《大数据技术原理与应用（第2版）》

http://dblab.xmu.edu.cn/post/bigdata

温馨提示：编辑幻灯片母版，可以修改每页PPT的厦大校徽和底部文字

第13章 大数据在不同领域的应用

（PPT版本号：2017年2月版本）

林子雨

厦门大学计算机科学系

E-mail: ziyulin@xmu.edu.cn

主页：http://www.cs.xmu.edu.cn/linziyu

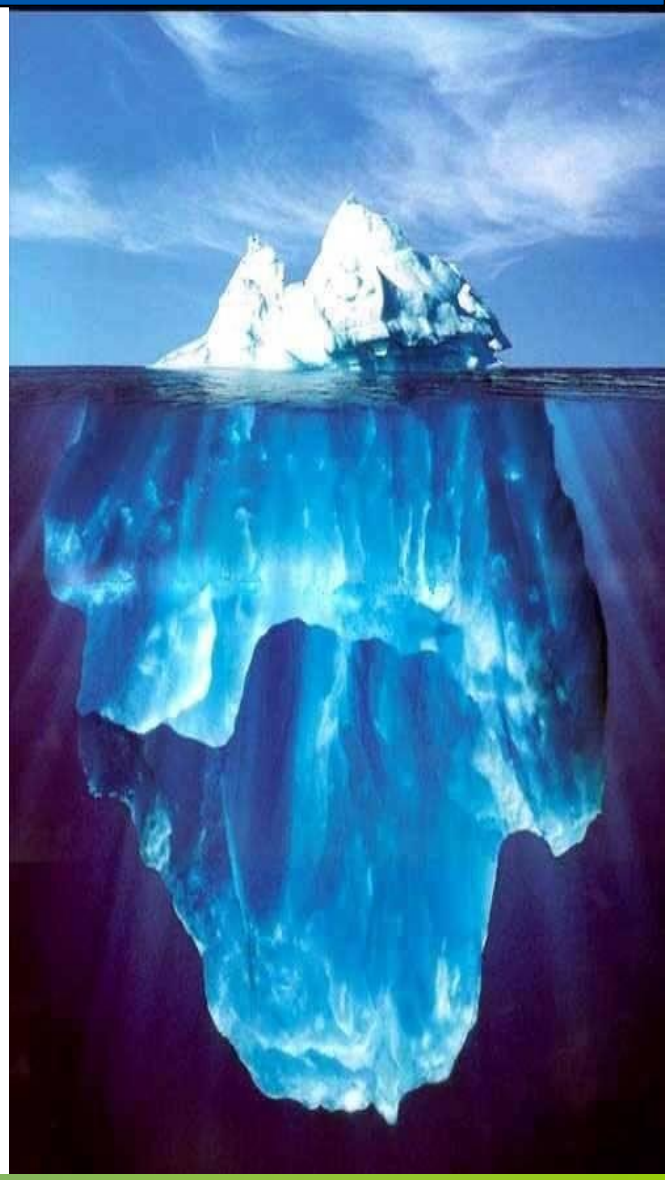
本章配套教学视频

《大数据技术原理与应用（第2版）》
《第13章 大数据在不同领域的应用》在线视频观看地址

http://dblab.xmu.edu.cn/post/bigdata-online-course/#lesson13

提纲

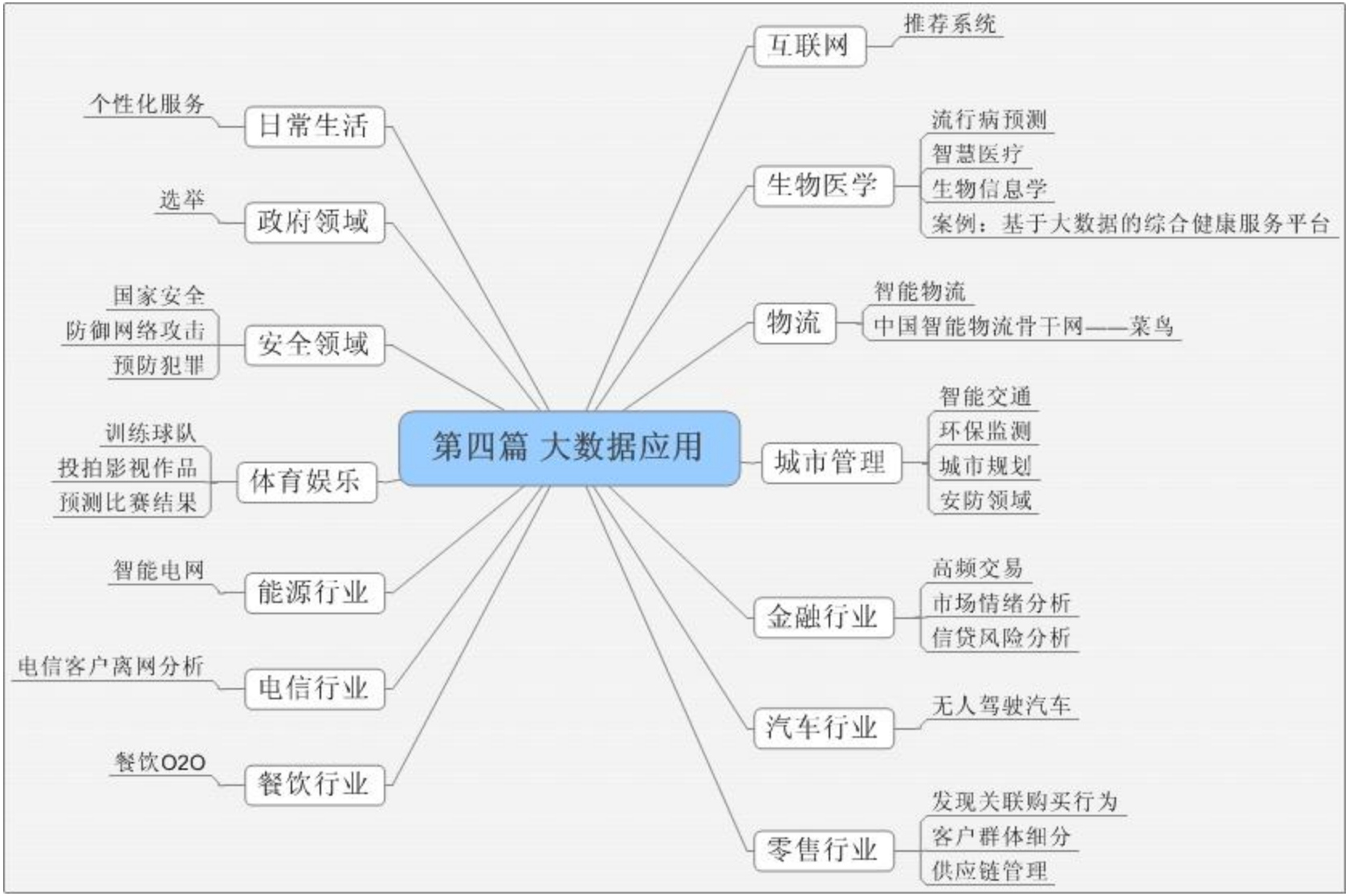
- 大数据应用概览
- 第13章 大数据在互联网领域的应用
- 13.1 推荐系统概述
- 13.2 推荐算法 – 协同过滤
- 13.3 协同过滤实践 – 电影推荐系统
- 第14章 大数据在生物医学领域的应用
- 14.1 基于大数据的综合健康服务平台
- 第15章 大数据的其他应用
- 15.1 大数据在物流领域中的应用



