1. Introduction

从协同过滤的特点和挑战、基于Memory的CF、基于Model的CF、混合CF（Hybrid CF）、评价指标对CF展开介绍。

CF的挑战：高度离散的数据、指数级增长的用户和物品数量（用户-物品矩阵的爆炸性增长）、需要在极短的时间内完成推荐、同义词处理、数据噪声等等

Memory-base CF：根据用户的评分计算用户或物品之间的相似度并做出预测和推荐，实现简单。

Model-base CF：基于模型的CF通常使用机器学习算法来构建预测模型，例如矩阵分解、神经网络、决策树等。这些算法可以通过学习评分数据中的模式和关系来预测新用户对新物品的评分。与基于记忆的方法相比，基于模型的方法通常具有更好的预测性能和可扩展性，但需要更多的计算资源和时间来训练模型。

除了协同过滤外，基于内容过滤的推荐方法也非常重要。简单介绍了collaborative filter和content-base recommender的异同

协同过滤：是基于用户历史行为数据的，它会分析用户对物品的评分或行为数据，然后找到与之相似的用户，将推荐给他们喜欢的物品。

基于内容过滤的推荐系统：是基于物品本身的特征进行推荐，它会分析物品的属性和特征，然后将与用户历史行为相似的物品推荐给他们。

这两种方法在推荐系统中都有自己的优势和不足。协同过滤可以在没有物品数据的情况下进行推荐，但需要足够的用户评分数据。而基于内容的推荐系统可以在没有足够用户评分数据时进行推荐，但需要准确的物品元数据。因此，在实际应用中，这两种方法通常会结合起来使用，以达到更好的效果。

2. Characteristics and Challenges of Collaborative Filtering

推荐系统为用户提供快速、精确的内容，从而吸引用户使用这个平台，转化为商品的购买率等效益。构建的推荐算法有以下特点和挑战：

2.1 数据稀疏

系统内通常有海量的内容和用户，因此会造成user-item矩阵极大且非常稀疏。稀疏的数据矩阵会导致很难找到足够相似的用户或物品进行推荐等问题：

冷启动问题：在有的文献中也被称为 new user problem or new item problem，当有新的用户或者新的物品内容加入系统后，由于新内容没有足够的数据，无法找到相似的数据，从而无法对新内容进行推荐。

低覆盖率问题：reduced coverage 当用户的评分数据非常少，而推荐内容的数量非常多时，就会出现reduced coverage rate的问题，推荐系统可能无法为这些用户生成推荐。

邻居传递问题：在推荐系统中，邻居传递是指当两个用户之间没有共同评分时，即使他们具有相似的品味，也无法被识别为相似用户。这可能会降低基于比较用户对的预测来生成推荐的推荐系统的有效性。

邻居传递问题通常是由于稀疏数据库中缺乏数据而引起的。为了解决这个问题，可以使用更复杂的算法来处理，例如基于物品的推荐系统，该系统使用物品的属性和特征来进行推荐，而不是仅仅依赖于用户之间的比较。

解决手段

降低矩阵纬度： 通过奇异值分解、主成分分析等数学手段去掉不重要的用户或物品信息。舍弃的信息可能会导致与推荐相关的信息也丢失了，降低了推荐质量

基于Hybrid CF的冷启动解决方案： 内容增强CF算法利用外部内容信息，有利于解决冷启动问。或通过结合协作信息和内容信息，提出冷启动的推荐方法。或将推荐内容分成不同的组，基于用户评分的高斯分布对用户推荐内容进行预测。

Model-based CF:TAN-ELR是一种基于贝叶斯网络和逻辑回归的分类器，通过树增强和扩展逻辑回归来提高模型的精度和效率。

基于矩阵分解的协同过滤算法： 将用户-物品评分矩阵分解成用户特征矩阵和物品特征矩阵，其中用户特征矩阵包含了用户的隐含特征向量，物品特征矩阵包含了物品的隐含特征向量。通过计算用户特征向量和物品特征向量的内积，就可以得到用户对物品的评分预测值。 常见的如 SVD、NMF、PMF、MMMF。MMMF最大化不同分类的边际间隔使得分类器具有更强的泛化性。

多重插补CF方法和插补增强CF算法则是利用插补方法来填补缺失数据，从而提高推荐系统的性能。

2.2 Scalability

当现有用户和项目的数量急剧增长时，传统的CF算法将遭受严重的可伸缩性问题。例如，拥有数千万个客户(M)和数百万个不同的目录项(N)，一个复杂度为O (n)的CF算法太大了。此外，推荐系统要有极高的实时性，并向所有用户提供推荐数据，无论他们的行为数据历史如何，这要求推荐系统的高可伸缩性。

