

DOI: 10.11992/tis.201710028

网络出版地址: <http://kns.cnki.net/kcms/detail/23.1538.TP.20180425.1001.006.html>

加权高效用因子的两阶段混合推荐算法

张旭, 孙福振, 方春

(山东理工大学 计算机科学与技术学院, 山东 淄博 255049)

摘要:传统协同过滤算法大多是围绕如何降低评分误差展开研究, 未涉及用户评分过程。本文考虑到用户评分动机和用户本身评分倾向的情况, 将用户评分过程分为用户评分和物品选择两个阶段, 从预测用户兴趣概率和用户效用角度出发, 采用潜在狄利克雷分布模型(LDA)挖掘出用户潜在在高效用因子和物品被靶向概率因子, 进而将两种因子加权融合作为第一阶段; 第二阶段采用奇异值分解预测用户评分值并根据该评分值选择物品。综上, 本文提出一种加权高效用因子的两阶段混合推荐算法(hybrid recommendation algorithm based on two-phase weighted high utility factor, Htp_Uf)。在 MovieLens 数据集上, 实验结果表明, 该算法在归一化累计折损增益(NDCG)和 1-Call 两种评价标准下优于其他 4 种推荐算法, 能够有效提高推荐质量。

关键词:两阶段; 高效用因子; 靶向因子; 主题模型; 用户兴趣; 混合推荐; 用户效用; 评分倾向

中图分类号: TP311 **文献标志码:** A **文章编号:** 1673-4785(2019)03-0518-07

中文引用格式: 张旭, 孙福振, 方春. 加权高效用因子的两阶段混合推荐算法[J]. 智能系统学报, 2019, 14(3): 518-524.

英文引用格式: ZHANG Xu, SUN Fuzhen, FANG Chun. Two-phase weighted high-utility factor-based hybrid recommendation algorithm[J]. CAAI transactions on intelligent systems, 2019, 14(3): 518-524.

Two-phase weighted high-utility factor-based hybrid recommendation algorithm

ZHANG Xu, SUN Fuzhen, FANG Chun

(College of Computer Science and Technology, Shandong University of Technology, Zibo 255049, China)

Abstract: The traditional collaborative filtering algorithm focuses on methods for reducing the rating error, but it does not involve a user rating process. In this paper, we divide the user rating process into two phases, i.e., user rating and item selection, and take into account user-rating motivation and user-rating tendency. From the perspectives of predicting a user's interest probability and user utility, we adopt a latent dirichlet allocation (LDA) to mine a user latent high-utility factor and an article-targeted probabilistic factor, and then weight and fuse these two factors in the first phase. In the second phase, we use singular value decomposition to predict a user rating value, which we then use in user recommendation lists. Our main contribution is our proposal of a two-phase weighted high-utility factor-based hybrid recommendation algorithm, which we abbreviate as Htp_Uf. We compare the effectiveness of the Htp_Uf algorithm with that of others using the MovieLens dataset. The results show that the performance of the proposed algorithm is superior to that of four other recommendation algorithms in terms of two evaluation criteria, namely, the normalized discounted cumulative gain and 1-call, which can effectively improve the quality of the recommendation.

Keywords: two phases; high-utility factor; targeted factor; topic model; user interest; hybrid recommendation; user utility; user rating tendency

推荐系统的核心是推荐算法, 根据算法的学

习目的不同, 主要分为评分预测推荐算法和排序预测推荐算法两大类。文献[1-2]研究指出, 排序预测更符合 TopN 推荐的思想, 排序预测比评分预测能更好地为用户进行 TopN 推荐。随着互联

收稿日期: 2017-10-31. 网络出版日期: 2018-04-25.

基金项目: 国家自然科学基金项目(61602280); 山东省自然科学基金项目(ZR2014FQ028).

通信作者: 孙福振. E-mail: sunfuzhen@sdu.edu.cn.

