Information Fusion 99 (2023) 101858

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full length article

Cross-modal contrastive learning for aspect-based recommendation

Heesoo Won [a], Byungkook Oh [b], Hyeongjun Yang [a], Kyong-Ho Lee [a,*]

[a] Department of Computer Science, Yonsei University, 50, Yonsei-ro, Seodaemun-gu, 03722, Seoul, Republic of Korea
[b] Samsung Research, 56, Seongchon-gil, Seocho-gu, 06765, Seoul, Republic of Korea

ARTICLE INFO

Keywords:
Knowledge graph
Graph neural networks
Self-supervised learning
Aspect-based recommendation

ABSTRACT

Knowledge-enhanced recommender systems with aspects have improved recommendation performance by better profiling user preferences. Existing models can be divided into graph-based and text-based depending on the type of external knowledge: knowledge graph and review text. Since each knowledge provides different information from the scope and detail of aspects, it is necessary to integrate them for modeling sophisticated aspect-level preferences. However, it is difficult to directly fuse the aspects defined on two types of knowledge because they are expressed in different latent spaces. To tackle this problem, we explore self-supervised learning on multi-modal data. Specifically, we propose a novel model called **CO**ntrastive learning with cro**S**s-**MO**dal a**S**pects (**COSMOS**). To take the data imbalance between knowledge graph and review texts into consideration, we devise a cross-modal contrastive learning scheme, which generates multiple views of a user or an item based on inter-modal correlation. With the correlation between aspects, COSMOS captures the inherent dependency between graph and text data. The fine-grained aspect-level preference, which contains salient features (from review text) as well as general ones (from knowledge graph), leads to providing high-quality recommendation results, even if a user only has one of the two data. Experimental results on two datasets show that COSMOS outperforms state-of-the-art recommender systems.

1. Introduction

To better recommend personalized items for users, recommender systems have made significant contributions in many fields. Recently, knowledge-enhanced recommender systems have gained wide attention, where users and items share additional information as knowledge. Graph-based recommendations [1–4] improve accuracy and provide explainability by utilizing human knowledge. Specifically, recent works focus on GNN-based methods, which propagate information from multi-hop neighbors in a graph. Also, text-based recommendations [5,6] share word-level information between users and items, which compute context-aware representations that reflect semantic information.

Knowledge-enhanced recommender systems can be analyzed at an aspect level. Aspects enable the fine-grained elaboration of user preference from user–item interactions, where users with similar aspects will have a similar item preference. *Publication date*, *genre*, and *author* (or *artist*, *producer*, and *featured artists*) are examples of aspects in a book (or music) domain. Modeling interactions at the granularity of aspects helps understand why a user selects an item.

Prior efforts on the knowledge-enhanced recommendation with aspects can be categorized into graph and text-based methods. The graph-based method [7] defines each aspect with a distribution over relations of a Knowledge Graph (KG), considering the importance of every combination of relations. Therefore, we can extract the general preference of a user based on item characteristics. Take Fig. 1 as an example, where two users have interacted with the same books. By the interacted items and their connectivity, we can find out that the corresponding *author* and *genre* aspects have an influence on selecting them. However, a graph is insufficient in modeling aspects for the following reasons: (1) Since all the items interacted by users are regarded as preferred items, it is vulnerable to interaction noises; and (2) Aspects of the users who selected the same items are equally treated since it is impossible to reflect fine-grained context information. Existing graph-based model figures out that both users in Fig. 1 prefer the books written by Pratchett. However, as shown in the review texts, the first user emphasizes the *author* with positive emotions, while the other user expresses negative feelings about the same aspect. Instead, the user pays attention to the *character*, which means recommending Pratchett's book cannot satisfy both of them.

Another line of research [8–10] leverages review documents of a user. Users clearly express their opinions about a specific aspect in reviews, which makes extracting salient features possible. As shown in Fig. 1, we can visibly figure out the most important aspect *author*

* Corresponding author.
E-mail addresses: hswon97@yonsei.ac.kr (H. Won), byungkook.oh@samsung.com (B. Oh), edbm95@yonsei.ac.kr (H. Yang), khlee89@yonsei.ac.kr (K.-H. Lee).

https://doi.org/10.1016/j.inffus.2023.101858
Received 24 January 2023; Received in revised form 9 May 2023; Accepted 26 May 2023
Available online 3 June 2023

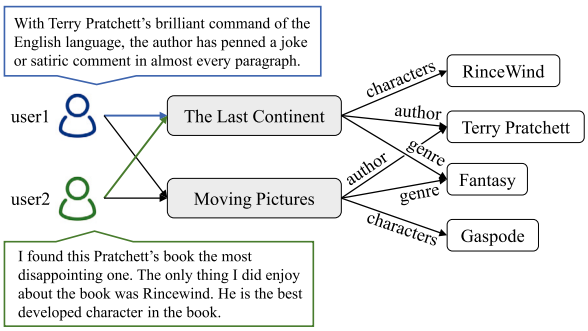

Fig. 1. An example of collaborative knowledge graph and review texts, where different users have interacted with the same items in book domain. The reviews shown are about the book *The Last Continent*.

or *character* from their review texts. Also, user representations with contextual information can be obtained. Despite fine-grained aspect modeling, we argue that text-based methods have fundamental problems. Especially, it cannot be applied to reviews that do not express a user's aspect. Moreover, since existing works assume that every user writes a review when interacting with an item, they cannot handle users without reviews.

In this paper, we aim to complement each limitation by developing a solution that integrates graph and text-based aspect modelings to identify the fine-grained level of aspects. Towards this end, we propose a new solution, named **CO**ntrastive learning with cro**S**s-**MO**dal a**S**pects (**COSMOS**), which models aspects based on the characteristics of the interacted items and also extracts major aspect features from reviews, considering the context. In fact, it is impossible to associate the aspects defined in two types of knowledge because they are represented in different latent spaces. We cannot directly map emotionally-expressed aspects from text and relation-based aspects from a graph, nor can we capture association scores between them. Therefore, to efficiently integrate multi-modal aspect modeling methods, we focus on aspect-level representations of users and items resulted from each information. Specifically, we apply contrastive learning on multi-modal representations, which should indicate the same aspect of a user or an item. As a result, an informative de-biased aspect-level representation can be learned by comparing signals from a KG and review text.

An important fact here is that not all users always write reviews when interacting with items. To address the data imbalance, we model cross-modal contrastive learning by computing two representations based on the correlation between multi-modal data. In the pre-training stage, COSMOS introduces text-based and graph-based encoders for cross-modal aspect-level representations by influencing each other. It learns correlations between latent aspects by comparing two views for each user (or item); After pre-training, we can fine-tune the model in the downstream recommendation task. We first use the pre-trained graph-based encoder to learn aspect-aware user (or item) representations and then predict their matching scores. Due to the aspect correlations, we can output the fine-grained representations even when only one data exists.

The contributions of this work are threefold:

- We propose a novel multi-modal contrastive learning approach for aspect-based recommendation, which integrates graph and text data by extracting informative aspect features for representation learning.
- Unlike typical contrastive models with multi-modal data, which simply align the views generated from each data, we create cross-modal views based on aspect correlations to resolve data imbalance.
- We conduct extensive experiments on two datasets to demonstrate the effectiveness of COSMOS, which considers cross-modal aspects.

2. Related work

2.1. Knowledge-enhanced recommendation

Previous studies on knowledge-enhanced recommender systems can be categorized into graph-based and text-based methods, according to the type of shared knowledge between users and items, which act as context information. Graph-based methods [1–4,11–15] make it easy to understand item connections and user preferences with a KG. Besides, text-based methods [5,6,16,17] reflect semantic information by sharing word-level information. Moreover, some studies [18,19] incorporate multi-modal data into a KG and then provide recommendations based on the reasoning relationship between entities.

