数据挖掘技术及在电子商务中的应用

刘兴雨 （首都经济贸易大学信息系 100026）

摘要：数据挖掘(Data Mining)被认为是解决"数据爆炸"和"数据丰富，信息贫乏(Data Rich and Information Poor)"的一种有效方法。本文对数据挖掘的基本原理作了阐述，分析了数据挖掘的流程及主要功能，介绍了数据挖掘工具的算法和特点，并结合具体实例说明了数据挖掘在电子商务中的应用。

关键词：数据挖掘 电子商务 预测 算法

1 信息“爆炸”与数据挖掘

随着互联网的出现，信息过量几乎成为人人需要面对的问题。如何才能不被信息的汪洋大海所淹没，从中及时发现有用的知识，提高信息利用率，变得十分重要。因此，数据挖掘技术应运而生，并且显示出越来越强大的生命力。

同样，一个电子商务网站每天需要搜集和处理大量的数据，利用数据挖掘技术可以帮助商家了解客户以往的需求趋势，并预测未来，从而给商家带来巨大的利润。那么什么是数据挖掘呢？

数据挖掘是按照既定的业务目标，对大量的企业数据进行探索、揭示隐藏其中的规律性并进一步将之模型化的先进、有效的方法。

数据挖掘（DataMining）和数据库知识发现（Knowledge Discovery in Database,KDD）是近年来随着数据库和人工智能技术的发展而出现的全新信息技术，同时也是计算机科学与技术，尤其是计算机网络的发展和普遍使用所提出的、迫切需要解决的重要课题。很多人将数据挖掘和KDD作为互换的术语来使用，其实它们是有区别的，KDD是一个综合的过程，包括实验记录、迭代求解、用户交互以及许多定制要求和决策设计等。而数据挖掘是指从数据中提取模式的过程。所以数据挖掘只是KDD中的一个具体但又关键的步骤。

2 数据挖掘的技术与方法

在数据挖掘中最常用的技术有：

(1) 人工神经网络（Artificial Neural Networks）：仿照生理神经网络结构的非线形预测模型，通过学习进行模式识别。

(2) 决策树方法(Decision Trees)：代表着决策集的树形结构。

(3) 遗传算法(Genetic Algorithms)：基于进化理论，并采用遗传结合、遗传变异、以及自然选择等设计方法的优化技术。

(4) 邻近搜索算法(Nearest Neighbor Method)：将数据集合中每一个记录进行分类的方法。

(5) 规则推理(Rule Induction)：从统计意义上对数据中的"如果-那么"规则进行寻找和推导。

另外还有集合论的粗集方法（Rough Set）、模糊逻辑(Fuzzy Logic)、公式发现等等。

数据挖掘常用的方法有：

(1) 关联分析（Associations）：其目的就是挖掘出隐藏在数据间的相互关系。

(2) 序列模式分析（Sequential Patterns）：侧重于挖掘数据的前后时间顺序关系。

(3) 分类分析（Classifiers）：可以用来描述一些记录的特征。

(4) 聚类分析（Clustering）：根据一定的规则合理地划分数据，是与分类互逆的过程。

3 实例

假如一个电子商务网站的数据分析员想要知道一些商品出售方面得问题，比如：30-40岁之间的女性对什么产品最感兴趣，她们在购买该产品时，通常还会购买什么其他产品……如果只靠以前传统得人工技术，从巨大的商品购买信息中找到答案几乎是不可能的。此时他需要的就是数据挖掘技术。他所要做的主要工作如下：

3.1 首先需要做的是要明白数据挖掘所能解决的典型问题是什么

数据挖掘所能解决的典型商业问题包括：数据库营销（Database Marketing）、客户群体划分（Customer Segmentation & Classification）、背景分析（Profile Analysis）、交叉销售（Cross-selling）等市场分析行为，以及客户流失性分析(Churn Analysis)、客户信用记分(Credit Scoring)、欺诈发现(Fraud Detection)等等。

明白这个问题后，他就可以知道他所解决的问题是否能从数据挖掘中找到满意的答案。

3.2 选择合适的数据挖掘工具

如果当这个数据分析员从上一步的分析中发现，他所要解决的问题恰好是用数据挖能够比较好地完成。那么他需要做的第二步就是选择合适的数据挖掘技术与方法。

首先，他要将商业问题转化成一系列数据挖掘的任务。主要有六种任务：分类，估值，预测，篮子分析(market basket analysis,关联技术的一种应用，其目的是发现同时发生的事件之间的分组)，聚集，描述。

例如：分析什么年龄段的客户对某种商品的购买最多，其任务就是聚集，但是可以采用的技术有很多：关联分析、聚集分析、决策树、人工神经网络。再比如分析客户流失的原因，其任务就是分类，同样可以采用的技术有很多：遗传算法，决策树，人工神经网络。从中选择了决策树，是因为分完类之后，我们需要知道每个类的流失的原因。

