融合几何信息瓶颈的可解释推荐系统

Hanqi Yan、Lin Gui、Menghan Wang、Kun Zhang、Yulan He

arXiv:2305.05331v2 [cs.IR] 5 Jan 2024

摘要——**可解释推荐系统能够说明其推荐决策，从而增强用户对系统的信任。大多数可解释推荐系统要么依赖人工标注的理由来训练模型进行解释生成，要么利用注意力机制从评论中提取重要文本片段作为解释。提取出的理由通常局限于单条评论，可能无法识别评论文本之外的隐含特征。为避免昂贵的人工标注过程，并生成超越单条评论的解释，我们提出将从用户-物品交互中学习到的几何先验分布融入变分网络，该网络从用户-物品评论中推断潜在因子。单个用户-物品对的潜在因子可同时用于推荐和解释生成，并自然继承先验知识中编码的全局特征。在三个电子商务数据集上的实验结果表明，与使用 Wasserstein 距离的变分推荐系统相比，我们的模型显著提升了可解释性，同时在推荐行为方面取得了与现有基于内容的推荐系统相当的性能。代码已公开发布于**
h
t
tps://github.co
m/hanqi-qi/Explainable-Recommender-GIANT。

索引词—**推荐系统、可解释性、信息瓶颈**

I. 引言

通常，推荐系统会将用户的偏好与物品特征（例如物品描述或与物品相关的评论）进行比较，或研究用户与物品之间的历史交互（例如评分、购买或点击行为），以找出用户可能感兴趣的物品。除了推荐物品外，可解释推荐系统还会生成推荐决策背后的理由 [1], [2]。现有的大多数可解释推荐系统要么生成理由，要么从给定的用户—物品评论中提取文本片段，用以解释模型决策。生成理由需要使用经过标注的数据进行训练，例如用户提供的简短评论，用来解释其与物品交互的行为。基于抽取的可解释推荐系统，要么依赖评论中带有情感倾向且经过标注的方面片段进行训练 [3]–[7]，要么基于注意力权重，从目标物品的评论中提取句子作为推荐理由 [8]–[15]。

我们认为，解释生成模型通常依赖人工标注的理由进行监督，而在实践中获取这些标注需要投入大量人力。基于抽取方法构建的可解释推荐器只能从单个用户-物品评论文档中抽取理由，因此存在以下局限。首先，有些评论可能过于笼统，无法解释用户评分，因而无法用于解释生成。例如，评论“我真的很喜欢这款智能手机，会推荐给朋友”并未提供任何线索来说明用户为何喜欢这款智能手机。其次，直接从评论文档中抽取的特征可能无法反映某些全局属性，而这些属性只能通过隐式的用户-物品交互来识别。对于用户没有直接购买或评分、但受到其他志趣相投的用户青睐的物品，我们仍然可以从相关评论中获得有意义的见解。

为解决上述局限性，我们提出一种**几**何**信**息**瓶**颈 框**架**（**GIANT**），将从用户-物品交互图中学习到的先验融入其中，以改进从评论中提取的潜在因子。本质上，我们利用从一种模态（即用户-物品交互图）中学习到的先验，约束另一种模态（即评论文本）中潜在因子的学习。如此一来，我们便能自然地解决基于变分自编码器（VAE）的方法中高斯先验假设过于简单的问题 [16]。具体而言，我们的先验源自用户-物品交互图中获得的用户/物品簇。每个用户或物品都具有软聚类分配，该分配表示为簇上的概率分布，并与其到各簇质心的距离成反比。因此，我们将该先验称为几何先验，因为它将用户/物品的几何属性转化为与簇相关的概率分布。随后，我们将该先验融入变分网络，以约束从评论文本中学习潜在因子的过程。对于一个用户-物品对（u，i），我们将用户的历史评论和该物品关联的评论（不包括 u对i的评论）输入变分网络，以推断与用户和物品相关的潜在因子，并将其用于评分预测和解释生成。

图1展示了如何利用所提出的几何先验来构建可解释推荐系统。为了将该先验施加于以用户评论集合和物品评论集合为输入的变分框架中的潜变量，我们应用 KL 散度，最小化从评论文本中学习到的用户潜在因子z_u和物品潜在因子 z_i的后验分布与从用户-物品交互图中推导出的几何先验之间的差异。变分网络推断出的潜在因子可以视为总结输入评论关键语义的主题。对于输入的用户-物品对，用户（或物品）潜在因子中最显著的维度被视为揭示用户偏好（或物品特征）的最重要主题。正如实验部分将展示的，

稿件收稿日期：2023年3月23日；修订日期：2023年11月7日；录用日期：2023年12月23日。

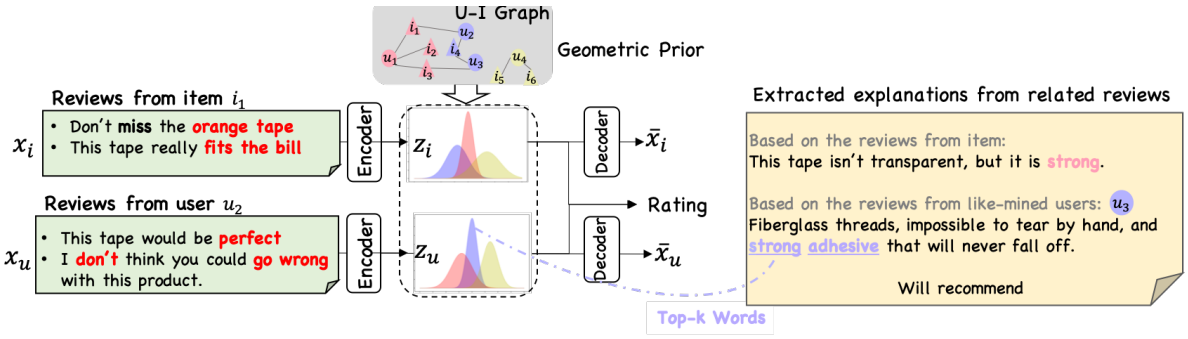


图 1. 所提出的几何先验反映了交互图中用户与物品之间的距离，频繁交互的用户/物品会被分配到同一个簇中（以节点颜色表示）。现有的抽取式方法只能从给定的评论对（x_i，x_u）中识别指示性词语（以红色显示），而我们的方法能够提取由其他志趣相投的用户撰写或针对相似物品的句子级评论；这些用户或物品被分配到同一个潜在维度/簇中。也就是说，我们提取用户 u_3 的评论中的句子，因为该用户与用户 u_2 属于同一个簇；而 u_2 的评论过于宽泛，无法概括关键特征。

对于每个主题（即潜在因子的一个维度），我们可以检索与其关联度最高的前-k 个词，作为该主题的表示。为了生成句子级解释，我们从位于同一用户/物品簇中的评论中找出与用户主题表示和物品主题表示最相似的句子。 [1]

在三个常用基准数据集上的实验结果表明，所提出的方法在推荐预测方面取得了与若干强基线方法相当的性能。此外，定量和定性分析均表明，与采用标准高斯先验或使用 Wasserstein 距离的变分自编码器相比，在几何先验正则化的潜在空间中，我们可以观察到清晰且有意义的簇。用户研究表明，我们生成的解释在相关性、忠实性 和信息性方面表现更佳。我们的贡献可概括如下：

- 我们提出了一种基于用户-物品交互推导出的几何先验分布，以融入相似用户/物品的全局特征，从而能够在显式用户-物品交互之外生成推荐解释。
- 我们证明，几何先验分布可以有效地融入变分框架；该框架通过一个上界，约束文本输入与受几何先验分布正则化的潜在表示之间的互信息。
- 实验结果表明，我们提出的方法在评论评分预测和解释生成方面均有效。

II. 相关工作

我们回顾推荐系统，尤其关注基于变分自编码器构建的系统，以及能够解释推荐决策的系统。

a) 使用变分自编码器的推荐系统：用户评论、物品或品牌描述等文本数据对于开发高质量的推荐系统可能很重要，因为这些数据可用于解决用户-物品交互中的稀疏度问题。HFT [17] 和 CTR [18] 采用了潜在狄利克雷分配（LDA） [19] 从评论文本中提取潜在主题。由于具备提取潜在且可解释特征的能力，变分自编码器也可用于文本建模 [20]–[22]。 [21] 认为变分自编码器中常用的各向同性高斯分布过于简化，并提出了 BiVAE，通过引入受约束的自适应先验分布来学习依赖于用户和物品的先验分布。最近，评论信息被用于以更细粒度解耦潜在用户意图，例如，DisenGCN [23] 和 DisenHan [22] 使用图注意力机制来区分多种关系和特征。

b) 可解释推荐系统：一种常见的可解释方法是使用提供的真实解释训练语言生成模型，这些解释可以是给定评论的第一句话，也可以是评论文本中由人工标注的文本片段 [4], [24]。其中一些方法首先提取用户-方面-观点元组和物品-方面-观点元组作为真实值，然后对这两种元组之间的交互进行建模，以推导评分和指示性特征 [25],[26]。基于检索的方法（又称抽取式方法）直接从目标物品的历史评论中选取代表性文本片段（理由），即词语或句子，用于解释推荐器行为。基于注意力的方法在基于评论的可解释推荐系统中十分常见 [14], [15]。除了利用注意力权重选择理由外， [27], [28] 还提出了不同的优化目标来改进选择结果。

然而，上述方法存在两个主要问题。一些方法需要先提取方面，以提供与方面相关的预测特征作为解释，因此严重依赖所采用的方面提取器的质量；另一些方法则依赖人工标注的理由来训练监督式语言生成器，以生成解释。我们提出的方法则只需利用来自评论评分的监督信号即可生成解释。

III. 几何信息瓶颈

为了纳入积极交互过的用户/物品的全局特征，以解释推荐行为，我们提出将从用户-物品交互图中推导出的几何先验融入变分网络，从评论文本中学习用户和物品的潜在因子。尽管

[1]关于如何提取句子级评论作为解释的详细信息，请参阅§V：解释生成。

已有许多工作提出在标准变分自编码器中采用更丰富的先验分布，以推断更复杂、更真实的后验分布 [29],[30], ，但我们的解决方案不同于现有方法，因为我们需要施加从图模态中学习到的先验知识，以约束文本模态中的变量。接下来，我们将展示如何在信息瓶颈理论框架下实现这一点。

基于信息瓶颈（*IB*）理论 [31], ，我们可以训练一个编码器-解码器架构 $x \xrightarrow{encode} z \xrightarrow{decode} \bar{x}$，其中 $x \in X$、$z \in Z$ 和 $\bar{x} \in X$ 分别依次表示输入、隐藏表示和重构表示 ¯，以保留关于 X 在 Z 中的有意义信息，同时最大限度地压缩 X 的信息。我们通过最大化以下目标来实现这一点：

$$O_{IB} = I(\bar{X}; Z) - \beta \cdot I(X; Z), \tag{1}$$

其中， $I(\cdot)$ 表示互信息， β 是拉格朗日乘子。第二项旨在最大限度地压缩 X，等价于最小化 X 与 Z 之间的互信息，记为 $I(X; Z) \leq I_c$， I_c 是上界。

