Empirical Evidence for Search Effectiveness Models

Alfan Farizki Wicaksono
The University of Melbourne
Melbourne, Australia

Alistair Moffat
The University of Melbourne
Melbourne, Australia

ABSTRACT

Given a SERP in response to a user-originated query, Moffat et al. (CIKM 2013; TOIS 2017) suggest that $C(i)$, the conditional continuation probability of the user examining the $(i+1)$ st element presented in the SERP, given that they are known to have examined the i th one, is positively correlated with both i and with the user's initial estimate of the volume of answer pages they are looking for, and negatively correlated with the extent to which suitable answer pages have been identified in the SERP at positions 1 through i. Here we first describe a methodology for specifying how $C(i)$ should be defined in practical (as against ideal) settings, and then evaluate the applicability of the approach using three large search interaction logs from two different sources.

KEYWORDS

User model; evaluation; adaptive metric; average precision

ACM Reference Format:
Alfan Farizki Wicaksono and Alistair Moffat. 2018. Empirical Evidence for Search Effectiveness Models. In *The 27th ACM International Conference on Information and Knowledge Management (CIKM '18), October 22–26, 2018, Torino, Italy*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3269206.3269242

1 INTRODUCTION AND BACKGROUND

In a sequence of recent work Moffat et al. consider how users interact with search engine result pages (SERPs), and the implications that has on search evaluation [2, 6–8]. They argue for weighted-precision metrics in which the i th ranked document has a weight $W(i)$, and propose that if a gain of $0 \le r_i \le 1$ is derived from that i th document, then the usefulness of the SERP is given by:

$$\mathsf{M}_W(\langle r_1, r_2, \ldots\rangle) = \sum_{i=1}^{\infty} r_i \cdot W(i)\,. \tag{1}$$

If it is assumed that the weights $W(i)$ sum to one, then M_W has units of "average gain per document inspected". If it is further assumed that the weights are non-increasing, then a second value

$$C(i) = \frac{W(i+1)}{W(i)} \tag{2}$$

can be associated with each depth $i \ge 1$. Based on a probabilistic user who starts at the top of the SERP and proceeds sequentially from one document to the next until they finish interacting with the SERP, $C(i)$ represents the *conditional continuation probability* of them viewing the $(i+1)$ th document after viewing the i th. Such approaches are also sometimes called *cascade models* [1, 3, 4]. Moffat et al. further postulate that $C(i)$ has three defining behaviors:

- positive correlation with i, so that as the user proceeds through the SERP, their conditional likelihood of continuing increases;
- positive correlation with T, the "volume" of relevance the user had in mind (consciously or subconsciously) when they commenced their search, so that the more the user is searching for, the more patient they will be while looking; and
- negative correlation with the extent to which relevance "volume" had been accumulated as a result of observing the results at positions 1 through i in the SERP, so that as the user accomplishes their goal, they become less patient.

In combination, these three relationships give rise to metrics that are *adaptive*, with the weights $W(i)$ not just a function of i, but also a function of the user's search goal, and of the extent to which the SERP addresses that goal. Note that to be consistent with the cascade assumption, adaptivity should be based solely on the elements in the SERP that the user has seen, and not on ones that they have not. This latter requirement is violated by average precision (AP) [7, 8] and also by the recent proposal of Jiang and Allan [5], see Moffat and Wicaksono [6].

As a "proof of existence" of such $C(i)$ functions being possible, Moffat et al. [7, 8] observe that defining a new metric "INST" via

$$C(i) = \left(\frac{i + T + T_i - 1}{i + T + T_i}\right)^2, \tag{3}$$

where $T_i = T - \left(\sum_{j=1}^{i} r_j\right)$ is an estimate of the relevance volume as yet unfound by depth i, meets the three correlation requirements.

Our Contribution. We consider how to define $C(i)$ in practice so as to reflect user behavior and allow weighted-precision metrics to be instantiated. Making use of three large interaction logs from two different sources, we then ask whether the three relationships postulated by Moffat et al. [7, 8] can be detected.