网信息的爆炸式增长,带来信息过载问题的同时,大量信息背后隐藏着用户潜在行为因子,而传统推荐算法包括业界运用成熟的协同过滤算法^[3],并没有考虑挖掘用户的潜在行为过程。

文献[4-5]运用边缘化去噪自编码的深度学习学习方法学习出一种高效潜在表征方法,没有将用户潜在行为用于深度学习建模中;文献[6]将用户的点击行为与转化率联系在一起构建了动态集体矩阵分解模型,有效利用了用户的显性点击行为而没有关注用户隐性行为;文献[7]提出了一种基于本体论与时序模式挖掘的混合推荐算法以此来提高 TopN 推荐质量;文献[8-9]采用概率矩阵分解技术,将用户评分信息和社交网络信息整合,挖掘出额外信息可以很好地解决预测精度低的问题,未考虑到用户兴趣因子情况;文献[10]使用物品协变量和物品标签等个性化物品潜在的主题,提出了随机变分贝叶斯(SVB)框架和一种协变量为导向的多异性监督主题模型(CGHSTM);文献[11]考虑到用户对物品各个属性面的偏好,利用 LDA 模型挖掘用户 K 个属性面提出了一种 SACF 算法;文献[12]为了提高推荐多样性,将用户的主题模型和应用的主题模型与 MF 相结合的 LDA_MF 模型融合提出了一种混合推荐算法。上述几个模型有效缓解了可扩展性问题,提高了推荐多样性,而未考虑兴趣因子以及高效用因子等用户潜在行为因素。

综上,为有效挖掘用户兴趣偏好,本文考虑到用户的潜在行为过程,即用户评分和物品选择两个阶段。通过分析用户历史行为数据,采用用户 topic 模型进行训练,挖掘出用户潜在高效用因子和物品被靶向概率因子,提出了一种加权用户潜在高效用因子的两阶段混合推荐算法。

1 算法描述

1.1 用户 topic 模型

考虑到一个用户不可能只对一种物品感兴趣,一种物品通常有多重属性,不仅归属于单一类型。一个用户不一定对所有物品评高分,而被评高分的物品通常不仅源自单一用户等情况。因此,本文选择 LDA 主题模型对用户历史动作特征进行训练,进而挖掘出用户潜在高效用因子与物品被靶向概率因子, LDA 描述文档的生成概率为^[13]

$$P(w|d) = \sum_z P(z|d) \cdot P(w|z) \quad (1)$$

式中: $P(w|d)$ 为文档 d 已知的情况下词 w 出现的条件概率; $P(z|d)$ 为文档 d 已知的情况下主题 z 出

现的条件概率; $P(w|z)$ 表示在主题 z 下选择词 w 的条件概率。本文将 LDA 融入到 Htp_Uf 算法中主要考虑 2 种因子潜在高效用因子和物品被靶向概率因子,即 $\text{user} \rightarrow \text{interest} \rightarrow \text{item}$ 和 $\text{user} \rightarrow \text{utility} \rightarrow \text{value}$, 详细流程如图 1 所示。

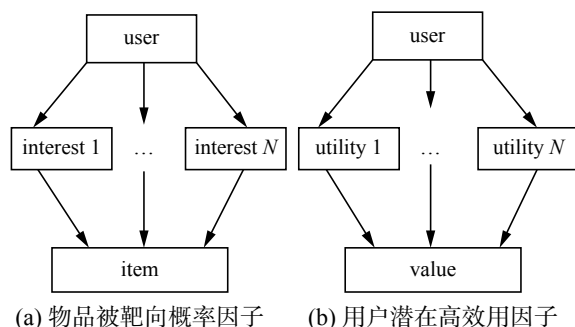


图 1 两种因子流程图

Fig. 1 Flowchart of two factors

图 1 中, user 表示用户, interest 表示用户兴趣主题, item 表示选择的物品, utility 表示效用, value 表示用户评分值。其中, 图 1(a) 表示物品被靶向概率因子, 图 1(b) 表示用户潜在高效用因子。

1.2 用户潜在高效用因子模型

本文认为用户选择物品后并决定对该物品评分之前还要有一个动作, 即衡量评分后这个物品会给用户自身带来多少潜在效用。然后根据这个效用再结合该物品是否符合用户兴趣选择物品给出具体评分值。

定义 1 用户潜在高效用因子。在推荐系统中, 用户在花销时间成本的条件下对物品进行评分时, 不仅会考虑物品是否符合自己的兴趣, 还会考虑评分给自身带来的效用, 即用户满意度。显然效用直接决定用户评分值。而本文将这种效用定义为用户潜在效用因子 (latent utility factor, LUF)。为避免抽取单个效用因子造成的偏置现象, 本文抽取其中两个高概率因子进行累加, 累加的结果称为用户潜在高效用因子 (latent high utility factor, LHUF)。综上, 可根据用户的历史评分行为来预测用户的潜在高效用因子。用户潜在高效用因子数学范式为

$$h_pf(u) \triangleq \sum_{v \in [V-1, V]}^u P(r_v) \leftarrow (U, I, V, \{R_{u,i}\}_{i \in I, u \in U}, \{P(r_v)\}_{v \in V}) \quad (2)$$

