浙江大学
ZHEJIANG UNIVERSITY

How to write a paper?

Jiawei Chen

Oct 2024

Why paper written is so important?

- Easily get accepted
  --- easily to read, making the reviewers get key contributions
- Making the paper more influential
  --- Papers that are highly cited are always well-written
- Experience to improve the ability to express opinions
  --- the ability to promote ideas and logical rigor

Different levels of paper-written

第一境界
利剑

手拿好的idea，
写的一言难尽，
差点被拒

第二境界
软剑

胡丽花哨，
堆了很多东西，
不着重点

第三境界
重剑

重剑无锋，
大巧不工

第四境界
木剑

不在拘泥于领域

第五境界
无剑

？

How to write an introduction

❑基本宗旨(Principles)：简洁性(concise) and 易读性(readable)

- First paragraph: the importance and background of the research topic.
  - E.g., Recommender systems (RS) have been widely applied in various personalized services ….
  - Not easy. Should be **tailored** for the paper
    --- top-k optimization, should emphasize this thing
    --- diversity, personalized would be highlighted

How to write an introduction

❑基本宗旨(Principles)：简洁性(concise) and 易读性(readable)

- Second paragraph: the limitation of recent studies
  ---- new task? the importance of this task, better with examples
  ---- new model? The limitations of existing methods

- Point-by-point to illustrate the limitations
  ---- e.g., existing methods can be categorized into the XX types.
  ---- e.g., existing methods suffer from the following limitations.

Recent efforts have proposed *surrogate losses* (Lapin et al., 2016; 2017) to optimize NDCG@K, yet these approaches exhibit significant limitations:

How to write an introduction

❑基本宗旨(Principles)：简洁性(concise) and 易读性(readable)

- Third paragraph: (make a summarize of limitations) and introduce the proposed method

  - One sentence to illustrate the key motivation
    --- e.g., to address xxx, we propose to xxx
    --- e.g., the key challenge lies on xxx. To xx, we propose xxx
  - Introduce the proposed method
    --- **how does the proposed method address these limitations**
    --- imply the novelty
    --- (theoretical) analyses to show the advantages (optimal)

How to write an introduction

❑基本宗旨(Principles)：简洁性(concise) and 易读性(readable)

- Fourth paragraph: make a summarize of the experiments (optional)

  - Give some impressive and informative results
    --- e.g., percent of improvements
    --- e.g., do something interesting experiments
    --- e.g., covering xx scenarios

- Claim contributions

How to write a method

❑基本宗旨(Principles)：严谨性(rigorous) and 易读性(readable)

- One section (preliminary) to give notations, problem definition and some important analyses

  --- definition and notations are important; only those that are utilized should be adopted.

  --- analyses are also important; to give the motivations of the proposed methods; to give the challenges or limitations

How to write a method

❑基本宗旨(Principles)：严谨性(rigorous) and 易读性(readable)

- One section (methodology) to detail the proposed method
  - give a short introduction of this section, or give a bird-view of the proposed method.
  - must be logical; point-by-point
    --- e.g., we propose xx, leveraging xx strategies:
    --- e.g., our proposed method consists of xx parts
  - Do not forget the analyses
    --- insights behind theorems
    --- the benefits of the proposed methods

How to write a method

❑基本宗旨(Principles)：严谨性(rigorous) and 易读性(readable)

- Some skills
  - Lemmas are companied with analyses
  - List the advantages of the proposed methods
  - Give the key words before the paragraph

4.3 Discussion

We show that the proposed Adap-τ satisfies the following three desirable properties:

Personalization. As for user-wise adaption, owing to the capability of *Superloss*, our model could calculate the specific τ in terms of their cumulated loss. The larger train loss suggests the data may contain more noises, and thus drives the model to be more conservative. τ would become larger accordingly to slow the pace of hard-mining and down-weigh the contribution of this user.

Adaption. As for data-wise adaption, our Equation (9) automatically computes proper τ without extra manual intervention, thus avoiding the notorious hyper-parameter search for τ.

