Axiometrics: An Axiomatic Approach to Evaluation Metrics

Stefano Mizzaro
Udine University, Italy
mizzaro@uniud.it

Let’s be clear from the start: I. Won’t. Go. Overtime.

Aims – 1

- To present both:
  - basic material (what you find in books)
  - advanced material (recently published, even not yet published)
- Links with Julio, Enrique, Evangelos
  - not (yet!) fully integrated
  - a bit disorganized...

Aims – 2

- Intro to IR Evaluation
- Intro to IR Evaluation Measures / Metrics
- The Link between Measurement Theory and Metrics
  - Intro to Measurement Theory (Scales)
  - Metrics Analysis
  - The Axiometrics Framework (*)

The Monday effect...

You don’t find it funny?!
Outline

- Evaluation [5’]
- Measures / metrics [15’]
- Measurement theory [15’]
- Metrics analysis [5’]
- Axiometrics framework [15’]

Outline

- Evaluation [5’]
- Measures / metrics [15’]
- Measurement theory [15’]
- Metrics analysis [5’]
- Axiometrics framework [15’]

What is evaluation (in IR)?

- Eh...
- Ideally: a machine telling you how good an IR system is
- "Good": effective, capable to retrieve relevant (useful?!?) documents
  - (efficiency is also studied, but focus is on effectiveness)

The importance of evaluation in IR

- Everybody agrees that evaluation is of paramount importance in IR
- One of the most evaluation-oriented disciplines in computer/information sciences
- We’re busy doing a lot of evaluation since the 60s
  - So this talk is relevant. I do not know if it is useful :-)

A short history of nearly everything about IR evaluation

The importance of evaluation in IR

- Everybody agrees that evaluation is of paramount importance in IR
- One of the most evaluation-oriented disciplines in computer/information sciences
- We’re busy doing a lot of evaluation since the 60s
  - And we don’t know (agree on) how to evaluate
(Other) Issues in IR evaluation

- Relevance
  - "Topicality"?
  - "Utility"?
  - ...

- Methodology
  - Test collection, benchmark, TREC-like
  - User study (--> Diane)
  - Large log analysis
  - ...

Outline

- Evaluation [5']
- Measures / metrics [15']
- Measurement theory [15']
- Metrics analysis [5']
- Axiometrics framework [15']

We go down the dark evaluation metrics rabbit hole

- (or, measure)
- """"A number telling us how effective an IR system is"
- Simple, isn't it?

Simple, isn't it?

IR effectiveness metric

- (or, measure)
- """"A number telling us how effective an IR system is"
- Simple, isn't it?
A shorter list
- Precision, Recall
- Precision-Recall curve
- MAP (Mean Average Precision)
- P@n (Precision at n)
- NDCG (Normalized Discounted Cumulative Gain)
- MRR (Mean Reciprocal Rank)
- RBP (Rank Biased Precision)
- TBG (Time Based Gain)

Precision & Recall
- \[ P = \frac{\text{relevant} \cap \text{retrieved}}{\text{retrieved}} \]
- \[ R = \frac{\text{relevant} \cap \text{retrieved}}{\text{relevant}} \]

P & R: probabilistic definition
- \[ P = p(\text{relevant} | \text{retrieved}) \]
- \[ R = p(\text{retrieved} | \text{relevant}) \]

Ranking!
- But today all IR systems rank the documents
- Limitations of P&R
  - 2 numbers, not just one
  - Not affected by the rank of retrieved docs.
- Solutions: (too?) many.
  - Precision/Recall curve
  - MAP (Mean Average Precision)
  - ...
- Let us see some examples

<table>
<thead>
<tr>
<th>Rank</th>
<th>Rel?</th>
<th>R</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0,25</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0,5</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0,5</td>
<td>0,67</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0,75</td>
<td>0,75</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0,75</td>
<td>0,6</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0,75</td>
<td>0,5</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1</td>
<td>0,57</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>1</td>
<td>0,5</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>1</td>
<td>0,44</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>1</td>
<td>0,4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rank</th>
<th>Rel?</th>
<th>R</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0,25</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0,5</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0,5</td>
<td>0,67</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0,75</td>
<td>0,75</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0,75</td>
<td>0,6</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0,75</td>
<td>0,5</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1</td>
<td>0,57</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>1</td>
<td>0,5</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>1</td>
<td>0,44</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>1</td>
<td>0,4</td>
</tr>
</tbody>
</table>
(Because we don’t want)

We prefer

(We could do but we don’t do)

So we average the step curves
Over N queries and we get

And of course on 11 levels of recall

We happily compare systems?

