The 46th International ACM SIGIR Conference on Research and Development in Information Retrieval

Collaborative Residual Metric Learning

Tianjun Wei, Jianghong Ma, Tommy W.S. Chow

July 26, 2023
Taipei

Collaborative Filtering

- Utilize **user** historical **interactions** with **items** to produce recommendations
- Consider the **implicit** feedback (Clicking and Purchasing)
  - User : $u \in U$
  - Item: $i \in I$
  - Interactions: $(u, i) \in T$
- Top-K recommendation:
  - Recommend K items from the item set I that user u has not interacted with before.

Interaction Matrix

- $\boldsymbol{R} \in \mathbb{R}^{|U| \times |I|}$

$$R_{ui} = \begin{cases} 1, & (u, i) \in T \\ 0, & Otherwise \end{cases}$$

Background

Metric Learning in Collaborative Filtering

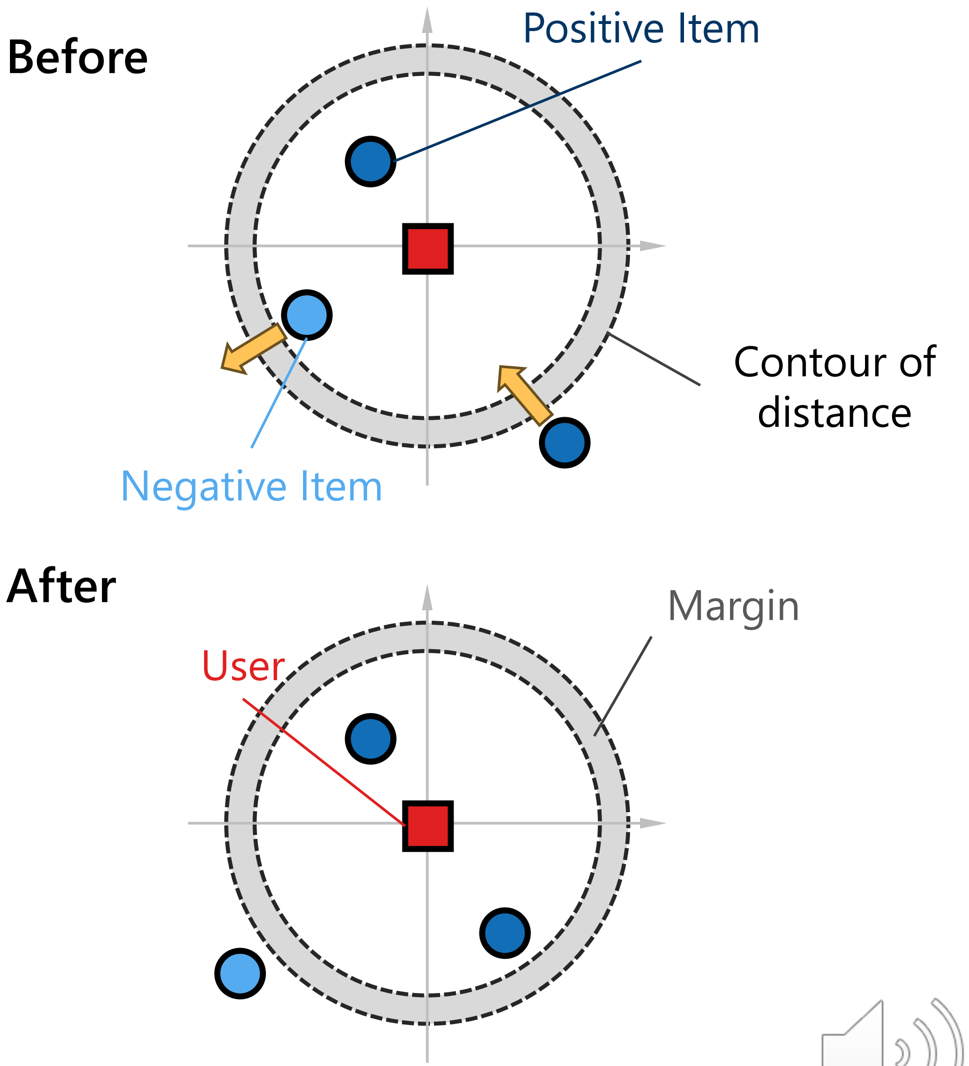
Metric Learning

- Learn a valid distance metric to:
  - Decrease the distance between similar points
  - Increase the distance between dissimilar points
- A valid distance metric satisfies triangle inequality:

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) \leq d(\boldsymbol{x}_i, \boldsymbol{x}_k) + d(\boldsymbol{x}_k, \boldsymbol{x}_j).$$

Collaborative Metric Learning (CML)

- Consider interacted items for a user as similar points
- Use Euclidean distance as the metric
- Update the latent representations of users and items other than distance metric



Cheng-Kang Hsieh, Longqi Yang, Yin Cui, Tsung-Yi Lin, Serge Belongie, and Deborah Estrin. 2017. Collaborative Metric Learning. In WWW'17

Signal-based Models

Generalized Mahalanobis Distance

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^T \boldsymbol{W} (\boldsymbol{x}_i - \boldsymbol{x}_j)}$$

- $\boldsymbol{W}$ must be symmetric positive semi-definite to satisfy triangle inequality
- Generalized Mahalanobis distance is equivalent to Euclidean distance when $\boldsymbol{W} = \boldsymbol{I}$

Representing User-item Interactions as Signals

- Defines the space with the same number of dimensions as $|I|$: $\boldsymbol{P} = \boldsymbol{D}_I^{-t} \boldsymbol{R}^T, \boldsymbol{Q} = \boldsymbol{D}_I^t$
- Calculate a preference score by weight matrix $\boldsymbol{C} \in \mathbb{R}^{|I| \times |I|}$: $y_{ui} = \boldsymbol{p}_u^T \boldsymbol{C} \boldsymbol{q}_i$

Different variants by specifying t

- Linear autoencoder ($\boldsymbol{t} = \boldsymbol{0}$):

$$\min_{\mathrm{C}} \frac{1}{2} ||\boldsymbol{R} - \boldsymbol{R}\boldsymbol{C}||_F^2, \qquad s.t.\, diag(\boldsymbol{C}) = 0$$

- Graph signal model ($\boldsymbol{t} = \boldsymbol{0.5}$):

$$\boldsymbol{Y} = \boldsymbol{R}\boldsymbol{D}_I^{-\frac{1}{2}} (\boldsymbol{V}\boldsymbol{V}^T) \boldsymbol{D}_I^{\frac{1}{2}} \quad (\boldsymbol{C} = \boldsymbol{V}\boldsymbol{V}^T)$$

Harald Steck. 2019. Embarrassingly Shallow Autoencoders for Sparse Data. In WWW'19.
Chao Chen, Dongsheng Li, Junchi Yan, Hanchi Huang, and Xiaokang Yang. 2021. Scalable and Explainable 1-Bit Matrix Completion via Graph Signal Learning. In AAAI '21.

Incorporate Signal-based Models with ML

Ensuring the PSD property of weight matrix

- Graph filtering model: Always satisfied ($\boldsymbol{C} = \boldsymbol{V}\boldsymbol{V}^T$)
- Linear autoencoder: Almost impossible to satisfy because of the **diagonal zero** constraint, $diag(\boldsymbol{C}) = 0$

