Cumulated Relative Position: A Metric for Ranking Evaluation

Marco Angelini[3], Nicola Ferro[1], Kalervo Järvelin[2], Heikki Keskustalo[2], Ari Pirkola[2], Giuseppe Santucci[3], and Gianmaria Silvello[1]

[1] University of Padua, Italy
{ferro,silvello}@dei.unipd.it
[2] University of Tampere, Finland
{kalervo.jarvelin,heikki.keskustalo,ari.pirkola}@uta.fi
[3] "La Sapienza" University of Rome, Italy
{angelini,santucci}@dis.uniroma1.it

Abstract. The development of multilingual and multimedia information access systems calls for proper evaluation methodologies to ensure that they meet the expected user requirements and provide the desired effectiveness. IR research offers a strong evaluation methodology and a range of evaluation metrics, such as MAP and (n)DCG. In this paper, we propose a new metric for ranking evaluation, the CRP. We start with the observation that a document of a given degree of relevance may be ranked too early or too late regarding the ideal ranking of documents for a query. Its relative position may be negative, indicating too early ranking, zero indicating correct ranking, or positive, indicating too late ranking. By cumulating these relative rankings we indicate, at each ranked position, the net effect of document displacements, the CRP. We first define the metric formally and then discuss its properties, its relationship to prior metrics, and its visualization. Finally we propose different visualizations of CRP by exploiting a test collection to demonstrate its behavior.

1 Introduction

Designing, developing, and evaluating an *Information Retrieval (IR)* system is a challenging task, especially when it comes to understanding and analyzing the behavior of the system under different conditions in order to tune or to improve it as to achieve the level of effectiveness needed to meet the user expectations.

The development of information access systems calls for proper evaluation methodologies to ensure that they meet the expected user requirements and provide the desired effectiveness. IR research offers a strong evaluation methodology based on test collections [1]. A range of evaluation metrics, such as MAP and nDCG, are widely used within this methodology [2]. These metrics are particularly suitable to the evaluation of IR techniques in terms of the quality of the output ranked lists, and often to some degree suitable to the evaluation of user experience regarding retrieval. Unfortunately, the traditional metrics do

not take deviations from optimal document ranking sufficiently into account. For example, the *Mean Average Precision (MAP)* only considers precision at relevant document ranks and employs binary relevance. MAP nor its extensions to graded relevance [3,4] offer no explicit method for penalizing for suboptimal documents ranked early. Further, the original *Normalized Discounted Cumulated Gain ((n)DCG)* [5] only discounts the relevance gain of late-arriving relevant documents without penalizing for suboptimal documents ranked early. While the extension of (n)DCG [6] penalizes for retrieving non-relevant documents, it does not generally handle ranking suboptimal documents early and does not explicitly take into account the severity of document mis-ranking. We think that a proper evaluation metric for ranked result lists in IR should: (a) explicitly handle graded relevance including negative gains for unhelpful documents, and (b) explicitly take into account document misplacements in ranking either too early or too late given their degree of relevance and the optimal ranking. In the present paper, we propose such a new evaluation metric, the *Cumulated Relative Position (CRP)*.

We start with the observation that a document of a given degree of relevance may be ranked too early or too late regarding the ideal ranking of documents for a query. Its relative position may be negative, indicating too early ranking, zero indicating correct ranking, or positive, indicating too late ranking. By cumulating these relative rankings we indicate, at each ranked position, the net effect of document displacements, the CRP.

The novel CRP metric is related to prior metrics, such as sliding ratio [7], normalized recall [7,8], the satisfaction frustration total measure [7], and (n)DCG. However, CRP differs from these in explicitly handling: (a) graded relevance, and (b) document misplacements either too early or too late given their degree of relevance and the ideal ranking. Thereby, CRP offers several advantages in IR evaluation:

- at any number of retrieved documents examined (rank) for a given query, it is obvious to interpret and it gives an estimate of ranking performance as a single measure relative to the ideal ranking for the topic;
- it is not dependent on outliers since it focuses on the ranking of the result list;
- it is directly user-oriented in reporting the deviation from ideal ranking when examining a given number of documents; the effort wasted in examining a suboptimal ranking is made explicit;
- it allows conflation of relevance grades of documents and therefore more or less fine-grained analyses of the ranking performance of an IR technique may be produced;
- it can be summarized by four synthesis indicators describing the ranking quality of the IR system under investigation;
- it is possible to point out several graphical representations by stressing one of the different aspects of measurement allowed by CRP.