直观地说，我们假设输入在不同模态下具有表示 $x^t \in X^t$ 和 $x^g \in X^g$ [2] ，并满足约束 $I(X^t; X^g) \geq I_x$。由于 X^t 和 X^g 源自同一个输入，因此它们应当相关。随后，在两种模态中分别独立训练、并通过公式（1）进行优化的两个编码器-解码器架构，受到 $I(X^t; Z^t) \leq I_c$ 和 $I(X^g; Z^g) \leq I_c$ 的约束。我们为多模态信息瓶颈推导出如下新的上界：[3]

$$I(X^t; Z^g) \leq H(X^t) - H(X^g) + H(X^g|X^t) + I_c, \tag{2}$$

其中， $H(\cdot)$ 和 $H(\cdot|\cdot)$ 分别表示熵和条件熵。因此，我们可以使用来自文本模态的输入来最大化目标函数，并通过另一种模态进行正则化， [4]

$$O^t_{IB} = I(\bar{X}^t; Z^t) - \beta \cdot I(X^t; Z^g). \tag{3}$$

根据 [32],，上述公式可通过采用重参数化技巧 [33]，并加入随机高斯噪声 ϵ来近似：

$$\begin{aligned} O^t_{IB} = \mathbb{E}_{x^t \sim X^t} \mathbb{E}_{\epsilon \sim p(\epsilon)} [- \underbrace{\log q(\bar{x}^t | f(x^t, \epsilon))}_{\text{Recons. term: } Q(x^t)} \\ + \beta \cdot \underbrace{\mathrm{KL}(p(Z^t|x^t, \epsilon) | r(Z^g))}_{\text{KL-div. term: KL}(Z^t|Z^g)}, \end{aligned} \tag{4}$$

其中， x^t 是来自输入 X^t的一个训练样本， $\bar{x}^t$ 是重构输出； $q(\cdot)$ 和 $p(\cdot)$ 分别是由解码器和编码器估计的后验概率， $r(Z^g)$ 是由用户-物品交互图估计的分布。这里，公式 (4) 的形式与 β-VAE 类似，其中包含由后验概率定义的重构项 $Q(x^t)$，以及 KL 散度项 KL($Z^t|Z^g$)，后者由来自另一种模态的先验分布 Z^g进行正则化。

然而，将几何先验分布纳入变分网络时，需要解决两个关键挑战：

(1) 如何从用户-物品交互图中推导先验分布 $r(Z^g)$；(2) 如何计算 $Q(x^t)$，其在公式 (4) 中被定义为 $\log q(\bar{x}^t | f(x^t, \epsilon))$ 的对数条件概率。

a) 使用高斯核聚类得到的几何先验： 现在我们讨论如何从用户-物品交互图中推导几何先验。在该图中，节点可以是用户或物品，而频繁交互的节点更可能具有相似的偏好或特征。我们可以使用图神经网络，以评论评分作为监督信号，学习用户/物品的节点表示。随后，相似的用户/物品会自然地聚集到 K 个簇之一，记作 $\boldsymbol{C}_k$，$k \in \{1, 2, \cdots, K\}$. [5]

接下来，我们展示如何根据聚类结果定义几何先验分布。这里，我们使用高斯核计算用户或物品与其所属簇的质心之间的距离 ρ^g_k ，如公式（5）所示 $\boldsymbol{C}_k$ 。

$$\rho^g_k = \frac{\exp\left(-\|\boldsymbol{e}^g - \boldsymbol{C}_k\|^2 / 2\alpha^2\right)}{\sum_{k'=1}^{K} \exp\left(-\|\boldsymbol{e}^g - \boldsymbol{C}_{k'}\|^2 / 2\alpha^2\right)}, \tag{5}$$

其中， $\boldsymbol{e}^g$ 是用户或物品的图节点嵌入， $\boldsymbol{C}_k$ 是对应的第 k个簇质心向量， α 是一个超参数，用于根据数据密度调整分布。

可以很容易地证明 $\Sigma_k \rho^g_k = 1$。也就是说， $\{\sqrt{\rho^g_k}\}$ 的分布位于均匀 K球上的一个超球冠区域内；当 K 较大时，它可以弱近似为高斯分布 [34]。我们用 $r(Z^g)$表示从用户-物品交互图中获得的用户或物品在各簇上的概率分布，即 $r(Z^g) = \{r(z^g_u)\ r(z^g_i)\}$，其中 $r(z^g_u) = [\rho^g_{u1}, \cdots, \rho^g_{uK}]$，且 $r(z^g_i) = [\rho^g_{i1}, \cdots, \rho^g_{iK}]$。

*b)*条件概率的后验分布： 式（4）中的重构项 $Q(x^t)$ 可通过 Nadaraya-Watson 估计量 [35]推导得到：

$$\begin{aligned} Q(x^t) &= \log \frac{\hat{p}(\bar{x}^t_\epsilon, x^t)}{\hat{p}(x^t)} \\ &= \log \frac{\frac{1}{N} \sum_{j=1}^{N} \kappa\left(\frac{x^t - x^t_j}{h}\right) \cdot \kappa\left(\frac{\bar{x}^t_\epsilon - x^t_j}{h}\right)}{\frac{1}{N} \sum_{j=1}^{N} \kappa\left(\frac{x^t - x^t_j}{h}\right)}, \end{aligned} \tag{6}$$

其中， $\kappa(\cdot)$是用于密度估计的核函数， $\bar{x}^t_\epsilon$ 是加入随机噪声 ϵ后重构得到的 x^t ， x^t_j 是第 j个训练样本。实际应用中，我们选择径向基函数核$\kappa = \exp(-||x - x'||^2)$)，并忽略分母，因为目标是优化 $\bar{x}^t$的解码。然后，根据学习到的度量空间中的三角不等式，我们可以得到以下界： $Q(x^t) \propto \log(\exp(-||\bar{x}^t_\epsilon - x^t||^2))$。通过最小化损失，我们既能保留相邻表示的关键局部相似性，又能确保不同特征特征所对应的不同表示 [36], [37]。

因此，我们能够保证，采用我们提出的几何先验分布的信息瓶颈为：

- **可迁移的** – 根据公式（2），文本输入与经几何先验正则化的压缩潜变量之间的互信息仍受全局信息瓶颈的上界约束，即我们可以

[2]在本文其余部分中，我们使用上标 t 表示文本，使用 g 表示图。[3] 证明见附录 B。[4]证明见附录 B。

[5]下文中，由于用户和物品的处理方式相同，我们省略其下标 u、 i 。

利用一个模态的上界来指导另一个模态的训练。

•**可行** – 利用式 (5) 中的方案，我们无需使用预定义的先验分布或过度简化的先验分布，而是采用基于高斯核的距离分布来正则化潜在空间。表 III 中的实证推荐结果显示了几何先验的优越性。

• **可解释** – 不同模态的表示与其几何空间概率之间的对齐，使我们能够利用从评论文本中推断出的潜在因子（即表 VII 所示、嵌入于不同潜在维度中的语义簇）来生成推荐解释。

IV. GIANT 框架

本节将介绍几何信息瓶颈的一种实用实现，并利用它实现物品推荐和解释生成。

a) 图节点嵌入上的聚类。: 我们构建一个用户-物品交互图，其中节点为用户或物品；若用户对物品的评分高于该物品的平均评分，则在两者之间建立边。随后，我们使用*LightGCN* [38] 获取用户/物品嵌入。选择该方法是因为它省去了特征变换和非线性激活，因此成为近年来成功的推荐系统中广受欢迎的选择，相关研究 [39]–[41]也证明了这一点。之后，我们对学习得到的图节点嵌入应用 K-means，以得到簇质心向量。

b) 文本特征提取器。: 对于给定的用户-物品对 $(u,\ i)$，我们的推荐系统的输入是由 u 撰写的一组评论，以及关于 i 的一组评论（不包括 u对 i的评论）。由于每个用户或物品都关联多条评论，且部分评论可能很长，因此采用卷积神经网络（CNN）对评论进行编码，因为它在编码长序列时效率较高，相关研究见 [8]。随后，我们通过注意力机制聚合多个用户/物品的评论，分别得到最终的文本向量，记作 x_u^t或 x_i^t。给定文本向量，我们采用两个共享权重、经过几何正则化的变分网络（即编码器-解码器结构），以推断分别从用户和物品评论文本中得到的潜在因子 z_u^t 和 z_i^t。[6]

c) 最小化分布差异。: 为最小化式 (4) 中的 KL 散度，我们需要将 Z^g 和Z^t 投影到大小相同的度量空间中。具体而言，我们通过应用重参数化技巧 [33]，得到潜变量 $z^t = f(x^t, \epsilon)$。然而，由于初始化具有随机性，学习到的因子 Z^t 可能无法在 Z^g中找到对应项。因此，在基于单层 MLP 的编码器中，我们提出一种先验分布中心化技巧，通过使编码器权重接近从图中得到的簇质心 $\boldsymbol{C} \in \mathbb{R}^{d\times k}$，在簇（由用户-物品交互图得到）与编码器之间建立联系：

$$\mathcal{R}_{centroid} = \sum_{k=1}^{K} \left\| W_{en}^k - \boldsymbol{C}_k \right\|^2, \tag{7}$$

[6]为简洁起见，以下省略下标 u 和 i 。

其中， W_{en}^k 表示编码器权重矩阵 W_{en} 的第 k列， $\boldsymbol{C}_k$ 表示第 k个簇质心向量的表示。请注意， $R_{centroid}$ 会迫使 Z^t 的分布逼近先验分布，并忽略输入。因此，实践中，这种正则化仅用于训练初期的几个训练轮次。

一旦建立了编码器权重与簇质心之间的联系，文本潜变量 z^t 的学习便可视为将文本表示 x^t 软分配到不同簇的过程。受 [42],启发，我们采用 Student t-分布来缓解拥挤问题。这里，我们使用 τ 调整分布尾部，并将其取值映射到 $[0, 1]$范围内。文本潜变量 z_k^t 的第 k维表示相应用户或物品被分配到第 k个簇的概率：

$$\eta_k^t = \frac{1 + e^{\varphi(z_k^t)/\tau}}{\sum_{k'=1}^{K} e^{\varphi(z_{k'}^t)/\tau}}, \tag{8}$$

其中， τ 是 softmax 函数中的温度， $\varphi(z^t)$是对 z^t的线性变换。随后， Z^t 的后验分布与 Z^g 的先验分布之间的 KL 散度定义为：

$$\mathrm{KL}(p(Z^t|x^t,\epsilon)|r(Z^g)) = \sum_{k=1}^{K} \eta_k^t \log \frac{\eta_k^t}{\rho_k^g} \tag{9}$$

其中， η 是式（8）中定义的派生簇分布， ρ 是根据用户-物品交互图计算出的先验分布，定义见式（5）。

d) 重构损失: 如第三节所述，式（4）中的重构项具有如下性质： $Q(x^t) \propto \log(\exp(-||\bar{x}_\epsilon^t - x^t||^2))$。因此，它可以定义为均方误差函数：

$$Q(x^t) = \mathbb{E}_{x^t\sim X^t}\mathbb{E}_{\epsilon\sim p(\epsilon)}[||\bar{x}_\epsilon^t - x^t||^2], \tag{10}$$

