# Summaries, Ranked Retrieval and Sessions: A Unified Framework for Information Access Evaluation

Tetsuya Sakai
Microsoft Research Asia, Beijing, P.R.C.
tetsuyasakai@acm.org

Zhicheng Dou
Microsoft Research Asia, Beijing, P.R.C.
zhichdou@microsoft.com

## ABSTRACT

We introduce a general information access evaluation framework that can potentially handle summaries, ranked document lists and even multi-query sessions seamlessly. Our framework first builds a *trailtext* which represents a concatenation of all the texts read by the user during a search session, and then computes an evaluation metric called *U-measure* over the trailtext. Instead of discounting the value of a retrieved piece of information based on ranks, U-measure discounts it based on its *position* within the trailtext. U-measure takes the document length into account just like *Time-Biased Gain* (TBG), and has the diminishing return property. It is therefore more realistic than rank-based metrics. Furthermore, it is arguably more flexible than TBG, as it is free from the *linear traversal* assumption (i.e., that the user scans the ranked list from top to bottom), and can handle information access tasks other than ad hoc retrieval. This paper demonstrates the validity and versatility of the U-measure framework. Our main conclusions are: (a) For ad hoc retrieval, U-measure is at least as reliable as TBG in terms of rank correlations with traditional metrics and discriminative power; (b) For diversified search, our diversity versions of U-measure are highly correlated with state-of-the-art diversity metrics; (c) For multi-query sessions, U-measure is highly correlated with Session nDCG; and (d) Unlike rank-based metrics such as DCG, U-measure can quantify the differences between linear and nonlinear traversals in sessions. We argue that our new framework is useful for understanding the user's search behaviour and for comparison across different information access styles (e.g. examining a direct answer vs. examining a ranked list of web pages).

## Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

## Keywords

diversity, evaluation, metrics, sessions, test collections


*SIGIR'13,* July 28–August 1, 2013, Dublin, Ireland.


## 1. INTRODUCTION

Evaluation is central to the progress of Information Retrieval (IR) research. It helps researchers build better systems to achieve their ultimate goal, namely, to satisfy the user's information need. In this era of digital information overload, *system-oriented* evaluation is a necessity as a complement to *user-oriented* evaluation, as the latter is difficult to scale and to generalise. However, system-oriented evaluation in IR is facing a few serious challenges:

**Challenge 1** System-oriented evaluation tends to oversimplify real search tasks in such a way that high effectiveness values may not guarantee high user performance or satisfaction. For example, whenever IR researchers rely entirely on metrics like Average Precision and Discounted Cumulative Gain (DCG), they implicitly assume that the user scans a ranked list of document IDs (not even documents) from top to bottom, like a machine; most metrics do not consider the user's actual effort for finding information from snippets and from documents of various lengths. Hence the criticism from the user-oriented camp: "*Where's the user?*"

**Challenge 2** "Search" has become a commodity, and as a result, IR tasks have diversified. It is no longer just about a list of documents. We want to evaluate a system that returns a single *multi-document summary* in response to a query (e.g. [19]), a *diversified* search result page (e.g. [20]), an *aggregated* search result page (e.g. [29]), or a system that tries to satisfy the user in a *session* that accommodates query reformulations (e.g. [14]). For every new task, an appropriate evaluation methodology needs to be designed, and virtually every such methodology faces **Challenge 1**. Furthermore, the IR community lacks a "common language" across these diverse tasks, which makes it difficult for us to compare across them and to build a unified IA system.

In the present study, we tackle the above two challenges: we believe that our proposed evaluation framework is a small but significant step towards that goal.

Figure 1 introduces *trailtexts*, the central concept in our evaluation framework. Part (a) shows a single textual query-biased summary being shown to the user. Suppose that we have observed (by means of, say, *eyetracking* [13]) that the user read only the first and the last sentences of this summary. In this case, we define the trailtext as a simple concatenation of these two sentences: "`Sentence1 Sentence2`." Part (b) shows an aggregated search output: the user reads a snippet in the *news* panel, then reads an *ad*, and finally reads a snippet in the *web* panel. In this case, the trailtext is defined as "`Snippet1 Ad2 Snippet3`." Part (c) is a more traditional search engine result page: the user reads the first two snippets, and then visits the second URL to read the full text. In

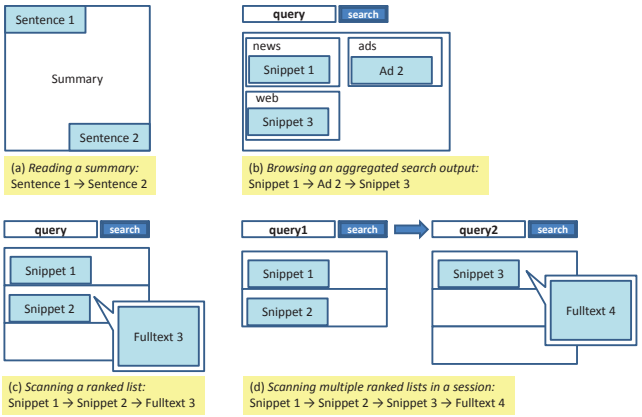


**Figure 1: Constructing trailtexts for various tasks.**

this case, the trailtext is "`Snippet1 Snippet2 Fulltext3`." Note that such trailtexts may possibly be constructed systematically under a certain user behaviour model instead of actual user observation. Finally, Part (d) shows a session that involves one query reformulation: the user reads two snippets in the original ranked list, reformulates the query, reads one snippet in the new ranked list, and finally visits the actual document. The trailtext is then "`Snippet1 Snippet2 Snippet3 Fulltext4`."

Our new evaluation framework comprises two steps:

**Step 1** Generate a trailtext, or multiple possible trailtexts, by either observing the actual user or assuming a user model;

**Step 2** Evaluate the trailtext(s), based on relevant *information units* (e.g. documents, passages, nuggets) found within it, while discounting the value of each information unit based on its *position* within the trailtext.

Our proposal was inspired by the *S-measure* framework of Sakai, Kato and Song [19], and the *Time-Biased Gain* (TBG) framework of Smucker and Clarke [23]. The S-measure framework, which only considered the evaluation of textual summaries (See Figure 1(a)), assumes that the user reads an entire summary from beginning to end at a constant speed, and that the value of an information unit wears out linearly according to its offset position in the summary. Thus, their key idea was to use *position*-based discounting, in contrast to the rank-based discounting employed by DCG for ranked retrieval. On the other hand, the TBG framework as described at SIGIR'12 was designed for ranked retrieval, and uses the *time spent by the user* as the basis for discounting the value of relevant information instead of the document rank[1].

While the work of Smucker and Clarke [23] primarily focused on estimating the *time to reach a document at rank* $k$, and assumed *linear traversal*, i.e., that all users scan the ranked list from top to bottom sequentially, our new framework is potentially more general. To be more specific, **Step 1** above says nothing about the user model: our framework can handle *nonlinear traversal*, i.e., cases where the user clicks a document at rank $k$ and then one at rank $j(< k)$. Our framework is also more general in that it can handle various tasks such as Parts (a), (b) and (d) of Figure 1, and in that it provides a common language across these tasks by means of trailtexts. Potentially, this may be useful for addressing questions such as: "Given this query, which is a better IA system? One that returns a single multi-document summary, or one that returns a list of URLs with snippets?" although we leave this for future work. The objective of this paper is to demonstrate the validity and versatility of the U-measure framework.

[1]S-measure's position-based discounting is in fact a form of time-based discounting, as it assumes that the user's reading speed is constant [19].

# 2. RELATED WORK

## 2.1 Evaluating Textual Output

In summarisation and Question Answering evaluation, *comparison units* for computing precision, recall and the like are created either manually (e.g. *semantic content units* [16]) or automatically (e.g. N-grams). The matching between the system output and the gold standard may also be performed either manually or automatically: ROUGE [15] is an example of the completely automatic approach. In contrast, the S-measure framework [19] relies on manual *information unit* extraction and manual matching, and uses the *positions* of the information units for discounting their values. S-measure encourages systems to present important pieces of information first and to minimise the amount of text the user has to read.

In the present study, we borrow the idea of position-based discounting from S-measure, but devise a general framework for evaluating summaries, ranked retrieval, search sessions and other textual information seeking activities. In our *U-measure* framework, an information unit may be a nugget, a search engine snippet, or an entire document as in traditional *document* retrieval. Both automatic and manual evaluation approaches are possible with this framework, although we only consider automatic evaluation based on document relevance and clicks in this paper.

