Multi-Perspective Learning to Rank to Support Children's Information Seeking in the Classroom

Garrett Allen
Delft University of Technology
Delft, The Netherlands
G.M.Allen@tudelft.nl

Katherine Landau Wright
Boise State University
Boise, Idaho, United States
katherinewright@boisestate.edu

Jerry Alan Fails
Boise State University
Boise, Idaho, United States
jerryfails@boisestate.edu

Casey Kennington
Boise State University
Boise, Idaho, United States
caseykennington@boisestate.edu

Maria Soledad Pera
Delft University of Technology
Delft, The Netherlands
M.S.Pera@tudelft.nl

***Abstract*—We introduce a re-ranking model that augments the functionality of standard search engines to aid *classroom* search activities for *children* (ages 6–11). This model extends the known listwise learning-to-rank framework by balancing risk and reward. Doing so enables the model to prioritize Web resources of high educational alignment, appropriateness, and adequate readability by analyzing the URLs, snippets, and page titles of Web resources retrieved by a mainstream search engine. Experimental results demonstrate the value of considering multiple perspectives inherent to the classroom when designing algorithms that can better support children's information discovery.**

***Index Terms*—children's web search, ranking**

I. Introduction

Children in elementary classrooms (Kindergarten–5^{th} grade, typically 6–11 years old) use search engines (**SE**) to find Web resources needed to complete their school assignments [1], [2]. SE built specifically for children's use in a classroom environment, such as EdSearch, require regular maintenance, e.g., curating resources (text or media) manually to identify educational value or offering resources from allow-listed sites using Google's Custom Search (**GCS**) platform which utilizes the *SafeSearch* feature to filter out pornographic resources. Maintaining an up-to-date allow-list becomes burdensome as the Web proliferates. Moreover, children's SE based on GCS are known to return less relevant results 30% of the time, trading relevance for safer results [3]. Further, specialized SE must overcome the barrier of adoption: children prefer to use the popular mainstream options, e.g., Google, which are known to dominate the market [4], [5].

Mainstream SE are designed and optimized for adults and can overlook unique factors that impact children's use [6]–[8]. This causes barriers to children identifying relevant resources among those presented on a search engine result page (**SERP**) generated in response to their inquiries [9], [10]. Children struggle to recognize what and how much information is available online, seldom looking past the first six SERP resources [11]. Children have trouble understanding the content of retrieved resources due to the complexity of their texts, which leads to uncertainty when selecting relevant resources [12]. When turning to mainstream SE, children may inadvertently be exposed to inappropriate resources, even when using mainstream functionalities (like Google's *SafeSearch*) as these primarily filter for pornography [13] and do not account for other potentially harmful content, e.g., violence. Safe search functionality also suffers from over-filtering by preventing resources from being returned if they contain terms that might be mistaken as inappropriate [14].

To better enable children's access to online information via SE, we focus on tailoring SERP for specific *audiences* and *contexts*. To scope our work, we turn to the framework introduced in Landoni et al. [15] that allows for the comprehensive design and assessment of search systems for children through four pillars. In our case: children aged 6–11 in grades Kindergarten–5^{th} (**K–5**) as the *user group*, classrooms as the *environment*, information discovery as the *task*, and re-ranking of resources to fit audience and context as the *strategy*.

We pose the following research question (**RQ**): *Does adapting a learning-to-rank model to account for multiple traits lead to prioritizing resources relevant to children and the classroom setting?* We posit that a learning to rank (**LTR**) strategy can be augmented to simultaneously consider multiple traits of online resources to yield a SERP that prioritizes *educationally valuable* and *comprehensible* resources while minimizing those that are *objectionable*. As such, we introduce `REdORank`, a novel re-ranking framework based on multi-perspective LTR meant to support children's use of their preferred SE to complete classroom-related tasks. This framework leverages the optimization process of LTR to learn a *balance* between the *risks* of inappropriate resources and the *rewards* of contextually relevant resources. In the interest of reproducibility, we share the implementation of `REdORank` in `https://github.com/Neelik/REdORank`.

II. Related Work

When using mainstream SE, children tend to explore SERP sequentially from top to bottom and click higher-ranked results [5], [11], [16]. Consequently, existing solutions addressing this behavior re-rank resources according to a user-defined

reading level [17], promote child-friendly sites or those of educational value [18]–[21]. Unfortunately, these strategies prioritize resources using a single perspective, which might not be sufficient when serving particular user groups and contexts.

