A Method to Anonymize Business Metrics to Publishing Implicit Feedback Datasets

Yoshifumi SEKI (Gunosy.inc, Japan)
Takanori MAEHARA (RIKEN, Japan)

Background

Gunosy

Dataset publication is important for recsys studies.

Datasets have contributed to develop recommendation system studies.

- MovieLens, Netflix Prize
- In recent years, some data science competitions, such as Kaggle, KDD Cup, and Recsys Challenges, promote dataset publications.

Implicit feedback datasets from commercial services are not enough.

- Recommendation systems have adopted in many and various service, so many and various datasets are needed.

Background

Gunosy

Dataset publication is important for recsys studies.

Datasets have contributed to develop recommendation system studies.

- **MovieLens, Netflix Prize**
- **In recent years, some data science competitions, such as Kaggle, KDD Cup, and Recsys Challenges, promote dataset publications.**

Implicit feedback datasets from commercial services are not enough.

- Recommendation systems have adopted in many and various service, so many and various datasets are needed.

Background

Gunosy

Dataset publication is important for recsys studies.

Datasets have contributed to develop recommendation system studies.

- MovieLens, Netflix Prize
- In recent years, some data science competitions, such as Kaggle, KDD Cup, and Recsys Challenges, promote dataset publications.

Implicit feedback datasets from commercial services are not enough.

- **Recommendation systems have adopted in many and various service, so many and various datasets are needed.**

Motivation

Gunosy

We would like to make it easier for commercial services to publish datasets.

There are some business risks to publish dataset.

- Leaking confidential business metrics.
- Some reputation risks.

Before publishing a dataset, researchers must get approval by a business manager.

- many business managers are not specialists in machine learning or recommender system.
- The researchers should be responsible for explaining the risks and benefits.

We focus on an implicit feedback datasets.

- Implicit feedback datasets include confidential business information and users' personal information.
- Explicit feedback datasets are often constructed by crawling public web resources, such as user reviews and ratings available online.

Motivation

Gunosy

We would like to make it easier for commercial services to publish datasets.

There are some business risks to publish dataset.

- Leaking confidential business metrics.
- Some reputation risks.

Before publishing a dataset, researchers must get approval by a business manager.

- many business managers are not specialists in machine learning or recommender system.
- The researchers should be responsible for explaining the risks and benefits.

We focus on an implicit feedback datasets.

- Implicit feedback datasets include confidential business information and users' personal information.
- Explicit feedback datasets are often constructed by crawling public web resources, such as user reviews and ratings available online.

Motivation

Gunosy

We would like to make it easier for commercial services to publish datasets.

There are some business risks to publish dataset.

- Leaking confidential business metrics.
- Some reputation risks.

Before publishing a dataset, researchers must get approval by a business manager.

- many business managers are not specialists in machine learning or recommender system.
- The researchers should be responsible for explaining the risks and benefits.

We focus on an implicit feedback datasets.

- Implicit feedback datasets include confidential business information and users' personal information.
- Explicit feedback datasets are often constructed by crawling public web resources, such as user reviews and ratings available online.

Motivation

Gunosy

We would like to make it easier for commercial services to publish datasets.

There are some business risks to publish dataset.

- Leaking confidential business metrics.
- Some reputation risks.

Before publishing a dataset, researchers must get approval by a business manager.

- many business managers are not specialists in machine learning or recommender system.
- The researchers should be responsible for explaining the risks and benefits.

We focus on an implicit feedback datasets.

- Implicit feedback datasets include confidential business information and users' personal information.
- Explicit feedback datasets are often constructed by crawling public web resources, such as user reviews and ratings available online.

Contributions

Gunosy

- We summarize the challenges of building and publishing datasets from commercial service
- We formulate the problem of building and publishing a dataset as a optimization problem that seeks the sampling weight of users.
- We applied our method to build datasets from the raw data of our real-world mobile news delivery service Gunosy, which is a popular news delivery service in Japan
  - The raw data has more than 1,000,000 users with 100,000,000 interactions.
- The implementation of our proposed method and a dataset built by our proposed method are public

https://github.com/gunosy/publishing-dataset-recsys20

Contributions

Gunosy

- **We summarize the challenges of building and publishing datasets from commercial service**
- We formulate the problem of building and publishing a dataset as a optimization problem that seeks the sampling weight of users.
- We applied our method to build datasets from the raw data of our real-world mobile news delivery service Gunosy, which is a popular news delivery service in Japan
  - The raw data has more than 1,000,000 users with 100,000,000 interactions.
- The implementation of our proposed method and a dataset built by our proposed method are public

https://github.com/gunosy/publishing-dataset-recsys20

Contributions

Gunosy

- We summarize the challenges of building and publishing datasets from commercial service

- **We formulate the problem of building and publishing a dataset as a optimization problem that seeks the sampling weight of users.**

- We applied our method to build datasets from the raw data of our real-world mobile news delivery service Gunosy, which is a popular news delivery service in Japan
  - The raw data has more than 1,000,000 users with 100,000,000 interactions.

- The implementation of our proposed method and a dataset built by our proposed method are public

https://github.com/gunosy/publishing-dataset-recsys20

Contributions

Gunosy

- We summarize the challenges of building and publishing datasets from commercial service
- We formulate the problem of building and publishing a dataset as a optimization problem that seeks the sampling weight of users.
- **We applied our method to build datasets from the raw data of our real-world mobile news delivery service Gunosy, which is a popular news delivery service in Japan**
    - **The raw data has more than 1,000,000 users with 100,000,000 interactions.**
- The implementation of our proposed method and a dataset built by our proposed method are public

https://github.com/gunosy/publishing-dataset-recsys20

Contributions

Gunosy

- We summarize the challenges of building and publishing datasets from commercial service
- We formulate the problem of building and publishing a dataset as a optimization problem that seeks the sampling weight of users.
- We applied our method to build datasets from the raw data of our real-world mobile news delivery service Gunosy, which is a popular news delivery service in Japan
    - The raw data has more than 1,000,000 users with 100,000,000 interactions.
- **The implementation of our proposed method and a dataset built by our proposed method are public**

https://github.com/gunosy/publishing-dataset-recsys20

Tasks

Gunosy

We only focus on the following three data to simplify the situation.

- **User behavior logs:** When user *u* clicks article *a* at time *t*, the triplet *(u, a, t)* is recorded as a log
- **User attributes**: each user has attributes, such as age and gender.
- **Article category**: Each news articles has a category, such as sports, entertainment, and politics.

Our task is to publish a subset of the user behavior logs.

We build dataset by “sampling user” approach.

1. Samples users from user behavior logs.
2. Collects all the user behavior logs associated with the sampled users

Tasks

Gunosy

We only focus on the following three data to simplify the situation.

