Quality-Aware Neural Complementary Item Recommendation

Yin Zhang, Haokai Lu, Wei Niu, James Caverlee
Department of Computer Science and Engineering, Texas A&M University
zhan13679,caverlee@cse.tamu.edu,lu.haokai,weiniu.2010@gmail.com

ABSTRACT

Complementary item recommendation finds products that go well with one another (e.g., a camera and a specific lens). While complementary items are ubiquitous, the dimensions by which items go together can vary by both product and category, making it difficult to detect complementary items at scale. Moreover, in practice, user preferences for complementary items can be complex combinations of item quality and evidence of complementarity. Hence, we propose a new neural complementary recommender Encore that can jointly learn complementary item relationships and user preferences. Specifically, Encore (i) effectively combines and balances both stylistic and functional evidence of complementary items across item categories; (ii) naturally models item latent quality for complementary items through Bayesian inference of customer ratings; and (iii) builds a novel neural network model to learn the complex (non-linear) relationships between items for flexible and scalable complementary product recommendations. Through experiments over large Amazon datasets, we find that Encore effectively learns complementary item relationships, leading to an improvement in accuracy of 15.5% on average versus the next-best alternative.

CCS CONCEPTS

• **Information systems** → *Social recommendation*; *Multimedia information systems*;

KEYWORDS

Complementary Item; Quality-aware; Recommendation

ACM Reference Format:
Yin Zhang, Haokai Lu, Wei Niu, James Caverlee. 2018. Quality-Aware Neural Complementary Item Recommendation. In *Twelfth ACM Conference on Recommender Systems (RecSys '18), October 2–7, 2018, Vancouver, BC, Canada,* Jennifer B. Sartor, Theo D'Hondt, and Wolfgang De Meuter (Eds.). ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3240323.3240368

1 INTRODUCTION

Complementary items that "go well" with one another abound. Examples include a camera that requires a specific lens or a laptop that works well with only certain chargers (see Figure 1). While these complements are strictly compatible – that is, they have particular requirements that allow them to work together – other complements are more loosely related. For example, an aesthetically matching shirt and pants outfit. Different from substitutes items that are interchangeable, complementary items are those that might be purchased together [23].

Figure 1: Complementary item examples and high-quality complementary items (with red mark).

And yet, it can be challenging to identify complementary items, especially considering large and varied item populations (e.g., Amazon boasts around 500 million unique items). For example, one method is to first find exactly compatible items [8]. However, a sample of 500,000 items from Electronics on Amazon finds only 20% explicitly mention compatibility with other items [23, 36], with even rarer occurrences of such mentions in categories like books, movies, and fashion. Hence, in this paper we aim to create new methods for *complementary item recommendation* that can uncover complementary items across items and categories.

In particular, we identify three critical challenges for accurately recommending complementary items: First, the dimensions of how items complement each other vary by item and by category. For example, previous work has shown how to uncover complements based on visual style [24], but some items match based on size or on specific common interfaces. Identifying these features requires adaptive methods that can integrate multiple (possibly conflicting) sources of evidence like images and product descriptions. Second, even if a set of complementary candidate items can be identified, which ones will actually be preferred by users? For example, dozens of iPhone adapters may be identified as complements to an iPhone, meaning that complementary item recommenders must carefully model item quality to discern user preferences. Third, complementary relationships among items are complex, with potential non-linear relationships among item features and item quality. Yet, many existing methods rely on linear combinations of single source features (like visual style) [23, 24, 36], meaning that adapting these methods to complementary items may lead to poor performance.

With these challenges in mind, we address the problem of uncovering complementary item relationships through the creation of a new Neural COmplementary REcommender called Encore. The proposed model is characterized by three unique features:

- Encore can effectively model both stylistic and functional evidence of complementary items through careful balancing of

RecSys '18, October 2–7, 2018, Vancouver, BC, Canada

ACM ISBN 978-1-4503-5901-6/18/10…$15.00
https://doi.org/10.1145/3240323.3240368

high-level visual features learned by a convolutional neural network and text-based embeddings of titles and descriptions;

- Encore naturally models latent item quality through Bayesian inference over user ratings, leading to an item-relationship based quality-aware ranking method; and
- Encore builds a novel neural item-relationship based model to learn the complex complementary relationships between items. That is, the interplay of style, function, and quality can be learned for different categories of items, leading to more flexible and scalable complementary item recommendations.

Through experiments over large Amazon datasets, we quantitatively and qualitatively evaluate the performance of Encore versus a suite of state-of-the-art baselines. We find that Encore effectively learns complementary relationships between items, leading to an improvement in accuracy of 15.5% on average versus the next-best alternative across multiple categories of items. We further evaluate how different aspects of the model (e.g., images, text, ratings, neural recommendation) impact the final complementary item recommendation in different categories. We also show examples to illustrate the recommended items by Encore. Ultimately, we find that Encore's careful combination of different sources of complementary evidence is necessary for high-quality recommendation.

2 RELATED WORK

Item-to-item recommendation. Item recommendation often focuses on finding related items that are similar to an item of interest, rather than complements. Such item-to-item recommendation often uses collaborative filtering [18, 20, 30, 32, 37] with similarity functions [2, 7, 21] such as Pearson similarity [26], cosine-based similarity [6], conditional probability-based similarity [12], or simultaneous regression [SLIM] [28]. Recently, Kabbur et al. [11] introduced a model called FISM that uses latent factor matrices to learn item-item similarity. Shambour [33] used Euclidean distance to measure item-item similarity and showed such a method is better than traditional similarity approaches. Li et al. [17] used the proportion of same users who rated items to measure item similarity. Moreover, many approaches seek to find similar items by incorporating user ratings [1, 16, 38] or images [5, 10, 22]. Similarly, context-based recommenders [13] and phrase-level sentiment analysis [4] have been proposed to capture additional item features for improved recommendation.

Item-relationship based recommendation. Recent research has focused on detecting relationships between items – such as substitutes or complements [19, 23–25, 36] – that go beyond traditional item similarity. For example, [34] employed an association rule to find implicit relationships between items and used it as a regularization term in matrix factorization. [25] used user reviews to find relationships between items such as "albums that are similar with Taylor Swift's 1989". McAuley et al. showed how complementary fashion items like dresses and shoes can be recommended by projecting item images into a common visual style space [23]. Image-based recommendations are also discussed to discover substitutable items in a style space [24]. Another improved model based on images of items was proposed by He et al. [8], in which a mixtures-of-experts framework is built to model the relative importance of different image aspects.

Notation	Explanation
$\mathcal{I}$	item set, where $\mathcal{I} = \{I_1, I_2, \ldots I_{\lvert\mathcal{I}\rvert}\}$
$\mathbf{m}_i$,$\mathbf{t}_i$	image/text feature vector for item I_i
r_{ik}, q_{ik}	the k_{th} rating for item I_i and its binary
η_d, η_r	the thresholds for distance and ratings
θ_i	the expected value of q_{ik} after observing ratings
l_{ij}	the relationship (link) between two items $\mathbf{I}_i$ and $\mathbf{I}_j$
C_i	the set of items that are complementary with I_i
$\mathbf{E}_M$, $\mathbf{E}_T$	embedding matrix for image and for text
$\mathbf{W}_k$,$\mathbf{b}_k$	neural network weight matrix/bias term in layer k
$\mathbf{d}^{(n)}_{j\lvert i}(I_i, I_j\lvert\theta_j)$	neural item distance from query item I_i to I_j

Table 1: Notation.

