

Metrics, Statistics, Tests

Tetsuya Sakai¹

Waseda University, Japan
tetsuyasakai@acm.org

Abstract. This lecture is intended to serve as an introduction to Information Retrieval (IR) effectiveness metrics and their usage in IR experiments using test collections. Evaluation metrics are important because they are inexpensive tools for monitoring technological advances. This lecture covers a wide variety of IR metrics (except for those designed for XML retrieval, as there is a separate lecture dedicated to this topic) and discusses some methods for evaluating evaluation metrics. It also briefly covers computer-based statistical significance testing. The takeaways for IR experimenters are: (1) It is important to understand the properties of IR metrics and choose or design appropriate ones for the task at hand; (2) Computer-based statistical significance tests are simple and useful, although statistical significance does not necessarily imply practical significance, and statistical insignificance does not necessarily imply practical insignificance; and (3) Several methods exist for discussing which metrics are “good,” although none of them is perfect.

1 Introduction

This lecture is intended to serve as an introduction to Information Retrieval (IR) effectiveness metrics and their usage in IR experiments using test collections. Evaluation metrics are important because they are inexpensive tools for monitoring technological advances. Forty years ago, Cooper [36, 37] said: *“the best way to evaluate a retrieval system is, in principle at least, to elicit subjective estimates of the system’s utility to its users, quantified in terms of the number of utiles (e.g. dollars) they would have been willing to give up in exchange for the privilege of using the system.”* He also described this hypothetical evaluation scheme as follows: *“The system users in the sample are chosen at random from among the patrons as they enter the library and are about to make use of the retrieval system.”* Now in the 21st Century, it is very difficult to find “the users in the library,” observe them and ask them questions!

Sections 2 and 3 define and discuss “traditional” and “advanced” IR metrics, respectively. By traditional metrics, I mean those designed for evaluating a set of items or a ranked list of items based on relevance. By advanced metrics, I mean those designed for handling diversity, multi-query sessions, and IR systems that go beyond the ranked-list paradigm. (This lecture does not cover evaluation metrics specifically designed for XML retrieval, as there is a separate lecture dedicated to this topic.) Section 4 briefly describes computer-based statistical

significance tests that are useful for IR evaluation. Section 5 discusses tests for “evaluating evaluation metrics”: one ultimate goal of IR researchers is to build systems that completely and efficiently satisfy the user’s information needs, and we often regard evaluation metrics as crude indicators of user satisfaction or user performance. What are “good” metrics? Finally, Section 6 summarises this lecture.

A word of warning: in this lecture, I will present my personal views on IR effectiveness metrics and on methods for evaluating evaluation metrics. I discuss a lot of my own work because I know a lot about it. Hence I encourage the reader to go back to the original papers listed up in the references.

2 Traditional IR Metrics

Historically, IR was about *set retrieval*: should each document be retrieved or not? Section 2.1 describes some basic evaluation metrics for set retrieval, including the widely-used *recall*, *precision* and *F-measure* [68]. But with the advent of the digital information overload era, *ranked retrieval* has become the norm, so that the user can examine retrieved documents sequentially from the top and stop at her convenience. Section 2.2 describes a wide range of evaluation metrics for ranked retrieval, including *normalised Discounted Cumulative Gain* [49] (nDCG) which has been used widely not only in the IR research community but also for tuning commercial web search engines. These “traditional” set retrieval and ranked retrieval metrics require a gold standard (i.e. “right answers”): to be more specific, for each *search topic* (or *query*), a set of *relevant* documents is required. Note that “document” is a generic term that may refer to any *retrieval unit*: for example, it could be a web page, a textual passage, a multimedia file, a cluster of items and so on. Section 2.3 provides information for further reading.

2.1 Set Retrieval Metrics

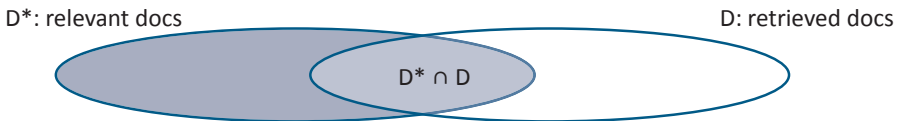


Fig. 1. Relevant/retrieved documents.

Recall and Precision Figure 1 is a Venn diagram that shows a set of relevant documents for a topic (D^*), a set of retrieved documents for that topic (D), and the intersection between the two ($D^* \cap D$). $D^* - D$ represents the documents that the retrieval system missed, while $D - D^*$ represents the nonrelevant documents retrieved. *Recall* (Rec) and *Precision* ($Prec$) directly reflect these properties, respectively: $Rec = |D^* \cap D|/|D^*|$, and $Prec = |D^* \cap D|/|D|$.

E-measure While it is clear that recall and precision have a trade-off relationship, we generally want both high recall and high precision. It would be useful to have a single, summary metric that incorporates this trade off. Let us first start with a basic version of *E-measure* [68], using Figure 1:

$$E\text{-measure} = \frac{|D^* \cup D| - |D^* \cap D|}{|D^*| + |D|}. \quad (1)$$

Using the aforementioned definitions of recall and precision, the above can alternatively be expressed as:

$$E\text{-measure} = 1 - \frac{1}{0.5 \frac{1}{\text{Prec}} + 0.5 \frac{1}{\text{Rec}}}. \quad (2)$$

But now it is clear that this version of E-measure assumes that recall and precision are equally important; let us generalise it by introducing a parameter α ($0 \leq \alpha \leq 1$):

$$E\text{-measure} = 1 - \frac{1}{\alpha \frac{1}{\text{Prec}} + (1 - \alpha) \frac{1}{\text{Rec}}}. \quad (3)$$

Furthermore, by letting $\alpha = 1/(\beta^2 + 1)$, the generalised E-measure can be rewritten as:

$$E\text{-measure} = 1 - \frac{(\beta^2 + 1)\text{PrecRec}}{\beta^2\text{Prec} + \text{Rec}}. \quad (4)$$

Here, the assumption is that the user attaches $\beta (\geq 0)$ times as much importance to recall as precision¹.

F-measure *F-measure* [28], which is simply *one minus E-measure*, is much more widely used than E-measure, probably because we want the evaluation metric value to be large for an effective retrieval system:

$$F\text{-measure} = \frac{(\beta^2 + 1)\text{PrecRec}}{\beta^2\text{Prec} + \text{Rec}}. \quad (5)$$

F-measure with $\beta = b$ is often expressed as F_b ; note that F_1 is a harmonic mean of precision and recall.

2.2 Ranked Retrieval Metrics

nDCG *Normalised Discounted Cumulative Gain* [49] (nDCG) has become one of the most widely-used evaluation metric for traditional ranked retrieval over the past decade. It is similar to a metric from the 1960s called the *Normalised Sliding Ratio* [67] (NSR), and handles *graded* relevance assessments unlike many other metrics that were used earlier in the IR community. For example, a topic may have some judged nonrelevant documents (relevance level 0), some partially

¹ $\frac{dE}{d\text{Rec}} = \frac{dE}{d\text{Prec}}$ when $\frac{\text{Prec}}{\text{Rec}} = \beta$ [68].

relevant documents (relevance level 1) and highly relevant documents (relevance level 2). We decide in advance the *gain value* gv_x for each relevance level x : for example, we could simply let $gv_1 = 1$, $gv_2 = 2$, and $gv_3 = 3$, by assuming that the raw value of each relevant document is proportional to its relevance level. Also, it is common to let $gv_0 = 0$: a nonrelevant document is of no value.

For a given ranked list of documents, let $g(r) = gv_x$ if the relevance level of the document at rank r is x . In particular, let $g^*(r)$ denote the gain value at rank r of an *ideal list*², obtained by sorting all relevant documents in decreasing order of the relevance level. A few versions of nDCG exist, but the one described here [17] is probably the most widely-used:

$$nDCG = \frac{\sum_{r=1}^l g(r) / \log(r+1)}{\sum_{r=1}^l g^*(r) / \log(r+1)} \quad (6)$$

where l is the *measurement depth*, also known as the *document cutoff*. Note that the logarithm base b cancels out in the above definition: for convenience let us use $b = 2$ here. The key feature of nDCG is that the gain value of each retrieved relevant document is discounted based on its rank: for example, if we set the gain value of each highly relevant document to be 3, then for a highly relevant document at rank 1, its discounted gain is $3 / \log(1+1) = 3$; but for a highly relevant document at rank 7, its discounted gain is $3 / \log(1+7) = 1$.

The use of the *original* nDCG, which regards the logarithm base b as a *user patience parameter* [49], is not recommended. The problem is that discounting is not applied when $r \leq b$. For example, when $b = 10$, this version of nDCG cannot tell the difference between a system that returns a relevant document at rank 1 and one that returns a relevant document at rank 10. To address this, Järvelin *et al.* [50] have described yet another version of nDCG, which discounts the raw gain by $1 + \log_b r$ instead of $\log(1+r)$.

11-point Average Precision This binary-relevance metric is a single-value summary of the *recall-precision curve* [108], but has been replaced in the early 1990s by (noninterpolated) *Average Precision*, which is described next. Although 11-point Average Precision is no longer popular, how to draw a recall-precision curve is perhaps still worth mentioning here.

Figure 2 shows how to compute *interpolated precision* for 11 recall points. In this example, the number of known relevant documents is five, and the system has managed to retrieve four of them. The recall ($Rec(r)$) and the precision ($Prec(r)$) at each rank r are shown on the left. For each recall point $i (\in \{0, 0.1, \dots, 1\})$, interpolated precision is given by:

$$IP_i = \max_{r, Rec(r) \geq i} Prec(r) . \quad (7)$$

That is, for a given recall point i , the actual recall values that satisfy this level are first obtained, and then the highest precision value among these actual recall points is obtained.

² Pollock, who proposed NSR in 1968, called it the *master list* [67].

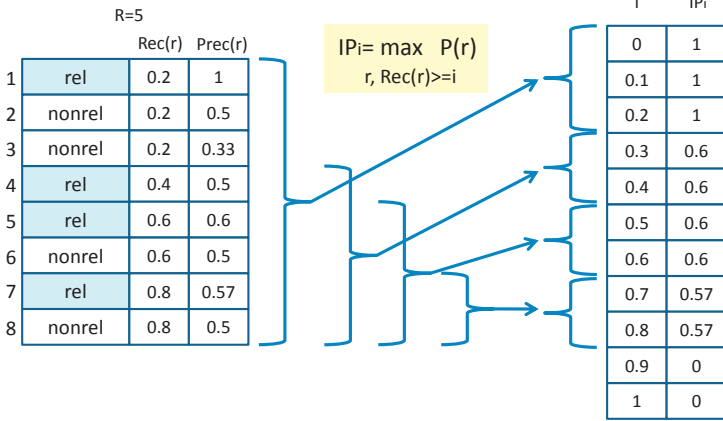


Fig. 2. Computing interpolated precision for the 11 recall points.

The recall-precision curve is obtained by plotting the interpolated precision value for each i . Moreover, 11-point average precision is simply given by:

$$11pt-AP = \frac{\sum_{i \in \{0, 0.1, \dots, 1\}} IP_i}{11}. \quad (8)$$

This averaging is not desirable for many IR applications, as the precisions at low recall points and those at high recall points are considered equally important.

Average Precision *Average Precision* (AP) was one of the most widely-used evaluation metric for ranked retrieval during the 1990s, since it was introduced at the Second Text Retrieval Conference (TREC-2) [108]. Let R denote the number of known relevant documents for a topic. For a given ranked list of documents, let $I(r)$ be 0 if the document at rank r is nonrelevant, and 1 otherwise. Let $C(r) = \sum_{k=1}^r I(k)$: this is the number of relevant documents within top r . Hence the precision at r is given by $Prec(r) = C(r)/r$. Then AP is defined as:

$$AP = \frac{1}{R} \sum_r I(r) Prec(r) = \frac{1}{R} \sum_r I(r) \frac{C(r)}{r}. \quad (9)$$

One of the strengths of AP over 11-point average precision and other metrics is that it is *top heavy*: that is, it is sensitive to changes near the top ranks. For example, suppose that, through a system improvement, a relevant document has moved up by one rank from rank 2 to 1. Before this improvement, this document contributes a precision of 0.5 to AP; after the improvement, it contributes a precision of 1. In contrast, suppose a relevant document has moved from rank 100 to 99 (and that there is no other relevant document in the ranked list). This has little impact on AP, as the contributed precisions are $1/100 = 0.0100$ and $1/99 = 0.0101$, respectively.

Robertson [71] provided a user model for AP. There is a user population, and all users scan the ranked list from top to bottom, but different users stop

scanning the list at different relevant documents (probably due to satisfaction). In AP, this probability distribution is assumed to be uniform across all relevant documents: that is, the probability that the user stops at each relevant document is $1/R$. Moreover, for each stopping point r , AP measures the *utility* of the top r documents in terms of precision $Prec(r)$. Hence, AP can be regarded as the *expected utility* for the user population.

The above formulation of AP and its user model assume that the document ranking is infinite, which may seem unrealistic. For those who want to use a small measurement depth l , the following variant of AP may be used:

$$AP = \frac{1}{\min(l, R)} \sum_r I(r) Prec(r) . \quad (10)$$

This ensures that the maximum possible AP is 1 even if $l < R$. Moreover, the user’s stopping probability distribution can now be interpreted as either uniform over all relevant documents (if $l \geq R$) or uniform over the first l retrieved relevant documents (if $l < R$).

Unlike nDCG, AP cannot handle graded relevance. While the use of binary-relevance metrics such as AP is still common in the IR community, it should be noted that, with such metrics, it is impossible to design retrieval systems that can retrieve, say, highly relevant documents before marginally relevant ones. In light of this, several graded-relevance versions of AP have been proposed. One of them is called *Q-measure* [75, 74] (or simply “Q”), which is discussed below. *Graded Average Precision* (GAP) [73] is a more recently-proposed alternative, which we shall omit in this paper as it is a little more complex than others. In contrast to Q which combines the ideas of nDCG and AP, GAP is based on a novel interpretation of graded relevance: more specifically, it assumes that the user has a binary notion of relevance, but that different users have different thresholds over the relevance levels. Sakai and Song [94] have compared Q and GAP in terms of *discriminative power* [77] (discussed in Section 5.1) and reported that Q outperformed GAP in some cases. In an earlier study, Sakai [80] compared Q with nDCG and Kishida’s *generalised Average Precision* [57] (gAP), yet another graded-relevance version of AP, and demonstrated the advantage of Q’s *user persistence parameter*, which gAP lacks.

Q-measure *Q-measure* [75, 74], a graded-relevance version of AP, replaces the precision $Prec(r)$ with the *blended ratio* $BR(r)$ which can handle graded relevance. Let $cg(r) = \sum_{k=1}^r g(k)$ and $cg^*(r) = \sum_{k=1}^r g^*(k)$: these are the (nondiscounted) *cumulative gains* [49] for the ranked list to be evaluated and for the ideal list, respectively. Then, for a given value of the *user persistence parameter* $\beta (\geq 0)$:

$$BR(r) = \frac{C(r) + \beta cg(r)}{r + \beta cg^*(r)} . \quad (11)$$

$BR(r)$ inherits the properties of $Prec(r) = C(r)/r$ and the *normalised Cumulative Gain* [49] $nCG(r) = cg(r)/cg^*(r)$. Moreover, in a binary-relevance eval-

uation environment (regardless of β), it is easy to prove that $BR(r) = Prec(r)$ holds if and only if $r \leq R$, while $BR(r) > Prec(r)$ holds if and only if $r > R$.

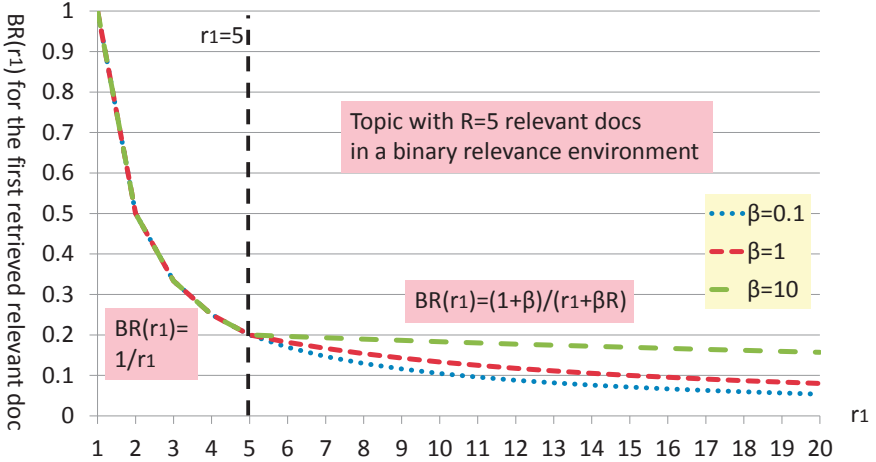


Fig. 3. Effect of β on $BR(r_1)$ in a binary relevance environment ($R = 5$).