How to write a experiment

❑基本宗旨(Principles)：防御性（defensive）

- No Bug in the experiments
- List the research questions (optimal)
- Experimental setup
  - Datasets: sufficiently large? Conventional?
  - Metrics: conventional?
  - Baselines: SOTA methods? Give the years of the publications
  - Hyperparameters: details both for the proposed method and compared method, refer to appendix xx.
  - One important skill: **we closely follow xx**

How to write a experiment

❑基本宗旨(Principles)：防御性（defensive）

- Experimental results
  - Comparing with SOTAs
  - **Ablation study** to demonstrate the impact of the component
  - Hyperparemeter study
  - Case study
  - Step-by-step with research questions

How to write a conclusion

❑基本宗旨(Principles)：有洞察力的（insightful）

- Conclusion

  - Follow your introduction
  - Show the insights of the paper
  - Show the benefits of the proposed method
  - Give some insightful future directions
  - Discuss the limitations (optimal)

Checklist for the paper written?

- Something reviewers care about:

  - Clear motivations?
    --- the importance of the studied problem
    --- the limitations of existing methods

  - Novelty of the proposed method?
    --- how is the proposed method differs from existing methods
    --- the advantages of the proposed method

Checklist for the paper written?

- Something reviewers care about:

  - Comprehensive experiments
    --- proper datasets
    --- sufficient baselines, SOTA methods
    --- details of the implemental details, e.g., hyperparameters
    --- sufficient ablation studies
    --- sufficient hyperparameter studies

An example of the paper

- Introduction:

为优化目标研究铺垫，突出独特性

Nowdays, recommender systems (RS) have permeated various personalized services [1, 2]. What sets recommendation apart from other machine learning tasks is its distinctive emphasis on *ranking* [3]. Specifically, RS aims to retrieve positive items in higher ranking positions (*i.e.*, giving larger prediction scores) over others and adopts specific ranking metrics (*e.g.*, DCG [4] and MRR [5]) to evaluate its performance.

The emphasis on ranking inspires a surge of research on loss functions in RS. Initial studies treated recommendation primarily as a classification problem, utilizing pointwise loss functions (*e.g.*, BCE [6], MSE [7]) to optimize models. Recognizing the inherent ranking nature of RS, pairwise loss functions (*e.g.*, BPR [8]) were introduced to learn a partial ordering among items. More recently, *Softmax Loss* (SL) [9] has integrated contrastive learning paradigms [10, 11], augmenting positive items as compared with negative ones, achieve state-of-the-art (SOTA) performance.

一句话讲清了重要性

An example of the paper

- Introduction:

While SL has proven effective, it still suffers from two limitations: 1) SL can be approximated to ranking metrics (*e.g.*, DCG, MRR [9, 12]), but their relationships are not sufficiently tight. Specifically, SL uses the exponential function $\exp(\cdot)$ as the *surrogate activation* to approximate the Heaviside function in DCG, resulting in a notable gap, especially when the function takes larger values. 2) SL is sensitive to noise (*e.g.*, false negatives [13]). Gradient analysis reveals that SL assigns higher weights to negative instances with larger prediction scores, while the weights are rather skewed and governed by an exponential function. This characteristic renders the model highly sensitive to noise. Specifically, false negative instances are common in RS, as a user's lack of interaction with an item might stem from unawareness rather than disinterest [14]. These instances would receive

An example of the paper

❑ Introduction:

To address these challenges, we propose a new family of loss functions, termed *Pairwise Softmax Loss* (PSL). PSL first reformulates SL in a pairwise manner, where the loss is applied to the score gap between positive-negative pairs. Such *pairwise* perspective is more fundamental to recommendation as the ranking metrics are also pairwise dependent. Recognizing that the primary weakness of SL lies in its use of the exponential function, PSL replaces this with other surrogate activations. While this extension is straightforward, it brings significant theoretical merits:

1) Potentially better surrogate for ranking metrics: We establish theoretical connections between PSL and conventional ranking metrics, *e.g.,* DCG. By choosing appropriate surrogate activations, such as ReLU or Tanh, we demonstrate that PSL achieves a tighter DCG surrogate loss than Softmax.

