Qualitative and Quantitative Methods in Libraries (QQML) 1:93 100, 2013

Towards New Methodologies for Assessing Relevance of Information Retrieval from Web Search Engines on Children's Queries

Dania Bilal[1] and Meredith Boehm[2]

[1]School of Information Sciences, University of Tennessee, 1345 Circle Park, 451 College of Communication and Information, Knoxville, TN 37996.
[2]School of Information Sciences, University of Tennessee, 1345 Circle Park, 451 College of Communication and Information, Knoxville, TN 37996.

Abstract: This paper expands on the results of a previous study by Bilal (2012) where she employed benchmarking and intellectual relevance judgment to evaluate and compare the ranking and relevancy of hits retrieved by five web search engines on thirty queries formulated by children to find information for given tasks. Given the dynamic nature of the Web and based on the findings of the study, Bilal called for new approaches for judging relevancy of information retrieval by the engines on children's queries. In the present paper, Bilal and Boehm propose a new multi-tiered research method that could produce a more nuanced and context-based relevance assessment. This method, *Reconciled Relevance* (RR) combined with *Reconciled Relevance Ranking* (RRa) is described and challenges for implementing it are outlined. The method has implications for child-driven relevance judgment, ranking algorithms, and relevance theory, as well as for the roles of mediators in maximizing children's web experiences.

Keywords: Web Search Engines, Children, Youth, Benchmarking, Relevance, Reconciled Relevance, Reconciled Relevance Ranking, Methodologies, Information Retrieval, Evaluation.

1. Introduction

The dynamic nature of web search engines makes the notion of time and place a critical aspect in evaluating relevancy of information retrieved by the engines on user queries. While existing "relevance" literature provides good understanding of the notion of relevance, of classic and newer evaluation metrics, and of the role of the user in judging relevance, there is a continuous need to develop new evaluation methods that takes into account the ever changing "ranking" of retrieved information on user queries by the engines. Today's children hardly use search engines designed for their age levels; and rely on Google as their gateway for

Received: 11.4.2012 / Accepted: 20.9.2012 ISSN 2241-1925
© ISAST

finding information (Bilal, 2012, Druin, et al., 2010; Bilal, 2004). While existing research on relevance judgment and ranking of retrieved output by search engines is abundant, there is very little literature that has evaluated and compared the ranking and relevancy of output by large-scale engines used often by young users (e.g., Google, Yahoo, Bing) to engines that are specifically developed for their age levels (e.g., Yahoo Kids, and Ask Kids). The fact that children are developing more affinity to large-scale engines (Google, Yahoo, and Bing) that are geared for the general public makes the evaluation of the retrieval performance of the engines on children's queries a much needed area of investigation. One needs to develop understanding of what children expect to find in these engines, how is it ranked, and how much of what is ranked at the top of retrieved hits is relevant to the context of the tasks and information needs reflected in children's queries.

Bilal & Ellis (2011) used Google and Yahoo Kids as benchmarks to compare overlap in retrieved results by the engines that had the same ranking as Google's and Yahoo Kids' top five ranked hits retrieved on the first results page using thirty queries children formulated to find information for given tasks. In a recent study, Bilal (2012) built on the findings of the previous research, and in concert with a trained graduate assistant, she and the assistant judged relevancy of each hit retrieved for a given query by a given engine on a three-point relevance scale. In addition, Bilal calculated recall and precision ratios of relevant and partially relevant hits retrieved by a given engine for a given query. One of the recommendations Bilal (2012) made for future research is to involve children in judging relevancy of information retrieval by the search engines for their queries. Nonetheless, the dynamic nature of relevance from the user's perspectives, and in this case children, may pose additional challenges due to the fact that human-based relevance judgment is characterized as dynamic, situational, cognitive, and emotional (Saracevic, 1975; Schamber, Eisenberg, & Nilan, 1990; Harter, 1992; Barry, 1994; Schamber, 1994; Wang & Soergel; 1998; Saracevic, 2007a-b). While a myriad of studies have evaluated the retrieval performance of web search engines (Spink, Jansen, Blakely, & Koshman (2006); Thelwall (2008); Lewandowski (2008; 2011); Spink & Jansen, (2004); Spink, A., and Greisdorf, 2001; and Vaughan (2004), very little research has investigated the performance of search engines on queries formulated by children using benchmarking and intellectual relevance judgment.