2 INTERACTION LOGS

This section describes the interaction data we have employed.

Data Resources. Two sources of interaction data are summarized in Table 1. The first is a sample of user activity logs from the English-language service of `Seek.com`[1], a job search site covering a range of Australian and international markets. The `Seek.com` data covers two modalities, search on mobile devices using a purpose-built app that has no pagination and continuous scrolling, and search using a

CIKM '18, October 22–26, 2018, Torino, Italy

ACM 978-1-4503-6014-2/18/10…$15.00
https://doi.org/10.1145/3269206.3269242

[1] http://seek.com.au. For privacy and commercial reasons this data cannot be made public. The terms of service and privacy policies of `Seek.com` were followed during the collection and analysis of this data.

	Seek.com, iOS/Android	Seek.com, browser	Yandex
Users	1,141	4,626	unknown
SERPs	31,741	21,682	1,147,815
SERP size	unlimited	paginated, 20	truncated, 10
Search domain	jobs	jobs	web
Click-throughs	yes	yes	yes
Impressions	yes	yes	no
Rel. judgments	no	no	yes

Table 1: Search interaction logs. The Seek.com data is a sample drawn from 15 October 2017 to 8 April 2018 for iOS/Android users, and 5 February 2018 to 8 February 2018 for browser users.

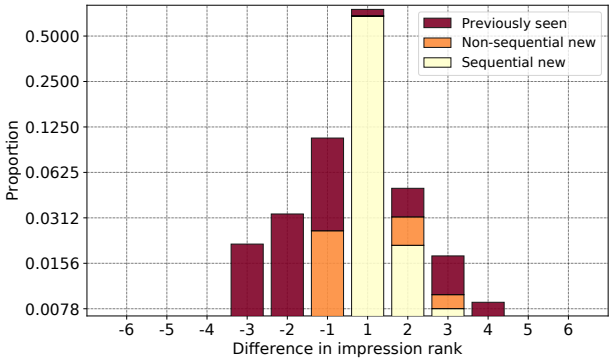

Figure 1: The distribution of impression jumps in the Seek.com iOS/Android interaction log. Note the logarithmic vertical scale.

web browser and a full-keyboard computer that has more traditional SERPs served, each containing 20 job summaries. Browser-based sessions initiated from mobile phones are not included in either of the two datasets.

The Seek.com logs record the activities of people seeking employment, or seeking information about employment situations even if not actively planning to apply, and include *impression* information derived from scrolling behavior. An impression is associated with any job advertisement that is fully visible on-screen for a minimum of 0.5 seconds, an interpretation that is valid because each job advertisement displayed has a relatively large screen footprint, and typically only one advert can be fully visible at any given instant, even in a browser window. The *impression sequences* are chronologically ordered. For example, the sequence $\langle 1, 2, 3, 2, 3, 4, 5, 7\rangle$ indicates that the user viewed the first result, then the second, then the third, then went backwards to view the second one again, and so on, exiting the SERP after viewing its seventh item, having skipped its sixth. Note that the impression point can shift by more than one if the intervening items are on-screen for less than 0.5 seconds.

The second source of interaction data is a publicly-available query log from Russian search company Yandex, containing ordered click-through sequences (rather than impressions). Relevance outcomes for more than a million of the SERPs appearing in the log are also included, based on judgments made a year after the logs had been collected[2].

Impression Sequences. The Seek.com interaction logs cover a set of users $U = \{u_1, \ldots, u_{|U|}\}$. Each user u_i is associated with a set of impression sequences $\mathcal{P}(u_i) = \{P_1, \ldots, P_{|\mathcal{P}(u_i)|}\}$, the lists of job advertisements they perused during their searches. Each sequence $P = \langle p_1, \ldots, p_{n(P)}\rangle$ consists of a list of $n(P)$ ranks p_k, where k is the index in the impression sequence, and p_k is an ordinal rank position in the SERP, starting at 1. After the $n(P)$ th impression the user ceased examining the SERP and initiated some other action.

