Better Effectiveness Metrics for SERPs, Cards, and Rankings

Paul Thomas
Microsoft
Canberra, Australia

Alistair Moffat
University of Melbourne
Melbourne, Australia

Peter Bailey
Microsoft
Canberra, Australia

Falk Scholer
RMIT University
Melbourne, Australia

Nick Craswell
Microsoft
Bellevue, WA, USA

ABSTRACT

Offline metrics for IR evaluation are often derived from a user model that seeks to capture the interaction between the user and the ranking, conflating the interaction with a ranking of documents with the user's interaction with the search results page. A desirable property of any effectiveness metric is if the scores it generates over a set of rankings correlate well with the "satisfaction" or "goodness" scores attributed to those same rankings by a population of searchers.

Using data from a large-scale web search engine, we find that offline effectiveness metrics do not correlate well with a behavioural measure of satisfaction that can be inferred from user activity logs. We then examine three mechanisms to improve the correlation: tuning the model parameters; improving the label coverage, so that more kinds of item are labelled and hence included in the evaluation; and modifying the underlying user models that describe the metrics. In combination, these three mechanisms transform a wide range of common metrics into "card-aware" variants which allow for the gain from cards (or snippets), varying probabilities of clickthrough, and good abandonment.

ACM Reference Format:
Paul Thomas, Alistair Moffat, Peter Bailey, Falk Scholer, and Nick Craswell. 2018. Better Effectiveness Metrics for SERPs, Cards, and Rankings. In *23rd Australasian Document Computing Symposium (ADCS '18), December 11–12, 2018, Dunedin, New Zealand*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3291992.3292002

1 SEARCH ENGINE RESULT PAGES

Modern web search engine result pages (SERPs) contain advertisements, "instant answers" and factoids, "deep" links into a site, maps, rows of images or video frames, suggested query refinements, and other elements. Even where the results are traditional links, they are typically illustrated by text extracted from the underlying resource such as titles, query-biased captions, or representative images. More generally *cards*—distinct elements onscreen, each with their own self-contained information—might be static (for example, a factoid answer with no supporting link), might be clickable (for example, a link to an underlying document), or might be interactive in some other way. Different cards can convey different amounts of information, can be differently attractive or unattractive relative to the search that led to them being displayed, and can lead to different documents or to none at all.

Conventional evaluation approaches that focus on document relevance are a poor match for modern SERPs with varying kinds of card. Established metrics typically model a user who reads all linked documents returned in an answer list, and benefits only from those documents. They do not allow for the additional gain that might arise from other interface elements, and nor do they deduct the gain from documents that—for whatever reason—do not get looked at by the user. This is clearly a gross simplification, and prevents accurate measurement of SERP usefulness.

As an example, consider the SERPs in Figure 1. Card **A** is excellent, and provides a direct answer. This card has a high gain, potentially leading to "good abandonment": cases where there is no further interaction between the user and the SERP, because they have already achieved their goal [15]. There is no underlying document for card **A**, and nothing to click on; it is meaningless to talk about "document **A**", or about the (further) gain that might be associated with it. Card **B** is similar although, with older data, the gain may be partial. In this case there is a synthetic document available (that is, one that was constructed by the search service rather than being a native crawled document published on a web site). Card **C** summarises a relevant document, but does so by extracting exactly the information that the user sought. As was the case with cards **A** and **B**, good abandonment is possible, even though it is a conventional extractive summary based on a published webpage.

Cards **D** and **E** are both attractive, and a searcher might be drawn to click either one and browse the linked documents. As it turns out, the document behind card **D** is not useful while that behind document **E** is; note that there is no necessary connection between the attractiveness of the card and the gain that accrues from some further element that is activated if the card is selected.

In this work we develop "card-aware" metrics which allow for:

- gain from cards, as well as from documents;
- varying probability of clicking through to examine a document, due to differing attractiveness, gain in the card, or whether a document even exists; and
- the possibility of good abandonment, that is, satisfaction without any clickthrough at all.

The approach we employ is applicable to any of the weighted precision family of metrics, including precision (P@k); scaled discounted cumulative gain (that is, DCG@k normalised into the range zero

ADCS '18, December 11–12, 2018, Dunedin, New Zealand

ACM ISBN 978-1-4503-6549-9/18/12…$15.00
https://doi.org/10.1145/3291992.3292002

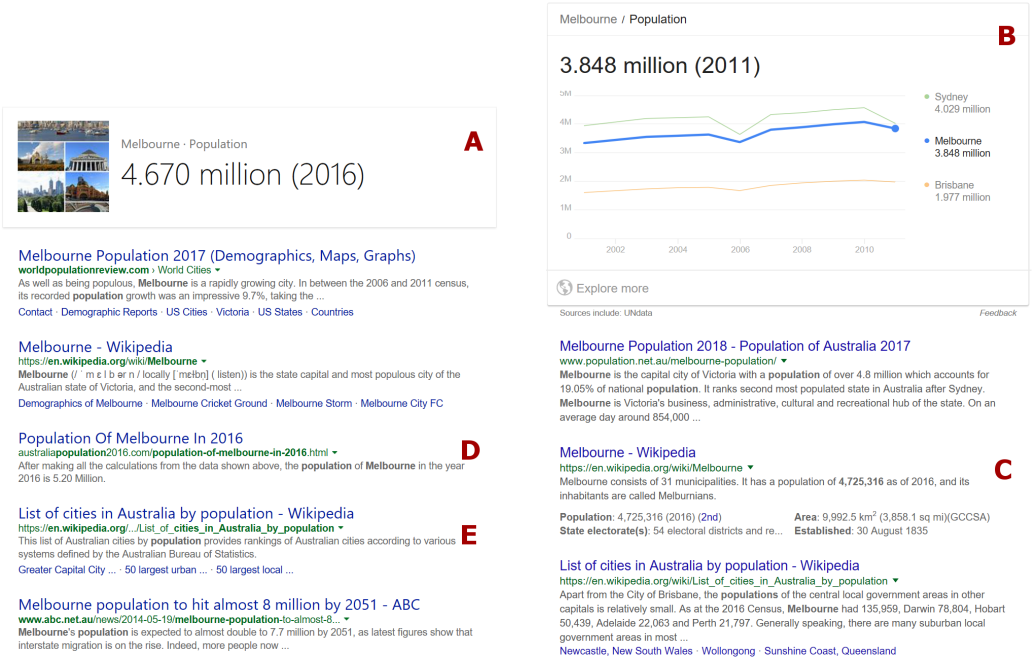

Figure 1: Partial SERPs from Bing (left) and Google (right) for the query [population of melbourne], captured January 2018.

to one assuming there are k full-gain documents); reciprocal rank (RR); rank-biased precision (RBP) [20]; and INST [22].

