What and How Children Search on the Web

Sergio Duarte Torres
University of Twente
Drienerlolaan 5
Enschede, The Netherlands
duartes@cs.utwente.nl

Ingmar Weber
Yahoo! Research Barcelona
Avda. Diagonal 177
E-08018 Barcelona, Spain
ingmar@yahoo-inc.com

ABSTRACT

The Internet has become an important part of the daily life of children as a source of information and leisure activities. Nonetheless, given that most of the content available on the web is aimed at the general public, children are constantly exposed to inappropriate content, either because the language goes beyond their reading skills, their attention span differs from grown-ups or simple because the content is not targeted at children as is the case of ads and adult content. In this work we employed a large query log sample from a commercial web search engine to identify the struggles and search behavior of children of the age of 6 to young adults of the age of 18. Concretely we hypothesized that the large and complex volume of information to which children are exposed leads to ill-defined searches and to disorientation during the search process. For this purpose, we quantified their search difficulties based on query metrics (e.g. fraction of queries posed in natural language), session metrics (e.g. fraction of abandoned sessions) and click activity (e.g. fraction of ad clicks). We also used the search logs to retrace stages of child development. Concretely we looked for changes in the user interests (e.g. distribution of topics searched), language development (e.g. readability of the content accessed) and cognitive development (e.g. sentiment expressed in the queries) among children and adults. We observed that these metrics clearly demonstrate an increased level of confusion and unsuccessful search sessions among children. We also found a clear relation between the reading level of the clicked pages and the demographics characteristics of the users such as age and average educational attainment of the zone in which the user is located.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Query formulation, Search process; H.1.2 [**User/Machine Systems**]: Human Factors

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.
Copyright 2011 ACM 978-1-4503-0717-8/11/10 ...$10.00.

General Terms

Experimentation, Human Factors, Measurement

Keywords

children, query logs, session analysis, topic classification, web search

1. INTRODUCTION

Both the fraction of children using the web and the amount of time they spend online has increased significantly in the past years. For instance, in the U.K. 63%, 76% and 83% percent of users 5 to 7, 8 to 11 and 12 to 15 years old respectively have access and use the Internet at home.[1] In the U.S. 32.4 million of children under the age of 18 years old were active users of the Internet in 2008, accounting for up to 19% of the online population.[2] Similar trends have been reported in other developed countries.

Given the small amount of content carefully designed for this audience and the lack of specialized search engines dedicated to help children find appropriate content on the web, there is an increasing need for research aimed at understanding the current difficulties these users experience searching for information on the Web.

To exemplify the difficulties that children encounter during the search process, consider the following two search session derived from the query log sample studied.

(1) A 10 years old girl submits the query *what is love*, the search engine triggers advertisements related to dating and casual encounters. Thinking that this ad is a result to the query, the girl clicks on it, after spending few seconds trying to understand what it is happening she goes back and then clicks on the first web result, which explains the chemical processes involved when people feel love. The content of this website goes beyond her reading skills and she quits the search session, most likely frustrated.

(2) When a 9 years old boy submits the query *hun*, the search engine suggests queries such as *hun school* (Princeton college), *hun sen* (primer minister of Cambodia) and *hun empire* (former empire ruled by Attila). Although this user is targeting the last topic suggested by the search engine, he does not seem to notice any of the query suggestions and simply continues with his initial query. Then he clicks on the first web result which happens to be a web directory of

[1] `http://stakeholders.ofcom.org.uk/binaries/research/media-literacy/ukchildrensml1.pdf`
[2] `http://www.iab.net/insights_research/530422/1675/600835`

links with adult content. Hun is also a popular term used to refer to a specific type of adult content. The user who is probably confused by the content he found decides to go back and then he clicks on the second web result which is the Wikipedia article of the Hun empire. As it was the case with our previous example, the article is dense and its language is too advanced for the reading capabilities of this user who after few seconds aborts the search session.

In these two examples, we observed that young users have a tendency to click on higher ranked results, spend short time on each url and in general have shorter sessions than those observed in older users. Although we observed that children do use the query suggestions provided by the search engine, these are less frequently used when they are displayed while the user is typing. In the first part of this paper, we quantify the struggle children experience while using a search engine using both established and novel query log metrics. Concretely, we present detailed information about queries, clicks and sessions. Details include aspects of query structure, click duration and session duration, among other features. All results are aggregated on a per-user basis and macro-averaged over age ranges that reflect human development stages. We explore how search suggestions can influence the success in their search process. We also investigated the likelihood of children to click on ads and to click on adult content by accident.

In the second part of the paper, we employ the query log as a mean to retrace the stages of child development. We point out differences in the topic distribution of what users search at different ages and gender. We show that the reading level of the pages clicked also varies according to the age and demographics such as the average income of the user's location.

The paper is organized as follows: in Section 2 we present the most relevant related work of previous studies on query logs and children search behavior. In Section 3.1 the data set used by us is described and the methodology followed to estimate the measures that we report is explained. In Section 4 we present the results and discussion of children's search difficulty. Section 5 discusses our findings on retracing child development stages using query logs. Both Section 4 and 5 are subdivided according to concrete research questions. We conclude in Section 6 with a discussion of our main findings and how they could be applied.

2. RELATED WORK

The most relevant literature on children search behavior on the web and query log analysis are described in the following paragraphs.

2.1 Information seeking on the web by children

Bilal [3, 4] investigated the behavior of children using the web directory Yahooligans![3] to solve open and well-defined informational tasks. The author reported that children are more successful in finding information when they used a navigational approach instead of keyword search. The author also reported that children are often ineffective at finding information as a consequence of frequent looped searches, hyperlink backtracking and poor query formulation. Druin et al. [7] characterized the search roles that children adopt during the search process and studied how these roles depend on the children's environment and their motivation. Our work differs from theirs in that we quantify the search characteristics and search difficulty of children based on aggregated results of thousands of users across a broad age range, which makes our observations more representative on a web-scale. Additionally we report topic interest trends over a population with diverse demographic characteristics.