A popular line of research focuses on identifying user preferences at a fine-grained level of aspects. The first type relies on a graph for modeling users' aspects. In KGIN [7], it is possible to infer general aspects based on item characteristics, by mapping each aspect with a combination of KG relations. Information from multi-hop entities can also be encoded in the representations. However, it is not only vulnerable to interaction noises but also does not reflect fine-grained context information.

An alternative type of research extracts context-aware word representations from texts, which learn users' latent aspect-level representations together with their opinions or sentiments. ANR [8] models fine-grained aspects from review texts regarding the importance of a user for each aspect. Following ANR, CATN [9] conducts cross-domain recommendations by mapping aspect-level preferences between two domains. Recently, after extracting aspects from reviews, GERA [10] builds a KG of users, items, and aspects and then provides recommendations by making use of graph embedding techniques. Although these studies utilize fine-grained context information from review data, a critical limitation is that all users should write reviews and express significant aspects.

Such limitations may hurt the aspect modeling abilities in each type of method. Towards this end, we propose a model to effectively integrate multi-modal aspect modeling methods to extract features for informative representations.

2.2. Self-supervised learning for recommendation

Self-supervised Learning (SSL) has been widely studied in natural language processing (NLP) [20–22] and computer vision (CV) [23,24]. It aims to train a network with an auxiliary task, where a pre-trained model can easily be adapted to downstream tasks. We can boost the performance of downstream tasks, by using pre-trained embeddings as an initial parameter for further training. A recent technical trend is contrastive modeling, which learns similarities between data instances to extract informative features. [24] applies data augmentation techniques to an anchor image and then performs discrimination between positive and negative pairs to learn generalizable representations.

SSL has also been applied on graph data [25–27]. It captures the structural and semantic properties of a graph by designing a pretext task to provide supervision from graph data itself. Then it can be easily generalized to downstream tasks with a few fine-tuning steps.

To the best of our knowledge, very limited works [28–32] perform SSL with graph for downstream recommendation task. For instance, pre-training GNN with a pretext task that reconstructs cold-start user (or item) embeddings [28] improves the downstream recommendation performance. SGL [29] conducts contrastive learning based on self-discrimination, which constructs supervision signals from the correlation within a given user–item graph. SEPT [30] is proposed to reflect homophily in recommender systems, which obtains signals from social networks. More recently, CML [33] captures multi-behavior patterns of different users from diverse behavior views for a customized recommendation, based on a contrastive self-supervised learning paradigm.

A little further, knowledge-aware SSL methods, which use knowledge graphs for side information, also focus on improving recommendation performance. CKGC [34] fully exploits the descriptive and structural information of a KG, by maximizing the agreement between descriptive attributes and structural connections. In addition, KGCL [35] proposed a KG augmentation schema to cope with the sparsity and noise issue in real-world KG.

However, current methods do not address aspects from multi-modal data. Our work is the first to learn informative de-biased representations by extracting common aspect features from multi-modal data with self-supervised learning.

2.3. Self-supervised contrastive learning on multi-modal data

There are impactful research and development of multi-source information with various modalities, such as vision-language [36], language-graph, and audio-vision [37,38], for visual question answering [39]. Some works [40,41] utilize relationships among multi-modal information of each entity as well as relationships among entities, called a multi-modal knowledge graph.

Traditional models generate joint representations of multi-modal data by merely fusing each vector representation based on concatenation operation, symmetric pooling, and attention mechanism. However, simple multi-modal models could be biased to a particular modality, resulting in an overfitting problem and worsening performance.

More recently, self-supervised contrastive learning has been widely adopted to bridge the semantic gap between two encoders for each multi-modal data, rather than generating joint representations of multiple modalities. Contrastive learning that utilizes multi-modal data [36,42,43] mainly uses image and text encoders to learn informative visual representations and applies them to several downstream tasks (e.g., image classification and object detection). Specifically, the vector representations of multi-modal data are directly compared with their other views (i.e., ones of negative samples) to maximize mutual information based on contrastive objectives such as InfoNCE [23]. By doing so, each pre-trained encoder can incorporate semantic correlations on the feature of other modality, and then can be tailored for a variety of downstream tasks, particularly zero-shot image classification [43].

In the latest research, MMCPR [44] performs self-supervised contrastive learning on multi-modal data by targeting the recommendation task's performance. MMCPR considers two modalities for users and items to fully exploit the multimodality of side information; review texts and user graph for users, and description texts, images, and item graph for items.

Unlike many existing studies, we conducted aspect-based multi-modal contrastive learning on text and graph data. Our work provides recommendations based on self-supervised contrastive learning between language-interaction, in the perspective of an aspect. By deeply associating language (user reviews) and interaction (user–item interaction), the performance of downstream recommendation task can be improved.

3. Problem formalization

We use review documents and a Collaborative Knowledge Graph (CKG) to model the aspects of a user and an item. We use D_u and D_i to denote user and item documents, respectively. D_u is a review set that contains all reviews written by a user, and D_i contains all reviews that an item receives. A CKG consists of a user–item interaction graph and a KG. We represent interaction data as a user–item bipartite graph, which is defined as $\{(u, y_{ui}, i)|u \in \mathcal{U}, i \in \mathcal{I}\}$, where $\mathcal{U}$ denotes user set and $\mathcal{I}$ denotes item set. Each interaction y_{ui} is mapped with a review of a user towards an item. Additionally, a KG is a heterogeneous graph consisting of item entities and relations between them, presented as $\{(h, r, t)|h, t \in \mathcal{E}, r \in \mathcal{R}\}$, where $\mathcal{E}$ is set as entity set and $\mathcal{R}$ as relation set. Each (h, r, t) indicates that a relation r exists between a head entity h and a tail entity t. Table 1 summarizes the key notations used throughout this paper.

Table 1
Notations and their definitions.

Notation	Definition
D_u	User document
$E_{u,a}$	Aspect-specific embedding of user u for aspect a
$L_{u,a}$	Local contextual feature embedding of user u for aspect a
$attn_{u,a}$	Text-based aspect attention score of user u for aspect a
$V_{a,t}$	Text-based aspect embedding vector for aspect a
$T_{u,a}$	Text-based representation of user u for aspect a
$\acute{T}_{u,a}$	Text-based representation with graph of user u for aspect a
T_u	Text-based representation of user u
$\acute{T}_u$	Text-based representation with graph of user u
$\alpha_{r,a}$	Aspect attention score of relation r for aspect a
$\beta_{u,a}$	Graph-based aspect attention score of user u for aspect a
$V_{a,g}$	Graph-based aspect embedding vector for aspect a
$G_{u,a}$	Graph-based representation of user u for aspect a
$\acute{G}_{u,a}$	Graph-based representation with text of user u for aspect a
G_u	Graph-based representation of user u
$\acute{G}_u$	Graph-based representation with text of user u
C	Aspect correlation matrix

Problem Definition. Given review documents and a CKG, our goal is to pre-train the encoding function to apply it to the downstream recommendation task. Specifically, contrastive learning in the pre-training stage based on data correlation will offer a valuable angle to estimate how likely a user would adopt an item in the fine-tuning stage.

4. Contrastive learning with cross-modal aspects

Fig. 2 illustrates the overall architecture in the pre-training stage. In the pre-training stage, we pre-train the COSMOS via contrastive learning. Specifically, the self-supervised task is to obtain supervision from the correlation within the multi-modal input data. We use cross-modal representations from graph and text data, which results in informative aspect-level representations. Moreover, aspects defined in each data can be mapped by learning cross-modal representations, which helps resolve the data imbalance. Since the same procedure is applied to users and items, we will take a user as an example in the following sections.