Focus instead on the difference of distances

- Recommendation task focuses on the relative relationship of preference scores (distance)

$$
\begin{aligned}
\Delta d^2 &= d^2(\boldsymbol{p}_u, \boldsymbol{q}_i) - d^2(\boldsymbol{p}_u, \boldsymbol{q}_j) \\
&= \boldsymbol{q}_i^T \boldsymbol{W} \boldsymbol{q}_j - \boldsymbol{q}_j^T \boldsymbol{W} \boldsymbol{q}_j - 2\boldsymbol{p}_u^T \boldsymbol{W}(\boldsymbol{q}_i - \boldsymbol{q}_j) + (\boldsymbol{p}_u^T \boldsymbol{W} \boldsymbol{p}_u - \boldsymbol{p}_u^T \boldsymbol{W} \boldsymbol{p}_u) \\
&= W_{ii}(d_i^{2t} - 2R_{ui}) - W_{jj}(d_j^{2t} - 2R_{uj}) - 2\underbrace{\boldsymbol{p}_u^T \boldsymbol{H}}_{y_{ui} \text{ in linear autoencoder}}(\boldsymbol{q}_i - \boldsymbol{q}_j)
\end{aligned}
$$

where $\boldsymbol{H} = \boldsymbol{W} - diag(\boldsymbol{W})$ is called the *Hollow matrix.*

- Eliminate redundant terms in the distances
- Separate diagonal and non-diagonal entries

Incorporate Signal-based Models with ML

After separating diagonal and non-diagonal values of $\boldsymbol{W}$, we have proved the following theorem:

> Suppose the i-th row sum of absolute value in $\boldsymbol{H}$ is $h_i = \sum_{1 \le j \le n} |H_{ij}|$,
> for any $\boldsymbol{H}$, there exists an $\omega = \max_{1 \le i \le n} \frac{h_i}{d_i^{-2t}}$, such that $\boldsymbol{W} = \boldsymbol{H} + \omega \boldsymbol{D}_I^{-2t}$ is PSD.

Then, we can find the relationships between the distances and the preference scores derived by $\boldsymbol{H}$.

$$\Delta d^2 = W_{ii}\left(d_i^{2t} - 2R_{ui}\right) - W_{jj}\left(d_j^{2t} - 2R_{uj}\right) \; - \; 2\boldsymbol{p}_u^T \boldsymbol{H}(\boldsymbol{q}_i - \boldsymbol{q}_j)$$

$$\frac{1}{2}\Delta d^2 = -\omega(d_i^{-2t} R_{ui} - d_j^{-2t} R_{uj}) - \underbrace{\boldsymbol{p}_u^T \boldsymbol{H}(\boldsymbol{q}_i - \boldsymbol{q}_j)}_{\tilde{y}_{uij}\text{ preference residual}}$$

I. When $R_{ui} = R_{uj} = 0$ (both items are uninteracted)

- $\Delta d^2 \cong -\tilde{y}_{uij}$, the distance relationship of uninteracted items is reflected by $\tilde{y}_{uij}$
- Critical to the model **prediction** process

II. When $R_{ui} = 1, R_{uj} = 0$ (only i is interacted)

- $\Delta d^2 + \omega d_i^{-2t} \cong -\tilde{y}_{uij} \Rightarrow$ when $y_{ui} = y_{uj}$, $d(\boldsymbol{p}_u, \boldsymbol{q}_i)$ is smaller than $d(\boldsymbol{p}_u, \boldsymbol{q}_j)$ by a positive margin
- Useful in the model **training** process

How does t affect the recommendations?