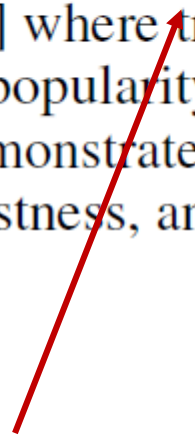
2) Theoretical connections with BPR loss: Our analyses reveal that optimizing PSL is equivalent to performing distributionally robust optimization (DRO, [15]) over the BPR loss. DRO is a theoretically sound framework where a model is optimized not only on a fixed empirical distribution but also across a set of distributions with adversarial perturbations. This DRO characteristic endows PSL with stronger generalization and robustness against out-of-distribution (OOD), especially given that such distribution shifts are common in RS, *e.g.,* shifts in user preference and item popularity [14, 16].

3) Control over the shape of the weight distribution: PSL provides flexibility in choosing surrogate activation that control the weight distribution. Given appropriate surrogate activations, such as ReLU or Tanh, it can mitigate the excessive impact of false negatives, thus enhancing robustness to noise.

An example of the paper

- Introduction:

Our analyses underscore the theoretical effectiveness and robustness of PSL. To empirically validate these advantages, we implement PSL with typical surrogate activations (ReLU, Tanh, Atan) and conduct extensive experiments on four real-world datasets across three experimental settings: 1) IID setting [17] where training and test distributions are identically distributed [18]; 2) OOD setting [19] with item popularity distribution shifts; 3) Noise setting [13] with a certain ratio of false negatives. Results demonstrate the superiority of PSL over existing losses in terms of recommendation accuracy, OOD robustness, and noise resistance.



实验亮点

An example of the paper

- Method:

2 Preliminaries

Task formulation. We will conduct our discussion in the scope of collaborative filtering (CF) [20], a widely-used recommendation scenario. Given the user set $\mathcal{U}$ and item set $\mathcal{I}$, Let $\mathcal{D} \subset \mathcal{U} \times \mathcal{I}$ be a collection of observed interactions, where $(u, i) \in \mathcal{D}$ means that user u has interacted with item i (*e.g.*, clicks, purchases, reviews, *etc.*). For each user u, we denote $\mathcal{P}_u = \{i \in \mathcal{I} : (u, i) \in \mathcal{D}\}$ as the set of positive items of u, while $\mathcal{I} \setminus \mathcal{P}_u$ represents negative item set. The goal of recommendation is

Recommendation metrics. Discounted Cumulative Gain (DCG) [4] is a prominent metric for assessing the quality of recommendations. Formally, DCG for each user u is calculated as follows:

Recommendation loss. Recent work on loss functions can be mainly classified into three types:

An example of the paper

- Method:

3 Analyses on Softmax Loss from Pairwise Perspective

In this section, we aim to first represent the Softmax Loss (SL) in a pairwise form, followed by an analysis of its relationship with the DCG metric, where two limitations of SL are exposed.

Drawback 1: SL is not tight enough as a DCG surrogate loss. There remains a significant gap between the Heaviside function $\delta(\cdot)$ and the exponential function $\exp(\cdot)$, especially when d_{uij} reaches a relatively large value, where $\exp(\cdot)$ becomes substantially larger than $\delta(\cdot)$. This gap is further exacerbated by the temperature parameter τ. Practically, we find the optimal τ is usually chosen to be less than 0.2, given the explosive nature of $\exp(\cdot)$, the gap becomes extremely large.

Drawback 2: SL is highly sensitive to noise (*e.g.*, false negative instances). False negative instances are common in a typical RS. This is often due to exposure bias [14], where a user's lack of interaction with an item might stem from unawareness rather than disinterest. However, SL is highly sensitive to these instances. On one hand, these instances, which may exhibit patterns similar to true positive ones, are difficult to differentiate by the model and often obtain larger predicted values, *i.e.*, bringing potentially larger d_{uij}. These instances can enlarge the gap between SL and DCG, causing the optimization to deviate from the DCG metric.

