Inf Retrieval J (2017) 20:506–523
DOI 10.1007/s10791-017-9303-0

SEARCH AS LEARNING

# Optimizing search results for human learning goals

**Rohail Syed**[1] · **Kevyn Collins-Thompson**[1]

Received: 30 October 2016 / Accepted: 28 April 2017 / Published online: 12 May 2017


**Abstract** While past research has shown that learning outcomes can be influenced by the amount of effort students invest during the learning process, there has been little research into this question for scenarios where people use search engines to learn. In fact, learning-related tasks represent a significant fraction of the time users spend using Web search, so methods for evaluating and optimizing search engines to maximize learning are likely to have broad impact. Thus, we introduce and evaluate a retrieval algorithm designed to maximize educational utility for a vocabulary learning task, in which users learn a set of important keywords for a given topic by reading representative documents on diverse aspects of the topic. Using a crowdsourced pilot study, we compare the learning outcomes of users across four conditions corresponding to rankings that optimize for different levels of keyword density. We find that adding keyword density to the retrieval objective gave significant learning gains on some topics, with higher levels of keyword density generally corresponding to more time spent reading per word, and stronger learning gains per word read. We conclude that our approach to optimizing search ranking for educational utility leads to retrieved document sets that ultimately may result in more efficient learning of important concepts.

**Keywords** Retrieval models and ranking · Intrinsic diversity · Assessment of learning in search

This paper significantly extends an earlier workshop paper that appeared at the Searching as Learning workshop (SAL) at SIGIR 2016 (Syed and Collins-Thompson 2016).

✉ Rohail Syed
rmsyed@umich.edu

Kevyn Collins-Thompson
kevynct@umich.edu

[1] School of Information, University of Michigan, 105 S. State St., Ann Arbor, MI 48109, USA

## 1 Introduction

The Web has become a primary source of online information for learning-related tasks (Bailey et al. 2012). While current Web search engines are tuned to give fast, high-quality results for single queries, they are optimized for generic relevance, not learning outcomes: many tasks involving educational goals require significant time and multiple queries to complete with current Web search engines (Bailey et al. 2012), and ideally, personalized retrieval that can exploit representations of user history and learning goals to be most effective. Developing a search algorithm that is optimized for the learning process is a natural prerequisite to encouraging more Web-based learning.

Exploring new topic areas and learning important domain vocabulary is one popular instance of a learning task (Bailey et al. 2012). Ideally, a retrieval algorithm optimized for this task would not only be effective at teaching a user the important keywords for a given topic by finding highly relevant representative documents, but also enable them to do so efficiently. While user effort itself could be defined in many ways when ranking search results based on factors such as reading difficulty (Collins-Thompson et al. 2011) or other text properties (Yilmaz et al. 2014), we consider the total amount of text to be read in the search result documents as a simple proxy for effort. Given a desired count of exposure for each keyword, by returning documents with higher keyword density per document, we obtain more efficient keyword coverage, reducing effort by reducing the total amount of text that needs to be read. Thus, we explore the role of keyword density as one example of a potentially influential component of educational retrieval. While human learning can be evaluated in terms of different levels of cognitive complexity (Bloom 1956), in this study we specifically focus on a low-complexity form of learning ("Remember" learning) which is assessed based on how well a user can recall or remember basic knowledge of new facts or definitions. The results of our study could also inform future work which may adapt to assess higher forms of cognitive complexity as per Bloom's taxonomy (Bloom 1956).

Toward that goal, the main contributions of this work are a novel search algorithm that re-ranks for optimized educational utility using keyword density as a novel extension of an existing retrieval algorithm and a study that evaluates the effectiveness of this approach on actual learning outcomes.

## 2 Related work

While research on ranking algorithms to maximize the relevance of generic or personalized search results is well-established, few studies have focused on algorithms that can optimize results with utility for an educational goal as the retrieval objective. Researchers have recognized the importance of going beyond traditional retrieval evaluation measures to consider user progress over time (Smucker and Clarke 2012) as well as degree of effort (Yilmaz et al. 2014). Yilmaz et al. (2014) conducted a recent study in investigating the appropriateness of existing relevance measures for assessing the usefulness of a document for users. They show that existing measures are not fully measuring document utility as they don't incorporate an element of effort in defining the true "relevance" of a document. As *effort* itself can be defined in different ways, the authors carefully define effort, or high-effort documents to be those "where people need to work relatively hard to extract relevant information" (Yilmaz et al. 2014). In their work, the authors operationalize this definition with two general measures (document length and readability) containing nine

specific features. Regression analysis shows that a gap between coded relevance judgments and implicit document utility can be explained, with statistical significance, by both readability features such as the LIX index and by document length features such as the total words in the document. In our current work, we test the effect of one of these effort indicators, total words, by altering how much importance our algorithm puts on finding a threshold number of required keywords as quick as possible (fewest total words).

Verma et al. (2016) more recently built on the work by Yilmaz et al. (2014) by directly getting "effort" judgements from crowdworkers rather than only getting relevance judgments as was done earlier (Yilmaz et al. 2014). They further specify their definition of effort as consisting of three components: (1) findability—how easy it is to quickly find what you were looking for in a document, (2) readability—how easy is the vocabulary in the document to understand and (3) understandability—how easy it was to actually learn something from the document. They show that of these factors, findability and relevance both predict user satisfaction with statistical significance, thus bolstering the earlier claim that effort does impact the user's "true" relevance judgment. The authors also found the the document length, measured as total words, was a strong and negative indicator of relevance, possibly suggesting that we should avoid longer documents where possible.

Eickhoff et al. (2014) 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. They also did not develop or assess new retrieval algorithms that could be adapted to improve learning outcomes. Collins-Thompson et al. (2011) incorporated a form of effort criterion into Web search ranking by incorporating reading difficulty as a personalized ranking feature, but did not assess its effectiveness for actual learning outcomes. Similarly, Raman et al. (2013) showed how 'intrinsically diverse' (ID) sessions for exploring and learning about a new, specific topic could be identified and supported using a new diversity-based retrieval algorithm, but without assessing learning outcomes. A subsequent study by Collins-Thompson et al. (2016) examined the effectiveness of ID results presentation on actual high- and low-level learning outcomes. We build on both of these previous studies by exploring a modified variant of the ID algorithm in the context of a vocabulary learning task.

