文章编号:1671 - 6833(2019)02 - 0048 - 07

融合社交信息的跨域时序兴趣预测方法

郝志峰, 申 策, 蔡瑞初, 温 雯

(广东工业大学 计算机学院,广东 广州 510006)

摘 要: 引入用户的社交网络信息是解决冷启动问题提升推荐效果的有效方法. 现有引入社交关系的静态方法忽略用户兴趣的变化导致预测结果滞后,针对这一问题提出一种跨域时序兴趣预测方法,引入用户社交行为信息. 首先提出跨域个性化排序模型,实现社交特征和购物特征的跨域融合. 其次按时间段划分用户历史行为,提出一种时序特征建模方法. 在真实数据集上的验证结果表明,所提出的方法能更有效地预测用户兴趣.

关键词: 兴趣预测; 跨域推荐; 社交信息; 时序行为; 排序学习

中图分类号: TP391 **文献标志码**: A **doi**:10. 13705/j. issn. 1671 - 6833. 2019. 02. 024

0 引言

个性化商品推荐系统通过分析用户购买信息来建模用户兴趣,并根据用户兴趣推荐相应商品解决信息过载问题. 虽然个性化商品推荐系统在一定程度上取得了较好的推荐效果,但是仍然面临着用户冷启动和预测结果滞后等问题.

随着社交网络的发展,用户在社交网站上留下了大量社交信息,为预测用户兴趣提供了额外的信息来源. 引入用户的社交信息,能够有效解决用户购物信息不足带来的预测精度低、冷启动等问题[1]. 依据社交关系理论[2-4]. Yang 等[5]提出了基于信任与被信任社交关系的推荐模型,Tang 等[6]从不同视角引入局部与全局社交信息并进行融合. Guo 等[7]将用户信任关系与评分信息一同看作隐式反馈信息提出了 TrustSVD 模型. 文献[8 - 9]通过分析社交邻近方法提出了低秩社交推荐模型,上述研究引入了的社交关系信息,一定程度上解决了用户冷启动问题.

笔者研究引入用户社交行为信息提升对用户兴趣的预测效果. 社交行为本身就包含着用户的兴趣,能够提供更准确的用户兴趣信息;并且社交行为与购物行为都具有时序性. 通过跨域的方式引入用户的社交行为信息存在两方面的挑战:①通过引入时序信息[10-11],能够使模型有效捕捉兴趣的动态变化,解决兴趣漂移问题[12]. 用户的社交、购物数据一般来自不同的网站,通常两个域中只有部分用户是重叠的,其余用户只有一个域的信息. 跨域引入用户的社交信息,需要以重叠用户作为信息传递的桥梁,而重叠用户数量不足必然会影响预测效果. ②两种行为统一时序建模困难. 用户的社交行为和购物行为都能反映用户当时的兴趣,在预测用户兴趣时可以互相作为补充. 但是,从时间上看两种行为的变化通常是不同步的,因此引入时序信息时需要尽量避免行为数据不同步对兴趣预测带来的干扰. 为此笔者提出一种跨域时序兴趣预测(cross-domain temporal interest prediction, CDTIP)方法,首先提出跨域兴趣预测模型,构建社交特征与购物特征间的跨域映射关系. 然后对用户两个域的时序行为进行特征建模,将用户行为数据按时间段分割,分别根据每个时间段的行为数据构造相应的社交特征和购物特征,根据这些特征训练模型预测用户兴趣. 实验结果表明,所提出的方法能够有效提升对用户兴趣的预测效果.

收稿日期:2018 - 08 - 17;**修订日期**:2018 - 10 - 11

基金项目:国家自然科学基金资助项目(61472089、61572143);NSFC - 广东联合基金(U1501254);广东省自然科学基金(2014A030306004、2014A030308008);广东省科技计划项目(2013B051000076、2015B010108006、2015B010131015);广东特支计划(2015TQ01X140);广州市珠江科技新星(201610010101)

作者简介:郝志峰(1968—),男,广东广州人,广东工业大学教授,博士,主要从事机器学习、人工智能方向的研究,E-mail: zfhao@gdut. edu. cn.

1　跨域时序兴趣预测方法

1.1　问题定义

在电商域中,所有的用户由集合 U_1 表示,所有的商品由集合 I 来表示. 商品推荐系统通过分析用户的购物历史记录,预测用户的兴趣. 购物记录由集合 $E = \{(u,i,t)\}$ 来表示,其中每个三元组 (u,i,t) 为用户的一条真实购买记录,其中 $u \in U_1, i \in I$ 分别表示用户和商品,$t \in T = \{t_1,\cdots,t_l\}$ 为对应的购物时间.

