# Retrieval Algorithms Optimized for Human Learning

Rohail Syed
School of Information
University of Michigan
105 S. State St.,
Ann Arbor, Michigan 48109
rmsyed@umich.edu

Kevyn Collins-Thompson
School of Information
University of Michigan
105 S. State St.,
Ann Arbor, Michigan 48109
kevynct@umich.edu

## ABSTRACT

While search technology is widely used for learning-oriented information needs, the results provided by popular services such as Web search engines are optimized primarily for generic relevance, not effective learning outcomes. As a result, the typical information trail that a user must follow while searching to achieve a learning goal may be an inefficient one involving unnecessarily easy or difficult content, or material that is irrelevant to actual learning progress relative to a user's existing knowledge. We address this problem by introducing a novel theoretical framework, algorithms, and empirical analysis of an information retrieval model that is optimized for learning outcomes instead of generic relevance. We do this by formulating an optimization problem that incorporates a cognitive learning model into a retrieval objective, and then give an algorithm for an efficient approximate solution to find the search results that represent the best 'training set' for a human learner. Our model can personalize results for an individual user's learning goals, as well as account for the effort required to achieve those goals for a given set of retrieval results. We investigate the effectiveness and efficiency of our retrieval framework relative to a commercial search engine baseline ('Google') through a crowdsourced user study involving a vocabulary learning task, and demonstrate the effectiveness of personalized results from our model on word learning outcomes.

## KEYWORDS

Retrieval Models and Ranking; Intrinsic diversity; Assessment of learning in search

## 1 INTRODUCTION

Considerable research attention has focused on how computer-based learning can supplement traditional classroom learning environments [15]. As a supplement to traditional forms of learning, a computerized approach via technology such as intelligent tutoring systems (ITS) offers numerous benefits in the form of personalized instruction, modern and up-to-date educational material and the comfort of self-paced learning [15]. However, with all these benefits come a number of challenges. A computational teaching approach must somehow represent the state of a student's knowledge, and then, as in traditional classroom instruction, infer the knowledge required for students to progress in the curriculum to a certain goal. In particular, the right learning resources must be selected for each student that enables them to learn material effectively and efficiently, enabling them to do well in future assessments on the material.

At the same time, Web search is now a key information source that learners use to obtain their own resources for many education-related tasks [1]. Algorithms have been developed to enhance Web search for learners by providing intrinsically diverse results for topic exploration [17] or by personalizing results so that they are at an appropriate level of reading difficulty [5]. However, all of these previous retrieval methods used algorithms trained to optimize relevance, not learning, and their evaluations similarly focused on improvements in generic relevance, not actual learning outcomes.

In this study, we bring together these two fields of research to develop and evaluate the first retrieval framework that attempts to optimize directly for learning gains. We do this by first developing a theoretical model that casts information retrieval as the search for a optimal 'training set' of documents that optimize a utility function that is the composition of a computational model of learning with a relevance-based objective function. While this results in a difficult optimization problem, inspired by developments in *machine teaching* [22] we give an approximate, efficient algorithm to solve the retrieval problem that finds document sets that are personalized for the prior knowledge of individual learners, cover all constituent aspects of a topic, and optimize the learning gains of the student on future tests for that topic. Our model also naturally incorporates the notion of effort as a key component of the retrieval objective. We evaluate our approach with an empirical study that examines both the effectiveness and efficiency of learning outcomes on a vocabulary learning task. In effect, we show how to build a search engine that optimizes retrieval utility as seen through the 'lens' of human cognition.

## 2 RELATED WORK

Our work brings together three different research areas: core retrieval algorithms and evaluation metrics, search engine support for learning, and intelligent tutoring systems.

**Core retrieval algorithms and evaluation metrics.** The vast majority of research on retrieval algorithm optimization and evaluation has assumed metrics based on generic relevance. In contrast, we believe that ours is the first study to develop theoretical models and retrieval algorithms that explicitly optimize results for an educational goal as a key retrieval objective. Researchers have recognized the importance of going beyond traditional retrieval evaluation measures to consider user progress over time [19] as well as degree of effort [21], but little, if any, of that work has involved modeling or assessing human learning.


*SIGIR '17, August 07-11, 2017, Shinjuku, Tokyo, Japan*

ACM ISBN 978-1-4503-5022-8/17/08…$15.00
https://doi.org/10.1145/3077136.3080835

Prior work on understanding users' information needs has categorized the types of web search queries users issue as being navigational, transactional or informational [9][4]. Of particular interest here are the informational queries, whose intent is characterized as "to acquire some information assumed to be present on one or more web pages" [4]. Their study found that such queries account for the most queries of the three categories. A later study using the same search engine and breaking the taxonomy down to a finer granularity found similar results [18]. The study more specifically found that the predominant type of informational queries were undirected informational queries, characterized as: "I want to learn anything/everything about my topic. A query for topic X might be interpreted as 'tell me about X.'" [18]. Studies have shown that a majority of queries are in fact informational [9][4] and that many search sessions involve multiple queries [15]. This finding has been consistent across multiple popular search engines [9][1]. The focus in this paper is on queries that a student issues that are meant to explore a topic and are intrinsically diverse (ID) [16]. [17]. Such queries represent a specific topic of interest but their underspecified nature allows for retrieval of multiple sub-topics, or knowledge components. We formulate a strategy that estimates these sub-topics beforehand and provides the student a re-ranked set of documents that optimally teaches them about the topic by effectively covering of the sub-topics.

**Search engine support for learning.** Exploring new topic areas and learning important domain vocabulary is one popular instance of learning tasks that have been identified in search [1]. Because search engines are a primary conduit of educational information for learners, a number of studies have focused on understanding and supporting learning tasks in search, particularly on the Web. A study by Eickhoff et al. [10] investigated learning behaviors of Web search users, but used only indirect evidence via implicit indicators derived from Web search logs, rather than direct assessment of users. That study also did not develop or assess new retrieval algorithms that could be adapted to improve learning outcomes. Collins-Thompson et al. [5] explored how reading difficulty – a component of user effort – could be effective as a personalized ranking feature, but did not optimize, or assess its effectiveness, for actual learning tasks or outcomes.

While Raman et al. [17] demonstrate the effectiveness of an intrinsically diverse search ranking relative to a baseline in terms of standard IR measures, their study did not assess how well an intrinsically diverse ranking helped users learn. Therefore, subsequent work by Collins-Thompson et al. [6] implemented the intrinsic diversity algorithm formulated in [17] and evaluated improvements to learning in terms of low and high levels of cognitive complexity, as specified by the revised Bloom's taxonomy [12]. While the study in [6] did evaluate the learning gains offered by an intrinsically diverse search engine, they did not specifically develop a re-ranking objective to optimize for learning purposes, as we do here. We build on this prior work by modifying the objective function initially specified by Raman et al. to incorporate an emphasis on better learning effectiveness, and by exploring a modified variant of the ID algorithm in the context of a vocabulary learning task.