Within the educational domain, Yilmaz et al. [22] label queries that align with educational subjects and use the labels as an indicator for re-ranking resources in the Turkish language. Usta et al. [23] train an LTR model for a query-dependent ranking strategy to prioritize resources for students in the 4^{th}–8^{th} grades. The authors extract features from the query logs of a Turkish educational platform through feature engineering. This approach differs from ours because the features originate from a domain-specific SE that includes course and grade information on the resources. In contrast, we design a re-ranker that is SE agnostic. Additionally, the features used in [23] include click data originating from children, which is not readily or publicly available for our user group.

Similar to `REdORank`, *Korsce* [10] examines the appropriateness, curriculum alignment, objectivity, and reading comprehensibility of resources to identify those fitting children in the 3^{rd} – 5^{th} grades. Korsce considers resources referring to pornography and hate speech inappropriate but overlooks other objectionable topics. Curriculum alignment is based on topic modelling, which involves a word-level and semantic space exploration but disregards contextual information that can be garnered from resource text in its entirety. Readability estimation is based on the Flesch-Kincaid formula. Yet, other formulas have been reported to be more effective when predicting the readability of K–12 resources [24]. Korsce requires a user's expected grade, which is rarely available for mainstream SE. It also ranks resources according to a static set of optimal weights, manually chosen as the result of empirical exploration of near-optimal rankers [25]. The selected ranker generates scores resource-by-resource (akin to pointwise methods). Alternatively, we utilize a listwise approach, allowing for absolute relevance comparisons among resources.

III. Methodology

`REdORank` is a multi-perspective LTR framework that re-ranks resources through examining *in tandem* the Readability, Educational alignment, and Objectionability of each resource R retrieved by a mainstream SE in response to a child's query inquiring on classroom-related concepts. Leveraging the power of mainstream SE, `REdORank` prioritizes resources intended for K–5 classrooms and students by taking advantage of its three modules: *reward*, *risk*, and *balance*.

A. Reward

The reward module determines the interaction between "positive" perspectives for resource analysis.

Readability. For resources to be useful, children must be able to decode and comprehend the information within them [26]. Readability, or "the overall effect of language usage and composition on a readers' ability to easily and quickly comprehend the document" [27], aids in identifying resources that children can understand. As shown in Eq. 1, in `REdORank`, the readability score S_{read} of R, inferred using its snippet R_S[1] is determined by `Spache-Allen` [24]. This formula utilizes a large vocabulary comprised of a broad range of terms that children acquire as they age and was empirically found to be effective for estimating the reading difficulty of children's online resources.

$$S_{read}(R) = \textit{Spache-Allen}(R_S) \tag{1}$$

Educational Resources. Not all resources aligned with children's reading abilities are suitable for the classroom. To explicitly respond to our environment, `REdORank` considers the educational alignment of resources and aims to promote those with educational value, as previous research has shown that ranking educational resources higher in search results has the potential to increase learning efficiency [28]. We focus on educational resources that inform on subjects targeted for children in grades K–5, such as developmentally appropriate language arts, science, and social studies. As shown in Eq. 2, to capture the degree to which a web resource R is educationally aligned, we employ `BiGBERT` [29], the Bidirectional Gated Recurrent Unit with BERT model. `BiGBERT` examines the URL (R_U) and snippet (R_S) of R based on known educational standards, such as the United States' Common Core State Standards and the Next Generation Science Standards. S_{edu} has a range of $[0, 1]$.

$$S_{edu}(R) = BiGBERT(R_S, R_U) \tag{2}$$

B. Risk

The risk module looks at "negative" perspectives that identify resources as inappropriate for the user group.

Given the user group and environment of interest, `REdORank` must mitigate the risk of presenting towards the top of SERP resources that could be deemed inappropriate. Doing so while avoiding over-filtering results that may appear objectionable but are not, e.g., an article on breast cancer [3], requires a solution beyond safe search. Inspired by prior strategies to detect objectionable resources [10], [30], we treat as objectionable for children in the classroom resources that relate to any category in ObjCat: Abortion, Drugs, Hate Speech, Illegal Affairs, Gambling, Pornography, and Violence. The Drugs category refers to resources over-arching *drugs*, but also *alcohol*, *tobacco*, and *marijuana*. Violence focuses on violent content, as well as *weapons*; Hate Speech accounts for *racism* and hateful/offensive content.

