2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT) | 978-1-6654-1924-6/20/$31.00 ©2020 IEEE | DOI: 10.1109/WIIAT50758.2020.00021

A Ranking Strategy to Promote Resources Supporting the Classroom Environment

Ashlee Milton*, Oghenemaro Anuyah†§, Lawrence Spear*, Katherine Landau Wright‡ and Maria Soledad Pera*
*People and Information Research Team, Dept. of Computer Science, Boise, ID, USA 83725, ashleemilton@u.boisestate.edu
†Dept. of Computer Science and Engineering, University of Notre Dame Notre Dame, IN, USA, oanuyah@nd.edu
‡College of Education, Boise State University, Boise, ID, USA 83725, katherinewright@boisestate.edu
§Work conducted while a student at Boise State University

***Abstract*—Popular search engines (SE) favored by children are optimized neither to respond to their search behavior and abilities, nor retrieve resources that align with classroom standards. Further, attempts to adapt SE to support children's inquiries have been one-dimensional, e.g., satisfy users' reading skills or search expertise. To ease the process of locating resources relevant to *children* in the *classroom*, we introduce KORSCE, a ranking strategy designed to complement the functionality of existing SE. KORSCE employs a multi-objective approach to re-rank resources retrieved by popular SE to fit a specific *target audience* and *setting* based on varied criteria: appropriateness, readability, objectivity, and curriculum-alignment. Experimental results and insights from an expert appraiser showcase KORSCE's ability to prioritize resources that assist children's information-seeking activities at school.**

***Index Terms*—search, children, classroom**

I. Introduction

Previous studies have pointed out gaps that the information retrieval community needs to address in the pursuit of the democratization of search and information access regardless of users' search ability [1]. This manifests on popular search engines (**SE**) which fall short when catering to *children* [2] in the *classroom* setting [3]. From a user perspective, SE do not explicitly consider children's distinct search style or comprehension skills [4]. From a setting perspective, SE do not promote resources that can foster learning while having an objective viewpoint [5]. Instead, resources targeting mainstream audiences are prevalent in SE Result Pages (**SERP**) [3]. Search result curation and filters could handle audience and setting concerns, but at a price. Curation can be labor intensive and result in limited resource availability; while safe-search filters can be too restrictive for the classroom [6]. In the end, curriculum-related resources are available online, but given SE tendency to favor resources based on PageRank and other commercial optimization strategies, they might not be easy to find.

We argue for the need to set a structure in place that allows children to continue using their preferred SE in the classroom but in such a way that the SE explicitly supports them, both in this particular context and as a niche type of user. As a means to respond to such need, we introduce **KORSCE**–**K**ids' **O**ptimized **R**anker for **S**earches in the **C**lassroom **E**nvironment–a ranking strategy tailored to school-aged children seeking curriculum-related information. **KORSCE** is not meant to be a new SE, instead it complements the functionality of existing SE to (i) leverage SE access to a rich set of resources, (ii) explicitly acknowledge children's preference for the use of known SE, and (iii) ease class-related information discovery tasks. We are mindful of the fact that the needs and abilities of children, along with classroom requirements, vary across ages. Consequently, we **scope** our work based on the framework presented in [7], which establishes 4 pillars for design and assessment of information retrieval systems: search strategy, user group, environment, and task. In our case, (1) existing SE enhanced with **KORSCE**, (2) children in the 3^{rd} to 5^{th} grades, (3) classroom, and (4) curriculum-related inquiries.

KORSCE's goal is to facilitate children's access to resources matching their skills and classroom constraints. Thus, it considers different criteria that simultaneously inform **resource relevance**: *appropriateness* addresses concerns with children being exposed to pornographic or hate-speech content; *readability* accounts for the need for content to be understandable; *objectivity* acknowledges children's struggle discerning fact from speculation, and their tendency to linearly explore SERP; and *curriculum-alignment* ensures that resources retrieved align with existing educational standards, such as Common Core State Standards[1] and Next Generation Science Standards[2]. We model our ranking strategy as a Multi-Objective (**MO**) problem that considers multiple criteria to be optimized. This differs from more traditional strategies that are optimized to rank resources from a single perspective, such as reading levels [8] or resource appropriateness [9].

Our work responds to areas requiring attention in SE design and offers preliminary insights that can inform development of personalized SE for the classroom. Our methodology has implications for the Search as Learning community, focused on facilitating learning to search, while searching to learn, by creating an environment where children can search at their own pace, even with limited search literacy skills, and still get resources that address their information needs [10]. Our research **contributions** include (i) a MO strategy that interacts with existing SE to support the retrieval and ranking

Work partially funded by NSF award #1565937.

[1]http://www.corestandards.org/
[2]https://www.nextgenscience.org/

978-1-6654-1924-6/20/$31.00 ©2020 IEEE
DOI 10.1109/WIIAT50758.2020.00021

Authorized licensed use limited to: University of Minnesota. Downloaded on May 23,2022 at 21:37:34 UTC from IEEE Xplore. Restrictions apply.

of resources that can best suit children searching for online resources to complete classroom-related inquiry tasks; and (ii) an analysis that showcases the importance of simultaneously considering multiple lenses for determining resource relevance for the classroom.