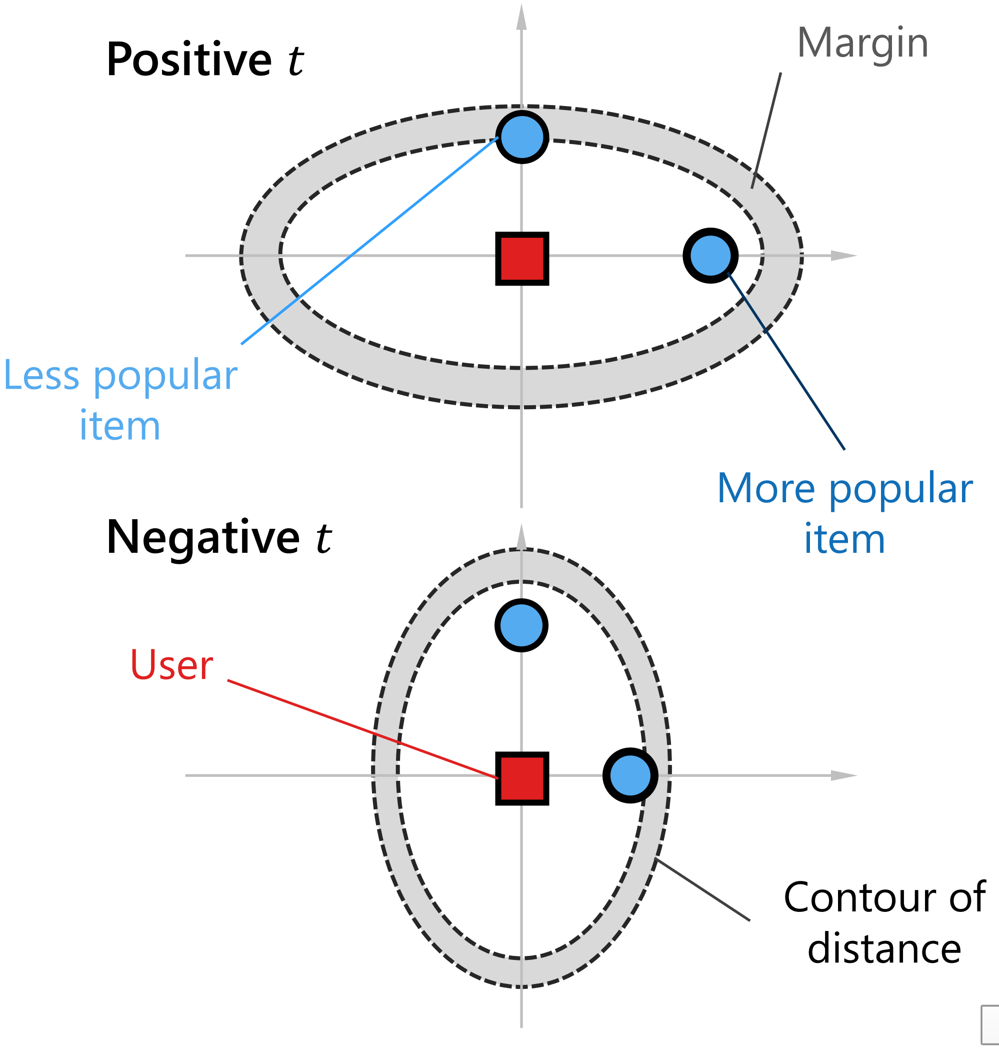
$$\frac{1}{2}\Delta d^2 = -\omega(d_i^{-2t}R_{ui} - d_i^{-2t}R_{uj}) - \boldsymbol{p}_u^T\boldsymbol{H}(\boldsymbol{q}_i - \boldsymbol{q}_j)$$

III. When $R_{ui} = R_{uj} = 1$ (both items are uninteracted)
$\Delta d^2 \cong -\tilde{y}_{uij}$, only when $t = 0$
When $t \neq 0$, preference score will introduce bias when representing the distance relationship of interacted items

Normalization strength t

$$y_{ui} = \boldsymbol{p}_u^T(d_i^t \cdot \boldsymbol{c}_i)$$

- Larger t will increase the chance of popular items to be recommended when regularization (i. e. l_2) is acted on $\boldsymbol{C}$
- Improve the **novelty** of the recommendations of signal-based models by decreasing $\boldsymbol{C}$



Collaborative Residual Metric Learning (CoRML)

Triplet Residual Margin Loss

Triplet Margin Loss:

$$L = \sum_{(u,i^+,i^-)} (d_{ui^+} - d_{ui^-} + \zeta)_+$$

- Replace $d_{ui^+} - d_{ui^-}$ with $-\tilde{y}_{ui^+i^-} + \omega d_{i^+}^{-2t}$
- Set ζ to $\omega d_{i^+}^{-2t}$ as an active margin

$$L_{TRM} = \sum_{(u,i^+,i^-)} (-\tilde{y}_{ui^+i^-})_+$$

$$= \sum_u \sum_{i^+} \alpha_{ui^+} y_{ui^+} + \sum_{i^-} \beta_{ui^-} y_{ui^-}$$

α and β are coefficients determined by the ranking of the value of y_{ui} for a user u.

