"Is Sven Seven?": A Search Intent Module for Children

Nevena Dragovic
Computer Science Dept.
Boise State University
Boise, Idaho, USA

Ion Madrazo Azpiazu
Computer Science Dept.
Boise State University
Boise, Idaho, USA

Maria Soledad Pera
Computer Science Dept.
Boise State University
Boise, Idaho, USA

{nevenadragovic, ionmadrazo, solepera}@boisestate.edu

ABSTRACT

The Internet is the biggest data-sharing platform, comprised of an immeasurable quantity of resources covering diverse topics appealing to users of all ages. Children shape tomorrow's society, so it is essential that this audience becomes agile with searching information. Although young users prefer well-known search engines, their lack of skill in formulating adequate queries and the fact that search tools were not designed explicitly with children in mind, can result in poor outcomes. The reasons for this include children's limited vocabulary, which makes it challenging to articulate information needs using short queries, or their tendency to create queries that are too long, which translates to few or irrelevant retrieved results. To enhance web search environments in response to children's behaviors and expectations, in this paper we discuss an initial effort to verify well-known issues, and identify yet to be explored ones, that affect children in formulating (natural language or keyword) queries. We also present a novel search intent module developed in response to these issues, which can seamlessly be integrated with existing search engines favored by children. The proposed module interprets a child's query and creates a shorter and more concise query to submit to a search engine, which can lead to a more successful search session. Initial experiments conducted using a sample of children queries validate the correctness of the proposed search intent module.

CCS Concepts

•**Information systems → Query intent;** *Web search engines;* •**Social and professional topics → Children;**

Keywords

Query Intent; Children; Search engines;

1. INTRODUCTION

The Web is an essential forum for gathering information for different purposes. To assist end users in locating materials from the Web, different tools have been developed. Among these tools, general-purpose search engines, such as Bing, Google, and Yahoo!, are the most prominent ones, as users of all ages turn to them on a daily basis to retrieve desired information. As one of the largest communities that search for online resources, children are introduced to the Web at increasingly young ages. Trends in the U.S. indicate that by age 5, $half$ of all children will turn to the Web on the daily basis, a number that increases to *two-thirds* by age 8 [9]. While early exposure can help them build foundational skills vital in a knowledge-rich society, search tools were not designed with children in mind nor do retrieved results explicitly target children, which causes many of them to fail to complete successful searches [8]. Most engines do not explicitly support, or offer weak support, for children's inquiry approaches. This is important to address given that early experiences can affect attitudes in using the Web, skill development in making adequate use of resources for personal and educational interests, and the ability to leverage information and use it to make contributions into adulthood [14].

As described in the 2013 book, Search Engine Society, "Children growing up in the 21^{st} century have only ever known a world in which search engines could be queried, and almost always provide some kind of an answer, even if it may not be the best one" [10]. However, like most adults, children struggle with describing their information needs in a concise query [6]. Moreover, they are known to sometimes use long natural language queries that can often lead to retrieving irrelevant resources [4]. To facilitate the search task, search engines offer query suggestions. Unfortunately, children often bypass suggested queries altogether, which could eliminate possible misspellings or better reflect their information needs. Even if children, as inexperienced users, struggle with stating the right queries to initiate successful searches, search engines are expected to retrieve relevant information in response to their requirements. While search intent modules have been developed to automatically generate queries that capture users' needs [11, 12], none of these specifically target children's queries. With that in mind, in this paper we introduce $QuIK$ (Query Intended for Kids), a simple, yet effective, search intent module tailored towards children ages 6-15. Its goal is to transform an initial children query Q into Q', a query that captures the intended meaning of Q. In accomplishing this task, we analyze the terms in Q to: (1) distinguish misspelled terms from the ones used by children that are not included in traditional dictionaries, (2) identify cultural references transmitted via mass media (children trendy terms), and (3) alter terms that

SIGIR '16, July 17-21, 2016, Pisa, Italy
© 2016 ACM. ISBN 978-1-4503-4069-4/16/07...$15.00
DOI: http://dx.doi.org/10.1145/2911451.2914738

could diminish the probability of a successful search session. Thereafter, *QuIK* identifies representative (updated) terms to be included in *Q*' which is submitted to Google, the preferred search engine for children [1].