4.1. Text-based aspect encoding

As illustrated in Fig. 3, review texts are used to learn representations related to aspects defined on graphs, based on the aspect correlations between two data. Following ANR [8], we first generate contextual representations $T_{u,a}$ from review texts in an aspect level, which offer more detailed opinions and experiences of a user u. Then, a cross-modal representation learning layer performs graph-to-text attention to compute text-based representations with graph $\acute{T}_{u,a}$. Thus, the text-based aspect encoding can extract features such as sentiment and opinion that are hard to reveal from a knowledge graph.

4.1.1. Embedding layer

A user document D_u is transformed into a matrix $E_u \in \mathbb{R}^{n \times d}$, where n is the document length and d is the word embedding dimension. Aspect-specific representation for the ith word $E_{u,a}[i]$ is derived by an aspect-specific word projection matrix $W_a \in \mathbb{R}^{d \times h_1}$, where h_1 is the aspect embedding dimension.

4.1.2. Aspect attention layer

Each aspect $a \in \{1, 2, \ldots, A\}$ defined on text is represented as an embedding vector $V_{a,t}$. By concatenating local context window, we obtain $L_{u,a}$ to calculate aspect attention scores:

$$attn_{u,a} = softmax(V_{a,t}(L_{u,a}[i])^{\top}) \quad (1)$$

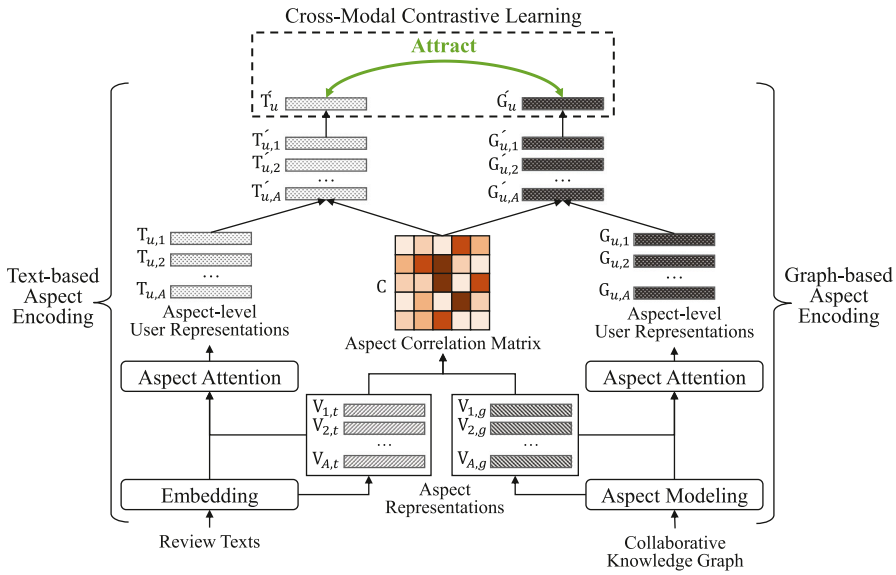

Fig. 2. Overall architecture of the pre-training stage, which consists of three modules: text-based aspect encoding, graph-based aspect encoding and cross-modal contrastive learning. We can effectively integrate aspect-level representations obtained from each data and learn the correlation between aspects. This process proceeds for each user and item.

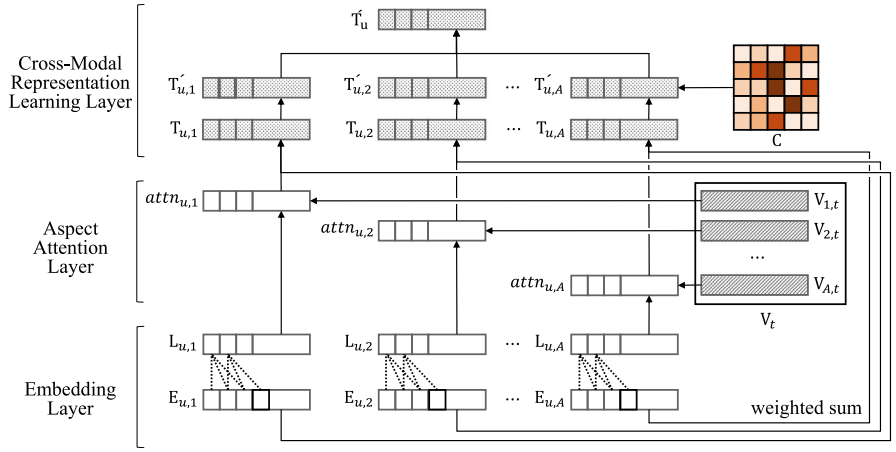

Fig. 3. Architecture of text-based aspect encoding, where reviews are used to learn representations related to aspects defined on graphs.

Here, $attn_{u,a}$ represents the importance of user u for aspect a. The importance of each word towards the target aspect is used to learn user representations:

$$\mathrm{T}_{u,a} = \sum_{i=1}^{n} (attn_{u,a}[i]\mathrm{E}_{u,a}[i]) \tag{2}$$

$\mathrm{T}_{u,a}$ is a text-based aspect-level representation of user u for aspect a. By considering context information, the representation reflects fine-grained aspects with opinions or sentiments.

4.1.3. Cross-modal representation learning layer

To compute text-based representation with graph, we first calculate the global aspect correlation matrix C as follows:

$$\mathrm{C} = LeakyReLU(\mathrm{V}_t^{\top}\mathrm{W}\mathrm{V}_g) \tag{3}$$

where V_t and V_g are global aspect representations in each text and graph data. We adopted the LeakyReLU function to calculate correlation scores to handle sparse aspect correlations across different modalities and extract aspect features better. $\mathrm{C}_{p,q}$ reflects the correlation between pth aspect from text and qth aspect from a graph, where W is a learnable weight matrix. $\mathrm{T}'_{u,q}$, a text-based aspect-level representation with graph for qth aspect, is calculated as follows:

$$\mathrm{T}'_{u,q} = \frac{\sum_{j=1}^{A} \mathrm{C}_{j,q} \times \mathrm{T}_{u,j}}{\sum_{i=1}^{A} \mathrm{C}_{i,q}} \tag{4}$$

We regard $\mathrm{C}_{j,q}$ as an attention weight to aggregate text-based representations $\mathrm{T}_{u,j}$. Therefore, associations between all aspects in text and qth aspect in a graph are reflected in qth text-based representation with graph. Finally, we add all aspect-level representations for a final text-based user representation with graph T'_u.

4.2. Graph-based aspect encoding

Contrary to text-based aspect encoding, graph-based aspect encoding leverages a collaborative knowledge graph to learn graph-based representations related to aspects from review texts. As shown in Fig. 4, we first obtain contextual representations $\mathrm{G}_{u,a}$ based on the domain knowledge of the interacted items. Then, graph-based representations with text $\mathrm{G}'_{u,a}$ are learned in the cross-modal representation learning layer via text-to-graph attention. Thus, the graph-based aspect encoding can better understand user preferences for semantically similar items having similar attributes (e.g., genre, director, actors).

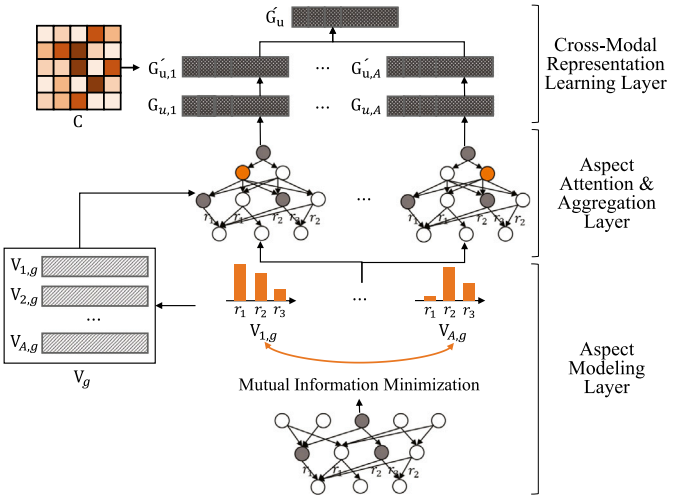

Fig. 4. Architecture of graph-based aspect encoding, where graph is used to learn representations related to aspects from texts.

