What to Read Next?: Making Personalized Book Recommendations for K-12 Users

Maria Soledad Pera
Computer Science Department
Brigham Young University
Provo, Utah, U.S.A.
mpera@cs.byu.edu

Yiu-Kai Ng
Computer Science Department
Brigham Young University
Provo, Utah, U.S.A.
ng@compsci.byu.edu

ABSTRACT
Finding books that children/teenagers are interested in these days is a non-trivial task due to the diversity of topics covered in huge volumes of books with varied readability levels. Even though K-12 readers can turn to book recommenders to look for books, the recommended books may not satisfy their personal needs, since they could be beyond/below their readability levels or fail to match their topics of interest. To address these problems, we introduce BReK12, a book recommender that makes personalized suggestions tailored to each K-12 user \( U \) based on books available on a social bookmarking site that (i) are similar in content to the ones that are known to be of interest to \( U \), (ii) have been bookmarked by users with reading patterns similar to \( U \)’s, and (iii) can be comprehended by \( U \). BReK12 is an asset to its users, since it suggests books that are appealing to its users and at grade levels that they can cope with, which can increase their reading selection choices and motivate them to read. We have also developed ReLAT, the readability analysis tool employed by BReK12 to determine the grade level of books. ReLAT is novel, compared with existing readability formulas, since it can predict the grade level of a book even if an excerpt of the book is not available. We have conducted empirical studies which have verified the accuracy of ReLAT in predicting the grade level of a book and the effectiveness of BReK12 over existing baseline recommendation systems.

Categories and Subject Descriptors
H.3.3 [Information Search and Retrieval]: Information Filtering, Retrieval Models, Selection Process

Keywords
Book recommendation system, readability, K-12

1. INTRODUCTION
Reading, an essential skill required for acquiring knowledge, is an integral part of the educational system. It is imperative to encourage K-12 students to read, since research studies have confirmed the enormous influence of reading on students’ development as learners and members of society, especially at an early age. Finding books that K-12 readers are interested in, however, is a difficult task due to the diversity of topics covered in the huge volume of books that target readers of different backgrounds and age groups. Recommender systems, which have been developed to suggest items of interest to their users, are supposed to alleviate the book-finding problem. However, existing recommenders employed at well-known book-affiliated websites, such as Novelist (www.ebscohost.com/novelist) and Amazon.com, adopt a “one-size-fits-all” strategy, which makes the same suggestions to different users on a given book without considering their individual preferences [14]. On the contrary, recommenders that offer personalized book suggestions overlook the readability levels of their users, since they are conceived with a general audience in mind. As a result, even if a book recommended to a user \( U \) matches \( U \)’s interests, the book might include complex (simple, respectively) content that is beyond (below, respectively) the readability level of \( U \), which fails to sustain the mission of matching users’ reading abilities with the suggested literature [1]. Moreover, these recommenders rely heavily on the existence of personal ratings [26] assigned to books by users, which are rarely provided by K-12 users of the existing social bookmarking sites established for them.

To address the aforementioned design problems of book recommendation systems, we have developed BReK12, a book recommender that makes personalized suggestions for K-12 users. To locate books for a user \( U \) based on \( U \)’s reading ability and interests, BReK12, which is designed to be integrated into a social bookmarking site on books, analyzes \( U \)’s profile, i.e., books that have been bookmarked by \( U \) on the site. If \( U \) is a new user, BReK12 treats a book provided by \( U \) as \( U \)’s profile. In doing so, BReK12 bypasses the cold-start problem often encountered by recommenders [24]. BReK12 first infers \( U \)’s readability level by analyzing the grade levels of books in his/her profile, which are determined using ReLAT, a robust readability level analysis tool that we have developed. Hereafter, BReK12 identifies a set of candidate books, among the ones archived at the site, with grade levels compatible to the inferred readability level of \( U \). For each candidate book, BReK12 determines its content similarity and readership similarity with books in the profile of \( U \) based on the brief descriptions of books that are publicly available online and users’ bookmarking patterns on the site, respectively. An aggregation strategy [3] is adopted...
so that the top-10 candidate books, with grade levels appropriate for \( U \) and with the highest combined content- and readership-similarity scores, are recommended to \( U \).

BReK12 is unique, since its design goal is to suggest books to K-12 users that simultaneously match their interests and reading abilities, which in turn can motivate them to read. Unlike state-of-the-art recommenders \([24]\), BReK12 simply employs readily available data, i.e., user bookmarks and brief descriptions of books, accessible from the social bookmarking sites where BReK12 is installed and book-affiliated websites, respectively to make recommendations. Moreover, BReK12 applies similarity, besides exact, matching on words to recommend books that are similar in content to users’ bookmarks, which otherwise could be ignored.

