

DOI: 10.7652/xjtuxb201708007

融合 Fisher 线性判别分析的多维特征 融合情景感知推荐方法

赵志华^{1,2}, 陈莉¹

(1. 西北大学信息科学与技术学院, 710127, 西安; 2. 山西水利职业技术学院信息工程系, 044004, 山西运城)

摘要: 针对采用单维特征建立用户的偏好模型所导致的推荐结果无法有效覆盖用户潜在偏好特征而影响推荐质量的问题,提出了一种基于 Fisher 线性判别分析的多维特征融合情景感知推荐方法。该方法建立了包含时间衰减度、属性偏好、偏好可影响程度等多维特征的偏好样本空间;采用特征融合、投影变换等方法,在最佳鉴别矢量空间基于 Fisher 判别准则融合用户的多维特征;采用拉格朗日乘子法求解最优投影方向,建立起多维特征优化的偏好获取模型。在 BookCrossing 与 Netfilix 数据集上的实验结果表明:与现有方法相比,所提方法的推荐准确率平均提高了 16.61%,多样性平均提高了约 38.01%,能够有效地覆盖用户的潜在偏好特征,并取得更好的推荐质量。

关键词: 多特征融合;Fisher 线性判别分析;属性偏好;时间衰减;情景感知推荐

中图分类号: TP391.9 **文献标志码:** A **文章编号:** 0253-987X(2017)08-0040-07

A Context-Aware Recommendation Method with Multi-Feature Fusion Based on Fisher Linear Discriminant Analysis

ZHAO Zhihua^{1,2}, CHEN Li¹

(1. School of Information Science and Technology, Northwest University, Xi'an 710127, China;
2. Department of Information Engineering, Shanxi Water Technical Professional College, Yuncheng, Shanxi 044004, China)

Abstract: A context-aware recommendation method with multi-feature fusion based on Fisher linear discriminant analysis is proposed to solve the problem that the recommendation result does not cover user's potential preference so the recommendation quality is influenced when prediction methods only acquire user's preference from the single view data. This method establishes a sample space of preference data, including the degree of time attenuation, attribute preference and the degree of behavior influence. The methods of feature fusion and projection transformation are used to fuse users' multidimensional features in an optimal vector space based on Fisher discriminant criterion. Then, the Lagrange multiplier method is employed to compute the optimal projection direction, and a users' preference model is constructed. Experimental results on data sets of BookCrossing and Netfilix and comparisons with existing methods show that the recommendation accuracy and diversity of the proposed method improve by 16.61% and 38.01%, respectively. These results indicate that the proposed method can effectively cover users' potential preference and achieve better prediction quality.

Keywords: multi-feature fusion; Fisher linear discriminant analysis; attribute preference; time attenuation; context-aware recommendation

智能推荐系统是大数据背景下缓解信息过载问题的主要支撑技术,可以分为基于内容的推荐方法与基于数学模型的推荐方法。具体研究包括:融合项目内容过滤的推荐策略、基于项目间隐性信任关系的推荐策略、结合图模型的推荐策略等。上述方法所面临的数据稀疏性^[1]和冷启动^[2]等问题,一直困扰着研究人员和服务提供商。由于这些推荐算法仅采用“用户-项目”二元关系为用户建立偏好模型,未考虑上下文环境对于用户偏好的影响,也在一定程度上影响了推荐质量。

随着普适计算、泛在计算等新型计算模式的发展,系统能够实时地获取用户状态、环境信息等内容,用户的偏好会受其所在上下文环境的影响而产生波动^[2],在获取偏好时考虑用户属性、环境信息等能够提高偏好预测的准确性^[3]。本文提出了一种融合用户多维特征的情景感知推荐方法,将多视图学习相关算法引入情景感知推荐流程,改变了现有情景感知算法仅采用单维特征的计算方式,提高了影响用户偏好的上下文信息融合程度,建立起更加全面的用户偏好模型。

1 相关工作

在情景感知推荐领域,大量的研究者都提出了不同的解决方案^[4-5]。Wang 等通过分析用户间的语义关系模型,提出了一种情景感知音乐推荐算法,提高了音乐推荐质量^[6];Amato 等结合云环境与上下文信息建模,提出了一种旅游线路推荐算法,提高了旅游线路推荐效率与准确度^[7]。这些方法多是从用户或者项目等角度结合上下文环境获取用户偏好,即从单维特征的角度生成推荐结果,由于现有数据存在的稀疏性问题,单维特征无法有效地覆盖用户的潜在偏好特征和隐藏的兴趣点,影响了偏好预测的适用性与准确度。