其中， x^t 表示输入的用户或物品表示， $\bar{x}_\epsilon^t$ 表示重构后的表示。

e) 评分预测与最终目标函数: 获得用户 z_u^t 和物品 z_i^t 的正则化潜在文本表示后，我们将对应的 ID 特征添加到其中，以获得用于评分预测的最终用户和物品嵌入 ζ_u 和 ζ_i ， $\zeta_u = W_u z_u^t + \epsilon_u$， $\zeta_i = W_i z_i^t + \epsilon_i$，其中，ID 特征 ϵ_u 和 ϵ_i 由将用户 ID 和物品 ID 输入嵌入层生成。受推荐系统中潜在因子模型的启发，我们在最终预测层中引入用户和物品的全局偏置，得到 $\hat{r}_{ui} = f_{\text{cls}}(\zeta_u,\ \zeta_i) + b_u + b_i$，其中， f_{cls} 将用户和物品特征组合为一个标量，b_u 和 b_i 是由ϵ_u 和 ϵ_i导出的偏置。预测评分的回归损失 $\mathcal{L}_r$ 按照给定用户-物品对的均方误差计算。结合上述所有组成部分，我们推导出如下训练目标函数：

$$\mathcal{L} = \mathcal{L}_r + O_{IB}^t + \mathcal{R}_{centroid} \tag{11}$$

f) 时间复杂度与可扩展性分析: 在本小节中，我们对 **GIANT** 和我们选定的基础图模型 LightGCN 进行复杂度分析。假设用户节点数为 m，物品节点数为

n，数据稀疏度定义为 α，用户-物品交互图中的边数为 $E = \alpha \cdot m \cdot n$。用户和物品的平均评论数分别为 $|D_u|$ 和 $|D_i|$。令 G 表示 GCN 层数。d 是图节点嵌入的维度，也是簇质心向量 $\mathcal{C}$的维度。K_u 是用户簇的数量，K_i 是物品簇的数量，T 是 K-Means [7]的最大迭代次数。为简单起见，假设 CNN 中所有核的大小均为 c，核的总数为 f。每条评论文本的填充长度表示为 L。

我们使用经过充分训练的 LightGCN 得到的用户和物品嵌入。其复杂度分析结果已在 [43] 中报告。额外计算包括几个步骤：使用 K-means 对用户/物品进行聚类、采用带有评分预测损失$\mathcal{L}_r$的变分自编码器框架进行最终评分预测、信息瓶颈正则化 O^t_{IB}以及质心正则化 R_{centroid}。对于文本特征提取，与单个用户或物品相关的所有评论会聚合为一条，再输入 CNN。因此，用户和物品的预期输入通道大小分别为 $m|D_u|$ 和 $n|D_i|$。由此，预期总计算复杂度可表示为 $\mathcal{O}(m|D_u|Lc^2f)+\mathcal{O}(n|D_i|Lc^2f)$。对于评分预测，我们只计算所有用户-物品交互的点积，复杂度为 $\mathcal{O}(E)$。对于式（9）中 KL 散度的计算，每个用户-物品对的复杂度由 (K_u+K_i) 决定。由于用户和物品潜变量的维度分别与 K_u 和 K_i相同，因此每个 $(u,\ i)$ 对的重构复杂度也为 $(K_u + K_i)$。由于数量

LightGCN	邻接矩阵 GCN BRP 损失	$\mathcal{O}(\lvert E\rvert)$ $\mathcal{O}(\lvert E\rvert^2Gd)$ $\mathcal{O}(\lvert E\rvert d)$
K-均值	聚类	$\mathcal{O}(K_umT)$+$\mathcal{O}(K_inT)$
CNN	特征提取器	$\mathcal{O}(m\lvert D_u\rvert Lc^2f) + \mathcal{O}(n\lvert D_i\rvert Lc^2f)$
VAE	评分预测 O_{IB} R_{centroid}	$\mathcal{O}(E)$ $\mathcal{O}\ (EK_u + EK_i)$ $\mathcal{O}(dEK_u + dEK_i)$

表 I GIANT 的解析时间复杂度。

簇的数量远小于物品或用户的数量，因此复杂度的预期增长主要来自 K-Means 聚类、文本特征提取和评分预测步骤，粗略来说：

$$\mathcal{O}((K_uT + |D_u|Lc^2f)m + (K_iT + |D_i|Lc^2f)n)+ \mathcal{O}((dK_u + K_u + dK_i + K_i + 1)E) \quad (12)$$

在最高配置中，我们将 d 设为 64，K_u 并将 K_i 设为 64，$|D_u|$ 并将 $|D_i|$ 设为 100，L 设为 32，核大小设为 3，核的数量设为 3。因此，主要项是 m和 n对应的第二个系数，约为 86，$400\cdot m$和 86，$400\cdot n$。据此，我们可以将计算复杂度简化为 $\mathcal{O}(|D_u|Lc^2fm+|D_u|Lc^2fn+(dK_u+dK_i)E)$。考虑到数据集的稀疏度，如表 II 所示，我们可以将 E 近似为 $(3.5e-3)mn$，

[7]请注意，我们使用 Scikit-Learn 中的 K-Means 包进行聚类，表中记录了平均计算复杂度。

总体复杂度主要由 $|D_u|Lc^2fm+|D_i|Lc^2fn+14mn$ 决定。该复杂度随每个用户或物品的评论数量线性增长。

将我们的计算成本与 LightGCN 的计算成本进行比较，后者表示为 $2|E|^2G$，其中 G 表示 GCN 层数。我们发现，只有当 $|D_u|Lc^2fm+7mn$ 的结果接近（αmn）[2]时，我们的时间复杂度才与 LightGCN 相当。实际上，在我们的实验配置中，这两个结果处于同一水平。因此，我们的时间复杂度几乎等同于现有基于图的推荐系统。

V. 句子级解释生成

我们的变分框架可以识别每个潜在维度的前 k 个词。为了让人们更容易理解模型决策背后的理由，我们从所有候选评论中提取最相关的 评论句子 （提取出的解释见图 1，示例见图 3），作为特定用户-物品对（u-i）的解释。

具体而言，我们的 **GIANT** 模型推断出潜在因子 z^t。潜在因子中最显著的维度表示用户 u 或物品 i所属的簇。召回率：在图空间中得到的簇数量等于潜在维度的数量。基于此，我们可以将用户候选评论确定为由与 u属于同一簇的用户所撰写的、针对物品i 的所有历史评论。类似地，我们从物品 i 的历史评论中识别物品候选评论，这些评论的主要主题与该物品的潜在主题相同。随后，我们使用每个主题中关联度最高的 5 个词来表示该主题，并通过聚合词级 GloVe [8] 词嵌入来生成主题表示。[9] 这些词是根据同一主题簇中评论的 TF-IDF 值选取的，并过滤掉停用词。评论句子的表示也通过聚合其组成词的 GloVe 词嵌入得到。然后，可以根据用户候选评论中各句子与用户潜在主题表示之间的余弦相似度，提取最相关的句子，作为用户 u偏好的摘要。类似地，根据物品候选评论中各句子与该物品潜在主题表示之间的余弦相似度，提取最相关的句子，作为物品 i特征的摘要。

VI. 推荐实验结果

A. 实验设置

a) 数据集和数据处理：我们选择了三个常用的电子商务数据集，包括*BeerAdvo-cate* [44] 以及两个亚马逊评论数据集，数字音乐 和办公用品 [45], 。这些数据集不仅包含用户与物品之间的交互，还包含评论文本。亚马逊评论数据 [10] 是一个热门数据集，包含 24 类产品。由于我们的方法从用户/物品中推导出.....

[8]https://nlp.stanford.edu/projects/glove/[9]我们还使用了更具代表性的词语进行了实验，但发现主题聚类的区分度较低。[10] https://jmcauley.ucsd.edu/data/amazon/

主要依据评论，我们使用亚马逊评论数据集的稠密子集（5-core），提取用户-物品对，使得剩余的每个用户和物品都至少有 5 条评论。表 II 总结了这三个数据集的统计信息。为了提升推荐系统的性能，我们对包含评论长度显著较长或评论数量异常多的条目（用户/物品）采用过滤机制。这是因为评论长度和评论数量都呈长尾分布，如 [8] 所述。具体而言，我们过滤掉评论 / 交互超过 100 次的用户或物品，并将这三个数据集的评论长度分别截断至 [200,400,300]。此外，我们随机抽取 80% 、 10% 、 10 % 作为训练集/验证集/测试集。

b) 评估指标：我们使用均方根误差（RMSE）和平均绝对误差（MAE）来评估评分预测准确率。我们还报告基于排序的评估结果，即精确率和召回率，用于衡量推荐物品与目标物品之间的重合程度。

数据集	#训练数	#用户数	#物品数	平均评分	稀疏度
BeerAdvocate	35450	6939	13122	3.74	4.4e-4
数字音乐	51768	5541	3568	4.22	2.9e-3
办公用品	35999	4902	2364	4.33	3.5e-3

表 II 所有数据集的评分范围均为 1-5。THEBEER 数据集的稀疏度最低。

*c)*基线方法： 我们将其与多种开源推荐系统进行比较，包括*HFT* [17]，*DeepCoNN* [46]，和*NARRE* [8]。*HFT* 将评论与评分相结合，并使用指数变换函数将评论文本与评分关联起来。*DeepCoNN* 在分别用于用户和物品的两个编码器之上，采用一个共享层对用户和物品进行交互建模。*NARRE* 是一种可解释推荐系统，它引入评论级注意力机制来选择重要评论，并将用户和物品 ID 作为评分预测的判别特征，同时将注意力权重较高的文本片段用作推荐依据。我们的**GIANT** 同样基于*NARRE*构建，但在编码器-解码器结构中引入了所提出的几何先验分布。最近，一种名为*RGCL* [47] 的图模型被提出，该模型引入了对比损失，并在其节点特征中同时考虑用户-物品交互和评论语义。我们的模型同样强调用户-物品交互在基于评论的推荐系统中的重要性，但引入了一种基于聚类的新型正则化技术。此外，为了突出我们所提出的正则化方法的差异，我们还将其与在*NARRE* 之上采用不同正则化方法来学习潜变量的其他几种编码器-解码器结构进行比较，并将其作为基线方法，即自编码器、*StandPrior* 、*WasserstainVAE* 和*IndivPrior*:

- *StandPrior* （又称 *StandVAE*）分别为用户和物品使用两个变分自编码器，每个变分自编码器都以标准高斯分布 $\mathcal{N}(0,1)$ 作为先验分布。
- *WasserstainVAE* 与 *StandPrior* 的不同之处在于，它使用 Wasserstein 度量 [48] 来计算后验分布 η 与先验分布 ρ之间的分布差异。
- *IndivPrior* 为每个用户或物品分配一个独立的高斯先验，其均值根据从用户-物品交互图中得到的相应用户或物品表示计算得出，即 $p(z_u^t) = \mathcal{N}(W_u\xi_u,\ \boldsymbol{I})$，$p(z_i^t) = \mathcal{N}(W_i\xi_i,\ \boldsymbol{I})$，其中 $\xi_u \in \mathbb{R}^d$, $\xi_i \in \mathbb{R}^d$ 表示用户和物品图节点，W_u 和 W_i 是大小为 $(d \times d)$ 的线性层中的可学习参数，d是图节点嵌入的维度。

d) 训练过程： 我们通过最小化式 (11) 中定义的目标函数来训练模型。KL 散度项的 β 设为 0.01，并对所有模型参数采用 L_2 正则化，权重为 0.001。我们仅在 0.5 的比例范围内训练先验中心化项，以便在后续训练中近似聚类中心。应在 O_{IB} 中稍后引入 KL 项，直到编码器中心得到充分训练。为此，我们采用类似的循环调度 [49] ，逐步调整每个训练轮次中的退火因子 λ 。具体而言，我们首先在 0.5 的比例范围内不使用 KL 项训练模型，然后在 0.25 的比例范围内将其从 0.5 退火至 1，最后固定 $\lambda = 1$。 [11]

