

# Projected Hamming Dissimilarity for Bit-Level Importance Coding in Collaborative Filtering

Christian Hansen\*  
University of Copenhagen  
c.hansen@di.ku.dk

Jakob Grue Simonsen  
University of Copenhagen  
simonsen@di.ku.dk

Casper Hansen\*  
University of Copenhagen  
chrh@di.ku.dk

Christina Lioma  
University of Copenhagen  
c.lioma@di.ku.dk

## ABSTRACT

When reasoning about tasks that involve large amounts of data, a common approach is to represent data items as objects in the Hamming space where operations can be done efficiently and effectively. Object similarity can then be computed by learning binary representations (hash codes) of the objects and computing their Hamming distance. While this is highly efficient, each bit dimension is equally weighted, which means that potentially discriminative information of the data is lost. A more expressive alternative is to use real-valued vector representations and compute their inner product; this allows varying the weight of each dimension but is many magnitudes slower. To fix this, we derive a new way of measuring the dissimilarity between two objects in the Hamming space *with* binary weighting of each dimension (i.e., disabling bits): we consider a field-agnostic dissimilarity that projects the vector of one object onto the vector of the other. When working in the Hamming space, this results in a novel projected Hamming dissimilarity, which by choice of projection, effectively allows a binary importance weighting of the hash code of one object through the hash code of the other. We propose a variational hashing model for learning hash codes optimized for this projected Hamming dissimilarity, and experimentally evaluate it in collaborative filtering experiments. The resultant hash codes lead to effectiveness gains of up to +7% in NDCG and +14% in MRR compared to state-of-the-art hashing-based collaborative filtering baselines, while requiring no additional storage and no computational overhead compared to using the Hamming distance.

## KEYWORDS

importance coding; hash codes; collaborative filtering

### ACM Reference Format:

Christian Hansen, Casper Hansen, Jakob Grue Simonsen, and Christina Lioma. 2021. Projected Hamming Dissimilarity for Bit-Level Importance Coding in Collaborative Filtering. In *Proceedings of the Web Conference 2021 (WWW '21)*, April 19–23, 2021, Ljubljana, Slovenia. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3442381.3450011>

\*Both authors share the first authorship.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.

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

© 2021 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.

ACM ISBN 978-1-4503-8312-7/21/04.

<https://doi.org/10.1145/3442381.3450011>

## 1 INTRODUCTION

Hashing-based learning aims to find short binary representations of data objects (called *hash codes*) that allow effective and efficient computations. For tasks involving multiple interacting objects, hash codes must be chosen carefully to ensure that certain properties of the codes (e.g., their inner products or mutual distance) carry information relevant to the task. For example, in hashing-based collaborative filtering, binary user and item representations must be learned so that the distance between them reflects how much user  $u$  likes item  $i$ . Currently, the most efficient way of doing this is to compute their Hamming distance, which is the sum of differing bits between two hash codes. However, by definition, the Hamming distance weighs each bit equally. This is a problem because the importance of the underlying properties encoded by each bit may differ. An alternative is to use real-valued vectors and compute their inner product, which allows varying the weight of each dimension, and thus enables a dimension-specific importance weighting not possible using the Hamming distance for binary codes. However, binary codes allow large storage reductions compared to floating point representations, while the Hamming distance enables massively faster computations compared to the inner product (e.g., real-time brute-force search in a billion items [23]). Motivated by this, we ask: can we introduce bit-level importance weighting on binary representations without compromising efficiency?

We reason that, by the definition of the inner product, the distance between two vectors  $u$  and  $i$  should be identical to the difference in length between vector  $u$  and vector  $i$ 's projection on  $u$ . This observation can be exploited by using a vector space where projections and lengths can be computed several magnitudes more efficiently than in Euclidean space. We show that performing the exact same projection and length computations in the Hamming vector space over  $\mathbb{Z}_2$  results in a novel projected Hamming dissimilarity, which corresponds to traditional measures used in real-valued representations. By choice of projection in the Hamming space, the projected Hamming dissimilarity effectively allows a bit-level binary weighting (corresponding to disabling bits) of  $i$ 's hash code via the hash code of  $u$ , but without decreasing efficiency compared to the Hamming distance. We propose a variational hashing model for learning hash codes optimized for our projected Hamming dissimilarity, which we experimentally evaluate in collaborative filtering experiments. Compared to state-of-the-art baselines using the Hamming distance, we observe effectiveness gains of up to +7% in NDCG and +14% in MRR, while also significantly improving the

convergence rate during model training compared to optimizing with the Hamming distance.

In summary, we **contribute** the first alternative to the Hamming distance for hashing-based learning that allows bit-level importance coding, while requiring no additional storage and no computational overhead. We make our code publicly available.<sup>1</sup>

## 2 RELATED WORK

Matrix factorization is one of the most popular collaborative filtering methods [14], but to reduce storage requirements and speed up computation, hashing-based collaborative filtering has been researched. For hashing-based methods, the users and items are represented as binary hash codes (as opposed to real-valued vectors), which traditionally have used the highly efficient Hamming distance (as opposed to the inner product) for computing user-item similarities. In the following, we review the literature on both binary representation learning and importance coding of such binary representations.

### 2.1 Learning binary representations (hash codes)

Early hashing-based collaborative filtering methods include two stages: First, real-valued representations of the data (vectors) are learned, and then the real-valued vectors are transformed into binary hash codes. Zhang et al. [34] use matrix factorization initially, followed by a binary quantization of rounding the real-valued vectors, while ensuring that the hash code is preference preserving with respect to observed properties of the data using their Constant Feature Norm constraint. Zhou and Zha [35] and Liu et al. [19] both explore binary quantization strategies based on orthogonal rotations of the real-valued vectors, which share similarities with Spectral Clustering [28]. However, the two-stage approaches often suffer from large quantization errors [17, 30], because the hash codes are not learned directly, but rather based on different quantization procedures. More recently, hash codes are learned directly: this has been done using relaxed integer optimization while enforcing bit balancing and decorrelation constraints [30]; and (ii) using an autoencoder to learn the codes in an end-to-end manner [7]. The latter approach is most similar to the variational hashing model proposed in our work for optimizing hash codes for our projected Hamming dissimilarity, but their work is designed for cold-start recommendation based on generalizing hash codes from item descriptions. In contrast, our hashing model is designed to work based purely on user-item ratings, without assuming any additional knowledge of users or items.

