Suggesting Simple and Comprehensive Queries to Elementary-Grade Children

Meher T. Shaikh [#1], Maria Soledad Pera [*2], Yiu-Kai Ng [#3]

[#] *Computer Science Department, Brigham Young University, Provo, Utah, 84602, U.S.A.*
[1] mtalats@gmail.com, [3] ng@compsci.byu.edu

[*] *Department of Computer Science, Boise State University, Boise, Idaho, 83725, U.S.A.*
[2] solepera@boisestate.edu

***Abstract*—Query suggestions (QS) tailored specifically for children are slowly gaining research attention in response to the growth in Internet use by children. Even though QS offered by popular search engines adequately meet the information needs of the general public, they do not achieve equivalent effectiveness from a child's perspective. This is because children's search behaviors, interests, cognitive levels, and ability to read and understand complex content are different from adults. Given the ubiquitous nature of the Web, its importance in today's society, and its increasing use in education, it is an urgent need to help children search the Web effectively. In this paper, we present a QS module, denoted CQS, which assists children in finding appropriate query keywords to capture their information needs by (i) analyzing content written for/by children, (ii) examining phrases and other metadata extracted from reputable (children's) websites, and (iii) using a supervised learning approach to rank suggestions that are appealing to children. CQS offers suggestions with vocabulary that can be comprehended by children and with topics of interest to them. Empirical studies conducted using keyword queries initiated by children, in addition to feedback gathered through crowdsourcing, have verified not only the effectiveness of CQS, but also the fact that children favor CQS-generated suggestions over the suggestions provided by Google, Yahoo!, and Bing.**

***Index Terms*—Query suggestion; children; backpropagation**

I. Introduction

Children regularly use search engines as the starting point in their quest for information [18]. Unfortunately, their search experiences can be negatively influenced by their lack of skill in formulating adequate search queries. While query suggestion (QS) modules designed for widely-used search engines facilitate query creation for a general audience, they were never designed from a child's perspective. A 2013 survey shows that children spend an average of six hours a day online,[1] making them active web users. Furthermore, approximately 5% of popular search engine users are children,[2] that explains why Google plans to create a children version of its search engine.[3] With the growth of this segment of Internet users, there is a demand for the development of QS modules tailored towards children.

Suggestions made by existing general-purpose QS modules may require advanced reading level on complex topics which children have difficulty understanding [3]. The discrepancies between children's and adults' search behaviors/interests were further verified by the study conducted by Torres et al. [16] who identified significant differences between queries for a general audience and queries that seek information on children's content, such as average length of queries (3.2 words for children versus 2.5 for regular users). Even though search engines designed specially for children, such as safesearchkids.com, kidsclick.org, and kidrex.org, exist, majority of them are not equipped with a QS module. To aid children with their quest for information that satisfy their needs, we have developed CQS, a QS module that offers suggestions for children who are up to twelve-year old.

Existing query recommendation/transformation techniques attempt to improve a submitted keyword query through word replacements, insertions, and deletions [5]. CQS, on the other hand, minimizes the effort required by a child in specifying his/her search intent by providing query recommendations, which are N-gram suggestions, that yield the suffix to the user's initial keyword query. Suggestions made by CQS for a child's query Q follow the common design methodology of existing web search engines, such as Google, Yahoo!, and Bing. Instead of reformulating Q, these popular engines offer suggestions by appending keywords to the end of Q.[4] The same applies to children search engines, including Kidzsearch.com, a leading kids' safe search engine.

CQS relies on bigrams extracted from multiple reputable websites that include content written for or by children, and differs from existing QS modules targeting children [18], [17] which rely on tags assigned by adults to describe children/teenager's websites for making query suggestions. CQS also considers bigrams extracted from Simple.Wikipedia.com, denoted SimpWiki, which is an evolving collection of documents written in basic English. SimpWiki targets young readers and adults learning English as a second language. Based on the content, which includes simple vocabulary and is written so that children can understand, CQS can suggest meaningful and useful phrases to children.

CQS, a unique and novel QS module, (i) requires neither

[1] http://goo.gl/QQAH2l
[2] http://goo.gl/xe4Ygq
[3] http://goo.gl/TfxcEr

[4] We examined hundreds of suggestions made by Google, Yahoo!, and Bing, and all of them were generated by a completion-based approach.

query logs, ontologies, nor user-feedback to make query suggestions, (ii) enhances children's web search experience by considering the potential multiple interpretations of keywords specified in a user-initiated query, (iii) offers suggestions that are free from the influence of adult or generic intent, since the main sources of information used by CQS in generating suggestions[5] are children's documents and Library of Congress Subject Headings that are known to be associated with children literature, (iv) generates queries on-the-fly, as opposed to applying expansion techniques or depending on archived queries, to offer suggestions for a child-initiated query that change overtime, as more up-to-date children content is considered, and (v) ensures that suggested queries target topics of information appealing to children in an attempt to better simulate their *search intent*, and facilitate their quest for useful information. For example, by considering topics of information, CQS correctly assumes that a child looking for information using the query "tiger" favors "tiger population" or "tiger cub" as suggestions instead of "tiger woods net worth."

