UNDR: User-Needs-Driven Ranking of Products in E-Commerce

Andrea Papenmeier
andrea.papenmeier@uni-due.de
University of Duisburg-Essen
Cologne, Germany

Daniel Hienert
daniel.hienert@gesis.org
GESIS – Leibniz Institute for the Social Sciences
Cologne, Germany

Firas Sabbah
firas.sabbah@uni-due.de
University of Duisburg-Essen
Duisburg, Germany

Norbert Fuhr
norbert.fuhr@uni-due.de
University of Duisburg-Essen
Duisburg, Germany

Dagmar Kern
dagmar.kern@gesis.org
GESIS – Leibniz Institute for the Social Sciences
Cologne, Germany

ABSTRACT

Online retailers often offer a vast choice of products to their customers to filter and browse through. The order in which the products are listed depends on the ranking algorithm employed in the online shop. State-of-the-art ranking methods are complex and draw on many different information, e.g., user query and intent, product attributes, popularity, recency, reviews, or purchases. However, approaches that incorporate user-generated data such as click-through data, user ratings, or reviews disadvantage new products that have not yet been rated by customers. We therefore propose the *User-Needs-Driven Ranking* (*UNDR*) method that accounts for explicit customer needs by using facet popularity and facet value popularity. As a user-centered approach that does not rely on post-purchase ratings or reviews, our method bypasses the cold-start problem while still reflecting the needs of an average customer. In two preliminary user studies, we compare our ranking method with a *rating-based ranking* baseline. Our findings show that our proposed approach generates a ranking that fits current customer needs significantly better than the baseline. However, a more fine-grained usage-specific ranking did not further improve the ranking.

CCS CONCEPTS

• **Information systems** → **Retrieval models and ranking**; *Personalization*; • **Human-centered computing** → *Empirical studies in HCI*.

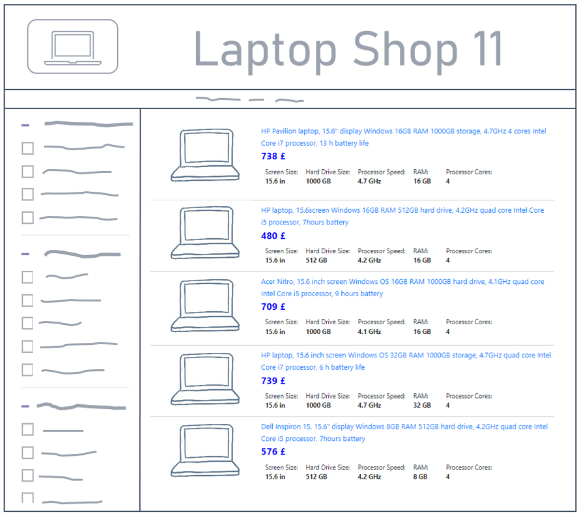
KEYWORDS

E-Commerce, Information Retrieval, Product Search, Ranking, Cold Product

1 INTRODUCTION

Ranking methods for product search are very complex. Aiming to combine business goals with user needs [4, 19], ranking methods usually combine a large number of features, such as product descriptions, user ratings, reviews, and behavioral data like clicks or purchases [3, 9, 14, 18]. From reviews, ranking algorithms can extract for example information about the importance of individual product attributes for users and subsequently account for those user needs in the ranking score [21, 25].

However, for "cold" products [16], i.e., products that are new or have not yet been purchased and evaluated, user-generated data is not available. Those products pose a problem for ranking methods that aim to reflect user needs based on user-generated data. Moreover, user ratings, reviews and click-through data carry implicit information about user needs and need to be carefully cleaned and preprocessed. To overcome the problem of cold products, literature suggested estimating missing data based on reviews from similar products [7, 12] or using alternative information sources such as social media to gather insights about cold products [16, 27].


Laptop Shop 11

HP Pavilion laptop, 15.6" display Windows 16GB RAM 1000GB storage, 4.7GHz 4 cores Intel Core i7 processor, 13 h battery life
738 £

Screen Size:	Hard Drive Size:	Processor Speed:	RAM:	Processor Cores:
15.6 in	1000 GB	4.7 GHz	16 GB	4

HP laptop, 15.6screen Windows 16GB RAM 512GB hard drive, 4.2GHz quad core Intel Core i5 processor, 7hours battery
480 £

Screen Size:	Hard Drive Size:	Processor Speed:	RAM:	Processor Cores:
15.6 in	512 GB	4.2 GHz	16 GB	4

Acer Nitro, 15.6 inch screen Windows OS 16GB RAM 1000GB hard drive, 4.1GHz quad core Intel Core i5 processor, 9 hours battery
709 £

Screen Size:	Hard Drive Size:	Processor Speed:	RAM:	Processor Cores:
15.6 in	1000 GB	4.1 GHz	16 GB	4

HP laptop, 15.6 inch screen Windows OS 32GB RAM 1000GB storage, 4.7GHz quad core Intel Core i7 processor, 6 h battery life
739 £

Screen Size:	Hard Drive Size:	Processor Speed:	RAM:	Processor Cores:
15.6 in	1000 GB	4.7 GHz	32 GB	4

Dell Inspiron 15, 15.6" display Windows 8GB RAM 512GB hard drive, 4.2GHz quad core Intel Core i5 processor, 7hours battery
576 £

Screen Size:	Hard Drive Size:	Processor Speed:	RAM:	Processor Cores:
15.6 in	512 GB	4.2 GHz	8 GB	4

Figure 1: Screenshot of a fictive laptop shop with the *UNDR* method.

To provide an alternative method for gathering information about attribute-level user needs, we propose the *User-Needs-Driven Ranking* (*UNDR*) score that accounts for current user needs by

(1) Gathering explicit, structured information about current user needs for a product category by collecting facet selections of users.
(2) Driven by the user needs, calculating the average popularity of each facet and its values.
(3) Using the popularity weights to assign popularity scores to each product and rank them accordingly.

To evaluate the *UNDR* method, we collected the multi-faceted user needs for a new laptop in a crowdsourcing experiment with N = 277 participants. We calculated the popularity weights for laptop attributes and their values, representing the average needs of all users. Subsequently, we computed the product ranking with the *UNDR* method and compared it to a *rating-based ranking* method in two user studies. In the first user study with N = 59 participants, participants were confronted with screenshots of fictive laptop shops displaying the top 5 ranked laptops of each method (see Figure 1). Participants assessed how well the offered laptops fit their needs as well as which store they would like to visit first. To explore the potential of more distinct user profiles, we conducted a second user study with N = 144 participants. We compared a "basic" user profile, i.e., popularity weights calculated from basic laptop users, and an "advanced" user profile, i.e., popularity weights based on advanced laptop users, with the baseline.

