This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/MIC.2021.3059027, IEEE Internet Computing

Exploring Cross-Modality Utilization in Recommender Systems

Quoc-Tuan Truong, Aghiles Salah, Thanh-Binh Tran, Jingyao Guo, Hady W. Lauw

Abstract—Multimodal recommender systems alleviate the sparsity of historical user-item interactions. They are commonly catalogued based on the type of auxiliary data (modality) they leverage, such as preference data plus user-network (social), user/item texts (textual), or item images (visual) respectively. One consequence of this categorization is the tendency for virtual walls to arise between modalities. For instance, a study involving images would compare to only baselines ostensibly designed for images. However, a closer look at existing models' statistical assumptions about any one modality would reveal that many could work just as well with other modalities. Therefore, we pursue a systematic investigation into several research questions: which modality one should rely on, whether a model designed for one modality may work with another, which model to use for a given modality. We conduct cross-modality and cross-model comparisons and analyses, yielding insightful results pointing to interesting future research directions for multimodal recommender systems.

Index Terms—Multimodal Recommender Systems, Multimodality, Cross Modality

1 Multimodal Recommender Systems

Modern online applications, such as e-commerce websites and online sharing platforms, rely heavily on recommender systems to guide their users in browsing the myriad of options offered to them. The goal is to provide every user with a relatively short list of items according to her preferences.

There are various approaches to recommender systems. The predominant model-based approach associates each user and item with a vector representation in some latent space so as to reflect affinities between both sets of objects—the closer the user's and item's representations, the higher the affinity. To learn these representations, classical approaches rely on historical behavioral data (or preference data), such as ratings, clicks, purchases, etc.

Preference data however tends to be very sparse, i.e., only few user-item interactions are observed, often less than 1% out of all possible interactions. This arises naturally in practice due to the large number of users and items, yet most users may have had the opportunity to interact with relatively few items. Similarly, the vast majority of items are in the long tail. Moreover, in a dynamic system, new user or item appears continually. With very few observations in place, it is difficult to learn a prediction model accurately.

One promising direction to alleviate the sparsity is to leverage auxiliary data, i.e., information beyond user-item interactions that can supplement the lack of preference signals. Examples of auxiliary data, also referred to as *modalities*, include product descriptions in text, product images, related items, etc., which often hold a clue on how users consume items. Subsequently, we use *multimodal recommender systems* to broadly refer to models relying on other modalities – in addition to preference information – to infer either user representations or item representations.

1.1 Present Siloization Along Modality Lines

Over the last decade or so, considerable efforts have been expended by the community to develop multimodal recommender systems. As a result, significant advances have been made in terms of architectures and solutions to incorporate auxiliary information into recommender systems [1]. An interesting phenomenon is how these advances in the literature tend to develop in streams along modality lines.

For instance, a stream of *textual* recommender models incorporates item information such as product descriptions, reviews, or article content [2], [3], [4], [5], [6]. In turn, deep learning advances in computer vision, and the availability of pretrained image models, facilitate a stream of *visual* recommender models, capable of leveraging perceptual signals from product images to explain user-item affinities [7], [8], [9]. Meanwhile, a stream of *graph* recommender models make use of social interactions, e.g., friendship or trust relationships, to better explain user preferences [10], [11], [12]. Analogously, other contributions have considered item relatedness – the graph of item-to-item connections – to extend a user's preferences to other items of similar aspect thereby alleviating the sparsity issue [9], [13], [14].

This apparent siloization along modality lines presumes virtual walls among different modalities such as texts, images, or graphs. A model ostensibly designed for images, e.g., [8], would experiment with only the image modality, and compare to other models also purportedly designed for images. In turn, a text-based model, e.g., [6], would be compared to another text-based model, e.g., [4], similarly with user graph [12]. This would have been fine if indeed there are impermeable partitions between these modalities. However, as we investigate in this article, there would be benefits in using a model for a modality different from the one it was originally designed for.

1.2 Investigating Cross-Modality Utilization

One key observation is that most multimodal recommendation algorithms are innately machine learning models that fit the preference data, aided by the auxiliary data as features in some form. While the raw representations of modalities may differ, the eventual representations used in the learning process may have commonalities in form (textual product description may be represented as term vectors, related items as a vector of adjacent graph neighbors, etc.). Indeed, if we peel off the layer of pre-processing steps specific to a modality, we find that, for most models, the underlying representation can accommodate other modalities.