4.2.1. Aspect modeling layer

Following KGIN [7], each aspect $V_{a,g}$ is assigned with a distribution over the combination of relations in a KG:

$$V_{a,g} = \sum_{r \in R} \alpha_{r,a} e_r \tag{5}$$

where e_r is the relation embedding, and $\alpha_{r,a}$ represents its attention score towards an aspect a, which is calculated by a trainable weight of relation r and aspect a. For independence modeling of aspects, we minimize the mutual information among aspects. By associating each aspect with item–item relations, a user–item relation can be decomposed into A aspects.

4.2.2. Aspect attention and aggregation layer

Graph-based user representation is calculated by aggregating aspect-aware information from a CKG. To explore information by taking high-order connectivity into account, we stack aggregation layers:

$$G^{l}_{u,a} = \frac{1}{|N_u|} \sum_{(a,i) \in N_u} \beta_{u,a} V_{a,g} \odot G^{l-1}_{i,a} \tag{6}$$

where relational paths are modeled to capture neighboring nodes' paths. To learn user representations of each aspect, we consider relatedness between each aspect and interacted items. $\beta_{u,a}$ indicates a user's attention towards the target aspect, and $|N_u|$ represents aspect-aware interaction history.

$$\beta_{u,a} = \frac{\exp(V^{\top}_{a,g}, G^{(l-1)}_{u,a})}{\sum_{a' \in \{1,2,\ldots,A\}} \exp(V^{\top}_{a',g}, G^{(l-1)}_{u,a})} \tag{7}$$

Then representations of user u at l different layers are summed up for a graph-based representation:

$$G_{u,a} = G^{(0)}_{u,a} + G^{(1)}_{u,a} + \cdots + G^{(l)}_{u,a} \tag{8}$$

As a result, graph-based aspect modeling leads to learning general representations based on item characteristics, reflecting information from multi-hop entities. The only difference when applied to items is the aggregation over a KG, not over an interaction graph.

4.2.3. Cross-modal representation learning layer

Similar to calculating text-based representations with graph, textual information is reflected in graph-based representations. We create $\acute{G}_{u,p}$, a graph-based aspect-level representation with text for pth aspect as:

$$\acute{G}_{u,p} = \frac{\sum^{A}_{j=1} C_{p,j} \times G_{u,j}}{\sum^{A}_{i=1} C_{p,i}} \tag{9}$$

$C_{p,j}$ acts as an attention weight to aggregate graph-based representations $G_{u,j}$. Therefore, when learning pth graph-based representation with text, associations between all aspects in a graph and pth aspect in text are considered. $\acute{G}_u$, a final graph-based user representation with text, is the sum of all aspect-level representations.

4.3. Cross-modal contrastive learning

To integrate graph and text data for effective aspect modeling, we utilize two aspect-level representations of a user. Since we cannot directly map aspects from each data expressed in different latent spaces, aspect-level representations of a user derived from each data are used to combine multi-modal data. A possible reason is that a user's aspect should be the same regardless of the data type. We can learn informative preferences from graph and text data by performing contrastive learning on aspect-level representations.

Having established two views of users, we treat the views of the same user as the positive pairs (i.e., $\{(z'_u, z''_u) | u \in \mathcal{U}, z'_u \in \acute{T}_u, z''_u \in \acute{G}_u\}$), and the views of any different users as the negative pairs (i.e., $\{(z'_u, z''_v) | u, v \in \mathcal{U}, u \neq v, z'_u \in \{\acute{T}_u, \acute{G}_u\}, z''_v \in \{\acute{T}_v, \acute{G}_v\}\}$). We maximize the agreement of positive user pairs and minimize that of negative pairs:

$$L^{user}_{ssl} = \sum_{u \in \mathcal{U}} -\log \frac{\exp(s(z'_u, z''_u)/\tau)}{\sum_{v \in \mathcal{U}} \exp(s(z'_u, z''_v)/\tau)} \tag{10}$$

where $s(\cdot)$ is set as cosine similarity function, which measures similarity between two representations; τ is the hyper-parameter, same as temperature in softmax. Item contrastive loss L^{item}_{ssl} is calculated in the same way. These two losses are combined for the self-supervised objective function:

$$L_{ssl} = L^{user}_{ssl} + L^{item}_{ssl} \tag{11}$$

Traditional multi-modal contrastive learning methods embed each data first and then align them into the same vector space, which leads to shallow embeddings. In contrast, we embed each information based on the correlation between latent aspects and then align them to make mapping possible. This leads to deep embeddings, which helps our model to understand data correlation better and thus resolve data imbalance.

4.4. Fine-tuning for aspect-based recommendation

After pre-training with cross-modal contrastive learning, we can fine-tune it in the downstream recommendation task. Specifically, we use the pre-trained graph-based encoding for each target user to produce a user embedding $\acute{G}_u$. An item embedding $\acute{G}_i$ is generated in the same way, using item documents and relation-aware information from a KG. To predict how likely a user would adopt an item, we employ the inner product on user (or item) representations:

$$y(u, i) = \acute{G}_u{}^{\top} \acute{G}_i \tag{12}$$

Then we calculate Bayesian Personalized Ranking (BPR) loss [45], which assumes that a prediction score of observed interactions should be scored higher than unobserved ones:

$$L_{BPR} = \sum_{(u,i,j) \in O} -\ln \sigma(y(u, i) - y(u, j)) + \lambda \| \Theta_{ssl} \|^2_2 \tag{13}$$

where $O = \{(u, i, j) | (u, i) \in O^+, (u, j) \in O^-\}$ denotes the training data with observed interactions O^+ and unobserved interactions O^-. By leveraging aspect correlations between multi-modal data in representation learning, text-based aspects are considered with a user's historical interactions, which is applicable to users without reviews. From characteristics of interacted items and high-order neighbors in a CKG, we can learn fine-grained representations based on aspects of text data, which in turn helps to handle imbalance between multi-modal data.

Table 2
Statistics of the datasets.

		Amazon-book	Yelp
User–item interaction	#Users	70,679	45,919
	#Items	24,915	45,538
	#Interactions	847,733	1,185,068
	Density	0.0005	0.0006
Knowledge graph	#Entities	88,572	90,961
	#Relations	39	42
	#Triples	2,557,746	1,853,704

5. Experiments

We conducted extensive experiments and ablation studies to verify the effectiveness of the proposed COSMOS. The detailed analysis results are reported in this section.

5.1. Experimental setup

5.1.1. Datasets

To evaluate the proposed model against several state-of-the-art baseline models, we conducted experiments on the *Amazon-book* and *Yelp* datasets. For each dataset, we combined multi-modal data to construct a CKG with review texts. For experiments, we selected 80% of the interaction history as the training set, and the remaining as the test set.

- **Amazon-book**[1] : We expanded the CKG of Book domain released by KGAT [1], which consists of user–item interactions and a Book KG. This interaction graph is constructed based on the Amazon-review dataset, regarding each review as an interaction. Therefore, by mapping the IDs of users and items in interaction records to those in reviews, we obtained review data of each interaction and constructed the Book CKG related to user reviews.
- **Yelp**[2] : This dataset also utilizes the CKG of Business domain from KGAT. Similar to Amazon-Book, we got review data of all interactions by mapping IDs between reviews from the Yelp challenge dataset and user–item interactions from a Business CKG.

The statistics of the two datasets are summarized in Table 2. Since all interactions are mapped with user reviews, the number of user–item interactions is the same as that of review texts.