**Intelligent tutoring systems.** Prior studies have demonstrated that personalized traditional education can yield enormous gains in learning [3]. However, personalized instruction can be both expensive and difficult to scale, motivating the exploration of effective automated approaches, typically through the development of intelligent tutoring systems. Corbett [7] showed that existing intelligent systems can already get quite close to the improvement offered by personalized human tutors but so far there has been little evidence that this generalizes to learning any topic with a vast unstructured set of resources (e.g. learning through Web searches). Pirolli et al. [15] demonstrate an ITS model that yields highly significant learning gains and can strongly predict a learner's future performance on test questions. In their system, the tutor model manually decides which documents would be best for a learner to read for the given set of test questions. The expert model automatically decides which documents most closely provide answers to the test question. Their system makes the assumption that the test questions selected accurately test knowledge of subject $J$ and that the knowledge needed to answer these questions can be represented in some finite quantitative form. We make these same assumptions in our model. Our study is primarily different from [15] in that our focus is on investigating how information retrieval (IR) algorithms can be directly optimized to improve a student's learning. In contrast, in [15], learners could find and choose the documents themselves and the choice of search algorithm was not controlled: while learners could use the tutor's recommendations, they were not required to.

## 3 LEARNING-OPTIMIZED RETRIEVAL MODELS

In a typical intelligent tutoring system, a student wants to learn about some subject $J$ and the system delivers resources that optimally teach subject $J$. As in traditional education, the learner's performance would be measured by how well they perform on a test that covers knowledge of subject $J$. Given a set of test questions and an estimate of the user's current knowledge state, an effective system would be able to select a set of teaching resources that updated the learner's knowledge state so that they maximized their expected score on a future test. This individualized set of teaching resources would also be selected to minimize the effort required to obtain this optimal score. The objective of the ideal ITS system is then twofold: (1) to develop an expert model that can accurately determine the knowledge required to answer a set of test questions and (2) to develop a tutor model that can retrieve an optimal set of personalized resources.

In this paper we describe both elements, but focus on the retrieval algorithm needed for an effective tutor model. Mathematically, we start by defining a high-level retrieval objective inspired by recent work in *machine teaching* [22], which finds optimal training sets for a given learning algorithm. Our overall goal here is to find a set of documents $\mathcal{D}$ that maximizes a user's utility $U$, which is defined in terms of their learning outcome $H(\mathcal{D})$ while minimizing the effort $E(\mathcal{D})$ associated with the retrieved document set. This high-level objective is

$$\arg\max_{\mathcal{D}} U(\mathcal{D}) \quad = H(\mathcal{D}) - \lambda\, E(\mathcal{D}) \tag{1}$$

where the parameter $\lambda$ controls the tradeoff between learning outcome and effort. In this framework, there are many possible ways to define learning outcomes $H(\cdot)$ and effort $E(\cdot)$, some of which we discuss later.

In general, given the combinatorially large number of potential document sets $\mathcal{D}$, solving Eq. 1 quickly enough for real users requires us to approximate this problem. To do this, we build on an

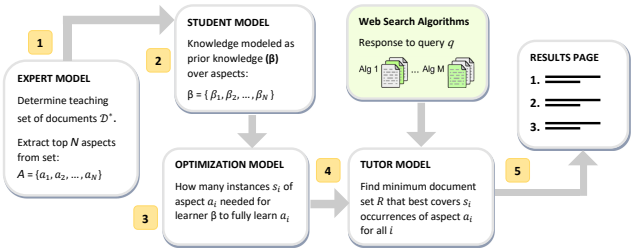


**Figure 1: High-level learning-optimized retrieval process**

approach described by Zhu [22], in which we proceed in two steps. First, we find a set of optimal sufficient statistics $S$ that maximizes the combined learning and effort objective, which any candidate document set to be considered should satisfy. Second, we obtain a pool of candidate documents using an existing retrieval framework, and select a subset of them that approximately satisfies the desired sufficient statistics $S$.

To solve the first step, finding optimal sufficient statistics $S$, we solve an adjusted form of Eq. 1, but in terms of $S$

$$\arg\max_{S} U(S) \quad = H(S) - \lambda E(S) \tag{2}$$

We show how to derive specific definitions for $H(\cdot)$ and $E(\cdot)$ in Sec. 3.1 and Sec. 3.2, using assumptions based on simple computational learning models from cognitive psychology, culminating in the complete specific form of the objective in Eq. 5 in Section 3.2. We solve the second step, using the optimal $S$ from the first step to select documents for the tutor model, in Sec. 3.3. As part of developing this retrieval model, we now walk through the various stages of our learning-optimized retrieval process as shown in Figure 1.

## 3.1 Retrieval Components

While human learning is a rich, complex phenomenon whose goals can be categorized into a multi-level hierarchy [2], in this paper we focus on one of the lower-level form of learning task and outcome: *vocabulary learning*, since it is a key precursor to other higher-level learning activities and has relatively more straightforward and well-understood assessment measures. We also make a number of simplifying assumptions to make the retrieval model tractable.

We must first choose a knowledge representation for the vocabulary learning process. We assume the widely-used *knowledge-tracing model*, implemented so that knowledge of a subject area $J$ is modeled in the form of a weighted multinomial distribution over knowledge aspects $A_k$. For example, in learning about the subject of 'igneous rocks' might involve learning about aspects that include their formation, location, and so on. For vocabulary learning, we assume these aspects correspond to $K$ specific technical keywords.

Then, we can say that the knowledge to answer test question $T_k$ is a function of how many units of relevant information the student has seen and the relative importance of that increase in knowledge. As a student only needs to read a finite amount of information about $J$ to achieve mastery of the subject, we compute a target distribution $S = \{S_1, \ldots, S_K\}$ of how many examples of each of the $K$ aspects the student needs to read in context for their learning to be considered complete. (We explain in Section 3.2.1 how these $S_k$ are estimated.)

**Expert Model**. The expert model, (Step 1 of Figure 1), is responsible for curating the set of documents $\mathcal{D}^*$ that best represents the knowledge aspects $A_k$ of the subject. In this process, the expert will generate a frequency multinomial of the $A_k$ keywords given by the bag-of-words contents of $\mathcal{D}^*$. We define this target distribution to be our aspect weights $W$.

While any number of possible algorithms could be used in this process, we focus on using a term frequency measure weighted by the log of a global corpus frequency.[1] Specifically, in constructing $W$, we first represent $\mathcal{D}^*$ as a bag-of-words model and determine the important keywords in the document set by considering the term frequencies normalized by their log total occurrences from a large corpus. This approach will give us vocabulary words that the student likely hasn't learned yet and which occur frequently in $\mathcal{D}^*$. To avoid getting keywords that may be rare but not topically relevant, we weighted these values by their word2vec [14] similarity to the first term in the base query $q$. Specifically, for each unique word $T_i$ in the bag-of-words of $\mathcal{D}^*$ we determine the important words by the score:

$$Score(T_i) = \frac{TF_i}{\log Total_i} \cdot word2vec(T_i, q)$$

We next extract the top $K$ highest-scored keywords and generate $W$ as the maximum likelihood estimation of these keywords' TF values[2]. Specifically, for $i = 1, 2, \ldots K$:

$$W_i = TF_i \cdot \left( \sum_{j=1}^{K} TF_j \right)^{-1}$$

For example, for the subject "igneous rocks" and $K = 5$, we get the distribution `W={'igneous':0.302, 'magma':0.178, 'felsic':0.057, 'mafic':0.069, 'rocks':0.394}`. Table 1 shows the top 5 keywords out of $K = 10$ for five different topics along with their corresponding weights.

Let us now consider the fact that different students may have different learning rates and different background knowledge. To avoid offering suboptimal reading resources, we must account for these differences by defining a *student model*.

