文章编号:2096-1472(2023)10-0015-04
DOI:10.19644/j.cnki.issn2096-1472.2023.010.004

一种基于遗忘机制与余弦相似度的智能推荐算法

许 馨,郭家赫,乔 宇,舒万能

(中南民族大学计算机科学学院,湖北 武汉 430074)

✉ 2020120367@scuec.edu.cn; 2020120335@scuec.edu.cn; 2020120344@scuec.edu.cn; shuwanneng@whu.edu.cn

摘 要:为了解决传统的协同过滤推荐算法计算用户之间相似性度量时,忽略用户与物品之间的相似关系导致推荐性能下降的问题,设计了一种结合遗忘机制与用户相似度的推荐算法。该算法基于用户-用户和物品-物品余弦相似度值和关系二元性,同时引入遗忘机制,根据用户对物品的评分以及记忆留存率进行偏好权重计算,再通过仔细合并相似度的值提高系统的覆盖率和点击率。通过在数据集 MovieLens 上与其他链接预测算法进行对比实验,结果证明该算法的命中率相较于其他算法提高了约 7%,覆盖率略高于现有算法。

关键词:推荐算法;余弦相似度;兴趣漂移;链接预测

中图分类号:TP312 **文献标志码**:A

An Intelligent Recommendation Algorithm Based on Forgetting Mechanism and Cosine Similarity

XU Xin, GUO Jiahe, QIAO Yu, SHU Wanneng

(*School of Computer Science, South-Central Minzu University, Wuhan 430074, China*)

✉ 2020120367@scuec.edu.cn; 2020120335@scuec.edu.cn; 2020120344@scuec.edu.cn; shuwanneng@whu.edu.cn

Abstract: Traditional collaborative filtering recommendation algorithm tends to ignore the similarity between users and items when calculating the similarity measure between users, which leads to the decline of recommendation performance. In order to solve this problem, this paper proposes to design a recommendation algorithm combining forgetting mechanism and user similarity. The algorithm is based on the user-user and item-item cosine similarity values and relationship duality. At the same time, the forgetting mechanism is introduced. The preference weight is calculated according to the user's score on the item and the memory retention rate, and then the system coverage and click rate are improved by carefully merging the similarity values. Through comparative experiments with other link prediction algorithms on the dataset MovieLens, the results show that the hit rate of this algorithm is about 7% higher than that of other algorithms, and the coverage rate is slightly higher than that of the existing algorithms.

Key words: recommendation algorithm; cosine similarity; interest drift; link prediction

0 引言(Introduction)

近年来,用户在线活动产生的信息量呈指数级增长,如何过滤日渐庞大的信息成为一个难题。推荐系统[1]是一种特殊的过滤方法,该方法根据用户的过往兴趣计算出类似的物品,再推荐给用户,以处理信息过载[2]问题。近年来,许多学者开展了推荐算法的研究,由于商品不断增加,用户对推荐性能和功能的需求也不断增长,因此学者们对于该领域的研究热情依然高涨。

基于用户相似性的链接预测方法,通过观测用户过往偏好的方式,研究人员设计了模型计算用户对新物品的潜在喜好度,并将潜在喜好度高的物品个性化地推荐给用户[3]。但在现实生活中,随着新物品出现时间的推移,用户会产生兴趣漂移,进而影响推荐的准确性。传统的链接预测算法只关注用户-用户或者物品-物品之间的相似度,忽略了用户与物品之间的相

收稿日期:2023-02-07

似关系，对推荐算法的准确性产生了不利影响。

为了解决这一问题，本文提出一个依赖于复数的实部和虚部的推荐模型。推荐模型中，首先将相似或不相似的链接用实数加权，而喜欢或不喜欢的链接用复数加权，由于复数在实数和虚数之间提供了一个自然的代数联系，因此推荐问题转化成一个链接预测问题。其次结合遗忘机制，利用遗忘度为用户对物品的偏好权重进行加权。最后计算两个节点之间的相似性，生成 TOPN(前 N 个最高的数据)推荐列表。通过在真实数据集中进行实验，验证推荐算法的性能。