本PPT是如下教材的配套讲义：
《大数据技术原理与应用
——概念、存储、处理、分析与应用》
（2017年2月第2版）
ISBN:978-7-115-44330-4
厦门大学 林子雨 编著，人民邮电出版社

欢迎访问《大数据技术原理与应用》教材官方网站：
http://dblab.xmu.edu.cn/post/bigdata

大数据应用概览


第四篇 大数据应用
互联网
推荐系统
生物医学
流行病预测
智慧医疗
生物信息学
案例：基于大数据的综合健康服务平台
物流
智能物流
中国智能物流骨干网——菜鸟
城市管理
智能交通
环保监测
城市规划
安防领域
金融行业
高频交易
市场情绪分析
信贷风险分析
汽车行业
无人驾驶汽车
零售行业
发现关联购买行为
客户群体细分
供应链管理
餐饮行业
餐饮O2O
电信行业
电信客户离网分析
能源行业
智能电网
体育娱乐
训练球队
投拍影视作品
预测比赛结果
安全领域
国家安全
防御网络攻击
预防犯罪
政府领域
选举
日常生活
个性化服务

大数据应用概览

- 推荐系统：为用户推荐相关商品
- 生物医学
  - 流行病预测
  - 智慧医疗：利用医疗大数据，促进优质医疗资源共享、避免患者重复检查、促进医疗智能化
  - 生物信息学：利用生物大数据，深入了解生物学过程、疾病致病基因等
- 物流：基于大数据和物联网技术的智能物流，可以提高物流信息化和智能化水平，降低物流成本和提高物流效率
- 城市管理
  - 智能交通：利用交通大数据，实现交通实时监控、交通智能诱导、公共车辆管理、旅行信息服务、车辆辅助控制等各种应用
  - 环保监测：监测分析大气和水污染情况，为污染治理提供依据
  - 城市规划：比如，利用住房销售和出租数据，可以评价一个城区的住房分布
  - 安防领域：基于视频监控、人口信息、地理数据信息等，利用大数据技术实现智能化信息分析、预测和报警

大数据应用概览

- 金融
  - 高频交易：是指从那些人们无法利用的极为短暂的市场变化中寻求获利的计算机化交易。采用大数据技术决定交易
  - 市场情绪分析和信贷风险分析
- 汽车：无人驾驶汽车，实时采集车辆各种行驶数据和周围环境，利用大数据分析系统高效分析，迅速做出各种驾驶动作，引导车辆安全行驶
- 零售行业：发现关联购买行为、进行客户群体细分
- 餐饮行业：利用大数据为用户推荐消费内容、调整线下门店布局、控制店内人流量
- 电信行业：客户离网分析
- 能源行业：智能电网，以海量用户用电信息为基础进行大数据分析，可以更好理解电力客户用电行为，优化提升短期用电负荷预测系统，提前预知未来2-3个月的电网需求电量、用电高峰和低谷，合理设计电力需求响应系统
- 体育娱乐：2014巴西世界杯，基于海量比赛数据和球员训练数据，指定有针对性球队训练计划，帮助德国国家队问鼎2014世界杯冠军
- 安全领域：应用大数据技术防御网络攻击，警察应用大数据工具预防犯罪
- 政府领域：利用大数据改进选举策略

13.1 推荐系统概述

- 13.1.1 什么是推荐系统
- 13.1.2 长尾理论
- 13.1.3 推荐方法
- 13.1.4 推荐系统模型
- 13.1.5 推荐系统的应用

13.1.1 什么是推荐系统

- 互联网的飞速发展使我们进入了信息过载的时代，搜索引擎可以帮助我们查找内容，但只能解决明确的需求
- 为了让用户从海量信息中高效地获得自己所需的信息，推荐系统应运而生。推荐系统是大数据在互联网领域的典型应用，它可以通过分析用户的历史记录来了解用户的喜好，从而主动为用户推荐其感兴趣的信息，满足用户的个性化推荐需求
- 推荐系统是自动联系用户和物品的一种工具，和搜索引擎相比，推荐系统通过研究用户的兴趣偏好，进行个性化计算。推荐系统可发现用户的兴趣点，帮助用户从海量信息中去发掘自己潜在的需求

13.1.2 长尾理论

- 推荐系统可以创造全新的商业和经济模式，帮助实现长尾商品的销售
- “长尾”概念于2004年提出，用来描述以亚马逊为代表的电子商务网站的商业和经济模式
- 电子商务网站销售种类繁多，虽然绝大多数商品都不热门，但这些不热门的商品总数量极其庞大，所累计的总销售额将是一个可观的数字，也许会超过热门商品所带来的销售额
- 因此，可以通过发掘长尾商品并推荐给感兴趣的用户来提高销售额。这需要通过个性化推荐来实现

13.1.2 长尾理论

- 热门推荐是常用的推荐方式，广泛应用于各类网站中，如热门排行榜。但热门推荐的主要缺陷在于推荐的范围有限，所推荐的内容在一定时期内也相对固定。无法实现长尾商品的推荐
- 个性化推荐可通过推荐系统来实现。推荐系统通过发掘用户的行为记录，找到用户的个性化需求，发现用户潜在的消费倾向，从而将长尾商品准确地推荐给需要它的用户，进而提升销量，实现用户与商家的双赢

13.1.3 推荐方法

- 推荐系统的本质是建立用户与物品的联系，根据推荐算法的不同，推荐方法包括如下几类：
  - 专家推荐：人工推荐，由资深的专业人士来进行物品的筛选和推荐，需要较多的人力成本
  - 基于统计的推荐：基于统计信息的推荐（如热门推荐），易于实现，但对用户个性化偏好的描述能力较弱
  - 基于内容的推荐：通过机器学习的方法去描述内容的特征，并基于内容的特征来发现与之相似的内容
  - 协同过滤推荐：应用最早和最为成功的推荐方法之一，利用与目标用户相似的用户已有的商品评价信息，来预测目标用户对特定商品的喜好程度
  - 混合推荐：结合多种推荐算法来提升推荐效果

