Automating Readers' Advisory to Make Book Recommendations for K-12 Readers

Maria Soledad Pera
Computer Science Department
Brigham Young University
Provo, Utah 84604, U.S.A.
mpera@cs.byu.edu

Yiu-Kai Ng
Computer Science Department
Brigham Young University
Provo, Utah 84604, U.S.A.
ng@compsci.byu.edu

ABSTRACT

The academic performance of students is affected by their reading ability, which explains why reading is one of the most important aspects of school curriculums. Promoting good reading habits among K-12 students is essential, given the enormous influence of reading on students' development as learners and members of society. In doing so, it is indispensable to provide readers with engaging and motivating reading selections. Unfortunately, existing book recommenders have failed to offer adequate choices for K-12 readers, since they either ignore the reading abilities of their users or cannot acquire the much-needed information to make recommendations due to privacy issues. To address these problems, we have developed Rabbit, a book recommender that emulates the readers' advisory service offered at school/public libraries. Rabbit considers the readability levels of its readers and determines the facets, i.e., appeal factors, of books that evoke subconscious, emotional reactions on a reader. The design of Rabbit is unique, since it adopts a multi-dimensional approach to capture the reading abilities, preferences, and interests of its readers, which goes beyond the traditional book content/topical analysis. Conducted empirical studies have shown that Rabbit outperforms a number of (readability-based) book recommenders.

Categories and Subject Descriptors

Information Systems [**Retrieval tasks and goals**]: [Recommender systems]

Keywords

Recommendation system; books; K-12; readers' advisory

1. INTRODUCTION

Besides watching TV, text messaging, and playing computer games, children and teenagers these days often spend spare time browsing through the Internet, looking for something fun to do. YouTube and Facebook are examples of popular sites among young audiences that offer free entertainment on demand anytime and anywhere. Recent statistics[1] have shown that children/teenagers spend an average of seven hours a day on (smart) devices. This is alarming, since a significant number of children/teenagers are underachieving at school, especially in *reading*. According to the 2013 National Assessment of Educational Progress,[2] only 32% of American 4^{th} graders are proficient in reading. Kids should allocate some of their free time on reading to enhance their educational experience. To turn the tide, educators, parents, government agents, and private organizations must join force to encourage kids to read. Unfortunately, very few existing websites/(non-)government agents are equipped with the resources/technologies to cope with the problem.

To motivate K-12 readers, it is necessary to avoid presenting these readers with books that are either too easy/difficult to read or involve topics unappealing to them which could diminish their interest in reading. In fact, finding the right books for the right audience is not easy. Even though existing recommenders can assist readers in finding books, they rely either on large historical data, e.g., personal tags/ratings, (which might not be available) or readers' interactions on social sites (which may not be accessible involving K-12 readers due to privacy issues). Furthermore, these systems *ignore* the reading abilities of the respective readers in recommending books. To address the shortcomings of these design methodologies, we have developed Rabbit, a Readers' advisory-based book recommendation tool that makes personalized, appealing suggestions on books to K-12 readers.

Rabbit is unique, since it simulates readers' advisory (RA), a service offered at school/public libraries which are established to champion and encourage reading. RA involves knowledgeable professionals in finding reading materials of interest for their patrons [14]. During the search process, librarians identify the *topics*, *contents*, and *appeal factors*, i.e., literary elements, appealing to each reader and suggest books accordingly. By offering this service, which is in high demand even in the era of the Web 2.0, libraries provide "a vital link between library materials and readers" [14]. Unfortunately, (young) readers may not approach readers' advisors to ask for suggestions, feel their interest are *obscure* or *low-brow*, or not even visit libraries in person [5]. By automating the RA process, we replace RA in finding books appealing to K-12 readers, which eliminates the interaction with professionals and at the same time handles any num-

RecSys'14, October 6–10, 2014, Foster City, Silicon Valley, CA, USA.
Copyright 2014 ACM 978-1-4503-2668-1/14/10$15.00.
http://dx.doi.org/10.1145/2645710.2645721.

[1] nytimes.com/2010/01/20/education/20wired.html?_r=0
[2] http://goo.gl/Ad1VbE.

ber of readers simultaneously that cannot be achieved by traditional RA.

Rabbit is novel, since besides analyzing the reading ability of a reader R, it examines appeal factors to capture the reasons why a book Bk is appealing to R. Rabbit determines the facets of Bk that instigate a reaction to Bk, which in turn impacts the perception of R on Bk. Rabbit explores the literary elements of books that identify the *rate* in which the stories unfold (pace), their *overall structure* (storyline), the *feelings* that these stories evoke on a reader (tone), and subject *matters* that some readers might find unpleasant or offensive (special topics), in addition to the *qualities* of the *characters* (characterization) and the *language* and *level of details* (writing style) of the stories. Rabbit is neither affected by the cold-start problem nor requires feedback from its users, and its design surpasses the content or reading patterns explored by state-of-the-art recommenders.

While topics and content descriptions of books are freely and publicly available from reputable online sources, such as the Library of Congress,[3] appeal-factor/-term descriptions, which are fundamental for our recommendation strategy, are either determined by professionals on-the-fly or accessed through a paid subscription to RA databases. To automate the process of extracting appeal-term descriptions of books, we have developed ABET (Appeal-based extraction tool), a component of Rabbit, which relies on book reviews retrieved from book-related websites, such as Amazon(.com) and Powells.com. ABET is based on rules that examine typed dependencies and part-of-speech tags of words in reviews to identify appeal factors and their terms. ABET relies on reviews for extracting appeal-term descriptions on books, since they are readily available online and capture readers' varied opinions on the literary elements of a book.