An example of the paper

- Method:

4 Methodology

4.1 Pair-wise Softmax Loss

Recognizing the limitations of SL, particularly its reliance on the unsatisfactory exponential function, we propose to extend SL with a more general family of losses, termed *Pair-wise Softmax Loss* (PSL). In PSL, the exponential function $\exp(\cdot)$ is replaced by other activation functions $\sigma(\cdot)$:

$$\mathcal{L}_{\mathrm{PSL}}(u) = \sum_{i \in \mathcal{P}_u} \log \left(\sum_{j \in \mathcal{I}} \sigma(d_{uij})^{1/\tau} \right) \tag{4.1}$$

An example of the paper

- Method:

PSL as a Better Surrogate for Ranking Metrics. To highlight the advantages of replacing $\exp(\cdot)$ with alternative activation functions, we present the following lemma:

Lemma 4.1. *When the condition*

$$\delta(d_{uij}) \leq \sigma(d_{uij}) \leq \exp(d_{uij}) \tag{4.2}$$

is satisfied for any $d_{uij} \in [-1, 1]$, *then PSL is a tighter DCG surrogate loss than SL.*

The proof is presented in Appendix A.1. This lemma reveals that PSL could be a tighter surrogate loss of DCG compared to SL. Besides, it also provides guidance on the selection of a proper activation — we may choose the function that lies between $\exp(\cdot)$ and $\delta(\cdot)$. As demonstrated in Figure 1a, our chosen activations σ_{Relu}, σ_{Atan}, and σ_{Tanh} adhere to this principle.

分析很关键

An example of the paper

- Method:

性质分析

4.2 Discussions

Comparisons of two extension forms. We highlight three advantages of the form that positions the temperature outside (i.e., $\sigma(d_{uij})^{1/\tau}$) over the inside (i.e., $\sigma(d_{uij}/\tau)$): 1) As previously discussed, the outside form can be regarded as a DRO-empowered BPR loss, while the inside form cannot; 2) PSL with the outside form can degenerate into BPR as $\tau \to \infty$, ensuring that PSL performs at least as well as BPR, while the inside form does not have this property. 3) The outside form facilitates the selection of activation functions. For example, to ensure Lemma 4.1 holds, we only need to consider the range of d_{uij} as input, *i.e.*, $[-1, 1]$. However, for the inside form, this range would be expanded by $1/\tau$, complicating the selection of activation functions.

Connections with other losses. 1) **Connection with AdvInfoNCE [31]:** According to Theorem 3.1 of Zhang et al. [31], AdvInfoNCE can indeed be considered as a special case of PSL with $\sigma(\cdot) = \exp(\exp(\cdot))$. We argue that this activation is not a good choice as it would enlarge the gap with DCG. In fact, we have $\log \text{DCG} \leq \mathcal{L}_{\text{PSL}} \leq \mathcal{L}_{\text{SL}} \leq \mathcal{L}_{\text{AdvInfoNCE}}$ (*cf.* Appendix A.2 for proof).

区别联系

An example of the paper

- Experiments:

5 Experiments

5.1 Experimental setup

5.1.1 Datasets and testing paradigms.

5.1.2 Metrics. We closely refer to Wu et al. [13] and Zhang et al. [31], adopting NDCG@K [4] and Recall@K for performance evaluation. Where NDCG is the normalized DCG where DCG is

5.1.3 Compared methods. Five representative loss functions are compared in our experiments, including: 1) the representative pairwise loss **BPR** (UAI'09 [8]); 2) **Softmax Loss** (TOIS'24 [9]) and its two enhancements **AdvInfoNCE** (NIPS'23 [31]) and **BSL** (ICDE'24 [13]); 3) another state-of-the-art loss **LLPAUC** (WWW'24 [32]) that optimizes the Lower-Left Partial AUC. Readers may refer to Appendix B.3 for more details about these baselines.

An example of the paper

- Experiments:

5.1.4 Hyperparameter settings. A grid search is utilized to find the optimal hyperparameters. For all compared methods, we closely refer to configurations provided in their respective publications to ensure their optimal performance. As we also highly fine-tune the basic Softmax Loss, the improvements of existing methods over it are not as significant as presented in their papers. The hyperparameter settings are provided in the appendix Appendix B.5, where the detailed optimal hyperparameters for each method on each dataset are reported.

