

融合社交因素和评论文本卷积网络模型的汽车推荐研究

冯永^{1,2} 陈以刚^{1,2} 强保华³

摘要 汽车作为较高价值和个性化的消费品,使得用户购车决策过程较一般商品更为复杂. 本文主要研究社交环境和评论文本两方面对用户购车决策过程的影响,提出了融合社交因素和评论文本卷积网络模型(Social and comment text CNN model based automobile recommendation, SCTCMAR). SCTCMAR 首先定义了基于购买用途需求的社交圈,在此基础上提出了个人偏好计算方法,并引入了偏好相似度;其次,设计了卷积神经网络学习汽车评论文本的隐特征;然后将社交影响量化因素和评论文本特征有机融合注入推荐模型,并采用低阶矩阵分解技术进行模型计算. 另外,本文使用 GloVe 预训练词嵌入模型,产生了 SCTCMAR 的另一个版本 SCTCMAR+. 最后,将 SCTCMAR、SCTCMAR、FMM (Flexible mixture model)、TR (Trust rank)、Random sampling 在课题组爬取后经清理、去重和整合的 266 995 个用户、702 辆汽车信息的真实数据集上进行精确率、召回率和平均倒序排名三个指标的多粒度实验比较,结果表明本文提出的 SCTCMAR+ 和 SCTCMAR 具有良好的推荐性能.

关键词 汽车推荐, 卷积神经网络, 社交圈, 矩阵分解

引用格式 冯永, 陈以刚, 强保华. 融合社交因素和评论文本卷积网络模型的汽车推荐研究. 自动化学报, 2019, 45(3): 518–529

DOI 10.16383/j.aas.2018.c170245

Social and Comment Text CNN Model Based Automobile Recommendation

FENG Yong^{1,2} CHEN Yi-Gang^{1,2} QIANG Bao-Hua³

Abstract This paper mainly studies the influence of social environment and comment text on the user decision making process in purchasing an automobile, and proposes an automobile recommendation model named SCTCMAR (social and comment text CNN model based automobile recommendation). First, SCTCMAR defines a social circle of users in terms of user's purpose in choosing an automobile; on this basis, a method to calculate personal preference is put forward. and preference similarity is put forward. Second, a convolution neural network model is designed to learn the hidden features from the automobile comment texts. Afterwards, both the social circle and comment text features are integrated into the recommendation model, and low-rank matrix decomposition technology is used to solve the problem. In addition, by applying the GloVe pre-training word embedded model to SCTCMAR, an improved version, SCTCMAR+ is further proposed. Finally, a performance of comparison SCTCMAR+, SCTCMAR with FMM (flexible mixture model), TR (trust rank), and Random sampling is performed on a real dataset with 266 995 users and 702 automobiles. Experimental results show that SCTCMAR+ and SCTCMAR models outperform other counterparts in precision, recall and ARHR (average reciprocal hit rank).

Key words Automobile recommendation, convolution neural network (CNN), social circle, matrix factorization

Citation Feng Yong, Chen Yi-Gang, Qiang Bao-Hua. Social and comment text CNN model based automobile recommendation. *Acta Automatica Sinica*, 2019, 45(3): 518–529

收稿日期 2017-05-10 录用日期 2017-11-06
Manuscript received May 10, 2017; accepted November 6, 2017
国家自然科学基金(61762025), 国家重点研究发展计划(2017YFB1402400), 重庆市基础与前沿研究计划(cstc2017jcyjAX0340), 广西可信软件重点实验室开放课题(kx201701), 广西云计算与大数据协同创新中心开放课题(YD16E01), 重庆市重点产业共性关键技术创新专项(cstc2017zdcy-zdyxx0047), 重庆市社会事业与民生保障科技创新专项(cstc2017shmsA20013) 资助
Supported by National Natural Science Foundation of China (61762025), National Key Research and development Program of China (2017YFB1402400), Frontier and Application Foundation Research Program of Chongqing CSTC (cstc2017jcyjAX0340), Guangxi Key Laboratory of Trusted Software (kx201701), Guangxi Cooperative Innovation Center of Cloud Computing and Big Data (YD16E01), Key Industries Common Key Technologies Innovation Projects of Chongqing CSTC (cstc2017zdcy-zdyxx0047), and Social Undertakings and Livelihood Security

汽车相对于普通生活用品,属于特殊商品,一是价格高,在购买时用户相对更谨慎;二是用户购买需求复杂,决策过程的总体时间相对较长. 由于汽车自

Science and Technology Innovation Funds of Chongqing CSTC (cstc2017shmsA20013)

本文责任编辑 周涛

Recommended by Associate Editor ZHOU Tao

1. 重庆大学计算机学院 重庆 400030 2. 重庆大学信息物理社会可信服务计算教育部重点实验室 重庆 400030 3. 桂林电子科技大学广西云计算与大数据协同创新中心 桂林 541004

1. College of Computer Science, Chongqing University, Chongqing 400030 2. Key Laboratory of Dependable Service Computing in Cyber Physical Society, Ministry of Education, Chongqing University, Chongqing 400030 3. Guangxi Cooperative Innovation Center of Cloud Computing and Big Data, Guilin University of Electronic Technology, Guilin 541004

身具有复杂度高、交易和消费频次低等特点,所以在推荐任务执行过程中,会遇到用户相关领域知识匮乏、用户历史交易数据少、商品属性复杂等困难。

针对相关领域知识匮乏和商品属性复杂,研究者一般通过构建相关知识领域专家系统应对。这些专家系统通过与用户进行信息交互来准确获得用户的信息及需求,在此基础上结合已有的领域专业知识体系协同计算分析,从而对用户提供更高效的推荐^[1]。然而,专家系统的构建过程相对繁琐,在系统构建初期,需要将领域专家提供的信息和知识体系转化成系统能理解的数据模型,这些数据模型主要包括专业知识库、规则集合等,这个过程需要花费大量的时间和人力。同时,在整理专业知识和规则时,由于人为将需求和属性特征关联,导致专家系统的知识库和规则库产生一定的片面性和主观性,最终对推荐结果形成干扰。针对用户历史交易数据少,一些研究者利用社交网络来获取用户更多的相关信息来弥补信息缺失,而另外一些研究者则使用项目信息(例如项目描述、简介、评论文本信息)进行处理。虽然在社交网络和评论文本这两个领域的研究已分别有了瞩目的成果,但是将社交网络和评论文本相互结合的研究却相对较少。社交网络为用户提供了丰富的社交信息,可缓解用户历史交易记录少的困难,而评论文本分析则可缓解用户相关领域知识匮乏和商品属性复杂带来的负面影响。

本文首次将评论文本与社交网络信息结合应用于汽车商品推荐领域,一方面构建购买用途需求社交圈,在此社交圈的基础上分析个人偏好和偏好相似度两种社交因素,并给出合理的量化计算方法。另一方面,以评论文本为汽车的附属信息,利用卷积神经网络提取文本隐特征。在设计汽车推荐模型时,融合社交环境因素和评论文本卷积网络获得的附加属性信息,从而提升汽车推荐系统的效果。

综上,在对社交网络和评论行为两方面的数据充分挖掘的基础上,设计汽车推荐模型这一研究在技术理论和现实应用层面都具有重要的意义和价值。

1 相关工作

近年来,研究人员提出了几种基于社会信任的推荐系统来提高推荐准确性。事实上,用户可以信任不同类别中的不同朋友。基于这一概念,最近有研究引入了“推断的信任圈”概念^[2],用于社交网络中的推荐,考虑人际信任的影响。他们专注于推断类别特定的社交信任圈。然而,一些社交网络用户喜欢选择与他们的个人偏好密切相关的产品,很少考虑人际影响。研究者探索使用矩阵分解工具设计社交推荐从而解决“冷启动”问题。Jamali等^[3]在社交网络中探索了这类方法,将信任传播机制加入到推荐模