多特征融合主要适用于样本空间可以采用多维特征描述的学习问题,研究起源于 Yarowsky 用多维特征解决图像样本特征分类问题^[8],目前是样本分类、模型优化与半监督学习等领域的热门研究方向。由于推荐的过程可以看作是以用户偏好为分类特征,将相符的项目分类到相应的用户偏好集合中,即推荐的产生可以转换为相关数据的多维特征融合与优化。

多维特征的选取需要全面的涵盖用户的偏好特征才能够取得较好的推荐效果,而用户的偏好是其行为在具体项目上的选择性体现,因此偏好特征应

包含用户偏好、项目 2 个层面。从用户的角度而言,一方面用户的偏好会随时间的变化而变化,在建立用户偏好模型时,需要考虑历史偏好数据产生的时间信息;另一方面,根据协同进化理论,通过对用户偏好影响关系的分析和建模,能够预测用户未来的偏好选择,因此在用户层面本文选取时间衰减和行为可影响程度作为其主要特征。从项目的角度而言,可通过用户对于项目的历史评分值获取用户对于该项目的偏好程度。项目一般由多属性描述,将上述偏好程度转化为用户对项目具体属性的偏好值,在一定程度上能够提高数据密度,因此从项目层面,本文采用属性偏好作为其主要特征。

由于上述特征数据具有非线性可分的特点,导致难以直接找到合适的分类面实现对上述特征进行有效的分类^[9]。线性判别分析能够利用已知类别的样本建立判别模型,当遇到新的样本点时,只要根据相应的判别公式和判别准则,就能判别该样本点所属的类别,即能够有效地解决上述问题。基于上述分析,本文方法采用线性判别分析融合多维特征数据来生成推荐结果,并在真实数据集上进行了实验验证。

2 多维特征融合的情景感知推荐方法

本文方法首先获取时间衰减度、属性偏好、偏好可影响程度等多维特征,通过特征融合、投影变换,在最佳鉴别矢量空间通过 Fisher 判别准则融合用户的多维特征,并采用拉格朗日乘子法求解最优投影方向,建立起多维特征优化的偏好获取模型。

2.1 多维特征获取

本文采用的多维特征包括基于时间衰减的用户偏好、属性偏好、用户基本信息、项目基本信息、用户偏好可影响度等,具体定义描述如下。

定义 1 时间衰减度指用户历史偏好对当前偏好影响衰减程度的度量值,即由于用户的偏好会随时间而改变,用户的历史偏好对当前偏好的影响应考虑该历史偏好产生的时间上下文。影响时间衰减度的因素包括时间偏差、初始偏好值、偏好偏差等。时间衰减度主要描述用户对某类项目的偏好衰减幅度以及总体的衰减度量值两部分。

由于用户的初始偏好越大,随时间衰减的速度越快,因此用户对某类项目的偏好衰减主要包含以下 2 个影响因素:用户对该类项目的偏好值在该用户整体偏好值的分布水平;对该类项目的历史偏好值随时间的变化程度。若 T_{u_i} 表示用户 u 对于项目

类 I_j 的偏好的衰减程度,则

$$T_{uI_j} = (A_{I_j} - A_u) / (A_{I_j} + A_u) + \alpha \left(\sum_{m,n \in I_j} \frac{|P_{um} - P_{un}|}{|t_1 - t_2|} \right) / (C(I_j)(C(I_j) - 1)) \quad (1)$$

式中: A_{I_j} 表示用户 u 对于项目类 I_j 中项目的历史偏好均值; A_u 表示用户 u 对于所有项目的历史偏好均值; α 为修正两影响因素的修正因子; P_{um} 、 P_{un} 分别表示用户 u 对于任意项目 m 与 n 的历史偏好值; t_1 与 t_2 分别表示上述偏好产生的时间; $C(I_j)$ 为 u 对项目类 I_j 中存在历史偏好行为的项目数量。

用户总体的偏好衰减程度则为用户在所有项目类别上的偏好衰减均值,具体度量方法为

$$\overline{T}_u = \sum_{I_j \subset I} T_{uI_j} / (C(I)) \quad (2)$$

式中: $C(I)$ 表示项目集合 I 中项目类别数。

定义 2 属性偏好描述了用户对于描述项目的各个属性的偏好值。由于将用户对于项目的历史偏好转换为对于其相关属性的偏好能够建立更加准确的偏好模型^[7], 本文采用衰减后的偏好值作为度量用户对于属性偏好的基础数据, 若 a_i 表示用户 u 存在历史偏好项目中所包含的任意属性, 则

$$Q_{ua_i} = \sum_{v \in I_{a_i}} P_{uv}^* / (C(P_{ua_i})) \quad (3)$$

式中: Q_{ua_i} 表示用户 u 对于属性 a_i 的属性偏好值; P_{ua_i} 表示用户 u 的历史偏好中包含属性 a_i 项目的历史偏好值; P_{uv}^* 表示衰减后的用户偏好

