The Twist Measure for IR Evaluation: Taking User's Effort Into Account

Nicola Ferro and Gianmaria Silvello
Department of Information Engineering, University of Padua, Via Gradengo 6/b, Padua 35131, Italy. E-mail: {ferro, silvello}@dei.unipd.it

Heikki Keskustalo, Ari Pirkola, and Kalervo Järvelin
School of Information Sciences, University of Tampere, Kanslerinrinne 1, Tampere FI-33014, Finland. E-mail: {heikki.keskustalo, ari.pirkola, kalervo.jarvelin}@uta.fi

We present a novel measure for ranking evaluation, called Twist (τ). It is a measure for informational intents, which handles both binary and graded relevance. τ stems from the observation that searching is currently a that searching is currently taken for granted and it is natural for users to assume that search engines are available and work well. As a consequence, users may assume the utility they have in finding relevant documents, which is the focus of traditional measures, as granted. On the contrary, they may feel uneasy when the system returns nonrelevant documents because they are then forced to do additional work to get the desired information, and this causes avoidable effort. The latter is the focus of τ, which evaluates the effectiveness of a system from the point of view of the effort required to the users to retrieve the desired information. We provide a formal definition of τ, a demonstration of its properties, and introduce the notion of effort/gain plots, which complement traditional utility-based measures. By means of an extensive experimental evaluation, τ is shown to grasp different aspects of system performances, to not require extensive and costly assessments, and to be a robust tool for detecting differences between systems.

Introduction

Information retrieval (IR) systems are pervasive in our society, and they are no longer perceived as technological advances or support tools but as staples of everyday life. The foremost examples are search engines, the main tools used by millions of people for searching, retrieving, and accessing information: How many people rely on having a search engine always available and how many think that if you cannot find something with a search engine, it does not exist?

Experimental evaluation (Cleverdon, 1997; Harman, 2011) has been and still is the main methodology adopted in IR for improving its systems and, as Tague-Sutcliffe pointed out its ultimate goal is "satisfying users not just in individual cases, but collectively, for all actual and potential users in the community" (Tague-Sutcliffe, 1996, p. 1). This has produced a wealth of research, ranging from lab-style evaluation (Harman, 2011; Harman & Voorhees, 2005; Sanderson, 2010) to interactive IR evaluation (Kelly, 2009) and the study of what user (search) tasks are (Ingwersen & Järvelin, 2005; Toms, 2011).

Here, we focus on lab-style evaluation and search tasks with informational intents (Broder, 2002), which account for the vast majority (80%) of all queries on the web (Jansen, Booth, & Spink, 2008). In particular, we investigate how the idea of search as a commodity leads to the definition of an evaluation measure grasping a different angle with respect to already existing measures.

The main measures for evaluating informational tasks—namely, average precision (AP) (Buckley & Voorhees, 2005), cumulated gain measures, for example, discounted cumulated gain (DCG) and its normalized version nDCG (Kekäläinen & Järvelin, 2002), and rank-biased precision (RBP) (Moffat & Zobel, 2008)—evaluate IR systems in terms of the quality of the ranked list output and, to some extent, also the user experience with the system. These measures adopt different user models all centered around the concept of *utility* and basically account for the level of utility (or gain) that a user can achieve in scanning a ranked result list. They differ in how they compute utility and when the total utility satisfies the user, that is, at which rank or after how many encountered relevant documents the user is

satisfied, as well as in how the measures take into account user behavior and experience (Carterette, 2011; Moffat, Thomas, & Scholer, 2013).

However, when *search* is considered a commodity, users may presume that their needs will always be served by systems, that is, they may consider systems as equivalent when seen from the viewpoint of traditional evaluation measures, and they may focus more on whether systems impose on them some avoidable effort to achieve a certain level of utility.

Therefore, in this paper, we introduce the Twist (τ) measure, which evaluates systems from the viewpoint of the avoidable effort for their users by accounting for their fatigue while visiting a nonideal ranking of documents. Twist allows us to answer the question "How much did a given utility/gain cost in terms of effort?" distinguishing between systems that have similar utility/gain but require different amounts of user effort to attain it. As a consequence, Twist is an ideal companion to traditional utility measures for informational tasks when you look at search as a commodity because Twist spots those systems that make it more difficult for their users to effortlessly benefit from effective information retrieval functionalities.

Twist adopts a user model common to several measures (Moffat et al., 2013), such as DCG and RBP, where the user scans the ranked list from top to bottom until s/he stops, and returns an estimate of the effort required to the user to traverse the ranked list. Twist builds on the notion of document misplacement, that is, how far a document is from its expected position range in an ideal ranking, assuming that a document deviating from its ideal range puts an additional burden on the user during the visit causing avoidable effort. More in detail, if we consider the curve where, at each rank position, all the document misplacements up to that rank are cumulated, Twist combines the first rank position at which no additional effort is required with the total amount of effort during the visit. The computation of the former involves considering at which rank position the first zero of the curve (if any) is; the computation of the latter involves considering the area under the curve.

Twist fits in the conceptual framework proposed by Carterette (2011), which defines system effectiveness, user models, and user utility, by substituting the concept of utility with effort: the *browsing model* is a sequential scan of the list where, at each rank position, there is the same probability of the user either stopping or going ahead to the next rank; the *document utility model* is actually a document *effort* model, based on the notion of document misplacement as explained previously; and the *utility accumulation model* is actually an *effort* accumulation model, based on the cumulation of the document misplacements up to each rank position. Note that Twist and DCG have the same browsing model which is similar to RBP when its persistence parameter is set to $p = 0.50$. They adopt the same utility accumulation model based on cumulative sum while they differ in the document utility model which is based on the document misplacement for Twist and on discounted gain for DCG.

We provide a straightforward and sound definition of Twist by using basic set and function theory (Section "Definition of the Twist Measure section"), and we formally prove some of its properties and, based on these properties, we introduce system archetypes that provide templates of performance behavior in different retrieval scenarios (Run Archetypes section). This latter aspect builds on the work by Egghe (2008), which studied precision, recall, and other metrics in a set of predefined retrieval scenarios because the properties of Twist allow us to provide a finer-grain distinction for several cases.

We introduce the notion of effort/gain plots, where Twist is plotted against traditional measures (AP, nDCG, RBP) to help interpreting system performance not only in terms of utility/gain but also in terms of the effort actually required to achieve that utility/gain (Section "Effort/Gain Plots section").

We propose using the cumulated relative position (CRP) curve (i.e., the curve where, at each rank position, all the document misplacements up to that rank are cumulated), as a visual tool both to provide a quick and intuitive idea of system behavior and to ease the interpretation and understanding of plots of other rank-by-rank measures, such as DCG (CRP as a Visual Tool section).

We conduct a thorough experimental evaluation (Experiments section) with Text REtrieval Conference (TREC)[1] collections to analyze:

- *Correlation with other metrics*: the Kendall tau correlation analysis (Kendall, 1948; Voorhees, 2001) shows that Twist is lowly correlated with existing metrics and confirms that it is looking at system performances from a different point of view;
- *Pool downsampling*: the analysis when pools are downsampled (Buckley & Voorhees, 2004; Sakai, 2007a) shows that Twist remains stable and thus extensive, and costly assessments are not mandatory in order to effectively compare systems by using it;
- *Discriminative power*: the analysis of sensitivity (Sakai, 2006, 2012, 2014) shows that Twist is, in general, as sensitive as the other metrics and improves in certain cases, confirming it as a robust tool for comparing system performances.

Finally, we discuss the advantages of Twist with respect to other related metrics (Related Metrics and Discussion section), draw conclusions, and provide an outlook for future work (Conclusions section).

Definition of the Twist Measure

Preliminary Definitions

In this section, we introduce some relevant concepts regarding experimental evaluation that are necessary to define τ. We rely on the formalization of these concepts as introduced in (Angelini, Ferro, Santucci, & Silvello, 2014)

---

[1]http://trec.nist.gov/

and reported here in short. However, in "Preliminary Definitions" we fully report it for completeness and support the demonstration of the properties of the proposed measure.

Let D bet a finite set of **documents**; $d \in D$ a document, i.e. the basic information unit; T a finite set of **topics**; and, $t \in T$ a topic, that is, the materialization of a user information need.

Let $(REL, \preccurlyeq)$ be a finite and totally ordered set of **relevance degrees** where we call **nonrelevant** the relevance degree $nr = \min(REL)$.

The **ground truth** function GT associates a relevance judgment $rel \in REL$ to each document d for each topic t.

The recall base RB_t is the total number of relevant documents for a given topic t, where by relevant document is meant any document with relevance degree above nonrelevant.

A **run** is a set of vectors of documents, where each vector $\mathbf{r}_t = (d_1, d_2, \ldots, d_N)$ of length N represents the ranked list of documents retrieved for a topic t with the constraint that no document is repeated in the ranked list. The **relevance score** $\hat{\mathbf{r}}_t = (rel_1, rel_2, \ldots, rel_N)$ associates the corresponding relevance degree to each element of a run vector.

The definition of *run* allows us to build on the work of Egghe (2008) who classified global curves obtained by considering traditional metrics such as precision and recall based on the number of retrieved documents. In a binary relevance world, he distinguished between perfect retrieval (first return all the relevant documents, then all the nonrelevant), perverse retrieval (first return all nonrelevant documents, then the relevant ones), and random retrieval (randomly returns documents without regard for their relevance). Egghe (2008) also treated the special case of a

retrieval where the number of relevant documents decreases with the number of retrieved ones; we do not explicitly handle this scenario assuming the existence of a retrieval density function (Guns, Lioma, & Larsen, 2012) based on binary relevance which is out of the scope of this article. In the following, we formally define several types of runs by extending the binary scenarios considered in Egghe (2008) and by modeling them in a graded relevance context.

The perfect retrieval scenario is defined by the **ideal run** $\mathbf{i}_t$ which contains the best ranking of all the relevant documents for all the topics; that is all the retrieved documents are arranged in the vectors in descending order of relevance.

The **worst run** $\mathbf{w}_t$ defines a set of permutations, all of which consist of solely nonrelevant documents. Note that the worst run exists only if there are at least N nonrelevant documents in D. The worst run was not considered in Egghe (2008) and here it is defined for the first time.

The **full-scale run** $\mathbf{fs}_t$, which is an extension of the perverse retrieval case of Egghe (2008), contains the worst ranking of the documents, still retrieving all the relevant documents for all the topics; in other words, it reverses the order of the ideal run.

Consider the following example: the set $REL = \{nr, pr, fr, hr\}$ contains four relevance degrees where nr stands for "nonrelevant," pr for "partially relevant," fr for "fairly relevant" and hr stands for "highly relevant" with the following ordering $nr \preccurlyeq pr \preccurlyeq fr \preccurlyeq hr$; the recall base is $RB_t = 7$; the length of the vectors is $N = 15$, and there are two different systems A and B. According to the previous definitions we have the following vectors, where $\hat{\mathbf{a}}_t$ and $\hat{\mathbf{b}}_t$ are two additional vectors of two hypothetical systems "A" and "B."

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
$\hat{\mathbf{i}}_t$	= (	hr,	hr,	fr,	fr,	pr,	pr,	pr,	nr,	nr,	nr,	nr,	nr,	nr,	nr,	nr	)
$\hat{\mathbf{w}}_t$	= (	nr,	nr,	nr,	nr,	nr,	nr,	nr,	nr,	nr,	nr,	nr,	nr,	nr,	**nr**,	nr	)
$\hat{\mathbf{fs}}_t$	= (	nr,	nr,	nr,	nr,	nr,	nr,	nr,	nr,	pr,	pr,	pr,	fr,	fr,	hr,	hr	)
$\hat{\mathbf{a}}_t$	= (	hr,	hr,	fr,	nr,	pr,	fr,	nr,	nr,	nr,	pr,	nr,	nr,	nr,	nr,	nr	)
$\hat{\mathbf{b}}_t$	= (	hr,	nr,	pr,	nr,	fr,	nr,	nr,	nr,	fr,	pr,	nr,	nr,	hr,	pr,	nr	)

The **minimum rank** $\min_{\mathbf{i}_t}(rel)$ is the first position at which we find a document with relevance degree equal to rel in the ideal run while the **maximum rank** $\max_{\mathbf{i}_t}(rel)$ is the last position at which we find a document with relevance degree equal to rel in the ideal run.

In our example, we have $\min_{\mathbf{i}_t}(hr) = 1$, $\max_{\mathbf{i}_t}(hr) = 2$, $\min_{\mathbf{i}_t}(fr) = 3$, $\max_{\mathbf{i}_t}(fr) = 4$, $\min_{\mathbf{i}_t}(pr) = 5$, $\max_{\mathbf{i}_t}(pr) = 7$, $\min_{\mathbf{i}_t}(nr) = 8$, and $\max_{\mathbf{i}_t}(nr) = 15$.

Relative Position

We can now introduce the relative position (RP), which quantifies the effect of the misplacement of relevant

documents. It was first proposed in Ferro, Sabetta, Santucci, and Tino (2011) as a support to the creation of an interactive system for exploring DCG plots via the addition of a bar visually showing the effect of the misplacement at each rank position. It was then exploited for a visual interactive failure analysis system as reported in Angelini, Ferro, Santucci, and Silvello (2012b, 2014), Di Buccio et al. (2011a, 2011b), and Ferro et al. (2011). In this paper, we exploit RP as a stepping stone for introducing the new Twist measure as well as formally proving some of its properties.

Figure 1 provides an intuitive view of the functioning of RP, relying on the previous example. In an ideal ranking,
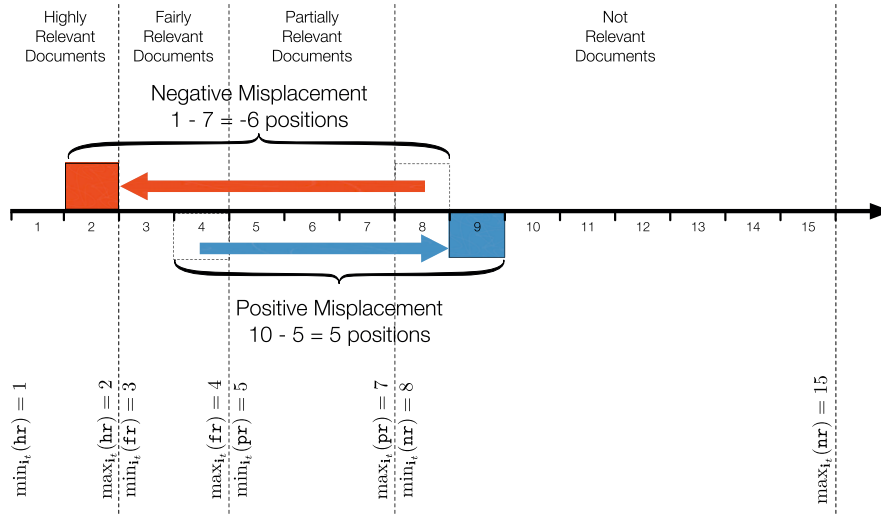
FIG. 1. Intuitive view of relative position. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

which is provided by the ideal run $\mathbf{i}_t$, the "highly relevant" documents would come first, followed by the "fairly relevant" ones, then the "partially relevant" ones, and finally the "not relevant" ones would come. Suppose now that at rank position 2 there is a "not relevant" document, in the area where "highly relevant" documents would be expected. In the most optimistic view, that is, the one that minimizes the misplacement, this document comes from the beginning of the area of the "not relevant" documents at rank 8, which is provided by $\min_{\mathbf{i}_t}(nr)$: therefore, there is a negative misplacement by 6 positions. Note that at rank 1 there is a "highly relevant" document as expected, thus the misplacement is zero. In a similar way, at rank position 9, there is a "fairly relevant" document, in the area where "not relevant" documents would be expected. Again in the most optimistic view, that is, the one that minimizes the misplacement, this document comes from the end of the area of the "fairly relevant" documents at rank 4, which is provided by $\max_{\mathbf{i}_t}(fr)$: Therefore, there is a positive misplacement by 5 positions.

As noted in Figure 1, RP departs from gain-based measures such as DCG; indeed, at each rank position, the former compares document misplacements, whereas the latter compares (and discounts) gain values. Moreover, they also rely on different scales of measurements (Krantz, Luce, Suppes, & Tversky, 1971; Stevens, 1946): Although RP is based on an ordinal scale, DCG can use both ordinal and ratio scales because the weights can be assigned to the relevance degrees and then interpreted as relevance ratios. For example, with weights $nr = 0$, $pr = 1$, $fr = 2$, and $hr = 3$, highly relevant documents can be interpreted as three times as relevant as the partially relevant ones.

Definition 1. *Given a run* R(*t*)*, the* **Relative Position** *(RP) is a function*

$$
\begin{aligned}
\text{RP:} \quad & T \times D^N \quad \rightarrow \quad \mathbb{Z}^N \\
& (t, \mathbf{r}_t) \quad \mapsto \quad \mathbf{rp}_{\mathbf{r}_t} = (rp_1, rp_2, \dots, rp_N)
\end{aligned}
$$

where

$$
\mathbf{rp}_{\mathbf{r}_t}[j] = \begin{cases} 0 & \textit{if } \min_{\mathbf{i}_t}(\hat{\mathbf{r}}_t[j]) \le j \le \max_{\mathbf{i}_t}(\hat{\mathbf{r}}_t[j]) \\ j - \min_{\mathbf{i}_t}(\hat{\mathbf{r}}_t[j]) & \textit{if } j < \min_{\mathbf{i}_t}(\hat{\mathbf{r}}_t[j]) \\ j - \max_{\mathbf{i}_t}(\hat{\mathbf{r}}_t[j]) & \textit{if } j > \max_{\mathbf{i}_t}(\hat{\mathbf{r}}_t[j]) \end{cases}
$$

RP points out the local and instantaneous effect of misplaced documents and how much they are misplaced with respect to the ideal case $\mathbf{i}_t$. Zero values denote documents which are within the ideal interval; positive values denote documents which are ranked below their ideal interval, i.e. documents of higher relevance degree that are in a position of the ranking where less relevant ones are expected; and, negative values denote documents which are above their ideal interval, i.e. less relevant documents that are in a position of the ranking where documents of higher relevance degree are expected. Note that the greater the absolute value of RP is, the greater is the distance of the document from its ideal interval; this constitutes avoidable effort put on the user while scanning the result list.

