Measuring the Utility of Search Engine Result Pages
An Information Foraging Based Measure

Leif Azzopardi  
University of Strathclyde  
Glasgow, UK  
lefos@acm.org

Paul Thomas  
Microsoft  
Canberra, Australia  
pathom@microsoft.com

Nick Craswell  
Microsoft  
Bellevue, WA, USA  
nickcr@microsoft.com

ABSTRACT
Web Search Engine Result Pages (SERPs) are complex responses to queries, containing many heterogeneous result elements (web results, advertisements, and specialised "answers") positioned in a variety of layouts. This poses numerous challenges when trying to measure the quality of a SERP because standard measures were designed for homogeneous ranked lists.

In this paper, we aim to measure the utility and cost of SERPs. To ground this work we adopt the C/W/L framework which enables a direct comparison between different measures in the same units of measurement, i.e. expected (total) utility and cost. Within this framework, we propose a new measure based on information foraging theory, which can account for the heterogeneity of elements, through different costs, and which naturally motivates the development of a user stopping model that adapts behaviour depending on the rate of gain. This directly connects models of how people search with how we measure search, providing a number of new dimensions in which to investigate and evaluate user behaviour and performance.

We perform an analysis over 1000 popular queries issued to a major search engine, and report the aggregate utility experienced by users over time. Then in a comparison against common measures, we show that the proposed foraging based measure provides a more accurate reflection of the utility and of observed behaviours (stopping rank and time spent).

CCS CONCEPTS
• Information systems → Retrieval effectiveness; • Human-centered computing → User models;

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SIGIR'18, July 8-12, 2018, Ann Arbor, MI, USA  
© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-1-4503-5657-2/18/07 ... $15.00  
https://doi.org/10.1145/3209978.3210027

1 INTRODUCTION
Most Information Retrieval (IR) evaluation measures focus on estimating the quality of a ranked list of results, where each result is a simple link to another web page [32]. However, modern web search engine result pages (SERPs) are complex, composite, responses with curated and computationally selected elements, consisting of algorithmic web results, advertisements and a variety of answer cards. Furthermore, result elements are positioned in different layouts on the SERP: e.g. in the header, left rail, core, right rail, or footer. Consequently, the assumptions implicit in many retrieval measures no longer hold in the context of evaluating modern SERPs [2]. While there has been a number of studies investigating which specialist answers (“verticals”) to include [1, 3, 39], which verticals are preferred [2], and how they affect search behaviour and satisfaction [11, 23], in this work we attempt to directly measure the quality of a whole SERP. To do so a number of sub-questions first need to be addressed: (i) what are the different elements on a SERP, (ii) in what order are the elements examined, (iii) what is the cost of inspecting different elements, (iv) what is the benefit of those elements, and ultimately, (v) what is the expected (total) utility of a SERP?

To start addressing these questions, we studied SERPs from a major web search engine analysing the different types of elements shown for one thousand popular queries, and evaluated different possible “orderings” given the layout of the SERPs. From this analysis, we were then able to estimate the time spent per element type, which was used to estimate the cost of processing SERPs. Given relevance assessments for the elements appearing on the top thousand queries, we then inferred an aggregate utility curve experienced over time—that is, the total utility experienced by the population of users given the cost they incurred (time spent). We argue that a “good” measure will approximate observed behaviours (e.g. time spent and stopping rank) and inferred utility.

Since different measures provide different perspectives, and are in different units, it is not possible to directly compare them to the inferred utility curves. To address this problem, this paper draws upon the recent theoretical developments regarding the measurement and modelling of IR systems [9, 25, 26] where measures can be expressed as either the expected utility (EU) or expected total utility (ETU) [9], depending on how the stopping/continuation function that defines the user stopping model is expressed [25, 26]. We then draw upon Information Foraging Theory [30] to develop a forager based user stopping model that adapts its stopping behaviour based on the gain accrued during the search process: directly connecting the theory of how we model people’s search behaviour with how we measure it. We show that our forager-based user model provides a better approximation of the utility experienced than do other common models.
2 BACKGROUND