CQS can easily be adopted to handle suggestions for different niche of users. For example, CQS can make suggestions for teenagers by gathering content written for/by teenagers besides identifying topics of interest to them, which can be extracted from various web directories, such as Dmoz.com, and websites targeting teenagers. The QS strategy developed for CQS can also be adopted to make suggestions targeting specific themes/topics which demonstrates the flexibility in its applicability and generality. For example, suggestions pertaining to "sports" can be made by an extended CQS which extracts information of specific sport categories, such as ”tennis” and ”soccer”, in addition to other sport-related contents from sports websites.

II. Related Work

QS in itself is a non-trivial task for web search engine designers, since it requires disambiguating user's search intent using very few query keywords, i.e., 2.8 words on the average. If a QS module is designed for addressing children's information needs, as opposed to a general, i.e., more mature, audience, then it has to analyze children's search intents and behaviors, which are different from those of an adult [6].

While research on QS systems targeting children is limited, research work on QS systems for a general audience is rich and well-documented. Existing QS approaches for a general audience [14] either adopt probabilistic methodologies, examine query logs, apply strategies based on random walks, or rely on ontologies, to name a few.

In suggesting queries for young audiences, Duarte et al. [17] introduce a QS module based on tags created at Delicious.com. The team constructs a bipartite graph using tags and their corresponding URLs, and suggests queries as a result of a random walk on the bipartite graph that is biased towards children's content. Later research [18] presents further enhancements based on topical and language modeling features, such as topic-sensitive Page Rank and children-related vocabulary distribution, to more effectively suggest queries for children. Similar to the approaches described above, CQS does not rely on query logs to generate suggestions. CQS, however, differs from these QS modules for children, since, instead of using a bipartite graph, CQS considers diverse features that aim to precisely capture children's intents. More importantly, CQS relies on content written for/by children to suggest queries as opposed to relying on tags that are often provided by adults and may be poorly defined due to the lack of quality control on user tags and thus can be inherently noisy [5].

Eickhoff et al. [7] present a two-step query expansion strategy for children. Given a query Q, it retrieves top-n results from various search engines and uses tags assigned to each retrieved web page at Delicious as keywords to expand Q. In addition, the name of high-level semantic categories (inferred from Wikipedia and the DMOZ.org taxonomy) associated with tags are treated as expansion terms as well. While CQS generates cohesive phrases to guide users in formulating queries, the approach in [7] simply provides tag-related terms to add to the given query to locate children-related content.

TABLE I
Categories defined and used by CQS based on information available at children websites

Adventure	Animals	Books	Comedy
Did You Know	Education	Entertainment	Health
History	Music	Nature	Science
Space	Sports	Video	World

III. Our Query Suggestion Approach

Using bigrams extracted from children's websites and well-established probabilistic/information retrieval models, CQS identifies each *candidate suggestion* for a user query Q and the closest *categories*, i.e., topics, to which Q belong. Table I shows the list of categories defined at well-known children's websites and considered by CQS. All the websites, which include various types of information extracted by CQS for generating query suggestions, are shown in Table II.

For each candidate suggestion CS, CQS computes its *ranking score* using a backpropagation (BP) model [13] on a number of features, which are described below. The top-ranked phrases, which are simple, easy to read, and better capture topics of interest to children, are offered as suggestions for Q.

A. Candidate Suggestions

To determine the candidate suggestions for query Q with m (≥ 1) words, i.e., terms, CQS examines the *frequency of occurrence* of words that follow the last word t_m in Q in a category c. CQS identifies the frequencies of the top-five,[6]

[5] Relying on content archived at children's websites assures that the generated suggestions will not favor content not specifically targeting young audiences. For example, the suggestion "basketball wives" generated by general-purpose QS modules for the query "basketball" might lead to content above the maturity level of children.

[6] Given that suggestions including seldom-occurring words are less likely to make it to the top ranking positions among the suggestions for Q, CQS considers only the top-five most frequent words. In doing so, CQS speeds up its processing time without affecting its accuracy, a fact that has been empirically verified.

TABLE II
WEBSITES USED BY CQS FOR GENERATING QUERY SUGGESTIONS FOR CHILDREN

Website	URL	Data Used by CQS
Spaghetti Book Club	www.spaghettibookclub.org/	Training phrases for BP
Good Book Recommendations	best-kids-books.com/good-book-recommendations.html	Training phrases for BP
Mother Daughter & Son Book Reviews	motherdaughterbookreviews.com/	training phrases for BP
American Literature: The Children's Library	americanliterature.com/childrens-library	Bigrams
Reader Views: reviews, by kids, for kids	readerviewskids.com/reviews-by-age/	Bigrams
Dogo news: Fodder for young minds	www.dogonews.com/	Bigrams, Categories info. & likelihood, Naïve Bayes (NB) feature (kid class)
Time for Kids	timeforkids.com	Bigrams, NB feature (kid class)
Kidworld	www2.bconnex.net/∼kidworld./	Bigrams, NB feature (kid class)
National Geographic: Kids	kids.nationalgeographic.com/kids/	Bigrams, Categories info., NB feature, (kid class)
Simple English Wikipedia	Simple.Wikipedia.org	Bigrams, Phrase simplicity
Stone Soup: The Magazine by Young Writers and Artists	www.stonesoup.com/archive/stories	Bigrams, Categories info., Category likelihood, NB feature (kid class)
BookHive: Your Guide To Children's Literature	www.cmlibrary.org/bookhive/books/	Bigrams, Categories info., Category likelihood
Reading Rockets	www.readingrockets.org/article/22366	Children's vocabulary
BigIQkids	bigiqkids.com/SpellingVocabulary/Lessons/wordlist SpellingFirstGrade.shtml	Children's vocabulary
The Game Gal	http://www.thegamegal.com/printables/	Children's vocabulary
Children's Library	archive.org/details/iacl	NB feature (kid class)
Free Kids Books	freekidsbooks.org/	Bigrams
Mighty Books	mightybooks.com/	Category likelihood
Poetry for Kids	www.poetry4Kids.com	Bigrams
Gutenberg - Children's Fiction	gutenberg.org/wiki/Children's_Literature_(Bookshelf)	Bigrams
Randomly Selected 10,900 Wiki Documents	www.wikipedia.org/	NB feature (generic class)

