小树苗相似性：一种可执行、可解释的基于记忆的推荐工具

Giambattista Albora,Lavinia Rossi Mori,Andrea Zaccaria

摘要

《小树苗相似性：一种可执行、可解释的基于记忆的推荐工具》介绍了一种用于推荐的基于记忆的工具——Sapling Similarity。在推荐系统中，基于记忆的协同过滤通过测量用户或物品之间的相似性来进行推荐。现有的基于共同邻居的方法存在局限性，而受决策树启发的 Sapling Similarity 方法允许相似度取负值，能够更好地处理不同用户或物品之间的差异。通过实验，Sapling Similarity 在多个数据集上的表现优于现有相似度指标。与基于模型的方法相比，Sapling Similarity 具有更高的可解释性，同时在推荐准确性上也具有竞争力，在 Amazon - Book 数据集上表现尤其突出。文章还介绍了 Sapling Similarity 的计算方法和相关实验设置，为推荐系统的研究和实践提供了新的思路和方法。

关键词：推荐算法；基尼不纯度；决策树

1 引言

在当今数字化时代，复杂系统中不同对象间的交互广泛存在，众多复杂系统可简化为两类不同对象间的交互，如经济系统中出口产品与国家或企业的关联、生物系统中疾病与患者或微生物的联系、社会系统中用户与 Facebook 页面或演员与参演电影的关系等。这些交互关系通常通过二分网络进行有效表示，其中节点分属不同集合，链接仅连接不同集合中的节点。二分网络中的链接情况多样，包括连续评分、区间评分、二元评分和一元评分等。在信息系统领域，基于二分网络的协同过滤（CF）是构建推荐系统的常用方法，其目的在于依据用户与物品的过往交互信息，向用户推荐可能感兴趣的物品。CF 主要分为基于记忆和基于模型两类方法。基于记忆的 CF 通过计算用户或物品间的相似度来寻找近邻，进而基于近邻进行推荐，具有实现简单直观、无需超参数且结果易解释等优点；基于模型的 CF 则运用图神经网络（GCN）和矩阵分解（MF）以提升推荐质量，但依赖优化超参数来最大化性能。在实际应用中，如经济复杂性框架，基于记忆的 CF 因其简单性和可解释性而备受青睐，被广泛用于预测和推荐。然而，当前基于记忆的 CF 在处理一元（无权）链接的二分网络时，其相似度度量方法存在局限性。传统基于共现的相似度度量方法通常为正定的，仅考虑节点间的正相似性，忽视了节点可能存在的负相似性。例如在经济领域，日本擅长高科技产品，赞比亚专注原材料，两者出口产品差异大，存在负相关，但传统方法无法体现这种关系。这种对负相似性的忽视，可能导致在推荐系统中无法准确评估信息对推荐的影响，从而影响推荐效果。

为突破上述困境，本文创新性地提出一种全新的局部相似度度量方法——Sapling Similarity。该方法基于信息理论与概率方法构建，其核心在于允许相似度取值为负，这一特性使

它能够有效识别节点间的反相关或不相似关系。同时，Sapling Similarity 在计算过程中充分考虑网络整体规模信息，这是相较于传统方法的又一重大突破。传统方法往往仅依赖节点共现次数等局部信息，忽略了网络整体结构对相似度判断的潜在影响。Sapling Similarity 的提出，为二分网络节点相似性度量提供了更为全面、合理的解决方案，有望显著提升推荐系统性能，推动其在经济、社交、娱乐等众多领域的广泛应用与深度发展。

2 相关工作

2.1 协同过滤

协同过滤 (CF) 是 20 世纪 90 年代中期引入的一种非常流行的推荐系统技术，广泛应用于多个领域。例如，亚马逊（Amazon）使用基于物品的 CF 为用户推荐产品，在经济复杂性框架中，基于物品的 CF 常被用于衡量国家与产品出口之间的关联性。CF 主要在二分网络上进行操作，其中一个层的节点可视为用户，另一个层的节点视为物品，节点间的链接表示用户对物品的评分或交互行为。目前协同过滤分为基于记忆的协同过滤和基于模型的协同过滤，它们的特点如下：

基于记忆的 CF：基于记忆的 CF 的核心在于测量用户或物品之间的相似度，以此找到近邻，进而基于近邻进行推荐。在处理一元（unary）数据（即链接仅表示存在或不存在，无权重信息）时，最简单的相似度度量方法是计算共现次数，也就是两个节点共同连接到的其他节点数量。然而，由于度高的节点相较于度低的节点更容易产生共现，所以通常会使用节点的度对共现次数进行归一化，从而得到不同的相似度度量指标。基于记忆的 CF 方法具有简单直观、无需复杂超参数且结果易解释等优点，但传统基于共现的相似度度量方法存在局限性，即均为正定的，无法考虑节点间可能存在的负相似性，如在经济领域中不同国家出口产品的互补或竞争关系可能导致负相关，但这些方法无法捕捉。

