Top-N Recommendation with Missing Implicit Feedback

Daryl Lim
Univ. of California, San Diego
dklim@ucsd.edu

Julian McAuley
Univ. of California, San Diego
jmcauley@ucsd.edu

Gert Lanckriet
Univ. of California, San Diego
gert@ece.ucsd.edu

ABSTRACT

In implicit feedback datasets, non-interaction of a user with an item does not necessarily indicate that an item is irrelevant for the user. Thus, evaluation measures computed on the observed feedback may not accurately reflect performance on the complete data. In this paper, we discuss a missing data model for implicit feedback and propose a novel evaluation measure oriented towards Top-N recommendation. Our evaluation measure admits unbiased estimation under our missing data model, unlike the popular Normalized Discounted Cumulative Gain (NDCG) measure. We also derive an efficient algorithm to optimize the measure on the training data. We run several experiments which demonstrate the utility of our proposed measure.

Categories and Subject Descriptors:
H.3.3 [Information Search and Retrieval]

Keywords: Recommender Systems, Ranking, Evaluation

1. INTRODUCTION

Personalized recommendation of relevant content is a common task in many retrieval systems. Many collaborative filtering approaches [3] attempt to identify user preferences based on *explicit feedback* such as user ratings. However, *implicit feedback*[1], in which a user's preferences are expressed through item interactions such as views or purchases, is often more common than explicit feedback.

In both explicit and implicit feedback systems, the presence of missing data poses a challenge to the evaluation of a recommendation system. In explicit feedback datasets, ratings can be Missing-not-at-Random (MNAR)[8], so systems trained only on observed ratings may give biased predictions. On the other hand, in implicit feedback datasets, non-interaction of a user with an item does not necessarily indicate that the item is irrelevant for the user. If we view unobserved but relevant user-item pairs as missing data, then measures which do not take the missing data mechanism into consideration may also exhibit bias when evaluated on the complete data.

RecSys'15, September 16–20, 2015, Vienna, Austria.

ACM 978-1-4503-3692-5/15/09 ...$15.00.
DOI: http://dx.doi.org/10.1145/2792838.2799671.

To address the MNAR problem in explicit feedback systems, a missing data model was proposed in [8], and it was shown that the Top-N recall and the Area-under-the-Top-N-curve(ATOP) measures evaluated on the observed data provided an unbiased estimate of performance on the complete data under the missing data model. Due to the close relationship between ATOP and AUC, surrogate loss functions to minimize AUC on the training data were proposed.

However, the Top-N recall is known to be difficult to maximize directly, while it has been shown in several recent works ([10, 7, 4]) that optimizing for AUC may not yield the best results on performance measures such as NDCG or MAP which focus on the top of the ranking. Thus, there is a need for a performance measure which admits efficient optimization and is aligned with top-of-the-ranking metrics.

In this work, we first present a missing observation model for implicit feedback data. Next, we present a new performance measure, the Average Discounted Gain (ADG), which focuses on top-of-the-ranking performance and can be estimated without bias on the observed relevance data under our missing data model. Finally, we present an efficient optimization algorithm to optimize the ADG, and evaluate our proposed method on several datasets.

2. DATA MODEL

In our setting, we assume that we are given a set of users $\mathcal{U} = \{u_1, u_2 \cdots u_m\}$ and a set of candidate items $\mathcal{I} = \{i_1, i_2 \cdots i_n\}$. We are also given implicit feedback in the form of a user-item relevance matrix $X \in \mathbb{R}^{|\mathcal{U}| \times |\mathcal{I}|}$ where

$$X(j,k) = \begin{cases} 1, & i_k \text{ is relevant to } u_j \\ 0, & \text{otherwise} \end{cases}$$

Accordingly, we can define the lists of *relevant* and *irrelevant* items for each user:

$$\mathcal{O}_u^+ = \{i_k : X(u, i_k) = 1\}; \quad \mathcal{O}_u^- = \mathcal{I} \setminus \mathcal{O}_u^+.$$

