Information Retrieval Metrics

Useful Resources

1. https://en.wikipedia.org/wiki/Evaluation_measures_(information_retrieval)
2. Alternatives to Bpref
3. https://hosseiniec.wordpress.com/2014/07/28/retrieval-evaluation-with-incomplete-relevance-data/
4. http://people.cs.georgetown.edu/~nazli/classes/ir-Slides/Evaluation-12.pdf
5. https://link.springer.com/article/10.1007/s10791-008-9059-7
6. https://people.eng.unimelb.edu.au/jzobel/fulltext/acmtois08.pdf
7. https://pdfs.semanticscholar.org/6ed8/b5935be55fdc0eb2756fc376f73e098cbd39.pdf
8.

Document Retrieval

Good Docs: D1, D2, D3, D5, D8, D13, D21, D34

Bad Docs: D4, D6, D7, D9, D10, D11, D12, D14, D15, D16, D17, D18, D19, D20

Everything else unlabelled

Example order of retrieval and ranking:

D1 D100 D6 D8 D34 D101 D302 D9 D4 D10 D105 D20 D400 D13 D500 D21 D2

Metrics - Familiar Metrics - Precision

1. How many did the system get right?
2. Precision@k => |True positives till k| / k
3. 100% Precision? - All the retrieved k documents are relevant

Good Docs: D1, D2, D3, D5, D8, D13, D21, D34

Bad Docs: D4, D6, D7, D9, D10, D11, D12, D14, D15, D16, D17, D18, D19, D20

D1 D100 D6 D8 D34 | D101 D302 D9 D4 D10 D105 D20 D400 D13 D500 D21 D2

Precision@5 - ⅗

100% Precision possible order: D1 D13 D21 D8 D34 Precision@5 - 5/5

Metrics - Familiar Metrics - Average Precision

$$\text{AveP} = \frac{\sum_{k=1}^{n} (P(k) \times \text{rel}(k))}{\text{number of relevant documents}}$$

Good Docs: D1, D2, D3, D5, D8, D13, D21, D34

Bad Docs: D4, D6, D7, D9, D10, D11, D12, D14, D15, D16, D17, D18, D19, D20

D1 D100 D6 D8 D34 | D101 D302 D9 D4 D10 D105 D20 D400 D13 D500 D21 D2

Average Precision - 1/1 + 2/4 + 3/5 Precision@5 - 3/5

100% Precision possible order: D1 D13 D21 D8 D34 Precision@5 - 5/5

Metrics - Familiar Metrics - Recall

1. How many right answers did the system get?
2. Recall@k => |True positives till k| / |True Positives|
3. 100% Recall? - All the relevant documents are retrieved by the position k

Good Docs: D1, D2, D3, D5, D8, D13, D21, D34 - |Good Docs| = 8

Bad Docs: D4, D6, D7, D9, D10, D11, D12, D14, D15, D16, D17, D18, D19, D20

D1 D100 D6 D8 D34 | D101 D302 D9 D4 D10 D105 D20 D400 D13 D500 D21 D2

Recall@5 - 3/8

100% Recall possible order*: D1, D2, D3, D5, D8, D13, D21, D34

Metrics - Familiar Metrics - Recall

100% Recall order[*]: * D1, *, D2, *, D3, *, D5, *, D8, *, D13, *, D21, *, D34 *,

1. Any number of documents can appear in between
2. Order of relevant documents **doesn’t matter**
3. Recall@k can be 100% iff all labelled relevant documents are retrieved at k
4. 100% Precision doesn’t imply 100% Recall and vice versa
5. Optimize for P/R based on requirements of the downstream task

Metrics - Familiar Metrics - F Score

$$F_\beta = \frac{(1 + \beta^2) \cdot (\text{precision} \cdot \text{recall})}{(\beta^2 \cdot \text{precision} + \text{recall})}$$

1. Balances P&R
2. Higher beta - Recall is more imp
3. Common versions
   a. F1
   b. F2
   c. F1.5
   d. F0.5 - Precision is more imp