To identify objectionable resources, we extend a state-of-the-art model to produce $Judge_{bad}$, a lexicon-based classification model that scrutinizes their terminology [10]. The original model uses vocabulary from the pre-defined lists sourced from Google's archive [31] and the Hate Speech Movement's website [32] for the Pornography and Hate Speech categories. Since we extend to other categories, defining further lists of

[1]Due to the complexities of gathering, computing resources, and storage needs for processing the full content of Web pages, we use snippets as proxy.

'objectionable' terms is necessary. Without curated term lists for the remaining categories, we generate them through a novel process called category understanding via label name replacement [33]. For this, we use websites from Alexa Top Sites [34] known to belong to categories in ObjCat as our corpus. For each category, the occurrence of the category name (and sub-category names, if available) within a website is masked, and a pre-trained BERT encoder is used to produce a contextualized vector representation h with the masked category name. BERT's masked language model head yields a probability distribution that a term w from within BERT's vocabulary will occur at the location of the masked category name. As terms can occur in different contexts within the same corpus, extracted terms are ranked by their probability of occurrence and by how many times they can replace a category name in the corpus while maintaining context. As in [33], the top 100 terms per category are used as the representative list.

We represent R with a collection of 16 text-based features extracted from its snippet R_S that account for the prevalence (i.e., frequency of occurrence) of objectionable terms in R_S; scenarios where a term could be misconstrued as objectionable depending on context; and the fact that producers of objectionable online content are known to introduce intended misspellings as an attempt to bypass safe search filters [10].[2]

We use a Random Forest model to identify objectionable resources based on its effectiveness in similar classification tasks [10]. Using the feature representation of R as input, a trained Random Forest model (max leaf node, min leaf samples, and min sample split are set to 32; max depth to 8) produces a binary probability distribution $\hat{\mathbf{y}}$ over each class–objectionable and not–such that $\hat{\mathrm{y}} \in [0, 1]$ for R. To serve as the sensitivity score exploited by the risk module, we define S_{bad} as the probability value of R being associated with the objectionable class (Eq. 3).

$$S_{bad}(R) = Judge_{bad}(R_S) \tag{3}$$

C. Balance

The balancing module trades off the outputs of the *risk module* (a value that acts as cost and therefore decreases resource prioritization) and the *reward module* (a value meant to increase resource prioritization in the ranking), resulting in a final ranking score by which resources are reordered.

Listwise LTR & AdaRank. LTR is a machine learning strategy that, when applied to Information Retrieval, refers to the task of automatically constructing "a ranking model using training data, such that the model can sort new objects according to their degrees of relevance, preference, or importance" [36]. LTR models accept more than one resource as input, resulting in pointwise, pairwise, or listwise variations [37] that consider either a single resource, a pair of resources, or a list of resources, resp., during the optimization of the loss function. When used for Web search, listwise models are reported to be more effective than the pointwise and pairwise counterparts [38], [39]. Popular listwise models [40], [41], however, optimize their ranking functions on a single relevance measure. In practice, relevance does not always depend on a single trait, as relevance judgments are associated with concepts like usefulness, utility, pertinence, etc. [42]. To better align with such real-world scenarios, multi-objective LTR strategies that optimize loss functions for multiple measures of relevance have been brought forth [25], [43]. Yet, such approaches opt for the pairwise variation [44], [45].

When accounting for multiple objectives, listwise approaches like AdaRank [41] are rarely considered, despite being one of the more prevalent algorithms in LTR research [46], [47]. AdaRank learns a ranking function through the optimization of an evaluation measure. The metric most commonly used for optimization is Normalized Discounted Cumulative Gain (**NDCG**), which measures the agreement between a predicted ranked list and the ground truth for a query. This style of LTR is geared towards a single relevance value with respect to a query and does not account for "risk" factors of resources.

Cost-sensitive Optimization. The goal of a search system is to retrieve resources from a collection that have the highest relevance with regard to a user's query. In some cases, these collections contain resources not meant to be seen by all users, such as private medical documents or top secret missives in the case of a government system. These types of resources are known as sensitive resources. To avoid presenting sensitive materials in response to online inquiries, Sayed and Oard [48] introduced an extended version of the DCG metric, called Cost Sensitive Discounted Cumulative Gain (**CS-DCG**). This new metric (Eq. 4) introduces a cost penalty, or risk factor, for displaying a sensitive document within a k retrieved resources ranking.

$$CS - DCG_k = \sum_{i=1}^{k} \frac{g_i}{d_i} - c_i \tag{4}$$

where i is a position in the ranking, g_i is the relevance gain of the i^{th} resource, and d_i is the discount for the i^{th} resource.