[3]Today known as Yahoo! kids: http://kids.yahoo.com/

2.2 Related query log analysis

Duarte et al. [9, 8] compared users accessing general purpose information with users accessing information aimed at children. They reported significant differences between these type of users and show that some of their results are in line with previous case studies on children search behavior. Nonetheless, the authors argued that is not possible to assure that the users studied were children.

In [23] query logs are employed to study how search differs on users with different demographics. They used demographic information derived from the US-census and user profile information to describe search patterns and behaviors for population segments with different demographic characteristics. In this work, we employed an analogous methodology to show that the reading level of the urls clicked by children also varies across demographic features.

In [24] the authors tried to explain how the "who searches", the "what he searches", and the "how he searches" interact. Related to our work, they also gave details about topical distributions as function of a user's age. Though we apply similar methodological techniques, such as analyzing session characteristics, the main difference is one of breadth vs. depth. With our focus on a particular demographic group, namely children, we can go into far more details and paint a fine-grained picture of children's struggles and search behavior online.

3. DATA SET AND METHODOLOGY

3.1 Data set

The data set employed in this study was extracted from a large sample of the Yahoo! search logs of May to August of 2010. The following restrictions were employed to filter out search log data:

- Log entries of users without a valid Yahoo! account
- Log entries of users with unspecified birth year, gender and zip code in the Yahoo! profiles. Ill defined fields were also excluded (e.g. invalid zip codes)
- Queries containing personally identifiable information, such as credit card numbers or full street addresses
- Queries that were issued by only a single user
- Queries containing only a single token consisting exclusively of non-alphanumerical characters

For users below 10 years old we collected search volume in the order of hundreds of thousands of queries from tens of thousands of users. For users aged 10 or more years old we employed search volumes in the order of millions of queries from hundreds of thousands of users. As certain aspects of this data set are considered business sensitive, for various metrics we report *relative* differences between age groups, as opposed to *absolute* differences including, say, actual click-through-rates on ads.

The motivation of this study is to characterize the search activity of young web searchers and identify crucial differences between the search behavior across children of different ages and adults. For this purpose, we aggregated entries from the log data according to the user's birth year. Concretely, we estimate the age of the users by setting the date of birth as the 31 of December of the birth year provided in the user profiles and considering that the search was carried out in 2010. The following age ranges were created

- early elementary: 6-7 years old
- readers: 8-9 years old
- old children: 10-12 years old
- teenagers: 13-15 years old
- old teenagers: 16-18 years old
- adults: Above 18 years old

Children from 5 to 6 years old refine their motor skills and start to be involve in social games. Children from 6 to 8 years old start to expand their vision of the world beyond their immediate surroundings. Children from 8 to 12 years old acquire the ability to represent the entities of the world in terms of concepts and abstract representations. Teenagers on the other hand become more interested in social interactions [16]

Our selection of the age groups follows the development changes present in these stages of life. As we will show in this work, theses stages also have an impact on what children search on the web and on the way they interact with the search engine.

It can be argued that it is unreliable to trust the user information provided in the Yahoo! profiles since people can lie about their age, gender or geographical location. Nonetheless, since at last as early as 2007 Yahoo! has required the consent of a parent or legal responsible for users under 13 years old to create an account [4]. Currently Yahoo! charges a symbolic amount of $.50 to confirm that a guardian is responsible for the child creating the account. Apart from the (small) financial cost, the corresponding time and effort increases the chances of having veridical information for these age groups. Note that even if a small fraction of supposed child users lied about their true age, this is less problematic for general *trends* to be present or not, though the actual *absolute* numbers will be affected. It is also interesting to notice that in social networks, children tend to lie to make themselves appear older and this practice is often backed by parents [19].

As mentioned, we only used search and click events of users for whom we could obtain (self-provided) age, gender and US ZIP code. We then used the ZIP code in combination with US census information[5] to further annotate users with demographic estimates about their education level (the fraction of the population in a certain age range holding a bachelor's degree or higher). This techniques has been previously used in the context of query log analysis in [23, 24].

[4]`http://info.yahoo.com/privacy/us/yahoo/family/details.html`

[5]`http://factfinder.census.gov/`

3.2 Methodology

In all of our work, we take a user-centric approach as we want to provide insights into how children search online. This means that *all* of our statistics are macro-averages, where things are averaged with each user contributing equally, as opposed to micro-averages, where things are averaged over all query instances and heavy users will have a bigger importance.

For various parts of our analysis, we also make use of the notion of a *search session*. To break sequences of queries and clicks into sessions, we used a very simple approach that splits sequences of query and click events into sessions using a sliding window of 30 minutes. A similar approach has been used in several query log studies [14, 22].

For the queries and clicked documents (if any) we computed various metrics which will be explained in the sections where they are analyzed. However, the distinction between navigational and non-navigational queries [5] is used in several sections and so we describe here how this distinction was computed. We used two different approaches in parallel. First, we used the *click entropy* [24] to get estimates about how diverse the clicked results in response to a particular query were. Queries that had a sufficient support, a minimum of 2 occurrences, were judged as navigational if the click entropy was no larger 1.0. This approach works well for head queries and, e.g. detects the query "utube" as a navigational query. Additionally, we used a simple heuristics on given (query, click) pairs. Note that this heuristics does not label the query as such as navigational, but rather individual (query, click) pairs. So *facebook* could be non-navigational if the user clicks `http://en.wikipedia.org/wiki/Facebook`. Our heuristics works as follows: First query and url are tokenized (by white-spaces and dot characters respectively), then tokens are sorted and plurals are stemmed. We label the pair as navigational if the query contains a domain extension (i.e. www, .com, .org), the domain of the URL is contained in the query (or vice versa), or the edit distance between the query and the domain of the url is smaller that an threshold value (in the results reported we used 2 as threshold for queries containing more than 4 characters). For instance, this method is able to detect the navigational intent of the pair (kids abercrombi, `www.abercrombiekids.com/`).