$$P_{uv}^* = P_{uv} T_{uI_j} \quad (4)$$

定义 3 行为可影响程度指用户的偏好行为受其他相关用户的影响程度, 是从行为层面来度量用户偏好间的关联关系。通过度量用户与 Top-N 近邻的偏好行为关系, 能够反映出当前上下文环境下所建立偏好模型的可信程度, 以及该用户偏好受其他用户的可影响程度。目前行为相似度的度量方法主要有余弦相似度、修正的余弦相似度、泊松相似度等。由于余弦相似度能够减少用户评分尺度对于偏好差异的影响, 本文采用衰减行为相似度, 令 S_u 表示行为可影响程度, $S'(u, d)$ 表示衰减行为相似度, 则

$$S_u = \sum_{d \in U_u} S'(u, d) / (C(U_u)) \quad (5)$$

$$S'(u, d) = \left(\sum_{v \in I_{ud}} (P_{uv}^* - \overline{T}_u)^2 \right)^{-1/2} \cdot \left(\sum_{v \in I_{ud}} (P_{uv}^* - \overline{T}_u)^2 \right)^{-1/2} \cdot \sum_{v \in I_{ud}} (P_{uv}^* - \overline{T}_u)(P_{dv}^* - \overline{T}_u) \quad (6)$$

式中: U_u 表示与用户 u 的历史偏好交集非空的 Top-N 近邻用户集合; I_{ud} 表示用户 u 与 d 有共同偏好的项目集合。

2.2 基于 Fisher 线性判别分析的多维特征融合

由于推荐结果的生成, 可以看作把符合用户偏好的项目分类至用户推荐列表的过程, 因此采用线性判别分析可以获取最佳推荐效果的分类型条件。在获取的基于用户的时间衰减度、属性偏好、行为可影响程度等多维数据的基础上, 将上述数据转换成描述用户偏好的多维特征数据, 通过特征融合、投影变换, 在最佳鉴别矢量空间通过 Fisher 判别准则 (Fisher linear discriminant analysis, FLDA) 融合用户的多维特征, 并采用拉格朗日乘子法求解最优投影方向, 即具有最佳推荐效果的分类型方向。

基于用户的多维特征 (特征样本数为 k) 构成 m 维空间向量, 并可抽象为 m 行 n 列的矩阵 $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_m)^T$, 其中 $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_m$ 表示对矩阵 $\mathbf{A}$ 按行进行分组。本节的目的在于找到最优的线性变换 $\mathbf{G}$, 使得 $\mathbf{G}^T$ 可将 m 维向量空间的任意向量 $\mathbf{a}_i$ ($\mathbf{a}_i$ 为 $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_m$ 中任意行向量) 通过投影变换映射到 l 维空间中的向量 $\mathbf{a}_i^l$, 具体表示如下

$$\mathbf{G}^T: \mathbf{a}_i \rightarrow \mathbf{a}_i^l \quad (7)$$

在给定高维样本空间上, 各分类样本中的类间离散度 $\mathbf{S}_b$ 、类内离散度 $\mathbf{S}_w$ 和总的离散度 $\mathbf{S}_t$ 的定义如下

$$\mathbf{S}_b = \frac{1}{k} \sum_{x=1}^k n_x (\mathbf{c}^{(x)} - \mathbf{c})(\mathbf{c}^{(x)} - \mathbf{c})^T \quad (8)$$

$$\mathbf{S}_w = \frac{1}{k} \sum_{x=1}^k \sum_{\mathbf{a}_i \in A} (\mathbf{a}_i - \mathbf{c}^{(x)})(\mathbf{a}_i - \mathbf{c}^{(x)})^T \quad (9)$$

$$\mathbf{S}_t = \frac{1}{k} \sum_{\mathbf{a}_i \in A} (\mathbf{a}_i - \mathbf{c})(\mathbf{a}_i - \mathbf{c})^T \quad (10)$$

式中: n_x 为 k 个特征样本中第 x ($1 \leq x \leq k$) 个样本的元素数; $\mathbf{c}^{(x)} = \frac{1}{n_x} \mathbf{a}_x \mathbf{e}$ 表示第 x 个样本的样本均值, $\mathbf{e}$ 为单位列向量; $\mathbf{c} = \frac{1}{k} \sum_{x=1}^k \mathbf{c}^{(x)}$ 表示整体的样本均值。那么, 显然存在如下等式成立

$$\mathbf{S}_t = \mathbf{S}_b + \mathbf{S}_w \quad (11)$$

若令

$$\mathbf{H}_b = n^{-1/2} [n_1^{1/2} (\mathbf{c}^{(1)} - \mathbf{c}), n_2^{1/2} (\mathbf{c}^{(2)} - \mathbf{c}), \dots, n_k^{1/2} (\mathbf{c}^{(k)} - \mathbf{c})] \quad (12)$$

