Displaying Relevance Scores for Search Results

Guy Shani
Information Systems Engineering
Ben Gurion University
shanigu@bgu.ac.il

Noam Tractinski
Information Systems Engineering
Ben Gurion University
noamt@bgu.ac.il

ABSTRACT

Internet search engines typically compute a relevance score for webpages given the query terms, and then rank the pages by decreasing relevance scores. The popular search engines do not, however, present the relevance scores that were computed during this process. We suggest that these relevance scores may contain information that can help users make conscious decisions. In this paper we evaluate in a user study how users react to the display of such scores. The results indicate that users understand graphical displays of relevance, and make decisions based on these scores. Our results suggest that in the context of exploratory search, relevance scores may cause users to explore more search results.

1. INTRODUCTION

Searching for information over the web has become an everyday task for practically everybody in the modern world. Users of popular search engines, such as Google, Bing, or Yahoo, come from a diverse population of ages, occupations, and education. Thus, search engines must display their results in a simple manner.

It has become customary for search engines to present the webpage results that were computed in the form of a list, ranked by decreasing relevance to the query terms. When the list becomes too long to fit a single page, the user may choose to pass to the next page, observing another portion of the list. Indeed, users have learned to scan the list from the top down, searching for a webpage that would contain the information that they seek [8].

To sort the list by decreasing relevance, search engines typically compute a numeric relevance score to the query terms for each webpage in the pool of candidates. An example is Google's pagerank, which computes a numeric score based on incoming links to a webpage [1]. Modern search engines use complicated methods balancing multiple relevance scoring algorithms, but the result of these advanced algorithms is still typically a list of webpages with a numeric relevance score. Then, the search engine orders the list by decreasing relevance score and presents the sorted list to the user.

Even though the relevance scores are an important part of the computation of the search results, all popular search engines do not present these scores to the user. This is strange, as such scores may help users to make better decisions concerning the process of finding the webpage that contains the needed information. For example, consider two cases, one in which relevance scores slowly diminish, and another in which there are a few very relevant webpages, following which the relevance scores decrease considerably. One might imagine that given slowly decreasing scores, users that don't find the information in the first two webpages may continue to look at more search results, while given rapidly decreasing scores, users who did not find information in the first two webpages may choose not to look at the results with the much lower relevance and instead consider other options, e.g., revising their query.

Displaying relevance scores is not a new idea. Hearst [3] indicates that such displays were popular in the past, but fell out of favor. Commercial search engines do not disclose the reasons for not displaying these scores. Hearst [3] suggests several possible reasons for this practice. First, search engines may consider their relevance scores to be a trade secret, and may fear competitors or spammers would gain some advantage from these scores. A second reason may be the desire to keep the interface simple and clean, with as little diversion as possible. Finally, search engines may consider these scores not to be understandable or useful for users. While we are not aware of published studies that directly assessed the understandability and the usefulness of relevance scores to users, research shows that how search results are displayed makes a difference in how users make use of those results and in users' attitudes towards those results (e.g., [6]).

Thus, the objective of this paper is to evaluate whether different types of relevance score presentation produce differential user behavior (i.e., in terms of number of results examined per query) and attitudes towards the search engine. To this effect, we conduct a study in which we compare three different displays — a list of results without any relevance score display, a list with the actual numeric relevance score, and a list with horizontal bars whose lengths represent their respective relevance scores.

2. BACKGROUND

With the growing use of information systems in practically all aspects of everyday life, the art of finding information in vast databases has become a critical issue. The need for smart information retrieval methods has become even more important as the internet became our major source of information.

Commercial search engines, such as Google, Bing, and Yahoo, are now used by billions of people worldwide. As such, these search engines have also become a major source of income for companies, and significant amounts of research is dedicated to making these search engines more accurate. The accuracy of the search engine is mostly dependent on its ability to estimate the relevance of a webpage to a given user query. Then, the webpages are or-

dered (ranked) by decreasing relevance, and displayed to the user by that order. Such a presentation cues the users to view the results by order of appearance until they find the information that they are looking for. Research has shown that the ordered list has a strong influence on users' selection of links returned by the search engine, in that users are biased towards links at the top of the list [8].

