基于文本注意力的推荐系统可解释性研究

朱 芮[1], 刘布楼[2], 刘艺语[1], 邹鑫雨[1], 李晨亮[1]

[1]武汉大学国家网络安全学院 武汉 中国 430072

[2]清华大学计算机科学与技术系 北京 中国 100084

摘要 可解释性能够提高用户对推荐系统的信任度并且提升推荐系统的说服力和透明性, 因此有许多工作都致力于实现推荐系统的可解释性。由于评论中包含了丰富的信息, 能够体现用户偏好与情感信息, 同时包含了对应商品所具有的特性, 最近的一些基于评论的深度推荐系统有效地提高了推荐系统的可解释性。这些基于评论的深度推荐系统中内置的注意力机制能够从对应的评论中识别出有用的语义单元(例如词、属性或者评论), 而推荐系统通过这些高权重的语义单元做出决策, 从而增强推荐系统的可解释性。但可解释性在很多工作中仅作为一个辅助性的子任务, 只在一些案例研究中来做出一些定性的比较, 来说明推荐系统是具有可解释性的, 到目前为止并没有一个能够综合地评估基于评论推荐系统可解释性的方法。本文首先根据在注意力权重计算机制的不同, 将这些具有可解释性的基于评论的推荐系统分为三类: 基于注意力的推荐系统, 基于交互的推荐系统, 基于属性的推荐系统, 随后选取了五个最先进的基于评论的深度推荐系统, 通过推荐系统内置的注意力机制获得的评论权重文档, 在三个真实数据集上进行了人工标注, 分别量化地评价推荐系统的可解释性。标注的结果表明不同的基于评论的深度推荐系统的可解释性是具有优劣之分的, 但当前的基于评论的深度推荐系统都有超过一半的可能性能够捕捉到用户对目标评论的偏好信息。在评估的五个推荐系统中, 并没有哪个推荐系统在所有的数据中具有绝对的优势。也就是说, 这些推荐系统在推荐可解释性方面是相互补充的。通过进一步的数据分析发现, 如果推荐系统具有更精确的分数预测结果, 那推荐系统通过注意力机制获得的高权重的信息确实更能够体现用户的偏好或者商品特征, 说明推荐系统内置的注意力机制在提高可解释性的同时也能够提高预测精度; 并且发现相较于长评论, 推荐系统更容易捕捉到较短的评论中的特征信息; 而可解释性评分高的推荐系统会更可能地为形容词赋予较高的权重。本文也为推荐系统可解释性评估进一步研究和探索更好的基于评论的推荐系统解决方案提供了一些启示。

关键词 推荐系统; 注意力机制; 可解释性; 用户评论; 深度学习

中图法分类号 TP391 **DOI 号** 10.19363/J.cnki.cn10-1380/tn.2021.09.10

Research on Interpretability of Recommendation System based on Text Attention Mechanism

ZHU Rui[1], LIU Bulou[2], LIU Yiyu[1], ZOU Xinyu[1], LI Chenliang[1]

[1] School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China

[2] Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

Abstract Interpretability can enhance users' trust in the recommendation systems, and improve the persuasion and transparency of the latter. So far, many efforts have been devoted to achieve recommendation interpretability. The rich information provided in user reviews can reflect user's preference and consumption experience, as well as the corresponding item's features. Hence, recent deep review-based recommendation systems capitalize reviews for accurate and interpretable recommendation and have advanced this purpose significantly. The built-in attention module devised in these deep review-based recommendation systems models can identify semantic units (e.g., words, aspects, or individual reviews) from the corresponding reviews, which also facilitate the interpretability of the recommendation systems. However, interpretability is typically taken as an auxiliary evaluation subtask in some works, where examples are used as case studies for some qualitative comparison to show that the recommendation system is interpretable. Right now, there is no comprehensive evaluation towards how good the interpretability delivered by these review-based recommendation systems are. In this paper, according to the different calculation methods of attention weight, we first summarize existing deep review-based recommendation systems into three categories: attention-based recommendation system, interaction-based recommendation system, and aspect-based recommendation system. Then, we perform a human evaluation based on the built-in attention mechanism of five state-of-the-art deep review-based recommendation systems across three real-world datasets, covering all three categories for interpretability evaluation. The annotation results suggest that the interpretability of different deep review-based recommendation systems is different, but the current deep review-based recommendation

通讯作者: 李晨亮, 博士, 教授, Email: cllee@whu.edu.cn

本课题得到国家自然科学基金(No. 61872278)资助。

收稿日期: 2021-4-30; 修改日期: 2021-08-05; 定稿日期: 2021-08-10

systems can successfully uncover more than half of user's preference for the target item with higher chance. We also note that there is no absolute winner in discovering user preference from all cases, among the five recommendation systems evaluated. That is, the models are complementary to each other in terms of recommendation interpretability. Through further data analysis, it is found that a higher recommendation accuracy often indicates that the highlighted information in the reviews is indeed relevant to the user's preferences or item's features. It shows that the built-in attention mechanism of the recommendation systems can not only enhance the interpretability, but also improve the prediction accuracy. Moreover, we found that compared with long reviews, recommendation systems are easier to capture the feature information in shorter reviews; and recommendation systems with high interpretability scores are more likely to give adjectives a higher weight. Overall, this work sheds some light on further research towards the development of interpretability evaluation and better review-based recommendation system solutions.

Key words recommendation systems; attention mechanism; interpretability; user reviews; deep learning

1 引言

推荐系统可以根据用户对商品的历史消费行为中挖掘用户个性化的兴趣爱好和商品独特的特征, 并根据这些挖掘到的特征为用户推荐可能感兴趣的产品或者服务。但这些基于交互记录的推荐系统存在着一些问题, 推荐系统通常只能给用户推荐某个用户感兴趣的商品, 但无法准确地捕捉到用户的兴趣点, 换个角度说, 推荐系统不能清晰地传递给用户为什么推荐这个商品, 即不具有可解释性。

可解释性是通过对推荐系统的决策提供合理的解释, 能够有效地提升推荐系统的透明度、说服力、可信赖性, 也能够提升用户的使用体验。用户评论是指用户在购买商品或者接受服务后, 对商品服务的文字性反馈, 评论中包含着丰富的关于用户个性喜好和商品特征的信息, 比如对商品性能的描述(例如商品的规格、质量)或者一些明显的情感倾向, 这为推荐系统能够更好地提取用户信息提供了数据支撑, 因此基于评论的推荐系统能够有效地提高推荐的准确性和可解释性。

早期的基于评论的推荐工作在主题模型的基础上在评论中提取用户和商品的潜在语义主题[1-3]。随着深度学习技术的发展, 为了提高推荐的可解释性, 有很多基于评论的推荐系统为评论文档中不同的评论、单词等语义单元赋予不同的注意力权重, 将注意力机制应用到推荐系统中[4], 通过这种方式使推荐系统更具有可解释性, 从而更有效地捕捉到评论文本中的"有用的"信息。例如, 图 1 是本文根据 ANR[5] 内置的注意力机制学习到的评论文档中单词的注意力权重所作的可视化表示, 其中颜色为红色的词是在本篇文档中注意力权重大于整体权重中值的词, 在这基础上, 红色字体越深, 代表着此单词的注意力权重越大。本条评论来自 Amazon 的 Music Instruments 数据集, 是一个用户对一个音乐电缆的评论, 从图 1 中可以看出, ANR 可以有效地为一些能描述商品特征的词赋予较高的权重, 例如"useful"、"different color", 也能够着重学习到描述用户情感的词, 例如"Great", 并且从这些高权重的单词中, 可以推测出此用户对这个商品是比较满意的, 事实上用户也确实为这条音乐电缆标记了 5 分的评分(5 分为最高分数)。

I like it. Very useful for connecting rank components between. Great idea in different colors. I will buy it again.

图 1 ANR 获得的带权重的评论实例

Figure 1 An instance of weighted review obtained by ANR

但在大部分推荐系统的衡量指标中通常只追求推荐结果的高准确率, 而忽视推荐的可解释性。事实上, 这些工作普遍将可解释性仅作为一个辅助性的评估子任务, 比如作为示例出现在一些案例研究中来做出一些定性的比较, 来表明推荐系统是具有可解释性的。但是到目前为止, 并没有一个通用的办法来对现存的先进的推荐系统的可解释性进行系统的分析和定量的评估。

本文提出了一种基于文本注意力机制的推荐系统的可解释性定量评估方法, 能够通用于任意应用了注意力机制的基于评论的深度推荐系统。本文通过判断基于评论的深度推荐系统内置的注意力机制, 是否能够真正捕捉到目标评论中所反映的用户偏好或者商品特征信息, 来对推荐系统的可解释性做出量化的评估分数。

本文的贡献包含下面 3 个内容:

(1) 本文提出了一种定量评价推荐系统可解释性的方法, 基于最先进的 5 个基于评论的深度推荐系统, 在 3 个现实的数据集 Amazon-Musical Instruments、Amazon-Office Products、Yelp 上各选取 200 条实例, 根据推荐系统内置的注意力机制获得的评论权重文档, 标注了总计 3000 条实例;

(2) 本文通过对可解释性评分结果分析发现, 当

前的基于评论的深度推荐系统有超过一半的可能性能够捕捉到用户对目标评论的偏好, 或者商品的特征信息;

(3) 本文探究了可解释性评分与评论的长度、推荐系统的预测精度、高权重词语之间的关联关系, 并通过进一步分析发现: 推荐系统在更短的评论上更容易捕捉到有用的信息; 对于可解释性较好的实例, 推荐系统对于这条实例的预测分数很有可能也会更加准确; 可解释性评分高的推荐系统会偏向于为更多的形容词赋予较高的权重。