基于模型的 CF：随着机器学习技术在不同研究领域的成功应用，越来越多基于模型的 CF 技术被开发出来。其中，Graph Convolution Networks（GCN）及其衍生模型在构建 CF 方面表现出色。例如，2019 年提出的 NGCF 在 Gowalla、Yelp2018 和 Amazon - Books 等数据集上取得了较好性能；2020 年的 LightGCN 通过简化 NGCF 架构进一步提升了推荐质量。此后，基于 LightGCN 又发展出了 LT - OCF、SimpleX、UltraGCN 等模型，它们旨在以简单架构实现最优结果。此外，2022 年 Choi 等人受基于分数的生成模型启发提出的 BSPM - EM 和 BSPM - LM 模型，在上述三个数据集上也取得了当前先进的结果。这些基于模型的方法通常依赖多个超参数，这些超参数需要在每个数据集上进行优化以获得最佳推荐分数。

已有研究中，基于记忆的 CF 方法虽简单易用，但相似度度量的局限性影响了其在复杂数据关系下的推荐效果；基于模型的 CF 方法虽性能较高，但依赖大量超参数优化且结果解释性相对较差。本文旨在提出一种新的相似度度量方法 Sapling Similarity，以克服传统基于记忆的 CF 方法在处理一元数据时无法考虑负相似性的问题，并将其应用于协同过滤中，构建 SSCF 模型，与现有方法进行比较，验证其在推荐系统中的有效性和优势，同时探索其在二分网络分析及其他相关领域的潜在应用价值。

$$M_{i\alpha} = \begin{cases} 1 & \text{如果 } (i,\alpha) \in E \\ 0 & \text{如果 } (i,\alpha) \notin E \end{cases}$$

图 1. 矩阵元素值

2.2 二分图网络

在协同过滤推荐系统中，二分图技术发挥着关键作用。通过构建用户与物品的二分图，系统能够依据用户 - 物品的历史交互信息，精准预测用户对未接触物品的兴趣，从而实现个性化推荐。二分网络被定义为一个图 G=(U,V,E)，其中 U 和 V 是两个不同的节点集合（也称为层），E 是节点 i 与 a 之间所有连接（i,a）的集合。设 |U| 是集合 U 的维度 (基数)，|V| 是集合 V 的维度 (基数)，二分网络可以用一个的二进制矩阵（邻接矩阵）|U|×|V| 来表示，其元素的定义如图一所示：

节点 i 或 a 的度是与之相连的数量，分别表示为：$k_i = \sum_{\lambda=1}^{|V|} M_{i\lambda}$和$k_a = \sum_{l=1}^{|U|} M_{la}$。其中节点 i,j 或 a， 之间的共现次数为：$CO_{IJ}^{(users)} = \sum_{\lambda=1}^{|V|} M_{i\lambda}M_{j\lambda}$ 和 $CO_{\alpha\beta}^{(items)} = \sum_{l=1}^{|U|} M_{la}M_{l\beta}$。CO 矩阵定义了两个单部图网络，其节点仅属于一个集合。通常，二分网络描述了用户与物品之间的交互（例如，在国家 - 出口产品网络的情况下，我们可以将前者视为用户，后者视为物品）。在我们的符号表示中，集合 U 对应于用户，集合 V 对应于物品。重点关注两个用户之间的 Sapling Similarity，并且将定义（物品的总数）N=|V|。物品之间相似性的情况是等效的，唯一的变化是 N=|U|（用户的总数），并且共现次数和度是指物品层中的节点。

2.3 决策树苗

决策树是树状相似性的基础构建块，用于量化一个用户与另一个用户连接或不连接某一物品时，对估计第一个用户连接该物品概率的影响。它由三个盒子组成，每个盒子分为两个区域，分别表示用户连接和不连接物品的情况。通过比较不同区域中连接和不连接物品的比例，可以直观地判断两个用户之间的相似性正负。例如，若已知用户连接某物品会增加用户连接该物品的概率，则两者相似性为正；反之，若降低概率，则相似性为负；若不改变概率，则相似性为零。如图一所示，为构建的决策树苗，它分为三个部分，根部及左右节点，接下里介绍其作用。