We employ session-based user interaction data from bing.com to demonstrate the validity of this approach, focusing on a set of common queries, and the user actions associated with the corresponding SERPs. We use *query reformulation rate* as a surrogate for satisfaction, arguing that query reformulation within multi-query sessions is a measurable signal of dissatisfaction. By independently constructing judgements for the main cards present in each SERP and for the documents underlying those cards, we are able to compare card-oblivious and card-aware metric variants, and demonstrate that the scores generated by the card-aware approaches correlate better with user satisfaction.

2 RELATED WORK

Weighted-precision metrics. Many effectiveness metrics including precision at k documents (P@k), reciprocal rank (RR), rank-biased precision (RBP) [20], and average precision (AP), are *weighted-precision* metrics [30]. Sometimes also known as *cascade* approaches [8], metrics in this family assume that each user looks at the document at rank one (and in general, the document at rank i), gets some gain from doing so (possibly zero), and then decides whether to continue on to the document at rank two (and in general, the document at rank $i+1$) with some probability. Over a universe of users, each document in the ranking is assigned a *weight* that corresponds to the fraction of attention it gets, with the weights of necessity being non-increasing as the depth in the ranking increases.

In this linear framework the user model (and metric) is defined entirely by the behaviour at each decision point. Following Moffat et al. [22], we use C_i to denote the conditional probability of any single user continuing from rank i to rank $i+1$, given they just viewed the document at rank i. Weighted-precision metrics are then defined solely by the corresponding values for C_i. For example, for RBP, C_i is a fixed value, $C_i = \phi$, for some constant ϕ chosen according to the anticipated persistence of the user; and for RR, $C_i = 1$ when the i th document is not relevant, and $C_i = 0$ when it is. Moffat et al. [22] provide other examples.

Given the vector $\mathbf{C}$, the corresponding weight vector $\mathbf{W}$ is readily computed [22]. The final metric value is then the inner product of $\mathbf{W}$ and a vector of gain at each rank. If the continuation probability C_i depends not only on i, but also on the gain values at ranks 1 to i, then $\mathbf{W}$ and the user model it corresponds to are said to be *adaptive*. In the case of RR, for example, $W_i = 1/k$ for $i \le k$ when the first relevant document is at rank k, and $W_i = 0$ when $i > k$.

Moffat et al. [22] describe a further adaptive weighted-precision effectiveness metric, INST. They argue that the continuation probability C_i might be correlated with rank, i; with the user's target relevance, T; and with the gains in the listing to depth i.

Beyond documents. Various researchers have examined the role of the SERP as an artefact in the search process. Li et al. [15] introduced "good abandonment", describing how information on a SERP might completely satisfy a searcher, and investigated its prevalance using both PC and mobile search query logs. Chilton and Teevan [7] further examined how direct answers in the SERP affect follow-on interactions with SERP elements, and also in repeated search behaviours. Search success metrics have also incorporated good abandonment: see, for example, Khabsa et al. [13]. Bailey et al. [3] introduced "whole page relevance" and described an evaluation method SASI that allows labelling of various facets of SERP elements, beyond simple topical relevance, as well as capturing holistic properties such as freshness or coherence. Arguello et al. [1] examined image and web results, and find that lack of coherence between

the two has a significant effect on user behaviour. Kim et al. [14] examined how aspects of the SERP affect the judging process in side-by-side preference evaluation methods, finding that multiple dimensions beyond topical relevance are factored into decisions.

Enhanced user modelling within metrics. Traditional evaluation with test collections makes use of a set of search topics, ranked lists of documents returned by a retrieval system, and a relevance judgement for each topic-document combination This approach assumes that users view documents, and that the content of relevant documents provides gain. However, most IR systems don't display documents directly, but first show a search results page, typically including short summaries of each document. Based on the summary, the user then decides whether or not to view the underlying document. This interaction is explored by Turpin et al. [26], who conclude that system orderings may change substantially when summaries are taken into account.

Using interaction data from users of a commercial search engine, Yilmaz et al. [28] developed a search user model that incorporates explicit probabilities for clicking on summaries, based on both the summary quality and the relevance of the underlying document, leading to the Expected Browsing Utility metric, EBU. Their experiments demonstrate that EBU is more closely correlated with user behaviour as indicated by clicks than are other metrics.

Smucker and Clarke [25] cast the search process into a framework based on time, explicitly modelling the durations of stages such as the relatively shorter period required to read a document summary compared to reading the actual document, and define the *time-biased gain* effectiveness metric. Azzopardi et al. [2] also considered time, deriving a **C** function which takes into account the time needed to read cards of different types and which is based on models from information foraging.

Finally, Zhang and Zhai [29] proposed the interface card model, which represents the retrieval model as a sequence of "interface cards" that should be presented to the user to maximize their gain and minimise the effort required. Their framework uses Markov Decision Processes and reinforcement learning to solve the optimization problem.

Other user considerations. Of ongoing interest is the relationship between IR effectiveness metrics and user-based measures of system performance. Jiang and Allan [11] investigated the correlation between user-reported perceptions (performance and task difficulty) and session-based metrics. They found that a variant of sDCG (session-DCG, see Järvelin et al. [10]) normalised by the number of queries in the session showed the strongest relationship.

