# **Recommendation based on collaborative filtering and user tags**

Haipeng Wang Huang Zhou Cheng Wang Sijie Zhou

Wuhan Textile University

**Abstract**

*推荐系统已在多个电子商务应用中使用。推荐系统有三种类型: 基于内容的过滤、协同过滤和混合推荐系统。在本文中，使用包含100万分数的Movielens数据集评估了两种类型的协作过滤技术，即item-cf和user-cf。这两种类型是基于余弦相似度函数的矩阵分解和用户协同过滤。此外，在本文中，我们还采用了基于用户标签的推荐，使用hetrec2011数据集。*

# **Introduction**

推荐系统是用于预测用户对选项的响应的数据挖掘应用程序之一。推荐系统分类如下:

* 基于内容的过滤: 推荐与用户喜欢的项目内容相似的项目。
* 协同过滤: 推荐类似用户首选的项目。
* 混合方法: 以几种不同的方式结合基于内容和协作过滤方法。

## **Utility Matrix**

在推荐系统中，用户对项目的偏好称为评分。通常一个评级是整数。将用户映射到评级的实体称为效用矩阵。所有推荐系统都基于效用矩阵。在效用矩阵中，行表示用户，列表示项目。

有两种填充效用矩阵的方法。第一种方法是要求用户对项目进行评分。例如，Netflix要求用户对电影进行评分。几乎所有推荐系统的应用都使用这种方法。另一种方法是根据用户行为得出评级。这些派生的评级通常是布尔值。例如，在电影应用程序中，如果用户观看电影，我们可以假设用户喜欢该电影。但是我们不知道用户对电影的喜欢程度。这就是为什么基于用户行为的预测评级通常以布尔值给出。

推荐系统的目标是预测效用矩阵的缺失元素或空白。我们知道，通过询问用户的评分来填充效用矩阵。很明显，用户不会对所有项目进行评分。所以，每个用户比效用矩阵中的填充元素有更多的空白。推荐系统的目标是预测这些空白。在大多数情况下，预测用户将给予高度评价的空白就足够了。

## **Content Based Filtering (CBF)**

在基于内容的过滤中，基于用户c分配给与s “相似” 的项目的实用程序来估计用户c的项目的实用程序。换句话说，基于内容的过滤是推荐与用户偏好的项目相似的项目。基于内容的过滤包含三个实体: 实用程序矩阵，项目配置文件和用户配置文件。

## **Collaborative Filtering (CF)**

通常，协作过滤是使用涉及多个数据源之间协作的技术对信息或模式进行过滤的过程。

基于用户的协作过滤 (UBCF): 在基于用户的协作过滤中，基于实用程序由与用户U “相似” 的用户u分配给项目i。以下是为用户u推荐项目的基于用户的协同过滤算法:

对于每个其他用户w，计算u和w之间的相似度s，保留顶级用户，按相似度排序为邻域n；

对于n中的某个用户偏爱的每个项目i，但是u还没有偏爱；

对于n中每个对i具有偏好的用户v，将v对i的偏好合并到运行平均中，返回按加权平均排名的顶部项目。

根据上述算法，第一个for循环计算一组相似用户n，称为邻域。第二个for循环预测用户u尚未评级的项目的评级。为了测量两个用户之间的相似性，可以使用诸如余弦距离之类的距离度量。

邻域有两种类型: 固定大小的邻域和基于阈值的邻域。固定大小的邻域由n个最相似的用户组成。通常，当n增加时，推荐的准确性增加仅仅是因为有更多类似的用户。基于阈值的邻域由相似用户组成，其相似性度量大于值t，其中-1 <t <1。通常，当阈值增加时，推荐的准确性降低仅仅是因为太少的类似用户最终在附近。

矩阵分解 (MF): 在矩阵分解中，用户和项目都映射到潜在因子空间中，从而将用户-项目交互建模为该空间中的内积。每个项目i与向量 相关联，每个用户u与向量 相关联。用户u对项目i的评级由表示，其估计如下:

用余弦函数评估基于用户的协同过滤

# **Evaluating user based collaborative Filtering**

基于用户的协同过滤算法主要包含两个步骤：

a. 计算用户之间的相似度：找到和目标用户兴趣相似的用户集合。

b. 根据用户相似度及用户对物品的评价为物品打分：找到这个集合中的用户喜欢的，且目标用户没有听说过的物品推荐给目标用户。

计算用户相似度

根据协同过滤算法的定义，这里主要是利用用户行为的相似度来计算兴趣的相似度。给定用户u和用户v，令和分别表示用户u和用户v曾经有过正反馈的物品集合，则有如下三种方式计算相似度：

Jaccard公式

余弦相似度（UserCF算法）

改进的余弦相似度（UserIIF算法）

在UserIIF算法中，是物品i的热度，可见其对热门物品进行了惩罚，因为两个用户对冷门物品采取过同样的行为更能说明他们兴趣的相似度。

这里要强调一个工程实现上的Trick。在计算用户行为之间的相似度时，如果按照定义实现的话，需要对两两用户的行为集合进行统计，这样的时间复杂度为，但用户行为往往是十分稀疏的，很多用户之间的行为并没有交集，导致时间浪费在这些不必要的计算上。这时就可以建立Item-User的倒排表，这样在同一个Item下面的User两两之间一定是在这个Item上有交集的，所以只需要遍历所有的Item，对其下所有的User两两进行统计即可，这样可以极大降低时间复杂度。

