Notes on what to measure in INEX

Gabriella Kazai Mounia Lalmas

Queen Mary University of London, UK

Second edition!

- www.dcs.qmul.ac.uk/~gabs
- Publications page

Outline

- What to measure
  - Retrieval task
  - User behaviour
- Requirements for a metric
- A small experiment
- Conclusions

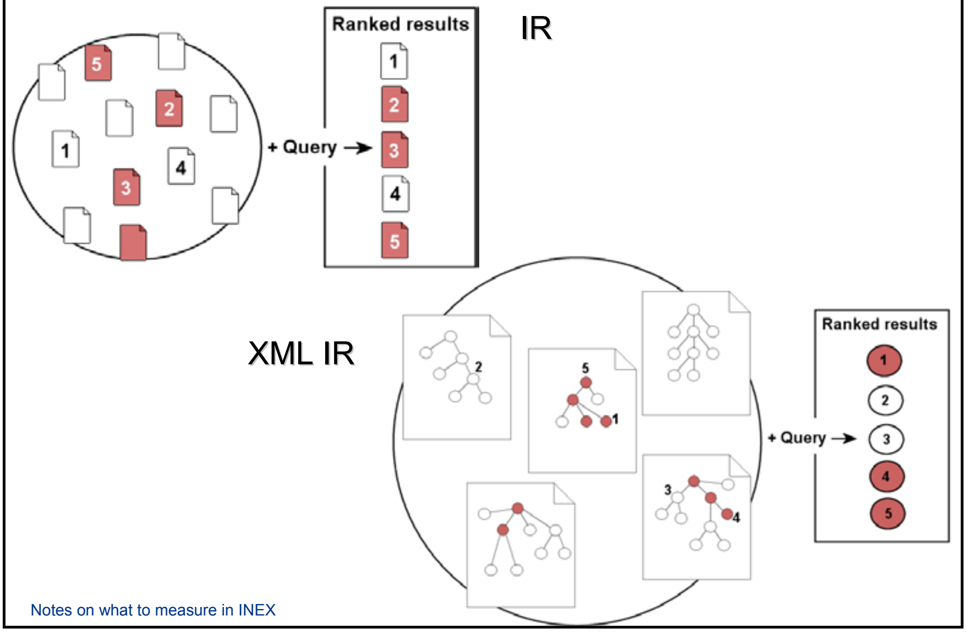
What to measure?

- Retrieval effectiveness
- Rank systems according to how well they satisfy a user's query given a

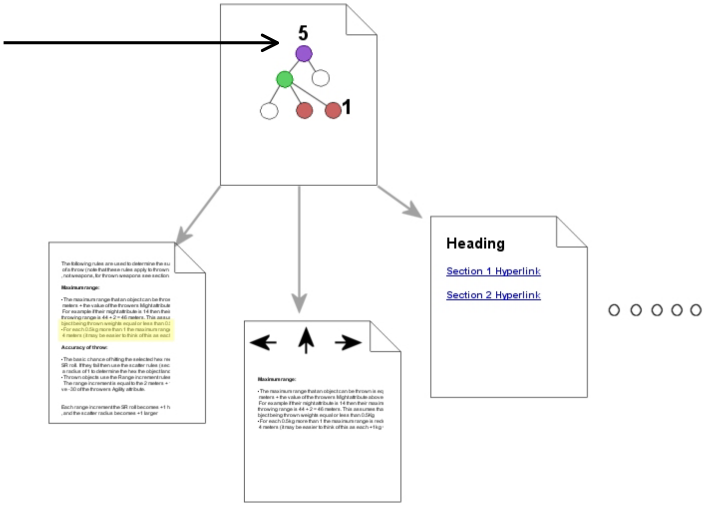
  retrieval task and a

  model of user behaviour.