B. 评分预测结果

表 III 给出了我们的模型在所有数据集上的评分预测结果，并与基线方法进行了比较。我们有以下观察结果。(1) *DeepCoNN* 基于非线性神经网络堆叠构建，用于评论语义建模，其表现优于*HFT* ； 后者利用指数变换函数，将评论文本中的主题分布与从评分中提取的潜在因子关联起来。这表明深度学习在特征提取方面具有优势。(2)自编码器 在*BeerAdvocate*上的表现略优于*NARRE* ，这说明它在提取用于评分预测的关键上下文信息方面是有效的。(3) *RGCL* 成为表现最佳的基线模型；与我们的*GIANT*相比，它甚至在啤酒 和音乐 数据集上取得了有竞争力的结果。然而，在办公用品数据集上，其整体表现不及我们的方法。三个数据集上的平均 RMSE/MAE 值分别为：*RGCL* 为 84.49/63.49，*GIANT*为 84.18/62.87，这表明我们的模型优于*RGCL*。这些结果证明了我们提出的信息瓶颈正则化应用于潜在语义空间的有效性和稳健性。通过这种方式，用户和物品可以根据交互数据划分到不同的簇中，而其他基线推荐算法并非如此。

a) 我们提出的信息瓶颈正则化的效果：随后，我们研究不同信息瓶颈的效果，即 *StandPrior*、*WasserstainVAE* 和 *IndivPrior*。从表 III 最后一组结果可以看出，*WasserstainVAE* 和 *StandPrior* 的整体表现优于*IndivPrior* 以及表 III 中的所有基线方法。这种提升可以用变分自编码器的特性来解释，它】【。

[11]超参数设置见附录 A .

在处理可用观测值较少的稀疏数据时，这一点尤其有益。然而，*IndivPrior* 对单个用户/物品施加独立的高斯先验分布，其性能比其他三种变分框架更差，甚至不如自编码器。我们的模型受益于从用户/物品簇中提取的丰富先验分布，同时避免使用全局正态先验（*StandPrior* 和 *WasserstainVAE*），或为每个用户/物品单独设置实例级先验分布（*IndivPrior*）。

模型	*BeerAdvocate*		数字音乐		办公用品	
	均方根误差(↓)	平均绝对误差(↓)	均方根误差(↓)	平均绝对误差(↓)	均方根误差(↓)	平均绝对误差(↓)
HFT	79.81	62.42	96.42	74.76	89.46	68.57
DeepCoNN	77.28	59.45	94.69	71.03	85.14	64.82
NARRE	76.80	58.94	93.69	69.30	84.40	63.40
RGCL	**74.65**	**57.12**	**92.02**	**68.49**	86.81	64.87
自编码器	75.94	58.93	93.97	69.13	85.03	64.32
StandPrior	75.80	58.73	93.60	69.21	84.69	63.97
WassVAE	75.62	58.06	93.81	69.32	84.20	63.43
IndivPrior	76.43	59.88	94.57	70.51	85.33	65.05
GIANT	75.36*	57.87*	92.87*	68.68*	**84.32**	**62.05***

表 III 所有方法在均方根误差（%）和平均绝对误差（%）上的性能比较。* 表示与最佳基线方法相比，$p < 0.01$ 具有统计显著性。所有行中的最佳结果以粗体标出，每组中的最佳结果以下划线标出。

b) 基于排序的结果：除准确度指标（即均方误差和平均绝对误差）外，我们还使用基于排序的指标（即命中率@10和归一化折损累计增益@10）评估模型，并比较表现最好的两个基线方法（表 IV）。由于没有对整个物品列表进行评估，我们遵循 [50], [51] 中的常见做法，随机抽取用户未交互过的 100 个负样本，并将测试物品与这 100 个物品一起排序。随后，对于命中率@10和归一化折损累计增益@10，均将排序后的列表截断至前 10 个。不同数据集上的性能趋势与准确度指标所反映的趋势有所不同。*RGCL* 在 Office 数据集上的表现相当，而在 *BeerAdvocate* 数据集上的表现则差得多，命中率为 0.47，而***GIANT***为 0.62。不过，我们的*GIANT*在基于排序的指标上总体仍然表现更优。具体而言，在命中率@10这一指标上，我们的模型得分为 0.65，而*RGCL* 得分为 0.61。*WaserstainVAE* 似乎是第二组中表现第二好的编码器—解码器模型。表 III 中两组模型的相对排序表现有所不同，而编码器—解码器结构模型内部的相对表现与准确度指标相似，即均方误差；总体而言，**GIANT** 在这两类指标上都取得了最佳结果。

C. 各损失项的贡献

我们研究了三个损失项对评分预测准确性和潜变量多样性的影响。由于变分网络很容易坍缩为无条件生成模型，即在极端情况下，所有输入都会被映射到同一个潜在编码 [52]。我们使用相应潜在值最大的维度索引作为

模型	*BeerAdvocate*		数字音乐		办公用品	
	命中率	NDCG	热门歌曲率	NDCG	热门歌曲率	NDCG
纯矩阵分解	0.43	0.23	0.78	0.52	0.61	0.33
SVD++	0.41	0.21	0.52	0.28	0.55	0.28
NeuMF	0.41	0.21	0.78	0.52	0.63	0.35
NARRE	0.61	0.52	0.65	0.59	0.59	0.64
RGCL	0.47	0.55	**0.72**	0.63	**0.64**	**0.72**
自编码器	0.52	0.49	0.62	0.60	0.61	0.62
StandPrior	0.55	0.54	0.68	0.63	0.61	0.62
WassVAE	0.55	0.54	0.69	0.65	0.61	0.60
GIANT	**0.62**	**0.57**	0.69	**0.67**	**0.64**	0.64

表 IV 基于排序的评估结果，即所有基线方法和我们提出的**GIANT**的命中率@10 和归一化折损累计增益@10。

簇分配 ID。然后，我们得到簇分配结果 $\mathcal{A} \in \mathbb{R}^{N \times K}$，其中 N 是测试样本的数量，K 是簇数。多样性根据熵 $H(X) = -\sum_{k \in K} p(k) \log(p(k))$ 计算，其中 $p(k)$ 表示落入第 k 个簇的样本数占所有样本数的比例。数值越大，表示多样性越好。我们记录用户和物品潜变量之间的最大多样性。结果如表 V 所示。

变体	*BeerAdvocate*			数字音乐			办公用品		
	NDCG	RMSE	Div	NDCG	RMSE	Div	NDCG	RMSE	Div
完整模型	0.57	75.36	1.89	0.67	92.87	3.46	0.64	84.32	1.58
-w/o$\mathcal{R}_{cent}$	0.55	75.96	0.04	0.65	92.62	0.07	0.62	84.58	0.02
-w/o KL项	0.55	75.64	1.43	0.65	93.83	2.86	0.61	84.23	1.21
-不使用 $Q(x^t)$	0.52	76.02	1.04	0.61	93.95	3.10	0.59	84.94	0.93
$Q(x^t)$ 余弦相似度	0.56	75.35	0.11	0.67	92.85	1.76	0.63	84.97	0.07

表 V 评分预测结果，包括归一化折损累计增益@10、均方根误差(%) 和不同模型变体的多样性 等指标。多样性越低，说明不同样本学习到的潜变量越倾向于坍缩为相似的潜在编码。

我们观察到，移除 $Q(x^t)$ 会导致最大的性能下降。移除 $\mathcal{R}_{centroid}$ 后，潜变量的多样性降至接近零，这表明几乎所有潜变量都落入同一个簇。这凸显了我们的先验中心化项能够使潜变量 z^t 有效反映软聚类分配的能力。我们还发现，在 $Q(x^t)$ 中用余弦相似度（$1 - cosine$）替换均方误差，会显著降低潜变量的多样性。这在一定程度上可以解释为，余弦相似度侧重于衡量输入 x^t 与重建输出 $\bar{x}^t$之间的夹角，而忽略了向量的大小；然而，大小在我们的场景中很重要。

VII. 潜变量与可解释性评估

除了评分预测外，我们还设计了实验来评估生成的潜变量并探究性能提升的来源。首先，我们评估生成的潜在簇在语义一致性和可分离性方面的表现。然后，我们考察所识别的重要潜在维度对评分预测的影响。最后，我们进行人工评估，以研究其可解释性。

A. 簇一致性与可分性

a) 语义一致性：为验证创建语义连贯的簇的能力，我们提出将簇一致性度量为簇内每对评论文档之间余弦相似度的平均值。我们首先将测试集中的评论输入预训练CNN编码器，获得文档级评论表示。随后，我们计算簇内每对评论表示之间余弦相似度的平均值，并进一步对所有簇取平均（表 VI）。我们采用在学习得到的图节点嵌入上应用 KMeans 所生成的簇作为对比，即式 (5) 中的 ρ。

模型	*BeerAdvocate*		数字音乐		办公用品	
	User	Item	User	Item	User	Item
图	0.092	0.085	0.155	0.214	0.245	0.200
GAINT	0.490	0.481	0.457	0.471	0.492	0.407

表 VI 相似度值越大，连贯性越好。**GIANT**生成的簇比图聚类的连贯性显著更高。

b) 聚类可分性：由于潜变量 z^t 的每个维度都对应一个簇，对于第 k 个维度，我们可以查找其中 z_k^t 值最高的相关评论，并将评论集中的高频词列为代表性主题词[12]。表 VII 展示了三个数据集上随机选取的三个簇中的代表性词语。对于我们的模型，不同簇中最突出的词语各不相同（因此用不同颜色标出），即“棕色烘烤”、“松木、焦糖”和“风味不错”。我们分别在数字音乐 和办公用品 数据集中识别出“*ABBA*”及其热门歌曲，以及“网线”和“路由器”等计算机外围设备。

我们很难从*WasserstienVAE* 的结果中看出明确的主题区分。例如，簇 2 和簇 3 的前 4 个词中有 3 个相同：“甜”、“清淡”和“麦芽”。此外，簇 2 中的“碳酸化”和簇 3 中的“白色泡沫”都与啤酒泡沫有关。结果表明，融入由用户/物品簇导出的先验分布后，我们提出的方法能够学习评论语义空间中的潜变量，从而生成可分性更好的主题聚类。

B. 通过扰动潜变量进行全面性评估

为了探究我们所识别簇的重要性，即取值较大的潜在维度，我们计算移除特定维度前后的性能变化，并将全面性 定义为：$\sum_i^N \left(r(z^i) - r\ z^i_{\setminus k}\right)^2$ ()，其中，$r(\cdot)$表示预测评分，z^i 表示第i个评估用户-物品对的潜变量，$z^i_{\setminus k}$ 表示从 z^i中移除所识别出的排名前k 的潜在维度。为此，我们将排名前k 个维度中的值替换为潜在变量的平均值