The relationship between online and offline effectiveness metrics and user satisfaction was studied by Chen et al. [6], who demonstrated that offline metrics are more strongly correlated with user-based measures in a homogeneous search environment such as the traditional "ten blue links", while online metrics show a higher correlation in a heterogeneous environment such as when the results from different verticals are incorporated into a results page. Other work has investigated the extension of click models to account for the presence of specialised vertical results on a search engine result page. Wang et al. [27] developed a click model that accounts for different user behaviours depending on the type of result (for example, multimedia or application) that is displayed. These models were extended by Markov et al. [18] into vertical-aware effectiveness metrics, and shown to correlate more strongly with online signals such as click behavior than other metrics, depending on the kind of vertical that was present in the answer list. To account for the presence of verticals and cards in mobile search results, Luo et al. [16] proposed Height-Biased Gain, a metric that takes the height of different answer items into account when calculating gain.

Other indicators have also been employed. For example, Shokouhi and Guo [24] determined good abandonment by tracking the length of time each card is visible on the screen, calling it relevant if visible for longer than some minimum time threshold, in the expectation that the user paused at that point to obtain the card's gain. Kelly [12] surveyed work on developing and deploying signals based on user actions. Finally, note that document "usefulness" may be different to document relevance [17].

3 EVALUATING EFFECTIVENESS METRICS

Queries. We work with a set of 994 common English-language queries issued to bing.com, from the United States. The queries are used as they were typed and include alternative phrasings of what is likely to be the same need (for example, [yahoo mail] and [yahoo mail login]), as well as clear misspellings (example: [facebok]). The query set represents a large number of clear navigational tasks, but also broader queries such as [crossword puzzles], [taylor swift], and [women scientists].

Cards. Each query was re-issued to bing.com, the (first) resulting SERP was captured, and the resulting cards were recorded. The median number of cards per SERP was 12 (interquartile range 10–13), including "organic" results, advertisements, multimedia (video and images), and a range of rich results spanning maps, business information, reviews, download links, share tickers, sports results, lottery results, dictionary lookups, and others. We excluded small, largely content-free elements such as pagination, result counts, or privacy notices, but included query suggestions as these may also be useful to a searcher.

A median of 9 cards per SERP pointed to other web pages (median 1 advertisement and 8 organic results). Those referenced pages were labelled for relevance on a four-point scale using labels "bad", "fair", "good", and "excellent". (A fifth label, "perfect", was combined with "excellent".) We used trusted crowd workers who were subject to quality control checks and were paid by the hour. Results that did not point to a web page did not receive relevance labels. Around 10% of pages were labelled by two judges, and under 3% of pages received two different labels. In these cases we used the "lower". Pages which were without labels, e.g. if they could not be loaded, were assumed non-relevant.

Reformulation. Since satisfaction cannot be observed at scale, we use *query reformulations* as a proxy measure [9]. We say that a query is reformulated if, at a later point in the same session, a second query is issued which has at least one-third of the terms in common: for example, if [Star Wars] is followed at some stage by [Star Wars movie], or if [restaurant near me] is followed by [cafe near me] or by [restaurants New York]. Reformulation is a signal that the response to the first query was inadequate, and is an indicator of failure in regard to the first query in the pair.

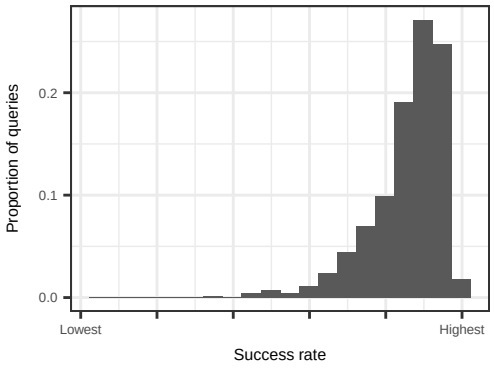

Figure 2: Success rate across 994 distinct queries at bing.com. The figure shows the distribution, but for commercial reasons we do not give actual rates (they are not from 0 to 1).

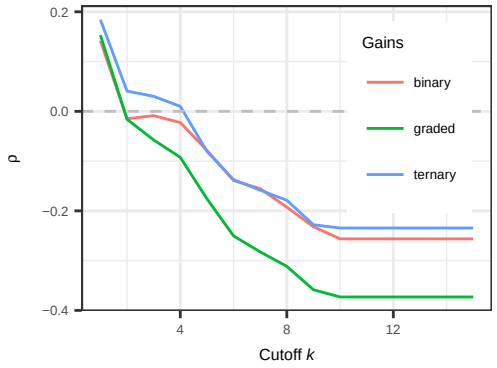

Figure 3: Correlation between metric score and success rate (Spearman's ρ) for P@k as a function of k, for three gain mappings.

Terms that varied only slightly from one query to the next, for example small typographical fixes, were counted as "in common". We also counted a reformulation if a later query in the same session was [google], [yahoo], [bing], or any of a few variants. Sessions were segmented by 30 minutes or more of inactivity.

The *reformulation rate* of a query Q is the proportion of times it is reformulated; the *success rate* the proportion of times it is *not* reformulated. This is a relatively crude measurement, since reformulation might still mean success, and failure might not lead to reformulation, but is a reasonable first-order estimate. Success rate is somewhat skewed (skewness = 2.3) but does vary from query to query, as shown in Figure 2.

Queries for which the standard error on the reformulation rate was greater than 0.05 were removed. This left almost 10 million query instances, with median 2628 instances per query (mean 8914).

Relatively few queries had high reformulation, because the search engine works well for these 994 popular queries, and because reformulation isn't seen in every failure case. Some larger proportion of queries had low (close to zero) reformulation.