In developing an ITS system for search-based learning, the objective would be to select a set of optimal teaching resources, given a set of test questions and information about the user's current knowledge. This individualized set of teaching resources would then be given to a learner who should be able to perform optimally on the predetermined test. The intuition is that if a learner reads more about a given topic, they are more likely to correctly answer questions about it in the future. 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 select resources that would enable a particular user to perform optimally on that test.

Pirolli and Kairam (2013) demonstrate an ITS model that yields 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 ITS system is a special case of the ideal framework where the resources are exclusively Web documents. Furthermore, their system makes the assumption that the test questions selected accurately test their knowledge of subject and that the knowledge needed to answer these questions can be represented in some finite quantitative form. We will make these same assumptions in our model.

In our study, we relax the automatic approach for the expert model as our focus is more on evaluating the tutor model. In doing so, our expert model will manually decide a subset of documents $\mathcal{D}^*$ that most closely cover a single test topic. Furthermore, unlike the previous study, our tutor model does not require manual intervention and provides the learner resources algorithmically. While the past study (Pirolli and Kairam 2013) was designed to accommodate multiple test topics, we investigate the specific case of optimizing for a single test topic at a time.

Our study is different primarily in that our focus is on investigating how different information retrieval (IR) algorithms can affect a student's learning. This is different from the work by Pirolli and Kairam (2013) where the learners could find and choose the documents themselves and the choice of search algorithm was not controlled. In that study, while learners could use the tutor's recommendations, they were not required to. In our study, we control for this by making the assumption that the various search algorithms represent various tutors. We make the further assumption that the learner subsequently reads through all documents returned by the algorithm until they have reached a set learning goal.

Specifically, we evaluate how our algorithm, using different emphasis on the student's reading effort, performs in terms of improvements in a student's learning gain. For comparison purposes, we will use four levels of effort emphasis which we describe in the next section.

We extend earlier work by Syed and Collins-Thompson (2016) in several key ways: (1) we have added significantly more details, clarifications and figures regarding the study method, design, evaluation and results; (2) we have conducted further analysis on the effects on learning outcome of image coverage and keyword density; (3) we include results regarding how theoretical and actual time spent differ; and (4) we include details about the study participants' survey responses.

## 3 Method

Our retrieval approach has three stages: (1) given a topic expressed as a query, selecting appropriate aspects to be learned for each topic, with each aspect represented by a keyword, (2) for each aspect (keyword), determining the total frequency with which the keyword should occur in the retrieval results, and (3) developing a retrieval algorithm for vocabulary learning that finds documents to 'cover' the selected keywords efficiently by including the keyword density of the documents as an adjustable sub-objective.

As an early work in the area of search as learning, this initial study will only assess the simplest level of learning, in terms of cognitive complexity, which is the 'Remember' level (Anderson et al. 2001). At this level, we only assess how well the student is able to recall basic facts or definitions they have read. As such, the nature of our assessment will be a vocabulary learning test. As our focus is on vocabulary learning, we chose to represent the topic aspects as the top $N$ most representative unigrams. Furthermore, since we are assessing *factual knowledge* rather than *conceptual knowledge* (Anderson et al. 2001), we can justify considering only the frequency of the keywords as an approximation for

learning "knowledge of terminology" which Anderson et al. (2001) characterize as including "knowledge of specific verbal and nonverbal labels and symbols".

### 3.1 Selecting topic aspects

For each topic in our study, we manually collected a set of exemplar Web documents $\mathcal{D}^*$ that were deemed to be representative of useful knowledge about that topic. In this step, we considered the top four documents from a Google search of the topic represented by base query $q$ and added them to our Expert set $\mathcal{D}^*$. We explicitly did not add any Wikipedia articles because such articles may bias the experiment since we used the Wikipedia article corresponding to $q$ as part of our retrieval objective, as we describe in Sect. 3.3. We similarly skipped any documents that primarily contained content in non-textual formats (videos, animations, picture galleries, etc.).

Next, we represented the vocabulary learning goal for a given topic as a weighted set $K = \{k_1, \ldots, k_N\}$ of keywords, which we call the *target keywords*, derived from the topic's exemplar set. For this study, we chose the top $N = 10$ most representative keywords for each topic, using a measure of term frequency in $\mathcal{D}^*$ weighted by inverse term log-frequency in a global corpus. As different aspects of a topic may have greater or less relevance in understanding the topic, each keyword is assigned an associated weight $w_i$, where $w$ are the parameters of a multinomial distribution estimated from the frequency counts of the keywords in the representative set $\mathcal{D}^*$. Table 1 shows the top 5 out of 10 keywords generated for each topic, along with their relative weight $w_i$ (in parentheses).

### 3.2 Determining total words to read

We assume that a student's knowledge of each topic keyword $k_i$ monotonically increases with each instance of it that they read. Now let $T$ be the total number of keywords the learner reads. The distribution of $T$ among the $N$ keywords will be proportional to the importance of each keyword, given by $w_i$. Let $s = \{s_1,\ldots,s_N\}$ be the vector of total instances of each keyword the learner has to read. Then, if $s_i$ is the total instances of $k_i$ the learner has to read, we have: $s_i = T \cdot w_i$.

**Table 1** Top 5 (out of 10) selected keywords per topic, sorted by descending keyword weights $w_i$

| Topic | Keyword 1 | Keyword 2 | Keyword 3 | Keyword 4 | Keyword 5 |
|---|---|---|---|---|---|
| Igneous rock | Rock (.382) | Igneous (.171) | Magma (.102) | Mineral (.070) | Earth (.056) |
| Tundra | Tundra (.374) | Arctic (.094) | Alpine (.087) | Temperature (.083) | Permafrost (.075) |
| DNA | DNA (.385) | Cell (.132) | Base (.084) | Strand (.071) | Acid (.064) |
| Cytoplasm | Cytoplasm (.376) | Cell (.276) | Membrane (.076) | Cellular (.071) | Organelle (.071) |
| GSM | GSM (.246) | Mobile (.181) | System (.122) | Network (.098) | Telecommunication (.092) |

The keywords to be learned range from easy ('rock') to technical ('organelle')

Ideally, a student would learn the most with unlimited instances of each keyword ($T = \infty$). However, in reality a student's time and effort will limit the amount of training they experience, so the $T$ value for each topic was manually chosen to produce small document sets (less than 15 documents).