2. RELATED WORK

Numerous approaches have been developed to identify and extract either features, (the polarity of) opinions, or feature-opinion pairs from reviews based on bootstrapping, natural language processing, machine learning, extraction rules, latent semantic analysis, statistical analysis, and information retrieval. (An in-depth review of state-of-the-art approaches adopted for opinion mining and extraction can be found in [11].) A product review describes the product's actual features, such as the "zoom" of a camera, which is unlike a book review that evaluates "organization and writing style, possible market appeal, and cultural, political, or literary significance" of a book [15]. A book review is a form of literary criticism in which the work is analyzed based on its content, style, and merit. We have observed that existing information extraction approaches on product reviews are ill-equipped for book reviews, since sentences expressed in book reviews tend to be more elaborated than the ones used to describe products. For this reason, ABET is not designed using any existing extraction approach. Instead, it relies on simple rules to perform linguistic and semantic analysis of the content and writing style of book reviews.

A number of book recommendation systems have been developed in the past. Amazon's recommender suggests books based on the purchase patterns of its users [8], whereas Yang et al. [17] analyze users' access logs to infer their preferences and apply the traditional collaborative-filtering (CF) strategy to make book recommendations. Givon and Lavrenko [4] combine CF and social tags to capture the content of books for recommendation. Sieg et al. [16] rely on the standard user-based CF framework and incorporate semantic knowledge in the form of a domain ontology to capture the topics of interest to a user. The hybrid-based recommenders in [4, 16, 17], in addition to Rabbit, overcome the cold-start problem. Unlike Rabbit, however, the others require (i) historical data on the users in the form of ratings, which may not always be available, or (ii) an ontology, which can be labor-intensive and time-consuming to construct.

Unlike Rabbit, PReF [12] examines users' connections as part of the recommendation process, which may not be accessible as they involve K-12 readers due to privacy imposed on children. BReK12 [12], which is the closest book recommender compared with Rabbit, is based on content and readability analysis. The former, however, relies on the availability of bookmarking information offered by social bookmarking site users to analyze reading patters of users. Furthermore, with the exception of BReK12, neither of the aforementioned recommenders considers the readability level of their users as part of their recommendation strategies. Although Rabbit is not a recommendation system for learning, its design goal is to enhance reading selections for K-12 users. (An in-depth description of existing recommenders in the educational domain can be found in [9].)

3. READERS' ADVISORY (RA)

Rabbit emulates the readers' advisory (RA) service available at public libraries since the late 1800's [3, 14]. RA offers (non-)fiction materials of potential interest with "the help of knowledgeable and non-judgmental library staff" [14]. While traditional RA involves face-to-face discussions between patrons and librarians, a number of public libraries, such as Williamsburg Regional,[4] take advantage of existing technologies and replace human interactions with online forms filled out by patrons to capture their interests [5].

Either through face-to-face conversations or filled-out online forms, a RA librarian's task is to identify the type of books preferred by readers based on the reasons behind their preferences. Besides analyzing the *topical areas* and *content descriptions* of books favored by a reader R, during the RA process, librarians examine the *appeal factors* of books that R is interested in [14]. Appeal factors, such as the pacing or description of characters in books, are "the elements of a book—whether definable or just understood—that make readers enjoy the book" [14]. These factors capture general traits of a book that attract the attention of a reader and are considered in answering one of the most important RA questions, "Why is the reader interested in a given book?" For example, some readers might enjoy the Harry Potter books (by J.K. Rowling) because of the established friendships among students and the boarding school setting, whereas others like the fantasy aspect of the story.

To the best of our knowledge, there is no consensus on the set of appeal factors that must be considered during the RA process. The most prominent appeal factors, as articulated in RA-related literature [3, 5, 14] include: (i) characterization, (ii) frame, (iii) pacing, (iv) storyline, (v) language and writing style, (vi) tone, and (vii) special topics. The first six appeal factors are well-known literary elements of

[3] catalog.loc.gov/

[4] wrl.org/books-and-reading/adults/looking-good-book

Table 1: Sample appeal terms associated with each of the appeal factors considered by Rabbit

Appeal Factors	Appeal Terms
Characterization	Believable, distant, dramatic
Frame	Bittersweet, contemporary, descriptive
Language and Writing Style	Candid, complex, conversational, extravagant, poetic, prosaic
Pacing	Easy, fast, slow
Special Topics	Addiction, bullying, violence
Storyline	Action-oriented, character-centered
Tone	Dark, happy, surreal

(non-)fiction books [2], whereas the latter identifies subjects addressed in a book that can cause emotional stress to some readers but tolerated/enjoyed by others [3]. Each appeal factor is associated with a vocabulary, which is a set of keywords, called *appeal terms*, employed to describe the factor, which we have defined based on well-known RA literature [3, 14]. The appeal factors considered by Rabbit and a sample of their respective appeal terms are shown in Table 1.

Based on the contents, topics, and appeal terms that describe the appeal factors of books $preferred$ by a reader R, librarians suggest other books matching (to a certain degree) the interests/preferences of R. However, due to the amount of books being published on a regular basis these days, it is an impossible task for a librarian to be familiar with every existing book to determine if it could be a potential relevant recommendation for R. For this reason, librarians turn to RA databases, which are available at NoveList, Fiction Connection, Which Book, and Readers' Advisory Online, to conduct fact-based, appeal factor-oriented, and read-alike searches in locating books to suggest to a reader [3].

4. APPEAL-TERM DESCRIPTIONS

While Rabbit conducts topical and content analysis of books using data from book-related websites, appeal-term descriptions are only available through RA databases or determined by professionals. Unfortunately, accessing reputable RA databases, such as NoveList Plus, comes with a price tag, i.e., paid subscription, whereas professionals might not have read a particular book and thus it would not be possible for them to infer the corresponding appeal-term description on-the-fly. To address these constraints, we have developed ABET, a tool that automatically extracts appeal-term descriptions of books from reviews available at well-known book-related websites, such as Amazon, Bertrams, Bookfinder4u, Bookmooch, Dogobooks and Fishpond.As reading is a personal experience, it is anticipated that a book is (not) appealing to a reader for various reasons. By analyzing reviews, we extract diverse readers' opinions on a book based on appeal terms that describe the corresponding appeal factors of the book, which in turn facilitates the task of identifying why books are appealing to a reader based on their literary elements.