The rest of the paper is organized as follows: Section 2 presents the CRP and its properties. Section 3 presents a comparison between CRP and previous met-

rics and considerations about their ability of addressing the bi-directional nature of search results. Section 4 presents a visualization of the CRP by comparing it with the DCG and a visualization of the CRP synthesis indicators based on parallel coordinates. Lastly, Section 5 draws some final remarks and points-out future developments of CRP.

2 Cumulated Relative Position

2.1 Definition of the Metric

We define the set of *relevance degrees* as $(REL, \leq)$ such that there is an order between the elements of REL. For example, for the set $REL = \{\texttt{nr}, \texttt{pr}, \texttt{fr}, \texttt{hr}\}$, `nr` stands for "non relevant", `pr` for "partially relevant", `fr` for "fairly relevant", `hr` stands for "highly relevant", and it holds $\texttt{nr} \leq \texttt{pr} \leq \texttt{fr} \leq \texttt{hr}$.

We define a function $\mathrm{RW} : REL \to \mathbb{Z}$ as a monotonic function[4] which maps each relevance degree ($rel \in REL$) into an *relevance weight* ($w_{rel} \in \mathbb{Z}$), e.g. $\mathrm{RW}(\texttt{hr}) = 3$. This function allows us to associate an integer number to a relevance degree; much of the previous work studied the impact of varying these weights on *Cumulated Gain (CG)*, *Discounted Cumulated Gain (DCG)*, and (n)DCG measures [5,6].

We define with D the set of documents we take into account, with $N \in \mathbb{N}$ a natural number, and with D^N the set of all possible vectors of length N containing different orderings of the documents in D. We can also say that a vector in D^N represents a ranking list of length N of the documents D retrieved by an IR system. Let us consider a vector $\mathbf{v} \in D^N$, a natural number $j \in [1, N]$, and a relevance degree $rel \in REL$, then the *ground truth* function is defined as:

$$\begin{aligned} \mathrm{GT} : & D^N \times \mathbb{N} \to REL \\ & \mathbf{v}[j] \mapsto rel \end{aligned} \tag{1}$$

Equation 1 allows us to associate a relevance degree to the document $d \in D$ retrieved at position j of the vector $\mathbf{v}$, i.e. it associates a relevance judgment to each retrieved document in a ranked list.

In the following, we define with $\mathbf{r} \in D^N$ the vector of documents retrieved and ranked by a run r, with $\mathbf{i} \in D^N$ the ideal vector containing the best ranking of the documents in the pool (e.g. all highly relevant documents are grouped together in the beginning of the vector followed by fairly relevant ones and so on and so forth), and with $\mathbf{w} \in D^N$ the worst-case vector containing the worst rank of the documents retrieved by the pool (e.g. all the relevant documents are put in the end of the vector in the inverse relevance order).

In the following we use an example to explain the equations we introduce. Let us consider an ideal vector $\mathbf{i}$ composed of k intervals of documents sharing the same rel. We assume to have a pool composed by 20 elements where $k = 4$ and

[4] This means that $\forall \{rel_1, rel2\} \in REL \mid rel_1 \leq rel_2 \Rightarrow \mathrm{RW}(rel_1) \leq \mathrm{RW}(rel_2)$.

the recall base is $R = 10$. Let $\mathbf{i}$ be $[\mathtt{hr}, \mathtt{hr}, \mathtt{hr}, \mathtt{fr}, \mathtt{fr}, \mathtt{fr}, \mathtt{pr}, \mathtt{pr}, \mathtt{pr}, \mathtt{pr}, \mathtt{nr}, \ldots, \mathtt{nr}]$. The worst-case vector $\mathbf{w}$ is $[\mathtt{nr}, \ldots, \mathtt{nr}, \mathtt{pr}, \mathtt{pr}, \mathtt{pr}, \mathtt{pr}, \mathtt{fr}, \mathtt{fr}, \mathtt{fr}, \mathtt{hr}, \mathtt{hr}, \mathtt{hr}]$. Then, let us consider two systems A and B such that:

$$\mathbf{r_A} = [\mathtt{hr}, \mathtt{hr}, \mathtt{fr}, \mathtt{nr}, \mathtt{pr}, \mathtt{fr}, \mathtt{nr}, \mathtt{nr}, \mathtt{nr}, \mathtt{pr}, \mathtt{hr}, \mathtt{nr}, \ldots, \mathtt{nr}]$$

$$\mathbf{r_B} = [\mathtt{hr}, \mathtt{hr}, \mathtt{pr}, \mathtt{nr}, \mathtt{fr}, \mathtt{pr}, \mathtt{nr}, \mathtt{nr}, \mathtt{fr}, \mathtt{pr}, \mathtt{fr}, \mathtt{nr}, \mathtt{hr}, \mathtt{pr}, \mathtt{nr}, \ldots, \mathtt{nr}]$$

The recall of system A is $\frac{7}{10}$, whereas the recall of system B is 1.

