Hans Journal of Data Mining 数据挖掘, 2025, 15(2), 201-211
Published Online April 2025 in Hans. https://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2025.152017

Hans汉斯

基于用户评分数据的多维度电影推荐系统研究

余振洋，俞　婷，陈鑫磊，肖　炼

嘉兴南湖学院信息工程学院，浙江　嘉兴

收稿日期：2025年3月23日；录用日期：2025年4月16日；发布日期：2025年4月23日

摘　要

随着个性化推荐系统在各类应用中的广泛应用，电影推荐作为其中的典型场景，受到了越来越多的关注。推荐系统的核心任务是根据用户的历史行为数据，特别是评分数据，来为用户提供个性化的推荐。推荐效果的好坏与相似度计算方法的选择密切相关。常见的相似度计算方法包括余弦相似度、皮尔逊相似度、欧几里得距离和Jaccard相似度。每种方法有其独特的特点和适用场景，但单一使用某种相似度方法往往会受到数据特性和环境的限制，导致推荐性能的下降。本文通过系统地比较和分析这四种相似度计算方法在不同环境下的表现，探讨了它们在电影推荐中的应用效果。研究表明，在不同数据场景下(如稀疏数据、新用户、活跃用户等)，合理组合不同相似度计算方法的比例，能够克服单一方法的局限性，显著提高推荐系统的准确性和质量。通过实验验证，我们发现基于加权组合的多维度推荐方法，相较于单一相似度方法，能够在不同推荐场景中提升推荐系统的综合表现。

关键词

个性化推荐，相似度计算，余弦相似度，皮尔逊相似度，欧几里得距离，Jaccard相似度，电影推荐，多维度推荐算法

Research on Multi-Dimensional Movie Recommendation System Based on User Rating Data

Zhenyang Yu, Ting Yu, Xinlei Chen, Lian Xiao

School of Information Engineering, Jiaxing Nanhu University, Jiaxing Zhejiang

Received: Mar. 23rd, 2025; accepted: Apr. 16th, 2025; published: Apr. 23rd, 2025

Abstract

With the widespread adoption of personalized recommendation systems in various applications, movie recommendation as a typical scenario has attracted increasing attention. The core task of

文章引用: 余振洋, 俞婷, 陈鑫磊, 肖炼. 基于用户评分数据的多维度电影推荐系统研究[J]. 数据挖掘, 2025, 15(2): 201-211. DOI: 10.12677/hjdm.2025.152017

recommendation systems lies in providing personalized suggestions based on users' historical behavioral data, particularly rating data. The effectiveness of recommendations is closely tied to the selection of similarity computation methods. Common approaches include cosine similarity, Pearson correlation, Euclidean distance, and Jaccard similarity. While each method has unique characteristics and applicable scenarios, relying solely on a single similarity measure often leads to performance degradation due to data-specific limitations and environmental constraints. This study systematically compares and analyzes the performance of these four similarity computation methods under different environmental conditions, exploring their application effectiveness in movie recommendations. The research demonstrates that rationally combining multiple similarity measures with weighted proportions can overcome the limitations of individual methods and significantly enhance recommendation accuracy and quality across diverse data scenarios (e.g., sparse data, new users, and active users). Experimental results verify that the proposed multi-dimensional recommendation method based on weighted combinations outperforms single similarity approaches in improving comprehensive system performance across various recommendation scenarios.

Keywords

Personalized Recommendation, Similarity Computation, Cosine Similarity, Pearson Correlation, Euclidean Distance, Jaccard Similarity, Movie Recommendation, Multi-Dimensional Recommendation Algorithm

1. 引言

随着互联网和大数据技术的迅速发展，个性化推荐系统已经成为提升用户体验和商业价值的关键工具。在众多推荐任务中，电影推荐作为典型应用，广泛应用于各大平台，成为个性化推荐的重要组成部分。推荐系统的核心任务是基于用户的历史行为数据(如评分、点击、购买等)，为用户提供个性化的内容建议。

为实现精准的电影推荐，推荐系统常通过计算用户和电影之间的相似度来建模用户兴趣。常见的相似度计算方法包括余弦相似度、皮尔逊相似度[1]、Jaccard 相似度[2]和欧几里得距离[3]。每种方法有其独特的计算原理和适用场景。在实际应用中，余弦相似度特别适用于高维稀疏数据，尤其在大规模用户–电影评分矩阵的计算中表现高效；皮尔逊相似度则适用于用户评分趋势相似的情况，能够有效消除评分偏差；欧几里得距离适用于密集评分数据，但在稀疏矩阵中性能较弱；Jaccard 相似度则适用于二值数据，尤其在判断用户是否观看某个电影时效果较好。