Although often...

P/R curve \Rightarrow MAP

- P/R curve
  - It is not a number
  - It can be transformed into a number by measuring the area below the curve
- \Rightarrow AP (Average Precision)
- \Rightarrow MAP (Mean Average Precision)
- Good property: top-heavyness

| Rank | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | ...
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Rel</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
</tbody>
</table>

User model?

P/n

- Simply count how many relevant documents are retrieved in the first n positions of the rank
- P@10 useful for classical Web search engines
- P@1 for "Feeling lucky"
Non binary relevance

- Some documents are "more relevant" than others
- Discounted Cumulative Gain (DCG, NDCG)
  - Different relevance --> different gain for the user
    - E.g., H --> 3, R --> 2, P --> 1, N --> 0
  - Sum of the gains while walking down the rank
  - Discounting more and more: late rank positions give less gain even if of equal relevance ("top-heaviness")

---

**Example**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Rel</th>
<th>Gain</th>
<th>CG</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>H</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>P</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>N</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>R</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>H</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>R</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>N</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>P</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>N</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>P</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

---

**Example**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Rel</th>
<th>Gain</th>
<th>CG</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>H</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>P</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>N</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>R</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>H</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>6</td>
<td>R</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>7</td>
<td>N</td>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td>8</td>
<td>P</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>9</td>
<td>N</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>10</td>
<td>P</td>
<td>1</td>
<td>13</td>
</tr>
</tbody>
</table>

---

**Example**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Rel</th>
<th>Gain</th>
<th>CG</th>
<th>Discount</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>H</td>
<td>3</td>
<td>log(1)</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>P</td>
<td>1</td>
<td>log(2)</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>N</td>
<td>0</td>
<td>log(3)</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>R</td>
<td>2</td>
<td>log(4)</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>H</td>
<td>3</td>
<td>log(5)</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>R</td>
<td>2</td>
<td>log(6)</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>N</td>
<td>0</td>
<td>log(7)</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>P</td>
<td>1</td>
<td>log(8)</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>N</td>
<td>0</td>
<td>log(9)</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>P</td>
<td>1</td>
<td>log(10)</td>
<td>0</td>
</tr>
</tbody>
</table>

---

**Example**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Rel</th>
<th>Gain</th>
<th>CG</th>
<th>Discount</th>
<th>DG</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>H</td>
<td>3</td>
<td>log(1)</td>
<td>1</td>
<td>3/1=3</td>
</tr>
<tr>
<td>2</td>
<td>P</td>
<td>1</td>
<td>log(2)</td>
<td>1</td>
<td>1/1=1</td>
</tr>
<tr>
<td>3</td>
<td>N</td>
<td>0</td>
<td>log(3)</td>
<td>0</td>
<td>0/0=0</td>
</tr>
<tr>
<td>4</td>
<td>R</td>
<td>2</td>
<td>log(4)</td>
<td>0</td>
<td>2/2=1</td>
</tr>
<tr>
<td>5</td>
<td>H</td>
<td>3</td>
<td>log(5)</td>
<td>0</td>
<td>3/3=1</td>
</tr>
<tr>
<td>6</td>
<td>R</td>
<td>2</td>
<td>log(6)</td>
<td>0</td>
<td>2/2=1</td>
</tr>
<tr>
<td>7</td>
<td>N</td>
<td>0</td>
<td>log(7)</td>
<td>0</td>
<td>0/0=0</td>
</tr>
<tr>
<td>8</td>
<td>P</td>
<td>1</td>
<td>log(8)</td>
<td>0</td>
<td>1/1=1</td>
</tr>
<tr>
<td>9</td>
<td>N</td>
<td>0</td>
<td>log(9)</td>
<td>0</td>
<td>0/0=0</td>
</tr>
<tr>
<td>10</td>
<td>P</td>
<td>1</td>
<td>log(10)</td>
<td>0</td>
<td>1/1=1</td>
</tr>
</tbody>
</table>
### Axiometrics