BReK12 relies on ReLAT to determine the grade level of a book \( B \) based on the subject areas addressed in \( B \), the readability level of the intended audience of books written by the author of \( B \), subject headings assigned to \( B \), and the grammatical/sentence structures in (an excerpt of) \( B \), if any is available. Unlike existing readability formulas/tools, such as Lexile Analyzer and Flesch-Kincaid, ReLAT can predict the readability level of a book even if (a sample of) the text of the book is unavailable, which is its novelty.

Besides serving social bookmarking site users, BReK12 can also recommend books for each K-12 patron of a school/public library, assuming that the list of books of interest provided by the library patron and the book catalog used by the library are given. In addition, BReK12 can be adapted to make recommendations for users of any book-affiliated website based on books searched by the users on the site.

The remaining of this paper is organized as follows. In Section 2, we discuss existing readability formulas and book recommenders. In Sections 3 and 4, we introduce ReLAT and BReK12, respectively. In Section 5, we present the results of the empirical studies conducted to (i) assess the performance of ReLAT and BReK12 and (ii) compare their performances with existing approaches. In Section 6, we give a concluding remark and directions for future work.

2. RELATED WORK

In this section, we present a number of widely-used readability formulas/analysis tools and book recommenders and compare them with ReLAT and BReK12, respectively.

2.1 Readability Formulas/Analysis Tools

For almost a century, hundreds of readability formulas have been developed for predicting the readability level of a text \([4, 7]\). Traditional readability formulas, such as Flesch-Kincaid and Coleman-Liau, rely on shallow features, which include word frequency and sentence length, to compute the grade level of a text. These formulas, however, often provide a rough estimation of text difficulty, which “(may) judge a nonsense passage as quite readable if the text’s jumbled words are frequent, short, and organized into brief sentences” \([4]\). Recently-developed formulas are based on (i) linguistic features, such as the ones introduced by Feng \et al. \([9]\) and Collins-Thompson and Callan \([7]\), (ii) pre-defined word lists, such as Lexile and Revised Dale-Chall formula, (iii) enhanced shallow features, such as the Advantage-TASA Open Standard for Readability (ATOS) formula, and (iv) non-textual features, such as SVM-Ranker \([15]\), which examines images on books to predict their grade levels, and ReadAid \([22]\), which considers information about the authors of a book and US Curriculum topics addressed in the book in addition to exploring the lexicographical and syntactical structures of the book. While most of these formulas are widely-used and popular, none of them can predict the readability level of a book if (a sample of) its text is not available, which can be achieved by ReLAT. (See \([4]\) for an in-depth discussion of existing readability formulas.)

2.2 Book Recommenders

A number of book recommenders \([14, 20, 26]\) have been proposed in the past. Amazon’s recommender \([14]\) suggests books based on the purchase patterns of its users. Yang \et al. \([26]\) analyze users’ access logs to infer their preferences and apply the traditional collaborative-filtering (CF) strategy, along with a ranking method, to make book suggestions. Givon and Lavrenko \([10]\) combine the CF strategy and social tags to capture the content of books. Similar to the recommenders in \([10, 26]\), the one in \([25]\) adopts the standard user-based CF framework and uses a domain ontology to determine the users’ topics of interest. The recommenders in \([10, 25, 26]\) overcome the problem that arises due to the lack of initial information to perform the recommendation task, i.e., the cold-start problem. However, unlike BReK12, they are constrained by requiring historical data in the form of ratings to make recommendations, which may not always be available. Moreover, the recommender in \([25]\) relies on the existence of a book ontology, which can be labor-intensive and time-consuming to construct. In making recommendations, Park and Chang \([20]\) analyze individual/group behaviors, such as clicks and shopping habits, and features describing books, such as their library classification, whereas PRf \([21]\) suggests books bookmarked by connections of a LibraryThing user. Similar to BReK12, PRf adopts a similarity-matching strategy, which differs from the exact-matching constraint imposed in \([20]\) and a number of content-based recommenders \([11, 18]\). However, neither PRf nor any of the aforementioned recommenders considers the readability level of their users as part of their recommendation strategies. While BReK12 is not a recommender system for learning \([24]\) per se, its design goal is to aid children/teenagers in developing good reading habits so that they can succeed at school and in the living of a good life. With that in mind, BReK12 is designed as an educational enrichment tool. (An in-depth discussion of existing recommender systems in the educational domain can be found in \([16]\).)

3. A GRADE LEVEL PREDICTION TOOL

As previously stated, existing readability formulas/analysis tools rely on at least a sample of a text to compute its readability level, which is a severe constraint, since text in a book is not always freely accessible due to the copyright laws. To alleviate the text constraint, we have developed ReLAT, which determines the grade level of any book using metadata on books, which accessible from reputable online sources, in addition to sample texts of books only if they are available. Analyzing a book using multiple perspectives, ReLAT can predict its grade level even if an excerpt on the book is missing. Furthermore, ReLAT instantly produces the grade level of a book, which is not always possible using commercial readability analysis tools. For example, Lexile offers scores for only approximately 150,000 out of the millions of books published worldwide in English, and requires
The grade level of a book takes as an input a unique identifier of BLAT. To determine the grade level of a book, ReLAT examines up to fifty-nine predictors to determine the grade level of B. The predictors are treated as contributing factors, i.e., evidences, which are used in analyzing the grade level of B. Due to the space constraint, we do not define each predictor. Instead, we present the nature of predictors in each category (as shown in Figure 1) below.