1 相关工作(Related work)

1.1 相关推荐算法

推荐算法通常被分为基于内容的个性化推荐算法和协同过滤推荐算法。基于内容的推荐算法依赖于内容的相似度，而协同过滤推荐算法依赖于相似用户提供的评分，但这些推荐算法有各自的缺陷，例如“冷启动”“兴趣漂移”“数据稀疏”等问题极大地影响了推荐算法的准确性和推荐的多样性。在传统的基于用户或基于物品的协同过滤相似性计算方法下，新用户没有邻居可以参考，文献[4]提出用户-物品偏好矩阵是高度稀疏的，因此会导致不准确的推荐。

为了解决数据稀疏问题，提高推荐准确性，研究人员提出了基于用户-物品交互图的模型[5]。用户-物品交互图中存在两种节点类型，即物品和用户。用户-物品交互图中的推荐可以被转换为链接预测问题，链接预测主要是根据特征的相似性和节点之间的其他连接预测两个节点之间发生连接的概率。目前，用户-用户或物品-物品链接的类型被标记为相似或不相似，用户和物品之间的链接类型被标记为喜欢或不喜欢。经过这样的调整，因为只需要将物品推荐给用户，所以喜欢或不喜欢的物品链接变得更加重要。

为了解决兴趣漂移问题，文献[6]提出了基于艾宾浩斯遗忘规律的算法模型，首先通过引入一个指数类型的遗忘函数量化用户对物品的兴趣衰减程度，其次根据当前时间计算出用户对物品的偏好权重，最后进行排序推荐。文献[7]提出了一种结合遗忘曲线和改进相似度的组合推荐算法，通过引入大范围加权因子改进用户相似度，提升高稀疏数据下用户的相似度，再引入遗忘曲线跟踪用户的兴趣漂移，识别用户的短期兴趣与长期兴趣。

1.2 余弦相似度

余弦相似度[8]是常用的相似度算法之一，它被广泛地应用在数据处理的各个领域，通常在协同过滤算法中，评分矩阵的每行向量都代表一个用户，其值代表用户对物品的评分，余弦相似度计算函数定义如下：

$$\cos(\boldsymbol{A},\boldsymbol{B})=\frac{\boldsymbol{A}\times\boldsymbol{B}}{\|\boldsymbol{A}\|\times\|\boldsymbol{B}\|}=\frac{\sum_{i=1}^{n}(A_i\times B_i)}{\sqrt{\sum_{i=1}^{n}A_i^{\,2}}\times\sqrt{\sum_{i=1}^{n}B_i^{\,2}}} \tag{1}$$

1.3 遗忘机制

人脑对新物品产生印象后，会随着时间的推移而慢慢遗忘该物品，而且遗忘速度最初很快，后期会慢慢减缓[9]。根据科学家对遗忘规律的研究，将遗忘函数定义如下：

$$F(t)=\mathrm{e}^{-\frac{\ln 2\times(t-ect)}{hv}} \tag{2}$$

其中，遗忘函数 $F(t)$ 为当前时间用户对新物品记忆的留存程度，t 为当前时间，ect 为用户对物品产生印象的时间，hv 为用户遗忘的速率，根据函数定义，记忆留存程度与 hv 呈反比，hv 越大，则遗忘速度越慢[10]。本文将 hv 定义为当前时间与用户第一次对物品产生记忆的时间之间的差值。将上述理论代入推荐算法，假设用户在 t_1 时刻选择物品 A，在 t_2 时刻选择物品 B，t_2 时刻即推荐列表产生的时间，当前时刻为 t，那么该用户对于物品 B 的感兴趣程度公式如下：

$$F(t)=\mathrm{e}^{-\frac{\ln 2\times(t-t_1)}{t-t_2}} \tag{3}$$

2 算法设计(Algorithm design)

