Personalizing Web Search Results by Reading Level

Kevyn Collins-Thompson
Paul N. Bennett
Ryen W. White
Microsoft Research
One Microsoft Way
Redmond, WA 98052
{kevynct, pauben, ryenw}
@microsoft.com

Sebastian de la Chica
Microsoft Bing
One Microsoft Way
Redmond, WA 98052
sedelach@microsoft.com

David Sontag
Microsoft Research
New England
One Memorial Drive
Cambridge, MA 02142
dsontag@csail.mit.edu

ABSTRACT

Traditionally, search engines have ignored the reading difficulty of documents and the reading proficiency of users in computing a document ranking. This is one reason why Web search engines do a poor job of serving an important segment of the population: children. While there are many important problems in interface design, content filtering, and results presentation related to addressing children's search needs, perhaps the most fundamental challenge is simply that of providing relevant results at the right level of reading difficulty. At the opposite end of the proficiency spectrum, it may also be valuable for technical users to find more advanced material or to filter out material at lower levels of difficulty, such as tutorials and introductory texts.

We show how reading level can provide a valuable new relevance signal for both general and personalized Web search. We describe models and algorithms to address the three key problems in improving relevance for search using reading difficulty: estimating user proficiency, estimating result difficulty, and re-ranking based on the difference between user and result reading level profiles. We evaluate our methods on a large volume of Web query traffic and provide a large-scale log analysis that highlights the importance of finding results at an appropriate reading level for the user.

Categories and Subject Descriptors: H.3.3 [**Information Retrieval**]: Retrieval Models; **General Terms:** Algorithms, Experimentation; **Keywords:** Reading difficulty, re-ranking, personalization.

1. INTRODUCTION

Our goal is to show how modeling reading proficiency of users and the reading difficulty of documents can be used to improve the relevance of Web search results. This goal is motivated by the fact that content on the Web is written at a wide range of different reading levels: from easy introductory texts and material written specifically for children, to difficult, highly-technical material for experts that requires advanced vocabulary knowledge to comprehend. Web users differ widely in their reading proficiency and ability to understand vocabulary, depending on factors such as age, educational background, and topic interest or expertise. Web search engines, however, typically use algorithms optimized for the 'average' user, not specific individuals. These facts currently impair the ability of users to carry out successful searches by finding material at an appropriate level of reading difficulty for them.

As an example of the need, and potential, for personalization by reading level, consider the query [*insect diet*], whose actual top-ranked results from a major search engine are shown in Table 1. While a younger child may be more likely to have chosen query terms like [*bug diet*] or [*what do bugs eat?*], the choice of [*insect diet*] could have been made by a child doing a class project, or an elementary school teacher searching for low-difficulty material; parents or middle-school science students may require intermediate material, and more advanced high school and college users may require sites describing entomology research, as the top results do here. Only one result (`www.tutorvista.com`), at rank position eight, is at a low level of difficulty (American school grade level of 5.0). There are several results, however, including the top two, that contain highly technical research-oriented content most appropriate for specialists only. Clearly there is a need for improvement in ranking search results at an appropriate level of reading difficulty.

To address this problem, we describe a tripartite approach based on user profiles, document difficulty, and re-ranking. First, we discuss how snippets and Web pages can be labeled with reading level and combined with Open Directory Project (ODP, `www.dmoz.org`) category predictions. Second, we describe how a user's reading proficiency profile may be estimated automatically from their current and past search behavior. Third, we use this profile to train a re-ranking algorithm that combines both relevance and difficulty in a principled way, and which generalizes easily to broader tasks such as expertise-based re-ranking. In this view, the overall relevance of a document is a combination of two factors: a general relevance factor, provided by an existing ranking algorithm, and a user-specific reading difficulty model, based on the gap between a user's proficiency level and a document's difficulty level. While users may self-identify their desired level of result difficulty, such information may not always be provided. We therefore investigate methods for estimating a reading proficiency profile for users based on their online search interaction patterns.

We structure our study as follows. In Section 2 we review related work in reading difficulty prediction, modeling user expertise, and search systems for children. We then describe three key problems that must be addressed in using reading level to improve Web search relevance: estimating

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.

Rank	URL Domain	Title	Category	Reading Level (Grade level)
1	insectdiets.com	Insect Diet & Rearing Research	Technical/Research	High (10.0)
2	imfc.cfl.rncan.gc.ca	Insect Diet	Technical/Research	High (10.0)
3	www.sugar-glider-store.com	Insect-Eater Diet	Commercial	Medium (7.0)
4	insectdiets.com	Insect Rearing Research	Technical/Research	High (10.0)
5	insectrearing.com	Bio-Serv Entomology Division	Commercial, Technical	Medium (8.0)
6	www.ehow.com	Aquatic Insects & Diet	Educational	Medium (7.0)
7	www.exoticnutrition.com	Insect-eater Diet...	Commercial	Medium (6.0)
8	www.tutorvista.com	Insect diet: Questions & Answers	Educational	Low (5.0)
9	www.encarta.msn.com	Dictionary	Not relevant (empty)	N/A
10	deltafarmpress.com	Producers may put fish on insect diet	Technical/News	High (10.0)

Table 1: Top ten results, in rank order, for the query [*insect diet*] from a commercial search engine, showing the wide variation in reading level that can occur for material retrieved on the same query. Reading level here is estimated using the statistical model described in Section 3.1 and shown in brackets. (Query issued on January 20, 2011.)

a profile of the user's reading proficiency, estimating a document's reading difficulty, and re-ranking algorithms that can effectively combine relevance and difficulty signals to improve search quality. Section 3 then develops the theoretical models and algorithms we use to address each of these three areas. In Section 4 we contrast the search behavior of two groups: users looking for 'kids'-related material and general users, by performing a large-scale log analysis of query, session, and result properties. This analysis helps provide insights into features that may be useful for improving re-ranking by reading-level. Section 5 evaluates the effectiveness of those algorithms according to implicit relevance judgments obtained from actual Web search logs comprising queries, search result clicks, and post-click navigation events. Finally, in Sections 6 and 7 we discuss areas for future research and summarize our findings.

2. RELATED WORK

Personalizing search using reading level touches on several research areas with relevant prior work: search systems for children and students; modeling user expertise and topic familiarity; algorithms for predicting reading difficulty; and personalization or re-ranking methods based on additional user-specific relevance signals.