型中。潘涛涛等^[4]利用置信系数区分评分值之间的可信度。信任传播已被证明是社会网络分析和信任推荐中的关键因素。Yang等^[5]提出社交网络中“信任圈”的概念。他们的模型优于BasicMF^[6]和SocialMF^[3]。用户之间的信任值由矩阵 S 表示,用户 u 与用户 v 的有向加权社会关系由正值 $S_{u,v}^{c*} \in [0, 1]$ 表示。其基本思想是用户特征应该类似于他/她的好友的特征的平均值,其种类为 c 中的权重 $S_{u,v}^{c*}$ 。除了文献[5]中的人际影响因素外,Jiang等^[7]提出另一个重要因素:个人偏好。他们对中国的人人网数据集和腾讯微博数据进行了实验,结果表明社会语境因素(个人偏好和人际影响)在模型中有重要的意义。例如张燕平等^[8]利用历史记录获取用户声誉来建立声誉推荐系统。Qian等^[9]提出了一种结合用户人际兴趣相似性、人际影响力和个人兴趣因素的个性化推荐模型(Personalized recommender model, PRM)。Wang等^[10]提出使用社会传播模拟和内容相似性分析来更新用户内容矩阵。他们还构建了一个联合的社会内容空间来衡量用户和视频之间的相关性,为视频导入和重新分享推荐提供了高度的准确性。Feng和Qian^[2]提出了一个推荐模型,以满足用户(特别是对有经验的用户)的个性。此外,他们的方法不仅考虑个人偏好,而且结合了人际信任影响和人际兴趣相似性。

隐含狄利克雷分布(Latent Dirichlet allocation, LDA)和栈式降噪自编码器(Stacked denoising auto-encoder, SDAE)等基于文档建模的方法主要是利用了评论、摘要或概要的项目描述文档^[11-14]。具体来说,Wang等^[11]提出了协同主题回归(Collaborative topic regression, CTR),在概率方法中结合了LDA和协同过滤。而CTR的变体是将LDA集成到协同过滤中通过使用不同的集成方法来分析项目描述文档^[12-13]。Wang等^[6]提出了协同深度学习(Collaborative deep learning, CDL),将SDAE整合到概率矩阵分解(Probabilistic matrix factorization, PMF)中,从而在评分预测精度方面产生更准确的隐义模型^[14]。

但是,现有的集成模型不能完全捕获文档信息,因为它们采用词袋模型,该模型假定忽略如周围单词和单词顺序文档的上下文信息。但文档内的这种细微差异也是成为更深入理解文档的重要因素,并且进一步理解有助于改进评分预测准确度。为了解决上述问题,部分研究者尝试使用卷积神经网络(Convolutional neural network, CNN)。CNN通过本地接收场、共享权重和亚采样的建模组件有效地捕获图像或文档的局部特征^[15-16]。因此,CNN的使用有助于更深入地理解文档,并且产生比LDA和SDAE更好的隐义模型,特别是缺乏评分的描述

文档. 此外, CNN 能够利用预训练的词嵌入模型, 例如 Glove^[17], 以便更好地理解项目描述文档, 而 LDA 和 SDAE 不能利用预训练的词嵌入模型.

然而, 现有的 CNN 不适合推荐任务. 具体来说, 常规 CNN 主要解决预测词、短语或文档的标签分类任务. 相反, 推荐的目的通常是回归任务, 旨在准确地逼近用户对项目的评分. 因此, 现有的 CNN 不能直接应用于本文的推荐任务. 为了解决该技术问题, Kim 等^[18] 提出了一种文档上下文感知推荐模型—卷积矩阵分解 (Convolutional matrix factorization, ConvMF), 利用 CNN 捕获项目描述文档的上下文信息, 进一步增强评分预测精度.

2 SCTCMAR 模型

本文对汽车推荐系统进行研究, 为了更好地理解用户的评分行为, 需要挖掘用户的个性化特征. 本文从两方面入手: 1) 社交环境对用户的影响, 在特定的车型下构建购买用途需求的社交圈, 在此社交圈的基础上, 分析用户的个人偏好和偏好相似度两个社交因素, 并计算量化其社交关系; 2) 分析评论文本对项目的影响, 通过深度学习技术构建了四层卷积神经网络来学习文本特征. 最终, 本文提出了新颖的 SCTCMAR 汽车推荐模型, 该模型融合了社交影响和评论文本特征, 即用户的个人偏好、偏好相似度以及项目的评论文本特征, 最终将这三个因素注入到推荐模型中体现社交因素和评论文本特征因素对用户决策过程的影响.

2.1 汽车推荐框架

本文提出的模型如图 1 所示, 中间的点线框为 PMF, 左边线框是社交因素, 右边线框是 CNN. 该模型在概率矩阵分解 (PMF) 的基础上, 融合社交因素对用户的影响和评论文本特征对项目的影响, 提出了 SCTCMAR 模型. 推荐模型中的符号定义见表 1.

假设有 N 个用户和 M 个项目以及真实评分矩阵 $R = [R_{u,i}]^{N \times M}$, 目标是找到用户和项目的隐义模型 ($U^T \in \mathbf{R}^{n \times k}$ 和 $V \in \mathbf{R}^{m \times k}$), 求其乘积 ($U^T V$) 来重构评分矩阵 R . 通过概率矩阵分解模型使用 W , F 和 Q 不断地影响用户和项目的隐义模型, 从而求其乘积得到预测 u 到 i 的评分值, 不断逼近预测值与实际值之间的差距, 从而提高推荐精准度. 在概率上真实评分的条件概率分布如下:

$$p(R|U, V, \sigma_R^2) = \prod_u \prod_i \prod_{i'} N(r_{ui} | u_u^T v_i, \sigma_R^2)^{I_{ui}} \quad (1)$$

其中, $N(x|u, \sigma^2)$ 表示均值为 u , 方差为 σ^2 的高斯分布概率密度函数, I_{ui} 是指示函数, 如果用户 u 对

项 i 评价过, 则 $I_{ui} = 1$, 反之为 0.

表 1 推荐框架的符号定义

Table 1 Notations in recommender framework

表示符号	概念描述
$U = \{u_1, \dots, u_N\}$	用户集合
$F = [F_{u,v}]^{M \times M}$	偏好相似度矩阵
$V = \{i_1, \dots, i_M\}$	汽车集合
$F_{u,v} \in [0, 1]$	用户 u 和用户 v 之间的偏好相似度
$R = [R_{u,i}]^{N \times M}$	评分矩阵
$Q = [Q_{u,i}]^{N \times M}$	用户个人偏好
$R_{u,i}$	用户 u 对汽车 i 的评分
$Q_{u,i} \in [0, 1]$	用户 u 的偏好和汽车 i 的主题之间的相关性
W	卷积神经网络 (CNN) 中的内部权重