To generate the appeal-term description for a given book, ABET relies on the taxonomy defined in Section 3, which despite being comprehensive, cannot account for a given appeal factor/term being specified differently in readers' reviews. For example, a reviewer may refer to the "Storyline" of a book as "story" or "narrative", and (s)he may also use either "quick" or "fast" to describe its "Pace". For this reason, we extend the appeal factors/terms by including the (stemmed) synonyms of each term/factor, which are identified using WordNet.[5] (The complete list of appeal terms for each appeal factor can be found at goo.gl/BSwuPw.)

While the taxonomy can serve as an aid to identify potential appeal factors/terms in reviews, it is imperative to properly associate these appeal terms and appeal factors in the reviews so that appeal factor-appeal term pairs can be correctly extracted to generate an accurate appeal-term description for a given book. To accomplish this task, we have defined a number of extraction rules[6] (as given in Table 2) for ABET based on typed dependency relations between word pairs in sentences extracted from reviews. It is natural for ABET to turn to typed dependencies, since they capture the *semantic connection*, i.e., association, between words in sentences. (A detailed discussion on typed dependencies is available at http://goo.gl/31Puwm.)

Rules used by ABET are based on written patterns identified in book reviews and capture the semantic link between appeal factors and their corresponding terms. Consider sentence S_A, "The narrative of the book is dramatic", and sentence S_B, "He creates believable characters". In S_A the *subject* of the sentence, i.e., "narrative," is characterized as being "dramatic", whereas in S_B its *object*, i.e., "characters", is described as "believable". As illustrated by the aforementioned examples, if the subject/object of a sentence is an appeal factor, then a word in the sentence that is semantically directly linked to the mentioned object/subject is often an appeal term. ABET captures these connection patterns using Rules 1 and 2 as defined in Table 2.

In a sentence, an appeal terms can also be indirectly connected with an appeal factor. Consider sentence S_C, "The descriptions included are extravagant". "Extravagant" is *indirectly* related to the subject of S_C, i.e., "descriptions", through the word "included". Using Rule 3, ABET examines pairs of grammatical relations that involve indirect connections among words. Now consider sentence S_D, "The characters are not simple". Based on Rule 1, ABET would mistakenly connect "Characterization" with "simple". This example reveals the need to examine pairs of grammatical relations in the presence of negated terms. ABET applies Rule 4, which identifies a negated term as a modifier of an appeal term t and then extracts as the appeal term for the corresponding factor the antonym of t (if it is included in the vocabulary defined in ABET's taxonomy for the factor).

Rules 3 and 4 take precedence over Rules 1 and 2, since once a typed dependency in a sentence is used by either of the former rules, it cannot be considered by the latter ones.

While majority of other relations (beyond the ones captured by Rules 1 to 4) seldom appeared in reviews, we observed three *special cases* that facilitate the extraction of appeal terms for "Special Topics" and "Frame", respectively, which we defined in Rules 5 to 7. Consider S_E, "It is about violence at schools", S_F, "Bullying is depicted in the book", and S_G, "The action is set in a school", which include special written patterns pertaining to the "Frame" and "Special Topics" appeal factors that are based on prepositions, subjects, and objects identified in sentences in reviews. The preposition "about" in S_E captures an appeal term employed to describe the factor "Special Topics", i.e., "violence", whereas

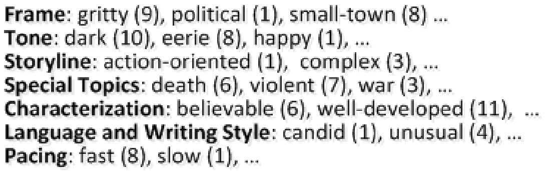
[5] wordnet.princeton.edu

[6] ABET performs linguistic/semantic analysis on sentences in reviews using Stanford Parser (http://goo.gl/Pxj2QW).

Table 2: Rules considered by ABET to identify appeal factor-appeal term pairs in book reviews

Notations
$rel(A, B)$ is a *grammatical relation* between a *dominant*, i.e., *governor* or *head*, word (A) and a *subordinate*, i.e., *dependent* or *modifier*, word (B) L_F, L_T, EL_F, and EL_T are the list of appeal factors, list of appeal terms, extended list of appeal factors, and extended list of appeal terms, respectively w_f is an appeal factor in L_F, and w_t is an appeal term in L_T $w \rightsquigarrow w_f$ ($w \rightsquigarrow w_t$, respectively) denotes that w is a synonym of w_f (w_t, respectively) $POS(w)$ is the part-of-speech tag of w which is a verb (adverb, respectively) if $POS(w)$ = "VB" ("RB", respectively) Abbreviation: adv(erbial)mod(ifier), a(djectival)mod(ifier), c(lausal)comp(lement), d(irect)obj(ect), neg(ation modifier), nn (noun compound modifier), n(ominal)subj(ect), nsubjpass (passive nominal subject), prep(_*) (Prepositional modifier) ABET extracts a pair $< w_f, w_t >$ only if w_t is in the corresponding vocabulary defined for w_f

