论行业信息资源的数据挖掘

雷　宇
（重庆市烟草公司，重庆　400060）

加入WTO，使作为世界烟草第一大国的中国变成国内外烟草企业激烈竞争的主战场。“大力推进信息化建设，用现代技术和管理手段，推动烟草行业现代化水平的提高，使信息化建设在行业改革与发展中发挥更大的作用”也成为了近年来重庆烟草大力推进的工作内容。那么随着行业信息技术的发展，各种应用系统已经在重庆烟草的许多业务上发挥着不可或缺的重要作用，这些业务系统通过网络的互联，使得计算机拥有了海量的信息资源，但是由于各应用系统相对独立存在，信息孤岛现象随之产生。为此数据仓库与数据挖掘技术被引入重庆烟草，成为一项重要的研究课题，并且“行业数据中心及辅助决策支持系统” 项目作为今年信息中心的重要工作正在实施中。

数据挖掘的目的在深入分析数据信息特点的基础上，着重从功能的角度给出一种用于数据挖掘的过程框架，并且伴随着数据挖掘的进行，相应的数据信息也是从单纯的数据到知识模式的演进，使之为重庆烟草辅助决策支持做出重要的贡献。

1　什么是数据挖掘

数据挖掘(ＤａｔａＭｉｎｉｎｇ)是指从数据集中提取潜在的、人们感兴趣的知识，并把提取的知识表示为概念(Ｃｏｎｃｅｐｔｓ)、规则(Ｒｕｌｅｓ)、规律(Ｒｅｇｕ ｌａｒｉｔｉｅｓ)、模式(Ｐａｔｔｅｒｎｓ)等形式，更广义的说法是:数据挖掘是指在一些事实或观察数据集合中寻找模式的过程。

2　数据挖掘的起源和研究背景

近年来，随着大规模的工业生产过程的自动化、商务贸易电子化及企业和政府事务电子化的迅速普及以及科学计算的日益增长，产生了大规模的数据源。计算机网络技术的长足进步也为数据的传输和远程交互提供了技术手段。日益成熟的数据库系统和数据库管理系统都为这些海量数据的存储和管理提供了技术保证，为步入信息时代奠定了基础，这些庞大的数据库及其中的海量数据是极其丰富的信息源。

在这些信息源中隐含了许多有潜在价值的知识，如何发现这些有用的知识是人工智能、数据库等领域的研究焦点。但是仅仅依靠传统的数据检索机制和统计分析方法已经远远不能满足需要了。因此，近年来出现了一门新兴的知识获取提取技术——数据挖掘。数据挖掘旨在从数据库中提取正确的、非平常的、未知的、有潜在应用价值的并最终可为用户理解的模式。它的出现为自动和智能地把海量的数据转化成有用的信息和知识提供了手段。数据挖掘涉及到诸如机器学习、模式识别、统计学、数据库和人工智能等等众多学科，是数据库理论和机器学习的交叉学科。

3　数据挖掘的特点

与传统的数据库查询系统相比较,数据挖掘技术有以下不同。

ａ．传统的数据库查询一般都具有严格的查询表达式，可以用ＳＱＬ语句描述，而数据挖掘则不一定具有严格的要求，常常表现出即时、随机的特点，查询要求也不确定。整个挖掘过程也无法仅用ＳＱＬ语言就能完整表达，实际上，

数据挖掘常常用一种类ＳＱＬ语言来描述。

ｂ．传统的数据库查询一般生成严格的结果集，但数据挖掘可能并不生成严格的结果集。挖掘过程往往基于统计规律，产生的规则并不要求对所有的数据项总是成立，而是只要达到一定的事先给定的阈值就可以了。

ｃ．通常情况下，数据库查询只对数据库的原始字段进行；而数据挖掘则可能在数据库的不同层次上发掘知识规则，传统的数据挖掘基于关系数据库或数据仓库，所处理数据具有完整的结构，ＤＢＭＳ还提供了查询和统计的快速响应。

4　数据挖掘在系统中的全过程

数据挖掘的任务就是在海量的数据中发现有用的数据。但是仅仅发现数据那是不够的。我们必须对这种模型做出一定的反应，并采取行动，最后将有用的数据转换成信息，信息变成行动，行动转换成价值。

下面给出数据挖掘过程的步骤：

ａ．定义商业问题。

有些问题的产生是显然的，如：开辟新品牌的市场；为现存的品牌和新品牌定价；了解客户情况等等。但有些问题是不明显的，那么就要和各类业务人员进行交流，当他们了解了数据挖掘之后，他们就有可能提出更好的问题。为此，必须对他们进行启发式的介绍，开阔他们的思路。