式中: U 表示用户集合; I 为物品集合; $R_{u,i}$ 为用户 u 对物品 i 的具体评分值; $P(r_v)$ 为编号 u 的用户对每个评分值预测可能评价的概率值; V 为评分区间最大值。由式 (2) 可知, 核心是 $P(r_v)$ 数值的预测, 下面将通过 LDA 模型对 $P(r_v)$ 值进行预测。

受启发于 LDA 模型在文本挖掘中加入主题

维度,在推荐系统用户-评分值二分关联关系中加入效用维度,则变为用户、效用、评分值三分关联关系。相应的用户打分值概率为

$$\hat{P}(\text{user, value})_{\text{LDA}} = P(\text{value}|\text{user}) = \sum_{\text{benefits}} P(\text{benefits}|\text{user}) \cdot P(\text{value}|\text{benefits}) \quad (3)$$

式中: $P(\text{benefits}|\text{user})$ 表示用户 user 选择效用为 benefits 的概率; $P(\text{value}|\text{benefits})$ 表示在效用为 benefits 的情况下用户 user 评分为 value 的概率。

在上述过程中,一个观测数据 (value|user) 的生成过程解释为用户 user 根据效用评价具体分数为 value 的过程。

该过程为由已知隐含变量参数,生成观测数据的过程,为了充分挖掘这些信息,本文采用 LDA 模型训练数据。在 LDA 模型中的训练方法有变分法 (variational inference) 和 Gibbs 抽样法 (gibbs sampling)^[14]等,本文采用 Gibbs 抽样法。Gibbs 用户效用抽样算法详细描述如下。

1) 随机初始化: 为所有用户评分行为中的具体评分值随机分配一个初始的效用。

2) 重新扫描观测数据,对其中的每个评分过程进行记录,基于式 (4) 利用观测数据中的其他行为信息,估计其条件概率,重新采样该评分值的效用。同时更新模型参数。

$$P(\sigma_{u\xi} = \sigma | \sigma_{\neg <u, \xi>}, \xi) \propto \frac{SV_{u\sigma, \neg <u, \xi>} + \delta}{\sum_{k=1}^{T_p} SV_{uk, \neg <u, \xi>} + T_p \delta} \cdot \frac{SW_{\sigma\xi, \neg <u, \xi>} + \eta}{\sum_{j=1}^n SW_{\sigma j, \neg <u, \xi>} + n\eta} \quad (4)$$

式中: $\sigma_{u\xi}$ 为用户 u 给出的评分值为 ξ 所对应的效用; $SV_{u\sigma}$ 为用户 u 所给出的所有评分值被抽样为效用 σ 的数量; $SW_{\sigma j}$ 为评分值 j 被抽样为效用 σ 的次数; $\neg <u, \xi>$ 为不考虑用户 u 对具体评分值为 ξ 时的动作。

3) 重复执行上述采样过程直到 Gibbs 抽样收敛。

4) 统计 Gibbs 抽样收敛后的效用-分共存概率矩阵,利用式 (5) 和式 (6) 计算 LDA 模型的参数。

$$\lambda_{u\sigma} = P(\sigma|u) = \frac{SV_{u\sigma} + \delta}{\sum_{k=1}^{T_p} SV_{uk} + T_p \delta} \quad (5)$$

$$\omega_{\sigma\xi} = P(\xi|\sigma) = \frac{SW_{\sigma\xi} + \eta}{\sum_{j=1}^n SW_{\sigma j} + n\eta} \quad (6)$$

式中: δ 和 η 分别为 λ 和 ω 的超参数。

综上,预测用户 user 根据效用 utility 评价评分值为 value 的概率记为 $\hat{P}_{\text{LDA}}^{\Delta \text{LUF}}$, 见式 (7) 和式 (8):

$$\hat{P}_{\text{LDA}}^{\Delta \text{LUF}} = \sum_{v \in [V-1, V]} P(r_v) \quad (7)$$

$$P(r_v) = \sum_{\text{utility}} p(\text{utility}|\text{user}) \cdot p(\text{value}|\text{utility}) = \sum_{\sigma=1}^{T_p} \lambda_{u\sigma} \cdot \omega_{\sigma\xi} \quad (8)$$

式中 v 为评分区间最大值。

1.3 物品被靶向概率因子模型

根据用户的历史行为预测用户对每个物品感兴趣的概率。

定义 2 物品被靶向概率因子。每个用户不可能对所有物品都有过评分行为,为挖掘用户潜在兴趣,根据用户历史数据预测用户对每个物品进行评价的概率,本文将此概率定义为物品被靶向概率因子。

