The Need for a Comprehensive Strategy to Evaluate Search Engine Performance in the Classroom

Oghenemaro Anuyah, Michael Green, Ashlee Milton
People & Research Information Team (PIReT)
Dept. of Computer Science
Boise State University, Boise, ID
{oghenemaroanuyah,michaelgreen, ashleemilton}@u.boisetstate.edu

Maria Soledad Pera
People & Research Information Team (PIReT)
Dept. of Computer Science
Boise State University, Boise, ID
solepera@boisetstate.edu

ABSTRACT

Given how ingrained Search Engines (**SEs**) are in educational environments, it is essential to evaluate their performance in response to inquiries that pertain to the classroom setting. In this position paper, we discuss the limitations of relying solely on traditional Information Retrieval metrics and usability studies. We argue in favor of a new comprehensive strategy for assessment that integrates other aspects (i.e., the *search context*) with new and existing measures, in order to better quantify positive and negative factors that influence SEs outcomes targeting children.

CCS CONCEPTS

• **Information systems** → **Web searching and information discovery**; **Evaluation of retrieval results**; • **Social and professional topics** → *Children*.

KEYWORDS

Search in the classroom, children, usability, evaluation

ACM Reference Format:
Oghenemaro Anuyah, Michael Green, Ashlee Milton and Maria Soledad Pera. 2019. The Need for a Comprehensive Strategy to Evaluate Search Engine Performance in the Classroom. In KidRec Workshop, ACM IDC*KidRec '19: Workshop in International and Interdisciplinary Perspectives on Children & Recommender and Information Retrieval Systems, Co-located with ACM IDC, June 15, 2019, Boise, ID.* Boise, ID, USA, 4 pages.

CHILDREN, SEARCH ENGINES, AND THE CLASSROOM

The use of search engines (**SEs**) in the classroom has widely spread, being that SEs are a "valuable asset for children's education, as it encourages learning, enhances the class environment, and introduces children in the early stages of their lives to today's information society" [2]. Popular SEs utilized in the classroom include Google and Bing, as well as child-oriented counterparts such as Kidrex, Kiddle, and Kidzsearch [1, 6]. Unfortunately, prior works have demonstrated that children do not always have successful experiences when using their preferred SEs (such as Google or Bing). This is often attributed to their insufficient skills in formulating effective queries, as well as lack of proficiency in judging the relevance of and comprehending the content of resources retrieved in response to their search inquiries [3, 7]. Hence, it is imperative that the right evaluation measures are employed to effectively quantify the correctness and adequacy of these systems when it comes to responding to this specific audience.

EVALUATION STRATEGIES

In evaluating SEs, we consider different aspects, such as their ranking algorithm, ability to capture user query intent, and other usability features (e.g., perception of the search interface). Several techniques used for assessment take advantage of benchmarks such as TREC [11] and CLEF [10], which provide labeled datasets along with defined user-generated ground-truth. Unfortunately, in the case of children, obtaining ground truth is a challenge, as data collected from this audience is subjected to stringent privacy rules [5], hence, limiting the number of publicly available datasets. Moreover, standard metrics often adopted to assess the aforementioned benchmarks may not always be sufficient in measuring the overall performance of SEs that tend to children's educational information needs.

Are Standard IR Evaluation Metrics Enough? For evaluating ***ranking tasks***, common metrics include Mean Reciprocal Rank (*MRR*), Normalized Cumulative Discount Gain (*NDCG*), and Precision-at-k (*P@K*). MRR measures the position of the first relevant result on the ranked list of resources, NDCG penalizes a system that consistently positions relevant resources low on the ranked list, and P@K quantifies the proportion of resources in a top-k set that are relevant. While these metrics tell us how good a given ranking algorithm is, their use may not help to address the problems of offering irrelevant, or in some cases, unsuitable resources to children.[1] Even though top-ranked resources may be relevant to the keywords present in a query, it may be irrelevant to a young child conducting an educational search if: (i) he cannot comprehend the content of a resource or (ii) the content of the resource does not pertain to the educational domain.

For evaluating ***filtering tasks***, e.g., disregarding inappropriate resources for children using safe search filters, the most common metric used is *Accuracy* [9]. Accuracy captures the fraction of correctly labeled resources out of those retrieved in response to a query (i.e., resources that were

[1] In this context, we refer to irrelevant resources as those that do not match the information needs expressed in a query, whereas, unsuitable resources are those that are above/below the readability levels of the user, opinion-based, or considered inappropriate for the target audience.

Figure 1: Kidrex's safe search fails to retrieve resources for the query *organ pipe cactus monument*.

correctly filtered or retained). Unfortunately, while accuracy may be effective for quantifying how useful the system is in identifying child-unsafe resources, relying on this metric alone may not tell us if the SE is too strict when it comes to handling education-relevant resources (see Figure 1 for an educational query that led to no results retrieved on a child-oriented SE). Moreover, using accuracy does not guarantee that those resources retained align with children's developmental capability, are fact-checked, or if they would be beneficial to the child.

Are Usability Studies Sufficient? A more participatory strategy for SE evaluation directly involves users representative of the target audience [1]. This approach may depend upon some form of survey responses, questionnaires, or a facilitator guiding the user through the evaluation process [8]. These practices are usually done to better capture *user preference* or *satisfaction*. However, depending on the audience involved in the user study (e.g., young children or adults) and the IR task assigned, satisfaction may mean different things: it may refer to whether the user liked the system or not, or interpreted as asking if the system improved their search experience. A recent study showed that children tend to go with the former, i.e., those options in the user study that they liked rather than accurate ones [1]. This is anticipated, due to children's lack of skills judging the quality and accuracy of information they are presented with. We believe such an evaluation approach to be insufficient on its own given that it would not necessarily quantify suitability of resources.

