到额外内容之前是老师画的重点。

数据挖掘的背景、起源、为什么会有这件事情?

- 二十世纪末以来，全球信息量以惊人的速度急剧增长，据估计，每二十个月将增加一倍。四V特征如下：
  - 大量（volume）：量很大，虽然我们现在不用量来区分了。
  - 高速（velocity）：更新很快，现在每时每刻都在产生着新的数据，我们希望得到实时的分析结果。
  - 多样（variety）：数据是几种的结合（结构化、非结构化、半结构化）
  - 真实（veracity）：数据是真实数据，当前采用的数据时用户**真实意图**的反映（所以需要**数据的甄别**）
- 目前的数据库系统虽然可以高效实现数据的录入、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。
- 大数据挖掘技术应运而生并显示出强大生命力
  - **大数据时代，变革医疗卫生**
  - **大数据时代，变革军事决策**
  - **大数据时代，变革生活出行**
- 大数据和普通数据的区别就是其多态性，实际上不会用量来区分。

数据挖掘与传统的数据分析有什么区别?

- **数据分析：**是指用适当的统计分析方法对收集来的大量数据进行分析，提取有用信息和形成结论而对数据加以详细研究和概括总结的过程。这一过程也是质量管理体系的支持过程。在实用中，数据分析可帮助人们作出判断，以便采取适当行动。它是对数据的一种操作手段。或者算法。目标是针对先验的约束，对数据进行整理、筛选、加工，由此得到信息。
- **数据挖掘：**是数据库知识发现中的一个步骤。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。它是对数据分析手段后的信息，进行价值化的分析。

- **数据分析和数据挖掘的区别**在于：
  - 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，这个属于数据分析。你要分析人脸。数据通过你的先验的方法，就是出来个猫脸。你的数据分析也没有问题。你需要默默的承受结果，并 且尊重事实。因此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。
  - 数据挖掘，是对信息的价值化的获取。价值化自然不考虑数据本身，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挖掘。评估，则就是数据挖掘。此时对比数据分析，最大的特点就是，你需要调整你的不同的先验约束，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。而是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。

大数据时代的数据挖掘有什么变化?

- **传统的数据挖掘（Data Mining）定义：**
  - 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。

- **大数据时代的数据挖掘三大关键转变：**
  - **转变一：从数据抽样到全数据**
    - 在大数据时代，我们可以进行数据分析，有时甚至可以处理和某个特别现象相关的所有数据，而不再依赖于随机抽样。
    - **全数据：**
      - 样本 = 总体
      - 传统的**随机抽样**自身存在缺陷
      - 目前已有足够强大的数据采集能力、存储能力和计算能力
  - **转变二：从尽可能精确到允许不精确**
    - 研究数据如此之多，以至于我们不再热衷于追求精确度。

- **允许不精确：**
  - 大数据由于自身特征，必然存在不精确数据，无法避免
  - **量变引起质变：**用数量来纠正错误
  - 对于小规模或特定数据，仍需要精确

- **转变三：从因果关系到相关关系**
  - 我们不再热衷于寻找因果关系，而是寻找相关关系。
  - **相关关系/联系性：**
    - 如果A和B经常同时发生，那么B发生了，就可以预测A也发生了
    - 探求“**是什么**”而不是“为什么”
    - “相关关系提供清晰的视角，而加上因果关系时，很多视角就可能被屏蔽掉。”