尽管这些方法在各自的应用场景中有其优势，但它们也存在明显的局限性。单一使用某种方法可能会受到数据特性和推荐环境的限制，导致推荐性能的下降。特别是当面对稀疏数据、评分偏差等问题时，单一的相似度计算方法往往无法全面捕捉用户的多维度偏好。因此，如何在不同环境下合理选择和组合相似度计算方法，成为提升推荐系统效果的关键问题。

本文旨在通过对余弦相似度、皮尔逊相似度、欧几里得距离和 Jaccard 相似度四种相似度方法进行加权组合搭配，探索其在不同推荐环境下的表现。通过实验评估，将比较这些方法在数据稀疏性、冷启动等问题上的应对能力。通过这种多维度的组合方式，能够有效克服单一相似度方法的局限性，提升推荐系统的整体性能，并为实际应用中的电影推荐系统提供优化方案。

2. 相关工作

近年来，基于多维度相似度融合的推荐方法逐渐成为提升协同过滤性能的关键研究方向。经典研究如Sarwar *et al*. (2001) [3]提出的基于物品的协同过滤算法，通过皮尔逊相关系数缓解数据稀疏性问题，验证了相似度度量对推荐质量的影响。后续研究进一步探索了多指标融合策略：例如，Huang *et al*. (2022) [4]通过加权融合余弦相似度与 Jaccard 系数，验证了混合相似度在冷启动场景下的有效性；Zhou *et al*. (2023) [5]则在社交推荐中结合信任度与欧氏距离，证明了多维度相似度的场景适应性。然而，现有方法多针对单一场景(如冷启动或密集数据)设计权重分配规则，缺乏对多场景下相似度组合策略的系统性分析。

针对上述问题，本文提出一种基于加权混合相似度矩阵的构建方法，旨在探索不同相似度组合对推荐性能的场景敏感性。具体地，选择四种经典相似度度量：余弦相似度：捕捉用户评分向量的方向一致性；皮尔逊相关系数：消除用户评分偏置，衡量线性相关性；Jaccard 相似度：量化用户间共同评分项的占比；欧氏距离：反映评分差异的绝对值大小。

通过赋予四种相似度差异化权重，本方法构建混合相似度矩阵，并分别在稀疏数据、密集数据、新用户和活跃用户场景下评估其性能。实验结果表明，不同场景下最优相似度组合存在显著差异：例如，稀疏数据场景中，余弦相似度与 Jaccard 系数的组合能够有效缓解数据稀疏性带来的偏差；而在活跃用户场景下，皮尔逊相关系数与欧氏距离的组合对评分多样性具有更好的适应性。这一发现与文献[4] [5]中多相似度融合的思路一致，进一步揭示了权重分配需结合场景特征的必要性。

需要指出的是，尽管多相似度组合策略提升了推荐系统的鲁棒性，但其普适性仍受限于两方面因素：其一，不同相似度对数据分布的敏感性差异显著(如 Jaccard 对长尾物品的过度惩罚)；其二，权重的固定分配可能导致场景迁移时的性能下降。未来研究可探索基于数据特征的自适应权重生成机制，以进一步提升方法的泛化能力。

3. 基于用户评分数据的多维度电影推荐系统研究

本文旨在通过对用户评分数据进行分析，结合余弦相似度(Cosine Similarity)、皮尔逊相关系数(Pearson Correlation)、欧几里得距离(Euclidean Distance)和杰卡德相似度(Jaccard Similarity)四种相似度度量方法，构建加权混合相似度矩阵，并在四种模拟环境(数据稀疏环境、新用户环境、活跃用户环境和密集数据环境)下进行了多组实验，探索不同相似度组合下的效果，并根据准确率、召回率、F1 值指标对推荐结果进行评估。

3.1. 方法介绍

3.1.1. 推荐系统模型设计

本文旨在根据用户的评分数据，为用户提供个性化的电影推荐。

(1) 评分数据的影响

用户对电影的评分反映了用户的偏好。评分越高(如 4 或 5 分)，表示用户对该电影越喜欢[6]。为确保推荐的电影符合用户的偏好，在测试集上只选择评分大于等于 4 的电影作为用户的实际偏好。

(2) 评分矩阵的构建

数据经过预处理后，使用 pandas.pivot()函数将评分数据转换为用户－电影评分矩阵，矩阵的行是用户，列是电影，值是评分，未评分的地方填充为 0。根据该评分矩阵，我们对用户或电影进行相似度计算，以基于相似的用户或电影进行推荐。例如：原始评分数据以长格式(Long Format)存储，包含用户 ID、电影 ID 和评分三列，部分示例如表 1 所示，通过 pandas.pivot()函数将其转换为用户－电影评分矩阵(宽格式，Wide Format)，生成如表 2 所示的矩阵。