One of the key assumptions of the cascade model is that users scan the SERP from top to bottom, starting at rank 1, then viewing ranks 2, 3, 4, and so on, until they stop looking at results. Figure 1 shows the distribution of *impression jumps*, computed as $p_{k+1} - p_k$ for $1 \le k < n(P)$ and then aggregated over all impression sequences P in the Seek.com log for iOS/Android users. The observation that "−1" impression jumps are the second most common after "+1" impression jumps has been made in connection with a range of previous log data, suggesting that users pursue a "one step backwards, two steps forwards" approach to viewing the SERP (see, for example, Thomas et al. [9]). To confirm that supposition, each bar in Figure 1 is broken into three components, representing, respectively: transition to a rank position that had been previously viewed; transition to a rank position that had not previously been viewed, and is *not* the rank one greater than the previous maximum rank viewed; and transition to a rank position that *is* exactly one greater than the previous maximum rank viewed. That is, when a SERP element is visited for the first time, with high probability it is the next one in sequence that has not yet been viewed. Similar analysis of the browser-based impression sequences showed an even stronger pattern of the next new result visited being the one after the current maximum rank, and a very strong dominance of "+1 to next new" and "+2 to next new" moves. The difference between the two can be explained by noting that scrolling using a computer is usually a more precise action than scrolling on a touch-screen.

Filtering Out Non-Views. While exploring the browser-based search interactions we noted a disproportionate number of very large backward jumps in connection with page boundary transitions, for example $\langle \ldots, 19, 20, 1, 21, \ldots\rangle$; and at the end of the sequence, for example, $\langle \ldots, 34, 35, 22, 21\rangle$. These arise when users decide to finish viewing their current page of results, and use the mouse and browser scrollbar to quickly return to the top of the page to initiate a fresh action, such as loading the next page of results or to reformulate the query via the search box at the top of the page. The imprecision associated with such gross scrolling movements may result in results near the top of the page being on display for more than 0.5 seconds, but not actually being the target of the user's attention at that time. Before proceeding further with these logs we filtered them to remove decreasing subsequences that started with a long backwards jump greater than 10 and that ended at the first element of a page. In the case of the two sequences shown as examples, they would be rewritten as $\langle \ldots, 19, 20, 21, \ldots\rangle$ and $\langle \ldots, 34, 35\rangle$ respectively. Approximately 13.255% and 0.844% of the impression sequences from browser and iOS/Android apps users, respectively, were altered in this way.

[2]https://academy.yandex.ru/events/data_analysis/relpred2011/

Rule	Continuation contributions for p_k, for $1 \le k \le n(P)$
L	$N_L(p_k, P)$ += 1, if and only if $k < n(P)$; $D_L(p_k, P)$ += 1.
M	$N_M(p_k, P)$ += 1, if and only if $p_k < \max_{1 \le i \le n(P)} p_i$; $D_M(p_k, P)$ += 1.
G	$N_G(p_k, P)$ += 1, if and only if $p_k < \max_{k < i \le n(P)} p_i$; $D_G(p_k, P)$ += 1.

Table 2: Three alternative definitions of "continue" for the impression sequence $P = \langle p_1, p_2, \ldots, p_{n(P)} \rangle$, described in terms of $N(p_k, P)$ and $D(p_k, P)$, the respective numerator and denominator contributions arising from the k th impression in P. Both $N(p_k, P)$ and $D(p_k, P)$ are accumulated by iterating over $1 \le k \le n(P)$. This process leaves $N(i, P) = D(i, P) = 0$ if there is no value $1 \le k \le n(P)$ for which $p_k = i$.

