电子商务中的 WEB 数据挖掘与 XML

WEB Data Mining and XML in Electronic Commerce

孙 巍 （机械工业信息中心 100823）

摘要：电子商务（EC，Electronic Commerce）就是借助于计算机网络技术，通过电子交易手段来完成金融、物资、服务和信息等价值交换，快速而有效地从事各种商务活动的新方法。这种商业电子化的趋势不仅为客户提供了便利的交易方式和广泛的选择，同时也为商家提供了更加深入地了解客户需求信息和购物行为特征的可能性。无论是 B2B、B2C 还是 B2G 电子商务模式，都需要通过 Web 方式建立信息流的交互。如何把 Web 上的数据经过提取、洗涤、加工转换生成潜力巨大的价值信息，激发了数据挖掘技术在电子商务中的应用。

关键词：电子商务 数据挖掘 web xml

1 引言

随着电子商务的蓬勃发展，商业 Web 网站面临越来越激烈的竞争。面对大量的电子商务信息，找出用户感兴趣的信息加以组织利用，加强客户关系的管理，提高客户满意度，从而改进 Web 站点的设计、改善企业与客户的关系成为电子商务发展必须要解决的问题。数据挖掘概念就是从这样的商业角度开发出来的。对于企业而言，数据挖掘有助于发现业务发展的趋势，帮助企业做出正确的决策，使企业处于更有利的竞争位置。

2 数据挖掘

数据挖掘（Data Mining）就是从大量的、不完全的未知数据中提取隐含在其中的对人们分析有用的有价值信息、模式和趋势，然后以易于理解的可视化形式表达出来，其目的是为了提高市场决策能力、检测异常模式、控制可预见风险、在经验模型基础上预言未来趋势等，从而为企业提供决策依据。数据挖掘的主要步骤有数据清洗、数据集成、数据转换、数据挖掘、模式评估、知识表示等。

2.1 数据挖掘的功能

数据挖掘综合了各个学科技术，有很多的功能，当前的主要功能如下：

（1）概念描述。概念描述就是对某类对象的内涵进行描述，并概括这类对象的有关特征。概念描述分为特征性描述和区别性描述，前者描述某类对象的共同特征，后者描述不同类对象之间的区别。生成一个类的特征性描述只涉及该类对象中所有对象的共性。生成区别性描述的方法很多，如决策树方法、遗传算法等。例如：银行部门根据以前的数据将客户分成了不同的类别，现在就可以根据这些来区分新申请贷款的客户，以采取相应的贷款方案。

（2）聚类。数据库中的记录可被化分为一系列有意义的子集，即聚类。聚类增强了人们对客观现实的认识，是概念描述和偏差分析的先决条件。聚类技术主要包括传统的模式识别方法和数学分类学。80 年代初，Mchalski 提出了概念聚类技术牠其要点是，在划分对象时不仅考虑对象之间的距离，还要求划分出的类具有某种内涵描述，从而避免了传统技术的某些片面性。例如：将申请人分为高度风险申请者，中度风险申请者，低度风险申请者。

（3）关联规则和序列模式的发现。数据关联是数据库中存在的一类重要的可被发现的知识。若两个或多个变量的取值之间存在某种规律性，就称为关联。关联可分为简单关联、时序关联、因果关联。关联分析的目的是找出数据库中隐藏的关联网。有时并不知道数据库中数据的关联函数，即使知道也是不确定的，因此关联分析生成的规则带有可信度。例如：每天购买啤酒的人也有可能购买香烟，比重有多大，可以通过关联的支持度和可信度来描述。与关联不同，序列是一

种纵向的联系。例如:今天银行调整利率,明天股市的变化。

(4) 预测。数据挖掘自动在大型数据库中寻找预测性信息,以往需要进行大量手工分析的问题如今可以迅速直接由数据本身得出结论。一个典型的例子是市场预测问题,数据挖掘使用过去有关促销的数据来寻找未来投资中回报最大的用户,其它可预测的问题包括预报破产以及认定对指定事件最可能作出反应的群体。

(5) 偏差的检测。数据库中的数据常有一些异常记录,从数据库中检测这些偏差很有意义。偏差包括很多潜在的知识,如分类中的反常实例、不满足规则的特例、观测结果与模型预测值的偏差、量值随时间的变化等。偏差检测的基本方法是,寻找观测结果与参照值之间有意义的差别。例如:在银行的 100 万笔交易中有 500 例的欺诈行为,银行为了稳健经营,就要发现这 500 例的内在因素,减小以后经营的风险。