Rule	Objective	Conditions	Identified Factors/Terms
1	To capture the written patterns based on a keyword, i.e., appeal term, that immediately precedes/ follows the subject or object of a sentence S, i.e., appeal factor	$A \in EL_T, B \in EL_F, rel \in$ {nn, nsubj}, (If A is a synonym of a term that applies to "Characterization" or "Storyline", then $POS(A) \notin$ {"VB", "RB"})	$B \rightsquigarrow w_f$ $A \rightsquigarrow w_t$
2		$A \in EL_F, B \in EL_T, rel \in$ {advmod, amod, prep_in, prep_about}	$A \rightsquigarrow w_f$ $B \rightsquigarrow w_t$
3	To identify an appeal term that qualifies its indirectly-related appeal factor in S	$rel \in$ {nn, nsubj}, $B \in EL_F$, and $\exists rel_2(C, D) \in$ {amod, dep, ccomp}, $A = C, D \in EL_T$	$B \rightsquigarrow w_f$ $D \rightsquigarrow w_t$
4	To explicitly consider *negated* appeal terms in S	$B \in EL_F, rel \in$ {nn, nsubj}, $\exists neg(C, D)$, A $(= C)$ is an antonym of $\bar{A} \in EL_T, D$ is a negation term	$B \rightsquigarrow w_f$ $\bar{A} \rightsquigarrow w_t$
5	To account for the multiple ways in which a reviewer can describe the setting of books or peeves/favored subject matters in books to handle special cases of "Special Topic" and "Frame" factors in S	$A \in EL_T, rel \in$ {prep_about}, A is a synonym of a term that describes "Special Topics"	w_f = "Special Topics" $A \rightsquigarrow w_t$
6		$B \in EL_T, rel \in$ {dobj, nsubj, nsubjpass}, $POS(A)$ = "VB", B is a synonym of an appeal term that describes "Special Topics"	w_f = "Special Topics" $B \rightsquigarrow w_t$
7		$A \rightsquigarrow$ "Frame" $\in EL_F, B$ (a synonym of an appeal term that describes "Frame") $\in EL_T, rel \in$ {prep_in}	w_f = "Frame" $B \rightsquigarrow w_t$



Frame: gritty (9), political (1), small-town (8) ...
Tone: dark (10), eerie (8), happy (1), ...
Storyline: action-oriented (1), complex (3), ...
Special Topics: death (6), violent (7), war (3), ...
Characterization: believable (6), well-developed (11), ...
Language and Writing Style: candid (1), unusual (4), ...
Pacing: fast (8), slow (1), ...

Figure 1: ABET-generated appeal-term description for "The Hunger Games"

"in" in S_G is connected with an appeal term, i.e., "school", that describes the factor "Frame". Moreover, "bullying" in S_F, which is assigned a "VB" part-of-speech tag, is an appeal term describing the factor "Special Topics".

ABET creates the appeal-term description for a book Bk by applying rules defined in Table 2 on (up to) 500 distinct reviews of Bk, if they are available. In generating the appeal-term description of Bk, ABET considers not only the appeal terms extracted from reviews on Bk, but also their *frequency of occurrence*. The latter captures the relative *degree of significance* of an appeal term in describing its corresponding factor based on reviewers' varied opinions on appeal factors that apply to Bk. (A sample of the appeal term description generated using ABET for the book "The Hunger Games" by Suzanne Collins is shown in Figure 1.)

5. OUR PROPOSED RECOMMENDER

In making suggestions, Rabbit first analyzes the *profile* of a reader R, which consists of a set of N ($\geq$ 1) books either given or bookmarked[7] by R on a social bookmarking site. Based on the profile, Rabbit identifies books that are compatible with the readability level of R, which are treated as *candidate books* to be considered for recommendation. These books are selected among those available at a book repository, including (i) *reputable websites*, such as OpenLibrary.org or WorldCat.org, which are two of the largest online library catalogs, (ii) *school/public libraries*, and (iii) book-related *bookmarking sites*, such as BiblioNasium.com, which is a website that encourages reading among children/teenagers. Rabbit computes a ranking score, which quantifies the *degree of relevance*, of each candidate book with respect to (books in) the profile of R using a regression model applied to the analytical results of the book using diverse publicly available information.

5.1 Candidate Books

It is imperative for Rabbit to locate books with grade levels adequate for each individual reader, since "reading for understanding cannot take place unless the words in the text are accurately and efficiently decoded" [10]. To accomplish this task, Rabbit first determines the readability level of a reader R by analyzing the grade levels of books in R's profile. The readability level of R is determined by averaging the grade level of each book P_B in R's profile, computed using TRoLL [12], a tool for regression analysis of literacy levels,

[7] Only books bookmarked by a user during the most recent academic year are examined to account for the fact that the users enhance their reading comprehension skills over time.

which captures the *central tendency* of the grade levels of books that have been read by R. Unlike popular prediction formulas/tools (such as Flesch-Kincaid, Lexile, and ATOS [1]), which rely on text of a book to compute its grade level (a severe constraint, since text is not always freely accessible due to copyright laws), TRoLL computes the grade level of any book using book metadata publicly accessible from reputable online sources, even in the absence of sample text.

Rabbit applies Equation 1 to determine the set of candidate books considered for recommendation.

$$SCB(R) = \{CB \mid CB \in Rep \wedge$$
$$TRoLL(CB) \in [\frac{\sum_{P_B \in P} TRoLL(P_B)}{|P|} \pm 0.25]\} \quad (1)$$

where CB is a candidate book available at a book repository Rep, $|P|$ denotes the number of books in R's profile, and $TRoLL(CB)$ ($TRoLL(P_B)$, respectively) is the grade level of CB (P_B, respectively) determined by TRoLL. By selecting books within *half a grade level* of the *mean* readability level of R, Rabbit considers books for recommendation within an appropriate level of (text) complexity for R based on the grade levels of books in R's profile.