**Example**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Sel</th>
<th>Gain</th>
<th>CG</th>
<th>Discount</th>
<th>DG</th>
<th>DCGB</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>H</td>
<td>3</td>
<td>3</td>
<td>log(1)</td>
<td>3</td>
<td>3.0</td>
</tr>
<tr>
<td>2</td>
<td>P</td>
<td>1</td>
<td>4</td>
<td>log(2)</td>
<td>1</td>
<td>4.0</td>
</tr>
<tr>
<td>3</td>
<td>N</td>
<td>0</td>
<td>6</td>
<td>log(3)</td>
<td>0</td>
<td>4.0</td>
</tr>
<tr>
<td>4</td>
<td>R</td>
<td>2</td>
<td>6</td>
<td>log(4)</td>
<td>2</td>
<td>5.0</td>
</tr>
<tr>
<td>5</td>
<td>H</td>
<td>3</td>
<td>9</td>
<td>log(5)</td>
<td>3</td>
<td>6.3</td>
</tr>
<tr>
<td>6</td>
<td>R</td>
<td>2</td>
<td>12</td>
<td>log(6)</td>
<td>2</td>
<td>7.1</td>
</tr>
<tr>
<td>7</td>
<td>N</td>
<td>0</td>
<td>12</td>
<td>log(7)</td>
<td>0</td>
<td>7.1</td>
</tr>
<tr>
<td>8</td>
<td>P</td>
<td>1</td>
<td>12</td>
<td>log(8)</td>
<td>1</td>
<td>7.4</td>
</tr>
<tr>
<td>9</td>
<td>N</td>
<td>0</td>
<td>12</td>
<td>log(9)</td>
<td>0</td>
<td>7.4</td>
</tr>
<tr>
<td>10</td>
<td>P</td>
<td>1</td>
<td>13</td>
<td>log(10)</td>
<td>1</td>
<td>7.7</td>
</tr>
</tbody>
</table>

**Example**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Sel</th>
<th>Gain</th>
<th>CG</th>
<th>Discount</th>
<th>DG</th>
<th>DCGB</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>H</td>
<td>3</td>
<td>3</td>
<td>log(1)</td>
<td>3</td>
<td>3.0</td>
</tr>
<tr>
<td>2</td>
<td>P</td>
<td>1</td>
<td>4</td>
<td>log(2)</td>
<td>1</td>
<td>4.0</td>
</tr>
<tr>
<td>3</td>
<td>N</td>
<td>0</td>
<td>6</td>
<td>log(3)</td>
<td>0</td>
<td>4.0</td>
</tr>
<tr>
<td>4</td>
<td>R</td>
<td>2</td>
<td>6</td>
<td>log(4)</td>
<td>2</td>
<td>5.0</td>
</tr>
<tr>
<td>5</td>
<td>H</td>
<td>3</td>
<td>9</td>
<td>log(5)</td>
<td>3</td>
<td>6.3</td>
</tr>
<tr>
<td>6</td>
<td>R</td>
<td>2</td>
<td>11</td>
<td>log(6)</td>
<td>2</td>
<td>7.1</td>
</tr>
<tr>
<td>7</td>
<td>N</td>
<td>0</td>
<td>11</td>
<td>log(7)</td>
<td>0</td>
<td>7.1</td>
</tr>
<tr>
<td>8</td>
<td>P</td>
<td>1</td>
<td>12</td>
<td>log(8)</td>
<td>1</td>
<td>7.4</td>
</tr>
<tr>
<td>9</td>
<td>N</td>
<td>0</td>
<td>12</td>
<td>log(9)</td>
<td>0</td>
<td>7.4</td>
</tr>
<tr>
<td>10</td>
<td>P</td>
<td>1</td>
<td>13</td>
<td>log(10)</td>
<td>1</td>
<td>7.7</td>
</tr>
</tbody>
</table>