5.1.2. Implementation details

For text-based aspect modeling (including text-based encoding in our model), we applied pre-trained *word2vec*[3] to initialize word embeddings, following existing text-based aspect modeling methods which use 300 dimensions and set local context window size as 3. Meanwhile, for graph-based aspect modeling (including graph-based encoding in our model), we used Xavier initialization by setting the dimension of ID embeddings as 64 and the model depth l as 3, following what KGIN has shown to be the most effective. We adopted Adam [46] optimizer, where the learning rate was set as 0.001 and the batch size as 64. The number of aspects A between user–item relations was 4 in Amazon-Book and 6 in Yelp. We adopted Precision@K, Recall@K, F1@K and NDCG@K metrics to evaluate the top-K recommendation, where K is set as 20 as default and reported the average results. The higher values of all the metrics indicate better performance.

[1] http://jmcauley.ucsd.edu/data/amazon/index.html.
[2] https://www.yelp.com/dataset.
[3] https://code.google.com/archive/p/word2vec/.

5.1.3. Baseline approaches

We compared COSMOS with knowledge-enhanced recommender systems, covering aspect-free (NARRE, KGAT, MMCPR) and aspect-aware (ANR, KGIN, GERA) methods. Specifically, aspect-aware models are divided into text-based (ANR, GERA) and graph-based (KGIN) methods, according to the knowledge type of aspect modeling:

- **NARRE** [5] is a text-based recommendation model that utilizes CNN and a Feedforward neural network to learn representations of users and items from relevant reviews.
- **KGAT** [1] is a GNN-based recommendation method that explores high-order connectivity with regard to relations in a CKG.
- **MMCPR** [44] is a contrastive learning recommendation model that considers and aggregates cross-modality side information both on the user and item side.
- **ANR** [8] is a text-based recommendation model, which extracts aspects from reviews by estimating aspect-level importance of users and items.
- **KGIN** [7] is a state-of-the-art graph-based recommendation model that captures aspects from user–item interactions based on item–item relations in a KG.
- **GERA** [10] is a neural aspect-based recommendation model, which figures out aspects with opinions based on reviews, and then constructs a KG to apply KG embedding techniques.

For text-based methods (NARRE, ANR, GERA), we leveraged the review set of users towards items. Besides, for graph-based models (KGAT, KGIN), we regarded user reviews towards items as interactions. We set the integration of user–item relationships and a KG as the input of the experiments. Additionally, MMCPR utilized metadata of items, including text and image, and regarded it as user–item interaction.

5.2. Overall performance comparison

Table 3 shows our experimental results on two datasets. The results of aspect-based recommender systems and multi-modal contrastive learning based recommender system varying K are also shown in Fig. 5. We reported Recall@K and NDCG@K on Amazon-Book and Yelp datasets, ranging the recommended item numbers K in {10, 20, 30, 40}. We had the following findings:

- In most cases, aspect-aware recommender systems showed better performance than aspect-free methods except MMCPR, indicating that aspects profile preferences at a fine-grained level. By decomposing a single user–item interaction into multiple aspects, fine-grained user preferences and item characteristics lead to better deciding why a user likes an item.
- Exceptionally, although MMCPR does not consider fine-grained aspects of a user or an item, it showed the highest recommendation performance among existing studies. With the contrastive learning paradigm, various information modalities can be fused for better correlating users and items.
- Text-based and graph-based models achieved similar performances among aspect-based recommender systems, while text-based methods showed slight improvements. This is probably because reviews include notable aspects, which means that it is possible to extract fine-grained aspect-level preferences than graph-based general ones.
- COSMOS outperformed all baselines across two datasets. Specifically, COSMOS showed improvements over the strongest baselines by 5.39% and 2.63% on Recall, and 6.04% and 2.82% on NDCG in Amazon-Book and Yelp, respectively. Since COSMOS captures fine-grained aspect-level preferences from both graph and text data, it achieved competitive results over state-of-the-art knowledge-aware recommender systems. In short, our model leverages both salient and general features from multi-modal data to figure out a more fine-grained preference, by effectively integrating latent aspects from different knowledge with a contrastive learning paradigm.

Table 3
Overall performance comparison.

	Amazon-book				Yelp			
	Precision	Recall	F1	NDCG	Precision	Recall	F1	NDCG
NARRE	0.0158	0.1503	0.0286	0.0898	0.0076	0.0722	0.0138	0.0729
KGAT	0.0154	0.1485	0.0279	0.0893	0.0073	0.0709	0.0132	0.0865
MMCPR	0.0186	0.1754	0.0344	0.0965	0.0084	0.0815	0.0153	0.0932
ANR	0.0182	0.1691	0.0329	0.0918	0.0083	0.0801	0.0150	0.0912
KGIN	0.0179	0.1684	0.0324	0.0910	0.0082	0.0796	0.0149	0.0905
GERA	0.0184	0.1725	0.0333	0.0941	0.0086	0.0799	0.0155	0.0922
COSMOS	**0.0198**	**0.1854**	**0.0358**	**0.1027**	**0.0089**	**0.0837**	**0.0161**	**0.0959**

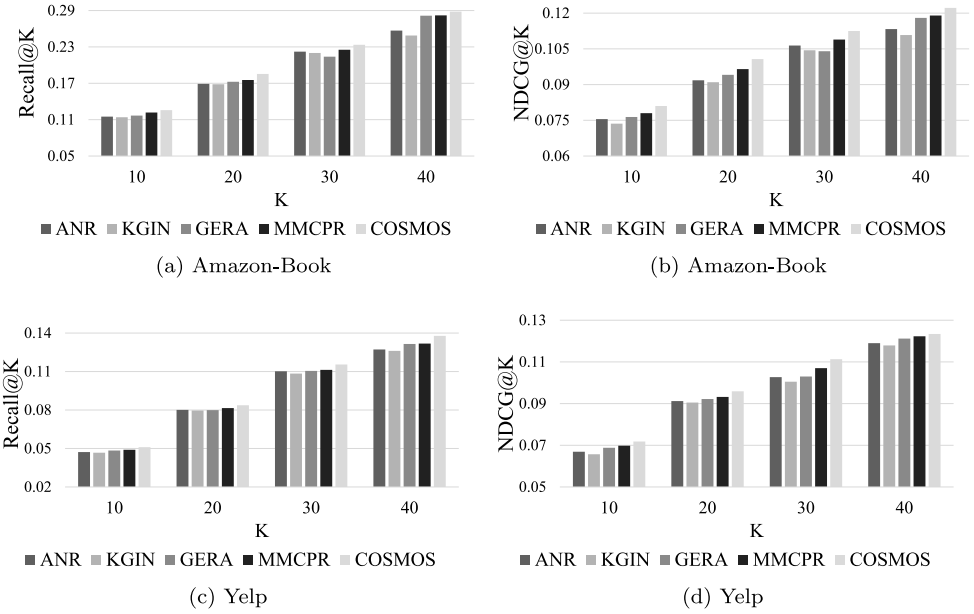

Fig. 5. Recommendation performance on Recall@K and NDCG@K.

5.3. Further analysis

5.3.1. The number of aspects

We conducted experiments to examine the impact of the number of aspects, varying A from 1 to 8. Based on the Recall metric, the number of aspects A between user–item relations was adopted as 4 in Amazon-Book and 6 in Yelp. The results are shown in Fig. 6, where fewer aspects imply coarse-grained aspects, whereas more aspects indicate fine-grained aspects. Our study indicates that:

- When A increased, the model's performance improved in both datasets, until it reached a certain level. It emphasizes that modeling user–item relations at multiple fine-grained level of aspects leads to better finding out the reason for a user's choice of an item.
- In Amazon-Book, the model yielded the best performance when there were 4 aspects between user–item interactions, while in Yelp, it performed best at 6. The performance decreased when the aspect number increased continuously, which means too fine-grained aspects are inappropriate to identify users' preferences. In particular, compared to Yelp, setting A larger made the performance worse in Amazon-Book. This supports the comment in KGIN that since Freebase entities constitute a Book KG, item relations that are irrelevant to a user's choice (e.g., *kg.object_profile.prominent_type* and *rdf-schema#label*) may make lower correlation scores towards text-based aspects.