From definition 1, it follows that in the case of the ideal run $\mathbf{rp}_{\mathbf{i}_t}[j] = 0$, $\forall j \in [1, N]$, $\forall t \in T$; and in the context of the previous example, we can determine the following RP vectors:

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
$\mathbf{rp}_{i_t}$	= (	0,	0,	0,	0,	0,	0,	0,	0,	0,	0,	0,	0,	0,	0,	0	)
$\mathbf{rp}_{w_t}$	= (	−7,	−6,	−5,	−4,	−3,	−2,	−1,	0,	0,	0,	0,	0,	0,	0,	0	)
$\mathbf{rp}_{fs_t}$	= (	−7,	−6,	−5,	−4,	−3,	−2,	−1,	0,	+2,	+3,	+4,	+8,	+9,	+12,	+13	)
$\mathbf{rp}_{a_t}$	= (	0,	0,	0,	−4,	0,	+2,	−1,	0,	0,	+3,	0,	0,	0,	0,	0	)
$\mathbf{rp}_{b_t}$	= (	0,	−6,	−2,	−4,	+1,	−2,	−1,	0,	+5,	+3,	0,	0,	+11,	+7,	0	)

where, for example, $\mathbf{rp}_{w_t}[1] = 1 - \min_{i_t}(\hat{\mathbf{w}}_t[1]) = 1 - \min_{i_t}(\text{nr}) = 1 - 8 = -7$, $\mathbf{rp}_{fs_t}[15] = 15 - \max_{i_t}(\hat{\mathbf{fs}}_t[15]) = 15 - \max_{i_t}(\text{hr}) = 15 - 2 = +13$, $\mathbf{rp}_{a_t}[6] = 6 - \max_{i_t}(\hat{\mathbf{a}}_t[6]) = 6 - \max_{i_t}(\text{fr}) = 6 - 4 = +2$, and $\mathbf{rp}_{b_t}[6] = 6 - \min_{i_t}(\hat{\mathbf{b}}_t[6]) = 6 - \min_{i_t}(\text{nr}) = 6 - 8 = -2$. Note that RP values at rank 2 and 9 for the $\mathbf{b}_t$ run correspond to those shown in Figure 1. As we can see, RP values for the considered runs (i.e., $\mathbf{a}_t$ and $\mathbf{b}_t$) are all nonnegative after the recall base (i.e., $RB_t = 7$), whereas they can be both positive and negative before it. In the ideal case, after the recall base there should not be any relevant document, thus every relevant document encountered in this part of the ranking is an effect of previous misplacements and it is assigned to a positive RP value which represents a recovery of user efforts. On the other hand, before the recall base the user expects to find only relevant documents (in decreasing relevance order in the ideal case) and thus a document misplacement in this part of the ranking is assigned to a negative RP value representing avoidable effort put on the user.

The previous example and the definition of RP highlight an important point about its behavior: When documents are placed within their expected range, each rank position looks "equivalent"; on the other hand, when documents fall outside their expected range, each rank position looks "different." For example, in the former case, putting a "highly relevant" document retrieved by run $\mathbf{b}_t$ either at position 1 or 2 would give the same contribution to RP; in the latter case, putting a "highly relevant" document retrieved by run $\mathbf{b}_t$ at position 13 gives a bigger contribution to RP than putting it at, say, position 9. Note that this behavior differentiates RP from other measures such as DCG where each rank position always looks "different" since a different discount is applied to it.

The following proposition provides closed formulas for calculating the minimum and maximum of the RP of the full-scale run and demonstrates that, for each topic, they represent, respectively, the lower and upper bounds for the RP of any run.

Proposition 1. *Let* R(*t*) *be a generic run and* FS(*t*) *be the full-scale run. It holds:*

(1) lowerbound: $\forall t \in T$, $\min(\mathbf{rp}_{fs_t}) = \mathbf{rp}_{fs_t}[\min_{i_t}(\max(REL))]$
$= -RB_t$
and $\forall h \in [1, N], \min(\mathbf{rp}_{fs_t}) \leq \mathbf{rp}_{r_t}[h]$

(2) upperbound: $\forall t \in T$, $\max(\mathbf{rp}_{fs_t}) = \mathbf{rp}_{fs_t}[\max_{i_t}(\min(REL))]$
$= N - \max_{i_t}(\max(REL))$
and $\forall h \in [1, N], \max(\mathbf{rp}_{fs_t}) \geq \mathbf{rp}_{r_t}[h]$

The demonstration of proposition 1 is reported in the List of Symbols and Acronyms section in the Appendix.

According to the example above, we have that $N = 15$, $RB_t = 7$, $\min_{i_t}(\max(REL)) = \min_{i_t}(\text{hr}) = 1$, $\max_{i_t}(\min(REL)) = \max_{i_t}(\text{nr}) = 15$, and $\max_{i_t}(\max(REL)) = \max_{i_t}(\text{hr}) = 2$ therefore, $\min(\mathbf{rp}_{fs_t}) = \mathbf{rp}_{fs_t}[1] = -7$ and $\max(\mathbf{rp}_{fs_t}) = \mathbf{rp}_{fs_t}[15] = 15 - 2 = +13$.

Cumulated Relative Position

As previously discussed, RP detects instantaneous and local effects of relevant document misplacement at each rank position. On the other hand, we need, for each rank position, to account for all the avoidable effort up to that rank position. An immediate way to achieve this result is to compute the integral of the RP function or, in other terms being a discrete function, its cumulative sum. This is exactly what CRP does in the next definition. An initial version CRP was proposed for the first time in Angelini et al. (2012a) paired together with some visualizations to interact with it; in this paper, we concentrate on its full formalization as well as on thoroughly studying its properties.

Definition 2. *Given a run* R(*t*)*, the **Cumulated Relative Position (CRP)** is a function*

$$\text{CRP}: T \times D^N \to \mathbb{Z}^N$$
$$(t, \mathbf{r}_t) \mapsto \mathbf{crp}_{r_t} = (crp_1, crp_2, \ldots, crp_N)$$

where

$$\mathbf{crp}_{r_t}[j] = \sum_{k=1}^{j} \mathbf{rp}_{r_t}[k]$$

For each position j, CRP sums the values of RP up to position j included. From definition 2, it follows that for the ideal run $\mathbf{crp}_{i_t}[j] = 0, \forall j \in [1, N], \forall t \in T$.

In our example, we can determine the RP vectors shown above where, for example, $\mathbf{crp}_{w_t}[3] = (-7) + (-6) + (-5) = -18$, $\mathbf{crp}_{fs_t}[10] = (-7) + (-6) + (-5) + (-4) + (-3) + (-2) + (-1) + 0 + 2 + 3 = -23$, $\mathbf{crp}_{a_t}[6] = 0 + 0 + 0 + (-4) + 0 + 2 = -2$, and $\mathbf{crp}_{b_t}[6] = 0 + (-6) + (-2) + (-4) + 1 + (-2) = -13$.

As in the case of other cumulated metrics such as DCG, it is possible to provide a recursive definition of CRP:

$$\mathbf{crp}_{r_t}[j] = \begin{cases} \mathbf{rp}_{r_t}[j] & \text{if } j = 1 \\ \mathbf{crp}_{r_t}[j-1] + \mathbf{rp}_{r_t}[j] & \text{if } j > 1 \end{cases}$$

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
$\mathbf{crp_{i_t}}$	= (	0,	0,	0,	0,	0,	0,	0,	0,	0,	0,	0,	0,	0,	0,	0	)
$\mathbf{crp_{w_t}}$	= (	−7,	−13,	−18,	−22,	−25,	−27,	−28,	−28,	−28,	−28,	−28,	−28,	−28,	−28,	−28	)
$\mathbf{crp_{fs_t}}$	= (	−7,	−13,	−18,	−22,	−25,	−27,	−28,	−28,	−26,	−23,	−19,	−11,	−2,	+10,	+23	)
$\mathbf{crp_{a_t}}$	= (	0,	0,	0,	−4,	−4,	−2,	−3,	−3,	−3,	0,	0,	0,	0,	0,	0	)
$\mathbf{crp_{b_t}}$	= (	0,	−6,	−8,	−12,	−11,	−13,	−14,	−14,	−9,	−6,	−6,	−6,	+5,	+12,	+12	)

An alternative interpretation of the relation between RP and CRP is to consider CRP as the space the system is covering. Indeed, when CRP is moving either upwards on the negative side or downwards on the positive side, it indicates it is again moving toward the ideal case, by visiting relevant documents. Moreover, in this interpretation, being the first derivative of CRP, RP represents the speed the system is moving backward and forward in this space or, in other terms, the rate at which avoidable effort is generated.

The following proposition demonstrates that, for each topic, the minimum and maximum of the CRP in the full-scale run represent, respectively, the lower and upper bound for the CRP of any run.

Proposition 2. *Let* R(t) *be a generic run and* FS(t) *be the full-scale run. It holds:*

(1) lowerbound: $\forall t \in T$, $\min(\mathbf{crp_{fs_t}}) = \mathbf{crp_{fs_t}}[RB_t] = -\dfrac{RB_t(RB_t+1)}{2}$

and $\forall h \in [1, N]$, $\min(\mathbf{crp_{fs_t}}) \le \mathbf{crp_{r_t}}[h]$

(2) upperbound: $\forall t \in T$, $\max(\mathbf{crp_{fs_t}}) = \mathbf{crp_{fs_t}}[N] =$

$$= RB_t(N - RB_t) - \sum_{rel_i \succ \min(REL)} \Delta_{rel_i} \sum_{\substack{rel_k \succ \min(REL) \\ rel_k \circ rel_i}} \Delta_{rel_k}$$

and $\forall h \in [1, N]$, $\max(\mathbf{crp_{fs_t}}) \ge \mathbf{crp_{r_t}}[h]$

where $\Delta_{rel_i} = \max_{\mathbf{i}_t}(rel_i) - \min_{\mathbf{i}_t}(rel_i) + 1$.

The demonstration of proposition 2 is reported in the Preliminary Definitions section in the Appendix.

In our example, we have $RB_t = 7$, $N = 15$, $\Delta_{hr} = 2$, $\Delta_{fr} = 2$, and $\Delta_{pr} = 3$; therefore, $\min(\mathbf{crp_{fs_t}}) = \mathbf{crp_{fs_t}}[7] = -\dfrac{7 \cdot 8}{2} = -28$ and $\max(\mathbf{crp_{fs_t}}) = \mathbf{crp_{fs_t}}[15] = 7 \cdot (15 - 7) - (2 \cdot 2 + 2 \cdot 4 + 3 \cdot 7) = 56 - 33 = 23$.

Recovery Ratio

As discussed in the previous sections, at each rank position, CRP indicates how far a system is from the ideal point 0. Therefore, an intuitive indicator of how well a system is performing is the earliest rank position, if any, at which it passes through the ideal point 0 with respect to the recall base RB_t that represents the earliest rank position at which, in an ideal case, the system would have had the chance of retrieving all the relevant documents.

Definition 3. *Given a run* R(t)*, the recovery ratio is a function*

$$\rho : T \times D^N \to [0, 1]$$
$$(t, \mathbf{r}_t) \mapsto \rho_{\mathbf{r}_t}$$

*where, given the set of the **crossings** of the run with respect to the x axis*

$$x_{\mathbf{r}_t} = \{ j \in [1, N-1] \mid (\mathbf{crp_{r_t}}[j] \le 0 \wedge \mathbf{crp_{r_t}}[j+1] \ge 0) \vee$$
$$(\mathbf{crp_{r_t}}[j] \ge 0 \wedge \mathbf{crp_{r_t}}[j+1] \le 0) \}$$

*and the **balance point***

$$\beta_{\mathbf{r}_t} = \max(RB_t, \min(x_{\mathbf{r}_t}))$$

we have

$$\rho_{\mathbf{r}_t} = \frac{RB_t}{\beta_{\mathbf{r}_t}}$$

The crossings of the run identify the positions in the vector where the CRP crosses the x axis in either the upward or downward direction by looking at two consecutive elements in the vector where the first is negative and the second is not negative (or vice versa) taking this first one as cross-point of the x axis. We have not used the condition $CRP_{\mathbf{r}_t}[j] = 0$ because CRP assumes discrete values and it often does not get the actual value zero.

The balance point gets the maximum between the first of the crossings of the run and the recall base RB_t. It basically tries to look at when the system stops moving in the negative region of the space and passes close to the ideal point, thus looking at the first chance the system had to stop to provide avoidable effort to the user.

It may happen that a run never crosses the x axis. In this case, we would have $x_{\mathbf{r}_t} = \varnothing$ and we assume[2] that $\min(\varnothing) = +\infty$.

---

[2]This assumption is well founded. Let $S \subset \mathbb{N}$ be a finite subset of the natural numbers, as happens in the case of the crossings $x_{\mathbf{r}_t}$. Let $n \subset \mathbb{N}$ be a natural number and let us define the minimum as a recursive function

$$\min(S \cup n) = \begin{cases} n & \text{if } n < \min(S) \\ \min(S) & \text{otherwise} \end{cases}$$ where the recursion is well founded

since $|S|$, the size of the argument to the recursive call, is strictly smaller than $|S \cup n|$, the size of the initial input to the function. To make this recursion terminate, we need a base case. The only base case which works is $\min(\varnothing) = +\infty$. Indeed, if we assigned $\min(\varnothing)$ to any other value $v \in \mathbb{N}$, this would lead to unsoundness, as then $\min(\{v+1\}) = \min(\varnothing \cup \{v+1\}) = v$ instead of $v+1$.

Assuming that the length of a run is greater or equal to the recall base (i.e., $N \geq RB_t$), the recovery ratio estimates how close the balance point is to the recall base RB_t, where $\rho_{\mathbf{r}_t} = 1$ indicates a perfect ranking from the recovery ratio point-of-view and $\rho_{\mathbf{r}_t} \to 0$ a progressively worse ranking; $\rho_{\mathbf{r}_t} \to 0$ when $\beta_{\mathbf{r}_t} \to +\infty$. This is reflected by the recovery ratio of the ideal run which is always one since this run corresponds to the x axis and the balance point is always set to RB_t. On the other hand, the recovery ratio is always zero for the worst run because the balance point is always set to infinity. In the case of a full-scale run, the balance point is always set after the RB_t.

Space Ratio

The interpretation of CRP as space leads to two questions. First, how much "positive" and "negative" space has the system made the user walk through as an effect of deviations from the ideal path? Secondly, how great is that covered space?—remembering that, in the ideal case, the user would not waste any effort and the total space traveled due to deviations is zero. Definition 4 answers the first question while Definition 5 answers the second one.

Definition 4. *Given a run* R(t), *the* **forward space** *and the* **backward space** *are, respectively, a function*

$$s^+: \quad T \times D^N \quad \to \quad \mathbb{N} \qquad s^-: \quad T \times D^N \quad \to \quad \mathbb{N}$$
$$(t, \mathbf{r}_t) \quad \mapsto \quad s^+_{\mathbf{r}_t} \qquad (t, \mathbf{r}_t) \quad \mapsto \quad s^-_{\mathbf{r}_t}$$

where

$$s^+_{\mathbf{r}_t} = \sum_{k | \mathbf{rp}_{\mathbf{r}_t}[k] > 0} \mathbf{rp}_{\mathbf{r}_t}[k] \quad s^-_{\mathbf{r}_t} = \sum_{k | \mathbf{rp}_{\mathbf{r}_t}[k] < 0} |\mathbf{rp}_{\mathbf{r}_t}[k]|$$

The forward and backward space functions consider that RP is the first derivative (speed) of CRP (space) and so the integral of RP where it assumes either only positive or only negative values gets the total space covered by the system going forward and backward, respectively.

To define the space ratios, we adopt a different approach than the recovery ratio. Indeed, in that case, we compared the behavior of the run to the recall base which is something independent of any given run. In the case of space we do not have such an external reference point and we need to resort to some specific run as the point of comparison for a given topic; the full-scale run is the most natural candidate since it is the one which causes the maximum avoidable effort to the user.

Definition 5. *Given a run* R(t) *and the full-scale run* FS(t), *the* **forward space ratio** *and the* **backward space ratio** *are, respectively, a function*

$$\sigma^+: \quad T \times D^N \quad \to \quad [0,1] \qquad \sigma^-: \quad T \times D^N \quad \to \quad [0,1]$$
$$(t, \mathbf{r}_t) \quad \mapsto \quad \sigma^+_{\mathbf{r}_t} \qquad (t, \mathbf{r}_t) \quad \mapsto \quad \sigma^-_{\mathbf{r}_t}$$

where

$$\sigma^+_{\mathbf{r}_t} = 1 - \frac{s^+_{\mathbf{r}_t}}{s^+_{\mathbf{fs}_t}} \qquad \sigma^-_{\mathbf{r}_t} = 1 - \frac{s^-_{\mathbf{r}_t}}{s^-_{\mathbf{fs}_t}}$$

The **space ratio** *is a function:*

$$\sigma: \quad T \times D^N \quad \to \quad [0,1]$$
$$(t, \mathbf{r}_t) \quad \mapsto \quad \sigma_{\mathbf{r}_t}$$

where

$$\sigma_{\mathbf{r}_t} = \frac{2\sigma^+_{\mathbf{r}_t}\sigma^-_{\mathbf{r}_t}}{\sigma^+_{\mathbf{r}_t} + \sigma^-_{\mathbf{r}_t}}$$

is the harmonic mean of the forward and backward space ratios.

Both the forward and backward space ratios measure how close the space is covered by a run with respect to the space covered by the full-scale run in the same direction. The closer a run is to the full-scale run, the smaller the value of the space ratios is. Indeed, the full-scale run provides us with the maximum amount of space the system can cover in both directions which is exactly the opposite of the ideal case in which the system would avoid deviating in space from the ideal path at all. Therefore, the closer a run is to the full-scale one, the worse it is.

The forward space ratio correctly identifies the ideal run as the best possible and the full-scale run as the worst possible but incorrectly equates the ideal case and the worst case. On the other hand, the backward space ratio correctly distinguishes between the ideal and worst runs and considers the worst and full-scale runs as equivalent; this is a sensible assumption since the worst run is actually worse than the full scale one, never retrieving relevant documents.

It can be noted that the worst and full-scale runs are the only two cases in which the backward space ratio assumes exactly the value zero. This observation is exploited in the space ratio where the product between the forward and backward space ratios returns zero when the backward space ratio is zero, thus compensating in this way the misinterpretation of ideal and worst runs made by the forward space ratio. Moreover, from a methodological point of view Ferger (1931) and Stevens (1955), among others, point out that the harmonic mean[3] is the most suitable for ratio scales and it gives more weight to the smaller values in a sample and this is exactly the behavior needed to compensate for the misinterpretation of the forward space ratio.

Figure 2 shows the CRP curve of a typical run and of the full-scale run to intuitively illustrate the meaning of the recovery and space ratios. The recovery ratio basically measures how close the (first) crossing of the x axis of the CRP

---

[3]Note that the geometric mean, which gives more weight to the smaller values in a sample as well, is more appropriate when the involved quantities have a logarithmic relation (Stevens, 1955), which is not the case of the space rations.
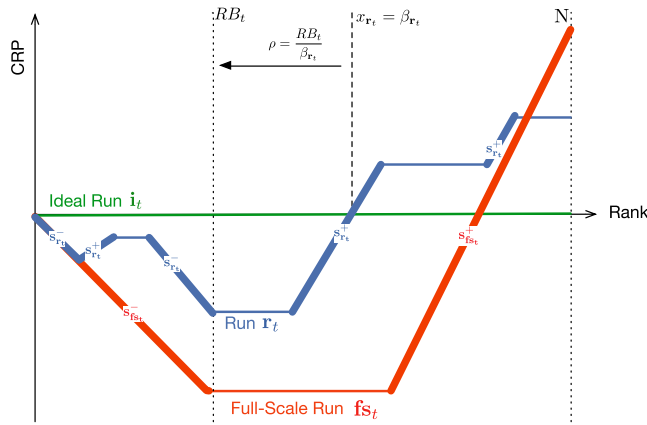
FIG. 2.   Intuitive view of the recovery and space ratios. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

curve is to the recall base, since a closer crossing indicates a better run which either has lost less in the initial rank positions or has recovered faster from an initial loss. The space ratios measure how much area there is under the RP curve, that is, the length of the CRP curve, for both positive and negative misplacements, where smaller values indicate better runs which have accumulated less document misplacements; then, they compare it with the corresponding values of the full-scale run, which is the worst case possible; the closer the behavior of a run is to the one of the full-scale run, the worse it is.