- **User behavior logs:** When user *u* clicks article *a* at time *t*, the triplet *(u, a, t)* is recorded as a log
- **User attributes**: each user has attributes, such as age and gender.
- **Article category**: Each news articles has a category, such as sports, entertainment, and politics.

Our task is to publish a subset of the user behavior logs.

We build dataset by “sampling user” approach.

1. Samples users from user behavior logs.
2. Collects all the user behavior logs associated with the sampled users

Tasks

Gunosy

We only focus on the following three data to simplify the situation.

- **User behavior logs:** When user *u* clicks article *a* at time *t*, the triplet *(u, a, t)* is recorded as a log
- **User attributes**: each user has attributes, such as age and gender.
- **Article category**: Each news articles has a category, such as sports, entertainment, and politics.

Our task is to publish a subset of the user behavior logs.

We build dataset by “sampling user” approach.

1. Samples users from user behavior logs.
2. Collects all the user behavior logs associated with the sampled users

Tasks

Gunosy

We only focus on the following three data to simplify the situation.

- **User behavior logs:** When user *u* clicks article *a* at time *t*, the triplet *(u, a, t)* is recorded as a log
- **User attributes**: each user has attributes, such as age and gender.
- **Article category**: Each news articles has a category, such as sports, entertainment, and politics.

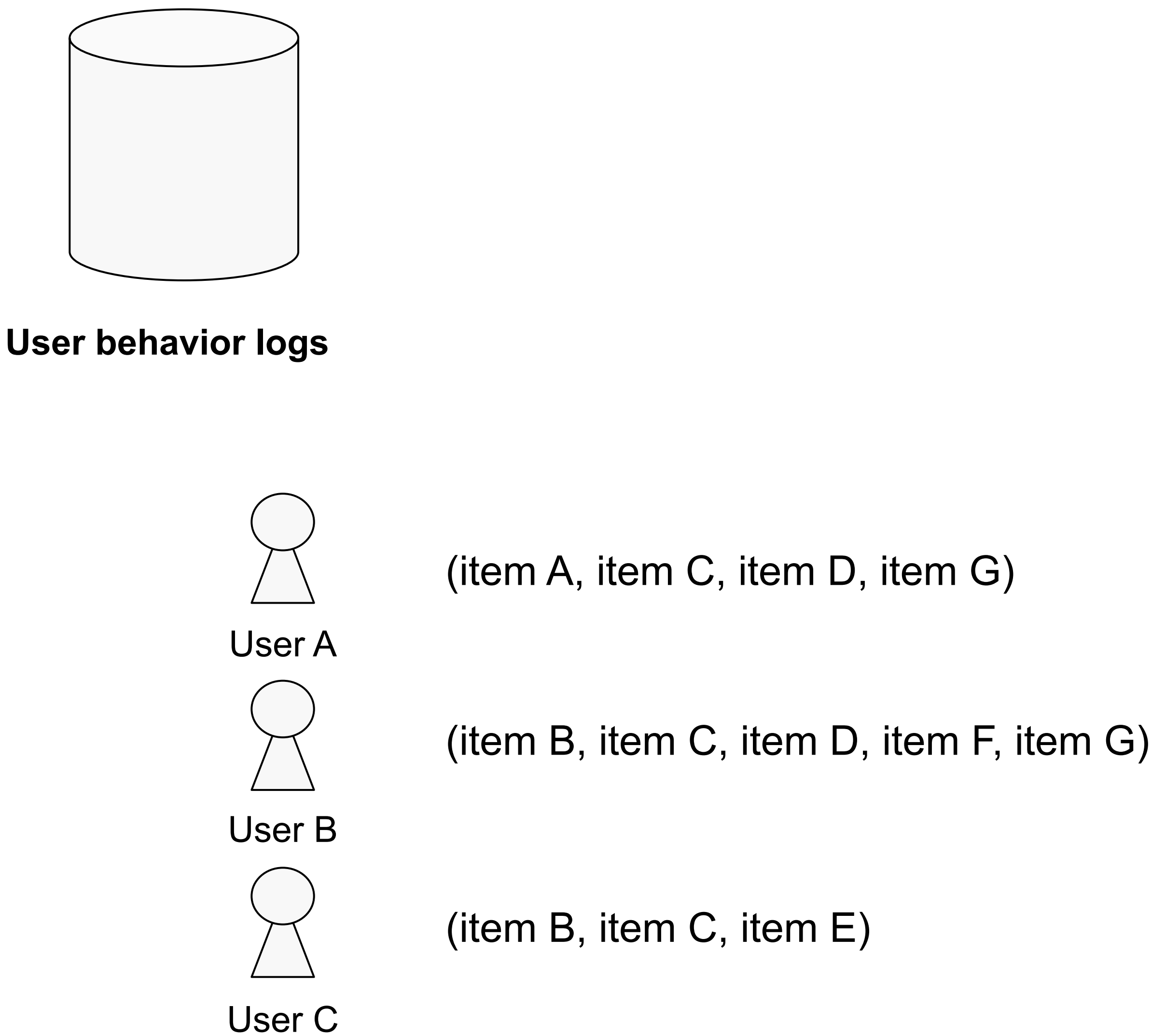
Our task is to publish a subset of the user behavior logs.

We build dataset by "sampling user" approach.

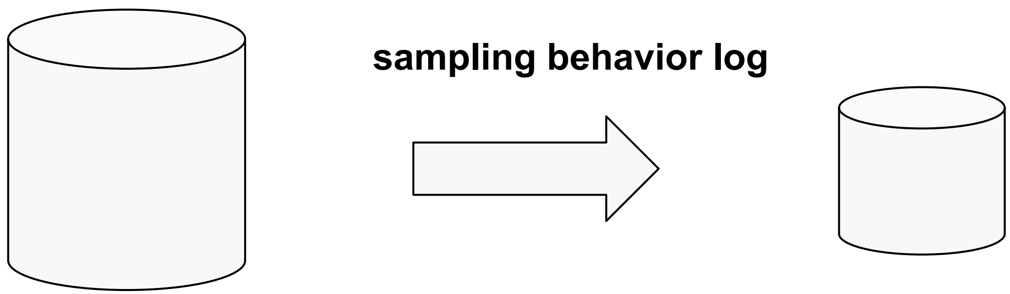
1. **Samples users from user behavior logs.**
2. **Collects all the user behavior logs associated with the sampled users**

Sampling Approach

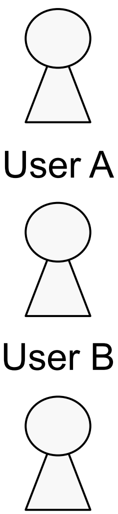
Gunosy



Sampling Approach



The consumption histories of the users are missing.