Figure 3 illustrates the role of β for a topic with $R = 5$ relevant documents in a binary relevance environment. Here, the x axis represents r_1 , the rank of the *first* relevant document found in the ranked list; the y axis represents the value of $BR(r_1)$. In a binary relevance environment, since $BR(r) = Prec(r)$ holds for $r \leq R$, note that $BR(r_1) = 1/r_1$ holds for $r_1 \leq R$. On the other hand, in a binary relevance environment, it is easy to show that $BR(r_1) = (1 + \beta)/(r_1 + \beta R)$ for $r_1 > R$. It can be observed from the figure that a large β represents a user who is very tolerant to relevant documents retrieved at low ranks; In practice, β is often set to 1, although this is an arbitrary choice.

Q can be defined as follows:

$$Q\text{-measure} = \frac{1}{R} \sum_r I(r) BR(r) = \frac{1}{R} \sum_r I(r) \frac{C(r) + \beta cg(r)}{r + \beta cg^*(r)} \quad (12)$$

or, for a given measurement depth l ,

$$Q\text{-measure} = \frac{1}{\min(l, R)} \sum_r I(r) BR(r) . \quad (13)$$

Following Robertson’s interpretation of AP [71], Q can be regarded as an evaluation metric which (a) assumes, just like AP, that the user’s stopping probability distribution is uniform over all (or l) relevant documents; and (b) measures the utility at a given stopping rank in terms of the blended ratio [92]. Also, it is clear that Q reduces to AP when $\beta = 0$.

While Q is not as widely-used as nDCG, it has been used as one of the official metrics in the NTCIR Crosslingual IR (CLIR) task [58], Advanced Crosslingual

Information Access (ACLIA) task [93] and the Geotemporal Information Retrieval (GeoTime) task [45].

Sakai and Robertson [92] have explored a few extensions of the above interpretation: they considered non-uniform stopping probability distributions, namely, a distribution based on relevant documents seen so far, and a distribution that takes the relevance levels into account. The family of these metrics is collectively known as *Normalised Cumulative Utility* (NCU).

R-precision, R-measure *R-precision* [108] is a binary-relevance, early-TREC metric, defined for each topic with R relevant documents as $R\text{-prec} = \text{Prec}(R)$. That is, this is the precision (or equivalently, recall) at the measurement depth of R . Similarly, *R-measure* [75, 74], a variant of Q, is defined as $R\text{-measure} = BR(R)$. These metrics can be regarded as a type of NCU where *all* users stop scanning the ranked list at rank R . Although R-measure leverages graded relevance, it gives a score of one to any system as long as the top R documents are all relevant, even if marginally relevant documents are ranked above the highly relevant ones.

RR The basic assumption behind all of the above ranked retrieval metrics is that the user wants as many relevant documents as possible. While they may be suitable for *informational* search intents, there are also *navigational* search intents [14], which basically require just one document: in this case, we can assume that retrieving multiple relevant documents do not help the user.

Reciprocal Rank (RR) is a metric suitable for navigational intents. For a ranked list that does not contain a relevant document, we let $RR = 0$. Otherwise, let r_1 be the rank of the first relevant document in the ranked list: then $RR = 1/r_1$.

RR can also be seen as a member of the aforementioned NCU family: it is assumed that *all* users stop at rank r_1 , and the utility at rank r_1 is measured by precision: $\text{Prec}(r_1) = C(r_1)/r_1 = 1/r_1$. Just like AP, it cannot handle graded relevance.

P⁺ There are a few graded-relevance versions of RR: here, we discuss P⁺ [76], which is a variant of Q and therefore a member of the NCU family. For a ranked list that does not contain a relevant document, we let $P^+ = 0$. Otherwise, let r_p be the rank of the document that is highest-ranked among the *most relevant* documents within the measurement depth l . For example, if a ranked list contains a marginally relevant document at rank 2, a highly relevant document at rank 4 and another highly relevant document at rank 6, then $r_p = 4$. (Whereas, note that $r_1 = 2$.) Then P⁺ is defined as:

$$P^+ = \frac{1}{C(r_p)} \sum_{r=1}^{r_p} I(r)BR(r) . \quad (14)$$

Thus, P^+ is an NCU metric that (a) assumes that the user’s stopping probability distribution is uniform over the top $C(r_p)$ relevant documents, i.e., all relevant documents at or above r_p ; and (b) measures the utility at a given stopping rank in terms of the blended ratio just like Q .

Sakai [76] have discussed the advantages of P^+ over other graded-relevance versions of RR such as *Weighted Reciprocal Rank* (WRR) [44], *P-measure* (defined as $BR(r_p)$) and *O-measure* (defined as $BR(r_1)$). While P^+ itself is not a well-known metric, together with Q , it forms the basis of another metric for evaluating diversified search called $P+Q$, which we shall discuss in Section 3.1. Probably the most well-known graded-relevance metric that is suitable for navigational intents is *Expected Reciprocal Rank* [27] (ERR), which we shall discuss next.

ERR Let $Pr(r)$ denote the probability that the user is satisfied at a document at rank r . ERR assumes that the user stops scanning the ranked list as soon as she is satisfied with a document, and that this satisfaction probability depends directly and solely on the relevance level of each document. For example, we can assume that $Pr(r) = 0$ if the document at r is nonrelevant; if we have marginally relevant, partially relevant and highly relevant documents (i.e. three relevance levels), we may let $Pr(r)$ be $(2^1 - 1)/2^3 = 1/8$, $(2^2 - 1)/2^3 = 3/8$ and $(2^3 - 1)/2^3 = 7/8$, respectively [27]. Under the *linear traversal* assumption (i.e. the user scans the list from top to bottom), the probability that the user is still unsatisfied at rank r is given by $dsat(r) = \prod_{k=1}^r (1 - Pr(k))$. ERR is then given by:

$$ERR = \sum_r dsat(r-1)Pr(r)\frac{1}{r} . \quad (15)$$

ERR can also be regarded as an instance of NCU, which (a) assumes that the user’s stopping probability over ranks is given by $dsat(r-1)Pr(r)$, i.e. the probability that the user is dissatisfied with all documents between ranks 1 and $r-1$ and finally satisfied at r ; and (b) uses the RR at r to measure the utility. Note that RR is used rather than precision, since the document at r is considered to be the only useful one.

What distinguishes ERR from most of the other metrics discussed so far is its *diminishing return* property [26]: whenever a relevant document is found, the value of another relevant document found later in the list is discounted. For example, given the three-level probability setting as described above, the stopping probability for a highly relevant document at rank 2 would be $(1 - 0) * 7/8 = 0.8750$ if the document at rank 1 is nonrelevant; but it would be $(1 - 7/8) * 7/8 = 0.1094$ if the document at rank 1 is also highly relevant. The interpretation is that the second highly relevant document in the latter case is *redundant*, which aligns well with the definition of a navigational intent. This property is in contrast with other metrics such as nDCG and Q that discount the value of each relevant document based solely on its rank. Sakai and Robertson [92] have described another NCU metric that also possesses the diminishing return property: the stopping probability distribution of their metric is designed based

on the assumption that “it is probably more likely that a user would stop after few relevant documents than after many.”

Another interesting feature of ERR is that it does not have a recall component, unlike other graded-relevance metrics such as Q and nDCG: note that even though nDCG does not directly depend on R , the number of relevant documents, it can still be regarded as a *recall-dependent* metric as it relies on an ideal ranked list which requires enumeration of relevant documents³. ERR, on the other hand, does not rely on the notion of ideal list, and is not normalised.

RBP *Rank-Biased Precision* [63] (RBP) is another recall-independent metric with a clear user model: like all other metrics discussed so far, the model assumes linear traversal, and furthermore assumes that, after the user examines rank r , she will either move on to rank $(r+1)$ with probability p or stop scanning the list with probability $1-p$. The user behaves this way irrespective of the relevance of the documents, and p is a constant. This p can be regarded as a user persistence parameter: the higher p is, the more persistent she is.

RBP can handle graded relevance, with gain values $g(r) = gv_x$ set within the 0-1 range. It can be expressed as:

$$RBP = (1-p) \sum_r p^{r-1} g(r) . \quad (16)$$

Note that RBP discounts the value of a relevant document based solely on the rank, just like other metrics such as nDCG and Q. Thus, unlike ERR, it does not possess the diminishing return property.

While Moffat and Zobel [63] have discussed the strengths of RBP such as its recall-independence, Sakai and Kando [89] have demonstrated a few of its shortcomings: for example, the maximum possible value of RBP varies widely depending on the parameter p ⁴; RBP has low *discriminative power* [77] (discussed in Section 5.1).

TBG The metrics discussed so far treat a ranked list of documents as if they are just a list of document IDs with relevance levels. In modern IR contexts such as web search, however, the user often examines snippets (a.k.a. summaries) before reading the actual documents, and the document lengths vary. In light of this, Smucker and Clarke [104] have proposed to use the *time spent by the user* as the basis for discounting the value of a document instead the document rank.

While the general framework TBG proposes to accumulate the gains of relevant documents over *time*, the instantiation of TBG discussed by Smucker and Clarke [104] actually performs a *rank-based* gain accumulation as follows:

$$TBG = \sum_r g(r) \exp(-T(r) \frac{\ln 2}{h}) \quad (17)$$

³ Of course, we also have *Discounted Cumulative Gain* (DCG) [49], which is not normalised.

⁴ In a binary relevance environment, the maximum RBP for a topic with R relevant documents is given by $(1-p) \sum_{r=1}^R p^{r-1}$.

where $T(r)$ is the *expected time to reach rank r* and h is the half-life for the time-based decay (i.e. discounting) function. This instantiation of TBG [104] is based on binary relevance: the gain value $g(r)$ for every relevant document is estimated as the probability of click on a relevant summary times the probability of judging the actual document as relevant; that for every nonrelevant document is zero. As for $T(r)$, let T_S be the time to read any summary in seconds; let $L(r)$ be the length of the document at r in terms of the number of words; and let $Pr_{click}(r)$ be the probability of click at r , which depends on whether the document at r is relevant or nonrelevant. Then $T(r)$ is estimated as:

$$T(r) = \sum_{k=1}^{r-1} (T_S + Pr_{click}(k)T_D(k)) \quad (18)$$

where $T_D(r) = 0.018L(r) + 7.8$ is the estimated time to read the document at r .

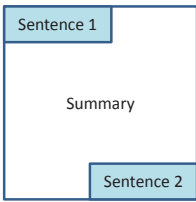
As the summation over previous ranks in Eq. 18 shows, TBG relies on the linear traversal assumption. Moreover, as the formula for $T_D(r)$ shows, TBG further assumes that the document reading time grows linearly with the document length.

It is of note that TBG as defined above does not guarantee diminishing return, even though it discounts documents by taking relevance into account. Suppose we have a nonrelevant document at rank 1, and a relevant document at rank 2, and imagine that the nonrelevant document at rank 1, whose document length is 1000 words, is replaced with a new *relevant* document whose length is 10 words. Moreover, following the calibration results from Smucker and Clarke [104], let $Pr_{click}(r) = 0.64$ if the document at r is relevant and $Pr_{click}(r) = 0.39$ otherwise. Then, according to Eq. 18, the time to reach the relevant document at rank 2 *before* the replacement is $T(2) = T_S + 0.39 * (0.018 * 1000 + 7.8) = T_S + 10.062$, while the corresponding time *after* the replacement is $T(2) = T_S + 0.64 * (0.018 * 10 + 7.8) = T_S + 5.107$. Thus, by replacing a long nonrelevant document at rank 1 with a short relevant document, the time required to reach rank 2 has *decreased*, which means that the relevant document at 2 receives *more* weight according to the exponential decay function in Eq. 17. On the other hand, if the document length variance is relatively small, we can expect TBG to follow the diminishing return pattern most of the time.

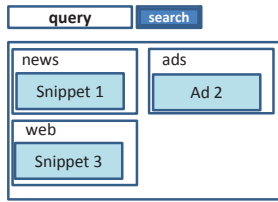
Smucker and Clarke [102, 103] have extended their TBG ideas in the context of stochastic simulation of user behaviours.

Prior to the proposal of TBG, Turpin *et al.* [105] and Yilmaz *et al.* [117] have also explored incorporating the snippet examination phase into IR evaluation. Several forms of time-based evaluation have also been proposed previously: for example, Dunlop [42] proposed a time-based evaluation method based on Cooper's *expected search length* [35].

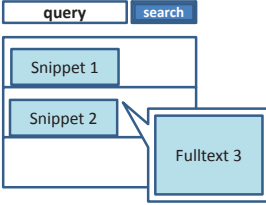
U-measure Sakai and Dou [85] recently proposed a general information access evaluation framework that can potentially handle not only ranked retrieval discussed here but also summaries, diversified search, multi-query sessions etc. that



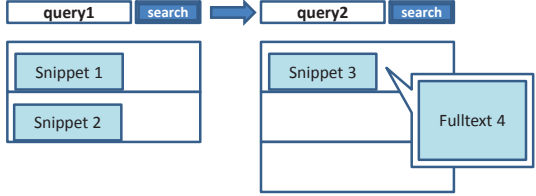
(a) Reading a summary:
Sentence 1 → Sentence 2



(b) Browsing an aggregated search output:
Snippet 1 → Ad 2 → Snippet 3



(c) Scanning a ranked list:
Snippet 1 → Snippet 2 → Fulltext 3



(d) Scanning multiple ranked lists in a session:
Snippet 1 → Snippet 2 → Snippet 3 → Fulltext 4

Fig. 4. Constructing trailtexts for various information access tasks.

will be discussed in Section 3. Their *U-measure* framework is similar to TBG in that it takes document lengths into account, but unlike TBG (as instantiated by Smucker and Clarke [104]), it does not depend on the linear traversal assumption.

Figure 4 illustrates the construction of *trailtext*, which represents all the text the user has read during an information seeking process. This could be obtained from direct user observation with eyetracking, or from user behaviour models with relevance assessments or click data. Since we are now discussing ranked retrieval, let us focus on Part (c) of this figure: here, the user scans a search result page, reads the first snippet, reads the second snippet and then visits the full text of the second document. The trailtext in this case is represented as a concatenation of these texts: “Snippet 1 Snippet 2 Fulltext 3.” The key idea of the U-measure framework is to define an evaluation metric over the trailtext rather than document ranks, so that any textual information seeking activities may be evaluated on a common ground.

Formally, a trailtext tt is a concatenation of n strings: $tt = s_1 s_2 \dots s_n$. These strings may be documents, parts of documents, snippets, sentences, or any other fragments of text that have been read. We define the *offset position* of s_k ($1 \leq k \leq n$) as $pos(s_k) = \sum_{j=1}^k |s_j|$ where the length of each string is measured in terms of the number of characters. Furthermore, we define the *position-based gain* as $g(pos(s_k)) = 0$ if s_k is considered nonrelevant and $g(pos(s_k)) = gv_x$ if its relevance level is considered to be x . Then U-measure is given by:

$$U\text{-measure} = \frac{1}{\mathcal{N}} \sum_{pos=1}^{|tt|} g(pos)D(pos) \quad (19)$$

where \mathcal{N} is a normalisation factor, which is set to zero if normalisation is not required. Here, $D(pos)$ is a position-based decay function, which may be defined as:

$$D(pos) = \max(0, 1 - \frac{pos}{L}) \quad (20)$$

where $L (> 0)$ is a parameter, which represents the amount of text read at which all relevant pieces of information become worthless for any user (set to $L = 132000$ by Sakai and Dou [85] based on web search session data). While an exponential decay function like the one used with TBG (Eq. 17) is also possible, the above simple linear function has been inherited from *S-measure* [91] which was proposed for summary evaluation. S-measure will be discussed in Section 3.3.