Establishing a baseline. Using the relevance labels, we calculated a number of standard metrics for each SERP:

- Precision at 10, using binary gains;
- RR, using binary gains;
- ERR [5], using graded gains;
- SDCG at 10 (that is, DCG@10, scaled by a fixed denominator to lie in the range 0–1 [21, 30]), using graded gains;
- RBP [20], with persistence $\phi = 0.7$ and using graded gains;
- INST [22], with relevance target $T = 1$ and using graded gains.

Binary gains were formed from the graded gains by mapping "poor" to $r_i = 0$, and the other three relevance levels to $r_i = 1$, in keeping with evaluation campaigns such as TREC. Where we used graded gains, the categorical labels were mapped using $r_r \in \{0, 1/4, 1/2, 1\}$, the exponential scheme often used with DCG. The reference point for all results is a baseline that includes "organic" results only—that is, only the conventional "ten blue links" results, each with a title and caption, and each linking to another page on the web.

For each metric we measure how closely the metric predicts searcher satisfaction—assessed, as already described, via the proxy of success (non-reformulation) rate. Spearman's rank correlation ρ between scores and success rate is used to assess their correspondence. That is, we ask of each metric the extent to which increases in score correspond, on average, to more successful query outcomes. If ρ is high, then changes in that metric's scores indicate changes in success, and the metric is a useful proxy of success.

These baseline results are summarised in the left-hand column of Table 1 ("baseline parameters, organic"), measured on a 20% subset of the queries, stratified according to success rate (that is, correlations are reported against queries 1, 6, 13, …when ordered by success rate). The other 80% of the queries are used for training in later experiments, and not used as test data at any stage.[1]

The three shallow metrics—RR, ERR, and INST—correlate best with success, although the correlation is not substantial, and, in the case of INST, not significantly different to zero. The three deeper metrics have negative correlations, a coincidence of the fold examined here, with other folds showing slight positive correlation or no correlation. On average, for these deep metrics, there is little signal either way, and the variation from fold to fold suggests that no general conclusion about system effectiveness should be drawn based on changes in these untuned metrics.

The correlations also vary considerably across parameter choices. Figure 3 shows how ρ changes for precision (P@k) as the cutoff k is varied: the conventional choice $k = 10$ in fact gives the lowest correlation with success rate, and P@1 fares much better. There is also a difference in correlations when shifting from binary gains to ternary gains (with partial credit provided for the two middle relevance labels), and to the full range of graded gains. There is similar variation with other metrics. If metric scores are to be used to predict reformulation behaviour, it is clear that the parameter values must be carefully selected, including both the metric's headline parameter and also the four values that define the gain mapping.

4 TUNING PARAMETERS

We next tuned the metric parameters, to find their best performance. In particular, we tuned:

- for P@k, $k \in \{1, 2, 3, \ldots, 15\}$;

[1] Correlation coefficients for the other four folds for all results shown in Table 1 will be made available via an online appendix after publication. While correlation coefficients do vary somewhat from fold to fold, the trends in Table 1 are consistent in each case.

Metric	Baseline parameters			Tuned parameters			Tuned+card-aware		
	Org.	Org.+ads	All	Org.	Org.+ads	All	Org.	Org.+ads	All
P@k	−0.256	−0.187	−0.110	0.184	0.154	0.196	0.183	0.154	0.220
RR	0.091	0.109	0.140	0.154	0.154	0.197	0.241	0.154	0.221
ERR	0.109	0.114	0.167	0.140	0.144	0.181	0.146	0.161	0.231
RBP	−0.111	−0.011	0.071	0.178	0.161	0.198	0.179	0.151	0.217
SDCG@k	−0.217	−0.114	−0.032	0.184	0.154	0.196	0.183	0.154	0.220
INST	0.049	0.100	0.147	0.086	0.131	0.171	0.172	0.152	0.236
See section	§3	§5	§5	§4	§5	§5	§6	§6	§6

Table 1: Correlations of metric scores with success rate (Spearman's ρ, higher is better). "Baseline" uses standard metric parameters; "tuned" selects them to maximise ρ; and then "Tuned+card-aware" adds card attractiveness labels. Training for tuned varieties was on 80% of queries; ρ is reported over the remaining 20% in all cases. In each case the "org." columns are organic results only ("ten blue links"); "org+ads" includes advertisements; and "all" includes all items on the SERP. Reported correlations all ±approx. 0.06 with 95% confidence.

- for SDCG@k, $k \in \{1, 2, 3, \ldots, 15\}$;
- for RBP, $\phi \in \{0.05, 0.10, 0.15, \ldots, 0.95\}$;
- for INST, $T \in \{1, 2, 3, 4, 5\}$.

Reciprocal rank and ERR do not have parameters.

Gains were also fitted to the four relevance levels. The highest gain was fixed at 1 and the lowest at 0; the other two gains could take any values in $\{0.0, 0.1, 0.2, \ldots, 1.0\}$ subject to the partial ordering being preserved. This has the effect of turning precision into graded precision, although if binary gains gave good correlation this would be allowed. Parameters and gains were fitted to maximise correlation against the per-query success rate. (We note that some of these parameters, for example T, might best be modelled per-query, or even per-user per-query, but we do not attempt this here.) As the optimisation was complex, and in general the surface was not smooth, we used evolutionary algorithms from the `rgenoud` R package [19]. Fitting used an 80% stratified subset, and the reported correlations are for the remaining 20%.

Allowing each metric to align with measured success behaviour makes a substantial improvement (see "tuned parameters, organic" in Table 1). All metrics now positively correlate: that is, a metric improvement does tend to mean an improvement in searcher behaviour. The improvement is small in some cases—for example, INST improves only by 0.037 points—but dramatic in others. Precision and SDCG each move from $\rho \approx -0.2$ to $\rho \approx 0.2$.