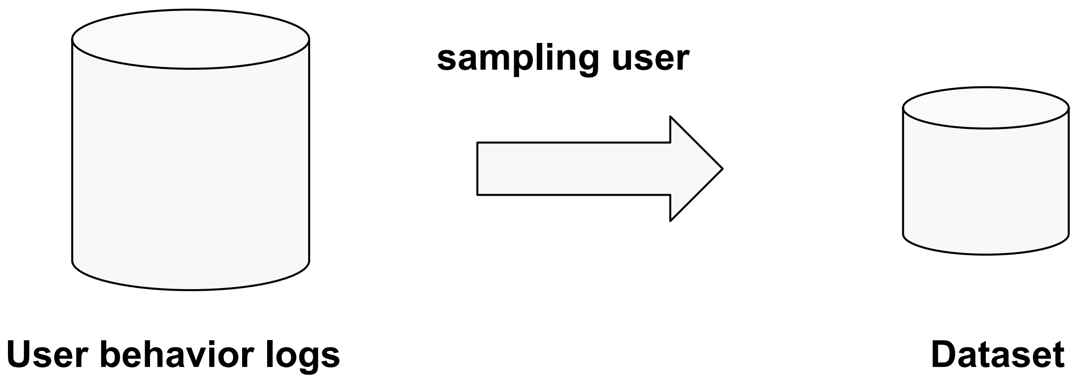


(item A, ~~item C~~, item D, item G)

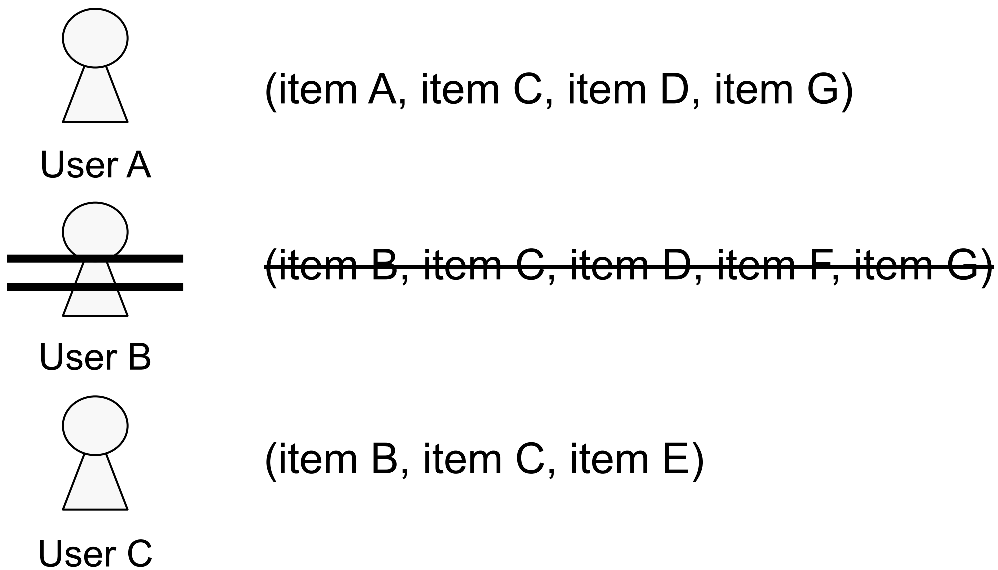
(item B, item C, ~~item D~~, item F, ~~item G~~)

(~~item B~~, item C, ~~item E~~)

Sampling Approach



The consumption histories of the users are keeping.



Challenges

We pose the following three challenges.

1. Anonymize the Business Metrics
2. Maintain Faireness
3. Reduce Popularity Bias

Challenges

Gunosy

We pose the following three challenges.

1. **Anonymize the Business Metrics**
2. Maintain Faireness
3. Reduce Popularity Bias

- Do not want to disclose confidential business metrics.
  - operating income
  - the average number of clicks
  - the average active rate of users
- If the users are sampled uniformly, some business metrics could be easily estimated.
  - the average number of clicks
  - the average active rate of users
- **We must sample users with a non-uniform distribution.**

Challenges

Gunosy

We pose the following three challenges.

1. **Anonymize the Business Metrics**
2. Maintain Faireness
3. Reduce Popularity Bias

- **Do not want to disclose confidential business metrics.**
  - **operating income**
  - **the average number of clicks**
  - **the average active rate of users**
- If the users are sampled uniformly, some business metrics could be easily estimated.
  - the average number of clicks
  - the average active rate of users
- **We must sample users with a non-uniform distribution.**

Challenges

Gunosy

We pose the following three challenges.

1. **Anonymize the Business Metrics**
2. Maintain Faireness
3. Reduce Popularity Bias

- Do not want to disclose confidential business metrics.
  - operating income
  - the average number of clicks
  - the average active rate of users
- **If the users are sampled uniformly, some business metrics could be easily estimated.**
  - **the average number of clicks**
  - **the average active rate of users**
- **We must sample users with a non-uniform distribution.**

Challenges

Gunosy

We pose the following three challenges.

1. **Anonymize the Business Metrics**
2. Maintain Faireness
3. Reduce Popularity Bias

- Do not want to disclose confidential business metrics.
  - operating income
  - the average number of clicks
  - the average active rate of users
- If the users are sampled uniformly, some business metrics could be easily estimated.
  - the average number of clicks
  - the average active rate of users
- **We must sample users with a non-uniform distribution.**

Challenges

Gunosy

We pose the following three challenges.

1. Anonymize the Business Metrics
2. **Maintain Faireness**
3. Reduce Popularity Bias

- Publishing a fair dataset is very important.
  - Some existing methods that maintain fairness use user attributes; hence the user attributes cause de-anonymization.
  - Publishing unfair dataset indirectly contributes to creating unfair machine learning models.
- **This risk will damage the company's reputation.**

Challenges

Gunosy

We pose the following three challenges.

1. Anonymize the Business Metrics
2. **Maintain Fairness**
3. Reduce Popularity Bias

- Publishing a fair dataset is very important.
  - Some existing methods that maintain fairness use user attributes; hence the user attributes cause de-anonymization.
  - Publishing unfair dataset indirectly contributes to creating unfair machine learning models.
- **This risk will damage the company's reputation.**

Challenges

Gunosy

We pose the following three challenges.

1. Anonymize the Business Metrics
2. **Maintain Faireness**
3. Reduce Popularity Bias

- Publishing a fair dataset is very important.
    - Some existing methods that maintain fairness use user attributes; hence the user attributes cause de-anonymization.
    - Publishing unfair dataset indirectly contributes to creating unfair machine learning models.
- **This risk will damage the company's reputation.**

Challenges

Gunosy

We pose the following three challenges.

1. Anonymize the Business Metrics
2. Maintain Faireness
3. **Reduce Popularity Bias**

- Recommender systems are expected to match long-tailed items with users; thus, algorithms suffering the popularity bias cannot achieve their role.
- **We believe popularity bias is a problem in building dataset.**
  - If the dataset is built by the uniform sampling, the items of unpopular categories are less frequently sampled.
  - Because researchers cannot increase the number of interactions, the publisher must keep a certain amount of interactions with unpopular category items.

