《大数据技术原理与应用》

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11章 大数据在互联网领域的应用

（PPT版本号：2015年1月29日版本）

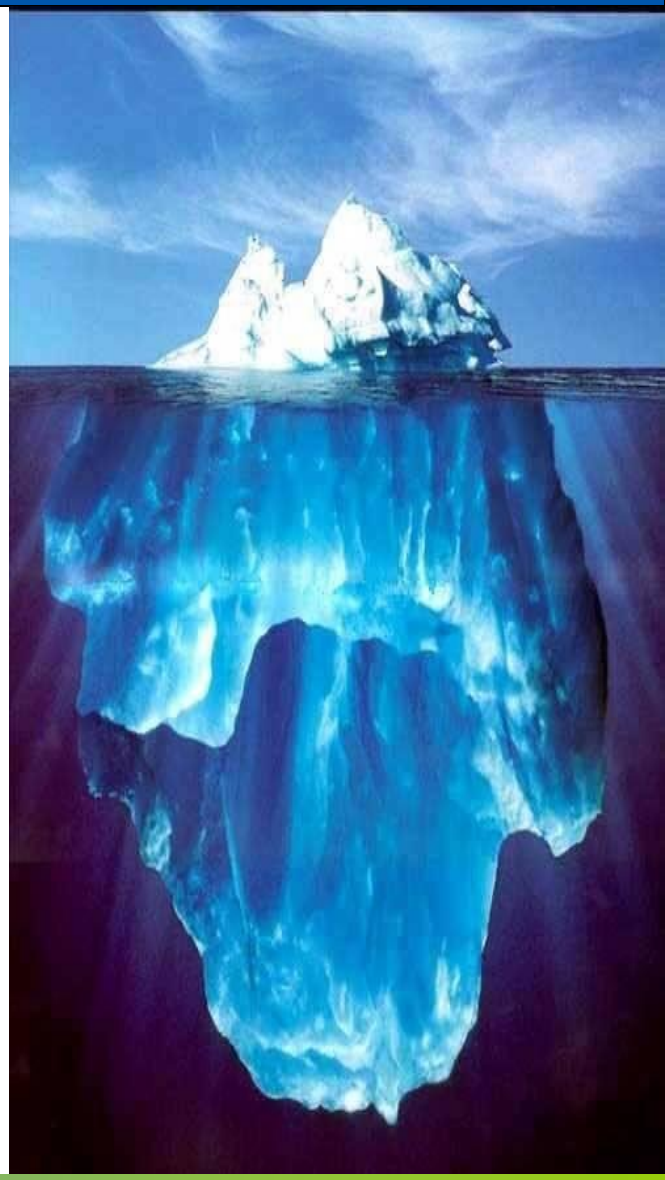
林子雨

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**

提纲



本PPT是如下教材的配套讲义：
21世纪高等教育计算机规划教材
《大数据技术原理与应用
——概念、存储、处理、分析与应用》
（2015年6月第1版）
厦门大学 林子雨 编著，人民邮电出版社
ISBN:978-7-115-39287-9

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

11.1 推荐系统概述

- 11.1.1 什么是推荐系统
- 11.1.2 长尾理论
- 11.1.3 推荐方法
- 11.1.4 推荐系统模型
- 11.1.5 推荐系统的应用

11.1.1 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

11.1.2 长尾理论

- “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

11.1.2 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
- 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