2 相关工作

交互数据是由用户和商品之间的交互行为构成, 基于交互数据的推荐系统的核心方法是协同过滤技术[6], 推荐系统会向用户推荐与其相似的人购买过的商品或者与其购买过的商品相似的商品。而矩阵分解是协同过滤技术中最为常见的方法, 原始的矩阵分解模型[7]是将用户和商品映射到潜在特征空间中来将用户对商品的评分进行建模, 然后通过用户和商品潜在特征的点乘的结果挖掘一对用户与商品之间的关系。在这基础上, 很多工作对原始的矩阵分解模型做了优化, 比如将矩阵分解与邻域模型相结合[8], 认为用户对商品的评分不仅依赖于这对用户-商品对中的潜在特征, 也受用户对其他商品的评分的影响, 或者将其扩展到可以更加泛化地对特征进行建模的分解机模型[9], 通过对用户或商品间的时序行为进行建模, 来找寻影响当前用户或商品最大的邻居集合。但基于矩阵分解的推荐系统都是使用点乘操作作为最终的评分预测结果, 点乘操作只能确保潜在特征的线性结合, 但不能考虑到高级的特征交互。一些工作通过使用神经网络来学习用户的评分特征, NeuMF[10]为协同过滤提供了一个通用的深度神经框架, 将用户和商品的特征向量作为输入, 通过神经网络匹配来替代传统协同过滤中使用点乘计算评分。但矩阵的稀疏性仍然限制了推荐系统效果的提升, 协同过滤技术不能推荐那些具有较少评论的商品而且不能向具有较少评论的用户进行推荐, 而且无法向用户解释为什么推荐这些商品, 即不具有可解释性。

2.1 基于评论的推荐系统

用户评论、商品特征描述这类文本信息是推荐系统中常见的辅助性信息, 使用文本信息能够在一定程度上缓解推荐系统固有的局限性。在早期工作中, 基于评论的推荐系统主要采用主题模型来从评论中分别为用户和商品学习潜在的语义主题。HFT[1]和 CTR[3]使用隐狄利克雷分布(Latent Dirichlet Allocation, LDA)[11]来推测文本中的潜在主题, RBLT[2]认为具有高分评价的商品的评论中会包含更多正面的商品特性, 因此使用重复高分评论来构建基于分数增强的文本, 进一步从基于分数增强的评论文本中来提取商品的主题特征。CDL[12]使用堆叠去噪自动编码器(Stacked Denoising Autoencoders, SADE)学习文本中的潜在特征, 并输入到概率矩阵分解模型(Probabilistic Matrix Factorization, PMF)[13]来得到用户和商品的潜在矩阵。TLFM[14]提出了两个独立的因子学习模型, 来挖掘用户和商品共同的情感一致性和文本一致性, 然后将两个模型结合到一起来对评分进行预测。上述的方法虽然能够利用评论文本中的信息, 但这些方法都是基于词袋的模型, 忽略了词序信息和局部的语义信息, 丢失了句子中有价值的信息。

在最近几年, 深度学习模型被逐渐应用到基于评论的推荐系统中, 并且有效地提高了基于评论的推荐系统的效能。为了能够结合上下文信息从而达到更好的推荐效果, 一些工作使用卷积神经网络(Convolutional Neural Network, CNN)[15]或循环神经网络(Recurrent Neural Network, RNN)[16]将包含语义上下文的信息映射为具有连续值的向量表示。ConvMF[17]使用 CNN 学习商品评论中的局部语义信息, 来获得基于评论的潜在语义表示。DeepCoNN[18]分别根据用户和商品的评论使用 CNN 学习用户和商品的语义表示, 随后将两个语义表示进行拼接来进行分数预测。TransNet[19]相对于 DeepCoNN, 添加了一个额外的目标网络层来推测那些不能获得对应评论的实例的潜在语义, 进而提高评分预测的精度。SSG[20]将评论信息融合到图神经网络(GNN)中, 有利于图神经网络更有效地提取用户商品间的交互信息, 并采取了一种多模式的建模方式, 取得了很好的效果。

2.2 具有可解释性的推荐系统

虽然上述的应用了神经网络的推荐系统取得了不错的效果, 但深度推荐系统对输出的结果不能赋予很好的可解释性。有些工作通过注意力机制来学习评论序列中潜在特征的重要性, 期望以这种方式来获得评论中关键的信息, 在提高推荐系统的准确性的同时提高可解释性。D-Attn[4]结合全局和本地的注意力机制来识别用户和商品评论中重要的词, 来获得更精确的语义特征表示。NARRE[21]认为针对不同的目标用户商品对, 评论的重要性是不同的, 因此 NARRE 设计了一种注意力机制能够挑选出用户

商品评论文档中重要的评论。MPCN[22]采用了基于指针的共注意力模式来实现多层次的信息选择，保证能够挑选出重要的评论与这条评论中重要的词。CARL[23]采用 CNN 来获得评论中词的语境特征，并通过共注意力机制来获得每个词重要程度，最后以一个动态的线性的融合机制预测评分。ACF[24]采用了两个注意力机制，分别可以用来从多个商品以及用户购买过的有代表性的商品中选择有用的信息。TARMF[25]使用了一种基于注意力的门控循环单元(GRU)为 PMF 的表示获得语义解释。CARP[26]提出使用胶囊网络来进行评分预测并在更加细粒度的层面上提供可解释性，即分别衡量用户对商品的喜爱、排斥程度。

KPRN[27]利用知识图谱来建模具有可解释性的推荐系统，知识图谱的路径被用来推测用户-商品交互的潜在的原因。AMCF[28]提出了一种新型的特征映射方法，能够将不具有可解释性的特征映射到具有可解释性的属性特征上，进而使得不需要外部数据的传统模型具有可解释性，同时 AMCF 提出了一种基于属性的评估推荐系统可解释性的手段，但这种评估方式具有一定的局限性，不能应用于一般的推荐系统上。

3 问题定义

大部分基于评论的深度推荐系统通过将一个用户所有的评论拼接得到用户评论文档，与用户相类似，商品评论文档是将一个商品所有评论进行拼接。用户、商品评论文档分别表示为 $D_u=(t_1,t_2,\cdots,t_m)$，$D_i=(t_1,t_2,\cdots,t_n)$，其中 t_j 表示对应评论文档中的第 j 个词，m 和 n 分别代表用户和商品评论文档中词的个数，由基于评论的推荐系统预测出的分数记为 $\hat{r}_{u,i}=F(D_u,D_i)$，表示用户 u 对商品 i 的偏好程度，这也是推荐系统的预测目标。

3.1 基于评论的深度推荐系统

很多神经模型使用注意力机制来学习出评论文本中重要的词、属性或者评论等语义单元，从而能够更好地学习潜在语义特征，基于文本注意力也能够使得推荐系统更具有可解释性。本文根据在注意力权重计算机制的不同，将这些具有可解释性的基于评论的推荐系统分为三类：基于注意力的推荐系统，基于交互的推荐系统，基于属性的推荐系统。

3.1.1 基于注意力的推荐系统

基于注意力的推荐系统分别通过对应的评论文档得到用户和商品的语义表示。在特征提取过程中，基于注意力的推荐系统使用注意力机制来为对应文档中的每个词都分配一个注意力权重，如果某个词能够突出地表现用户偏好或者商品特征，那么这个词将会被赋予一个比较大的注意力权重值，如图 2 所示。例如，D-Attn[4]使用一个局部的注意力机制和一个全局的注意力机制来识别重要的词；在 NARRE 中，使用注意力机制来计算评论级别的权重值，本文接下来会详细介绍 NARRE 是如何获得这个注意力权重值。

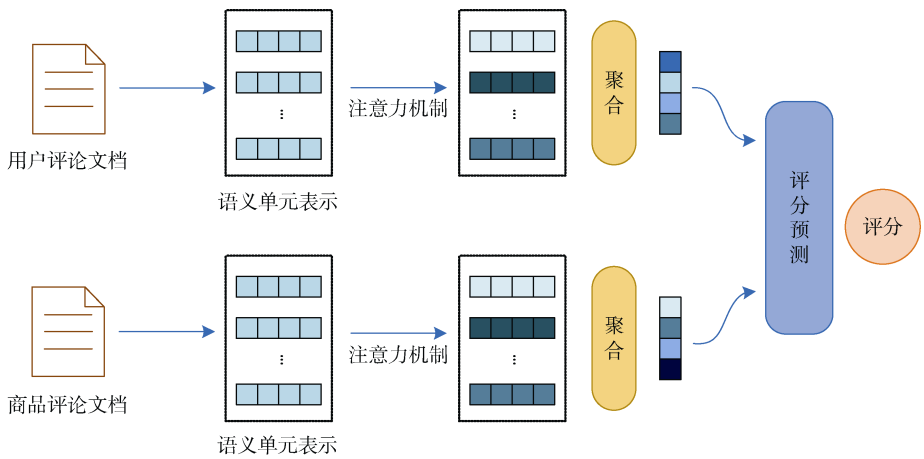


图 2 基于注意力的推荐系统

Figure 2 Attention-based recommendation system

NARRE 首先通过 CNN 得到用户和商品的评论级别包含语境信息的特征向量：$O^u=(O_1,O_2,\cdots,O_M)$ 和 $O^i=(O_1,O_2,\cdots,O_N)$，其中 M 和 N 分别代表用户和商品对应的评论条数。随后 NARRE 通过一个两层的神经网络来为每条评论计算注意力权重值，从用户评论文档中选择出能够真正反映用户偏好的评论，最终聚集成为一个特征向量作为这个用户的表示。NARRE 将上一步得到的第 l 条评论的表示和

对应的商品的 ID embedding i_l 输入到注意力神经网络中:

$$a_l^* = h^T \mathrm{ReLU}(W_1^{na} O_l + W_1^{na} i_l + b_1^{na}) + b_2^{na} \quad (1)$$

