第1章　数据仓库与数据挖掘概述

1.1　数据仓库的兴起

1.1.1　从数据库到数据仓库

由数据库(DB)发展到数据仓库(DW)主要在于如下几点。

- 数据太多,信息贫乏(data rich, information poor):随着数据库技术的发展,企事业单位建立了大量的数据库,数据越来越多,而辅助决策信息却很贫乏,如何将大量的数据转化为辅助决策信息成了研究的热点。
- 异构环境数据的转换和共享:由于各类数据库产品的增加,异构环境的数据也随之增加,如何实现这些异构环境数据的转换和共享也成了研究的热点。
- 利用数据进行事务处理转变为利用数据支持决策:数据库用于事务处理,若要达到辅助决策,则需要更多的数据。例如,如何利用历史数据的分析来进行预测。对大量数据的综合得到宏观信息等均需要大量的数据。

数据仓库概念提出后,在不到几年的时间内就得到了迅速的发展。数据仓库产品也不断出现并陆续进入市场。

1. 数据库用于事务处理

数据库存储大量的共享数据,作为数据资源用于管理业务中的事务处理,已经成为了成熟的信息基础设施。

数据库中存放的数据基本上是保存当前数据,随着业务的变化随时更新数据库中的数据。例如,学生数据库,随着新生的入校,数据库中要增加新学员的数据记录;随着毕业学生的离校,数据库中要删除这些学员的数据记录。数据库总是保持当前的数据记录。

不同的管理业务需要建立不同的数据库。例如,银行中储蓄业务要建立储蓄数据库,记录所有储蓄用户的存款及使用信息;信用卡业务要建立信用卡数据库,记录所有用户信用卡的存款及使用信息;贷款业务要建立贷款数据库,记录贷款用户的贷款及使用信息。

数据库是为事务处理需求设计和建立的,从而使计算机在事务处理上发挥极大的效果。但是,数据库在帮助人们进行决策分析时就显得不适应了。例如,银行想了解用户的经济状态(收入与支出情况)以及信誉如何(是否超支,还贷情况等)?是否继续贷款给他?单靠一个数据库是无法完成这种决策分析的。必须将储蓄数据库、信用卡数据库、贷款数据库集中起来,对某一个人进行全面分析,才能准确了解他的存款及收支情况、信用卡使用情况以及贷款和还贷情况。这样,银行才能有效地决定是否给此人继续贷款。

同时使用3个数据库进行操作并非是一件简单的事,由于3个管理业务各自独立,在建

立数据库时对同一个人可能使用了不同的编码，对于他的姓名可能有的用汉字，有的用汉语拼音，有的用英文。这为使用 3 个数据库地共同进行决策分析带来了困难。

2. 数据仓库用于决策分析

随着决策分析的需求扩大，兴起了支持决策的数据仓库。它是以决策主题需求集成多个数据库，重新组织数据结构，统一规范编码，使其有效地完成各种决策分析。

从数据库到数据仓库的演变，体现了以下几点。

(1) 数据库用于事务处理，数据仓库用于决策分析

事务处理功能单一，数据库完成事务处理的增加、删除、修改、查询等操作。决策分析要求数据较多。数据仓库需要存储更多的数据，不需要修改数据，主要提取综合数据的信息，以及分析预测数据的信息。

(2) 数据库保持事务处理的当前状态，数据仓库既保存过去的数据又保存当前的数据

数据库中数据随业务的变化一直在更新，总保存当前的数据，如学生数据库。数据仓库中数据不随时间变化而变化，但保留大量不同时间的数据，即保留历史数据和当前数据。

(3) 数据仓库的数据是大量数据库的集成

数据仓库的数据不是数据库的简单集成，而是按决策主题，将大量数据库中数据进行重新组织，统一编码进行集成。如银行数据仓库数据是由储蓄数据库、信用卡数据库、贷款数据库等多个数据库按“用户”主题进行重新组织、编码和集成而建立的。可见，数据仓库的数据量比数据库的数据量大得多。

(4) 对数据库的操作比较明确，操作数据量少。对数据仓库操作不明确，操作数据量大

一般对数据库的操作都是事先知道的事务处理工作，每次操作(增加、删除、修改、查询)涉及的数据量也小，如一个或几个记录数据。

对数据仓库的操作都是根据当时决策需求临时决定而进行的。如比较两个地区某个商品销售的情况。该操作所涉及的数据量很大，不是几个记录数据，而是两个地区多个商店的某商品的所有销售记录。