## 2.2 Evaluating Ranked Document Retrieval

In the evaluation of ranked document retrieval, the common assumption is that the (average) user scans the ranked list from top to bottom until he stops at a certain rank [9, 12, 13]. This *linear traversal* assumption forms the basis of virtually all existing IR metrics, including those designed for diversified search, where systems are required to achieve not only high relevance but also high diversity across possible search intents [9, 10, 20].

Unlike traditional rank-based metrics, TBG pays attention to the fact that the time spent at each rank differs, depending especially on the document length and document novelty [23]. However, TBG as described in that work relies on estimating the time to reach rank $k$, which in turn relies on the linear traversal assumption on a single ranked list. In contrast, our U-measure is arguably more general in that it can naturally handle multi-query sessions and nonlinear traversals. In fact, we drop the notion of "document ranks" altogether once we have constructed the trailtext (Recall Figure 1). Just like TBG, U-measure can take into account not only the documents visited but also search engine snippets, which has recently been recognised as important in search engine evaluation (e.g. [24, 28]). On the other hand, one potential advantage of TBG (and other time-based approaches [5]) over ours is that it may also be useful for evaluating retrieval of nontextual information[2].

Evaluating diversified search has received attention recently, and several diversity metrics have been proposed and compared: Chapelle *et al.* [9] and Clarke *et al.* [10] independently showed that $\alpha$-*nDCG*

[2]One SIGIR reviewer commented: "*it is better to convert material read to a measure of time and stay in the TBG framework [..] TBG can be nicely adapted to non-ranked list evaluation.*" However, it is not immediately obvious to us exactly how TBG may be extended beyond "estimating the time to reach rank $k$" and to handle nonlinear traversal. Moreover, we argue that evaluating *textual* information access based on *texts* rather than time has benefits: for example, U-measure with information units should be useful for evaluating the quality of search engine snippets.

and *ERR-IA* (intent-aware expected reciprocal rank) have similar properties; Sakai and Song [20, 21] showed several advantages of *D-nDCG* over ERR-IA and $\alpha$-nDCG. In this study, we experiment with a "D-" version and an "IA" version of U-measure for diversity evaluation, as we shall describe later.

For both traditional and diversified IR, we believe that information units that are finer than "relevant documents" are required to properly handle information novelty and redundancy (e.g. [17]). As we discussed in Section 2.1, our framework can take any textual pieces of information as information units. However, in the present study, we limit ourselves to considering snippets and relevant/clicked documents. Note that our U-measure is a generalisation of S-measure, and that the latter has already been used successfully for evaluating query-biased summaries based on information units [19].

Recently, *user simulation* has received attention as a method for bridging the gap between system-oriented and user-oriented evaluation [8, 22]. Our framework is agnostic to whether the trailtext is generated based on user observation or simulation. While the evaluation methods explored in this paper are *deterministic* in that one particular user model is considered to generate exactly one trailtext for each search scenario, it would be possible to incorporate simulation to generate a population of trailtexts that reflect different user behaviours. But this is beyond the scope of this paper[3].

## 2.3 Evaluating More Complex Tasks

The IR community has considered the evaluation of tasks that are more complex than returning a single ranked list of documents. Session-based IR evaluation [5, 14] is one example. Our proposed framework can evaluate sessions as well, as all user actions are encoded as a trailtext. However, unlike Baskaya, Keskustalo and Järvelin [5], we do not explicitly consider the cost of user actions such as query (re)formulation and clicking on "next page." We assume that the text read by the user is an adequate representation of the user effort. Azzopardi [4] views interactive IR applications as a stream of documents and proposes evaluation metrics such as the "frequency of observing a relevant document." His document stream is similar to our trailtext, but the latter can potentially handle arbitrary pieces of text.

Other IR tasks are something of a mix between summarisation and ranked retrieval: character-based *bpref* (binary preference) has been used for evaluating a ranked list of passages [2]; Yang and Lad [26] proposed a nugget-based evaluation method that models *utility* as *benefit* minus *cost of reading* for evaluating multiple ranked lists of passages for a standing information need. Arvola, Kekäläinen and Junkkari [3] have proposed an evaluation method for an XML retrieval task where the user first sees a list of documents and then jumps to relevant passages of a document selected from that list. Their proposal is also similar to ours in that it also considers the amount of text read by the user as well as the actual reading order (within each document). The key differences are that they treat document list scanning and document browsing as two separate modes, and that they evaluate the former by average-precision-like metrics, which assume linear traversal.

# 3. PROPOSED FRAMEWORK

As was mentioned in Section 1, our evaluation framework first generates a trailtext based on user observation (e.g. eyetracking or click logging) or a user model, and then defines an evaluation metric over the trailtext by applying position-based discounting. Section 3.1 defines the general U-measure framework, which computes a score for a given trailtext and relevance information associated with it. Then Section 3.2 discusses how we actually derive trailtexts from document relevance and clicks in this study.

## 3.1 U-measure

A trailtext $tt$ is a concatenation of $n$ strings: $tt = s_1 s_2 \ldots s_n$. Each string $s_k (1 \leq k \leq n)$ could be a document title, snippet, full text, or even some arbitrary part of a text (e.g. nugget). We assume that the trailtext is exactly what the user actually read, in the exact order, during an information seeking process. We define the offset position of $s_k$ as $pos(s_k) = \sum_{j=1}^{k} |s_j|$. We measure lengths in terms of the number of *characters* [19]. Each $s_k$ in a trailtext $tt$ is considered either $l$-relevant, i.e. relevance level of $l (> 0)$, or nonrelevant. For example, in the case of summarisation evaluation based on nuggets, a relevant $s_k$ may be a string that has been found (either manually or automatically) to be a match with a gold standard nugget [19]. Alternatively, in an evaluation environment where only document relevance assessments are available, a relevant $s_k$ may be the full text of a relevant document, where it is assumed that the user actually read the entire document, as we shall discuss in Section 3.2. We define the *position-based gain* as $g(pos(s_k)) = 0$ if $s_k$ is considered nonrelevant, and $g(pos(s_k)) = gv_l$ if it is considered $l$-relevant, where $gv_l$ is a *gain value* for relevance level $l$. Note that a string $s_k$ that is *considered* nonrelevant may in fact be relevant: for example, if the user reads duplicate documents, it is possible to count only the first one as relevant [23]; similarly, for information needs that do not require exhaustive pieces of information, it is possible to treat only the first relevant piece of information as relevant, as in the Reciprocal Rank metric.

The general form of U-measure is given by:

$$U = \frac{1}{\mathcal{N}} \sum_{pos=1}^{|tt|} g(pos) D(pos) \quad (1)$$

where $\mathcal{N}$ is a normalisation factor (which we simply set to $\mathcal{N} = 1$ in this study, following recent evaluation studies [5, 23]) and $pos$ is an offset position within $tt$ and $D(pos)$ is a *position-based* decay function. Following the S-measure framework [19], here we assume that the value of a relevant information unit decays with the amount of text the user has read, and adopt a linear function:

$$D(pos) = \max(0, 1 - \frac{pos}{L})) \, . \quad (2)$$

Here, $L$ is the amount of text at which all relevant information units become worthless. Note that, if the user's reading speed is constant, Eq. 2 is also equivalent to linear discounting *by time*.

In this study, we interpret $L$ as the largest *Maximal Trailtext Length* (MTL) across all possible search sessions, where the MTL of a session is the sum of the lengths of (a) all snippets above the lowest click and (b) all documents clicked by the user in that session. Thus, $L$ represents the largest amount of text that the user may have had to read in one session. This is the point where we consider that all information units become worthless. Note that we do *not* rely on the linear traversal assumption: the snippets may have been read in any order, and the documents may have been clicked in any order. We simply sum up the snippet and document lengths.

To estimate $L$, we first obtained 21,911,694 sessions (partitioned based on 30-minute inactivities) from Microsoft's Bing (September 7, 2012, US market), under the constraint that every query in the session received at least one click. This constraint is convenient for evaluating multi-query sessions, as we shall discuss in Section 6.

[3] In our diversity experiments (Section 5), we do consider an Intent-Aware version of U-measure, which considers multiple trailtexts that represent different user intents.