13.1.4 推荐系统模型

- 一个完整的推荐系统通常包括3个组成模块：用户建模模块、推荐对象建模模块、推荐算法模块：
  - 用户建模模块：对用户进行建模，根据用户行为数据和用户属性数据来分析用户的兴趣和需求
  - 推荐对象建模模块：根据对象数据对推荐对象进行建模
  - 推荐算法模块：基于用户特征和物品特征，采用推荐算法计算得到用户可能感兴趣的对象，并根据推荐场景对推荐结果进行一定调整，将推荐结果最终展示给用户

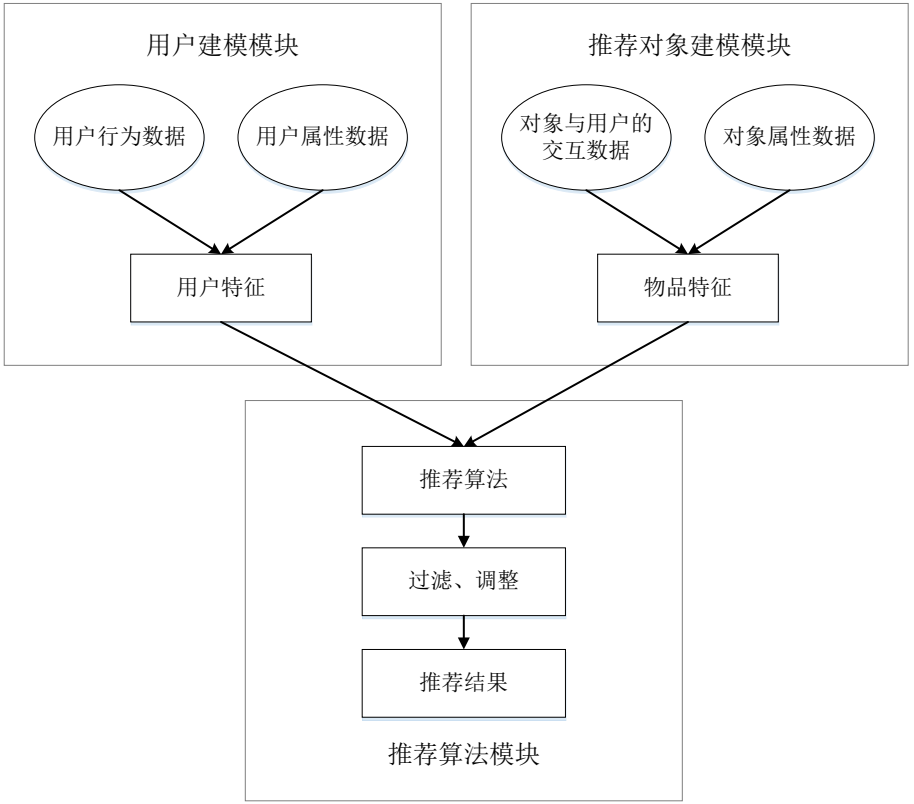


图11-1 推荐系统基本架构

13.1.5 推荐系统的应用

- 目前推荐系统已广泛应用于电子商务、在线视频、在线音乐、社交网络等各类网站和应用中
- 如亚马逊网站利用用户的浏览历史记录来为用户推荐商品，推荐的主要是用户未浏览过，但可能感兴趣、有潜在购买可能性的商品

图 亚马逊网站根据用户的浏览记录来推荐商品

13.1.5 推荐系统的应用

- 推荐系统在在线音乐应用中也逐渐发挥作用。音乐相比于电影数量更为庞大，个人口味偏向也更为明显，仅依靠热门推荐是远远不够的
- 虾米音乐网根据用户的音乐收藏记录来分析用户的音乐偏好，以进行推荐。例如，推荐同一风格的歌曲，或是推荐同一歌手的其他歌曲

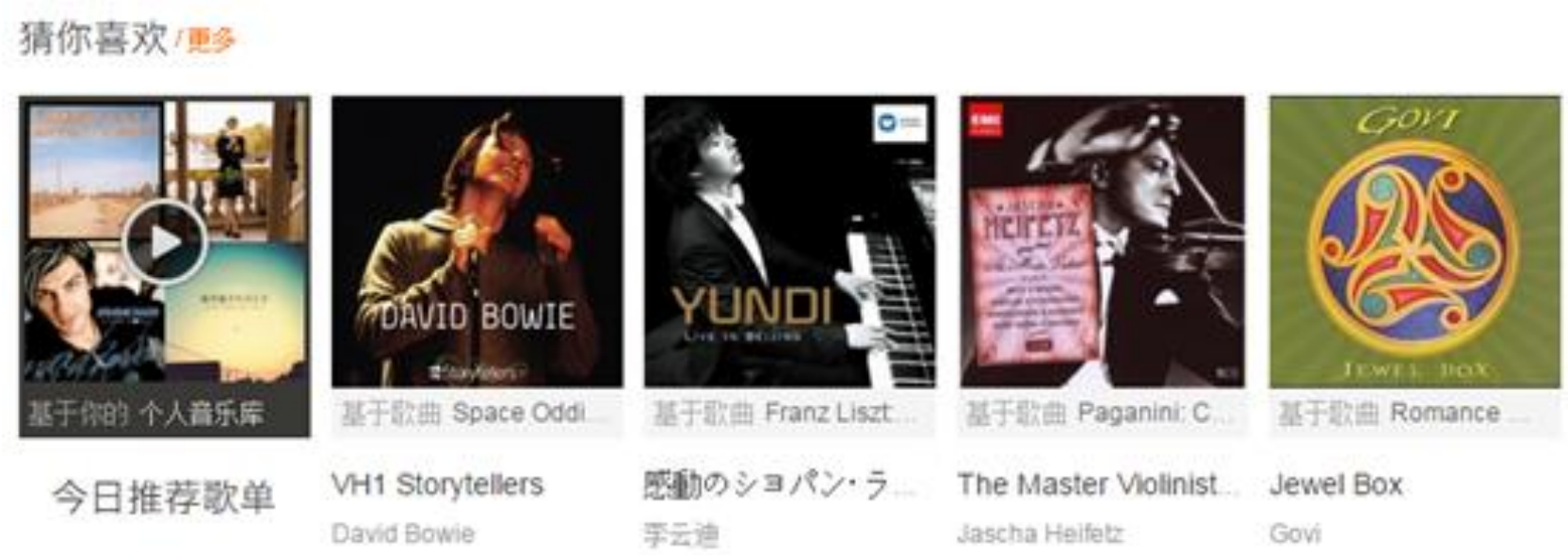


图 虾米音乐网根据用户的音乐收藏来推荐歌曲

13.2 协同过滤

- 推荐技术从被提出到现在已有十余年，在多年的发展历程中诞生了很多新的推荐算法。协同过滤作为最早、最知名的推荐算法，不仅在学术界得到了深入研究，而且至今在业界仍有广泛的应用
- 协同过滤可分为基于用户的协同过滤和基于物品的协同过滤

