数据挖掘

摘要：

顾名思义，数据挖掘就是从大量的数据中挖掘出有用的信息。它是根据人们的特定要求，从浩如烟海的数据中找出所需的信息来，供人们的特定需求使用。2000 年 7 月，IDC发布了有关信息存取工具市场的报告。1999 年，数据挖掘市场大概约为 7.5 亿美元，估计在下个 5 年内市场的年增长率为 32.4%，其中亚太地 区为 26.6%。到 2002 年，该市场会发展到 22 亿美元。据国外专家预测，随着数据量的日益积累和计算机的广泛应用，在今后的 5—10 年内，数据挖掘将 在中国形成一个新型的产业。

为了帮助大家了解数据挖掘的基本理论和方法，我们从“数据挖掘讨论组”网站上整理加工了一组有关该概念的基本知识，省却了纷繁的技术方法，供读者学习参考。

URL：http://www.stcsm.gov.cn/learning/lesson/xinxi/20021125/lesson.asp

第一课 数据挖掘技术的由来

1.1 网络之后的下一个技术热点

我们现在已经生活在一个网络化的时代，通信、计算机和网络技术正改变着整个人类和社会。如果用芯片集成度来衡量微电子技术，用CPU处理速度来衡量计 算机技术，用信道传输速率来衡量通信技术，那么摩尔定律告诉我们，它们都是以每 18 个月翻一番的速度在增长，这一势头已经维持了十多年。在美国，广播达到 5000 万户用了 38 年；电视用了 13 年；Internet拨号上网达到 5000 万户仅用了 4 年。全球IP网发展速度达到每 6 个月翻一番，国内情况亦然。 1999 年初，中国上网用户为 210 万，现在已经达到 600 万。网络的发展导致经济全球化，在 1998 年全球产值排序前 100 名中，跨国企业占了 51 个， 国家只占 49 个。有人提出，对待一个跨国企业也许比对待一个国家还要重要。在新世纪钟声刚刚敲响的时候，回顾往昔，人们不仅要问：就推动人类社会进步而 言，历史上能与网络技术相比拟的是什么技术呢?有人甚至提出要把网络技术与火的发明相比拟。火的发明区别了动物和人，种种科学技术的重大发现扩展了自然人 的体能、技能和智能，而网络技术则大大提高了人的生存质量和人的素质，使人成为社会人、全球人。

现在的问题是：网络之后的下一个技术热点是什么？让我们来看一些身边俯拾即是的现象：《纽约时报》由 60 年代的 10～20 版扩张至现在的 100～200 版，最高曾达 1572 版；《北京青年报》也已是 16～40 版；市场营销报已达 100 版。然而在现实社会中，人均日阅读时间通常为 30～45 分钟，只能浏览一份 24 版的报纸。大量信息在给人们带来方便的同时也带来了一大堆问题：第一是信息过量，难以消化；第二是信息真假难以辨识；第三是信息安 全难以保证；第四是信息形式不一致，难以统一处理。人们开始提出一个新的口号：“要学会抛弃信息”。人们开始考虑：“如何才能不被信息淹没，而是从中及时 发现有用的知识、提高信息利用率？”

面对这一挑战，数据开采和知识发现（DMKD）技术应运而生，并显示出强大的生命力。

1.2 数据爆炸但知识贫乏

另一方面，随着数据库技术的迅速发展以及数据库管理系统的广泛应用，人们积累的数据越来越多。激增的数据背后隐藏着许多重要的信息，人们希望能够对其 进行更高层次的分析，以便更好地利用这些数据。目前的数据库系统可以高效地实现数据的录入、查询、统计等功能，但无法发现数据中存在的关系和规则，无法根 据现有的数据预测未来的发展趋势。缺乏挖掘数据背后隐藏的知识的手段，导致了“数据爆炸但知识贫乏”的现象。

1.3 支持数据挖掘技术的基础

数据挖掘技术是人们长期对数据库技术进行研究和开发的结果。起初各种商业数据是存储在计算机的数据库中的，然后发展到可对数据库进行查询和访问，进而 发展到对数据库的即时遍历。数据挖掘使数据库技术进入了一个更高级的阶段，它不仅能对过去的数据进行查询和遍历，并且能够找出过去数据之间的潜在联系，从 而促进信息的传递。现在数据挖掘技术在商业应用中已经可以马上投入使用，因为对这种技术进行支持的三种基础技术已经发展成熟，他们是：

- - 海量数据搜集
- - 强大的多处理器计算机
- - 数据挖掘算法

Friedman[1997]列举了四个主要的技术理由激发了数据挖掘的开发、应用和研究的兴趣：

- - 超大规模数据库的出现，例如商业数据仓库和计算机自动收集的数据记录；
- - 先进的计算机技术，例如更快和更大的计算能力和并行体系结构；
- - 对巨大量数据的快速访问；
- - 对这些数据应用精深的统计方法计算的能力。

商业数据库现在正在以一个空前的速度增长，并且数据仓库正在广泛地应用于各种行业；对计算机硬件性能越来越高的要求，也可以用现在已经成熟的并行多处 理机的技术来满足；另外数据挖掘算法经过了这 10 多年的发展也已经成为一种成熟，稳定，且易于理解和操作的技术。

1.4 从商业数据到商业信息的进化

从商业数据到商业信息的进化过程中，每一步前进都是建立在上一步的基础上的。见下表。表中我们可以看到，第四步进化是革命性的，因为从用户的角度来看，这一阶段的数据库技术已经可以快速地回答商业上的很多问题了。

进化阶段	商业问题	支持技术	产品厂家	产品特点
数据搜集(60 年代)	“过去五年中我的总收入是多少？”	计算机、磁带和磁盘	IBM,CDC	提供历史性的、静态的数据信息
数据访问(80 年代)	“在新英格兰的分部去年三月的销售额是多少？”	关系数据库（RDBMS），结构化查询语言（SQL），ODBC Oracle 、 Sybase 、 Informix 、 IBM 、 Microsoft	Oracle 、 Sybase 、 Informix 、 IBM 、 Microsoft	在记录级提供历史性的、动态数据信息
数据仓库；决策支持(90 年代)	“在新英格兰的分部去年三月的销售额是多少？波士顿据此可得出什么结论？”	联机分析处理（OLAP）、多维数据库、数据仓库	Pilot、Comshare、Arbor 、 Cognos 、 Microstrategy	在各种层次上提供回溯的、动态的数据信息
数据挖掘（正在流行）	“下个月波士顿的销售会怎么样？为什么？”	高级算法、多处理器计算机、海量数据库	Pilot 、 Lockheed 、 IBM、SGI、其他初创公司	提供预测性的信息

表一、数据挖掘的进化历程。

数据挖掘的核心模块技术历经了数十年的发展，其中包括数理统计、人工智能、机器学习。今天，这些成熟的技术，加上高性能的关系数据库引擎以及广泛的数据集成，让数据挖掘技术在当前的数据仓库环境中进入了实用的阶段。

1.5 数据挖掘逐渐演变的过程

数据挖掘其实是一个逐渐演变的过程,电子数据处理的初期,人们就试图通过某些方法来实现自动决策支持,当时机器学习成为人们关心的焦点.机器学习的过 程就是将一些已知的并已被成功解决的问题作为范例输入计算机,机器通过学习这些范例总结并生成相应的规则,这些规则具有通用性,使用它们可以解决某一类的 问题.随后,随着神经网络技术的形成和发展,人们的注意力转向知识工程,知识工程不同于机器学习那样给计算机输入范例，让它生成出规则，而是直接给计算机 输入已被代码化的规则，而计算机是通过使用这些规则来解决某些问题。专家系统就是这种方法所得到的成果，但它有投资大、效果不甚理想等不足。80 年代人们 又在新的神经网络理论的指导下，重新回到机器学习的方法上，并将其成果应用于处理大型商业数据库。随着在 80 年代末一个新的术语，它就是数据库中的知识发 现，简称KDD(Knowledge discovery in database).它泛指所有从源数据中发掘模式或联系的方法，人们接受了这个术语，并用KDD来描述整个数据发掘的过程，包括最开始的制定业务目标到 最终的结果分析，而用数据挖掘（data mining）来描述使用挖掘算法进行数据挖掘的子过程。但最近人们却逐渐开始使用数据挖掘中有许多工作可以由统计方法来完成，并认为最好的策略是将统计 方法与数据挖掘有机的结合起来。

数据仓库技术的发展与数据挖掘有着密切的关系。数据仓库的发展是促进数据挖掘越来越热的原因之一。但是，数据仓库并不是数据挖掘的先决条件，因为有很多数据挖掘可直接从操作数据源中挖掘信息。