Incorporating CS-DCG into an LTR model such as AdaRank empowers the model to learn to rank sensitive documents lower than those that are not sensitive. This aligns with what we seek to do with the objectionability perspective of `REdORank`: eradicate from top-ranking positions resources that can be perceived as sensitive for the user group and environment that are the focus of our work. Thus, instead of depending upon the traditional NDCG for optimizing its LTR re-ranker, `REdORank` uses CS-DCG. In this case, we use as the sensitivity cost c_i, S_{bad} (Eq. 3).

CS-DCG accounts for objectionable resources but still only considers a single signal for relevance gain. In the context of our work, however, it is imperative to leverage the influence that both educational alignment and readability have on determining a given resource's relevance. It is not sufficient to simply linearly combine the respective grade level and

[2]Feature implementation details are excluded due to space limitations, and the fact that they are not the main contribution of this work. Implementation can be found on Github (§I), with a detailed description in [35]

educational alignment scores, S_{edu} and S_{read}. Instead, it is important to understand these two scores' interdependence in dictating relevance gain.

We take inspiration from the popular TF-IDF weighting scheme to model the connection between educational alignment and readability. Intuitively, we treat S_{edu} as representative of the content of R (in terms of matching the classroom setting) and readability as the discriminant factor with respect to resources considered for ranking purposes. Given the often high readability levels of online resources [14], we use 13 as the readability level representative of the collection and, therefore, use it as the max readability in the numerator for IDF. The mixer score for R informed by the two aforementioned signals of relevance is computed as in Eq. 5.

$$mixer(R) = S_{read}(R) \times log_2(\frac{13}{S_{edu}(R)}) \quad (5)$$

By incorporating multiple signals of relevance into determining relevance gain and expanding DCG with a cost-sensitivity factor, we have defined an updated metric that ensures `REdORank` explicitly learns to respond to the user group, task, and environment requirements.

IV. Experimental Set-up

While **datasets** like MQ2007 are available for evaluating LTR models, none comprises queries, resources, and "ideal" labels on our user group (children ages 6–11) and environment (classroom). In addition, these datasets do not include known objectionable resources, which are crucial for assessing the validity of `REdORank`'s design. Thus, we construct our own dataset: RankSet. Beginning with a known "ideal" resource, we use its title as a query to retrieve other resources to produce a ranked list. The ideal resource is always positioned at the top of the ranking, as it is treated as the ground truth. The remaining top-N ranked resources (excluding the one originating the search, if available) are used to complete the ranked list. To evaluate `REdORank`'s ability to push objectionable resources towards the bottom of the ranking, we append at the end of the list a known "bad" resource. To act as ideal resources, we use 9,540 articles from NewsELA [49] with known reading levels and educational value targeted for children on various topics. For bad resources, we turn to ObjSet. We use Google's API to retrieve up to 20 resources, their titles, search snippets, and rank positions (we drop queries that lead to no resources or resources with missing content) for each query inferred from ideal articles. We assign relevance labels of 2 to the ideal resources, 0 to the bad resources, and 1 to all other resources. This results in RankSet containing 2,617 queries and 46,881 resources.

To demonstrate the correctness of `REdORank`'s design and its applicability, we undertake an **ablation study**. To further contextualize the performance of `REdORank`, we perform a **comparison** with a baseline and a state-of-the-art counterpart.

To **measure** performance, we use NDCG@10 and Mean Reciprocal Rank (**MRR**). Given the importance of positioning objectionable resources very low among retrieved results, we also compute an alternative version of MRR, in which we account for the position of the first objectionable item. We call this $\mathbf{MRR}_B$, where a lower value indicates better performance. The significance of results is verified using a two-tailed student t-test with $p<0.05$; all results reported and discussed in the following section are significant unless stated otherwise.