注：

如何处理缺失数据（不完全的数据）？

用平均值替代、删除该记录、用该记录的其他数据寻找相似记录填补……

但大数据时代并不需要对缺失数据进行填补。

噪声是什么？

噪声数据未必是坏的数据，只是与一般行为不太一致的数据点（被称为噪声点）

噪声点一般情况下来说是要被识别的，通常代表着不同的意义。

比如，大家都考60分的课，有个人考了90，他就是个噪声。

大数据挖掘的定义？

- 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。

大数据挖掘的主要功能？

七个：关联分析、分类与预测、聚类分析、概念/类别描述、时间序列分析、离群点分析、

演变分析。

1. 关联分析

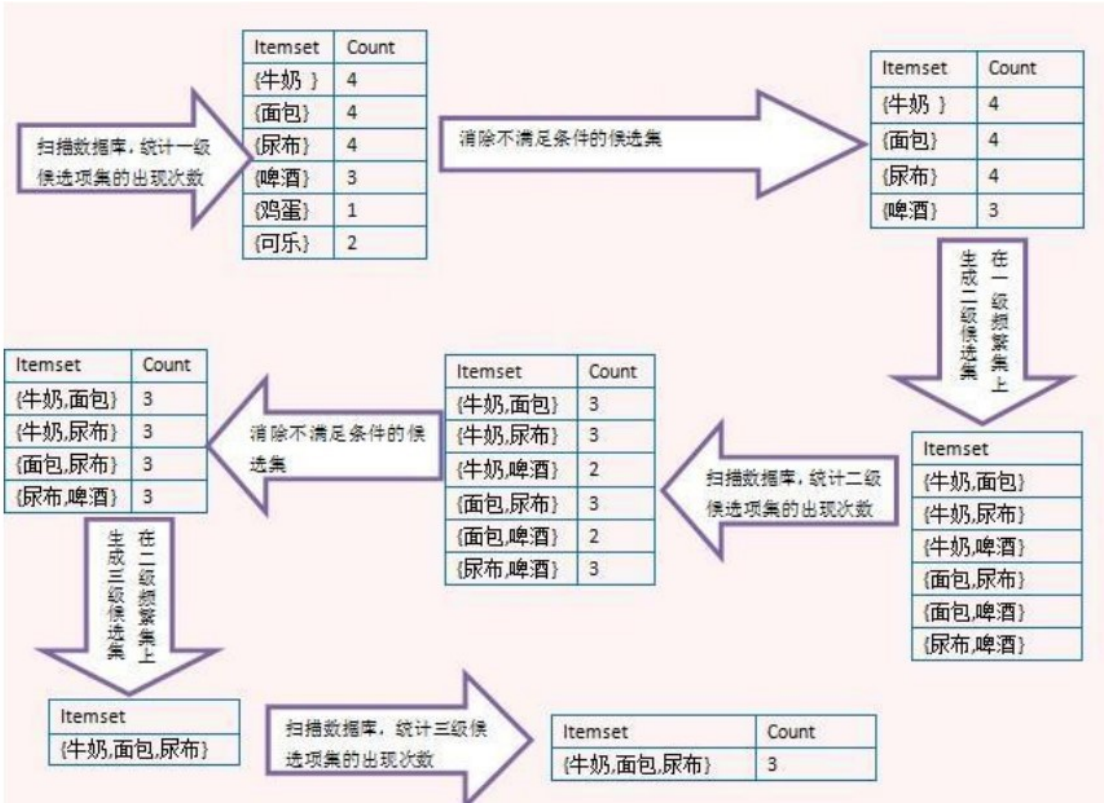
- **定义：**从一个项目集中发现关联规则，该规则显示了给定数据集中经常一起出现的 属性 - 值 条件元组，其在交易数据分析、支持定向市场、商品目录设计和其他业务决策等方面有着广泛的应用。
    - **Apriori算法：**
        i. 一种挖掘关联规则的**频繁**项集算法，（频繁是有定义的，比如一段时间内出现了多少次，由业务经理指定）
        ii. 算法应用广泛，可用于消费习惯分析推荐，在移动通讯中指导业务运营、制定决策等
    - **FP-Growth算法：**
        i. 使用了频繁模式树的数据结构加快整个挖掘过程
        ii. 算法时间和空间复杂度较低
    - **其他算法：**
        i. 基于划分的关联规则挖掘算法
- 例子：
    - 示例1，如下是一个超市几名顾客的交易信息。

TID	Items
001	Cola, Egg, Ham
002	Cola, Diaper, Beer
003	Cola, Diaper, Beer, Ham
004	Diaper, Beer