Motivated by this insight, we set out to investigate several research questions surrounding the different modalities. Our scope is the three modalities most frequently used with preference data, namely: text, image, and graph.

- **RQ#1**: *As there are increasingly more datasets with multiple modalities, which modality should one rely on?* A rich dataset may have multiple modalities available. For instance, in addition to preference data, a dataset may have product descriptions, product photos, as well as graph of co-purchased products.
- **RQ#2**: *Could a model designed for one modality (e.g., text) potentially perform better with another modality (e.g., images)?* This is an intriguing, yet under-explored issue. We hypothesize that the answer may well be positive, as modalities could be differentially informative. If affirmative, this would motivate the development of a model's capacity to inter-operate with various modalities.
- **RQ#3**: *In face of the multiplicity of models for a given modality (e.g., text), should we consider a model which is designed for a different modality?* Conventionally, one would look into only models designed for the same modality. We postulate that one should look beyond and consider models designed for other modalities as well.

Contribution and Scope Our primary contribution is a systematic analysis on the comparative values of the modality behind models, as well as the cross-modality utilization of a model for a modality different than the one it was originally designed for. The research questions above offer significant value of practical and academic impact. Their answers would inform whether we should continue perceiving and developing multimodal recommender systems in separate modality streams, or we should approach multimodality in a holistic and inter-operable manner. It is worth noting that our scope revolves around substituting one modality for another, thus widening the usability of current models. It is not our intention to make a statement about a specific algorithm or model, nor to propose yet another specific model. Moreover, while we would touch on the issue of joining modalities simultaneously, we would keep a fuller study of that issue to future work.

2 Experimental Setup

To investigate the research questions outlined in Section 1.2 systematically, we conduct a series of experiments involving comparisons across modalities as well as across models.

TABLE 1
Data statistics.

Dataset	#users	#items	#ratings	#docs	#imgs	#rels
Cellphones	3,383	2,170	9,214	2,170	2,170	2,012
Clothing	5,377	3,393	13,689	3,393	3,393	9,198
Electronics	55,930	30,074	212,863	30,074	30,074	63,242
MoviesTV	28,566	10,116	196,277	10,116	10,116	19,763
Office	24,232	13,520	99,255	13,520	13,520	206,719
Tools	19902	12,522	58,419	12,522	12,522	44,509

2.1 Datasets

The primary consideration is for the datasets to have user-item preferences (e.g., ratings) as well as the three modalities in scope. We rely on six Amazon datasets [7], [8], [15], which contain three auxiliary data modalities: product textual descriptions, images, and a symmetric product graph extracted from the *Also-Viewed* information. We pre-process each dataset so as to keep only those items for which the above three modalities are available simultaneously. For sufficient statistics, we retain users with at least three observed ratings, and items with at least two interactions.

The sizes of the resulting datasets are summarized in Table 1, including the number of users (#*users*), items (#*items*), ratings (#*ratings*), item text documents (#*docs*), item image features (#*imgs*), and relations (#*rels*) or the number of edges per graph. For each dataset, we randomly select 80% of the observed user-item interactions as training data and the remaining 20% as test data.

2.2 Models

We include several representative recommender models spanning various modalities, covering rating prediction as well as ranking objectives. Where possible, we keep the variants comparable (e.g., sharing some underlying base model). The selection of representatives is not fixated on which model is currently state-of-the-art as that changes dynamically and it is not our intent to argue on behalf of specific models. Rather, the selection convenes comparable and well-validated models for the purpose of speaking on the metapoint involving the modalities behind the models.

Two are models originally designed for augmenting preference data with **text** information.

- *CDL*: Collaborative Deep Learning [4] composes matrix factorization, to model user preferences, and stacked denoising autoencoder (SDAE), to represent item textual descriptions.
- *CDR*: Collaborative Deep Raking [16] is analogous to CDL, with a ranking loss instead of a Gaussian likelihood.

Two are models originally designed for augmenting preference data with **visual** information.

- *VMF*: Visual Matrix Factorization [9] leverages item visual features to explain user preferences.
- *VBPR*: Visual Bayesian Personalized Ranking [8] employs a ranking loss for learning preferences.

The final model is originally designed for augmenting preference data with **graph** information.

