数据存储模式的选择策略

由于存在 MOLAP 和 ROLAP 两种在线分析的处理技术，在应用 OLAP 时，必然面临选择哪种数据存储模式的问题。这里分别从查询性能、数据加载性能、空间占用、分析能力、维的管理以及维护能力等方面来分析这两种模式的特点，以帮助针对具体的应用，选择合适的数据存储模型。

（1）查询性能：由于 MOLAP 直接处理存放在多维数组总的数据，因此一般而言，MOLAP 的查询性能要优于 ROLAP，查询响应速度较快且较稳定。而 ROLAP 的查询响应速度这不够稳定，有时很快，有时这比较慢。

（2）数据加载性能：在数据加载的操作中，MOIAP 除要完成数据的装载外，还需要对所有立方体中的所有值进行计算。这样 MOIAP 所需要的数据加载时间就比较长。而对于 ROLAP 来说，在数据加载过程中所要完成的操作是数据加载、索引和概要表的创建。由于在 ROIAP 中所进行的概要表创建量一般较少，因此 ROIAP 的加载时间要比 MOIAP 的短。

（3）空间占用：如果所有的维成员组合都存在相应的度量值，则采用 MOLAP 时比较节省存储空间。但在实际应用中，许多维成员的组合不存在相应的度量值，从而形成稀疏矩阵，此时采用 MOLAP，就造成了空间的大量浪费。随着维数的增加，这种空间的浪费呈爆炸性的增长。

（4）分析能力：MOLAP 在分析过程中的精度较高，具有分析的优势；而 ROLAP 的分析结果往往由于 SQL 语言的约束，使 ROLAP 的分析效果往往不如 MOLAP。

究竟选择 MOLAP 还是 ROLAP 主要看应用的规模。如果要建立功能复杂、规模较大的企业级数据仓库，则一般选择 ROIAP 方式；而如果是建立功能单一、小型的数据集市则更宜采用 MOIAP 方式。

决策分析及应用

4 决策分析及应用

4.1 在线分析处理

建立数据仓库的目的是为了对数据仓库中的数据进行灵活多样的查询分析。数据仓库中数据的组织方式为进行这种查询分析提供了可能 ,但是仅仅依靠数据仓库本身 并不能完成这种复杂的数据查询分析。为了对数据仓库中数据进行多角度、多视图的查询，方便地获得概括性的或详细的信息，需要采用在线分析处理技术，用于辅 助决策。

在进行在线分析处理技术过程中，使用基于维表和事实表的多维数据模型，通过对井组的多维数据进行切片，切块、旋转、钻取等分析性处理，可以从多个角度、多 个侧面观察油田生产的各类数据（如气油比、含水量、日配注水量等），从而更加深入地了解包含在数据中的信息。如图 3 所示。

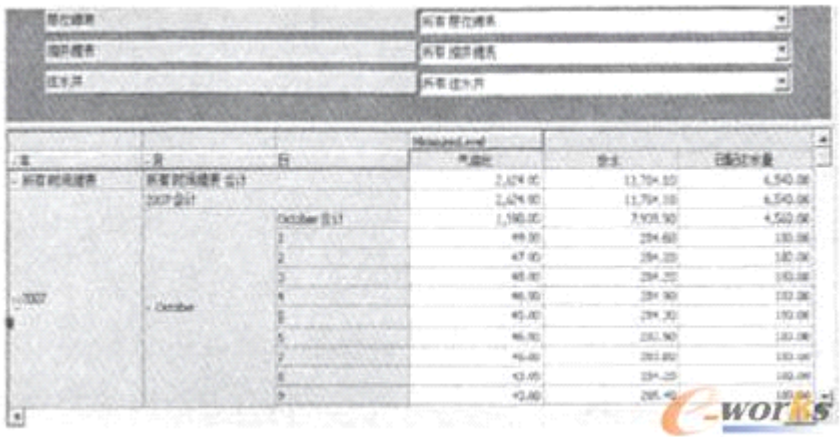


图 3 井组多维数据集在线分析

另外，往往有些有意义的生产参数在构建的多维数据集中是隐含的，可以通过在线分析技术以度量值或维度成员（统称计算成员）的形式创建这些参数。如井组多维 数据集中并没有每口油井的日产油量的数据，若想获取这类数据可以借助参数关系公式：日产油量 = 日产气量*（1/气油比）来进行创建该计算成员，如图 4 所 示。如此便可以使用计算成员将原始数据建模为有意义的业务指示符来增加分析的价值。

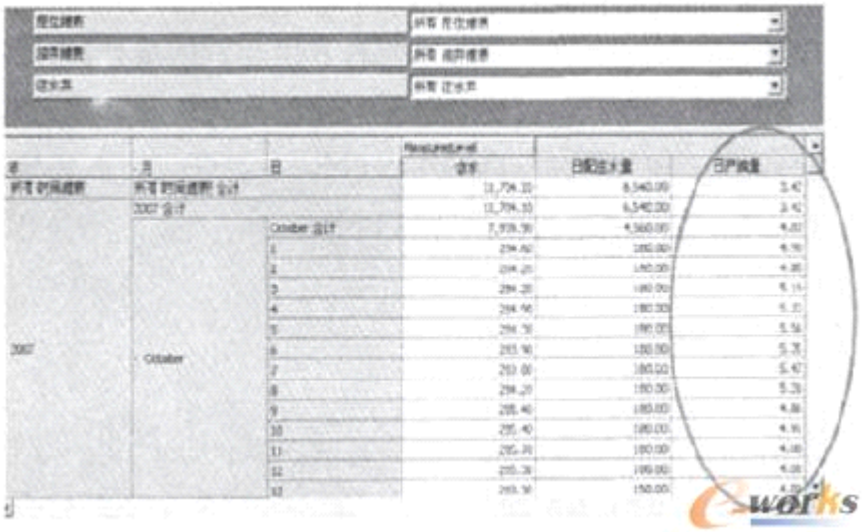


图 4 增加系统分析参数

综合运用上述方法，可以从不同角度、不同的层次观察分析数据，有助于获得有价值的信息，从

而起到辅助决策的作用。

4.2 数据挖掘及应用

数据挖掘采用基于人工智能来分析数据的技术，通过对数据仓库中数据的分析去发现隐含的模式和数据关系。有效构建高效的数据挖掘模型，是成功实施数据挖掘任 务的关键。主要建模方法包括：关联规则、决策树、粗糙集、统计分析、神经网络、支持向量机、聚类、贝叶斯预测等。而在实际建模过程中，需要结合具体问题对 多种建模方法进行综合比较和分析。因此，结合油田生产的实际情况，这里采用基于井组生产数据仓库的聚类分析算法来建立数据挖掘模型。

在 n 维空间中应用聚类数据挖掘时，采用明考斯基距离：

$$d(i,j) = (|xi1 - xj1|^q + |xi2 - xj2|^q + \ldots + |xip - xjp|^q)^{1/q}$$

其中 i =（xi1，xi2，…，xip），j =（xj1，xj2，…，xjp）是两个 p 维的数据对象，即数据库中有 p 个字段的第 i 条记录与第 j 条记录。在聚类分析中，有的生产参数数据值根据聚类需要给予较大的权重。此时加权明考斯基距离计算公式为：

$$d(i,j) = (w1|xi1 - xj1|^q + w2|xi2 - xj2|^q + \ldots + wp|xip - xjp|^q)^{1/q}$$