11.1.3 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

11.1.4 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

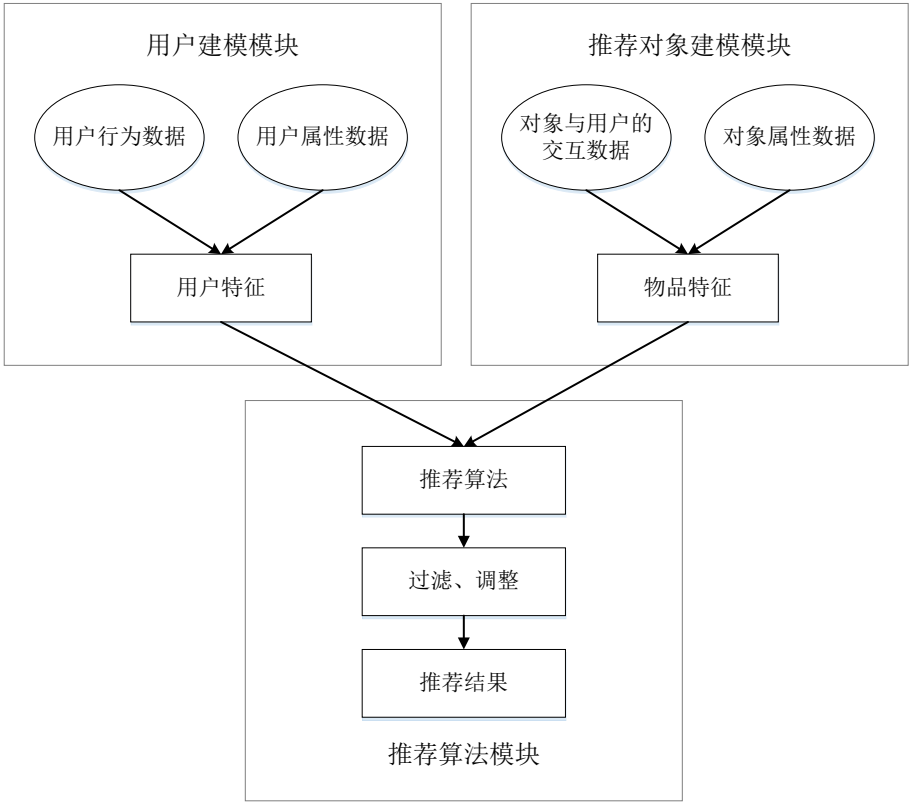


图11-1 推荐系统基本架构

11.1.5 推荐系统的应用

- 目前在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

11.1.5 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

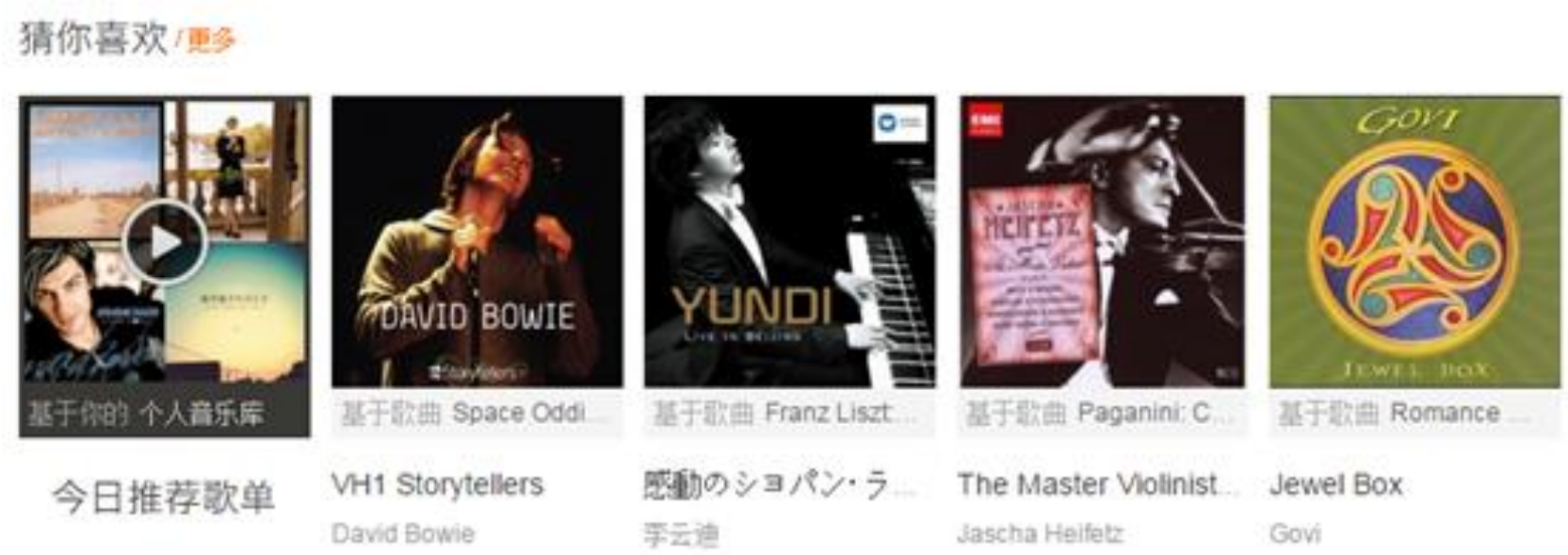


图11-3 虾米音乐网根据用户的音乐收藏来推荐歌曲

11.2 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

- 11.2.1 基于用户的协同过滤（UserCF）
- 11.2.2 基于物品的协同过滤（ItemCF）
- 11.2.3 UserCF算法和ItemCF算法的对比