Effective search systems for children and students have been the focus of increased interest in recent years. Progress in improved user interfaces, crawling and indexing strategies, and models of child-centered relevance are all important in creating a better search experience for children. The PuppyIR project [15] has begun to examine these types of important questions around the design of search engines, especially user interfaces and finding appropriate Web sites for children [7]. To date, however, there has been little, if any, published work on user modeling and re-ranking algorithms based on reading level and their deployment and evaluation in a commercial-scale search engine. Gyllstrom and Moens [10] proposed a binary labeling of Web documents: material for children versus adults, where the label is inferred using a PageRank-inspired graph walk algorithm called AgeRank. For queries, recent work has explored query expansion methods for queries formulated by children [18]. Our approach operates at a lower level and assumes that common operations such as spelling correction have already been performed by the retrieval system to obtain the top-ranked documents, although such additional processing may well be improved using the methods and features that we develop in this paper. Torres *et al.* [17] performed an analysis of the AOL query log to characterize so-called 'Kids' queries. A query was labeled as a Kids query if and only if it had a corresponding clicked document whose domain was listed as an ODP entry in the 'Kids&Teens' ODP top-level category. Our study includes analysis based on the same definition, but we also explore the important dimension of reading level.

An emerging community of human-factors researchers has been focusing on children's experiences in searching for information online. Hirsh [11] carried out a detailed study of the relevance criteria that children employ when searching for information online. The study offers important findings on what criteria matter to children as they search the Web (e.g., topicality was viewed as much more important than authority). Bilal [2] investigated children's cognitive, affective, and physical behaviors as they use the Yahooligans! search engine to find information on a specific search task. Bilal found that childrens' search processes were ineffective and inefficient, as well as of low quality, suggesting that children need to be better trained in how to search – or search engines need to adapt, as we are advocating for in this paper. More recently, Druin *et al.* [6] studied how children use keyword-based Web search and found that children exhibit a number of different roles (e.g., content searcher, distracted searcher) that have implications for the design of new search interfaces tailored toward children's information needs and search behaviors.

Search engines have attempted to adapt to children's use in other ways. For example, many search engines provide a degree of parental control filtering, which blocks inappropriate material. Other sites provide a corpus of high-quality but highly controlled 'white-listed' sites that is curated by human editors, but limited in scope and recency compared to the standard Web. Since these other areas are either existing technologies or outside the scope of our research, we focus on the core problem of improving search result relevance for users, given an estimate of their reading proficiency.

Estimating reading difficulty has been studied for decades, but traditional formulae such as Flesch-Kincaid provide only a crude combination of vocabulary and syntactic difficulty estimates [5]. Recent progress has been made in applying statistical modeling and machine learning to improve reading difficulty estimation for non-traditional documents [5][12] such as Web pages or short snippets. An earlier study on predicting query readability level [14] attempted automatic recognition of reading levels from user queries by using Support Vector Machines with syntactic and vocabulary-based features. A study by Clarke *et al.* [4] showed that the features of snippets provided by search engines have the potential for significant influence on clickthrough behavior. Their study included an ad-hoc readability measure for each snip-

pet: the percentage of words that occur in a list of the 100 most frequent English words. They did not experimentally validate this measure, but it is related to the Dale readability feature we include (Section 3.1.2).

Reading difficulty is related to, but separate from, the *topic familiarity* of a document to a user. Kumaran *et al.* [13] examined re-ranking Web search results with respect to topic familiarity. Their study results suggested that traditional reading difficulty features and formulas such as Flesch-Kincaid alone could not predict whether a document was an introductory or advanced text for a given topic. However, our study retains a focus on general language proficiency, where it is somewhat easier to distinguish between levels. We also do not rely exclusively on traditional difficulty measures as these have been shown to perform poorly for Web texts [5]. Instead, we apply a robust statistical modeling approach that is able to capture detailed, per-word distinctions in usage across grades.

White *et al.* [22] examined how domain expertise influences Web search behavior, and their analysis of numerous query and session features inspired our own choice of features for result re-ranking. Earlier work by Teevan *et al.* [16] examined general personalization based on a variety of user behavior and content-based features, and re-ranked using a simple interpolation formula. A number of studies have investigated combining a base relevance score with auxiliary user- or group-associated features to perform personalized re-ranking. This includes subtopic coverage and novelty [24], which gives a multiplicative re-ranking update but also assumes a particular retrieval model. We emphasize that this paper is about solving multiple problems to provide an end-to-end solution that assumes little about the base ranking score or underlying retrieval model and thus can be applied in a wide variety of systems.

With this research we extend previous work in a number of ways. First, we introduce a document labeling methodology that assigns reading level and ODP category predictions to both documents and the corresponding query-biased snippet that searchers use when making search engine result page (SERP) clickthrough decisions. These labels play an important role in re-ranking and the evaluation of re-ranking. Second, we describe how a user's reading proficiency profile may be estimated automatically from their current and past search behavior. Finally, we train and evaluate at Web scale a re-ranking algorithm that combines both relevance and difficulty in a principled way, and allows proficiency profile features to be used to re-rank Web search results.

3. PROBLEM FORMULATION

There are three key problems to solve in order to incorporate reading level as a relevance signal for Web search: (i) estimating reading level of documents and snippets, (ii) estimating reading proficiency of users, and (iii) ranking documents based on reading level of users and documents. We now treat each of these in turn.

3.1 Estimating reading difficulty of documents and snippets

We represent the reading difficulty of a document or text as a random variable R_d taking values in the range $[1, 12]$. In this study, these values correspond to American school grade levels, although they could easily be modified for finer or coarser distinctions in level, or for different tasks or populations. We computed reading level predictions for two different representations of a page: the combined title and summary text, which we call a 'snippet', that appears for that page in the search engine results page; and the full body text extracted from the HTML of the underlying page. The snippet text and full-page text are complementary sources of information. While the snippet provides a relatively short sample of content for the underlying page, it is query-specific, and is what users see in choosing whether or not the corresponding page may be relevant and thus whether to click the result. The full-page text, in contrast, is not affected by a query, and is what users see *after* clicking on a result hyperlink on the search result page. We were interested in the interaction of snippet and page level estimates as well as their individual effectiveness. Indeed, we discovered a strong interaction of snippet-page difficulty difference with page dwell time (see Section 4.4 for more details).

3.1.1 Prediction using language models

The reading difficulty prediction method that we use for this study, summarized in this section, has been shown to be effective for both short, noisy texts, and full-page Web texts. Unlike traditional measures that compute a single numeric score, methods based on statistical language modeling provide extra information about score reliability by computing the likely *distribution* over levels, which can be used to compute confidence estimates. Moreover, language models are vocabulary-centric and can capture fine-grained patterns in individual word behavior across levels. Thus, they are ideal for the noisy, short, fragmented text that occurs on the Web in queries, titles, result snippets, image or table captions. Because of this short, noisy nature of Web snippets we applied a robust, vocabulary-oriented reading difficulty prediction methods that is a hybrid of the original smoothed unigram approach [5] and a more recent model based on estimated age of word acquisition [12].

