#TheHorror: Evaluating Information Retrieval Systems for Kids

Ashlee Milton
People and Research Information Team (PIReT)
Boise State University
Boise, ID 83702, USA
ashleemilton@u.boisestate.edu

Maria Soledad Pera
People and Research Information Team (PIReT)
Boise State University
Boise, ID 83702, USA
solepera@boisestate.edu

Abstract
Evaluation of information retrieval systems (IRS) is a prominent topic among information retrieval researchers–mainly directed at a general population. Children require unique IRS and by extension different ways to evaluate these systems, but as a large population that use IRS have largely been ignored on the evaluation front. In this position paper, we explore many perspectives that must be considered when evaluating IRS; we specially discuss problems faced by researchers who work with children IRS, including lack of evaluation frameworks, limitations of data, and lack of user judgment understanding.

Author Keywords
Children; Evaluation; Information Retrieval

CCS Concepts
•**Social and professional topics** → **Children;** •**Information systems** → *Information retrieval; Evaluation of retrieval results;* •**Human-centered computing** → User studies;

The issue
Evaluation in information retrieval (**IR**) remains a core research interest. Unlike most disciplines, evaluation of information retrieval systems (**IRS**), including popular ones like recommender systems and search engines, is not limited to effectiveness, as IR strives to give users what makes them

happy not necessarily what they asked for [3]. IR evaluation is an ongoing topic of interest among researchers and practitioners, with the focus being on metrics and their applicability for different tasks (e.g., multilingual IR), different domain (e.g., medical and legal), and applicability concepts like bias and fairness [22, 21, 16, 4, 14]. The target users of these explorations on evaluation have been a traditional user [3]. Thus, the frameworks and benchmarks used to evaluate IRS only account for a general user, who are not the only stakeholders.

It is our stance that the lack of evaluation frameworks for children's IRS is a problem that needs to be addressed by not only researchers working with children but the IR community. Evaluating IRS for children presents new challenges that have yet to be fully addressed. There have been a few evaluation *frameworks* presented in recent years that attempt to create structure to enable assessment for kids' IRS [15, 1]. Unfortunately, they are not always general enough to be applied to the varying IRS that are available for children. Even when these frameworks can be used, there is an overlying issue of *data*. Data from children is hard to get in several respects including collecting, accessing, and storing due to laws protecting children [6, 19, 20]. Even though some data may be accessible, how children *judge* IRS differs from the general population [11, 13, 12, 24]. We aim at showcasing (i) issues that ourselves and other researchers and practitioners face with data, user judgments, and frameworks, as well as (ii) the fact that evaluation that involves protected users is an issue with bigger implications that requires attention and must be informed by different areas of the IR community and beyond.

The trouble with frameworks

Existing frameworks have set out to make the best of the data that is available to evaluate children's IRS. The one presented in [15] brings context to evaluating and designing children's search systems with their four pillars: search strategy, user group, environment, and task. While originally designed for search systems, it is general enough to be used for context in other IR tasks [8, 7]. However, the proposed framework does not explicitly address the limitations that arise due to lack of data. In [1], the authors present a framework to address the issue of relevance judgement with children regarding search. This framework is task-specific and would need manipulation to work with other IR tasks and areas. A useful framework that lends itself to IRS evaluation and that does not require data from children is the Cranfield paradigm [5]. Cranfield uses known suitable resources as ground truth. However, it has been dropped and deemed outdated when evaluating IRS for general populations in favor of more state-of-the-art alternatives. The latter tend to require large amounts of data that is not always available with children making them not practical. Works like [23], have defended the Cranfield paradigm as a viable framework, but the IR community is divided on this issue. In the end, existing frameworks are a helpful foundation but each has constraints that render them insufficient for the IR community at large.