Most existing methods are based on a single source of item information – such as images or textual information. However, in practice, the item relationships are complex and the interaction of items in the relationship varies according to different categories. For example, style-based methods do well for clothing but not for books. In this paper, *we incorporate multiple, possible conflicting sources of item complementarity into a novel neural-based framework that can learn the contributions of each source.* Moreover, we improve upon item-relationship based recommendations that have typically focused on how closely items are related to each other (relevance ranking) by *incorporating latent item quality into complementary item recommendation (via quality-aware ranking).* In the following, we present the design of our Encore approach.

3 OVERALL APPROACH: ENCORE

We assume we have a set of items $\mathcal{I} = \{I_1, I_2, \ldots I_{\lvert\mathcal{I}\rvert}\}$ and sets of links $C_i = \{l_{ic_1}, l_{ic_2}, \ldots l_{ic_{\lvert C_i\rvert}}\}$, $i \in \{1, 2, \ldots \lvert\mathcal{I}\rvert\}$ that describe relationships between a query item I_i and its complementary items $I_{c_1}, I_{c_2}, \ldots I_{\lvert C_i\rvert} \in \mathcal{I}$. Inspired by [24], our goal is to design a *complementary distance function* $d_{j\lvert i}(I_i, I_j)$ that captures a user's preferences for complementary items given $I_i \in \mathcal{I}$. Specifically, we propose the quality-aware Neural Complementary Item Recommendation framework Encore that decomposes the problem into three phases (see Figure 2):

- **Detect Complementary Items.** First, we aim to construct a distance function $d^{(c)}_{j\lvert i}(I_i, I_j)$ that assesses how well an item $I_j \in \mathcal{I}$ complements the seed item I_i based on stylistic properties (via the embedding $\mathbf{E}_M$) and functional properties (via the embedding $\mathbf{E}_T$).
- **Quality-Aware Recommendation.** Second, we augment the first distance with a quality-aware distance $d^{(r)}_{j\lvert i}(I_i, I_j\lvert\theta_j)$ to capture user preferences. Specifically, We show how to estimate the latent item quality θ_j and then asymmetrically incorporate it for ranking candidate complementary items (i.e. given an item I_i, find the nearest high-quality complementary items I_j).
- **Transform via Neural Model.** Finally, to capture the complex relationships between item properties and ratings, we

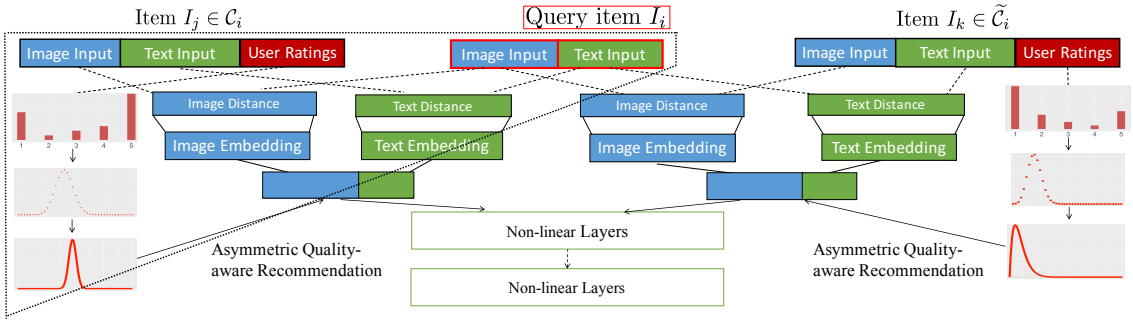

Figure 2: Overall ENCORE Model Framework.

build a neural-based distance $d^{(n)}_{j|i}(I_i, I_j)$ that can jointly learn complementary item relationships ($d^{(c)}_{j|i}(I_i, I_j)$) and user preference in $d^{(r)}_{j|i}(I_i, I_j|\theta)$, leading to high-quality complementary item recommendation.

3.1 Detecting Complementary Items

In this section, we focus on detecting complementary items from two perspectives: style and function. Our aim is to construct a distance function $d^{(c)}_{j|i}(I_i, I_j)$ that balances these two perspectives across different categories. For example, complementary fashion items may mainly match on style (that is, they go well visually with each other). In contrast, a Mac Pro and its charger need to functionally match based on a common interface (that is, the charger needs to fit specifically with the laptop, regardless of style). In practice, these notions of style and function vary across categories and can both be necessary in many cases. For example, while complementary fashion items may need to be stylistic matches, they also need to have similar functional sizes (e.g., identifying a woman's shirt and not one for a toddler). Ultimately, we propose a joint embedding model that captures both perspectives and the model can be customized for different product categories.

Style-Based Complements. As the first step, we exploit the image-based relationship between complementary items to find stylistically-related items. Following [24], we first use the high-level visual features extracted from a convolutional neural network (CNN) proposed by [14]. The CNN is pre-trained by Caffe 1.2 million ImageNet (ILSVRC12 challenge). Particularly, the features that we use are the output of the second fully connected layer in CNN based on their strong performance in previous work [8, 24], and the feature vector length is $f_m = 4096$. After extracting high-level image features, we can learn a low-rank Mahalanobis transformation for image embedding [24] (we refer to this method as **LMT**) and then calculate the Euclidean distance between the high-level image feature vector $\mathbf{m}_i$ and $\mathbf{m}_j$ in the embedding space. The image distance is used to represent the distance between items I_i and I_j, that is:

$$d^{(cm)}_{j|i}(I_i, I_j) = ||(\mathbf{m}_i - \mathbf{m}_j)^T \mathbf{E}_M||_2^2, \quad (1)$$

where $\mathbf{E}_M \in R^{f_m \times f_{em}}$ is the low-rank Mahalanobis transformation matrix and f_{em} is the embedding dimension of image. Based on the distance, a shifted sigmoid function is used to calculate the probability that two items belong to a certain relationship:

$$P(l_{ij} \in C_i) = \sigma(-d^{(cm)}_{j|i}(I_i, I_j)) = \frac{1}{1 + e^{d^{(cm)}_{j|i}(\mathbf{I}_i, \mathbf{I}_j) - \eta_d}}.$$

Based on this probability, we can use maximum likelihood to train $\mathbf{E}_M$ so that it can identify style-based complements [24].

Figure 3: Image-Confusing Items.

Functional Complements. Such an image-based approach is well suited for fashion-related items that demonstrate clear visual style. However, since it relies solely on image-based features, there may be significant errors introduced for complementary relationship when it is applied to other product categories. For example, Figure 3 shows several items that can confuse image-only approaches, such as a Panda USB battery and Baymax flash drive which are complementary with laptops. If they are mis-classified as toys according to their visual appearance, they will be deemed complementary with other toys rather than a laptop. And even when the images themselves are identical (as in Figure 3), an image-only recommender could mistakenly recommend uncomplementary MacBook Pro chargers for a MacBook Air.