### 3.3 Developing the retrieval algorithm

As a baseline retrieval algorithm, we used the *intrinsic diversity* algorithm developed by Raman et al. (2013), since it was designed to provide optimal exploration of topics with multiple sub-aspects. The intrinsic diversity objective essentially rewards high quality documents from relevant and representative subtopics, while penalizing redundant documents and subtopics.[1] To account for user effort, we added a new sub-objective term ($e^{\alpha\epsilon_i}$) to the existing intrinsic diversity objective that influences the keyword density (and thus, the efficiency of keyword coverage) for results. So the objective we want to maximize involves four components: (1) the relevance of the $i^{\text{th}}$ document $d_i$ to the base query $q$ (2) the relevance of the same document to the subtopic query $q_i$ that retrieved it (3) the novelty $\eta_i$ offered by $d_i$ relative to other documents already encountered in the set $\mathcal{D}$ and (4) the effective keyword density contained in $d_i$. This gives us the following optimization problem:

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

where the topic we want to teach is given by the base query $q$, $\mathcal{D}$ is the resulting document set, $\eta_i$ is a redundancy penalty, $q_i$ is the $i^{\text{th}}$ sub-topic query and $Rel(d_i|q_i)$ is the reciprocal rank of document $d_i$ in the results page returned for query $q_i$. The redundancy penalty is specifically given as the Maximal Marginal Relevance (Carbonell and Goldstein 1998) trade-off between relevance and novelty as:

$$\eta_i = \lambda[\cos(\text{snip}(q_i), \text{snip}(q))] - (1 - \lambda) \max_{j<i} [\cos(d_i, d_j)]$$

where $\cos(a, b)$ is the cosine similarity of $a$ and $b$ and snip(x) is the bag of words representation of the top 10 snippets returned by query $x$. This algorithm is largely based on work by Raman et al. (2013). The two main exceptions are that (1) our novelty measure $\eta_i$ considers the cosine similarity between documents instead of only SERP snippets and (2) we added the keyword density term $e^{\alpha\epsilon_i}$. Thus, with this extension of the original intrinsic diversity algorithm, setting $\alpha = 0$ largely recovers the original intrinsic diversity algorithm, while higher values of $\alpha$ result in document sets with increasingly dense keyword coverage.

[1] We chose operational parameter settings $\beta = 10, \lambda = 0.2$.

**Algorithm 1** IntrinsicTeacher algorithm that ranks documents for the vocabulary learning task

Initialize with document set $D_i$ as Google search results for subtopic query $q_i$ for all $\mathcal{Q}$.
$\mathcal{D}$ given as output document set.
$Count_{di}$ given as vector of keyword counts in document $d_i \in D_i$.
$Count_S$ given as cumulative keyword counts covered in $\mathcal{D}$.
$Count_T$ given as total keyword counts required for each aspect $a \in A$.
$\mathcal{D} \leftarrow \emptyset$
$Count_{Sj} \leftarrow 0 \ \forall j \in Count_S$
**while** $\exists Count_{Sj} : Count_{Sj} < Count_{Tj}$ **do**
  $maxS \leftarrow 0$
  $maxD \leftarrow \emptyset$
  $Count_D \leftarrow \emptyset$
  **for all** $q_i \in \mathcal{Q}$ **do**
    **for all** $d_i \in D_i, d_i \notin \mathcal{D}$ **do**
      $docS \leftarrow Rel(d_i|q) \cdot Rel(d_i|\, q_i) \cdot e^{\beta \eta_i} \cdot e^{\alpha \epsilon_i}$
      **if** $docS > maxS$ **then**
        $maxS \leftarrow docS$
        $maxD \leftarrow d_i$
        $Count_D \leftarrow Count_{di}$
      **end if**
    **end for**
  **end for**
  $\mathcal{D} \leftarrow \mathcal{D} \cup maxD$
  **for all** $Count_{Sj} \in Count_S$ **do**
    $Count_{Sj} \leftarrow Count_{Sj} + Count_{Di}$
  **end for**
**end while**
**return** $\mathcal{D}$

More specifically, $\epsilon_i$ is the normalized contribution that document $d_i$ offers in terms of how much closer it brings the student towards reading the total required number of keyword instances (the $s$ counts for each of the $N$ keywords). Let $C_{\mathcal{D}} = \{C_{\mathcal{D}1}, C_{\mathcal{D}2}, \ldots C_{\mathcal{D}N}\}$ be the set of keyword counts the student has cumulatively seen so far from documents in set $\mathcal{D}$, let $C_i = \{C_{i1}, C_{i2}, \ldots C_{iN}\}$ be the set of keyword counts in document $d_i$ and $|d_i|$ be the total word count of $d_i$. Then we have:

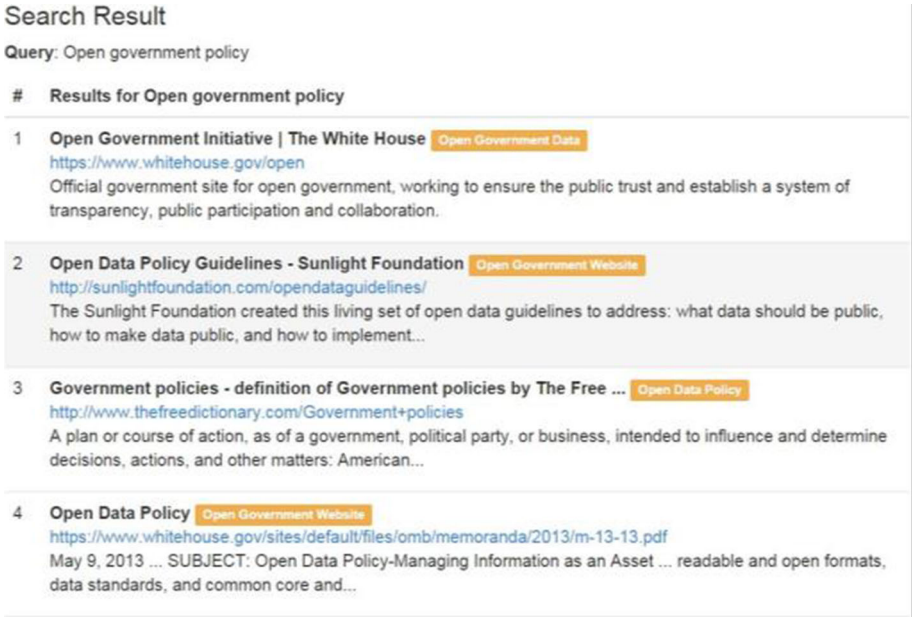
$$\epsilon_i = \frac{1}{|d_i|} \sum_{j=1}^{N} \begin{cases} C_{ij} & C_{ij} + C_{\mathcal{D}j} \leq s_j \\ \max(0, s_j - C_{\mathcal{D}j}) & \text{otherwise} \end{cases} \tag{2}$$

the term $\epsilon_i$ is a measure of the keyword density in $d_i$ with respect to the target keywords for the topic. Unlike a simple keyword density measure, $\epsilon_i$ is a piecewise linear function to avoid giving importance to documents that have high coverage of keywords that prior documents in the $\mathcal{D}$ have already covered the required $s$ times. By rewarding documents that have higher density, via the choice of a higher $\alpha$ setting, we enable the learner to reach the target $s$ counts faster.