From function GT we can point out a set called *relevance support* defined as:

$$RS(\mathbf{v}, rel) = \{j \in [1, N] \mid \mathrm{GT}(\mathbf{v}, j) = rel\} \tag{2}$$

which, given a vector $\mathbf{v} \in D^N$ – it can be a run vector $\mathbf{r}$, the ideal vector $\mathbf{i}$, or the worst-case vector $\mathbf{w}$ – and a relevance degree rel, contains the indexes j of the documents of $\mathbf{v}$ with which the given relevance degree (rel) relevance is associated. For instance, in the presented example we have $RS(\mathbf{r_A}, \mathtt{hr}) = \{1, 2, 11\}$ and $RS(\mathbf{r_B}, \mathtt{hr}) = \{1, 2, 3\}$.

Given the ideal vector $\mathbf{i}$ and a relevance degree rel, we can define the *minimum rank* in $\mathbf{i}$ as the first position in which we find a document with relevance degree equal to rel. In the same way, we can define the *maximum rank* in $\mathbf{i}$ as the last position in which we find a document with relevance degree equal to rel. In formulas, they become:

$$\begin{aligned} \min_{\mathbf{i}}(rel) &= \min\big(RS(\mathbf{i}, rel)\big) \\ \max_{\mathbf{i}}(rel) &= \max\big(RS(\mathbf{i}, rel)\big) \end{aligned} \tag{3}$$

In the context of our example, we can say that : $\min_{\mathbf{i}}(\mathtt{hr}) = 1$, $\max_{\mathbf{i}}(\mathtt{hr}) = 3$, $\min_{\mathbf{i}}(\mathtt{fr}) = 4$, $\max_{\mathbf{i}}(\mathtt{fr}) = 6$, $\min_{\mathbf{i}}(\mathtt{pr}) = 7$, $\max_{\mathbf{i}}(\mathtt{pr}) = 10$, $\min_{\mathbf{i}}(\mathtt{nr}) = 11$, and $\max_{\mathbf{i}}(\mathtt{nr}) = 20$.

Given a vector $\mathbf{v}$ and a document at position $j \in [1, N]$, we can define the *Relative Position (RP)* as:

$$\mathrm{RP}(\mathbf{v}, j) = \begin{cases} 0 & \text{if } \min_{\mathbf{i}}\big(\mathrm{GT}(\mathbf{v}, j)\big) \leq j \leq \max_{\mathbf{i}}\big(\mathrm{GT}(\mathbf{v}, j)\big) \\ j - \min_{\mathbf{i}}(\mathrm{GT}\big(\mathbf{v}, j)\big) & \text{if } j < \min_{\mathbf{i}}\big(\mathrm{GT}(\mathbf{v}, j)\big) \\ j - \max_{\mathbf{i}}(\mathrm{GT}\big(\mathbf{v}, j)\big) & \text{if } j > \max_{\mathbf{i}}\big(\mathrm{GT}(\mathbf{v}, j)\big) \end{cases} \tag{4}$$

RP allows for pointing out misplaced documents and understanding how much they are misplaced with respect to the ideal case $\mathbf{i}$. Zero values denote documents which are within the ideal interval, positive values denote documents which are ranked below their ideal interval, and negative values denote documents which are above their ideal interval. Note that the greater the absolute value of $\mathrm{RP}(\mathbf{v}, j)$ is, the bigger is the distance of the document at position j from its ideal interval. From equation 4, it follows that $\mathrm{RP}(\mathbf{i}, j) = 0$, $\forall j \in [1, N]$.

In our example we can determine the following RP vectors:

$$\mathrm{RP}(\mathbf{r_A}) = [0, 0, -1, -7, -2, 0, -4, -3, -2, 0, +8, 0, \ldots, 0]$$

$$\mathrm{RP}(\mathbf{r_B}) = [0, 0, -4, -7, 0, -1, -4, -3, +3, 0, +5, 0, +10, +4, 0, \ldots, 0]$$

Given a vector $\mathbf{v}$ and a document at position $j \in [1, N]$, we can define the *Cumulated Relative Position (CRP)* as:

$$\mathrm{CRP}(\mathbf{v}, j) = \sum_{k=1}^{j} \mathrm{RP}(\mathbf{v}, k) \tag{5}$$

For each position j, CRP sums the values of RP up to position j included. From equation 5, it follows that $\mathrm{CRP}(\mathbf{i}, j) = 0, \forall j \in [1, N]$. In our example, $\mathrm{CRP}(\mathbf{r_A}, 20) = -11$ and $\mathrm{CRP}(\mathbf{r_B}, 20) = +3$.