Once tuned, several of the metrics have similar correlation at about $\rho = 0.18$, because they have learned equivalent parameters. Precision and SDCG are both best with cutoff $k = 1$, RBP with $\phi = 0.1$, and INST with $T = 1$, which makes them all extremely top-heavy and is not surprising given the situation (web search) and mix of queries (popular, and often navigational). Tuned gains give partial but small credit for the middle two relevance levels: precision and RBP use $r_i \in \{0, 0.2, 0.2, 1\}$, reciprocal rank uses $\{0, 0, 1, 1\}$ (recall that reciprocal rank only uses binary gains), while ERR and INST use $\{0, 0.5, 0.5, 1\}$ and $\{0, 0.6, 0.6, 1\}$ respectively. These suggest that searchers do not make the same fine distinctions between the categories "fair" and "good" that trained judges do.

Although there is good correlation with P@1, there may still be reason to choose another metric. In particular, P@1 and SDCG@1 (which are equivalent) only take four values—or three when the middle two relevance levels are the same. Users may well behave differently as P@1 changes, but it is a metric which is insensitive to most ranking changes. The best trade-off between validity, sensitivity, and other attributes will vary from one evaluation to the next. Regardless, the simple tuning exercise we carried out here gives much better performance across the board.

We must note an important source of uncertainty: metrics were tuned to the success rate of a set of queries, but the relevance labels are only from a single instance. If the SERP varied from instance to instance, even for the same query, different behaviour might emerge, and might not be captured in the labels and hence metrics we have. In practise, this is not likely to be a significant issue as SERPs for popular queries are fairly consistent.

5 CONSIDERING RICH RESULTS

The metrics discussed so far have used only so-called "organic" results, that is, pointers to web pages or the conventional "ten blue links"', and do not take into account the full variety of results searchers see (Figure 1). Including more result types might lead to better correlations. We investigate this possibility in two stages: first including advertisements, then including all cards.

Reading order. Results in a modern SERP do not appear in a list. As well as headers and footers with (for example) pagination and other controls, there are two main areas in current SERP designs: a main sequence or "core" where most results appear, and a "right rail" which is commonly used for summaries of entities and related features. The metrics considered here, however, require a simple list of labels in reading order. If only organic results are considered—which only appear in the core—we can safely assume that searchers read top to bottom. When a right rail is present, that assumption must be revisited. For these experiments we assume the reading order shown in Figure 4: first, the top two cards in the core are examined; then anything in the right-hand rail; then ranks 3 and beyond in the core. This adapts the results of Azzopardi et al. [2], who examine click times and mouse movements.

Advertisements. Ads are not often considered in effectiveness evaluation, perhaps because they do not feature in many search engines, and perhaps because searchers are often considered "blind"

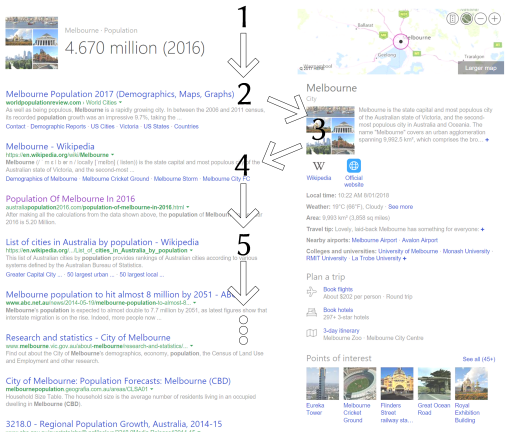

Figure 4: Assumed reading order for rich SERPs. Readers are modelled as looking at the first two cards in the core; then the right rail; then any remaining cards in the main sequence.

to them [23]. Advertisements can clearly change a searcher's experience, however, both because they are in fact attended to [4], and because they can provide links to relevant documents, particularly for navigational or shopping needs.

We collected relevance labels for the documents referenced by advertisements, and had them judged by the trusted crowd workers. The "org+ads" columns in Table 1 summarise the correlations that emerge when the previous methodology is then repeated. For the baseline (untuned) variants, including advertisements is useful across the board, with gains in ρ of up to 0.1 when the metrics are able to "see" the advertisements. As advertisements typically appear before organic results, and the untuned metrics are rather deep, this may also correspond to reduced noise at deeper ranks.

In the tuned case, adding advertisements makes little difference at best and leads to degradations in general, an effect consistent across all folds. The degradation is worst in the shallow metrics (P@1, for example), and occurs because advertisements are more variable than organic results, as they depend not just on topic and a degree of personalisation but also on advertiser changes, auctions, and fairness constraints, all of which enforce variety. These effects lead to more noise at top ranks and in more missed labels, which, in keeping with tradition, we treat as non-relevant. The best-fit metrics are still shallow, but the extra noise hurts correlations.

Adding all cards. The captured SERPs contain almost 90 other kinds of card including specialisations for dates, maps and directions, definitions, stock prices, social media, generic factoids, and many others. These substantially change the searchers' experience, and even though there is often no linked "document", many of these richer cards are designed to efficiently expose information.

Adding all cards to the evaluation leads to marked improvement in ρ (columns "All" in Table 1), and even with default parameters, RR, ERR, and INST now correlate at $\rho \approx 0.15$. With tuned parameters, correlations are around 0.2 for precision, RR, RBP, and SDCG, with ERR and INST at 0.18 and 0.19 respectively. Labelling more kinds of result gives a list which more closely matches what searchers