在社交域中,所有的用户由集合 U_2 表示,用户的社交信息主要包括用户个人信息、社交行为信息、社交关系信息等. 用户个人信息包括用户的性别、年龄等,可以通过特征向量表示这些信息,对应的特征向量集合为 $\boldsymbol{S}_{\mathrm{A}}$;笔者研究的用户社交行为信息主要为用户发送的文本信息,用户发送的所有文本由集合 $\boldsymbol{S}_{\mathrm{D}} = \{(text,t)\}$ 表示. 社交域和电商域的重叠用户由集合 $U = U_1 \cap U_2$ 表示,$t \in T$ 为文本发表时间.

根据上述符号定义,跨域兴趣预测问题可以形式化地描述为:通过电商域中的 $\{U_1,I,E\}$ 以及社交域中的 $\{U,\boldsymbol{S}_{\mathrm{A}},\boldsymbol{S}_{\mathrm{D}}\}$ 学习兴趣预测模型 f,对于任意用户 $u \in U_1 \cup U_2$ 模型预测用户的兴趣 $f_u: I \rightarrow \boldsymbol{y}_u$,其中 $\boldsymbol{y}_u$ 为 $|I|$ 维实值向量,每个维度的值 $\boldsymbol{y}_{ui}$ 表示用户对于对应商品 i 的兴趣值.

1.2　跨域个性化排序模型

笔者引入因子分解机(factorization machines, FM)模型,并将其扩展为跨域个性化排序模型来解决用户社交信息与购物信息的融合问题. 下面将简单讲解 FM 模型,随后介绍如何构建跨域个性化排序模型.

FM 模型由 Rendle 等[13]提出,是一个基于特征的预测模型. 给定 n 维特征向量 $\boldsymbol{x}$,其第 i 维的特征用 x_i 表示. 2 阶 FM 模型可以定义如下:

$$\hat{y}(\boldsymbol{x}) = w_0 + \sum_{i=1}^{n} w_i \boldsymbol{x}_i + \sum_{i=1}^{n}\sum_{j=i+1}^{n} \boldsymbol{x}_i \boldsymbol{x}_j \sum_{k=1}^{l} v_{ik} \cdot v_{jk}, \quad (1)$$

式中:$w \in \mathbf{R}^n, V \in \mathbf{R}^{n\times k}$ 为待学习参数. FM 模型能够有效的模拟向量 $\boldsymbol{x}$ 中不同特征之间的交互关系,通过改写为等价形式可以在线性时间复杂度内完成计算,等价公式为:

$$\hat{y}(\boldsymbol{x}) = w_0 + \sum_{i=1}^{n} w_i \boldsymbol{x}_i - \frac{1}{2}\sum_{k=1}^{l}\Big(\Big(\sum_{i=1}^{n} v_{ik}\boldsymbol{x}_i\Big)^2 - \sum_{i=1}^{n} v_{ik}^2 \boldsymbol{x}_i^2\Big). \quad (2)$$

用户的历史购物记录直接反映了用户对商品的兴趣,但也由于用户购买记录的隐式反馈[14]特性(用户未购买的商品夹杂着不感兴趣和感兴趣但还未浏览到的商品),因此直接将购买的商品标为正类 $y_{ui} = 1, (u,i) \in E$;没有购买的商品标为负类 $y_{ui} = 0, (u,i) \notin E$ 所训练的预测模型会存在过拟合问题. 笔者借鉴个性化排序技术的思想,不再直接预测用户对每个商品 $i \in I$ 的兴趣值 y_{ui};而是每次抽取一对商品 $a,b \in I$(其中 $y_{ua} \neq y_{ub}$),预测用户对于该对商品的相对偏好 $y_{uab} = y_{ua} - y_{ub}$,也即预测用户喜欢商品 a 胜过商品 b 的概率 $p_{uab} = \frac{1}{2}(y_{uab} + 1)$. 假设用户的特征向量为 $\boldsymbol{x}_u$,两个商品的特征向量分别为 $\boldsymbol{z}_a$、$\boldsymbol{z}_b$,则模型通过用户和商品的特征,预测用户对商品的相对兴趣,形式化描述如下:

$$\hat{y}_{uab}(\boldsymbol{x},\boldsymbol{z}) = \sum_{i=1}^{m} u_i \boldsymbol{x}_i + \sum_{i=1}^{n} v_i \boldsymbol{z}_i + \frac{1}{2}\sum_{k=1}^{l}\Big(\Big(\sum_{i=1}^{m} u_{ik}\boldsymbol{x}_i + \sum_{i=1}^{n} v_{ik}\boldsymbol{z}_i\Big)^2 - \sum_{i=1}^{m} u_{ik}^2 \boldsymbol{x}_i^2 - \sum_{i=1}^{n} v_{ik}^2 \boldsymbol{z}_i^2\Big), \quad (3)$$

