Web信息处理与应用

第九节

个性化检索（下）

徐童

2024.11.7

上节回顾

- **推荐系统典型案例：Netflix Prize**

- BPC队的夺冠秘诀（3）捕捉情境化规律
  - 用户行为存在着情境化的规律，不同情境可能造成不同影响
    - 例如，用户周末情绪较好，打分偏高；而工作日情绪较差，打分偏低
    - 又如，用户不同时间的评分行为对当下呈现不同权重的影响
      - 旧的偏好会随着时间流逝而逐渐衰减

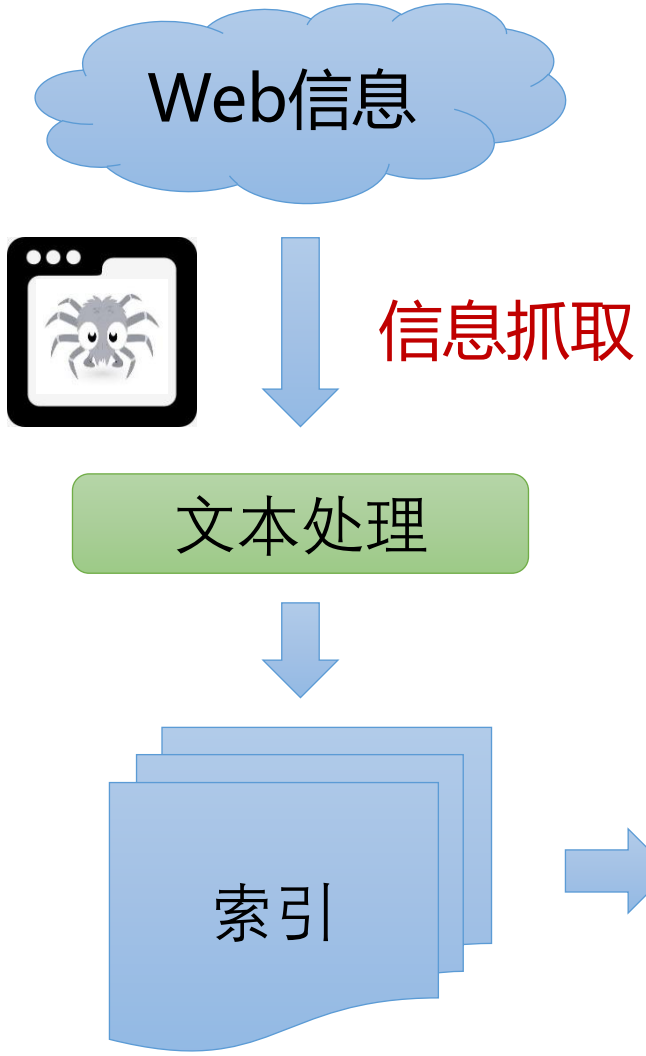
本节内容

- **本节教学内容**

- 在上一节中，我们学习了个性化检索 / 推荐系统的常见模型
- 在本节课中，我们还将进一步学习以下内容：
  - 情境感知的个性化检索与推荐
  - 数据处理中的小技巧
    - 如何采样数据
    - 如何规约数据
    - 如何离散数据

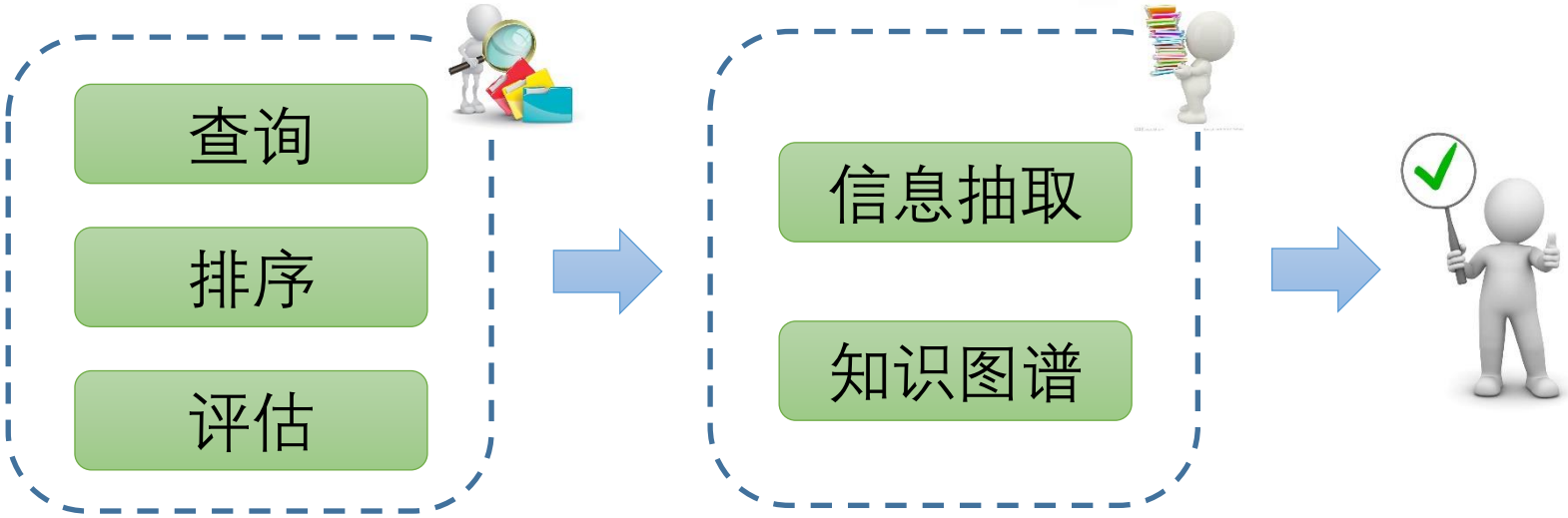
本节主题

- 本课程所要解决的问题



第九个问题：

哪些小技巧能够有效提升个性化检索服务的精度？



本节目录

- **情境感知的查询理解**
- 情境感知的推荐任务
- 数据预处理
  - 数据采样
  - 数据归约
  - 数据离散

- **回顾：用户查询条件的常见问题**

- 在许多时候，用户输入的查询条件缺乏足够的精准性
  - 用户查询条件存在歧义，难以判断真实意图
  - 用户查询条件过于精简、语义信息不够完整
- 在缺少直接来自用户的反馈时，往往需要借助其他信息来协助判断

- **用户查询条件的常见问题**

- 用户查询条件存在歧义，难以判断真实意图

or

- **用户查询条件的常见问题**

- 此时，查询上下文能够帮助我们判断用户的真实意图

+ 搜索历史

=

- **用户查询条件的常见问题**

- 用户查询条件过于精简、语义信息不够完整

or

- **用户查询条件的常见问题**

- 基于查询时的环境信息，可以填补查询条件中的缺失要素

北京 + 非秋季 →

杭州 + 非冬季 →

• 情境的概念与意义

- 在上述两个例子中，我们都是用了其他信息帮助我们理解用户意图。这一类信息被我们称作“**情境信息**” （Context Information）
  - 从计算机学科的视角出发，“情境”一词可定义为“所有与人机交互相关，用于区分标定当前特殊场景的信息”。
  - 基于这一定义，服务提供者可借助情境信息，为用户提供更精确的信息检索和过滤服务。

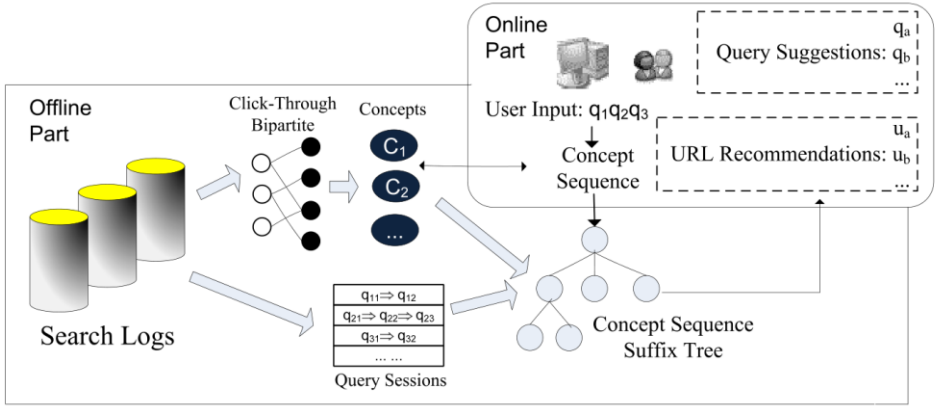
- **搜索中的基础情境：上下文**

- 直观上，查询上下文可以帮助更好的理解用户查询词。
  - 用户的搜索行为往往具有一定的连贯性
  - 相应的，同一查询会话中的查询词和点击的 URL 往往是相关的
    - 一个小问题：如何拆分查询会话?

查询会话 ID	查询会话
S_1	Beautiful mind ⇒ Gladiator ⇒ Russel Crowe ⇒ Russel Crowe movies ↓ www.imdb.com/title/tt0172495
S_2	Roma ⇒ Roma history ⇒ Gladiator ↓ www.exovedate.com/the_real_gladiator_one.html

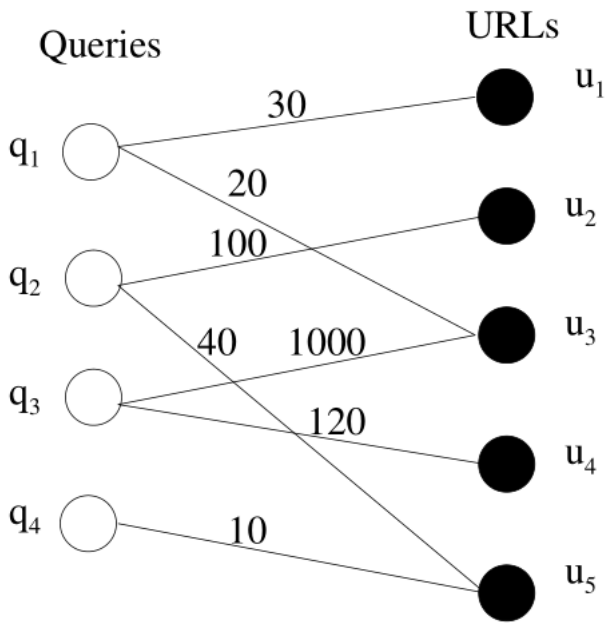
• 基于上下文感知的搜索：基本流程

- 线下训练阶段（模型准备阶段）
  - 首先，将查询词归纳为查询概念，从而避免查询词稀疏性的影响
  - 其次，建立模型，描述查询概念之间的关联关系，支撑上下文感知
- 线上服务阶段（感知查询阶段）
  - 首先，切分会话，判断与当前查询相关的上下文查询与点击记录
  - 其次，根据已有查询记录理解用户当前意图，并进行精准查询