In our arguments, comparisons between (macro-)averages computed for different groups, say session lengths of children between 6 and 7 and adults between 40 and 70, are core elements in our arguments. Hence, we were careful to test the various differences we report for statistical significance, using a two-tailed t-test for the equality of means with unequal variance and sample size.[6] We consider a difference to be statistically significant if the probability of the null hypothesis, i.e. the two means being equal, is smaller than 5%. Recall that the averages for each group are *macro*-averages across the users in the corresponding group.

4. SEARCH DIFFICULTY

Query, click and session characteristics were collected to identify differences in the search process between users of different ages and gender. In the following paragraph we analyze each one of these types of metrics. Our focus here is on finding metrics that give insight into the *search difficulty*

[6]`http://en.wikipedia.org/wiki/Student\%27s_t-test#Unequal_sample_sizes.2C_unequal_variance`

Table 1: Query length averages by query intent

Age	Global		Non-Nav.		Nav.	
	T. length	C. length	T. length	C. length	T. length	C. length
6 to 7	2.55	16.49	2.80	17.40	2.14	15.16
8 to 9	2.56	16.59	2.77	17.27	2.24	15.91
10 to 12	2.56	16.62	2.81	17.54	2.17	15.49
13 to 15	2.60	16.82	2.84	17.67	2.18	15.67
16 to 18	2.64	17.08	2.86	17.92	2.19	15.68
19 to 25	2.71	17.34	3.03	18.72	2.22	15.16
26 to 30	2.68	17.34	3.02	18.83	2.25	15.32
31 to 40	2.65	17.43	3.00	18.88	2.26	15.76
> 40	2.80	19.05	3.12	20.09	2.43	17.73

that children face. In particular, we are interested in metrics related to *confusion*.

4.1 Do children pose longer queries?

The formulation of a well-defined query is an crucial part of the search process in IR systems [6]. The correlation between query length and IR effectiveness has widely been explored before [2, 13]. On TREC ad-hoc settings it has been found that longer queries lead to better search performance and user satisfaction [2]. Nonetheless, recent studies show that this result does not always hold on the Web-scale [6]. Query length has also been associated to the specificity of the user's query intent, longer queries representing more specific and less ambiguous information needs [17]. In this work, the obvious two query length metrics were considered: token length and character length. Token length is measured as the number of tokens separated by white-spaces and character length is simply the number of characters (including white-spaces) in the query.

Table 1 summarizes the results obtained by age range and query intent. A clear increasing trend of length was observed from younger to older ages. This result suggests that younger users tend to formulate simpler information's goals. Given that the difference margin is larger for non-navigational queries, this result may also indicate that younger users have difficulties finding the right keywords to formulate more elaborated information needs.

4.2 Do children pose queries using natural language?

Children have been observed to pose queries in natural language given their lack of familiarity with the keyword approach of search engines. Moreover, at younger stages children typically have a greater sense of curiosity which we hypothesized is reflected in the searches they performed. The following query types were created to quantify these phenomena.

1. Question queries: Queries for which the first token is a question word (how, where, what, ...), or the last character of the query is a question mark (e.g. what is the only immortal animal?)

2. Modal queries: Queries containing auxiliary verbs as will, won't, don't or modal verbs as shall, should, can, etc. (e.g., I don't want to go school)

3. Knowledge questions: Queries containing the words *describe*, *about*, *explain*, *define* or *interesting*

Table 2: Fraction of query types

Age	quest.	modal	knowl. quest.	superl.	for kids
6 to 7	2.07%	0.41%	0.16%	0.91%	2.36%
8 to 9	2.56%	0.29%	0.08%	1.48%	1.74%
10 to 12	3.53%	0.58%	0.11%	1.46%	0.97%
13 to 15	3.84%	0.71%	0.16%	1.33%	0.43%
16 to 18	3.33%	0.69%	0.20%	1.15%	0.34%
19 to 25	2.80%	0.49%	0.20%	1.23%	0.32%
25 to 30	2.54%	0.44%	0.16%	1.16%	0.54%
31 to 40	2.19%	0.33%	0.14%	1.09%	0.68%
>40	1.69%	0.24%	0.11%	1.07%	0.31%

4. Superlatives: Queries containing superlative adjectives (e.g, the fastest dog)

5. Kids targeted queries: Queries with the terms *for kids* or *for children*

Knowledge queries attempt to measure the fraction of queries intended to fetch a specific explanation about an issue or topic. Superlative queries are commonly employed to satisfy the curiosity about certain topic such as in "fastest animal". These queries were detected by looking at tokens with the suffix *est* and filtering out those matched tokens that are not listed as adjectives in Wordnet[7] or that have a locational meaning (e.g., west). Kids targeted queries are employed to focus the search on content oriented for children. Table 2 summarizes the results obtained by age for the set of non-navigational queries. Although the age range 10 to 18 has the highest fraction of question queries, the 6 to 9 years group does not have a noticeably higher fraction than, say, the 31 to 40 years age range. Similarly, *lacks* of clear trends over the age groups can be observed for the other features. Only the "for kids" query type came close to behaving as expected. The fraction of superlative queries peaks for children in the 8 to 12 years age range.