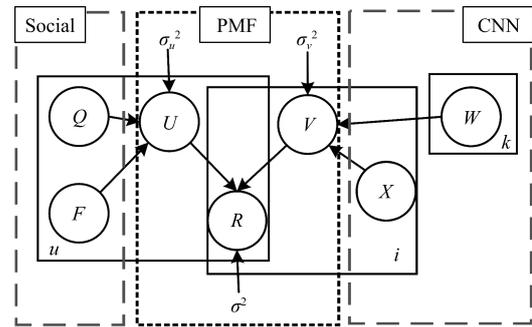


图 1 SCTCMAR 的图模型

Fig. 1 Graphical model of SCTCMAR

对于用户隐义模型有三个因素: 1) 为了防止过拟合的 0 均值高斯先验概率; 2) 在已知该用户社交圈中朋友的偏好隐特征时, 用户隐特征的条件概率; 3) 在已知该用户社交圈中个人偏好隐特征时, 用户隐特征的条件概率. 具体定义为

$$p(U|F, Q, \sigma_U^2, \sigma_F^2, \sigma_Q^2) \propto p(U|\sigma_U^2) \times p(U|F, \sigma_F^2) \times p(U|Q, \sigma_Q^2) = \prod_u^N N(u_u | 0, \sigma_U^2 I) \times \prod_u^N N(u_u | \sum_{v \in C_u^c} F_{u,v}^{c*} U_v^c, \sigma_F^2 I) \times \prod_u^N N(u_u | Q_{u,i}^{c*} U_i^c, \sigma_Q^2 I) \quad (2)$$

与常规 PMF 中的项隐义模型不同, 我们假设从三个变量生成项目隐义模型. 1) CNN 中的内部权重 W ; 2) 表示项 i 的文档 X_i ; 3) ε_i 变量作为高

斯噪声. 这使我们能够进一步优化项隐义模型的评分. 因此, 最终项隐义模型为

$$v_i = cnn(W, X_i) + \varepsilon_i \quad (3)$$

$$\varepsilon_i \sim N(0, \sigma_V^2 I) \quad (4)$$

其中, cnn 表示卷积运算.

对于 W 中的每个权重 w_k , 设置最常用 0 均值高斯先验概率. 具体定义如下:

$$p(W|\sigma_W^2) = \prod_k N(W_k|0, \sigma_W^2) \quad (5)$$

于是, 项隐义模型的条件概率如下:

$$p(V|W, X, \sigma_V^2) = \prod_i^M N(v_i|cnn(W, X_i), \sigma_V^2 I) \quad (6)$$

其中, X 是项目的描述文档集合. 为了建立 CNN 和 PMF 之间的桥梁, 使用从 CNN 模型获得的评论隐向量作为高斯分布均值, 并且项的高斯噪声作为高斯分布的方差, 这有助于完全分析描述文档和评分.

接下来将按以下过程详细介绍:

- 1) 为了确定社交环境的影响范围, 需要构建基于购买用途需求的社交圈;
- 2) 在该社交圈的基础上分别分析用户的个人偏好和偏好相似度;
- 3) 使用卷积神经网络提取评论隐特征.

2.1.1 社交圈

给定用户 u , 类别 c 和用途 t , 本文使用 Yang 等^[2] 提出的推理规则为当前类别评分用户建立社交圈.

定义 1. 基于购买用途需求的社交圈 (C_u^t). 在特定类别 c 下, 对于用途 t , 原始社交网络 SN (Social network) 满足以下条件的所有用户 v 构成 u 的社交圈 C_u^t , 即 $v \in C_u^t$.

1) 在 SN 中用户 u 和用户 v 有直接社交关系, 即 $SN_{u,v} = 1$;

2) $t \in u^t \ \&\& \ t \in v^t$. 其中, u^t 代表用户 u 的购买用途集合, $t \in u^t$ 表示用户 u 的购买用途需求集合中含有用途 t . 如果 v 符合上述条件, 则记为 $v \in C_u^t$, 否则记为 $v \notin C_u^t$.

按照给出的说明, 图 2 展示了购买用途需求的社交圈的构造过程. 每个用户购买用途需求使用括号标注在用户旁边, 例如用户 u 的购买用途需求为 t_1, t_2, t_3 . 用户 u 为了构造特定购买用途需求 t_1 下的社交圈 $C_u^{t_1}$, 首先需要在 SN 中找出与 u 有直接社交关系的用户集合 $\{u_1, u_2, u_3, u_4, u_5\}$, 再按照条件 $t \in u^t \ \&\& \ t \in v^t$ 进行筛选, 最后找到用户集合 $\{u_1, u_2, u_3\}$ 是用户 u 在购买用途需求 t_1 下的社交圈 $C_u^{t_1}$. 同理, 可推出 u 在 t_2 和 t_3 下的社交圈 $C_u^{t_2}$ 和 $C_u^{t_3}$. 下文对社交影响中的社交因素的有关研究全部限制在购买用途需求的社交圈范畴内.

2.1.2 个人偏好

Feng 和 Qian^[2] 提出了一种基于项类别分布向量来衡量用户个人偏好的方法, 但忽略了历史用户评分值. 然而, u 对项 i 的评分值反映了 u 喜欢 i 的程度. 因此, 将项类别分布与用户评分数据结合, 可以更精确地学习用户的个人偏好. 关于类别 c , 可以从评分矩阵 R^c 获得用户评分数据. 矩阵 R^c 的每一行归一化如下:

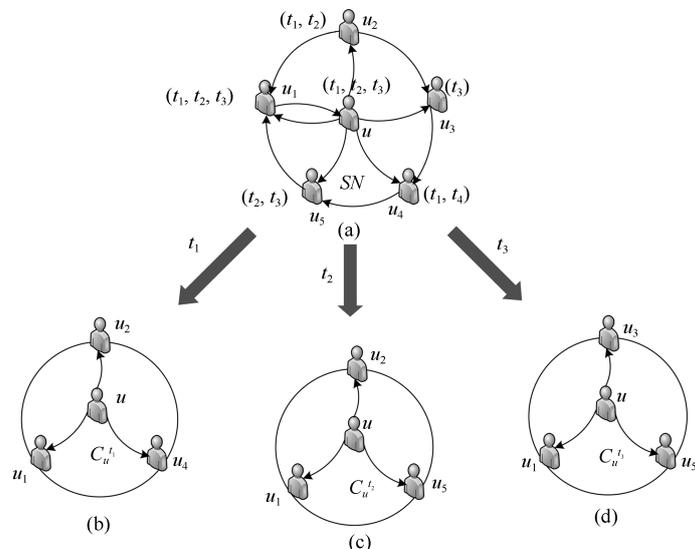


图 2 构造购买用途需求的社交圈

Fig. 2 The inference process of purpose-based social circle of users

$$R_{u,i}^{c*} = \frac{R_{u,i}^c}{\sum_{i \in H_u^c} R_{u,i}^c} \quad (7)$$

其中, H_u^c 是在类别 c 用户 u 评分的项集合. 类似于 Feng 和 Qian^[2] 的方法, 因为一个项属于不同的类别, 使用类别分布向量 D_i 表示项 i 的主题.

$$D_i = [I_{c_1}, I_{c_2}, \dots, I_{c_n}] \quad (8)$$

其中, I_{c_n} 是指示函数, 如果项 i 属于类别 c_n , 则 $I_{c_n} = 1$, 反之亦然. n 是目录的数量.

定义 2. 个人偏好是用户 u 在特定类别 c 下对汽车主题的偏爱分布情况, 用 D_u^c 表示.

如果 u 在类别 c 中具有评分, 则其个人偏好应该类似于项类别分布向量的加权平均值. u 在类别 c 中的个人偏好为

$$D_u^c = \sum_{i \in H_u^c} R_{u,i}^{c*} D_i \quad (9)$$

如果 u 从来没有为类别 c 中的项评分, 作为其个人偏好的近似值, 使用属于他的社交圈的朋友的个人偏好的平均值. 从而

$$D_u^c = \frac{1}{|C_u^c|} \sum_{v \in C_u^c} D_v^c \quad (10)$$

其中, $|C_u^c|$ 代表社交圈 C_u^c 的用户数.

在实践中, 用户 u 的个人偏好与项目 i 的主题的相关性被视为用户 u 在项目 i 上的隐实际评分值, 用 $Q_{u,i}$ 表示.

$$Q_{u,i} = \text{Cosin}(D_u^c, D_i) \quad (11)$$

其中, Cosin 表示两个向量间的余弦相似度.