N

[12]我们排除了停用词以及出现在所有簇中的高频词。

根据 [53] 确定变量。在图 2 的折线图中，随着更多重要维度被移除，我们的模型表现出更大的性能变化。我们进一步随机删除k 潜在维度，并通过减去随机删除所导致的变化来计算相对性能变化（图 2 下方的表格）。我们的模型具有最显著的相对变化，其次是 *WasserstienVAE*；而 *StandPrior* 的相对变化为负，这表明随机移除潜在维度可能会对那些占主导地位的维度带来更大的变化。

C. 生成的解释的人工评估

我们进行人工评估，以验证所提出方法的可解释性。为了让人类更容易理解模型决策背后的理由，我们从用户/物品评论中提取最相关的句子，作为特定用户-物品对（用户 u和物品i）的解释。生成的解释示例如图 3 所示，同时还展示了模型的推荐建议：如果模型的预测评分高于平均预测评分，则将推荐；否则不会推荐。

我们提出了三项评估指标：相关性、忠实性和信息性，并邀请三位精通英语的人工评估者，对来自三个数据集的每个数据集中随机选取的 120 个用户-物品对所对应的 *WasserstienVAE*、*NARRE* 和*GIANT* 生成的解释进行 1 至 5 分评分。评估者不知道模型的标识。

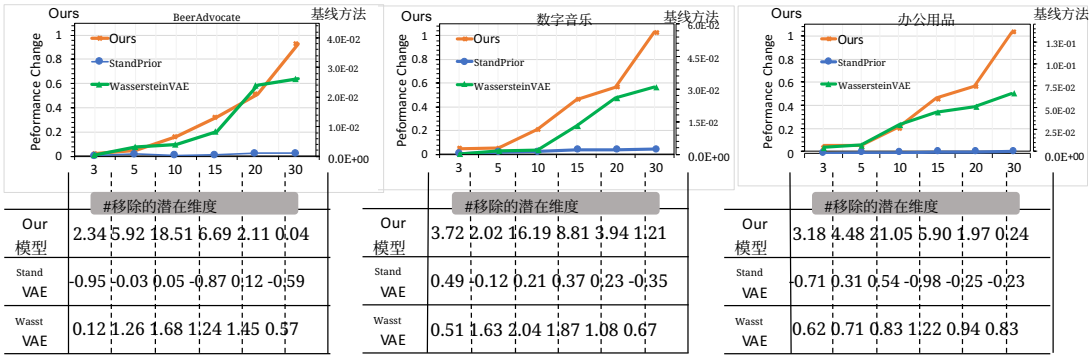
- 相关性。从用户评论和物品评论中提取的句子是否与相关主题、方面或内容有关？涉及方面重合度更高的解释应获得更高分数。
- 忠实性。从用户评论和物品评论中提取的句子是否会影响模型给出的推荐建议？请同时考虑所描述的情感倾向和方面。例如，如果用户和物品的所选评论都对产品的“价格”表达了积极的情感倾向，则忠实性得分应为 5。情感倾向与所提及方面之间的任何不一致都可能降低此分数。
- 信息性。所提供的解释是否有效地反映了用户偏好和物品特征？此项标准应与前几项分开，只关注提取的评论中对各种方面及其相关情感倾向的讨论程度。方面覆盖程度较低会导致信息性得分降低，例如“我真的很喜欢这个产品！”这样的表述。

表 VIII 中展示的人工评估结果表明，*WasserstienVAE* 和 **GIANT** 在相关性方面表现更好。这在一定程度上可以解释为：当当前评论不包含任何关键信息时，它们能够从同一用户簇划分结果中志趣相投的用户的评论中提取句子。相比 *WasserstienVAE*的优异表现进一步表明了我们提出的几何先验的有效性。

	GIANT	*WassersteinVAE*
Beer	**倾倒**，颜色，**棕色**，口感，**烘烤**，麦芽，几乎，中等，咖啡，深色 **不错**，**风味**，深色，**更好**，酿造，风格，差，加拿大，香气，饮用 **松木**，**焦糖**，挂杯，**柑橘**，口感，一丝，印度淡色艾尔，风味，酒体，浓郁	浅色，颜色，麦芽，**玻璃杯**，倒入，**饮用**，酸，倾倒，口感，**酒精** 香气，清淡，甜，麦芽，**白色泡沫**，口感，**香气**，一丝，**酿造**，风味 甜，**咖啡**，清淡，麦芽，香气，**碳酸化**，倒酒，口感，一丝，颜色
Music	**Dr. Dre**，**Prince**，**曾以知名**，西海岸，镜子，《爱》，恨，西海岸 故事，Jadakiss，总是	quot，专辑，**询问鲁弗斯**，歌曲，**放克**，**嗓音**，音乐，乐队，灵魂，爱
	手镯乐队，**维姬**，**声乐**，地方，现场演出，英雄，采取，跌倒，利物浦，披头士	歌曲，**余晖**，专辑，乐队，**美丽**，曲目，**歌词**，悬置，唱片， Christian ，
	《**舞后**》，《**抓住机会**》，《**妈妈咪呀**》，ABBA，精选集，《金钱》 《金钱》，《给我，给我》，《超级巡演》，《心知肚明》，黄金	唱片、专辑、歌曲、乐队、曲目、时间、声音、**头脑与心灵**、**民谣**、**现场演出**、《爱》
Office	**设置**、**计算机**、**无线**、网络、电缆、路由器、编辑、铅笔、墨水	**扫描**、**软件**、铅笔、**Mac OS**、工作、打印机、小型、已打印、颜色、文件
	照片、**书桌**、**激光**、扫描仪、工作、扫描、爱普生、扫描、足够、容易	易于使用、**白板书写**、文件夹、**鼠标垫**、办公室、**墨盒**、有效、已使用、 便利贴
	铅笔，**胶带**，**墨水**，钢笔，总之，钢笔，**橡皮**，夹子，还不错，橡皮擦	**推荐**，打印机，文件夹，**纸**，易于使用，**盒子**，功能，订书机，铅笔， 佳能，颜色

表 VII

从**GIANT** 和*WassersteinVAE* 在**数字音乐** 及**办公用品** 数据集中随机选取的三个簇中，按出现频率排序的最显著词语。对于每个数据集，我们突出显示其他两个簇中未出现的前三个词语。**GIANT** 生成的主题可分离性更好，而*WassersteinVAE* 未能生成清晰的主题模式。



	#移除的潜在维度					
Our 模型	2.34	5.92	18.51	6.69	2.11	0.04
Stand VAE	-0.95	-0.03	0.05	-0.87	0.12	-0.59
Wasst VAE	0.12	1.26	1.68	1.24	1.45	0.57

	#移除的潜在维度					
Our 模型	3.72	2.02	16.19	8.81	3.94	1.21
Stand VAE	0.49	-0.12	0.21	0.37	0.23	-0.35
Wasst VAE	0.51	1.63	2.04	1.87	1.08	0.67

	#移除的潜在维度					
Our 模型	3.18	4.48	21.05	5.90	1.97	0.24
Stand VAE	-0.71	0.31	0.54	-0.98	-0.25	-0.23
Wasst VAE	0.62	0.71	0.83	1.22	0.94	0.83

图 2. **顶部**：移除识别出的最重要的前 k 个潜在维度后得到的全面性 数值，$k \in \{3, 5, 10, 15, 20, 30\}$。我们的模型使用左侧 y 轴，两个基线方法使用右侧 y 轴。**底部表格**：减去随机移除造成的变化后，性能的相对变化。

The interpretations between 46 U-I pair. (GIANT)

Considering **your preferences** and **those of users who share your tastes**:
- It seems *sturdy enough* for classroom use.
- It is much thinner than duct tape and seems *stronger*.

Furthermore, based on **the reviews of this product**:
- You *can't just accidentally tear through this tape* you need to *use a sharp blade on purpose* to break the seal.
- I did use some on my grandson's toy lawn mower and *it is holding the broken handle on really well*.

We will recommend this item.

The interpretations between 46 U-I pair. (WassersteinVAE)

Considering **your preferences** and **those of users who share your tastes**:
- *I don't think you could go wrong* with this product.
- This tape isn't transparent, but it is *strong*.

Furthermore, based on **the reviews of this product**:
- I used the tape to secure some boxes i m using to store items. *I won't be using for a while*.
- Fiberglass threads, *impossible to tear by hand*, and *strong adhesive* that will never fall off.

We will recommend this item.

图 3. 从**GIANT** （上）和*WassersteinVAE* （下）中提取的对胶带这一物品的解释。在**GIANT**生成的示例中，用户和物品侧的评论/理由都聚焦于“结实耐用”，并揭示了许多对推荐很重要的方面。而*WassersteinVAE*提取的用户评论（例如“出问题”）未能捕捉关键因素，提取出的物品评论甚至自相矛盾，例如“不会再用了”与推荐行为相悖。

指标		相关性	忠实性	信息性
Beer	*NARRE*	2.7	2.4	3.1
	WasserVAE	**3.6**	3.5	**3.5**
	GIANT	3.5	**3.7**	**3.5**
Music	*NARRE*	4.4	3.2	3.8
	WasserVAE	4.0	3.3	3.0
	GIANT	**4.7**	**3.8**	**4.1**
Office	*NARRE*	2.7	2.8	3.9
	WasserVAE	3.5	2.8	3.8
	GIANT	**3.8**	**3.3**	**4.2**

表 VIII 对生成的解释的相关性、忠实性 和信息性进行的人工评估结果。评估对象是从三个数据集中随机选取的 120 个用户-物品对。每一对标注者针对这三项指标的 Cohen's Kappa 系数分别为 0.42、0.38 和 0.47。

VIII. 结论

在本文中，我们利用从用户-物品交互图中获得的共享共同兴趣/特征的用户/物品簇，通过以下方式细化评论文本的潜在因子

我们提出的几何信息瓶颈（**GIANT**）。实证结果表明，**GIANT** 在学习语义连贯且可解释的潜在空间方面表现更佳，生成的解释也更忠实于模型决策，同时在三个常用数据集上实现了具有可比性的评分预测准确率。

IX. 致谢

本研究得到英国工程与物理科学研究委员会资助（资助编号：EP/T017112/1、EP/V048597/1、EP/X019063/1）。HY 获得由华威大学和中国国家留学基金管理委员会共同资助的博士奖学金。YH 获得由英国研究与创新署资助的图灵人工智能奖学金（资助编号：EP/V020579/1）。