(I) Predictors based on an excerpt of B examine:

- **Grammar Concepts.** ReLAT analyzes the complexity of the grammar usage in B by counting the occurrences of various grammatical concepts in its sentences, which are present perfect, modal verbs, past progressive tense, parts of speech, phrases, suffixes, prefixes, and key vocabulary words.

- **Shallow Features.** ReLAT considers a number of well-established textual features commonly used by traditional readability formulas: the average number of syllables per word, average sentence length, percentage of words with at least three syllables, average number of characters per word, and absolute number of words.

- **Subject Areas.** The US Curriculum dictates which subject areas should be taught at each K-12 grade level. For example, multiplication is taught at the 3rd grade while trigonometry at the 12th grade. ReLAT relies on a Latent Dirichlet Allocation (LDA) model [5] to identify a set of representative keywords that best describes the content of B. Thereafter, ReLAT calculates the resemblance (using word-correlation factors [12]) between these keywords and each of the pre-defined subject areas established by the US Curriculum. The subject area SA with the highest degree of resemblance is treated as the subject area of B and the grade level associated with SA is used for predicting the grade level of B.

(II) Predictors based on topical information of B analyze its Subject Headings, e.g., "Biography", which are short phrases that capture the topics covered in books and are used by the US Library of Congress to categorize books. Based on a mapping between Subject Headings and grade levels (that we have already determined using Subject Headings associated to books with a known grade level), ReLAT identifies the grade levels that correspond to each of the Subject Headings of B. These grade levels are taken into account by ReLAT to determine the grade level of B.

(III) Predictors that consider author information of B are based on the fact that, in general, K-12 authors write for a particular group of readers at a certain grade level. For this reason, ReLAT treats the audience level metric of an author defined by WorldCat, which captures the "intellectual level of the audience for which the item is intended," in addition to the topical information and subject areas of the author’s other books (as introduced in I and II, respectively) as additional predictors that determine the overall grade level of B.

Since the information required to compute the value of a predictor can be missing, it may not be possible to use all the predictors for predicting the grade level of B. ReLAT considers various combinations of the 59 predictor coefficients and applies multiple linear regression analysis [17] (given below) on predictors applicable to B to predict its grade level.

\[
\begin{align*}
  c_0 + c_1 x_1 + c_2 x_2 + \ldots + c_n x_n &= y
\end{align*}
\]  

where \(c_0\) is the intercept term, \(c_i\) (1 ≤ i ≤ n) is the regression coefficient of predictor \(x_i\), and \(y\) is the predicted grade level.

Prior to applying Equation 1 to predict the grade level of a book, the intercept and coefficients associated with each applicable predictor are computed through a one-time training process using the ordinary least squares estimation method [17] on a training set of 8,737 K-12 books written by different authors that cover diverse topics and were extracted from various publishers’ websites and the Children’s Literature Comprehensive Database (CLCD.com). Each training instance consists of the (values of) predictors that apply to a book B and the grade level of B determined by its publisher. Since publishers usually suggest a range of readability levels for each of their published books, ReLAT considers the average grade level of the range defined for B as its grade level during the training process. In doing so, ReLAT avoids any bias that might occur by assigning books their lower/upper grade bound during the regression training.

**EXAMPLE 1.** Consider the book “Five Little Kittens” by Nancy Geller Jewell, which is a 32-page picture book with some long sentences. As stated in [23], unlike existing readability formulas that often overstate the difficulty of books for emergent readers, Accelerated Reader (AR) is a decent measure on the readability levels of books. Even though “Five Little Kittens” is appropriate for grades K-3, as suggested by its publisher, its Flesch-Kincaid grade level is 4.6 and its Lexile score is 970 (which corresponds to grades 6-7). The AR grade level for the book is 2.6, which matches the target audience for the book. The AR score, however, sug-

![Figure 1: The grade-level prediction process of ReLAT, our proposed readability level analysis tool](image-url)
gests that children should be at least in the 2nd grade to read “Five Little Kittens,” whereas the grade level predicted by ReLAT, which is 0.9, indicates that Kindergartners can read the book, providing a grade level more compatible (than AR’s) with the book publisher’s. (See Section 5.3 for the performance analysis of ReLAT.) □

4. THE BOOK RECOMMENDER

In this section, we present the design of BReK12. Given the profile $P$ of a user $U$, BReK12 selects a set of candidate books, which are compatible with the readability level of $U$ (determined in Section 4.1). Hereafter, BReK12 assigns a ranking score to each candidate book $B$, which is computed using an aggregation strategy (introduced in Section 4.4) on the content and readership similarity of $B$ with respect to the books in $P$ (defined in Sections 4.2 and 4.3, respectively).