这里 W_1^{na}, W_2^{na}, b_1^{na}, b_2^{na} 都是可以学习的参数, ReLU 是非线性激活函数。

根据上述得到的注意力分值, 可以通过一个 softmax 函数进行正则从而得到用户的第 l 条评论对应的注意力权重值:

$$w_l = \frac{\exp(a_l^*)}{\sum_{j=1}^{M} \exp(a_j^*)} \quad (2)$$

用户 u 的特征向量是由上述得到的权重值与评论的特征向量加权而得: $O_w^u = \sum_{j=1}^{M} w_j O_j$, 随后再经过一个全连接层来得到用户最终的特征向量。NARRE 通过相同的计算过程得到商品的特征表示, 并与用户的特征表示一起用于后续的评分预测中。

3.1.2 基于属性的推荐系统

基于属性的推荐系统首先在用户和商品评论文档中提取多个属性, 随后通过计算用户端和商品端提取出来的属性的匹配程度来计算评分。每一个属性实际是一种高层次的语义特征, 能够覆盖用户关心的或者商品含有的某个特定的属性或主题。在这些方法中, 注意力机制被设计应用在为每个属性识别具有代表性的词, 基于属性的推荐系统如图 3 所示。例如, TARMF 使用一个基于注意力机制的 RNN 来分别在用户评论文档和商品评论文档提取基于属性的特征表示; ANR 通过利用注意力机制来为每个基于词语境窗口的属性计算每个单词的重要性; CARP 使用自注意力机制来从各自对应的评论文档中提取用户的观点和商品的属性。下文将详细介绍 ANR 和 CARP 如何获得属性级别的注意力权重值及属性级别的表示。

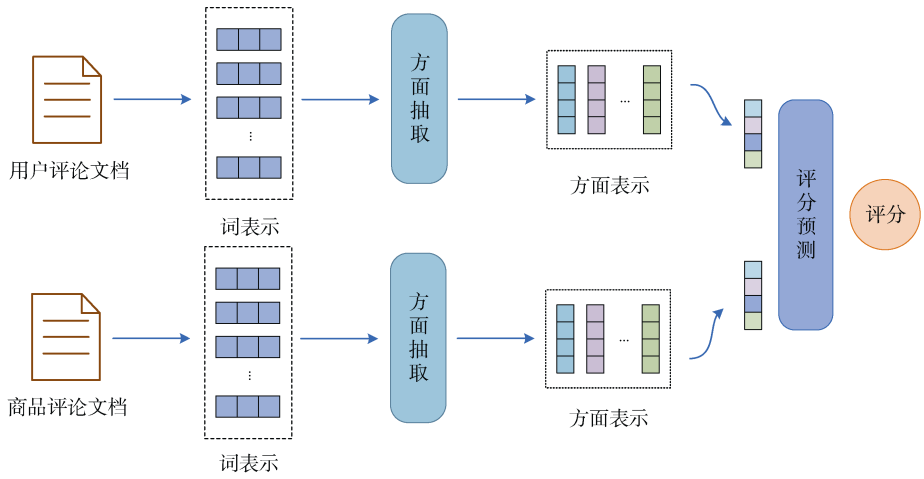


图 3 基于属性的推荐系统

Figure 3 Aspect-based recommendation system

ANR 首先得到用户评论文档的词向量表示 $\boldsymbol{H}^u = (h_1, h_2, \cdots, h_m)$, 其中 m 代表用户评论文档中的词的个数, ANR 使用一个特定于属性的单词映射矩阵 $\boldsymbol{W}_a$ 来区分不同词的对应的属性:

$$\boldsymbol{H}_j^{u,a} = H_j^u \boldsymbol{W}_a \quad (3)$$

H_j^u 是评论文档中第 j 个词原本的词向量表示, $H_j^{u,a}$ 是在给定的用户 u 和属性 a 下的特定于属性的词表示, 由于有 K 个不同的属性, 因此存在 K 个这样的映射操作。

为了更好地捕捉语义信息, ANR 使用局部的语境窗口来计算文档中每个词的重要性, 设每个属性都可以用一个向量 $\boldsymbol{v}_a \in \mathbb{R}^{c*d}$ 来表示, 其中 c 是一个超参数, 代表了基于语境的窗口大小, 那么评论文档中第 j 个词的语义窗口表示为:

$$z_j^{u,a} = (H_{j-c/2}^{u,a}; \ldots; H_j^{u,a}; \ldots; H_{j+c/2}^{u,a}) \quad (4)$$

其中 $(;,\cdot)$ 代表连接操作, 随后通过一个 softmax 函数得到这个词在对应属性下的注意力权重值, 再进行加权求和得到这个词的特征向量:

$$w_j^{u,a} = \frac{v_a (z_j^{u,a})^T}{\sum_{h=1}^{m} v_a (z_h^{u,a})^T} \quad (5)$$

$$O^{u,a} = \sum_{h=1}^{m} w_j^{u,a} H_j^{u,a} \quad (6)$$

ANR 在商品侧以相同的计算方式得到属性级别的特征表示 $O^{i,a}$。

为了聚集属性级别的特征表示, ANR 使用一种协同训练的方式来学习属性的重要性。根据上述得到属性级别的用户表示 O^u 和商品表示 O^i 来计算一个相似度矩阵:

$$\boldsymbol{S} = \mathrm{ReLU}(O^u W_s O^i) \tag{7}$$

其中, W_s 是可学习的参数, 根据相似度矩阵 $\boldsymbol{S}$ 得到用户和商品的属性的注意力权重值 β_u 和 β_i:

$$\beta_u = \mathrm{softmax}(\mathrm{ReLU}(O^u W_1^{anr} + S^T(O^i W_2^{anr}))v_1) \tag{8}$$

$$\beta_i = \mathrm{softmax}(\mathrm{ReLU}(O^i W_2^{anr} + S^T(O^u W_1^{anr}))v_2) \tag{9}$$

其中, W_1^{anr}、W_2^{anr}、v_1、v_2 都是可学习的参数, 值得一提的是, 这样学习出的属性重要性在不同的用户、商品之间都是互不相同的。ANR 通过计算得到的属性重要性 β_u、β_i 与上述得到的用户商品属性级别的特征向量 $O^{u,a}$、$O^{i,a}$ 聚合, 用于最终的评分预测中。

与 ANR 获取属性级别的注意力权重方法不同, CARP 首先分别将用户、商品评论文档经过一个卷积操作, 以 ReLU 作为激活函数来得到融合了语境的特征向量: $O^u = [c_1^u; \cdots; c_m^u]$、$O^i = [c_1^i; \cdots; c_n^i]$, 其中 m 和 n 分别代表用户和商品评论文档长度。

对于用户评论来说, CARP 为了识别出用户评论文档中哪些词与用户观点是紧密相关的, 添加了一个识别用户有用观点的门控机制:

$$p_{u,x,j} = W_p(c_j \odot \sigma(W_{x,1} c_j + W_{x,2} q_{u,x} + b_x)) \tag{10}$$

其中, $q_{u,x}$ 是第 x 个观点对应的特征向量, 这个特征表示是所有用户共享的, 通过模型进行学习更新, $\odot$ 是元素积操作。

CARP 提出使用自注意力机制来从用户的评论文档中提取用户观点, 首先在用户评论文档中将上述得到的每个词特征向量取均值得到 $\overline{p}_{u,x} = \frac{1}{m}\sum_j p_{u,x,j}$, 那么在给定的用户 u 和观点 x 下, 从评论文档中提取出观点的特征向量表示为:

$$v_{u,x} = \sum_j \alpha_{u,x,j} p_{u,x,j} \tag{11}$$

其中 $\alpha_{u,x,j} = \mathrm{softmax}(p_{u,x,j}^T \overline{p}_{u,x})$, 是用户评论文档值相对于观点 x 的注意力权重值, 确保在用户评论中提取观点时, 能够捕捉到对于不同的观点始终重要的特征。

CARP 以相同的方式从商品评论文档中提取出 k 个属性的特征表示。CARP 定义用户的第 x 个观点与商品的第 y 个属性组成一个逻辑单元, 对应的特征表示 $g_{x,y}$ 为:

$$g_{x,y} = [(v_{u,x} - a_{i,y}) \oplus (v_{u,x} \odot a_{i,y})] \tag{12}$$

那么对于用户 k 个观点、商品 k 个属性能够组成 $k*k$ 个不同的逻辑单元。CARP 通过胶囊网络结构来识别哪些逻辑单元是重要的, 以及基于逻辑单元来得到用户对商品的喜好程度。CARP 分别使用正向胶囊网络和负向胶囊网络来提取逻辑单元中用户对商品的正向情感和负面情感:

$$S_{s,u,i} = \sum_{x,y} \beta_{s,x,y} t_{s,x,y} \tag{13}$$

其中, $t_{s,x,y} = W_{s,x,y} g_{x,y}$, $s \in \{pos, neg\}$, 即正面情感和负面情感, $\beta_{s,x,y}$ 是耦合系数, 表示逻辑单元 $g_{x,y}$ 在决定情感 s 时的贡献度, CARP 设计了一种双向协议路由(Routing by Bi-Agreement)算法来对 $\beta_{s,x,y}$ 进行更新计算。

随后通过一个非线性的挤压函数将 $S_{s,u,i}$ 压缩到 (0,1) 范围内:

$$o_{s,u,i} = \frac{\| s_{s,u,i} \|^2}{1 + \| s_{s,u,i} \|^2} \frac{s_{s,u,i}}{\| s_{s,u,i} \|} \tag{14}$$