2.1.3 偏好相似度

定义 3. 偏好相似度是在特定类别 c 下, 用户 u 和 v 之间的主题偏好向量相似度, 用 $F_{u,v}^c$ 表示. 具体来说, 本文采用 Feng 和 Qian^[2] 提出的方法来度量偏好相似度.

$$F_{u,v}^c = \text{Cosin}(D_u^c, D_v^c) \quad (12)$$

归一化后如下:

$$F_{u,v}^{c*} = \frac{F_{u,v}^c}{\sum_{v \in C_u^c} F_{u,v}^c} \quad (13)$$

2.1.4 评论隐特征

CNN 的目标是从汽车评论文本数据中学习文本的隐特征向量. 图 3 显示了 CNN 的结构, 包括 Embedding 层、卷积层、Pooling 层和输出层.

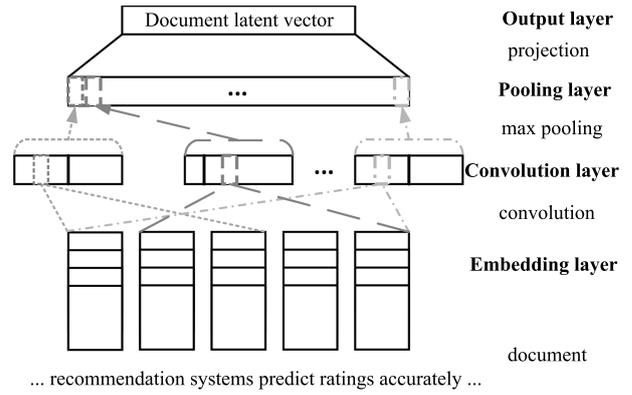


图 3 卷积神经网络结构

Fig. 3 Structure of convolutional neural network

1) Embedding 层

该层将预处理的汽车评论文本变换成表示下一个卷积层文档的密集矩阵. 具体来说, 对于作为 l 个词的序列文档, 通过连接文档中词的向量将文档表示为矩阵. 词向量由随机初始化或用词向量嵌入模型 GloVe 预先训练好, 通过优化过程进一步训练, 文档矩阵 $D \in \mathbf{R}^{p \times l}$ 变为

$$D = \begin{bmatrix} \cdots & | & | & | & \cdots \\ \cdots & w_{i-1} & w_i & w_{i+1} & \cdots \\ \cdots & | & | & | & \cdots \end{bmatrix} \quad (14)$$

其中, l 为文档长度, p 为 w_i 的嵌入维度大小.

2) 卷积层

卷积层提取上下文特征.

定义 4. 上下文特征向量用来提取汽车评论文本的上下文信息的向量, 用 c^j 表示.

具体地, 上下文特征 $c_i^j \in \mathbf{R}$ 由第 j 个共享权重 $W_c^j \in \mathbf{R}^{p \times ws}$ 提取, 其窗口大小 ws 确定周围词的数量.

$$c_i^j = f(W_c^j * D_{(:,i:(i+ws-1))} + b_c^j) \quad (15)$$

其中, $*$ 是卷积算子, $b_c^j \in \mathbf{R}$ 是 W_c^j 的偏差, f 是非线性激活函数. 在诸如 sigmoid、tanh 和 ReLU 的非线性激活函数中, 使用 ReLU 来避免梯度消失的问题, 导致优化收敛慢并且可能导致差的局部最小值^[19-20]. 带有权重 W_c^j 文档的上下文特征向量 $c^j \in \mathbf{R}^{l-ws+1}$ 由下式构成:

$$c^j = [c_1^j, c_2^j, \dots, c_i^j, \dots, c_{l-ws+1}^j] \quad (16)$$

然而, 一个共享权重只能捕获一种类型的上下文特征. 因此, 使用多个共享权重来捕获多种类型的上下文特征, 能够生成与 W_c 数量 n_c 一样多的上下文特征向量. 例如 W_c^j 中的 $j = 1, 2, 3, \dots, n_c$.

3) Pooling 层

Pooling 层从卷积层提取表征特征, 并且通过构建固定长度特征向量的池化操作来处理文档的可变长度. 在卷积层之后, 文档被表示为 n_c 个上下文特征向量, 其中每个上下文特征向量具有可变长度 (即 $l - ws + 1$ 个上下文特征). 然而, 这种表示存在两个问题: 1) 存在太多的上下文特征 c_i , 其中大多数上下文特征可能不会帮助增强其性能; 2) 上下文特征向量的长度变化, 使得难以构建接下来的层. 因此, 利用 max-pooling, 通过从每个上下文特征向量中仅提取最大上下文特征, 将文档的表示减少到 n_c 个固定长度向量中.

$$\mathbf{d}_f = [\max(\mathbf{c}^1), \dots, \max(\mathbf{c}^j), \dots, \max(\mathbf{c}^{n_c})] \quad (17)$$

其中, \mathbf{c}^j 是从第 j 个共享权重 W_c^j 抽取的长度为 $l - ws + 1$ 的上下文特征向量.

4) 输出层

通常, 在输出层, 从前一层获得的高级特征应该被转换用于特定任务. 因此, 在推荐任务的用户和项目隐义模型的 k 维空间上投影 \mathbf{d}_f , 通过使用常规非线性投影最终产生评论隐向量.

定义 5. 评论隐向量是表示汽车评论文本中的隐义特征向量, 又作为评论隐特征, 用 \mathbf{s}_i 表示.

$$\mathbf{s}_i = \tanh(W_{f_2} \{ \tanh(W_{f_1} + \mathbf{b}_{f_1}) \} + \mathbf{b}_{f_2}) \quad (18)$$

其中, $W_{f_1} \in \mathbf{R}^{f \times n_c}$, $W_{f_2} \in \mathbf{R}^{k \times f}$ 为投影矩阵, 并且在 $\mathbf{s}_i \in \mathbf{R}^k$ 中, $\mathbf{b}_{f_1} \in \mathbf{R}^f$, $\mathbf{b}_{f_2} \in \mathbf{R}^k$ 分别为 W_{f_1} , W_{f_2} 的偏置向量.

通过上述过程, CNN 架构以预处理的汽车评论文本为输入函数, 返回每个文档的隐向量作为输出.

$$\mathbf{s}_i = \text{cnn}(W, X_i) \quad (19)$$

其中, W 表示所有权重和偏差变量, 以防止混乱. X_i 表示项目 i 的文档, \mathbf{s}_i 表示项目 i 的评论隐向量.

3 汽车推荐模型

3.1 SCTCMAR 模型

本文采用低阶矩阵分解技术将用户-评分矩阵分为用户隐特征矩阵 $U^T \in \mathbf{R}^{n \times k}$ 和汽车隐特征矩阵 $V \in \mathbf{R}^{m \times k}$, k 是隐空间的向量维度, 对于每个类别 c , 汽车推荐模型 SCTCMAR 的目标函数为

$$\begin{aligned} L^c(R^c, U^c, V^c, W^c, F^{c*}, Q^{c*}) = & \\ & \frac{1}{2} \sum_{(u,i) \text{ obs}} (R_{u,i}^c - \hat{R}_{u,i}^c)^2 + \frac{\lambda_U}{2} \|U^c\|_F^2 + \\ & \frac{\lambda_V}{2} \|V^c - \text{cnn}(W^c, X^c)\|_F^2 + \end{aligned}$$

$$\begin{aligned} & \frac{\lambda_W}{2} \sum_k \|W_k^c\|_F^2 + \\ & \frac{\beta}{2} \sum_{u \in U^c} \left((U_u^c - \sum_{v \in C_u^c} F_{u,v}^{c*} U_u^c) (U_u^c - \sum_{v \in C_u^c} F_{u,v}^{c*} U_u^c)^T \right) + \\ & \frac{\gamma}{2} \sum_{(u,i) \text{ obs}} |H_u^{c*}| (Q_{u,i}^{c*} - U_u^c V_i^{cT})^2 \quad (20) \end{aligned}$$

其中, $|H_u^{c*}|$ 指已评分汽车集合经过正则化后的汽车数量, $R_{u,i}^c$ 表示用户 u 对汽车 i 的实际评分, $\hat{R}_{u,i}^c$ 为通过目标函数捕获的预测评分. 一般来说, 评分项目的数量越多, 用户 u 越可能依赖自己个人的偏好做出选择. 式 (20) 中, 第 2 行表示评论文本因素对汽车的影响; 第 3 行表示用户的朋友偏好对用户的影响; 最后一行表示用户个人偏好的因素影响.