The woes of data

Due to privacy laws like the Children's Online Privacy Protection Act, Family Educational Rights and Privacy Act, and General Data Protection Regulation [19, 20, 6], children's data is highly protected. While these safeguards are of the upmost importance to keep children safe online, it makes collecting or finding children's data difficult. Thus, unless researchers have a means of gathering the needed data within the bounds of the child privacy laws themselves, it is impossible to get such data. Even then, there are additional rules as to what can be collected (e.g., demographics), and how it must be stored. These stringent rules (i) make it ex-

tremely burdensome to share data and (ii) can lead to insufficient data for evaluating IRS for children.

The ambiguity of judgment

Existing ground truth may be misleading–it is not always what it seems. Studies exploring children's behavior as they interact with IRS and evaluation strategies for kid-friendly IRS [18, 11] revealed that young users do not act in the same manners as adults do when interacting with or evaluating IRS [2, 11]. For example, kids tend to click on the first search result on a search result page regardless of its relevance [9, 10]. Naturally, the thought is to use the child's clicked link as the relevant result but since they tend to favor the first result, regardless of relevance, does it work as ground truth? Even if you ask children what their judgments are, instead of trying to infer them, you can still have problems. Consider when asking kids to rate on a standard 1 to 5 scale, studies have shown they tend to only rate 4's or 5's regardless of what they think [11]. Behaviors like these make it hard to define what the ground truth of collected data is and how applicable it is for conducting offline evaluations of IRS.

The unknown

The current state of evaluating IRS for children is in its infancy and it is indeed a complex undertaking driven by multiple perspectives [17]. There is not general framework that can be used consistently and is accepted by the community as a whole; there is no reliable and/or standard way to obtain data for evaluation; and ground truth requires a unique perspective of relevance and that is just not the case when it comes to IRS for children. These issues showcase not only the importance of developing frameworks without the need for massive amounts of data but also why involving the larger community to create it is key. The reason for engaging with the IR community (and beyond) is two-fold. First, if the community is involved, they will become aware of the issues attached to the development and evaluation of IRS for children. Second, the researchers and practitioners can bring in their experiences on evaluation, especially from other areas working with protected populations. Working together we can learn from each other and hopefully come up with ways to facilitate the development of evaluation in different areas of study and bring the issues of evaluation of IRS for kids into the spotlight.

REFERENCES

[1] Dania Bilal and Meredith Boehm. 2017. Towards new methodologies for assessing relevance of information retrieval from web search engines on children's queries. *Qualitative and Quantitative Methods in Libraries* 2, 1 (2017), 93–100.

[2] Dania Bilal and Joe Kirby. 2002. Differences and similarities in information seeking: children and adults as Web users. *Information processing & management* 38, 5 (2002), 649–670.

[3] Jamie Callan. 2020. Better Representation of Search Tasks. Available at: `https://www.youtube.com/watch?v=eHJTkFUxJgg&feature=youtu.be&t=18047`. (2020). (accessed May 6).

[4] Rocío Cañamares, Pablo Castells, and Alistair Moffat. 2020. Offline evaluation options for recommender systems. *Information Retrieval Journal* (2020), 1–24.

[5] Cyril Cleverdon. 1967. The Cranfield tests on index language devices. In *Aslib proceedings*. MCB UP Ltd.

[6] Federal Trade Commission and others. 1998. Children's online privacy protection act of 1998. (1998).

[7] Brody Downs, Oghenemaro Anuyah, Aprajita Shukla, Jerry Fails, Maria Soledad Pera, Katherine Landau

Wright, and Casey Kennington. 2020a. KidSpell: A Child-Oriented, Rule-Based, Phonetic Spellchecker. In *Proceedings of the The 11th International Conference on Language, Resources, and Evaluation (LREC '20)*.

[8] Brody Downs, Aprajita Shukla, Mikey Krentz, Maria Soledad Pera, Casey Kennington, Jerry Fails, and Katherine Landau Wright. 2020b. Guiding the Selection of Child Spellchecker Suggestions using Audio and Visual Cues. In *Proceedings of the The 19th International Conference on Interaction Design and Children (IDC '20)*. Association for Computing Machinery, New York, NY, USA.

[9] Sergio Duarte Torres, Djoerd Hiemstra, and Pavel Serdyukov. 2010. Query Log Analysis in the Context of Information Retrieval for Children. In *Special Interest Group on Information Retrieval*. ACM, 847–848.