V. Results and Discussion

We begin our evaluation of adapting LTR to children searching in the classroom by looking at how a known listwise LTR algorithm, AdaRank, optimized for a standard metric, performs when trained to rank according to our perspectives. We train variations of AdaRank with each perspective, educational alignment, readability, and objectionability, each acting as a single feature. We refer to these variations with the suffixes -E, -R, and -O, resp. We train the same set of variations for `REdORank` with the addition of ones that use the mixer to combine the educational alignment and readability perspectives into a single feature. We refer to these with the suffixes -M, where the mixed values are the only feature, and -MER, where the mixed values are used alongside the individual perspectives. Results are presented in Tables I and II.

As anticipated, AdaRank-O performed the worst, i.e., lower NDCG and MRR scores but higher MRR_B. We attribute this to AdaRank-O optimizing for the "risk" perspective and thus learning to prioritize the known bad resource above the known ideal. When optimizing on the "reward" perspectives, AdaRank-E and AdaRank-R outperform AdaRank-O. These models place objectionable resources around the 10^{th} position according to MRR_B; ideal ones around the 5^{th} position, according to MRR (Rows 1–3 in Table I). This indicates these models are learning to focus on the types of resources well-suited for our user group and environment. When considering

TABLE I
Ablation study using RankSet. * indicates significance w.r.t. REdORank and bold indicates best performing for each metric.

Row	Algorithm	Optimization Metric	NDCG	MRR	MRR_B
1	AdaRank	NDCG	0.778*	0.226*	0.097*
2	AdaRank-E	NDCG	0.765*	0.209	0.110*
3	AdaRank-R	NDCG	0.774*	0.222	0.101*
4	AdaRank-O	NDCG	0.675*	0.148*	0.537*
5	REdORank-E	nCS-DCG	0.765*	0.209	0.110*
6	REdORank-R	nCS-DCG	0.774*	0.222	0.101*
7	REdORank-O	nCS-DCG	0.675*	0.148*	0.537*
8	REdORank-M	nCS-DCG	0.765*	0.209	0.110*
9	REdORank-MER	nCS-DCG	0.777	0.218	**0.089***
10	REdORank	nCS-DCG	**0.779**	**0.228**	0.097

TABLE II
Assessment using RankSet. * indicates significance w.r.t. REdORank and bold indicates best performing.

Algorithm	Optimization Metric	NDCG	MRR	MRR_B
LambdaMART	NDCG	**0.784**	**0.228**	**0.081***
Korsce	N/A	0.753*	0.209	0.163*
REdORank	nCS-DCG	0.779	**0.228**	0.097

all of the features together, AdaRank outperforms each individual variation, showcasing that the multi-perspective design choice for REdORank is well-founded.

We surmise that the AdaRank models are learning to rank objectionable resources lower as a beneficial side-effect of optimizing on the educational alignment and readability. To account for objectionable as an explicit signal of cost and to balance that risk with the reward of the other perspectives, we turn to REdORank, optimized for normalized CS-DCG (nCS-DCG). For REdORank-E, REdORank-R, and REdORank-O, we see similar performances to those of their AdaRank counterparts (Rows 5–7 and 2–4 in Table I, resp.). This further highlights that the perspectives matter. We posit that the interconnection of educational alignment and readability will serve as a beneficial composite signal for the relevance of resources. Hence, we utilize the mixer to combine the two perspectives. Surprisingly, REdORank-M performs worse in all metrics when compared to REdORank-R and performs the same as REdORank-E. To fully investigate whether this combined perspective could provide value to the re-ranking, we created REdORank-MER. Lending credence to the idea of incorporating a combined perspective, REdORank-MER outperformed each of the individual perspective variations. While this variation performed significantly better than REdORank in terms of MRR_B, it performed worse for the other two metrics. This highlights that the explicit consideration of a sensitivity cost factor, alongside multiple perspectives of relevance, has beneficial effects on re-ranking resources for children searching in the classroom.

To attain a better understanding of how REdORank performs, we also compare it to both a state-of-the-art counterpart, Korsce (see §II), and a baseline LTR algorithm, in LambdaMART, a popular pairwise LTR model that utilizes Multiple Additive Regression Trees [50]. The results of these two models ranking the resources in RANKSET can be seen in Table II. We see that REdORank performs significantly better than Korsce for all metrics. This is visually represented in Figure 1. We attribute the difference in performance to the fact that Korsce ranks in a pointwise, weighted objective manner. In an unexpected outcome, LambdaMART's performance was comparable to that of REdORank–except for MRR_B, differences in performance were not significant. Recall that RANKSET contains a single ideal resource, which can be more easily "located" by a pairwise algorithm due to the one-to-one comparisons made. In contrast, a listwise approach looks at all resources simultaneously, allowing a single ideal resource to get lost in the crowd. We attribute this characteristic of our dataset to be the cause of the performance differences. However, a listwise approach is better suited to the re-ranking task in real-world scenarios, where more than one ideal resource is likely in a single list [38].

