前沿技术

区块链

- 概念
  - 区块链起源于比特币，是一个信息技术领域的术语。从本质上讲，它是一个共享数据库，存储于其中的数据或信息，具有"不可伪造""全程留痕""可以追溯""公开透明""集体维护"等特征。基于这些特征，区块链技术奠定了坚实的"信任"基础，创造了可靠的"合作"机制，具有广阔的运用前景。
  - 区块链是分布式数据存储、点对点传输、共识机制、加密算法等计算机技术的新型应用模式。
- 类型
  - 公有区块链
    - 世界上任何个体或者团体都可以发送交易，且交易能够获得该区块链的有效确认，任何人都可以参与其共识过程。
  - 联合（行业）区块链
    - 由某个群体内部指定多个预选的节点为记账人，每个块的生成由所有的预选节点共同决定（预选节点参与共识过程），其他接入节点可以参与交易，但不过问记账过程(本质上还是托管记账，只是变成分布式记账，预选节点的多少，如何决定每个块的记账者成为该区块链的主要风险点），其他任何人可以通过该区块链开放的API进行限定查询
  - 私有区块链
    - 仅仅使用区块链的总账技术进行记账，可以是一个公司，也可以是个人，独享该区块链的写入权限，本链与其他的分布式存储方案没有太大区别。
- 特征
  - 去中心化
    - 区块链技术不依赖额外的第三方管理机构或硬件设施，没有中心管制，除了自成一体的区块链本身，通过分布式核算和存储，各个节点实现了信息自我验证、传递和管理。去中心化是区块链最突出最本质的特征
  - 开放性
    - 区块链技术基础是开源的，除了交易各方的私有信息被加密外，区块链的数据对所有人开放，任何人都可以通过公开的接口查询区块链数据和开发相关应用，因此整个系统信息高度透明
  - 独立性
    - 基于协商一致的规范和协议(类似比特币采用的哈希算法等各种数学算法)，整个区块链系统不依赖其他第三方，所有节点能够在系统内自动安全地验证、交换数据，不需要任何人为的干预
  - 安全性
    - 只要不能掌控全部数据节点的51%，就无法肆意操控修改网络数据，这使区块链本身变得相对安全，避免了主观人为的数据变更
  - 匿名性
    - 除非有法律规范要求，单从技术上来讲，各区块节点的身份信息不需要公开或验证，信息传递可以匿名进行
- 核心技术
  - 分布式账本
  - 非对称加密
  - 共识机制
  - 智能合约
- 应用
  - 金融/物联网/物流/公共服务/数字版权/保险/公益

物联网

- 概念
  - 物联网是一个基于互联网、传统电信网等的信息承载体，它让所有能够被独立寻址的普通物理对象形成互联互通的网络
  - 物物相连，万物互联
    - 第一，物联网的核心和基础仍然是互联网，是在互联网基础上的延伸和扩展的网络
    - 其用户端延伸和扩展到了任何物品与物品之间，进行信息交换和通信。
- 基本特征
  - 整体感知
    - 射频识别，二维码，智能传感器感知获取物体各类信息
  - 可靠传输
    - 通过对互联网、无线网络的融合，将物体的信息实时、准确地传送，以便信息交流、分享。
  - 智能处理
    - 用各种智能技术，对感知和传送到的数据、信息进行分析处理，实现监测与控制的智能化。
- 关键技术
  - 射频识别技术
  - 传感网
  - M2M系统框架
  - 云计算
- 应用
  - 智能交通
  - 智能家居
  - 公共安全
- 挑战
  - 技术标准的统一与协调
  - 管理平台问题
  - 成本问题
  - 安全性问题