Challenges

Gunosy

We pose the following three challenges.

1. Anonymize the Business Metrics
2. Maintain Faireness
3. **Reduce Popularity Bias**

- Recommender systems are expected to match long-tailed items with users; thus, algorithms suffering the popularity bias cannot achieve their role.
- **We believe popularity bias is a problem in building dataset.**
    - If the dataset is built by the uniform sampling, the items of unpopular categories are less frequently sampled.
    - Because researchers cannot increase the number of interactions, the publisher must keep a certain amount of interactions with unpopular category items.

Mathematical Formulation

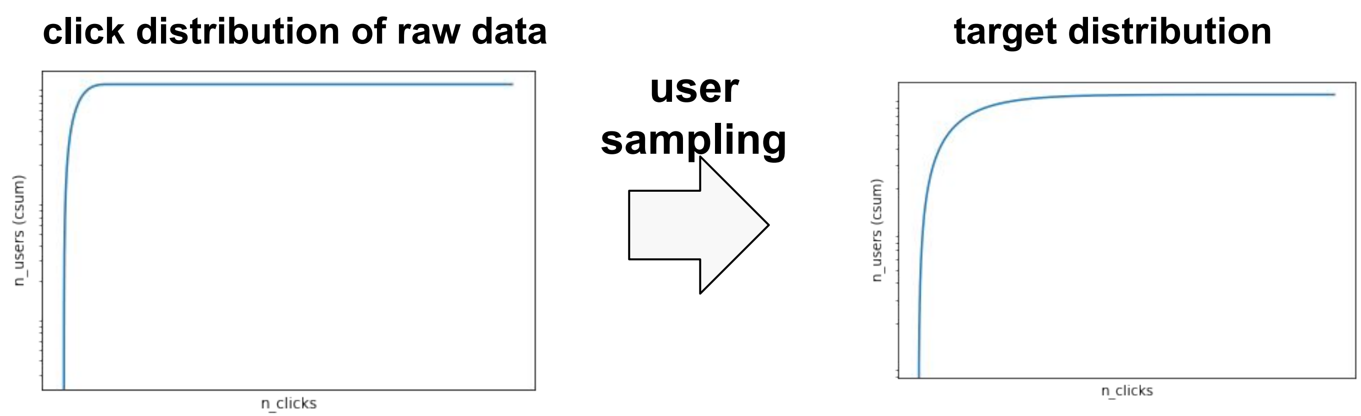
Gunosy

We formulate our task as a problem of finding the sampling weight of users: w(u).

We assume that our business metric are anonymized if the distribution of the number of clicks in the dataset is different from one in the raw data.

- formulating this challenge is impossible because it needs to enumerate all the metrics that we should anonymize.
- several important metrics are strongly correlated with the distribution of the number of clicks.

We sample users to make the distribution of datasets closer to a target distribution.



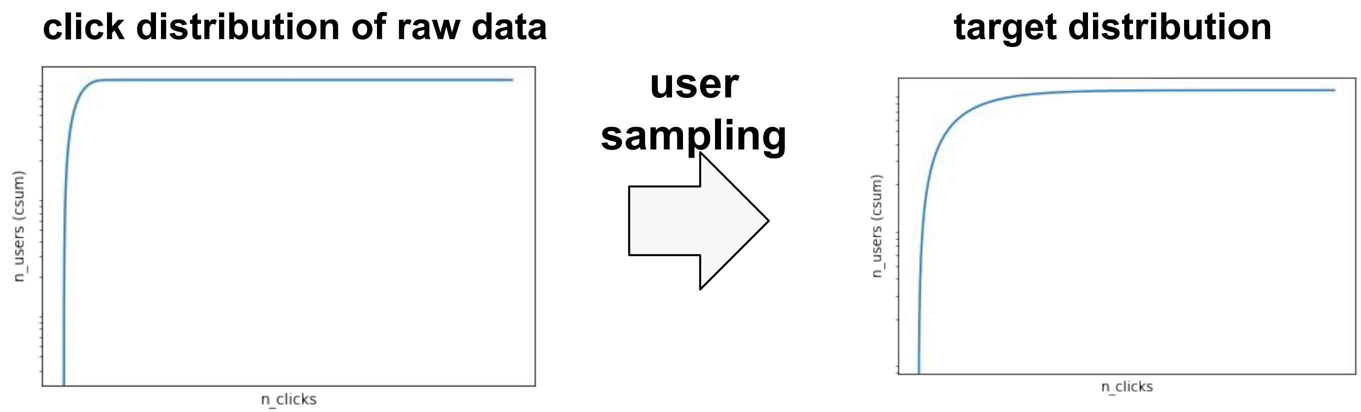
Mathematical Formulation

We formulate our task as a problem of finding the sampling weight of users: w(u).

We assume that our business metric are anonymized if the distribution of the number of clicks in the dataset is different from one in the raw data.

- **formulating this challenge is impossible because it needs to enumerate all the metrics that we should anonymize.**
- **several important metrics are strongly correlated with the distribution of the number of clicks.**

We sample users to make the distribution of datasets closer to a target distribution.



Mathematical Formulation

Gunosy

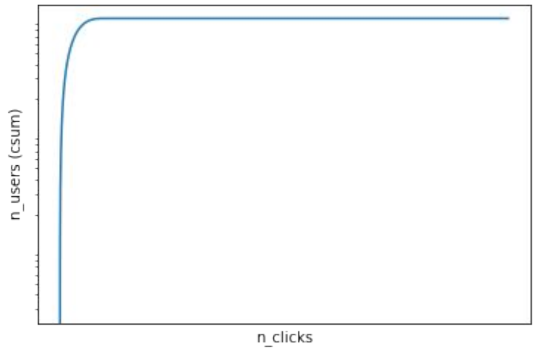
We formulate our task as a problem of finding the sampling weight of users: w(u).

We assume that our business metric are anonymized if the distribution of the number of clicks in the dataset is different from one in the raw data.

- formulating this challenge is impossible because it needs to enumerate all the metrics that we should anonymize.
- several important metrics are strongly correlated with the distribution of the number of clicks.