4.3 Do children have a position click bias?

We collect the macro-averaged distribution of the result ranks clicked by the users. The macro rank distribution is computed by estimating the probability of each user to click at each rank position, only taking into account query instances with click, and then averaging the distributions across users belonging to the same age range. Figure 1 presents the distribution of clicks for the first five results

[7] `http://wordnet.princeton.edu/`

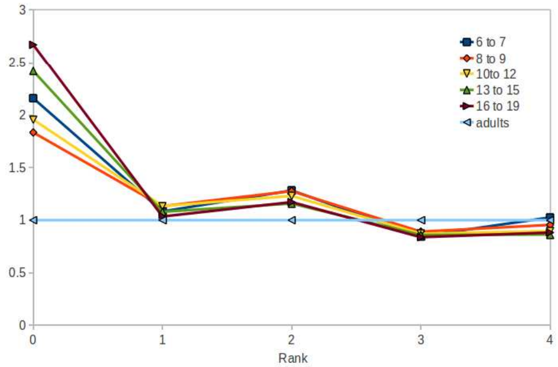

Figure 1: Relative rank frequency distribution across age ranges

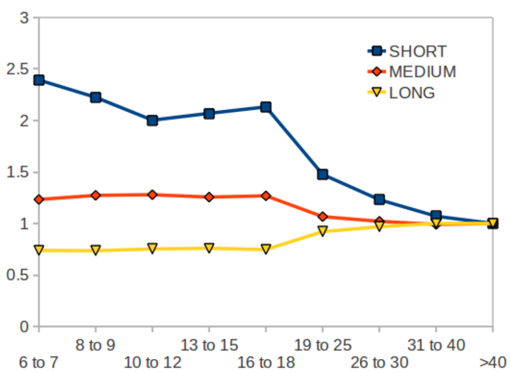

Figure 2: Distribution of click length across the age groups

where all click-through-rates are relative to those of adults. A rank of "0" refers to any kind of "special result" which includes links to current news, shopping results or any other high quality content which is typically only shown for high volume queries. Not surprisingly we found that younger users tend to click on higher ranked results, clicking twice as often as adults on the special rank 0 results. Similarly, for positions 1 and 2 were a factor between 1.1 and 1.3 as likely as adults to click. For lower positions this behavior is reversed.

4.4 Do children have more "long" clicks?

Previous work showed that one signal use to detect search success occurs in the form of *long clicks*. Here a long click is a click on a result such that after the click the user does not issue a new query or click on another result for at least 100 seconds. However, before the session times out (time limit of 30 minutes) he does submit another event so that the click duration can be estimated. Clicks at the end of a session have unknown click duration. We broke down non-final clicks into the three classes short (0-10 seconds), medium (11-99 seconds) and long ($\geq$ 100 second) [12]

Fig 2 shows that the fraction of long clicks is comparatively low for children of all ages, before it suddenly jumps to a higher level for users in the 19 to 25 years age range. This result indicates search frustration in younger users since they tend to abort the clicked pages sooner than adults.

Table 3: Relative ad click through rates

Age	Click ratio
6 to 7	1.28
8 to 9	1.14
10 to 12	1.04
13 to 15	0.89
16 to 18	0.9
19 to 25	0.84
25 to 30	0.8
30 to 40	0.86
$>$ 40	1.20

4.5 Are children more or less likely than adults to click on ads?

We employed the macro-fraction of ad-clicks to quantify how likely it is for an user of a given age range to click on an ad. Since not all the queries trigger advertisements, the estimation was performed only for clicks on results that were generated by queries that had triggered at least one click on an ad. Table 3 reports the fraction ratios of ad-clicks in respect to the group of adults users between 30 and 40 years old. Values greater than *1* means that users are more likely to click on ads than this age range of adults. Surprisingly, we observed higher ratios of ad-clicks for users at very young ages (6 to 12) which suggests disorientation during the search process for these users since ads are most of the time not targeted at this demographic segment. This observation is in line with previous research that showed that in the context of online games children are also more likely to click on ads as they fail to recognize them as such [1, 18]. It also reconfirms the findings concerning the position click bias from Section 4.3.

4.6 Are they relying more on search suggestions?

Druin et al. [7] reported in a detailed case study with 12 participants that children aged 7 to 12 often ignore the auto-completion and query suggestion facilities provided by search engines. This behavior occurs as a consequence of their longer attention on the typing instead of on the screen which make children ignore the queries suggested by the search engine. Figure 3 shows the fraction of queries that were submitted to the search engine as a product of a query suggestion or query correction. Query suggestions are triggered by the search engine when the user is typing the query (i.e. query auto-completion) or as the form of related searches right after the user has submitted the query. The automatic query correction functionality is triggered by spelling mistakes and are commonly displayed by the search engine by informing the user *We have included "britney spears" results - Show only "brittnay spears"*.

Figure 3 shows that, somewhat surprisingly, children are *not* more likely to make use of query suggestions or query corrections. However, we did observe the trend that the younger a user is, the more likely is he to *undo* a query suggestion, i.e. to insist on the (incorrect) spelling by clicking on the option like "Show only 'brittnay spears"'. The fraction for users aged 6 to 7 to click such an in-correction was a factor of 1.62 higher than for a user in the age range 26 to 30.