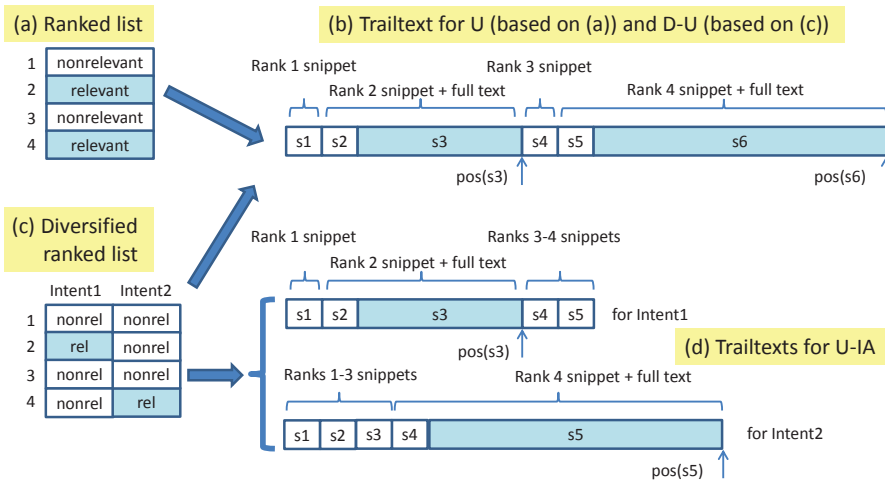


Fig. 5. Automatically constructing trailtexts from relevance assessments of traditional and diversified IR test collections.

Figure 5 Part (a) shows a ranked list where the documents at ranks 2 and 4 are known to be relevant; Part (b) shows a possible trailtext for this list, under the linear traversal assumption. It is assumed that the four snippets plus the two relevant documents are read. In practice, it is assumed that only $F\%$ of every relevant document is read; $F = 20$ has been shown to be a reasonable choice [85].

Like ERR, U-measure possesses the diminishing return property. Suppose that, in Part (a), the nonrelevant document at rank 3 is replaced by a relevant document. Then, since it is now assumed that the document at rank 3 is also read, the trailtext shown in Part (b) will be longer, and the fourth document is pushed back towards the end of the trailtext. That is, the gain value for the fourth document has diminished.

	nDCG	AP	Q	P+	ERR	RBP	TBG	U
(a) Graded relevance								
(b) Normalised								
(c) Recall-independent								
(d) Discriminative power								
(e) Diminishing return								
(f) Snippet & doc length								
(g) Nonlinear traversal								

Fig. 6. Comparison of traditional ranked retrieval metrics.

Ranked Retrieval Metrics: Summary Figure 6 provides a quick summary of the properties of the traditional ranked retrieval metrics. Some additional comments:

(a) The original AP cannot handle graded relevance, but a few graded-relevance versions exist (e.g. [57, 73]). TBG as described by Smucker and Clarke [104] is binary-relevance-based.

and (c) These properties are two sides of the same coin. nDCG, Q and P⁺ depend on an ideal ranked list, which requires the enumeration of all known relevant documents. AP and Q depend directly on the number of relevant documents. On the other hand, ERR, RBP, TBG and U are unnormalized (See the discussion on normalization below).

(d) nDCG, AP and Q are top-heavy metrics suitable for informational search intents, as they have been designed to consider many relevant documents. Hence, in terms of discriminative power (discussed in Section 5.1), they outperform other metrics such as P⁺ (See Sakai [76]), ERR (See Sakai and Song [94]), RBP (See Sakai and Kando [89]), TBG (See Smucker and Clarke [104] and Sakai and Dou [85]), and U (See Sakai and Dou [85]).

(e) ERR and U possess the diminishing return property, which is intuitive. TBG also shows this property unless the document lengths do not vary wildly. Diminishing return means that when a relevant document is found, the value of the next relevant document diminishes: this generally has a negative impact on discriminative power (See (d)).

(f) Besides TBG and U, a few other studies have considered the user’s snippet reading behaviour (e.g. [105, 117]). But only TBG and U consider the document length.

(g) The instantiation of TBG as described by Smucker and Clarke [104] depends on the time to reach rank r . This relies on the linear traversal assumption. Sakai and Dou [85] have demonstrated that U can quantify the difference

between linear and nonlinear traversals in the context of click-based web search evaluation where click timestamps are available (See Section 3.2).

Normalisation and Averaging Given a test collection with a topic set and relevance assessments for each topic, it is common to discuss the *arithmetic* mean of an evaluation metric over the topic set. For this purpose, *normalised* metrics, that range fully between 0 and 1, are convenient. Normalising before averaging implies that every topic is of equal importance, while not normalising sometimes implies that every user effort (e.g. finding one relevant document) is of equal importance. When using unnormalised metrics, researchers should be aware that the upperbound is different for every topic, and that topics with certain properties (e.g. those with many relevant documents) may heavily influence the mean.

A useful alternative to the arithmetic mean is the *geometric* mean: for example, while the arithmetic mean of AP is known as *MAP* (Mean AP), the geometric mean version is known as *GMAP* [70]. Taking a geometric mean is equivalent to taking the log of the metric for each topic and then taking the arithmetic mean, thereby emphasising the lower end of the metric scale. Thus this is useful for examining poor retrieval performance.

Condensed-list Metrics Many modern large-scale test collections were built based on *pooling* [106, 108], and therefore the relevance assessments are *incomplete* [16]: the target corpus probably contains more relevant documents that have never been assessed. Formally, let \mathcal{D} denote the target corpus, and for a particular topic, let C_j denote the *contributions* (e.g. top-100 retrieved documents) from the j -th *contributor* to the test collection (e.g. a TREC participant). Then, the pool for this topic is given by $P = \bigcup_j C_j$, where $|P| \ll |\mathcal{D}|$, and the documents in $\mathcal{D} - P$ are never judged for this topic (See also Section 5.4). Moreover, the incomplete relevance assessments may also be *biased* towards particular types of relevant documents or towards particular types of retrieval systems. The incompleteness is a problem particularly when one wants to evaluate a system that did not contribute to the pools: the documents returned by such a system are either (I) judged relevant (possibly with relevance levels); (II) judged nonrelevant; or (III) unjudged. We do not know whether each unjudged document is relevant or not.

A standard practice in the IR community is to regard both documents of Types (II) and (III) as nonrelevant. However, a simple and useful alternative is to first create a *condensed list* from the raw ranked list by *removing all unjudged documents* from it, and then compute the evaluation metrics for the condensed list [79]. For example, if a raw ranked list contains an unjudged document at rank 1, a judged nonrelevant document at rank 2, and a judged relevant document at rank 3, the corresponding condensed list would have the judged nonrelevant document at rank 1 and the judged relevant document at rank 2. Thus condensing a ranked list *promotes* judged documents.

A condensed-list version of Metric M is referred to as M' [79]: in particular, AP' is also known as *Induced AP* [116]. Let r' denote the rank of a document in a condensed list. Then, from Eq. 9:

$$AP' = \frac{1}{R} \sum_{r'} I(r') \frac{C(r')}{r'}. \quad (21)$$

Buckley and Voorhees [16] designed a family of metrics known as *bpref* (binary preference) specifically for the purpose of conducting IR evaluation that is robust to incomplete relevance assessments. The basic idea is to evaluate systems based on their ability to prefer judged relevant documents over judged nonrelevant ones. However, Sakai [79] showed that *bpref* is equivalent to AP' except that it lacks the *top heaviness* property, and that some condensed-list metrics are in fact more robust to incompleteness than *bpref*. Subsequently, Sakai and Kando [89] generalised his experiments.

Let N denote the number of judged nonrelevant documents for a topic, and recall that R is the number of judged relevant documents. Using our notations, *bpref* can be expressed as:

$$bpref_R = \frac{1}{R} \sum_{r'} I(r') \left(1 - \frac{\min(R, r' - C(r'))}{R}\right) \quad (22)$$

if $R \leq N$, and

$$bpref_N = \frac{1}{R} \sum_{r'} I(r') \left(1 - \frac{r' - C(r')}{N}\right) \quad (23)$$

if $R \geq N$. Let us consider a case where $R = N = 500$, so that $bpref = bpref_R = bpref_N$, and recall our discussion of the top heaviness of AP . Thus, when a relevant document moves up from 2 to 1 in the condensed list, its contribution of precision to AP' changes from 0.5 to 1; whereas, when a relevant document moves up from 100 to 99, its contribution of precision to AP' increases from 0.0100 to 0.0101. The latter change is negligible and hence AP' is top heavy. In contrast, when a relevant document moves up from 2 to 1 in the condensed list, the contribution to *bpref*, $1 - (r' - C(r'))/N$, changes from $1 - (2 - 1)/500 = 0.9980$ to $1 - (1 - 1)/500 = 1$ and the difference is only 0.002; when a relevant document moves up from 100 to 99, the contribution to *bpref* changes from $1 - (100 - 1)/500 = 0.8020$ to $1 - (99 - 1)/500 = 0.8040$ and the difference is 0.002 again. It can be observed that this lack of top heaviness arises from the large constants R and N used as the denominator in Eqs. 22 and 23. Compare these with Eq. 21, which uses r' as the denominator.

Büttcher *et al.* [18] advocated the use of a metric called *RankEff* [2] for robust evaluation with incomplete and biased relevance assessments. However, Sakai [81] pointed out that *RankEff* is none other than *bpref_N*, whose limitation has already been discussed above. De Beer and Moens proposed graded-relevance versions of *bpref* called *rpref* [40]: one of them is similar to *bpref_N* and therefore suffers from the same problem; the other has a minor flaw, which can be fixed [79].

While condensed-list metrics handle incomplete relevance assessments more elegantly and robustly than bpref [79, 89], they do not necessarily provide accurate evaluation results if the relevance assessments are biased. More specifically, Sakai [81] showed that, while standard metrics tend to underestimate *non-contributors* (i.e. systems that did not contribute to the pools), condensed-list metrics tend to overestimate them. This is because new systems return many unjudged documents: they are removed when the ranked list is condensed, which results in promotion of many relevant documents in the list.

2.3 Further Reading

Kekäläinen and Järvelin [56] have discussed graded-relevance versions of recall and precision called *generalised recall* and *generalised precision*. Several researchers have discussed appropriate decay functions for ranked retrieval evaluation [20, 53, 119].

Some ranked retrieval tasks require high recall. Patent search would be an example. Magdy and Jones [62] have recently proposed a recall-oriented evaluation metric specifically designed for patent search. In the context of patent invalidation search, Sakai [78] pointed out that *conditional relevance* [37] in a ranked list may be handled using an approach related to the condensed list: If Patent 1 at rank 1 and Patent 2 at rank 10 can invalidate a new patent application only if they are used together, then an evaluation metric that treats only Patent 2 in the ranked list as relevant may be useful.

On handling incompleteness and bias: in contrast to the simple condensed-list approach which can be used with any evaluation metric (See Section 2.2), there are also statistical approaches to estimating binary-relevance AP, such as *infAP* [116] and *statAP* [23]; Webber and Park [113] describe a *score adjustment* approach, which requires some new relevance assessments for the non-contributors.

Della Mea and Mizzaro's *Average Distance Measure* [41] is a metric that requires systems to estimate the absolute relevance score for each document, and is not a ranked retrieval metric per se. For ranked retrieval, it is not suitable as it lacks the top heaviness property [74].

The evaluation metrics discussed in this lecture assume per-document relevance assessments. An alternative would be to design evaluation metrics based on *preference judgments* [22]: is this document more relevant than another?

3 Advanced IR Metrics

Section 2 discussed set retrieval and ranked retrieval metrics: the evaluation target was a set or a ranked list of documents, where each document is either (graded) relevant or nonrelevant.

In this section, we discuss evaluation metrics for more diverse information access tasks. Section 3.1 discusses evaluation metrics for *diversified search*, which is especially important for web search where queries tend to be *ambiguous* and/or

underspecified [32]. Section 3.2 discusses evaluation metrics for multi-query sessions (i.e. multiple ranked lists), and Section 3.3 discusses those for systems that generate a textual output in response to a query. Section 3.4 provides information for further reading.

3.1 Diversified Search Metrics

Given an ambiguous and/or underspecified query, diversified search aims at covering different search *intents* with a single, short list of retrieved documents. To evaluate diversified search, it is usually assumed that each topic has a set of known intents (or *subtopics*)⁵. In contrast to traditional IR evaluation where relevance assessments are obtained for each *topic*, in diversity evaluation, relevance assessments are obtained for each *intent*. Note that a document may be relevant to multiple intents of a given topic, with different degrees of relevance.

A diversified search test collection may consist of the following:

- (a) A target corpus;
- (b) A topic set $\{q\}$ that contains ambiguous or underspecified topics;
- (c) A *topic type label* for each topic, e.g. “ambiguous”, “underspecified (faceted)”, etc. (*optional*);
- (d) A set of intents $\{i\}$ for each topic;
- (e) Intent probabilities $Pr(i|q)$ (*optional*);
- (f) An *intent type label* for each intent, e.g. “informational”, “navigational”, etc. (*optional*); and
- (g) (Graded) relevance assessments for each intent.

Subtopic Recall, or Intent Recall *Subtopic recall* [118], also known as *intent recall* [94] (I-rec), is the proportion of intents covered by a search output. In the context of ranked retrieval, one way to express it would be as follows. Let $I_i(r)$ be 0 if the document at rank r is nonrelevant to Intent i , and 1 otherwise; let $isnew_i(r)$ be 1 if $I_i(k) = 0$ for $1 \leq k \leq r - 1$, and 0 otherwise; and let $newint(r) = \sum_i isnew_i(r)I_i(r)$. This is the number of new intents covered at rank r . Then intent recall for ranked retrieval may be expressed as:

$$I-rec = \frac{\sum_r newint(r)}{|\{i\}|} . \quad (24)$$

This metric by itself is not sufficient for diversity evaluation as it is actually a set retrieval metric.

⁵ “office” may be an *ambiguous* query, which may have intents such as “microsoft office” and “workplace”; “harry potter” may be an *underspecified* query, which may have intents such as “harry potter books”, “harry potter films”, “harry potter the character” and so on.

α -nDCG α -nDCG [32] was probably the first metric to have considered the trade-off between relevance and diversity for ranked retrieval. It is an extension of nDCG: the key difference is that, prior to rank-based discounting, each document relevant to a particular intent is discounted based on the number of relevant documents already seen. Because *redundancy* within each intent is discouraged, the overall diversity of the ranked list is encouraged.

Let $C_i(r) = \sum_{k=1}^r I_i(k)$. α -nDCG is computed by replacing the standard gain values $g(r)$ in Eq. 6 with *novelty-biased* gains $ng(r)$:

$$ng(r) = \sum_i I_i(r)(1 - \alpha)^{C_i(r-1)} \quad (25)$$

where α is a parameter that can be interpreted as the probability that the user judges a nonrelevant document to be relevant to intent i by mistake ($0 \leq \alpha < 1$)⁶. Unlike the standard nDCG, however, computing the ideal list based on $ng(r)$ and thereby obtaining the ideal novelty-biased gains $ng^*(r)$ is NP-complete, and a greedy approximation is required in practice.

It should be noted that α -nDCG cannot handle per-intent graded relevance. According to Eq. 25, the relevance level of a document (before discounting) is defined simply as the number of intents it covers⁷. For example, if $\alpha = 0.5$ (the setting used at the TREC diversity task [29]), a document relevant to only one intent will receive an $ng(r)$ of 1 if this is the first relevant one found for the intent, 0.5 if this is the second relevant one found, and 0.25 if this the third relevant one found, and so on. Also, the above version of α -nDCG does not consider the intent probabilities $Pr(i|q)$: Clarke *et al.* [30] extended the α -nDCG framework to incorporate them.

Leenanupub, Zuccon and Jose [59] proposed to set the parameter α of α -nDCG on a per-topic basis. Clarke, Kolla and Vechtomova [33] combined the ideas of RBP and α -nDCG and proposed another diversity metric called *Novelty- and Rank-Biased Precision* (NRBP).