5.2 Multiple-Evidence Analysis

Rabbit suggests books that not only readers can comprehend, but also they are interested in by analyzing diverse publicly accessible metadata as described below

5.2.1 Exploring Topical Information

Rabbit examines the *topical description* (i.e., topic) of a book defined by Library of Congress Subject Headings (LCSH) assigned to the book by professional cataloguers. LCSH, a de facto controlled vocabulary, constitute the largest general indexing vocabulary in the English language. Subject headings, which are *terms* or *phrases* that denote concepts, events, or names, are used by librarians to index books according to their themes. Examples of LCSH include "Juvenile Fiction," and "Archaeology–History–18th century".

Rabbit, which explores the topical resemblance between CB and books in R's profile P, examines the degree to which the distribution of topics in CB matches the distribution of topics of books in P. This *topical similarity* measure, which is defined using the vector space model (VSM) and computed in Equation 2, prioritizes candidate books that have been assigned LCSH which match the LCSH favored by R, i.e., most frequent LCSH assigned to books in P.

$$TSim(CB, P) = \frac{\vec{CB} \cdot \vec{P}}{||\vec{CB}|| \times ||\vec{P}||} \quad (2)$$

where CB and P are represented as n-dimensional vectors $\vec{CB} = < W_{CB_1}, \ldots, W_{CB_n} >$ and $\vec{P} = < W_{P_1}, \ldots, W_{P_n} >$, respectively, n is the number of distinct subject headings assigned to CB and books in P, W_{CB_i}, the weight of CB_i ($1 \leq i \leq n$) is "1" if CB_i is a subject heading of CB, and is "0" otherwise, and W_{P_i}, which is the weight of P_i ($1 \leq i \leq n$), is computed as a proportion between the number of books in P that have been assigned P_i and the total number of books in P, i.e., $|P|$.

In computing the *topical similarity* measure, in addition to other similarity measures presented below, we rely on VSM, since VSM handles frequency distributions, which is essential in comparing candidate books based on multiple perspectives with the profile of a reader. We have empirically verified that given the short length of the descriptions for each book based on its topics, content, and appeal factors (which include very few LCSHs, words, and terms, respectively), VSM is a more reliable distance measure compared with its probabilistic counterparts, such as KL-divergence.

5.2.2 The Book Content Analysis

Besides determining the topics that are of interest to a reader R, Rabbit also identifies written matters covered in books that are generally appealing to R by analyzing the content description of each book in P (and each candidate book CB), which is extracted from reputable book-related websites, such as Amazon and the Library of Congress.

Rabbit computes the *content similarity* of each CB with respect to P based on the "bag-of-words" representation of the *brief descriptions* of CB and (books in) P. Rabbit favors candidate books with contents compatible with the contents that are most commonly addressed in books belonged to R's profile. To compute $CSim(CB, P)$, the *content similarity* score defined in Equation 3, Rabbit employs an *enhanced* version of the cosine similarity measure based on word-correlation factors (WCF) [6], which relaxes the *exact matching* constraint imposed by the cosine measure by exploring words in the content description of CB that are analogous to, besides the same as, words in the content description of books in P. Rabbit relies on WCF, as opposed to other popular similarity metrics applied to WordNet, since we have empirically verified that word-similarity scores predicted by using WCF correlate with human assessments on word similarity. Using WordSim353,[8] which is a test collection for measuring word relatedness, and STS,[9] which provides human assessments on sentence similarity for 750 pairs of sentences, we compared the performance of WCF with FaITH[10] [13], which is a feature and information theoretic-based similarity measure. (Only FaITH was considered, since as reported in [13], it outperforms other well-known information-theoretic, ontology, and hybrid-based approaches that exploit word-related information available at WordNet to estimate the degree of similarity between pairs of words.) The results of the experiments verified that the performance of WCF is comparable to that of FaITH.

$$CSim(CB, P) = \frac{\vec{CB} \cdot \vec{P}}{||\vec{CB}|| \times ||\vec{P}||} \quad (3)$$

where the content of CB and P are represented as vectors $\vec{CB}$ = <W_{CB_1}, …, W_{CB_n}> and $\vec{P}$ = <W_{P_1}, …, W_{P_n}>, respectively, n is the number of distinct non-stop, stemmed keywords in the brief descriptions of CB and each book in P, and W_{P_i} and W_{CB_i} are the *weights* of keywords P_i and CB_i, respectively, which are calculated using Equations 4 and 5 such that the frequency distributions of words in $\vec{CB}$ and $\vec{P}$ are determined based on the frequency distributions of non-stop, stemmed words among the brief descriptions of CB and all the books in P, respectively.

$$W_{CB_i} = \begin{cases} \frac{f_{CB_i, CB}}{max_{CB_i \in CB}(f_{CB_i, CB})} & \text{if } CB_i \in CB \\ \frac{\sum_{c \in HS_{P_{B_i}}} f_{c, CB}}{|HS_{P_{B_i}}|} & \text{otherwise} \end{cases} \quad (4)$$

[8] cs.technion.ac.il/~gabr/resources/data/wordsim353/

[9] goo.gl/KzMHgn

[10] Available at grid.deid.unical.it/similarity

$$W_{P_i} = \begin{cases} \frac{f_{P_i,P}}{max_{P_i \in P}(f_{P_i,P})} & \text{if } P_i \in P \\ \frac{\sum_{c \in HS_{CB_i}} f_{c,P}}{|HS_{CB_i}|} & \text{otherwise} \end{cases} \quad (5)$$

where $|HS_w|$ is the size of HS_w, $f_{w,D}$ is the *frequency of occurrence* of w in D, and HS_w is the set of words that are *highly similar* to, but not the same as, a given word w in the brief description D of either CB or P. (A word is highly similar if it is included in a reduced WCF matrix [12].)