Our implementation of the intrinsic diversity algorithm determines the base query's subtopics by analyzing the corresponding Wikipedia article on that query's topic. It generates sub-topic queries by extracting the main header topics in the article and appending them to the base query. For example, for the query "DNA", some sub-topic queries were: "DNA Properties" and "DNA Biological functions". We then fetch the top 70 Google search results for the base query and the top 70 results for each of the sub-topic queries and run

optimization problem (1). We can see why an exhaustive search of the possibility space of candidate documents is very inefficient. If we consider one base query and four subtopic queries, each providing 70 unique documents, we have a total of 350 candidate documents to consider. If we need to fill ten SERP slots based on the maximization criteria, optimization with an exhaustive search would need to test approximately $\binom{350}{10} = 6.676 \times 10^{18}$ unique combinations, which would clearly be intractable. Instead, we use a greedy heuristic shown in Algorithm 1, which, for $Q$ total queries, $N$ total links per query and an output SERP of $R$ links, has a computational complexity of $\mathcal{O}(Q \cdot N \cdot R)$ which is far more efficient. Our algorithm is very similar to that implemented by Raman et al. (2013). Figure 1 shows an example of intrinsically diverse SERP results for the topic "Open government policy" and with $\alpha = 0$ (Collins-Thompson et al. 2016). Our algorithm differs in that we modified it to consider more candidate documents in each iteration and to terminate adding documents to $\mathcal{D}$ once the document set contains documents with a cumulative keyword count greater or equal to $s_i$ for all keywords $k_i$.

We intend to refine our subtopic extraction methods to generalize beyond those available in Wikipedia topics in future work. In general, many different variables can simultaneously influence learning. Some students may learn better with multimedia aids, some will learn better with pure text documents, some will benefit from more technically-worded documents and so on. In this paper, we will specifically evaluate only Web documents that contain only text and, at most, supplementary pictures.


Search Result

**Query**: Open government policy

| # | Results for Open government policy |
|---|---|
| 1 | **Open Government Initiative \| The White House** Open Government Data<br>https://www.whitehouse.gov/open<br>Official government site for open government, working to ensure the public trust and establish a system of transparency, public participation and collaboration. |
| 2 | **Open Data Policy Guidelines - Sunlight Foundation** Open Government Website<br>http://sunlightfoundation.com/opendataguidelines/<br>The Sunlight Foundation created this living set of open data guidelines to address: what data should be public, how to make data public, and how to implement... |
| 3 | **Government policies - definition of Government policies by The Free ...** Open Data Policy<br>http://www.thefreedictionary.com/Government+policies<br>A plan or course of action, as of a government, political party, or business, intended to influence and determine decisions, actions, and other matters: American... |
| 4 | **Open Data Policy** Open Government Website<br>https://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf<br>May 9, 2013 ... SUBJECT: Open Data Policy-Managing Information as an Asset ... readable and open formats, data standards, and common core and... |

**Fig. 1** Example of intrinsically diverse search results using the algorithm from Collins-Thompson et al. (2016)

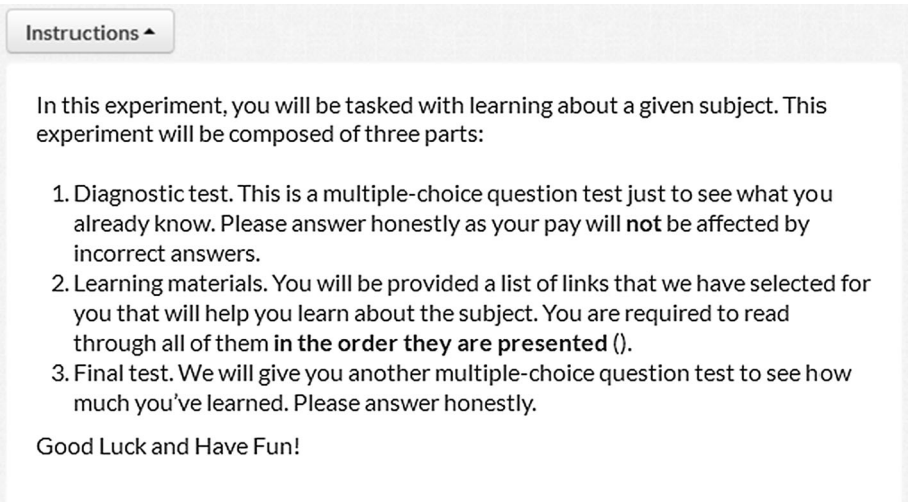



**Fig. 2** User study instructions (Always visible throughout the experiment)

## 4 Evaluation

To assess the potential effect on learning outcomes of retrieved documents optimized using different levels of keyword density (choices of $\alpha$), we ran a crowdsourced user study that involved a vocabulary learning task: learning the target keywords. The task consisted of three stages: (1) participants first completed a multiple-choice pre-test to assess their existing knowledge of the keywords; (2) then, based on the condition, they read through a provided retrieval set of documents containing the keywords to be learned; (3) finally, they completed an immediate post-test to assess their updated keyword knowledge (Figs. 2, 3). Participants had to complete these stages in the ordered sequence and after progressing to the next stage, could not return to a previous stage. In the reading stage, participants had to click on and read all the links they were provided. There was no time limit explicitly provided to the participants but we manually excluded any who spent less than four minutes on the entire task as they likely did not take the task seriously.

We ran five separate crowdsourced jobs corresponding to five different topics selected to cover a range of scientific topics: Igneous rocks (geology), Tundra (environmental science), DNA (genetics), Cytoplasm (biology) or GSM (telecommunications). For each of these topic jobs, a participant was randomly[2] assigned to one of four different keyword density conditions, corresponding to $\alpha$ settings of $[0, 80, 120, \infty]$. We chose the specific values of 80 and 120 based on empirical analysis of the average maximum variation in the document sets produced by different levels of $\alpha$ across multiples of 40 when compared with the $\alpha = 0$ condition. The $\alpha = \infty$ condition simply means that we give full weight only to the keyword density $\epsilon_i$ term and ignore all other terms in the ID retrieval objective.