$$\mathbf{H}_w = n^{-1/2} [\mathbf{a}_1 - \mathbf{c}^{(1)} \mathbf{e}^T, \mathbf{a}_2 - \mathbf{c}^{(2)} \mathbf{e}^T, \dots, \mathbf{a}_k - \mathbf{c}^{(k)} \mathbf{e}^T] \quad (13)$$

$$\mathbf{H}_t = n^{-1/2} [\mathbf{a}_1 - \mathbf{c}, \mathbf{a}_2 - \mathbf{c}, \mathbf{a}_3 - \mathbf{c}, \dots, \mathbf{a}_k - \mathbf{c}] \quad (14)$$

其中 $n_1, n_2, \dots, n_k$ 分别表示 k 个多维样本的元素数, n 为矩阵 $\mathbf{A}$ 的列数, 那么各离散度矩阵可转换为如下形式

$$\mathbf{S}_b = \mathbf{H}_b \mathbf{H}_b^T; \mathbf{S}_w = \mathbf{H}_w \mathbf{H}_w^T; \mathbf{S}_t = \mathbf{H}_t \mathbf{H}_t^T \quad (15)$$

因为

$$\mathbf{H}_b = n^{-1/2} [n_1^{1/2} (\mathbf{c}^{(1)} - \mathbf{c}), n_2^{1/2} (\mathbf{c}^{(2)} - \mathbf{c}), \dots, n_k^{1/2} (\mathbf{c}^{(k)} - \mathbf{c})] \quad (16)$$

那么

$$\mathbf{H}_b^T = n^{-1/2} \begin{bmatrix} n_1^{1/2} (\mathbf{c}^{(1)} - \mathbf{c})^T \\ n_2^{1/2} (\mathbf{c}^{(2)} - \mathbf{c})^T \\ \vdots \\ n_k^{1/2} (\mathbf{c}^{(k)} - \mathbf{c})^T \end{bmatrix} \quad (17)$$

$$\mathbf{H}_b \mathbf{H}_b^T = n^{-1/2} [n_1^{1/2} (\mathbf{c}^{(1)} - \mathbf{c}), n_2^{1/2} (\mathbf{c}^{(2)} - \mathbf{c}), \dots, n_k^{1/2} (\mathbf{c}^{(k)} - \mathbf{c})] \cdot \begin{bmatrix} n_1^{1/2} (\mathbf{c}^{(1)} - \mathbf{c})^T \\ n_2^{1/2} (\mathbf{c}^{(2)} - \mathbf{c})^T \\ \vdots \\ n_k^{1/2} (\mathbf{c}^{(k)} - \mathbf{c})^T \end{bmatrix} = \frac{1}{n} \sum_{x=1}^k n_x (\mathbf{c}^{(x)} - \mathbf{c}) (\mathbf{c}^{(x)} - \mathbf{c})^T = \mathbf{S}_b \quad (18)$$

同理可知, $\mathbf{S}_w$ 与 $\mathbf{S}_t$ 可被转换为如下形式

$$\mathbf{S}_w = \mathbf{H}_w \mathbf{H}_w^T \quad (19)$$

$$\mathbf{S}_t = \mathbf{H}_t \mathbf{H}_t^T \quad (20)$$

基于离散度矩阵的性质, 各矩阵的迹可表示分类样本的样本结构, 以及各样本空间与总体样本中心(类间离散度)间的偏差程度, 定义如下

$$\begin{aligned} \text{tr}(\mathbf{S}_b) &= \frac{1}{n} \sum_{x=1}^k n_x (\mathbf{c}^{(x)} - \mathbf{c})^T (\mathbf{c}^{(x)} - \mathbf{c}) = \\ &= \frac{1}{n} \sum_{x=1}^k n_x \|\mathbf{c}^{(x)} - \mathbf{c}\|^2 \end{aligned} \quad (21)$$

类内离散度可转换为

$$\begin{aligned} \text{tr}(\mathbf{S}_w) &= \frac{1}{n} \sum_{x=1}^k \sum_{a_i \in A} (a_i - \mathbf{c}^{(x)})^T (a_i - \mathbf{c}^{(x)}) = \\ &= \frac{1}{n} \sum_{x=1}^k \sum_{a_i \in A} \|a_i - \mathbf{c}^{(x)}\|^2 \end{aligned} \quad (22)$$