most frequently-occurred t_{m+1} words following t_m, denoted $f(t_m, t_{m+1})$, in c. For each one of the top-five words t_{m+1} in c, CQS considers the next top-five $f(t_{m+1}, t_{m+2})$ frequency values and so on to determine candidate suggestions in different categories for Q. To obtain the frequency distribution of bigrams that are used in generating candidate suggestions, CQS examines the consecutive word occurrences in the 82,000 documents belonging to the 16 categories extracted from children's websites (see Table II). Using word occurrences, CQS considers $f(t_m, t_{m+1})$ and creates phrases as suffixes of Q, which yield candidate suggestions for Q.

B. Category Likelihood

Given a query Q, CQS computes the likelihood of keyword(s) in Q matching the contents of different categories. To determine the *category likelihood* of Q, CQS employs the multinomial model, along with the well-known Bayes' rule [5], to compare the probability distribution of terms in different categories using the m ($\geq$ 1) terms in Q as shown below.

$$P(c|Q) = \frac{\prod_{i=1}^{m} P(k_i|c)P(c)}{\sum_{c \in C} \prod_{i=1}^{m} P(k_i|C=c)P(C=c)} \tag{1}$$

where $P(c)$ is the probability of observing c, which is the ratio of documents in c to the total number of documents used to train the multinomial model, C is the set of pre-defined 16 categories considered by CQS, and $P(k_i|c)$ is the probability that the i^{th} term in Q is observed in c as determined using the multinomial model and is defined below.

$$P(k|c) = \frac{tf_{k,c} + 1}{|c| + |V|} \tag{2}$$

where $|c|$ is the number of non-stop, stemmed keywords in the training documents of c, and $|V|$ is the number of distinct non-stop, stemmed keywords in 41,847 training documents extracted from children's websites, and $tf_{k,c}$ is the *frequency of occurrence* of term k in c.

CQS treats the likelihood value computed in Equation 1 as one of the measures to determine the significance of each candidate suggestion CS and computes the *category likelihood* score of CS with respect to c, denoted $CL(CS, c)$, as

$$CL(CS, c) = P(c|Q). \tag{3}$$

If the probability of keywords in Q belonged to category c is *high*, then a candidate suggestion originated from c is treated as a *more promising* suggestion for Q. CQS offers diverse query suggestions by considering various categories to which a given *ambiguous* query can be interpreted.

C. Bigram and N-gram Frequencies

CQS relies on frequency distribution of bigrams within categorized documents to provide statistics of consecutive term occurrences in different categories, i.e., categories shown in Table I. We define a *term* as a non-stopword keyword, which can be preceded by a sequence of *connection words*[7]. For example, if a user enters the query keyword "information," a suggestion "information about animals" makes more sense than the suggestion "information animals," since in the latter case the relationship between the two keywords is missing. To obtain the frequency distribution of bigrams, CQS examines the consecutive term occurrences in the aforementioned 82,000 children's documents (including SimpWiki documents), which are distributed across 16 categories.

[7] A *connection word* [2] is either a preposition, a conjunction, or an article, which is treated as a stopword and is *not* counted as words in a suggestion but is retained to capture the precise meaning of a suggestion.

We consider *bigram* frequency distribution of terms so that searching for related terms to a user query becomes *more efficient* than examining the frequencies of occurrence of sequences of more than two terms which increases the database size [2] for the document collection and significantly impacts the search time for terms related to a given query.

For a given query Q with the only keyword A, CQS generates candidate suggestions for Q by considering phrases, which are N-grams that includes A, such as ABC and $ABCD$ in which AB, BC, and CD are bigrams belonging to documents of the same category c. In generating phrases, CQS first considers all the frequent bigrams with the leading A. Based on the statistical data as to the *frequency of occurrence* of words B that follow A, and words that follow B, and so on, CQS concatenates the bigrams such that the 2^{nd} word in the preceding bigram is the 1^{st} word in the subsequent bigram to generate candidate suggestions. Thus, candidate suggestions are N-grams generated from across different categories, which are constructed based on the frequencies of word co-occurrences. CQS computes the N-Gram frequency score for each candidate suggestion CS.

Given a CS, which is sequence of words, t_1, …, t_n ($n > 1$) extracted from documents in c, CQS computes the average (N-Gram) frequency of the sequence below.

$$f(t_1, \ldots, t_n) = \frac{f(t_1, t_2) + \ldots + f(t_{n-1}, t_n)}{n - 1} \tag{4}$$

where $f(t_{i-1}, t_i)$, $1 < i \leq n$, is the frequency of the bigram made up by the word t_{i-1} followed by word t_i. A high N-gram frequency value of CS indicates that, in general, CS includes highly co-occurring bigrams in c and is treated as a favorable suggestion.