**Example**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Sel</th>
<th>Gain</th>
<th>CG</th>
<th>Discount</th>
<th>DG</th>
<th>DCGB</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>H</td>
<td>3</td>
<td>3</td>
<td>log(1)</td>
<td>3</td>
<td>3.0</td>
</tr>
<tr>
<td>2</td>
<td>P</td>
<td>1</td>
<td>4</td>
<td>log(2)</td>
<td>1</td>
<td>4.0</td>
</tr>
<tr>
<td>3</td>
<td>N</td>
<td>0</td>
<td>6</td>
<td>log(3)</td>
<td>0</td>
<td>4.0</td>
</tr>
<tr>
<td>4</td>
<td>R</td>
<td>2</td>
<td>6</td>
<td>log(4)</td>
<td>2</td>
<td>5.0</td>
</tr>
<tr>
<td>5</td>
<td>H</td>
<td>3</td>
<td>9</td>
<td>log(5)</td>
<td>3</td>
<td>6.3</td>
</tr>
<tr>
<td>6</td>
<td>R</td>
<td>2</td>
<td>11</td>
<td>log(6)</td>
<td>2</td>
<td>7.1</td>
</tr>
<tr>
<td>7</td>
<td>N</td>
<td>0</td>
<td>11</td>
<td>log(7)</td>
<td>0</td>
<td>7.1</td>
</tr>
<tr>
<td>8</td>
<td>P</td>
<td>1</td>
<td>12</td>
<td>log(8)</td>
<td>1</td>
<td>7.4</td>
</tr>
<tr>
<td>9</td>
<td>N</td>
<td>0</td>
<td>12</td>
<td>log(9)</td>
<td>0</td>
<td>7.4</td>
</tr>
<tr>
<td>10</td>
<td>P</td>
<td>1</td>
<td>13</td>
<td>log(10)</td>
<td>1</td>
<td>7.7</td>
</tr>
</tbody>
</table>

**Example**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Sel</th>
<th>Gain</th>
<th>CG</th>
<th>Discount</th>
<th>DG</th>
<th>DCGB</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>H</td>
<td>3</td>
<td>3</td>
<td>log(1)</td>
<td>3</td>
<td>3.0</td>
</tr>
<tr>
<td>2</td>
<td>P</td>
<td>1</td>
<td>4</td>
<td>log(2)</td>
<td>1</td>
<td>4.0</td>
</tr>
<tr>
<td>3</td>
<td>N</td>
<td>0</td>
<td>6</td>
<td>log(3)</td>
<td>0</td>
<td>4.0</td>
</tr>
<tr>
<td>4</td>
<td>R</td>
<td>2</td>
<td>6</td>
<td>log(4)</td>
<td>2</td>
<td>5.0</td>
</tr>
<tr>
<td>5</td>
<td>H</td>
<td>3</td>
<td>9</td>
<td>log(5)</td>
<td>3</td>
<td>6.3</td>
</tr>
<tr>
<td>6</td>
<td>R</td>
<td>2</td>
<td>11</td>
<td>log(6)</td>
<td>2</td>
<td>7.1</td>
</tr>
<tr>
<td>7</td>
<td>N</td>
<td>0</td>
<td>11</td>
<td>log(7)</td>
<td>0</td>
<td>7.1</td>
</tr>
<tr>
<td>8</td>
<td>P</td>
<td>1</td>
<td>12</td>
<td>log(8)</td>
<td>1</td>
<td>7.4</td>
</tr>
<tr>
<td>9</td>
<td>N</td>
<td>0</td>
<td>12</td>
<td>log(9)</td>
<td>0</td>
<td>7.4</td>
</tr>
<tr>
<td>10</td>
<td>P</td>
<td>1</td>
<td>13</td>
<td>log(10)</td>
<td>1</td>
<td>7.7</td>
</tr>
</tbody>
</table>

### Example

<table>
<thead>
<tr>
<th>Rank</th>
<th>Sel</th>
<th>Gain</th>
<th>CG</th>
<th>Discount</th>
<th>DG</th>
<th>DCGB</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>H</td>
<td>3</td>
<td>3</td>
<td>log(1)</td>
<td>3</td>
<td>3.0</td>
</tr>
<tr>
<td>2</td>
<td>P</td>
<td>1</td>
<td>4</td>
<td>log(2)</td>
<td>1</td>
<td>4.0</td>
</tr>
<tr>
<td>3</td>
<td>N</td>
<td>0</td>
<td>6</td>
<td>log(3)</td>
<td>0</td>
<td>4.0</td>
</tr>
<tr>
<td>4</td>
<td>R</td>
<td>2</td>
<td>6</td>
<td>log(4)</td>
<td>2</td>
<td>5.0</td>
</tr>
<tr>
<td>5</td>
<td>H</td>
<td>3</td>
<td>9</td>
<td>log(5)</td>
<td>3</td>
<td>6.3</td>
</tr>
<tr>
<td>6</td>
<td>R</td>
<td>2</td>
<td>11</td>
<td>log(6)</td>
<td>2</td>
<td>7.1</td>
</tr>
<tr>
<td>7</td>
<td>N</td>
<td>0</td>
<td>11</td>
<td>log(7)</td>
<td>0</td>
<td>7.1</td>
</tr>
<tr>
<td>8</td>
<td>P</td>
<td>1</td>
<td>12</td>
<td>log(8)</td>
<td>1</td>
<td>7.4</td>
</tr>
<tr>
<td>9</td>
<td>N</td>
<td>0</td>
<td>12</td>
<td>log(9)</td>
<td>0</td>
<td>7.4</td>
</tr>
<tr>
<td>10</td>
<td>P</td>
<td>1</td>
<td>13</td>
<td>log(10)</td>
<td>1</td>
<td>7.7</td>
</tr>
</tbody>
</table>