3.2 模型训练

针对推荐模型 SCTCMAR 的目标函数, 使用梯度下降法进行优化, 将 U_u^c 和 V_i^c 看作变量并求偏导.

$$\begin{aligned} \frac{\partial L^c}{\partial U_u^c} = & \sum_{i:i \in C_u^c} I_{u,i}^{R^c} (r_m^c + U_u^c V_i^{cT} - R_{u,i}^c) V_i^c + \\ & \lambda_U U_u^c + \beta (U_u^c - \sum_{v \in C_u^c} F_{u,v}^{c*} U_u^c) - \\ & \beta \sum_{v:u \in C_u^c} F_{u,v}^{c*} (U_u^c - \sum_{w \in C_u^c} F_{u,w}^{c*} U_w^c) + \\ & \gamma \sum_{i:i \in H_u^c} I_{u,i}^{R^c} |H_u^{c*}| (U_u^c V_i^{cT} - Q_{u,i}^{c*}) V_i^c \quad (21) \end{aligned}$$

$$\begin{aligned} \frac{\partial L^c}{\partial V_i^c} = & \sum_{u \in U^c} I_{u,i}^{R^c} (r_m^c + U_u^c V_i^{cT} - R_{u,i}^c) U_u^c + \\ & \lambda_V (V_i^c - \text{cnn}(W^c, X^c)) + \\ & \gamma \sum_{u \in U^c} I_{u,i}^{R^c} |H_u^{c*}| (U_u^c V_i^{cT} - Q_{u,i}^{c*}) U_u^c \quad (22) \end{aligned}$$

其中, $I_{u,i}^{R^c}$ 表示指示函数, $I_{u,i}^{R^c} = 1$ 代表用户 u 对汽车 i 有过评分, 否则 $I_{u,i}^{R^c} = 0$.

求偏导后同时更新用户隐特征向量和汽车隐特征向量, 然后把更新后的隐特征向量代入目标函数中重新计算, 使目标函数值最快速度下降, 循环迭代该过程直到目标函数收敛为止, 最终经过训练得到用户隐特征矩阵 U_u^{cT} 和汽车隐特征矩阵 V_i^c .

然而, 由于 W 与 CNN 结构中的特征 (例如 max-pooling 层和非线性激活函数) 密切相关, 所以

W 不能像对 U 和 V 一样由梯度下降的方式来优化. 然而, 当 U 和 V 是暂时恒定时, 我们观察到 L 可以被解释为具有如下的 $L2$ 正则化项的平方误差函数.

$$\begin{aligned} \varepsilon(W) = & \frac{\lambda_V}{2} \sum_i^M \|(v_i - \text{cnn}(W, X_i))\|^2 + \\ & \frac{\lambda_W}{2} \sum_k^{|w_k|} \|w_k\|^2 + \text{constan}(t) \end{aligned} \quad (23)$$

其中, constan 表示返回一个常量.

W 是每一层的权重和偏置, 为了优化 W , 使用反向传播算法. 重复整个优化过程 (U, V, W 交替地更新) 直到收敛. 通过优化 U, V, W , 最终可以预测用户对汽车的未知评分.

$$\hat{R} = \hat{r} + u_u^T v_i = \hat{r} + u_u^T (\text{cnn}(W, X_i) + \varepsilon_i) \quad (24)$$

4 实验及结果分析

4.1 数据集

本文实验主要使用网络爬虫从 Internet 抓取的汽车基本信息、用户信息、用户对汽车评论信息和与汽车有关的语料文本集. 汽车和用户数据主要来源于汽车之家和网易汽车, 预训练词嵌入模型主要将语料数据集作为训练数据, 通过使用 GloVe 工具¹来构建. 该工具使用的是一种无监督学习算法, 主要用于获取词特征向量, 本质上是具有加权最小二乘的 log-bilinear 模型, 其模型的主要来源于观察的现象, 即词-词共同出现可能性的比例具有编码某种形式的意义的潜力. 此外, 分词过程中使用了数据堂和搜狗输入法提供的汽车领域字典以及爬取的互联网上的汽车字典资源, 如表 2 所示.

汽车评论文本预处理的主要任务包括, 数据清理和中文分词两部分. 1) 清理: 由于实验数据主要是通过网络爬虫技术从相关的汽车网站上抓取得到的, 所以数据信息的格式有很多不规范的地方, 存在结构化和非结构化数据. 得到的数据中存在 HTML 标签、Javascript 代码、广告标语等不同形式的噪点, 需要进行去噪、清洗, 转为合适的数据结构, 便于使用和理解. 去噪主要采用正则表达式, 对广告标语主要靠人工清理. 2) 分词: 在自然语言处理的相关研究工作中, 分词占据着重要的地位. 对于中文语料, 分词工作更是如此. 本文采用 ICTCLAS² 系统作为研究工作中的分词工具, 主要使用基于词典的分词功能.

爬取的数据经过清理、去重和整合处理之后, 最终数据集中有 266 995 个用户信息、702 辆汽车信息以及用户购买的汽车信息, 其中每一个用户的购车用途需求有多个, 例如上班、接送孩子、自驾游等; 汽车车型有 11 种类型, 如跑车、小轿车、SUV、皮卡等. 用户对其购买汽车的评分主要是 1~5 的整数. 本文将数据集分成 5 组, 进行 5-fold 交叉验证来评估算法的性能.

4.2 评价指标体系

针对汽车的推荐性能, 主要从三个方面进行评价: 1) 汽车推荐列表中用户真正购买的比例; 2) 在用户-汽车测试集中向用户成功推荐的比例; 3) 汽车推荐列表中用户最终购买汽车的排名. 评价指标为精确率 (Precision) 和召回率 (Recall).

$$\text{Precision} = \frac{H}{N \times T} \quad (25)$$

$$\text{Recall} = \frac{H}{T} \quad (26)$$

表 2 数据集信息列表
Table 2 The list of datasets

名称	来源/网址	用途
汽车之家	http://www.autohome.com.cn	汽车及用户的相关数据信息
网易汽车	http://auto.163.com	汽车及用户的相关数据信息
搜狗实验室	http://www.sogou.com/labs/	语料数据集
汽车之家	http://www.autohome.com.cn	语料数据集
网易汽车	http://auto.163.com	语料数据集
腾讯汽车	http://auto.qq.com/	语料数据集
新浪汽车	http://auto.sina.com.cn/	语料数据集
数据堂	http://www.datatang.com/	汽车领域字典
输入法字典	搜狗输入法	汽车领域字典

¹GloVe, available: <https://nlp.stanford.edu/projects/glove>, 2017

²ICTCLAS, available: <http://ictclas.nlpir.org>, June 10, 2017

其中, T 为用户数, N 表示系统为用户推荐的汽车列表长度, H 指最终购买汽车推荐列表中的汽车的用户数.