### 2.2 Importance coding of binary representations

Approaches for coding importance into hash codes have been studied for the Hamming distance in several applications. Zhang et al. [32] present an image ranking approach that uses a weighted Hamming distance, where bit-level real-valued weights are learned based on both the discriminative power of each bit across all hash codes,

but also dynamically computed based on the hash code used for querying. The bit-level weights are multiplied on each differing bit between two hash codes, such that the distance is computed as the sum of the weights. Different ways of defining bit-level weights have been explored based on fixed weights per bit [27], fixed weights based on byte-level block differences between hash codes [5], and query-adaptive weights [12, 31]. While fixed weights enable faster retrieval than dynamic weights, they all share the same limitation of being significantly less efficient than the Hamming distance, because they can no longer be expressed using highly efficient Boolean hardware-level operations. Furthermore, in addition to the increased storage requirement due to the weights, transferring the weights to the lowest level of memory (i.e., the register) adds additional computational overhead compared to the Hamming distance.

More recent work addresses the problem that hash codes have reduced representational power compared to real-valued vectors, but increasing the hash code dimensionality to match the amount of bits used in the real-valued case hurts model generalization [17]. An alternative, in the task of collaborative filtering, is Compositional Coding for Collaborative Filtering (CCCF) [17], which is a broadly similar method to learning compositional codes for (word) embedding compression [3, 25]. CCCF is a hybrid of hash codes and real-valued weights: each hash code is split into  $k$  blocks of  $r$  bits each, and each block is associated with a real-valued scalar indicating the *weight* of the block. The distance between two CCCF hash codes is then computed as a weighted sum of the Hamming distances of the individual blocks, where each weight is the product of each block's weight. The problem with this approach is that each block requires an individual Hamming distance computation, as well as floating point multiplications of the block weights. In fact, the CCCF block construction no longer allows for highly efficient Boolean operations because the distance computation is weighted by each block's weight.

In contrast to the above approaches, our projected Hamming dissimilarity can exploit the same highly efficient Boolean operations as the Hamming distance, while enabling a bit-level binary weighting on hash codes without reducing efficiency.

## 3 BIT-LEVEL IMPORTANCE CODING IN HASH CODES

### 3.1 Preliminaries

Given two data objects  $u$  and  $i$ , let  $z_u \in \{-1, 1\}^m$  and  $z_i \in \{-1, 1\}^m$  denote their hash codes, where  $m$  is the number of bits in the hash code, which is typically chosen to fit into a machine word. The Hamming distance  $d_H$  between  $z_u$  and  $z_i$  is defined as:

$$d_H(z_u, z_i) = \sum_{j=1}^m 1[z_u^{(j)} \neq z_i^{(j)}] = \text{SUM}(z_u \text{ XOR } z_i) \quad (1)$$

which can be computed very efficiently due to the Boolean operations on the word level, and the SUM which is computed using the *popcnt* bit string instruction. Because Hamming distance is integer-valued, the Hamming distances between several pairs of objects can be linear-time sorted using e.g. radix sort (Hamming distances must be elements of  $[0, m]$ , hence they are bounded) in ascending

<sup>1</sup>The code is available at <https://github.com/casperhansen/Projected-Hamming-Dissimilarity>

order to create a ranked list, allowing for very fast object similarity computation in data-heavy tasks like information retrieval [23].

The above definition clearly shows the efficiency strength of the Hamming distance, but also its weakness in weighting each bit-dimension equally, even though the underlying data properties encoded by some bits may be more discriminative than those of others. In contrast, representations using real-valued vectors and the inner product allow varying the weight of each dimension, and thus enable a dimension-specific importance weighting not possible using the Hamming distance for binary codes. Next, we show how we derive a new way of measuring the dissimilarity between  $u$  and  $i$  in the Hamming space *with* binary weighting of each dimension.

### 3.2 Projected Hamming dissimilarity for bit-level importance coding

Let  $V$  be a vector space over any field  $F$ , and let  $P_{\vec{u}}(\cdot)$  be a projection operator on  $V$ , i.e., for each fixed  $\vec{u}, \vec{i} \in V$ ,  $P_{\vec{u}}(\cdot) : V \rightarrow V$  is a linear map such that

$$P_{\vec{u}}(P_{\vec{u}}(\vec{i})) = P_{\vec{u}}(\vec{i}). \quad (2)$$

In the following, we consider an asymmetric relationship between two objects  $\vec{u}$  and  $\vec{i}$ , such that  $\vec{u}$  is considered a query object used for searching among data items denoted as  $\vec{i}$ . We consider both query and data items as elements of  $V$ . Intuitively, each dimension of  $V$  corresponds to a property of potential importance to a query (e.g., classical music); the projection of each query on the dimension corresponds to the strength of importance, and the projection of each item on the dimension corresponds to how much the item scores on that dimension.

Let  $\|\cdot\| : V \rightarrow \mathbb{R}$  be a norm on  $V$ ; we define the *dissimilarity* between  $\vec{u}$  and  $\vec{i}$ , denoted  $\delta(\vec{u}, \vec{i})$ , as the norm of the projection of  $\vec{i}$  on  $\vec{u}$ :

$$\delta(\vec{u}, \vec{i}) = \|\vec{u} - P_{\vec{u}}(\vec{i})\| \quad (3)$$

Similarly to Eq. 1, the more different  $u$  and  $i$  are, the higher their dissimilarity  $\delta(\vec{u}, \vec{i})$  should be. The dissimilarity is a natural concept: several existing notions of distance or similarity can be seen as special cases of Eq. 3. For example, in the standard Euclidean space the often-used cosine distance<sup>2</sup>,  $1 - \cos(\vec{i}, \vec{u})$ , is one instance of Eq. 3 if we assume unit length vectors.

In hashing-based search, we are particularly interested in the binary vector space  $V = \{-1, 1\}^m$ , with bitwise addition and scalar multiplication over the field  $F = \mathbb{Z}_2$ , where the projection operator is given by:

$$P_{z_u}(z_i) = z_u \text{ AND } z_i \quad (4)$$

for  $z_u, z_i \in V$ , i.e., masking the item hash code  $z_i$  by the query hash code  $z_u$ . Due to working in the Hamming space, the norm  $\|\cdot\|$  is chosen as the Hamming norm (sometimes also called the zero norm or  $L_0$ ). Using this, we obtain the projected Hamming dissimilarity  $\delta_H^P$ , defined as:

$$\delta_H^P(z_u, z_i) = \|z_u - P_{z_u}(z_i)\| = \text{SUM}(z_u \text{ XOR } \underbrace{(z_u \text{ AND } z_i)}_{\text{projection}}) \quad (5)$$

<sup>2</sup>We use the conventional naming of the cosine distance, even though it is not a proper distance as it does not satisfy the triangle inequality.