Evaluation has played a central role in the development of IR systems. Over the years, increasingly sophisticated measures have been developed to use test collections—document collections labelled for relevance—to approximate the system’s performance and the user’s search satisfaction (see e.g. Sanderson [32] for an excellent overview). Measures have evolved from precision- and recall-based to utility- and cost-based—with more focus being directed towards how the user interacts with the search results, what they gain and at what cost. In conjunction with these developments numerous efforts have shown how the different measures are mathematically related [9, 18, 26, 38], and how they can be housed within a utility based framework [25]. This is an important development, meaning that we are in the process of standardising the units of measurement, and that it is possible to compare measurements directly (which we capitalised on in this work). Also, the utility based framework naturally connects with Information Foraging Theory, which we are in the process of standardising the units of measurement, Carterette shows that different measures can be formulated in this framework, depending on how the different models are instantiated. Relevant to the present work are “model 1” measures such as Carterette’s interpretation of RBP that estimate the expected utility, and “model 2” measures such as DCG that estimate the expected total utility. The difference is whether the utility is either extracted from one of the documents, or from all the documents found. Put another way, model 1 measures assume the user gets value from only one document, while model 2 measures assumes the user gets some value from all documents. In terms of estimating the utility of a SERP, model 1 measures are likely to underestimate the actual utility (unless there is only one click), while model 2 measures are likely to overestimate actual utility (unless a user clicks on everything).

C/W/L. Moffat et al. [26] further formalize the relationship between measures, such that the Expected Utility (EU) of any arbitrary “weighted precision” measure can be generalized as:

$$EU = M(r) = \sum_{i=1,\ldots,\infty} W_M i, r_i$$

where $r$ is the relevance (gain) vector for each rank $i$, and $W_M$ is a metric specific weighting vector. $W_M i$ can be interpreted as the expected proportion of attention a user gives to rank $i$. For example, for precision at rank 5, the weight vector would be $W_{PR@5} = (0.2, 0.2, 0.2, 0.2, 0.2, \ldots)$, while for RBP with persistence parameter $\phi$ we have $W_{RBP, i} = (1 - \phi)\phi^{i-1}$. $M(r)$ is thus the expected utility per document inspected.

The $W$ vector can be interpreted in the model shown in Figure 1. In this model, a user reads the document at rank $i$; accumulates some gain, e.g. $r_i$; then chooses either to continue to rank $i+1$, or stop. This decision can be captured in a vector $C$, for continue. The conditional probability of continuing past rank $i$, $C_i$, directly relates to the weight vector such that:

$$C_i = \frac{W_{i+1}}{W_i}.$$  

This continuation probability is easy to interpret and to reason about. For example, $C_{RBP}$ is the constant $\phi$; $C_{PR@k}$ is 1 for ranks $1 \ldots k - 1$, and 0 thereafter; and $C_{RBP}$ is 1 when $r_i = 0$, and 0 when $r_i = 1$. Finally we can derive the probability that the $i$th document in the ranking is the last one observed by the user with $L_i$:

$$L_i = \left( \prod_{j=1,\ldots,i-1} C_j \right) (1 - C_i) = \frac{W_i - W_{i+1}}{W_i}.$$  

Later, we will use $L_i$ in our evaluations to predict the stopping rank. Here we also note how the C/W/L Framework can be extended to estimate the Expected Total Utility (ETU) using $L$ as follows:

$$ETU = M_{Total}(r) = \sum_i L_i \sum_{j=1,\ldots,i} r_j$$

that is, the sum over all ranks of the gain accrued by reading that far, times the probability that this is where they will stop. The distributions defined by C, W, and L are mathematically related, and can be defined to instantiate a variety of measures including precision, RBP, RR, and AP [25]. In this paper, we build our measure

---

1We will use “gain”, “benefit” and “utility” interchangeably, depending on the context.

2In what follows, for simplicity we drop the subscript $M$ unless context requires it.
within the utility based C/W/L framework—because it: (i) provides a way to represent the different user stopping models of different measures and (ii) means we can directly compare measurements in the same units, i.e. compare expected (total) utility.

**Effort and Time.** The next major evolution in IR measures has been the introduction of costs—as it has been argued and shown that the effort required and/or time spent during the search process affects user interaction and search satisfaction [5, 18, 30, 34, 35, 37]. Smucker and Clarke [34] formalize the idea within Time Biased Gain to create a measure that reports the amount of (discounted) gain experienced over time (i.e. the rate of gain). They introduce a reading model where the time taken to reach a particular rank i is based upon the cost of reading result elements and the cost of reading each document up to and including i. Of note is that the time spent on result elements in this model is fixed, while the time spent reading a document is proportional to its length.

Rather than focus on time, Sakai and Dou [31] present a related measure based on text trails. Again a reading model is based upon an exponential decay function, such that as user reads through text they are less and less likely to continue to the end of the document. The U-measure is essentially the rate of gain over that text that gets read. More recently, Jiang and Allan [18] propose measures that also consider the costs (as approximated by time), by calculating the ratio of gain to cost. They show that by including the cost, a higher, but only moderate, correlation to session based user satisfaction can be achieved. While these measures can be seen as different ways to consider the cost within the evaluation measure, they assume that documents are homogeneous (e.g. all news documents), and that the benefit is spread across the text. However, in the context of measuring a SERP, which is composed of heterogeneous elements (text, images, etc.) which link to pages and other resources, the time spent on landing pages is quite varied and dependent on the result element in question (i.e. news, video, game, homepage, etc.). Consequently, in this work, we focus solely on evaluating the SERP elements, the time taken to process those elements, and the value that each element adds to the SERP—in order to predict the utility and cost of the SERP itself.

