Cumulative Relative Position: A Metric for Ranking Evaluation

Nicola Ferro, *Gianmaria Silvello*
University of Padua, Italy

Kalervo Järvelin, Heikki Keskustalo, Ari Pirkola
University of Tampere, Finland

Marco Angelini, Giuseppe Santucci
"La Sapienza" University of Rome, Italy

CLEF 2012 - Information Access Evaluation meets Multilinguality, Multimodality, and Visual Analytics
Rome, Italy, September 17-20, 2012

Outline

- Motivations
- Overview of CRP
- Properties of CRP
- Synthesis Indicators and Visualizations
- On-Going Work

Motivations

- Design and develop an IR system is challenging and testing it is time consuming
  - Analyze the behavior of the system under different conditions in order to tune or improve the system
  - **Meet user expectations (!)**

- We need **proper evaluation methodologies** to ensure IR systems meet user requirements

- We can do it evaluating the **quality of the output ranked lists**

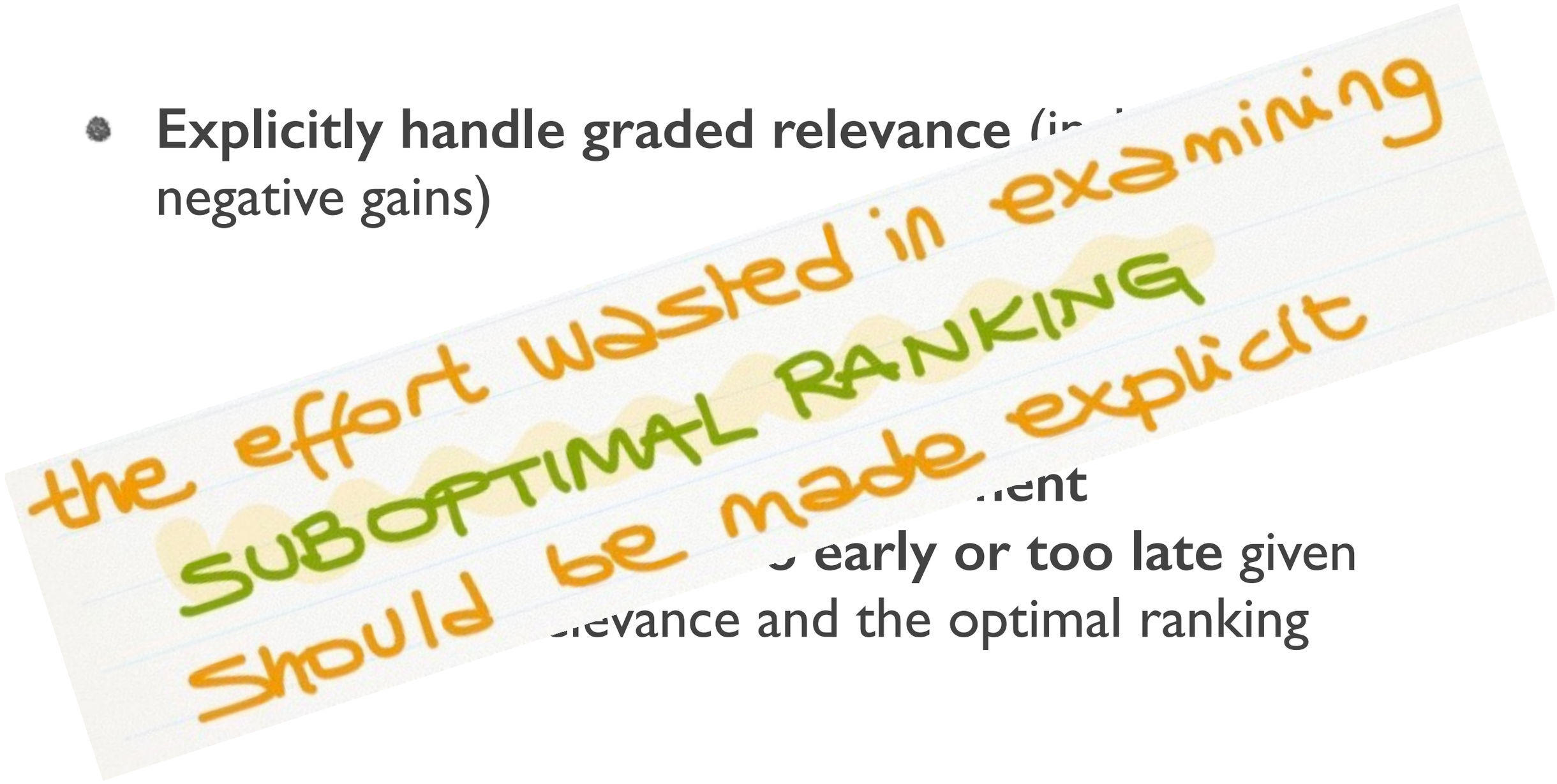
A couple of things that a metric should do...

- **Explicitly handle graded relevance** (including negative gains)

- Explicitly **take into account document misplacements either too early or too late** given their degree of relevance and the optimal ranking

A couple of things that a metric should do...

- **Explicitly handle graded relevance** (in
  negative gains)



ient

early or too late given

levance and the optimal ranking

... but what about the very good metrics we have?