式中:u_i、v_i、u_{ik}、v_{ik} 为模型参数;$\boldsymbol{x}$ 即为用户向量 $\boldsymbol{x}_u$;$\boldsymbol{z}_{ab} = \boldsymbol{z}_a - \boldsymbol{z}_b$,为两个不同商品对应特征向量的差向量. $\hat{y}_{uab}(\boldsymbol{x},\boldsymbol{z})$ 为模型预测的相对兴趣值,通过 sigmoid 回归函数可以将预测结果转化为用户喜欢商品 a 胜过商品 b 的概率表示形式,

$$\hat{p}_{uab} = \sigma(\hat{y}_{uab}) = \frac{1}{1 + \mathrm{e}^{-\hat{y}_{uab}}}. \quad (4)$$

通过交叉熵(cross-entropy, CE)损失函数计算损失 $-p_{uab}\log(\hat{p}_{uab}) - (1 - p_{uab})\log(1 - \hat{p}_{uab})$,并根据随机梯度下降法(stochastic gradient descent, SGD)更新模型参数,模型各参数对应的更新公式如下:

$$u_i \leftarrow u_i - \eta(\sigma(\hat{y}_{uab}) - p_{uab})\boldsymbol{x}_i, \quad (5)$$

$$v_i \leftarrow v_i - \eta(\sigma(\hat{y}_{uab}) - p_{uab})\boldsymbol{z}_i, \quad (6)$$

$$u_{ik} \leftarrow u_{ik} - \eta(\sigma(\hat{y}_{uab}) - p_{uab})\Big(x_i \sum_{j=1}^{m} u_{jk}\boldsymbol{x}_j - u_{ik}\boldsymbol{x}_i^2\Big), \quad (7)$$

$$v_{ik} \leftarrow v_{ik} - \eta(\sigma(\hat{y}_{uab}) - p_{uab})\Big(z_i \sum_{j=1}^{n} v_{jk}\boldsymbol{z}_j - v_{ik}\boldsymbol{z}_i^2\Big). \quad (8)$$

1.3　基于兴趣区间特征建模

用户的社交行为和购物行为都是时序行为,但不同行为间的时序关系复杂,导致同时建模两种时序行为并预测用户兴趣非常困难. 为此,笔者提出一种基于兴趣区间的行为数据划

分与时序特征建模方法.

1.3.1 兴趣区间定义

虽然用户的行为随着兴趣发生变化具有时序性,但是可以假设在一个时间段内用户的兴趣和相应行为是稳定的;虽然社交行为、购物行为以及用户兴趣三者的变化都不同步,但是可以假设在一个时间段内,用户的社交、购物行为所反应的兴趣与该段时间的实际兴趣是一致的. 基于上述两点定义了用户的兴趣区间.

定义 1 兴趣区间. 将用户的历史行为时间按照长度 τ 分割为首尾相接的时间段,假设每个时间段都满足:①时间段内用户兴趣基本不变;②不同的时间段之间用户的兴趣相互独立;③时间段内用户的社交、购物行为能且只能反映用户该段时间内的兴趣. 则称这些时间段 $\{h_1,\cdots,h_l\}$,$h \subseteq T$ 为用户的兴趣区间.

根据兴趣区间的定义,可以将用户的历史社交行为和购物行为分到不同的兴趣区间中,如图 1 所示. 传统的为每个用户构造特征 $\boldsymbol{x}_u$ 并预测用户兴趣 $\boldsymbol{y}_u$ 的方法,被转化为每个时间段都为用户构造相应特征 $\boldsymbol{x}_{uh}$ 并预测相应时间用户兴趣 $\boldsymbol{y}_{uh}$ 的方法. 因此传统方法可以认为是本方法的一个特例,当时间段的长度 τ 大于用户历史行为时间时,由于无法分割出兴趣区间,本方法便退化为传统方法.

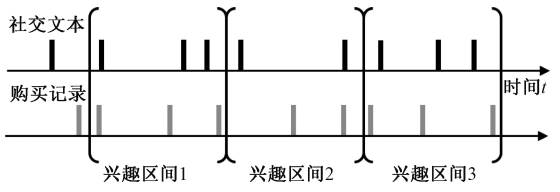

图 1 用户兴趣区间

Fig. 1 User's interest period

1.3.2 基于兴趣区间的特征建模

根据兴趣区间的定义,用户的社交行为数据和购物行为数据被分成了不同的子集合,根据新划分的数据,对用户进行特征建模.