其中的 wp 为对应的 |xip - xjp| 权重，其值在 0.1 之间，但是所有的权重之和应为 1。

由于传统的聚类技术是无监督学习过程，因而易产生两种极端情况：一种情况是把数据库中的每一条记录看作一个类，这样当然达到了把记录分类的目的，但是却与 聚类技术是为了可以更清楚地理解数据库中的记录这个最终目的相违背。另一种极端情况是把所有的记录归入一个类，虽然实现了概括数据库内容的目的，但是不能 提供任何有用的信息。因此，这里采用分层聚类的方法实现，该技术的一个优点就是允许最终用户指定最后生成的类的数目。把分层聚类技术生成的目录结构建立成 树型结构，由此就可以决定合适的类的数目，既概括了数据库内容，同时又能提供有用的信息。并且这棵树的生成过程可以是从上到下分裂而成，也可以是从下往上 逐步合并而成。

由此，可获得油田井组生产决策系统的挖掘模型，如图 5 所示。

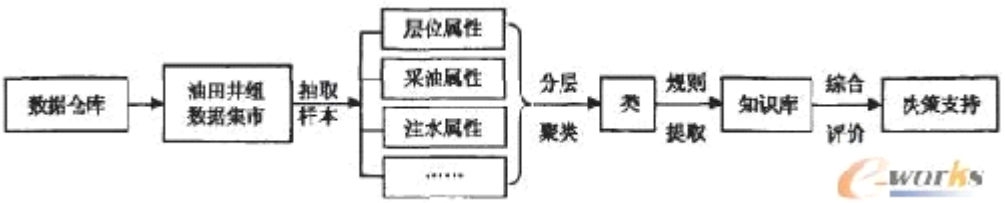


图 5 油田井组生产决策系统挖掘模型

系统的软件方案是利用 Analysis Services 构建油田生产数据仓库，利用 DTS（数据转换服务）把需要的数据（油井生产信息、注水井生产信息及层位信息等）从油田企业数据源（如 ERP 等）导入到油田生产数据仓库，进而针对油田生产多维数据模型开展联机分析和数据挖掘，以便识别各类井组的特征，根据井组的不同参数属性，为后续生产 方案的制定提供有力的决策支持。

这里使用聚类算法将研究对象的井组划分为 6 个类别。油井、注水井和层位是要调查的维度。然后选择想要在算法中表示各个井组类别特性的统计特征列表，然后训 练此模型，最终使其能够浏览受训练数据并从中分析六种井组类别。根据每种井组类别的统计属性，就可以选择调整合适的井组生产参数。

经过挖掘分析发现，随着层位、采油方式以及注水方式等生产参数的不同，对于油田生产关键指标参数（如日产量、含水量以及日配注水量等）的影响可以获得定量 的认识，并且还可以分别进行单参数和多参数的分析评价，这对于油田生产调整和优化具有重要的指导意义，并为实现油田的智能化决策提供了可靠的技术支持。

5 总 结

数据挖掘是实现油田生产智能化决策的现代化油藏管理必不可少的技术。在建立油田企业生产数据仓库的基础上，采用多种分析挖掘策略并实施多主题的数据挖掘是比较有意义的，可以为油田企业决策分析提供强有力的技术支持，并进一步提高油田的市场竞争力。

::ttnn::

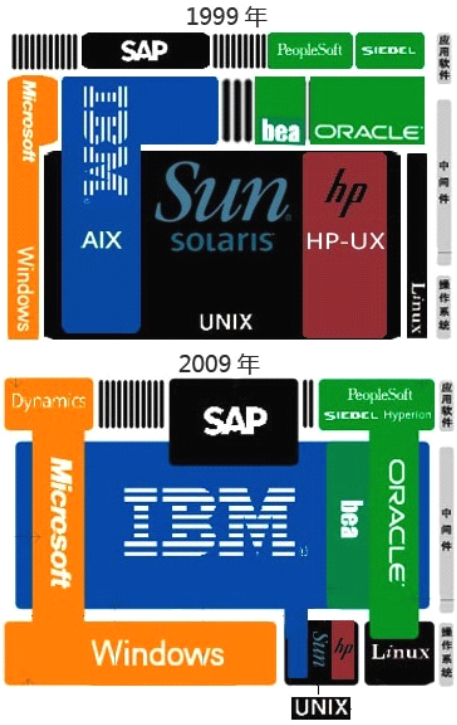
『业界观察』

十年动荡

作者：Qing 20090522

看图说话罗，弹指 10 年，IT 格局变化。
谁知道这幅图的原始出处？

http://docs.google.com/File?id=dgkpsn4m_153d4xndsf4_b



::ttnn::

云计算，一种市场

作者：Qing　20090605

移动的研究院最近号称要搞云计算，名字叫大云，bigcloud。移动研究院的这个 bigcloud，看上去暂时还没什么特别之处，底层的平台用一些现成软件，比如 hadoop。而 bigcloud 的几个重要构件的名称，看上去也非常眼熟，HyperDFS，不由让人想起 GFS、HDFS，HugTable 不由联想到 BigTable。所以，看起来移动现在研究的主要是云计算的应用模式，而非基础设施的技术。

关于什么是云计算，我相信就像对 BI 这个概念的众说纷纭一样，会有很多种理解。虽然，有机构对此作出定义，但未必是权威的，最后的权威是市场说了算。如果 google 的模式用的人多，那么他就是云计算的标准，如果 amazon 做得好，就是他的定义。对于那些流行的词汇，云、网格、SaaS（甚至是 DaaS 等任何 XaaS）似乎都有些重叠之处。我觉得他们的共同之处就在于将一种能力变成服务，而且要将他们标准化。云，集中式地提供数据存储、计算能力的服务，网格，分布式地提供计算能力的服务，SaaS，将软件当作服务。

能力，是一种你有什么本事，属于内在的东西。
服务，是一种你可以为别人做什么，属于外在的东西。

一个性能力强的人，如果仅仅是为了肉体欢愉去做爱，没有创造社会价值。如果他将这种能力变成一种服务，那就创造和社会价值。为国家 gdp 作出贡献了。将能力服务化，事情变得很简单了。你说你可以提供服务，别人可以直接问多少钱，很简洁。

云计算大概也就是这种模式的体现吧，亚马逊当初是卖书的，但是他的底层 IT 技术能力很强啊，于是也把这能力服务化了。按使用收费，数据没 G 多少钱，耗用 CPU 多少算多少钱，明码标价，童叟无欺。

其实想想，这种将能力变成服务的模式很早就有的。比如电子邮件，以前企业要搞自己的邮件服务器，后来可以用 web 邮箱，那是很早以前的事情了吧，那时候可没人提 SaaS、云之类的。但我想将能力能够转换成一种简洁的服务模式并且量化价格恐怕也是不容易的事情，不然这事儿早就泛滥开了。比如云计算提供计算服务，可是你怎么衡量这种含含糊糊的能力呢？但后来终于有了衡量的方法。所以，从另一个角度将，云计算其实提供了一种市场的功能，将能力变成服务，使服务可以交易。