In the present paper, we review the two research methods Bilal employed in her 2012 study (benchmarking and intellectual relevance judgment), and propose a new research framework for a multi-tiered approach, *Reconciled Relevance (RR)* that combined with *Reconciled Relevance Ranking* (RRa) could produce a more nuanced relevance assessment of information retrieval by search engines, particularly for children's queries. The proposed method has implications for chid-driven relevance judgment, ranking algorithms, and relevance theory, as well as for the roles of mediators in assisting children in judging relevance of search engine outputs.

2. Research Question

Given the fact that children resort to mediators for assistance (e.g., information specialists, teachers, parents) in understanding tasks and finding information on the web, and based on the literature review and synthesis of the study findings by Bilal (2012), we addressed this overarching question: *What evaluation methods should be employed to judge the ranking and relevancy of information retrieved by search engines on children's queries?*

3. Related Studies

Many studies have evaluated the retrieval performance of web search engines using different research methodologies and queries. These studies were reviewed by Bilal (2012) and will not be covered in this section. Rather, benchmarking and intellectual relevance judgment Bilal (2012) employed in her recent study are synthesized within the context of published literature. In the latter study, Bilal extracted thirty queries from published literature on children's information behaviour and interaction with digital interfaces and input each of the queries into each of the five search engines (Google, Yahoo, Bing, Yahoo Kids, and Ask Kids). Results retrieved by a given engine for a given query was benchmarked to each of Google's and Yahoo Kids' top five ranked results it retrieved on the first results page consisting of ten hits per page.

Benchmarking: Google was used as a benchmark due to its popularity among children. Yahoo Kids was selected as a benchmark because it is targeted for children ages 7-12. Only hits retrieved by a given engine for a given query that overlapped with the ranked output by a benchmark was evaluated and compared, and the percentage in overlap in hits that had the same ranking as a benchmark was calculated.

Intellectual Relevance Judgment: Relevance judgment of retrieved output on the given queries produced by the five search engines was made by Bilal and a trained graduate teaching assistant (jurors). Each retrieved hit was evaluated in relation to its contribution to resolving a given task either partially or fully for which children constructed a given query. In addition, each retrieved hit for a given query by a given engine was judged for relevancy by each juror based on the query itself, respective task, topic, and context using Mizzaro's framework (1997). Relevance of a given retrieved hit (titles of links, summaries, URLs, and respective pages) for a given query by a given engine was evaluated and scored on a three-point relevant scale, 1=relevant, 0.5=partially relevant, and 0=not-relevant. Submission of the thirty queries to the five engines resulted in 1500 hits, 500 for one-word queries, 500 for two-word queries, and 500 for phrase or natural language queries.

The findings of Bilal's (2012) study showed that Yahoo and Bing were similar in their retrieval performance on the queries in relation to the overlap they produced that benchmarked to Google's top five ranked hits. In addition, the two engines yielded similar precision ratios of relevant hits, though the ratio by Bing was slightly higher than that by Yahoo. Ask Kids outperformed Yahoo Kids in finding overlap that benchmarked to Google's top five ranked hits, and also surpassed it on the precision of relevant hits it produced across the queries.

Using two differing search engines as benchmarks allowed for a contextual understanding of the differences between the engines in terms of what children would expect to retrieve for their queries. In the case of Bilal's (2012) study, we see that while use of Google as a benchmark resulted in hits that overlapped with Yahoo, Bing, and Ask Kids, utilizing Yahoo Kids as a benchmark yielded very few hits. In terms of relevancy, Ask Kids resulted in much more relevant hits on children's queries than did Yahoo Kids. In fact, much of the "unique" content that Yahoo Kids retrieved across the queries was either not relevant or partially relevant. Subsequently, we ask the following question: *Do unique processes in coverage and indexing of content change the effectiveness of benchmarking as a method?* In this case, Bilal's study demonstrates the capability of benchmarking to pick out the discrepancy in content across search engines. However, benchmarking alone falls short of providing understanding of the capabilities of a given engine in retrieving results relevant to a given query. In addition, the changing nature of ranking of retrieved output by search engines makes benchmarking, especially when used manually, very challenging.