- 13.2.1 基于用户的协同过滤（UserCF）
- 13.2.2 基于物品的协同过滤（ItemCF）
- 13.2.3 UserCF算法和ItemCF算法的对比

13.2.1 基于用户的协同过滤（UserCF）

- 基于用户的协同过滤算法（简称UserCF算法）在1992年被提出，是推荐系统中最古老的算法
- UserCF算法符合人们对于“趣味相投”的认知，即兴趣相似的用户往往有相同的物品喜好：当目标用户需要个性化推荐时，可以先找到和目标用户有相似兴趣的用户群体，然后将这个用户群体喜欢的、而目标用户没有听说过的物品推荐给目标用户
- UserCF算法的实现主要包括两个步骤：
  - 第一步：找到和目标用户兴趣相似的用户集合
  - 第二步：找到该集合中的用户所喜欢的、且目标用户没有听说过的物品推荐给目标用户

13.2.1 基于用户的协同过滤（UserCF）

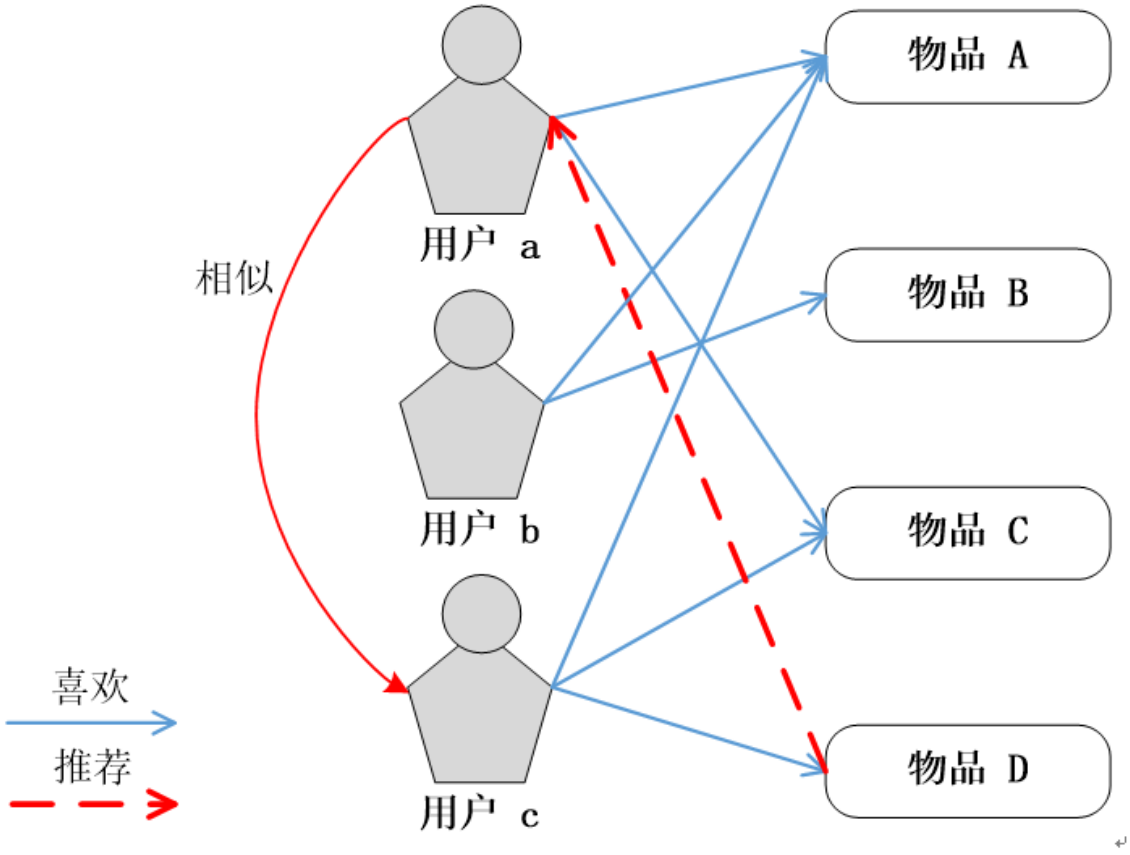


图 基于用户的协同过滤（User CF）

13.2.1 基于用户的协同过滤（UserCF）

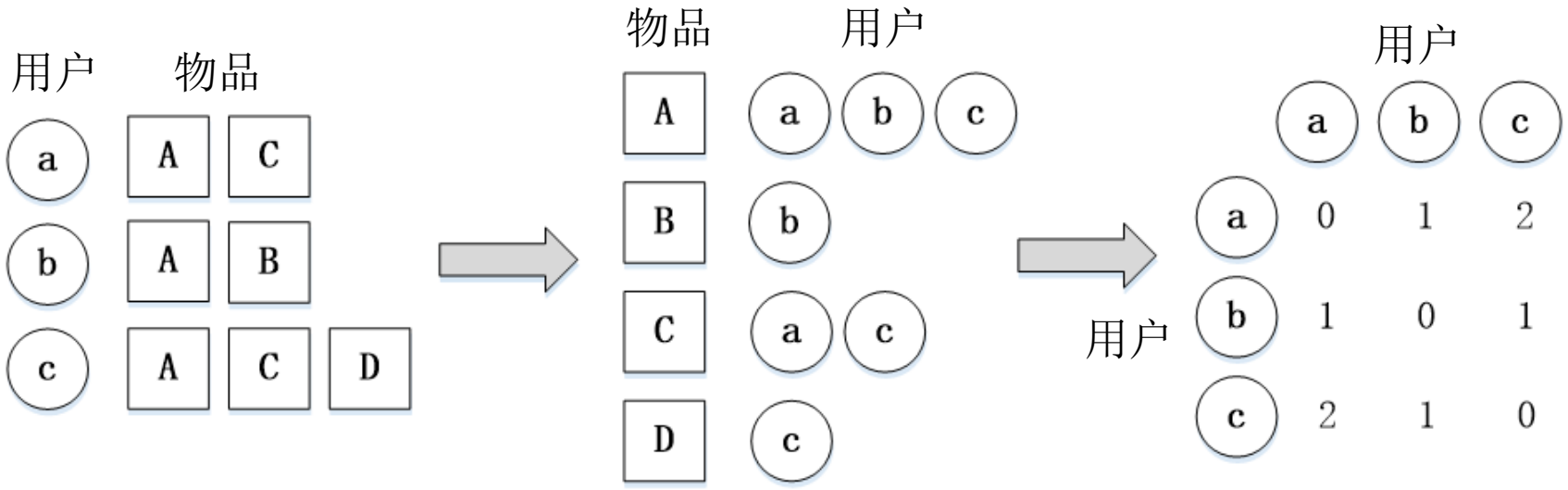
- 实现UserCF算法的关键步骤是计算用户与用户之间的兴趣相似度。目前较多使用的相似度算法有：
  - 泊松相关系数（Person Correlation Coefficient）
  - 余弦相似度（Cosine-based Similarity）
  - 调整余弦相似度（Adjusted Cosine Similarity）