对于汽车这一特定商品, 还采用平均倒数排名 (Average reciprocal hit rank, ARHR)^[21] 对推荐排序准确度进行评价. 排序准确度评估指标衡量的是系统产生的推荐集中商品的排序与用户个人偏好的相似度, 平均倒数排名衡量推荐系统向用户输出的效用性, 用户最终购买的商品在推荐列表中的排名越高, 该推荐系统的效用性越强. 在推荐列表中, 排序位置 p_i 的成功推荐效果值设置为 $1/p_i$, 然后进行迭代, 迭代后求平均值得到平均倒数排名值. 具体如下:

$$ARHR = \frac{1}{T} \sum_{i=1}^H \frac{1}{p_i} \quad (27)$$

4.3 与其他推荐算法的对比

为了验证提出的汽车推荐模型 SCTCMAR 的有效性, 选用具有代表性的推荐算法与 SCTCMAR 进行比较, 被选入实验对比的四种推荐算法如下:

1) FMM (Flexible mixture model). 由 Si 等^[22] 提出的灵活混合模型. FMM 通过将用户和项目同时聚类在一起, 同时进行协同过滤来扩展现有的分区/聚类算法, 而不必假定每个用户和项目只应属于单个集群. 实验中, 按照该算法的主要思想, 将用户和汽车以购买用途需求分别进行聚类, 实现了 FMM 模型, 最终完成测试用户的汽车推荐任务.

2) TR (Trust rank). 由 Zou 等^[23] 提出的算法. 该算法利用可用于电子商务的用户信任网络来处理冷启动问题给推荐系统带来的限制. 用户信任网络由用户在其中指定的友谊或信任关系形成, 主要通

过应用个性化的 PageRank 算法, 扩展了给定用户的朋友, 即其他具有类似购买记录的人的朋友. 在实验中, 根据该模型的主要想法, 构造一个以购车用户组成的社交网络, 并且实现了该模型和算法.

3) RS (Random sampling). 从整个汽车信息集中随机抽取 N 辆汽车向用户推荐, 作为其他算法的对比实验.

4) SCTCMAR+. 本文提出的推荐算法的另外一个版本, 该方法使用了 GloVe 预训练词嵌入模型.

现有的深度学习平台和工具种类繁多^[24], 本文使用 Python 和 Keras³ 库来实现本文模型 SCTCMAR. 良好的学习率策略可以显著提高深度学习模型的收敛速度, 减少模型的训练时间^[25]. 为了训练 CNN 的权重, 本文使用小批量的 RMSprop (自动调节学习速率的深度学习优化方法), 每个小批量包括 128 个训练项目. 至于详细的 CNN 架构, 本文使用以下设置: 1) 将文档的最大长度设置为 3000. 2) SCTCMAR: 随机初始化维度大小为 50 的词潜在向量. SCTCMAR+: 通过尺寸大小为 50 的预训练词嵌入模型初始化词潜在向量. 两种词潜在向量都将通过优化过程训练. 3) 在卷积层中, 使用的各种窗口大小为 3, 4 和 5, 并且使用每个窗口大小的 100 个共享权重. 4) 代替与 CNN 的权重相关的 L_2 正则化器, 使用 dropout 和设置 dropout 率为 0.2 以防止 CNN 过度拟合. 本文按照原文的调优成果将 FMM 算法中的用户和汽车种类分别设为 10 和 20, 对于 TR 算法和其他参数的设置设为默认值.

在实验对比中, 首先对比 N 取值为 5, 10, 15, 20 的情况下的精确率和召回率. 表 3、表 4 和图 4 是各算法的精确率和召回率的比较, 以及 SCTCMAR+ 相对于其他方法提升百分比的实验结果.

表 3、表 4 和图 4 表明, SCTCMAR 相较于其

表 3 本文提出的算法与其他算法在推荐精确率上的比较及提升百分比

Table 3 Precision comparison and improvement between our algorithm and other algorithms

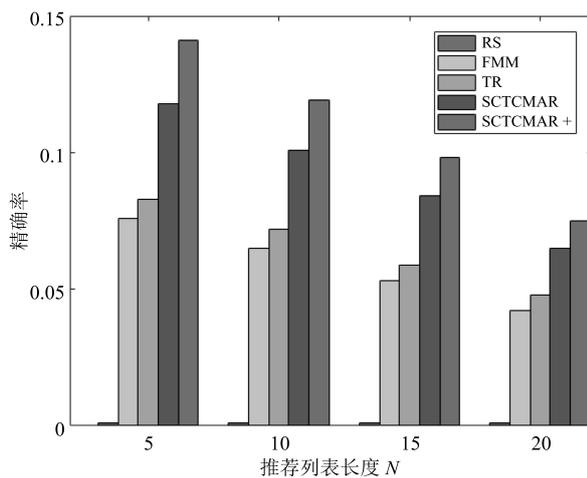
算法	$N = 5$	$N = 10$	$N = 15$	$N = 20$
FMM	0.076	0.065	0.053	0.042
	46.10 %	45.38 %	45.92 %	44.00 %
TR	0.083	0.072	0.059	0.048
	41.13 %	39.50 %	39.80 %	36.00 %
RS	0.001	0.001	0.001	0.001
	99.29 %	99.16 %	98.98 %	98.67 %
SCTCMAR	0.118	0.101	0.084	0.065
	16.31 %	15.13 %	14.29 %	13.33 %
SCTCMAR+	0.141	0.119	0.098	0.075

³F.Chollet.Keras. <http://github.com/fchollet/keras>, 2017

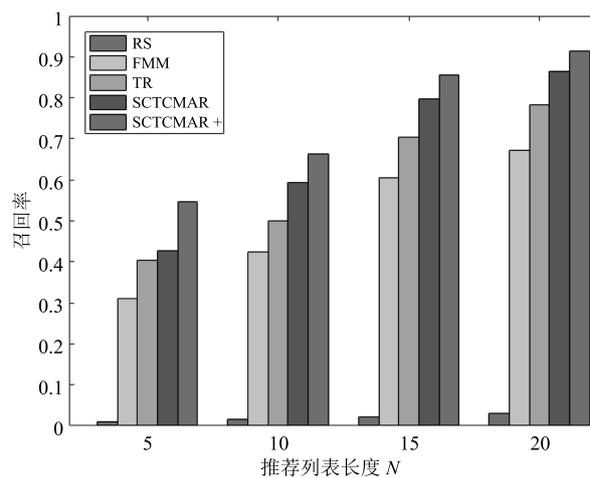
表 4 本文提出的算法与其他算法在推荐召回率上的比较及提升百分比

Table 4 Recall comparison and improvement between our algorithm and other algorithms

算法	$N = 5$	$N = 10$	$N = 15$	$N = 20$
FMM	0.311 43.04 %	0.425 35.80 %	0.604 29.52 %	0.672 26.56 %
TR	0.405 25.82 %	0.501 24.32 %	0.705 17.74 %	0.783 14.43 %
RS	0.010 98.17 %	0.016 97.58 %	0.023 97.32 %	0.032 96.50 %
SCTCMAR	0.472 13.55 %	0.593 10.42 %	0.798 6.88 %	0.864 5.57 %
SCTCMAR+	0.546	0.662	0.857	0.915



(a) Comparison of precision of each algorithm



(b) Comparison of recall of each algorithm

图 4 在精确率和召回率上本文提出的算法与其他算法的对比

Fig 4 Precision and recall comparisons between our algorithm and other algorithms

他方法在精确率和召回率上的性能好, 进一步验证了算法对于汽车这个特殊商品的推荐有效性. 无论是精确率还是召回率, 随机抽取算法比其他各个算法都低很多. 具体来说, 在精确率方面, 从图 4(a) 可以看出, SCTCMAR 汽车推荐模型比 FMM 和 TR 推荐算法性能要好, 相对于 FMM 在 $N = 5$ 时提升了 46.1%, 而相对于 TR 在 $N = 10$ 时提高了 41.1%; 召回率方面, 从图 4(b) 可以看出, 当 $N = 5$ 时, SCTCMAR 达到 54.6%, 相对于 FMM 上升了 43%, TR 上升了 25.8%. 从图 4 可以看出, N 值越大, 各个算法的推荐精准率逐渐下降, 即推荐列表中的汽车越来越多, 算法错误推荐的概率也越大; 而召回率的走势与精确率相反, 随着 N 值增大, 召回率逐渐上升, N 值越大, 算法预测到用户最终购买汽车的概率越大.