Twist

The Twist grasps the overall angle and outlook of CRP about a run by combining the recovery and space ratios through their arithmetic mean.

Definition 6. *Given a run* R(t)*, the* **Twist** *is a function:*

$$\tau: \quad T \times D^N \quad \to \quad [0,1]$$
$$(t, \mathbf{r}_t) \quad \mapsto \quad \tau_{\mathbf{r}_t}$$

where

$$\tau_{\mathbf{r}_t} = \frac{\rho_{\mathbf{r}_t} + \sigma_{\mathbf{r}_t}}{2}$$

Run Archetypes

In this section, we present the plots of CRP for several kinds of archetypal runs that illustrate the behavior of CRP for these different categories of runs belonging to different retrieval scenarios and, as discussed in the Preliminary Definitions section they extend the work of Egghe (2008). Figure 3 shows the plots for the reference runs defined in the previous section.

In particular, Figure 3A shows the plot of the ideal run—that is, the perfect retrieval scenario of Egghe (2008)—which, as previously stated, is always zero and

causes no effort to the user. In the previous section, we interpreted CRP as the total space up to a given rank position walked through by the system while progressing in list, that is, how distant the system is at a given rank position from the ideal point in space that represents the best ranking. The ideal point coincides with the x axis and both negative and positive values of CRP indicate a departure from this point.

Figure 3B illustrates the case of the worst run—a kind of zero recall retrieval scenario of Egghe (2008)—where only nonrelevant documents are retrieved and the system is departing more and more from the ideal point and it will never be able to get back toward it, causing a major effort to the user due to only retrieving not-relevant documents.

Figure 3C shows the case of the full-scale run—that is, the perverse retrieval scenario of Egghe (2008)—which, in the initial part of the rank, behaves as the worst run and then starts to retrieve all the relevant documents, in increasing order of relevance. This causes the system to move again towards the ideal point, which is crossed, and then it continues to depart from it in the positive direction. The fact that, even after crossing the ideal point, the system continues to move away from it in the positive direction is not an indicator of performances better than the ideal run but rather the signal that, even though something positive is happening because relevant documents are eventually retrieved, this is happening (too) late in the ranking and so it will leave the system far away from the ideal point, still in a positive region of the space. Therefore, this archetype causes the user the greatest effort due to both retrieving not-relevant documents and misplacing all the relevant ones in the worst possible way.

Figure 3D–H present several archetype runs that can be collocated in the random retrieval scenario of Egghe (2008). Although in Egghe (2008) there is no distinction between the curves in this scenario, CRP provides richer information and a wider spectrum of curves, allowing us to be more accurate in addressing this scenario.

Figure 3D and E demonstrate the sensitivity of CRP by showing the plots for two slightly different archetypes of excellent runs, runs able to cross the x axis once before the recall base RB_t (the vertical dotted line). The one of Figure 3D misplaces relevant documents of different relevance degrees, that is, it is a permutation of the ideal run up to the recall base RB_t. The one of Figure 3E misplaces relevant documents and misses very few of them, ranking high some nonrelevant ones instead; we can note that after the recall base RB_t, it stays constant which means it is no longer retrieving relevant documents even if it would be possible to retrieve the few missed ones and thus still grow after RB_t. These are both excellent runs since they commit marginal or very few errors in the positions of the ranking up to RB_t and they recover very quickly.

Note that excellent runs cannot happen in the case of binary relevance. Indeed, in this case, just one misplaced relevant document causes a loss before the recall base RB_t which can be recovered only after RB_t because no more
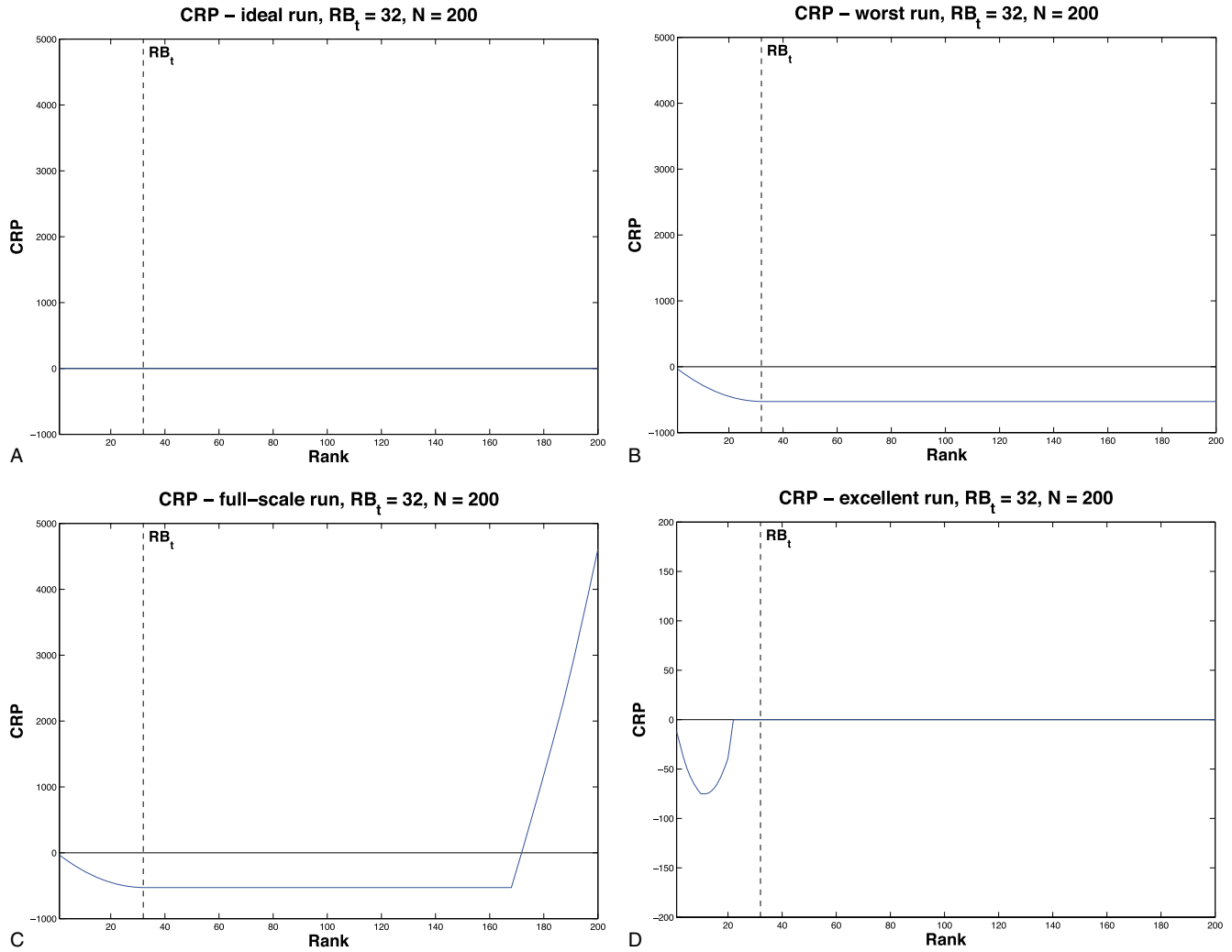
FIG. 3.   Archetypes of runs. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

misplacements before RB_t will in any case add zero to CRP and so it will never go up again before RB_t.

Figure 3F presents the case of the up-and-down runs which start similar to an excellent run but before the recall base RB_t they drop down again and, eventually, stay stable.

Figure 3G and H show the plots for two different archetypes of typical runs, meant as runs that may cross the x axis after the recall base RB_t and that represent the most common runs found in the experimental sets. The one in Figure 3G misplaces relevant documents with nonrelevant ones and with ones of different relevance degrees and then, after the recall base RB_t, starts to retrieve relevant documents again up to a positive region of the space. The one in Figure 3H is similar to the one in Figure 3G except that, after the recall base RB_t, it does not retrieve relevant documents enough to move up to the positive region of the space.

Recovery Ratio

Figure 4 shows how the recovery ratio ρ_{r_t} changes according to the different run archetypes we previously

discussed; it also provides the values for the balance point β_{r_t} and the minimum of the set of crossings x_{r_t}.

As expected, in the case of the ideal run we have $\rho_{i_t} = 1$ and in the case of the worst run we have $\rho_{w_t} = 0$; note that in the former case the balance point β_{i_t} is equal to the recall base RB_t while in the latter one it is $+\infty$, since the worst run never crosses the x axis.

As far as the full-scale run is concerned, the following proposition determines its upper bound.

Proposition 3. *Let* FS(t) *be the full-scale run and let N be its length. It holds:*

$$\rho_{fs_t} < \frac{1}{2}$$

where $N = \omega RB_t$ *with* $\omega \geq 3$ *real number.*

The demonstration of proposition 3 is reported in the Properties section of the Appendix.
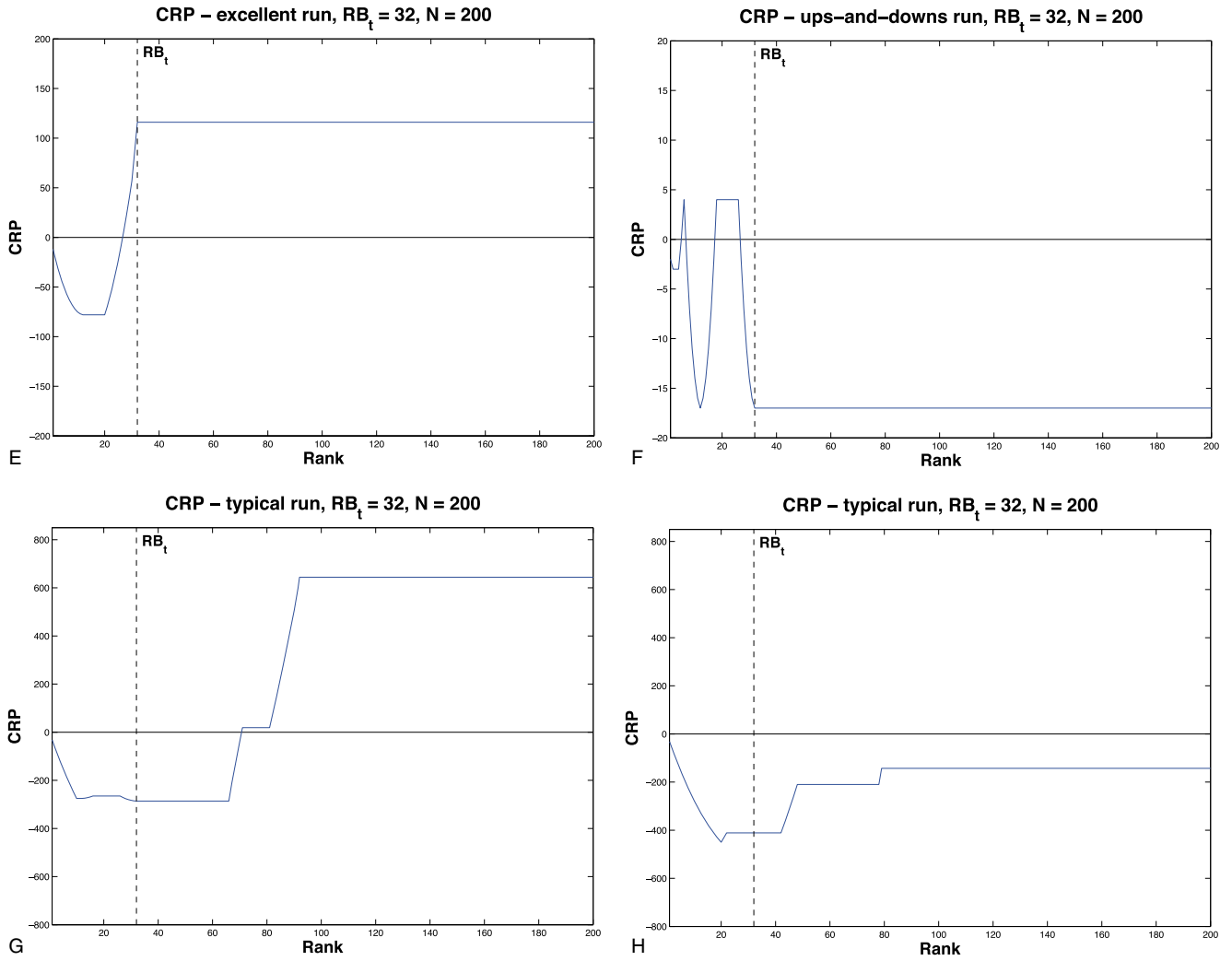
FIG. 3.   *(Continued)*

As shown in Figure 4, the recovery ratio for both the excellent and up-and-down runs is equal to 1, since they both cross the *x* axis not later than the recall base RB_t.

Typical runs of type A, that is, those crossing the *x* axis after the recall base RB_t like the one shown in Figure 3G, may achieve a recovery ratio very close to one since the earliest rank position at which they can recover from previous misplacements is RB_t+1; note that, in this case, the bigger the recall base is, that is, $RB_t \to +\infty$, the closer to 1 the recovery ratio will be, that is, $\rho_{r_t} \to 1$.

Typical runs of type B, that is, those never crossing the *x* axis, such as the one shown in Figure 3H, have recovery ratio $\rho_{r_t} = 0$, as in the case of the worst run.

Summing up the discussion, the recovery ratio:

• Discriminates full-scale runs and typical runs of type A well;
• Conflates ideal, excellent (type A and B), and up-and-down runs considering them equivalent;
• Conflates worst runs and typical runs of type B considering them equivalent.

Space Ratio

As shown in Figure 5, in the case of excellent and up-and-down runs, all the different space ratios tend towards one since all these run archetypes are typically characterized by relatively small movements forward and backward, even if they are correctly detected as not as good as the ideal run.

As expected, the typical runs of type A are the case in which the space ratios can assume all the range of their values.

When it comes to typical runs of type B, we can note that the forward space ratio fails to recognize them and assimilates them to the ideal run. On the other hand, the backward space ratio correctly assesses their behavior. Overall, by using the harmonic mean, the space ratio ameliorates this situation because it weights the backward space ratio more.

The cases of ideal, worst, and full-scale runs have already been discussed in the Definition of the Twist Measure Section.

	Ideal	Worst	Full Scale	Excellent	Ups-and-Downs	Typical A	Typical B
Plot Template							
$\min(x_{\mathbf{r}_t})$	1	$+\infty$	$\geq N - RB_t + 1 \Leftrightarrow \geq (\omega - 1)RB_t + 1$	$\leq RB_t$	$< RB_t$	$\geq RB_t + 1$	$+\infty$
$\beta_{\mathbf{r}_t}$	RB_t	$+\infty$	$\geq (\omega - 1)RB_t + 1$	RB_t	RB_t	$\geq RB_t + 1$	$+\infty$
$\rho_{\mathbf{r}_t}$	1	0	$< \dfrac{1}{2}$	1	1	$\leq \dfrac{RB_t}{RB_t + 1} < 1$	0

FIG. 4. Behavior of the crossings, balance point, and recovery ratio for the different run archetypes. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

	Ideal	Worst	Full Scale	Excellent	Ups-and-Downs	Typical A	Typical B
Plot Template							
$\sigma_{\mathbf{r}_t}^{+}$	1	1	0	$\to 1$	$\to 1$	< 1	1
$\sigma_{\mathbf{r}_t}^{-}$	1	0	0	$\to 1$	$\to 1$	< 1	< 1
$\sigma_{\mathbf{r}_t}$	1	0	0	$\to 1$	$\to 1$	< 1	< 1

FIG. 5. Behavior of the space ratios for the different run archetypes. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Summing up the discussion, the space ratio:

- Discriminates typical runs of type A and B well;
- Discriminates ideal, excellent (type A and B), and ups-and-downs runs.

Twist

As shown in Figure 6, the Twist correctly distinguishes between ideal and worst runs and detects excellent and up-and-down runs as slightly less than the ideal one, even if a little more than the space ratio alone.

The full-scale run is distinguished from the worst case and its Twist value can range only in the lower quarter of the possible values.

As far as typical runs of type both A and B, they are correctly detected and the Twist of type B ones can range only in the lower half of the possible values.

Note that other kinds of averages, such as harmonic or geometric means, would have not worked so well. Indeed, both of them require multiplying the two averaged factors and this would lead again to conflating: (i) worst runs and typical runs of type B; (ii) worst and full-scale runs.

Empirical Analysis

The archetypes previously presented identify the main categories into which we can classify a CRP curve. In the following we consider four public test collections, whose characteristics are reported in Table 1: (a) TREC 10, 2001, Web Track (Hawking & Craswell, 2001); (b) TREC 14, 2005, Robust Track (Voorhees, 2005); TREC 20, 2011, Web Track (Clarke, Craswell, & Voorhees, 2012); and (c) TREC 21, 2012, Web Track (Clarke, Craswell, & Voorhees, 2013).

	Ideal	Worst	Full Scale	Excellent	Ups-and-Downs	Typical A	Typical B
Plot Template							
$\rho_{\mathbf{r}_t}$	1	0	$< \dfrac{1}{2}$	1	1	$\leq \dfrac{RB_t}{RB_t + 1} < 1$	0
$\sigma_{\mathbf{r}_t}$	1	0	0	$\to 1$	$\to 1$	< 1	< 1
$\tau_{\mathbf{r}_t}$	1	0	$< \dfrac{1}{4}$	$\to 1$	$\to 1$	< 1	$< \dfrac{1}{2}$

FIG. 6. Behavior of the Twist for the different run archetypes. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

TABLE 1. Features of the adopted experimental collections.

Feature	TREC 10	TREC 14	TREC 20	TREC 21
Track	Web	Robust	Web	Web
Corpus	WT10g	AQUAINT	ClueWeb09	ClueWeb09
# documents	1.7M	1.0M	1,040.0M	1,040.0M
# topics	50	50	50	50
# runs	95	74	37	27
Run length	1,000	1,000	10,000	10,000
Relevance degrees	3	3	4	4
Pool depth	100	55	25	30 and 25

TABLE 2. Classification of the CRP curves for the considered test collections; note that there are no full-scale runs and that up and down and excellent (type A and B) runs are conflated into the excellent category.