Our findings show that the top 5 laptops of the *UNDR* method are perceived to better fit the user needs of our participants compared to the *rating-based* baseline. Consistently, significantly more participants decided to first visit the *UNDR* shop. However, the second user study showed that offering a more specialized ranking adapted to a user profile (basic, advanced) does not further improve how well the products fit participants' needs.

With our approach, we contribute a ranking method that reflects current user needs and can be deployed by shops without collecting post-purchase generated user data (therefore avoiding the cold-start problem). With the user evaluations, we further contribute empirical evidence in support of the proposed ranking method.

2 RELATED WORK

Online shops facilitate the product discovery and purchase of customers in e-commerce. Compared to information discovery in web search, product search brings up unique challenges, such as integrating user needs with business needs [4, 19], new performance evaluation criteria [23], and data and data annotation challenges [19, 23]. In online product search, search systems are confronted with users' complex, multi-faceted information needs [1, 20, 25]. That is, users have preferences for multiple product attributes, which is often accounted for with providing facets in the search interface [1, 8]. While some product aspects are highly important to customers, others are less relevant [10, 21, 25], e.g., when searching for a laptop, a user might find the price and brand very important but the processor brand less important. Besides product attributes, social attributes like the average product rating influence customer's purchase decision [11, 15].

Information about the importance of product aspects can be used to improve the **ranking of products**, i.e., the order in which the products are sorted and shown to the users. Often, users only consider the first products in the result list [4], making the ranking method one of the most crucial and complex issues in this field [19]. Previous literature suggested either asking users to input and control their individual levels of importance of product aspects [10], or extracting the importance levels from user-generated product content such as user reviews [21, 25]. Taking a binary approach in which users can select important features (rather than weighting them), Sabbah et al. [17] showed that aspect-level sentiment information from reviews can support ranking performance. Other works have mined review texts and user ratings (e.g., the 5-star rating that users assign to a product) to extract information for individual product features to improve the ranking [3, 14]. Besides user-generated reviews and ratings, other user data can be used to improve the ranking: Wu et al. [23] leveraged click-through data and information about purchases to optimize product ranking, and Derakhshan et al. [4] used the "consideration set" of users (i.e., the set of products that users have encountered during their search) in their ranking model. Overall, ranking algorithms are often complex models that combine a number of features both from users and from sellers, such as user queries, click-through data, add-to-carts, revenue information, or order rates [9]. For a more complete overview over ranking algorithms used in e-commerce, the reader is referred to Najib et al. [13] and Santu et al. [9].

Besides comparing several ranking methods, Santu et al. [9] also derived a set of challenges for the field of product search. One is the "presence of uncertain features" which especially applies for new products: "**Cold products**" [16], i.e., new products that were just released in an online shop, initially do not have user reviews, ratings, or click-through data, hence introducing uncertain features. To overcome the cold-start problem, literature proposes several approaches. One line of research estimates missing information based on attributes and information a new product shares with other products in the same or a similar domain (e.g., in [7, 12, 16]). Similarly, Zhao et al. [27] aggregated data from social media and use them as substitutions for missing reviews of cold products in product recommendation, while Pourgholamali [16] collected product data across several additional sources to fill for missing reviews and ratings. Other works use continuous data collection during search, for example, Bi et al. [2] recorded a user's clicking behavior on the first result page for re-ranking later result pages within a single search session.

User behavior data is also used in the related field of recommendation systems to improve **personalized** product recommendations [5, 28] or personalized product rankings [26]. Additionally, structured knowledge about the landscape of products, i.e., hierarchical relations between product categories, can be deployed to improve recommendations [24] and to build user profiles [6].

3 UNDR: USER-NEEDS-DRIVEN RANKING

In this section, we introduce our *User-Needs-Driven Ranking* (*UNDR*) method which assigns user-centered ranking scores to products without the need for user-generated post-purchase data such as ratings or reviews.

While previous ranking and recommendation methods suggest harnessing reviews and ratings [3, 14, 17, 25], we propose gathering information about users' facet selection behavior as indicator for user needs in a product browsing context. This data can either be collected independent of an online shop with controlled user surveys or, if an online shop with facets is already in use, by logging users' facet behavior directly. In contrast to ratings and reviews, a facet-based approach does not require actual purchases and an extra effort of rating and reviewing the product. From the facet selection behavior, we can derive (1) how popular a facet or product attribute is for users, and (2) how popular a specific facet value or attribute value is for users. We assume that:

- **A1** The more often a facet is selected, the more important the attribute is to users.
- **A2** The more often a facet value is selected, the stronger this value represents the current average user need.

For a set of facets F, we can then compute a user-needs-driven ranking score for a product as follows:

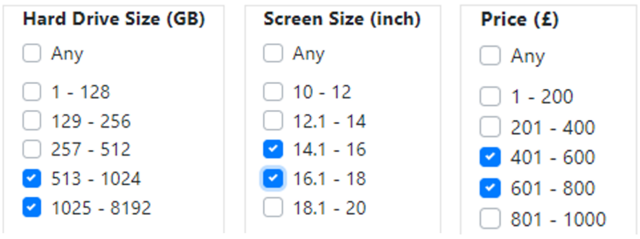
$$\mathrm{UNDR}_{\mathrm{score}} = \sum_{f \in F} w_f \cdot w_{f_v} \tag{1}$$

where f is a facet or product attribute in F, w_f the popularity weight of a facet (computed as the number of users that used f divided by the total number of observed users), and w_{f_v} the popularity weight of a facet value (computed as the number of users that selected value f_v divided by the number of users that used f).

To evaluate our proposed ranking approach, we explore the use case of laptop shops. Laptops are regularly used by many people (simplifying the recruitment of target users), are usually described by multiple attributes (hence representing a multi-faceted, complex need), and are already sold online so that collecting a dataset with user-generated information as a baseline is feasible. In the following sections, we describe how the *UNDR* method can be used in the context of an online laptop shop.

3.1 Collecting User Needs

The *UNDR* method uses information about the popularity of facets and their values. Data of facet selections can be collected either from usage logs of an existing online shop, or by a structured user survey. In our experiment, we set up a crowdsourcing task to collect information about facet usage. We asked workers to first answer questions about their demographic background (age, gender, domain knowledge about laptops) and for which tasks they usually use their laptop (multiple choice of ten common tasks, see Figure 3). Workers then read a scenario ("Imagine your computer broke down. Now, you are planning to buy a new laptop.") and were asked to make a facet selection for ten common laptop attributes (price, RAM size, operating system, brand, hard drive size, screen size, CPU cores, CPU speed, CPU brand, battery life). See Figure 2 for an exemplary facet selection. We asked workers to select the "any" option, if the attribute was not important to them.



