Anchoring and Adjustment in Relevance Estimation

Milad Shokouhi
Microsoft
Cambridge, United Kingdom
milads@microsoft.com

Ryen W. White
Microsoft
Redmond, Washington
ryenw@microsoft.com

Emine Yilmaz
University College London
London, United Kingdom
e.yilmaz@cs.ucl.ac.uk

ABSTRACT

People's tendency to overly rely on prior information has been well studied in psychology in the context of *anchoring and adjustment*. Anchoring biases pervade many aspects of human behavior. In this paper, we present a study of anchoring bias in information retrieval (IR) settings. We provide strong evidence of anchoring during the estimation of document relevance via both human relevance judging and in natural user behavior collected via search log analysis. In particular, we show that sequential relevance judgment of documents collected for the same query could be subject to anchoring bias. That is, the human annotators are likely to assign different relevance labels to a document, depending on the quality of the last document they had judged for the same query. In addition to manually assigned labels, we further show that the implicit relevance labels inferred from click logs can also be affected by anchoring bias. Our experiments over the query logs of a commercial search engine suggested that searchers' interaction with a document can be highly affected by the documents visited immediately beforehand. Our findings have implications for the design of search systems and judgment methodologies that consider and adapt to anchoring effects.

1. INTRODUCTION

Consider a simple experiment; which involves estimating the total value of the following mathematical equation in five seconds:

$$1 \times 2 \times 3 \times 4 \times 5 \times 6 \times 7 \times 8$$

What was your estimate? Now, how about the following equation?

$$8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1$$

It is immediately apparent that the second equation is the same as the first but in reverse order, thus the estimate should not change. In 1974, Tversky and Kahneman [12] conducted the same experiment and presented the first sequence to a group of subjects and the second to another group. They noted that the median estimate was 512 for the subjects in the first group and 2250 for those in the second.[1] This drastic difference between the median estimates of the two groups can be explained by the *anchoring effect* – also referred to as *anchoring bias* [12]. Anchoring – or focalism – is a cognitive bias that explains the human tendency to rely heavily on first presented information (*anchor*) when making decisions. In the example above, given the short amount of time permitted for calculation, the subjects were subconsciously biased by the first few numbers in equations and reached significantly different estimates across the two groups. Anchoring effects have been studied in a range of settings. Northcraft and Neale [7] showed that the listed price of a property affects how much people – including experts – are willing to pay for it despite having access to comprehensive information about the quality factors. Wansink et al. [14] demonstrated that consumers are likely to purchase more products when they are presented in multiple-unit prices and purchase limits (e.g., "On sale – 6 cans for $3" versus "On sale – 50¢"). Here, the number of units acts as the anchor and biases consumer behavior.

In this paper, we study the anchoring effect in IR. In particular, we focus on how anchoring affects the *relevance* ratings of documents. Relevance labels for documents are usually obtained either (1) in a batch form by soliciting explicit judgments from human assessors with respect to a query, or (2) in an online form where the document relevance is inferred by using implicit feedback e.g., time spent on a document (dwell time) from search log data. We first demonstrate that the relevance judgment of a document is affected by the judgment that was made on an immediately preceding document. We further focus on relevance inferred from dwell time and show that the relevance of the previously-clicked document can also have a significant impact on the time searchers spend on the current document, which could significantly affect relevance inferred via dwell time. While the notion of relevance in IR has been studied for decades [8], we believe that our study is the first investigation of anchoring bias on implicitly inferred relevance labels.

Our results suggest that (1) the biases introduced by the relevance of the last labeled/clicked document should be considered both for batch and online evaluation, (2) that dwell time estimates should be used with caution as proxy for relevance, in particular for the documents that are presented towards the bottom of a ranked list (which are likely to be considered only after higher-ranked alternatives), and (3) the models that infer relevance from dwell time could be improved by incorporating the relevance label (or dwell time) of the previously judged/clicked document, if available.