Approximated Ranking Weights

- α and β are dynamically updated during model training
- Use y_{ui} to approximate α and β

$$\tilde{\alpha}_{ui^+} = \phi_u y_{ui^+} - 1, \tilde{\beta}_{ui^-} = \phi_u y_{ui^-}$$

- Scaling factor

$$\phi_u = \epsilon \left(\frac{d_u}{\max_u d_u} \right)^{-t_u}$$

  - Global scaling: scale y_{ui^+} to obtain a negative $\tilde{\alpha}_{ui^+}$
  - User-degree scaling: reduce the effects of different number of non-zero entries in $\boldsymbol{p}_u$ of each user

Loss Function

$$L_{CoRML} = \sum_u \sum_i y_{ui}(\phi_u y_{ui} - R_{ui}) = tr((\boldsymbol{Y}^T(\boldsymbol{\Phi Y} - \boldsymbol{R})))$$

Hybrid Preference Score

$$\boldsymbol{Y} = \boldsymbol{R}(\lambda \underline{\boldsymbol{D}_I^{-t}\boldsymbol{H}\boldsymbol{D}_I^{-t}} + (1-\lambda)\underline{\boldsymbol{D}_I^{-\frac{1}{2}}\boldsymbol{G}\boldsymbol{D}_I^{\frac{1}{2}}})$$

Extension from linear autoencoder with adjustable t

Extension from graph signal model

$$\boldsymbol{G} = \left(\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}} - \mathrm{diag}\left(\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}\right)\right)_+$$

Optimization Problem

$$\min_{\boldsymbol{H}} tr\left(\boldsymbol{Y}^T(\boldsymbol{\Phi Y} - \boldsymbol{R})\right),$$
$$s.t.\quad diag(\boldsymbol{H}) = 0, \boldsymbol{H} \geq 0, \boldsymbol{H} = \boldsymbol{H}^T$$

- Optimized through Alternating Directions Method of Multipliers (ADMM)