第二课 数据挖掘的定义

2.1 技术上的定义及含义

数据挖掘（Data Mining）就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。

与　　数据挖掘相近的同义词有数据融合、数据分析和决策支持等。这个定义包括好几层含义：数据源必须是真实的、大量的、含噪声的；发现的是用户感兴趣的知 识；发现的知识要可接受、可理解、可运用；并不要求发现放之四海皆准的知识，仅支持特定的发现问题。

----何为知识?从广义上理解，数据、信息也是知识的表现形式，但是人们更把概念、规则、模式、规律和约束等看作知识。人们把数据看作是形成知识的源泉，好像从矿石中采矿或淘金一样。原始数据可以是结构化的，如关系数据库中的数据；也可以是半结构化的，如文本、图形和图像数据；甚至是分布在网络上的 异构型数据。发现知识的方法可以是数学的，也可以是非数学的；可以是演绎的，也可以是归纳的。发现的知识可以被用于信息管理，查询优化，决策支持和过程控 制等，还可以用于数据自身的维护。因此，数据挖掘是一门交叉学科，它把人们对数据的应用从低层次的简单查询，提升到从数据中挖掘知识，提供决策支持。在这 种需求牵引下，汇聚了不同领域的研究者，尤其是数据库技术、人工智能技术、数理统计、可视化技术、并行计算等方面的学者和工程技术人员，投身到数据挖掘这 一新兴的研究领域，形成新的技术热点。

这里所说的知识发现，不是要求发现放之四海而皆准的真理，也不是要去发现崭新的自然科学定理和纯数学公式，更不是什么机器定理证明。实际上，所有发现 的知识都是相对的，是有特定前提和约束条件，面向特定领域的，同时还要能够易于被用户理解。最好能用自然语言表达所发现的结果。

2.2 商业角度的定义

数据挖掘是一种新的商业信息处理技术，其主要特点是对商业数据库中的大量业务数据进行抽取、转换、分析和其他模型化处理，从中提取辅助商业决策的关键性数据。

简而言之，数据挖掘其实是一类深层次的数据分析方法。数据分析本身已经有很多年的历史，只不过在过去数据收集和分析的目的是用于科学研究，另外，由于 当时计算能力的限制，对大数据量进行分析的复杂数据分析方法受到很大限制。现在，由于各行业业务自动化的实现，商业领域产生了大量的业务数据，这些数据不 再是为了分析的目的而收集的，而是由于纯机会的(Opportunistic)

商业运作而产生。分析这些数据也不再是单纯为了研究的需要，更主要是为商业决策提供真正有价值的信息，进而获得利润。但所有企业面临的一个共同问题是：企业数据量非常大，而其中真正有价值的信息却很少，因此从大量的数据中经过深 层分析，获得有利于商业运作、提高竞争力的信息，就像从矿石中淘金一样，数据挖掘也因此而得名。

因此，数据挖掘可以描述为：按企业既定业务目标，对大量的企业数据进行探索和分析，揭示隐藏的、未知的或验证已知的规律性，并进一步将其模型化的先进有效的方法。

2.3 数据挖掘与传统分析方法的区别

数据挖掘与传统的数据分析(如查询、报表、联机应用分析)的本质区别是数据挖掘是在没有明确假设的前提下去挖掘信息、发现知识.数据挖掘所得到的信息应具有先未知,有效和可实用三个特征.

先前未知的信息是指该信息是预先未曾预料到的,既数据挖掘是要发现那些不能靠直觉发现的信息或知识,甚至是违背直觉的信息或知识,挖掘出的信息越是出 乎意料,就可能越有价值.在商业应用中最典型的例子就是一家连锁店通过数据挖掘发现了小孩尿布和啤酒之间有着惊人的联系.

2.4 数据挖掘和数据仓库

大部分情况下,数据挖掘都要先把数据从数据仓库中拿到数据挖掘库或数据集市中（见图 1）。从数据仓库中直接得到进行数据挖掘的数据有许多好处。就如我 们后面会讲到的，数据仓库的数据清理和数据挖掘的数据清理差不多，如果数据在导入数据仓库时已经清理过,那很可能在做数据挖掘时就没必要在清理一次了，而 且所有的数据不一致的问题都已经被你解决了。

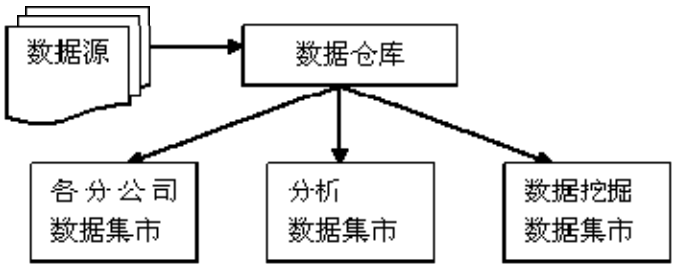


图 1：数据挖掘库从数据仓库中得出

数据挖掘库可能是你的数据仓库的一个逻辑上的子集,而不一定非得是物理上单独的数据库。但如果你的数据仓库的计算资源已经很紧张，那你最好还是建立一个单独的数据挖掘库。

当然为了数据挖掘你也不必非得建立一个数据仓库，数据仓库不是必需的。建立一个巨大的数据仓库，把各个不同源的数据统一在一起，解决所有的数据冲突问 题，然后把所有的数据导到一个数据仓库内，是一项巨大的工程，可能要用几年的时间花上百万的钱才能完成。只是为了数据挖掘，你可以把一个或几个事务数据库 导到一个只读的数据库中，就把它当作数据集市，然后在他上面进行数据挖掘。

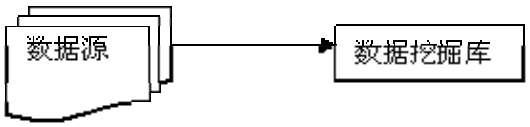


图 2：数据挖掘库从事务数据库中得出

2.5 数据挖掘和在线分析处理（OLAP）

一个经常问的问题是，数据挖掘和OLAP到底有何不同。下面将会解释，他们是完全不同的工具，基于的技术也大相径庭。

OLAP是决策支持领域的一部分。传统的查询和报表工具是告诉你数据库中都有什么（what happened），OLAP则更进一步告诉你下一步会怎么样（What next）、和如果我采取这样的措施又会怎么样（What if）。用户首先建立一个假设，然后用OLAP检索数据库来验证这个假设是否正确。比如，一个分析师想找到什么原因导致了贷款拖欠，他可能先做一个初始的 假定，认为低收入的人信用度也低，然后用OLAP来验证他这个假设。如果这个假设没有被证实，他可能去察看那些高负债的账户，如果还不行，他也许要把收入 和负债一起考虑，一直进行下去，直到找到他想要的结果或放弃。

也就是说，OLAP分析师是建立一系列的假设，然后通过OLAP来证实或推翻这些假设来最终得到自己的结论。OLAP分析过程在本质上是一个演绎推理 的过程。但是如果分析的变量达到几十或上百个，那么再用OLAP手动分析验证这些假设将是一件非常困难和痛苦的事情。

数据挖掘与OLAP不同的地方是，数据挖掘不是用于验证某个假定的模式（模型）的正确性，而是在数据库中自己寻找模型。他在本质上是一个归纳的过程。比 如，一个用数据挖掘工具的分析师想找到引起贷款拖欠的风险因素。数据挖掘工具可能帮他找到高负债和低收入是引起这个问题的因素，甚至还可能发现一些分析师从来没有想过或试过的其他因素，比如年龄。

数据挖掘和OLAP具有一定的互补性。在利用数据挖掘出来的结论采取行动之前，你也许要验证一下如果采取这样的行动会给公司带来什么样的影响，那么OLAP工具能回答你的这些问题。

而且在知识发现的早期阶段，OLAP工具还有其他一些用途。可以帮你探索数据，找到哪些是对一个问题比较重要的变量，发现异常数据和互相影响的变量。这都能帮你更好的理解你的数据，加快知识发现的过程。

2.6 数据挖掘，机器学习和统计

数据挖掘利用了人工智能（AI）和统计分析的进步所带来的好处。这两门学科都致力于模式发现和预测。

数据挖掘不是为了替代传统的统计分析技术。相反，他是统计分析方法学的延伸和扩展。大多数的统计分析技术都基于完善的数学理论和高超的技巧，预测的准 确度还是令人满意的，但对使用者的要求很高。而随着计算机计算能力的不断增强，我们有可能利用计算机强大的计算能力只通过相对简单和固定的方法完成同样的 功能。

一些新兴的技术同样在知识发现领域取得了很好的效果，如神经元网络和决策树，在足够多的数据和计算能力下，他们几乎不用人的关照自动就能完成许多有价值的功能。

数据挖掘就是利用了统计和人工智能技术的应用程序，他把这些高深复杂的

技术封装起来，使人们不用自己掌握这些技术也能完成同样的功能，并且更专注于自己所要解决的问题。

2.7 软硬件发展对数据挖掘的影响

使数据挖掘这件事情成为可能的关键一点是计算机性能价格比的巨大进步。在过去的几年里磁盘存储器的价格几乎降低了 99%，这在很大程度上改变了企业界 对数据收集和存储的态度。如果每兆的价格是￥10，那存放 1TB的价格是￥10,000,000,但当每兆的价格降为 1 毛钱时,存储同样的数据只有 ￥100,000!