- 给定用户u和用户v，令N(u)表示用户u感兴趣的物品集合，令N(v)为用户v感兴趣的物品集合，则使用余弦相似度进行计算用户相似度的公式为：

$$w_{uv} = \frac{|N(u) \cap N(v)|}{\sqrt{|N(u)||N(v)|}}$$

13.2.1 基于用户的协同过滤（UserCF）

- 由于很多用户相互之间并没有对同样的物品产生过行为，因此其相似度公式的分子为0，相似度也为0
- 我们可以利用物品到用户的倒排表（每个物品所对应的、对该物品感兴趣的用户列表），仅对有对相同物品产生交互行为的用户进行计算



（a）用户喜欢的物品列表　（b）物品对应的用户列表　（b）相似度矩阵W

图 物品到用户倒排表及用户相似度矩阵

13.2.1 基于用户的协同过滤（UserCF）

- 得到用户间的相似度后，再使用如下公式来度量用户u对物品i的兴趣程度P_{ui}：

$$p(u,i)=\sum_{v\in S(u,K)\bigcap N(i)} w_{uv}r_{vi}$$

- 其中，S(u, K)是和用户u兴趣最接近的K个用户的集合，N(i)是喜欢物品i的用户集合，W_{uv}是用户u和用户v的相似度，r_{vi}是隐反馈信息，代表用户v对物品i的感兴趣程度，为简化计算可令r_{vi}=1

- 对所有物品计算P_{ui}后，可以对P_{ui}进行降序处理，取前N个物品作为推荐结果展示给用户u（称为Top-N推荐）

13.2.2 基于物品的协同过滤（ItemCF）

- 基于物品的协同过滤算法（简称ItemCF算法）是目前业界应用最多的算法。无论是亚马逊还是Netflix，其推荐系统的基础都是ItemCF算法
- ItemCF算法是给目标用户推荐那些和他们之前喜欢的物品相似的物品。ItemCF算法主要通过分析用户的行为记录来计算物品之间的相似度
- 该算法基于的假设是：物品A和物品B具有很大的相似度是因为喜欢物品A的用户大多也喜欢物品B。例如，该算法会因为你购买过《数据挖掘导论》而给你推荐《机器学习实战》，因为买过《数据挖掘导论》的用户多数也购买了《机器学习实战》