2.1 Generation of observed ratings

Due to the scarcity of resources (for example, time, money or both), users may not be able to consume all items in $\mathcal{I}$ in which they are interested. We therefore assume that each user has a partially observed *prior relevant* set $\mathcal{P}_u^+ \subseteq \mathcal{I}$, which contains *all* items in $\mathcal{I}$ which are relevant to the user. We can then view $\mathcal{O}_u^+ \subseteq \mathcal{P}_u^+$ as a subset of items that a user has chosen to consume (i.e., interact sufficiently with so that it is identified as relevant). In our model, we will assume that the observed items $\mathcal{O}_u^+$ are a simple random sample of unknown size drawn from $\mathcal{P}_u^+$. Equivalently, for a given user

u, each item in $\mathcal{P}_u^+$ has the same (but unknown) probability of being in $\mathcal{O}_u^+$. It may be argued that in real-world settings, such a missing data model may be unrealistic; however, selecting test-set items uniformly from $\mathcal{O}_u^+$ to evaluate implicit feedback methods is a common practice (e.g. [6, 1]).

Our model has close connections to the model in [8] which was originally proposed for explicit data. In fact, when the non-relevant explicit feedback is discarded, the model in [8] is mathematically equivalent to our model, albeit with a different underlying interpretation.

3. AVERAGE DISCOUNTED GAIN (ADG)

In this section, we will present the Average Discounted Gain, a new evaluation measure which can give us an unbiased estimate of performance on $\mathcal{P}_u^+$ under our missing data model. We first assume that we are given user and item sets $\mathcal{U}$ and $\mathcal{I}$, and for each user are given relevant/irrelevant item sets $\mathcal{R}_u^+$ and $\mathcal{R}_u^-$. We also assume a prediction function $f_\theta(u,i)$ parameterized by θ which assigns a score to each user-item pair (u,i). In this paper, we learn a k-dimensional vector for each user and item, as well as a per-item bias:

$$\theta = \{\theta^{user} \in \mathbb{R}^{|\mathcal{U}|\times k}, \theta^{item} \in \mathbb{R}^{|\mathcal{I}|\times k}, \theta^{bias} \in \mathbb{R}^{|\mathcal{I}|}\}$$

and define

$$f_\theta(u,i) = \theta_u^{user} \cdot \theta_i^{item} + \theta_i^{bias}. \tag{1}$$

For a given user u, we define the *rank* of item i under the prediction function $f_\theta(u,i)$ as

$$\text{rank}(i) = \sum_{i' \in \mathcal{I}\setminus i} \mathbf{I}(f_\theta(u,i') - f_\theta(u,i)),$$

where $\mathbf{I}(k) = 1$ if $k > 0$ and 0 otherwise. Essentially, $\text{rank}(i)$ is the number of items (both relevant and irrelevant) with a higher predicted score than item i for a given u. Then, we can define the ADG:

Definition 1. The Average Discounted Gain (ADG) is defined as

$$\frac{1}{|\mathcal{R}_u^+|} \sum_{i^+ \in \mathcal{R}_u^+} \frac{1}{\log_2(\text{rank}(i^+)+2)} \tag{2}$$

where $\mathcal{R}_u^+$ is the set of all relevant items to the user u.

Using this definition, we define the ADG on the observed and complete data respectively:

$$\text{ADG}^{obs} = \frac{1}{|\mathcal{O}_u^+|} \sum_{i^+ \in \mathcal{O}_u^+} \frac{1}{\log_2(\text{rank}(i^+)+2)}$$

$$\text{ADG}^{comp} = \frac{1}{|\mathcal{P}_u^+|} \sum_{i^+ \in \mathcal{P}_u^+} \frac{1}{\log_2(\text{rank}(i^+)+2)}$$

THEOREM 1. *Under the assumption that $\mathcal{O}_u^+$ is a simple random sample from $\mathcal{P}_u^+$, ADG^{obs} is an unbiased estimator of ADG^{comp}.*

PROOF. Given a fixed θ, each relevant item $i_p \in \mathcal{P}_u^+$ is associated with a discounted gain value $\frac{1}{\log_2(\text{rank}(i_p)^+)+2)}$, which depends *only* on the rank of i_p. Now note that every observed item $i_o \in \mathcal{O}_u^+$ has the same rank, and therefore the same discount value, as the corresponding item in $\mathcal{P}_u^+$. Thus if $\mathcal{O}_u^+$ is a random sample from $\mathcal{P}_u^+$, then the mean discounted gain (i.e. ADG) can be estimated without bias. □

In the next section, we show how ADG is related to the NDCG measure, and also show that under our missing data model, NDCG^{obs} is a biased estimator of NDCG^{comp}.