2.2 Properties of the Metric

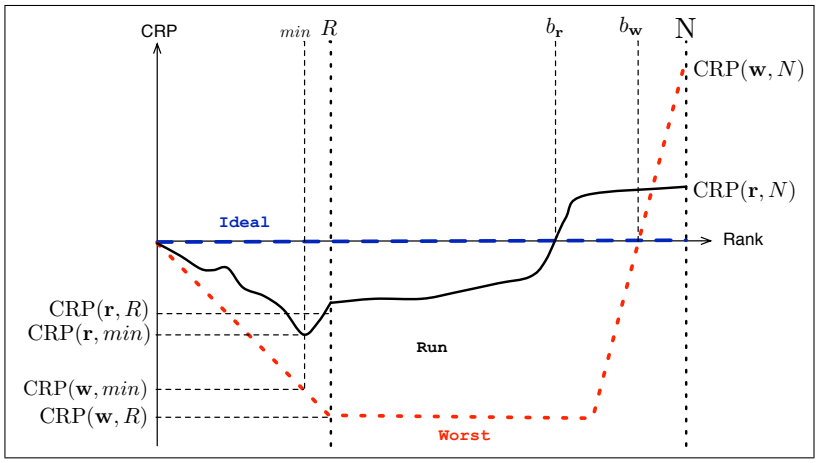

Fig. 1. Cumulative Relative Position sketch for a topic of a given run: min is the rank of the turn-around point of the run, R is the rank of the recall base, b_r is the rank of the balance point of the run, b_w is the rank of the balance point of the worst-case, N indicates the number of retrieved documents, $\mathrm{CRP}(\mathbf{r}, R)$ is the loss value of the run at R, $\mathrm{CRP}(\mathbf{r}, min)$ is the minimum CRP value of the run, $CRP(\mathbf{w}, min)$ is the CRP value of the worst-case at the minimum CRP of the run, $CRP(\mathbf{w}, R)$ is the loss value of the worst-case at R, $\mathrm{CRP}(\mathbf{w}, N)$ is the maximum CRP value of the worst-case, and $\mathrm{CRP}(\mathbf{r}, N)$ is the maximum CRP value of the run.

We can point out the following properties for CRP:

- CRP can only be zero or negative before reaching the rank of the recall base (R);
- the faster the CRP curve goes down before R, the worse the run is;
- after R the CRP curve is non-decreasing;

– after that the last relevant document has been encountered, CRP remains constant;
– the sooner we reach the x-axis (balance point: $b_{\mathbf{r}}$), the better the run is.

In Figure 1 we can see a sketch of the CRP for a topic of a run. For a given topic there are two fixed values which are the rank of recall base (R) and the number of retrieved documents (N); this allows us to compare systems on the R basis. There are significant points both on the y-axis and in the x-axes. Given the run $\mathbf{r}$, in the y-axes we point out three values:

1. CRP($\mathbf{r}, R$): the **loss value** of the run measured at R;
2. CRP($\mathbf{r}, min$): the minimum CRP of the run;
3. CRP($\mathbf{r}, N$): the CRP value at N, it is the maximum value of CRP for the run.

In the x-axes we point out two points:

1. min: the **turn-around** point of the CRP curve, which is the most relevant point of inflection of the curve[5];
2. $b_{\mathbf{r}}$: the **balance point** of the curve. It indicates the point where CRP has re-gained the value lost from 1 to min.

We can define four synthesis indicators describing the CRP curve of a topic for a given run. These indicators characterize the CRP curve and allow us to understand the behaviour of the system under examination for a given topic. Furthermore, these indicators are exploited to produce alternative visualizations of CRP; we can exploit them to read the CRP along different dimensions, each one representing a different aspect of the measurement.

The first indicator is the *recovery value* (ρ) defined as the ratio between R and $b_{\mathbf{r}}$:

$$\rho = \frac{R}{b_{\mathbf{r}}} \tag{6}$$

The recovery-value is always between 0 and 1 ($0 < \rho \leq 1$) where $\rho = 1$ indicates a perfect ranking and $\rho \to 0$ a progressively worse ranking. Please note that $\rho \to 0$ when $b_{\mathbf{r}} \to \infty$.