Traditional metrics do not take deviations from optimal ranking sufficiently into account

MAP (extended to graded relevance)

Discounted Cumulative Gain

1) no explicit way for penalizing early-ranked docs

2) penalization (only) for non-relevant documents (DCG with negative gains)

3) they do consider the **severity** of document mis-ranking

Relative Position (RP)

how it works

	Ideal		Run		RP
j = 1	HR	min(HR) = 1	1	HR	ideal	0
2	HR		2	HR	ideal	0
3	HR	max(HR) = 3	3	FR	too early	-1
4	FR	min(FR) = 4	4	NR	too early	-7
5	FR		5	PR	too early	-2
6	FR	max(FR) = 6	6	FR	ideal	0
7	PR	min(PR) = 7	7	NR	too early	-4
8	PR		8	NR	too early	-3
9	PR		9	NR	too early	-2
10	PR	max(PR) = 10	10	PR	ideal	0
11	NR	min(PR) = 11	11	HR	too late	+8
⋮			12	NR		0
20	NR	max(PR) = 20	20	NR		0

Relative Position

Let us determine **how much** a document is misplaced with respect to its ideal rank

Cumulative Relative Position

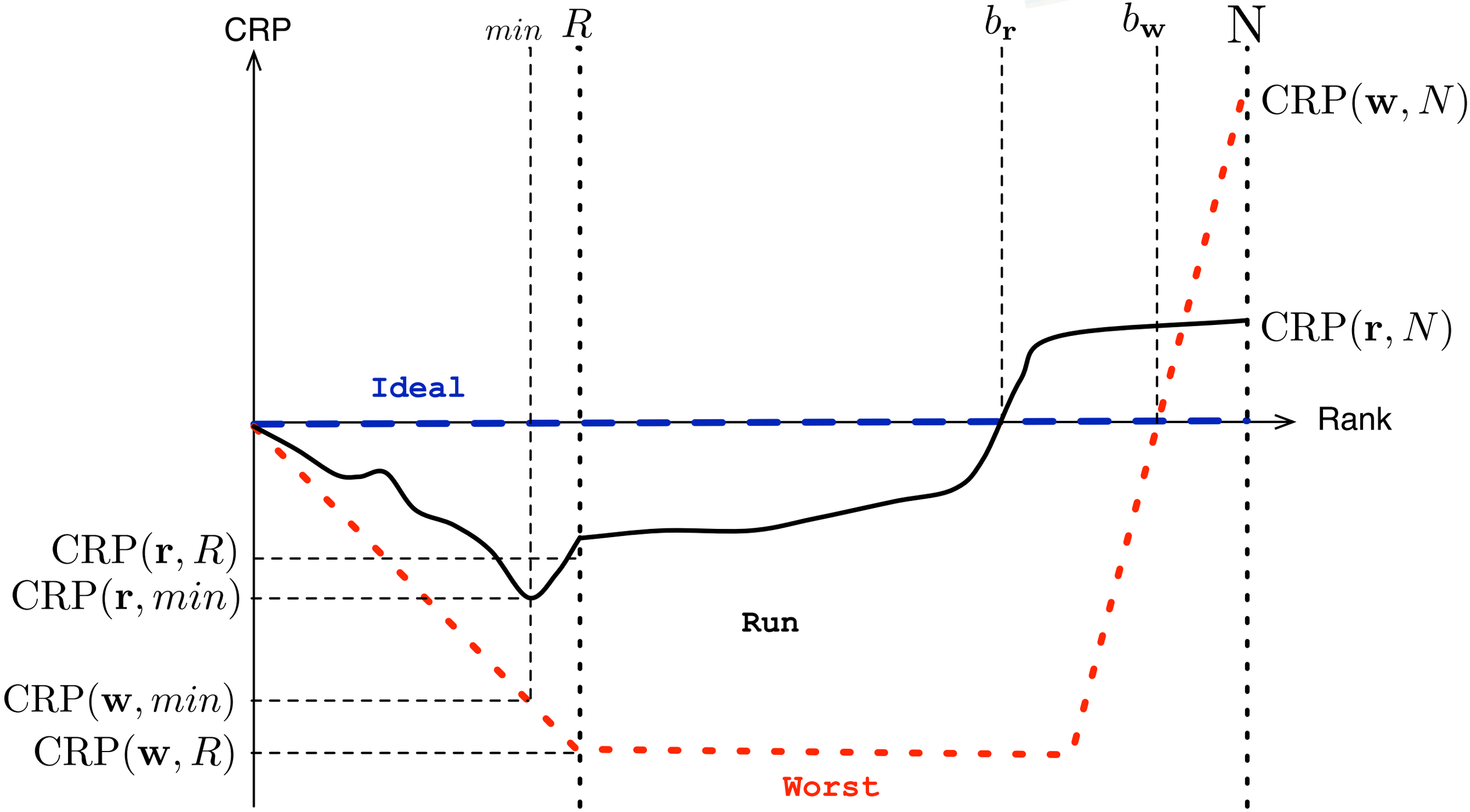
how it works

CRP cumulates the RP values

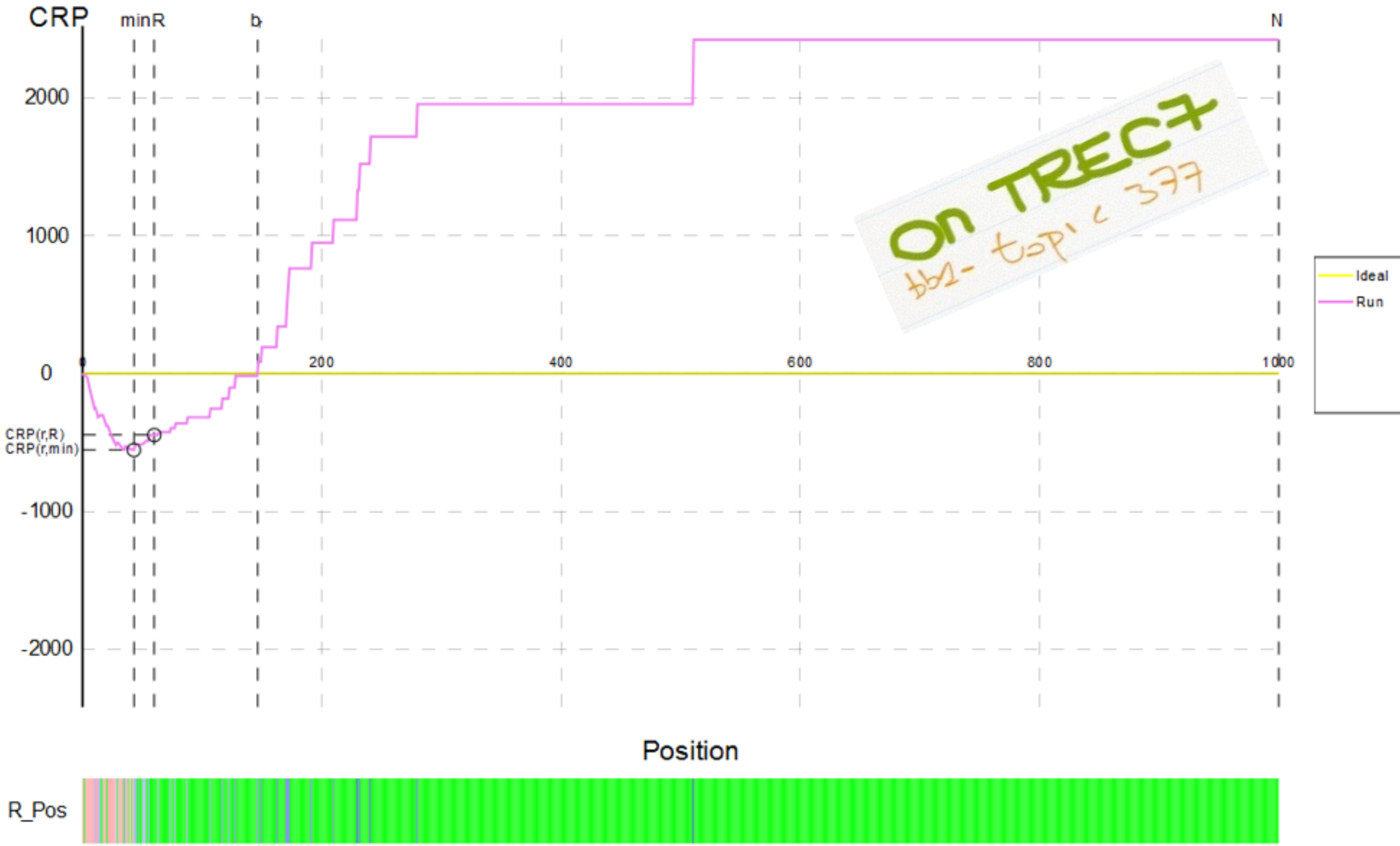
$$\mathrm{CRP}(\mathbf{v}, j) = \sum_{k=1}^{j} \mathrm{RP}(\mathbf{v}, k)$$

