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**
- Explicitly take into account document misplacements either too early or too late given their degree of relevance and the optimal ranking
... but what about the very good metrics we have?
... but what about the very good metrics we have?

Traditional metrics do not take deviations from optimal ranking sufficiently into account
... 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**
... 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)
Relative Position (RP)

Ideal

\[ j = 1 \]
- HR

\[ j = 2 \]
- HR

\[ j = 3 \]
- HR

\[ j = 4 \]
- FR

\[ j = 5 \]
- FR

\[ j = 6 \]
- FR

\[ j = 7 \]
- PR

\[ j = 8 \]
- PR

\[ j = 9 \]
- PR

\[ j = 10 \]
- PR

\[ j = 11 \]
- NR

\[ \vdots \]

\[ j = 20 \]
- NR
Relative Position (RP)

Ideal

\[ j = 1 \]

\[ \text{min}(HR) = 1 \]

\[ \text{max}(HR) = 3 \]

\[ \vdots \]

\[ 20 \]

\[ \text{NR} \]
Relative Position (RP)

Ideal

j = 1

HR

min(HR) = 1

max(HR) = 3

min(FR) = 4

max(FR) = 6

venerdì 14 settembre 12
Ideal

\[
\begin{align*}
&j = 1 & \text{min}(HR) = 1 \\
&2 & \max(HR) = 3 \\
&3 & \text{min}(FR) = 4 \\
&4 & \max(FR) = 6 \\
&5 & \text{min}(PR) = 7 \\
&6 & \max(PR) = 10 \\
&7 & \text{min}(PR) = 11 \\
&8 & \max(PR) = 20 \\
&9 & \\
&10 & \\
&11 & \\
&20 &
\end{align*}
\]
### Relative Position (RP)

<table>
<thead>
<tr>
<th>Ideal</th>
<th>Run</th>
</tr>
</thead>
<tbody>
<tr>
<td>j = 1</td>
<td>1</td>
</tr>
<tr>
<td>HR</td>
<td>HR</td>
</tr>
<tr>
<td></td>
<td>min(HR) = 1</td>
</tr>
<tr>
<td>j = 2</td>
<td>2</td>
</tr>
<tr>
<td>HR</td>
<td>HR</td>
</tr>
<tr>
<td>j = 3</td>
<td>3</td>
</tr>
<tr>
<td>HR</td>
<td>FR</td>
</tr>
<tr>
<td>j = 4</td>
<td>4</td>
</tr>
<tr>
<td>FR</td>
<td>NR</td>
</tr>
<tr>
<td>j = 5</td>
<td>5</td>
</tr>
<tr>
<td>FR</td>
<td>PR</td>
</tr>
<tr>
<td>j = 6</td>
<td>6</td>
</tr>
<tr>
<td>FR</td>
<td>FR</td>
</tr>
<tr>
<td>j = 7</td>
<td>7</td>
</tr>
<tr>
<td>PR</td>
<td>NR</td>
</tr>
<tr>
<td>j = 8</td>
<td>8</td>
</tr>
<tr>
<td>PR</td>
<td>NR</td>
</tr>
<tr>
<td>j = 9</td>
<td>9</td>
</tr>
<tr>
<td>PR</td>
<td>NR</td>
</tr>
<tr>
<td>j = 10</td>
<td>10</td>
</tr>
<tr>
<td>PR</td>
<td>PR</td>
</tr>
<tr>
<td>j = 11</td>
<td>11</td>
</tr>
<tr>
<td>NR</td>
<td>HR</td>
</tr>
<tr>
<td>j = 12</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>NR</td>
</tr>
<tr>
<td>j = 13</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>NR</td>
</tr>
<tr>
<td>j = 14</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>NR</td>
</tr>
<tr>
<td>j = 15</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>NR</td>
</tr>
<tr>
<td>j = 16</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>NR</td>
</tr>
<tr>
<td>j = 17</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>NR</td>
</tr>
<tr>
<td>j = 18</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>NR</td>
</tr>
<tr>
<td>j = 19</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>NR</td>
</tr>
<tr>
<td>j = 20</td>
<td>20</td>
</tr>
<tr>
<td>NR</td>
<td>NR</td>
</tr>
</tbody>
</table>

Ideal values:
- min(HR) = 1
- max(HR) = 3
- min(FR) = 4
- max(FR) = 6
- min(PR) = 7
- max(PR) = 10

Run values:
- max(PR) = 20
Relative Position (RP)