Following [12], we say that a document D has an (r, s)-reading level t if at least s percent of the words in D are familiar to at least r percent of the general population[1]. We say that a word has r-acquisition level $\mu_w(r)$ if r percent of the population have acquired the word by grade μ_w. For a fixed (but large) vocabulary V of distinct words, we define an approximate age-of-acquisition for all words $w \in V$ using a truncated normal distribution with parameters (μ_w, σ_w). We estimated (μ_w, σ_w) from a corpus of labeled Web content from [5]. With these word parameters, we can then apply the above definition of (r, s)-readability. To compute the readability distribution of a text passage, we accumulate individual word predictions into a stepwise cumulative density function (CDF). Each word contributes in proportion to its frequency in the passage. The reading level of the text is then the grade level corresponding to the s-th percentile of the text's word acquisition CDF. Details are given in Kidwell *et al.* [12].

3.1.2 Prediction using traditional semantic variables

We also compute a traditional measure of vocabulary-based difficulty: the fraction of unknown words in a query or snippet (which we call the 'Dale readability measure') relative to the Dale 3000 word list, which is the semantic component of the Dale-Chall reading difficulty measure [3].

3.1.3 Category prediction

We used automatic classification techniques to assign a category label to each page. A logistic regression classifier using an L2 regularizer was trained over each of the ODP topics identified: for the experiments reported in this study we had available the 219 topical categories from the top two levels of the ODP hierarchy, although we focus primarily on the Kids&Teens category in this study. Our classi-

[1]In that study, the authors found that setting $r = 0.80$ and $s = 0.65$ provided the greatest reduction in training error, and so we use the same settings for r and s here.

fier assigned one or more labels using a similar approach to that described in [1]. We chose to use the Open Directory Project (ODP) for classification because of its broad, general purpose topic coverage; the availability of reasonably high-quality training data; and of special interest is the Kids&Teens category, which was created in Nov. 2000 with its own set of editorial guidelines with the goal of providing kids-safe content for the under-18 age group.

3.2 Estimating reading proficiency of users

One approach is to have users self-identify their level of reading proficiency. For example, this is the approach that Google has recently used as part of their advanced search tools: users may choose to filter the results to show only 'basic', 'intermediate', or 'advanced' reading level. This approach has as its advantages both simplicity of user interaction and transparency of search behavior. One disadvantage of such an approach is that it may be difficult for users to properly calibrate their reading level. Also, reading proficiency may change over time, and it may be dependent on the actual query issued. Thus, we can consider ways to construct a reading proficiency profile automatically from search behavior. This may include the previous queries and click-throughs in either the session, or the user's long-term history. We discuss one such approach now.

To match the difficulty distribution $p(R_d)$ of a document given in Sec. 3.1.1, we define a proficiency profile for user u to be a distribution $p(R_u)$ over levels, allowing us to compute the probability that a user understands a document. As with the document, R_u can take values in the range $[1, 12]$. Here, we give a generative model for a user's search sessions that we will use in estimating $p(R_u)$ from a user's search behavior. Although the prior distribution $p(R_u)$ is assumed to be the same for all of a user's search sessions, the posterior $p(R_u \mid \text{query})$ depends on the query and may differ between sessions. Let Q denote the set of queries that the user has issued in this session, and let D_q denote the documents that the user clicks on in response to the query. We make the assumption that a user clicks and *dwells* on a document only if they can understand it - or more generally, if they like the page because the reading level is appropriate to their intent[2]. A session is generated as follows:

1. $r_d \sim p(R_d)$ (given in Sec. 3.1.1)
2. $r_u \sim p(R_u)$ (to estimate)
3. For all $q \in Q$:
   (a) $q \sim p(\text{query} \mid r_u)$
   (b) For all $d \in D_q$:
   SAT-click $= 1 \sim p(u \text{ likes reading level of } d \mid r_u, r_d)$

Typically, users will like reading documents whose difficulty is at or below their average proficiency level, and dislike documents more and more as their difficulty increases above this average level. To reflect this, we may choose a definition such as

$$p(u \text{ likes level of } d \mid r_u, r_d) = \exp\left(-\max(0, r_d - r_u)\right). \quad (1)$$

However, the appropriate form of Eq. 1 may vary depending on the user and their intent. For example, some expert users looking for high-difficulty technical material may actually want to penalize easier documents, to avoid introductory material. Other users, such as students, may also want material that is neither too difficult nor too easy relative to their level. Thus, we can instead define

$$p(u \text{ likes level of } d \mid r_u, r_d) = \exp\left(-(r_d - r_u)^2\right).$$

In this study we use Eq. 1 but exploring other forms or learning this from data is a topic for future work.

The distribution $p(\text{query} \mid R_u)$ would ideally be a language model that is directly estimated using query logs. An alternative is to use the language model that we developed for document classification. However, query readability may be very different from document readability. The distinction is that the words a user *recognizes* may be different from the words that they choose to use in queries. A much simpler approach is to simply model the length of the query, ignoring the actual words. As we show in Figure 1, the query length can be informative about a user's reading proficiency.

We use the above ideas to compute a session-based query difficulty feature based on the average reading level of the satisfied clicks that a user enacts in previous queries within the session.

If we also have access to the set of web sites that a user frequently visits, we could use these to help predict the reading proficiency of a user. For example, Club Penguin and Funbrain are two sites often visited by children, but less frequently visited by adults. This is a topic for future work.

3.3 Re-ranking based on reading difficulty

To learn effective re-rankings and to explore the importance of features related to reading level, we use the LambdaMART [23] algorithm, a state-of-the-art ranking algorithm based on boosted regression trees. Compared with other ranking approaches LambdaMART is typically more robust to sets of features with widely varying ranges of values, such as categorical features. Since LambdaMART produces a tree-based model, it can be used as a feature selection algorithm or to rank features by their importance (Section 5.1.2). Based on the results of our analysis in Section 4, along with evidence on the effectiveness of specific features gathered by other authors in previous studies of expertise [22], we chose the following set of features for study.

Query features. These features rely only on the query string and include query length in characters and query length in space-delimited words.

Query/session features. If previous queries were present in a session, we estimate a dynamic reading level for a user by taking the average reading level of the clicked snippets from previous queries in the same user search session. Because of the sparse nature of clicks we also compute a confidence value for this query level that increases with the sample size of clicked snippets. We also include a measure of the length of a session, in terms of the number of previous queries.

Snippet features. We compute the estimated reading difficulty of a page's snippet using the algorithm described in Section 3.1.1, and the Dale-Chall semantic variable from Section 3.1.2. We also include the relative difficulty of the snippet compared to the levels of the other top-ranked result snippets: snippets are sorted by descending reading level, and then the reciprocal rank of the snippet is computed with respect to that ranking.

Page features. Using the same reading level prediction algorithm used for snippets, we compute reading difficulty for the body text of the Web page corresponding to a snippet. We also include a confidence feature for this prediction.