Table 1. Example of raw rating data
表 1. 原始评分数据示例

用户 ID	电影 ID	评分
1	101	5.0
1	102	3.0
2	101	4.0
2	103	2.0
3	102	4.0
3	103	5.0

Table 2. Transformed user-movie rating matrix
表 2. 转换后的用户－电影评分矩阵

Movie-id	101	102	103
User-id			
1	5.0	3.0	0.0
2	4.0	0.0	2.0
3	0.0	4.0	4.0

(3) 推荐方法

基于用户评分数据，通过计算用户与用户、电影与电影之间的相似度来进行推荐。针对不同场景(如稀疏数据、新用户、活跃用户等)，通过不同的抽样策略来测试推荐系统的表现。

3.1.2. 相似度计算方法

本文采取了四种相似度计算方法：

(1) 余弦相似度(Cosine Similarity)

余弦相似度衡量的是两个向量夹角的余弦值，用于计算用户或物品之间的相似性。其基本思想是，如果两个向量的夹角越小，则它们之间的相似度越高；反之，如果夹角越大，则相似度越低。具体公式如下：

$$\text{Cosine Similarity} = \frac{\boldsymbol{A} \cdot \boldsymbol{B}}{\|\boldsymbol{A}\| \|\boldsymbol{B}\|} \tag{1}$$

其中， $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 分别是两个用户或物品的评分向量， $\|\boldsymbol{A}\|$ 和 $\|\boldsymbol{B}\|$ 是它们的欧几里得范数。

(2) 皮尔逊相似度(Pearson Correlation)

皮尔逊相关系数是一种衡量两个变量线性相关程度的统计方法。它通过计算两个用户或物品评分的相关性来衡量相似度，公式为：

$$\text{Pearson Similarity} = \frac{\sum \left(A_i - \bar{A}\right)\left(B_i - \bar{B}\right)}{\sqrt{\sum \left(A_i - \bar{A}\right)^2 \sum \left(B_i - \bar{B}\right)^2}} \tag{2}$$

其中， A_i 和 B_i 是用户或物品的评分， $\bar{A}$ 和 $\bar{B}$ 分别是它们的平均评分。

(3) 欧几里得距离(Euclidean Distance)

欧几里得距离是最直观的距离度量之一，用于计算两个向量之间的“直线”距离，公式如下：

$$\text{Euclidean Distance} = \sqrt{\sum \left(A_i - B_i\right)^2} \tag{3}$$

其中 A_i 和 B_i 表示用户 A, B 对电影 i 的评分。

(4) Jaccard 相似度(Jaccard Similarity)

Jaccard 相似度主要用于衡量两个集合之间的相似度，尤其适合评估两个集合中元素的交集与并集的比率。对于用户推荐系统，Jaccard 相似度一般用于处理二值数据(如评分是否为正分)，其公式为：

$$\text{Jaccard Similarity} = \frac{|A \cap B|}{|A \cup B|} \tag{4}$$

其中，A 和 B 是两个集合，$|A \cap B|$ 是交集的大小，$|A \cup B|$ 是并集的大小。

3.1.3. 数据集

本文使用了经典的 MovieLens 数据集[7]。这是一个公开的电影推荐数据集，由 GroupLens 研究小组发布，主要用于推荐系统的研究和测试。数据集包括两个文件：movies.dat：包含电影的 ID、标题和类型信息和 ratings.dat：包含用户对电影的评分数据。首先进行数据预处理：数据加载时，使用 pandas.read_csv()函数读取数据，并通过 sep = '::'参数进行分隔(MovieLens 数据集中的默认分隔符为::)。对评分数据进行清洗，仅保留评分大于 0 的记录。由于我们希望使用具有足够互动数据的用户来提高推荐系统的准确性，因此筛选出至少对 10 部电影评分的用户。还对电影进行了筛选，确保仅选择至少被 5 名用户评分的电影，以减少数据稀疏性对模型的影响。

3.1.4. 性能指标

本文分别采用准确率、召回率、F1 值指标[8]对推荐结果进行评估。以下是各个指标的简单介绍：

(1) 准确率(Precision)

定义为推荐结果中实际喜欢的电影所占的比例，适用于衡量推荐系统在推荐内容时的“准确性”。

$$\text{Precision} = \frac{\text{推荐命中的电影数量}}{\text{推荐的电影总数}} \tag{5}$$

(2) 召回率(Recall)