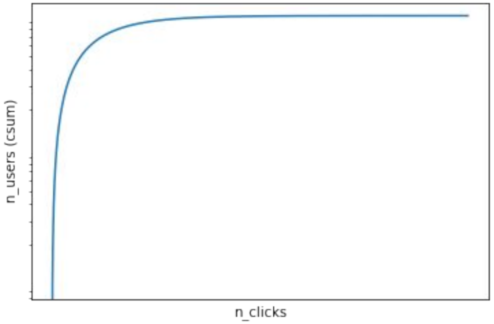
We sample users to make the distribution of datasets closer to a target distribution.

click distribution of raw data



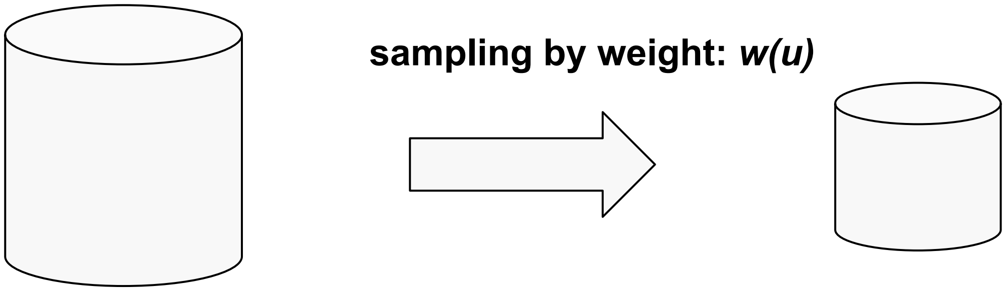
user sampling

target distribution



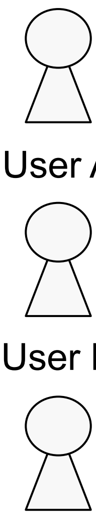
Finding Sampling Weight

sampling by weight: ***w(u)***



User behavior logs **Dataset**

Finding optimal w(u) to close target distribution



w(User A)	User A	(item A, item C, item D, item G)
w(User B)	User B	(item B, item C, item D, item F, item G)
w(User C)	User C	(item B, item C, item E)

Mathematical Formulation

Gunosy

We sample users to make the distribution of datasets closer to a target distribution.

$$L_{\text{click}}(w) = W(\overline{A_{\text{click}} w}, p_{\text{click}})$$

$\overline{A_{click} w}$: the expected click distribution on the dataset.

p_{click} : the target distribution

W : Wasserstein distance.

Mathematical Formulation

Gunosy

We sample users to make the distribution of datasets closer to a target distribution.

$$L_{\text{click}}(w) = W(\overline{A_{\text{click}} w}, p_{\text{click}})$$

$\overline{A_{click} w}$: the expected click distribution on the dataset.

p_{click} : the target distribution

W : Wasserstein distance on the real line.

We also sample users to make the distribution of user attributes and clicks in each article categories to a specific distribution.

$$L_{\text{attribute}}(w) = D(\overline{A_{\text{attribute}} w}, p_{\text{attribute}}),$$

$$L_{\text{category}}(w) = D(\overline{A_{\text{category}} w}, p_{\text{category}}),$$

D is the KL divergence.
Each expected distribution is simply calculated using sampling weight.

Mathematical Formulation

Gunosy

We sample users to make the distribution of datasets closer to a target distribution.

$$L_{\text{click}}(w) = W(\overline{A_{\text{click}} w}, p_{\text{click}})$$

$\overline{A_{click} w}$: the expected click distribution on the dataset.

p_{click} : the target distribution

W : Wasserstein distance on the real line.

We also sample users to make the distribution of user attributes and clicks in each article categories to a specific distribution.

$$L_{\text{attribute}}(w) = D(\overline{A_{\text{attribute}} w}, p_{\text{attribute}}),$$

$$L_{\text{category}}(w) = D(\overline{A_{\text{category}} w}, p_{\text{category}}),$$

D is the KL divergence.

Each expected distribution is simply calculated using sampling weight.

Mathematical Formulation

Gunosy

We find a sampling weight at which all the loss functions have small values.

$$\min_{w \in \mathbb{R}^{U}_{\geq 0}} L(w) := \alpha_{\text{click}} L_{\text{click}}(w) + \alpha_{\text{attribute}} L_{\text{attribute}}(w) + \alpha_{\text{category}} L_{\text{category}}(w),$$

We apply the gradient descent-type algorithms to minimize loss function.

Experiments

Gunosy

We built a dataset from the raw data in our news delivery services.

We built eight dataset from user behavior logs

- sample 60,000 users from raw data.
- two type target click distributions.
  - Zipf(1) and Zipf(2)
- controlled/un-controlled target distribtion of Attributes and Category

Table 1: Statistics of constructed datasets.

Name	#Clicks	Attributes	Category	#Users	#Interactions	#Items	Density
1UU	Zipf(1)			60,000	5,406,392	29,219	0.00308
1CU	Zipf(1)	o		60,000	5,372,623	29,054	0.00308
1UC	Zipf(1)		o	60,000	4,211,982	29,671	0.00236
1CC	Zipf(1)	o	o	60,000	4,255,355	29,740	0.00238
2UU	Zipf(2)			60,000	301,293	13,377	0.00037
2CU	Zipf(2)	o		60,000	318,525	13,406	0.00039
2UC	Zipf(2)		o	60,000	148,337	11,940	0.00020
2UU	Zipf(2)	o	o	60,000	146,711	11,965	0.00020

Experiments

Gunosy

We built a dataset from the raw data in our news delivery services.

We built eight dataset from user behavior logs

- sample 60,000 users from raw data.
- two type target click distributions.
  - Zipf(1) and Zipf(2)
- controlled/un-controlled target distribtion of Attributes and Category

Table 1: Statistics of constructed datasets.

Name	#Clicks	Attributes	Category	#Users	#Interactions	#Items	Density
1UU	Zipf(1)			60,000	5,406,392	29,219	0.00308
1CU	Zipf(1)	o		60,000	5,372,623	29,054	0.00308
1UC	Zipf(1)		o	60,000	4,211,982	29,671	0.00236
1CC	Zipf(1)	o	o	60,000	4,255,355	29,740	0.00238
2UU	Zipf(2)			60,000	301,293	13,377	0.00037
2CU	Zipf(2)	o		60,000	318,525	13,406	0.00039
2UC	Zipf(2)		o	60,000	148,337	11,940	0.00020
2UU	Zipf(2)	o	o	60,000	146,711	11,965	0.00020

Experiments

Gunosy

We built a dataset from the raw data in our news delivery services.

We built eight dataset from user behavior logs

- sample 60,000 users from raw data.
- two type target click distributions.
  - Zipf(1) and Zipf(2)
- controlled/un-controlled target distribtion of Attributes and Category

Table 1: Statistics of constructed datasets.

Name	#Clicks	Attributes	Category	#Users	#Interactions	#Items	Density
1UU	Zipf(1)			60,000	5,406,392	29,219	0.00308
1CU	Zipf(1)	o		60,000	5,372,623	29,054	0.00308
1UC	Zipf(1)		o	60,000	4,211,982	29,671	0.00236
1CC	Zipf(1)	o	o	60,000	4,255,355	29,740	0.00238
2UU	Zipf(2)			60,000	301,293	13,377	0.00037
2CU	Zipf(2)	o		60,000	318,525	13,406	0.00039
2UC	Zipf(2)		o	60,000	148,337	11,940	0.00020
2UU	Zipf(2)	o	o	60,000	146,711	11,965	0.00020

Experiments

Gunosy

We built a dataset from the raw data in our news delivery services.