TID代表交易流水号，Items代表一次交易的商品。

我们对这个数据集进行关联分析，可以找出关联规则{Diaper}→{Beer}。
它代表的意义是，购买了Diaper的顾客会购买Beer。这个关系不是必然的，但是可能性很大，这就已经足够用来辅助商家调整Diaper和Beer的摆放位置了，例如摆放在相近的位置，进行捆绑促销来提高销售量。

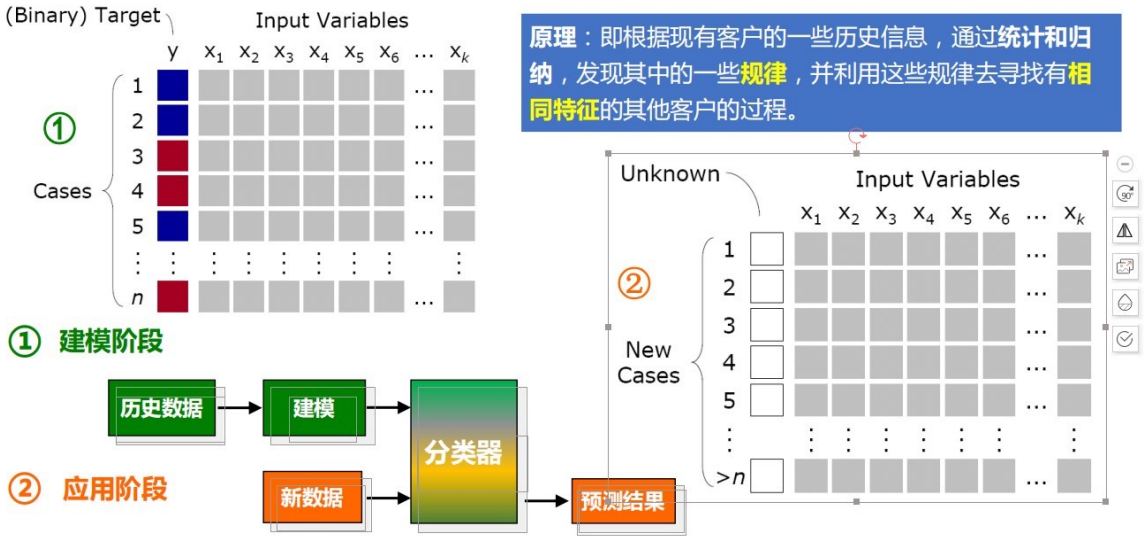
1、事务：每一条交易称为一个事务，例如示例1中的数据集就包含四个事务。
2、项：交易的每一个物品称为一个项，例如Cola、Egg等。
3、项集：包含零个或多个项的集合叫做项集，例如{Cola, Egg, Ham}。
4、k−项集：包含k个项的项集叫做k-项集，例如{Cola}叫做1-项集，{Cola, Egg}叫做2-项集。
5、**支持度计数**：一个项集出现在几个事务当中，它的支持度计数就是几。例如{Diaper, Beer}出现在事务 002、003和004中，所以它的支持度计数是3。
6、**支持度**：支持度计数除于总的事务数。例如上例中总的事务数为4，{Diaper, Beer}的支持度计数为3，所以它的支持度是3÷4=75%，说明有75%的人同时买了Diaper和Beer。
7、频繁项集：支持度大于或等于某个阈值的项集就叫做频繁项集。例如阈值设为50%时，因为{Diaper, Beer}的支持度是75%，所以它是频繁项集。
8、前件和后件：对于规则{Diaper}→{Beer}，{Diaper}叫做前件，{Beer}叫做后件。
9、**置信度**：对于规则{Diaper}→{Beer}，{Diaper, Beer}的支持度计数除于{Diaper}的支持度计数，为这个规则的置信度。例如规则{Diaper}→{Beer}的置信度为3÷3=100%。说明买了Diaper的人100%也买了Beer。
10、强关联规则：大于或等于最小支持度阈值和最小置信度阈值的规则叫做强关联规则。关联分析的最终目标就是要找出强关联规则 [2] 。