2.1 融合遗忘机制的加权二部图

推荐算法可以简单地表示为一个二部图，在有向网络中，顶点 V 表示物品，U 代表用户，E 表示用户购买物品或是对物品的评分数据等。设 U 为用户集合，G 为物品集合，则 V 为所有用户和物品的并集 $V=U\cup G$。

二部图中的节点会不断更新，其关联程度也会不断改变。在用户-物品链接中，引入了遗忘因子，用来表示用户对当前物品的感兴趣程度。感兴趣程度低，则根据其特征进行推荐的权重变低，感兴趣程度高，则优先进行推荐。

$$F(t)=\mathrm{e}^{-\frac{\ln 2\times(t-et_{ij})}{hv_j}} \tag{4}$$

其中，et_{ij} 表示用户第一次购买基准物品的时间，hv_j 表示遗忘速率，为推荐时间与当前时间的差值。将遗忘机制计算的偏好权重与用户对物品的评级结合，其公式表达如下：

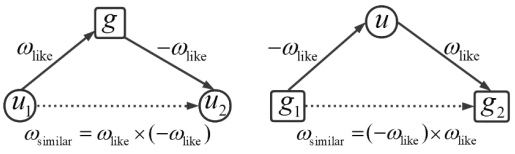
$$\delta_{ij}=\omega_{ij}f_{ij} \tag{5}$$

其中，ω_{ij} 是用户 i 对物品 j 的评级，计算结果代表了推荐权重。

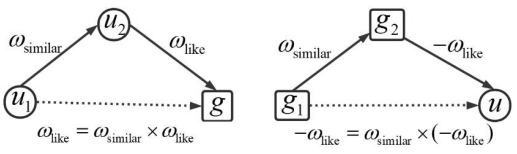
2.2 链接预测算法

在二部图中，如果两个节点之间有连接，那么总是有两个相反方向的链路连接这个节点对，可以通过链路预测图中用户和特定物品之间是否可能会产生链接，之后通过链接预测算法[11]计算任何物品与特定用户之间的相关性。

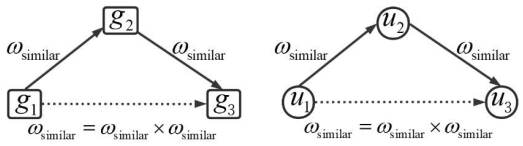
用户-物品二部图的节点或许有两种类型的关系。首先，对于“用户-用户”链接和“物品-物品”链接，两个实体之间都存在一个相似因子，同时由于用户会对物品产生偏好，所以用户 U 对物品 G 之间的链接总有一个从物品 G 到用户 U 的反向链接与之对应。该模型的链接关系如图 1 所示。



(a)用户-物品的相同兴趣产生相似性



(b)类似的用户-物品会有类似的兴趣



(c)用户-物品相似度可传递

图 1　三角形向量乘法规则

Fig. 1 Triangular vector multiplication rule

该模型的链接关系具体表现如下。

(1)用户 u_1 对物品 g 感兴趣,用户 u_2 也对物品 g 感兴趣,那么用户 u_1 和 u_2 可能相似,说明对相同物品感兴趣的用户可能具有相似性链接;用户 u 对物品 g_1 感兴趣,同时对物品 g_2 感兴趣,那么物品 g_1 和物品 g_2 可能相似,说明同一个用户喜欢的不同物品可能具有相似性链接;其公式表达如下:

$$\omega_{similar}=-\omega_{like}^2 \tag{6}$$

(2)用户 u_1 和用户 u_2 相似,用户 u_2 对物品 g 感兴趣,那么用户 u_1 也可能对物品 g 感兴趣,说明相似用户可能会喜欢同一个物品;物品 g_1 和物品 g_2 相似,用户 u 对物品 g_2 感兴趣,那么用户 u 也可能对物品 g_1 感兴趣,说明用户可能会喜欢相似的物品;其公式表达如下:

