How and why do users change their assessment of search results over time?

Maayan Zhitomirsky-Geffet  
Bar-Ilan University  
Ramat-Gan, Israel  
Maayan.Zhitomirsky-Geffet@biu.ac.il

Judit Bar-Ilan  
Bar-Ilan University  
Ramat-Gan, Israel  
Judit.Bar-Ilan@biu.ac.il

Mark Levene  
Birkbeck University  
London, UK  
Mark@dcs.bbk.ac.uk

ABSTRACT
In this study we investigate whether and why users change their preferences when assessing search engine results over time. We conducted a study with 35 subjects who were asked to rank and assign relevance scores to the same set of search results for three times, with a few weeks period between each round. The subjects were then exposed to the differences in their judgements and were asked to explain them. A new coefficient to measure change was introduced to assess the results of the experiment. We found that all the subjects judge the vast majority of the results differently in every round. However, there was less change in relevance judgements than in rankings. Most of the subjects were satisfied with their changes, and did not perceive them as mistakes but rather as a legitimate phenomenon, since they believe that time influences the relevance assessment. Our analysis reveals that the main factors that caused these changes were due to categorical thinking, influence of the learnt information, and environmental and emotional changes.

KEYWORDS
Ranking, relevance judgements, search engines, change over time, locality, categorical thinking

INTRODUCTION
Search engines provide users with a major gateway to the vast amount of information available on the Internet. For almost every query there are numerous results matching the query, yet users are usually content only with the 10-20 items displayed on the first results page (Jansen & Spink, 2006). The top search engine results are presumed to be the most relevant for the given query. In (Hariri, 2011) the author studied Google rankings and asked whether top results are considered more relevant by the users. The fifth ranked result was judged to be of highest relevance slightly more than the top ranked result. Moreover, previous research reveals quite a high level of disagreement between the ranking of search engines and user-produced rankings (Bar-Ilan et al., 2007; Bar-Ilan & Levin, 2011). Inter-user agreement on ranking of search results has also been shown to be quite low due to subjectivity in human judgements (Bar-Ilan et al., 2007; Teevan et al., 2011). These studies only asked the users to rank the results, without asking for their relevance judgements, and the users were asked to rank the results only once.

Scholer, Turpin & Sanderson (2011) studied repeated relevance judgements of TREC evaluators. They found that quite often (for 15-24% of the documents) the evaluators were not consistent in their decisions, and considered these inconsistencies to be errors made by the assessors. Subjectivity and disagreements in user evaluation of search results may stem from numerous factors, such as different backgrounds, fields of expertise and interests.

Another important but currently under-explored factor is time. Thus, if users were asked to repeat the ranking exercise after a certain time period, would their ranking remain the same? Moreover, if we ask users to choose a relevance grade for each result, where relevance grades come in, say, four categories from very relevant to not relevant, would the relevance assessments remain stable over time? Also, it is not well-understood to what extent user preferences change in time and how such knowledge may inform search engines’ rankings.

In this study we aim to explore these questions relating to the influence of time on user assessment of search results. We also try to understand the factors and reasons that influence users’ opinions and cause the change in their preferences in ranking search results. To this end, we conducted a user study with 35 subjects who were asked to rank and assess the relevance of the same set of results for the same given query three times, a few weeks apart. Then the differences between the three rankings and relevancies were shown to the subjects and they were asked to explain them. Their answers are further analysed and classified to understand the factors that influence the user preferences and cause the change in their judgements. To quantitatively evaluate the results we propose a new measure to assess the proportion of change between the user judgements in different rounds of the experiments.

As opposed to the previous studies, we measure changes in users’ rather than experts’ judgements, and also for ranking of results and relevance judgements within four categories, rather than just for binary relevance judgements. We
believe that the changes in ranking and relevance judgements are not necessarily errors, as it is well-known that relevance is subjective and situation dependent (Saracevic, 1996; Mizzaro, 1998), and therefore a document judged to be relevant in a certain situation and time might not be judged to be relevant later on.

**EXPERIMENTAL SETUP**

Thirty five information science students participated in the study. They were presented with the following scenario: “Your aim is to learn about the given topic based only on the search results, in order to be able to prepare a good summary of the topic”. The presented topic (query) was “Cyber warfare” (in English). The participants were then presented with a randomly ordered list of 20 search results for the given query. The search results were presented in a similar style to SERPs, i.e. title, URL and snippet as displayed by the search engine.

The users were asked to assess the relevance of all 20 items on a four level ordinal scale: not relevant (1), slightly relevant (2), somewhat relevant (3) and relevant (4). In addition, they were asked to rank the 10 best results in their opinion, where no ties were allowed. There were two reasons we decided to ask users to rank only the top 10: 1) search engines present 10 results as the default setting, and 2) it requires considerable effort to sift through search results and to rank them without ties. Google Forms were used to collect the answers. The result set comprised the Google results displayed on the first and the tenth result page (i.e. results 1 to 10 and 101 to 110).

