Evaluating Leading Web Search Engines on Children's Queries

Dania Bilal and Rebekah Ellis

School of Information Sciences
University of Tennessee, 1345 Communications 451, Knoxville, TN 37996
{dania,rellis17}@utk.edu

Abstract. This study compared retrieved results, relevance ranking, and overlap across Google, Yahoo!, Bing, Yahoo Kids!, and Ask Kids on 15 queries constructed by middle school children. Queries included one word, two words, and multiple words/phrases/natural language, and the results were benchmarked against Google and Yahoo Kids! top 5 and top 10 retrieved results using a new relevance ranking metric. Yahoo! and Bing yielded similar results on all queries, but their relevance ranking differed on one-word queries. Ask Kids outperformed Yahoo Kids! on all queries, and a modest percentage of results had the same relevance ranking as Google. Yahoo Kids! and Ask Kids returned unique results that were not retrieved by the other three engines on the first results page. Yahoo! and Bing produced the highest percentage in overlap with Google followed by Ask Kids. Implications are made for children and mediators concerning the use of search engines on children's queries.

Keywords: Children, queries, query construction, web search engines, evaluation, information retrieval, relevance ranking, ranking comparison, overlap, unique results, Google, Yahoo!, Bing, Yahoo Kids!, Ask Kids.

1 Introduction

Children use the Web on a daily basis to support their information needs, find materials for research projects, and communicate with others using social media. A recent study shows that the time children aged 2-11 spend online increased 63% between 2004 and 2009 [1]. Children's use of Web search engines accounts for a portion of this time increase. Today's children rank Google as their top choice for finding information, followed by Yahoo!, Bing, and Ask.com [2]. Previous studies of Google, Yahoo!, Live Search/MSN, and AskJeeves revealed inconsistencies in retrieved results, overlap [3][4], and a high percentage (84.9%) of results that were unique to one specific engine [5]. To a child, this means that certain relevant results to a query will most likely be missed if the child uses only one search engine. As children are increasingly using leading search engines designed for the general public (Google, Yahoo!, Bing) rather than engines that are specifically designed for their age levels (i.e., Yahoo Kids! and Ask Kids), we need to develop an understanding of how these search engines differ in retrieval and relevance ranking of results for children's

queries. Such an understanding is significant because these engines are not equal; they have their own search capabilities, ranking algorithms, limitations, and complexity. As Thelwall notes, "although search engines are entirely logical because they are computer programs, their complexity means that the results they present often have inconsistencies and unexpected variability" [3]. Results gained from this study should be useful for children and their mediators (parents, teachers, practitioners) in determining which search engines to use for child-driven content on specific types of queries (one word, two words, and multiple word/phrases/natural language). In addition, the results could help generate improvements in the interface design of these search engines in support of children's effective Web interaction.

2   Related Studies

2.1   Evaluation of Search Engines Designed for the General Public

In a study of 1,587 single-word queries, Thelwall [3] compared Google, Yahoo!, and Live Search (previously MSN) for hit counts, consistencies, and matching URLs. Google and Yahoo! provided five to six times more hits than Live Search, and Yahoo! returned slightly more matching URLs than Google. The engines were more consistent in the top-level domains represented in the URLs, although Yahoo! returned the most. For hit counts, Google outperformed Yahoo! and Live Search, but for the range of domains and sites represented in the results, Yahoo! surpassed the other two engines. Lewandowski [4] compared the retrieval performance of five Web search engines (i.e., Yahoo!, Ask, Google, MSN, and Seekport) on single-word queries. He found that Google and Yahoo! outperformed the others, but neither was found to be superior overall to the other, although Google held a slight advantage when only the top three results were considered. In an earlier study, Spink, Jansen, Blakely, and Koshman [5] compared retrieved results from Google, Yahoo!, MSN, and Ask Jeeves. They found that 84.9% of the retrieved results from the first results page were unique to 1 of the 4 search engines and that the percentage of unique results was higher than the percentage of results that overlapped across the engines.