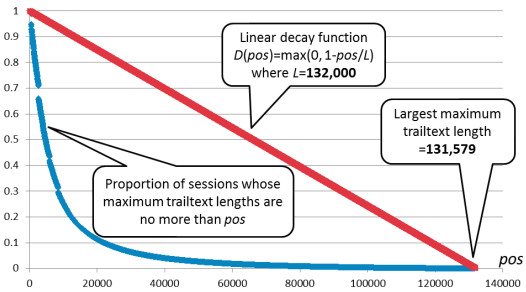


**Figure 2: Proportion of sessions whose MTLs are no more than $pos$ (after removing 0.5% of the sessions with the largest MTLs), and a linear decay function derived from it.**

By assuming that every snippet is 200-character long (which on average is a valid assumption for Bing) and automatically counting the number of characters for every clicked document, we computed the MTL for each session, and discarded 0.5% sessions with the largest MTL values (extremely long sessions for some unknown reasons). As the largest MTL among the remaining 21,802,136 sessions was 131,579, we chose to set $L = 132,000$. While it is difficult to imagine a user who *actually* reads over 100,000 characters within a session, it is important that $L$ is sufficiently large for evaluation purposes, as discussed below.

For the 21,802,136 sessions, the curve in Figure 2 shows the proportion of sessions whose MTLs are no more than $x$. There is a sudden drop near the 70% line, which is caused by extremely popular "navigational" sessions whose MTLs were identical (computed as the length of the first snippet plus that of the top page of the popular website). The curve seems to be in line with existing click-based studies, and that an exponential function may be a good approximation to it (e.g. [6, 23]). We leave the curve fitting for future work, and use the simple linear decay function shown in the figure: perhaps what really should be done is to segment queries into several user behaviour types (e.g. informational, navigational or even more fine-grained types), and to devise different decay functions that match these types. The raw top-heavy curve is just an average of various search behaviours, and not necessarily appropriate for evaluation purposes where, for example, we are also interested in handling informational queries. For example, the average document length computed based on our full session data is 5,445.0 characters (averged over 39,716,443 clicked web pages); that computed based on the relevant news articles of the TREC 2005 robust track test collection (which we use in Section 4 to replicate the TBG experiment by Smucker and Clarke [23]) is 3,672.0 characters (averaged over 334,079 articles). Whereas, according to the curve, the decay value is 50% when $x = 4,717$ characters. Thus, if we use the curve directly as a decay function, the value of a relevant document is halved after only one or two relevant documents have been found. In the present instantiation of U, we use the linear function to pay attention to more relevant documents. Also, since $L$ determines the gradient of the linear function, we set it to a large value for the same reason.

For evaluating summarisation and traditional ranked retrieval, the gain value for an $l$-relevant information unit could be set, for example, as $gv_l = (2^l - 1)/2^H$, where $H$ is the highest relevance level [9, 20]. Whereas, in a diversity evaluation environment where multiple possible intents $i$ are known for each query $q$ and per-intent relevance assessments as well as the intent probabilities $P(i|q)$ are available [20, 21], we can define the "global gain" for each relevant $s_k$ as:

$$g(pos(s_k)) = \sum_i P(i|q) g_i(pos(s_k)) \quad (3)$$

where $g_i(pos(s_k)) = gv_l$ if $s_k$ is $l$-relevant to the $i$-th intent. Thus, this is the overall value of a document obtained by combining the "local" (i.e. per-intent) gain values. We plug Eq. 3 into Eq. 1 to obtain $D$-$U$, an extension of *D-measure* for diversity evaluation [20].

Another natural way to handle diversity is to compute a U-measure value ($U_i$) for each intent $i$ separately, and finally combine them using the intent probabilities $P(i|q)$. This follows the *Intent-Aware* (IA) approach to diversity evaluation [1, 9]:

$$U\text{-}IA = \sum_i P(i|q) U_i \; . \quad (4)$$

## 3.2 Deriving Trailtexts from Document Relevance and Clicks

In Section 3.1, we discussed the general case where each string $s_k$ could be *any* piece of text that the user has read. Without eyetracking studies, however, it is difficult to construct a trailtext based on user observation. Therefore, in this paper, we consider special cases where we assume that $s_k$ is either a search engine *snippet* or (a part of) the *full text* of the document. For each query (or a session), we construct a trailtext or several trailtexts automatically, by leveraging either *document relevance assessments* of existing test collections or *document clicks* that we obtained from a commercial search engine. In the latter case where *click order* information is available, we can drop the linear traversal assumption.

Figure 3 illustrates one way to automatically construct trailtexts based on *relevance assessments* and document rankings. Part (a) could be a ranked list from a TREC run: its second and fourth documents are relevant. (For simplicity we consider binary relevance here, although we actually leverage graded relevance.) Based on the linear traversal assumption, we can build a trailtext as shown in Part (b). Then, U can easily be computed using the relevant positions $pos(s_3)$ and $pos(s_6)$.

An important point to note here is that the U-measure computed based on trailtexts such as those described above satisfies the *diminishing return* property, similar to ERR [9]. Suppose that, in Figure 3 Part (a), the nonrelevant document at rank 3 is replaced by a relevant document. Then, since we now assume that the third document is also read, the trailtext will be longer than the one shown in Part (b), and the full text of the *fourth* document is pushed back towards the end of the new trailtext. As a result, the value of the fourth document *diminishes* according to the decay function shown in Figure 2. Whereas, many rank-based metrics such as nDCG lack this property: the value of the relevant document at rank 4 is determined absolutely by its gain value and its rank, no matter what the ranked list has above that rank.

Figure 3(c) shows a diversified ranked list of documents for a query that is known to have two intents. The second document is relevant to Intent 1 but not to Intent 2, while the fourth one is relevant to Intent 2 but not to Intent 1. In the $D$-$U$ methodology discussed in Section 3.1, a single "global gain" value is computed for every document using Eq. 3, and a single trailtext is created as shown in Part (b). It is assumed that both of the two relevant documents are read. On the other hand, in the aforementioned *U-IA* methodology, a trailtext is created for each intent, as shown in Part (d). For example, the trailtext for Intent 1 is created by assuming that only the document at rank 2 is read.

Figure 4 illustrates one way to automatically construct trailtexts based on *click data*, which is in line with the way we compute MTLs (Section 3.1). Using click data, we can conduct session evaluation involving query reformulations and nonlinear traversals. Given the lack of eyetracking evidence, we assume that, in every session, the user reads every document he clicks, and that he reads

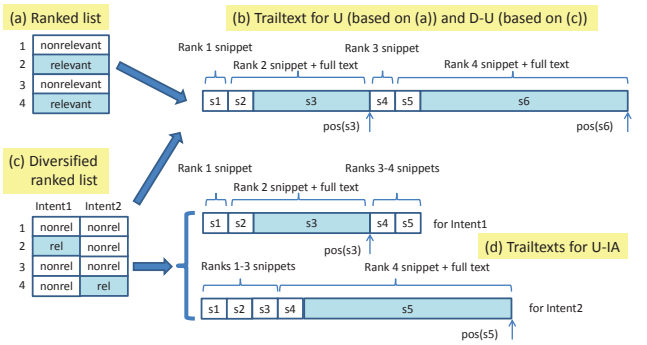


**Figure 3: Automatically constructing trailtexts from relevance assessments of traditional and diversified IR test collections.**

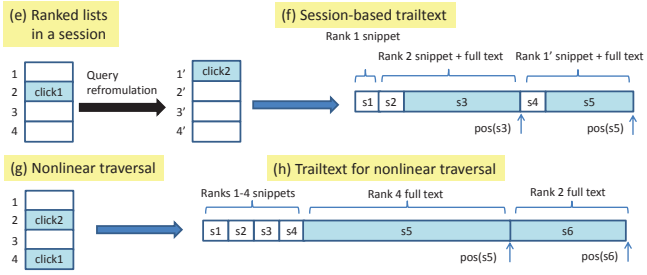


**Figure 4: Automatically constructing trailtexts from clicks for nonlinear traversals and sessions.**

every snippet ranked above the lowest clicked rank in every ranked list shown during the session[4]. Part (e) illustrates a session involving one query reformuation and therefore two ranked lists: the user clicks the second document in the first list and then the first document in the second list. Part (f) shows a possible trailtext for this behaviour. Part (g) shows a nonlinear traversal in which the user clicks the document at rank 4 and then one at rank 2; Part (h) shows a possible trailtext for this user, by assuming that snippets between ranks 1 and 4 are read before the full texts of the two documents are read. In this paper, we only consider sessions in which every ranked list contains at least one click: without eyetracking we cannot tell how the user examined a ranked list that did not result in any clicks [14].