11.2.1 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

11.2.1 基于用户的协同过滤（UserCF）

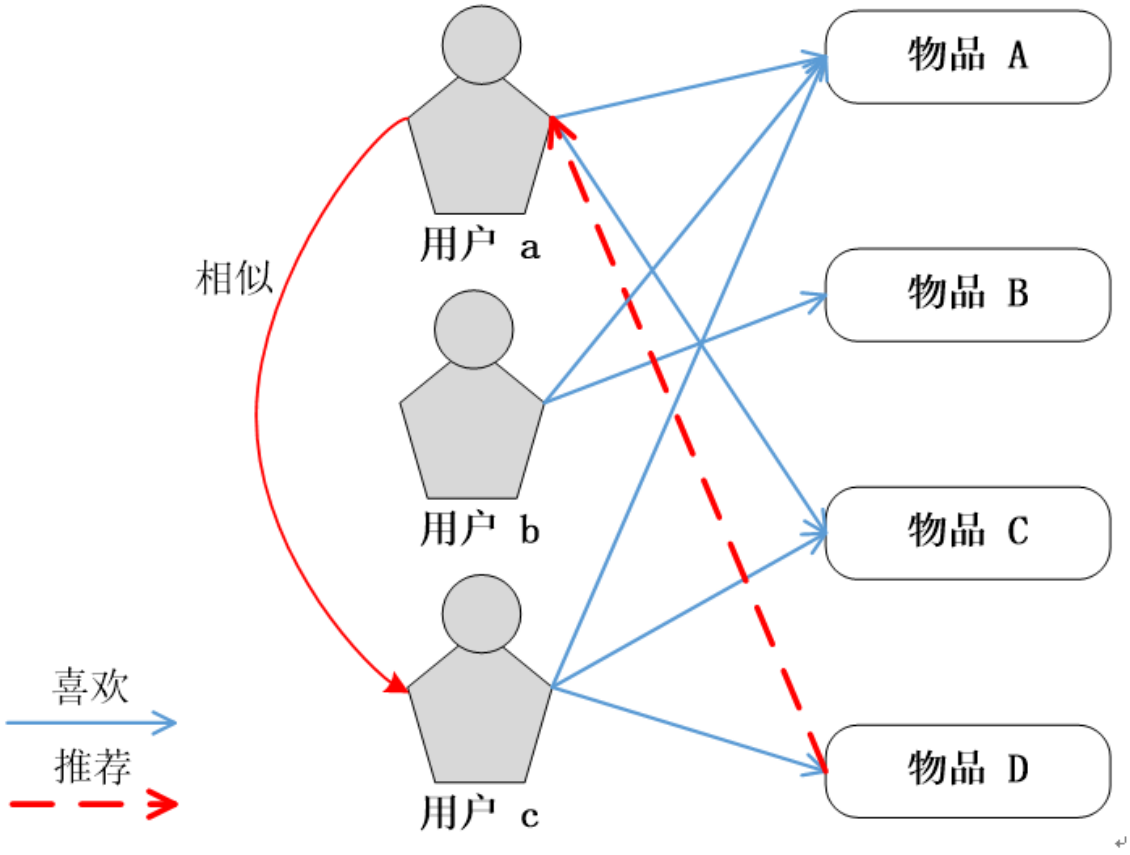


图11-4 基于用户的协同过滤（User CF）

11.2.1 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：

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

11.2.1 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

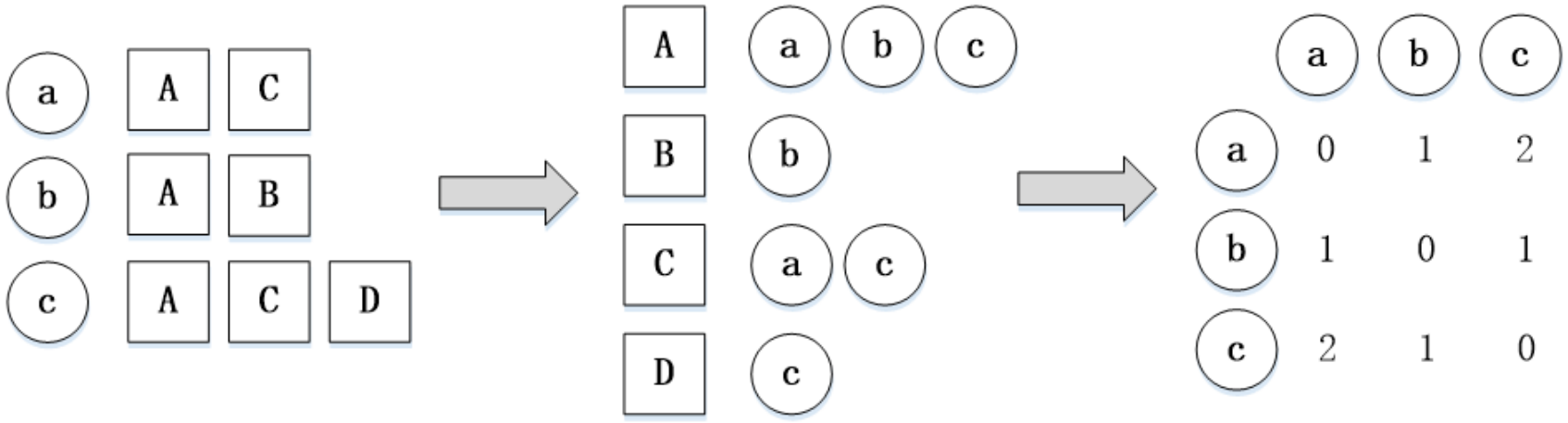


图11-5 物品到用户倒排表及用户相似度矩阵

11.2.1 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）