$$\omega_{like}=\omega_{similar}\times\omega_{like} \tag{7}$$

(3)物品 g_1 和物品 g_2 相似,物品 g_2 和物品 g_3 相似,那么物品 g_1 和物品 g_3 可能相似,说明物品之间的相似性可传递;用户 u_1 和用户 u_2 相似,用户 u_2 和用户 u_3 相似,那么用户 u_1 和用户 u_3 可能相似,说明用户之间的相似性可传递。其规则可用数学表达式表示如下:

$$\omega_{similar}=\omega_{similar}^2 \tag{8}$$

因此,只有找到两个不同的非零整数,才能解上述方程组。假设 $\omega_{similar}=1$ 且 $\omega_{like}=j$,j 是虚数单位,引入复数的概念后,公式(6)至公式(8)分别可以转换成 $1=-j^2$,$j=1\times j$ 和 $1=1^2$。

当用户不喜欢物品时,从用户到物品之间的链接用 $-j$ 进行加权,从物品到用户之间的链接则用 j 加权。相对于相似的链接,只有同时知道链接权重的符号和链接的方向,才能区分用户喜欢与不喜欢,而权重的值则代表了用户不喜欢的程度。最后,将遗忘因子与链接权重相乘,即可以得到新的兴趣度权重。

3　实验结果与分析(Experimental results and analysis)

3.1　实验数据集

本文使用 MovieLens 数据集[12]进行实验,该数据集由 943 名用户、1 682 部电影以及 100 000 条用户对电影的评分组成。

首先,随机选取每个用户评分的 10%的项目创建临时测试集,而临时训练集则包含其他评分。其次,将临时测试集中的"5 星"评分筛选出来作为最终测试集中的评分,将临时测试集中剩余的评分合并到临时训练集中作为最终训练集中的评分。

3.2　评价指标

本文研究的重点是测试集中有多少相关的物品可以推荐给用户,还计算了向所有用户推荐物品的总体比例。因此,比较方法的性能可以通过命中率和覆盖率进行衡量,在 TOPN 推荐的情况下,总体命中率和覆盖率通过平均所有测试用例的结果进行描述。

$$hitsrate(N)=\frac{\sharp hits}{|T|} \tag{9}$$

$$converage(N)=\frac{|\bigcup recommed(N,u)|}{\sharp items} \tag{10}$$

当用户 u 的推荐列表中出现物品 g 时,测试集中的每一对相关的用户-物品对都将获得一次命中,总的命中数用 $\sharp hits$ 表示,测试对的数量用 $|T|$ 表示。因此,命中率可以被定义为向用户推荐物品的能力,覆盖率为系统可以推荐物品的百分比,当这两个指标的值较高时,表示算法的性能较好。

3.3　实验结果

首先,将数据集中的评级转换为复数,得到一个邻接矩阵。其次,应用余弦相似度理论测量数据集中的用户-物品评级矩阵,再引入遗忘机制为用户对物品的喜好度加权。最后,得到用户-用户余弦相似度矩阵和物品-物品余弦相似度矩阵。组合这些矩阵后,将这两个数据集的主要邻接矩阵构造为方阵。节点之间的紧密度值是通过邻接矩阵的幂进行衡量的,因此可以利用双曲正弦函数作为链路预测函数。计算二部图中奇次幂的和,并给出不同长度的最短路径。连接两个节点的路径越多,该函数的得分就越高,两个节点之间的关系也越重要。因此,本文设计了实验测试在不同路径长度下,FMS(Forgetting Mechanism and Cosine)链接预测方法的推荐性能。图 2 和图 3 分别展示了 FMS 算法在路径长度为 3、5、7 和 9 的命中率和覆盖率对比。图 2 表明,随着路径长度的增加,算法的命中率有所降低。显然,当路径长度为 3 时,算法的综合推荐性能更好。