So far, we have assumed that the user reads (I) all snippets above the lowest relevant or clicked document; and (II) entire full text of every relevant/clicked document. In practice, we make a more realistic assumption than (II), namely, that (II′) the user reads only $F\%$ of each relevant/clicked document. In the present study, we assume that the user reads only 20% of each clicked document (i.e. $F = 0.2$) by default. In our comparative experiments with TBG using the TREC 2005 robust data (Section 4), we show that the choice of $F$ has a direct impact on the correlation with ranked-based metrics and on discriminative power, and that $F = 0.2$ is a reasonable choice. In Section 4.3, we show that our initial attempt at estimating $F$ from click data also supports this choice. On the other hand, we stick to the choice of $L = 132,000$, based on the largest MTL discussed in Section 3.1. This is because, even if an *average* user reads only $F\%$ of each clicked document, we want to accommodate users who read the entire documents as well.

Figure 5 shows our instantiation of U-measure based on click data. The pseudocode reads a file where each line is a triple consisting of *querynumber* (e.g. 1 for the first query in a session), *clickedrank* and *doclen* (length of clicked document), and the lines are chronologically ordered. It handles multi-query sessions and nonlinear traversals as we have illustrated in Figure 4. The use of the array *snippetdone* reflects our assumption that the user does not read the same snippet twice, although alternatives are possible.

[4]An eyetracking study by Joachims *et al.* supports this assumption [13]. For example, they report that when rank 5 is clicked, then snippets between ranks 1 and 4 are read 54.5-81.8% of the time; that at rank 5 is read 100% of the time, and that at rank 6 is read only 18.2% of the time.

$snippetlen = 200$;
$g = 0.5$; // gain of a clicked document: $(2^l - 1)/2^H = (2^1 - 1)/2^1$.
$pos = 0$; $U = 0$;
**while** read $< querynumber, clickedrank, doclen >$ sorted by time
  **if** $querynumber$ is new **then** initialise array $snippetdone[]$;
  // stores whether or not snippet at rank $r$ has already been read.
  **for**( $r = 1$; $r \leq clickedrank$; $r$++)
    **if** $snippetdone[r] == 0$ **then**
      $pos$ += $snippetlen$; //reads all snippets above a click.
      $snippetdone[r] = 1$;
    **end if**
  $pos$ += $F * doclen$; // reads $F\%$ of clicked document.
  $U$ += $g * \max(0, 1 - pos/L)$;
**end while**
**return** $U$;

**Figure 5: Algorithm for computing U-measure by reading a session data file, which consists of *querynumber*, *clickedrank* and *doclen* sorted by time.**

# 4. EVALUATING TRADITIONAL IR

## 4.1 Experimental Setting

To demonstrate that U-measure is a useful alternative to TBG, we first compare them in an experimental setting very similar to that of Smucker and Clarke [23]: we evaluate 74 TREC 2005 Robust Track runs with 50 topics [25], using the document length statistics from the AQUAINT corpus[5]. Smucker and Clarke estimated several parameters required to instantiate TBG, based on a user study involving eight TREC 2005 Robust Track topics and 48 participants, where each search session ran up to 10 minutes. We copy all parameter values from their study: hence our instantiation of TBG can be described as follows.

$$TBG = \sum_{r=1}^{\infty} g(r) \exp(-T(r)\frac{\ln 2}{224}) \quad (5)$$

where the exponential factor is the time-based decay function[6] and $T(r)$ is the estimated time to reach rank $r$, computed as the time to read snippets plus the time to read clicked documents:

$$T(r) = \sum_{m=1}^{r-1} 4.4 + (0.018 l_m + 7.8) p_{click}(m) \; . \quad (6)$$

Eq. 6 relies on two important assumptions, namely, the linear traversal assumption (as the summation over previous ranks suggests), and that the user's reading speed is constant (the time required to read a full text is linear with respect to its length $l_m$ as measured by the number of *words*). Also, according to Smucker and Clarke's calibration, $p_{click}(m) = 0.64$ if the document at rank $m$ is relevant and $p_{click}(m) = 0.39$ otherwise; the gain value was estimated to be $g(r) = 0.4928$ for a relevant document and otherwise zero [23]. Thus this TBG is binary-relevance-based.

[5]http://www.ldc.upenn.edu/Catalog/docs/LDC2002T31/

[6]Smucker and Clarke [23] estimated this half-life of $h = 224$ from an MSN search engine query log that contains about five million searches.

While U and TBG are similar, we note that diminishing return is not guaranteed with TBG. Consider what happens to a relevant document at rank 2 when a 1000-word nonrelevant document at rank 1 is replaced by a 10-word relevant document. According to Eq. 6, the gain for rank 2 actually *increases*[7]. But this is not a problem if the document lengths do not vary wildly.

We note that while Smucker and Clarke used automatic duplicate document detection to account for the fact that users tend to read "redundant" documents quickly, our experiments do not employ any special treatment of such documents. As we mentioned earlier, we plan to address novelty and redundancy issues based on information units smaller than documents in our future work.

As was discussed in Section 3, we use $L = 132,000$ and $F = 0.2$ for computing U-measure by default. We compare U with TBG, AP, nDCG@10 and nDCG@1000. As U and nDCG can handle graded relevance, we used the standard gain value setting of $gv_l = (2^l - 1)/2^H$, where the highest relevance level $H$ is 2 for the Robust test collection. In addition, we also experimented with their binary versions, which we denote by $\mathrm{U}_{bin}$ and $\mathrm{nDCG}_{bin}$, as TBG and AP do not utilise graded relevance. AP and nDCG were computed using NTCIREVAL[8]; the document lengths (in words for TBG and in characters for U) were computed using the HEADLINE and TEXT fields of the AQUAINT collection.

## 4.2 Main Traditional IR Results

Table 1 compares the run rankings produced by different metrics in terms of Kendall's $\tau$ and symmetric $\tau_{ap}$. The latter is similar to $\tau$ but is more sensitive to top ranks [27]. It can be observed that U is highly correlated with nDCG@1000, AP and TBG, and that it is more highly correlated with nDCG@1000 and with AP than TBG is. For example, the $\tau$ between U and nDCG@1000 is .819, while that between TBG and nDCG@1000 is .780. Whereas, TBG is more highly correlated with nDCG@10 than U is (.734 vs. .653 in $\tau$), which is probably because the exponential decay of TBG is more top heavy than the linear decay of U. Table 2 shows the correlation results for $\mathrm{U}_{bin}$ and $\mathrm{nDCG}_{bin}$, which suggest that reducing graded relevance to binary relevance has little impact on the rankings. Indeed, although not shown in these tables, the $\tau$ ($\tau_{ap}$) between U and $\mathrm{U}_{bin}$ is .920 (.803).

We also compare the metrics in terms of *discriminative power* [18], which has been used in a number of recent studies for comparing the stability of metrics (e.g. [10, 20, 23]). While discriminative power does not say anything about whether the metric is right or wrong, low discriminative power means that the metric is not useful for drawing conclusions from an experiment. Table 3 compares the metrics in terms of discriminative power at the 0.05 significance level, based on a randomised version of two-sided Tukey's Honestly Significant Differences test. (This test is more conservative than standard pairwise tests, as it considers the entire set of runs [7].) Here, "required $\Delta$" is the minimal performance difference that is usually statistically significant [18]. For example, Table 3 shows that nDCG@1000 detected a significant difference for 26.9% of the $(74*73/2 =)2,701$ run pairs, and that a difference of 0.15 in terms of Mean nDCG@1000 over 50 topics is usually statistically significant. Figure 6 provides a more general picture for some of the metrics in Table 3, in the form of *ASL (Achieved Significance Level) curves* for $0 < ASL \leq 0.1$ [18]. These results show that while U is substantially less discriminative than nDCG@1000 and AP, it is at least as discriminative as TBG, and much more discriminative than nDCG@10. Also, the results are consistent with the discriminative power experiments of Smucker and Clarke [23], who reported that TBG lies between AP and nDCG@10. The diminishing return property of U and TBG is probably one reason why they are not as discriminative as AP and nDCG@1000: once a relevant document is found, they tend to ignore additional relevant documents, which means observing fewer data points.