We consider a special case, i.e., $f(t_{i-1}, \text{EOS})$, where EOS stands for "End Of Sentence". $f(t_{i-1}, \text{EOS})$ captures the frequency with which term t_{i-1} is the *last* term in a sentence. CQS considers all the terms following t_{i-1}, including EOS.

CQS applies N-gram frequency to determine which candidate suggestions should be considered favorable. Given the keyword query "football," the N-grams "football star," "football team," and "football league division" are multiple suggestions to be examined, which turn out to occur more often than "football jersey" and "football kit" in the children's documents. Subsequently, the former are more appealing as query suggestions to children than the latter straightly based on their *frequencies of occurrence*.

D. Children's Vocabulary

CQS determines "children-friendliness" of the vocabulary used in a candidate suggestion by consulting a *vocabulary dictionary* comprised of words appropriate for children that were downloaded from children's word lists posted at a number of children's websites (see Table II for details). If a term in a suggestion is found in the dictionary, it is assigned the value of 1; otherwise, it is given the value of 0. For a candidate suggestion CS with multiple terms, CQS *averages* the values over all the terms in CS and obtains a single value between 0 and 1, which is called the *Vocabulary score* for CS.

E. Phrase Simplicity

As indicated earlier, one of the design goals of CQS is to offer suggestions that children can understand, i.e., simple query suggestions. To determine the *simplicity* of a candidate suggestion CS, CQS measures not only whether non-stop words in CS appear in children vocabulary, but also how often they are seen in web pages that are written in a simple style. Given that texts consisting of short sentences and simple words are deemed easier to read than those including longer sentences and rare words [1], it is assumed that web pages written using *basic* English vocabulary and *shorter* sentences are tailored for children. SimpWiki is such a website. Hence, we maintain a count of all the words in the entire collection of documents archived at SimpWiki.

A candidate suggestion CS is ranked higher if it includes keywords in the SimpWiki documents. CQS considers the *normalized frequencies* for the occurrences of keywords in SimpWiki documents, with values between 0 and 1. A value closer to 1 for a non-stop word in a phrase indicates that the word is very often found in SimpWiki, which reflects its *degree of simplicity*. For example, the word "call" with a value 0.7 indicates that it is commonly found in simple text documents, as compared with "torrent", which is assigned a value 0.0003, is less-frequently-used in kid's simple text documents. The values of the keywords in CS are *averaged*, and the averaged value is referred as the *simplicity score* of CS.

Simplicity vocabulary differs from the children vocabulary introduced in Section III-D. First, words in the children vocabulary are *not* extracted from a collection of text documents. Instead, they are words that a child is expected to know. Second, a word in the simplicity vocabulary is simple but may be absent in the children vocabulary. For example, kids are familiar with "Nintendo" and "Transformers," which are *simple* based on their *frequency of exposure* to children. However, they are not commonly-occurred words in children's literature, and hence are not assigned to the children vocabulary. Quite often to distinguish which one of the two candidate suggestions with all the words in the children vocabulary is more appealing to a kid, CQS must rely on their *simplicity* scores. Consider the suggestions "animal park" and "animal liberation" for the query "animal." Based on the higher frequency of occurrence of "park" than "liberation" in children's literature, "animal park" is assigned a higher *simplicity score* than "animal liberation." The generated simplicity scores are reasonable, since a child is more likely looking for "parks" to take their pets to than information about animal "liberation" movement.

F. Children Phrase Distribution

While *phrase simplicity* indicates how often keywords in a query suggestion CS are used by kids, it does not show whether the words are more likely to be found in documents pertaining to kids than in documents belonging to generic audience. To measure the word distribution of CS within kids' documents more likely than in general-audience's documents, CQS determines its *Naïve Bayes (NB) classification score* using the NB model. This score captures the probability distribution of keywords in CS that are also in *children's* content. We trained a *multinomial model* [5] using 10,900 documents randomly chosen from Wikipedia.org for the non-kid's class and 25,115 documents from children's websites (see Table II) for the kid's class. Although the number of documents in kid's class is more than the number in non-kid's class, it does not create an imbalance, since documents in the former are comparatively shorter than the Wikipedia documents, resulting in a vocabulary smaller than the non-kid's class. Given CS, the NB feature calculates the likelihood of CS using the Naïve Bayes' rule [5].

G. Locality

By concatenating different bigrams in a category into an N-gram phrase, some undesirable phrases, such as "greek language french cyclist" which consists of frequent bigrams 'greek language', 'language french', and 'french cyclist', can be created and should be avoided. To eliminate their creations, CQS determines the *locality* score for each candidate suggestion CS. The locality score, as defined below, which is based on the *Lennon Similarity measure* [12], captures the *likelihood* of all the bigrams in CS being extracted from the same document(s) within a given category c, such that the *smaller* the number of documents in c in which all the bigrams in CS occur, the *less likely* CS is an appealing one.

$$locality(CS, c) = \frac{S_n}{Min\{S_{l_1 l_2} - S_n, \ldots, S_{l_{n-1} l_n} - S_n\} + S_n} \tag{5}$$

where n is the number of terms in CS, l_i $(1 \leq i \leq n)$ is a term in CS, S_n is the number of documents in c that include all the bigrams in CS, and $S_{l_i l_j}$ $(i, j > 0)$ is the number of documents in c that include bigram $l_i l_j$. It is easy to see that the *more* bigrams in CS, the *fewer* the number of documents that include all the bigrams in CS which yields the *lower* the locality score. By using *Min* in Equation 5, we normalize the locality score of CS without penalizing CS on its length.