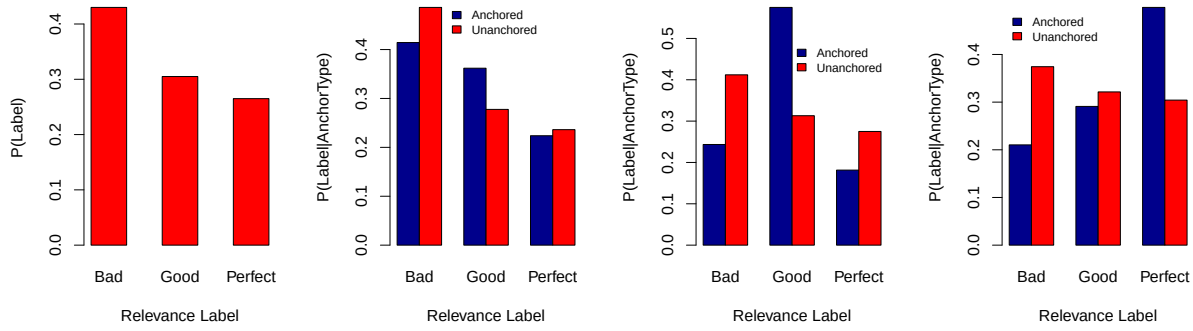
2. ANCHORED RELEVANCE RATINGS

Relevance judgments for a query are typically collected in batches where each assessor may rate multiple documents for the same query [13]. The problem with this common practice is that the

[1]The correct value is 40320.

SIGIR'15, August 09 - 13, 2015, Santiago, Chile.

DOI: http://dx.doi.org/10.1145/2766XXX.XXXXXXX.



(a) Anchor-Type: None (b) Anchor-Type: Bad (c) Anchor-Type: Good (d) Anchor-Type: Perfect

Figure 1: Conditional probability of observing a relevance label for a document given the relevance label assigned to its anchor. (a) probability distribution when the document labels are collected in the absence of anchors. (b-d) the dark blue bars show the labels assigned to anchored documents separately for each class of anchor. In each plot, the red columns depict the label distribution over the same subset of documents based on the unanchored judgments.

document(s) judged early for a given query can *anchor* judges' remaining judgments on that topic. Carterette and Soboroff [1] were the first to report that sequential labeling of documents may affect the assigned relevance label. They observed that sequentially judged documents tend to receive the same labels, and referred to this phenomenon as "autocorrelation". However, the authors did not provide any explanation for their observations nor compare their results with unanchored datasets to establish a ground truth.

Scholer et al. [9] performed the first analysis of anchoring in relevance assessments by measuring the "priming effect" in relevance judgments.[2] They used the relevance of the top k labeled documents by an assessor as the anchor and analyzed how the relevance of these documents may affect the labels assigned to the documents judged subsequently. They showed that if the judges are presented with many documents of high relevance when they start assessing for a query, they tend to assign lower relevance labels to documents labeled later on (and vice versa). In our work, part of our focus is to analyze how the relevance of the previously-labeled document can affect the relevance label assigned to the current document. In contrast to Scholer et al. [9], our results show that when labeling a document, judges tend to assign the same label as the previous document they have labeled.

Our conclusions do not necessarily reject those reached by Scholer et al. [9]. The differences are caused by the type of anchoring that is considered in the two studies. Scholer et al. [9] focus on long-term anchoring (top k labeled documents as the anchor) and analyze how this affects the relevance labels assigned to the documents judged later. In our work, we focus on the short-term anchoring (last labeled document as the anchor) and analyze how this affects the relevance labels assigned to the document judged immediately after. The relevance of a single document is unlikely to have a significant effect on judges' overall expectation of relevance. However, having seen a document of a certain relevance level, the judges might subconsciously expect the relevance level of the next document judged to be similar, and hence their judgments may be affected.

For our experiments, we sampled 400 queries from a ranker training dataset of the Bing search engine. For each query in this dataset there are tens of documents from which we randomly selected three. The first two documents are used as anchors ($\mathcal{A}_1$, $\mathcal{A}_2$) and the last document is used as the target (anchored) (τ) document to measure the impact of the anchoring effect. We hired three mutually exclusive groups of professional relevance assessors from the crowdsourcing platform of the Bing search engine, and provided them with an identical judging guidelines for rating the relevance of documents in three levels (Bad, Good, Perfect). In total we collected our judgments from 220 assessors. We used the first group of judges (81 assessors) to collect *unanchored* relevance labels for our target documents. Each document is judged by three different assessors from this group and is annotated based on the majority vote (although other consensus models can be also used instead). In 8.5% of cases (34/400 queries) there were ties, which were broken by selecting one of the labels at random. Figure 1(a) depicts the distribution of relevance labels for target documents assigned by these judges.

The assessors in the other two groups are used to collect anchored judgments. We refer to these assessors as *anchored judges*. The anchored judges first rate an anchor document ($\mathcal{A}_1$ or $\mathcal{A}_2$ depending on the group) and then rate the target document (τ) for a query. Therefore, the target documents are rated by both groups of anchored judges, while the anchor documents are different across the two groups. Again, each anchor and target document is rated by three different judges in each group, and the final label is determined by the majority vote and tie-breaking as with the unanchored set. Judges rated a maximum of 20 (query, anchor, target) tuples. They always judged an anchor document ($\in \mathcal{A}_1$ or $\mathcal{A}_2$) then the target (anchored) document ($\in \tau$). For each task, a tuple was randomly sampled, without replacement, from the pool of 400 total. To ensure that there are no systematic differences between the anchored groups we compared the overall distribution of labels. An unpaired t-test found no statistically significant differences ($p = 0.979$).