需要注意的是:数据挖掘的各项功能不是独立存在的,在数据挖掘中互相联系,发挥作用。

数据挖掘的方法

(1) 关联分析法。关联分析法从关系数据库中提取关联规则是几种主要的数据挖掘方法之一。挖掘关联是通过搜索系统中的所有事物,并从中找到出现条件概率较高的模式。关联实际上就是数据对象之间相关性的确定,用关联找出所有能将一组数据项和另一组数据项相联系的规则,这种规则的建立并不是确定的关系,而是一个具有一定置信度的可能值,即事件发生的概率。关联分析法直观、易理解,但对于关联度不高或相关性复杂的情况不太有效。

(2) 人工神经元网络(ANN)。人工神经元网络是数据挖掘中应用最广泛的技术。神经网络的数据挖掘方法是通过模仿人的神经系统来反复训练学习数据集,从待分析的数据集中发现用于预测和分类的模式。神经元网络对于复杂情况仍能得到精确的预测结果,而且可以处理类别和连续变量,但神经元网络不适合处理高维变量,其最大的缺点是不透明性,因为其无法解释结果是如何产生的,及其在推理过程中所用的规则。神经元网络适合于结果比可理解性更重要的分类和预测的复杂情况,可用于聚类、分类和序列模式。

(3) 决策树(DT)。决策树是一种树型结构的预测模型,其中树的非终端节点表示属性,叶节点表示所属的不同类别。根据训练数据集中数据的不同取值建立树的分支,形成决策树。与神经元网络最大的不同在于其决策制定的过程是可见的,可以解释结果是如何产生的。决策树一般产生直观、易理解的规则,而且分类不需太多计算时间,适于对记录分类或结果的预测,尤其适用于当目标是生成易理解、可翻译成 SQL 或自然语言的规则时。决策树也可用于聚类、分类及序列模式,其应用的典型例子是 CART(回归决策树)方法。

(4) 遗传算法(GA)。遗传算法是一种基于生物进化理论的优化技术。其基本观点是"适者生存"原理,用于数据挖掘中则常把任务表示为一种搜索问题,利用遗传算法强大的搜索能力找到最优解。实际上遗传算法是模仿生物进化的过程,反复进行选择、交叉和突变等遗传操作,直至满足最优解。遗传算法可处理许多数据类型,同时可并行处理各种数据,常用于优化神经元网络,解决其他技术难以解决的问题,但需要的参数太多,对许多问题编码困难,一般计算量大。

(5) 近邻算法。近邻算法将数据集合中每一个记录进行分类。

(6) 传统统计方法。传统统计方法包括:抽样技术:我们面对的是大量的数据,对所有的数据进行分析是不可能的也是没有必要的,就要在理论的指导下进行合理的抽样; 多元统计分析:因子分析,聚类分析等; 统计预测方法,如回归分析,时间序列分析等。

除了上述的常用方法外,还有粗集方法,模糊集合方法等。

3 WEB 数据挖掘的流程

Web 挖掘是指从大量 Web 文档的集合 C 中发现隐含的模式 p. 如果将 C 看作输入,将 p 看作输出,那么 Web 挖掘的过程就是从输入到输出的一个映射 N: C→p 。与传统数据和数据仓库相比,Web 上的信息是非结构化或半结构化的、动态的、并且是容易造成混淆的,从数据库研究的角度出发,Web 上网站的信息也可以看作是一个数据库,一个更大的、复杂性更高的

数据库。所以很难直接以 Web 网页上的数据进行数据挖掘，而必须经过必要的数据处理。典型 Web 挖掘的处理流程如下：

（1）查找资源。任务是从目标 Web 文档中得到数据，值得注意的是有时信息资源不仅限于在线 Web 文档，还包括电子邮件、电子文档、新闻组，或者网站的日志数据甚至是通过 Web 形成的交易数据库中的数据。

（2）数据清洗（data cleaning）和事务识别（transaction identification）。包括对 Web 日志进行清洗、过滤和转换以及无关记录的剔除，判断是否有重要的访问没有被记录，并从中抽取感兴趣的数据；并将 URL、资源的类型、大小、请求的时间、在资源上停留的时间、请求者的 Internet 域名、用户、服务器状态作为数据 cube 的维数变量；再将对模块、页面和文件请求次数，来自不同 Internet 域请求次数、事件、会话、带宽、错误次数、不同浏览器种类、用户所在组织作为度量变量建立 data cube；而将文件、图像脚本及多媒体等其他文件转换成可用于 Web 使用挖掘的数据格式，从而可将数据挖掘技术用于 Web 流量分析、典型的事件序列分析和用户行为模式分析及事务分析。