**Student model**. The student model (Step 2 of Figure 1), represents the knowledge state of the student who is learning about subject $J$. To find documents that teach the student, we can simply find the set of documents that minimally reaches the required set of counts $S$. However, this approach ignores: (1) the fact that document length or keyword coverage is not necessarily indicative of topical relevance or quality and (2) different students may already know about certain aspects of $J$ and their time would be better spent learning the aspects that they don't know.

We assume that we can measure a student's learning outcome in terms their performance on a test on the given subject, so that we can assess learning by measuring a learner's performance on a set of $K$ test questions $T = \{T_k\}$ on those aspects (keywords). We code the learner's responses via the set $Y$ of binomial variables $Y_k$ such that:

$$Y_k = \begin{cases} 1 & \text{student answered } T_k \text{ correctly} \\ 0 & \text{otherwise} \end{cases}$$

[1] We used the British National Corpus (BNC)

[2] Words that were semantically the same as an earlier ranked word (word2vec similarity > 0.3) but with different tense/form were removed. Example: 'rock', 'rocks'

| Topic | Keyword 1 | Keyword 2 | Keyword 3 | Keyword 4 | Keyword 5 |
|---|---|---|---|---|---|
| Igneous rock | rocks (.308) | igneous (.236) | magma (.139) | minerals (.081) | basalt (.056) |
| Tundra | tundra (.346) | arctic (.211) | plants (.127) | permafrost (.088) | soils (.075) |
| Phrenology | phrenology (.382) | brain (.164) | skull (.103) | science (.079) | perception (.073) |
| Pottery | pottery (.517) | clay (.145) | pots (.105) | potters (.061) | ceramic (.057) |
| Synapse | neurons (.385) | electrical (.169) | axon (.100) | synapse (.077) | membrane (.069) |

**Table 1: Top 5 (out of 10) selected keywords for five topics, sorted by descending keyword weights $W_i$. The keywords to be learned range from easy ('rock') to technical ('permafrost').**

We also make the assumptions that the student is a Bayesian learner and has no memory loss (post-reading knowledge is never less than pre-reading knowledge). We further assume that reading an instance of keyword $A_k$ will monotonically increase the student's knowledge of that aspect (Step 3 of Figure 1). As such, we define the student's prior knowledge $\beta$ to be a vector of how many instances of each keyword we expect them to have read before being provided $\mathcal{D}$. Then, we have:

$$\beta = \{\beta_1, \beta_2, \ldots, \beta_K\}$$

We assume the widely-used *item response theory* (IRT) model [11] as our cognitive learning model that defines the probability of a correct response $Y_k$ on test $T$ as a logistic function of user and item parameters:

$$P(Y_k = 1 \mid U, W_k, \beta_k, L_k, S_k) = \left(1 + e^{-f(U, W_k, \beta_k, L_k, S_k(\mathcal{D}))}\right)^{-1}$$

Here, the IRT model parameters are:

- $U$ - The user's individual learning rate. This is defined such that the faster a student can learn, the less resources they will require to complete their understanding of $J$. In this study, we assume a fixed $U$ for all users.
- $W_k$ - The weight given to term $k$ where $W$ is the weight multinomial defined in the Expert model. Terms with higher weight assigned are more important for the student to learn and hence are assigned higher number of $S_k$.
- $\beta_k$ - The student's prior knowledge of keyword $k$, measured by the number of instances of $k$ the student has already seen before being provided the document set $\mathcal{D}$.
- $L_k$ - A parameter that quantifies the ease of learning for keyword $k$.
- $S_k(\mathcal{D})$ - The target instances of keyword $k$ the student sees in document set $\mathcal{D}$.

Now we define the function $f(\cdot)$ to be the log weighted sum of the total instances of the keyword the student has learned (prior knowledge + post-reading knowledge):

$$f(U, W_k, \beta_k, L_k, S_k(\mathcal{D})) = \log\left((\beta_k + S_k) \cdot W_k \cdot L_k \cdot U\right)$$

With these operational settings, we can then more specifically define the expected learning for the $k^{th}$ term as:

$$P(Y_k = 1 \mid U, W_k, \beta_k, L_k, S_k) = \left(1 + e^{-\log((\beta_k + S_k) \cdot W_k \cdot L_k \cdot U)}\right)^{-1}$$

## 3.2 Optimization model

We are now ready to define the learning outcome and effort components of Eq. 2 to obtain our final optimization problem for finding the optimal sufficient statistics $S = \{S_k\}$.

*3.2.1 Learning objective.* Prior to providing the document set, we assume that we already know the user's learning rate $U$, the weights multinomial $W$ and the learner's prior knowledge of the topic $\beta$. Then, our objective is to find the total instances of each aspect keyword the student *still* needs to read in order to have learned the subject. Aggregating across all $K$ aspect terms, we can now define the learning outcome part of the objective $H(S)$ in Eq. 2.

$$H(S) = \sum_{k=1}^{K} \left(1 + e^{-\log((\beta_k + S_k \cdot W_k \cdot U)}\right)^{-1} \tag{3}$$

*3.2.2 Effort objective.* There are two types of effort we could attempt to quantify in the $E(\mathcal{D})$ term of Eq. 1 for a given document set $\mathcal{D}$ (and thus for the proxy objective for $S$ in Eq. 2). First, we could define the effort (or cost) involved on the part of the search engine in obtaining $\mathcal{D}$, e.g. from content access fees, retrieval latency, or other factors. Second, we could define the effort involved on the part of the user in reading and understanding $\mathcal{D}$. In this paper we focus only on the latter, i.e. the user's effort, as the most relevant factor for our study, but future systems could easily capture both types in our model.

Various studies (e.g. [21]) have explored user effort factors associated with processing retrieved documents. For example, depending on the specific application, we might define $E(\mathcal{D})$ as a function of the vocabulary difficulty levels of all words in the documents [5]. For simplicity, in this study we simply define $E(\mathcal{D})$ in terms of the total amount of subject reading a student has to do, i.e. the total count of keywords $S_k$ in the document set $\mathcal{D}$. The tradeoff parameter, $\lambda$, sets the (possibly student-specific) preference for less or more reading material as a tradeoff against learning utility. In this study, we keep the effort tradeoff parameter $\lambda$ constant[3] but in future work, we will investigate how student-specific penalties can change learning outcomes. So we can express effort in terms of total keywords as:

$$E(S) = \sum_{k=1}^{K} S_k \tag{4}$$

*3.2.3 Combined objective.* Combining the learning and effort functions as in Eq. 2 gives us the final optimization problem for obtaining the desired keyword counts $S = \{S_k\}$:

$$\begin{aligned} &\arg\max_S H(S) - \lambda E(S) \\ &= \arg\max_S \sum_{k=1}^{K} \left(1 + e^{-\log((\beta_k + S_k) \cdot W_k \cdot L_k \cdot U)}\right)^{-1} - \lambda \sum_{k=1}^{K} S_k \end{aligned} \tag{5}$$

[3]The parameter $\lambda$ was set to 0.0060 based on simulated resultant $S_k$ values. We have similarly kept the keyword weights $W$ as uniform to avoid confounding variables in our experiment design.

