How to Make an Outlier? Studying the Effect of Presentational Features on the Outlierness of Items in Product Search Results

Fatemeh Sarvi
AIRLab, University of Amsterdam, The Netherlands
f.sarvi@uva.nl

Mohammad Aliannejadi
University of Amsterdam, The Netherlands
m.aliannejadi@uva.nl

Sebastian Schelter
University of Amsterdam & Ahold Delhaize, The Netherlands
s.schelter@uva.nl

Maarten de Rijke
University of Amsterdam, The Netherlands
m.derijke@uva.nl

ABSTRACT

In two-sided marketplaces, items compete for attention from users since attention translates to revenue for suppliers. Item exposure is an indication of the amount of attention that items receive from users in a ranking. It can be influenced by factors like position bias. Recent work suggests that another phenomenon related to inter-item dependencies may also affect item exposure, viz. *outlier items* in the ranking. Hence, a deeper understanding of outlier items is crucial to determining an item's exposure distribution. In this work, we study the impact of different presentational e-commerce features on users' perception of outlierness of an item in a search result page. Informed by visual search literature, we design a set of crowdsourcing tasks where we compare the observability of three main features, viz. price, star rating, and discount tag. We find that various factors affect item outlierness, namely, visual complexity (e.g., shape, color), discriminative item features, and value range. In particular, we observe that a distinctive visual feature such as a colored discount tag can attract users' attention much easier than a high price difference, simply because of visual characteristics that are easier to spot. Moreover, we see that the magnitude of deviations in all features affects the task complexity, such that when the similarity between outlier and non-outlier items increases, the task becomes more difficult.

CCS CONCEPTS

• **Information systems** → *Users and interactive retrieval.*

KEYWORDS

Product search; Fairness; Outliers

ACM Reference Format:
Fatemeh Sarvi, Mohammad Aliannejadi, Sebastian Schelter and Maarten de Rijke. 2023. How to Make an Outlier? Studying the Effect of Presentational Features on the Outlierness of Items in Product Search Results. In *ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '23), March 19–23, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3576840.3578278

CHIIR '23, March 19–23, 2023, Austin, TX, USA

ACM ISBN 979-8-4007-0035-4/23/03.
https://doi.org/10.1145/3576840.3578278

1 INTRODUCTION

In two-sided marketplaces items compete for attention from users since attention translates to revenue for suppliers. Item exposure is an indication of how much attention each item receives from users. Effective estimation of item exposure is crucial for challenges such as item fairness [4, 5, 13, 14, 17, 20, 21, 25] and bias in counterfactual learning to rank [1, 8, 9, 15, 16, 23]. Various modeling assumptions have been proposed for item exposure estimation in ranking. The position-based assumption [8, 9] is a widely accepted model according to which the higher the position of an item is, the more exposure it receives.

To the best of our knowledge, the modeling assumptions made to estimate item exposure include inter-item independence and define exposure as a function of an item's position in a ranking. However, recent research has introduced another phenomenon that accounts for a particular type of inter-item dependency [18]: the existence of *outlier items* in a ranked list may affect the exposure that all items in the list receive. Here, outlier items are items that observably deviate from the other items in a ranked list w.r.t. task-specific, presentational features, like the price of a product in product search.

Sarvi et al. [18] show that the presence of outlier items may result in more (or less) attention being focused on and around the outlier item compared to a list without any outliers. E.g., on an e-commerce search result page, adding a red discount tag as a discriminative feature, to only one product can attract more attention to it irrespective of its position or relevance to the query, thereby changing the exposure distribution estimated based on position-based assumptions. Although Sarvi et al. [18] provide critical intuitions on how outliers affect user behavior, they do not study how different presentational features are actually perceived by users in this context.