（3）模式发现和模式分析。在经过数据清洗和事务识别后，即可根据不同的需求选择模式发现技术，如统计分析、关联规则、时序模式、路径分析（path analysis）及聚类、分类技术。其中统计分析通过分析会话文件可对网页视图、浏览时间和导航路径长度给出描述性的统计分析。该分析有助于改进系统性能，增强系统安全性，便于站点修改并可提供决策支持。路径分析可用于发现 Web 站点中最经常被访问的路径，从而可调整站点结构。基于 Web 日志的关联规则挖掘则可发现用户与站点各页面的访问关系，可找出在某次服务器会话中经常出现的一些相关网页，即支持度超过预设阈值的一组网页。聚类则多指客户群体聚类和 Web 网页聚类。客户群体聚类指将具有相似浏览模式的用户分在一组，从而方便电子商务网站为用户提供个性化服务，而 Web 页面聚类则提供有针对性的网络服务应用。时序模式发现是根据一段时间的 Web 使用记录分析是否存在一定趋势，以预测未来的访问模式。

在逻辑上，我们可以把 Web 看作是位于物理网络之上的一个有向图 G = （N，E），其中节点集 N 对应于 Web 上的所有文档，而有向边集 E 则对应于节点之间的超链。对节点集作进一步的划分，N = {N l，N nl}。所有的非叶节点 N nl 是 HTML 文档，其中除了包含文本以外，还包含了标记以指定文档的属性和内部结构，或者嵌入了超链以表示文档间的结构关系。叶节点 N l 可以是 HTML 文档，也可以是其它格式的文档，以及图形、音频等媒体文件（如图 1 所示）。N 中每个节点都有一个 URL，其中包含了关于该节点所位于的 Web 站点和目录路径的结构信息。

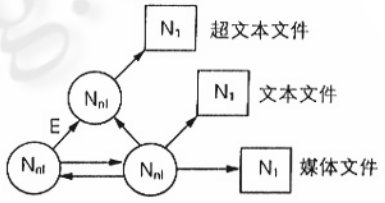

图 1　Web 的逻辑结构

Web 上信息的多样性决定了 Web 挖掘任务的多样性。

4　WEB 数据挖掘的分类

一般地，Web 挖掘可分为三类（如图 2 所示）：

（1）Web 内容挖掘。从原始的文本及其描述或多媒体中抽取知识的过程

（2）Web 结构挖掘。从网站的组织结构和链接关系中推导知识

（3）Web 使用挖掘。从 Web 的访问记录中抽取感兴趣的模式

其中 Web 使用挖掘是目前研究的重点。

5　WEB 数据挖掘的任务与面临的关键问题

数据挖掘的主要任务是进行数据描述和预测，描述数据的一般特性，对数据进行合并分组，再在当前数据上进行推断预测。这就要求数据挖掘系统要能够挖掘多种类型的数据模式，以适应不同的用户需求。Web 数据挖掘是数据挖掘技术在 Web 环境下的应用，通过对客户访问 Web 网站的数据分析，从中获取有价值的电子商务信息，从中得到详细的商务行为细节，用于商业决策，但如何解决不同语言和交易平台、不同的

协议和数据结构的集成与查询问题，成为了电子商务系统集成的关键。 解决 Web 上的异构数据的集成与查询问题，就必须要有一个模型来清晰地描述 Web 上的数据。针对 Web 上的数据半结构化的特点，寻找一个半结构化的数据模型是解决问题的关键所在。除了要定义一个半结构化数据模型外，还需要一种半结构化模型抽取技术，即自动地从现有数据中抽取半结构化模型的技术。面向 Web 的数据挖掘必须以半结构化模型和半结构化数据模型抽取技术为前提。所幸的是标记语言 XML 的出现为为解决这一难题提供了新的思路。

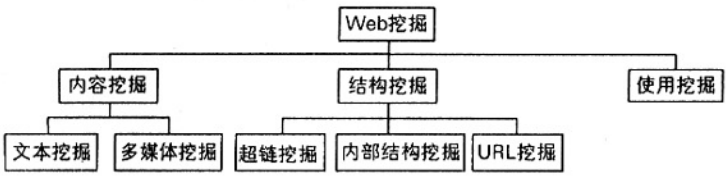

图 2　Web 挖掘的分类

6　XML 与 Web 数据挖掘技术

以 XML 为基础的新一代 WWW 环境是直接面对 Web 数据的，不仅可以很好地兼容原有的 Web 应用，而且可以更好地实现 Web 中的信息共享与交换。XML 可看作一种半结构化的数据模型，可以很容易地将 XML 的文档描述与关系数据库中的属性一一对应起来，实施精确地查询与模型抽取。