II. Related Work

Personalizing retrieved resources to satisfy diverse users is a challenging task. This typically requires the analysis of implicit feedback or user behavioral information, which depends upon the existence or inference of a user profile. Access to this information in the case of children is difficult, as online privacy rules like the Children's Online Privacy Protection Act and General Data Protection Regulation, prevent archiving identifiable information. Still, several alternatives have been explored for identifying and ranking the *right set of resources* in response to children's SE inquiries [11], [12]. One of the perspectives most-widely explored for our target audience is readability. Most strategies for readability-based ranking explore html-based features [8], which are known to offer a limited perspective, as well as traditional readability formulas [11], [13], which examine shallow features to estimate the complexity of a resource. Also responding to the target audience, we find strategies that analyze SERP-related features, e.g., result presentation and ease of navigation, as well as ethical-related data e.g., presence of ads [11], [14]. Others prevent children from accessing inappropriate content, i.e., pornography, hate speech, and vulgar terms, by using term weighting techniques [15] or by reformulating children's queries to educational interests instead of blocking resources [12]. Researchers have dedicated efforts to better meet the information needs of users in varied domains. One of the most prolific areas is the prioritization of objective, i.e., non-opinionated, resources [16]. Other attempts involve identifying terminology and/or resources aligning with medical and educational domains [17], [18]. Most of these strategies leverage existing ontologies and labeled resources to inform design, which are seldom publicly available when it comes to the K–12 domain. To the best of our knowledge, there is a gap in the literature regarding ranking bias towards resources aligning exclusively with K–12 context. While sites like Newsela.com offer annotated resources matching classroom requirements, these sites do not appear prominently on SERP generated by popular SE [3].

Even-though the aforementioned strategies can inform detection of resources that align with inquiries associated with classroom search, none of them simultaneously account for the target audience and setting of the proposed work.

III. KORSCE

Given a query Q formulated by a user U in grade G, **KORSCE** applies a MO strategy that examines each candidate resource R retrieved by a popular SE for Q from various perspectives for prioritization purposes. This allows **KORSCE** to leverage the indexing and retrieval capabilities of popular SE to re-rank resources in a manner that best suits our target audience (children) and setting (classroom).

A. Gathering Candidate Resources

Successfully completing the search process depends on the retrieval of resources that address the information needs expressed in users' queries, thus presenting users' resources that match their search intent is a must. SE core functionality is meant to address this premise [9], which is why we treat the set of resources identified by a popular SE for Q as candidate resources that potentially align with U's information need. In other words, this is the set of resources that **KORSCE** examines for re-ranking purposes.

Recent reports reveal that safe-search versions of popular SE can be too restrictive; resources that are likely relevant to a query can still be mistakenly filtered out, such as those related to biology and human anatomy [6]. At the same time, safe-search is known to overlook filtering resources that might not be "safe". For example, consider the term *boobies*, a kind of bird as well as a slang term for female breasts. In response to a query including this term, Google SafeSearch presents resources about birds, which we anticipated. However, it also includes resources associated with the slang interpretation of the term which, while not porn-related, are not applicable for the classroom, e.g., an Urban Dictionary definition. Thus, for candidate resource identification we defer to a popular SE without safe-search but pay special attention to deterring resources that do not respond to the needs of our target audience and setting.

B. Deterring Inappropriate Resources

Inappropriateness is a broad term, one that is challenging to objectively define. To control scope, we treat as inappropriate resources that contain *explicit content* like pornography (inspired by the premise of safe-search) and *violence-related content* like hate speech and violent acts (motivated by the current social context).

To gather evidence on R's inappropriateness for the classroom, **KORSCE** examines its content, meta-tags, and anchor-tags. This results in the 13 features described below.

- *Unique count* of sexually explicit and hate-speech words in R's content (i, ii), meta-tags (iii, iv), and anchor-tags (v, vi) that correlate with words in Google's bad words (code.google.com/archive/p/badwordslis) and Hate Speech Movement's hate term lists (HateSpeechMovement.org), respectively.
- *Proportion* of sexually explicit and hate-speech words, computed based on the number of words in R's content (vii, viii), meta-tags (ix, x), and anchor-tags (xi, xii) that match terms in the aforementioned bad/hate-based term lists, over the total number of words in R's content.
- *Proportion* of misspelled inappropriate words in R's content (xiii) that match terms in the aforementioned bad/hate-based term lists (python's enchant library).

We compute a single score to quantify the degree to which R is inappropriate: **App(R)** = TRF($\vec{R_{FS}}$). *App* is in the

Authorized licensed use limited to: University of Minnesota. Downloaded on May 23,2022 at 21:37:34 UTC from IEEE Xplore. Restrictions apply.

range [0,1], 0 indicates inappropriate, R_{FS} is the vector representation of R, and TRF is a Random Forests model[3].

C. Fostering Reading Comprehension

To be considered relevant to the target audience and setting, resources included early in a SERP should match children's reading comprehension abilities. Readability, the measure of the complexity of a text, is an essential perspective to ensure that resources presented to children can be comprehended by them. In an educational context, the most important facets of reading that inform readability are reading development and comprehension [19]. For information discovery, comprehension is imperative when evaluating a resource, since retaining new information is a goal in the classroom. Hence, it is important to examine the consequences of including a resource in a SERP that is above, at, and below the user's grade to aid in information retention. Users can easily understand resources that are below their grade; while this does not help reading development, it is not a major concern in our case. When users encounter texts above their expected grade, they experience a significant decrease in comprehension when compared to texts that match their grade [19]. Thus, we design **KORSCE** to explicitly consider the degree to which U can comprehend R.