- *MCF*: Matrix Co-Factorization [9] exploits relationships among products (item graph). It composes

two matrix factorization models with shared item factors to jointly factorize user-item and item-item interaction matrices.

In addition, as a reference point to assess the contribution of auxiliary data, for every model we include its variant (*Base*) [17], [18], [19] relying on **only preference** data (*sans* auxiliary data). Implementations of the above models are available in the Cornac[1] recommendation framework [20].

2.2.1 Cross-Modality Utilization

Originally, each of the above models assumes a specific modality. One of the research questions (RQ#2) calls for experiments where a model is used with a different modality. This is feasible, because the statistical assumptions and architectures of these models allow them to work with any of the modalities in our scope, as elaborated below.

- *Text models with visual or graph data.* CDL and CDR model jointly the user-item preferences and item textual information. The latter is organized into a document-word matrix $\mathbf{X}$, where the ith row $\mathbf{x}_i$ of this matrix, corresponding to item i, is a binary vector indicating the words occurring in the item's textual description. Items visual information (images) are also represented by a matrix $\mathbf{V}$, whose row $\mathbf{v}_i$ corresponds to the image of the ith item. Analogously the item graph is represented by its adjacency matrix $\mathbf{C}$ indicating which items are connected. Hence, to use CDL and CDR with visual (resp. graph) modality we substitute the visual feature matrix $\mathbf{V}$ (resp. adjacency matrix $\mathbf{C}$) for the document-word one $\mathbf{X}$.
- *Visual models with text or graph information.* The visual models assume a vector-based representation of perceptual information. Hence, we follow an analogous strategy as with the text models to use VBPR and VMF with text and graph auxiliary data.
- *Leveraging text/visual information using MCF.* In addition to user-item interactions, MCF integrates item-to-item relationships (item graph) represented by an adjacency matrix $\mathbf{C} = (c_{ij})$, where $c_{ij} = 1$ if items i and j are related, and $c_{ij} = 0$ otherwise. To leverage item textual information with MCF, every item is associated with a binary bag-of-word vector indicating the words appearing in its description. We then use the cosine to measure the similarity between these vectors and build a nearest neighbor graph of items, i.e., $c_{ij} = 1$ if j belongs to the set of nearest neighbors of i, and $c_{ij} = 0$ otherwise. We follow the same approach in the case of visual information, where every item is associated with its image represented as an n-dimensional vector. In all our experiments, we set the number of nearest neighbors to 5. We found that beyond five neighbors, we only gain slight improvement at the cost of more intensive computations.

2.2.2 Model Hyper-Parameter Settings

All the above models include two main types of hyper-parameters: a number of factors K or dimension of the latent space, and a set of regularization parameters. For the former we experiment with different values of K, ranging from 10 to 100 with steps of 10, and report the best performance for every model. For the regularization parameters, we start with the values originally recommended by the authors of each model, and further conduct a pilot study, where we explore values in the set $\{0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000\}$, to select the final values for these hyper-parameters.

For the learnable parameters, e.g., user and item representations, we use the same random initial point for all models where it is possible.

1. https://cornac.preferred.ai

2.3 Evaluation Measures

We focus on Top-N recommendation and adopt two widely used evaluation measures in this context, namely Recall@N and NDCG@N (Normalized Discount Cumulative Gain). Both measures vary from 0.0 to 1.0 (higher is better). Importantly, none of the above models optimizes these measures, making them good external measures for cross-model comparisons.

3 Result Analyses

Figures 1 and 2 report the results of every model across all modalities, datasets and measures. We clarify that preference data is omnipresent. In these figures we use *Base* to refer to scenario in which we only rely on preference data, without use of any auxiliary data. In this case, every model collapses to its preference or collaborative component.

3.1 Is Auxiliary Data Useful?

First, we make a few general observations that help validate the utility of auxiliary data for multimodal recommenders.

In most cases, the *Base* model performs worse than the model augmented by auxiliary data of any modality of interest. Interestingly, the outperformance does not depend on pairing up a model with its original modality. For instance, though CDL was originally designed for text, its performance with image or graph still outperforms the *Base* model substantially. This result holds regardless the modality type a model is used with, thereby emphasizing the importance of cross-modal utilization of models.

We now turn to the three research questions first outlined in Section 1.2 and seek deeper insights from the results.