- 

- **关联规则生成：**得到了频繁项集，而此时的任务就是在频繁项集里面挖掘出大于最小置信度阈值的关联规则。怎么挖呢？把频繁项集分成前件和后件两部分，然后求规则前件→后件的置信度，如果大于最小置信度阈值，则它就是一条强关联规则。但是把频繁项集分成前件和后件的情况有很多，我们可以对其进行一些优化。

2. 分类与预测

- **分类定义：**分类指通过分析一个类别已知的数据集（训练集）的特征来建立一组模型，该模型可用以预测类别未知的数据项（测试集）的**类别**。（常用分类树）
  - 注意：训练集和测试集的属性集合和类型必须是完全一致的，匹配上才行。
- **预测定义：**预测与分类类似，只不过它要预测的不是类别，而是一个**连续的数值**。
- **分类器生成过程原理：**即根据现有客户的一些历史信息，通过统计和归纳，发现其中的一些**规律**，并利用这些规律去寻找有**相同特征**的其他客户的过程。

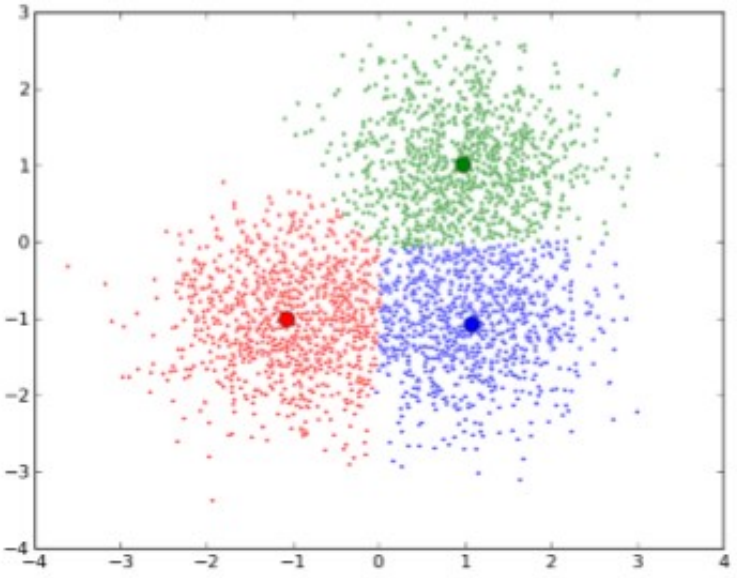


- 通过分析**训练集**中的数据，为每个类别建立分类分析模型，用这个分类模型对数据库中其他记录（**测试集**）进行分类
  - 决策树预测技术
    - 以ID3、C4.5为代表的决策树预测模型
    - Random Forests（随机森林），准确性和健壮性较高
  - 回归预测技术
    - Logistic回归模型，主要用于二元离散型变量的预测
  - 其他预测技术
    - 神经网络：Kohonen 神经网络
    - SVM支持向量机：可用于二元分类的SVM模型