Since the complement criteria varies across products (recall Figure 1), instead of using existing methods to find functional topics of each product, we propose to directly learn the functional complementary features.[1] Specifically, we propose to exploit text-based embeddings which can model these more nuanced relationships. That is, we aim to find a compatible text distance $d^{(ct)}_{j|i}(I_i, I_j)$ by $\mathbf{t}_i$ and $\mathbf{t}_j$, where $\mathbf{t}_i$ includes the title and description of item I_i.

Based on that, we propose to extract $\mathbf{t}_i$ by using a distributed representation [15, 27]. Specifically, we train a representation with a window size of 20 and learning rate 0.1. The final representation is a fixed-length feature vector $\mathbf{t}_i \in R^{f_t}$. Through experimental

[1] One initial idea is to mine mentions of complements directly from the text of each product description – e.g., to seek phrases such as "this charger is compatible with MacBook Pro". However, only 20% of products in Electronics (99,304/498,196) contain such explicit mentions [23], with even rarer occurrences of such mentions in categories like Digital Music and Clothing. The method also can not cover all situations for complementary relationships.

Rating	(A)	(B)	(C)
1	1	2	39
2	1	0	53
3	0	0	80
4	0	1	150
5	36	4	341

Figure 4: A "Bella Ladies" hoodie and three complementary pants (A,B,C) with their ratings (on a 1-5 scale).

validation, we find that different dimensions of the text vector, such as 4096, don't strongly impact the overall results, so we use $f_t = 100$. So the distance between items I_i and I_j is calculated by:

$$d^{(ct)}_{j|i}(I_i, I_j) = ||(\mathbf{t}_i - \mathbf{t}_j)^T \mathbf{E}_T||^2_2. \quad (2)$$

where matrix $\mathbf{E}_T \in R^{f_t \times f_{et}}$ is the trained text embedding to learn text features that are related to the complementary relationship and f_{et} is the embedding dimension of text.

3.2 Quality-Aware Recommendation

By carefully combining $d^{(cm)}_{j|i}(I_i, I_j)$ and $d^{(ct)}_{j|i}(I_i, I_j)$, we could immediately begin to recommend the nearest complementary items. E.g., we could combine the two factors as $d^{(c)}_{j|i}(I_i, I_j) = q d^{(cm)}_{j|i}(I_i, I_j) + \rho d^{(ct)}_{j|i}(I_i, I_j)$, where q and ρ are hyper-parameters. In practice, however, users may choose a relatively high-rated complementary item [31], rather than the strictly nearest complementary one. For example, Figure 4 shows three complementary pants for a user who bought a Bella Ladies hoodie. The nearest pants – (A) and (B) – found by $d^{(c)}_{j|i}(I_i, I_j)$ are from the same fashion line Bella Ladies. However, the user's actual choice is (C), a pair of Spandex pants that are more distant by $d^{(c)}_{j|i}(I_i, I_j)$ but that are rated more highly than (A) and (B) (as shown on the right of Figure 4).

Therefore we hypothesize that these complementary items purchase decisions are driven by both perceived match (stylistic and functional) and by item ratings. However, in practice, item ratings are noisy and the number of ratings is different across items, which makes it hard to capture user purchase preference. So we propose to model each item's latent quality through careful consideration of item rating distributions [9, 31, 34]. Concretely, we model *item latent quality* as the expectation θ that a user will highly rate an item, and so it may be easier to capture user purchase preference. Then based on this θ, we propose a quality-aware distance function $d^{(r)}_{j|i}(I_i, I_j)$ that can provide rich user preference information. In the following, we first show an example (Figure 5 and 6) to illustrate why Bayesian inference is preferred to estimate θ here. Then we discuss details of θ_i estimation and build a quality-aware complementary distance $d^{(r)}_{j|i}(I_i, I_j|\theta_i)$.

Figure 5 shows ratings for three items. Item 1 and item 2 have same average ratings, while item 2 has been rated many more times (yielding more confidence in its underlying quality). So there is high probability that users would prefer item 2 than item 1. By Bayesian inference, the posterior distributions for the three example items are shown in Figure 6. We find that the posterior rating distribution for item 2 is more narrow and close to 1 while the distribution

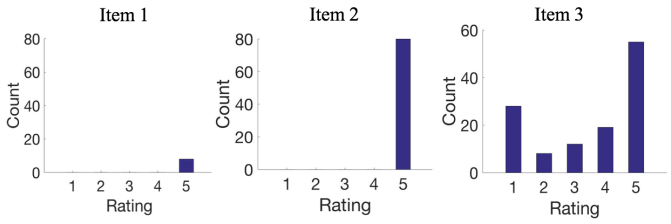

Figure 5: Ratings distributions for three example items.

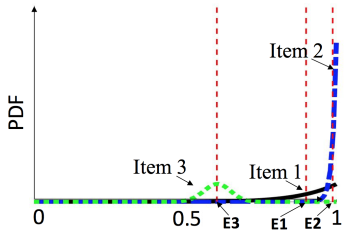

Figure 6: Posterior distribution for the example items. Ei is the expectation of the i^{th} item.

of item 1 is more spread out, which means the posterior rating distribution can properly indicate users have a higher probability to highly rate item 2 than item 1. So we leverage it to estimate θ_i. Concretely, the steps for generalizing $d^{(r)}_{j|i}(I_i, I_j|\theta_i)$ are:

Ratings-Based Bayesian Inference. First, suppose item I_i has ratings $r_{i1}, r_{i2} \ldots r_{i|I_i|}$ by different users and we treat each $r_{ik} \in \mathbb{Z}$ as a random variable for product I_i ratings. Since users have different evaluation scales and there is no big difference between 4-star or 5-star when a 5-scale rating is used [35], we first smooth r_{ik} as a binary random variable q_{ik}. Let η_r be a binary threshold to separate good ratings and bad ratings. If $r_{ik} > \eta_r$, it means r_{ik} is a good rating; otherwise it means the user thinks there are drawbacks of item I_i. So the probability (q_{ik}) that "the k_{th} rating of item I_i is a good rating" can be represented as:

$$q_{ik} = \begin{cases} 1 & \text{if } r_{ik} > \eta_r \\ 0 & \text{otherwise,} \end{cases}$$

and the p.d.f. of q_{ik} for item I_i is:

$$f(q_{ik}|\theta_i) = \begin{cases} \theta^{q_{ik}}(1-\theta)^{1-q_{ik}} & \text{for } q_{ik} = 0, 1 \\ 0 & \text{otherwise.} \end{cases}$$

Thus q_{ik} is Bernoulli distributed $q_{ik} \sim \mathbf{B}(1, \theta_i)$. The expectation that item I_i can get a good rating is $\mathbb{E}(q_i|\theta_i) = \sum_{k=1}^{\infty} q_{ik} f(q_{ik}|\theta_i) = \theta_i$ and $\theta_i \in [0, 1]$ for each item I_i. So θ_i can be used to measure user expectations (refer as quality) towards item I_i. If the value of θ_i is high, it means there is a high probability that the item I_i can get a good rating (i.e. item I_i has high quality).