2.2   Evaluation of Search Engines Designed for Children

The earliest study that evaluated the retrieval performance of search engines designed for children was conducted by Bilal [6]. She compared Yahooligans!, Ask Jeeves Kids, and Super Snooper (now a defunct engine) on a set of children's queries (single words, multiple words, and natural language) using the following four criteria: retrieval output, relevance, overlap, and redundancy. Yahooligans! was effective on queries with single terms and ineffective on queries with multiple terms or natural language. Ask Jeeves Kids retrieved results from Yahooligans! for queries using natural language that the latter failed to retrieve. There was an overlap in returned results between Yahooligans! and Ask Jeeves Kids.

Taking a different approach, Large, Beheshti, and Rahman [7] evaluated Yahooligans!, Ask Jeeves Kids, KidsClick, and Lycos Zone from the perspective of middle school children. The authors generated four child-centered design criteria:

goals, visual design, information architecture, and personalization that they considered essential for implementation in the design of children's search engines. In addition, they found that the four engines were more effective on single-word searches than on queries with phrases or natural language.

Bilal [8][9][10] evaluated Yahooligans! based on middle school children's information-seeking behavior on fact-based, research-based, and fully self-generated tasks. Overall, children were more successful on the fully self-generated task than on the other two tasks. Additionally, children clicked more on results displayed near the top of the first results page than in the middle or bottom of the page. Children's use of single-word queries was more effective than queries with multiple words or natural language. Most of the breakdowns children experienced resulted from the inadequate interface design of Yahooligans!

In a recent study, Druin, et al. [11] explored the keyword searching of children ages 7, 9, and 11 in using Google on four assigned tasks. A small group of children demonstrated strong search expertise and very few children were successful in formulating complex queries. Seven search roles were identified based on children's information seeking. Children's success varied depending on search role. For example, p*ower searchers* were the most successful in accessing information and in assessing whether they had found what they needed.

In summary, previous studies of search engines, namely Google, Yahoo!, MSN/Live Search (now Bing), and Ask Jeeves (now Ask) uncovered inconsistencies in the range of URLs, domain coverage, relevance ranking, and overlap across the engines. Studies of children's information behavior on the Web revealed that they were more successful in finding information from simple queries than for complex queries. One study, evaluated the retrieval performance of three child-driven search engines. However, this study is dated in the late 1990s. Today, children prefer to use Google and like engines over Yahoo Kids! and Ask Kids. Nevertheless, there has been little comparison between these two search engine groups on children's queries. This study begins to examine differences and similarities between these engine groups in retrieving and ranking results on the first results page for queries with one word, two words, and multiple words/phrases/or natural language.

3   Research Questions

This study addressed these research questions:

1. Using Yahoo Kids! as the benchmark, what are the differences in relevance ranking of the top five retrieved results on the first results page across Google, Yahoo!, Bing, and Ask Kids on children's queries (one word, two words, and multiple words/phrases/natural language)?
2. Using Google as the benchmark, what are the differences in relevance ranking of top five retrieved results on the first results page across Yahoo!, Bing, Yahoo Kids!, and Ask Kids on children's queries (one word, two words, and multiple words/phrases/natural language)?
3. Using Google as the benchmark, what is the percentage of overlap in retrieved results on the first results page for the top 5 and top 10 results across Yahoo!, Bing,

Yahoo Kids!, and Ask Kids on children's queries (one word, two words, and multiple words/phrases/natural language)?

4   Method

This study employed the quantitative method for comparing children's queries across the five search engines (Google, Yahoo, Bing, Yahoo Kids!, and Ask Kids).

Query Set. We examined the published literature from 1989 to present on children's interaction with digital tools (e.g., OPACs, CD-ROMs, and search engines). We identified 130 tasks that were assigned to children and/or self-selected by them, and we focused on studies of children in grade levels 5-9. For these tasks, we examined how children queried a given digital tool (i.e., search statements using one word, two words, multiple words/phrases/natural language) so that we could query each of the five search engines using these statements.

Query Sample. We selected five queries with single words, five with two words, and five with multiple words/phrases/natural language (Table 1) to use in this study.