Hard Drive Size (GB)

□ Any

□ 1 - 128
□ 129 - 256
□ 257 - 512
☑ 513 - 1024
☑ 1025 - 8192

Screen Size (inch)

□ Any

□ 10 - 12
□ 12.1 - 14
☑ 14.1 - 16
☑ 16.1 - 18
□ 18.1 - 20

Price (£)

□ Any

□ 1 - 200
□ 201 - 400
☑ 401 - 600
☑ 601 - 800
□ 801 - 1000

Figure 2: Example of facet value selections for laptop attributes "Hard Drive Size", "Screen Size", and "Price".

We recruited N = 304 crowd workers on *Prolific*[1], who had to be English native speakers currently residing in the United Kingdom without literacy difficulties. We restricted the participation to a single country to reduce effects of the current economic situation and accompanying variance in the conception of laptop prices. We excluded 14 responses due to low quality and 13 responses because workers did not own a laptop, which could affect their frame of reference. The final sample consisted of 277 workers (160 female, 110 male, 3 non-binary, 4 prefer not to say) with an average age of M = 36.5 years (SD = 11.4 years) and a medium domain knowledge about laptops (on a scale of 1 = "low knowledge about laptops" to 5 = "high knowledge", M = 3.6, SD = 1.0). Figure 3 shows the distribution of domain knowledge levels per laptop usage task. Some purposes, like basic tasks, streaming, and video conferencing are done by workers across all domain knowledge levels, whereas more advanced tasks (e.g., software development, high-level gaming) are mostly done by workers with high knowledge of laptops.

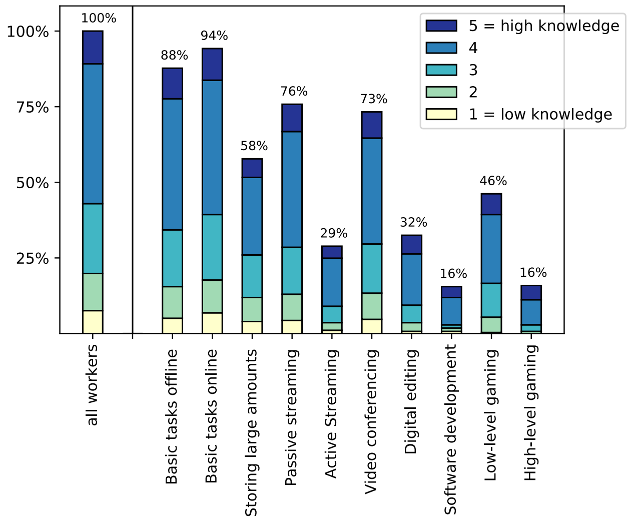


Figure 3: Participants laptop usage habits per task (multiple choice) in percent, divided into self-reported levels of domain knowledge about laptops.

3.2 UNDR Popularity Weights

Based on the responses, we calculated the overall popularity weight of each attribute as the percentage of users who selected a specific value, i.e., percentage of users who did not select the "any" option. Table 1 lists the attribute-level popularity weights to give an understanding of the relative popularity of the ten laptop attributes. For example, for the attribute "Screen Size", 41 users (15%) selected the "any" option. The overall popularity weight of "Screen Size" is therefore $w_f = 0.85$. Furthermore, out of the $277 - 41 = 236$ users who selected specific screen size values, 153 users (40%) selected the value "14.1 - 16" inches. Consequently, the value-level popularity weight of "14.1 - 16" in "Screen Size" is $w_{f_v} = 0.40$. A laptop with a screen size of 14.9 would therefore receive a score of

[1] https://www.prolific.co

$0.85 \cdot 0.40 = \mathbf{0.34}$ for the screen size attribute. Screens smaller than 12 inches are less popular and received a value-level popularity of $w_{f_v} = 0.03$ in our experiment. A laptop with an 11 inches screen would therefore be assigned a score of $0.85 \cdot 0.03 = \mathbf{0.026}$ for the screen size attribute.

attribute	"any" count out of 277	overall popularity weight
Price	24	0.91
Brand	77	0.72
Operating system	35	0.87
Screen size	41	0.85
Hard drive size	55	0.80
RAM size	33	0.88
CPU cores	126	0.55
CPU speed	90	0.68
CPU brand	141	0.49
Battery life	38	0.86

Table 1: Overall popularity weights per laptop attribute.

3.3 Product Dataset Collection

We collected a dataset of 1,445 laptops from *Amazon*[2] in December 2021, containing product information such as the technical descriptions, prices, and user-generated data such as ratings and reviews. Prices were collected in US dollars and converted to pound sterling using the exchange rate of those days of 0.75. We discarded data points that did not carry information about all ten common laptop attributes (price, RAM size, operating system, brand, hard drive size, screen size, CPU cores, CPU speed, CPU brand, battery life) and duplicates. Furthermore, to allow for comparison with a rating-based baseline, we only kept data points with at least ten customer ratings and calculated for each laptop the average rating score. After the reduction, the final dataset contains 182 laptops. Using the popularity weights for attributes and their values (see Section 3.2), we then calculated the *UNDR* score for each laptop and ranked the laptops, starting with the laptop with the highest *UNDR* score at the first rank.

4 USER STUDY 1: UNDR VS. BASELINE

The *UNDR* method aims to be a user-centered ranking approach without needing user-generated post-purchase data such as ratings or reviews. To evaluate whether our method can substitute or even improve on common product popularity signals such as star ratings, we designed a user study answering the following research question:

RQ1 How well does the *UNDR* method perform compared to a *rating-based ranking* method in the eyes of the users?

In this preliminary study, we chose the "sort-by-average customer rating"-function as our baseline because this function is well-known to users and commonly offered by most online-shops as a sorting feature. We compared the two rankings (*UNDR*, *rating-based ranking*) using a within-subject design. Here, we focused on the "first impression" of rankings and investigated whether users are more drawn towards a shop using the *UNDR* method for ranking as opposed to a shop using a *rating-based ranking.* Before investing effort and time into the development of a fully interactive prototype, we decided to start our research with a simplistic, preliminary user study. We used static screenshots (showing the top 5 results of each ranking method) to collect initial insights into the potential of the *UNDR* score.

