Constructing Better Evaluation Metrics by Incorporating the Anchoring Effect into the User Model

Nuo Chen [1] Fan Zhang [2] Tetsuya Sakai [1]

[1]Waseda University, Japan

[2]Wuhan University, China

July 2022

Summary

- **Behavioral Economics + Interactive IR**
- Information searchers are not always rational. They are influenced by cognitive biases, including **the anchoring effect**
- Incorporating **the anchoring effect** into the user model of current evaluation metrics can make them correlate better with the user satisfaction feedback

Anchoring Effect

Cognitive Biases

- Systematic pattern of deviations in thinking.[1]
- Decisions can be different from cases where the decision maker is rational[2][3]
- **Anchoring**: judgements and decisions are "**biased toward the initial values**" and "**different starting points yield different estimates**"[4]

---

[1] Amos Tversky and Daniel Kahneman. "Judgment under Uncertainty: Heuristics and Biases". In: *Science* 185.4157 (1974), pp. 1124–1131.

[2] Amos Tversky and Daniel Kahneman. "Loss Aversion in Riskless Choice: A Reference-Dependent Model". In: *Quarterly Journal of Economics* 106 (1991), pp. 1039–1061.

[3] Amos Tversky and Daniel Kahneman. "Advances in prospect theory: Cumulative representation of uncertainty". In: *Journal of Risk and Uncertainty* 5 (1992), pp. 297–323.

[4] Tversky and Kahneman, "Judgment under Uncertainty: Heuristics and Biases".

Anchoring Effect

An Example

- An experiment by Tversky & Kahneman(1974)[5]:
  - $1 * 2 * 3 * 4 * 5 * 6 * 7 * 8 =?$
  - $8 * 7 * 6 * 5 * 4 * 3 * 2 * 1 =?$
- The median estimate was **512** for the subjects in the first group and **2250** for those in the second group

[5] Tversky and Kahneman, "Judgment under Uncertainty: Heuristics and Biases".

Anchoring Effect

In Interactive IR

- Shokouhi *et al.* (2015)[6]:
  - If previous document is highly or somewhat relevant, the same relevance label is more likely to be assigned to the next document.
- Thomas *et al.* (2022)[7]:
  - The first relevance label has a relatively more frequent appearance in the subsequent labels.
  - The second label is more likely to match the first.

---

[6] Milad Shokouhi, Ryen White, and Emine Yilmaz. "Anchoring and Adjustment in Relevance Estimation". In: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '15. 2015, pp. 963–966.

[7] Paul Thomas et al. "The Crowd is Made of People: Observations from Large-Scale Crowd Labelling". In: *ACM SIGIR Conference on Human Information Interaction and Retrieval*. CHIIR '22. 2022, pp. 25–35.

Proposed Framework

Simulating the Anchoring Effect

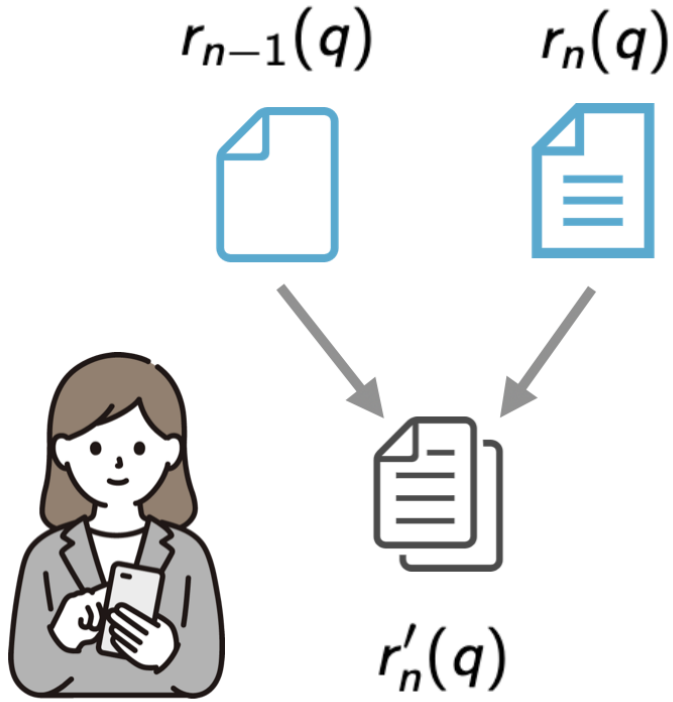
- We define the relevance level users perceived under anchoring effect as the following:
-

$$r'_n(q) = \begin{cases} r_n(q) & \text{for} \quad n = 1, \\ \alpha \cdot r_{n-1}(q) + (1-\alpha) \cdot r_n(q) & \text{otherwise} \end{cases} \tag{1}$$

