# 目录


- 第一章 绪论
- 第二章 搜索技术
- 第三章 知识表示
- 第四章 推理技术
- 第五章 机器学习
- 第六章 数据挖掘
- 第七章 计算智能
- 第八章 智能体技术

- 数据挖掘定义与发展
- 数据挖掘方法
- 数据挖掘技术
- Web数据挖掘
- 大数据

# 数据挖掘定义与发展

数据库知识发现（Knowledge Discovery in Databases， KDD）

数据挖掘(Data Mining DM)

数据分析（Data Analysis）

数据融合(Data Fusion)

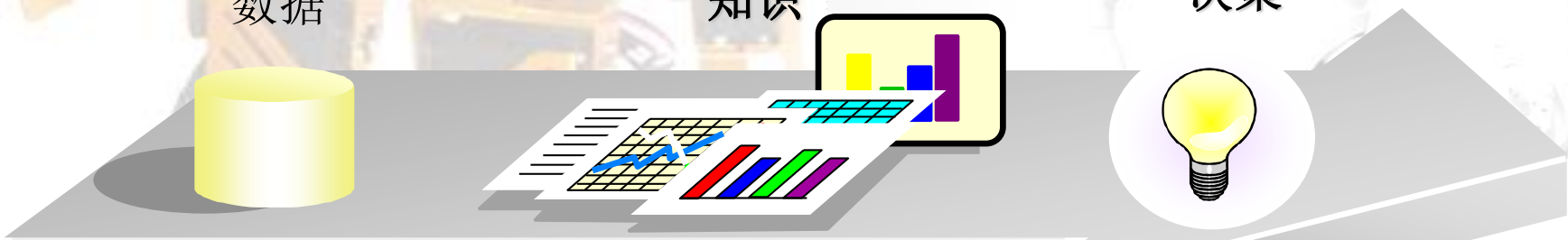
决策支持（Decision Supporting）

# 数据挖掘定义与发展

## 数据挖掘的产生和发展

苦恼：淹没在数据中 ； 不能制定合适的决策！

数据 知识 决策



- 金融
- 经济
- 人口统计
- 生命周期

- 模式
- 趋势
- 模型
- 关联规则
- 序列

- 目标市场
- 资金分配
- 销售的地理位置

**数据爆炸，知识贫乏**

# 数据挖掘定义与发展

- 1989 IJCAI会议： 数据库中的知识发现讨论专题
  - Knowledge Discovery in Databases (G. Piatetsky-Shapiro and W. Frawley, 1991)
- 1991-1994 KDD讨论专题
  - Advances in Knowledge Discovery and Data Mining (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1996)
- 1995-1998 KDD国际会议 (KDD'95-98)
  - Journal of Data Mining and Knowledge Discovery (1997)
- 1998 ACM SIGKDD, SIGKDD'1999-2002 会议,以及SIGKDD Explorations
- 数据挖掘方面更多的国际会议
  - PAKDD, PKDD, SIAM-Data Mining, (IEEE) ICDM, DaWaK, SPIE-DM, etc.

# 数据挖掘定义与发展

## 知识发现的定义

**Fayyad, Piatetsky-Shapiro和Smyth在KDD96国际会议的会议论文《From Data Mining to Knowledge Discovery》一文中将KDD定义为：**

**“The nontrivial process of identifying valid，novel，potentially useful，and ultimately understandable patterns in data.”**

**KDD指大量数据中获取有效的、新颖的、有潜在作用的和最终可理解的模式的非平凡过程。**

# 数据挖掘定义与发展

（1）数据集：是指一个有关事实$F$的集合，它是用来描述事物有关方面的信息，是进一步发现知识的原材料。数据可以是一个或一组数据库、数据仓库、电子表格或其他类型的信息库，在数据上往往需要进行数据清理、集成和规约等预处理。

（2）新颖：经过知识发现提取出的模式必须是新颖的，至少对系统来说应该如此。模式是否新颖可以通过两个途径来衡量：其一是在所得到的数据方面，通过对比当前得到的数据和以前的数据或期望得到的数据之间的比较，来判断该模式的新颖程度；其二是在其内部所包含的知识方面，通过对比，发现的模式与已有的模式的关系来进行判断。