参考文献

[1] A. Ghazimatin、O. Balalau、R. Saha Roy 和 G. Weikum，“Prince：推荐系统中的提供方可解释性与反事实解释”，载于 第 *13* 届国际网络搜索与数据挖掘会议论文集，2020 年，第 196–204 页。[2] Y. Zhang、X. Chen 等，“可解释推荐：综述与新视角”，《信息检索基础与趋势®》，第 14 卷，第 1 期，第 1–101 页，2020 年。[3] Y. Zhang、G. Lai、M. Zhang、Y. Zhang、Y. Liu 和 S. Ma，“基于短语级情感分析的可解释推荐显式因子模型”，载于 第 *37* 届国际 *ACM SIGIR* 信息检索研究与发展会议，*SIGIR '14*，澳大利亚昆士兰州黄金海岸，*2014* 年 *7* 月 *6–11* 日，S. Geva、A. Trotman、P. Bruza、C. L. A. Clarke 和 K. J¨arvelin 编。ACM，2014 年，第 83 92 [ ] 10 1145 2600428 2609579 – 页。在线[4]提供：https://doi.org/ 。 / 。J. Ni、J. Li 和 J. J. McAuley，“利用远程标注的评论和细粒度方面为推荐提供依据”，载于 *2019* 年自然语言处理实证方法会议暨第 *9* 届国际自然语言处理联合会议论文集，*EMNLP-IJCNLP 2019*，中国香港，*2019* 年 *11* 月 *3–7* 日，K. Inui、J. Jiang、V. Ng 和 X. Wan 编。计算语言学协会，2019 年，第 188–197 页。 [ ] 10 18653 1 19 1018 在线提供：[5]https://doi.org/ 。 /v /D - Z. Chen、X. Wang、X. Xie、T. Wu、G. Bu、Y. Wang 和 E. Chen，“用于可解释推荐的协同注意力多任务学习”，载于 第二十八届国际人工智能联合会议论文集，*IJCAI 2019*，中国澳门，*2019* 年 *8* 月 *10–16* 日，S. Kraus 编。ijcai.org，2019 年，第 2137–2143 页。 [在线]。提供：https://doi.org/ 10 24963 2019 296 。 /ijcai。 / L. Li、Y. Zhang 和 L. Chen[6]，“为推荐生成神经模板解释”，丛书：CIKM '20。美国纽约州纽约市：计算机协会，2020 年，第 755–764 页。 [在线]。提供：https://doi.org/ 10 1145 3340531 3411992 。 / 。J. Tan、S. Xu、Y. Ge、Y. Li、X. Chen 和 Y. Zhang，反事实可解释推荐。美国纽约州纽约市：计算机协会，2021 年，第 1784–1793 页。 [在线]。提供：https://doi.org/ 10 1145 3459637 3482420 。 / 。C. Chen、M. Zhang、Y. Liu 和 S. Ma[8]，“带评论级解释的神经注意力评分回归”，载于 *2018* 年万维网大会论文集，*WWW 2018*，法国里昂，*2018* 年 *4* 月 *23–27* 日，P. Champin、F. Gandon、M. Lalmas 和 P. G. Ipeirotis 编。ACM，2018 年，第 1583–1592 页。 [在线]。提供：https://doi.org/ 10 1145 3178876 3186070 。 / 。P. Wang、R. Cai 和 H. Wang[9]，“用于推荐的基于图的抽取式解释器”，载于 *WWW '22*：*ACM Web 2022* 会议，虚拟活动，法国里昂，*2022* 年 *4* 月 *25–29* 日，F. Laforest、R. Troncy、E. Simperl、D. Agarwal、A. Gionis、I. Herman 和 L. M´edini 编。ACM，2022 年，第 2163–2171 页。 [在线]。提供：https://doi.org/ 10 1145 3485447 3512168 。 / 。C. Wu、F. Wu、J. Liu 和 Y. Huang[10]，“利用三层注意力进行分层用户与物品表示的推荐”，载于*2019* 年计算语言学协会北美分会会议：人类语言技术论文集，*NAACL-HLT 2019*，*2019* 年第 *7* 卷，*2019* 年，第 *2 1* 号，明尼阿波利斯，美国，*6* 月——（长篇与短篇论文）J. Burstein、C. Doran 和 T. Solorio 编。计算语言学协会，2019 年，第 1818 1826 [ ] 10 18653 1 19 1180 – 页。在线提供：[11]https://doi.org/ 。 /v /n - Y. Tay、A. T. Luu 和 S. C. Hui，“用于推荐的多指针协同注意力网络”，载于 第 *24* 届 *ACM SIGKDD* 国际知识发现与数据挖掘会议论文集，*KDD 2018*，英国伦敦，*2018* 年 *8* 月 *19–23* 日，Y. Guo 和 F. Farooq 编。ACM，2018 年，第 2309–2318 页。 [在线]。提供：https://doi.org/10.1145/3219819.3220086

[12] C. Li、C. Quan、L. Peng、Y. Qi、Y. Deng 和 L. Wu，“用于推荐并解释你喜欢与不喜欢内容的胶囊网络”，载于第 42 届国际 ACM SIGIR 信息检索研究与发展会议论文集，SIGIR 2019，法国巴黎，2019 年 7 月 21–25 日，B. Piwowarski、M. Chevalier、 ´E. Gaussier、Y. Maarek、J. Nie 和 F. Scholer 编。ACM，2019 年，第 275–284 页。 [ ] 10 1145 3331184 3331216 在线。提供：[13]https://doi.org/ 。 / 。X. Chen、Y. Zhang 和 Z. Qin，“基于神经注意力模型的动态可解释推荐”，载于第三十三届 AAAI 人工智能会议，AAAI 2019，第三十一届人工智能创新应用会议，IAAI 2019，第九届 AAAI 人工智能教育进展研讨会，EAAI 2019，美国夏威夷檀香山，2019 年 1 月 27 日至 2 月 1 日。AAAI Press，2019 年，第 53–60 页。 [在线]。提供：https://doi.org/ 10 1609 33 01 330153 。 /aaai.v i 。H. Liu、W. Wang、[14]Q. Peng、N. Wu、F. Wu 和 P. Jiao，“通过三层注意力网络实现全面的用户与物品表示”，《ACM 信息系统汇刊》，第 39 卷，第 3 期，第 25:1–25:22 页，2021 年。 [ ] 10 1145 3446341 在线。提供：[15]https://doi.org/ 。 / T. Zhang、C. Sun、Z. Cheng 和 X. Dong，“Aenar：面向方面的可解释神经注意力推荐模型，用于评分预测”，《专家系统与应用》，第 198 卷，文章 116717，2022 年。[16] D. P. Kingma 和 M. Welling，“自编码变分贝叶斯”，载于第 2 届国际学习表征会议，ICLR 2014，加拿大艾伯塔省班夫，2014 年 4 月 14–16 日，会议论文集，Y. Bengio 和 Y. LeCun 编，2014 年。 [在线]。提供：http://arxiv.org/abs/ 1312 6114 。J. McAuley 和 J. Leskovec，“隐因子与隐主题：利用评论文本理解评分维度”，载于第 7 届 ACM 推荐系统会议论文集，丛书：RecSys '13。美国纽约州纽约市：计算机协会，2013 年，第 165 172 [ ] 10 1145 2507157 2507163 – 页。[17]在线提供：https://doi.org/ 。 / 。C. Wang 和 D. M. Blei，“用于推荐科学文章的协同主题建模”，载于第 17 届 ACM SIGKDD 国际知识发现与数据挖掘会议论文集，美国加利福尼亚州圣迭戈，2011 年 8 月 21–24 日，C. Apt´e、J. Ghosh 和 P. Smyth 编。ACM，2011 年，第 448–456 页。 [在线]。提供：https://doi.org/ 10 1145 2020408 2020480 。 / 。D. M. Blei、A. Y. Ng 和 M. I. Jordan[19]，“潜在狄利克雷分配”，《机器学习研究期刊》，第 3 卷，第 993–1022 页，2003 年。 [在线]。提供：http://jmlr.org/papers/v3/blei03a.html[20] H. Fei、Y. Ren、S. Wu、B. Li 和 D. Ji，“作为文档级情感分类先验的潜在目标—观点：一种从细粒度视角出发的变分方法”，载于 WWW '21：2021 年万维网大会，23，2021 年，19 虚拟活动／斯洛文尼亚卢布尔雅那，4 月——J. Leskovec、M. Grobelnik、M. Najork、J. Tang 和 L. Zia 编。ACM／IW3C2，2021 年，第 553–564 页。 [在线]。提供：https://doi.org/ 10 1145 3442381 3449789 。 / 。Q. Truong、A. Salah 和 H. W. Lauw[21]，“用于协同过滤的双边变分自编码器”，载于 WSDM '21，第十四届 ACM 网络搜索与数据挖掘国际会议，虚拟活动，以色列，2021 年 3 月 8–12 日，L. Lewin-Eytan 、 D. Carmel 、 E. Yom-Tov 、 E. Agichtein 和 E. Gabrilovich 编。ACM，2021 年，第 292 300 [ ] 10 1145 3437963 3441759 – 页。[22]在线提供：https://doi.org/ 。 / 。X. Wang、H. Jin、A. Zhang、X. He、T. Xu 和 T. Chua，“解耦图协同过滤”，载于第 43 届国际 ACM SIGIR 信息检索研究与发展会议论文集，SIGIR 2020，虚拟活动，中国，2020 年 7 月 25–30 日，J. Huang、Y. Chang、X. Cheng、J. Kamps、V. Murdock、J. Wen 和 Y. Liu 编。ACM，2020 年，第 1001–1010 页。 [在线]。提供：https://doi.org/ 10 1145 3397271 3401137 。 / 。J. Ma、P. Cui、K. Kuang、X. Wang 和 W. Zhu[23]，“解耦图卷积网络”，载于第 36 届国际机器学习会议，ICML 2019，2019 年 6 月 9–15 日，美国加利福尼亚州长滩，丛书：机器学习研究论文集，K. Chaudhuri 和 R. Salakhutdinov 编，第 97 卷。PMLR，2019 年，第 4212–4221 页。 [在线]。提供：http://proceedings.mlr.press/v97/ma19a.html[24] L. Li、Y. Zhang 和 L. Chen，“用于可解释推荐的个性化提示学习”，CoRR，第 abs/2202.07371 卷，2022 年。 [ ] 2202 07371 在线。提供：[25]https://arxiv.org/abs/ 。J. Tan、S. Xu、Y. Ge、Y. Li、X. Chen 和 Y. Zhang，“反事实可解释推荐”，载于 CIKM '21：第 30 届 ACM 信息与知识管理国际会议，虚拟活动，澳大利亚昆士兰州，2021 年 11 月 1–5 日