**A shorter list**

- Precision, Recall
- Precision-Recall curve
- MAP
- P@n
- NDCG
- MRR (Mean Reciprocal Rank)
- RBP (Rank Biased Precision)
- TBG (Time Based Gain)

**... even continuous relevance**

- Dynamometer, Hand force grip, even physiological data
- IR System showing a "relevance bar" close to each document
  - "estimation of the amount of relevance"
  - Magnitude estimation (paper @ last SIGIR)
- ...
**Metrics classification**

- Metrics could be classified on the basis of underlying notions of:
  - relevance (binary, ranking, continuous)
  - retrieval (binary, ranking, continuous)
  - 3 x 3 (or N x N) grid, ...

**Classification (incomplete!)**

<table>
<thead>
<tr>
<th>Relevance</th>
<th>Sliding ratio</th>
<th>ADJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rank</td>
<td>NDCG, Kendall</td>
<td>Spearman?</td>
</tr>
<tr>
<td>Binary</td>
<td>PBR, E, F, ...</td>
<td></td>
</tr>
<tr>
<td>Retrieval</td>
<td>RP curve, MAP, R-prec</td>
<td></td>
</tr>
</tbody>
</table>

---

**Take home messages so far**

- In IR, Evaluation is important.
- There are many (100+) metrics
  - System-oriented metrics only
    - (R, P, MAP, NDCG, ...)
  - Let alone human/user-oriented metrics
    - (user satisfaction, fatigue, ...)
  - And still IR only, no clustering, filtering...
  - Just a few are indeed used
    - (which ones? Why those? -- Julio)
  - Definitions of some of them

---

**Outline**

- Evaluation [5’]
- Measures / metrics [15’]
- Measurement theory [15’]
- Metrics analysis [5’]
- Axiometrics framework [15’]
Measurement

- Definition: A process aimed at determining a relationship between a physical quantity and a unit of measurement
- Typically, one assigns numbers to objects/events
- Studied in Measurement Theory
  - Reasonably settled

Measurement Theory

- One important concept: which numbers can I select when measuring? What properties do they have?
- Which measurement scale?
  - (or "level")
  - E.g., to measure length:
    - Meters
    - Inches
    - "Longer than" (?)
4. Ratio scale
- I'm twice taller than him
- He is twice richer than me
  - (Both "how much" & "how many")
- I'm twice older than you
  - Years, months (*12), days, ...
- Zero
  - Age starts from zero!
- But not
  - Today is twice as hot as yesterday?

3. Interval scale
- Today is twice as hot as yesterday?
- In the last two days we had the same increase of 5°C in temperature
  - The difference between today and yesterday temperature is the same as ...
- Dates are another example
  - Ratio: KO (2000 is not 2 * 1000)
  - (ratios of differences: OK)

2. Ordinal scale
- A measure is not an amount but a **rank**
- It is a form of measurement!
- Ex: Lines ranked according to their length
- It does **not** mean that:
  - the first is twice as long than the second
  - the length difference between the 1st and the 2nd is the same as the 2nd and the 3rd

1. Nominal scale
- Qualitative
- Categories
- Names, gender, nationality, ...
- Can be Dichotomous or Non-dichotomous
- No assumptions on ratios, distances, ranks.