Metrics - Familiar Metrics - Success

1. Did the system get **at least one** correct at k?

D1 D100 D6 D8 D34 | D101 D302 D9 D4 D10 D105 D20 D400 D13 D500 D21 D2

Success@1 is 100%

Success@5 is 100%

Success@k is 100% for any k is 100% as the first doc is a good doc

Success@3 is 0% but Success@5 is 100% D9 D4 D10 D6 D8

Success is 0 for D9 D4 D10

Metrics - Familiar Metrics - nDCG

$$\mathrm{DCG_p} = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{\log_2(i+1)}$$

$$\mathrm{nDCG_p} = \frac{DCG_p}{IDCGp}.$$

1. Ranking metric
2. Graded Labels
3. Encourages to rank great results at the top
4. Condensed
    a. Calculate using only labelled points
5. Global - Ideal DCG - Use all the labelled data points
6. Local - Ideal DCG - Sorted retrieved labels

Metrics - Familiar Metrics - nDCG

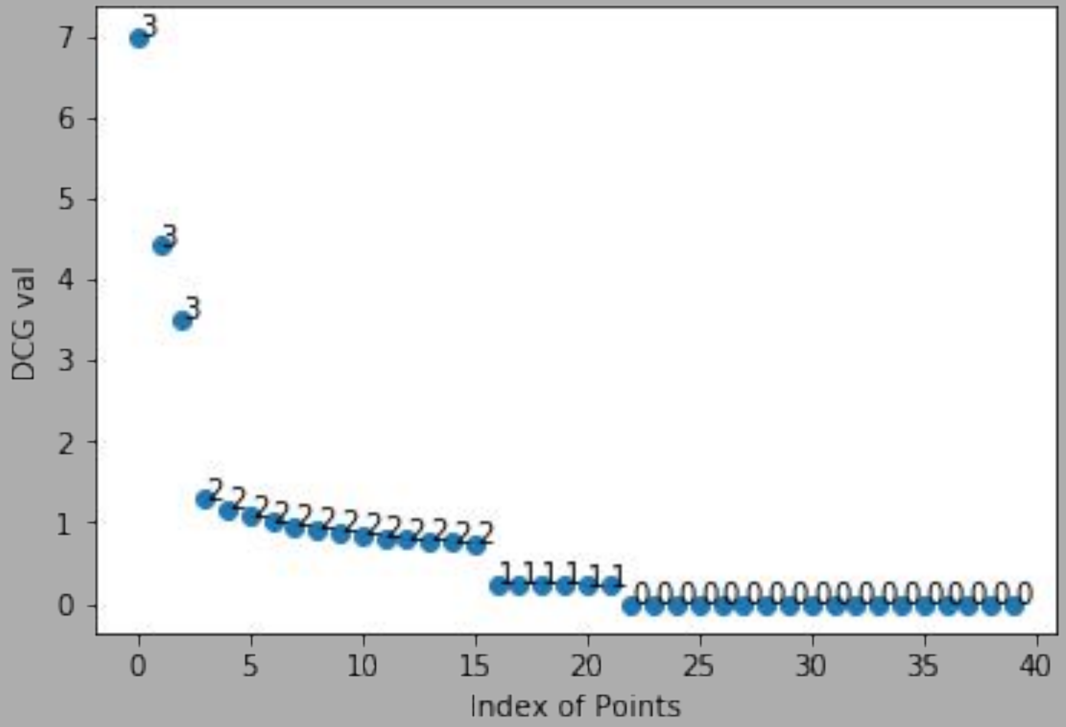
nDCG global 0.4274189765288345

nDCG global condensed 0.5244899222593769

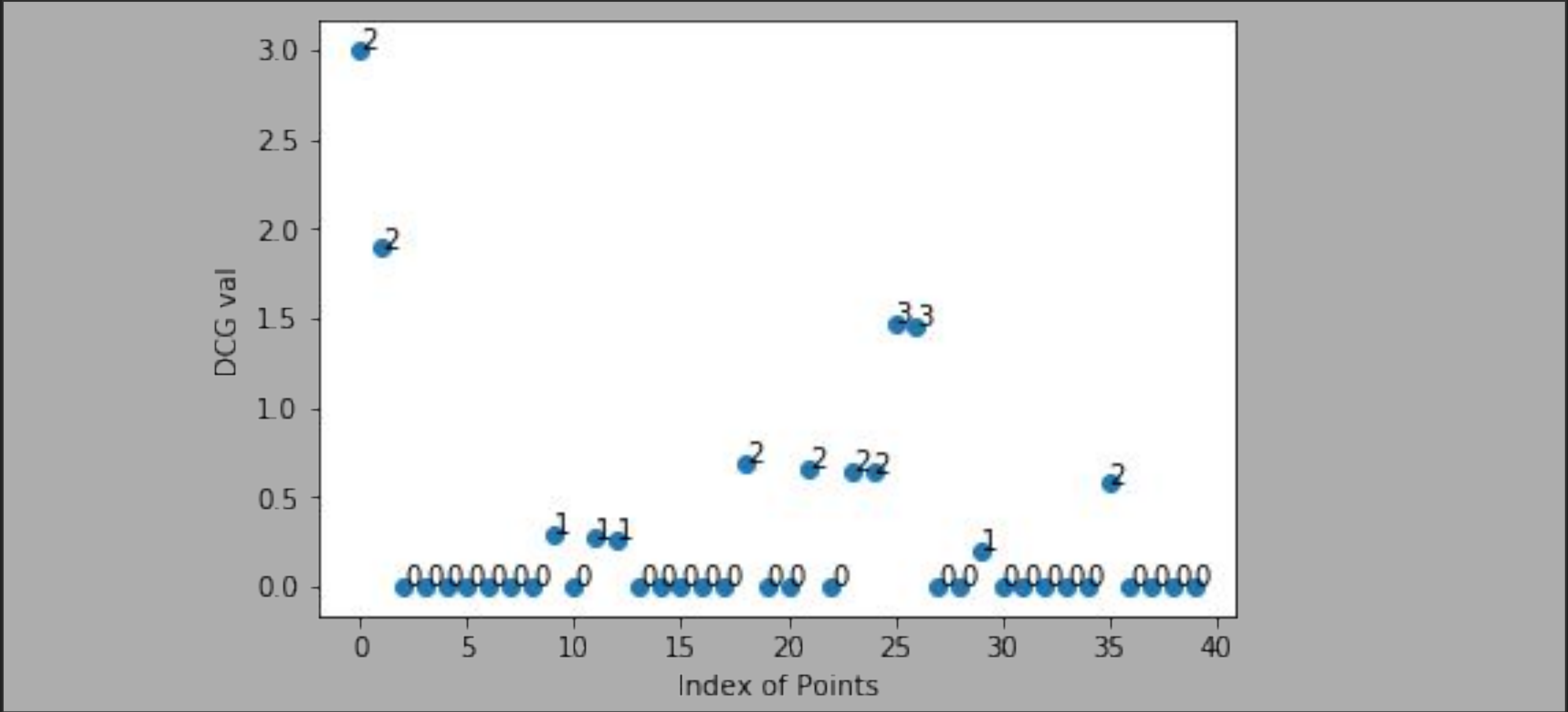
nDCG local 0.5916351066479533

nDCG local condensed 0.7260011092904043

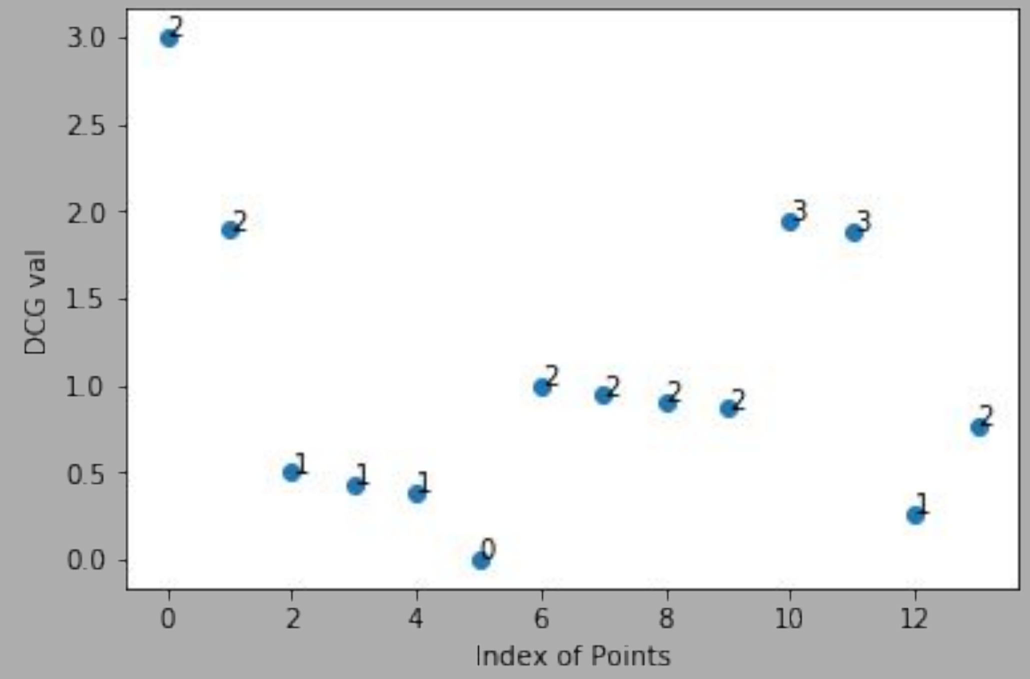
Metrics - Familiar Metrics - nDCG - Ideal BB



Metrics - Familiar Metrics - nDCG - Random Datum



Metrics - Familiar Metrics - nDCG - Random Datum - Condensed



Metrics - New Metrics - bpref

$$\text{bpref} = \frac{1}{R}\sum_{r} 1 - \frac{|n \text{ ranked higher than } r|}{R}$$

$$\text{bpref-10} = \frac{1}{R}\sum_{r} 1 - \frac{|n \text{ ranked higher than } r|}{10 + R}$$

1. Consider only **judged** documents
2. Approximately the % of wrongly ordered pairs
3. If **no labelled irrelevant** documents are retrieved
   a. bpref = Recall
4. If **no labelled relevant** documents are retrieved
   a. bpref = 0
5. If **all labelled relevant > all labelled irrelevant**
   a. bpref = Recall
6. Binary Labels

Metrics - New Metrics - bpref

Good Docs: D1, D2, D3, D5, D8, D13, D21, D34 - |Good Docs| = R = 8

Bad Docs: D4, D6, D7, D9, D10, D11, D12, D14, D15, D16, D17, D18, D19, D20

D1 D100 D6 D8 D34 | D101 D302 D9 D4 D10 D105 D20 D400 D13 D500 D21 D2

bpref@5 - 1/8 [(1 - 0/8) + (1-1/8) + (1-1/8)]

D6 D1 D100 D302 D8 D9 D34 D105 D4 D13

bpref@10 - 1/8 [(1-1/8) + (1-1/8) + (1-2/8) + (1-3/8)]