The perception and visual search communities have conducted many studies into how the human brain can immediately identify recognizable objects like outliers in an image and how different visual attributes (e.g., color, shape) can add to the complexity of this task [7, 19, 22, 24]. Informed by these studies, we design our task as a visual search process where the objective is to find a target (i.e., outlier items) among distractors (i.e., non-outlier items), as fast as possible.

Since presentational features can be composed of multiple visual attributes, it is not trivial how they are perceived by users when they act as a discriminative feature. To fill this gap, we compare different presentational features from the e-commerce search domain. Our goal is to provide insights into how efficiently and accurately

users detect outlier items w.r.t. these features. We consider three observable features commonly used in e-commerce, viz. price, star rating, and discount tag. Previous work has shown the importance of these features in influencing users' purchase decisions [2, 10].

Research questions. We aim to answer the following research questions: (RQ1) How do deviations in the presentational features of an item in a ranked list contribute to users perceiving the item as an outlier? (RQ2) To what extent do the visual attributes (color, shape, value) of each presentational feature affect the detectability of the outlier?

We find that presentational features have different degrees of observability and influence on items' outlierness. E.g., a bright red background color of a discount tag makes it easier to spot compared to price which is shown as a number with a regular font size and color. Moreover, the magnitude of deviations of observable features affects the task complexity. This means that when the similarity between outlier and non-outlier items increases (e.g. closer values for price in a list), the task becomes more difficult.

2 METHOD

In this section, we describe the details of our crowdsourcing tasks.

2.1 Overview

We design our tasks as a visual search process [7, 19, 22, 24], where the objective is to find a target among distractors. We focus on the domain of e-commerce search, where the distractors are non-outlier products in a ranking, and the target is the outlier products that differ in at least one presentational feature. We compare three presentational features, namely, price, star rating, and discount tag. When considering a discount tag, our task is close to a disjunctive search process known from visual search [22] that focuses on detecting a target that differs from the distractors in terms of a unique visual feature, e.g., the discount tag [12]. When regarding price and star rating, our task is closer to conjunction search [22], where the distractors exhibit at least one common feature with the target [19]. However, unlike conjunction search, in our task the difference between the target and the distractors is in the values of the features not the features themselves (e.g., the value of the product's price). In the rest of this paper we refer to the target item as *outlier*.

Following previous work [6, 22], we use *reaction time* (RT) and *accuracy* to measure the effort it takes to detect the target (outlier) among its distractors. The goal of this task is to examine and determine which presentational features are easier to detect by the workers, i.e., the shorter the RT to find the outlier, the easier it is.

We perform our experiments using two tasks, where we build several synthetic product search result pages and examine how each feature contributes to the outlierness of an item, both separately and simultaneously. Below, we describe different stages of our two tasks.

2.2 Instructions

In the instructions, we describe the overall goal of the research and the concept of an outlier in a search result page, providing tangible examples. We ask participants to scan and compare all items in a list and flag outliers as *fast* as they can. We also ask them to fill out a questionnaire after completing the tasks.

2.3 Task I

In the first task, we evaluate how fast any of the three presentational features (price, star rating, discount tag) can be spotted on a search result page. To this end, we explicitly describe and mention the one feature at a time to the participants and ask them to scan the list and find up to two outlier items, *only* with respect to the given feature. For instance, after providing a definition of outliers in the instruction, we mention that there are one or two outliers in terms of different values for price in the list and that they have to find them as fast as they can. We place one of the outliers at the top of the list and the other at the bottom. To avoid position and randomness bias, we keep the position of the outlier items fixed while other items are randomly placed in the list.

Wolfe [24] suggests that visual features including luminance, color, and orientation affect the RT in a visual search task. Following this work and inspired by the experiments in [22], we tested two variations of Task I, namely Type I and Type II, where we change the shape, color and value of the presentational features to study different magnitudes of deviation of the outliers from the rest. In Type I, we use features that more strongly discriminate between the outlier and the rest compared to Type II. For example, an outlier w.r.t. price can be 10 or 2 times more expensive than other products. We use the former in Type I and the latter in Type II. The same goes for star rating. Regarding the discount tag feature we use the suggestions by Wolfe [24] to distinguish between the outlier items of Type I and Type II. In Type I we use a bright red color as background with a bold white font stating that there is a special deal on the product, whereas, in Type II, we use a light green text without any background stating a 10% discount.