计算机计算能力价格的降低同样非常显著。每一代芯片的诞生都会把CPU的计算能力提高一大步。内存RAM也同样降价迅速，几年之内每兆内存的价格由几百块钱降到现在只要几块钱。通常PC都有 64M内存，工作站达到了 256M，拥有上G内存的服务器已经不是什么新鲜事了。

在单个CPU计算能力大幅提升的同时,基于多个CPU的并行系统也取得了很大的进步。目前几乎所有的服务器都支持多个CPU，这些SMP服务器簇甚至能让成百上千个CPU同时工作。

基于并行系统的数据库管理系统也给数据挖掘技术的应用带来了便利。如果你有一个庞大而复杂的数据挖掘问题要求通过访问数据库取得数据,那么效率最高的办法就是利用一个本地的并行数据库。

所有这些都为数据挖掘的实施扫清了道路，随着时间的延续，我们相信这条道路会越来越平坦。

第三课 数据挖掘的研究历史和现状

3.1 研究历史

从数据库中发现知识（KDD）一词首次出现在 1989 年举行的第十一届国际联合人工智能学术会议上。到目前为止，由美国人工智能协会主办的KDD国际 研讨会已经召开了 8 次，规模由原来的专题讨论会发展到国际学术大会（见表 1），研究重点也逐渐从发现方法转向系统应用，注重多种发现策略和技术的集成，以及多种学科之间的相互渗透。1999 年，亚太地区在北京召开的第三届PAKDD会议收到 158 篇论文，空前热烈。IEEE的Knowledge and Data Engineering会刊率先在 1993 年出版了KDD技术专刊。并行计算、计算机网络和信息工程等其他领域的国际学会、学刊也把数据挖掘和知识发现列 为专题和专刊讨论，甚至到了脍炙人口的程度。

表1 历届KDD国际学术会议一览表

时间	会议名称	会议地址	收录论文比例数	参加会议人数
1989.6	Workshop on KDD	Detroit, Michigan, USA	2:1	30
1991.7	Workshop on KDD	Anaheim, California	3.5:1	46
1993.7	Workshop on KDD	Washington, USA	3:1	40
1995	KDD95	Montreal, Canada	4.5:1	340
1996.8	KDD96	Portland, Oregon, USA	5:1	450
1997.2	PAKDD97	Singapore	3.5:1	97
1997.8	KDD97	California, USA	6:1	600
1998.4	PAKDD98	Melbourne, Australia	110:31	120
1998.8	KDD98	New York, USA	247:68	773
1999.4	PAKDD99	Beijing, China	158:66	150
1999.8	KDD99	San Diego, CA	280:27	600

3.2 出版物及工具

此外，在Internet上还有不少KDD电子出版物，其中以半月刊Knowledge Discovery Nuggets最为权威（http://www.kdnuggets.com/subscribe.html）。在网上还有许多自由论坛，如DM Email Club等。至于DMKD书籍，可以在任意一家计算机书店找到十多本。目前，世界上比较有影响的典型数据挖掘系统有：SAS公司的Enterprise Miner、IBM公司的Intelligent Miner、SGI公司的SetMiner、SPSS公司的Clementine、Sybase公司的Warehouse Studio、RuleQuest Research公司的See5、还有CoverStory、EXPLORA、Knowledge Discovery Workbench、DBMiner、Quest等。读者可以访问http://www.datamininglab.com.网站，该网站提供了许多数 据挖掘系统和工具的性能测试报告。

3.3 国内现状

与国外相比，国内对DMKD的研究稍晚，没有形成整体力量。1993 年国家自然科学基金首次支持我们对该领域的研究项目。目前，国内的许多科研单位和 高等院校竞相开展知识发现的基础理论及其应用研究，这些单位包括清华大学、中科院计算技术研究所、空军第三研究所、海军装备论证中心等。其中，北京系统工 程研究所对模糊方法在知识发现中的应用进行了较深入的研究，北京大学也在开展对数据立方体代数的研究，华中理工大学、复旦大学、浙江大学、中国科技大学、 中科院数学研究所、吉林大学等单位开展了对关联规则开采算法的优化和改造；南京大学、四川联合大学和上海交通大学等单位探讨、研究了非结构化数据的知识发 现以及Web数据挖掘。

3.4 国内现状

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。”

数据挖掘研究内容和本质

----随着DMKD研究逐步走向深入，数据挖掘和知识发现的研究已经形成了三根强大的技术支柱：数据库、人工智能和数理统计。因此，KDD大会程序委 员会曾经由这三个学科的权威人物同时来任主席。目前DMKD的主要研究内容包括基础理论、发现算法、数据仓库、可视化技术、定性定量互换模型、知识表示方 法、发现知识的维护和再利用、半结构化和非结构化数据中的知识发现以及网上数据挖掘等。

----数据挖掘所发现的知识最常见的有以下四类：

4.1 广义知识（Generalization）

----广义知识指类别特征的概括性描述知识。根据数据的微观特性发现其表征的、带有普遍性的、较高层次概念的、中观和宏观的知识，反映同类事物共同性质，是对数据的概括、精炼和抽象。

----广义知识的发现方法和实现技术有很多，如数据立方体、面向属性的归约等。数据立方体还有其他一些别名，如“多维数据库”、“实现视图”、“OLAP”等。该方法的基本思想是实现某些常用的代价较高的聚集函数的计算，诸如计数、求和、平均、最大值等，并将这些实现视图储存在多维数据库中。既然很多聚集函数需经常重复计算，那么在多维数据立方体中存放预先计算好的结果将能保证快速响应，并可灵活地提供不同角度和不同抽象层次上的数据视图。另一 种广义知识发现方法是加拿大SimonFraser大学提出的面向属性的归约方法。这种方法以类SQL语言表示数据挖掘查询，收集数据库中的相关数据集， 然后在相关数据集上应用一系列数据推广技术进行数据推广，包括属性删除、概念树提升、属性阈值控制、计数及其他聚集函数传播等。

4.2 关联知识（Association）

----它反映一个事件和其他事件之间依赖或关联的知识。如果两项或多项属性之间存在关联，那么其中一项的属性值就可以依据其他属性值进行预测。最为著名的关联规则发现方法是R.Agrawal提出的Apriori算法。关联规则的发现可分为两步。第一步是迭代识别所有的频繁项目集，要求频繁项目集的 支持率不低于用户设定的最低值；第二步是从频繁项目集中构造可信度不低于用户设定的最低值的规则。识别或发现所有频繁项目集是关联规则发现算法的核心，也 是计算量最大的部分。

4.3 分类知识(Classification&Clustering)

----它反映同类事物共同性质的特征型知识和不同事物之间的差异型特征知识。最为典型的分类方法是基于决策树的分类方法。它是从实例集中构造决策树，是一种有指导的学习方法。该方法先根据训练子集（又称为窗口）形成决策树。如果该树不能对所有对象给出正确的分类，那么选择一些例外加入到窗口中，重 复该过程一直到形成正确的决策集。最终结果是一棵树，其叶结点是类名，

中间结点是带有分枝的属性，该分枝对应该属性的某一可能值。最为典型的决策树学习系 统是ID3，它采用自顶向下不回溯策略，能保证找到一个简单的树。算法C4.5 和C5.0 都是ID3 的扩展，它们将分类领域从类别属性扩展到数值型属性。

----数据分类还有统计、粗糙集（RoughSet）等方法。线性回归和线性辨别分析是典型的统计模型。为降低决策树生成代价，人们还提出了一种区间分类器。最近也有人研究使用神经网络方法在数据库中进行分类和规则提取。

4.4 预测型知识（Prediction）

----它根据时间序列型数据，由历史的和当前的数据去推测未来的数据，也可以认为是以时间为关键属性的关联知识。

----目前，时间序列预测方法有经典的统计方法、神经网络和机器学习等。1968 年Box和Jenkins提出了一套比较完善的时间序列建模理论和 分析方法，这些经典的数学方法通过建立随机模型，如自回归模型、自回归滑动平均模型、求和自回归滑动平均模型和季节调整模型等，进行时间序列的预测。由于 大量的时间序列是非平稳的，其特征参数和数据分布随着时间的推移而发生变化。因此，仅仅通过对某段历史数据的训练，建立单一的神经网络预测模型，还无法完 成准确的预测任务。为此，人们提出了基于统计学和基于精确性的再训练方法，当发现现存预测模型不再适用于当前数据时，对模型重新训练，获得新的权重参数，建立新的模型。也有许多系统借助并行算法的计算优势进行时间序列预测。