4.1 Identifying Candidate Books

BReK12 recognizes that “reading for understanding cannot take place unless the words in the text are accurately and efficiently decoded” [19] and only recommends books with readability levels appropriate to its users.

BReK12 applies Equation 2 to estimate the readability level of a user $U$, denoted $RL(U)$, based on the grade level of each book $P_B$ in $U$’s profile predicted by ReLAT, denoted $RLAT(P_B)$. Note that only books bookmarked in a user’s profile during the most recent academic year are considered, since it is anticipated that the grade levels of books bookmarked by users gradually increase as the users enhance their reading comprehension skills over time.

$$RL(U) = \frac{\sum_{P_B \in P} RLAT(P_B)}{|P|} \tag{2}$$

where $|P|$ denotes the number of books in $U$’s profile and average is employed to capture the central tendency on the grade levels of books bookmarked by $U$.

Having established $U$’s readability level, BReK12 creates CandBks, the subset of books archived at the bookmarking site that are within-one-grade-level range from $U$’s. By considering books within one grade level above/below $U$’s mean readability level, BReK12 recommends books with an appropriate level of complexity for $U$ and grade levels approximate to the grade levels of books that have been read by $U$ (as of the most recent academic year).

**Example 2.** Consider a BiblioNasium.com user $U$ who has bookmarked a number of books from Dav Pilkey’s “Captain Underpants” series. Based on the grades levels predicted by ReLAT for the books archived at BiblioNasium (see a sample of BiblioNasium books in Table 1) and $U$’s readability level, which is 4, BReK12 does not include $B_{k1}$ or $B_{k3}$ in CandBks, since their grade levels are below/beyond the range deemed appropriate for $U$. □

4.2 Content Similarity Measure

BReK12 depends on the profile $P$ of $U$ to infer $U$’s interests/preferences. To determine the degree to which the content of a book $B$ in CandBks appeals to $U$, BReK12 computes the content similarity between $B$ and each book $P_B$ in $P$, denoted $CSim(B, P)$ as defined in Equation 3, using a “bag-of-words” representation on the brief descriptions of $B$ and $P_B$ obtained from book-affiliated websites, such as Amazon. To compute $CSim(B, P)$, BReK12 employs an enhanced version of the cosine similarity measure based on word-correlation factors [12], which relaxes the exact-matching constraint imposed by the cosine measure and explores words in the description of $B$ that are analogous to, besides the same as, words in the description of $P_B$.

Word-correlation factors in the correlation matrix introduced in [12] reflect the degree of similarity between any two words according to their (i) frequencies of co-occurrence and (ii) relative distances in Wikipedia (.org) documents. Approximately 880,000 documents covering a wide range of topics and written by more than 89,000 authors with varied writing styles were used to construct the matrix. Compared with synonyms/related words compiled by WordNet (wordnet.princeton.edu) in which pairs of words are not assigned similarity weights, word-correlation factors offer a more sophisticated measure of word similarity. In addition, word-correlation factors have been successfully employed to solve a number of content-similarity problems [21, 22].

$$CSim(B, P) = \max_{P_B \in P} \frac{\sum_{i=1}^{n} V_{B_i} \times V_{P_B_i}}{\sqrt{\sum_{i=1}^{n} V_{B_i}^2} \times \sqrt{\sum_{i=1}^{n} V_{P_B_i}^2}} \tag{3}$$

where $B$ and $P_B$ are represented as $n$-dimensional vectors $VB = \langle V_{B1}, ..., V_{Bn} \rangle$ and $VP_B = \langle V_{P_B1}, ..., V_{P_Bn} \rangle$, respectively, $n$ is the number of distinct words in the descriptions of $B$ and $P_B$, and $V_{Bi}$ ($V_{P_Bi}$, respectively) which is the weight assigned to word $B_i$ ($P_{Bi}$, respectively), is calculated as shown in the equations in Table 2.