3.2 Which Modality Should One Rely On?

To investigate this question (RQ#1), Table 2 summarizes the results from Figures 1 and 2 by reporting the best performing modality for each dataset-model pair.

Based on this table, in most cases, it seems that *graph* provides the strongest information, followed by *text* and then *images*. Recall that in our experiment *graph* comes from the *Also-Viewed products* information, i.e., products that tend to be browsed within the same session. By nature, this information can encode contextual relationships among items regarding various aspects such as visual appearance, specification, compatibility, etc. Some of these aspects, e.g., compatibility, are hard to capture based on textual and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/MIC.2021.3059027, IEEE Internet Computing

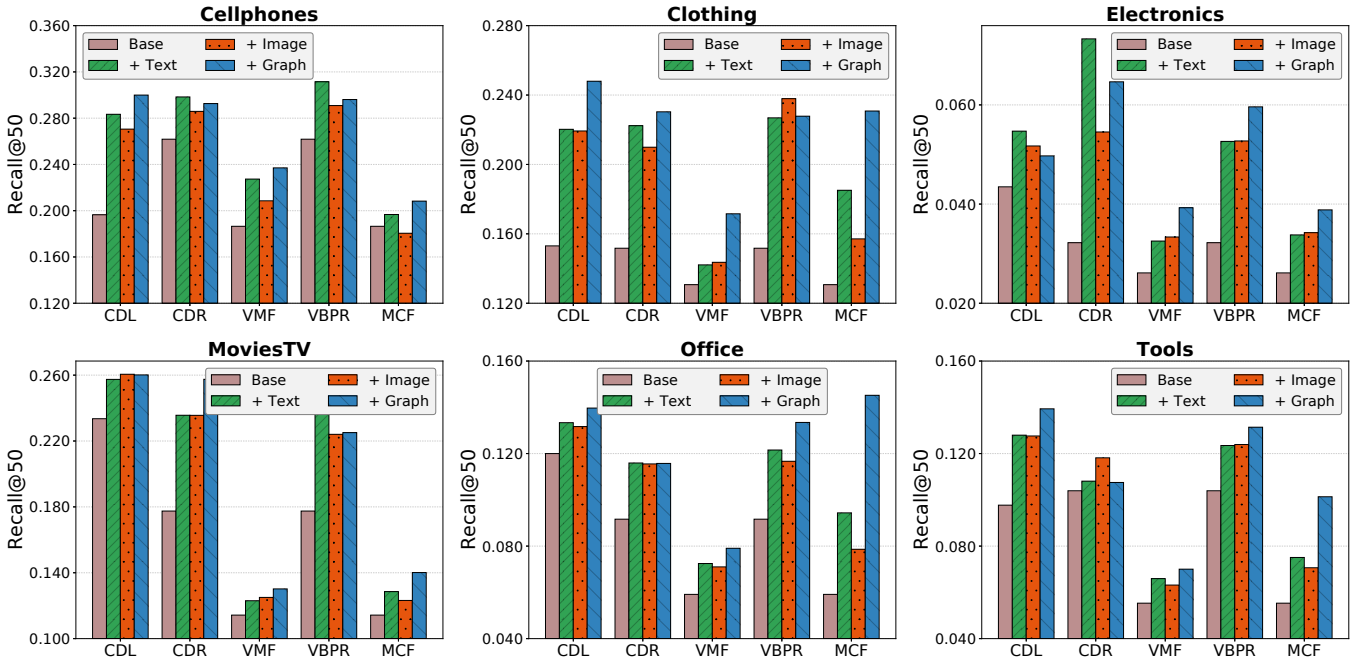


Fig. 1. Recall@50 performance achieved by every model for each modality, across datasets.