如今 BI 也大肆地 SaaS，主要是应用层面，报表、仪表盘等绩效管理方面的应用很多，当然，随着 amazon、salesforce 服务的兴起，也开始有数据整合、数据分析之类的服务。估计在 salesforce 里面还会有其他的一些行业应用模型服务，比如客户的生命周期管理啊，离网预警之类的。

由此可见云计算跟 ttnn 研究院的“决策分析市场”项目也有些关联，后者也是希望将分析能力变成分析服务，建立一种市场的模式，只不过，这个市场买的是跟分析相关的服务。现在还没有想好分析服务到底是个什么样子，如果能够对“分析”建立一种消费者可以接受的衡量方式（比如可以有规格定义，有价格，有效果），那这个市场算是建成了。

作者：Magic.jiangmj

企业用户采用云计算模式的六大优势：
1.降低成本
云技术是逐渐支付的，从而为企业节省了开支。
2.增加存储能力
使用云技术，企业可以比使用私有服务器存储更多的数据。
3.高度自动化
IT 人员不必担心软件软件最新问题。
4.灵活性高
云计算要比传统的计算模式提供更多的灵活性。
5.更高的流动性
员工可以随时随地获取信息，而不是像过去那样必须坐在办公室的电脑前。
6.解放 IT 部门
IT 部门再也不用担心不断服务器升级和其它计算问题，从而可以将更多的精力投入创新中。

缺点如下：
计算机用户使用模式的改变：如果一想到把自己的数据存储在远地、应用程序不安装在自己的计算机上，就觉得失去了太多的控制权，那么云计算可能根本吊不起用户的胃口。

安全问题：在云计算领域，数据安全性方面还存在让人担心的一些重大问题。大批的数据存储在异地的大型“数据中心”，这对黑客及认为掌握了信息如同掌握了权力的其他人来说是个诱人的目标。公司需要确信自己的数据真正得到了安全保护。
可靠性仍是个问题：如果网络瘫痪，或者如果接入网络的那条连接瘫痪，除非问题得到了解决，否则你就无法访问自己的数据。另外，如果存放你那些应用程序和数据的数据中心出现了故障或关闭，除非问题得到了解决，否则也存在数据和程序丢失的风险。

能耗问题：云计算从用户的角度来看更加环保，但所有那些数据和应用程序都必须存储在某个地方的机器上；但这些机器处于随时运行状态，因而能耗相当大。

::ttnn::

『数据仓库』

CDBMS

作者： Qing 20090611

昨天编写 BI 简报的时候，有一则新闻是 illumnate 发布其新版本分析数据库。这种数据库叫做 Correlation DBMS，在五月份的时候，ttnn 也有人曾经问起过这个东东，当时没注意。我一时拿不准应该如何翻译这个词，关联数据库？相关数据库？后来选用了后者。

从网上找了些资料看，发现，这种 CDBMS 的实现几乎只有 illumnate 一家。这是一家西班牙的小厂商。

如今的数据库产品已经不再是关系数据库一统天下的局面，因为分析应用越来越受到重视，关系数据库、列式数据库、相关数据库，甚至还有键值数据库，层出不穷。CDBMS 的主要特点在于底层数据存储，而上层依然可以用 SQL、ODBC 来访问。

关系数据库是按照记录存储数据的，不同类型的值放在一起；
列式数据库是按照每列数据存储数据。一列统一类型的值放在一起；
CDBMS 呢，是按照值存储的，每个不同的值只出现一次，而所有对这个值，记录引用的索引；比如有个客户的姓名叫 ABC，而还有一个公司名字也叫 ABC，那么在 CDBMS 里面，只存有一个 ABC 这个值，但在索引里面记录了有两个地方引用它。显然，在索引里面，必定还要区分”客户“和”公司“不同的实体，由此可见，这种数据库的元数据设计得要相对丰富些。关系数据库的物理和逻辑视图是能够匹配的，但这种数据库，逻辑上是关系实体，物理上不是，得用元数据做映射。

这种数据库是专门为分析而设计的。因为不存储冗余数据，所以他对于海量数据，非常节省空间，如果说这个有点不太吸引人的话，另一个据称的优点就是做那种 adhoc 查询，是非常快的。不过这点我还没想到是为什么，留待大家求证。还有一个优点，就是扩展性，因为数据库结构只是中间层元数据定义的，物理结构是恒定的，基于集合的。所以，改变表结构这类事情对他就很 eazy 地说。（但说是这么说，如果元数据结构太复杂，本身就是一种阻碍。灵活性跟易用性很难兼得。）

推荐阅读：
http://en.wikipedia.org/wiki/Correlation_database

作者：syfins

跟 teradata 的类似设计嘛
adhoc 的大量内连接查询，对于几个表相同的 value 其实是对物理上同一个存储 item 的引用，那么对于内连接的大量 I/O 吞吐就可以避免了
Teradata 的优势也差不多，根据 hash 分布，同一个 value 存储在一个 node 上，能不快啊

::ttnn::

GDBMS

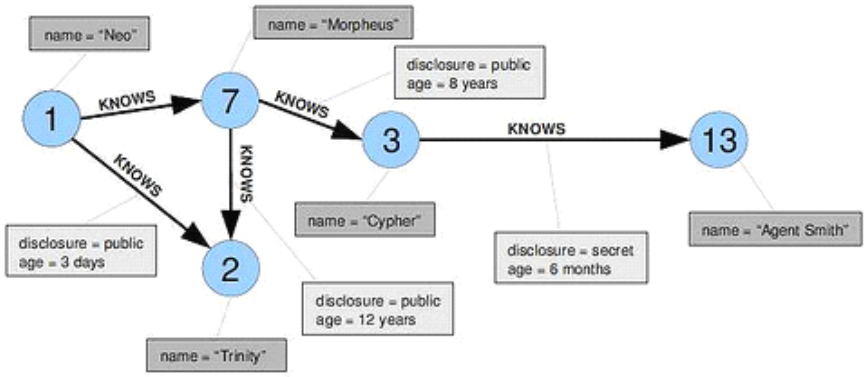
作者：Qing 20090618

前端时间介绍了一种 CDBMS 的相关数据库技术，按照数据库值唯一存储。关系数据库称霸江湖几十年，现在似乎遇到越来越大的挑战，对数据的需求越来越丰富，传统 RDBMS 最擅长的关系实体的事务处理，在如今的数据环境里面有些吃力。比如在俺们数据分析领域，要不得为传统关系数据库增强 MPP 能力，而另外还寻求新的数据存储方法，以迎合数据分析独特的查询需求，列式存储、相关存储、键值存储等等。