**Adaptive and Constrained Models.** More recently proposed IR measures include adaptive (also referred to as dynamic) user stopping models as well as incorporate constraints. Rather than the naive assumption that users simply walk down a ranked list according to some pre-defined and fixed probability distribution, adaptive measures consider what has been encountered so far, what is desired and what are the constraints [18, 24, 26, 38]. For example, the key idea in the INST measure [24] is that the user has some idea of the number of documents that they want, which can be considered a soft constraint. As they examine documents, and find relevant documents, then the probability of stopping increases as they get closer to their target T. Within the C/W/L framework, Moffat et al. encode this idea with a new continuation function C that depends on T, i, and the number of relevant documents found. A further development by Moffat and Wicaksono [27] allows for “egregiously non-relevant” results, which reduce the chance of further interaction. This allows more nuance than simply modelling zero gain, and in simulations it gives more plausible measurement in the presence of notably poor-quality documents.

In the Bejeweled player model [38], a similar approach is taken, but where both cost and gain constraints are imposed—which influences the user stopping probability. The proposed measures are described in a manner similar to the C/W/L framework, and the authors show that common IR measures can be derived using a common framework. They propose two variants: (i) a static measure, where cost and gain are fixed constraints, and (ii) an adaptive measure, where when a user encounters relevant documents, their desire for relevant documents increases (i.e. T increases), and so too does their willingness to spend more time; while if they encounter non-relevant documents, their desire for relevance and tolerated cost both decrease. While this stopping intuition may be relevant in the context of a game, it is less intuitive for certain search tasks. Zhang et al. [38] show that their measures correlate to session based satisfaction substantially higher for informational queries, than for navigational queries, but their adaptive measures only provided marginally better correlations to user satisfaction judgments.

**Session Based.** Another innovation in the evaluation of search systems has been the development of session based measures which consider performance over a series of related queries [e.g. 4, 17, 20, 22]. Typically, standard measures are linearly combined with a discount to formulate an overall measure. For example, one of the first session based measures proposed was Session DCG [17], where an exponent is introduced to discount subsequent queries, such that if the user finds a relevant document on the nth query, it is considered less valuable than if they found it at the (n – 1)th query, assuming it was seen at the same rank.

In this paper, we will focus on evaluating individual queries, and leave session based adaptations for future work. Instead, we explain how the different constraints on searching and the adaptive behaviours which have been introduced in the various measures naturally arises within Information Foraging Theory, and how they can be encoded within the user stopping model, to define the basis of a family of new theoretically underpinned IR measures.

### 3 A FORAGING BASED STOPPING MODEL

Information Foraging Theory (IFT) [30] models how people search for information by applying ideas from optimal foraging theory. IFT assumes that when searching for information, people adopt instinctive foraging mechanisms that evolved to help our ancestors find food. IFT has been widely used within interactive information retrieval [e.g. 7, 29, 33, 36] and provides an intuitive and formal way to describe and predict search behaviour [7, 30].

A central component of IFT is the Information Patch model, which considers how long a forager ought to stay in a “patch”—here, a SERP—before moving on. The model predicts that a forager...
will move to a new patch when the rate of gain from the current patch falls below some tolerated rate. This is Charnov’s Marginal Value Theorem [10]. The intuition is as follows: if the yield from the current patch is lower than what the forager could obtain elsewhere, on average, then they should move on. However, if the current patch is yielding a higher rate of gain than average, then the forager should keep exploiting the current patch. Of course, this is subject to how much they require, and subject to how much time they have available to forage. Given the patch model from IFT, we formalize a foraging based user stopping model within the C/W/L framework—connecting the theory on how we model people’s search behaviour with how we measure search performance.

We first assume that a searcher wants to consume a certain amount of gain and that once they reach that desired level, they are more likely to stop (e.g. the model is goal-sensitive). However, second, they won’t naïvely continue trying to reach this goal; instead we assume that searcher will only continue exploiting a patch so long as the rate of gain is sufficiently high (so the model is rate-sensitive: see Figure 2). This second rule restates Charnov’s Marginal Value Theorem, while the first imposes a sufficiency/satiation constraint.

To encode these conditions within the continuation function, we use sigmoid functions to provide a probability distribution. This gives us enough flexibility to represent a number of possible stopping behaviours, including a number of existing measures.