Metrics - New Metrics - bpref

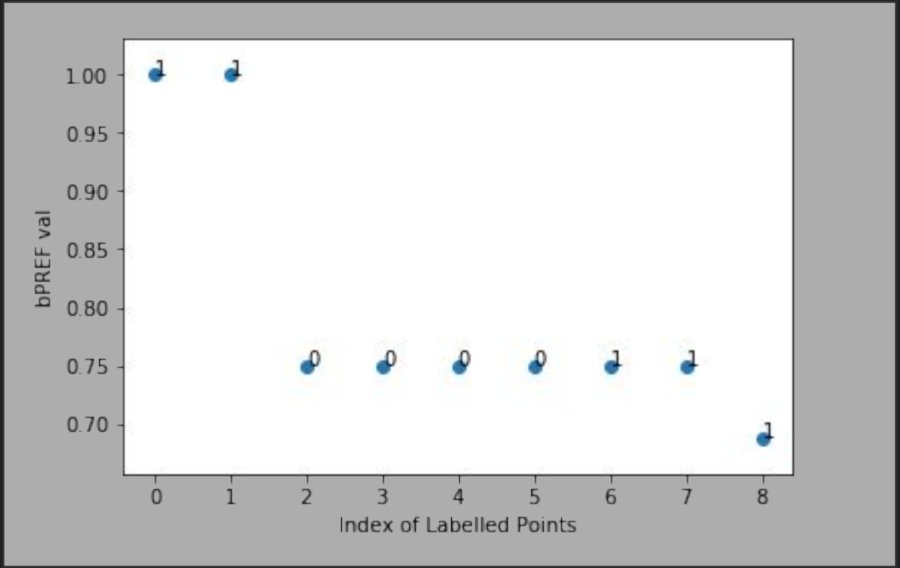
Condensed Labels

[1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 0, 1]

Num Labelled Docs = R = 16

B-pref - 0.45

Compared to Precision, a good document at a better rank is better



Metrics - New Metrics - Q-Measure

$$Q\text{-}measure = \frac{1}{R} \sum_{1 \leq r \leq L} isrel(r) \frac{\beta cg(r) + count(r)}{\beta cig(r) + r}$$

r - rank R - num of relevant documents

isrel(r) - Indicator flag 1 if labelled relevant, 0 for everything else

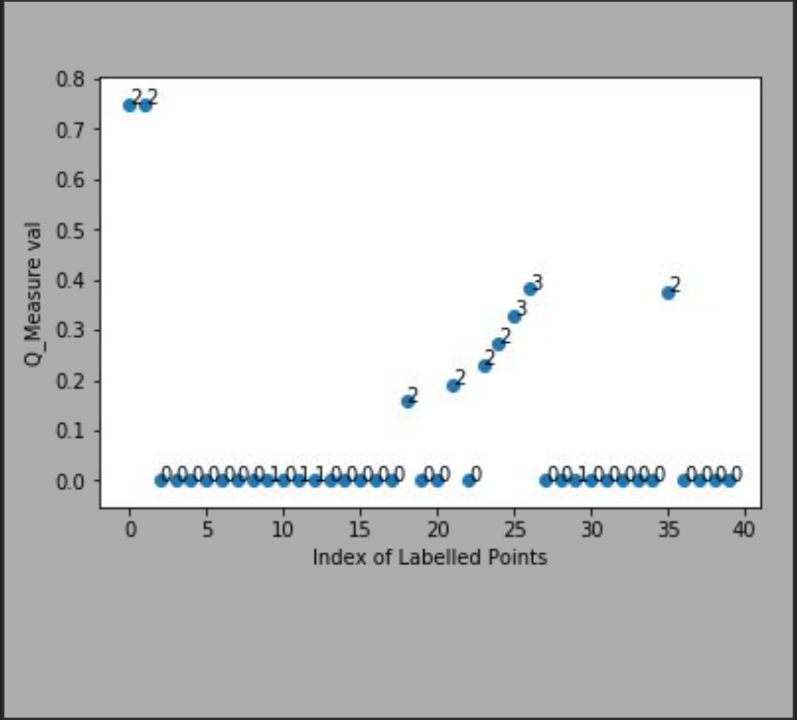
Gain - Quality of relevance - 3, 2, 1, 0

Cumulative Gain (cg) - gain(r) + cg(r-1)