Cumulative Relative Position

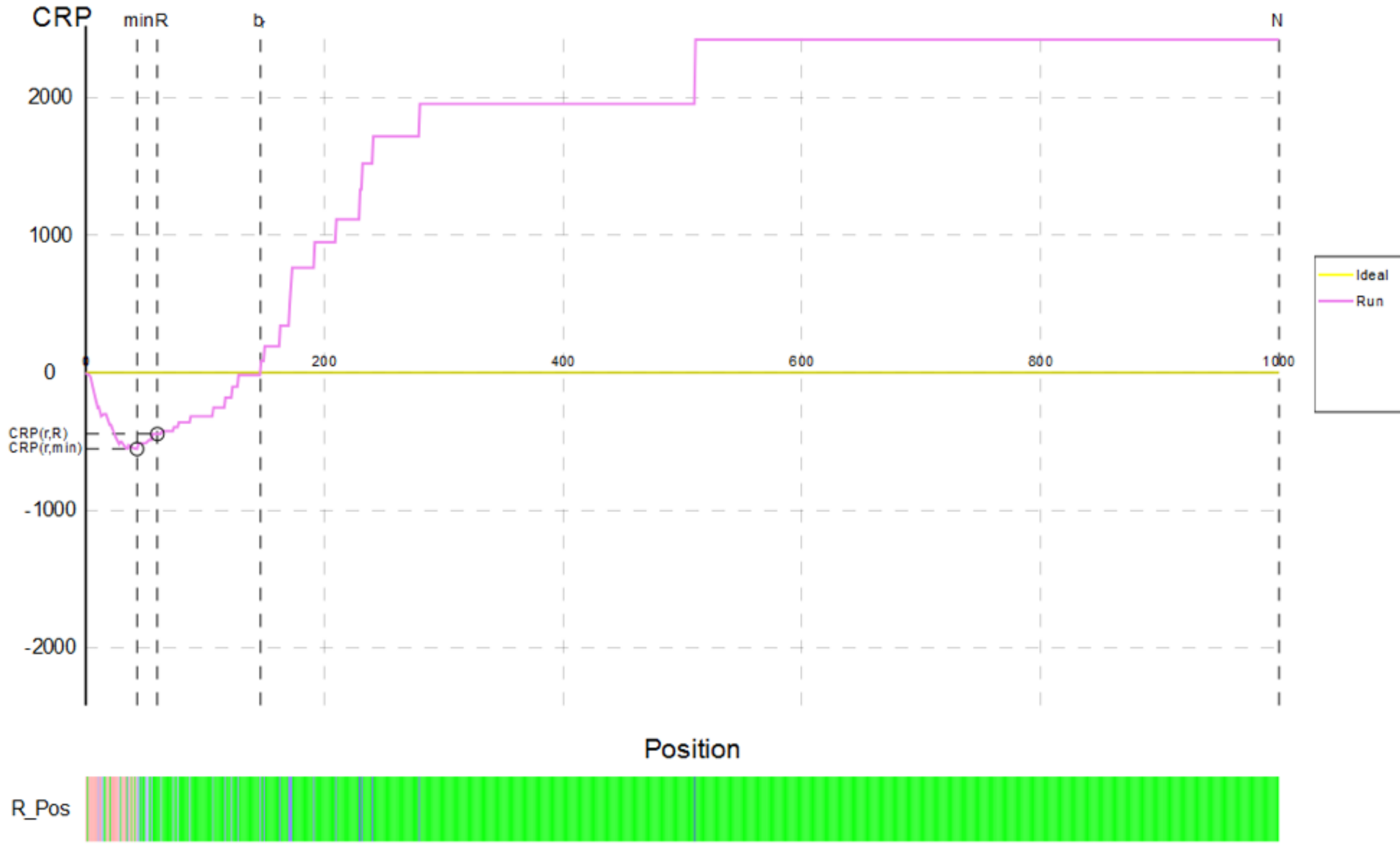
an intuitive view



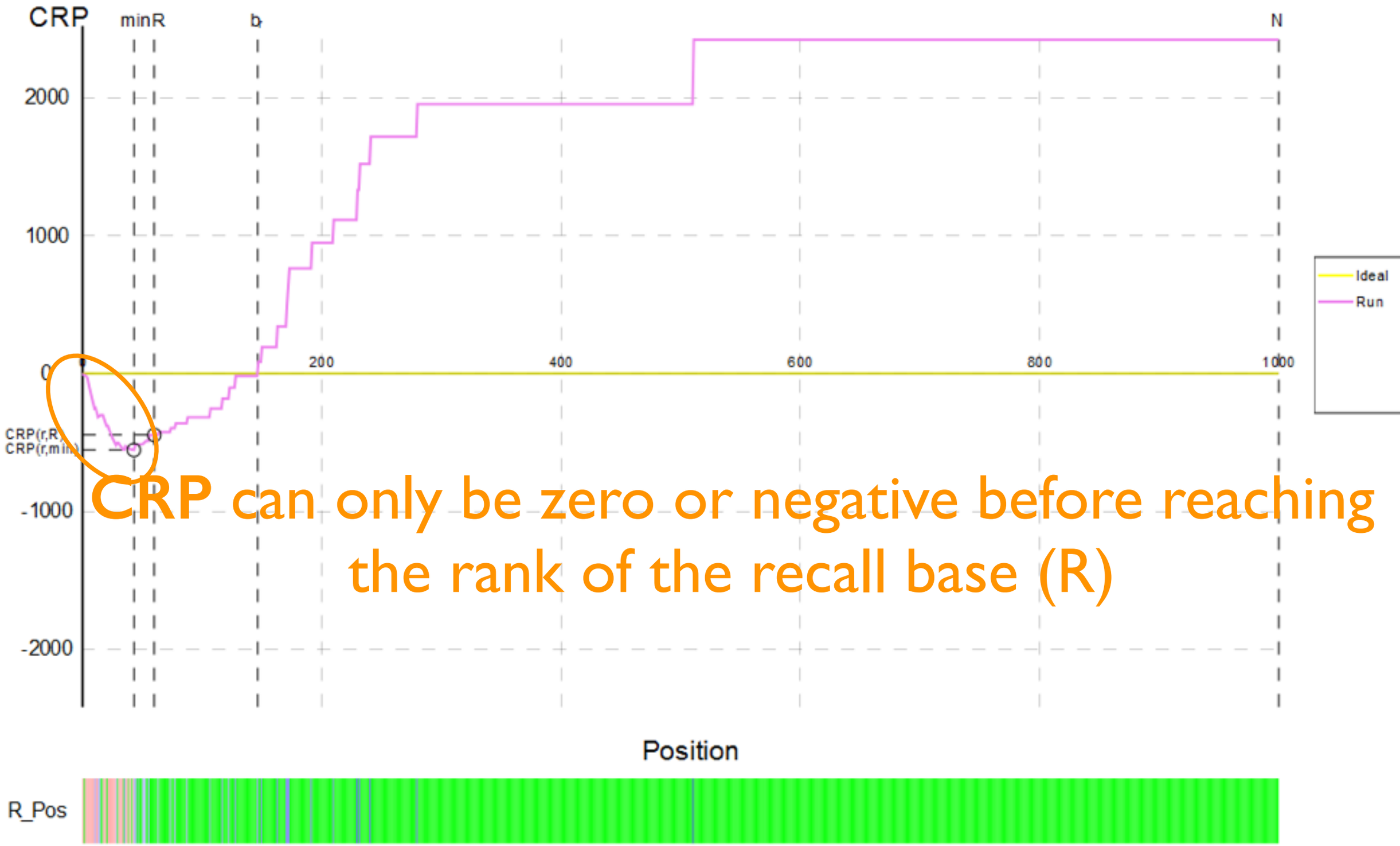