Model-based CF的相关工作提供了一些有效的方案：item-based Person correlation CF：不计算所有项目对之间的相似度，而是仅计算用户共同评价的项目对之间的相似度;简单的贝叶斯CF算法通过基于观察到的用户评分行为进行预测；聚类CF算法：通过在更小和高度相似的集群中而不是在整个数据库中寻求用户的推荐来解决可伸缩性问题；

2.3 Synonymy

同义词指的是许多相同或非常相似的项目具有不同的名称或条目。推荐系统应该发现词意之间的关联，从而能够推荐类似的内容。这块涉及NLP相关内容。

3. Memory-Based Collaborative Filtering Techniques

基于内存的CF算法使用用户-物品数据库的整个或一个示例来生成预测。每个用户都是一群有相似兴趣的人中的一部分。通过计算一个新用户（或活跃用户）的邻居，可以产生对他或她的新项目的偏好的预测。

The neighborhood-based CF algorithm是一种基于邻域的协同过滤算法，用于推荐系统中的个性化推荐。在这种算法中，用户和物品之间的相似度是通过计算它们之间的距离或相似性来确定的。然后，根据用户之前的评分和相似物品的评分，为用户推荐最相关的物品。这种算法通常用于处理大型数据集，因为它不需要计算所有用户和物品之间的相似度，而只需要计算与目标用户或物品最相似的一小组邻居。因此，它可以提高推荐系统的效率和准确性。

在实际应用中，通常使用基于余弦相似度或皮尔逊相关系数等技术来计算用户和物品之间的相似度。此外，还可以使用KNN（K-Nearest Neighbor）算法来确定与目标用户或物品最相似的邻居。

3.1 Similarity Computation

3.1.1 Correlation-Based Similarity

基于Pearson correlation：

Pearson相关系数（Pearson correlation coefficient）是一种衡量两个变量之间线性关系强度的统计量。它用于衡量两个连续变量之间的相关性，取值范围在-1和1之间，其中0表示没有线性相关性，正值表示正相关性，负值表示负相关性。Pearson相关系数可以用来衡量两个变量之间的线性关系，但并不能说明它们之间是否存在其他类型的关系。Pearson相关系数通常用于探索两个变量之间的关系，例如在数据挖掘、机器学习和统计分析中。它可以帮助研究人员确定两个变量之间是否存在相关性，并在建立预测模型时帮助选择最相关的变量。但需要注意的是，Pearson相关系数只能衡量线性关系，对于非线性关系则不适用。

Pearson相关系数的扩展算法有：constrained Pearson correlation、Spearman rank correlation

后两者：都是通过比较两个变量之间的等级顺序来计算的，而不是直接比较它们的数值。用来衡量两个变量之间的单调关系，即它们是否在一起变化，但不能说明它们之间是否存在其他类型的关系。它们通常用于探索两个变量之间的关系，不需要假设数据呈正态分布或线性相关。

3.1.2. Vector Cosine-Based Similarity.

Adjusted cosine similarity：向量余弦相似度计算两个物品之间的夹角，而调整后的余弦相似度则考虑了不同用户之间使用不同评分尺度的问题。由于不同用户可能使用不同的评分尺度，因此向量余弦相似度无法考虑这种差异。为了解决这个问题，可以使用调整后的余弦相似度。该方法从每个用户的评分中减去该用户的平均评分，然后计算物品之间的余弦相似度。调整后的余弦相似度与Pearson相关系数具有相同的公式，但是它对用户的评分进行了标准化，以纠正评分尺度的差异。

3.2 Prediction and Recommendation Computation

在邻域方法中，预测过程主要包括两个步骤：选择邻居和计算预测评分。通过这两个步骤，可以为当前用户生成个性化的推荐结果。

3.2.1 Weighted Sum of Others’Ratings 加权邻居评分和

首先需要找到与当前用户最相似的一组用户，这些用户被称为邻居。然后，计算每个邻居对当前用户未评价物品的预测评分。

具体而言，对于每个邻居，可以使用其评价过的物品与当前用户未评价物品之间的相似度来计算其对当前用户未评价物品的影响权重。然后将每个邻居的影响权重与其对当前用户未评价物品的评分相乘，得到每个邻居对当前用户未评价物品的加权评分。最后将所有邻居的加权评分相加，得到当前用户对未评价物品的预测评分。这种方法相比于简单平均或加权平均方法，更加准确地反映了邻居之间的差异性。同时，由于使用了邻居之间的相似度来计算权重，因此可以更好地处理邻居之间的差异性。