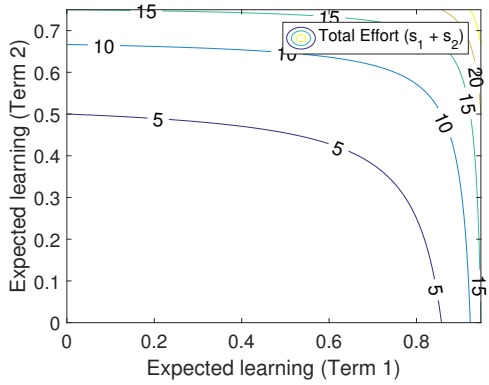


**Figure 2: Possible tradeoffs in expected learning for each of two keywords (Term 1, Term 2) in a topic. Isolines show points of constant effort (total keyword instances read). Expected learning for each keyword is based on the logistic IRT definition above. (Ease of learning parameters for each keyword are set to $L_1 = 1.2$ and $L_2 = 0.2$ respectively.)**

Note that without an effort penalty, we could simply maximize the learning utility by training the user with an unlimited number of documents, so effort plays an important role as a 'soft' constraint in the overall optimization problem. This particular instance of the problem can be solved using standard nonlinear optimization techniques to obtain the $S_k$ values[4].

Figure 2 illustrates a simple instance of this optimization problem that shows the learning/effort tradeoff based on a topic with two keywords, using the sigmoid objective above. For a fixed total number of keywords to be read (shown by the isolines with total effort of 5, 10, 15, etc), there is a opportunity cost: for every additional unit assigned to one keyword, there also may be an expected potential loss in knowledge caused by *not* assigning the user's attention to the other keyword. In general these tradeoffs will also be affected by the ease of learning $L_k$ for the $k^{th}$ keyword such that words that are easier to learn will result in a higher expected learning for the same total units assigned. For example, Fig. 2 shows that for a total effort of reading $S_1 + S_2 = 5$ keyword instances, to get the same expected learning for both keywords (i.e. the point that intersects the line $y = x$), we would assign $S_1 = \frac{1}{1.4}$ and $S_2 = \frac{6}{1.4}$ to get an expected learning outcome (probability of a correct test result) of 0.461 for both keywords.

## 3.3 Final retrieval stage: tutor model

Once the student model has determined how many instances of each target word the student must read (Step 3 of Figure 1), the last step is to design a retrieval algorithm that finds a set of Web documents that best covers these $S$ optimal target word instances. Because users are learning the expert keywords solely from exposure to them in context, the quality and diversity of these contexts is important for learning. Thus, it is important to use a retrieval approach that finds documents that show the target keywords being used in a diverse set of contextually relevant contexts for a given topic. Toward this goal, we chose intrinsically diverse ranking [17] as our retrieval framework, because of its ability to jointly identify key subtopics along with the subtopics' most relevant documents.

**Tutor model.** The tutor model decides what is the best set of documents $\mathcal{D}$ to provide the student to learn from (Step 4 of Figure 1). The tutor will first need to choose a retrieval algorithm to fetch a ranked list of candidate documents. In particular, as we will show in this section, our optimization objective very closely matches that used by [17] in their objective function. Next, the tutor will initiate with an empty set of documents $\mathcal{D}$ and will then greedily add candidate documents from the ranked list we retrieved. The stopping criteria for adding documents is determined by the cumulative keywords count in the bag-of-words of $\mathcal{D}$, given as vector $C_{\mathcal{D}}$. We will keep adding documents to $\mathcal{D}$ (Step 5 of Figure 1) until the following condition is reached:

$$C_{\mathcal{D}k} \geq S_k \quad k = 1, \ldots, K$$

In developing our choice of document retrieval algorithm, we required that it must incorporate: (1) a measure of document relevance to both the base query $q$ and the subtopic queries $q_i$ (that is, $Rel(d|q)$ and $Rel(d|q_i)$ respectively); (2) a measure $\eta$ to avoid selecting redundant documents; (3) and a measure $\epsilon$ to maximize keyword density so that the student is provided with concise but relevant material.

We note that the intrinsic diversity objective developed by Raman et al. [17] meets most of these criteria and our generalization of it in [20] meets all the criteria with the objective:

$$\arg\max_{\mathcal{D}} \sum_{i=1}^{|\mathcal{D}|} Rel(d_i|q) \cdot Rel(d_i|q_i) \cdot e^{\delta \eta_i} \cdot e^{\alpha \epsilon_i} \tag{6}$$

Our generalization of this objective function differs from that proposed in [17] in two ways. Firstly, the $\eta_i$ term that we use incorporates a measure of document novelty which the function by Raman et al. [17] did not.

Secondly, in our generalization, we add the $\epsilon_i$ term which is defined as the contribution document $d_i$ offers in terms of how much closer it brings the student towards the required keyword counts $S$. We measure the contribution in terms of the keyword density of the document - that is, the useful keywords this document offers normalized by the total words the student has to read in this document. By normalizing the contribution, our algorithm will prefer documents that cover the content in less overall text.

We note that the implementation of the intrinsic diversity algorithm in [17] is essentially a special case of (6) where $\epsilon_i = 0$. A higher value of $\alpha$ gives a ranking that prioritizes finishing the reading in as few total words possible. This ensures the student expends the least possible effort but comes with the tradeoff of quality of the documents the student gets.

A recent study evaluated the learning gains using a similar objective function and showed that for several topics, putting the entire emphasis on the $\epsilon_i$ term (by setting $\alpha = \infty$) resulted in the highest learning gains per word read [20]. As they state in their study, a more rigorous method for selecting $\alpha$ values to test is needed and will be an area of future work.

While we have considered factors of relevance, document length and keyword coverage, we have not yet considered the reading difficulty of the given document. In particular, given two documents $d_1$ and $d_2$ that are identical in the above measures but differ in their vocabulary levels (one may be for a more expert audience and the

[4]The general form of this objective is a sum-of-sigmoids, which for more sophisticated models may require sigmoidal optimization methods.

other for a novice audience), we want to choose the document that is easier for the participant to learn from. So we normalize the keyword coverage instead by the *weighted* document length where for word $w_k$, we have Age-of-Acquisition readability score $r_k$. We get these scores from the study by Kuperman et al. [13].

Specifically, let $C_{\mathcal{D}}$ represent the cumulative keyword counts the student has seen so far and $C_R$ represent the keywords frequency distribution of $d_i$. we have:

$$\epsilon_i = \left(\sum_{k=1}^{|d_i|} r_k\right)^{-1} \sum_{j=1}^{K} \begin{cases} C_{Rj} & C_{Rj} + C_{\mathcal{D}j} \leq S_j \\ \max(0, S_j - C_{\mathcal{D}j}) & \text{else} \end{cases} \quad (7)$$

The $\alpha$ variable in the objective function (Equation 6) controls how much importance we give to optimizing for keyword density versus intrinsic diversity. While we assume values of $\alpha$ are chosen manually, an automated approach to finding the optimal $\alpha$ is a natural next step for future work.

Algorithm 1 gives the final retrieval process which takes as input the vector $S$ and a set of candidate documents $D_i$ for each subtopic query where $D = \{D_i\}$. It then iteratively determines the candidate document that best improves the objective function (6), adding it to $\mathcal{D}$. All Web search queries issued in any part of this study were issued through the Google Search API[5] and we collected the top 70 links from each query as candidate documents.

**Algorithm 1** IntrinsicTeacher algorithm that ranks documents for the vocabulary learning task. First developed in [20].