$o_{s,u,i}$ 最终会被输入到分数预测层, 用于预测用户 u 对商品 i 的评分。

3.1.3 基于交互的推荐系统

基于注意力的推荐系统以独立和静态的方式通过识别与用户偏好和用户特征相关的重要信息, 也就是说用户评论文档中词的重要性并没有考虑到商品评论文档中的相关信息, 在商品评论文档中也存在同样的问题。但在实际应用中, 用户-商品对的关联信息更能充分表现用户对商品的偏好信息。因此, 基于交互的推荐系统通常通过共注意力机制加权的方式来充分捕捉用户商品的交互信息, 从而使得在用户和商品评论文档中关联语义信息被模型捕捉到, 并赋予比较高注意力权重, 如图 4 所示。例如, 在 MPCN 中, 通过添加评论级别和词级别的共注意力机制来提取在用户商品对中关联度最高的信息; CARL 使用带平均池化的共注意力机制来获取词的重要性程度; 与 CARL 相似, DAML[29]在共注意力机制中利用了基于语境的词语义表示计算得的欧几里得距离; 值得注意的是, 在基于属性的推荐系统中提起到的 CARP, 同样也可以被视为一种基于交互的推荐系统, CARP 中的逻辑单元的表示是通过用户观点和商品属性的特征表示的交互得来。接下来本文将详细介绍 MPCN 和 CARL 是如何通过特征交互获得词级别的注意力权重。

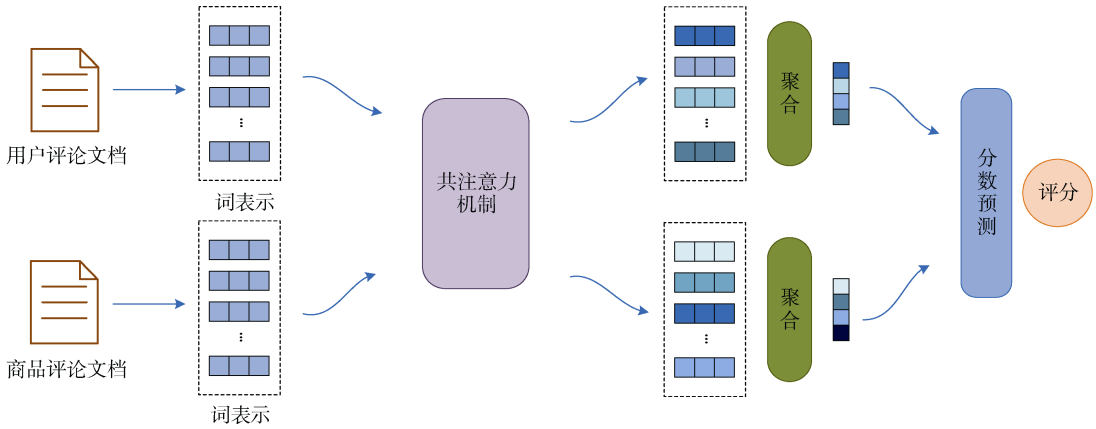


图 4　基于交互的推荐系统

Figure 4　Interaction-based recommendation system

MPCN 作为多层次结构, 输入是评论的序列, 每条评论又是词构成的序列。在经过词向量层后, MPCN 将组成评论的所有词向量求和, 得到每条评论的向量表示 $\boldsymbol{h}$, 随后通过一个门机制来决定评论中多少信息传递到下一个阶段, 得到 $\boldsymbol{h}$。

与上述介绍的 ANR 类似, MPCN 通过用户评论和商品评论之间的相似度矩阵来计算评论级别的共注意力:

$$S = F(\bar{\boldsymbol{h}}_{\boldsymbol{u}})^T W_s F(\bar{\boldsymbol{h}}_{\boldsymbol{i}}) \tag{15}$$

其中, $F(\cdot)$ 是1层的前馈神经网络, $\bar{\boldsymbol{h}}_{\boldsymbol{u}}$ 和 $\bar{\boldsymbol{h}}_{\boldsymbol{i}}$ 分别是用户 u 和商品 i 对应的所有评论的特征向量。为了能够从用户评论和商品评论中获取到对预测目标评分真正有意义的评论, MPCN 通过公式(16)、(17)来得到评论指针:

$$p_u = G(\max_{col}(S)) \tag{16}$$

$$p_i = G(\max_{row}(S)) \tag{17}$$

$G(\cdot)$ 是经过 Gumbel-softmax 函数后的 argmax 操作, Gumbel-softmax 的计算方法为:

$$Gumbel-softmax(x_i) = \frac{\exp(\frac{\log(x_i)+g_i}{\tau})}{\sum_{j=1}^{k}\exp(\frac{\log(x_j)+g_i}{\tau})} \tag{18}$$

其中 $g_i = -\log(-\log(u_i))$, $u_i \sim uniform(0,1)$, τ 是温度系数。通过上述操作, MPCN 可以挑选出用户 u 的第 p_u 条评论的特征向量 $\bar{\boldsymbol{h}}'_{\boldsymbol{u}}$, 商品 i 的第 p_i 条评论的特征向量 $\bar{\boldsymbol{h}}'_{\boldsymbol{i}}$。

MPCN 作为多层次的推荐系统, 不仅能够挑选出重要的评论, 也可以挑选出评论中重要的词。与评论级别的共注意力机制类似, 首先计算词级别的相似度矩阵:

$$S_w = F(\bar{\boldsymbol{h}}'_{\boldsymbol{u}})^T W_w F(\bar{\boldsymbol{h}}'_{\boldsymbol{i}}) \tag{19}$$

与评论级略有不同的是, 词级别的共注意力使用了平均池化层来达到更稳定的效果:

$$\bar{O}^u = (softmax(\underset{col}{mean}(S_w)))^T \bar{\boldsymbol{h}}'_{\boldsymbol{u}} \tag{20}$$

$$\bar{O}^i = (softmax(\underset{row}{mean}(S_w)))^T \bar{\boldsymbol{h}}'_{\boldsymbol{i}} \tag{21}$$

MPCN 使用了多指针组合机制, 即以相同的方式实施了 k 次评论级别的共注意力机制, 分别为用户 u 和商品 i 挑选出 k 条评论, 随后通过词级别的共注意力机制得到每条评论的加权特征向量, 那么经过多次指针操作后可以得到:

$$O^u = \{\bar{O}_1^u, \bar{O}_2^u, \cdots, \bar{O}_k^u\} \tag{22}$$

$$O^i = \{\bar{O}_1^i, \bar{O}_2^i, \cdots, \bar{O}_k^i\} \tag{23}$$

最后 MPCN 使用连接、相加或者全连接神经网络三种方式来聚集这 k 次操作后的特征, 进而得到用户 u、商品 i 最终的特征向量, 用于后续的分数预测中。

与 CARP 相类似, CARL 首先分别将评论文档经过一个卷积操作来得到融合了语境的特征向量: $\boldsymbol{O}^u = [c_1^u; \cdots; c_m^u]$、$\boldsymbol{O}^i = [c_1^i; \cdots; c_n^i]$, 其中 m 和 n 分别代表用户评论文档和商品评论文档的长度。CARL 使用了一个注意力矩阵 $\boldsymbol{T}$ 来从 $\boldsymbol{O}^u$ 和 $\boldsymbol{O}^i$ 中得到每个特征向量的重要性, 计算用户评论文档和商品评论文档每对词与词之间的关联关系。

$$R_{j,k} = \tanh((\boldsymbol{c}_j^u)^T T \boldsymbol{c}_k^i) \tag{24}$$

$R_{j,k}$ 能够反映 c_j^u 与 c_k^i 之间的关联程度, 其中 $\boldsymbol{c}_j^u$ 是用户文档表示中第 j 个词的特征向量, $\boldsymbol{c}_k^i$ 是商品文档表示中第 k 个词的特征向量。随后 CARL 采用平均池化操作来聚集 $\boldsymbol{R}$ 矩阵每行每列的特征向量, 再

通过 softmax 函数分别得到用户、商品评论文档中每个特征向量的注意力权重值, 从而能够辨别出哪些词相对于评论文档是更加重要的:

$$a_j^u = softmax(\underset{col}{mean}(R)) \tag{25}$$

$$a_k^i = softmax(\underset{row}{mean}(R)) \tag{26}$$

因此CARL可以通过注意力矩阵 $\boldsymbol{T}$, 获得基于用户-商品对交互的词的注意力权重值:

$$w^u = [w_1^u, \cdots, w_m^u] \tag{27}$$

$$w^i = [w_1^i, \cdots, w_n^i] \tag{28}$$

由于上述的注意力权重值是基于用户评论文档和商品评论文档一起算出的, 因此如果一个词具有更高的权重值, 那么意味着这个词对应的特征向量与待预测的用户-商品对的相关性越高, 基于注意力权重得到用户、商品评论文档的加权特征表示:

$$O_w^u = diag(w^u)O^u \tag{29}$$

$$O_w^i = diag(w^i)O^i \tag{30}$$

这里的 $diag(w^*)$ 是指对角线元素为 w^* 的对角矩阵。

随后, CARL 使用一个带平均池化层的 CNN 网络分别进一步提取用户和商品评论文档高层次特征, 并再各自通过一个全连接网络得到最终的评论文档的特征表示, 用于后续的评分预测。

3.2 问题设定

本文选取五个当前最先进的推荐系统来进行可解释性的量化评估, 这五个基于评论的深度推荐系统分别是 ANR、CARL、CARP、MPCN 和 NARRE, 5 个推荐系统覆盖了上述介绍的三类基于文本注意力的推荐系统, 但在对评论建模方式互不相同。值得注意的是这五个模型分别为了得到更好推荐结果, 从评论中在不用的粒度级别标注出与用户偏好或者商品特征最相关的语义信息: ANR 将每个词都分配到各个不同的属性; CARP 进一步将情感添加到用户情感-商品属性对中; MPCN 和 CARL 基于语义相关联的词来推测用户的偏好; 而 NARRE 能够识别与用户偏好和商品特征最相关的评论。本文将识别重要的评论或者属性视为一种计算词级别注意力权重的特殊形式, 因此通过各个推荐系统中的所使用的注意力机制, 可以很容易地得到在评论文档中每个词的重要程度, 即注意力权重值。值得一提的是, 本文作为一个探究基于评论的推荐系统可解释性的初步探索与尝试, 从词级别来对推荐系统的可解释性进行评估。