Goal Sensitive. We model the first condition with C1, and say that the chance of continuing decreases as gain accumulates:

\[
C_{1i} = 1 - \left( 1 + b_1 \cdot e^{(T - y_i) R_1} \right)^{-1}
\]  

where \(T\) is the target and \(y_i\) is the gain accrued so far (i.e. \(y_i = \sum_j r_j\)). A rational searcher with target \(T\) in mind would continue whenever \(T > y_i\) and stop as soon as \(T \leq y_i\), assuming the other conditions are not violated. However, we can imagine that there might be some uncertainty regarding \(T\) or \(y_i\), and so the searcher may stop earlier, or continue longer. So, we include a “rationality” parameter \(R_1\), such that: when \(R_1 = \infty\) we have the “perfectly rational” user model, and at \(R_1 = 0\) we have an “agnostic” user model, where the gain accumulated makes no difference to behaviour (and thus C1 doesn’t depend on \(y_i\) or \(T\)). When \(T = y_i\), i.e. when the target is reached, \(C_1 = 1 - \frac{1}{1 + e^{R_1}}\) and therefore we can use \(b_1\) to model the probability of stopping when the searcher has found enough. Note, the \(1 - (\ldots)\) in Eq. 5 turns the stopping function into a continue function, as needed by the C/W/L framework.

Rate Sensitive. We model the second constraint with C2, and say that the chance of continuing decreases as the rate of gain decreases:

\[
C_{2i} = \left( 1 + b_2 \cdot e^{(A - \frac{y_i}{R_2}) R_2} \right)^{-1}
\]

where \(A\) is the tolerated rate of gain, \(y_i\) is the cost so far (i.e. \(y_i = \sum_j t_j\), where \(t_j\) is time spent on element \(i\)), \(y_i\) is the gain so far, and thus \(y_i / y_i\) is the rate of gain so far. As above, a rational searcher who will tolerate a rate of gain of \(A\) would continue only if \(y_i / y_i > A\); otherwise they would stop. However, again, we can imagine that there would be some uncertainty regarding the rate of gain or the tolerated rate of gain, and so we also include a rationality parameter \(R_2\), where \(R_2 = \infty\) implies the searcher is perfectly rational and \(R_2 = 0\) implies the user is agnostic of the rate of gain. \(b_2\) operates in the same way as \(b_1\), changing the continue probability when the rate of gain is exactly at the searcher’s threshold.

To illustrate how the parameters of the sigmoid function affect the probability of continuing, we have plotted a number of examples of the C2 function in Figure 3. From the plots, as rationality \(R_2\) is increased (left to right), the probability of continuing depends more upon the rate of gain. When \(R_2 = 0\), C2 is a constant, whereas when \(R_2 = 100\) C2 is close to a step function. As \(A\), the tolerated rate of gain, increases (top to bottom), the threshold for the step function or the centre of the curve also increases. The third parameter, \(b_2\), shifts the curves up and down: shown here is \(b_2 = 1\), meaning that when the user is exactly at \(A\) they have a 50% chance of continuing (intercept shown with the vertical dashed lines). As \(b_2\) tends to \(\infty\) then the chance of continuing tends to 0%, while if \(b_2\) tends to zero, then the chance of continuing tends to 100%. \(C_1\) is controlled in the same way as \(C_2\), but has opposite shape: as total gain increases, the probability of continuing decreases.

IFT Continuation Function. Given the two conditions C1 and C2 we can formulate the IFT continuation function: \(C_{IFT,i} = C_1 \cdot C_2\), where we assume that the conditions are independent. Given the combined continuation function, it can be embedded within the C/W/L framework to provide a measure of search performance.

The IFT continuation function has the flexibility to adapt to various user behaviours—and certain edge cases result in existing measures. For example, when \(R_1 = \infty, T = 1\) in C1, and C2 is set to one (i.e. \(A = 0\) and \(b_2 = 0\)) we have a user who will stop exactly when they have accumulated one unit of gain, but will tolerate any amount of cost: this is reciprocal rank. With \(R_1 = R_2 = 0\) the continuation function is controlled only by \(b_1\) and \(b_2\), and is a constant, which therefore models rank biased precision. If instead, we set the parameters of \(C_2\) such that it equals one, and so only the goal sensitive condition (C1) is in effect, then the measure behaves like INST. Finally, the Bejewelled player model can be indirectly modelled by setting the tolerated rate of gain (\(A\)), which is based on the gain divided by the cost, and the target (\(T\)). Alternatively, an extra condition could be added to represent cost sensitivity that encodes the maximum cost the searcher is willing to incur (\(K\)), such that they continue if \(K > \kappa_i\). This is left for further work.
While these queries are predominately head queries and naviga-
tional in nature, it provides us with sufficient data to provide robust
measures, we selected $b_1 = b_2 = 0.25$, because if $R$ was set to zero,
the continuation function would approximate RBP close to $\phi = 0.1$,
the best tuned setting—and so by setting the other parameters, we
can see how the IFT user model improves over RBP (or not). We
set $R_1 = R_2 = 10$, as a middle ground between being agnostic (i.e.
$R = 0$) and being perfectly rational ($R = \infty$). Finally, we set $T = 0.2$
and $A = 0.1$ to simulate a casual web searcher, who is looking for a
reasonable (not bad) page, and is willing to examine a few items. We
shall denote this as the IFT user model. To evaluate the influence of
each of the continuation functions independently, we also evaluate
IFT-C1 and IFT-C2, which have the same parameters for C1 and C2
as above, but where either $C_2$ or $C_1$, respectively, is held constant.
We leave parameter estimation and tuning for future exploration.
Figure 4: A SERP is typically composed of four sections: header, footer, core, and right rail. Elements of different kinds are found in each section.