3. 数据库与数据仓库对比

数据库与数据仓库的对比如表 1.1 所示。

表 1.1　数据库与数据仓库对比

数据库	数据仓库
面向应用	面向主题
数据是详细的	数据是综合的或提炼的
保持当前数据	保存过去和现在的数据
数据是可更新的	数据不更新
对数据操作是重复的	对数据的操作是启发式的
操作需求是事先可知	操作需求是临时决定的

续表

数据库	数据仓库
一个操作存取一个记录	一个操作存取一个集合
数据非冗余	数据时常冗余
操作比较频繁	操作相对不频繁
查询的是原始数据	查询的是经过加工的数据
事务处理需要的是当前数据	决策分析需要过去、现在的数据
很少有复杂的计算	很多复杂的计算
支持事务处理	支持决策分析

1.1.2 从 OLTP 到 OLAP

1. 联机事物处理(on line transaction processing,OLTP)

联机事物处理是在网络环境下的事务处理工作,利用计算机网络技术,以快速的事务响应和频繁的数据修改为特征,使用户利用数据库能够快速地处理具体的业务。OLTP 是事务处理从单机到网络环境发展的新阶段。OLTP 应用要求多个查询并行,以便将每个查询分布到一个处理器上。

OLTP 的特点在于事务处理量大,但事务处理内容比较简单且重复率高。大量的数据操作主要涉及的是一些增加、删除、修改、查询等操作。每次操作的数据量不大且多为当前的数据,OLTP 的数据组织的数据模型采用实体-关系(E-R)模型。

OLTP 处理的数据是高度结构化的,涉及的事务比较简单,数据访问路径是已知的,至少是固定的。事务处理应用程序可以直接使用具体的数据结构,如表、索引等。

OLTP 面对的是事务处理操作人员和低层管理人员。

在过去三十多年中,OLTP 系统发展的目标就是能够处理大量的数据。每时间单位能够处理更多的事务,能支持更多的并发用户,且有更好的系统健壮性。大型的系统每秒能够处理 1000 个以上的事务。有些系统,像机票预订系统,每秒能够处理的事务峰值可以达到 2 万个。

数据库存储的数据量很大,经常每天要处理成千上万的事务,OLTP 在查找业务数据时是非常有效的。但是为高层领导者提供决策分析时,则显得力不从心。

2. 联机分析处理(on line analytical processing,OLAP)

关系数据库之父 E. F. Codd 在 1993 年认为,联机事务处理已经不能满足终端用户对数据库决策分析的需要,决策分析需要对多个关系数据库共同进行大量的综合计算才能得到结果。为此,他提出了多维数据库和多维分析的概念,即联机分析处理概念。关系数据库是二维数据(平面),多维数据库是空间立体数据。

近年来,人们利用信息技术生产和搜集数据的能力大幅度提高,大量的数据库被用于商业管理、政府办公、科学研究和工程开发等,这一势头仍将持续发展下去。于是,一个新的挑战被提出来:在信息爆炸的时代,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或者规律，提高信息利用率呢？要想使数据真正成为一个决策资源，只有充分利用它为一个组织的业务决策和战略发展服务才行，否则大量的数据可能成为包袱，甚至成为垃圾。OLAP 是解决这类问题的最有力的工具之一。

OLAP 专门用于支持复杂的分析操作，侧重对分析人员和高层管理人员的决策支持，可以应分析人员的要求快速、灵活地进行大数据量的复杂处理，并且以一种直观易懂的形式将查询结果提供给决策制定人，以便他们准确掌握企业（公司）的经营情况，了解市场需求，制定正确方案，以增加效益。OLAP 软件以它先进的分析功能和以多维形式提供数据的能力，正作为一种支持企业关键商业决策的解决方案而迅速崛起。

OLAP 的基本思想是决策者从多方面和多角度以多维的形式来观察企业的状态和了解企业的变化。

3. OLTP 与 OLAP 的对比

OLAP 是以数据仓库为基础，其最终数据来源与 OLTP 一样均来自底层的数据库系统，但由于二者面对的用户不同，OLTP 面对的是操作人员和低层管理人员，OLAP 面对的是决策人员和高层管理人员，因而数据的特点与处理也明显不同。

