Optimizing Search Results for Educational Goals: Incorporating Keyword Density as a Retrieval Objective

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

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

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.

1. INTRODUCTION

The Web has become a primary source of online information for learning-related tasks [1]. 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 [1], 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 [1]. 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 [2] or other text properties [8], 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, thus reducing effort by reducing the total amount of text that needs to be read. Thus, we explore the role of keyword density as a component of educational retrieval.

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 proxy for effort, 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 [7] as well as degree of effort [8], but little, if any, of that work has involved learning assessment. Eickhoff et al. [5] 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. [2] 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. [6] 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. [3] 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

Search as Learning (SAL), July 21, 2016, Pisa, Italy

context of a vocabulary learning task.

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.

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. We then 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 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 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. Then, if s_i is the total instances of k_i the learner reads, we have: $s_i = T \cdot w_i$.

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. [6], 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:

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

[1]We chose operational parameter settings $\beta = 10, \lambda = 0.2$ where the topic we want to teach is given by the base query q, $\mathcal{D}$ is the resulting document set, $Div(q_i, \mathcal{Q})$ is a redundancy penalty, q_i is the i^{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.

With this extension, setting $\alpha = 0$ recovers the original intrinsic diversity results, while higher values of α result in document sets with increasingly dense keyword coverage.

More specifically, ϵ_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_j counts). Let $C_{\mathcal{D}}$ represent the cumulative keyword counts the student has seen so far from documents higher in the ranking, and C_i represents the keyword frequency distribution of d_i. Then we have:

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

The term ϵ_i effectively is a measure of the keyword density in d_i with respect to the target keywords for the topic. By rewarding documents that have higher density, via the choice of a higher α setting, we enable the learner to reach the target s_j counts faster.

Our implementation of the intrinsic diversity algorithm determines the base query's sub-topics 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-topics queries and run optimization problem (1).

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.

4. EVALUATION

To assess the potential effect on learning outcomes of retrieved documents optimized using different levels of keyword density (choices of α), we ran a crowdsourced user study that involved a vocabulary learning task: learning the target keywords. Participants first completed a multiple-choice pre-test to assess their existing knowledge of the keywords, then based on the condition, read through a provided retrieval set of documents containing the keywords to be learned, and then completed an immediate post-test to assess their updated keyword knowledge. 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,

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

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)

Table 1: Top 5 (out of 10) selected keywords per topic, sorted by descending keyword weights w_i. The keywords to be learned range from easy ('rock') to technical ('organelle').

Topic	α=0	α=80	α=120	α=∞	p-value
Igneous rock	1.312	1.094	1.333	**1.529**	p=.562
Tundra	1.406	**1.829**	1.800	1.514	p=.346
DNA	1.481	**1.576**	1.438	1.483	p=.977
Cytoplasm	1.719	**3.067**	1.286	1.333	p<.001***
GSM	1.654	**2.478**	1.258	1.967	p=.0126*
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

Table 2: ANOVA analysis for learning gains across different α conditions. Bold values are maximum across conditions.

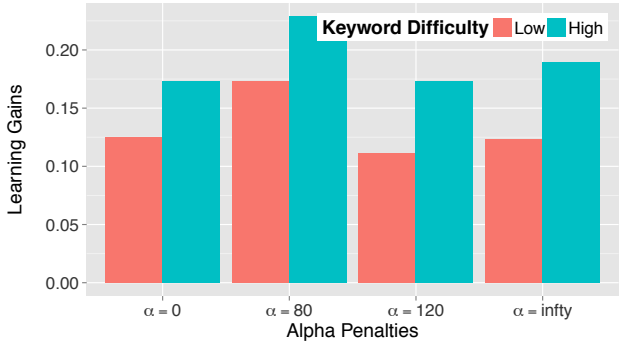

Figure 1: Learning gains were greater for keywords in the 'higher difficulty' category.

corresponding to α settings of $[0, 80, 120, \infty]$. The $\alpha = \infty$ condition simply means that we give full weight only to the keyword density ϵ_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 vocabulary term by looking at the difference in scores[3].

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.

The particular set of documents shown to each participant was based on which α condition they were assigned. We gathered data for 35 participants per α 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 and those who didn't complete the full task, we ended up with 616 total participants.

5. RESULTS

Overall, our analysis showed that different choices of α 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 α conditions. Retrieval results incorporating higher keyword density gave statistically significant mean learning gains for two out of the five topics[4] (Table 2). 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 α and that choosing an α 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 α 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 1 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.

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

[4]For all ANOVA analysis reported, the same significance ranges were found using bootstrapped ANOVA and the Kruskal-Wallis test.

[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.