Table 1. Queries submitted to search engines

One-word queries	Two-word queries	Multiple words/phrases/ natural language queries
Ozone Diabetes The Simpsons Spiders Vegetarians	Spanish armada Speed skating Endangered animals Ancient numerals Olympic hockey	Why dolphins migrate? Women in space Clock using sun and stick What are the three most common crimes in California? Women at war

Data. We submitted each query to each search engine, retrieved the results on the first results page, printed out the results, and coded and analyzed them manually. Each set of queries was submitted within a minute of each other, and all 15 queries occurred on January 31, 2011, between 10:00 and 11:30 p.m. to avoid any occurrence of possible changes in retrieved results due to search engine updates.

We used Yahoo Kids! and Google as benchmarks for comparing the retrieved results, relevance ranking of the results, and overlap across the engines. Yahoo Kids! was selected as a benchmark because it is specifically designed for children ages 7-12, and Google was selected as a benchmark because it is the most used by children.

The first results page on a given query in a given search engine was printed out and the top five retrieved results were considered for comparison against the benchmark. We coded the top 5 ranked results retrieved by Yahoo Kids! and Google using the codes, 1-5 (1=first ranked result; 5=fifth ranked result). Our submission of the 15 queries to the 5 search engines resulted in a total of 75 printouts of retrieved results (25 printouts for queries with 1 word, 25 for queries with 2 words, and 25 for queries with multiple words/phrases/natural language). We excluded advertisements from the

results each search engine retrieved. We employed a *relevance ranking metric* with four values: 0=same ranking as the rank of the benchmark; (+n) (e.g., +1,+2,+3) means that the retrieved result by a given search engine is ranked 1, 2, or 3 positions (as applicable) below the rank of a retrieved result by the benchmark; (-n) (e.g., -1, -2,-3) means that the retrieved result is ranked 1, 2, or 3 positions (as applicable) above the retrieved result by the benchmark; and NR means that no results were retrieved for a given query by a given search engine on the first results page that matched with the top 5 or top 10 results retrieved by a benchmark. We coded the 75 printouts of retrieved results manually using the values explained above and shown in Table 2. We compared the ranking position of the results retrieved for a given query by a given search engine against the top five ranked results retrieved by the benchmarks, one time using Google as the benchmark and another time using Yahoo Kids! as the benchmark.

Table 2. Values for ranking of retrieved results across search engines

---

0, where a result has the same ranking position as the ranking by the benchmark.

+n, where n=ranking position below the benchmark rank. For example, a result with a +1 means it was ranked one position below the ranking by the benchmark.

-n, where n=ranking position above the benchmark rank. For example, a result with a -1 means it was ranked one position above the ranking by the benchmark.

NR = Not retrieved.

---

5   Results

In reporting the results, we use the code (YK-R) to indicate Yahoo Kids! relevance ranking and the code (G-R) to denote Google relevance ranking of a given retrieved result.

Differences in Relevance Ranking of Retrieved Results across Search Engines – Yahoo Kids as Benchmark on One-Word Queries. No results were retrieved by Google, Yahoo!, and Bing that matched with the top five results retrieved on one-word queries by Yahoo Kids! Ask Kids retrieved one matching result for the query *diabetes*, but it did not have the same ranking as Yahoo Kids (+3 vs. YK-R=3, respectively).

Differences in Relevance Ranking of Retrieved Results across Search Engines – Google as Benchmark on One-Word Queries. On the query *ozone*, Yahoo Kids! did not return any results on its first results page that matched with the top five results retrieved by Google. Two results retrieved by Yahoo! and two retrieved by Bing were identical and had the same ranking as Google (0 and G-R=1, respectively) and

(0 and G-R=3, respectively). Ask Kids retrieved two matched results, but their ranking was different from Google (-2, +5 vs. G-R=3 and G-R=4, respectively). For the query *diabetes*, Yahoo! retrieved three results that matched with Google, but only two had the same ranking as Google (0, 0 and G-R=2, G-R=3, respectively). Ask Kids retrieved only one result that matched with Google, but it did not have the same ranking as Google (-1 vs. G-R=2, respectively). For the query *spiders*, Yahoo! and Bing retrieved three results each that matched the results Google retrieved, but only one result had the same ranking as Google (0 and G-R=1, respectively). For the query *vegetarians*, Yahoo! and Bing retrieved four results each that matched with Google, but only one result that Yahoo! retrieved had the same ranking as Google (0 and G-R=1, respectively). Ask Kids retrieved one matched result, but it did not have the same ranking as Google (-3 vs. G-R=4, respectively). For the query *the Simpsons*, Yahoo! retrieved four results that matched with and had the same ranking as Google. Bing retrieved four results of which only two had the same ranking as Google.