Cumulative Relative Position



CRP Properties

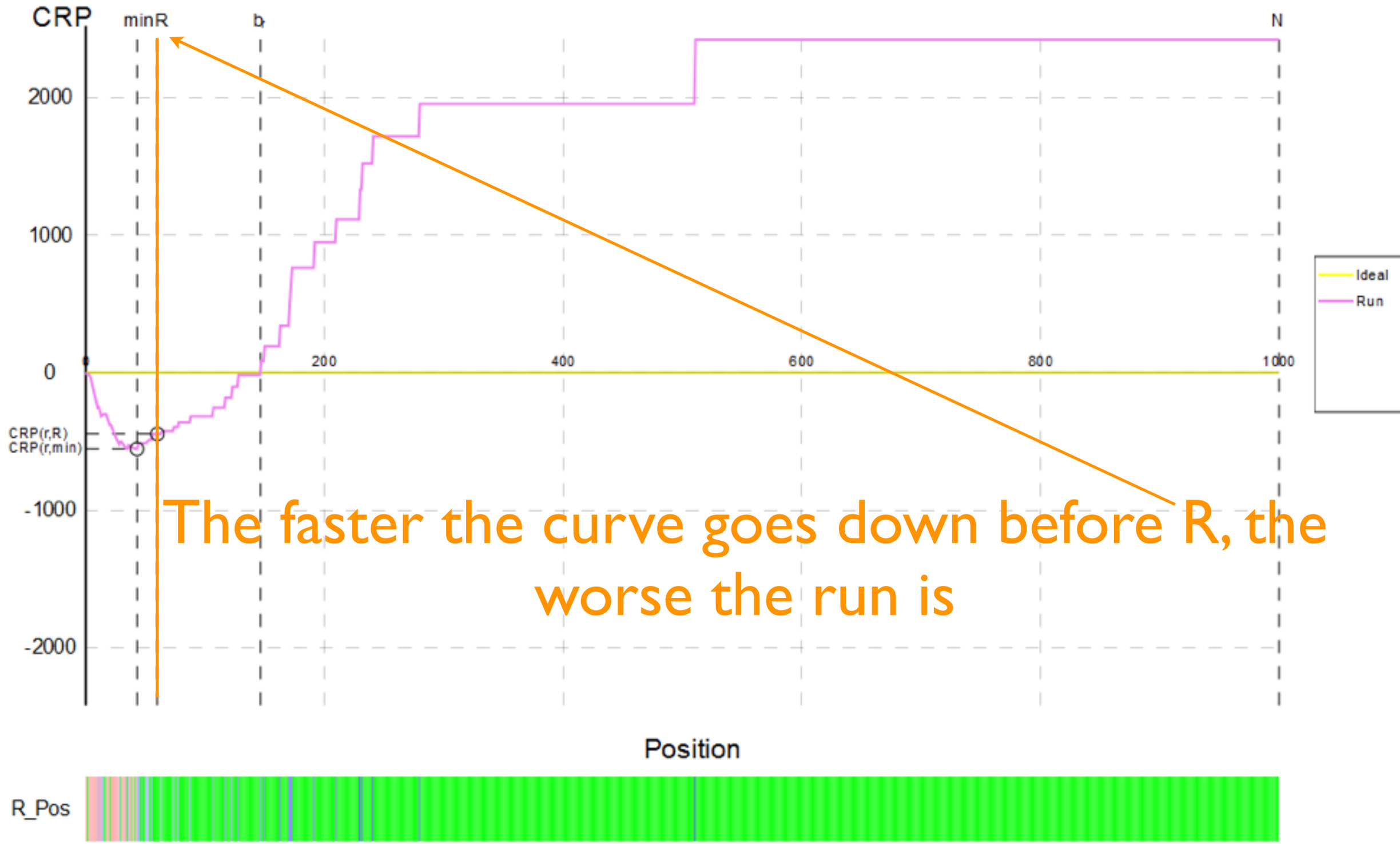


CRP Properties