But how can we estimate θ_i? Many previous methods do not consider ratings for recommendation, so they assume the quality of each item is randomly distributed, that is $\theta_i \in U[0, 1]$ for all $I_i \in \mathcal{I}$. In contrast, we use Bayes' theorem [3, 9] to estimate the posterior p.d.f of θ_i of item I_i based on users' ratings (as we previously

illustrated), where this uniform distribution is treated as a prior for items when we have no rating information:

$$\xi(\theta_i|q_{i1}, q_{i2}...q_{i|I_i|}) = \frac{f_i(q_{i1}, q_{i2}...q_{i|I_i|}|\theta_i)\xi(\theta_i)}{h_i(q_{i1}, q_{i2}...q_{i|I_i|})} \propto f_n(q_{i1}, q_{i2}...q_{i|I_i|}|\theta_i)\xi(\theta_i), \tag{3}$$

where $h_i(\cdot)$ is the marginal joint p.d.f of q_{i1}, $...q_{in_i}$. $\xi(\theta_i)$ is the prior p.d.f. of θ_i. Here it is the p.d.f of uniform distribution $U[0, 1]$. $f_i(q_{i1}, q_{i2}...q_{i|I_i|}|\theta_i)$ is the likelihood function:

$$f_i(q_{i1}, q_{i2}...q_{i|I_i|}|\theta_i) = \prod_{k\in[1,...|I_i|]} f(q_{ik}|\theta_i) = \theta_i^{\sum_{k=1}^{|I_i|} q_{ik}}(1-\theta_i)^{|I_i|-\sum_{k=1}^{|I_i|} q_{ik}}. \tag{4}$$

Let $\Phi(q_{i1}, q_{i2}...q_{i|I_i|}) := \frac{\Gamma(|I_i|+2)}{\Gamma(\sum_{k=1}^{|I_i|} q_{ik}+1)\Gamma(|I_i|-\sum_{i=1}^{|I_i|} q_{ik}+1)}$, where $\Gamma(z) = \int_0^1 x^{z-1}e^{-x}dx$. According to Equation 3, the p.d.f of posterior distribution $\xi(\theta_i|q_{i1}, q_{i2}...q_{i|I_i|})$ is:

$$\xi(\theta_i|q_{i1}, q_{i2}...q_{i|I_i|}) = \Phi(q_{i1}, q_{i2}...q_{i|I_i|})\theta_i^{\sum_{k=1}^{|I_i|} q_{ik}}(1-\theta)^{|I_i|-\sum_{k=1}^{|I_i|} q_{ik}}.$$

So the posterior distribution of θ_i is a Beta distribution

$$\theta_i \sim \textbf{Beta}(\sum_{k=1}^{|I_i|} q_{ik} + 1, |I_i| - \sum_{k=1}^{|I_i|} q_{ik} + 1). \tag{5}$$

Based on the posterior distribution, the expectation of θ_i is $\mathbb{E}(\theta_i|q_{i1}, q_{i2}...q_{i|I_i|}) = \frac{\sum_{k=1}^{|I_i|} q_{ik}+1}{|I_i|+2}$, which we can use to estimate users expectation for item I_i (same as Figure 6 examples).

Recommendation with Asymmetric Ratings. In practice, users care more about the quality of a complementary candidate item I_j, rather than the quality of a query item I_i. So we consider the latent quality for item I_j in $d_{j|i}^{(r)}(I_i, I_j)$. Given I_j's quality estimate, the recommended item quality should be inversely proportional to item distances: the lower the quality of the candidate I_j, the larger the distance from the query item. That is:

$$d_{j|i}^{(r)}(I_i, I_j|\theta_j) \propto \mathbb{E}(1-\theta_j|q_{j1}, q_{j2}...q_{j|I_j|}) = \frac{|I_j| + 1 - \sum_{k=1}^{|I_j|} q_{jk}}{|I_j| + 2},$$

when item I_i is queried and I_j is recommended. But how do we incorporate $\mathbb{E}(1-\theta_i|q_{i1}, q_{i2}...q_{i|I_i|})$ to item relationship distance for our complementary recommendation?

3.3 Neural Recommendation

As shown in Figures 3 and 4, complementary relationships vary greatly across categories. Moreover, users may choose a relatively high quality complementary item rather than the strictly nearest complementary one according to previous analysis of Figure 4. Hence, instead of directly combining three sources of information, we propose a neural-based complementary recommender that can bring some attractive characteristics for complementary item recommendation: (i) Neural methods may capture the variability of complementary relationships for different categories; (ii) Neural models can also offer more flexibility in balancing the style/functional complementary matches with item quality through activations; and (iii) Many neural methods may be easily parallelized for scalable computation [29], which can be beneficial for large item populations with high-dimensional visual and text features. Concretely, we transform Encore's complementary distance into a non-linear space $d_{j|i}^{(n)}(I_i, I_j|\theta_j)$ to capture these complex complementary relationships (see Figure 2).

To calculate $d_{j|i}^{(n)}(I_i, I_j|\theta_j)$ between I_i and I_j, we first extract features of query item I_i and its candidates complementary item I_j in each space, then use embedding to learn visual and functional complementary features separately. Instead of directly using their distances, we concatenate the embedded features with the expectation quality of candidate item I_j into a multi-modal space. So Encore can learn complementary relationships by feature differences in each source:

$$\mathbf{c}_{j|i}(I_i, I_j|\theta_j) = [(\mathbf{m}_i - \mathbf{m}_j)^T\mathbf{E}_M, (\mathbf{t}_i - \mathbf{t}_j)^T\mathbf{E}_T, \mathbb{E}(1-q_j|\theta_j))]^T,$$

where $\mathbf{c}_{j|i}(I_i, I_j|\theta_j) \in \mathbb{R}^{(f_{em}+f_{et}+1)}$. We use it as our asymmetric merged layer of our neural network (because we only consider the quality of candidate item I_j). Then with adding $\mathbf{W}_1 \in \mathbb{R}^{(f_{em}+f_{et}+1)\times m_1}$ and bias $\mathbf{b}_1 \in \mathbb{R}^{m_1}$ in the layer:

$$\mathbf{h}_{j|i}(I_i, I_j|\theta_j) = \mathbf{c}_{j|i}(I_i, I_j|\theta_j) \times \mathbf{W}_1 + \mathbf{b}_1.$$

Then we add the activation function and now the distance becomes:

$$d_{j|i}^{(n)}(I_i, I_j|\theta_j) = ||\tanh(\mathbf{h}_{j|i}(I_i, I_j|\theta_j))\mathbf{W}_2||^2, \tag{6}$$

where $\tanh(-\mathbf{h}_{j|i}(I_i, I_j|\theta_j)) = \frac{e^{-\mathbf{h}_{j|i}(I_i,I_j|\theta_j)} - e^{-\mathbf{h}_{j|i}(I_i,I_j|\theta_j)}}{e^{-\mathbf{h}_{j|i}(I_i,I_j|\theta_j)} + e^{-\mathbf{h}_{j|i}(I_i,I_j|\theta_j)}} \in \mathbb{R}^{m_1}$. And $\mathbf{W}_2$ in Equation 6 is a weight vector when features are put into non-linear space. So we can calculate the probability that item I_i and I_j are complementary when I_i is queried as: $P(l_{ij} \in C_i) = \frac{1}{1+e^{d_{j|i}^{(n)}(I_i,I_j\theta_j)-\eta_d}}$, where η_d is a learned complementary threshold.