OLTP 和 OLAP 是两类不同的应用，它们的各自特点见表 1.2 所示。

表 1.2　OLTP 与 OLAP 对比表

OLTP	OLAP
数据库数据	数据库或数据仓库数据
细节性数据	综合性数据
当前数据	历史数据
经常更新	不更新，但周期性刷新
一次性处理的数据量小	一次性处理的数据量大
对响应时间要求高	响应时间合理
用户数量大	用户数量相对较少
面向操作人员，支持日常操作	面向决策人员，支持决策需要
面向应用，事务驱动	面向分析，分析驱动

1.1.3　数据字典与元数据

1. 数据库的数据字典

数据字典是数据库中各类数据描述的集合，在数据库设计中占有很重要的地位。数据字典通常包括数据项、数据结构、数据流、数据存储和处理过程 5 个部分，其中数据项是数据的最小组成单位。若干个数据项可以组成一个数据结构。数据字典通过对数据项和数据结构的定义来描述数据流、数据存储的逻辑内容。

（1）数据项

数据项是不可再分的数据单位。对数据项的描述通常包括数据项名、数据项含义说明、数据类型、长度、取值范围和取值含义等。

(2) 数据结构

数据结构反映了数据之间的组合关系。一个数据结构可以由若干个数据项组成，也可以由若干个数据结构组成。数据结构的描述通常包括数据结构名、含义说明和数据项等。

(3) 数据流

数据流是数据结构在系统内传输的路径，对数据流的描述通常包括数据流名、说明、数据流来源、数据流去向和平均流量等。其中“数据流来源”是说明该数据流来自哪个过程。“数据流去向”是说明该数据流将到哪个过程去。“平均流量”是指单位时间(如每天)传输的次数。

(4) 数据存储

数据存储是数据结构保存数据的地方，数据存储的描述通常包括数据存储名、说明、编号、输入的数据流、输出的数据流、数据量、存取频度和存取方式。其中“存取频度”指每小时或每天或每周存取几次、每次存取多少数据等信息。“存取方式”包括是批处理还是联机处理；是检索还是更新；是顺序检索还是随机检索等。另外，“输入的数据流”要指出其来源，“输出的数据流”要指出其去向。

(5) 处理过程

处理过程一般用判定表或判定树来描述。数据字典中只需要描述处理过程的说明性信息，通常包括处理过程名、说明、输入、输出和处理。其中“处理”中主要说明该处理过程的功能及处理要求。

可见，数据字典是关于数据库中数据的描述，而不是数据本身。

2. 数据仓库的元数据

数据仓库远比数据库复杂。在数据仓库中引入了“元数据”的概念，它不仅是数据仓库的字典，而且还是数据仓库本身信息的数据。

元数据(meta data)定义为关于数据的数据(data about data)，即元数据描述了数据仓库的数据和环境。

元数据在数据仓库中不仅定义了数据仓库有什么，还指明了数据仓库中信息的内容和位置，刻画了数据的抽取和转换规则，存储了与数据仓库主题有关的各种商业信息，而且整个数据仓库的运行都是基于元数据的，如数据的修改、跟踪、抽取、装入、综合以及使用等。由于元数据遍及数据仓库的所有方面，已成为整个数据仓库的核心。

数据仓库的元数据除对数据仓库中数据的描述(数据仓库字典)外，还有以下 3 类元数据。

(1) 关于数据源的元数据

数据仓库的数据源包含了很多不同的数据结构，为数据仓库选取的数据元素字段长度和数据类型而有所不同。为数据仓库挑选数据时，得将记录拆分，并将来自不同源文件的记录的某些部分组合起来，还要解决编码和字段长度不同的问题。当将这些信息传递给最终用户的时候，必须把这些数据与原始数据联系起来。

(2) 关于抽取和转换的元数据

这类元数据包含了源数据系统的数据抽取方法、数据抽取规则，以及抽取频率等数据转

换的所有信息。

(3) 关于最终用户的元数据

最终用户元数据是数据仓库的导航图,使最终用户可以从数据仓库中找到自己需要的信息。

1.1.4 数据仓库的定义与特点

数据仓库的概念是由 W. H. Inmon 在《建立数据仓库(Building the Data Warehouse)》一书中提出的。数据仓库的提出是以关系数据库、并行处理和分布式技术为基础的信息新技术。