4.1 Task and Procedure

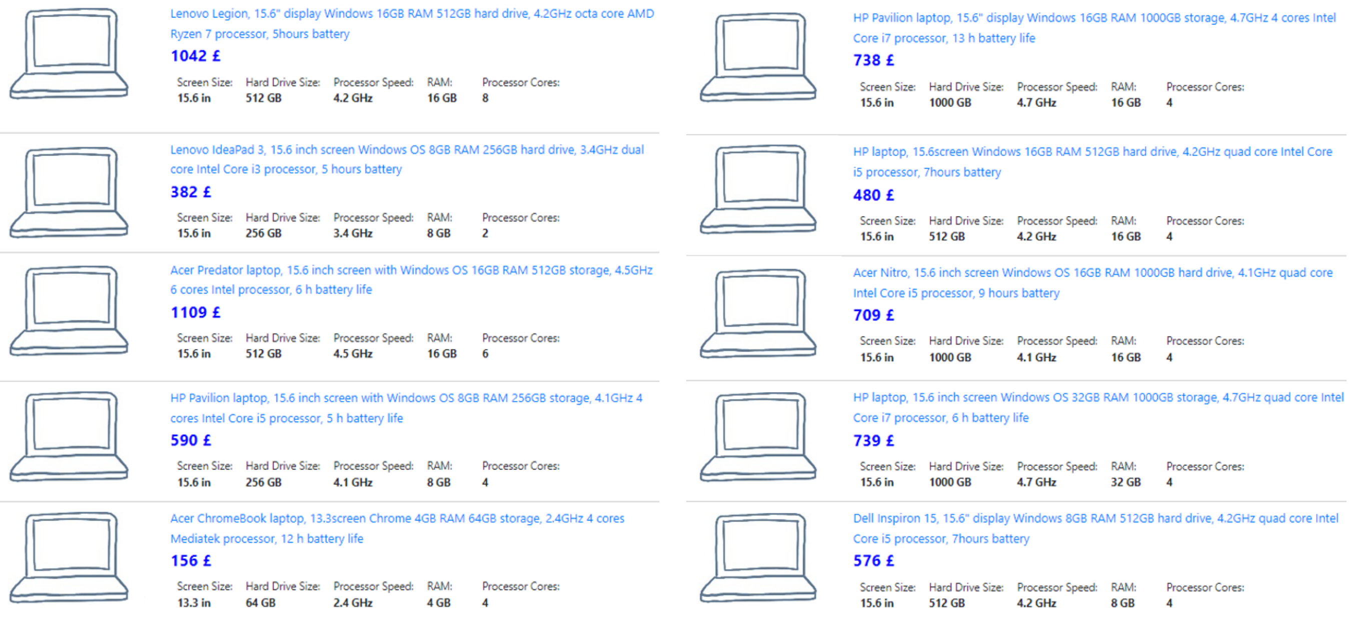
We asked participants of our user study to first give informed consent and answer some demographic questions (age, gender, domain knowledge about laptops). Subsequently, they read the same scenario description as the crowdworkers (see Section 3.1) that prompted them to imagine needing a new laptop that fits their current needs. Furthermore, the scenario told them that they find two online shops on the internet. Participants then saw two static screenshots: One of the *rating-based* shop and one of the *UNDR* shop. The exact result lists presented to the participants are shown in Figure 4a (baseline) and Figure 4b (*UNDR*). We constructed the screenshots such that they resemble the common online shop structure with shop branding at the top, a facet column on the left, and a result list in the center (see Figure 1). We replaced the facet column with scribbles and product images with a generic image to reduce the effect of visual confounders [22]. All laptops in the result list display the same information and have a standardized title (model, brand, screen size, operating system, RAM size, storage size, processor information, battery life). To simulate shops with "cold" products, i.e., products without ratings or reviews, we did not display the user ratings – the ratings were only used for determining the ranking of the *rating-based* shop. The order in which the screenshots were presented was randomised. For each shop, participants were asked to indicate how well the shop fits their needs and how likely it is that they would visit this store. Finally, participants selected which shop they would visit first and described the reasons for their decision in an open text box.

4.2 Measures and Analysis

To investigate how the two ranking methods perform in the eyes of the users, we analyzed three performance indicators:

(1) **Fitness**: Rating of agreement with the statement "The laptops of shop X fit my needs." on a scale of 1 ("strongly disagree") to 5 ("strongly agree"). We tested for significant differences using the Wilcoxon signed-rank test for paired samples at $\alpha = 0.05$.
(2) **Visit likelihood**: Rating of agreement with the statement "I would like to visit the website of shop X." on a scale of 1 ("strongly disagree") to 5 ("strongly agree"). We tested for significance with the Wilcoxon signed-rank test at $\alpha = 0.05$.
(3) **Shop selection**: Number of times a shop was selected in the question "Which of the two laptop shops would you visit first?". We determined significant differences using the cumulative distribution function of the binomial distribution, testing whether the observed distribution is different from a 50%-50% distribution at $\alpha = 0.05$.

[2] https://www.amazon.com



Lenovo Legion, 15.6" display Windows 16GB RAM 512GB hard drive, 4.2GHz octa core AMD Ryzen 7 processor, 5hours battery
1042 £
Screen Size: 15.6 in | Hard Drive Size: 512 GB | Processor Speed: 4.2 GHz | RAM: 16 GB | Processor Cores: 8

Lenovo IdeaPad 3, 15.6 inch screen Windows OS 8GB RAM 256GB hard drive, 3.4GHz dual core Intel Core i3 processor, 5 hours battery
382 £
Screen Size: 15.6 in | Hard Drive Size: 256 GB | Processor Speed: 3.4 GHz | RAM: 8 GB | Processor Cores: 2

Acer Predator laptop, 15.6 inch screen with Windows OS 16GB RAM 512GB storage, 4.5GHz 6 cores Intel processor, 6 h battery life
1109 £
Screen Size: 15.6 in | Hard Drive Size: 512 GB | Processor Speed: 4.5 GHz | RAM: 16 GB | Processor Cores: 6

HP Pavilion laptop, 15.6 inch screen with Windows OS 8GB RAM 256GB storage, 4.1GHz 4 cores Intel Core i5 processor, 5 h battery life
590 £
Screen Size: 15.6 in | Hard Drive Size: 256 GB | Processor Speed: 4.1 GHz | RAM: 8 GB | Processor Cores: 4

Acer ChromeBook laptop, 13.3screen Chrome 4GB RAM 64GB storage, 2.4GHz 4 cores Mediatek processor, 12 h battery life
156 £
Screen Size: 13.3 in | Hard Drive Size: 64 GB | Processor Speed: 2.4 GHz | RAM: 4 GB | Processor Cores: 4

(a) Laptop list with the *rating-based ranking* method.