11.2.2 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》

11.2.2 基于物品的协同过滤（ItemCF）

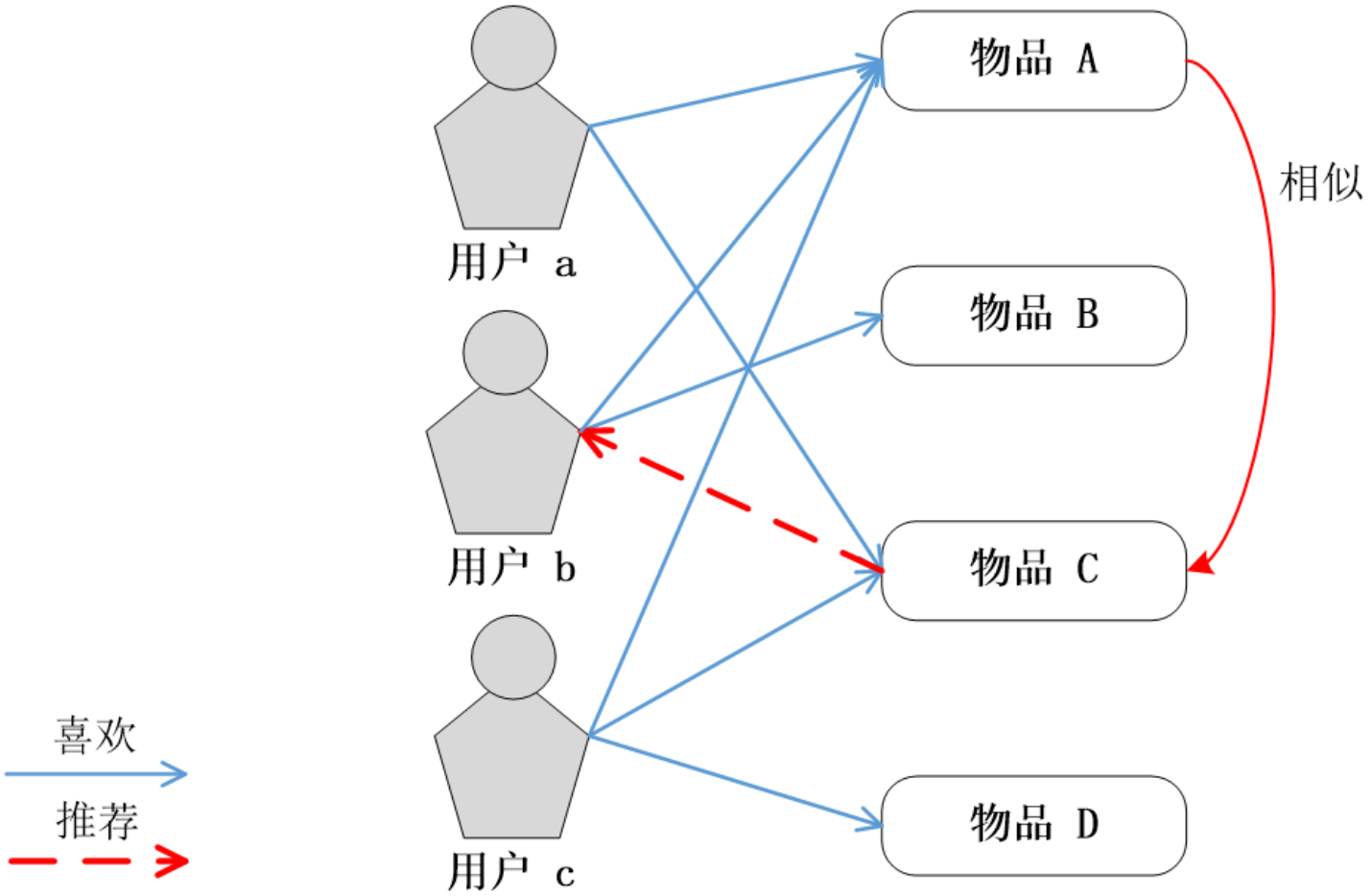


图11-6 基于物品的协同过滤（Item CF）

11.2.2 基于物品的协同过滤（ItemCF）

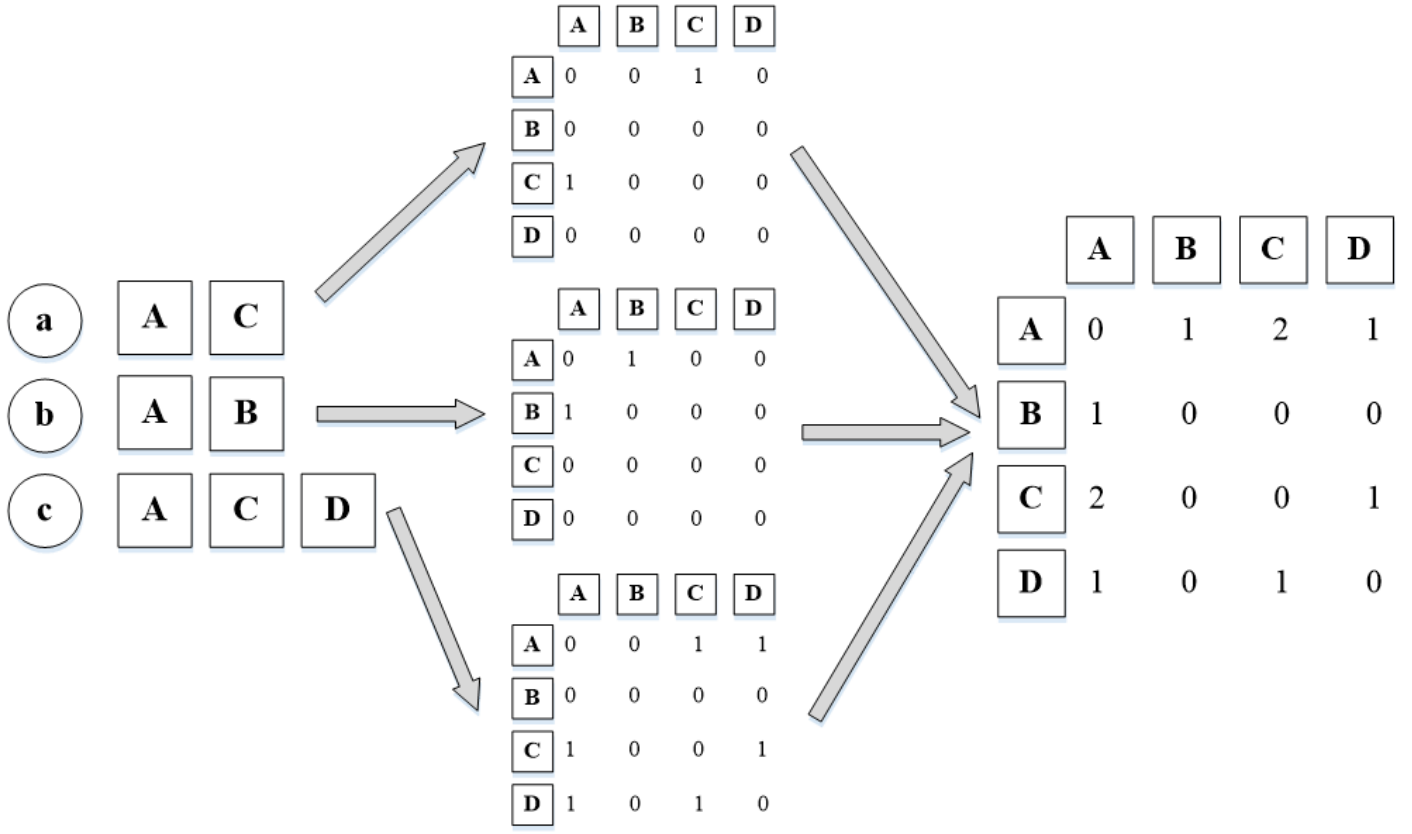
- ItemCF算法与UserCF算法类似，计算也分为两步：
  - 第一步：计算物品之间的相似度；
  - 第二步：根据物品的相似度和用户的历史行为，给用户生成推荐列表。

- 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)：

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

11.2.2 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11-7用户到物品倒排表及物品相似度矩阵