While having a similar formulation as the Hamming distance (see Eq. 1), the projection of the item hash code  $z_i$  unto the query hash code  $z_u$  in Eq. 5 allows a binary importance weighting of  $z_i$ , which corresponds to disabling unimportant bits as defined by the query hash code  $z_u$  (corresponding to the bit-dimensions where the query hash code is -1). We consider bits to be disabled since a -1 bit in  $z_u$  leads to *all* item hash codes also having a -1 in that bit after the projection. Note that due to the choice of projection, the projected Hamming dissimilarity is asymmetric (i.e., in general  $\delta_H^P(z_u, z_i) \neq \delta_H^P(z_i, z_u)$ ), whereas the Hamming distance is symmetric.

Compared to the Hamming distance, the projected Hamming dissimilarity fundamentally changes the purpose of the query hash code: instead of each dimension encoding a positive or negative preference for a property, it now encodes which properties of the item are important to the query (-1's from the query hash code are copied to the item due to the AND operation). Thus, this formulation can produce query-specific item representations while still only using a single code for each query and item respectively.

**3.2.1 Speeding up the projected Hamming dissimilarity.** The projected Hamming dissimilarity in Eq. 5 requires one additional Boolean operation compared to the Hamming distance. However, because the item codes are static once learned, we can reduce the time complexity to the same as the Hamming distance by observing that:

$$\begin{aligned} \delta_H^P(z_u, z_i) &= \text{SUM}(z_u \text{ XOR } (z_u \text{ AND } z_i)) \\ &= \text{SUM}(z_u \text{ AND } (\text{NOT } z_i)) \end{aligned} \quad (6)$$

where  $(\text{NOT } z_i)$  can be precomputed and stored instead of the original item hash codes, thus requiring no additional storage and the same number of Boolean operations as the Hamming distance.

Next, we present how the projected Hamming dissimilarity can be used for learning hash codes in collaborative filtering, where, notationally, a *user* takes the role of the query, and *items* are to be recommended based on their relevance to the user.

## 4 PROJECTED HAMMING DISSIMILARITY IN COLLABORATIVE FILTERING

We propose a variational hashing model for collaborative filtering that learns user and item hash codes optimized for our projected Hamming dissimilarity. To derive a variational framework for hashing-based collaborative filtering, we define the likelihood of a user  $u$  as the product over the likelihoods of the observed user specified ratings:

$$p(u) = \prod_{i \in I_u} p(R_{u,i}), \quad p(i) = \prod_{u \in U_i} p(R_{u,i}) \quad (7)$$

where  $I_u$  is the set of items rated by user  $u$ , and  $U_i$  is the set of users who have rated item  $i$ . This formulation enforces a dual symmetric effect of users being defined by all their rated items, and items being defined by the ratings provided by all the users. To maximize the likelihood of all observed items and users, we need to maximize the likelihood of the observed ratings  $p(R_{u,i})$ . Note that instead of maximizing the raw likelihood, we consider the log-likelihood to derive the objective below. We assume that the likelihood of a rating,  $p(R_{u,i})$ , is conditioned on two latent vectors: a user hash code  $z_u$ , and an item hash code  $z_i$ . We sample the hash code of

a user or item by performing  $m$  Bernoulli trials, which as a prior is assumed to have equal probability of sampling -1 and 1 ( $p(z_i)$  and  $p(z_u)$  below). This yields the following log-likelihood to be maximized:

$$\log p(R_{u,i}) = \log \left( \sum_{z_i, z_u \in \{-1, 1\}^m} p(R_{u,i} | z_u, z_i) p(z_i) p(z_u) \right) \quad (8)$$

However, this is intractable to compute, hence we derive a lower bound. First, we define the learned approximate posterior distributions for the user and item hash codes as  $q_\phi(z_i|i)$  and  $q_\psi(z_u|u)$ , respectively, where  $\psi$  and  $\phi$  are the distribution parameters. Next, we multiply and divide with the approximate posterior distributions, rewrite to an expectation, and finally apply Jensen's inequality to obtain a lower bound on the log-likelihood:

$$\begin{aligned} \log p(R_{u,i}) &\geq \mathbb{E}_{q_\phi(z_i|i), q_\psi(z_u|u)} \left[ \log [p(R_{u,i} | z_u, z_i)] \right. \\ &\quad \left. + \log p(z_i) - \log q_\phi(z_i|i) + \log p(z_u) - \log q_\psi(z_u|u) \right] \end{aligned} \quad (9)$$

Since  $z_i$  and  $z_u$  will be sampled independently, then  $q_\phi(z_i|i)$  and  $q_\psi(z_u|u)$  are independent and we can rewrite to the variational lower bound:

$$\begin{aligned} \log p(R_{u,i}) &\geq \mathbb{E}_{q_\phi(z_i|i), q_\psi(z_u|u)} \left[ \log [p(R_{u,i} | z_u, z_i)] \right. \\ &\quad \left. - \text{KL}(q_\phi(z_i|i) || p(z_i)) - \text{KL}(q_\psi(z_u|u) || p(z_u)) \right] \end{aligned} \quad (10)$$

where  $\text{KL}(\cdot || \cdot)$  is the Kullback-Leibler divergence. Thus, to maximize the expected log-likelihood of the observed rating, we need to maximize the conditional log-likelihood of the rating, while minimising the KL divergence between the approximate posterior and prior distribution of the two latent vectors. Maximizing the expected conditional log-likelihood can be considered as a reconstruction term of the model, while the KL divergence can be considered as a regularizer.

Next we present the computation of the approximate posterior distributions  $q_\phi(z_i|i)$  and  $q_\psi(z_u|u)$  (Section 4.1) and the conditional log-likelihood of the rating  $p(R_{u,i} | z_u, z_i)$  (Section 4.2).

#### 4.1 Computing the approximate posterior distributions

The approximate posterior distributions can be seen as two encoder functions modelled as embedding layers in a neural network. Each encoder function maps either a user or an item to a hash code. We present below the derivation of the encoder function for the user (the item function is computed in the same way). The probability of the  $j$ 'th bit is given by:

$$q_\psi^{(j)}(z_u|u) = \sigma(E_u^{(j)}) \quad (11)$$

where  $E_u^{(j)}$  is the  $j$ 'th entry in a learned real-valued embedding  $E$  for user  $u$ , and  $\sigma$  is the sigmoid function. The  $j$ 'th bit is then given by:

$$z_u^{(j)} = 2\lceil \sigma(E_u^{(j)}) - \mu^{(j)} \rceil - 1 \quad (12)$$

where  $\mu^{(j)}$  is either chosen stochastically by sampling  $\mu^{(j)}$  from a uniform distribution in the interval  $[0,1]$ , or chosen deterministically to be 0.5 (deterministic choice allows to obtain fixed hash

codes for later evaluation). As the sampling is non-differentiable, a straight-through estimator [1] is used for backpropagation.

