Controllable Gradient Item Retrieval

Haonan Wang[12*], Chang Zhou[2], Carl Yang[3], Hongxia Yang[2†], Jingrui He[1†]
[1] University of Illinois Urbana-Champaign, [2]Alibaba Group, [3]Emory University
[1]{haonan3, jingrui}@illinois.edu, [2]{eirczhou.zc, yang.yhx}@alibaba-inc.com, [3]{j.carlyang}@emory.edu

ABSTRACT

In this paper, we identify and study an important problem of gradient item retrieval. We define the problem as retrieving a sequence of items with a gradual change on a certain attribute, given a reference item and a modification text. For example, after a customer saw a white dress, she/he wants to buy a similar one but more floral on it. The extent of "more floral" is subjective, thus prompting one floral dress is hard to satisfy the customer's needs. A better way is to present a sequence of products with increasingly floral attributes based on the white dress, and allow the customer to select the most satisfactory one from the sequence. Existing item retrieval methods mainly focus on whether the target items appear at the top of the retrieved sequence, but ignore the demand for retrieving a sequence of products with gradual change on a certain attribute. To deal with this problem, we propose a weakly-supervised method that can learn a disentangled item representation from user-item interaction data and ground the semantic meaning of attributes to dimensions of the item representation. Our method takes a reference item and a modification as a query. During inference, we start from the reference item and "walk" along the direction of the modification in the item representation space to retrieve a sequence of items in a gradient manner. We demonstrate our proposed method can achieve disentanglement through weak supervision. Besides, we empirically show that an item sequence retrieved by our method is gradually changed on an indicated attribute and, in the item retrieval task, our method outperforms existing approaches on three different datasets.

KEYWORDS

information retrieval; recommendation system; weakly-supervised learning; disentangled representation learning; variational autoencoder

1 INTRODUCTION

Controllable recommendations are essential for enhancing the customer experience in real-world recommendation scenarios. An example is shown in Figure 1: customers are inspired by one white dress and want to purchase a dress with some degree of differences in certain attributes to the white one. In offline shopping, it is easy for the customer to make the salesperson promote a series of products that only differ in certain attributes indicated by the customer in a gradient manner. Then the customer can select the most favorite product from the series of products conveniently. However, it is hard for current recommendation systems to present a sequence of products in a gradient form on a certain attribute based on a reference product. The controllable recommendation as a new type of interaction paradigm can solve the problem. In our work, we define controllable recommendation as a two-stage process. In the first stage, a product will be promoted by the recommendation system along with several modification options for each customer. In the second stage, based on the product and the customer-selected modification, a sequence of products with gradient change on a certain attribute will be retrieved. As this is a new type of interaction with a lot of uncertainty, we need to verify in prototype whether the gradient retrieval is feasible. To make it simple and clean, we keep the discussion of the impact of the customers and the performance of the overall controllable recommendation in the future works. As a first step to approach the controllable recommendation, in this work, we only study the problem of gradient item retrieval with a reference item and a modification as query.

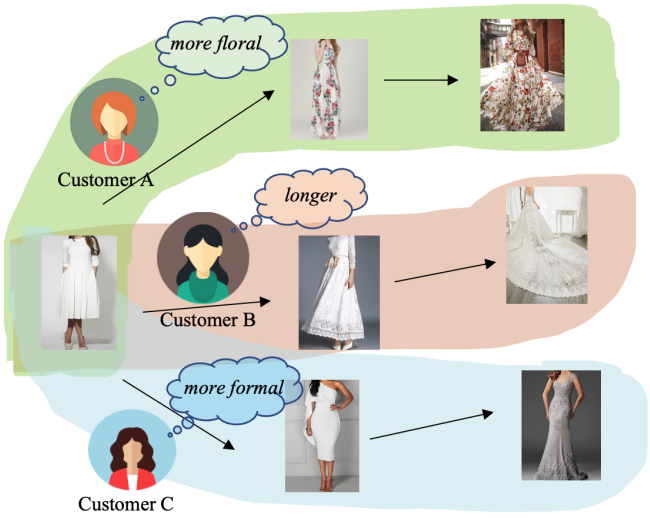

Figure 1: A motivating example of our proposed framework. After browsing one white dress, different users want to purchase a dress with some degree of differences in a certain attribute to the white one.

Current methods usually formulate the second stage as a retrieval problem with a text as a query [33, 38]. Those methods mainly care about whether the target items are retrieved at the top of the retrieved item sequence. Thus, the items in a retrieved item sequence are ranked by the similarities between the input query and

*Work done when he was a research intern at Alibaba Group.
†Co-corresponding authors.

WWW '21, April 19–23, 2021, Ljubljana, Slovenia

ACM ISBN 978-1-4503-8312-7/21/04.
https://doi.org/10.1145/3442381.3449963

items. The demand of retrieving a list of items with gradual change on a certain attribute is largely ignored. The key limitation of these methods is that they only try to model the similarity between the query and target item in their common representation space. In contrast, our method regards a modification text as a "walk" starting from a certain item in the hidden space. By gradually increasing the "step size", a sequence of items can be retrieved in a gradient manner.

Furthermore, we aim to retrieve a sequence of items with gradual change on a certain attribute with weak supervision. Specifically, the goal is to retrieve a sequence of items, where the relevance of a certain attribute is in increasing/decreasing order and other attributes keep the same level. Note the desired attributes (e.g. "floral", "formal") and modification actions ("more" or "less") are indicated by a modification text. To solve the problem, we propose a novel **C**ontrollable **G**radient **I**tem **R**etrieval framework, called *CGIR* , which learns disentangled item representations with semantic meanings. In the training stage, we only need to know whether a certain product has this attribute or not in order to ground the semantic meanings of each attribute to dimensions of the factorized representation space. This type of weak supervision alleviates the burden of obtaining hand-labeled item sequences with gradual change for an attribute. Thanks to the disentanglement property of learned item representations, we can modify the value on dimensions associated with an indicated attribute to form queries without affecting irrelevant attributes. In the inference stage, by using the queries with different modification strength, a sequence of items can be retrieved in a gradient manner.

Unlike previous unsupervised disentanglement methods which have been demonstrated to rely heavily on model inductive bias and require careful supervision-based hyper-parameter tuning [21], in this work, we propose a weakly supervised setting to learn disentangled item representations. Specifically, to achieve disentanglement, our method grounds the semantic meanings of attributes to different dimensions of the factorized representation. Following the previous discussion about disentanglement [28], we decompose disentanglement into two distinct concepts: *consistency* and *restrictiveness*. Specifically, *consistency* means only when the hidden factor of one attribute changes, the attribute will change accordingly; and *restrictiveness* means when one hidden factor changes, irrelevant attributes will keep the same [28]. By enforcing the disentangled factors to match the oracle hidden factors and encoding them into separate dimensions of representation, our proposed method can satisfy the two properties, which allow us to retrieve items with gradual changes along a certain attribute by tuning the value of relevant dimensions.

To summarize, the main contributions of this paper are:

- We identify and define the task of gradient item retrieval.
- For the first time, we propose a weakly-supervised disentanglement framework that can ground semantic meanings to dimensions of a disentangled representation space.
- We demonstrate that our weakly-supervised method can achieve the desired representation disentanglement with semantic meanings, and empirically show that our method can achieve gradient retrieval on both public and industrial datasets.

The rest of this paper is organized as follows. The proposed *CGIR* is introduced in Section 2. Qualitative and quantitative experiments are given in Section 3. Section 4 reviews the related work. Finally, we conclude this work in Section 5.