Intent-Aware Metrics Agrawal *et al.* [1] proposed the *intent-aware* (IA) approach to diversity evaluation. Let M_i be the value of a traditional IR metric computed for each intent i , using the per-intent relevance assessments for i . Then the IA version of this metric, denoted by M -IA, is simply defined as:

$$M\text{-IA} = \sum_i Pr(i|q)M_i . \quad (26)$$

More specifically, Agrawal *et al.* considered nDCG, AP and RR for M_i . Note that, to compute *nDCG-IA*, an ideal list needs to be created *for each intent*

⁶ Whereas, it is assumed that the user never judges a relevant document to be non-relevant by mistake [32].

⁷ To be more precise, α -nDCG defines the relevance level of a document as the number of *nuggets* it covers [32], but in practice, each intent (subtopic) is considered as a single nugget.

based on *per-intent* relevance assessments, so that $nDCG_i$ is computed prior to taking the expectation over the intents. The per-intent gain values $gv_{i,x}$ used for computing $nDCG_i$ are sometimes referred to as *local* gains, and the per-intent ideal list used as the denominator of $nDCG_i$ is sometimes referred to as *locally ideal* lists [94].

While IA metrics are simple to understand and to compute, they have several shortcomings. First, they do not range fully between 0 and 1: note, for example, that it is usually impossible for a system output to be locally ideal for every intent at the same time when computing nDCG-IA. Second, IA metrics generally tend to heavily reward relevance-oriented systems rather than diversity-oriented systems [30, 94]. Third, they underperform other diversity metrics in terms of *discriminative power* [77] (discussed in Section 5.1).

Perhaps the most useful (and the most popular) of the IA metrics is *ERR-IA*, the IA version of ERR [27]. A version of ERR-IA was used as the primary metric at the TREC Web Track Diversity Task [34]. As we discussed in Section 2.2, ERR has the diminishing return property, which, when used with the IA approach, serves as a mechanism for penalising redundancy for each intent i , just like the novelty-biased gain of α -nDCG does. Thus, unlike the other IA metrics, ERR-IA can reward diversity-oriented systems as it is supposed to. Clarke *et al.* [30] and Chapelle *et al.* [26] have independently shown that α -nDCG and ERR-IA can be formulated within a single framework.

D-measures Sakai and Song [94] proposed the *D-measure* approach to diversity evaluation. Let *rel* be a random binary variable, which can either be 1 (relevant) or 0 (nonrelevant). According to the *Probability Ranking Principle* [69] (PRP), systems should rank the documents $\{d\}$ by $Pr(rel = 1|q, d)$. In the context of diversity evaluation where the query q has a set of intents $\{i\}$, we let $rel = 1$ for (q, d) if and only if there exists at least one intent i such that $rel = 1$ for (i, d) . If we assume that the intents for query q are mutually exclusive, then the PRP reduces to ranking documents by $\sum_i Pr(i|q)Pr(rel = 1|i, d)$, where $Pr(rel = 1|i, d)$ is the probability that d is relevant to intent i . If we further assume that the local gain value $gv_{i,x}$ for each (i, d) pair is proportional to this probability, then the systems should rank documents by the *global gain*, given by $\sum_i Pr(i|q)gv_{i,x}$. The resultant list is called the *globally ideal* list. This can be understood as the requirement that documents highly relevant to many major intents should be ranked higher than those marginally relevant to few minor intents, which is intuitive.

Let $GG^*(r)$ denote the global gain value for the document at rank r in the globally ideal list. On the other hand, for a given diversified ranked list to be evaluated, let $g_i(r) = gv_{i,x}$ if the document at r is x -relevant to intent i , and let the global gain at r be defined as:

$$GG(r) = \sum_i Pr(i|q)g_i(r) . \quad (27)$$

By replacing the $g(r)$ of nDCG in Eq. 6 with $GG(r)$, D -nDCG can be defined as:

$$D\text{-nDCG} = \frac{\sum_{r=1}^l GG(r)/\log(r+1)}{\sum_{r=1}^l GG^*(r)/\log(r+1)}. \quad (28)$$

Similarly, based on the globally ideal list, other “D-measures” such as D - Q (a D -version of Q -measure) can be defined [94].

Note that while nDCG-IA requires multiple locally ideal lists, D -nDCG defines one globally ideal list, achieves the maximum value of 1 when the evaluated list is identical to the ideal list for ranks $[1, l]$. D -measures are “overall relevance” metrics that combine per-intent relevance assessments and intent probabilities.

At the NTCIR INTENT tasks [88], D -nDCG (overall relevance) was plotted against I -rec (pure diversity) for each participating system, which is useful for seeing which systems are relevance-oriented and which systems are diversity-oriented. Furthermore, to combine the two axes to provide a summary metric, the INTENT tasks also used $D\sharp$ -nDCG:

$$D\sharp\text{-nDCG} = \gamma I\text{-rec} + (1 - \gamma) D\text{-nDCG} \quad (29)$$

where γ is a parameter ($0 \leq \gamma \leq 1$), simply set to 0.5 at NTCIR.

Sakai and Song [94, 95] have demonstrated the advantages of the D -measure framework over α -nDCG and the IA metrics in terms of *discriminative power* [77] (discussed in Section 5.1) and the *concordance test* [82] (discussed in Section 5.3).

Sakai and Dou [85] have combined the idea of U -measure (See Section 2.2) with the above D -measure approach and with the IA approach to handle diversity evaluation. Figure 4(b)-(d) (See Section 2.2) illustrate how trailtexts can be constructed in the context of diversity evaluation: recall that U -measure can reflect the snippet/document reading behaviour of the user, and has the diminishing return property. Let s_k be a string (i.e. a snippet or part of full text), and let $pos(s_k)$ be the offset position of s_k within a trailtext. Then, using position-based local gain values $g_i(s_k)$ for each i , the position-based global gain can be defined as

$$g(pos(s_k)) = \sum_i Pr(i|q)g_i(pos(s_k)). \quad (30)$$

Plugging in Eq. 30 to Eq. 19 gives D - U , the D -measure version of U -measure. Similarly, the IA version of U can be computed by first computing a “local” U -measure U_i for each intent, and then combining them across the intents:

$$U\text{-IA} = \sum_i Pr(i|q)U_i. \quad (31)$$

In fact, it can be shown analytically that D - U and U -IA behave similarly [85]⁸.

⁸ In contrast, D -nDCG and nDCG-IA do *not* behave similarly, as normalisation is involved [94]: while D -nDCG normalises for the entire topic, nDCG-IA normalises per-intent (and is not normalised in its final form).

Intent-Type-Sensitive Metrics All of the diversity metrics discussed above are *intent-type-agnostic*: they do not consider the informational/navigational intent type labels⁹. One could argue that, just as diversified search systems should try to allocate more space within the top search result page to popular intents (i.e. those with high $Pr(i|q)$ values), they should also try to allocate more space to the informational intents, while reserving one document slot for each popular navigational intent. Sakai’s *intent-type-sensitive* diversity metrics do just that [82].

In the context of intent-type-sensitive diversity evaluation, we denote the sets of informational and navigational intents for query q as $\{i\}$ and $\{j\}$, respectively. One simple idea for intent-type-sensitive evaluation would be to completely ignore “redundant” relevant documents for each navigational intent, by assuming that only the first relevant document found will be useful for that intent¹⁰. In accordance with this view, let us modify Eq. 27 as follows:

$$GG^{DIN}(r) = \sum_i Pr(i|q)g_i(r) + \sum_j isnew_j(r)Pr(j|q)g_j(r) . \quad (32)$$

That is, we “turn off” all “redundant relevant” documents for each navigational intent. Note that we do this only for the ranked list being evaluated: the globally ideal list remains unchanged. *DIN-nDCG* can now be defined as:

$$DIN-nDCG = \frac{\sum_{r=1}^l GG^{DIN}(r)/\log(r+1)}{\sum_{r=1}^l GG^*(r)/\log(r+1)} . \quad (33)$$

Since the modified global gain ignores some relevant documents for navigational intents, $GG^{DIN}(r) \leq GG(r)$ holds in general, and the maximum value of *DIN-nDCG* may be less than one if at least one navigational intent has multiple relevant documents. Clearly, *DIN-nDCG* is a generalisation of *D-nDCG*: if all of the intents for q are informational, it reduces to *D-nDCG*.

Another approach to intent-type-sensitive diversity evaluation is to borrow the IA approach, but to use two different metrics for handling the two intent types. More specifically, let us use *Q*-measure (Eq. 13) for each informational intent, and P^+ (Eq. 14) for each navigational intent: recall that the only difference between these two metrics is that while *Q* assumes a uniform stopping probability distribution over R (or l) relevant documents, P^+ assumes a uniform stopping probability distribution over the top r_p relevant documents. Then, our second intent-type-sensitive metric, $P+Q$, can be defined as:

$$P+Q = \sum_i Pr(i|q)Q_i + \sum_j Pr(j|q)P_j^+ . \quad (34)$$

Finally, Eqs. 33 or 34 may be combined with *I-rec* using a formula similar to Eq. 29: the resultant metrics are called *DIN#-nDCG* and $P+Q\#$, respectively.

⁹ For query “harry potter”, “I want to know various facts about harry potter’s characters” is probably an informational intent; “I want to visit pottermore.com” is probably navigational.

¹⁰ Even navigational intents generally have multiple relevant documents in diversity test collections that have been constructed at TREC and NTCIR.

























































	α -nDCG	ERR-IA	D $\#$ -nDCG	D-U	U-IA	DIN $\#$ -nDCG	P+Q $\#$
(a) Per-intent graded relevance							
(b) Intent probabilities							
(c) Normalised							
(d) Recall independent							
(e) Discriminative power							
(f) Per-intent diminishing return							
(g) Snippet & doc length							
(h) Concordance test							

Fig. 7. Comparison of diversity metrics.

Diversity Metrics: Summary Figure 7 provides a quick summary of the diversity metrics discussed above. Some additional comments:

- (a) We have discussed Eq. 25: α -nDCG defines the graded relevance of a document as the number of intents it covers, and does not have a mechanism for directly handling per-intent graded relevance.
- (b) The original α -nDCG [32] did not consider $Pr(i|q)$, but later it was incorporated [30].
- and (d) Again, these are two sides of the same coin. α -nDCG requires an approximation of an ideal ranked list; there is a version of ERR-IA used at TREC that is normalised in a way similar to α -nDCG [30]. Normalisation generally implies the knowledge of all relevant documents, so the normalised metrics are recall-dependent. D($\#$)-nDCG, DIN($\#$)-nDCG and P+Q all require a globally ideal list which also implies the knowledge of all relevant documents. DIN($\#$)-nDCG is “almost” normalised, but may not reach one if at least one navigational intent has multiple relevant documents.
- (e) In terms of discriminative power, D($\#$)-nDCG and α -nDCG outperform ERR-IA [94]; D($\#$)-nDCG outperform D-U, U-IA and ERR-IA [85]¹¹.
- (f) α -nDCG, ERR-IA and U-IA possess the per-intent diminishing return property: for each intent, “redundant” relevant documents are penalised, so that

¹¹ These two studies [94, 85] used a version of ERR-IA, which is an “IA version of normalised ERR.”

diversity across intents is encouraged. D-U behaves similarly to U-IA, as the original U-measure already has the *per-topic* diminishing return property [85].

- (g) To date, D-U and U-IA are the only diversity metrics that take the user’s snippet and full text reading behaviour into account.
- (h) Let “ $M_1 \gg M_2$ ” denote the relationship: “ M_1 outperforms M_2 in terms of the concordance test with some gold standard metrics.” In terms of simultaneous concordance with I-rec and *effective precision*¹², $D\#-nDCG \gg D\#-nDCG \gg P+Q\# \gg \alpha-nDCG$ [82] while $DIN\#-nDCG \gg D-nDCG \gg P+Q$ [96]; in terms of simultaneous concordance with I-rec and precision and *Precision for the Most Popular Intent* (PMP)¹³, $D\#-nDCG \gg D-nDCG \gg$ (a version of) $ERR-IA$ [95]; In terms of simultaneous concordance with I-rec and precision, $D\#-nDCG \gg U-IA \gg D-U \gg D-nDCG \gg \alpha-nDCG \gg ERR-IA$ [83]¹⁴.

It is worth noting that ERR-IA performs relatively poorly in terms of both discriminative power and the concordance test.

Chandar and Carterette [24] analysed α -nDCG, ERR-IA and the intent-aware version of AP using multi-way analysis of variance. Sakai, Dou and Clarke [86] have investigated the effect of the choice of intents on diversity evaluation with α -nDCG, ERR-IA and $D(\#)$ -nDCG. Golbus, Aslam and Clarke [46] have combined the ideas of α -nDCG, IA metrics and D-measure and proposed a family of metrics called $\alpha\#-IA$ measures, which emphasise inherently difficult topics and subtopics. Brandt *et al.* [13] have proposed a dynamic tree-like presentation of diversified search results and discussed an evaluation method for it.

Sakai *et al.* [87] and Sakai [84] have experimented with condensed-list versions of $D(\#)$ -nDCG and ERR-IA to investigate the possibility of evaluating *non-contributors* (See Section 2.2) with existing diversity test collections. The results suggest that condensed-list diversity metrics provide better estimates of the non-contributors’ true performances than the raw-list metrics.

3.2 Session Metrics

In this section, we discuss evaluation metrics for multi-query sessions, which involve multiple ranked lists of documents.

Session DCG Here, we define a *multi-query session* as a user’s search activity involving at least one *query reformulation* (which could be done manually or possibly through a click on a query suggestion) and therefore multiple ranked lists of documents, but with an unchanging underlying information need. That is, there is a static set of (graded) relevant documents for this need.

¹² Precision that ignores redundant relevant documents for navigational intents.

¹³ Only documents that are relevant to the intent with the highest intent probability are considered relevant. This gold standard metric is meant to represent the diversity metrics’s ability to emphasise important intents.

¹⁴ This study [83] used the official ERR-IA performance values from TREC 2011.

In the above setting, the idea of nDCG can be extended as follows. First, arrange the m multiple ranked lists in chronological order, and concatenate the top l documents from the lists. (Alternatively, if the data contains click information, then each ranked list could first be truncated at the lowest click and then be concatenated [85].) Let \mathbf{r} be the rank of a document in the concatenated list. (The list may contain duplicate documents: one possible approach to handling this is to simply keep only the first occurrence of each document in the list and remove all other duplicates, in a way similar to the construction of a condensed list [54].) The gain at \mathbf{r} , i.e. $g(\mathbf{r})$, may be defined based on relevance assessments, clicks, or possibly both. Let $qnum(\mathbf{r})$ be a function that maps the document at \mathbf{r} in the concatenated list to its query number: for example, if the document at \mathbf{r} originally comes from the ranked list for the second query issued, then $qnum(\mathbf{r}) = 2$. Then a version of *session Discounted Cumulative Gain* [54] (sDCG) can be defined as:

$$sDCG = \sum_{\mathbf{r}} \frac{g(\mathbf{r})}{\log_4(qnum(\mathbf{r}) + 3) \log_2(\mathbf{r} + 1)}. \quad (35)$$

Thus the value of a relevant document is discounted not only by the rank in the concatenated list, but also by how many queries had to be issued in order to reach the document. In the original definition of sDCG [50], documents in later ranked lists could receive higher discounted gains than ones in the earlier lists, but the above formulation solves the problem.

The above sDCG is unnormalised: in a way similar to Eq. 6, it could be normalised based on a single ideal ranked list, which represents a situation where the user could obtain all relevant documents in decreasing order of relevance without ever reformulating a query. Note that in this case, duplicate relevant documents in the concatenated list obtained from the system should be removed: the same relevant documents should not be rewarded twice. (Järvelin *et al.* [50] describe a different normalisation scheme that involves concatenation of the top l documents from m ideal ranked lists, allowing duplicates.)

Click-based U U-measure, which was discussed in Section 2.2, can handle the evaluation of multi-query sessions. If click data with timestamps are available, it can handle nonlinear traversals as well [85]. Figure 8 Parts (e) and (f) show how a trailtext may be constructed from clicks that involve two queries (i.e. two ranked lists), by assuming that clicked documents are relevant. Parts (g) and (h) show how a trailtext may be constructed from a nonlinear traversal: in this example, the click data shows that the document at rank 4 was clicked first, and then the one at rank 2 was clicked; here, we assume that the user read the four snippets first and then read (parts of) the two clicked documents. More generally, Figure 9 provides a pseudocode of a click-based version of U-measure, for a search engine whose average snippet length is 200 characters. Note that this is just a straightforward implementation of Eq. 19 from Section 2.2.

