Query Log Analysis in the Context of Information Retrieval for Children

Sergio Duarte Torres
University of Twente
The Netherlands
duartes@cs.utwente.nl

Djoerd Hiemstra
University of Twente
The Netherlands
hiemstra@cs.utwente.nl

Pavel Serdyukov
Delft University
The Netherlands
p.serdyukov@tudelft.nl

ABSTRACT

In this paper we analyze queries and sessions intended to satisfy children's information needs using a large-scale query log. The aim of this analysis is twofold: i) To identify differences between such queries and sessions, and general queries and sessions; ii) To enhance the query log by including annotations of queries, sessions, and actions for future research on information retrieval for children. We found statistically significant differences between the set of general purpose and queries seeking for content intended for children. We show that our findings are consistent with previous studies on the physical behavior of children using Web search engines.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Query formulation

General Terms

Experimentation, Measurement

Keywords

query log analysis, query intent, query representation

1. INTRODUCTION

The Internet today is widely used by children for information and communication purposes. Unfortunately, most of the current Information Retrieval (IR) systems are designed for adults who have shown to have different search approaches and cognitive skills than children [1]. Thus, there is an increasing need for research aimed at understanding children's search characteristics and to provide them suitable IR systems. Query logs are valuable resources to explore the search behavior of users and have been found highly useful to improve their search experience. In this study, we employed the AOL query log and the DMOZ *kids&teens* directory to identify differences in the queries and sessions employed by users to retrieve children and general-purpose information. This comparison also allows us to confirm on a large-scale environment previous findings in children physical search on Web search engines [1]. It is important to mention that we are aware of the controversy of the usage of this query log in the research community. For this reason we clarify that no actual user identification is carried out in any of the experiments performed in our study.

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.
ACM 978-1-60558-896-4/10/07.

Table 1: Most frequent children queries

1) nickjr.com	4) coloring pages	7) the wiggles
2) elmo	5) postopia	8) starfall.com
3) nick jr	6) candystand	9) dora the explorer

2. CHILDREN QUERIES & SESSIONS

The AOL query log [5] contains approximately 36 million entries. We represent this query log as the set:

$$Log = \langle \langle u_i, q_i, t_i, d_i, r_i \rangle \mid 1 \leq i \leq n \rangle \qquad (1)$$

where u, q, t, d, r refers to the user ID, query, time of submission, domain clicked and its rank position respectively, n defines the size of the query log.

The identification of queries employed to retrieve content for children was performed by matching the DMOZ *kids&teens* entries tagged for kids (which point to appropriate Web content for children up to 12 years old) with the clicked domains of the query log. Given that the query log does not include the entire URL visited, matches were restricted to the cases in which only the domain is listed as DMOZ entry. The set of these queries is represented by Equation 2.

$$Kids = \langle \langle u_i, q_i, t_i, d_i, r_i \rangle \mid d_i \in DMOZ_{kids} \rangle \qquad (2)$$

Sessions were constructed by grouping contiguous queries from the same user that are submitted with a time difference smaller than θ. A formal definition of session is shown in Equation 3.

$$S = \langle \langle q_{i_1}, u_{i_1}, t_{i_1} \rangle, ..., \langle q_{i_k}, u_{i_k}, t_{i_k} \rangle \rangle \qquad (3)$$

where $u_{i_1} = ... = u_{i_k}$, $t_{i_1} \leq ... \leq t_{i_k}$ and $t_{i_{j+1}} - t_{i_j} \leq t_\theta$ for all $j = 1, 2..., k-1$ The parameter t_θ was set to 30 minutes because it is the most common value employed in the literature [4]. This time window is also suitable for sessions expressing children's information needs since it has been shown that on average children spend up to 16 minutes to fulfill an information need [2]. We define a *children session* as a session that contains at least one children query entry.

The data collected contains 485,561 query entries (10,252 unique queries) and 21,009 sessions. The most frequent queries are shown in Table 1. Note that although it is not possible to establish if these queries were submitted by children, we are still able to study the characteristics of the queries and sessions for which the underlying information need is related to children content. We consider that this assumption is reasonable since the DMOZ kids directory employed to identify these queries is realistic and of high quality.