An implementation of the IFT measure and the other measures is available at https://github.com/Microsoft/irmetrics-r.

5 SERP COMPOSITION AND LAYOUT

Most effectiveness metrics, including all those used here, assume results (or SERP elements) are read strictly in order; but a modern SERP has a complex two-dimensional layout where the "right" order is not obvious. Figure 4 displays a typical layout adopted by major search engines, where the SERP consists of:

- a **header**, where the query box, and query statistics are displayed,
- the **core**, where the main set of algorithmic results are shown along with advertisements and other answers e.g. navigational entity cards, image, video and news elements, etc.,
- often, a **right rail**, where entity cards, advertisements, related searches, etc. are shown, and,
- a **footer**, with navigational cues to the next or previous page.

Typically, algorithmic web results are only shown in core, while advertisements typically appear at the top of core, bottom of core or in the right rail, though they sometimes appear elsewhere. News, video, image and other answer elements typically only appear in core, after any advertisements, and possibly after some algorithmic results. Related searches typically appear at the bottom of core or bottom of the right rail. In the following subsections, we first extract the different elements, then infer a reading order, before estimating the time spent per element.

### Table 1: Approximate breakdown of element types appearing on pages in the sample of popular queries.

<table>
<thead>
<tr>
<th>Element type</th>
<th>% of pages</th>
<th>% of elements</th>
<th>Appears</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithmic web results</td>
<td>100%</td>
<td>64.3%</td>
<td>Core</td>
</tr>
<tr>
<td>Advertisements</td>
<td>43%</td>
<td>8%</td>
<td>Both</td>
</tr>
<tr>
<td>News</td>
<td>70%</td>
<td>6%</td>
<td>Core</td>
</tr>
<tr>
<td>Query suggestions</td>
<td>44%</td>
<td>3%</td>
<td>Core</td>
</tr>
<tr>
<td>Images</td>
<td>2%</td>
<td>&lt;1%</td>
<td>Core</td>
</tr>
<tr>
<td>Videos</td>
<td>15%</td>
<td>1%</td>
<td>Core</td>
</tr>
<tr>
<td>Entity cards</td>
<td>89%</td>
<td>7%</td>
<td>Both</td>
</tr>
<tr>
<td>Result disambiguation</td>
<td>4%</td>
<td>&lt;1%</td>
<td>Right</td>
</tr>
<tr>
<td>Stock</td>
<td>12%</td>
<td>1%</td>
<td>Core</td>
</tr>
<tr>
<td>Other</td>
<td>100%</td>
<td>9%</td>
<td>Both</td>
</tr>
</tbody>
</table>

5.1 Page Elements

Within the set of elements extracted, we observed over 100 different types of elements—which types depend upon the query. For example, for the query "Walmart", a SERP might include an element with links to Walmart’s website and within its site, another element may show a map to nearby branches, etc. However, for the query “typing test”, the result list includes a specialized answer element with an embedded typing test. Rather than modelling each element type individually, we considered the nine most popular element types and assigned the rest to an "other" element category. The resulting types were algorithmic web results, advertisements, and several kinds of answers: news, query suggestions, images, entity lookups, result disambiguation, video, and stocks. For each element on each page, we recorded whether it appeared in the core or the right rail, and the rank where it appeared.

Table 1 provides an overview of the percentage of pages that contained each type, the percentage of the elements of each type, and where they appeared on the page. As expected, algorithmic web results appeared on all pages, with entities, news, query suggestions, and advertisements being the next most prevalent. Other elements appeared depending on the query (as mentioned above). The number of elements per page varied from 2 to 60, with mean 12.8 and median 12 elements per page. Sixty elements on a page may seem unlikely, but certain very popular queries cover many intents: for example “Disney” covers holidays to Disney resorts, movies and TV shows, shopping, characters, news, social media streams, and so on as well as a very varied audience. In contrast, a query such as “Bank of America”, typically houses the median number of elements. This breakdown shows how the composition of pages varies over the sampled popular queries—now given these elements, we turn our attention to approximating the order in which they are inspected.