Kanoulas *et al.* [54] proposed more complex evaluation metrics for sessions, which consider multiple possible browsing paths over the multiple ranked lists. U-

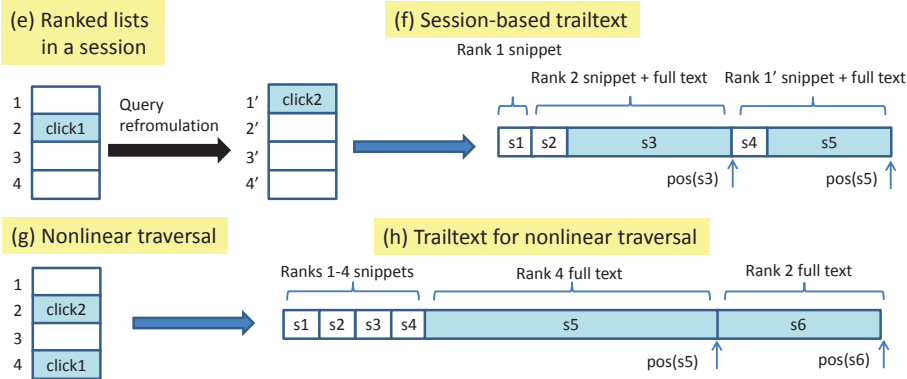


Fig. 8. Automatically constructing trailtexts from clicks or nonlinear traversals and sessions.

```

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 ≤ 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;

```

Fig. 9. Algorithm for computing U-measure by reading a session data file, which consists of *querynumber*, *clickedrank* and *doclen* sorted by time.

measure may also be extended along this line. Baskaya, Keskustalo and Järvelin [10] proposed an evaluation framework for sessions where the cost of various user actions such as query (re)formulations and clicking on “next page” are taken into account. This is in contrast to U-measure which assumes that the text that the user has read is an adequate representation of the user effort. Azzopardi [9] viewed interactive IR applications as a stream of documents and proposed evaluation metrics such as “frequency of observing a relevant document.”

3.3 Summarisation and QA Metrics

Query-focused text summarisation and question answering are types of information access where the output provided by the system is *textual*, in contrast to the information access tasks previously discussed where the output was in essence a set of document IDs, a ranked list of document IDs or multiple ranked lists of

document IDs (although TBG and U-measure consider snippets and document contents in addition). The textual output could be a single text, a ranked list of texts or a combination of document IDs with texts, but here let us consider the simplest case of evaluating a single text produced in response to a query.

ROUGE *ROUGE* [60] (Recall-Oriented Understudy for Gisting Evaluation) is a family of metrics that have been used widely for evaluating summaries. Here we discuss a few from the family to understand its basic principles. In summarisation, summaries are evaluated by means of comparison with one or more *reference* summaries, which represent the gold standard. The reference summaries could be prepared, for example, by hiring multiple people to construct summaries manually. For simplicity, here we discuss the case where there is only one reference summary s^* . Let s denote the summary to be evaluated, and let $gram_N(s)$ denote the set of word N-grams generated from s . Let e denote an N-gram, and let $Count(e, s)$ denote the frequency of e within s . Then the most basic version of ROUGE, known as *ROUGE-N*, can be expressed as follows [60, 64]:

$$ROUGE-N = \frac{\sum_{e \in gram_N(s) \cap gram_N(s^*)} \min(Count(e, s), Count(e, s^*))}{\sum_{e \in gram_N(s^*)} Count(e, s^*)} \quad (36)$$

It is clear that ROUGE-N is basically an N-gram recall measure: it was inspired by a machine translation evaluation metric called *BLEU*, which is based on N-gram precision [66].

Another version of ROUGE, called ROUGE-S, uses *skip bigrams* as the basic matching unit instead of N-grams, to allow more flexible matching between the system’s summary and the reference summaries. For a given summary s , let $skip_2(s)$ denote the set of skip bigrams, that is, any word pair extracted from the text that preserves the word order, including bigrams¹⁵. Then ROUGE-S can be expressed as follows [60, 64]:

$$Rec-S = \frac{\sum_{e \in skip_2(s) \cap skip_2(s^*)} \min(Count(e, s), Count(e, s^*))}{\sum_{e \in skip_2(s^*)} Count(e, s^*)} \quad (37)$$

$$Prec-S = \frac{\sum_{e \in skip_2(s) \cap skip_2(s^*)} \min(Count(e, s), Count(e, s^*))}{\sum_{e \in skip_2(s)} Count(e, s)} \quad (38)$$

$$ROUGE-S = \frac{(\beta^2 + 1)Prec-SRec-S}{\beta^2 Prec-S + Rec-S} \quad (39)$$

It is clear that ROUGE-S is an F-measure (Eq. 5) based on skip bigrams. Lin [60] proposed a variant of ROUGE-S called *ROUGE-SU*, which uses unigrams in addition.

¹⁵ In practice, a word distance constraint may be imposed in order to avoid pairs of words that are too far apart.

It can be observed that in summarisation evaluation, essentially IR metrics such as recall and F-measure are computed based on small textual units. (Manually constructed *semantic content units* [65] may be used instead of automatically extracted units such as those mentioned above.) This is also true for *question answering* evaluation, where the small textual unit is referred to as *nuggets*: atomic pieces of information that address a certain aspect of the question [39].

Suppose that a set of gold-standard nuggets V^* is available for a question, and that we hired a group of assessors who independently labelled each nugget $v \in V^*$ as either *vital* or *okay* (i.e. non-vital). Then, using the vital labels as votes, a weight $w(v)$ can be assigned to each v . Furthermore, given a system’s answer of length l (in characters, excluding white spaces), it can be manually compared with the nuggets from V^* , so that a set of *matched* nuggets $V (\subseteq V^*)$ is obtained. Let $allow = 100 * |V|$. Then the answer may be evaluated as follows [39]:

$$W-Rec = \frac{\sum_{v \in V} w(v)}{\sum_{v \in V^*} w(v)} \quad (40)$$

$$Prec_{allow} = 1 - \frac{\max(0, l - allow)}{l} \quad (41)$$

$$F-measure_{QA} = \frac{(\beta^2 + 1)Prec_{allow}W-Rec}{\beta^2Prec_{allow} + W-Rec} \quad (42)$$

Note that $Prec_{allow} = 1$ if $l \leq allow$. Thus it is assumed that each matched nugget in V is entitled to use up 100 characters. On the other hand, if $l > allow$, then the $(l - allow)$ characters in the answer is treated as noise.

Lin and Demner-Fushman [61] proposed an automatic unigram-matching method for replacing the aforementioned manual matching between the answer and the gold standard nuggets, and called their F-measure-based evaluation metric *POURPRE*. It should be noted that while automatic matching methods like ROUGE and POURPRE enable efficient evaluation for *extractive* systems, they may not be able to fully handle *abstractive* systems: for example, an intelligent summariser might *paraphrase* the information obtained from source documents, causing the automatic matching to fail.

S-measure, T-measure Sakai, Kato and Song [91] defined a task related to multi-document summarisation and question answering called *one click access* and proposed an extension of nugget-based weighted recall (Eq. 40) called *S-measure*. Sakai and Kato [90] extended this framework and introduced a precision-like metric called *T-measure*, and an F-measure-like metric called $S\ddagger$.

Figure 10 illustrates the concept of one click access evaluation. One click access systems are required to present important pieces of information first, and to minimise the amount of text the user has to read to obtain the information. If traditional nugget-based weighted recall is used, Outputs (a) and (b), which cover the same information, would receive exactly the same score. In contrast, S-measure prefers (b) over (a). On the other hand, T-measure imposes a length

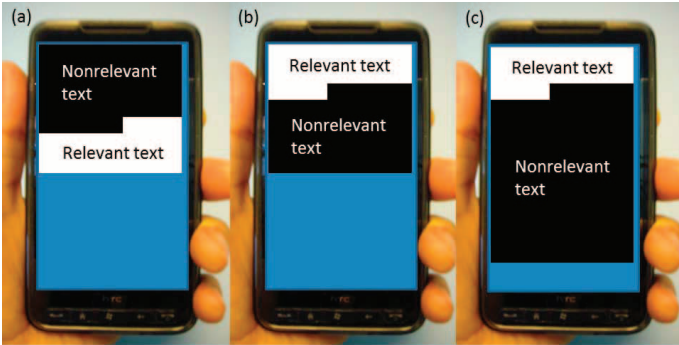


Fig. 10. Comparison of one click access systems.

penalty and prefers (b) over (c). $S_{\#}^{\ddagger}$ reflects both of these properties, as shown below.

In the one click access evaluation framework, the basic evaluation unit is called the *iUnit*. Let V^* denote the set of gold-standard iUnits for a query, and let $w(v)$ denote the weight assigned to an iUnit $v \in V^*$. Each iUnit n has a *vital string* $vs(v)$, which represents a minimal textual expression required in order to convey the information of the iUnit to the user [91]. For example, suppose that v represents a fact: “Paul McCartney was born on June 18 *cv*, 1942.” Then the vital string for v could possibly be defined as “born 6/18/1942.” Thus the vital string defines how much minimal space the iUnit requires. For a given query, we first define a *Pseudo Minimal Output* (PMO) by sorting all $vs(v)$ where $v \in V^*$ by using $w(v)$ as the first key and $|vs(v)|$ as the second key and concatenating them. PMO approximates an ideal output that presents important and concise iUnits first. Let $pos^*(v)$ denote the *offset position* (end position in characters) of $vs(v)$ within the PMO.

Let $V(\subseteq V^*)$ be the set of iUnits identified within a system output. In one click access evaluation, a system output is manually compared with the gold-standard iUnits, and the *position* of each iUnit found within the system output is recorded. For each $v \in V$, let $pos(v)$ denote its offset position (end position in characters) within the system output. Then S-measure, a position-aware version of weighted recall (See Eq. 40), is defined as:

$$S\text{-measure} = \frac{\sum_{v \in V} w(v) \max(0, 1 - pos(v)/L)}{\sum_{v \in V^*} w(v) \max(0, 1 - pos^*(v)/L)} \quad (43)$$

$$= \frac{\sum_{v \in V} w(v) \max(0, L - pos(v))}{\sum_{v \in V^*} w(v) \max(0, L - pos^*(v))} \quad (44)$$

where L is a parameter representing how quickly the user’s patience runs out [91]. For example, Sakai, Kato and Song [91] considered a Japanese one click access task with $L = 1000$: as the average reading speed of Japanese text is known to be around 500 characters per minute, this task means that the user needs to gather information within two minutes: after that, the value of any nugget becomes zero.

As S is only a position-aware version of recall, it gives the same score to Outputs (b) and (c). In order to introduce a length penalty to handle such cases, Sakai and Kato [90] introduced *T-measure*:

$$T\text{-measure} = \frac{\sum_{v \in V} |vs(v)|}{l} \quad (45)$$

where l is the system output length in characters. In contrast to the nugget precision used for question answering which uses an arbitrary allowance parameter (Eq. 41), T reflects the fact that different pieces of information require different amount of space. Finally, $S\sharp$ is a version of F-measure that is built on S and T :

$$S\sharp = \frac{(1 + \beta^2)T\flat S\flat}{\beta^2 T\flat + S\flat} \quad (46)$$

where $S\flat = \min(1, S\text{-measure})$ and $T\flat = \min(1, T\text{-measure})$ as the raw metrics are not theoretically bounded above by 1. These metrics have been used at the NTCIR One Click Access (1CLICK) task [55].

3.4 Further Reading

Recently, Arguello *et al.* [5] and Zhou *et al.* [120] have proposed evaluation methods for *aggregated search*, where not only web search results but also vertical search results (e.g. news, images, videos) need to be selectively presented. Here, the users' *vertical orientations* are taken into account: for example, for a given topic, some users might generally prefer images to textual web pages regardless of relevance. Zhou *et al.* [120] discuss the connection between diversity evaluation and aggregated search evaluation. So far, aggregated search in the research community has been considered to be the problem of arranging blocks of web search results and selected verticals on top of one another, although a more general and practical formulation would involve presentation in a two-dimensional space.

There are also information access tasks that are something of a mix between ranked retrieval and summarisation, and some evaluation methods have been proposed accordingly. Character-based *bpref* has been used for evaluating a ranked list of passages [4]; Yang and Lad [114] 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. Character-based precision and recall have been used for evaluating XML passages [52]; Arvola, Kekäläinen and Junkkari [6] 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. But as was mentioned earlier, XML retrieval evaluation is beyond the scope of this lecture.

The aforementioned U-measure [85] can potentially handle various information access tasks seamlessly by means of *trailtext*; it is easy to see that U (Eqs. 19 and 20) is a generalisation of an unnormalised version of S -measure (Eq. 43).

4 Computer-based Significance Tests

4.1 Basics

As was mentioned earlier, evaluation metrics are typically computed over a set of topics (or search requests), and it is common to compare systems based on *Mean AP* (MAP), *Mean nDCG* etc. Significance test results or confidence intervals should accompany evaluation metric values: there are arguments against statistical significance testing (e.g. [48, 51]), but reporting p -values is at least more informative than just saying “Our system’s MAP was 0.333, while the baseline’s MAP was 0.300.” Is this difference likely to be substantial or due to chance?

Statistical significance testing starts with a *null hypothesis* H_0 : in IR experiments, a typical null hypothesis would be that all systems that are being evaluated are equivalent. Then we try to compute and discuss the p -value: this is the *probability of the observed or even more extreme data, under H_0* . That is, “Assuming that the null hypothesis is true, how rare would this observation be?” Table 1 shows a contingency table that is used in significance testing: here, an arbitrary threshold called α is introduced. If the p -value is less than α , then what we have observed is something extremely rare, so we reject H_0 : that is, we decide that the systems are probably *not* equivalent.

Table 1. Type I and Type II errors in significance testing.

	Accept H_0	Reject H_0
H_0 is actually true (systems are actually equivalent)	correct conclusion (probability: $1 - \alpha$)	Type I error (probability: α)
H_0 is actually false (systems are actually different)	Type II error (probability: β)	correct conclusion (probability: $1 - \beta$)

The α is called the *significance level*, and is typically set to 0.05 (95% confidence level) or 0.01 (99% confidence level). However, note that this threshold directly affects our conclusions: consider what happens when the p -value is 0.03. Thus, it is better to report the actual p -value instead of saying “the difference is significant at $\alpha = 0.05$.” It is important to remember that statistical significance does not necessarily imply practical significance, and that statistical insignificance does not necessarily imply practical insignificance [47]. For example, Algorithm A may consistently and significantly outperform Algorithm B for any given topic, but each of the performance improvements may be too small for the user no notice; Algorithm A may have fail to significantly outperform Algorithm B, but your experiment may have used a small number of topics.

Classical significance tests may be used in IR experiments: when comparing two systems using a common topic set, for example, standard tests such as Student’s t -test (a parametric test), Wilcoxon signed-rank test and the sign test

(nonparametric tests) may be used¹⁶. In general, parametric tests rely on more assumptions but have higher *statistical power* ($1 - \beta$ in Table 1) [110]. But these tests can be found in any textbooks on statistics.

In this lecture, I will mention a few simple and useful significance testing methods that rely on computer power instead of assumptions on the underlying distributions (which often do not hold). Computer-based significance tests rely on fewer assumptions than classical tests, and are applicable to test statistics other than the mean. Here I quote Efron and Tibshirani who described the *bootstrap*, a very useful and versatile computer-based statistical framework [43]: “*The use of the bootstrap either relieves the analyst from having to do complex mathematical derivations, or in some instances provides an answer where no analytical answer can be obtained.*”

4.2 Paired Bootstrap Test

This section briefly describes the *paired bootstrap test* [43, 77, 100] which may be used instead of the *t*-test: suppose we have two systems X and Y that we want to compare using a test collection with n topics. Unlike the *t*-test, the bootstrap test does not require the normality assumption, and yet is as powerful.