4.5 偏差型知识(Deviation)

----此外，还可以发现其他类型的知识，如偏差型知识(Deviation)，它是对差异和极端特例的描述，揭示事物偏离常规的异常现象，如标准类 外的特例，数据聚类外的离群值等。所有这些知识都可以在不同的概念层次上被发现，并随着概念层次的提升，从微观到中观、到宏观，以满足不同用户不同层次决 策的需要。

第五课 数据挖掘的功能

数据挖掘通过预测未来趋势及行为，做出前摄的、基于知识的决策。数据挖掘的目标是从数据库中发现隐含的、有意义的知识，主要有以下五类功能。

5.1 自动预测趋势和行为

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。

5.2 关联分析

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。

5.3 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。

5.4 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。

5.5 偏差检测

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。

第六课 数据挖掘常用技术

6.1 人工神经网络

神经网络近来越来越受到人们的关注，因为它为解决大复杂度问题提供了一种相对来说比较有效的简单方法。神经网络可以很容易的解决具有上百个参数的问题 （当然实际生物体中存在的神经网络要比我们这里所说的程序模拟的神经网络要复杂的多）。神经网络常用于两类问题：分类和回归。

在结构上，可以把一个神经网络划分为输入层、输出层和隐含层（见图 4）。输入层的每个节点对应一个个的预测变量。输出层的节点对应目标变量，可有多个。在输入层和输出层之间是隐含层（对神经网络使用者来说不可见），隐含层的层数和每层节点的个数决定了神经网络的复杂度。

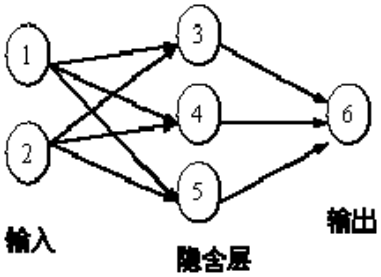


图 4：一个神经元网络

除了输入层的节点，神经网络的每个节点都与很多它前面的节点（称为此节点的输入节点）连接在一起，每个连接对应一个权重 Wxy，此节点的值就是通过它 所有输入节点的值与对应连接权重乘积的和作为一个函数的输入而得到，我们把这个函数称为活动函数或挤压函数。如图 5 中节点 4 输出到节点 6 的值可通过如下计 算得到：

W14*节点 1 的值+W24*节点 2 的值

神经网络的每个节点都可表示成预测变量（节点 1，2）的值或值的组合（节点 3-6）。注意节点 6 的值已经不再是节点 1、2 的线性组合，因为数据在隐含 层中传递时使用了活动函数。实际上如果没有活动函数的话，神经元网络就等价于一个线性回归函数，如果此活动函数是某种特定的非线性函数，那神经网络又等价 于逻辑回归。

调整节点间连接的权重就是在建立（也称训练）神经网络时要做的工作。最早的也是最基本的权重调整方法是错误回馈法，现在较新的有变化坡度法、类牛顿 法、Levenberg-Marquardt 法、和遗传算法等。无论采用那种训练方法，都需要有一些参数来控制训练的过程，如防止训练过度和控制训练的速 度。

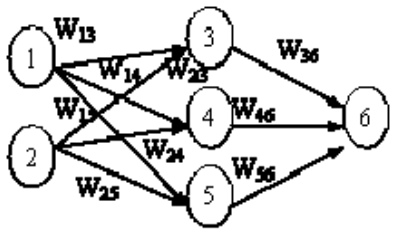


图 5：带权重 W_{xy} 的神经元网络

决定神经网络拓扑结构（或体系结构）的是隐含层及其所含节点的个数，以及节点之间的连接方式。要从头开始设计一个神经网络，必须要决定隐含层和节点的数目，活动函数的形式，以及对权重做那些限制等，当然如果采用成熟软件工具的话，他会帮你决定这些事情。

在诸多类型的神经网络中，最常用的是前向传播式神经网络，也就是我们前面图示中所描绘的那种。我们下面详细讨论一下，为讨论方便假定只含有一层隐含节点。

可以认为错误回馈式训练法是变化坡度法的简化，其过程如下：

前向传播：数据从输入到输出的过程是一个从前向后的传播过程，后一节点的值通过它前面相连的节点传过来，然后把值按照各个连接权重的大小加权输入活动函数再得到新的值，进一步传播到下一个节点。

回馈：当节点的输出值与我们预期的值不同，也就是发生错误时，神经网络就要 “学习”（从错误中学习）。我们可以把节点间连接的权重看成后一节点对前一节点的“信任” 程度（他自己向下一节点的输出更容易受他前面哪个节点输入的影响）。学习的方法是采用惩罚的方法，过程如下：如果一节点输出发生错误，那么他看他的错误是 受哪个（些）输入节点的影响而造成的，是不是他最信任的节点（权重最高的节点）陷害了他（使他出错），如果是则要降低对他的信任值（降低权重），惩罚他 们，同时升高那些做出正确建议节点的信任值。对那些收到惩罚的节点来说，他也需要用同样的方法来进一步惩罚它前面的节点。就这样把惩罚一步步向前传播直到 输入节点为止。

对训练集中的每一条记录都要重复这个步骤，用前向传播得到输出值，如果发生错误，则用回馈法进行学习。当把训练集中的每一条记录都运行过一遍之后，我 们称完成一个训练周期。要完成神经网络的训练可能需要很多个训练周期，经常是几百个。训练完成之后得到的神经网络就是在通过训练集发现的模型，描述了训练 集中响应变量受预测变量影响的变化规律。

由于神经网络隐含层中的可变参数太多，如果训练时间足够长的话，神经网络很可能把训练集的所有细节信息都“记”下来，而不是建立一个忽略细节只具有规 律性的模型，我们称这种情况为训练过度。显然这种“模型”对训练集会有很高的准确率，而一旦离开训练集应用到其他数据，很可能准确度急剧下降。为了防止这 种训练过度的情况，我们必须知道在什么时候要停止训练。在有些软件实现中会在训练的同时用一个测试集来计算神经网络在此测试集上的正确率，一旦这个正确率 不再升高甚至开始下降时，那么就认为现在神经网络已经达到做好的状态了可以停止训练。

图 6 中的曲线可以帮我们理解为什么利用测试集能防止训练过度的出现。在图中可以看到训练集和测试集的错误率在一开始都随着训练周期的增加不断降低，而 测试集的错误率在达到一个谷底后反而开始上升，我们认为这个开始上升的时刻就是应该停止训练的时刻。

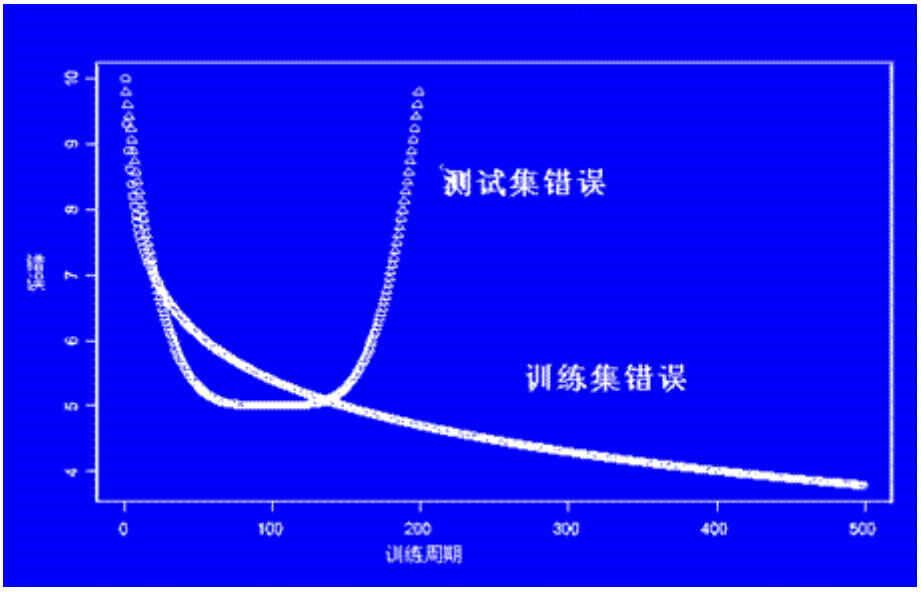


图 6：神经网络在训练周期增加时准确度的变化情况

神经元网络和统计方法在本质上有很多差别。神经网络的参数可以比统计方法多很多。如图 4 中就有 13 个参数（9 个权重和 4 个限制条件）。由于参数如此之 多，参数通过各种各样的组合方式来影响输出结果，以至于很难对一个神经网络表示的模型做出直观的解释。实际上神经网络也正是当作“黑盒”来用的，不用去管 “盒子”里面是什么，只管用就行了。在大部分情况下，这种限制条件是可以接受的。比如银行可能需要一个笔迹识别软件，但他没必要知道为什么这些线条组合在 一起就是一个人的签名，而另外一个相似的则不是。在很多复杂度很高的问题如化学试验、机器人、金融市场的模拟、和语言图像的识别，等领域神经网络都取得了 很好的效果。