	Total	Ideal	Worst	Typical A	Typical B	Excellent
TREC10	4,750	0.17%	3.09%	83.87%	12.72%	0.15%
TREC14	3,700	0.00%	1.86%	80.24%	17.89%	0.00%
TREC20	1,850	0.11%	8.86%	60.70%	30.16%	0.16%
TREC21	1,350	0.00%	4.15%	64.59%	31.26%	0.00%

These collections have been chosen because of their different characteristics, which allows us to evaluate the proposed measure in a heterogeneous setting: different size of the document corpus, large number of runs, runs of different length, difficult topics, Web scale and real topics, shallow pools, and different number of relevance grades. TREC 10 is a collection of English web documents with graded relevance judgments for which a large number of experiments have been submitted. The length of the submitted runs is 1,000 documents. TREC 14 has a smaller document corpus than TREC 10 and 50 topics with graded relevance judgments that had been demonstrated to be difficult in previous TREC campaigns; the length of the submitted runs is 1,000 documents. The goal of the Robust track was to focus research on improving the consistency of retrieval technology by concentrating on poorly performing topics (Voorhees, 2005). TREC 20 presents a huge multilingual web corpus—that is, more than one billion documents—topics are created from the logs of a commercial search engine and it allows us to evaluate up-to-date IR systems working on a Web scale. In the ad hoc task we consider here, it has 50 topics with six-level relevance assessment (i.e., spam, not relevant, relevant, highly relevant, keyword and navigational); we conflated the spam degree into not relevant and keyword into the highly relevant one, thus working with

four relevance degrees. Note that with respect to previous TREC campaigns (i.e., TREC 18 and 19) where topics were chosen to be of medium-to-high frequency, TREC 20 attempted to work with more obscure topics, which may still be underspecified (i.e., faceted), but should be less ambiguous (Clarke et al., 2012). The length of the submitted runs is 10,000 documents, all the submitted runs were judged to depth 25. Two types of runs were submitted to this track: category A runs, which used the whole collection, and category B runs, which used a subset of about 50 million English-language pages. In the following, we consider all the submitted runs. Last, TREC 21 has the same characteristics of TREC 20, but it introduced 50 new topics; 25 topics were judged to depth 30 and 25 to depth 20 (Clarke et al., 2013). Both TREC 20 and 21 allow us to experiment with actual shallow pools.

In Table 2, we report the CRP curve archetypes classification for the runs of each considered test collection. We can see that the vast majority of runs—that is, about 95% for each collection—are classified as typical A or typical B archetypes confirming the fact that these are the most common CRP curves type we may encounter when evaluating a system from the effort point-of-view; the ratio between typical A and typical B runs is unbalanced towards typical A archetypes for all the collections, even if the difference between the two is smaller in recent collections—that is, TREC 20 and TREC 21—where we find up to 30% of
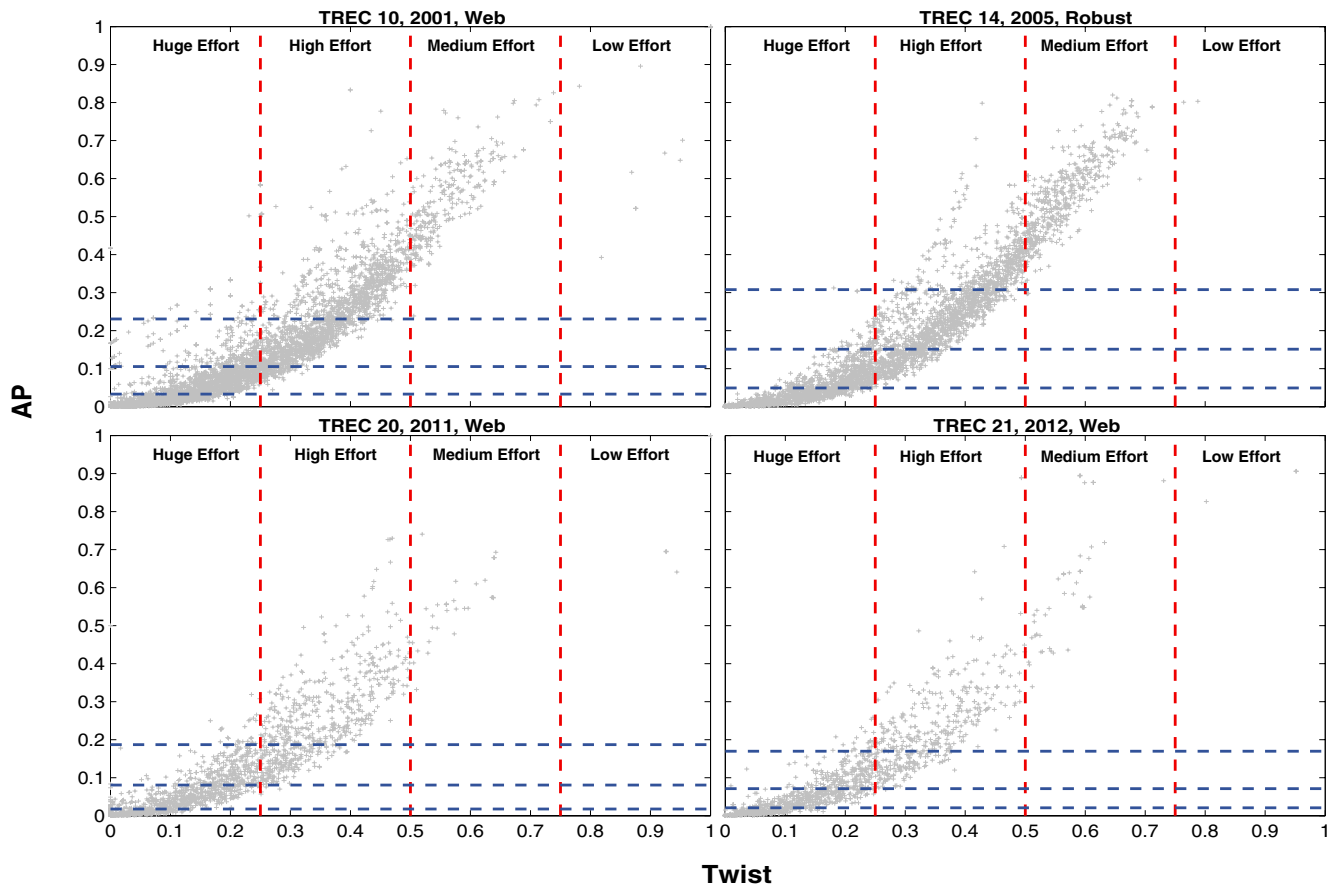
Effort/Gain: AP over Twist

FIG. 7. Effort/gain plot of AP against Twist for all the runs in the four test collections. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

typical B curves. TREC 20 and 21 also report a higher percentage of worst curves than TREC 10 and TREC 14. From the analysis of the CRP curve archetypes we can see that TREC 10 presents all kinds of curve archetypes—the full-scale is not present in any collection because it is an abstraction—as well as TREC 20; whereas TREC 14 and TREC 21 present mostly typical A and typical B curves along with some worst runs.

Effort/Gain Plots

In this section we answer to the question introduced above: "How much did a given utility/gain cost in terms of effort?" which allows us to differentiate among runs which look similar from a utility/gain perspective. For example, two or more runs that have almost similar AP and thus look almost equivalent from the utility/gain perspective may have quite different values for Twist and thus look quite different from the effort perspective.

Therefore, we compare the effort-based Twist measure with four widely adopted gain-based measures, which are AP (Buckley & Voorhees, 2005), binary preference (bref) (Buckley & Voorhees, 2004), rank-biased precision (RBP) (Moffat & Zobel, 2008) with persistence value set to 0.8 and normalized discounted comulated gain (nDCG) (Kekäläinen & Järvelin, 2002), calculated at the last retrieved document with discount function set to log base 10 and a base 5 weighting scheme (e.g., hr = 10, pr = 5, nr = 0 for a collection with three relevance degrees); the comparison between Twist and AP is shown in Figure 7, Twist and bpref in Figure 8, Twist and RBP in Figure 9, and Twist and nDCG in Figure 10.

Each figure reports four scatter plots comparing one of the considered gain-based measures—such as AP and DCG—and the Twist measure and the Twist measure for each test collection presented previously. In these plots, we draw a point for each topic of each run in the given test collection; the *x* coordinate reports the Twist value, whereas the *y* coordinate the gain-based measure value. Each plot
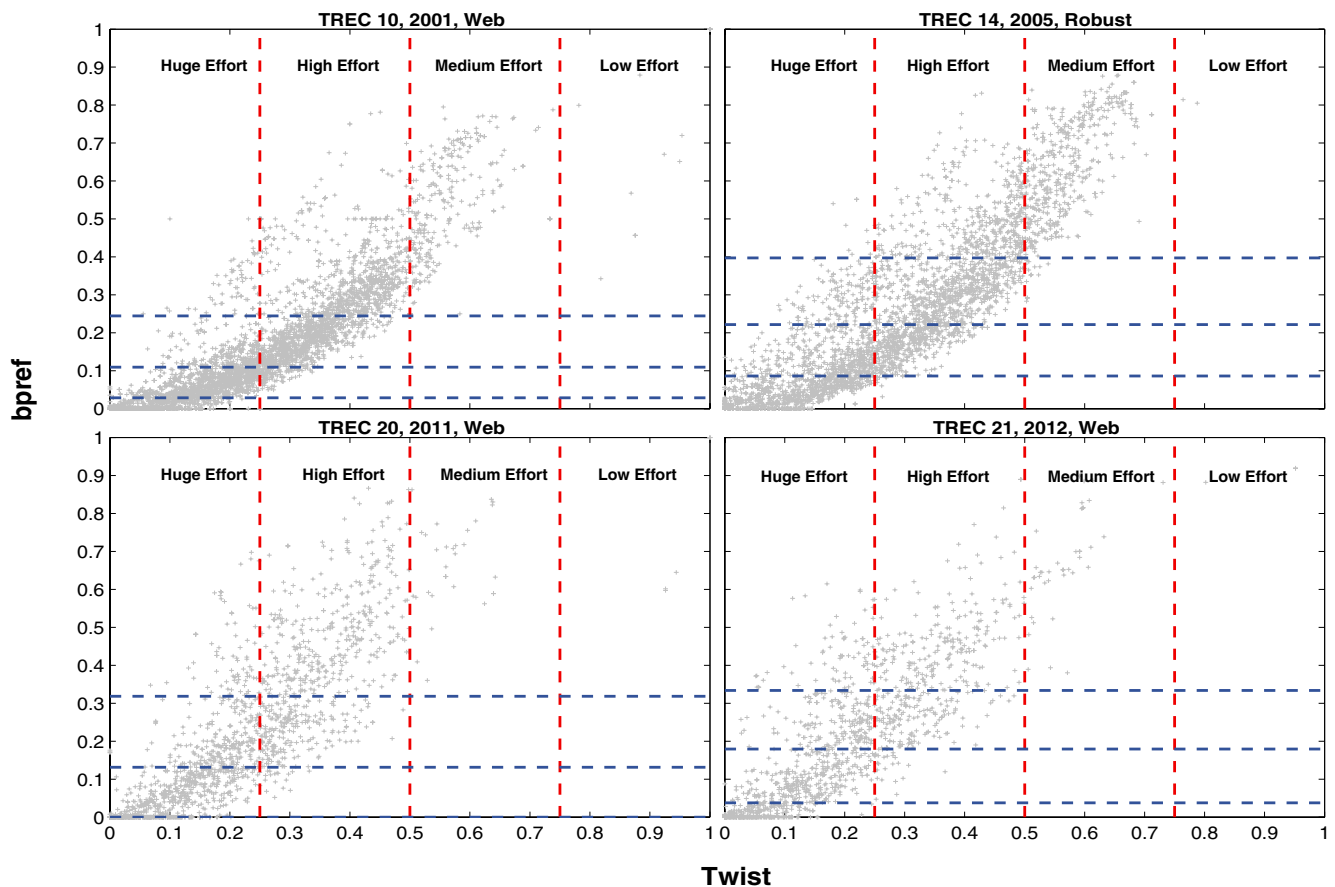
FIG. 8.   Effort/gain plot of bpref against Twist for all the runs in the four test collections. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

shows a grid composed of four rows and four columns; the rows correspond to the four quantiles of the gain-based values such that the points in the lower row correspond to runs which achieved a low overall gain, those in the second row a medium gain, those in the third row a high gain and those in the topmost row a huge gain. Whereas, the columns correspond to the bounds of the Twist measure summarized in Figure 6, where we can see that the "worst" and the "full-scale" like runs, which translate into a huge effort for the user, have a Twist value between 0 and 0.25. "Typical B" runs—that is, those for which the CRP curve never crosses the x axis thus requiring a high effort for the user—have a Twist value always lower than 0.5. The remaining runs are the good ones classified into the "typical A," "excellent," or "ups-and-down" archetypes; we can distinguish between good runs requiring medium effort when the Twist value is between 0.5 and 0.75 (i.e., "typical A" run) and excellent runs requiring low effort when the Twist value is above the 0.75 threshold.

In this way each plot is divided into 16 quadrants, each one containing the points representing the runs classified on an effort/gain basis. The runs lying in the diagonal quadrants going from the lower left to the upper right are those which are evaluated in the same way, both from the effort and the utility/gain point-of-view; indeed, the runs in the lower left quadrant are those with low gain-based measure and low Twist measure, thus they require a huge effort to achieve low gain, whereas the runs in the upper right quadrant are those with a huge gain achieved requiring low effort. Below and above the diagonal quadrants we find all the runs that needs to be evaluated both from the gain and the effort point-of-view to be correctly discriminated. As an example, the runs in the upper left quadrant are excellent from the gain point-of-view, but very bad from the effort perspective; in the quadrant next to this (same row, but second column) we find excellent runs from the gain point-of-view, but still demanding considerable effort from the user. A good system should lie in the four quadrants in the upper-right part of the plot, where we achieve a high or huge gain demanding medium or low effort by the user.

However, with the intent of showing the complementarity of Twist with respect to other gain-based measures, we are
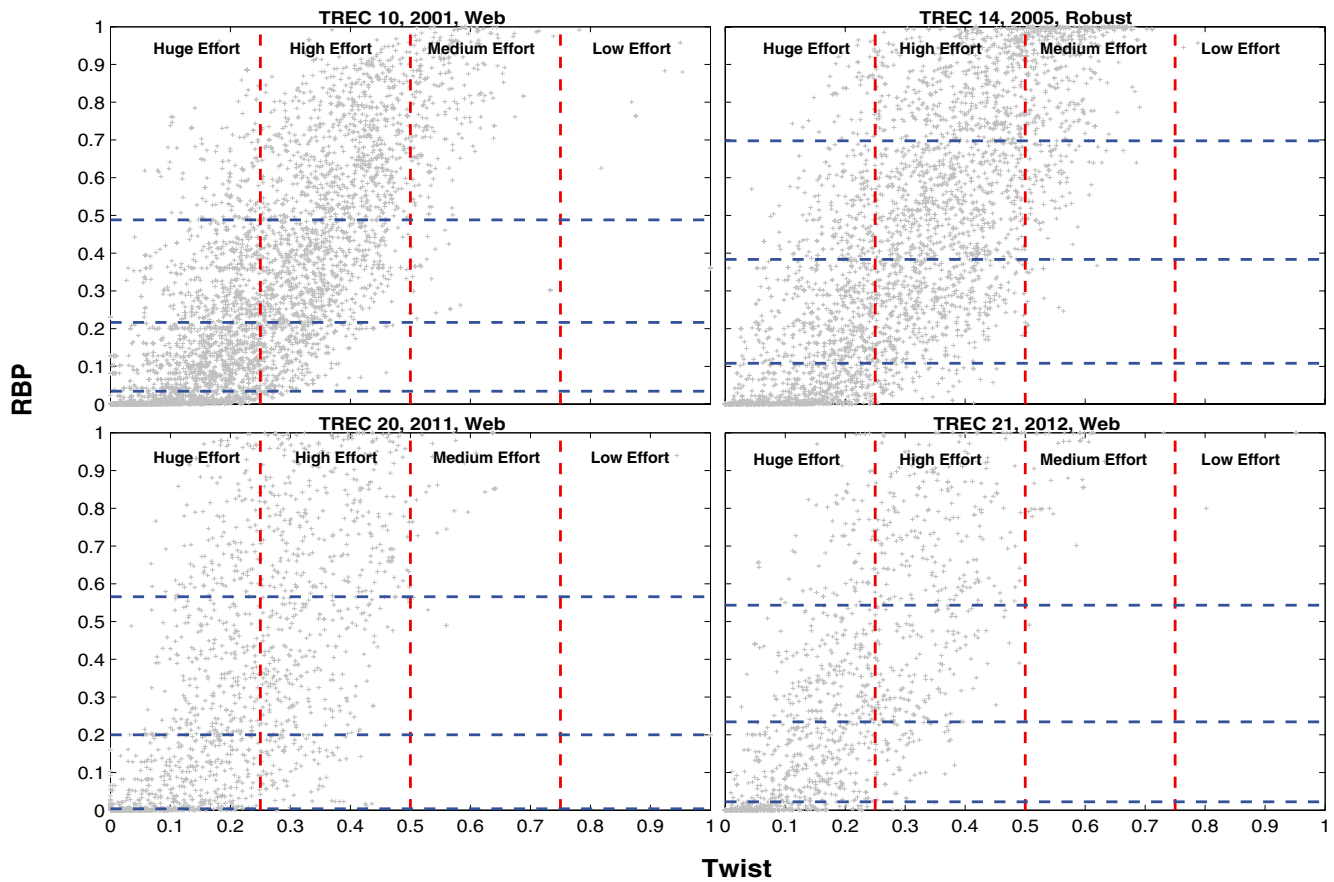
Effort/Gain: RBP over Twist

FIG. 9.   Effort/gain plot of RBP against Twist for all the runs in the four test collections. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

here mainly interested in the runs in the four upper-left quadrants—that is, those for which considerable effort is required for achieving a high or huge gain; for simplicity, in the following we call these four quadrants the "high-high quadrants." The runs in the high-high quadrants cannot be properly evaluated from the utility/gain perspective alone or from the effort perspective alone, but they need both these aspects to correctly discriminate between good or bad runs. In the following, we see that the percentage of runs that can be correctly discriminated by using only a gain-based or the Twist measure is always lower than the percentage of runs that requires both the gain-based and the Twist measure.

In Figure 7 we see the plot of the Twist measure against AP for the four considered collections; the behavior of the runs is consistent across all the collections and we can identify a concave-up trend of the points in the scatter plot. For TREC 10 about 28% of runs lie in the diagonal quadrants and thus can be correctly discriminated either by AP or Twist. Some 22% lie in the quadrant where the runs achieve medium gain requiring a huge effort (i.e., second row, first column) and about 6% lie in the quadrant where

they achieve huge gain requiring medium effort (i.e., fourth row, third column); these runs are in a gray area where a gain-based measure or the Twist measure still can evaluate their behavior fairly well. There are 44% of runs that reside in the high-high quadrants and thus are considered very good if evaluated from the AP perspective only and very bad if evaluated from the Twist perspective only. Note that less than a quarter of the runs achieving the highest AP values for TREC 10 (i.e., the ones in the topmost row of the plot) requires medium or low effort by the user and thus should be considered the best performing runs for TREC 10.