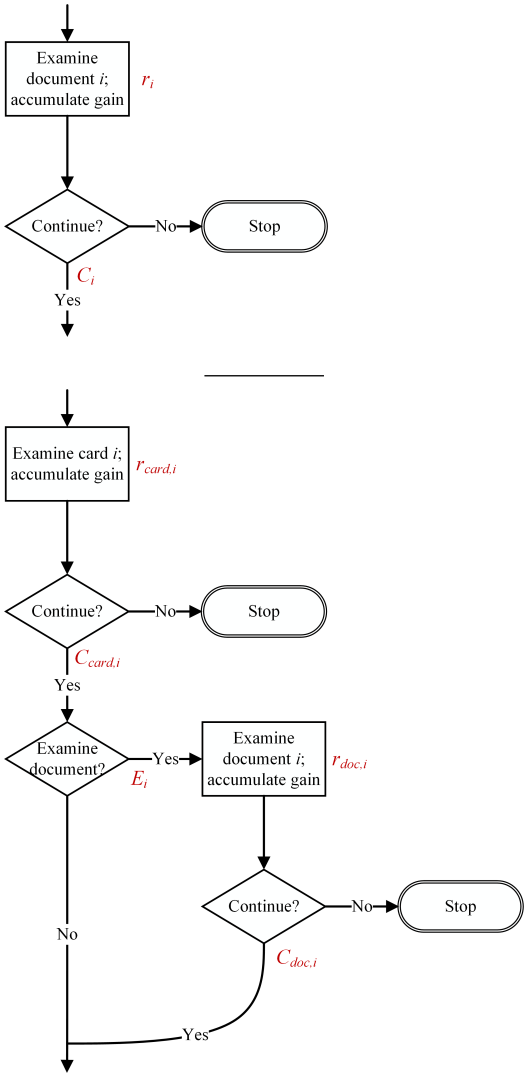

Figure 5: Models of users actions and decisions. Top: the simple model behind weighted-precision metrics, characterised by relevance (gain) vector **r** and continuation vector **C**. Bottom: extended model with cards, examination (clickthrough), and documents.

actually see, and lets metrics align better with searcher behaviour—even in cases where there is no "document" present.

6 CARD-AWARE READING MODELS

Including advertisements and richer cards means metrics consider more results, but they still do not take into account the range of actions possible on a SERP, nor the information in a card itself. In this section we describe a simple technique that allows any weighted precision metric to be made "card-aware".

The upper part of Figure 5 illustrates the user model behind weighted-precision metrics. At each rank i, the user examines a document; accumulates gain r_i; then with probability C_i proceeds to document r_{i+1}, or stops with probability $1 - C_i$.

The lower flowchart extends the model by making three changes: gain can come from cards as well as from documents; not all cards get a click, and so not all documents are examined; and the user might stop reading after a card, not just after a document.

Gain. Conventional metrics assign some degree of gain (or utility) to each document. To account for information in the card itself, we break this in two: $r_{\text{card},i}$ is the gain from the ith card without any clickthrough, and $r_{\text{doc},i}$ the *further* gain from the underlying document should the user click through. We assume summaries are extractive, so any information in the card comes from the document, and so $r_{\text{card},i} + r_{\text{doc},i} \leq 1$. A card and document between them cannot contribute more than a "perfectly informative" document.

We do not make any assumptions regarding the source of $\mathbf{r}_{\text{card}}$ and $\mathbf{r}_{\text{doc}}$. In a typical case they will be assigned by third-party judges after the fact, but they may come from observed behaviours, think-aloud protocols, or anywhere else, and they may be on any scale so long as the inequality above is preserved.

Examination. The searcher's choice to click or not is modelled by $\mathbf{E}$ ("examine"), where E_i represents the chance of them examining document i. Again we make no assumptions about where $\mathbf{E}$ comes from or what it is conditioned on, although it is reasonable to assume it will vary with the gain from, and also the attractiveness of, the card. The combination of $\mathbf{E}$ and $\mathbf{r}_{\text{card}}$ allows two interesting cases to be modelled: good abandonment, when $r_{\text{card},i}$ is high and E_i is zero; and $E_i = 0$ when there is a card but no clickable link.

Inner product form. As can be seen in Figure 5, the extended user model has an extra branch/merge, and two places to stop. To resemble a weighted-precision metric, and be computable with an inner product, the model is reduced to a single stopping point, with a revised continuation function $\mathbf{C}$ and revised gain vector $\mathbf{r}$.

We first define two continuation probabilities for each rank i, taking $C_{\text{card},i}$ to be the chance of continuing past a card, either to read the underlying document or to look at the next card, given that the user is looking at card i; and taking $C_{\text{doc},i}$ to be the chance of continuing, and reading the next card, given the user is looking at document i. Each of these will be defined by the underlying metric via the appropriate function. For example, for P@k both $C_{\text{card},i}$ and $C_{\text{doc},i}$ depend on i, while for RBP both are equal to ϕ.

For a given card/document pair, there are four possible outcomes:

(1) stop after the card, with probability $1 - C_{\text{card},i}$, or
(2) continue past the card ($C_{\text{card},i}$), click (E_i), and stop after the document ($1 - C_{\text{doc},i}$), or
(3) continue past the card ($C_{\text{card},i}$), click (E_i), and continue past the document ($C_{\text{doc},i}$), or
(4) continue past the card ($C_{\text{card},i}$), not click ($1 - E_i$), and continue to the next card.

Combining cases 3 and 4 defines a single continuation function:

$$C_i = C_{\text{card},i} \ \left(E_i \ C_{\text{doc},i} + (1 - E_i)\right) \ ,$$

and a single value for the expected gain from the pair,

$$r_i = r_{\text{card},i} + C_{\text{card},i} \ E_i \ r_{\text{doc},i} \ .$$

From $\mathbf{C}$ we can derive $\mathbf{W}$ as before, to create a card-aware metric. Note that if $\mathbf{r}_{\text{card}} = \mathbf{0}$, $\mathbf{C}_{\text{card}} = \mathbf{1}$, and $\mathbf{E} = \mathbf{1}$, each card-aware metric is identical to the original, and the user models are also similar (searchers get no gain from cards, always click through to read documents, and get gain only from those documents).

Implementation. Algorithm 1 calculates the revised continuation vector $\mathbf{C}$, the revised gain vector $\mathbf{r}$, and the final effectiveness score.[2] It takes as input the gains from cards $\mathbf{r}_{\text{card}}$ and documents $\mathbf{r}_{\text{doc}}$, as well as a vector of examination probabilities $\mathbf{E}$. The underlying metric is represented by its C-FUNCTION.