The first question that we investigate is if relevance judgments are subject to anchoring effects. That is: *Does anchor quality affect the labels assigned to target documents?* We grouped the anchored judgments collected for the target documents according to the labels assigned to their anchor documents. The dark blue bars in Figure 1 represent the probability of observing a relevance label for a target document, depending on the rating assigned to its anchor. For instance, Figure 1(b) includes only cases where the target document was judged immediately after an anchor page which was rated as *Bad* by the assessor. Figure 1(c-d) are generated in a similar manner but for *Good* and *Perfect* anchors respectively. The red bars in Figure 1(b-d) represent the distribution of labels assigned to the same set of documents but by unanchored judges. It is clear from these

[2]There is a subtle difference between the anchoring and priming paradigms; in the latter, the priming information (or anchor) is often externally provided by the experimenter, while in the former it is internally generated by the participants themselves [11].

graphs that the anchored and unanchored judgments have different distributions despite the fact that they are computed over the same set of target documents. We applied Chi-squared tests to compare the label distributions for each of the three anchor types in Figure 1(b-d). The results confirm that the differences between the anchored and unanchored groups are statistically significant (Bad anchor: $\chi^2(2) = 8.91, p = 0.0028$; Good anchor: $\chi^2(2) = 80.49, p < 0.0001$; Perfect anchor: $\chi^2(2) = 6.76, p = 0.0093$). This provides supporting evidence for the presence of anchoring biases that can be introduced in collecting sequential relevance judgments from human assessors. Overall, there is a stronger anchoring effect for Good and Perfect judgments. This is expected as Good and Perfect documents both represent relevant documents and are more similar to each other than to Bad documents, which are irrelevant. The next question that naturally arises from these findings is: *Can the anchoring direction be predicted?* That is, given the label of the anchor document can one make any predictions about the label that will be assigned to the target (anchored) document?

The conditional probabilities in Figure 1(c-d) clearly suggest that the most likely label for a relevant (Good, Perfect) target document is the rating assigned to its anchor. In other words, the anchoring direction is *towards* the anchor. However, Figure 1(b) shows that non-relevant anchors (Bad) have the opposite effect, with anchored judges being observed to be less likely to select a rating of Bad. Understanding all factors that affect the anchoring direction in relevance judgments is an interesting direction for future research.

3. ANCHORED CLICK BEHAVIOR

Search result clickthrough rate was once commonly used to infer the relevance of documents [3]. Subsequently, it has been shown that clickthrough statistics are often highly affected by issues such as presentation bias and perceived relevance of the documents. The perceived relevance of a document is mainly based on the summary (snippet) of the document presented on the result page, and can be different than the actual relevance of the document; hence, searchers may end up clicking on a document and discover that it is not relevant [2]. In order to overcome this problem, dwell time, the time spent examining a document, has been proposed as an implicit signal of relevance. Dwell time has been examined in a number of previous studies to infer searcher satisfaction [4, 6] and relevance [5] from observed search behavior. Over many years, a dwell time of 30 seconds has become a standard threshold from which to infer document relevance from document examination behavior [10]. Document visits with dwell times exceeding that threshold have been regarded as implicit indications of relevance.

Beyond controlled experimental settings, such as that employed in the previous section, we were also interested in whether there was any evidence of anchoring effects in the wild, in naturalistic search settings such as Web search. In such settings, anchoring might affect relevance labels inferred implicitly from search behaviors. We used six weeks of search logs from the Bing search engine. These logs contained millions of query-URL pairs on which to perform our study. To remove geographic and linguistic variations we focused on queries generated by searchers within the United States geographic locale. The logs contained queries, the time-ordered sequence of result clicks for each query, and dwell time on each of the landing pages reached through a result click.

In this analysis, we focused on the effects of anchoring on landing page dwell time. We calculated dwell time based on the time between subsequent search engine interactions, including re-visits to the search result page and query reformulations. Dwell times could not be computed for the last clicks for impressions (since there was no subsequent event on which to base dwell time estimates), and these clicks were ignored in our analysis. We divided the landing page dwell times into two groups: (i) Quickback: Dwell time of 15s or less, and (ii) Satisfied: Dwell time of 30s or more. These thresholds were derived from previous research on implicit feedback and satisfaction modeling in Web search settings [4]. Documents with dwell times ranging from 16-29s (inclusive) were excluded in order to simplify our analysis since it is less clear whether such dwell times are associated with satisfaction or dissatisfaction.