The novelty of the proposed module is highlighted in its ability to modify children queries to initiate the retrieval of kid-related materials, which can lead to improving search engines' performance. The design of the proposed module explicitly considers different patterns kids use while searching the Web to adequately capture the intended meaning of their original queries. Furthermore, the proposed module has the ability to distinguish misspellings from terms commonly used by kids, as is the case of terms that refer to children trendy terms to which they are exposed on a daily basis (e.g., *Xbox*), emphasized terms (e.g., *aammaazziinngg* can indicate that something is very amazing), or diminutives (e.g., *daddy*). *QuIK* also handles long queries or questions, which is a must, given that children between 10 and 15 years old represent the largest group of users who tend to create natural language queries, some of which are in the form of a question [7]. Furthermore, by replacing an original query without imposing the burden on the child of analyzing the suggestions provided by search engines' "did you mean" feature (that eliminates misspellings from originally-submitted queries), *QuIK* considers children's habit to scroll pass the provided suggestions, due to the lack of visual cues [4]. To the best of our knowledge, no module that can be coupled with existing search engines can capture search intent exclusively from children's queries and regenerate queries on-the-fly to foster the retrieval of child-friendly results.

2. RELATED WORK

As described by Bilal et al. [2], to adequately serve children, search engines must address the fact that children are seldom successful in formulating succinct queries. In fact, researchers have observed that children tend to use long (natural language) queries, as opposed to keyword queries, when performing a web search [4]. Unfortunately, the longer the query, the less likely a web search engine can retrieve relevant resources in response [4] to it, which can be very frustrated for them. Search intent can help capture users' information needs by understanding the purpose behind the used words, which is important not only for children, but for all groups of users with a web search issue.

A number of studies [11, 12, 15] address the issue of search intent. Their strategies focus on: creating a hierarchical taxonomy using a tree to find representations of generic intents from user queries [15], examining bias between users' search intent and the query generated in each search session [11], or investigating query intent when users search for cognitive characteristics in documents [12]. Mentioned strategies, however, fail to specifically consider issues pertaining directly to children, which demonstrates a clear need for new modules that automatically create queries on their behalf. *QuIK* simultaneously addresses misspellings, queries with terms that children use but that are harder to match by search engines, purposely emphasized terms, and trendy topics that are more popular among children than among adults. Such topics are yet to be addressed in the literature.

3. QUIK, A SEARCH INTENT MODULE

To understand children's writing patterns along with the manner in which they search the Web, it is imperative to examine children query logs. Unfortunately, the Yahoo dataset used in [7] is not publically available, while other query logs available through Webscope[1] contain only the most popular queries among users, which do not give a real representation of child-formulated queries. Consequently, we consider children's search sessions extracted from the AOL query log following the strategy in [5]. These sessions (6471 queries), however, are outdated and lack references to trendy terms. Knowing this limitation, we also consider an alternative to query logs that captures recent content of interest to children and treat content (i.e., sentences) written by children as queries. In doing so, we identify common information needs, writing styles and issues that may influence the manner in which queries are formulated by kids. We relied on reviews (275004 sentences) written by elementary school-aged children archived at Spaghetti Book Club[2].

Based on the analysis of AOL and Spaghetti, we observed that 76% and 54% of the queries contained one or more terms[3] not recognized by WordNet[4], a large lexical database of English words. Deeper analysis allowed us to identify a number of patterns related to the way in which children express their information needs (see Figure 1). Children are not always proficient typists, and many look at the keyboard as they type. Unfortunately, looking down instead of up at the screen causes them to miss search engine query suggestions [4]. Given that *misspellings* negatively impact a search session and children tend to make more spelling mistakes than adults, correction is a must [4]. However, it is an imperative to differentiate misspelled words from words favored by children, including *diminutives* and *exaggerated* words. As described in [3], "diminutives are a characteristic of child-directed speech and diminutive suffixes are among the first morphemes that a child acquires and uses in his/her speech". Exaggerated words refer to intentionally misspelled terms to provide emphasis. Another characteristic of children search, is explicitly referencing *trendy* terms among their peers. Some of those terms can be treated as misspellings by search engines (e.g., Google suggests "*seven*" instead of "*Sven*" a character in Frozen, a popular Disney movie) which causes poor retrieval of desired resources.