HP Pavilion laptop, 15.6" display Windows 16GB RAM 1000GB storage, 4.7GHz 4 cores Intel Core i7 processor, 13 h battery life
738 £
Screen Size: 15.6 in | Hard Drive Size: 1000 GB | Processor Speed: 4.7 GHz | RAM: 16 GB | Processor Cores: 4

HP laptop, 15.6screen Windows 16GB RAM 512GB hard drive, 4.2GHz quad core Intel Core i5 processor, 7hours battery
480 £
Screen Size: 15.6 in | Hard Drive Size: 512 GB | Processor Speed: 4.2 GHz | RAM: 16 GB | Processor Cores: 4

Acer Nitro, 15.6 inch screen Windows OS 16GB RAM 1000GB hard drive, 4.1GHz quad core Intel Core i5 processor, 9 hours battery
709 £
Screen Size: 15.6 in | Hard Drive Size: 1000 GB | Processor Speed: 4.1 GHz | RAM: 16 GB | Processor Cores: 4

HP laptop, 15.6 inch screen Windows OS 32GB RAM 1000GB storage, 4.7GHz quad core Intel Core i7 processor, 6 h battery life
739 £
Screen Size: 15.6 in | Hard Drive Size: 1000 GB | Processor Speed: 4.7 GHz | RAM: 32 GB | Processor Cores: 4

Dell Inspiron 15, 15.6" display Windows 8GB RAM 512GB hard drive, 4.2GHz quad core Intel Core i5 processor, 7hours battery
576 £
Screen Size: 15.6 in | Hard Drive Size: 512 GB | Processor Speed: 4.2 GHz | RAM: 8 GB | Processor Cores: 4

(b) Laptop list with the *UNDR* method.

Figure 4: Screenshot of the top 5 results of both fictive laptop shops.

Furthermore, we added a qualitative measure to gain additional insights into users' decision-making processes:

(4) **Shop selection reason**: Open text answer to the question "Why do you think this shop best fits your needs?". We analyzed the data with a qualitative coding and clustering approach with one annotator.

4.3 Participants

Similar to the recruitment in Section 3.1, we invited N = 60 participants on *Prolific* to take part in our user study. We applied the same prescreening aspects as before (English native speakers, UK residents, no literacy difficulties) to reduce effects of cultural or economic confounders. Within the N = 59 valid responses (one discarded due to low quality), age distribution (M = 36.6 years, SD = 13.4 years) was similar to the crowdsourcing experiment and the gender distribution was balanced (29 female, 30 male, 0 non-binary, 0 prefer not to say). Participants had, again, a medium average domain knowledge about laptops (on a scale of 1 to 5, M = 3.2, SD = 1.1).

4.4 Results

To investigate how well the *UNDR* method performs compared to a *rating-based ranking* method in the eyes of the users (**RQ1**), we first looked at how well both ranking methods fit users' needs. Figure 5 depicts the distribution of agreement levels with the statement "The laptops of shop X fit my needs" in the baseline shop (M = 3.5, SD = 1.0) and *UNDR* shop (M = 4.0, SD = 0.8). The difference in means is significant (U = 172, p = .003), showing that participants found the offer of the *UNDR* shop to better fit their needs than the laptops of the *rating-based* baseline.

We further looked at the visit likelihood as a second measure of how participants perceive and assess the rankings. On average, participants reported a significantly higher visit likelihood (U = 94, p < .001) for the *UNDR* shop (M = 4.1, SD = 0.7) than for the baseline shop (M = 3.4, SD = 0.9). Figure 6 visualizes participants' responses in the visit likelihood measure.

At the end of the experiment, participants had to make a decision about which of the two stores to visit first. 39 participants (66%) decided for the *UNDR* shop, whereas 20 participants (34%) selected

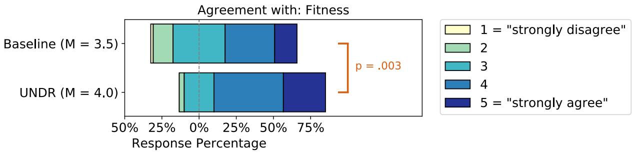


Figure 5: Likert plot of fitness measure for both ranking interfaces and Wilcoxon's test result, p-value corrected with Bonferroni method.

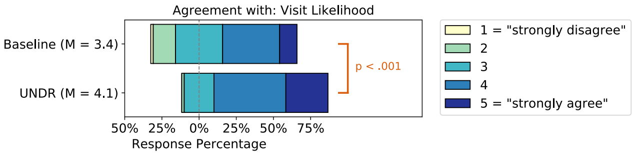


Figure 6: Likert plot of visit likelihood measure for both ranking interfaces and Wilcoxon's test result, p-value corrected with Bonferroni method.

the baseline shop. This distribution differs significantly (p = .009) from an expected 50%-50% distribution of both shops would draw an equal number of visitors. Participants also explained their decision in an open answer. 37 participants mentioned the price of the selected shop to either better fit their needs or to be closer to the price range they would expect for a laptop at the time. Most of the times, participants set the observed prices in relation to the hardware specifications: *"They seem better value for money"* or *"The pricing of shop 4 seems to be higher for the spec"*. Furthermore, participants often mention specific laptop attributes that seem to be important for them: hard drive size (15), RAM size (5), CPU (4), brand (4), battery life (3), and screen size (1). Similarly, others do not describe specific attributes, but note that the specifications fit their specific usage goal such as *"help with daily tasks"*, *"for personal use"*, while 15 say it fits their needs in general: *"Better selection for my needs"* or *"match my requirements"*. Although participants report foremost price and fitness for their (individual) needs as the deciding factors, more subtle factors seem to effect the decision as well. While 6 participants liked the broader price range in the baseline shop (*"Broad range of price points"*), others take the narrow price range in the *UNDR* shop as a sign of a well-curated laptop offer: *"the prices and descriptions are similar which suggest they are of a certain standard"* and *" It looks like they are more evenly priced, which means they are likely acceptable quality"*.

5 USER STUDY 2: PROFILING

The *UNDR* method can potentially provide a ranking tailored to specific user groups with different usage habits, depending on whose data the popularity weights are based on. In the first user study (Section 4), we used a ranking based on the data of all crowd workers. However, while some crowd workers used their laptops only for "basic" tasks, others, more knowledgeable crowd workers, also had "advanced" usage habits (see Figure 3). We therefore set up a second user study to answer the following research questions:

RQ2 Do user-group-specific (basic, advanced) *UNDR* result lists better fit the users' needs than the general *UNDR* result list or the *rating-based* baseline?

RQ3 To what extent does the classification into an incorrect user profile harm the user's view of the ranking performance?