现在的数据环境里面，结构化和非结构化数据共存，数据膨胀迅速，各种网络应用催生独特形态结构的数据，还有大趋势云计算的鼓动。为了存储非结构化数据，google 用自己的 BigTable，而网络应用方面，有人宣称 web3.0 时代是网络社交时代，表达这种社交网络，虽然说 ER 理论也可以，但并非很好适合。还记得以前在面向对象领域有个“阻抗不匹配”的著名说法吧，因为编程语言跟数据存储的理论不匹配，因此对于编程者总是要花费一些精力在这两种理论，OO 和 ER 之间转换。所以，在类库里面出现了持久层，用来自动映射关系表和对象，这种做法似乎不够彻底；而另外一种做法，就是专门设计一种面向对象数据库，试图彻底改变这种不匹配，让数据库理论跟编程理论一致不就 OK 了。但从实际效果看，用的多的还是持久层，面向对象数据库并没有多少市场。这跟关系数据库众多的开发者有关，也跟那些面向对象数据库的不成熟有关吧。但你看现在 google 的 app engine，并没有使用关系数据库，开发者可以不用考虑后台的存储，提供了一种类似面向对象数据库的能力（但应该不是）。

网络社交现在很火热，facebook、twitter、qq、msn，甚至是普通的电信通讯、邮件，都是一种社交网络。传统我们都是用关系实体来存储这种数据的，我跟你打电话，“我”是一个“用户”的实体，“你”是另一个“用户”的实体，我们之间存在了“通话”的关系；“你”还可能跟“她”发生了关系…显然，这是一种很独特的数据结构，而且对这种数据的分析需求也越来越多，人们要从中找到人与人之间的关系、圈子，是不是一个家庭的，是不是一个公司的，是不是情侣关系。而且还要去发现一个人的重要程度。用普通关系数据库，干这事儿吃力。

于是乎，现在又专门出来一种 graph database，图数据库。看下图：



在这种数据库里面，数据按照节点、关系和属性键值存储。现在有一个开源产品，Neo4j，基本上这也是一种键值数据库，也就是说其最底层数据存储都是按照 key-value 存放的，一般这种存储方式是比较适合并行处理的。而graph database 的特点应该就似乎在这一层上增添了 graph 的特点，内置一些常见的 graph 算法。不知道是否有更上层的 SNS 分析功能。这种产品已经有点靠近应用层面了。

这也许是如今技术发展的一种趋势，但出现一种热门应用，就会出现一种后台技术变革的一条龙服务。

作者：hawking_bin

这个 neo4j 的 graph 似乎不是人际关系图这个 graph。前者是模型级的，后其是实例级的，不是同一范畴。其实在关系数据库模型之前就有网络数据库模型（还有层次数据库模型，都成了恐龙化石了）。这个 neo4j 是在复古吧。不过随着 SSD 单价的下降，数据库技术可能会有很大的变化，新环境又变得
适合恐龙了也说不定。

再说社会关系分析，分析两个人之间的关系还是小事，当进行递推时就要命了。比如 A 猪把流感

传给 B 猪 C 猪，B 猪又传给 D 猪 E 猪...这不是一般的索引能优化的。

::ttnn::

『分析 模型』

RFM 在电信的应用

作者： Qing 20090521

以前在坛子里面讨论过 RFM 的东西，我没有太多关注，因为总觉得这是一个零售行业的市场分析模型。最近看到一篇《RFM，挖掘先驱》的文章，才发现原来这玩意儿并非局限于零售行业。所谓零售，是广义的，按照现在的说法，B2C 都算是零售，所以，电信行业的市场营销也可以用到此模型。

先来介绍一下这个模型，RFM，分别代表三个首写字母：

- Recency：最近购买
- Frequency：购买频率
- Monetary：消费金额

这三个特性可以用来衡量客户对未来营销活动的是否响应，按照这个模型——越是最近购买的，越是频繁购买的，越是花费大的，越可能响应你的营销活动。

这个模型已经用了 40 多年，可以看到，非常简洁，很强大。简单几句话就可以概括。如果我们通常的挖掘模型能够做到这一步，那就牛逼了。以前项目里面几十个模型，没有哪个模型能够抽象成几句话，几个变量就能表达的。当然，这可能也跟 RFM 是个通用模型有关系吧。就算对于一些很细的领域，比如电信客户流失，那个著作等身的专家 rob，也没有抽象出这等模型。我估计，40 年前，发明这个模型的人没有用挖掘技术，最多也就是统计技术。而且还有一件事很奇怪，在电信行业似乎很少有用这个去做营销分析的。难道也是因为认为这并非电信行业的模型才不用的？

可以用的。在电信营销里面，同样需要考察客户对营销的响应程度，每次做营销，当然都希望接触到的客户尽量都能够响应。所谓响应，就是有明确的购买倾向。但电信行业的营销似乎对消费金额，也就是 M 最感兴趣。对 R 和 F 的考虑其实并不看重。以前遇到过一个接触响应模型，其实跟这个 RFM 有同样的目的，分析客户的响应程度如何。那个模型的基本思路是跟天气预报学的——最近响应良好的，下一次的响应也会良好。因此，可以说，这个模型考虑的是 R，而 F 和 M，

没有考虑。没有考虑 F，恐怕跟电信的业务也有关系，到底什么算一次真正的购买？办理一项收费新业务当然算？但办理一项免费的新业务（常有的事）算不算？充值算不算？

不过这也不算什么大问题，总归是可以产生一种定义的。而接下来，就可以综合应用 RFM 来进行营销响应的预测了。

在上面提到的文章中，包括实际的操作方法，挺好的——如何排序，如何选取目标客户数，如何选取分组大小。三个变量，但最后总归只能有一个排序标准。

但结合实际情况，发现这个模型也存在新问题，这个模型是在帕累托原理下，但如今，在长尾理论的影响下，产生很多低成本的精确营销手段，RFM 能够定位20%的高响应客户，但也总不能放弃另外的 80%吧，而对于后者，如果能在内容偏好、时机偏好上有些考虑的话，也许同样有很好的响应率。

作者：笨笨

我之前接触过的几个电信行业客户都有在关注这个模型阿

我觉得该模型最大的价值在于精简出重要的三个因素，或者说三个维度。其实也就是一个精简了的聚类模型。当然，它采用了线性划分的方式，近似于 Cube，
而非采用非线性划分。

另外关键的一点是，不少分类预测模型可以将 RFM 作为预测因子，引入模型之中，这为建模人员变量的选择扩大了范围。

作者：QiCici

记得 2000 年看 Rob 的一本有关电信营销策略的书（英文版的，没见过中文版的），上面有专门介绍 RFM 模型的。

相信 Rob 没有抽象出这等模型，并非是他的不能，我想，还是在于 RFM 确实可能更适用于快速消费品行业的消费品购买分析，而非电信行业电信产品的购买分析。（Rob 是我的偶像，很多年前，他那本电信数据仓库的书将我引进数据仓库和数据挖掘的门。）