And indeed
- Nicholas Chrisman
- 1998
- Proposes 10 scales, not just 4

Anyway "Good Old Fashioned" Measurement scales
1. Nominal
2. Ordinal
3. Interval
4. Ratio
Legit operations
- Given a scale, only some operations make sense
  - Arithmetic: +, -, *, /
  - Statistic: Mean, median, mode, ...
- Ex:
  - Average height, weight, ...: OK
  - Average gender: KO

Legit relational/math operations
- \(=, \neq, >, <, +, -, \times, \div\)

<table>
<thead>
<tr>
<th>Scale</th>
<th>=</th>
<th>#</th>
<th>&gt;</th>
<th>&lt;</th>
<th>+</th>
<th>-</th>
<th>\times</th>
<th>\div</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ordinal</td>
<td>✓</td>
<td>✓</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interval</td>
<td>✓</td>
<td>✓</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ratio</td>
<td>✓</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Permissible transformations
- \(a \cdot x\)
- \(a \cdot x + b\)
- Monotonic
- 1-to-1

<table>
<thead>
<tr>
<th>Scale</th>
<th>(a \cdot x)</th>
<th>(a \cdot x + b)</th>
<th>Monotonic</th>
<th>1-to-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Ordinal</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Interval</td>
<td>✓</td>
<td>✓</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Ratio</td>
<td>✓</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>

Meaning
- If you transform the measure, are you still measuring the same thing?
- Nationality
- Rank
- Temperature
- Money

Examples
- Nominal scale for nationality
  - Greek = 1
  - Italian = 2
  - Spanish = 3
  - Japanese = 4
  - ...
  - 3 - 1 = 4 - 2. Uh?!?
- Whereas interval scale for temperature \(^\circ\)C
  - 30\(^\circ\) - 10\(^\circ\) = 40\(^\circ\) - 20\(^\circ\): ok

Legit statistics
- Mode
- Median
- Mean
- Arithmetic
- Geometric, Harmonic

<table>
<thead>
<tr>
<th>Scale</th>
<th>Mode</th>
<th>Median</th>
<th>Mean</th>
<th>Arithmetic</th>
<th>Geometric, Harmonic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal</td>
<td>✓</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ordinal</td>
<td>✓</td>
<td>✓</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interval</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Ratio</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
</tbody>
</table>
Examples

- Nationality, mode, ok
- Mean rank? Uh?!
  - (think of ranking ten students...)
- Mean temperature, ok

Why are you telling me this?

Two reasons

1. Measurement theory and scales can be used to directly analyze IR metrics
2. Because IR can be seen as a measurement. Of relevance --> Language to define axioms on metrics

Let's see 1. first

Outline

- Evaluation [5']
- Measures / metrics [15']
- Measurement theory [15']
- Metrics analysis [5']
- Axiometrics framework [15']
MRR (Mean Reciprocal Rank)
- Take the rank of the 1st relevant doc.
- Take the reciprocal (…)
  \[ RR = \frac{1}{\text{rank}(i)} \]
- Then take the mean (…) over some topics
  \[ MRR = \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{\text{rank}(i_q)} \]

Even worse than that... NDCG
- H,R,P,N --> 3, 2, 1, 0 ?
  - Linear, most common
  - H,R,P,N --> 100, 10, 1, 0 ?! (or 4, 2, 1, 0)
  - Exponential, sometimes used, actually
  - H,R,P,N --> 100, 99, 90, 0 ?!?
    - “Crazy”, never heard of... why?
  - Arbitrary choice!
  - By transforming relevance levels into gains, we transform an 2. Ordinal scale --> 4. Ratio scale!
  - And we also discount
    - Dividing by log(rank)...
    - All metrics that can be modeled as gain/discount...

MRR?!
- It is used
  - In many papers
  - Even in some TREC tracks
- When analyzed with "measurement theory glasses", is is not "measurement theory proof"
- A reciprocal is taken on an ordinal scale...
- ... then it is averaged...

Classification (more principled, still incomplete!)

To summarize, but not to conclude
- 100+ metrics
  - Measurement theory seems a useful tool
    - Metric classification
    - Some arbitrary choices
    - Some metrics are not "Measurement theory proof"
  - Metric engineering seems more an art (artisan) than a science...
- A more principled approach?

Outline
- Evaluation [5’]
- Measures / metrics [15’]
- Measurement theory [15’]
- Metrics analysis [5’]
- Axiometrics framework [15’]
2. IR as measurement of relevance

- Both an IR system and a human assessor **measure** the **relevance** of a document
- Maybe on different **scales**:
  - Ranked SE output: ordinal scale
  - Human TREC-like relevance assessor: nominal (ordinal) scale (binary, relevant/nonrelevant)
  - ... IR related tasks like categorization, filtering, ...