Using the data from the crowdsourcing experiment, we calculated a *basic UNDR* laptop order based on the data of the 83 workers with only basic usage habits (no digital editing, no software engineering, no high-level gaming). Equally, we computed an *advanced UNDR* order on the data of the 194 workers with advanced usage habits (at least one of: digital editing, software engineering, high-level gaming). We used a study design similar to the setup in the first user study: We compared *UNDR* shops with the *rating-based* baseline shop in a within-subject experiment. Moreover, we made sure to gather at least 30 valid responses for any of the following groups: The correct profile classifications (1) basic user - *basic UNDR* shop and (2) advanced user - *advanced UNDR*, and the incorrect profile classifications (3) basic user - *advanced UNDR* shop and (4) advanced user - *basic UNDR*.

5.1 Task and Procedure

For comparability, we kept the task description, procedure and questions of the first user study and only exchanged the screenshots of the *UNDR* shops.

5.2 Measures and Analysis

In this experiment, we focused only on the quantitative measures of **fitness** to needs as described in Section 4.2. For comparison with the baseline (within-subject comparison), we again used the Wilcoxon signed-rank test for paired samples. To compare the fitness of two *UNDR* shops (e.g., general *UNDR* from the first user study vs *basic UNDR* or *advanced UNDR*), we used the Mann-Whitney U test. All significance tests are evaluated at a significance level of $\alpha = 0.05$.

5.3 Participants

We again used the recruitment procedure and prescreening factors described in Section 4.3 for this second user study. We stopped recruiting once we had at least 30 participants for each user - profile combination, leading to N = 144 valid responses in total (n = 36 in group (1), n = 38 in group (2), n = 33 in group (3), n = 37 in group (4)). The sample group was approximately gender-balanced (76 female, 63 male, 3 non-binary, 2 prefer not to say) and on average slightly older (M = 40.5 years, SD = 14.5 years) than the participants of the first user study (but not significantly, Mann-Whitney U = 3754, p = .070). The average domain knowledge about laptops was M = 3.2 (SD = 1.1).

5.4 Results

We first analyzed how the user-group-specific *UNDR* shops are perceived by the respective user groups (**RQ2**). Figure 7 gives an overview of how well the shops fitted users' needs in all four groups. We did not find a significant difference between the *basic UNDR* shop and the baseline for basic users in group (1) at $\alpha = 0.05$ (U = 66, p = .088). For advanced users, the *advanced UNDR* shop was a better fit than the baseline shop (U = 62, p = .005). However, comparing how well the profiled *UNDR* shops matched the needs of their users with how well the general *UNDR* shop from the first study, we did not find a significant difference (U = 2183, p = .431). Therefore, we cannot conclude that the profiling brings an advantage over the general *UNDR* result list.

Besides the general potential for improvement with usage profiles, we investigated whether an incorrect profiling can harm how well the results fit users' needs (**RQ3**). That is, what happens if a basic user is mistaken as advanced user and subsequently sees the *advanced UNDR* shop? On the one hand, for the basic user group (3), the *advanced UNDR* shop is perceived to fit significantly less to the user needs than the *basic UNDR* shop (U = 409, p = .010). That is, mistaking a basic user for an advanced user and subsequently showing the wrong *UNDR* ranking can potentially harm the user experience. On the other hand, the advanced user group (4) did not find the offer of the *basic UNDR* shop to be a worse fit for their needs (U =596, p = .114).

6 DISCUSSION

In this paper, we introduce the *User-Needs-Driven Ranking* (*UNDR*) and take a first step to evaluate its potential as a user-centered

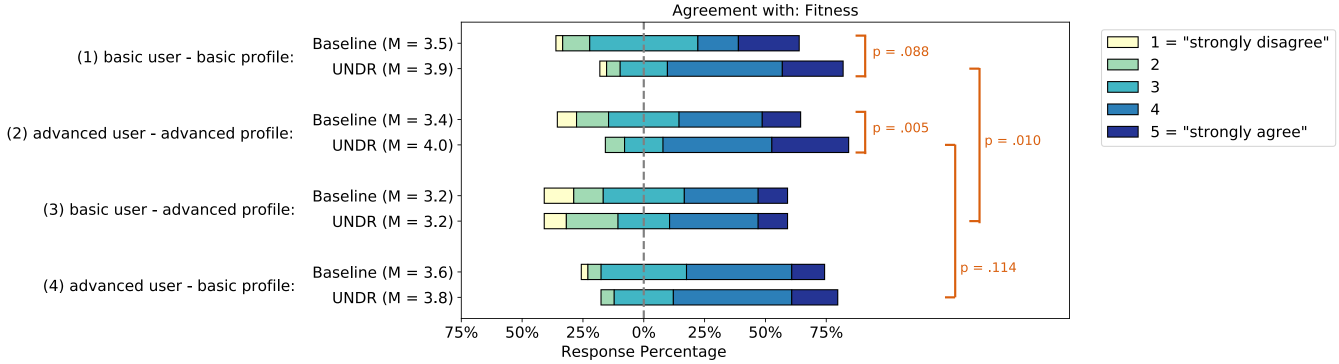


Figure 7: Likert plot of fitness measure in the second user study, grouped by actual usage habit (user group) and profile shown (basic, advanced). Each group assessed both the shop using *UNDR* and the shop using the *rating-based* baseline.

ranking in a product browsing context. Our goal was to find a ranking that accounts for current user needs without requiring user-generated post-purchase data like ratings or reviews. In a first user study in the laptop domain, we explored how the *UNDR* method compares to a *rating-based* baseline at a "first impression" stage, similar to a situation in which a user enters an online shop and is confronted with an initial ranking of the products. Our findings show that the *UNDR* method brought about products that are more fitting to users' needs than the baseline. Participants also reported to more likely visit the *UNDR* shop than the baseline shop, and more often chose to visit the *UNDR* shop first, indicating that this shop provides a higher chance of success in their eyes. Although the first user study was only a starting point, we derive from it that the *UNDR* method delivers acceptable initial results and is worth exploring further in follow-up experiments. The *UNDR* method could be especially valuable for online shops that have little user-generated data, e.g., shops that just went online, shops with fewer visitors, or shops that have less customers in some product categories than in others. It could also provide an advantage in product domains with fast-changing user needs (e.g., bikes, for which new features such as e-mobility were added over time) because user needs can be collected in a fast and structured way via surveys.

In the second user study, we investigated the potential of a more fine-grained, user-group-specific ranking. We could not find an indication that a basic laptop ranking and an advanced laptop ranking with the *UNDR* method provides added value over a general user profile. For participants with basic usage habits, showing the wrong profile could even reduce the positive effects of the *UNDR* method. In our experiments, we had a well-controlled task with well-controlled recruiting procedure, which might have led to participants with similar needs. It is also possible that the classification into "basic" and "advanced" is not a suitable grouping factor; however, our findings also show that the profiling was not worse than the baseline. In other settings, needs might be more varied: Distinguishing for example between "vegetarians" and "meat eaters" in recipe search or between "electric vehicle" and "car with combustion engine" in car search might have a stronger effect.