定义 1.基于评论的推荐系统的可解释性评价。给定用户评论文档 $D_u = (t_1, t_2, \cdots, t_m)$ 和商品评论文档 $D_i = (t_1, t_2, \cdots, t_n)$, 通过对应的推荐系统内置的注意力机制获得的包含词级别权重的用户评论权重文档 $D_u^w = (<t_1, w_1>, <t_2, w_2>, \cdots, <t_n, w_m>)$ 和商品评论权重文档 $D_i^w = (<t_1, w_1>, <t_2, w_2>, \cdots, <t_n, w_n>)$, 其中 $w_j \in (0,1)$, 代表了文档中的第 j 个词对推荐系统给出评分预测贡献程度, $\sum_{j=1}^{n} w_j = 1$, $\sum_{k=1}^{m} w_k = 1$。本文提出的对基于评论的推荐系统的可解释性评估任务, 是通过衡量被扩展的带权重的评论文档 D_u^w 和 D_i^w 中的高权重词, 能否准确地表达出用户 u 对商品 i 的评论(即目标评论)中所表现出来的用户偏好、商品特征来得到一个量化分数。

根据 3.1 节介绍, CARL 和 MPCN 能够直接获得词级别的注意力权重, 本文首先来介绍如何得到 NARRE、ANR、CARP 三个推荐系统的词级别的注意力权重。由于 NARRE 能够得到评论文档中每条评论的注意力权重, 因此本文视为同一条评论中的每个词具有相同权重。对于 ANR, 令 β_k 和 $\alpha_{k,j}$ 分别表示属性 k 的重要性, 在用户评论文档中词 t_j 相对于属性 k 的注意力权重, 那么可以得到 ANR 对应用户评论文档的词的注意力权重为 $w_j = \sum_k \beta_k \cdot \alpha_{k,j}$, 对于商品评论文档也是同样的操作。在 CARP 中, 令 $\beta_{s,v,k}$ 代表由情感 s 对应的胶囊网络得到的用户观点 v 和商品属性 k 所组成的逻辑单元的重要程度, 那么用户评论文档中词的重要性为 $w_j = \sum_s \sum_v \beta_{s,v,*} \cdot \alpha_{v,j}$, 其中 $\alpha_{v,j}$ 是指词 t_j 相对于用户观点 v 的注意力权重值, $\beta_{s,v,*} = \sum_k \beta_{s,v,k}$; 商品评论文档中词的注意力权重值也采取相同的计算方式。

4 数据标注

根据上述的问题定义, 本文在三个数据集上进行人工标注工作。三个数据集分别是来自 Amazon-5cores①的 Office Products 和 Musical Instrument[30]和 Yelp Challenge②的 Yelp16-17, 对于 Yelp 数据集, 与之前的工作[26, 31]相类似, 本文选取了 2016 年到 2017 年时间跨度内的数据来组成 Yelp16-17, 并且为了保持数据的统一, 与 Amazon 数据集相类似, 本文也对

① http://jmcauley.ucsd.edu/data/amazon/
② https://www.yelp.com/dataset/challenge

Yelp16-17 进行 5-core 处理, 即每个用户和每个商品都至少有五条评论。3 个数据集都包含用户对商品的一个[1, 5]的评分以及文本评论。本文将一个用户所有的评论以一个特殊符号“[SEP]”作为间隔拼接起来以形成用户评论文档, 并以相同的方式形成商品评论文档。随后本文通过截断长评论来将每个评论文档的长度限制在 300 个词以内。

用户在购买商品后的评论通常比较简洁[31], 但较短的评论不能包含充分的信息, 从而不能很详细地了解用户对评论的偏好信息, 并不利于可解释性的评估工作。因此对于每个数据集, 本文随机选取 200 个用户-商品对, 且保证每个用户-商品对所对应的目标评论至少包含 20 个词。关于 3 个数据集的数据统计如表一所示。本文总共探究了在 3 个数据集上 5 个基于评论的推荐系统, 因此有 3000 条实例需要人工进行标注。本文招募了 4 位受过良好教育的标注人员, 他们精通英语并且对这 3 个数据集的商品信息较为熟悉。标注人员均分为两组, 即每一组都需要标注 1500 条实例数据, 每条实例都由两位标注人员同时进行标注。值得一提的是, 在分发给标注人员数据时, 标注人员并不能得知标注的数据是由哪个基于评论的推荐系统形成的。

表 1 在 3 个数据集上随机选取的数据统计

Table 1 Statistics of randomly selected data on three datasets

数据集	用户数	商品数	评论数	平均评论长度
Musical Instruments	80	162	200	86.56
Office Products	53	177	200	155.6
Yelp 16-17	34	197	200	121.48

表 2 推荐系统在随机选取的数据上的表现

Table 2 The performance of the recommendation system on randomly selected datasets

数据集	Musical Instruments		Office Products		Yelp 16-17	
推荐系统	MSE	MAE	MSE	MAE	MSE	MAE
ANR	0.56	0.62	0.54	0.57	1.05	0.77
CARL	0.53	0.60	0.52	0.54	1.02	0.76
CARP	0.54	0.57	0.49	0.52	1.02	0.77
MPCN	0.59	0.62	0.64	0.64	1.15	0.82
NARRE	0.70	0.62	0.53	0.57	1.04	0.77

标注人员为每条实例标注一个[1, 5]范围内的可解释性评分, 其中 1 分代表推荐系统学习到的高权重词与目标评论中用户的偏好或者商品的特征信息完全无关; 2 分代表推荐系统得到的高权重的词大部分都是无意义的词, 只有少量的词能够反映目标评论; 3 分代表推荐系统赋予较高权重的词较少的是完全没有意义的词, 有明确反映用户偏好或商品特征的词; 4 分代表推荐系统赋予高权重的词有多个能够描述用户商品对对应的用户偏好或者商品特征的词; 5 分代表被推荐系统突出强调的词大部分都是描述用户偏好或商品特征并且情感与对应的实例相一致。

需要特别说明的是, 标注人员在标注数据时会严格考虑推荐系统得到的高权重词, 也就是说, 如果一个词与目标评论高度相关, 能够充分地反应用户的偏好或者商品的特征, 但这个词并没有被推荐系统赋予比较高的权重, 那么这样的实例也会被标注一个低的分数。在标注过程中, 标注人员可以通过一个可视化的工具来得到推荐系统习得的高权重词, 如图 5 所示。本文会基于对应的推荐系统内置的注意力机制获得的权重文档, 计算一个权重中值 w_{med}, 将注意力权重值小于 w_{med} 的单词的前景颜色置为黑色, 背景颜色置为白色; 将注意力权重值大于 w_{med} 的词进行加粗且背景色置为浅灰色, 并且使用渐变的前景色来表现词的重要程度, 即如果词的注意力权重值越高那么前景色的着色越深。由于用户标注过程是一个主观倾向性很大的过程, 因此在正式标注前添加一个预标注阶段: (1)首先每一个组都随机标注相同的 20 条实例; (2)两位标注人员互相检查标注结果, 在标注结果的基础上进行讨论最终达到一个统一的标注标准和一致的标注结果; (3)同一组内的两位标注人员再独立标注全部的数据。

在人工标注工作结束后, 本文计算了同一组内的两位标注人员标注结果在各个数据集、各个推荐系统内的皮尔逊相关系数, 结果如表 3 所示, 其中 G1、G2 分别代表两个标注组。皮尔逊相关系数能够反映两个变量之间的数值线性相关关系, 对于总体的两个随机变量 X, Y 之间的皮尔逊相关系数可以被定义为:

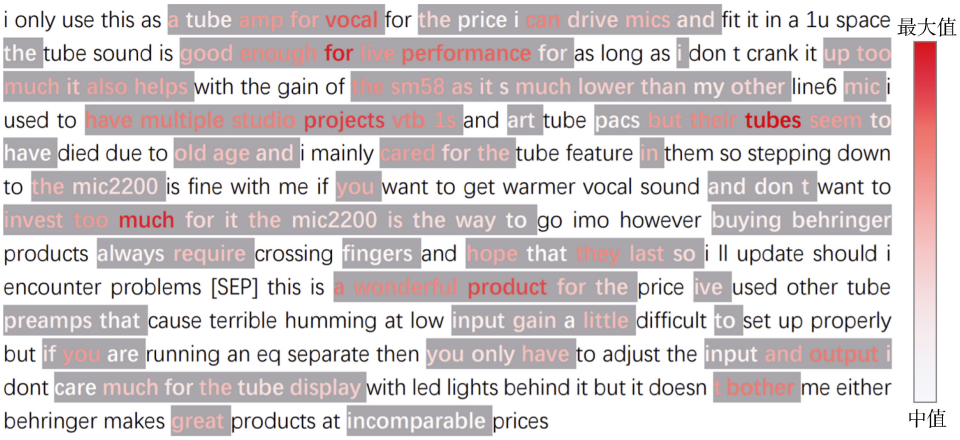


图 5　带权重的商品评论文本可视化示例

Figure 5　Visualization with an item review snippet

$$\rho_{X,Y}=\frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y} \tag{31}$$

整体标注的皮尔逊相关系数均值为 0.70, 即两组标注人员在标注结果上都达到了比较高的一致性。对于每一条实例, 本文将为这条数据标注分数的两位标注人员标注的结果的均值作为这条实例的最终的可解释性评分, 并用于后续的进一步分析工作中。

表 3　同一组内两位标注人员标注结果的皮尔逊相关系数(G1、G2 分别代表两个标注组)

Table 3　Pearson's correlation coefficient of the annotation results of two annotators in the same group (G1 and G2 respectively represent the two annotation groups)

数据集	Musical Instruments		Office Products		Yelp 16-17	
推荐系统	G1	G2	G1	G2	G1	G2
ANR	0.75	0.79	0.68	0.73	0.63	0.70
CARL	0.69	0.80	0.55	0.66	0.72	0.75
CARP	0.62	0.63	0.54	0.69	0.65	0.71
MPCN	0.64	0.73	0.61	0.67	0.57	0.73
NARRE	0.70	0.67	0.63	0.73	0.67	0.71