用户特征向量 $\boldsymbol{x}_{uh}$ 构建. ①本研究筛选出了性别、地域(精确到省份)、等级、信用、注册日期、关注数、粉丝数等 7 个有代表性的个人信息,构造个人信息特征向量. 其中性别和地域等标签类型的特征采用了 one-hot 向量表示法,其余的数值类型的特征则使用 max-min 归一化后的值进行表示,最后所有的特征组合成用户的个人信息特征向量 $\boldsymbol{a}_u$. ② 用户的社交行为主要为发送的文本信息,将用户每个兴趣区间内发送的文本合并,然后通过 word2vec 工具中训练好的 Skip-gram 模型,将文档中的每个词转化为相应的向量形式,最后获得每个句子包含的所有词向量的平均向量;使用 TF-IDF 对句子的每个词进行词频统计,并通过 PCA 算法[15]对句子的稀疏向量表示进行降维,最后与句子的词向量拼接,作为文本特征向量 $\boldsymbol{d}_{uh}$. ③ 将用户 - 商品交互矩阵中用户的维度信息作为用户与商品的交互特征. 最终,用户特征向量 $\boldsymbol{x}_{uh}$ 如下所示,

$$\boldsymbol{x}_{uh} = (\underbrace{1,\cdots,0.2,\cdots,0.5}_{\text{个人信息特征}\boldsymbol{a}_u},\underbrace{-0.7,\cdots,-0.1,0.3}_{\text{社交行为特征}\boldsymbol{d}_{uh}},\underbrace{0,\cdots,0,1,0,\cdots,0}_{\text{用户对应维度为1}}).$$

商品特征向量 $\boldsymbol{z}_i$ 构建. ①为了方便用户查找和筛选,通常商品被分为不同的类别,如服装、食品等,使用 one-hot 向量 $\boldsymbol{a}_i$ 表示商品的类别特征;②将用户的购物记录按照购买时间先后排序,并构造用户购买商品的序列,将每个商品作为一个单词,然后使用 word2vec 工具中的 Skip-gram模型训练商品对应的特征向量 $\boldsymbol{d}_i$; ③将用户 - 商品交互矩阵中商品的维度信息作为商品与用户的交互特征. 最终,商品特征向量 $\boldsymbol{z}_i$ 如下所示,

$$\boldsymbol{z}_i = (\underbrace{0,\cdots,1,\cdots,0}_{\text{类别特征}\boldsymbol{a}_i},\underbrace{0.4,\cdots,-0.6,0.2}_{\text{商品特征}\boldsymbol{d}_i},\underbrace{0,\cdots,0,1,0,\cdots,0}_{\text{用户对应维度为1}}).$$

用户对商品的兴趣评分特征 $\boldsymbol{y}_{uh}$ 构建. 根据用户在每个兴趣区间上的购买记录集合 $E_h \subseteq E$,用户对商品的兴趣表示为;若 $(u,i,t) \in S_h$,则 $\boldsymbol{y}_{uh}$ 对应商品 i 维度的值 $y_{uhi}=1$;反之若 $(u,i,t) \notin S_h$ 则用户对商品不感兴趣 $y_{uhi}=0$. 用户兴趣特征向量 $\boldsymbol{y}_{uh}$ 如下所示,

$$\boldsymbol{y}_{uh} = (\underbrace{0,1,1,\cdots,0,1,1,\cdots,1,1,0}_{\text{购买过的商品值为1}}).$$

根据用户特征向量 $\boldsymbol{x}_{uh}$;商品特征向量 $\boldsymbol{z}_i$;用户兴趣评分向量 $\boldsymbol{y}_{uh}$,用户在兴趣区间的行为与相应兴趣的映射关系可以表示为:$f(\boldsymbol{x}_{uh},\boldsymbol{z}_i) \rightarrow y_{uhi}$.

1.4 模型训练算法

首先将用户历史行为时间分段构造兴趣区间,然后按照每个区间为用户构造特征向量以及相应的兴趣向量,在每个区间随机抽取用户兴趣商品和不感兴趣商品组成训练样本对,根据 1.2 节提出的跨域个性化排序模型,预测用户时序兴趣偏好,最后通过梯度下降法更新模型参数,具体如下.

