数据挖掘概念与特性

一． 基本概念

数据挖掘（*Data mining*）又译为资料探勘、数据采矿。它是数据库知识发现（英语: Knowledge-Discovery in Databases，简称：KDD)中的一个步骤。数据挖掘一般是指从大量的数据中通过算法搜索隐藏于其中信息的过程。数据挖掘通常与计算机科学有关，并通过统计、在线分析处理、情报检索、机器学习、专家系统（依靠过去的经验法则）和模式识别等诸多方法来实现上述目标

----百度百科

以上是数据挖掘的广义概念，在大数据专业中，数据挖掘一般是指从大量的数据中通过算法搜索隐藏于其中价值的过程，其所使用的技术就是大数据处理技术栈，业界内当前使用最多的是 Hadoop 生态圈工具，一整套的大数据分析处理工具。
为什么 Hadoop 使用率最高?
因为它是开源，免费的

二． 数据挖掘定义

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。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。

与数据挖掘相近的同义词有数据融合、人工智能、商务智能、模式识别、机器学习、知识发现、数据分析和决策支持等。

----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. 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

三． 大数据挖掘和大数据分析的区别?

大数据挖掘是对信息的价值化的获取。价值化自然不会考虑数据本身，而是考虑数据是否有价值。由此，一批数据，你尝试对它做不同的价值评估，这是大数据挖掘
大数据分析是以输入的数据为基础，通过先验的约束，对数据进行处理，但是不以结论何如为调整。例如你需要图像识别，这个属于大数据分析。你要分析人脸，数据就是通过先验的 方法，就是出来个猫脸。你的数据分析也没有问题，你需要默默承受结果，并且尊重事实。因此大数据分析的重点在于数据的有效性、真实性和先验约束的正确性
此时对比大数据分析，最大的特点就是你需要调整你不同的先验约束，再次对数据进行分析。而先验的约束已经不是针对数据来源自身的特点，例如信噪比处理算法。你期望得到的一个有价值的内容，做先验的约束，以观测，数据根据这个约束，是否有正确的反馈

四． 成功案例

数据挖掘帮助 Credilogros Cía Financiera S.A.改善客户信用评分

Credilogros Cía Financiera S.A. 是阿根廷第五大信贷公司，资产估计价值为 9570 万美元，对于 Credilogros 而言，重要的是识别与潜在预先付款客户相关的潜在风险，以便将承担的风险最小化。

该公司的第一个目标是创建一个与公司核心系统和两家信用报告公司系统交互的决策引擎来处理信贷申请。同时，Credilogros 还在寻找针对它所服务的低收入客户群体的自定义风险评分工具。除这些之外，其他需求还包括解决方案能在其 35 个分支办公地点和 200 多个相关的销售点中的任何一个实时操作，包括零售家电连锁店和手机销售公司。 [3]
最终 Credilogros 选择了 SPSS Inc.的数据挖掘软件 PASWModeler，因为它能够灵活并轻松地整合到 Credilogros 的核心信息系统中。通过实现 PASW Modeler，Credilogros 将用于处理信用数据和提供最终信用评分的时间缩短到了 8 秒以内。这使该组织能够迅速批准或拒绝信贷请求。该决策引擎还使 Credilogros 能够最小化每个客户必须提供的身份证明文档，在一些特殊情况下，只需提供一份身份证明即可批准信贷。此外，该系统还提供监控功能。Credilogros 目前平均每月使用 PASW Modeler 处理 35000 份申请。仅在实现 3 个月后就帮助 Credilogros 将贷款支付失职减少了 20%

数据挖掘帮助 DHL 实时跟踪货箱温度

DHL 是国际快递和物流行业的全球市场领先者，它提供快递、水陆空三路运输、合同物流解决方案，以及国际邮件服务。DHL 的国际网络将超过 220 个国家及地区联系起来，员工总数超过 28.5 万人。在美国 FDA 要求确保运送过程中药品装运的温度达标这一压力之下，DHL 的医药客户强烈要求提供更可靠且更实惠的选择。这就要求 DHL 在递送的各个阶段都要实时跟踪集装箱的温度。
虽然由记录器方法生成的信息准确无误，但是无法实时传递数据，客户和 DHL 都无法在发生温度偏差时采取任何预防和纠正措施。因此，DHL 的母公司德国邮政世界网（DPWN）通过技术与创新管理（TIM）集团明确拟定了一个计划，准备使用 RFID 技术在不同时间点全程跟踪装运的温度。通过 IBM 全球企业咨询服务部绘制决定服务的关键功能参数的流程框架。DHL 获得了两方面的收益：对于最终客户来说，能够使医药客户对运送过程中出现的装运问题提前做出响应，并以引人注目的低成本全面切实地增强了运送可靠性。对于 DHL 来说，提高了客户满意度和忠实度；为保持竞争差异奠定坚实的基础；并成为重要的新的收入增长来源

五． 常用的分析算法

1. C4.5：是机器学习算法中的一种分类决策树算法，其核心算法是 ID3 算法。
2. K-means 算法：是一种聚类算法。
3.SVM：一种监督式学习的方法，广泛运用于统计分类以及回归分析中
4.Apriori ：是一种最有影响的挖掘布尔关联规则频繁项集的算法。
5.EM：最大期望值法。
6.pagerank：是 google 算法的重要内容。
7. Adaboost：是一种迭代算法，其核心思想是针对同一个训练集训练不同的分类器然后把弱分类器集合起来，构成一个更强的最终分类器。
8.KNN：是一个理论上比较成熟的的方法，也是最简单的机器学习方法之一。
9.Naive Bayes：在众多分类方法中，应用最广泛的有决策树模型和朴素贝叶斯（Naive Bayes）
10.Cart：分类与回归树，在分类树下面有两个关键的思想，第一个是关于递归地划分自变量空间的想法，第二个是用验证数据进行减枝