2 PROPOSED CGIR METHOD

In this section, we first formally define the notation and the gradient item retrieval problem. Then we introduce the proposed framework, followed by discussions about how the proposed method can learn the disentangled item representations with semantic meanings. After that, we show that our weakly-supervised method can achieve disentangled representation with consistency and restrictiveness theoretically.

2.1 Notation and Problem Formulation

Notation In this problem, we are provided with a set of users $\mathcal{U}$, a set of items $\mathcal{I}$, a set of attribute strings $\mathcal{T}$, interaction data $\mathbf{X}$ between users and items, and item-attribute relation data $\mathbf{A}$ between attributes and items. Specifically, the interaction data $\mathbf{X}$ consists of the interactions between N users and M items. An interaction between user u and item i is denoted by $x_{u,i} \in \{0, 1\}$, where $x_{u,i} = 1$ indicates that user u adopts item i, whereas $x_{u,i} = 0$ means there is no recorded interaction between them. For convenience, we use $\mathbf{x}_{u,:}$ to represent the items adopted by user u and $\mathbf{x}_{:,i}$ to denote the users who interacted with item i. The item-attribute relation data $\mathbf{A}$ consists of relations between M items and T attributes, $T = |\mathcal{T}|$. If item i has attribute t, then $a_{i,t} = 1$, otherwise $a_{i,t} = 0$. The attribute vector of item i is denoted as $\mathbf{a}_{i,:}$. Besides, the attribute difference data $\mathbf{Y}$ is composed of attribute difference vector $\mathbf{y}_{i,i'} = \mathbf{a}_{i,:} - \mathbf{a}_{i',:}$, $\mathbf{y}_{i,i'} \in R^T$. Each element of the difference vector $y^t_{i,i'} \in \{-1, 0, 1\}$ indicates the difference between item i and i' on a certain attribute t. Triple data $\mathcal{D}$ is constructed using previously mentioned data and it is composed of $(i, \mathbf{y}_{i,i'}, i')$ triples where i denotes reference item, $\mathbf{y}_{i,i'}$ denotes modification and i' is the desired target item.

Problem Definition We define the gradient item retrieval problem as follows: *based on a reference item and a modification, retrieve a sequence of items in which relevance for a certain desired attribute is in increasing or decreasing order, and relevance for other attributes remains the same.* To make it simple, we consider that a query consists of a reference item and a modification about only one attribute. Note that, if multiple attributes are required to be modified, we can apply the atomic modification several times. Mathematically, we define the query as $(i, \alpha t)$ where i indicates the reference item, $\alpha \in \{1, -1\}$ is the modification action and t is the desired modification attribute. Note that there is a bijection between α and the modification words "more" and "less". For the gradient item retrieval problem, it can be defined as: for a query $(i, \alpha t)$ and its corresponding retrieval sequence *Seq-i*, we want to maximize the probability of the sequence satisfying the constraint: $\alpha \cdot relevance(Seq\text{-}i@k, t) < \alpha \cdot relevance(Seq\text{-}i@k + 1, t)$ and $relevance(Seq\text{-}i@k, t') = relevance(Seq\text{-}i@k + 1, t')$, for any other $t' \in \mathcal{T}, t' \neq t$, where the *relevance* function measures the relevance score between a retrieved item *Seq-i@k* and a certain attribute.

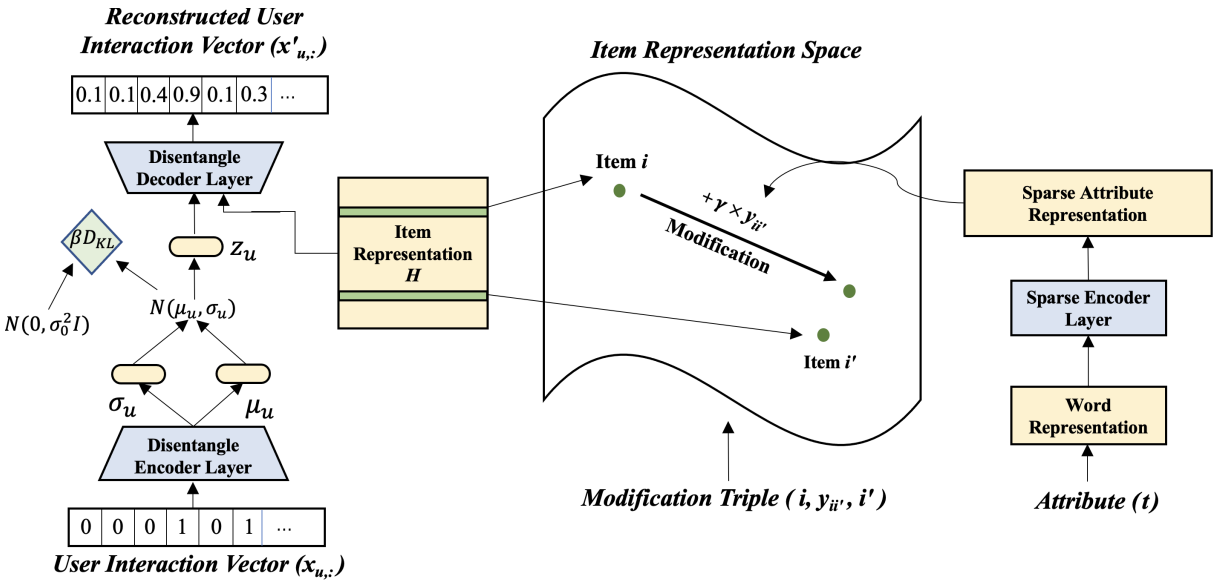

Figure 2: Overview of our proposed *CGIR* framework. It includes three major parts – the left is for disentangled item representation, the right part aims at enforcing representation of attributes to be sparse, and the middle part is for aligning the disentangled item representation space and the sparse word representation space. They are trained in an end-to-end manner.

2.2 Proposed Framework of CGIR

The general framework of the proposed method is shown in Figure 2. It includes three major parts.

The left part is designed based on Variational Autoencoder framework [17], which learns a disentangled item representation from user activities. For each user u, we encode the interaction vector $x_{u,:}$ to the user hidden representation $\mathbf{z}_u \in R^D$. After calculating the interaction probability between user u and all items $\mathbf{H} \in R^{M \times D}$, we reconstruct the interaction vector $x'_{u,:} \in R^M$. The reconstruction loss can be calculated between user interaction vector $x_{u,:}$ and its relevant reconstructed vector $x'_{u,:}$. The disentanglement loss is computed using the mean μ_u and variance σ_u. We keep the dimensionality of μ_u and σ_u the same as $\mathbf{z}_u$.

The right part aims to encode attribute strings to a space where attribute representations are sparse. In that space, each representation of an attribute string has only a few activated dimensions. Our intuition is, for each item, the information of its disentangled representation includes the information of all its attributes. Therefore, each attribute representation should only correspond to some dimensions of the disentangled representation. The input of the sparse encoder model is pre-trained word vectors. We use GloVe [26] as initial features for English words and pre-trained Chinese Word Vectors [20] as initial features for Chinese characters (attribute data of Alishop-attribute dataset is in Chinese). If an attribute only has one word or phrase, we use the relevant sparse word representation as the attribute representation. For attributes including multiple words, a sum pooling is applied over sparse representations of words to obtain the attribute representation.

The middle part is for aligning the disentangled item representation space and the sparse attribute representation space. By leveraging the VAE framework, representations are factorized, where dimensions tend to be independent [22]. However, the meaning of each dimension or the composition of some dimensions remains unclear. The goal of this part is to ground the semantic meanings of attributes to dimensions of factorized item representations.

