第 20 卷第 3 期
2003 年 9 月

测 绘 学 院 学 报
Journal of Institute of Surveying and Mapping

Vol. 20 No. 3
Sept. 2003

文章编号:1009-427X(2003)03-0180-03

基于 Web GIS 的电子商务数据挖掘研究

毛克彪[1,2]，覃志豪[1]，陈晓燕[3]，李 昕[3]

(1. 南京大学 国际地球系统科学研究所，江苏 南京 210093；2. 南京大学 城市资源系，江苏 南京 210093；
3. 南京大学 计算机系，江苏 南京 210093)

摘要:文中对 Ansari 等提出的基本的电子商务应用框架做了进一步的细化和补充，详细分析了数据收集和数据挖掘问题，包括数据来源、数据类型、数据收集器工作、数据转化、数据库建立和搜寻规则构建等。在此基础上，探讨了数据挖掘技术在基于 Web GIS 的电子商务中的具体应用。最后讨论了将数据挖掘技术应用于电子商务数据挖掘的最终目的和目前尚面临的一些困难。

关 键 词:Web GIS;电子商务;数据挖掘;应用框架

中图分类号:P282　　**文献标识码**:A

数据挖掘是一种从海量数据中发现蕴藏在数据内的规律的技术。电子商务是数据挖掘最理想的应用领域之一，因为它能提供大量的具有丰富属性的数据，一方面在于它的理想性，另一方面在于它的实用性和迫切性。将数据挖掘应用到电子商务，需要一个完善的应用框架的支撑。Suhail Ansari 等人[2]提出了一个电子商务数据挖掘的应用框架，文中对该框架作了细化和补充，并讨论了该框架的各个组成部分。然后在此基础上探讨数据挖掘技术在基于 Web GIS 的电子商务中的具体应用。最后讨论了将数据挖掘技术应用于商务数据挖掘的最终目的和目前尚面临的一些困难。

1　应用框架

文中所讨论的数据挖掘和电子商务的集成系统框架如图 1 所示。该框架主要由 4 部分组成，即商务数据定义部分、系统－顾客接口部分、数据收集器部分和数据挖掘分析部分(包括数据转换、数据仓库的建立和挖掘分析)。连接这 4 个部分的是 3 个主要的数据流，即商务数据的组织表示、顾客信息和应用结果。

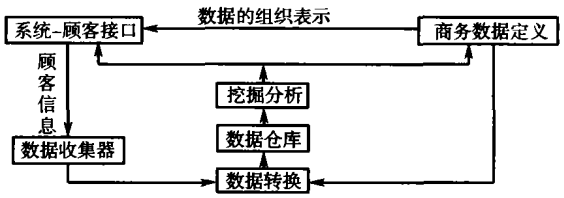

图 1　电子商务数据挖掘系统框架

1.1　商务数据定义

在商务数据定义部分，主要对产品属性等相关的数据信息进行定义，如产品说明、产品价格等，在电子商务系统中，使用者一般通过等级和分类技术建立全体商品的架构。从数据挖掘观点来看，这种基于分类架构并为产品定义丰富属性的做法在建立数据仓库阶段是很有用处的。

1.2　系统－顾客接口

系统－顾客接口部分是顾客与系统的接口。从顾客角度看，它是一系列通过超级链接组织在一起的网页。通过这些网页，顾客可以浏览商品、将商品放入或移出购物篮、检索商品或获得帮助等。从设计者角度看，这部分不仅应当设计出方便实用的网页，而且应当为每个网页定义一些属性，如登记性网页、浏览性网页、搜索性网页等。从数据挖掘角度看，方便实用的网页能够更好地搜集顾客信息。通过为网页定义属性，可以更好地记录顾客的浏览路径，以便进行分析。

1.3　数据收集器

数据收集器是数据挖掘分析流程中的核心部分，它运用各种模式来发掘算法和建立规则或模型。在电子商务领域内，它不仅可以在很大范围内自动搜集数据，而且可以得到许多在其它领域内不可能得到的数据信息。数据收集器主要收集与顾客有关的所有信息，这些信息有 4 个来源：顾客特征、顾客会话的特征记录、顾客在会话中的浏览路径以及顾客在会话中的购物过程和购物结