For a topic set of size n , let $\mathbf{x} = (x_1, \dots, x_n)$ and $\mathbf{y} = (y_1, \dots, y_n)$ denote the per-topic performances as measured by some metric M . Thus the per-topic *differences* are given by $\mathbf{z} = (z_1, \dots, z_n)$ where $z_i = x_i - y_i$. The *sample means*, defined as $\bar{x} = \sum_i x_i/n$ and $\bar{y} = \sum_i y_i/n$, are what are often reported in IR papers, e.g. MAP of X , MAP of Y , and so on. But what we really want to know is whether the *population means* of X and Y , which we denote by μ_X and μ_Y , are any different. Hence, let $\mu = \mu_X - \mu_Y$ and let us set up the following hypotheses for a two-tailed test:

$$H_0 : \mu = 0 \quad vs. \quad H_1 : \mu \neq 0. \quad (47)$$

Thus the null hypothesis H_0 says that the population means of X and Y are actually the same.

Just like classical significance tests, the bootstrap assumes that \mathbf{z} is an independent and identically distributed sample drawn from an unknown distribution. Figure 11 shows how to obtain B *bootstrap samples* of the per-topic differences that obey H_0 . For simplicity, let us assume that $n = 5$, $\mathbf{w} = (z_1 - \bar{z}, \dots, z_5 - \bar{z}) = (0.2, 0.0, 0.1, 0.4, 0.0)$ and the b -th random sample of integers is $(1, 3, 1, 2, 4)$. Then, $\mathbf{w}^{*b} = (0.2, 0.1, 0.2, 0.0, 0.4)$.

¹⁶ Some IR history: in the late 1970s, Van Rijsbergen wrote [68]: “*parametric tests are inappropriate because we do not know the form of the underlying distribution. [...] One obvious failure is that the observations are not drawn from normally distributed populations.*” He then wrote: “*the sign test [...] can be used conservatively.*” In the early 1990s, Hull wrote [47]: “*While the errors may not be normal, the t-test is relatively robust to many violations of normality. Only heavy skewness [...] or large outliers [...] will seriously compromise its validity.*”


```

w = ( $z_1 - \bar{z}, \dots, z_n - \bar{z}$ );
for  $b = 1$  to  $B$ 
  from a set of integers  $(1, \dots, n)$ ,
  obtain a random sample of size  $n$  by sampling with replacement;
  for  $i = 1$  to  $n$ 
     $j = i$ -th element of the sample of integers;
     $w_i^{*b} = j$ -th element of w;
  end for
end for

```

Fig. 11. Algorithm for creating B bootstrap samples $\mathbf{w}^{*b} = (w_1^{*b}, \dots, w_n^{*b})$ for the Paired Test.

Now let us consider the *studentized* statistic of \mathbf{z} :

$$t(\mathbf{z}) = \frac{\bar{z}}{\bar{\sigma}/\sqrt{n}} \quad (48)$$

where $\bar{z} = \sum_i z_i/n$, and $\bar{\sigma}$ is the standard deviation of \mathbf{z} , given by:

$$\bar{\sigma} = \sqrt{\sum_i (z_i - \bar{z})^2 / (n - 1)}. \quad (49)$$

Each bootstrap sample \mathbf{w}^{*b} can be studentised in a similar way. Then, the p -value, or the *Achieved Significance Level* [43] (ASL), can be obtained as shown in Figure 12: this is simply the proportion of $t(\mathbf{w}^{*b})$ that are larger than $t(\mathbf{z})$. The p -value thus obtained should be reported together with the MAP values, etc.

```

count = 0;
for  $b = 1$  to  $B$ 
  if ( $|t(\mathbf{w}^{*b})| \geq |t(\mathbf{z})|$ ) then count++;
ASL = count/ $B$ ;

```

Fig. 12. Algorithm for estimating the Achieved Significance Level based on the Paired Test.

4.3 Unpaired Bootstrap Test

The bootstrap test described above was for a one-sample problem: we knew that x_i corresponds to y_i and we could discuss the per-topic performance differences z_i . More generally, however, there are times when we cannot assume that x_i corresponds to y_i . For example, suppose we have a set of AP values computed over a certain topic set, and another set of AP values computed over a *different* topic set. These topics may or may not differ in size. This section describes a simple bootstrap test that is applicable to such *two-sample* problems: are the two sets of performances substantially different?

Let $\mathbf{x} = (x_1, \dots, x_n)$ and $\mathbf{y} = (y_1, \dots, y_m)$ denote the per-topic performances as measured by some metric M , where m may or may not be equal to n . Then the observed difference between the two overall performances is given by $\hat{d} = M(\mathbf{x}) - M(\mathbf{y})$, where, for example, $M(\mathbf{x})$ denotes some summary statistic computed based on \mathbf{x} . But what we really want to know is whether the *true* difference d between X and Y is substantial. Hence our hypotheses for a two-tailed test would be:

$$H_0 : d = 0 \quad \text{vs.} \quad H_1 : d \neq 0. \quad (50)$$

As with classical significance tests, we assume that \mathbf{x} and \mathbf{y} are independently and identically distributed samples from unknown distributions F and G , respectively. Since we now need a distribution that obeys H_0 , let us assume that $F = G$, that is, that the observed per-topic performances all come from the same distribution. Figure 13 shows how to obtain B bootstrap samples \mathbf{x}^{*b} and \mathbf{y}^{*b} that obey H_0 . For simplicity, suppose that $\mathbf{x} = (0.1, 0.3)$, $\mathbf{y} = (0.2, 0.0, 0.0)$ and therefore that $\mathbf{v} = (0.1, 0.3, 0.2, 0.0, 0.0)$. If the b -th random sample of integers is $(1, 3, 1, 2, 4)$, then $\mathbf{x}^{*b} = (0.1, 0.2)$ and $\mathbf{y}^{*b} = (0.1, 0.3, 0.0)$. Thus, per-topic performance values are sampled with replacement without looking at whether they come from \mathbf{x} or \mathbf{y} .

```

v = (x1, ..., xn, y1, ..., ym);
for b = 1 to B
    from a set of integers (1, ..., n + m),
        obtain a random sample of size n + m by sampling with replacement;
    for i = 1 to n
        j = i-th element of the sample of integers;
        xi*b = j-th element of v;
    end for
    for i = n + 1 to n + m
        j = i-th element of the sample of integers;
        yi-n*b = j-th element of v;
    end for
end for

```

Fig. 13. Algorithm for creating bootstrap samples $\mathbf{x}^{*b} = (x_1^{*b}, \dots, x_n^{*b})$ and $\mathbf{y}^{*b} = (y_1^{*b}, \dots, y_m^{*b})$ for the Unpaired Test.

Figure 14 shows how to compute the ASL based on the unpaired bootstrap test. Note that the ASL is the proportion of the bootstrap-based overall differences that are larger than the observed difference.

Webber, Moffat and Zobel [109] have demonstrated that *score standardisation* is useful for making the evaluation metric values such as \mathbf{x} and \mathbf{y} comparable across different test collections.

```

count = 0;
for b = 1 to B
    if( |M(x*b) - M(y*b)| ≥ |d̂| ) then count++;
ASL = count/B;

```

Fig. 14. Algorithm for estimating the Achieved Significance Level based on the Unpaired Test.

4.4 Randomised Tukey’s HSD Test

When more than two systems are being evaluated in an experiment, then significance tests suitable for that purpose should be used instead of conducting a pairwise test such as the t -test or the bootstrap test one at a time. If a pairwise test with a significance level of α is conducted for k system pairs, then the *family-wise error rate* amounts to $1 - (1 - \alpha)^k$: this is the probability of detecting at least one significant difference for a pair of systems that are in fact equivalent. Carterette [21] describes a simple computer-based test suitable for multiple comparisons, which is a randomised version of the *Tukey’s Honestly Significant Differences (HSD) test*. The main idea behind Tukey’s HSD is that if the largest mean difference observed is not significant, then none of the other differences should be significant either; the null hypothesis is that there is no difference between *any* of the systems.

For an experimental environment where we have n topics and m systems (where $k = m(m - 1)/2$), let \mathbf{U} be an n -by- m matrix whose element (i, j) represents the performance of the j -th system for topic i according to some metric M . Figure 15 shows how to obtain the ASL for each run pair based on the randomised Tukey’s HSD test. The outcome of this test will generally be more conservative than that of pairwise tests conducted independently, as the family-wise error rate is now bounded above by α .

4.5 Further Reading

For one-sample problems, Smucker, Allan and Carterette [101] reported that the paired bootstrap test, the randomisation test (a.k.a. permutation test) and the t -test have little practical difference. Nevertheless, they advocate the use of the randomisation test, partly because the test does not require the assumption that the IR test topics are a random sample from a population of topics. They also argue that the use of the Wilcoxon and sign tests should be discontinued.

Robertson and Kanoulas [72] recently proposed a new methodology for significance testing in IR experiments, which views a document collection of a test collection as a sample from some larger population of documents. Thus, they discuss the interaction between a sampling of topics and a separate sampling of documents. A related approach has been described earlier by Cormack and Lynam [38].

```

foreach pair of runs  $(X, Y)$ 
   $count(X, Y) = 0$ ;
for  $b = 1$  to  $B$ 
  for  $i = 1$  to  $n$  // i.e. for every topic (every row of  $\mathbf{U}$ )
     $i$ -th row of  $\mathbf{U}^{*b} =$  random permutation of the  $i$ -th row of  $\mathbf{U}$ ;
     $max^{*b} = \max_j \bar{\mathbf{u}}_j^{*b}$ ;  $min^{*b} = \min_j \bar{\mathbf{u}}_j^{*b}$  where
       $\bar{\mathbf{u}}_j^{*b}$  is the mean of  $j$ -th column vector of  $\mathbf{U}^{*b}$ ;
    foreach pair of runs  $(X, Y)$ 
      if  $(max^{*b} - min^{*b} > |\bar{\mathbf{u}}(X) - \bar{\mathbf{u}}(Y)|)$  where
         $\bar{\mathbf{u}}(\cdot)$  is the mean of the column vector for a given run in  $\mathbf{U}$ 
      then  $count(X, Y) ++$ ;
  end for
foreach pair of runs  $(X, Y)$ 
   $ASL(X, Y) = count(X, Y)/B$ ;

```

Fig. 15. Algorithm for obtaining the Achieved Significance Level with the two-sided, randomised Tukey’s HSD given a performance value matrix \mathbf{U} whose rows represent topics and columns represent runs [21].

5 Testing IR Metrics

One ultimate goal of IR researchers is to build systems that completely and efficiently satisfy the user’s information needs, and we often regard evaluation metrics as crude indicators of user satisfaction or performance. But what are “good” metrics? There is no perfect method that answers this question. In general, it is difficult to involve real users in determining which metrics are good: we are using metrics instead of directly asking the users because it is difficult to involve real users! Below, we discuss some (imperfect) methods that have been used to “evaluate” evaluating metrics.

5.1 Discriminative Power

Suppose that two systems X and Y are being compared with evaluation metrics M_1 and M_2 . According to M_1 , X outperforms Y and the p -value is 0.0001; according to M_2 , X outperforms Y but the p -value is 0.3. If these two metrics are compared while the probability of Type I Error α (i.e. probability of concluding that two systems are different even though they are in fact equivalent) is held constant (e.g. $\alpha = 0.05$), M_1 provides a statistically significant result while M_2 does not. If this trend can be observed for different systems pairs, then one might prefer to use M_1 in IR experiments. This property of M_1 reflects its consistency or stability across the topics.

More specifically, suppose that m systems are being compared; this gives us $m(m - 1)/2$ system pairs. We can obtain a p -value for each of these pairs and for each metric, and draw *Achieved Significance Level (ASL) curves* [77] like the ones shown in Figure 16. Here, the y -axis represents the ASL (i.e. p -values), and the x -axis represents the system pairs sorted by ASL. Sakai [77] originally used the pairwise bootstrap test for producing ASL curves, but this example [85] uses

the randomised version of the Tukey’s HSD test. Metrics whose curves are close to the origin are the ones with high *discriminative power* [77, 79]: they produce smaller p -values for many run pairs than other metrics do.

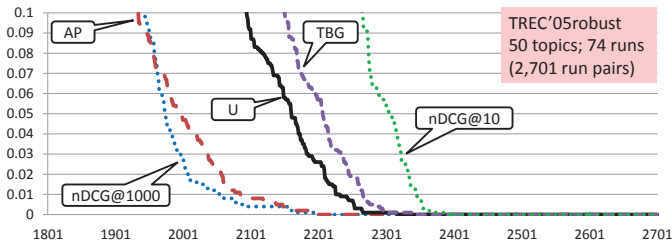


Fig. 16. ASL curves from Sakai and Dou [85].

The discriminative power method may also be used for estimating the minimum performance delta required that gives a statistically significant result, given a topic set of size n [77]. With the randomised Tukey’s HSD test, this can simply be estimated as the smallest value among the performance deltas that were actually found to be significant [82].

Discriminative power measures the consistency or stability of metrics based on significance testing¹⁷. It does *not* tell whether the metrics are measuring what we want to measure. Moreover, as was discussed earlier, statistical significance does not necessarily imply practical significance (while statistical insignificance does not necessarily imply practical insignificance). Despite this limitation, discriminative power is a moderately popular method for evaluating evaluation metrics (e.g. [30, 46, 53, 59, 73, 104, 111]).

Prior to the proposal of the discriminative power method, Buckley and Voorhees [15] and Voorhees and Buckley [107] proposed methods that are related to discriminative power. The “swap method” [107] splits the topic set of a given test collection in half, uses these two topic sets to evaluate systems independently, and asks how consistent the pairwise evaluation outcomes are. However, their methods do not consider statistical significance. Sanderson and Zobel [99] used the t -test for *filtering* run pairs before conducting the swap method. Unlike the discriminative power method, however, the swap method cannot directly estimate the performance delta between two systems that can be considered substantial for the *full* topic set: for example, if the topic set contains $n = 50$ topics, then it needs to be split into two sets of 25 topics [77]. A similar split-topic method was used by Zobel in the 1990s [121].

¹⁷ We assume that a metric is a function of some gold standard data and a system output – and nothing else. For example, something that *knows* that Output X is from Google and Output Y is Bing and uses this information to say that (say) “ Y is better than X ” for *any* query [97] is *not* a metric.

5.2 Rank Correlation

Rank correlation compares two rankings. Thus, to evaluate the sanity of an evaluation metric M , it is possible to produce a system ranking according to M , and compare it with another system ranking according to a “well-established” metric M^* . (Here, it is assumed that the two metrics rank the same set of systems.) This is also an imperfect method for evaluating evaluation metrics: we want new metrics to correlate relatively well with “established” metrics: an extremely low correlation would suggest that either previous IR research or the new metric is wrong; an extremely high correlation would suggest that it is not necessary to introduce the new metric.

Rank correlation statistics can be regarded as a special type of ranked retrieval metrics where the gold standard data also take the form of a ranked list. The most widely-used rank correlation statistic in the IR community is *Kendall’s* τ . Let m be the size of the two ranked lists, so that there are $m(m-1)/2$ pairs of ranked items within each list. Let *conc* denote the number of item pairs for which the two ranked lists are concordant (e.g. if Item X is ranked above Item Y in one list, Item X is also ranked above Item Y in the other list); similarly, let *disc* denote the number of item pairs for which the two lists are discordant. Then τ is simply given by:

$$\tau = \frac{\text{conc} - \text{disc}}{m(m-1)/2}. \quad (51)$$

One of the problems with τ in the context of IR evaluation is that the swaps near the top of the ranks and those near the bottom of the ranks are treated equally, even though what happens near the top of the ranks is generally more important. Thus several researchers have proposed alternative rank correlation statistics that have the *top heaviness* property. Here, we describe a relatively widely-used variant of τ , known as τ_{ap} [115], which is easy to compute.

The raw τ_{ap} interprets one of the two ranked lists as the gold standard (i.e. correct ranking). Let *correct*(r) denote the number of items above rank r in the evaluated list that are *correctly* ranked with respect to the item at rank r . For example, suppose that Item Y is at rank r in the evaluated list, and that Item X is ranked above it. If the gold-standard list also has X above Y , then Item Y contributes to *correct*(r). Then τ_{ap} is given by:

$$\tau_{ap} = \frac{2}{m-1} \sum_{r=2}^m \left(\frac{\text{correct}(r)}{r-1} \right) - 1. \quad (52)$$

While Kendall’s τ is a monotonic function of the probability that a randomly chosen pair of ranked items is ordered concordantly, τ_{ap} is a monotonic function of the probability that a randomly chosen item and *one ranked above it* are ordered concordantly; unlike τ , the raw τ_{ap} is asymmetric. However, a symmetric version can easily be obtained by averaging two correlation values when each list is treated as the gold standard [115]. Both τ and τ_{ap} lie between -1 and 1 .