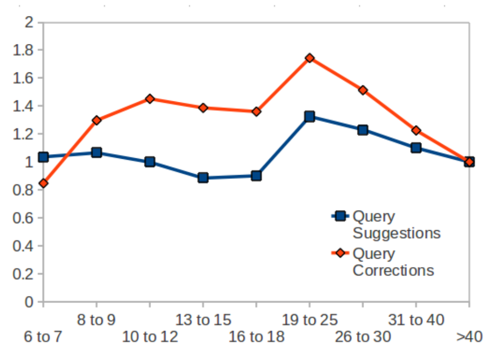

Figure 3: Query suggestions and correction usage

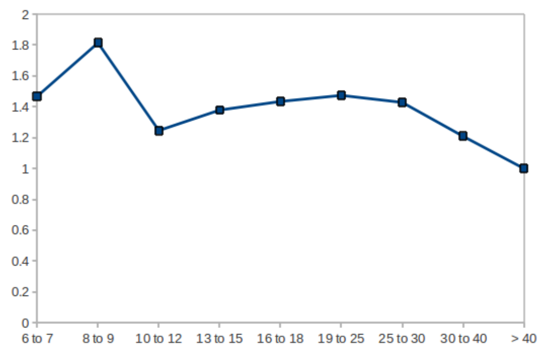

Figure 4: Relative likelihoods of accidental clicks on adult content websites.

4.7 Are children more likely to accidentally click on adult content, only to immediately correct their choice?

Children are potentially exposed to adult and explicit material on the Web given its large volume and the lack of parental supervision. Although, we observed lower volume of queries accessing adult explicit content for users below 13 years old (as it will be depicted in Section 5.1.1) it is important to quantify how often this content is accessed accidentally.

We hypothesized that users clicking by accident on a website with adult content would immediately go back and click on a different web result. Thus, we estimate the likelihood of having a click on a website without adult content after a short clicked on a website with adult content was registered during the same user session. Note that this process may occur more that once during the same user session. The last event of the sessions were ignored in the calculations since their click duration are unknown.

Figure 4 shows the relative frequencies for the event of an accidental short and immediately reverted click on adult content. Note that even though children in the 6 to 9 years age range have a comparatively high probability of immediately reverting to a different result after a (supposedly accidental) click on adult content, their *absolute* probability of clicking on this type of result or of issuing a related query is very low. The fact that the probability of these accidents-with-immediate correction is higher for children aged 6 to 7 than for children aged 8 to 9 can potentially be explained by the fact that the youngest children might take too long

Table 4: Session characteristics

Age	S.duration	S. length	Query ref.	Click ref.
6 to 7	3.79	3.76	0.24	0.17
8 to 9	3.51	3.71	0.24	0.15
10 to 12	3.63	3.71	0.23	0.14
13 to 15	3.91	3.76	0.26	0.14
16 to 18	4.04	3.82	0.26	0.13
adults <25	8.20	5.45	0.32	0.22
adults <30	8.45	5.43	0.30	0.20
adults <40	8.39	5.28	0.29	0.20
adults <70	8.42	5.25	0.34	0.24

to read an entry page explaining that the site contains adult material and that the visitor needs to be of legal age (typically 18) to view the content.

4.8 Are children's search sessions ill defined?

One indication for a user struggling with a query is the fact that a user goes back to a question issued earlier in the same session after temporarily exploring different queries. As our sessions were quite short, with an average of 3.51 minutes for ages between 8 and 9 years old, it is unlikely that the second occurrence of a query is indicative of a renewal of the earlier information need. More likely, it indicates that the user has not yet fulfilled the earlier information need. We call such queries that repeated with a session "query refindings and their fraction is computed as follows. For each user we estimate the fraction of refinding queries inside a session (in respect to the total number of queries inside the session). Then, we averaged this fractions for all the sessions of the same user to generate a per users estimate of query refinding usage. As with all the other metrics reported in this work, we report the macro-averages across users. Similarly, if a user clicks the same URL repeatedly (which can be interspersed with other events) this can be seen as an indication that he is struggling and trying to make up his mind about the most relevant result.

Apart from the fraction of refinding queries and clicks, Table 4 also shows two simple measures for the average session length, one for the length measured in minutes and one counting the number of events (queries, clicks and next result page) in a session. It is important to clarify that these estimations exclude sessions containing only one entry (i.e. sessions in which a query was submitted and no clicks were registered)

Table 4 shows that search sessions of children are considerably shorter that for adults. Surprisingly, this statement includes the 16 to 19 age range and the jump to "adulthood" occurs suddenly in the group 20 to 25.

The fraction of query refinding and, in particular, click refinding sessions is *lower* for children. However, rather than to be taken as an indication for a lower level of confusion it is more likely to be a result of the fact that children have (considerably) shorter sessions and so there is simply less opportunity to issue the same query or click the same URL again.

5. TRACING CHILDREN DEVELOPMENT STAGES

Previous work showed that, given enough search history of a user, attributes such as gender, age and location can