3.1 Comparison with NDCG

The (binary) NDCG with logarithmic discount factor for a user u can be defined as

$$\frac{1}{\text{IDCG}(\mathcal{R}_u^+)} \sum_{i^+ \in \mathcal{R}_u^+} \frac{1}{\log_2(\text{rank}(i^+)+2)}$$

where $\text{IDCG}(k) = \sum_{j=1}^{k} \frac{1}{\log_2(j+2)}$. We can see that the ADG is equivalent to the NDCG with a different per-user weighting function; thus, we expect that the ADG will focus on the top of the ranking just like the NDCG.

THEOREM 2. *Under the assumption that $|\mathcal{O}_u^+|$ is a simple random sample from $\mathcal{P}_u^+$, $NDCG^{obs}$ is an unbiased estimator of $NDCG^{comp}$ only when $|\mathcal{O}_u^+| = |\mathcal{P}_u^+|$.*

PROOF. First note that

$$\text{NDCG}^{comp} = \frac{1}{\text{IDCG}(|\mathcal{P}_u^+|)} \sum_{i^+ \in \mathcal{P}_u^+} \frac{1}{\log_2(\text{rank}(i^+)+2)},$$

then

$$\begin{aligned}
\mathbb{E}\left[\text{NDCG}^{obs}\right] &= \frac{|\mathcal{O}_u^+| \cdot \mathbb{E}\left[\frac{1}{|\mathcal{O}_u^+|} \sum_{i \in \mathcal{O}_u^+} \frac{1}{\log_2(\text{rank}(i^+)+2)}\right]}{\text{IDCG}(|\mathcal{O}_u^+|)} \\
&= \frac{|\mathcal{O}_u^+|}{\text{IDCG}(|\mathcal{O}_u^+|)} \text{ADG}^{comp} \quad \text{(from Theorem 1)} \\
&= \frac{|\mathcal{O}_u^+| \cdot \text{IDCG}(|\mathcal{P}_u^+|)}{|\mathcal{P}_u^+| \cdot \text{IDCG}(|\mathcal{O}_u^+|)} \text{NDCG}^{comp}
\end{aligned}$$

which is only unbiased when $|\mathcal{O}_u^+| = |\mathcal{P}_u^+|$. □

Since $\mathcal{O}_u^+ \subseteq \mathcal{P}_u^+$, this means that NDCG^{obs} will always be a biased estimate of NDCG^{comp} unless the user consumes *all* items in $\mathcal{P}_u^+$.

4. OPTIMIZATION

We now present an efficient algorithm to optimize the ADG for a given dataset. Since the ADG is bounded between 0 and 1, instead of maximizing the ADG, we will minimize $(1 - \text{ADG})$. First, we note that

$$\begin{aligned}
1 - \text{ADG} &= 1 - \frac{1}{|\mathcal{O}_u^+|} \sum_{i^+ \in \mathcal{O}_u^+} \frac{1}{\log_2(\text{rank}(i^+)+2)} \\
&= \frac{1}{|\mathcal{O}_u^+|} \sum_{i^+ \in \mathcal{O}_u^+} \mathcal{C}(\text{rank}(i^+)) \qquad (3)
\end{aligned}$$

where

$$\mathcal{C}(k) = 1 - \frac{1}{\log_2(k+2)}. \tag{4}$$

It can be shown (omitted for brevity) that $\forall k \in \{1 \cdots |\mathcal{I}|\}$,

$$\mathcal{C}(k) = \sum_{1}^{k} \alpha_k, \;\; \exists\, \vec{\alpha} \in \{\vec{\alpha} \in \mathbb{R}^{|\mathcal{I}|} : \alpha_1 > \alpha_2 \cdots > \alpha_{|\mathcal{I}|} > 0\}.$$

Thus, we can transform $\text{rank}(i^+)$ into a loss, and use the approximation

$$\mathcal{C}(\text{rank}(i^+)) \approx \sum_{i^- \in \mathcal{V}_{u,i^+}} \mathcal{C}(|\mathcal{V}_{u,i^+}|) \frac{f_\theta(u,i^-) - f_\theta(u,i^+) + 1}{|\mathcal{V}_{u,i^+}|} \tag{5}$$

(see [10] for a related derivation) where

$$\mathcal{V}_{u,i^+} = \{i^- \in (\mathcal{I} \setminus i^+) : f_\theta(u, i^-) - f_\theta(u, i^+) + 1 > 0\}.$$