To achieve the goal, one direct way is to leverage the item-attribute relation data $\mathbf{A}$, which adopted by some existing GAN-based methods [13, 19, 27, 39, 40]. However, those methods ignore the relationship between items, which contradicted with the essence of item retrieval. Instead, we implicitly align the two space by minimizing the distance between the target item i' and the modification result which computed by adding a correct modification $\mathbf{y}_{i,i'}$ on the reference item i. Note, overlapping is allowed between corresponding dimension sets of two attributes, because two attributes may have the same semantic primitives which are separately encoded into different dimensions of item representation. And to keep a linear relationship in the hidden space, we directly add an item representation and an attribute representation without any non-linear transformation. The coefficient γ controls the strength of modification. In the training stage, we set $\gamma = 1$ since we only use the information of whether one item has a certain attribute or not. During the inference stage, in order to retrieve a sequence of items in a gradient manner, we change the strength coefficient γ by increasing a fraction number at each step and keep the top one retrieved item for each step to form the retrieval sequence. The three parts are trained in an end-to-end manner.

2.3 Weakly-Supervised Disentangled Representation Learning with Semantic Meaning

Weakly-Supervised Variational Auto-Encoder. we leverage the VAE framework [17] to enforce item representations to be factorized. And, to involve the information of attribute data, as we stated in the previous section, we model the relation between item pairs

and attributes instead of item-attribute data. Specifically, we model the joint distribution of observed variables $\mathbf{X}$ and $\mathbf{Y}$ by joint distribution $p_\theta(\mathbf{X}, \mathbf{Y})$ where θ denotes parameters of *CGIR* . Our generative model assumes that the observed data are generated from the following distribution:

$$p_\theta(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}) = \iint p_\theta(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}|\mathbf{Z}, \mathbf{H}) p_\theta(\mathbf{Z}, \mathbf{H})\, d\mathbf{Z}\, d\mathbf{H} \tag{1}$$

$\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$ are variables sampled from a distribution parameterized by hidden variables $\mathbf{Z} \in R^{N\times D}$ and $\mathbf{H} \in R^{M\times D}$. The meanings of $\mathbf{Z}$ and $\mathbf{H}$ are described in the previous subsection. As shown in Figure 3, $\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}$ are independent when conditional on $\mathbf{Z}$ and $\mathbf{H}$. Therefore, we have,

$$p_\theta(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}) = \iint p_\theta(\tilde{\mathbf{X}}|\mathbf{Z}, \mathbf{H}) p_\theta(\tilde{\mathbf{Y}}|\mathbf{Z}, \mathbf{H})\, d\mathbf{Z}\, d\mathbf{H} \tag{2}$$

We assume interactions between users and items are independent and identically distributed (*i.i.d.*), and vectors in attribute difference data are also *i.i.d.*. Therefore, for the two terms in equation 2, we have $p_\theta(\tilde{\mathbf{X}}|\mathbf{Z}, \mathbf{H}) = \prod_{u,i} p_\theta(\tilde{x}_{u,i}|\mathbf{z}_u, \mathbf{h}_i)$ and $p_\theta(\tilde{\mathbf{Y}}|\mathbf{Z}, \mathbf{H}) = \prod_{i,i'} p_\theta(\tilde{\mathbf{y}}_{i,i'}|\mathbf{h}_i, \mathbf{h}_{i'})$ separately. Following the paradigm of variational autoencoder (VAE) [5, 25], we introduce a variational dsitribution to alleviate computational burden of integral of equation2 and maximize the lower bound of $\ln p_\theta(\tilde{x}_{u,i}, \tilde{\mathbf{y}}_{i,i'})$ by:

$$\begin{aligned} \ln p_\theta(\tilde{x}_{u,i}, \tilde{\mathbf{y}}_{i,i'}) \geq\ & \mathbb{E}_{q_\theta(\mathbf{z}_u, \mathbf{h}_i|x_{u,i})}\left[\ln p_\theta(\tilde{x}_{u,i}|\mathbf{z}_u, \mathbf{h}_i)\right] \\ & - \mathcal{D}_{KL}\left(q_\theta(\mathbf{z}_u, \mathbf{h}_i|x_{u,i})||p(\mathbf{z}_u, \mathbf{h}_i)\right) \\ & + \mathbb{E}_{q_\theta(\mathbf{z}_u, \mathbf{h}_i|x_{u,i}), q_\theta(\mathbf{z}_u, \mathbf{h}_{i'}|x_{u,i'})}\left[\ln p_\theta(\tilde{\mathbf{y}}_{i,i'}|\mathbf{h}_i, \mathbf{h}_{i'})\right]. \end{aligned} \tag{3}$$

The expectation $\mathbb{E}_{q_\theta(\mathbf{z}_u, \mathbf{h}_i|x_{u,i})}[\cdot]$ is still intractable. As shown in figure 4, we have $\mathbf{z}_u \perp \mathbf{h}_i|x_{u,i}$, according to the Common cause decomposition of graphical models[3]. Therefore, we have the following decomposition:

$$q_\theta(\mathbf{z}_u, \mathbf{h}_i, |x_{u,i}) = q_\theta(\mathbf{z}_u, |x_{u,i}) q_\theta(\mathbf{h}_i, |x_{u,i}). \tag{4}$$

Instead of computing $\mathbb{E}_{q_\theta(\mathbf{z}_u, \mathbf{h}_i|x_{u,i})}[\cdot]$ directly, we use the Gaussian re-parameterization trick[17] to solve $\mathbb{E}_{q_\theta(\mathbf{z}_u, |x_{u,i}) q_\theta(\mathbf{h}_i, |x_{u,i})}[\cdot]$.

Factorization via Regularization. A natural strategy to encourage factorization is to force statistical independence between dimensions. As demonstrate in the previous work [12], if the prior satisfies factorization, penalizing the Kullback–Leibler term of equation 3 would encourage independence between the dimensions. In here, we choose two standard multivariate normal distributions as priors for $\mathbf{z}_u$ and $\mathbf{h}_i$. For the Kullback–Leibler divergence part of equation 3, we can decompose it as:

$$\begin{aligned} & \mathcal{D}_{KL}\left(q_\theta(\mathbf{z}_u, \mathbf{h}_i|x_{u,i})||p(\mathbf{z}_u, \mathbf{h}_i)\right) \\ & = \mathcal{D}_{KL}\left(q_\theta(\mathbf{z}_u|x_{u,i}) q_\theta(\mathbf{h}_i|x_{u,i})||p(\mathbf{z}_u) p(\mathbf{h}_i)\right) \\ & = \mathcal{D}_{KL}\left(q_\theta(\mathbf{z}_u|x_{u,i})||p(\mathbf{z}_u)\right) + \mathcal{D}_{KL}\left(q_\theta(\mathbf{h}_i|x_{u,i})||p(\mathbf{h}_i)\right) \end{aligned} \tag{5}$$

The two KL terms in equation 5 aim at enforcing factorization of user and item representations separately. Due to the time-efficient requirement of recommendation system, we keep a representation table for items, instead of inferring them from interaction matrix at each time. Therefore, we only keep the first term of equation 5 in the final objective. Although this simplification has been used in the previous work[25], we also empirically show that this simplification can enforce item representations to be factorized in our experiments.

Besides, We follow β-VAE[12] to strengthen the KL divergence by a factor of β.