如果 $\mathbf{G}^T$ 表示任意的线性变换, 基于各离散度的定义, 在样本空间 R^l 中, 投影变换后的各离散度定义形式如下

$$\mathbf{S}_b^l = \mathbf{G}^T \mathbf{S}_b \mathbf{G} \quad (23)$$

$$\mathbf{S}_w^l = \mathbf{G}^T \mathbf{S}_w \mathbf{G} \quad (24)$$

$$\mathbf{S}_t^l = \mathbf{G}^T \mathbf{S}_t \mathbf{G} \quad (25)$$

对于任意的 $\mathbf{G}^T$, 最优的 $\mathbf{G}^*$ 应能取得最大的 $\text{tr}(\mathbf{S}_b^l)$ 以及最小的 $\text{tr}(\mathbf{S}_w^l)$, 即需要满足如下的线性判别准则

$$\mathbf{G}^* = \arg \max \{ \text{tr}((\mathbf{S}_w^l)^{-1} \mathbf{S}_b^l) \} \quad (26)$$

也就是

$$\mathbf{G}^* = \arg \max \{ \text{tr}((\mathbf{S}_t^l)^{-1} \mathbf{S}_b^l) \} \quad (27)$$

线性判别分析的目的就在于采用带 Fisher 判别准则的拉格朗日乘子法求解上述等式, 获得最优的 $\mathbf{G}^*$ 。拉格朗日函数定义如下

$$L(\mathbf{G}, \delta) = \mathbf{G}^T \mathbf{S}_b^l \mathbf{G} - \delta (\mathbf{G}^T \mathbf{S}_w^l \mathbf{G} - b) \quad (28)$$

对 $\mathbf{G}$ 求偏导

$$\frac{\partial L(\mathbf{G}, \delta)}{\partial \mathbf{G}} = 2\mathbf{S}_b^l \mathbf{G} - 2\delta \mathbf{S}_w^l \mathbf{G} \quad (29)$$

令偏导为 0, 得

$$\mathbf{S}_b^l \mathbf{G}' = \delta \mathbf{S}_w^l \mathbf{G}' \quad (30)$$

两边同乘 $\mathbf{S}_w^{l-1}$, 得

$$\mathbf{S}_w^{l-1} \mathbf{S}_b^l \mathbf{G}' = \delta \mathbf{G}' \quad (31)$$

原问题也就转换为求解矩阵 $\mathbf{S}_w^{l-1} \mathbf{S}_b^l$ 的特征值问题, 其最大特征值所对应的特征向量即所求的最佳投影方向, 由此也可推出具有最优分类效果的投影方向, 在此方向上, 对用户偏好项目的分类效果最好, 并基于分类后的项目为用户生成推荐列表。

3 实验设计与结果分析

3.1 数据集

由于在公开的数据集中还没有包含上下文信息的数据集, 本文采用的是扩充 BookCrossing 与 Netflix 数据集。对于 BookCrossing 数据集, 通过加入合理的上下文生成规则, 建立了一个模拟真实数据集 BookCrossing-MN, 其中 BookCrossing 数据集由 Book-Crossing 图书社区上获取的真实用户的图书行为数据组成, 包含 278 858 名读者对于 271 379 本图书的行为记录。

Netflix 数据集是从 Netflix 电影网站上获取的超过 48 万个随机选取的匿名用户, 对于 17 000 多部电影的评分记录, Netflix 主要包括 Training set 和 Movie titles 两部分。Training set 包含电影 ID、用户 ID、评分(1~5)、日期 4 个特征值; Movie titles 包含电影 ID、上映年份、电影标题 3 个特征值。

3.2 评价准则

本文采用平均绝对误差 M 、平衡准确率与召回率的综合评价指标 F_1 作为推荐准确性的度量指标, 其中 M 为预测行为与实际行为间的差异, 用以评价所提出推荐算法的优劣, 其值越小说明提出的推荐算法的预测结果与用户实际偏好间的偏差越小, 推荐准确率越高, F_1 越高说明推荐结果越理想。假设采用推荐算法获取的用户偏好预测值为 $\{p_1, p_2, p_3, \dots, p_N\}$, 而其对应的真实用户偏好分值为

$\{q_1, q_2, q_3, \dots, q_N\}$, 则

$$M = \sum_{i=1}^N |p_i - q_i| / r \tag{32}$$

$$F_1 = 2Pr / (P + r) \tag{33}$$

式中: P 为准确率; r 为召回率。

多样性 D 能够反映推荐结果覆盖用户潜在偏好的概率, 其值越大相应的概率也就越大。假设 $S(i, j)$ 定义了项目 I_i 和 I_j 之间的相似度, $|R|$ 表示推荐列表的长度, 则

$$D = 1 - 2 \sum_{I_i, I_j \in R} s(i, j) / (|R| (|R| - 1)) \tag{34}$$

3.3 实验设计与结果分析

本文所采用的 BookCrossing-MN 数据集比 Netflix 数据集拥有更多的属性特征, 因此本文实验以 Book Crossing-MN 数据集为主, 在 BookCrossing-MN 进行参数检验, 在 2 个数据集上都与其他算法进行对比, 具体实验过程与分析如下。