我认为，RFM 的适用基础应该是产品消费具有一定的频率，根据频率能够推导出规律。

电信产品不具备这个特点，无论从每次打电话，使用电信产品的频率来说，还是从营销活动来说。

最根本的区别就在于，耐用消费品的使用是专门针对人的某一具体需求而言，比如牙膏，不管品牌，每次的用量、每天的用量、以至每月每年的用量都是一定的，因为其需求的单一和产品使用的可重复性；电话的使用却没这么规律，试想，或许你每周的每个固定时间会打电话给外地的父母，但其他更多的电话呢，它具备这种规律吗？不具备规律的话，你能够找得到规律，并应用到你的产品分析上吗？

营销活动也一样，每次营销活动都不一样，满足客户的需求不一样，响应的人群也都不一致。咱们没办法从每次营销活动的响应情况，因此推论下一次营销活动的响应情况，除非你从消费心理的角度去分析营销响应之间的关联性，然后从具备关联关系的某营销活动推论下一个营销活动的响应情况。

我们当年看到 Rob 介绍的这个模型的时候，也试图应用在电信产品的使用或营销活动的响应分析上，但是未果。正是因为电信产品和营销活动的上述特点。

真的非常希望对这个问题有进一步的探讨，没准真有应用的可能也不一定，哈！

作者： QiCiCi

应该是快速消费品，晕了！

刚才突然想到，对那种特有规律的电信产品，比如大家都要定期给父母或亲友打的电话等，没准倒是可以用上 RFM 来作相关的分析。

作者： Qing

也许可以对电信行业的购买行为做个分类，对于某些类别的购买行为可用 rfm 来做响应分析。

购买充值卡充值；
购买新业务；
开通新功能；
购买新终端；
办理新套餐；
。。。

发现电信和普通零售的一个区别，电信的消费是有协议在先的，消费者不容易转到竞争对手那里。比如定期给父母亲友打电话，这种行为如果具备规律性的话，电信运营商大可不必操心用户是否会持续这种行为。

但运营商确实也需要预测营销活动的响应度，按照 rfm 的理论。那些最近有购买记录的（比如以上几种），并且周期内次数多，金额大的，更容易接受营销活动。也许这也是成立的，但如果要更加精确的，恐怕还得结合具体营销内容，在零售业应该也是这样吧。比如要销售手机证券业务，思路是先找到所有可能对手机证券感兴趣的，然后用 rfm 来预测他们的响应级别，侧重将营销资源向高级别用户倾斜。这样说说的通，但需要实践论证。

作者：hunterdong

我的理解，快速消费品的个体消费者并非是特别适合 RFM 的，顶多是胃口大点和小点的区别，最牛的顾客也就比普通人多用 5 倍 10 倍到头了吧。就这点来说觉得和电信产品有类似。

作者：QiCici

> 为什么说快速消费品的分析比较适合用 rfm，是因为其消费的规律性；为什么要找规律，是因为需要用实际购买的情况与其该有的规律相比较，如果不符合规律，这就发现经营问题或商业机会了。

rfm 的分析并非价值分析，所以不在于产品的贵贱。

为什么说如果电信产品呈现规律性或许可适用 rfm 的分析，也即这个道理。举个最简单的例子，比如你每周都给父母打电话报平安，可是发现这段时间这些电话不见了，为什么不见了？我们可以假设话务去对手那儿了，话务转到对手哪儿了，使大家都熟悉的什么问题呀？流失呀！

rfm 的分析当然不仅是为 rfm 而作，rfm 的分析其实是其他分析的基础。

说明白了没？哈！

::ttnn::

自动模型的设想

作者：Qing 20090602

我们现在面临的数据分析产物，要不是研究报告，要不一个分析模型，如果做的好的，可以从他

们演化出更具抽象的业务规则（知识）。愈抽象的业务规则适用面愈大，可用性强，而精准性可能稍稍下降。比如 rfm 模型。

研究报告、分析模型是产物，而撰写、研发，是过程。在应用中，更加重要的是产物。能不能将过程和产物结合起来呢？也就是将过陈自动化。昨天走在路上，我想到这点。后来想想，其实这并非一个新点子，以前也曾经设想过一种自我进化的分析模型。通常的分析模型，比如挖掘模型吧，按照一定的方法论，先理解业务、再整理修补数据、训练模型、评估模型。最后，模型完成了，运行的时候将制定的数据输入，可以完成诸如客户预测之类的目的。而在现实的一些项目中，对于这类模型，还需要一个重新训练的周期，比如每个季度重新训练一次，以确保模型的精度。

我想，是否可以将这个方法论优化一下，有的可以人工进行，有的可以机器自动进行。比如业务理解是需要人的智慧，整理数据需要人的智慧，但修补数据、清洗数据、建立模型、评估模型等等，这些机器智慧恐怕也是足够的。而且，业务理解和数据准备的工作的频繁性未必很高。

因此我们设想最后一种分析模型的应用场景，为了叙述方面，称之为“自动模型”。

自动模型的初始化工作，是在业务理解的环节，对模型目标作出准确定义，准备了数据（还没有进行清洗、修补），将这些输入到自动模型中。自动模型的输出，则是对目标预测。

自动模型的输入输出就是如此。而输入还可以配置、调整，比如对模型目标作出更加精确的量化定义。比如，增加了新的变量。将这些都扔进自动模型，甚至，如果可以的话，将人工的经验判断（比如若干专家的评分）也作为输入输进自动模型。所有这些，都导致更加精确的预测结果（当然，这是理想的，现实情况可能是更加不精确，那原因就多了）。

至于自动模型的内部运作，每次有新的数据输入，他都会进行数据清洗、修补，自动进行数据探索，发现关键变量，自动利用多种模型算法重新训练模型，并且根据评估结果来择优选取最终模型用以打分预测。整个过程是自动的。

虽然有人说，其实在建模工作中，准备数据的工作量最大，占去 80%，所以，这种自动化只是减少了其他 20%不到的工作量。不过也已经足够，至少这会让分析师将主要工作放在更重要的事情——业务理解、数据准备上面。而且这样的东西可以让不太懂挖掘的人可以做深入分析了，在建模这项工作上面，可以配备两个角色，一是业务顾问，一是数据处理员（可以算是助手），省去挖掘建模人员。业务顾问设立分析目标，作出分析假设和潜在的变量，以及模型结果出来以后的解读工作；而数据处理员根据目标、假设、变量去准备数据，然后，其他的事情交给自动模型解决。而现在我们看到的很多挖掘工程师，要不转向去搞业务分析，要不就去搞挖掘工具的开发，比如这里提到的自动模型，或者模型算法之类的。

作者： Luke Zhang

我看好这个想法，但是如何将其实现，从事实到概念，从数据到信息，再到知识、概念、模型，如何真的让其自动呢？那一步能够自动，哪一步不能。

作者： hawking bin

这个在技术上是毫无难度，但在政治上则阻力重重。当今身边的事就不方便说了，说历史吧。在工业革命初期，机器经常遭到工人破坏。因为工人以为机器的发明和采用，是他们失业和贫困的根源。这就叫做机器吃人。直到有伟大导师指出生产关系才是生产力提高的瓶颈，才终止了工人对机器的仇恨。