To summarise our experiments with the TREC Robust Track data (following Smucker and Clarke [23]), we have shown that U is highly correlated with nDCG@1000, AP and TBG, and that it is at least as discriminative as TBG. Hence we believe that it is fair to regard U as a good alternative to TBG for the purpose of traditional ad hoc IR evaluation. In the remainder of this paper, we demonstrate the usefulness of U in other IR settings. We shall not discuss TBG any further as the instantiation of TBG we used was calibrated based on a particular user study [23], and its appropriateness for other IR settings is unknown.

**Table 1: TREC 2005 Robust $\tau$/$\tau_{ap}$ rank correlation (74 runs; 50 topics): AP, TBG and graded-relevance metrics.**

| | nDCG@1000 | AP | TBG | U |
|---|---|---|---|---|
| nDCG@10 | .730/.591 | .726/.592 | .734/.638 | .653/.531 |
| nDCG@1000 | - | .913/.843 | .780/.635 | .819/.680 |
| AP | - | - | .792/.644 | .816/.685 |
| TBG | - | - | - | .834/.692 |

**Table 2: TREC 2005 Robust $\tau$/$\tau_{ap}$ rank correlation (74 runs; 50 topics): all metrics use binary relevance information.**

| | $\mathrm{nDCG}_{bin}$@1000 | AP | TBG | $\mathrm{U}_{bin}$ |
|---|---|---|---|---|
| $\mathrm{nDCG}_{bin}$@10 | .720/.565 | .715/.561 | .732/.663 | .648/.494 |
| $\mathrm{nDCG}_{bin}$@1000 | - | .903/.817 | .781/.635 | .825/.742 |
| AP | - | - | .792/.644 | .814/.736 |
| TBG | - | - | - | .846/.717 |

**Table 3: TREC 2005 Robust discriminative power at $\alpha = 0.05$ (74 runs; 50 topics).**

| metric | disc. power | required $\Delta$ |
|---|---|---|
| nDCG@1000 | 26.9% | 0.15 |
| $\mathrm{nDCG}_{bin}$@1000 | 26.7% | 0.16 |
| AP | 26.0% | 0.12 |
| U | 20.0% | 5.28 |
| $\mathrm{U}_{bin}$ | 19.6% | 11.00 |
| TBG | 18.3% | 1.66 |
| $\mathrm{nDCG}_{bin}$@10 | 16.4% | 0.21 |
| nDCG@10 | 14.6% | 0.17 |

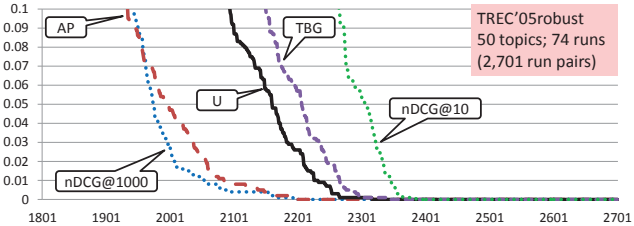


**Figure 6: ASL curves for TREC 2005 Robust runs. The $y$ axis represents the ASL and the $x$ axis represents run pairs sorted by the ASL. nDCG and U utilise graded relevance.**

## 4.3 Effect of F

While we have demonstrated that our default version of U performs at least as well as TBG in terms of rank correlation and discriminative power, these results are in fact dependent on the choice of $F$, the percentage of text that the user is expected to read for each relevant document. The main reason U and TBG are highly correlated with rank-based metrics is that they retain the rank-based penalty mechanism in the form of snippet length or snippet reading time; the main reason U and TBG differ from the ranked-based metrics is that they more realistically reflect the effort spent on

[7] $T(2)$ before and after the replacement are 14.5 and 9.5.
[8] http://research.nii.ac.jp/ntcir/tools/ntcireval-en.html

reading each relevant document, and have the diminishing return property (on average for TBG). Thus, if we emphasize the snippet reading feature and suppress the document reading features of U or TBG, we get a metric that is more like a ranked-based one. As for discriminative power, a highly discriminative metric is required to (a) examine a wide range of ranks; and (b) emphasize the top ranks[9]. In the case of U, it is relatively easy to boost rank correlation with existing metrics and discriminative power at the same time, by lowering the value of $F$. Recall that lowering $F$ means accumulating shorter strings $s_k$ for trailtexts: this results in examination of deeper ranks, as illustrated below.

Given a ranked list to be evaluated, let the *Effective Measurement Depth* (EMD) be the rank at which the decay $D(pos)$ falls below 0.0001, so that the ranks further down will actually be ignored. Furthermore, let the *Average EMD* (AEMD) be the EMD averaged across all runs and all topics. In our TREC Robust experiments, the AEMD of TBD is 198.2. That is, TBD examines about top 200 documents on average. Whereas, the AEMD of U is shown in Table 4 as a function of $F$, together with the $\tau$ and $\tau_{ap}$ between U and other metrics: it can be observed that, as $F$ is increased, the AEMD of U goes down monotonically. Moreover, as $F$ is increased, the impact of the snippet reading feature (i.e. rank-based penalty) is reduced and the correlations with nDCG and AP go down monotonically as well. For example, with nDCG@1000, the $\tau$ is .876 when $F = 0.1$ but .471 when $F = 1.0$. On the other hand, $F = 0.2$, our default version of U, appears to be the most similar to TBG: the $\tau$ values with TBG for $F = 0.1$ and $F = 0.4$ are lower than that for $F = 0.2$, namely, $\tau = .834$.

Figure 7 shows the effect of $F$ on the ASL curve of U: it can be observed that the discriminative power of U can be enhanced by setting $F$ to a value even smaller than 0.2 (which also enhances the correlations with nDCG and AP as shown in Table 4). Interestingly, however, the relationship between $F$ and discriminative power is not monotonic: for example, U with $F = 1.0$ appears to be a little more highly discriminative than U with $F = 0.6$.

We have also explored ways to estimate $F$ from actual click data, but decided not to use these estimates directly in the present study for two reasons. First, we found that $F$ is heavily dependent on clicked ranks: users spend a lot of time on top ranked documents. While the U-measure framework is open to incorporating a variable $F$ for constructing the trailtext, this does make the metric more complex. Second, we found that the estimate of $F$ varies considerably depending on how it is estimated. One problem we encountered is that users apparently do other things besides reading text when they dwell on a clicked page or a search engine result page.

One of the more successful methods we tried was as follows. First, to estimate the user's reading speed ($SP$), which we assume to be constant (as does TBG), we computed the Snippet Reading Time $SRT = t/c$ for every query $q$, where $c$ is the rank of the document that received the first click and $t$ is the time spent after issuing the query and until this first click. By averaging over 46,526,324 queries from Bing's query log (September 7, 2012, US market), we obtained $SRT = 14.002$ seconds[10]. Hence, the reading speed can be estimated as $SP = 200/14.002 = 14.283$ chars/sec. Second, to estimate $F$, we obtained 1,789,636 instances from the same session data where ranks $k$ and $(k+1)$ in the first result page were clicked consecutively, in this exact order. For each instance, let $t'$ be the time spent between these clicks (i.e. time to read the document at rank $k$ and then the snippet at rank $(k+1)$), and let $dl$ be the length of the document at rank $k$. Then the amount of text read at rank $k$ can be estimated as $SP * t' - 200$, and therefore $F = (SP * t' - 200)/dl$. The average $F$ over all instances was 0.282.

In summary, U with $F = 0.2$ is a reasonable choice, as it is highly correlated with TBG as well as with nDCG and AP, achieves relatively high discriminative power, and is supported by real click data. Hereafter, we use $L = 132,000$ and $F = 0.2$ in all of our experiments.

[9]For example, Precision@1000 satisfies (a), but not (b).

[10]This is considerably longer than 4.4 seconds in Eq. 6, the "time to evaluate a summary" in the TBG framework [23].