The second indicator is the *balance ratio* (b_{ratio}) defined as *one* minus the ratio between $b_{\mathbf{r}}$ (i.e. balance point of the run) and $b_{\mathbf{w}}$ (i.e. balance point of the worst-case):

$$b_{ratio} = 1 - \frac{b_{\mathbf{r}}}{b_{\mathbf{w}}} \tag{7}$$

The balance ratio is always between 0 and 1 ($0 \leq b_{ratio} < 1$) where $b_{ratio} = 0$ indicates the worst possible ranking because $b_{\mathrm{r}} = b_{\mathrm{w}}$ and $b_{ratio} \to 1$ a progressively better ranking. Basically, the b_{ratio} points out the correlation between the ranking of the run and the worst-case ranking.

[5] An inflection point is a point on a curve at which the sign of the curvature (i.e., the concavity) changes.

The third indicator is the minimum CRP value ratio (CRP_{min}) defined as *one* minus the ratio between the minimum CRP value of the run and the CRP value of the worst-case calculated in correspondence with the minimum CRP value of the run (please see Figure 1).

$$\mathrm{CRP}_{min} = 1 - \frac{\mathrm{CRP}(\mathbf{r}, min)}{\mathrm{CRP}(\mathbf{w}, min)} \tag{8}$$

The minimum CRP value ratio is always between 0 and 1 ($0 \leq \mathrm{CRP}_{min} \leq 1$) where $\mathrm{CRP}_{min} = 1$ indicates a perfect ranking because it means that $R = min$ and that $\mathrm{CRP}(\mathbf{r}, min) = 0 = \mathrm{CRP}(\mathbf{i}, min)$; on the other hand, $\mathrm{CRP}_{min} = 0$ indicates the worst possible ranking because it means that $\mathrm{CRP}(\mathbf{r}, min) = \mathrm{CRP}(\mathbf{w}, min)$.

The fourth indicator is the CRP value ratio at N (CRP_N) defined as *one* minus the ratio between the CRP value at N of the run and the CRP value of the worst-case at N (please see Figure 1).

$$\mathrm{CRP}_N = 1 - \frac{\mathrm{CRP}(\mathbf{r}, N)}{\mathrm{CRP}(\mathbf{w}, N)} \tag{9}$$

The CRP value ratio at N is always between 0 and 1 ($0 \leq \mathrm{CRP}_N \leq 1$) where $\mathrm{CRP}_N = 1$ indicates a good ranking because it means that $\mathrm{CRP}(\mathbf{r}, N) = 0$, and $\mathrm{CRP}_N = 0$ indicates the worst possible ranking because it means that $\mathrm{CRP}(\mathbf{r}, N) = \mathrm{CRP}(\mathbf{w}, N)$.

We consider an IR system which produces a set of runs defined as $RUN = \{\mathbf{r_1}, \mathbf{r_2}, \ldots, \mathbf{r_T}\}$ where $T \in \mathbb{N}$ is the number of considered topics; every topic has its own recall base R, so for topic t_1 there is a recall base R_1, for topic t_2 there is a recall base R_2 and so on and so forth until topic t_T with recall base R_T. Now, we can define the average recovery-value (ρ_{avg}) as:

$$\rho_{avg} = \frac{1}{T} \sum_{t=1}^{T} \frac{R_t}{b_{\mathbf{r_t}}} \tag{10}$$

The closer ρ_{avg} is to one, the better the system under examination behaves.

3 Comparison with Previous Metrics

The novel CRP metric has several advantages when compared with several previous and related measures. The *Normalized Recall (NR)* metric [8], the *Sliding Ratio (SR)* metric [7], and the *Satisfaction–Frustration–Total (SFT)* metric [7] all seek to take into account the order in which documents are presented to the user. The NR metric compares the actual performance of an IR technique to the ideal one (when all relevant documents are retrieved first). Basically, it measures the area between the ideal and the actual curves. NR does not take the degree of document relevance into account and is highly sensitive to the last relevant document found late in the ranked order.

The SR metric takes the degree of document relevance into account and actually computes the cumulated gain and normalizes this by the ideal cumulated gain for the same retrieval result. The result thus is quite similar to the *Normalized Cumulated Gain ((n)CG)* metric (see below). SR is dependent on the retrieved list size: with a longer list the ideal ranking may change essentially and this affects all values of the metric from rank one onwards. Improving on normalized recall, SR is not dependent on outliers, but it is sensitive to the actual retrieved set size.