The pre- and post-vocabulary tests consisted of a series of multiple-choice questions, one for each of the $K$ keywords. Both the pre- and post-reading tests were constructed with identical questions so that we could investigate the participants' learning gain for each

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

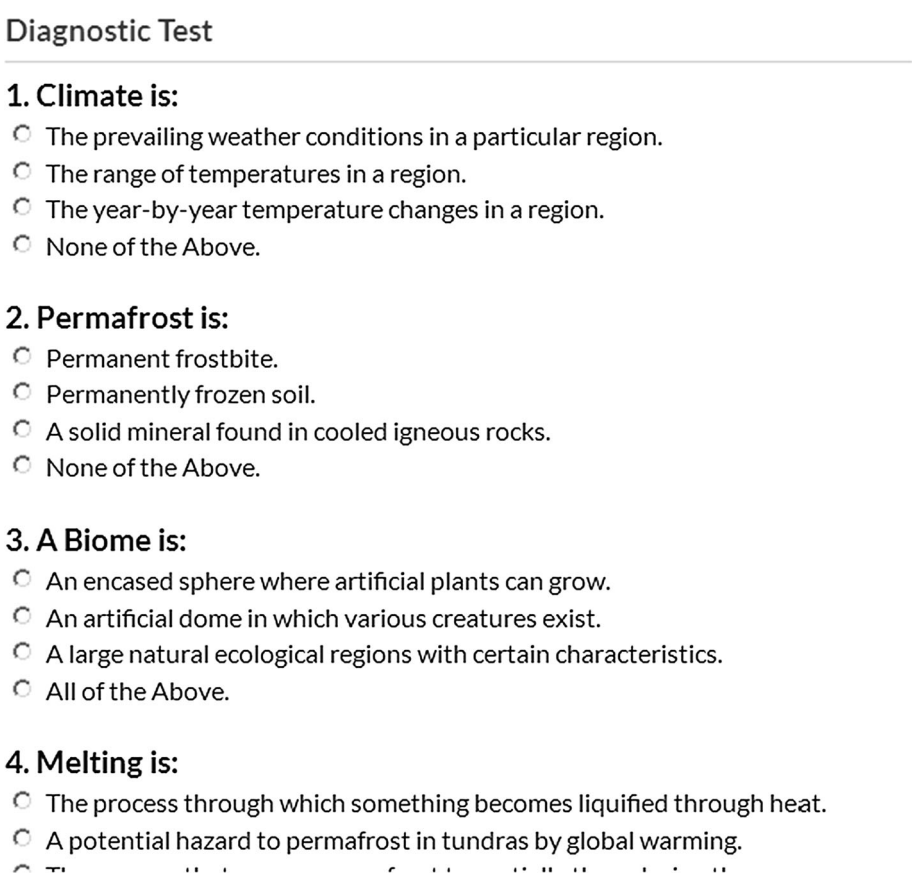



**Fig. 3** User study first part: the pre-reading vocabulary test

vocabulary term by looking at the difference in scores.[3] If the participant incorrectly answered the definition of a term in the pre-reading test but got it right in the post-reading test, they would score a learning outcome of 1 for that keyword. Otherwise, they score 0 for that keyword. We then aggregate their total learning outcomes as a measure of their overall learning gains.

We used the Crowdflower platform for this study. Participants were offered US$0.04 per page (the equivalent of US$3.20/h) 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.

The particular set of documents shown to each participant was based on which $\alpha$ condition they were assigned. We gathered data for 35 participants per $\alpha$ condition per topic, resulting in a total of 140 participants per topic and 700 participants overall. After excluding those who didn't pass the test questions, those who didn't complete the full task

[3] In measuring 'learning gain', we assume no memory loss so the learning gain is always non-negative.

and those who didn't spend at least four minutes on the task, we ended up with 447 total participants.

## 5 Results

Overall, our analysis showed that different choices of $\alpha$ were in fact associated with differences in learning, as measured by both absolute and normalized gains from pre-test to post-test.

We first analyzed learning gains (sum of learning gains for all $K$ keywords) across the four $\alpha$ conditions. Retrieval results incorporating higher keyword density gave statistically significant higher mean learning gains for two out of the five topics[4] (Table 2). As the remaining topics didn't show statistically significant differences across conditions, we consider the two topics that did show strong differences. Both of these topics showed a peak learning gain at the $\alpha = 80$ condition, suggesting that a combination of lowering effort via the keyword density parameter and rewarding intrinsic diversity in documents offers better learning gains than either factor alone. However, we also found that the setting of $\alpha = 120$ yielded the worst learning gains in those same topics. This suggests that the learning gains are quite sensitive to the particular choice of $\alpha$ and that choosing an $\alpha$ that combines both the ID objective and the keyword density objective is not always going to improve learning utility. It's not entirely clear why the specific value of $\alpha = 80$ offered better performance but we intend to investigate this further and how to algorithmically choose $\alpha$ in future work, using an extended set of topics.

Since the target keywords ranged from more familiar to more technical, and learning gains could be expected to interact with keyword difficulty, we faceted the learning gain results by low- and high-difficulty keyword categories.[5] Figure 4 shows the result of averaging the learning gains for each keyword in the two difficulty categories and then averaging the results across the five topics. We see that there were learning gains in all conditions for both low- and high-difficulty keywords, but as expected, learning gains were higher for the higher-difficulty (and thus initially less familiar) keywords (one-way ANOVA differences in means between high and low difficulty words was statistically significant at the $p < .05$ level—tested for all four conditions).

### 5.1 Learning gains per word read

Next, as a measure of learning efficiency, we evaluated absolute learning gain normalized by the total words read. We assume that participants are reading the entire documents that we provide them so we consider the total words read to be the document length. 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. ANOVA analysis of the different $\alpha$ levels shows that most topics had strongly significant differences in means. There was a general trend of increasing gains with increasing $\alpha$ and several topics achieved maximum gains at $\alpha = \infty$ (Table 3).

[4] For all ANOVA analysis reported, the same significance ranges were found using bootstrapped ANOVA (2000 iterations).

[5] Keywords were split into two groups of five keywords according to their age of acquisition (AoA) score in a standard psychometric database. If a keyword didn't have an AoA score, it was assumed to be maximum difficulty.