**Table 4: $\tau$ (top row) and $\tau_{ap}$ (bottom row) with the TREC 2005 Robust data (74 runs and 50 topics): effect of $F$ on U vs. other metrics. The $F = 0.2$ column has been copied from Table 1. Highest values across columns are shown in bold.**

| $F$ | 0.1 | 0.2 | 0.4 | 0.6 | 0.8 | 1.0 |
|---|---|---|---|---|---|---|
| AEMD of U | 462.9 | 384.0 | 281.3 | 221.8 | 182.0 | 155.7 |
| nDCG@10 | **.686** | .653 | .590 | .476 | .413 | .348 |
| | **.525** | .531 | .450 | .332 | .282 | .235 |
| nDCG@1000 | **.876** | .819 | .698 | .571 | .491 | .471 |
| | **.776** | .680 | .500 | .371 | .302 | .241 |
| AP | **.876** | .816 | .690 | .562 | .485 | .411 |
| | **.795** | .685 | .506 | .355 | .289 | .228 |
| TBG | .811 | **.834** | .801 | .697 | .634 | .562 |
| | .682 | **.692** | .608 | .489 | .428 | .364 |

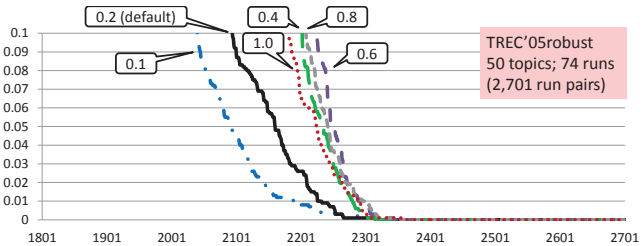


**Figure 7: Effect of $F$ on the ASL curves of U for TREC 2005 Robust runs.**

## 5. EVALUATING DIVERSIFIED IR

In this section, we compare our trailtext-based diversity metrics *D-U* and *U-IA* with existing rank-based diversity metrics. For this purpose, we use the TREC 2011 Web Track Diversity Task data [11] as it contains per-intent *graded* relevance assessments (with $H = 3$) unlike its predessessors. We used its 50 topics and 17 Category A runs. The rank-based metrics we consider are *I-rec* (a.k.a. subtopic recall), *D-nDCG* and *D♯-nDCG* (i.e. simply an average of I-rec and D-nDCG), and per-intent-normalised version of ERR-IA [9] which we call *nERR-IA* [20, 21]. The first three are the official metrics at the NTCIR INTENT task; A version of ERR-IA was used as the primary metric at the TREC diversity task. Again, we use NTCIREVAL to compute these rank-based metrics, using the exponential gain value setting (See Section 4). Following TREC, the intent probabilities $P(i|q)$ are assumed to be uniform. As diversification concerns the very top of the ranked list, we evaluate the top ten documents (D-nDCG@10, etc.).

Table 5 compares the TREC 2011 Diversity rankings according to different metrics in terms of $\tau$ and symmetric $\tau_{ap}$. It can be observed that D-U and U-IA are highly correlated with existing ranked-based diversity metrics (e.g. the $\tau$ between D-U and D-nDCG is .809), and that D-U and U-IA are extremely highly correlated with each other ($\tau = .985$). Below, we explain the latter observation.

As diversity metrics are generally more complex than traditional metrics, Figure 8 provides an example ranked list from an actual TREC 2011 Diversity run (Topic=137; Run=uwBA), with how D-U and U-IA are actually computed. Understanding this example should also help the reader see why D-U and U-IA are very similar.

**Table 5: TREC 2011 Diveristy $\tau$/$\tau_{ap}$ rank correlation (17 Category A runs; 50 topics).**

| | D-nDCG | D♯-nDCG | D-U | nERR-IA | U-IA |
|---|---|---|---|---|---|
| I-rec | .809/.761 | .897/.890 | .706/.560 | .676/.626 | .691/.551 |
| D-nDCG | - | .912/.859 | .809/.668 | .779/.689 | .794/.659 |
| D♯-nDCG | - | - | .750/.610 | .750/.696 | .735/.601 |
| D-U | - | - | - | .647/.528 | .985/.991 |
| nERR-IA | - | - | - | - | .632/.519 |

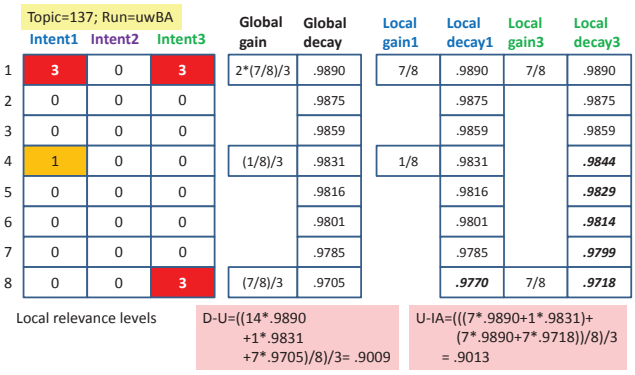


**Figure 8: Examples of how D-U and U-IA are computed for the TREC 2011 Diversity runs.**

This topic has three intents (i.e. subtopics), and the run returned three relevant documents: the one at rank 1 is 3-relevant to Intents 1 and 3; the one at rank 4 is 1-relevant to Intent 1 only, and so on. D-U assumes that all of the relevant documents are read, and computes "global gains" using Eq. 3 as shown in the figure. For example, for the document at rank 1, since $gv_3 = (2^3 - 1)/2^3 = 7/8$ for both Intents 1 and 3, the global gain is $2 * (7/8)/3$. On the other hand, the $D(pos)$ values are shown in the "global decay" column: for example, as the length of the document at rank 1 was found to be 6,279 characters, the estimated trailtext length after reading (a part of) this document is $200 + 0.2 * 6279 = 1455.8$. Thus $D(pos) = 1 - 1455.8/132000 = .9890$. Whereas, at rank 2, since we assume that only the snippet is read, $D(pos) = 1 - (1455.8 + 200)/132000 = .9875$. The final value of D-U is .9009. On the other hand, U-IA is computed as shown on the right of Figure 8. For Intent 1, it is assumed that only the documents at ranks 1 and 4 are read; for Intent 3, it is assumed that only the documents at ranks 1 and 8 are read. Then a U value is computed separately for Intents 1 and 3, and the final value of U-IA is .9013. Note, for example, that the local decay for Intent 3 starts deviating from the global decay at rank 4 (.9844 vs. .9831), as the document at rank 4 is not relevant to this intent.

More generally, let $I$ be the set of known intents for a topic, and let $I'(\subseteq I)$ be the set of intents covered by a ranked list (so that $I\text{-}rec = |I'|/|I|$). We say that a document in the list is *strictly locally relevant* if it is relevant to at least one intent from $I'$ *and* nonrelevant to at least one intent from $I'$. In Figure 8, the documents at ranks 4 and 8 are strictly locally relevant with respect to the ranked list. It is easy to see that if there is no strictly locally relevant document in the ranked list, then $D_i(pos) = D(pos)$ holds for any $(i, x)$. That is, any local decay value would be identical to the global one. Whereas, $D\text{-}U = \sum_{pos}(\sum_i P(i|q)g_i(pos))D(pos) = \sum_i P(i|q)(\sum_{pos} g_i(pos)D(pos))$ from Eqs. 1 and 3, and $U\text{-}IA = \sum_i P(i|q)(\sum_{pos} g_i(pos)D_i(pos))$ from Eq. 4. Hence it is clear that *if there is no strictly locally relevant document in the ranked list, then D-U = U-IA holds*. A corollary is that *if the ranked list covers only one intent (i.e. $|I'| = 1$), then D-U = U-IA holds*.

**Table 6: TREC 2011 Diveristy discriminative power at $\alpha = 0.05$ (17 Category A runs; 50 topics).**

| metric | disc. power | required $\Delta$ |
|---|---|---|
| D-nDCG | 27.2% | 0.12 |
| D♯-nDCG | 23.5% | 0.15 |
| I-rec | 19.1% | 0.19 |
| nERR-IA | 17.6% | 0.14 |
| D-U | 14.7% | 0.35 |
| U-IA | 14.7% | 0.35 |

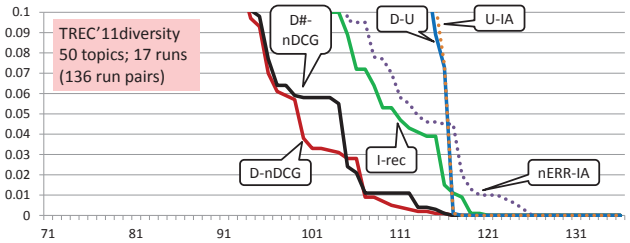


**Figure 9: ASL curves for TREC 2011 Diversity runs. The $y$ axis represents the ASL and the $x$ axis represents run pairs sorted by the ASL.**

Table 6 compares the discriminative power of the diversity metrics at $\alpha = 0.05$. Figure 9 visualises the discriminative power for $0 < ASL \leq 0.1$. From these results, it can be observed that D-U and U-IA are substantially less discriminative than D(♯)-nDCG. On the other hand, while they are less discriminative than nERR-IA at $\alpha = 0.05$, they are actually more discriminative at $\alpha = 0.01$. Again, the diminishing return property of D-U, U-IA and nERR-IA is one reason why they are not as discriminative as D(♯)-nDCG[11].