13.2.2 基于物品的协同过滤（ItemCF）

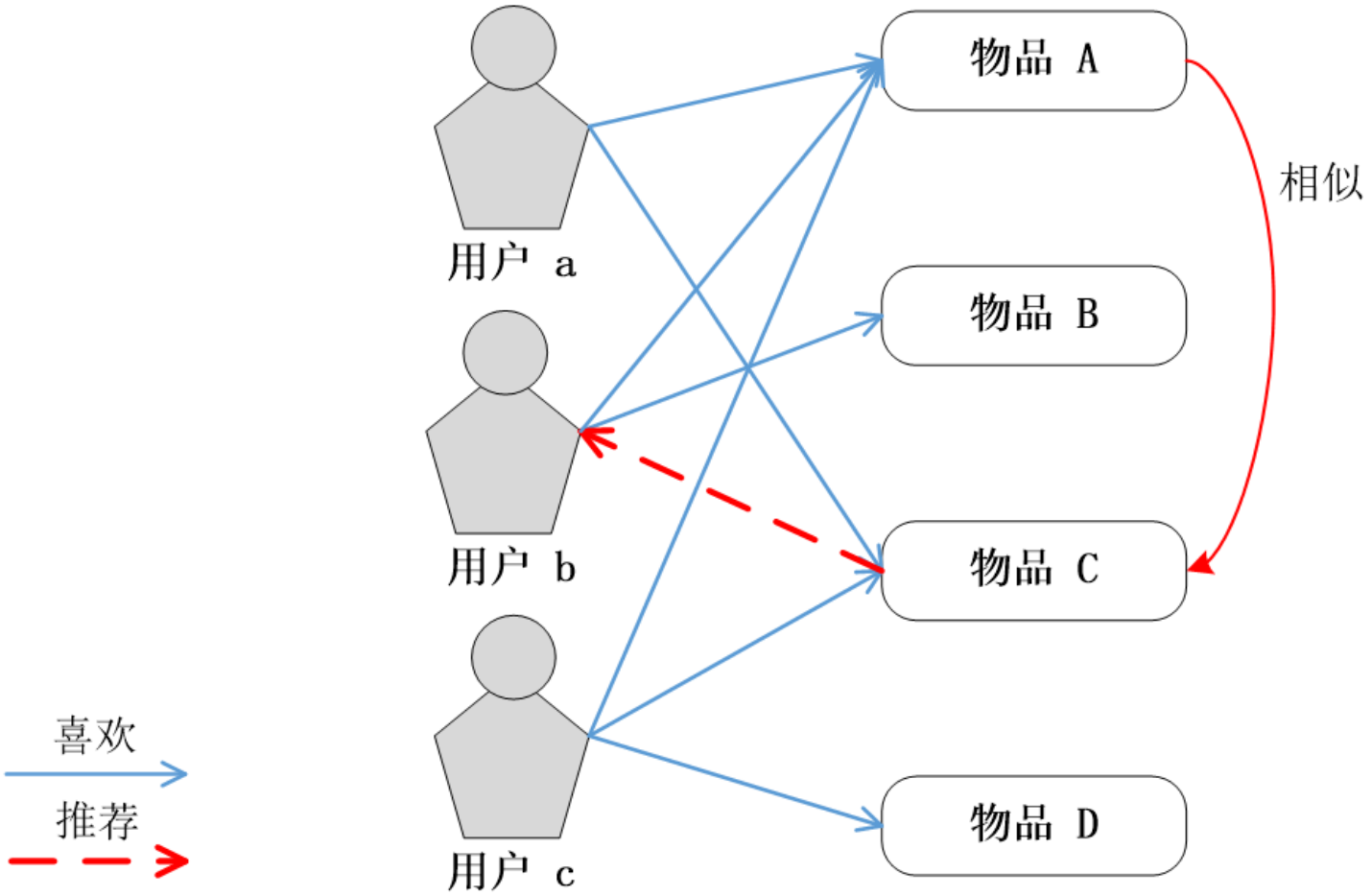


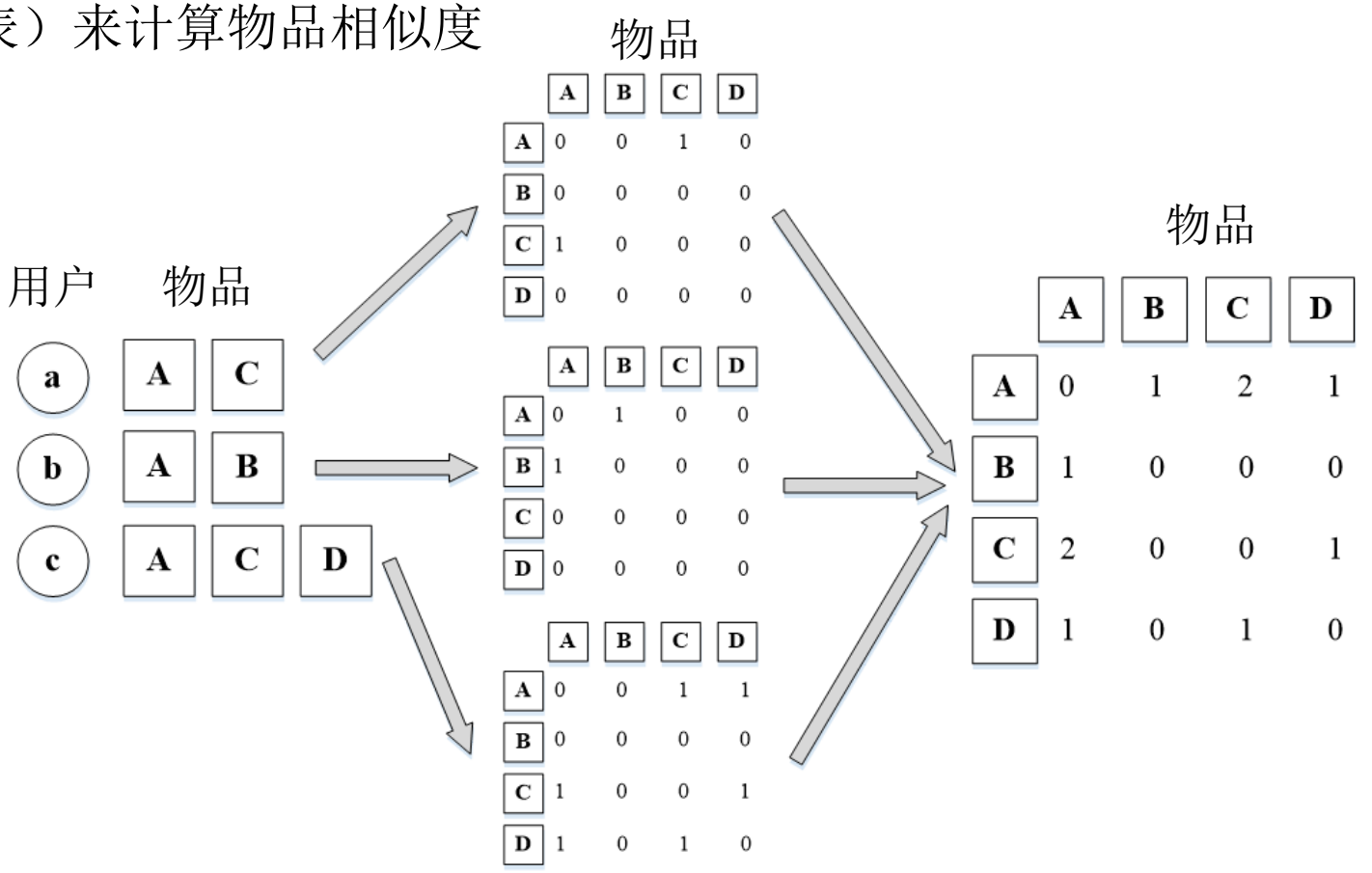
图 基于物品的协同过滤（Item CF）

13.2.2 基于物品的协同过滤（ItemCF）

- ItemCF算法与UserCF算法类似，计算也分为两步：
  - 第一步：计算物品之间的相似度；
  - 第二步：根据物品的相似度和用户的历史行为，给用户生成推荐列表。

13.2.2 基于物品的协同过滤（ItemCF）

- ItemCF算法通过建立用户到物品倒排表（每个用户喜欢的物品的列表）来计算物品相似度



（a）用户喜欢的物品列表　（b）物品相似度矩阵M　（c）物品相似度矩阵R

图 用户到物品倒排表及物品相似度矩阵

13.2.2 基于物品的协同过滤（ItemCF）

- ItemCF计算的是物品相似度，再使用如下公式来度量用户u对物品j的兴趣程度P_{uj}(与UserCF类似)：

$$p_{uj} = \sum_{i \in N(u) \cap S(j,K)} w_{ji} r_{ui}$$

其中，S(j, K)是和物品j最相似的K个物品的集合，N(u)是用户u喜欢的物品的集合，w_{ji}物品i和物品j的相似度，r_{ui}是隐反馈信息，代表用户u对物品i的感兴趣程度，为简化计算可令r_{vi}=1

13.2.3 UserCF算法和ItemCF算法的对比

- UserCF算法和ItemCF算法的思想、计算过程都相似
- 两者最主要的区别：
  - UserCF算法推荐的是那些和目标用户有共同兴趣爱好的其他用户所喜欢的物品
  - ItemCF算法推荐的是那些和目标用户之前喜欢的物品类似的其他物品
- UserCF算法的推荐更偏向社会化，而ItemCF算法的推荐更偏向于个性化