11.2.3 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

11.2.3 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

11.3 协同过滤实践

- 11.3.1 实践背景
- 11.3.2 数据处理
- 11.3.3 计算相似度矩阵
- 11.3.4 计算推荐结果
- 11.3.5 展示推荐结果

11.3.1 实践背景

- 我们选择以MovieLens公开数据集作为实验数据，采用ItemCF算法，使用Python语言来实现一个简易的电影推荐系统
- 具体采用的MovieLens 100k 数据集包括了1000名用户对1700部电影的评分记录，每个用户都至少对20部电影进行过评分，一共有100000条电影评分记录
- 基于这个数据集，我们解决的是一个评分预测问题，即如何通过已知的用户评分记录来预测未知的用户评分
- 对于用户未进行评分的电影，我们希望能够预测出一个评分，而这个评分反过来也可以用于猜测用户是否会喜欢这部电影，从而决定是否给用户推荐该电影

11.3.2 实践数据

- 用户对电影评分的数据格式如下，包含了用户ID、电影ID、评分、评分时间戳
- 通过评分数据，我们便可以采用如余弦相似度来计算用户之间的相似度

用户ID	电影ID	评分	评分时间戳
196	242	3	881250949
186	302	3	891717742
22	377	1	878887116
244	51	2	880606923
166	346	1	886397596
298	474	4	884182806
115	265	2	881171488
253	465	5	891628467
305	451	3	886324817
6	86	3	883603013
……			

图11-8 用户对电影的评分数据

11.3.3 实践流程

- 具体实现流程如下（具体代码见教材）：

1. 预处理：读取数据，提取评分
2. 计算相似度：使用余弦相似度计算电影间的相似度
3. 计算推荐结果：针对目标用户，对该用户未评分的电影计算预测评分
4. 展示推荐结果：对计算的评分进行降序排序，取Top-N个结果，作为最终的推荐结果