5.3.2. Impact of the interaction sparsity

Table 3 shows that all models, including baseline approaches and ours, have a low score in both datasets. Each Amazon book and Yelp

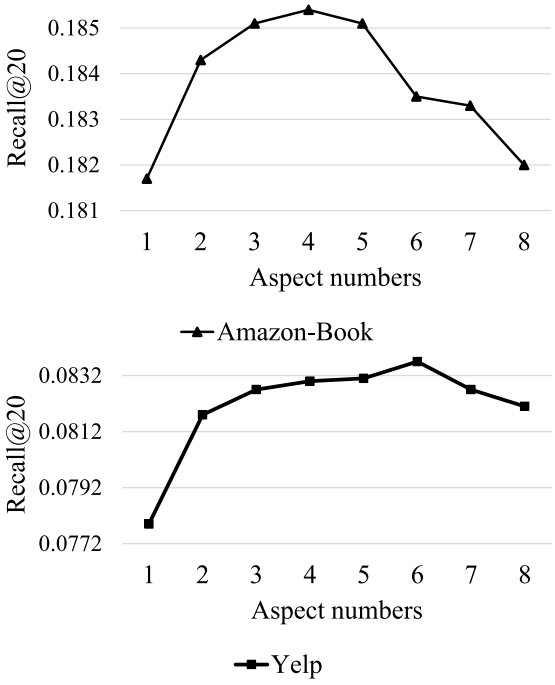

Fig. 6. Impact of the number of aspects.

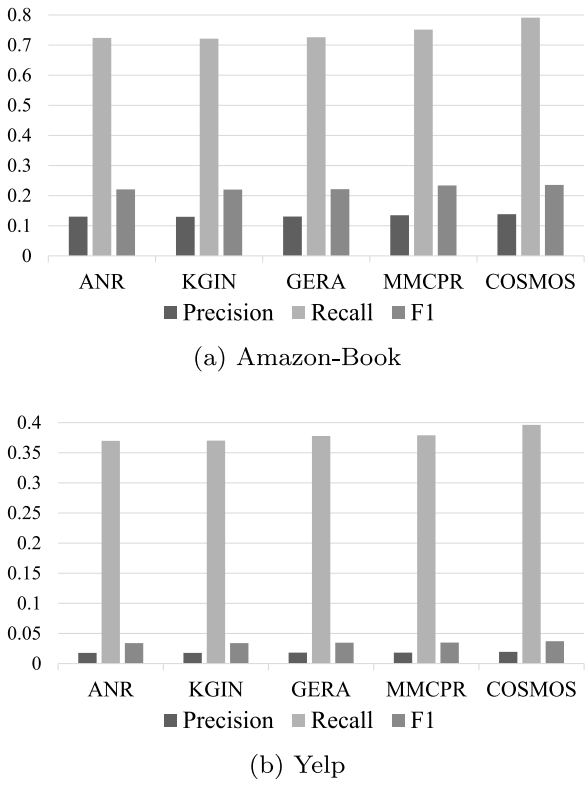

Fig. 7. Performance comparisons of knowledge-aware recommendations based on new datasets from Table 4.

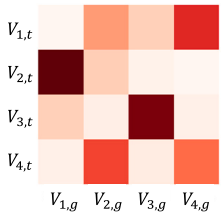

Fig. 8. A global correlation matrix that shows the relatedness between the aspects defined in text and graph, when there are four aspects.

Table 4
Statistics of the datasets, where cold-start users and items are removed.

		Amazon-book	Yelp
User–item interaction	#Users	12,746	11,058
	#Items	24,480	44,212
	#Interactions	348,951	531,361
	Density	0.0011	0.0011

dataset has a low density of 0.0005 and 0.0006, as shown in Table 2. To compare the model performance according to interaction sparsity levels, we conducted an additional experiment with a dataset excluding cold-start users and items. For Amazon-Book, we removed the case of fewer than 10 interactions, and for Yelp, we dismissed the case of fewer than 20 interactions. As a result, both Amazon-Book and Yelp have a density of 0.0011 in the newly created dataset. Table 4 summarizes the statistics of the datasets where cold-start users and items are removed, and Fig. 7 shows the Precision, Recall, and F1 score of aspect-based recommendations (ANR, KGIN, GERA, COSMOS) for the new dataset:

- In both datasets, the performance of aspect-based recommendation models greatly improved. Density refers to the number of user–item interactions among the product of the total number of users and the number of items. Therefore, we can conclude that as the density increases, the performance improves by better grasping a user's preference for an item.
- Even in the case of excluding cold-start users or items, the performance of COSMOS is higher than that of the state-of-the-art models. The lower the density, the lower the performance, but we can observe that COSMOS performs the best under any circumstances.

5.3.3. Case study

We presented the aspects defined on each type of data in Amazon-Book; top-5 words of an aspect in the text and top-2 relations of an aspect in the graph. In the situation where there are four aspects, each aspect of text/graph and global aspect correlation matrix are shown in Table 5 and Fig. 8, respectively. Also, an example recommendation of Amazon-Book domain is described in Tables 6 and 7.

- Some reviews left by u_1 and each item i_1 and i_2 received are shown in Table 6. From u_1's review, we can infer that this user mainly prefers the genre of 'love'. Looking at the aspect attention scores in Table 7, when we perform an aspect-based recommendation based only on the graph, items i_1 and i_2 are both recommended to a user u_1. It is due to the first graph-based aspect $V_{1,g}$, related to 'theater.play.genre' and 'theater.plays in this genre' relations of a KG.
- However, in the aspect-based recommendation based on text, item i_1 is recommended to user u_1, but item i_2 is not. Given that the attention score of $V_{2,t}$, which represents the genre aspect, has been significantly lowered, we can verify that the recommendation results for two items of entirely different genres have changed. Since texts reflect a more detailed opinion than a graph, we can conclude that a user's specified aspect leads to better recommendation results.

5.4. Ablation study

5.4.1. Impact of contrastive learning

Existing aspect-based recommendations model aspects of a user or an item based on one type of knowledge (e.g., review text or a KG). Because we are the first to integrate different types of data for fine-grained user preferences, there is no way to combine user preferences in different latent spaces without contrastive learning. Therefore, we just concatenated the aspect-level preferences from text and graph data. Then we measured the performance on the recommendation task in an end-to-end manner, named COSMOS$_{w/o\ CL}$, to demonstrate the efficiency of contrastive learning. We evaluated the performance for two things: using both graph and text and only graph without text. The results are shown in Table 8:

- When COSMOS$_{w/o\ CL}$ used both graph and text data, although it performed better than the existing baselines, which only utilized one, there was no significant difference in the performance. Since we have employed all user preferences identified in each data, we expect the performance to be much better than the methods learned to represent the user by utilizing only one data. However, looking at the actual experimental results, we can conclude that concatenation does not integrate the two data effectively.
- However, the performance of COSMOS$_{w/o\ CL}$ degrades when only graph data is used as an input for the recommendation task (w/o text). It is because user preferences are identified only in one data, and the performance is similar to baseline using only graphs (KGIN).
- COSMOS shows the best performance even when we only use a graph without text. It is because COSMOS does not simply integrate graph and text data but integrates both data while learning interdependent deep representations. Thus, contrastive learning is essential for learning representations that combine informative features from two data.

Table 5
The aspects defined on each type of data, where top-5 words are listed for each text-based aspect, and top-2 KG relations are listed for each graph-based aspect.