查询概念归纳

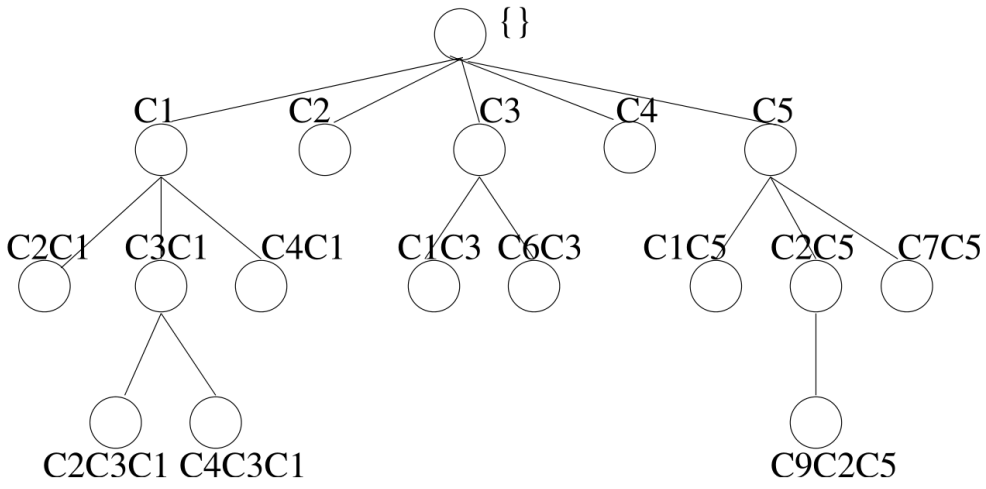
- 用户可能用不同的查询词描述同样的信息需求
  - 例如，中国科学技术大学 和 中科大 代指同一所大学
- **查询概念**：一组有着相同语义的查询词
  - 可以解决查询词的稀疏性问题
  - 同时，更规范地解释查询上下文
- 一种启发式方法：Query-URL的二部图聚类



Cao H, et al., *Context-Aware Query Suggestion by Mining Click-Through and Session Data*, KDD 2008

基础的上下文感知方法

- 借助查询概念归纳，我们将会话内的查询序列转化为了查询概念序列
- 接下来，核心任务在于如何从序列中抽取查询概念的序列模式
  - 考虑特定长度以内的所有序列模式
  - 保留频率高于阈值的模式，并存储为后缀树（Suffix Tree）
  - 当面临上下文感知任务时，根据已有序列找到相应节点，从而获得候选查询建议



Cao H, et al., *Context-Aware Query Suggestion by Mining Click-Through and Session Data*, KDD 2008

• 细化的上下文感知不同类型

- 我们已经证实了上下文有助于更好理解用户的查询意图。然而，上下文信息是否具有不同类型？各种类型所起到的作用有何不同？
- 几种常见的查询上下文类型：
  - 查询重组
  - 查询特化
  - 查询泛化
  - 一般关联

Xiang B, et al., *Context-Aware Ranking in Web Search*, SIGIR 2010

• 细化的上下文感知不同类型

- 类型1：查询重组
  - 用户的后续查询仅仅只是先前查询的重新表述，目的不变或类似
  - 在此情况下，先前点击的内容往往不再被点击（即使内容相关）

Query 1: "homes for rent in atlanta"		Query 2: "houses for rent in atlanta"	
×	**Atlanta homes for rent - home rentals - houses for ren...** Rentlist is directory of Atlanta home rentals featuring links to... *http://www.rentlist.net*		**Atlanta homes for rent - home rentals - houses for ren...** Rentlist is directory of Atlanta home rentals featuring links to... *http://www.rentlist.net*
	Homes For Rent, lease in Atlanta suburbs. Can't sell ... Atlanta homes for rent, homes for lease in Gwinnett and north... *http://atlantahomesforrent.com*		**Homes for Rent in Atlanta, GA** Houses, Apartments and Homes for Rent in Atlanta, GA Find ... *http://www.usrentallistings.com/ga/atlanta*
	Rentals.com - Homes for Rent, Apartments, Houses ... Atlanta Home Rentals; Austin Home Rentals; Charlotte Home... *http://www.rentals.com*		**Atlanta Home Rentals, Homes for Rent in Atlanta ...** Atlanta Rentals - Homes for Rent in Atlanta, Apartments, Re... *http://www.rentals.com/Georgia/Atlanta*
×	**Atlanta Home Rentals, Homes for Rent in Atlanta ...** Atlanta Rentals - Homes for Rent in Atlanta, Apartments, Re... *http://www.rentals.com/Georgia/Atlanta*		**Homes For Rent, lease in Atlanta suburbs. Can't sell ...** Atlanta homes for rent, homes for lease in Gwinnett and north... *http://atlantahomesforrent.com*
	Homes for Rent in Atlanta, GA Houses, Apartments and Homes for Rent in Atlanta, GA Find ... *http://www.usrentallistings.com/ga/atlanta*	×	**Atlanta Homes for Rent, Rental Properties, Houses for ...** Search for Homes for Rent in Atlanta, Georgia for free. View li... *www.rentalhouses.com/find/GA/AtlantaArea/ATLANTA*

• 细化的上下文感知不同类型

- 类型2：查询特化
  - 在用户的后续查询中，对先前查询中部分内容进行了更为具体、深入的查询
  - 在此情况下，先前查询中较为泛指的内容将被略过

Query 1: “time life music”		Query 2: “time life Christian CDs”	
×	**Welcome to TimeLife.com | Homepage** TimeLife.com: The best in music & video from a name you can... *http://www.timelife.com*		**Welcome to TimeLife.com | Homepage** Enjoy 138 romantic classics on 9 CDs from top artists like John... *http://www.timelife.com*
	Time-Life - Wikipedia, the free encyclopedia Time-Life is a creator and direct marketer of books, music, vid... *http://en.wikipedia.org/wiki/Time-Life_Music*		**Time Life Music & Video As Seen On TV** Christian ... Time Life Music & Video CD & DVD Collections ... *http://www.asseenontvmusic.com/timelife.html*
	Welcome to TimeLife.com | Music Shop online for exclusive music CDs, music collections, & musi... *http://www.timelife.com/webapp/wcs/stores/servlet/Categor...*		**Welcome to TimeLife.com | Music** Shop online for exclusive music CDs, music collections, & musi... *http://www.timelife.com/webapp/wcs/stores/servlet/Categor...*
	Contemporary Country (Time-Life Music) - Wiki... Contemporary Country was a 22-volume series issued by Time-... *http://en.wikipedia.org/wiki/Contemporary_Country_(Time-...*	×	**Songs ... Time Life 10 CD Collection... Christian Music** CD/Album review of Songs 4 Ever Time Life 10 CD Collection... *http://www.titletrakk.com/album-cd-reviews/songs-4...*
	Time Life Canada | Homepage The most comprehensive country music collection dedicated to... *http://www.timelife.ca*	×	**Christian Band - Newsong - More Life - CD Review of ...** Christian Band - Newsong - More Life CD Review ... Three yea... *http://christianmusic.about.com/cs/cdreviews/fr/aafpr09080...*

• 细化的上下文感知不同类型

- 类型3：查询泛化
  - 在用户的后续查询中，对先前查询中部分内容进行了更泛化的查询
  - 体现了用户对于该查询更广泛的兴趣，而不是局限在某一特定话题
    - 在本实例中，用户从单纯的游戏网站转为查询游戏介绍和历史（如维基百科）

Query 1: "Free online Tetris"		Query 2: "Tetris game"	
×	**Tetris Friends Online Games - Play Free Games Featuri...** Play free online games featuring Tetris. Play single-player and ... *http://tetrisfriends.com*		**Tetris Friends Online Games - Play Free Games Featuri...** Play free online games featuring Tetris. Play single-player and ... *http://tetrisfriends.com*
×	**Play Free Tetris Game Online** Play this classic, original, Flash Tetris Game online for free. *http://www.gametetris.com*		**Tetris game** Free online game: Make lines with falling blocks! Russia's finest... *http://www.play.vg/games/6-Tetris.html*
	Free Tetris Game Free tetris game - Play free tetris games online, learn about tet... *http://www.tetrislive.com*	×	**Tetris (Game Boy) - Wikipedia, the free encyclopedia** Tetris was a pack-in title included with the Game Boy at the ha... *http://en.wikipedia.org/wiki/Tetris_(handheld_game)*
	4FreeOnlineGame.com - Free Online Tetris Game 4FreeOnlineGame - Free Online Tetris Game ... This is the all ... *http://www.4freeonlinegame.com/Tetris*	×	**Tetris - non-stop puzzle action** Tetris logo, Tetris theme song and Tetriminos are trademarks of... *http://www.tetris.com*
	Tetris - Play Tetris. Free online games @ Adoption Media, LLC 1995 - 2010 This site should not subst... *http://games.adoption.com/free-online-games/Tetris*		**Free Tetris Game** Free tetris game - Play free tetris games online, learn about tetr... *http://www.tetrislive.com*

• 细化的上下文感知不同类型

- 类型4：一般关联
  - 借助先前查询，可以补全用户在当前搜索中的特定意图
  - 更接近之前所笼统叙述的“上下文”情境感知的概念
    - Xbox+fifa ⇨ 足球游戏

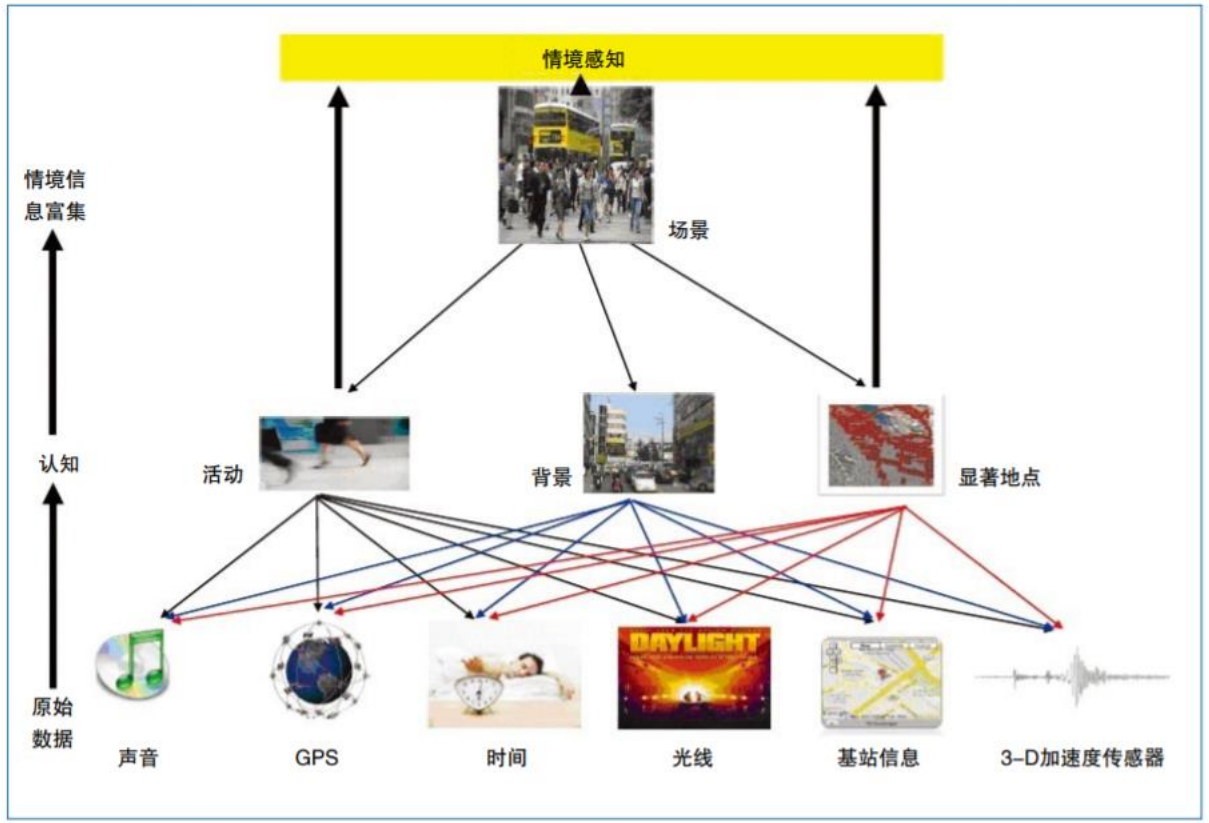
Query 1: “Xbox 360”		Query 2: “FIFA 2010”	
×	**Xbox.com | Home** Xbox.com is your ultimate source for all things Xbox and Xb... *http://www.xbox.com*		**FIFA.com - The Official Website of the FIFA World Cup** The Official Website of the 2010 FIFA World Cup South Africa™ *http://www.fifa.com/worldcup/index.html*
	Xbox 360 - Wikipedia, the free encyclopedia The Xbox 360 is the second video game console produced by ... *http://en.wikipedia.org/wiki/Xbox_360*		**2010 FIFA World Cup - Wikipedia, the free encyclopedia** The template below has been deprecated (see discussion), and ... *http://en.wikipedia.org/wiki/2010_FIFA_World_Cup*
×	**Xbox.com | Xbox 360** Find out more about Xbox 360, the awesome lineup of games ... *http://www.xbox.com/en-US/hardware*		**FIFA.com - Fédération Internationale de Football Associa...** The official site of the international governing body of the sport ... *http://fifa.com*
	Microsoft Xbox Xbox 360 delivers the most powerful console, the next genera... *http://www.microsoft.com/xbox*	×	**FIFA 10 Soccer : FIFA 2010 - EA Sports Games** Improvement in Management Mode, Flick Passes, Ball Physics, ... *http://www.ea.com/games/fifa-soccer*
	Xbox 360 - Gizmodo This No-Name HTPC Remote Has a Keyboard, Can Work W... *http://gizmodo.com/tag/xbox-360*		**FIFA 2010 World Cup in South Africa** A surprise in the 2007 Asian Cup! The Iraqis win it! In spite of ... *http://southafrica2010.wordpress.com*