H. Subject Headings

It has been shown [8] that searching information on the Web can be facilitated by searching for the topics of the desired information. With that in mind, we have designed CQS to examine the topics of information addressed in candidate suggestions and penalize suggestions that are associated with topics/themes that are not commonly associated with children content. To accomplish this task, CQS relies on *Library of Congress Subject Headings* (LCSH), which is a de facto universal controlled vocabulary and constitutes the largest general indexing vocabulary in the English language. LCSH, which are keywords or phrases that denote concepts, events, or names, are employed by librarians to categorize and index books according to their themes, i.e., topics. Examples of LCSH include "Fairy tales" and "Fear of the dark-Fiction".

To identify, among the large number of LCSH, the subject headings that address topics of interest to children, we (i) examined LCSH assigned to 30,000 randomly selected books known to be suitable for children with readability levels between the K-6 grades defined by publishers and (ii) generated a list of 10,749 children's LCSH, denoted $cLCSH$, that describe text content in children's literature using subject keywords.

In examining the topical information of a candidate suggestion CS, CQS employs Equation 6 to determine the *degree of closeness* of CS and children's subject headings, denoted $SHScore$. To compute the SHScore feature score of CS, CQS compares CS against each subject heading SH in $cLCSH$, and chooses the *highest* similarity value between CS and the subject headings as the value that quantifies the degree to which CS addresses themes suitable for children. The degree of similarity between CS and SH is computed using the *word correlation factor* (wcf)[8] [15] of each (non-stop, stemmed) word in CS with respect to (non-stop, stemmed) words in SH.

$$SHScore(SH, CS) =$$

$$MAX_{SH \in cLCSH} \frac{\sum_{i=1}^{n} Min\{\sum_{j=1}^{m} wcf(CS_i, SH_j), 1\}}{n} \tag{6}$$

where n (m, respectively) is the number of distinct (non-stop, stemmed) words in CS (SH, respectively), CS_i (SH_j, respectively) is a (non-stop, stemmed) word in CS (SH, respectively), and $wcf(CS_i, SH_j)$ is the correlation factor of CS_i and SH_j.

The *Min* function in Equation 6 imposes a constraint on summing up the correlation factors of words in the description of CS and SH. Even if a word in the description of CS (i) matches exactly one of the words in SH and (ii) is similar to some of the remaining words in SH, which yields a value greater than 1.0, CQS limits the sum of their similarity measure to 1.0, which is the word-correlation factor of an exact match. This constraint ensures that if CS contains a dominant word w in its description which is highly similar to a *few* words in SH, w alone cannot dictate the content resemblance value of CS with respect to SH. Words in SH that are similar to most of the words in CS should yield a greater $SHScore$ value than the $SHScore$ value of words in SH that are similar to only one dominant word in CS. The *Max* function, on the other hand, ensures that the $SHScore$ of CS reflects the highest similarity of CS among all the subject headings which most effectively captures the topics/theme of CS.

A candidate suggestion CS with a *high* $SHScore$ illustrates that CS is closely related to contents in children's literature and hence is treated by CQS as *more favorable* compared with other suggestions with *lower* $SHScore$.

I. Ranking Candidate Queries

Using the individual scores of the features introduced in Sections III-B through III-H, which are computed for each candidate suggestion CS, CQS ranks the candidate suggestions belonged to multiple categories so that the top-k sugges-

[8] wcf reflects the *degree of similarity* between any two words CQS relies on wcf, as opposed to WordNet-based similarity measures, since it has been empirically verified that the former correlates with human assessments on word similarity more accurately than the latter [15].

tions[9] are recommended to its user by CQS. CQS relies on a backpropagation model to generate a single score for each candidate suggestion CS that reflects the cumulative effect of each of the seven features computed for CS and determines the degree to which CS is a suggestion suitable for children. BP is a machine learning algorithm based on neural networks, which learns weights associated with different inputs, i.e., features in our case, and is often used to perform categorization and/or ranking tasks [9].

In training the BP model for CQS, 138,579 training instances were used. Each instance includes a given noun-phrase, in lieu of a query, and is associated with the seven different feature scores computed for the corresponding noun-phrase and a label, which is either 1 or 0, to designate whether the noun-phrase is a children or generic query, respectively. In gathering the training instances of children queries, noun-phrases from a number of children websites, including spaghettibookclub.org, motherdaughterbookreviews.com, and best-kids-books.com,[10] were extracted. Training instances associated with generic queries, on the other hand, included queries extracted from the AOL query log,[11] a well-known source of general-audience queries.

IV. Experimental Results

In this section, we discuss the results of the empirical studies conducted to assess the design of CQS.

A. Evaluation Framework

Due to the lack of benchmark datasets to evaluate the design methodology and performance of QS modules for children, we turned to a number of 7- to 12-year-old children who are 1^{st}- to 6^{th}-grade students at a local school. During the month of April 2014, we asked the students to first create keyword queries that they would like to use to conduct their searches.

We used *eight*, five unigram and three bigram, queries randomly chosen out of the 127 unique queries provided by 137 elementary school students to evaluate the performance of CQS. For each query Q, we applied CQS to generate the top-4 query suggestions,[12] which were mixed with the top-4 suggestions of Q offered by Google, Yahoo!, and Bing, the three widely-used search engines.