根节点（下部盒子）：全面展示用户 i 在整个物品集合中的连接状况。右侧区域记录用户 i 连接的物品数量 ki，左侧区域记录未连接的物品数量 N-ki，比例计算以物品总数为 N 分母。这为后续判断提供了基础的概率参考，即先验概率，表示在不考虑其他用户连接信息时，物品与用户 i 连接的可能性。

右叶子节点(右侧盒子)：聚焦于用户 j 连接的物品子集。右侧区域记录用户 i 和 j 共同连接的物品数量 COij(users)，左侧区域记录用户 j 连接但用户 i 未连接的物品数量 kj-COij(users)，计算比例时以用户连接的物品数量 kj 为分母。通过对比右叶子节点与根节点的比例变化，可以判断用户 j 的连接情况对用户 i 连接概率的影响方向。若该子集中用户 i 连接的比例高于根节点概率，说明用户 j 的连接对用户 i 有正向影响，相似性倾向于正；反之则倾向于负。

左叶子节点（左侧盒子）：从用户 j 未连接的物品角度进行分析。右侧区域记录用户 i

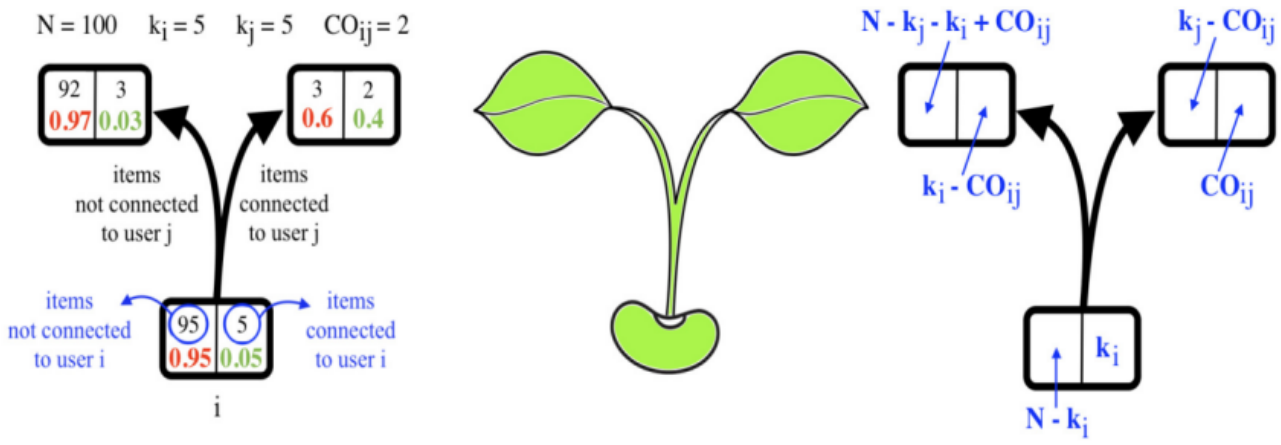

图 2. 决策树苗

连接但用户未连接的物品数量 ki-COij(users)，左侧区域记录两者均未连接的物品数量 N-kj-ki+COij(users。左叶子节点的比例变化同样用于辅助判断相似性，从另一个侧面反映用户 j 的连接信息对用户 j 连接概率的影响。

3 本文方法

3.1 基于 Sapling Similarity 的协同过滤

在评估了 Sapling Similarity 矩阵（包括用户间 B（user）的和物品间的 B(item) 之后，可利用它们构建用户 - 基于和物品 - 基于的协同过滤（CF）模型，进而定义 SSCF，以实现更精准的推荐。对于用户 - 基于 CF，向用户推荐物品的置信度值计算公式为：$S_{i\alpha}^{(item)} = \frac{\sum_\lambda B_{\alpha\lambda}^{(item)} M_{i\lambda}}{\sum_\lambda |B_{\alpha\lambda}^{(item)}|}$

该公式通过对与用户 i 相似的用户 l 对物品 a 的连接情况进行加权求和，并除以相似度绝对值的总和来计算推荐置信度。这意味着如果与相似的用户大多连接到物品 a，则会较高 Sia(item)，从而更倾向于向 i 推荐 a。

对于物品 - 基于 CF，推荐物品 a 给用户 i 的置信度值计算公式为：$S_{i\alpha}^{(item)} = \frac{\sum_\lambda B_{\alpha\lambda}^{(item)} M_{i\lambda}}{\sum_\lambda |B_{\alpha\lambda}^{(item)}|}$

这里是根据物品 a 与其他物品　的相似度以及用户 i 对物品　的连接情况来计算，若与 a 相似的物品大多被连接，则 Sia(item) 较高，就更可能被推荐给 i 用户。