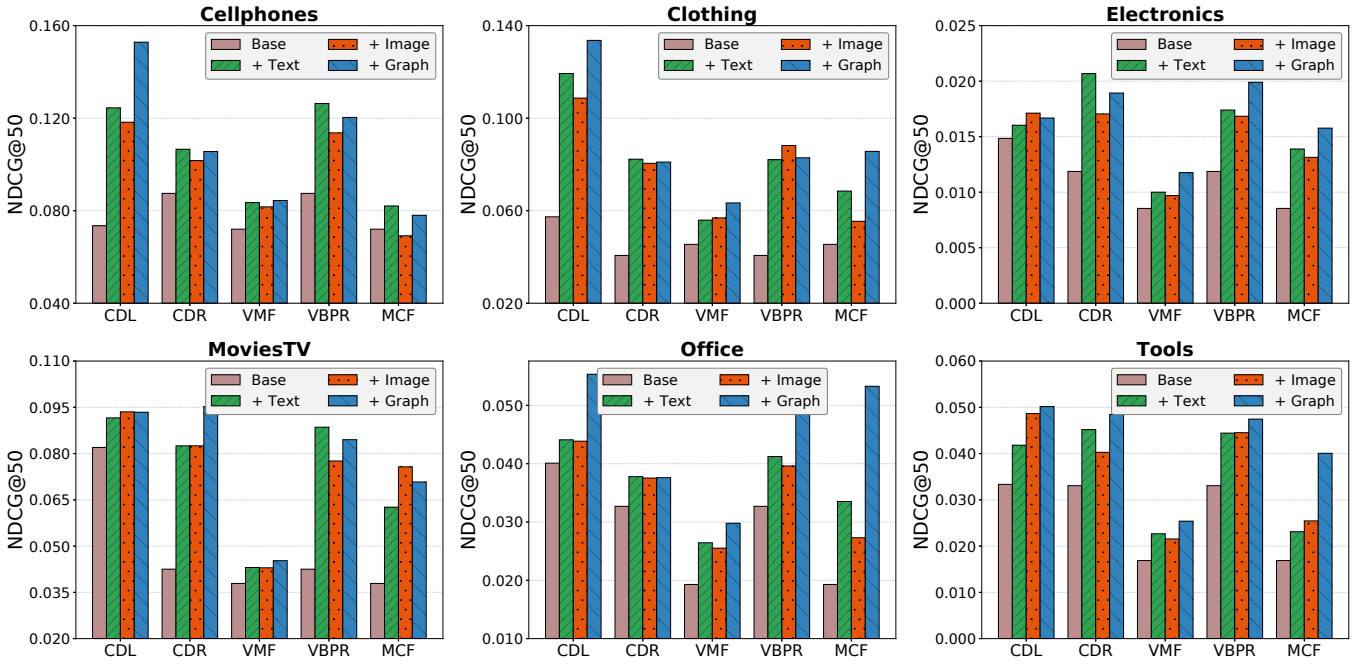


Fig. 2. NDCG@50 performance achieved by every model for each modality, across datasets.

visual features. For instance, a t-shirt and a matching pair of jeans are related to one another and likely to be consumed together, yet they have different visual and textual features. Between two other modalities, *text* tends to help more than *images* in most cases, even when visual appearance is important (Clothing). This suggests that text often carries more information than product image features.

On Cellphones, the behaviour is slightly different in terms of NDCG. Here, *graph* is even more sparse (only few relations are available), which could explain the low improvement in recommendation performance, while *text* seems to carry more information.

Note that the best-performing modality is not the same for all datasets. Thus, the answer to this research question is largely data-dependent, which is reasonable because modalities may be differentially informative.

TABLE 2
Best-performing modality with Recall@50 and NDCG@50, for each model (with their originally designed modality) across datasets.

	Dataset	CDL (text)	CDR (text)	VMF (image)	VBPR (image)	MCF (graph)
Recall@50	Cellphones	graph†*	text	graph†*	text†*	graph*
	Clothing	graph†*	graph†*	graph†*	image*	graph*
	Electronics	text*	text*	graph†*	graph†*	graph*
	MoviesTV	image†	graph†*	graph†*	text†*	graph*
	Office	graph†*	text	graph†*	graph†*	graph*
	Tools	graph†*	image†*	graph†*	graph†*	graph*
NDCG@50	Cellphones	graph†*	text	graph†*	text†*	text
	Clothing	graph†*	text*	graph†*	image*	graph*
	Electronics	image†*	text*	graph†*	graph†*	graph*
	MoviesTV	image†	graph†*	graph†*	text†*	image†*
	Office	graph†*	text	graph†*	graph†*	graph*
	Tools	graph†*	graph†*	graph†*	graph†*	graph*

† statistically significant as compared to the original modality (paired t-test with p-value < 0.05).
* statistically significant as compared to all other modalities (paired t-test with p-value < 0.05).