Each child who participated in the study was asked to choose *four* useful suggestions for each of the eight test queries, which are *arctic circle*, *british*, *chocolate chip*, *football*, *greek*, *ice cream*, *information*, and *snow*. The top-4 most frequently chosen suggestions for each test query Q, among the choices provided by the 43 children who participated in the evaluation, were treated as the *gold standard* of Q.

We acknowledge that the number of queries considered for the evaluation of CQS by school children is relatively small. However, given that (i) evaluations involving children are difficult to conduct due to privacy constraints [17] and (ii) we only had access to students for a limited amount of time as each student involved in the assessment was given 15 minutes to complete the evaluation imposed by their school administrators, we limited the number of queries to be assessed to eight which allowed each student to spend an average of at most 2 minutes on evaluating suggestions for a query.

To determine the effectiveness of CQS and existing QS modules (considered for comparison purpose) in making useful suggestions to children, we have computed the *Normalized Discounted Cumulative Gain (nDCG)* value [5] on their corresponding top-4 suggestions for each test query. nDCG *penalizes* relevant suggestions that are ranked *lower* in the list of suggested queries.

B. Performance Evaluation

To verify the correctness of CQS, we first assessed its overall performance which we compared with the performance of a number of existing QS modules. Thereafter, we evaluated the effectiveness of each of its features.

1) Combination Strategies: For each candidate suggestion CS, CQS generates seven different scores, one for each of the features presented in Section III. In order to combine the seven scores into a single one which determines the *ranking* of CS, CQS considers different combination strategies: (i) CombMNZ [11], which is a linear combination measure frequently used in fusion experiments [4], (ii) Reciprocal Rank Fusion (RRF) [4], and (iii) the BP model presented in Section III-I.

We have empirically verified that BP is significantly better than CombMNZ and RRF in combining different features of a candidate suggestion based on the Wilcoxon signed-rank test.

2) CQS versus QS Modules: We compared CQS with the QS modules employed by Google, Yahoo!, and Bing, which is an evaluation framework similar to the one adopted by the authors of [18], [17]. One of the strengths of our evaluation strategy lies on the fact that we rely on children's assessments, and there is no room for adult-based bias. This is because we use keyword queries initiated by children as test queries and the top-4 suggestions selected as the gold standard are the ones chosen by children.

Figure 1 shows the performance of Google, Yahoo!, Bing, and CQS using the nDCG measure. The results have verified that suggestions made by CQS are more appealing to children than the ones offered by Google, Yahoo!, and Bing. CQS shows a statistically significant improvement ($p \leq 0.01$) in nDCG with respect to Yahoo! and Bing.

Besides analyzing the overall performance of CQS, Google, Yahoo!, and Bing using nDCG, we also examined their performance at the *query level*. As shown in Figure 2, CQS outperforms Google in making suggestions in *four* out of the 8 test queries. More importantly, CQS-offered suggestions are placed at *higher* ranking positions compared to Google.

We attempted to compare CQS against other children QS modules [7], [18], [17]. Unfortunately, implementing these modules requires setting up different parameters which are not

[9] k in top-k suggestions is determined by the software developer who implements CQS and is recommended to be in the range of 4 and 10.

[10] These sources include diverse content for creating sample children queries addressing multiple topics.

[11] http://goo.gl/TOIcz5

[12] The top-4 suggestions of CQS were used, since Google offers four suggestions for each query, for comparison purpose.

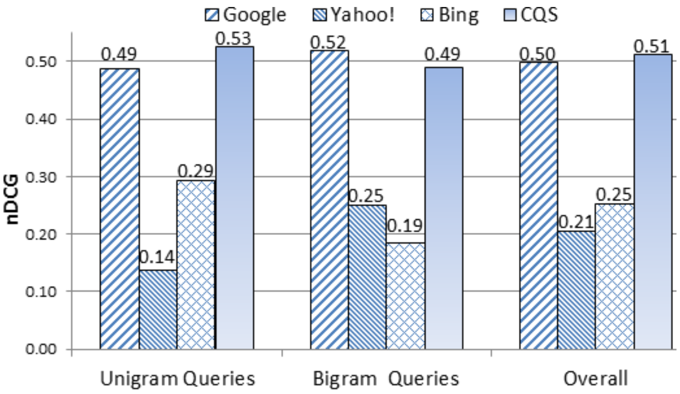

Fig. 1. The nDCG scores for Google, Yahoo!, Bing, and CQS respectively determined using their top-4 suggestions against the gold standards

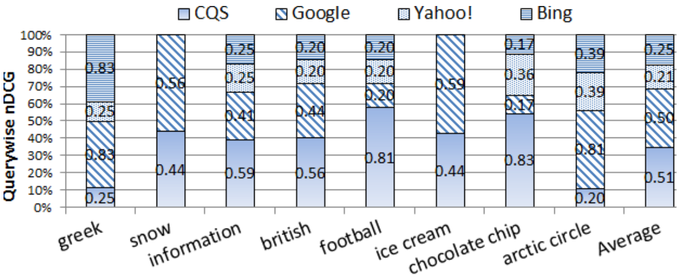

Fig. 2. Per-query distribution of nDCG scores for CQS, Google, Yahoo!, and Bing, respectively

explicitly articulated in [7], [18], [17]. Furthermore, datasets presented in [7], [18], [17] are not available to the research community. For this reason, fair comparisons between CQS and these children QS modules are not possible.

