中国大学MOOC

《大数据技术原理与应用》

http://www.icourse163.org/course/XMU-1002335004

中国大学MOOC　2018年春季学期

第13讲 大数据在不同领域的应用

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

中国大学MOOC《大数据技术原理与应用》课程地址:
http://www.icourse163.org/course/XMU-1002335004

课程重要资料

大数据软件安装和编程指南

读者在学习《大数据技术原理与应用》MOOC课程时，需要进行安装Linux系统和各种大数据软件，并开展基础编程实践。这个实践过程，如果没有配套的指南，将会耗费读者大量的时间，而且在实践过程中的大量障碍，网络上都没有现成的答案，会给读者带来很大的挫折感，感觉学习大数据是一件“痛苦的事情”。为了帮助读者实现“零基础”学习大数据并顺畅完成实验环境搭建和开展基础编程，课程团队建设了与本课程配套的《大数据软件安装和基础编程实践指南》，读者即使没有Linux系统知识，没有学习过任何大数据软件的使用，也可以在自己熟悉的Windows操作系统上顺利安装Linux虚拟机和各种大数据软件，顺利完成基础的编程实践，让学习入门级大数据技术变得“相对容易”。课程团队在过去5年时间里建设的在线免费资源，大大降低了大数据技术学习门槛，已经较好地帮助很多大数据初学者顺利完成了大数据基础实践，获得了读者的好评，目前，厦门大学数据库实验室网站上的这些在线免费大数据教学资源每年访问量超过100万次，在国内高校形成了广泛的影响力。

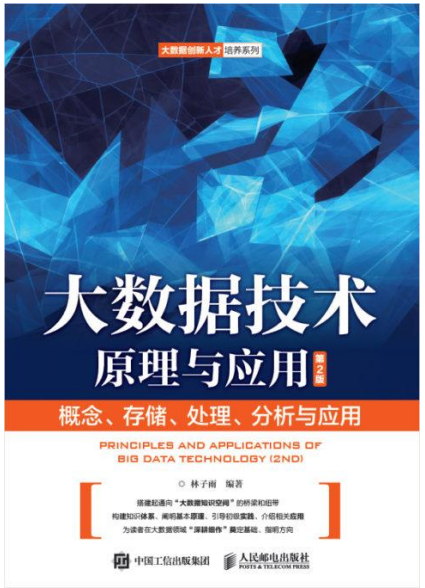
重要提示：读者在学习《大数据技术原理与应用》MOOC课程时，在中国大学MOOC课程的栏目中，有一个名称为“大数据软件安装和编程指南”的子栏目，进入这个栏目，可以帮助读者顺利完成大数据上机环境的安装和开展基础编程实践。在观看每个章节的MOOC视频时，可以充分利用该栏目辅助自己完成上机实验操作。

欢迎访问教材官网获取教学资源

《大数据技术原理与应用——大数据概念、存储、处理、分析与应用》

教材官网：http://dblab.xmu.edu.cn/post/bigdata

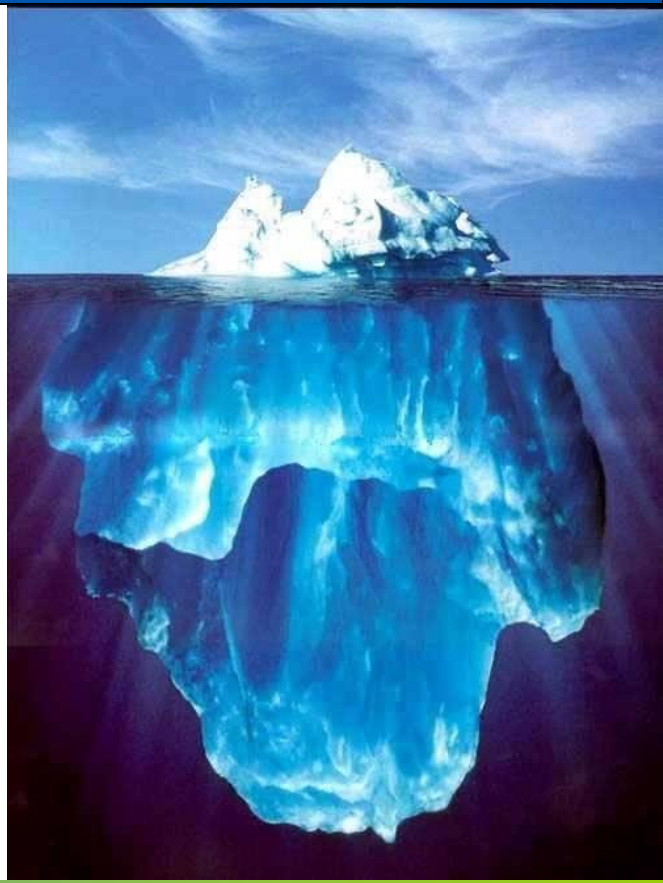
厦门大学 林子雨编著，人民邮电出版社，2017年1月第2版
ISBN:978-7-115-44330-4