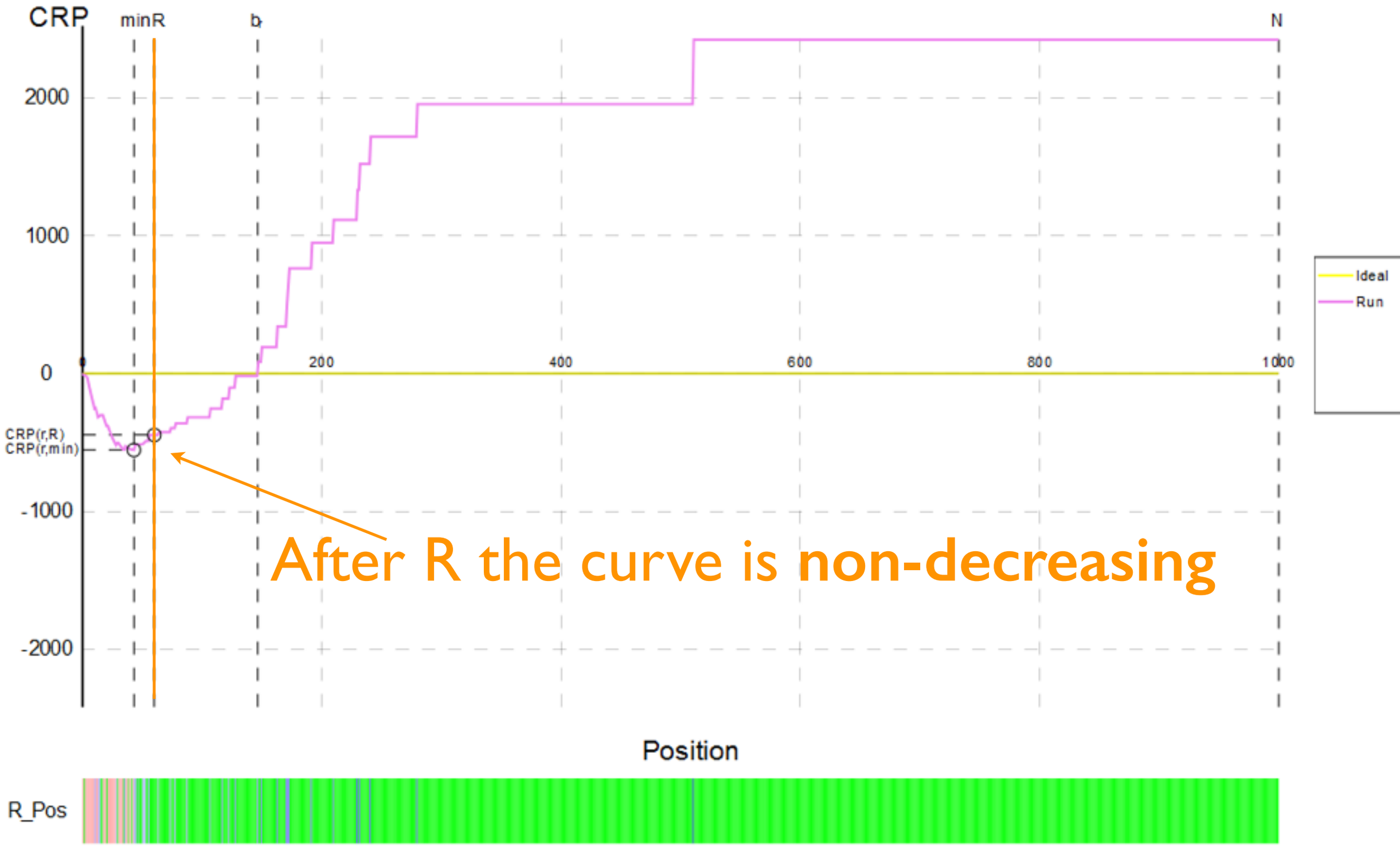


CRP can only be zero or negative before reaching the rank of the recall base (R)