We use Equation 1 to compute *Rd*(R), the reading comprehension score associated with each resource. This score penalizes resources according to the degree to which their readability level, i.e., $RL(R)$, deviates from G (a proxy for U's reading abilities). We base *Rd* on the findings presented in [19], which indicate that students reading at grade level have a 75.79% comprehension rate, dropping to 31% for text above grade level. As such, a resource at the expected grade level carries a weight of 1.

$$\textbf{Rd(R,G)} = \begin{cases} 1 & RL(R) = G \\ \frac{cos(0.79 * RL(R) - (G - 0.21 * G)) + 1}{2} & G < RL(R) < G + 4 \\ \frac{cos(0.5236 * RL(R) - 0.5236 * G) + 1}{2} & G - 6 < RL(R) < G \\ 0 & \text{Otherwise} \end{cases} \quad (1)$$

For $RL(R) > G$, we use a cosine curve to mimic the approximate averages in [19] for above grade level. Cosine provides a gentle slope when close to 1 and becomes more severe as it grows farther away, which serves the purpose of our penalties. Similarly, we employ another curve for $RL(R) < G$, but it is less steep overall as there are no reported negative consequences to comprehension at lower levels. We still use a cosine curve in this context, to recognize that if a text is deemed too easy for a user, there can be a negative effect on reading growth. Estimating readability is non-trivial, as there is a plethora of formulas. For online resources, this is compounded by the lack of a standard formula for this type of text. Flesch-Kincaid is the formula used by several organizations, including corporate and military, and it has recently been applied specifically for the context for online resources and children [20]. For these reasons, we use Flesch-Kincaid to estimate $RL(R)$.

[3]TRF was chosen via an empirical exploration of various machine learning alternatives to determine the most adequate for this task. Maximum depth=8; maximum leaf node, minimum leaf samples, and minimum sample split are all 32. Model parameters are set via grid search.

D. Quantifying Resource Objectivity

Children may not have sufficient skills to determine if resource content is opinion or fact [21]. This prompts **KORSCE** to prioritize resources that are objective, i.e., non-opinionated. To do so, we adapt the approach first introduced in [16] based on language models. This approach examines resource content to determine if it is more likely to contain vocabulary frequent in either objective or subjective texts. As per [16], we build three language models. For the subjective and objective models, C_S and C_O are subjective and objective document collections, V_{C_S} and V_{C_O} represent the respective vocabulary of these collections, w_i serves as the target word, w_j is a word in the respective collection, and $c(w_x, C_y)$ represents the count of the words in a collection where, w_x stand for either w_i or w_j and C_y is either the objective or subjective collection. For R's language model, $P(w_i|\hat{\theta}_R)$ is the probability distribution for words in R, α(=1) is Jelinek-Mercer's smoothing parameter, and Z is R's vocabulary size.

Obj: θ_S:$\{P_{ML}(w_i|\theta_S)$=$\frac{c(w_i,C_S)}{\sum_{w_j \varepsilon V_{C_S}} c(w_j,C_S)}$=$\frac{C(w_i,C_S)}{|C_S|}\}_{i=1}^{|V_{C_S}|}$

Subj: θ_O:$\{P_{ML}(w_i|\theta_O)$=$\frac{c(w_i,C_O)}{\sum_{w_j \varepsilon V_{C_O}} c(w_j,C_O)}$=$\frac{C(w_i,C_O)}{|C_O|}\}_{i=1}^{|V_{C_O}|}$

R: θ_R: $P(w_i|\hat{\theta}_R) = \frac{c(w,R)+\alpha}{\sum_{w_i \varepsilon Z} c(w_i,R)+\alpha|Z|}$

To estimate R's degree of objectivity, we use **Obj(R)** = D($\theta_S||\theta_R$) - D($\theta_O||\theta_R$). $D(\theta_O||\theta_R)$ and $D(\theta_S||\theta_R)$ are computed using KL–divergence (in Equation 2). KL–divergence quantifies the similarity between R's language model and a reference language model θ_1, which is either the subjective or objective language model in our case.

$$D(\Theta_1||\Theta_R) = \sum_{w \epsilon V} P(w|\theta_1) log \frac{P(w|\theta_1)}{P(w|\theta_R)} \quad (2)$$

E. Aligning Resources to the Curriculum

KORSCE explores resources to determine their alignment with the K-12 curriculum. Unlike objectivity, it is not feasible to rely solely on word-level explorations, as in this case obtaining a higher-level representation of what resources are about is the goal [22]. This requires a common semantic space that enables estimating the relative degree to which resources are related to curriculum-relevant topics.

We turn to Latent Dirichlet allocation (LDA) for topic modeling (python's Gensim library). From the education hierarchy introduced in [3], we identify 5 subject areas in the K-12 curriculum: Science, Math, Social Studies, Geography, and English (inferred from educational standards like Common Core State Standards). We use these subject areas to bias topic modeling generation by setting the number of topics of our LDA model to 6: 5 to represent known education subject areas and 1 to capture broader matters. Note that by using the LDA model, we are not trying to determine the specific topic of a

Authorized licensed use limited to: University of Minnesota. Downloaded on May 23,2022 at 21:37:34 UTC from IEEE Xplore. Restrictions apply.

resource. Rather, we want to find if it is more likely that the resource addresses varied areas of K-12 curriculum as opposed to general concepts.

We use our LDA model to generate $\vec{T_R}$=<p_1,$\cdots$,p_6>. This vector captures R's probability distribution across the 6 pre-defined topics (the first 5 related to educational subject areas, and the remaining one broad) such that p_t is the probability that R belongs to the t^{th} topic. Thereafter, **CA(R)** = $\sum_{t\in[1,5]} p_t$ yields a score that reflects the degree to which R references curriculum-related topics.

F. Ranking Resources

Traditional rankings are optimized based on a single aspect for relevance, e.g., readability [23]. In the context of our work, however, the goal is to prioritize resources that simultaneously address the needs of our target audience and setting. In other words, for a resource to be relevant it is not sufficient for it to match the reading abilities of a user, it must also be suitable for the classroom. Thus, we model our re-ranking strategy as a multi-objective (**MO**) problem.