Inferring $C(i)$. To compute empirical estimates of $C(i)$ a binary indicator variable is associated with each of the $n(P)$ impressions in each sequence P. Table 2 indicates three ways in which this might be done. Heuristic "L" records only the last impression in P as being a non-continuation, incrementing the continuation "numerator count" $N_L(p_k, P)$ once for every rank p_k that is not that final impression. The second approach, heuristic "M" assigns non-continuation to all instances of the maximum rank recorded in the impression sequence. Finally, heuristic "G" is a hybrid of the two, and records a continuation whenever an impression is followed by one in the sequence that is at a higher rank position.

For example, consider $P = \langle 1, 2, 1, 3, 4, 2, 1, 3, 2 \rangle$. Using method L the sequence of count increments would result in (among other values) $N_L(1, P) = D_L(1, P) = 3$; $N_L(2, P) = 2$ and $D_L(2, P) = 3$; and $N_L(6, P) = D_L(6, P) = 0$. If the same sequence were to be processed using method M, (some of) the outcomes would be $N_M(2, P) = D_M(2, P) = 3$ and $N_M(4, P) = 0$ and $D_M(4, P) = 1$; and if method G was employed, both $N_G(2, P)$ and $N_G(3, P)$ would be one less that the corresponding $D_G(2, P)$ and $D_G(3, P)$ values, because the last two values in P would both be deemed to have presented an instance of non-continuation.

The per-sequence numerator and denominator values can then be aggregated in two ways to form an overall estimate $\hat{C}(i)$: by micro-averaging across users and queries,

$$\hat{C}(i) = \frac{\sum_{u \in U} \sum_{P \in \mathcal{P}(u)} N(i, P)}{\sum_{u \in U} \sum_{P \in \mathcal{P}(u)} D(i, P)}; \tag{4}$$

or by macro-averaging across users,

$$\hat{C}(i) = \frac{1}{|U'(i)|} \sum_{u \in U'(i)} \frac{\sum_{P \in \mathcal{P}(u)} N(i, P)}{\sum_{P \in \mathcal{P}(u)} D(i, P)}. \tag{5}$$

In both computations $N(i, P)$ and $D(i, P)$ are computed via the operational definitions given in Table 2; and in Equation 5 the set of non-zero indicators is used, $U'(i) = \{u \in U \mid \sum_{P \in \mathcal{P}(u)} D(i, P) > 0\}$.

Figure 2 shows the resulting empirical $\hat{C}(i)$ functions for phone app-based search (top) and for browser-based search (bottom). As a reference the $C(i)$ curves for two static metrics are also shown: SDCG@50, a scaled version of DCG in which the maximum score is adjusted to 1.0; and INSQ with $T = 3$ [7]. For app-based search, the $\hat{C}(i)$ distribution increases as a function of i for all three methods (Table 2), and is relatively smooth. The inferred $\hat{C}(i)$ distribution for the browser-based search shows the same background pattern, but with marked punctuation at page boundaries, highlighting user reluctance to step to new pages. All three estimation methods give rise to the same general trend, differing in detail but not in their overall behavior.

