6-3-cn
1 本节我们讨论推荐系统
2
数据处理系统提供大数据计算处理能力和应用开发平台。从计算架构的角度，将数据处理系统分为数据算法层、计算模型层、计算平台层、计算引擎层等。
与大数据相关的计算算法包括机器学习算法和数据挖掘算法。计算模型是指不同类型的大数据在不同场景下的处理方式，包括批处理、流计算、结构化数据
的大规模并发处理(MPP)模型、内存计算模型和数据流图模型。
计算平台和引擎提供各种开发套件和操作环境，我们选取 Spark Mllib 和 TensorFlow 为例
数据应用系统我们以推荐系统和社交网络分析为例进行讲解，本节我们讨论推荐系统
3
推荐系统是信息过滤系统的一个子类，它试图预测用户对某项物品的评级或偏好。
简单地说，就是向用户推荐相关物品的算法。推荐系统被广泛应用于许多领域，如 Netflix，亚马逊，京东，淘宝，QQ 音乐等。
推荐算法大致可以分为三类。协同过滤，基于内容和基于知识的过滤算法。而协同过滤算法和基于内容的算法最大的区别在于，基于内容的算法依赖于物品本身的特征，而协同过滤依赖于其他用户对同一物品的反应。
协同过滤可以进一步分为基于邻域和基于模型，基于邻域的算法包括基于用户和基于物品的算法，我们将在这一节讨论。基于模型的算法包括隐语义模型、图模型等。在隐语义模型中，我将以矩阵分解为例进行说明。
基于内容的推荐算法基于项目特征对项目进行推荐，包括结构化特征和非结构化特征。
第三类是基于知识的算法。
4 协同过滤和基于内容的过滤
让我们看一下这个图，
左边是协同过滤，以基于用户的过滤算法为例。
这表明，相似的用户会对文章有相似的偏好，
相似的用户是根据他们阅读的共同文章(绿色和红色)定义的。
因为他们是相似的用户，所以女读者读过的文章(蓝色的文章)可以推荐给男读者。
右边是基于内容的过滤推荐，红色文章和绿色文章根据文章的特点定义为相似的文章，当红色文章被用户阅读时，绿色文章就会被推荐给同一用户。
5 基于用户的协同过滤和基于物品的协同过滤
这张图展示了基于用户的过滤和基于项目的过滤。
在基于用户的过滤中，我们根据他们之前选择的共同物品找到相似的用户，
这里蒂姆选择了蛋卷冰淇淋、巧克力、蛋筒冰淇淋和甜甜圈，
John 选择了巧克力冰淇淋蛋筒，Tim 和 John 有一些共同的东西，那么我们定义他们为相似的用户。
对于 Tim 选择的和 John 没有选择的，我们可以推荐给 John，这里有蛋卷和甜甜圈的冰淇淋。
在基于物品的过滤中，我们根据用户选择的行为将带有蛋卷冰淇淋和蛋筒冰淇淋定义为相似的项目，
约翰选了蛋卷冰淇淋，所以我们向他推荐蛋卷冰淇淋。
我们将详细解释基于用户的过滤和基于物品的过滤。
6 在这节课中，我们将解释 3 种主要的推荐系统算法，
基于用户的协同过滤、基于物品的协同过滤和基于内容的过滤。
7 基于用户的协同过滤计算步骤
首先让我们看看基于用户的过滤推荐系统算法。
基于用户的过滤步骤如图所示，
1）我们收集用户购买记录，
2）根据用户购买记录构建物品-用户的表。

3）基于物品到用户表构建用户-用户矩阵;
4）基于 User- User 矩阵构建了 User-User 相似度矩阵;
5）然后计算用户对不同物品的兴趣程度。
6）根据用户对不同商品的兴趣程度获得推荐结果。