Snippet-page features. We compute the (signed) difference between the full page reading level and the snippet reading level.

Query-page features. These features capture the strength of a query-document match: the main signals here are the

[2]A widely-used dwell time satisfaction threshold in Web search is 30 seconds, termed a 'satisfied' click or SAT-click. In reality, dwell time may vary with a number of factors, such as the age or reading proficiency of the user.

Source	Feature Name	Description
Query	query_char_len	Query length (in characters)
	query_word_len	Query length (in words)
Query (Session)	session_user_level	Session-based user reading level estimate
	session_user_level_confidence	Confidence estimate for user reading level
	prev_queries_in_session	Number of previous queries in current search session
Snippet	snippet_difficulty	Reading level of snippet
	relative_snippet_difficulty	Relative snippet difficulty in top 10 results
	dale_snippet_difficulty	Dale difficulty level
Page	full_page_difficulty	Reading level of page body text
	full_page_difficulty_confidence	Confidence level for full-page reading level
Snippet-Page	snippet_page_level_difference	Difference between reading levels of snippet and full page
	snippet_page_difference_confidence	Confidence level for snippet-page level difference
Query-Page	norm_production_score	Normalized ranker score for a page
	reciprocal_rank_score	Reciprocal rank of page
Query-Snippet	snippet_query_diff	Signed difference in reading level between query and snippet
	snippet_query_diff_abs	Absolute difference in reading level between query and snippet

Table 2: The features used by LambdaMART reranking. Features that potentially make use of previous queries in a session are denoted (Session).

normalized ranker score from the search engine, and the reciprocal rank of a page in the top ten results.

Query-snippet features. We compute the absolute and signed differences between the estimated user reading level and the estimated snippet reading level.

Table 2 summarizes the set of features used for ranking.

4. LARGE-SCALE QUERY LOG ANALYSIS

In this section we perform a summary analysis of search log data in order to contrast the properties of Kids and non-Kids users and data sources. We conjectured that Kids sessions and queries would exhibit differences, especially with respect to the reading level of preferred result snippets. Our analysis also estimates coverage for Kids queries to assess the likely impact of personalization in this area, and shows that having both snippet- and page-level reading level predictions is valuable.

4.1 Data set and evaluation methodology

The primary source of data for this study was a proprietary data set containing the anonymized logs of URLs visited by users who consented to provide interaction data through a widely-distributed browser plug-in. The data set contained browser-based logs with both searching and browsing episodes from which we extract search-related data. These data provide us with examples of real-world searching behavior that may be useful in understanding and modeling kids-related search. Log entries include a browser identifier, a timestamp for each page view, and the URL of the Web page visited. To remove variability caused by geographic and linguistic variation in search behavior, we only include log entries generated in the English-speaking United States locale. The results described in this paper are based on URL visits during the first week of October 2010 representing millions of Web page visits from hundreds of thousands of unique users. From these data we extracted search sessions from a major commercial Web search engine, using a session extraction methodology similar to [20]. Search sessions begin with a query, occur within the same browser and tab instance (to lessen the effect of any multi-tasking that users may perform), and terminate following 30 minutes of user inactivity.

From these search sessions we extracted search queries and for each query, we obtained the top ten search results retrieved by the Web search engine and the titles and the snippets for each result that were displayed on the search engine's result page at query time. We then estimated the grade level distribution for each of those results using the snippet text and the full text of the corresponding Web page, per the method described in Section 3.1.1.

In addition, we also obtained binary relevance judgments for each result in the top 10 using a methodology similar to that in [9]. We define a satisfied (SAT) click in a similar way to previous work [19] (i.e., with either a dwell time post-click of 30 seconds or the last SERP click in the session). Advantages of these log-based judgments are that many judgments can be easily gathered, and that they are personalized to the user and the query, which is important in the evaluation of personalized search algorithms. With these judgments, we define two evaluation scenarios: 'Last-SAT', which assigns a positive judgment to one of the top 10 URLs if it is the *last* satisfied SERP click in the session (by click time); and 'All-SAT', which assigns a positive judgment to *any* satisfied click in a session. (In both cases, the remaining top-ranked URLs receive a negative judgment.) While Last-SAT has been shown to be highly indicative of a user's goal [9], we also examine All-SAT since informational queries, which are more likely to have multiple relevant results, may exhibit different performance qualities.

For the Last-SAT case, the Mean Reciprocal Rank (MRR) of the positive judgment is used to evaluate retrieval performance before and after re-ranking. For All-SAT, we evaluate average precision on the top 10 results.

Queries for which we cannot assign a positive judgment to any top-10 URL are excluded from the feature set. We also excluded queries corresponding to twelve very high frequency navigational queries[3]. Although there are benefits to including these queries, such as detecting engine switching behavior [21], their highly predictable nature across users makes them less interesting for a personalization study, and removing them also greatly reduces data processing demands.

After this filtering, 759,671 queries remained. Of these, 555,048 queries (73%) had no previous queries in the same session and 205,623 had at least one previous query, leaving 27% of queries potentially amenable to session-based improvements. The 27% is low, since previous work has found

[3]facebook, google, myspace, gmail, bing, yahoo, yahoomail, craigslist, youtube, ebay, aol, hotmail.

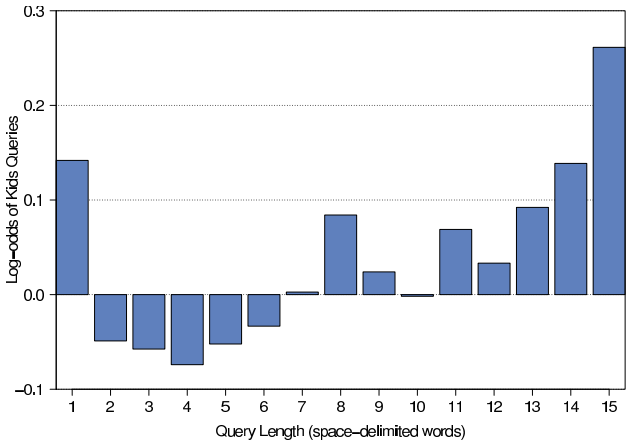

Figure 1: Log-odds of query length (in words) for Kids vs. All queries, showing that single-word queries and queries with eight words or more are more likely for Kids queries than general queries.