CoRML

Experiments

Performance Comparison

Dataset

- 4 real-world public datasets

Evaluation Metrics

- NDCG@K
- MRR@K

Baselines

- CML models
- Signal-based Models
- GCN models

Dataset	Metric	CML	L-CML	DPCML	SLIM	EASE	RecVAE	GFCF	UltraGCN	SimGCL	CoRML
Pinterest	NDCG@5	0.0509	0.0594	0.0563	0.0488	0.0558	0.0516	0.0620	0.0572	0.0616	***0.0655**
	NDCG@10	0.0665	0.0766	0.0724	0.0630	0.0704	0.0668	0.0785	0.0729	0.0783	***0.0824**
	NDCG@20	0.0897	0.1021	0.0965	0.0841	0.0921	0.0895	0.1031	0.0962	0.1031	***0.1076**
	MRR@5	0.1018	0.1186	0.1133	0.0957	0.1125	0.1024	0.1239	0.1146	0.1237	***0.1306**
	MRR@10	0.1164	0.1343	0.1283	0.1084	0.1262	0.1164	0.1390	0.1292	0.1390	***0.1458**
	MRR@20	0.1261	0.1444	0.1381	0.1171	0.1353	0.1258	0.1488	0.1387	0.1488	***0.1556**
Gowalla	NDCG@5	0.0853	0.0985	0.0999	0.1100	0.1211	0.0890	0.1174	0.1108	0.1229	***0.1317**
	NDCG@10	0.0953	0.1093	0.1087	0.1156	0.1268	0.0978	0.1257	0.1181	0.1295	***0.1383**
	NDCG@20	0.1125	0.1281	0.1261	0.1302	0.1412	0.1140	0.1440	0.1348	0.1460	***0.1554**
	MRR@5	0.1533	0.1743	0.1811	0.1912	0.2186	0.1613	0.2121	0.2001	0.2235	***0.2334**
	MRR@10	0.1682	0.1899	0.1957	0.2043	0.2323	0.1752	0.2269	0.2144	0.2377	***0.2479**
	MRR@20	0.1768	0.1984	0.2040	0.2118	0.2393	0.1832	0.2352	0.2225	0.2454	***0.2558**
Yelp2018	NDCG@5	0.0483	0.0574	0.0556	0.0535	0.0611	0.0525	0.0587	0.0585	0.0646	***0.0690**
	NDCG@10	0.0521	0.0617	0.0592	0.0554	0.0628	0.0558	0.0617	0.0621	0.0676	***0.0716**
	NDCG@20	0.0629	0.0742	0.0709	0.0644	0.0722	0.0663	0.0731	0.0737	0.0795	***0.0832**
	MRR@5	0.1007	0.1188	0.1156	0.1117	0.1277	0.1106	0.1236	0.1234	0.1349	***0.1435**
	MRR@10	0.1149	0.1345	0.1304	0.1245	0.1413	0.1247	0.1380	0.1385	0.1499	***0.1586**
	MRR@20	0.1241	0.1443	0.1399	0.1327	0.1496	0.1336	0.1472	0.1478	0.1594	***0.1679**
ML-20M	NDCG@5	0.2319	0.2731	0.2620	0.2785	0.3025	0.3045	0.2718	0.2365	0.2675	***0.3189**
	NDCG@10	0.2326	0.2689	0.2588	0.2710	0.2934	0.3033	0.2671	0.2280	0.2644	***0.3103**
	NDCG@20	0.2486	0.2832	0.2725	0.2813	0.3036	0.3204	0.2799	0.2369	0.2794	***0.3212**
	MRR@5	0.3761	0.4341	0.4190	0.4478	0.4829	0.4777	0.4356	0.3919	0.4310	***0.4967**
	MRR@10	0.3932	0.4494	0.4347	0.4621	0.4963	0.4923	0.4506	0.4063	0.4466	***0.5098**
	MRR@20	0.4002	0.4554	0.4409	0.4677	0.5014	0.4976	0.4566	0.4124	0.4527	***0.5149**