3.2.2. Simple Weighted Average简单加权平均

对于item-based的预测，用户u对i的评分如下：

其中r{u,n},n∈N为用户n对其他物品n的评级，w{i,n}为物品i 和n的相似度。

3.3. Top-N Recommendations

3.3.1 User-Based Top-N Recommendation Algorithms

用户基础的Top-N推荐算法是一种通过分析用户行为数据来为特定用户推荐最有可能引起其兴趣的N个物品的技术。这种算法在电子商务、社交网络和音乐视频等应用中得到了广泛应用，并且随着深度学习技术的发展，其研究和实践也取得了显著进展。

基本原理:用户基础的Top-N推荐算法主要基于协同过滤（Collaborative Filtering, CF）技术。协同过滤算法可以分为两大类：基于邻域的（neighborhood-based）和基于模型的（model-based）。基于邻域的协同过滤算法通过找到与目标用户兴趣相似的其他用户，然后将这些用户喜欢的物品推荐给目标用户。而基于模型的协同过滤算法则通过建立一个全局的用户-物品评分矩阵，利用矩阵分解技术（如奇异值分解，SVD）来预测用户对未知物品的评分。

3.3.2. Item-Based Top-N Recommendation Algorithms

基于物品的Top-N推荐算法首先根据物品之间的相似度计算出每个物品的k个最相似物品；然后通过取k个最相似物品的并集并移除用户已经购买的物品集合U中的每个物品，来确定推荐物品的候选集合C；最后计算集合C中每个物品与集合U中物品之间的相似度。排序后得到的C集合中的物品，按照相似度降序排列，即为推荐的基于物品的Top-N列表。这种方法的一个问题是，当一组物品的联合分布与该组中各个单独物品的分布不同时，上述方案可能会产生次优的推荐结果。为了解决这个问题，Deshpande和Karypis 开发了更高阶的基于物品的Top-N推荐算法，当确定要向用户推荐的项目集时，使用所有项组合，直到特定大小为止。

3.4. Extensions to Memory-Based Algorithms

3.4.1. Default Voting

在许多协作过滤算法中，仅从用户都评过分的物品的交集中计算成对相似度[5, 27]。当评价数太少以至于无法生成相似度值时，这种方法将不可靠。此外，只关注交集相似度会忽略反映用户整个评分历史中的全局评分行为。

相关研究通过假设一些默认的投票值来填补缺失的评分可以提高CF预测性能。Herlocker等人通过减少具有少于50个共同项目的用户的权重来解决小交集问题。Chee等人使用团体（或小组）的平均值作为默认投票，以扩展每个用户的评分历史。Breese等人对未观察到的评分使用中性或略为负面的偏好，然后在生成的评分数据上计算用户之间的相似度。

3.4.2. Inverse User Frequency (IUF)

用于调整用户评分矩阵中的权重，以提高推荐性能。IUF的基本思想是，普遍出现的物品对于区分用户之间的偏好不如罕见的物品重要。因此，IUF通过对用户评分矩阵中的每个物品进行加权来调整物品的重要性，其中权重f\_j=log(n/n\_j)与使用该物品的用户数量成反比，n\_jn为对j物品进行评分的用户总数。这意味着罕见的物品将具有更高的权重，而普遍出现的物品将具有较低的权重。

3.4.3. Case Amplification

Case Amplification通过对权重进行变换来增强相似度高的用户之间的关系，并减弱相似度低的用户之间的关系。这可以通过将权重进行指数变换来实现，其中较高的权重将得到更大的增强，而较低的权重将得到更大的惩罚。

其中ρ \rhoρ是Case Amplification的放大系数，ρ≥1，典型的ρ值为2.5 。Case Amplification可以减少数据中的噪声。它倾向于支持高权重，因为小值的幂次方变得可以忽略。例如，如果权重很高，wi,j = 0.9，则它仍然很高（ 0.9^{2.5} ≈ 0.8）；如果权重很低，例如wi,j = 0.1，则它将是可以忽略的（ 0.1^{2.5} ≈ 0.003）。

3.4.4. Imputation-Boosted CF Algorithms (IBCF)

Imputation-Boosted CF Algorithm是一种基于物品的协同过滤推荐算法，它使用缺失值填充来提高预测性能。在这个算法中，预测用户对某个物品的评分是通过分析其他用户对类似物品的评分来实现的。与传统的协同过滤算法不同的是，Imputation-Boosted CF Algorithm使用缺失值填充来处理评分矩阵中的缺失值，以提高预测精度。

IBCF-NBM，结合了使用朴素贝叶斯的IBCF和使用均值填补的IBCF。这种混合算法被设计用于不同密度的数据集。对于密集的数据集，使用朴素贝叶斯的IBCF可以更好地处理数据。对于稀疏的数据集，使用均值填补的IBCF可以更好地处理数据。通过将两种算法组合起来，可以提高推荐算法的准确性。