但是, 从表 3 和图 4 可以发现, 相对于一般消费商品来说, 本文中各个算法精确率都很低, 主要是由精确率的定义和汽车这种特定商品的属性造成的. 从精确率的定义来看, 它计算推荐列表中用户真正购买的汽车比例. 对汽车而言, 用户通常情况下只会选购列表中的一辆, 即所占比例不会超过 $1/N$.

为了对系统推荐准确性进行更好的评估, 本文还采用了平均倒序排名 ARHR^[21] 这一评价标准. 平均倒序排名值越大, 表明用户真正购买的汽车在推荐列表中的排名越靠前, 对于购买汽车用户的推荐效果越优. 实验结果如表 5 和图 5 所示.

从图 5 可以看出, 各算法远比随机抽取的平均倒序排名高. 在汽车这种特定领域商品 Top- N 推荐任务中, SCTCMAR 相比于 FMM 和 TR 算法在平均倒序排名上有更好的推荐准确度. 进一步可知, N

值越大, 各推荐算法的平均倒序排名越高, 最终慢慢趋于平稳. 说明该评价标准对于汽车商品推荐的有效性, 也进一步说明将社交因素与评论文本特征融合到推荐模型的有效性.

表5 各算法在平均倒序排名上的对比

Table 5 ARHR comparison between our algorithm and other algorithms

算法	$N = 5$	$N = 10$	$N = 15$	$N = 20$
FMM	0.082	0.106	0.122	0.131
TR	0.119	0.137	0.148	0.154
RS	0.002	0.003	0.005	0.006
SCTCMAR	0.171	0.182	0.201	0.210
SCTCMAR+	0.196	0.218	0.233	0.241

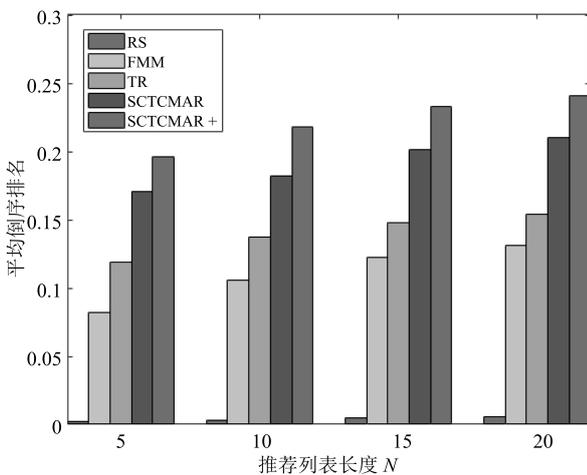


图5 各算法在平均倒序排名上的对比

Fig 5 ARHR comparison between our algorithm and other algorithms

4.4 进一步分析

1) 预训练词嵌入模型的影响

与协同过滤的主题回归 (CTR) 和协同深度学习 (CDL) 不同, SCTCMAR 能够利用预先训练的词语嵌入模型, 例如 GloVe^[17]. 因此, 通过使用 GloVe 预先训练的词嵌入模型初始化 SCTCMAR 的 CNN 的嵌入层来分析预训练的词嵌入模型对汽车推荐任务的影响. 表 3~5 显示了 SCTCMAR+ 与 SCTCMAR 在数据集上的实际变化, 精确率、召回率和平均倒序排名都有不错的提升, 但是可以看出词嵌入模型对 SCTCMAR 的影响与推荐列表长度 N 没有直接关系. 为了进一步分析预训练词嵌入模型的影响, 实验选取皮卡 (pika)、面包 (mianbao)、小型车 (small) 和紧凑型车 (compact) 四种车型在

$N = 5$ 时的平均倒序排名影响. 其中车型的评分数据稀疏度是 $pika > mianbao > small > compact$. 实验结果如表 6 所示.

表6 本文算法使用词嵌入模型在平均倒序排名上的影响

Table 6 ARHR comparison between our algorithm with pre-trained word embedding model

算法	pika	mianbao	small	compact
SCTCMAR	0.171	0.169	0.184	0.186
SCTCMAR+	0.188	0.183	0.180	0.176
提升 (%)	9.04	7.65	-2.22	-5.68

在表 6 中, 尽管有微小的变化, 可以观察到一致的趋势, 评分数据越来越稀疏, 预先训练的词嵌入模型有助于提高 SCTCMAR 的性能. 因为预先训练的词嵌入模型的语义和句法信息补充了评分的不足. 还可以观察到, 给定足够数量的评分来训练, 则预先训练的词嵌入模型使 SCTCMAR 的性能恶化. 换句话说, 由于评分直接反映了用户和项目之间的关系, 最好是充分利用评分信息, 而不是在评估数据相对密集的时候, 使用从外部文档获得的预先训练的词嵌入模型.

2) 因素组合的影响

在实验中将研究的因素的参数设为一定值, 将剩余其他的因素的参数设为 0. 在 SCTCMAR 模型中, 对于个人偏好因素, 设定 $\gamma = 5$ 和 $cnn(W^c, X^c) = \lambda_w = \beta = 0$, 对于偏好相似度因素, 设定 $\beta = 5$ 和 $cnn(W^c, X^c) = \lambda_w = \gamma = 0$; 对于评论隐特征因素设定 $\beta = \gamma = 0$. SCTCMAR 模型有两个社交因素, 因此性能比较时设置相同的参数.

我们系统地比较了不使用因素以及使用 1 个、2 个、3 个因素的各种方法, 比较结果如图 6 所示. UP, PS, CTC 分别表示 SCTCMAR 中的个人偏好、偏好相似度和评论隐特征. + 表示因素组合. 通过比较可以发现, 所有因素对提高平均倒序排名都有积极影响. 进一步, SCTCMAR 模型中的每一个因素比 PMF 中的每个社会因素都有更好的表现. 此外, SCTCMAR 的多因素组合比其他方法更好, 例如 FMM, Trust rank 和 Random sampling.

5 结语

本文主要针对汽车推荐进行研究, 重点探索社交因素和评论文本因素对用户购车决策过程的影响, 融合以上两方面提出了新颖的 SCTCMAR 推荐模型. 虽然实验结果表明 SCTCMAR 模型在一定程度上改善了推荐性能, 但是还没有考虑用户购车需求与决策过程中更多的关联关系以及更多的汽车属

性信息. 下一步工作将考虑将更多的汽车属性信息和交互关系信息融入到推荐模型中, 进一步提高推荐精度, 并找到更有趣的规律.