An example of the paper

- Experiments:

5.2 Performance Comparisons

5.2.1 Results under IID Scenerio. Table 1 presents the performance of our PSL compared with

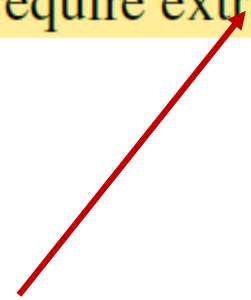
5.2.2 Results under OOD Scenerio.
Given the nearly consistent behavior

5.2.3 Results under Noisy Scenerio.
ratio of imputed false negative instanc

An example of the paper

- Related work:

Loss-related Recommendation Research. Existing losses can mainly categorized into point-wise loss [6, 21], pair-wise loss [8], and Softmax loss [9], as discussed in section 2. Given the highly effectiveness of SL, Recently some researchers propose to enhance SL from different perspectives. For example, BSL [13] considers to enhance the positive distribution robustness by leveraging DRO; Advinfonce [31] leverages adversarial learning to enhance the robustness of SL; [16] proposes to incorporating bias-aware Margins in SL to tackle popularity bias. Besides these three types of losses, some other losses are also explored in recent year. For example, [52] proposes auto-loss to leverage auto machine learning technique to search the optimal loss; [32] proposes LLPAUC to approximate Recall@K metrics. The major concerns of these losses lie on the lack of theoretical connections with DCG metrics, making them may not always outperforms the basic SL. Besides, the auto-loss and LLPAUC require extra iterative learning, which would incur extra computational time and increase instability.



要讨论他们的缺点！要多引，不要漏。 Reviewer可能是作者

An example of the paper

- Conclusion:

7 Conclusion and Limitation

In this work, we deliver a new family of loss functions, termed as Pairwise Softmax Loss (PSL). PSL enjoys three theoretical advantages: 1) it can be a better surrogate for ranking metrics with proper activation functions; 2) it can be understood as a specific BPR loss empowered by distributionally robust optimization; 3) it can flexible control the shape of the instance weight distribution. These three properties demonstrate PSL has better effectiveness and robustness over the basic Softmax Loss. Our extensive experiments on three testing scenarios validate the superiority of PSL over existing methods. A limitation of PSL and SL is inefficiency. Both PSL and SL require a relatively large number of negative instances in a epoch. How to address to this issue would be an interesting furture direction.

How to write a review paper

❑基本宗旨(Principles)：全面性 (comprehensive) and 新颖性(novelty)

- First point: study on an important area
  --- why this topic is so important?

- Second point: why this topic needs a review?
  --- the flourishing publications
  --- comparing with existing survey/review paper

- Third point: the taxonomy of this area
  --- the unique nature of this area
  --- how to classify existing methods, their connections?

- Fourth point: the future directions

How to write a review paper

- An example: Bias and Debias in Recommender System: A Survey and Future Directions (Tois'23)

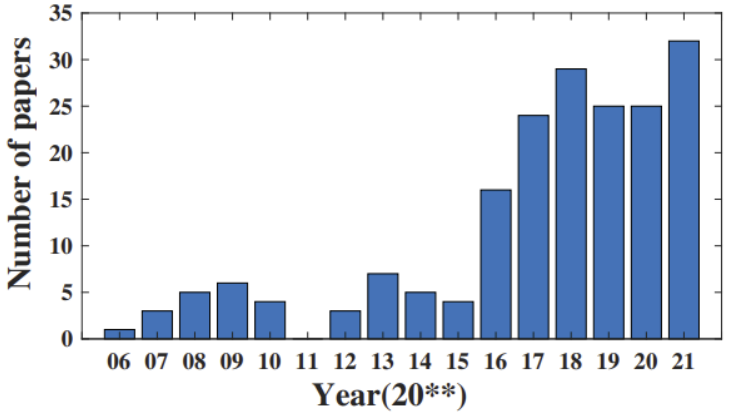
Ubiquity of Biases in RS. Although RS has generated large impacts in a wide range of applications, it faces many bias problems which are challenging to handle and may deteriorate the recommendation effectiveness. Bias is common in RS for the following factors. (1) User behavior data, which lays the foundation for recommendation model training, is observational rather than experimental. The main reason is that a user generates behaviors on the basis of the exposed items, making the observational data confounded by the exposure mechanism of the system and the self-selection of the user. (2) Items are not evenly presented in the data, e.g., some items are more popular than others and thus receive more user behaviors. As a result, these popular items would have a larger impact on the model training, making the recommendations biased towards them. The same situation applies to the user side. (3) One nature of RS is the feedback loop — the exposure mechanism of the RS determines user behaviors, which are circled back as the training data for the RS. Such feedback loop not only creates biases but also intensifies biases over time, resulting in "the rich get richer" Matthew effect.