To better understand children's information needs, *QuIK* infers the search intent from a child's query Q and automatically replaces it with a shorter, more concise query Q' that can lead to better search outcomes. To do so, *QuIK* analyzes every non-stop term t in Q to determine if it is a good candidate to capture the intended meaning of Q and thus should be included in Q'. To verify if t is misspelled, *QuIK* uses the popular Apache Lucene SpellChecker[5]. However, not all $t's$ (not recognized by WordNet) are treated as misspellings, if t ends with a common diminutive suffix, we retrieve the root form of t by eliminates last two letters of t. If t is an exaggerated term, then its true meaning is determined by eliminating multiple repeating characters, which generates a new, normalized version of t. Furthermore, it is

[1] https://webscope.sandbox.yahoo.com/

[2] https://wordnet.princeton.edu/

[3] For practical reasons, stopwords in AOL and Spaghetti were removed, terms were lemmatized and URLs were ignored, since search engines adequately handle queries including the names of popular sites.

[4] http://www.spaghettibookclub.org/

[5] https://lucene.apache.org/

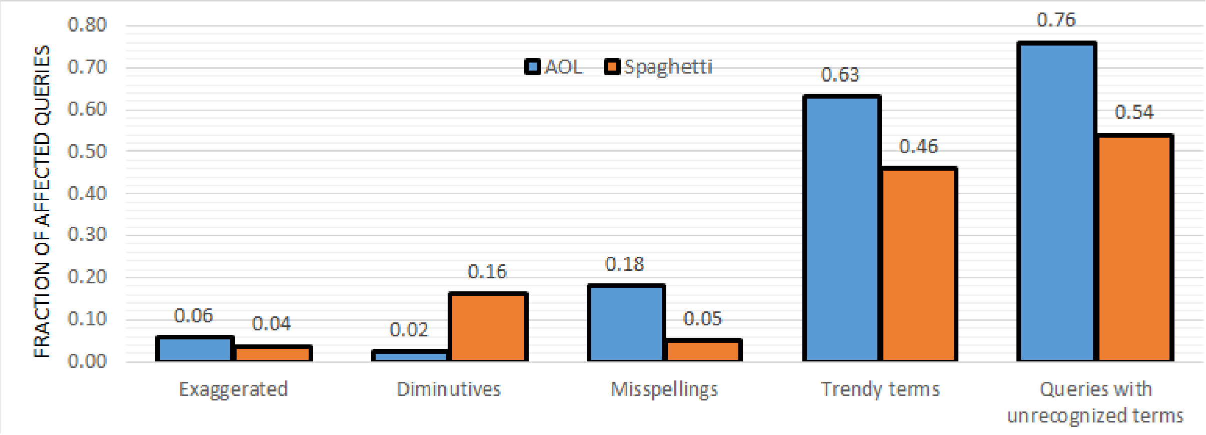


Figure 1: Writing patterns observed on AOL and Spaghetti data sources

very important to capture trendy terms and enable $QuIK$ to understand their significance and provide relevant results for children. Hence, $QuIK$ considers existing categories on well-known websites with children content (e.g., Amazon[6] and Common Sense Media[7]) to determine if a t is misspelled or if it is in fact a trendy, child-related term. We cannot assume that kids will use only child-friendly vocabulary in formulating Q. To address this issue, we study term associations and consider terms related to the ones in Q that can steer the search towards retrieving child-related resources. For example, a query that uses the term *doctor* instead of *surgeon* is better tailored towards the retrieval of suitable children content, since the former is a simpler term. To precisely understand the complexity of terms used in queries, we created a children's dictionary CD. CD includes close to 100,000 lemmatized, non-stop words, extracted from texts retrieved from a sample of various children-related websites (where content is written for and by children), such as Long Long Time Ago, KidPub and Teenreads. If t is contained in CD, then $QuIK$ treats it as child-related, otherwise it turns to WordNet to find a more suitable term by selecting one of t' hypernyms and use it instead of t[8].