An orthogonal line of research deals with the design of intuitive user interfaces for internet search engines[3]. Currently, all commercial search engines present their results as a list. For each item on the list the search engine presents its title, the url of the webpage, and a snippet of the text in the webpage. Noticeably missing from the information provided for each item by current search engines is the estimated relevance of the webpage to the query, or to the query for the specific individual in a personalized search [10].

Various ideas of presenting users with additional information have been investigated in the past. White et al. [11] suggest showing alongside the regular search results, a list of webpages that users typically go to, when searching for the same terms. Alongside these webpages, they add a graphical bar representing the amount of users who go to that webpage (the popularity of that webpage).Users who experimented with the system reported that they did not find the popularity bars interesting. However, their study did not compare the popularity bars to a standard system (i.e., without popularity bars) or to a numeric display of popularity.

Tanin et al. [9] investigated results display for information retrieval in databases, presenting additional attributes relevant to the query, and the distribution of information for these attributes, assuming that supplying users with additional information would help them in reformulating their queries. Golbeck and Hu [2] investigated the use of various visualization methods (stars, colors, bars) for presenting rating or popularity (but not relevance) of items such as movies or songs, concluding that there was no difference given various visualization techniques, but that users liked the visual presentation of scores nevertheless. Mann and Reiterer [7] studied methods for presenting the overall relevance of the result set. They used bars for presenting the relevance of each keyword to the search results, as well as the overall relevance to the query.

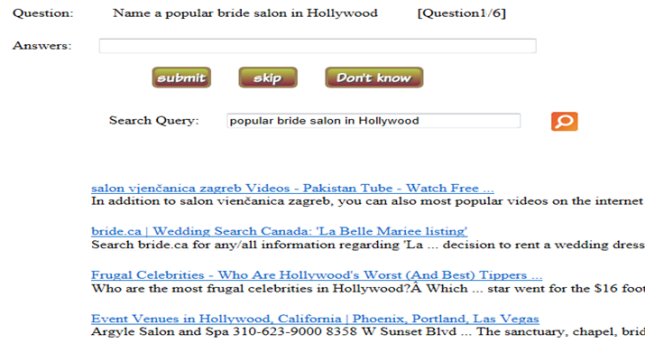
Hoeber and Yang [4] suggest a complex visualization technique called HotMap, to provide better information than the standard results lists. HotMap presents relevance information over the entire set of results in a graphical representation. In a user study they show that HotMap helps users to find the relevant documents, and that the users like the system. HotMap suggests a radically different user interface, and its helpfulness may not apply to presenting relevance scores on standard lists of search results. Iwata et al. [5] also suggest a complex mechanism for displaying relevance in a multi-aspect task, showing increased effectiveness.

Overall, clearly many researchers have considered the presentation of relevance scores of results, and it is thus surprising that there is no recent study that evaluates simple relevance displays within modern search engines. Our objective was to evaluate user's behavior given the presentation of relevance score and their attitudes towards different presentation formats of the scores. For this purpose we conducted a user study, which is described below.

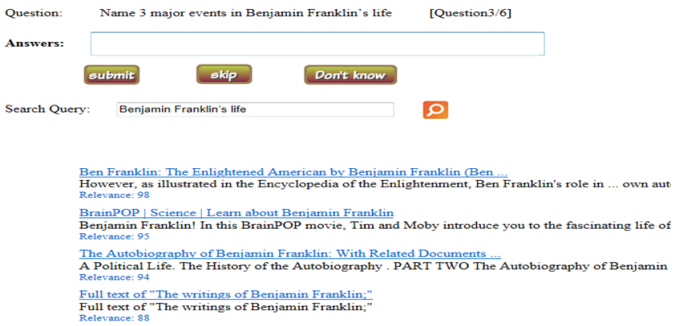
3. METHOD

For the study we created a search engine UI over the Bing API. The UI allowed us to add, below each search result, a numeric or graphical display of the relevance of that result.

