Rendering Expressions to Improve Accuracy of Relevance Assessment for Math Search

Matthias S. Reichenbach, Anurag Agarwal, Richard Zanibbi
Rochester Institute of Technology
1 Lomb Drive, Rochester NY, 14620
{msr5919, axasma, rlaz}@rit.edu

ABSTRACT

Finding ways to help users assess relevance when they search using math expressions is critical for making Mathematical Information Retrieval (MIR) systems easier to use. We designed a study where participants completed search tasks involving mathematical expressions using two different summary styles, and measured response time and relevance assessment accuracy. The control summary style used Google's regular hit formatting where expressions are presented as text (e.g. in LaTeX), while the second summary style renders the math expressions. Participants were undergraduate and graduate students. Participants in the rendered summary style ($n = 19$) had on average a 17.18% higher assessment accuracy than those in the non-rendered summary style ($n = 19$), with no significant difference in response times. Participants in the rendered condition reported having fewer problems reading hits than participants in the control condition. This suggests that users will benefit from search engines that properly render math expressions in their hit summaries.

Keywords

Mathematical Information Retrieval (MIR); Search Engine Results Page (SERP); Relevance assessment

Categories and Subject Descriptors

H.3.3. [**Information Storage and Retrieval**]: Search and Retrieval — Search process, Selection process; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces — User-centered design.

1. INTRODUCTION

Activity in the area of Mathematical Information Retrieval (MIR) has been increasing in recent years [15]. In 2013 a math retrieval competition was held as part of the NTCIR Workshop [1] and advances have been made in techniques for retrieving expressions by visual appearance [8, 9] and operator structure [11]. Work has also begun on integrating search results obtained from independent text and expression indices [11] and creating search interfaces that simplify math entry, for example by allowing handwriting, images, keyboard and mouse to be used for input [13].

An important open problem in MIR is how best to present search hits for mathematical documents. Little work on this problem is available, and we know of no published human evaluations. Youssef [14] describes an implementation of hit content summarization, where documents are fragmented into small units (e.g.: equations, sentences, tables and graphs), with summaries defined by the top matching fragments for a query using a combination of metrics. One expects that rendering math expressions in search hits, as opposed to presenting them textually (e.g. using their LaTeX encoding in a document) will make it easier for users to read and understand search hits that contain math, but this has not been tested.

For text search interfaces [7], different result presentations (summary styles) have been shown to affect the user's ability to assess relevance [2][10]. In addition, different summary styles are most effective for different information needs. For example, when users want to find a specific piece of information (i.e. an *information need* [3]), longer result summaries are more effective [6]. However, when users want to find a specific website or resource (i.e. a *navigational need* [3]) short summaries are more effective [6]. Previous research has described the specific information needs for math search [16], distinguishing *informational* (e.g. finding a proof) from *resource* needs (a form of navigational need, e.g. locating a tutorial on a topic).

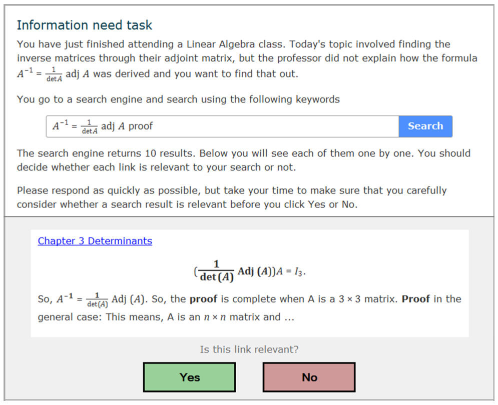
In the presented study we compared two summary styles based on Google search hits, with math expressions rendered in one style and not in the other. Similar to Aula [2], we created pre-defined search tasks and search results, formatted in the two different summary styles and measured relevance assessment accuracy and response time. To test for differences arising from information need, we used two search tasks designed to require two different information needs (the *informational* vs. *resource* needs identified by Zhao, et al. [16]). The following section describes the methodology of our experiment, followed by the experimental results, their discussion, and our conclusions.

2. EXPERIMENTAL DESIGN

Hypotheses. Two different summary styles were considered. The Control summary style is a modification of the Google search result format, and the experimental (Rendered) summary style in which math expressions are rendered (i.e. properly formatted). We hypothesize that the properly formatted expressions will help with readability and thereby allow users to assess relevance faster, similar to what Aula [2] found for text search.

In addition, we wanted to test if participants' ability to assess relevance was affected by search task. We consider two search tasks with different mathematical information needs as identified

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2257-7/14/07 ...$15.00.
.