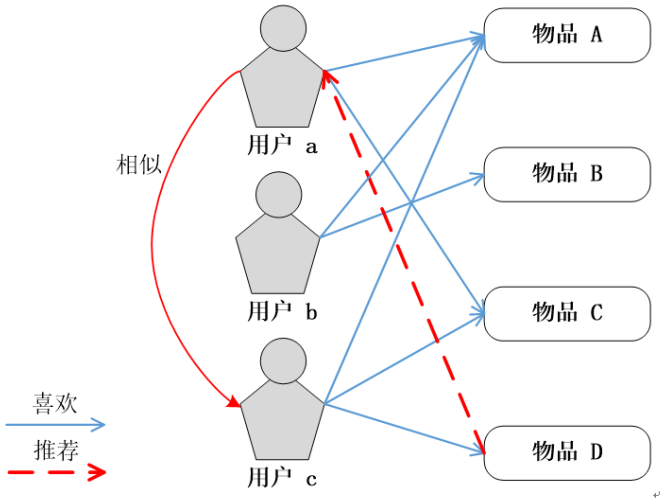


图 基于用户的协同过滤（User CF）

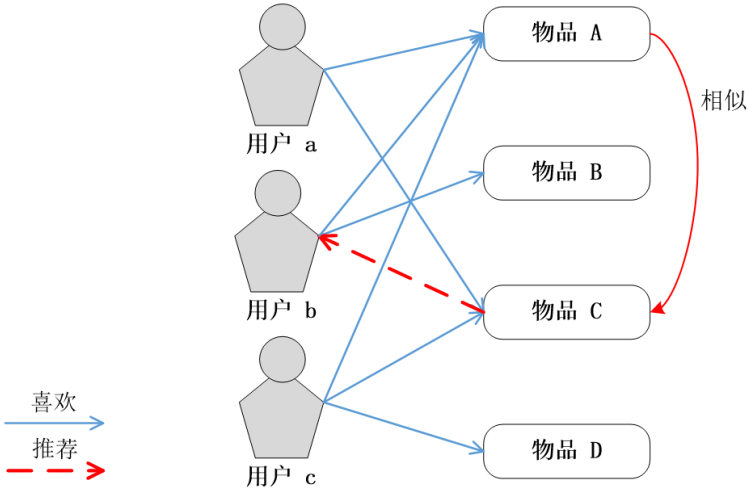


图 基于物品的协同过滤（Item CF）

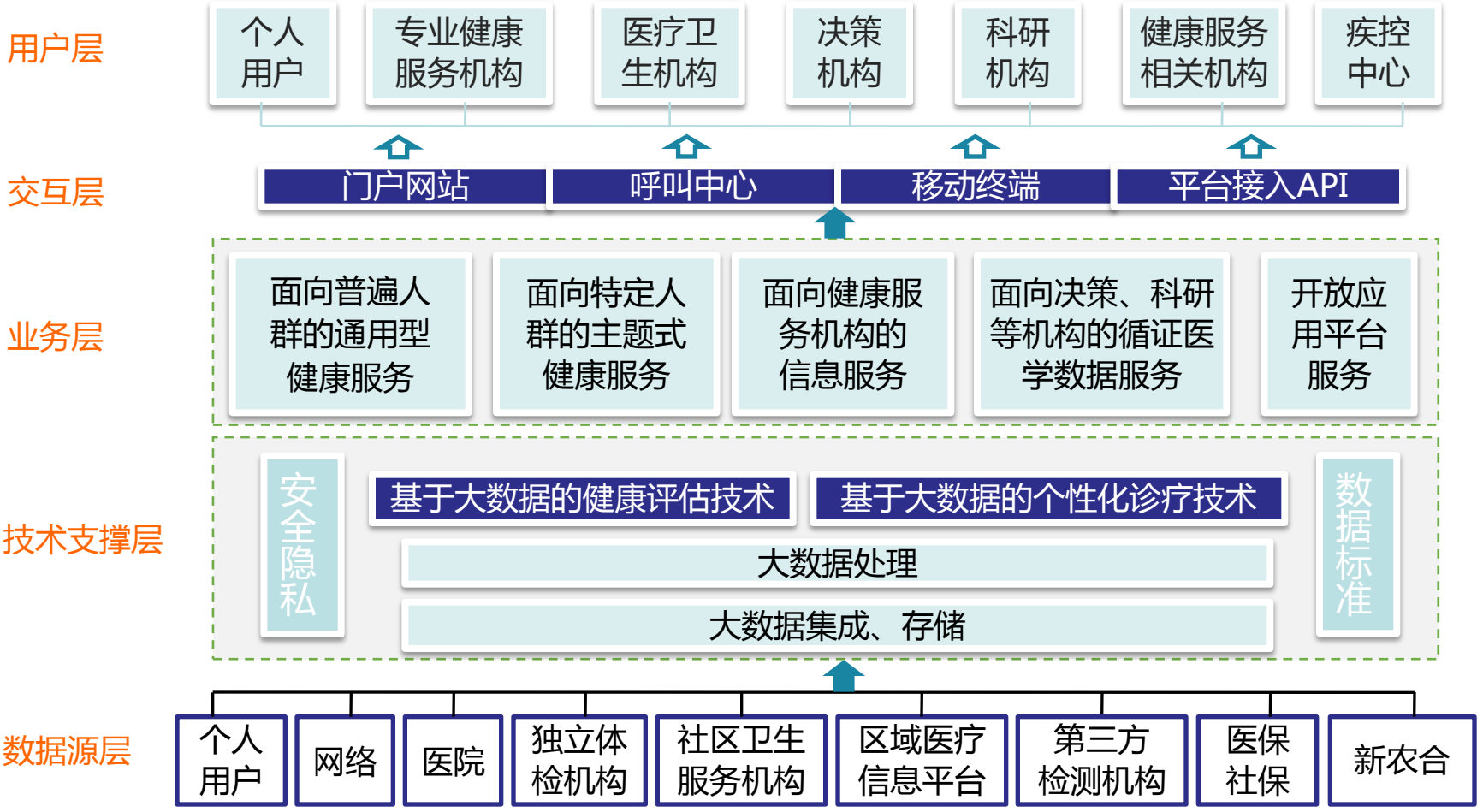
13.2.3 UserCF算法和ItemCF算法的对比

- UserCF算法的推荐更偏向社会化：适合应用于新闻推荐、微博话题推荐等应用场景，其推荐结果在新颖性方面有一定的优势
- UserCF缺点：随着用户数目的增大，用户相似度计算复杂度越来越高。而且UserCF推荐结果相关性较弱，难以对推荐结果作出解释，容易受大众影响而推荐热门物品

- ItemCF算法的推荐更偏向于个性化：适合应用于电子商务、电影、图书等应用场景，可以利用用户的历史行为给推荐结果作出解释，让用户更为信服推荐的效果
- ItemCF缺点：倾向于推荐与用户已购买商品相似的商品，往往会出现多样性不足、推荐新颖度较低的问题

14.1基于大数据的综合健康服务平台

目标：构建覆盖全生命周期、内涵丰富、结构合理的以人为本全面连续的综合健康服务体系，利用大数据技术和智能设备技术，提供线上线下相结合的公众健康服务，实现“未病先防、已病早治、既病防变、愈后防复”，满足社会公众多层次、多方位的健康服务需求，提升人民群众的身心健康水平。



15.1 大数据在物流领域的应用

智能物流集成商案例：阿里巴巴的中国智能物流骨干网（地网）

菜鸟网络到底是什么?