Going back to our RQ, given its visibly higher lower bound on NDCG@10 over its counterparts (see Figure 1), its successful performance regarding ranking known educational and readable resources high in the rankings, and its expected generalizability to real-world re-ranking scenarios, we consider the design of REdORank to be an appropriate model for providing re-ranking to search systems supporting children's online inquiry activities in the classroom.

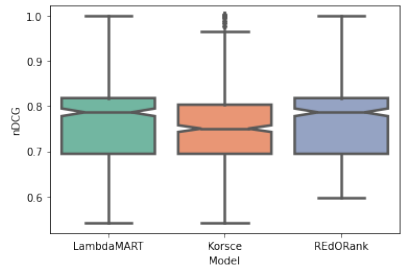

Fig. 1. NDCG@10 for different re-ranking models using RANKSET.

VI. CONCLUSIONS, LIMITATIONS, AND FUTURE WORK

REdORank, the novel re-ranking strategy presented in this manuscript advances Information Retrieval for Children–centered on the design, development, and assessment of strategies that enable children's information discovery. REdORank examines resources retrieved by commercial SE and prioritizes them based on three perspectives, i.e., educational alignment, readability, and objectionability, so that those best suited for the context and user group at hand are ranked higher. In turn, it serves as a means to ease SERP exploration when children interact with the SE they favor. Given its promising offline evaluations, we plan to study REdORank in a realistic environment.

Currently, REdORank depends upon readability, which applies to English language resources. However, considering the vastness of the web, exploring multilingual readability formulas and other estimation methods that account for the presence of media elements, e.g., images and charts on web pages, could offer valuable insights. We also suggest increasing the granularity of $Judge_{bad}$ in identifying objectionable content based on specific age groups. It is worth researching the benefits of combining additional relevance signals beyond text, such as the origin or authorship of a resource. Such factors contribute to the credibility of a resource. Unfortunately, children are known not to judge the credibility of online resources [51], making credibility a valuable extra perspective to bring into the fold for re-rankers. Ongoing research in HCI has explored the impact of visual elements of a SERP on children's search behavior [9], [52]. REdORank provides further avenues of exploration regarding identifying resource types and elements that can serve as cues. Integrating visual elements that align with the ranking process can enhance the transparency of search systems. This can impact the ease of use and understandability of a system, and integrating such elements could benefit users learning to search [1], [53].

ACKNOWLEDGMENT

Work partially funded by NSF Award #1763649.

REFERENCES

[1] I. M. Azpiazu, N. Dragovic, M. S. Pera, and J. Fails, "Online searching and learning: Yum and other search tools for children and teachers," *IRJ*, vol. 20, no. 5, pp. 524–545, 2017.

[2] R. Rajalakshmi, H. Tiwari, J. Patel, R. Rameshkannan, and R. Karthik, "Bidirectional gru-based attention model for kid-specific url classification," in *Deep Learning Techniques and Optimization Strategies in Big Data Analytics*, 2020, pp. 78–90.

[3] V. Figueiredo and E. M. Meyers, "The false trade-off of relevance for safety in children's search systems," *ASIS&T*, vol. 56, no. 1, pp. 651–653, 2019.

[4] D. Bilal and L.-M. Huang, "Readability and word complexity of serps snippets and web pages on children's search queries: Google vs bing," *Aslib*, 2019.

[5] J. Gwizdka and D. Bilal, "Analysis of children's queries and click behavior on ranked results and their thought processes in google search," in *CHIIR*, 2017, pp. 377–380.

[6] I. Madrazo, N. Dragovic, O. Anuyah, and M. S. Pera, "Looking for the movie seven or sven from the movie frozen? a multi-perspective strategy for recommending queries for children," in *CHIIR*, 2018, pp. 92–101.

[7] N. Vanderschantz and A. Hinze, "How kids see search: A visual analysis of internet search engines," in *HCI 2017*. BISL, 2017.

[8] M. Landoni, M. Aliannejadi, T. Huibers, E. Murgia, and M. S. Pera, "Right way, right time: Towards a better comprehension of young students' needs when looking for relevant search results," in *UMAP*, 2021, pp. 256–261.