本节目录

- 情境感知的查询理解
- **情境感知的推荐任务**
- 数据预处理
  - 数据采样
  - 数据归约
  - 数据离散

• 更复杂的情境信息

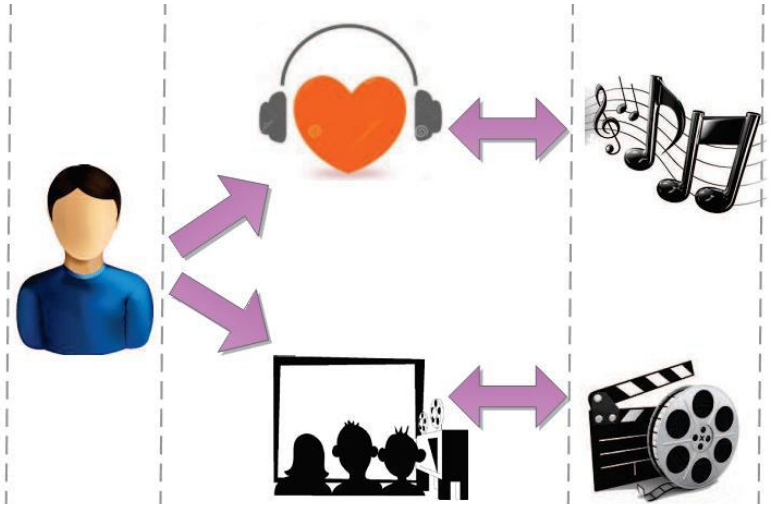
- 随着智能终端与传感器技术的发展，我们所能够获得的情境信息更为丰富。
  - 对于用户所处状态描述更为完整、准确
  - 用户的信息需求类型也更为多样、复杂



陈恩红等，*移动情境感知的个性化推荐技术*，中国计算机学会通讯，2013(3):18-24

- **推荐系统中的情境有哪些?**

- 情境是那些在重复进行同一活动时可能发生变化的因素
  - 看电影：时间、地点、同伴等
  - 听音乐：时间、地点、情绪、场合等
  - 派对或餐厅：时间、地点、场合等
  - 旅行：时间、地点、天气、交通状况等



- 根据应用场景，收集和整理相关的信息，可以更加精确地捕捉用户需求

• 情境感知的用户需求

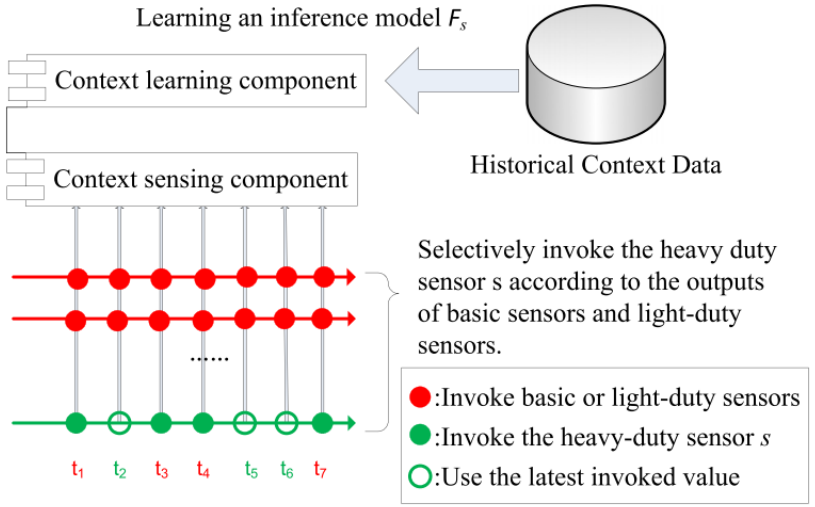
- 描述用户所处状态及意图的，也不再是单一的上下文（查询或查询概念），而是一组复杂且相互关联的情境信息。
- 单一解读某一种情境要素，可能无法得到用户当前状态的准确描述

Timestamp	Context record
t_1	{(Holiday?:No),(Time range: AM8:00-9:00),(Speed: High),(Audio level: Low)}
t_2	{(Holiday?:No),(Time range: AM8:00-9:00),(Speed: High),(Audio level: Middle)}
t_3	{(Holiday?:No),(Time range: AM8:00-9:00),(Speed: High),(Audio level: Middle) ,(Interaction: Music)}
	……
t_{58}	{(Holiday?:Yes),(Time range: AM10:00-11:00),(Movement: Move)(Location: Shop),(Audio level: Middle)}
t_{59}	{(Holiday?:Yes),(Time range: AM10:00-11:00),(Movement: Move)(Location: Shop),(Audio level: Middle)}
t_{60}	{(Holiday?:Yes),(Time range: AM10:00-11:00),(Movement: Move)(Location: Shop),(Audio level: Middle)}
	……

Zhu H, et al., *Mining Mobile User Preferences for Personalized Context-Aware Recommendation*, ACM TIST, 5(4): 58, 2014

• 情境感知的用户需求

- 我们通过一个有趣的案例，来看看个性化服务中的情境感知能做些什么
  - 通过情境识别模式化路径，减少传感器（如GPS）收集次数并降低能耗
  - 发散思维：这一思想是否能够用在缓存机制的使用上？



Li X, et al., *Learning to Infer the Status of Heavy-duty Sensors for Energy-efficient Context-sensing*, ACM TIST, 3(2): 35, 2012

- **如何获取用户所处的情境?**

- 显式获取
  - 直接询问用户 或 预先定义情境，并让用户从中进行选择
- 隐式获取
  - 通过应用的使用数据获取用户当前的位置等情境信息
- 推断获取
  - 从用户的近期行为（序列）中进行推断和归纳

- **情境筛选**

- 显然，并非所有的情境信息都会影响到用户当下的决策
  - 直接询问用户哪些情境信息对于当下是重要的
  - 通过特征选择
    - 例如，主成分分析（PCA）和线性判别分析（LDA）
  - 通过统计分析
    - 统计测试，如信息增益、互信息、Freeman-Halton Test检验等

- **如何将情境信息整合到推荐系统中?**

- 传统推荐：Users × Items ⟶ Ratings

- 情境感知的推荐：Users × Items × Contexts ⟶ Ratings

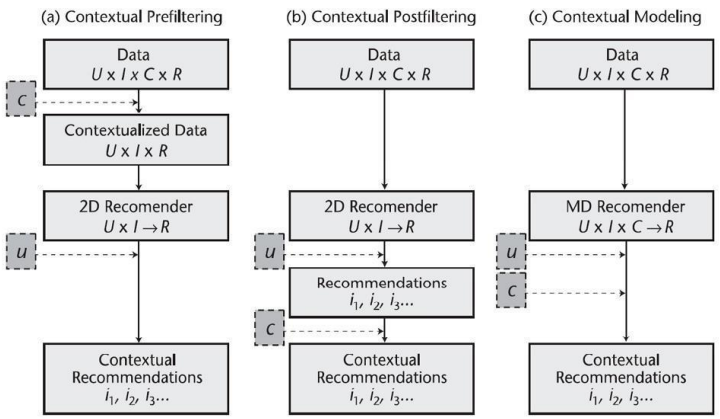
User	Item	Rating	Time	Location	Companion
U1	T1	3	Weekend	Home	Kids
U1	T2	5	Weekday	Home	Partner
U2	T2	2	Weekend	Cinema	Partner
U2	T3	3	Weekday	Cinema	Family
U1	T3	1	Weekend	Cinema	Kids

- 本案例中的新增情境维度：时间，位置，同伴

情境感知

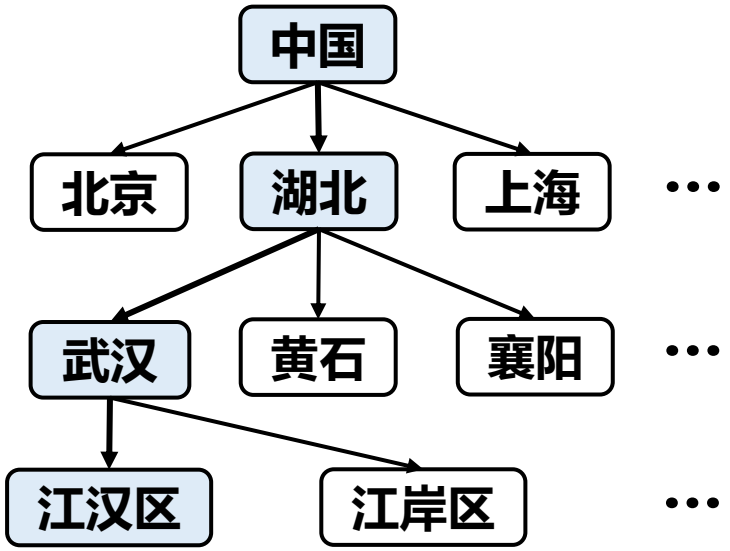
- **如何将情境信息整合到推荐系统中?**

- 情境感知推荐的三种范式
  - 情境预过滤：事先根据情境信息筛选、分隔数据记录
  - 情境后过滤：产生推荐结果后，根据情境信息进行筛选
  - **情境建模**：将情境信息纳入到推荐系统的模型中，在推荐的各个阶段更好地理解和利用情境信息