Differences in Relevance Ranking of Retrieved Results across Search Engines – Yahoo Kids! as Benchmark on Two-Word Queries. On the query *Spanish armada*, Google retrieved four results that matched with Yahoo Kids!, but they did not have the same ranking as Yahoo Kids! (+6, +2, +5,+1 vs. YK-R=1, YK-R=2, YK-R=3, YK-R=4, respectively). Yahoo! and Bing retrieved one identical result on this query that matched with Yahoo Kids!, but it did not have the same ranking as Yahoo Kids! (+6 vs. YK-R=2). Ask Kids also retrieved one matched result that did not have the same ranking as Yahoo Kids! (+1 vs. YK-R=1, respectively). For the query *speed skating*, only Ask Kids retrieved two matched results but they did not have the same ranking as Yahoo Kids! (+3, +2 vs. YK-R=1, YK-R=2, respectively). None of the four engines returned results that matched with Yahoo Kids! for the query *ancient numerals* or *Olympic hockey*.

Differences in Relevance Ranking of Retrieved Results across Search Engines – Google as Benchmark on Two-Word Queries. Yahoo! and Bing retrieved four identical results each for the query Spanish armada, that matched with Google of which one result had the same relevance ranking as Google (0 and G-R=1). Yahoo Kids! retrieved two results that matched with Google, but none of them had the same ranking as Google.

For the query *endangered animals*, Yahoo! and Bing retrieved one result each that had the same ranking as Google (0 and G-R=2). Ask Kids retrieved three matched results, but none of them had the same ranking as Google. For the query *speed skating*, Yahoo! and Bing retrieved three matched results, but none of them had the same ranking as Google. Yahoo Kids! did not retrieve any results that matched with Google. Ask Kids retrieved one result that matched with and had the same ranking as Google (0 vs. G-R=1).

For the query *ancient numerals*, Yahoo! and Bing retrieved three identical results each that matched with Google, but none of them had the same ranking as Google (-2, +6, +4 vs. G-R=3, G-R=4, G-R=5, respectively). Ask Kids retrieved only one matched result, but it did not have the same ranking as Google (+6 vs. G-R=5). Yahoo Kids! retrieved no results from its first results page that matched with Google.

For the query *Olympic hockey*, Yahoo! and Bing retrieved three identical results each of which one result had the same ranking as Google (0 vs. G-R=1). Neither Yahoo Kids! nor Ask Kids retrieved results that matched with Google.

Differences in Relevance Ranking of Retrieved Results across Search Engines – Yahoo Kids! as Benchmark on Multiple Words/Phrases/Natural Language Queries. No engine retrieved results on its first results page that matched with Yahoo Kids!, with the exception of Google that returned one result for *the query women in space*, but it did not have the same ranking as Yahoo Kids! (+7 vs. YK-R=1, respectively).

Differences in Relevance Ranking of Retrieved Results across Search Engines – Google as Benchmark on Multiple Words/Phrases/Natural Language Queries. For *the query women in space*, Yahoo! and Bing retrieved four identical results each that matched with Google, but they did not have the same ranking as Google. Ask Kids retrieved three results that matched with Google, but only one result had the same ranking as Google (0 and G-R=4; +1, -2 vs. G-R=5, G-R=1, respectively). Yahoo Kids! did not retrieve any result that matched with Google.

For another query, *women at war*, Yahoo! and Bing retrieved one identical result that had the same ranking as Google (0 and G-R=1) and one result that matched with but did not have the same ranking as Google (+1 vs. G-R=3). Bing retrieved one result that had the same ranking as Google (0 and G-R=3). Yahoo Kids! retrieved one result that matched with Google, but it did not have the same ranking as Google (-2 vs. G-R=4) and Ask Kids retrieved one matched result that had the same ranking as Google (0 and G-R=3). For the query *clock using sun and stick*, Yahoo! and Bing retrieved one identical result but it did not have the same ranking as Google (-1 vs. G-R=3).