Geometric Relationship of Item Representation. As shown in the middle part of Figure 2, to implicitly align item space and attribute space, we leverage the geometric relationship between items. For a reference-target item pair, their distance will be minimized when a correct modification is added on the reference item. Based on the intuition, we define the third term of equation 3 as:

$$p_\theta(\tilde{\mathbf{y}}_{i,i'}|\mathbf{h}_i, \mathbf{h}_{i'}) = \frac{q_\theta(\mathbf{h}_{i'}|\mathbf{x}_{:,i'})\left(q_\theta(\mathbf{h}_i|\mathbf{x}_{:,i}) + \gamma \cdot \sum_{t\in\mathcal{T}} \tilde{y}^t_{i,i'} \cdot F_\theta(t)\right)}{\sum_{j'\in[1,M]} q_\theta(\mathbf{h}_{i'}|\mathbf{x}_{:,i'})\left(q_\theta(\mathbf{h}_i|\mathbf{x}_{:,i}) + \gamma \cdot \sum_{t\in\mathcal{T}} \tilde{y}^t_{i,j'} \cdot F_\theta(t)\right)} \tag{6}$$

In whole, $\gamma \cdot \sum_{t\in\mathcal{T}} \tilde{y}^t_{i,i'} \cdot F_\theta(t)$ represents the modification $\mathbf{y}_{i,i'}$ scaled by a factor γ. During training stage, we set the modification strengthen coefficient γ equals one. And during inference, γ will be gradually changed to retrieve item in gradient manner. The $\tilde{y}^t_{i,i'}$ indicates the modification direction for attribute t, $F_\theta(\cdot) : R^K \rightarrow R^D$ is the sparse attribute encoder which encode the attribute t to a sparse representation. The equation 6 represents the probability of one triple $(i, \mathbf{y}_{i,i'}, i')$ in $\mathcal{D}$. To align two representation spaces, we maximize the equation 6.

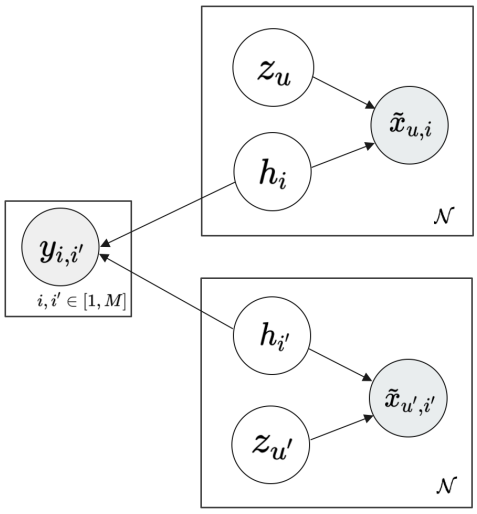

Figure 3: The decoder model, $p(\mathbf{X}, \mathbf{Y}|\mathbf{Z}, \mathbf{H})$.

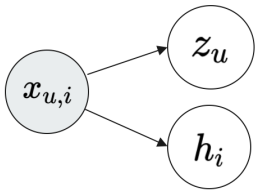

Figure 4: The encoder model, $p(\mathbf{Z}, \mathbf{H}|\mathbf{X})$.

Sparse Attribute Representation Following our intuition that one item's attribute has less information than the whole item and should only be grounded to part of disentangled item representations, we enforce the attribute representation to be sparse before the alignment of attribute and item representation. Function $F_\theta(\cdot)$

is an attribute encoder which maps a attribute string to a sparse representation space. Specifically,

$$F_\theta(t) = \sum_{w \in \mathcal{W}(t)} f_\theta(w) \tag{7}$$

where $\mathcal{W}(t)$ represents the set of words used in attribute string t. Function $f_\theta(\cdot)$ upscale the word representation to another representation space. To enforce the word representation has only a few activated dimensions a sparse loss (*SL*) is applied:

$$\begin{aligned} SL = \frac{1}{D}\sum_{d=1}^{D}\Big(\max(\frac{1}{|\mathcal{W}|}\sum_{w\in\mathcal{W}} f_\theta^d(w) - \rho, 0)^2 \\ + \frac{1}{|\mathcal{T}|}\sum_{w\in\mathcal{W}} f_\theta^d(w) \times (1 - f_\theta^d(w))\Big). \end{aligned} \tag{8}$$

The first term is an Average Sparsity Loss (ASL) which penalizes any deviation of the observed average activation value $f_\theta^d(w)$ from the desired average activation value ρ which is usually set to a small value. The second term is a Partial Sparsity Loss (PSL) that facilitates the value of each dimension of $f_\theta(w)$ to be close to either 0 or 1[29].

Overall Objective Function The above equations bring us to the following training objective. Parameter θ is optimized by maximizing the objective:

$$\begin{aligned} \mathbb{E}_{p_{data}(\mathrm{X})}\Big[&\mathbb{E}_{q_\theta(\mathbf{z}_u,\mathbf{h}_i|x_{u,i})}\big[\ln p_\theta(\tilde{x}_{u,i}|\mathbf{z}_u,\mathbf{h}_i)\big] \\ &- \mathcal{D}_{KL}\big(q_\theta(\mathbf{z}_u|x_{u,i})||p(\mathbf{z}_u)\big) \\ &+ \mathbb{E}_{q_\theta(\mathbf{z}_u,\mathbf{h}_i|x_{u,i}),q_\theta(\mathbf{z}_u,\mathbf{h}_{i'}|x_{u,i'})}\big[\ln p_\theta(\tilde{\mathbf{y}}_{i,i'}|\mathbf{h}_i,\mathbf{h}_{i'})\big]\Big] \\ &- \frac{1}{D}\sum_{d=1}^{D}\Big(\max(\frac{1}{|\mathcal{W}|}\textstyle\sum_{w\in\mathcal{W}} f_\theta^d(w) - \rho, 0)^2 \\ &+ \frac{1}{|\mathcal{T}|}\textstyle\sum_{w\in\mathcal{W}} f_\theta^d(w) \times (1 - f_\theta^d(w))\Big). \end{aligned} \tag{9}$$

2.4 Disentanglement with Guarantee

As demonstrated by Locatello et al.[21], VAE-based unsupervised learning methods fundamentally cannot achieve disentanglement without model inductive biases. Therefore, a natural question is can our method deliver a disentanglement without the help of model inductive bias? Shu et al.[28] gives a theoretical analysis which shows disentanglement can be achieved with guarantee under proper weak supervision. Within their analysis framework, three types of weakly supervised settings were considered, which are restricted labeling, matching pairing, and rank pairing. In our case, attributes of items are considered as hidden factors. For item i and item i', we construct the attribute difference vector $\mathbf{y}_{i,i'}$ by comparing them under each attribute t, $y_{i,i'}^t = a_{i,t} - a_{i',t}$. If attribute t belongs to item i but not for item i', then the ranking of i is higher than i' and $y_{i,i'}^t$ equals 1. Therefore, the $(i, \mathbf{y}_{i,i'}, i')$ triple data can be understood as a special type of ranking-pair where the ranking is binarized and semantic meaningful. Then according to the *Weak Supervision Disentanglement Theorem*[28], the disentangled representation learned under three types of weak supervision is distribution-matching an oracle disentangled representation in which the consistency property of hidden factors, considered by weak supervision signal, can be guaranteed. In our setting, we consider the ranking of one attribute between two items at each triple, because of the restriction $\sum_{t\in\mathcal{T}} |y_{i,i'}^t| = 1$. Further, empirically we have $\sum_{i,i'\in[1,N]} |y_{i,i'}^t| > 1, \forall t \in \mathcal{T}$ which means all attributes are considered by the weak supervision signal. Further, the consistency of all attributes can be guaranteed. By the Full Disentanglement Rule[28], the consistency of all factors further implies the restrictiveness property is guaranteed in disentangled representation.

$$\bigwedge_{t\in\mathcal{T}} C(t) \iff \bigwedge_{t\in\mathcal{T}} D(t), \bigwedge_{t\in\mathcal{T}} D(t) \iff \bigwedge_{t\in\mathcal{T}} R(t) \tag{10}$$

where $C(t)$ denotes the consistency of hidden factor t, $R(t)$ denotes restrictiveness of hidden factor t and $D(t)$ denotes the disentanglement of hidden factor t.