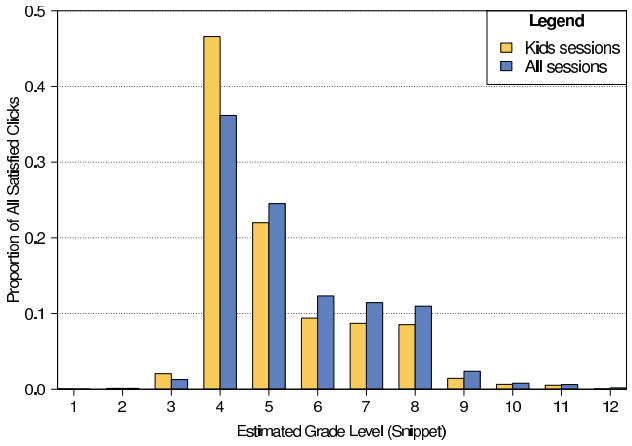

Figure 2: The estimated reading difficulty of snippets for pages with satisfied clicks, for Kids sessions and all sessions.

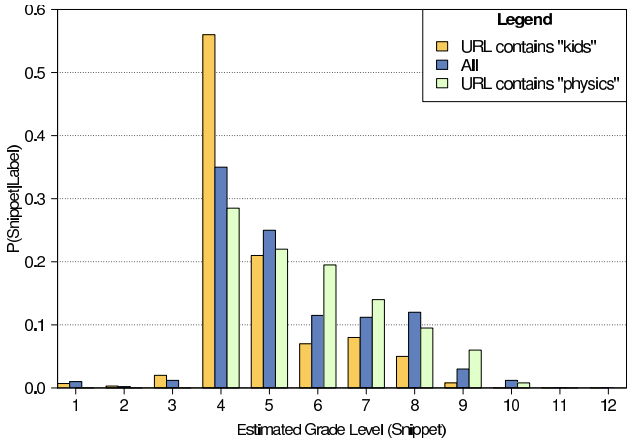

Figure 3: The estimated reading difficulty of snippets of pages with different substrings in their URL: 'kids' and 'physics' have sharply different difficulty distributions, with the All distribution lying in between.

that 60% of search sessions contained multiple queries [20]. This may be explained in part by differences in the search behavior of users issuing Kids queries and perhaps even the removal of the small set of high-frequency navigational queries.

4.2 Query properties

We used our development set (Section 5.1) to extract 29,498 total queries and 19,601 unique queries having at least one click on a SERP result labeled with the Kids&Teens ODP category. We compared this to the properties of the full development set, with no filtering, which we call the All queries set.

Many of the most frequent queries found in a recent AOL query log analysis by Torres *et al.* [17] also appeared highly ranked in our list in various forms, such as [*nick jr*], [*starfall.com*], [*coloring pages*] and [*dora*]. We found that the distribution of query lengths for Kids sessions was also similar to Torres *et al.* , with longer queries being more likely on average, as shown in Figure 1, possibly due to a greater frequency of natural language queries. Query lengths above the line (positive log-odds) are more likely for Kids queries, and those below the line are more likely for All queries. Our data also showed that single-word queries were more likely, as the result of a larger proportion of navigational queries.

4.3 Reading level distribution of snippets

Of all SERP clicks, 3.6% were satisfied clicks in our Kids sessions sample. The distribution of snippet reading levels for all SAT clicks for Kids and All queries is shown in Figure 2. The figure shows that the estimated reading difficulty of snippets associated with satisfied clicks is skewed toward lower grade levels for queries coming from Kids sessions, compared to all queries.

As an example of the connection between simple syntactic features within a page's URL string and the estimated reading level of its snippet, we extracted URLs containing the substring 'kids' and another set containing the substring 'physics'. Figure 3 shows that the two reading level distributions are very different: the 'physics' distribution is shifted toward a much higher level of difficulty. The All distribution lies in an intermediate area between the two. Our comparison here makes the assumption that documents about physics are of higher difficulty whereas the vast majority of kids documents are of lower difficulty. Indeed, we can see this reflected in the actual distributions in Figure 3.

Note that Figure 2 is computed from user clicks while Figure 3 is computed from properties of URL strings, which are user-agnostic. Thus, Figure 2 shows whether users take reading difficulty into consideration in deciding whether they like a page. For example, it could be that for the query [*physics*] all users click on documents with reading level 10 – this would be captured in the type of histogram shown in Figure 2 but not in Figure 3.

4.4 Snippet-page difficulty gap predicts average dwell time

To assess the importance of using both snippet and page representations of reading level, we examined the correlation between the *difference* in the predicted reading levels of a SERP snippet vs. the full text of the corresponding Web page, and the average user dwell time (in seconds) for that Web page. We found a strong relationship between these quantities: for example, the more difficult the underlying page is, compared to the clicked snippet for that page, the

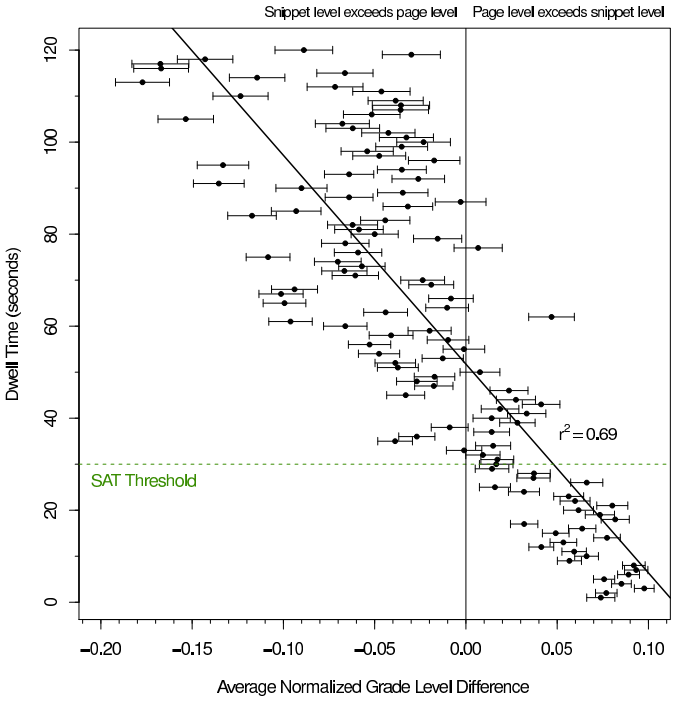

Figure 4: Average user dwell time is strongly predicted by the difference between the query-specific snippet reading difficulty and the underlying page reading difficulty. As actual page difficulty increases relative to the displayed snippet, user will be more likely to abandon that page quickly. Error bars denote 95% confidence intervals.

more likely it became that the user would be unsatisfied and leave that page quickly (e.g., spend less than 30 seconds reading it). This is summarized in Figure 4. The means and confidence intervals for each dwell time, conditioned on snippet-page reading level difference, are shown in Figure 4 and were computed by the bootstrap method [8], using 100 iterations over the search log data. Each iteration sampled $N = 630,000$ click instances with replacement. The Pearson correlation coefficient over all dwell times of 120 seconds or less was $r^2 = 0.69$: a strong signal about the behavior of a user population. This is an important finding with a number of implications. First, it provides clear evidence that modeling *both* snippet and page level provides a valuable generic relevance signal. As we show later in Section 5.1.2, the difference between snippet and page level does have high relative weight among effective ranking features. Second, it suggests that snippet quality could be improved by constraining a snippet's expected reading level to closely match the underlying page reading level.