The scatter plot of TREC 14 runs follows the same trend of TREC 10 where only 37% of the run reside in the high-high quadrants and thus require both AP and Twist to be correctly evaluated. For TREC 20 the percentage of runs lying in the diagonal quadrants is consistent with the other collections (i.e., 26%), but about half of the runs lie in high-high quadrants and thus cannot be correctly discriminated by AP or Twist alone. Note that in this case, a very small number of runs achieving the best AP values
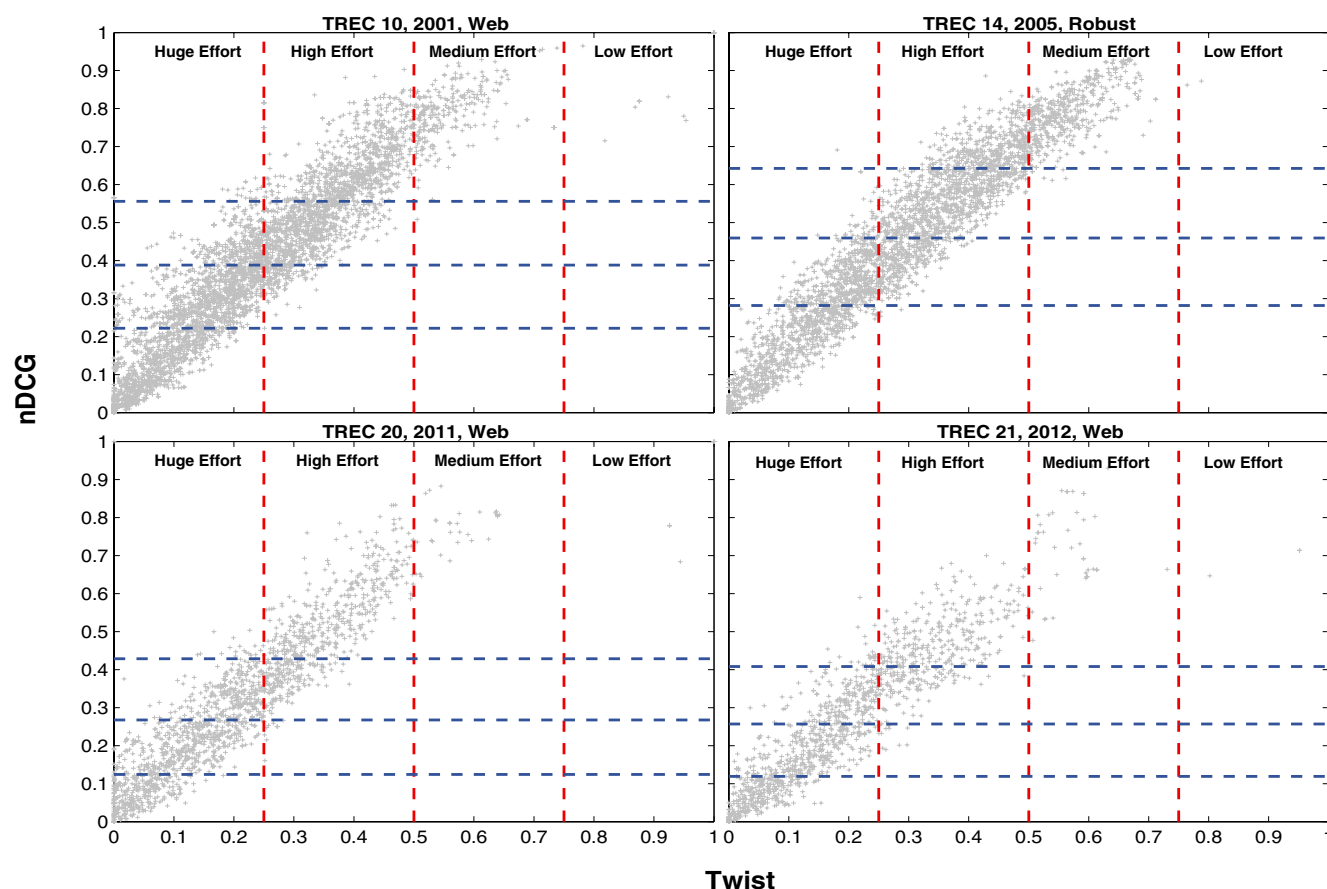
Effort/Gain: nDCG over Twist

FIG. 10. Effort/gain plot of nDCG against Twist for all the runs in the four test collections. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

demands medium or low effort by the user (i.e., less than 10%). TREC 21 follows a similar trend with a quarter of runs lying in the diagonal quadrants and half in the high-high quadrants. From this analysis we can conclude that the TREC 14 Robust track presents the majority of runs that can be correctly evaluated considering only the AP perspective (even though they are less than half of the total), whereas the three Web tracks (TREC 10, 20, and 21) collections present about half of their runs in the high-high quadrants, which require evaluation both from the AP and the Twist perspectives.

In general, for all the collections the percentage of runs in the high-high quadrants is higher than those in the diagonal ones, showing that the joint use of AP and Twist can improve our understanding of run behavior in the majority of cases.

In Figure 8 we report the scatter plots for bpref against Twist. We see that, as expected, these plots are very similar to those of AP against Twist; indeed, AP and bpref are highly correlated measures (Buckley & Voorhees, 2004) and evaluate very similar aspects of a run. The points in the plots

of bpref are a little bit more scattered than the AP ones, but the percentage of runs in the diagonal quadrants and in the high-high quadrants are the same as reported previously; the sole relevant exception is that the percentage of runs for TREC 20 in the diagonal quadrants goes from 26% of AP against Twist to 1.24% of bpref because many runs that were in the diagonal for the AP against Twist plot, in this case moved slightly upward in the gray area below the high-high quadrants.

In Figure 9 we report the RBP over Twist scatter plots. We can see that the points in this case are widely scattered across the 16 quadrants grid and it is hard to recognize a clear trend as we have done before for AP and bpref against Twist. RBP against Twist plots follow a very similar distribution of values as the previous ones. For TREC 10 we have 31% of the total runs in the diagonal quadrants and 44% in the high-high ones. For TREC 14, 38% of the total runs are in the diagonal quadrants and 37.4% in the high-high ones; in this case the percentage of runs achieving the best RBP values, which requires medium or low effort, is 10% lower than in the case of AP and bpref. For TREC 20 one third of

runs are in the diagonal quadrants (similar to AP) and half of them are in the high-high ones. Finally, TREC 21 follows exactly the same distribution as TREC 20 with a third of runs in the diagonal quadrants and half of them in the high-high ones.

For these plots as well as for the AP and bpref ones, the percentage of runs in the diagonal quadrants is significantly lower than those in the high-high quadrants for all the considered test collections.

In Figure 10 we report the scatter plot of nDCG against Twist. In this case we can recognize a slightly concave-down curve of the points rather different from the trend of AP and bpref. We can also notice more ample quantiles than in the previous cases. Despite these considerations, the distribution of runs across the collections follows exactly the same trend as in the AP against Twist and bpref against Twist plots.

From this analysis we can conclude that three largely adopted measures which evaluate utility/gain aspects of a run, if taken alone, do not discriminate well for 37% up to 48% of the runs submitted to four TREC ad hoc campaigns. This shows that utility is one important aspect, but it must be weighted by also considering the avoidable effort required.

CRP as a Visual Tool

The Twist measure is the ideal companion of the gain-based measures because it helps to spot those systems requiring considerable effort from the user to achieve a certain level of gain. The CRP curve of a run can be paired with the DCG to evaluate the behavior of a system from the qualitative point-of-view by considering both utility and effort aspects. Previously we discussed some effort/gain plots showing how Twist can discriminate between equally good runs from the utility/gain viewpoint in terms of avoidable effort required on the part of the user; such plots can be the starting point for a thorough analysis of IR systems behavior by integrating them with plots of the CRP and DCG curves of a run as shown in Figure 11.

In Figure 11 we report the effort/gain plot of nDCG against Twist for the TREC 10 test collection, where we highlighted four runs belonging to the top 25% nDCG values. All of these runs are considered good from the utility/gain point of view, but if we take into account the avoidable effort they require to achieve that gain we see that they are different from each other. The CRP curve serves as a visual aid for an in-depth exploration of the behavior of such runs and the pairing with the DCG curve allows for a rank-by-rank comparison in terms of effort versus gain.

Two runs created by the "jscbtawtl4" system are highlighted (i.e., the square and triangle) for topic 539 and topic 544. From the nDCG point-of-view both runs are in the fourth quantile and thus they are considered to be very good, whereas the first requires huge effort and the second only medium effort. From the comparison between the CRP and the DCG curves, we can see that for topic 539, DCG shows a curve composed of several steps indicating a (slow) growing trend, whereas CRP shows a full-scale like curve where very few (or no) relevant documents are retrieved in the first positions and most of the relevant documents are found among the subsequent ranks. By looking to the DCG plot we cannot say where there are misplaced documents and how much they are misplaced. Furthermore, it is difficult to observe the differences in early ranks, whereas in CRP they are easily recognizable. Indeed, we can see that in the first twenty positions DCG remains flat whereas CRP shows a negative trend suggesting that relevant documents have been misplaced with nonrelevant ones. The second plot regarding this system (i.e., plot (c) for topic 544) reports a DCG curve with a more evident growth trend than the previous one; when this curve is compared with the CRP one, we can see that in the first 250 ranks the system misplaces many relevant documents and requires a certain effort by the user to achieve the utility shown in the DCG graph.

Now, if we compare this plot (plot (c), jscbtawtl4, topic 544) with the plot referring to the run highlighted with a circle (plot (b), hum01t, topic 544), we can appreciate the role of the CRP curve as a fundamental tool for discriminating between equally good gain-oriented runs for the same topic. Indeed, the DCG curves are almost identical, showing two very good runs; on the other hand, the CRP curves are very different from each other, the first (plot(b), hum01t, topic 544) cross the x axis after rank 900 and it is a de facto typical B run, whereas the second crosses the x axis after around rank 600 showing the trend of a typical A run. Furthermore, the typical B-like run shows a slow growing trend in the second part of the curve, with large parts where it stays flat; whereas the typical A curve shows a curve that grows constantly in the second part of the graph, meaning that it retrieves enough relevant documents to move up in the positive region of the space.

The last plot we report (plot (d), pir1wt2, topic 504) reports CRP and DCG for a very good run from both the gain and the effort point-of-view. We can see that the nDCG of this run is very close to the one reported in plot (c) (they are both higher than 0.8) and also the DCG plots show two positive curves even though the second grows much faster. On the other hand, CRP is quite different; plot (d) resembles an excellent run, whereas plot (c) a typical A run. Plot (d) reports the CRP curve of a run which misplaces very few documents in the first positions and that recovers the misplacements soon enough to cross the x axis within the first 50 ranks. Also in this case CRP complements DCG curves adding important information for discriminating between equally good runs from the utility/gain point-of-view.

Therefore, CRP can be exploited as a visual tool that can be employed for supporting and improving failure analysis (Buckley, 2004; Harman, 2008; Savoy, 2007), which is deemed a fundamental activity in experimental evaluation and system development. The effort/gain plots can be used for spotting the runs that achieve equivalent utility for the
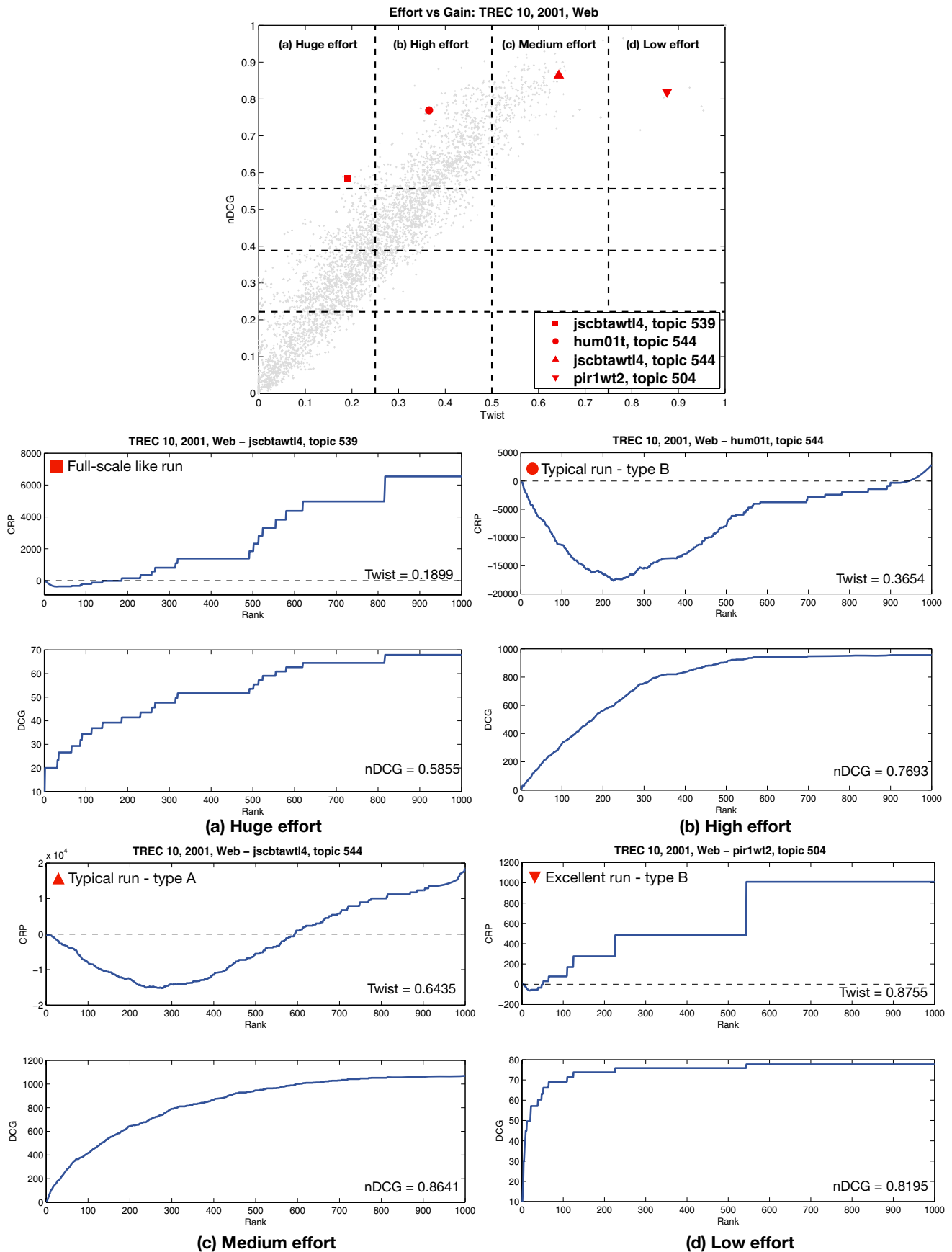
FIG. 11.   Visual analysis by means of the CRP and the DCG curves of four runs with fixed gain and effort ranging from huge to low. For sake of readability the plots have different scales on the vertical axes.
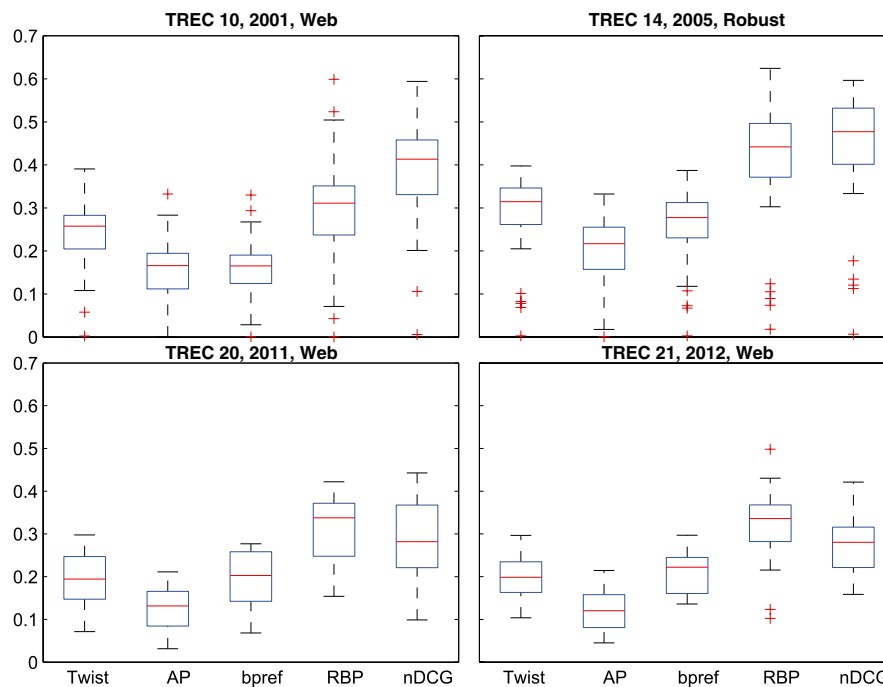
FIG. 12. Distribution of the average values of the considered measures for all the employed test collections. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

user, but report a great variation from the avoidable effort perspective. Once these runs have been identified, the CRP together with the DCG curve allow for a deeper analysis of a run by helping the researcher to understand its behavior rank-by-rank from both the utility and the effort viewpoint. Such a tool can be seen as an extension of visual systems that help the researcher to perform failure analysis; in particular, the state-of-the-art Visual Information Retrieval Tool for Upfront Evaluation (VIRTUE) system (Angelini et al., 2014) could benefit from the CRP visual tool because it could extend the use of RP, which is currently employed in the system for spotting critical regions of a ranking and grasp possible causes of a failure.

Experiments

This section compares the Twist measure with the widely used gain-based IR measures we presented previously: AP, bpref, RBP, and nDCG. To this purpose, we test Twist across three dimensions by determining:

- The correlation among measures using Kendall's Tau (Kendall, 1948; Voorhees, 2001);
- The robustness of the measures to downsampled pools according to the stratified random sampling method (Buckley & Voorhees, 2004);
- The discriminative power of the measures by employing the paired bootstrap test (Sakai, 2006, 2012, 2014).

We conducted these experiments on the four test collections described previously: TREC 10 2001 Web Track,

TREC 14 2005 Robust Track, TREC 20 2011 Web Track, and TREC 21 2012 Web Track, whose features are reported in Table 1 on page 19. All these collections use graded relevance judgments. For those measures that rely on binary relevance—namely, AP, bpref, and RBP—we adopted a "lenient" mapping, that is, every document above not relevant is considered as binary relevant.

There are systems that perform very poorly; to prevent these uninteresting systems from affecting the experiments, we consider only the top 75% of runs (Voorhees & Buckley, 2002; Webber, Moffat, Zobel, & Sakai, 2008) as measured by mean average precision (MAP); see Figure 12 for the distribution of the average values for all the considered measures.

The full source code of the software used to conduct the experiments is available for download[4] to ease comparison and verification of the results.