After each t in Q has been updated based on the identified patterns, if needed, or recognized as a trendy term, $QuIK$ proceeds to create a new, "cleaner" version of Q. Given the results from previous studies, we know that general audiences use just 2.3 words on average for a query generation [13], while the number increases to 3.8 when considering children [5]. For this reason, $QuIK$ uses the 3 most representative terms to correctly capture the information need expressed in Q. To quantify the degree to which each non-trendy term t in Q expresses children's information need, we calculate a score for t using the content retrieved from children-related sites (used in the creation of the children dictionary), as shown in Equation 1.

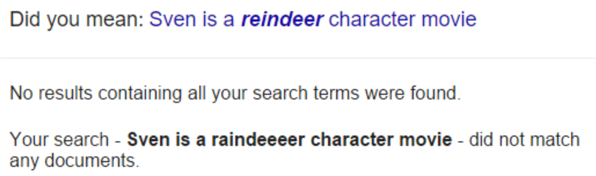
$$RepScore(t) = \frac{\sum_{d \in C} \frac{freq_{t,d}}{|d|}}{|C|} \qquad (1)$$

where C represents a collection of documents of size $|C|$, d is a document in C, $|d|$ is the number of words in d, and $freq_{t,d}$ represents frequency of occurrence of t in d.

In creating Q', $QuIK$ gives priority to children trendy terms in order to better understand the search intent. The rest of the terms to be included in Q' are the terms in Q that yield the highest $RepScore(t)$. Lastly, Q' is submitted to Google to initiate a search on behalf of a child for Q.

4. QUIK IN ACTION

We applied $QuIK$ on 55000 queries from Spaghetti (created as discussed in Section 3 and disjoint from the set of queries used for development purposes) and observed that the number of queries with one or more words non-recognized by WordNet decreased from 16% to 2.3%. At the same time, we submitted sampled children queries to Google that included a number of the aforementioned identified patterns. We observed that when submitting queries including exaggerated and trendy terms, Google retrieves no results (see Figure 2). Furthermore, after submitting "sven movie", Google provided suggestions such as "seven movie online", "seven movie ending", "seven movie deaths" and "seven movie trailer". As we can see, none of these suggestions is kid-related, and obviously does not pertain to the popular Disney movie Frozen. These examples provide further evidence on the need for modules such as $QuIK$, which are tailored towards satisfying children's needs.



Did you mean: Sven is a ***reindeer*** character movie

No results containing all your search terms were found.

Your search - **Sven is a raindeeeer character movie** - did not match any documents.

Figure 2: Screen capture of Google in response to the query "Sven is a raindeeeer character movie"

5. INITIAL ASSESSMENT

We conducted an initial assessment to evaluate both (i) the quality of the updated queries generated by $QuIK$ given an initial child-formulated query and (ii) the degree to which

[6] http://www.amazon.com/

[7] https://www.commonsensemedia.org/

[8] If the returned term does not belong to CD, then t remains unchanged.

$QuIK$-generated queries influence the retrieval of documents relevant to children. Due to the lack of benchmark datasets, we created a survey in which, for a given child-generated query CQ, we asked appraisers to (1) choose the query reformulation that best captures the intended meaning of CQ and (2) choose the document that is the most relevant given CQ. Among the provided query reformulations, we included (in random order) the $QuIK$-generated query for CQ as well as two query suggestions provided by Google[9] for CQ. Similarly, among the provided results, we included (in random order) the top-two ones retrieved using CQ as well as the top-retrieved one using $QuIK$-generated query for CQ. We asked 25 elementary-school educators to complete the surveys using a small sample of children queries (including "Sven is a reindeer character in a movie" and "how does the baby kengaroos stay inside the pouch"[10]).

Based on the collected evaluations (the results of which are summarized in Table 1), we observed that in 59% of the responses, queries generated by $QuiK$ were selected as the ones that best capture the intended meaning of a given children query. Moreover, 88% of the responses indicated that $QuIK$-generated queries lead to the retrieval of relevant resources, as opposed to resources retrieved by Google in response to the original given query.