#### 4.2 Computing the conditional log-likelihood

The conditional log-likelihood can be considered a reconstruction of the rating, given the user and item hash codes. Similarly to [16], we model the observed ratings as a ground truth rating with additive Gaussian distributed noise, which is then discretized to the observed categorical rating. The conditional log-likelihood can then be computed as:

$$p(R_{u,i} | z_u, z_i) = \mathcal{N}(R_{u,i} - f(z_u, z_i), \sigma^2) \quad (13)$$

where  $f(z_u, z_i)$  is a function that reconstructs the rating given the user and item hash codes. Maximising the log-likelihood  $\log p(R_{u,i} | z_u, z_i)$  corresponds to minimising the mean squared error (MSE) between  $R_{u,i}$  and  $f(z_u, z_i)$ , which is done for training the model. Existing work on hashing-based collaborative filtering [17, 30] also employs a MSE objective, and thus makes the same Gaussian distribution assumption as done here.

We define the reconstruction function to be our proposed projected Hamming dissimilarity:

$$f(z_u, z_i) = g(\delta_H^P(z_u, z_i)) \quad (14)$$

where  $g$  is a fixed affine transformation that maps the interval of the projected Hamming dissimilarity to the interval of the ratings, such that the minimum and maximum of the dissimilarity correspond to the minimum and maximum of the ratings. It should be noted that while variational autoencoders are generative models, we do not explicitly utilize this in the model, as we are primarily concerned with the reconstruction of the observed ratings. This is standard in the related domain of semantic hashing [6, 8, 9, 24].

### 5 EXPERIMENTAL EVALUATION

We evaluate the effectiveness and efficiency of the projected Hamming dissimilarity for bit-level importance coding in hash codes in collaborative filtering experiments, where the task is to recommend relevant items to users. Items and users are represented as learned hash codes, and user-item relevance is approximated by operations (such as the Hamming distance or the projected Hamming dissimilarity) on those hash codes.

#### 5.1 Datasets

We use 4 publicly available datasets commonly used in prior work [15, 17, 18, 20, 30, 33] and summarized in the bottom of Table 1. The datasets comprise two movie rating datasets, Movielens 1M (ML-1M)<sup>3</sup> and Movielens 10M (ML-10M)<sup>4</sup>; a Yelp dataset with ratings of e.g., restaurant and shopping malls<sup>5</sup>; and a book rating dataset from Amazon<sup>6</sup> [10]. Following Rendle et al. [20], we remove users and items with less than 10 ratings. Following Zhang et al. [30], for each user 50% of the ratings are used for testing, 42.5% for training, and the last 7.5% for validation. If a user rates an item multiple times, only the first rating is kept.

<sup>3</sup><https://grouplens.org/datasets/movielens/1m/>

<sup>4</sup><https://grouplens.org/datasets/movielens/10m/>

<sup>5</sup><https://www.yelp.com/dataset/challenge>

<sup>6</sup><http://jmcauley.ucsd.edu/data/amazon/>

## 5.2 Evaluation metrics

We measure the effectiveness with the mean Normalised Discounted Cumulative Gain (NDCG) [11], which is often used to evaluate recommender systems with non-binary ratings (or relevance values). We use the average NDCG at cutoffs {5, 10} over all users and report the average for each cutoff value. We also compute the Reciprocal Rank (RR) of the highest rated item per user, averaged over all users, which represents how well the approaches are at getting a highly relevant item to the top of the ranked list:

$$\text{DCG@k} = \sum_{i=1}^k \frac{2^{\text{rel}_i} - 1}{\log_2(i + 1)}, \quad \text{NDCG@k} = \frac{\text{DCG@k}}{\text{DCG@k}^{(\text{opt})}}, \quad \text{RR} = \frac{1}{\text{rank}} \quad (15)$$

where  $\text{rel}_i$  is the relevance of the item in position  $i$  of the ranked list of items,  $\text{DCG@k}^{(\text{opt})}$  is the DCG@k of the optimal ranking, and rank is the position of the first highest rated item for a user. These measures are averaged across all users, and following the standard notation, the mean reciprocal rank is denoted as MRR.

## 5.3 Baselines

We learn hash codes optimized for the projected Hamming dissimilarity by incorporating it in the variational hashing model as described in Section 4 (denoted  $\text{VH}_{\text{PHD}}$ ). We compare this to standard and state-of-the-art baselines (listed below) that use Hamming distance in different ways. For reference, we also include real-valued Matrix Factorization (MF)<sup>7</sup> [14]. MF is not directly comparable to the hashing approaches (it learns latent real-valued vectors for users and items, instead of hash codes, and computes their inner product), but we include it as a point of reference to a real-valued collaborative filtering baseline. In fact, MF has been shown to be both more efficient and highly competitive in effectiveness compared to neural real-valued collaborative filtering approaches [21]. For a fair comparison of MF to the hashing approaches, we set the latent dimension in MF to be the same as the number of bits used in the hashing approaches. We experiment with hash code lengths of 32 and 64 bits, which correspond to the common machine word sizes. The hashing baselines are:

- **DCF**<sup>8</sup> [30] learns user and item hash codes through binary matrix factorization solved as a relaxed integer problem, while enforcing bit balancing and decorrelation constraints. The codes can be used directly for computing the Hamming distance.
- **CCCF**<sup>9</sup> [17] learns hash codes of  $k$  blocks, where each block has  $r$  bits. A floating point weight is given to each block for computing user-item similarities as a weighted sum of block-level Hamming distances. In [17], the floating point weights are mistakenly not counted towards the amount of bits used, thus leading to an unfair advantage. For a fair comparison, we count each floating point weight only as 16 bits (rather than the typical 32 or 64 bits used for single or double precision, respectively).
- **MFmean** and **MFmedian** are based on matrix factorization (MF), but use each dimension's mean or median for doing the

binary quantization to bits [29]. Similar to DCF, these codes can be used directly for computing the Hamming distance. We include these to highlight the large quantization loss occurring when the hash codes are not learned directly.

- **VH** is the same variational hashing model that we use for learning hash codes to be used with the projected Hamming dissimilarity, but here the codes are learned using the Hamming distance.

## 5.4 Tuning

All hyper parameters for the baselines are tuned using the same set of possible values as in the original papers. For CCCF, we use block sizes of {8, 16, 32, 64} and each floating point weight counts for 16 bits. We try all possible combinations that fit within the bit budget, and if a single block is chosen, then the weight is not included in the bit calculation. Using a Titan X GPU, we train our variational hashing model from Section 4 using the Adam optimizer [13], and tune the learning rate from the set {0.005, **0.001**, 0.0005} and the batch size from the set {100, 200, **400**, 800} (best values in bold). As noise injection has been found beneficial to reduce over-fitting in variational neural models [26], we add Gaussian noise to the ratings during training; we initially set the variance of the Gaussian to 1 and reduce by a factor of  $1 - 10^{-4}$  in every training iteration.