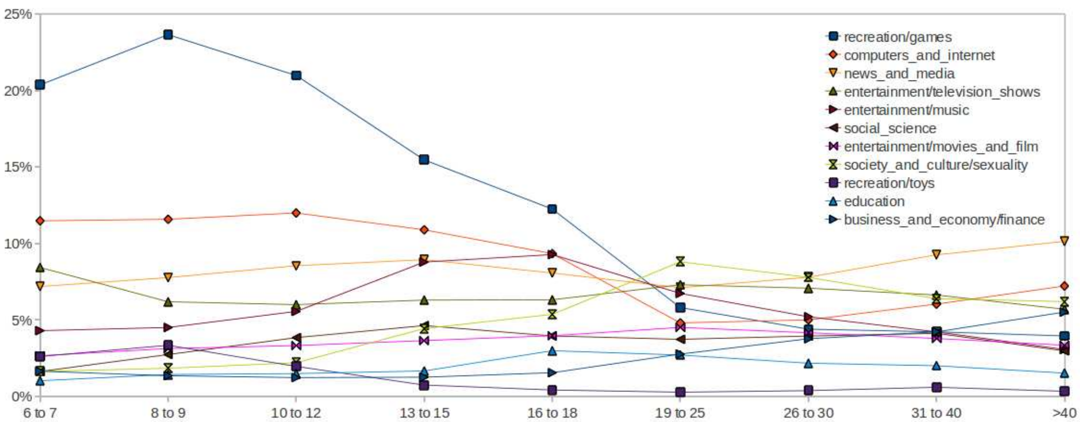

Figure 5: Topic progression through the ages

Table 5: Examples of queries and their mapped entities

Query	entity
facebook, facebook login	`en.wikipedia.org/wiki/Facebook`
disney cars games, 2011 cars, cars 2 2011	`en.wikipedia.org/wiki/Cars_2`
what to do with hummus, ideal protein, hummus recipe	`en.wikipedia.org/wiki/Hummus`
download b.o.b. airplanes, bob airplanes lyrics, airplanes part 2	`en.wikipedia.org/wiki/Airplanes_(song)`
youtuyoutube, youtui, youtuyoutube	`en.wikipedia.org/wiki/Youtube`
back to school clothes, london school uniforms, dress code for christian school	`en.wikipedia.org/wiki/School_uniform`

be estimated [15]. In this work, we looked at a related but different problem: can we find hints in the query logs that give indications about a child's development stage? In particular, can we confirm existing hypothesis and knowledge about child psychology?

5.1 What do children search for?

We investigated what children search for and how this evolves along two dimensions. First, which high level topics do they search for. Second, which concrete entities do they search for and which are typical characteristics of these entities. In both cases we tried to link our findings back to existing knowledge about child psychology, such as the development of gender differences or the orientation of children in certain age groups towards idols/heroes.

5.1.1 Which topics are children interested in?

We used a proprietary classifier to map web pages to entries in the Yahoo! Directory[8]. To obtain a classifier for *queries*, not pages, we then used a weighted majority voting scheme on the top 10 organic results returned by the Yahoo! search engine. See [24] for details. In total, there were 95 different topics. Figure 5 presents the average topic fractions for the 11 most frequent topics searched by users below 18 years old.

The behavior of Figure 5 is intuitive. Kids have a much higher fraction of queries falling into recreation/games than adults and the same holds, though at a lower level, for recreation/toys. The interest in music is most expressed in the teenage age range (13 to 18). The fraction of business/finance increases steadily as users grow older.

[8] `http://dir.yahoo.com/`

While we are most interested in understanding age-related differences, there are also important gender-related differences, even in children [10, 11]. We were interested in how gender differences evolve as children grow up. Are gender differences more pronounced in, say, teenagers than in adults? To answer this question we quantified gender differences by looking at the topical distribution for particular age groups. Each such topical distribution corresponds to a probability distribution, summing to 100%. We used the 1-norm to quantify the differences between the probability distributions belonging to boys and, respectively, girls for a given age group.

The blue line in Figure 6 shows that the gender differences for children are a lot smaller than for adults. However, a lot of these gender differences are due to a gender bias in the topic for adult content. The red line shows the gender differences when this topic has been removed and the remaining topics renormalized. As can be seen in the plot, this modification removes a large part of the age-related increase in gender differences.

The largest differences between genders were observed in the categories business and economy, computers and Internet and society and culture/sexuality. Nonetheless, this differences were significantly higher for users male and females above 16 years old, which is the trend that is observed in Figure 6.

5.1.2 Which entities are children interested in?

As the topics we used were fairly broad, such as "music" or "finance & investment", we were also interested in obtaining more fine-grained information by looking at the (main) *Wikipedia entity* a query refers to. To map queries to Wikipedia articles we used the following simple, yet effec-

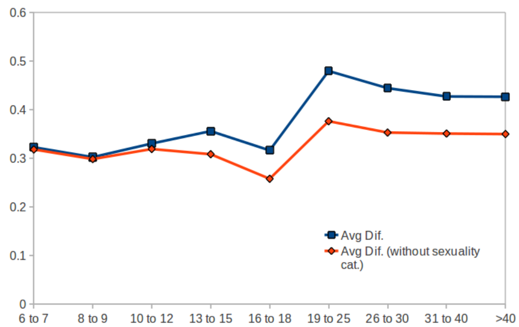

Figure 6: Average topic difference between genders through the ages as measured by the $||1||$-norm

Table 6: Entity fractions for Child related content and Living People according to the Wikipedia categories

Age	Children and Kids	Living people
6 to 7	5.81%	8.11%
8 to 9	5.49%	6.97%
10 to 12	3.38%	7.59%
13 to 15	1.46%	9.47%
16 to 18	0.95%	10.86%
19 to 25	0.62%	11.96%
26 to 30	0.63%	11.54%
31 to 40	0.89%	11.09%
>40	0.62%	10.88%

tive approach: we ran each of the millions of distinct queries on the Yahoo! search engine and limited the results to results from `http://en.wikipedia.org/wiki/`. The first result was then used as the entity representation for the query. Note that these queries against Wikipedia were only run fairly recently, though the original queries were submitted about one year earlier. This ensured that even for recent events almost always a Wikipedia page could be found. Table 5 shows some examples of this mapping.

An overview of the entities searched by young and adults users is presented by the tag clouds in Figure 7 and 8 respectively. These entities correspond only to the non-navigational queries found in our data set. Entities related to adult content were also manually removed.

One of the advantages of mapping queries to Wikipedia pages is that Wikipedia pages come with a categorical classification and that this classification is both more fine-grained and in a certain sense orthogonal to our own topic classification (see Section 5.1.1). For example, pages about current celebrities almost always belong to the "Living People" category. Similarly, there are many child-related categories such as "Early childhood education". We used a simple pattern match for the prefixes "child" and "kid" to identify these pages. In Table 6 we present the fraction of entities associated to children content in Wikipedia and the fraction of famous people found in the queries for the age groups.