3 EXPERIMENTS

We evaluate *CGIR* on real-world datasets with the aim to answer the following research questions (RQs):

RQ 1 Does *CGIR* achieve gradient item retrieval?

RQ 2 Does *CGIR* outperform other competitors in the item retrieval task?

RQ 3 Can *CGIR* achieve factorized item representation?

	AiShop-tag	ML-25M	ML-20M
# of users	465,573	160,775	136,677
# of items	1,02,746	38,715	20,660
# of interaction	4.4M	12.5M	10.0M
# of tags	263	1086	1086
# of tagged items	1,02,746	29,133	13,025
avg. # of tags per item	4.16	12.61	13.46
# of available tags in Y	263	1086	1086

Table 2: Attributes of datasets after preprocessing.

3.1 Experimental Settings

Datasets We experimented with two publicly accessible MovieLens data sets[1] MovieLens-25M and MovieLens-20M, as well as an industrial internal dataset from Alibaba. For both MoviesLens data and Alishop dataset, we regard tags of a movie or an item as its attributes. In this section "tag" and "attribute" refer to the same thing and will be used interchangeably. For the user movie rating data, we follow MacridVAE, in which ratings are binarized by keeping ratings of four or higher and users who have watched at least five movies. For the tag data, we clean the user provided tags and keep those appeared in the official genome tag table. Additionally we collect a dataset, named AliShop-tag, from Alibaba's e-commerce platform Taobao. All items in AliShop-tag has tags as well as titles and images. Every user in this dataset clicks at least ten items. The characteristics of the three datasets are summarized in Table2. Note, to show all attributes can be considered by the modification data Y even we restrict $\sum_{t\in\mathcal{T}} |y_{i,i'}^t| = 1$, we analyze the number of available tags in $\mathcal{T}$, where a tag t is called available if $\sum_{i,i'\in[1,N]} |y_{i,i'}^t| > 1$.

Query Construction Queries are created as following: pairs of

[1]The MovieLens data set: https://grouplens.org/datasets/movielens/

	ALiShop-tag				ML-25M				ML-20M			
Models	**Hit@20**	**Hit@50**	**MRR**	**MGS**	**Hit@20**	**Hit@50**	**MRR**	**MGS**	**Hit@20**	**Hit@50**	**MRR**	**MGS**
CBIR	0.0211	0.0409	0.01601	0.1811	0.2651	0.3328	0.2116	0.1629	0.3127	0.4638	0.2481	0.1494
DSCMR	0.0239	0.0591	0.01769	0.1934	0.2974	0.3471	0.2292	0.1683	0.3271	0.4810	0.2622	0.1543
TIRG	0.0581	0.0831	0.02418	0.2364	0.4328	0.4801	0.3094	0.1903	0.4733	0.5497	0.3286	0.2085
CGIR	**0.0626**	**0.1019**	**0.02638**	**0.2796**	**0.4412**	**0.4891**	**0.3164**	**0.2588**	**0.4986**	**0.5572**	**0.3374**	**0.2359**
CGIR w/o VAE	0.0572	0.0810	0.02421	0.2314	0.4286	0.4731	0.2981	0.2094	0.4729	0.5334	0.3196	0.1938
CGIR w/o Sparse	0.0628	0.1021	0.02641	0.2607	0.4462	0.4905	0.3188	0.2361	0.4990	0.5578	0.3375	0.2162

Table 1: Gradient Item Retrieval Performance Evaluation on three datasets.

products that have one attribute difference in their descriptions are selected as the query item and target item pairs; and the modification query is composed by a modification action word ("more" or "less") and the different attribute, e.g. more floral. By this way, triple data is constructed, where the head is a reference item, the tail is a target item and the middle is the modification query. As conventional practice, we hold 20% of triple data for testing and 80% for training. The constructed data is close to what will be used in the real-world scenario, where possible modifications will be made offline for each item and then be prompted to a customer who browsed the item just now.

Baselines. We introduce a set of baselines in our experiments:
1. Content-based item Retrieval(CBIR): We train a fully connected network to predict a matching score between a query (a reference item an a modification) and items. We embed item representation from its interaction history with all users. Besides, we encode the query as a concatenation of the representation of a reference item and the modification text.
2. Text Image Residual Gating [33] (TIRG): We adapted this method for item retrieval. TIRG encodes the interaction between one item and users to the item representation. The method aims to map the representation of item and representation of text into the same space and combine them through residual connection. Then we estimate the matching score between the target item and the combination between reference item and modification.
*3. Deep Supervised Cross-modal Retrieval [38] (DSCMR):*We adapt this cross-modal text-image retrieval method to our setting. As previously, we use the interaction history of each item as an input feature, and tag texts and text input. This method tries to find a common representation space, in which the samples from different modalities can be compared directly.
We also introduce two variants for ablation study to analyze the impact of different components of *CGIR* to the performance.
1. CGIR w/o VAE: In this method, instead of using VAE as item representation encoder, we use interaction history as input and a fully connected network as an encoder to encode the interaction history of an item as its representation. This part is same as those baselines. For the remaining, we keep it same as original *CGIR*.
2. CGIR w/o Sparse: We drop the partial sparsity loss and average sparsity loss as shown in 8. For the other parts, we keep them same as *CGIR*.

Evaluation Metrics. We use the following metrics to evaluate the performance of our proposed model. We use two commonly used evaluation criteria in our experiments to evaluate the performance of item retrieval. **Hit Rate** at K (HR@K) computed as the percentage of test queries where target item is within the top K retrieved items. **Mean Reciprocal Rank** (MRR) measures the mean of reciprocal rank of target item in the retrieved list. Besides, to qualitatively measure the gradient effect of retrieval result, we design a new metric, named **Mean Gradient Score** (MGS), to evaluate the degree of gradient for retrieved item list. We use the following equation to define the Mean Gradient Score:

$$MGS = \frac{1}{|test|} \sum_{(i,\alpha t)\in test} \Big(Consistency\_Score(Seq\text{-}i, t) \cdot \big(1 - \frac{1}{|\mathcal{T}|} \sum_{\substack{t'\in\mathcal{T} \\ t'\neq t}} Restrictiveness\_Score(Seq\text{-}i, t')\big)\Big) \tag{11}$$