The SFT metric consists of three components similar to the SR measure. The satisfaction metric only considers the retrieved relevant documents, the frustration metric only the irrelevant documents, and the total metric is a weighted combination of the two. Like SR, also SFT assumes equally long lists of retrieved documents, which are obtained in different orders by the IR techniques to be compared. This is a critical assumption for comparison since for any retrieved list size n, when $n << N$ (the database size), different IR techniques may retrieve quite different documents. A strong feature of SFT comes from its capability of penalizing an IR technique for retrieving irrelevant documents while rewarding for the relevant ones. CRP allows for comparison of equally long list of retrieved documents (e.g. n) by exploiting the $CRP(\mathbf{v}, n)$ value; but, at the same time it allows for comparisons based on the recall base (i.e. the recovery value) which are – to a reasonable degree – independent by the retrieved list size.

The cumulated gain–based metrics, the CG, DCG, (n)CG and (n)DCG [5], give at any rank examined, an estimate of the (normalized, discounted) cumulated gain as a single figure no matter what the recall base size is. They 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 weight down the gain received through documents found later in the ranked results. However, the gain values grow monotonically unless negative gain values [6] are used. The metrics do not explicitly handle ranking suboptimal documents early – this only shows lower gain values. Like the CRP, the normalized versions compare the ranking quality to each topics entire recall base (qrel) 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 ideal ranking), ranking is suboptimal and less relevant documents are retrieved earlier than they should. The CG-based metrics do not directly address ranking optimality but cumulate gain and loss (or negative gain). Both metrics address the bi-directional nature of searching which may be seen as a process alternating between success and failure.

In traditional test collection-based evaluation, the evaluation task is simplified by abstracting away users, their situations and tasks [9], and relevance is assumed as topical, stable and binary. This neglects user experiences in real life with dynamic, multiple-dimension and multi-graded relevance [10,11] and user

experiences caused by browsing sequences of non-relevant or suboptimal documents. Keskustalo and colleagues [6] analyze negative aspects in higher-order (above topical) relevance e.g. due to suboptimal ranking:

- Cognitive relevance: Not receiving pertinent information;
- situational relevance: Time pressure, effort;
- motivational/Affective relevance: Frustration, lack of accomplishment.

Both the CRP and the CG-based metrics, in particular their visualizations, facilitate the identification and analysis of these higher-order aspects of relevance. The dips in both kinds of graphs make this explicit in retrieval context (see Figure 2).

4 Experiments and Visualization

For the experimental analysis we adopted a test collection based on data from the TREC7 Ad-hoc test collection. A subset of all the *topics* 351-400 is considered, specifically those re-assessed in [5]. Indeed, the *relevance judgments* adopted are those obtained by the evaluation activity carried out in that paper. All the relevant documents of 20 TREC7 topics and 18 TREC8 topics were re-assessed together with 5% of documents judged as not relevant, where assessment was performed using a four graded relevance scale; details on the re-assessment procedure can be found in [12]. We developed a visual analytics prototype to visualize and interact with the various metrics adopted. In particular, we build on a first version of this prototype described in [13] to add the CRP visualizations. Figure 2 shows a screen shot of the running prototype comparing the CRP curve with the DCG curve (with negative weights).

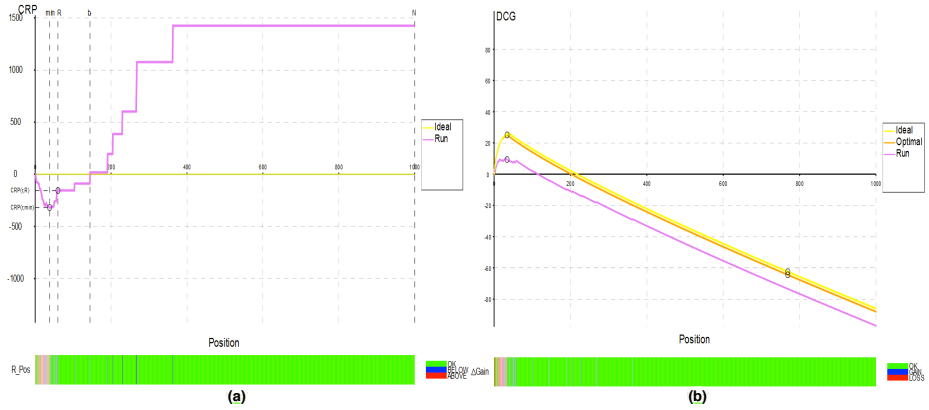

Fig. 2. Comparison between (a) CRP and (b) DCG with negative weights (i.e. $-1, 1, 2, 3$), on topic 351 of the "bb1" run.