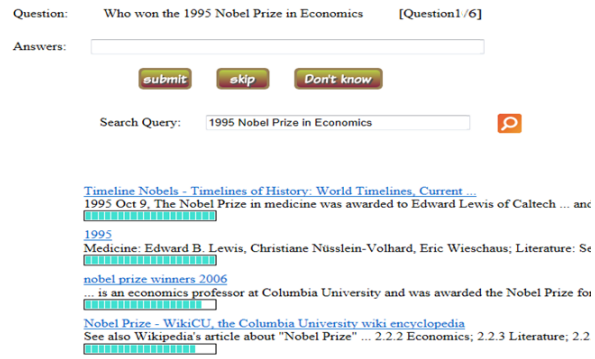
Sample: Ninety two Information Systems Engineering students volunteered to participate in the study, following an email invitation that was sent to all students in the department. To increase participation rate, all participants could enter a raffle for a digital camera.



(a) No relevance display.



(b) Numeric relevance display.



(c) Graphical relevance display.

Figure 1: The results of searching for a query with relevance displays. Showing only the first 4 of the 10 search results.

The email contained a link to the study webpage, and participants were asked to read instructions and complete several tasks online.

Procedure: Participants used a link in the email invitation to enter the study webpage. After reading the instructions on the webpage, the participants continued to performing the study's tasks. Participants were asked to complete 6 information search tasks, to respond to questions about their experience during the study, and then to add their name into the digital camera raffle. Throughout the study we did not inform the participants about its objective. Users became aware of the goal of the study only when answering the questions at the end.

Tasks: Participants were asked to respond to 6 queries which were randomly selected from a pool of 15 different tasks. The tasks were of two types: Queries with a single correct answer and queries with multiple correct answer. For each presented task, users had to type a search query, run the search, review the search results,

and write an answer. The search queries were then run through the Bing API[1], and the results were presented to the users. To make the tasks more challenging, we have removed from the returned search results all the Wikipedia webpages, which for most queries appear at the top of the results' list and contain all the needed information for a correct answer.

Presentation Formats: As explained above, like all other popular search engines, the Bing API does not return a relevance score for the search results. We therefore created artificial relevance scores for the returned items. These artificial scores were randomly decreasing from the first score downwards, as is the case in a ranked list. To gauge the effects of presenting relevance scores we used three different presentations for the relevance scores. As a baseline we showed the query results with no relevance scores, as they appear in popular search engines. A second type of presentation included numeric display of relevance, where scores are presented on a 0 to 100 scale. The third type of presentation included graphical display of relevance, where blue bars represent the relevance. Figure 1 shows some query results with and without relevance scores.

Upon receiving the query's results, users could click on search results, and a new tab would open with the chosen webpage. When the information is found, the user inserts the information into the Answers text box on the study tab, and submit the answer.

4. RESULTS

69 of the participants completed all 6 tasks, and answered the questions at the end. All users together completed 525 tasks (including "don't know" answers but excluding skipping). Overall, participants made 1022 queries (or 1.95 queries per task).

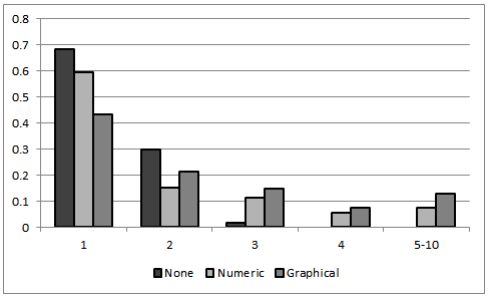


Figure 2: Portion of clicks for various positions in the ranked list of results. Positions 5 through 10 are aggregated.

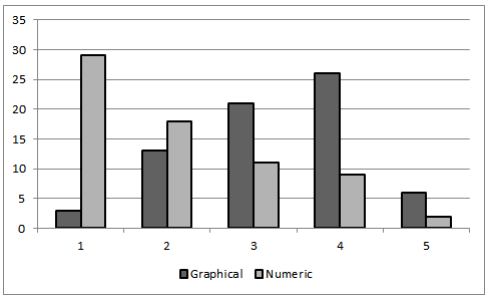


Figure 3: Deemed helpfulness of the relevance scores on a scale of 1 to 5, where 1 means "not at all" and 5 means "very helpful".