[9] M. Aliannejadi, M. Landoni, T. Huibers, E. Murgia, and M. S. Pera, "Children's perspective on how emojis help them to recognise relevant results: Do actions speak louder than words?" in *CHIIR*, 2021, pp. 301–305.

[10] A. Milton, O. Anuyah, L. Spear, K. L. Wright, and M. S. Pera, "A ranking strategy to promote resources supporting the classroom environment," in *WI-IAT*, 2020.

[11] E. Foss, A. Druin, R. Brewer, P. Lo, L. Sanchez, E. Golub, and H. Hutchinson, "Children's search roles at home: Implications for designers, researchers, educators, and parents," *JASIST*, vol. 63, no. 3, pp. 558–573, 2012.

[12] S. Amendum, K. Conradi, and M. Liebfreund, "The push for more challenging texts: An analysis of early readers' rate, accuracy, and comprehension," *Reading Psychology*, vol. 37, no. 4, pp. 570–600, 2016.

[13] J. Zeniarja, R. Sani, A. Luthfiarta, H. Susanto, E. Hidayat, A. Salam, and L. Mahendra, "Search engine for kids with document filtering and ranking using naive bayes classifier," in *ISemantic*, 2018, pp. 560–564.

[14] O. Anuyah, A. Milton, M. Green, and M. S. Pera, "An empirical analysis of search engines' response to web search queries associated with the classroom setting," *Aslib*, 2019.

[15] M. Landoni, D. Matteri, E. Murgia, T. Huibers, and M. S. Pera, "Sonny, cerca! evaluating the impact of using a vocal assistant to search at school," in *CLEF*, 2019, pp. 101–113.

[16] T. Gossen and A. NüRnberger, "Specifics of information retrieval for young users: A survey," *IPM*, vol. 49, no. 4, pp. 739–756, 2013.

[17] K. Collins-Thompson, P. N. Bennett, R. W. White, S. De La Chica, and D. Sontag, "Personalizing web search results by reading level," in *CIKM*, 2011, pp. 403–412.

[18] K. Gyllstrom and M.-F. Moens, "Wisdom of the ages: toward delivering the children's web with the link-based agerank algorithm," in *CIKM*, 2010, pp. 159–168.

[19] M. Iwata, Y. Arase, T. Hara, and S. Nishio, "A children-oriented re-ranking method for web search engines," in *WISE*, 2010.

[20] J. Sanz-Rodriguez, J. M. M. Dodero, and S. Sánchez-Alonso, "Ranking learning objects through integration of different quality indicators," *IEEE TLT*, vol. 3, no. 4, pp. 358–363, 2010.

[21] A. Segal, K. Gal, G. Shani, and B. Shapira, "A difficulty ranking approach to personalization in e-learning," *International Journal of Human-Computer Studies*, vol. 130, pp. 261–272, 2019.

[22] T. Yilmaz, R. Ozcan, I. S. Altingovde, and Ö. Ulusoy, "Improving educational web search for question-like queries through subject classification," *IPM*, vol. 56, no. 1, pp. 228–246, 2019.

[23] A. Usta, I. S. Altingovde, R. Ozcan, and O. Ulusoy, "Learning to rank for educational search engines," *IEEE TLT*, 2021.

[24] G. Allen, A. Milton, K. L. Wright, J. Fails, C. Kennington, and M. S. Pera, "Supercalifragilisticexpialidocious: Why using the "right" readability formula in children's web search matters," in *ECIR*, 2022, pp. 3–18.

[25] J. van Doorn, D. Odijk, D. M. Roijers, and M. de Rijke, "Balancing relevance criteria through multi-objective optimization," in *SIGIR*, 2016, pp. 769–772.

[26] S. Amendum, K. Conradi, and E. Hiebert, "Does text complexity matter in the elementary grades? a research synthesis of text difficulty and elementary students' reading fluency and comprehension," *Educational Psychology Review*, vol. 30, no. 1, pp. 121–151, 2018.

[27] C. Meng, M. Chen, J. Mao, and J. Neville, "Readnet: A hierarchical transformer framework for web article readability analysis," *Advances in Information Retrieval*, vol. 12035, p. 33, 2020.

[28] R. Syed and K. Collins-Thompson, "Optimizing search results for human learning goals," *IRJ*, vol. 20, no. 5, pp. 506–523, 2017.