<table>
<thead>
<tr>
<th>Ideal</th>
<th>Run</th>
</tr>
</thead>
<tbody>
<tr>
<td>j = 1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>20</td>
<td>20</td>
</tr>
</tbody>
</table>

- HR
- FR
- PR
- NR

Ideal:
- min(HR) = 1
- max(HR) = 3
- min(FR) = 4
- max(FR) = 6
- min(PR) = 7
- max(PR) = 10
- min(PR) = 11

Run:
- ideal
- too early
- ideal
- too early
- too early
- ideal
- too early
- too early
- too early
- ideal
- too late
- NR

venerdì 14 settembre 12
Relative Position (RP)

<table>
<thead>
<tr>
<th>Ideal</th>
<th>Run</th>
</tr>
</thead>
<tbody>
<tr>
<td>HR</td>
<td>HR</td>
</tr>
<tr>
<td>HR</td>
<td>FR</td>
</tr>
<tr>
<td>FR</td>
<td>PR</td>
</tr>
<tr>
<td>PR</td>
<td>NR</td>
</tr>
<tr>
<td>NR</td>
<td>HR</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>min(HR) = 1</td>
<td>ideal</td>
</tr>
<tr>
<td>max(HR) = 3</td>
<td>too early</td>
</tr>
<tr>
<td>min(FR) = 4</td>
<td>too early</td>
</tr>
<tr>
<td>max(FR) = 6</td>
<td>ideal</td>
</tr>
<tr>
<td>min(PR) = 7</td>
<td>too early</td>
</tr>
<tr>
<td>max(PR) = 10</td>
<td>ideal</td>
</tr>
<tr>
<td>min(PR) = 11</td>
<td>too late</td>
</tr>
<tr>
<td>max(PR) = 20</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RP</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
</tr>
<tr>
<td>-1</td>
<td></td>
</tr>
<tr>
<td>-7</td>
<td></td>
</tr>
<tr>
<td>-2</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
</tr>
<tr>
<td>-4</td>
<td></td>
</tr>
<tr>
<td>-3</td>
<td></td>
</tr>
<tr>
<td>-2</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
</tr>
<tr>
<td>+8</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Run: ideal, too early, too late
Relative Position

Let us determine **how much** a document is misplaced with respect to its ideal rank.
CRP cumulates the RP values

\[ \text{Cumulative Relative Position} \]

\[ \text{CRP}(v, j) = \sum_{k=1}^{j} \text{RP}(v, k) \]
Cumulative Relative Position
Cumulative Relative Position

CRP

Ideal

Rank

N

On intuitive view
Cumulative Relative Position

CRP

Ideal

Worst

N

Rank

an intuitive view
Cumulative Relative Position

CRP

min $R$

Ideal

Run

Worst

$N$

$b_r$

$N$

$N$

$N$

Rank
Cumulative Relative Position

\[ \text{CRP}(r, R) \]
\[ \text{CRP}(r, \text{min}) \]
\[ \text{CRP}(w, \text{min}) \]
\[ \text{CRP}(w, R) \]

\[ \text{Ideal} \]

\[ \text{Run} \]

\[ \text{Worst} \]

\[ \text{CRP}(w, N) \]

\[ \text{CRP}(r, N) \]

\[ \text{min} \quad R \quad b_r \quad b_w \quad N \]

\[ \text{R} \]

venerdì 14 settembre 12
CRP Properties

![Graph showing CRP properties with position R_Pos and ideal and run lines.](image)

**CLEF 2012 - Information Access Evaluation meets Multilinguality, Multimodality, and Visual Analytics**
Rome, Italy, September 17-20, 2012
CRP can only be zero or negative before reaching the rank of the recall base (R)
The faster the curve goes down before R, the worse the run is.
After R the curve is non-decreasing
After that the last relevant document has been encountered, CRP remains constant.
The sooner we reach the balance point, the better the run is.
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

\[ R = \text{rank of the recall base} \]
\[ \rho_w = \frac{R}{b_w} \]
\[ \rho = \frac{R}{b} \]
\[ b_{\text{ratio}} = 1 - \frac{b}{b_{\text{ratio}}} \]
\[ \text{CRP}_{\text{min}} = 1 - \frac{\text{CRP}(r,\text{min})}{\text{CRP}(w,\text{min})} \]
\[ \text{CRP}_N = 1 - \frac{\text{CRP}(r,N)}{\text{CRP}(w,N)} \]
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)