## 5.5 Effectiveness results

Table 1 reports the effectiveness results measured with NDCG and MRR, where the highest NDCG and MRR per column among the hashing-based baselines is shown in **bold**. Results statistically significantly better than the other Hamming distance baselines per column, using a paired two tailed t-test at the 0.05 level and Bonferroni correction, are indicated by an asterisk \*. The Amazon results for CCCF are not included because the released implementation requires >128GB of RAM on this dataset due to the larger amount of items and users.

There are 4 main findings in Table 1: (1) Hash codes optimized for the projected Hamming distance ( $\text{VH}_{\text{PHD}}$ ) outperform all hashing baselines at all times. (2) The gains of  $\text{VH}_{\text{PHD}}$  are larger for MRR than for NDCG, which means that the bit-level importance coding impacts the very top of the ranking list (i.e., the recommendations that matter the most). (3) The best hashing baselines (CCCF, DCF, and VH) have overall similar scores, which indicates a potential ceiling in effectiveness with standard Hamming distance on hash codes. (4) MF with real-valued vectors (i.e., no hash codes) using the inner product outperforms all the hashing approaches, which is to be expected as the representational power of 32/64 floating point numbers is notably higher than that of 32/64 bits. However,  $\text{VH}_{\text{PHD}}$  bridges a large part of the effectiveness gap between the hashing baselines and MF, such that the NDCG differences between  $\text{VH}_{\text{PHD}}$  and MF in 9 out of 16 cases are below 0.01, while the MRR differences in 4 out of 8 cases are close to 0.01.

**5.5.1 Impact of user difficulty.** Given MF as the best performing method, we consider each user's MF performance to be an indicator of difficulty for modeling that particular user. To see how this type of user difficulty impacts recommendation performance, we sort all users (per dataset) increasingly according to their 64-dimensional MF NDCG@10 scores (x axis), and plot the average NDCG@10

<sup>7</sup>Included as baseline in the CCCF repository <https://github.com/3140102441/CCCF>

<sup>8</sup><https://github.com/hanwangzhang/Discrete-Collaborative-Filtering>

<sup>9</sup><https://github.com/3140102441/CCCF>

**Table 1: NDCG@k and MRR scores. \* marks statistically significant gains over the other Hamming distance baselines per column using Bonferroni correction.  $\Delta\%$  shows the gain of  $VH_{PHD}$  over the best hashing-based baseline per column.**

32 bit/dim.	Yelp			ML-1M			ML-10M			Amazon		
	NDCG@5	NDCG@10	MRR	NDCG@5	NDCG@10	MRR	NDCG@5	NDCG@10	MRR	NDCG@5	NDCG@10	MRR
<b>Hamming distance</b>												
CCCF	.7286	.8000	.6250	.6867	.7110	.6493	.5491	.5987	.5683	-	-	-
DCF	.7412	.8095	.6368	.6791	.7092	.6382	.5645	.6120	.5843	.8256	.8714	.7759
MFmean	.6912	.7712	.5815	.5631	.5950	.5053	.4111	.4688	.4271	.7899	.8452	.7342
MFmedian	.6935	.7734	.5769	.5631	.5952	.5085	.4082	.4665	.4225	.7899	.8452	.7343
VH	.7467	.8132	.6473	.6851	.7123	.6419	.5669	.6157	.5815	.8254	.8712	.7758
<b>Projected Hamming dissimilarity</b>												
$VH_{PHD}$	<b>.8036*</b>	<b>.8547*</b>	<b>.7406*</b>	<b>.7135*</b>	<b>.7360*</b>	<b>.6940*</b>	<b>.5939*</b>	<b>.6358*</b>	<b>.6235*</b>	<b>.8479*</b>	<b>.8877*</b>	<b>.8062*</b>
$\Delta\%$	<b>+7.6%</b>	<b>+5.1%</b>	<b>+14.4%</b>	<b>+3.9%</b>	<b>+3.3%</b>	<b>+6.9%</b>	<b>+4.8%</b>	<b>+3.3%</b>	<b>+6.7%</b>	<b>+2.7%</b>	<b>+1.9%</b>	<b>+3.9%</b>
<b>Inner product</b>												
MF	.8071*	.8573*	.7513*	.7352*	.7502*	.7370*	.6029*	.6427*	.6385*	.8586*	.8954*	.8262*
64 bit/dim.	NDCG@5	NDCG@10	MRR	NDCG@5	NDCG@10	MRR	NDCG@5	NDCG@10	MRR	NDCG@5	NDCG@10	MRR
<b>Hamming distance</b>												
CCCF	.7371	.8060	.6329	.7016	.7259	.6716	.5645	.6134	.5837	-	-	-
DCF	.7497	.8155	.6574	.7049	.7285	.6766	.5865	.6316	.6088	.8299	.8747	.7825
MFmean	.6912	.7712	.5810	.5666	.5981	.5172	.4104	.4675	.4257	.7902	.8458	.7340
MFmedian	.6954	.7752	.5780	.5649	.5966	.5105	.4113	.4679	.4270	.7902	.8457	.7334
VH	.7537	.8185	.6561	.7103	.7338	.6759	.5860	.6328	.6013	.8300	.8746	.7828
<b>Projected Hamming dissimilarity</b>												
$VH_{PHD}$	<b>.8075*</b>	<b>.8577*</b>	<b>.7540*</b>	<b>.7267*</b>	<b>.7459*</b>	<b>.7136*</b>	<b>.6034*</b>	<b>.6427*</b>	<b>.6373*</b>	<b>.8521*</b>	<b>.8908*</b>	<b>.8147*</b>
$\Delta\%$	<b>+7.1%</b>	<b>+4.8%</b>	<b>+14.7%</b>	<b>+2.3%</b>	<b>+1.7%</b>	<b>+5.5%</b>	<b>+2.9%</b>	<b>+1.6%</b>	<b>+4.7%</b>	<b>+2.7%</b>	<b>+1.8%</b>	<b>+4.1%</b>
<b>Inner product</b>												
MF	.8096*	.8591*	.7573*	.7424*	.7552*	.7439*	.6188*	.6562*	.6594*	.8586*	.8954*	.8259*
<b>Dataset properties</b>				22,087 users			6,040 users			69,878 users		
				14,873 items			3,260 items			9,708 items		
				602,517 ratings			998,539 ratings			9,995,471 ratings		
				0.18% density			5.07% density			1.47% density		
										158,650 users		
										128,939 items		
										4,701,968 ratings		
										0.02% density		

score per user smoothed by averaging the 500 nearest users (y axis). We do this for the three best Hamming distance baselines (CCCF, DCF, and VH),  $VH_{PHD}$ , and MF, which can be seen in Figure 1.