3) Feature Evaluation: To determine which feature(s) of CQS, as presented in Sections III-B - III-H, contribute(s) the most in making children suggestions, we relied on a test dataset, denoted *TestData*. TestData consists of 12,000 labeled instances, which include phrases and their corresponding scores computed for each of CQS features. These phrases are uniformly distributed among children/non-children categories and are disjoint from the instances presented in Section III-I. We analyzed the capability of each (group of) feature(s) in distinguishing (non-)children phrases, which are potential candidate queries.

To further demonstrate the correctness of the features considered by CQS, we computed the nDCG scores of each feature using the dataset discussed in Section IV-A, in addition to the overall nDCG score of CQS computed using backpropagation as a combination strategy. As shown in Figure 3, each individual feature underperforms the combined features used by CQS. By combining all the features, CQS takes the advantage of their individual strengths and greatly improves the relevance and suitability of its generated suggestions for children. The overall nDCG of CQS as shown in Figure 3, which is 0.51, is a statistically significant improvement ($p < 0.001$) over the nDCG score achieved by any single feature.

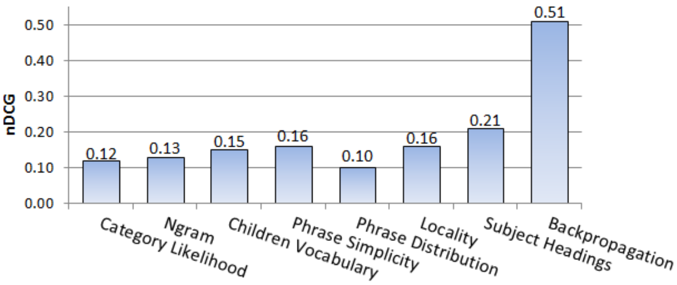

Fig. 3. Performance evaluation of (each of the features of) CQS

C. Mechanical Turk's Evaluation

As previously stated, there are no benchmark datasets that can be used to assess the performance of QS modules for children. For this reason, we turned to Mechanical Turk[13] to conduct empirical studies that allow us to further evaluate the performance of CQS. We relied on Amazon's Mechanical Turk, since it is a "marketplace for work that requires human intelligence", which allows individuals or businesses to programmatically access thousands of diverse, on-demand workers and has been used in the past to collect user feedback on various information retrieval tasks.

1) Relevance of CQS-generated Suggestions: We conducted a survey on Mechanical Turk in which we asked appraisers to examine a set of test queries and their corresponding suggestions created by CQS. For each query Q, appraisers were required to identify, among a provided set of four suggestions generated by CQS for Q, the ones (if any) that were suitable and relevant for children.

While the ten test queries (with five queries in each evaluation form) included in the survey, which are "Disney", "Lego", "Pet", "Transformers", "National football", "Art", "Dog", "Minecraft", "Video", and "Basketball player", were selected among the query set introduced in Section IV-A that address varied topics of interests for children at diverse school grade levels, the corresponding suggestions were generated using CQS. The goal of this survey is to quantify the degree to which queries suggested by CQS are appealing to children (from the adults' points of view). Based on the feedback collected through Mechanical Turk in July 2014, we have observed that, on the average, (close to) 50% of the recommendations generated by CQS were deemed suitable for children.

We are aware that each Mechanical Turk appraiser must be over 18 years old. We solicited appraisers of all walks of life and assessed the performance of CQS by separating the opinions of appraisers known to be educators or parents of young children,[14] who have a more direct knowledge on the interests/preferences of children in terms of selecting suitable query suggestions, from the opinions of general appraisers. The accuracy ratios computed based on parents/educators'

[13] https://www.mturk.com/mturk/welcome

[14] Mechanical Turk appraisers were asked to voluntarily answer a question which inquired whether they were parents/educators. Overall, 57% of the appraisers who assessed the performance of CQS were parents/educators.

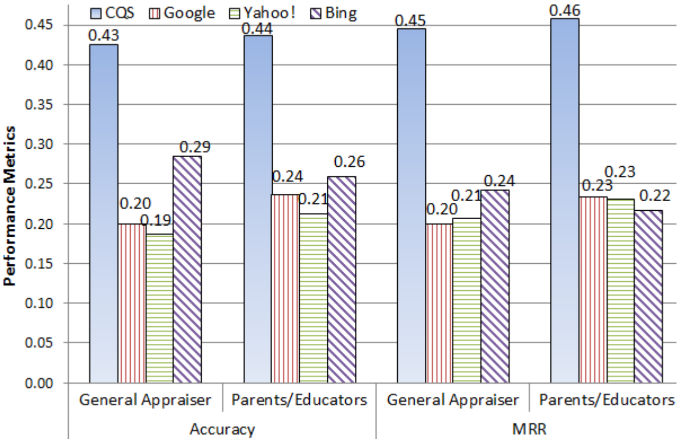

Fig. 4. Evaluations based on Mechanical Turk appraisers' responses

responses yield a statistically significant improvement ($p < 0.05$) over the ones based on the responses of the general appraisers, which is determined using the Wilcoxon test.