神经网络的另一个优点是很容易在并行计算机上实现，可以把他的节点分配到不同的 CPU 上并行计算。

在使用神经网络时有几点需要注意：第一，神经网络很难解释，目前还没有能对神经网络做出显而易见解释的方法学。

第二，神经网络会学习过度，在训练神经网络时一定要恰当的使用一些能严格衡量神经网络的方法，如前面提到的测试集方法和交叉验证法等。这主要是由于神经网络太灵活、可变参数太多，如果给足够的时间，他几乎可以“记住”任何事情。

第三，除非问题非常简单，训练一个神经网络可能需要相当可观的时间才能完成。当然，一旦神经网络建立好了，在用它做预测时运行时还是很快得。

第四，建立神经网络需要做的数据准备工作量很大。一个很有误导性的神话就是不管用什么数据神经网络都能很好的工作并做出准确的预测。这是不确切的，要 想得到准确度高的模型必须认真的进行数据清洗、整理、转换、选择等工作，对任何数据挖掘技术都是这样，神经网络尤其注重这一点。比如神经网络要求所有的输 入变量都必须是 0-1（或-1 -- +1）之间的实数，因此像“地区”之类文本数据必须先做必要的处理之后才能用作神经网络的输入。

返回顶部

6.2 决策树

决策树提供了一种展示类似在什么条件下会得到什么值这类规则的方法。比如，在贷款申请中，要对申请的风险大小做出判断，图 7 是为了解决这个问题而建立的一棵决策树，从中我们可以看到决策树的基本组成部分：决策节点、分支和叶子。

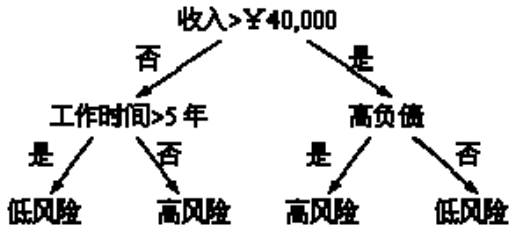


图 7：一棵简单的决策树

决策树中最上面的节点称为根节点，是整个决策树的开始。本例中根节点是“收入>￥40,000”，对此问题的不同回答产生了“是”和“否”两个分支。

决策树的每个节点子节点的个数与决策树在用的算法有关。如CART算法得到的决策树每个节点有两个分支，这种树称为二叉树。允许节点含有多于两个子节点的树称为多叉树。

每个分支要么是一个新的决策节点，要么是树的结尾，称为叶子。在沿着决策树从上到下遍历的过程中，在每个节点都会遇到一个问题，对每个节点上问题的不 同回答导致不同的分支，最后会到达一个叶子节点。这个过程就是利用决策树进行分类的过程，利用几个变量（每个变量对应一个问题）来判断所属的类别（最后每 个叶子会对应一个类别）。

假如负责借贷的银行官员利用上面这棵决策树来决定支持哪些贷款和拒绝哪些贷款，那么他就可以用贷款申请表来运行这棵决策树，用决策树来判断风险的大 小。“年收入>￥40,00”和“高负债”的用户被认为是“高风险”，同时“收入<￥40,000”但“工作时间>5 年”的申请，则被 认为“低风险”而建议贷款给他/她。

数据挖掘中决策树是一种经常要用到的技术，可以用于分析数据，同样也可以用来作预测(就像上面的银行官员用他来预测贷款风险)。常用的算法有CHAID、CART、 Quest 和C5.0。

建立决策树的过程，即树的生长过程是不断的把数据进行切分的过程，每次切分对应一个问题，也对应着一个节点。对每个切分都要求分成的组之间的“差异”最大。

各种决策树算法之间的主要区别就是对这个“差异”衡量方式的区别。对具体衡量方式算法的讨论超出了本文的范围，在此我们只需要把切分看成是把一组数据 分成几份，份与份之间尽量不同，而同一份内的数据尽量相同。这个切分的过程也可称为数据的“纯化”。看我们的例子，包含两个类别--低风险和高风险。如果 经过一次切分后得到的分组，每个分组中的数据都属于同一个类别，显然达到这样效果的切分方法就是我们所追求的。

到现在为止我们所讨论的例子都是非常简单的，树也容易理解，当然实际中应用的决策树可能非常复杂。假定我们利用历史数据建立了一个包含几百个属性、输 出的类有十几种的决策树，这样的一棵树对人来说可能太复杂了，但每一条从根结点到叶子节点的路径所描述的含义仍然是可以理解的。决策树的这种易理解性对数 据挖掘的使用者来说是一个显著的优点。

然而决策树的这种明确性可能带来误导。比如，决策树每个节点对应分割的定义都是非常明确毫不含糊的，但在实际生活中这种明确可能带来麻烦（凭什么说年收入￥40,001 的人具有较小的信用风险而￥40,000 的人就没有）。

建立一颗决策树可能只要对数据库进行几遍扫描之后就能完成，这也意味着需要的计算资源较少，而且可以很容易的处理包含很多预测变量的情况，因此决策树模型可以建立得很快，并适合应用到大量的数据上。

对最终要拿给人看的决策树来说，在建立过程中让其生长的太“枝繁叶茂”是没有必要的，这样既降低了树的可理解性和可用性，同时也使决策树本身对历史数 据的依赖性增大，也就是说这是这棵决策树对此历史数据可能非常准确，一旦应用到新的数据时准确性却急剧下降，我们称这种情况为训练过度。为了使得到的决策 树所蕴含的规则具有普遍意义，必须防止训练过度，同时也减少了训练的时间。因此我们需要有一种方法能让我们在适当的时候停止树的生长。常用的方法是设定决 策树的最大高度（层数）来限制树的生长。还有一种方法是设定每个节点必须包含的最少记录数，当节点中记录的个数小于这个数值时就停

止分割。

与设置停止增长条件相对应的是在树建立好之后对其进行修剪。先允许树尽量生长，然后再把树修剪到较小的尺寸，当然在修剪的同时要求尽量保持决策树的准确度尽量不要下降太多。

对决策树常见的批评是说其在为一个节点选择怎样进行分割时使用“贪心”算法。此种算法在决定当前这个分割时根本不考虑此次选择会对将来的分割造成什么 样的影响。换句话说，所有的分割都是顺序完成的，一个节点完成分割之后不可能以后再有机会回过头来再考察此次分割的合理性，每次分割都是依赖于他前面的分 割方法，也就是说决策树中所有的分割都受根结点的第一次分割的影响，只要第一次分割有一点点不同，那么由此得到的整个决策树就会完全不同。那么是否在选择 一个节点的分割的同时向后考虑两层甚至更多的方法，会具有更好的结果呢？目前我们知道的还不是很清楚，但至少这种方法使建立决策树的计算量成倍的增长，因 此现在还没有哪个产品使用这种方法。

而且，通常的分割算法在决定怎么在一个节点进行分割时，都只考察一个预测变量，即节点用于分割的问题只与一个变量有关。这样生成的决策树在有些本应很 明确的情况下可能变得复杂而且意义含混，为此目前新提出的一些算法开始在一个节点同时用多个变量来决定分割的方法。比如以前的决策树中可能只能出现类似 “收入<￥35,000”的判断，现在则可以用“收入<(0.35*抵押)”或“收入>￥35,000 或抵押<150,000” 这样的问题。

决策树很擅长处理非数值型数据，这与神经网络只能处理数值型数据比起来，就免去了很多数据预处理工作。

甚至有些决策树算法专为处理非数值型数据而设计，因此当采用此种方法建立决策树同时又要处理数值型数据时，反而要做把数值型数据映射到非数值型数据的预处理。

6.3 遗传算法

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。

6.4 近邻算法

将数据集合中每一个记录进行分类的方法。

6.5 规则推导

从统计意义上对数据中的“如果-那么”规则进行寻找和推导。

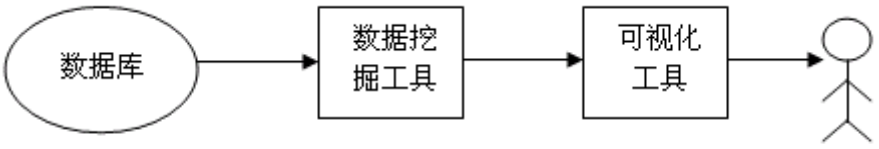
采用上述技术的某些专门的分析工具已经发展了大约十年的历史，不过这些工具所面对的数据量通常较小。而现在这些技术已经被直接集成到许多大型的工业标准的数据仓库和联机分析系统中去了。

第七课 数据挖掘的流程

7.1 数据挖掘环境

数据挖掘是指一个完整的过程,该过程从大型数据库中挖掘先前未知的,有效的,可实用的信息,并使用这些信息做出决策或丰富知识.