https://arxiv.org/pdf/2010.03240

How to write a review paper

- An example: Bias and Debias in Recommender System: A Survey and Future Directions (Tois'23)

Increasing Importance of Biases in RS Research. Recent years have seen a surge of research effort on recommendation biases. Figure 1 shows the number of related papers in top venues increases significantly since the year of 2015. The prestigious international conference on information retrieval, SIGIR, has organized specific sessions in 2020 and 2021 to discuss topics on bias elimination[1]. SIGIR even presents the Best Paper award to the paper on this topic in 2018 [23], 2020 [121] and 2021 [124, 208], respectively. The conferences Recsys and WWW also organized tutorial on this topic in 2021 [33, 182]. Biases not only draw increasing attention from the information retrieval academia, but also from the industry. For example, one competing task of KDD Cup 2020 organized by Alibaba is to handle the long-tail bias in E-commerce recommendation[2].



https://arxiv.org/pdf/2010.03240

How to write a review paper

- An example: Bias and Debias in Recommender System: A Survey and Future Directions (Tois'23)

Necessity of this Survey. Although many papers are published on this topic recently, to the best of our knowledge, none of them has provided a global picture of the RS biases and corresponding debiasing techniques. Particularly, we find that current studies on this topic are rather fragmented — despite the wide usage of the terminology "bias" in the literature, its definition is usually vague and even inconsistent across papers. For example, some work use "selection bias" to denote the bias of observed rating values [140], while others use "observational bias" to refer to the same meaning instead [66]. More confusingly, the same terminology "selection bias" has been conceptualized differently in different publications [125, 140, 169]. Moreover, a considerable number of researchers do not explicitly mention "bias" or "debias" in the paper (e.g. [29, 102, 174]), but they indeed address one type of biases in RS; these significant related work is difficult to be retrieved by the researchers interested in the bias topic. Given the increasing attention of biases in RS, the rapid development of debiasing techniques, and the flourishing but fragmented publications, we believe it is the right time to present a survey of this area, so as to benefit the successive researchers and practitioners to understand current progress and further work on this topic.

https://arxiv.org/pdf/2010.03240

How to write a review paper

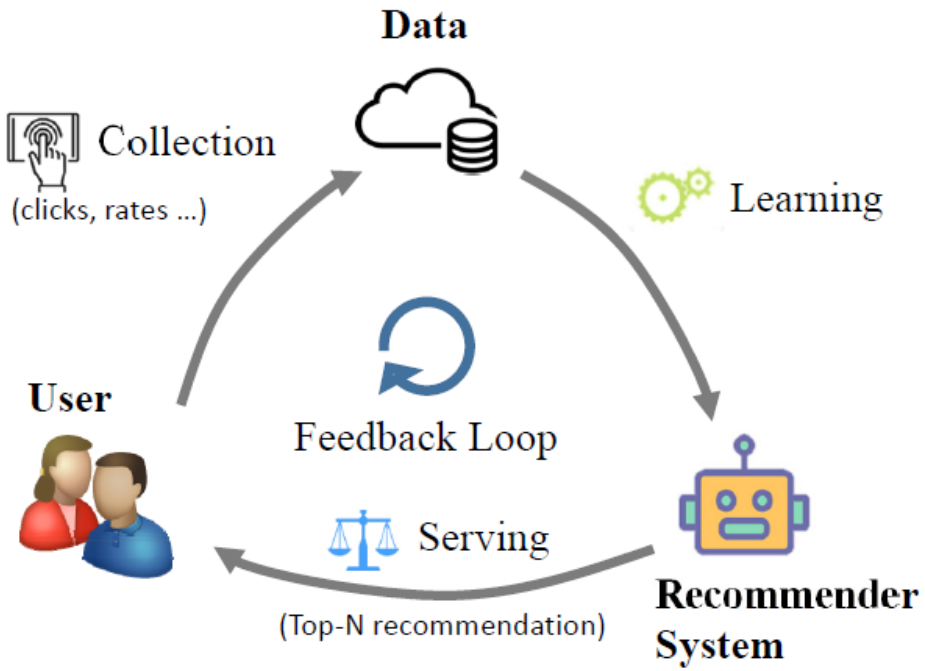
- An example: Bias and Debias in Recommender System: A Survey and Future Directions (Tois'23)