8 基于用户的协同过滤 相似度计算

基于用户的协同过滤是一种算法框架，根据与当前用户的相似度来识别相似用户，然后根据相似用户的评分对物品进行评分，然后由推荐系统根据物品的评分对物品进行推荐。

多个用户选择许多相同的物品时，我们可以称这些用户为相似用户。

在基于用户的协同过滤算法中，我们认为:一个用户会以更高的概率喜欢他或她的相似用户喜欢的物品。

如何计算用户的相似度？

通常使用 Jaccard 相似度、余弦相似度、欧氏距离和皮尔逊距离来计算两个用户之间的相似度。

设 N(u)是用户 u 喜欢的物品集合，N(v)是用户 v 喜欢的物品集合。

这里给出了 Jaccard 相似度、余弦相似度的计算方法。

9 基于用户的协同过滤如何获得相似用户

让我们以使用基于用户的协同过滤来获得推荐结果为例。

这里大写的 ABCD 是用户，小写的 abcde 是项目，

基于用户购买记录，我们构建了物品-用户表，该表显示了每个物品的所有购买过该物品的用户。

例如，商品 a 由用户 A 和用户 B 购买，商品 b 由用户 A 和用户 C 购买，等等。

然后基于物品到用户表构建用户-用户矩阵，每个元素值都是行用户和列用户购买的共同物品的数量。

例如，第一行和第二列的元素是由用户 A 和 B 的公共物品数计算的，他们只有一个公共物品，就是物品 a，所以值为 1。

10 基于用户的协同过滤 如何获得相似用户

基于我们刚刚计算的用户-用户矩阵，使用余弦相似度，我们可以计算用户-用户相似度矩阵，

这个矩阵中的每个元素都是对应行中的用户和对应列中的用户的相似度。（这里以余弦相似度为例）

这里 N(u)是用户 u 喜欢的物品的集合，N(v)是用户 v 喜欢的物品的集合。

11 基于用户的协同过滤 如何获得用户对不同物品的兴趣程度

在我们有了用户-用户相似度矩阵之后，

对于每个候选项 i，用户 u 对其感兴趣的程度 P(u,i)可以用上述公式计算。

r 表示的是用户 v 对商品 i 的兴趣，因为我们使用的是单一行为的隐反馈数据，因此在这里 r=1

12 基于用户的协同过滤 如何获得推荐结果

假设我们想向用户 A 推荐物品，

select K = 3 个相似用户，

根据用户-用户相似度矩阵，相似用户为 B、C 和 D

那么 B C D 选择而 A 没有选择的项目是 c 和 e，

然后分别计算用户 A 对物品 c 的兴趣，P (A, c)和用户 A 对物品 e 的兴趣，P (A, e) ，

它们都是 0.7416，这意味着用户 A 喜欢 c 和 e 一样。

13 现在我们看一下基于物品的协同过滤

14 基于物品的协同过滤

计算物品之间的相似度，推荐与用户选择物品相似度高的其他物品。

在基于物品的协同过滤算法中，我们认为:物品 a 和物品 b 的相似度是基于喜欢物品 a 的用户也喜欢物品 b。计算也是基于用户购买记录。

在此基础上计算物品-物品矩阵。在物品-物品矩阵中，每个元素都是既选择相应行中物品又相应列中物品的用户的数量。

例如，在右边的物品-物品 Item - Item 矩阵中，第一行第二列的元素的值，同时选择 a 和 b 的用户的数量是 1，只有用户 A。

矩阵中其他元素也可以用同样的方法计算得出。

15 基于物品的协同过滤 相似度计算
设N(i)是喜欢物品i的用户的集合，N(j)是喜欢物品j的用户的集合，
N(i)和N(j)的交集为同时喜欢物品i和物品j的用户数量。
物品i和物品j的余弦相似度，Wij可计算为:
N(i)和N(j)的交集的个数除以 (N(i))绝对值和 (N(j))绝对值的乘积开方。

16 基于物品的协同过滤 相似度计算
根据左侧矩阵所示的物品-物品矩阵，利用余弦相似度计算出物品-物品相似度矩阵。
17 基于物品的协同过滤 如何计算用户对不同物品的兴趣程度
对于每个候选物品j，用户U对物品j感兴趣的程度，Puj可以按照如下公式计算:
其中，N(u)表示用户U喜欢的商品的集合，S(j, k)是与商品j最相似的k个商品的集合，
Wij为商品i和商品j的相似度，Rui表示用户U对商品i的兴趣。

18 基于物品的协同过滤 如何得到推荐结果
现在我们可以得到推荐结果。
假设我们想向用户B推荐商品，选择K = 3个相似的商品，
所以用户B之前没有选择的项目是B d和e。
然后分别计算用户B 对物品b, d 和 e 的感兴趣程度P (B, d)， P (B, b)和P (B, e)，分别得到 1 ,0.5, 0.5，
那么我们认为用户B最喜欢d，用户B同样喜欢b和e

19 比较基于用户和基于物品的协同过滤算法
对于基于用户的过滤
如何根据用户来计算用户对候选物品的兴趣程度?
我们需要在用户和候选物品之间建立一个连接，这个连接就是用户相似度。
例如左边，我们计算用户A对商品c的兴趣程度P (A, c)，用户已经购买的商品是B和D，所以加上W.AB和W.AD的相似度。
以及基于物品的过滤
如何根据物品计算用户对候选物品的兴趣程度?
我们仍然需要在用户和候选物品之间建立一个连接，这个连接就是物品的相似性。
例如在右边，我们计算用户B对商品d的兴趣程度P (B, d)，已经购买的商品是商品a和商品c，所以加上W.da和W.dc的相似度。

20 现在我们看下基于内容的过滤
21 基于内容的过滤推荐，是最早的推荐算法，它推荐与用户过去喜欢的项目相似的项目。
这里的关键是物品相似度的度量，这是算法应用过程的核心。
三个主要步骤:
项目表示:为每个项目提取一些特征(内容)来表示它们
档案学习: 使用用户过去最喜欢(或不喜欢)物品的特征数据来学习用户的偏好特征(即用户档案)
生成推荐:通过将上一步获得的用户配置文件与候选项目的特征进行比较，将相关性最高的一组项目推荐给该用户。
22 本节学习了推荐系统的三个主要算法，基于用户的协同过滤和基于物品的协同过滤，和基于内容的过滤算法，今天的学习就到这里，谢谢大家