Line 4 computes the chance of continuing past the card at rank i. At this point the user has accrued gain from previous ranks ($\mathbf{r}$) and from card i ($r_{\text{card},i}$). Line 5 then computes the chance of continuing past the ith document, assuming the user has examined it: here the gain vector includes all previous gain, plus the gain from card i and document i. Line 6 computes the overall probability of continuing past rank i, as above. Finally, line 7 computes the expected gain from rank i, and updates the gain vector. Once $\mathbf{C}$ has been computed, line 8 coverts it to a weight vector in the usual way.

Algorithm 1 can be used with any weighted-precision metric, provided it is defined by a $\mathbf{C}$ vector. For example, AP could also be handled if the rank positions of all cards and documents with non-zero gain were known [21, 22].

Data and method. As before, four-level relevance labels for each page referenced in the SERP are used to give $\mathbf{r}_{\text{doc}}$. A second set of trusted crowd workers assigned labels to each card, choosing between labels for good abandonment (*good abandon*); attractive cards (*would click*); and unattractive cards (*would skip*). These labels were assigned to the whole range of cards on the SERP.

Just as relevance labels are mapped to gain levels, card labels were mapped to $\mathbf{E}$ (probability of click) and $\mathbf{r}_{\text{card}}$ (gain from the card) values, again with ordering contraints to be observed. Then, for each set of gain mappings, the $\mathbf{r}_{\text{doc}}$ vector was decreased by $\mathbf{r}_{\text{card}}$. If the extended model improves correlation, high E_i should match *would click*; as well as high $r_{\text{card},i}$ for *good abandon*. If the extended model does not help, we should see either low $\mathbf{E}$ throughout, or high $\mathbf{E}$ but low $\mathbf{r}_{\text{card}}$.

Results. The final columns of Table 1 ("tuned + card-aware") report the resulting correlations. Once tuned, metrics correlate similarly; again because the best-fit parameters give rise to similar models, with ranks 1 and 2 heavily weighted, and modest gains assigned to the middle two relevance levels. We also see the same improvements from considering all card types.

Further improvements arise from appearance labels and the card-aware user model, although more so when "All" cards are considered. Improvements are more limited when only organic results are considered and the lack of variety is the most likely explanation: organic results typically look reasonable, but do not answer the searcher's need up front, so get labelled *would click* rather than *good abandon* or *skip*, meaning that there is no extra information to improve the metrics.

7 REMARKS AND CONCLUSIONS

Limitations and extensions. There are several limitations in the card-aware reading model presented here, and corresponding possible extensions. First, our formulation assumes that summaries are

[2] Software covering a range of metrics including P@k, RR, ERR, RBP, SDCG@k, and INST will be made available via github.org/Microsoft/irmetrics-r.

Algorithm 1 Calculate card-based metric value from gains $\mathbf{r}_{\text{card}}$ and $\mathbf{r}_{\text{doc}}$, probabilities $\mathbf{E}$, and a metric defined by C-FUNCTION($\mathbf{r}, i$), assuming that $|\mathbf{r}_{\text{card}}| = |\mathbf{E}| = |\mathbf{r}_{\text{doc}}|$, that $\forall i : 0 \le E_i \le 1$, that $\forall i : r_{\text{card},i} + r_{\text{doc},i} \le 1$, and that $\forall \mathbf{r}, i : 0 \le \text{C-FUNCTION}(\mathbf{r}, i) \le 1$.

function CARD-METRIC($\mathbf{r}_{\text{card}}$, $\mathbf{r}_{\text{doc}}$, $\mathbf{E}$, C-FUNCTION)
  $\mathbf{r} \leftarrow \langle\rangle$ ▷ Need to compute r_i, the expected gain at rank i
  for $i \leftarrow 1 \ldots |\mathbf{r}_{\text{doc}}|$ **do**
    $C_{\text{card},i} \leftarrow \text{C-FUNCTION}(\langle \mathbf{r}, r_{\text{card},i} \rangle, i)$ ▷ Continue past this card?
    $C_{\text{doc},i} \leftarrow \text{C-FUNCTION}(\langle \mathbf{r}, r_{\text{card},i} + r_{\text{doc},i} \rangle, i)$ ▷ Continue past this document?
    $C_i \leftarrow C_{\text{card},i}(E_i C_{\text{doc},i} + (1 - E_i))$ ▷ Get probability of continuing past this (card, document) pair
    $\mathbf{r} \leftarrow \langle \mathbf{r}, r_{\text{card},i} + E_i r_{\text{doc},i} \rangle$ ▷ Extend $\mathbf{r}$ vector to include expected gain at rank i
  $\mathbf{W} \leftarrow \text{C-TO-W}(\mathbf{C})$ ▷ Weight vector $\mathbf{W}$ is derived from $\mathbf{C}$ following Moffat et al. [22]
  return $\mathbf{W} \cdot \mathbf{r}$ ▷ Inner product of gain and weight vectors

extracted from documents (where there is a backing document at all), so that $r_{\text{card},i} + C_{\text{card},i} E_i r_{\text{doc},i} \le 1$. If a card adds information not in the document (perhaps annotations regarding the source, or comparisons with other documents) this would not hold. The consequences of allowing gain > 1 vary according to the metric.

Second, the model here does not penalise mismatches. A card that encourages clicks but leads to a useless document is bad for users, wasting their time. To some extent this can be captured in the choice of C function, for example by saying that viewing poor documents is likely to lead to abandonment. If the underlying metric allows it, an alternative would be assigning negative $r_{\text{doc},i}$ in these cases, so expected gain for the document is also negative.

Third is the limitation of weighted-precision metrics: they have a simplistic notion of effort, measuring expected gain per document (card/document pair), but not accounting for features such as reading time. It may be worthwhile to blend the approach here with notions from patch theory [2] or time-biased gain [25].