Then, the experiment was repeated two more times in a period of a few weeks between each round. The second and third time the participants were given exactly the same instructions and exactly the same query and set of search results, but these were displayed in a different random order, to avoid participants just copying their previous assessments. In each round all the participants saw the results in the same random order. To prevent presentation bias (Bar-Ilan et al., 2009) the participants were also instructed to evaluate every result independently of its position in the list.

**Measures for result aggregation and evaluation**

For each query and result set, we assess what proportion of the results in the set was not given identical ranks or relevance judgements by a specified user, on two different rounds of the experiment. This measures the amount of change at the exact match level (i.e. we consider results with distance 0 between the two ranks or relevance values as identically judged). Further, we consider the case when the rankings or relevance judgements were not precisely identical in both rounds but still sufficiently close.

Formally, we define the non-local change coefficient for a category, $c$, at a distance $d$, with $0 \leq d \leq |c|$, for a given set of results, $r_1, r_2, ..., r_k$, evaluated twice by a user, $u$, either with ranking or relevance values, $r_1$ and $r_2$, as follows:

$$\Omega(c, d) = 1 - \frac{\sum_{i=1}^{k} n(r_i)}{|c|}$$

where

$$n(r) = \begin{cases} 1, & |r_1(r) - r_2(r)| \leq d \\ 0, & \text{otherwise} \end{cases}$$

Thus, $\Omega(c, d)$ is the proportion of the results that are judged to be relevant in the two rounds at distance greater than $d$ in the set $c$.

Note that for $d=0$ the change coefficient reduces to the exact match case, while $d>0$ defines the more general case. For relevance all the 20 results in $R$ were judged by the users and thus all of them are considered in the calculation of the change coefficient. However, for ranking only the top-10 results were actually assessed by the users. Therefore, for ranking only, as there are more results than ranks, the unranked results are technically assigned the rank of 11. Only results with at least one of the ranks being lower than 11 are considered. This is because results that were assigned rank 11 were not actually ranked by the users.

**RESULTS AND DISCUSSION**

We investigated the changes of individual users’ rankings and relevance judgements for the same 20 results between three pairs of rounds of the experiment: 1) the first and the second rounds; 2) the second and the third rounds; 3) the first and the third rounds. To this end, we calculated the average of the change coefficients, $\Omega(d)$, with $0 \leq d \leq 3$ over the individual users’ rankings and relevance judgements. The results are presented in Table 1. We observe that, in general, similar values were obtained for the different round pairs, recalling that when $\Omega(d)=0$, then no change occurred. (We do not show the change coefficient for relevance grades with $d>1$, since they are very close to or equal 0; these are denoted by N/A in Table 1.). As expected, we can see a consistent decrease in the change coefficient, especially the considerable decrease between distance 0 and distance 1, which indicates that most of the changes were local between the rounds. (We consider as local changes within a close range of ranks and relevance grades). We also observe that the relevance grades change less than the rankings.

This provides some evidence for categorical thinking (Mullainathan, 2000) of users when evaluating the search results. Thus, users could not easily distinguish between small differences of close ranks, however they were much better at distinguishing between relevant (category 4) and non-relevant (category 1) of the results. Therefore, we also calculated the proportion of categorical change between the ranked (top-10) and unranked (last-10) results. To this end, we computed the cross-category change as the proportion of results that switched between these categories in different rounds, i.e.
results that were judged as part of the top-10 on one round but not on the other. The averaged results (over all users) are presented in Table 2. As can be observed the vast majority of the results (about 80%) remained in the same category on different rounds.

The majority of the subjects viewed as most important the binary division between top-10 (ranked) and last-10 (unranked/non-relevant) result subsets. The changes of ranks within these two subsets were considered as local and insignificant by the subjects: “First of all, I notice that in spite of the changes in ranks between the rounds, I included the same results in the top-10 ranks, of course, without preserving the previous order within the top-10 results”. This is an explicit argument for categorical thinking (Mullainathan, 2000) of the subjects when ranking search results. It is also supported by our quantitative analysis above shown in Tables 1 and 2, showing the locality of changes and the relative stability of top-10 and last-10 subsets.

In particular, the following repetitive factors have been detected as causing the change in user judgements:

1) **categorical thinking** – binary division of relevant vs non-relevant, the local changes of ranks within the relevant groups are not significant for users: “In most cases there was no significant differences in my judgements – most of the unranked results in the first round remained unranked in the next rounds as well, and most of the ranked results were around the same ranks in all the three rounds.”;

2) **self-improvement** – correcting the mistakes from the first rounds or trying to improve the ranking in every next round: “I acted differently in every round of the experiment, and the reason for that is the feeling that there is something tricky about ranking the results and I have to try and improve my way of ranking every time.”;

3) **knowledge acquisition** – during the first rounds of the experiment the subjects were exposed to information on the topic, which caused them in later rounds to rank the results differently: “The reason for differences in my rankings is that in the beginning of the experiment I only had a very general idea on the topic while from round to round my knowledge expanded and thus my ranking had changed accordingly.”;

4) **irrational factors** that influence the subjective assessment of relevance, which depend on place (of doing the ranking), environment, mood, focus/anxiety

### Table 1. The average change coefficient values for ranking and relevance with different distances and round pairs. Standard deviation values are shown in parentheses following the average.