Based on the probability, we use the maximum likelihood function to find the maximum observed complementary relationship of set C_i for each I_i. Then the complementary relationship for the item set $\mathcal{I}$ is $C_{\mathcal{I}} = \{C_1, C_2, ...C_{|\mathcal{I}|}\}$. The log-likelihood function for all items in $\mathcal{I}$ is:

$$\begin{aligned} l(\mathbf{E}, \mathbf{W}, \mathbf{b}, \eta_d|C_{\mathcal{I}}, \widetilde{C}_{\mathcal{I}}) &= -\sum_{I_i\in\mathcal{I}}\sum_{I_{ij}\in C_i} lnP(l_{ij} \in C_i|I_i, I_j) - \sum_{I_i\in\mathcal{I}}\sum_{I_{ij}\in\widetilde{C}_i}(1 - lnP(l_{ij} \in \widetilde{C}_i|I_i, I_j))) \\ &= -\sum_{I_i\in\mathcal{I}}\sum_{I_{ij}}(y_{ij}lnP(r_{ij} \in C|I_i, I_j) + (1-y_{ij})(1 - lnP(l_{ij} \in \widetilde{C}|I_i, I_j))), \end{aligned} \tag{7}$$

where y_{ij} indicates whether there is a complementary relationship between item I_i and I_j. If $l_{ij} \in C_i$, then $y_{ij} = 1$; otherwise it is 0. In $l(\mathbf{E}, \mathbf{W}, \mathbf{b}, \eta_d|C_{\mathcal{I}}, \widetilde{C}_{\mathcal{I}})$, $\mathbf{E}$ represents $\{\mathbf{E}_M, \mathbf{E}_T\}$. $\mathbf{W}$ represents $\{\mathbf{W}_1, \mathbf{W}_2\}$. $\tilde{C}_{\mathcal{I}} = \{\tilde{C}_1, \tilde{C}_2, ...\tilde{C}_{|\mathcal{I}|}\}$. Each $\tilde{C}_i$ is a randomly selected negative set of non-complementary items. To train parameters, we generate $\tilde{C}_i$ such that $|\tilde{C}_i| = |C_i|$ [24].

4 EXPERIMENTS

In this section, we evaluate Encore's complementary item recommendations over large Amazon datasets in comparison with state-of-the-art baselines. Especially, we seek to address the following key research questions:

- How well does Encore perform versus baselines? And does this performance vary by item types? And also across complementary relationships (i.e. also-bought versus bought-together)?

Dataset	# Cat.	Also-bought			Bought-together		
		# Items	# Edges	Avg	# Items	# Edges	Avg
Digital Music	198	164,440	6,912,348	42	5,552	9,590	1.73
Movies	345	118,351	5,248,530	44	80,922	130,640	1.61
Cell phones & Accessory	81	122,031	2,985,220	24	100,567	138,815	1.38
Books	2,752	65,024	2,806,544	43	35,638	54,146	1.52
Electronics	786	140,922	4,446,609	32	140,020	194,309	1.39
Clothing	1,993	658,304	20,546,119	31	569,714	1,645,219	2.89

Table 2: Amazon datasets. The second column is the number of subcategories. The Avg column is the average number of linked items for each query item. For example, users also-bought 32 complementary items on average in Electronics and bought-together 1.39 items on average.

- What impact do the design choices of ENCORE have? For example, is textual-added complement more impactful than image-driven complement across different categories? What impact does the neural recommender model have versus a linear model for complementary recommendation?

Finally, we explore the complementary recommendations of ENCORE through several case studies.

4.1 Amazon Dataset

Concretely, we adopt a large real-world dataset from Amazon recently introduced in [8, 24]. The complete dataset contains over 1 million products and 42 million co-purchase relationships across around 20 top-level product categories. We focus on six main categories that display different complementary aspects: Electronics, Cell Phones & Accessories (C & A), Clothing, Books, Digital Music, and Movies (see Table 2 for details) [24]. Specifically, following with previous work [8, 24, 36], we adopt two relationships in Amazon data: the "Bought-together (BT)" relationship, where users bought item I_i and I_j simultaneously, and the "Also-bought (AB)" relationship, where users who bought item I_i also bought I_j' [20].

4.2 Experimental Setup

We make recommendation based on a single category at a time. We use item title and description for the functional embedding, and item image for the style embedding (which was collected in [24]). For non-complementary item, we randomly select a negative set such that $|\widetilde{C}_i| = |C_i|$. All experiments are trained using Nvidia GeForce GTX Tian X GPU with 12GB memory and 3072 cores using Tensorflow. Since it takes around one week to train a model over the full dataset, we randomly select 11,000 items from each training set as query items to do the five fold cross-validation for model training. We find similar performance between models trained over the full data and this approach [24]. [2]

Baselines. We consider a suite of state-of-the-art baselines. To evaluate the model structure of ENCORE, for fairness, we extend each approach to be trained over the exact same input as ENCORE – images, product text, and ratings:

- *Logistic Regression with Average Rating (LR_A)*: Our first baseline is a straightforward application of logistic regression. We concatenate the differences of images, text, and ratings between queried item I_i and item I_j as input: $\mathbf{f}'_{j|i}(I_i, I_j|\theta_j) = [(\mathbf{m}_i - \mathbf{m}_j)^T, (\mathbf{t}_i - \mathbf{t}_j)^T, \mathbb{E}'(1-\theta_j)]^T$, and calculate the probability that two items are complementary. Here $\mathbb{E}'(1-\theta_j)$ is the average ratings without using Bayesian approach.
- *Logistic Regression with Bayesian Rating (LR_B)*: This variant is similar to the previous but uses the Bayesian ratings inference to find $\mathbb{E}(1-\theta_j|q_j)$ rather than the average ratings. Hence, the input is $\mathbf{f}_{j|i}(I_i, I_j|\theta_j) = [(\mathbf{m}_i - \mathbf{m}_j)^T, (\mathbf{t}_i - \mathbf{t}_j)^T, \mathbb{E}(1-\theta_j|q_j)]^T$.
- *Weighted Nearest Neighbor (WNN)*: This method uses a weighted Euclidean distance to measure complement between items I_i and I_j: $d = ||\mathbf{f}_{j|i}(I_i, I_j|\theta_j) \circ \mathbf{w}||_2^2$ where $\circ$ is Hadamard product and $\mathbf{w}$ is a weight vector.
- *Feedforward Neural Network (FNN)*: We use a 3-layer neural network to measure the non-linear relationships of complement, where the input is the same as in logistic regression LR_B. We use $tanh$ and $softmax$ as activation functions for the second and third layer. We set the hidden and final dimensions to 10 in keeping with the other methods.
- *Low-rank Mahalanobis Transform (LMT) [24]*: This state-of-the-art method uses low-ranked Mahalanobis embedding matrix parameters [24]. Whereas the original approach in [24] relies on images only, we adapt it to use images, text, and ratings. The distance between queried item I_i and item I_j is calculated as $d = ||\mathbf{f}_{j|i}(I_i, I_j|\theta_j)^T \times \mathbf{E}_f||_2^2$ where $\mathbf{E}_f$ is the low-ranked Mahalanobis transform matrix. We set the embedding dimension $K = 10$. [24] also further splits top-categories into smaller categories (such as splitting Clothing into Men's, Women's, Boys, Girls).
- *Monomer [8]*: Another state-of-the-art method – Mixtures of Non-metric Embeddings method [8] that learns low-rank embeddings to uncover different aspects of complementary distance and uses a mixture of experts to find the final complementary distance. We adapt the original method to consider images, text and ratings as input: $\mathbf{f}_i^T = [\mathbf{m}_i^T, \mathbf{t}_i^T, \mathbb{E}(1-\theta_i|q_i)]$ rather than just images for each item I_i. The distance between queried I_i and I_j is calculated by $d = \sum P(n) d_n$ with mixture of weighted experts $P(n)$. $d_n = ||\mathbf{f}_i^T \mathbf{E}_{0f} - \mathbf{f}_j \mathbf{E}_{nf}||_2^2$ where $\mathbf{E}_{nf}$ is the n_{th} embedding matrix. Empirically, we set parameters $K = 10$ and $N = 3$.