	Text					Graph	
V_1	Animal	Flat	People	Personality	Relationships	theater.play.genre	theater.plays in this genre
V_2	Exploration	Time-travel	Love	Funny	Battle	book.date_written	book.short_story.genre
V_3	Resolution	Twist	Conflict	Event	Plot	date_of_first_performance	fictional_universe.
V_4	Time	Atmosphere	Culture	Society	Historical	theater.play.genre	book.illustrator

Table 6
User and item reviews, where a user review is a review set written by a user, and an item review is a review set that an item receives. Here we listed only a part of each review set.

	Review texts
u_1	This book is a true masterpiece of the **romance genre** - the **writing** is exquisite, the **characters** are fully realized, and the **love story** is both heartwarming and heartbreaking. It is a book that will make you believe in **the power of love**, even in the most difficult of circumstances.
i_1	An adrenaline-pumping **thriller** that had my heart racing from start to finish. The **relentless pacing, high-stakes plot, and complex characters** made for a truly exhilarating read. A perfect choice for fans of **nail-biting suspense**!
i_2	**Emotional and heart-wrenching**, this book is a **love story** that will make you laugh, cry, and feel all the feels. The **characters** are complex and multi-dimensional, the **plot** is gripping, and the **writing** is beautiful - a must-read for anyone who loves a **good romance**.

Table 7
Aspect attention scores in the example study based on aspects of text and graph, respectively.

User	Item	Text		Graph	
u_1	i_1	$V_{1,t}$	0.1642	$V_{1,g}$	**0.8042**
		$V_{2,t}$	**0.8107**	$V_{2,g}$	0.0232
		$V_{3,t}$	0.0233	$V_{3,g}$	0.1634
		$V_{4,t}$	0.0018	$V_{4,g}$	0.0092
	i_2	$V_{1,t}$	0.2311	$V_{1,g}$	**0.7616**
		$V_{2,t}$	0.2863	$V_{2,g}$	0.0450
		$V_{3,t}$	**0.3410**	$V_{3,g}$	0.1574
		$V_{4,t}$	0.1416	$V_{4,g}$	0.0360

5.4.2. Impact of cross-modal aspect correlations

We performed an ablation study to evaluate the efficiency of cross-modal representation learning. Instead of learning cross-modal aspect-level representations from text and graph-based aspect encoding, we generated two views from each data without considering their correlations, called COSMOS$_{\text{w/o AC}}$. Then a user's aspect-level representation is learned through contrastive learning with those two representations.

Existing works that utilize textual information remove the interaction records without review text in the preprocessing stage. However, it is unrealistic to suppose that users always write reviews when interacting with an item. Thus, we compared the performance of COSMOS$_{\text{w/o AC}}$ and COSMOS for the situation where only graph data is used in the downstream recommendation task. When we only use graphs in the fine-tuning stage, the pre-trained graph encoding stage learns aspect-level representations of users and items. As shown in Table 8, we had the following observations:

- The performance of COSMOS$_{\text{w/o AC}}$ was better compared to all baselines. Although it does not consider cross-modal aspects, it helps to model fine-grained aspects by attracting representations from graph and text, which can take advantage of both data.
- However, COSMOS$_{\text{w/o AC}}$ reported lower performance results than the basic COSMOS model. In COSMOS$_{\text{w/o AC}}$, we aligned representations embedded from each multi-modal data into the same vector space, which leads to shallow embeddings. This result shows the importance of cross-modal aspect correlations.
- Instead of merely aligning representations after obtaining them from each data, COSMOS associates multi-modal data in the embedding learning stage with inter-modal aspect relationships. Therefore, COSMOS can learn deep embeddings and handle data imbalance, which leads to high recommendation performance even when both data are not utilized.

5.4.3. Impact of text/graph in the fine-tuning stage

We conducted an ablation study to investigate the effect of text and graph data in the fine-tuning stage. Both data are used in the pre-training stage for capturing aspect correlations, and the main difference is the type of knowledge used in the fine-tuning stage. As in the original architecture, COSMOS$_{\text{w/o text}}$ represents fine-tuning only with a KG, while COSMOS$_{\text{w/o graph}}$ represents fine-tuning only with texts. The results are reported in Table 8, and we find that:

- The performance of COSMOS$_{\text{w/o text}}$ was a little better than COSMOS$_{\text{w/o graph}}$. Also, in each ablation study COSMOS$_{\text{w/o CL}}$ and COSMOS$_{\text{W/O AC}}$, fine-tuning with graph showed better performance than fine-tuning with text. From this results, we can infer that graph is more effective because it identifies the relationship between items and directly reflects its attributes.
- Also, there is an important reason for fine-tuning with graph data. As mentioned in the motivation part, not all users write reviews when interacting with items. When fine-tuning with graph's interaction data, context-aware aspects towards each interaction can be considered, which leads to the effect of using both text and graph with graph data alone.

6. Conclusions and future work

In this paper, we proposed a novel aspect-based recommendation model named COSMOS, which effectively integrates graph and text-based aspect modeling methods by maximizing mutual information between semantic representations. Although unifying aspect preferences from each knowledge is necessary to provide high-quality recommendations, aspects in different latent spaces cannot be directly mapped to each other. To address this issue, we applied contrastive learning on multi-modal data by generating two aspect-level representations and regarding them as multiple views of each user or item. Moreover, COSMOS considers correlations between aspects defined on graph and text data, which can help handle data imbalance between them. Experimental results on two datasets demonstrated the effectiveness of COSMOS, as well as representation learning with cross-modal

Table 8
Impact of contrastive learning and cross-modal aspect correlations. COSMOS$_{w/o\ CL}$ represents COSMOS without contrastive learning, and COSMOS$_{w/o\ AC}$ implies COSMOS without cross-modal aspect correlations.

		Amazon-book				Yelp			
		Precision	Recall	F1	NDCG	Precision	Recall	F1	NDCG
COSMOS$_{w/o\ CL}$	–	0.0190	0.1777	0.0343	0.0975	0.0087	0.0814	0.0157	0.0937
	w/o text	0.0181	0.1687	0.0327	0.0915	0.0083	0.0801	0.0150	0.0908
	w/o graph	0.0178	0.1702	0.0322	0.0899	0.0079	0.0798	0.0143	0.0898
COSMOS$_{w/o\ AC}$	w/o text	0.0187	0.1795	0.0339	0.0988	0.0084	0.0816	0.0152	0.0927
	w/o graph	0.0182	0.1704	0.0323	0.0927	0.0080	0.0803	0.0145	0.0908
COSMOS	w/o text	0.0198	0.1854	0.0358	0.1027	0.0089	0.0837	0.0161	0.0959
	w/o graph	0.0189	0.1760	0.0341	0.0964	0.0085	0.0824	0.0154	0.0929

aspects. In future work, we will explore cross-domain aspect-based recommendation, which not only considers aspect correlations between multi-modal data, but also between two relevant domains.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Kyong-Ho Lee reports financial support was provided by National Research Foundation of Korea. Kyong-Ho Lee has patent pending to Korean Intellectual Property Office.

Data availability

We have shared the link in the manuscript file

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP; Ministry of Science, ICT & Future Planning) (No. NRF-2022R1A2B5B01001835). Kyong-Ho Lee is the corresponding author.

References

[1] X. Wang, X. He, Y. Cao, M. Liu, T.-S. Chua, Kgat: Knowledge graph attention network for recommendation, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 950–958.

[2] Q. Zhu, X. Zhou, J. Wu, J. Tan, L. Guo, A knowledge-aware attentional reasoning network for recommendation, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 6999–7006.

[3] Z. Wang, G. Lin, H. Tan, Q. Chen, X. Liu, CKAN: Collaborative knowledge-aware attentive network for recommender systems, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020, pp. 219–228.

[4] J. Guo, Y. Zhou, P. Zhang, B. Song, C. Chen, Trust-aware recommendation based on heterogeneous multi-relational graphs fusion, Inf. Fusion 74 (2021) 87–95.