# 数据挖掘定义与发展

（3）潜在有用：提取出的模式应该是有意义的，有潜在的应用价值。这可以通过某些函数的值来衡量。

（4）可理解：知识发现的一个目标就是将数据库中隐含的模式以容易被人理解的形式表现出来，从而帮助人们更好地了解数据库中所包含的信息。

（5）模式：模式是指用语言来表示的一个表达式，它可用来描述数据集的特性，根据某种兴趣度度量，并于数据挖掘模块中进行交互挖掘，以便识别和表示知识的真正有趣的模式。

# 数据挖掘定义与发展

（6）过程：过程是在KDD中包含的步骤，如数据的预处理、模式搜索、知识表示及知识评估、过程优化等。

（7）非平凡：是对数据进行更深层处理的过程，已经超越了一般封闭形式的数量计算，包括对结构、模式和参数的搜索。

（8）有效性：通过KDD从当前数据所发现的模式必须有一定的正确程度，否则KDD就毫无作用。

# 数据挖掘定义与发展

## 知识发现的处理过程

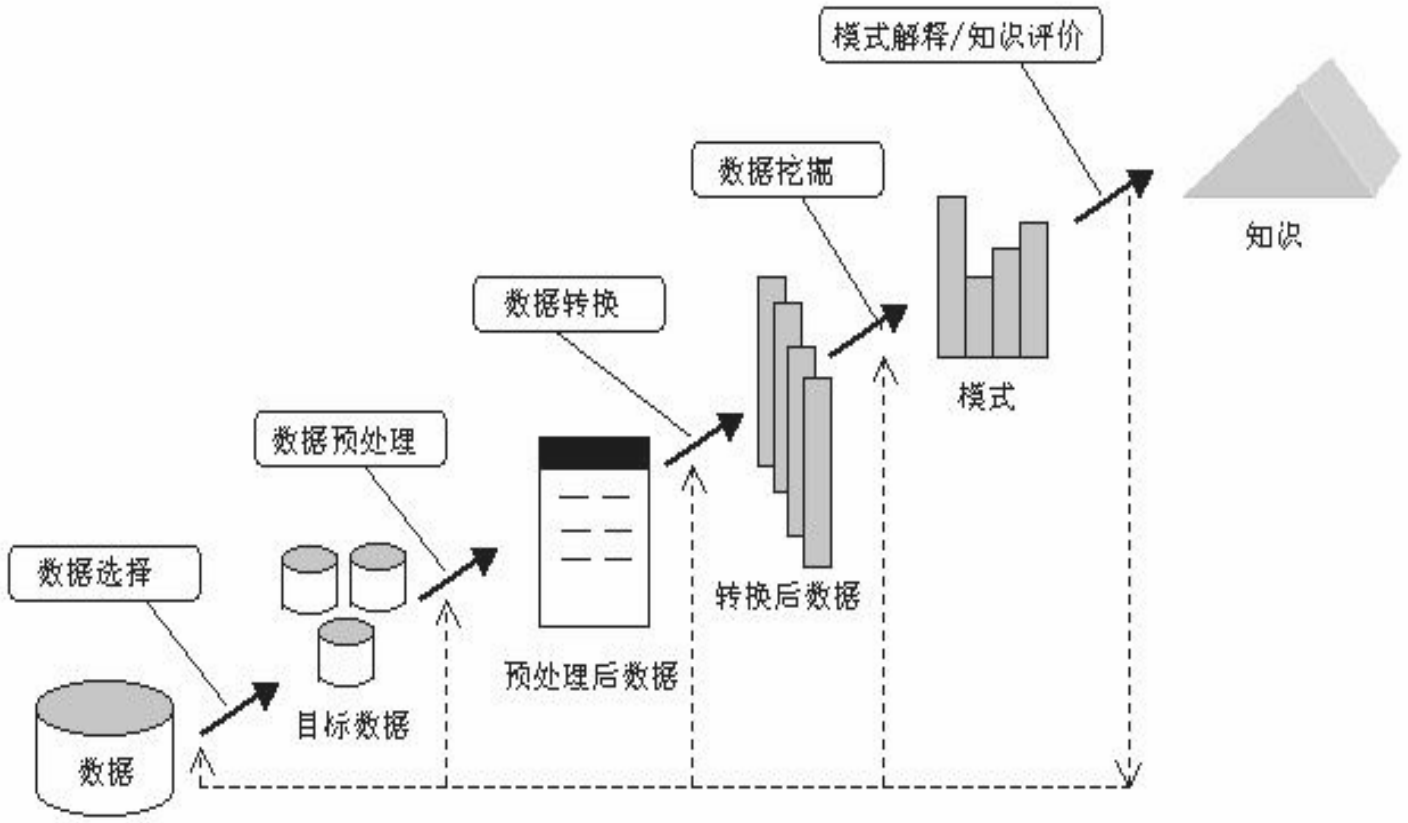



图　数据库中知识发现的处理过程模型

# 数据挖掘定义与发展

（1）数据选择。根据用户的需求从数据库中提取与KDD相关的数据。

（2）数据预处理。主要是对上述数据进行再加工，检查数据的完整性及数据的一致性，对丢失的数据利用统计方法进行填补，形成发掘数据库。

（3）数据转换。从发掘数据库里选择数据，即根据知识发现的任务对数据进行再处理，主要通过投影或数据库中的其他操作减少数据量。

# 数据挖掘定义与发展

（4）数据挖掘：

- 确定KDD目标： 根据用户要求，确定KDD发现的知识类型，因为对KDD的不同要求，会在具体的知识发现过程中采用不同的知识发现算法。
- 确定知识发现算法： 根据阶段5所确定的任务，选择合适的数据挖掘算法，包括选取合适的模型和参数，并使得挖掘算法与整个KDD的评判标准相一致。
- 数据挖掘： 运用选定的挖掘算法，搜索或产生一个特定的感兴趣的模式或数据集，从数据中提取出用户所需要的知识，这些知识可以用某种特定的方式表示或使用一些常用的表示方式，如产生式规则等。

# 数据挖掘定义与发展

（5）模式解释： 对发现的模式进行解释，去掉多余的不切题意的模式，转换成某个有用的模式，以使用户理解。在此过程中，为了取得更为有效的知识，可能会返回前面处理中的某些步骤，以便反复提取，从而提取出更有效的知识。

（6）知识评价：这一过程主要用于对所获得的规则进行价值评定，以决定所得的规则是否存入基础知识库。

上述KDD全过程的几个步骤可以进一步归纳为三个步骤，即数据挖掘预处理（数据挖掘前的准备工作）、数据挖掘、数据挖掘后处理（数据挖掘后的处理工作）。

# 数据挖掘定义与发展

## 数据挖掘软件

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、Microsoft SQL Server 2005等。