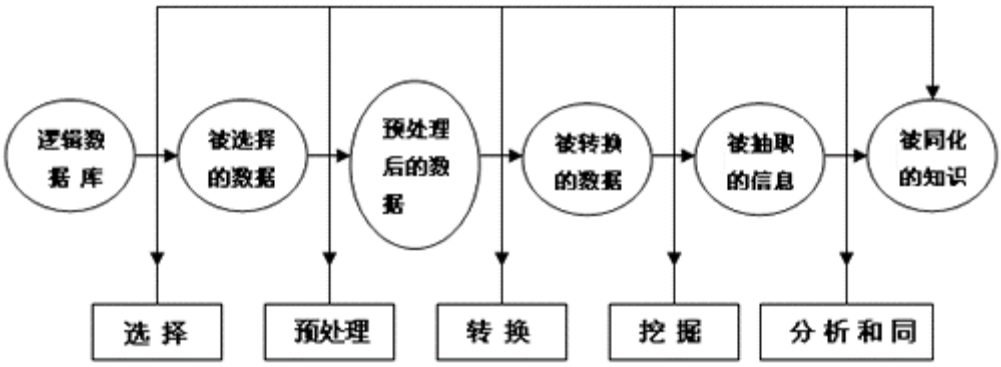
数据挖掘环境可示意如下图:



数据挖掘环境框图

7.2 数据挖掘过程图

下图描述了数据挖掘的基本过程和主要步骤



数据挖掘的基本过程和主要步骤

7.3 数据挖掘过程工作量

在数据挖掘中被研究的业务对象是整个过程的基础,它驱动了整个数据挖掘过程,也是检验最后结果和指引分析人员完成数据挖掘的依据和顾问.图 2 各步骤是 按一定顺序完成的,当然整个过程中还会存在步骤间的反馈.数据挖掘的过程并不是自动的,绝大多数的工作需要人工完成.图 3 给出了各步骤在整个过程中的工作 量之比.可以看到,60%的时间用在数据准备上,这说明了数据挖掘对数据的严格要求,而后挖掘工作仅占总工作量的 10%.

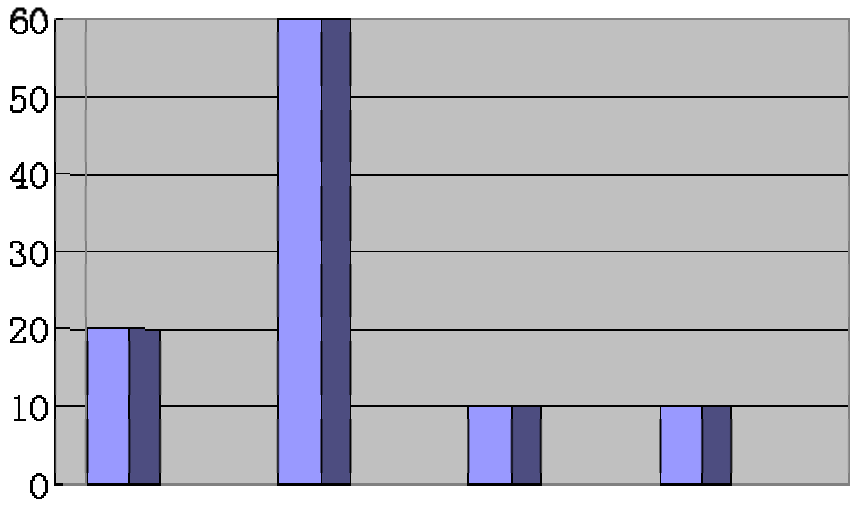


图 3 数据挖掘过程工作量比例

7.4 数据挖掘过程简介

过程中各步骤的大体内容如下:

1.1. 确定业务对象

清晰地定义出业务问题,认清数据挖掘的目的是数据挖掘的重要一步.挖掘的最后结构是不可预测的,但要探索的问题应是有预见的,为了数据挖掘而数据挖掘则带有盲目性,是不会成功的.

2.2. 数据准备

1)1) 数据的选择

搜索所有与业务对象有关的内部和外部数据信息,并从中选择出适用于数据挖掘应用的数据.

2)2) 数据的预处理

研究数据的质量,为进一步的分析作准备.并确定将要进行的挖掘操作的类型.

3)3) 数据的转换

将数据转换成一个分析模型.这个分析模型是针对挖掘算法建立的.建立一个真正适合挖掘算法的分析模型是数据挖掘成功的关键.

3.3. 数据挖掘

对所得到的经过转换的数据进行挖掘.除了完善从选择合适的挖掘算法外,其余一切工作都能自动地完成.

4.4. 结果分析

解释并评估结果.其使用的分析方法一般应作数据挖掘操作而定,通常会用到可视化技术.

5.5. 知识的同化

将分析所得到的知识集成到业务信息系统的组织结构中去.

7.5 数据挖掘需要的人员

数据挖掘过程的分步实现,不同的步会需要是有不同专长的人员,他们大体可以分为三类.

业务分析人员:要求精通业务,能够解释业务对象,并根据各业务对象确定出用于数据定义和挖掘算法的业务需求.

数据分析人员:精通数据分析技术,并对统计学有较熟练的掌握,有能力把业务需求转化为数据挖掘的各步操作,并为每步操作选择合适的技术.

数据管理人员:精通数据管理技术,并从数据库或数据仓库中收集数据.

从上可见,数据挖掘是一个多种专家合作的过程,也是一个在资金上和技术上高投入的过程.这一过程要反复进行牠在反复过程中,不断地趋近事物的本质,不 断地优先问题的解决方案。数据重组和细分添加和拆分记录选取数据样本可视化数据探索聚类分析神经网络、决策树数理统计、时间序列结论综合解释评价数据知识 数据取样数据探索数据调整模型化评价。

第八课 数据挖掘未来研究方向及热点

8．1 数据挖掘未来研究方向

----当前，DMKD研究方兴未艾，其研究与开发的总体水平相当于数据库技术在 70 年代所处的地位，迫切需要类似于关系模式、DBMS系统和SQL 查询语言等理论和方法的指导，才能使DMKD的应用得以普遍推广。预计在本世纪，DMKD的研究还会形成更大的高潮，研究焦点可能会集中到以下几个方面：

- 发现语言的形式化描述，即研究专门用于知识发现的数据挖掘语言，也许会像SQL语言一样走向形式化和标准化；
- 寻求数据挖掘过程中的可视化方法，使知识发现的过程能够被用户理解，也便于在知识发现的过程中进行人机交互；
- 研究在网络环境下的数据挖掘技术（WebMining），特别是在因特网上建立DMKD服务器，并且与数据库服务器配合，实现WebMining；
- 加强对各种非结构化数据的开采（DataMiningforAudio＆Video），如对文本数据、图形数据、视频图像数据、声音数据乃至综合多媒体数据的开采；

处理的数据将会涉及到更多的数据类型，这些数据类型或者比较复杂，或者是结构比较独特。为了处理这些复杂的数据，就需要一些新的和更好的分析和建立模型的方法，同时还会涉及到为处理这些复杂或独特数据所做的费时和复杂数据准备的一些工具和软件。

- 交互式发现；
- 知识的维护更新。

但是，不管怎样，需求牵引与市场推动是永恒的，DMKD将首先满足信息时代用户的急需，大量的基于DMKD的决策支持软件产品将会问世。

只有从数据中有效地提取信息，从信息中及时地发现知识，才能为人类的思维决策和战略发展服务。也只有到那时，数据才能够真正成为与物质、能源相媲美的资源，信息时代才会真正到来。

8．2 数据挖掘热点

就目前来看，将来的几个热点包括网站的数据挖掘（Web site data mining）、生物信息或基因（Bioinformatics/genomics）的数据挖掘及其文本的数据挖掘（Textual mining）。下面就这几个方面加以简单介绍。