算法 1　跨域时序兴趣预测算法

输入:电商平台用户集合 U_1;商品集合 I;历史时间 T;社交信息集合 S;历史购物集合 E;时间段长度 τ;算法迭代次数 γ;
输出:模型参数 $\theta = (\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{U}, \boldsymbol{V})$
算法步骤:
1:根据时间段长度 τ 划分历史时间,获得兴趣区间 $\{h_1, \cdots, h_n\} \leftarrow T$;
2:划分社交信息集合 $\{S_h\} \leftarrow S$,并为每个兴趣区间构造社交特征向量 $\boldsymbol{x}_{uh}$;
3:根据历史购物集合 E 构造商品特征向量 $\boldsymbol{z}_i$;
4:划分历史购物集合 $\{E_h\} \leftarrow E$,并为每个兴趣区间构造兴趣向量 $\boldsymbol{y}_{uh}$;
5:初始化参数 θ:$\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{U}, \boldsymbol{V} \leftarrow N(0, 0.1)$;
6:**for** $k \leftarrow 1$ to γ **do**
7:随机抽取一个用户 u,随机抽取一个兴趣区间 h;
8:随机抽取该区间内的一个样本对 (a, b),构造用户兴趣概率 p_{huab};
9:输入 $\boldsymbol{x}_{uh}, \boldsymbol{z} = \boldsymbol{z}_a - \boldsymbol{z}_b, p_{huab}$;
10:根据公式(5),(6)更新参数向量 $\boldsymbol{u}$ 和 $\boldsymbol{v}$;
11:根据公式(7),(8)更新参数矩阵 $\boldsymbol{U}$ 和 $\boldsymbol{V}$;
12:**end for**
13:**return** θ

模型训练完成后,对于任意用户 $u \in U_1 \cup U_2$,只需根据用户当前的社交信息或购物信息(重叠用户则是两种信息)构造相应的特征向量 $\boldsymbol{x}_{uh}$,即可预测用户当前的购买兴趣 $\boldsymbol{y}_{uh}$.

2　实验及分析

2.1　实验设置

通过与国内某电商平台的合作,我们获取了大约 1.3 万使用微博账号登录电商平台的用户数据,包括用户的微博 ID 和用户的购物数据. 其中购物数据包含了从 2014 年 1 月到 2017 年 7 月(共 42 个月)该批用户对 6 838 个不同品牌商品的共计 148 万条购物记录,其中每条购物记录包括用户 ID、品牌 ID 和订单的支付日期.

通过微博提供的 API 接口,我们获取了该批用户 2014 年 1 月到 2017 年 7 月的微博数据,包括用户的个人信息数据以及用户发送和转发的共计 213 万条微博. 用户的个人信息包含性别、所在省份、等级、信用、注册日期、关注数、粉丝数共 7 个部分. 两部分数据集的数据描述如表 1 所示,不同购买力的用户数量分布如图 2 所示,发送不同数量微博的用户数量分布如图 3 所示.

表 1　数据集描述

Tab. 1　Dataset description

数据来源	重叠用户	商品种类	数据	
			类型	数量/条
微博数据	13 165	—	微博文本	148 万
电商数据		6 838	购买记录	213 万

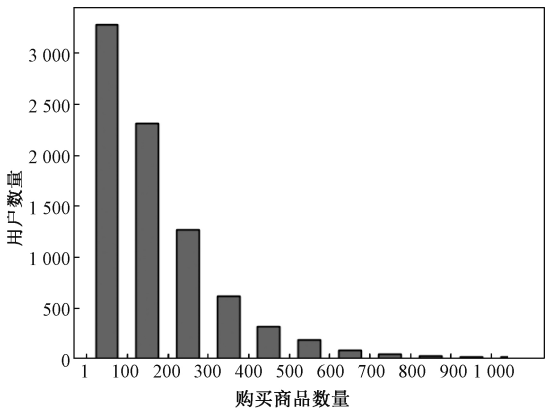

图 2　不同购买能力的用户分布

Fig. 2　Distribution of users with different purchases

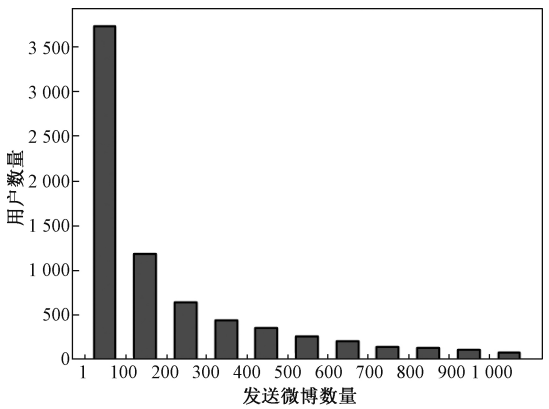

图 3　不同微博数量的用户分布

Fig. 3　Distribution of users with different blogs

通过将 2016 年 1 月设定为分界日期,获取到 2016 年初才开始出现购买记录的 327 位用户,将该部分用户作为冷启动用户,冷启动用户共有 19 433 条购物数据,该部分数据将不参与模型的训练,而是直接作为测试集用来测试模型对于冷启动用户的推荐效果.