5　可解释性分析

本文首先根据标注得到的可解释性分数, 来综合地定量评估选定的 5 个基于评论的深度推荐系统的可解释性, 并使用方差分析说明了推荐系统之间的可解释性分数差异是具有统计意义的。随后探究了可能影响推荐系统可解释性的因素, 以及可解释性高的推荐系统具有的一些性质。

5.1　可解释性表现

本文首先计算了在每个数据集、每个基于评论的推荐系统下的可解释性评分均值, 如表 4 所示。可以注意到对于每个模型, 整体的可解释性评分均值都在 3.0 分左右, 相较于本文制定的评分标准, 并不是非常高的分数。其中, CARL 在总体上获得了最高的可解释性评分, 在 3 个数据集上的均值为 3.21, 紧随其后的是 ANR, 可解释性评分为 3。结合获得的可解释性分数以及通过 CARL 获得的带有词注意力权重的评论文档, CARL 往往会为对较多的词赋予较高的权重值, 从另一个角度来说, 这样加大了 CARL 能够发现关键词的概率; 从表中可以看出 CARP 的可解释性最差, 在 3 个数据集上的可解释性评分均值仅为 2.75, 这可能是由于本文计算 CARP 的注意力权重时是将正向情感权重与负向情感权重加权而得, 这样得到的注意力权重本身在一定程度上削弱了 CARP 的可解释性。

为了排除造成上述推荐系统可解释性分数差异的结果是随机因素导致的, 本文将不同的推荐模型看作是一个研究因素, 分别基于 3 个数据集上各 200 条数据的可解释性分数结果进行方差分析。但根据预计算, 在相同数据集下方差齐性的显著性值都小

表 4　推荐系统的可解释性表现(最优、次优的表现分别用粗体字、下划线标出)

Table 4　The interpretability performance for recommendation systems (The best is in boldface and second best underlined)

推荐系统	Musical Instruments	Office Products	Yelp 16-17	均值
ANR	3.03	3.00	<u>2.98</u>	<u>3.00</u>
CARL	<u>3.05</u>	**3.39**	**3.20**	**3.21**
CARP	2.67	2.76	2.84	2.75
MPCN	2.68	2.84	2.80	2.77
NARRE	**3.06**	<u>3.01</u>	2.66	2.91
均值	2.90	3.00	2.89	2.93

于 0.05, 说明不同推荐系统间的可解释性分数的方差并不相齐, 因此本文采取一种多独立样本非参数检验方法: 克鲁斯卡尔-沃利斯检验。克鲁斯卡尔-沃利斯检验要求被检测的样本是独立或者不相关的, 由上文所述, 相同数据集下对不同的推荐系统可解释性评价结果满足推荐系统间互不相关的条件, 因此可以应用克鲁斯卡尔-沃利斯检验。本文计算了基于三个数据集上的克鲁斯卡尔-沃利斯检验的显著性值, 结果保留三位小数, 如表 5 所示。

表 5　不同数据集下克鲁斯卡尔-沃利斯检验显著性

Table 5　Kruskal-Wallis test significance for different datasets

数据集	Musical Instruments	Office Products	Yelp 16-17
显著性	0.000	0.000	0.000

克鲁斯卡尔-沃利斯检验的原假设是各样本服从的概率分布具有相同的中位数, 原假设被拒绝意味着至少一个样本的概率分布的中位数不同于其他样本。从表 5 中能够看出, 在本文的检测中, 各个不同数据集的显著性值都小于 0.05, 因此在 3 个数据集上都拒绝原假设, 说明推荐系统之间的可解释性分数差异是具有统计意义的, 而并非随机因素造成的, 这也为本文后续进一步的数据分析提供了支撑。

5.2　可解释性评分分布

本文探索了每个模型在每个数据集上的可解释性评分分布, 如图 6(a)到(c)所示。可以得出以下结论, 所有的推荐系统大部分的实例的可解释性评分集中到[2, 4]的分数范围内, 这与表 4 得到的结论是相一致的; 在 Yelp 数据集上, 5 个推荐系统能够有 50%以上的可能学习到目标评论中的相关含义, 即大部分的可解释性评分分布在[3, 4]范围内, 但在 Musical Instruments 和 Office Products 2 个数据集上, 与其他 4 个推荐系统不同, CARP 的评分分布集中在[2, 3]上, 说明 CARP 相对于其他推荐系统具有更差的可解释性; 从另一个方面来说, 基于所有的推荐系统和数据集, 平均有 6.6%的实例不能被学习到有用的信息(评分分布在[1, 2]范围内), 即不具有可解释性, 与之相对应的, 评分分布在[4, 5]范围内所占的实例比率

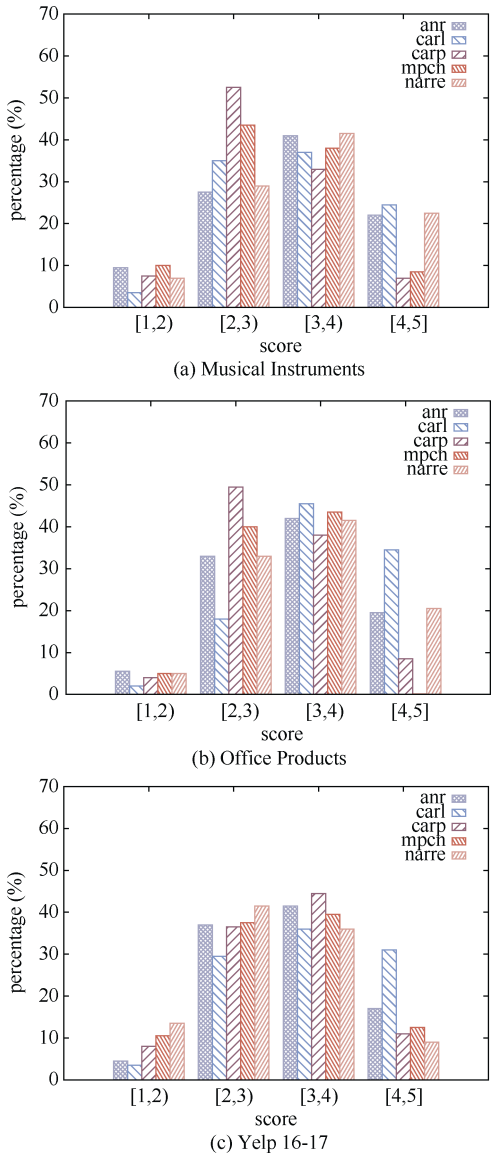


图 6　可解释性评分分布

Figure 6　Interpretability score distributions for each review-based recommendation system

均值为 17.3%。作为可解释性最好的模型, CARL 评分分布在[4, 5]范围内在 3 个数据集上平均比率为 30.0%, ANR 以 19.5%的比率紧随其后; 直观来看, 在大部分的数据集和推荐系统上, 有超过 50%的实例的可解释性评分分布在[3, 5]分数范围内, 也就是说, 当前基于评论的推荐系统有超过一半的可能性能捕捉到目标评论中用户观点或者商品特征。

5.3 可解释性评分因素分析

由于长评论中可能会包含更多细粒度的信息, 会涉及到用户关注的多个属性方面或者一个商品的多个特征, 而长评论中的一些内容也未必能够真正地反映评分。因此本文探究了目标评论长度与可解释性评分之间的关联关系, 如图 7(a)~(c)所示。对于每一条实例, 首先取基于 5 个推荐系统的最优可解释性评分作为这个实例的最终分数, 可以看出超过半数以上的实例可以达到 4.0 以上的评分。并且从图 7 中可以看出, 推荐系统很难全面地捕捉到比较长(目标评论长度在 200 以上)的实例中用户的观点或者商品的特征。

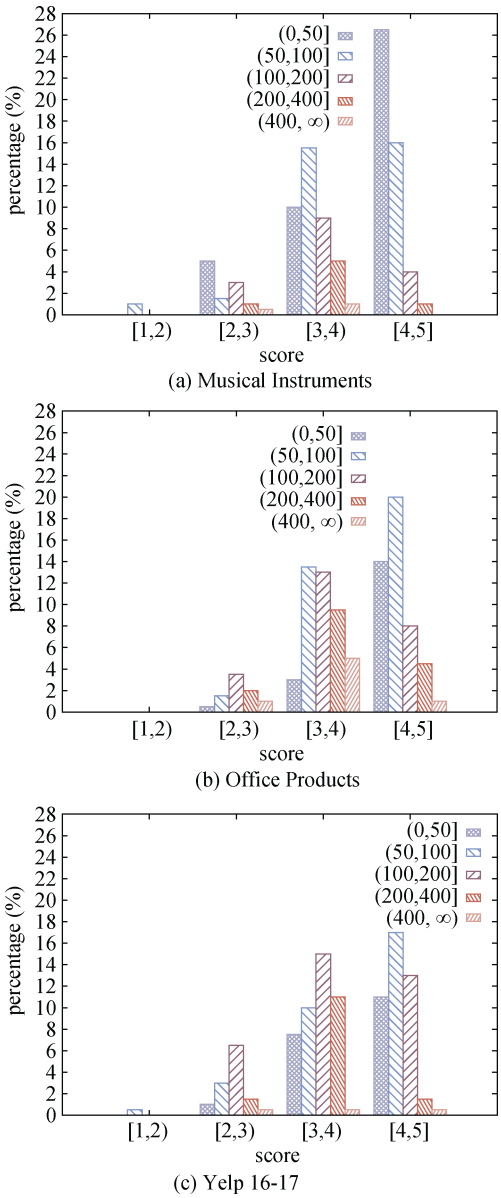
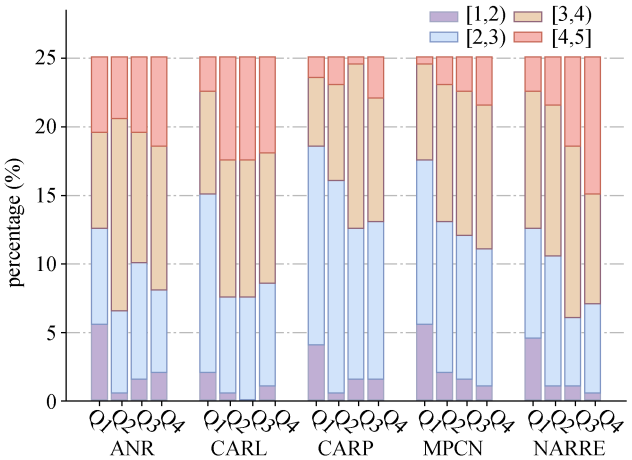