Here, *test* is a set of testing pairs. Each testing pair includes an item i and an desired modification αt. *seq-i* is an item sequence retrieved by increasing the strength coefficient γ by 0.1 in each step. The first term in equation 11 is the consistency score of the retrieved item sequence. It measures whether the relevant score of items in sequence *Seq-i* with respect to a certain attribute changes gradually. The second term is the restrictiveness score of the retrieved item sequence. It measures whether the modification on one attribute will influence the relevance between other attributes and items. We define the *Consistency_Score* and *Restrictiveness_Score* as the following:

$$Consistency\_Score(seq\text{-}i, t) = \Big(\frac{1}{N-1}\sum_{k=1}^{K-1} \mathbb{1}\big[\alpha \cdot Relevance(seq\text{-}i@k, t) < \alpha \cdot Relevance(seq\text{-}i@k+1, t)\big]\Big) \tag{12}$$

$$Restrictiveness\_Score(seq\text{-}i, t) = 1 - \Big(\frac{1}{N-1}\sum_{k=1}^{K-1} f\big(\alpha \cdot Relevance(seq\text{-}i@k, t) < \alpha \cdot Relevance(seq\text{-}i@k+1, t)\big)\Big) \tag{13}$$

where N is the length of the retrieved sequence *seq-i*, *seq-i@k* is the k-th item of the sequence. Specifically, *seq-i@k* is the top one item retrieved by the combination of the reference item representation and the modification with scaling coefficient $\gamma = 0.1 \times n$, the retrieved sequence *seq-i* is formed by increasing the coefficient. $Relevance(i, t)$ is to calculate the relevance score between item i and tag t. For "add/more" modification on certain tag, we expect the next item in the gradient sequence to have a higher relevance score regarding tag g. For "remove/less" modification, we expect a decrease in relevance score. Function $\mathbb{1}[\cdot]$ is an indicator function which map True and False to 1 and 0. And function $f(\cdot)$ map True and False to 1 and -1. For Restrictiveness Score, if the relevance between indicated tag and items of retrieved sequence in a random walk manner, it will converge to 1 as the length of sequence go to infinite. Note, for MovieLens dataset, a ground-truth relevance

	ALiShop-tag				ML-25M				ML-20M			
Models	**MGS**	**MGS-C**	**MGS-R**	**Ind.**	**MGS**	**MGS-C**	**MGS-R**	**Ind.**	**MGS**	**MGS-C**	**MGS-R**	**Ind.**
CBIR	0.1811	0.2765	0.7638	0.7627	0.1629	0.2481	0.7904	0.6944	0.1494	0.2575	0.7619	0.6791
DSCMR	0.1934	0.2919	0.7311	0.7398	0.1683	0.2634	0.7819	0.6819	0.1543	0.2763	0.7534	0.6637
TIRG	0.2364	0.3566	0.7193	0.7341	0.1903	0.2899	0.7403	0.6563	0.2085	0.3059	0.7264	0.6440
CGIR	**0.2796**	**0.3874**	**0.8329**	**0.9834**	**0.2588**	**0.3371**	**0.8961**	**0.9563**	**0.2359**	**0.3516**	**0.8674**	**0.9521**
CGIR w/o VAE	0.2314	0.3230	0.7893	0.7692	0.2094	0.2917	0.7388	0.6691	0.1938	0.2972	0.7309	0.6529
CGIR w/o Sparse	0.2607	0.3841	0.8114	0.9759	0.2361	0.3358	0.8755	0.9312	0.2162	0.3023	0.8448	0.9446

Table 3: Gradient Effect. To analysis the gradient effect, we provide a more comprehensive analysis using different metrics.

score between a movie and a tag is provided. The *Relevance* function directly output the ground-truth relevance score. However, for Alishop-tag dataset, labeling relevance for each item over every attribute is impossible. Therefore, we adopt a heuristic method. For each modification coefficient γ, instead of measuring the real relevance between top-one retrieved item and a certain attribute, we use the occurrence ratio of items, which has the attribute, on top 100 as the relevance score.

3.2 Gradient Item Retrieval Performance

To answer the first and second research questions, we conduct gradient item retrieval on AliShop-tag, MovieLens-20M and MovieLens-25M. The result is shown in table 1.

Item Retrieval Performance. We observe that our approach outperforms the baselines significantly. This is likely because the user interaction is noisy. Directly using interactions as fingerprint for items will include those noise. However, our method use a VAE [5] framework to extract information from user-item interaction. Interpreting from the information bottleneck view [31], the disentanglement loss enforces our model to forget those noisy part of data and compress those useful information. Therefore, the noisy user-item interaction is denoised by our method, which gives a high-quality item representation. We also notice that both our method and baselines have a drop on the AliShop-tag dataset. The main reason is likey because the industrial E-commercial dataset is more noisy which will influence the quality of item representation and the item set is larger which directly influences the evaluation metrics because we fix the number N in our experiment. Moreover, we observed that *CGIR* and TIRG outperfrom from CBIR and DSCMR by a significant margin. The improvement is likely because both our method and TIRG use a ranking loss, whereas CBIR and DSCMR use a matching loss.

Gradient Retrieval Performance. To measure the gradient retrieval performance, we apply a modification on a source item representation by increasing/decreasing the strength coefficient γ by 0.1 at each time. By analyzing the retrieved item sequences, we calculate the mean gradient score. We outperform all the other baselines methods on MGS. On AliShop-tag data, we achieve better mean gradient score. This is likely because the AliShop-tag dataset has larger number of items which can have a better coverage in the item representation space. During inference stage, less irrelevant items will be retrieved.

Ablation Study for Gradient Item Retrieval. We observe that without using VAE for disentangled item representation, there is a drop on both item retrieval performance and gradient retrieval performance and the impact on gradient retrieval performance is more serious. One reason is attribute-relevant information will appear at each dimension of distributional item representations. When a scaled modification is applied, more than one attributes' information will be changed. Another reason as we discuss previously, the VAE structure delivers a denoised item representation. Besides, we observe that without using sparse loss the item retrieval performance is competitive with the best *CGIR* , but the gradient item retrieval performance is affected obviously. This result is in the line with our expectation. The sparse loss will compress the semantic meaning of an attribute representation to several dimensions which avoid a modification on irrelevant attributes, however some information will be lost in the meanwhile.

3.3 Gradient Effect Study

In order to analyze the effect of disentanglement and answer the third research question. We provide two more detailed experiments. In the first one, we measure the consistency and restrictiveness. In the second one, we analyze the relation between independence level and mean gradient score(MGS).

Consistency and restrictiveness To validate the effect of disentangled representation, we measure the consistency and restrictiveness separately. More specifically, we calculate the mean value of restrictiveness score (equation 13) and consistency score (equation 12). Additionally, We quantify the level of independence by calculating the Uncorrelatedness of item representations [25]. We define Uncorrelatedness as:

$$Ind\_level(\mathbf{Z}) = 1 - \frac{1}{D(D-1)} \sum_{\substack{d_i, d_j \in [1,D] \\ d_i \neq d_j}} \left| CorrCoef(\mathbf{Z}_{:,d_i}, \mathbf{Z}_{:,d_j}) \right|. \tag{14}$$

The function $CorrCoef()$ measures the correlation coefficient between two variables. As shown in table 3, we denote restrictiveness score as **MGS-R**, mean consistency score as **MGS-C** and independence level as **Ind.**.

We observed that the independence level outperform all other methods, which indicates that our method can achieve factorized item representation. This directly answer the research question 3. We also observed that *CGIR* outperform other competitor on **MGS-R**. This improvement shows that *CGIR* has less influence on irrelevant hidden factor, when one factor was changed. This main reason is likely because by applying the disentangled loss, the item representation is factorized, so different hidden factors are encoded into different dimensions of the item representation, which allows us to only modify the value a few dimensions during inference. The performance on metric independence level also supports this explanation. Besides, we notice that although both VAE structure and sparse loss can impact the consistency and restrictiveness, the VAE is more important for important for disentangled representation.

Independence Level and Mean Gradient Score In order to analyze the relation between Mean Gradient Score and Independence Level achieved by our disentangled representation. We vary the hyper-parameters related with disentanglement (β and ρ for our method), and plot Figure 5 the relationship between the level of independence and Mean Gradient Score. We use all item representations on all three datasets to calculate the level of independence. By improving independence of item representations, we achieve a better result on gradient retrieval. This suggests that disentanglement loss can help improve the gradient item retrieval result.