大数据

- 概念
  - 大数据，指无法在一定时间范围内用常规软件工具进行捕捉、管理和处理的数据集合，是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。
- 特征
  - 容量
    - 数据的大小决定所考虑的数据的价值和潜在的信息
  - 种类
    - 数据类型的多样性；
  - 速度
    - 指获得数据的速度
  - 可变性
    - 妨碍了处理和有效地管理数据的过程
  - 真实性
    - 数据的质量
  - 复杂性
    - 数据量巨大，来源多渠道
  - 价值
    - 合理运用大数据，以低成本创造高价值
- 结构
  - 理论
    - 理论是认知的必经途径，也是较广泛认同和传播的基线
  - 技术
    - 技术是大数据价值体现的手段和前进的基石
  - 实践
    - 实践是大数据的最终价值体现
- 技术
  - Hadoop
    - Hadoop是Apache公司开发的一个开源框架,它允许在整个集群使用简单编程模型计算机的分布式环境存储并处理大数据。
    - 核心
      - 1、HDFS(分布式文件存储系统)
        - 数据以块的形式，分布在集群的不同节点。在使用HDFS时，无需关心数据是存储在哪个节点上、或者是从哪个节点从获取的，只需像使用本地文件系统一样管理和存储文件系统中的数据。
      - 2、MapReduce(分布式计算框架)
        - 分布式计算框架将复杂的数据集分发给不同的节点去操作，每个节点会周期性的返回它所完成的工作和最新的状态。大家可以结合下图理解MapReduce原理
      - 3、YARN(资源调度器)
        - 相当于电脑的任务管理器，对资源进行管理和调度
    - 其他重要模块
      - 4、HBASE(分布式数据库)
        - HBase是非关系型数据库(Nosql)，在某些业务场景下，数据存储查询在Hbase的使用效率更高。
      - 5、HIVE(数据仓库)
        - HIVE是基于Hadoop的一个数据仓库工具，可以用SQL的语言转化成MapReduce任务对hdfs数据的查询分析。HIVE的好处在于，使用者无需写MapReduce任务，只需要掌握SQL即可完成查询分析工作。
      - 6、Spark(大数据计算引擎)
        - Spark是专为大规模数据处理而设计的快速通用的计算引擎
      - 7、Mahout(机器学习挖掘库)
        - Mahout是一个可扩展的机器学习和数据挖掘库
      - 8、Sqoop
        - Sqoop可以将关系型数据库导入Hadoop的HDFS中，也可以将HDFS的数据导进到关系型数据库中
- 趋势
  - 数据资源化
    - 资源化是指大数据成为企业和社会关注的重要战略资源，并已成为大家争相抢夺的新焦点
  - 与云计算的深度结合
    - 大数据离不开云处理，云处理为大数据提供了弹性可拓展的基础设备，是产生大数据的平台之一
  - 科学理论的突破
    - 随着大数据的快速发展，就像计算机和互联网一样，大数据很有可能是新一轮的技术革命。
  - 数据科学与数据联盟的成立
    - 未来，数据科学将成为一门专门的学科，被越来越多的人所认知
  - 数据泄露泛滥

数据挖掘

- 概念
  - 数据挖掘是指从大量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。
- 特点
  - 基于大量数据
    - 小型数据可以人工分析总结规律
    - 小数据量无法反应真实世界普遍特性
  - 非平凡性
    - 数据挖掘的知识一定是不简单的
  - 隐含性
    - 数据挖掘要发现深藏数据内部的知识，而不是直接浮现在数据表面的知识
  - 新奇性
    - 挖掘的知识在以前是未知的，否则只是验证了经验
  - 价值性
    - 可以为企业带来直接间接的效益
- 数据挖掘步骤
  - 6评价模型和分析
  - 5建立模型
  - 4准备数据
  - 3分析数据
  - 2建立数据挖掘库
  - 1定义问题
- 经典算法
  - 神经网络法
    - 模拟生物神经系统的结构和功能，是一种通过训练来学习的非线性预测模型，它将每一个连接看作一个处理单元，试图模拟人脑神经元的功能，可完成分类、聚类、特征挖掘等多种数据挖掘任务
  - 决策树法
    - 决策树是根据对目标变量产生效用的不同而建构分类的规则，通过一系列的规则对数据进行分类的过程，其表现形式是类似于树形结构的流程图。比如，在贷款申请中，要对申请的风险大小做出判断。
  - 遗传算法
    - 遗传算法模拟了自然选择和遗传中发生的繁殖、交配和基因突变现象，是一种采用遗传结合、遗传交叉变异及自然选择等操作来生成实现规则的、基于进化理论的机器学习方法。
  - 粗糙集法
    - 粗糙集法也称粗糙集理论，是一种新的处理含糊、不精确、不完备问题的数学工具，可以处理数据约简、数据相关性发现、数据意义的评估等问题。
  - 模糊集法
    - 模糊集法是利用模糊集合理论对问题进行模糊评判、模糊决策、模糊模式识别和模糊聚类分析。模糊集合理论是用隶属度来描述模糊事物的属性。系统的复杂性越高，模糊性就越强
  - 关联规则法
    - 关联规则反映了事物之间的相互依赖性或关联性。
- 任务
  - 预测建模
    - 分类：预测离散目标变量
    - 回归：预测连续目标变量
  - 关联分析
    - 用于发现描述数据强关联特征模式
  - 聚类分析
    - 发现紧密相关的观测值族群，发现紧密相关的观测值群组，使得与属于不同族的观察值相比，同一族的观察值相互之间尽可能的类似
  - 异常检测
    - 子主题 1识别其特征显著不同于其他数据的观测值
- 挖掘过程
  - 数据准备
    - 数据集成
    - 数据选择
    - 预处理
  - 数据挖掘
  - 结构表达与解释
- 成功案例
  - 数据挖掘帮助Crediogros Cia Financiera S.A.改善客户信用评分
  - 数据挖掘帮助DHL实时跟踪货箱温度
  - 电信行业应用
- 存在问题
  - 设计保密，法律，伦理问题