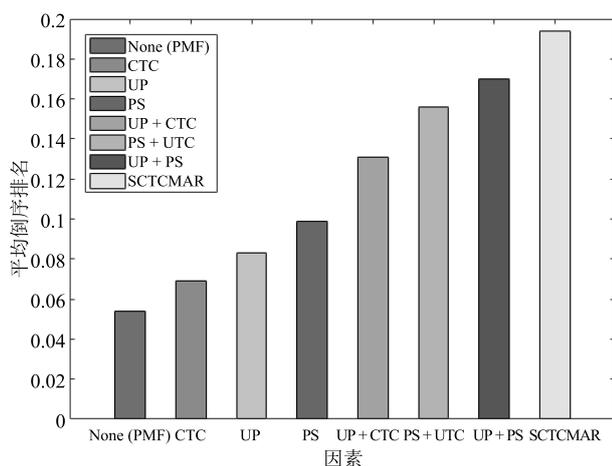


图 6 本文算法多因素联合对平均倒序排名的影响

Fig 6 ARHR line chart of impact of factors combination in our model

References

- Felfernig A, Burke R. Constraint-based recommender systems: technologies and research issues. In: Proceedings of the 10th International Conference on Electronic Commerce. Innsbruck, Austria: ACM, 2008.
- Feng H, Qian X M. Recommendation via user's personality and social contextual. In: Proceedings of the 22nd ACM International Conference on Information and Knowledge Management. San Francisco, California, USA: ACM, 2013. 1521–1524
- Jamali M, Ester M. A matrix factorization technique with trust propagation for recommendation in social networks. In: Proceedings of the 4th ACM Conference on Recommender Systems. Barcelona, Spain: ACM, 2010. 135–142
- Pan Tao-Tao, Wen Feng, Liu Qin-Rang. Collaborative filtering recommendation algorithm based on rating matrix filling and item predictability. *Acta Automatica Sinica*, 2017, **43**(9): 1597–1606
(潘涛涛, 文锋, 刘勤让. 基于矩阵填充和物品可预测性的协同过滤算法. *自动化学报*, 2017, **43**(9): 1597–1606)
- Yang X W, Steck H, Liu Y. Circle-based recommendation in online social networks. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Beijing, China: ACM, 2012. 1267–1275
- Salakhutdinov R, Mnih A. Probabilistic matrix factorization. In: Proceedings of the 20th International Conference on Neural Information Processing Systems. Vancouver, British Columbia, Canada: ACM, 2008. 1257–1264
- Jiang M, Cui P, Liu R, Yang Q. Social contextual recommendation. In: Proceedings of the 21st ACM international Conference on Information and Knowledge Management. Maui, Hawaii, USA: ACM, 2012. 45–54
- Zhang Yan-Ping, Zhang Shun, Qian Fu-Lan, Zhang Yi-Wen. Robust collaborative recommendation algorithm based on user's reputation. *Acta Automatica Sinica*, 2015, **41**(5): 1004–1012
(张燕平, 张顺, 钱付兰, 张以文. 基于用户声誉的鲁棒协同推荐算法. *自动化学报*, 2015, **41**(5): 1004–1012)
- Qian X M, Feng H, Zhao G S, Mei T. Personalized recommendation combining user interest and social circle. *IEEE Transactions on Knowledge and Data Engineering*, 2014, **26**(7): 1763–1777
- Wang Z, Sun L F, Zhu W W, Yang S Q, Li H Z, Wu D P. Joint social and content recommendation for user-generated videos in online social network. *IEEE Transactions on Multimedia*, 2013, **15**(3): 698–709
- Wang C, Blei D M. Collaborative topic modeling for recommending scientific articles. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego, California, USA: ACM, 2011. 448–456
- Ling G, Lyu M R, King I. Ratings meet reviews, a combined approach to recommend. In: Proceedings of the 8th ACM Conference on Recommender Systems. Foster City, Silicon Valley, California, USA: ACM, 2014. 105–112
- McAuley J, Leskovec J. Hidden factors and hidden topics: understanding rating dimensions with review text. In: Proceedings of the 7th ACM Conference on Recommender Systems. Hong Kong, China: ACM, 2013. 165–172
- Wang H, Wang N Y, Yeung D Y. Collaborative deep learning for recommender systems. In: Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Sydney, NSW, Australia: ACM, 2015. 1235–1244
- Lecun Y, Bottou L, Bengio Y, Haffner F. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998, **86**(11): 2278–2324
- Sun Xiao, Pan Ting, Ren Fu-Ji. Facial expression recognition using ROI-KNN deep convolutional neural networks. *Acta Automatica Sinica*, 2016, **42**(6): 883–891
(孙晓, 潘汀, 任福继. 基于 ROI-KNN 卷积神经网络的面部表情识别. *自动化学报*, 2016, **42**(6): 883–891)
- Pennington J, Socher R, Manning C D. GloVe: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: EMNLP, 2014. 1532–1543
- Kim D, Park C, Oh J, Lee S, Yu H. Convolutional matrix factorization for document context-aware recommendation. In: Proceedings of the 10th ACM Conference on Recommender Systems. Boston, Massachusetts, USA: ACM, 2016. 233–240

- 19 Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks. In: Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. Lauderdale, USA: AISTATS, 2011. 315–323
- 20 Maas A L, Hannun A Y, Ng A Y. Rectifier nonlinearities improve neural network acoustic models. In: Proceedings of the 30th International Conference on Machine Learning Workshop on Deep Learning for Audio Speech and Language Processing. Atlanta, Georgia, USA: ICML, 2013.
- 21 Deshpande M, Karypis G. Item-based top- N recommendation algorithms. *ACM Transactions on Information Systems (TOIS)*, 2004, **22**(1): 143–177
- 22 Si L, Jin R. Flexible mixture model for collaborative filtering. In: Proceedings of the 20th International Conference on Machine Learning (ICML-03). Washington, DC, USA: ICML, 2003. 704–711
- 23 Zou H T, Gong Z G, Zhang N, Zhao W, Guo J Z. TrustRank: a Cold-Start tolerant recommender system. *Enterprise Information Systems*, 2015, **9**(2): 117–138
- 24 Xi Xue-Feng, Zhou Guo-Dong. A survey on deep learning for natural language processing. *Acta Automatica Sinica*, 2016, **42**(10): 1445–1465
(奚雪峰, 周国栋. 面向自然语言处理的深度学习研究. *自动化学报*, 2016, **42**(10): 1445–1465)
- 25 He Yu-Yao, Li Bao-Qi. A combinatory form learning rate scheduling for deep learning model. *Acta Automatica Sinica*, 2016, **42**(6): 953–958
(贺昱曜, 李宝奇. 一种组合型的深度学习模型学习率策略. *自动化学报*, 2016, **42**(6): 953–958)

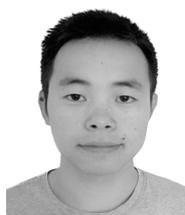


冯永 重庆大学计算机学院教授. 主要研究方向为大数据分析数据挖掘, 大数据管理与智能推荐, 大数据集成与人工智能. 本文通信作者.

E-mail: fengyong@cqu.edu.cn

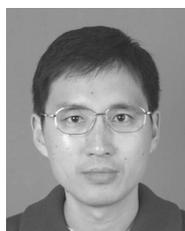
(FENG Yong Professor at the College of Computer Science, Chongqing University. His research interest covers

big data analysis and data mining, big data management and intelligent recommendation and big data integration and artificial intelligence. Corresponding author of this paper.)



陈以刚 重庆大学计算机学院硕士研究生. 主要研究方向为智能推荐与神经网络. E-mail: 20141413081@cqu.edu.cn

(CHEN Yi-Gang Master student at the College of Computer Science, Chongqing University. His research interest covers intelligent recommendation and neural networks.)



强保华 桂林电子科技大学广西云计算与大数据协同创新中心教授. 主要研究方向为大数据处理与信息检索.

E-mail: qiangbh@guet.edu.cn

(QIANG Bao-Hua Professor at the Guangxi Cooperative Innovation Center of Cloud Computing and Big Data, Guilin University of Electronic Technol-

ogy. His research interest covers big data processing and information retrieval.)