- 国内高校第一本系统介绍大数据知识专业教材
- 京东、当当等各大网店畅销书籍
- 大数据入门教材精品
- 国内多所高校采用本教材开课
- 配套目前国内高校最完备的课程公共服务平台
- 福建省精品在线开放课程

提纲

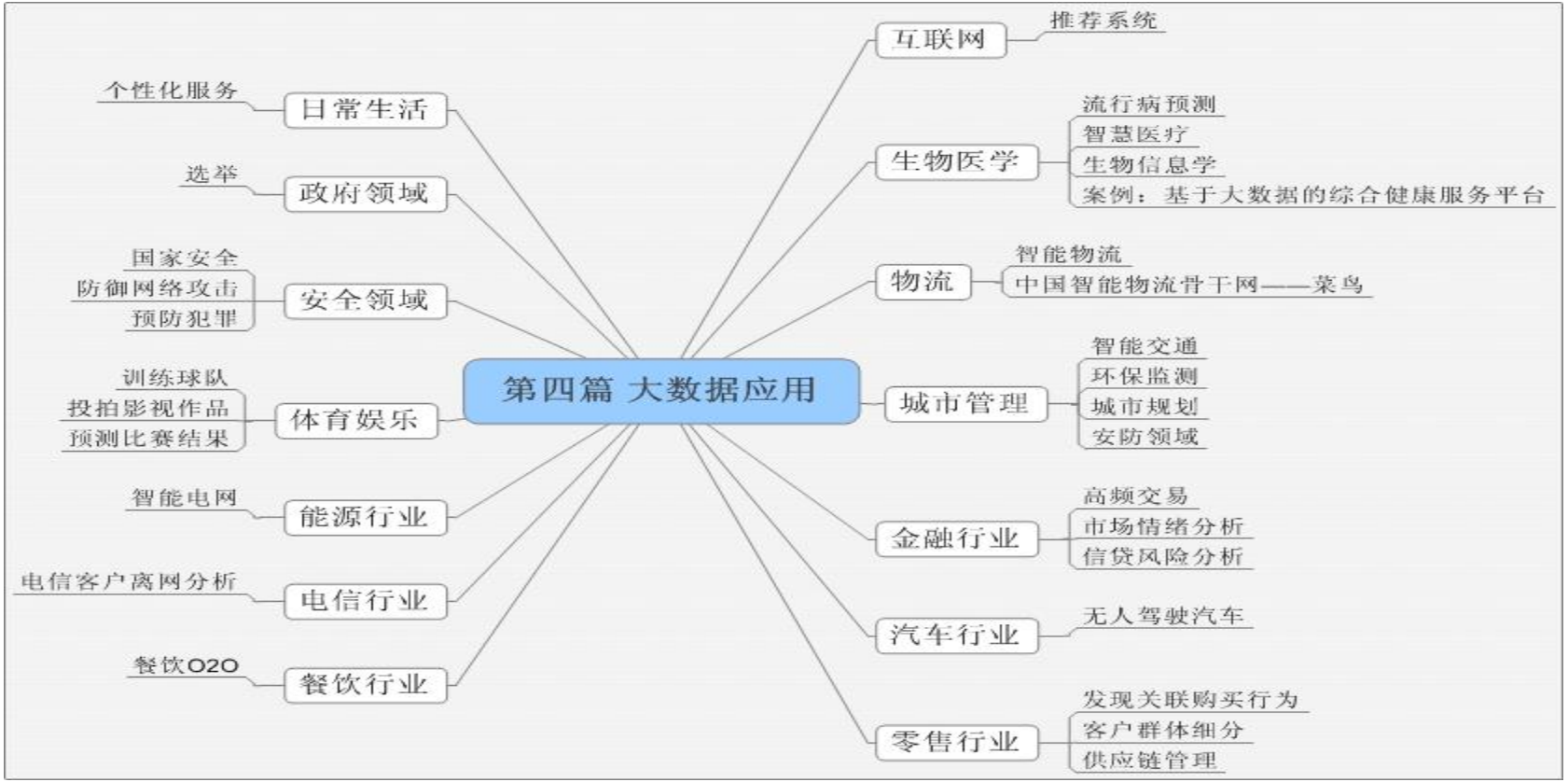
- **13.1 大数据应用概览**
- **13.2 大数据在互联网领域的应用——推荐系统**
- **13.3 大数据在智能医疗和智能物流领域的应用**