**Notation: \( \sigma \) and \( \alpha \)**

- 2 relevance measurements
  - by a system (\( \alpha \))
    - \( \sigma(q,d), \sigma(q,D), \sigma(Q,D) \)
    - e.g. SE ranked output
  - by a human (user/assessor) (\( \alpha \))
    - \( \alpha(q,d), \alpha(q,D), \alpha(Q,D) \)
    - e.g., TREC qrels

**Similarity**

- Given two relevance measurements \( \alpha \) and \( \sigma \), we can define a notion of similarity between them (given a query/topic \( q \) and a document \( d \))

\[
\text{sim}_{q,d} (\alpha, \sigma) 
\]

- And an IR system should provide a measurement \( \sigma \) that is similar to the human measurement \( \alpha \)
Features of Similarity
Amos Tversky
Hebrew University
Jerusalem, Israel

The metric and dimensional assumptions that underlie the geometric representation of similarity are questioned on both theoretical and empirical grounds. A new axiomatic approach to similarity is developed in which objects are represented as collections of features, and similarity is described as a feature-matching process. Specifically, a set of similarity assumptions is derived:

- Similarity is then a function that takes as arguments two objects.
- We do not need to refer to the metric itself but rather to the metric value.
- And we can compare similarities.
- For example,

\[ \text{sim}(\alpha, \sigma) < \text{sim}(\alpha, \sigma') \]

means that on the query \( q \) and the document \( d \) and given the human relevance judgment \( \alpha \), system \( \alpha \) is worse than \( \alpha' \) (i.e., less similar to \( \alpha \)).

So, to summarize (but not to conclude!)

- **Measurement** theory, Measurement scales, IR as relevance measurement
  - \( \sigma(q,d), \sigma(q,D), \sigma(Q,D), \alpha(q,d), \alpha(q,D), \alpha(Q,D) \)
- **Similarity**
  \[ \text{sim}(\alpha, \sigma) < \text{sim}(\alpha, \sigma') \]
- **Metric**

Generality: Across different measurement scales

- We can compute the similarity of two relevance judgments when they are on the same scale.
- ...but more than that...
- ... also when they are on different scales, in some cases.
  - E.g., the classical ad-hoc retrieval
    - scale(\( \alpha \)) = \([\text{ordinal}]\)
    - scale(\( \sigma \)) = \([\text{nominal}] \) (binary relevance. Ordinal)

Same scales: binary IR

\[ \text{sim}(\alpha, \sigma) > \text{sim}(\alpha, \sigma') \]
I don't know if they are complete. I think axioms are correct and consistent.

\[ \text{Non Rel} \quad \text{Rel} \]

\[ \bullet \quad \bullet \quad \bullet \]

\[ \bullet \quad \bullet \quad \bullet \]

\[ \text{sim}(\alpha, \alpha_1) < \text{sim}(\alpha, \alpha_2) \]

Some details...

- I do not trust our axioms too much yet...
- ... preliminary work...
- Actually:
  - I think axioms are correct and consistent
  - I don't know if they are complete
  - Stating axioms is also useful to "test the framework"
  - Measurement theory is an effective language to state them!

A First Axiom

**Axiom 3** (Similarity of two systems). Let \( q \) be a query, \( d \) a document, \( \alpha \) a human relevance measurement and \( \sigma \) and \( \sigma' \) two system relevance measurements such that

\[ \sigma(q, d) = \sigma'(q, d). \] (1)

Then

\[ \text{sim}_{q,d}(\alpha, \sigma) = \text{sim}_{q,d}(\alpha, \sigma'). \] (2)

Ok, ok. Second Axiom

**Axiom 6** (Overestimated documents). Let \( q \) be a query, \( d \) and \( d' \) two documents, \( \alpha \) a human relevance measurement and \( \sigma \) a system relevance measurement such that

\[ \alpha(d) > \alpha(d'), \]

\[ \text{sim}_{q,d}(\alpha, \sigma) < \text{sim}_{q,d'}(\alpha, \sigma) \]

and (6) and (8) hold (i.e., both \( d \) and \( d' \) are overestimated), then

\[ \text{metric}_{q,d}(\alpha, \sigma) < \text{metric}_{q,d'}(\alpha, \sigma). \]

- \( d \) is more relevant, \( \text{sim} \) on \( d \) is lower, then metric value on \( d \) has to be lower.
- \( d \) is both "more wrong" and "more visible" to the user.
Ok, ok. Third Axiom

Axiom 8 (System relevance). Let \( q \) be a query, \( d \) and \( d' \) two documents, \( \alpha \) a human relevance measurement and \( \sigma \) a system relevance measurement such that:

\[
\text{sim}_{q,d}(\alpha, \sigma) = \text{sim}_{q,d'}(\alpha, \sigma), \quad \sigma(d) > \sigma(d'), \quad \text{and}
\]

\[
\alpha(d) \geq \alpha(d').
\]

Then

\[
d \triangleright_{\text{metric}(\alpha, \sigma)} d'.
\]

(document \( d \) affects the metric value more than \( d' \))

Meaning?