- **情境预过滤**

- 代表性方法：LARS (Levandoski JJ, et al. ICDE 2012)
  - 情境信息：用户位置（所在国家，城市，区域等）
  - 具体做法
    - 根据用户位置将评分数据组织成树状结构
    - 生成从根节点到叶子节点的路径
    - 使用 ItemCF 为路径上的每个节点训练一个模型，最终的推荐结果是各节点的加权求和



- **情境后过滤**

- 代表性方法：LARS (Levandoski JJ, et al. ICDE 2012)
  - 情境信息：餐馆、商店、旅游景点等的位置
  - 具体做法：LARS 先忽略物品的位置信息，利用 ItemCF 算法计算用户对物品的兴趣 P(u, i)，最后加上一个距离的惩罚项 TravelPenalty(u, i)

$$\mathbf{RecScore}(u,i) = P(u,i) - \mathbf{TravelPenalty}(u,i)$$

其中，TravelPenalty(u, i) 是用户和物品之间的旅行距离，被归一化为与评分标准相同的数值范围（例如，[0，5]）

- **情境建模**

- 大名鼎鼎：Factorization Machines (Steffen Rendle, et al. ICDM 2010)
  - 预备知识：one-hot/multi-hot 编码
  - 性别特征：{“男”，“女”}（两种类别）
    - 男 => 1 0，女 => 0 1

性别	商品类别	是否点击
女	运动配件	0
男	化妆品	0
女	化妆品	1
男	运动配件	1

男	女	运动配件	化妆品	是否点击
0	1	1	0	0
1	0	0	1	0
0	1	0	1	1
1	0	1	0	1

- **情境建模**

- 大名鼎鼎：Factorization Machines (Steffen Rendle, et al. ICDM 2010)
  - 为什么不使用逻辑回归?
    - **特征组合的重要性**

性别	商品类别	是否点击
女	运动配件	0
男	化妆品	0
女	化妆品	1
男	运动配件	1

无特征组合

男	女	运动配件	化妆品	点击率
1	0	1	0	0.4

有特征组合

男	女	运动配件	化妆品	男&运动配件	女&化妆品	点击率
1	0	1	0	1	0	0.8

- **情境建模**

- 大名鼎鼎：Factorization Machines (Steffen Rendle, et al. ICDM 2010)
  - 二阶特征交叉
    - 朴素版本：$$y = \omega_0 + \sum_{i=1}^{n} \omega_i x_i + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \omega_{ij} x_i x_j$$

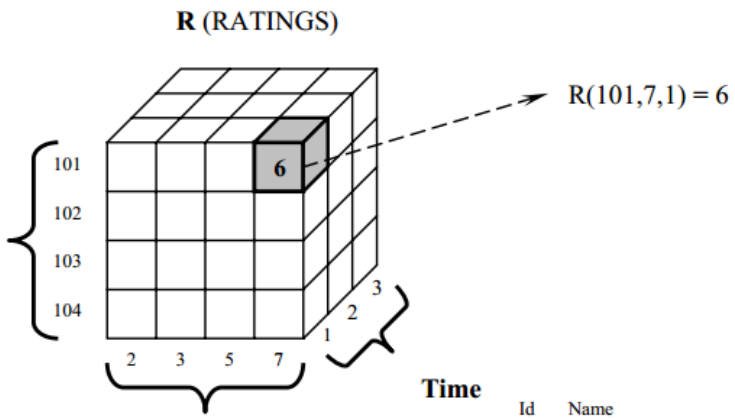
存在$\frac{n(n-1)}{2}$个二阶特征

矩阵分解

    - FM：$$y = \underbrace{\omega_0 + \sum_{i=1}^{n} \omega_i x_i}_{\text{线性回归}} + \underbrace{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} < v_i, v_j > x_i x_j}_{\text{二阶特征组合}}$$
    - 其中，$< \mathbf{v}_i, \mathbf{v}_j >= \sum_{f=1}^{k} v_{i,f} \cdot v_{j,f}$
    - 思考：为什么要将 W 矩阵分解为特征的隐向量？

• 情境建模

- 另一种有趣的方案：Tensor Factorization (Volodymyr Kuleshov , et al. PMLR 2015)
  - 多维空间：Users × Items × Contexts ⟶ Ratings



User

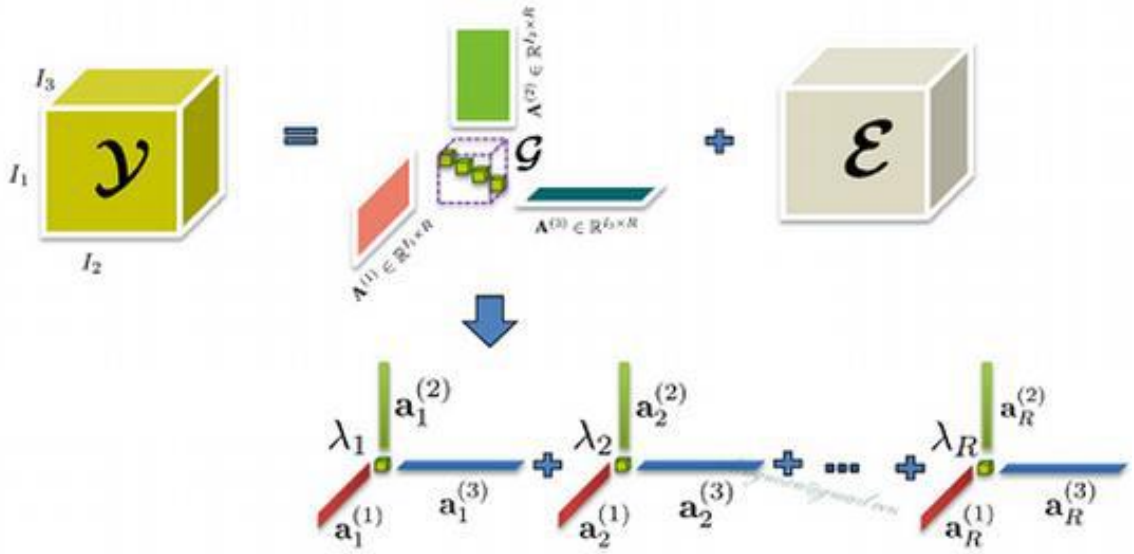
Id	Name	Age
101	John	25
102	Bob	18
103	Alice	27
104	Mary	21

Item

Id	Name	Cost
2	AB17	250.00
3	AB23	299.95
5	XY70	150.00
7	ZZ55	115.50

Time

Id	Name
1	Weekday
2	Weekend
3	Holiday



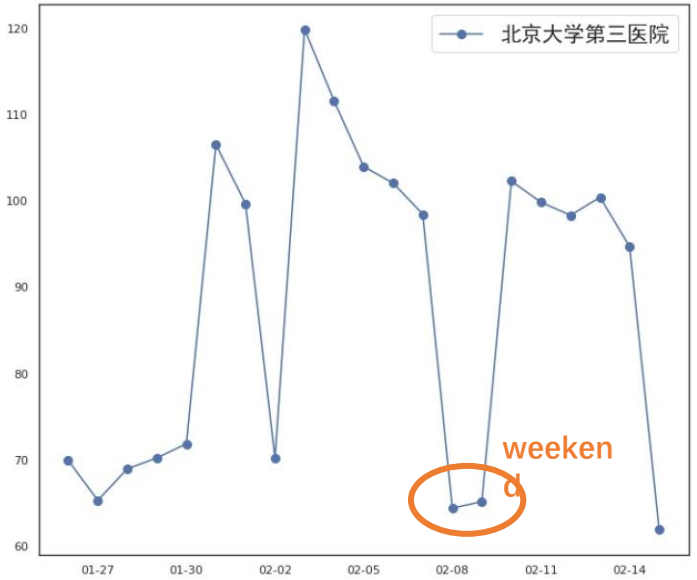
- $R_1 = R_2 = R_3 = R$.
- $\mathcal{G}$ is super diagonal.

$$\mathcal{Y} = \sum_r \lambda_r \mathbf{a}_r^{(1)} \circ \mathbf{a}_r^{(2)} \circ \mathbf{a}_r^{(3)} + \mathcal{E}$$

• 情境建模の应用案例

- 2020 CCF 大数据与计算智能“科技战疫”挑战赛冠军方案
  - 如何准确预测某个区域在某个时段的人流密度?
  - 观察：人流密度呈现鲜明的周期性（工作日/周末）
  - 解决方案：引入周级别的周期因子

$$\alpha_{i,d} = \frac{7 \cdot \sum_{h \in \{0,..,23\}} flow_{i,d-7,h}}{\sum_{j \in E_1} \sum_{h \in \{0,..,23\}} flow_{i,j,h}}$$



完整方案分享可参见公众号文章：https://mp.weixin.qq.com/s/I--keYwFgBuo74_9GPVnTA

- **实践资料**

- 公开数据集：https://github.com/irecsys/CARSKit/tree/master/context-aware_data_sets
  - 电影：AdomMovie, DePaulMovie, LDOS-CoMoDaData
  - 音乐：InCarMusic
- 工具包：CARSKit： https://github.com/irecsys/CARSKit
  - 使用指南：http://arxiv.org/abs/1511.03780

本节目录

- 情境感知的查询理解
- 情境感知的推荐任务
- **数据预处理**
  - 数据采样
  - 数据归约
  - 数据离散

- **为什么要进行数据处理?**

- 糟糕的数据质量将给数据分析过程造成严重的负面影响

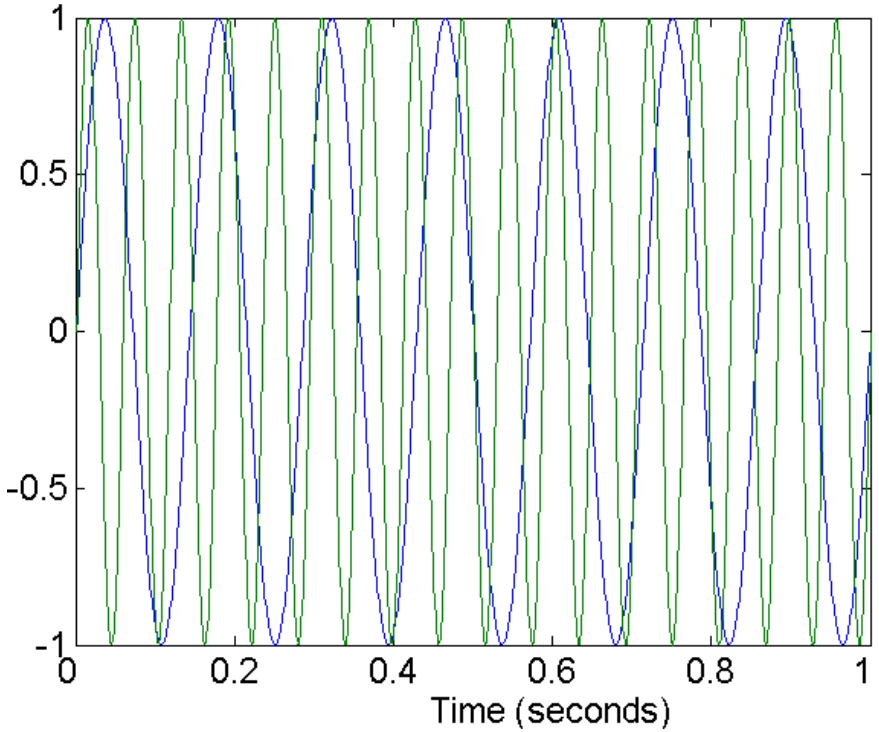
“The most important point is that poor data quality is an unfolding disaster. Poor data quality costs the typical company at least ten percent (10%) of revenue; twenty percent (20%) is probably a better estimate.”