4.5 Coverage for Kids-related queries

To understand what percentage of queries may be related to children's information needs and thus might be candidates for reading level-based personalization, we extracted all URLs from the Kids&Teens categories in the ODP and considered them to be children-friendly URLs given the ODP editing guidelines for these categories.

We used a query-click bipartite graph from anonymous search data containing a node for each query and each clicked URL for one month of 2010 query traffic in the English-speaking United States locale from a large commercial search engine. Query nodes are connected via edges to the URLs users actually clicked on for those queries during a search session. We looked up the ODP children-friendly URLs in this graph and extracted the queries leading to clicks on these URLs. While it is impossible to know if these queries were issued by users in the age range targeted by ODP with the Kids&Teens categories, we used this method to characterize such queries as having some degree of Kids&Teens intent. Using this technique, we estimated the coverage of Kids&Teens queries to be 13.62% of non-bot traffic. Summary statistics are shown in Table 3.

ODP Kids&Teens URLs	45,879
Unique Queries (with clicks on these URLs)	1,360,341
Impressions from these queries	286,174,932
Coverage	13.62%

Table 3: Kids&Teens-related traffic statistics (one month).

The dataset used to identify queries with Kids&Teens intent includes impressions collected from multiple revisions of the underlying Web search results ranking algorithm, meaning that there was significant variance Web search results returned for the same queries. As a result, we consider the 13.62% coverage presented here as a ceiling of opportunity for personalization of the search experience for children. In addition, our earlier analysis (from Section 4.1) shows that at least 27% of those Kids&Teens queries might be amenable to additional session-based improvements.

5. EVALUATION

Our evaluation section consists of three parts: re-ranking results, an analysis of relative feature importance, and robustness evaluation.

5.1 Re-ranking performance

In this section we confirm the effectiveness of re-ranking Web search results using reading level features. We examine not only basic retrieval performance in Section 5.1.1 but also the relative importance of different features (Section 5.1.2), and the robustness in rank gains and losses (Section 5.1.3).

We partitioned our log data as described in Section 4 randomly into two sets: 25% was used as a development set for LambdaMART parameter optimization, and the remaining 75% was used for training/test splits using 10-fold cross-validation. To avoid bias toward longer sessions, these sets were further subsampled to one random query per session.

5.1.1 Basic retrieval performance

In order to account for the possibility of query sets with relevant documents biased lower in the ranking (such as difficult queries), we trained a learned rank-only baseline reranking model using two features: the commercial search engine ranker score, and the rank position of a document. This model always performs at least as well as the original commercial ranker score. As described in Section 4, our main evaluation measures are the change in Mean Reciprocal Rank (MRR) of the last satisfied click ('Last-SAT'), and Mean Average Precision (MAP) using all satisfied clicks ('All-SAT'). In practice, because the vast majority of queries have only one click, there turns out to be almost no difference in performance between Last-SAT and All-SAT on this dataset, but we report both numbers for the full query set experiment for completeness. Because of the proprietary nature of our search engine system we do not report absolute MRR, but instead report relative change in MRR on a point scale from 0 to 100: $(MRR_{MODEL} - MRR_{BASE}) \cdot 100$, where MRR_{BASE} is the learned rank-only baseline MRR for

Experiment	Last-SAT (MRR Change)	All-SAT (MAP Change)
Rank-only Baseline	0.0	0.0
Query+Session	$+0.7^{\star}$	$+0.6^{\star}$
Query+Session+Page	$+0.8^{\star}$	$+0.8^{\star}$
Query+Session+Snippet	$+1.0^{\star}$	$+0.9^{\star}$
All Features	$+1.2^{\star}$	$+1.1^{\star}$

Table 4: Summary of the relative performance for different feature sets on full query set, in terms of change points in MRR (Last-SAT) and MAP (All-SAT). The Rank-only Baseline is used as the baseline for comparison and thus set to 0.0. Superscripts $\star$ and $+$ denote significance of change at $p < 0.01$ and $p < 0.10$ respectively using a paired t-test.

Query subset	Num. queries	% Total	Baseline Rel. MRR	Model Rel. MRR	Gain
Kids	15,796	4.4%	-4.1	-3.1	$+1.0^{\star}$
Science	23,059	6.8%	-9.0	-4.7	$+4.3^{\star}$
Sports	41,139	11.6%	$+7.2$	$+8.2$	$+1.0^{\star}$
Health	21,581	6.1%	-7.3	-7.3	0.0
All	545,255	100%	0.0	$+1.2$	$+1.2^{\star}$

Table 5: Summary of gains in MRR for Last-SAT click from re-ranking with reading-level features. To show the relative difficulty of different query sets, the Baseline Relative MRR Change is computed relative to the Rank-only Baseline for all queries in the test set, so that relative MRR for the 'All' query set is zero. Similarly, the Model Relative MRR Change is also computed relative to the Rank-only Baseline for all queries. Gain is the difference between the Model and Baseline. Superscripts $\star$ and $+$ denote significance of change at $p < 0.01$ and $p < 0.10$ respectively using a paired t-test.

all queries. We note that the baseline ranking, representing a highly-tuned commercial search engine, is very competitive, and even a 1-point change in MRR, when statistically significant, is considered a notable gain.

Table 4 compares the performance of the learned rank-only baseline to our model, for different features of the system. Across the full query set, there was a statistically significant +1.2 point MRR gain for Last-SAT clicks, and +1.1 point MAP gain for All-SAT clicks. Starting with the Rank-only baseline, we added a basic set of Query+Session features that made no use of snippet or page-level reading difficulty predictions, which gave an MRR gain of +0.7. To compare the relative utility of page vs. snippet features, we then added either all snippet-based features, or all page-based features, to the basic Query+Session set. The snippet-based features gave a slightly larger gain (+1.0) compared to adding full-page features (+0.8). Finally, we achieved the best overall performance when both snippet and page features were used together with the other features.

To investigate the effect of re-ranking using reading difficulty features on queries of different topics, we chose the ODP Kids&Teens, Science, Sports, and Health categories. For each of these categories, we extracted individual queries having at least one click on a URL belonging to that ODP category[4]. Table 5 shows gains and losses achieved by re-ranking for Kids, Science, Sports, and Health subsets of queries[5]. An increasingly negative (resp. positive) value for Base Relative MRR indicates a harder (resp. easier) retrieval task compared to the All Queries scenario.