5.2.3 Examining Appeal-Term Descriptions

Rabbit also examines the appeal elements of books preferred by R and captures the *overall degree of resemblance* between the appeal-term description of CB and (each book in) the profile P of R, which are generated using ABET. In calculating $ATSim(CB, P)$, the *appeal-term similarity* score of CB with respect to P, Rabbit adopts the cosine measure as defined in Equation 6.

$$ATSim(CB, P) = \frac{\sum_{f \in F} \frac{\vec{CB}_f \cdot \vec{P_f}}{||\vec{CB}_f|| \ \times ||\vec{P_f}||}}{|F|} \quad (6)$$

where F is the set of appeal factors in the appeal-term descriptions for CB and P, $|F|$ is the size of F, $\vec{CB}_f$ and $\vec{P_f}$ are the n-dimensional vector representations of the appeal-term distribution of an appeal factor f for CB and P, respectively, n is the number of distinct appeal terms in the distributions of the corresponding appeal factor for CB and P, $W_{CB_{f_i}} = \frac{freq_{i,CB_f}}{\max_{i \in CB_f} freq_{i,CB_f}}$ is the *weight* of the i^{th} $(1 \leq i \leq n)$ term in $\vec{CB}_f$, and $W_{P_{f_i}} = \frac{\sum_{P_B \in P} freq_{i,P_{B_f}}}{\max_{i \in P_f} \sum_{P_B \in P} freq_{i,P_{B_f}}}$ is the *weight* of the i^{th} term in P_f.

5.3 Ranking Candidate Books

To predict the ranking score of CB, Rabbit employs multiple linear regression, which is a statistical technique for building estimation models that accounts for the influence of multiple contributing factors, i.e., the topical, content, and appeal-term descriptions of CB in our case. The top-10 ranked books are recommended to R.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n \quad (7)$$

where Y is the dependent variable, i.e., ranking score of CB, β_0 is the intercept parameter, $\beta_1, \ldots, \beta_n$ are the coefficients of regression, $X_1, \ldots X_n$ are the independent variables (predictors), i.e., the scores defined in Section 5.2 for CB, and n is the number of predictors in the regression analysis.

In Equation 7, the intercept and coefficients of regression are estimated through a one-time training process using the Ordinary Least Squares method and the $Tset$ dataset (introduced in Section 6.2.1). $Tset$ consists of 1,663 instances, each of which is a book B that is either a relevant or non-relevant recommendation for a given reader R. Each instance is represented as a vector of the form $< B_1, B_2, B_3, rel_R >$, where B_i is the (value of the) i^{th} $(1 \leq i \leq 3)$ predictor computed for B, and rel_R is the target, which is "1" if B is a relevant recommendation for R, and is "0" otherwise.

6. EXPERIMENTAL RESULTS

In this section, we present the results of the empirical studies conducted to assess the performance of ABET and Rabbit. To perform these studies, we relied on a number of sample sets of books, i.e., SB_1 and SB_2. (Due to space constraints, we posted the sample sets under goo.gl/PWE9u2)

6.1 Assessing the Performance of ABET

Due to the lack of existing benchmark datasets for validating the performance of tools that automatically extract appeal factor-appeal term pairs, we have assessed the performance of ABET by (i) computing the precision and recall of appeal factor-appeal term pairs extracted from book reviews, (ii) analyzing the correctness of appeal-term descriptions created by ABET, and (iii) comparing appeal-term descriptions generated by ABET with respect to the ones extracted from NoveList on the same set of books.

We randomly selected a set of 100 books written for K-12 readers, and for each book we randomly examined a review. We *manually annotated* the appeal factor-appeal term pairs in each of the 100 examined reviews and compared the annotated pairs in each review with the ones extracted by ABET. The precision, recall, and F-measure achieved by ABET, which are 0.85, 0.82, and 0.83, respectively, verify the high accuracy of the rules defined for ABET in identifying appeal factors and their corresponding appeal terms in reviews. We have observed that majority of the pairs excluded by ABET were due to keywords used by reviewers to describe a given factor which are not included in the pre-constructed vocabulary of the corresponding factor defined for ABET (as described in Section 4). We have also observed that poor phrasing in reviews, which in turn yields nonsensical grammatical relations between pairs of words, and improper anaphora resolution in ABET have caused the majority of the extraction errors.

To further evaluate the appeal-term descriptions created by ABET, we relied on SB_1, a sample set of *eight books*, and conducted two surveys on Amazon Mechanical Turk.

In both surveys, we asked appraisers to select the keywords, i.e., appeal terms, that best describe each appeal factor for one of the books in SB_1. While the first survey includes the appeal terms considered by ABET for each appeal factor as the corresponding possible keyword choices, the second survey contains appeal terms defined by either ABET or NoveList. In the two surveys, we treated the appeal terms selected for each factor by appraisers as the "gold standard" for the factor and computed the accuracy of ABET (NoveList, respectively) based on the proportion of terms in the gold standard of a given factor defined by appraisers which match its counterpart identified by ABET (NoveList, respectively) for the factor. We relied on Mechanical Turk, since it is a "marketplace for work that requires human intelligence" that allows individuals/businesses to programmatically access thousands of diverse workers and has been used in the past to collect user feedback for information retrieval tasks [7]. Furthermore, we considered NoveList for the comparison purpose, since NoveList is a premier database for RA [3] and, to the best of our knowledge, is the only RA database that includes appeal-term descriptions for books.