We built eight dataset from user behavior logs

- sample 60,000 users from raw data.
- two type target click distributions.
  - Zipf(1) and Zipf(2)
- controlled/un-controlled target distribtion of Attributes and Category

Table 1: Statistics of constructed datasets.

Name	#Clicks	Attributes	Category	#Users	#Interactions	#Items	Density
1UU	Zipf(1)			60,000	5,406,392	29,219	0.00308
1CU	Zipf(1)	o		60,000	5,372,623	29,054	0.00308
1UC	Zipf(1)		o	60,000	4,211,982	29,671	0.00236
1CC	Zipf(1)	o	o	60,000	4,255,355	29,740	0.00238
2UU	Zipf(2)			60,000	301,293	13,377	0.00037
2CU	Zipf(2)	o		60,000	318,525	13,406	0.00039
2UC	Zipf(2)		o	60,000	148,337	11,940	0.00020
2UU	Zipf(2)	o	o	60,000	146,711	11,965	0.00020

Experiments

Gunosy

We built a dataset from the raw data in our news delivery services.

We built eight dataset from user behavior logs

- sample 60,000 users from raw data.
- two type target click distributions.
  - Zipf(1) and Zipf(2)
- controlled/un-controlled target distribtion of Attributes and Category

Table 1: Statistics of constructed datasets.

Name	#Clicks	Attributes	Category	#Users	#Interactions	#Items	Density
1UU	Zipf(1)			60,000	5,406,392	29,219	0.00308
1CU	Zipf(1)	o		60,000	5,372,623	29,054	0.00308
1UC	Zipf(1)		o	60,000	4,211,982	29,671	0.00236
1CC	Zipf(1)	o	o	60,000	4,255,355	29,740	0.00238
2UU	Zipf(2)			60,000	301,293	13,377	0.00037
2CU	Zipf(2)	o		60,000	318,525	13,406	0.00039
2UC	Zipf(2)		o	60,000	148,337	11,940	0.00020
2UU	Zipf(2)	o	o	60,000	146,711	11,965	0.00020

Zipf(2) datasets are more sparse than Zipf(1)

Experiments

Gunosy

We built a dataset from the raw data in our news delivery services.

We built eight dataset from user behavior logs

- sample 60,000 users from raw data.
- two type target click distributions.
  - Zipf(1) and Zipf(2)
- controlled/un-controlled target distribtion of Attributes and Category

Table 1: Statistics of constructed datasets.

Name	#Clicks	Attributes	Category	#Users	#Interactions	#Items	Density
1UU	Zipf(1)			60,000	5,406,392	29,219	0.00308
1CU	Zipf(1)	o		60,000	5,372,623	29,054	0.00308
1UC	Zipf(1)		o	60,000	4,211,982	29,671	0.00236
1CC	Zipf(1)	o	o	60,000	4,255,355	29,740	0.00238
2UU	Zipf(2)			60,000	301,293	13,377	0.00037
2CU	Zipf(2)	o		60,000	318,525	13,406	0.00039
2UC	Zipf(2)		o	60,000	148,337	11,940	0.00020
2UU	Zipf(2)	o	o	60,000	146,711	11,965	0.00020

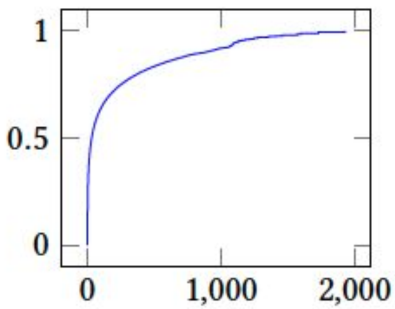
category controlled datasets are more sparse than uncontrolled datasets

Experiments

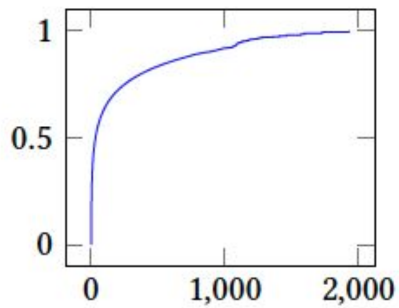
Gunosy

We successfully controlled the click distributions.