XML 已经成为正式的规范，开发人员能够用 XML 的格式标记和交换数据。XML 在三层架 构上为数据处理提供了很好的方法。使用可升级的三层模型，XML 可以从存在的数据中产生出来，使用 XML 结构化的数据可以从商业规范和表现形式中分离出来。

XML 给基于 Web 的应用软件赋予了强大的功能和灵活性，因此它给开发者和用户带来了许多好处。比如进行更有意义的搜索，并且 Web 数据可被 XML 唯一地标识。没有 XML，搜索软件必须了解每个数据库是如何构建的，但这实际上是不可能的，因为每个数据库描 述数据的格式几乎都是不同的。由于不同来源数据的集成问题的存在，现在搜索不兼容的数据库实际上是不可能的。XML 能够使不同来源的结构化的数据很容易地结合在一起。然后，数据就能被发送到客户或其他服务器做进一步的集合、处理和分发。XML 的扩展性和灵活性允许它描述不同种类应用软件中的数据，从描述搜集的 Web 页到数据记录，从而通过多种应用得到数据。同时，由于基于 XML 的数据是自我描述的，数据不需要有内部描述就能被交换和处理。利用 XML，用户可以方便地进行本地计算和处理，XML 格式的数据发送给客户后，客户可以用应用软件解析数据并对数据进行编辑和处理。XML 文档对象模式（DOM）允许用脚本或其他编程语言处理数据，数据计算不需要回到服务器就能进行。XML 可以被利用来分离使用者观看数据的界面，使用简单灵活开放的格式，可以给 Web 创建功能强大的应用软件。

XML 还可以通过以简单开放扩展的方式描述结构化的数据，XML 补充了 HTML，被广泛地用来描述使用者界面。HTML 描述数据的外观，而 XML 描述数据本身。由于数据显示与内容分开，XML 定义的数据允许指定不同的显示方式，使数据更合理地表现出来。本地的数据能够以客户配置、使用者选择或其他标准决定的方式动态地表现出来。CSS 和 XSL 为数据的显示提供了公布的机制。通过 XML，数据可以粒状地更新。每当一部分数据变化后，不需要重发整个结构化的数据。变化的元素必须从服务器发送给客户，变化的数据不需要刷新整个使用者的界面就能够显示出来。XML 应用于客户需要与不同的数据源进行交互时，数据可能来自不同的数据库，它们都有各自不同的复杂格式。但客户与这些数据库间只通过一种标准语言进行交互，那就是 XML。由于 XML 的自定义性及可扩展性，它足以表达各种类型的数据。客户收到数据后可以进行处理，也可以在不同数据库间进行传递。总之，在这类应用中，XML 解决了数据的统一接口问题。但是，与其他的数据传递标准不同的是，XML 并没有定义数据文件中数据出现的具体规范，而是在数据中附加 TAG 来表达数据的逻辑结构和含义。这使 XML 成为一种程序能自动理解的规范。XML 的自解释性使客户端在收到数据的同时也理解数据的逻辑结构与含义，从而使广泛、通用的分布式计算成为可能。

（下转第 24 页）

(上接第 28 页)

7 结束语

XML 的出现为解决 Web 数据挖掘的难题带来了机会。由于 XML 能够使不同来源的结构化的数据很容易地结合在一起,因而使搜索多样的不兼容的数据库能够成为可能。XML 的扩展性和灵活性允许 XML 描述不同种类应用软件中的数据,从而能描述搜集的 Web 页中的数据记录。随着 XML 作为在 Web 上交换数据的一种标准方式的出现,面向 Web 的数据挖掘将会变得非常轻松,从而极大的推动电子商务的发展。

参考文献

1 (加)Jiawei Han Micheline Kamber《数据挖掘概念与技术》,机械工业出版社,2001。

2 Chaudhri, A. B, Rashid, A, Zicari, R. XML 数据管理纯 XML 和支持 XML 的数据库系统,北京 清华大学出版社,2006。

3 Jackson J, Myllymakl J 基于 Web 的数据挖掘 http://www.IBM.com,2001