	QuIK	**Google**
Favored search intent	0.59	0.41
Favored relevant resources	0.88	0.12

Table 1: Initial assessment of $QuIK$

6. CONCLUSION AND FUTURE WORK

We discussed an initial analysis conducted to identify common struggles children face when expressing information needs while searching the Web. Knowing that children represent a large group of online users, and the importance that search outcomes have on their development, it is crucial to provide more attention to this task. While a number of studies identify issues and flaws of search engines while interacting with young users, very few research works explicitly focus on solving these problem, which is why research pertaining towards search intent modules tailored for this group is a must. $QuIK$ simultaneously considers newly discovered and well-known patterns identified in queries created by children, and creates queries that capture the information needs meant to be expressed by children.

Our discussed efforts represent the groundwork of an in-depth research. We plan to extend initial work by analyzing children query logs, and simultaneously addressing new (and already identified) issues pertaining to children querying to further enhance the design of proposed search intent module. A limitation of $QuIK$ is not considering extremely short queries often generated by young users. This is sometimes insufficient to retrieve needed information and as such creates more opportunity for improvement for $QuIK$. Another opportunity for $QuIK's$ enhancement is understanding the ambiguity of users' search intent during a session. This is important given that the same query may target different information for different users. As mentioned in Section 5, further assessment is a must. We plan on conducting online evaluations with a greater number of appraisers (including teachers and children) complemented with measuring the effect of performance of search engines coupled with $QuIK$ to enhance Web search environment.

7. REFERENCES

[1] D. Bilal. Ranking, relevance judgment, and precision of information retrieval on children's queries: Evaluation of google, yahoo!, bing, yahoo! kids, and ask kids. *JASIST*, 63(9):1879–1896, 2012.

[2] D. Bilal and R. Ellis. Evaluating leading web search engines on children's queries. In *Human-Computer Interaction. Users and Applications*, pages 549–558. Springer, 2011.

[3] I. Dabasinškienž. Intimacy, familiarity and formality: Diminutives in modern lithuanian. *Lituanus*, 55(1), 2009.

[4] 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*, pages 89–96, 2009.

[5] S. Duarte Torres, D. Hiemstra, and P. Serdyukov. Query log analysis in the context of information retrieval for children. In *ACM SIGIR*, pages 847–848, 2010.

[6] S. Duarte Torres, D. Hiemstra, I. Weber, and P. Serdyukov. Query recommendation for children. In *ACM CIKM*, pages 2010–2014, 2012.

[7] S. Duarte Torres, I. Weber, and D. Hiemstra. Analysis of search and browsing behavior of young users on the web. *ACM TWEB*, 8(2):7, 2014.

[8] E. Foss, A. Druin, R. Brewer, P. Lo, L. Sanchez, E. Golub, and H. Hutchinson. Children's search roles at home: Implications for designers, researchers, educators, and parents. *JASIST*, 63(3):558–573, 2012.

[9] A. L. Gutnick, M. Robb, L. Takeuchi, J. Kotler, L. Bernstein, and M. Levine. Always connected: The new digital media habits of young children. Joan Ganz Cooney Center at Sesame Workshop, 2011.

[10] A. Halavais. *Search engine society.* John Wiley & Sons, 2013.

[11] B. Hu, Y. Zhang, W. Chen, G. Wang, and Q. Yang. Characterizing search intent diversity into click models. In *WWW*, pages 17–26, 2011.

[12] M. P. Kato, T. Yamamoto, H. Ohshima, and K. Tanaka. Investigating users' query formulations for cognitive search intents. In *ACM SIGIR*, pages 577–586, 2014.

[13] T. Lau and E. Horvitz. *Patterns of search: analyzing and modeling web query refinement.* Springer, 1999.

[14] N. Vanderschantz, A. Hinze, and S. J. Cunningham. "sometimes the internet reads the question wrong": Children's search strategies & difficulties. *Am. Soc. Info. Sci. Tech.*, 51(1):1–10, 2014.

[15] X. Yin and S. Shah. Building taxonomy of web search intents for name entity queries. In *WWW*, pages 1001–1010, 2010.

[9] We consider Google for comparison purposes, since, as previously stated, is the search engine of choice for children.

[10] Any misspelled terms included in the queries provided in the survey were reproduced as per the original children queries.