- $r_n(q)$ is the "unbiased" relevance
- α is the factor of the anchoring effect. The larger α , the greater the anchoring effect

Proposed Framework

Simulating the Anchoring Effect



Proposed Framework

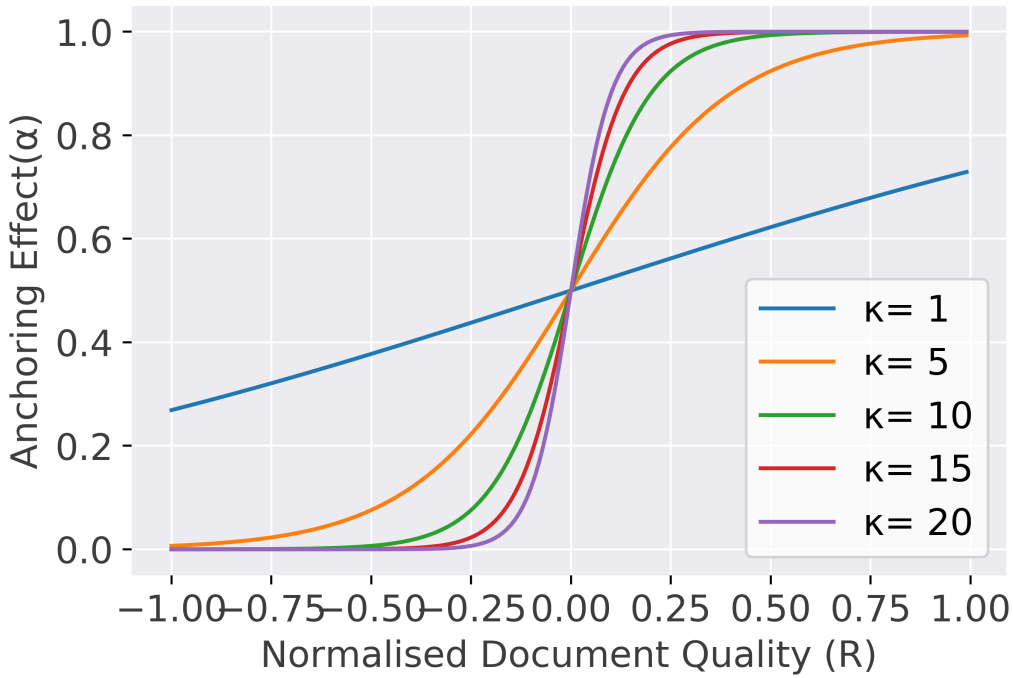
Simulating the Anchoring Effect

- Shokouhi et al. (2015) suggested that the factor of anchoring effect is related to the quality of the previous document:
- $$\alpha = \frac{\lambda}{1 + \exp(-\kappa R_{n-1}(q))} \tag{2}$$
- $R_{n-1}(q)$ is a normalization of the "unbiased" relevance level r_{n-1} ranging in $[-1, 1]$
- λ stands for the upper bound of the factor of the anchoring effect
- κ stands for how fast the anchoring effect grows as the document quality increases

Proposed Framework

Simulating the Anchoring Effect

The curve of the anchoring effect when $\lambda = 1$ and $\kappa = 0.05$, 0.15, 0.25, 0.35, 0.45 respectively.



Experiment

Dataset

THUIR-1[8]

- 2, 435 single-SERP sessions
- Click-through logs
- Query-level satisfaction feedbacks
- 4-level graded relevance labels

[8] Ye Chen et al. "Meta-evaluation of Online and Offline Web Search Evaluation Metrics". In: *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2017).

Experiment

Metrics

We apply the framework of **Anchoring-awared Metrics (AMs)** to the following metrics and compare the performance of the *vanilla* and the *AM* for each metric

- ERR
- Precision
- (scaled)DCG
- RBP
- INSQ
- INST

Experiment

Experiment 1

- Randomly split the data into 5 folds.
  - 4 folds for training, 1 fold for testing.
  - Repeat 10 times
- Tuning parameters
  - By User Behavior (impression depth)
  - By User Satisfaction
- Test

Experiment

Experiment 2

- For each query in the dataset, we compute the score of SERP given by the *vanilla* and the *AM* respectively
- Metric parameters: the mean of each parameter optimized in Experiment 1
- $2,391 * 2,390/2 = 2,857,245$ different SERP pairs in total.
- Collect the count of SERP pairs where the preferences given by the *vanilla* and the *AM* are different

Results

Analysis of the Parameters

	AM-ERR		AM-Precision		AM-scaled-DCG			AM-RBP			AM-INSQ			AM-INST		
	λ	κ	λ	κ	b	λ	κ	p	λ	κ	T	λ	κ	T	λ	κ
Mean	1.00	19.10	1.00	17.10	1.90	1.00	13.30	0.85	1.00	12.20	12.00	1.00	6.70	12.60	1.00	5.20
σ	0.000	0.316	0.000	1.853	0.037	0.000	6.325	0.000	0.000	8.230	9.298	0.00	5.165	6.667	0.00	1.033

- The large value of λ shows the strong influence of the anchoring effect
- κ is unstable and varies among different metrics.