Is a Task-centric Perspective Needed? As traditional metrics do not consider several important abstract factors that inform adequacy of SEs in responding to children's educational search inquiries, there is a need to go beyond traditional metrics for assessment purposes. We refer to these factors as task-centric elements that are specific to a child's **search context**. In the educational environment, task-centric elements include resource *readability*, *appropriateness*, *objectivity*, and most importantly, *educational value*. Readability measure would estimate the degree to which resources that children can comprehend are prioritized, but would not reflect the detriment of presenting materials above/beyond children comprehension level. Appropriateness would quantify the SE performance in filtering child-unsafe content (e.g., pornography or hate-speech); objectivity would reward prioritization of non-opinionated and reliable resources; educational value would help quantify the level to which the resource content is education-relevant. In this scenario, we anticipate that utilizing a measure that accounts for cases where no resource is retrieved for an educational search or incorrect, unsuitable resources are offered as an alternative is necessary, as this could lead to misinforming the child in terms of both validity of information and power of SEs.

THE NEED FOR A COMPREHENSIVE MEASURE

We have discussed the detriments of relying solely on traditional IR metrics and usability studies, or even task-centric measures that can, to a degree, quantify the correctness (and potential harm) of SEs

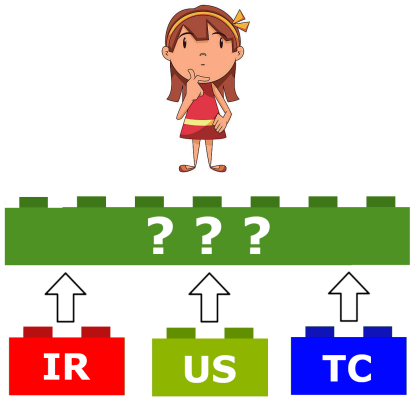

Figure 2: How do we aggregate different perspectives into a single evaluation metric that adequately quantifies SE performance in responding to children's educational tasks?

when responding to children's educational search tasks. With this in mind, we argue for the need to design a new, *comprehensive measure* that can simultaneously capture all three of the aforementioned perspectives (see Figure 2). Such a metric would be able to yield a *holistic* and *comparable* measurement that would capture overall performance of SEs that that are meant to satisfy children's information needs in a classroom setting.

Creating such a metric is a challenging endeavour, as it would have to account for aspects that are less objective than relevance or education value. For example, performance of SE would have to be judged based on not only its ability to locate and retrieve education-relevant resources, but also those that children can read and comprehend, are non-opinionated, age-appropriate, and more importantly, do not mislead its users – all factors that are dependent on each other. We believe that a starting point in this quest lies on the exploration of existing literature on retrieval and ranking in complex scenarios [4] and how it can be adopted for the specific target audience in scholastic environments.

ACKNOWLEDGMENTS

Work partially supported by NSF Award no. 1565937.

REFERENCES

[1] Oghenemaro Anuyah, Jerry Alan Fails, and Maria Soledad Pera. 2018. Investigating query formulation assistance for children. In *ACM IDC*. ACM, 581–586.

[2] Ion Madrazo Azpiazu, Nevena Dragovic, Maria Soledad Pera, and Jerry Alan Fails. 2017. Online searching and learning: YUM and other search tools for children and teachers. *Information Retrieval Journal* 20, 5 (2017), 524–545.

[3] Dania Bilal and Rebekah Ellis. 2011. Evaluating Leading Web Search Engines on Children's Queries. *Lecture Notes in Computer Science* 6764 (2011), 549–558.

[4] Toine Bogers, Marijn Koolen, Bamshad Mobasher, Alan Said, and Casper Petersen. 2018. 2nd workshop on recommendation in complex scenarios (complexrec 2018). In *Proceedings of the 12th ACM RecSys*. ACM, 510–511.

[5] Michael D Ekstrand. 2017. Challenges in evaluating recommendations for children. In *International Workshop on Children & Recommender Systems. Available at: https://drive. google. com/file/d/0B8S_mL8gIjZwVzZSLVZKN3Q4RG8/view.*

[6] Elizabeth Foss, Allison Druin, Robin Brewer, Phillip Lo, Luis Sanchez, Evan Golub, and Hilary Hutchinson. 2012. Children's Search Roles at Home: Implications for Designers, Researchers, Educators, and Parents. *JASIST* 63, 3 (2012), 558–573.

[7] Leah Graham and Panagiotis Takis Metaxas. 2003. Of course it's true; I saw it on the Internet!: Critical Thinking in the Internet Era. *Commun. ACM* 46, 5 (2003), 70–75.

[8] David Nicholas. 2003. *Assessing information needs: tools, techniques and concepts for the internet age.* Routledge.

[9] Deepshikha Patel and Prashant Kumar Singh. 2016. Kids Safe Search Classification Model. In *ICCES*. 1–7.

[10] Carol Peters. 2003. *Cross-Language Information Retrieval and Evaluation: Workshop of Cross-Language Evaluation Forum, CLEF 2000, Lisbon, Portugal, September 21-22, 2000, Revised Papers*. Vol. 2069. Springer.

[11] Ellen M Voorhees, Donna K Harman, et al. 2005. *TREC: Experiment and evaluation in information retrieval*. Vol. 63. MIT press Cambridge.