2.4 Task II

Unlike Task I, here we aim to examine the relative RT for the three features (price, star rating, discount tag) when presented to the users simultaneously. To better compare the three observable features, we jointly present the different combinations of these features and analyze the behavior of the users. While describing the three target features in this task, we do not mention to participants which features are being examined. Therefore, the workers are supposed to go through the list, examine all items with respect to all the features used in presenting the results, and then decide which items are outliers. Note that there are more than three features used to describe each item, for example, we used image, title and delivery information next to the price, discount tag and star rating. Moreover, we indicate that the workers have to mark a maximum of three items as outliers. Here, we also randomize the position of the outlier items while making sure that they appear both at the top and bottom of the list. We run the task for all combinations of at least two of the three features.

2.5 Page examination behavior

We record several signals related to participants' page examination behavior and their interaction data, such as mouse hovering, scrolling, viewed items, clicks, and time spent on task. To gain a more accurate estimate of RT, we ask the participants to click on a

Start button after reading the instructions. The search result page appears only after the Start button has been clicked. We compute the task completion time from the moment the workers click the Start button.

2.6 Post-task questionnaire

We ask participants to fill out a questionnaire after completing the task. To gain more insight into workers' background and online shopping behavior, we instruct them to fill out questions on their demographics and familiarity with online shopping. Moreover, to ensure that the workers understand outlier definition, we ask them to answer a question about the definition of an outlier in search.

2.7 Quality control

We follow three strategies for quality control. As part of the post-task questionnaire, we ask a multiple-choice question on the definition of outliers. All participants managed to answer this question correctly. Also, following Kittur et al. [11], we ask workers to justify their choice in a few keywords. We only remove the responses of those participants who entered random tokens as justifications of their answers. We also remove the responses of those who revisit the instructions more than two times while performing the task, since response time is crucial in this study.

2.8 Participants

We use Amazon Mechanical Turk as the platform for our crowd-sourcing experiments, with workers based only in the U.S., with an approval rate of 95% or greater. After quality control, we are left with 140 assignments (92 for Task I and 48 for Task II), submitted by 80 distinct participants. From the participants, 45% are female, 53% male, and 2% listed other genders. The majority of participants (74%) are between 25 and 44 years old, with 5% younger and 21% older workers.

3 RESULTS

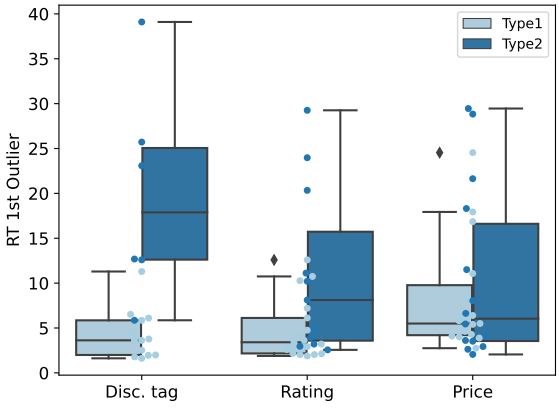
In this section, we present the results of our crowd-sourcing tasks in terms of the performance and behavior of the workers under different experimental conditions.

Table 1: Worker performance metrics in terms of RT and accuracy. Average (Avg.) and median (Med.) RT in seconds is reported for the first and second outlier (out. 1 & out. 2).