For what it is concerned with CRP, the visualization prototype focuses on three types of visualization: (1) the "CRP Graph" which shows the trend of the CRP curve calculated on a specific topic for a given run (Figure 2a); (2) the "CRP Aggregate Graph" which shows, in a Parallel Coordinates fashion (a visualization technique well-suited to give an insight on the correlation of various measures on a big collection of data), an aggregate view of all the topics for the given run, ordered by their recall base rank (R) (Figure 3); and (3) the "CRP vs DCG Graph" which eases the comparison of the CRP curve and the DCG curve (Figure 2).

For this analysis we consider the run named "bbn1" submitted to the TREC7 Ad-Hoc Track [14]. Figure 2a shows the CRP curve for topic 351 of "bb1". Both the ideal and run curve are reported; please note that the ideal curve coincides with the x-axis. In the bottom part of the graph an horizontal bar shows the RP values; this bar helps the analyst to understand the single contribution of each document to the CRP. The points corresponding to the synthesis indicators are reported to highlight the trend of the run and to pinpoint the areas in which the trend changes.

The visualization reported in Figure 2 allows the comparison, for the same topic (i.e. topic 351), of CRP and DCG curves. In order to facilitate the comprehension of the graph the bars showing the single contribution to the "cumulated" value of each document are also reported; the green color means exact positioning, the red color a position above the ideal, and the blue color a position below the ideal. This convention is also valid for DCG and its horizontal bar (i.e. the so-called ΔGain bar [13]): we use the color green for no gain, the blue color when a document is ranked below the ideal (we have a loss) and the red color when a document is ranked above the ideal (we have a gain). In this figure we can see that the distance between R and b_r gives us a visual measure of how much misplaced documents influence the initial part of the ranking list; the more relevant documents the system puts in high positions, the shorter the distance between R and b_r is. This fact is quantified by the ρ value; indeed, a high ρ value reflects a high number of relevant documents ranked in the expected position and a short distance between R and b_r. With respect to DCG, CRP allows for explicit considerations on the information value of late-ranked documents. After the b_r value, the CRP graph allows us to see in which positions misplaced documents are put and to which degree they contribute to the overall quality of the ranking. CRP increases by a step for every late-ranked document and the height of this step is proportional to the relevance of the document and to the position in which it lies.

Figure 3 shows the CRP Aggregate Graph for all the TREC 7 topics of the "bb1" run, by means of the Parallel Coordinates paradigm. It visualizes the four synthesis indicators of the CRP curve plus two parameters which characterize the topic under investigation: R which is the recall base and $\rho_w = \frac{R}{b_w}$ which is the recovery value for the worst-case.

All the values are in the $[0, 1]$ interval and a different color set is used to distinguish in a better way between the curves. This visualization allows us to

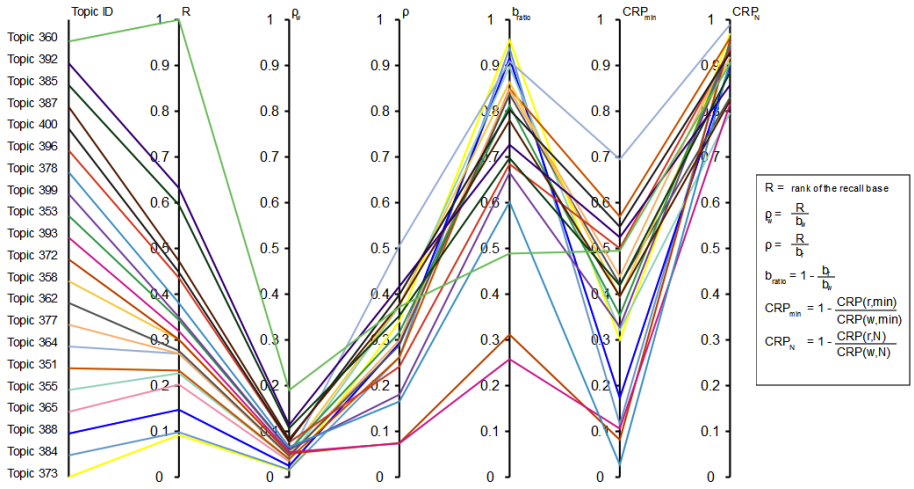

Fig. 3. CRP Parallel Coordinates Graph for the considered run (bb1 of TREC7).

see the correlation between the topics. We can point out the topic in which the run performs poorly (e.g. topic 393) and the topic for which it works fine (e.g. topic 394). The topics are ordered by their recall base in order to present a better overall visualization of the results.