ｂ．准备数据。

准备数据首先就要鉴别数据的可用性，以下一些情况可能影响数据挖掘的效果：

(1)不好的数据格式。

(2)另人费解的数据格式以及各个系统中数据含义的不一致。

(3)缺少相应可以实施的功能。

(4)挖掘出的结果缺乏充分的理由。

(5)企业内部组织的问题。

那么如果要分析、规范、建立全面满足行业需求的各种应用功能，保证挖掘结果的时效性，其系统应具备以下功能的集成：

(1)首先，决策支持系统的功能要全面，涵盖行业管理的所有层面，为公司上上下下的运作提供支持，为各层管理者及业务需求者所用。

(2)对系统的共性，建立行业有指导性的编码标准和基础数据库。如：单位代码基本信息库等等。行业内各基层单位可通过广域网实时下载基础库信息，保证整个行业标准的一致性。

(3)为了进一步保证系统的数据的安全性和准确性，对行业内各基层单位的上报数据要采取安全措施和审核措施，保证系统的安全性和上报数据的准确性。

(4)灵活的权限设置。系统内的任何一个单位，只要得到领导认可，就可以通过灵活的权限设置，看到系统内所有表中任意一张表，在没有领导认可的情况下，就完全看不到任何东西，也就是说，有多少权限，就能看或者制作多少东西。

(5)丰富的报表功能。除了要生成定制报表、异常报表和常规报表之外，还要根据需要动态生成临时报表，做临时的分析决策。

(6)简单丰富的领导查询系统。要将行业长时间积累的工作经验和管理方法相结合，形成一套具有实用价值，并且能够用图、文、表、曲线多种方法进行诠释，既清楚又直观的查询系统。为了使领导能够及时准确地掌握随时想查看的内容，必须提供一套灵活的动态查询方法。可以在任一单位查阅单个表或多个相关表中的数据内容。进行同类数据比

较、进行数据分析等。

(7)各处室及基层单位也应有自己的综合统计、条件查询和动态灵活查询功能，并且在权限的允许下，也可以自己进行动态统计分析。

(8)建立行业自己的行业网站，以便上传和下载行业内的各种工具和信息，并能通过自己的邮件进行交流，通过论坛等形式进行大范围的工作探讨。

c．根据信息建立相应的模型。

一旦所有数据准备好以后，就要选择数据挖掘的算法，接着建立模型。一旦当模型建立好以后，就要不断训练模型，对模型进行验证，验证时只使用系统中的一部分数据，然后使用系统中的另外一部分数据作为验证的数据集。

d．评价。

无论我们通过模拟方法计算出来的模型准确率有多高，都不能保证这个模型在面对现实世界中的真实数据时能取得好的效果。因此，对挖掘结果采取行动的过程中，各个不同的部门采取的行动应该是不相同的。其中有一个主要的方法是：不要一下子把面铺的很广，而是应该采取一种循序渐进的方法，逐渐扩大范围，使得结果具有可控性。

e．全面实施。

由分析人员对模型进行最终评定后，提出的行动方案的建议，然后付诸实施。在实施应用模型之后，还需要不断地对其进行监控。因为有一些分析是不断变化的，例如，客户消费行为、客户消费品牌，人民购买习惯等都会随着社会的发展而变化。

f．衡量结果。

衡量是对我们采取行动的结果给予的一个回馈。现在很多的实际情况是由于我们的工作很忙以至于忽略了对结果的衡量。这种行为是一个错误。因为每一个数据挖掘的过程，不管成功与否，都可以有一定的经验以用于将来的活动之中。现在的问题是如何进行衡量以便提供最好的效果。一个好的想法是我们可以把数据挖掘作为一个小的商业过程。通过比较将实际结果和预测的值进行比较，我们往往可以发现一些很好的问题用于下一个数据挖掘的过程中。

5　结束语

数据挖掘归根结底，它的本质就是把数据变为知识的过程。通过参与重庆烟草数据中心与决策与辅助支持系统项目，本人感觉到把分布在行业中，分散的，紊乱的不统一的数据，以科学的方法收集、存储、清洗、转换并进行分析挖掘后，就会成为也必然成为企业的宝贵财富，成为巨大的企业无形资产。

作者简介：　雷　宇，计算机科学与工程学院计算机科学与技术专业本科毕业，现在重庆市烟草公司信息中心工作。期间，全程参与了重庆市烟草公司计算机网络安全项目，现正在参与实施重庆市烟草公司行业数据中心及分析与辅助决策支持系统项目。