Correlations of Measures

Kendall's Tau rank correlation estimates the distance between two run rankings obtained by employing two different IR measures (Kekäläinen, 2005; Voorhees, 2001). This method is useful to show if and how the Twist measure derives different run rankings with respect to other measures; in related works (Sakai, 2007b; Voorhees, 2001) Kendall's Tau has been used to measure the correlation between two measures by analyzing the rankings of runs

---

[4]http://matters.dei.unipd.it/

TABLE 3. Kendall's Tau correlations between AP and the other utility-based measures.

	bpref	RBP	nDCG
AP (TREC10)	0.8506	0.6041	0.8334
AP (TREC14)	0.8390	0.7229	**0.9070**
AP (TREC20)	0.8498	0.5675	0.8962
AP (TREC21)	0.8461	0.5669	0.8388

Note. The reported values are statistically significant with *p* values well below 1%.

TABLE 4. Kendall's Tau correlations between Twist and the utility-based measures.

	AP	bpref	RBP	nDCG
Twist (TREC10)	0.8144	0.8173	0.6489	0.8791
Twist (TREC14)	0.8477	0.7547	0.6051	**0.9012**
Twist (TREC20)	0.8254	0.8148	0.3968	**0.9245**
Twist (TREC21)	**0.9157**	0.8526	0.3684	0.82105

Note. The reported values are statistically significant with *p* values well below 1%.

they produce. Measures with correlations greater than 0.9 should be considered similar and those with "less than 0.8 generally reflect noticeable changes in rankings, and suggest that the evaluation schemes have different emphasis" (Voorhees, 2001, p. 78). We study the run rankings produced by the gain-based measures presented previously and Twist; in previous studies (Kekäläinen, 2005; Sakai, 2007b) it has been shown how gain-based measures such as AP, bpref, and nDCG produce similar rankings showing a high correlation between them.

Table 3 reports the Kendall's Tau correlation between the base measures; it is consistent with previous findings, with a higher correlation between AP, nDCG, and bpref and a lower correlation for RBP. Indeed, RBP is the measure least correlated to the others because it is the only one not depending on the recall base. The correlation between AP and nDCG decreases in the case of the TREC 21 collection, where there are more grades of relevance with respect to the other collections and this increases the gap between the flat binary view of AP and the multi-graded one of nDCG.

Table 4 reports the Kendall's Tau correlations between AP, bpref, RBP, nDCG, and the Twist measure for the four test collections we considered.

As a general trend, Twist shows some correlation with the utility-based measures but never very high ones, suggesting that it takes a different perspective—the effort—with respect to these measures, since they have different document utility models, as discussed also in the introduction. Moreover, extremely low correlations would have been a symptom of possible misbehaviors of Twist (Sakai, 2014), but this is not the case with the present values.

Twist exhibits the highest correlation with nDCG and this can be explained by considering that (a) both are graded measures, (b) they adopt the same browsing model and stopping behavior, and (c) they somehow share a common approach based on cumulating either gains for nDCG or misplacements for Twist, which represents the document utility accumulation model. This behavior is also consistent with the effort/gain plots of Figure 10 which show a linear correlation. There is only one slight exception, the case of TREC 21, but this is probably more due to the fact that nDCG behaves differently on this collection with respect to the others, as will be seen below.

Twist and AP show a moderate Kendall's Tau rank correlation, highlighting the divide between binary and graded measures and the fact that the complementary approach of Twist based on effort can make it a good companion for AP, the de-facto standard measure in IR. Their difference, greater than that with nDCG, is also confirmed by the effort/gain plots of Figure 7 that shows a nonlinear correlation.

AP and bpref are known to be highly correlated (Buckley & Voorhees, 2004), and this is also reflected in their correlation with Twist, which exhibits a similar behavior, even though the Kendall's Tau of bpref and Twist are generally lower than one with AP, and the effort/gain plots of bpref in Figure 8 are slightly more scattered than those with AP.

Finally, Twist manifests the lowest correlations with RBP. This is consistent with the fact that RBP and Twist use two different browsing models, where $p = 0.8$ in the case of RBP indicates a more persistent user than the one entailed by Twist. In addition, they have two completely different document utility and document utility accumulation models, as explained in the introduction. Moreover, this can be explained by the fact that, unlike Twist, RBP does not take the recall base into account at all and it is very top heavy biased. This is further exacerbated for the TREC 20 and TREC 21 collections where the length of the runs is ten times the length of the runs in the other collections combined with the use of very shallow pools.

Robustness of Measures to Pool Downsampling

The stratified random sampling of the pools allows us to investigate the behavior of the Twist measure as relevance judgment sets become less complete following the methodology presented in Buckley and Voorhees (2004), which is here adapted to the case of multigraded relevance. For each topic, a separate list of documents at each relevance grade (not relevant, relevant, highly relevant, . . .) has been created from the original pool; for each sampling ratio *P*%, we selected $X = P\% \times D$ documents at the given relevance level, ensuring that at least 1 somehow relevant document and at least 10 not relevant documents are selected; the first max(1, *X*) documents from the random list at each relevant level have then been selected to constitute the new reduced pool; each smaller pool is a subset of each larger pool because we always select from the top of the lists. We used *P*% = [90, 70, 50, 30, 10]. Most effectiveness measures are known to be
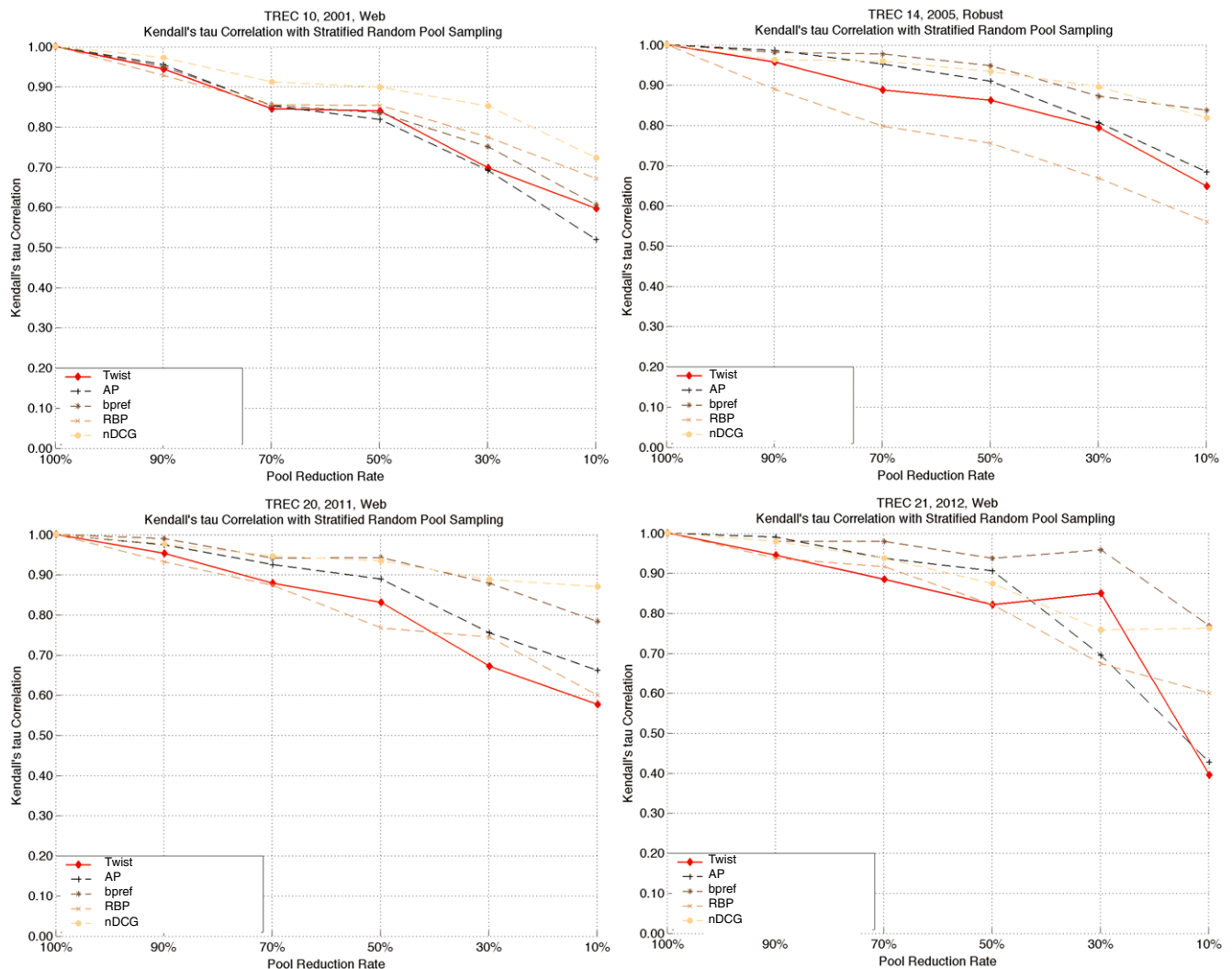
FIG. 13.   Change in Kendall's Tau correlation as judgment sets are downsampled. Each marker shows the value of the correlation between systems ranking at progressively reduced pools. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

"unstable for very small numbers of relevant documents" (Buckley & Voorhees, 2004, p. 29) and also for this reason, the bpref measure (implemented by following the definition given in Soboroff [2006], which has empirically shown to be robust to downsampling, has been introduced as a term of comparison with the Twist measure.

The plots in Figure 13 show the Kendall's Tau correlations between the system rankings produced using progressively downsampled pools from 100% (complete pool) to 10%. Each line shows the behavior of a measure; the flatter (and closer to 1.0) the line, the more robust the measure. In fact, a flat line indicates that the measure continues to rank systems in the same relative order with different levels of relevance judgments incompleteness.

We can see that the most robust measures are bpref and nDCG which achieve comparable levels of robustness for all the considered test collections. The Twist measure behaves quite well for all the collections; for TREC 10 up to a 50% downsampling it is as robust as bpref, AP, and RBP, whereas at 10% it is more robust than AP and as robust as bpref. For

TREC 14, Twist is more robust than RBP for all levels of incompleteness and it is very close to the behavior of AP. For TREC 20, it is more robust than RBP up to 50% and they are very close also at 10%. For TREC 21 all the measures are very close up to 50% and then Twist is more robust than AP, nDCG, and RBP at 30% while it behaves slightly worse than AP at 10%.

Twist proves to be fairly robust to downsampling and in several cases it is comparable to bpref, a measure explicitly built with this goal. It can be used as a companion of gain-based measures also when incomplete relevance judgments are considered.

Discriminative Power of Measures

The paired bootstrap test implemented as described in Sakai (2006) is used to examine the discriminative power of a metric which is desirable because it allows the experimenter to achieve reliable results with fewer topics (Webber, Moffat, & Zobel, 2010). The discriminative power
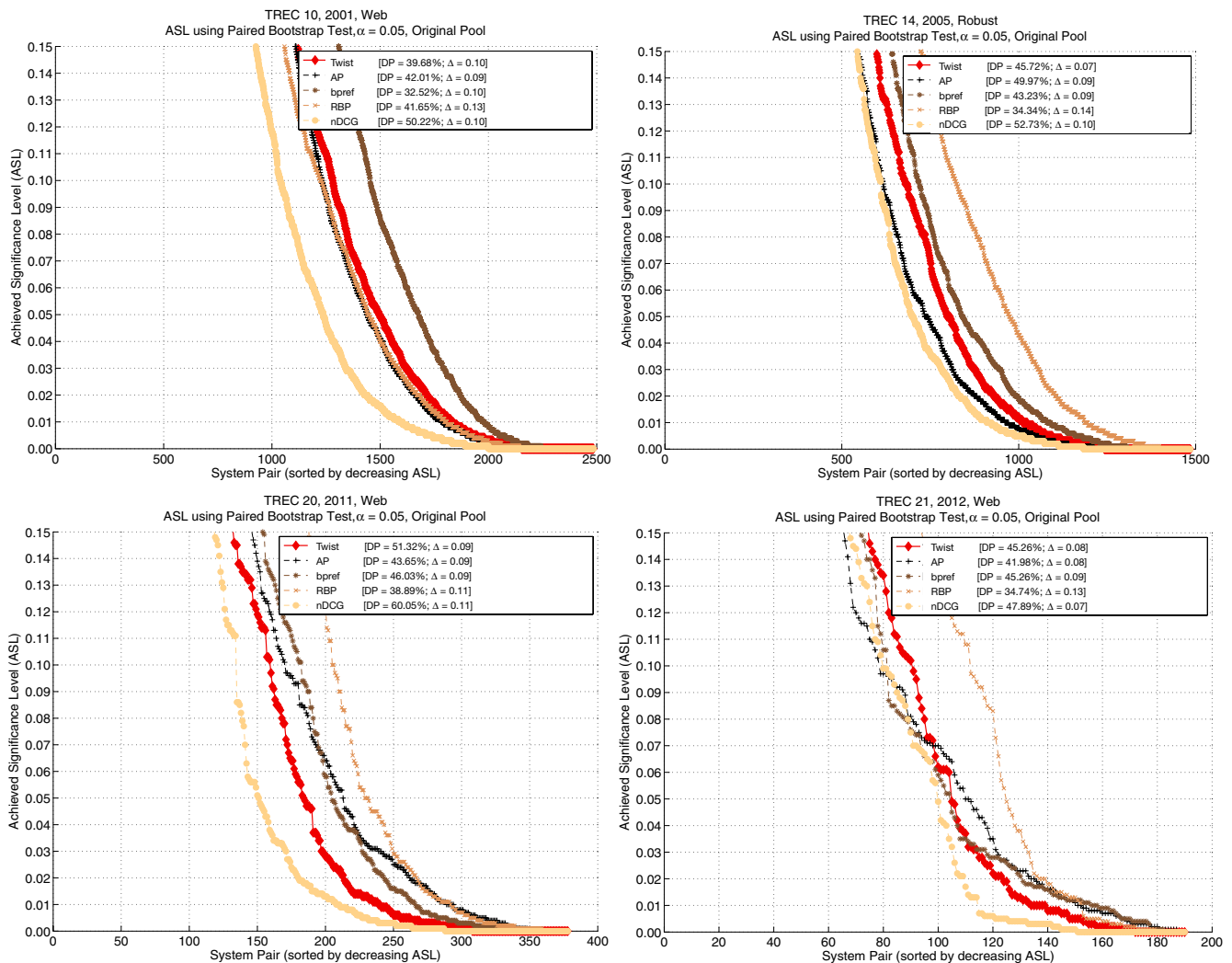
FIG. 14. Achieved significance level (ASL) curves based on paired bootstrap hypothesis tests for all test collections. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

of a measure indicates the achieved significance level (ASL), that is, the "proportion of statistically significant differences one can get out of a given experimental environment and therefore a measure of how reliable a metric is" (Sakai, 2012, p. 499). Given a test collection with a set of runs, the discriminative power is measured by conducting a boostrapped paired t test for every pair of runs and counting the number of significant differences; the test also estimates the performance Δ required in order to achieve statistical significance at a given confidence level α.

In Figure 14 we can see the ASL curves at $\alpha = 0.05$ of the considered measures for all the employed test collections. The curves on the left show a higher discriminative power than the curves on the right, since they report a larger number of system pairs achieving that value of ASL. In the legend of the plots we report the discriminative power (i.e., DP) indicating the percentage of system pairs discriminated by a measure (the higher the better) and Δ reporting the overall absolute difference required to state that one run is better than another with the given measure.

We can see that for TREC 10, Twist has a discriminative power comparable to RBP and AP, whereas it is more discriminative than bpref. For TREC 14, nDCG and AP are the most discriminative measures, Twist is in the middle, being more discriminative than both RBP and bpref. For TREC 20 and TREC 21 only nDCG is more discriminative than Twist which behaves better than RBP, AP, and bpref for both the collections. Twist reports low values for Δ for all the collections; in particular, for TREC 14 and TREC 20 Twist reports the lowest value, whereas for TREC 10 AP and for TREC 21 nDCG have slightly lower values.

In the end, Twist shows a good discriminative power; being always comparable to the one of the most used gain-based measures.

Related Metrics and Discussion

The novel Twist measure presented in this paper has several advantages when compared with previous metrics aimed at evaluating informational search intents—that is,

the user wants to find as many relevant documents as possible without wasting time on analyzing not-relevant documents (Broder, 2002). In the preceding section, we compared Twist with popular measures having similar purposes such as AP, bpref, RBP, and nDCG. In this section we discuss the differences between Twist and its predecessors. As stated in the Introduction, and as it will emerge from the following discussion, all these measures are centered around the idea of utility, measuring the level of utility provided to the user and reducing it in the case of nonoptimal rankings, whereas Twist revolves around the idea of search as a commodity and avoidable effort for the user.

It is out of scope for this paper to consider navigational tasks, the goal of which is to return a specific document to the user (Broder, 2002), and thus the main metrics for evaluating them—expected reciprocal rank (ERR) (Chapelle et al., 2011), normalized cumulative utility (NCU) (Sakai, Robertson, & Newswatch, 2008), and $P+$ (Sakai, 2012)—are not presented here.

Although AP does not handle graded relevance, it is worth considering as the starting point for other metrics which extend it and at the same time exploit its robustness. The main limitation of AP is that "it is based on the assumption that retrieved documents can be considered as either relevant or nonrelevant to a user's information need" (Robertson, Kanoulas, & Yilmaz, 2010, p. 603). Furthermore, as pointed out in Carterette, Kanoulas, and Yilmaz (2012, p. 116), AP "while interpretable in terms of an implicit user model, was originally defined in the absence of any underlying model of user interaction with the retrieved ranking of documents." For this reason, several extensions have been proposed in the literature. One of the principal ones is described in Robertson (2008) and Sakai et al. (2008); it is based on a user model considering users scanning a ranking list and stopping when a relevant document is found. A uniform distribution across all the relevant documents is assumed allowing for the calculation of a utility value for each document. In this context, AP can measure the expected utility of a ranking list for the considered population. This was improved in the definition of graded average precision (Robertson et al., 2010), which adopts a probabilistic user model defining to what "extent documents of different relevance grades account for the effectiveness score" (Robertson et al., 2010, p. 604). (Kekäläinen & Järvelin, 2002) propose a similar extension of AP to "generalized AP" to handle graded relevance assessments.

Buckley and Voorhees (2004) proposed bpref, which is highly correlated with AP when full relevance assessments are available and is yet more robust when the relevance assessments are reduced (Sakai & Kando, 2008). bpref as well as AP does not handle graded relevance judgments and it is based on the idea of measuring "the effectiveness of a system on the basis of judged documents only" (Buckley & Voorhees, 2004, p. 26). bpref has been described as "robust in the face of incomplete relevance information" (Buckley & Voorhees, 2004, p. 26), thus it is the most important measure to compare with when we deal with incomplete relevance

judgments. In the Experiments we noted that bpref is one of the more robust measures (overcome only by nDCG for some collection) and that Twist is comparable to it (for TREC 10 Twist is even more robust than bpref).

RBP (Moffat & Zobel, 2008) is a metric that addresses this very limitation; it is defined by starting from the observation that a user has no desire to examine every item in a ranking list. The idea is that a user progresses from a document to the other with probability p and, conversely, ends her/his examination of the list at a point with probability $1 - p$. This assumption allows for the definition of user models representing patient and impatient user populations by varying the probability p as DCG does by exploiting different document weighting schemas and log bases for the discounting function. The current definition of the Twist measure is parameter-free—as AP is—and it constitutes the main building block for an effort-oriented measure that can be extended in order to model different user populations as DCG and RBP do. As an example, patient and impatient users could be modeled by weighing document misplacements where RP values are boosted or discounted by tuning one or more parameters based on document relevance degrees, weighting schemas and/or rank positions. In this way it would be possible to distinguish the behavior of a single user and the expected behavior of a user population.