**Input:** $D_i$ as Google search results for subtopic query $q_i$ for all $Q$
$C_{dk}$ given as vector of keyword counts in document $d_k \in D_i$.
$C_{\mathcal{D}}$ given as cumulative keyword counts for aspect $j \in A$ covered in $\mathcal{D}$.
$S$ given as total required keyword counts for aspect $j \in A$.
**Output:** $\mathcal{D}$ given as output document set.

$\mathcal{D} \leftarrow \emptyset$
$C_{\mathcal{D}j} \leftarrow 0 \ \forall j \in C_{\mathcal{D}}$
**while** $\exists C_{\mathcal{D}j} : C_{\mathcal{D}j} < S_j$ **do** ▷ exit when all $C_{\mathcal{D}j} \geq S_j$
  $bestS \leftarrow 0$
  $bestD \leftarrow \emptyset$
  $C_D \leftarrow \emptyset$
  **for all** $q_i \in Q$ **do**
    **for all** $d_k \in D_i, d_k \notin \mathcal{D}$ **do**
      $docS \leftarrow Rel(d_k|q) \cdot Rel(d_k|q_i) \cdot e^{\beta \eta_i} \cdot e^{\alpha \epsilon_k}$
      **if** $docS > bestS$ **then**
        $bestS \leftarrow docS$
        $bestD \leftarrow d_k$ ▷ document with highest $bestS$
        $C_D \leftarrow C_{dk}$
      **end if**
    **end for**
  **end for**
  $\mathcal{D} \leftarrow \mathcal{D} \cup bestD$ ▷ append $bestD$ to output $\mathcal{D}$
  **for all** $C_{\mathcal{D}j} \in C_{\mathcal{D}}$ **do**
    $C_{\mathcal{D}j} \leftarrow C_{\mathcal{D}j} + C_{Dj}$ ▷ update keyword counts in $\mathcal{D}$
  **end for**
**end while**
**return** $\mathcal{D}$

[5]Queries used U.S. English, default API settings, and originated from Ann Arbor, MI.

## 3.4 Search results ranking algorithm

As discussed, the algorithm we use has the objective function given by Equation 6. While Raman et al. did a post-hoc study and were able to use historical search log data for capturing subtopic query signals, the same does not apply in a cold-start situation (where we have no input signals from the user besides for the query they issue). We operationalize each of the objective terms as follows:

(1) $q$ - The base search query is the topic the user specifies they want to search for. To help narrow the intent of the query, we prefix the topic with "Introduction to ". We don't do this prefix for subtopic queries.
(2) $q_i$ - A subtopic query for the base topic query $q$. In our implementation, we determine the set of these queries $Q$ using the Wikipedia article corresponding to $q$. From the article, we extract the main subtopic headers and prefix them with $q$ to generate the set of subtopic queries. We omit headers that are not related to the topic (e.g. "see also", "references", "further reading").
(3) $Rel(d_i|q)$ - The relevance of a document $d_i$ for the base query $q$ is given by its reciprocal rank in the commercial Web SERP results for $q$. That is, if $d_i$ is ranked at position $P$, its relevance for $q$ is given as $\frac{1}{P}$.
(4) $Rel(d_i|q_i)$ - Operationalized the same as $Rel(d_i|q)$ where the relevance is given by the reciprocal rank of $d_i$ in the SERP page returned by the commercial search engine in response to subtopic query $q_i$.
(5) $\eta_i = \lambda\left[\cos(\text{snip}(q_i), \text{snip}(q))\right] - (1-\lambda)\max_{j<i}\left[\cos(d_i, d_j)\right]$ where $\cos(a, b)$ is the cosine similarity of $a$ and $b$, $\text{snip}(x)$ is the bag of words representation of the top 10 snippets returned by query $x$ and $\lambda$ is a control parameter in the range [0,1] to let us decide how much importance we want to give to maintaining subtopic query relevance as opposed to maintaining document novelty.

We intend to refine our subtopic extraction methods to generalize beyond those available in Wikipedia topics in future work.

# 4 USER STUDY

To evaluate the learning effectiveness and efficiency of our model, we conducted a crowdsourced user study, focusing on the following research questions:

RQ 1. Does learning-optimized retrieval framework offer higher learning effectiveness or efficiency compared to traditional retrieval results of a baseline commercial Web search engine?
RQ 2. Do personalized search results that account for a user's prior knowledge improve learning effectiveness or efficiency?
RQ 3. How do learning effectiveness and efficiency vary across different topics (information needs) in different domains?

To explore these questions, we ran a crowdsourced user study that was based on a vocabulary learning task: learning the meaning of $K$ target keywords by reading them in context. Participants first completed a multiple-choice pre-test to assess their existing knowledge of the keywords. Then, based on the condition they were assigned, they were instructed to read through a set of documents containing the keywords to be learned. Last, they completed an immediate post-test to assess their updated keyword knowledge.

For information needs, we developed a set of ten topics that were selected from top-level categories of the Open Directory Project

to cover a range of areas, each having distinctive technical/expert vocabulary: Igneous rocks (geology), Tundra (environmental science), Cytoplasm (biology), Bioluminescence (biology), Phrenology (pseudo-science), Pottery (crafts), Cooking (food), Synapse (neuroscience), Refraction (optics) and Phenology (temporal phenomena).

We created one crowdsource job per topic, and for each topic job, each participant was randomly[6] assigned to one of three conditions:

(1) Commercial Web search baseline ('Web'). The participant was simply provided the top Web search results for the topic. Documents were only added until the same stopping criteria in Algorithm 1 was met.
(2) Non-personalized learning-optimized retrieval ($\alpha_N$). The participant was provided a document set retrieved through the objective function described earlier with $\alpha$ parameter set to $\infty$. The $\alpha = \infty$ condition simply means that the keyword density $\epsilon_i$ term becomes the only factor in the ID retrieval objective.
(3) Personalized learning-optimized retrieval ($\alpha_P$). The participant was provided a document set retrieved as defined above but with personalized $S_k$ values calculated based on their prior knowledge $\beta$, computed from their pre-test scores.

Note that for all conditions, we truncated the retrieved document set to a maximum of 10 documents, so as to not overwhelm the participants and to reflect typical search environments where users see 10 documents on the first SERP page.

The pre- and post- vocabulary tests consisted of a series of multiple-choice questions, with one question for each of the $K$ keywords. Both the pre- and post-test used identical questions, where we measured the participants' learning gain by the post- minus pre- scores[7]. To assess a participant's prior knowledge $\beta$, we looked at their pre-test answers for each of the $K$ keywords. If they answered correctly for keyword $k$, that keyword was assigned a $\beta_k$ value of 100 and a value of 0 otherwise. As the $\alpha_N$ condition is not personalized, we assume everyone has the same beginners prior knowledge (all $\beta_k = 0$). For the personalized condition, we used the optimized $\beta$ vector ($K$-dimensional vector of student's prior knowledge of aspects $A_k$) when computing result lists. For non-personalized results, we set all the $\beta$ values to 0. For the personalized condition, we pre-computed all possible sets of document/links that could be generated given different $\beta_k$ values. As each question had two possible states (correct or incorrect), there were $2^K$ possible retrieved document sets. Thi study used $K = 10$, resulting in 1024 possible document sets for the $\alpha_P$ condition. Thus, we were able to provide participants in the $\alpha_P$ condition personalized document sets in real-time.