图 7 目标评论长度段的最优可解释性评分分布

Figure 7 Interpretability score distributions for the best recommendation system over review length

值得一提的是, 本文同时计算了目标评论长度与在 5 个推荐系统可解释性评分的最大值之间的皮尔逊相关系数, 并且得到在 3 个数据集上的皮尔逊相关系数的均值是–0.295, 说明目标评论长度与可解释性分数确实存在着一定的负相关关系, 即目标评论越短, 推荐系统越可能能够学习到与目标评论中反映的用户观点和商品特征相一致的信息。

大部分的推荐系统在提高推荐系统的可解释性的同时更重要的目的是能够提高推荐系统的性能, 即提高评分预测的准确率。令 $\hat{r}^s_{u,i}$ 代表推荐系统 s 对于用户-商品实例对 (u,i) 的预测分数, 本文通过计算绝对误差 $\delta^s_{u,i}=|r^s_{u,i}-\hat{r}^s_{u,i}|$ 来表示推荐系统 s 能否从评论中捕捉到用户 u 对商品 i 的偏好, $\delta^s_{u,i}$ 越大, 代表推荐系统能够捕捉到信息越少。随后对于每个推荐系统 s, 依据 $\delta^s_{u,i}$ 在每个数据集内对所有的实例进行排序, 并将实例分为 4 等份: Q1~Q4, 其中 Q4 代表绝对误差低, 即评分预测表现好的前 25%, Q1 代表绝对误差高的 25%。图 8(a)~(c)展现了基于 3 个数据集和 4 个推荐系统在 Q1~Q4 的可解释性评分分布情况。通过图 8 可以推测出推荐系统的预测准确性与可解释性评分有正相关关系, 在每个数据集、每个推荐系统上, 都有超过 50%的实例既能够获得 3 分以上的可解释性评分, 又具有比较准确的预测评分。因此可以推测, 对于可解释性较好的实例, 推荐系统对于这条实例的预测分数很有可能会更加准确。



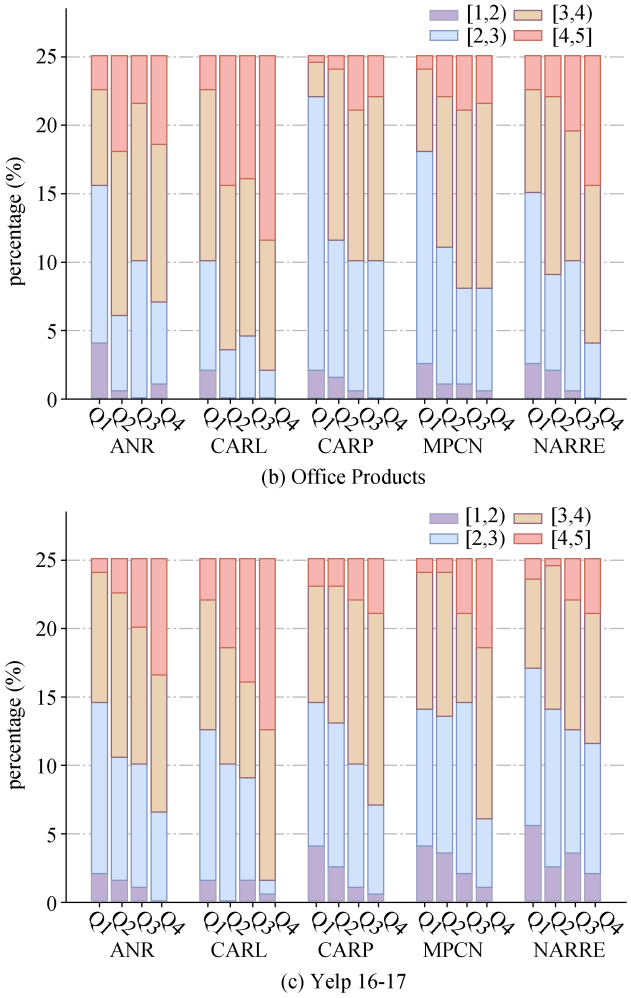


图 8　Q1~Q4 上的可解释性评分分布

Figure 8　Distribution of interpretability scores for Q1—Q4

为了进一步探究可解释性评分与分数预测的准确率之间的关系, 本文在每个数据集、每个推荐系统内计算了可解释性评分与绝对误差之间的皮尔逊相关系数, 结果如表 5 所示。而整体的相关系数均值为–0.342, 这也验证了推荐系统评分预测的准确性与可解释性之间确实存在正相关关联关系。因此可以推测一个更准确的预测结果意味着推荐系统更能“理解”用户的意图或者商品特征。

本文最后探究了词性分布与可解释性评分之间的关系。本文通过 ANR、CARL、CARP、MPCN 生成的带权重的评论文本, 将每个评论文本中的词的注意力权重进行排序, 选取注意力权重最大的前 5% 的词, 使用自然语言处理工具 NLTK[32]对所选定的词进行词性标注分析(POS tagging)。由于 NARRE 只能得到每条评论的注意力权重, 只能抽取出权重高的评论, 但本文认为关注单条评论的词性分布并没有很大的研究价值, 因此 NARRE 没有参与到词性分析中。根据挑选出的注意力权重高的词, 本文选取了出现频率比较高的 7 类词性: NN、JJ、RB、DT、IN、VBZ、PRP。其中“NN”为名词、“JJ”为形容词、“RB”为副词、“DT”为限定词、“IN”为介词、“VBZ”为动词第三人称、“PRP”为人称代词。本文计算了每类词性在对应实例选取的词中所占的比例, 并在每个可解释性分数段内计算均值, 得到的结果如图 9~11 所示。整体来看针对不同的数据集、不同的推荐系统, 高权重的词的词性分布有明显差别, 比如 MPCN 在 3 个不同的数据集上都会为更多的名词学习到高权重; 而在 Yelp 数据集中, 由于选取的实例大部分是用户对餐馆的评价, 有明确的对餐馆的味道、服务、环境等方面的描述, 因此本身数据集中包含的名词、形容词比较多, 也在图 11 中有所体现。尽管在同一个数据集、同一个推荐系统内, 不同分数段内所选取词的词性分布比较相似, 但是仍然能够观察出细微的区别: 在每个数据集、每个推荐系统内, 具有高权重的形容词所占的比例越高, 可解释性评分越高; 而可解释性评分在[1, 3)范围内的实例往往在限定词、介词、动词这几类词性中分布较多的高权重的词, 但可解释性评分相对较高的实例在这几类词性里的占比较低。形容词中包含了对商品的特征的描述, 还可能会有一些情感偏向明确的词, 因此可解释性评分高的推荐系统会更有可能为形容词赋予比较高的权重值。

表 6　可解释性评分与绝对误差间的皮尔逊相关系数

Table 6　Pearson’s correlation coefficient between interpretability score and absolute error

数据集	ANR	CARL	CARP	MPCN	NARRE
Musical Instruments	–0.34	–0.34	–0.30	–0.35	–0.33
Office Products	–0.36	–0.44	–0.38	–0.30	–0.32
Yelp 16–17	–0.36	–0.41	–0.36	–0.26	–0.29

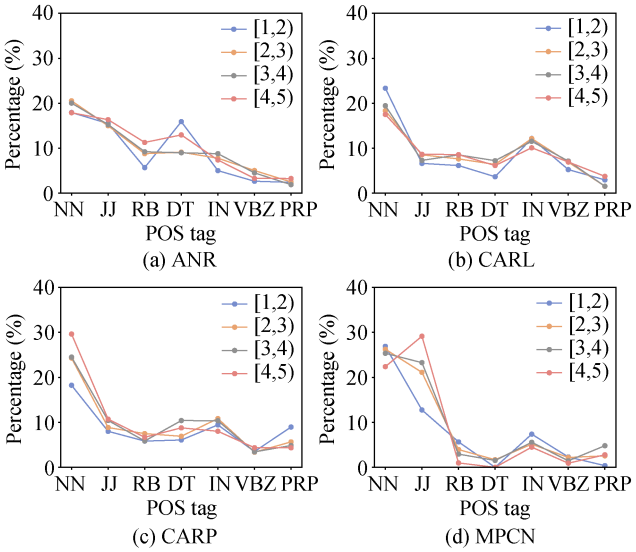


图 9　词性分布-Music Instruments

Figure 9　The part of speech distribution of Music Instruments

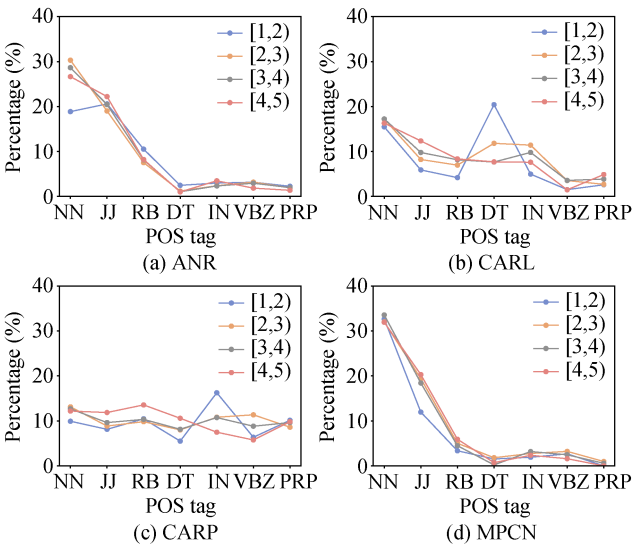


图 10　词性分布-Office Products

Figure 10　The part of speech distribution of Office Products

6　结论

本文提出了一种能够综合评价基于评论的推荐系统的可解释性的方法，通过对 5 个基于评论的深度推荐系统，在 3 个真实的数据集上进行的人工标注工作，发现当前的基于评论的深度推荐系统内置注意力机制有 50%以上的可能性能够精确地捕捉到用户对目标商品的偏好信息。本文通过对得到的可解释性评分进一步分析发现，推荐系统的可解释性表现在一定程度上与推荐系统的分数预测精度有正相关的关联关系；并且发现推荐系统在更短的评论