We expected young children or at least teenagers to have a large fraction of celebrity related fraction. However, that did not turn out to be the case and the highest fraction of such queries was observed for people of college age. The trend for child-related categories was much more intuitive and, as expected, this fraction is the more pronounced the younger the user is.

Table 7: Mean query sentiment values scores

Age	Positive	Negative	Diff
6 to 7	1.233	-1.211	0.0216
8 to 9	1.253	-1.237	0.0161
10 to 12	1.257	-1.248	0.009
13 to 15	1.284	-1.274	0.0101
16 to 18	1.274	-1.258	0.0165
19 to 25	1.300	-1.283	0.023
26 to 30	1.302	-1.275	0.026
31 to 40	1.322	-1.297	0.0248
>40	1.400	-1.376	0.0279

5.2 Do children express stronger sentiments in their queries?

To find out if children or adults are "more rational" in formulating a query we looked at the presence or absence of *sentiments* in queries. This was motivated by the fact that children in the age range of 9 to 12 tend to experience extreme changes of mood [20], which we hypothesized could be reflected in the formulation of queries.

To assign numerical scores to sentiments being present in queries, we used the SentiStrength[9] tool developed by Thelwall et al. [21]. This tool simultaneously assigns both a positive and a negative score to bits of English text, the idea being that users can express both types of sentiments at the same time such as in "I love you but I also hate you". Positive sentiment strength scores range from +1 (not positive) to +5 (extremely positive). Similarly, negative sentiment strength scores range from -1 to -5. The tool works by assigning scores to tokens in a dictionary which includes common emoticons. For example, "love" is mapped to $+3/-1$ and "stink" is mapped to $+1/-3$. Modifier words or symbols can boost the score such that "really love" is mapped to $+4/-1$ (the same for "love!!" or "looove"). The final positive sentiment strength for a bit of text is then computed by taking the maximum score among all individual positive scores. The negative sentiment strength is similarly calculated.

As can be seen in Table 7, sentiment analysis applied to individual queries did *not* reveal the expected trend. It did however reveal that the tendency to use both more positive and negative words in a question *increases* as users get older. This phenomenon is at least partly explained by the fact that they issue longer queries (see Table 1) and hence the probability of positive/negative sentiment words appearing is higher.

5.3 Does the reading level of the clicked results vary across ages and education levels?

One of the most noticeable factors in child development and its relation to web search behavior is an improvement in reading skills. As children improve their reading proficiency they will be able to (i) make sense of a wider range of web results, and (ii) potentially understand better the various elements of a web search engine, such as query suggestions or advertisements.

[9] `http://sentistrength.wlv.ac.uk/`

Figure 7: Entity tag cloud: 10 to 12 years old

Figure 8: Entity tag cloud: above 40 years old

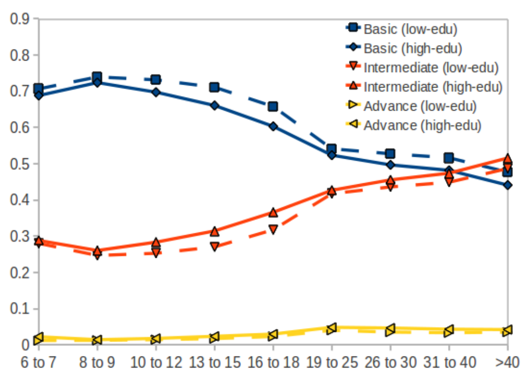

Figure 9: Reading level across age and average educational level

To retrace and quantify the improvement in reading level in our data set, we mapped clicked result pages to a 3-scale reading level using Google's "annotate results with reading levels" option.[10] Here, we simply issued the URL of the page of interest as a query to Google. In cases where the full URL did not return any results, or at least no results with an annotated reading level, we used backtracking by iteratively chopping of parts from the end of the URL, hopefully finding a shorter URL for which information could be obtained.

Table 8 gives a few examples for each of the three reading levels. Note that a single host such as `http://en.wikipedia.org` can host pages of all three reading levels. Averaged across all web pages, irrespective of the corresponding query volume, 51.6% of our URLs were classified as "basic", 35.8% as "intermediate" and 2.9% as advanced. For 10.2% we could not obtain a reading level with the current approach.

We observed a general and strong trend for the fraction of clicks on "basic" reading level pages to decline for older users. At the same time we observed a weak increase for the "advanced" level and a strong increase for the "intermediate" level.

To understand which other factors, apart from age, influence the preferred reading level of users, we also broke down users according to the education level in their self-report ZIP code. Here we used the census feature "percentage of population of the age of 25 or higher holding a bachelor degree or higher". We sorted users according to this features and investigated the lowest 20%-tile, and the highest 20%-tile. Figure 9 shows that, for the fraction of basic reading level pages, children from well-educated areas have about 3 years of advantage over children from poorly-educated areas. For example, a child from the age range 16 to 18 has a fraction of basic result clicks of 65%. This is slightly lower than the fraction for children in the age range 13 to 15 from well-educated areas, which is 66%, and much higher than the fraction of 60% for other children in the 16 to 18 age range also coming from well-educated areas.

6. CONCLUSIONS AND FUTURE WORK

In several aspects of our analysis, we observed a notable difference between children and adults but often, and this came as a surprise to us, the differences between the different age groups were quite small and children between 16 and 18 behaved more like children between 8 to 9 than young adults in the 19 to 25 range. This "sudden jump to adulthood", albeit not for all features, could potentially be explained by children leaving home and starting college or a job.