We utilize similar terminology as used in the previous section. For queries with multiple clicks, we define the set of *anchor* clicks as the first clicks and the set of *target* clicks as the second clicks. Note that we may observe separate instances of the same (query, clicked URL) pair in both sets depending on search activity, e.g. for a given query q, a searcher may click on a document d and terminate the search session (unanchored), while another searcher may click on a few documents first before clicking on d (anchored). The average rank position (r) of the anchor clicks and the target click is significantly different (anchor click: r = 2.28, target (anchored) click: r = 4.15, independent measures t-test: $p < 0.001$), signaling that searchers often adhere to a top-down examination strategy.

The first question that we sought to answer was whether there were differences in the dwell times for the target clicks given the nature of the anchor. Since we consider clicks within a query, we could be more confident that the searcher had the same intention with each observed click. Table 1 shows the distribution of dwell times (as percentages of the total count of query-click pairs in the four groups) across the four combinations of dwell times assigned to the first and the second click for a query. The table shows that in this analysis, consistent satisfaction (i.e., pairs of *Satisfied* clicks) is observed most frequently, whereas decreased satisfaction (i.e., transitions from *Satisfied* to *Quickback* clicks) occurs least often.

Given the raw counts that are used in computing the percentages reported in Table 1, we can also compute the extent to which they deviate from expected given independence between the first and second click. The numbers in brackets present the percentage deviations from the *expected* values given independence between the dwell time groups (i.e., for each cell, expected = (sum of row $\times$ sum of column) / overall total). It is clear that there is a significant deviation from the expected dwell time in the distribution of dwell times on the second (target) click for a query given the dwell time of the first (anchor) click (Chi-squared test over the frequency counts for the cells in Table 1 produces $\chi^2(1) = 3139222.15, p < .0001$). This suggests that there is a strong association between the nature of the dwell time on an anchor click and the dwell time on the target (anchored) click that follows. However, this analysis is only for queries with multiple clicks, and there may be limitations in considering such situations alone. For instance, low quality documents—that tend to have high Quickback rate—are more likely to appear in low quality search results. That is, if there is a low-relevance document returned in the top results, there might be a higher chance that the other top results are also low quality. To help address this concern, we identified a set of *unanchored* clicks comprising those with no preceding clicks (i.e., the first click for queries) from the same six weeks of data used in our log-based analysis. We then investigated the dwell times of the same query-URL pairs when they appear in our anchored sets. For each of the two anchor types (*Quickback* and *Satisfied*) we then computed the distribution of median [3] dwell times for the anchored clicks, and compared it against the distribution of median dwell times from the unanchored clicks.

Figure 2 shows the cumulative distribution of median dwell times across the 60 seconds immediately following the page load (including the region from 16-29s that excluded in the earlier analysis),

[3]The median is less sensitive to outliers than the mean.

Table 1: Distribution of dwell times across all combinations of the two dwell time groups (Quickback ($\leq$ 15s), Satisfied ($\geq$ 30s)). Values in brackets denote percentage deviations from expected given independence between dwell time bucket groups of first and second clicks. Arrows denote direction of the deviation (up=higher than expected, down=lower than expected).

		2nd click	
		Quickback	**Satisfied**
1st click	**Quickback**	19.74% [▲46.32%]	25.51% [▼19.68%]
	Satisfied	10.08% [▼38.28%]	44.68% [▲16.26%]

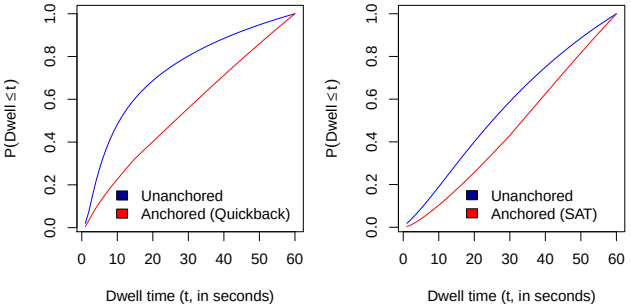


Figure 2: Cumulative dwell time distributions for anchored and unanchored clicks, per anchor type (Quickback and Satisfied).

grouped by anchor type. At any point on the x-axis of the figure, the value on the y-axis reflects P(Dwell) $\leq t$, where t denotes the dwell time in seconds.