实验 1 参数 α 对实验的影响。

参数 α 是偏好衰减程度的修正参数, 能够修正项目类以及单位时间内偏好波动幅度对于用户偏好衰减程度的影响。与原有历史偏好相比, 用户偏好更能够反映出用户真实的偏好信息, 也可以更加准确地反映出用户偏好间的相似关系。通过度量偏好衰减前后的行为相似度变化情况, 以获取当前实验环境下的最优 α 值。修正的余弦相似度度量方法能够排除用户评价尺度的影响, 采用修正的余弦相似度作为获取最优 α 值的行为相似度方法。经过反复实验对比, 选取出有代表性的 α , 实验结果如图 1~图 2 所示。

由于在每次实验中, 用户群组的选取为随机抽取, 因此在选取的有代表性的 2 组实验结果中, 衰减前的行为相似度会有一定程度的波动。由实验结果可以看出, 用户间的行为相似度随着用户数的增加并无统一的变化规律, 这是由于用户间的行为相似

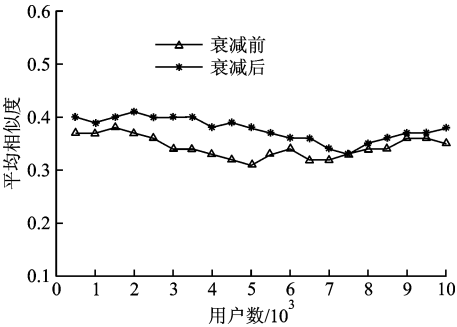


图 1 α 取值 0.12 时的实验结果

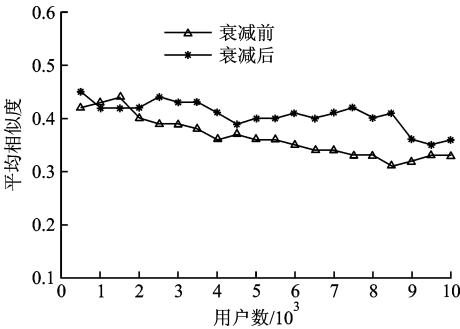


图 2 α 取值 0.71 时的实验结果

度只跟与其存在相关偏好行为的用户有关, 而与实验中选取的用户群组数无关。另外, 在不同数量的用户群组上, 衰减后的行为相似度值整体上大于原有值, 并且随着 α 的增加, 衰减后的行为相似度呈现出先增后减的趋势, 并在 $\alpha=0.71$ 时取得整体最优值, 说明在单位时间内偏好波动幅度是影响用户偏好衰减的主要因素。

实验 2 与现有方法的对比实验。

本文选取文献[10]中改进的协同过滤算法(OTCF)与文献[11]情景感知算法(GBCR)作为实验对比算法。OTCF 算法只采用评分数据, 所以在实验中只使用 BC-MN-Ratings 部分, 而 GBCR 算法与本文算法采用上下文信息与评分信息数据。

为了防止采用单一数据集导致的过拟合等问题, 将本文所用数据集划分成不同比例的测试数据集(数据提取方式为随机抽取), 并采用 ABO 算法随机隐藏数据。在这种情况下, 每次算法运行的数据集都不同, 实验结果会有一定程度的波动, 为提高实验结果客观性, 在每个数据节点运行 5 次, 将平均值作为每个节点的实验结果。经过反复实验对比, 在 BookCrossing-MN 数据集上, 选取 2 组有代表性的数据规模(用户数分别为 20 000、70 000), 并以 M 为度量标准, 对比实验结果如图 3~图 4 所示。在 Netflix 数据集上, 以 F_1 为度量标准, 选取同样比例的用户规模, 实验结果见表 1。

表 1 3 种方法的 F_1 对比结果

用户数	F_1		
	本文算法	GBCR	OTCF
10 000	0.18	0.12	0.94
20 000	0.21	0.15	0.96
70 000	0.22	0.14	0.89
100 000	0.19	0.15	0.87