Variations of ENCORE. In order to evaluate the impact of images, text, ratings, plus the appropriateness of adopting a neural approach, we consider several variations of our proposed approach:

- ENCORE$_{-M}$: This image-only method is based on [24] and uses the low-ranked Mahalanobis embedding matrix parameters in Equation 1. Note that we do not consider this refinement in any of the following alternatives.
- ENCORE$_{-MT}$: This method combines images and text, while ignoring ratings. The complementary distance between I_i and I_j is $q||(\mathbf{m}_i - \mathbf{m}_j)^T \mathbf{E}_M||_2^2 + \rho||(\mathbf{t}_i - \mathbf{t}_j)^T \mathbf{E}_T||_2^2$. q and ρ is decided by cross-validation with grid search in the range of {0.1, 0.5, 1, 1.5, 10, 100} in each datasets.
- ENCORE$_{-MTCos}$: This method is a simplified version of the previous one, replacing the text-based embeddings with a simpler cosine-based approach over the original text itself.

[2] All code and experimental results are available at: http://people.tamu.edu/~zhan13679/

- ENCORE$_{-MTR}$: This method considers images, text, and ratings, but uses a linear model: $d^{(r)}_{j|i}(I_i, I_j|\theta_j) = [||(\mathbf{m}_i - \mathbf{m}_j)^T \mathbf{E}_M||_2^2,\ ||(\mathbf{t}_i - \mathbf{t}_j)^T \mathbf{E}_T||_2^2,\ \mathbb{E}(1-\theta_j|q)]^T \mathbf{w}$, where $\mathbf{w} \in R^3$ is a model parameter to leverage the contributions of each source of information.
- ENCORE: Finally, we consider the full-blown ENCORE model that incorporates images, text, and ratings in a non-linear model as shown in Equation 6.

Metrics. For each method, we predict whether pairs of items are complementary or not, and measure the accuracy as:

$$ACC := \frac{\sum_i (\sum_j \mathcal{S}(P(l_{ij} \in C_i) - 0.5) + \sum_j \mathcal{S}(0.5 - P(l_{ij} \in \widetilde{C}_i)))}{\sum_i (|C_i| + |\widetilde{C}_i|)},$$

where $\mathcal{S}(\cdot)$ is a thresholding operator defined as: if $x > 0$, then $\mathcal{S}(x) = 1$; otherwise $\mathcal{S}(x) = 0$.

We also use Precision@k to measure the fraction of correctly predicted complementary items for each query item:

$$P@k := \frac{1}{|I|} \sum_i \frac{|GT(C_i) \cap Pred(C_i)@k|}{k}, \tag{8}$$

where $GT(C_i)$ is the ground truth set of items that are complementary with I_i and $Pred(C_i)@k$ is the predicted top-k recommended complementary items.

Parameter settings. For all models, the image and text latent factor dimensions, output dimensions are set to 10 empirically for a trade-off between performance and computational complexity, as well as for fair comparison across methods. For ratings, the threshold is $\eta_r = 3$. Other parameters are fine-tuned for all methods. Particularly, in each experiment, five fold cross-validation is used. Model parameters are first randomly initialized according to truncated normal distributions with mean 0. The standard deviation is decided by grid search in $\{0.1, 0.01, 0.001\}$, and updated by conducting stochastic gradient descent (SGD). The corresponding learning rate is determined by grid search in the range of $\{0.1, 0.05, 0.01, ..., 0.000001\}$. Generally, training for different categories of items converges within 30 iterations.

4.3 Evaluating Complementary Recommendation

We begin by investigating the model quality of ENCORE versus each baseline. Since each approach is built over the same information – images, text, and ratings – we can explore how each approach models and combines these factors for complementary item recommendation. We report the accuracy, precision@5, and precision@10 in Figure 7 for all methods. Table 3 shows the increase of ENCORE comparing with the next-best alternative ("AB" means "Also Bought". "BT" means "Bought Together". Again, here we have modified these original methods to incorporate text and ratings, beyond their original image-only approaches).

Focusing on accuracy (the top row of Figure 7), we observe that ENCORE results in the highest accuracy across both also-bought and bought-together items for all categories except for Books and Digital Music, resulting in an average improvement versus the next-best alternative of 15.5% (ACC row in Table 3). Since Books and Digital Music demonstrate a fairly weak notion of compatibility (e.g., phone chargers match with specific phones, but books of course can be bought with any other books), we see that ENCORE has difficulty, though performing as well as other sophisticated models like LMT and Monomer. Additionally, ENCORE outperforms the next-best alternative of the state-of-the-art LMT and Monomer by 16.9% on average. Since all methods consider same information, these results show the structure of ENCORE introduced via our neural framework results in an even greater improvement. We also observe that LR_A outperforms LR_B, which indicates the number of ratings closely influences a user's preference for complementary items (as in Figure 4).

Next, we focus on precision – see the middle and bottom rows of Figure 7, and increase of ENCORE comparing with the next best baselines in Table 3 row P@5 and P@10. We observe that for all also-bought categories, ENCORE results in the highest precision@5 and precision@10. ENCORE improves versus the next-best alternative an average of 18.4% for precision@5 and 17.8% for precision@10, and versus the best state-of-the-art alternative an average of 27.5% for precision@5 and 26.9% for precision@10 shown in Table 3. Here, we see further evidence of the importance of low-rank embedding and neural transformation in comparison with models like Logistic Regression and Weighted Nearest Neighbors. And for those models that do consider those factors, we see the importance of careful modeling of ratings and integrating each sources information separately in a lower rank non-linear spaces. Observe that precision values are low for all methods in bought-together categories – the data in this case is extremely sparse, with most items having fewer than three ground truth items in the complementary set.

4.4 Impact of ENCORE Model Choices

Given the good performance of ENCORE versus baselines, what impact do the specific design choices have on complementary item recommendation? Does the functional complement derived from text improve upon image-only approaches? Does adding a ratings-based recommender improve the quality of prediction? And what impact does the neural approach have? To anwser those questions, we focus here on accuracy as shown in Table 4; note that similar results hold over precision@5 and precision@10. We additionally report the relative improvement versus the image-only $Encore_M$.