根据购物记录条数是否超过 50 条对剩余用户进行筛选,并筛选出 1 423 位购物数量过少的用户作为长尾用户,长尾用户共有 38 612 条购物记录. 剩余的用户为普通用户,在模型训练过程

中,将分别从普通用户和长尾用户的购物记录中各随机抽取4/5混合后用于模型训练,剩余1/5的长尾用户和普通用户的数据将分别用来测试模型对于长尾用户和整体用户的推荐效果.最终,训练和测试数据的分布情况如表2所示.其中整体用户由普通用户、长尾用户和冷启动用户共同组成.

表2 用户数据描述

Tab.2 User data description

用户类别	用户数量	购物记录数量	发送微博数量	训练测试数据占比/%	
				用于训练	用于测试
冷启动用户	327	19 433	28 967	0	100
长尾用户	1423	38 612	55 369	80	20
普通用户	11 415	1 421 306	2 054 932	80	20

2.2 评价指标

平均绝对误差(mean absolute deviation, MAE)和均方根误差(root mean square error, RMSE)是评分预测任务中广泛使用的性能评价指标.N为样本数量、y_{ab}为实际评分值、$\hat{y}_{ab}$为模型预测评分值,则MAE的定义为:

$$MAE = \frac{1}{N}\sum_{a,b}|y_{ab}-\hat{y}_{ab}|. \tag{9}$$

$RMSE$的定义为:

$$RMSE = \sqrt{\frac{1}{N}\sum_{a,b}(y_{ab}-\hat{y}_{ab})^2}. \tag{10}$$

对于top-N推荐任务,准确率precision@N和召回率recall@N也是常用的性能评测指标.对用户u推荐的N个商品用集合$R(u)$表示,用户u真实购买过的商品用集合$T(u)$表示,则准确率的定义为:

$$\text{precision@ N} = \sum_{u}\frac{|R(u)\cap T(u)|}{|T(u)|}. \tag{11}$$

召回率的定义为:

$$\text{recall@ N} = \sum_{u}\frac{|R(u)\cap T(u)|}{|R(u)|}. \tag{12}$$

2.3 对比方法和参数设置

选取如下方法同笔者提出的CDTIP方法进行对比.

①SVD[16].同时考虑到用户特征、商品特征和二元交互特征3个方面的基于特征的矩阵分解模型.

②RFM[17].考虑到用户购物数据是隐反馈数据,因此使用了基于因子分解机改进的成对(pairwise)排序模型RFM模型.

③TrustSVD[7].对用户社交特征向量进行余弦相似度计算,相似度大于0.5的用户认为有社交关系.

④CDPR.笔者提出的跨域个性化排序模型,实验给出该模型与CDTIP的对比,用于验证引入兴趣区间这一时序信息的效果.

对于训练数据集,将每一条真实购物记录作为一个正例样本,并为每个正例随机抽取一个该用户未购买过的商品作为负例,使正负例的比例为1∶1组成训练样本对;对于测试数据集,将每一条购物记录作为一个正例,并为每个正例随机抽取50个负例,使正负例的比例为1∶50组成测试数据.

对于所有模型,统一采用梯度下降法进行优化,迭代次数统一设置为500次,即每条训练样本都会参与500次训练,模型的学习率为0.01,正则项设置为0.004,分解因子的数量统一设置为32.社交特征向量都使用了用户个人信息和用户的微博文本,文本统一使用Skip-gram模型转化为词向量,向量的维度统一为50.

2.4 实验效果对比

按照之前所述选取数据进行训练,将剩余未参与训练的数据分为整体用户、冷启动用户和长尾用户3个测试集,分别测试模型对于不同类型用户的推荐效果,各方法在Pre@N作为评价指标上的实验效果如图4所示;各方法在Rec@N作为评价指标上的实验效果如图5所示;各模型在不同类型用户上的MAE、$RMSE$等指标的测试效果对比如表3、4、5所示.