For the query *what are the three most common crimes in California*, no engine retrieved results that matched with Google. As to the query *why do dolphins migrate*, Yahoo! and Bing retrieved one identical result that had the same ranking as Google (0 vs. G-R=3), and three additional results that did not have the same ranking as Google (+1, +4, -3 vs. G-R=1, G-R=2, G-R=4, respectively). Both Yahoo Kids! and Ask Kids did not retrieve results on the first results page that matched with Google on these two queries.

5.1 Overlap in Results Across the Search Engines

We assessed the overlap in the top 5 and top 10 results retrieved by each search engine against Google's ranking of the top 5 retrieved results for a given query. To calculate the overlap in the top 10 results retrieved, we counted the results that matched with Google on each query by each search engine and totaled the matched results across all 5 queries by query type (one word, two words, and multiple words/phrases/natural language). We divided the total number of matched results for each query type by 50, where 50 is the maximum number of first results pages a given search engine would retrieve on the 5 queries by query type (10 results per page for a given query multiplied by 5 result pages).

Overlap in the top five and top ten ranked results.
One-word queries: Both Yahoo! and Bing had the same percentage in overlap with Google (36%) in the top 5 ranked results, followed by Ask Kids (12%). There was no overlap between Yahoo Kids! and Google on these queries. Yahoo! produced the highest percentage in overlap (84%) in the top 10 retrieved results benchmarked against Google, followed by Bing (36%) and Ask Kids (12%). Yahoo Kids! did not retrieve results that matched with Google, yielding no overlap (Table 3).
Two-word queries: Both Yahoo! and Bing resulted in an equal percentage in overlap with Google, 34% in the top 5 and 38% in the top 10 results, followed by Ask Kids, 22% in the top 5 and 32% in the top 10. Yahoo Kids! produced the lowest percentage in overlap with Google in both the top 5 and the top 10 (0.4%) (Table 3).
Multiple words/phrases/natural language queries: Both Yahoo! and Bing yielded the same percentage in overlap with Google, 22% in the top 5 and 40% in the top 10 results. Ask Kids resulted in 10% overlap with Google in the top 5 and 16% in the top 10. Yahoo Kids! showed (0.2%) overlap with Google in both the top 5 and top 10 results. Here too, Ask Kids surpassed Yahoo Kids! in finding matches that overlapped with Google.

Table 3. Overlap with Google across the search engines on children's queries

Engine	One-word queries Percent* in overlap		Two-word queries Percent* in Overlap		Multiple words/ phrases/natural language queries Percent* in overlap	
	Top 5	Top 10	Top 5	Top 10	Top 5	Top 10
Yahoo!	36%	84%	34%	38%	22%	40%
Bing	36%	36%	34%	38%	22%	40%
Yahoo Kids!	-0-	-0-	0.4%	0.4%	0.2%	0.2%
Ask Kids	12%	12%	22%	32%	10%	16%

*Percentage of overlap is based on the first results page retrieved by each engine on the five queries submitted.

6   Discussion and Conclusion

One of the key findings of this study is that using Google as a benchmark against which to compare the retrieved results and their relevance ranking across the four engines was more effective than using Yahoo Kids! as a benchmark. In fact, Yahoo Kids! is *unique* in its content, especially since it is targeted to children ages 7-12, as described in its online help. Conversely, Ask Kids which is also designed for children surpassed Yahoo Kids! in retrieving a number of results that not only matched with Google, but also had the same relevance ranking as Google. The difference in retrieval between Yahoo Kids! and Ask Kids may be attributed to the level of indexing the search engine employed and to the fact that unlike Yahoo!, Yahoo Kids!

is more of a directory than a search engine, and, therefore, its indexing methodology is less comprehensive than that employed by Ask Kids.

The overlap with Google across the four search engines varied by query type. Yahoo! and Bing retrieved identical results on most queries and had the highest percentage of overlap with Google. However, the relevance ranking of results between these two engines was different, especially on one-word queries. It appears that these engines have a common indexing methodology but use different ranking techniques.