作者： Qing

哈哈，bin 说的好夸张，要说阻力，一般都是大的利益既得者是最大阻碍。贵族反对自由，超大企业阻碍创新，工人们作为弱势群体也只能是搞点小破坏而已，无伤大雅。当然，如果工人们能够捆在一起，那就是农村包围城市，他们也成为一股强大力量了，不过这也不容易。看看这个自动模型，受益的其实那些卖算法的，以前，他们是 B2B，销售的是一种工具，而做成自动模型，他们是 B2C，开始销售一种零售商品了，更有利益。看上去暂时吃亏的，是现在这些搞模型开发的，自然会萎缩。可这种事情肯定也不会表现非常强烈，因为多少次的变革已经证明破坏工具阻碍不了大潮流，不然当计算机出现的时候，计算机也同样有被破坏的下场，不过看上去，这种破坏趋势逐步趋缓。

作者： hawking bin

计算机这个例子举得非常好。

计算机和网络具有两面性，一方面它们是提高效率的生产工具，另一方面它们又是破坏效率的娱乐设施。一般人买电脑上网往往堂皇的的号称是学习新工具新技术，但实际上是为了打游戏看水文泡网友练网游来打发时间。现在办公室里一般人手配一台电脑，还连网。但这电脑网络在作办公自动化工具的时间，也是效率的杀手----网页、聊天和无数小游戏（比如开心网）使员工心神经常游离于工作之外。而在业余时间电脑和网络作为娱乐设施的渗透更是深广。IT 对生产所带有的效益可能很大，但实在不是很好评估（我们做 BI 的效益评估时就很头疼）；但以它为基础所建立民用消费市场显而易见，深刻地改变了几乎所有人的生活方式，价值非常巨大。

因此可以说计算机的伟大在于：既提高了生产效率的时间，又创造了新的需求。这两面形成了一个良性循环。在这里面，我认为创造新需求是第一性，提高效率是第二性。如果在供需关系很稳

定的格局里，去做一项仅仅是提升效率的技术是有害的，至少是有短痛。对于卖机器和卖人力的公司，是不希望看到需求不变但效率提高的。因为效率越低，就需要越多的机器和人去折腾，赚的利润就越高。

长远来看当然不是这样，忍受了短痛之后能升华到更高得多的境界。但有些公司只看这一 Q 的事，和它说十年后或一百年后的事，它会觉得很荒谬。

::ttnn::

漫谈相关与回归

作者：刘飞燕 20090611

老师不断提醒我要对统计学的基本概念、定义及背景反复思考，这样才不会本末倒置，迷失方向。但是这个做起来很难，因为那些概念定义等看起来实在”太简单”、”没什么东西”，可能还是不能够平心静气吧！

最近静下来看了 David Freedman 等著的《统计学》的”相关与回归”部分，以及一篇关于直方图的文章，不免有些感慨！其实统计学中的很多概念、工具、方法等的实际意义或作用可能要比我们认为的要大很多，同时，当我们从一些概念定义等中发现出一些新东西时我们总会欣喜若狂。世界上的很多事物又何尝不是如此，人们对事物的了解总易受到传统或他人的影响仅仅停留在表面，很少达到全面而深刻，而一旦我们获得了那种深刻的洞察力，才发现真实世界是何等的精彩！一直以为直方图很简单，无非是一些代表频数的柱状图的组合而已，感觉没什么作用，但是看了一篇关于直方图制作方面的论文时，才认识到直方图的威力。直方图其实是非参数统计中估计总体分布特征的一项重要工具，选择好适当的组距和边界点（组距和最小边界点是关键），随着样本量的增大，它可以非常接近地反映数据的真实分布情况。其实，在统计中使用一种工具方法的目的也应该是使现有的数据尽可能多地反映出真实的信息，而这项工作往往是一个无底洞（这时又要考虑到效率问题了）。

散点图亦是如此。散点图给出了所有数据点的信息，但是如何从这些数据中获得结论或拟合模型，甚至用来预测？面对一张散点图，相关和回归应该是最容易想到的吧！这里主要谈谈两个变量间的相关和回归。

在研究两个变量的关系时，一般会先看看它们的散点图，在图中两变量的关系还是比较直观的，大致可以判断是否线性相关及相关性大小如何，是否是非线性相关等。而到底什么是相关呢？相关其实就是知道一件事对了解另一件事的帮助的大小。实际中，如果对某一事物不太了解，但是对与其有一定联系的另一事物有所了解，如果这种联系很强，那我们对于那件不了解的事物就有了更多的信息，或者说对这个不了解的事物有了更大的自信去预测。其实这也是研究中的一种常用的方法。

关于两个变量间的相关系数的计算。我们都知道两个变量 X 与 Y 的相关系数的计算公式为 $Cov(x,y)/(SD(x)*SD(y))$,然而这已经是一个结果性东西了 ,我更推崇 David Freedman 等著的《统计学》中计算方法：先分别对两个变量做标准化，比如对变量 X 做标准化 $(x_i-\bar{x})/SD(x)$, 然后对应的标准量相乘，最后加总再求平均. 这种求法反映到散点图中，相当于对散点图的坐标刻度标准化，从而使两个坐标轴具有了相同的刻度，同时在直观两个变量之间的相关性大小时不会受到各自的标准差大小的影响。这个新的坐标系把所有的点（数据对）分到了不同的象限，通过观察各个象限的点的个数和大致分布情况便可以对相关性的大小与正负有直观的了解，比如更多的点都分布在一、三象限且群集于一条直线周围，那么这两个变量的一般具有较强的正线性相关。

我们都知道相关系数是-1 到 1 之间的一个实数，那么相关系数为 0.8 是不是表示百分之八十的点群集在一条直线的周围吗？当然不是，相关系数是基于全体数据的一个综合信息，它反映的是所有点与某一条直线的群集程度，而不是一部分的点。由此也不免想到，我们在用到一些概念或定义时，也必须清楚这个概念或定义是基于怎样的对象，或有哪些局限条件或假定，比如概率论中的”事件”， “事件”是基于特定条件的，在具体使用过程中大家对这些特定条件都太”熟悉”以至于很少关注到它们，所以一旦条件改变了，大部分的人认为还是指同一个东西。

相关与因果也是这样，我们都知道相关不能同因果划等号，但实际过程中人们总”自然而然”地得出一些结论。一般来说，体重和升高相关，那体重大是不是就因为身高高呢？除非有一个固定的身材标准，而所有人都是这个标准。（随便提个问题：如果两个变量严格线性相关，即相关系数为 1，那是否可以说这两个变量中一个为因一个为果呢？)。其实两个变量的相关更经常的情况是它们同时受到另外的一个或多个因素的影响 ,在这里可以通过对照试验或观察研究来进一步研究。另外，相关是可逆的，而因果则不可以。所以我们分析相关时总是如此谨慎地说，某某变化，与此相关的某某”相应地”如何变化。研究发现，个人收入与教育水平相关，高教育水平是不是高收入的原因呢？实际情况是它们相互影响：教育水平高的人收入一般较高，收入高的一般也更有能力获得继续教育的机会。虽然相关不是因果，然而有时我们并不需要弄清所有的因果关系，盯住输入和输出，只要存在相关，即使不是因果关系也不妨碍人们利用这种关系来进行推断。比如