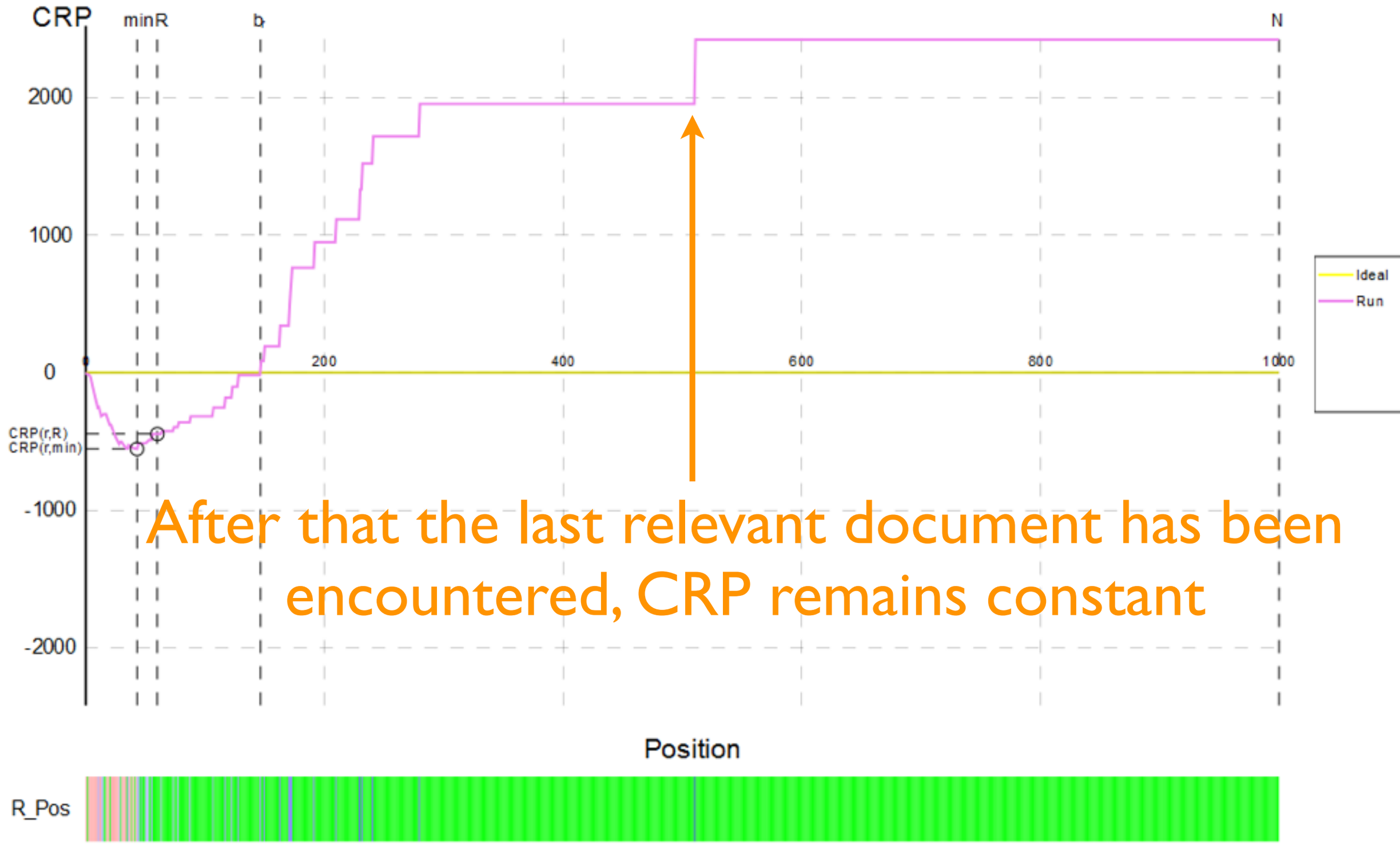
CRP Properties



CRP Properties

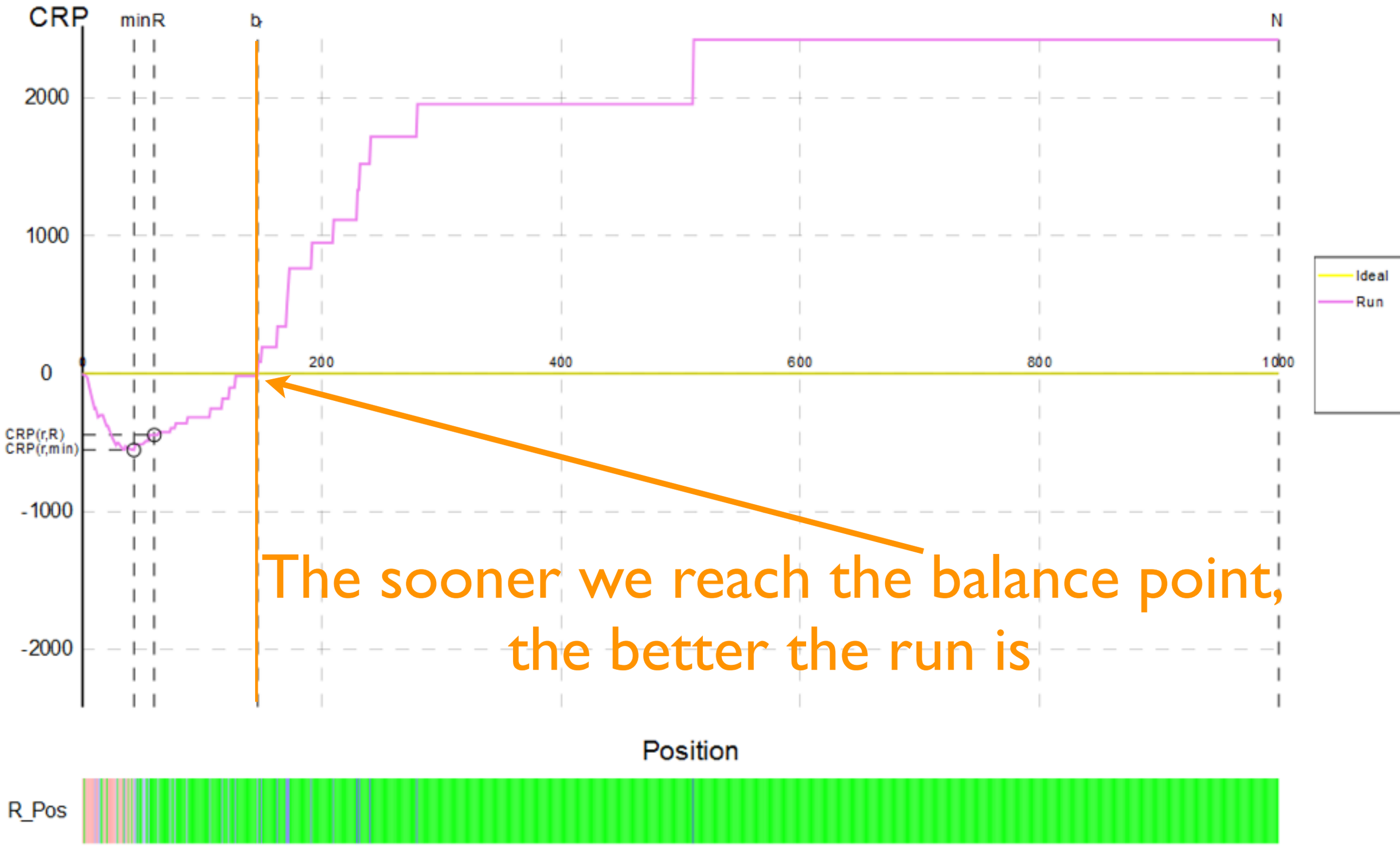


CRP Properties



After that the last relevant document has been encountered, CRP remains constant