We used the Crowdflower platform for this study. Participants were offered US$0.04 per page (the equivalent of US$3.20/hr) for completing the tasks. For quality control, in addition to Crowdflower's proprietary mechanisms and 'gold standard' questions, we limited the participant pool to users from the U.S. and Canada, given the vocabulary-centric nature of the task and reliance on English reading skills. We also offered the tasks only to workers in the highest quality (level 3) pool, and only kept responses from those workers who spent at least four minutes on the task. Participants who scored perfectly on the pre-test were also omitted because, as per our model, they would theoretically have no room for improvement.

[6]Participants were sorted into conditions based on Crowdflower's random assignment to tasks.

[7]In measuring 'learning gain', we assume no memory loss so the learning gain is always no less than zero.

| Measure | Web | $\alpha_N$ | $\alpha_P$ | p-value |
|---|---|---|---|---|
| Absolute Learning Gains | 1.721 | 1.831 | **1.982** | p=.046* |
| Learning Gain Per 1000 Words | 0.109 | 0.252 | **0.347** | p<.001! |
| Realized Potential Learning | 0.384 | 0.425 | **0.471** | p=.008† |
| Time Per Word | 12.007 | 29.176 | **35.022** | p<.001! |
| Signif. codes: 0 '!' 0.001 '†' 0.01 '*' 0.05 '·' 0.1 ' ' 1 | | | | |

**Table 2: Aggregated averages of key learning-related measures. Bold values are maximum across conditions. (All tables use same significance codes and bold meaning.)**

The particular set of documents[8] shown to each participant was based on which $\alpha$ condition they were assigned. We gathered data for 40 participants per condition per topic, resulting in a total of 120 participants per topic and 1200 participants overall. After excluding those who didn't pass the initial training questions, those who didn't complete the full task and those who spent less than four minutes on the entire task, we ended up with 863 total participants. Further description of the experiment design, along with screenshots, can be found in an earlier study [20].

## 5 RESULTS

Overall, our analysis showed that different choices of document retrieval algorithms were in fact associated with differences in learning, as measured by both absolute and normalized gains from pre-test to post-test. In the following analysis, all tests for statistical differences across the three conditions were done using the Kruskal-Wallis test[9]. The following analysis will focus on assessing how the averages for the three conditions we tested varied across different measures of learning outcomes. In particular, we use the following measures of learning effectiveness and efficiency.

**Absolute Learning Gains**: The number of keywords where post-test score was higher than pre-test. Measures learning effectiveness.

**Realized Potential Learning**: Absolute learning gains normalized by maximum possible improvement. For example, if two students had an absolute learning gain of 2, but one student had pre-test score of 8/10 and the other had 3/10, the realized potential learning for the first is 2/2 but for the second is 2/7.

**Learning Gains per Word Read**: Absolute learning gains normalized by how many words a participant had to read to achieve their learning gain. Measures learning efficiency. (Table results reported in terms of learning gains per 1000 words.)

**Time Per Word**: Total time spent on reading documents normalized by total words contained in those documents. Measures how much attention users are focusing on what they read.

### 5.1 Absolute Learning Gains

We first analyze the data aggregated across all topics to get an overall view of how each condition performed (Table 2). We see

[8]To avoid multimedia confounds, we only allowed Web documents that contain only text and, at most, supplementary pictures.

[9]While ANOVA analysis and bootstrapped ANOVA also yielded nearly the same results reported, we used this test instead as we found evidence that the data for various measures of learning had strongly non-normal distributions.

| | Learning Gain | | | Learning Gain/Word | | |
|---|---|---|---|---|---|---|
| **Topic** | **Web** | $\alpha_N$ | $\alpha_P$ | **Web** | $\alpha_N$ | $\alpha_P$ |
| Igneous rock | 1.769 | 2.533 | 2.364 | 0.081 | 0.118 | **0.255**$^*$ |
| Tundra | 2.115 | 1.655 | 2.231 | 0.108 | 0.233 | **0.266**$^*$ |
| Cytoplasm | 1.567 | 1.577 | 1.758 | 0.044 | 0.068 | **0.178**† |
| Biolumin. | 1.929 | 1.808 | 1.567 | 0.102 | 0.266 | **0.406**! |
| Phrenology | 1.156 | 1.424 | **2.097**† | 0.065 | 0.155 | **0.360**! |
| Phenology | 1.222 | **2.036**$^*$ | 2.033 | 0.132 | 0.281 | **0.316**! |
| Synapse | 2.071 | 2.233 | 2.267 | 0.036 | 0.084 | **0.142**† |
| Pottery | 2.156 | 1.710 | 1.600 | 0.101 | 0.417 | **0.611**! |
| Cooking | 1.407 | 1.471 | 1.957 | 0.049 | **0.293**! | 0.110 |
| Refraction | 1.824 | 1.957 | 2.107 | 0.135 | 0.207 | **0.230**· |

**Table 3: Absolute learning gains (left) and learning gains normalized per 1000 words (right) averaged across different conditions and topics.**

| **Topic** | **Web** | $\alpha_N$ | $\alpha_P$ | **p-value** |
|---|---|---|---|---|
| Igneous rock | 0.390 | 0.557 | **0.584** | p=.067· |
| Tundra | 0.386 | 0.355 | 0.319 | p=.667 |
| Cytoplasm | 0.283 | 0.350 | 0.390 | p=.367 |
| Bioluminescence | 0.416 | 0.358 | 0.393 | p=.633 |
| Phrenology | 0.325 | 0.402 | **0.617** | p=.003† |
| Phenology | 0.379 | **0.602** | 0.580 | p=.066· |
| Synapse | 0.410 | 0.463 | 0.492 | p=.597 |
| Pottery | 0.480 | 0.454 | 0.454 | p=.936 |
| Cooking | 0.498 | 0.350 | **0.601** | p=.053· |
| Refraction | 0.302 | 0.346 | 0.311 | p=.707 |

**Table 4: Averaged realized potential learning gains across different conditions. Bold values indicate maximums.**

that for all measures, the differences are statistically significant and that the $\alpha_P$ (personalized) condition always performs the best with the $\alpha_N$ (non-personalized) condition always performing second best. At the aggregate level, these findings already show that RQ1 and RQ2 are both supported. However, we need to look deeper to investigate RQ3, the consistency of the findings across topics.

We analyzed overall, or absolute, learning gains (sum of learning gains for all $K$ keywords) across the three conditions. We found that two out of the ten topics ('Phrenology' and 'Phenology') showed significantly different mean learning gains when compared with the three conditions. (Table 3). In both of these two topics, however, the mean learning gain in the $\alpha_P$ condition was almost double that of the commercial search baseline.

## 5.2 Realized Potential Learning Gains

While assessing absolute learning gains estimates how much a participant learned, it does not capture how well each student's *potential* learning gains were realized. A learning gain of 2 with a pre-test score of 8/10 is more significant that the same gain with a pre-test score of 3/10. Therefore, we also look at how much each student learned, relative to how much they *could* have learned. We calculate this by normalizing learning gains by ($K$ - pre-test score). We show these results in Table 4.

We see that one topic has a statistically significant gain ($p < 0.05$) in realized potential learning, with the $\alpha_P$ condition giving almost twice the performance of the commercial Web search baseline. These results lend support to RQ1 and RQ2, i.e. that the $\alpha = \infty$ condition does in fact give some improvement for both actual learning gains and realized potential learning. In general, the $\alpha_P$ condition almost always had a consistently larger (though not always statistically significant) gain across all topics.