3.3 Can a Model Designed for One Modality Perform Better with Another Modality?

This (RQ#2) is an interesting question because it opens up new possibilities. Table 2 indicates, for every model and dataset, under which modality it achieves its best performance. A priori, one would surmise that a model would perform the best with the modality it had originally been designed for (parenthesized in the table heading). Evidently, these tables show otherwise. With few exceptions (e.g., VBPR on Clothing, CDR on Electronics), the general tendency is that the best performing modality may not be the one a model was originally designed for. This emphasizes the importance of cross-modal utilization of models.

3.4 Given a Modality, Should we Consider a Model Designed for a Different Modality?

To explore this (RQ#3), we revisit Figures 1 and 2. This time controlling for a modality (e.g., image in red), we compare across models. For instance, on Office dataset, the best-performing model for image is CDL. This is intriguing for two reasons. First, the conventional approach is to consider only image-based models. Yet, here on several datasets, CDL (originally designed for text) is actually the best-performing model when using images. Second, the best-performing model for a given modality is still data-dependent. This further highlights the importance of cross-modal exploration for each model.

4 Conclusion and Perspectives

The investigation into research questions surrounding multimodality in recommender systems throws out surprising, yet insightful lessons. For one, in searching for solutions as well as baselines, researchers and practitioners alike should reach across the modality 'walls' to consider models that may well have been designed for a different modality. For another, we should encourage a more holistic and unified perspective of multimodality, so as to develop and evaluate a model based on its inter-operability across modalities. This is of interest as the modality that a given model is designed for may not always be *available* nor always the *best-performing* one as our experiment shows.

TABLE 3
Performance of CDL ($K = 100$) with Recall@50 and NDCG@50, under various modalities as well as their combination (simple concatenation).

	Dataset	Text	Image	Graph	Combination
Recall@50	Cellphones	0.2858	0.2748	0.3001	**0.3016**
	Clothing	0.2150	0.2081	0.2377	**0.2714**
	Electronics	0.0516	0.0500	0.0493	**0.0529**
	MoviesTV	0.2564	0.2589	0.2593	**0.2746**
	Office	0.1312	0.1296	0.1362	**0.1428**
	Tools	0.1299	0.1241	**0.1377**	0.1283
NDCG@50	Cellphones	0.1251	0.1179	**0.1542**	0.1290
	Clothing	0.1191	0.1106	**0.1398**	0.1119
	Electronics	0.0158	0.0157	0.0157	**0.0201**
	MoviesTV	0.0902	0.0910	0.0913	**0.1004**
	Office	0.0469	0.0479	0.0513	**0.0615**
	Tools	0.0431	0.0448	**0.0468**	0.0438

Moreover, our empirical findings point to interesting future research directions for multimodal recommender systems. For instance, since our results show that there is no such clear modality that is always better to rely on in all cases, it would be edifying to consider the integration of multiple modalities simultaneously. As a motivating instance, Table 3 reports the outcome of a preliminary experiment, where we fit CDL using the three modalities of interests as well as their combination. For the latter case, corresponding to the column *combined* in the above table, for every item we simply concatenate the vector representations of all the modalities of *text*, *image* and *graph*. For balance, feature values of each modality are scaled to lie in range [0, 1], before concatenation. Interestingly, even this naive approach to integration yields performance improvement over each individual modality on most datasets. In some cases (e.g., Tools) *combined* is better than *text* and/or *image* but stands behind *graph*. Hence, a promising question worth a further look in the future is how multiple modalities can be integrated in a coordinated manner, throwing more light into whether more modalities are always better, and whether there is a systematic approach for combining modalities.

Acknowledgments

This research is supported by the National Research Foundation, Prime Ministers Office, Singapore under its NRF Fellowship Programme (Award No. NRF-NRFF2016-07).

References

[1] S. Zhang, L. Yao, A. Sun, and Y. Tay, "Deep learning based recommender system: A survey and new perspectives," *ACM Comput. Surv.*, vol. 52, no. 1, 2019.
[2] C. Wang and D. M. Blei, "Collaborative topic modeling for recommending scientific articles," in *KDD*, 2011, pp. 448–456.
[3] J. McAuley and J. Leskovec, "Hidden factors and hidden topics: understanding rating dimensions with review text," in *ACM RecSys*, 2013, pp. 165–172.
[4] H. Wang, N. Wang, and D.-Y. Yeung, "Collaborative deep learning for recommender systems," in *KDD*, 2015, pp. 1235–1244.
[5] D. Kim, C. Park, J. Oh, S. Lee, and H. Yu, "Convolutional matrix factorization for document context-aware recommendation," in *ACM RecSys*, 2016, pp. 233–240.
[6] X. Li and J. She, "Collaborative variational autoencoder for recommender systems," in *KDD*, 2017, pp. 305–314.