云计算

- 概念
  - 狭义上讲，云计算就是一种提供资源的网络，从广义上说，云计算是与信息技术、软件、互联网相关的一种服务，这种计算资源共享池叫做"云"，云计算把许多计算资源集合起来，通过软件实现自动化管理，只需要很少的人参与，就能让资源被快速提供。
- 特点
  - 虚拟化技术
  - 动态可拓展
  - 按需部署
  - 可靠性高
  - 灵活性高
  - 性价比高
  - 可拓展性
- 服务类型
  - 基础设施即服务
  - 平台即服务
  - 软件即服务
- 技术
  - 体系结构
  - 资源监控
  - 自动化部署
- 应用
  - 存储云
  - 教育云
  - 医疗云
  - 金融云

人工智能

- 概念
  - 人工智能（Artificial Intelligence），英文缩写为AI。它是研究、开发用于模拟、延伸和扩展人的智能的理论、方法、技术及应用系统的一门新的技术科学。
- 应用
  - 机器视觉，指纹识别，人脸识别，专家系统，自动规划，智能搜索，定理证明，博弈，语言和图像理解等。人工智能是一门边缘学科，属于自然科学和社会科学的交叉。
- 研究方法
  - 大脑模拟
  - 符号处理
  - 统计学法
  - 集成方法
- 成果
  - 人际对弈
  - 模式识别
  - 自动驾驶
  - 知识工程
    - 智能搜索引擎
    - 计算机视觉/图像处理
    - 机器翻译
    - 数据挖掘
- 发展阶段
  - 计算机时代
  - 大量程序时代
    - 专家系统的出现，专家系统被用于股市预 测，帮助医生诊断疾病，以及指示矿工确定矿藏位置等。
  - 弱人工智能时代
    - 弱人工智能观点认为不可能制造出能真正地推理（REASONING）和解决问题（PROBLEM SOLVING）的智能机器，这些机器只不过看起来像是智能的，但是并不真正拥有智能，也不会有自主意识。主流科研集中在弱人工智能上，并且一般认为这一研究领域已经取得可观的成就。强人工智能的研究则处于停滞不前的状态下。
  - 强人工智能时代
    - 强人工智能观点认为不可能制造出真正能推理（REASONING）和解决问题（PROBLEM SOLVING）的智能机器，并且，这样的机器能将被认为是有知觉的，有自我意识的。强人工智能可以有两类：类人的人工智能，即机器的思考和推理就像人的思维一样。
    - 非类人的人工智能，即机器产生了和人完全不一样的知觉和意识，使用和人完全不一样的推理方式。
- 旗下领域
  - 机器学习
    - 深度学习
      - 深度学习是学习样本数据的内在规律和表示层次，这些学习过程中获得的信息对诸如文字，图像和声音等数据的解释有很大的帮助。它的最终目标是让机器能够像人一样具有分析学习能力，能够识别文字、图像和声音等数据。
      - 深度学习是一类模式分析方法的统称
        - 基于卷积运算的神经网络系统，即卷积神经网络(CNN)
        - 基于多层神经元的自编码神经网络，包括自编码(Auto encoder)以及近年来受到广泛关注的稀疏编码两类
        - 以多层自编码神经网络的方式进行预训练，进而结合鉴别信息进一步优化神经网络权值的深度置信网络(DBN)
      - 特点
        - 强调模式结构的深度，通常有5，5甚至10层隐层节点
        - 明确特征学习的重要性
      - 典型模型
        - 卷积神经网络模型
        - 深度信任网络模型
          - DBN可以解释为贝叶斯概率生成模型，由多层随机隐变量组成，上面的两层具有无向对称连接，下面的层得到来自上一层的自顶向下的有向连接，最底层单元的状态为可见输入数据向量
        - 堆栈自编码网络模型
          - 堆栈自编码网络的结构与DBN类似，由若干结构单元堆栈组成，不同之处在于其结构单元为自编码模型(auto-en-coder)而不是RBM
      - 训练过程
        - 自下上升的非监督学习
          - 就是从底层开始，一层一层地往顶层训练。采用无标定数据（有标定数据也可）分层训练各层参数，这一步可以看作是一个无监督训练过程，这也是和传统神经网络区别最大的部分，可以看作是特征学习过程。
        - 自顶向下的监督学习
          - 就是通过带标签的数据去训练，误差自顶向下传输，对网络进行微调。
      - 应用
        - 计算机视觉
        - 语音识别
        - 自然语言处理
          - word2vector能够更好地表达语法信息。深度学习在自然语言处理等领域主要应用于机器翻译以及语义挖掘等方面
  - 自然语言处理NLP
    - 自然语言是人类智慧的结晶，自然语言处理是人工智能中最为困难的问题之一
    - 主要范畴
      - 文本朗读/语音合成
      - 中文自动分词
        - 文本分类
        - 词性标注
      - 机器翻译
      - 问答系统
      - 信息检索
      - 信息抽取
        - 自动概要
        - 文本校对
      - 自然语言生成
      - 语音识别
    - 研究难点
      - 单词的边界界定
      - 词义的消歧
      - 句法的模糊性
      - 有瑕疵和不规范的输入
      - 语言行为和计划
    - 相关技术
      - 数据稀疏与平滑技术