<table>
<thead>
<tr>
<th></th>
<th>Round 1 vs round 2</th>
<th>Round 2 vs round 3</th>
<th>Round 1 vs round 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>ranking</td>
<td>Ω(0)</td>
<td>Ω(1)</td>
<td>Ω(2)</td>
</tr>
<tr>
<td>relevance</td>
<td>0.83(0.14)</td>
<td>0.80(0.17)</td>
<td>0.79(0.16)</td>
</tr>
<tr>
<td></td>
<td>0.59(0.25)</td>
<td>0.63(0.25)</td>
<td>0.49(0.27)</td>
</tr>
<tr>
<td></td>
<td>0.04(0.10)</td>
<td>0.12(0.11)</td>
<td>0.12(0.11)</td>
</tr>
<tr>
<td></td>
<td>N/A</td>
<td>0.51(0.25)</td>
<td>N/A</td>
</tr>
</tbody>
</table>

### Table 2. The average cross-category change for different pairs of experiment rounds.

<table>
<thead>
<tr>
<th></th>
<th>Round 1 vs round 2</th>
<th>Round 2 vs round 3</th>
<th>Round 1 vs round 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cross-category change</td>
<td>0.22(0.18)</td>
<td>0.19(0.16)</td>
<td>0.22(0.16)</td>
</tr>
</tbody>
</table>

Figure 1 displays the change coefficient values for ranking in round 1 vs 2 for each user at different distance values. As can be observed there were few users with a change of 0 even at distances 1 and 2.

Figure 1: The Ω values for ranking for each of the 35 users in the group and various distances, presented in descending order of Ω at d=0.

**Analysis of the main factors in user assessment changes**

The analysis of the answers that our subjects provided to explain the changes in their assessments of the results, reveals that most users believe that time influences the human relevance judgements. Only a few were surprised and confused by seeing the differences in their own judgements in different rounds of the experiment. Four out of 35 subjects assessed their judgements as consistent with a few local changes between close ranks. They explained it by adhering to the same criteria and type of thinking across all three rounds of the experiment.
and emotions: “The only way to explain this exceptional phenomenon is because the human ranking is influenced by many factors, such as the mood, the anxiety, the loss of patience, the external stimulus, even the design of the website, the place where I have done the rankings (the first and the third rounds were conducted at the campus while the second one was at home).”

Specific results intensely addressed in the subjects’ descriptions of their choices were: 1) Wikipedia (which is always ranked at the top by search engines) vs. scientific sites – many of the subjects felt that Wikipedia is not reliable enough and does not provide an in-depth analysis of the topic, and thus ranked it lower than academic sites, such as www.academia.edu, while others argued in favor of the information quality in Wikipedia and ranked it higher. Some of the subjects ranked Wikipedia at the top only in the first round and then decreased its rank in the further rounds: “In the first round I saw as more relevant the results that are more popular (e.g. Wikipedia), but in the next rounds I thought that this approach was not correct, and it is better to rely on sites that supply more authorized and scientific information, unlike the Wikipedia that gives a general idea of the topic but not necessarily reliable and sufficient”. One of the results was on the cyber warfare in a specific country (Russia), rather than in general. This result’s relevance score and ranking had also led to controversial opinions of the subjects.

Further, the subjects were also trying to understand how Google ranks the results. Most of them were not surprised that their judgements were different from those of Google, even though in many cases Google 100-110 ranked results were judged by the subjects as being in the top-10 ranks. Only three subjects viewed Google as a gold standard for ranking, and inconsistencies with its ranking as mistakes. In addition, presentation order of the results was considered as a cause of differences between the search engine ‘s and the subject’s results: “The list of the results that we were given was not ordered by Google rank. It might be that if we were provided the sorted list according to the Google rank, then my rankings were closer to those of Google.”

Finally, the subjects asserted that Google is based on “wisdom of the crowds” data, and thus ranks higher the popular sites with many incoming links, such as Wikipedia, but for human users the authority, the reliability and the depth of the analysis are more important.

CONCLUSION

The main contribution of this research is that we show how much and why time influences user assessment of search results. In addition, this is a first study that formally measures and investigates the factors that cause the change in user relevance and ranking judgements. We found that users do substantially change their assessment of rankings and relevancies over time, however most of the changes are within the same subset (or category) of results and seem insignificant to the users. The other factors for change were the influence of the information learnt on the topic, the tendency to self-criticism and improvement of the ranking, and the change in the emotional state of the subjects. This research may have practical implications for personalisation, as users’ preferences change over time and therefore the ranking of a search engine should adapt to this. For ranking, with the maximal locality threshold (d=3), only the minority of the changes (about 40%) were non-local, so it is these changed results that are especially in need of personalisation.

In future work we intend to conduct a large-scale experiment and to extend the setting to a set of several queries and various search engines’ results.

REFERENCES