$HS_w$ in Table 2 is the set of words that are highly similar to, but not the same as, a given word $w$ in the description of a book $Bk$, which is either $B$ or $P_B$. $|HS_w|$ is the size of $HS_w$, $tf_w, Bk = \frac{w_{Bk}}{\sum_{w \in Bk} w_{Bk}}$ is the normalized term frequency.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Weight Assignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>$B_i \in B$ and $P_{Bi} \notin P_B$</td>
<td>$V_{Bi} = tf_{Bi,B} \times idf_{Bi}$ and $V_{P_{Bi}} = tf_{P_{Bi},P} \times idf_{P_{Bi}}$</td>
</tr>
<tr>
<td>$B_i \in B$ and $P_{Bi} \notin P_B$</td>
<td>$V_{Bi} = tf_{Bi,B} \times idf_{Bi}$ and $V_{P_{Bi}} = \sum_{w \in HS_{P_{Bi}}} tf_{w,B} \times idf_{w}$</td>
</tr>
<tr>
<td>$P_{Bi} \notin P_B$</td>
<td>$V_{P_{Bi}} = \sum_{w \in HS_{P_{Bi}}} tf_{w,B} \times idf_{w}$</td>
</tr>
<tr>
<td>$B_i \notin B$ and $P_B$</td>
<td>$V_{Bi} = 0$ and $V_{P_B} = \sum_{w \in HS_B} tf_{w,B} \times idf_{w}$</td>
</tr>
<tr>
<td>$P_{Bi} \in P_B$</td>
<td>$V_{P_{Bi}} = tf_{P_{Bi},P} \times idf_{P_{Bi}}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ID</th>
<th>Book Title</th>
<th>Grade Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bk1</td>
<td>Mummies in the Morning</td>
<td>2.9</td>
</tr>
<tr>
<td>Bk2</td>
<td>Captain Underpants and the Big, Bad Battle of the Bionic Booger Boy</td>
<td>4.7</td>
</tr>
<tr>
<td>Bk3</td>
<td>The Hidden Boy</td>
<td>5.6</td>
</tr>
<tr>
<td>Bk4</td>
<td>Dragon’s Halloween</td>
<td>3.1</td>
</tr>
<tr>
<td>Bk5</td>
<td>Junie B. Jones Smells Something Fishy</td>
<td>3.0</td>
</tr>
</tbody>
</table>

Table 1: A number of BiblioNasium books

Table 2: TF-IDF weighting scheme used in the enhanced cosine similarity measure in Equation 3

Two words are highly similar if their correlation factor is included in a reduced version of the aforementioned word-correlation matrix which contains 13% of the most frequently-occurring words in the Wikipedia documents.
frequency of $w$ in $B_k$, and $idf_w = \log \frac{N}{n_w}$ is the inverse document frequency for $w$ in the collection of books $N$ archived at a social bookmarking site, where $n_w$ is the number of books in $N$ that include $w$ in their descriptions. Relying on the $tf-idf$ weighting scheme, BReK12 prioritizes discriminating words that capture the content of its respective book.

The max function in Equation 3 emulates the “most pleasure” strategy (commonly applied in game theory and group profiling [24]). Applying this strategy, BReK12 selects the highest possible score among the ones computed for each $P_B$ in $P$ and $B$. The larger the number of exact-matched or highly-similar words in the descriptions of both $B$ and $P_B$ is, the more likely $B$ is a relevant recommendation for $U$. Moreover, BReK12 adopts the widely-used cosine measure, which has been effectively applied to determine the degree of resemblance between any two items in content-based recommenders [18]. While content-similarity is computed by BReK12 using book descriptions, other textual information on books, such as their tag representations, can be used. Tags, however, are not always publicly available.

**Example 3.** To illustrate the merit of using the enhanced cosine similarity measure in Equation 3 to compute $CSim(B, P)$, consider the profile $P$ of user $U$ in Example 2 and two of the books, $B_{k_2}$ and $B_{k_4}$ as shown in Table 1. Using the traditional cosine measure, $B_{k_2}$ and $B_{k_4}$ yield the same content similarity score with respect to $P$. However, employing the enhanced cosine similarity measure, BReK12 obtains a more accurate content-similarity score for each book, since $CSim(B_{k_2}, P) = 0.57$ and $CSim(B_{k_4}, P) = 0.39$. These scores reflect that $U$ is likely more interested in books similar to the ones in the “Captain Underpants” series (by Dav Pilkey) than books about “dragons,” which we have verified by manually examining the profile of $U$. □

**4.3 Readership Similarity Measure**

$CSim$ locates books that are similar in content, to a certain degree, to the ones users have read in the past. This measure, however, does not consider books that are dissimilar in content but match users’ specific preferences/interests. Hence, BReK12 explores another dimension of resemblance between each book $B$ in CandBks and books in $U$’s profile $P$ by using the Lennon similarity measure [13] to perform co-readership analysis on users’ bookmarks on a social bookmarking site. This readership similarity measure (as defined in Equation 4) is based on the popular item-item similarity approach employed by collaborative-filtering recommenders to examine patterns of co-occurrence of items bookmarked by users to make recommendations [6].

$$RSim(B, P) = \max_{P_B \in P} \left( 1 - \frac{\min(|S_B - S_{B_k}|, |S_{P_B} - S_{B_k}|)}{\min(|S_B - S_{\emptyset}|, |S_{P_B} - S_{\emptyset}|) + |S_{\emptyset}|} \right)$$  \hspace{1cm} (4)

where $S_B$ ($S_{P_B}$, respectively) is the set of users who bookmarked $B$ ($P_B$, respectively), $S_{\emptyset} = S_B \cap S_{P_B}$, $|S_{\emptyset}|$ is the number of users who bookmarked both $B$ and $P_B$, $|S_B - S_{\emptyset}|$ ($|S_{P_B} - S_{\emptyset}|$, respectively) is the number of users who bookmarked $B$, but not $P_B$ ($P_B$ but not $B$, respectively) at a social bookmarking site, and the use of the max function was discussed in Section 4.2.