Considering both user studies, we conclude: (1) Our findings indicate that the *UNDR* method is a user-centered ranking that outperforms a *rating-based ranking* baseline in a "first impression" user study. (2) We do not find an indication that further profiling into usage-based profiles (basic, advanced) provides an additional improvement of how well the product offer fits the users' needs.

Since we present insights from preliminary studies, our findings are subject to several limitations. First, we used screenshots of online shops in our experiments, which eliminates the possibility for interactivity. We do not know whether the improved performance persists when adding facets and interactive elements such as textual search. However, a productive prototype of an online shop and bigger datasets (to avoid empty result lists) would be needed for an interactive experiment. In future work, we will develop an interactive prototype and evaluate ranking performance (e.g., using our fitness measure) at the end of a full search session. However, the insights from our user study show that shops could improve their "first impression" by using the *UNDR* method. Moreover, the *UNDR* method promotes products that conform to the average user need of a specific user group. Defining the boundaries of that target group, i.e., inclusion and exclusion criteria, clustering of similar users, is still an open question. Online shops that want to deploy the *UNDR* method should make an extensive target group analysis to avoid rankings that fit the data but not the end-user. Finally, in this preliminary evaluation, we have treated the *UNDR* score as an isolated metric. State-of-the-art ranking algorithms, however, are often more complex and consider multiple information sources. In a next step, the potential of integrating the *UNDR* score with existing methods should be explored, e.g., in learning-to-rank approaches, either as an additional feature or as a proxy for other popularity signals such as customer ratings for cold products. Despite those limitations, our initial experiments showcase the applicability of the *UNDR* method to the laptop domain and promising performance in preliminary user experiments.

The *UNDR* method is not only applicable to the laptop domain. In theory, it can be used in other product domains in which users have complex, multi-faceted information needs. The *UNDR* score represents how well a product matches user needs with respect to a set of attributes without making assumptions about the form (can be applied to both categorical or numerical facets) or content of an attribute. It is therefore highly flexible and adaptable. Future studies should investigate the performance of the *UNDR* score in various product domains such as technical products, clothes or furniture – domains that vary in number of attributes and number of values per attribute. However, to give a useful score, the *UNDR* method should only be applied in product domains where some attribute values are clearly more popular than others. In the laptop domain, for example, a screen smaller than 12 inches has a low popularity (.03) while 14 - 16 inches is much more popular (.40), which promotes 14 inches laptops and demotes smaller laptops. If all screen sizes were equally popular, the *UNDR* score would not provide distinguishing information.

Moreover, while we collected users' current facet selection behavior in a crowdsourcing experiment, it would be possible to collect such data from logs of productive systems. Future studies should explore how to collect and clean facet selection data to be used for calculating the *UNDR* popularity weights.

7 CONCLUSION

This paper introduced the *User-Needs-Driven Ranking* (*UNDR*) score, an approach that utilizes facet selection behavior to bypass the cold-start problem while still delivering a user-centered ranking. We presented two preliminary user studies to investigate the potential applicability of the *UNDR* method. Comparing our proposed method with a *rating-based ranking* showed that the *UNDR* method better addresses current user needs. However, further profiling the ranking method into usage-group specific rankings did not bring added value. If the *UNDR* method persists to bring similar or even better results than methods based on user-generated data (e.g., click-through data or reviews) in future user studies, facet selection data could be a valuable addition to state-of-the-art ranking algorithms. Especially online shops with little user data could then profit from including the *UNDR* score in their ranking to provide a user-centered ranking to their customers.

ACKNOWLEDGMENTS

This work was partly funded by the DFG, grant no. 388815326; the VACOS project at GESIS and the University of Duisburg-Essen.

REFERENCES

[1] Ori Ben-Yitzhak, Nadav Golbandi, Nadav Har'El, Ronny Lempel, Andreas Neumann, Shila Ofek-Koifman, Dafna Sheinwald, Eugene Shekita, Benjamin Sznajder, and Sivan Yogev. 2008. Beyond Basic Faceted Search. In *Proceedings of the 2008 International Conference on Web Search and Data Mining* (Palo Alto, California, USA) *(WSDM '08)*. Association for Computing Machinery, New York, NY, USA, 33–44. https://doi.org/10.1145/1341531.1341539

[2] Keping Bi, Choon Hui Teo, Yesh Dattatreya, Vijai Mohan, and W Bruce Croft. 2019. Leverage implicit feedback for context-aware product search.

[3] Samira Chaabna, Wang Hu, and Mohammed Lutf. 2015. Designing a ranking system for product search engine based on mining UGC. *J. Manag. Inf. Syst. E-Commerce* 2, 1 (2015), 23–65.

[4] Mahsa Derakhshan, Negin Golrezaei, Vahideh Manshadi, and Vahab Mirrokni. 2020. Product Ranking on Online Platforms. In *Proceedings of the 21st ACM Conference on Economics and Computation* (Virtual Event, Hungary) *(EC '20)*. Association for Computing Machinery, New York, NY, USA, 459. https://doi.org/10.1145/3391403.3399483

[5] Jingtao Ding, Guanghui Yu, Yong Li, Xiangnan He, and Depeng Jin. 2020. Improving Implicit Recommender Systems with Auxiliary Data. *ACM Trans. Inf. Syst.* 38, 1, Article 11 (feb 2020), 27 pages. https://doi.org/10.1145/3372338

[6] Yulong Gu, Zhuoye Ding, Shuaiqiang Wang, and Dawei Yin. 2020. *Hierarchical User Profiling for E-Commerce Recommender Systems.* Association for Computing Machinery, New York, NY, USA, 223–231. https://doi.org/10.1145/3336191.3371827

[7] Parth Gupta, Tommaso Dreossi, Jan Bakus, Yu-Hsiang Lin, and Vamsi Salaka. 2020. *Treating Cold Start in Product Search by Priors.* Association for Computing Machinery, New York, NY, USA, 77–78. https://doi.org/10.1145/3366424.3382705

[8] Marti Hearst, Ame Elliott, Jennifer English, Rashmi Sinha, Kirsten Swearingen, and Ka-Ping Yee. 2002. Finding the Flow in Web Site Search. *Commun. ACM* 45, 9 (sep 2002), 42–49. https://doi.org/10.1145/567498.567525