CRP Properties



We like CRP because:

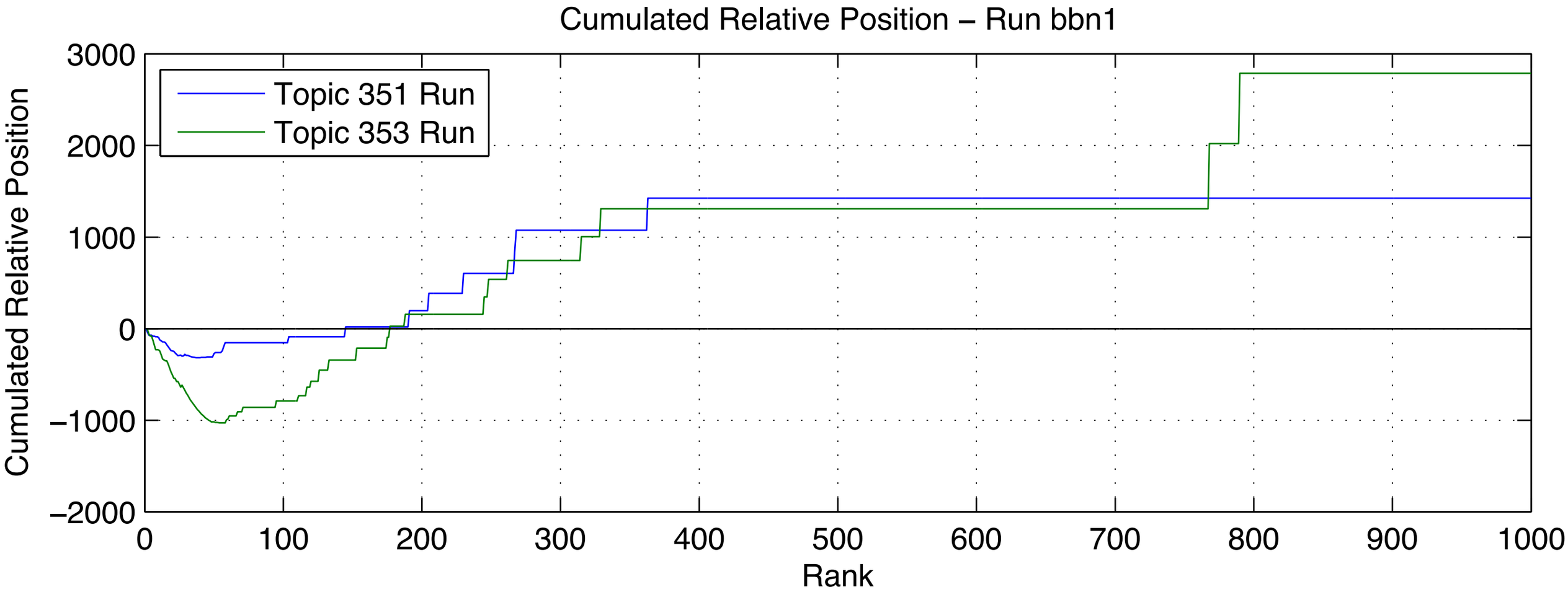
- **At any rank it gives an estimate of ranking performance** as a single measure relative to the ideal ranking