The MO ranking strategy introduced in [9], which inspires our work, first generates a set of near-optimal rankers to capture different trade-offs with respect to the different criteria for ranking purposes. Then, the optimal ranker is determined by examining trade-offs observed using nDCG and CLEF benchmarks in the health domain. This process depends on existing (query, and expert-labeled resources) samples [9], [24] and only considers 2 dimensions of relevance. Given the lack of benchmarks in our domain, and our focus on multiple dimensions, direct applicability of this strategy is not possible, yet, we can leverage its overarching architecture. We move beyond binary relevance and instead consider that resources can be deemed *ideal*, *veto*, and *subpar*, in terms of matching, or not, our target audience and setting (*ideal* and *veto*, resp.), as well as only meeting some of the criteria (in the case of *subpar*). Further, given the non-binary nature of our labeled resources, instead of nDCG as in [9], we rely on Precision@K.

Definitions. Let C be a set of criteria (presented in Sections III-B–III-E), a specific criteria c, and a set of samples I, such that each sample i is labeled as *ideal*, *veto*, or *subpar*. Further, let RW denote a set of rankers, each of the form $\vec{RW}$=<w_1,$\cdots$,$w_{|C|}$>, where w_c captures the weight of criteria c. Lastly, $rel(i, \vec{RW})$ = $\sum_{c\in C} x_{i,c}\times w_c$ is the overall relevance score of i based on a given ranker $\vec{RW}$, where $x_{i,c}$ is the relevance score of i given criteria c.

Near optimal rankers. To determine $RankedSet$, the set of N near-optimal rankers for **KORSCE**, we conduct an exhaustive ranker space exploration, as it is finite, using Algorithm 1 and the constraints defined below.

1) $\sum_{c\in C} w_c$=1; to ensure a linear combination of weights enabling relative comparison of criterion importance.
2) $w_c \geq 0.1\ \forall_{c\in C}$; to guarantee a minimum weight for all criteria for a ranker to be deemed near-optimal for our task, i.e., no criteria should be overlooked.
3) $x_{i,c}$, w_c, $rel(i, \vec{RW}) \in [0,1]$, $\forall_{i\in I}$ and $\forall_{c\in C}$; to enable relative comparison across corresponding scores.

The top-performing ranker may not be the one that balances all the criteria. For example, a ranker having a dominant readability weight with the remaining three weights at an acceptable minimum (Constraint 2), but does not live up to the expectations for the audience-setting focus of this study. We examine the different trade-offs in weights that inform resource relevance estimation and, consequently, overall ranking of resources [9]. This results in the optimal ranker $\vec{RW}_O$ for our task[4]. **KORSCE** computes a ranking score for R which dictates its position in the SERP using **RankScore(R)** = $\vec{RW}_O \cdot \vec{R_C}$, where $\vec{R_C}$ is the vector representation of R based on the scores computed for each criteria. The closer RankScore is to 1, the more suited R is for addressing the needs of our target audience and setting.

Algorithm 1 - *Selection Process for Near-Optimal Rankers*

```
Input: I, C, N, RW, K
Output: RankerSet
for RW⃗ in RW do
    if RW⃗ fulfills constraints 1-3 then
        for i in I do
            ScoredItems ← < i, rel(i, RW⃗)>
        end for
        RankerScore ← Compute(P@K(ScoredItems[ideal]),
                              P@K(ScoredItems[veto]))
        PosRankers ←<RankerScore,RW⃗>
    end if
end for
return RankerSet ← Select top-N from PosRankers sorted by RankerScore, where
ideal is maximized and veto is minimized.
```

IV. Experimental Results

In this section, we discuss experimental set up and results from our empirical examinations.

A. Data

Obtaining child-related data is not a simple task, especially due to protection regulations that make sharing children's data highly restrictive [7]. Unlike other search contexts, where turning to TREC or CLEF for benchmarks is possible, there are no public datasets we can use for design, development, and testing of ranking strategies such as **KORSCE**. Moreover, user studies involving children are not always feasible, given the large number of users required for findings to be significant. To address these limitations, we create three datasets: **QUANTASK**, **QUALTASK**, and **SIMULATEDTASK**, which we summarize in Table I.

With **QUANTASK**, we train **KORSCE** and its associated models for deployment and validation of its overall design. **QUANTASK** is comprised of samples labeled as (i) **ideal**, i.e., resources that match the abilities of the target users and the constraints of our context, (ii) *veto*, i.e., resources that do not align with our setting and audience, and (iii) **subpar**, i.e., resource samples that for various reasons do not live up to the expectations of **ideal** such as resources that do not match

[4]For details on the empirical analysis guiding the decision-making process required to select the optimal $\vec{RW}_O$, see Section IV-D

Authorized licensed use limited to: University of Minnesota. Downloaded on May 23,2022 at 21:37:34 UTC from IEEE Xplore. Restrictions apply.

TABLE I
OVERVIEW OF DATA SOURCES USED FOR DEVELOPMENT AND VALIDATION PURPOSES. DETAILED INFORMATION ON DATA SOURCES CAN BE FOUND IN TABLE II.