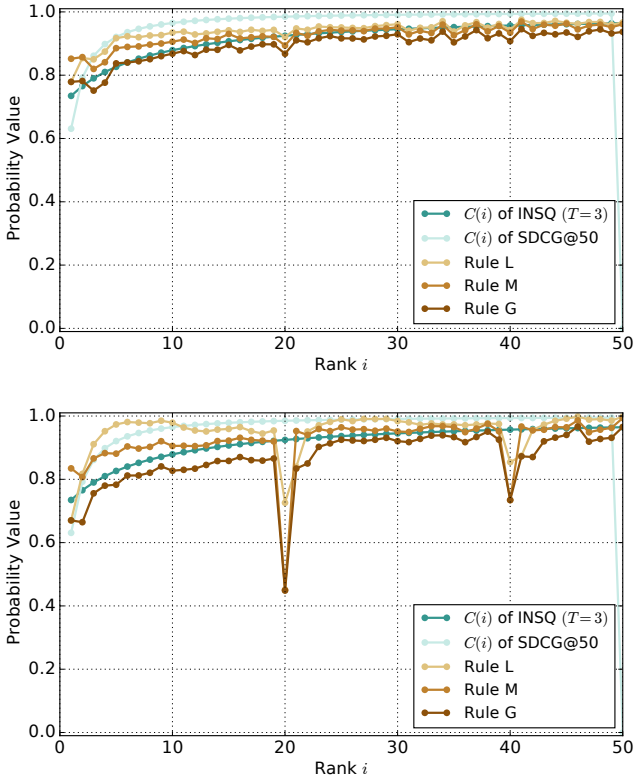

Figure 2: Empirical $\hat{C}(i)$ for iOS/Android users (top) and browser-based users (bottom), macro-averaged from the filtered `Seek.com` impression sequences, using three different heuristics for estimating $C(i)$. The micro-averaged performance was very similar.

3 MODELING $C(i)$

We consider two methods for fitting the inferred $\hat{C}(i)$ functions.

Best-Fit Static Parameters. Table 3 considers three non-adaptive weighted-precision effectiveness metrics, and lists the parameters that provide the best fit for method G, minimizing the weighted-by-frequency mean-squared-error between the inferred $\hat{C}(i)$ values and the $C(i)$ values that define each metric. As can be seen, among these three options INSQ provides the closest fit, with RBP also providing a better approximation than SDCG, which is a very deep metric. Browser-based search tends to be shallower than app-based search, and users are more persistent when job-searching using their phone than when using a computer, in part a consequence of the page breaks in the latter environment.

Model	iOS/Android		browser	
	parameter	MSE	parameter	MSE
SDCG	$k = 51$	1.20×10^{-2}	$k = 51$	1.50×10^{-2}
RBP	$p = 0.86$	0.40×10^{-2}	$p = 0.77$	0.82×10^{-2}
INSQ	$T = 3.70$	0.23×10^{-2}	$T = 1.80$	0.42×10^{-2}

Table 3: Best-fit parameters for three static user models, and weighted mean-squared errors between $C(i)$ and $\hat{C}(i)$ across the first 50 results in each SERP.

Factor	iOS/Android		browser	
	coeff.	p	coeff.	p
intercept	22.337	0.000	32.098	0.000
k	0.976	0.000	0.933	0.000
i	1.033	0.000	1.063	0.000
$i\%20 = 0$	0.970	0.858	0.366	0.000
serp_apps	1.096	0.012	1.752	0.000
sess_apps	1.024	0.112	1.006	0.507

Table 4: Multiplicative effect sizes and corresponding p values for factors affecting $\hat{C}(i)$ for the binary stop/continue indicators associated with impression sequences when categorized using method G.

Logistic Regression. The 0/1 stop/continue indicator variables $N(p_k, P)$ (described in Table 2) can be employed with logistic regression to measure the effect of possible contributing variables:

- the offset k within P, as an unbounded positive integer;
- the current rank, $i = p_k$, as an unbounded positive integer;
- whether i is a multiple of 20, as a binary indicator;
- the number of job applications made until this k th position in P, denoted serp_apps, a non-negative integer; and
- the number of job applications made until this point in the current session, denoted sess_apps, a non-negative integer.

To create sessions, the set $\mathcal{P}(u_i)$ associated with each user was segmented, and queries within a 30-minute interval of a previous action were regarded as being part of the same session. This definition is widely used for web search; we note that for job search, genuine sessions might in fact span days or weeks. When processed in this way the average session consisted of 2.916 and 2.691 queries/SERPS for app- and browser-based search respectively, amplifying the previous observation about relative user patience for the two search modes. To differentiate between "job seeking" and "information gathering" activities, only the sessions that led to at least one job application were considered in this analysis.