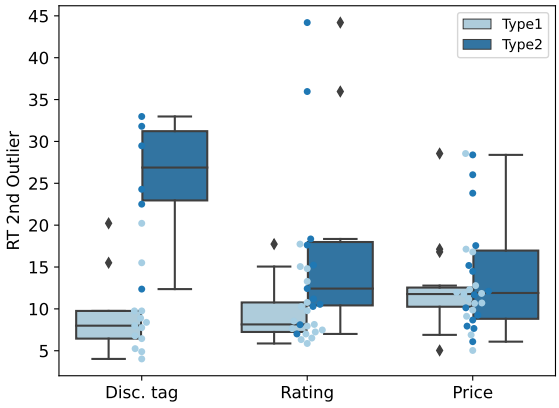
Type		RT out. 1		RT out. 2		Acc.
		Avg.	Med.	Avg.	Med.	
Type 1	Disc. tag	4.22	3.62	8.90	8.00	0.98
	Star rating	4.81	3.41	9.67	8.14	0.97
	Price	8.38	5.50	12.44	11.77	1.00
Type 2	Disc. tag	19.84	17.88	25.57	26.88	0.99
	Star rating	10.96	8.11	17.36	12.42	0.99
	Price	10.62	6.03	14.21	11.90	0.98

Reaction time and accuracy. Following [6, 22], we use reaction time (RT) and accuracy to measure the effort it takes to detect the outlier among non-outlier distractors. Table 1 summarizes participants' average responses to Task I in terms of RT and accuracy. Accuracy is high for all variations of Task I with a maximum of 1.00 for price in Type I and a minimum of 0.97 for star rating. We conclude from the high accuracy values that the workers grasped the concept of outlier in a ranked list and were able to accurately find them in the list. Next, we compare the time that the workers take to spot the outliers. Table 1 shows that for Type I outliers participants spotted the discount tags faster than the other two features in Task I. This is followed by star rating with a slightly higher recorded RT. As expected, we see that on average it took participants almost twice as long to find the price outliers. We conducted a one-way ANOVA test on RT for first outliers in Type I. Results show that the differences are statistically significant with p-value < 0.05.

Detecting discount tags is similar to a disjunctive visual search process, which has been shown to be easier to solve compared to conjunctive search (i.e., star rating and price) [22]. Moreover, users



(a)



(b)

Figure 1: Distribution of the RT for the (a) first and (b) second outliers in both variations of Task I.

favor simple options when they act under time pressure [3], which can lead to being biased towards easy-to-detect visual features such as discount tag. The higher RT for price can be attributed to the fact that certain visual features, including color and shape, are processed early in the brain using pre-attentive processes [22]. Star rating and discount tag have more visual characteristics regarding shapes and colors, however price is more simply presented in the product descriptions.

Another related aspect is the unknown range of the price values. This is less crucial for star rating or discount tag since the former has a range between 1 to 5 and latter is a binary feature.

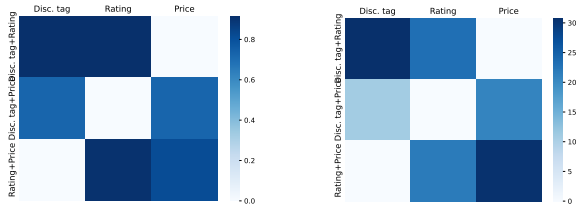
Type I vs. Type II. Next, we compare the results of Type I and Type II outliers. Our goal is to understand how much changing the magnitude of the deviations in terms of different features affect user performance. One can compare different ranges of deviations on specific features to model the relationship between the the deviations and user performance, but we leave this as future work and only compare two variations. The results in the upper and lower parts of Table 1 suggest that the reduced magnitude of deviations in all features leads to higher RT. Duncan and Humphreys [6] study the same effect by pointing out that when outlier to non-outlier similarity increases, the task becomes more difficult. Similarly, we see that RT increases for all the features, and for both the first and second outliers.