Cumulative Ideal Gain (cig) - Sum of ideal order of labelled relevant qualities

count(r) - num of relevant docs till r

Metrics - New Metrics - Q Measure



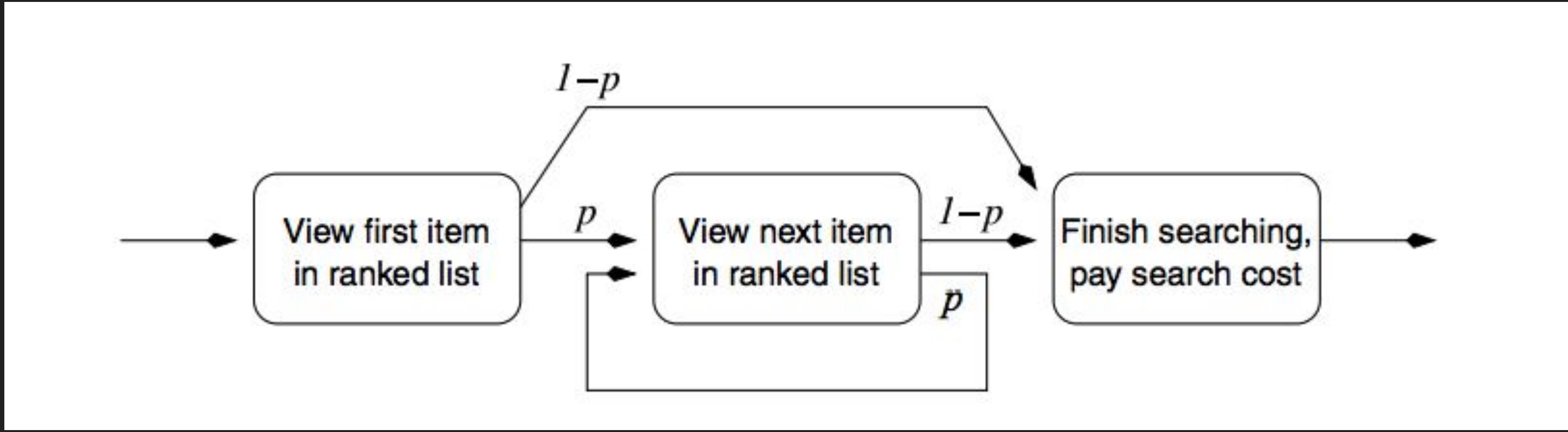
Labels - [2, 2, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 2, 0, 0, 2, 0, 2, 2, 3, 3, 0, 0, 1, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0]

Gain from a relevant document at a better rank is higher

Continuous/Multiple relevant labels increase the count in the numerator increases the value

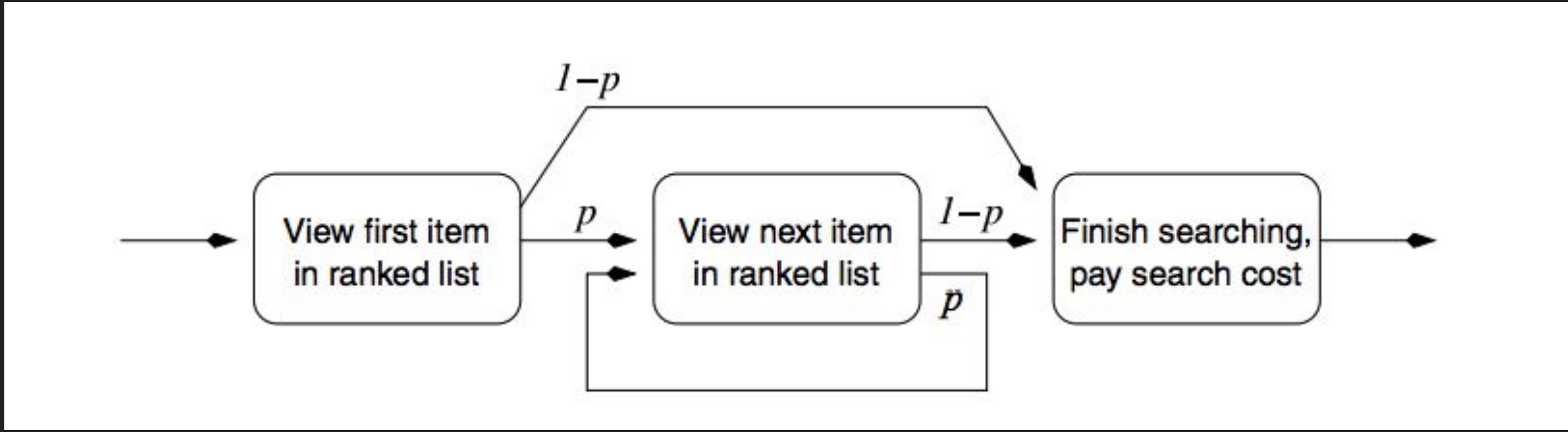
Metrics - New Metrics - Rank-Biased Precision(RBP)