物品被靶向概率因子数学范式为

$$I_{\text{Tf}}(u) \triangleq (U, I, \{P_{u,i}\}_{u \in U, i \in I}) \quad (9)$$

式中: U 表示用户集合; I 表示物品集合; $P_{u,i}$ 表示预测用户 u 对物品 i 进行评价的概率。借鉴于上节潜在高效用因子抽取,推导 $P_{u,i}$ 的计算范式,如式 (10)~(12):

$$P(z_{ui} = z | z_{\neg <u, i>}, i) \propto \frac{PT_{uz, \neg <u, i>} + \mu}{\sum_{k=1}^M PT_{uk, \neg <u, i>} + M\mu} \cdot \frac{PF_{zi, \neg <u, i>} + \psi}{\sum_{j=1}^n PF_{zj, \neg <u, i>} + n\psi} \quad (10)$$

$$\ell_{uz} = P(z|u) = \frac{PT_{uz, \neg <u, i>} + \mu}{\sum_{k=1}^M PT_{uk, \neg <u, i>} + M\mu} \quad (11)$$

$$\tau_{zi} = P(i|z) = \frac{PF_{zi, \neg <u, i>} + \psi}{\sum_{j=1}^n PF_{zj, \neg <u, i>} + n\psi} \quad (12)$$

式中: n 是物品的总数量; z_{ui} 表示用户 u 对物品 i 的评分记录对应的兴趣因子; PT_{uz} 是用户 u 的评分记录产品被抽样为兴趣因子 z 的个数; PF_{zj} 表示产品 j 被抽样为兴趣因子 z 的次数; $\neg <u, i>$ 表示不考虑用户 u 对产品 i 的评分行为; μ 和 ψ 分别为 ℓ 和 τ 的超参数。综上,通过模型训练可以计算物品被靶向概率因子,即

$$P_{u,i} = \sum_z P(z|u) \cdot P(i|z) = \sum_{z=1}^M \ell_{uz} \cdot \tau_{zi} \quad (13)$$

1.4 两种因子的融合模型

根据定义 1、2 分别计算出用户潜在高效用因子与物品被靶向概率因子。用户是否评高分不仅取决于被推荐物品与兴趣吻合度,更取决于被推荐物品给用户带来满足程度,即用户潜在高效用。因此,本文认为用户根据兴趣选择高概率物品进行评分且具有高效用因子的情况下预测物品评分,推荐效果会更好。

算法时间复杂度分析如下:由式(7)知,用户潜在高效用因子采用 topic 模型抽样收敛后所获得的值为 $\hat{P}_{LDA}^{\Delta LUF}$, 假设进行一次 Gibbs 抽样作为一个操作单位,记为 t_1 ,总计执行 M 次,则算法时间复杂度为 $O(M \times t_1)$ 。同理,由式(13)知,物品被靶向概率因子通过采样所获得的值为 $P_{u,i}$,倘若进行一次 Gibbs 抽样作为一个单位操作,记为 t_2 ,总计执行 N 次,则算法时间复杂度为 $O(N \times t_2)$ 。故两种因子线性相加之后,算法时间复杂度为 $O(M \times t_1 + N \times t_2)$ 。算法的时间复杂度并没有呈指数级明显增加。

本文将两种因子进行线性加权融合记为 HRe_FP,作为两阶段推荐中的第一阶段。

$$\text{HRe_FP} = \alpha \cdot \hat{P}_{LDA}^{\Delta LUF} + \beta \cdot P_{u,i} \quad (14)$$

式中参数 $\alpha = 0.5$ 、 $\beta = 0.5$ 时取得最优实验结果。

1.5 两阶段融合推荐

推荐系统的核心问题是给用户推荐他最感兴趣的几个物品,而不仅仅是预测用户对物品的评分,将评分高的物品推荐给用户。为了更好地挖掘出用户兴趣,重视评分过程,本文将用户评分过程分为用户评分和物品选择两个阶段,将兴趣度排序值作为用户推荐的依据更符合推荐算法设计的初衷。同时为不显著增加算法的时间复杂度,本文将两阶段线性融合。