[1]`http://www.bing.com/developers/`

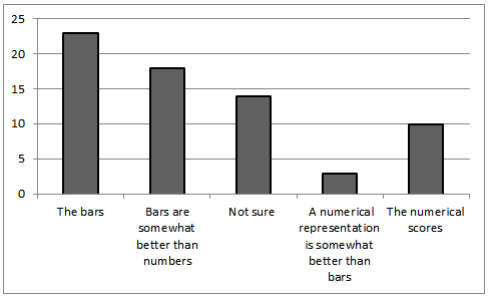


Figure 4: User preference over the relevance display types.

User were correct in about 65% of their answers overall, and clicked "don't know" in less than 20% of the tasks. Analyzing the portion of correct, incorrect, and "don't know" answers given the various relevance displays, our results do not show any statistically significant difference between the relevance bars and no relevance display (with the numeric scores slightly reducing the portion of correct answers). The relatively high proportion of correct answers suggests that users put an effort to fulfill their tasks and find the correct answers. Still, the correctness measure is not too informative for assessing the effects of the presentation formats, because the relevance scores attached to each result were artificial and uninformative.

In general, users took on average 95.7 seconds for answering correctly, 101.9 seconds for answering incorrectly, and 103.6 seconds for answering "don't know". These differences did not statistically significant. Again, these findings suggest that users invested a considerable effort towards finding the information.

4.1 Clicking Patterns

The main focus of this research was the behavior of users given the different presentation formats. The most direct measure of such behavior is the pattern in which search results were clicked given those different presentations. Figure 2 shows the portion of clicks on links in different positions in the result list.

As we can see, there is a considerable difference between the click patterns of the three presentation formats ($\chi^2 = 27.9$, $p < 0.001$). As expected, in all cases users click mostly on the first few results. When no relevance display is presented, users tend to click almost only on the first and second results (98% of the clicks). However, when a relevance score is presented, the first and second results receive only 70% of the clicks. The rest of the clicks involve results in lower positions on the list returned by the search engine. This effect was especially pronounced in the case of the graphical presentation format, in which about 44% of the clicks involved results that appeared below the first two on the list.

On average, participants explored (clicked on) 1.31 results per query. They made on average 1.46 clicks per query when no relevance score was displayed, 1.15 clicks when numeric scores were presented, and 1.33 clicks when displaying graphical relevance bars (differences significant using a paired t-test). Taken together, the results indicate that users do not click on more results when relevance scores are displayed; rather, they tend to click more often on results placed on lower positions in the result list.

This may be because users, upon observing the relatively high relevance score of the results at position 3 and 4, believe that these results may contain needed information, while users that do not observe a relevance score tend to believe that only the first or second result are worthwhile. The difference between numeric and graphical displays can be explained by the lower visibility of the numeric

scores (supported by the questionnaire's results below). When a user does not see the numeric scores, she behaves as in the case where no relevance score is being presented.

4.2 Questionnaire Analysis

We now analyze the results of the questionnaire that participants fulfilled at the end of the study. These results round up and complement the results obtained from the clicks analysis. The findings below reflect the answers of 69 participants who completed the questionnaire in its entirety.

We first wanted to know whether users noticed the relevance scores. We asked two questions — "Did you notice the blue bars below the search results?" and "Did you notice the numbers representing relevance below the search results?". About 97% of the participants noticed the bars, while only 55% of the participants noticed the numeric scores.

We then asked how helpful the two displays presenting relevance scores were. These questions were enabled only for users who reported to have noticed the displays. Figure 3 shows how useful people deem the relevance scores are. The average score for the graphical display was 3.27 while the average helpfulness score for the numeric display was 2.08 (difference is statistically significant using a paired t-test, $t(df) = 136, p < 0.001$). This finding was corroborated by another question in which we asked the participants which relevance display they preferred. The results can be seen in Figure 4. Clearly, users prefer the graphical bars to the numeric scores.

It is also important for the user interface to be as simple as possible, so that users will not have a learning curve for the new features. We hence asked users how fast did they understand what the blue bars represented. Most users (61 out of 69) indicated that they understood what the bars represent "immediately" or "quite quickly".