1. User oriented metric
2. p - persistence/probability that they will look at the next doc



Metrics - New Metrics - Rank-Biased Precision(RBP)

1. User oriented metric
2. p - persistence/probability that they will look at the next doc



Metrics - New Metrics - Rank-Biased Precision(RBP)

$$\text{RBP} = (1 - p) \cdot \sum_{i=1}^{d} r_i \cdot p^{i-1}.$$

1. Higher P - User will scroll/load more - Patient - Exhaustive search habits
2. Lower P - Impatient - Top 1 - Feeling Lucky
3. Many relevant docs at the top
   a. Similar gains - Doesn't matter if the user keeps going
4. Some relevant docs at the top - Many at the bottom
   a. Patient users find more documents
5. Bad results
   a. No one find anything

Metrics - New Metrics - Rank-Biased Precision(RBP)

$$\mathrm{RBP} = (1-p) \cdot \sum_{i=1}^{d} r_i \cdot p^{i-1}.$$

7. 1-p normalizes the effort, and metric reflects rate of finding results

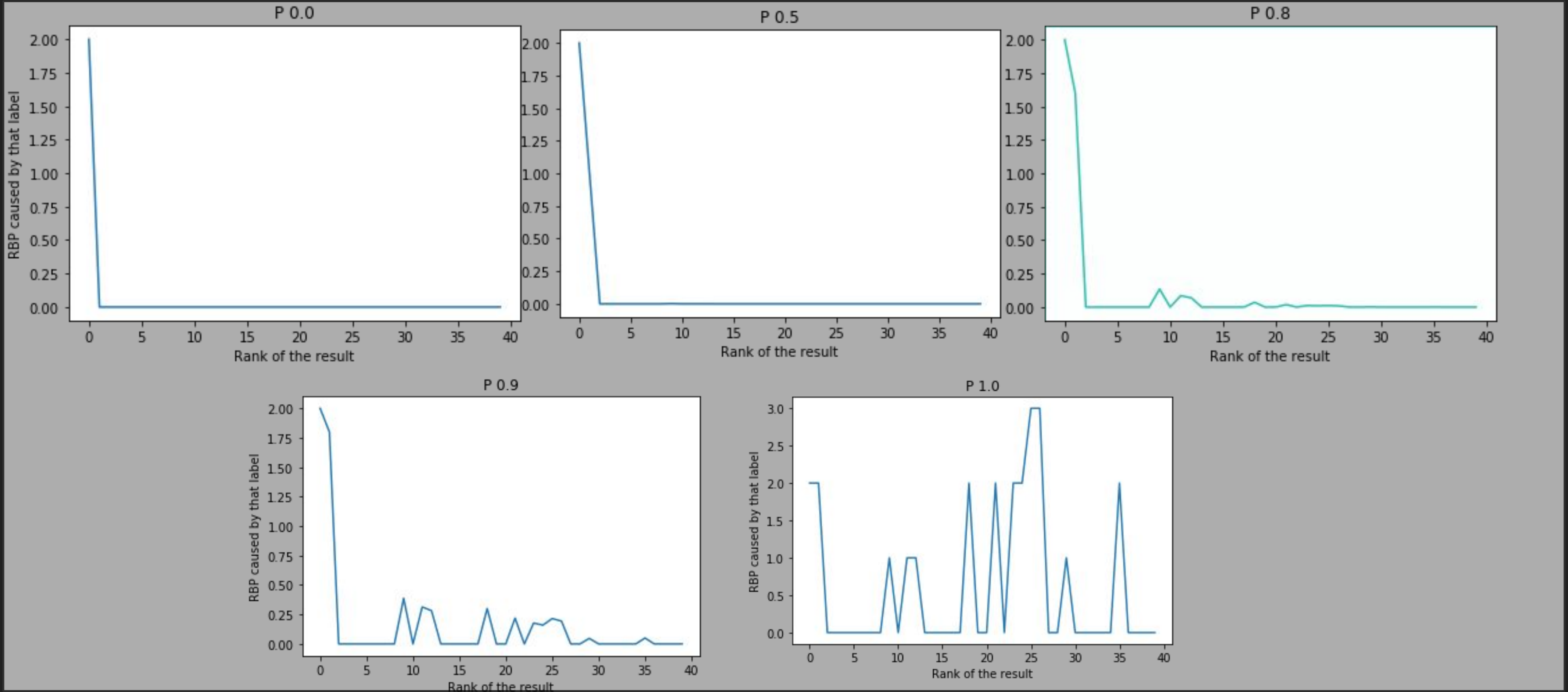
8. Higher RBP implies higher net satisfaction