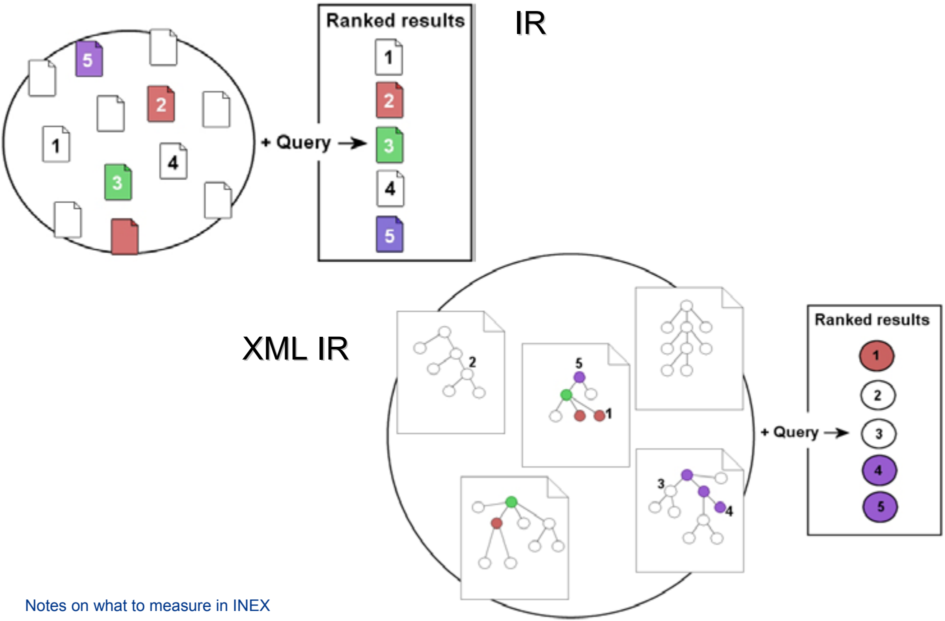
Retrieval task and user behaviour



Browsing/scrolling/etc?



Multiple degrees of relevance



Requirements

- Consider element size
- Allow partial score for near-misses
- Do not reward overlap nor penalise overlap-free runs
- Consider linear and other non-linear presentation
- Handle multiple dimensions (exh, spec)
- Handle multiple relevance degrees
- Ideal recall-base
- Normalisation

Metrics

- i2 (inex_eval)
- i3 (inex_eval_ng)
- XCG (cumulated gain for XML)
- PRUM (precision recall with user modeling)

- (T2I) (tolerance to irrelevance)
- (ERR) (expected ratio of relevant)

i2 metric

- Raghavan's precall [Gövert et al. 2002]:

$$P(rel\,|\,retr)(x) := \frac{x \cdot n}{x \cdot n + esl_{x \cdot n}} = \frac{x \cdot n}{x \cdot n + j + s \cdot i/(r+1)}$$

- Quantisation functions
  - Strict $f_{strict}(e,s) = \begin{cases} 1 \text{ if } (es) = 33 \\ 0 \text{ otherwise} \end{cases}$
  - Generalised

$$f_{gen}(e,s) = \begin{cases} 1.00 & \mathit{if} \quad (es) = 33 \\ 0.75 & \mathit{if} \quad (es) \in \{23,32,31\} \\ 0.50 & \mathit{if} \quad (es) \in \{13,22,21\} \\ 0.25 & \mathit{if} \quad (es) \in \{11,12\} \\ 0.00 & \mathit{if} \quad (es) = 00 \end{cases}$$

i3 metric

- E,S in ideal concept space [Gövert et al. 2005]:

$$r_o = \frac{\sum_{i=1}^{k} e(c_i) \cdot \frac{|c'_i|}{|c_i|}}{\mathrm{Re}\,l^U} \qquad p_o = \frac{\sum_{i=1}^{k} s(c_i) \cdot |c'_i|}{\sum_{i=1}^{k} |c'_i|}$$

- Quantisation functions
  - Strict

$$e_{strict}(e) = \begin{cases} 1 & \text{if } e = 3 \\ 0 & \text{otherwise} \end{cases}$$

$$s_{strict}(s) = \begin{cases} 1 & \text{if } s = 3 \\ 0 & \text{otherwise} \end{cases}$$