# 数据挖掘的方法

## 1. 统计方法

统计方法是从事物的外在数量上的表现去推断该事物可能的规律性。

（1）传统方法

渐近理论：当样本趋于无穷多时的统计性质

三个阶段：搜集数据、分析数据、推理

常用方法：

回归分析（多元分析、自回归）、 判别分析（贝叶斯判别、费歇尔判别、非参数判别）、聚类分析（系统聚类、动态聚类）、探索性分析（主元分析法、相关分析法）

# 数据挖掘的方法

（2）模糊集

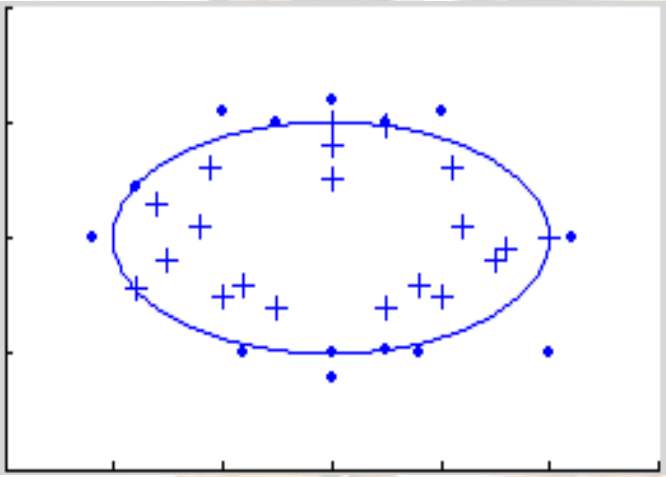
开发数据的不确定性模型

（3）支持向量机

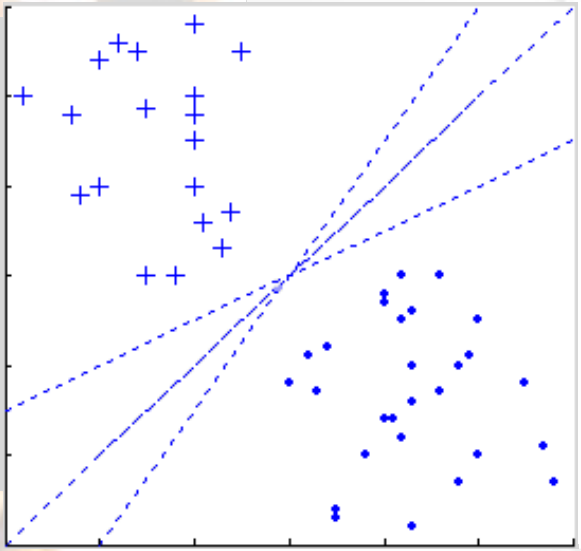
支持向量机（support vector machine, SVM）建立在统计学习理论和结构风险最小化原则之上，其主要思想是针对两类分类问题，在高维空间中寻找一个超平面作为两类的分割，以保证最小的分类错误。

# 数据挖掘的方法

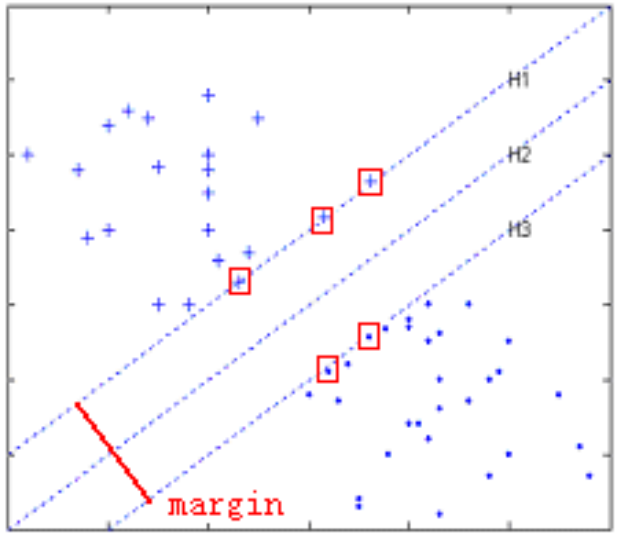
## （3）支持向量机



线性不可分



不同的分类超平面




最优分类超平面及其间隔

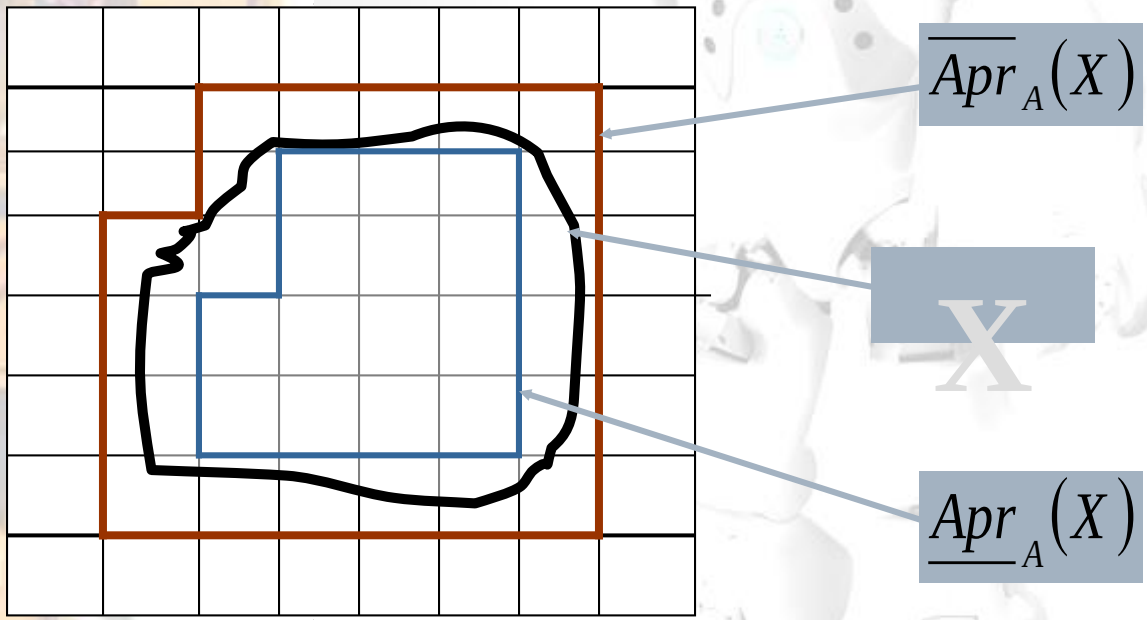
# 数据挖掘的方法

（4）粗糙集

粗糙集合理论(Rough Set, 也称为RS理论)由波兰数学家Pawlak.Z于1982年提出。粗糙集对不精确概念的描述是通过上近似（upper approximation）和下近似（lower approximation）这两个精确概念来实现的。一个概念（或集合）的下近似是指其中的元组肯定属于该概念；一个概念（或集合）的上近似是指其中的元组可能属于该概念。

粗糙集方法优点：不需要预先知道的额外信息，如统计中要求的先验概率和模糊集中要求的隶属度，算法简单，易于操作。

# 数据挖掘的方法




集合的上、下近似概念示意

# 数据挖掘的方法

## 2. 机器学习方法

可能用于机器发现的机器学习方法有：

（1） 规则归纳。

规则反映数据项中某些属性或数据集中某些数据项之间的统计相关性。

（2）决策树。

决策树的每一个非终叶节点表示所考虑的数据项的测试或决策。

（3）范例推理。

范例推理是直接使用过去的经验或解法来求解给定的问题。

# 数据挖掘的方法

（4） 贝叶斯网络。

贝叶斯信念网络是概率分布的图表示。贝叶斯网络基于后验概念的贝叶斯定理，是建立在数据进行统计处理基础上的方法，将不确定事件通过网络连接起来，可以对其他相关事件的结果进行预测，其网络变量可以是可见的，也可隐藏在训练样本中。贝叶斯网络具有分类、聚类、预测和因果关系分析的功能，其优点是易于理解，预测效果较好，缺点是对发生频率很低的事件预测效果不好。