—— Thomas C. Redman, DM Review, August 2004

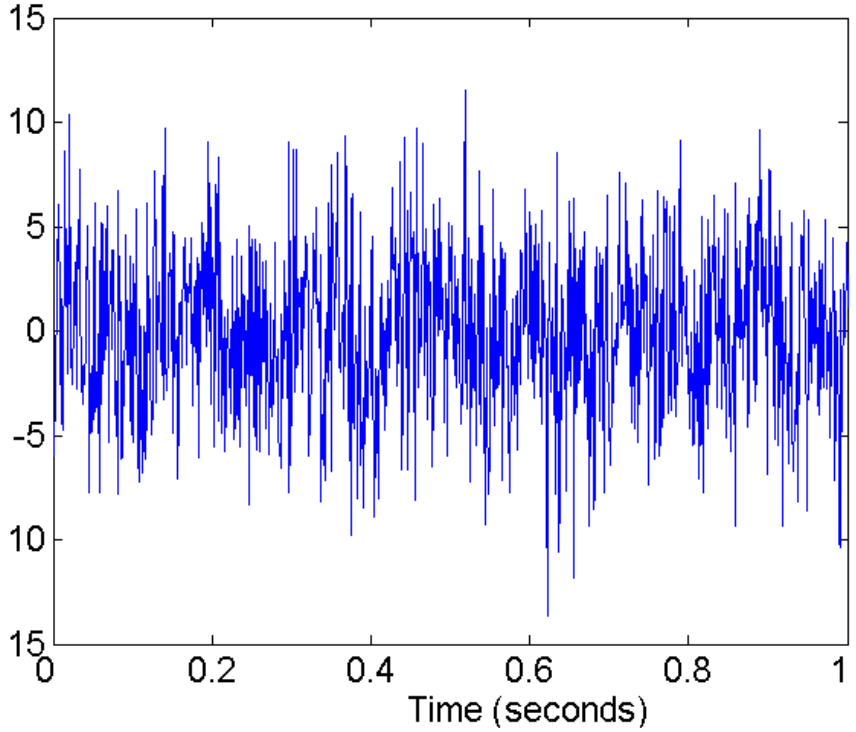
- 常见的数据质量问题
  - 数据测量、采集等过程中出现的错误
  - 噪声、离群点、缺失值等数据问题
  - 重复数据

• 常见数据问题（1）：噪声数据

- 噪声数据往往表现为原始数据的微调或者背景信号的堆叠
  - 例如，嘈杂环境下的通话声音，各种震动干扰下的地震波形



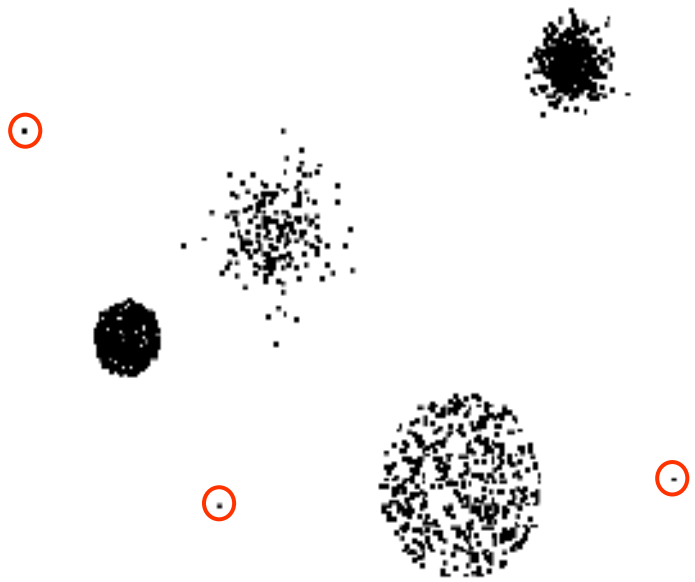
Two Sine Waves



Two Sine Waves + Noise

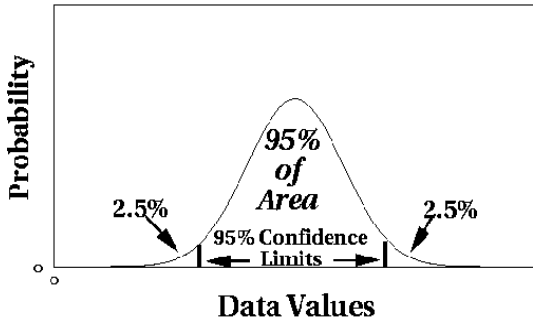
- **常见数据问题（2）：离群数据**

- 离群数据 / 异常点即与大部分其他对象不同的数据
  - 异常点既可能是我们研究的目标，也可能对我们的研究造成干扰



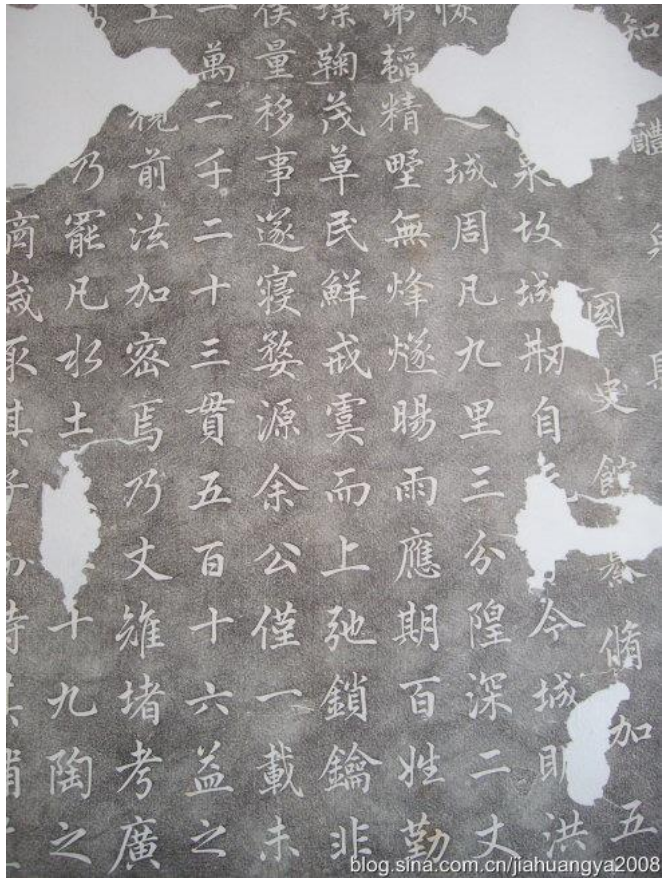
常见数据问题（2）：离群数据

- 离群数据 / 异常点即与大部分其他对象不同的数据
  - 一点题外话：离群检测技术专门面向异常点进行数据分析与挖掘
    - 异常数据不等于错误数据！而是包含着不同于寻常规律的数据
    - 例如，银行面向异常交易的风控、智慧医疗面向罕见病的诊断等场景
- 统计方法是离群检测的最基本方法，例如：基于正态分布的离群点判定
  - 在已知参数的前提下，可根据正态分布判定离群的概率
  - 一个有趣的应用：舆情监测
    - 严重不平衡条件下，通过学习正常值分布进行判定



- **常见数据问题（3）：缺失数据**

- 一方面，数据采集的不完整可能导致数据缺失
  - 例如，室内无法利用GPS采集位置坐标
- 另一方面，部分属性值在部分数据中不适用
  - 例如，儿童往往无法采集其收入信息
- 对待缺失值，往往采取删除和填补并重的方法
  - 一方面，基于已有数据填补缺失值
  - 另一方面，抛弃无法补全的数据记录



- **常见数据问题（4）：重复数据**

- 数据集中可能存在重复或近似重复的数据，这往往是由于多源数据归并所导致的
  - 例如，实体中的歧义现象，利君沙（琥乙红霉素片）
  - 又如，一个社交网络中用户的多个马甲账号
- 往往基于数据整合的方法加以解决
  - 例如，识别两个实体是否是含义相同，并合并相应的数据

- **常见预处理方法**

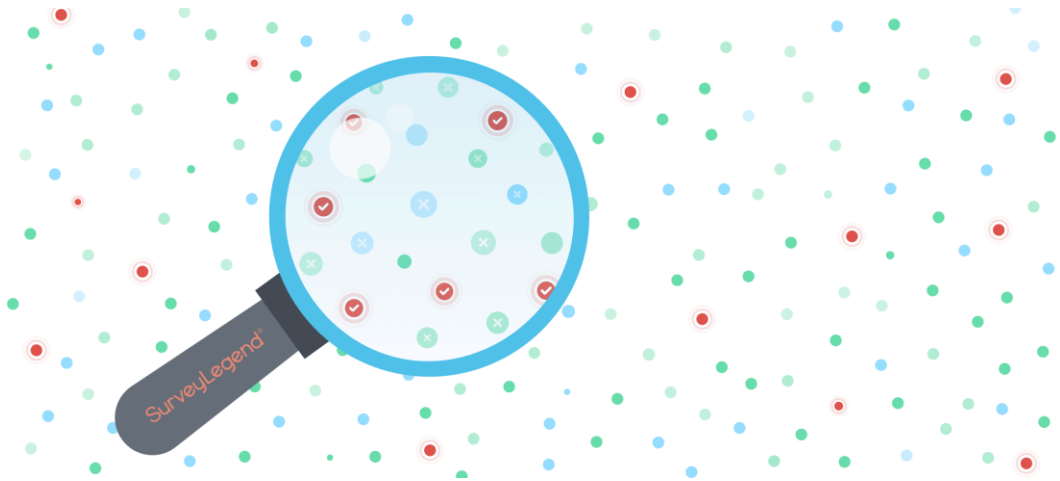
- 针对上述问题，需要通过数据预处理的方式提升数据质量
  - 数据采样（Sampling）
  - 维度归约（Dimensionality Reduction）
  - 数据离散（Discretization & Binarization）

本节目录

- 情境感知的查询理解
- 情境感知的推荐任务
- **数据预处理**
  - **数据采样**
  - 数据归约
  - 数据离散

- **数据采样的概念和目的**

- 数据采样，指选择一部分数据对象的子集进行分析的常用方法
  - 数据规模的急剧增加带来了计算能力的巨大负担
  - 通过采取小规模样本，可以起到近似的效果，同时降低开支
  - 即使在要求精确的场合，通过采取小规模样本进行初步分析，了解数据特性，也是有效的手段