- 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; the effort wasted in examining a suboptimal ranking is made explicit

- It allows the **conflation of relevance grades** of documents and therefore **more or less fine-grained analyses** of the ranking performances of an IR technique may be produced

...and because it's good for comparisons



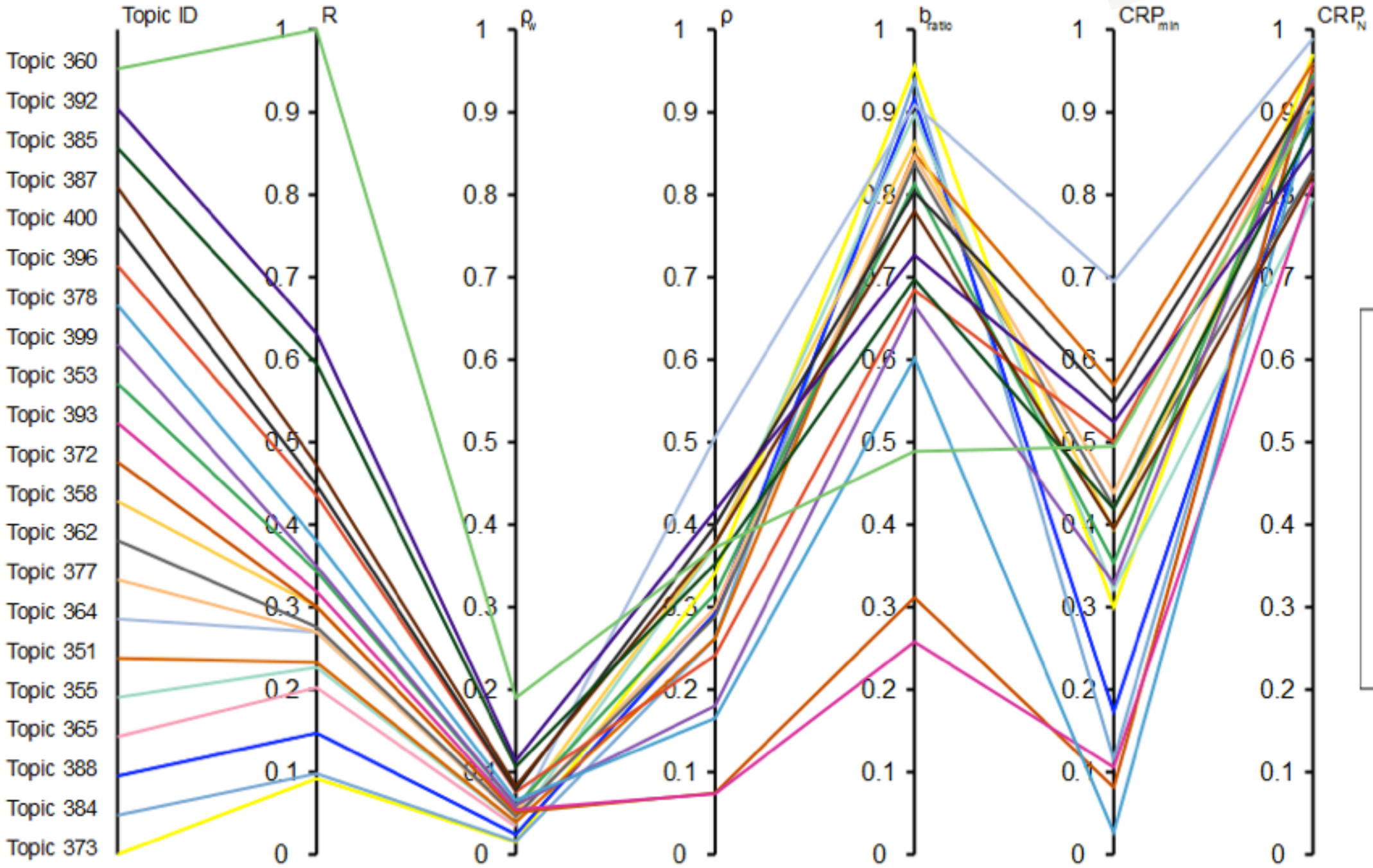
... and we also like it because:

- 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.

CRP synthesis indicators

at a glance



R = rank of the recall base

$\rho_w = \frac{R}{b_w}$

$\rho = \frac{R}{b_r}$

$b_{ratio} = 1 - \frac{b_r}{b_w}$

$CRP_{min} = 1 - \frac{CRP(r,min)}{CRP(w,min)}$

$CRP_N = 1 - \frac{CRP(r,N)}{CRP(w,N)}$

Ongoing Work

- A **Normalized** version of **CRP**
- **Reliability of CRP**: Stability and Sensitivity of the Synthesis Indicators
- **Extensive experimentation** and comparison with other (graded) metrics (e.g. DCG, R-measure, Q-measure) on different test collections (e.g. NTCIR-3 CLIR and TREC2011 Web Track)