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1月第2版）
厦门大学 林子雨 编著，人民邮电出版社
ISBN:978-7-115-44330-4

欢迎访问《大数据技术原理与应用》教材官方网站，免费获取教材配套资源：
http://dblab.xmu.edu.cn/post/bigdata

13.1 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
日常生活
个性化服务
政府领域
选举
安全领域
国家安全
防御网络攻击
预防犯罪
体育娱乐
训练球队
投拍影视作品
预测比赛结果
能源行业
智能电网
电信行业
电信客户离网分析
餐饮行业
餐饮O2O

13.2 大数据在互联网领域的应用——推荐系统

13.2.1 推荐系统概述
13.2.2 基于用户的协同过滤（UserCF）
13.2.3 基于物品的协同过滤（ItemCF）
13.2.4 UserCF算法和ItemCF算法的对比

13.2.1 推荐系统概述

- 13.2.1.1 什么是推荐系统
- 13.2.1.2 长尾理论
- 13.2.1.3 推荐方法
- 13.2.1.4 推荐系统模型
- 13.2.1.5 推荐系统的应用

13.2.1.1 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

13.2.1.2 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

13.2.1.2 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

13.2.1.3 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
  - 基于统计的推荐
  - 基于内容的推荐
  - 协同过滤推荐
  - 混合推荐

13.2.1.4 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

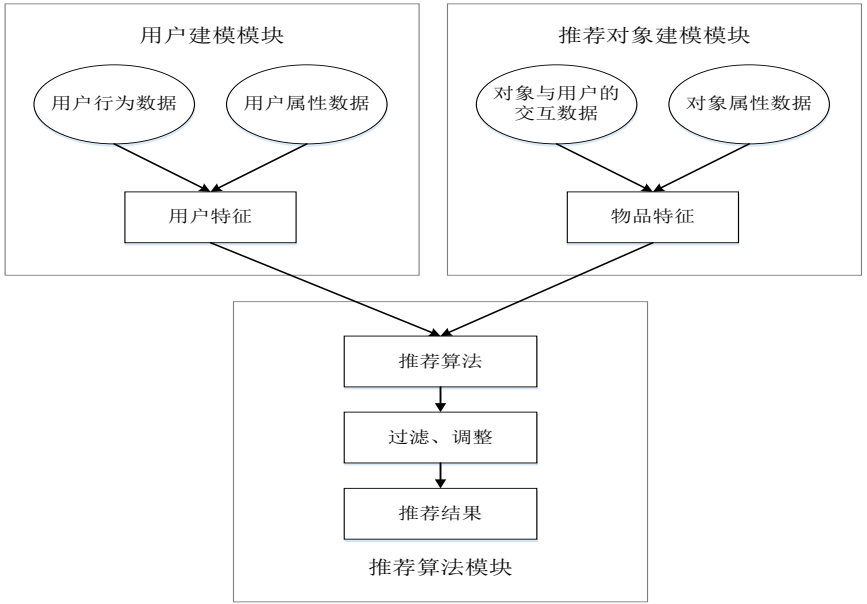


图11-1 推荐系统基本架构

13.2.1.5 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11-2 亚马逊网站根据用户的浏览记录来推荐商品