从目前的形势看,数据仓库技术已紧跟 Internet,成为信息社会中获得企业竞争优势的又一关键技术。

1. 数据仓库定义

(1) W. H. Inmon 对数据仓库的定义

数据仓库是面向主题的、集成的、稳定的、不同时间的数据集合,用于支持经营管理中决策制定过程。

(2) SAS 软件研究所的观点

数据仓库是一种管理技术,旨在通过通畅、合理、全面的信息管理,达到有效的决策支持。

从数据仓库的定义可以看出,数据仓库是为决策支持服务的,而数据库是为事务处理服务的。

2. 数据仓库特点

从数据仓库的定义可以看出数据仓库的特点如下。

(1) 数据仓库是面向主题的

主题是数据归类的标准,每一个主题基本对应一个宏观的分析领域。例如,保险公司的数据仓库的主题为客户、政策、保险金和索赔等。

基于应用的数据库组织则完全不同,它的数据只是为处理具体应用而组织在一起的。保险公司按应用组织的数据库是汽车保险、生命保险、健康保险和伤亡保险等。

(2) 数据仓库是集成的

数据进入数据仓库之前,必须经过加工与集成。对不同的数据来源进行统一数据结构和编码。统一原始数据中的所有矛盾之处,如字段的同名异义、异名同义、单位不统一和字长不一致等。总之,将原始数据结构做一个从面向应用到面向主题的大转变。

(3) 数据仓库是稳定的

数据仓库中包括了大量的历史数据。数据经集成进入数据仓库后是极少或根本不更新的。

(4) 数据仓库是随时间变化的

数据仓库内的数据时限在 5～10 年,故数据的键码包含时间项,并标明数据的历史时

期，以便适合决策分析时进行时间趋势分析。

而数据库只包含当前数据，即存储某一时间的正确的有效数据。

(5) 数据仓库中的数据量很大

通常的数据仓库的数据量为 10GB 级，相当于一般数据库 100MB 的 100 倍，大型数据仓库是一个 TB(1000GB)级数据量。

数据仓库中数据量的比重是：索引和综合数据占 2/3，原始数据占 1/3。

(6) 数据仓库软硬件要求较高

① 需要一个巨大的硬件平台；

② 需要一个并行的数据库系统。

1.2 数据挖掘的兴起

1.2.1 从机器学习到数据挖掘

学习是人类具有的智能行为，主要在于获取知识。机器学习是研究使计算机模拟或实现人类的学习行为，即让计算机通过算法自动获取知识。机器学习是人工智能领域中的重要研究方向。

20 世纪 60 年代开始了机器学习的研究。比较典型的成果有：Rosenblate 的感知机，它是最早用神经网络进行模式识别的方法；Sammel 的西洋跳棋程序，它用线性表达式的启发式方法，通过多次人机对弈，自动修改表达式中的系数，使程序逐渐聪明，该程序竟然达到胜过开发者和州冠军的成绩。

20 世纪 80 年代，机器学习取得了较大的成果。Michelski 等人的 AQ11 系统(1980)能从大量病例中归纳出大豆病症的判断规则。AQ11 是一个很成功的归纳学习方法；Quiulan 的 ID3(1983)决策树方法，影响很大，实用效果很强；Langley 等人的 BACON 系统(1987)能重新发现物理学的大量规律；Rumelhart 等人研制的反向传播神经网络 BP 模型(1985)为神经网络的学习开创了一个新阶段。

这些显著成果的出现，使“机器学习”逐渐形成了人工智能的主要学科方向之一。1980 年在美国召开了第一届国际机器学习学会研讨会，1984 年《机器学习》杂志问世。

中国在 1987 年召开了第一届全国机器学习研讨会。1989 年成立了中国人工智能学会机器学习学会。中国学者洪家荣研制的 AE1 系统(1985)采用了扩张矩阵方法；钟鸣等人研制的 IBLE 方法(1992)利用信道容量建立决策规则树，识别效果比 ID3 方法更高。本书作者研制的 FDD 经验公式发现系统(1998)，能发现含初等函数或复合函数的经验公式，发现的公式比 BACON 系统发现的公式范围更宽。

1989 年美国召开了第一届知识发现(knowledge discovery in database，KDD)国际学术会议，从数据库中发现知识形成了新概念。KDD 研究的问题有：定性知识和定量知识的发现；知识发现方法；知识发现的应用等。