Finally, we substitute Eq. (5) into Eq. (3), to get the final optimization problem:

$$\min_\theta \frac{1}{|\mathcal{U}|} \sum_U \sum_{\substack{i^+ \in \mathcal{O}_u^+ \\ i^- \in \mathcal{V}_{u,i^+}}} \mathcal{C}(|\mathcal{V}_{u,i^+}|) \frac{f_\theta(u, i^-) - f_\theta(u, i^+) + 1}{|\mathcal{O}_u^+||\mathcal{V}_{u,i^+}|}.$$

We follow a similar procedure to [10] to derive a stochastic gradient descent algorithm, and also use the early-stopping technique in [4] to speed up the optimization process. The pseudocode for the full algorithm is given in Algorithm 1.

Algorithm 1 The OPT-ADG algorithm

Input: user set $\mathcal{U}$, item set $\mathcal{I}$, relevance sets $\{\mathcal{O}_u^+ : u \in \mathcal{U}\}$

repeat
  Sample u uniformly from $\mathcal{U}$, i^+ uniformly from $\mathcal{O}_u^+$
  $N = 0$
  violatorFound = False
  repeat
    Sample i^- uniformly from $\mathcal{I} \setminus i^+$
    if $f_\theta(u, i^+) - f_\theta(u, i^-) < 1$ **then**
      violatorFound = True; $v = i^-$
      break
    end if
    $N = N + 1$
  until $N >= \frac{|\mathcal{I}|-1}{\gamma}$
  if violatorFound **then**
    Take gradient step on

$$\mathcal{C}\left(\left\lfloor \frac{|\mathcal{I}|-1}{N} \right\rfloor\right)(f_\theta(u, v) - f_\theta(u, i^+) + 1) \quad \text{(Eq. (4))}$$

  end if
until max iterations exceeded

5. DISCUSSION

One purported advantage of measures like the ATOP and the ADG is that their performance on $\mathcal{O}_u^+$ gives us an unbiased estimate of performance on $\mathcal{P}_u^+$ (henceforth, we shall refer to them as unbiased-to-missing-data (UBM) measures).

However, in practice, we cannot directly make use of this property if the ranking model to be evaluated is trained on data in $\mathcal{O}_u^+$, since items in $\mathcal{O}_u^+$ are no longer a random sample with respect to the ranking model. Thus, we cannot extrapolate performance on $\mathcal{P}_u^+$ by measuring performance on $\mathcal{O}_u^+$, as this is analogous to guessing test set performance based on training performance in a classification setting.

Nevertheless, we note that UBM measures still retain a nice property: if *a priori*, some relevant items per user are held out (i.e. not used for training by the ranking model) in disjoint test and validation sets which are both uniform random samples from $\mathcal{O}_u^+$, then we can expect the validation and test set performance to be similar *regardless of the number of validation or test items held out.*

Our claim is easy to prove: If we denote the relevant items in the validation set as $\mathcal{R}_{u,\text{val}}^+$ and the relevant items in the test set as $\mathcal{R}_{u,\text{test}}^+$, then we can view both $\mathcal{R}_{u,\text{val}}^+$ and $\mathcal{R}_{u,\text{test}}^+$ as uniform random samples from the set $\mathcal{R}_{u,\text{val}}^+ \cup \mathcal{R}_{u,\text{test}}^+$. Therefore from Theorem 1, we can expect that if we evaluate the validation and test sets with respect to any UBM measure, they would both yield unbiased estimates of performance on $\mathcal{R}_{u,\text{val}}^+ \cup \mathcal{R}_{u,\text{test}}^+$, *regardless of* $|\mathcal{R}_{u,\text{val}}^+|$ *and* $|\mathcal{R}_{u,\text{test}}^+|$. Using the same logic, if we are given an observation set $\mathcal{O}_u^+$ and prior observation set $\mathcal{P}_u^+$, then splitting $\mathcal{O}_u^+$ into $\mathcal{O}_{u,train}$ and $\mathcal{O}_{u,test}$, training a ranker on $\mathcal{O}_{u,train}$ then evaluating $\mathcal{O}_{u,test}$ on any UBM measure should yield similar performance to the same UBM measures given to $\{\mathcal{P}_u^+ \setminus \mathcal{O}_u^+\}$, which are exactly the unknown but relevant items we want to predict. Here, the intuition is somewhat analagous to the generalization ability of classifiers in classical machine learning settings when the validation and test set come from the same distribution.