Pollock [67], Carterette [19] and Webber, Moffat and Zobel [112] have also discussed top-heavy rank correlation statistics. Carterette’s d_{rank} measure in-

corporates correlations among system pairs; *Rank-Biased Overlap* by Webber *et al.* is applicable even to any pair of system rankings of different lengths.

5.3 Predictive Power and Concordance Test

Probably the most natural way to evaluate evaluation metrics is to “ask the user.” As was mentioned earlier, Cooper [36, 37], in the early 1970s, described a hypothetical interviewing method for users who “*enter the library.*” However, it is clear that such a method is not feasible for most of today’s IR systems such as web search engines.

Nevertheless, it is probably worthwhile to ask real people questions, and to check if evaluation metrics behave similarly to their judgments. Specifically, suppose a human participant is shown two outputs X and Y , and is asked to judge which is better. A collection of such *preference judgments* can be seen as the gold standard: if an evaluation metric agrees with the participant’s preference between X and Y , then that is a correct prediction. This can be performed for many pairs of outputs, and possibly for many participants. The ability to predict the correct preference has been referred to as *predictive power* [98]. Sanderson *et al.* [98] investigated the predictive power of traditional IR and diversity IR metrics, although they had to evaluate the latter type of metrics by treating each intent of a topic as an independent topic. Hence it may be difficult for the predictive power method to evaluate the ability of a diversity metric to actually reward diversity. Zhou *et al.* [120] reported on a similar experiment for aggregated search evaluation metrics. These studies leveraged Amazon Mechanical Turk (AMT). Similarly, in the context of diversity evaluation, Chandar and Carterette [25] used AMT to investigate what kind of novel document the user would want to see right after seeing a document relevant to a particular intent. While it should be remembered that the “Turkers” are not real users with an information need, these types of inexpensive, human-in-the-loop evaluation of evaluation metrics are probably good complements to “user-free” evaluation methods such as discriminative power.

In the context of evaluating diversity IR metrics, Sakai [82] described the *concordance test*, a user-free version of the predictive power test. Because diversity IR metrics are complex, the concordance test tries to examine how “intuitive” they are, by using some “gold-standard” metrics instead of the preference judgments. For example, for diversified search, since we want both high diversity and high relevance, it is possible to regard *intent recall* or *precision* as a gold standard. Moreover, simultaneous agreement with both of these metrics may also be examined. Note that these gold-standard metrics themselves are not good enough for diversity evaluation: these merely represent the basic properties of the more complex diversity metrics that should be satisfied.

Figure 17 shows a simple algorithm for comparing two candidate metrics M_1 and M_2 given a gold standard metric M^* : concordance with multiple gold standards may be computed in a similar way. Here, for example, $M_1(q, X)$ denotes the value of metric M_1 computed for the output of system X obtained in response to topic q . Note that this algorithm focusses on the cases where M_1

Table 2. Simultaneous concordance with intent recall and precision: TREC 2011 Web Track Diversity Task data; measurement depth $l = 10$ [83]. Statistically significant differences with the sign test are indicated by ‡ ($\alpha = 0.01$).

	D-nDCG	D-U	U-IA	ERR-IA	α -nDCG
D#-nDCG	48% /0%‡ (415)	47% /38%‡ (771)	45% /39%‡ (745)	70% /29%‡ (1106)	68% /35%‡ (913)
D-nDCG	-	42%/ 65% ‡ (562)	40%/ 67% ‡ (568)	66% /40%‡ (1044)	58% /48%‡ (974)
D-U	-	-	33%/ 80% ‡ (54)	66% /40%‡ (1472)	62% /45%‡ (1323)
U-IA	-	-	-	67% /38%‡ (1463)	63% /43%‡ (1299)
ERR-IA	-	-	-	-	19%/ 76% ‡ (292)

and M_2 disagree with each other. While it is clear that this is also an imperfect method for evaluating metrics as it assumes that the gold-standard metrics represent the real users' preferences, it is useful to be able to quantify exactly how often the metrics satisfy the basic properties such as "preference for a more diversified output" or "preference for a more relevant output" [82, 95].

```

Disagreements = 0; Conc1 = 0; Conc2 = 0;
foreach pair of runs (X, Y)
  foreach topic q
    ΔM1 = M1(q, X) - M1(q, Y);
    ΔM2 = M2(q, X) - M2(q, Y);
    ΔM* = M*(q, X) - M*(q, Y);
    if ( ΔM1 × ΔM2 < 0 ) then // M1 and M2 strictly disagree
      Disagreements ++;
      if ( ΔM1 × ΔM* ≥ 0 ) then // M1 is concordant with M*
        Conc1 ++;
      if ( ΔM2 × ΔM* ≥ 0 ) then // M2 is concordant with M*
        Conc2 ++;
    end if
  end foreach
Conc(M1|M2, M*) = Conc1/Disagreements;
Conc(M2|M1, M*) = Conc2/Disagreements;

```

Fig. 17. Concordance test algorithm for a pair of metrics M_1 and M_2 , given the gold-standard metric M^* .

Table 2 shows some examples of concordance test results, taken from Sakai [83]. Here, both intent recall and precision are used as the gold-standard metrics, and six diversity metrics are compared using the data from the TREC 2011 Diversity Task [31]. The α -nDCG and ERR-IA values are from the official TREC

results computed by `ndeval`¹⁸; the $D(\#)$ -nDCG values were computed using `NTCIREVAL`¹⁹; the D-U and U-IA values are from the Sakai and Dou [85]²⁰. This TREC data set contains 50 topics and 17 “Category A” runs [31], giving us $50 * 17 * 16/2 = 6800$ pairs of ranked lists. For example, the table shows the following information for $D(\#)$ -nDCG versus ERR-IA:

- $D(\#)$ -nDCG and ERR-IA disagree with each other for 1106 ranked list pairs out of 6800;
- Of the 1106 disagreements, $D(\#)$ -nDCG is concordant with both intent recall and precision 70% of the time, while ERR-IA is concordant with them only 29% of the time.
- The difference between $D(\#)$ -nDCG and ERR-IA is statistically significant at $\alpha = 0.01$ (though not shown in the table, $D(\#)$ -nDCG wins 592 times, while ERR-IA wins only 130 times)²¹.

It can be observed that, as was mentioned in Section 3.1, $D(\#)$ -nDCG \gg U-IA \gg D-U \gg D-nDCG \gg α -nDCG \gg ERR-IA holds, where “ \gg ” means “statistically significantly better than” in terms of simultaneous concordance with I-rec and precision.

5.4 Leave-One-Out Test

The Leave-One-Out (LOO) test [106, 121] is useful for testing the reusability of test collections that have been built based on pooling. It can also be used for comparing the robustness of evaluation metrics to incompleteness and system bias (e.g. [18, 89, 87]). Figure 18 shows how the LOO test works: the relevance assessments of a topic is a union of the *contributions* from each participating team (or *contributors*). Then a LOO relevance assessment set can be created by removing the *unique contributions* from one team (e.g. Team A). Then, if the runs from this team are evaluated based on the LOO set, it is similar to the situation where the original test collection is used for evaluating a *non-contributor*, i.e. a team that did not contribute to the pooling process.

Formally, let m be the number of contributors, and let C_j denote the contributions from the j -th team ($j = 1, \dots, m$). Each team may submit multiple runs²². The pool for this topic is given by $P = \bigcup_j C_j$, and the set of unique contributions from the j -th team is given by $U_j = C_j - \bigcup_{j' \neq j} C_{j'}$. Then the LOO set for the j -th team is given by $LOO_j = \bigcup_{j' \neq j} C_{j'} = P - U_j$. If the evaluation outcome for the j -th team based on LOO_j is similar to that based on the original

¹⁸ <http://trec.nist.gov/data/web/11/ndeval.c>

¹⁹ <http://research.nii.ac.jp/ntcir/tools/ntcireval-en.html>

²⁰ <http://research.microsoft.com/u/>

²¹ Thus $D(\#)$ -nDCG wins 54% of the time, while ERR-IA wins 12% of the time: whereas, the concordance percentages shown in the table include cases where $D(\#)$ -nDCG and ERR-IA are tied.

²² The original method of Zobel [121] left out one *run* at a time, but leaving out the entire team is more realistic and more stringent.

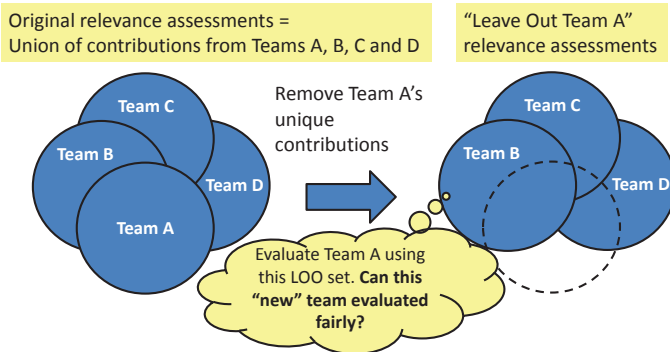


Fig. 18. Leaving out Team A.

relevance assessments P , then the test collection with that particular evaluation metric may be considered more or less reusable: the evaluation environment can properly evaluate systems that did not contribute to the pool.

5.5 Further Reading

There are other ways to evaluate evaluation metrics or evaluation environments. For example, Aslam, Yilmaz and Pavlu [8] have examined the *informativeness* of evaluation metrics; Ashkan and Clarke [7] have extended this approach to diversity evaluation metrics. *Generalisability theory* has been used for testing the reliability of evaluation environments [11, 23]. Before conducting experiments, it is always useful to discuss the theoretical properties of evaluation metrics: metrics may be studied or even designed using *measurement theory*, formal constraints and axioms (“axiometrics”) [3, 12]; just reformulating the definition of a known metric may reveal some of its (dis)advantages [79, 81].

6 Summary

This lecture covered a wide variety of IR metrics and discussed some methods for evaluating evaluation metrics. It also briefly described computer-based statistical significance test methods that are useful for IR evaluation. The take-aways for IR experimenters are: (1) It is important to understand the properties of IR metrics and choose or design appropriate ones for the task at hand; (2) Computer-based statistical significance tests are simple and useful, although statistical significance does not necessarily imply practical significance, and statistical insignificance does not necessarily imply practical insignificance; and (3) Several methods exist for discussing which metrics are “good,” although none of them is perfect.

Finally, the reader should be reminded that, to conduct good IR experiments, one should use a *competitive* baseline system (a statistically significant gain over an obsolete, fifty-year-old technique is unlikely to advance the state of the art),

multiple evaluation metrics (to evaluate systems from several angles), and multiple test collections (to see how consistent and generalisable the results might be).

Acknowledgements

I thank the organisers of PROMISE Winter School 2013, especially Dr. Nicola Ferro, for giving me the opportunity to participate in the event and for giving me the title of my lecture (and hence the title of this paper). I also thank Nobert Fuhr and the other anonymous reviewer for reading this lengthy paper and giving me good feedback. I thank Dr. Hidetsugu Nanba for checking my definition of ROUGE-N (Section 3.3).

References

1. Agrawal, R., Sreenivas, G., Halverson, A., Leong, S.: Diversifying search results. In: Proceedings of ACM WSDM 2009. pp. 5–14 (2009)
2. Ahlgren, P., Grönqvist, L.: Retrieval evaluation with incomplete relevance data: A comparative study of three measures. In: Proceedings of ACM CIKM 2006. pp. 872–873 (2006)
3. Allan, J., Aslam, J., Azzopardi, L., Belkin, N., Borlund, P., Bruza, P., Callan, J., Carman, M., Clarke, C.L., Craswell, N., Croft, W.B., Culpepper, J.S., Diaz, F., Dumais, S., Ferro, N., Geva, S., Gonzalo, J., Hawking, D., Jarvelin, K., Jones, G., Jones, R., Kamps, J., Kando, N., Kanoulas, E., Karlgren, J., Kelly, D., Lease, M., Lin, J., Mizzaro, S., Moffat, A., Murdock, V., Oard, D.W., de Rijke, M., Sakai, T., Sanderson, M., Scholer, F., Si, L., Thom, J.A., Thomas, P., Trotman, A., Turpin, A., de Vries, A.P., Webber, W., Zhang, X., , Zhang, Y.: Frontiers, challenges and opportunities for information retrieval: Report from SWIRL 2012. SIGIR Forum 46(1), 2–32 (2012)
4. Allan, J., Carterette, B., Lewis, J.: When will information retrieval be “good enough”? In: Proceedings of ACM SIGIR 2005. pp. 433–440 (2005)
5. Arguello, J., Diaz, F., Callan, J., Carterette, B.: A methodology for evaluating aggregated search results. In: ECIR 2011 (LNCS 6611). pp. 141–152 (2011)
6. Arvola, P., Kekäläinen, J., Junkkari, M.: Expected reading effort in focused retrieval evaluation. Information Retrieval 13(5), 460–484 (2010)
7. Ashkan, A., Clarke, C.L.: On the informativeness of cascade and intent-aware effectiveness measures. In: Proceedings of ACM SIGIR 2013. pp. 407–416 (2011)
8. Aslam, J.A., Yilmaz, E., Pavlu, V.: The maximum entropy method for analyzing retrieval measures. In: Proceedings of ACM SIGIR 2005. pp. 27–34 (2005)
9. Azzopardi, L.: Usage based effectiveness measures. In: Proceedings of ACM CIKM 2009. pp. 631–640 (2009)
10. Baskaya, F., Keskustalo, H., Järvelin, K.: Time drives interaction: Simulating sessions in diverse searching environments. In: Proceedings of ACM SIGIR 2012. pp. 105–114 (2012)
11. Bodoff, D., Li, P.: Test theory for assessing IR test collections. In: Proceedings of ACM SIGIR 2007. pp. 367–374 (2007)
12. Bollman, P., Cherniavsky, V.S.: Measurement-theoretical investigation of the MZ-metric. In: Proceedings of ACM SIGIR 1980. pp. 256–267 (1980)

13. Brandt, C., Joachims, T., Yue, Y., Bank, J.: Dynamic ranked retrieval. In: Proceedings of ACM WSDM 2011. pp. 247–256 (2011)
14. Broder, A.: A taxonomy of web search. SIGIR Forum 36(2) (2002)
15. Buckley, C., Voorhees, E.M.: Evaluating evaluation measure stability. In: Proceedings of ACM SIGIR 2000. pp. 33–40 (2000)
16. Buckley, C., Voorhees, E.M.: Retrieval evaluation with incomplete information. In: Proceedings of ACM SIGIR 2004. pp. 25–32 (2004)
17. Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., Hullender, G.: Learning to rank using gradient descent. In: Proceedings of ICML 2005. pp. 89–96 (2005)
18. Büttcher, S., Clarke, C.L., Yeung, P.C., Soboroff, I.: Reliable information retrieval evaluation with incomplete and biased judgements. In: ACM SIGIR 2007 Proceedings. pp. 63–70 (2007)
19. Carterette, B.: On rank correlation and the distance between rankings. In: Proceedings of ACM SIGIR 2009. pp. 436–443 (2009)
20. Carterette, B.: System effectiveness, user models, and user utility: A conceptual framework for investigation. In: Proceedings of ACM SIGIR 2011. pp. 903–912 (2011)
21. Carterette, B.: Multiple testing in statistical analysis of systems-based information retrieval experiments. ACM TOIS 30(1) (2012)
22. Carterette, B., Bennett, P.N., Chickering, D.M., Dumais, S.T.: Here or there: Preference judgments for relevance. In: ECIR 2008 (LNCS 4956). pp. 16–27 (2008)
23. Carterette, B., Pavlu, V., Kanoulas, E., Aslam, J.A., Allan, J.: Evaluation over thousands of queries. In: Proceedings of ACM SIGIR 2008. pp. 651–658 (2008)
24. Chandar, P., Carterette, B.: Analysis of various evaluation measures for diversity. In: Proceedings of DDR 2011. pp. 21–28 (2011)
25. Chandar, P., Carterette, B.: What qualities do users prefer in diversity rankings? In: Proceedings of DDR 2012 (2012)
26. Chapelle, O., Ji, S., Liao, C., Velipasaoglu, E., Lai, L., Wu, S.L.: Intent-based diversification of web search results: Metrics and algorithms. Information Retrieval 14(6), 572–592 (2011)
27. Chapelle, O., Metzler, D., Zhang, Y., Grinspan, P.: Expected reciprocal rank for graded relevance. In: Proceedings of ACM CIKM 2009. pp. 621–630 (2009)
28. Chinchor, N.: MUC-4 evaluation metrics. In: Proceedings of MUC-4. pp. 22–29 (1992)
29. Clarke, C.L., Craswell, N., Soboroff, I.: Overview of the TREC 2009 web track. In: Proceedings of TREC 2009 (2009)
30. Clarke, C.L., Craswell, N., Soboroff, I., Ashkan, A.: A comparative analysis of cascade measures for novelty and diversity. In: Proceedings of ACM WSDM 2011. pp. 75–84 (2011)
31. Clarke, C.L., Craswell, N., Soboroff, I., Voorhees, E.: Overview of the TREC 2011 web track. In: Proceedings of TREC 2011 (2012)
32. Clarke, C.L., Kolla, M., Cormack, G.V., Vechtomova, O., Ashkan, A., Büttcher, S., MacKinnon, I.: Novelty and diversity in information retrieval evaluation. In: Proceedings of ACM SIGIR 2008. pp. 659–666 (2009)
33. Clarke, C.L., Kolla, M., Vechtomova, O.: An effectiveness measure for ambiguous and underspecified queries. In: ICTIR 2009 (LNCS 5766). pp. 188–199 (2009)
34. Clarke, C.L., Craswell, N., Voorhees, E.: Overview of the TREC 2012 web track. In: Proceedings of TREC 2012 (2013)
35. Cooper, W.S.: Expected search length: A single measure of retrieval effectiveness based on the weak ordering action of retrieval systems. JASIS 19(1), 30–41 (1968)