5.2 Element Ordering

To apply current measures, we need to determine the order in which a user examines the elements housed on the SERP. Prior research has shown that the order of inspection can vary depending on various factors such as the layout, the attractiveness of elements, etc. [1, 12, 21]. However eye tracking studies have shown that the typical gaze distribution (scan path) is characterised by the "golden triangle" or "F-shaped pattern" [8, 21], where users examine result
Given the SERP is essentially composed of two lists (the core list and the right rail list), there are various ways in which we can obtain an ordering. For example, we could assume that users will examine all the core elements, then examine all the right rail elements, or vice versa. On the other hand, a user may start with either the core or right rail, and then alternate back and forth examining one or more from each; or may examine them in some other way altogether. Given past work, users tend to follows a top-down, left-to-right, examination pattern, where they start examining the core list first, then the right rail, then back to the core, and again to the right, and back to the core, creating the “F-shaped” pattern (see Fig. 5). Put more generally, we assume that a user examines \( n_{cf} \) elements from the core first, then \( n_{rf} \) elements from the right rail, and then adopts an interleaving view order by examining \( n_{cn} \) elements from core, and \( n_{rn} \) elements from the right (and so on, until all elements are examined). Typically, the number of elements displayed on the right is less than the number of elements in the core, so at some point the user would continue down the rest of core (with some probability according to \( C \)). By setting \( n_{cf}, n_{rf}, n_{cn}, n_{rn} \) we can define a variety of examination orders. Figure 5 shows two such orderings where users, either examine one element from each and then repeat (i.e. ordering (b) in Table 2), or examine two from core, and one from right rail, and then repeat (i.e. ordering (f) in Table 2).

To evaluate the different orderings produced, we plotted the probability of clicking on the element versus its estimated position, and fit an exponential function to the distribution to see which sequence produced an ordering most consistent with the click through data. Table 2 shows the different sequences examined, along with the Pearson’s correlation coefficient based on a least squares regression. We also tried modelling the case where a user “looks ahead”—i.e. inspects one or two elements past that which was clicked—and cases with 3 or more elements in a block (e.g. \( n_{cf} \geq 3 \), etc.), but these resulted in poorer overall fits. Not surprisingly, the least intuitive sequence, \( (a) \), which assumes the right rail element is inspected first, obtains the lowest correlation, \( r^2 = 0.669 \), while the best correlation was given by ordering \( (f) \), \( r^2 = 0.738 \). For the remainder of this paper, we use \( (f) \) to produce the ordering of results elements, and leave other orderings and more sophisticated models to future work.

### 5.3 Time Spent per Result Element

To determine the rate of gain for our C2 function, and estimate the total time spent on a SERP given each user model, we first needed to calculate the time a user spends on each element. To do this, we constructed a linear model that estimated the time spent on the SERP, given all the elements up to and including the element clicked. As independent variables to the model, we provided the number of times each element type appeared in the core and in the right rail, and the dependent variable was the time spent on the SERP before the click.
Table 3: Estimated time to read a result, by type and position.
Times relative to a standard algorithmic result in the core.
All estimates significant at $p < 0.0001$.

<table>
<thead>
<tr>
<th>Type</th>
<th>Position</th>
<th>Core</th>
<th>Right Rail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td></td>
<td>3.65</td>
<td></td>
</tr>
<tr>
<td>Algorithmic web results</td>
<td>1.00</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Advertisements</td>
<td>1.49</td>
<td>0.30</td>
<td>-</td>
</tr>
<tr>
<td>News</td>
<td>5.62</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Query suggest.</td>
<td>1.41</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Images</td>
<td>0.96</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Videos</td>
<td>3.91</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Entity cards</td>
<td>8.91</td>
<td>0.45</td>
<td>-</td>
</tr>
<tr>
<td>Result disambig.</td>
<td>-</td>
<td>1.81</td>
<td>-</td>
</tr>
<tr>
<td>Stock</td>
<td>0.97</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Other</td>
<td>3.22</td>
<td>0.96</td>
<td></td>
</tr>
</tbody>
</table>

Table 3 provides the estimated times for each element, where statistical testing indicates that all effects were significant ($p < 0.0001$, ANOVA $F$ test). Due to the commercial sensitivity of this data, we have reported the times relative to average time taken to process one algorithmic web result element.