1995 年在加拿大召开了第一届知识发现和数据挖掘(data mining，DM)国际学术会议。由于把数据库中的“数据”形象地比喻成矿床，“数据挖掘”一词很快流传开来。

数据挖掘是知识发现中的核心工作，主要研究发现知识的各种方法和技术。而这些方法和技术主要来自于机器学习。由于数据挖掘的发展，出现了一些新的数据挖掘方法，如大型数据库中关联规则的挖掘，以及利用粗糙集进行属性约简和规则获取等。

数据挖掘兴起时主要是在数据库中挖掘知识，随着数据仓库的出现和发展，很快将数据挖掘技术和方法用于数据仓库。典型的啤酒与尿布的故事（该两商品同时出售的出现概率很高）就是在数据仓库中挖掘出的关联知识。

1.2.2 数据挖掘的含义

按《人工智能辞典》的定义：信息是数据中所蕴涵的意义。知识是人们对客观世界的规律性认识。

数据库中每个数据记录的内涵代表了该记录的信息。而数据挖掘是从数据库中所有数据记录中归纳总结出知识。知识的数量大大少于数据记录量。这些知识代表了数据库中数据信息的规律，即用少量的知识能够覆盖数据库中所有的记录。

例如，人口数据库中存储各国人口的记录，它将是一个庞大的数据库。但是，通过数据挖掘，可以得出形式化表示的规则知识：

（头发＝黑色）∨（眼睛＝黑色）→亚洲人

其中∨表示“或”；→表示“蕴涵”，规则知识表示为：“若（条件）则（结论）”，即表示：若头发是黑色或者眼睛是黑色的人，则他是亚洲人。

该知识代表了亚洲人的特点，即覆盖了所有亚洲人的记录。

知识的获得是通过数据挖掘算法，如 AQ11 方法和 ID3 方法等经过计算得到的。

1.2.3 数据挖掘与 OLAP 的比较

1. OLAP 的多维分析

OLAP 是在多维数据结构上进行数据分析的。对多维数据进行分析是复杂的。一般从多维数据取出（切片、切块）二维或三维数据进行分析，或对层次的维进行钻取操作，向下钻取获得更详细的数据，向上钻取获得更综合的数据。

OLAP 要适应大量用户同时使用同一批数据，适应于不同地理位置的分散化的决策。OLAP 的功能和算法包括聚合、分配、比率、乘积等描述性的建模功能。

OLAP 平时需要查询大量的日常商业活动信息，如每周的布匹购买量、每周布匹的内部库存以及布匹的销售量等。OLAP 更需要查询商业活动的变化情况，如每周布匹购买量的变化值、衣服生产量的变化值、衣服销售价格的变化等。这些变化值对经理制定决策更重要。

经理往往从查询出的变化值中，通过 OLAP 追踪查询，找出存在的原因。例如，经理看到利润小于预计值的时候，可能会深入到各个国家查看整个产品利润情况。这样，他可能发现有些国家的利润明显低于其他国家，于是他自然就会查看这些国家中不同产品组的利润情况，总的目标就是寻找一些比较异常的数据来解释某个现象。经过一番观察之后，就会发现非直接成本在这些国家明显偏高。进一步对这些非直接成本分析，可以发现近期对于某

些产品的赋税明显增加，从而明显影响了最终的利润。这种分析查询要求时间响应快。

以上是 OLAP 的典型应用，通过商业活动变化的查询发现的问题，经过追踪查询找出问题出现的原因，达到辅助决策的作用。

2. 数据挖掘

OLAP 是在带层次的维度和跨维度进行多维数据分析的。数据挖掘则不同，是以变量和记录为基础进行分析的。

数据挖掘任务在于聚类（如神经网络聚类）、分类（如决策树分类）、预测等。这些是带有探索性的建模功能。

数据挖掘在于寻找不平常的且有用的商业运作模型。考察数据的不同类型或者找出变量之间的关系。数据挖掘需要查看海量数据，主要是详细数据和历史数据。为此经常将数据仓库中的数据复制到一个专门的存储器上，对数据的挖掘分析可能要花去大量的时间，即不要求快速分析。数据挖掘人员有时并不能精确地知道什么是必须分析的，有时数据挖掘一无所获。但是，有时通过数据挖掘会发现意外的、无价的信息“金块”。例如，如果能够确定一个高价值的客户或可能离开的客户特征，就可以要求公司采取措施保留这些客户，这比从竞争对手那儿重新争取曾经失去的客户费用少得多。

