

基于 Web 的市场营销数据挖掘*

夏火松 蔡淑琴

(武汉科技学院经济管理学院 武汉 430073)

摘 要 根据知识获取的标准探讨数据挖掘方法在市场营销知识获取中具有的特点,分析了 Web 上的数据挖掘系统的结构和实现方法、数据挖掘的经济性,并构造了灰色信息收集的成本模型。

关键词 Web 数据挖掘 市场营销 知识获取 粗集 灰色信息

1 引言

随着计算机技术的高速发展,特别是 Internet 技术的不断应用,Intranet、Extranet 已成为企业构建信息系统的网络模式。知识经济环境下使知识的含义已被赋予了新的意义,网络上具有丰实的信息,我们怎样才能对其进行分析、推理,发现数据间的关系,提取有用的特征,找出有效的、新颖的、有潜在用处的、易于理解的关系和模型;怎样才能利用一定的方法从数据中挖掘出复杂的模型,发现的知识能够为人所理解,能够作为先验知识被再利用,能够较少或不依赖于外部专家的主观知识;能够由于目标数据中存在数据丢失、失真等情况时,自然恢复正确的值,仅仅将噪音过滤;能够结合领域知识来高效地发现知识。数据挖掘正是关于从所收集的信息中获得知识的重要信息分析方法,它能够从存在的数据中找出有效的、新颖的、有潜在用处的、易于理解的关系模型。文^{[1][3]}在这方面进行了探讨,但对 Web 上的市场营销数据挖掘未作深入的分析。由于 WWW 提供了重要的商业资源,在 Web 上进行市场营销数据挖掘是非常有价值的。本文探讨 WWW 上利用 DM 进行获取知识系统结构和市场营销数据挖掘的知识获取的几个问题。

2 基于 Web 的市场营销知识数据挖掘结构

其系统结构如图 1 所示,下面作进一步的说明:

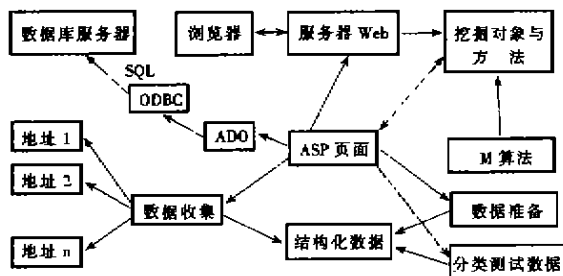


图 1 系统结构

浏览器是用户的人机界面接口,它可以是 IE5.0 或 Netscape Navigator 等和用户交流。

数据收集是收集与某一个系统目标有关的各个 Web 网址连通及信息下载有关的操作,同时创建数据库进行数据存储,其目的是将获取的未结构化的数据在使用前消除失效的信息,等待进一步处理。

ASP 页面是利用如 Visual InterDev6.0VBScript Java Script 创建专业的 HTML 和 ASP 页面,能在保留源代码格式的同时,轻松地实现多页面视图间的切换——包括所见即所得的设计、源代码视图和效果预览。如果在一台微机开发,而没有 Windows NT server,则必需在 windows95 或 Windows98 上安装 personal Web Server(个人 Web 服务器);而在 Windows NT server 中必须装 IIS,利用 IIS(VER.4.0)附带的 ASP。利用 InterDev6.0 可使和 Java script 等的声明完成,快速提示等功能可提高编程速度,还可以与各种 ODBC 兼容的数据源(如 Oracle,SQL sever, Visual foxpro, paradox 等)进行数据链接。

数据准备是把结构化的数据转化成适于数据挖掘算法进行机器的学习,用于训练和测试数据组的类别,以便于形成知识。

结构化数据:用数据库方法将 Web 上少结构化(less structured)数据重构成更结构化一些。

DM 算法是集成多种数据挖掘算法并能形成有效的模式和进行有效的解释。

分类测试数据主要是对形成的知识进行测试验证。

数据挖掘对象与方法:挖掘访问营销服务器的日志数据(Server logs; Error logs; Cookie logs; Cookie logs),特别是 Cookie logs 是自动标记和跟踪站点的访问信息,会话 cookie 是关于用户的加密信息,它被存储为简单的文本文件,包含了域名、过期时间、安全信息及路径信息。挖掘 Web 页面,从 HTML 到扩展标记语言 XML 元语言页面,如对页面内容摘要、分类、聚类及关联规则发现;挖掘 Web 页面超链接关系,分析用户的访问信息,如路径分析、访问手段、关联规则、序列模式发现聚类和分类。

3 市场营销数据挖掘的知识获取的几个问题

3.1 对取得信息理解是数据挖掘所面临的困难

* 国家自然科学基金资助项目(编号:798700727);湖北省教育厅重点研究项目(编号:2000B25010)

通过以上系统结构挖掘的市场营销如何使得知识的理解变得简单是我们必须考虑的问题。而粗糙集理论(Rough set theory)是由 E. Pawlak 于 1982 年提出的一种刻画不完整性和确定性的数学理论。它把知识看作一种对对象进行分类的能力,是关于论域的划分,从而认为知识是有粒度(granularity)的,知识的不精确性是由于组成论域知识的颗粒太大引起的。原子概念是信息系统知识表示的最小“颗粒”。“颗粒”越细,知识表示越清晰^[4]。如何改进方法,提高知识表示的准确性呢?在数据挖掘中增大特征集,扩大原子概念的规模,使知识表达的“最小”颗粒更加细化;同时在原子概念规模不变的前提下,改变原子的内涵,即从新的角度观察事物提取特征。在数据挖掘的探索假设空间(exploring hypothesis space)中,将需要较高维数进行知识描述时,进行分解细化,降低维数的工作,并利用粗糙集理论,使“颗粒”不断逼近、细化,从而使得知识的理解变得简单。