，G. Demartini ，G. Zuccon ，J. S. Culpepper ，Z. Huang ，and

H. Tong，编者。ACM，2021年，第1784–1793页。 [在线]。获取地址：https://doi.org/ 10 1145 3459637 3482420 。/。T. Chen、H. Yin、G. Ye、Z. Huang、Y. Wang 和 M. Wang，“试试这个：个性化且可解释的替代推荐”，载于《第43届国际 ACM SIGIR 信息检索研究与发展会议论文集》，SIGIR 2020，线上活动，中国，2020年7月25日至30日，J. Huang、Y. Chang、X. Cheng、J. Kamps、V. Murdock、J. Wen 和 Y. Liu，编者。ACM，2020年，第 891 900 [ ] 10 1145 3397271 3401042 –页。[26] 在线。获取地址：https://doi.org/。/。A. Yang、N. Wang、R. Cai、H. Deng 和 H. Wang，“ 推荐的比较解释 ”，载于《 WWW ’22：ACM Web 2022 会议》，线上活动，法国里昂，2022年4月25日至29日，F. Laforest、R. Troncy、E. Simperl、D. Agarwal、A. Gionis、I. Herman 和 L. M´edini，编者。ACM，2022年，第3113–3123页。 [在线]。获取地址：https://doi.org/ 10 1145 3485447 3512031 。/。S. Pan、D. Li、H. Gu、T. Lu、X. Luo 和 N. Gu，“通过评论合理化实现准确且可解释的推荐”，载于《ACM Web会议论文集2022》，2022年，第3092–3101页。[29] J. M. Tomczak 和 M. Welling，“带有Vamp先验的变分自编码器”，CoRR，第abs/1705.07120卷，2017年。 [在线]。获取地址：http://arxiv.org/abs/1705. 07120[30] J. J. Zhao、Y. Kim、K. Zhang、A. M. Rush 和 Y. LeCun，“用于生成离散结构的对抗正则化自编码器”，CoRR，第abs/1706.04223卷，2017年。 [在线]。获取地址：http://arxiv.org/abs/1706. 04223[31] N. Tishby 和 N. Zaslavsky，“深度学习与信息瓶颈原理”，2015年。 [在线]。获取地址：https://arxiv.org/abs/1503.02406[32]A. A. Alemi、I. Fischer、J. V. Dillon 和 K. Murphy，“深度变分信息瓶颈”，载于第5届国际学习表征会议，ICLR *2017,* *24 26,* 法国土伦，*2017,* 2017 [ ] 4月——会议论文集。OpenReview.net，。在线。获取地址：https://openreview.net/forum?id=HyxQzBceg [33] D. P. Kingma 和 M. Welling，“自编码变分贝叶斯”，2013 [ ] 1312 6114 。在线。获取地址：https://arxiv.org/abs/。P. Diaconis 和 D. Freedman，“寻求理论的十二个德·菲内蒂风格结果”，《Annales de l’I.H.P. Probabilités et statistiques》，第23卷，S2期，第397–423页，1987年。 [在线]。获取地址：http: // www.numdam.org/item/AIHPB 1987 23 S2 397 0/ [35] P. Hall、R. C. L. Wolff 和 Q. Yao，“条件分布函数的估计方法”，《美国统计学会杂志》，第94卷，第445期，第154–163页，1999年。 [在线]。获取地址：http://www.jstor.org/stable/ 2669691 [36] S. Czolbe、O. Krause、I. J. Cox 和 C. Igel，“一种基于Watson感知模型的生成式神经网络损失函数”，载于《神经信息处理系统进展33：神经信息处理系统年会2020，NeurIPS 2020》，2020年12月6日至12日，线上，H. Larochelle、M. Ranzato、R. Hadsell、M. Balcan 和 H. Lin，编者，2020年。 [在线]。获取地址：https://proceedings.neurips.cc/paper/2020/hash/165a59f7cf3b5c4396ba65953d679f17-Abstract.html[37] Z. Wang、A. C. Bovik、H. R. Sheikh 和 E. P. Simoncelli，“图像质量评估：从误差可见性到结构相似性”，《IEEE图像处理汇刊》，第13卷，第4期，第600–612页，2004年。 [在线]。获取地址：https://doi.org/ 10 1109 2003 819861 。/TIP。X. He、K. Deng、X. Wang、Y. Li、Y. Zhang 和 M. Wang，“LightGCN：简化并增强用于推荐的图卷积网络”，载于《第43届国际 ACM SIGIR 信息检索研究与发展会议论文集》，SIGIR 2020，线上活动，中国，2020年7月25日至30日，J. X. Huang、Y. Chang、X. Cheng、J. Kamps、V. Murdock、J. Wen 和 Y. Liu，编者。ACM，2020年，第639–648页。 [在线]。获取地址：https://doi.org/ 10 1145 3397271 3401063 。/。K. Mao、J. Zhu、X. Xiao、B. Lu、Z. Wang 和 X. He，“UltraGCN：用于推荐的图卷积网络极简化”，载于《第30届 ACM 信息与知识管理国际会议论文集》，2021年，第1253–1262页。[40] S. Peng、K. Sugiyama 和 T. Mine，“SVD-GCN：一种简化的推荐图卷积范式”，载于《第31届 ACM 信息与知识管理国际会议论文集》，2022年，第1625–1634页。[41] F. Wang、Z. Zheng、Y. Zhang、Y. Li、K. Yang 和 C. Zhu，“看得更远：面向推荐系统的知识图谱感知深度图卷积网络”，《信息科学》，第647卷，第119465页，2023年。 [在线]。获取地址：https://www.sciencedirect.com/ science/article/pii/S0020025523010502 [42] L. Van der Maaten 和 G. Hinton，“使用 t-SNE 可视化数据。”《机器学习研究期刊》，第 9 卷，第 11 期，2008年。[43] J. Wu、X. Wang、F. Feng、X. He、L. Chen、J. Lian 和 X. Xie，“用于推荐的自监督图学习”，载于第*44*届 *ACM SIGIR* 信息检索研究与发展国际会议论文集，2021年，第 726–735 页。[44] J. J. McAuley、J. Leskovec 和 D. Jurafsky，“从多方面评论中学习态度和属性”，载于第*12*届 *IEEE* 国际数据挖掘会议，*ICDM 2012*，布鲁塞尔，比利时，*2012*年*12*月*10*日至*13*日，M. J. Zaki、A. Siebes、J. X. Yu、B. Goethals、G. I. Webb 和 X. Wu 编。IEEE计算机学会，2012年，第 1020–1025 页。 [ ] 10 1109 2012 110 在线。可访问：https://doi.org/ 。 /ICDM. 。 R. He 和 J. J. McAuley，“起伏：使用单类协同过滤对时尚趋势的视觉演变进行建模”，载于第*25*届万维网国际会议论文集，*WWW 2016*，蒙特利尔，加拿大，*2016*年*4*月*11*日至*15*日，J. Bourdeau、J. Hendler、R. Nkambou、I. Horrocks 和 B. Y. Zhao 编。ACM，2016年，第 507 517 [ ] 10 1145 2872427 2883037 – 页。在线。可访问：https://doi.org/ 。 / 。 L. Zheng、V. Noroozi 和 P. S. Yu，“利用评论对用户和物品进行联合深度建模以实现推荐”，载于第十届 *ACM* 国际网络搜索与数据挖掘会议论文集，*WSDM 2017*，剑桥，英国，*2017*年*2*月*6*日至*10*日，M. de Rijke、M. Shokouhi、A. Tomkins 和 M. Zhang 编。ACM，2017年，第 425 434 [ ] 10 1145 3018661 3018665 – 页。在线。可访问：https://doi.org/ 。 / 。 J. Shuai、K. Zhang、L. Wu、P. Sun、R. Hong、M. Wang 和 Y. Li，“一种评论感知图对比学习推荐框架”，载于第*45*届 *ACM SIGIR* 信息检索研究与发展国际会议论文集，丛书 SIGIR ’22。纽约，美国：计算机协会，2022年，第 1283–1293 页。 [在线]。可访问：https://doi.org/ 10 1145 3477495 3531927 . / 。 I. O. Tolstikhin、O. Bousquet、S. Gelly 和 B. Sch¨olkopf，“Wasserstein自编码器”，*CoRR*，第 abs/1711.01558 卷，2017年。 [在线]。可访问：http://arxiv.org/abs/1711 01558 . H. Fu、C. Li、X. Liu、J. Gao、A. Celikyilmaz 和 L. Carin，“循环退火调度：缓解 KL消失的一种简单方法”，载于*2019*年计算语言学协会北美分会会议：人类语言技术论文集，*NAACL-HLT 2019，7，2019，2 1* 明尼阿波利斯，美国明尼苏达州，*6*月 - 卷（长篇和短篇论文），J. Burstein、C. Doran 和 T. Solorio 编。计算语言学协会，2019年，第 240 250 [ ] 10 18653 1 19 1021 – 页。在线。可访问：https://doi.org/ 。 /v /n - X. He、L. Liao、H. Zhang、L. Nie、X. Hu 和 T.-S. Chua，“神经协同过滤”，载于第*26*届万维网国际会议论文集，丛书 WWW ’17。日内瓦州和共和国，瑞士：国际万维网会议指导委员会，2017年，第 173–182 页。 [在线]。可访问：https://doi.org/ 10 1145 3038912 3052569 . / 。 A. M. Elkahky、Y. Song 和 X. He，“一种用于推荐系统跨领域用户建模的多视图深度学习方法”，载于第*24*届万维网国际会议论文集，*WWW 2015*，佛罗伦萨，意大利，*2015*年*5*月*18*日至*22*日，A. Gangemi、S. Leonardi 和 A. Panconesi 编。ACM，2015年，第 278–288 页。 [ ] 10 1145 2736277 2741667 在线。可访问：https://doi.org/ 。 / 。 X. Ma、C. Zhou 和 E. H. Hovy，“平均绝对误差：变分自编码器的互后验散度正则化”，载于第*7*届国际学习表征会议，*ICLR 2019*，新奥尔良，美国路易斯安那州，*2019*年*5*月*6*日至*9*日。OpenReview.net，2019年。 [在线]。可访问：https://openreview.net/forum?id=Hke4l2AcKQ [53] R. Fong 和 A. Vedaldi，“通过有意义的扰动对黑箱进行可解释性解释”，*CoRR*，第 abs/1704.03296 卷，2017年。 [在线]。可访问：http://arxiv.org/abs/1704.03296

附录 A 超参数设置

我们使用 Adam 进行优化，初始学习率设为 0.001。三个数据集的批大小分别设为 32、64 和 32，训练轮数分别为 10、10 和 15。图中的邻接矩阵为二值矩阵：交互评分高于语料库整体评分均值的用户-物品条目设为 1，否则设为 0。当图卷积网络在验证集上达到最佳召回率时，我们停止训练。簇质心向量由 K-means 得出。图节点特征维度 d 为 64，簇数量K在 $[16, 32, 64, 128, 256, 512]$中搜索。在 CNN 评论编码中，词嵌入和 ID 嵌入均由均匀分布初始化（-0.1,0.1），其维度分别设为 300 和 64。我们使用两个大小分别为 2 和 3、维度为 32 的 CNN 卷积核来编码评论。因此，CNN 编码器输出的评论表示维度为 64（即， 32×2）。为了合并每个用户或物品的多条评论，我们使用注意力机制，通过两个连续的线性变换层 $64 \to 32 \to 1$ 得到注意力权重。在 GIANT 中，潜变量维度设为与簇数量相同，即 K。softmax 温度 τ 在 $[1, 2, 4, 6, 8]$中搜索。编码器和解码器中的非线性激活函数均为 ReLU。我们使用具有 $[K, 64]$ 权重的线性层，将潜变量 z^t 映射到与 ID 特征相同的空间。对于用于评分预测的 f_{cls}，我们首先对用户特征和物品特征逐元素乘法的结果应用 ReLU 函数，然后使用形状为 $[64, 1]$ 的线性层将 ReLU 输出转换为评分标量。

附录 B 可迁移信息瓶颈

定理 假设我们使用独立且经过充分训练的编码器-解码器架构来对不同模态（例如图和文本）的特征建模，那么给定的输入x_n 在不同模态下具有表示 $x_n^t \in X^t$ 和 $x_n^g \in X^g$，并满足约束 $I(X^t; X^g) \geq I_x$，因为来自同一输入的 X^t 和X^g 应当相关。此外，假设 $I(Z^t; X^t) \leq I_c$ 且 $I(Z^g; X^g) \leq I_c$，则有以下性质：

$$I(X^t; Z^g) \leq H(X^t) - H(X^g) + H(X^g|X^t) + I_c \quad (13)$$

证明：我们先以不同顺序对以下三个来源 X^t、 X^g 和 Z^g 两次应用链式法则：

$$H(X^t, X^g, Z^g) = H(X^g|X^t, Z^g) + H(X^t|Z^g) + H(Z^g), \quad (14)$$
$$H(X^t, X^g, Z^g) = H(X^t|X^g, Z^g) + H(X^g|Z^g) + H(Z^g). \quad (15)$$