- **数据采样的代表性问题**

- 采样缺乏代表性，将影响对于原数据集的还原程度，进而产生误导
  - 采样数据应至少在统计指标上近似原数据集，例如均值和方差

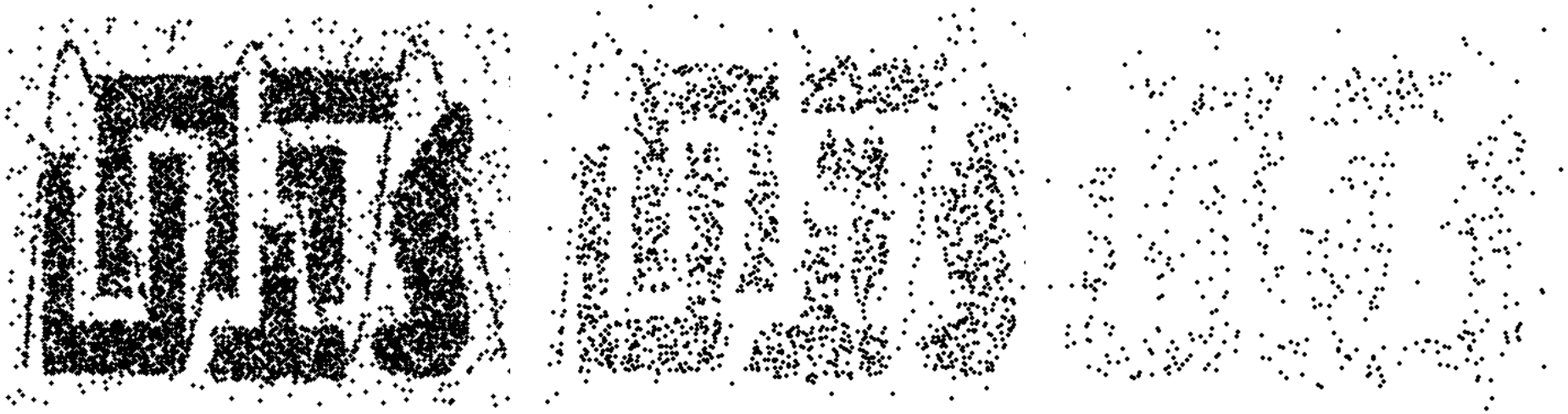
How the Media can manipulate our viewpoint

- **常用的数据采样方法**

- 最基本的采样技术为简单随机采用（Simple Random Sampling）
  - 对于所有对象，采用简单的等概率方式进行采样
  - 一般采用两种方式进行
    - 无放回采样：被采中的对象从整体中删除，仅可选中一次
    - 有放回采样：被选中的对象不会从整体中删除，可再次被选中
  - 两种方式无本质区别，但有放回采样较为简单（概率不变）
  - 更为复杂的方法包括分层采样（Stratified Sampling）
    - 对数据进行分组，从预先指定的组里进行采样

采样规模与信息损失

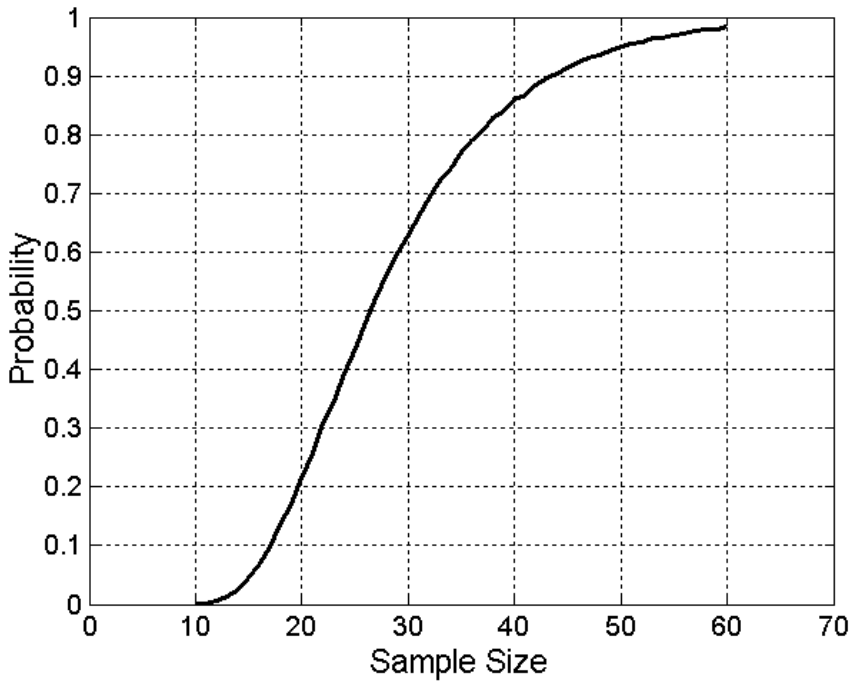
- 影响采样效果的重要因素之一是采样的样本容量
  - 较大的样本容量更能完整代表数据，但降低了采样的收益
  - 较小的样本容量在采样收益上更高，但可能造成信息的损失



8000 points　　2000 Points　　500 Points

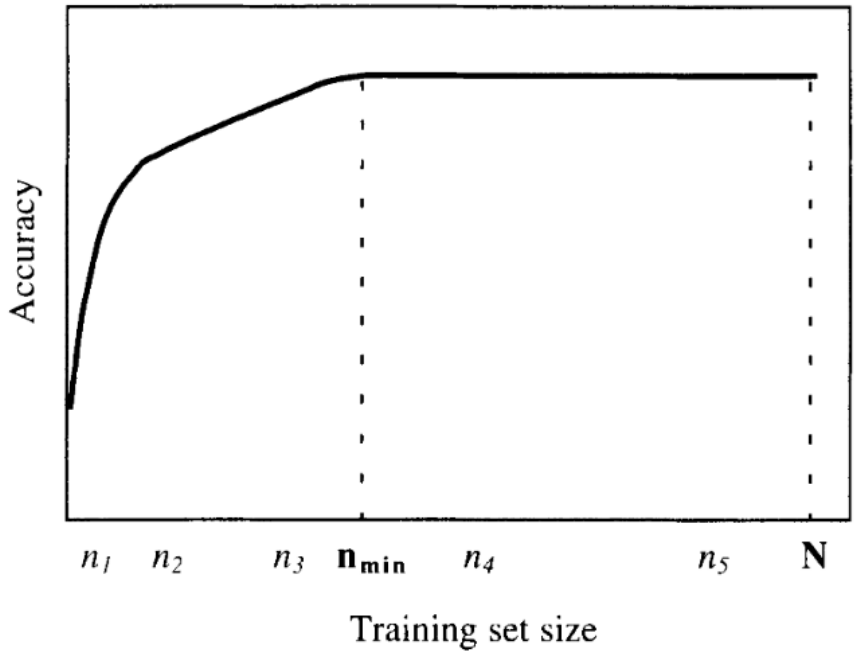
- **启发式的采样规模确定方法**

- 通过分组采样的方式，可以近似确定适当的样本容量
  - 例如，将一个数据集分为10组，每组数量大致相等
  - 组内的数据高度相似，而不同组的对象差异性较大
  - 右图展示了不同样本容量下，每组至少取到一个数据点的概率（体现代表性）
    - 可见，至少40个点，才能保证10组都取到的概率接近90%



- **渐进式采样**

- 渐进式采样从训练结果的角度确定采样规模：
  - 从一个小样本开始，然后逐步增大采样规模
  - 在模型的准确率趋于稳定的时候停止采样，从而确定采样规模
- 优点：不需要在一开始就确定采样的容量
- 缺点：计算开销大（需要多次迭代训练模型）



- **一些题外话**

- 数据采样不能成为数据造假的帮凶！
  - 一种不良的倾向：将数据采样等同于数据“筛选”
  - 采样不应该具有任何的倾向性
    - 所有算法的数据标准应该一视同仁
    - 不能以结果作为采样依据

本节目录

- 情境感知的查询理解
- 情境感知的推荐任务
- **数据预处理**
  - 数据采样
  - **数据归约**
  - 数据离散

• 维度归约的必要性

- 在数据集中，描述一个对象可能涉及到许多特征
  - 然而，并不是所有的特征都具有显著的区分作用

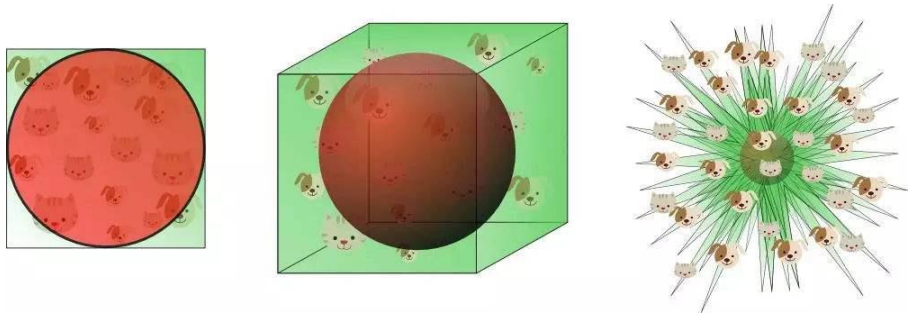
表 1

学生编号	语文	数学	物理	化学
1	90	140	99	100
2	90	97	88	92
3	90	110	79	83
...	...	...	...	...

  - 通过维度归约，可以删除不具有区分度的特征，同时可能降低噪声
  - 在避免维度灾难的同时，模型更容易理解，也更易于可视化

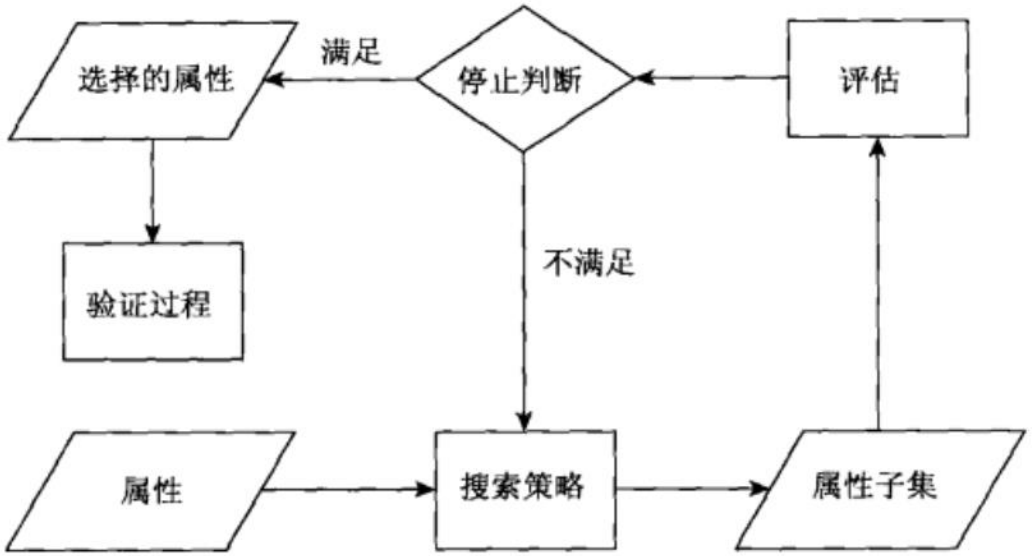
- **维度灾难的概念**

- 维度灾难，又称维度诅咒（Curse of Dimensionality）
  - 指随着数据维度的增加，数据分析困难程度大幅上升的现象
  - 可能由于以下两点原因导致
    - 计算复杂：计算量呈指数级增长，需要更多的计算资源来处理
    - 数据稀疏：随着维度增加，数据点在高维空间中的分布变得稀疏，大部分数据之间的距离都变得非常远，难以区分或识别模式



- **维度归约的基础思路：特征子集选择**

- 降低维度的基础思路是仅使用特征的一个子集（而不是归纳新特征）
  - 其目的在于去除冗余特征（重复信息，例如商品价格与商品税）和不相关特征（例如学生成绩与学生学号往往无关）
  - 除了直接删去多余特征外，为特征赋予不同权值也是一个可选择的方案



- **如何进行维度归约?**

- 在先前的例子中，缺乏区分度的维度可以直观地发现，因此人工删除即可
- 然而，更多时候，需要归约的维度难以通过简单的人工判断加以区分

表 2

学生编号	数学	物理	化学	语文	历史	英语
1	65	61	72	84	81	79
2	77	77	76	64	70	55
3	67	63	49	65	67	57
4	80	69	75	74	74	63
5	74	70	80	84	82	74
6	78	84	75	62	72	64
7	66	71	67	52	65	57
8	77	71	57	72	86	71
9	83	100	79	41	67	50
...	...	...	...	...	...	...