Table 4 lists multiplicative effect sizes computed in connection with method G (Table 2). As can be seen, with this interpretation of "continue", $\hat{C}(i)$ decreases with position in the impression sequence (k) and increases with the rank position in the SERP (i); for browser-based search is strongly affected by the page boundaries; and is also positively correlated with the number of applications lodged in regard to both the SERP and the session. The same fitting process for methods M and L showed similar overall patterns, but in some cases with the roles of i and k reversed.

The relationship between serp_apps and $C(i)$ is at odds with the INST assumption that $C(i)$ decreases as relevant documents are found, and job-search and web-search may differ in this regard – it might be, for example, that once a person takes the significant step of applying for one job on a SERP, they then maximize their likely gain by proceeding to make multiple applications.

Click-Based Evaluation. The Yandex data listed in Table 1 includes a set of 260 queries and more than a million SERPS for which relevance judgments are available for all clicked documents. We grouped these by query (covering multiple different users and served SERPS for each) and selected the click streams for the ten most frequent queries for which at least one relevant document was clicked in each SERP. This resulted in a set of 553,392 SERPS, covering a total of 719,889 clicks. Those click sequences were then processed using method G to develop binary $\hat{C}(i)$ indicators, and logistic regression again undertaken. We also estimated a value T_{est} for T as the average number of clicks on relevant documents across the set of SERPS associated with each query. The following factors and multiplicative effect sizes were then computed, all highly significant: k at 1.479; T_{est} at 27.310; and R_k at 0.739, where R_k is the number of clicks on relevant documents up to and including the k th click. That is, T_{est} (as estimated here) is strongly correlated with continuation likelihood, but in tension with that, as relevant documents are reached and clicked, users tend to stop searching.

4 CONCLUSIONS

We sought to explore the extent to which the assumptions associated with the INST $C/W/L$ metric could be supported through analysis of user interaction data. The supposition that $C(i)$ increases with i has been largely validated, but the complementary expectation that $C(i)$ would decrease as relevant documents were found was not supported in the job search interaction log. We conclude that either job searchers have different behaviors to web searchers; or that "relevance" in job search is different to "applications"; or that $C(i)$ actually increases as relevance is accumulated by the user. We hope to use further log data to distinguish between these possibilities.

Acknowledgment. This work was funded under the Australian Research Council's Linkage Projects funding scheme (project number LP150100252), with additional support from Seek.com.

REFERENCES

[1] A. Ashkan and C. L. A. Clarke. On the informativeness of cascade and intent-aware effectiveness measures. In *Proc. WWW*, pages 407–416, 2011.
[2] L. Azzopardi, P. Thomas, and N. Craswell. Measuring the utility of search engine result pages: An information foraging measure. In *Proc. SIGIR*, pages 605–614, 2018.
[3] B. Carterette. System effectiveness, user models, and user utility: A conceptual framework for investigation. In *Proc. SIGIR*, pages 903–912, 2011.
[4] O. Chapelle, D. Metzler, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *Proc. CIKM*, pages 621–630, 2009.
[5] J. Jiang and J. Allan. Adaptive persistence for search effectiveness measures. In *Proc. CIKM*, pages 747–756, 2017.
[6] A. Moffat and A. F. Wicaksono. Users, adaptivity, and bad abandonment. In *Proc. SIGIR*, pages 897–900, 2018.
[7] A. Moffat, P. Thomas, and F. Scholer. Users versus models: What observation tells us about effectiveness metrics. In *Proc. CIKM*, pages 659–668, 2013.
[8] A. Moffat, P. Bailey, F. Scholer, and P. Thomas. Incorporating user expectations and behavior into the measurement of search effectiveness. *ACM Trans. Inf. Sys.*, 35(3):24:1–24:38, 2017.
[9] P. Thomas, F. Scholer, and A. Moffat. What users do: The eyes have it. In *Proc. Asia Info. Retri. Soc. Conf.*, pages 416–427, 2013.