收稿日期:2003-01-05;**修回日期**:2003-06-11

基金资助:国家重点基础研究发展规划项目(2001CB309404);海外青年学者合作研究基金(40128001);教育部科学技术重点项目

作者简介:毛克彪(1977－)，男，湖南沅江人，硕士生，研究方向为空间数据挖掘、遥感数字图像以及高光谱应用。

果。顾客特征是顾客的登记信息,如年龄、性别、收入等。顾客一个会话的特征可记录并保存在服务器的日志文件中。在该文件中,一个会话对应着一个字串。该字串包含一些诸如客户端IP地址、会话建立时间、会话释放时间等信息。在应用该字串时,一个难点是怎样把不同的顾客区别开来。IP地址不能用来区分顾客,因为处在同一个防火墙之后的所有客户端都用同一个IP地址来登录服务器,更不用说那些共用同一个客户端的多位顾客。顾客在一个会话中的浏览路径可以形式地表示为一个有向图。图中的节点代表一个网页,而节点与节点之间的有向线段则代表网页之间的超级链接。这就为浏览路径的分析提供了可能性。顾客在一个会话中的购物过程和购物结果体现了顾客与系统发生的商务交易。记录交易过程的数据体现了电子商务系统特有的数据收集能力。与交易过程相关的数据主要体现为将商品放入或移出购物篮。另外,用户在搜索性网页内输入的关键字也是很重要的数据。通过这些关健字,可以发现用户的兴趣。

1.4 数据挖掘分析

1) 数据转化与数据库建立

用来建立电子商务数据库的数据有两个来源,一是在商务数据定义部分定义的数据,二是数据收集器收集到的与顾客有关的数据。前者直接用于数据库建立,后者则需要转化才能用于数据库建立。这些转化一般包括增加新属性、建立完善的概念等级、聚合、过滤、取样、删除属性等。

2) 挖掘分析结果表示与用户的交互

一般的数据挖掘模式在电子商务系统中都可以进行,包括关联规则、分类、聚类等。具有电子商务特色的一些基本的统计信息也是数据挖掘的对象。比如,哪些商品是最畅销的,哪些是最不畅销的,同类商品不同品牌的销售情况如何,哪些是顾客使用频率最高的搜索关键字,哪些关键字最有可能导致搜索失败,以及为什么会失败,等等。数据挖掘的结果应当用一些直观的、易于理解的可视化方法提交给使用者。此外,应当让用户能够以一种方便的方式参与挖掘分析过程。比如,允许用户微调规则的条件,并可以观察由于规则改变之后会产生的影响。这样可能会要求用户具有一些数据挖掘的知识,但却可以保证得到更好的、用户更乐于接受的结果。

3) 对用户浏览路径的分析

电子商务系统设计通常是按照常识和自己的习惯设计网页和网页之间的超级链接,但是这样的设计可能并不符合顾客的习惯。于是,后期的调整或重新设计就成为一件很重要的事情。Theusinger和Huber[3]介绍了他们在一个对顾客浏览路径进行数据挖掘项目中的经验和技术。他们认为,基于时序的关联规则和预测模式是最有用的,而在建立预测模式时,可以运用决策树、回归分析、神经网络等手段。Conen等在文献[4]中对更一般的网站讨论了怎样通过分析顾客的浏览路径建立自适应的网站。他们将自适应性调整分为战略性(长远考虑)和战术性(短期考虑)两种,但同时指出,应当谨慎地进行战略性调整,因为这会影响网站的稳定性。对于战术性调整,他提出了一个简单实用的算法,即树形数据结构法,用来从历史记录中发现使用频繁的超级链接,为浏览者提供建议。

2 数据挖掘在基于Web GIS电子商务中的应用

2.1 GIS对数据进行可视化分析

Web GIS是一个可以在不同的操作系统上提取地理空间及属性数据并能提供分析能力的一个系统。它在电子商务中的应用不仅表现为选址,而且表现为物流配送、客流分析等。因此基于Web GIS的电子商务数据挖掘是一个重要的研究方向。

利用GIS可以直观地显示商业人口、聚居人口、某个给定区域一定范围内的平均年收入,最好的客户群、在一个给定的地图区域,客户可能的竞争选择等,其中人口报告通常用来产生潜在的商店、主要的营销区域、个体客户的详细描述。这些信息在GIS里不会自动产生有决定性的决策,然而借助数据挖掘技术则可做到。