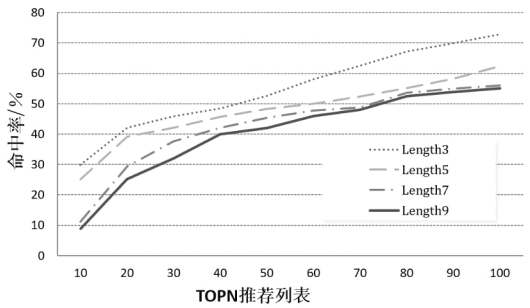


图 2　不同路径长度的命中率比较

Fig. 2 Hit rates comparison of different path lengths

图 3 表明,路径长度为 3 时,算法的覆盖率最高。但是,随着 TOPN 推荐列表的增加,路径长度为 5 的覆盖率逐渐增长到与路径长度为 3 的覆盖率持平,路径长度为 7 或者 9 时,算法覆盖率远远低于路径长度为 3 或者 5。

采用基于物品的 TOPN 推荐算法对 FMS 进行性能评估,将 TOPN 推荐列表的长度从 10 增加到 100,将这些结果与文献[11]中引入的一种复杂项目推荐的链接预测方法 CORLP (Complex Representation-based Link Prediction)进行对比,实验结果如下:图 4 和图 5 分别展示了 FMS 算法和 CORLP 算法在路径长度为 3 和 5 时的命中率和覆盖率的对比。显然,随着 TOPN 推荐列表的增加,命中率和覆盖率也持续增大,并且在路径长度为 3 或者 5 的情况下,FMS 算法的命中率均显著高于 CORLP 算法。

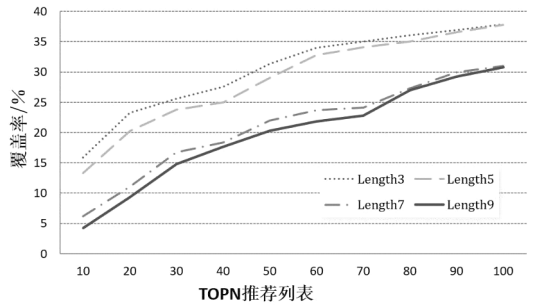


图 3 不同路径长度的覆盖率比较

Fig. 3 Coverage comparison of different path lengths

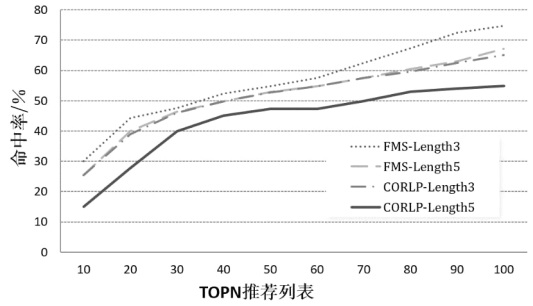


图 4 与 CORLP 算法的命中率比较

Fig. 4 Hit rates comparison of FMS and CORLP

图 5 表明，路径长度为 3 或者 5 的情况下，FMS 算法与 CORLP 算法的覆盖率几乎相同。FMS 算法在路径长度为 3 和 5 时均拥有较好的性能，而 CORLP 算法仅在路径长度为 3 时表现出的综合性能较好。

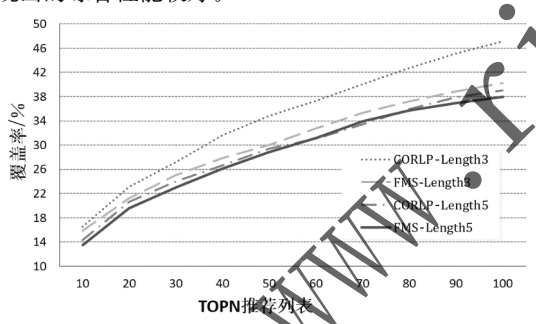


图 5 与 CORLP 算法的覆盖率比较

Fig. 5 Coverage comparison of algorithm FMS and CORLP

4 结论(Conclusion)