In Equation 4, the $\min(\text{imum})$ of the two differences between $|S_B - S_{\emptyset}|$ and $|S_{P_B} - S_{\emptyset}|$ is chosen, since by using the smaller of the two differences, we can more accurately capture the similarity between $B$ and $P_B$. As a difference reflects the number of users who bookmark $B$, but not $P_B$ (or vice versa), a smaller difference signifies that proportionally a larger number of users who bookmark one book also prefer the other book, which is a better indication of the degree of readership similarity between the two books.

**Example 4.** To illustrate the usefulness of readership similarity measure for making recommendations, consider $B_{k_3}$ in Table 1 and the book “The Adventures of Captain Underpants” by Dav Pilkey, denoted $P_B$, which is a book in the profile of user $U$ introduced in Example 2. The contents of the books differ, since $P_B$ is about the adventures of two fourth graders and a superhero, whereas $B_{k_3}$ details the events that occur when Junie, the main character in the book, takes her pet to school. The books, however, share some common thread of interest to a group of BiblioNasium users who have bookmarked both, since the books include characters of similar age, written in similar literary styles, and share the same genre. Relying partially on the readership similarity between $B_{k_3}$ and $P_B$, which is 0.67, BReK12 retains $B_{k_3}$ as one of the top-10 recommendations for $U$, which otherwise would have been ignored due to the lack of related content between the two books. □

**4.4 Rank Aggregation**

Using the computed content- and readership-similarity scores of each book $B$ (with a readability level appropriate for $U$) in CandBks, BReK12 applies the Borda Count voting scheme [3] to determine the ranking score for $B$. The Borda Count voting scheme is a positional-scoring procedure such that given $k$ ($\geq 1$) candidates, each voter casts a vote for each candidate according to his/her preference. A candidate that is given a first-place vote receives $k$-1 points, a second-ranked candidate $k$-2 points, and so on up till the last candidate, who is awarded no points. Hereafter, the points assigned to each candidate across all the voters are added up and the candidate with the most points wins.

The Borda Count strategy, which has been successfully applied to different information retrieval tasks [3], is employed by BReK12 to generate a single ranking score for $B$, denoted $Borda(B)$ as defined in Equation 5, that regards the content- and readership-similarity scores as equally important in determining the degree to which a user is interested in $B$. Using Equation 5, BReK12 assigns (i) $k = |\text{CandBks}|$, which is the number of candidate books selected for a user $U$, and (ii) $C = 2$, which is the number of voters, i.e., the two ranked lists of similarity scores on books computed in Sections 4.2 and 4.3, respectively. Candidate books with the top-10 Borda scores are recommended to $U$.

$$Borda(B) = \sum_{c=1}^{C} (k - S^B_c)$$  \hspace{1cm} (5)

where $S^B_c$ is the position on the ranking of $B$ based on the $c^{th}$ ranked list to be fused.

BReK12 adopts Borda as an aggregation strategy, since (i) its combination algorithm is simple and efficient, which requires neither training nor compatible relevance scores that may not be available and (ii) its performance is competitive with other existing combination strategies [3].

**5. EXPERIMENTAL RESULTS**

In this section, we first introduce the datasets, metrics, and evaluation protocol used for assessing the performance
of ReLAT and BReK12 (in Sections 5.1 and 5.2, respectively). Hereafter, we present the results of the empirical studies conducted for evaluating the effectiveness of ReLAT and BReK12 (in Sections 5.3 and 5.4, respectively).

5.1 The Datasets

To the best of our knowledge, there is no benchmark dataset that can be used for evaluating readability-level prediction tools on books. Thus, to evaluate ReLAT we constructed BookGL, using data extracted from CLCD.com, a website established to assist teachers, parents, and librarians in choosing appropriate books for K-12 readers, the Young Adults Book Central (YABC.com), and reputable publishers’ websites. BookGL consists of 2,248 books distributed among the K-12 grade levels with the grade level (range) of each book assigned by its publisher. (See Table 3 for the source websites and their number of books in BookGL.) Due to the lack of consensus among researchers on the accuracy of existing readability prediction tools [4], we consider publisher-provided grade levels as the “gold-standard,” since they are defined by human experts.