- Corollary:
  - By taking \( \text{scale}(\alpha) = \text{[Rank]} \) we derive that:
    - Early rank positions affect a metric value more than later rank positions
    - IR metrics should be "top-heavy"
  - Previous Axiom 8 states a more abstract/general principle, independent of the scales

Now, a last Axiom

Axiom 9 (User relevance). Let \( q \) be a query, \( d \) and \( d' \) two documents, \( \alpha \) a human relevance measurement and \( \sigma \) a system relevance measurement such that:

\[
\text{sim}_{q,d}(\alpha, \sigma) = \text{sim}_{q,d'}(\alpha, \sigma), \quad \sigma(d) > \sigma(d'), \quad \text{and}
\]

\[
\sigma(d) \geq \sigma(d').
\]

Then

\[
d \triangleright_{\text{metric}(\alpha, \sigma)} d'.
\]

(document \( d \) affects the metric value more than \( d' \))

Meaning?

- A metric should weigh more, and be more affected, by more relevant documents
  - "\( \alpha \) top heavyness", "human top heavyness"
  - Perhaps less intuitive than previous axiom,
  - but it does indeed seem natural in the framework
  - by symmetry (treat \( \alpha \) as \( \sigma \))
  - To evaluate a nonrelevant document as nonrelevant is an easy job (the vast majority of documents in a collection are nonrelevant)

Meaning?

- Consequence: linear gain values of 3, 2, 1, 0 in NDCG (for H, R, P, N) can be questioned
  - Exponential 100, 10, 1, 0 (or 4, 2, 1, 0) might be better
  - (already proposed in the original paper, but not much used...)
  - And "crazy" 100, 99, 90, 0 is wrong!

A theorem

Theorem 2 (Unbalanced query). Let \( Q \) be a query set, \( q \notin Q \) a query, \( D \) a document set, \( \alpha \) a human relevance measurement and \( \sigma \) and \( \sigma' \) two system relevance measurements such that:

\[
\text{metric}_{Q,D}(\alpha, \sigma) > \text{metric}_{Q,D}(\alpha, \sigma')
\]

and

\[
\text{metric}_{Q,\{q\}, D}(\alpha, \sigma) \leq \text{metric}_{Q,\{q\}, D}(\alpha, \sigma').
\]

Then

\[
\text{metric}_{Q,D}(\alpha, \sigma) < \text{metric}_{Q,D}(\alpha, \sigma').
\]
6.2 From Similarity to Metric

The first axioms represent basic constraints on similarity. Let $Q$ be a query set, $q \in Q$ a query, $D$ a document set, $\alpha$ a human relevance measurement, and $\sigma$ and $\sigma'$ two system effectiveness metrics such that

$$\text{metric}(\alpha, \sigma) > \text{metric}(\alpha, \sigma')$$

on the query set $Q$ and the document collection $D$, and given the human relevance judgment $\alpha$, $\sigma$ is more effective than $\sigma'$.

Then

$$\text{metric}(\alpha, \sigma) < \text{metric}(\alpha, \sigma')$$

$\sigma$ is less effective than $\sigma'$ also on the new query $q$.

**Theorem 2** (Unbalanced query). Let $Q$ be a query set, $q \in Q$ a query, $D$ a document set, $\alpha$ a human relevance measurement, and $\sigma$ and $\sigma'$ two system effectiveness metrics such that

$$\text{metric}(\alpha, \sigma) > \text{metric}(\alpha, \sigma')$$

and

$$\text{metric}(\alpha, \sigma) < \text{metric}(\alpha, \sigma')$$

Then

$$\text{metric}(\alpha, \sigma) < \text{metric}(\alpha, \sigma')$$

we add a new query $q$ and then $\sigma$ becomes less effective than $\sigma'$.

---

**Biblio**

(I'm here until Friday, and I'm also on facebook and Twitter :-) )

(Ahem... Instagram? Pinterest?!? Periscope?!? Snapchat?!?!)