5. DISCUSSION AND FUTURE WORK

The exclusion of relevance scores from search results is puzzling. Hearst [3] suggested that possible reasons for this exclusion may include attempts to conceal trade secrets, attempts to keep the interface simple, and concerns that users may not be able to understand those scores. The use of simple graphics to represent relevance scores can mitigate the first concern, as those graphics do not show detailed relevance scores (as opposed to the numeric presentations used in this study). The addition of the bars to the results detracts only slightly, if at all, from the simplicity of the user interface. Most importantly, we provided empirical evidence that contradicts the third reason. From the users point of view, relevance scores, presented as analog bars are not only well understood but also have influence on their search behavior.

It is obvious from the results of our user study that people observe and respond to displaying relevance of search results. Users both clicked more often on webpages below the first two results, and stated that the graphical relevance display is helpful. This is even more encouraging when we consider that the relevance scores were artificial. One may expect that the results will be even more pronounced when using true relevance scores. These findings summon further research on this topic. For example, in our study, relevance scores were not informative in order to isolate the net effects of presentation format on the *apparent* relevance of the link rather than on its *actual* relevance and accuracy. Future studies may be designed to test the interactive effects of presentation formats and actual relevance scores.

Another line of future research relates to how changes in the presentation of relevance scores are noticed and interpreted by users. Such research would include, for example, questions regarding how sensitive users are to changes in the length of the filled relevance bar, and are they equally sensitive to corresponding changes in different presentation formats (e.g., numeric vs. analog presentations). We will examine, e.g., whether the numeric scores had low impact because people did not notice them, which could be amended by larger fonts, or because people have difficulty interpreting numeric scores in general.

Our results show that users explore more results when presented with relevance scores. This may be important in the context of exploratory search, where the user needs to collect information from various sources, as opposed to known-item search, where the user is interested in specific information from a single source.

Our findings regarding the effects of presentation formats on search behavior are limited to bottom-line data obtained from users' clicks and answers to a questionnaire. Process data (e.g., by using eye-tracking instruments or think aloud protocols) is needed to improve our understanding of these effects, and provide insights into the processes that underlie the different behavior.

We have listed a few limitations of our study, which we intend to amend in future research. The external validity of the findings can be enhanced: First, there is a need to study a more diverse sample of the user population. Second, future research using true relevance scores will help corroborate the findings of this study.

Finally, this study only compared two methods of displaying relevance, a numeric scores and horizontal bars. Users' understanding and effective use of search results may benefit from other presentation formats or variations on the displays used in this study.

6. REFERENCES

[1] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the seventh international conference on World Wide Web 7*, WWW7, pages 107–117, 1998.

[2] J. Golbeck and C. Hu. Impact of Visualization Methods on Interaction with Search Results. In *CHI*, 2009.

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

[4] O. Hoeber and X. D. Yang. A comparative user study of web search interfaces: Hotmap, concept highlighter, and google. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, WI '06, pages 866–874, 2006.

[5] M. Iwata, T. Sakai, T. Yamamoto, Y. Chen, Y. Liu, J.-R. Wen, and S. Nishio. Aspectiles: tile-based visualization of diversified web search results. In *SIGIR '12*, pages 85–94, 2012.

[6] Y. Kammerer and P. Gerjets. How the interface design influences users' spontaneous trustworthiness evaluations of web search results: comparing a list and a grid interface. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, ETRA '10, pages 299–306, 2010.

[7] T. M. Mann and H. Reiterer. Evaluation of different visualizations of web search results. In *Proceedings of the 11th International Workshop on Database and Expert Systems Applications*, DEXA '00, pages 586–, 2000.

[8] B. Pan, H. Hembrooke, T. Joachims, L. Lorigo, G. Gay, and L. Granka. In google we trust: Users' decisions on rank, position, and relevance. *Journal of Computer-Mediated Communication*, 12(3):801–823, 2007.

[9] E. Tanin, B. Shneiderman, and H. Xie. Browsing large online data tables using generalized query previews. *Information Systems*, 32(3):402–423, 2007.

[10] J. Teevan, S. T. Dumais, and E. Horvitz. Potential for personalization. *ACM Transactions on Computer-Humamn Interacttions*, 17(1):4:1–4:31, April 2010.

[11] R. W. White, M. Bilenko, and S. Cucerzan. Studying the use of popular destinations to enhance web search interaction. In *SIGIR*, pages 159–166, 2007.