Information need task

You have just finished attending a Linear Algebra class. Today's topic involved finding the inverse matrices through their adjoint matrix, but the professor did not explain how the formula $A^{-1} = \frac{1}{\det A} \operatorname{adj} A$ was derived and you want to find that out.

You go to a search engine and search using the following keywords

$A^{-1} = \frac{1}{\det A} \operatorname{adj} A$ proof

The search engine returns 10 results. Below you will see each of them one by one. You should decide whether each link is relevant to your search or not.

Please respond as quickly as possible, but take your time to make sure that you carefully consider whether a search result is relevant before you click Yes or No.

Chapter 3 Determinants

$$\left(\frac{\mathbf{1}}{\mathbf{det}\,(A)}\ \mathbf{Adj}\ (A)\right)A = I_3.$$

So, $A^{-\mathbf{1}} = \frac{\mathbf{1}}{\det(A)}$ Adj (A). So, the **proof** is complete when A is a 3 × 3 matrix. **Proof** in the general case: This means, A is an $n \times n$ matrix and ...

Is this link relevant?

Figure 1: Online interface for collecting relevance assessments. The search task description and instructions are displayed in the top of the page, while search hits are displayed in the bottom. When the user presses 'Yes' or 'No,' the bottom panel is replaced by the next search hit.

by Zhao et al. [16]. We hypothesized that better readability should have a larger effect for the search task intended to satisfy an informational need (i.e. needing to ascertain specific information), versus the second search task intended to satisfy a resource need (i.e. a specific type of resource, such as a tutorial).

Design. Posters and email were used to recruit graduate and undergraduate student participants from the College of Computing and Information Science and the College of Science at the Rochester Institute of Technology (USA). Respondents completed a pre-screening questionnaire to assess whether they met the required level of math proficiency, defined as having completed two or more college-level math courses, and experience with computer systems and search engines. Participants were also required to have normal or corrected to normal vision and hearing.

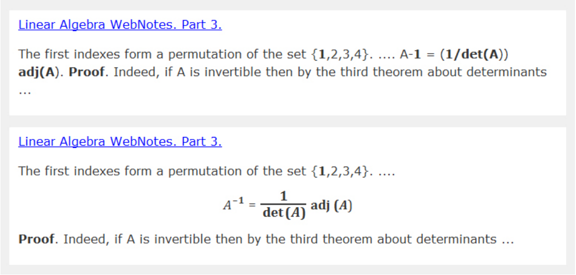
Participants were divided into two groups for each hit summary style. All participants performed three tasks: one familiarization task and the two search tasks with different information needs presented in a counterbalanced order.

The experiment design conforms to a mixed factorial design where the summary style condition was between subjects and the search task condition was within subjects. The dependent variables (DV), were participant response time for assessing whether a hit is relevant to a search task, and the accuracy of their relevance assessment. Further details of the experimental design are provided in the remainder of this section.

2.1 Independent Variables

Search Tasks. We designed two search tasks intended to have differing mathematical information needs. Zhao et al. [16] distinguishes between informational needs that require specific mathematical facts, and resource needs which require a specific resource (e.g. source code or a tutorial). As this was a preliminary study, and we anticipated that participants would find making relevance determinations difficult due to the subject matter, we chose to create a small number of tasks. For each search task, participants were prompted to the underlying information need by means of a short scenario.

Task 1: The task intended to satisfy an information need asked the participant to search for a proof of a linear algebra equation with the query "$A^{-1} = \frac{1}{\det A} \cdot \operatorname{adj} A$ proof." The full text for the task is shown in Figure 1.



Linear Algebra WebNotes. Part 3.

The first indexes form a permutation of the set {**1**,2,3,4}. .... A-**1** = (**1/det(A**)) **adj(A**). **Proof**. Indeed, if A is invertible then by the third theorem about determinants ...

Linear Algebra WebNotes. Part 3.

The first indexes form a permutation of the set {**1**,2,3,4}. ....

$$A^{-\mathbf{1}} = \frac{\mathbf{1}}{\mathbf{det}\,(A)}\ \mathbf{adj}\ (A)$$

Proof. Indeed, if A is invertible then by the third theorem about determinants ...

Figure 2: Illustration of summary styles. From top to bottom: summary style obtained from Google search (SS1, the Control condition), and then for the same hit but with the math expressions rendered (SS2, the Rendered condition).

Task 2: For the second task designed to satisfy a resource need, the participant was asked to search for a tutorial about derivatives of polynomials with the query "$\frac{d}{dx}ax^b = abx^{b-1}$ tutorial." Additional details are available elsewhere [12].