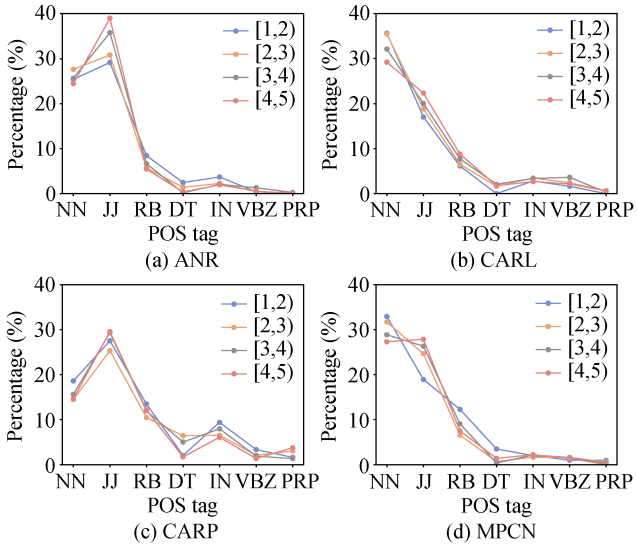


图 11　词性分布-Yelp

Figure 11　The part of speech distribution of Yelp

上更容易捕捉到有用的信息；可解释性评分高的推荐系统会偏向为更多的形容词赋予较高的权重。在未来的工作中，本文计划设计一种能够通过推荐系统内置的注意力机制得到的注意力权重值，可以自动评估推荐系统的可解释性的方法。总的来说，本文提供了一种评价推荐系统可解释性的新思路，也为探索更好的基于评论的推荐系统的解决方案提供了一些启示。

参考文献

[1] McAuley J, Leskovec J. Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text[C]. *The 7th ACM conference on Recommender systems*, 2013: 165-172.

[2] Tan Y, Zhang M, Liu Y, et al. Rating-boosted latent topics: Understanding Users and Items with Ratings and Reviews[C]. *IJCAI*, 2016, 16: 2640-2646.

[3] Wang C, Blei D M. Collaborative Topic Modeling for Recommending Scientific Articles[C]. *The 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011: 448-456.

[4] Seo S, Huang J, Yang H, et al. Interpretable Convolutional Neural Networks with Dual Local and Global Attention for Review Rating Prediction[C]. *The Eleventh ACM Conference on Recommender Systems*, 2017: 297-305.

[5] Chin J Y, Zhao K Q, Joty S, et al. ANR: Aspect-Based Neural Recommender[C]. *The 27th ACM International Conference on Information and Knowledge Management*, 2018: 147-156.

[6] Su X Y, Khoshgoftaar T M. A Survey of Collaborative Filtering Techniques[J]. *Advances in Artificial Intelligence*, 2009, 2009:

1-19.

[7] Koren Y, Bell R, Volinsky C. Matrix Factorization Techniques for Recommender Systems[J]. *Computer*, 2009, 42(8): 30-37.

[8] Koren Y. Factorization Meets the Neighborhood: A Multifaceted Collaborative Filtering Model[C]. *The 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2008: 426-434.

[9] Rendle S. Factorization Machines[C]. *2010 IEEE International Conference on Data Mining*, 2010: 995-1000.

[10] He X N, Liao L Z, Zhang H W, et al. Neural Collaborative Filtering[C]. *The 26th International Conference on World Wide Web*, 2017: 173-182.

[11] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation[J]. *the Journal of Machine Learning Research*, 2003, 3: 993-1022.

[12] Wang H, Wang N Y, Yeung D Y. Collaborative Deep Learning for Recommender Systems[C]. *The 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015: 1235-1244.

[13] Mnih A, Salakhutdinov R R. Probabilistic Matrix Factorization[J]. *Advances in Neural Information Processing Systems*, 2007, 20: 1257-1264.

[14] Song K S, Gao W, Feng S, et al. Recommendation Vs Sentiment Analysis: A Text-Driven Latent Factor Model for Rating Prediction with Cold-Start Awareness[C]. *The Twenty-Sixth International Joint Conference on Artificial Intelligence*, 2017: 2744-2750.

[15] Kim Y. Convolutional Neural Networks for Sentence Classification[C]. *The 2014 Conference on Empirical Methods in Natural Language Processing*, 2014: 1746–1751.

[16] Mikolov T, Karafiát M, Burget L, et al. Recurrent Neural Network Based Language Model[C]. *Eleventh Annual Conference of The International Speech Communication Association*, 2010: 1045-1048.

[17] Kim D, Park C, Oh J, et al. Convolutional Matrix Factorization for Document Context-Aware Recommendation[C]. *The 10th ACM Conference on Recommender Systems*, 2016: 233-240.

[18] Zheng L, Noroozi V, Yu P S. Joint Deep Modeling of Users and Items Using Reviews for Recommendation[C]. *The Tenth ACM International Conference on Web Search and Data Mining*, 2017: 425-434.

[19] Catherine R, Cohen W. TransNets: Learning to Transform for Recommendation[C]. *The Eleventh ACM Conference on Recommender Systems*, 2017: 288-296.

[20] Gao J Y, Lin Y, Wang Y S, et al. Set-Sequence-Graph: A Multi-View Approach towards Exploiting Reviews for Recommendation[C]. *The 29th ACM International Conference on Information & Knowledge Management*, 2020: 395-404.

[21] Chen C, Zhang M, Liu Y Q, et al. Neural Attentional Rating Regression with Review-Level Explanations[C]. *The 2018 World Wide Web Conference*, 2018: 1583-1592.

[22] Tay Y, Luu A T, Hui S C. Multi-Pointer Co-Attention Networks for Recommendation[C]. *The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018: 2309-2318.

[23] Wu L B, Quan C, Li C L, et al. A Context-Aware User-Item Representation Learning for Item Recommendation[J]. *ACM Transactions on Information Systems*, 2019, 37(2): 1-29.

[24] Chen J Y, Zhang H W, He X N, et al. Attentive Collaborative Filtering: Multimedia Recommendation with Item- and Component-Level Attention[C]. *The 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2017: 335-344.

[25] Lu Y C, Dong R H, Smyth B. Coevolutionary Recommendation Model: Mutual Learning between Ratings and Reviews[C]. *The 2018 World Wide Web Conference*, 2018: 773-782.

[26] Li C L, Quan C, Peng L, et al. A Capsule Network for Recommendation and Explaining what You Like and Dislike[C]. *The 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2019: 275-284.

[27] Wang X, Wang D X, Xu C R, et al. Explainable Reasoning over Knowledge Graphs for Recommendation[J]. *The AAAI Conference on Artificial Intelligence*, 2019, 33: 5329-5336.

[28] Pan D, Li X R, Li X, et al. Explainable Recommendation via Interpretable Feature Mapping and Evaluation of Explainability [EB/OL]. 2020: ArXiv Preprint ArXiv:2007.06133.

[29] Liu D H, Li J, Du B, et al. DAML: Dual Attention Mutual Learning between Ratings and Reviews for Item Recommendation[C]. *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019: 344-352.

[30] He R N, McAuley J. Ups and Downs: Modeling the Visual Evolution of Fashion Trends with One-Class Collaborative Filtering[C]. *The 25th International Conference on World Wide Web*, 2016: 507-517.

[31] Wu L B, Quan C, Li C L, et al. PARL: Let Strangers Speak out what You Like[C]. *The 27th ACM International Conference on Information and Knowledge Management*, 2018: 677-686.

[32] Bird S, Klein E, Loper E. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit[M]. “O'Reilly Media, Inc.”, 2009.

朱芮 于 2020 年在东北林业大学信息管理与信息系统专业获得学士学位。现在武汉大学网络空间安全专业攻读硕士学位。研究领域为信息检索。研究兴趣包括: 自然语言处理、数据挖掘。Email: ruizhu@whu.edu.cn

刘艺语 于 2019 年在中国海洋大学保密管理专业获得学士学位。现在武汉大学网络空间安全专业攻读硕士学位。研究领域为信息检索。研究兴趣包括: 自然语言处理、数据挖掘。Email: liuyiyu@whu.edu.cn

李晨亮 于 2012 年在南洋理工大学大学获得博士学位。现任武汉大学教授。研究领域为信息检索。研究兴趣包括: 数据挖掘、机器学习、社交网络分析、文本/网络挖掘和自然语言处理方面的研究。Email: cllee@whu.edu.cn

刘布楼 于 2020 年在武汉大学计算机科学与技术专业获得学士学位。现在清华大学计算机科学与技术专业攻读硕士学位。研究领域为网络信息检索。研究兴趣包括: 自然语言处理。Email: bulouliu@gmail.com

邹鑫雨 于 2020 年在武汉大学自动化专业获得学士学位。现在武汉大学网络空间安全专业攻读硕士学位。研究领域为信息检索。研究兴趣包括: 自然语言处理、数据挖掘。Email: zouxinyu@whu.edu.cn