13.2.1.5 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

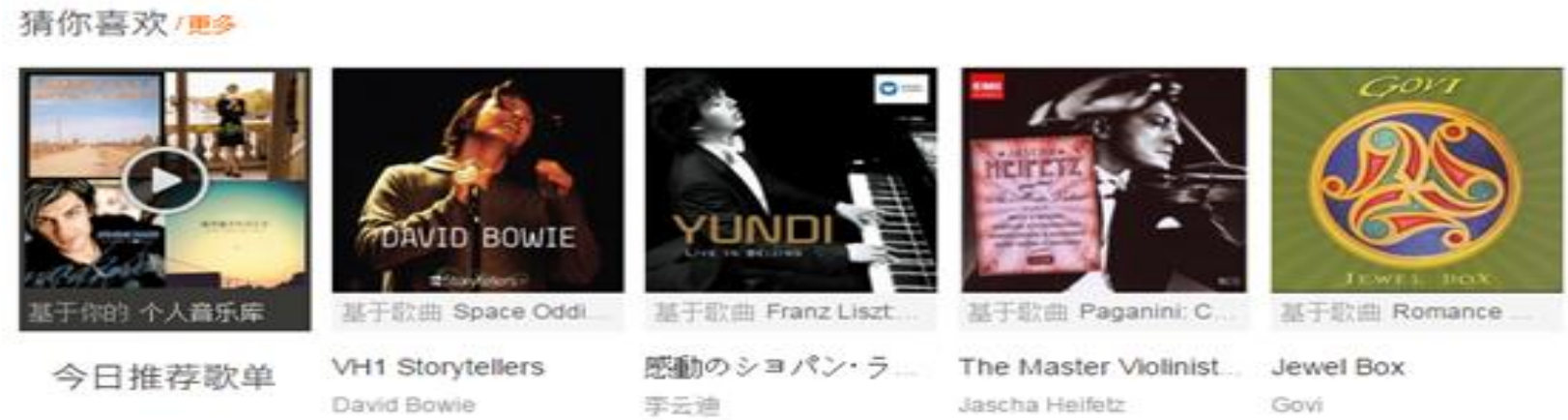


图11-3 虾米音乐网根据用户的音乐收藏来推荐歌曲

13.2.2 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

13.2.2 基于用户的协同过滤（UserCF）

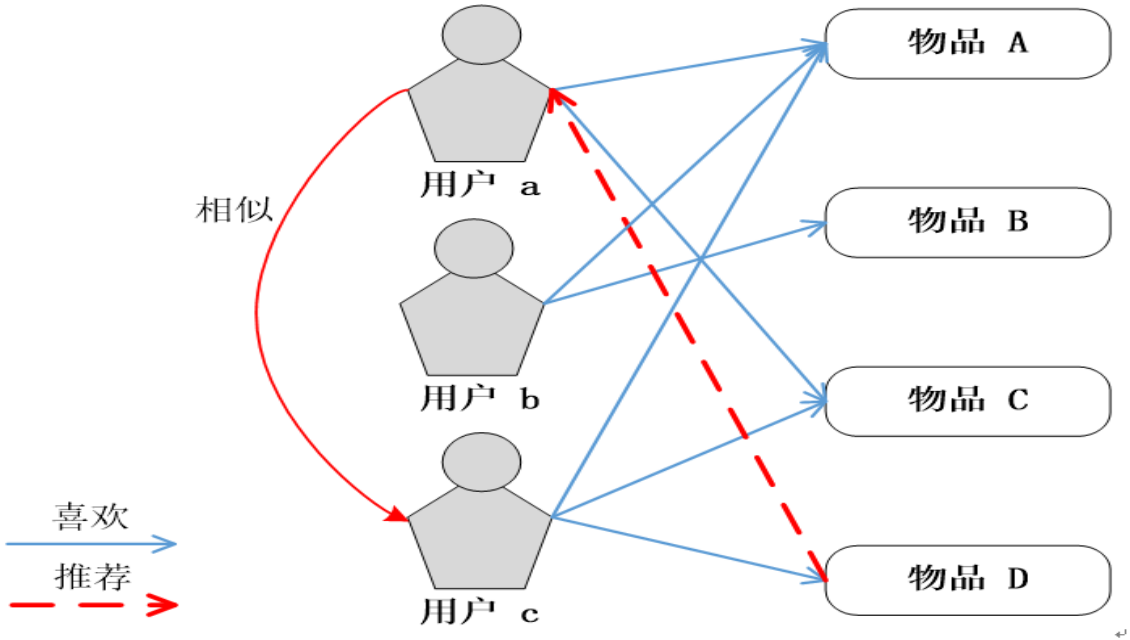


图11-4 基于用户的协同过滤（User CF）

13.2.2 基于用户的协同过滤（UserCF）

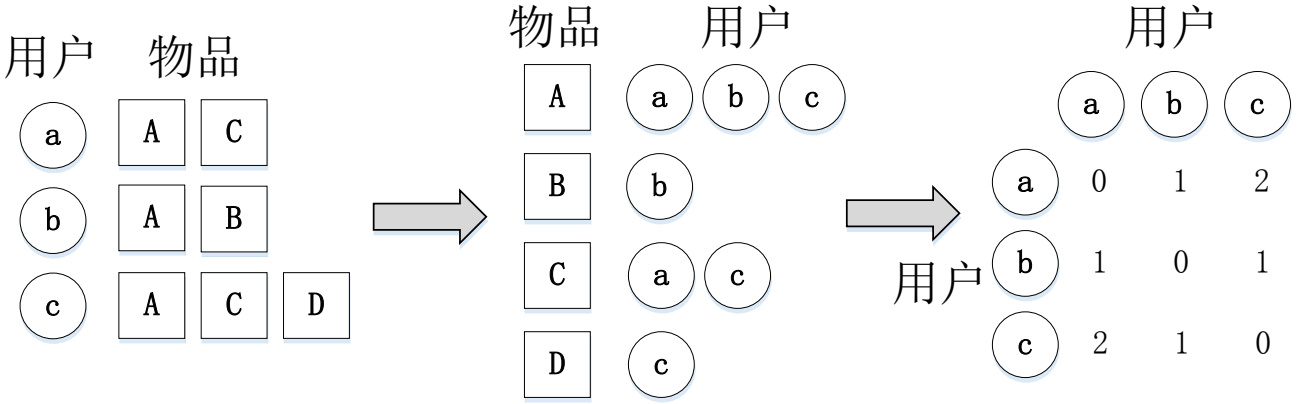
- 实现UserCF算法的关键步骤是计算用户与用户之间的兴趣相似度。目前较多使用的相似度算法有：
  - 泊松相关系数（Person Correlation Coefficient）
  - 余弦相似度（Cosine-based Similarity）
  - 调整余弦相似度（Adjusted Cosine Similarity）