- **维度归约的代表性方法：主成分分析**

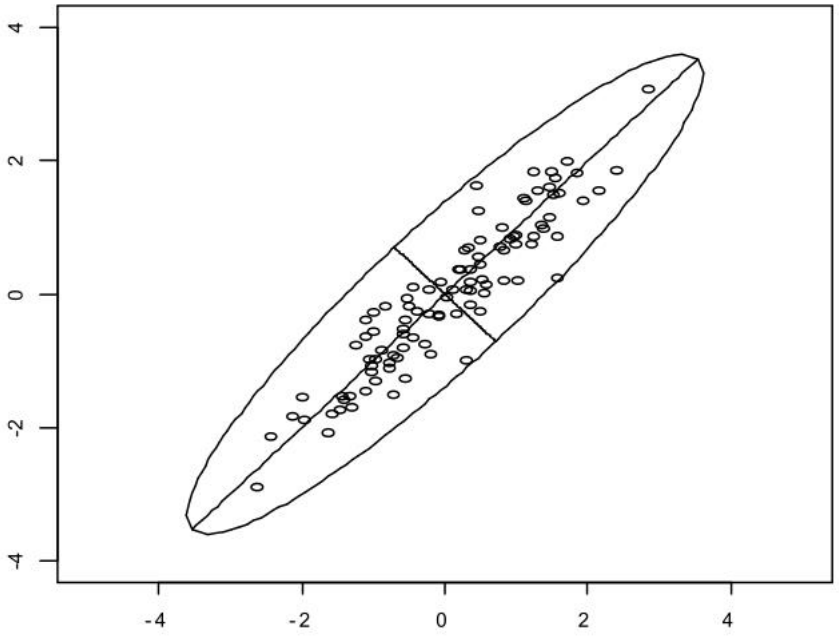
- 对于成绩而言，优秀的学生尚可用总分来加以衡量
- 然而，更多时候，属性更复杂、维度更多，且无法简单加和
  - 假定你是一个公司的财务经理，掌握了公司的所有数据，如固定资产、流动资金、借贷的数额和期限、工资支出、原料消耗、产值等信息。
  - 当你需要向上级汇报公司状况的时候，你可以原封不动地罗列所有指标和数字吗？当然不能。
  - 你必须对各个方面进行高度概括，进而用少数指标简单明了地介绍。

- **维度归约的代表性方法：主成分分析**

- 针对这一需求，我们希望能够从纷乱的属性中找到一些具有代表性、综合性的指标，从而包含丰富的信息量，帮助我们抓住主要矛盾
- 主成分分析（Principal Component Analysis，PCA）应运而生
- PCA的思路是通过正交变换将一组可能存在相关性的变量转换为一组线性不相关的变量，转换后的这组变量叫主成分。
  - 通过这种方式，可以采用较少的综合指标综合先前存在于各个属性（可能相关）中的各类信息，而综合指标之间彼此不相关。

- **维度归约的代表性方法：主成分分析**

- 获取主成分的过程，同时也是降维的过程。**如何进行?**
- 先看一个二维的实例，如果只有两维特征，我们可以将其表示为一个椭圆
  - 椭圆有长轴短轴，相比之下，短轴方向的数据变化较少，区分度偏低
  - 在这种情况下，我们删去短轴，再将坐标轴变换与长轴平行
- 对于高维椭球而言，思路是类似的，即找出主轴（方差最大的轴）与几个最长的轴作为新维度



- **维度归约的代表性方法：主成分分析**

- 现在的问题在于，如何选定轴（坐标系）？
- 我们意识到，选择轴的标准是轴上的投影方差尽可能大**（最大可分性）**
- 而一个有趣的现象是：最大特征值对应的特征向量可以最大化投影方差
  - 为什么呢？会作为作业中的问答题哦，否则估计你们也不会去查资料的。
- 因此，我们只要求得数据样本的最大的K个特征值，其特征向量所对应的线性组合就可以形成K个新的综合指标
  - K个特征值的比重反应了主成分的信息量，一般应大于0.85

参考资料：https://blog.csdn.net/zhongkelee/article/details/44064401

- **维度归约的代表性方法：主成分分析**

- 主成分分析的一个实例，以先前的考试成绩为例
  - 可见，头两个成分特征值累积占了总方差的81.142%，而后面的特征值的贡献越来越少

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.735	62.254	62.254	3.735	62.254	62.254
2	1.133	18.887	81.142	1.133	18.887	81.142
3	.457	7.619	88.761			
4	.323	5.376	94.137			
5	.199	3.320	97.457			
6	.153	2.543	100.000			

Extraction Method: Principal Component Analysis.

- **维度归约的代表性方法：主成分分析**

- 主成分分析计算实例
  - 数据介绍

$$X_{raw} = \begin{bmatrix} 1 & 99 \\ 1 & 101 \\ 2 & 101 \\ 2 & 102 \\ 4 & 102 \end{bmatrix}$$

Samples	Feature1	Feature2
1	1	99
2	1	101
3	2	101
4	2	102
5	4	102

  - 目标：将二维数据降到一维

- **维度归约的代表性方法：主成分分析**

- 主成分分析计算实例
  - 标准化
    - 计算每个特征的均值：
      - Feature1：$\mu_1 = 2$
      - Feature2：$\mu_2 = 101$
    - 原始特征减去均值：$X = \begin{bmatrix} 1-2 & 99-101 \\ 1-2 & 101-101 \\ 2-2 & 101-101 \\ 2-2 & 102-101 \\ 4-2 & 102-101 \end{bmatrix} = \begin{bmatrix} -1 & -2 \\ -1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 2 & 1 \end{bmatrix}$

Samples	Feature1	Feature2
1	1	99
2	1	101
3	2	101
4	2	102
5	4	102

- **维度归约的代表性方法：主成分分析**

- 主成分分析实例
  - 计算协方差矩阵C
    - 协方差： $\mathrm{cov}(x_i, x_j) = E[(x_i - \mu_i)(x_j - \mu_j)]$

标准化中已完成

$$X = \begin{bmatrix} 1-2 & 99-101 \\ 1-2 & 101-101 \\ 2-2 & 101-101 \\ 2-2 & 102-101 \\ 4-2 & 102-101 \end{bmatrix} = \begin{bmatrix} -1 & -2 \\ -1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 2 & 1 \end{bmatrix}$$

    - 协方差矩阵：

$$C = \begin{pmatrix} \mathrm{cov}(x_1, x_1) & \mathrm{cov}(x_1, x_2) \\ \mathrm{cov}(x_2, x_1) & \mathrm{cov}(x_2, x_2) \end{pmatrix} = \frac{1}{N-1} X^T X = \frac{1}{4}\begin{bmatrix} 6 & 4 \\ 4 & 6 \end{bmatrix} = \begin{bmatrix} 1.5 & 1 \\ 1 & 1.5 \end{bmatrix}$$

- **维度归约的代表性方法：主成分分析**

- 主成分分析实例
  - 计算协方差矩阵中的特征值和特征向量
    - 协方差矩阵：$C=\begin{bmatrix}1.5 & 1\\ 1 & 1.5\end{bmatrix}$
    - 计算协方差矩阵特征值：

      令 $|\lambda E-C|=\begin{vmatrix}\lambda-1.5 & -1\\ -1 & \lambda-1.5\end{vmatrix}=(\lambda-1.5)^2-1=0$ ，解得 $\lambda=2.5$或0.5
    - 计算特征向量：

$$C\alpha=\lambda\alpha\Rightarrow\begin{bmatrix}1.5 & 1\\ 1 & 1.5\end{bmatrix}\begin{bmatrix}a\\ b\end{bmatrix}=2.5\begin{bmatrix}a\\ b\end{bmatrix}\Rightarrow\begin{cases}1.5a+b=2.5a\\ a+1.5b=2.5b\end{cases}\Rightarrow a=b$$

    - 单位化之后即单位特征向量：$w_1=\left[\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}}\right]^{\bullet}$

• 维度归约的代表性方法：主成分分析

- 主成分分析实例
  - 对原数据降维
    - 计算投影矩阵：

$$W^* = (w_1) = \left[\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right]^T$$

    - 投影：

$$X' = X \times W^* = \begin{bmatrix} -1 & -2 \\ -1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 2 & 1 \end{bmatrix} \times \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} -\frac{3}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \\ 0 \\ \frac{1}{\sqrt{2}} \\ \frac{3}{\sqrt{2}} \end{bmatrix}$$

输入： 样本集 $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$;
低维空间维数 d'.

过程：

1: 对所有样本进行中心化: $\boldsymbol{x}_i \leftarrow \boldsymbol{x}_i - \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{x}_i$;
2: 计算样本的协方差矩阵 $\mathbf{X}\mathbf{X}^{\mathrm{T}}$;
3: 对协方差矩阵 $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ 做特征值分解;
4: 取最大的 d' 个特征值所对应的特征向量 $\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_{d'}$.

输出： 投影矩阵 $\mathbf{W} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_{d'})$.