Moreover, we see in Figure 1 how the RT distribution of the two outlier variants differ for Task I. As expected, the plots show a higher RT for all features, and both outliers. However, it is interesting to note that we observe the lowest relative effect on the price (26.73% increase), compared to star rating (127.86%) and discount tag (370.14%). We relate this to the visual nature of discount tag and star ratings. Reducing the color contrast of discount tag would have a greater impact on the user's ability to detect it among the distractors, compared to a different price ranges as the user still has to carefully check the prices to detect the outlier. Regarding the accuracy we see no drop, suggesting that even a more subtle deviation in observable feature can be detected by many users.

Feature combinations. Figure 2a shows recall values for combinations of features, where the y-axis indicates recall of a combination of features and x-axis indicates the value for a specific feature. As expected, detecting the outlier w.r.t. price is more difficult for participants (on average, 1.3% and 7.3% lower values than for discount tag and star rating). In terms of RT, Figure 2b confirms our findings in Table 1, except for the combination of discount tag and star rating, where on average participants found star rating ~7.5 seconds faster than discount tag. Perhaps, it is because the average position of the outlier w.r.t. discount tag is lower than star rating (12 and 9.5, respectively).

4 DISCUSSION & CONCLUSIONS

We compared three presentational features from the e-commerce search domain and study their effect on attracting users' attention in a search result page, hence making an item outlier. Informed by visual search processes [7, 19, 22, 24], we designed a series of tasks where the objective is to find a target among distractors. We instructed participants to find outliers in a result page, as fast as possible. Below, we answer our research questions.



(a) (b)

Figure 2: (a) Recall and (b) RT for combinations of observable features. Y-axis shows the metric of the corresponding feature and x-axis shows the second feature used in the combination.

RQ1. How do deviations in the presentational features of an item in a ranked list contribute to users perceiving the item as an outlier?. The complexity of observable features in terms of visual characteristics and value range affects their observability to users. Features that are more visually distinct from the rest of the list and/or have a discriminative feature have a better chance to stand out. Value range also plays an important role where features with a known range of values are easier to detect when deviating from the rest compared to features with an unknown value range. When different features deviate simultaneously, they lead to more complex scenarios where usually the most visually observable feature tends to dominate users' attention.

RQ2. To what extent do the visual attributes (color, shape, value) of each presentational feature affect the detectability of the outlier?. Different levels of a feature's deviation affect its observability; however, the phenomenon is complex to model. In the case of discount tags, a contrasting color plays a role. This can further be quantified, which we leave for future work. As for price and star rating, we modified the relative numerical difference with the rest of the list and observed different reactions from the participants, suggesting that future research in this area should aim at quantifying and modeling the definition of observability and its impact on user perception.

Although our work informs how different features contribute to outlierness of an item, it cannot directly be used to estimate outlier item exposure. In the future, we aim to design an in-situ user study to track and model the impact of outliers on item exposure.

ACKNOWLEDGMENTS

This research was supported by Ahold Delhaize and by the Hybrid Intelligence Center, a 10-year program funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

REFERENCES

[1] Aman Agarwal, Xuanhui Wang, Cheng Li, Michael Bendersky, and Marc Najork. 2019. Addressing Trust Bias for Unbiased Learning-to-rank. In *WWW*. 4–14.
[2] Praveen Aggarwal and Rajiv Vaidyanathan. 2016. Is Font Size a Big Deal? A Transaction–Acquisition Utility Perspective on Comparative Price Promotions.