- 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：

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

13.2.2 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　　（c）相似度矩阵W

图11-5 物品到用户倒排表及用户相似度矩阵

13.2.2 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}：

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

- 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）

13.2.3 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。

13.2.3 基于物品的协同过滤（ItemCF）

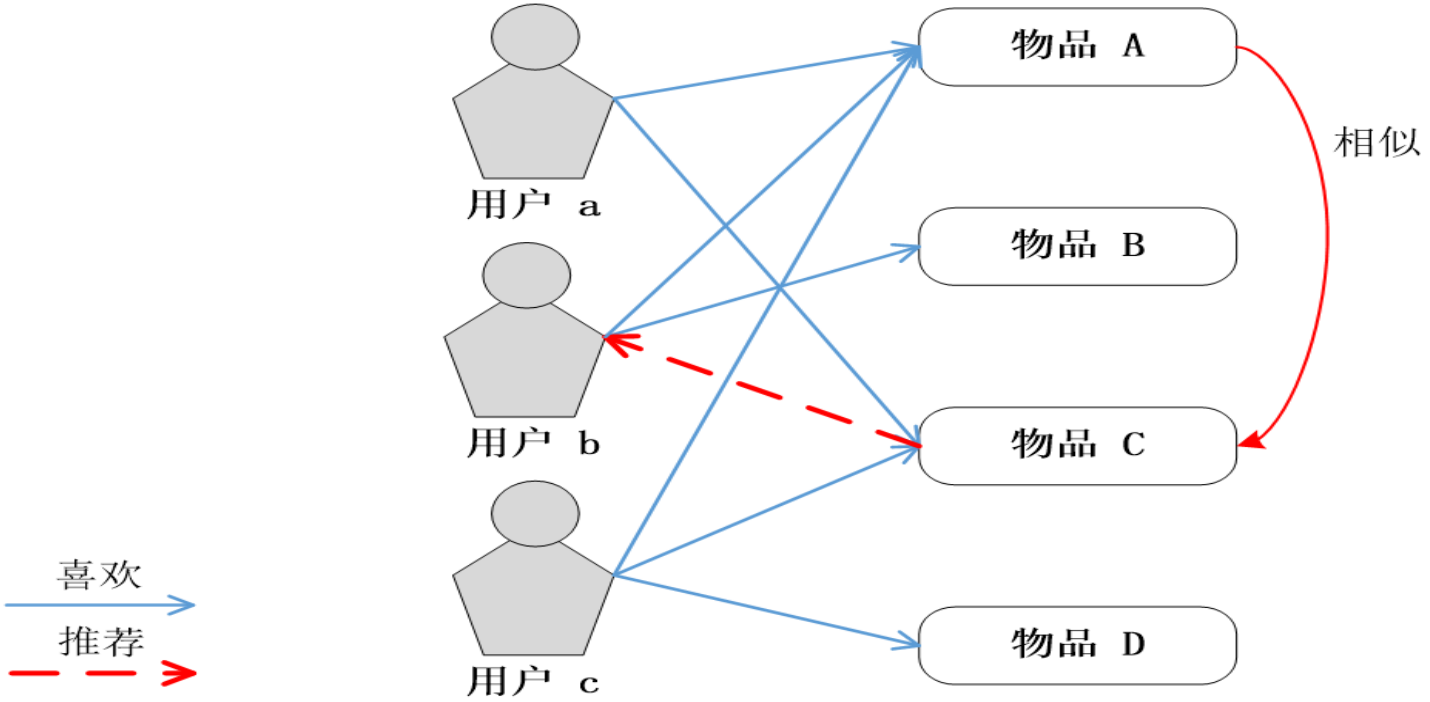


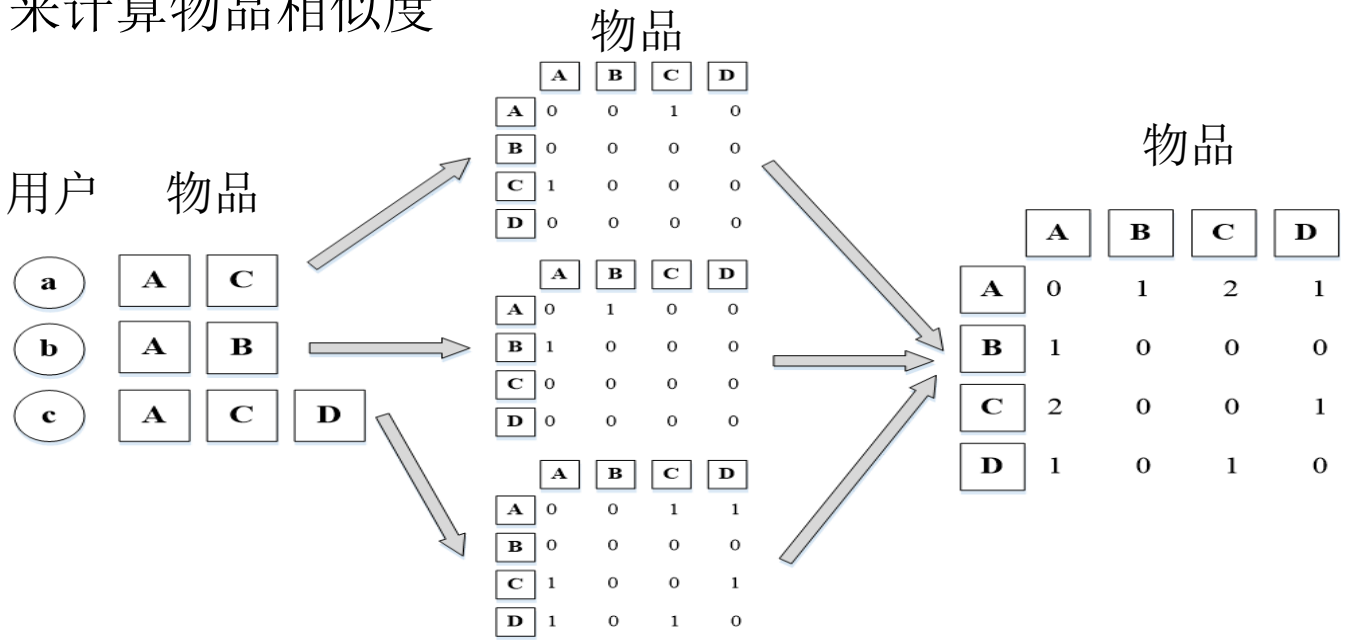
图11-6 基于物品的协同过滤（Item CF）

13.2.3 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

13.2.3 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11-7用户到物品倒排表及物品相似度矩阵

13.2.3 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$r_{vi}=1$

13.2.4 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

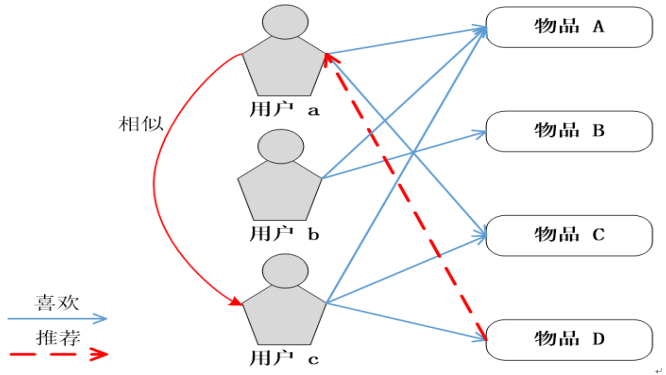


图11-4 基于用户的协同过滤（User CF）

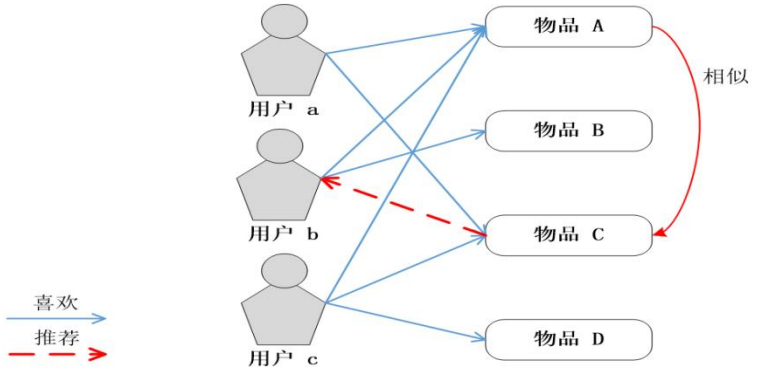