Across several useful subclasses of queries, re-ranking with reading level features gave consistent, statistically significant gains in MRR compared to the learned rank-only baseline. The most challenging query set in terms of low baseline (Science) had a rank-only relative baseline of -9.0 compared to the rank-only baseline of the full query set, while the 'easiest' subclass was Sports, whose rank-only baseline change was +7.2 over the full query set. After adding reading level features, our ranking model achieved net MRR point gains of +1.0 for Kids queries, +4.3 for Science queries, and +1.0 for Sports queries. Somewhat surprisingly, Health queries showed no gain - the reasons for this require further study.

It is encouraging to find a class of queries like Science queries that obtain a particularly large benefit from reading level features. The Science category contains a higher proportion of more technical material than most other ODP categories[6] and thus search results for those queries might be expected to have higher reading difficulty entropy, leading to greater potential for personalization. Kids&Teens pages, in contrast, are typically already tailored for children and thus have less variation in reading level, which may explain the reduced effect of reading level features on those queries.

5.1.2 *Relative effectiveness of ranking features*

To understand the relative contribution of our query, session, snippet, and page-based reading difficulty features to re-ranking effectiveness, we used LambdaMART to obtain scores representing relative feature importance. These scores are computed as the average reduction in residual squared error when applying the given feature, averaged over all trees and over all splits. The scores are then normalized relative to the most informative feature, which has a score of 1.000. Table 6 lists the top-scoring features resulting from our main experiment using all features over all queries, averaged over 10 cross-validation splits.

Examining the top five features in this list, the highest-weighted feature - perhaps not surprisingly - was the reciprocal rank score of a page, reflecting the extreme bias toward top-ranked clicks that is typical of Web search results. Rela-

[4]For classifying queries in this way, the click did not have to be a satisfied click.

[5]Note that according to our definition, a query can belong to potentially multiple ODP categories depending on the user's clicked pages, so the query counts in Table 5 do not sum to the total number of all queries. Also, total queries reported here are about 70% of the total queries reported in Section 4, due to combined effects of query subsampling to reduce session length bias, and 10-fold cross-validation split.

[6]For example, the average reading difficulty level estimated for snippets for Kids&Teens pages was 4.53, for Sports was 4.79 and for Science was 5.34.

Feature	Rel. Weight
Reciprocal rank score	1.000
Relative snippet difficulty	0.295
Query length in characters	0.237
Session-based user reading level	0.216
Snippet-page reading level difference	0.207
Dale snippet difficulty level	0.186
Normalized ranker score	0.183
Query-snippet reading level difference	0.142
Query length in words	0.116
Snippet-page level difference confidence	0.081
Snippet reading level	0.076
Page reading level	0.048
Number of previous queries in session	0.030
Session-based user reading level confidence	0.019

Table 6: The relative importance of features computed by LambdaMART reranking for all queries. Feature importance in the right column is the average reduction in residual squared error over all trees and over all splits, relative to the most informative feature.

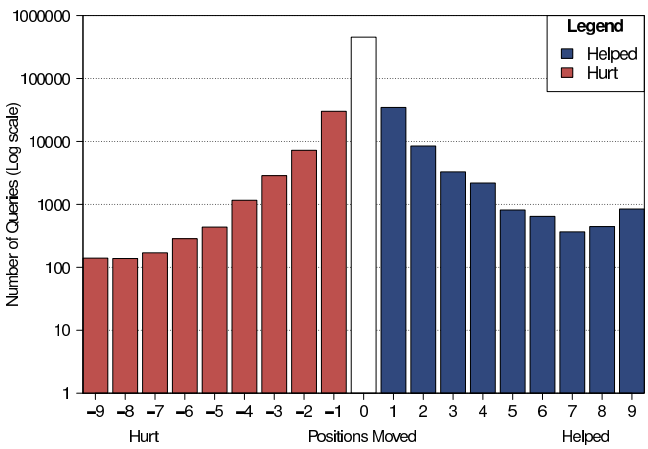

Figure 5: Histogram showing the variance of losses (left tail) and gains (right tail) using re-ranking by reading level features on All queries. The loss or gain in rank position of the last satified click is given on the x-axis. The y-axis denotes the number of queries (note the log scale).

tive snippet difficulty was the second-most important feature (0.295), which matches what we have observed informally: that when picking from a list of otherwise similar results, users tend to pick the snippet with lowest reading difficulty. Query length in characters was more influential (0.237) than query length in words (0.116), although both contributed to the prediction performance. The predictive power of query length is in accordance with the log-odds length distribution shown in Figure 1. Estimating the user's reading proficiency from their previous session queries proved to be another highly-ranked feature (0.216), showing the importance of user-specific personalization. As predicted by our analysis in Section 4.4, the reading level difference between a snippet and its corresponding Web page also carries valuable information, and this is reflected in its position as one of the top 5 features (0.207).

Among the remaining features, the Dale difficulty feature of the snippet, with a relative weight of 0.186, proved moderately effective as a complementary level prediction. Several base features, such as snippet-page level difference, have corresponding confidence features: one of these (for snippet-page level difference) did appear in the top 10 (0.081) while the others had much weaker feature scores.

5.1.3 Robustness of re-ranking

While improving a retrieval metric's average gain across queries is important, in the case of re-ranking an existing set of results, the *variance* of relative gains and losses compared to the initial ranking is also critical to measure. An algorithm may improve average performance, but also increase the number of queries dramatically hurt by re-ranking.

Figure 5 shows the distribution of gains and losses across all queries for the re-ranking model used in Section 3.3, as measured by the change in rank position of the last satisfied click. A robust algorithm is one that is able to achieve good on-average results, while having a minimal failure profile – as measured by a statistic like the number of queries hurt by re-ranking (the left half of histogram in Figure 5). Out of a total of 545,245 queries, 450,921 queries (82.7%) had no change; 51,759 (9.4%) were helped (rank of the last SAT click increased at least one position); and 42,565 (7.8%) were hurt (rank of the last SAT click decreased by at least one rank position). While more work is needed on methods to reduce the loss side of the histogram, the multi-position gains we obtain are encouraging. In particular, for changes of 6 or more rank positions, we achieve a ratio of queries helped to hurt of greater than 2 to 1.

6. DISCUSSION

Personalization by reading level can be considered somewhat orthogonal, but complementary, to other dimensions of personalization such as location, domain expertise, or topical interest. Our study is, to our knowledge, the first of its kind to study and evaluate personalization by reading level as an aspect of large-scale Web search. From our work it is clear that some problems, such as automatically estimating a reliable user profile of reading proficiency, and finding more accurate or effective features for re-ranking, are non-trivial and require further research. Furthermore, while our study included some features defined over a single user session, we believe learning and applying a longer-term user history could also be quite valuable.