  - Generalised

$$e_{gen}(e) = e/3 \qquad s_{gen}(s) = s/3$$

nXCG metric

- Cumulated Gain for XML [Kazai et al. 2004]

$$nXCG[i] = \frac{\sum_{j=1}^{i} XG[j]}{\sum_{j=1}^{i} XI[j]}$$

- Quantisation functions
  - Strict, generalised
  - Specificity-oriented generalised (SOG)

$$f_{SOG}(e,s) = \begin{cases} 1.00 & if \quad (es) = 33 \\ 0.9 & if \quad (es) = 23 \\ 0.75 & if \quad (es) \in \{13, 32\} \\ 0.5 & if \quad (es) = 22 \\ \ldots & \end{cases}$$

PRUM metric

- Precision recall with user modeling [Piwowarski et al. 2005]

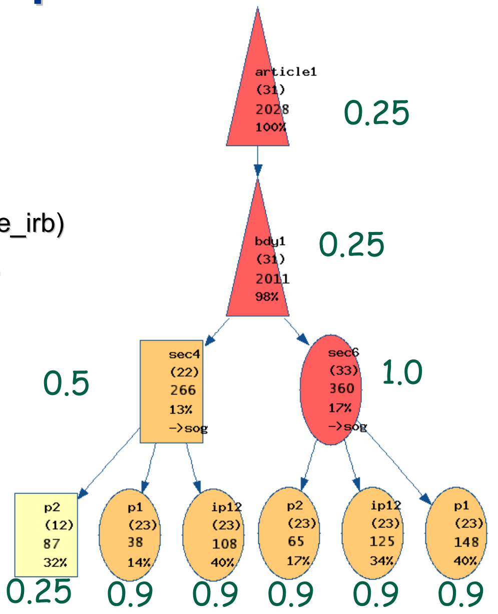
$$PRUM(l) = P(Lur \mid \mathrm{Re}\,tr, L = l, Q = q)$$

Metrics and requirements

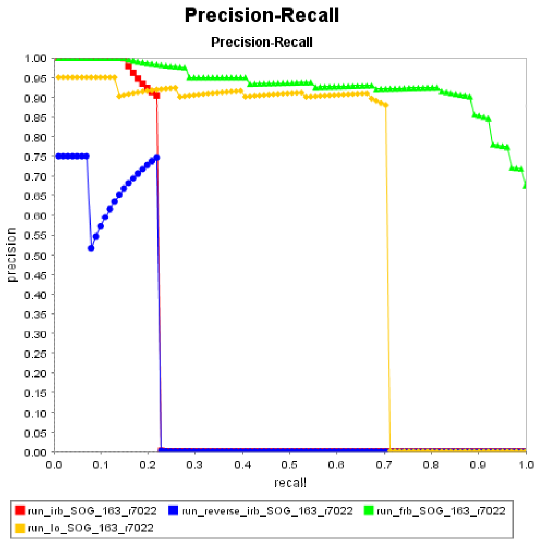
	i2	i3	XCG	PRUM
Element size	no	yes	indirectly	no
Ideal recall-base	no	ind.	yes	yes
Near-misses	no	ind.	yes	yes
Overlap	no	yes	yes	yes
Output: linear	yes	yes	yes	yes
Output: non-linear	no	no	no	no
Multiple dimensions	yes	yes	yes	yes
Multiple degrees	no	no	yes	no
Normalisation	no	no	yes	no

Experiment

- Runs (using SOG)
  - Full recall-base run (frb)
  - Ideal run (irb)
  - Reverse ideal run (reverse_irb)
  - Relevant leaf only run (lo)