定义为实际喜欢的电影中被推荐的比例，适用于衡量推荐系统是否覆盖了用户的真实偏好。

$$\text{Recall} = \frac{\text{推荐命中的电影数量}}{\text{用户实际喜欢的电影数量}} \tag{6}$$

(3) F1 值(F1 Score)

综合考虑准确率和召回率，是它们的调和平均数，用于平衡推荐的准确性和覆盖度。

$$\text{F1} = 2 \times \frac{\text{Precision} + \text{Recall}}{\text{Precision} \times \text{Recall}} \tag{7}$$

4. 实验设计与结果分析

本实验旨在评估不同相似度计算方法对电影推荐系统效果的影响，并通过调整相似度方法的权重组合来优化推荐结果。通过实验探索最佳的相似度组合，以提升在不同用户场景下推荐系统的准确率、召回率和 F1 值。

4.1. 实验设计

本实验采用了四种相似度度量方法：余弦相似度(Cosine Similarity)，皮尔逊相似度(Pearson Correlation Coefficient)，欧几里得距离(Euclidean Distance)和 Jaccard 相似度(Jaccard Similarity)，尝试了多种相似度比例组合，以权衡不同相似度方法的影响，比例组合如表 3 所示。

Table 3. Weight allocation table
表 3. 权重分配图

组合编号	Cosine	Pearson	Euclidean	Jaccard	备注
1	0.7	0.1	0.1	0.1	余弦主导
2	0.1	0.7	0.1	0.1	皮尔逊主导
3	0.1	0.1	0.7	0.1	欧式主导
4	0.1	0.1	0.1	0.7	Jaccard 主导
5	0.2	0.2	0.4	0.2	随机组合 1
6	0.5	0.1	0.1	0.3	混合组合
7	0.4	0.3	0.1	0.2	随机组合 2

在实验设计中，我们将 MovieLens 数据集按照不同的用户和电影评分特征划分为多个场景，以便评估不同相似度计算方法在各种数据密度下的推荐效果。这些场景包括稀疏数据场景、密集数据场景、新用户场景和活跃用户场景，具体划分的情况如表 4。

Table 4. Scenario classification table
表 4. 场景划分表

场景名称	用户数	电影数
稀疏数据(sparse_data)	2000	1000
密集数据(dense_data)	5000	2000
新用户(new_user)	500	500
活跃用户(active_user)	3000	1500

4.2. 实验结果及分析

通过对实验结果进行 Kruskal-Wallis 检验[9]和 Mann-Whitney U 检验[10]，得出四个场景(稀疏数据场景、密集数据场景、新用户场景和活跃用户场景)观察到的性能差异(基于 F1 (公式 7)值)在统计学上都具有显著意义。每个场景的 Kruskal-Wallis p 值 < 0.05 (均落在$[9.55 \times 10^{-13}, 1.56 \times 10^{-3}]$区间内)，表明 7 的 F1 值在每个场景内存在整体显著差异，且 Mann-Whitney U 检验确认了一些特定比例对之间的差异显著(corrected_p < 0.05)。这些差异主要集中在强调 Cosine 相似性与其他比例(Pearson、Euclidean 或 Jaccard)之间，反映了不同相似性权重对推荐系统性能的影响。在所有显著的 Mann-Whitney U 检验中，校正后 p 值(corrected_p)均<0.05，且落在$[1.64 \times 10^{-8}, 1.07 \times 10^{-2}]$的区间内，验证了这些差异的统计学意义。