Model	Resources (# of samples)	Section
Random Forest	*DMOZ* (7,000), *Alexa* (3,500), and *HateSpeechMovement* (3,500)	III-B
Objective & Subjective Language Models	*News* (44,940), *IDLA* (38,490), *FactCheck* (1,422), *YahooAnswers* (45,026), and *Blogger*(44,127)	III-D
LDA Topic Modeling	*Wikipedia* (181) and *Blogger*(500)	III-E
Multi-Objective Ranker	**QUANTASK**: *IDLA* (2,337), *Newsels* (540), *InTouch Magazine* (540), *Blogger* (900), and *LewdResults* (1,095)	III-F
KORSCE	**QUANTASK**	IV-C,IV-D
	QUALTASK: *IDLA* (259), *Newsels* (60), *InTouch Magazine* (60), *Blogger* (100), *News* (100), *YahooAnswers* (100) and *LewdResults* (121)	IV-D,IV-E
	SIMULATEDTASK: *IDLA* (25)	IV-F

classroom expectations, but are within the reading levels of a user. **QUALTASK** has a similar composition to **QUANTASK**, but it is used to validate **KORSCE**. Lastly, **SIMULATEDTASK**, which is created following a Cranfield-style paradigm, as a proxy for real-world assessment. In this dataset, we use the title of a known *ideal* resource as a query to retrieve web resources and treat the corresponding resource as the only known relevant result to the query. It is important to note that these three datasets are disjoint. We do so to ensure no cross-contamination between development and validation stages.

B. Setting Up KORSCE

To enable experimentation, we build each model required by **KORSCE** to inform its ranking using **QUANTASK**. Recall that resources used for training are disjoint from the ones used for testing purposes (in Sections IV-C - IV-F).

C. Considering Diverse Criteria

We have designed **KORSCE** to individually respond to different concerns that should be considered if SE are to better aid children in the classroom. To highlight the need for each criterion, we conduct an experiment using **QUANTASK**, where we compute each criterion score from Sections III-B - III-E for each resource in the dataset. Subsequently, we compute Pearson's correlation to capture the strength of association between each criterion. As shown in Figure 1, there is no strong correlation among the different criteria. In fact, the strongest one is between objectivity and curriculum alignment, which is expected as educational resources often focus on facts. The lack of strong correlations indicates that all criteria are essential and enable describing resources from multiple perspectives for better informed ranking.

We have shown that each criterion is useful for informing resource relevance for classroom-related inquiries, and that they are directly related to classroom context and users. We do not know, however, the consequences of bringing into the spotlight a single dimension. As an alternative indicator to the need to look beyond a one-dimensional relevance criterion, we rank each resource in **QUALTASK** by each of its calculated criteria scores individually. As depicted in Figure 2, appropriateness and readability fared the worst. The former

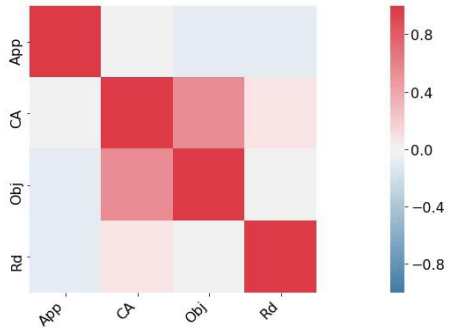

Fig. 1. Pearson correlation for **KORSCE** criteria. Red positive; blue negative. Computed using **QUANTASK**.

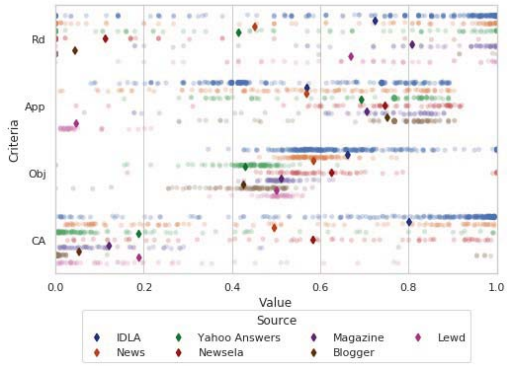

Fig. 2. Distribution of resources in **QUALTASK** ranked based on scores for readability (Rd), appropriateness (App), objectiveness (Obj), and curriculum alignment (CA). Criteria on Y-axis, ranking scores on X-axis.

tends to cluster all type of resources together–regardless of their suitability for our audience and setting–except for *veto*. This is not surprising as some curriculum-aligned resources can be marked as inappropriate due to vocabulary associated with both biology and adult content. The latter does not consistently prioritize *ideal* resources. Further, it promotes *veto* content, which is anticipated since most adults prefer to read at a 7^{th} grade, making content like magazines and adult materials have a higher readability score. While curriculum alignment and objectivity fare better, they do not provide a whole picture of our situation. Upon further inspection of the distribution of ranking scores displayed in Figure 2 for these two criteria, it comes across that some resources lead to higher ranking scores then they should. For instance, there are resources that, while curriculum aligned, contain hate-speech and thus should not be scored high. Alternatively, there is a plethora of objective samples that are definitely not suitable for the classroom, but unfortunately using objectivity in isolation make it to the top of the rank, e.g., samples colored red in Figure 2.

D. Prioritizing for the Classroom

Having illustrated the need for all the aforementioned criteria, we are now faced with the task of identifying the degree of influence each criterion should have in ranking.

Authorized licensed use limited to: University of Minnesota. Downloaded on May 23,2022 at 21:37:34 UTC from IEEE Xplore. Restrictions apply.

TABLE II
RESOURCES USED IN THE DEVELOPMENT AND ASSESSMENT OF **KORSCE**.