PCA 算法

- **维度归约的代表性方法：主成分分析**

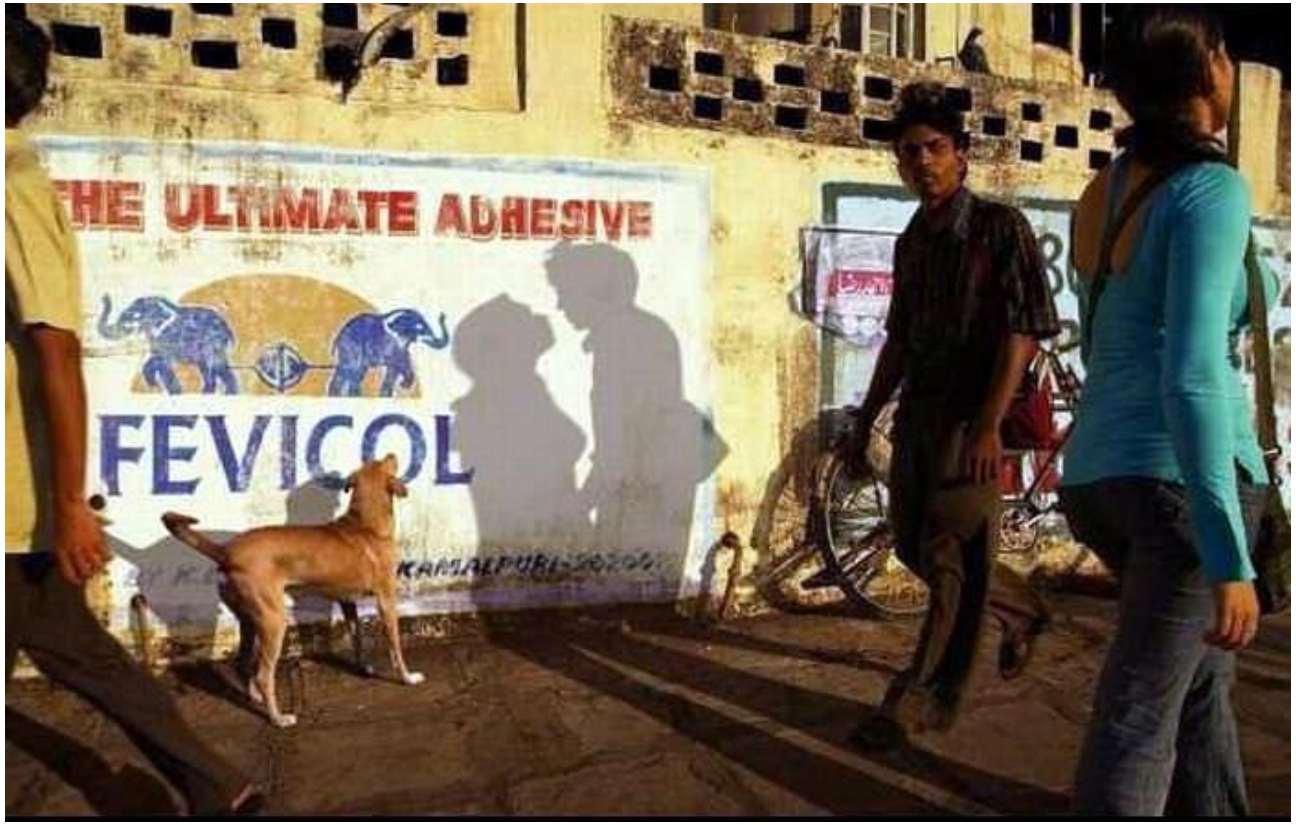
- 主成分分析有哪些需要注意之处?
  - 主成分分析依赖于原始变量，也只能反映原始变量的信息。因此，原始变量的选择很重要（对于原来的问题本身也很重要）
  - 主成分分析的内在假设之一是原始变量直接存在一定的关联性
    - 相应的，如果原始变量本质上相互独立，那么降维就有可能失败
    - 很难将独立变量用少数综合变量概况，数据越相关，降维效果越好
  - PCA的结果未必清晰可解释，与选取的原始变量及数据质量等都有关

- **维度归约中的信息损失**

- 需要注意的是，维度归约同样可能造成信息损失，甚至产生误导效果

- **维度归约中的信息损失**

- 需要注意的是，维度归约同样可能造成信息损失，甚至产生误导效果

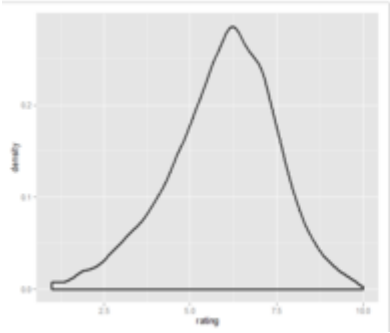


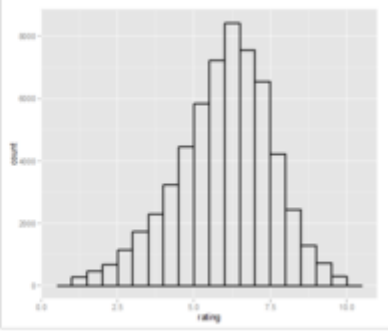
本节目录

- 情境感知的查询理解
- 情境感知的推荐任务
- **数据预处理**
  - 数据采样
  - 数据归约
  - **数据离散**

- **数据离散化的概念**

- 数据离散化（Discretization），旨在将连续属性变换为分类属性
  - 例如，病人的年龄是一个连续值，但是在实际治疗中，往往仅需要一些年龄段的信息即可（如成年/未成年，儿童/成人/老人等）
  - 也可用于分类属性值过多的情况，例如大学的专业较多，可以用文理工农加以概括
- 对于特定数据挖掘问题，特别是分类问题，通过合并减少类别数目是有益的

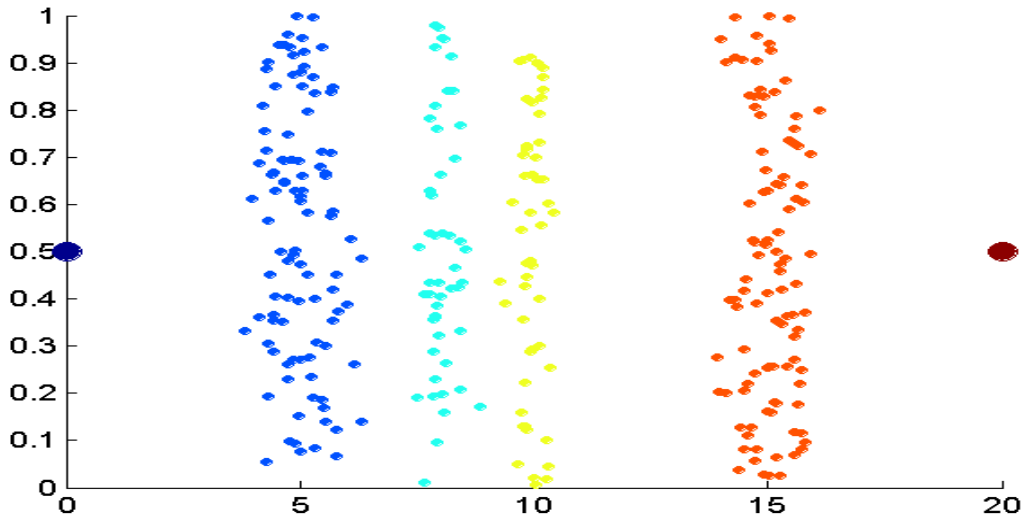




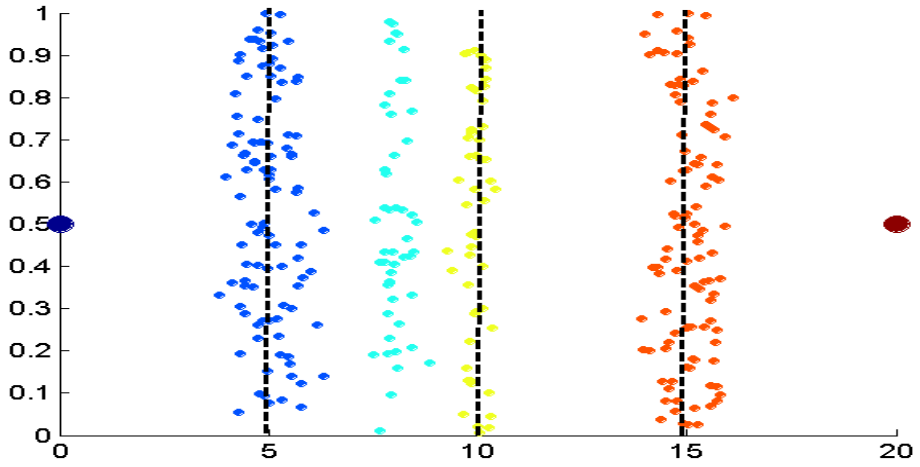
- **非监督离散化**

- 用于分类的离散化方法，最根本的区别在于其离散化过程是否有监督，即是否使用类别信息（Supervision）

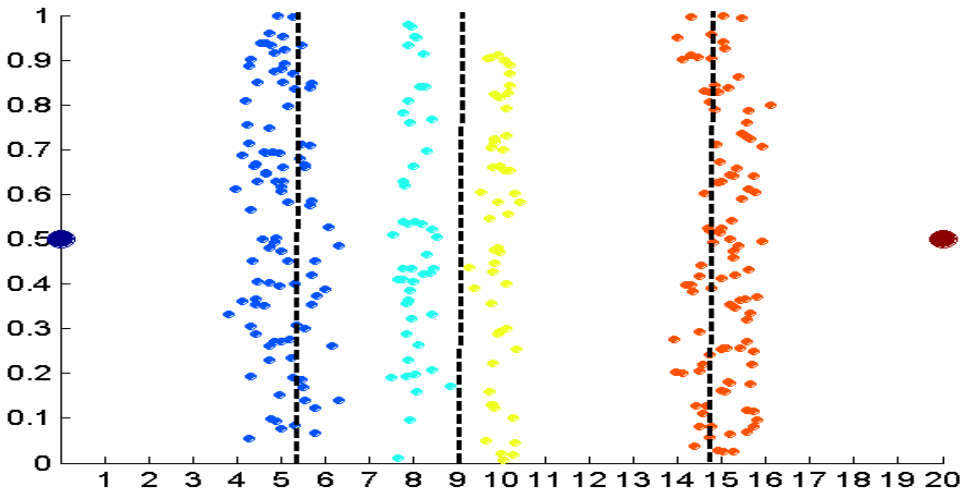
  - 对于不使用类别信息的非监督离散化方法，往往根据数据本身的特性进行离散
  - 常见的方法包括等宽、等频率、k均值、等深离散法等



• 非监督离散化

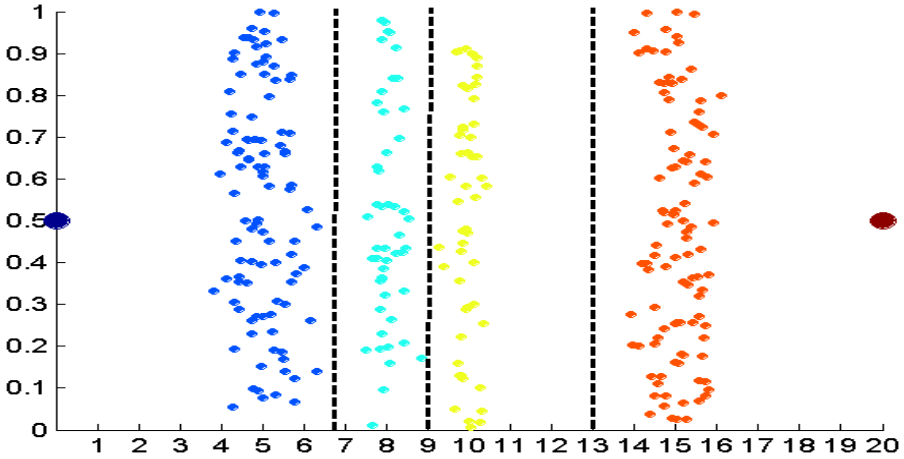


等宽离散化



等频率离散化

K均值离散化



此处仅依赖 x 轴进行离散化
Y 轴主要用于取不同值方便可视化

- **有监督离散化**

- 有监督离散化更注重问题导向，其目的在于取得更好的结果
- 基于熵（Entropy）的方法是最重要的有监督离散化方法之一
  - 采用如下方法定义某个区间的熵

$$e_i = -\sum_{j=1}^{k} p_{ij} log_2 p_{ij}$$

  - 总的熵可以定义为加权和

$$e = \sum_{i=1}^{n} w_i e_i$$

  - 显然，熵越小，区间内的纯度越高（标签越一致），越符合我们的要求
  - 因此，一种做法是先进行二分，选择熵最小的点进行第一轮分割。进而，对其中具有较大熵（即纯度不高，信息较混乱）的部分再进行下一轮分割，以此类推。

本章小结

个性化检索（下）

- 情境感知的查询理解
- 情境感知的推荐服务
- 数据预处理
  - 数据采样
  - 数据归约
  - 数据离散

tongxu@ustc.edu.cn