2.2 数据挖掘提供决策支持

预见性数据挖掘就是利用GIS分析技术,通过数据分析,将大量的数据精简为单个的预见或评分。在基于Web GIS电子商务数据挖掘的应用中,GIS能够联合客户的历史数据或商店的销售记录,以及企业的人口统计、商业、运输、市场研究数据,并根据这些数据建立预见性模型,评价有潜力的新区域和顾客、交叉性买卖、目标市场、客户摆动和其它相似的应用,具体应用流程见图2。

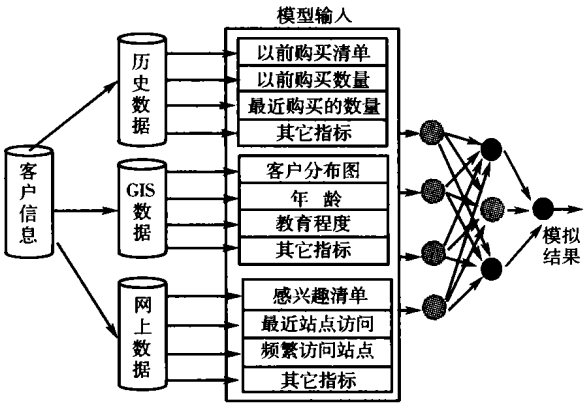

图 2 数据源模型训练示意图

2.3 制定客户关系管理目标

客户关系管理是市场的关键。其应用的核心是各种不同的数据，经常与地图联接在一起。其潜在智能来源于预见性模型。

选择商店最重要的是选择地点。为了建立一个应用，必需能够详细地分析影响地点好坏的所有因素。描述地点是 GIS 的特长，通过综合各区的人口统计数据、运输数据、商业人口统计、土地使用信息和其它公司数据，用户能够很容易地得到描述一个地点的数据。竞争对手的数据也非常重要。地点的数据用来描述在一个给定区域内商店的饱和度和竞争强度

3 基于 Web GIS 电子商务数据挖掘的目标

可以分为与商务有关和与商务无关两类，与商务有关的目标是通过发现有用的模式，使商务管理者能更好地分析和认识顾客群，以便及时、正确地调整经营策略。与商务无关的目的，在于通过顾客与系统的交互，发现其中的规律，使系统设计者能有的放矢地调整系统与用户的接口，即调整网页的内容和其间的链接关系[6]。比如，发现多数顾客，当他们浏览页面 $P1$ 时，下一个浏览的页面是 $P2$，那么可以重新设计页面 $P1$，在其中提供让顾客继续浏览页面 $P2$ 的建议。

4 基于 Web GIS 电子商务数据挖掘的困难

将基于 Web GIS 的数据挖掘与电子商务系统有机地结合是一项复杂的系统工程，有许多问题要考虑，主要包括数据收集方面的困难、挖掘分析方面的困难、挖掘算法的可扩展性和挖掘结果的可理解性。

1）数据收集方面的困难

主要表现在电子商务系统是一个虚拟的交易系统，很难保证收集到的信息的可靠性。

2）挖掘分析方面的困难

以多粒度级[7]的数据挖掘为例加以说明。数据收集器收集到的数据是分布在多个粒度级别上的：网页浏览是最低级别的；会话是次高级的，每个会话包含多个浏览过的网页；顾客是最高级的，每个顾客涉及到多个会话。如果在网页浏览级上进行挖掘，就会使用到有关的会话信息和顾客信息，从而违背了一般数据挖掘算法的基本原则，即数据记录的离散性和相互独立性。因此需要设计新的算法来支持多粒度级的数据挖掘。

3）挖掘结果被理解的困难

在电子商务系统中，数据挖掘的结果要提交给下列人员：商务决策者、DBA（数据库管理员）、网站设计师等，他们有不同的知识结构，使用挖掘结果的目的和方式也各不相同，从而应当针对他们的不同特点，分别用合适的形式表达挖掘结果。

5 结语

电子商务数据挖掘是数据挖掘的一个重要的应用领域，其研究刚刚起步。文中讨论了一个试图把数据挖掘集成到电子商务系统中的应用框架，分析了这一框架的各个组成部分和工作流程，并重点论述了其中的数据收集器部分和数据挖掘分析部分。分析了数据挖掘在基于 Web GIS 的电子商务中的具体应用。最后，总结了将数据挖掘技术应用到电子商务系统中的最终目的和目前面临的困难。

参 考 文 献：

[1] Han J, Kamber M. Data Mining :Concepts and Techniques [R]. San Mateo, California: Morgan Kaufmann Publishers, 2000:1301-1309.