These results for the first time show the relative times taken to process the different element types on a SERP. Advertisements in the core section take slightly longer to process than a web result (49% longer), while advertisements in the right rail attract much less attention and take 70% less time, on average. Images and stock tickers cost a similar amount of time as a web result, but video, news and entity cards (in the core) take much longer, at 3.91×, 5.62×, and 8.91× respectively. This is not too surprising as these elements tend to be much larger, as well, as more complex, mixing text, graphics and even interactive elements. However, when entity cards are on the right rail, they attract little attention, on average, and only take about 45% the time of a web result.

6 MEASURES, MODELS AND BEHAVIOURS
With models of reading order, and estimates of the cost as well as gain from examining each element, we are able to compare measures on modern web SERPs.

6.1 Comparing Measures
One limitation of using web search log data is that user satisfaction judgements are not available. Rather than trying to predict user satisfaction, we focus our evaluation on how well each measure predicts observables. Thus, to evaluate the quality of each of the measures, we focused on three aspects: (i) how well the measures estimate the user’s actual stopping rank (last element clicked), (ii) how well the measures estimate the inferred utility curves, and (iii) how well they estimate the time spent on the SERP.

Stopping Rank. Recall that the $L$ vector is a probability distribution modelling the probability of stopping at a particular rank. For each measure and impression we can easily derive the likelihood of a measure’s user model predicting they stop at rank $i$: it is simply $L_i$. For example, under RBP with $\phi = 0.1$, $L_1 = 0.9 \times 0.1^{i-1}$. The likelihood of this model, given we observed a searcher stopping at rank 1, is 0.9; while as the likelihood of this model if we observe a searcher stopping at rank 3 is only 0.009. For each impression, given the the last click, we can calculate the likelihood of the actual stopping rank for each measure. We report the mean likelihood over all impressions.

Utility. To infer the utility experienced by users, we assumed that a user only receives benefit from a element that that they clicked on. Thus, given the graded relevance labels (which approximate how useful/valuable the element is, if selected), and the empirical click data we can infer at each rank the total empirical utility experienced (referred to as inferred utility or inferred gain). This provides a picture of how useful the different elements were to the user for each impression. We then compared these inferred utility curves against the expected (and estimated) utility given each of our measures and report the mean absolute error between the estimated and observed.

Cost. Given the C/W/L framework, we can also estimate the expected total cost by replacing the relevance/gain vector ($r$) with the cost vector ($k$) as follows:

$$ETC = \sum_i L_i \sum_{j=1}^{\ldots i} k_j$$

where $k$ provides the times to process each element on the SERP. We derived these times in Section 5.3. We will thus be estimating the time spent on the page, but will refer more generally to this as the cost [c.f. 6, 18, 19, 38]. We compared the expected total time given each measure against the actual time spent on the SERP, and again report the mean absolute error.

6.2 Results
All else being equal, we should prefer a measure whose underlying user model matches searchers’ actual behaviour and enables us to estimate observables such as where the user will stop on the SERP, how long they will spend on the SERP, and (inferred from clicks and judgements) how much gain they will accrue from the SERP (i.e. the inferred utility or inferred gain).

Before inspecting the Tables of results, Figure 6 provides examples of the (expected and inferred) total utility over time for two individual query impressions, and for the average over 5,000 query impressions.

In the left plot, we show an impression with only a click at rank 1, which is highly relevant. The searcher accumulates one unit of gain in one unit of time, but no more after this (solid red line). RR and P@1 model this accurately, but P@10, which assumes the searcher will examine all of the top 10 result elements, grossly overestimates the total gain. RBP 0.7 also overestimates the total gain, but not to the same extent. IFT provides a closer estimate—because the goal condition is met, early on, and so the searcher is more likely to stop sooner—leading to a better approximation.

The middle plot of Figure 6 shows an impression with three clicks, at ranks 1–3; only the first two are relevant, but there are other relevant elements deeper in the listing. Since the first item was clicked, all measures are accurate initially. By the end of the interactions, RR (which only considers the first relevant item) and P@1 (which only considers rank 1) underestimate total gain. RBP and P@10 overestimate the total gain, P@10 grossly so. IFT goes to
Figure 6: Examples of total gain over time for several measures, plot against the searcher’s accumulated (inferred) gain based on observed clicks. Left: an impression with a single click at rank 1. Centre: an impression with three clicks, not all of which lead to gain. Right: the total gain over time, averaged over a random sample of 5,000 impressions. In the first case most measures overestimate the inferred gain. In the second and third cases some measures overestimate the inferred gain (P@10, RBP) and others underestimate the gain (P@1) while the IFT measure tends to better track the inferred gain.

an intermediate depth—the rate of gain is sufficiently high to keep the search going initially, but falls off quickly—and approximates the inferred gain almost exactly.