# 数据挖掘的方法

（5）遗传算法。

在求解过程中，通过最好解的选择和彼此组合，使期望解的集合愈来愈好。

**3. 神经计算方法**

**4. 可视化方法**

可视化（**visualization**）就是把数据、信息和知识转化为可视的表示形式的过程。

# 数据挖掘技术

按数据挖掘任务分类

- ⑩描述（Description）：了解数据中潜在的规律
- ⑩预言（Predication）：用历史预测未来

数据挖掘技术

- 概念/类描述
- 关联规则分析
- 分类（预言）
- 聚类
- 序列模式
- 异常检测

# 数据挖掘技术

## 1. 概念/类描述（Concept Description)

特征化和区分（Characterization and Comparision)

概念或类别描述使用汇总的、简洁的、精确的方式描述每个类和概念，可通过前面的方法得到：

（1）数据特征化，一般地汇总所研究类的数据；

（2）数据区分，将目标类与一个或多个比较类进行比较；

（3）数据特征化和比较，两者的结合。

数据特征的输出可以用多种形式输出，包括扇形图、条图、曲线、多位数据立方体和交叉表在内的多维表。结果描述也可以用概括关系或关联规则形式来表示。

## 2. 关联分析(Association Rules)

关联规则分析就是发现关联规则，在交易数据、关系数据或其他信息载体中，查找存在于项目集合或对象集合之间的频繁模式、关联、相关性、或因果结构。

规则形式：

**Body ® Head [support, confidence]**

例：

**buys(x, “diapers”) ® buys(x, “beers”) [0.5%, 60%]**

**major(x, “CS”) ^ takes(x, “DB”) ® grade(x, “A”) [1%, 75%]**

# 数据挖掘技术

支持度*s*，一次交易中包含{A,B}的可能性

$$Support(A \Rightarrow B) = P(A \cup B);$$

可信度 *c*，包含{A}的交易中也包含B的条件概率

$$Confidence\ (A \Rightarrow B) = P(B|A)$$

同时满足大于等于最小支持度阈值（min_support）和最小可信度（min_confidence）的规则称作强规则。

满足大于等于最小支持度(min_support), 称项目集$X \subseteq I$是频繁项目集(Frequent Itemset)。

# 数据挖掘技术

| 交易ID | 购买商品 |
|---|---|
| 2000 | A,B,C |
| 1000 | A,C |
| 4000 | A,D |
| 5000 | B,E,F |

最小支持度 50%
最小可信度 50%

| 频繁项集 | 支持度 |
|---|---|
| {A} | 75% |
| {B} | 50% |
| {C} | 50% |
| {A,C} | 50% |

对于 $A \Rightarrow C$：

**support = support({*A* 、 *C*}) = 50%**

**confidence = support({*A* 、 *C*})/support({*A*}) = 66.6%**

# 数据挖掘技术

## Apriori算法

基本思想: 频繁项集的任何子集也一定是频繁的。

算法的核心:

用频繁的($k-1$)-项集生成候选的频繁 $k$-项集

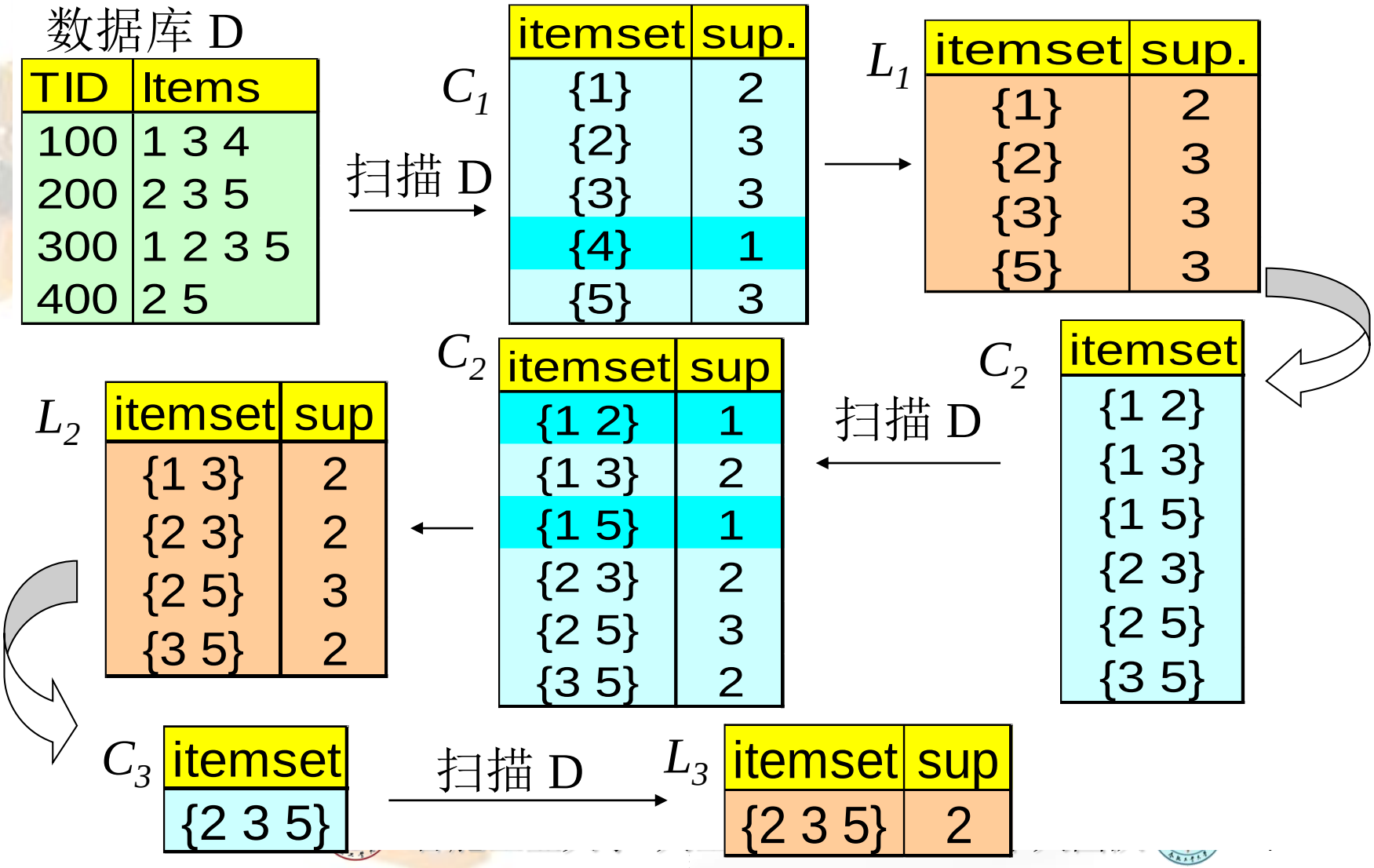
用数据库扫描和模式匹配计算候选集的支持度

算法瓶颈: 候选集生成

巨大的候选集:

多次扫描数据库:

# 数据挖掘技术


数据库 D
TID Items
100 1 3 4
200 2 3 5
300 1 2 3 5
400 2 5
扫描 D
$C_1$
itemset sup.
{1} 2
{2} 3
{3} 3
{4} 1
{5} 3
$L_1$
itemset sup.
{1} 2
{2} 3
{3} 3
{5} 3
$C_2$
itemset
{1 2}
{1 3}
{1 5}
{2 3}
{2 5}
{3 5}
扫描 D
$C_2$
itemset sup
{1 2} 1
{1 3} 2
{1 5} 1
{2 3} 2
{2 5} 3
{3 5} 2
$L_2$
itemset sup
{1 3} 2
{2 3} 2
{2 5} 3
{3 5} 2
$C_3$
itemset
{2 3 5}
扫描 D
$L_3$
itemset sup
{2 3 5} 2

## 3. 分类(Classification)

分类是找出描述并区分数据类或概念的分类函数或分类模型(也常常称作分类器)，该模型能把数据库中的数据项映射到给定类别中的某一个，以便能使用模型预测类标记未知的对象。

常用的分类方法：

（1） 信息论方法

ID3方法 ------决策树方法

利用信息论中信息增益寻找数据库中具有最大信息量的字段，建立决策树的一个节点，并根据字段的不同取值建立树的分枝，在每个分枝子集中重复建树的下层节点。

# 数据挖掘技术

（2） 集合论方法

粗集方法、概念格方法

（3） 人工神经网络方法

① 前馈网络：含感知机.反向传输模型.函数式网络。

② 反馈式网络：用于联想记忆和优化计算。

③ 自组织网络：用于聚类。

（4）遗传算法：模拟生物进化过程的方法。

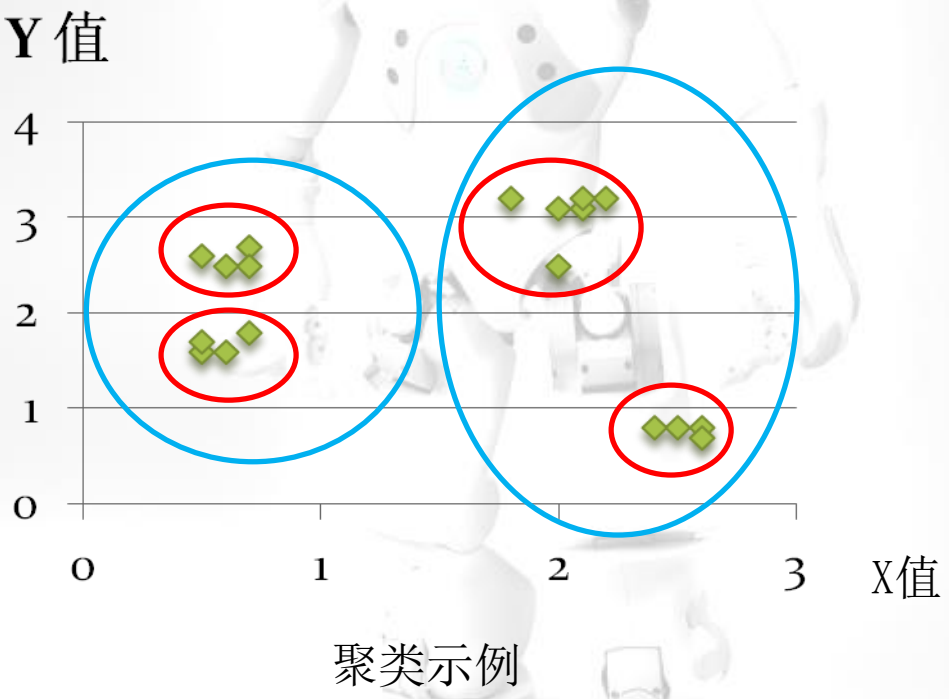
（5）统计分析方法：贝叶斯网，线性回归分析，线性判别分析，聚类分析，差异分析，因子分析等。