As shown in Figure 2, based on the 200 responses collected in September 2013, ABET achieves an overall 94% accuracy in identifying appeal terms (in reviews) that describe a book. More importantly, the accuracy on the identified appeal terms for each appeal factor considered by ABET is in the upper eighty percentile or higher. Figure 3, on the other hand, shows the accuracy ratios of ABET and Nov-

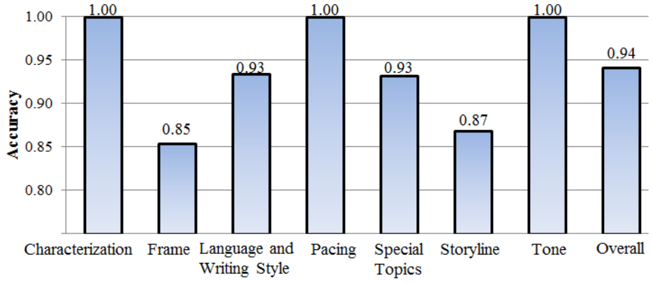


Figure 2: Performance evaluation of ABET conducted using Amazon Mechanical Turk

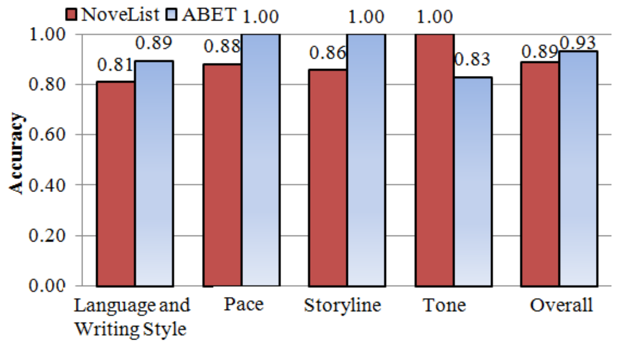


Figure 3: Performance evaluation of ABET and NoveList using Amazon Mechanical Turk

eList calculated using the 200 responses provided by Mechanical Turk appraisers for the second survey. NoveList describes books using only four appeal factors, as opposed to the seven considered by ABET. For this reason, we have compared the performance of ABET and NoveList based on their common appeal factors (as shown in Figure 3). Based on the appraisers' assessments, we claim that appeal-term descriptions provided by ABET are favoured over the ones defined by professionals who maintain the RA database at NoveList. Note that the improvement in overall accuracy ratio achieved by ABET over NoveList, in addition to the improvement on "Language and Writing Style", "Pacing" and "Storyline" factors, are statistically significant (p < 0.01).

6.2 Assessing the Performance of Rabbit

In this section, we discuss the empirical studies conducted to assess the design and performance of Rabbit.

6.2.1 Dataset and Evaluation Strategy

Even though BookCrossing[11] has been employed to evaluate book recommenders tailored to a general audience, it is not specifically designed for assessing the performance of recommenders for K-12 readers. We used data provided by BiblioNasium, which is a secure social networking site on books that targets children and teenagers, to evaluate Rabbit instead. The dataset consists of books that have been bookmarked by each one of the 5,580 BiblioNasium users who joined the site within the first four months of its establishment. A portion of the dataset, called $Tset$, which consists of 10% of the 5,580 users and their profiles, was employed for training Rabbit's regression model, whereas the remaining users and their profiles, called $Eset$, were used for evaluation purposes. The design methodology of Rabbit relies on *topical*, *brief content*, and *appeal-term* descriptions, in addition to the predicted grade levels of books. Thus, we retrieved the brief book descriptions and LCSH from reputable book-related websites, the appeal-term descriptions from book reviews using ABET, and the book readability levels using TRoLL.

We adopt the popular 5-fold cross validation strategy to evaluate recommenders. Since only withheld books are considered relevant, it is not possible to account for potentially relevant books a user has not bookmarked, a well-known limitation of this evaluation protocol. As the limitation applies to all the recommenders evaluated in the empirical studies, the results are consistent for the comparison purpose.

6.2.2 The Performance Evaluation of Rabbit

We assessed the performance of Rabbit using $Eset$ in terms of *Normalized Discounted Cumulative Gain* (nDCG), which determines the overall (ranking) performance of a recommender and penalizes relevant recommendation positioned lower in the recommendation list. We observed that ranking candidate books solely based on *appeal factors* or *topical information* yields the lowest nDCG scores, i.e., 0.19 and 0.18, respectively This is anticipated, since LCSH and appeal-term descriptions mainly identify the types of books preferred by a reader R from a general perspective. Even though the content-based approach yields a relatively high nDCG score (i.e., 0.24), the experimental results show that the multi-dimensional strategy of Rabbit, of which content-based analysis is a component, locates more relevant books, which is justified by the statistically significant improvement in nDCG (i.e., 0.32) generated by the latter ($p < 0.001$).

Linearly combining different similarity scores considered by Rabbit yields a lower nDCG value (i.e., 0.27) than the one obtained by using the regression model, which is statistically significant ($p < 0.001$). The results validate the necessity of accounting for the impact, i.e., weight, of each individual similarity score. Furthermore, we have verified that bypassing the candidate selection step would have a *negative* impact on the overall performance of Rabbit, since recommending books that only match the interests or general traits of books preferred by a reader R, without considering R's readability level, increases the number of non-relevant suggestions (as demonstrated by the 0.19 nDCG achieved by Rabbit if no candidate selection step is applied.)

6.2.3 Rabbit versus BReK12

We also compared the performance of Rabbit with BReK12 (as introduced in Section 2). We perform the comparison, since to the best of our knowledge BReK12 is the only available recommender that explicitly considers the readability level of its users in making personalized book recommendations. Furthermore, other state-of-the-art approaches for (book) recommendations are excluded for the comparison purpose using $Eset$, since they require either *personal ratings* on books provided by individual users or *social connections* established by social bookmarking site users, neither are available on social websites for K-12 readers nor in the $Eset$ dataset. Rabbit achieves a statistically significant improvement ($p < 0.001$) over BReK12 in terms of nDCG, which are 0.32 and 0.18, respectively.