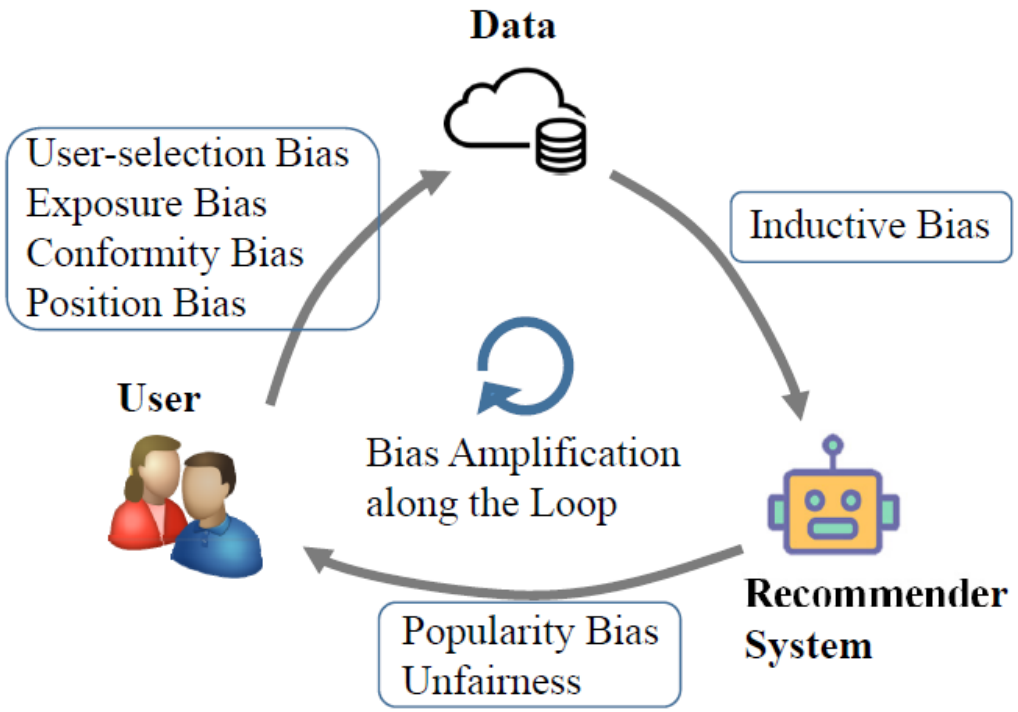
Difference with Existing Surveys. A number of surveys in recommendation have been published recently, focusing on different perspectives of RS. For example, [207] reviews explainable recommendation, [156] reviews knowledge-based recommendation, [203] and [209] summarize the recommendation methods based on deep learning and reinforcement learning, respectively. However, to our knowledge, the perspective of bias has not been reviewed in existing RS surveys. There are some surveys on the bias issues, but they are not on the recommendation task. For example, [16] recently reviews the bias issues in natural language processing, [159] reviews the sample selection bias on model estimation, [206] summarizes fairness in learning-based sequential decision algorithms. There are some surveys on the bias and fairness of general machine learning and artificial

https://arxiv.org/pdf/2010.03240

How to write a review paper

- An example: Bias and Debias in Recommender System: A Survey and Future Directions (Tois'23)





https://arxiv.org/pdf/2010.03240

How to write a review paper

- Literature review (70%)
  - Written assignment (>3000 words)
  - Covering at least 15 core papers
  - Giving a taxonomy of this area
  - Directly using GPT for generation is prohibited

- **Deadline: 10th Nov (11月10号）**

Thank You!