SSCF 的定义与计算 SSCF 定义为用户 - 基于和物品 - 基于估计的加权平均，公式为：

SSCF=(1-γ)$S^{(user)}$ + $\gamma S^{(item)}$ 其中 r 是模型中的唯一参数，用于调节在这种混合推荐方法中用户 - 基于和物品 - 基于推荐的相对权重。例如，当 r=0 时，SSCF 等同于用户 - 基于 CF；当 r=1 时，SSCF 等同于物品 - 基于 CF。通过调整的值，可以在不同数据集和应用场景中找到最优的加权组合，以平衡用户间相似性和物品间相似性对推荐结果的影响，从而提高推荐的准确性和有效性。SCF 结合了用户 - 基于和物品 - 基于 CF 的优点，能够更全面地利用用户与物品之间的关系信息进行推荐。相比于单一的用户 - 基于或物品 - 基于 CF，它能够适应更复杂的用户偏好和物品关系模式。

dataset	users	items	interactions	density
Country -Export	169	5040	120438	14.14%
Yelp2018	31,668	38,048	1,561,406	0.130%
Amazon - Product	52,643	91,599	2,984,108	0.062%

图 3. 数据集属性

4 复现细节

4.1 代码复现

主要为推荐系统的实现，主要功能是根据用户的历史行为数据，计算用户之间的相似度和物品之间的相似度，然后根据这些相似度进行推荐。具体步骤如下：1. 数据读取：根据用户指定的数据集名称，调用 utils.read$_data$

2. 计算基于用户的推荐：根据用户相似度矩阵 B 和评分矩阵 M，计算每个用户对所有物品的预测评分。这里使用的是加权平均的方法，即将每个用户对其他用户的评分进行加权求和，然后除以该用户对所有其他用户的相似度之和。

3. 计算基于物品的推荐：根据物品相似度矩阵 B 和评分矩阵 M，计算每个物品对所有用户的预测评分。这里使用的是加权平均的方法，即将每个物品对其他物品的评分进行加权求和，然后除以该物品对所有其他物品的相似度之和。

4. 计算最终推荐：将基于用户的推荐和基于物品的推荐进行加权求和，得到最终的推荐结果。这里的权重是用户指定的参数 g，表示基于用户的推荐和基于物品的推荐的相对重要性。

5. 输出结果：最后，代码输出了所有推荐方法的性能指标和运行时间。

4.2 数据集

在这个实验中,在三个数据集上进行了评估,分别是 Country-export,Yelp2018 和 Amazon-Product。这些数据集的一些属性如表图 2 所示展示这些数据集的关键属性，如用户数量（users）、物品数量（items）、交互数量（interactions）以及密度（Density）等，可以直观了解数据集的规模和稀疏程度。这些数据集的多样性允许在不同的协同过滤方法之间进行公平且全面的比较。评价模型优劣的性能指标为：

- precision@20：前 20 条推荐中相关元素的比例;
- recall@20：前 20 条推荐中相关元素的比例;
- ndcg@20（归一化折损累计增益）：计算时只考虑排名前 20 位的得分。

	model	precision@20	recall@20	ndcg@20
Country-Export	Jaccard	0.1840	0.0447	0.1988
	Adamic/adar	0.1663	0.0355	0.1768
	SSCF	**0.2118**	**0.0479**	**0.2259**

图 4

	model	precision@20	recall@20	ndcg@20
Yelp2018	Jaccard	0.0223	0.0987	0.0650
	Adamic/adar	0.0221	0.0972	0.0634
	SSCF	**0.0227**	**0.0995**	**0.0682**

图 5

4.3 引入其它协调过滤算法

为了与其它算法进行比较，引入除了 Sapling Similarity 之外的其他基于记忆的 CF 所使用的相似度度量方法。

1.Jaccard（杰卡德相似度）：公式为：$B_{ij}^{JA} = \frac{CO_{ij}}{k_i + k_j - CO_{ij}}$

该方法在计算相似度时考虑了节点和的度，通过将共现次数除以两个节点度之和减去共现次数来归一化相似度。在用户 - 物品二分网络中，可用于衡量用户或物品间的相似性，相较于 Common Neighbors，能在一定程度上减少大度节点对相似度的影响。

2. Adamic/Adar（亚当斯 / 阿达相似度）：$B_{ij}^{AD} = \sum_{\lambda} \frac{M_{i\lambda} M_{j\lambda}}{\log(k_\lambda)}$

其中 K 是节点的度。该方法对共现进行加权，权重为节点度的对数的倒数。其思想是稀有共同邻居（即度较小的共同邻居）对相似度的贡献更大，适用于区分不同重要性的共现关系。