利用公鸡打鸣来预报太阳升起，虽然公鸡打鸣绝对不是日出的原因(虽然打鸣发生在先)。
在对两变量的相关关系有一定了解后，接下来的自然想法便是拟合回归模型。”回归”这一词来自于高尔顿的父子两代身高的研究，身高较高的父亲其儿子的平均身高要比父亲矮些，身高较矮的父亲其儿子的平均身高要比父亲高些，用高尔顿的话说就是”回归到平常”。虽然现在统计学上的”回归”这一概念已经远远超出的当时的定义，但是回归的原始思想依然有着非常重要的作用。”回归”，个人认为其实就是向中心的回归。在知道某地区 18-24 岁男子的身高的大致情况时，如果没有其他信息，让我们估计该地区中某一特定区域 18-24 岁男子的平均身高时（当然不是侏儒或篮球运动员之类的人），自然是用平均数（包括中位数）去估计了，这便是回归，没有其它的辅助信息时我们总倾向于平均值，这当然是符合统计思想的。两个变量的相关系数绝对值为 1 时，那么知道一变量的值就立即知道了另一变量的值；相关系数为 0 时，那么知道一变量的值对预测另一变量没有任何意义，那么我们就估计其值为平均值；相关系数绝对值介于 0 与 1 之间时，相关程度越大，我们越不倾向于取平均值。其实回归模型也是基于平均意义的，让我们来看看回归的本质（暂以两个变量 x 和 y 为例），回归是对每一个 x 值的 y 的平均值的估计，所以用回归模型来预测或估计总是平均意义的（这也是回归的思想），而针对某个特别的个体的预测则就需要非常的慎重了。
有这样一个例子，某学前班在儿童入学和结业时均要做智商测验，结果发现前后两次测验的分数平均都接近于 100 分，标准差为 15 分。但是仔细观察发现入学分数低于平均值的儿童结业时分数平均提高了 5 分，相反入学分数高于平均值的儿童结业时分数平均降低了 5 分，难道学前班会使儿童的智商平均化？其实没那么夸张，这只是回归效应的一个表现，只要两次测验分数的散点图中所有点不在同一条直线（这条直线的斜率为 1）上，那么就会存在回归效应。观察得到的数据并不是真实值，都有或大或小的、或正或负误差，在大多数对称的概率分布中，观察值大于平均值的往往是其真实值加上了一个正的机会误差，观察值小于平均值的往往是其真实值加上了一个负的机会误差。所以在那个学前班中，入学分数较平均分低的儿童其真实分数一般是大于观察值的，因此在结业时的分数一般是要比入学时高，因为在向观察值的平均值，即真实值回归。

相关与回归是一定范围内的相关与回归，超出范围没有任何意义（经常实践的人应该会很少犯此类毛病的吧)。回归其实并不能增加信息量，它是一种结论（结论的准确性还有待评价），或对数据以某一种方式的总结，超出范围的估计预测是没有任何意义的。收入与教育水平有关，无休止的教育显然不会带来收入的持续的增加 ,所以人为地改变一个变量 ,希望通过回归模型的”魔力”来改变另一个变量是很荒谬的。另外，变量也是有范围或区域限制的，因此在使用回归模型做预测时是要非常谨慎的。

现在研究的回归往往都是多元回归，往往比较复杂，其实这是符合实际情况的，因此往往要用多个变量作为因子来拟合，但是这些变量是不是考察某一方面的较好指标呢，比如收入与教育水平有关，还可能与父母的社会地位有关，那这个”父母的社会地位”这一因子又该如何度量呢？这又是一个问题，尽管多元回归是一种非常有用的技术，但是永远代替不了对数据间内在关系的了

解。由此可见实践经验的重要性!

相关文章

- 使用回归分析，样本过少时不妨好先作图看看
- 不同版本的散点图矩阵
- 用局部加权回归散点平滑法观察二维变量之间的关系

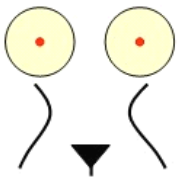
作者：Qing

简单的统计观察有大作用。

相关和因果虽然有区别，

但可以用来忽悠不假思索的人们。

直方图



三点图

::ttnn::

『工具 应用』

半精确营销

作者：Qing 20090608

当营销理论从 4P 走向 4C 的时候，其关键就是从面向产品的营销到面向客户的营销。名称不同，可基本思想不同，就差了那么一点点。而平时在运用这些理论的时候，也经常差了么一点点，而偏离了那些基本思想。

4P，产品，定价，促销和地点，他说，要做好市场营销，就要提供好的产品，合理定价，在适当的地点进行促销。而 4C 分别是客户、成本、便利和沟通，他说，要做好市场营销，就要对目标客户，以合理的成本提供便利的夫妇并进行深入沟通。你说谁说的有道理，也许在不同的环境里面有不同的适用性。上次谁说了一句，任何理论都有其适用范围，而忽视它的适用范围去谈论它，基本是扯淡。

对于这些理论，我没有研究太深入，毕竟不是搞市场营销的。但BI 有个很重要的作用就是支撑市场营销，所谓精确营销。所以，当去思考一个完整的体系的时候，发现其实确实是需要这些理论支撑的。我不大喜欢照搬现有理论，如果没有理解，宁愿构造一种新的，可能存在很大缺陷的理论，反正我想着，这些缺陷如果被发现了，也算是完善了。也许完善到后来，才突然发现它跟那些成型理论没什么两样的时候，也知足了。但这个过程是充满思考并完善的。

创造新理论或者新体系最直接的方法，就是给它起一个名字，然后去解释它。比如用一些关键概念的首写字母，去拼凑成一个有趣的名字。之前，我曾炮制过一个 PACE 模型，一个 FACE 体系，都是类似的技巧。最近在琢磨精确营销体系，也想起个有趣的名字，但想不到什么好的，于是放弃了，先给了一个普通的名字。

关键是能不能将其中的概念说圆，其实对于 4P、4C，难道大家不觉得很凑巧吗？为什么都是 4？为什么 4 个字母都是一样的？我想着，精确营销，应该是以客户为中心（为什么以客户为中心？恐怕多半是因为这是一个非常流行的趋势，而非真的认识到必须如此不可。），然后关注客户喜好什么产品，可以通过什么渠道接触它，可以用什么样的促销行动来刺激他，可以在什么时机刺激他。最终的目的，无非是让他响应我的营销。

很长一段时间里面，我不知道面临的问题究竟是什么，也许到现在还是不清楚的。跟现有的理论结合考虑，比如 4C、4P，总觉得有些差异。后来觉得，精确营销面临的问题，或者说我这里思考的“精确营销”面临的问题跟 4C、CP 面临的市场营销问题不同。后者是从广义的市场营销入手，从产品策划到营销策划的环节，是一个较长的周期过程。而这里的精确营销，可能仅仅是在于短期的接触到客户并营销。而之前，产品设计、定价都没有包含在精确营销里面。