Results

The Effectiveness of *AM*s

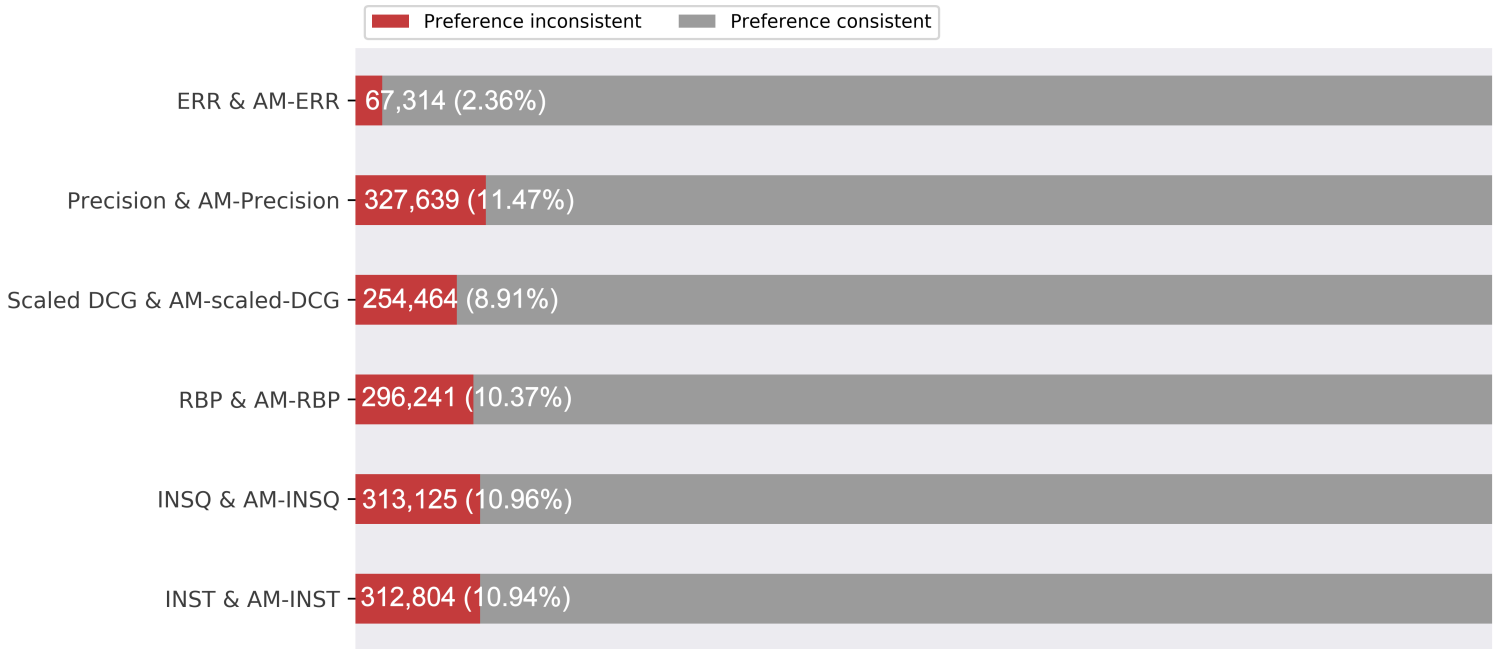
- All of the *AMs* outperformed the *vanilla*
- We can better predict satisfaction feedbacks by incorporating the anchoring effect

Metric	UB	US
ERR	0.311^{**}	
AM-ERR	**0.315**	
Precision	0.241^{***}	
AM-Precision	**0.306**	
scaled-DCG	0.322^{***}	0.324^{***}
AM-scaled-DCG	**0.348**	
RBP	0.313^{***}	0.331^{***}
AM-RBP	**0.347**	
INSQ	0.290^{***}	0.326
AM-INSQ	**0.333**	
ISNT	0.282^{***}	0.330
AM-INST	**0.332**	

Results

Different SERP preference between the *AM* and the *vanilla*

Around 2% to 11% of the conclusions drawn by current metrics are overridden after considering the anchoring effect.



Future Research

- A better simulation of the anchoring effect
- The anchoring effect in multi-queries session and diverse tasks
- Help users succeed through the understanding of cognitive biases

Thank You for Your Attention

Nuo Chen
pleviumtan@toki.waseda.jp
Real Sakai Lab,
Department of Computer Science and Engineering,
Waseda University