**Table 2** ANOVA analysis for learning gains across different $\alpha$ conditions

| Topic | $\alpha = 0$ | $\alpha = 80$ | $\alpha = 120$ | $\alpha = \infty$ | $p$ value |
|---|---|---|---|---|---|
| Igneous rock | **1.55** | 1.20 | 1.38 | **1.55** | $p = .727$ |
| Tundra | 1.44 | **1.852** | 1.815 | 1.37 | $p = .473$ |
| DNA | 1.71 | 1.55 | **1.76** | 1.57 | $p = .938$ |
| Cytoplasm | 1.86 | **2.90** | 1.45 | 1.58 | $p = .012$* |
| GSM | 1.60 | **2.50** | 1.45 | 2.33 | $p = .064$ |
| Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 | | | | | |

Bold values are maximum across conditions

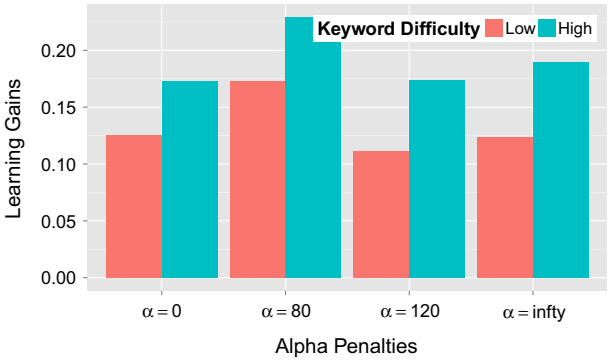



**Fig. 4** Learning gains were greater for keywords in the 'higher difficulty' category

**Table 3** ANOVA analysis for learning gains per 1000 words

| Topic | $\alpha = 0$ | $\alpha = 80$ | $\alpha = 120$ | $\alpha = \infty$ | $p$ value |
|---|---|---|---|---|---|
| Igneous rock | 0.176 | 0.116 | 0.174 | **0.316** | $p = .001$** |
| Tundra | 0.093 | 0.203 | 0.138 | **0.210** | $p = .007$** |
| DNA | 0.234 | 0.203 | 0.206 | **0.276** | $p = .546$ |
| Cytoplasm | 0.558 | **0.811** | 0.361 | 0.451 | $p = .006$** |
| GSM | 0.167 | 0.315 | 0.249 | **0.614** | $p < .001$*** |
| Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 | | | | | |

Bold values are maximum across conditions

We note that one topic, Cytoplasm, showed an opposite trend where higher alpha values mostly lead to worse normalized learning gains. We hypothesize that this may be because the total number of words used in each condition for Cytoplasm were significantly lower (almost half as many for $\alpha = 0$ and $\alpha = 80$) compared to the four other topics. It is thus possible that the positive impact of choosing higher $\alpha$ values is only effective after passing a certain threshold of minimum reading material.

## 5.2 Learning gains per unit time

When considering learning gains per unit time (Table 4) instead of per word, the results were much less conclusive: only one topic showed significant differences in mean learning per time and in that particular topic, the $\alpha = \infty$ condition showed more than three times as much learning per time spent as compared to any other condition. However, because these

**Table 4** ANOVA analysis for learning gains per time spent (s)

| Topic | $\alpha = 0$ | $\alpha = 80$ | $\alpha = 120$ | $\alpha = \infty$ | $p$ value |
|---|---|---|---|---|---|
| Igneous rock | **0.034** | 0.013 | 0.019 | 0.019 | $p = .400$ |
| Tundra | 0.024 | **0.027** | **0.027** | 0.020 | $p = .940$ |
| DNA | **0.120** | 0.051 | 0.080 | 0.022 | $p = .2565$ |
| Cytoplasm | 0.125 | **0.167** | 0.051 | 0.053 | $p = .195$ |
| GSM | 0.099 | 0.033 | 0.030 | **0.373** | $p = .011$* |
| Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 | | | | | |

Bold values are maximum across conditions

results were only significant for one topic, we can't generalize this finding to all topics. To better understand the factors affecting learning gain per unit time (denoted $\frac{\Delta L}{Time}$), consider the following decomposition:

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

This relationship is visualized in Fig. 5, 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= .49, p= .03) between these two components. However, while the slope of this approximately linear relationship (which is exactly $\frac{\Delta L}{Time}$, learning per unit time), is relatively stable across conditions, there are very different tradeoff regimes depending on the value of $\alpha$: the $\alpha = 0$ condition is characterized by some of the lowest reading times per word (second lowest for 4/5 topics) and learning gains per word (lowest for 3/5 topics), while the $\alpha = \infty$ condition is characterized by the highest times (first or second highest for 3/5 topics) and learning gains per word (highest for 3/5 topics). Thus, while the overall learning gain per unit time (ratio of the two components) may not change dramatically across conditions, the underlying two components, representing the tradeoff users choose between reading time and learning efficiency, vary greatly as keyword density changes greatly.

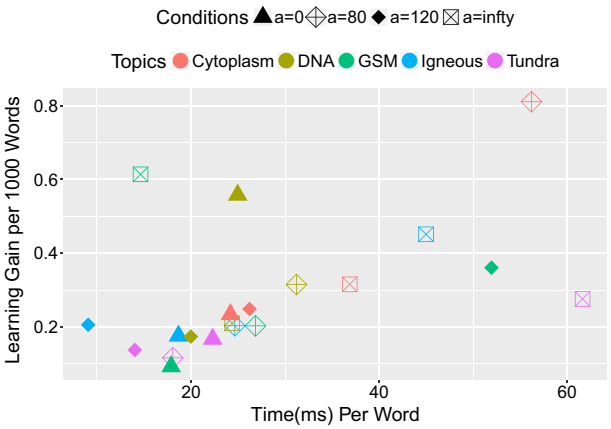



**Fig. 5** Learning gains per word generally increases with reading time per word

### 5.3 Theoretical time versus actual time spent

We also wanted to determine whether or not participants actually spent the amount of time a theoretical model would assume. In particular, we consider a simple linear model where the time spent reading a document was proportional to the length of the document in words. Borrowing the fixed value of time spent per word = 0.018 s from Smucker and Clarke (2012), we tested how well this could estimate the true time participants spent per condition. We found an opposite trend in general where participants seemed to spend more time than predicted as $\alpha$ values increased (Fig. 6). The difference between predicted and actual time spent is largest at $\alpha = \infty$ which is the condition with the least number of words to read on average. So while the theoretical model would predict proportionately less time spent, in reality, participants were willing to spend significantly more time. This might suggest that it is not the length of the document itself that predicts the duration patterns but rather, it could also be the perceived and actual ease of reading shorter documents.