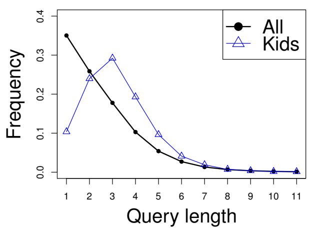

Figure 1: Length frequency

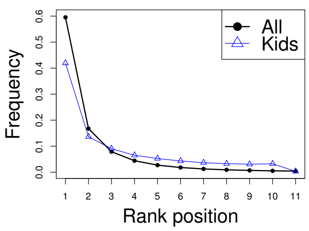

Figure 2: Rank frequency

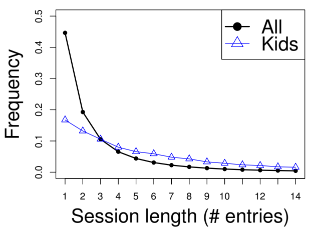

Figure 3: Session length

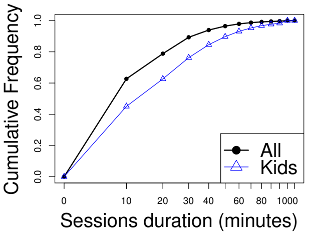

Figure 4: Session duration

3. FINDINGS SUMMARY

The analysis was carried out at the query and session level. Query length and domain rank data of the clicked domains were considered for the former and session length, duration and query reformulations for the latter. All the results found for the kid queries are significantly different from the queries of the whole log using the Wilcoxon signed-rank test and the t-test at the 95% confidence level.

3.1 Query length analysis

Query length is an indicator of the complexity of the query and the difficulty of the user to express information needs using keywords. Queries that were used to retrieve information for children were on average (3.23 words per query) significantly longer than the average of the queries in the whole query log (2.5 words per query) as illustrated Figure 1. Interestingly, this finding is in-line with Druin et al.[3] studies in which children aged 8 to 12 are found to formulate longer queries. A more frequent formulation of queries using natural language constructs have also been found in children search behavior [1]. We verified this observation by extracting adjectival and verbal phrases from the queries. We found that 65% of the children retrieval queries contain either of these phrases compared to 56% of the queries in the whole query log. We also found a greater use of questions in the former queries (3.92% vs 2.71%).

3.2 Click analysis

The rank distribution of clicks was collected to compare the retrieval performance between children and general purpose queries. Queries in which highly ranked domains are more often clicked indicate that information needs are more efficiently satisfied by the IR system. Figure 2 shows the rank frequency distribution of clicks. This figure demonstrates that on average the retrieval performance of the children queries is poorer than the queries used to retrieve general-purpose content since clicks on lower ranked results are more frequent in the children query set (5.77 vs 3.58 average rank). We also observed an important drop for the children queries on the clicks ranked as 10. This can be explained by the fact that children refuse to go beyond the first page more often than older users [3].

3.3 Sessions length analysis

Figure 3 shows that sessions used to retrieve information for children are longer than general-purpose sessions. The longer average length found for the children sessions (8.76 vs 2.8 query entries per session) suggests that these users were not certain of the relevance of the information found since they had to perform more queries and explore more documents. This result is consistent with Bilal's findings [2] in which children showed *nonlinear navigation style* when solving search tasks. This search style is characterized by the exploration of several choices before a final relevance judgment is made [2]. This result can also indicate that the documents retrieved by the search engine are not sufficient to satisfy the user's information need.

3.4 Session duration analysis

The duration distribution in minutes of children and general-purpose sessions are shown in Figure 4. This figure shows that users require more time to explore and complete information needs associated to children content, which suggests more difficulty to solve the information tasks associated to these sessions. This results is consistent with the greater amount of queries and clicks on lower ranked pages found in the children queries. Interestingly, the mean duration found for the children's sessions (20.38 minutes) is in line with the average time reported by Bilal et al.[2] for children that were unsuccessful completing fact-based information tasks (19.69 minutes).

4. ACKNOWLEDGEMENTS

This research is funded in by PuppyIR, a project in the European Union's 7th Framework ICT Programme under grant agreement no. 231507.

5. CONCLUSIONS AND FUTURE WORK

This work represents a valuable methodology to study the search behavior of user's pursuing children information needs given that our findings are in-line with previous studies of children information-seeking behavior on web search engines [1, 2, 3]. This method has also been proved adequate to corroborate on a large-scale setting the results drawn by case-studies on small group of users and can be applied to characterize the search behavior of different user groups. We have enriched the AOL query log by annotating the children queries and session information based on this study and as future work we will employ this resource to study query assistance methods to improve the search experience of children.

6. REFERENCES

[1] D. Bilal. Children's use of the yahooligans! web search engine: Ii. cognitive and physical behaviors on research tasks. *J. Am. Soc. Inf. Sci. Technol.*, 52(2):118–136, 2001.

[2] D. Bilal. Children's use of the yahooligans! web search engine. iii. cognitive and physical behaviors on fully self-generated search tasks. *J. Am. Soc. Inf. Sci. Technol.*, 53(13):1170–1183, 2002.

[3] A. Druin, E. Foss, L. Hatley, E. Golub, M. L. Guha, J. Fails, and H. Hutchinson. How children search the internet with keyword interfaces. In *IDC '09: Proceedings of the 8th International Conference on Interaction Design and Children*, pages 89–96, New York, NY, USA, 2009. ACM.

[4] R. Jones and K. L. Klinkner. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *CIKM '08*, pages 699–708. ACM, 2008.

[5] G. Pass, A. Chowdhury, and C. Torgeson. A picture of search. In *InfoScale '06: Proceedings of the 1st international conference on Scalable information systems*, page 1, New York, NY, USA, 2006. ACM.