(1) 密集数据环境的数据展示及分析

由图 1 可知在密集数据环境(用户数 5000，电影数 2000)下，实验结果表明，基于欧氏距离主导的组

合(权重 0.7)表现最优(F1 = 0.41)，其次是皮尔逊主导组合(F1 = 0.40)与 Jaccard 主导组合(F1 = 0.40)，而余弦相似度主导的组合表现最差(F1 = 0.37)，其中最高精确率达(公式 5) 0.56，召回率(公式 6) 0.41，F1 值 0.41。在评分数据稠密的情况下，欧氏距离通过绝对差异的平方和计算(公式 3)，在密集数据下(用户评分量高、共同项多)显著放大用户间偏好差异信号。由于数据填充率高，评分噪声的统计方差趋于稳定，其信噪比(SNR)提升，从而增强区分能力。与稀疏和活跃环境相比，F1 值小幅下降，说明在数据较为密集的情况下，传统协同过滤方法的改进空间受限。这一现象与 Xue 等人(2017) [11]的研究结论一致。他们指出，在密集数据环境中，深度矩阵分解(Deep Matrix Factorization, DMF)等基于深度学习的方法往往能比传统协同过滤算法提供更优的推荐效果，同时在 MovieLens-10M 数据集上验证，DMF 的 F1 值(0.48)显著高于传统协同过滤方法(0.41)，印证了传统方法的理论瓶颈传统协同过滤方法在数据密集时。虽然能够有效计算相似度，但容易受评分噪声和计算复杂度的影响，而 DMF 等方法可以通过学习更复杂的特征表示来增强推荐的准确性和稳定性。此外，从不同相似度度量的表现来看，皮尔逊相似度仍然能够较好地捕捉评分的线性关系，而欧几里得距离在衡量评分的绝对差异方面提供了一定的补充。余弦相似度在密集数据环境中表现最差的主要原因可能在于方向敏感性与评分强度信息丢失，归一化操作(公式 1)将用户评分向量投影至单位超球面，导致评分强度差异(如用户 A 评分为[5, 5, 5] vs 用户 B [3, 3, 3])被完全忽略，导致推荐结果偏离真实用户偏好。

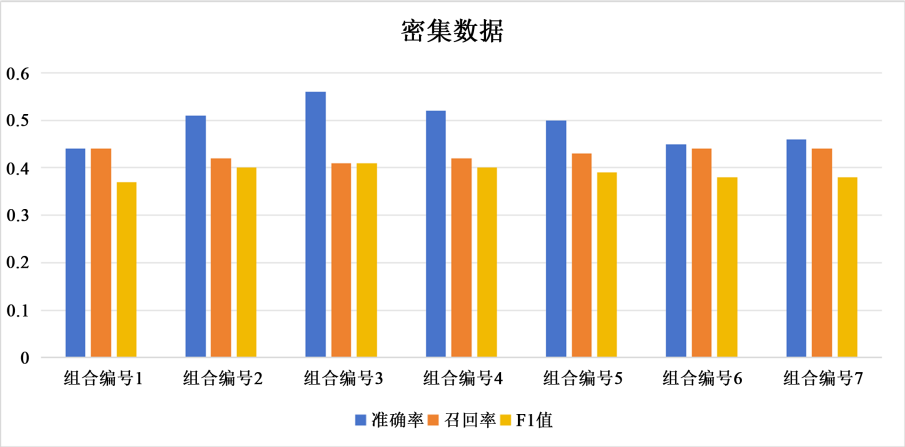

Figure 1. Bar chart of data in dense data scenario
图 1. 密集数据环境中的数据柱状图

(2) 新用户环境中的数据展示及分析

如图 2 可知在新用户环境(用户数 500，电影数 500)下，实验结果表明，基于欧几里得距离主导的组合在新用户场景下表现较优，其精确率达到 0.57，召回率 0.50，F1 值 0.48，相比其他相似度组合，具有更高的稳定性。在新用户场景下，数据的稀疏性可能较高，尤其是用户的评分信息较为稀缺。尽管欧几里得距离在稀疏数据环境下可能受到影响，但如果新用户的评分较为集中，且评分分布不是特别极端(即评分差异不大)，那么欧几里得距离可以有效地捕捉到用户之间的相似性，尤其是在多维评分的情况下。在多维度相似度组合的情境下，权重分配直接影响最终的推荐效果。欧几里得距离主导的组合，意味着它在计算相似度时占据了更高的比重，这可能加强了其对于用户偏好的识别能力，尤其是在新用户数据较为稀疏的情况下。在新用户稀疏场景下，欧氏距离主导的组合有效性的核心逻辑可归纳为：评分集中性加上差异有限性通过欧氏距离(公式 3)可使累积效应放大。这一过程表明，即使数据稀疏，只要评分分布集中且差异有限，欧氏距离通过其数学特性与权重设计的协同作用，仍能有效捕捉用户相似性，从而