Another desirable property of UBM methods is allowing us to make statements about the *absolute performance* of ranking models: For example, ADG can be interpreted as the mean discounted gain of relevant items. In contrast, we cannot predict the absolute performance of a ranking model on non-UBM measures such as NDCG on $\{\mathcal{P}_u^+ \setminus \mathcal{O}_u^+\}$ without knowing $|\{\mathcal{P}_u^+ \setminus \mathcal{O}_u^+\}|$.

6. EXPERIMENTS

To evaluate the performance of our proposed measure, we conducted experiments on 3 datasets: *Amazon Games*, a subset of customer reviews from the Video Games category on Amazon, MovieLens 10M data, and *last.FM* listening data for 110000 users from the Million Song Dataset Challenge hosted on Kaggle. For the Amazon Games and MovieLens data, we binarized the data and treated the 4 and 5 star reviews as relevant and the rest as irrelevant. For the last.FM data, we considered a song relevant to a user if the user listened to it at least 3 times, and irrelevant otherwise. Due to the sparsity of each dataset, we also densified the data by retaining the most popular items and users with the most reviews. Our datasets are summarized below.

Dataset	Users	Items	Interactions	Sparsity
last.FM	10000	10000	97727	0.097%
MovieLens	9888	5000	711084	1.44%
Amazon Games	17437	17915	201154	0.064%

To show the utility of optimizing for ADG over AUC, we implemented two similar matrix factorization methods, MF-ADG and MF-AUC, both with $f_\theta(u, i)$ defined as in Eq. (1). MF-ADG uses Algorithm 1, while MF-AUC tries to optimize the empirical AUC for each user by solving

$$\min_\theta \frac{1}{|\mathcal{U}|} \sum_U \frac{1}{|\mathcal{O}_u^+||\mathcal{O}_u^-|} \sum_{i+ \in \mathcal{O}_u^+} \sum_{i- \in \mathcal{O}_u^-} [f_\theta(u, i^-) - f_\theta(u, i^+) + 1]_+$$

where $[\cdot]_+ = \max(0, \cdot)$.

For each user, 10% of the relevant items were used for the validation set, while 20% of the relevant items were used for the test set, and both were uniformly sampled from $\mathcal{O}_u^+$. This process was repeated four times to create four folds and the mean performance was reported. For fairness, both methods were initialized with the same random parameters, and each algorithm was run for 1000000 iterations. We regularized the ℓ_2-norms of both the user and item latent vectors, and used a single regularization parameter λ whose optimal value was determined by performance on the validation set. The number of latent factors per item and user was fixed at 50, and for MF-ADG, the value of γ was fixed at 100. For each method, we report three UBM measures,

ATOP, ADG, Recall@10) and also two popular ranking measures, Mean Average Precision (MAP) and NDCG. As there was negligible difference between ATOP and AUC in our experiments (<0.1%) we chose to only report ATOP in the paper. As a baseline, we also computed the rank-k SVD on the user-item relevance matrix for different values of k for each dataset, but do not report the results as the performance even for the best value of k was significantly worse than both MF-ADG and MF-AUC on all metrics.

6.1 Results

Table 1 shows the performance of both methods on all experiments. For each dataset, MF-ADG performed better than MF-AUC on all ranking measures except ATOP, which is expected because of the close link between the ATOP objective and AUC optimization. This supports our claim that optimizing ADG on the training set improves performance at the top of the ranking.

	Amazon Games		last.FM		MovieLens	
	MF-AUC	MF-ADG	MF-AUC	MF-ADG	MF-AUC	MF-ADG
ATOP	0.7584 (0.0014)	0.7546 (0.0049)	0.7490 (0.0064)	0.7449 (0.0028)	0.8855 (0.0018)	0.8821 (0.0042)
MAP	0.0104 (0.0003)	**0.0124 (0.0003)**	0.0242 (0.0006)	**0.0281 (0.0006)**	0.0775 (0.0007)	**0.0858 (0.0003)**
NDCG	0.1460 (0.0006)	**0.1501 (0.0004)**	0.1701 (0.0007)	**0.1750 (0.0008)**	0.3718 (0.0010)	**0.3820 (0.0002)**
rec@10	0.0170 (0.0004)	**0.0211 (0.0007)**	0.0473 (0.0010)	**0.0539 (0.0019)**	0.0945 (0.0011)	**0.1025 (0.0013)**
ADG	0.1080 (0.0005)	**0.1110 (0.0003)**	0.1294 (0.0005)	**0.1332 (0.0006)**	0.1714 (0.0004)	**0.1768 (0.0001)**

Table 1: Mean performance on all datasets across 4 folds. The number in brackets is the standard error of the mean. Methods which performed significantly better are bolded.