To sum up: D-U and U-IA are highly correlated with existing ranked-based diversity metrics, and are very highly correlated with each other. While they are not as discriminative as D(♯)-nDCG, they may be useful for evaluating web search result diversification from the user's perspective, as they are the only ones that take the document lengths into account.

## 6. EVALUATING MULTI-QUERY SESSIONS

Sections 4 and 5 discussed TREC-style evaluations using relevance assessments. In this section and in Section 7, we utilise the session data mentioned in Section 3.1 to compute U-measure based on clicks, using the algorithm shown in Figure 5.

To test the validity of U for *multi-query* session evaluation, we first constructed records of the form *sessionID*, *querynum*, *clickdrank*, *doclen* for the aforementioned 21,911,694 sessions. For example, a record (S1,2,3,500) indicates that in Session S1, the third URL for the second query (i.e. after one query reformulation) was clicked and that the document length is 500 characters. Moreover, within each session, the records are sorted by time. 5,178,327 of these sessions (23.6%) contained multiple queries. From these sessions, we obtained a random sample containing 50,000 sessions. The average and the maximum number of queries per session for this sample are 2.649 and 50, respectively; the average and the maximum number of clicks per session are 3.566 and 124, respectively. Sessions with many queries and clicks are often to do with pornography, as we shall discuss later.

For comparison, we also computed a version of *Session DCG* (sDCG), similar to that instantiated by Kanoulas *et al.* [14]. Although they also proposed session evaluation metrics that consider many possible browsing paths over multiple ranked lists, we do not consider them, since in our experimental setting, we can deterministically construct a trailtext based on the lowest click in each

[11]As for I-rec, its discriminative power is known to vary widely across test collections [21].

ranked list in a session. Note that sDCG and other rank-based metrics rely on the linear traversal assumption.

We compute sDCG for a given session as follows. First, construct a single ranked list by (i) truncating each ranked list at the lowest clicked rank; (ii) concatenating the ranked lists. (We do not employ any special treatment for duplicate documents [14].) Then, sDCG is computed over this concatenated ranked list:

$$sDCG = \sum_r \frac{isrel(r)}{\log_4(querynum(r)+3)\log_2(r+1)} \quad (7)$$

where $isrel(r)$ is 1 if the document at rank $r$ in the concatenated list was *clicked* and 0 otherwise, and $querynum(r)$ returns $j$ when the document at $r$ comes from the ranked list for the $j$-th query in the session. In short, sDCG discounts relevant (or *clicked* in our case) documents not only by ranks, but also by the number of query reformulations that the user had to go through.

Figure 10 visualises the correlation between U and sDCG. As indicated in the figure, the Pearson's correlation (which takes into account the absolute scores) is .820, while kendall's $\tau$ (which compares the session rankings) is .600. **Example A** in this figure represents a session with extremely high U and sDCG values. This session actually contains two unrelated queries ("2000 Ranger 175" – a fishing boat, and "kfc in janesville wi"), but all but one of the clicks are with the former query: the user clicked 92 times to examine 39 unique documents. It is possible that (s)he was conducting a survey of the fishing boat. **Example B** is a session with 124 clicks for 4 porn queries. **Example C** represents a navigational information need: the queries are "check yahoo mail" and "check yahoo mail account," and the user clicked on the same yahoo mail page 12 times. With the first query, the user clicked the top ranked document (i.e. checked email) 11 times, which adds 11 to sDCG[12]; then, with the second query, the user clicked the top ranked document (i.e. the document at rank 2 in the concatenated list) once, which further adds $1/\log_4(2+3)\log_2(2+1) = .5435$ to sDCG. Thus $sDCG = 11.5435$. Whereas, owing to the diminishing return property, the value of a click on the yahoo mail page decays gradually with U: as the length of this page was estimated to be 539 characters, according to Figure 5, the decay value for the first click is $1-(200+0.2*539)/132000 = .9977$, while that for the eleventh click is $1-(200+11*0.2*539)/132000 = .9895$.

To sum up, click-based U is highly correlated with sDCG when used for multi-query session evaluation, but unlike sDCG, it has the diminishing return property similar to ERR [9], and can take document lengths into account. It is also possible to design trailtext-based metrics based on different browsing paths in a way similar to Kanoulas *et al.* [14], but this is beyond our current scope.

## 7. EVALUATING NONLINEAR TRAVERSAL

In this section, we demonstrate the potential of evaluating nonlinear traversals using click-based U. For this experiment, we first extracted a total of 2,015,311 sessions from our sample date (September 7, 2012, US market) that contained at least one nonlinear traversal somewhere in the session. From this set, we randomly sampled 50,000 sessions that contained at least one nonlinear traversal *before the first query reformulation*. Furthermore, to isolate the problem of evaluating nonlinear traversals from that of evaluating multi-query sessions, we *truncated* all records where $querynum > 1$, i.e. all clicks after the first query reformulation. The average and the

[12]Duplicate clicks can of course be filtered out, but we decided to reward every click in our experiment. In the case of **Example C**, note that the user may obtain new information every time he clicks on the yahoo mail page.

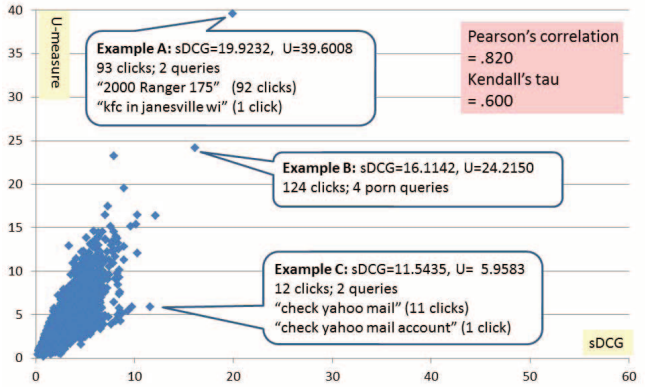


**Figure 10: Correlation between U and sDCG for the 50,000 multi-query sessions.**

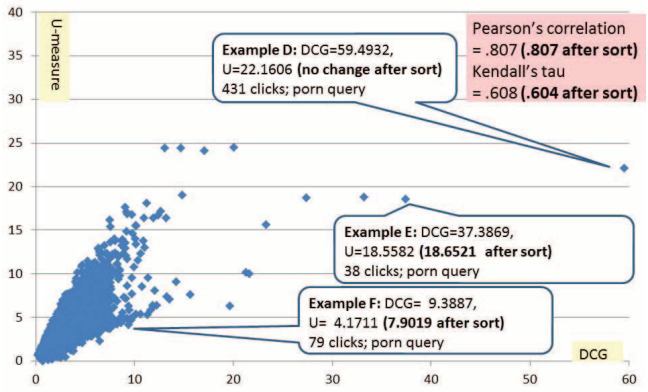


**Figure 11: Correlation between U and DCG for the 50,000 truncated nonlinear traversal sessions.**

maximum number of clicks per truncated session for this sample are 3.612 and 431, respectively. As we are now dealing with single ranked lists, sDCG reduces to the standard, binary-relevance version of DCG (Substitute $querynum(r) = 1$ to Eq. 7).

Figure 11 visualises the correlation between U and DCG. As shown in the figure, Pearson's correlation is .807, while Kendall's $\tau$ is .608: the values are remarkably similar to the multi-query case even though we are using a different sample here. In addition, we also computed U *after sorting each session data file by the clicked rank*, thereby obtaining artificial *linear* traversal sessions. Note that DCG cannot reflect this change as it simply discounts every clicked document based on the rank. As indicated in the figure, the effect of sorting on the correlation values is negligible, but in fact only 368 of the 50,000 sessions were unaffected by the sort in terms of U. Below, we discuss a few specific examples indicated in the figure.