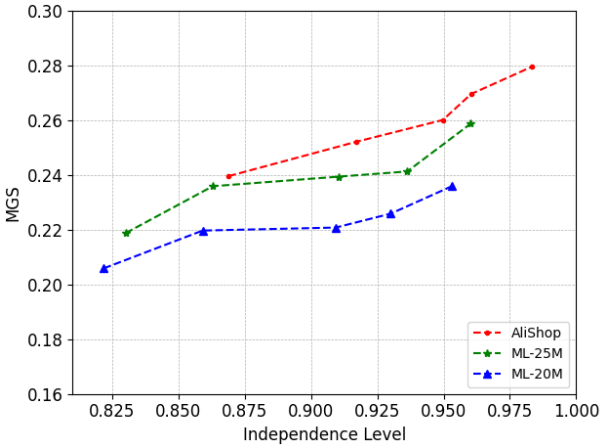

Figure 5: Independence Level vs. Mean Gradient Score

Case Study To illustrate the gradient effect achieved by our method, we visualize several cases as shown in Figure 6 and 7. For MovieLens datasets, because the ground truth relevance score between moives and movie tags are given, we show the relevance score under the poster of each movie. We retrieved those movies by changing the value of modification strength coefficient γ from 0.2 to 1.0 increasing 0.2 at each step. We only keep the top 1 movie into the gradient item retrieval list for each γ. We visualize "more" and "less" modification results in Figure 6a and Figure 6b, respectively. For Alishop-tag dataset, we show the top 4 items retrieved by different modification strength coefficients. This is because we do not have ground-truth relevance score between tags and products. A heuristic way to measure and show the relevance is to use the number of desired products appeared in top@K as the relevance score. As shown in Figures 7a and 7b, the number of desired items in top 4 of each retrieved list is increasing.

(a) Apply modification "more romantic" on movie Lions For Lambs (2007)

(b) Apply modification "less action" on movie Alita:Battle Angel(2019)

Figure 6: Apply modification on movie data to retrieve a sequence of movies in gradient manner.

(a) Apply modification "more sport" on a sneaker

(b) Apply modification "more thick" on a light shirt

Figure 7: Apply modification on Alishop-tag data, a sequence of items are retrieved after changing the value of γ each time

4 RELATED WORK

Product Search and Item Retrieval There are a lot of research has been done on product searches by incorporating text into the query, such as [11, 15] use the user's feedback to the search query. For the problem of image-based product search Vo et al. [33] proposed a method that regards an image and a text string as a query and allows attribute modification. Besides, for image-based fashion search, Zhao et al. [36] developed a memory-augmented deep learning system that can perform attribute manipulation based on the reference image. Moreover, there are a lot of cross-modal methods that deal with the item retrieval problem [34, 35, 38]. Cross-modal methods try to encode information from different modalities into a common representation space. To learn a high-quality common representation space, Zhen et al. [38] leveraged data pairs to match them in representation space. To deal with the unparalleled data scenario, [34] proposed an adversarial learning method to deal with it. We are approaching the item retrieval problem where image data are not available. Besides, unlike previous work which seldom shows its effectiveness of gradient retrieval, in this work, we also focus on how to retrieval items in a gradient manner with respect to certain attributes indicated by a modification.

Disentangled representation learning Disentanglement is an open problem in the realm of representation learning which aims to identify and disentangle the underlying explanatory factors [2].

There are a lot of works that focus on unsupervised disentanglement [7, 12, 16, 18, 25, 37]. β-VAE [12] demonstrates that disentanglement can emerge once the KL divergence term in the VAE [17] objective is aggressively penalized. Later, Zhao et al. [37] proposed InfoVAE which regarded VAE from the view of information theory. By maximizing the mutual information between the data variables and latent variables, the mutual information between the latent variables is minimized. However, Locatello et al. [22] theoretically demonstrate that unsupervised learning of disentanglement arises from model inductive bias and empirically shows that many existing methods for the unsupervised learning of disentangled representations are brittle, requiring careful supervision-based hyper-parameter tuning. Therefore, recently, the research attention has turned to forms of disentanglement in supervised or weakly supervised setting [5, 6, 8, 9]. To model pairwise similarities between data samples, Chen et al.[5] proposed a pairwise VAE that tries to capture a binary relationship(similar or not). And Feng et al.[8] proposed a Dual Swap Disentangling method to leverage binary similarity labels. Besides, a theoretical framework was given by Shu et al. [28], which guarantees consistency and restrictiveness can be achieved under three types of weakly supervised setting. Different from [5, 8], in this work, we focus on using ranking triples information as supervision. In the ranking triples, not only the ranking relation between two data samples were given, but we also provided the information about we compare the two data samples in which point of view. Besides, our method aims to ground the semantic meaning of the comparison view into the dimensions of disentangled representations.

Critiquing Recommender Systems Critiquing is a method widely used conversational recommendation[30] which supports a task-oriented, multi-turn dialogue with their users to discover the detailed and current preferences of the user[14]. In critiquing approaches, users are presented with a recommendation result during the dialogue and then apply pre-defined critiques on the result[4, 10]. Specifically, in this setting, a user is iteratively provided with an item recommendation and attribute description for that item; a user may either accept the recommendation, or critique the attributes in the item description to generate a new recommendation result[32]. Recently, there are some works introduce the critiquing method into the current deep learning recommendation system to improve the explainability of the system[1] which a system proposes to the user a recommendation with its keyphrases and the user can interact with the explanation and critique phrases. Furthermore, there are some work focus on the latent linear critiquing[23, 24] which built on existing linear embedding recommendation algorithm to co-embed keyphrase attributes and user preference embeddings and modulate the strength of multi-step critiquing feedback. By levering the linear structure of the embeddings, the number of interactions required to find a satisfactory item is reduced. Different from those methods, we think a better way is to provide a user with an item sequence with a gradual change on an indicated attributes in order to allow users to obtain satisfactory items with as few interactions as possible. Besides, in those methods, keyphrase frequency usage data is necessary to learn the strengthen of a critiquing. However, in our method, only attributes data is required.

5 CONCLUSION

In this paper, we identify and study a new problem – gradient item retrieval. It is defined as retrieving a sequence of items with gradual change with respect to a certain attribute indicated by a modification text. To solve this problem, we proposed a novel method Controllable Gradient Item Retrieval *CGIR*. Our method takes a product and a modification text, which indicates what attributes to change and how to change, as a query and retrieves a sequence of items with gradual change on the relevance between the indicated tag and items in the sequence. To achieve the gradient effect, our method learns a disentangled item representation with weak supervision and grounds semantic meanings to dimensions of the representation. We show that our method can achieve consistency and restrictiveness under a previously proposed theoretical framework. Empirically, we demonstrate that our method can retrieve items in a gradient manner; and in item retrieval tasks, our method outperforms existing approaches on three different datasets.

6 ACKNOWLEDGEMENT

This work is supported by National Science Foundation under Award No. IIS-1947203 and IIS-2002540. The views and conclusions are those of the authors and should not be interpreted as representing the official policies of the funding agencies or the government.

REFERENCES

[1] Diego Antognini, Claudiu Musat, and Boi Faltings. 2020. Interacting with Explanations through Critiquing. *arXiv e-prints* (2020), arXiv–2005.

[2] Yoshua Bengio, Aaron C. Courville, and Pascal Vincent. 2013. Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 8 (2013), 1798–1828. https://doi.org/10.1109/TPAMI.2013.50

[3] Wray L. Buntine. 2011. Operations for Learning with Graphical Models. *CoRR* abs/1105.2519 (2011). arXiv:1105.2519 http://arxiv.org/abs/1105.2519

[4] Robin D Burke, Kristian J Hammond, and BC Yound. 1997. The FindMe approach to assisted browsing. *IEEE Expert* 12, 4 (1997), 32–40.