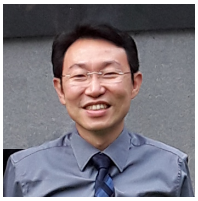
[7] J. McAuley, C. Targett, Q. Shi, and A. Van Den Hengel, "Image-based recommendations on styles and substitutes," in *SIGIR*, 2015, pp. 43–52.
[8] R. He and J. McAuley, "Vbpr: Visual bayesian personalized ranking from implicit feedback." in *AAAI*, 2016, pp. 144–150.
[9] C. Park, D. Kim, J. Oh, and H. Yu, "Do also-viewed products help user rating prediction?" in *WWW*, 2017, pp. 1113–1122.
[10] T. T. Nguyen and H. W. Lauw, "Collaborative topic regression with denoising autoencoder for content and community co-representation," in *CIKM*, 2017, pp. 2231–2234.
[11] W. Fan, Y. Ma, D. Yin, J. Wang, J. Tang, and Q. Li, "Deep social collaborative filtering," in *ACM RecSys*, 2019, pp. 305–313.
[12] X. Wang, X. He, L. Nie, and T.-S. Chua, "Item silk road: Recommending items from information domains to social users," in *SIGIR*, 2017, pp. 185–194.
[13] A. Salah and H. W. Lauw, "A bayesian latent variable model of user preferences with item context," in *IJCAI*, 2018, pp. 2667–2674.
[14] X. Xin, X. He, Y. Zhang, Y. Zhang, and J. Jose, "Relational collaborative filtering: Modeling multiple item relations for recommendation," in *SIGIR*, 2019, pp. 125–134.
[15] J. McAuley, R. Pandey, and J. Leskovec, "Inferring networks of substitutable and complementary products," in *KDD*, 2015, pp. 785–794.
[16] H. Ying, L. Chen, Y. Xiong, and J. Wu, "Collaborative deep ranking: A hybrid pair-wise recommendation algorithm with implicit feedback," in *PAKDD*, 2016, pp. 555–567.
[17] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "Bpr: Bayesian personalized ranking from implicit feedback," in *UAI*, 2009, pp. 452–461.
[18] A. Mnih and R. R. Salakhutdinov, "Probabilistic matrix factorization," in *NeurIPS*, 2008, pp. 1257–1264.
[19] Y. Hu, Y. Koren, and C. Volinsky, "Collaborative filtering for implicit feedback datasets," in *ICDM*, 2008, pp. 263–272.
[20] A. Salah, Q.-T. Truong, and H. W. Lauw, "Cornac: A comparative framework for multimodal recommender systems," *JMLR*, vol. 21, no. 95, pp. 1–5, 2020.

Quoc-Tuan Truong is a PhD candidate in Computer Science at Singapore Management University (SMU). He obtained his bachelor's degree in Computer Science from UET-VNU, Hanoi.

Aghiles Salah is a postdoctoral researcher at SMU. His research is in machine learning. Prior to joining SMU, Aghiles earned his PhD and was formerly an assistant professor (ATER) at Paris Descartes University.

Thanh-Binh Tran is a research engineer at SMU. He obtained a joint master degree from Eurecom and Télécom ParisTech University in France, and his bachelor degree in Computer Science from the VNU University of Science, Ho Chi Minh.

Jingyao Guo is a research engineer at SMU. She holds a master's degree in Computer Control and Automation from the Nanyang Technological University, and a bachelor degree in Electrical Engineering and Automation from the Beijing Jiaotong University.



Hady W. Lauw is an associate professor at SMU, where he leads the Preferred.AI research group working modeling preferences and recommender systems. Formerly, he served as post-doctoral researcher at Microsoft Research in Silicon Valley, as well as scientist at A*STAR's Institute for Infocomm Research. Earlier, he received his PhD from Nanyang Technological University.