Issues encountered during the study suggest that the benefits and convenience of benchmarking may not outweigh other factors of note. The jurors confronted problems when finding URLs and activating the links retrieved for the queries in the initial stage of data collection. Comparison of the hits can be tricky when time limits the ability to manually capture retrieved results for each query in each engine within an acceptable time frame to avoid any change that could occur due to updates by the engines. Google is constantly updating the retrieved results and hence, their ranking, due to the fast-paced nature of the web combined with the engine's ranking algorithm and the competitive element of SEO (Search Engine Optimization) functions for online business. Functions such as these are blessing and a curse for relevance judgment. *Is it possible to pin down benchmarks if the ranking algorithm of an engine is a trade secret and the initially retrieved content to be compared is in constant flux?*

Benchmarking as a research method did not covey a deep understanding of the retrieval performance of the search engines (Bilal, 2012). *Is what is* m*ost popular also most pertinent?* One issue that relates to the use of Google as a benchmark is the preconceived notion that Google is suited to all users. While the engine may be the preferred choice among all Internet users including children, the factor of relevancy vis-à-vis a given task or query is not applied by Google or other search engines. Hence, the most popular may not always be most pertinent to the nuances of a given need. This is especially critical in the case of the large-scale engines children often because these young users lack adequate skills in formulating effective search statements and experience difficulty differentiating and deciphering retrieved results (Druin, et al., 2010; Bilal, 2005).

Lack of Transparency: Search engines do not disclose ranking algorithms. This lack of understanding of how rankings are generated creates a problem for

using benchmarking to procure a valid picture of the actual relevance of retrieved results. As Ali & Sufyan (2011) point out, "The ranking algorithms of the search engines cannot be analyzed because these are kept secret due to the competition among the search engines and also, to avoid misuse by the mischievous users" (p. 840). The contextual base for the rationale behind ranking is left open to speculation. As the findings of Bilal (2012) indicate, hit overlap among the five engines varied depending on the topic, and the query language itself. Benchmarking hit overlap by given engines to Google's and Yahoo Kids' top five ranked hits revealed that the engines were more effective on the one-word queries than on the phrase or natural language queries. Conversely, the average precision ratio calculated for the five engines across the queries revealed that they were more effective on the two-word queries and the least effective on the one-word queries. Accordingly, the ranking of retrieved results for the queries was not in concert with intellectual relevance judgment.

Choice of benchmark: Based on the findings of Bilal (2012), utilizing Yahoo Kids as a benchmark in future research is not recommended. According to Bilal, replication of the research design of her study should focus on intellectual relevance judgment using graded relevance rather than on manual benchmarking of retrieved output by search engines.

Intellectual relevance Judgment may help us to ensure that a given search engine is indeed an effective tool for the parameters of evaluation on children's queries. Based on external parameters defined by knowledge of contextual definition, the nature of given tasks, and respective query formulations by the children, the results of Bilal's study (2012) showed that intellectual relevance judgment and calculation of recall and precision ratios of relevant and partially relevant results retrieved by the engines for the queries provided a more accurate picture for understanding the strengths and weaknesses of each search engine vis-à-vis specific types of queries. Google, for example, surpassed the other engines in precision on the natural language queries, whereas Bing outperformed Google and the other engines on the two words queries. Ask Kids produced a much higher precision on all queries than Yahoo Kids, though it is also designed for children (for additional results, see Bilal, 2012).