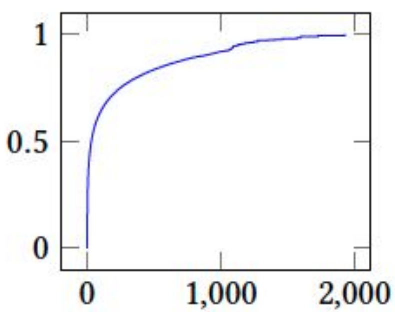
Zipf(1)'s distribution



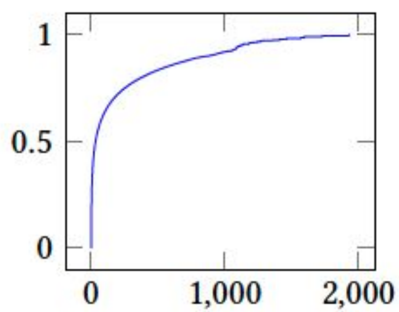
(a) 1CC



(b) 1CU

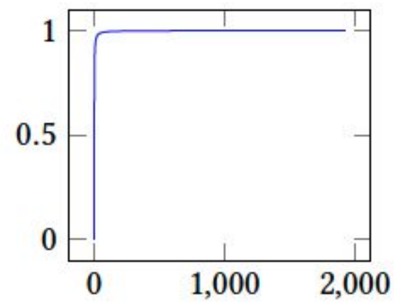


(c) 1UC

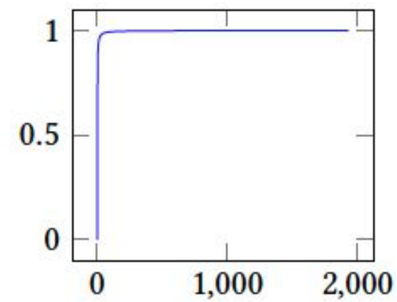


(d) 1UU

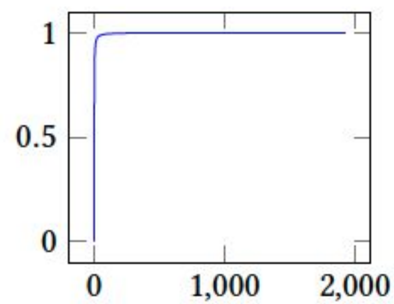
Zipf(2)'s distribution



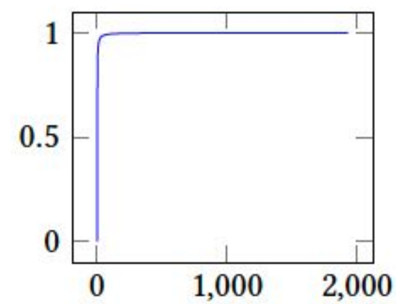
(e) 2CC



(f) 2CU



(g) 2UU



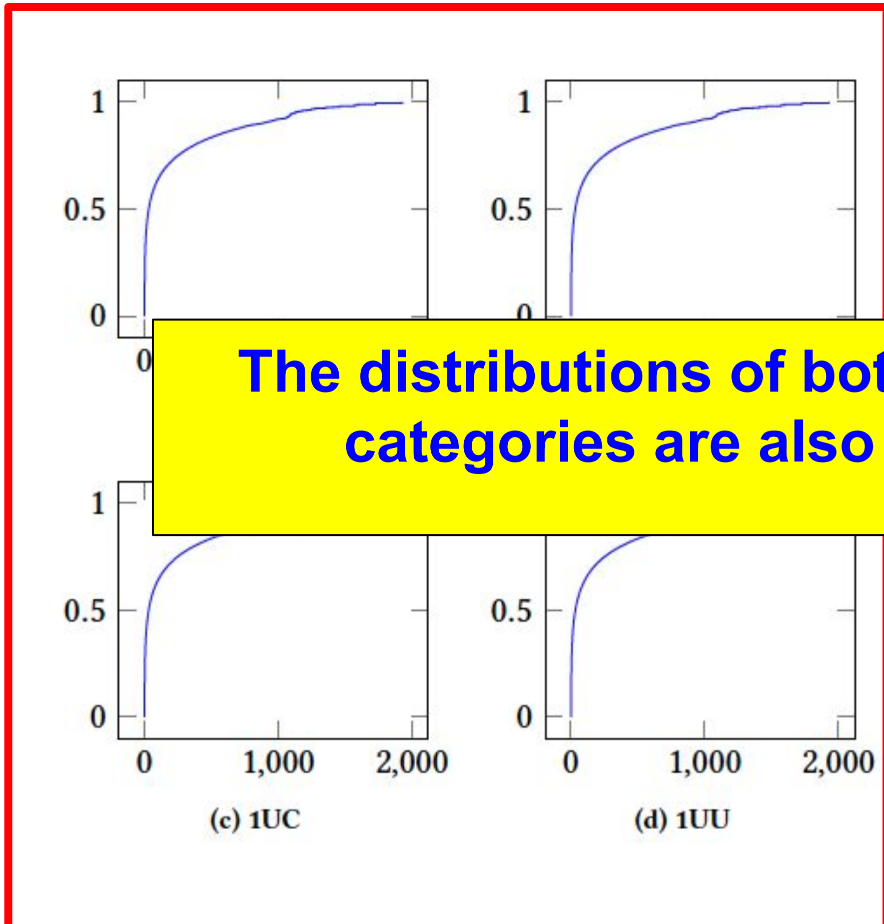
(h) 2UU

Experiments

Gunosy

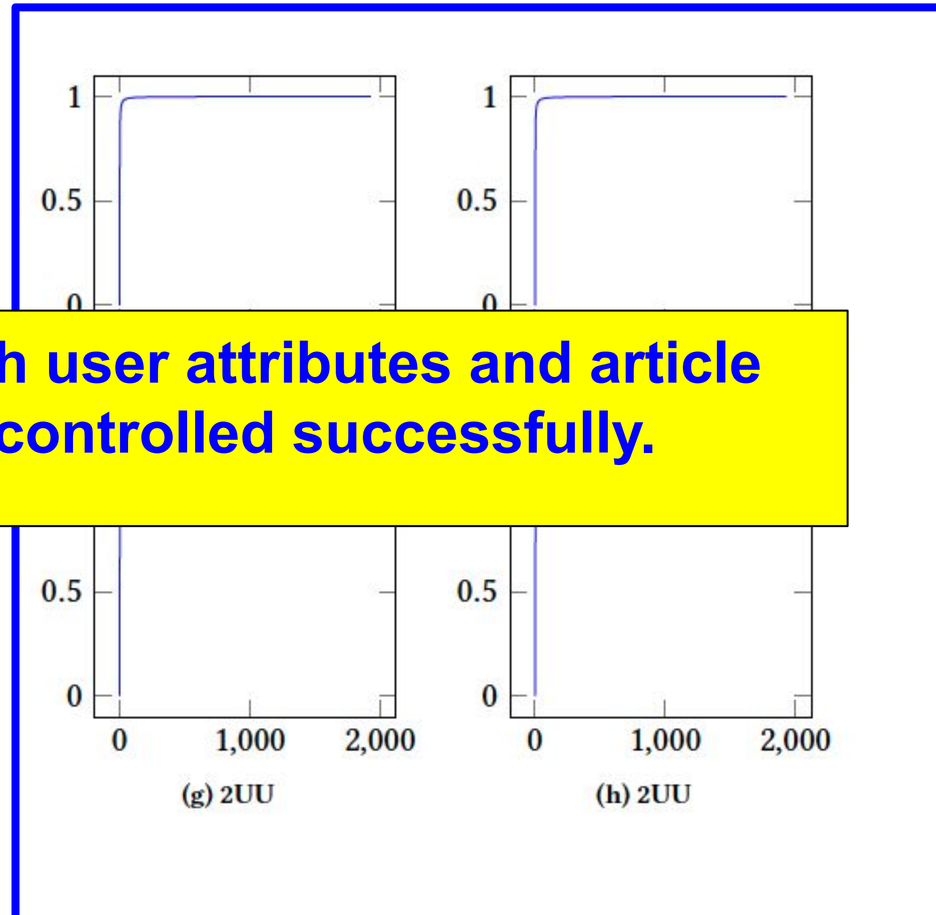
We successfully controlled the click distributions.

Zipf(1)'s distribution



(c) 1UC (d) 1UU

Zipf(2)'s distribution



(g) 2UU (h) 2UU

The distributions of both user attributes and article categories are also controlled successfully.

Experiment

Gunosy

Comparing algorithms evaluations for each dataset

The performance of the algorithms differed in how the datasets were built.