- 中国智能物流骨干网，又名“菜鸟”
- **菜鸟网络计划在5到8年内，打造一个全国性的超级物流网。**
- 这个网络能在24小时内将货物运抵国内任何地区，能支撑日均300亿元(年度约10万亿元)的巨量网络零售额。

1000亿元投资物流基础设施　强强联手共建智能骨干网络物流信息系统向所有的制造商、网商、快递公司、第三方物流公司完全开放

阿里物流体系

天网	地网
天猫牵头负责与各大物流快递公司对接的数据平台	即“菜鸟”，又称“中国智能物流骨干网（CSN）”

本章小结

- 本章内容首先介绍了推荐系统的概念，推荐系统可帮助用户从海量信息中高效地获得自己所需的信息
- 接着介绍了不同的推荐方法以及推荐系统在电子商务、在线音乐等网站中的具体应用
- 本章重点介绍了协同过滤算法，协同过滤算法是最早推出的推荐算法，至今仍获得广泛的应用，协同过滤包括基于用户的协同过滤算法（UserCF）和基于物品的协同过滤算法（ItemCF）。这两种协同过滤算法思想相近，核心是计算用户、物品的相似度，依据相似度来做出推荐。然而，这两种协同过滤算法各自适合的应用场景不同，UserCF适合社交化应用，可作出新颖的推荐，而ItemCF则适合用于电子商务、电影等应用。在具体实践中，常常结合多种推荐算法来提升推荐效果
- 最后介绍了大数据在医疗健康领域的应用和大数据在物流领域的应用

附录：主讲教师林子雨简介

主讲教师：林子雨

单位：厦门大学计算机科学系
E-mail: ziyulin@xmu.edu.cn
个人网页：http://www.cs.xmu.edu.cn/linziyu
数据库实验室网站：http://dblab.xmu.edu.cn

扫一扫访问个人主页

林子雨，男，1978年出生，博士（毕业于北京大学），现为厦门大学计算机科学系助理教授（讲师），曾任厦门大学信息科学与技术学院院长助理、晋江市发展和改革局副局长。中国计算机学会数据库专业委员会委员，中国计算机学会信息系统专业委员会委员，荣获“2016中国大数据创新百人”称号。中国高校首个“数字教师”提出者和建设者，厦门大学数据库实验室负责人，厦门大学云计算与大数据研究中心主要建设者和骨干成员，2013年度厦门大学奖教金获得者。主要研究方向为数据库、数据仓库、数据挖掘、大数据、云计算和物联网，并以第一作者身份在《软件学报》《计算机学报》和《计算机研究与发展》等国家重点期刊以及国际学术会议上发表多篇学术论文。作为项目负责人主持的科研项目包括1项国家自然科学青年基金项目(No.61303004)、1项福建省自然科学青年基金项目(No.2013J05099)和1项中央高校基本科研业务费项目(No.2011121049)，同时，作为课题负责人完成了国家发改委城市信息化重大课题、国家物联网重大应用示范工程区域试点泉州市工作方案、2015泉州市互联网经济调研等课题。中国高校首个“数字教师”提出者和建设者，2009年至今，“数字教师”大平台累计向网络免费发布超过100万字高价值的研究和教学资料，累计网络访问量超过100万次。打造了中国高校大数据教学知名品牌，编著出版了中国高校第一本系统介绍大数据知识的专业教材《大数据技术原理与应用》，并成为京东、当当网等网店畅销书籍；建设了国内高校首个大数据课程公共服务平台，为教师教学和学生学习大数据课程提供全方位、一站式服务，年访问量超过50万次。具有丰富的政府和企业信息化培训经验，厦门大学管理学院EDP中心、浙江大学管理学院EDP中心、厦门大学继续教育学院、泉州市科技培训中心特邀培训讲师，曾给中国移动通信集团公司、福州马尾区政府、福建龙岩卷烟厂、福建省物联网科学研究院、石狮市物流协会、厦门市物流协会、浙江省中小企业家、四川泸州企业家、江苏沛县企业家等开展信息化培训，累计培训人数达3000人以上。

附录：《大数据技术原理与应用》教材

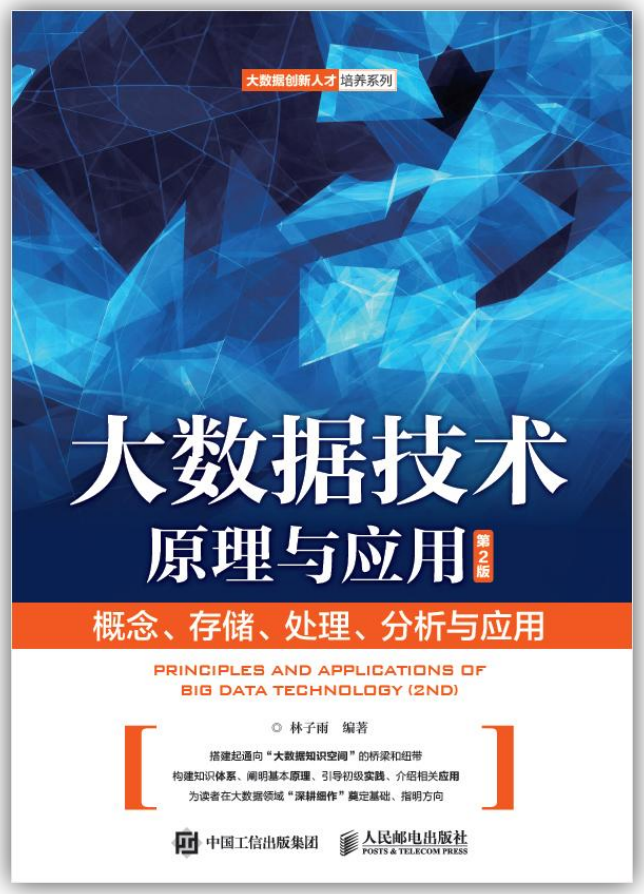
扫一扫访问教材官网

《大数据技术原理与应用——概念、存储、处理、分析与应用（第2版）》，由厦门大学计算机科学系林子雨博士编著，是中国高校第一本系统介绍大数据知识的专业教材。

全书共有15章，系统地论述了大数据的基本概念、大数据处理架构Hadoop、分布式文件系统HDFS、分布式数据 库HBase、NoSQL数据库、云数据库、分布式并行编程模型MapReduce、Spark、流计算、图计算、数据可视化以及大数据在互联网、生物医学和物流等各个领域的应用。在Hadoop、HDFS、HBase和MapReduce等重要章节，安排了入门级的实践操作，让读者更好地学习和掌握大数据关键技术。

本书可以作为高等院校计算机专业、信息管理等相关专业的大数据课程教材，也可供相关技术人员参考、学习、培训之用。

欢迎访问《大数据技术原理与应用——概念、存储、处理、分析与应用》教材官方网站：
http://dblab.xmu.edu.cn/post/bigdata



附录：中国高校大数据课程公共服务平台

中国高校大数据课程

公 共 服 务 平 台

http://dblab.xmu.edu.cn/post/bigdata-teaching-platform/

扫一扫访问平台主页

扫一扫观看3分钟FLASH动画宣传片


Thank You!