Detailed Analysis

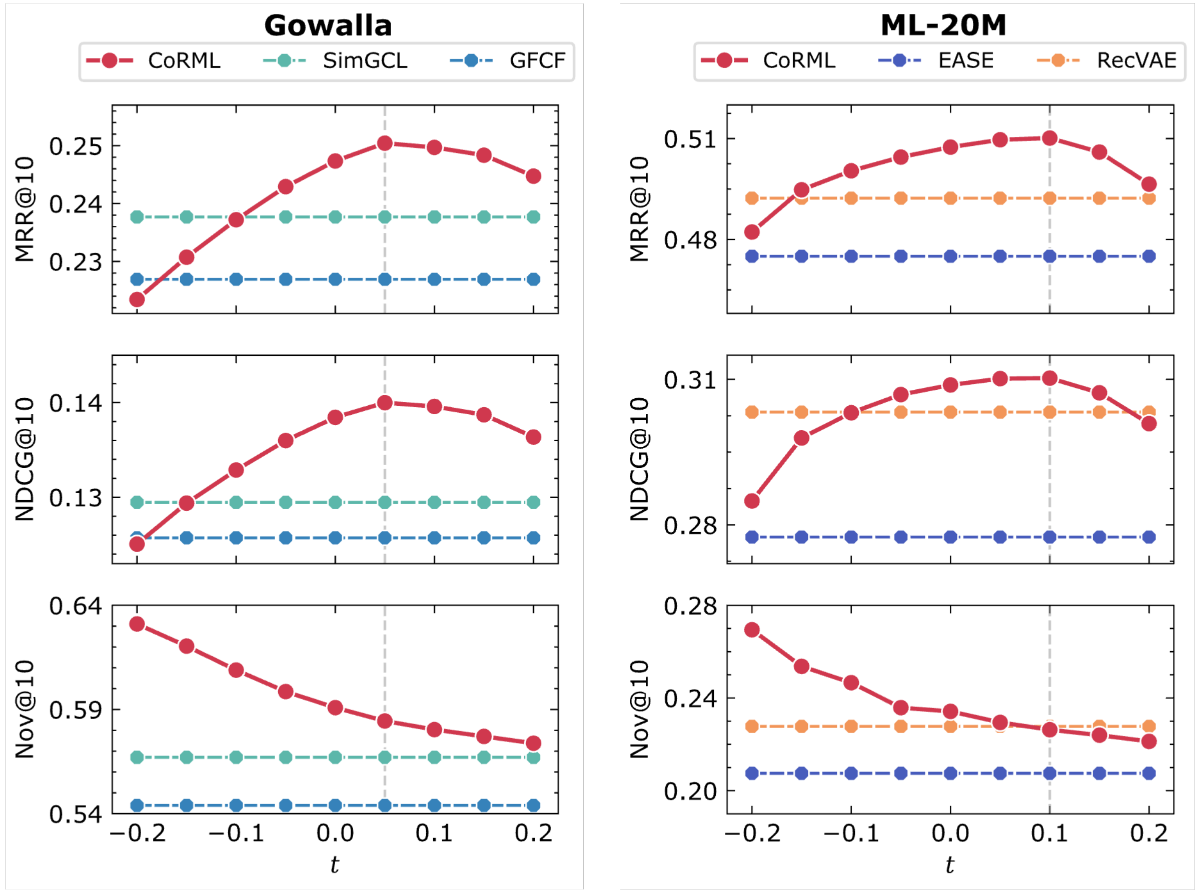
Mitigating Popularity Bias

Novelty

- Measure the popularity of top-K items recommended

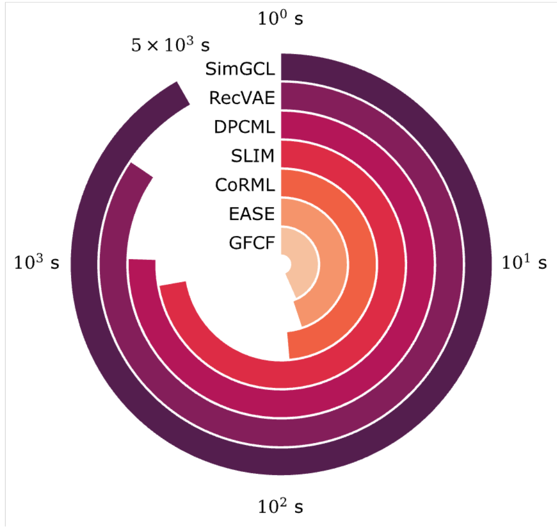
$$Nov@K = \frac{1}{|U|K}\sum_{u}^{|U|}\sum_{i}^{K} -\frac{1}{\log_2|U|}\log_2\frac{d_i}{|U|}$$

- Smaller t increases the novelty of recommendation
- Accuracy and novelty are not just trade-off