提升推荐效果。这一现象与 Bobadilla 等人(2013) [12]的研究结论一致，该论文指出，在用户评分稀疏且分布集中的场景中，欧氏距离能有效消除用户评分尺度差异(如用户 A 评分范围 4~5 星 vs 用户 B 评分范围 3~4 星)，从而更可靠地捕捉绝对差异，同时论文特别强调，余弦相似度未进行中心化处理(即未减去用户评分均值)，导致其对用户评分基线的敏感性(Baseline Sensitivity)。例如，用户 C (均分 4.8)与用户 D (均分 3.5)即使对同一电影的评分相同(如均评 4 星)，余弦相似度(公式 1)也会因基线差异而低估其相似性，这与实验中余弦组合 F1 = 0.42 的次优表现直接相关。此外，在“冷启动问题”章节中指出 Jaccard 系数(公式 4)仅依赖共同项存在性信号(而非评分值差异)，在稀疏数据(如用户共同评分项 ≤3)中容易产生误判。例如，用户 E 对某电影评 5 星与用户 F 评 1 星仍被视为相同存在性信号，这与用户实验中 Jaccard 主导组合 F1 = 0.46 的局限性一致。总体来看，在欧几里得主导的组合中，在组合策略中，欧氏距离主导捕捉评分强度差异，而皮尔逊相似度通过均值中心化缓解用户评分偏置，Jaccard 系数扩展行为存在性覆盖，三者协同提升鲁棒性。通过加权组合，它们互相补充，从而提升了推荐效果，尤其是在新用户场景下，组合策略可能通过互补优势部分缓解单一度量的局限。但整体推荐性能仍受到冷启动问题的影响，未来可尝试结合基于内容的推荐方法，以进一步提升推荐质量。

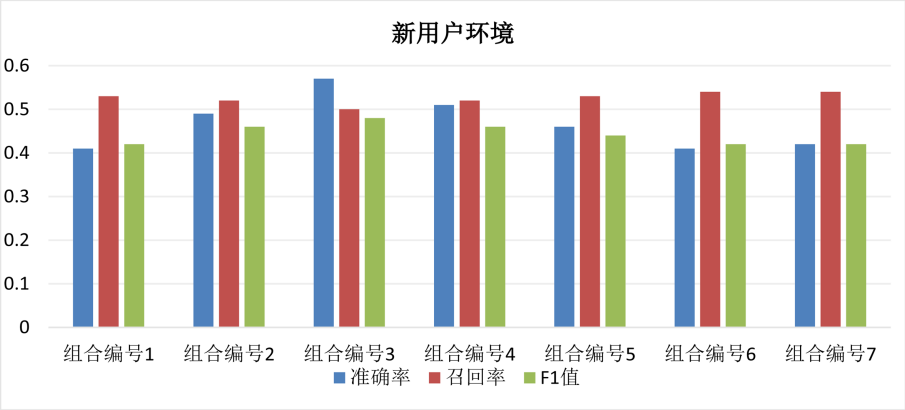

Figure 2. Bar chart of data in new user scenario
图 2. 新用户环境中的数据柱状图

(3) 稀疏数据环境中的数据展示及分析

通过图 3 可知在稀疏数据环境(用户数 2000，电影数 1000)下，皮尔逊和欧氏在稀疏环境下分别表现出准确率优势(0.54 和 0.56)，Jaccard 的表现最为均衡(F1 = 0.46，低于欧氏 0.48 但高于余弦 0.37)。在用户–物品评分矩阵密度仅为 0.5% (2000 用户 × 1000 电影)的极端稀疏场景下，不同相似度度量的性能差异本质上反映了其数学原理对数据稀疏性的适应能力差异。由皮尔逊(公式 2)可知其核心是通过均值中心化消除用户评分偏置(如严格型用户普遍低分、宽容型用户普遍高分)，仅保留评分模式的相对差异。在稀疏场景下，评分矩阵中的强关联用户往往具有线性评分模式(如某类用户对所有电影评分均比均值高固定分值)，皮尔逊对此类模式敏感，在稀疏场景下，皮尔逊的均值中心化对评分偏置的消除能力依赖于共同评分项的统计显著性。当共同评分项数 ≥3 时，均值估计趋于稳定，此时皮尔逊能有效保留“评分趋势一致性”信号，主导权重下准确率达 0.54，印证其对评分偏置的消除能力。欧式距离(公式 3)计算用户评分向量的绝对差异，其本质是绝对值差异的累积，对评分量级敏感。在共同评分项目较少时，用户对同一物品的评分绝对差异(如用户 A 评 5 分、用户 B 评 1 分)会被放大，形成显著的距离差异。这种特性在稀疏环境下有利于区分具有极端偏好差异的用户，新用户若对少数物品给出极端评分(如 5 分或 1 分)，欧氏距离能快速识别具有相似极端偏好的用户群，而皮尔逊可能因样本不足导致均值估计偏

差，相比需要协方差计算的皮尔逊，欧氏距离的计算复杂度更低，在实时推荐场景下更具优势。Jaccard系数(公式 4)定义为共同评分项目占比，完全忽略具体评分值，仅关注评分行为的存在性。当用户共同评分项目极少时(如仅 1 个共同评分)，Jaccard 仍能通过其他非共同项目的评分存在性(如用户 A 评过物品 M、用户 B 评过物品 N)挖掘潜在关联，而皮尔逊/欧氏因缺乏共同评分直接失效，均衡性原理在于通过扩大关联用户范围(包括评分模式不同但行为分布相似的用户)，提升推荐覆盖率。Breese 等人[13]的研究指出皮尔逊在评分偏置显著时(如严格型用户)表现更优，而欧氏在极端偏好差异下(如用户对某类电影全 5 分 vs 全 1 分)更具区分力，皮尔逊与欧氏分别通过评分偏置修正和绝对差异捕捉，在稀疏数据下具有互补优势。