i2 metric



irb

r1

$$r = \frac{1}{6.75} = 0.14$$

$$p = \frac{1}{1 + 0 + 1 \cdot 0/(1+1)} = 1$$

r2

$$r = \frac{1 + 0.5}{6.75} = 0.22$$

$$p = \frac{1.5}{1.5 + 0 + 0.5 \cdot 0.5/(0.5+1)} = 0.9$$

reverse_irb

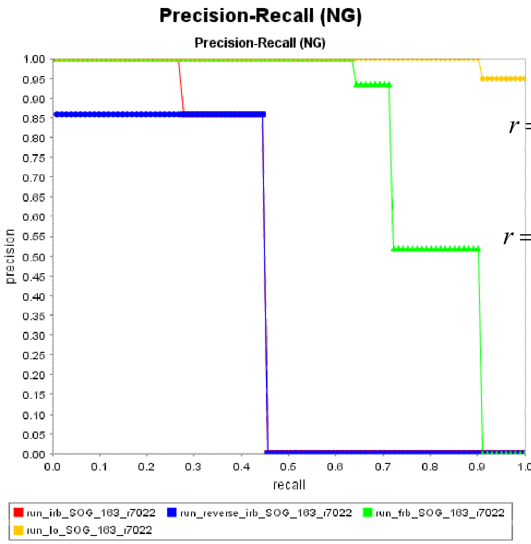
r1

$$r = \frac{0.5}{6.75} = 0.07$$

$$p = \frac{0.5}{0.5 + 0 + 0.5 \cdot 0.5/(0.5+1)} = 0.75$$

…

i3 metric



irb

r1

$$r = \frac{1\cdot(360/360)}{3.67} = 0.27 \qquad p = \frac{1\cdot 360}{360} = 1$$

r2

$$r = \frac{1\cdot(360/360) + 0.66\cdot(266/266)}{3.67} = 0.45$$

$$p = \frac{1\cdot 360 + 0.66\cdot 266}{360 + 266} = 0.86$$

reverse_irb

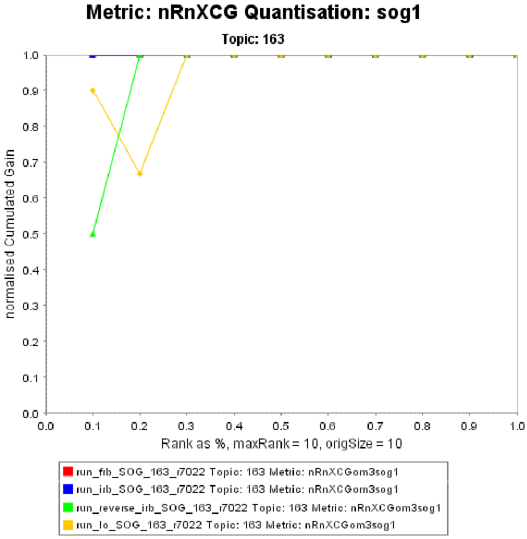
r1

$$r = \frac{0.66\cdot(266/266)}{3.67} = 0.18$$

$$p = \frac{0.66\cdot 266}{266} = 0.67$$

…

nXCG metric



irb

r1

$$nXCG = \frac{1}{1} = 1$$

r2

$$nXCG = \frac{1 + 0.5}{1.5} = 1$$

reverse_irb

r1

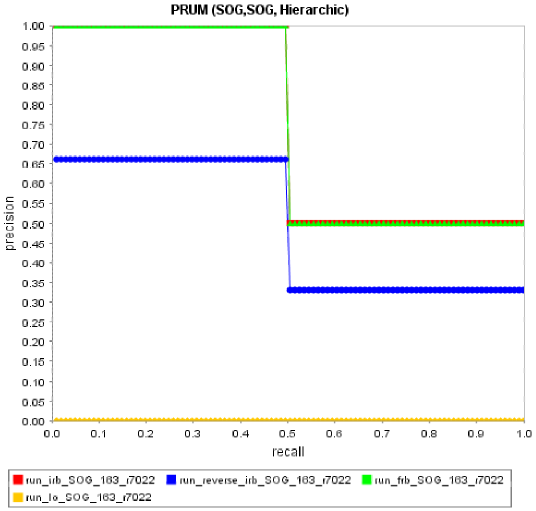
$$nXCG = \frac{0.5}{1} = 0.5$$

r2

$$nXCG = \frac{0.5 + 1}{1.5} = 1$$

PRUM metric

PRUM (SOG,SOG, Hierarchic)



Conclusions

- i2 metric - needs to go or address additional requirements (~PRUM)
- i3, XCG, PRUM (T2I) - which one to use as official or use all?

Thank you