首先采用奇异值分解计算预测评分值 \hat{r}_{ui} ;进而将第一阶段计算所得融合因子 HRe_FP 与 \hat{r}_{ui} 乘积作为用户对物品兴趣度排序值,见式(15)、式(16):

$$\hat{r}_{ui} = \mu + b_i + b_u + q_i^T p_u \quad (15)$$

$$\text{RecOrderValue}_{u,i} = \text{HRe_FP} \times \hat{r}_{ui} \quad (16)$$

式中: μ 表示所有物品的平均评分; b_u 表示第 u 个用户偏离平均分的程度; b_i 表示第 i 个物品偏离平均分的程度; p_u 表示第 u 个用户的主题偏好度; q_i 表示第 i 个物品的主题属向度。HRe_FP 表示 2.4 小节中第一阶段值。

2 实验分析

2.1 实验数据集

本文实验所用的数据集是 Minnesota 大学 GroupLens 小组开发的 MovieLens 数据集。实验所采用的数据集规模为 943 个用户对 1 682 部电影的 10 万条评分数据和 6 040 个用户对 3 900 部电影的 100 万条评分数据。

2.2 实验评价标准

推荐算法的评价标准总体上来说一般有两种:一种是与顺序无关的评价标准,例如均方根误差、平均绝对误差、K-Call^[15]等;另一种是与顺序相关的评价标准,例如归一化累计折损增益(normalized discounted cumulative gain, NDCG^[16])。本文分别采用 1-call($K=1$ 时)和 NDCG 作为实验评价标准。具体描述为:K-Call 表示 TopN 推荐结果中至少含有 K 个相关产品的用户所占有的比例。K-Call 的公式为

$$K\text{-Call}@N = \frac{1}{|U|} \sum_{u \in U} f(|L_N(u)| \geq K) \quad (17)$$

式中: $f(\cdot)$ 表示布尔函数,变量为真时函数取值为 1,为假时取值为 0。在实际应用 K-Call 中,通常选择 $K=1$ 作为评价标准,即 1-Call 表示评价推荐算法在 TopN 推荐中包含至少一个相关产品的能力。NDCG 评价标准见式(18)~(20):

$$\text{DCG}@N(u) = \sum_{i=1}^N \frac{2^{R(u,i)} - 1}{\log(1 + i)} \quad (18)$$

$$\text{IDCG}@N(u) = \text{Max}(\text{DCG}@N(u)) \quad (19)$$

$$\text{NDCG}@N = \frac{1}{|U|} \sum_{u \in U} \frac{\text{DCG}@N(u)}{\text{IDCG}@N(u)} \quad (20)$$

式中: $R(u,i)$ 是用户 u 对推荐列表中第 i 个产品的评分; $\text{IDCG}@N(u)$ 是 $\text{DCG}@N(u)$ 的最大值; U 是测试集中所有用户集合; N 表示推荐列表的长度。

2.3 实验结果

本小节主要是将本文提出的 Htp_Uf 算法与 4 种基线算法在两种评价标准下的对比,以及各个参数选取对 Htp_Uf 算法的影响。4 种基线算法分别为:基于用户的协同过滤推荐(User_Based),基于物品的协同过滤推荐(Item_Based)^[17]、SVD 推荐算法^[18]、主题模型和矩阵分解模型的混合推荐算法(HTMMF)^[19]。

2.3.1 Htp_Uf 算法与 4 种基线算法在 NDCG 评价标准下的对比

实验中 Htp_Uf 算法各参数取值如下: $\mu=0.2$, $\psi=0.1$, $\delta=0.2$, $\eta=0.1$ 。

图 2 主要是将本文提出的 Htp_Uf 算法与基于用户的协同过滤算法(User_Based, 近邻数为 50),基于物品的协同过滤算法(Item_Based),奇异值分解(SVD),以及主题模型和矩阵分解模型的混合算法(HTMMF)在 NDCG 评价标准下进行的比较。

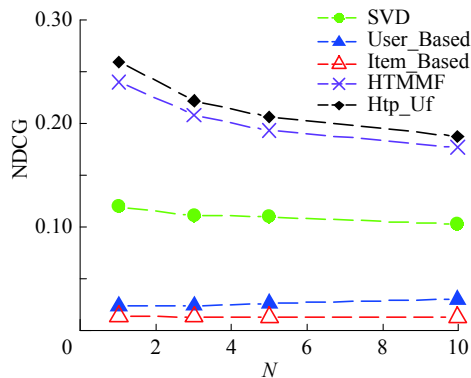


图2 Htp_Uf与4种基线算法NDCG对比

Fig. 2 Diagram comparing performances of Htp_Uf and four baseline algorithms by NDCG