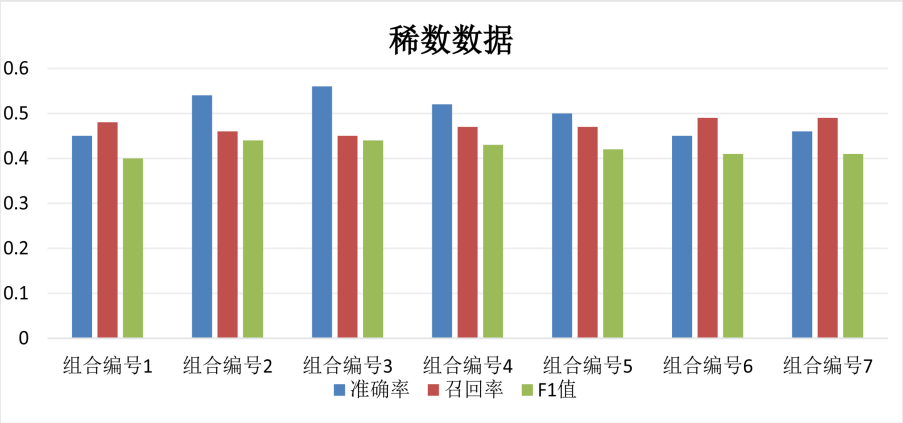

Figure 3. Bar chart of data in sparse data scenario
图 3. 稀疏数据环境中的数据柱状图

(4) 活跃用户环境的数据展示及分析

通过图 4 可知在活跃用户环境(用户数 3000，电影数 1500)下，欧式主导的准确率最高，这是因为在用户评分量充足时，差异累积的稳定性随数据量增加而增强，信噪比提升，对强偏好差异(如用户 A 全 5 分 vs 用户 B 全 1 分)形成显著区分信号，活跃用户的评分差异分布更清晰，欧氏距离通过绝对值差异的平方和，精确捕捉用户间的强偏好对立。Jaccard 系数通过行为存在性扩展了关联用户范围，覆盖更多潜在兴趣，使其召回率表现较好，但是忽略评分值的特性导致误关联(如用户 A 对电影 X 评 5 分，用户 B 评 1 分，但 Jaccard 仍认为他们相似)致使准确率略低。Bobadilla 等人(2013)也指出 Jaccard 系数在活跃环境下误关联率升高(如用户对热门电影的高/低评分被等同)，导致准确率下降(Section 5.4)。余弦相似度(权重 0.7 时)：准确率 0.44 (最差)，由其公式 1 可知方向对齐优先，忽略评分强度差异，归一化操作将用户评分向量投影至单位超球面，导致评分基准坍缩：严格评分者和宽容评分者对同一电影的评分差异被掩盖。余弦相似度在用户评分基准差异显著但相对偏好一致时可能更适用，但在活跃用户场景下，其缺陷如强度坍缩成为主要瓶颈。Pearson 公式 2 行为是捕捉线性相关性，消除评分均值影响，通过减去用户评分均值，缓解评分基准差异(如严格型与宽容型用户)，但是会存在线性假设局限：若用户 A 对某类电影评分随电影热度递增(非线性偏好)，而用户 B 递减，Pearson 可能低估其相关性。准确率与 Jaccard 持平(0.52)，优于余弦相似度，因均值中心化减少误对齐；但弱于欧氏距离，因无法捕捉非线性偏好。这一现象与 Aggarwal (2016)在《Recommender Systems: The Textbook》[2]一书中的研究结论一致，指出，皮尔逊与欧氏分别通过线性修正(均值中心化)和差异累积(L2 范数)捕捉用户偏好的不同维度，二者在活跃环境下的互补性可部分缓解单一度量的局限性。