[11] Informatik.unifreiburg.de/∼cziegler/BX

6.2.4 Rabbit versus Other Recommendation Modules

To further validate the performance of Rabbit, we conducted a survey using Mechanical Turk on 10 sample books in another test set SB_2, which evaluated the degree to which books recommended by Rabbit are preferred over those suggested by recommendation modules at well-known book-related websites. We have selected recommenders that adopt diverse strategies in making suggestions: (i) Amazon, which considers purchasing patterns of its users [8], GoodReads,[12] which combines multiple proprietary algorithms to "analyze 20 billion data points," and (iii) NoveList,[13] which examines book-related information, including title, publication date, and appeal factors for recommending books.

Each survey included the top-2 (for which some of them can be identical) recommendations made by the aforementioned recommenders for a given sample book Bk. Appraisers were asked to select, to the best of their knowledge, the top-two books most closely related to Bk, which were treated as the *gold standard* for Bk. Based on the 500 responses collected during November 2013, we computed the accuracy of the top-2 recommendations made by Amazon, GoodReads, NoveList, and Rabbit, which are 0.50, 0.18, 0.24, and 0.44, respectively. Recommendations made by Rabbit and Amazon are preferred over the suggestions made by GoodReads and NoveList. Furthermore, the improvement, in terms of accuracy ratios, achieved by Rabbit over GoodReads and NoveList is statistically significant ($p < 0.001$). In terms of the overall accuracy, Amazon outperforms Rabbit. However, their differences in nDCG are not statistically significant ($p < 0.001$).

Rabbit can make suggestions regardless of the number of books of interest to R. It differs from the recommendation modules employed at Amazon and NoveList, since the latter can only examine books of interest to a user one at the time. Rabbit is also different from GoodReads, since GoodReads processes either a given book or the entire profile of a user. Furthermore, Rabbit can treat a book as a candidate suggestion immediately after the book is published, whereas Amazon requires the existence of a number of purchasing transactions involving the new book in order to suggest it to a user. In addition, in making recommendations for children and teenagers, Rabbit considers books provided directly by K-12 readers to generate personalized suggestions. Recommendations generated by Amazon that target children and teenagers, on the other hand, are the result of extensive analysis of the purchasing patterns of adults.

7. CONCLUSIONS

We have introduced Rabbit, a recommender which makes personalized suggestions on books that match the *interests* and *reading abilities* of its K-12 users. Rabbit emulates the readers' advisory process offered at public/school libraries to recommend books that are similar in *contents*, *topics*, and *literary elements* of other books appealing to a reader, with the latter based on extracted appeal-term descriptions.

Rabbit can be a stand-alone tool used by readers (educators/parents, respectively) or adopted by (K-12) social bookmarking sites for providing suitable reading selections. Even though we have developed Rabbit with K-12 readers in mind, the readers' advisory-based methodology of Rabbit can also be used to suggest books for adults (with reading levels below/above the 12^{th} grade level) as well.

The results of the conducted experiments have (i) validated the correctness of the design methodology of Rabbit and (ii) demonstrated the superiority of Rabbit over other recommenders that either explicitly consider or ignore the reading ability of its users by using data from BiblioNasium and Mechanical Turk appraisers.

[12] goo.gl/99me5f

[13] support.epnet.com/knowledge_base/detail.php?id=4772

8. REFERENCES

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

[2] C. Coulter and M. Smith. The Construction Zone: Literary Elements in Narrative Research. *Educational Researcher*, 38(8):577–590, 2009.

[3] A. Cox and K. Horne. Fast-Paced, Romantic, Set in Savannah: A Comparison of Results from Readers' Advisory Databases in the Public Library. *Public Library Quarterly*, 31(4):285–302, 2012.

[4] S. Givon and V. Lavrenko. Predicting Social-Tags for Cold Start Book Recommendations. In *ACM RecSys*, pages 333–336, 2009.

[5] N. Hollands. Improving the Model for Interactive Readers' Advisory Service. *Reference & User Services Quarterly*, 45(3):205–212, 2006.

[6] J. Koberstein and Y.-K. Ng. Using Word Clusters to Detect Similar Web Documents. In *KSEM*, pages 215–228, 2006.

[7] M. Koolen, J. Kamps, and G. Kazai. Social Book Search: Comparing Topical Relevance Judgments and Book Suggestions for Evaluation. In *ACM CIKM*, pages 185–194, 2012.

[8] G. Linden, B. Smith, and J. York. Amazon.com Recommendations: Item-to-Item Collaborative Filtering. *IEEE Internet Computing*, 7(1):76–80, 2003.

[9] N. Manouselis, H. Drachsler, K. Verbert, and E. Duval. *Recommender Systems for Learning.* Springer Briefs in Electr. and Comp. Eng., 2013.

[10] J. Oakhill and K. Cain. The Precursors of Reading Ability in Young Readers: Evidence From a Four-Year Longitudinal Study. *SSR*, 16(2):91–121, 2012.

[11] B. Pang and L. Lee. Opinion Mining and Sentiment Analysis. *FTIR*, 2(1-2):1–135, 2008.

[12] M. Pera. *Using Online Data Sources to Make Recommendations on Reading Materials for K-12 and Advanced Readers.* PhD thesis, BYU, April 2014.

[13] G. Pirro and J. Euzenat. A Feature and Information Theoretic Framework for Semantic Similarity and Relatedness. In *ISWC*, pages 615–630, 2010.

[14] J. Saricks. *Readers' Advisory Service in the Public Library,* 3^{rd} *Ed.* ALA Store, 2005.

[15] R. Shaban. A Guide to Writing Book Reviews. *JEPHC*, 4(3):Article 11, 2006.

[16] A. Sieg, B. Mobasher, and R. Burke. Improving the Effectiveness of Collaborative Recommendation with Ontology-based User Profiles. In *ACM HetRec*, pages 39–46, 2010.

[17] C. Yang, B. Wei, J. Wu, Y. Zhang, and L. Zhang. CARES: a Ranking-oriented CADAL Recommender System. In *ACM/IEEE JCDL*, pages 203–212, 2009.