As far as useful lessons learned are concerned, the position click bias for children is worth pointing out. This bias also leads to a higher fraction of ad clicks and to a higher fraction of cases where (useful and correct) spelling corrections are *undone* by the user as he clicks on, say, "Show only 'brittnay spears"'. Both of these indicate that very young users have a tendency to "click whatever is presented at a prominent position" which has implications for the design of an appropriate search interface.

In the future, we plan to continue the line of thought that "child development can be observed through search logs". Concretely, we would like to investigate (i) how the differences between concrete and abstract entities depends on the

[10] `http://www.google.com/support/websearch/bin/answer.py?hl=en&answer=1095407`

Table 8: Examples of websites for each of the three reading levels

reading level	example urls
basic	www.cookingtips-recipes.com, www.funbrain.com, en.wikipedia.org/wiki/Toy_Story, www.pbteen.com
intermediate	en.wikipedia.org/wiki/John_Wooden, horoscopes.astrology.com, www.sprint.com/, www.foxnews.com
advanced	www.answers.com/topic/mathematics, www.medicinenet.com/tinnitus/article.htm, www.merriam-webster.com

age, (ii) if FSK levels of movies or suggested age limits for games match the age of the user, and (iii) if "how to" or "how can I" queries and their topics can be used to describe what type of problems children are facing at different development stages.

Our hope is that a deeper understanding of children's search behavior and their struggles will lead to a reduction of the prevalent "one size fits all" search interface. In particular, our work suggests that due to the strong position click bias a linear, ranked list might not be the best way of presenting search results and alternatives should be explored.

7. ACKNOWLEDGMENTS

We would like to thank Alejandro Jaimes for helpful discussions at an early stage of this work. The pointers to the New York Times articles [18, 19] inspired parts of this work. We are also thankful to Mounia Lalmas and Djoerd Hiemstra for their encouraging and valuable comments.

This research was partially funded by the European Community's Seventh Framework Programme FP7/2007-2013 under grant agreement no. 231507

8. REFERENCES

[1] S. An and S. Stern. Mitigating the effects of advergames on children; do advertising breaks work? *Journal of Advertising*, 40(1):43 – 56, 2011.

[2] N. J. Belkin, C. Cool, D. Kelly, G. Kim, J. y. Kim, H. j. Lee, G. Muresan, M. c. Tang, and X. j. Yuan. Query length in interactive information retrieval. In *In Proceedings of the 26th Annual ACM International Conference on Research and Development in Information Retrieval (SIGIR '03)*, pages 205–212, 2003.

[3] D. Bilal. Children's use of the yahooligans! web search engine: II cognitive and physical behaviors on research tasks. *Journal of the American Society for Information Science and Technology*, 52:118–136, January 2001.

[4] D. Bilal. Children's use of the yahooligans! web search engine: III cognitive and physical behaviors on fully self-generated search tasks. *Journal of the American Society for Information Science and Technology*, 53:1170–1183, November 2002.

[5] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36:3–10, September 2002.

[6] D. Downey, S. Dumais, D. Liebling, and E. Horvitz. Understanding the relationship between searchers' queries and information goals. In *Proceeding of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 449–458, 2008.

[7] 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 *Proceedings of the 8th International Conference on Interaction Design and Children*, IDC '09, pages 89–96, 2009.

[8] S. Duarte Torres, D. Hiemstra, and P. Serdyukov. An analysis of queries intended to search information for children. In *Proceeding of the third symposium on Information interaction in context*, IIiX '10, pages 235–244, 2010.

[9] S. Duarte Torres, D. Hiemstra, and P. Serdyukov. Query log analysis in the context of information retrieval for children. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 847–848, 2010.

[10] J. Eccles, A. Wigfield, R. D. Harold, and P. Blumenfeld. Age and gender differences in children's self- and task perceptions during elementary school. *Child Development*, 64(3):830–847, 1993.

[11] J. S. Eccles, J. E. Jacobs, and R. D. Harold. Gender role stereotypes, expectancy effects, and parents' socialization of gender differences. *Journal of Social Issues*, 46(2):183–201, 1990.

[12] A. Hassan, R. Jones, and K. L. Klinkner. Beyond DCG: user behavior as a predictor of a successful search. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 221–230, 2010.

[13] B. Jansen, M. Bernard, J. Jansen, A. Spink, and T. Saracevic. Real life, real users, and real needs: A study and analysis of user queries on the web. *Information Processing & Management*, 36:207–227, March 2000.

[14] R. Jones and K. L. Klinkner. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *Proceeding of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 699–708, New York, NY, USA, 2008. ACM.

[15] R. Jones, R. Kumar, B. Pang, and A. Tomkins. "i know what you did last summer": query logs and user privacy. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, pages 909–914, 2007.

[16] R. Kail. Children and their development. Pearson Education, 2009.

[17] N. Phan, P. Bailey, and R. Wilkinson. Understanding the relationship of information need specificity to search query length. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 709–710, 2007.

[18] M. Richtel. Children fail to recognize online ads, study says, April 2011. http://bits.blogs.nytimes.com/2011/04/21/children-fail-to-recognize-online-ads-study-says/?hpw.

[19] M. Richtel and M. Helft. Facebook users who are under age raise concerns. *New York Times*, page B1, March 2011. http://www.nytimes.com/2011/03/12/technology/internet/12underage.html?pagewanted=1&_r=1.

[20] P. Smith, H. Cowie, and M. Blades. Understanding children's developments. BlackWell, 2003.

[21] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas. Sentiment in short strength detection informal text. *Journal of the American Society for Information Science and Technology*, 61:2544–2558, 2010.

[22] S. K. Tyler and J. Teevan. Large scale query log analysis of re-finding. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 191–200, 2010.

[23] I. Weber and C. Castillo. The demographics of web search. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 523–530, 2010.

[24] I. Weber and A. Jaimes. Who uses web search for what: and how. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 15–24, 2011.