Hit Summary Styles. Two summary styles were used corresponding to the two levels of our summary styles independent variable. The first level (SS1) was used as a control. The hit results were styled based on how they appeared in the Google's results page, effectively using it as the "gold standard" (see Figure 2). Removing the result's URL and any other links besides the title were the only modifications to the original summaries. URLs were removed to prevent participants from making relevance assessments from the URL directly, rather than the content of the search hit result summary itself.

The second level (SS2) was our experimental condition. SS1 was used as the base for SS2, but with every math expression in it properly formatted (see Figure 2). Expressions in the result summaries were converted from their original code (e.g. LaTeX) when available, or visually when not, to MathML — a W3C standard for describing mathematical notation in XML — using MathJax[1]. The converted code was then rendered in our experiment website by Mozilla Firefox's native MathML rendering engine.

2.2 Search Hits and Data Collection

Search Hit Creation and Relevance Determination. The search results for each query were selected from a Google results page after searching with the task's pre-defined query. Query expressions were converted to LaTeX and then stripped of special characters to make them suitable for Google search.

A search result hit was only considered relevant if it contained: 1) at least some portion of the query expression, and 2) the accompanying text query term or a semantically equivalent word (i.e. 'proof' for Task 1, and 'tutorial' for Task 2). Five hits matching this criteria were selected from the search results. Non-relevant hits were selected from search hits that did not contain the query expression but did contain some other expression. In some cases, additional searches were made to generate hits that met the criteria.

Hit Presentation and Data Collection. Data collection was performed using the online system shown in Figure 1. The familiarization and two experimental tasks each had a "card" that slid into view from the right of the data collection web page. The card was split into two parts. The top half described the information need and showed the query terms in a mock-up search bar. This section was visible throughout the completion of the task so participants could refer to it if they needed to see the query terms or remem-

[1] http://www.mathjax.org/

ber something about the information need. The bottom half was used to display hit results and collect binary relevance assessment responses from the participants using ‘Yes’ and ‘No’ buttons. After the participant pressed a button to make a relevance assessment, the current hit result slid out of view towards the left of the screen and a new hit slid into view from the right.

Hit results were presented one-at-a-time to avoid the large effect that ordering in search result pages has on assessment accuracy, as shown by Cutrell and Guan [5]. Presenting hits one-at-a-time also forces participants to consider the contents of each search hit. The presentation order was counterbalanced across participants to avoid ordering effects. Using these two strategies, we hoped to obtain a consistent measurement of perceived relevance for the search hits across participants. A similar design was used earlier by Kickmeier and Albert [10].

The system was run on a server with Apache, PHP and MySQL. The client computers used by the participants had access to the server and were running Windows 7 and the Firefox browser, and had a standard keyboard and mouse.

2.3 Protocol

Participants were told to read the familiarization task and follow the instructions on the screen. The familiarization task had the same structure as the experimental tasks but with only four result hits. They were verbally told to “respond as quickly as possible, but take your time to make sure that you carefully consider whether a search result is relevant before you click Yes or No, even if it takes you longer than it usually does when you search, that is fine."

The experiment’s website then guided the participants through the experimental tasks (which were counter-balanced between participants). Participants were again asked to respond as quickly as possible, but take their time to make sure that they carefully consider whether a search result is relevant before clicking Yes or No, both verbally and with written instructions on-screen. After the familiarization task was completed and these instructions were provided verbally, the experimenter stated that he wouldn’t be able to answer any more questions because the tasks are timed.

Each task started with a short description of the information need and the pre-defined query used to meet the information need and generate the results. Participants were asked to read the tasks and, when ready, click a Start button. At this moment the system started measuring response times and relevance assessments. Each of the hit results related to the search task were displayed one by one — in a counter-balanced order among participants — until all 10 had been assessed.

After finishing the tasks, participants were taken to an online questionnaire. It was designed to measure subjective responses to the system, the summary styles and the tasks. Before leaving, participants were given $10.00 as compensation for their time.

3. RESULTS

A total of 38 participants completed the experiment. All participants reported having normal, or corrected to normal, vision and hearing. Additionally, all participants indicated not having any problems, such as dyslexia, when reading from a computer screen. 73.7% (n=28) of participants were male and 26.3% (n=10) were female. 92.1% (n=35) of participants reported being between the ages of 18 and 24 with the rest reporting being between 25 and 34. 76.3% (n=29) reported their highest level of education as some college with the rest reporting having earned a higher education title.