Even though the BookCrossing dataset (informatik.uni-freiburg.de/~cziegler/BX) has been employed to evaluate book recommenders tailored to a general audience, it is not specifically designed for assessing the performance of book recommenders for K-12 users. Hence, we used data provided by BibliNasium, which is a safe and secure social networking site on books developed exclusively for children and teenagers, to evaluate BReK12. The dataset consists of the profile of, i.e., books that have been bookmarked by, each of the 297 BibliNasium users who joined the site within its first month of being launched. As the design methodology of BReK12 relies on brief descriptions and predicted grade levels of books, we extracted the former from Amazon.com and predicted the latter using ReLAT.

5.2 Metrics and Evaluation Protocol

To assess the performance of ReLAT, we apply the (absolute) error rate (ER) [8], which is the absolute difference between an expected and a predicted grade level for a book $B$.

$$ER = \frac{1}{\text{BookGL}} \sum_{B \in \text{BookGL}} |PR(B) - GL(B)|$$

(6)

where $|\text{BookGL}|$ is the total number of books in BookGL, $GL(B)$ is the predicted grade level of $B$ by a readability formula/analysis tool, and $PR(B)$ is either the lower or upper bound of the grade level range of $B$ determined by its publisher, whichever is closest to $GL(B)$. (Recall that publishers often assign a grade range, not a level, to a book.)

We evaluate BReK12 using Precision@10, Mean Reciprocal Rank (MRR), and Normalized Discounted Cumulative Gain (nDCG) [8]. Precision@10 measures the fraction of the top-10 ranked recommendations that are relevant, whereas MRR computes the average ranking position of the first relevant recommended book. nDCG determines the overall ranking performance of BReK12 and penalizes relevant books ranked lower in the recommendation list. The penalization is based on a reduction, which is logarithmically proportional to the position of each relevant book in a ranked list.

We adopt the popular five-fold cross validation strategy to evaluate BReK12 (and recommender systems considered for the comparison purpose). In each of the five repetitions, 80% of the books bookmarked by a user $U$ in the BibliNasium dataset are used to create $U$’s profile and the remaining 20% are reserved for the testing purpose. A recommended book $B$ is treated as relevant to $U$ if it is included in the 20% of the books withheld for the testing purpose, and is non-relevant otherwise, which is a commonly-employed protocol for assessing the performance of recommendation systems [11]. Since only withheld books are considered relevant, it is not possible to account for potentially relevant books a user has not bookmarked, which is a well-known limitation of this evaluation protocol. As the limitation affects all the recommenders evaluated in the conducted empirical studies, the results are consistent for the comparative purpose.

5.3 Performance Evaluation of ReLAT

Using the 127 books in BookGL with excerpts, we compared the grade-level prediction accuracy of ReLAT with a number of well-known readability formulas: Coleman-Liau, Flesch-Kincaid, Rix, and Spache, which we have implemented. (See detailed discussion on these readability formulas in [4].) Figure 2(a) shows that (i) on the average the grade level predicted by ReLAT for a book with text (in BookGL) is about half a grade from the grade (range) determined by its publisher and (ii) ReLAT’s error rate is at least 33% lower than the error rate created by its counterparts.

We also evaluated the performance of ReLAT in predicting the grade level of books for which their excerpts cannot be obtained online. Among the 2,121 books in BookGL without sample text, the 0.82 error rate generated by ReLAT is less than one grade level off the ranges specified by the publishers of the books. This low error rate is not only an accomplishment of ReLAT, but also it cannot be achieved by any of the existing readability formulas/analysis tools, since none of them can predict the grade level of books without excerpts. The overall error rate of ReLAT on BookGL, in which 94% of the books are without text, is 0.81, which is within one grade level of the targeted grade level.

We further compare the performance of ReLAT with two popular readability analysis tools widely-accepted by grade schools and reading programs in the USA: Accelerated Reader (AR) and Lexile. Recall that the algorithms developed to compute AR and Lexile scores are not publicly accessible, but we were able to find 897 books with AR scores and 314 books with Lexile scores among the books in BookGL at ARbookfind.com and Lexile.com, respectively. As shown in Figure 2(b), ReLAT outperforms AR and is significantly more accurate than Lexile in predicting the grade level of the analyzed books (in BookGL).

5.4 Performance Evaluation of BReK12

In this section, we verify the correctness of the design methodology of BReK12 and compare its performance with a number of existing recommendation strategies.
5.4.2 Comparing BReK12 with Other Recommenders

5.4.1 Effectiveness of BReK12

The results of the study conducted to validate the methodologies applied by BReK12 for selecting and ranking books to be recommended for K-12 users are presented as follows:

- The Enhanced Cosine (EC) measure employed by BReK12 to perform content-similarity matching outperforms the Traditional Cosine measure (as shown in Figure 3), which has been verified based on the improvements in Precision@10, MRR, and nDCG that are statistically significant as determined by using the Wilcoxon signed-rank test [8] (with \( p < 0.05 \)).