[5] Junxiang Chen and Kayhan Batmanghelich. 2020. Weakly Supervised Disentanglement by Pairwise Similarities. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 3495–3502. https://aaai.org/ojs/index.php/AAAI/article/view/5754

[6] Mickaël Chen, Ludovic Denoyer, and Thierry Artières. 2018. Multi-View Data Generation Without View Supervision. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net. https://openreview.net/forum?id=ryRh0bb0Z

[7] Tian Qi Chen, Xuechen Li, Roger B. Grosse, and David Duvenaud. 2018. Isolating Sources of Disentanglement in Variational Autoencoders. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (Eds.). 2615–2625. https://proceedings.neurips.cc/paper/2018/hash/1ee3dfcd8a0645a25a35977997223d22-Abstract.html

[8] Zunlei Feng, Xinchao Wang, Chenglong Ke, Anxiang Zeng, Dacheng Tao, and Mingli Song. 2018. Dual Swap Disentangling. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (Eds.). 5898–5908. https://proceedings.neurips.cc/paper/2018/hash/fdf1bc5669e8ff5ba45d02fded729feb-Abstract.html

[9] Aviv Gabbay and Yedid Hoshen. 2019. Latent Optimization for Non-adversarial Representation Disentanglement. *CoRR* abs/1906.11796 (2019). arXiv:1906.11796 http://arxiv.org/abs/1906.11796

[10] Kristian Hammond, Robin Burke, Charles Martin, and Steven Lytinen. 1995. FAQ finder: a case-based approach to knowledge navigation. In *Proceedings the 11th Conference on Artificial Intelligence for Applications*. IEEE, 80–86.

[11] Xintong Han, Zuxuan Wu, Phoenix X. Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S. Davis. 2017. Automatic Spatially-Aware Fashion Concept Discovery. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 1472–1480. https://doi.org/10.1109/ICCV.2017.163

[12] Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net. https://openreview.net/forum?id=Sy2fzU9gl

[13] Ali Jahanian, Lucy Chai, and Phillip Isola. 2020. On the "steerability" of generative adversarial networks. arXiv:1907.07171 [cs.CV]

[14] Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2020. A survey on conversational recommender systems. *arXiv preprint arXiv:2004.00646* (2020).

[15] Lu Jiang, Alexander G. Hauptmann, and Guang Xiang. 2012. Leveraging high-level and low-level features for multimedia event detection. In *Proceedings of the 20th ACM Multimedia Conference, MM '12, Nara, Japan, October 29 - November 02, 2012*, Noboru Babaguchi, Kiyoharu Aizawa, John R. Smith, Shin'ichi Satoh, Thomas Plagemann, Xian-Sheng Hua, and Rong Yan (Eds.). ACM, 449–458. https://doi.org/10.1145/2393347.2393412

[16] Hyunjik Kim and Andriy Mnih. 2018. Disentangling by Factorising. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018 (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer G. Dy and Andreas Krause (Eds.). PMLR, 2654–2663. http://proceedings.mlr.press/v80/kim18b.html

[17] Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1312.6114

[18] Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. 2018. Variational Inference of Disentangled Latent Concepts from Unlabeled Observations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net. https://openreview.net/forum?id=H1kG7GZAW

[19] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. 2018. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*. 35–51.

[20] Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018. Analogical Reasoning on Chinese Morphological and Semantic Relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Melbourne, Australia, 138–143. https://doi.org/10.18653/v1/P18-2023

[21] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. 2019. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 4114–4124. http://proceedings.mlr.press/v97/locatello19a.html

[22] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. 2019. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 4114–4124. http://proceedings.mlr.press/v97/locatello19a.html

[23] Kai Luo, Scott Sanner, Ga Wu, Hanze Li, and Hojin Yang. 2020. Latent Linear Critiquing for Conversational Recommender Systems. In *Proceedings of The Web Conference 2020*. 2535–2541.

[24] Kai Luo, Hojin Yang, Ga Wu, and Scott Sanner. 2020. Deep Critiquing for VAE-Based Recommender Systems. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1269–1278.

[25] Jianxin Ma, Chang Zhou, Peng Cui, Hongxia Yang, and Wenwu Zhu. 2019. Learning Disentangled Representations for Recommendation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 5712–5723. https://proceedings.neurips.cc/paper/2019/hash/a2186aa7c086b46ad4e8bf81e2a3a19b-Abstract.html

[26] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1532–1543. https://doi.org/10.3115/v1/D14-1162

[27] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. 2020. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9243–9252.

[28] Rui Shu, Yining Chen, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2020. Weakly Supervised Disentanglement with Guarantees. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. https://openreview.net/forum?id=HJgSwyBKvr

[29] Anant Subramanian, Danish Pruthi, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Eduard H. Hovy. 2018. SPINE: SParse Interpretable Neural Embeddings. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, Sheila A. McIlraith and Kilian Q. Weinberger (Eds.). AAAI Press, 4921–4928. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17433

[30] Cynthia A Thompson, Mehmet H Goker, and Pat Langley. 2004. A personalized system for conversational recommendations. *Journal of Artificial Intelligence Research* 21 (2004), 393–428.

[31] Naftali Tishby and Noga Zaslavsky. 2015. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop, ITW 2015, Jerusalem, Israel, April 26 - May 1, 2015*. IEEE, 1–5. https://doi.org/10.1109/ITW.2015.7133169

[32] Frederich N Tou, Michael D Williams, Richard Fikes, D Austin Henderson Jr, and Thomas W Malone. 1982. RABBIT: An Intelligent Database Assistant.. In *AAAI*. 314–318.

[33] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. 2019. Composing Text and Image for Image Retrieval - an Empirical Odyssey. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 6439–6448. https://doi.org/10.1109/CVPR.2019.00660

[34] Xin Wen, Zhizhong Han, Xinyu Yin, and Yu-Shen Liu. 2019. Adversarial Cross-Modal Retrieval via Learning and Transferring Single-Modal Similarities. *CoRR* abs/1904.08042 (2019). arXiv:1904.08042 http://arxiv.org/abs/1904.08042

[35] Jun Yu and Xiao-Jun Wu. 2019. Unsupervised Concatenation Hashing with Sparse Constraint for Cross-Modal Retrieval. *CoRR* abs/1904.00726 (2019). arXiv:1904.00726 http://arxiv.org/abs/1904.00726

[36] Bo Zhao, Jiashi Feng, Xiao Wu, and Shuicheng Yan. 2017. Memory-Augmented Attribute Manipulation Networks for Interactive Fashion Search. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 6156–6164. https://doi.org/10.1109/CVPR.2017.652

[37] Shengjia Zhao, Jiaming Song, and Stefano Ermon. 2017. InfoVAE: Information Maximizing Variational Autoencoders. *CoRR* abs/1706.02262 (2017). arXiv:1706.02262 http://arxiv.org/abs/1706.02262

[38] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. 2019. Deep Supervised Cross-Modal Retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 10394–10403. https://doi.org/10.1109/CVPR.2019.01064

[39] Yao Zhou, Jianpeng Xu, Jun Wu, Zeinab Taghavi Nasrabadi, Evren Korpeoglu, Kannan Achan, and Jingrui He. 2020. GAN-based Recommendation with Positive-Unlabeled Sampling. arXiv:2012.06901 [cs.IR]

[40] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*.