## 5.3 Learning Gains per Word Read

While both learning gains and realized potential learning measured the effectiveness of each of the three conditions, we also wanted to assess the *efficiency* of each algorithm. We define learning efficiency as the learning gains per unit effort spent. As per our model, we defined effort to be the total words read - we make the assumption that participants read the full documents they were provided - so we will evaluate learning gains per word read[10]. This measure incorporates effort such that, for two students scoring the same absolute gain, the one who achieved this gain with less effort (reading less text) is rewarded more.

Analysis of the three conditions shows a very strong and consistent trend of the $\alpha_P$ condition offering the best learning gains per word read and with very strong statistical significance (p<.01 for seven out of ten topics) (Table 3). In particular, for each topic, the $\alpha_P$ condition offers an average 3.55 times improvement in learning gains per word read over the Web (commercial search) baseline. The $\alpha_N$ condition offers a 2.63 times improvement. These results suggest that even without personalization, the $\alpha = \infty$ condition offers significantly more efficient learning gains for the participants compared to the Web baseline. These results lend further support to both RQ1 and RQ2 along with RQ3 where normalized learning gains were found to be consistently better in the $\alpha = \infty$ condition compared to the baseline and were almost consistently better in the personalized condition versus non-personalized.

## 5.4 Learning Gains per Unit Time

Considered as a single variable, learning gains per unit time spent reading was associated with significant differences across conditions for only two topics at the (p<.05) level. To understand this further, we split learning gains $\Delta L$ per unit time into two subfactors using the following decomposition:

$$\frac{\Delta L}{Time} = \frac{\Delta L}{Words} \times \frac{Words}{Time} = \frac{\Delta L}{Words} \Big/ \frac{Time}{Words}$$

The relationship of these two subfactors is visualized in Figure 3, with $\frac{Time}{Words}$ on the x-axis and $\frac{\Delta L}{Words}$ on the y-axis. As the plot makes evident, there is a positive correlation (r=.374, p<.001) between these two subfactors. Moreover, while the slope of this approximately linear relationship (which is exactly $\frac{\Delta L}{Time}$, learning per unit time), is relatively stable across conditions – as the initial analysis showed – there are in fact very different tradeoff regimes for user efficiency that lead to similar learning gains per unit time, across the three retrieval conditions. For example, the Web baseline is largely characterized by having the lowest average reading time per word as well as lowest learning gain per word (7/10 topics). In contrast, the personalized $\alpha_P$ condition is characterized by typically having the highest learning gain per 1000 words (8/10 topics).

[10]For simplicity in presenting these findings, we show the means multiplied by 1000 (Learning gains per 1000 words)

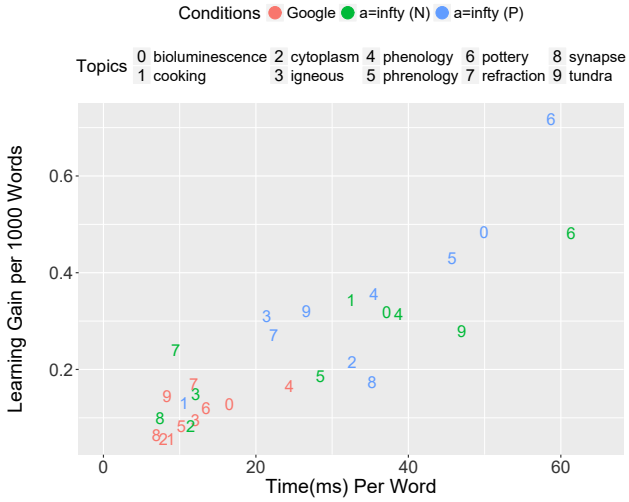


**Figure 3: Learning gains per word generally increases with reading time per word. $\alpha = \infty$ (N) is the non-personalized condition and $\alpha = \infty$ (P) is the personalized condition.**

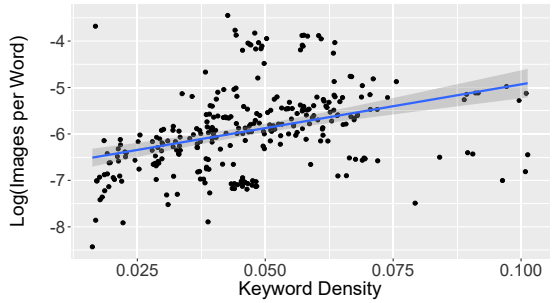


**Figure 4: Image coverage increases with keyword density. Each data point represents a unique document set shown to a study participant.**

## 5.5 Image Coverage

To gain more insight into why pages with increased keyword density might contribute to more efficient and effective learning, we investigated additional properties of the page content that might be correlated with keyword density. We found that while few result documents made use of multimedia such as animations, audio or video, some used images to supplement the text. Thus, the *picture superiority effect* [8], in which people tend to remember things better when they see pictures rather than words, could be relevant. We examined whether there was a relationship between image coverage – defined as total images divided by total words – and keyword density. We determined the number of relevant images manually for each page, excluding irrelevant images such as navigation icons and advertisements. We found that pages with higher keyword density did indeed tend to have increased image coverage. On average, participants saw 1.5, 4.8 and 3.9 images per 1000 words, for the Web, $\alpha_N$, and $\alpha_P$ conditions respectively. Thus, participants in either of the two $\alpha = \infty$ conditions saw almost three times as many images per word as those in the Web commercial search baseline. We observe informally that pages using a higher density of keywords tend to be those that give an overview of topic for instructional purposes, and thus are more likely to be supplemented with images by the author. The keyword density of the document set a participant read did indeed show a linear relationship to the log of the image coverage. Fig. 4 shows a linear correlation between these measures (r=.37, p<.001)[11]. We intend to investigate this phenomenon and other content properties that may interact with learning in future work.

## 5.6 Operational ranking features

The retrieval process described above uses a multi-stage custom pipeline. Here we investigate whether we can use our findings to generate new ranking features for use in a standard retrieval framework, such as a learning-to-rank algorithm, while giving result sets comparable to the full pipeline. Specifically, we look at approximations to the keyword contribution density $\epsilon_i$ (corresponding to document $d_i$). For example, we can replace the original $\epsilon_i$ term with a version that does not require computation of the set of required keyword counts ($s_k$):

$$\epsilon_i^* = \left(\sum_{k=1}^{|d_i|} r_k\right)^{-1} \sum_{j=1}^{K} C_{R_j} \cdot f(C_{\mathcal{D}_j}, C_{R_j})$$

where $C_R$ is the keyword counts in $d_i$ and $f(C_{\mathcal{D}_j}, C_{R_j})$ is a function that decays the contribution of keyword $j$ in the new document as a function of the new document's coverage of $j$ and the coverage in all previous documents. Here, we propose the decay function

$$f(C_{\mathcal{D}_j}, C_{R_j}) = \left(C_{\mathcal{D}_j} + C_{R_j}\right)^{-\gamma}.$$

This offers an approximation to the full objective in Equation 7 by down-weighting the contribution of keywords the user has likely already seen many times and is thus unlikely to benefit more from. The $\gamma$ parameter controls how fast the contribution of a given keyword will decay, which is related to the effort penalty $\lambda$ in the full model[12]. For the personalized condition, the keyword contribution of terms the student already knows is set to zero.

