PRUM
Precision Recall with User Modelling

Application to XML

B. Piwowarski
P. Gallinari
G. Dupret
Outline

- Why measuring?
- XML Information Retrieval
- Current metrics
  - GPR, PRng, XCG
- PRUM
  - A new user model
  - A new metric
- Comparison
Why measuring?

- The evaluation methodology started early (Cleverdon, 1967)
- Justification of pragmatic/theoretic developments
- In new IR paradigms (XML, video, etc.), measures have the same role to play... but none of the proposed metrics are satisfying.
XML Retrieval (and INEX...) 

“...more precise access by giving more specific answers...” (ie. XML nodes)

Assessments: a new scale

- Specificity (S): The extent to which a document component is focused on the information (4 values: 0 to 3)
- Exhaustivity (E): The extent to which the information contained in a document component satisfies the user's query need (4 values: 0 to 3)
Precision Recall

- Only exact answers are rewarded
- Too strict!
- Overlap problem in recall base
Generalised Precision Recall

- Assign a relevance value to near misses
  (quantisation)
- Recall base is expanded
- Overlap is a problem
PRng (N. Gövert)

- Exhaustivity and specificity based on the notion of an ideal concept space upon which precision and recall are defined.
- Considers overlap in retrieval results
  - Very sensitive to list order
- Does not consider overlap in recall-base (not completely true)
PRng (N. Gövert)

Recall(a,b,c)
= 0.75 + 0 \times 1 + 0 \times 0.5
= 0.75

Recall(c,b,a)
= 0.5 + .5 \times 1 + 1/3 \times 0.75
= 1.25

Recall(c,b,a)
= 0.5 + .5 \times 1 + 1/3 \times 0.75
= 1.25
XCG (G. Kazai)

- Based on cumulated gain measure for IR
- Accumulates gain obtained by retrieving elements at fixed ranks
- Based on the construction of an ideal recall-base
- Consider overlap in both retrieval results and recall-base
XCG (G. Kazai)

- The most consistent metric
  - Normalisation maximises the gain of near misses
  - Normalisation to ensure that performance is bounded by an “ideal” system
- ... but
  - No precise user model
  - No precision dimension
Towards user models

- Web and others
  - Quintana
  - Dunlop... leads to:
- Tolerance To Irrelevance (A. de Vries)
  - The first real user model in XML IR
  - Some theoretic problems
PRUM: main ideas

- Construction of an ideal recall base without overlap (like XCG)
- Definition of a precise user model (which includes T2I as a special case)
- Sound formal grounds (based on Raghavan's probabilistic precision-recall)
The PRUM user model

The user wants to see 4 relevant elements
The PRUM user model

The user has seen 0 relevant element

Consulted elements

P=0

Seen elements
The PRUM user model

The user has seen 1 relevant element
The PRUM user model

**Consulted elements**

```
| a | b | c |
```

\[ P = \frac{2}{3} \]

**Seen elements**

The user has seen 2 relevant elements
The PRUM user model

Consulted elements:

- a
- b
- c
- d

P = 2/4

Seen elements:

- a
- b
- c
- k

The user has seen 2 relevant elements
The PRUM user model

The user has **seen** 4 relevant elements
PRUM stochastic user model
PRUM stochastic user model
PRUM definition

- Extends Raghavan's definition

\[ P(L_{\text{ur}} | C_s, L=l, Q=q) \]

- The element \textbf{L}eads to an \textbf{U}nseen \textbf{R}elevant element
- The element is \textbf{C}onsulted by the user
- The user wants to see \( l \) % of the relevant elements
Computing PRUM

- Two probabilities have to be computed at each rank in the consulted list:
  - The probability that the user has seen $s$ distinct relevant elements
    $$P(F_i = s)$$
  - The probability that the user sees a new relevant element (knowing that ...)
    $$P(F_i > F_{i-1} | F_{i-1} = s)$$
Experiments (INEX 2004)

- 5 hypothetic systems
  1) ideal (I) = ancestors (A)
  2) parent (P) ~ biggest child (B)
  3) document (D)

- Orders
  - GRP: A >> I > P > D >> B
  - RPng: A > I ~ P > D >> B
  - PRUM: I = A >> B >> P >> D
Conclusion

- **PRUM**
  - Extension of Precision-Recall (probabilistic)
  - Explicit parameters ($\neq$ GPR, PRng, XCG)
  - Good behaviour of the metric
  - Adapted to several IR formalisms

- Work to be done...
  - More realistic user model parameters

- Shortcomings
  - Graded scale
  - Complexity