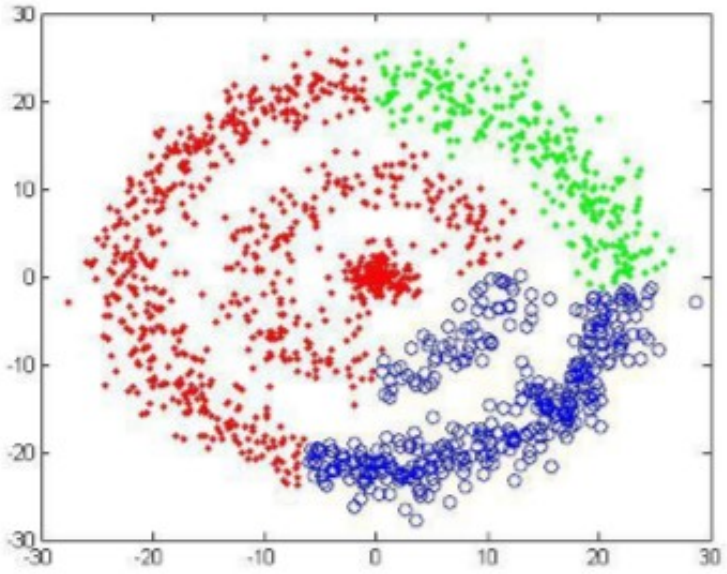
3. 聚类分析

- **定义：**又称为“同质分组”或者“无监督的分类”，指把一组数据分成不同的“簇”，簇中数据相似，簇间数据距离较远。相似性可以由用户或者专家定义的距离函数加以度量。
  - 集合里每个元素都有类标识的、或者是用于预测的，是分类。
  - 聚类中，把当前集合分成几个簇是不知道的。
  - 聚类时，所有属性都考虑在内，共同计算相似性的结果。**单个属性一定不是聚类。**
- **聚类技术：**
  - 基于约束的聚类

        - 有先验知识的情况下，加上约束的聚类。（手动设置相似度，比如给不想聚类的设置0，想聚类的设置无限大）
        - 聚类是先聚，再找某一簇的特征，与业务对接。
        - 例：
            - ①把班里同学分成男同学和女同学，这既不是分类也不是聚类，这根本就不是一个数据挖掘问题，是一个简单的查询而已。
            - ②根据当前同学的特征得到五个族群，目标是给这五个不同族群的同学制定个性化培养方案，这是聚类问题。
    - 基于距离的聚类
        - K-Means算法，特点是：所有簇的形成依据为簇成员之间的距离，根据不同的距离函数，算法适用性有所差异
    - 基于密度的聚类
        - DBSCAN算法，检测密度区域，可识别任意形状的簇并过滤噪声点
    - 其他聚类技术
        - 如基于划分的聚类
        - 基于模型的聚类（神经网络）等
    - 例子：
        - 正确例子：



- 反例：下面这张图分簇就错了（说明方法不对），应该是按环分簇。



注：分类与聚类的区别

①目标不同，应用场景不同

②分类是有监督（开始时的数据集有标识）的学习过程，聚类是无监督的学习过程。

③数据集不一样，分类有训练集（带标识）和测试集（不带标识），聚类无这种区分，只有一个数据集。

④类别标识的含义不同，分类操作开始前每个类别都标清了，聚类开始前不知道

⑤方法的原理不一样，分类：从训练集中提取隐含其中的对于类别的共同信息以分类器的模式展示出来，在数据集上完成操作；聚类：同质分组，根相似性度量的方式对一组训练集进行聚类，要求组间尽量差距大，组内尽量差异小。

4. 概念/类别描述

5. 时间序列分析

6. 离群点分析

a. 离群点（outlier），也叫孤立点，是行为表现和大多数点不一样的点。

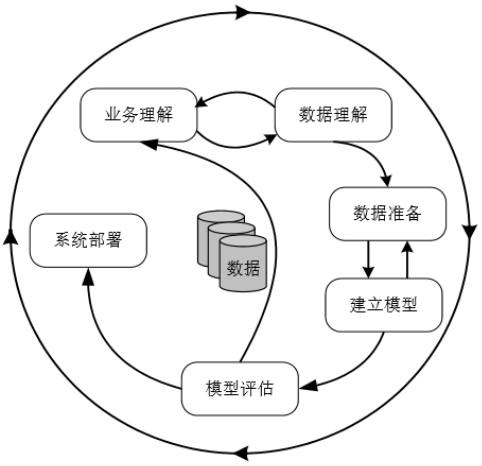
b. 业务分析时，离群点或许更应该被关注。

c. 离群点寻找：算与其他点的距离或局部密度

7. 演变分析

a. 比如体重的变化、成绩的改变……这是在学校里做成绩分析最合适的方法

大数据挖掘的标准流程?