Also note that the average actual time spent was nearly the same in the $\alpha = 0$ and $\alpha = \infty$ conditions despite the fact that the theoretical time spent should have been almost twice as high in the $\alpha = 0$ condition. This further suggests that users may, on average be willing to spend the same amount of time on document sets containing very different total number of words. This discrepancy may be explained by the possibility that shorter document sets simply require less overall effort in reading, where effort is defined by the total words read. We also consider the possibility that non-textual elements of the web page documents may have influenced the user's behavior. For example, the presence of accompanying images in the texts may have encouraged users to remain engaged with the content and spend more time. As we will show in the next section, the $\alpha = \infty$ condition had documents with the most images used per word of text.

### 5.4 Image coverage versus keyword density

To gain more insight into why pages with increased keyword density might contribute to more efficient 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, a number did use images to supplement the text. Thus, the *picture superiority effect* (De Angeli et al. 2005), in which people tend to remember things better when they see pictures rather than words, could be relevant, since

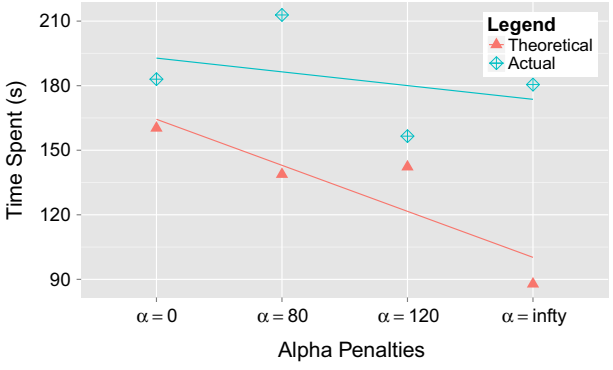



**Fig. 6** Comparison of theoretical versus actual time spent. Actual time tends to be more than what we would predict

we were testing fact-based learning, which relies at least partially on recall. We thus examined whether there was a relationship between image coverage—defined as total images divided by total words—as a function of $\alpha$. 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, as shown in Fig. 7. For three of the five topics, the highest image coverage is in the $\alpha = \infty$ condition.

We consider the possibility that a heavier coverage of images in teaching documents can improve learning outcomes regardless of condition. There is partial evidence of this in that ANOVA analysis of the topics "Igneous rock", "Tundra" and "DNA" showed no statistical significance in means (Table 2) and these three topics had the top three average image coverage (.0024, .0026 and .0034 respectively). On the other hand, the two topics that showed significant differences ("Cytoplasm" and "GSM") had the lowest coverage (.0015 and .0006 respectively). As such, it is possible that a higher image coverage can collectively improve or worsen learning gains regardless of conditions. Determining if the presence or absence of images actually has such an effect warrants further investigation.

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. We intend to investigate this phenomenon and other content properties that may interact with learning in future work.

Because each condition lacked any variation in keyword density or image coverage (each condition produced only one distinct set of documents), we could not determine with this information alone if keyword density or image coverage was responsible for the learning gains improvement. However, we did conduct a follow-up study (Syed and Collins-Thompson 2017) using the same framework but with some altered parameters where we tested three conditions, one of which was the $\alpha = \infty$ condition personalized relative to the participant's pre-reading scores (this simply means that the required $s$ counts were modified to reflect what the participant already knew). This allowed for many data points of different keyword densities, image coverages and learning gains. We aggregated all participants in the personalized condition and created a two-by-two split of learning gains by median image coverage (lower (n= 141) and higher (n= 142) than median) and median keyword density [lower (n= 141) and higher (n= 142)] of the assigned document sets. We then conducted a two-way ANOVA with learning gains as the dependent variable to test for interactions between keyword density and image coverage. We found that there were no significant interactions ($p = .36$) and that image coverage did not yield significant differences in learning gain ($p = .84$). However, we did find that keyword density did yield significant differences ($p = .01$), suggesting that it was in fact changes in keyword density that yielded the learning gain improvements.

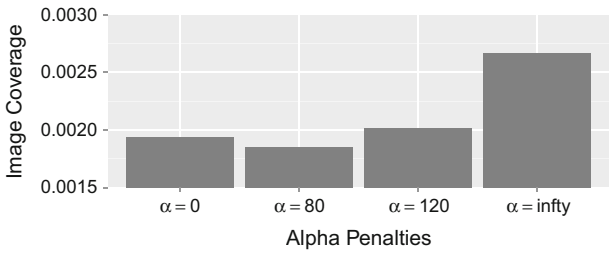



**Fig. 7** Higher $\alpha$ penalty generally results in documents with higher image coverage

We also note that both image coverage and keyword density are measures that are normalized by total words in the document set. By removing this normalization, we repeated the above analysis with total images seen versus total keywords seen. We found that the interaction was still insignificant ($p = .35$) but that total keywords was now insignificant as well ($p = .27$) whereas total images was strongly significant ($p < .001$). This suggests that if we don't factor in the effort the participant has to spend in learning, simply looking at the total keywords they have read won't have any predictable effect on learning outcomes. However, this also shows that regardless of how much a user has to read, the more images they get to see, the better their learning outcomes will be. It might be worth noting that in the follow-up study—from where we're getting this data—the keyword density term additionally penalized documents that had higher vocabulary difficulty levels.

## 6 Discussion

We note some of the limitations in this study and emphasize that this is an early work in the area of specifically optimizing search engines for human learning purposes. Firstly, we note that the objective in the implementation proposed in this study was focused on the simplest level of cognitive complexity ('Remember' learning). This required the participants to only know the definitions of a set of relevant keywords, but did not test nor was it optimized for, higher levels of complexity such as understanding or application of the relevant subject terms. For these more complex methods of learning, we would have to modify the objective function that we are maximizing along with possibly modifying the $\epsilon$ parameter to be defined by something besides keyword density. Furthermore, even at the 'Remember' level of learning, we made the assumption that the more times a participant sees the word they need to learn, the more likely it is that they can triangulate the meaning of the word. It may be possible, however, that factors such as context or surrounding words might influence linguistic learning. Furthermore, factors like vocabulary difficulty in the documents was not controlled for in this study but could certainly be incorporated in the objective function in later studies. Developing more robust models of learning is a compelling direction for future work.