由图 3、图 4 可以看出, 随着 $|R|$ 的增加, 3 种算

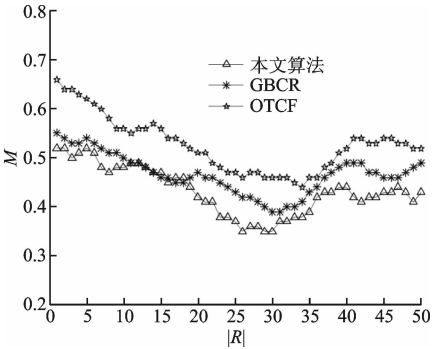


图 3 20 000 用户群组的推荐准确率对比的实验结果

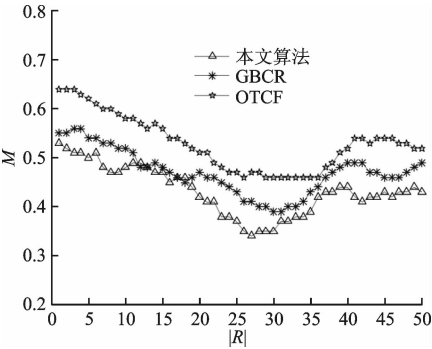


图 4 70 000 用户群组的推荐准确率对比的实验结果

法的准确率都有所提升,说明推荐列表长度越大,推荐结果越准确。随着 $|R|$ 的持续增加, M 逐渐升高,推荐准确率逐渐降低,原因在于推荐列表过长导致推荐算法采用了与待预测偏好用户相关度较低的相关数据,影响了推荐准确率。在 4 种数据规模上,与 OTCF 算法相比,本文算法与情景感知 GBCR 算法能够取得更好的推荐准确率,说明将上下文信息融入偏好获取流程,可以提高推荐算法的推荐质量。与 GBCR 和 OTCF 算法相比,本文算法准确率平均提升了 11.46%和 21.75%。由表 1 可以看出,采用 F_1 作为度量标准,本文算法能够取得更加准确的推荐结果,说明考虑用户偏好模型的多维特征并进行特征融合能够取得更好的推荐准确率,从而在准确率层面验证了本文算法的有效性。

多样性能够反映所获取的推荐结果中项目间的不相似程度,度量的是推荐算法能否挖掘用户潜在偏好,不同算法多样性对比的实验结果见表 2。

由表 2 可以看出,整体上 GBCR 算法的多样性比 OTCF 算法提高了 2.69%,说明考虑上下文信息仅能提升算法的准确率,对于多样性则无明显优势。通过对比还可以发现,与 GBCR 和 OTCF 算法相比,本文算法的多样性分别提升了 36.18%和 39.84%,即本文算法能够有效地挖掘用户的潜在偏好。本文

表 2 3 种算法的多样性对比结果

$ R $	D		
	本文算法	GBCR	OTCF
10	0.64	0.41	0.44
20	0.65	0.43	0.45
30	0.65	0.42	0.46
40	0.67	0.42	0.46
50	0.67	0.43	0.52

算法考虑了时间衰减度、属性偏好、行为可影响度等多维特征,能够建立更加全面的用户偏好模型,在一定程度上减弱了历史偏好的影响,提高了推荐列表的多样性。

本文对 3 种算法的时间复杂度进行对比,测试针对具体用户生成单次推荐列表所需的时间。随机选取 20 个用户,计算各推荐节点并取时间均值。实验的运行环境为 Intel(R)Core(TM)i5-4560 CPU,主频 3.3 GHz,内存 16 GB,实验结果如图 5 所示。

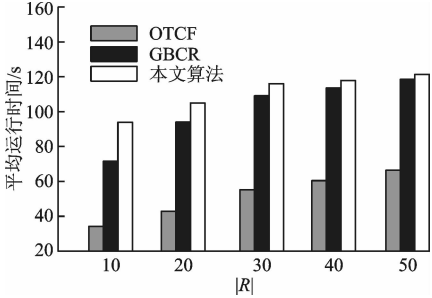


图 5 3 种算法运行时间的对比结果

由图 5 可以看出,OTCF 算法的时间复杂度最低,GBCR 算法的时间开销比 OTCF 算法增加约 111.49%,而本文算法比 GBCR 算法增加了约 9.46%。本文算法时间开销略高于 GBCR 算法,但增加幅度较少。因为本文算法和 GBCR 算法考虑了额外的上下文信息,所以会增加时间开销。本文算法的时间复杂度与现有的情景感知算法基本保持一致,说明了本文算法能够在可接受的时间开销内大幅提高推荐结果的准确率与多样性。

4 结 论

由于数据稀疏性等问题,现有的基于用户历史偏好记录的推荐方法所建立的偏好模型无法覆盖用户潜在的偏好特征,影响了推荐的准确率与多样性。本文将多视图学习理论引入偏好获取领域,提出了一种基于 Fisher 线性判别分析的多维特征融合情景感知推荐方法,提出了时间衰减度、属性偏好、偏

好可影响程度等多维特征的获取方法,并据此生成偏好样本空间,通过特征融合、投影变换,在最佳鉴别矢量空间通过 Fisher 判别准则融合用户的多维特征,并采用拉格朗日乘子法求解最优投影方向,并建立起多维特征优化的偏好获取模型。在大规模真实数据集上的对比实验结果表明,与现有方法相比,在时间开销提升约 9.46% 的基础上,推荐准确率平均提升 16.61%,多样性平均提高近 38.01%,验证了本文算法能够显著地提高推荐准确率与推荐质量,并且能够有效地挖掘用户的潜在偏好特征。