With this approximation, we computed the set overlap (number of elements in the intersection) between equal numbers of documents retrieved using the original $\epsilon_i$ score, and the approximate $\epsilon_i^*$ score, averaging the overlap across topics (and in the personalization case, all personalized retrieval sets). We found that the approximate $\epsilon^*$ approach produced an average set overlap of 71.0% with the non-personalized document sets and a set overlap of 61.8% for the personalized document sets. Setting the decay function to a constant (1) results in a further drop to an average of 45.3% and 43.3% document sets overlap respectively. For comparison, the average set overlap between the Web condition document set and the personalized document set was only 6.4%. This suggests that at least some of the effect of our general approach could be incorporated into standard ranking frameworks using decay-controlled keyword density as an easily-computed feature.

[11] Documents with no images were omitted from the log calculation.
[12] The value $\gamma = 1.5$ yielded the best set overlap with documents produced via Eq. 7 and tested over all topics.

## 6 DISCUSSION

This study represents a first step in a new research direction in which computational models of human cognition are incorporated directly into the objective functions used to optimize retrieval. In a more general view, in our model the search engine becomes a mechanism for effectively choosing content that will update a user's cognitive state from a prior knowledge state, toward a goal state. Beyond supporting more sophisticated forms of learning, this could also suggest personalized mechanisms for addressing some kinds of algorithmic bias in search results, or helping users with reading disabilities. We made a number of simplifying assumptions that future work could seek to enhance and extend.

Our approach is also suited for applications that could be integrated with existing search engines. For example, query intent detection in search engines could be used to find queries focused on learning, and to trigger use of the keyword density parameter as a ranking feature for those queries. Using the heuristic approximation discussed above, this would be straightforward to accomplish and could yield increased learning benefits for users.

Our empirical results show evidence that most of our research questions were successfully answered by the empirical analysis, with RQ1 and RQ2 being supported and RQ3 being partially supported. However, we also note a few limitations to the current study. First, this study only assesses vocabulary learning, which covers only one level of cognitive complexity per Bloom's taxonomy [12]. More complex objectives and document representations may be able to capture the more subtle cues needed for higher-level learning tasks such as drawing correct analogies, or synthesizing multiple points of evidence. Second, we focused on the text content of pages, filtering out any images, video, or other non-text content. While we did provide an analysis of image coverage, the contribution of video in particular would be interesting to study. Third, our reliance on Wikipedia to mine subtopics limits our topics to those with an adequate Wikipedia entry: exploring methods for finding appropriate subtopics for arbitrary queries is a topic for future work.

## 7 CONCLUSION

We introduced a novel retrieval framework for optimizing search results for a learning-oriented objective by composing simple computational models of word learning with an intrinsically diverse ranking objective. We ran a crowdsourced user study to assess human learning gains on a vocabulary learning task using this framework, comparing the effectiveness and efficiency of personalized and non-personalized retrieval with a strong commercial search baseline. Across the topics that we tested, our personalized retrieval algorithm did in fact offer statistically significant overall learning gains per word read, for both absolute and realized potential learning scores. We did not extensively tune the parameters used in our models, so further gains may be possible with more exploration of the model and parameter spaces.

**Acknowledgements.** We thank the anonymous reviewers for their comments. This work was supported in part by the Michigan Institute for Data Science (MIDAS), and by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A140647 to the University of Michigan. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

## REFERENCES

[1] Peter Bailey, Liwei Chen, Scott Grosenick, Li Jiang, Yan Li, Paul Reinholdtsen, Charles Salada, Haidong Wang, and Sandy Wong. 2012. User task understanding: a web search engine perspective. In *NII Shonan Meeting on Whole-Session Evaluation of Interactive Information Retrieval Systems, Kanagawa, Japan.*

[2] Benjamin S Bloom. 1956. *Taxonomy of educational objectives: The classification of educational goals.* New York, Longmans, Green.

[3] Benjamin S Bloom. 1984. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher* 13, 6 (1984), 4–16.

[4] Andrei Broder. 2002. A Taxonomy of Web Search. *SIGIR Forum* 36, 2 (Sept. 2002), 3–10.

[5] Kevyn Collins-Thompson, Paul N. Bennett, Ryen W. White, Sebastian de la Chica, and David Sontag. 2011. Personalizing Web Search Results by Reading Level. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM '11).* ACM, New York, NY, USA, 403–412.

[6] Kevyn Collins-Thompson, Soo Young Rieh, Carl C. Haynes, and Rohail Syed. 2016. Assessing Learning Outcomes in Web Search: A Comparison of Tasks and Query Strategies. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval (CHIIR '16).* ACM, New York, NY, USA, 163–172.

[7] Albert Corbett. 2001. Cognitive computer tutors: Solving the two-sigma problem. In *International Conference on User Modeling.* Springer, 137–147.

[8] Antonella De Angeli, Lynne Coventry, Graham Johnson, and Karen Renaud. 2005. Is a picture really worth a thousand words? Exploring the feasibility of graphical authentication systems. *International Journal of Human-Computer Studies* 63, 1 (2005), 128–152.

[9] Sergio Duarte Torres, Djoerd Hiemstra, and Pavel Serdyukov. 2010. An Analysis of Queries Intended to Search Information for Children. In *Proceedings of the Third Symposium on Information Interaction in Context (IIiX '10).* ACM, New York, NY, USA, 235–244.

[10] Carsten Eickhoff, Jaime Teevan, Ryen White, and Susan Dumais. 2014. Lessons from the Journey: A Query Log Analysis of Within-session Learning. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining (WSDM '14).* ACM, New York, NY, USA, 223–232.

[11] Brian W Junker. 1999. Some statistical models and computational methods that may be useful for cognitively-relevant assessment. *Prepared for the National Research Council Committee on the Foundations of Assessment. Retrieved April* 2 (1999), 81.

[12] David R Krathwohl. 2002. A revision of Bloom's taxonomy: An overview. *Theory into Practice* 41, 4 (2002), 212–218.

[13] Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods* 44, 4 (2012), 978–990.

[14] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[15] Peter Pirolli and Sanjay Kairam. 2013. A knowledge-tracing model of learning from a social tagging system. *User Modeling and User-Adapted Interaction* 23, 2-3 (2013), 139–168.

[16] Filip Radlinski and Susan Dumais. 2006. Improving Personalized Web Search Using Result Diversification. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '06).* ACM, New York, NY, USA, 691–692.

[17] Karthik Raman, Paul N Bennett, and Kevyn Collins-Thompson. 2014. Understanding intrinsic diversity in web search: Improving whole-session relevance. *ACM Transactions on Information Systems (TOIS)* 32, 4 (2014), 20.

[18] Daniel E Rose and Danny Levinson. 2004. Understanding User Goals in Web Search. In *Proceedings of the 13th International Conference on World Wide Web (WWW '04).* ACM, New York, NY, USA, 13–19.

[19] Mark D. Smucker and Charles L.A. Clarke. 2012. Time-based Calibration of Effectiveness Measures. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '12).* ACM, New York, NY, USA, 95–104.

[20] Rohail Syed and Kevyn Collins-Thompson. 2017. Optimizing search results for human learning goals. *Information Retrieval Journal* (2017), 1–18.

[21] Emine Yilmaz, Manisha Verma, Nick Craswell, Filip Radlinski, and Peter Bailey. 2014. Relevance and Effort: An Analysis of Document Utility. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM '14).* ACM, New York, NY, USA, 91–100.

[22] Xiaojin Zhu. 2013. Machine teaching for bayesian learners in the exponential family. In *Advances in Neural Information Processing Systems.* 1905–1913.