Table 2: Comparison of the algorithm NDCG@10 evaluations for each datasets

	#Click	Attributes	Category	Random	TopPopular	Item-KNN	Item2vec	BPR-MF	GRU4REC
uniform		-		0.219	0.338	0.58	0.581	0.43	0.429
1UU	Zipf(1)			0.218	0.289	0.552	0.583	0.418	0.392
1CU	Zipf(1)	o		0.221	0.291	0.549	0.592	0.442	0.451
1UC	Zipf(1)		o	0.219	0.223	0.518	0.584	0.415	0.423
1CC	Zipf(1)	o	o	0.219	0.205	0.52	0.583	0.402	0.415
2UU	Zipf(2)			0.219	0.283	0.413	0.333	0.349	0.437
2CU	Zipf(2)	o		0.213	0.306	0.412	0.331	0.364	0.337
2UC	Zipf(2)		o	0.201	0.260	0.408	0.315	0.344	0.344
2CC	Zipf(2)	o	o	0.217	0.264	0.438	0.332	0.335	0.346

Experiment

Gunosy

Comparing algorithms evaluations for each dataset

Evaluations on Zipf(1)'s datasets were **similar to uniform.**

Table 2: Comparison of the algorithm NDCG@10 evaluations for each datasets

Second Best

	#Click	Attributes	Category	Random	TopPopular	Item-KNN	Item2vec	BPR-MF	GRU4REC
uniform		-		0.219	0.338	0.58	0.581	0.43	0.429
1UU	Zipf(1)			0.218	0.289	0.552	0.583	0.418	0.392
1CU	Zipf(1)	o		0.221	0.291	0.549	0.592	0.442	0.451
1UC	Zipf(1)		o	0.219	0.223	0.518	0.584	0.415	0.423
1CC	Zipf(1)	o	o	0.219	0.205	0.52	0.583	0.402	0.415
2UU	Zipf(2)			0.219	0.283	0.413	0.333	0.349	0.437
2CU	Zipf(2)	o		0.213	0.306	0.412	0.331	0.364	0.337
2UC	Zipf(2)		o	0.201	0.260	0.408	0.315	0.344	0.344
2CC	Zipf(2)	o	o	0.217	0.264	0.438	0.332	0.335	0.346

Experiment

Gunosy

Comparing algorithms evaluations for each dataset

Evaluation results on Zipf(2)'s datasets were worse than Zipf(1)'s.
This may because Zipf(2) datasets were sparse.

Table 2: Comparison of the algorithm NDCG@10 evaluations for each datasets

	#Click	Attributes	Category	Random	TopPopular	Item-KNN	Item2vec	BPR-MF	GRU4REC
uniform		-		0.219	0.338	0.58	0.581	0.43	0.429
1UU	Zipf(1)			0.218	0.289	0.552	0.583	0.418	0.392
1CU	Zipf(1)	o		0.221	0.291	0.549	0.592	0.442	0.451
1UC	Zipf(1)		o	0.219	0.223	0.518	0.584	0.415	0.423
1CC	Zipf(1)	o	o	0.219	0.205	0.52	0.583	0.402	0.415
2UU	Zipf(2)			0.219	0.283	0.413	0.333	0.349	0.437
2CU	Zipf(2)	o		0.213	0.306	0.412	0.331	0.364	0.337
2UC	Zipf(2)		o	0.201	0.260	0.408	0.315	0.344	0.344
2CC	Zipf(2)	o	o	0.217	0.264	0.438	0.332	0.335	0.346

Best

Experiment

Comparing algorithms evaluations for each dataset

It is necessary to select sampling settings according to the purpose, and it may be important to publish datasets with various settings.

Table 2: Comparison of the algorithm NDCG@10 evaluations for each datasets

	#Click	Attributes	Category	Random	TopPopular	Item-KNN	Item2vec	BPR-MF	GRU4REC
uniform		-		0.219	0.338	0.58	0.581	0.43	0.429
1UU	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
1CU	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
1UC	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
1CC	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
2UU	Zipf(2)			0.219	0.283	0.413	0.333	0.349	0.437
2CU	Zipf(2)	o		0.213	0.306	0.412	0.331	0.364	0.337
2UC	Zipf(2)		o	0.201	0.260	0.408	0.315	0.344	0.344
2CC	Zipf(2)	o	o	0.217	0.264	0.438	0.332	0.335	0.346

It is necessary to select sampling settings according to the purpose, and it may be important to publish datasets with various settings.

Conclusion

Gunosy

This study is the first attempt to reduce business risks in publishing datasets

1. summarizing the challenges of building and publishing datasets from commercial service.
2. formulating the problem of building and publishing a dataset as a optimization problem that seeks the sampling weight of users.
3. appling our method to build datasets from the raw data of our real-world mobile news delivery service

Limitations & Future Works

- We did not give a theoretical guarantee if the impossibility of the estimation. Providing such an impossibility is an important.
- This study only considered the user-item interactions. However real world services may have different types of behavior logs.

Conclusion

Gunosy

This study is the first attempt to reduce business risks in publishing datasets

1. **summarizing the challenges of building and publishing datasets from commercial service.**
2. **formulating the problem of building and publishing a dataset as a optimization problem that seeks the sampling weight of users.**
3. **appling our method to build datasets from the raw data of our real-world mobile news delivery service**

Limitations & Future Works

- We did not give a theoretical guarantee if the impossibility of the estimation. Providing such an impossibility is an important.
- This study only considered the user-item interactions. However real world services may have different types of behavior logs.

Conclusion

Gunosy

This study is the first attempt to reduce business risks in publishing datasets

1. summarizing the challenges of building and publishing datasets from commercial service.
2. formulating the problem of building and publishing a dataset as a optimization problem that seeks the sampling weight of users.
3. appling our method to build datasets from the raw data of our real-world mobile news delivery service

Limitations & Future Works

- **We did not give a theoretical guarantee if the impossibility of the estimation. Providing such an impossibility is an important.**
- **This study only considered the user-item interactions. However real world services may have different types of behavior logs.**

Conclusion

Gunosy

This study is the first attempt to reduce business risks in publishing datasets

Previously, researchers has not disclosed how to build the dataset and has not shared the knowledge with the community.

We hope that our work will lead to more discussions on the process of building and publishing datasets and that many datasets will be published.

Feel free to contact me: yoshifumi.seki@gunosy.com

our implementation and dataset avaiavle

https://github.com/gunosy/publishing-dataset-recsys20

Gunosy

情報を世界中の人に最適に届ける

Mathematical Formulation

Gunosy

We formulate our task as a problem of finding the sampling weight of users.

$u \in U$: a user in user behavior logs

$m \ll |U|$: number of users in the building dataset

$w(u)$: non-negative weight for each user,

We sample m users without replacement, where user u is included in the samples with probability proportional w(u)

Mathematical Formulation

Gunosy

We formulate our task as a problem of finding the sampling weight of users.

$u \in U$: a user in user behavior logs

$m \ll |U|$: number of users in the building dataset

$w(u)$: non-negative weight for each user,

We sample m users without replacement, where user u is included in the samples with probability proportional w(u)

We represent our three challenges as three loss functions.

L_{click} : anonymizing business metrics

$L_{attribute}$: maintaining fairness

$L_{categories}$: removing popularity bias

minimize the weighted sum of the loss functions.