未来作者将会致力于解决多特征融合在推荐系统领域的应用中所存在的小样本及秩限制问题,以期进一步的提高推荐质量。

参考文献:

- [1] 王立才,孟祥武,张玉洁. 上下文感知推荐系统 [J]. 软件学报, 2012, 23(1): 1-20.
WANG Licai, MENG Xiangwu, ZHANG Yujie. Context-aware recommender systems [J]. Journal of Software, 2012, 23(1): 1-20.
- [2] 王智圣,李琪,汪静,等. 基于隐式用户反馈数据流的实时个性化推荐 [J]. 计算机学报, 2016, 39(1): 52-64.
WANG Zhisheng, LI Qi, WANG Jing, et al. Real-time personalized recommendation based on implicit user feedback data stream [J]. Chinese Journal of Computers, 2016, 39(1): 52-64.
- [3] 郑麟,朱福喜,姚杏. 基于属性提升与推荐采样的推荐评分预测 [J]. 计算机学报, 2016, 39(8): 1501-1514.
ZHENG Lin, ZHU Fuxi, YAO Xing. Recommendation rating prediction based on attribute boosting with partial sampling [J]. Chinese Journal of Computers, 2016, 39(8): 1501-1514.
- [4] 涂丹丹,舒承椿,余海燕. 基于联合概率矩阵分解的上下文广告推荐算法 [J]. 软件学报, 2013, 24(3): 454-464.
TU Dandan, SHU Chengchun, YU Haiyan. Using unified probabilistic matrix factorization for contextual advertisement recommendation [J]. Journal of Software, 2013, 24(3): 454-464.
- [5] 王玉祥,乔秀全,李晓峰,等. 上下文感知的移动社交网络服务选择机制研究 [J]. 计算机学报, 2010, 33(11): 2126-2135.
WANG Yuxiang, QIAO Xiuquan, LI Xiaofeng, et al. Research on context-awareness mobile SNS service selection mechanism [J]. Chinese Journal of Computers,

2010, 33(11): 2126-2135.

- [6] WANG Ming, KAWAMURA T, SEI Y, et al. Context-aware music recommendation with serendipity using semantic relations [M]. Berlin, Germany: Springer, 2014: 17-32.
- [7] AMATO F, MAZZO A, MOSCATO V, et al. Exploiting cloud technologies and context information for recommending touristic paths [M]. 3rd ed. Berlin, Germany: Springer, 2014: 281-287.
- [8] YAROWSKY D. Unsupervised word sense disambiguation rivaling supervised methods [C]// Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics. New York, USA: ACM, 1970: 189-196.
- [9] COVER T M. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition [J]. IEEE Transactions on Electronic Computers, 1965, 14(3): 326-334.
- [10] ZHANG Weinan, CHEN Tianqi, WANG Jun, et al. Optimizing top-*n* collaborative filtering via dynamic negative item sampling [C]// Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM, 2013: 785-788.
- [11] WU Hao, YUE Kun, LIU Xiaoxin, et al. Context-aware recommendation via graph-based contextual modeling and postfiltering [J]. International Journal of Distributed Sensor Networks, 2015, 2015: 613612.

[本刊相关文献链接]

- 赵长伟,彭勤科,张志勇. 混合因子矩阵分解推荐算法. 2016, 50(12): 87-91. [doi:10.7652/xjtuxb201612014]
- 郑修猛,陈福才,吴奇,等. 一种利用多群组智慧的协同推荐算法. 2016, 50(10): 118-124. [doi:10.7652/xjtuxb201610018]
- 孙忱,奚宏生,高荣. 邻域线性最小二乘拟合的推荐支持度模型. 2015, 49(6): 77-83. [doi:10.7652/xjtuxb201506013]
- 秦继伟,郑庆华. 面向推荐系统的音乐内涵空间建模研究. 2014, 48(4): 31-34. [doi:10.7652/xjtuxb201404006]
- 秦继伟,郑庆华,郑德立,田锋,等. 结合评分和信任的协同推荐算法. 2013, 47(4): 100-104. [doi:10.7652/xjtuxb201304017]
- 温彦,刘晨,韩燕波. 一种用户主导的跨组织数据按需集成方法. 2013, 47(2): 116-123. [doi:10.7652/xjtuxb201302020]
- 吴茜媛,张云强,郑庆华,等. e-Learning 用户心理体验量化评价研究. 2012, 46(12): 109-115. [doi:10.7652/xjtuxb201212109]

(编辑 武红江)