The right-hand plot of Figure 6 shows the average gain and cost over a random sample of 5,000 impressions with clicks. We can see that the average accumulated gain grows fastest early on, then only slowly as people tend not to click further down the SERP. Most of the illustrated measures overstate the cumulated gain, as most measures put too much weight on higher ranks; the exceptions being P@1 and RR. Again IFT provides a very close fit, tracking the inferred gain.

Stopping Rank Likelihoods. Table 4 reports the mean likelihood of the stopping rank for each measure over all 673k query impressions (NB a higher likelihood is better). As shown in previous work a “shallower” or more “top-heavy” measure performs much better than “deeper” ones—and in fact in almost all cases searchers only click on the first one or two elements on the SERP. Conventional measures/settings such as P@10 or RBP with ϕ = 0.7, which assumes a user examines further down the list (to depth 10, or about 3–4, respectively), perform poorly as predictions of final stopping depth. However, our IFT measures perform substantially better that the other measures: suggesting that the foraging user model can better predict when/where users stop.

Gain and Cost. Table 4 also reports the mean absolute error (MAE) between the inferred total utility and the expected total utility (Eqn. 4), and the mean absolute error between the actual total cost and the expected total cost (Eqn. 7, in relative time units). Again this was calculated over all 673k query impressions. Here a lower error indicates a closer fit. In general, the measures best able to predict the stopping rank are also good at predicting the gain and cost. Again, the deeper measures overestimate the gain and the cost, with P@5 for example being out by 0.47 points of gain on average (equivalent to two partially relevant documents) and 4.31 units of time (the time to inspect an additional 4.31 algorithmic results). Again, the IFT measure provides the closest fit, quite substantially, indicating that the two constraints capture more accurately the behaviour of searchers on this sample of popular queries.

Goal vs. Rate Sensitivity. To illustrate the contributions of C1 (goal sensitivity) and C2 (rate sensitivity) we also computed the likelihood and error figures for two variants of IFT, “IFT-C1” and “IFT-C2”. Interestingly, both continuation functions perform very well alone—and while this is perhaps not surprising, as a good result ranking will provide relevant material early on, they both still outperform top-heavy measures like P@1, RR and SDCG@1. This suggests that the rationality parameter, which was set at 10, provided some additional flexibility to cater for the instances when users go deeper. Also of note is that the combination of C1 and C2 in IFT leads to small, but not negligible, improvements in terms of estimating the stopping and total cost.

We also note that the data set used here is from common queries to a web search engine. In other settings searchers may have different examination or stopping behaviours, due to differences in goals or in document collections. Alternative parameter settings would be appropriate in these cases, for IFT and for standard measures.

Caveats. While some of the measures evaluated above have tuned parameters, we emphasise that the parameters for IFT have not been tuned in any way—instead, where possible, we have grounded the selection of parameters for IFT based on its relationship to existing measures. Nonetheless, the above analysis does reveal that
encoding both goal and rate sensitivity in the user stopping model as motivated by IFT makes a material difference to our predictions of observed search behaviours motivating further analysis.

7 CONCLUSIONS AND FUTURE WORK

In this paper, our goal was to measure the utility (and cost) of SERPs. In the process, we introduced a measure, derived from Information Foraging Theory, where we have shown that the inclusion of continuation functions that encode goal and rate sensitivity, and account for the different cost of elements, has the potential to result in a better fit to observables such as stopping rank, time on SERP and inferred utility. This required a number of challenges to be addressed before we could even perform such measurements. We needed to develop methods to infer the scan path of elements to create an ordering (ranking) and then we needed to estimate the costs of different elements—both of which could be explored in significantly more depth. Further, we developed a new methodology for comparing and evaluating IR measures based on observable behaviour (stopping rank, time on page, and inferred utility) and so, for the first time we have shown how well common measures predict such observables in the context of web search. This was achieved by housing all measures within the C/W/L framework where it is possible to directly compare measures to each other and crucially to the inferred utility curves—leading to meaningful and comparable values (rank, gain, time). This is a significant step forward in IR evaluation and moves towards a more standardised and intuitive way to evaluate performance and behaviour.

Finally, the introduction of the foraging based user model not only directly connects the theory on search behaviour to how we measure search performance, but also opens up a new way to evaluate search performance—where we can examine the differences between users and user populations, in different tasks and settings—and explore not only their patience, but also their rationality, tolerance, and how goal-directed they are when searching.

ACKNOWLEDGMENTS

We thank Bodo von Billerbeck, Widad Machmouchi, David Hawking, Peter Bailey, and Alexander Moore for their assistance and suggestions. We also thank the SIGIR reviewers and committee for their effort and for their useful comments and feedback.

REFERENCES