11.3.3 实践流程

- 例如我们对用户ID为1的用户，取10个推荐结果如下：

```
film: Titanic (1997),      rating: 3.89483842578
film: Air Force One (1997),    rating: 3.86732319398
film: L.A. Confidential (1997),    rating: 3.8255074736
film: Winter Guest, The (1997),    rating: 3.77287384289
film: Postman, The (1997),     rating: 3.7726583083
film: Wag the Dog (1997),      rating: 3.76181477552
film: Fire Down Below (1997),      rating: 3.75018044393
film: Anna Karenina (1997),    rating: 3.74833946959
film: Leaving Las Vegas (1995),    rating: 3.74818188227
film: Shadow Conspiracy (1997),    rating: 3.7351481072
```

图11-10 推荐结果

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
- 本章最后通过一个具体的实例，介绍了如何使用Python语言实现一个简易的电影推荐系统，深化对推荐系统的认识

附录：主讲教师

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，编著出版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》并成为畅销书籍，编著并免费网络发布40余万字中国高校第一本闪存数据库研究专著《闪存数据库概念与技术》；主讲厦门大学计算机系本科生课程《数据库系统原理》和研究生课程《分布式数据库》《大数据技术基础》。具有丰富的政府和企业信息化培训经验，曾先后给中国移动通信集团公司、福州马尾区政府、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、福建龙岩卷烟厂等多家单位和企业开展信息化培训，累计培训人数达2000人以上。

附录：大数据学习教材推荐

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》，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13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、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

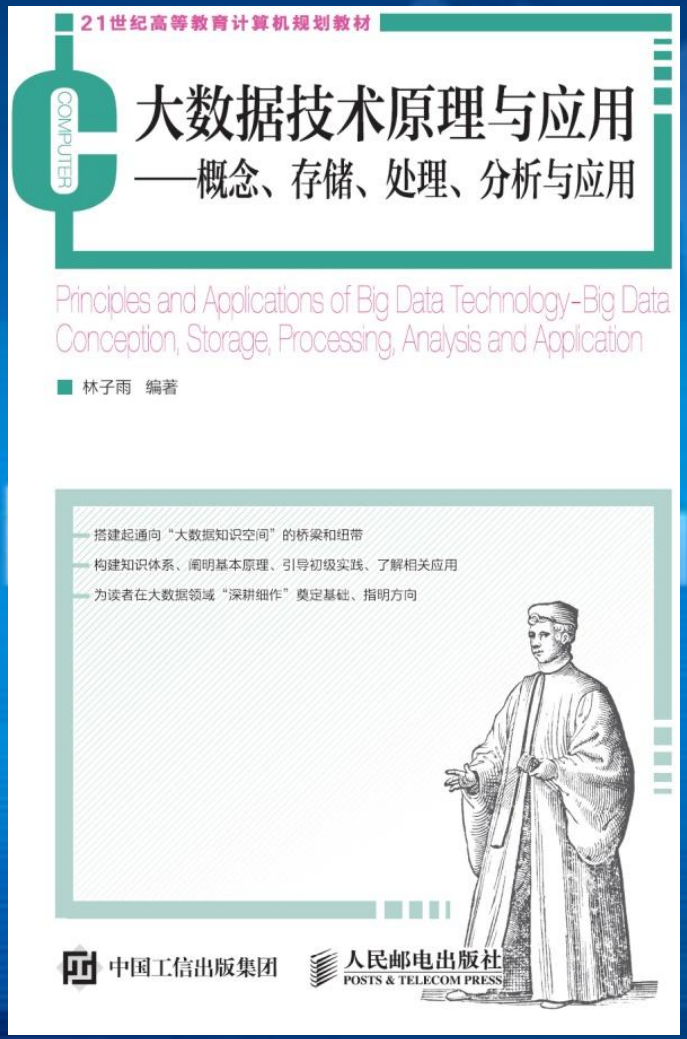
附录：中国高校大数据课程公共服务平台

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



Department of Computer Science, Xiamen University, 2016