2) Evaluations on QS Modules: We also turned to Mechanical Turk to validate our claim that queries suggested by CQS for children are more favorable than the ones generated by Google, Yahoo!, and Bing. To verify this claim, we conducted another survey on Mechanical Turk, which requested the appraisers to identify for each test query the *two* suggestions that, to the best of their knowledge, were most suitable for children. The test queries in this survey are the same queries as presented in Section IV-C1, and the corresponding suggestions are the top-2 suggestions generated by CQS, Bing, Google, and Yahoo!. (Note that due to overlapped suggestions offered by the four QS modules, there can be less than eight suggestions for each of the test queries.) We treated the two suggestions chosen for each test query Q by each appraiser as the *gold standard* for Q. Based on the chosen suggestions, we computed the *accuracy ratio* and *Mean Reciprocal Rank* (MRR). While the former quantifies the proportion of relevant suggestions generated by a QS module, the latter computes the average ranking position of the first relevant suggestion provided by the corresponding QS module.

The accuracy and MRR scores computed according to the responses collected during the month of July of 2014 are shown in Figure 4. The results (which are statistically significant with $p < 0.05$) show that appraisers often preferred children query suggestions provided by CQS over the suggestions created by Google, Yahoo!, or Bing. These findings are consistent among the 65% of appraisers who were either educators or parents of young children.

V. Conclusions

A Statistical report published in 2012 shows that 76% of children searched information on the Internet [10]. To enhance the children's web search experience, it is critical to design a query suggestion module tailored towards children's information needs. In this paper, we have proposed a query suggestion module, called CQS, to suggest queries for children. Instead of following existing query suggestion approaches that rely on frequently-used queries in query logs or children's query suggestion approaches that count on snippets and titles given by search engines to (obtain tags that can be used to) generate candidate suggestions, CQS considers sentences in children's writing, children's vocabulary/phrases, simplicity of words, children's subject headings, and children's categories (i.e., subject areas) extracted from various children's websites, to generate simple and comprehensible phrases as query suggestions. The novelty of CQS is its reliance on freely and easily accessible online content/documents written by or for children. These resources not only allow CQS to generate age-appropriate suggestions, but they also offer different quantitative measures to be considered for capturing children's information needs, creating cohesiveness and simplicity of keywords in suggestions, and enriching the coverage of various topics in suggestions. Experiments conducted to evaluate the performance of CQS demonstrate the correctness of the design methodology of CQS and show that children prefer suggestions offered by CQS over Google/Yahoo!/Bing's.

References

[1] R. Benjamin. Reconstructing Readability: Recent Developments and Recommendations in the Analysis of Text Difficulty. *Educational Psychology Review*, 24(1):63–88, 2012.

[2] S. Bhatia, D. Majumdar, and P. Mitra. Query Suggestions in the Absence of Query Logs. In *ACM SIGIR*, pages 795–804, 2011.

[3] D. Bilal, S. Sarangthem, and I. Bachir. Toward a Model of Children's Information Seeking Behavior in Using Digital Libraries. In *IIiX*, pages 145–151, 2008.

[4] G. Cormack, C. Clarke, and S. Buettcher. Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods. In *ACM SIGIR*, pages 758–759, 2009.

[5] W. Croft, D. Metzler, and T. Strohman. *Search Engines: Information Retrieval in Practice*. Addison Wesley, 2010.

[6] C. Eickhoff, P. Dekker, and P. de Vries. Supporting Children's Web Search in School Environments. In *IIIX*, pages 129–137, 2012.

[7] C. Eickhoff, T. Polajnar, K. Gyllstrom, S. Torres, and R. Glassey. Web Search Query Assistance Functionality for Young Audiences. In *ECIR*, pages 776–779, 2011.

[8] K. Flowers and N. Cookie. Knowledge Structure and Subject Access. In *ACM SIGCHI*, page 3, 1992.

[9] T. Jayalakshmi and A. Santhakumaran. Statistical Normalization and Back propagation for Classification. *Computer Theory and Engineering*, 3(1):1793–1798, 2011.

[10] Y. Kammerer and M. Bohnacker. Children's Web Search with Google: The Effectiveness of Natural Language Queries. In *IDC*, pages 184–187, 2012.

[11] J. Lee. Analyses of Multiple Evidence Combination. In *ACM SIGIR*, pages 267–276, 1997.

[12] J. Lennon, P. Koleff, J. Greenwood, and K. Gaston. The Geographical Structure of British Bird Distributions: Diversity, Spatial Turnover and Scale. *Animal Ecology*, 70:966–979, 2001.

[13] T. Mitchell. *Machine Learning*. McGraw Hill, 1997.

[14] U. Ozertem, O. Chapelle, P. Donmez, and E. Velipasaoglu. Learning to Suggest: A Machine Learning Framework for Ranking Query Suggestions. In *ACM SIGIR*, pages 25–34, 2012.

[15] M. Pera and Y.-K. Ng. What to Read Next?: Making Personalized Book Recommendations for K-12 Users. In *RecSys*, pages 113–120, 2013.

[16] S. Torres, D. Hiemstra, and P. Serdyukov. Query Log Analysis in the Context of Information Retrieval for Children. In *ACM SIGIR*, pages 847–848, 2010.

[17] S. Torres, D. Hiemstra, I. Weber, and P. Serdyukov. Query Recommendation for Children. In *ACM CIKM*, pages 2010–14, 2012.

[18] S. Torres, D. Hiemstra, I. Weber, and P. Serdyukov. Query Recommendation in the Information Domain of Children. *JASIST*, 65(7):1368–1384, 2014.