Journal of Consumer Marketing (2016).
[3] Leif Azzopardi. 2021. Cognitive Biases in Search: A Review and Reflection of Cognitive Biases in Information Retrieval. In *CHIIR*. ACM, 27–37.
[4] Asia J Biega, Krishna P Gummadi, and Gerhard Weikum. 2018. Equity of Attention: Amortizing Individual Fairness in Rankings. In *SIGIR*. 405–414.
[5] Fernando Diaz, Bhaskar Mitra, Michael D Ekstrand, Asia J Biega, and Ben Carterette. 2020. Evaluating Stochastic Rankings with Expected Exposure. In *CIKM*. 275–284.
[6] John Duncan and Glyn W Humphreys. 1989. Visual Search and Stimulus Similarity. *Psychological review* 96, 3 (1989), 433.
[7] Loann Giovannangeli, Romain Bourqui, Romain Giot, and David Auber. 2022. Color and Shape Efficiency for Outlier Detection from Automated to User Evaluation. *Visual Informatics* (2022).
[8] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2005. Accurately Interpreting Clickthrough Data as Implicit Feedback. In *SIGIR*. 154–161.
[9] Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. 2017. Unbiased Learning-to-rank with Biased Feedback. In *WSDM*. 781–789.
[10] Karen C Kao, Sally Rao Hill, and Indrit Troshani. 2020. Effects of Cue Congruence and Perceived Cue Authenticity in Online Group Buying. *Internet Research* (2020).
[11] Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing User Studies with Mechanical Turk. In *Proceedings of the 2008 Conference on Human Factors in Computing Systems, CHI*. ACM, 453–456.
[12] Brian McElree and Marisa Carrasco. 1999. The Temporal Dynamics of Visual Search: Evidence for Parallel Processing in Feature and Conjunction Searches. *Journal of Experimental Psychology: Human Perception and Performance* 25, 6 (1999), 1517.
[13] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. 2018. Towards a Fair Marketplace: Counterfactual Evaluation of the Trade-off between Relevance, Fairness & Satisfaction in Recommendation Systems. In *CIKM*. 2243–2251.
[14] Marco Morik, Ashudeep Singh, Jessica Hong, and Thorsten Joachims. 2020. Controlling Fairness and Bias in Dynamic Learning-to-rank. In *SIGIR*. 429–438.
[15] Alamir Novin and Eric M. Meyers. 2017. Making Sense of Conflicting Science Information: Exploring Bias in the Search Engine Result Page. In *CHIIR*. ACM, 175–184.
[16] Zohreh Ovaisi, Ragib Ahsan, Yifan Zhang, Kathryn Vasilaky, and Elena Zheleva. 2020. Correcting for Selection Bias in Learning-to-rank Systems. In *WWW*. 1863–1873.
[17] Piotr Sapiezynski, Wesley Zeng, Ronald E Robertson, Alan Mislove, and Christo Wilson. 2019. Quantifying the Impact of User Attentionon Fair Group Representation in Ranked Lists. In *WWW*. 553–562.
[18] Fatemeh Sarvi, Maria Heuss, Mohammad Aliannejadi, Sebastian Schelter, and Maarten de Rijke. 2022. Understanding and Mitigating the Effect of Outliers in Fair Ranking. In *WSDM*. 861–869.
[19] Jiye Shen, Eyal M Reingold, and Marc Pomplun. 2003. Guidance of Eye Movements during Conjunctive Visual Search: The Distractor-ratio Effect. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale* 57, 2 (2003), 76.
[20] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of Exposure in Rankings. In *KDD*. 2219–2228.
[21] Ashudeep Singh and Thorsten Joachims. 2019. Policy Learning for Fairness in Ranking. In *NeurIPS*.
[22] Anne M Treisman and Garry Gelade. 1980. A Feature-integration Theory of Attention. *Cognitive psychology* 12, 1 (1980), 97–136.
[23] Ali Vardasbi, Harrie Oosterhuis, and Maarten de Rijke. 2020. When Inverse Propensity Scoring does not Work: Affine Corrections for Unbiased Learning to Rank. In *CIKM*. 1475–1484.
[24] Jeremy M Wolfe. 1998. What Can 1 Million Trials Tell Us about Visual Search? *Psychological Science* 9, 1 (1998), 33–39.
[25] Himank Yadav, Zhengxiao Du, and Thorsten Joachims. 2019. Policy-Gradient Training of Fair and Unbiased Ranking Functions. *arXiv preprint arXiv:1911.08054* (2019).