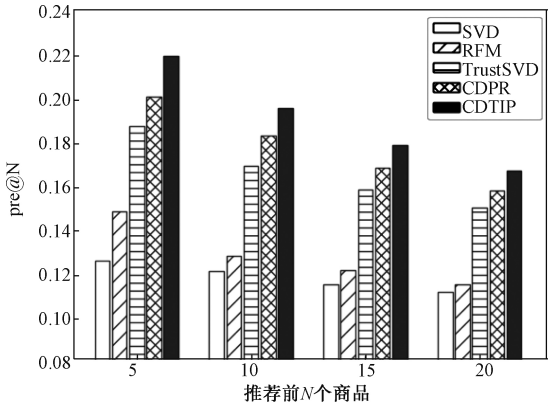

图4 精度指标对比

Fig.4 Precision@N

在参与对比的方法中,笔者提出的跨域时序兴趣预测CDTIP方法,无论是在精度指标上还是召回率指标上都比加入社交信息的TrustSVD方

法

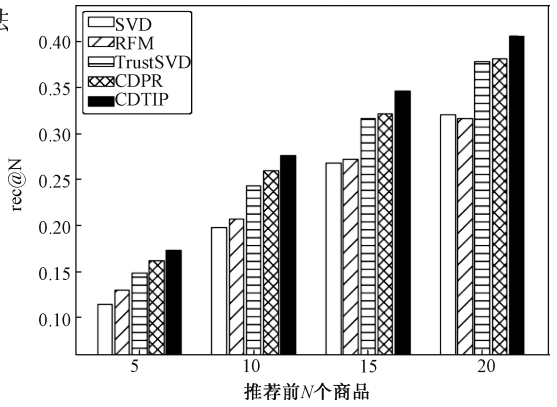

图 5　召回率指标对比

Fig. 5　Recall@N

表 3　整体用户测试结果

Tab. 3　Performance for all users

方法	p@5	p@10	r@5	r@10	*MAE*	*RMSE*
SVD	0.125	0.121	0.114	0.197	0.643	0.712
RFM	0.158	0.147	0.130	0.217	0.597	0.669
TrustSVD	0.186	0.168	0.148	0.242	0.545	0.610
CDPR	0.191	0.172	0.151	0.250	0.505	0.587
CDTIP	**0.209**	**0.185**	**0.163**	**0.266**	**0.503**	**0.582**

表 4　冷启动用户测试结果

Tab. 4　Performance for cold start users

方法	p@5	p@10	r@5	r@10	*MAE*	*RMSE*
SVD	0.089	0.082	0.067	0.157	0.729	0.783
RFM	0.116	0.107	0.083	0.174	0.668	0.718
TrustSVD	0.129	**0.126**	0.102	0.185	0.641	0.698
CDPR	0.134	0.119	0.101	0.193	0.635	0.689
CDTIP	**0.141**	0.123	**0.115**	**0.201**	**0.624**	**0.682**

表 5　长尾用户测试结果

Tab. 5　Performance for long tail users

方法	p@5	p@10	r@5	r@10	*MAE*	*RMSE*
SVD	0.115	0.109	0.091	0.169	0.675	0.740
RFM	0.133	0.115	0.128	0.183	0.639	0.703
TrustSVD	0.161	0.147	0.132	0.209	0.598	0.647
CDPR	0.170	0.165	0.148	0.238	0.570	0.616
CDTIP	**0.184**	**0.172**	**0.154**	**0.240**	**0.563**	**0.599**

效果有所提升,本方法在精度上平均提升 11%,在召回率上平均提升 9%. 与不按照兴趣区间划分数据的 CDPR 方法进行对比,在精度上平均提升 7%,在召回率上平均提升 6%,*MAE* 和 *RMSE* 指标也有所降低.

与其他用户部分数据参与模型训练部分数据用于测试不同,冷启动用户的所有数据都不参与模型训练,因此对模型提出了更大的挑战. 得益于使用 word2vec 训练得到的商品向量能够包含一定的用户购物相似性信息,以及用户社交网络信息对用户特征的补充,笔者提出的方法对冷启动用户的推荐效果优于其他对比方法.

对长尾用户的推荐效果低于对整体用户的推荐效果,一个可能的原因是长尾用户购买的商品量过少,对模型预测产生了影响,但是从实验结果可以看出,笔者提出的方法对长尾用户的推荐效果优于其他对比方法.

2.5　兴趣区间范围大小的影响

根据 1.3 节中所述,时间段长度 τ 规定了兴趣区间的时间范围,本研究分别选取了周(7 天)、月(30 天)、季度(91 天)、半年(182 天)和一年(365 天)作为 τ 的待选值,通过实验效果对比,确定 τ 的合理取值. 采用了 pre@5 和 pre@10 作为评价指标,结果如图 6 所示.

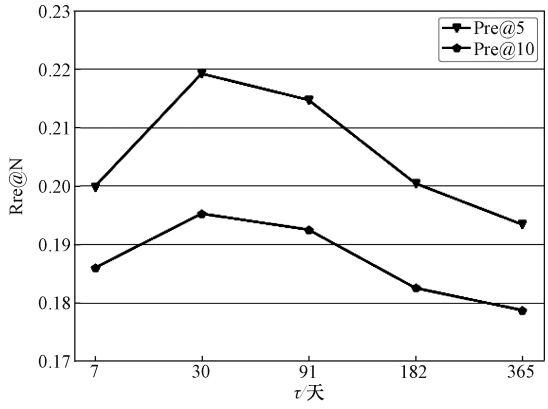

图 6　时间段长度的影响

Fig. 6　Effects of period length

由图 6 可知,兴趣区间范围越小越能得到精确的用户社交与购物行为之间的关系,但是兴趣区间范围过小,会因为无法获取足够的信息,而导致效果更差. 当时间段长度 τ 的取值为一个月(30 天)时效果优于其余选择.