- **CRISP - DM 模型（Cross Industry Standard Process for Data Mining）**最先在1996年被提出
  - 目的：把数据挖掘的过程标准化，使数据挖掘项目的实施速度更快、成本更低、更可靠并且更容易管理。
- **六个步骤：**业务理解、数据理解、数据准备、建立模型、模型评估、系统部署。
  - 

- **业务理解**
  - **确定业务目标**
    - 分析项目的背景，从业务视点分析项目的目标和需求，确定业务角度的成功标准。
  - **项目可行性分析**
    - 分析拥有的资源，条件和限制，风险估计，成本和效益估计。
  - **确定数据挖掘目标**
    - 明确确定数据挖掘的目标和成功标准，数据挖掘的目标和业务目标是不一样的，前者指技术上的，例如生成一棵决策树等。
  - **提出项目计划**
    - 对整个项目做一个计划，初步估计用到的工具和技术。

- **数据理解**
  - **最好就按部就班过一遍，最稳妥**
    - **收集原始数据**
      - 收集本项目所涉及到的数据，如有必要，把数据装入数据处理工具，并作一些初步的数据集成的工作，生成相应报告。
    - **描述数据（熟练掌握）**
      - 对数据做一些大致的描述，例如记录数、属性数等，给出相应报告。
        - 如：数据表的数据来源、数据的规模有多大（多少行多少列）、每个属性是什么类型（连续型？离散型？）
        - 离散型又可分二元（取值只有两个）和非二元
        - 二元又可分对称（如果两个属性的取值权重、地位、作用一样，则是对称的二元变量）和非对称
    - **探索数据**
      - 对数据做简单的统计分析，例如关键属性的分布等。
    - **检查数据质量**
      - 包括数据是否完整、数据是否有错、是否有缺失值等问题。
- **数据准备**
  - **数据选择**
    - 根据数据挖掘目标和数据质量选择合适的数据，包括表的选择、记录选择和属性选择。
  - **数据清洁**
    - 提高选择好的数据的质量，例如去除噪音，估计缺失值等。
  - **数据创建**
    - 在原有数据的基础上是生成新的属性或记录。
  - **数据合并**
    - 利用表连接等方式将几个数据集合并在一起。
  - **数据格式化**
    - 把数据转换成适合数据挖掘处理的格式。如：

        - （建议数据挖掘前做）**归一化**：数据都转到0~1之间，防止属性取值单位不同导致结果不同导致影响。
        - **归一化公式：x' = (x - min)/(max-min)**
    - **……**
- **建立模型**
    - **选择建模技术**
        - 确定数据挖掘算法和参数，可能会利用多个算法。
    - **测试方案设计**
        - 设计某种测试模型的质量和有效性的机制。
    - **模型训练**
        - 在准备好的数据集上运行数据挖掘算法，得出一个或者多个模型。
    - **模型测试评估**
        - 根据测试方案进行测试，从数据挖掘技术的角度确定数据挖掘目标是否成功。（这里所有的检验都是技术层面的）
        - **模型的评价标准案例（以一个分类器为例）**
            1. 精确度
            2. 效率高
            3. 复杂度
            4. 可解释性
- **模型评估**
    - **结果评估**
        - 从商业角度评估得到的模型，甚至实际试用该模型测试其效果。
    - **过程回顾**
        - 回顾项目的所有流程，确定每一个阶段都没有失误。（失误了就回溯到上一步，甚至可以回溯到第一步）
    - **确定下一步工作**
        - 根据结果评估和过程回顾得出的结论，确定是部署该挖掘模型还是从某个阶段

开始重新开始。

- **系统部署**
  - **部署计划**
    - 对在业务运作中部署模型作出计划。
    - 监控和维护计划：如何监控模型在实际业务中的使用情况，如何维护该模型
  - **做出最终报告**
    - 项目总结，项目经验和项目结果。
    - 项目回顾：回顾项目的实施过程，总结经验教训；对数据挖掘的运行效果做一个预测。

大数据挖掘的应用范围?