5 Conclusions

In the present paper, we have proposed a new evaluation metric for Information Retrieval, the *Cumulated Relative Position (CRP)*. We started with the observation that a document of a given degree of relevance may be ranked too early or too late regarding the ideal ranking of documents for a query. Its relative position may therefore be negative, indicating too early ranking, zero indicating correct ranking, or positive, indicating too late ranking. By cumulating these relative rankings we indicate the net effect of document displacements, the CRP. We defined the CRP, and discussed its properties, formally. We also presented visualizations of the CRP that help analyze individual query performance, aggregate query performance, and compare the CRP performance with other IR metrics such as the DCG.

The CRP metric differs from prior standard IR metrics in explicitly handling document ranking misplacements either too early or too late given their degree of relevance and the ideal ranking. We believe that the CRP offers several advantages in IR evaluation because (a) it is obvious to interpret and it gives an estimate of ranking performance as a single measure relative to the ideal ranking for the topic; (b) it is independent on outliers since it focuses on the ranking of the result list; (c) it directly reports the effort wasted in examining suboptimal

rankings; (d) it is based on graded relevance; (e) it can easily be summarized by four synthesis indicators; (f) it works fine with graphical representations.

Good evaluation metrics are required for progress in IR. We believe that the CRP metric is a useful tool in the IR evaluators tool box.

Acknowledgements The work reported in this paper has been supported by the PROMISE network of excellence (contract n. 258191) project as a part of the 7th Framework Program of the European commission (FP7/2007-2013).

References

1. Sanderson, M.: Test Collection Based Evaluation of Information Retrieval Systems. Foundations and Trends in Information Retrieval (FnTIR) **4** (2010) 247–375
2. Harman, D.K.: Information Retrieval Evaluation. Morgan & Claypool Publishers, USA (2011)
3. Kekäläinen, J., Järvelin, K.: Using Graded Relevance Assessments in IR Evaluation. Journal of the American Society for Information Science and Technology (JASIST) **53** (2002) 1120—1129
4. Robertson, S.E., Kanoulas, E., Yilmaz, E.: Extending Average Precision to Graded Relevance Judgments. In: Proc. 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2010), ACM Press, New York, USA (2010) 603–610
5. Järvelin, K., Kekäläinen, J.: Cumulated Gain-Based Evaluation of IR Techniques. ACM Transactions on Information Systems (TOIS) **20** (2002) 422–446
6. Keskustalo, H., Järvelin, K., Pirkola, A., Kekäläinen, J.: Intuition-Supporting Visualization of User's Performance Based on Explicit Negative Higher-Order Relevance. In: Proc. 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008), ACM Press, New York, USA (2008) 675–681
7. Korfhage, R.R.: Information Storage and Retrieval. Wiley Computer Publishing, John Wiley & Sons, Inc., USA (1997)
8. Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval. McGraw-Hill, New York, USA (1983)
9. Voorhees, E.M.: TREC: Continuing Information Retrieval's Tradition of Experimentation. Communications of the ACM (CACM) **50** (2007) 51–54
10. Cosijn, E., Ingwersen, P.: Dimensions of Relevance. Information Processing & Management **36** (2000) 533–550
11. Saracevic, T.: Relevance reconsidered. In Ingwersen, P., Pors, N.O., eds.: Proc. 2nd International Conference on Conceptions of Library and Information Science – Integration in Perspective (CoLIS 2), Royal School of Librarianship, Copenaghen, Denmark (1996) 201–218
12. Sormunen, E.: Liberal Relevance Criteria of TREC: Counting on Negligible Documents? In: Proc. of the 25th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press (2002) 324–330
13. Ferro, N., Sabetta, A., Santucci, G., Tino, G.: Visual Comparison of Ranked Result Cumulated Gains. In: Proc. 2nd International Workshop on Visual Analytics (EuroVA 2011), Eurographics Association, Goslar, Germany (2011) 21–24
14. Voorhees, E., Harman, D.: Overview of the Seventh Text REtrieval Conference (TREC-7). In: NIST Special Publication 500-242: The Seventh Text REtrieval Conference (TREC 7), Springer-Verlag, Heidelberg, Germany (1999)