8.2.1 网站的数据挖掘（Web site data mining）

需求

随着Web技术的发展，各类电子商务网站风起云涌，建立起一个电子商务网站并不困难，困难的是如何让您的电子商务网站有效益。要想有效益就必须吸引客 户，增加能带来效益的客户忠诚度。电子商务业务的竞争比传统的业务竞争更加激烈，原因有很多方面，其中一个因素是客户从一个电子商务网站转换到竞争对手那 边，只需点击几下鼠标即可。网站的内容和层次、用词、标题、奖励方案、服务等任何一个地方都有可能成为吸引客户、同时也可能成为失去客户的因素。而同时电 子商务网站每天都可能有上百万次的在线交易，生成大量的记录文件（Logfiles）和登记表，如何对这些数据进行分析和挖掘，充分了解客户的喜好、购买 模式，甚至是客户一时的冲动，设计出满足于不同客户群体需要的个性化网站，进而增加其竞争力，几乎变得势在必行。若想在竞争中生存进而获胜，就要比您的竞 争对手更了解客户。

电子商务网站数据挖掘

在对网站进行数据挖掘时，所需要的数据主要来自于两个方面：一方面是客户的背景信息，此部分信息主要来自于客户的登记表；而另外一部分数据主要来自浏 览者的点击流（Click-stream），此部分数据主要用于考察客户的行为表现。但有的时候，客户对自己的背景信息十分珍重，不肯把这部分信息填写在 登记表上，这就会给数据分析和挖掘带来不便。在这种情况之下，就不得不从浏览者的表现数据中来推测客户的背景信息，进而再加以利用。

就分析和建立模型的技术和算法而言，网站的数据挖掘和原来的数据挖掘差别并不是特别大，很多方法和分析思想都可以运用。所不同的是网站的数据格式有很 大一部分来自于点击流，和传统的数据库格式有区别。因而对电子商务网站进行数据挖掘所做的主要工作是数据准备。目前，有很多厂商正在致力于开发专门用于网 站挖掘的软件。

8.2.2 生物信息或基因的数据挖掘

生物信息或基因数据挖掘则完全属于另外一个领域，在商业上很难讲有多大的价值，但对于人类却受益非浅。例如，基因的组合千变万化，得某种病的人的基因 和正常人的基因到底差别多大？能否找出其中不同的地方，进而对其不同之处加以改变，使之成为正常基因？这都需要数据挖掘技术的支持。

对于生物信息或基因的数据挖掘和通常的数据挖掘相比，无论在数据的复杂程度、数据量还有分析和建立模型的算法而言，都要复杂得多。从分析算法上讲，更 需要一些新的和好的算法。现在很多厂商正在致力于这方面的研究。但就技术和软件而言，还远没有达到成熟的地步。

8.2.3 文本的数据挖掘（Textualmining）

人们很关心的另外一个话题是文本数据挖掘。举个例子，在客户服务中心，把同客户的谈话转化为文本数据，再对这些数据进行挖掘，进而了解客户对服务的满 意程度和客户的需求以及客户之间的相互关系等信息。从这个例子可以看出，无论是在数据结构还是在分析处理方法方面，文本数据挖掘和前面谈到的数据挖掘相差 很大。文本数据挖掘并不是一件容易的事情，尤其是在分析方法方面，还有很多需要研究的专题。目前市场上有一些类似的软件，但大部分方法只是把文本移来移 去，或简单地计算一下某些词汇的出现频率，并没有真正的分析功能。

随着计算机计算能力的发展和业务复杂性的提高，数据的类型会越来越多、越来越复杂，数据挖掘将发挥出越来越大的作用。

第九课 数据挖掘应用

9.1 数据挖掘解决的典型商业问题

需要强调的是，数据挖掘技术从一开始就是面向应用的。目前，在很多领域，数据挖掘(data mining)都是一个很时髦的词，尤其是在如银行、电信、保险、交通、零售（如超级市场）等商业领域。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。

9.2 数据挖掘在市场营销的应用

数据挖掘技术在企业市场营销中得到了比较普遍的应用，它是以市场营销学的市场细分原理为基础，其基本假定是“消费者过去的行为是其今后消费倾向的最好说明”。

通过收集、加工和处理涉及消费者消费行为的大量信息，确定特定消费群体或个体的兴趣、消费习惯、消费倾向和消费需求，进而推断出相应消费群体或个体下 一步的消费行为，然后以此为基础，对所识别出来的消费群体进行特定内容的定向营销，这与传统的不区分消费者对象特征的大规模营销手段相比，大大节省了营销 成本，提高了营销效果，从而为企业带来更多的利润。

商业消费信息来自市场中的各种渠道。例如，每当我们用信用卡消费时，商业企业就可以在信用卡结算过程收集商业消费信息，记录下我们进行消费的时间、地 点、感兴趣的商品或服务、愿意接收的价格水平和支付能力等数据；当我们在申办信用卡、办理汽车驾驶执照、填写商品保修单等其他需要填写表格的场合时，我们 的个人信息就存入了相应的业务数据库；企业除了自行收集相关业务信息之外，甚至可以从其他公司或机构购买此类信息为自己所用。

这些来自各种渠道的数据信息被组合，应用超级计算机、并行处理、神经元网络、模型化算法和其他信息处理技术手段进行处理，从中得到商家用于向特定消费 群体或个体进行定向营销的决策信息。这种数据信息是如何应用的呢？举一个简单的例子，当银行通过对业务数据进行挖掘后，发现一个银行帐户持有者突然要求申 请双人联合帐户时，并且确认该消费者是第一次申请联合帐户，银行会推断该用户可能要结婚了，它就会向该用户定向推销用于购买房屋、支付子女学费等长期投资 业务，银行甚至可能将该信息卖给专营婚庆商品和服务的公司。数据挖掘构筑竞争优势。

在市场经济比较发达的国家和地区，许多公司都开始在原有信息系统的基础上通过数据挖掘对业务信息进行深加工，以构筑自己的竞争优势，扩大自己的营业 额。美国运通公司(American Express)有一个用于记录信用卡业务的数据库，

数据量达到54亿字符，并仍在随着业务进展不断更新。运通公司通过对这些数据进行挖掘，制定了“关联 结算(Relation ship Billing)优惠”的促销策略，即如果一个顾客在一个商店用运通卡购买一套时装，那么在同一个商店再买一双鞋，就可以得到比较大的折扣，这样既可以增 加商店的销售量，也可以增加运通卡在该商店的使用率。再如，居住在伦敦的持卡消费者如果最近刚刚乘英国航空公司的航班去过巴黎，那么他可能会得到一个周末 前往纽约的机票打折优惠卡。

基于数据挖掘的营销，常常可以向消费者发出与其以前的消费行为相关的推销材料。卡夫(Kraft)食品公司建立了一个拥有3000万客户资料的数据 库，数据库是通过收集对公司发出的优惠券等其他促销手段作出积极反应的客户和销售记录而建立起来的，卡夫公司通过数据挖掘了解特定客户的兴趣和口味，并以此为基础向他们发送特定产品的优惠券，并为他们推荐符合客户口味和健康状况的卡夫产品食谱。美国的读者文摘(Reader's Digest)出版公司运行着一个积累了40年的业务数据库，其中容纳有遍布全球的一亿多个订户的资料，数据库每天24小时连续运行，保证数据不断得到实 时的更新，正是基于对客户资料数据库进行数据挖掘的优势，使读者文摘出版公司能够从通俗杂志扩展到专业杂志、书刊和声像制品的出版和发行业务，极大地扩展 了自己的业务。

基于数据挖掘的营销对我国当前的市场竞争中也很具有启发意义，我们经常可以看到繁华商业街上一些厂商对来往行人不分对象地散发大量商品宣传广告，其结 果是不需要的人随手丢弃资料，而需要的人并不一定能够得到。如果搞家电维修服务的公司向在商店中刚刚购买家电的消费者邮寄维修服务广告，卖特效药品的厂商 向医院特定门诊就医的病人邮寄广告，肯定会比漫无目的的营销效果要好得多。

9.3 成功案例

9.3.1 电话收费和管理办法

加拿大BC省电话公司要求加拿大Simon Fraser大学KDD研究组根据其拥有的十多年的客户数据，总结、分析并提出新的电话收费和管理办法，制定既有利于公司又有利于客户的优惠政策。

9.3.2 竞技运动中的数据挖掘

美国著名的国家篮球队NBA的教练，利用IBM公司提供的数据挖掘工具临场决定替换队员。想象你是NBA的教练，你靠什么带领你的球队取得胜利呢？当 然，最容易想到的是全场紧逼、交叉扯动和快速抢断等具体的战术和技术。但是今天，NBA的教练又有了他们的新式武器：数据挖掘。大约20个NBA球队使用 了IBM公司开发的数据挖掘应用软件Advanced Scout系统来优化他们的战术组合。例如Scout就因为研究了魔术队队员不同的布阵安排，在与迈阿密热队的比赛中找到了获胜的机会。