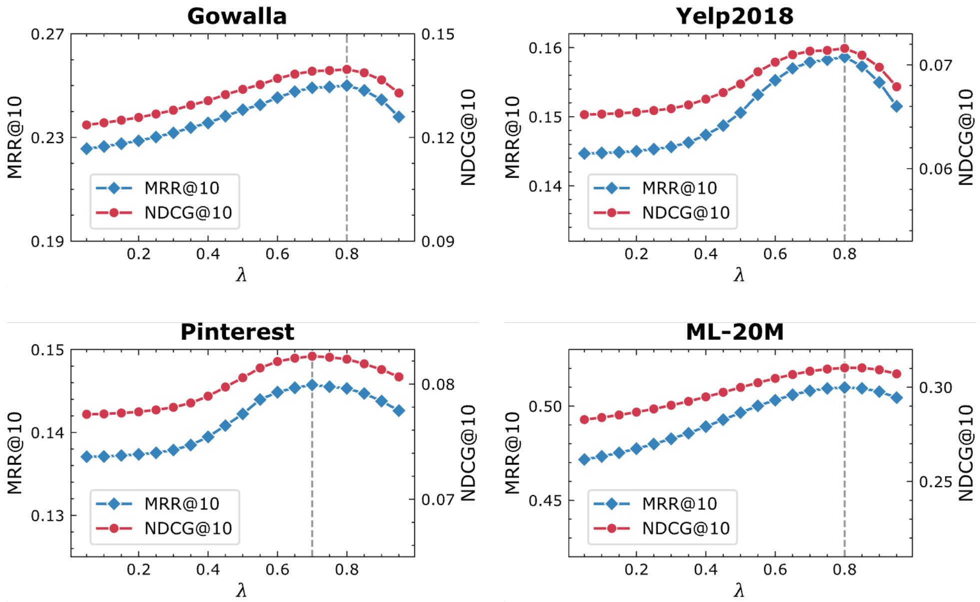


Tao Zhou, Zoltán Kuscsik, Jian-Guo Liu, Matúš Medo, Joseph Rushton Wakeling, and Yi-Cheng Zhang. 2010. Solving the apparent diversity-accuracy dilemma of recommender systems. Proceedings of the National Academy of Sciences 107, 10 (2010), 4511–4515.

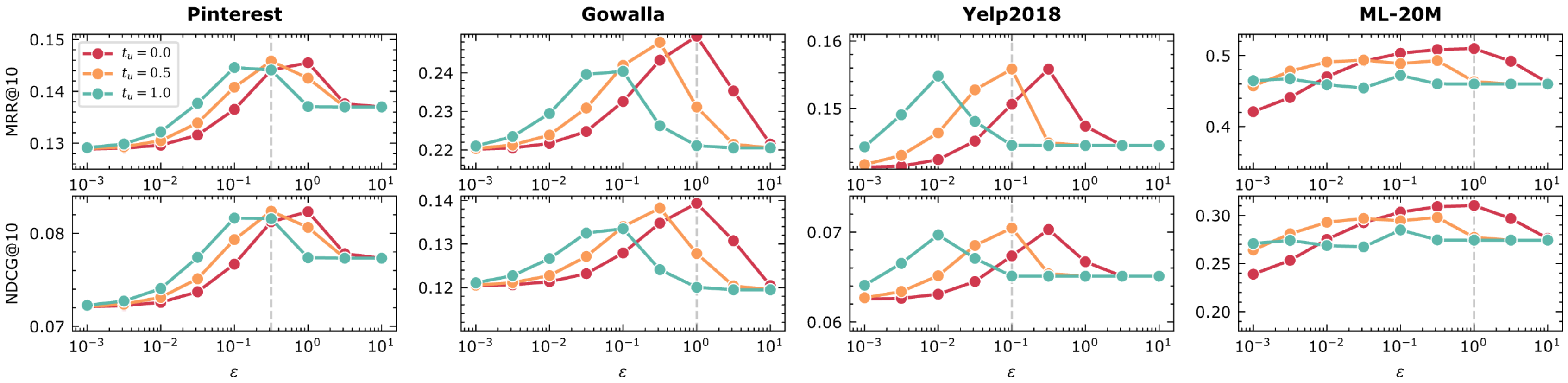
Detailed Analysis



Training Time



Effect of λ



Effect of scaling factors

Thank you!

The code is available at GitHub: *Joinn99/CoRML*

Paper

Presenter:

Tianjun Wei

City University of Hong Kong

tjwei2-c@my.cityu.edu.hk

Code