[10] Jacek Gwizdka, Preben Hansen, Claudia Hauff, Jiyin He, and Noriko Kando. 2016. Search as learning (SAL) workshop 2016. In *39th International SIGIR conference on Research and Development in Information Retrieval*. ACM, 1249–1250.

[11] Lynne Hall, Colette Hume, and Sarah Tazzyman. 2016. Five Degrees of Happiness: Effective Smiley Face Likert Scales for Evaluating with Children. In *Proceedings of the The 15th International Conference on Interaction Design and Children (IDC '16)*. Association for Computing Machinery, New York, NY, USA, 311–321. DOI: http://dx.doi.org/10.1145/2930674.2930719

[12] Theo Huibers and Thijs Westerveld. 2019. Relevance and utility in an educational search environment. In *3rd International and Interdisciplinary Perspectives on Children & Recommender and Information Retrieval Systems, KidRec 2019: what does good look like?*

[13] Hanna Jochmann-Mannak, Leo Lentz, Theo Huibers, and Ted Sanders. 2016. How Interface Design and Search Strategy Influence Children's Search Performance and Evaluation. In *Web Design and Development: Concepts, Methodologies, Tools, and Applications*. IGI Global, 1332–1379.

[14] Liadh Kelly, Lorraine Goeuriot, Hanna Suominen, Mariana Neves, Evangelos Kanoulas, Rene Spijker, Leif Azzopardi, Dan Li, João Palotti, Guido Zuccon, and others. 2019. CLEF ehealth 2019 evaluation lab. In *European Conference on Information Retrieval*. Springer, 267–274.

[15] Monica Landoni, Davide Matteri, Emiliana Murgia, Theo Huibers, and Maria Soledad Pera. 2019. Sonny, Cerca! evaluating the impact of using a vocal assistant to search at school. In *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 101–113.

[16] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. 2018. Towards a Fair Marketplace: Counterfactual Evaluation of the Trade-off between Relevance, Fairness & Satisfaction in Recommendation Systems. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*. Association for Computing Machinery, New York, NY, USA, 2243–2251. DOI: http://dx.doi.org/10.1145/3269206.3272027

[17] Emiliana Murgia, Monica Landoni, Theo Huibers, Jerry Alan Fails, and Maria Soledad Pera. 2019. The Seven Layers of Complexity of Recommender Systems for Children in Educational Contexts. In *Workshop on Recommendation in Complex Scenarios, co-located with ACM RecSys*. 5–9. DOI: http://dx.doi.org/Vol-2449/paper1.pdf

[18] Maria Soledad Pera, Emiliana Murgia, Monica Landoni, and Theo Huibers. 2019. With a Little Help from My Friends: Use of Recommendations at School. (2019).

[19] Family Educational Rights. 1974. Privacy Act or 1974. (1974).

[20] Family Educational Rights. 2017. Privacy Act of 2017. (2017).

[21] Pablo Sánchez, Rus M. Mesas, and Alejandro Bellogín. 2018. New Approaches for Evaluation: Correctness and Freshness: Extended Abstract. In *Proceedings of the 5th Spanish Conference on Information Retrieval (CERI '18)*. Association for Computing Machinery, New York, NY, USA, Article 14, 2 pages. DOI: http://dx.doi.org/10.1145/3230599.3230614

[22] Daniel Valcarce, Alejandro Bellogín, Javier Parapar, and Pablo Castells. 2018. On the Robustness and Discriminative Power of Information Retrieval Metrics for Top-N Recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems (RecSys '18)*. Association for Computing Machinery, New York, NY, USA, 260–268. DOI: http://dx.doi.org/10.1145/3240323.3240347

[23] Ellen M Voorhees. 2019. The evolution of cranfield. In *Information Retrieval Evaluation in a Changing World*. Springer, 45–69.

[24] SD Wentzel. 2019. *Evaluating Information Retrieval Systems for Children in an Educational Context*. B.S. thesis. University of Twente.