- We have observed statistically significant improvement (\( p < 0.05 \)) when books with unsuitable readability levels for the respective BiblioNasium users are excluded prior to applying the content-similarity matching on potential book recommendations. (See EC versus EC + ReLAT Filtering in Figure 3.)

- The statistically significant improvements (\( p < 0.01 \)) on various performance metrics achieved by EC + Readership over EC indicate that examining both the content and readership similarity of candidate books with respect to books in the profile of each BiblioNasium user increases the accuracy of the recommendations.

- We have verified (based on the results of a conducted empirical study) that recommending books that match users’ reading abilities without considering their individual preferences is not beneficial. This is anticipated, since users might not find particular topics addressed in books appealing even if the books are suitable to their readability levels.

- When books beyond/below users’ readability levels are not chosen as candidate books by BReK12, its overall effectiveness increases, according to the statistically significant improvements (\( p < 0.01 \)) achieved by BReK12 over EC + Readership. This comparison validates the correctness of BReK12’s design methodology, i.e., to recommend books of interest to individual users that they can read and understand.

5.4.2 Comparing BReK12 with Other Recommenders

As previously stated, no other personalized book recommender explicitly considers the reading abilities of its users. Thus, we have compared the performance of BReK12 with a number of recommenders developed for a general audience.

Tag Vector Similarity (TVS) [11], L-Cosine (L-Cos) [18], and Item-Based Collaborative Filtering (ICF) [6] were employed for comparison purposes, as opposed to other state-of-the-art book recommenders introduced in Section 2, since the latter require personal ratings on (K-12) books provided by individual users or social connections established by social bookmarking site (K-12) users, neither of which are archived by BiblioNasium. To determine which books should be recommended to a user, TVS applies the cosine similarity measure on TF-IDF tag vector representations of books\(^5\), whereas L-Cos considers the weighted frequency of each keyword in the description or title of a book. ICF, on the other hand, calculates the degree of similarity between any two books based on the number of users who have bookmarked both books on a social bookmarking site, which is a variation of the popular collaborative filtering strategy commonly adopted for making recommendations.

As shown in Figure 3, BReK12 outperforms its counterparts based on the evaluation metrics introduced in Section 5.2. The improvements achieved by BReK12 over the others are statistically significant (with \( p < 0.01 \)). According to the computed MRR, on the average BReK12 users are required to browse through about one (\( \frac{1}{1.9} = 0.52 \)) recommended book before locating a relevant one, whereas users of TVS, L-Cos, and ICF are required to scan through at least 4 (\( \frac{1}{0.25} = 4 \)) recommended books, respectively. The Precision@10 values reflect that, in general, close to 8 (out of 10) books suggested by BReK12 are relevant, as opposed to close to 4, 2, and 6 relevant books recommended by TVS, L-Cos, and ICF, respectively. The nDCG scores indicate the superiority of BReK12 over TVS, L-Cos, and ICF in ranking relevant books to be recommended higher in the list of suggested books.

While TVS, L-Cos, and ICF consider only textual descriptions of books or bookmarking patterns of users on a social site, BReK12 examines multiple contributing factors to identify potential recommendations, which increases the number of relevant reading selections for the users.

6. CONCLUSIONS AND FUTURE WORK

We have introduced BReK12, a unique recommender tailored to K-12 readers, which makes personalized suggestions on books that satisfy both the preferences and reading abilities of its users. Unlike current state-of-the-art recommenders that rely on the existence of users’ historical data in the form of ratings, which are missing among the K-12 users, BReK12 simply considers readily available brief descriptions on books, patterns of co-occurrence among books bookmarked on a social bookmarking site on which BReK12 is installed, and grade levels of books computed using our newly-developed ReLAT. ReLAT is novel, since it can determine the grade level of any book (even if a sample of the text of a book is unavailable) by analyzing the Subject Headings of books, US Curriculum subject areas identified in books, and information about the authors of books. As children continue to read more books if they can choose what to read [2], a significant contribution of BReK12 is to provide K-12 readers a selection of suitable books to choose from that are not only appealing to them, but can be comprehended.

\(^5\)Tag descriptions on books can be extracted from the INEX 2012 Social Book Search Track dataset (inex.mmci.unisalzburg.de/tracks/books/).
Figure 3: Performance evaluation of TVS, L-Cos, IICF, and BReK12 using the BiblioNasium dataset

by them. The conducted experiments demonstrate (i) the accuracy of ReLAT and its superiority over existing readability formulas/analysis tools, and (ii) the effectiveness of BReK12, which outperforms baseline recommenders in suggesting books for K-12 users.

As part of our future work, we plan to extend BReK12 so that it can suggest reading materials for struggling readers, i.e., readers with learning disabilities and those who learn English as a second language, for whom the grade level of a recommended book is an important factor to be considered.

7. REFERENCES