The mean response time taken by all participants to assess relevancy for each hit was 12.93 seconds ($\sigma = 5.77, n = 757$) and mean relevance assessment accuracy for all participants was 75.57%. A Pearson Correlation test was performed to test for learning effect across both summary styles. A small correlation between presentation order and time was found for the Rendered condition ($r = -0.143, p < 0.01$) but not for the Control condition ($p > 0.05$). Search task presentation order was counter-balanced, and so this effect arises from practice during the experiment, and not the presentation order. No correlation was found between accuracy and presentation order for both summary styles ($p > 0.05$). A small negative correlation between time and accuracy was found for the Control ($r = -0.114, p < 0.05$) but not for the Rendered condition.

Table 1: Relevance assessment accuracies and response times. Task 1 required locating a proof; Task 2 required locating a tutorial. Groups: Control $n = 19$**; Rendered** $n = 19$**; Total** $n = 76$

		Accuracy (%)		Response Time (s)	
Task	Summary	μ	σ	μ	σ
1	Control	69.47	13.11	12.58	4.55
	Rendered	83.10	12.01	14.06	5.11
2	Control	69.71	20.78	12.39	4.79
	Rendered	80.00	15.63	12.70	4.35
1 & 2	(Total)	75.57	16.60	12.93	4.66

Data collected from the experiment was summarized by participant and task. An accuracy score was calculated as the percentage of correct assessments and the response time was calculated as the average time to make a relevance assessment for the hits in the task. The mean time to decide was 12.93 seconds ($\sigma = 4.66, n = 76$) and the mean accuracy was 75.57% ($\sigma = 16.60\%, n = 76$). Table 1 presents the mean and standard deviation for accuracy and timing metrics for each combination of summary style and search task, along with the same metrics for all participants.

A 2 (Search Task) x 2 (Summary Style) mixed-effects factorial ANOVA was performed on accuracy scores. Accuracy scores were found to not change by search task ($F(1, 36) = 0.211, p > 0.05$) and no interaction effect was shown ($F(1, 36) = 0.286, p > 0.05$). However, accuracy scores did change based on summary style ($F(1, 36) = 8.730, p < 0.01$). On average, the percentage of correct relevance assessments by participants in the Rendered condition was 17.18% higher than those in the Control. No effects were observed for response times ($p > 0.05$). A post-hoc power analysis for assessment accuracy was performed ($\pi = 0.82$), which is above the level of 0.80 normally considered adequate. A post-hoc power test for task time was much lower, as the distributions for task time are much more similar than those for assessment accuracy.

Exit Questionnaire. 73.68% ($n = 28$) identified Task 1 (shown in Figure 1) as easier, with 82.14% of participants saying they were more familiar with the math used (linear algebra vs. calculus). A Mann-Whitney Independent Samples test found no significant difference between summary style groups for the questions “I’m familiar with the math involved in these tasks” and “I have had information needs similar to the tasks I just completed” ($p > 0.05$).

A significant difference was found for the question “I had no problems reading the results presented” ($p < 0.005$). Figure 3 provides the histogram of responses from the participants. 15 (78.9%) of participants in the Rendered condition Agreed or Strongly Agreed that they had no problem reading the hit summaries, in comparison with only 5 (26.3%) of the participants in the Control condition.

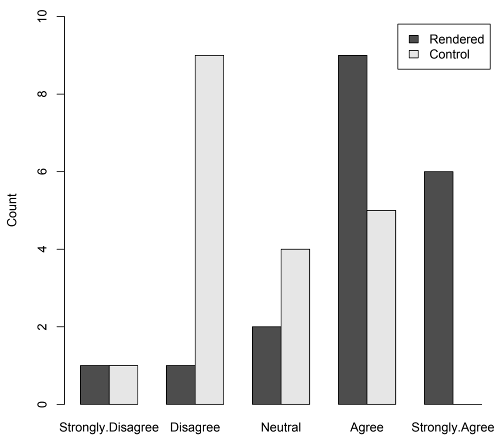


Figure 3: Participant responses from the Rendered and Control summary style conditions for the statement "I had no problems reading the results presented."

4. DISCUSSION

The results support our hypothesis that the users' ability to assess relevance for search hits in math search improves when expressions are rendered. Participants in the rendered condition had on average 17.18% better relevance assessment accuracy, and reported having greater ease with reading hit summaries. Only participants in the Rendered condition showed a learning effect that, if extrapolated, could mean even shorter response times once users are more practiced in math search.

Additionally, the small negative correlation between time and accuracy in the Control presents a violation of the speed-accuracy trade-off that may occur when the ability to discriminate between correct and incorrect alternatives is low [4]. When discriminability is high, reducing speed increases accuracy, whereas with low discriminability reducing speed does not increase accuracy — in fact, in the Control accuracy decreases slightly as response time increases. This is consistent with the participants' self-reporting of how difficult it was to read the search hits (Figure 3).