Another key finding is that the percentage of results that had the same relevance ranking as Google across the four engines was the highest on one-word queries and the lowest on multiple word/phrases/natural language queries. This finding is congruent with the results of previous studies on search engines designed for the general public [3][4][5] and those designed for children [8][9][10], which revealed that the engines were more effective on single-word queries than on two-words or natural language queries. This is an area of research that needs further investigation to uncover these underlying retrieval problems across these engines.

Although it is designed for children, Ask Kids! outperformed Yahoo Kids! in retrieving results that overlapped with Google on all types of queries. This finding may be attributed to the customized type of documents indexed by Yahoo Kids! that tend to be highly filtered and judged based on the reading abilities of elementary and middle school children. This finding could mean that children and their mediators may want to use Ask Kids over Yahoo Kids! to find child-centric information for their queries without sifting through the many results retrieved by Google.

Another finding is that both Yahoo Kids! and Ask Kids retrieved *unique* results that were not shared by Google, Yahoo!, or Bing on the first results page for the queries. These results are child driven, and, therefore, children and their mediators may miss exposure to these results that could be relevant to the queries by focusing only on Google or like engines. Due to the fact that we focused on system-driven relevance judgment and ranking of retrieved results across the five engines, we did not judge relevance of the *unique* results that both Yahoo Kids! and Ask Kids retrieved on the queries. In addition, we did not judge relevance of the results retrieved by Google, Yahoo!, and Bing against children's reading abilities or reading comprehension.

This study provided an understanding of differences and similarities between search engines designed for adult users and engines designed for children. It uncovered strengths and weaknesses of retrieved results, relevance ranking, and overlap across the engines on different types of children's queries. Much remains to be learned about these search engines to optimize their use by children. Some questions to address in future studies are: 1. What are the search capabilities and retrieval performance of Yahoo Kids! and Ask Kids on different types of children's queries? 2. Of the five search engines used in this study, what engine is the "best" on children's queries in specific subject domains (e.g., science, social science), and 3. To what extent does the system-driven relevance ranking of retrieved results across the five engines vary from the user-based relevance ranking of the same results on different types of queries?

References

1. Nielsen, J.: Children's Websites: Usability issues in designing for kids, `http://www.useit.com/alertbox/childen.html`
2. Olsen, S.: Google seeks to help children search better, `http://www.nytimes.com/2009/12/26/technology/internet/26kidsearch.html`
3. Thelwall, M.: Quantitative comparisons of search engine results. J. Am. Soc. Inform. Sci. 59, 1702–1710 (2008)
4. Lewandowski, D.: The retrieval effectiveness of Web search engines: Considering results descriptions. J. Doc. 64, 915–937 (2008)
5. Spink, A., Jansen, B.J., Blakely, C., Koshman, S.: A study of overlap and uniqueness among major Web search engines. Inform. Process. Manag. 42, 1379–1391 (2006)
6. Bilal, D.: Web search engines for children: A comparative study and performance evaluation of Yahooligans!, Ask Jeeves Kids, and Super Snooper. In: Kids, J., Snooper, S. (eds.) 62nd Am. Soc. Inform. Sci. Meeting, pp. 70–82. Information Today, Inc., Medford (1999)
7. Large, A., Beheshti, J., Rahman, T.: Design criteria for children's Web portals: The users speak out. J. Am. Soc. Inform. Sci. Tech. 53, 79–94 (2002)
8. Bilal, D.: Children's use of the Yahooligans! Web search engine. I. Cognitive, physical, and affective behaviors on fact-based tasks. J. Am. Soc. Inform. Sci. 51, 646–665 (2000)
9. Bilal, D.: Children's use of the Yahooligans! Web search engine. III. Cognitive and physical behaviors on fully self-generated tasks. J. Am. Soc. Inform. Sci. Tech. 53, 1170–1183 (2002)
10. Bilal, D.: Children's use of the Yahooligans! Web search engine. II. Cognitive and physical behaviors on research tasks. J. Am. Soc. for Inform. Sci. Tech. 52, 118–137 (2001)
11. Druin, A., et al.: How children search the internet with keyword interfaces. In: 8th International Conference on Interaction Design and Children, Como, Italy, pp. 89–96 (2009)