这是一种妥协，因为现实的营销通常并非是以客户为中心的，往往要达到一个非客户的目的。比如要提升某种产品的渗透率，这是从产品出发，哪怕这个产品非常烂，也得硬着头皮上。还有可能是要提升电子渠道的活跃度，这是从渠道出发的目的。管理思想上面没有到精确营销的地步，所以现实也没到精确营销的地步，很多时候只能说是半精确营销。所以，后来我发现，我要解决的是如何进行半精确营销的问题。

客户、产品、渠道、促销、时机，要搞好半精确营销，得从这五个方面入手，无论你是以什么样的目的，是推荐产品，还是提升渠道，还是一次具体促销，都可以找到合适的目标客户，对他们营销，可以达到具备高响应率的结果。

“为什么要‘时机’？他看起来不太一样。时机在系统上不好实现。”

当我跟同事探讨的时候，他提出这个问题。我也觉得有些不太一样，客户、产品、渠道、促销都算是有实体的，客户可以跟产品匹配，渠道可以跟产品匹配，而时机，看起来却不太想实体，是一种随机的，出现在某个地点，发生了某个事件，在什么时间点。不可能枚举所有的可能，因此，他不像实体，而是依附某个实体的属性一般。但我没有立即去掉它，因为如果不做刚才这番形而上的分析的话，从业务逻辑上是能够说的通的。所以我说，“没关系，我想个可以说圆乎的理由，反正要说服人不在乎理论是否完整，而是在于说者的态度。”

作者：seamyhometown

客户，网购的越来越普及使客户的个性化信息明晰和容易收集；
成本，全球化的生产模式使得产品的成本日益透明和可比；
便利，物流业的快速发展使得产品的投放日益快捷；
沟通，多样化的沟通方式和客户化营销使得沟通可以更深入；
4C 是大势所趋 ,但很多传统企业仍停留在 4P 阶段 ,用 BI 的思想和方法论对 4C 进行整合和创新，相信前景广阔！

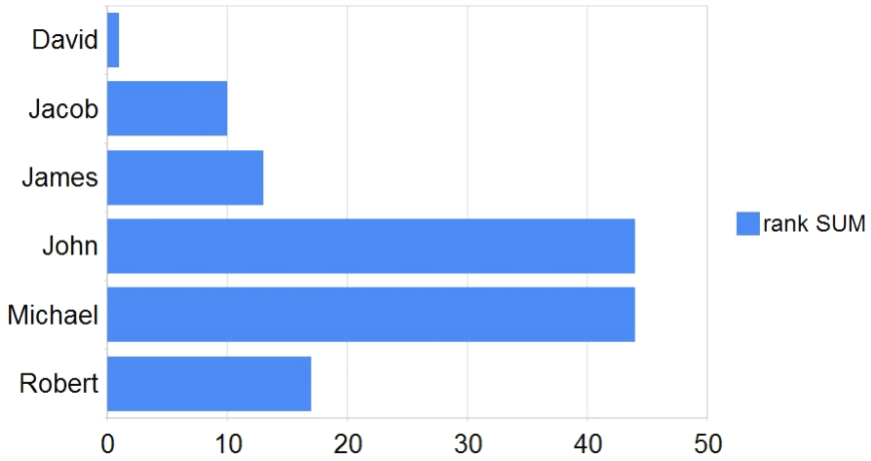
::ttnn::

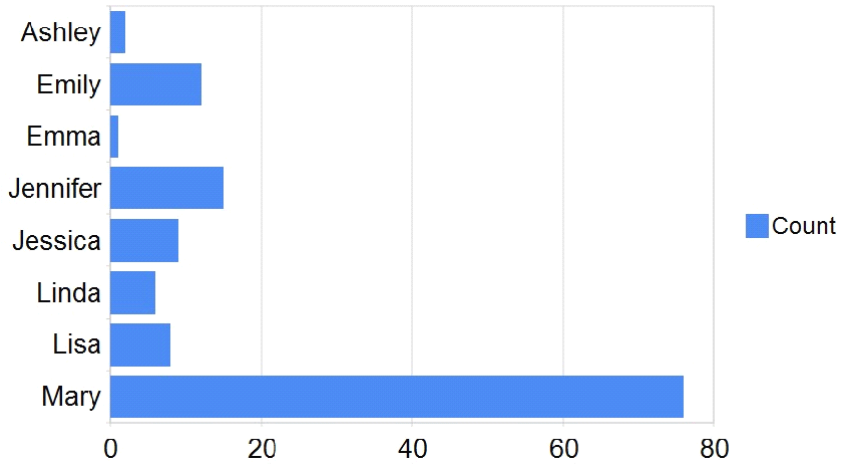
google 的数据表共享服务

作者：Qing 20090617

上周 google 推出一个 fusion tables 的实验室产品，只要用 google 帐户登录就可以使用，他可以让用户每次上传 100M 大小的数据文件；可以跟邀请其他人分享此数据文件，比如让其他人对这些数据进行评论，这种评论可以到行级的；可以对数据进行可视化查看。我试了一下，可视化功能还不是非常强，有通常关系数据上的处理，如条件过滤、聚集。但我想 google 恐怕会逐步增强这种能力的。

本身，fusion tables 已经提供了公共数据集，比如最近 120 年每年出生 BB 的姓名排行榜。按照男性和女性分别看历年来排名第一的，如下图：





当然，这些数据应该都是米国的。男性，约翰和迈克尔双塔奇兵，女性，玛丽一枝独秀。这些图可以生成 html 代码嵌入到其他网页里面去。

看起来这玩意跟 google 的 spreadsheet 在功能上很接近，都能处理表格数据，都能可视化啊，好像区别仅仅在于文件的大小。不过目前 fusion tables 还仅仅是实验室产品，测试阶段。但他们应该是有明显的定位不同，spreadsheet 重点在于处理表格文档，而 fusion table 重点将是大型数据分析。

我将 ttnn group 的成员列表导出成 csv 文件，然后上传到 fusion tables 上（出于隐私考虑，删除了成员姓名和邮件字段），如果大家有兴趣，可以登录去玩玩。
http://tables.googlelabs.com/DataSource?dsrcid=25059/25059

基于这个东西，可以实现在线分析、web2.0 式的分析应用，也同样是云计算的思路。有点意思。大家也可以设想一下，可以在这上头做点啥数据增值的东西。

我想，如果能够增加一下自动数据探索的功能，应该是比较好的。目前的功能，excel 也能干，并且能干的更好。现实当中，如果他的功能有限，恐怕他只能充当一个数据交换的工具，权当可以交换 100M 的数据，其他人还是将这些数据下载下来放到 excel 或者数据分析工具里面去。所以，下一步的数据分析功能可能也是 google 要考虑的，按照他们一贯的做法，可能会公布一些插件 API，让网络好事者编写一些 gadget，针对一个数据文件，只需要点击一个按钮，就可以生成一个数据分析报告。这个思路在当初一个 ttnn 研究院的项目提案中也提到过。甚至可以再进一步，会有一些数据挖掘的插件。虽然比不上专业分析工具那般强大，但应付一些常见的中型数据量分析已经绰绰有余。关键是，简单。这会吸引很多用户，毕竟希望从数据里面找点什么的人

很多，而懂得那些分析工具的人又很少。

::ttnn::