We observe that  $VH_{PHD}$  outperforms all Hamming distance baselines, showing that the projected Hamming dissimilarity is robust across users. Note that, for the 20,000 users with the lowest MF NDCG@10 on ML-10M, all hashing-based methods outperform MF, highlighting that MF is not always consistently better than hashing-based alternatives (despite allowing for much richer (real-valued versus binary) data representations). In addition, for Yelp and Amazon,  $VH_{PHD}$  obtains near-identical performance as MF for a majority of the users. Interestingly, on Amazon the hashing-based approaches generally perform worse than MF on the [80000, 120000] user interval. We argue that this observation is due to those users not being expressed well by the limited representational power of hash codes using the Hamming distance, compared to real-valued vectors using the inner product, but the projected Hamming dissimilarity reduces a large part of this gap.

**5.5.2 Impact of the number of items per user.** We investigate the impact of a user’s activity as defined by their number of rated items and plot the NDCG@10 scores per user for 64 bit hash codes and 64-dimensional vectors for MF (See Figure 2).

Generally for all methods, we observe higher performance for users with few rated items, and the performance drops as the number of rated items increases. On all datasets except Yelp, the hashing-based baselines perform similar to the real-valued MF initially, but the performance gap occurs as the number of rated items increases, especially for the users with the highest number of rated items for ML-10M and Amazon. The hash codes may perform worse on users with a high number of rated items due to their limited representational power (compared to a real-valued vector), however, using the projected Hamming dissimilarity in our  $VH_{PHD}$  enables to reduce this gap significantly. In fact, our  $VH_{PHD}$  performs almost identically to MF on Yelp and most of the users on ML-10M and Amazon, except those with the highest number of rated items.

**5.5.3 Impact of the average item popularity per user.** We now consider how a user’s average item popularity impacts performance. We denote an item’s popularity as the fraction of users who have rated the item, such that a user’s average item popularity can vary between 0.0 to 1.0, where 1.0 corresponds to only having rated items all other users have rated as well. We order users by their average item popularity and plot the NDCG@10 scores per user for 64 bit hash codes and 64-dimensional vectors for MF (See Figure 3).

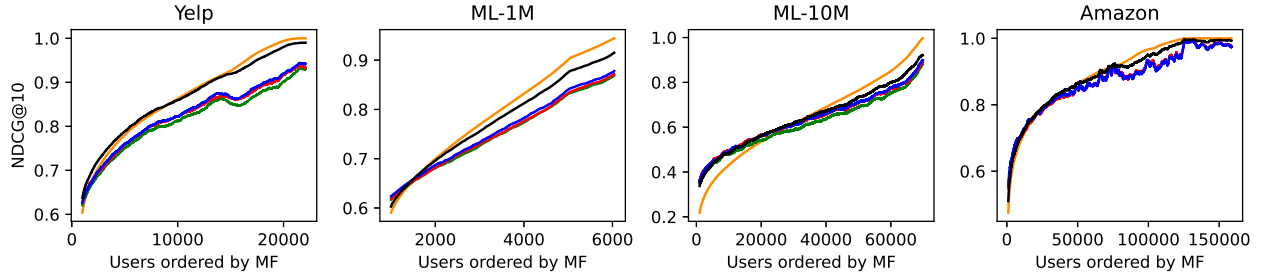


Figure 1: Users are ordered by NDCG@10 for MF and the user-level performances are plotted.

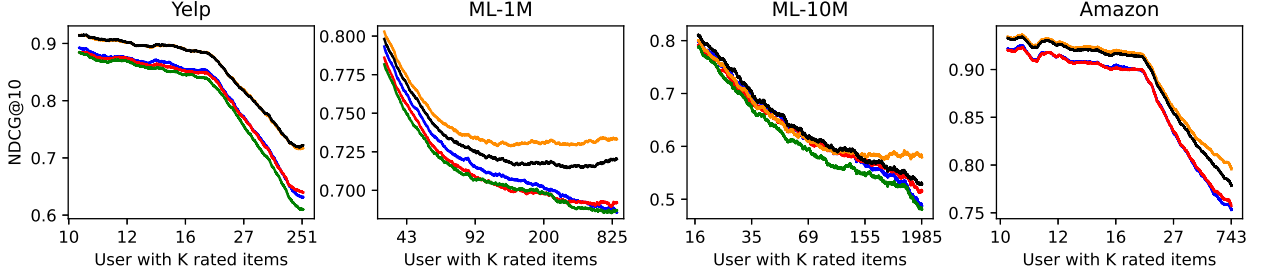


Figure 2: Users are ordered by their number of rated items and the user-level NDCG@10 are plotted.

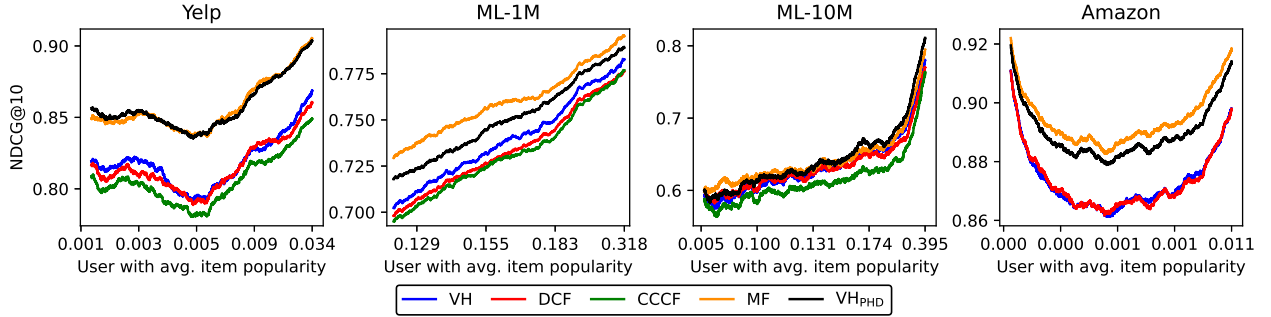


Figure 3: Users are ordered by their average item popularity and the user-level NDCG@10 are plotted.

Generally for all methods, users with a high average item popularity obtain the highest performance scores, whereas users with a lower average item popularity tend to be harder to model. This can be explained by popular items appearing often during training, thus making it easier to learn high quality item representations (both binary and real-valued) matching the properties of those items. Interestingly, on Amazon and partly Yelp, the performance increases for users with the lowest average item popularity. We argue this is primarily due to the high sparsity of those datasets, meaning that some items are rated by very few users, thus making it possible to fit to a small set of user types. Similarly to the analysis on the number of rated items in Section 5.5.2, we overall observe a similar trend that our  $VH_{PHD}$  significantly reduces the gap between the existing hashing-based methods and the real-valued MF.