- 亚马逊的信息公司
- 运营商分析用户行为
- Twitter看用户兴趣与情绪
- 医疗中的数据挖掘
- 生活中的数据挖掘如赌徒预测、时尚预测、堵车语言、音乐推荐
- 京东全业务数据分析和应用：用户画像、供应链优化、智能网站
- 智能交通
- 决策管理
- 智能应用：电力、医疗、城市、交通、供应链、银行业

额外内容

大数据的难题：如何将半结构化和非结构化的数据进行数据的整合。

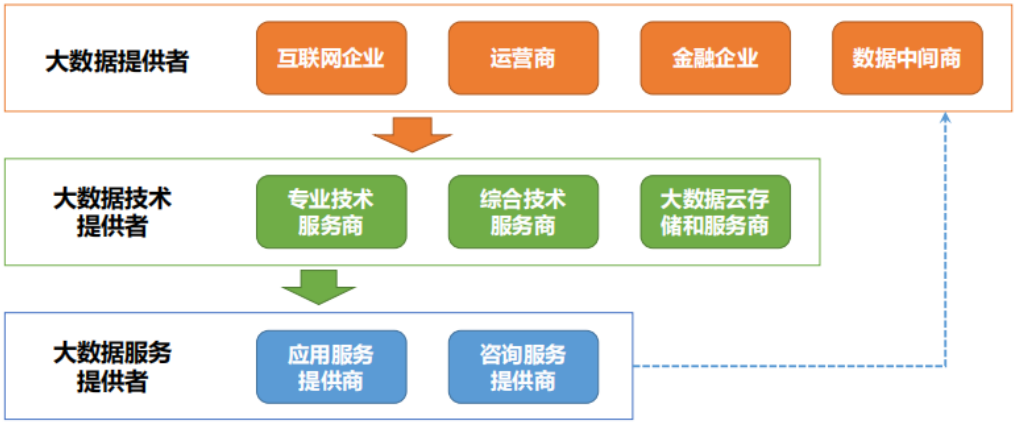
可考虑做降维处理，比如将对商品的描述降维一个数组。文本：文章→段落→句子→词→词频。但降维处理一定会有数据的丢失。

现在可以分开处理，结构化与结构化处理，半结构化与半结构化处理……

好的聚类结果

类内具有较小相异度，类间具有较大相异度，这样的聚类结果是比较好的。

大数据产业链与体系架构



大数据安全问题及应对策略

网络化易攻击

- 网络化使各个行业领域实现资源共享和数据互通；
- 正因为平台的互通使蕴含海量数据和潜在价值的大数据更容易吸引黑客的攻击；
- 对于攻击者而言，相对低的成本可以获得“滚雪球”的收益。

非结构化数据发展

- 大数据之前通常将数据存储分为关系型数据库和文件服务器两种。
- NoSQL数据存储力不从心：一是NoSQL无法沿用SQL的模式，适应难；二是NoSQL软件使用新代码，漏洞多多；三是由于NoSQL服务器软件没有内置足够的安全。

技术发展因素

- 大数据本身安全存在漏洞，权限控制以及密钥生成、存储和管理方面的不足都可能造成数据泄漏。
- 被隐藏在大数据中的恶意软件和病毒代码很难发现，形成永久可持续攻击。
- 攻击技术提高，攻击者利用大数据技术进行攻击。

存储安全	应用安全	管理安全
• 采用虚拟化海量存储技术来存储数据资源。 • 数据加密，通过SSL加密实现移动保护大数据。 • 分离密钥和加密数据。 • 使用过滤器。通过过滤器的监控，一旦发现数据离开用户网络自动阻止数据传输。 • 数据备份。	• 防止APT攻击，设计全流量审计方案，提醒隐藏有病毒的应用程序。 • 用户访问控制。根据大数据的密级程度和用户需求的不同设定不同的权限等级，严格控制访问权限。 • 整合工具和流程，确保大数据应用安全处于大数据系统的顶端。 • 数据实时分析引擎，从大数据中第一时间挖掘各类安全事件发出警告响应。	• 规范建设，一套规范的运行机制、建设标准和共享平台建设至关重要。 • 建立以数据为中心的安全系统。 • 融合创新，积极创造大数据公司技术融合平台，特别是在数据挖掘、人工智能 、机器学习等新技术的创新应用融合创新。

大数据发展趋势

- 超级运营商
- 移动互联网定向推荐
- 多类型数据挖掘
- 大数据与人工智能
- 大数据可视化
- 其他数据挖掘研究领域如基因组学、运动科学、生物科学、心理健康、天气预报……