Topic	α=0	α=80	α=120	α=∞	p-value
Igneous rock	0.149	0.106	0.168	**0.312**	p<.001***
Tundra	0.091	0.201	0.137	**0.232**	p<.001***
DNA	0.203	0.207	0.168	**0.261**	p=.258
Cytoplasm	0.516	**0.857**	0.320	0.381	p<.001***
GSM	0.173	0.312	0.216	**0.517**	p<.001***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

Table 3: ANOVA analysis for learning gains per 1000 words. Bold values are maximum across conditions.

Topic	α=0	α=80	α=120	α=∞	p-value
Igneous rock	**0.044**	0.017	0.019	0.018	p=.048*
Tundra	0.038	0.033	**0.042**	0.029	p=.816
DNA	**0.147**	0.081	0.087	0.028	p=.068.
Cytoplasm	**0.231**	0.202	0.086	0.095	p=.111
GSM	0.112	0.074	0.046	**0.279**	p=.008**
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

Table 4: ANOVA analysis for learning gains per time spent (s). Bold values are maximum across conditions.

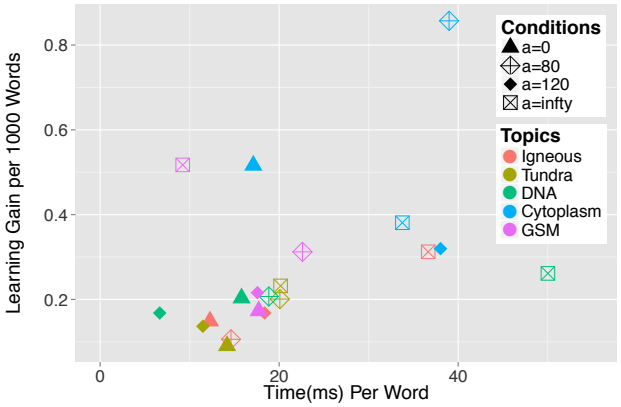

Figure 2: Learning gains per word generally increases with reading time per word.

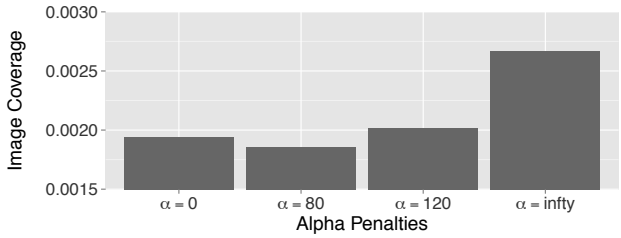

Figure 3: Higher α penalty generally results in documents with higher image coverage.

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. 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 α levels shows that most topics had strongly significant differences in means. There was a general trend of increasing gains with increasing α and several topics achieved maximum gains at $\alpha = \infty$ (Table 3).

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 high α values is only effective after passing a certain threshold 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: for example, two topics showed significant differences in mean learning per time, but with opposite extremes of α values (0 and ∞). To better understand the factors affecting learning gain per unit time (denoted $\frac{\Delta L}{Time}$), consider the following decomposition:

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

This relationship is visualized in Figure 2, 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=.42, p=.06) 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 α: the $\alpha = 0$ condition is characterized by some of the shortest reading times per word and lowest learning gains per word, while the $\alpha = \infty$ condition is characterized by the highest times and learning gains. 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.

5.3 Image Coverage vs. 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* [4], in which people tend to remember things better when they see pictures rather than words, could be relevant, since 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 α. 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. 3. 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 im-

ages 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.

6. 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 α 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 α 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 α. We found that for some topics participants did in fact show stronger learning gains per word with non-zero α 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- vs high keyword density conditions. In future work we intend to explore criteria for selecting optimal operational settings of α, 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

[1] Bailey P., Chen L., Grosenick S., Jiang L., Li Y., Reinholdtsen P., Salada C., Wang H., and Wong S. 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] Collins-Thompson K., Bennett P. N., White R. W., Chica S., de la, and 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)*. ACM, New York, NY, USA, 403–412. DOI:http://dx.doi.org/10.1145/2063576.2063639

[3] Collins-Thompson K., Rieh S. Y., Haynes C. C., and 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)*. ACM, New York, NY, USA, 163–172. DOI:http://dx.doi.org/10.1145/2854946.2854972

[4] De Angeli A., Coventry L., Johnson G., and 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 (2005), 128–152.

[5] Eickhoff C., Teevan J., White R., and 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 (WSDM '14)*. ACM, New York, NY, USA, 223–232. DOI:http://dx.doi.org/10.1145/2556195.2556217

[6] Raman K., Bennett P. N., and 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)*. ACM, New York, NY, USA, 463–472. DOI:http://dx.doi.org/10.1145/2484028.2484089

[7] Smucker M. D. and 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 (SIGIR '12)*. ACM, New York, NY, USA, 95–104. DOI:http://dx.doi.org/10.1145/2348283.2348300

[8] Yilmaz E., Verma M., Craswell N., Radlinski F., and Bailey P. 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. DOI:http://dx.doi.org/10.1145/2661829.2661953