----系统分析显示魔术队先发阵容中的两个后卫安佛尼.哈德卫(Anfernee Hardaway)和伯兰.绍(Brian Shaw)在前两场中被评为－17分,这意味着他俩在场上，本队输掉的分数比得到的分数多17分。然而，当哈德卫与替补后卫达利尔.阿姆斯创(Darrell Armstrong)组合时，魔术队得分为正14分。

----在下一场中，魔术队增加了阿姆斯创的上场时间。此着果然见效：阿姆斯创得了21分，哈德卫得了42分，魔术队以88比79获胜。魔术队在第四 场让阿姆斯创进入先发阵容，再一次打败了热队。在第五场比赛中，这个靠数据挖掘支持的阵容没能拖住热队，但Advanced Scout毕竟帮助了魔术队赢得了打满5

场，直到最后才决出胜负的机会。

----Advanced Scout是一个数据分析工具，教练可以用便携式电脑在家里或在路上挖掘存储在NBA中心的服务器上的数据。每一场比赛的事件都被统计分类，按得分、助 攻、失误等等。时间标记让教练非常容易地通过搜索NBA比赛的录像来理解统计发现的含义。例如：教练通过Advanced Scout发现本队的球员在与对方一个球星对抗时有犯规纪录，他可以在对方球星与这个队员“头碰头”的瞬间分解双方接触的动作，进而设计合理的防守策略。

----Advanced Scout的开发人，因德帕尔.布罕德瑞，开发该应用时他正在IBM的ThomasJ.Watson研究中心当研究员，他演示了一个技术新手应该如何使用数据挖掘。布罕德瑞说：“教练们可以完全没有统计学的培训，但他们可以利用数据挖掘制定策略”。与此同时，另一个正式的体育联盟，国家曲棍球联盟，正在开 发自己的数据挖掘应用NHL－ICE，联盟与IBM建立了一个技术型的合资公司，去年 11 月推出一个电子实时的比赛计分和统计系统。在原理上是一个与Advanced Scout相似的数据挖掘应用，可以让教练、广播员、新闻记者及球迷挖掘NHL的统计。当他们访问NHL的Web站点时，球迷能够使用该系统循环看联盟的比赛，同时广播员和新闻记者可以挖掘统计数据，找花边新闻为他们的实况评述添油加醋。

----当然，所有系统都有其局限性。所以不要期望这样的数据挖掘可以帮助一支球队找到赢得足球世界杯的策略。

9.3.3 数据挖掘技术在商业银行中的应用

数据挖掘技术在美国银行金融领域应用广泛。金融事务需要搜集和处理大量数据，对这些数据进行分析，发现其数据模式及特征，然后可能发现某个客户、消费 群体或组织的金融和商业兴趣，并可观察金融市场的变化趋势。商业银行业务的利润和风险是共存的。为了保证最大的利润和最小的风险，必须对帐户进行科学的分 析和归类，并进行信用评估。Mellon银行使用Intelligent Agent数据挖掘软件提高销售和定价金融产品的精确度，如家庭普通贷款。零售信贷客户主要有两类，一类很少使用信贷限额（低循环者），另一类能够保持较 高的未清余额（高循环者）。每一类都代表着销售的挑战。低循环者代表缺省和支出注销费用的危险性较低，但会带来极少的净收入或负收入，因为他们的服务费用几乎与高循环者的相同。银行常常为他们提供项目，鼓励他们更多地使用信贷限额或找到交叉销售高利润产品的机会。高循环者由高和中等危险元件构成。高危险分 段具有支付缺省和注销费用的潜力。对于中等危险分段，销售项目的重点是留住可获利的客户并争取能带来相同利润的新客户。但根据新观点，用户的行为会随时间 而变化。分析客户整个生命周期的费用和收入就可以看出谁是最具创利潜能的。Mellon银行认为“根据市场的某一部分进行定制”能够发现最终用户并将市场 定位于这些用户。但是，要这么做就必须了解关于最终用户特点的信息。数据挖掘工具为Mellon银行提供了获取此类信息的途径。Mellon银行销售部在 先期数据挖掘项目上使用Intelligence Agent寻找信息，主要目的是确定现有Mellon用户购买特定附加产品：家庭普通信贷限额的倾向，利用该工具可生成用于检测的模型。据银行官员称： Intelligence Agent可帮助用户增强其商业智能，如交往、分类或回归分析，依赖这些能力，可对那些有较高倾向购买银行产品、服务产品和服务的客户进行有目的的推销。 该官员认为，该软件可反馈用于分析和决策的高质量信息，然后将信息输入产品的算法。Intelligence Agent还有可定制能力。

美国Firstar银行使用Marksman数据挖掘工具，根据客户的消费模式预测何时为客户提供何种产品。Firstar银行市场调查和数据库营销 部经理发现：公共数据库中存储着关于每位消费者的大量信息，关键是要透彻分析消费者投入到新产品中的原因，在数据库中找到一种模式，从而能够为每种新产品 找到最合适的消费者。Marksman能读取 800 到 1000 个变量并且给它们赋值，根据消费者是否有家庭财产贷款、赊帐卡、存款证或其它储蓄、投资产 品，将它们分成若干组，然后使用数据挖掘工具预测何时向每位消费者提供哪种产品。预测准客户的需要是美国商业银行的竞争优势。

9.3.4 因特网筛选

最近，还有不少DMKD产品用来筛选因特网上的新闻，保护用户不受无聊电子邮件和商业推销的干扰，很受欢迎。

第十课 实施数据挖掘项目考虑的问题

谈到数据挖掘应从以下三方面加以考虑，一是用数据挖掘解决什么样的商业问题，二是为进行数据挖掘所做的数据准备，三是数据挖掘的各种分析算法。

数据挖掘的分析算法主要来自于以下两个方面：统计分析和人工智能（机器学习、模式识别等）。数据挖掘研究人员和数据挖掘软件供应商，在这一方面所做的 主要工作是优化现有的一些算法，以适应大数据量。另外需要强调的是，任何一种数据挖掘的算法，不管是统计分析方法、神经元网络、各种树分析方法，还是遗传 算法，没有一种算法是万能的。不同的商业问题，需要用不同的方法去解决。即使对于同一个商业问题，可能有多种算法，这个时候，也需要评估对于这一特定问题 和特定数据哪一种算法表现好。

做数据挖掘研究的人，往往把主要的精力用于改进现有算法和研究新算法上。人们都知道数据准备是必不可少的一步，但很少有人去真正花时间和精力去研究。 其实数据挖掘最后成功与失败，是否有经济效益，数据准备起到了至关重要的作用。数据准备包含很多方面：一是从多种数据源去综合数据挖掘所需要的数据，保证 数据的综合性、易用性、数据的质量和数据的时效性，这有可能要用到数据仓库的思想和技术；另一方面就是如何从现有数据中衍生出所需要的指标，这主要取决于 数据挖掘者的分析经验和工具的方便性。

众所周知，SQL 是广泛用于数据库查询的语言，有很多数据挖掘软件提供商利用 SQL 来为数据挖掘做数据准备，但就笔者多年来的分析经验和同其他专家探 讨感觉到，SQL 在很多时候有些力不从心，因为数据挖掘和分析的一些算法通常要求数据具有一定的格式和规范性。

还需要强调的一点是，人们通常把数据挖掘工具看得过份神秘，认为只要有了一个数据挖掘工具，就能自动挖掘出所需要的信息，就能更好地进行企业运作，这 是认识上的一个误区。其实要想真正做好数据挖掘，数据挖掘工具只是其中的一个方面，同时还需要对企业业务的深入了解和数据分析经验。一个企业要想在未来的 市场中具有竞争力，必须有一些数据挖掘方面的专家，专门从事数据分析和数据挖掘工作。再同其他部门协调，把挖掘出来的信息供管理者决策参考，最后把挖掘出 的知识物化。在国内的企业中，还很少有决策人员认识到这一点。如果管理者没有这方面的意识，数据挖掘和数据分析就很难发挥应有的作用，很容易走向两个极 端，一是认为数据挖掘没有用处，二是开始认为数据挖掘是万能的。如此得到的结果往往与初始期望相去太远。

具体地说，应考虑以下八个问题：

1. 超大规模数据库和高维数据问题；

2. 数据丢失问题；
3. 变化的数据和知识问题；
4. 模式的易懂性问题；
5. 非标准格式的数据、多媒体数据、面向对象数据处理问题；
6. 与其他系统的集成问题；
7. 网络与分布式环境下的 KDD 问题。
8. 个人隐私问题