Van Couvering (2007) considers how "technological schemas (e.g. market criteria, engineering criteria) constrain both the possible interpretations of quality and the mobilization of resources around alternate frameworks by which search engine quality might be assessed" (p. 334). This concept is especially important as it relates to children and their information seeking needs. The search engines function (in the eyes of the administrator/companies) not to inform or relay the most equitable information for instruction. Instead, relevancy is related to other factors including customer satisfaction and efficiency. Subsequently, assessing the user's perspective of relevancy is at the core of judging relevance.

The findings from other studies partially confirm those by Bilal (2012) where *Google is found to come out on top during testing have been used.* For example, Lopez & Ribeiro (2011), in a comparative study of web search engines in

health information retrieval, found that the general web search engines surpassed health-specific engines. "Google is users' preferred search engine and it is also the one with better precision. Differences in this search engine are more expressive at the top of the rankings which means Google's first results page is a good place to start a health search session" (p.889). In the of findings in Bilal's study, Google did indeed come out on top in terms of producing the highest precision ratio but only on natural language queries; it was surpassed by Bing on the two words queries and also on the total precision ratio it produced for relevant and partially relevant hits it retrieved across the queries. Google's strong retrieval capability on natural language queries, a syntax that is most commonly employed by young users, is one of the justifications for its popularity among these users.

4. Reconciled Relevance (RR)

The case for the multi-tiered methodology. Bilal (2012) has shown, use of both benchmarking and intellectual relevance judgment with a graded relevance scale combined with the calculation of recall and precision ratios added a new dimension to contextual understanding of relevance the engines produced on children's queries. Arguably, given the dynamic nature of the web and the dynamic nature of user-driven relevance judgment, is relevance assessment a *benign* notion? If the answer is no, then we need to develop innovative approaches to assess relevance. One of these approaches is Reconciled Relevance (RR) combined with Reconciled Relevance Ranking (RRa).

Reconciled Relevance and Reconciled Relevance Ranking: From the perspective of Bilal (2012), it becomes clear that by using a multi-tiered evaluation approach, researchers should be able to develop a study of greater granularity that lends itself to solving some issues that are particularly to user-driven relevance assessment. The proposed RR approach nuances both ranking of retrieved output and intellectual relevance judgment by involving users, experts, and mediators in the evaluation process. This approach includes:

> a. Users (children) who will evaluate relevancy of retrieved results for given queries and rank them based on their own perspectives of relevance;
> b. Mediators who will evaluate the results and rank them based on their own assessment of relevance. Mediators may include information specialists, teachers, and/or parents. In all cases, a graded relevance scale should be employed for judging relevance;
> c. Researchers or experts who will evaluate the same retrieved results and rank them based on the context and requirements of given tasks for which children formulated queries;

The rankings of retrieved outputs by the role players (i.e., a-c) will be analysed and compared to the rankings of the outputs by given engines in time and space. The average mean value of the rankings by the players (*RRa) Ranking)* will be generated. Similarly, the mean value of average relevance ratings of retrieved outputs by the players will be calculated to produce the *Reconciled Relevance (RR)*

ratings. RRa could be compared to RR to identify *gaps* between rankings by given engines and the players' ratings of relevance.

Future research using this approach will be needed. Though it may be demanding and time consuming, the approach could provide a holistic understanding of the *gaps* in relevance assessments and rankings between and among not only the role players, but also between their judgments and retrieval by given engines. Taking into account the dynamic nature of the web, retrieval by search engines for given queries should be "frozen" in time and space to avoid changes that could occur due to updates.

5. Conclusions

Two issues emerge based on the evaluation methods currently employed in the relevance assessment area of study. First, how do we design research for the aforementioned method using software rather than relying on manual data collection and analysis? Second, how do we evaluate retrieved results by search engines that are relevant to given queries but not relevant to the children's reading levels or other cognitive dimensions? Future work should focus on these issues of which we lack understanding.

References

Ali, R., and Beg, S. (2011). *An overview of Web search evaluation methods*. doi:10.1016/j.compeleceng.2011.10.005.

Barry, C. (1994). User-defined relevance criteria: An exploratory study. *Journal of the American Society for Information Science*, 45(3), 149-159.