On the other hand, we have shown (see Figure 14), in line with Sakai and Kando 2008), that RBP has a low discriminative power if compared to other gain-based measures and to Twist for evaluation with complete relevance assessments. With high values of p (i.e., $p = 0.95$) RBP has a high discriminative power, but it is not stable when downsampled pools are employed, whereas Twist has been proven to strike a good balance between sensitiveness and robustness to incomplete judgments. Unlike Twist, RBP does not depend on the recall base and independence from it is a feature of the metric because it allows for evaluation also with incomplete relevance judgments (Moffat & Zobel, 2008). On the other hand, this aspect can turn out to be a weakness because RBP "may give a very low score even to an ideal ranked output: the fact that it does not rely on recall implies that it denies the very existence of an ideal ranked output" (Sakai & Kando, 2008, p. 453).

RBP shares some similarities with the DCG and nDCG (Järvelin & Kekäläinen, 2002; Moffat & Zobel, 2008), because for any rank examined, it gives an estimate of the (normalized, discounted) cumulated gain as a single figure no matter what the recall base size. Cumulated gain metrics are not heavily dependent on relevant documents found late in the ranked order since they focus on the gain cumulated from the beginning of the result up to any point of interest. The discounted versions realistically weigh down the gain received through documents found later in the ranked results.

From a qualitative point of view, the CRP curves indicate upward and downward swings depending on the misplacements in document ranking. However, the gain values of DCG grow monotonically unless negative gain values

(proposed by Keskustalo, Järvelin, Pirkola, & Kekäläinen [2008]) are used. Like the CRP, the normalized versions compare the ranking quality to each topic's entire recall base allowing statistical comparability. Both the CRP and the CG-based metrics with negative weights address the issue of suboptimal ranking of search results but in different ways. The CRP indicates suboptimal ranking directly through the CRP curve; when this curve deviates from the x axis (representing the ideal ranking), ranking is suboptimal and less relevant documents are retrieved earlier than they should be. The CG-based metrics do not directly address ranking optimality but cumulate gain and loss (or negative gain), whereas CRP quantifies the misplacements and thus, the suboptimality of a ranking.

In summary, the Twist measure explicitly handles graded relevance and takes into account document misplacements either too early or too late given their degree of relevance and the ideal ranking. The CRP curve, from which Twist is derived, can be analyzed rank-by-rank and to determine if a document is ranked too early or too late (i.e., deviation from the ideal) from the graphical—that is, by examining the plot—and the quantitative point of view—that is, the Twist measure. The differences between Twist and the other measures are highlighted by the low correlations reported in the Experiments section; no other metric explicitly weights the document misplacements in a ranking and this fact directly reflects on the low correlation between them.

Conclusions

In traditional test collection-based evaluation, the evaluation task is simplified by abstracting away users, their situations and tasks; this neglects user experiences caused by browsing sequences of nonrelevant or suboptimal documents. In this work, we have argued that nowadays search is considered as a commodity and that many users require high quality documents but simultaneously risk losing (some of) them due to the limited number of ranks inspected when users become weary. In this context, it is important to provide novel metrics that are more sensitive than the traditional ones in expressing the effects of document misplacements (with respect to the ideal) at any rank.

To this end, we developed a novel measure—the Twist τ—which explicitly addresses these aspects and presents good properties of robustness to shallow pool and discriminative power. We believe that, in modern large and increasingly complex environments, the proposed metric should be employed whenever possible, because it provides an ideal companion to traditional measures, allowing us to understand not only which systems provide high utility/gain to the user but also at what price.

In future work, we will investigate how Twist behaves when not only ordinal scales are used but also interval and ratio ones, i.e., when instead of using the relevance degrees, relevance weights are used to amplify document misplacement. We foresee the definition of a weighted version of the Twist measure that will allow us to model different user populations.

We also plan to conduct a deeper investigation of the Twist measure. As suggested by Tague-Sutcliffe (1992), statistical significance tests play a key role in experimental evaluation and, considering the effort/gain plots, they could be applied to understand whether and when significant differences in a gain/utility-based measures such as AP or nDCG correspond or not to significant differences in the Twist measure both within the same quadrant and across different quadrants.

Moreover, we would like to investigate how Twist is related to other approaches for taking user effort into account. More specifically, the time spent by a user in carrying out an information access task is usually considered as an indicator of the effort required of her/him (Ingwersen & Järvelin, 2005; Järvelin, 2013). This will also call for studies that explore the correlation between Twist and actual user behavior to understand to what extent the effort reported by Twist is a good predictor of the user's way of interacting with the ranked result list. It would be interesting to know whether there is a correlation between Twist and the effort measures in an interactive user study as, for example, has been done by Smucker and Jethani (2010) to relate human performance in information access with the precision measure. In this respect, another avenue is to explore whether it is possible to make Twist a time-calibrated measure, as for example do Smucker and Clarke (2012a, 2012b) by substituting the notion of rank with the notion of time.

---

[5]http://www.promise-noe.eu/

[6]http://www.cultura-strep.eu/

[7]http://www.preforma-project.eu/

References

Angelini, M., Ferro, N., Järvelin, K., Keskustalo, H., Pirkola, A., Santucci, G., & Silvello, G. (2012a). Cumulated relative position: A metric for ranking evaluation. In T. Catarci, P. Forner, D. Hiemstra, A. Peñas, & G. Santucci (Eds.), Information access evaluation. Multilinguality, multimodality, and visual analytics. Proceedings of the Third International Conference of the CLEF Initiative (CLEF 2012) (pp. 112–123), Lecture Notes in Computer Science (LNCS) 7488. Heidelberg, Germany: Springer.

Angelini, M., Ferro, N., Santucci, G., & Silvello, G. (2012b). Visual interactive failure analysis: Supporting users in information retrieval evaluation. In J. Kamps, W. Kraaij, & N. Fuhr (Eds.), Proceedings of the Fourth Symposium on Information Interaction in Context (IIiX 2012; pp. 195–203). New York, NY: ACM Press.

Angelini, M., Ferro, N., Santucci, G., & Silvello, G. (2014). VIRTUE: A visual tool for information retrieval performance evaluation and failure analysis. Journal of Visual Languages and Computing, 25(4), 394–413.

Broder, A. (2002). A taxonomy of Web search. SIGIR Forum, 36(2), 3–10.

Buckley, C. (2004). Why current IR engines fail. In M. Sanderson, K. Järvelin, J. Allan, & P. Bruza (Eds.), Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004) (pp. 584–585). New York, NY: ACM Press.

Buckley, C., & Voorhees, E.M. (2004). Retrieval evaluation with incomplete information. In M. Sanderson, K. Järvelin, J. Allan, & P. Bruza (Eds.), Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004) (pp. 584–585). New York, NY: ACM Press.

Buckley, C., & Voorhees, E.M. (2005). Retrieval system evaluation. In D.K. Harman & E.M. Voorhees (Eds.), TREC. Experiment and evaluation in information retrieval (pp. 53–78). Cambridge, MA: MIT Press.

Carterette, B., Kanoulas, E., & Yilmaz, E. (2012). Evaluating web retrieval effectiveness. In D. Lewandowsky (Ed.), Web search engine research, library and information science (pp. 105–138). Bingley, UK: Emerald Group Publisher Limited.

Carterette, B.A. (2011). System effectiveness, user models, and user utility: A conceptual framework for investigation. In W.-Y. Ma, J.-Y. Nie, R. Baeza-Yaetes, T.-S. Chua, & W.B. Croft (Eds.), Proceedings of the 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2011) (pp. 903–912). New York, NY: ACM Press.

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

Clarke, C.L.A., Craswell, N., & Voorhees, H. (2012). Overview of the TREC 2011 Web track. In E.M. Voorhees & L.P. Buckland (Eds.), The 20th Text REtrieval Conference Proceedings (TREC 2011) (pp. 1–8), Special Publication 500-295. Washington, DC: National Institute of Standards and Technology.

Clarke, C.L.A., Craswell, N., & Voorhees, H. (2013). Overview of the TREC 2012 web track. In E.M. Voorhees & L.P. Buckland (Eds.), The 21st Text REtrieval Conference Proceedings (TREC 2012) (pp. 1–8), Special Pubblication 500-298. Washington, DC: National Institute of Standards and Technology.

Cleverdon, C.W. (1997). The Cranfield tests on index languages devices. In K. Spärck Jones & P. Willett (Eds.), Readings in information retrieval (pp. 47–60). San Francisco, CA: Morgan Kaufmann Publisher, Inc.

Di Buccio, E., Dussin, M., Ferro, N., Masiero, I., Santucci, G., & Tino, G. (2011a). Interactive analysis and exploration of experimental evaluation results. In M.L. Wilson, T. Russell-Rose, B. Larsen, & J. Kalbach (Eds.), Proceedings of the First European Workshop on Human-Computer Interaction and Information Retrieval (EuroHCIR 2011) (pp. 11–14). Retrieved from http://ceur-ws.org/Vol-763/

Di Buccio, E., Dussin, M., Ferro, N., Masiero, I., Santucci, G., & Tino, G. (2011b). To re-rank or to re-query: Can visual analytics solve this dilemma? In P. Forner, J. Gonzalo, J. Kekäläinen, M. Lalmas, & M. de Rijke (Eds.), Multilingual and multimodal information access evaluation. Proceedings of the Second International Conference of the Cross-Language Evaluation Forum (CLEF 2011) (pp. 119–130), Lecture Notes in Computer Science (LNCS) 6941. Heidelberg, Germany: Springer.

Egghe, L. (2008). The measures precision, recall, fallout and miss as a function of the number of retrieved documents and their mutual interrelations. Information Processing and Management, 44(2), 856–876.

Ferger, W.F. (1931). The nature and use of the harmonic mean. Journal of the American Statistical Association, 26(173), 36–40.

Ferro, N., Sabetta, A., Santucci, G., & Tino, G. (2011). Visual comparison of ranked result cumulated gains. In S. Miksch & G. Santucci (Eds.), Proceedings of the Second International Workshop on Visual Analytics (EuroVA 2011) (pp. 21–24). Goslar, Germany: Eurographics Association.

Guns, R., Lioma, C., & Larsen, B. (2012). The tipping point: F-score as a function of the number of retrieved items. Information Processing and Management, 48(6), 1171–1180.

Harman, D.K. (2008). Some thoughts on failure analysis for noisy data. In D. Lopresti, S. Roy, K. Schulz, & L. Venkata Subramaniam (Eds.), Proceedings of the Second Workshop on Analytics for Noisy unstructured text Data (AND 2008) (pp. 1–1). New York, NY: ACM Press.

Harman, D.K. (2011). Information retrieval evaluation. San Rafael, CA: Morgan & Claypool Publishers.

Harman, D.K., & Voorhees, E.M. (Eds.). (2005). TREC. Experiment and evaluation in information retrieval. Cambridge, MA: MIT Press.

Hawking, D., & Craswell, N. (2001). Overview of the TREC-2001 Web track. In E. Voorhees & D.K. Harman (Eds.), The Tenth Text REtrieval Conference (TREC 2001) (pp. 61–67), IST Special Publication 500-250. Washington, USA: Department of Commerce, National Institute of Standards and Technology.

Ingwersen, P., & Järvelin, K. (2005). The turn: Integration of information seeking and retrieval in context. Heidelberg, Germany: Springer.

Jansen, B.J., Booth, D.L., & Spink, A. (2008). Determining the informational, navigational, and transactional intent of web queries. Information Processing and Management, 44(3), 1251–1266.

Järvelin, K. (2013). User-oriented evaluation in IR. In M. Agosti, N. Ferro, P. Forner, H. Müller, & G. Santucci (Eds.), Information retrieval meets information visualization—PROMISE winter school 2012, revised tutorial lectures (pp. 86–91), Lecture Notes in Computer Science (LNCS) 7757. Heidelberg, Germany: Springer.

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems, 20(4), 422–446.

Kekäläinen, J. (2005). Binary and graded relevance in IR evaluations-comparison of the effects on ranking of IR systems. Information Processing and Management, 41(5), 1019–1033.

Kekäläinen, J., & Järvelin, K. (2002). Using graded relevance assessments in IR evaluation. Journal of the American Society for Information Science and Technology, 53(13), 1120–1129.

Kelly, D. (2009). Methods for evaluating interactive information retrieval systems with users. Foundations and Trends in Information Retrieval, 3(1–2), 1–224.

Kendall, M. (1948). Rank correlation methods. UK: Griffin.

Keskustalo, H., Järvelin, K., Pirkola, A., & Kekäläinen, J. (2008). Intuition-supporting visualization of user's performance based on explicit negative higher-order relevance. In T.-S. Chua, M.-K. Leong, D.W. Oard, & F. Sebastiani (Eds.), Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008). New York, NY: ACM Press.

Krantz, D.H., Luce, R.D., Suppes, P., & Tversky, A. (1971). Foundations of measurement. Additive and polynomial representations (Vol. 1). New York, NY: Academic Press.

Moffat, A., & Zobel, J. (2008). Rank-biased precision for measurement of retrieval effectiveness. ACM Transactions on Information Systems, 27(1), 1–27.

Moffat, A., Thomas, P., & Scholer, F. (2013). Users versus models: What observation tells us about effectiveness metrics. In A. Iyengar, Q. He, J. Pei, R. Rastogi, & W. Nejdl (Eds.), Proceedings of the 22nd International Conference on Information and Knowledge Management (CIKM 2013) (pp. 659–668). New York, NY: ACM Press.

Robertson, S. (2008). A new interpretation of average precision. In T.-S. Chua, M.-K. Leong, D.W. Oard, & F. Sebastiani (Eds.), Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008). New York, NY: ACM Press.

Robertson, S.E., Kanoulas, E., & Yilmaz, E. (2010). Extending average precision to graded relevance judgments. In F. Crestani, S. Marchand-Maillet, E.N. Efthimiadis, & J. Savoy (Eds.), Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2010). New York, NY: ACM Press.

Sakai, T. (2006). Evaluating evaluation metrics based on the bootstrap. In E.N. Efthimiadis, S. Dumais, D. Hawking, & K. Järvelin (Eds.), Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006). New York, NY: ACM Press.

Sakai, T. (2007a). Alternatives to bpref. In W. Kraaij, A.P. de Vries, C.L.A. Clarke, N. Fuhr, & N. Kando (Eds.), Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007) (pp. 71–78). New York, NY: ACM Press.

Sakai, T. (2007b). On the reliability of information retrieval metrics based on graded relevance. Information Processing and Management, 43(2), 531–548.

Sakai, T. (2012). Evaluation with informational and navigational intents. In A. Mille, F.L. Gandon, J. Misselis, M. Rabinovich, & S. Staab (Eds.), Proceedings of the 21st International Conference on World Wide Web (WWW 2012) (pp. 499–508). New York, NY: ACM Press.

Sakai, T. (2014). Metrics, statistics, tests. In N. Ferro (Ed.), Bridging between information retrieval and databases—PROMISE winter school 2013, revised tutorial lectures (pp. 116–163), Lecture Notes in Computer Science (LNCS) 8173. Heidelberg, Germany: Springer.

Sakai, T., & Kando, N. (2008). On information retrieval metrics designed for evaluation with incomplete relevance assessments. Information Retrieval, 11(5), 447–470.

Sakai, T., Robertson, S., & Newswatch, I. (2008). Modelling a user population for designing information retrieval metrics. In T. Sakai, M. Sanderson, N. Kando, & M. Sugimoto (Eds.), Proceedings of the Second Workshop on Evaluating Information Access (EVIA 2008) (pp. 30–41). Tokyo, Japan: National Institute of Informatics.

Sanderson, M. (2010). Test collection based evaluation of information retrieval systems. Foundations and Trends in Information Retrieval, 4(4), 247–375.

Savoy, J. (2007). Why do successful search systems fail for some topics. In Y. Cho, Y. Wan Koo, R.L. Wainwright, H.M. Haddad, & S.Y. Shin (Eds.), Proceedings of the 2007 ACM Symposium on Applied Computing (SAC 2007) (pp. 872–877). New York, NY: ACM Press.

Smucker, M.D., & Clarke, C.L.A. (2012a). Stochastic simulation of time-biased gain. In X. Chen, G. Lebanon, H. Wang, & M.J. Zaki (Eds.), Proceedings of the 21st International Conference on Information and Knowledge Management (CIKM 2012) (pp. 2040–2044). New York, NY: ACM Press.

Smucker, M.D., & Clarke, C.L.A. (2012b). Time-based calibration of effectiveness measures. In W. Hersh, J. Callan, Y. Maarek, & M. Sanderson (Eds.), Proceedingds of the 35th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2012) (pp. 95–104). New York, NY: ACM Press.

Smucker, M.D., & Jethani, C.P. (2010). Human performance and retrieval precision revisited. In F. Crestani, S. Marchand-Maillet, E.N. Efthimiadis, & J. Savoy (Eds.), Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2010). New York, NY: ACM Press.

Soboroff, I. (2006). Dynamic test collections: Measuring search effectiveness on the live Web. In E.N. Efthimiadis, S. Dumais, D. Hawking, & K. Järvelin (Eds.), Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006). New York, NY: ACM Press.

Stevens, S.S. (1946). On the theory of scales of measurement. Science, New Series, 103(2684), 677–680.

Stevens, S.S. (1955). On the averaging of data. Science, New Series, 121(3135), 113–116.

Tague-Sutcliffe, J.M. (1992). The pragmatics of information retrieval experimentation, revisited. Information Processing and Management, 28(4), 467–490.

Tague-Sutcliffe, J.M. (1996). Some perspectives on the evaluation of information retrieval systems. Journal of the American Society for Information Science, 47(1), 1–3.

Toms, E. (2011). Task-based information searching and retrieval. In I. Ruthven & D. Kelly (Eds.), Interactive information seeking, behaviour and retrieval (pp. 43–59). London, UK: Facet Publishing.

Voorhees, E. (2001). Evaluation by highly relevant documents. In D.H. Kraft, W.B. Croft, D.J. Harper, & J. Zobel (Eds.), Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001) (pp. 74–82). New York, NY: ACM Press.

Voorhees, E.M., & Buckley, C. (2002). The effect of topic set size on retrieval experiment error. In K. Järvelin, M. Beaulieu, R. Baeza-Yates, & S. Hyon Myaeng (Eds.), Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2002) (pp. 316–323). New York, NY: ACM Press.

Voorhees, H. (2005). Overview of the TREC 2005 robust retrieval track. In E.M. Voorhees & L.P. Buckland (Eds.), The 14th Text REtrieval Conference Proceedings (TREC 2005) (pp. 11–20), Special Publication 500-266. Washington, DC: National Institute of Standards and Technology.

Webber, W., Moffat, A., Zobel, J., & Sakai, T. (2008). Precision-at-ten considered redundant. In T.-S. Chua, M.-K. Leong, D.W. Oard, & F. Sebastiani (Eds.), Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008). New York, NY: ACM Press.