[5] C. Chen, M. Zhang, Y. Liu, S. Ma, Neural attentional rating regression with review-level explanations, in: Proceedings of the 2018 World Wide Web Conference, 2018, pp. 1583–1592.

[6] Z. Qiu, X. Wu, J. Gao, W. Fan, U-BERT: Pre-training user representations for improved recommendation, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, pp. 4320–4327.

[7] X. Wang, T. Huang, D. Wang, Y. Yuan, Z. Liu, X. He, T.-S. Chua, Learning intents behind interactions with knowledge graph for recommendation, in: Proceedings of the Web Conference 2021, 2021, pp. 878–887.

[8] J.Y. Chin, K. Zhao, S. Joty, G. Cong, ANR: Aspect-based neural recommender, in: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, 2018, pp. 147–156.

[9] C. Zhao, C. Li, R. Xiao, H. Deng, A. Sun, Catn: Cross-domain recommendation for cold-start users via aspect transfer network, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020, pp. 229–238.

[10] I. Cantador, A. Carvallo, F. Diez, Rating and aspect-based opinion graph embeddings for explainable recommendations, 2021, arXiv preprint arXiv:2107.03385.

[11] C. Wang, Y. Zhu, H. Liu, W. Ma, T. Zang, J. Yu, Enhancing user interest modeling with knowledge-enriched itemsets for sequential recommendation, in: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, 2021, pp. 1889–1898.

[12] K. Tu, P. Cui, D. Wang, Z. Zhang, J. Zhou, Y. Qi, W. Zhu, Conditional graph attention networks for distilling and refining knowledge graphs in recommendation, in: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, 2021, pp. 1834–1843.

[13] R. Huang, C. Han, L. Cui, Entity-aware collaborative relation network with knowledge graph for recommendation, in: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, 2021, pp. 3098–3102.

[14] R. Togashi, M. Otani, S. Satoh, Alleviating cold-start problems in recommendation through pseudo-labelling over knowledge graph, in: Proceedings of the 14th ACM International Conference on Web Search and Data Mining, 2021, pp. 931–939.

[15] S.-J. Park, D.-K. Chae, H.-K. Bae, S. Park, S.-W. Kim, Reinforcement learning over sentiment-augmented knowledge graphs towards accurate and explainable recommendation, in: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, 2022, pp. 784–793.

[16] S. Luo, X. Lu, J. Wu, J. Yuan, Aware neural recommendation with cross-modality mutual attention, in: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, 2021, pp. 3293–3297.

[17] K. Xiong, W. Ye, X. Chen, Y. Zhang, W.X. Zhao, B. Hu, Z. Zhang, J. Zhou, Counterfactual review-based recommendation, in: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, 2021, pp. 2231–2240.

[18] R. Sun, X. Cao, Y. Zhao, J. Wan, K. Zhou, F. Zhang, Z. Wang, K. Zheng, Multi-modal knowledge graphs for recommender systems, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 2020, pp. 1405–1414.

[19] Y. Wei, X. Wang, L. Nie, X. He, R. Hong, T.-S. Chua, MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 1437–1445.

[20] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.04805.

[21] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI Blog 1 (8) (2019) 9.

[22] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, Albert: A lite bert for self-supervised learning of language representations, 2019, arXiv preprint arXiv:1909.11942.

[23] A.v.d. Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, 2018, arXiv preprint arXiv:1807.03748.

[24] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International Conference on Machine Learning, PMLR, 2020, pp. 1597–1607.

[25] Z. Hu, Y. Dong, K. Wang, K.-W. Chang, Y. Sun, Gpt-gnn: Generative pre-training of graph neural networks, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 1857–1867.

[26] J. Qiu, Q. Chen, Y. Dong, J. Zhang, H. Yang, M. Ding, K. Wang, J. Tang, Gcc: Graph contrastive coding for graph neural network pre-training, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 1150–1160.

[27] J. Zhang, H. Zhang, C. Xia, L. Sun, Graph-bert: Only attention is needed for learning graph representations, 2020, arXiv preprint arXiv:2001.05140.

[28] B. Hao, J. Zhang, H. Yin, C. Li, H. Chen, Pre-training graph neural networks for cold-start users and items representation, in: Proceedings of the 14th ACM International Conference on Web Search and Data Mining, 2021, pp. 265–273.

[29] J. Wu, X. Wang, F. Feng, X. He, L. Chen, J. Lian, X. Xie, Self-supervised graph learning for recommendation, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 726–735.

[30] J. Yu, H. Yin, M. Gao, X. Xia, X. Zhang, N.Q.V. Hung, Socially-aware self-supervised tri-training for recommendation, 2021, arXiv preprint arXiv:2106.03569.

[31] Y. Liu, S. Yang, C. Lei, G. Wang, H. Tang, J. Zhang, A. Sun, C. Miao, Pre-training graph transformer with multimodal side information for recommendation, 2020, arXiv preprint arXiv:2010.12284.

[32] X. Xia, H. Yin, J. Yu, Y. Shao, L. Cui, Self-supervised graph co-training for session-based recommendation, in: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, 2021, pp. 2180–2190.

[33] W. Wei, C. Huang, L. Xia, Y. Xu, J. Zhao, D. Yin, Contrastive meta learning with behavior multiplicity for recommendation, in: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, 2022, pp. 1120–1128.

[34] X. Cao, Y. Shi, J. Wang, H. Yu, X. Wang, Z. Yan, Cross-modal knowledge graph contrastive learning for machine learning method recommendation, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 3694–3702.

[35] Y. Yang, C. Huang, L. Xia, C. Li, Knowledge graph contrastive learning for recommendation, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 1434–1443.

[36] S. Uppal, S. Bhagat, D. Hazarika, N. Majumder, S. Poria, R. Zimmermann, A. Zadeh, Multimodal research in vision and language: A review of current and emerging trends, Inf. Fusion 77 (2022) 149–171.

[37] M. Brousmiche, J. Rouat, S. Dupont, Multimodal Attentive Fusion Network for audio-visual event recognition, Inf. Fusion 85 (2022) 52–59.

[38] L.A. Passos, J.P. Papa, J. Del Ser, A. Hussain, A. Adeel, Multimodal audio-visual information fusion using canonical-correlated Graph Neural Network for energy-efficient speech enhancement, Inf. Fusion 90 (2023) 1–11.

[39] S. Zhang, M. Chen, J. Chen, F. Zou, Y.-F. Li, P. Lu, Multimodal feature-wise co-attention method for visual question answering, Inf. Fusion 73 (2021) 1–10.

[40] E. Wang, Q. Yu, Y. Chen, W. Slamu, X. Luo, Multi-modal knowledge graphs representation learning via multi-headed self-attention, Inf. Fusion 88 (2022) 78–85.

[41] W. Zheng, L. Yan, C. Gou, Z.-C. Zhang, J.J. Zhang, M. Hu, F.-Y. Wang, Pay attention to doctor–patient dialogues: multi-modal knowledge graph attention image-text embedding for COVID-19 diagnosis, Inf. Fusion 75 (2021) 168–185.

[42] X. Yuan, Z. Lin, J. Kuen, J. Zhang, Y. Wang, M. Maire, A. Kale, B. Faieta, Multi-modal contrastive training for visual representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 6995–7004.

[43] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.

[44] Z. Liu, Y. Ma, M. Schubert, Y. Ouyang, Z. Xiong, Multi-modal contrastive pre-training for recommendation, in: Proceedings of the 2022 International Conference on Multimedia Retrieval, 2022, pp. 99–108.

[45] S. Rendle, C. Freudenthaler, Z. Gantner, L. Schmidt-Thieme, BPR: Bayesian personalized ranking from implicit feedback, 2012, arXiv preprint arXiv:1205.2618.

[46] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, arXiv preprint arXiv:1412.6980.