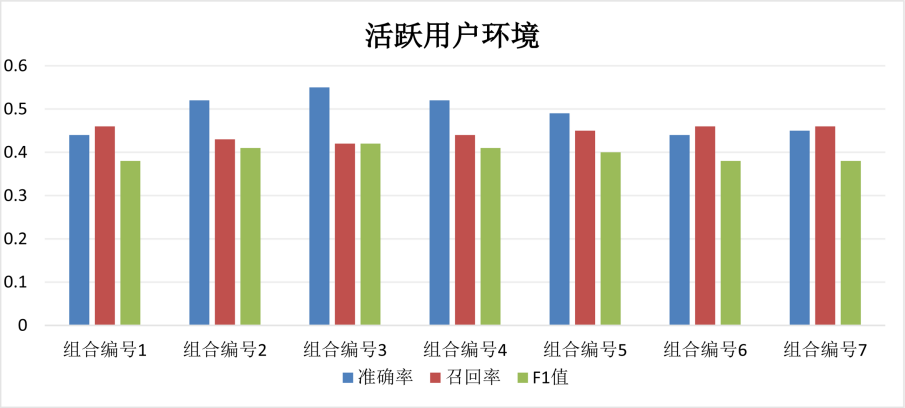

Figure 4. Bar chart of data in active user scenario
图 4. 活跃用户环境的数据柱状图

5. 结束语

本文研究了基于用户评分数据的多维度电影推荐系统，采用了余弦相似度、皮尔逊相似度、欧几里得距离和 Jaccard 相似度四种常见的相似度计算方法，并对其在不同场景下的表现进行了评估。通过实验分析发现，不同的相似度计算方法适用于不同的推荐场景，且各自具有其优势和局限性。在稀疏数据和冷启动问题上，基于内容的推荐方法和混合推荐系统可以有效地改善推荐效果。未来我们将通过引入深度学习或强化学习进一步提高推荐系统的精度和实时性，尤其是在用户兴趣变化较快的情况下。除了用户评分数据外，我们计划结合社交网络数据、行为数据等信息，进一步丰富用户的偏好模型，提升推荐的准确性。

基金项目

嘉兴南湖学院大学生研究训练计划(No. 8517233215)。

参考文献

[1] 马超, 张力, 赵祥, 等. 基于协同过滤的个性化推荐算法研究[J]. 计算机科学, 2017, 44(5): 125-130.

[2] Aggarwal, C.C. (2016) Recommender Systems: The Textbook. Springer. https://doi.org/10.1007/978-3-319-29659-3

[3] Sarwar, B., Karypis, G., Konstan, J. and Riedl, J. (2001). Item-Based Collaborative Filtering Recommendation Algorithms. *Proceedings of the* 10*th International Conference on World Wide Web*, New York, 1-5 May 2001, 285-295. https://doi.org/10.1145/371920.372071

[4] 黄向春, 刘洋, 王浩. 基于多相似度融合的协同过滤冷启动优化方法[J]. 计算机学报, 2022, 45(10): 2105-2114.

[5] 周正乾, 张伟, 陈林. 社交网络中融合信任与多维度相似度的推荐模型[J]. 软件学报, 2023, 34(3): 567-580.

[6] Cremonesi, P., Koren, Y. and Turrin, R. (2010) Performance of Recommender Algorithms on Top-N Recommendation Tasks. *Proceedings of the* 4*th ACM Conference on Recommender Systems*, Barcelona, 26-30 September 2010, 39-46. https://doi.org/10.1145/1864708.1864721

[7] GroupLens Research (2009) MovieLens Dataset. https://grouplens.org/datasets/movielens/

[8] Powers, D. (2011) Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation. *Journal of Machine Learning Technologies*, **2**, 37-63.

[9] Kruskal, W.H. and Wallis, W.A. (1952) Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*, **47**, 583-621. https://doi.org/10.1080/01621459.1952.10483441

[10] Mann, H.B. and Whitney, D.R. (1947) On a Test of Whether One of Two Random Variables Is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, **18**, 50-60. https://doi.org/10.1214/aoms/1177730491

[11] Xue, H., Dai, X., Zhang, J., Huang, S. and Chen, J. (2017) Deep Matrix Factorization Models for Recommender Systems.

Proceedings of the 26*th International Joint Conference on Artificial Intelligence* (*IJCAI*), 19-25 August 2017, Melbourne, 3203-3209. https://doi.org/10.24963/ijcai.2017/447

[12] Bobadilla, J., Ortega, F., Hernando, A. and Gutiérrez, A. (2013) Recommender Systems Survey. *Knowledge-Based Systems*, **46**, 109-132. https://doi.org/10.1016/j.knosys.2013.03.012

[13] Breese, J.S., Heckerman, D. and Kadie, C. (1998) Empirical Analysis of Predictive Algorithms for Collaborative Filtering. *Proceedings of the* 14*th Conference on Uncertainty in Artificial Intelligence* (*UAI*), San Francisco, 24-26 July 1998, 43-52.