Webber, W., Moffat, A., & Zobel, J. (2010). The effect of pooling and evaluation depth on metric stability. In T. Sakai, M. Sanderson, W. Webber, N. Kando, & K. Kishida (Eds.), Proceedings of the Third International Workshop on Evaluating Information Access (EVIA 2010) (pp. 7–15). Kyoto, Japan: National Institute of Informatics.

Appendix

Preliminary Definitions

Let D be a finite set of **documents**; $d \in D$ a document, i.e. the basic information unit; T a finite set of **topics**; and, $t \in T$ a topic, i.e. the materialization of a user information need.

Definition 7. *Let REL be a finite set of **relevance degrees** and let $\preceq$ be a total order relation on REL so that*

$$(REL, \preceq)$$

*is a totally ordered set. We call **nonrelevant** the relevance degree* nr $\in REL$ *such that*

$$\text{nr} = \min(REL)$$

Being a finite totally ordered set, the set of relevance degrees admits the existence of a minimum and a maximum.

Definition 8. *Let D be a finite set of* documents *and T a finite set of* topics. *The **ground truth** is a function*

$$\begin{aligned} GT: \quad T \times D &\rightarrow REL \\ (t, d) &\mapsto rel \end{aligned}$$

Definition 9. *The **recall base** is a function*

$$\text{RB}: \quad T \quad \to \quad \mathbb{N}$$
$$t \quad \mapsto \quad RB_t = |\{d \in D \mid GT(t, d) \succ \min(REL)\}|$$

Note that usually the size of the set of documents is much much larger than the recall base for any given topic, i.e. $|D| \gg 2RB_t, \forall t \in T$.

Definition 10. *Given a natural number $N \in \mathbb{N}^+$ called the* length of the *run, a **run** is a function*

$$\text{R}: \quad T \quad \to \quad D^N$$
$$t \quad \mapsto \quad \mathbf{r}_t = (d_1, d_2, \dots, d_N)$$

such that $\forall t \in T, \forall j, k \in [1, N] \mid j \neq k \Rightarrow \mathbf{r}_t[j] \neq \mathbf{r}_t[k]$ where $\mathbf{r}_t[j]$ denotes the j-th element of the vector $\mathbf{r}_t$, vectors start with index 1, and vectors end with index N.

Definition 11. *Given a run $\text{R}(t) = \mathbf{r}_t$, the **relevance score** of the run is a function:*

$$\hat{\text{R}}: \quad T \times D^N \quad \to \quad REL^N$$
$$(t, \mathbf{r}_t) \quad \mapsto \quad \hat{\mathbf{r}}_t = (rel_1, rel_2, \dots, rel_N)$$

where

$$\hat{\mathbf{r}}_t[j] = GT(t, \mathbf{r}_t[j])$$

It is worth noting that, in general, the relevance score function is not injective since two different run vectors for two different topics may map to the same vector of relevance degrees; this is also intuitive from the fact that $|D| \gg |REL| \Rightarrow |D|^N \gg |REL|^N$ and so there are many more vectors of documents than vectors of relevance degrees.

Definition 12. *The **ideal run** $\text{I}(t) = \mathbf{i}_t$, where $N \geq RB_t$, is a run which satisfies the following constraints*

(1) recall base: $\forall t \in T, \quad |\{j \in [1, N] \| GT(t, \mathbf{i}_t[j]) \succ \min(REL)\}| = RB_t$
(2) ordering: $\forall t \in T, \quad \forall j, k \in [1, N] \mid j < k \Rightarrow \mathbf{i}_t[j] \succeq \mathbf{i}_t[k]$

Condition (1) ensures that all the relevant documents are retrieved in the ideal run while condition (2) guarantees that they are in descending order of relevance. Note that the ideal run actually defines a whole set of permutations of the documents with the same relevance degree.

From Definition 12 it follows that, for each topic $t \in T$, the relevance score of the ideal run $\hat{\mathbf{i}}_t$ is a monotonic non-increasing function by construction. Therefore, the maximum of the function is at $j = 1$ and it is equal to $\hat{\mathbf{i}}_t[1] = \max(REL)$ and the minimum is at $j = N$ and it is equal to $\hat{\mathbf{i}}_t[N] = \min(REL)$.[8]

---

[8]For the sake of accuracy, instead of max(*REL*), which is the maximum relevance degree possible, we should consider the maximum relevance degree actually present in the ground truth for a given topic, since it may happen that not all topics have documents for each relevance degree. This can be expressed as $max_t = \max GT(t, d), d \in \{d \in D\} | GT(t, d) > \min(REL)\}$. However, this difference does not impact the following results and so, with a slight abuse of notation, we will continue to use max(*REL*)

Definition 13. *The **worst run** $\text{W}(t) = \mathbf{w}_t$, where $N \geq RB_t$, is a run which satisfies the following constraint*

$$\forall t \in T, \forall j \in [1, N] \Rightarrow \hat{\mathbf{w}}_t[j] = \min(REL)$$

The worst run defines a set of permutations, all of which consist of single nonrelevant documents. Note that the worst run exists only if there are at least N nonrelevant documents in D.

Definition 14. *The **full-scale run** $\text{FS}(t) = \mathbf{fs}_t$, where $N \geq 2RB_t$, is a run which satisfies the following constraints*

(1) recall base: $\forall t \in T, \quad |\{j \in [1, N] \| GT(t, \mathbf{fs}_t[j]) \succ \min(REL)\}| = RB_t$
(2) ordering: $\forall t \in T, \quad \forall j, k \in [1, N] \mid j < k \Rightarrow \hat{\mathbf{fs}}_t[j] \succeq \hat{\mathbf{fs}}_t[k]$

Condition (1) ensures that all the relevant documents are retrieved in the full-scale run while Condition (2) guarantees that they are in descending order of relevance starting from the end of the vector. It is called full-scale run since it produces the minimum and maximum values of RP and CRP for a run of length N. It reverses the order of the ideal run, that is $\forall t \in T, \forall j \in [1, N] \Rightarrow \mathbf{fs}_t[j] = \mathbf{i}_t[N - j + 1]$.

From Definition 14 it follows that, for each topic $t \in T$, the relevance score of the full-scale run $\hat{\mathbf{fs}}_t$ is a monotonic non-decreasing function by construction. Therefore, the maximum of the function is at $j = N$ and it is equal to $\hat{\mathbf{fs}}_t[N] = \max(REL)$ and the minimum is at $j = 1$ and it is equal to $\hat{\mathbf{fs}}_t[1] = \min(REL)$.

Definition 15. *Given the ideal run $\text{I}(t)$ and a relevance degree $rel \in REL$ such that $\exists j \in [1, N] \mid \hat{\mathbf{i}}_t[j] = rel$, the **minimum rank** and the **maximum rank** are, respectively, a function*

$$\min_{\mathbf{i}_t}(rel): \quad T \times D^N \times REL \quad \to \quad \mathbb{N}^+$$
$$(t, \mathbf{i}_t, rel) \quad \mapsto \quad \min\left(\{j \in [1, N] \mid \hat{\mathbf{i}}_t[j] = rel\}\right)$$
$$\max_{\mathbf{i}_t}(rel): \quad T \times D^N \times REL \quad \to \quad \mathbb{N}^+$$
$$(t, \mathbf{i}_t, rel) \quad \mapsto \quad \max\left(\{j \in [1, N] \mid \hat{\mathbf{i}}_t[j] = rel\}\right)$$

The minimum rank is the first position at which we find a document with relevance degree equal to *rel* while the maximum rank is the last position at which we find a document with relevance degree equal to *rel* in the ideal run. Note that, by construction, we have: $\min_{\mathbf{i}_t}(\max(REL)) = 1$; $\min_{\mathbf{i}_t}(\min(REL)) = RB_t + \underline{1}$; $\max_{\mathbf{i}_t}(\min(REL)) = N$; and, given a relevance degree $rel \in REL$ strictly above $min(REL)$ and below any other relevance degree, i.e. $\underline{rel} \in REL \mid \overline{rel} \succ \min(REL) \wedge \forall rel_i \in REL, rel_i \neq \min(REL) \Rightarrow \overline{rel} \preceq rel_i$, $\max_{\mathbf{i}_t}(\overline{rel}) = RB_t$.

---

instead, for the sake of simplicity. Moreover, this problem does not apply to min(*REL*) since we can safely assume that each topic has at least a "not relevant" document.

This appendix reports the proof of propositions presented in Section "Definition of the Twist Measure".

Relative Position The following is the proof of Proposition 1 which provides the upper and lower bounds for RP.

Proof. We separately demonstrate each condition.

Condition 1. The lowest value possible for RP is given by the biggest misplacement of a "not relevant" document in the position of the first "most relevant" document, that is the position $j = \min_{\mathbf{i}_t}(\max(REL))$ where a run gets the value $\mathbf{rp}_{\mathbf{r}_t}[j] = j - \min_{\mathbf{i}_t}(\min(REL))$. By construction, this is the case of the full-scale run where $j = 1$ and $\mathbf{fs}_{\mathbf{r}_t}[1] = 1 - \min_{\mathbf{i}_t}(\mathrm{nr}) = 1 - (RB_t + 1) = -RB_t$.

Condition 2. The highest value possible for RP is given by the largest misplacement of a "most relevant" document in the position of the last "not relevant" document, that is the position $j = \max_{\mathbf{i}_t}(\min(REL))$ where a run gets the value $\mathbf{rp}_{\mathbf{r}_t}[j] = j - \min_{\mathbf{i}_t}(\max(REL))$. By construction, this is the case of the full-scale run where $j = N$ and $\mathbf{fs}_{\mathbf{r}_t}[N] = N - \max_{\mathbf{i}_t}(\max(REL))$. $\square$

Proposition 1 shows us that the range of RP for a run $\mathbf{r}_t$ is, typically, not balanced around zero since the absolute value of its upper bound $\max(\mathbf{rp}_{\mathbf{fs}_t})$ can be bigger than the absolute value of its lower bound $\min(\mathbf{rp}_{\mathbf{fs}_t})$. Indeed, in general, it holds $|N - \max_{\mathbf{i}_t}(\max(REL))| \ge |-RB_t| \Leftrightarrow N - \max_{\mathbf{i}_t}(\max(REL)) \ge RB_t$ since $N \ge 2RB_t$ and $\max_{\mathbf{i}_t}(\max(REL)) \le RB_t$; they are equal only in the case of binary relevance when $\max_{\mathbf{i}_t}(\max(REL)) = RB_t$ and of runs of length when $N = 2RB_t$. Therefore, RP usually may have larger absolute values in the positive region than in the negative one; this is also intuitive from the fact that the misplacement of a "not relevant" document in the interval of the relevant ones up to

RB_t is usually smaller than the misplacement of a relevant document in the interval of the "not relevant" ones from $RB_t + 1$ up to $N \ge 2RB_t$.

Cumulated Relative Position

The following is the proof of Proposition 1 which provides the upper and lower bounds for CRP.

Proof. We separately demonstrate each condition.

Condition 1. The lowest value possible for CRP of a run is achieved when all the "not relevant" documents are ranked up to the recall base, i.e. when any relevant document is missing in the first positions of the ranking up to RB_t. This is exactly the case of the full-scale run which leads to:

$$\mathbf{crp}_{\mathbf{fs}_t}[RB_t] = \sum_{k=1}^{RB_t} \mathbf{rp}_{\mathbf{fs}_t}[k] \overset{1}{=} \sum_{k=1}^{RB_t} [k - \min_{\mathbf{i}_t}(\min(REL))] =$$
$$\overset{2}{=} \sum_{k=1}^{RB_t} [k - (RB_t + 1)] = -RB_t(RB_t + 1) + \sum_{k=1}^{RB_t} k =$$
$$= -RB_t(RB_t + 1) + \frac{RB_t(RB_t + 1)}{2} = -\frac{RB_t(RB_t + 1)}{2}$$

where (1) comes from the definition of RP and the fact all the "not relevant" documents up to RB_t are above their ideal interval and so we are in the case $k < \min_{\mathbf{i}_t}(\min(REL))$; and, (2) comes from a previous observation which noted that $\min_{\mathbf{i}_t}(\min(REL)) = RB_t + 1$.

Condition 2. The highest value possible for CRP of a run is achieved when relevant documents are retrieved in the interval of the ranking corresponding to "not relevant" documents and the more relevant the documents are the more they are placed towards the end of the vector. This is exactly the case of the full-scale run which leads to:

$$\mathbf{crp}_{\mathbf{fs}_t}[N] = \sum_{k=1}^{N} \mathbf{rp}_{\mathbf{fs}_t}[k] \overset{1}{=} \sum_{k=1}^{RB_t} \mathbf{rp}_{\mathbf{fs}_t}[k] + \sum_{k=RB_t+1}^{N-RB_t} \mathbf{rp}_{\mathbf{fs}_t}[k] + \sum_{k=N-RB_t+1}^{N} \mathbf{rp}_{\mathbf{fs}_t}[k] =$$
$$\overset{2}{=} \sum_{k=1}^{RB_t} [k - \min_{\mathbf{i}_t}(\min(REL))] + \sum_{k=RB_t+1}^{N-RB_t} 0 + \sum_{k=N-RB_t+1}^{N} [k - \max_{\mathbf{i}_t}(\widehat{fs}_t[k])] =$$
$$\overset{3}{=} -\frac{RB_t(RB_t + 1)}{2} + \sum_{k=N-RB_t+1}^{N} k - \sum_{k=N-RB_t+1}^{N} \max_{\mathbf{i}_t}(\widehat{\mathbf{fs}}_t[k]) =$$
$$\overset{4}{=} -\frac{RB_t(RB_t + 1)}{2} + \sum_{p=1}^{RB_t} (p + N - RB_t) - \sum_{k=N-RB_t+1}^{N} \max_{\mathbf{i}_t}(\widehat{\mathbf{fs}}_t[k]) =$$
$$= -\frac{RB_t(RB_t + 1)}{2} + \frac{RB_t(RB_t + 1)}{2} + RB_t(N - RB_t) - \sum_{k=N-RB_t+1}^{N} \max_{\mathbf{i}_t}(\widehat{\mathbf{fs}}_t[k]) =$$
$$\overset{5}{=} RB_t(N - RB_t) - \sum_{rel_i \succ \min(REL)} \Delta_{rel_i} \sum_{\substack{rel_k \succ \min(REL) \\ rel_k \preceq rel_i}} \Delta_{rel_k}$$

where (1) splits the definition of RP in its three different intervals—above, ok, below—and (2) substitutes the corresponding values; (3) applies at the first interval—the upper one—the results just demonstrated for Condition 1; (4) substitutes the index of the sum with $p = k - (N - RB_t)$ to make it more explicit; and, (5) considers that $\sum_{k=N-RB_t+1}^{N} \max_{\mathbf{i}_t}(\widehat{\mathbf{fs}}_t[k])$ corresponds to sum, for each relevance degree above "not relevant," the total number of documents with that relevance degree multiplied by its maximum rank. This can be reformulated by noting that the total number of documents with a given relevance degree can be written as $\Delta_{rel_i} = \max_{\mathbf{i}_t}(rel_i) - \min_{\mathbf{i}_t}(rel_i) + 1$ and the maximum rank of a given relevance degree can be written as the sum of the total number of documents for all the relevance degrees above "not relevant" and up to the given one, that is $\sum_{\substack{rel_k \succ \min(REL) \\ rel_k \preceq rel_i}} \Delta_{rel_k}$. $\square$

Both propositions 1 and 1 show how the lower bounds of RP and CRP are only functions of the recall base RB_t while the upper bounds depend on the recall base RB_t, the length of the run N, and the intervals for the different relevance degrees.

Recovery Ratio

The following is the proof of Proposition 1 which provides the upper bounds for the recovery ratio of the full scale run.

Proof. By construction, the first position in the ranking at which the full-scale run retrieves a document with the smaller relevance degree above "not relevant" is $j = N - RB_t + 1$ and that is also the first chance that CRP has to become positive, if it gains enough, otherwise the crossing may happen later on, if it happens at all.

If there was no crossing, by definition of recovery ratio, we have $\rho_{\mathbf{fs}_t} = 0 < \dfrac{1}{2}$.

Otherwise, the crossing for the full-scale run is at $N - RB_t + 1 + k = (\omega - 1)RB_t + 1 + k$ where $k = 0, 1, \ldots, RB_t - 1$ allows us to look for a crossing up to the last rank position. By definition of recovery ratio, we have $\rho_{\mathbf{fs}_t} = \dfrac{RB_t}{N - RB_t + 1 + k} = \dfrac{RB_t}{(\omega - 1)RB_t + 1 + k}$. Ab absurdo, suppose that, in this case, it holds that $\rho_{\mathbf{fs}_t} \geq \dfrac{1}{2}$, so:

$$\frac{RB_t}{(\omega - 1)RB_t + 1 + k} \geq \frac{1}{2} \Leftrightarrow 2RB_t \geq (\omega - 1)RB_t + 1 + k$$

$$\Leftrightarrow (\omega - 3)RB_t \leq -(1 + k)$$

The left hand side is always greater than or equal to zero, since $\omega \geq 3$ and $RB_t \geq 0$; the right hand side is always strictly less than zero, since $k \geq 0$; so the inequality is never satisfied. Therefore, it must be $\rho_{\mathbf{fs}_t} < \dfrac{1}{2}$.

List of Symbols and Acronyms

This appendix reports the list of symbols and acronyms used throughout the paper.

List of Symbols

Symbol	Description
t	Topic
T	Set of topics
d	Document
D	Set of documents
rel	Relevance degree
REL	Set of relevance degrees
hr	Higly relevant relevance degree
fr	Fairly relevant relevance degree
pr	Partially relevant relevance degree
nr	Not relevant relevance degree
GT	Ground truth
RB_t	Recall base of topic t
$\mathbf{r}_t$	Run for topic t
$\mathbf{w}_t$	Worst run for topic t
$\mathbf{fs}_t$	Full-scale run for topic t
$\mathbf{i}_t$	Ideal run for topic t
N	Length of the run
$\min_{\mathbf{i}_t}(rel)$	First position at which we find a document with relevance degree equal to rel in the ideal run for topic t
$\max_{\mathbf{i}_t}(rel)$	Last position at which we find a document with relevance degree equal to rel in the ideal run for topic t
ρ	Recovery ratio
$x_{\mathbf{r}_t}$	Set of the crossings of the run with respect to the x axis
$\beta_{\mathbf{r}_t}$	Balance point of run $\mathbf{r}_t$
s^+	Forward space
s^-	Backward space
σ^+	Forward space ratio
σ^-	Backward space ratio
σ	Space ratio
τ	Twist measure

List of Acronyms

Acronym	Description
AP	Average Precision
bpref	Binary Preference
CRP	Cumulated Relative Position
DCG	Discounted Cumulated Gain
ERR	Expected Reciprocal Rank
IR	Information Retrieval
MAP	Mean Average Precision
NCU	Normalized Cumulative Utility
nDCG	normalized Discounted Cumulated Gain
RBP	Rank-Biased Precision
RP	Relative Position
TREC	Text REtrieval Conference
VIRTUE	Visual Information Retrieval Tool for Upfront Evaluation