We suggest a couple of explanations for this result. The first is obvious, in that it easier to see the structure of an expression if it is rendered. The second is that formatting expressions, particularly offset expressions such as shown in Figure 2 segments the hit into smaller regions, making them easier to read. Along those lines, Kickmeier obtained a surprising result that making words bold at random in hit summaries (up to a certain frequency) tended to increase assessment accuracy for textual search hits [10].

Our results do not support our hypothesis that relevance assessment accuracy would be influenced by search task. There is a confound raised by participants' higher familiarity with the math in one of the tasks. Also, a larger number of search tasks would be needed to properly test this.

5. CONCLUSION

Users are accustomed to search result hits containing mostly text and links. Our results suggest that rendering mathematical expressions rather than leaving them in textual form (e.g. LaTeX) significantly increases relevance assessment accuracy for math search hits without significantly increasing assessment time. Given this, search engine designers should make a concerted effort to properly render mathematical expressions presented in hit summaries.

As we knew that evaluating search hits with expressions would be challenging for participants, we chose to consider only two search tasks in this first study, and to present hits one-at-a-time in a counterbalanced order to avoid biases arising from placement in a search results page. Follow-on studies are needed to test whether our findings hold when users consider hits within search result pages, and to examine whether a larger set of search tasks will show information need influencing which hit result summary styles produce the most accurate relevance assessments by users.

In the future, we are interested in testing different summary styles, such as modifying hit summaries to increase the amount of document context that surround the matched expression in the document, showing math expressions that surround a matched expression, or varying the proportion of expressions to text.

6. ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation (USA) under Grant No. IIS-1016815.

7. REFERENCES

[1] A. Aizawa, M. Kohlhase, and I. Ounis. NTCIR-10 math pilot task overview. In *Proc. NII Testbeds and Community for Information access Research (NTCIR)*, pages 654–661, Tokyo, Japan, 2013.

[2] A. Aula. Enhancing the readability of search result summaries. In *Proc. HCI 2004: Design for Life*, pages 1–4, 2004.

[3] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, Sept. 2002.

[4] J. R. Busemeyer. Violations of the speed-accuracy tradeoff relation. In *Time Pressure and Stress in Human Judgment and Decision Making*, pages 181–193. 1993.

[5] Z. Guan and E. Cutrell. An eye tracking study of the effect of target rank on web search. In *Proc. SIGCHI*, pages 417–420, 2007.

[6] Z. Guan and E. Cutrell. What are you looking for? an eye-tracking study of information usage in web search. In *Proc. SIGCHI*, pages 407–416, 2007.

[7] M. Hearst. *Search User Interfaces*. Search User Interfaces. Cambridge University Press, 2009.

[8] X. Hu, L. Gao, X. Lin, Z. Tang, X. Lin, and J. B. Baker. Wikimirs: A mathematical information retrieval system for Wikipedia. In *Proc. Joint Conf.Digital Libraries (JCDL)*, pages 11–20, 2013.

[9] S. Kamali and F. W. Tompa. Retrieving documents with mathematical content. In *Proc. SIGIR*, pages 353–462, Dublin, Ireland, Aug. 2013.

[10] M. Kickmeier and D. Albert. The effects of scanability on information search: An online experiment. In *Proc. HCI*, pages 33–36, 2003.

[11] T. T. Nguyen, K. Chang, and S. C. Hui. A math-aware search engine for math question answering system. In *Proc. Information and Knowledge Management*, pages 724–733, 2012.

[12] M. Reichenbach. Improving accuracy of relevance assessment for math search using rendered expressions. Master's thesis, Rochester Institute of Technology (RIT), NY, USA, 2013.

[13] C. Sasarak, K. Hart, R. Pospesel, D. Stalnaker, L. Hu, R. LiVolsi, S. Zhu, and R. Zanibbi. min: A multimodal web interface for math search. *Symp. Human-Computer Interaction and Information Retrieval*, pages online, 4pp, 2012.

[14] A. S. Youssef. Methods of relevance ranking and hit-content generation in math search. In *Calculemus 0́7 / MKM 0́7*, pages 393–406, 2007.

[15] R. Zanibbi and D. Blostein. Recognition and retrieval of mathematical expressions. *Intl. J. Document Analysis and Recognition (IJDAR)*, 15(4):331–357, 2012.

[16] J. Zhao, M.-Y. Kan, and Y. L. Theng. Math information retrieval: user requirements and prototype implementation. In *Proc. JCDL*, pages 187–196, 2008.