图2中, N 表示推荐列表长度。对比发现Htp_Uf算法相比其他4种算法效果更好,特别是与User_Based和Item_Based相比,提升效果呈指数增长。其主要原因为: Htp_Uf算法将用户潜在高效因子与物品被靶向因子用于推荐算法第一阶段,不仅预测用户对物品评分的概率,还考虑到用户潜在高效情况,更适合进行TopN推荐,从而提升推荐算法性能。而HTMMF算法次之,原因是: HTMMF算法在第一阶段中考虑到用户对物品评分的概率,而没有考虑到用户潜在高效因子情况。

由表1中数据知, Htp_Uf算法比Item_Based、User_Based、SVD、HTMMF这4种算法均有不同幅度的提高, HTMMF算法仅次于Htp_Uf算法,而对比Item_Based算法, Htp_Uf算法提升效果呈指数增长,验证了该算法的有效性。从侧面说明了Item_Based算法不适合TopN推荐,其主要原因是: Item_Based算法为了给用户推荐与该用户历史偏好物品相似的物品,而不一定是用户最感兴趣的物品。

表1 Htp_Uf与4种基线算法NDCG提升率比较

Table 1 Comparison of increasing rates by Htp_Uf and four baseline algorithms by NDCG %

算法	$N=1$	$N=3$	$N=5$	$N=10$
Item_Based	2 919	2 359	2 288	2 042
User_Based	1 214	1 015	829	620
SVD	121.29	103.38	91.23	86.40
HTMMF	8.14	6.81	6.82	5.87

2.3.2 Htp_Uf算法与4种基线算法在1-Call评价标准下对比

实验中Htp_Uf算法各参数取值为: $\mu=0.2$, $\psi=0.1$, $\delta=0.5$, $\eta=0.01$ 。

由图3对比可知, Htp_Uf算法在1-Call评价标准下,相比SVD算法至少提高33.12%,相比HTMMF算法提高5.8%,相比User_Based与

Item_Based呈指数级地提升。综上, Htp_Uf算法在至少推荐一个相关物品的能力上性能更好,也验证了该算法的有效性。

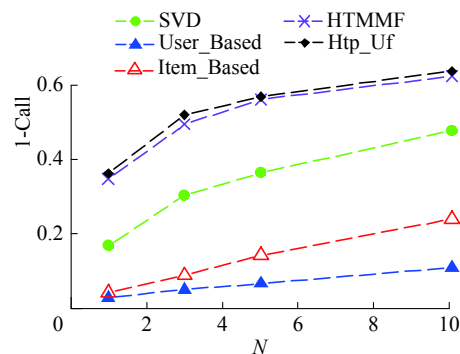


图3 Htp_Uf与4种基线算法1-call对比图

Fig. 3 Diagram comparing Htp_Uf and four baseline algorithms by 1-Call

图4为在MovieLens 1M数据集上, Htp_Uf算法与其他4种算法在NDCG评价标准下效果对比图。该数据集包含6 040个用户对3 900个物品100万条评分记录,由图4知, Htp_Uf算法在NDCG评价标准下相比其他4种算法有显著提高。

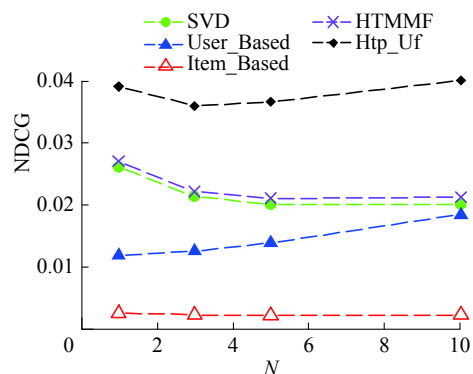


图4 Htp_Uf与4种基线算法NDCG对比图

Fig. 4 Diagram comparing Htp_Uf and four baseline algorithms by NDCG

由表2知, Htp_Uf算法在不同的推荐列表长度下,采用1-Call评价标准,相比其他4种基线算法均有不同程度的提高。由表1和表2联合说明, Htp_Uf算法在与顺序有关的评价标准下的提升率要明显高于顺序无关下的提升率,符合该算法的设计初衷。

表2 Htp_Uf与4种基线算法1-Call提升率比较

Table 2 Comparison of increasing rates for Htp_Uf and four baseline algorithms by 1-Call %

算法	$N=1$	$N=3$	$N=5$	$N=10$
Item_Based	1 821	1 156	881	529
User_Based	1 041	544	321	171
SVD	125.31	71.67	57.18	33.12
HTMMF	5.80	4.04	1.25	1.28