Conclusions. Modern SERPs and user responses to them have advanced beyond the behavioural models used for traditional offline metrics. In particular, users may get information from the SERP directly; may or may not click through to read each resulting document; and may not even have the option of clicking. If we take reformulation rate as indicator of success, then standard metrics are demonstrably poor at predicting search outcomes.

We can improve this by: tuning the parameters of metrics including the allocation of gain; assigning relevance labels to more classes of object; and building user models which allow for good abandonment, gain from cards, and varying click probability.

Acknowledgments. We thank Bodo von Billerbeck and Alex Moore for their help wrangling data. This work was partially supported by the Australian Research Council (Project DP180102687).

REFERENCES

[1] J. Arguello, W.-C. Wu, D. Kelly, and A. Edwards. Task complexity, vertical display and user interaction in aggregated search. In *Proc. SIGIR*, pages 435–444, 2012.
[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] P. Bailey, N. Craswell, R. W. White, L. Chen, A. Satyanarayana, and S. M. M. Tahaghoghi. Evaluating search systems using result page context. In *Proc. IIiX*, pages 105–114, 2010.
[4] G. Buscher, S. T. Dumais, and E. Cutrell. The good, the bad, and the random: An eye-tracking study of ad quality in web search. In *Proc. SIGIR*, pages 42–49, 2010.
[5] O. Chapelle, D. Metzler, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *Proc. CIKM*, pages 621–630, 2009.
[6] Y. Chen, K. Zhou, Y. Liu, M. Zhang, and S. Ma. Meta-evaluation of online and offline web search evaluation metrics. In *Proc. SIGIR*, pages 15–24, 2017.
[7] L. B. Chilton and J. Teevan. Addressing people's information needs directly in a web search result page. In *Proc. WWW*, pages 27–36, 2011.
[8] C. L. A. Clarke, N. Craswell, I. Soboroff, and A. Ashkan. A comparative analysis of cascade measures for novelty and diversity. In *Proc. WSDM*, pages 75–84, 2011.
[9] A. Hassan, X. Shi, N. Craswell, and B. Ramsey. Beyond clicks: Query reformulation as a predictor of search satisfaction. In *Proc. CIKM*, pages 2019–2028, 2013.
[10] K. Järvelin, S. L. Price, L. M. L. Delcambre, and M. L. Nielsen. Discounted cumulated gain based evaluation of multiple-query IR sessions. In *Proc. ECIR*, pages 4–15, 2008.
[11] J. Jiang and J. Allan. Correlation between system and user metrics in a session. In *Proc. CHIIR*, pages 285–288, 2016.
[12] D. Kelly. Methods for evaluating interactive information retrieval systems with users. *Found. & Trends in IR*, 3(1-2):1–224, 2009.
[13] M. Khabsa, A. Crook, A. H. Awadallah, I. Zitouni, T. Anastasakos, and K. Williams. Learning to account for good abandonment in search success metrics. In *Proc. CIKM*, pages 1893–1896, 2016.
[14] J. Kim, G. Kazai, and I. Zitouni. Relevance dimensions in preference-based IR evaluation. In *Proc. SIGIR*, pages 913–916, 2013.
[15] J. Li, S. Huffman, and A. Tokuda. Good abandonment in mobile and PC internet search. In *Proc. SIGIR*, pages 43–50, 2009.
[16] C. Luo, Y. Liu, T. Sakai, F. Zhang, M. Zhang, and S. Ma. Evaluating mobile search with height-biased gain. In *Proc. SIGIR*, pages 435–444, 2017.
[17] J. Mao, Y. Liu, K. Zhou, J.-Y. Nie, J. Song, M. Zhang, S. Ma, J. Sun, and H. Luo. When does relevance mean usefulness and user satisfaction in web search? In *Proc. SIGIR*, pages 463–472, 2016.
[18] I. Markov, E. Kharitonov, V. Nikulin, P. Serdyukov, M. de Rijke, and F. Crestani. Vertical-aware click model-based effectiveness metrics. In *Proc. CIKM*, pages 1867–1870, 2014.
[19] W. R. Mebane Jr and J. S. Sekhon. Genetic optimization using derivatives: The rgenoud package for R. *J. Stat. Soft.*, 42(11):1–26, 2011.
[20] A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Sys.*, 27(1):2.1–2.27, 2008.
[21] A. Moffat, P. Thomas, and F. Scholer. Users versus models: What observation tells us about effectiveness metrics. In *Proc. CIKM*, pages 659–668, 2013.
[22] 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.
[23] J. W. Owens, B. S. Chaparro, and E. M. Palmer. Text advertising blindness: The new banner blindness? *J. Usability Stud.*, 6(3):172–197, 2011.
[24] M. Shokouhi and Q. Guo. From queries to cards: Re-ranking proactive card recommendations based on reactive search history. In *Proc. SIGIR*, pages 695–704, 2015.
[25] M. D. Smucker and C. L. A. Clarke. Time-based calibration of effectiveness measures. In *Proc. SIGIR*, pages 95–104, 2012.
[26] A. Turpin, F. Scholer, K. Järvelin, M. Wu, and J. S. Culpepper. Including summaries in system evaluation. In *Proc. SIGIR*, pages 508–515, 2009.
[27] C. Wang, Y. Liu, M. Zhang, S. Ma, M. Zheng, J. Qian, and K. Zhang. Incorporating vertical results into search click models. In *Proc. SIGIR*, pages 503–512, 2013.
[28] E. Yilmaz, M. Shokouhi, N. Craswell, and S. E. Robertson. Expected browsing utility for web search evaluation. In *Proc. CIKM*, pages 1561–1564, 2010.
[29] Y. Zhang and C. Zhai. A sequential decision formulation of the interface card model for interactive IR. In *Proc. SIGIR*, pages 85–94, 2016.
[30] Y. Zhang, L. A. F. Park, and A. Moffat. Click-based evidence for decaying weight distributions in search effectiveness metrics. *Inf. Retr.*, 13(1):46–69, 2010.