[29] G. Allen, B. Downs, A. Shukla, C. Kennington, J. Fails, K. L. Wright, and M. S. Pera, "Bigbert: Classifying educational webresources for kindergarten-12^{th} grades," in *ECIR*, 2021, pp. 176–184.

[30] L.-H. Lee, Y.-C. Juan, H.-H. Chen, and Y.-H. Tseng, "Objectionable content filtering by click-through data," in *CIKM*, 2013, pp. 1581–1584.

[31] Google, "Bad word list," Retrieved from: https://code.google.com/archive/p/badwordslist/downloads.

[32] H. S. Movement, "Reports," Retrieved from: https://nohatespeechmovement.org/.

[33] Y. Meng, Y. Zhang, J. Huang, C. Xiong, H. Ji, C. Zhang, and J. Han, "Text classification using label names only: A language model self-training approach," in *EMNLP*, 2020.

[34] Amazon, "Alexa top sites," https://www.alexa.com/topsites/category, 2020, (accessed September 17, 2020).

[35] G. Allen, K. L. Wright, J. A. Fails, C. Kennington, and M. S. Pera, "A multi-perspective learning to rank approach to support children's information seeking in the classroom," 2023.

[36] T.-Y. Liu, *Learning to rank for information retrieval*, 2011.

[37] H. Li, "A short introduction to learning to rank," *IEICE TIS*, vol. 94, no. 10, pp. 1854–1862, 2011.

[38] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li, "Learning to rank: from pairwise approach to listwise approach," in *ICML*, 2007, pp. 129–136.

[39] N. Tax, S. Bockting, and D. Hiemstra, "A cross-benchmark comparison of 87 learning to rank methods," *IPM*, vol. 51, no. 6, pp. 757–772, 2015.

[40] L. Pang, J. Xu, Q. Ai, Y. Lan, X. Cheng, and J. Wen, "Setrank: Learning a permutation-invariant ranking model for information retrieval," in *SIGIR*, 2020, pp. 499–508.

[41] J. Xu and H. Li, "Adarank: a boosting algorithm for information retrieval," in *SIGIR*, 2007, pp. 391–398.

[42] P. Borlund, "The concept of relevance in ir," *JASIST*, vol. 54, no. 10, pp. 913–925, 2003.

[43] S. Bruch, S. Han, M. Bendersky, and M. Najork, "A stochastic treatment of learning to rank scoring functions," in *WSDM*, 2020, pp. 61–69.

[44] D. Carmel, E. Haramaty, A. Lazerson, and L. Lewin-Eytan, "Multi-objective ranking optimization for product search using stochastic label aggregation," in *The WebConf*, 2020, pp. 373–383.

[45] M. Momma, A. Bagheri Garakani, N. Ma, and Y. Sun, "Multi-objective ranking via constrained optimization," in *Companion of the WebConf*, 2020, pp. 111–112.

[46] S. Kuzi, S. Labhishetty, S. K. Karmaker Santu, P. P. Joshi, and C. Zhai, "Analysis of adaptive training for learning to rank in information retrieval," in *CIKM*, 2019, pp. 2325–2328.

[47] R. McBride, K. Wang, Z. Ren, and W. Li, "Cost-sensitive learning to rank," in *AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 4570–4577.

[48] M. F. Sayed and D. W. Oard, "Jointly modeling relevance and sensitivity for search among sensitive content," in *SIGIR*, 2019, pp. 615–624.

[49] Newsela, "Newsela article corpos," 2016. [Online]. Available: https://newsela.com/data

[50] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[51] E. K. Hämäläinen, C. Kiili, M. Marttunen, E. Räikkönen, R. González-Ibáñez, and P. H. Leppänen, "Promoting sixth graders' credibility evaluation of web pages: an intervention study," *Computers in Human Behavior*, vol. 110, p. 106372, 2020.

[52] G. Allen, B. L. Peterson, D. K. Ratakonda, M. Sakib, J. Fails, C. Kennington, K. L. Wright, and M. S. Pera, "Engage!: Co-designing search engine result pages to foster interactions," in *IDC*, 2021, pp. 583–587.

[53] S. Y. Rieh, K. Collins-Thompson, P. Hansen, and H.-J. Lee, "Towards searching as a learning process: A review of current perspectives and future directions," *Journal of Information Science*, vol. 42, no. 1, pp. 19–34, 2016.