2.3.3 参数 ψ 对 Htp_Uf 算法的影响

固定参数 $\mu=0.2, \delta=0.2, \eta=0.1$ 时, 检验参数 ψ 的变化对 Htp_Uf 算法的影响。

由图5知, 参数 ψ 从 0.1 开始, 以步长为 0.1 递增至 0.9。纵向分析, 随着推荐列表长度的增加, NDCG 值递减, 反之亦然, 即 NDCG 值的大小与列表长度成反比。横向分析, 当推荐列表长度为 1, ψ 为 0.1 时, NDCG 值最高, ψ 为 0.6 时次之。当推荐列表长度为 3、5、10, ψ 从 0.1 增加到 0.5 时, NDCG 值呈递减趋势, 即 NDCG 值的大小与参数 ψ 大小成反比。因此, 不论推荐列表长度为多少, 参数 ψ 取值为 0.1 最好。

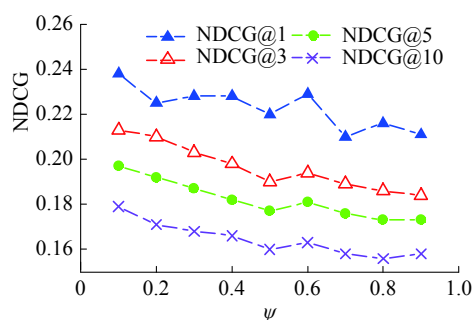


图5 参数 ψ 在 NDCG 标准下对 Htp_Uf 的影响

Fig. 5 Influence of ψ parameter variation to model of Htp_Uf by NDCG

2.3.4 参数 μ 对 Htp_Uf 算法的影响

固定参数 $\psi=0.2, \delta=0.2, \eta=0.1$ 时, 检验参数 μ 的变化对 Htp_Uf 算法的影响。

由图6知, 当推荐列表长度为 1 时, 为使 NDCG 值最高, $\mu=0.1$ 为最佳选择。当推荐列表长度分别为 3、5、10 时, NDCG 值呈先增后减的趋势; 当推荐列表长度为 3、5 时, $\mu=0.2$ 最优; 当推荐列表长度为 10 时, $\mu=0.3$ 最佳。

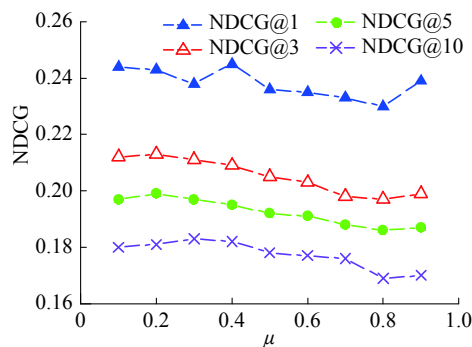


图6 参数 μ 在 NDCG 标准下对 Htp_Uf 的影响

Fig. 6 Influence of μ parameter variation on NDCG's model of Htp_Uf

3 结束语

受经济学领域中效用函数的启发, 同时考虑

到用户对物品感兴趣但不一定评高分的情况, 本文将用户评分过程分为用户评分和选择产品两个阶段。将用户评分行为中抽象出用户潜在高效用因子和物品被靶向概率因子, 并且将这两种因子加权作为两阶段推荐的第一阶段, 采用奇异值分解因子模型来预测具体评分值作为第二阶段, 将两阶段融合形成一种加权高效用因子的两阶段混合推荐算法, 实验结果表明该算法提高了 TopN 推荐质量。

本文提出的两阶段混合推荐算法在挖掘高效用因子和靶向物品概率因子计算方面, 都不同程度地受到用户历史评分矩阵数据缺失的影响, 也存在用户冷启动, 物品冷启动, 数据稀疏性等问题, 这些问题将是我们未来研究工作的重点。

参考文献:

- [1] WEIMER M, KARATZOGLOU A, LE Q V, et al. COFIR-ANK maximum margin matrix factorization for collaborative ranking[C]//Proceedings of the 20th International Conference on Neural Information Processing Systems. Vancouver, British Columbia, Canada: Curran Associates Inc., 2007: 1593-1600.
- [2] LIU N N, ZHAO Min, YANG Qiang. Probabilistic latent preference analysis for collaborative filtering[C]//Proceedings of the 18th ACM Conference on Information and Knowledge Management. Hong Kong, China: ACM, 2009: 759-766.
- [3] BREESE J S, HECKERMAN D, KADIE C. Empirical analysis of predictive algorithms for collaborative filtering[C]//Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence. Madison, Wisconsin: Morgan Kaufmann Publishers Inc., 1998: 43-52.
- [4] LI Sheng, KAWALE J, FU Yun. Deep collaborative filtering via marginalized denoising auto-encoder[C]//Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. Melbourne, Australia: ACM, 2015: 811-820.
- [5] KAŠŠÁK O, KOMPAN M, BIELIKOVÁ M. Personalized hybrid recommendation for group of users: Top-N multimedia recommender[J]. Information processing and management, 2016, 52(3): 459-477.
- [6] LI Sheng, KAWALE J, FU Yun. Predicting user behavior in display advertising via dynamic collective matrix factorization[C]//Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information. Santiago, Chile: ACM, 2015: 875-878.
- [7] TARUS J K, NIU Zhendong, YOUSIF A. A hybrid knowledge-based recommender system for e-learning based on

- ontology and sequential pattern mining[J]. *Future generation computer systems*, 2017, 72: 37–48.
- [8] KOREN Y, BELL R, VOLINSKY C. Matrix factorization techniques for recommender systems[J]. *Computer*, 2009, 42(8): 30–37.
- [9] MA Hao, ZHOU Dengyong, LIU Chao, et al. Recommender systems with social regularization[C]//Proceedings of the 4th ACM International Conference on Web Search and Data Mining. Hong Kong, China: ACM, 2011: 287–296.
- [10] ANSARI A, LI Yang, ZHANG J Z. Probabilistic topic model for hybrid recommender systems: a stochastic variational Bayesian approach[D]. New York: Columbia Business School, 2017.
- [11] 彭敏, 席俊杰, 代心媛, 等. 基于情感分析和 LDA 主题模型的协同过滤推荐算法[J]. *中文信息学报*, 2017, 31(2): 194–203.
- PENG Min, XI Junjie, DAI Xinyuan, et al. Collaborative filtering recommendation based on sentiment analysis and LDA topic model[J]. *Journal of Chinese information processing*, 2017, 31(2): 194–203.
- [12] 黄璐, 林川杰, 何军, 等. 融合主题模型和协同过滤的多样化移动应用推荐[J]. *软件学报*, 2017, 28(3): 708–720.
- HUANG Lu, LIN Chuanjie, HE Jun, et al. Diversified mobile app recommendation combining topic model and collaborative filtering[J]. *Journal of software*, 2017, 28(3): 708–720.
- [13] BLEI D M, NG A Y, JORDAN M I, et al. Latent dirichlet allocation[J]. *Journal of machine learning research*, 2003, 3(4/5): 993–1022.
- [14] GRIFFITHS T L, STEYVERS M. Finding scientific topics[J]. *Proceedings of the national academy of sciences of the United States of America*, 2004, 101(S1): 5228–5235.
- [15] SHI Yue, KARATZOGLOU A, BALTRUNAS L, et al. CLiMF: learning to maximize reciprocal rank with collaborative less-is-more filtering[C]//Proceedings of the 6th ACM conference on recommender systems. Dublin, Ireland: ACM, 2012: 139–146.
- [16] JÄRVELIN K, KEKÄLÄINEN K. Cumulated gain-based evaluation of IR techniques[J]. *ACM transactions on information systems (TOIS)*, 2002, 20(4): 422–446.
- [17] LINDEN G, SMITH B, YORK J. Amazon. com recommendations: item-to-item collaborative filtering[J]. *IEEE internet computing*, 2003, 7(1): 76–80.
- [18] SYMEONIDIS P. Content-based dimensionality reduction for recommender systems[M]//PREISACH C, BURKHARDT H, SCHMIDT-THIEME L, et al. *Data Analysis, Machine Learning and Applications*. Berlin, Heidelberg: Springer, 2008: 619–626.
- [19] ZHAO Xiangyu, NIU Zhendong, CHEN Wei, et al. A hybrid approach of topic model and matrix factorization based on two-step recommendation framework[J]. *Journal of intelligent information systems*, 2015, 44(3): 335–353.

作者简介:



张旭, 男, 1991 年生, 硕士研究生, 主要研究方向为智能信息处理、推荐系统。



孙福振, 男, 1978 年生, 副教授, 博士, 主要研究方向为信息检索与数据挖掘、推荐系统、话题检测与热点跟踪。授权国家发明专利 6 项。发表学术论文 30 余篇。



方春, 女, 1981 年生, 讲师, 博士, 主要研究方向为智能计算、模式识别、生物医学研究。发表学术论文 10 余篇。