From Figure 2 we can observe clear differences in dwell time distributions between the documents in the anchored and unanchored sets. These differences are significant according to the Kolmogorov Smirnov test within each anchor type (Quickback Anchor: $\mathcal{D} = 0.6333$, $p < 0.001$; SAT Anchor: $\mathcal{D} = 0.2833$, $p = 0.0162$). This demonstrates that the dwell time on the anchor document is related to the dwell time on the target page. Research on dwell time typically considers the time spent on each document independently for applications such as satisfaction modeling, e.g., [6]. Our findings in this section suggest that models that interpret dwell times of a document from median dwell time across all searchers may be affected by anchoring biases. These models also need to consider the dwell time of any anchor that may be present when making inferences about the target (anchored) click, for applications such as relevance and satisfaction estimation.

The trends in our findings are clear and we analyze the aggregate behavior of millions of searchers, improving our confidence in the robustness of our conclusions. However, we should also acknowledge that there are other factors that may influence the dwell time beyond the nature of the anchor, such as those associated with searcher traits (e.g., people with a tendency to review pages quickly) or task constraints (e.g., pressing deadlines leading to the rapid review of content). Further work is required to understand the significance of these additional factors on the generalizability of our conclusions about anchoring effects and dwell times.

4. CONCLUSIONS

Anchoring is a cognitive bias that explains the human tendency to rely heavily on first presented information (*anchor*) when making decisions. In this paper, we studied how anchoring may affect human perception of relevance during relevance judging and in the examination of search results in naturalistic search settings. In particular, we focused on (1) relevance labels obtained from relevance assessors by obtaining explicit judgments, and (2) relevance labels inferred from dwell time, the time spent on a document. We showed that relevance of the last labeled document can have a significant effect on the relevance label assigned to the current document during relevance assessment. Our results demonstrate that judges tend to assign different labels depending on the relevance of the previously labeled document. We showed that the impact of short-term anchoring based on preceding documents could be different to those reported previously based on the longer term anchoring [9]. Determining the point where one effect diminishes and the other becomes dominant is an interesting future direction. We further demonstrated that the time searchers spend dwelling on a document is highly related to the amount of time they have spent on the last document that they clicked on, which can lead to significant biases on relevance labels inferred from dwell time.

Our findings can significantly impact the evaluation of retrieval systems and the training of learning-to-rank algorithms. Our results suggest that when human generated or implicit relevance labels are used, labels assigned to previous documents need to be considered.

References

[1] B. Carterette and I. Soboroff. The effect of assessor error on ir system evaluation. In *SIGIR*, 2010.

[2] G. Dupret and C. Liao. A model to estimate intrinsic document relevance from the clickthrough logs of a web search engine. In *WSDM*, 2010.

[3] G. Dupret and B. Piwowarski. A user browsing model to predict search engine click data from past observations. In *SIGIR*, 2008.

[4] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. Evaluating implicit measures to improve web search. *ACM Trans. Inf. Syst.*, 23(2):147–168, Apr. 2005.

[5] D. Kelly and N. J. Belkin. Display time as implicit feedback: Understanding task effects. In *SIGIR*, 2004.

[6] Y. Kim, A. Hassan, R. W. White, and I. Zitouni. Modeling dwell time to predict click-level satisfaction. In *WSDM*, 2014.

[7] G. B. Northcraft and M. A. Neale. Experts, amateurs, and real estate: An anchoring-and-adjustment perspective on property pricing decisions. *Organizational behavior and human decision processes*, 39(1):84–97, 1987.

[8] T. Saracevic. Relevance: A review of the literature and a framework for thinking on the notion in information science. part ii: nature and manifestations of relevance. *JASIST*, 58(13): 1915–1933, 2007.

[9] F. Scholer, D. Kelly, W.-C. Wu, H. S. Lee, and W. Webber. The effect of threshold priming and need for cognition on relevance calibration and assessment. In *SIGIR*, 2013.

[10] D. Sculley, R. G. Malkin, S. Basu, and R. J. Bayardo. Predicting bounce rates in sponsored search advertisements. In *SIGKDD*, 2009.

[11] F. Strack and T. Mussweiler. Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility. *Journal of Personality and Social Psychology*, 73(3):437, 1997.

[12] A. Tversky and D. Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1974.

[13] E. M. Voorhees, D. K. Harman, et al. *TREC: Experiment and evaluation in information retrieval*. MIT Press, 2005.

[14] B. Wansink, R. J. Kent, and S. J. Hoch. An anchoring and adjustment model of purchase quantity decisions. *Journal of Marketing Research*, pages 71–81, 1998.