# 数据挖掘技术

## 4. 聚类(Clustering)

聚类是把数据按照相似性归纳成若干类别，同一类中的数据彼此相似，不同类中的数据相异。

聚类是一种无监督分类法， 没有预先指定的类。




聚类示例

# 数据挖掘技术

**与分类的区别：**

分类依赖于预先定义的类和带类标号的训练实例，是一种观察式 的学习；而聚类是找到这个簇的特征或者标号的过程。

**一个有效的聚类算法必须满足两个条件：**

类内数据对象的强相似性，通常用紧致度描述；

类间数据对象的弱相似性，常采用分离度描述。

# 数据挖掘技术

## 聚类算法的分类

聚类分析算法取决于数据的类型、聚类的目的和应用。

### （1）基于划分方法

给定一个包含n个对象的数据集和要构建的划分数目k，划分方法首先创建一个初始划分，然后采用一种迭代的重定位技术，尝试通过对象在划分间的移动来改进划分

### （2）基于层次方法

层次聚类是将数据集分解成几级进行聚类，层的分解可以用树形图来表示以任一样本

# 数据挖掘技术

## （3）基于密度的方法

点为基础，当该点的给定邻域内包含的数据点个数超过某一给定阈值时，就以其邻域中的数据点为基础继续进行广度或深度探索，扩展簇的大小。

## （4）基于网格的方法

基于网格的聚类算法的特点是采用一个多分辨率的网格数据结构，从而在该网格结构上进行聚类。

## （5）基于模型的方法

基于模型的方法为每个类假定了一个模型，并试图寻找数据对给定模型的最佳拟合。

## K-means算法

（1）从D中随机取k个元素，作为k个簇的各自的中心。

（2）分别计算剩下的元素到k个簇中心的相似度，将这些元素分别划归到相似度最高的簇。

（3）根据聚类结果，重新计算k个簇各自的中心。

（4）将D中全部元素按照新的中心重新聚类。

（5）重复第4步，直到聚类结果不再变化。

（6）将结果输出。

# 数据挖掘技术

例：现有一个数据集{1，2，30，15，10，18，3，9，8，25}，用K-means算法将这些数据聚类。

解：设k=3，即将数据集聚成3类。随机选取3个数作为初始簇均值：m1=9，m2=8，m3=25，开始迭代。

相似度度量采用的距离值为两个数的差的绝对值。

第一次迭代得到3个簇是

K1={1，2，3，8}， k2={9，10，15}， k3={18，25，30}

重新计算每个簇的均值，则均值更新为m1=3.5，m2=11.3，m3=24.3

第二次迭代 得到3个簇

K1={1，2，3}， k2={8， 9，10，15}， k3={18，25，30}

新的均值为m1=3.5，m2=11.3，m3=24.3

# 数据挖掘技术

第三次迭代得到3个簇是

K1={1，2，3}， k2={8， 9，10，15，18}， k3={25，30}

新的均值为m1=2，m2=12，m3=27.5

第四次迭代 得到3个簇

K1={1，2，3}， k2={8， 9，10，15，18}， k3={25，30}

每个簇的数据不再变化，达到稳定，算法终止。

# 数据挖掘技术

## 相似性度量

**(1)欧几里德距离(Euclidean Distance)**

$$Dist(x_i, x_j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{id} - x_{jd})^2}$$

**(2)曼哈顿距离(Manhattan Distance)**

$$Dist(x_i, x_j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{id} - x_{jd}|$$

# 数据挖掘技术

**(3)明考斯基距离(Minkowski Distance)**

$$Dist(x_i, x_j) = \left( \left| x_{i1} - x_{j1} \right|^h + \left| x_{i2} - x_{j2} \right|^h + \cdots + \left| x_{id} - x_{jd} \right|^h \right)^{\frac{1}{h}} \quad (h>0 \wedge h \in Z)$$

**(4)夹角余弦距Ig(Cosine Distance)**

$$Dist(x_i, x_j) = 1 - sim(x_i, x_j) = 1 - \frac{\sum_{k=1}^{d} (x_{ik} \cdot x_{jk})}{\sqrt{\sum_{k=1}^{d} x_{ik}^2 \cdot \sum_{k=1}^{d} x_{jk}^2}}$$

## 5.序列(Sequence)模式

序列模式是指通过时间序列搜索出的重复发生概率较高的模式。

时间序列模式根据数据随时间变化的趋势预测将来的值。这里要考虑到时间的特殊性质，像一些周期性的时间定义如星期、月、季节、年等，以及不同的日子如节假日可能造成的影响，日期本身的计算方法，还有一些需要特殊考虑的地方如时间前后的相关性(过去的事情对将来有多大的影响力)等。

# 数据挖掘技术

例：顾客租借影碟的一个典型的顺序是先租“星球大战”，然后是“帝国反击战”，再是“杰达武士归来”(这三部影片是以故事发生的时间先后而情节连续的)。值得注意的是租借这三部电影的行为并不一定需要是连续的。在任意两部之间插租了任何电影，仍然满足这个序列模式，并且扩展一下，序列模式的元素也可以不只是一个物品(如一部电影)，它也可以是一个物品的集合。