9. Choosing p - Reverse engineer to gather impact from the kth result of choice.

10. # documents the user is likely to see

k=30 - Top 3 pages based on our user behavior

Metrics - New Metrics - RBP



Metrics - New Metrics - r-pref

$$rpref\_N = \frac{1}{R'} \sum_{D_k \in J,\ \phi_k > 0} \phi_k \left(1 - \frac{penalty_k}{N'}\right) \tag{2}$$

where

$$R' = \sum_{D_k \in J} \phi_k \tag{3}$$

$$N' = \sum_{D_k \in J} (1 - \phi_k) \tag{4}$$

and

$$penalty_k = \sum_{D_l \in J,\ r_l < r_k,\ \phi_l < \phi_k} \frac{\phi_k - \phi_l}{\phi_k}\ . \tag{5}$$

ϕ_k = [0, 1] - Normalized grades (TODO - Need to check)

Penalty - For a retrieved relevant document, compute penalty for every judged **less** relevant docs ranked above it

Normalized by ϕ_k implies that the penalties incurred by improper ranking of a great document is more than that of a good document

Directly equate to Recall if perfect ranking is achieved

Metrics - What does everything mean?

- High Precision - User is seeing relevant results - But can’t decide if the results are exhaustive
- High Recall - User is seeing relevant results interleaved with possibly irrelevant results
- High Precision & Recall doesn’t imply perfect ranking, as they are rank agnostic
- High Local nDCG - Great relevant results appear near the top - User is seeing relevant results - But can’t decide if the results are exhaustive
- Really High Global nDCG - Great relevant results appear near the top - User is seeing relevant results - More confidence that the majority of **labelled** relevant documents are retrieved

Metrics - What does everything mean?

- High b-pref
  - High Recall - Almost agnostic to retrieval of judged irrelevant documents
  - Good relative ranking - If there were multiple judged irrelevant documents
- Low b-pref
  - Low Recall - Low relevant labelled documents are seen
  - Lower - Bad Ranking
- High q-measure
  - High frequency of relevant labelled documents at higher ranks
- Low q-measure
  - Highly relevant labelled documents are found at lower ranks
  - Proportion of relevant labelled documents is lower - Not enough recall

Metrics - What does everything mean?

- High RBP - User finds **a** relevant result quickly based on their browsing habits
- Low RBP
  - Not enough relevant documents shown
  - Relevant documents are surfaced in later pages
  - Can take forever to find the relevant documents
  - User might end the search without seeing any relevant results
- High R-Pref
  - Highly relevant grades documents are ranked above judged lower graded documents
  - Lower irrelevant document recall if ranking is perfect
- Low R-Pref
  - Ranking mistakes
  - Only Low quality result recall