[9] Shubhra Kanti Karmaker Santu, Parikshit Sondhi, and ChengXiang Zhai. 2017. On Application of Learning to Rank for E-Commerce Search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Shinjuku, Tokyo, Japan) *(SIGIR '17)*. Association for Computing Machinery, New York, NY, USA, 475–484. https://doi.org/10.1145/3077136.3080838

[10] Dagmar Kern, Wilko van Hoek, and Daniel Hienert. 2018. Evaluation of a Search Interface for Preference-Based Ranking: Measuring User Satisfaction and System Performance. In *Proceedings of the 10th Nordic Conference on Human-Computer Interaction* (Oslo, Norway) *(NordiCHI '18)*. Association for Computing Machinery, New York, NY, USA, 184–194. https://doi.org/10.1145/3240167.3240170

[11] Boying Li, Eugene Ch'ng, Alain Yee-Loong Chong, and Haijun Bao. 2016. Predicting online e-marketplace sales performances: A big data approach. *Computers & Industrial Engineering* 101 (2016), 565–571. https://doi.org/10.1016/j.cie.2016.08.009

[12] Paul Missault, Arnaud de Myttenaere, Andreas Radler, and Pierre-Antoine Sondag. 2021. Addressing Cold Start With Dataset Transfer In E-Commerce Learning To Rank. In *Proceedings of Knowledge Management in e-Commerce* (Ljubljana, Slovvenia) *(KMM-ecomm '21)*. Association for Computing Machinery, New York, NY, USA, 1–5.

[13] Achmad Choirun Najib and Nur Aini Rakhmawati. 2019. A Systematic Literature Review on the Product Ranking Methods. *Khazanah Informatika: Jurnal Ilmu Komputer dan Informatika* 5, 1 (2019), 88–98.

[14] Erfan Najmi, Khayyam Hashmi, Zaki Malik, Abdelmounaam Rezgui, and Habib Ullah Khan. 2015. CAPRA: A Comprehensive Approach to Product Ranking Using Customer Reviews. *Computing* 97, 8 (aug 2015), 843–867. https://doi.org/10.1007/s00607-015-0439-8

[15] Robin S. Poston and Cheri Speier. 2005. Effective Use of Knowledge Management Systems: A Process Model of Content Ratings and Credibility Indicators. *MIS Quarterly* 29, 2 (2005), 221–244. http://www.jstor.org/stable/25148678

[16] Fatemeh Pourgholamali. 2016. Mining Information for the Cold-Item Problem. In *Proceedings of the 10th ACM Conference on Recommender Systems* (Boston, Massachusetts, USA) *(RecSys '16)*. Association for Computing Machinery, New York, NY, USA, 451–454. https://doi.org/10.1145/2959100.2959102

[17] Firas Sabbah and Norbert Fuhr. 2021. A Transparent Logical Framework for Aspect-Oriented Product Ranking Based on User Reviews. In *Advances in Information Retrieval*, Djoerd Hiemstra, Marie-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani (Eds.). Springer International Publishing, Cham, 558–571.

[18] Daria Sorokina and Erick Cantu-Paz. 2016. Amazon Search: The Joy of Ranking Products. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Pisa, Italy) *(SIGIR '16)*. Association for Computing Machinery, New York, NY, USA, 459–460. https://doi.org/10.1145/2911451.2926725

[19] Manos Tsagkias, Tracy Holloway King, Surya Kallumadi, Vanessa Murdock, and Maarten de Rijke. 2021. Challenges and Research Opportunities in ECommerce Search and Recommendations. *SIGIR Forum* 54, 1, Article 2 (feb 2021), 23 pages. https://doi.org/10.1145/3451964.3451966

[20] Damir Vandic, Flavius Frasincar, and Uzay Kaymak. 2013. Facet Selection Algorithms for Web Product Search. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management* (San Francisco, California, USA) *(CIKM '13)*. Association for Computing Machinery, New York, NY, USA, 2327–2332. https://doi.org/10.1145/2505515.2505664

[21] Martin Voigt, Artur Werstler, Jan Polowinski, and Klaus Meißner. 2012. Weighted Faceted Browsing for Characteristics-Based Visualization Selection through End Users. In *Proceedings of the 4th ACM SIGCHI Symposium on Engineering Interactive Computing Systems* (Copenhagen, Denmark) *(EICS '12)*. Association for Computing Machinery, New York, NY, USA, 151–156. https://doi.org/10.1145/2305484.2305509

[22] Mengyue Wang, Xin Li, and Patrick Y.K. Chau. 2016. The Impact of Photo Aesthetics on Online Consumer Shopping Behavior: An Image Processing-Enabled Empirical Study. In *37th International Conference on Information Systems (ICIS 2016) Proceedings.* Association for Information Systems, Atlanta, USA, 1005–1016.

[23] Liang Wu, Diane Hu, Liangjie Hong, and Huan Liu. 2018. Turning Clicks into Purchases: Revenue Optimization for Product Search in E-Commerce. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (Ann Arbor, MI, USA) *(SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 365–374. https://doi.org/10.1145/3209978.3209993

[24] Xuzheng Yu, Tian Gan, Yinwei Wei, Zhiyong Cheng, and Liqiang Nie. 2020. *Personalized Item Recommendation for Second-Hand Trading Platform.* Association for Computing Machinery, New York, NY, USA, 3478–3486. https://doi.org/10.1145/3394171.3413640

[25] Zheng-Jun Zha, Jianxing Yu, Jinhui Tang, Meng Wang, and Tat-Seng Chua. 2014. Product Aspect Ranking and Its Applications. *IEEE Transactions on Knowledge and Data Engineering* 26, 5 (2014), 1211–1224. https://doi.org/10.1109/TKDE.2013.136

[26] Yan Zhang, Defu Lian, and Guowu Yang. 2017. Discrete Personalized Ranking for Fast Collaborative Filtering from Implicit Feedback. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* (San Francisco, California, USA) *(AAAI '17)*. AAAI Press, Palo Alto, California, 1669–1675.

[27] Wayne Xin Zhao, Sui Li, Yulan He, Edward Y. Chang, Ji-Rong Wen, and Xiaoming Li. 2016. Connecting Social Media to E-Commerce: Cold-Start Product Recommendation Using Microblogging Information. *IEEE Transactions on Knowledge and Data Engineering* 28, 5 (2016), 1147–1159. https://doi.org/10.1109/TKDE.2015.2508816

[28] Meizi Zhou, Zhuoye Ding, Jiliang Tang, and Dawei Yin. 2018. Micro Behaviors: A New Perspective in E-Commerce Recommender Systems. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (Marina Del Rey, CA, USA) *(WSDM '18)*. Association for Computing Machinery, New York, NY, USA, 727–735. https://doi.org/10.1145/3159652.3159671