Many improvements and interesting directions remain to help users find and understand material appropriate to their needs and reading level. For example, while personalization seeks to adapt the content to the user, we can also consider the reverse goal: adapting the *user* to the content. By this we mean applying models of reading level and vocabulary difficulty to identify *learning opportunities* that would help reduce the gap between the user's reading proficiency and document reading level. For example, a search engine might identify critical 'words to learn' on a topic, such as *bronchitis* for coughs in a home medical care page. A similar idea could be used to help non-native language learners.

The role of reading level features in improving query and document representations is also a rich area for further study. As our findings on dwell time in Section 4.4 suggest, identifying differences in vocabulary or reading level distribution between the different representation streams of a Web page, such as anchor text and captions, and those of the underlying page could help identify problems with snippet or document quality, or even distinctions in usage between different user groups. Also, the variation of dwell times across different ages or reading proficiency levels may also be interesting to investigate further. When processing a likely Kids query, the search engine could provide more child-appropriate snippets

or query suggestions for that query. Since the reading level distribution of a page can be pre-computed and stored in the index, improvements in reranking that require reading level could be applied quickly and reliably. We also believe that the interaction of reading level with topic is important and intend to explore this combination in future work.

In general, reading difficulty level can serve as a valuable contextual signal to improve the ranking of documents presented to individual users during a search session. Document relevance may be improved using the language models in this paper because they can characterize user intent based on vocabulary usage during and across search sessions. Search engines could further leverage these models to personalize all aspects of the user experience, including captions, ads, images, videos, or even pure presentation features such as font size, site previews, and page composition. While the actual impact of such personalization efforts on overall user satisfaction remains a point for further investigation, reading level modeling may provide a powerful tool to bridge the vocabulary gap between search engines and their users.

7. CONCLUSION

We have shown how incorporating reading level features for users and documents can provide a valuable new signal for relevance in Web search. We explored three key problems in improving relevance for search using reading difficulty: estimating models of user reading proficiency, estimating models of result difficulty, and combining relevance and difficulty signals to re-rank based on the difference between user and result reading level. We also provided a large-scale analysis of log data to characterize certain aspects of user behavior and classes of features and queries that were likely to be effective in personalization using reading difficulty predictions. Our results show that statistically significant gains may be obtained with a commercial search engine, even for general queries, by incorporating reading difficulty features. Furthermore, we found specific sub-classes of queries, such as science-oriented queries, that are particularly amenable to improvement. Our work could easily be generalized to model domain expertise in specific subject areas, such as those defined in the Open Directory Project. For example, Web search results could be ranked with the introductory material first, followed by increasingly technical material. Other advances such as level-appropriate query suggestions, result snippets, and site recommendations are also possible.

Acknowledgements

We thank Dan Liebling and Misha Bilenko for technical assistance, and Sue Dumais, Jaime Teevan, and the anonymous reviewers for their valuable suggestions.

8. REFERENCES

[1] P. N. Bennett, K. Svore, and S. T. Dumais. Classification-enhanced ranking. In *Proc. of WWW 2010*, 111–120.

[2] D. Bilal. Children's use of the yahooligans! web search engine: Cognitive, physical, and affective behaviors on fact-based search tasks. *J. Am. Soc. Inf. Sci.*, 51(7):646–665, 2000.

[3] J. Chall, E. Dale. *Readability Revisited: The New Dale Chall Readability Formula*. Brookline Books, 1995.

[4] C. L. Clarke, E. Agichtein, S. Dumais, and R. W. White. The influence of caption features on clickthrough patterns in web search. In *Proc. of SIGIR 2007*, 135–142.

[5] K. Collins-Thompson and J. P. Callan. A language modeling approach to predicting reading difficulty. In *HLT-NAACL*, 193–200.

[6] A. Druin, E. Foss, H. Hutchinson, E. Golub, and L. Hatley. Children's roles using keyword search interfaces at home. In *Proc. of CHI 2010*, 413–422.

[7] C. Eickhoff, P. Serdyukov, and A. de Vries. A combined topical/non-topical approach to identifying web sites for children. In *Proc. of WSDM 2011*, 505–514.

[8] B. Efron, R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, New York, 1994.

[9] J. Gao, W. Yuan, X. Li, K. Deng, and J.-Y. Nie. Smoothing clickthrough data for web search ranking. In *Proc. of SIGIR 2009*.

[10] K. Gyllstrom and M.-F. Moens. Wisdom of the ages: toward delivering the children's web with the link-based agerank algorithm. In *Proc. of CIKM 2008*, 159–168.

[11] S. Hirsh. Children's relevance criteria and information seeking on electronic resources. *JASIST*, 50(14):1265–1283.

[12] P. Kidwell, G. Lebanon, and K. Collins-Thompson. Statistical estimation of word acquisition with application to readability prediction. In *Proc. of EMNLP 2009*.

[13] G. Kumaran, R. Jones, and O. Madani. Biasing web search results for topic familiarity. In *Proc. of CIKM 2005*, 271–272.

[14] X. Liu, W. B. Croft, P. Oh, and D. Hart. Automatic recognition of reading levels from user queries. In *Proc. of SIGIR 2004*, 548–549.

[15] PuppyIR. PuppyIR: An open source environment to construct information services for children. 2011. http://www.puppyir.eu/.

[16] J. Teevan, S. T. Dumais, and E. Horvitz. Personalizing search via automated analysis of interests and activities. In *Proc. of SIGIR 2005*, 449–456.

[17] S. D. Torres, D. Hiemstra, and P. Serdyukov. An analysis of queries intended to search information for children. In *IIiX 2010*, 235–244.

[18] M. van Kalsbeek, J. de Wit, D. Trieschnigg, P. van der Vet, T. Huibers, and D. Hiemstra. Automatic reformulation of children's search queries. Technical Report TR-CTIT-10-23, June 2010.

[19] K. Wang, T. Walker, and Z. Zheng. Estimating relevance ranking quality from web search clickthrough data. In *Proc. of SIGKDD 2009*, 1355–1364.

[20] R. W. White, P. N. Bennett, and S. T. Dumais. Predicting short-term interests using activity-based search context. In *Proc. of CIKM 2010*, 1009–1018.

[21] R. W. White and S. T. Dumais. Characterizing and predicting search engine switching behavior. In *Proc. of CIKM 2009*, 87–96.

[22] R. W. White, S. T. Dumais, and J. Teevan. Characterizing the influence of domain expertise on web search behavior. In *Proc. of WSDM 2009*, 132–141.

[23] Q. Wu, C. J. C. Burges, K. M. Svore, and J. Gao. Adapting boosting for information retrieval measures. *Information Retrieval*, 3(13):254–270, 2010.

[24] C. X. Zhai, W. W. Cohen, and J. Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *Proc. of SIGIR 2003*, 10–17.