## 6.异常(Outlier)检测

异常检测是用来发现”小的模式”(相对于聚类)，即数据集中间显著不同于其它数据的对象。

常用方法：

基于统计（statistical-based)的方法

基于距离 (distance-based)的方法

基于偏差(deviation-based)的方法

基于密度(density-based)的方法

# 6.8 知识发现

## 6.8.7 Web数据挖掘

### 1. Web数据挖掘定义

网络数据资源种类：

第一类是内容(Content)，即网页上的真正数据；

第二类是结构(Structure)，即描述内容组织的数据；结构信息包括各种HTML或XML标记及其出现的序列等，其中最主要的结构信息是网页之间的超链接；

第三类是使用(Usage)，是网页被人浏览的记录，如IP地址、访问时间等，这些信息可以从Web服务器的日志文件获得。

第四类是用户资料(User Profile)，是某个网站中记录的用户资料。

# Web数据挖掘

## 1. Web数据挖掘定义

定义：

**Web挖掘是对Web文档的内容、Web上可利用资源的使用情况以及资源之间的关系进行分析，从中发现有效的、新颖的、潜在有用的、并且最终可理解的模式。**

# Web数据挖掘

## 2. Web数据挖掘流程

查找资源、信息选择及预处理、模式发现、模式分析。

查找资源 从各种Web数据源中得到数据，数据可以来自于Web文档、电子邮件、新闻组或Web日志等；

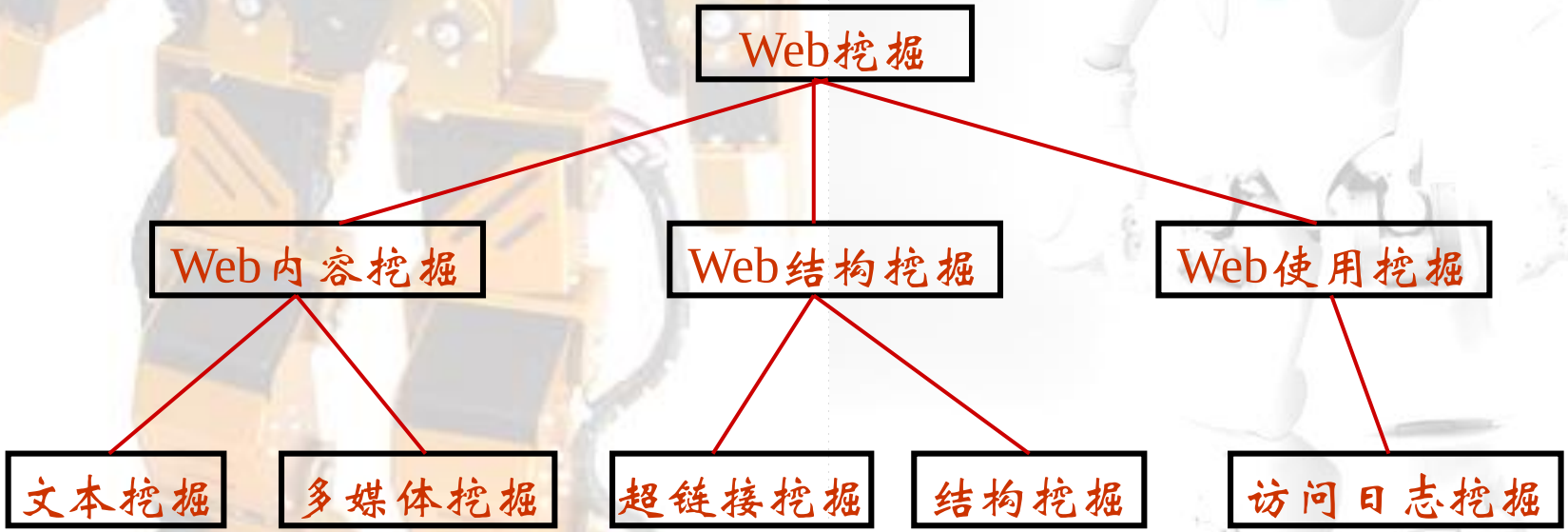
信息选择及预处理 从查找得来的资源中除去无用信息，保留有用信息，并将信息进行必要的整理；

模式发现 在一个站点内部或在多个站点间自动进行模式发现；

模式分析 验证、解释所发现的模式，它可以通过与分析人员进行交互或者由机器自动完成。

# Web数据挖掘

## 3. Web数据挖掘分类



# Web数据挖掘

**Web内容挖掘** 从Web文档内容或其描述中发现有用信息的过程。

常用的Web内容挖掘方法：

（1）改进的WWW的搜索引擎，包括Web Crawler,Lycos；

（2）数据库方法：把半结构化的Web信息重构，使得Web信息更结构化，然后就可以使用标准化的数据库查询机制和数据挖掘方法进行分析；

（3）对页面中的文本进行特征描述，特征描述的模型有很多种，向量空间模型（VSM），布尔逻辑模型，概率模型等等。继而对特征向量进行挖掘，对页面中的多媒体信息进行多媒体信息挖掘，具体方法有页面内容摘要、分类、聚类以及关联规则发现等。

# Web数据挖掘

**Web结构挖掘** 从web结构中发现潜在链接模式的过程。由于文档之间存在着超链接，WWW可以通过这种超链接揭示出文档内容之外的一些有价值的信息。例如指向一个页面的超链接数目就表明了该文档受欢迎的程度，而其包含的超链接数目就表明该文档主题的丰富程度。

结构挖掘的功能是通过分析一个Web页面链接和被链接数量以及链接对象的重要性来建立Web的链接结构模式，并为户提供与请求相关度较大的Web页面，提高搜索引擎的精度和查全率。主要有PageRank和Hub/Authority两种算法。

# Web数据挖掘

**Web使用挖掘** 通过对用户在访问WWW服务器时留下的访问记录进行挖掘，从而获得有关用户的访问模式。服务器日志包括访问日志、引用日志和代理日志。

访问日志记录了用户的标识、访问时间、方法、请求的页面、协议、服务器状态及传输字节数等；

引用日志记录的是被请求页面的存放位置；

代理日志记录了用户使用的浏览器和操作系统的类型。根据三者的内在关系，可以将它们拼接成完整的日志纪录并以关系表形式保存在数据库中。

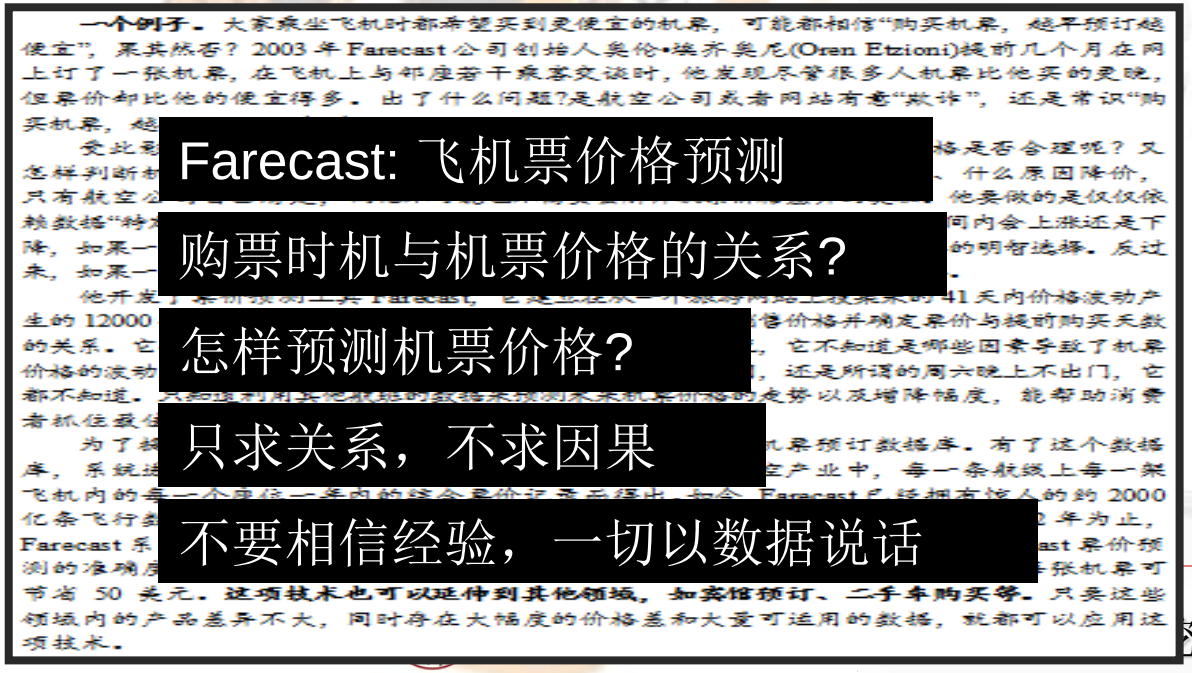
# Web数据挖掘

这些信息中隐含着用户对特定内容的需要。Web使用记录挖掘是通过处理服务器日志文件，以发现用户的浏览模式，如序列模式、关联规则、用户聚类和页面聚类等，理解用户的行为，从而实现：（1）寻找用户的兴趣，进行网页预测推荐，为用户提供个性化服务；（2）改进和优化Web站点结构。