任务 / 技术、方法	分类	估值	预测	篮子分析	聚集	描述
关联分析			是	是	是	是
遗传算法	是		是			
聚集分析					是	
决策树	是		是		是	是
人工神经网络	是	是	是		是	

其次，准备数据。他首先要从企业大量数据中取出一个与要搜索的问题相关的样板数据子集，而不是动用全部企业数据。通过对数据样本的精选，不仅能减少数据处理量，节省系统资源，而且能通过对数据的筛选，使数据更加具有规律性。

第三步，数据分析。就是通常所进行的对数据深入调查的过程。此时他需要从样本数据集中找出规律和趋势，用聚类分析区分类，这时要尽可能对问题解决的要求能进一步明确化、进一步量化。针对问题的需求要对数据进行增删，按照对整个数据挖掘过程的新认识组合或生成一个新的变量，以体现对状态的有效描述。理解可以获得的数据的信息：内容、字段类型、记录之间的关系。最终要达到的目的就是搞清楚多因素相互影响的、十分复杂的关系，发现因素之间的相关性。

可能影响技术选择的一些数据性质：

(1) 种类字段占优势。关联分析和连接分析只适用于种类字段。决策树也可以很容易的用于种类字段。但是，当种类的值较多的时候，效果可能就会比较的差，当然如果限制分支的个数的时候，决策树的效果还是不错的。神经元网络可以将种类字段转化成数值字段，但是这样就给种类字段强加了一个先后次序。也可以将种类字段作为多个输入，但是当值很多时，这种方法就成问题了。

(2) 数值字段占优势。人工神经网络将所有输入转化到0-1之间。聚集分析通过距离函数来处理数值字段。决策树可以通过 splitter 数值来处理。对于关联分析，必须将数值变量区间化成种类变量。但是区间的选择是一个很困难的问题。

(3) 每个记录都有大量的字段（独立）神经元网络和MBR技术会受其影响，关联规则挖掘也会受影响。而决策树受其影响的程度就比较的小。

(4) 多个目标字段（非独立）。神经元网络是最佳的选择。

(5) 记录是变长的。只有关联规则和连接分析可以直接处理。

(6) 有时间顺序的数据。人工神经网络，关联规则对时间顺序的数据的处理能力比较的好。决策树也能处理时间顺序，但是需要的数据准备就相对的比较多一点。

这样通过明确数据挖掘任务、分析数据性质，选择适合的数据挖掘技术和方法。

3.3 选择合适的数据挖掘产品

对于数据挖掘产品的选择可以从以下几个方面考虑：

商业评价：更多地考虑市场和资金方面。

应用评价：比较在某一领域某种产品更为适用。

算法评价：从数据挖掘的最底层比较这些技术。

数据分析员可以从以上三个方面综合地考虑，选择适合自己电子商务网站的产品。

3.4 建模

利用选择好的数据挖掘方法和数据挖掘工具，进一步明确问题，进一步调整数据结构和内容，运用神经网络、决策树、时间序列分析等确定的一种或几种方法来建立模型。这一步是数据挖掘的核心环节。

（下转第 51 页）

(上接第 47 页)

3.5 评价

从上述过程中将会得出一系列的分析结果、模式和模型，多数情况会得出对目标问题多侧面的描述，这时就要综合它们的规律性，提供合理的决策支持信息。评价的一种办法是直接使用原先建立模型样本和样本数据来进行检验。另一种办法是另找一批数据并对其进行检验，已知这些数据能反映客观实践的规律性。再一种办法是在实际运行的环境中取出新鲜数据进行检验。如果分析人员对分析结果不满意可递归地执行以上几个过程，直到满意为止。

4 数据挖掘的意义与展望

对于给定的数据库，数据挖掘技术能够自动趋势预测、自动探测以前未发现的模式、还可以让现有的软件和硬件更加自动化，并且在升级的或者新开发的平台上执行，从而产生巨大的商业利润。

最近，Gartner Group 的一次高级技术调查将数据挖掘和人工智能列为"未来三到五年内将对工业产生深远影响的五大关键技术"之首，并且还将并行处理体系和数据挖掘列为未来五年内投资焦点的十大新兴技术前两位。根据最近 Gartner 的 HPC 研究表明，"随着数据捕获、传输和存储技术的快速发展，大型系统用户将更多地需要采用新技术来挖掘市场以外的价值，采用更为广阔的并行处理系统来创建新的商业增长点。"■

参考文献

1 王珊，数据仓库技术与联机分析处理

2 林尧瑞，人工智能导论

3 Data Mining