Source	Description	Samples instances describing
DMOZ	7,000 resources from children and teenager categories available at DMOZ	Child friendly online materials
HateSpeechMovement	3,500 resources from "Hate-speech movement", a site known to compile violence-related materials [25], [26]	Hateful online materials
Alexa	3,500 sexually-explicit materials from the "adult content" section of Alexa [27]	Sexually explicit materials
Yahoo Answers	45,126 posts and comments from subjective categories (Food & Drink, Family & Relationships, etc.) for Yahoo L6 webscope question and answering dataset [28] (*subpar*)	Subjective materials
Idaho Digital Learning Academy (IDLA)	41,111 curriculum relevant resources (*ideal*)	Readability-appropriate, objective, and curriculum-aligned material
Newsela	600 curriculum relevant resources [29] (*ideal*)	Readability-appropriate, objective, and curriculum-related material
Blogger	45,627 blog posts extracted from blogger.com [30] (*subpar*)	Subjective language in online environments
FactCheck	1,422 fact checked resources [31]	Subjective language
Wikipedia	181 Wikipedia entries about topics outlined in an existing educational hierarchy [32]	Curriculum-related content
InTouch Magazine	600 celebrity news that we extracted from InTouch Magazine (*subpar*)	General content
News	45,040 news articles collected from American news publications [30] (*subpar*)	Objective content
Lewd Results	1,216 results from queries for adult content inferred from Google Bad Words and Hate Speech Terms (*veto*)	Sexually-explicit and hate-related material

In pursuit of this goal, we follow the framework defined in [9] to explore the compromises involved in simultaneously considering multiple relevance criteria when ranking resources within complex domains. For this we use **QUANTASK**: 90% of its instances are used to identify N=3 near-optimal rankers (i.e., each ranker in $RankerSet$ as defined in Algorithm 1); the remaining 10% are used to explore the trade-offs of prioritizing criteria across the near-optimal rankers. To helps us contextualize the fact that some criteria do help enforce certain constraints that cannot be overlooked, e.g., lewd content has no place in the classroom, we also consider a ranker consisting of uniform weights across criteria, which serves as a baseline.

For near-optimal and uniform rankers, we compute a relevance score for each instance in the 10% split. We then sort instances based on their corresponding relevance score and compute Precision@K. We set K={10,143,287}, as there are 287 known *ideal* instances in the 10% split, thus capturing performance at the top 10, i.e., the number of results on a SERP, and when half, as well as all, *ideal* instances should have been found. As shown in Table III, the ranker consisting of uniform criterion weights yields the worst performance, which is expected. The ranker that obtained the best performance is the one that prioritizes objectivity. However, it does so to the detriment of appropriateness and readability. Given that both criteria are *essential* to our user, this compromise is not acceptable. There is a similar issue with the ranker prioritizing curriculum alignment. While the corresponding weighting scheme has a more even distribution, it allows for the greatest number of *veto* instances to appear high in the ranking, which is unacceptable.

TABLE III
PERFORMANCE TRADE-OFF AMONG RANKERS USING **QUANTASK**. BOLD INDICATES OPTIMAL FOR **KORSCE**.

Weights				Metrics		
App	CA	Obj	Rd	p@10	p@143	p@287
0.25	0.25	0.25	0.25	1.0	0.78	0.71
0.18	0.49	0.16	0.17	1.0	0.88	0.82
0.10	0.23	0.57	0.10	0.8	0.90	0.83
0.29	**0.31**	**0.15**	**0.25**	1.0	0.82	0.73

E. Performance in the Classroom

To take a deeper look into **KORSCE**'s performance, we use **KORSCE** (with optimal ranker from Section IV-D) to compute the ranking score that would be assigned to each resource in **QUALTASK**. Noticeably from Figure 3, most *ideal* resources are consistently positioned high on the ranking, followed closely by *subpar* resources, i.e., news articles and magazines, which we attribute to their reading ease. *Veto* resources and blogs have the lowest scores which we expected, as these resources do not align with information seeking activities at school. Instances from Yahoo Answers have, on average, a lower ranking than their educational counterparts. Nevertheless, the scores for Yahoo Answers resources are distributed across the board, as they do contain some educational content depending on the question.

We were concerned with some *ideal* resources ranking low and *veto* content ranking higher than expected. This prompted us to sample and manually examine resources. Among *ideal* resources with low appropriateness scores we found an article on "Beowulf" (an old English heroic poem). We believe this is due to the overlap of some words considered hate speech, such as "Anglo Saxon" which is used in white supremacist rhetoric, appearing in the article at a high frequency. Among the *veto* resources that ranked high, we saw some related to the workings of the judicial system, which can relate to social studies.

F. Supporting SE in the Classroom

Recall that the research question that prompted our work was exploring whether existing popular SE could be adapted to better serve our target audience and setting. Thus far we have validated **KORSCE**'s design premises, but we are also curious on its overall performance when used to complement

Authorized licensed use limited to: University of Minnesota. Downloaded on May 23,2022 at 21:37:34 UTC from IEEE Xplore. Restrictions apply.

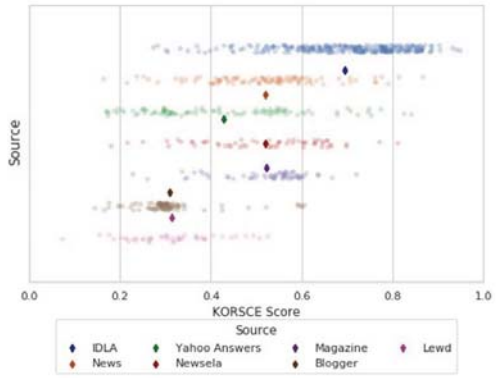

Fig. 3. Rank scores assigned by **KORSCE** to resources in **QUALTASK**; resources grouped by type to showcase score distribution. Y-axis captures source, X-axis ranking score assigned by **KORSCE**.

a mainstream SE: Bing. We explore the relevance of top resources ranked by Bing vs. **KORSCE** (i.e., resources retrieved by Bing and re-ranked using **KORSCE**). For this experiment, we use **SIMULATEDTASK** and turn to an educator to act as an expert assessor and establish ground truth. For each query in **SIMULATEDTASK**, we asked the appraiser to select among two resources (the top-one retrieved by Bing, along with the top-one re-ranked using **KORSCE**) the one that best suited the classroom. Based on appraiser responses, resources prioritized by **KORSCE** were favored for 65% of the queries. From this we infer that **KORSCE**'s MO strategy promotes more adequate resources.