**5.5.4 Stochastic or deterministic hash codes.** We investigate the effect of the sampling strategy for hash codes (see Eq. 12) during training and evaluation. The sampling can either be deterministic ( $\mu^{(j)} = 0.5$ ) or stochastic ( $\mu^{(j)}$  is sampled uniformly at random from  $[0, 1]$ ), and does not have to be the same for training and evaluation. Figure 4 shows the performance for the four configurations

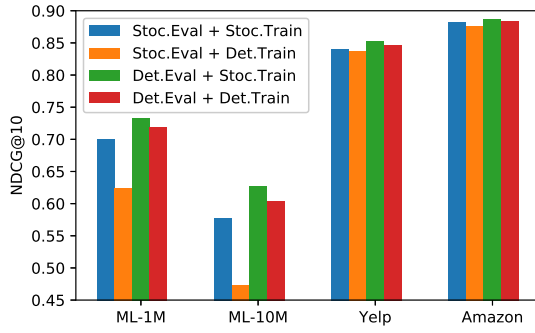
of stochastic sampling or deterministic output across all datasets. We observe that stochastic training with deterministic evaluation consistently performs the best, while deterministic training and deterministic evaluation perform second best. As expected, stochastic sampling at evaluation performs significantly worse than the deterministic option (even more so when trained deterministically), as every item has a small probability of being sampled such that it has a small distance to a user, even though it has a low rating (and vice versa for highly rated items).

## 5.6 Efficiency results

We measure the efficiency of the projected Hamming dissimilarity in terms of convergence rate (when integrated into the variational hashing model) and runtime overhead compared to the Hamming distance.

**5.6.1 Convergence rate.** Figure 5 shows the convergence rate for the variational hashing model using either the Hamming distance or the projected Hamming dissimilarity for 64 bit hash codes. We see that training with the projected Hamming dissimilarity significantly improves the convergence rate compared to the model





**Figure 4: NDCG@10 of VHPHD when varying whether 64 bit hash codes are sampled stochastically or deterministically.**

with the Hamming distance. The time to run a single batch is the same for both the Hamming distance and the projected Hamming dissimilarity (approximately 0.007 seconds for a batch size of 400), from which we conclude that using and optimizing for the projected Hamming dissimilarity not only improves NDCG and MRR, but also drastically reduces training time. We argue that the masking done in the projected Hamming dissimilarity makes the hash codes easier to learn: During training, any update in the item hash code that makes a bit flip will change the distance to all users. However, for the projected Hamming dissimilarity, an update to an item only influences the projected Hamming dissimilarity to a user if the user’s corresponding bit is 1 (as opposed to -1). Thus, the Hamming distance has a global reach for each bit, compared to a more localised reach for the projected Hamming dissimilarity, which effectively makes the hash codes easier to learn.

**5.6.2 Runtime analysis.** The projected Hamming dissimilarity has the same number of Boolean operations as the Hamming distance (see Eq. 1 and 6), due to exploiting that the negation of the item hash codes in the projected Hamming dissimilarity can be precomputed and stored instead of the original item hash codes (i.e., requiring no additional storage). We now verify the actual runtime of both the Hamming distance and the projected Hamming dissimilarity, when considering a fixed set of hash codes. Note that this is a comparison of generated hash codes, thus all approaches using the Hamming distance (or projected Hamming dissimilarity) would have the same runtime in the following experiment. We implement both the Hamming distance and projected Hamming dissimilarity efficiently in C on a machine with a 64 bit instruction set. A test environment was made with 100M randomized 64 bit hash codes, where we measure the time taken to compute 100M Hamming distances and projected Hamming dissimilarities (averaged over 1000 repeated runs). All experiments were run on a single thread<sup>10</sup>, with all hash codes loaded in RAM. The source code was compiled with the highest optimization level utilizing all optimization flags applicable to the hardware.

As reported in Table 2, the mean experiment time was 0.07401 seconds using both the Hamming distance the projected Hamming dissimilarity. Thus, the projected Hamming dissimilarity add no computational overhead, but still allows learning hash codes enabling significant effectiveness improvements as reported in Table

**Table 2: Runtime in seconds and runtime overhead compared to the Hamming distance for 100M computations.**

	Runtime (s)	Runtime overhead
Hamming distance	0.07401	-
Projected Hamming dissimilarity	0.07401	+0.0%
Inner product	4.71414	+6269.6%

1. Finally, Table 2 also reports the runtime of computing the inner product of floating point vectors of length 64: the computation time is 4.71414 seconds, thus being significantly slower than the Hamming space operations.

## 6 CONCLUSION

We presented the projected Hamming dissimilarity, which allows bit-level binary importance weighting (i.e., disabling bits), to produce hash codes that accurately represent dissimilarity between data objects and allow for very efficient subsequent processing. Next, we proposed a variational hashing model for learning hash codes to be optimized for the projected Hamming dissimilarity, and experimentally evaluated it in collaborative filtering experiments. Compared to state-of-the-art hashing-based baselines, we obtained effectiveness improvements of up to +7% in NDCG and +14% in MRR, across 4 widely used datasets. These gains come at no additional cost in storage or recommendation time, as the projected Hamming distance has the same extremely fast computation time as the Hamming distance. Compared to the Hamming distance, we further find that model optimization using the projected Hamming dissimilarity significantly improves the convergence rate, thus speeding up model training.

In future work, we plan to investigate the projected Hamming dissimilarity, and possible adaptations of it, in symmetric retrieval settings consisting of item-item similarity search, as opposed to asymmetric user-item search explored in this work. One such example is document similarity search, which in the hashing setting is known as Semantic Hashing [22], where current work has focused on using the Hamming distance for measuring document similarities [2, 4, 6, 8, 9, 24].

## REFERENCES

- [1] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. 2013. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432* (2013).
- [2] Suthee Chaidaroon and Yi Fang. 2017. Variational Deep Semantic Hashing for Text Documents. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 75–84.
- [3] Ting Chen, Martin Renqiang Min, and Yizhou Sun. 2018. Learning K-way D-dimensional Discrete Codes for Compact Embedding Representations. In *Proceedings of the International Conference on Machine Learning*. 853–862.
- [4] Wei Dong, Qinliang Su, Dinghan Shen, and Changyou Chen. 2019. Document Hashing with Mixture-Prior Generative Models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 5226–5235.
- [5] Bin Fan, Qingqun Kong, Xiaotong Yuan, Zhiheng Wang, and Chunhong Pan. 2013. Learning weighted Hamming distance for binary descriptors. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2395–2399.
- [6] Casper Hansen, Christian Hansen, Jakob Grue Simonsen, Stephen Alstrup, and Christina Lioma. 2019. Unsupervised Neural Generative Semantic Hashing. In *Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval*. 735–744.