[2] Suhail Ansari, Ron Kohavi, Llew Mason, et al. Integrating e-commerce and data ming:architecture and challenges [A]. ICDM'01: The 2001 IEEE International Conference on Data Mining[C],2001:1-12.

[3] Christiane Theusinger, Klaus-Peter Huber. Analyzing the footsteps of your customers[EB/OL]. http://citeseer.nj.nec.com/354100.html.

[4] Filip Conen, Gilbert Swinnen, Koen Vanhoof, et al. A Framework for Self Adaptive Websites: Tactical versus Strategic Changes[EB/OL]. http://citeseer.nj.nec.com/353699.html.

（下转第 185 页）

即可。当在某些工程地点只有内线电话而无法登陆因特网时,就可以考虑采用GPRS的方式。其次是传输的速率不同:ADSL的标称速率是上行8 Mb/s、下行2 Mb/s,而GPRS的标称速率为150 kb/s,与ADSL相比较低,但也能满足远程监控的需要。

4 结语

介绍了在自动监测系统中实现远程监视与控制的两种现代化的通讯方式。借助这些现代化的手段,就可以在各种自动化测量工程中实现远程监控管理,不仅方便了整个系统的维护,而且还降低了整个系统的运行成本,节约了大量的人力物力。系统实现了远程监控,为技术人员及时有效地维护提供了有利的条件,也为相关领导的视察与决策提供了极大的方便。

远程监控也可以在工程单位内部的局域网中实现,所以在某些工程项目中设计通讯网络时,要考虑可能会有其他系统使用通讯网,最好预留出足够的通讯端口,这样就可以更加方便地组建远程监控系统。

参考文献:

[1] 包 欢,徐忠阳,张良琚. 自动变形监测系统在地铁结构变形监测中的应用[J]. 测绘学院学报,2003,(2).

[2] 包 欢. 自动极坐标实时差分测量系统及其在大坝外部变形监测中的应用[D]. 郑州:信息工程大学测绘学院,2000.

The Application of Modern Communication Technology in Automatic Surveying System

BAO Huan[1], XU Zhong-yang[1], ZHANG Ji-nian[2]

(1. *Institute of Surveying and Mapping, Information Engineering University, Zhengzhou* 450052, *China*;
2. *Department of Educational Administration, Information Engineering University, Zhengzhou* 450002, *China*)

Abstract: This paper introduces the application of remote monitor control for Automatic surveying system based on ADSL or GPRS. It also offers some good suggestion for those who study the correlative field.

Key words: surveying system; remote monitor control system; real-time measurement; real-time processing

责任编辑 陶大欣

(上接第182页)

[5] Paul Duke. Geospatial Mining for Market Intelligence[EB/OL]. http://www.tdan.com/i016hy02.htm.

[6] Michael J, Berry A, Gordon Linoff. Data Mining Techniques: For Marketing, Sales, and Customer Support [M]. John Wiley & Sons, 2000.

[7] Saharon Rosset, Uzi Murad, Einat Neumann, et al. Discovery of Fraud Rules for Telecommunications: Challenges and Solutions[A]. Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C], 1999.

A Preliminary Study of Data Mining Technique for E-commerce Application

MAO Ke-biao[1,2], QIN Zhi-hao[1], CHEN Xiao-yan[3], LI Xin[3]

(1. *International Institute for Earth System, Nanjing University, Nanjing* 210093, *China*;
2. *Department of Urban and Resource Science, Nanjing University, Nanjing* 210093, *China*;
3. *Department of Computer Science, Nanjing University, Nanjing* 210093, *China*)

Abstract: E-commerce is an important application of data mining technique. This application requires integrating the advantages of Web, GIS and E-commerce into an entity, which involves in such components as data obtaining, database creating, and data mining. Ansari et al. proposed a basic architecture to integrate these components for applying data mining technique to the E-commerce. The paper expands and improves this architecture. Detailed discussions have been given to such issues as data sources, data types, data obtainer, data transformation, database generation and searching rules establishment. Examination is also given to the specific application of data mining in Web GIS-based E-commerce. Finally, the paper discusses the goals of data mining technique for E-commerce application and the difficulties of implementation such technique in the real world.

Key words: WEBGIS; E-commerce; data mining; architecture

责任编辑 安 敏