Overall, we see the full-blown ENCORE improves upon all of its variations across all categories, with an average accuracy of 14.0%.

From column ENCORE$_{-MT}$ in Table 4, Text-based evidence is a strong indicator of functional complement in addition to what images can provide, especially in Books (Δ 22%), Digital Music (Δ 32%), and Electronics (Δ 16%) for the bought-together relationship. For also-bought, the relative accuracy improvement is smaller, with Electronics being the one category with worse accuracy (Δ -1%).

Our careful modeling of ratings makes a key positive impact for both also-bought and bought-together items. The accuracy improvement of ENCORE$_{-MTR}$ shows that Electronics, Cell Phones & Accessories, and Digital Music are all highly influenced by user ratings. Indeed, we calculate the average rating for predicted complementary items in Electronics and find that ENCORE$_{-MTR}$ recommends items with ratings higher than ENCORE$_{-M}$ by 4.6%, ENCORE$_{-MTCos}$ by 6.1% and ENCORE$_{-MT}$ by 1.2%.

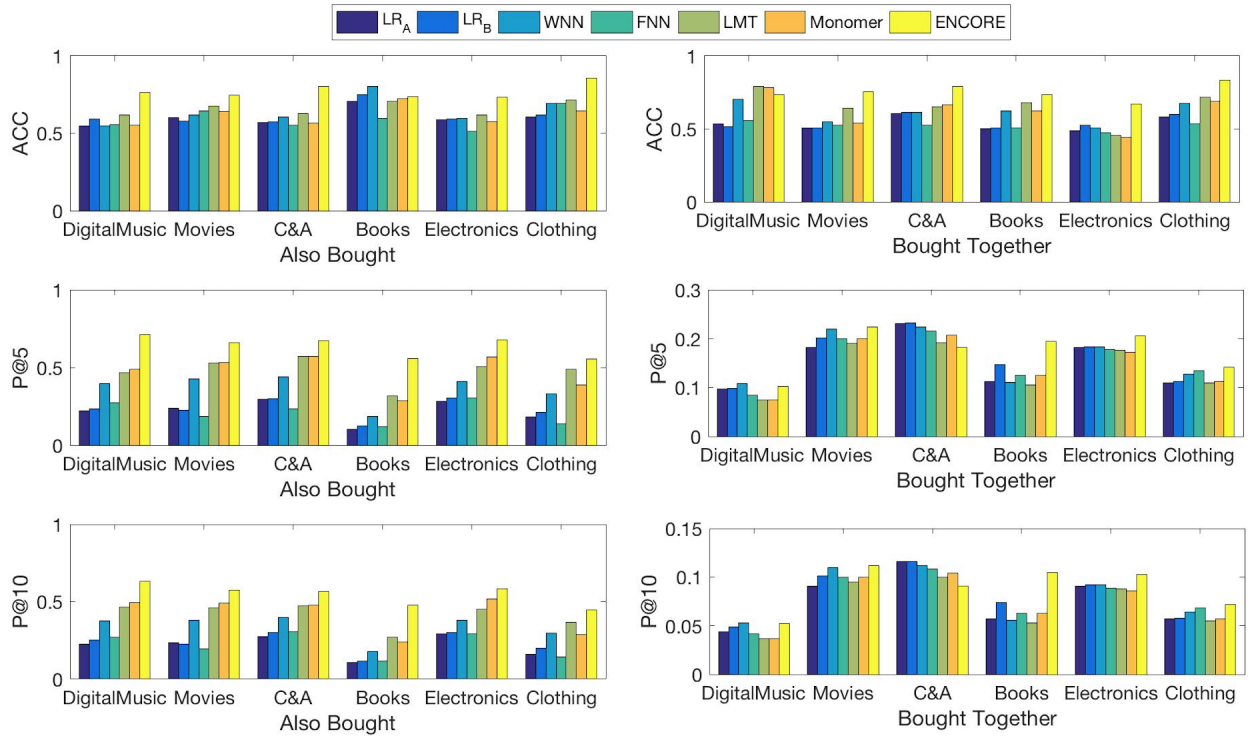

Figure 7: Accuracy and Precision of ENCORE and baselines.

Δ	Digital Music		Movies		C & A		Books		Electronics		Clothing		Average
	AB	BT	AB	BT	AB	BT	AB	BT	AB	BT	AB	BT	
ACC	23%	-7%	10%	17%	29%	19%	-8%	9%	18%	40%	19%	16%	15.5%
P@5	46%	-5%	24%	2%	17%	-22%	76%	33%	20%	12%	13%	4%	18.4%
P@10	27%	-1%	18%	2%	18%	-22%	76%	42%	13%	12%	21%	6%	17.8%

Table 3: Accuracy and Precision increase of ENCORE comparing with the next-best alternative.

Dataset		ENCORE$_{-M}$	ENCORE$_{-MT}$		ENCORE$_{-MTR}$		ENCORE	
		ACC	ACC	Δ	ACC	Δ	ACC	Δ
Digital Music	AB	0.661	0.660	0%	0.755	14%	**0.763**	15%
	BT	0.525	0.693	32%	0.722	37%	**0.738**	40%
Movies	AB	0.686	0.722	5%	0.733	7%	**0.746**	9%
	BT	0.680	0.736	8%	0.748	10%	**0.756**	11%
C & A	AB	0.780	0.791	1%	0.797	2%	**0.806**	3%
	BT	0.722	0.749	4%	0.784	9%	**0.791**	10%
Books	AB	0.702	0.712	1%	0.725	3%	**0.738**	5%
	BT	0.580	0.706	22%	0.726	25%	**0.737**	27%
Electronics	AB	0.713	0.703	−1%	0.712	0%	**0.733**	3%
	BT	0.503	0.583	16%	0.623	24%	**0.670**	33%
Clothing	AB	0.844	0.845	0%	0.855	1%	**0.855**	1%
	BT	0.757	0.810	7%	0.827	9%	**0.833**	10%

Table 4: Prediction accuracy of ENCORE variations. Δ is the change in accuracy compared with ENCORE$_{-M}$. ENCORE outperforms the other methods in each experiment for both also-bought and bought-together relationships.

Finally, we see that the non-linearity of ENCORE plays a significant role to identify complementary relationships. On average, ENCORE results in an improvement of 6.29% for also-bought and 21.91% for bought-together in comparison over $Encore_M$. The impact is especially large in the Electronics categories since the complement relationship is quite complex as discussed.

4.5 ENCORE Recommendations

We also generate predictions by ENCORE for several other items to give additional insights. For a domain like electronics, we see in Figure 8 that ENCORE generates different recommendations for different query items. For example, for the computer in first row, it recommends a keyboard cover, laptop sleeve, and external DVD writer. For an iPhone 5 in last row, ENCORE can recommend iPhone 5 screen protectors and cases.