Bilal, D. (2012, in press). Ranking, relevance judgment, and precision of information retrieval on children's queries: evaluation of Google, Yahoo, Bing, Yahoo Kids, and Ask Kids. *Journal of the American Society for Information Science and Technology.*

Bilal, D., (2004). Research on children's information seeking on the web. In: *Youth Information-Seeking: Theories, Models, and Approaches*, Mary Kay Chelton and Colleen Cool (Eds.). Lanham, MD: Scarecrow Press (pp. 271-291).

Bilal, D., and Ellis, R. (2011). Evaluating leading Web search engines on children's queries. (2011). *HCI 2011 International, Proceedings of the 14th International Conference on Human-Computer Interaction, Part IV, Lecture Notes in Computer Science 6764*, July 9-14, Orlando, FL (pp. 549-558).

Druin, A., Foss, E., Hutchinson, H., Golub, E., and Hatley, L. (2010). Children's rolesusing keyword search interfaces at home. *CHI'10, Proceedings of the 28th International Conference on Human Factors in Computing Systems*. New York, NY, (pp. 413-422).

Harter, S. (1992). Psychological relevance and information science. *Journal of the American Society for Information Science*, 53(4), 602–615.

Lewandowski, D. (2011). The retrieval effectiveness of search engines on navigational queries. *Aslib Proceedings*, 63, 354-363.

Lewandowski, D. (2008). The retrieval effectiveness of Web search engines: considering results descriptions. *Journal of Documentation* 64(6), 915-937.

Lopes, C., and Ribeiro, C. (2011). Comparative evaluation of Web search engines in health information retrieval. *Online Information Review*, 35, 869-892. http://dx.doi.org/10.1108/14684521111193175.

Mizzaro, S. (1997). Relevance: The whole history. *Journal of the American Society for Information Science*, 48(9), 810-832.

Saracevic, T. (1975). Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, 26(6), 321–343.

Saracevic, T. (2007a). Relevance: A review of the literature and a framework for thinking on the notion in information science, Part II: Nature and manifestations of relevance. *Journal of the American Society for Information Science and Technology*, 8(13): 1915–1933.

Saracevic, T. (2007b). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance. *Journal of the American Society for Information Science and Technology*, 58 (13), 2126–2144.

Schamber, L. (1994). Relevance and information behavior. *Annual Review of Information Science and Technology*, 29, 3–48.

Schamber, L., Eisenberg, M. and Nilan, M. (1990). A re-examination of relevance: Toward a dynamic, situational definition. *Information Processing & Management*, 6(6), 755–776.

Spink, A., and Greisdorf, H. (2001), Regions and levels: Measuring and mapping users' relevance judgments. *Journal of the American Society for Information Science and Technology*, 52: 161–173.

Spink, A., and Jansen, B. (2004). *Web Search: Public Searching of the Web*. Boston, MA: Kluwer Academic Publishers.

Spink, A., Jansen, B., Blakely, C., and Koshman, S. (2006). A study of overlap and uniqueness among major Web search engines. *Information Processing and Management* 42, 1379-1391.

Tamine-Lechani, L., Boughanem, M., and Daoud, M. (2010). Evaluation of contextual information retrieval effectiveness: overview of issues and research. *Knowledge Information Systems*, Vol. 24, 1–34. DOI 10.1007/s10115-009- 0231-1.

Thelwall, M. (2008). Quantitative comparisons of search engine results. *Journal of the American Society for Information Science and Technology*, 59(11), 1702-1710.

Van Couvering, E. (2007) Is Relevance Relevant? Market, Science, and War: Discourses of Search Engine Quality, *Journal of Computer-Mediated Communication*, Vol. 12, Issue 3, Article first published online: 6 JUN 2007.

Vaughan, L. (2004). New measurements for search engine evaluation proposed and tested. *Information Processing and Management* 40, 677–691.

Wang, P., and Soergel, D. (1998). A cognitive model of document use during a research project. Study I. Document selection. *Journal of the American Society forInformation Science*, 49(2), 115–133.