为了解决基于相似度的推荐算法在复杂域中未考虑到用户对物品喜好度以及用户“兴趣漂移”的问题，本文研究了用户和物品之间的相似性因素，并通过遗忘因子对用户相似性进行度量，设计了一种结合遗忘机制和用户相似性的推荐算法，通过在 MovieLens 数据集上与其他算法进行对比实验，本文提出的 FMS 算法在覆盖率和命中率方面均优于其他算法。基于图 4 的对比结果，本文提出的算法在 TOP100(前 100 个相似物品)的情况下命中率显著优于 CORLP 算法，在路径长度为 3 时，命中率提高了 10%，在路径长度为 5 时，命中率提高了 12%。在 TOP40 下优势不明显，但在路径长度为 3 和 5 时，命中率依然分别提高了 2%和 6%，经过 20 组数据对比取平均值，本文提出的 FMS 算法相较于其他算法命中率提高了约 7%。通过对链路预测函数进行缩放参数的修改，该方法获得了更高的命中率，在较小的参数规模下具有较高的覆盖率。实验证明，经过改进后的算法在推荐准确性上有了较大的提升，说明用户与物品之间的相似度度量以及遗忘因子对推荐算法性能具有一程度的影响。接下来，可以考虑在算法中融入图像之间的语义关系，设计基于图像语义的推荐系统，进一步提升推荐算法的性能。

参考文献(References)

[1] ISINKAYE F O，FOLAJIMI Y O，OJOKOH B A. Recommendation systems：principles，methods and evaluation[J]. Egyptian Informatics Journal，2015，16(3)：261-273.

[2] MATTHES J，KARSAY K，SCHMUCK D，et al. "Too much to handle"：impact of mobile social networking sites on information overload，depressive symptoms，and well-being[J]. Computers in Human Behavior，2020，105：106217.

[3] 吕琳媛. 复杂网络链路预测[J]. 电子科技大学学报，2010，39(5)：651-661.

[4] ADOMAVICIUS G，TUZHILIN A. Toward the next generation of recommender systems：a survey of the state-of-the-art and possible extensions[J]. IEEE Transactions on Knowledge and Data Engineering，2005，17(6)：734-749.

[5] 陈小辉，高燕. 基于优化欧氏距离的协同过滤推荐[J]. 计算机与现代化，2015(3)：7-40，47.

[6] 金禧，王瑞琴，陆悦聪. 基于艾宾浩斯遗忘曲线和注意力机制的推荐算法[J]. 电信科学，2022，38(10)：89-97.

[7] LI T Y，JIN L L，WU Z B，et al. Combined recommendation algorithm based on improved similarity and forgetting curve[J]. Information，2019，10(4)：130.

[8] 黄川林，鲁艳霞. 基于协同过滤和标签的混合音乐推荐算法研究[J]. 软件工程，2021，24(4)：10-14.

[9] 于洪，李转运. 基于遗忘曲线的协同过滤推荐算法[J]. 南京大学学报(自然科学版)，2010，46(5)：520-527.

[10] 刘晓光，谢晓尧. 一种结合遗忘机制与加权二部图的推荐算法[J]. 河南科技大学学报(自然科学版)，2015，36(3)：48-53.

[11] XIE F，CHEN Z，SHANG J X，et al. A link prediction approach for item recommendation with complex number[J]. Knowledge-Based Systems，2015，81：148-158.

[12] DOOMS S，BELLOGÍN A，DE PESSEMIER T，et al. A framework for dataset benchmarking and its application to a new movie rating dataset[J]. ACM Transactions on Intelligent Systems and Technology，2016，7(3)：1-28.

作者简介：

许 馨(1999-)，女，硕士生。研究领域：推荐算法，智能计算。本文通信作者。

郭家赫(1997-)，男，硕士生。研究领域：推荐算法，智能监控。

乔 宇(1996-)，男，硕士生。研究领域：推荐算法，推荐系统。

舒万能(1981-)，男，博士，副教授。研究领域：智能计算，云计算。