3.2 市场营销数据挖掘中的灰色信息获取

我们知道企业采取保密措施,并能为企业带来经济利益的信息,从法律角度这类信息归属于“商业秘密”的范畴,而另一类是既未公开发表,又在企业内外传播的非“商业秘密”的一类信息被称为“灰色信息”。获取“灰色信息”对于市场营销是非常重要的。灰色信息收集的正当方式与途径主要有:通过电话咨询、问卷调查、信函式面访等方式从第三方获取,录用竞争对手离职人员,通过自身企业内部员工形成的人际网络收集信息等。灰色信息与商业秘密不仅法律性质不同,而且收集方式与采取不正当手段(窃、取、利、诱、贿赂等)非法获取商业秘密的活动是不同的。商业秘密收集方式有:获取商业秘密的物质载体(未经公开的机密文件);通过贿赂或威胁等手段诱导对手企业内部相关人员泄密;非法进入对手企业内部的电子文件、数据库或通信系统;伪装成某种人员而成为对手企业的法定职工来获取情报等^[2]。灰色信息是相对于公开发表的信息而言的,它在性质上是未经大众媒体公开发表,但也未经持有主体采取保密措施,而在组织内外传播的一类信息,收集过程中一方面自觉判断与控制自己获取信息的正当行为,另外还要判断信息的可靠性、时效性^[2]。

那么在 WWW 上能否通过 DM 获取灰色信息,从而得到灰色市场营销知识呢?上面的系统结构中的 cookie 是一种比较好的方法。

3.3 外部因素对数据挖掘获取的知识所产生的影响

我们基于网络上利用各种 DM 算法挖掘知识,然而对外部因素考虑较少。我们认为,应建立企业的 Extranet 收集外部环境的市场信息,对挖掘的知识进行修正。

4 市场营销数据挖掘的经济性

4.1 信息获取的成本模型

信息获取可以增加人们更加有利决策的能力,从而提高市场营销的效率。因为在不完全信息的基础上做出的决策极可能是达不到预期目标或者是不满意决策,甚至是错误的决策。有效信息常能增加人们对事件发生的可能性判断,人们愿意为信息支付费用,目的是用于其增加期望收益。设 P 为事件发生

的概率, $1-P$ 为事件不发生的概率, C_p 为事件发生的成本, C_{1-p} 为事件不发生的成本, K 为不完全信息条件下采取行动的成本,那么人们愿意为获取信息支付费用 T 应满足下列不等式:

$$P * C_p + (1 - P) * C_{1-p} + T < K$$

在这个问题中,不对称市场可以通过各种方式如数据挖掘、担保、额外获取信息、政府法令等进行调节。然而对上述参数的估计难以确立,人们已探讨了隐藏行为的道德风险模型,隐藏信息的道德风险模型;逆向选择模型、信息传递模型和信息甄别模型等 5 种模型来探讨这种信息获取的得失。

4.2 灰色信息收集的成本模型

对于灰色信息成本模型公式我们作如下修改:设信息灰度为 g ; $0 \leq g \leq 1$, 当 $g=1$ 时,其信息的透明度最大,我们能够随时到各个网站利用 DM 工具进行数据挖掘,获取有用知识;当 $g=0$ 时,其信息的透明度最小,即可能涉及到商业秘密或国家安全的消息,这种信息获取的成本是很大的,也是难以挖掘的;当 $0 < g < 1$ 时,我们能够获取未结构化、其信息的透明度不大的数据,但是要付出一定的成本。因此,我们构造成本公式:设灰色信息获取成本为 T , K 为没有获取灰色信息条件下所需要的成本,则:

$$\frac{(P * C_p + (1 - P) * C_{1-p}) + T}{g} \leq K$$

5 结论

通过以上几个方面的探讨,为市场营销和电子商务提供了支持有效决策的参考知识。我国网民现在已接近 1000 万,而且增长速度快:1997 年 10 月底,网民 62 万;1998 年 12 月底,网民 210 万;1999 年 6 月底,网民达 400 万;1999 年底,网民已到 890 万。在 Web 上挖掘信息,寻求适合网上交易的各类物品,建立满足消费者的消费欲望,达到网上交易的网站,使得市场定位、产品选择(检索)、陈列与描述、价格的制订、物流的配送和付款方式等更加专业化、智能化和合理化。按照以上探讨的系统结构和几个问题的分析,我们正在建立模拟系统,而其真正的实用系统还有大量的工作要做。

参考文献

- 1 夏火松. 基于数据挖掘技术的市场营销智能决策支持系统. 武汉纺织工学院学报, 1999; (4)
- 2 李正中, 许 蕾. 竞争情报行为的正当性与灰色信息收集方式的研究. 情报学报, 2000; (1)
- 3 夏火松. 市场营销专家知识的获取. 武汉纺织工学院学报, 1999; (3)
- 4 胡 涛等. 基于粗糙集的不确定知识表示方法. 计算机科学, 2000; (3)
- 5 Kenneth c Laudon, Jane Price Laudon. Information Systems and The Internet. A problem - Solving Approach - 4th ed. Dryden Press, 1998
- 6 Johnny S. K Wong, Risti Nayart Armin R. Mikler. A framework for a world wide web - based Data Mining System. Journal of Network and Computer Applications, 1998
- 7 Chung, H. Michael and Gray, paul. Special Section: Data Mining. Journal of Management Information Systems, 1999; (1)

(责编:王京韵)