<sup>10</sup>We used a Intel Core i9-9940X @ 3.30GHz and had 128GB RAM available.



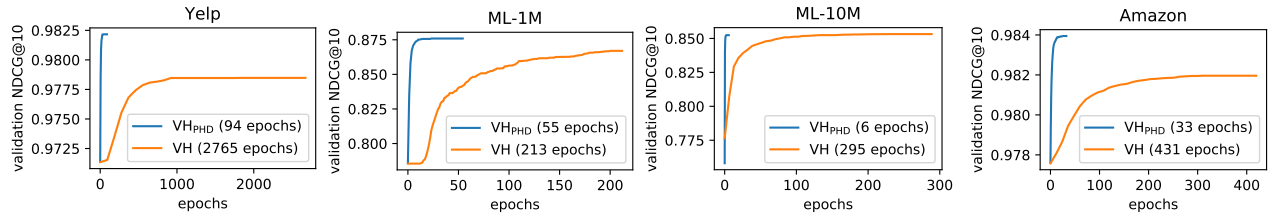


Figure 5: Convergence rate optimizing projected Hamming dissimilarity or Hamming distance.

- [7] Casper Hansen, Christian Hansen, Jakob Grue Simonsen, Stephen Alstrup, and Christina Lioma. 2020. Content-aware Neural Hashing for Cold-start Recommendation. In *Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval*. 971–980.
- [8] Casper Hansen, Christian Hansen, Jakob Grue Simonsen, Stephen Alstrup, and Christina Lioma. 2020. Unsupervised Semantic Hashing with Pairwise Reconstruction. In *SIGIR*. 2009–2012.
- [9] Christian Hansen, Casper Hansen, Jakob Grue Simonsen, Stephen Alstrup, and Christina Lioma. 2021. Unsupervised Multi-Index Semantic Hashing. In *Proceedings of The Web Conference 2021*. In print.
- [10] Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the International Conference on World Wide Web*. 507–517.
- [11] Kalervo Järvelin and Jaana Kekäläinen. 2000. IR evaluation methods for retrieving highly relevant documents. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 41–48.
- [12] Tianxu Ji, Xianglong Liu, Cheng Deng, Lei Huang, and Bo Lang. 2014. Query-adaptive hash code ranking for fast nearest neighbor search. In *Proceedings of the 22nd ACM international conference on Multimedia*. 1005–1008.
- [13] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*.
- [14] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix Factorization Techniques for Recommender Systems. *Computer* 42, 8 (Aug. 2009), 30–37.
- [15] Defu Lian, Rui Liu, Yong Ge, Kai Zheng, Xing Xie, and Longbing Cao. 2017. Discrete Content-aware Matrix Factorization. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 325–334.
- [16] Dawen Liang, Rahul G Krishnan, Matthew D Hoffman, and Tony Jebara. 2018. Variational autoencoders for collaborative filtering. In *Proceedings of the 2018 World Wide Web Conference*. 689–698.
- [17] Chenghao Liu, Tao Lu, Xin Wang, Zhiyong Cheng, Jianling Sun, and Steven C.H. Hoi. 2019. Compositional Coding for Collaborative Filtering. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 145–154.
- [18] Han Liu, Xiangnan He, Fuli Feng, Liqiang Nie, Rui Liu, and Hanwang Zhang. 2018. Discrete Factorization Machines for Fast Feature-based Recommendation. In *Proceedings of the International Joint Conference on Artificial Intelligence*. 3449–3455.
- [19] Xianglong Liu, Junfeng He, Cheng Deng, and Bo Lang. 2014. Collaborative hashing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2139–2146.
- [20] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*. 452–461.
- [21] Steffen Rendle, Walid Krichene, Li Zhang, and John Anderson. 2020. Neural Collaborative Filtering vs. Matrix Factorization Revisited. In *Fourteenth ACM Conference on Recommender Systems (RecSys '20)*. 240–248.
- [22] Ruslan Salakhutdinov and Geoffrey Hinton. 2009. Semantic hashing. *International Journal of Approximate Reasoning* 50, 7 (2009), 969–978.
- [23] Ying Shan, Jie Zhu, JC Mao, et al. 2018. Recurrent binary embedding for gpu-enabled exhaustive retrieval from billion-scale semantic vectors. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2170–2179.
- [24] Dinghan Shen, Qinliang Su, Paidamoyo Chapfuwa, Wenlin Wang, Guoyin Wang, Ricardo Henao, and Lawrence Carin. 2018. NASH: Toward End-to-End Neural Architecture for Generative Semantic Hashing. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. 2041–2050.
- [25] Raphael Shu and Hideki Nakayama. 2018. Compressing Word Embeddings via Deep Compositional Code Learning. In *Proceedings of the International Conference on Learning Representations*.
- [26] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*. 3483–3491.
- [27] Qifan Wang, Dan Zhang, and Luo Si. 2013. Weighted hashing for fast large scale similarity search. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. 1185–1188.
- [28] Stella X. Yu and Jianbo Shi. 2003. Multiclass Spectral Clustering. In *Proceedings of the IEEE International Conference on Computer Vision - Volume 2*. 313–319.
- [29] Dell Zhang, Jun Wang, Deng Cai, and Jinsong Lu. 2010. Self-taught Hashing for Fast Similarity Search. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 18–25.
- [30] Hanwang Zhang, Fumin Shen, Wei Liu, Xiangnan He, Huanbo Luan, and Tat-Seng Chua. 2016. Discrete collaborative filtering. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 325–334.
- [31] Jian Zhang and Yuxin Peng. 2018. Query-adaptive image retrieval by deep-weighted hashing. *IEEE Transactions on Multimedia* 20, 9 (2018), 2400–2414.
- [32] Lei Zhang, Yongdong Zhang, Jinhu Tang, Ke Lu, and Qi Tian. 2013. Binary code ranking with weighted hamming distance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1586–1593.
- [33] Yan Zhang, Defu Lian, and Guowu Yang. 2017. Discrete personalized ranking for fast collaborative filtering from implicit feedback. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*.
- [34] Zhiwei Zhang, Qifan Wang, Lingyun Ruan, and Luo Si. 2014. Preference Preserving Hashing for Efficient Recommendation. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. 183–192.
- [35] Ke Zhou and Hongyuan Zha. 2012. Learning Binary Codes for Collaborative Filtering. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 498–506.