A second limitation is in the possibility of non-textual components of the website interfering with the learning process. In particular, we showed earlier how conditions that had better image coverage tended also matched the $\alpha$ conditions that had better learning per words read. However, as we determined from a later study using an extended set of topics, the effect of higher or lower image coverage on learning gains was not statistically significant whereas the effect of keyword density was, suggesting that it was in fact the keyword density that was likely affecting the learning gains and not the image coverage. We choose to provide participants with the full documents as-is to emulate the natural learning experience for the end-user. In a future study, we may consider stripping out the textual content from the documents and only providing them with plain text to remove any confounds of interference but at the expense of creating an artificial UI design.

As this study is one of the first to our knowledge to use a crowdsourcing platform for a complicated educational teaching task, one of our concerns was about user satisfaction and whether or not users would respond favorably to the task conditions. The Crowdflower platform gives each worker an optional satisfaction survey at the end of the task. Collecting all the results, we found a roughly 20% completion rate (145/700 responses) with the

following aggregated averages: (1) overall satisfaction was 4.26/5.00 (2) ease of the task was 3.72/5.00 (3) payment satisfaction was 4.00/5.00 and the clarity of instructions was 4.40/5.00. These results suggest that the crowdworkers were generally satisfied with the overall task. It is understandable that the ease of task score was not very high since this task was inherently an unusually complex task compared to the micro-tasks that crowd platforms are typically used for. While this survey was designed and administered by Crowdflower themselves, future evaluations of our framework might involve our own user satisfaction surveys to see how self-reported satisfaction and perceived learning correlate to keyword density and actual learning gains.

We also considered whether or not the document sets being retrieved in different $\alpha$ conditions were actually different from each other. We found that, in terms of Google's own SERP rankings, documents in different conditions did, on average, have significantly different average SERP rankings. The $\alpha = 0$ case resulted in retrieved documents that were closest to Google's top SERP ranks (rank 2.86 on average across topics). This was to be expected since the $\alpha = 0$ condition puts more emphasis on Google's relevance criteria. On the other hand, the remaining three $\alpha$ levels gathered documents that were further away from Google's idea of what is relevant (average ranks of 4.46, 7.65 and 13.81 for $\alpha = 80$, $\alpha = 120$ and $\alpha = \infty$ respectively). As such, each condition is fetching different documents and the $\alpha = \infty$ is finding documents that were very likely not being considered when keyword density was disabled ($\alpha = 0$).

## 7 Conclusion

We introduced a novel algorithm for optimizing Web search results for a learning-oriented objective—a vocabulary learning task—by extending intrinsically diverse ranking to incorporate a keyword density sub-objective. This keyword density was controlled by a parameter $\alpha$ that rewarded documents containing a high density of topic-relevant keywords. The result was an algorithm that not only gave relevant, diverse results to explore new topics, but also emphasized efficient keyword coverage in the results content, thus allowing learners to potentially expend less effort toward their learning goal. We hypothesized that changing the keyword density $\alpha$ would be associated with positive changes in users' vocabulary learning outcomes. We tested this hypothesis with a crowdsourced pilot study based on five topics, across four conditions that varied keyword density by using different values of $\alpha$. We found that for some topics participants did in fact show stronger learning gains per word with non-zero $\alpha$ settings. Of the four topics that showed significant differences of means, three were maximized at $\alpha = \infty$. This is an interesting finding as the $\alpha = \infty$ condition *only* considers the keyword density as its objective which means that our findings suggest that a search algorithm that is blind to the rank or implicit quality of a document is offering better results than an algorithm that explicitly considers such measures. We also examined learning gains per word and per unit time, finding that users showed very different tradeoffs between reading time per word and learning gains per word in low- versus high keyword density conditions. In future work we intend to explore criteria for selecting optimal operational settings of $\alpha$, and to incorporate more personalized components in the retrieval model.

**Acknowledgements** This research was supported in part 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

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., et al. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives complete edition*. New York: Longman.

Bailey, P., Chen, L., Grosenick, S., Jiang, L., Li, Y., Reinholdtsen, P., et al. (2012). User task understanding: A web search engine perspective. In *NII Shonan Meeting on Whole-Session Evaluation of Interactive Information Retrieval Systems*, Kanagawa, Japan.

Bloom, B. S. (1956). *Taxonomy of educational objectives: The classification of educational goals*. New York: Longmans, Green.

Carbonell, J., & Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 335–336). ACM.

Collins-Thompson, K., Bennett, P. N., White, R. W., de la Chica, S., & Sontag, D. (2011). Personalizing web search results by reading level. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management, CIKM '11* (pp. 403–412). New York, NY: ACM.

Collins-Thompson, K., Rieh, S. Y., Haynes, C. C., & Syed, R. (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* (pp. 163–172). New York, NY: ACM.

De Angeli, A., Coventry, L., Johnson, G., & Renaud, K. (2005). Is a picture really worth a thousand words? Exploring the feasibility of graphical authentication systems. *International Journal of Human-Computer Studies*, *63*(1), 128–152.

Eickhoff, C., Teevan, J., White, R., & Dumais, S. (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* (pp. 223–232). ACM.

Pirolli, P., & Kairam, S. (2013). A knowledge-tracing model of learning from a social tagging system. *User Modeling and User-Adapted Interaction*, *23*(2–3), 139–168.

Raman, K., Bennett, P. N., & Collins-Thompson, K. (2013). Toward whole-session relevance: Exploring intrinsic diversity in web search. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '13* (pp. 463–472). New York, NY: ACM.

Smucker, M. D., & Clarke, C. L. (2012). Time-based calibration of effectiveness measures. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 95–104). ACM.

Syed, R., & Collins-Thompson, K. (2016). Optimizing search results for educational goals: Incorporating keyword density as a retrieval objective. In *Second International Workshop on Search as Learning (SaL 2016)*. ACM. http://ceur-ws.org/Vol-1647/SAL2016_paper_21.pdf.

Syed, R., & Collins-Thompson, K. (2017). Retrieval algorithms optimized for human learning. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '17*. New York, NY: ACM (to appear).

Verma, M., Yilmaz, E., & Craswell, N. (2016). On obtaining effort based judgements for information retrieval. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining* (pp. 277–286). ACM.

Yilmaz, E., Verma, M., Craswell, N., Radlinski, F., & Bailey, P. (2014). Relevance and effort: An analysis of document utility. In *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management* (pp. 91–100). ACM.