图11-6 基于物品的协同过滤（Item CF）

13.2.4 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

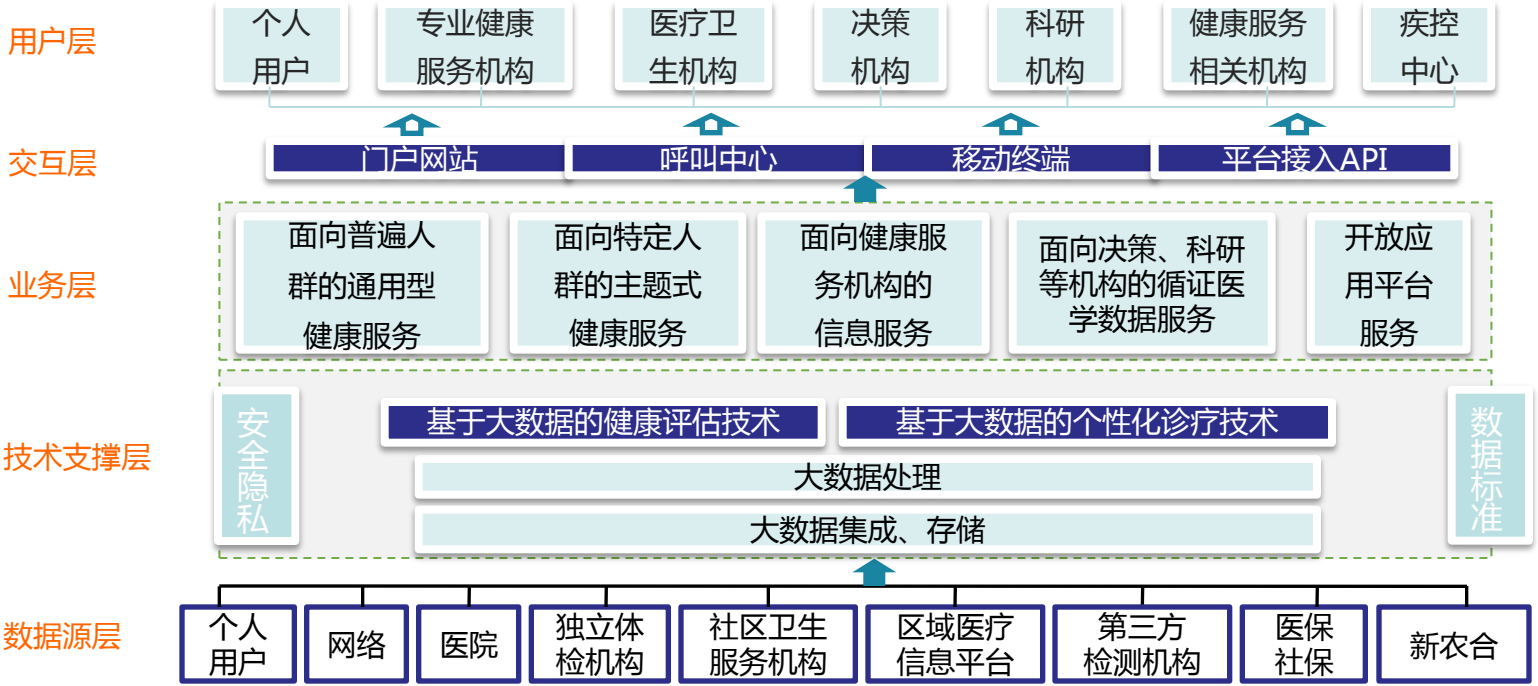
- ItemCF算法的推荐更偏向于个性化：适合应用于电子商务、电影、图书等应用场景，可以利用用户的历史行为给推荐结果作出解释，让用户更为信服推荐的效果
- ItemCF缺点：倾向于推荐与用户已购买商品相似的商品，往往会出现多样性不足、推荐新颖度较低的问题

13.3 大数据在智能医疗和智能物流领域的应用

- 13.3.1 基于大数据的综合健康服务平台
- 13.3.2 大数据在物流领域的应用

13.3.1 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。



13.3.2 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”将物流资源重组，欲将运力变得更集中、高效