G. Discussion

Traditional ranking models for the setting and audience we study have room for improvement. The use of single criterion paints a one-dimensional picture of what a user requires, and children are in no way one-dimensional. As previously stated, popular SE are not tailored to address the requirements that should be accounted for when used by young searchers within the class environment. Though our evaluations, we demonstrated that **KORSCE**'s varied relevance criteria are meaningful and necessary. The proposed MO ranking can help SE better support children by providing resources that are better suited for them and the classroom. Via our experiments, we corroborated the results in [9], as the ranker we choose does not have the best overall Precision@K but gave **KORSCE** the ability to prioritize resources that are relevant to curriculum-related inquiries.

With the help of an expert appraiser, we assessed **KORSCE** from an education standpoint compared to a mainstream SE. Using **KORSCE** with SE improves ranking for the context of our study, which is vital given children usually do not go beyond the first SERP when conducting searches. Educators' observations on resource selection brought to light gaps that future iterations of **KORSCE** should consider. For example, concise resources that include sentences directly addressing queries' information needs, lead to selection of resources retrieved by Bing. Further, there currently not a way to distinguish curriculum-aligned resources for children from those for teachers in preparation for classroom instruction.

Insights emerging from our analysis, along with observations from expert appraisers, while preliminary, serve as a first step towards personalization of SE for children in the classroom. Deployment of **KORSCE** in classrooms could lead to more information retention for children, as resources presented to them from online inquires would match their cognitive abilities. Improved recall could result in better performance at school. Additionally, teachers and parents would worry less about what their children are being exposed to online while in the classroom. Teachers could better utilize their time and energy to help children without having to monitor resources being accessed by this audience. For these reasons, next steps in our research agenda include a user study with children in the classroom directly interacting with **KORSCE**.

V. SCOPE & LIMITATIONS

We outline some limitations emerging from our work.

Users. We view **KORSCE** as a means to adapt existing SE to better aid young users in 3^{rd} – 5^{th} grade. We do recognize, however, that users can differ in needs and abilities even within the same grade. In future work, we will explore how to adapt **KORSCE** so that it can respond to not only classrooms in general, but individual users.

Language. **KORSCE** is designed to work with resources written in English but could potentially be expanded to other languages based on the availability of relevant corpora.

Readability. There are many state-of-the-art readability formulas. Depending on context and language, a different formula may be more appropriate. Due to context, language, and resource type, we settled on Flesch-Kincaid.

Simulated task. In the absence of benchmarks, we followed a Cranfield-style paradigm for dataset generation. This paradigm, while old, is still used today and can be the only option when assessing systems personalized to unique users, especially when data available for evaluation is sparse [33]. Using **SIMULATEDTASK**, we can judge **KORSCE**'s effectiveness in prioritizing resources for the classroom setting. In the future, we plan to deploy **KORSCE** in a real classroom-setting and conduct a user study for examining **KORSCE** applicability in the classroom. Exploring how users interact with resources retrieved by existing SE complemented by **KORSCE** vs. safesearch functionality would allow us to gather direct feedback from users to quantify applicability for the classroom.

Interface. Emergent searchers (below 3^{rd} grade) may find interacting with traditional text-based interfaces difficult, and instead favor graphical interfaces. **KORSCE** is meant to complement existing SE, hence, we did not account for interface constraints, which are beyond our scope.

Safe search. It is true that the use of sexually explicit and hate words has been shown to be not entirely effective in determining the appropriateness of resources. We use a lexicon-based strategy in **KORSCE**, but will explore alternatives in the future. SE specifically geared to kids are available, mainly

Authorized licensed use limited to: University of Minnesota. Downloaded on May 23,2022 at 21:37:34 UTC from IEEE Xplore. Restrictions apply.

Kiddle and KidRex, but the do not offer API's for comparisons to be possible and have been shown to have too restrictive in terms of safe search for our context [32].

Efficiency. There are many components that inform **KORSCE**. This presents a challenge in terms of efficiency for live deployment. Where a SE such as Bing will return results in tenth of seconds, **KORSCE** can take minutes to do the same. For this preliminary stage, we did not focus on efficiency but rather on the criteria that were needed to make **KORSCE** prioritize results for children in a classroom setting. In future iterations, we wish to address efficiency as it is a key factor for successful live deployment.

VI. Conclusions

Children usually do not go beyond the first SERP result when searching, and quite often click the very first resource regardless of its relevancy [10]. To better support children searching in the classroom, we introduced **KORSCE**, a novel strategy meant to complement existing SE functionality by enhancing its ranking. **KORSCE** allows SE to respond to the specific needs that arise when it comes to identifying resources that are suited to search tasks conducted by children, grades 3^{rd} to 5^{th}, in a class setting. **KORSCE** explicitly and simultaneously considers several relevance criteria in order to (i) foster resource readability to allow for better comprehension, (ii) assist children's shortcomings when identifying non-opinionated online resources, (iii) deter access to inappropriate content to prevent exposure to derogatory terminology, and (iv) prioritize information related to the classroom.

Results from our evaluation reveal the importance of the proposed criteria for ranking. They also demonstrate the value of looking beyond a one-dimensional aspect for evaluation: overall precision was not the determinant factor for deciding which set of weights better captured the needs of target audience and setting. As next steps, we will expand our research to include different tests of our strategy, such as having users test the system. Our preliminary findings could inform further research and implementation of technologies being utilized at school. These technologies can then offer the scaffolding and empower children and teachers to take better advantage of the learning environment.