36. Cooper, W.S.: On selecting a measure of retrieval effectiveness. *JASIS* 24(2), 87–100 (1973)
37. Cooper, W.S.: On selecting a measure of retrieval effectiveness: Part II. implementation of the philosophy. *JASIS* 24(6), 413–424 (1973)
38. Cormack, G.V., Lynam, T.R.: Statistical precision of information retrieval evaluation. In: *Proceedings of ACM SIGIR 2006* (2006)
39. Dang, H., Lin, J.: Different structures for evaluating answers to complex questions: Pyramids won't topple, and neither will human assessors. In: *Proceedings of ACL 2007*. pp. 768–775 (2007)
40. De Beer, J., Moens, M.F.: Rpref: A generalization of bpref towards graded relevance judgments. In: *Proceedings of ACM SIGIR 2006*. pp. 637–638 (2006)
41. Della Mea, V., Mizzaro, S.: Measuring retrieval effectiveness: A new proposal and a first experimental validation. *JASIST* 55(6), 503–543 (2004)
42. Dunlop, M.D.: Time, relevance and interaction modelling for information retrieval. In: *Proceedings of ACM SIGIR '97*. pp. 206–213 (1997)
43. Efron, B., Tibshirani, R.J.: *An Introduction to the Bootstrap*. Chapman & Hall/CRC (1993)
44. Eguchi, K., Oyama, K., Ishida, E., Kando, N., Kuriyama, K.: Overview of the web retrieval task at the third NTCIR workshop. In: *NII Technical Reports NII-2003-002E* (2003)
45. Gey, F., Larson, R., Machado, J., Yoshioka, M.: NTCIR9-GeoTime overview - evaluating geographic and temporal search: Round 2. In: *Proceedings of NTCIR-9*. pp. 9–17 (2011)
46. Golbus, P.B., Aslam, J.A., Clarke, C.L.: *Increasing evaluation sensitivity to diversity*. Information Retrieval (2013)
47. Hull, D.: Using statistical testing in the evaluation of retrieval experiments. In: *Proceedings of ACM SIGIR 1993*. pp. 329–338 (1993)
48. Ioannidis, J.P.: Why most published research findings are false. *PLoS Med* 2(8) (2005)
49. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems* 20(4), 422–446 (2002)
50. Järvelin, K., Price, S.L., Delcambre, L.M.L., Nielsen, M.L.: Discounted cumulated gain based evaluation of multiple-query IR sessions. In: *ECIR 2008 (LNCS 4956)*. pp. 4–15 (2008)
51. Johnson, D.H.: The insignificance of statistical significance testing. *The Journal of Wildlife Management* 63(3), 763–772 (1999)
52. Kamps, J., Pehcevski, J., Kazai, G., Lalmas, M., Robertson, S.: INEX 2007 evaluation measures. In: *INEX 2007 (LNCS 4862)*. pp. 24–33 (2008)
53. Kanoulas, E., Aslam, J.A.: Empirical justification of the gain and discount function for nDCG. In: *ACM CIKM 2009*. pp. 611–620 (2009)
54. Kanoulas, E., Carterette, B., Clough, P.D., Sanderson, M.: Evaluating multi-query sessions. In: *Proceedings of ACM SIGIR 2011*. pp. 1053–1062 (2011)
55. Kato, M.P., Sakai, T., Yamamoto, T., Iwata, M.: Report from the NTCIR-10 1CLICK-2 Japanese subtask: Baselines, upperbounds and evaluation robustness. In: *Proceedings of ACM SIGIR 2013* (2013)
56. Kekäläinen, J., Järvelin, K.: Using graded relevance assessments in IR evaluation. *JASIST* 53(13), 1120–1129 (2002)
57. Kishida, K.: Property of average precision and its generalization: An examination of evaluation indicator for information retrieval. In: *NII Technical Reports NII-2005-014E* (2005)

58. Kishida, K., Chen, K.H., Lee, S., Kuriyama, K., Kando, N., Chen, H.H.: Overview of CLIR task at the sixth NTCIR workshop. In: Proceedings of NTCIR-6. pp. 1–19 (2007)
59. Leenanupab, T., Zuccon, G., Jose, J.M.: A comprehensive analysis of parameter settings for novelty-biased cumulative gain. In: Proceedings of ACM CIKM 2012. pp. 1950–1954 (2012)
60. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Proceedings of the ACL 2004 Workshop on Text Summarization Branches Out (2004)
61. Lin, J., Demner-Fushman, D.: Methods for automatically evaluating answers to complex questions. *Information Retrieval* 9(5), 565–587 (2006)
62. Magdy, W., Jones, G.J.: PRES: A score metric for evaluating recall-oriented information retrieval applications. In: Proceedings of ACM SIGIR 2010. pp. 611–618 (2010)
63. Moffat, A., Zobel, J.: Rank-biased precision for measurement of retrieval effectiveness. *ACM TOIS* 27(1) (2008)
64. Nanba, H., Hirao, T.: Automatic evaluation in text summarization (in Japanese). *Transactions of the Japanese Society for Artificial Intelligence* 22(1), 10–16 (2008)
65. Nenkova, A., Passonneau, R., McKeown, K.: The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing* 4(2), Article 4 (2007)
66. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: IBM Research Report RC22176 (2001)
67. Pollock, S.M.: Measures for the comparison of information retrieval systems. *American Documentation* 19(4), 387–397 (1968)
68. Rijsbergen, C.J.V.: *Information Retrieval (Second Edition)*. Butterworths (1979)
69. Robertson, S.E.: The probability ranking principle in IR. *Journal of Documentation* 33, 130–137 (1977)
70. Robertson, S.E.: On GMAP: and other transformations. In: Proceedings of ACM CIKM 2006. pp. 78–83 (2006)
71. Robertson, S.E.: A new interpretation of average precision. In: Proceedings of ACM SIGIR 2008. pp. 689–690 (2008)
72. Robertson, S.E., Kanoulas, E.: On per-topic variance in IR evaluation. In: Proceedings of ACM SIGIR 2012. pp. 891–900 (2012)
73. Robertson, S.E., Kanoulas, E., Yilmaz, E.: Extending average precision to graded relevance judgments. In: Proceedings of ACM SIGIR 2010. pp. 603–610 (2010)
74. Sakai, T.: New performance metrics based on multigrade relevance: Their application to question answering. In: Proceedings of NTCIR-4 (Open Submission Session) (2004)
75. Sakai, T.: Ranking the NTCIR systems based on multigrade relevance. In: AIRS 2004 (LNCS 3411). pp. 251–262 (2005)
76. Sakai, T.: Bootstrap-based comparisons of IR metrics for finding one relevant document. In: AIRS 2006 (LNCS 4182). pp. 429–444 (2006)
77. Sakai, T.: Evaluating evaluation metrics based on the bootstrap. In: Proceedings of ACM SIGIR 2006. pp. 525–532 (2006)
78. Sakai, T.: For building better retrieval systems : Trends in information retrieval evaluation based on graded relevance (in Japanese). *IPSJ Magazine* 47(2), 147–158 (2006)
79. Sakai, T.: Alternatives to bpref. In: Proceedings of ACM SIGIR 2007. pp. 71–78 (2007)
80. Sakai, T.: On penalising late arrival of relevant documents in information retrieval evaluation with graded relevance. In: Proceedings of EVIA 2007. pp. 32–43 (2007)

81. Sakai, T.: Comparing metrics across TREC and NTCIR: The robustness to system bias. In: Proceedings of ACM CIKM 2008. pp. 581–590 (2008)
82. Sakai, T.: Evaluation with informational and navigational intents. In: Proceedings of WWW 2012. pp. 499–508 (2012)
83. Sakai, T.: How intuitive are diversified search metrics? Concordance test results for the diversity U-measures. In: Proceedings of AIRS 2013 (2013)
84. Sakai, T.: The unreusability of diversified test collections. In: Proceedings of EVIA 2013 (2013)
85. Sakai, T., Dou, Z.: Summaries, ranked retrieval and sessions: A unified framework for information access evaluation. In: Proceedings of ACM SIGIR 2013. pp. 473–482 (2013)
86. Sakai, T., Dou, Z., Clarke, C.L.: The impact of intent selection on diversified search evaluation. In: Proceedings of ACM SIGIR 2013 (2013)
87. Sakai, T., Dou, Z., Song, R., Kando, N.: The reusability of a diversified test collection. In: AIRS 2012 (LNCS 7675). pp. 26–38 (2012)
88. Sakai, T., Dou, Z., Yamamoto, T., Liu, Y., Zhang, M., Kato, M.P., Song, R., Iwata, M.: Summary of the NTCIR-10 INTENT-2 task: Subtopic mining and search result diversification. In: Proceedings of ACM SIGIR 2013 (2013)
89. Sakai, T., Kando, N.: On information retrieval metrics designed for evaluation with incomplete relevance assessments. *Information Retrieval* 11, 447–470 (2008)
90. Sakai, T., Kato, M.P.: One click one revisited: Enhancing evaluation based on information units. In: AIRS 2012 (LNCS 7675). pp. 39–51 (2012)
91. Sakai, T., Kato, M.P., Song, Y.I.: Click the search button and be happy: Evaluating direct and immediate information access. In: Proceedings of ACM CIKM 2011. pp. 621–630 (2011)
92. Sakai, T., Robertson, S.: Modelling a user population for designing information retrieval metrics. In: Proceedings of EVIA 2008. pp. 30–41 (2008)
93. Sakai, T., Shima, H., Kando, N., Song, R., Lin, C.J., Mitamura, T., Sugimoto, M., Lee, C.W.: Overview of NTCIR-8 ACLIA IR4QA. In: Proceedings of NTCIR-8. pp. 63–93 (2010)
94. Sakai, T., Song, R.: Evaluating diversified search results using per-intent graded relevance. In: Proceedings of ACM SIGIR 2011 (2011)
95. Sakai, T., Song, R.: Diversified search evaluation: Lessons from the NTCIR-9 INTENT task. *Information Retrieval* (2013)
96. Sakai, T., Song, Y.I.: On labelling intent types for evaluating search result diversification. In: Proceedings of AIRS 2013 (2013)
97. Sanderson, M.: Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval* 4, 247–375 (2010)
98. Sanderson, M., Paramita, M.L., Clough, P., Kanoulas, E.: Do user preferences and evaluation measures line up? In: Proceedings of ACM SIGIR 2010. pp. 555–562 (2010)
99. Sanderson, M., Zobel, J.: Information retrieval system evaluation: Effort, sensitivity, and reliability. In: Proceedings of ACM SIGIR 2005. pp. 162–169 (2005)
100. Savoy, J.: Statistical inference in retrieval effectiveness evaluation. *Information Processing and Management* 33(4), 495–512 (1997)
101. Smucker, M.D., Allan, J., Carterette, B.: A comparison of statistical significance tests for information retrieval evaluation. In: Proceedings of ACM CIKM 2007. pp. 623–632 (2007)
102. Smucker, M.D., Clarke, C.L.A.: Modeling user variance in time-biased gain. In: Proceedings of ACM HCIR 2012 (12)

103. Smucker, M.D., Clarke, C.L.A.: Stochastic simulation of time-biased gain. In: Proceedings of ACM CIKM 2012. pp. 2040–2044 (2012)
104. Smucker, M.D., Clarke, C.L.A.: Time-based calibration of effectiveness measures. In: Proceedings of ACM SIGIR 2012. pp. 95–104 (2012)
105. Turpin, A., Scholer, F., Järvelin, K., Wu, M., Culpepper, J.S.: Including summaries in system evaluation. In: Proceedings of ACM SIGIR 2009. pp. 508–515 (2009)
106. Voorhees, E.M.: The philosophy of information retrieval evaluation. In: CLEF 2001 (LNCS 2406). pp. 355–370 (2002)
107. Voorhees, E.M., Buckley, C.: The effect of topic set size on retrieval experiment error. In: Proceedings of ACM SIGIR 2002. pp. 316–323 (2002)
108. Voorhees, E.M., Harman, D.K. (eds.): TREC: Experiment and Evaluation in Information Retrieval. The MIT Press (2005)
109. Webber, W., Moffat, A., Zobel, J.: Score standardization for inter-collection comparison of retrieval systems. In: Proceedings of ACM SIGIR 2008. pp. 51–58 (2008)
110. Webber, W., Moffat, A., Zobel, J.: Statistical power in retrieval experimentation. In: Proceedings of ACM CIKM 2008. pp. 571–580 (2008)
111. Webber, W., Moffat, A., Zobel, J.: The effect of pooling and evaluation depth on metric stability. In: Proceedings of EVIA 2010. pp. 7–15 (2010)
112. Webber, W., Moffat, A., Zobel, J.: A similarity measure for indefinite rankings. ACM TOIS 28(4) (2010)
113. Webber, W., Park, L.A.: Score adjustment for correction of pooling bias. In: Proceedings of ACM SIGIR 2009. pp. 444–451 (2009)
114. Yang, Y., Lad, A.: Modeling expected utility of multi-session information distillation. In: ICTIR 2009 (LNCS 5766). pp. 164–175 (2009)
115. Yilmaz, E., Aslam, J., Robertson, S.: A new rank correlation coefficient for information retrieval. In: Proceedings of ACM SIGIR 2008. pp. 587–594 (2008)
116. Yilmaz, E., Aslam, J.A.: Estimating average precision with incomplete and imperfect judgments. In: ACM CIKM 2006 Proceedings. pp. 102–111 (2006)
117. Yilmaz, E., Shokouhi, M., Craswell, N., Robertson, S.: Expected browsing utility for web search evaluation. In: Proceedings of ACM CIKM 2010. pp. 1561–1564 (2010)
118. Zhai, C., Cohen, W.W., Lafferty, J.: Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In: Proceedings of ACM SIGIR 2003. pp. 10–17 (2003)
119. Zhang, Y., Park, L.A.F., Moffat, A.: Click-based evidence for decaying weight distributions in search effectiveness metrics. Information Retrieval 13(1), 46–69 (2010)
120. Zhou, K., Cummins, R., Lalmas, M., Jose, J.M.: Evaluating aggregated search pages. In: Proceedings of ACM SIGIR 2012. pp. 115–124 (2012)
121. Zobel, J.: How reliable are the results of large-scale information retrieval experiments? In: Proceedings of ACM SIGIR 1998. pp. 307–314 (1998)