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A：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。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年度和2017年度厦门大学教学类奖教金获得者。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)，主持的教改课题包括1项2016年福建省教改课题和1项2016年教育部产学协作育人项目，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500万字高价值的研究和教学资料，累计网络访问量超过500万次。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100万次。

附录B：大数据学习路线图

大数据学习路线图访问地址：http://dblab.xmu.edu.cn/post/10164/

附录C：《大数据技术原理与应[illegible]材

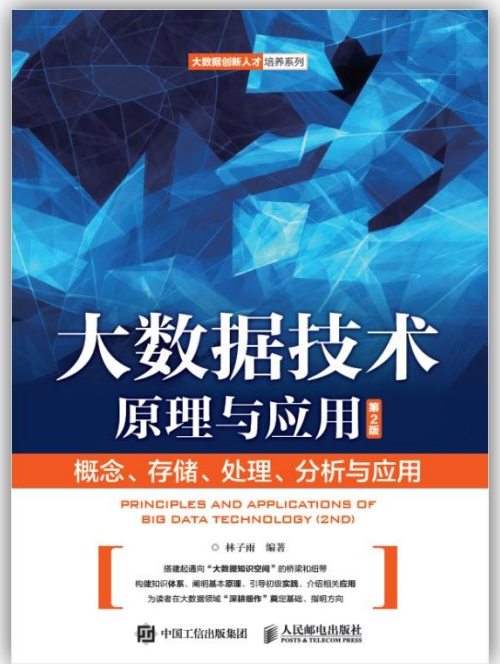
扫一扫访问教材官网

《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D：《大数据基础编程、实验和案例教程》

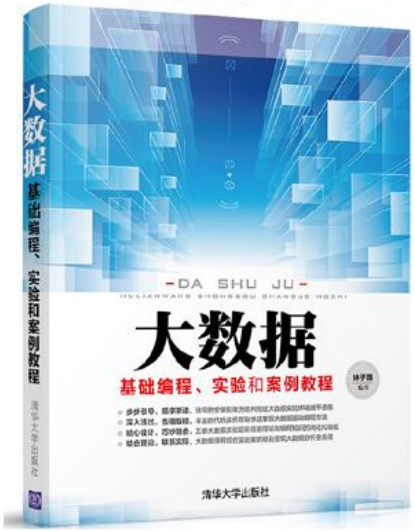
本书是与《大数据技术原理与应用（第2版）》教材配套的唯一指定实验指导书

大数据教材

1+1黄金组合

厦门大学林子雨编著

配套实验指导书



- 步步引导，循序渐进，详尽的安装指南为顺利搭建大数据实验环境铺平道路
- 深入浅出，去粗取精，丰富的代码实例帮助快速掌握大数据基础编程方法
- 精心设计，巧妙融合，五套大数据实验题目促进理论与编程知识的消化和吸收
- 结合理论，联系实际，大数据课程综合实验案例精彩呈现大数据分析全流程

清华大学出版社 ISBN:978-7-302-47209-4

附录E：《Spark编程基础》教材

《Spark编程基础》

厦门大学 林子雨，赖永炫，陶继平 编著

披荆斩棘，在大数据丛林中开辟学习捷径
填沟削坎，为快速学习Spark技术铺平道路
深入浅出，有效降低Spark技术学习门槛
资源全面，构建全方位一站式在线服务体系

人民邮电出版社出版发行，ISBN:978-7-115-47598-5

教材官网：http://dblab.xmu.edu.cn/post/spark/

授课视频： http://dblab.xmu.edu.cn/post/10482/

本书以Scala作为开发Spark应用程序的编程语言，系统介绍了Spark编程的基础知识。全书共8章，内容包括大数据技术概述、Scala语言基础、Spark的设计与运行原理、Spark环境搭建和使用方法、RDD编程、Spark SQL、Spark Streaming、Spark MLlib等。本书每个章节都安排了入门级的编程实践操作，以便读者更好地学习和掌握Spark编程方法。本书官网免费提供了全套的在线教学资源，包括讲义PPT、习题、源代码、软件、数据集、授课视频、上机实验指南等。

附录F：高校大数据课程公共服务平台

国内高校首个大数据课程公共服务平台，为全国高校教师和学生提供大数据教学资源一站式“免费”在线服务，包括课程教材、讲义PPT、课程习题、实验指南、学习指南、备课指南、授课视频、技术资料、实验案例、在线教程等，目前平台每年访问量超过100万次，成为全国高校大数据教学知名品牌

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!