Table 2 shows the mean performance of both methods on the test and validation subsets of the MovieLens dataset respectively. As discussed in Section 5, we can see that the UBM measures (ATOP, REC@10 and ADG) show broadly consistent performance across the test and validation sets, while the MAP and NDCG measures vary greatly. Similar observations were made for the other two datasets which we have omitted due to space constraints. This supports our claim that measuring these performance measures on a validation set can allow us to make confident predictive statements about the performance of the model on the unseen test data, even when the number of test items is unknown.

7. RELATED WORK

Many predictive models have been proposed for both explicit feedback [3] and implicit feedback [1, 6]. [5, 2] have studied the MNAR assumption in terms of model fitting with different missing data models but the evaluations do not take the missing data models into account. Furthermore, to the best of our knowledge, no one has formally proposed a missing data model for implicit feedback models. Our work is most closely related to [8, 9].

8. CONCLUSION

In this work, we proposed a missing data model for implicit feedback and a novel evaluation measure which allows unbiased estimation with respect to our missing data model. We also showed that ranking models trained to maximise our evaluation measure have improved performance on top-of-the-ranking measures. In future work, we plan to explore different models of missing data generation.

	MF-AUC			MF-ADG		
Measure	Test	Valid	Diff%	Test	Valid	Diff%
ATOP	0.8855	0.8849	-0.06	0.8821	0.8817	-0.05
ADG	0.1714	0.1709	-0.29	0.1768	0.1768	-0.00
REC@10	0.0945	0.0945	0.00	0.1025	0.1030	0.49
MAP	0.0775	0.0586	-24.38	0.0858	0.0657	-23.43
NDCG	0.3718	0.2957	-20.47	0.3820	0.3046	-20.42

Table 2: Test vs. validation performance on MovieLens dataset. Performance on UBM measures is consistent across test and validation sets.

9. ACKNOWLEDGMENTS

The authors acknowledge support from Yahoo!, Inc., the Sloan Foundation, and NSF Grants CCF-0830535 and IIS-1054960. Daryl Lim was supported by a fellowship from the Agency for Science, Technology and Research (A*STAR), Singapore.

10. REFERENCES

[1] Y. Hu, Y. Koren, and C. Volinsky. Collaborative filtering for implicit feedback datasets. In *Proc. IEEE ICDM (2008)*, pages 263–272, 2008.

[2] Y. Kim and S. Choi. Bayesian binomial mixture model for collaborative prediction with non-random missing data. In *RecSys '14*, pages 201–208, 2014.

[3] Y. Koren, R. M. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *IEEE Computer*, 42(8):30–37, 2009.

[4] D. Lim and G. Lanckriet. Efficient learning of mahalanobis metrics for ranking. In *Proc. ICML 2014*, pages 1980–1988, 2014.

[5] B. M. Marlin and R. S. Zemel. Collaborative prediction and ranking with non-random missing data. In *RecSys '09*, pages 5–12, 2009.

[6] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme. BPR: bayesian personalized ranking from implicit feedback. In *UAI 2009*, pages 452–461, 2009.

[7] Y. Shi, A. Karatzoglou, L. Baltrunas, M. Larson, A. Hanjalic, and N. Oliver. Tfmap: optimizing map for top-n context-aware recommendation. In *Proc. ACM SIGIR*, 2012.

[8] H. Steck. Training and testing of recommender systems on data missing not at random. In *Proc. ACM SIGKDD*, pages 713–722, 2010.

[9] H. Steck. Evaluation of recommendations: rating-prediction and ranking. In *RecSys '13*, pages 213–220, 2013.

[10] J. Weston, S. Bengio, and N. Usunier. Large scale image annotation: learning to rank with joint word-image embeddings. *Machine Learning*, 81:21–35, 2010.