3　结论

提出了一种跨域时序兴趣预测方法,实验结果验证了本方法同时引入社交信息和时序信息的有效性. 虽然取得了一定的成功,但是本方法只是假设用户的每个兴趣区间内的兴趣是相互独立的,而事实上区间内的兴趣之间存在一定的相关性. 我们将在接下来的工作中研究如何利用神经网络[18]等模型建模这种相关特性.

参考文献:

[1]　刘华锋, 景丽萍, 于剑. 融合社交信息的矩阵分解推荐方法研究综述[J]. 软件学报, 2018(2):340 - 362.

[2] MARSDEN P V, FRIEDKIN N E. Network studies of social influence. [J]. Sociological Methods & Research, 1993, 22(1):127 – 151.

[3] WASSERMAN S, FAUST K. Social Network Analysis [J]. Encyclopedia of Social Network Analysis & Mining, 1994, 22(Suppl 1):109 – 127.

[4] MA H, ZHOU D, LIU C, et al. Recommender systems with social regularization[C]// Forth International Conference on Web Search & Web Data Mining. DBLP, 2011:287 – 296.

[5] YANG B, LEI Y, LIU J, et al. Social Collaborative Filtering by Trust. [C]// International Joint Conference on Artificial Intelligence. AAAI Press, 2013: 2747 – 2753.

[6] TANG J, HU X, GAO H, et al. Exploiting local and global social context for recommendation[C]// International Joint Conference on Artificial Intelligence. AAAI Press, 2013:2712 – 2718.

[7] GUO G, ZHANG J, YORKE-SMITH N. TrustSVD: collaborative filtering with both the explicit and implicit influence of user trust and of item ratings[C]// Twenty-Ninth AAAI Conference on Artificial Intelligence. AAAI Press, 2015:123 – 129.

[8] SEDHAIN S, SANNER S, BRAZIUNAS D, et al. Social collaborative filtering for cold-start recommendations[J]. 2014:345 – 348.

[9] SEDHAIN S, MENON A K, SANNER S, et al. Low-Rank linear cold-start recommendation from social data. In: Proc. of the 31th AAAI Conf. on Artificial Intelligence. AAAI Press, 2017:1502 – 1508.

[10] REN Y, ZHU T, LI G, et al. Top-N Recommendations by Learning User Preference Dynamics[C]// Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, 2013:390 – 401.

[11] WANG X, ZHU J, ZHENG Z, et al. A Spatial-Temporal QoS Prediction Approach for Time-aware Web Service Recommendation[J]. Acm Transactions on the Web, 2016, 10(1):7.

[12] 孙光福, 吴乐, 刘淇,等. 基于时序行为的协同过滤推荐算法[J]. 软件学报, 2013(11): 2721 – 2733.

[13] RENDLE S. Factorization Machines[C]// IEEE, International Conference on Data Mining. IEEE, 2011: 995 – 1000.

[14] RENDLE S, FREUDENTHALER C, GANTNER Z, et al. BPR: Bayesian personalized ranking from implicit feedback[J]. 2012:452 – 461.

[15] 董建华, 王国胤, 雍熙,等. 基于 Spark 的标准化 PCA 算法[J]. 郑州大学学报(工学版), 2017, 38(5):7 – 12.

[16] CHEN T, ZHANG W, LU Q, et al. SVDFeature: a toolkit for feature-based collaborative filtering [J]. Journal of machine learning research, 2012, 13(1): 3619 – 3622.

[17] YUAN F, GUO G, JOSE J M, et al. LambdaFM: Learning Optimal Ranking with Factorization Machines Using Lambda Surrogates[J]. 2016:227 – 236.

[18] 孙峰, 龚晓玲, 张炳杰,等. 一种基于共轭梯度法的广义单隐层神经网络[J]. 郑州大学学报(工学版), 2018,39(2):28 – 32.

A Cross-domain Temporal Interest Prediction Method by Integrating Social Information

HAO Zhifeng, SHEN Ce, CAI Ruichu, WEN Wen

(School of Computer Science, Guangdong University of Technology, Guangzhou 510006, China)

Abstract: Integrating user's social information was an appropriate way to solve the user-cold start problem. Among various prediction models focusing on integrating social relation information, few noticed the dynamic change of the user's interest. Thus, in this paper, we propose a cross-domain temporal interest prediction approach was proposed by integrating social activity information. Firsty a cross-domain personized ranking model was constructed which can map the feature from social space into the purchase space. Further, we propose a feature modeling method based on data grouped by time period was proposed. Experiments on the dataset verified that the proposed method could predict user's interest more effectively.

Key words: interest prediction; cross-domain recommendation; social information; sequential behavior; learning to rank