# **Evaluating item based collaborative Filtering**

与基于用户的协同过滤算法一样，基于物品的协同过滤算法也是基于邻域的一种做法。它也可以分为两步：

a. 计算物品之间的相似度。

b. 根据物品的相似度和用户的历史行为为用户生成推荐列表。

计算物品相似度主要还是利用用户的行为数据，即比较对两个物品有过正反馈的用户集合的相似性。令N(i)为喜欢物品i的用户集合，则有如下几种相似度计算方法：

购买了该商品的用户也经常购买的其他商品

余弦相似度（ItemCF算法）

上面的公式在计算的时候会导致物品与热门物品的相似度都很高，因此可以加上物品j的热度惩罚项，变成了如下的余弦相似度的形式：

改进的余弦相似度（ItemIUF算法）

与UserIIF算法类似，这里也对热门用户进行了惩罚，即活跃用户对物品相似度的贡献应该小于不活跃的用户。

# **Evaluate user tag-based recommendations**

最简单的算法（SimpleTagBased）

在拿到用户标签行为数据之后，一个最容易想到的方法就是：

统计每个用户最常用的标签

对于每个标签，统计被打过这个标签次数最多的物品

对于每个用户，首先找到他常用的标签，然后找到具有这些标签的最热门物品进行推荐

用公式表示用户u对物品i的兴趣为，其中，是用户u打过标签b的次数，是物品i被打过标签b的次数。

利用TF-IDF进行改进（TagBasedTFIDF和TagBasedTFIDF++）

上面的公式倾向于给热门标签对应的热门物品很大的权重，从而不能反映用户个性化的兴趣，因此可以借鉴TF-IDF思想，对上述公式进行改进（TagBasedTFIDF）：

其中记录了标签b被多少个不同的用户使用过。可见这个公式是对热门标签进行惩罚，当然对于热门物品也可以进行同样的惩罚（TagBasedTFIDF++）：

其中记录了物品i被多少个不同的用户打过标签。

# **Conclusion**

推荐系统已在多个电子商务应用程序中使用，例如Amazon和Netflix。推荐系统有三种类型: 基于内容的过滤，协作过滤和混合技术。在本文中，使用包含100万评级的Movielens数据集评估了两种类型的协作过滤技术。这两种类型是矩阵分解和基于余弦相似性函数的基于用户的协作过滤。两种类型的评估基于完整数据集的RMSE和完整数据集的不同分区。分区取决于年龄，类型和时间。对于这两种技术，结果表明，完整集合的RMSE小于每个分区的RMSE。我们在分区中出现更多错误的原因是，一个分区中电影的属性取决于其他分区中电影的属性。在分区中，我们最大化一个分区的贡献，而忽略其他分区的贡献。因此，这种方法是减法的。我们需要的是一种加法方法，在这种方法中，我们最大化每个分区的贡献，同时保持其他分区的贡献。

# **References**

1. U. Panniello and M. Gorgoglione, “A contextual modeling approach to context-aware recommender systems,” in Proc. the 3rd Workshop on Context-Aware Recommender Systems, October 2013.
2. A. Rajaraman, J. Leskovec, and J. D. Ullman, “Recommendation systems,” in Mining of Massive Datasets, Cambridge University Press, 2012, pp. 303-323.
3. G. Badaro, H. Hajj, W. El-Hajj, and L. Nachman, “A hybrid approach with collaborative filtering for recommender systems,” in Proc. 9th International Wireless Communications and Mobile Computing Conference, 2013, pp. 349-354.
4. R. Burke, “Hybrid recommender systems: Survey and experiments,” User Modeling and User-Adapted Interaction, vol. 12, no. 4, pp. 331-370, November 2002.
5. J. Fan, W. Pan, and L. Jiang, “An improved collaborative filtering algorithm combining content-based algorithm and user activity,” in Proc. International Conference on Big Data and Smart Computing, 2014, pp. 88-91.
6. F. Fouss and M. Saerens, “Evaluating performance of recommender systems: An experimental comparison,” presented at the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2008.
7. H. Ji, J. Li, C. Ren, and M. He, “Hybrid collaborative filtering model for improved recommendation,” in Proc. IEEE International Conference on Service Operations and Logistics, and Informatics, 2013, pp. 142-145.
8. Roy Deepjyoti,Dutta Mala. An Improved Cat Swarm Search-Based Deep Ensemble Learning Model for Group Recommender Systems[J]. Journal of Information & Knowledge Management,2022,21(03).
9. Chen Yida,Qiu Xiaoyu,Ma Chuanjiang,Xu Yunfeng,Sun Yang. A recommender system fused with implicit social information through network representation learning[J]. Computers and Electrical Engineering,2022,100.
10. Fernández-García Antonio Jesús,Rodriguez-Echeverria Roberto,Preciado Juan Carlos,Perianez Jorge,Gutiérrez Juan D.. A hybrid multidimensional Recommender System for radio programs[J]. Expert Systems With Applications,2022,198.
11. Khoali Mohamed,Laaziz Yassin,Tali Abdelhak,Salaudeen Habeeb. A Survey of One Class E-Commerce Recommendation System Techniques[J]. Electronics,2022,11(6).
12. Horasan, Fahrettin. Latent Semantic Indexing-Based Hybrid Collaborative Filtering for Recommender Systems[J]. Arabian Journal for Science and Engineering,2022(prepublish).
13. Tavakoli Mohammadreza,Faraji Abdolali,Vrolijk Jarno,Molavi Mohammadreza,Mol Stefan T.,Kismihók Gábor. An AI-based open recommender system for personalized labor market driven education[J]. Advanced Engineering Informatics,2022,52.
14. Sumathy B.,Kumar Anand,Sungeetha D.,Hashmi Arshad,Saxena Ankur,Kumar Shukla Piyush,Nuagah Stephen Jeswinde. Machine Learning Technique to Detect and Classify Mental Illness on Social Media Using Lexicon-Based Recommender System[J]. Computational Intelligence and Neuroscience,2022,2022.