1.2.4 数据挖掘与统计学

1. 统计学的发展过程

统计学是一门有悠久历史的学科。统计学开始于 17、18 世纪，与国家政治有紧密的关系。英国 W. Petty(1623—1682 年)的《政治算术》一书中第一次用计量和比较的方法，对英国与法、意、荷等国进行国力比较。J. Graunt(1620—1674 年)通过统计计算，发现男女人数占人口数的比例大致相等、出生儿中男婴比例稍高、婴幼儿的死亡率较大等规律性的现象。

17 世纪，B. Pascal 等人提出“概率”概念，用来描述某一事件发生的可能性。18 世纪，在观测天体运动时会有误差产生，虽然多次测量，由于有误差，得到的总是和真值不同的值。高斯(Gauss，1777—1855 年)提出误差值落在(a,b)区间的概率等于该区间上正态分布曲线下的面积，称误差服从正态分布（高斯分布）。比利时的凯特勒(A. Quetelet，1796—1874 年)称“支配着社会现象的法则和方法是概率论”。

近代统计学重视社会调查。通过对全部对象（总体）进行调查，为制定计划和决策提供依据，如果对总体的某些分布情况有一定把握，就不必搞全面调查，可以搞部分调查，即抽样调查，由部分推断全部。概率论和数理统计理论起着重要的作用。现在各国在经济统计、国事调查、社会调查、收视率调查、民意测验等采用的几乎都是抽样调查。

现代统计学从线性到非线性、从低维到高维、从显在到潜在、从连续到离散等方面有较完备的理论和方法。统计软件包 SPSS、SAS 等已经普及，统计工作基本上利用计算机来完成。

2. 统计学中应用于数据挖掘的内容

(1) 常用统计

在大量数据中求最大值、最小值、总和、平均值等。

(2) 相关分析

通过求变量间的相关系数来确定变量间的相关程度。

(3) 回归分析

建立回归方程(线性或非线性)以表示变量间的数量关系,再利用回归方程进行预测。

(4) 假设检验

在总体存在某些不确定情况时,为了推断总体的某些性质,提出关于总体的某些假设,对此假设利用置信区间来检验,即任何落在置信区间之外的假设判断为"拒绝",任何落在置信区间之内的假设判断为"接受"。

(5) 聚类分析

将样品或变量进行聚类的方法,具体方法是把样品中每一个样品看成是 m 维空间的一个点,聚类是把"距离"较近的一个点归为同一类,而将"距离"较远的点归为不同的类。

(6) 判别分析

建立一个或多个判别函数,并确定一个判别标准。对未知对象利用判别函数将它划归某一个类别。

(7) 主成分分析

主成分分析是把多个变量化为少数的几个综合变量,而这几个综合变量可以反映原来多个变量的大部分信息。

主成分分析的一种推广是因子分析,即用少数几个因子(F_i)去描述许多变量(X_j)之间的关系。变量(X_j)是可以观测的显在变量,而因子(F_i)是不可观测的潜在变量。

3. 统计学与数据挖掘的比较

统计学主要是对数量数据(数值)或连续值数据(如年龄、工资等)进行数值计算(如初等运算)的定量分析,得到数量信息。如常用统计量(最大值、最小值、平均值、总和等)、相关系数、回归方程等。

数据挖掘主要对离散数据(如职称、病症等)进行定性分析(覆盖、归纳等),得到规则知识。例如,如果某人的眼睛是黑的或者头发是黑的,则可以认为他是亚洲人。

在统计学中有聚类分析和判别分析,它们与数据挖掘中的聚类和分类相似。但是,采用的标准不一样,统计学的聚类采用的"距离"是欧式距离,即两点间的坐标(数值)距离。而数据挖掘的聚类采用的"距离"是海明距离,即属性取值是否相同,相同者距离为 0,不相同者距离为 1。

总之,统计学与数据挖掘是有区别的,但是,它们之间是相互补充的。不少数据挖掘的著作中均把统计学的不少方法引入到数据挖掘中,与将机器学习中不少方法引入到数据挖掘中一样,作为从数据获取知识的一大类方法。

虽然统计学的不少方法可以归入到数据挖掘中,但统计学仍然是一门独立的学科。