5 CONCLUSION

In this paper, we have focused on finding "complementary" relationships of items based on user preferences. We proposed a new neural item relationship-based recommender – ENCORE – which carefully combines multiple sources of complement evidence. We saw how stylistic complements (via images) and functional complements (via text-based titles and descriptions) could be combined in a

Figure 8: ENCORE predictions examples for a computer, iPad Air, Camcorder, Camera and iPhone 5. Query items are to the left of the line. Predictions are on the right.

quality-aware framework for uncovering high-quality complementary recommendations. Quantitative and qualitative results show that ENCORE improves upon a state-of-the-art baseline by 15.5% on average, even when all models are built over the exact same input. In our continuing work, we are interested in personalizing ENCORE for considering individual user personality, in addition to the aggregate perspective in the current version. We are also interested to explore more nuanced models of functional complements to improve the quality of our recommendations.

REFERENCES

[1] Yang Bao, Hui Fang, and Jie Zhang. 2014. TopicMF: Simultaneously Exploiting Ratings and Reviews for Recommendation. In *AAAI*.

[2] Iván Cantador, Alejandro Bellogín, and David Vallet. 2010. Content-based recommendation in social tagging systems. In *RecSys*. ACM.

[3] Bradley P Carlin and Thomas A Louis. 2000. *Bayes and empirical Bayes methods for data analysis*. Vol. 17. Chapman & Hall/CRC Boca Raton, FL.

[4] Xu Chen, Zheng Qin, Yongfeng Zhang, and Tao Xu. 2016. Learning to rank features for recommendation over multiple categories. In *SIGIR*. ACM.

[5] Xu Chen, Yongfeng Zhang, Qingyao Ai, Hongteng Xu, Junchi Yan, and Zheng Qin. 2017. Personalized key frame recommendation. In *SIGIR*. ACM.

[6] Mukund Deshpande and George Karypis. 2004. Item-based top-n recommendation algorithms. *ACM Transactions on Information Systems (TOIS)* 22, 1 (2004).

[7] Christian Desrosiers and George Karypis. 2011. A comprehensive survey of neighborhood-based recommendation methods. *Recommender systems handbook* (2011).

[8] Ruining He, Charles Packer, and Julian McAuley. 2016. Learning Compatibility Across Categories for Heterogeneous Item Recommendation. In *ICDM*. IEEE.

[9] Bryan Hooi, Neil Shah, Alex Beutel, Stephan Günnemann, Leman Akoglu, Mohit Kumar, Disha Makhija, and Christos Faloutsos. 2016. Birdnest: Bayesian inference for ratings-fraud detection. In *SDM*. SIAM.

[10] Vignesh Jagadeesh, Robinson Piramuthu, Anurag Bhardwaj, Wei Di, and Neel Sundaresan. 2014. Large scale visual recommendations from street fashion images. In *SIGKDD*. ACM.

[11] Santosh Kabbur, Xia Ning, and George Karypis. 2013. FISM: factored item similarity models for top-n recommender systems. In *SIGKDD*. ACM.

[12] George Karypis. 2001. Evaluation of item-based top-n recommendation algorithms. In *CIKM*. ACM.

[13] Jonghun Kim, Daesung Lee, and Kyung-Yong Chung. 2014. Item recommendation based on context-aware model for personalized u-healthcare service. *Multimedia Tools and Applications* 71, 2 (2014).

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*.

[15] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *ICML*.

[16] Hsin-Hsien Lee and Wei-Guang Teng. 2007. Incorporating multi-criteria ratings in recommendation systems. In *IRI*. IEEE.

[17] Chenyang Li and Kejing He. 2017. CBMR: An optimized MapReduce for item-based collaborative filtering recommendation algorithm with empirical analysis. *Concurrency and Computation: Practice and Experience* 29, 10 (2017).

[18] Dongsheng Li, Chao Chen, Qin Lv, Li Shang, Yingying Zhao, Tun Lu, and Ning Gu. 2016. An algorithm for efficient privacy-preserving item-based collaborative filtering. *Future Generation Computer Systems* 55 (2016).

[19] Ting Li, Anfeng Liu, and Changqin Huang. 2016. A similarity scenario-based recommendation model with small disturbances for unknown items in social networks. *IEEE Access* 4 (2016).

[20] Greg Linden, Brent Smith, and Jeremy York. 2003. Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing* 7, 1 (2003).

[21] Gregory D Linden, Jennifer A Jacobi, and Eric A Benson. 2001. Collaborative recommendations using item-to-item similarity mappings. US Patent 6,266,649 (to Amazon.com).

[22] Yihui Ma, Jia Jia, Suping Zhou, Jingtian Fu, Yejun Liu, and Zijian Tong. 2017. Towards Better Understanding the Clothing Fashion Styles: A Multimodal Deep Learning Approach. In *AAAI*.

[23] Julian McAuley, Rahul Pandey, and Jure Leskovec. 2015. Inferring networks of substitutable and complementary products. In *SIGKDD*. ACM.

[24] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *SIGIR*. ACM.

[25] Julian McAuley and Alex Yang. 2016. Addressing complex and subjective product-related queries with customer reviews. In *WWW*. International World Wide Web Conferences Steering Committee.

[26] Prem Melville, Raymond J Mooney, and Ramadass Nagarajan. 2002. Content-boosted collaborative filtering for improved recommendations. In *AAAI*.

[27] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *ICLR Workshop*.

[28] Xia Ning and George Karypis. 2011. Slim: Sparse linear methods for top-n recommender systems. In *ICDM*. IEEE.

[29] Wei Niu, James Caverlee, and Haokai Lu. 2018. Neural Personalized Ranking for Image Recommendation. In *WSDM*. ACM.

[30] Parivash Pirasteh, Jason J Jung, and Dosam Hwang. 2014. Item-based collaborative filtering with attribute correlation: a case study on movie recommendation. In *Asian Conference on Intelligent Information and Database Systems*. Springer.

[31] Bella Rozenkrants, S Christian Wheeler, and Baba Shiv. 2017. Self-Expression Cues in Product Rating Distributions: When People Prefer Polarizing Products. *Journal of Consumer Research* (2017).

[32] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *WWW*. ACM.

[33] Qusai Shambour, Mou'ath Hourani, and Salam Fraihat. 2016. An item-based multi-criteria collaborative filtering algorithm for personalized recommender systems. *International Journal of Advanced Computer Science and Applications* 7, 8 (2016).

[34] Jianshan Sun, Gang Wang, Xusen Cheng, and Yelin Fu. 2015. Mining affective text to improve social media item recommendation. *Information Processing & Management* 51, 4 (2015).

[35] Peter D Turney. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics.

[36] Zihan Wang, Ziheng Jiang, Zhaochun Ren, Jiliang Tang, and Dawei Yin. 2018. A Path-constrained Framework for Discriminating Substitutable and Complementary Products in E-commerce. In *WSDM*. ACM.

[37] Jian Wei, Jianhua He, Kai Chen, Yi Zhou, and Zuoyin Tang. 2017. Collaborative filtering and deep learning based recommendation system for cold start items. *Expert Systems with Applications* 69 (2017).

[38] Shouxian Wei, Xiaolin Zheng, Deren Chen, and Chaochao Chen. 2016. A hybrid approach for movie recommendation via tags and ratings. *Electronic Commerce Research and Applications* 18 (2016).