由于上述两个等式右侧相同，因此有：

$$\begin{aligned} &H(X^g|X^t, Z^g) + H(X^t|Z^g) + H(Z^g) \\ &= H(X^t|X^g, Z^g) + H(X^g|Z^g) + H(Z^g) \end{aligned} \quad (16)$$

可简化为：

$$H(X^t|Z^g) = H(X^g|Z^g) + H(X^t|X^g, Z^g) - H(X^g|X^t, Z^g). \quad (17)$$

由于 Z^g、 X^t 和 X^g 均由同一组样本在不同模态或表示下得出，因此显然并非相互独立。因此，我们有 $H(X^g|X^t, Z^g) \leq H(X^g|X^t)$（因为 X^g 和 Z^g 相互依赖），且 $H(X^t|X^g, Z^g) \geq 0$（因为 X^t 和 X^g 相互依赖）。因此，我们可以用 $H(X^g|X^t)$ 替换 $H(X^g|X^t, Z^g)$，并将 $H(X^t|X^g, Z^g)$ 替换为 0，从而推导出以下不等式：

$$H(X^t|Z^g) \geq H(X^g|Z^g) - H(X^g|X^t) \quad (18)$$

应用条件微分熵的性质，可得 $H(X^g|Z^g) = H(X^g) - I(X^g; Z^g)$，从而上述公式可简化为：

$$H(X^t|Z^g) \geq H(X^g) - I(X^g; Z^g) - H(X^g|X^t) \quad (19)$$

因此，我们有：

$$\begin{aligned} I(X^t; Z^g) &= H(X^t) - H(X^t|Z^g) \\ &\leq H(X^t) - H(X^g) + I(X^g; Z^g) + H(X^g|X^t) \\ &\leq H(X^t) - H(X^g) + H(X^g|X^t) + I_c \end{aligned} \quad (20)$$

□

式（20）表明，当两个模态的分布相同，即 $X^t = X^g$，且 $H(X^g|X^t)$ 的条件熵为 0 时，可迁移信息瓶颈的最低上界为 $I(X^t; Z^g) \leq I_c$。令 $H = H(X^g) - H(X^t) + H(X^t|X^g)$，则式（20）可简化为 $I(X^g; Z^t) \leq H + I_c$，其中 I_c 是两个模态的信息瓶颈上界。回顾基于拉格朗日乘子的目标函数：

$$O_{IB} = I(\bar{X}; Z) - \beta \cdot I(X; Z), \quad (21)$$

其优化目标是：

$$\max_{\theta} I(\bar{X}; Z) \quad \text{s.t.} \quad I(X; Z)) \leq I_c. \quad (22)$$

现在，我们聚焦于文本模态中编码器-解码器框架的优化，并用上标 t 标记变量。若使用从图模态中学习到的潜在分布 Z^g 对 Z^t 施加约束，则目标函数可改写为：

$$\max_{\theta} I(\bar{X}^t; Z^t) \quad \text{s.t.} \quad I(X^t; Z^g) \leq H + I_c. \quad (23)$$

由于 $H = H(X^g) - H(X^t) + H(X^t|X^g)$ 与我们的优化无关，基于拉格朗日乘子的目标函数 O_{IB}^t 可改写为：

$$O_{IB}^t = I(\bar{X}^t; Z^t) - \beta \cdot I(X^t; Z^g), \quad (24)$$

也就是说，无需为文本模态定义先验分布。相反，我们可以使用图模态中潜变量的后验分布作为正则化，引导文本数据的训练。

附录 C 可迁移信息瓶颈的目标函数优化

信息瓶颈： 回顾一下，本文中的信息瓶颈目标函数形式为 $I(Z^t; \bar{X}^t) - \beta \cdot I(X^t; Z^g)$，其中

$$I(Z^t; \bar{X}^t) = \int p(\bar{x}^t, z^t) \cdot \log \frac{p(\bar{x}^t|z^t)}{p(\bar{x}^t)} d\bar{x}^t dz^t, \quad (25)$$

其中 $p(\cdot)$ 是未被观测到的真实分布。根据 [32], 给出的证明，我们考虑解码器中的后验估计 $q(\bar{x}_t|z_t)$。根据吉布斯不等式，我们有：

$$\begin{aligned} I(Z^t; \bar{X}^t) &\geq \int p(\bar{x}^t, z^t) \cdot \log \frac{p(\bar{x}^t|z^t)}{q(\bar{x}^t)} d\bar{x}^t dz^t \\ &= \int p(\bar{x}^t, z^t) \cdot \log p(\bar{x}^t|z^t) d\bar{x}^t dz^t - H(\bar{X}^t). \end{aligned} \quad (26)$$

由于解码结果 $H(X^t)$ 的熵与模型优化无关，因此我们只需考虑基于后验估计的下界。根据 [32], , 通过重参数化技巧，我们得到以下经验近似:

$$\mathbb{E}_{x^t\sim X^t}\mathbb{E}_{\epsilon\sim p(\epsilon)}[\log q(\bar{x}^t|f(x^t,\epsilon)] \tag{27}$$

对于拉格朗日乘子项 $I(X^t,\ Z^g)$，它不同于标准的信息瓶颈方法所基于的 β-变分自编码器，我们首先考虑 $I(X^t,\ Z^t)$ 的信息瓶颈，因为根据 B 节给出的证明，$I(X^t,\ Z^g) \leq H + I(X^t,\ Z^t)$。在此，应用吉布斯不等式可得以下界:

$$I(X^t;Z^t) = \int p(x^t,z^t)\log p(z^t|x^t)dx^t dz^t - \int p(z^t)\log p(z^t)dz^t. \tag{28}$$

由于估计 Z_t 的先验分布可能很困难，根据可迁移信息瓶颈的定义，我们应用吉布斯不等式，得到:

$$I(X^t;Z^g) \leq \int p(z^t|x^t)p(x^t)\log\frac{p(z^t|x^t)}{r(z^g)}dx^t dz^t + H, \tag{29}$$

其中， $r(z^g)$ 是基于图的潜在表示的后验分布，而 H 由两种模态的先验分布决定，且与模型优化无关。结合这两个结果，上述界可通过重参数化技巧 [33] ，利用高斯随机变量 ϵ 进行近似:

$$\begin{aligned} O_{IB} = \mathbb{E}_{x^t\sim X^t}\mathbb{E}_{\epsilon\sim p(\epsilon)}[\underbrace{-\log q(\bar{x}^t|f(x^t,\epsilon)]}_{\text{Recons. term: Q(x^t)}} \\ + \beta\cdot\underbrace{\mathrm{KL}(p(Z^t|x^t,\epsilon)|r(Z^g))}_{\text{KL-div. term: KL(Z^t|Z^g)}}, \end{aligned} \tag{30}$$

进一步讨论重构项 $Q(x^t)$

在第二节中，我们提出使用 Nadaraya-Watson 估计量 [35] 来近似条件概率 $q(\bar{x}^t|f(x^t,\ \epsilon)]$。其思路是将解码结果 $\bar{x}^t$ 插入训练样本空间 $\{x^t\}$，并采用核密度估计方法，根据贝叶斯法则近似 $Q(x^t)$:

$$Q(x^t) = \log\frac{\hat{p}(\bar{x}^t_\epsilon,x^t)}{\hat{p}(x^t)} = \log\frac{\frac{1}{N}\sum_{j=1}^N \kappa(\frac{x^t-x^t_j}{h})\cdot\kappa(\frac{\bar{x}^t_\epsilon-x^t_j}{h})}{\frac{1}{N}\sum_{j=1}^N\kappa(\frac{x^t-x^t_j}{h})}, \tag{31}$$

由于 x^t 的输入特征是固定的，我们只关注通过编码器-解码器架构对重构结果的更新。因此，我们有:

$$Q(x^t) \propto \frac{1}{N}\sum_{j=1}^N \kappa(\frac{x^t-x^t_j}{h})\cdot\kappa(\frac{\bar{x}^t_\epsilon-x^t_j}{h}) \tag{32}$$

为简化上述估计，我们在估计中应用径向基函数核和三角不等式，得到:

$$\begin{aligned} Q(x^t) &\propto \log\frac{1}{N}\sum_{j=1}^N \kappa(\frac{x^t-x^t_j}{h})\cdot\kappa(\frac{\bar{x}^t_\epsilon-x^t_j}{h}) \\ &= \log\frac{1}{N}\sum_{j=1}^N \exp(-||x^t-x^t_j||^2)\cdot\exp(-||\bar{x}^t_\epsilon-x^t_j||^2) \\ &= \log\frac{1}{N}\sum_{j=1}^N \exp(-||x^t-x^t_j||^2 - ||\bar{x}^t_\epsilon-x^t_j||^2) \\ &\leq \log\frac{1}{N}\sum_{j=1}^N \exp(-||\bar{x}^t_\epsilon-x^t||^2) \\ &= \log[\exp(-||\bar{x}^t_\epsilon-x^t||^2)] \end{aligned} \tag{33}$$

因此，如果选择自然对数函数，我们便能通过均方误差优化 $Q(x^t)$:

$$-Q(x^t) \geq -\ln(\exp(-||\bar{x}^t_\epsilon-x^t||^2)) = ||\bar{x}^t_\epsilon-x^t||^2 \tag{34}$$

因此，我们可以将下界以及学习目标 O_{IB} 改写为:

$$O_{IB} = \mathbb{E}_{x^t\sim X^t}\mathbb{E}_{\epsilon\sim p(\epsilon)}[||\bar{x}^t_\epsilon-x^t||^2] + \beta\cdot\mathrm{KL}(p(Z^t|x^t,\epsilon)|r(Z^g)), \tag{35}$$

附录D 附加实验结果

我们还将使用 *StandPrior* 生成的簇进行了比较。结果如表 IX 所示。由于不同簇中的高频词大多重叠，很难在 *StandPrior* 中观察到清晰的主题模式。前几个词都是“啤酒花”、“麦芽”和“良好”。其结果甚至比 *WassersteinVAE* 更差，后者至少能够生成具有多样性的簇，这一点体现在每个簇中各不相同的词语上。

Beer	hop，良好，麦芽，香气，**提示**，倒入，**极佳**，良好，口感，很多
	啤酒花，香气，良好，甜，清淡，麦芽，余味，**酿造**，倒入，**玻璃杯**
	啤酒花、麦芽、清淡、甜、许多、**碳酸化**、良好、充分、口感、倒入
音乐	伟大、最佳、乐队、《爱》、**热门歌曲**、良好、第一、制作、歌词、到来
	《爱》、伟大、最佳、乐队、**声音**、良好、第一、甚至、制作、歌词
	伟大、《爱》、最佳、良好、乐队、歌词、制作、粉丝、第一
Office	使用、工作、良好、活页夹、很棒、将会、**简单**、**不错**、标签、**墨水**
	标签、简单、胶带、颜色、很棒、使用、工作、制作、良好、非常
	使用、标签、产品、胶带、颜色、活页夹、**质量**、非常、**扫描仪**、很棒

表 IX

来自*StandPrior* 、*BeerAdvocate*、*DigitalMusic* 和*OfficeProducts* 数据集的三个随机选取簇中最显著的词语（按出现频率排序）。对于每个数据集，我们突出显示其他两个簇中未出现的排名前三的词语。*StandPrior* 生成的主题所包含的词语大多高度重叠。