# 大数据

## 大数据价值发现

**2008年全球产生的数据量为0.49ZB($2^{50}$MB)**
**2009年的数据量为0.8ZB**
**2010年增长为1.2ZB**
**2011年的数量更是高达1.82ZB**
**2012年为止，人类所有印刷材料的数据量是200PB**
**预计到2020年，全世界的数据规模将达今天的44倍。**



**一个例子。**大家乘坐飞机时都希望买到更便宜的机票，可能都相信“购买机票，越早预订越便宜”，果真然否？2003年Farecast公司创始人奥伦•埃齐奥尼(Oren Etzioni)提前几个月在网上订了一张机票，在飞机上与邻座若干乘客交谈时，他发现尽管很多人机票比他买的更晚，但票价却比他的便宜得多。出了什么问题?是航空公司或者网站有意“欺诈”，还是常识“购买机票，越

受此

怎样判断

只有航空

赖数据“特

降，如果一

来，如果一

他开发

生的 12000

的关系。它

价格的波动

都不知道。

为了

库，系统

飞机内的

化条飞行

Farecast系

测的准确

节省 50 美元。这项技术也可以延伸到其他领域，如宾馆预订、二手车购买等。只要这些领域内的产品差异不大，同时存在大幅度的价格差和大量可运用的数据，就都可以应用这项技术。

格是否合理呢？又

什么原因降价，

他要做的是仅仅依

间内会上涨还是下

的明智选择。反过

。

41天内价格波动产

卷价格并确定票价与提前购买天数

。它不知道是哪些因素导致了机票

，还是所谓的周六晚上不出门，它

的走势以及增降幅度，能帮助消费

机票预订数据库。有了这个数据

空产业中，每一条航线上每一架

的每一个座位一类的经验是你记录无得出，如今，Farecast已经拥有惊人的约2000

年为止，

ast票价预

张机票可

Farecast票价预测的准确度

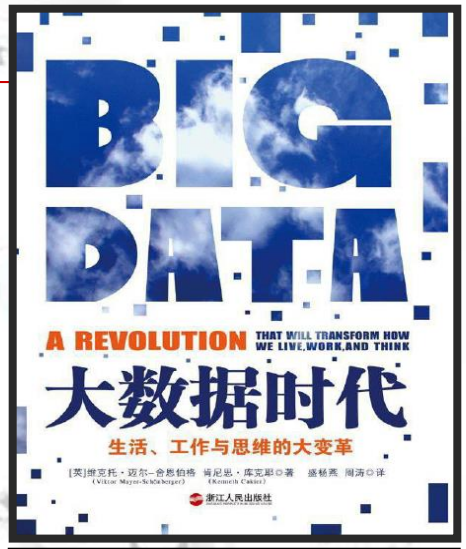
Farecast: 飞机票价格预测

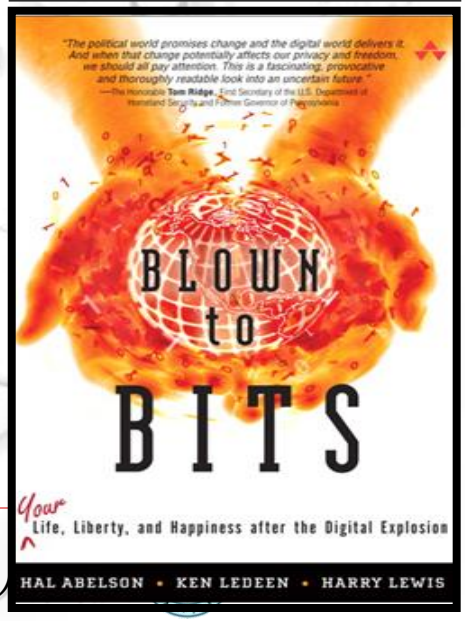
购票时机与机票价格的关系?

怎样预测机票价格?

只求关系，不求因果

不要相信经验，一切以数据说话



# 大数据

## 大数据价值发现

- 华尔街金融家利用电脑程序分析全球3.4亿微博账户的留言，根据民众情绪抛售股票：
- 银行根据求职网站的岗位数量，推断就业率；
- 投资机构搜集并分析上市企业声明，从中寻找破产的蛛丝马迹；
- 美国总统奥巴马的竞选团队依据选民的微博，实时分析选民对总统竞选人的喜好，基于数据对竞选议题的把握，成功赢得总统大选。
- 中国网民发动的“人肉搜索”，已成功地使若干“表哥”“表叔”“房叔”“房妹”等腐败官员落入法网。
- ……

# 大数据

大数据特点— 5 个“V”：

- 数据体量巨大（Volume）：从TB级别，跃升到PB级别；
- 数据类型繁多（Variety）：日志、视频、图片、位置信息等等。
- 处理速度快（Velocity）：1秒定律，实时分析与处理。
- 价值密度低（Value），商业价值高：如连续不间断视频监控过程中，可能有用的数据仅仅有一两秒。
- 真实性（Veracity）：数据质量对于分析和决策相当重要。

# 6.8 知识发现——大数据

HACE理论

3

**Autonomous**

- 自治的分布式数据源
- 去中心化的控制

**Complex**

- 复杂的数据类型和表示
- 复杂的数据间关系

2

4

**Evolving**

- 数据类型的不断丰富
- 数据量的动态增长
- 数据间关系的动态演化

**Heterogeneous**

- 异构的数据类型
- 稀疏的维度

1

# 大数据

**大数据（big data）**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。

**大数据技术：**包括大规模并行处理（MPP）数据库、数据挖掘、分布式文件系统、分布式数据库、云计算平台、互联网和可扩展的存储系统。

**大数据思维：**

第一由样本到全量思维；第二由精确到模糊思维；第三由因果到关联思维。

它是一种基于数据量化和互联，通过数据分析，挖掘，应用，以达到整个世界高度智能化甚至智慧化的思维和方法。