References

[1] O. Le Deuff, "Search engine literacy," in *ECIL*, 2017, pp. 359–365.

[2] D. Bilal, "Ranking, relevance judgment, and precision of information retrieval on children's queries: Evaluation of google, y ahoo!, b ing, y ahoo! k ids, and ask k ids," *Journal of the American Society for Information Science and Technology*, vol. 63, no. 9, pp. 1879–1896, 2012.

[3] O. Anuyah, A. Milton, M. Green, and M. S. Pera, "An empirical analysis of search engines' response to queries associated with the classroom setting," *Aslib*, vol. 72, no. 1, pp. 88 – 111, 2019.

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

[5] D. Porcello and S. Hsi, "Crowdsourcing and curating online education resources," *Science*, vol. 341, no. 6143, pp. 240–241, 2013.

[6] 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.

[7] 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*. Springer, 2019, pp. 101–113.

[8] K. Collins-Thompson, P. Bennett, R. White, S. De La Chica, and D. Sontag, "Personalizing web search results by reading level," in *20th ACM CIKM*. ACM, 2011, pp. 403–412.

[9] J. van Doorn, D. Odijk, D. Roijers, and M. de Rijke, "Balancing relevance criteria through multi-objective optimization," in *39th ACM SIGIR*. ACM, 2016, pp. 769–772.

[10] J. Gwizdka, P. Hansen, C. Hauff, J. He, and N. Kando, "Search as learning (sal) workshop 2016," in *39th ACM SIGIR*, 2016, pp. 1249–1250.

[11] C. Eickhoff, P. Serdyukov, and A. P. de Vries, "Web page classification on child suitability," in *19th ACM CIKM*, 2010, pp. 1425–1428.

[12] N. Gupta and S. Hilal, "Algorithm to filter & redirect the web content for kids," *IJET*, vol. 5, pp. 88–94, 2013.

[13] D. Bilal and J. Gwizdka, "Children's eye-fixations on google search results," *ASIS&T*, vol. 53, no. 1, pp. 1–6, 2016.

[14] N. Vanderschantz and A. Hinze, "Do internet search engines support children's search query construction: a visual analysis," 2017.

[15] D. Patel and P. K. Singh, "Kids safe search classification model," in *ICCES*. IEEE, 2016, pp. 1–7.

[16] S. Karimi and A. Shakery, "A language-model-based approach for subjectivity detection," *Journal of Information Science*, vol. 43, no. 3, pp. 356–377, 2017.

[17] L. Soldaini, A. Yates, E. Yom-Tov, O. Frieder, and N. Goharian, "Enhancing web search in the medical domain via query clarification," *IRJ*, vol. 19, no. 1-2, pp. 149–173, 2016.

[18] R. Cortinovis, A. Mikroyannidis, J. Domingue, P. Mulholland, and R. Farrow, "Supporting the discoverability of open educational resources," *Education and Information Technologies*, vol. 24, no. 5, pp. 3129–3161, 2019.

[19] 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.

[20] D. Bilal, "Comparing google's readability of search results to the flesch readability formulae: A preliminary analysis on children's search queries," *ASIS&T*, vol. 50, no. 1, pp. 1–9, 2013.

[21] H. M. Walker, "Homework assignments and internet sources," *ACM Inroads*, vol. 4, no. 4, pp. 16–17, 2013.

[22] M. Pavlinek and V. Podgorelec, "Text classification method based on self-training and lda topic models," *Expert Systems with Applications*, vol. 80, pp. 83–93, 2017.

[23] B. Carterette, "System effectiveness, user models, and user utility: a conceptual framework for investigation," in *34th ACM SIGIR*, 2011, pp. 903–912.

[24] G. Zuccon, "Understandability biased evaluation for information retrieval," in *ECIR*. Springer, 2016, pp. 280–292.

[25] H. S. Movement, "Reports," Retrieved from: https://nohatespeechmovement.org/, Accessed: July 2018.

[26] M. ElSherief, V. Kulkarni, D. Nguyen, W. Y. Wang, and E. Belding, "Hate lingo: A target-based linguistic analysis of hate speech in social media," in *12th AAAI CWSM*, 2018.

[27] A. Internet, "Alexa: About us," Available at: https://www.alexa.com/topsites, 2019, (accessed October 2, 2019).

[28] Yahoo!, "Yahoo! datasets," Available at: https://webscope.sandbox.yahoo.com/catalog.php?datatype=l&did=11, 2019.

[29] W. Xu, C. Callison-Burch, and C. Napoles, "Problems in current text simplification research: New data can help," *TACL*, vol. 3, pp. 283–297, 2015.

[30] J. Patrick, "News and blog data crawl," Available at: https://www.kaggle.com/patjob/articlescrape, 2019, (accessed July, 2019).

[31] A. Moore, B. Adair, A. Mantzarlis, T. Cai, C. Yu, and R. Guha, "Academia, publishers and tech come together to open up fact check data," https://www.datacommons.org/docs/download.html, May 2018.

[32] O. Anuyah, I. Madrazo Azpiazu, and M. S. Pera, "Using structured knowledge and traditional word embeddings to generate concept representations in the educational domain," in *Companion proc. of the 2019 World Wide Web Conference*. ACM, 2019, pp. 274–282.

[33] E. M. Voorhees, "The evolution of cranfield," in *Information Retrieval Evaluation in a Changing World*. Springer, 2019, pp. 45–69.

Authorized licensed use limited to: University of Minnesota. Downloaded on May 23,2022 at 21:37:34 UTC from IEEE Xplore. Restrictions apply.