5 实验结果分析与未来展望

在本次研究中，我们精心选取了三个相对较小但极具代表性的数据集，即 Country-Export 数据集、Yelp2018 数据集以及 Amazon-Product 数据集，展开了一系列严谨且细致的实验。这些数据集涵盖了不同领域的丰富信息，为我们的实验提供了多样化的数据支撑和场景模拟。基于 Sapling Similarity 的 SSCF 模型在多个关键指标上展现出了卓越的性能优势。具体而言，

在精度这一重要指标上，SSCF 模型能够以更高的准确性捕捉到用户真正感兴趣的内容，相较于其他传统相似度度量方法等模型，其精度提升显著，能够更加精准地定位和推荐，大大减少了误推的情况，为用户提供了更为可靠的推荐结果。在召回率方面，SSCF 模型同样表现出色。它能够更全面地挖掘出用户可能感兴趣的项目，不放过任何一个潜在的相关内容，从而使得召回率得到了明显的提高。这意味着用户能够在更大程度上发现那些原本可能被遗漏但实际上符合其兴趣和需求的信息，极大地丰富了用户的体验和选择范围。而 NDCG（归一化折损累计增益）指标更是进一步凸显了 SSCF 模型的优越性。该指标综合考虑了推荐结果的排序质量和相关性，SSCF 模型在这一指标上的优异表现，充分证明了它不仅能够准确地推荐相关内容，还能将最相关的内容优先呈现给用户，极大地提升了推荐系统的整体质量和用户满意度。此外，Sapling Similarity 的应用潜力远不止于此。它不仅在协同过滤推荐系统中为用户提供精准且个性化的推荐服务，还能够广泛应用于二分网络投影等领域，如社区检测和聚类等。在社区检测中，Sapling Similarity 能够敏锐地捕捉到网络中节点之间的相似性和关联性，从而准确地划分出不同的社区结构，揭示出网络中隐藏的社区模式和节点关系，为深入理解网络的组织结构和动态演化提供了有力的工具和方法。

展望未来，Sapling Similarity 的应用前景一片光明。我们可以进一步探索利用它进行特征选择的可能性和方法。在面对经济复杂性等相关领域中大规模高维数据的挑战时，通过 Sapling Similarity 进行特征选择，有望大幅减少机器学习算法训练样本的特征数量。这不仅能够显著降低训练时间，提高模型的训练效率，还能增强模型的可解释性，使我们更容易理解模型是如何基于关键特征做出决策和预测的。这一探索将为解决复杂数据处理问题提供新的思路和途径，推动相关领域的发展和进步，为数据驱动的决策和应用带来更多的可能性和价值。

参考文献

[1] G. Albora and A. Zaccaria. Machine learning to assess relatedness: the advantage of using firm-level data. *Complexity*, 2022, 2022.

[2] M. Del Vicario, F. Zollo, G. Caldarelli, A. Scala, and W. Quattrociocchi. Mapping social dynamics on facebook: The brexit debate. *Social Networks*, 50:6–16, 2017.

[3] K. I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A. L. Barabási. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690, 2007.

[4] C. A. Hidalgo. Economic complexity theory and applications. *Nature Reviews Physics*, pages 1–22, 2021.

[5] S. Li, M. Xie, and X. Liu. A novel approach based on bipartite network recommendation and katz model to predict potential micro-disease associations. *Frontiers in Genetics*, page 1147, 2019.

[6] A. L. Schmidt, F. Zollo, A. Scala, C. Betsch, and W. Quattrociocchi. Polarization of the vaccination debate on facebook. *Vaccine*, 36(25):3606–3612, 2018.

[7] M. Stracamore, L. Pietronero, and A. Zaccaria. Which will be your firm's next technology? comparison between machine learning and network-based algorithms. *arXiv preprint arXiv:211002004*, 2021.

[8] A. Tacchella, M. Cristelli, G. Caldarelli, A. Gabrielli, and L. Pietronero. A new metrics for countries' fitness and products' complexity. *Scientific reports*, 2(1):1–7, 2012.

[9] D. J. Watts and S. H. Strogatz. Collective dynamics of'small-world' networks. *Nature*, 393(6684):440–442, 1998.

[10] A. Zaccaria, M. Del Vicario, W. Quattrociocchi, A. Scala, and L. Pietronero. Poprank: Ranking pages' impact and users' engagement on facebook. *PloS one*, 14(1):e0211038, 2019.

article
[style=authoryear]biblatex output.bib