**Example D** and **Example E** are porn queries with 431 and 38 clicks, respectively. For the latter, the value of U increases slightly after the sort. **Example F** is also a porn query, with 79 clicks, but we examine this closely as the value of U increased considerably from 4.1711 to 7.9019 after the sort. This user clicked the document at rank 426 (which was his ninth click), and then clicked one at rank 58 (his tenth click). Moreover, after clicking on the document at rank 402 (which was his 40th click), he clicked one at rank 399. Thus there were two nonlinear traversals within this truncated session. The document at rank 58 was the highest-ranked clicked document, so the artificial linear traversal file regards this as the first clicked document. This document happened to be very long: 20,044 characters. Thus, while the value of U for the original nonlinear traversal session reflects the user who examined nine rel-

atively short documents before reaching the long document at rank 58, the U for the artificial linear traversal session represents a user who had to read this long document first. This property of encouraging systems to return *relevant and concise* information first has been inherited from S-measure for summary evaluation [19]. While other variants of U are possible, we believe that handling nonliner traversals and different document lengths as we do is a useful step towards understanding the real users' search behaviour.

# 8. CONCLUSIONS AND FUTURE WORK

We introduced a general information access evaluation framework that can potentially handle summaries, ranked document lists and even multi-query sessions seamlessly. Our framework first builds a *trailtext* which represents a concatenation of all the texts read by the user during a search session, and then computes *U-measure* over the trailtext, based on *position*-based discounting. U-measure takes the document length into account just like TBG, and has the diminishing return property. It is therefore more realistic than rank-based metrics. Furthermore, it is arguably more flexible than TBG, as it is free from the *linear traversal* assumption, and can handle information access tasks other than ad hoc retrieval. Our main conclusions are: (a) For ad hoc retrieval, U-measure is at least as reliable as TBG in terms of rank correlations with traditional metrics and discriminative power; (b) For diversified search, our diversity versions of U-measure are highly correlated with state-of-the-art diversity metrics; (c) For multi-query sessions, U-measure is highly correlated with Session nDCG; and (d) Unlike rank-based metrics like DCG, U-measure can quantify the differences between linear and nonlinear traversals in sessions. We argue that our new framework is useful for understanding the user's search behaviour and for comparison across different information access styles (e.g. examining a direct answer vs. examining a ranked list of web pages).

Our future work includes the following: (1) Exploring setting and varying $F$ based on real search logs, while maintaining the simplicity of U; (2) Devising alternative decay functions for accommodating various types of information needs; (3) Comparing different information access styles, as was discussed above; (4) Combining the U-measure framework with eyetracking and/or information unit approaches, so that, for example, the difference between two search engines with similar DCG values but different snippet qualities can be quantified.

To make our work as reproducible as possible, we have made the following publicly available at `http://research.microsoft.com/u/` : (I) Length statistics for all relevant documents in the TREC 2005 Robust and 2011 Diversity data; (II) Multi-query and Nonlinear session records of the form *sessionID*, *querynum*, *clickedrank*, *doclen* (50,000 sessions each); and (III) Raw evaluation metric values from all experiments reported in this paper.

# 9. REFERENCES


[1] R. Agrawal, G. Sreenivas, A. Halverson, and S. Leong. Diversifying search results. In *Proceedings of ACM WSDM 2009*, pages 5–14, 2009.

[2] J. Allan, B. Carterette, and J. Lewis. When will information retrieval be "good enough"? In *Proceedings of ACM SIGIR 2005*, pages 433–440, 2005.

[3] P. Arvola, J. Kekäläinen, and M. Junkkari. Expected reading effort in focused retrieval evaluation. *Information Retrieval*, 13(5):460–484, 2010.

[4] L. Azzopardi. Usage based effectiveness measures. In *Proceedings of ACM CIKM 2009*, pages 631–640, 2009.

[5] F. Baskaya, H. Keskustalo, and K. Järvelin. Time drives interaction: Simulating sessions in diverse searching environments. In *Proceedings of ACM SIGIR 2012*, pages 105–114, 2012.

[6] B. Carterette. System effectiveness, user models, and user utility: A conceptual framework for investigation. In *Proceedings of ACM SIGIR 2011*, pages 903–912, 2011.

[7] B. Carterette. Multiple testing in statistical analysis of systems-based information retrieval experiments. *ACM TOIS*, 30(1), 2012.

[8] B. Carterette, E. Kanoulas, and E. Yilmaz. Simulating simple user behavior for system effectiveness evaluation. In *Proceedings of ACM CIKM 2011*, pages 611–620, 2011.

[9] O. Chapelle, S. Ji, C. Liao, E. Velipasaoglu, L. Lai, and S.-L. Wu. Intent-based diversification of web search results: Metrics and algorithms. *Information Retrieval*, 14(6):572–592, 2011.

[10] C. L. A. Clarke, N. Craswell, I. Soboroff, and A. Ashkan. A comparative analysis of cascade measures for novelty and diversity. In *Proceedings of ACM WSDM 2011*, pages 75–84, 2011.

[11] C. L. A. Clarke, N. Craswell, I. Soboroff, and E. M. Voorhees. Overview of the TREC 2011 web track. In *Proceedings of TREC 2011*, 2012.

[12] G. Dupret and B. Piwowarski. A user behavior model for average precision and its generalization to graded judgments. In *Proceedings of ACM SIGIR 2010*, pages 531–538, 2010.

[13] T. Joachims, L. Granka, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM TOIS*, 25(2), 2007.

[14] E. Kanoulas, B. Carterette, P. D. Clough, and M. Sanderson. Evaluating multi-query sessions. In *Proceedings of ACM SIGIR 2011*, pages 1053–1062, 2011.

[15] C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the ACL 2004 Workshop on Text Summarization Branches Out*, 2004.

[16] A. Nenkova, R. Passonneau, and K. McKeown. The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing*, 4(2):Article 4, 2007.

[17] S. Rajput, M. Ekstrand-Abueg, V. Pavlu, and J. Aslam. Constructing test collections by inferring document relevance via extracted relevant information. In *Proceedings of ACM CIKM 2012*, pages 145–154, 2012.

[18] T. Sakai. Evaluating evaluation metrics based on the bootstrap. In *Proceedings of ACM SIGIR 2006*, pages 525–532, 2006.

[19] T. Sakai, M. P. Kato, and Y.-I. Song. Click the search button and be happy: Evaluating direct and immediate information access. In *Proceedings of ACM CIKM 2011*, pages 621–630, 2011.

[20] T. Sakai and R. Song. Evaluating diversified search results using per-intent graded relevance. In *Proceedings of ACM SIGIR 2011*, pages 1043–1042, 2011.

[21] T. Sakai and R. Song. Diversified search evaluation: Lessons from the NTCIR-9 INTENT task. *Information Retrieval*, 2013.

[22] M. D. Smucker and C. L. A. Clarke. Modeling user variance in time-biased gain. In *Proceedings of HCIR 2012*, 2012.

[23] M. D. Smucker and C. L. A. Clarke. Time-based calibration of effectiveness measures. In *Proceedings of ACM SIGIR 2012*, pages 95–104, 2012.

[24] A. Turpin, F. Scholer, K. Järvelin, M. Wu, and J. S. Culpepper. Including summaries in system evaluation. In *Proceedings of ACM SIGIR 2009*, pages 508–515, 2009.

[25] E. M. Voorhees. Overview of the TREC 2005 robust retrieval track. In *Proceeings of TREC 2005*, 2006.

[26] Y. Yang and A. Lad. Modeling expected utility of multi-session information distillation. In *Proceedings of ICTIR 2009 (LNCS 5766)*, pages 164–175, 2009.

[27] E. Yilmaz, J. Aslam, and S. Robertson. A new rank correlation coefficient for information retrieval. In *Proceedings of ACM SIGIR 2008*, pages 587–594, 2008.

[28] E. Yilmaz, M. Shokouhi, N. Craswell, and S. Robertson. Expected browsing utility for web search evaluation. In *Proceedings of ACM CIKM 2010*, pages 1561–1564, 2010.

[29] K. Zhou, R. Cummins, M. Lalmas, and J. M. Jose. Evaluating aggregated search pages. In *Proceedings of ACM SIGIR 2012*, pages 115–124, 2012.